Dual-Stream Recurrent Convolutional Neural Networks as Models of Human Audiovisual Perception

by Michael Joannou

UNIVERSITY OF BIRMINGHAM

A thesis submitted to the University of Birmingham
for the degree of DOCTOR OF PHILOSOPHY

School of Psychology
College of Life and Environmental Sciences
University of Birmingham
March 2022

Abstract

Multisensory perception allows humans to operate successfully in the world. Increasingly, deep neural networks (DNNs) are used as models of human unisensory perception. In this work, we take some of the first steps to extend this line of research from the unisensory to the multisensory domain, specifically, audiovisual perception. First, we produce a highly-controlled, large, labelled dataset of audiovisual action events for human vs DNN studies. Next, we introduce a novel deep neural network architecture that we name a 'dual-stream recurrent convolutional neural network' (DRCNN), consisting of 2 component CNNs joined by a novel 'multimodal squeeze unit' and fed into an RNN. We develop a series of these architectures, leveraging a number of pretrained state-of-the-art CNNs, and train a number of instances of each, producing a series of classifiers. We find that, after optimising 12 classifier instances on audiovisual action recognition, all classifiers are able to solve the audiovisual correspondence problem, indicating that this ability may be a consequence of the task constraints. Further, we find that these classifiers are highly affected by signals in the unattended to modality during unimodal classification tasks, demonstrating a high level of integration across modalities. Further experiments revealed that dual-stream RCNN classifiers perform significantly worse than humans on a visual-only action recognition task when stimuli was clean or distorted by Gaussian noise or Gaussian blur. Both classifiers and humans were able to leverage audio information to increase their levels of performance in the clean condition, and to significantly decrease the effect of visual distortion on their audiovisual performances. Indeed, 5/6 classifiers performed within the range of human performance on clean audiovisual stimuli, and 3/6 maintained human level performance when low levels of Gaussian noise were introduced.

Acknowledgements

There are a number of people I would like to acknowledge for their support throughout the PhD, starting with my supervisors. Uta Noppeney, a source of great knowledge without whom this PhD would not have happened. With my computer science & engineering background, thank you for baptising me into the area of cognitive science. Bernd Bohnet, whose positive attitude helped me stay calm, thank you for our coffee meetings throughout the years. Pia Rotshtein, thank you for all of your helpful guidance and of course for reading the thesis. But most of all thank you for motivating and uplifting me whilst also encouraging me to leave my desk and socialise as the PhD took its hold. Thank you for your care and support.

Thank you also to my previous lab group, who left within the first 2 years of my PhD, but whose mutual support during that time was highly valued, and whose company made my PhD much more enjoyable. Thank you to Dr Dietmar Heinke and Dr Frank Guerin for kindly agreeing to examine this thesis. I would also like to thank the Engineering and Physical Sciences Research Council (EPSRC) for funding the PhD.

To my mum, Sue Jones, thank you for providing a caring voice when things were difficult, and for love and support throughout my education and PhD. I am sorry that you had to endure my PhD chatter over the past 4 years. Although I have not visited as much as I would have liked in the past 12 months, as the PhD came to an end, I promise to visit much more now. Thank you to both you and Frank for sharing your lovely seaside home for me to escape to and decompress, and for the lovely company. Frank, thank you for our nice walks, let's have many more.

Finally, I thank my girlfriend, Ioanna Papadaki, who has lived with me during the last year of the PhD. Thank you for bringing love and joy to my life and providing me with support and good company (and of course delicious Greek food) during a challenging period. Sorry for my

low energy and stress at times, particularly towards the end of the PhD, and for spending every evening at my desk, let’s go and enjoy some trips together.

Contents

List of Figures

List of Tables

LIST OF ABBREVIATIONS

- **ANN**: Artificial neural network
- **AVC**: Audiovisual correspondence
- **AVMIT**: Audiovisual Moments in Time (dataset name)
- **CNN**: Convolutional neural network
- **DNN**: Deep neural network
- **DRCNN**: Dual-stream recurrent convolutional neural network
- **MIT**: Moments in Time (dataset name)
- **RNN**: Recurrent neural network
- **RCNN**: Recurrent convolutional neural network
- **SC**: Superior colliculus
- **SVM**: Support vector machine
- **VEGAS**: Visually Engaged and Grounded AudioSet (dataset name)

CHAPTER 1

GENERAL INTRODUCTION

"We know of no animal with a nervous system in which the different sensory representations are organised so that they maintain exclusivity from one another." - Stein and Meredith, 1993

"If you want to understand a really complicated device, like a brain, you should build one." - Geoffrey Everest Hinton, 2018

To perceive the world, humans make use of information across a number of sensory modalities. The process of extracting useful signals from noise and combining these signals across modalities is not a simple one. Indeed, our understanding of these mechanisms in the animalian brain has grown since Aristotle first pondered sensory mechanisms (384–322 B.C.) but it is still growing today.

More generally, in the quest to further understand sensory perception, some have looked to Deep Neural Networks (DNNs). DNNs are now capable of human levels of performance on a number of unisensory recognition tasks (Krizhevsky et al., 2012; Cireşan et al., 2012; Wan et al., 2013; Sun, Chen, et al., 2014; Taigman et al., 2014; Russakovsky et al., 2015; He et al., 2015; McLoughlin et al., 2015; Zhang, McLoughlin, et al., 2015; Phan et al., 2016; Takahashi et al., 2016; Laffitte et al., 2016; Parascandolo et al., 2016). This has been followed by the use of DNNs to explore human visual perception (Dodge and Karam, 2016; Dodge and Karam, 2017; Wichmann et al., 2017; Geirhos, Temme, et al., 2018; Dodge and Karam, 2019; Stabinger et al., 2016; Heinke et al., 2021; Yamins, Hong, Cadieu, and Dicarlo, 2013; Yamins, Hong,

Cadieu, Solomon, et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014; Güçlü and Van Gerven, 2015; Cichy, Khosla, et al., 2016), human auditory perception (Kell, Yamins, et al., 2018) and multisensory perception (Rideaux et al., 2021).

Much of the DNN literature is centred on unisensory learning. Particularly in the area of video recognition, the majority of action recognition leaderboard positions, for benchmarks such as Kinetics-700 (Smaira et al., 2020) and Moments in Time (Monfort et al., 2019), are visual-only solutions. But although there has been a focus on unisensory tasks, there have been a number of DNNs in the literature realising multisensory performance gains on tasks such as speech recognition (Petridis et al., 2017), voice activity detection (Tao and Busso, 2017), speech separation (Gogate et al., 2018), emotion recognition (Zhang, Wang, et al., 2019) and action recognition (Nagrani et al., 2021; Akbari et al., 2021).

In this work, I wish to extend the area of study comparing DNNs and humans to the audiovisual action recognition domain by developing audiovisual DNNs and exploring their behaviour. The following chapter introduces the literature to support and motivate the empirical work documented in this thesis. I first describe multisensory integration and its basic principles. Next, I describe the core operation of artificial neural networks. I then consider the relationship between artificial neural networks and human intelligence. Finally, I provide an overview of this thesis.

1.1 Multisensory Integration in Biological Life

The use of multiple sensory modalities considerably increases an organisms chance of survival and success by providing more information about the environment or substituting for one another when information in one modality is unreliable (Stein and Meredith, 1993). When provided with multiple ‘views‘ of the environment, an organism must combine this information into a coherent percept such that it can operate effectively in its environment (Stein and Meredith, 1993). The drive to combine signals effectively results in neural and behavioural distinctions between multisensory and unisensory processing in the brain (Stein, 2012). The study of this

phenomena is known as multisensory integration. In this section, we outline the multisensory causal inference problem, the principles of multisensory integration and behavioural responses to multisensory stimuli.

1.1.1 Principles of Multisensory Integration

Our sensory organs provide us with several 'views' of the environment that can be used simultaneously or can provide recourse when some sensory data becomes unreliable (for instance touching a wall to navigate down a hallway in the dark). In order to effectively combine signals from different sensory modalities, a sensory system must first solve the correspondence problem (otherwise known as the binding problem or causal inference problem) (Shams and Beierholm, 2010), namely, which sensory signals have a common source and should be integrated and which signals are from separate sources and should be segregated. In order to solve the problem, humans make use of a number of cues. When signals arriving from different sensory modalities are approximately synchronous and occur in close spatial proximity to one another, they are more likely to have a common source. In the case that unisensory signals have large spatial or temporal disparities, they are less likely to have a common source. These spatiotemporal cues are made use of by the human brain (Munhall et al., 1996; Slutsky and Recanzone, 2001; Lewald and Guski, 2003; Wallace et al., 2004) alongside higher-order cues (Laurienti et al., 2004; Parise and Spence, 2009; Calvert et al., 2000; Doehrmann and Naumer, 2008; Noppeney, Ostwald, et al., 2010; Krugliak and Noppeney, 2016). The temporal and spatial rules are considered fundamental principles of multisensory integration and are known as the *temporal rule* and *spatial rule* respectively. Indeed a considerable amount of work has explored the spatiotemporal effects of multisensory stimuli in the superior colliculus of rodents and cats (Stein and Meredith, 1993).

In the case that signals from different modalities are judged to have a common cause, human adults have been shown to integrate the information near-optimally in accordance with Maximum Likelihood Estimation (MLE) (Ernst and Banks, 2002). MLE provides an 'ideal observer' model of multisensory integration whereby redundant information is weighted according to its reliability

and integrated to provide an unbiased estimate (this could be an estimate of the spatial position for instance) (Ernst and Banks, 2002; Alais and Burr, 2004). Further, where all component unisensory data is unreliable, it is integrated to a greater extent, reducing the multisensory variance below that of the unisensory variances. This is known as the principle of inverse effectiveness (PoIE) and is considered a third fundamental principle of multisensory integration. Indeed, in this case, there are clear increases in perceptual salience and behavioural performance. Using the MLE model, however, follows the 'unity assumption', the assumption that the signals have a common cause (Welch and Warren, 1980). In recent years, Bayesian Causal Inference (BCI; Körding et al., 2007) models of multisensory integration have provided a mathematical description of how observers solve the binding problem according to the causal uncertainty and sensory noise. Like the MLE model, the BCI model provides an ideal observer that can provide multisensory location estimates while also taking into account the causal structure of the stimuli (whether the stimuli had a common cause of different causes). Using the BCI model framework, the unity assumption exists as the prior of common cause. The MLE and BCI models are not explored in this work, their descriptions serve to communicate to the reader that humans are highly optimised to solve the binding problem and integrate sensory data.

1.1.2 Multisensory behaviour

Integration of redundant signals from multisensory stimuli provides a number of behavioural benefits. For instance, for a number of decades it has been widely accepted that bisensory stimuli speed up reactions (Hershenson, 1962; Morrell, 1967; Gielen et al., 1983; Diederich and Colonius, 2004). This reduction in reaction time is known as the redundant signals effect (RSE). It is worth noting that the RSE can be partially explained by *statistical facilitation* (also known as *probability summation*) (Raab, 1962) which could be simply described as 'always taking the fastest unimodal answer' in a multimodal task. This was described by Raab (1962) using a separate-activation model (or *race model*) that builds up separate activations for the stimuli on each channel until one reaches a criterion and produces a response, the reaction time is thus the shortest of those individual reaction times on every trial, and faster than either unimodal response

overall. Others have shown race models not to adequately describe the RSE on a number of tasks (Miller, 1982; Diederich, 1995; Townsend and Nozawa, 1995; Miller, 2004), instead opting for *coactivation models* that build up a pooled activation using the signals from both modalities and results in shorter reaction times than either of the reaction times to the individual component modalities alone. The race model, however, still provides a useful 'baseline winner's advantage' (Miller, 2016).

Faster response time are not the only advantage of multisensory stimuli, for example humans are able to locate multisensory stimuli faster than unisensory stimuli (Alais and Burr, 2004; Wallace et al., 2004) and recognise items more accurately (Giard and Peronnet, 1999; Molholm et al., 2004; Stefanics et al., 2005). Humans and other animals make use of a number of cues in order to enjoy these performance benefits, beyond just the spatiotemporal. Where in the last section we described the spatial rule and temporal rule, explaining how these dimensions affect the strength of integration, we only briefly mentioned higher-order cues. Indeed, the spatial rule and temporal rule are well-established laws of multisensory integration (Stein and Meredith, 1993), while the effects of higher-order cues, such as *semantic congruence*, on multisensory integration (Laurienti et al., 2004) are less established. In this realm, semantic congruence refers to the matching of semantic meaning across sensory modalities (for example the sound of a dog bark and the image of a dog), *incongruence* then refers to the mismatch of sensory content (the sound of a dog bark is presented alongside the image of a cat). It has been shown that semantically congruent multisensory stimuli lead to increases in behavioural performance over unisensory stimuli (Laurienti et al., 2004; Molholm et al., 2004). Although these performance benefits could also be modulated by attending to a single modality. For instance, Yuval-Greenberg and Deouell (2007) found that there was an interaction between attended to modality and semantic congruency, whereby multisensory benefit was amplified when the auditory modality was attended to on an object-recognition task. This could be due to the *modality appropriateness hypothesis* (Welch and Warren, 1980) as the visual modality likely carries more useful information for object-recognition than the audio modality. The accuracy benefits of these semantic cues are of interest throughout all study chapters in this work.

1.2 Artificial Neural Networks

Where the previous section summarised the ideas within the field of multisensory integration that will be explored in this work, this section provides readers with an overview of artificial neural network (ANN) models. It is not intended to provide a comprehensive educational resource, but rather provide a basis of understanding of the general concepts and ideas. This section outlines the component artificial neuron, neural network architectures and the backpropagation learning algorithm.

1.2.1 The artificial neuron

Artificial neural networks (ANNs) are connectionist models consisting of artificial neurons organised in layers. These layers can have many different configurations, organised such that signals travel from the input layer to the output layer. Each artificial neuron has a set of inputs; either a portion of the model input (such as image pixel values in an image recognition task) or the outputs of some preceding neurons. Each artificial neuron performs a weighted sum of its inputs (and a bias term) before passing the value through an activation function to produce a single-valued output, or 'activation'. This activation can then be used as the input to another artificial neuron or used as model output. For classification problems, the output of an ANN is a probability distribution over the possible classes.

The activation function in a modern artificial neuron is often a non-linear function, but this was not always the case for artificial neural networks. Current artificial neurons are an extension of the single-layer perceptron (Figure 1.1) introduced by Rosenblatt (Rosenblatt, 1957) and based on the McCulloch-Pitts neuron (McCulloch and Pitts, 1943). The original single-layer perceptron is a linear supervised learning algorithm, capable of learning binary classification problems. Specifically it is capable of learning a threshold function (Equation 1.1).

$$f(x) = \begin{cases} 1 & \text{if } w \cdot x + b > 0 \\ 0 & \text{otherwise} \end{cases} \tag{1.1}$$

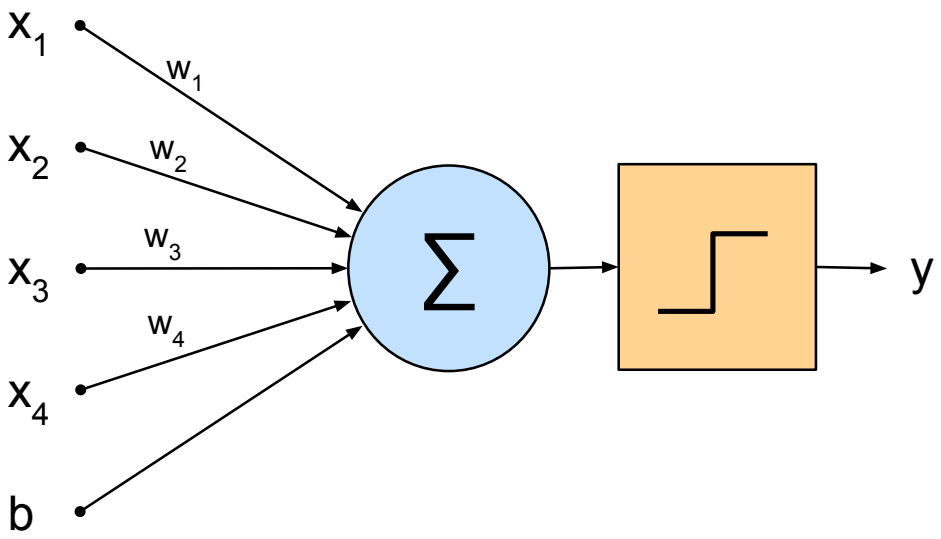


Figure 1.1: The single-layer perceptron. The model performs a dot product between the input vector, $[x_1, x_2, x_3, x_4]$, and the weights vector, $[w_1, w_2, w_3, w_4]$. The model then sums the weighted inputs alongside a bias term, b. The sum is then passed through a Heaviside step function. The output y is a binary value pertaining to 1 of the 2 possible classes.

Where $f(x)$ is the single-layer perceptron output, w is the learned weights vector, x is the input vector and b is the learned bias term. As this model is only capable of learning linearly separable problems, it was unable to solve the exclusive-or (XOR) problem, leading to a loss of interest in this area of research and the so-called 'AI winter'.

By combining 2 or more layers of artificial neurons (3 layers including the input nodes) and replacing the Heaviside step function with another non-linear activation function (or 'non-linearity') we can build artificial neural networks that can be trained with backpropagation (Rumelhart et al., 1986) and gradient descent. These models are known as deep neural networks (DNNs) and allow us to overcome the previous problems of the perceptron. The Heaviside step function must be replaced as it is non-differentiable (backpropagation would not be able to calculate the partial derivatives) at x=0 and has a gradient of zero at every other point (gradient descent would not be able to update the weights). This combination of backpropagation and gradient descent are further described in Section 1.2.4. The activation, a_i, of an artificial neuron,

i, with activation function, f, can thus be calculated by finding the dot product of the weights vector, W_i, and the inputs, X_i, plus the bias term, b_i (Equation 1.2).

$$a_i = f(W_i \cdot X_i + b_i) \tag{1.2}$$

Traditionally the non-linearities used in DNNs were simple, saturating functions such as the sigmoid function (Equation 1.3; Hinton et al., 2012) and hyperbolic tangent. The Universal Approximation Theorem shown that any feed-forward neural network with at least 1 *hidden layer* (a layer between the input and output layer of an ANN), using saturating activation functions with sufficient neurons in its hidden layer is capable of approximating any continuous function between two Euclidean spaces (Cybenko, 1989; Hornik et al., 1989). The long-standing issue, however, with the use of these saturating functions as activation functions in DNNs is that gradient updates become problematically small around the saturating zones (the so-called 'vanishing gradient problem').

$$f(x) = \frac{1}{1+e^{-x}} \tag{1.3}$$

The non-saturating ReLu activation function (Equation 1.4) allowed DNNs to be trained faster and in an end-to-end fashion (Krizhevsky et al., 2012; He et al., 2016). Indeed, deep neural networks have far surpassed the state-of-the-art performance in a number of domains such as; object recognition (Krizhevsky et al., 2012), hand-written digit recognition (Cireşan et al., 2012; Wan et al., 2013) and face recognition (Sun, Chen, et al., 2014; Taigman et al., 2014). The success of deep neural networks (DNNs) has also extended into other domains, including but not limited to audio event recognition (Zhang, McLoughlin, et al., 2015; McLoughlin et al., 2015) with DNNs surpassing the state of the art (Takahashi et al., 2016; Phan et al., 2016; Laffitte et al., 2016; Parascandolo et al., 2016).

$$f(x) = max(0,x) \tag{1.4}$$

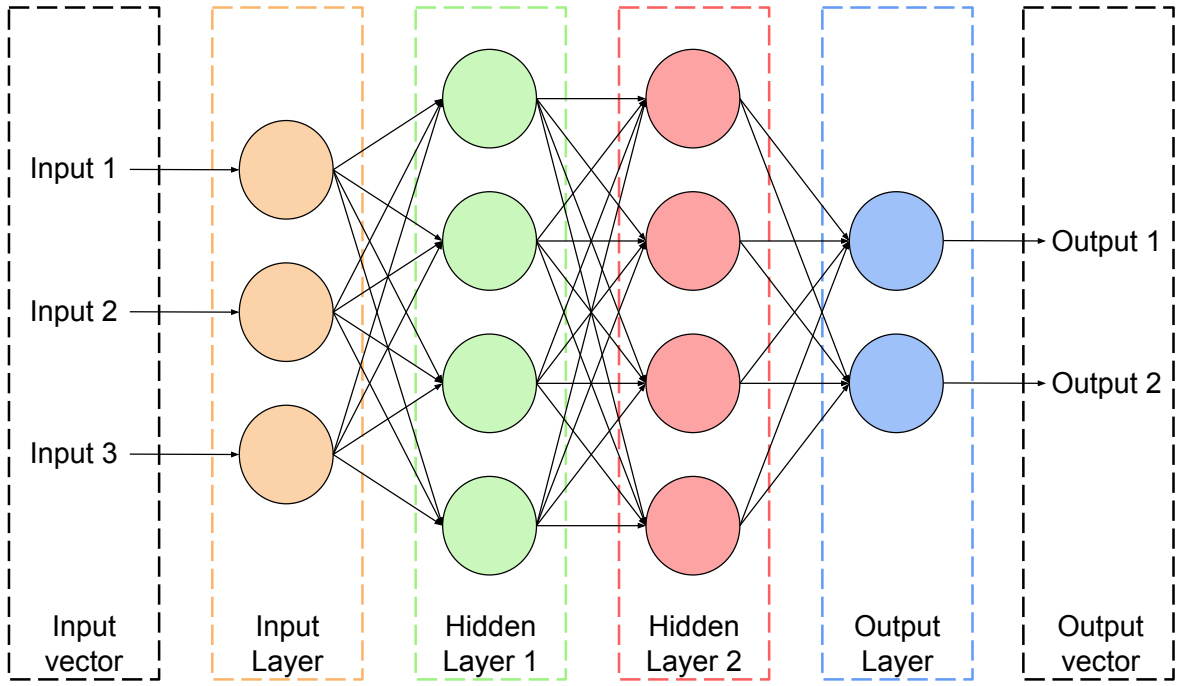


Figure 1.2: A multi-layer perceptron with 2 hidden layers. Each circle represents an artificial neuron. Connections between neurons flow from each layer to every neuron in the consecutive layer.

1.2.2 Artificial neural network architectures

The artificial neurons described in the last section are the building blocks of modern artificial neural networks. In this section, we describe how these neurons can be configured into ANNs. Broadly speaking, artificial neurons can be organised in two different ways; feedforward or recurrent. Feedforward neural networks have purely feedforward connections, these are connections travelling in one direction, from input to output, without any feedback connections. These are used extensively in pattern recognition with clear input-output pairs (such as image inputs and object label outputs in an object recognition task). There are a number of different types of feedforward ANN extending from the multi-layer perceptron (MLP, Figure 1.2) to the more modern convolutional neural network (CNN) (LeCun, Boser, et al., 1989; Lecun et al., 2015) or transformer (Vaswani et al., 2017) models.

Of particular interest in this work are CNNs. LeCun, Boser, et al. (1989) detailed one of the first neural networks with convolutional layers that could work directly from pixels, it was a zip code recognition system and was also one of the first ever practical applications of

backpropagation. CNNs have since been widely used for image recognition tasks (Krizhevsky et al., 2012; Cireşan et al., 2012; Wan et al., 2013; Sun, Chen, et al., 2014; Tan and Le, 2019) and are used throughout our work in Sections 2, 4 and 5. A CNN is a neural network with a constrained architecture, largely characterised by the component *convolutional layers* and *pooling layers* with the model terminating with some number of fully connected layers.

A convolutional layer contains artificial neurons whose receptive fields are local such that each neuron only receives a spatially local portion (either an image patch or a group of activations from the preceding layer of neurons) of the input (Lecun et al., 2015). The receptive field of each neuron overlaps the receptive field of neighbouring neurons in the same layer, such that together, the neurons in the convolutional layer cover the entire input space (Lecun et al., 2015). This is in contrast to the artificial neurons of the previously described MLP whereby each neuron was connected to all preceding neurons. CNN neurons are organised into filters, with each filter consisting of neurons with identical receptive fields spanning the entire input space. In this way, after learning, these filters are each selective for a particular feature in the input space. With numerous filters, these layers are 3D and capable of detecting a selection of features across the input. Convolutional layers are shift equivariant operations (where stride is equal to 1, above which these layers are approximately equivariant), with the output of the layer equivalent whether the translation is made to the input prior to the operation or afterwards according to Equation 1.5.

$$F(T(x)) = T(F(x)) \tag{1.5}$$

Where F is the convolutional layer operation, T is the translation function and x is the input to the convolutional layer.

By stacking convolutional layers (with other operations, notably pooling, interspersed) the receptive fields effectively cover increasingly large portions of the input and so detect increasingly higher-level features.

Pooling layers are another important characteristic of a CNN. They provide a representation of each feature at each location by subsampling. This reduces the size of the representation and reduces the number of necessary parameters (and thus the amount of required computation).

By pooling the activations of neighbouring convolutional layer neurons, the pooling operation provides invariance to small translational shifts according to Equation 1.6

$$P(T(x)) = P(x) \tag{1.6}$$

Where P is the pooling layer operation, T is the translation function (for shifts smaller than the pooling kernel) and x is the input to the pooling layer.

The output of the pooling layer is equivalent only if the corresponding shift in activations fall within the same pooling kernel (and so must be smaller than the kernel size). This is because shifts in the input will still be captured by the same pooling neurons, and still give the same output (either the maximum or average activation).

In contrast to feedforward neural networks, recurrent neural networks (RNNs) are those models with feedback connections. In this way, signals travel in both directions between the input and output layers. There are a number of configurations for RNNs, but they all have a common feature, *hidden states* are fed back into the model to be used at the next timestep. By means of this feedback connection, RNNs maintain a state throughout the input sequence, altering this state according to new input information. This is the reason they are used for input sequence problems. There are a number of RNN architectures, extending from the Fully Recurrent Neural Network (FRNN, or 'vanilla' RNN, Figure 1.3) to more modern Gated Recurrent Units (GRUs) and Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) networks. Feedforward models can be used to solve problems with sequential data too, but their fixed input size and lack of internal state necessitates building a model for particular sequence lengths, or using engineering solutions such as padding. For very large sequences, feedforward models are impractical.

To find the hidden state and output of an FRNN (the most simple case), the equations are, understandably, similar to those of a simple feedforward artificial neuron (Equation 1.2). In fact, one may set the weights matrix associated to the feedback connection to be a zero matrix, in which case Equation 1.7 is equal to Equation 1.2 where the activation function is a hyperbolic tangent.

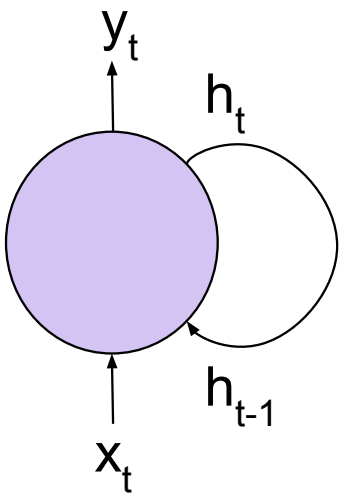


Figure 1.3: RNN Schematic. With input X_t at timestep t, receiving the hidden state h_{t-1} from the previous timestep $t-1$ and producing the output y_t at timestep t. The model updates the hidden state according to x_t and sends the new hidden state h_t through its feedback loop, ready for the next timestep.

$$h_t = \tanh\left(W \cdot x_t + U \cdot h_{t-1} + b\right) \tag{1.7}$$

Where h_t is the hidden state at time t and h_{t-1} is the hidden state at the prior timestep $t-1$. W is the input-to-hidden weight matrix and U is the hidden-to-hidden weight matrix. x_t is the input at timestep t and b is the bias value. For an FRNN, the hidden state at each timestep is also the output at that timestep.

1.2.3 Recurrent convolutional neural networks

The CNN and RNN architectures, described in the previous section, have been combined in a number of ways by researchers attempting to model spatiotemporal data (Donahue et al., 2015; Shi et al., 2015; Tsironi et al., 2016; Ning et al., 2017; Çakır et al., 2017; Yang et al., 2019; Sabir et al., 2019; Khaki et al., 2020; Gupta et al., 2021). Collectively, these are referred to as recurrent convolutional neural networks (RCNNs). One particular RCNN configuration, of interest in this work, consists of a CNN model to extract spatial features at each time-step of a spatiotemporal data sequence and an RNN to resolve over the temporal dimension (Donahue et al., 2015; Tsironi et al., 2016; Ning et al., 2017; Çakır et al., 2017; Yang et al., 2019; Sabir et al., 2019; Khaki et al.,

2020; Gupta et al., 2021). In this way, RCNNs can be built that are deep in both the spatial and temporal dimensions. As such, the model is particularly well suited to video processing tasks, where the CNN component provides embeddings at each time point and the RNN component provides a video level embedding. Indeed, these are the models implemented and studied in this thesis (Chapters 4 and 5). Given enough computational resources and data, RCNNs are end-to-end trainable for classification tasks (Donahue et al., 2015).

Where feedforward models require fixed-size input, RNNs do not. This means that those building feedforward ANNs for spatiotemporal modelling tasks must utilise engineering workarounds such as input data padding or windowing. Feedforward models spanning the temporal dimension scale poorly with input layer size growing according to the maximum sequence length. By utilising both CNNs and RNNs in an RCNN, however, one may enjoy the convenience of both for spatiotemporal modelling tasks. The spatial features are still abstracted using state-of-the-art CNNs but the RNNs allow the model to process variable length inputs and provide variable length outputs without these engineering workarounds.

1.2.4 Learning

Previous sections described the operations and architectures of ANNs, in this section we provide some description of how those models are trained. The weights (used to multiply input values during the weighted sum) and biases of an ANN are learned parameters. During learning (known as 'training'), ANNs are tuned on input-target pairs. For *supervised learning* problems, these targets are labels, and for the classification problems explored in this work, those labels are discrete (rather than continuous in the case of a regression problem). The model outputs can be considered probability distributions, with each value corresponding to the probability of a particular class. This is produced by a final softmax activation function (Equation 1.8) in the model.

$$o_j = \frac{e^{z_j}}{\sum_{i=1}^{N} e^{z_i}} \tag{1.8}$$

Where o_j is the softmax output (model output) of neuron j corresponding to a single class, z_j is the input to the activation of neuron j (the 'preactivation') and N is the number of neurons in the softmax layer (the number of classes). The loss can then be calculated as a measure of the distance between the actual and the desired output of the model. The categorical cross-entropy loss function provides a measure of loss between output distribution and the target distribution (Equation 1.9).

$$L = -\sum_{i=1}^{N} y_i \cdot \ln o_i \tag{1.9}$$

Where L is the loss, y_i is the target output of neuron i for one value of the output (corresponding to a class). The loss calculated here provides a measure of how distinguishable the 2 discrete probability distributions are from each other (the output distribution and the target distribution). For single-label classification problems (where each sample has a single ground-truth) the target distributions are 'one-hot encodings', consisting of a vector of 0s for all classes other than a single value of 1 at the position corresponding to the ground-truth.

Finding the optimal set of parameters to minimise this loss forms an optimisation problem. However, as ANNs are made up of a number of organised transformations, we can think of them as large composite functions, with each layer depending on the output of the previous layer. One implication of this, is that even the composition of two convex functions is not necessarily convex, complicating the optimisation problem. Consider a simple, three-layer neural network (with one hidden layer). Permuting the neurons in the hidden layer (and making the corresponding change in the output layer) would still give the same model output. This holds at any local minima, thus there are several minima with the same value. In this way, the loss surface is non-convex.

In order to minimise the loss, we may find the relationship between each trainable parameter and the loss (the partial derivative of the loss with respect to that parameter) and then tune the parameter by some small value in the direction of the negative gradient. This is the well established combination of backpropagation and mini-batch gradient descent (Rumelhart et al., 1986; Goodfellow, Bengio, et al., 2016). The training data is shuffled and organised into mini-batches. For each example in the mini-batch, the data is input to the model (the 'forward pass'),

giving a loss value for that example. The backward pass uses backpropagation as a method of finding the partial derivatives of the loss with respect to the model output activation, and then using the chain rule (Equation 1.10) to find partial derivatives of the loss with respect to the trainable parameters earlier in the model.

$$\frac{\partial y}{\partial x} = \frac{\partial y}{\partial u} \cdot \frac{\partial u}{\partial x} \tag{1.10}$$

The chain rule allows backpropagation (Rumelhart et al., 1986) to take steps from the output to the input of an ANN (the 'backwards pass'), using previously calculated partial derivatives to find new partial derivatives. In the context of ANNs, we can implement this to find the partial derivative of the loss function with respect to the preactivations of the softmax layer (Equation 1.11). In doing so, we must consider the individual partial derivatives according to each output neuron.

$$\frac{\partial L}{\partial z_j} = \sum_{i=1}^{N} \frac{\partial L}{\partial o_i} \cdot \frac{\partial o_i}{\partial z_j} \tag{1.11}$$

The calculated gradients for each parameter are calculated across examples in the mini-batch. Finally there is a parameter update step whereby gradient descent is used to update each parameter in the direction of the negative (averaged across mini-batch) gradient. Starting at the output of the model, the partial derivative of the loss function (Equation 1.9) with respect to a particular softmax output, i, can be obtained (Equation 1.12).

$$\frac{\partial L}{\partial o_i} = -\frac{y_i}{o_i} \tag{1.12}$$

Finding the partial derivative of the softmax function is a little more involved. The quotient rule for derivatives provides us with a clear method to find the derivative of the softmax but there are 2 possible outcomes according to whether $i = j$ vs $1 \neq j$ (the effect of the corresponding logit to the output of the softmax is much larger than the effect of the other logits). In the case where $i = j$, the derivative of e^{z_j} (the numerator of Equation 1.8) with respect to z_i is e^{z_j}, but when $i \neq j$, the derivative of e^{z_j} with respect to z_i is now 0 (because it is now a constant). This

means that we need to calculate for both cases (Equation 1.13).

$$\frac{\partial o_i}{\partial z_j} = \begin{cases} o_i(1-o_i) & \text{if } i = j \\ -o_i \cdot o_j & \text{if } i \neq j \end{cases} \tag{1.13}$$

Where i and j are used to refer to softmax neurons (in the case when they are the same or different neurons). Where i and j refer to the same neuron, we can find the product of the following partial derivatives (Equation 1.14).

$$\frac{\partial L}{\partial o_j} \cdot \frac{\partial o_j}{\partial z_j} = -y_j + o_j \cdot y_j \tag{1.14}$$

Where i and j refer to different neurons, we can find the product of these partial derivatives (Equation 1.15).

$$\frac{\partial L}{\partial o_i} \cdot \frac{\partial o_i}{\partial z_j} = y_i \cdot o_j \tag{1.15}$$

From this we have the summation to solve Equation 1.11.

$$\frac{\partial L}{\partial z_j} = \sum_{i=1}^{N} \frac{\partial L}{\partial o_i} \cdot \frac{\partial o_i}{\partial z_j} = \sum_{i=1}^{N} y_i \cdot o_j - y_j \tag{1.16}$$

As the target, y is a one-hot vector, whose summed elements always equals 1, we can write (Equation 1.17).

$$\frac{\partial L}{\partial z_j} = o_j - y_j \tag{1.17}$$

This is the first backpropagated term for a single output neuron's preactivation. The use of the chain rule allows these calculations to continue backwards through the ANN towards the input layer, finding the partial derivatives of all trainable parameters prior to tuning. Practically, these are matrix operations (Jacobian matrices of first-order derivatives) as modern DNNs can have millions of trainable parameters. This also allows large-scale parallelisation of the algorithm. But here we have demonstrated how these partial derivatives can be backpropagated.

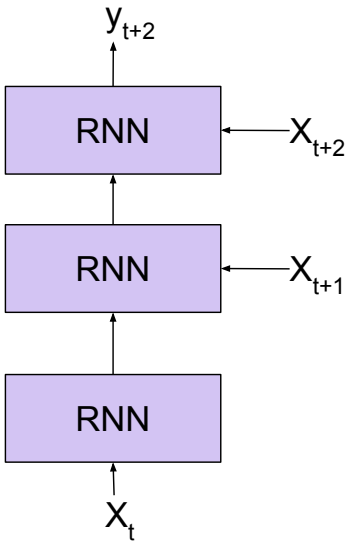


Figure 1.4: An unrolled RNN. Backpropagation Through Time (BPTT) can be implemented to treat the RNN model at each time step as if it is a different layer of the same feed-forward network in order to propagate the error through time. Although, as it is the same model at each timestep, the weights are the same.

To train RNNs, a variation of the backpropagation algorithm can be used whereby the the RNN is 'unrolled', with the models input, output and state at each timestep displayed side-by-side such that the feedback connection appears to be a simple feedforward connection between instance of the same model. This is known as Backpropagation Through Time (BPTT, Figure 1.4).

This is necessary as, for RNNs, the hidden layer affects the loss function both *directly* (through its connection to the output layer) but also *indirectly* (through its feedback connection affecting the hidden layer at the next timestep). BPTT allows the algorithm to propagate the errors from the last timestep to the prior timesteps. In this way, the algorithm must unroll and propagate the errors back through a number of timesteps equal to the length of the input sequence. This can be rather costly in terms of computational resources as, for each timestep, the RNN activations need to be stored in memory. Truncated Backpropagation Through Time is a common solution to this problem, splitting the input sequences up into a number of smaller input sequences, and running BPTT on each one individually. This has the disadvantage that dependencies larger than the size of the new sequence length will not be learned.

1.2.5 Audiovisual Deep Learning

With an overview of the general concepts of artificial neural networks provided in previous sections, this section extends the topic to focus on the area of audiovisual learning. Audiovisual learning tasks can be divided into the following three categories; audiovisual separation and localisation, audiovisual recognition, audiovisual generation and audiovisual correspondence/representation learning.

Audiovisual separation considers the problem of obtaining audio signals, each pertaining to a single source. This is a problem that humans must solve in the environment where there is an auditory signal to be perceived amongst auditory noise. Sounds from different sources mix in the air before arriving at the ear, this mixture must then be separated by the brain such that the listener is capable of attending to particular auditory signals. When paired with visual data, this further aids separation by providing additional source information via audiovisual correspondences (Gabbay et al., 2018; Ephrat et al., 2018; Afouras et al., 2018). The visual data also allows localisation of the auditory streams within the visual frame. This is problem is posed in a similar way to the cocktail party scenario (Cherry, 1953) in multisensory integration; whereby a listener must perceive speech in a noisy social setting.

Audiovisual recognition tasks provide audiovisual data pertaining to some event(s) that must be allocated one or more labels. This is also a task that is ecologically relevant to humans. With access to data across a number of sensory streams, humans must recognise events in their environment in order to respond effectively. Audiovisual speech recognition (Gogate et al., 2018) and audiovisual action recognition (Akbari et al., 2021; Nagrani et al., 2021) tasks are common examples in the field of deep learning and are ecologically relevant to humans. We use an action recognition task throughout this work for optimisation and often for testing.

Unlike the other discriminative models in this section, there are also generative audiovisual models whereby audio is generated from visual data or vice versa. Audiovisual generation tasks include those tasks that generate data in one modality in accordance with presented data in another. This is parallel to the ability of humans to imagine corresponding data in other modalities (for example imagining the sound of a laugh in response to an image of someone

laughing). These models can be trained in a variety of ways, Zhou et al. (Zhou, Wang, et al., 2018) for instance created a series of models to create raw waveforms from videos using an encoder-decoder type architecture. Generative models can also be pitted against an adversary in the case of a generative adversarial networks (GANs) (Goodfellow, Pouget-Abadie, et al., 2014). A GAN uses a generative model that generates data and a discriminative model that discriminates between data from the model distribution and data from the data distribution (real vs generated).

Audiovisual correspondence learning include all those tasks that involve matching audio to visual streams. These are self-supervised learning problems that do not require the expensive annotation regimes that many datasets employ. Instead, originally sourced audiovisual videos can be obtained and labelled as 'corresponding', then negative data-label pairs can be generated in two different ways. The first is the audiovisual correspondence (AVC) task (Arandjelovic and Zisserman, 2017) where the audio and visual streams are shuffled *between* videos and no longer correspond semantically. The second is the audiovisual synchronisation (AVS) task (Cheng et al., 2020), where the audio and visual streams still belong to the original video but a temporal misalignment has been introduced *within* the video (for instance a dog may open and close its mouth and then a second later there is a bark sound). Although these models have been shown to learn representations that generalise well to recognition tasks (Arandjelovic and Zisserman, 2017; Cheng et al., 2020), we broaden the definition of audiovisual representation learning tasks here to include other models, such as audiovisual autoencoders, that are not explicitly optimised on a correspondence task but could still be used to learn audiovisual representations. These representations can then be used on other tasks such as unisensory or multisensory recognition tasks. Audiovisual correspondence tasks are used in this Chapter 4 for out-of-domain testing of DNNs optimised on action recognition.

1.2.6 Early and late multimodal fusion

For multisensory recognition tasks, signals from different data modalities may carry both redundant and complementary information. It is in this realm of multisensory perception that accuracy and reliability gains can be found, particularly when unisensory data contains

ambiguities or noise. However, it is a difficult practical challenge to learn to combine data from multiple modalities in a way that useful for perception. An outline of some of the challenges faced when developing models of multimodal perception can be found in Atrey et al. (2010) which provides considerations on the wider problem of multimedia analysis. In particular, we focus here on the processing depth of the first multimodal operation or 'level of fusion'. This could be early, late or some hybrid that attempts to leverage the advantages of both Mervitz et al. (2020).

Early fusion in deep learning typically involves combining features from each modality and building a model that processes both, together, from beginning to end. This model may not be well suited for both modalities and some engineering may be required to properly align the data. However, these models have the opportunity to model signal-level interactions in the data of different modalities.

The alternative to early fusion of multimodal data in deep learning models is of course late fusion. Late fusion strategies typically involve using 2 unimodal submodels that perform a significant amount of the processing before any shared processing stages which occur towards the end of the sequence. Typically the unimodal submodel abstractions are aggregated using concatenation or addition/ average before these final multimodal processing steps are carried out. We may call this an *additive approach* in line with the terminology in Liu et al. (2018). As the representations are far abstracted from the raw data, correspondences between the data of each modality is likely to be far less clear and more difficult to model. In the work outlined in this thesis, we use late fusion methods of audiovisual fusion as it allowed us to leverage unisensory CNNs, pretrained on unisensory tasks, to achieve high levels of audiovisual recognition performance.

Additive approaches to multimodal perception make an assumption that all modalities are useful in every sample and should be weighted equally across samples. In reality, samples sometimes present weak modalities where a signal may be no use at all. For instance, when classifying the visual signal of a video showing a dog running across the screen, backing music would not aid classification at all but additive methods of multimodal fusion would erroneously

use this data to make a classification. Liu et al. (2018) states that another assumption of additive models of multimodal fusion is that neural networks built on top of the fused data should be able to learn to determine the quality of the data in a modality and recover the classification but that this is difficult in practice. The paper then goes on to introduce a multiplicative multimodal method that explicitly assumes that some modalities of a sample are less informative than others. The method suppresses any high penalties on a unimodal model when another unimodal models assign a high probability to the correct class and can be used during the training regime of a multimodal artificial neural network. Attention methods are also commonly used to solve this problem of assigning a weighting to activations in a deep learning model according to the sample.

1.3 Humans vs Artificial Neural Networks

Artificial Neural Networks have reached and, in a number of cases, surpassed human levels of performance on a number of naturalistic classification tasks (Krizhevsky et al., 2012). These performances, however, refer to specific benchmarks that are used for training and testing models. As discussed at the end of Section 1.3.4, these DNNs are often optimised to perform, not only a specific task, but inference on a specific dataset, which, in the case of image recognition, is only some small subset of the set of possible images. Despite those current limitations, for the first time, cognitive scientists have computational models capable of classifying naturalistic stimuli. Interest in using artificial neural networks as models of the human brain has grown alongside these advancements in the field of deep learning, in particular as models of sensory perception. An important criterion of models of sensory perception is that they should be able to solve the same tasks that humans solve. In the following section we will outline the transformational and architectural similarities between humans and ANNs, how ANNs may be used to explore biological forms of intelligence and where behaviour has been found to diverge between humans and ANNs.

1.3.1 Computations in Deep Neural Networks and the human brain

The representational transformations throughout an ANN are a potential source of similarity with human behaviour and activation patterns. Fundamentally, ANNs are connectionist models just like the animalian brain. They are made up of a number of units named artificial neurons which coarsely model biological neurons, receiving inputs (either model input data or the activations of preceding neurons), performing a weighted sum (a parallel to integration of postsynaptic action potentials in biological dendrites), and often passing them through a non-linear activation function (a parallel to the firing action potential of the biological soma). Modern deep neural networks also often leverage other operations that are computationally plausible in the brain, including; convolution, threshold non-linearities, pooling and normalisation (Carandini and Heeger, 2012).

Although the artificial neuron is biologically inspired and its organisation in modern day artificial neural networks can result in human-like performance and error patterns and generate predictions of animalian brain responses, there is a neuron-level performance gap. Some biological neurons in the human brain have been shown to solve the XOR problem for example (Gidon et al., 2020), a task that even a single layer of perceptrons were famously unable to solve, causing the research area of artificial neural networks to dwindle and fall into the 'AI winter'.

The reasons for the performance gaps between artificial and biological neurons is not well understood, but we may glean some insight from briefly reviewing some known differences. For instance, artificial neurons in feedforward models, such as those used in state-of-the-art CNN models, generally do not maintain a resting potential. Further while they provide a parallel to spatial summation of postsynaptic potentials (they perform a weighted sum, and in CNNs the potentials are spatially local) they do not perform a temporal summation, whereby postsynaptic potentials occurring in the same place but slightly different times are integrated. In other words, once a CNN has been trained and is in 'inference mode', its activations and output in response to a stimulus will in no way be adjusted according to the previous stimulus. As such, these neurons are not activating to sequences of inputs as are their biological counterparts, and have no perception of time or sequence. Of course, an organised ANN sequentially processes inputs

from input layer to output layer, but the artificial neuron itself provides one response to one set of inputs.

The list of differences between ANNs and the parts of the animalian brain they represent (e.g. object recognition models and the primate visual ventral stream) does not only include neuron-level differences. Backpropagation, used to train ANNs, is considered a biologically implausible learning algorithm for the brain as neurons do not form synapses according to the synapses between other neurons. In accordance with Hebbian learning, a biological learning algorithm must ensure that neurons are only able to access information from neighbouring neurons.

ANNs are most commonly trained via supervised learning techniques whereby labelled data is used to backpropagate through the network, finding the partial derivative of each parameter with respect to the loss, and tuning it in the direction of the negative gradient. This supervised learning strategy is often reliant on enormous labelled datasets, often at least tens of thousands samples or even millions of samples are used. ImageNet, for example, is the most commonly used labelled image dataset today and contains over 14 million labelled images. To what extent the encodings in the human brain are learnt by experience or hard-coded in our genome is an ongoing question in neuroscience, but one thing is certain, we do not learn from millions of labelled examples.

1.3.2 Hierarchical Processing

As DNNs are hierarchical systems, with the output of one layer forming the input to the subsequent layer, researchers have been able to assess the hierarchical nature of the visual ventral stream by considering which DNN layers best predict responses at different regions of the visual cortex using fMRI and MEG. In other words, researchers are able to explore the question; 'are the regional differences in activity due to sequential processing?'.

A model that is highly predictive of any particular brain region must be able to achieve a similar level of performance on relevant tasks. The human visual ventral stream, terminating at inferotemporal (IT) cortex, is involved with visual recognition tasks such as object recognition.

As the neural coding of inferotemporal (IT) cortex gives rise to high performance on object recognition tasks, any highly predictive model of IT cortex must also perform to a similar standard (Yamins, Hong, Cadieu, Solomon, et al., 2014).

Studies have shown that in a number of deep CNNs that reach human performance levels on object recognition tasks, the final layers are highly predictive of IT cortex in both macaques (Yamins, Hong, Cadieu, and Dicarlo, 2013; Cadieu et al., 2014) and humans (Yamins, Hong, Cadieu, and Dicarlo, 2013; Yamins, Hong, Cadieu, Solomon, et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014; Güçlü and Van Gerven, 2015; Cichy, Khosla, et al., 2016). Further, in these same models, there have been further hierarchical correspondences found. For example Yamins, Hong, Cadieu, Solomon, et al. (2014) found that the penultimate layer (and sole input to the final layer) of their deep CNN was highly predictive of V4, the dominant cortical input to IT cortex. Others have found early deep CNN layers to correspond to early parts of the human visual ventral stream (V1 and V2) (Güçlü and Van Gerven, 2015). In their fMRI and ANN experiment, (Güçlü and Van Gerven, 2015) found that the hierarchical correspondences between the ventral stream and object recognition optimised CNNs also corresponded to an increase in feature complexity, invariance and size as signals move downstream.

Although these deep CNN models were not explicitly developed to predict activations in IT cortex, by training the model on a task for which this brain area exhibits high decoding performance, object recognition, the model was highly predictive of IT cortex. Further, due to the hierarchical nature of CNNs, the inductive bias to learn a hierarchical solution led to correspondences with even upstream areas of the afferent visual ventral stream. Whilst the choice of task and hierarchical architecture of the model are important features of the visual ventral stream in primates, the exact order of neurologically inspired operations such as convolution, pooling and normalisation seem to be less important. This is reflected by the variety of architectures examined in this literature and the finding in Yamins, Hong, Cadieu, Solomon, et al. (2014) that object-recognition performance correlated strongly with a model's ability to predict IT cortex activations.

Artificial Neural Networks have not only been used to model the visual ventral stream, for

instance they have also been used as models of the auditory cortex (Kell, Yamins, et al., 2018). Here, the researchers applied the same principles used in the visual ventral stream literature to demonstrate the hierarchical nature of the auditory cortex using ANNs. This underlines an important potential use for artificial neural networks as tools for investigating the hierarchical nature of the sensory cortices.

1.3.3 Deep Neural Networks as Approximations of Ideal Observers in psychology

Even where these models deviate from biological neural networks, there is still utility in a model, optimised on an ecologically relevant task, that performs to a similar level to its biological counterparts. Indeed, in Section 1.3.2 we explained how biologically plausible operations, organised in a hierarchy and optimised on a task that is important to some region of the sensory cortices can lead to a model with human-level performance, capable of predicting neural responses. But here, we focus on one element of this recipe; the ecologically relevant task.

Despite the many differences between artificial and biological neural networks, deep neural networks are valuable tools of investigation in sensory neuroscience. As deep learning models can be optimised on real-world, ecologically important tasks, it is possible to use them to explore the role of task-constraints on neural systems and behaviour in the animalian brain, particularly in tasks where provably optimal ideal observers cannot be obtained (Kell and McDermott, 2019).

An ideal observer is a mathematical model that performs a specific task in an optimal way given the stimuli. Ideal observer models are used to investigate how information is processed in a perceptual system by providing an upper bound on performance. In particular, ideal observer models have been used to investigate human perceptual processing, the assumption being that biological organisms often solve ecologically important tasks optimally. These ideal observer models can provide explanations for perceptual behaviour, such as the response to illusions presented to animals during perceptual tasks by showing that they are optimal in particular circumstances.

1.3.4 Deep Neural Networks as Models of Human Perceptual Judgement

DNNs have now achieved state-of-the-art performance on a number of real-world classification tasks, rivalling human performance (Krizhevsky et al., 2012). To effectively model human perception, DNNs must first reach human levels of task performance. As described in previous sections, we know that one potential source of performance similarity between DNNs and humans is that both systems have reached the natural limits of performance on the task. If this is the case, then one may investigate further and consider error patterns. If error patterns are dissimilar, then that would reflect some computational dissimilarity, if not, the source of the performance similarity could be some other algorithmic similarity. For a classification task we may ask, 'do DNNs and humans confuse the same categories as one another?'.

Much of the literature exploring the sensory cortices in humans and other primates using ANNs has focussed on the ventral stream of the visual system, extending on work that sought to compare other computer vision models to the ventral stream. For instance Borji and Itti (2014) benchmarked 14 computer vision models on scene and object recognition datasets for which human data was already available in order to provide an overview of the progress towards achieving human-level vision in 2014. But following the success of Krizhevsky et al. (2012), the computer vision literature progressively came to focus on artificial neural networks. Here, deep CNNs reached, and even exceeded human levels of performance. In 2012, deep CNNs surpassed human performance on hand-written digit recognition tasks (Cireşan et al., 2012; Wan et al., 2013), in 2014 deep CNNs surpassed human performance on face recognition tasks (Taigman et al., 2014; Sun, Chen, et al., 2014), and by 2015, the performance of deep CNNs surpassed that of humans on the ImageNet challenge (Russakovsky et al., 2015; He et al., 2015).

Although DNNs have demonstrated human-level performances on a number of visual tasks, this is not always the case with tasks outside of the original domain for which the DNN was trained. Stabinger et al. (2016) for instance sought to investigate LeNet and GoogLeNet against human participants on a series of 23 visual reasoning tasks used in Fleuret et al. (2011). In each task an image was presented containing 2 or more generated closed-contours, the human or computer vision model must then select 1 of 2 possible categories to describe the image. In

order to solve each task, the observer must be able to detect some organisational principle (e.g. proximity, similarity, symmetry). The researchers found that on most tasks that required the comparison of shapes, both CNNs performed poorly. This was with the exception of 4 tasks for which the CNNs were found to be using unintended patterns in the data. Highlighting an issue that humans may incorrectly conclude that these models have learned human-like concepts. Another study by Heinke et al. (2021) shows that a series of popular CNNs were unable to solve a geometrically possible vs. impossible shape task like human participants. In Funke et al. (2021), researchers found that CNNs were able to solve the closed vs. open contour problem at performances similar to humans. However, in this study, it was found that the CNNs were 'cheating' and identifying edges in order to detect open contours, rather than using a concept of 'closedness'. These studies are important examples of DNNs failing to solve problems that humans solve (or solving them in a different way), particularly when those tasks are outside of the training domain.

Expanding the problem of image recognition to image *interpretation*, whereby a visual system must recognise and localise primitive object features, allows another avenue of investigation when considering computational models of human vision. Humans are able to solve this problem effortlessly, identifying object components that carry additional information about identity and configuration, but not much is currently understood about the problem (Ben-Yosef et al., 2018). Using a minimal recognisable images task (Ullman et al., 2016), Ben-Yosef et al. (2018) was able to investigate the image properties used by humans and computational models to recognise and localise primitive features. Minimal recognisable images are image crops that are reliably recognised by humans but are unrecognisable if they are cropped further i.e. the patch containing approximately the minimal features for recognition (Ullman et al., 2016). DNNs can typically recognise and localise objects but draw coarse bounding boxes and do not identify the object's semantic components (Ben-Yosef et al., 2018). The interpretation model built by Ben-Yosef et al. (2018) consisted of a DNN to obtain candidate primitives followed by a relations calculation and then a decision tree classifier to select the most compatible configuration. The model was found to use properties that were important to human performance and similarly experienced a

sharp performance drop when they were removed. This better reflected human performance than previous bottom-up models (Ben-Yosef et al., 2018). It is worth noting, however, that DNNs are largely trained on image-scale tasks (with coarser labels/ bounding boxes). It is therefore possible that primitive labelling of ImageNet, for example, may bring about a similar dependence on features humans use, and a similar set of minimally recognisable image patches for DNNs that are trained on it.

Another work (George et al., 2017) sought to incorporate other inductive biases into their computer vision models than are currently used in CNNs. Specifically, the ability of the model to segregate contour and surface representations and to code for border-ownership. These are considered to have an important role in the visual cortex (DeYoe and Van Essen, 1988; Lamme et al., 1999; Craft et al., 2007). George et al. (2017) introduces a hierarchical model named a Recursive Cortical Network (RCN) that uses a compositional hierarchy of features to model contours and a Conditional Random Field to model surfaces. The work finds that the RCN is more data efficient and robust than CNNs on a range of tasks. More specifically, RCN was able to break a series of CAPTCHA tasks; reCAPTCHA 66.6%, BotDetect 64.4%, Yahoo 57.4% and PayPal 57.1%. For comparison, humans score 87.4% on reCAPTCHA (George et al., 2017) and scores above 1% are considered to render the CAPTCHA ineffective (Bursztein et al., 2011). This was achieved with few training samples per character on each task. To train a CNN (Goodfellow, Bulatov, et al., 2014) on the CAPTCHA task, however, required the researchers over 2.3 million unique training images to obtain an accuracy rate of 89.9% (George et al., 2017). This trained CNN then failed on string lengths not present in the training data, and the performance decreased drastically in response to imperceptible perturbations to the input. The RCN was also found to be more robust to clutter in one-shot and few-shot MNIST than CNN models LeNet-5 (LeCun, Bottou, et al., 1998) and VGG-fc6 (Simonyan and Zisserman, 2015) pretrained on ImageNet (Deng et al., 2009). This study highlights the role of adding biologically-informed inductive biases to computer vision models could play in the pursuit of human-level generalisation ability.

In considering deep neural network models of human perception, the literature thus far has focussed on feedforward models (Dodge and Karam, 2016; Dodge and Karam, 2017; Wichmann

et al., 2017; Geirhos, Temme, et al., 2018; Dodge and Karam, 2019; Stabinger et al., 2016; Heinke et al., 2021; Yamins, Hong, Cadieu, and Dicarlo, 2013; Yamins, Hong, Cadieu, Solomon, et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014; Güçlü and Van Gerven, 2015; Cichy, Khosla, et al., 2016; Kell, Yamins, et al., 2018). Particularly in the area of visual perception, this is perhaps unsurprising. Research demonstrating that humans are capable of object recognition with only brief stimuli presentations (Potter, 1976; Thorpe et al., 1996) has led researchers to believe that 'core object recognition' is largely solved in the human visual cortex via feedforward connections. Indeed, systems with primarily feedforward operations are sufficient to solve challenging object recognition problems (DiCarlo et al., 2012). This is further evidenced by the success of feedforward DNNs, reaching human level performances on object recognition (Krizhevsky et al., 2012). However, observations that neural and behavioural responses are delayed when occlusions are introduced to stimuli on object recognition tasks or that the recognition of occluded objects is disrupted by masking (Johnson and Olshausen, 2005; Wyatte, Curran, et al., 2012) indicates that solving this problem requires recurrent processing. Indeed, research is starting to examine the advantages of recurrent processing in visual recognition tasks (O'Reilly et al., 2013; Spoerer, McClure, et al., 2017; Bergen and Kriegeskorte, 2020). Spoerer, McClure, et al. (2017) in particular shows that recurrence improves recognition performance of a set of DNNs on a number of challenging tasks, such as occluded digit recognition. It has also been shown that recurrent models can outperform purely feedforward models on image recognition tasks with fewer parameters (Liang and Hu, 2015; Spoerer, McClure, et al., 2017). Indeed, the work presented in this thesis considers recurrent convolutional neural network models of perception (Chapters 4 and 5).

An opinion piece by Yuille and Liu (2021) provides an overview of the current progress of DNNs as models of human visual perception. Whilst the authors outline the great achievements of DNNs in the computer vision domain, they also provide some criticisms, largely around supervised learning with large, labelled datasets. In particular, the authors reason that DNNs, optimised on a dataset, may generalise poorly outside of that dataset as the set of possible images is infinitely large. (We note here, that there is a fixed number of possible images, given a fixed

resolution with pixels ranging from 0 to 255, but this is intractable with 256^3 possible RGB values). Resultant biases have been studied in popular benchmarks such as ImageNet (Zhu et al., 2017). The authors go as far as to suggest that limited and biased datasets could be the cause of adversarial vulnerability, causing the DNNs to draw 'short-cut' decision boundaries. As current large, labelled datasets do not wholly represent the world of possibilities, they are also poor measures of real-world performance 'an algorithm is only as good as the dataset it is evaluated on and the performance measures used' (Yuille and Liu, 2021). Further criticisms include; the problem specificity of many DNNs, that supervised DNNs need many more examples in order to learn than human participants and that DNNs are less robust to occlusions and perturbations than humans.

1.3.5 Deep Neural Networks as Models of Human Visual Perception as the Signal Gets Weaker

In life, there are a number of circumstances where visual signals are unreliable; perhaps the sun is setting and there is little light, or perhaps there is too much light on a winter's day when the sun is low in the sky, perhaps there is a thick fog or heavy rain, it could even be the case that one is swimming underwater. In these circumstances, it is important that we are still able to operate in our environment.

There are additional opportunities for visual noise to occur in digital media, perhaps when images are first photographed or when they undergo lossy compression techniques. A smudged or out of focus camera lens can result in a blurred photograph for instance. More recently, some types of filter will add a creative blur to particular parts of an image to make the image more aesthetic (such as Instagram). Impulse noise (such as salt and pepper noise where pixels are randomly assigned to be black or white) is also common, and may present itself during acquisition, transmission or storage. Humans, to some extent, are able to still perceive noisy images, any model of human visual perception must be able to do the same.

A number of studies have examined the classification behaviour of deep CNNs in response to images with a variety of degradations including; Gaussian noise (Dodge and Karam, 2016;

Dodge and Karam, 2017; Dodge and Karam, 2019), Gaussian blur (Dodge and Karam, 2016; Dodge and Karam, 2017; Dodge and Karam, 2019), contrast reduction (Dodge and Karam, 2016; Wichmann et al., 2017; Geirhos, Temme, et al., 2018), greyscaling (Geirhos, Temme, et al., 2018), salt and pepper noise (Geirhos, Temme, et al., 2018), JPEG compression (Dodge and Karam, 2016), JPEG2000 compression (Dodge and Karam, 2016) and a number of others. In all of these cases, researchers trained CNNs on ImageNet and used some held-out images to be used as test data for both models and humans.

When trained on ImageNet; AlexNet, GoogLeNet, VGG-16 and ResNet-50 have been shown to have similar performance and confusion matrices to human observers on greyscale (Geirhos, Temme, et al., 2018), colour (Dodge and Karam, 2017; Wichmann et al., 2017; Geirhos, Temme, et al., 2018) and to a lesser extent reduced-contrast images (Wichmann et al., 2017; Dodge and Karam, 2017; Geirhos, Temme, et al., 2018). However, humans have been shown to generalise to weak signals much better then these CNNs under a variety of distortions (Wichmann et al., 2017; Dodge and Karam, 2017; Geirhos, Temme, et al., 2018). In particular, CNNs seem to perform poorly on high frequency noise such as; low-pass filters, uniform noise and Gaussian noise, and low frequency noise such as; high-pass filters and blur (Dodge and Karam, 2016; Dodge and Karam, 2017; Geirhos, Temme, et al., 2018)

It is difficult to ascertain which of these CNNs are more robust to visual distortions. Dodge and Karam (2016) and Wichmann et al. (2017) report VGG-16 as clearly outperforming AlexNet and GoogLeNet under all of these noise conditions except high levels of JPEG compression. On a separate dog-breed classification task, (Dodge and Karam, 2019) reports GoogLeNet outperforming VGG-16 on high levels ($\sigma > 100$) of Gaussian noise and low levels of Gaussian blur ($\sigma = 2$). When categories were semantically much further apart and easy to recognise in a broader animal classification task (Dodge and Karam, 2019), both humans and models obtained higher accuracy except ResNet50 which performed considerably worse in the Gaussian noise case. Geirhos, Temme, et al. (2018) carried out an extensive investigation on a wider selection of distortions and some more recent, deeper CNN models including; GoogLeNet, VGG-19 and ResNet-152. Here, ResNet-152 largely outperformed all other models on every distortion

(greyscale, false colour, uniform noise, low-pass filter, contrast reduction, 3 types of Eidolon noise, phase noise, power equilisation and rotation) other than the high-pass filter distortion on which it performed worse than the other models. Although ResNet-152 additionally performed better than the other models and human participants on clean images, so its performance would have to deteriorate to a larger extent to fall below that of the others. This demonstrates how performance can vary across tasks and models.

As mentioned in the previous section (Section 1.3.4), research in this area has focussed on feedforward models of human visual perception. The previously mentioned advantages of adding recurrence, however, extends to this area of recognition in unreliable conditions. Indeed, the work by Spoerer, McClure, et al. (2017) that tested deep neural networks on an occluded digit recognition task, further investigated the ability of DNNs to classify when stimuli contained additive Gaussian noise, finding that those models with top-down and/or lateral (feedback) connections were more robust than the feedforward models. Although there is strong support in the literature to say that core object recognition is largely solved by feedforward processes (see Section 1.3.4), there is mounting evidence that feedback connections from extrastriate regions provide additional functionality, beyond attention, including grouping, associational reinforcement and filling-in of features (see Wyatte, Jilk, et al., 2014 for a review). In particular, these recurrent processes are considered important for recognition beyond core object recognition, for instance when visual stimuli are degraded.

1.4 Thesis Overview

In the following chapters I aim to extend the research exploring DNNs as models of human perception to the audiovisual domain. Chapter 2 outlines the methodological foundations of the work. Chapter 3 presents a large-scale data sorting study with trained participants to produce a large, labelled video dataset of *audiovisual* action events suitable for examining DNNs and human participants. Chapter 4 aims to produce DNN models of audiovisual perception, and to investigate whether the ability to solve the audiovisual correspondence problem (AVC) arises

implicitly due to optimisation on an audiovisual action recognition task. We develop a series of novel DNN architectures, run hyperparameter searches on them and optimise them on audiovisual action recognition before using support vector machines (SVMs) to test them on the AVC task. We further carry out a series of 'selective-attention' tasks on the models to explore the interaction between auditory and visual signals in the learnt representations. In Chapter 5 we add a series of visual distortions to the test stimuli of our audiovisual classifiers in order to explore the robustness of the models, and to better understand their ability to use audio data to reduce any negative effects on performance. We further carry out online experiments with human participants in order to better understand similarities and differences in performances and error patterns. Chapter 6 provides a general discussion of the work carried out in the thesis, the research findings, the limitations and where research should go from here.

CHAPTER 2

METHODS

The following chapter provides an overview of the experiment methodologies employed in this thesis. First, I outline the action recognition task used to optimise and test models (see Chapter 3). Next, I outline our use of state-of-the-art neural networks via transfer learning and how they are employed in our own models (see Chapter 4). Finally, I present the use of held-out test sets and hypothesis tests to compare classifiers to one another and to human participants in Chapters 4 and 5.

2.1 Action recognition

Although humans can live comfortably without fully functional sensory systems in today's nurturing societies, evolutionary pressures would have required humans to be able to reliably perceive their environments. Of particular importance would have been the ability to recognise actions, our ancestors certainly would have had to recognise an approaching threat in order to increase their chance of survival. In this way, recognising action events provides a problem that is ecologically relevant to humans, and by optimising deep neural networks on this task, we are able to learn about how the constraints of this task may have formed neural systems and behaviour. Action recognition also provides a perfect test for audiovisual perception, as the event captured by either modality often requires a temporal sequence of data in order to recognise it. This is contrary to an audiovisual object recognition task, for instance, in which only a single frame would be required alongside the audio sequence. Further, there are several large, labelled

action recognition datasets publicly available that can be utilised, In this work, we leverage the Moments in Time (MIT) dataset (Monfort et al., 2019) and the Visual Engaged and Grounded AudioSet (VEGAS) dataset (Zhou, Wang, et al., 2018) and further prepare a clean training and test set with controlled levels of audiovisual correspondence. The preparation of the Audiovisual Moments in Time (AVMIT) dataset and the extended version, AVMIT-VEGAS, is the subject of Chapter 3. These datasets are used throughout this thesis, with the AVMIT dataset being used in Chapter 4 and AVMIT-VEGAS being used in Chapter 5.

2.2 Transfer learning

A common problem when training DNNs is the required amount of data. Modern deep neural networks use very large, labelled datasets and often this data is not available outside of popular benchmarks such as ImageNet (Deng et al., 2009). Particularly for the benchmark we produce in Chapter 3, the size of the training set can be a limiting factor for training DNNs. The models we develop in Chapter 4, however, utilise advances in the unisensory DNN literature by using CNNs previously trained on a unisensory problem. This is known as *transfer learning*, whereby models trained for one task are redeployed for another, either as feature extractors or to be fine-tuned and used as an effective starting point. In this work, in order to run hyperparameter searches and train many classifiers, and to test across many conditions and tasks, we opt to use the former method. Indeed, to solve the audiovisual problems in this work we implement 2 CNNs (1 auditory and 1 visual) in each architecture we develop. Training 2 deep CNN architectures, together, as part of a larger hybrid model, is outside of our compute capability. Running the CNNs in 'inference mode' and training the audiovisual components was quite possible however, resulting in less than 1 million trainable parameters in all cases. The models developed throughout this work use a novel extension of the recurrent convolutional neural networks (RCNNs) described in Section 1.2.3, leveraging four pretrained CNNs. We create a VGG-based (Simonyan and Zisserman, 2015) audiovisual feature extractor and a MobileNet-based (Howard et al., 2017) audiovisual feature extractor. Both have few parameters (by deep CNN standards) with one representing

earlier, simpler operations and the other representing more modern, efficient operations. We outline the architectures of the CNNs used for transfer learning below. The development and behavioural examination of the models make up the study in Chapter 4.

2.2.1 VGG-16

Simonyan and Zisserman (2015) introduced VGG-16 which is a simple CNN implementation consisting largely of 2D convolutional layers and max pooling layers. The convolutional layers have a kernel size of 3x3, a stride of 1 and varying numbers of filters that increase as the model becomes deeper. Where the original model has 16 layers, we use the Global Average Pooling TensorFlow operation applied to the final convolutional block of the model. Where modern CNNs often have hundreds of layers. Hence we use this architecture as a well performing yet primitive CNN to contrast with the modern, deep CNNs used in our other models.

2.2.2 VGGish

VGGish (Hershey et al., 2017) is an Audio Set (Gemmeke et al., 2017) trained 11-layer VGG architecture from Simonyan and Zisserman (2015). Unlike the visual implementation described above, this model contains the final 3 fully connected layers as in the original publication (Simonyan and Zisserman, 2015), although the final fully-connected layer is reduced from 1,000 units to 128 units to provide a more compact audio embedding.

2.2.3 YamNet

YamNet (Plakal and Ellis, 2020) is an implementation of the MobileNetV1 architecture (Howard et al., 2017) pretrained on Audio Set (Gemmeke et al., 2017). Howard et al. (2017) demonstrated that MobileNetV1 was able to outperform a number of previous state-of-the-art CNN models such as GoogLeNet (**Szegedy2015a**) and VGG-16 (Simonyan and Zisserman, 2015) on the ImageNet challenge despite using considerably less multiplication and add operations and less parameters. MobileNetV1 was able to achieve this through the use of depthwise separable convolutional

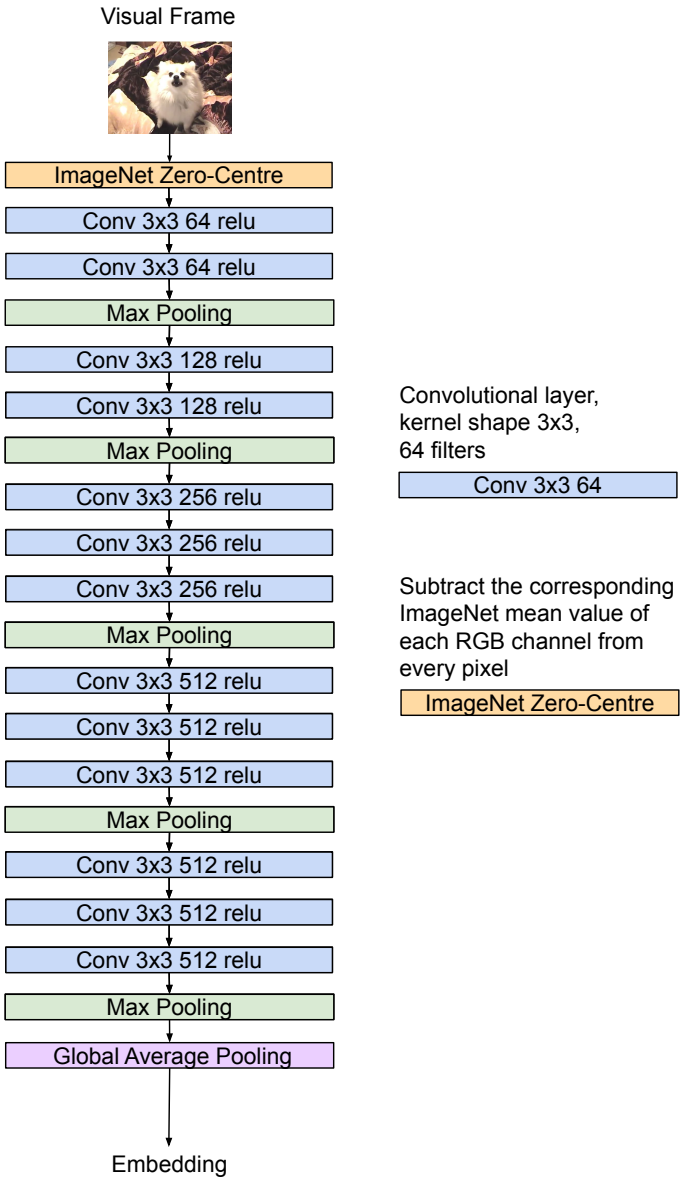


Figure 2.1: VGG-16 (Simonyan and Zisserman, 2015). A visual frame is fed to a series of 2D convolutional and pooling layers. The final classification layer is removed in order to provide feature embeddings.

layers, whereby a normal 2D convolution operation is decomposed into a depthwise (channel-wise) and pointwise convolutions. These convolutions where interspersed with BathchNorm operations and ReLu activation functions. The MobileNetV1 models also used a width and resolution multiplier that could be used to scale the networks to a smaller size, if desired, whilst optimising performance, although YamNet uses the full-sized model.

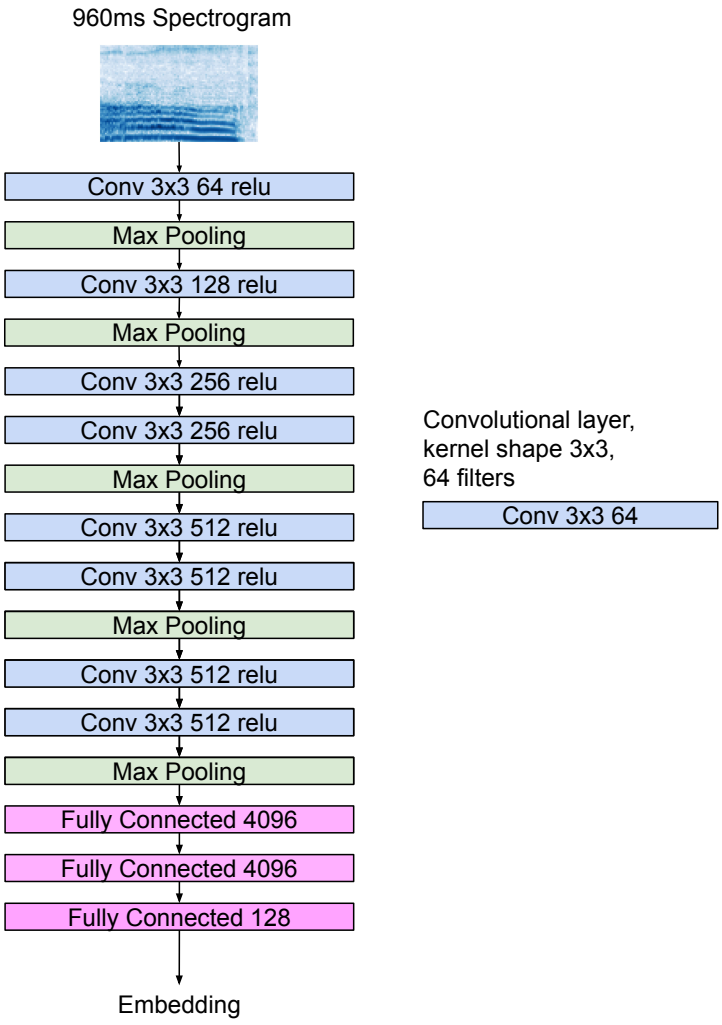


Figure 2.2: VGGish (Hershey et al., 2017) (VGG-11 (Simonyan and Zisserman, 2015) architecture). A log mel spectrogram corresponding to 960ms is fed to a series of 2D convolutional and pooling layers. The final classification layer is removed in order to provide feature embeddings.

2.2.4 EfficientNet

Tan and Le (2019) developed a set of deep CNNs, each one corresponding to a previous state-of-the-art CNN, designed to preserve (or even exceed) performance levels but reduce the number of parameters and required floating point operations (FLOPs) to run the model. The number of FLOPs are used to measure the required computational resources for a deep learning model to perform inference. The work emphasised the importance of scaling across all dimensions in a deep convolutional neural network rather than some previous efforts to scale across single dimensions (particularly depth) and the researchers introduce a novel method of scaling across

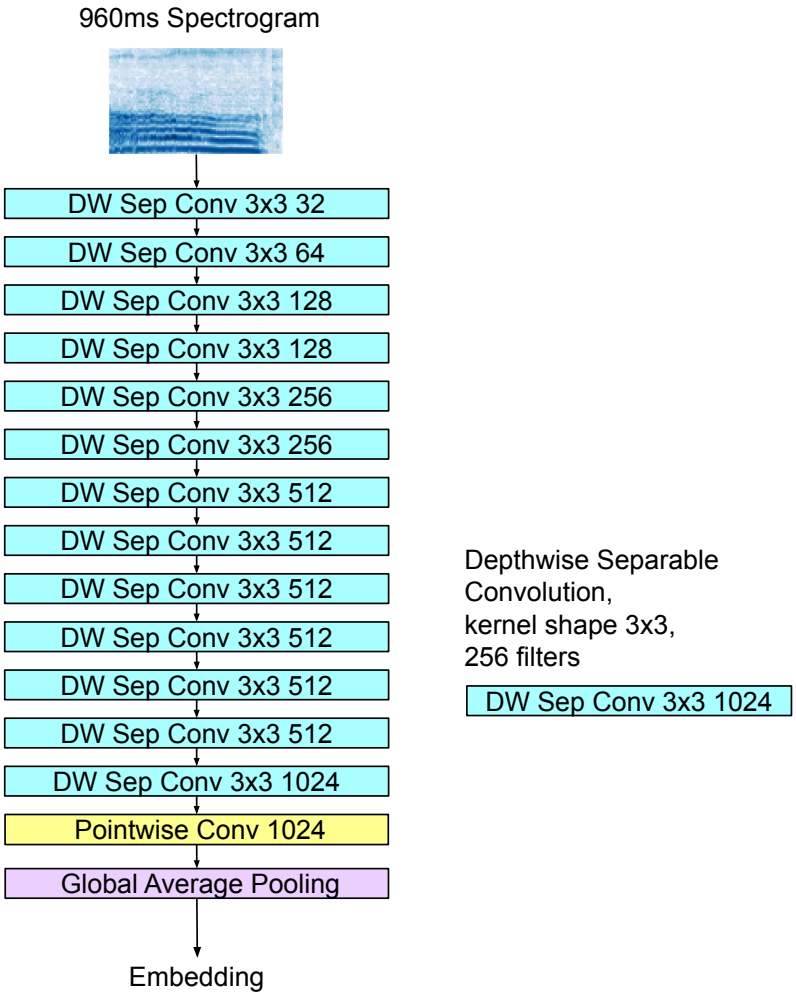


Figure 2.3: YamNet (Plakal and Ellis, 2020) (MobileNetV1 (Howard et al., 2017) architecture). A log mel spectrogram is fed to the to a series of depthwise separable convolutional layers. The final classification layer is removed in order to provide feature embeddings.

all dimensions using a compound coefficient which is used to produce the new family of CNNs named EfficientNets.

As the CNNs selected were to be used as components of larger models, we selected the smallest available model of the EfficientNet (Figure 2.4) family, EfficientNet-B0, as a visual feature extractor. EfficientNet-B0 makes use of inverted bottleneck residual blocks that were used in MobileNetV2 (Sandler et al., 2018). These are based on the depthwise-separable convolutions in the original MobileNetV1 architecture (Howard et al., 2017),

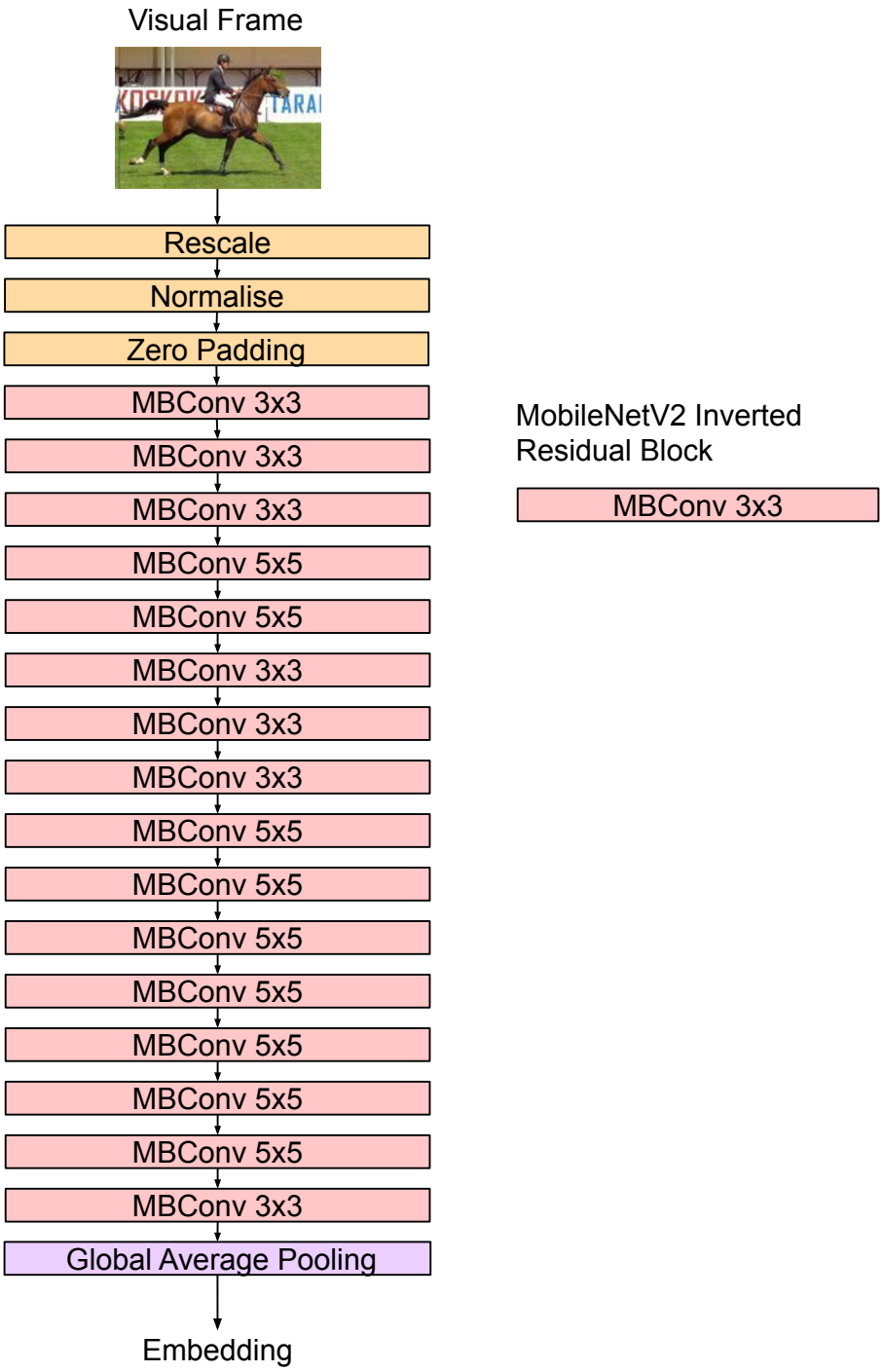


Figure 2.4: EfficientNet-B0, the smallest of the EfficientNet model series (Tan and Le, 2019). A visual frame is fed to the input preprocessing layers of the model and then subsequently processed by a series of MobileNetV2 inverted residual blocks. The final classification layer is removed in order to provide feature embeddings.

2.3 Model Testing

2.3.1 Using a Held-Out Test Set

In the field of deep learning, researchers often explore a number of statistical questions, and often they pertain to assessing the performance of a supervised learning algorithm in a single domain (e.g. performance on ImageNet; Krizhevsky et al. (2012)). This requires that the researcher uses the dataset to both train and test a supervised learning algorithm using resampling methods such as leave-one-out cross-validation, K-fold cross-validation and the bootstrap method, to estimate generalisation performance. It is at this fundamental level in which our research deviates. In this thesis, we study DRCNN and human behaviour on a limited number of audiovisual test videos with a high level of audiovisual correspondence and controllability that we obtain ourselves through a large-scale human sorting task (AVMIT/AVMIT-VEGAS test sets; Chapter 3). The training sets, however, are of a different level of cleanliness and audiovisual correspondence (as shown in Chapter 3), and so the tasks are measuring out of domain performance before any modifications are even made to the videos. Further, we did not endeavour to estimate the performance of particular learning algorithms, and instead focus on the learnt behaviour of trained classifier instances (refer to Dietterich (1998) for further explanation).

2.4 Hypothesis testing

Hypothesis tests are seldom carried out in deep learning literature, but where they are, they are often used alongside K-fold cross-validation to assess the performance of particular learning algorithms on a particular dataset (an estimation of the ability to generalise to within domain data). As previously explained in Section 2.3.1 our investigations are centred on testing outside of the training distribution, in this way we use all of the available test data and all of the available training data in one investigation, without sampling from the same dataset. Dietterich (1998) provides a clear description about two possible hypothesis test objectives; selecting the best learning algorithm and selecting the best classifier, where the supervised learning algorithm,

provided with examples and corresponding classes, produces a classifier and the classifier, given some input example, provides an output classification.

Hypothesis tests can be utilised to study effects in the results of group design experiments and individual effects in single-subject experiments. With deep learning experiments, often a selection of DNNs are selected for study. However, these models have not been uniformly sampled from the set of possible models under investigation (for example the set of all possible CNNs) and so one must exercise caution when generalising conclusions beyond the given scenario (model architecture, training set etc.). Indeed, the work by Funke et al. (2020) and Funke et al. (2021) shows the danger of drawing conclusions about families of DNNs that generalise beyond reasonable bounds. In our studies, we are conservative with the conclusions we make, and focus on individual variability of trained model instances such as in the work by Geirhos, Janssen, et al. (2017) and Geirhos, Temme, et al. (2018). The hypothesis tests we carry out are used to study individual, trained classifier behaviour, to compare these classifiers to humans or to each other, but never to treat the group of classifiers as a single sample.

2.4.1 One-sample permutation tests

Throughout Chapters 4 and 5 we frequently use permutation tests to detect significant effects. Permutation tests provide us with a null distribution for a test statistic by permuting data and can be used alongside a chosen test statistic to detect significant performances (Ojala and Garriga, 2010). The use of permutation tests provides a method to assess the significance of classifier performance (Ojala and Garriga, 2010; Hsing et al., 2003; Golland et al., 2005). In particular, Ojala and Garriga (2010) uses permutation tests to answer the questions 'how can we trust that the classifier has learned a significant predictive pattern in the data and that the chosen classifier is appropriate for the specific classification task?'. The former question, here, addresses the performance of the single classifier instance on the test data.

We carry out one-sample permutation tests by permuting test set labels and studying hypotheses centred on the test results of single, trained classifier instances, rather than generalised across particular model architectures in line with method 'Test 1' in Ojala and Garriga (2010). We

are then able to consider individual significant results across a range of recurrent convolutional neural networks to *indicate* trends without making bold assertions. The one-sample permutation tests used in this work have the following null hypothesis 'the predictions and ground truth labels are independent'. The null distribution can therefore be simulated by permuting the labels of the test set as in Ojala and Garriga (2010), the pseudocode is presented in Algorithm 1.

Algorithm 1 One-sample permutation test with classification accuracy test statistic (two-tailed)

```
actual_stat = Count(actual_labels == predictions)/num_predictions   ▷ class. acc.
simulated_stats = []                                                ▷ create an empty list
for num_iter do                                                     ▷ number of iterations
    simulation = Shuffle(actual_labels)
    sim_stat = Count(simulation == predictions)/num_predictions     ▷ simulated class. acc.
    simulated_stats.append(sim_stat)                                ▷ Add simulated test stat to list
p = Count(simulated_stats ≥ actual_stat)/num_iter                   ▷ p-value
```

2.4.2 Paired one-sample permutation tests

To detect effects between two related groups, one can use a paired one-sample permutation test. Paired tests are used when there is an element-wise dependency between samples. In psychology, this could be that the samples are before and after scores for participants on a test, with two scores (before and after) for each participant forming a pair. In our work, this can be used to compare classifiers and humans on the same test stimuli or to compare a single classifier's performance on a set of videos under particular conditions.

As with the one-sample permutation test, the null distribution of the test statistic is simulated by permuting the data and measuring the test statistic over a very large number of iterations. The p-value is then the number of instances that the actual measured test statistic is larger than or equal to the simulated test statistics. For test statistics that can be positive or negative, measuring the p-value as described will give the one-tailed result, using *absolute* values can be used to detect the effect in either direction for the two-tailed result. In our work we use the two-tailed result, which is naturally given by the McNemar statistic (Everitt (1977); Section 2.4.3), which has a lower bound of 0.

Under the null hypothesis, the values within each pair are interchangeable. Thus to simulate

the distribution of the test statistic under the null hypothesis, the data is shuffled *within-pairs* not *between-pairs*. For example, where a classifier has been tested on two test sets, the null hypothesis could be that there is no significant difference in classification performance. Where one test set is an element-wise modification of the other, the sets of predictions now constitute 'paired data'. In this case, the null hypothesis is that the test set modification does not lead to a significant effect in classifier performance, thus each prediction is equally likely to be in response to the test set element as the modified test set element.

Algorithm 2 Paired permutation test on a series of test results (two-tailed)

```
acc_1                                                          ▷ e.g. human:[94, 91, 91, 89]
acc_2                                                          ▷ e.g. model:[91, 85, 89, 80]
diff = acc_2 - acc_1                                           ▷ e.g. [3, 6, 2, 9]
actual_stat = Mean(diff)                                       ▷ e.g. 5
paired_data = Pair(acc_1, acc_2)                               ▷ e.g. [[94,91], [91,85], [91,89], [89,80]]
simulated_stats = []                                           ▷ create an empty list
for num_iter do
    simulation = Shuffle(paired_data)                          ▷ e.g. [[91,94], [91,85], [91,89], [80,89]]
    sim_acc_1 = Slice(simulation, first)                       ▷ e.g. [91, 91, 91, 80]
    sim_acc_2 = Slice(simulation, second)                      ▷ e.g. [94, 85, 89, 89]
    sim_diff = sim_acc_2 - sim_acc_1                           ▷ e.g. [-3, 6, 2, -9]
    sim_stat = Mean(sim_diff)                                  ▷ e.g. -4
    simulated_stats.append(sim_stat)                           ▷ Add simulated test stat to list
p = Count(abs(actual_stat) ≥ abs(simulated_stats))/num_iter    ▷ p-value
```

In order to compare a classifier to a sample of human participants, we test the classifiers on the same randomly sampled test stimuli presented to participants, producing a series of classification accuracies for both human participants and classifiers that could then be used in a paired permutation test (Algorithm 2). Given these paired performances (human and classifier accuracies on each test set) we then find the mean difference between those performances for our test statistic. The paired permutation test is then used to obtain the null distribution and compare the test statistic to obtain the *p*-value. For comparing the test performances of two classifiers, however, we utilise the McNemar statistic (Everitt, 1977) further explained in Section 2.4.3. In this case, the permutation test instead shuffles binarised prediction pairs, rather than classification accuracy pairs, to obtain the McNemar test statistic null distribution as in Dietterich (1998) and defined in Algorithm 3.

Algorithm 3 Paired permutation test with McNemar test statistic (two-tailed)

```
pred_1 = Binarise(predictions1)                          ▷ e.g. [1, 1, 1, 0]
pred_2 = Binarise(predictions2)                          ▷ e.g. [1, 0, 0, 0]
actual_stat = McNemar(pred_1, pred_2)                    ▷ e.g. 2
paired_data = Pair(pred_1, pred_2)                       ▷ e.g. [[1,1], [1,0], [1,0], [0,0]]
simulated_stats = []                                     ▷ create an empty list
for num_iter do
    simulation = Shuffle(paired_data)                    ▷ e.g. [[1,1], [0,1], [1,0], [0,0]]
    sim_pred_1 = Slice(simulation, first)                ▷ e.g. [1, 0, 1, 0]
    sim_pred_2 = Slice(simulation, second)               ▷ e.g. [1, 1, 0, 0]
    sim_mcnemar = McNemar(sim_pred_1, sim_pred_2)        ▷ e.g. 0
    simulated_stats.append(sim_mcnemar)                  ▷ Add simulated test stat to list
p = Count(actual_stat ≥ simulated_stats)/num_iter        ▷ p-value
```

2.4.3 McNemar test statistic

Investigating the difference between a single trained classifier's performance on a test set under multiple conditions, we sought to obtain a suitable test statistic to best measure any significant changes in performance. Dietterich (1998) found that, for algorithms executed only once (a single trained classifier instance), the only hypothesis test with an acceptable Type 1 error (incorrectly rejecting the null hypothesis) when comparing supervised classification algorithms was the McNemar test (Everitt, 1977).

To obtain the McNemar test statistic for a set of predictions on the same test set (thus the two predictions corresponding to a particular example constitute a pair) one must first obtain the contingency table. For a multiclass classification problem such as ours, we must first binarise the predictions according to whether they were correct.

Table 2.1: Contingency table

(a) number of examples classified correctly by both classifiers	**(b)** number of examples classified correctly by classifier A and misclassified by classifier B
(c) number of examples classified correctly by classifier B and misclassified by classifier A	**(d)** number of examples misclassified by both classifiers

A McNemar statistic can then be obtained using the discordant cells of the contingency table (Table 2.1; top right and bottom left cells are discordant). The equation is shown in Equation 2.1.

$$statistic = \frac{(b-c)^2}{b+c} \tag{2.1}$$

Where b and c correspond to the discordant cells in Table 2.1. This test determines whether the row and column marginal frequencies are equal. Once the test statistic is obtained, it can be compared to the χ^2 distribution, or as in our case, compared to a null distribution generated via a paired permutation test 3.

2.4.4 Bonferroni correction

Most researchers test a null hypothesis with an α level of 0.05, thus accepting a maximum type 1 error rate of 5% (erroneously accepting the null hypothesis). When conducting multiple hypothesis tests on the same sample of data, the family-wise error rate increases (Equation 2.2).

$$\alpha_{family-wise} = 1 - (1 - \alpha_{percomparison})^n \tag{2.2}$$

Where n is the number of comparisons. How to deal with this is a source of disagreement in the literature; a question can be raised about what constitutes a family of analyses (for instance the Bonferroni correction is not generally applied to ANOVA tests) or whether the increase in family-wise error is important (as it is concerned with the global null hypothesis and not the hypothesis in question (Perneger, 1998)). Despite the disagreements in this area, researchers often err on the side of caution against egregious type 1 errors and use a more stringent α criterion for which to compare their p-value (or equivalently, adjust the p-value as we do in this work). The Bonferroni correction multiplies the calculated p-value by the number of comparisons (or divides the α rate). We opt to apply the Bonferroni correction to our p-values in this work.

CHAPTER 3

AUDIOVISUAL MOMENTS IN TIME: A VIDEO BENCHMARK OF AUDIOVISUAL EVENTS FOR MAN AND MACHINE

Contributions: All work including programming, modelling, data collection, analysis and writing were carried out by Michael Joannou with Pia Rotshtein and Uta Noppeney performing supervisory roles.

3.1 Abstract

Exploring natural intelligence with deep neural networks (DNNs) is a growing area of research. Investigations have thus far focussed on the primate visual system by optimising and testing DNNs on image recognition tasks, learning a great deal about visual ventral stream processing and the constraints imposed by biological vision tasks. However, organisms that learn from experience learn from *multisensory* data *sequences*, rather than unimodal data at a single time-point (stand-alone images). Large, labelled action-recognition datasets, leveraging big data on online platforms such as YouTube, provide an interesting opportunity to extend the exploration of biological intelligence via DNNs to the audiovisual domain. In this work, however, we find that even those datasets that contain audio and visual events, seldom focus on *audiovisual* events, where signals are perceived to have a common cause. Thus, where researchers would like to compare DNNs to humans on audiovisual recognition problems, they will not be able to uniformly sample a held-out test set. To this end, we introduce the Audiovisual Moments in Time dataset (AVMIT) for human-DNN comparison; a training dataset of 11,109 videos and held-out test set of 960 videos, across 16 audiovisual action event classes. Candidate videos for AVMIT were selected from the Moments in Time (MIT) dataset and each was classified by 3 trained participants according to whether the video depicted the labelled audiovisual event as a 'dominant' feature of the video. We further provide an extended version of the dataset (AVMIT-VEGAS) that we obtained using clipped videos from the Visually Engaged and Grounded AudioSet (VEGAS) and cleaned with our voting system. AVMIT-VEGAS contains 17,578 audiovisual training videos across 23 event classes and 1,380 held-out test videos.

3.2 Introduction

Deep Neural Networks (DNNs) are now commonly used as predictive models of human behaviour (Cichy and Kaiser, 2019). DNNs require large amounts of labelled data for training, and indeed this has driven many researchers to collect and provide large, labelled datasets (Lin et al., 2014; Russakovsky et al., 2015; Gemmeke et al., 2017). However, in those cases where researchers would like to compare human participants against deep neural networks with naturalistic stimuli corresponding to trained classes, they are confronted with a decision to use a held-out test set from the training dataset itself, or to elsewhere obtain a set of naturalistic stimuli. Indeed, much of the work investigating the human visual system has opted for the former solution; using ImageNet (Deng et al., 2009; Russakovsky et al., 2015) to train and compare test DNNs against human participants with great success (Seibert et al., 2016; Wichmann et al., 2017; Rajalingham et al., 2018; Geirhos, Temme, et al., 2018; Geirhos, Michaelis, et al., 2019; Singer et al., 2020). In part, the successful use of ImageNet as a common ground test set can be attributed to its cleanliness, with quality control implemented via human-annotation. Indeed, researchers often obtain large sets of candidate data samples by crawling several online search engines and then utilising a crowd-sourcing tool such as Amazon Mechanical Turk (Crowston, 2012) to allow human participants to sort or annotate it (Deng et al., 2009) in order to assure a level of cleanliness to their datasets. To account for participant mistakes and disagreements, researchers often have multiple users sort the same samples independently (Deng et al., 2009).

Where researchers would like to extend this area of research to other domains outside of image recognition challenges, they will be confronted with the same decision about whether to use a held-out test set from their training dataset or procure other stimuli for comparison. For instance Kell, Yamins, et al. (2018) compared human and DNN performance on held-out test sets on a music-genre recognition task, The Million Song Dataset (Bertin-Mahieux et al., 2011), and a word-recognition task, TIMIT (Garofolo et al., 1993) and Wall Street Journal (Paul and Baker, 1992) speech corpora, but used an alternate natural sounds test set for their fMRI experiment, of which many samples were not speech or music. Extending into the area of audiovisual perception, current action-recognition video datasets do not provide complementary

stimuli sets for human comparison (Heilbron et al., 2015; Gu et al., 2018; Monfort et al., 2019; Li, Thotakuri, et al., 2020; Smaira et al., 2020). Those that wish to compare humans and deep neural networks in this domain, will thus have to follow the literature and select a held-out test set or somehow procure a set of suitable stimuli (which may or may not belong to the trained labels).

Although large, labelled video datasets (Heilbron et al., 2015; Gu et al., 2018; Monfort et al., 2019; Li, Thotakuri, et al., 2020; Smaira et al., 2020) have been collected in a similar manner to image-recognition datasets (Deng et al., 2009) (using majority votes by human participants) these videos may not be as suitable for use as test stimuli in behavioural or neuroimaging studies. The extension from single images to audiovisual sequences provides additional types of noise. Aside from frame-level noise that can present itself in image datasets (motion blur, dead pixels, addition of watermarks etc.) video datasets contain video-level visual noise such as dropped frames, time-lapses and multiple video panes (e.g. showing a narrator in a small frame overlaid on to the video). There is also audio noise to consider; such as white noise, backing music or narration unrelated to the video label. Indeed, the most popular large, labelled action recognition datasets contain many videos with no audio stream at all or an audio stream containing only contain digital silence.

Even where videos have clear audio and visual signals, however, those signals may not correspond. Many video datasets, such as the Kinetics datasets (Smaira et al., 2020) were annotated by Amazon Mechanical Turkers (Crowston, 2012) according to whether they could *see* the labelled action, with no reference to sound. In the ActivityNet video dataset (Heilbron et al., 2015), Turkers were instructed to select whether activities were present, without referring to audio or visual streams. As such, the videos in modern, large, labelled video datasets are often composed of videos that have qualified as containing visual and/or audio events, but have not been organised in such a way as to contain *audiovisual* events. One video dataset, notable for its activity classes that rely on audio data is the Moments in Time (MIT) dataset (Monfort et al., 2019). Whilst obtaining a vocabulary of actions for which they would collect videos, the researchers building the MIT dataset excluded those that were unlikely to be visual *or* audible

(e.g. thinking), but kept verbs that were unlikely to be visual *and* audible (e.g. knitting, waving or humming). Further, during annotation with Amazon Mechanical Turk (Crowston, 2012), participants selected whether the labelled event was happening in each video, and this included those instances in which the event could only be heard or seen. In this way, the MIT dataset contains many videos depicting *audio* and/or *visual* events, and far fewer depicting *audiovisual* events.

The prevalence of noise and lack of correspondences make many videos in large, labelled video datasets (Heilbron et al., 2015; Gu et al., 2018; Monfort et al., 2019; Li, Thotakuri, et al., 2020; Smaira et al., 2020) unsuitable for controlled human experiments in the audiovisual domain. Indeed, in this work we show that a majority of rated videos from a modern audiovisual video dataset (Monfort et al., 2019) were voted as lacking clear audiovisual correspondences by our trained participants. The additional noise modes provides a clear motivation to support researchers and progress the literature by providing a clean video test set that can be used alongside a modern large, labelled video dataset to compare humans and deep neural networks. Indeed, providing a complementary held-out test set of clean videos depicting *audiovisual* events (where the audio and visual signals pertain to the same event(s)) is the primary objective for this work.

We first sought to select one of the large, labelled action recognition video datasets used in the deep learning literature for which we would provide a held-out test set for human experimentation. In particular, we considered the video length and cleanliness. Audiovisual events occur over a range of different time-periods, which is reflected in the size of the videos across current labelled video datasets with YouTube-8m (Abu-El-Haija et al., 2016) containing videos with an average duration of 230 seconds, Sports-1m (Karpathy et al., 2014) an average of 336 seconds, Kinetics-700-2020 (Smaira et al., 2020) a fixed length of 10 seconds and Moments in Time (Monfort et al., 2019) a fixed length of 3 seconds. Of particular interest in the area of research comparing deep neural networks to humans are those short, 3-second videos, as this duration corresponds to the length of human working memory (Baddeley, 1992; Barrouillet et al., 2004). Indeed, this was the intention behind collecting videos of this length (Monfort et al., 2019). Short

videos are also ideal for human based experiments where repetition of the same type of trial is often required to establish a reliable signal. We thus elect to obtain our candidate videos from the Moments in Time datasets. To obtain our held-out test set of audiovisual events, we carry out an extensive sorting task on these candidate videos using trained participants in a controlled environment.

We further identify a smaller video dataset, previously made available as part of a study whereby researchers focussed on the development of a model that could synthesise sound from visual frames (Zhou, Wang, et al., 2018). In developing the model, the authors found that the cleanliness of the dataset was of paramount importance for the model to be able to generate "convincing" audio for image data that was semantically congruent and somewhat temporally synchronised. To obtain the data they required, the researchers carried out a dataset cleaning task using Amazon Mechanical Turk (Crowston, 2012) to clean AudioSet videos (Gemmeke et al., 2017), producing a dataset named VEGAS (Visually Engaged and Grounded AudioSet). Although these videos are 10 seconds in length, we clip these videos and add them to our own sorting task, producing an extended version of our training and held-out test set, which we call AVMIT-VEGAS.

3.3 Methods

3.3.1 Participants

Eleven participants (10 females; mean age 26.18, range 19-63 years) were recruited and gave informed consent to take part in the video sorting task. No participants were excluded. All reported normal hearing and normal or corrected-to-normal vision. Participants were reimbursed for their participation in the task at a rate of £6 per hour, plus a bonus of 10p paid for correct classification of randomly interspersed ground truths (Further detailed in Section 3.3.5). Participants on average earned a total (hourly payment + bonus) of less than £7 per hour. The research was given a favourable opinion by the University of Birmingham Ethical Review Committee.

3.3.2 Experiment setup

Participants were seated at a desk in an experiment cubicle or quiet area to complete this task. The experiment was presented on a Dell Latitude 5580 laptop with 15.6" screen and Linux Ubuntu 18.04.2 LTS operating system, with no chin rest or other controls for viewing distance or angle. Auditory stimuli were presented via a pair of Sennheiser HD 280 Professional over-ear headphones. The experiment was programmed in Python 2 (Van Rossum and Drake Jr, 1995) and Psychopy 2020.2.10 (Peirce et al., 2019).

3.3.3 Stimuli

All original video stimuli were originally sourced from the training and validation sets of the MIT (Monfort et al., 2019) dataset and the VEGAS (Zhou, Wang, et al., 2018) dataset. We first obtained the labelled training (802,264 videos) and validation (33,900 videos) sets of the MIT dataset. The events depicted in these videos unfold over 3 seconds. For many of the classes in the MIT dataset, audio data would not help recognition of the labelled event (e.g. "imitating", "knitting", "measuring"). We select a subset of 41 audiovisual classes that we consider to have informative audio and visual correspondences such that integration of these signals would aid classification (corresponding to 88,579 training videos and 4,100 validation videos). These action classes are listed in the results section (Figure 3.2). To ensure that videos were audiovisual we removed videos without audio streams or whose amplitude did not exceed 0 (digital silence).

As the sorting task progressed it became clear that surprisingly few MIT videos were classified as clean by our participants, we then sought to increase the number of candidate videos added to the sorting task. Of particular interest were MIT classes that were similar to those selected for AVMIT, we chose to relabel and add those videos to the sorting task. Incorrectly relabelled videos would be filtered from the dataset as participants sorted them according to presence and dominance of the labelled audiovisual event, whereas those videos that were correctly relabelled and clean would be added to AVMIT. Table 3.1 displays those AVMIT classes alongside the other MIT classes that were relabelled and added to the sorting task.

Table 3.1: ‘Bolstering’ MIT classes relabelled as AVMIT classes prior to participant sorting

AVMIT class	Bolstering class
Giggling	Laughing
Frying	Cooking, Boiling
Inflating	Blowing
Pouring	Spilling, Drenching, filling
Diving	Swimming, Splashing
Raining	Dripping

To provide an extended version of the dataset, we prepared video stimuli from the VEGAS (Zhou, Wang, et al., 2018) dataset to be sorted by participants. We first selected 7 of the 10 classes baby crying, fireworks, rail transport, helicopter, printer, snoring, chainsaw to be included in the sorting task with some relabelling. The 3 classes we chose not to include (‘dog‘, ‘water flowing‘ and ‘drum‘) were excluded because of the clear overlap with existing MIT classes (‘barking‘/‘howling‘, ‘pouring‘ and ‘drumming‘) for which enough videos had been classified as ‘clean‘.

As described previously for MIT videos, if participants did not consider the audiovisual events in the relabelled videos to be well described by the label, this would be captured by the sorting task. The VEGAS class ‘chainsaw‘ was relabelled as ‘sawing‘ to bolster the ‘sawing‘ class if participants considered this to be an adequate labelling. Further, while the MIT dataset provides verb labels, the VEGAS dataset uses a mixture of noun and verb labels. As the VEGAS dataset has been built around audiovisual correspondences however, those noun-related labels correspond to some noun-related audiovisual event, rather than just the presence of some object in the videos. With 4 remaining noun labels fireworks, rail transport, printer, helicopter we simply replaced ‘printer‘ with ‘printing‘ prior to adding these videos to the sorting task and left the others in place with no obvious, simple replacements.

VEGAS videos are between 2 and 10 seconds in length and have an average length of 7 seconds. We took a 3 second clip from the centre of videos corresponding to the remaining, relabelled classes, baby crying, fireworks, rail transport, helicopter, printing, snoring, sawing and removed those that less than 3 seconds in duration. Although every 2 second interval of each

Figure 3.1: Video rating task screen displaying a chopping video and accumulated bonus.

original VEGAS video had been cleaned, we added these 3 second clips to our sorting task for homogeneity and allowed our trained participants to vote whether each video clip was suitable for our dataset.

3.3.4 Procedure

With our candidate video set (Section 3.3.3), we next created a video sorting task that could be carried out by multiple trained participants to identify which videos contained the labelled audiovisual event and whether the audiovisual event was a dominant presence in the video. This procedure was similar to annotation procedures carried out in Zhou, Wang, et al. (2018) to produce the VEGAS dataset. In that work, researchers selected a subset of Audio Set (Gemmeke et al., 2017) and used Amazon Mechanical Turk (Crowston, 2012) to verify the presence of the labelled data in both audio and visual streams. The researchers found that this was necessary as their models were unable to generate audio from visual frames due to the lack of correspondences in the original dataset.

Participants observed a series of videos and were instructed to provide a button response after each had finished playing. On each trial, participants were presented with a 3 second video and then classified it as 1:“unclean”, 2:“moderately clean” or 3:“very clean”. To provide a classification, participants were trained to use the following logic:

1. Does the action described by the label appear in the video and do you hear it?:

   Yes: move to question 2

   No: give a 1 rating

2. Did the visual event cause the audio event?:

   Yes: move to question 3

   No: give a 1 rating

3. Is the on-screen, labelled action dominant in both the audio and visual streams?:

   Yes: give a 3 rating

   No: give a 2 rating

For this task, a dominant event was considered to have a longer duration and higher intensity than other events in the same video. Participants were instructed that these labelled audiovisual events should be the focus of the video, in order to ensure that videos whose labelled event was only visible for a short period of the video were removed. Each video was viewed at least 3 times by different participants.

Each screen presented to the participant (Figure 3.1) consisted of the label along the top, the video below and slightly to the left of the label (videos had different resolutions so they were each given a common left edge position and bottom edge position) and a bonus counter in the bottom right. The video would play, and once finished, would disappear and the program would halt until the participant pressed a key. The options were; 1, 2, 3, space, where the numbers referred to the classification system described above and the space key would replay the video. Participants were able to replay the video any number of times they like before making a classification. If the participant made a classification while the video was still playing, a warning screen would fill the display, instructing the participant not to press a key too early. This was particularly important given that the task was asking the participant to assess the "dominance" of the labelled activity in the visual and auditory modalities and so each video should be viewed in its entirety to assess

relative duration of events. After a classification was made, the bonus counter would be updated, and the new label title and video would appear and play as before.

3.3.5 Bonus payments

We provided bonus payments to participants in order to ensure engagement and to provide positive feedback for desired answers. A bonus payment of 10p (GBP) was given for each classification of a video for which a ground truth was available. To obtain ground truths, 2,000 videos were uniformly sampled from the set of candidate videos prior to the sorting task and then classified by the author. These videos were distributed throughout the sorting task and participants were unaware of the possibility of a bonus when completing a trial. If the participant gave a matching classification for one of these previously classified videos, they would receive an bonus, which was added to their total in the bottom right of the screen (Figure 3.1). This bonus accumulated over their sessions and was paid at the end of participation alongside their hourly compensation.

3.3.6 Participant training

In order to ensure the quality of the AVMIT dataset and held-out test set, we opted to use trained participants in a controlled environment rather than Amazon Mechanical Turk. Participants were required to complete the training exercise, detailed here, before they could participate in the sorting task. Before starting, each participant was given a set of instructions to read, outlining the task. These instructions were then verbally explained to them. The participants then undertook a training exercise whereby a video from each class was presented and the possible classification and reasoning was discussed with the author of the study. The participant then went on to classify another set of training videos corresponding to each class under the observation of the author. Of these videos, the participants needed to classify 38 of the 41 videos according to the author's ground truth. Of the 11 participants that completed the training and testing exercise, all participants passed and went on to take part in the cleaning task.

3.4 Results

We present the Audiovisual Moments in Time (AVMIT) dataset and its complementary held-out test set containing labelled audiovisual action events to be used in experiments with deep neural networks and humans. We additionally present the extension to this dataset, AVMIT-VEGAS, containing additional video classes. We further report the participant ratings to reveal characteristics of the MIT dataset and clipped VEGAS videos.

3.4.1 AVMIT

Participants provided 232,593 video ratings throughout the course of this sorting problem, providing data on 77,531 videos across 41 MIT classes and 7 VEGAS classes. The outcome of the participant sorting task on the MIT videos revealed that across all video classes, a considerable number of videos were not rated as containing the labelled audiovisual event by a single participant (Figure 3.2a). Indeed, in those cases where all participants agreed, less than half of the videos in each class were rated as containing the audiovisual event.

Following other datasets in the literature (Deng et al., 2009; Russakovsky et al., 2015; Monfort et al., 2019) we use majority votes as criteria for acceptance into the training dataset. Specifically, we obtain those videos in which the labelled audiovisual event was perceived and considered dominant by the majority of participants. Using this criterion, only 17,904 videos were rated as clean out of the 61,248 MIT videos rated. In this set of clean videos, only 16 action classes contained over 500 videos (Figure 3.3a). Videos from those classes with less than 500 videos were discarded to help ensure that all classes in AVMIT had enough videos to both train and test a deep neural network. Although repeated videos were not reported in Monfort et al., 2019, we find and remove 8 from our clean video subset.

In order to provide a fixed video test set of audiovisual actions, we set as a criterion that all participants must agree that the audiovisual event was perceived and dominant in the video. In order to ensure a level of homogeneity in the dataset, we obtained those videos with a visual frame rate of 30fps and further cleaned them, removing videos that:

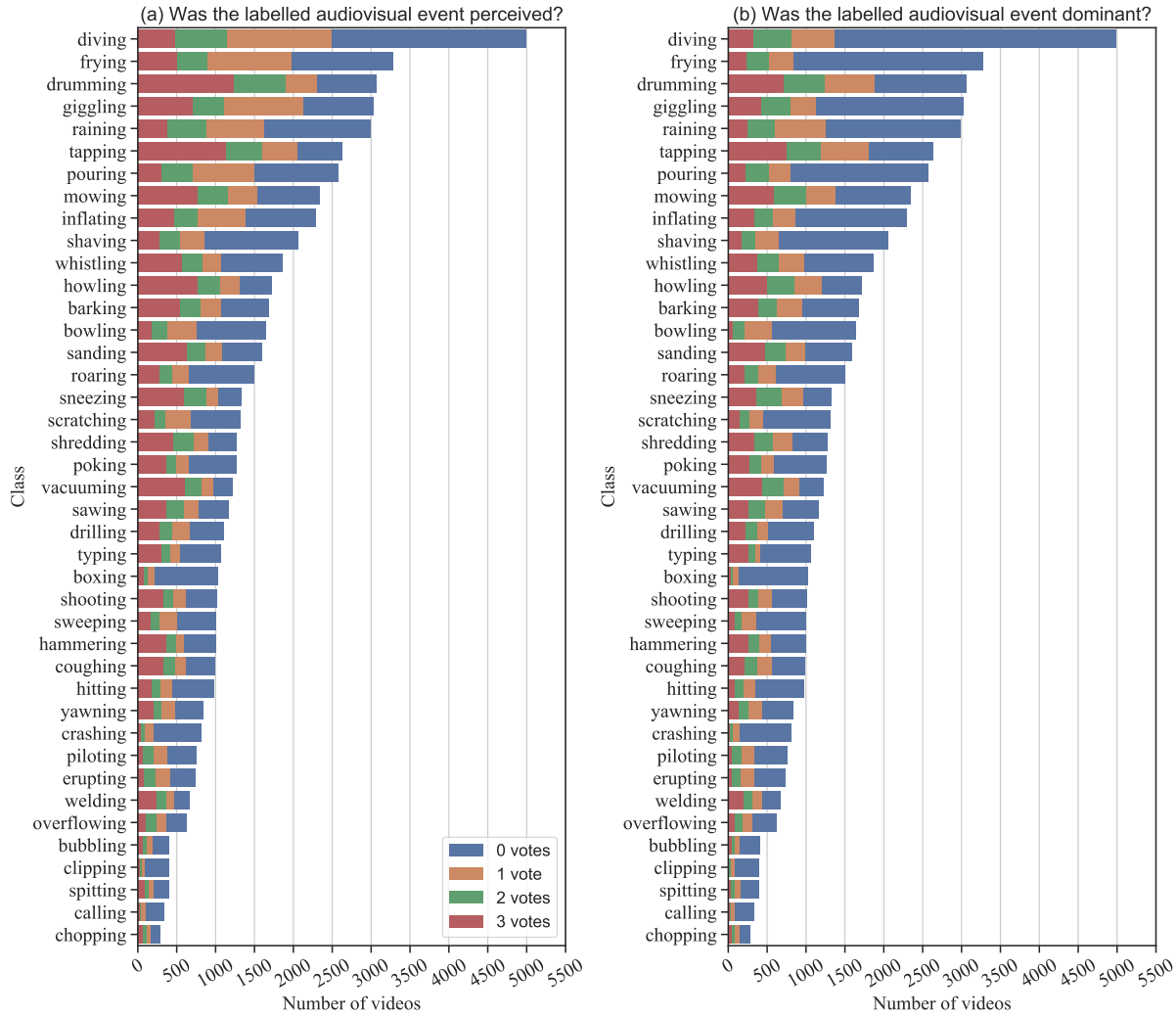
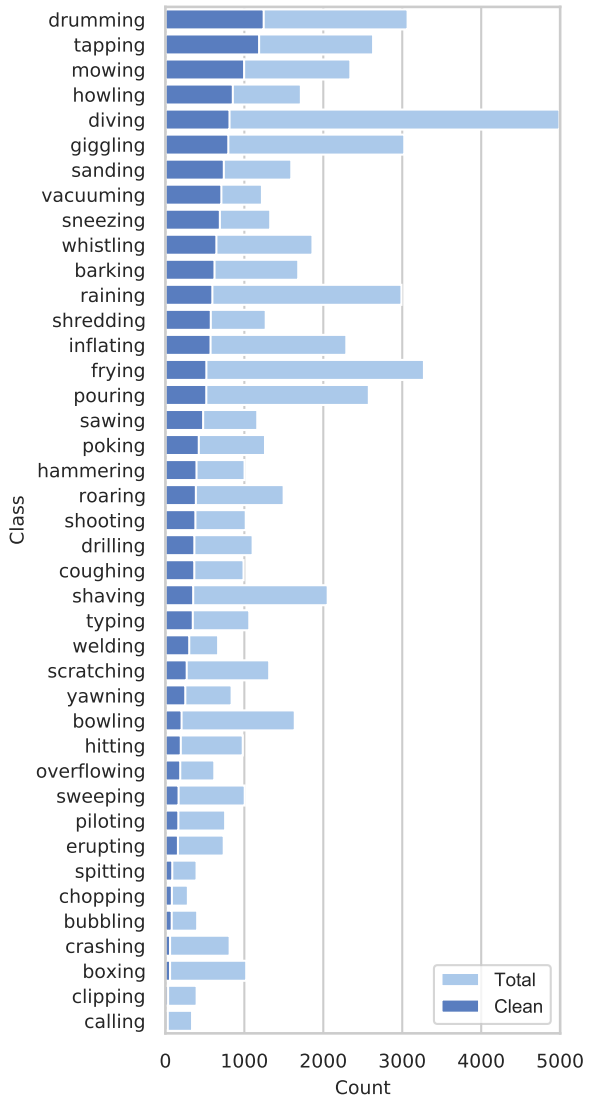


Figure 3.2: Number of MIT videos in each class that obtained a 'yes' vote from 0,1,2 or 3 participants when asked the following questions: (a) Was the labelled audiovisual event perceived? (b) Was the labelled audiovisual event dominant?

- Had been edited to appear as though something supernatural had occurred (such as something appearing or disappearing instantaneously)
- Had an excessive number of time-lapses
- Contained frames with excessive watermarks or writing on the frames
- Consisted of 2 video streams
- Were not naturalistic (depicting cartoons or simulations)

From this subset, 60 videos were uniformly sampled from each class and used to provide a



(a) Proportion of clean videos (majority vote).

Class	Training set video count
drumming	1185
tapping	1127
mowing	937
howling	793
diving	748
giggling	737
sanding	680
vacuuming	649
sneezing	630
whistling	586
barking	562
raining	534
shredding	515
inflating	512
pouring	458
frying	456

(b) Training set video count.

Figure 3.3: AVMIT training set. (a) Number of videos rated as 'clean' by majority of participants as a proportion of the total number of videos rated. (b) The number of training set videos in each class after removal of held-out test set.

complementary AVMIT held-out test set. The remaining videos were returned to the training dataset. The AVMIT training dataset produced in this work contains 11,109 audiovisual videos confirmed by a majority of trained participants to depict the labelled *audiovisual* action as a dominant presence. The final number of AVMIT training videos for each class can be observed in Table 3.5b. The corresponding held-out test set contains a total of 960 audiovisual videos (60 videos per class, 16 action classes), confirmed by all trained viewers to depict the labelled *audiovisual* action as a dominant presence. Balancing across classes provides 7,296 training

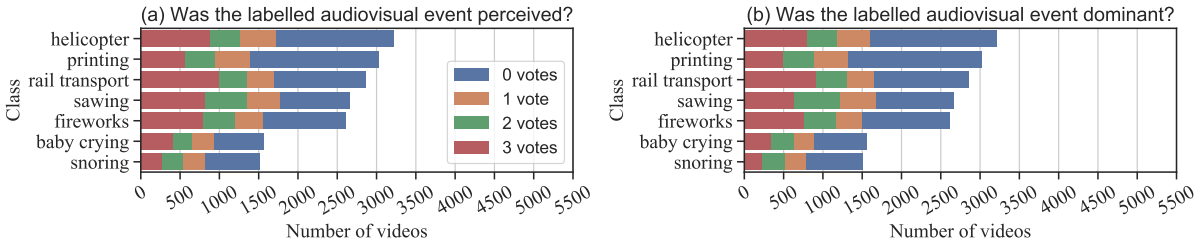
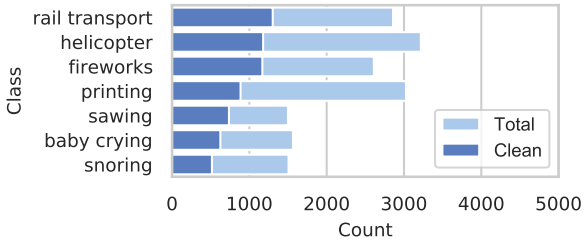


Figure 3.4: Number of VEGAS videos in each class that obtained a 'yes' vote from 0,1,2 or 3 participants when asked the following questions: (a) Was the labelled audiovisual event perceived? (b) Was the labelled audiovisual event dominant?



(a) Proportion of clean videos (majority vote).

Class	Training set video count
rail transport	1242
sawing	1154
helicopter	1118
fireworks	1108
printing	826
baby crying	565
snoring	456

(b) Training set video count.

Figure 3.5: VEGAS training set extension (relabelled). (a) Number of videos rated as 'clean' by majority of participants as a proportion of the total number of videos rated. (b) The number of training set videos in each class after removal of held-out test set.

videos (456 videos per class).

3.4.2 AVMIT-VEGAS

A larger proportion of clipped videos from the VEGAS dataset were rated as containing the labelled audiovisual event than the MIT videos. As 478 'sawing' videos from the MIT dataset had received a majority vote of 'clean', only 22 videos of the relabelled 'chainsaw' VEGAS video clips were required to qualify 'sawing' into the AVMIT-VEGAS dataset. 736 relabelled 'chainsaw' clips were added, providing a total of 1214 'sawing' videos for the AVMIT-VEGAS dataset (1154 training videos and 60 test videos). All VEGAS categories contained over 500 'clean' videos (by majority vote) and so qualified for the AVMIT-VEGAS dataset.

We again applied the same criteria to the VEGAS clips as we did to AVMIT and obtained

a further 420 test set videos (60 videos per action class for 7 classes). In total, this provided a held-out test set for the AVMIT-VEGAS dataset of 1,380 videos. The AVMIT-VEGAS dataset produced in this work contains a total of 17,578 training set videos across 23 action classes, where all videos have been rated as focussing on the labelled audiovisual event. Balancing across classes, there are still 10,488 audiovisual videos in the training set (456 videos per class).

3.5 Discussion

As deep neural networks are increasingly used as investigative tools in cognitive science, researchers will move beyond the area of unimodal recognition tasks into multimodal research. Much of the literature has thus far considered DNNs as models of the ventral visual stream in humans (Yamins, Hong, Cadieu, and Dicarlo, 2013; Yamins, Hong, Cadieu, Solomon, et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014; Güçlü and Van Gerven, 2015; Cichy, Khosla, et al., 2016) and non-human primates (Yamins, Hong, Cadieu, and Dicarlo, 2013; Cadieu et al., 2014), focussing on image recognition tasks. A natural progression from this area would be to extend into the audiovisual domain, given that audiovisual integration is an established area of research in cognitive science (Stein and Meredith, 1993; Stein, 2012) and there is an abundant source of audiovisual data online from sources such as YouTube. Several large, labelled video datasets for action recognition have been released (Heilbron et al., 2015; Gu et al., 2018; Monfort et al., 2019; Li, Thotakuri, et al., 2020; Smaira et al., 2020) which can be used to train DNNs and these could be compared against humans.

Current action recognition datasets (Heilbron et al., 2015; Gu et al., 2018; Monfort et al., 2019; Li, Thotakuri, et al., 2020; Smaira et al., 2020), however, have collected videos of labelled audio and/or visual events, but not specifically audiovisual events. It is this important distinction, particularly in participant instruction during dataset annotation, that likely leads to low levels of audiovisual correspondence. For instance, the Moments in Time dataset (Monfort et al., 2019) instructed Amazon Mechanical Turkers to ‘press a Yes or No key signifying if the action is happening in the scene‘ without requesting that the activity be detectable in both audio and visual

modalities. It should be noted that the intention of the researchers collecting the MIT dataset was not to collect a dataset of audiovisual events, so these instruction served their purpose and resulted in the collection of a large dataset with wide coverage and diversity of action events. Another example is the second dataset studied in this work, the VEGAS dataset (Zhou, Wang, et al., 2018). Participants on Amazon Mechanical Turk were instructed to verify the label in the audio *and* visual domain in each video, but the signals in each modality could both potentially pertain to the label but not be caused by the same event. In this way, the sound of a dog barking in the background of a video followed by a dog moving its mouth the in the visual stream could correctly be labelled 'dog' although the auditory signal came from a different location at a different time to the visual signal.

Our study found that, despite filtering for action classes considered to have useful audio and visual signals, and despite filtering out videos without audio streams or with digital silence, only 17,904 videos out of 61,248 MIT videos were classified as containing a properly labelled, dominant, audiovisual event by a majority of our trained participants. Perhaps more surprising was the low number of VEGAS (Zhou, Wang, et al., 2018) videos containing properly labelled, dominant audiovisual events. With just 6,411 videos out of 16,283 videos verified by a majority of our trained participants to contain properly labelled, dominant, audiovisual events. This is despite the Amazon Mechanical Turk annotation for each 2 second clip throughout each video.

Those researchers who would like to study DNNs and humans using videos depicting audiovisual events are thus without a clear benchmark. This means that their research must either; use a held-out test set from the a popular labelled video dataset such as ActivityNet (Heilbron et al., 2015) or Moments in Time (Monfort et al., 2019) or they must collect or otherwise obtain a set of videos depicting audiovisual events. The primary aim of this work was to provide researchers with a clear, held-out test set that could be used alongside a large, labelled action recognition training dataset to compare DNNs and human participants. We achieved this by selecting action classes from the MIT dataset that were deemed to contain useful audio and visual signals, removing those video files with no audio signals and then further carrying out a exhaustive video sorting task to ensure the labelled audiovisual event was both present and

dominant according to 3 trained participants. We further utilise the data collected during our sorting task to create a corresponding training dataset. The dataset was constructed using action classes with over 500 videos that had been voted by the majority of participants as containing the labelled audiovisual event as the dominant presence. This training dataset (11,109 videos) and corresponding held-out test set (960 videos) contain 16 audiovisual action event classes. We name these video sets the Audiovisual Moments in Time (AVMIT) dataset and held-out test set.

By utilising another published dataset, VEGAS (Zhou, Wang, et al., 2018), we were able to provide an extended version of this dataset, AVMIT-VEGAS. A selection of 7 VEGAS classes were clipped (obtaining the central 3 seconds of the video), in some cases relabelled, and then added to the sorting task for participants to rate. With 1 of these classes 'chainsaw' bolstering an existing MIT class 'sawing', this provided an additional 7 classes (as 'sawing' now qualified for the dataset) in this extended version. This added an additional 6,469 training videos (where 456 of these videos are from the 'sawing' MIT class) and an additional 420 held-out test videos.

Although this work successfully produced a held-out video test set of audiovisual events for humans and DNNs, the corresponding training dataset has some limitations to its usage. One notable limitation is the size of the training dataset, the balanced extended dataset contains 10,488 videos across 23 classes, but for multimodal problems with large feature spaces it is perhaps preferable to use larger datasets. The original MIT dataset (Monfort et al., 2019) for example contains 1 million labelled videos across 339 classes. Researchers will have to consider the required dataset size for their DNNs and their ultimate goals for learnt behaviour (how important the audiovisual correspondences in the data are for example). One potential solution would be to use a large dataset to pretrain the model and then fine-tune the model on AVMIT or AVMIT-VEGAS, indeed transfer learning (Pan and Yang, 2010; Tan, Sun, et al., 2018; Wang, Gao, et al., 2019) is a common strategy to overcome small datasets.

CHAPTER 4

MULTISENSORY INTEGRATION IN DUAL-STREAM RECURRENT CONVOLUTIONAL NEURAL NETWORKS

Contributions: All work including programming, modelling, data collection, analysis and writing were carried out by Michael Joannou with Pia Rotshtein, Uta Noppeney and Bernd Bohnet performing supervisory roles.

4.1 Abstract

Humans utilise data from multiple sensory organs to operate in a multisensory world. To operate successfully, they must use this data to recognise events unfolding in their environment. Although humans solve this problem effortlessly, multisensory perception is not a trivial problem. Data arriving across different sensory modalities can be redundant (pertaining to the same event) and thus integrated, or complementary (pertaining to different events) and thus segregated in the brain. This is known as the multisensory correspondence problem or causal inference problem. The brain is never explicitly presented with the ground truth (the causal structure) of the world, yet the ability to solve this problem using spatiotemporal and higher order cues is either learnt through experience or encoded in the genome. In this work, we develop a set of 6 novel dual-stream recurrent convolutional neural networks (DRCNNs) and ask 'Is it possible that a DRCNN, optimised to solve an audiovisual action recognition task, implicitly learns to solve the audiovisual correspondence (AVC) task using semantic cues?'. Our findings show that all action recognition trained DRCNNs were capable of solving the AVC task, including those instances optimised on data with lower levels of audiovisual correspondence. We further explore the interaction of audio and visual signals in the learnt audiovisual embeddings using a cross-modal learning task and a shared-representation learning task. We observe that by fitting the SVM to the embedded signals of one modality, information presented in the other modality is still captured by the SVM. We further explore this interaction by introducing two of our tasks; the congruent and incongruent selective-attention tasks, parallel to those in the area of psychology. We find a significant decrease in performance accuracy when incongruent information is presented alongside the attended-to modality, and a significant effect in 21/24 instances when congruent information is presented. The interaction of the signals across modalities in the audiovisual embedding is demonstrative that activations are not held separate throughout processing.

4.2 Introduction

Our lives are inherently multisensory, with our brains making use of data collected from a number of sensory organs in order to perceive the environment and take actions within it. Using multisensory data to perceive the environment is not a simple problem, indeed the area of multisensory integration is an active area of research in the fields of cognitive science (Stein and Meredith, 1993; Stein, 2012; Noppeney, Jones, et al., 2018; Mihalik and Noppeney, 2020) At its core, the problem is that redundant and complementary multisensory information are received by the human brain, which must then infer the causal structure of the data (the binding problem), and integrate data from common sources into a percept that maximises its effectiveness.

An unresolved question in cognitive science is how the brain solves the multisensory binding problem (Mihalik and Noppeney, 2020) (also known as the causal inference or correspondence problem). Although humans solve this problem effortlessly, the brain has no explicit information about the source of a signal, yet must make inferences based on sensory data. Research suggests that human observers solve the binding problem in line with Bayesian causal inference (BCI) (Körding et al., 2007) taking into account the uncertainty about the source of the signals and, in the case of integration, to what extent signals should be integrated according to their reliability. Indeed, humans are highly optimised for this task, but how did the solution come to be encoded in our brains?

It is possible that the ability to solve the multisensory causal inference task can arise implicitly from optimising on an ecologically-relevant multisensory recognition tasks. However, where traditional *ideal observer* models are typically used to explore statistically optimal solutions to behavioural tasks in psychology, they are often intractable for ecologically-relevant tasks with naturalistic stimuli. In order to explore the extent to which a learner is capable of solving a multisensory correspondence problem, after only being optimised on a multisensory action recognition problem, we leverage developments in the field of *deep learning* (Krizhevsky et al., 2012; Cireşan et al., 2012; Wan et al., 2013; Sun, Chen, et al., 2014; Russakovsky et al., 2015).

Deep learning describes a class of techniques for learning hierarchical representations and complex, composite functions using connectionist models known as *deep neural networks*

(DNNs). DNNs are composed of artificial neurons, organised into layers, interspersed with other operations such as pooling (Riesenhuber and Poggio, 1999; Krizhevsky et al., 2012) and BatchNorm (Ioffe and Szegedy, 2015) that can be trained end-to-end using backpropagation and gradient descent. Each neuron in the network has a set of trainable parameters (akin to synaptic connections in the brain) that provide a weighting of the input components by performing a dot product, the weighted inputs are then summed and passed through a non-linear activation function. The models are described as 'deep' because of the many layers of neurons utilised in order to obtain human-level performance on a number of naturalistic classification tasks (Krizhevsky et al., 2012; Cireşan et al., 2012; Wan et al., 2013; Russakovsky et al., 2015; He et al., 2015; Zhang, McLoughlin, et al., 2015; McLoughlin et al., 2015; Phan et al., 2016; Takahashi et al., 2016; Laffitte et al., 2016; Parascandolo et al., 2016)

The ability of DNNs to obtain human-level performance has further motivated researchers to explore the extent to which they model human intelligence. Indeed, work in the area of neuroscience has revealed a number of DNNs to be highly predictive of the visual ventral stream (Yamins, Hong, Cadieu, and Dicarlo, 2013; Cadieu et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014; Güçlü and Van Gerven, 2015; Cichy, Khosla, et al., 2016) and auditory cortex (Kell, Yamins, et al., 2018), revealing hierarchical correspondences between early and late stages of the DNNs and the cortices. Although neuroimaging is outside the scope of this work, the emergence of correspondences between task-optimised DNNs and task-associated brain regions adds credence to DNNs as models of biological intelligence. Indeed, DNNs have also been used to explore human visual (Dodge and Karam, 2016; Dodge and Karam, 2017; Wichmann et al., 2017; Geirhos, Temme, et al., 2018; Dodge and Karam, 2019; Heinke et al., 2021), auditory (Kell, Yamins, et al., 2018) and multisensory behaviour (Rideaux et al., 2021) beyond just task performance by considering error patterns and generalisation ability to tasks outside the training domain.

As DNN models can be optimised on naturalistic classification tasks and reach human-levels of performance, we are able to treat DNNs as approximations of ideal observers (although they are not provably optimal). Exploring their learned behaviour can thus help us understand the

role of task constraints in shaping behaviour (Kell and McDermott, 2019). In this work, we explore the extent to which a DNN is capable of solving an audiovisual correspondence (AVC) problem, after being optimised on an audiovisual action recognition problem. Of course, learned behaviour could also be a product of the algorithm itself. We do not try to tease apart these factors in this work and instead develop a set of (similar) DNNs and train them on two different datasets in order to help understand these effects outside of specific architectures.

Obtaining behavioural data for a DNN in response to input data outside its training domain is a simple task once the stimuli have been selected (given that the input data is of the same shape). An example of this is a DNN, trained to classify cats and dogs, presented at test time with images of lions and wolves, or perhaps just degraded images of cats and dogs. This process can provide valuable data on the ability of the DNN to generalise to data given some particular domain shift (or distributional shift). However, in the case that the DNN is to be tested on a different task with different outputs (*transfer learning*), a change must be made to the model. This is the case in our work, we would like to train a DNN on an audiovisual action recognition task and then test the model on an audiovisual correspondence task. To successfully understand the utility of learnt DNN representations to a new task, one must replace the final softmax classification layer, responsible for providing the final output probability distribution for the trained task, with another model trained to map the embeddings to the new task. However, to ensure that there is no additional non-linear fitting capacity added at this stage, a linear *support vector machine* (SVM) can be utilised (Ngiam et al., 2011).

While this provides a means by which we can explore the ability of an audiovisual recognition trained model to solve the audiovisual correspondence task, if the DNN learns to hold separate activations for audio and visual signals, an SVM would only have to learn a simple AND logic function (i.e. audio drumming AND visual drumming corresponds). Acknowledging this, we explore the extent to which audio and visual representations are integrated by implementing the 'hearing to see' and 'seeing to hear' tasks from (Ngiam et al., 2011) and further explore how effective the audio and visual representations are alone by implementing the'cross-modality learning' tasks (Ngiam et al., 2011).

To further explore audiovisual integration in our classifiers, we carry out two novel tasks in the area of deep learning that we consider to provide a parallel to *selective-attention* tasks in the area of psychology (Yuval-Greenberg and Deouell, 2007). In Yuval-Greenberg and Deouell (2007), participants were presented with image and audio clip combinations and asked 'Which animal do you see?' or 'Which animal do you hear?'. In each presentation, the image and the audio clips were either semantically congruent (corresponding to the same animal) or incongruent (corresponding to different animals). In the selective-attention tasks we introduce, the classifiers are also tasked with classifying a single modality of presented multimodal stimuli, as the participants are in Yuval-Greenberg and Deouell (2007) (thus 'selective-attention'). But our tasks are distinct to those in Yuval-Greenberg and Deouell (2007) in that the stimuli are *video* and audio clip combinations and the classification task is *action recognition*.

To implement this with DNNs, we train individual SVM instances on embedding-action label pairs, where one SVM instance is trained on embeddings of audio-only data (attend audio) and the other SVM instance is trained on the visual-only embeddings (attend visual). The SVMs are then tested on their trained modality alongside congruent or incongruent data in the other modality to assess the effect on performance.

In this work, we introduce a hybrid DNN called a 'dual-stream recurrent convolutional neural network' (DRCNN). The DRCNNs we develop in our work are 'dual-stream' because we utilise a convolutional neural network (CNN; LeCun, Boser, et al., 1989; Krizhevsky et al., 2012) for visual feature extraction and another CNN for audio feature extraction before creating a joint audiovisual embedding that is passed to a recurrent neural network (RNN) at each time-step. This extends on a line of hybrid architectures that combine convolutional and recurrent operations in a number of ways such as; adding feedback connections to CNN layers (Wang, Lei, et al., 2020), replacing fully-connected nature of LSTMs with convolutional operations (Shi et al., 2015) and connecting the output of a single CNN to an RNN (Donahue et al., 2015; Ning et al., 2017; Çakır et al., 2017; Sabir et al., 2019; Khaki et al., 2020; Gupta et al., 2021). The name of hybrid architectures composed of a CNN followed by an RNN is inconsistent in the literature, and have been referred to as CNN-RNNs (Khaki et al., 2020), RCNNs (Ning et al., 2017; Gupta

et al., 2021), CRNNs (Çakır et al., 2017) and recurrent convolutional models (Sabir et al., 2019). In this work, we use the term RCNN to broadly refer to any architecture implementing both convolutional and recurrent operations.

We develop a series of 6 DRCNNs in our work, each with a combination of 1 of 2 audiovisual feature extractors and 1 of 3 RNNs; fully-recurrent neural network (FRNN, also known as a 'basic' or 'vanilla' RNN), gated recurrent unit (GRU; Cho et al., 2014) or a long short-term memory unit (LSTM; Hochreiter and Schmidhuber, 1997). The first audiovisual feature extractor is built by first obtaining VGG-16 (Simonyan and Zisserman, 2015) as a visual feature extractor and VGGish (Hershey et al., 2017) as an audio feature extractor. The second audiovisual feature extractor uses EfficientNetB0 (Tan and Le, 2019) for visual features and YamNet (Plakal and Ellis, 2020) for audio features. Both visual CNNs were trained on ImageNet (Deng et al., 2009; Russakovsky et al., 2015) and audio CNNs were trained on (Gemmeke et al., 2017). Audio and visual embeddings were both fused at each timestep using a novel audiovisual 'squeeze' bottleneck before input to the RNN.

To train the models, we use the balanced version of the AVMIT dataset (Chapter 3) of 16 action classes, verified by a majority of trained participants to contain the labelled audiovisual action as the dominant event. We consider that the high level of audiovisual correspondences in the training data may be a necessary component for the DNNs to learn about important semantic correspondences across the audio and visual streams. To explore whether the models will implicitly learn to solve the audiovisual correspondence problem even when these correspondences in the training data are reduced, we create a corresponding video dataset from Moments in Time (MIT) (Monfort et al., 2019), by sampling the largest possible balanced dataset according to the AVMIT action classes. We call this dataset MIT-16. An instance of each DNN is then trained on AVMIT and a second instance is trained on MIT-16. All trained classifiers are tested on the controlled AVMIT test dataset, allowing us to have fine-grained control over the level of correspondence in the test videos.

In this work, we investigate whether a series of dual-stream RCNNs, trained on an audiovisual action recognition task, are able to solve the audiovisual correspondence problem. We further

explore whether the increased audiovisual correspondence of the AVMIT dataset is a necessary component to implicitly learn to solve the audiovisual correspondence problem. Next, we explore the behavioural interaction between the audio and visual data using a series of (aforementioned) selective-attention tasks. We assess the unimodal classification accuracy and then use this as an informative baseline to understand the behavioural consequences of the presenting stimuli in the unattended to modality. Data is provided in the unattended to modality either alone (shared-representation learning) or alongside data in the attended to modality (this can be congruent or incongruent). We repeat this procedure to explore effect of each modality on the other. Our hypotheses for this work are that; DRCNNs trained on AVMIT will be able to solve the audiovisual correspondence problem, DRCNNs trained on MIT-16 will not be able to solve the audiovisual correspondence problem, DRCNNs will obtain above chance performance on the shared representation task, DRCNNs will achieve higher classification accuracies when congruent stimuli is provided in the unattended to modality, DRCNNs will obtain lower classification accuracies when incongruent stimuli is provided in the unattended to modality. In the following sections, we outline our methods, present our results and then discuss this work.

4.3 Methods

4.3.1 Software packages

All models were developed and tested with Python 3.7.9 (Van Rossum and Drake, 2009) and TensorFlow 2.3.1 (Abadi et al., 2015). For audio preprocessing we used the python packages Pydub 0.24.1 (Robert, Webbie, et al., 2018), SciPy.signal (SciPy version 1.5.4) (Virtanen et al., 2020) and to resample we use resampy (McFee, 2016). For visual preprocessing we used OpenCV 4.4.0.44 (Bradski, 2000). For testing with SVMs, we use sklearn (Pedregosa et al., 2011).

4.3.2 Model development

We present a set of dual-stream RCNNs developed as part of this study. At each timestep, each model preprocessed a video frame and produced a spectrogram from an audio clip, passed preprocessed data to corresponding unimodal CNNs to extract feature embeddings, flattened 2D embeddings and produced an audiovisual embedding using our 'multimodal-squeeze' unit. Across the whole video, these audiovisual representations at each timestep provide a sequence which is then passed to an RNN. The final output of the RNN is passed to a fully-connected softmax layer to provide the output distribution over the action classes.

The CNNs were first selected in order to obtain 2 audiovisual feature extractors. For the sake of variety, we obtain 2 more 'simple' VGG style architectures to make the first audiovisual feature extractor, and 2 more modern MobileNet/EfficientNet style architectures to make the second. VGG-16 (Simonyan and Zisserman, 2015) and EfficientNetB0 (Tan and Le, 2019) were chosen as the visual feature extractors, the audio feature extractors were VGGish (Hershey et al., 2017) and YamNet (Plakal and Ellis, 2020). Both visual CNNs were trained on ImageNet (Deng et al., 2009; Russakovsky et al., 2015) and both audio CNNs were trained on Audio Set (Gemmeke et al., 2017). For each CNN, we remove the final softmax layer, responsible for the output distribution on the original trained task. Where those CNNs terminate in 2D embeddings, we further add a Global Average Pooling operation to reduce it to a 1D embedding, in the case of the VGGish model, output embeddings are already 1D after a fully connected layer.

Much of the audio preprocessing was identical prior to processing by the VGGish and YamNet feature extractors. For stereophonic audio, we first obtain a monophonic stream using pydub.AudioSegment.set_channels() (Robert, Webbie, et al., 2018), which produces a new audio stream equal to the mean of the left and right channels (Equation 4.1).

$$S_{new} = 0.5 \cdot S_L + 0.5 \cdot S_R \tag{4.1}$$

Where S_{new} is the new monophonic audio sample, S_L is the original left sample and S_R is the original right sample.

The audio data was then cast to a depth of 16 bits if that was not already its current bit depth using pydub.AudioSegment.set_sample_width() (Robert, Webbie, et al., 2018). These int16 audio samples are then mapped from the range [-32768, 32767] (2^{15} with one bit dedicated to sign) to the range [-1.0, 1.0] by dividing by the maximum value of 32768.0. They are finally resampled to 16 kHz before spectrograms are calculated.

We next carry out a short-time Fourier transform (STFT) to provide a frequency decomposition over time. We use a frame size of 25ms (the period over which signals are assumed to be stationary) and a 10ms stride (the frequency with which we obtain a frame). The overlapping frames help to ensure that any frequency in the signal that may exist between otherwise non-overlapping frames are captured in the spectrum. A Hann filter is then applied to each of the frames before a fast Fourier transform (FFT) is carried out. A log mel spectrogram is then obtained by using a mel filter bank of 64 filters, over the range 125-7500 Hz, and then finding the logarithm of each spectrum (plus a small delta of 0.01 to avoid taking the log of 0; Equation 4.2).

$$log\ mel\ spectrogram = log(mel\ spectrogram + 0.01) \tag{4.2}$$

The log mel spectrograms are then windowed into smaller 960ms spectrograms, ready for the CNN. But this last point of audio preprocessing, prior to input into the CNNs, is where the operations deviate between the VGGish model and the YamNet model. For VGGish, the stride is 960ms between windows, such that the input spectrograms (and thus the audio feature outputs) have no overlap. For YamNet, the stride is 480ms, such that there is a 50% overlap between input spectrograms and YamNet output embeddings.

To preprocess the visual video frames before input to the CNN, we first sampled frames to align with the audio sample rate. For EfficientNetB0, we sampled a visual frame every 480ms (the same sample rate as the audio frames), for VGG-16 we sampled frames every 960ms. If there was an additional sample taken in either modality, we would clip this from the stream. Frames were then resized to dimensions of 224x224x3 using OpenCV (Bradski, 2000) in line with the expected input size of the CNN models. For VGGish the images were then zero centred, but for EfficientNetB0, images were rescaled, normalised and then zero-padded.

Once data is preprocessed and passed through the unimodal CNNs, the unimodal embeddings must be joined to create an audiovisual embedding. This must happen at each timestep in order to create a sequence of audiovisual representations that can be modelled by the RNN. Further, so as not to bias a particular modality, we required that audio and visual representations were of equal size prior to concatenation such that an equal number of trainable parameters (and thus fitting capacity) is provided to each modality during audiovisual processing in the RNN. To solve this problem, we implement a series of operations (not unlike the initial 'squeeze' of a 'squeeze-excitation' block used throughout EfficientNet models (Tan and Le, 2019) and other state of the art CNNs) to reduce the audio and visual embeddings down to a common bottleneck size. We refer to this set of operations as a *multimodal squeeze unit* and they are presented in our general model diagrams (Figures 4.1 and 4.2). First, we expand each embedding to have 2 additional dimensions of size 1 and utilise a 1x1 2D convolution with batch normalisation and an activation function to allow the model to learn a non-linear mapping to a different embedding size, before again running a Global Average Pooling operation and concatenating the audio and visual embeddings ready for input into the RNN. The exact size of the squeeze was not decided during architecture development but rather the result of a later hyperparameter search (Section 4.3.4).

The audiovisual representations are then fed at each timestep an RNN. We use 3 different RNN models; FRNN, GRU and LSTM. Alongside the variation in audio and visual CNNs, this provides us with 6 different dual-stream RCNN architectures that better serve us to understand the ability of a dual-stream RCNN, optimised on audiovisual action recognition task, to implicitly learn to solve the audiovisual correspondence task. We additionally add dropout to the input units of the RNN at a rate decided via the hyperparameter search (Section 4.3.4. The RNN is then followed by a fully connected SoftMax classification layer with 16 units, corresponding to each of the 16 AVMIT action classes, to provide an output distribution.

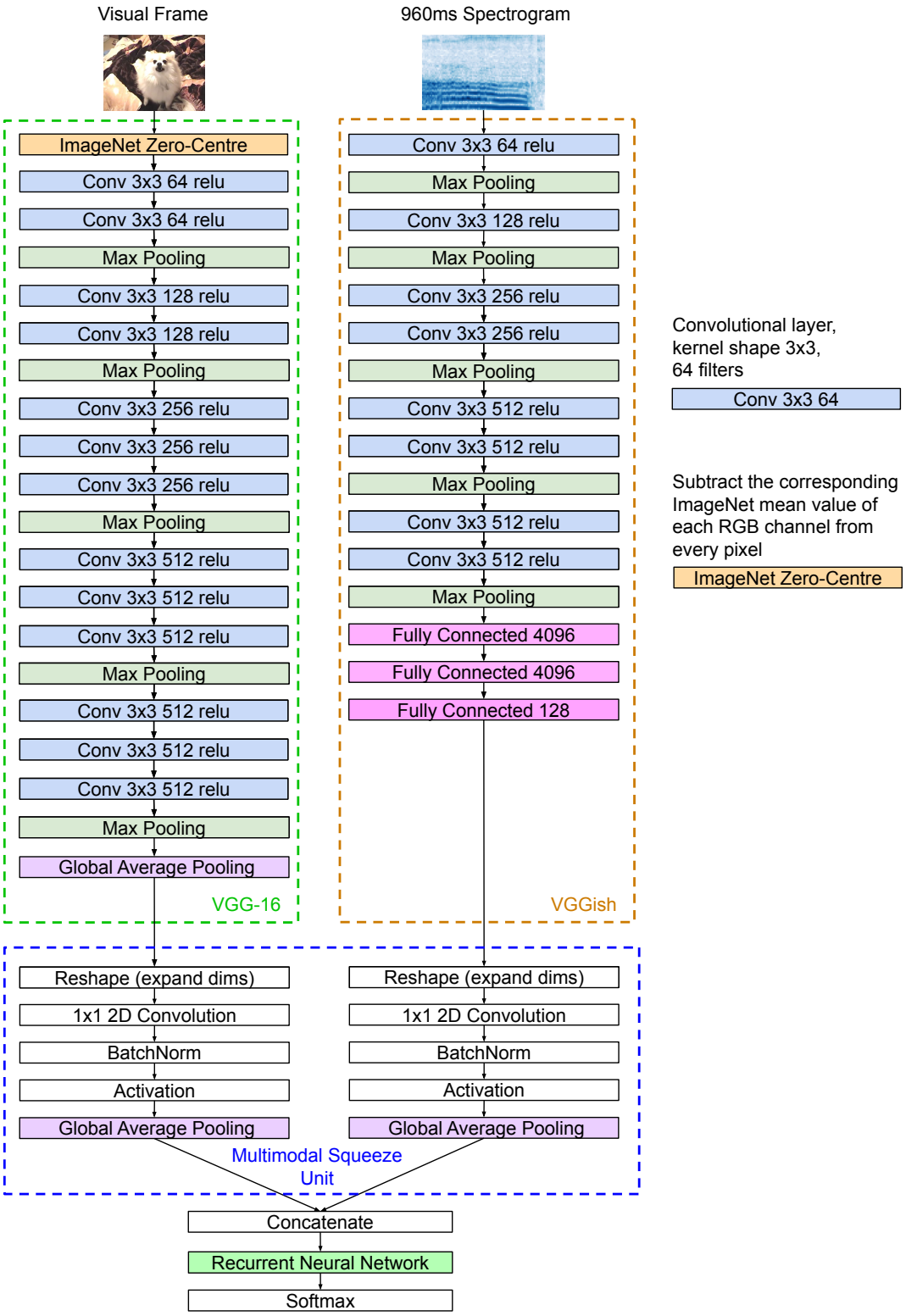


Figure 4.1: VGG-16 + VGGish Dual-Stream Recurrent Convolutional Neural Network.

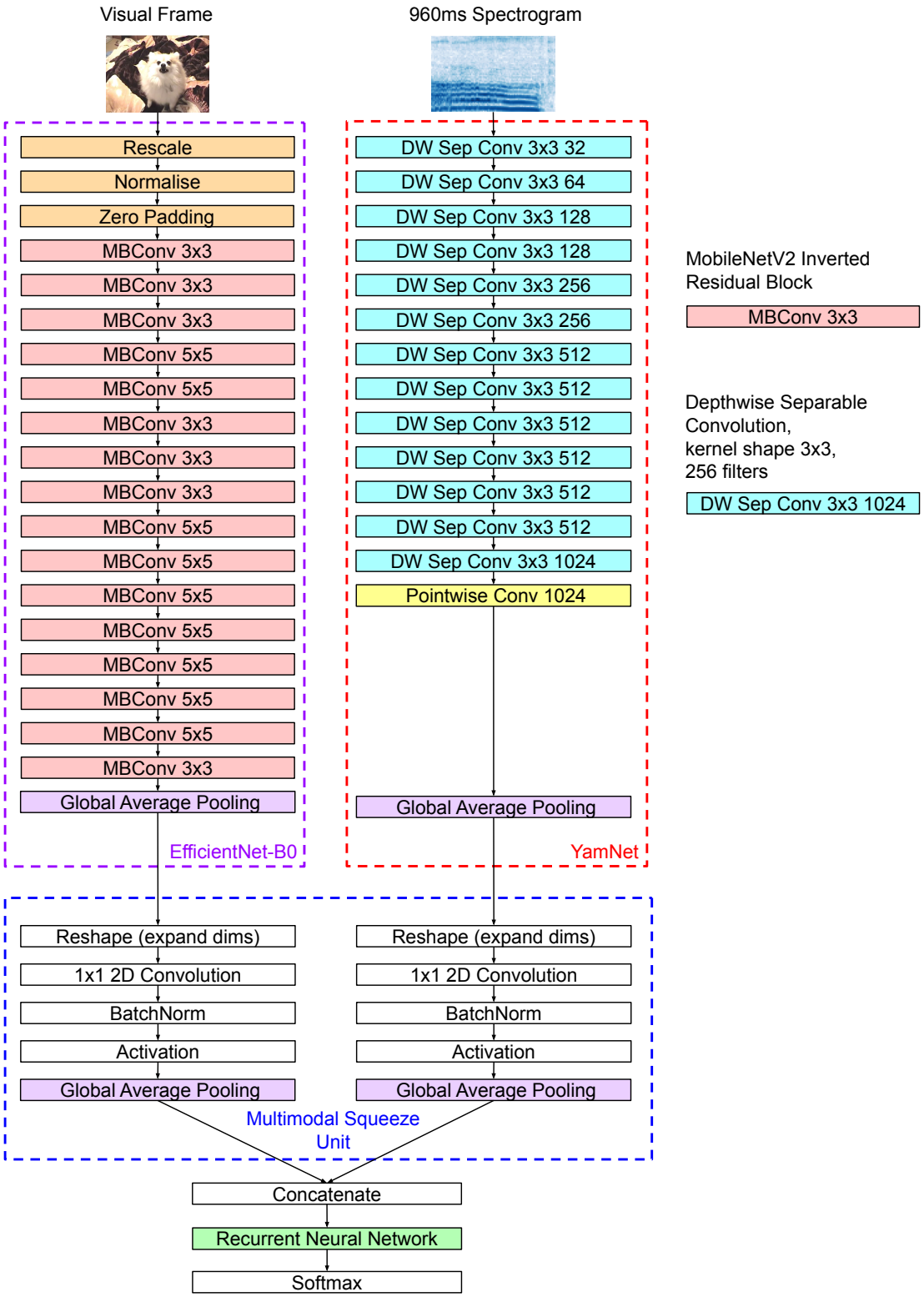


Figure 4.2: EfficientNet-B0 + YamNet Dual-Stream Recurrent Convolutional Neural Network.

4.3.3 Training data

As previously mentioned, we utilise the balanced version of the AVMIT training set for its high level of audiovisual correspondences (Chapter 3), which we assume is a necessary component of our DRCNNs to implicitly learn about semantic correspondences across modalities. To obtain a similar training dataset with lower levels of correspondence, we first obtain the original MIT videos corresponding to the AVMIT classes. From this MIT subset, we sample the maximum number of videos from each class, such that the subset is balanced across classes (1,406 videos) and refer to this training set as 'MIT-16' throughout our work. An instance of each DRCNN model is trained on AVMIT and another instance of the model is trained on MIT-16, providing us with 12 classifiers for our experiments. This should allow us to observe whether the higher level of correspondence in the AVMIT training set is necessary to implicitly learn to solve the audiovisual correspondence problem.

4.3.4 Hyperparameter search

The classifiers trained in this work are all tested on the AVMIT test dataset, which contains a higher level of correspondence than the complementary AVMIT training dataset and MIT-16 as voted by a trained participants (Chapter 3). Although our experiment was to measure out-of-domain performance, we sought to find a single set of hyperparameters for each model that would increase the likelihood of high performance levels on the audiovisual action recognition task. We carried out a hyperparameter search (random search with bootstrapping (Efron and Tibshirani, 1986)), creating 300 surrogate models per DRCNN, each with a particular combination of hyperparameter values that were uniformly sampled from provided sets or intervals.

We searched over the following hyperparameters; number of filters, $n_{bottleneck}$, in the 1x1 2D Convolution in the audiovisual bottleneck where $n_{bottleneck} \in \{32, 64, 128, 256\}$, the activation function, a, of the audiovisual bottleneck, where $a \in \{relu, swish\}$, the number of Recurrent Neural Network units, n_{RNN}, where $n_{RNN} \in \{32, 64, 128, 256\}$, the dropout rate, d, for the RNN, where $d \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, the learning rate, l, of the model, where

$l \in [1.0x10^{-5}, 5.0x10^{-4}]$. During the random search, models were trained in the same manner (Adam optimiser (Kingma and Ba, 2015) and exponential learning rate decay) as during final training, the only exception being that the early stopping patience was reduced from 20 epochs to 8 in order to save time during the random search. The best performing configurations for each model (Table 4.1) were then selected for all experiments.

Table 4.1: Hyperparameter Search Results: Selected Hyperparameters

Feature extractor	RNN	Units	Bottle	Act.	Drop	LR	Trainable Params
YamNet + EffNetB0	FRNN	128	256	swish	0.3	$7.05x10^{-5}$	675,472
YamNet + EffNetB0	GRU	128	64	swish	0.5	$7.25x10^{-5}$	248,976
YamNet + EffNetB0	LSTM	64	256	swish	0.3	$4.10x10^{-5}$	740,112
VGGish + VGG-16	FRNN	256	256	swish	0.4	$1.05x10^{-4}$	366,352
VGGish + VGG-16	GRU	128	256	relu	0.5	$3.92x10^{-4}$	413,968
VGGish + VGG-16	LSTM	256	256	swish	0.5	$1.74x10^{-4}$	956,944

4.3.5 Model training regime

All classifiers were trained once on the audiovisual action recognition problem. During training, CNN parameters were locked, preserving their object-recognition/audio event recognition representations. This allowed for the audiovisual elements of the model to be trained faster, and required less computational resources. An instance of each model was trained on AVMIT, and another instance was trained on MIT-16. The cross-entropy loss function was used as a measure of loss, and the model was trained with backpropagation and the Adam optimiser (Kingma and Ba, 2015). Each model was trained for up to 200 epochs with a batch size of 16 samples, although with an early stopping patience of 20 epochs, all models executed training before that point.All learned parameters were then fixed in place throughout testing.

4.3.6 Model testing

Action recognition

The AVMIT controlled test set was used in all testing in this chapter. The first results

collected were naturally those of the task used for training; action recognition. Here, the trained softmax layer is already trained to give an output probability distribution across each of the 16 action classes. The loss, top 1 classification accuracy (the proportion of trials in which the model gave the highest probability to the correct action class) and the top 5 classification accuracy (the proportion of trials in which the correct action class was assigned one of the top five probabilities) was used to measure performance.

SVMs to assess behaviour on new tasks

To explore the extent to which the final learnt audiovisual RNN representations were able to solve other tasks, we needed to replace the softmax classification layer with a model that could map these representations to the labels of the new task. A consideration here is that the new model must only provide a simple, linear mapping to the new task, so that results can be attributed to the audiovisual representations. For the binary tasks, a simple SVM with a linear kernel meets these requirements. For the following multi-class classification problems, we then use an SVM for each class in a one-vs-rest strategy, using the well established sklearn OneVsRestClassifier (Pedregosa et al., 2011). In this way, we were able to see to what extent the learnt representations of the RNN were able to solve a new task, without adding any non-linear fitting capacity. To create an analogy between this model training and testing procedure and human participants we could say; initial training of the deep neural network on an ecologically relevant task is parallel to life experience, training an SVM on the model's embeddings and a new task is parallel to an instruction to the participant (such as 'answer the question: do the audio and visual stimuli correspond?') and testing the SVM is parallel to the participant observing the stimuli and trying to follow the previously given instruction from the researcher.

Audiovisual correspondence task

In deep learning, the audiovisual correspondence task (Arandjelovic and Zisserman, 2017) is a binary classification task in which data from 2 modalities is provided and the classifier must provide a label of 'corresponds'/'does not correspond'. To explore the extent to which the model could solve the audiovisual correspondence task (without training on that task), we trained SVMs on the audiovisual RNN embeddings and labels of our audiovisual correspondence task.

To prepare the training set for the SVM, we shuffled the visual data amongst the AVMIT training set videos to create the incongruent portion, before concatenating it to the original training set (the congruent portion) before passing the data through the DRCNN to obtain the embeddings. Videos were labelled according to whether they were 'congruent' or 'incongruent', providing the binary classification task to be learned by the SVM.

To help ensure the classifier did not learn unhelpful strategies, we balanced the number of incongruent combinations. In this way, the videos of each class are matched with every other class an equal number of times. In order to achieve this balance, the number of videos per class must be a multiple of the number of possible incongruent classes (number of incongruent classes = $16 - 1 = 15$).

The AVMIT training dataset contains 456 videos per class, so in order to ensure each visual stream was matched with an equal number of audio streams from each class, we sampled 450 videos from each class for use in the incongruent dataset and discarded the remaining 6 videos, ensuring that each visual stream is accompanied by audio belonging to each and every class in exactly 30 instances. To ensure a balance between the congruent and incongruent portions of the AVC training set, we used the same 450 sampled videos in the congruent portion of the training set. Generating an incongruent test dataset did not require discarding videos. Altogether, this resulted in a training set containing 7,200 congruent videos and 7,200 incongruent videos and a test set of 960 congruent videos and 960 incongruent videos.

Selective-attention tasks

We extend our set of tasks to include what we refer to as a bank of selective-attention tasks. These are 16-way action recognition tasks, as before, except this time linear SVMs are utilised to map the final embeddings to a unimodal classification task. The combination of training and testing stimuli provided to the DNN + SVM classifier defines the task. The first two tasks are the *cross-modal learning* and *shared-representation learning* tasks (Table 4.2; introduced in Ngiam et al. (2011)). The third and fourth tasks are novel tasks in the area of deep learning that we introduce here; the *congruent selective-attention task* and the *incongruent selective-attention task*. One commonality amongst all of these tasks is that the SVM is trained on a the data of a unimodal

Table 4.2: Selective-Attention Tasks.

	Feature Learning	SVM Training	SVM Testing
Cross-Modal Learning	audiovisual audiovisual	audio visual	audio visual
Shared-Representation Learning	audiovisual audiovisual	audio visual	visual audio
Congruent Selective-Attention Task	audiovisual audiovisual	audio visual	cong. audiovisual cong. audiovisual
Incongruent Selective-Attention Task	audiovisual audiovisual	audio visual	incong. audiovisual incong. audiovisual

stimuli in a multimodal embedding (i.e. we present a visual-only video to a DRCNN, then use those embeddings to train the SVM). These tasks to provide a parallel to *selective-attention* tasks in psychology. All of these selective-attention tasks are intended to explore the behavioural interaction between representations according to modality. For each DRCNN, we train one SVM to 'attend audio' and train another SVM to 'attend visual' by replacing the unattended to modality by zeros. We then test each of these SVMs on the four test cases (same-modality data, alternate modality data, congruent audiovisual, incongruent audiovisual) corresponding to our bank of 4 selective attention tasks.

The cross-modality learning task allows us to first assess the unimodal performance of each DRCNN. The data at test time for each SVM is from the same modality as training time, and in the same way, the other modality is replaced by zeros. The remaining 3 selective-attention tasks all provide different modalities of data to the SVM at training and test time and are used to further explore the extent to which these unimodal representations interact on a behavioural level (as the SVM has not been trained to use this data for classification). The shared-representation task replaces the attended to modality, A, with zeros and preserves the unattended to modality, B, at test time in order to assess the extent to which modality B affects representations that the SVM had learnt to classify modality A. Our novel congruent/incongruent selective-attention tasks are intended to provide a parallel to the selective-attention recognition tasks in psychology (Yuval-Greenberg and Deouell, 2007) in which participants are instructed to report 'what they saw' or 'what they heard' whilst congruent or incongruent stimuli are

presented alongside the target. Each SVM (trained to attend to a single modality) is presented with the test stimuli with either congruent or incongruent stimuli in the other modality. The incongruent audiovisual condition was generated using the same audiovisual combinations from the previously described audiovisual correspondence task, except rather than providing boolean correspondence labels, class labels are provided according to the attended to modality. These tasks allow us to further understand the behavioural implications of the shared representations in the audiovisual embeddings.

4.4 Results

4.4.1 Action recognition

The high performance of all models on the AVMIT audiovisual action recognition test set is perhaps unsurprising given the high level of audiovisual correspondence. There was very little variation in the top 5 classification accuracy of the models on the action recognition task with all models scoring almost perfect accuracy and one model scoring 100%. All models trained on AVMIT obtained a lower loss and higher top 1 accuracy than their MIT-16 trained counterpart on the audiovisual action recognition problem (Table 4.3). Similarly, according to the loss and top 1 accuracy, the performance of all DRCNNs using the YamNet+EfficientNet-B0 audiovisual feature extractor were higher than that of the DRCNNs using VGGish+VGG-16 audiovisual feature extractors when trained on the same dataset. Those MIT-16 trained models utilising LSTMs obtained lower loss and higher top 1 classification accuracy than their MIT-16 trained counterparts utilising FRNNs or GRUs.

4.4.2 Audiovisual correspondence

After being optimised on an audiovisual action recognition task, all DRCNNs were able to perform above chance level (50%) on the audiovisual correspondence task ($p < 0.00001$ for each individual classifier; Table 4.4) despite not explicitly being trained to do so. A p-value was

Table 4.3: Action Recognition Performance on AVMIT Test Set.

Training Set	Feature extractor	RNN	Loss	Top 1 Acc. (%)	Top 5 Acc. (%)
AVMIT	YamNet + EffNetB0	FRNN	0.1841	94.58	99.90
MIT 16	YamNet + EffNetB0	FRNN	0.2973	89.79	99.90
AVMIT	YamNet + EffNetB0	GRU	0.1600	95.73	99.90
MIT 16	YamNet + EffNetB0	GRU	0.2430	92.29	99.90
AVMIT	YamNet + EffNetB0	LSTM	0.1674	95.52	99.79
MIT 16	YamNet + EffNetB0	LSTM	0.2366	92.81	100
AVMIT	VGGish + VGG-16	FRNN	0.2980	90.73	99.79
MIT 16	VGGish + VGG-16	FRNN	0.4388	84.79	99.58
AVMIT	VGGish + VGG-16	GRU	0.2917	91.04	99.79
MIT 16	VGGish + VGG-16	GRU	0.4108	85.83	99.69
AVMIT	VGGish + VGG-16	LSTM	0.2892	90.94	99.90
MIT 16	VGGish + VGG-16	LSTM	0.3527	86.98	99.90

obtained for each classifier performance using a one sample permutation test (100,000 iterations). Model rank according to action recognition performance did not directly correspond to rank according to audiovisual correspondence task performance. In particular, some models (e.g. YamNet + EfficienetNet-B0 + LSTM model) performed relatively well on the action recognition task and relatively poorly on the audiovisual correspondence task, while some models (e.g. VGGish + VGG-16 + LSTM model) performed relatively poorly on the action recognition task and relatively well on the audiovisual correspondence task when considering the performance of all models. We can also observe that some models (YamNet + EfficientNet-B0 + FRNN and VGGish + VGG-16 + GRU models) trained on AVMIT obtained a lower classification accuracy on the audiovisual correspondence task than their MIT-16 counterparts, which never occurred across the action recognition task.

4.4.3 Selective-attention tasks

The DRCNNs were able to solve the cross-modal learning task above chance accuracy (6.25%) in both the audio and visual domains (Figures 4.3 and 4.4). These performances were all found, by individual one-sample permutation tests, to be significant (100,000 iterations; $p < 0.0001$, Bonferroni corrected for 4 comparisons). The classification accuracies ranged from 69.27% to

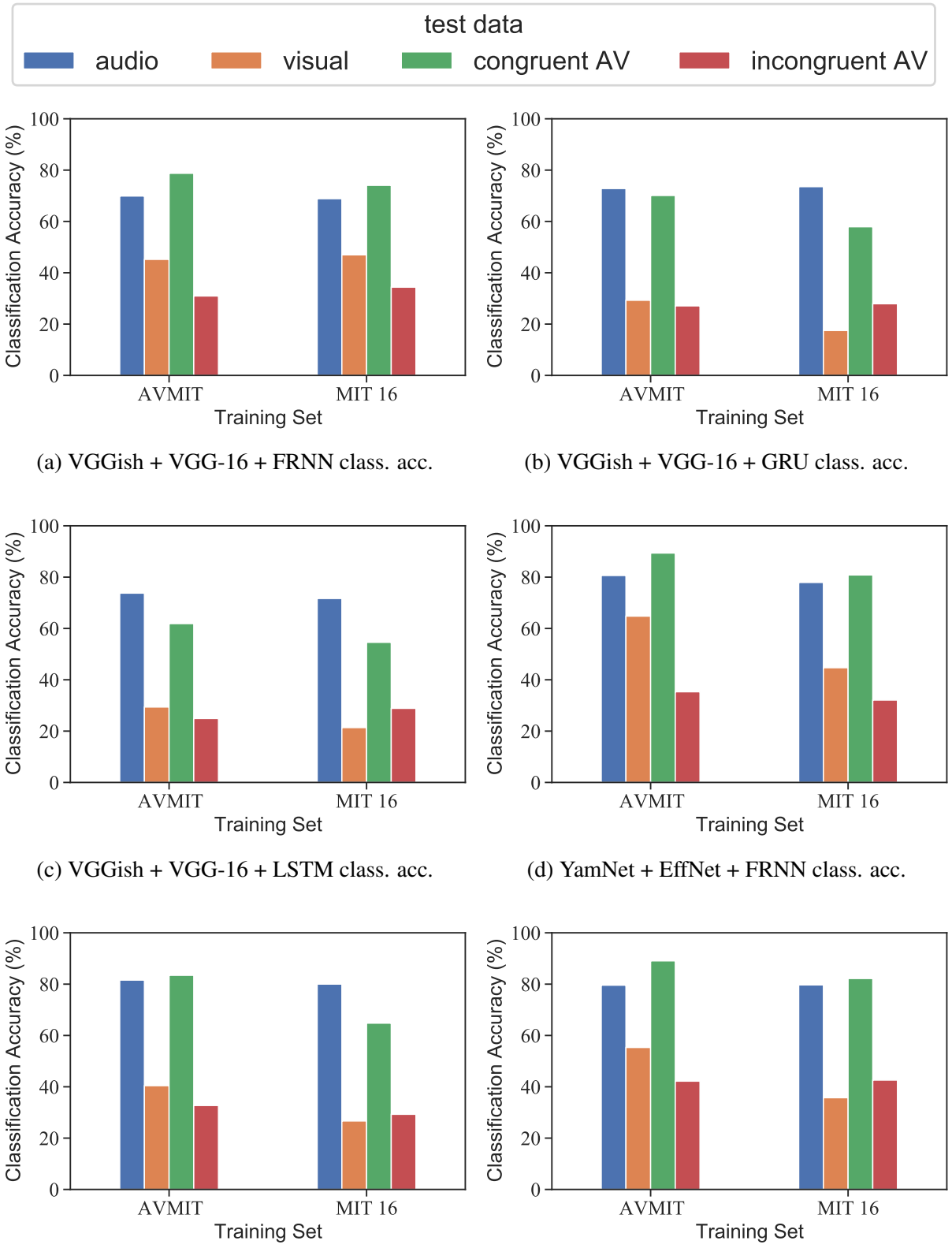
Table 4.4: Audiovisual correspondence task performance

Training Set	Feature extractor	RNN	Accuracy (%)
AVMIT	YamNet + EffNetB0	FRNN	63.39
MIT 16	YamNet + EffNetB0	FRNN	68.23
AVMIT	YamNet + EffNetB0	GRU	72.40
MIT 16	YamNet + EffNetB0	GRU	71.62
AVMIT	YamNet + EffNetB0	LSTM	58.33
MIT 16	YamNet + EffNetB0	LSTM	59.06
AVMIT	VGGish + VGG-16	FRNN	69.38
MIT 16	VGGish + VGG-16	FRNN	68.44
AVMIT	VGGish + VGG-16	GRU	70.73
MIT 16	VGGish + VGG-16	GRU	72.87
AVMIT	VGGish + VGG-16	LSTM	71.77
MIT 16	VGGish + VGG-16	LSTM	71.88

81.88% for the audio-only case and from 68.75% to 76.98% for the visual-only case

The shared-representation learning task results were significantly above chance performance (6.25%) for all DRCNNs as revealed by a series of one-sample permutation tests (100,000 iterations; $p < 0.0001$ Bonferroni corrected for 2 comparisons). The VGGish + VGG-16 + GRU model trained on MIT-16 providing the lowest score of 17.50% on the 'attend audio classify visual' or 'hearing to see' task. The performances on shared-representation learning tasks were revealed by a series of paired permutation tests to be significantly lower than the corresponding cross-modal performance for each ('attend audio' or 'attend visual') trained SVM classifier (100,000 iterations; $p < 0.0001$ Bonferroni corrected for $4 \cdot 2$ comparisons). Here, a single paired permutation test was run for each classifier individually.

The addition of congruent stimuli in the unattended to modality in the selective-attention tasks did not always result in an increase in performance (Figures 4.3 and 4.4). Although this was the case in the majority of examples, 8 of the 24 instances resulted in a decrease in classification accuracy when congruent stimuli was presented alongside the attended to stimuli. According to a paired permutation test carried out for each of the 24 SVM classifiers, however, the effect size was significant in 21 cases ($p \leq 0.05$, Bonferroni corrected for 4 comparisons). The 3 classifiers, for which the addition of congruent stimuli in the unattended modality did not produce a significant effect, were; AVMIT-trained YamNet + EfficientNet-B0 + GRU attend-



(a) VGGish + VGG-16 + FRNN class. acc.

(b) VGGish + VGG-16 + GRU class. acc.

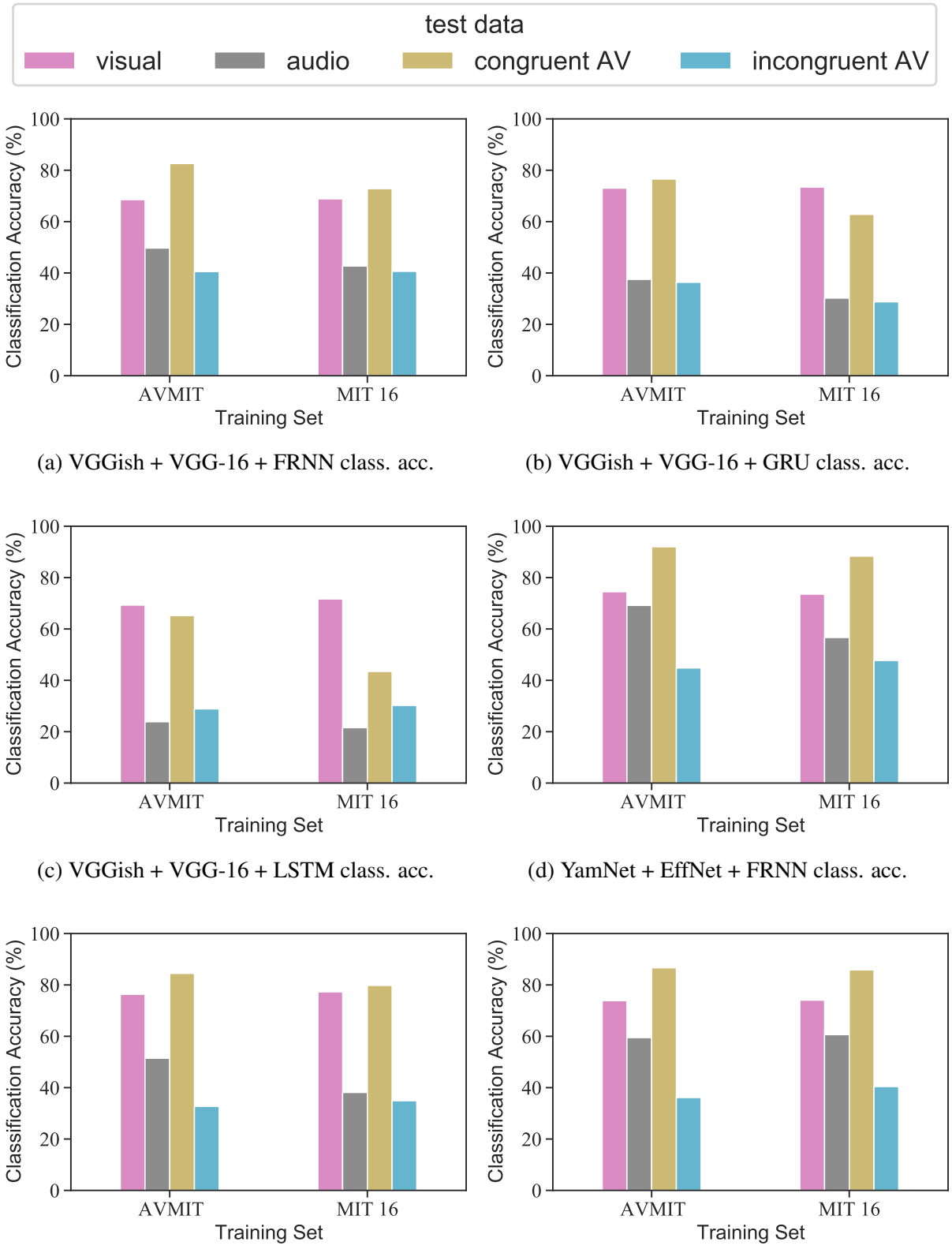
(c) VGGish + VGG-16 + LSTM class. acc.

(d) YamNet + EffNet + FRNN class. acc.

(e) YamNet + EffNet + GRU class. acc.

(f) YamNet + EffNet + LSTM class. acc.

Figure 4.3: Classification accuracy of DRCNNs during an audio-selective task when presented with audio-only, visual-only, congruent audiovisual or incongruent (where audio is correctly labelled) audiovisual stimuli.



(a) VGGish + VGG-16 + FRNN class. acc.

(b) VGGish + VGG-16 + GRU class. acc.

(c) VGGish + VGG-16 + LSTM class. acc.

(d) YamNet + EffNet + FRNN class. acc.

(e) YamNet + EffNet + GRU class. acc.

(f) YamNet + EffNet + LSTM class. acc.

Figure 4.4: Classification accuracy of DRCNNs during a visual-selective task when presented with visual-only, audio-only, congruent audiovisual or incongruent (where visual is correctly labelled) audiovisual stimuli.

audio (1.88% gain), MIT-16-trained YamNet + EfficientNet-B0 + GRU attend-visual (2.50% gain) and MIT-16-trained YamNet + EfficientNet-B0 + LSTM attend-audio (2.50% gain). These p-values are reported individually in Table 4.5.

Table 4.5: Classification accuracy gain and corresponding p values, Bonferroni corrected for multiple comparisons, when congruent stimuli is provided alongside the stimuli in the attended to modality (cross-modal vs congruent selective-attention performance). Each p value is produced from a paired permutation test with 100,000 iterations.

Training set	Feature extractor style	RNN	Attend audio		Attend visual	
			Accuracy gain (%)	p value	Accuracy gain (%)	p value
AVMIT	VGGish+VGG16	FRNN	8.85	< 0.0001	14.06	< 0.0001
MIT 16	VGGish+VGG16	FRNN	5.21	< 0.0001	3.96	0.00144
AVMIT	VGGish+VGG16	GRU	-2.71	0.05	3.54	0.00652
MIT 16	VGGish+VGG16	GRU	-15.63	< 0.0001	-10.63	< 0.0001
AVMIT	VGGish+VGG16	LSTM	-11.88	< 0.0001	-4.06	0.00068
MIT 16	VGGish+VGG16	LSTM	-17.08	< 0.0001	-28.23	< 0.0001
AVMIT	YamNet+EffNet	FRNN	8.75	< 0.0001	17.50	< 0.0001
MIT 16	YamNet+EffNet	FRNN	2.92	0.03128	14.79	< 0.0001
AVMIT	YamNet+EffNet	GRU	1.88	0.39112	8.13	< 0.0001
MIT 16	YamNet+EffNet	GRU	-15.21	< 0.0001	2.50	0.08156
AVMIT	YamNet+EffNet	LSTM	9.48	< 0.0001	12.81	< 0.0001
MIT 16	YamNet+EffNet	LSTM	2.50	0.08372	11.77	< 0.0001

The final task completed by each classifier was the incongruent selective-attention task. In this case, the presence of the incongruent stimuli in the unattended to modality had a considerably detrimental effect on performance (Figures 4.3 and 4.4). An effect that was found to be significant, for every classifier, by a series of paired permutation tests (100,000 iterations; $p < 0.0001$ Bonferroni corrected for 4 comparisons).

We further present a series of confusion matrices to visualise the confusions made by each unimodal case, and the final multisensory case by the DRCNNs (Figures 4.5, 4.6, 4.7 and 4.8). The audiovisual confusion matrices here, are those of the complete DRCNN with trained softmax layer, to better understand the non-linear combination of audio and visual data into a single classification. In this way we can observe the confusions made in the unimodal domain and those in the audiovisual domain, showing how the DRCNNs effectively utilise representations in both modalities in order to reduce the number of confusions in the audiovisual domain.

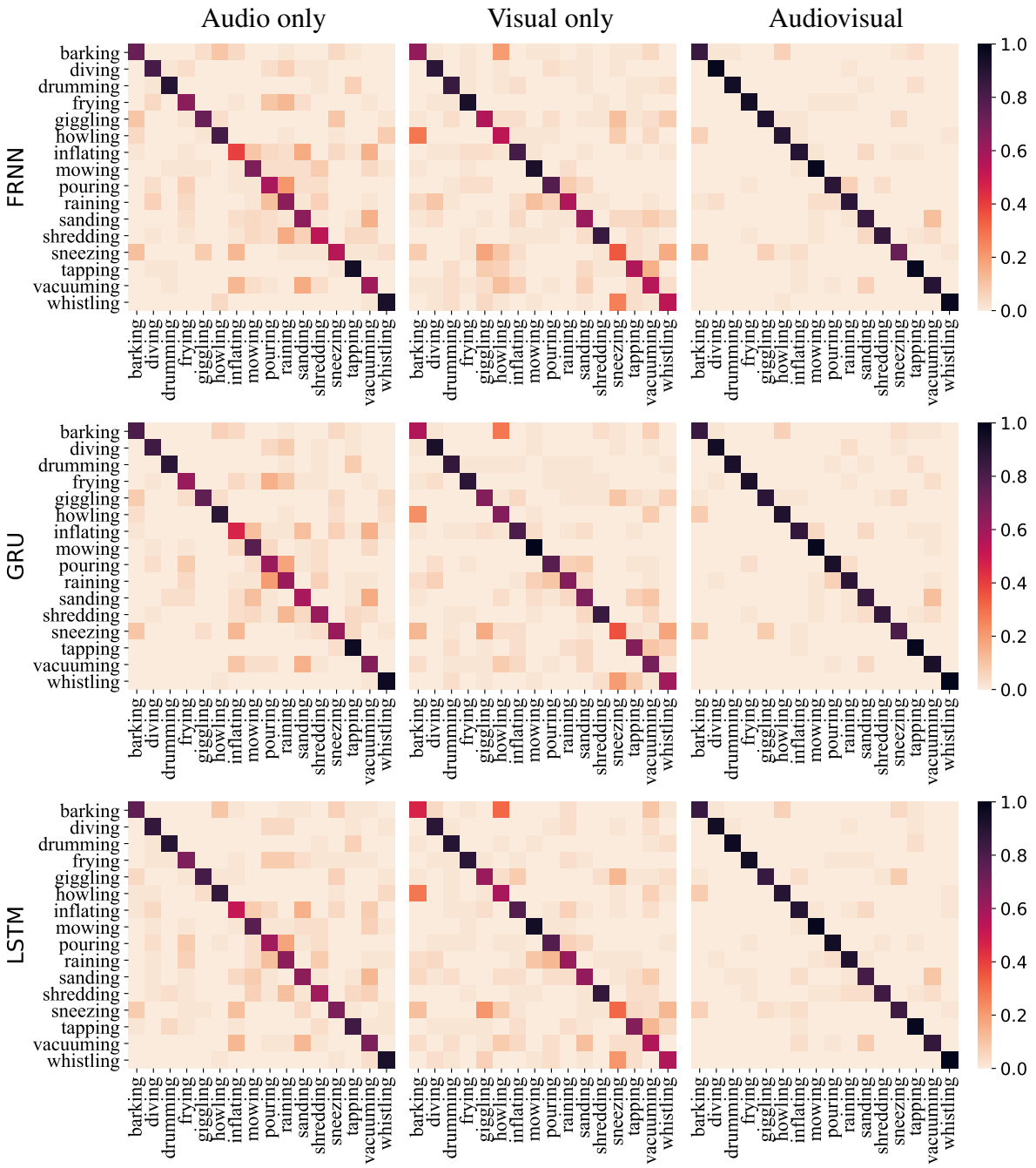


Figure 4.5: Confusion matrices for AVMIT-trained VGGish + VGG-16 DRCNNs on AVMIT test data in the cross-modal learning condition (audio or visual only) and the audiovisual condition. The audiovisual condition shows performance enhancement over the unimodal condition, demonstrating effective use of signals in both modalities.

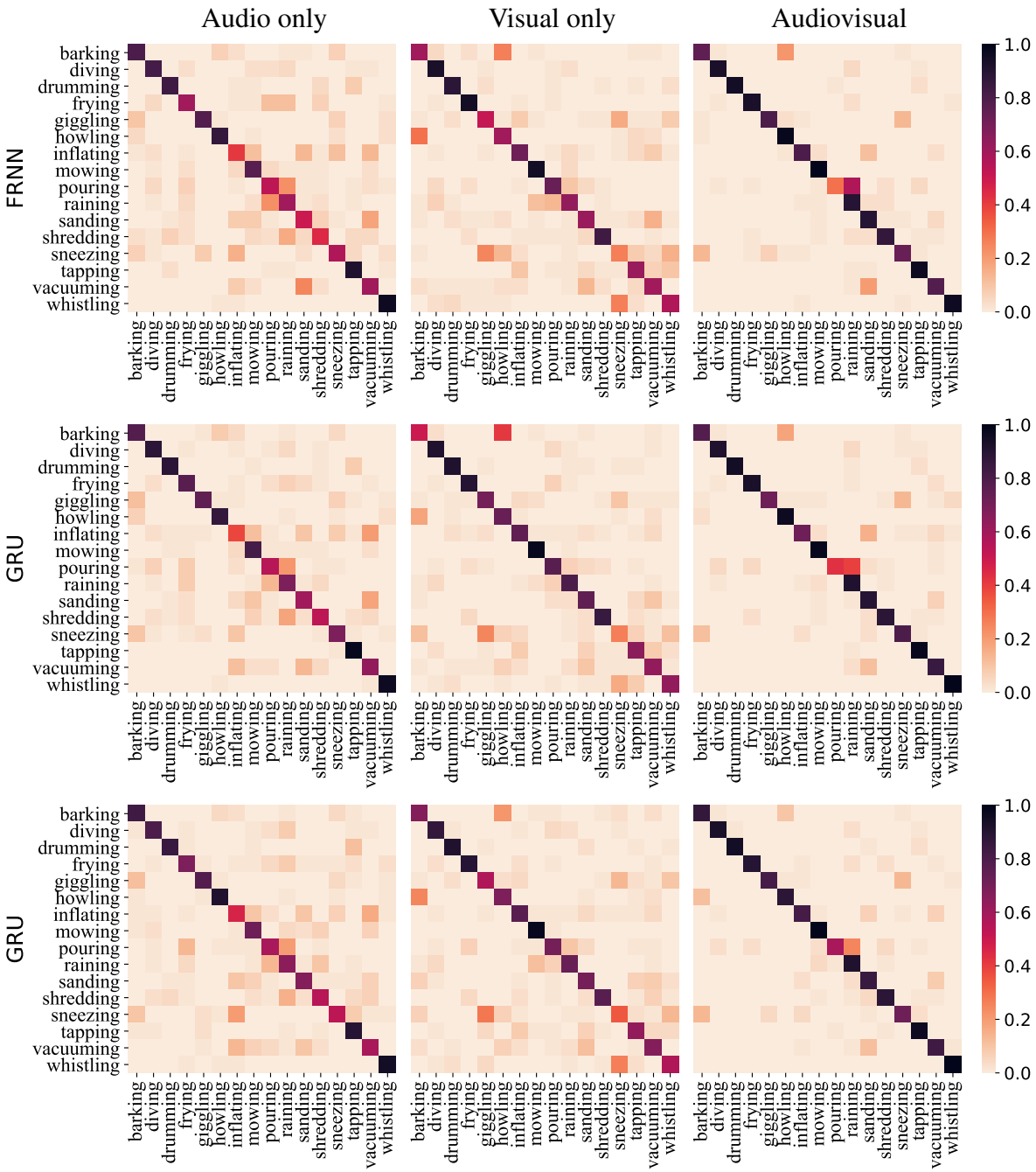


Figure 4.6: Confusion matrices for MIT-16-trained VGGish + VGG-16 DRCNNs on AVMIT test data in the cross-modal learning condition (audio or visual only) and the audiovisual condition. The audiovisual condition shows performance enhancement over the unimodal condition, demonstrating effective use of signals in both modalities.

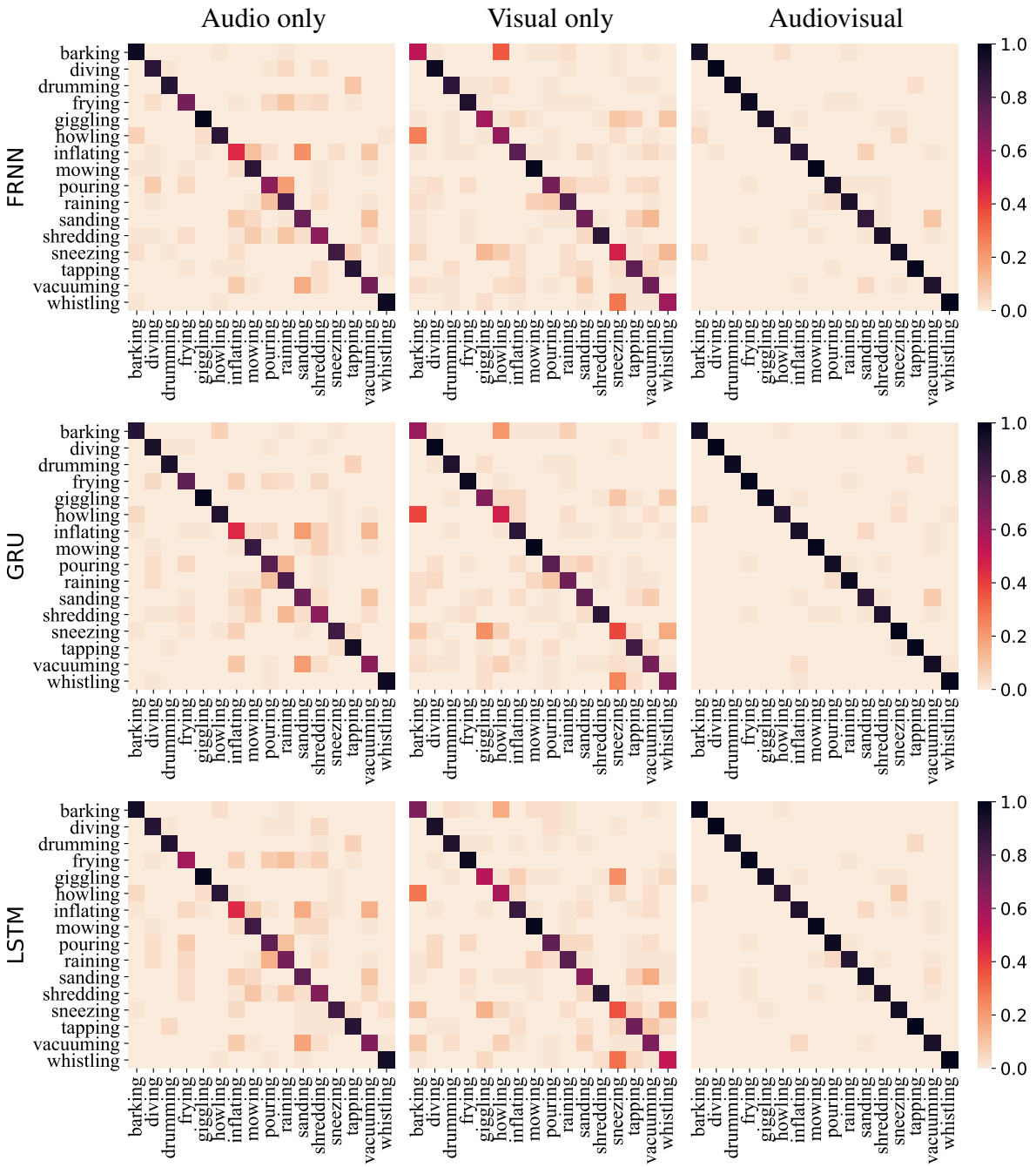


Figure 4.7: Confusion matrices for AVMIT-trained YamNet + EfficientNet-B0 DRCNNs on AVMIT test data in the cross-modal learning condition (audio or visual only) and the audiovisual condition. The audiovisual condition shows performance enhancement over the unimodal condition, demonstrating effective use of signals in both modalities.

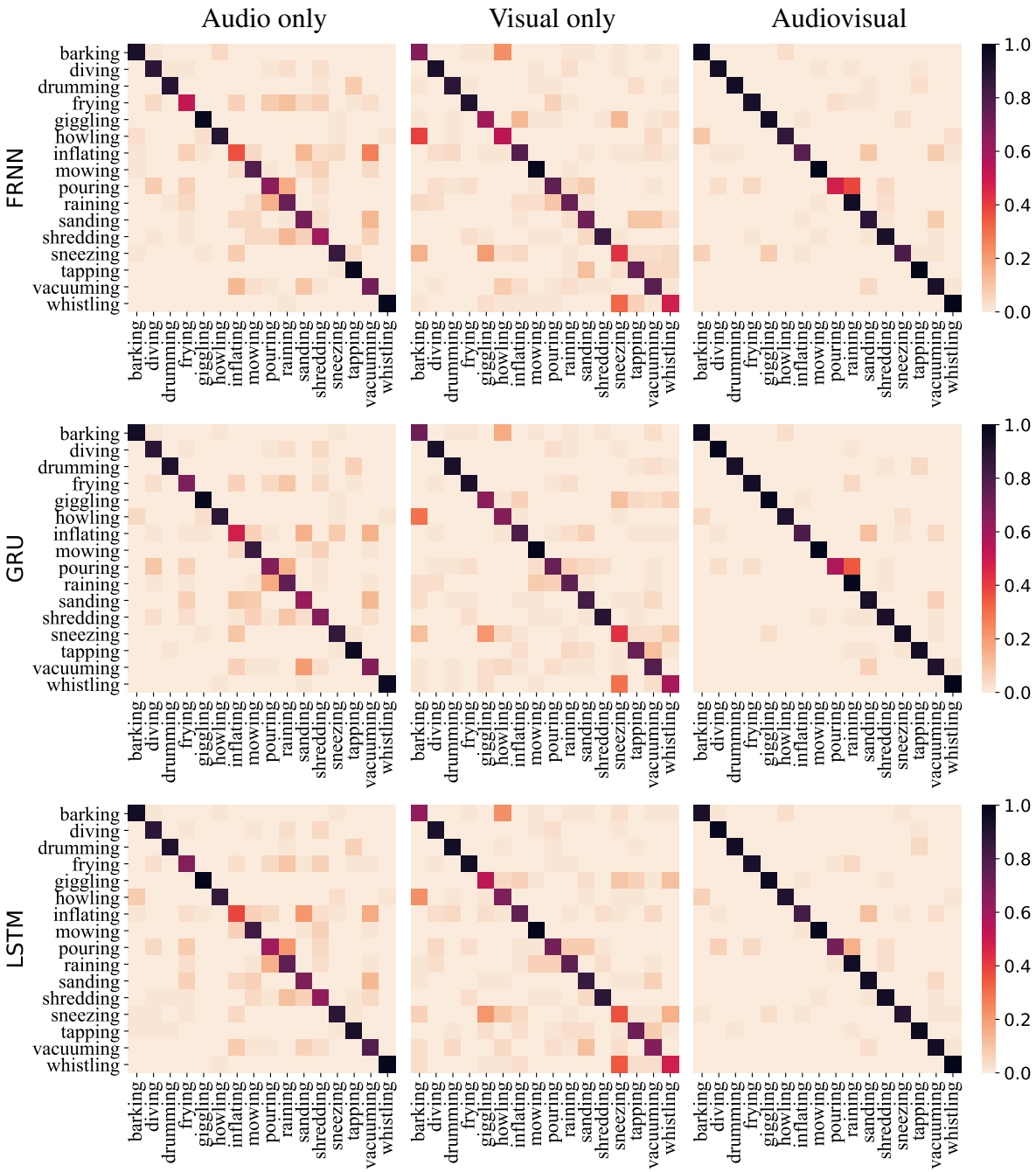


Figure 4.8: Confusion matrices for MIT-16-trained YamNet + EfficientNet-B0 DRCNNs on AVMIT test data in the cross-modal learning condition (audio or visual only) and the audiovisual condition. The audiovisual condition shows performance enhancement over the unimodal condition, demonstrating effective use of signals in both modalities.

To better visualise the differences between the unimodal classifier error patterns and that of the corresponding audiovisual models, we present a series of confusion difference matrices (CDMs; Figures 4.9 and 4.10; Geirhos, Janssen, et al. (2017) and Dyck and Gruber (2020)). These highlight the differences between audiovisual confusions and unimodal confusions, indicating a strategy by the DRCNN to leverage both modalities to reduce the number of confusions. It can be observed that the AVMIT-trained audiovisual classifiers all predicted correct (diagonal) labels more than AVMIT-trained unimodal classifiers. Indeed, there are no off-diagonal cells that were predicted by the AVMIT-trained audiovisual classifiers more than the corresponding unimodal classifiers. There are several misclassifications made more commonly by the unimodal classifiers, however. For instance, in the visual-only domain, classifiers often confused 'howling' with 'barking', a confusion that is seldom made in the audio domain, these confusions are reduced as the model utilises audio data in the audiovisual domain. MIT-16-trained classifiers are less exact, but still show a general trend of audiovisual classifiers making more correct classifications than unimodal classifiers. One notable exceptions in the MIT-16 case is the common confusion of the 'pouring' and 'raining' classes. As the word 'pour' can also mean 'rain' in English (*Cambridge Dictionary* 2022), we consider that this confusion likely reflects a labelling overlap in MIT-16 and the wider MIT dataset (Monfort et al., 2019) that is less prevalent in the AVMIT dataset.

4.5 Discussion

This study sought to investigate whether the ability to solve the audiovisual correspondence problem can arise implicitly from optimisation on an audiovisual action recognition problem, extending on the idea that humans are highly optimised to solve ecologically relevant tasks. More specifically, we developed a series of Recurrent Convolutional Neural Networks, trained an instance of each on one of two datasets (with different levels of correspondence), and tested them on our own audiovisual correspondence problem. We further explored the behavioural consequences of audio and visual interactions in the audiovisual embeddings of the model with a set of selective-attention tests, two taken from the literature (Ngiam et al., 2011) and two novel

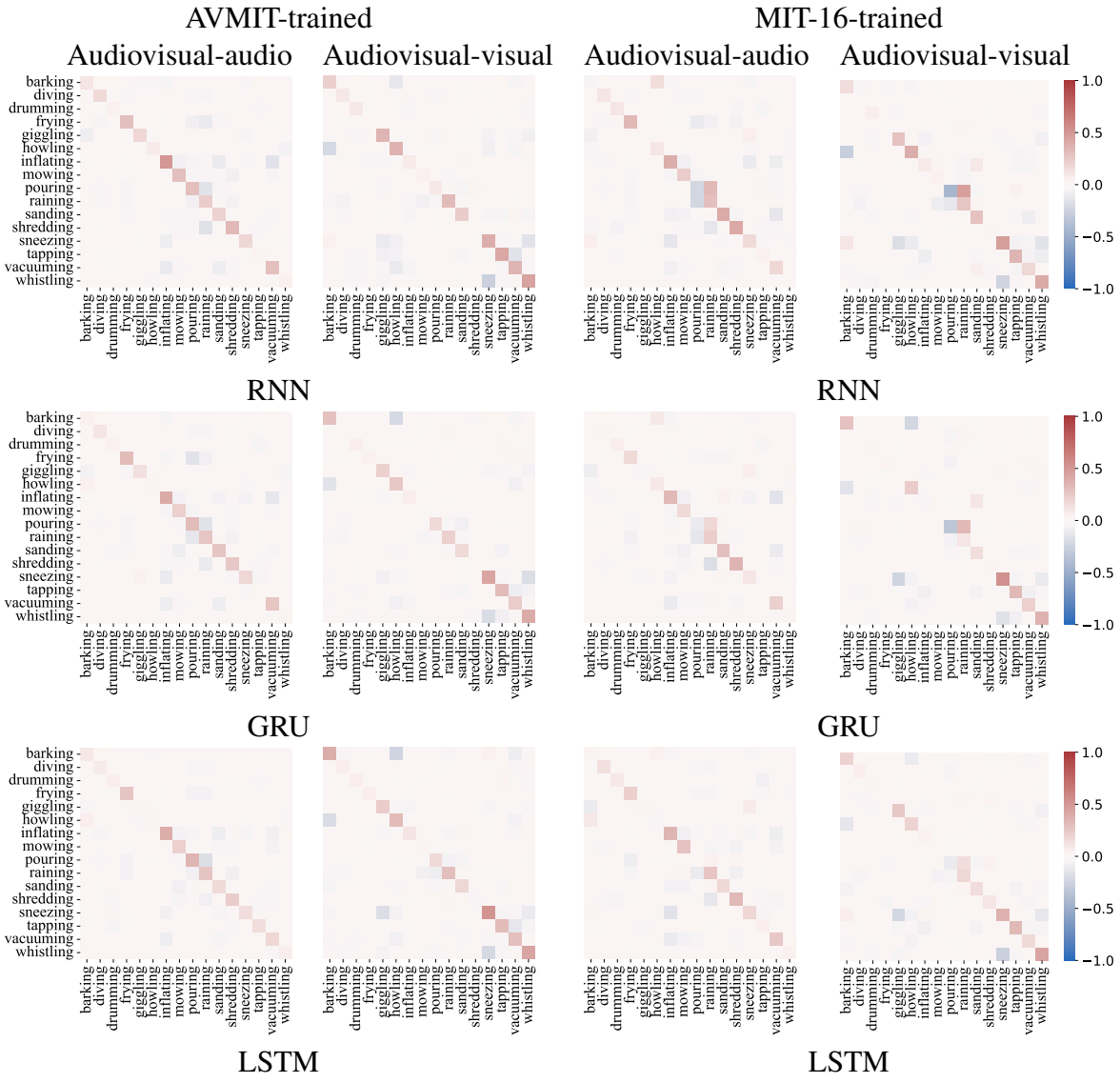


Figure 4.9: Confusion Difference Matrices (CDMs) for VGGish + VGG-16 models. Each matrix shows the difference between the audiovisual confusion matrix and that of a single modality (left:audio, right:visual) for each classifier. A cell value of 1.0 (red) indicates that the audiovisual classifier made this prediction for all videos in the given class (row) and the unimodal classifier did not make this prediction in any of the trials for that label. Similarly, a negative cell value (blue) indicates that the unimodal classifier made that prediction more than the audiovisual classifier.

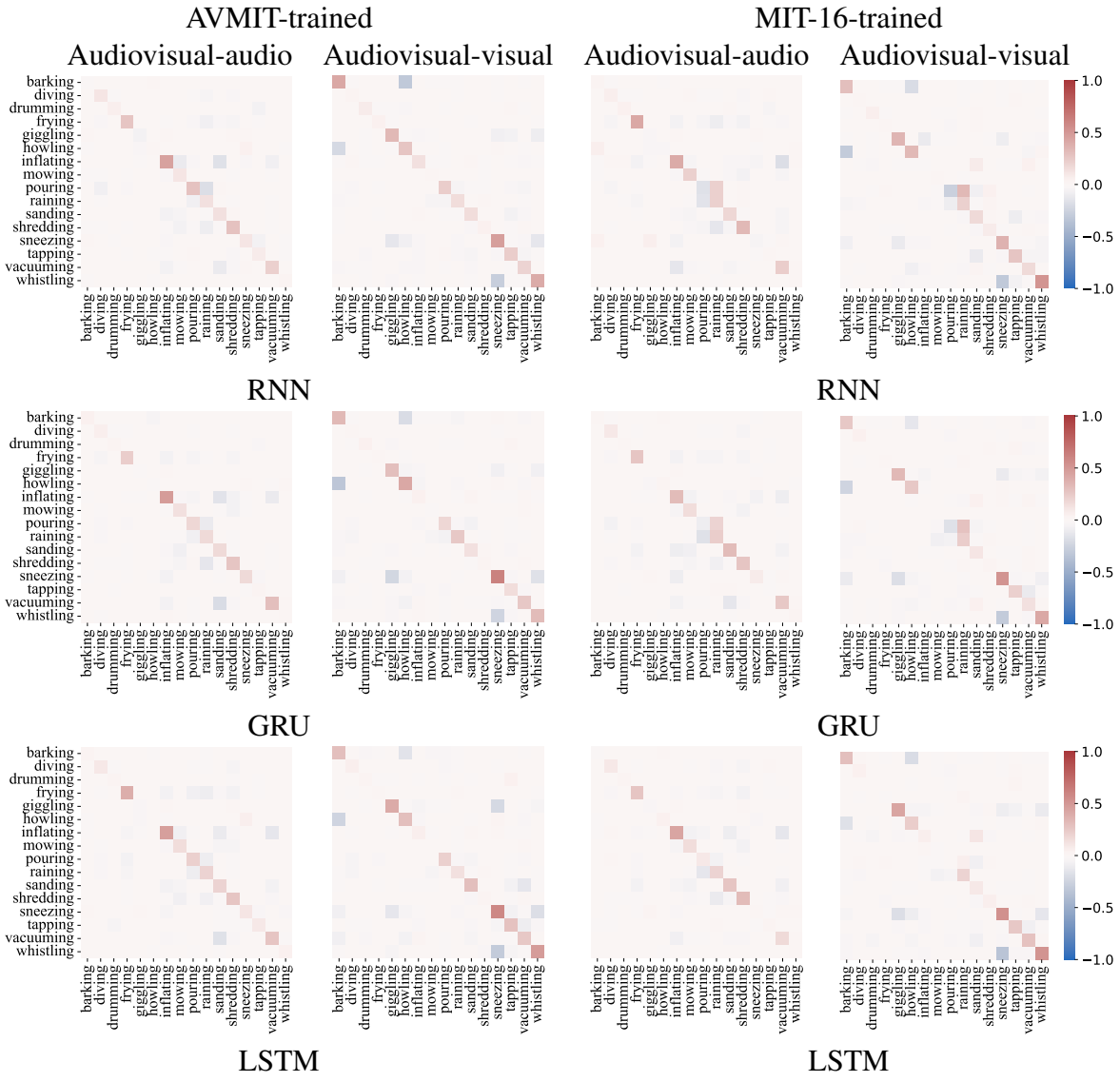


Figure 4.10: Confusion Difference Matrices (CDMs) for YamNet + EfficientNet-B0 models. Each matrix shows the difference between the audiovisual confusion matrix and that of a single modality (left:audio, right:visual) for each classifier. A cell value of 1.0 (red) indicates that the audiovisual classifier made this prediction for all videos in the given class (row) and the unimodal classifier did not, in even a single instance, make this prediction. Similarly, a negative cell value (blue) indicates that the unimodal classifier made that prediction more than the audiovisual classifier.

tests we introduce here. Each selective-attention task involved training a support vector machine on the audiovisual RNN embeddings of the DRCNN where only unimodal data was provided (thus the SVM models only unimodal data in the multimodal embedding) and then providing data in one or both stream, and observing the effect on classification accuracy.

In line with our principal hypothesis, all AVMIT trained DRCNN models were able to solve the audiovisual correspondence task at performance levels significantly above chance (50%), as revealed by a series of one-sample permutation tests, despite not explicitly being trained on this problem. Our hypothesis that the higher levels of audiovisual correspondence in AVMIT over the MIT-16 training dataset was necessary to implicitly learn to solve the AVC task, however, was not true. MIT-16 trained DRCNN models too were able to solve the AVC task, with each performance confirmed to be significantly above chance by a one-sample permutation test. This was because MIT-16 either contained an adequate level of audiovisual correspondence, or the correspondence was not at all necessary for the emergence of the ability to solve the AVC task in our DRCNNs. If the level of audiovisual correspondence in the training data was unimportant here, that would suggest that the unimodal representations from the pretrained CNNs are well preserved despite bottlenecking in the multimodal squeeze unit and the RNN (the so-called 'separate activation' model described earlier). In this case the final softmax activation function (responsible for providing the final action classification) could provide simple simple weighting to each unimodal representation. Then the ability to solve the audiovisual correspondence task would require the SVM replacing the softmax layer to learn a simple AND function (e.g. 'audio diving' AND 'visual diving' gives 'correspond'; 'audio diving' AND 'visual laughing' gives 'does not correspond'). A separate activation model such as this would not make use of data across modalities (such as semantic congruence) and would make a poor model of audiovisual perception in humans. Our bank of selective-attention tasks revealed this not to be the case, however.

The selective-attention tasks first used a unimodal/'cross-modal learning' problem to provide an uninterrupted baseline of unimodal performance prior to the introduction of data in the unattended to modality. One-sample permutation tests performed on the predictions of each

model showed all cross-modal learning performances to be significantly above chance (6.25%). Shared-representation performances were also significantly above chance across both auditory and visual modalities, as hypothesised, revealing that the linear mapping to classify unimodal data from the audiovisual embedding captures signals from the other modality. The performance of each DRCNN on both the 'attend audio' and 'attend visual' cross-modal learning tasks were revealed to be significantly higher than the corresponding shared-representation tasks, which was unsurprising given that the SVMs had fit unimodal signals in the audiovisual embeddings.

The hypothesis that congruent stimuli provided in the 'unattended to' modality would increase unimodal classification accuracy was not true in every case, with unexpected individuality amongst DRCNNs on the congruent selective-attention task. There were 13 significant performance increases and 8 significant performance decreases when congruent stimuli was presented alongside the stimuli in the attended to modality. Where a significant effect was detected in both the 'attend audio' and 'attend visual' congruent tasks, this tended to be in the same direction, with congruent stimuli increasing performance in both modalities in 5/12 DRCNNs, and decreasing performance in both modalities in 3/12 DRCNNs. 3/12 DRCNNs did not provide a significant effect in both modalities and only 1/12 DRCNNs provided a significant decrease in one task and a significant increase in the other. We do not present here an explanation for why not all trained DRCNNs improved their classification accuracy, as hypothesised, when congruent stimuli was provided alongside the attended to modality. However, we do acknowledge that in 21 of the 24 congruent tests, the effect was significant, adding further evidence to the non-linear intertwining of audio and visual signals in the audiovisual representation. In the incongruent selective-attention task all DRCNNs scored significantly reduced classification accuracies when incongruent stimuli was presented in the unattended to modality, in line with our hypotheses.

In Yuval-Greenberg and Deouell (2007), the researchers report that participants were more accurate in the congruent trials than in the incongruent trials, with no unimodal baseline obtained. This matches the results in our study in every instance. Although as we previously mentioned, there were variations in behaviour across DRCNN classifiers, the congruent accuracy was always considerably higher than the incongruent accuracy. Yuval-Greenberg and Deouell (2007) does

not obtain a unimodal baseline, however, such that they could measure the effect of adding the unattended signal. In our work we obtain this additional information to help understand how these unattended signals can assist or harm classification performance. To extend our research, we could run a human experiment alongside our study for comparison against our DRCNN classifiers. The study could include the congruent/incongruent selective-attention tasks, as in Yuval-Greenberg and Deouell (2007) with the additional unimodal classification task to obtain a baseline performance.

The linear SVMs, trained to classify unimodal signals captured in the audiovisual RNN embeddings of the DRCNNs, were affected by the information in the other modality. This is reflected in the significant effects between the classification accuracies on the cross-modal learning tasks and the corresponding classification accuracies on the shared-representation learning, congruent and incongruent selective-attention tasks. Specifically, the addition of signals from the unattended modality to the audiovisual embedding affected those embedding features that captured signals in the attended to modality, that were used to optimise a maximally separating hyper-plane. In this way, we can conclude that the DRCNN does not maintain clearly separated activations of audio and visual signals in its RNN embeddings.

To better understand the normal operation of the DRCNN models as they were initially optimised with a softmax layer prior to our transfer learning investigations with SVMs, we presented audiovisual confusion matrices on the action recognition problem alongside that of the unimodal task (cross-modal learning). Here, common confusions could be seen amongst the models. Some of these confusions exist largely in one modality but not in the other, and it could be observed that the audiovisual domain largely combines these representations in a way such that it leverages both modalities and reduces confusions. One clear example was the confusion between ‘barking’ and ‘howling’ in the visual domain that was much less prominent in the auditory domain and the product audiovisual domain, the DRCNN would perform much worse if it weighed the visual representation more than the auditory representation in this scenario. Confusion difference matrices presented in this work highlighted differences between the audiovisual and unimodal predictions, with audiovisual predictions occurring more frequently

along the diagonal (correct) and unimodal predictions occurring more frequently off-diagonal (incorrect). That the audiovisual DRCNNs do not perform worse than their respective unimodal performances (if for example, the DRCNN adopted both sets of unimodal confusions) shows that the model does not weight each modality uniformly in every scenario. Thus the DRCNNs effectively weigh each modality according to some approximate measure of reliability. Although this is not necessarily an example-level reliability, and could in fact be a class-level reliability (e.g. weigh the audio domain heavier than the visual domain with 'barking'/'howling' classifications).

Where this work considered the ability to solve the audiovisual correspondence task as an emergent property of optimisation on an audiovisual recognition task, the correspondences were purely semantic. In humans, this problem is solved using spatiotemporal cues (Munhall et al., 1996; Slutsky and Recanzone, 2001; Lewald and Guski, 2003; Wallace et al., 2004) as well as higher order cues (Laurienti et al., 2004; Parise and Spence, 2009; Calvert et al., 2000; Doehrmann and Naumer, 2008; Noppeney, Ostwald, et al., 2010; Krugliak and Noppeney, 2016). Indeed, there is a considerable body of work exploring the use of spatiotemporal information in the brains of rodents and cats (Stein and Meredith, 1993) and the resultant neural and behavioural consequences. In particular, these lower level cues are processed in the superior colliculus (SC) of the brain, where there exists a number of overlapping sensory and motor maps that provide an architectural basis of multisensory integration in the deep layers of SC. Further work could obtain video training sets whose audio is stereophonic, and explore the emergent multisensory behaviour in select DNNs and whether they learn to use spatial cues for example. This could also be investigated with embodied agents in simulated environments.

Further studies could equally explore DRCNN output over time. The output of the RNN component can be observed over time-steps, and any accumulation of evidence towards particular classes can be observed in the sequence of output distributions. Any accumulation of evidence over time-steps would provide a parallel to human intelligence, the accumulation of audiovisual evidence would be particularly interesting here (Noppeney, Ostwald, et al., 2010). Parallels to reaction time could be explored using thresholds in the entropy of the output distribution as in (Spoerer, Kietzmann, et al., 2020). In particular, a researcher could ask 'do these trained

DRCNNs provide the correct prediction sooner with multimodal stimuli than unimodal stimuli?'.

Although we do not investigate the activations themselves and instead focus on behavioural phenomena, we postulate that the audiovisual interactions observed in this work could occur on both a scale of neuron subpopulations and the single neuron. Further work could use maximal activation analyses, ablation analyses, activation perturbations and feature permutation analyses amongst other methods to explore the tuning properties of neurons in the embeddings of these DRCNNs. These explorations could reveal particular neurons/regions of the audiovisual embedding that are entirely unimodal/ multisensory, and measure the proportion of each. Representational analyses could also attempt to detect congruent and opposite neurons as in the work by Rideaux et al. (2021).

Investigations could also be carried out to explore audiovisual integration behaviours in other action recognition models. Indeed, a number of audiovisual action recognition models have been introduced in the literature (review can be found at Sun, Ke, et al., 2022). For instance, Xiao et al. (2020) introduced SlowFast Networks for the problem of audiovisual video recognition. SlowFast Networks have two pathways for visual information (one at a low frame rate, one at a high frame rate) and another, faster, pathway for audio data. Lateral connections are implemented to allow audio streams to inform visual streams at multiple processing depths, although the audio pathway is sometimes dropped out during training to enable joint learning across modalities. This approach gave state-of-the-art performance across a number of benchmarks including the Kinetics action recognition dataset (Carreira and Zisserman, 2017). Another work by Kazakos et al. (2019) introduced the Temporal Binding Network (TBN) that uses RGB, optical flow and audio modalities and integrates them within some adjustable temporal binding window. As in our studies, confusion difference matrices in this work display the types of class confusions that are increased or reduced with the addition of audio, and indeed the overall performance increases when more than one modality is used. This shows the utility of using multiple modalities in the model, and indeed this could be followed by a number of selective-attention experiments to better understand integration at the behavioural level. Multi-stream CNN models could also be investigated for multisensory integration behaviours and help understand the architectures

introduced in our work. A three-stream CNN, utilising audio, optical flow and RGB features to recognise human actions was introduced by Wang, Yang, et al. (2016). Although the work studies two levels of fusion, these involve far less computation after the fusion point than the models investigated here (thus we would expect less integration). Nonetheless, research into the multisensory integration behaviour of basic CNN models would add to our understanding of inductive biases/architectures and the multisensory behaviours produced. Another multimodal CNN for human action recognition was introduced by Owens and Efros (2018). This was a two-stream CNN, self-supervised on the temporal alignment of audio and visual signals. The researchers fine-tuned these representations to solve sound source localisation, audiovisual action recognition and on/off-screen audio source separation problems. Whilst the work we present in this chapter demonstrated that the ability to solve the audiovisual correspondence task can emerge from optimisation on an audiovisual recognition task, the researchers in Owens and Efros (2018) show that the inverse is also true. These studies together suggest that the solution spaces of recognition problems and correspondence problems are closely related.

CHAPTER 5

DUAL-STREAM RECURRENT CONVOLUTIONAL NEURAL NETWORKS AS MODELS OF HUMAN AUDIOVISUAL PERCEPTION AS THE SIGNAL GETS WEAKER

Contributions: All work including programming, modelling, data collection, analysis and writing were carried out by Michael Joannou with Pia Rotshtein, Uta Noppeney and Bernd Bohnet performing supervisory roles.

5.1 Abstract

Convolutional neural networks (CNNs) have been investigated as models of human sensory perception due to their human-level performance on specific naturalistic classification benchmarks such as ImageNet. Research has thus far revealed a vulnerability of ImageNet-trained CNNs to a number of visual distortions, causing a deviation from human performance and error patterns. In this work, we explore the ability of recurrent neural networks (RNNs) to accumulate evidence across CNN embedding sequences (corresponding to video image sequences) to overcome visual distortions in the video recognition domain. We further investigate the ability of RNNs, with audio CNN embeddings alongside visual, to dynamically leverage audio information to dampen deteriorating performances when these visual distortions are introduced. We obtain a series of dual-stream RCNN models (Chapter 4), train them on the AVMIT-VEGAS dataset and test them in the visual and audiovisual domain on the following distortions; Gaussian noise, Gaussian blur, salt and pepper noise and contrast reduction. We then carry out a series of online experiments to obtain human performance and error patterns on the same action recognition task with Gaussian noise and Gaussian blur distortions to compare against our classifiers. We find that in the visual domain, the dual-stream RCNN classifiers become increasingly biased and suffer deteriorating performance that reaches/approaches random chance on the studied distortion levels. Although human performance decreases across these same distortion levels, classification accuracy is significantly higher than that of the studied classifiers. The addition of audio alongside visual data led to a significant increase the overall classification accuracies of dual-stream RCNN classifiers in all cases (distortion types/levels) and humans in all cases other than the lowest level of Gaussian noise where the increase was too small to detect a significant effect. We further observe significant decreases in the *rate of performance degradation* for humans and classifiers when audio is provided alongside distorted visual data. Showing that both dual-stream RCNN classifiers and humans alike are able to rely more heavily on clean auditory data when visual data becomes unreliable in order to preserve performance. Thereby extending the human vs artificial neural network literature to the audiovisual domain, and the DNN robustness literature to include audiovisual action recognition and dual-stream RCNNs.

5.2 Introduction

Humans make use of a number of sensory modalities in order to build more reliable percepts and to provide recourse when sensory data becomes unreliable. In some instances, this utility is obvious when there is seemingly no perceived signal in a given modality. For example, imagine you are sitting in the passenger seat of a vehicle and you see a passenger of another vehicle gesturing through the windows. You are unable to hear the passenger, but once you attend to them, you see them mouth the words 'your headlight is out'. In this scenario, you were able to use visual information (by lip reading) when the audio information was completely unavailable, despite the visual modality not being the most 'appropriate' modality for speech recognition (*modality appropriateness*; Welch and Warren, 1980). But our many sensory organs do not only exist to provide reserve sensory data for when unisensory perception in the most appropriate modality is not possible.

An important benefit of multisensory perception is the ability to create more reliable and effective percepts. This idea has thus far permeated the deep learning literature largely in the audiovisual speech recognition domain to obtain improved recognition rates over audio-only systems (Zhou, Yang, et al., 2019; Yu et al., 2020; Aldeneh et al., 2021) particularly when audio data is unreliable. One particularly famous speech recognition problem is the *cocktail party* scenario (Cherry, 1953); given an environment with more than one speaker, how does one selectively attend to a single speaker? This problem has been studied for decades in the area of psychology (Cherry, 1953; Brungart and Simpson, 2007; Bronkhorst, 2015; Li, Wang, et al., 2018) and more recently addressed in the area of deep learning (Gabbay et al., 2018; Ephrat et al., 2018). Indeed, overcoming noisy environments to recognise speech is a well established problem in the deep learning literature with dedicated benchmarks focussing on other forms of auditory noise as well (Reddy, Gopal, et al., 2020; Reddy, Dubey, et al., 2021). DNN solutions are often trained on distorted training data (Fang et al., 2021) or built with some mechanism to explicitly weigh sensory modalities or remove unisensory noise (Yu et al., 2020; Zhang, Li, et al., 2021) to solve these problems.

Unreliable stimuli pose a problem to perceptual systems beyond speech recognition tasks

though. For instance in the area of video recognition, where many researchers build systems to recognise visual image streams, noise can present itself at multiple stages prior to reception by the algorithm. Heavy rainfall or low light conditions could obscure videos, motion blur or dropped frames could occur at video capture as could blur due to a smudged lens, lossy compression techniques could create noise during transmission or impulse noise could occur during storage. Humans (or algorithms) may even edit videos deliberately to add noise, such as blurring faces/vehicle license plates or adding proprietary watermarks. Accurate models of human perception should show some level of robustness to this noise, and from an engineering perspective, those deploying computer vision algorithms must ensure a level of resilience too. Just as those in the audiovisual speech recognition domain have sought to use visual processing to ensure resilience to auditory noise, we consider the use of audio processing to ensure resilience to visual noise in the action recognition domain.

The effect of a number of visual distortions on ImageNet-trained deep convolutional neural networks has been explored thus far in the literature; Gaussian noise (Dodge and Karam, 2016; Dodge and Karam, 2017; Dodge and Karam, 2019), Gaussian blur (Dodge and Karam, 2016; Dodge and Karam, 2017; Dodge and Karam, 2019), contrast reduction (Dodge and Karam, 2016; Wichmann et al., 2017; Geirhos, Temme, et al., 2018), greyscaling (Geirhos, Janssen, et al., 2017; Geirhos, Temme, et al., 2018), salt and pepper noise (Geirhos, Temme, et al., 2018) amongst others. With the exception of greyscaling, contrast-reduction and colour distortions, humans have obtained higher classification accuracies than DNNs such as AlexNet (Krizhevsky et al., 2012), GoogLeNet (Szegedy et al., 2015), VGG-16 (Simonyan and Zisserman, 2015) and ResNet-50 (He et al., 2016) on the aforementioned image distortions. This is despite often outperforming human participants on clean ImageNet stimuli (Geirhos, Temme, et al., 2018). We ask whether image sequences (videos) and recurrent connections could provide recourse in these unisensory classification tasks when noise is presented.

We obtain a series of dual-stream recurrent convolutional neural network (DRCNN) models (Chapter 5) and train them on the AVMIT-VEGAS audiovisual action recognition problem (Chapter 3). These classifiers have been shown to obtain high classification accuracies on

the AVMIT test set and 'cross-modal learning' tasks (unisensory classification). Given their high unisensory performances, these classifiers provide an ideal opportunity to assess whether the processing of multiple frames and recurrent connections could provide some resilience to frame-level noise. We test these classifiers under four different visual distortions; Gaussian noise, Gaussian blur, salt and pepper noise and contrast reduction. All at varying noise levels. Further, as these classifiers are optimised on audiovisual action recognition, we additionally test them on these distortions alongside clean audio, and assess their ability to dynamically weight signals from each modality, without being explicitly trained to do so. To consider if the classifiers show human like resilience to visual distortion in the visual and audiovisual domains, we carry out a number of online classification experiments on Pavlovia (Peirce et al., 2020), with participants recruited from Prolific (*Prolific* 2014) on Gaussian noise and Gaussian blur. These distortion types, in particular, have been shown to be particularly destructive to ImageNet trained DNNs on image recognition tasks (Dodge and Karam, 2016; Dodge and Karam, 2017; Dodge and Karam, 2019).

To our knowledge, DNN performance on audiovisual action recognition tasks under visual noise has not been explored in the literature. Nor has audiovisual video recognition been used as a medium to compare humans and DNNs. In the following sections, we outline our methods to prepare test stimuli, run DNN/human experiments and analyse results, we then present our results on each subexperiment before discussing this study as a whole.

5.3 Methods

Here we discuss the experimental paradigm, models, creation of our dataset, stimuli reliability manipulations and information about participant observers.

5.3.1 Software packages

Human experiments were developed using Psychopy builder 2021.1.4 (Peirce et al., 2019) with Python 3.6.6 (Van Rossum and Drake, 2009) and the following python packages; NumPy 1.18.1

(Harris et al., 2020), SciPy 1.4.1 (Virtanen et al., 2020), Matplotlib 3.3.0 (Hunter, 2007). These human experiments were hosted online using Pavlovia (Peirce et al., 2020). Participants were recruited using the online participant recruitment tool Prolific (*Prolific* 2014).

The distorted videos were created using Python 3.7.9 (Van Rossum and Drake, 2009), For audio preprocessing we used the python packages Pydub 0.24.1 (Robert, Webbie, et al., 2018), SciPy.signal (SciPy version 1.5.4) (Virtanen et al., 2020) and to resample we use resampy (McFee, 2016). For visual preprocessing we used OpenCV 4.4.0.44 (Bradski, 2000). For compiling audio and video together into mp4 files we used MoviePy 1.0.3 (*MoviePy* 2017).

All models were developed and tested with Python 3.7.9 (Van Rossum and Drake, 2009), NumPy 1.18.5 (Harris et al., 2020) and TensorFlow 2.3.1 (Abadi et al., 2015). For audio preprocessing we used the python packages Pydub 0.24.1 (Robert, Webbie, et al., 2018), SciPy.signal (SciPy version 1.5.4) (Virtanen et al., 2020) and to resample we use resampy (McFee, 2016). For visual preprocessing we used OpenCV 4.4.0.44 (Bradski, 2000). For testing with SVMs, we use sklearn (Pedregosa et al., 2011).

5.3.2 Experimental Paradigm and Procedure

Principally, we developed a number of perceptual models and compared them to each other and human participants on two 23-way action video classification tasks. In the first task, visual data was distorted by Gaussian noise, in the second task Gaussian blur was used to distort the visual data. For both tasks, human participants watched a series of 3 second videos. After each video, a menu screen was presented with 23 buttons; each corresponding to a label description. Participants were tasked to click the button, using the mouse or touchpad on their computer, corresponding to the label that best describes the video they just watched. The maximum response time was set to be 5 minutes to detect inactivity and terminate the experiment.

Participants first had to complete a practice routine (Figure 5.1a). The practice routine was used to familiarise participants with the task but also used as a screening tool to ensure a level of accuracy on undistorted videos. The set of practice videos contained 2 undistorted videos for each class, uniformly sampled from a set of 10 designated practice videos for each class. One

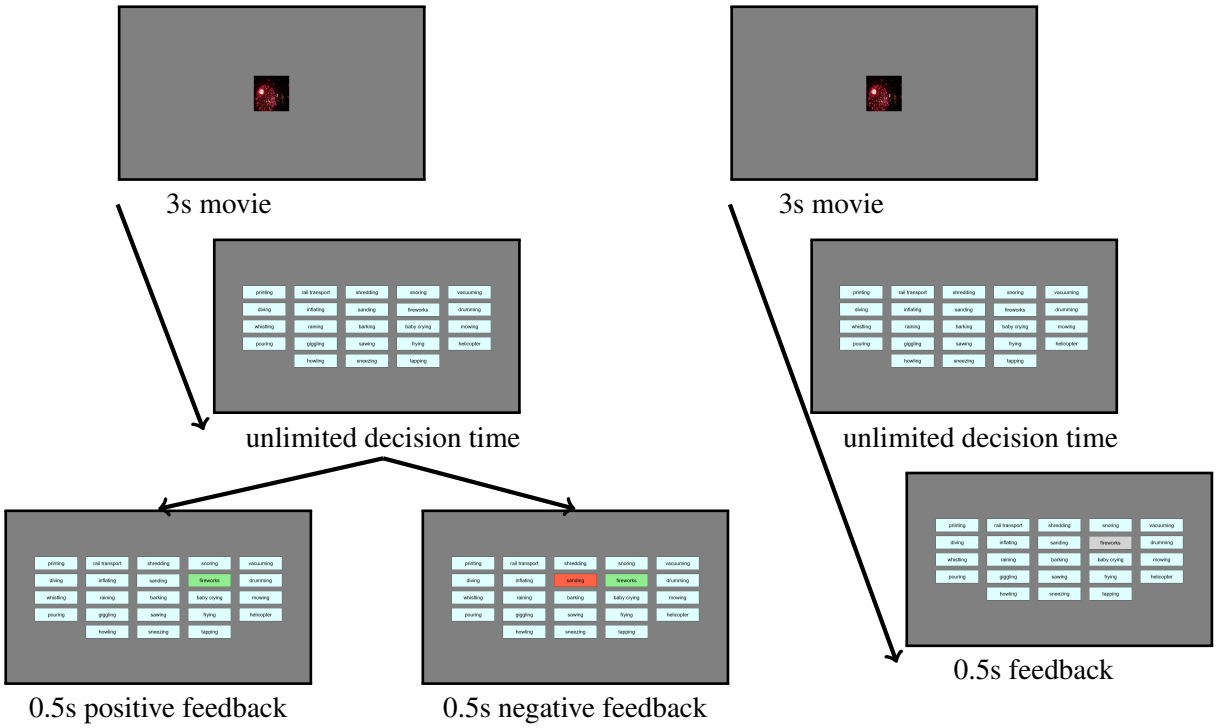
of these practice videos in each class was uniformly sampled and designated to be silent, the remaining video was audible. If participants gave 4 wrong answers (out of 46 practice videos) then the experiment was terminated and the participant was excluded from participating in the experiment. After each button response, feedback was given to participants to show them the correct answer (Figure 5.1a).

If participants passed the screening criteria, the first session would start automatically, first informing the participant that feedback would no longer be required before then displaying the first 230 videos to the participant. Participant button selections were given grey feedback so as not to guide participants on the accuracy of their answers (Figure 5.1b). After successful completion of the whole practice routine and experiment session, participants were invited to participate in a second session. This second session contained the experiment routine for the remaining 230 videos and could be completed by the participant within 3 days of completing the first session.

The DRCNN models developed in Chapter 4 were adopted for this work, but were instead trained and tested on AVMIT-VEGAS rather than AVMIT (Chapter 3) to provide a larger set of comparisons against humans. These models were tested on the exact same videos as human participants, under the same conditions (audio/silent; Gaussian noise, Gaussian blur). The DRCNNs were then tested on additional noise levels as it was found that they were more sensitive to small increases in noise than human participants, whose responses during piloting remained relatively unchanged in these noise ranges. We further tested the models on low contrast and salt and pepper noise.

5.3.3 Training and test videos

The balanced AVMIT-VEGAS dataset was selected as the training dataset to optimise our DRCNNs, with the held-out test set providing a set of clean, controllable audiovisual action videos that allow for direct comparison of DNN and human performance and behaviour. The AVMIT-VEGAS test set contains action videos corresponding to 23 classes, each with 60 videos, with a total of 1,380 videos. According to our pilot studies, participants typically classified



(a) Practice trial (b) Experiment trial

Figure 5.1: Schematic of the timelines of each trial in the experiment

around 400-500 videos per hour. Classifying all 1,380 videos would thus require around 3 hours of participation per participant. In order to obtain reliable estimates of human behaviour, we instead distribute our participant hours across more participants and fewer videos. To do this, 20 AVMIT-VEGAS test videos were uniformly sampled for each class and used in this study, providing a test set of 460 videos that could be classified in approximately 1 hour.

5.3.4 Distortions

Dodge and Karam (2016) found that deep neural networks were particularly poor at generalising to both additive Gaussian noise and Gaussian blur, we focus our experiments on these distortions when comparing humans against our models. We use the same methods of Gaussian noise and Gaussian blur distortion in our work as Dodge and Karam (2016). We further observe the behaviour of our models on salt and pepper noise and contrast reduction. As our work uses video data, we apply distortions independently to each frame. We also consider the undistorted case for human participants and our models. All images are colour and the RGB pixels are in the range

[0, 255]. We additionally consider these videos with and without complementary audio data.

We select a fixed visual frame resolution of 224x224 pixels for test stimuli as is required by EfficientNet-B0 and VGG-16. All Gaussian noise and Gaussian blur stimuli that are presented to humans are presented to the DRCNNs, the only alteration is the visual frame rate. We reduce the visual frame rate from 30fps for humans, to 2.08fps (1 frame every 480ms) for our YamNet + EfficientNet-B0 DRCNNs or 1.04fps (1 frame every 960ms) for our VGGish + VGG-16 DRCNNs. This reduction in visual frame rate is to allow alignment between the audio and visual embeddings from the CNN feature extractors. Where additional samples are taken in either modality, they were clipped from the stream.

For the Gaussian noise condition, we add to each channel of every pixel a noise value sampled from a Gaussian distribution centred at 0 with a standard deviation, $\sigma \in [100, 200, 300, 400]$. Although both human and model performance was examined at these noise levels, we additionally tested our models on the additional noise levels, $\sigma \in [10, 20, 30, 40, 50, 60, 70, 80, 90]$ as their performance deteriorates drastically in this range whilst human performance remains relatively unchanged. Examples of this method of adding Gaussian noise to coloured images in test convolutional neural networks can be found in Zhou, Song, et al. (2017), Dodge and Karam (2016), Dodge and Karam (2018), and Dodge and Karam (2019).

The Gaussian blur condition includes 12 different levels of noise. Again, we use the same procedure as Zhou, Song, et al. (2017), Dodge and Karam (2016), Dodge and Karam (2018), and Dodge and Karam (2019). The blur is applied using a Gaussian kernel with a standard deviation, $\sigma \in [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]$ with the kernel truncated at a distance of 4 times the standard deviation in each direction. To be more specific:

$$kernel\ shape = (8\sigma + 1, 8\sigma + 1) \tag{5.1}$$

As behavioural data collection time and associated cost is lower for our models than for human participants, we elect to further the behavioural examination of our models to contrast reduction and salt and pepper noise modes. Salt and pepper noise is generated by randomly assigning a number of pixels to have a 0 or 255 (white/salt or black/pepper) value; where salt

and pepper are added in equal amounts. Each pixel was assigned a new value according to probability p, where $p \in [10, 20, 30, 40, 50, 60, 70, 80, 90]\%$. For the low contrast condition, we use the method and formula in Geirhos, Temme, et al. (2018) but apply this method to all colour channels. For all contrast levels $c \in [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]$:

$$low\ contrast\ frame = c \cdot frame + \frac{255(1-c)}{2} \tag{5.2}$$

5.3.5 Human Observers

The human classification task was developed using Psychopy builder (Peirce et al., 2019) so that it could be hosted online using Pavlovia (Peirce et al., 2020). Participants were recruited using the participant recruitment service Prolific (*Prolific* 2014) and gave informed consent to take part in the study. For each participant, only mouse click locations and times were recorded alongside anonymous Prolific IDs, no sensitive or personal information was stored.

Altogether we recruited 20 participants, 10 for the Gaussian noise experiment and 10 for the Gaussian blur experiment. The 'exclusion list' function on prolific was used to prevent participants from one experiment participating in another. We used a number of other Prolific screening criteria to select participants; they had to be first language English speakers, have no hearing difficulties, have normal or corrected to normal vision, have a minimum approval rating from other Prolific studies of 80% and be using a computer with a Windows 10 operating system. Further, the participants had to be using a desktop computer with audio to participate in the study.

In addition to the Prolific screening criteria, we added our own screening task, outlined in section 5.3.2. This was to ensure participants understood the language used in the buttons, that their computers could properly present the video and that they understood the task before they were recruited on to the main experiment. Altogether, 14 participants were rejected from the experiment at this stage, a further 2 participants did not complete the whole experiment and their session timed out, we continued recruitment until 20 participants had successfully completed the entire study.

Each participant viewed all 460 videos (corresponding to 20 test videos per class) exactly once. Each participant viewed a video under a unique noise level, such that no 2 participants viewed the same video in that condition. All classes and noise levels were counterbalanced such that each participant viewed a video from each class under every condition (distortion level, audio/silent) exactly twice.

Participants were compensated £7.50 per hour, which was considered to be a reasonable rate by Prolific (*Prolific* 2014). This study was given a favourable opinion by the University of Birmingham Ethical Review Committee.

5.3.6 Accuracy and response distribution entropy

As in Geirhos, Temme, et al. (2018) we observe both classification accuracy and output distribution entropy. Each DRCNN is tested on every stimulus provided to participants (every video, distortion type, distortion level combination). Error bars provided on classification accuracy plots correspond to the range of human accuracies.

Taking the Shannon entropy of the output distribution for each model, distortion type, distortion level combination allowed us to investigate whether there were any biases present in the classifications made by human participants or computational models. The Shannon entropy, H, of the output distribution χ is obtained as follows:

$$H(\chi) = -\sum_{i=1}^{23} p(x_i) log_2(p(x_i)) \tag{5.3}$$

Entropy is a measure of how similar a distribution is to the uniform distribution (where there is no bias). For our 23-way classification task, the maximum value for Shannon entropy (indicating the uniform distribution) is approximately 4.52.

5.4 Results

On visual-only test stimuli without distortion, it can be observed that all RCNN classifiers obtained lower classification accuracies than all human participants (Figures 5.2a, 5.3a and

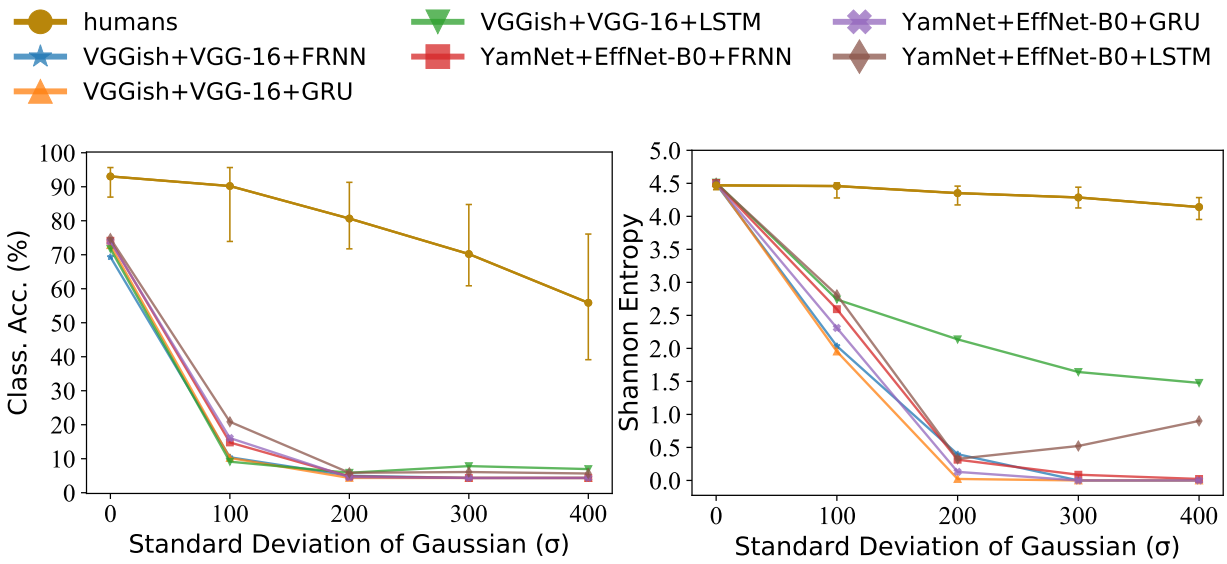
5.4a). Indeed, the performance of each individual classifier was revealed to be significantly different to the average human performance by a series of paired one-sample permutation tests (100,000 iterations; $p < 0.005$; Bonferroni corrected for multiple comparisons). Although the classifier errors were found to be significantly different to those of the human participants, it can be observed in Figures 5.2b, 5.3b and 5.4b that the entropy was approximately at the maximum value of 4.52 as for all human observers. Thus the classifiers were not performing worse at this condition due to class bias in their predictions.

Mean human performance increased by 3.48% when audio data was presented alongside clean visual data, an effect which was revealed to be significant by a paired permutation test (all 2^{20} permutations; $p = 0.00036621$; Bonferroni corrected for multiple comparisons). All DRCNN classifiers also obtained a significant increase in classification accuracy when audio was presented alongside the visual stimuli (100,000 iterations; $p < 0.00001$; Bonferroni corrected for multiple comparisons).

Unlike the visual-only task, 5 classifiers performed within the range of human performance when classifying clean audiovisual stimuli (Figures 5.2c, 5.3c and 5.4c). Only the VGGish+VGG-16+FRNN model performed outside of the range of human audiovisual performance with a classification accuracy of 88.45%, falling just outside the minimum human performance of 89.13%, although no significant effect was detected. Additionally, the entropy of each classifier's performance on the clean test stimuli is approximately maximum (4.52) for both the visual-only and audiovisual classification tasks as it is for each individual human participant.

5.4.1 Gaussian noise results

The introduction of Gaussian noise at σ=100 and above led to a large deterioration in DRCNN classifier performance in the visual-only condition (Figure 5.2a). Likewise the human average performance decreased as the noise level was increased. These visual-only classification accuracies were revealed by a series of paired permutation tests to be significantly lower for the DRCNN classifiers than the human participants at all examined Gaussian noise levels (100,000 iterations, $p < 0.005$, Bonferroni corrected for multiple comparisons). Similarly, in the audiovi-



(a) Visual class. acc.

(b) Visual Shannon entropy

(c) Audiovisual class. acc.

(d) Audiovisual Shannon entropy

Figure 5.2: Classification accuracy and Shannon entropy of the response distribution for visual object-only and audiovisual DRCNNs on Gaussian noise at a coarse resolution in the range $\sigma \in [100, 200, 300, 400]$.

sual condition, human average classification accuracy was found to be significantly higher than that of all DRCNN classifiers across all Gaussian noise levels (100,000 iterations, $p < 0.005$, Bonferroni corrected for multiple comparisons).

Human participants obtained higher classification accuracies when audio was present than when the stimuli were silent on all levels of Gaussian noise distortion (Figures 5.2a and 5.2c). At σ=100, however, this increase in classification accuracy was not found to be significant. Though via a series of paired permutation tests, we detected significant classification accuracy increases of 10.22%, 16.74% and 22.83% at $\sigma \in [200, 300, 400]$ respectively ($p = 0.005$, Bonferroni

corrected for multiple comparisons).

In the case of the DRCNN classifiers, there was a marked difference between the visual-only and audiovisual test performances. For each classifier, at all Gaussian noise levels, the difference between classification accuracy in the visual-only case and the audiovisual case (Figures 5.2a, 5.2c) was significant. This was revealed by a series of paired permutation tests, utilising the McNemar test statistic, carried out for each classifier/noise-level combination (100,000 iterations, $p < 0.0001$, Bonferroni corrected for multiple comparisons). It can be observed that the visual-only performance of all DRCNN classifiers decreases considerably at $\sigma = 100$ and reaches approximately chance performance (4.35%) at $\sigma \geq 200$. In the audiovisual condition, however, DRCNN classifiers reach some performance floor at $\sigma \in [100\ 200]$, from which point the reduction in performance as the noise level is increased is approximately 0.

As the visual-only model performance reduces drastically in the interval $\sigma \in [0, 100]$, we further tested the DRCNN classifiers at a finer noise resolution in this interval (Figure 5.3). The performance degradation in this noise interval in both the visual-only and audiovisual domain was found to be highly linear (Table 5.1). The relationship between visual and audiovisual classifier performances in this interval are in line with the previously observed relationship (i.e. audiovisual performance are larger than visual-only performance). Again, these performance differences were found to be significant (100,000 iterations, $p < 0.0001$, Bonferroni corrected for multiple comparisons).

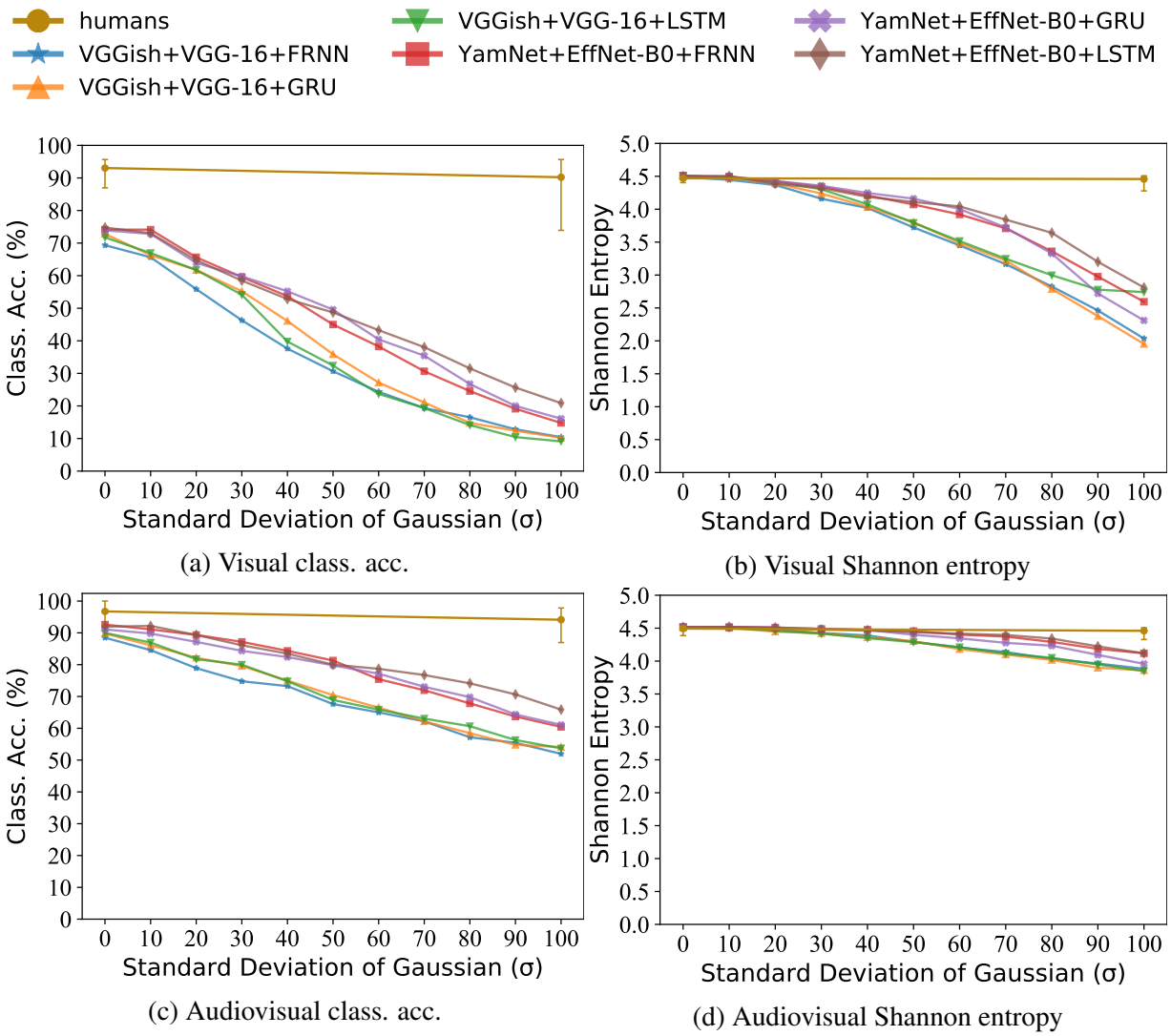
Table 5.1: Pearson correlation coefficient for the performances of each classifier on the Gaussian noise problem in the noise interval $\sigma \in [0100]$.

Classifier	Visual		Audiovisual	
	r	p-value	r	p-value
VGGish+VGG16+FRNN	-0.9821	$7.94x10^{-8}$	-0.9960	$9.62x10^{-11}$
VGGish+VGG16+GRU	-0.9901	$5.60x10^{-9}$	-0.9969	$2.99x10^{-11}$
VGGish+VGG16+LSTM	-0.9841	$4.67x10^{-8}$	-0.9958	$1.13x10^{-10}$
YamNet+EffNet+FRNN	-0.9962	$7.92x10^{-11}$	-0.9885	$1.09x10^{-8}$
YamNet+EffNet+GRU	-0.9954	$1.80x10^{-10}$	-0.9895	$7.26x10^{-9}$
YamNet+EffNet+LSTM	-0.9983	$1.93x10^{-12}$	-0.9919	$2.24x10^{-9}$

The mode of failure can be observed in the entropy plots (Figures 5.2b, 5.2d, 5.3b and 5.3d)

with classifier prediction sets decreasing in entropy alongside classification accuracy decreases. This shows the classifiers becoming increasingly biased as Gaussian noise is increased. Indeed, it is notable that the audiovisual predictions are less biased than the visual-only counterparts, this is alongside audiovisual increases in performance. There are 2 notable irregularities in the visual-only case (Figure 5.2b) with YamNet+EfficientNet-B0+LSTM becoming less biased as σ is increased beyond 200, and the entropy of the predictions made by VGGish+VGG-16+LSTM decreasing at a much more gradual pace than the other DRCNNs. This, however, was not sufficient to retain performance on the task. It can be observed that the entropy of the human prediction sets did not decrease considerably despite decreases in the classification accuracy, showing that error patterns were well distributed across classes, but revealing differences in error patterns between the DRCNN classifiers and human participants.

Although we ascertained that audiovisual performance was larger than visual performance for classifiers and humans at all noise levels, this could potentially be explained by the higher audiovisual accuracy in the clean condition. I.e. the audiovisual performance could be higher than the visual performance across noise levels because it was higher before noise was applied. We carried out a series of paired permutation tests to study the performance drop at each condition (the difference from the clean condition). The null hypothesis then becomes: There is no significant difference in the accuracy decrease due to Gaussian noise in the visual and audiovisual domains. The test statistic is then the mean difference between audiovisual performance drop and visual performance drop (difference-in-differences). The performance drop in the visual domain was significantly larger than the audiovisual domain for all classifiers at all conditions other than $\sigma = 10$ (100,000 iterations per test, $p < 0.005$, Bonferroni corrected for multiple comparisons). Similarly, these same permutation tests revealed human visual-only performance drop to be larger than audiovisual performance drop (100,000 iterations per test, $p < 0.005$, Bonferroni corrected for multiple comparisons).



(a) Visual class. acc.

(b) Visual Shannon entropy

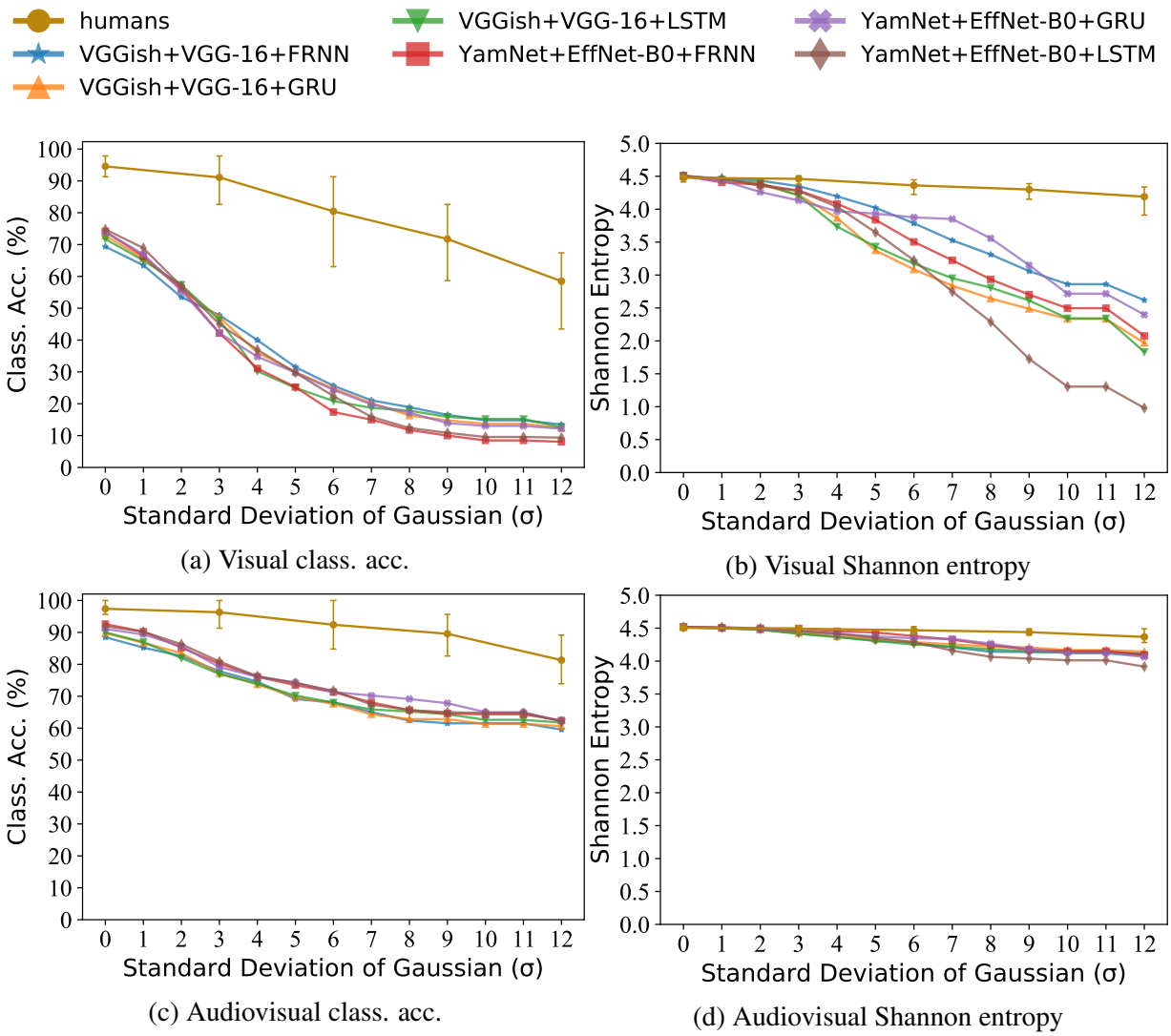
(c) Audiovisual class. acc.

(d) Audiovisual Shannon entropy

Figure 5.3: Classification accuracy and Shannon entropy of the response distribution for visual object-only and audiovisual DRCNNs on Gaussian noise at a fine resolution in the range $\sigma \in [10, 20, 30, 40, 50, 60, 70, 80, 90, 100]$.

5.4.2 Gaussian blur results

The visual-only classification accuracy of human participants was significantly higher than that of all DRCNN classifiers at all Gaussian blur distortion levels (100,000 iterations; $p < 0.005$; Bonferroni corrected for multiple comparisons). The difference in performance between DRCNNs and humans decreased in the audiovisual domain at all distortion levels, however these differences were still found to be significant by a series of paired permutation tests (100,000 iterations; $p < 0.005$; Bonferroni corrected for multiple comparisons).



(a) Visual class. acc.

(b) Visual Shannon entropy

(c) Audiovisual class. acc.

(d) Audiovisual Shannon entropy

Figure 5.4: Classification accuracy and Shannon entropy of the response distribution for visual object-only and audiovisual DRCNNs on Gaussian blur.

As in the Gaussian noise experiment, humans were able to utilise the audio data to increase their performance on the classification task when visual stimuli was degraded by Gaussian blur as shown by a series of paired permutation tests between visual-only and audiovisual performances (100,000 iterations, $p < 0.005$, Bonferroni corrected for multiple comparisons). The DRCNN classifiers also utilised the available audio information to increase their respective audiovisual performances above their visual-only performances (100,000 iterations, $p < 0.0001$, Bonferroni corrected for multiple comparisons).

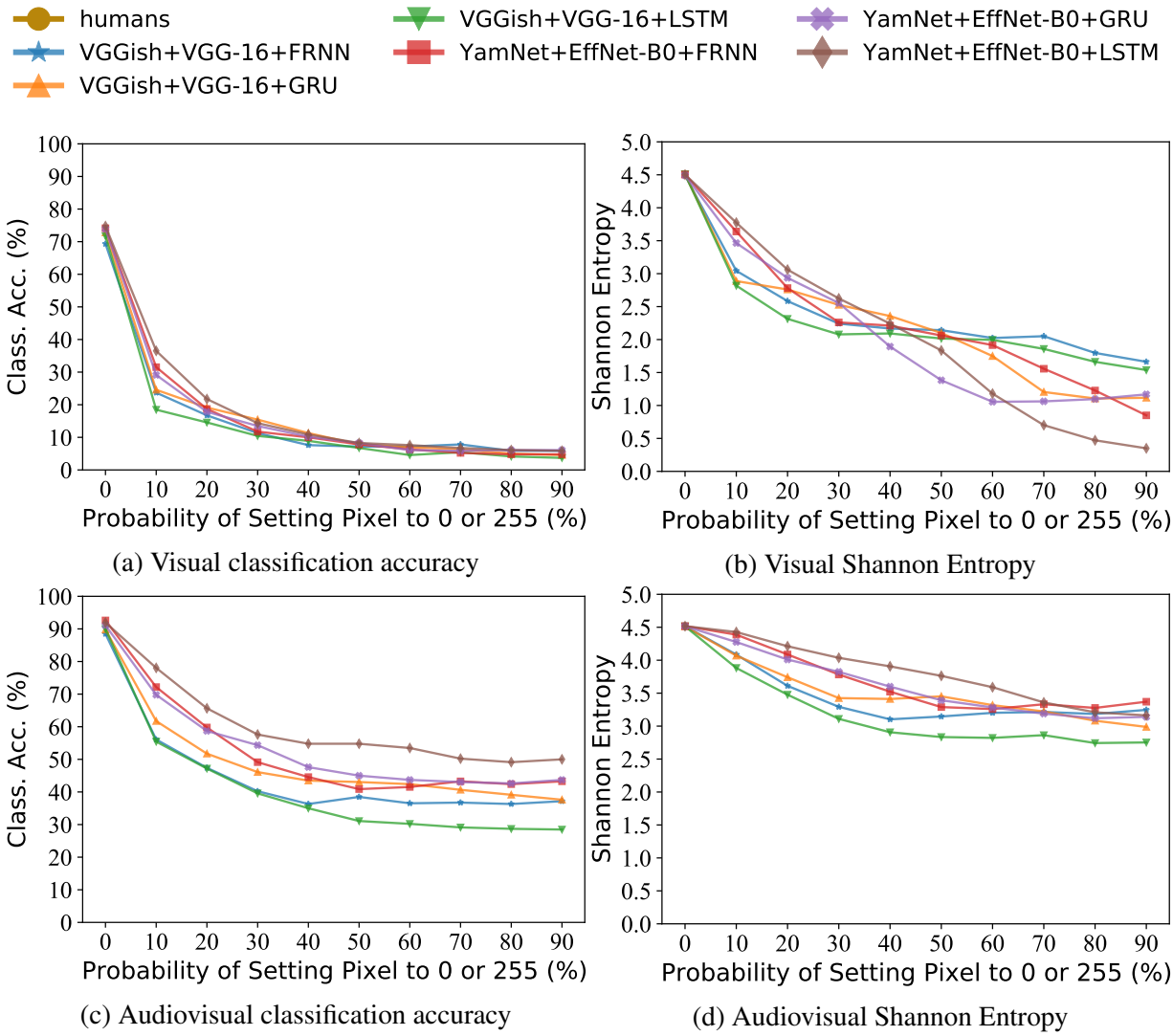
Following the Gaussian noise experiment, we ran further paired permutation tests to study

the difference in visual-only performance drop and audiovisual performance drop at all Gaussian blur distortion levels for humans and classifiers. These tests revealed a significant difference between the visual-only and audiovisual performance drop of each classifier at each Gaussian blur condition (100,000 iterations per test, $p < 0.005$, Bonferroni corrected for multiple comparisons). This was also found to be the case for human participants whose performance drop in the audiovisual domain was significantly smaller than that of the visual-only domain at all Gaussian blur levels (100,000 iterations per test, $p < 0.005$, Bonferroni corrected for multiple comparisons).

5.4.3 Salt and Pepper results

The salt and pepper noise was particularly destructive to classifier accuracy in our experiments at the given noise levels. Indeed, the classification accuracy decrease towards random chance over the examined noise levels (Figure 5.5a) is similar to that of the Gaussian noise experiment (Figure 5.3a). It can be observed that the entropy continues to decrease as salt and pepper noise is increased, despite small or zero change in classification accuracy whereas this is not the case for the Gaussian noise condition (Figure 5.5b). As in previous experiments, the DRCNN classifiers successfully utilise the audio data in the audiovisual condition to improve their classification accuracy when compared to the visual-only condition (Figures 5.5a and 5.5c). Indeed, the audiovisual classification accuracy was revealed to be significantly higher than the corresponding visual-only classification accuracy for the same distortion level by a series of paired permutation tests (100,000 iterations, $p < 0.0001$, Bonferroni corrected for multiple comparisons).

The performance drop of each classifier at each salt and pepper noise level, relative to clean performance, was revealed to be significantly different in the visual-only and audiovisual domain (100,000 iterations per test, $p < 0.005$, Bonferroni corrected for multiple comparisons). Equally, humans experienced a significantly smaller performance drop in the audiovisual domain than the visual-only domain across all distortion levels (100,000 iterations per test, $p < 0.005$, Bonferroni corrected for multiple comparisons).



(a) Visual classification accuracy

(b) Visual Shannon Entropy

(c) Audiovisual classification accuracy

(d) Audiovisual Shannon Entropy

Figure 5.5: Classification accuracy and Shannon entropy of the response distribution for visual object-only and audiovisual DRCNNs on salt and pepper noise.

5.4.4 Low Contrast results

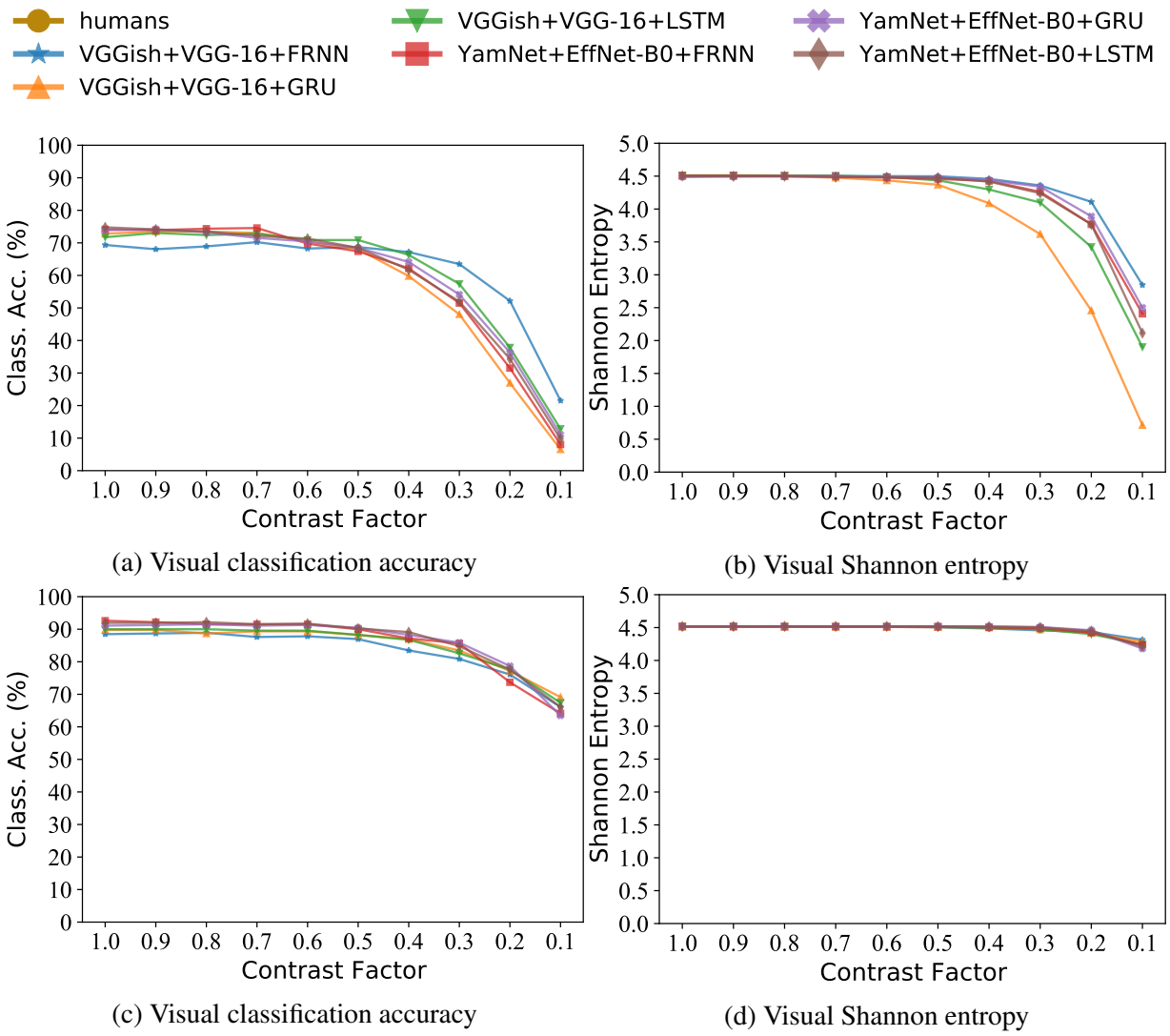
DRCNN classifiers tested in the visual domain on reduced contrast stimuli were resilient over the examined distortion levels, maintaining close to clean stimuli performance until the contrast factor reached 0.4 (Figure 5.6a). It can be observed in the complementary entropy plot (Figure 5.6b) that the distribution of predictions over all classes was approximately uniform until contrast levels of around 0.5 or 0.5, then decreasing in a similar fashion to the classification accuracy. As in the previous experiments, the DRCNN classifiers were able to increase their classification accuracies by utilising audio data in the audiovisual domain. A series of paired

permutation tests showed that the difference between visual-only and audiovisual classification accuracy was significant at all distortion levels (100,000 iterations, $p < 0.0001$, Bonferroni corrected for multiple comparisons). Further permutation tests revealed performance drops in the audiovisual domain where significantly different than that of the visual-only domain at $c \leq 0.3$ for all classifiers other than VGGish+VGG-16+FRNN (100,000 iterations per test, $p < 0.005$, Bonferroni corrected for multiple comparisons). Although the VGGish+VGG-16+FRNN classifier was more resilient in the visual-only domain (with the highest visual-only accuracies at $c \leq 0.3$), paired permutation tests revealed these visual-only performance drops to be significantly larger than audiovisual performance drops at $c \leq 0.2$ (100,000 iterations per test, $p < 0.005$, Bonferroni corrected for multiple comparisons).

5.5 Discussion

This study sought to address 3 experiment questions: Can dual-stream recurrent convolutional neural networks (DRCNNs) tested in the visual domain retain clean-stimuli performance when frame-level noise is introduced? To what extent can DRCNNs leverage audio signals to preserve audiovisual performance when frame-level noise is introduced? Are the classification accuracy scores and error patterns of DRCNN classifiers similar to that of human participants? These experiment questions arise from the current research where DNN and human performance have been compared on distorted images (Dodge and Karam, 2016; Geirhos, Janssen, et al., 2017; Wichmann et al., 2017; Dodge and Karam, 2017; Geirhos, Temme, et al., 2018; Dodge and Karam, 2019) and where visual data has been utilised to increase robustness of DNNs on speech recognition tasks (Gabbay et al., 2018; Ephrat et al., 2018; Zhou, Yang, et al., 2019; Yu et al., 2020; Aldeneh et al., 2021).

Our experiments are the natural next step in the literature. Thus far, the literature has uncovered a number of vulnerabilities of ImageNet-trained CNNs to visual distortions (Dodge and Karam, 2016; Geirhos, Janssen, et al., 2017; Wichmann et al., 2017; Dodge and Karam, 2017; Geirhos, Temme, et al., 2018; Dodge and Karam, 2019). We first explore the ability



(a) Visual classification accuracy

(b) Visual Shannon entropy

(c) Visual classification accuracy

(d) Visual Shannon entropy

Figure 5.6: Classification accuracy and Shannon entropy of the response distribution for visual object-only and audiovisual DRCNNs on low contrast.

of RNNs to accumulate noisy evidence in the form of CNN embeddings to distorted image sequences (videos). A model that uses a CNN to extract spatial features at each time point and an RNN to resolve over the temporal dimension is known as a RCNN. However, in order to provide comparison to performance in the audiovisual domain (experiment question 2) we use the dual-stream RCNNs of Chapter 4. These DRCNNs have 2 CNNs, 1 audio and 1 visual, to extract feature embeddings at each time-point. These models were all trained on the AVMIT-VEGAS training set (Chapter 3), and all test data was sourced from the AVMIT-VEGAS test set prior to distortion. Distortions examined where Gaussian noise, Gaussian blur, salt and pepper noise and

contrast reduction.

The visual-only classification accuracies of all DRCNN classifiers were reduced to chance (4.35%) or close to chance performance under high levels of Gaussian noise, Gaussian blur, low contrast or salt and pepper noise with few exceptions. Performance on Gaussian noise was reduced to 9-21% at $\sigma \geq 100$ but degraded to chance or marginally above chance level performance on larger distortion levels. Gaussian blur reduced classifier performance to 8-14% accuracy at the highest distortion level of $\sigma = 12$. Salt and pepper noise reduced the performance of all classifiers to just above random chance as noise increased to a pixel reassignment rate of 50% and above. Classifiers scored between 6-13% accuracy at the lowest contrast level (highest distortion level) except the VGGish+VGG-16+FRNN classifier which obtained 21.52% accuracy. The DRCNN classifiers examined here did not preserve their clean-stimuli performances of 69-75%. Where performance decreased, this was almost always coupled with an increase in class bias (decrease in response entropy) in line with reported failure modes in the image recognition domain (Geirhos, Temme, et al., 2018).

Next, we advance the research from the visual to the audiovisual domain to examine the extent to which dual-stream RCNN classifiers are able to dynamically weigh sensory modalities when visual distortions are present. This is parallel to the advancements in audiovisual speech recognition (Gabbay et al., 2018; Ephrat et al., 2018; Zhou, Yang, et al., 2019; Yu et al., 2020; Aldeneh et al., 2021) where systems utilise visual information to overcome auditory noise. The audiovisual performance of all DRCNN classifiers was found to be significantly higher than the corresponding visual-only performance at each distortion level for all examined distortions and at every distortion level. Indeed, audiovisual performance was higher than visual-only performance on the clean stimuli condition, so we further explored the performance decrease in each domain to ensure that audiovisual performance was not higher purely due to higher performances before degradation. We studied performance decreases at each distortion level (relative to undistorted performance) and found that the difference between these decreases in the visual and audiovisual domains was also significant for Gaussian noise, Gaussian blur, salt and pepper noise and high levels of contrast reduction ($C \leq 0.2$). Thus, for all dual-stream RCNN

classifiers, informative audio data alongside visual data did not only improve performance, but also increased robustness to visual noise (decreased the rate of degradation). It could be observed from Shannon entropy plots that the bias of predictions increased for all classifiers as distortion level was increased, although this was comparatively less than in the visual domain in line with the change in classification accuracy.

Where the literature has also began to compare the robustness of human perception to that of DNNs, we extend our behavioural experiments to humans as the final contribution of this study. Human participants were recruited on Prolific (*Prolific* 2014) and carried out an online classification task that we developed with Psychopy (Peirce et al., 2019) and hosted on Peirce et al. (2020). Humans were specifically tested on Gaussian noise and Gaussian blur distortion types that have been shown to be particularly destructive to CNN performance (Dodge and Karam, 2016; Dodge and Karam, 2017; Geirhos, Temme, et al., 2018) to observe whether dual-stream RCNNs better match human performance. On the clean stimuli condition, human participants significantly outperformed dual-stream RCNN classifiers in the visual domain, but in the audiovisual domain, classifiers performed within the range of human performance.

Across both Gaussian noise and Gaussian blur distortion types and all measured distortion levels, human participants outperformed our dual-stream RCNN classifiers. Additional tests of the classifiers at finer resolutions of Gaussian noise in the $\sigma \in [0\ 100]$ interval, however, suggests that the YamNet+EfficientNet-B0 based classifiers perform within the range of human performance at $\sigma = 10$ in the audiovisual domain (as their performance did not degrade at this distortion level). Human participants were also found to effectively leverage audio information to significantly improve their performance in the audiovisual domain over their performance in the visual domain. Indeed, human audiovisual accuracy was significantly higher at all levels of Gaussian blur and Gaussian noise levels above $\sigma = 100$ (there was an increase in audiovisual performance at this level but no significant effect was detected). The performance decreases in the audiovisual domain were shown to be significantly smaller than those of the visual domain at Gaussian noise levels above $\sigma = 100$ and all levels of Gaussian blur, demonstrating an increase in robustness to visual noise when audio was presented alongside visual stimuli as in the dual-stream

RCNN classifiers.

DNNs in the literature have been shown to increasingly bias particular classes as the signal gets weaker (Dodge and Karam, 2017; Wichmann et al., 2017; Geirhos, Temme, et al., 2018) whereas human error patterns appear to be more distributed (Geirhos, Temme, et al., 2018). It is worth noting, however, that DNNs give an output distribution, but the classification is often taken to be the maximum probability (top 1) or if the answer is in the top 5 probabilities (top 5). In this way, the output probability distribution is not fully considered. Dodge and Karam (2016) for instance reported that DNNs were 'less confident' of the correct classification as the signal gets weaker, referring to the observed reduction in assigned probability to that class. Geirhos, Temme, et al. (2018) shows how the temperature parameter could be adjusted to sample from the softmax output distribution rather than using the argmax function, but reports that the increase in response distribution entropy comes at the expense of classification accuracy. This observation in the literature, that DNNs become increasingly biased to particular classes as distortion level is increased, was certainly observed in this study on all distortions, but we provide a method to mitigate this through the use of other modalities.

This study has multiple implications for the literature. Firstly, the FRNN, GRU and LSTMs were unable to compensate for the noisy VGG-16 or EfficientNet-B0 representations to retain visual-only performance when distortions were introduced, despite processing multiple frames and accumulating evidence via recurrence. But despite this lack of resilience in the visual-only domain, the dual-stream RCNN classifiers were able to utilise audio information to better preserve performance under visual distortion. Thus, those implementing video event recognition systems for automatic video captioning, generating video descriptions, surveillance, robots amongst other uses may choose to implement dual-stream RCNN classifiers (or use our classifiers) following the results of this study. Whilst multimodal DNN performance gains have been realised in the literature (Petridis et al., 2017; Tao and Busso, 2017; Gogate et al., 2018; Zhang, Wang, et al., 2019) including some in action recognition specifically (Nagrani et al., 2021; Akbari et al., 2021) these studies have focussed solely on accuracy rather than reliability under noisy conditions. We look to the area of audiovisual speech recognition where researchers have used

the addition of visual information to resolve issues with auditory noise (Gabbay et al., 2018; Ephrat et al., 2018; Zhou, Yang, et al., 2019; Yu et al., 2020; Aldeneh et al., 2021) and show that our dual-stream RCNN classifiers are able to utilise audio to dampen the performance decreases due to visual noise. We further implicate the research by demonstrating deviations in the classification performance and error patterns of human and DNNs when RNNs are added to CNNs (RCNNs) in the image sequence recognition domain and that the ability to leverage audio data to decrease the rate of deterioration still does not match human performance.

CHAPTER 6

GENERAL DISCUSSION

The work presented in this thesis contributes to the area of research exploring human intelligence with DNNs, specifically in the area of audiovisual perception. I produced a large, labelled video dataset of *audiovisual* action events suitable for examining DNNs and human participants and documented the process and results in Chapter 3. In Chapter 4 I investigated the ability of dual-stream recurrent convolutional neural network (DRCNN) classifiers, optimised on audiovisual action recognition, to solve the audiovisual correspondence task. The final study chapter (Chapter 5) detailed the comparison of DRCNN classifiers and human participants, in the visual and audiovisual domain, when visual distortions are applied. In this General Discussion chapter, I will summarise the findings of empirical chapters and how they advance the field before providing directions for future research.

6.1 Findings

6.1.1 Creating a large, labelled dataset of audiovisual action events

The overarching aim of the work presented in this thesis was to extend the area of research that uses deep neural networks as investigative tools in cognitive science into the realm of audiovisual perception. Although audiovisual integration is an established area of research in cognitive science (Stein and Meredith, 1993; Stein, 2012) others have previously focussed on unisensory systems such as the human visual (Yamins, Hong, Cadieu, and Dicarlo, 2013; Yamins, Hong,

Cadieu, Solomon, et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014; Güçlü and Van Gerven, 2015; Cichy, Khosla, et al., 2016; Stabinger et al., 2016; Seibert et al., 2016; Dodge and Karam, 2017; Geirhos, Janssen, et al., 2017; Rajalingham et al., 2018; Geirhos, Temme, et al., 2018; Dodge and Karam, 2019; Jacobs and Bates, 2019; Singer et al., 2020; Heinke et al., 2021) or auditory cortex (Kell, Yamins, et al., 2018). In these cases, researchers either use a held-out test set from the dataset or alternate stimuli sourced elsewhere.

We elected to use an audiovisual action recognition problem, given the clear ecological relevance to humans, as the principal problem for which we would optimise our DNNs and test them against human participants. Particularly as there was an abundance of large, labelled video datasets for action recognition that have been released (Heilbron et al., 2015; Gu et al., 2018; Monfort et al., 2019; Li, Thotakuri, et al., 2020; Smaira et al., 2020). These datasets, however, contain videos of labelled audio and/or visual events, but not specifically *audiovisual events*. This is the first problem that presented itself; there were no large, labelled video datasets of *audiovisual events*.

The aim of the first study, detailed in Chapter 3, was to obtain a clear, held-out test set that could be used alongside a large, labelled training dataset to compare DNNs and human participants on an action recognition task. To solve this, we sourced candidate videos from the Moments in Time (MIT; Monfort et al., 2019) dataset and carried out a large-scale sorting task with trained participants. Additional candidate videos were then sourced from the Visually Engaged and Grounded AudioSet (VEGAS; Zhou, Wang, et al., 2018) to extend the dataset. The sorting task provided dataset characteristics that could further be used to create training/test sets with videos that met selected criteria.

Our study found that only 17,904 videos out of the 61,248 sorted MIT videos were classified as containing a properly labelled, dominant, audiovisual event by a majority of our trained participants. This was despite removing videos without audio streams or with digital silence, and despite removing classes considered to not have informative audio and visual signals. Similarly, 6,411 video clips out of the 16,283 sorted VEGAS video clips were verified by a majority of our trained participants to contain properly labelled, dominant, audiovisual events. This was despite

the VEGAS annotation task in which each 2-second excerpt was verified to contain the labelled event in the audio and visual streams by Amazon Mechanical Turkers (Crowston, 2012).

We used total agreement of trained participants about the presence and dominance of labelled audiovisual action events as criteria for videos from the MIT dataset to be included in the clean test set for human vs DNN comparison. We further required that all test videos had a frame rate of 30fps and removed videos with excessive noise (e.g. watermarks). From this process a test set of 960 videos (16 classes, 60 videos per class) was produced. A further complementary training set of audiovisual action events was produced by using majority votes of trained participants, altogether containing 11,109 videos (7,296 for balanced dataset; 456 videos per class). We name the training and test set the Audiovisual Moments in Time (AVMIT) dataset. An extended training and test set, named AVMIT-VEGAS, was produced using the video clips originally sourced from the VEGAS dataset, producing a test set of 1,380 videos (23 classes; 60 videos) and a training set of 17,578 videos (10,488 for balanced dataset; 456 videos per class).

6.1.2 Developing deep neural network models of audiovisual perception

With a benchmark for human vs DNN comparison prepared, we required DNN models of human audiovisual perception that could be optimised on the task. Recurrent convolutional neural networks (RCNNs; Donahue et al., 2015; Tsironi et al., 2016; Ning et al., 2017; Çakır et al., 2017; Yang et al., 2019; Sabir et al., 2019; Khaki et al., 2020; Gupta et al., 2021), utilising a convolutional neural network (CNN; LeCun, Boser, et al., 1989; Krizhevsky et al., 2012) to extract spatial embeddings at each time-point and a recurrent neural network (RNN) to process them in sequence, provided an ideal architecture type for a number of reasons. Firstly, RCNNs allowed us to extend on the current literature that had compared CNNs to human visual (Dodge and Karam, 2016; Dodge and Karam, 2017; Wichmann et al., 2017; Geirhos, Temme, et al., 2018; Dodge and Karam, 2019; Stabinger et al., 2016; Heinke et al., 2021; Yamins, Hong, Cadieu, and Dicarlo, 2013; Yamins, Hong, Cadieu, Solomon, et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014; Güçlü and Van Gerven, 2015; Cichy, Khosla, et al., 2016) and auditory perception (Kell, Yamins, et al., 2018). Secondly, RCNNs have both feedforward and recurrent connections,

making them more biologically realistic than CNN models of perception (Spoerer, McClure, et al., 2017). Thirdly, we found a way that we could extend these model architectures into the audiovisual domain, whilst leveraging transfer learning and state-of-the-art unimodal CNNs to obtain high performances on the task. Lastly, Spoerer, McClure, et al. (2017) considered the use of RCNNs as models of biological object recognition, finding that recurrent neural networks outperformed purely feedforward networks on an object recognition task including in the presence of occlusions or Gaussian noise.

As our work was in the audiovisual domain, we introduced the dual-stream RCNN architecture to process audiovisual sequences in Chapter 4. DRCNNs utilise two CNNs, one to extract features from the visual frame and the other to extract features from an audio clip at each time-step. The two embeddings are then fused together into a single audiovisual embedding using a series of operations that we name a 'Multimodal Squeeze Unit' (similar to the 'squeeze' of a 'squeeze-excitation' block used in modern CNNs; Tan and Le, 2019). These embeddings form a sequence that are then fed to an RNN for sequence classification.

In order to understand some of the architectural similarities between our DRCNN models and human systems of multisensory perception we can consider the fusion point of the unimodal data and the subsequent operations. In the field of deep learning, 'early' and 'late' fusion can refer to very different systems. While Atrey et al. (2010) attempts to formalise the nomenclature in deep learning according to *feature* level fusion and *decision* level fusion, we instead consider the utility of the unimodal features prior to fusion to marry these terms with the field of neuroscience. The DRCNN algorithms developed in this work would be considered 'early fusion' by Atrey et al. (2010), but in neuroscience, these architectures are better described as 'late fusion'. This is because the entire unimodal solution space is activated before any multisensory processing, i.e. the CNN components extract unimodal embeddings that are capable of solving object/event recognition problems before they are fused. In the human brain, this would be analogous to fusing visual features after processing in IT cortex at the end of the human visual ventral stream (late fusion).

In the human brain, multisensory processing occurs at multiple levels of processing. Tra-

ditionally, multisensory integration was thought to only occur after considerable unisensory processing at late stages in the cortical hierarchy, such as parietal and prefrontal cortices (Calvert et al., 2000; Macaluso, Driver, et al., 2003; Barraclough et al., 2005; Stevenson and James, 2009). Many studies carried out on the brains of cats and rodents have revealed multisensory integration occurs in the superior colliculus of the mammalian midbrain (Stein and Meredith, 1993) and modern studies suggest that integration takes place in early sensory areas of the cortex too (Foxe et al., 2002; Lee and Noppeney, 2011; Lee and Noppeney, 2014). Not only does multisensory integration take place in multiple places in the human brain, but there are suggestions that different *types* of integration take place at different stages of cortical processing (Noppeney, Jones, et al., 2018). For instance, stimuli from one modality can enhance or suppress responses to preferred stimulus in sensory areas, or add to the information content in those areas (Noppeney, Jones, et al., 2018). While higher order integration such as that mediated by superior temporal sulcus has been shown to integrate higher order features in categorisation tasks (Amedi et al., 2005; Werner and Noppeney, 2010) and speech recognition tasks (Calvert et al., 2000; Wright et al., 2003). The literature also reveals that these higher-order association areas tend to be less sensitive to the exact timing of stimuli (Werner and Noppeney, 2011) and the spatial displacement of visual and auditory stimuli (Macaluso, George, et al., 2004).

The suggestion that late multisensory fusion is concerned with higher-order features makes intuitive sense. The representations become further abstracted from the raw signal and can become increasingly rich in semantic information deeper into a hierarchical perceptual model, so any multisensory fusion that occurs late in the processing hierarchy seems likely to be semantic in nature. For instance, the output embedding from the visual CNN in our DRCNNs may contain the information for 'dog' and the audio CNN the information for 'barking' before these representations are fed into the multimodal squeeze unit. This would be consistent with late fusion in higher-order areas of the cortex. To follow the literature (Noppeney, Jones, et al., 2018) and better model audiovisual integration in the human brain, one could introduce lateral connections between the CNN components of the DRCNNs to also allow information to flow at earlier stages of processing. This may provide the ability to model signal-level interactions of

the data in different modalities i.e. more sensitive to timing and spatial displacement.

One may also consider accumulator models of audiovisual perception in the human brain and compare them to our DRCNN models. The work by Noppeney, Ostwald, et al. (2010) suggest that the left inferior frontal sulcus (IFS) accumulates audiovisual evidence, utilising recurrent loops with auditory and visual cortices. Our DRCNN models also carry out considerable unisensory processing at each timestep (CNNs) and then are fused (multimodal squeeze unit) before they are processed in accordance with fused audiovisual data at previous timesteps (RNN). The inductive biases of the recurrent neural networks mean that the audiovisual embeddings are received sequentially and that they are not stored in memory. Thus the RNN must be able to compress past audiovisual information (or evidence). This is in line with the suggested system of audiovisual evidence accumulation in prefrontal cortex (Noppeney, Ostwald, et al., 2010). Indeed, further research has shown that RNNs (and RCNNs) can be adjusted for speed/accuracy trade-offs, much like human participants, using output entropy thresholds (Spoerer, Kietzmann, et al., 2020).

In order to better understand the learnt behaviour of dual-stream RCNN classifiers and better generalise our results beyond a single architecture, we developed a set of 6 DRCNNs in our work. These were each made up of 1 of 2 audiovisual feature extractors (the CNNs) and 1 of 3 RNNs. The audiovisual feature extractors were either VGG-based with an ImageNet-trained VGG-16 (Simonyan and Zisserman, 2015) and an Audio Set-trained VGGish (Hershey et al., 2017), or were MobileNet-based with ImageNet-trained EfficientNetB0 (Tan and Le, 2019) and Audio Set-trained YamNet (Plakal and Ellis, 2020). The RNNs were either fully-recurrent neural network (FRNN, also known as a 'basic' or 'vanilla' RNN), gated recurrent unit (GRU; Cho et al., 2014) or a long short-term memory unit (LSTM; Hochreiter and Schmidhuber, 1997). These 6 DRCNN models, developed and outlined here, were the focus of this thesis.

6.1.3 Audiovisual correspondence encoded in dual-stream recurrent convolutional neural network classifier embeddings

With the audiovisual action recognition benchmark obtained and the dual-stream RCNNs developed, we were able to move on to explore perceptual behaviours and learnt abilities of the models. In our first DNN investigation in Chapter 4, we sought to explore how audiovisual perceptual behaviours arise due to the constraints of ecologically relevant tasks. Namely, we asked whether the ability to solve the audiovisual correspondence (AVC) problem arises implicitly as a consequence of optimising on an audiovisual recognition problem. The ability of humans to solve the audiovisual correspondence problem (and indeed the more general multisensory correspondence problem) is a well researched topic in cognitive science (Körding et al., 2007; Shams and Beierholm, 2010; Aller and Noppeney, 2019; Mihalik and Noppeney, 2020), yet how the brain came to solve this problem without explicitly being presented with the ground truth (the causal structure) of the world is unknown.

Although the means by which the solution to the audiovisual correspondence problem came to be encoded in the brain is unknown, many of the cues that humans use to solve the problem are, including spatiotemporal cues (Munhall et al., 1996; Slutsky and Recanzone, 2001; Lewald and Guski, 2003; Wallace et al., 2004) and higher-order cues (Laurienti et al., 2004; Parise and Spence, 2009; Calvert et al., 2000; Doehrmann and Naumer, 2008; Noppeney, Ostwald, et al., 2010; Krugliak and Noppeney, 2016). It is these latter cues, specifically the semantic cues, that we focus on in this work. Indeed, late fusion methods of audiovisual integration, such as is employed in our dual-stream RCNN models, lend themselves much more to leveraging semantic cues than early fusion methods. Particularly such high-level semantic concepts such as the action events in our benchmark are those typically represented at higher levels of the cortical hierarchy and deep layers of DNNs (Yamins, Hong, Cadieu, and Dicarlo, 2013; Cadieu et al., 2014; Yamins, Hong, Cadieu, Solomon, et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014; Güçlü and Van Gerven, 2015; Cichy, Khosla, et al., 2016). So it would stand to reason that early fusion methods would not rely so heavily on semantic cues.

Specifically, our aim in this work was to answer the following question ‘Is it possible that a

dual-stream RCNN classifier, optimised on an audiovisual action recognition task, will implicitly learn to solve the audiovisual correspondence problem?'. To explore the topic, we optimised an instance of each of the 6 dual-stream RCNN models on the AVMIT dataset, and optimised another instance on a similar dataset with lower levels of audiovisual correspondence, MIT-16. We obtained MIT-16 by sampling the largest possible, balanced subset across the 16 AVMIT classes. In this way, we could explore whether AVMIT's high levels of audiovisual correspondence were necessary for a dual-stream RCNN to learn to solve the AVC problem. Altogether, this left 12 classifiers (2 instances of each dual-stream RCNN).

The results revealed that all dual-stream RCNN classifiers implicitly learnt to solve the AVC problem, even if they were trained on MIT-16, demonstrating that the higher audiovisual correspondences were not a necessity as hypothesised. However, if the multimodal squeeze unit and RNN kept the audio and visual representations completely separate, the ability to solve the AVC task would be less interesting, as it reduces to a logical AND problem. Thus we further explored the extent to which the audio and visual signals were integrated in the audiovisual embeddings produced by the RNN before classification.

6.1.4 Multisensory integration in dual-stream recurrent convolutional neural network classifiers

To explore the extent to which audio and visual signals were integrated in the RNN embeddings of our dual-stream RCNN classifiers, we carried out a series of behavioural experiments. First we looked to the work of Ngiam et al. (2011) which provided 2 behavioural experiments we could carry out on our audiovisual embeddings; the cross-modal learning task and the shared-representation learning task. The cross-modal learning task assesses the extent to which a multisensory system has encoded unisensory information. The shared-representation learning task assesses the extent to which units that encode the signal of one modality also encode the other, the *shared representation*.

More generally though, the experiments introduced by Ngiam et al. (2011) provided a framework for assessing the learnt behaviour captured in multisensory embeddings that we

extended to introduce 2 novel tasks; the congruent selective-attention task and the incongruent selective attention task. In these experiments, linear support vector machines (SVMs) were implemented, as before, and trained to map the unimodal signal present in the multimodal embedding (thus 'attending') to the correct unimodal answer. At test time, however, we provide multimodal stimuli, either congruent or incongruent in the 'unattended to' modality. In this way, the performance difference between the congruent/incongruent selective-attention task and the cross-modal learning task informs of how the congruent or incongruent stimuli affects the unimodal classification. This further informs us of the extent to which the audio and visual signals are integrated in the dual-stream RCNN embeddings, and the extent to which they interact to affect behavioural performance. These selective-attention tasks have been important in the area of psychology, where they have been used to explore multisensory integration (Yuval-Greenberg and Deouell, 2007; Noppeney, Ostwald, et al., 2010; Leo and Noppeney, 2014; Krugliak and Noppeney, 2016). In the area of multimodal learning with deep neural networks, they provide a behavioural measurement of the extent to which portions of an embedding encode multimodal data. Indeed, in the context of this thesis, these novel tasks for DNNs form an important method of exploring multisensory behaviour.

Although the dual-stream RCNN classifiers were capable of solving the cross-modal learning task at performance levels (¿68% accuracy) significantly above random chance (6.25%) in both the audio and visual domain, this unisensory performance was affected by data in the other modality. The shared-representation learning, congruent and incongruent selective-attention tasks all resulted in significant changes in the classification accuracy, other than 3/24 congruent test cases where significant effects were not detected for small performance gains. But in these cases of small classification accuracy changes that were not identified as significant, those classifiers obtained significant performance changes in the other modality on the congruent task, and across all other tasks. Thus the representations constructed by all dual-stream RCNN classifiers in our studies integrated signals from audio and visual modalities and did not keep activations completely separate. One may further consider these results alongside those of Yuval-Greenberg and Deouell (2007). In both studies, congruent performances were significantly higher than

incongruent performances (in their human participants and our DRCNN classifiers). Although one must consider the difference in task (animal classification vs. action recognition) and stimuli (image & audio vs. video & audio) this shows that both human participants and our DRCNN classifiers are affected by signals in the unattended to modality according to congruency. These results conclude our findings in Chapter 4 around the audiovisual correspondence problem and audiovisual integration.

6.1.5 Dual-stream RCNNs and humans: visual perception as the visual signal gets weaker

In Chapter 5 we sought to further explore the dual-stream RCNNs as models of human perception. There is a growing body of literature examining the vulnerabilities of CNNs to visual noise and in some cases comparing to human performance (Dodge and Karam, 2016; Geirhos, Janssen, et al., 2017; Wichmann et al., 2017; Dodge and Karam, 2017; Geirhos, Temme, et al., 2018; Dodge and Karam, 2019). One question that then arises is, given the CNN components of RCNNs, will they suffer the same vulnerabilities to visual noise in image sequences, or will the RNN components be able to utilise multiple frames and persistent signals to accumulate evidence towards a particular classification. Further, how will performance and error patterns compare to that of human participants. To explore this question, we elected to use AVMIT-VEGAS benchmark rather than the previously used AVMIT (Chapter 4) benchmark, as we no longer required classifiers optimised on data with lower audiovisual correspondences, and the additional classes of the extended dataset provided a more comprehensive test set. This produced 6 dual-stream RCNN classifiers, 1 per architecture. These were tested in the visual-only domain using SVMs (cross-modal learning task).

The 6 dual-stream RCNN classifiers and human participants were tested on Gaussian noise and Gaussian blur distortions, with classifiers further examined on salt and pepper noise and contrast reduction. These experiments were visual-only, to isolate the effects of the distortion on visual accuracy. On clean, visual-only stimuli, classifiers obtained classification accuracies of 69-75% accuracy, significantly below the human average accuracy of 93.80%.

When Gaussian noise was introduced to the videos, classifier performance decreased linearly to 9-21% at $\sigma = 100$ and then further decreased to approximately chance performance as the distortion level was increased to $\sigma = 400$. Human performance, however, deteriorated much less, decreasing to an average accuracy of 55.87% at $\sigma = 400$. Indeed, not only were the classifier performances across all Gaussian noise conditions significantly lower than the human average performance, but this also extended to degradation as a proportion of clean performance. The accuracy at each distortion level, as a proportion of that classifiers performance on clean stimuli, was significantly lower than that of the human participants. Showing that significant differences between humans and classifiers across Gaussian distortion levels were not solely due to lower initial differences in performance before noise was introduced, but rather due to differing resilience to noise.

When Gaussian blur was introduced to the test stimuli, classifiers performance was reduced to 8-12% at the highest distortion level of $\sigma = 12$ whereas human average performance only decreased to 58.48%. Indeed, difference between classifier and human performance across all Gaussian blur conditions were significant. As in the Gaussian noise condition, we also found that the classification accuracy at each distortion level, as a proportion of that classifier's accuracy before distortion is introduced, was significantly lower than that of human participants.

The classifiers were further tested on salt and pepper noise, which reduced accuracy to slightly more than chance performance as the pixel reassignment rate increased beyond 50%, and contrast reduction, in which classifiers scored between 6-13% accuracy at the lowest contrast level (highest distortion level) except the VGGish+VGG-16+FRNN classifier which obtained 21.52% accuracy.

This vulnerability of DRCNNs to visual distortion follows from the finding that ImageNet trained CNNs are vulnerable to visual distortion on image recognition tasks (Dodge and Karam, 2016; Geirhos, Janssen, et al., 2017; Wichmann et al., 2017; Dodge and Karam, 2017; Geirhos, Temme, et al., 2018; Dodge and Karam, 2019). In light of recent evidence that increasing the shape bias vs texture bias of CNNs can reduce their vulnerability to image distortion (Geirhos, Michaelis, et al., 2019), we explore this line of literature for possible explanations. There is

a large body of evidence from the areas of psychology and neuroscience that human object recognition primarily uses shape features (Tanaka, 1996; Pasupathy and Connor, 2001; Quiroga et al., 2005; Wagemans et al., 2008). Given the similar performance of CNNs and humans on the ImageNet benchmark (Krizhevsky et al., 2012; Russakovsky et al., 2015; He et al., 2015) and that the ImageNet trained CNN representations are highly predictive of the human visual ventral stream (Yamins, Hong, Cadieu, and Dicarlo, 2013; Yamins, Hong, Cadieu, Solomon, et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014; Güçlü and Van Gerven, 2015; Cichy, Khosla, et al., 2016), it is understandable that researchers may expect that CNNs primarily use shape features. This idea that CNNs learn primitive object features/shapes and then hierarchically learn combinations of these features to perform inference (Lecun et al., 2015; Kriegeskorte, 2015) is termed the *shape hypothesis* by Geirhos, Michaelis, et al. (2019). Indeed, the shape hypothesis is supported by a number of findings including visualisations with Deconvolutional Networks that show hierarchies of object parts (Zeiler and Fergus, 2014). Kubilius et al. (2016) also added to evidence in the direction of the shape hypothesis, showing that CaffeNet (Caffe implementation of a simple CNN), VGG-19 (Simonyan and Zisserman, 2015) and GoogLeNet (Szegedy et al., 2015) can recognise objects based on shape cues alone (like humans) and that the deep layers show a high sensitivity for shape. This is despite being trained on ImageNet alone, without any explicit training to recognise shape. However, Kubilius et al. (2016) did find that these CNNs underperformed on a silhouette classification task (shape cues only) when compared to humans by approximately 20%.

More recent empirical evidence, however, suggests that ImageNet trained CNN models are biased towards texture rather than shape (Brendel and Bethge, 2019; Geirhos, Michaelis, et al., 2019). Brendel and Bethge (2019) introduced a model named BagNet that detects local features without considering their spatial ordering across the image (similar to bag-of-features; BoF) and yet still obtains competitive performance on ImageNet. This leads to the consideration that object recognition (ImageNet in particular) could be solved with local, texture information. This is the so-called *texture hypothesis*. In order to explore the validity of the texture hypothesis, Geirhos, Michaelis, et al. (2019) used style transfer (Gatys et al., 2016) to create stimuli with

shape/texture conflicts that could be used to quantify bias in CNNs. The results of the experiment supports the texture hypothesis, with VGG-16 (Simonyan and Zisserman, 2015), GoogLeNet (Szegedy et al., 2015), AlexNet (Krizhevsky et al., 2012) and ResNet-50 (He et al., 2016) largely classifying according to texture cues rather than shape cues and humans classifying largely according to shape cues (95.9% of correct decisions). To obtain a CNN classifier that would better leverage shape cues and not rely on texture, Geirhos, Michaelis, et al. (2019) further used style transfer to create 'Stylized-ImageNet', replacing the texture of all images with randomly selected painting. By optimising on this training set, ResNet-50 performed with a much stronger shape bias in the cue-conflict experiment, approximating human bias to a much higher degree. Further, the researchers showed that Stylized-ImageNet trained ResNet-50 outperforms ImageNet trained ResNet-50 on uniform noise, low contrast, high-pass filtering, phase noise and Eidolon distortions, better approximating human performance (although ImageNet-ResNet-50 obtained higher performances at low levels of distortion and the low-pass filtering condition). These findings have important implications for the robustness of CNN classifiers. Given these findings, further experiments could be carried out to pretrain the CNN components of DRCNN classifiers on Stylized-ImageNet and observe whether these models are better able to preserve their clean stimuli performance once visual distortions are introduced to frames.

6.1.6 Dual-stream RCNNs and humans: audiovisual perception as the visual signal gets weaker

The final question we study in this thesis is an extension of the previous question asked in the same chapter (Chapter 5); 'To what extent can dual-stream RCNNs leverage audio signals to preserve audiovisual performance when frame-level noise is introduced?'. This is a parallel line of research to that in the area of audiovisual speech recognition (Gabbay et al., 2018; Ephrat et al., 2018; Zhou, Yang, et al., 2019; Yu et al., 2020; Aldeneh et al., 2021) where researchers seek to leverage visual cues to overcome auditory noise and increase recognition rates.

The 6 dual-stream RCNN classifiers and human participants were again tested on Gaussian noise and Gaussian blur distortions, with classifiers further examined on salt and pepper noise

and contrast reduction. Although this time, visual stimuli were presented alongside clean audio stimuli (regardless of the visual noise condition). On clean, audiovisual stimuli, human increased their accuracy from an average of 93.80% on visual-only stimuli to 97.28% in the audiovisual case. A difference for which a significant effect was detected. Similarly, all dual-stream RCNN classifiers significantly increased their classification accuracy with the presence of audio stimuli. In contrast to the results in the visual-only domain, however, 5 of the 6 classifiers performed within the range of human performance in the clean audiovisual condition. The VGGish+VGG-16+FRNN classifier obtained a classification accuracy of 88.45%, where the lowest performing human participant scored 89.13%, although no significant difference was detected.

The classification accuracies obtained by the dual-stream RCNN classifiers were significantly higher than those obtain in the visual-only domain on all distortion types and levels. It also seems likely that for very small levels of Gaussian noise ($\sigma \leq 10$), the YamNet+EfficientNet-B0 based DRCNN classifiers obtained human levels of performance, given that classification accuracy was approximately equal to undistorted accuracy. Although we cannot confirm this as we measured humans on much coarser distortion levels due to their robustness to noise. For all other Gaussian noise levels and all Gaussian blur levels though, humans significantly outperformed all classifiers. Humans also significantly increased their classification accuracy with the addition of audio at every examined distortion level other than Gaussian noise at $\sigma = 100$, where the increase was too small for a significant effect to be detected.

In order to assess the extent to which the classifiers were able to leverage audio data, beyond an overall step increase in performance, we studied the performance decreases, relative to clean performance, at each distortion level. The decrease in classification accuracy was significantly larger in the visual domain than the audiovisual domain for all distortion types and levels. Thus increased performances were not only due to an overall increase in clean stimuli performance, but rather an increase in robustness as the dual-stream RCNN classifier more heavily utilised the audio information as distortion increased. Demonstrating the ability of these dual-stream RCNN classifiers as dynamic and robust audiovisual recognition systems.

6.2 Contributions, limitations and future directions

The work presented in this thesis provides some of the first steps in exploring deep neural networks as models of human audiovisual perception. The AVMIT and AVMIT-VEGAS datasets we presented provides 2 of the first large, labelled video sets of audiovisual action events, and the first for DNN, human comparison. A notable limitation of the training set, though, is that by holding such high inclusion criteria, the dataset may be too small to train very deep neural networks. In our studies, we utilised transfer learning, adding pretrained CNNs to our dual-stream RCNNs. Another strategy could be to use the original MIT (Monfort et al., 2019) dataset for pretraining, before further fine-tuning on AVMIT. Even if models are trained entirely on the original MIT dataset though, the AVMIT held-out test set still provides a highly controlled test for visual or audiovisual behaviour against human participants.

The dual-stream recurrent convolutional neural networks developed in this work utilised developments from the audio and visual recognition domains. Further the development of the multimodal squeeze units and RNNs provided a simple mechanism for those wishing to develop multisensory RNN models from individual unisensory systems. There are, however, plenty of advancements that can be made to create more robust models and further explore DNN and human perception on other research questions.

As DNNs operating in the audiovisual domain often require the use of visual sequences to complement the audio data (which is inherently temporal) there is a new source of visual information than in the image recognition domain; the information *between* frames. Optical flow, introduced by Gibson (2017) is the observable motion between frames, an approximation of 3D motion projected on to a 2D surface. An optical flow algorithm takes 2 frames as input and produce 'optical flow fields ' or 'flow images' that contain a single vector at each location across the frame (although in reality a single location can have multiple flows). DNNs could be used to produce flow images on the fly and another CNN could be added to the dual-stream RCNN to extract flow features.

Another advancement could include the use of stereophonic audio. Indeed, as the audio data in the AVMIT and AVMIT-VEGAS videos is monophonic, there is no available spatial

information that could be provided by stereophonic audio. With a DNN capable of localising in both the visual and auditory domain, one could examine any ventriloquist effect that may present itself (Alais and Burr, 2004).

One clear next step in the research would be to build more dual-stream RCNN, utilising other CNN models, and test on more distortion types/levels against human participants. These distortion investigations could extend into the audio domain, which may further our understanding of the relationship between distortion type and CNN/RNN/RCNN behaviour. Pairing these distortions with incongruent stimuli may also allow researchers to better understand the relationship between modality reliability and that modality's influence on classification.

One could also measure the response times of the dual-stream RCNNs developed in this work using the entropy thresholding methods in Spoerer, Kietzmann, et al. (2020). This would allow researchers to assess our dual-stream RCNNs further on distortions or within/without audio or visual data and assess how these factors affect response times. This behavioural data could then be compared to human response times on the same stimuli.

At this point in time, large, labelled datasets and supervised learning are used to produce the DNN classifiers for use as models of human perception. Whilst this allows for very particular research questions centred around specific tasks, ultimately, humans do not learn from large, labelled datasets. Although some unknown portion of human behaviour may be hard-coded in the genome, humans learn from embodied experience. The use of naturalistic simulations and ecologically relevant tasks alongside using embodied agents to train DNNs (such as in the realm of deep reinforcement learning; Mnih et al., 2015) could be used to build more realistic models of perception. Examining the differences between the learnt behaviour of supervised and unsupervised techniques certainly deserves more research attention (Khaligh-Razavi and Kriegeskorte, 2014).

6.3 Conclusions

In conclusion, the work described in this thesis provided a large, labelled training set and held-out test set of audiovisual action events for human vs DNN investigations and provided a series of dual-stream RCNN models/classifiers. The work further studied the learnt behaviours of these DRCNN classifiers, finding that they were all capable of solving the audiovisual correspondence task after only being optimised on an audiovisual action recognition task. Further, the behaviour of each classifier on unisensory classification tasks was highly affected by signals present in unattended to modalities, indicating that signals were highly integrated.

Lastly, we found that another set of dual-stream RCNN classifiers were vulnerable to visual distortion, with significant differences in behaviour to that of human participants. However, the addition of clean audio data allowed the classifiers to both increase classification accuracy, and reduce performance degradation as distortion level increased. The dual-stream RCNN classifiers reached human levels of performance on clean audiovisual stimuli and low levels of Gaussian noise, but on higher levels of noise or under Gaussian blur distortions, performance was significantly lower than that of human participants.

References

McCulloch, Warren S and Walter Pitts (1943). "A logical calculus of the ideas immanent in nervous activity". In: *The bulletin of mathematical biophysics* 5.4, pp. 115–133. ISSN: 1522-9602. DOI: 10.1007/BF02478259. URL: https://doi.org/10.1007/BF02478259.

Cherry, E. Colin (1953). "Some Experiments on the Recognition of Speech, with One and with Two Ears". In: *Journal of the Acoustical Society of America* 25.5, pp. 975–979. ISSN: NA. DOI: 10.1121/1.1907229.

Rosenblatt, Frank (1957). *The Perceptron - A Perceiving and Recognizing Automaton*.

Hershenson, Maurice (1962). "Reaction time as a measure of intersensory facilitation". In: *Journal of Experimental Psychology* 63.3, pp. 289–293.

Raab, David H (1962). "Statistical facilitation of simple reaction times". In: *Transactions of the New York Academy of Sciences* 24, pp. 574–590.

Morrell, Lenore (1967). "Intersensory facilitation of reaction time". In: *Psychonomic Science* 8.2, pp. 77–78. ISSN: 00333131. DOI: 10.3758/BF03330675.

Potter, M. C. (1976). "Short-term conceptual memory for pictures." In: *Journal of experimental psychology. Human learning and memory* 2.5, pp. 509–522. ISSN: 00961515. DOI: 10.1037/0278-7393.2.5.509.

Everitt, Brian (1977). *The Analysis of Contingency Tables*. London: Chapman and Hall.

Welch, Robert B. and David H. Warren (1980). "Immediate perceptual response to intersensory discrepancy". In: *Psychological Bulletin* 88.3, pp. 638–667. ISSN: 00332909. DOI: 10.1037/0033-2909.88.3.638.

Miller, Jeff (1982). *Divided Attention: Evidence for Coactivation with Redundant Signals*. Tech. rep., pp. 247–279.

Gielen, Stan C.A.M., Richard A. Schmidt, and Pieter J.M. Van Den Heuvel (1983). "On the nature of intersensory facilitation of reaction time". In: *Perception & Psychophysics* 34.2, pp. 161–168. ISSN: 00016918. DOI: 10.1016/0001-6918(84)90040-4.

Efron, B and R Tibshirani (1986). "Bootstrap Methods for Standard Errors and Confidence Intervals, and Other Measures of Statistical Accuracy". In: *Statistical Science* 1.1, pp. 54–77.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams (1986). "Learning representations by back-propagating errors". In: *Nature* 323.6088, pp. 533–536.

DeYoe, E A and D C Van Essen (1988). "Concurrent processing streams in monkey visual cortex". In: *Trends in Neurosciences* 11.5, pp. 219–226. ISSN: 0166-2236. DOI: https://doi.org/10.1016/0166-2236(88)90130-0. URL: https://www.sciencedirect.com/science/article/pii/0166223688901300.

Cybenko, G (1989). "Approximations by superpositions of sigmoidal functions". In: *Mathematics of Control, Signals, and Systems* 2.4, pp. 303–314. ISSN: 10009221. DOI: 10.1007/BF02836480.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White (1989). "Multilayer feedforward networks are universal approximators". In: *Neural Networks* 2.5, pp. 359–366. ISSN: 08936080. DOI: 10.1016/0893-6080(89)90020-8.

LeCun, Y, B Boser, et al. (Aug. 1989). "Backpropagation Applied to Handwritten Zip Code Recognition". In: *Neural Computation* 1.4, pp. 541–551.

Baddeley, Alan (1992). "Working Memory Components of Working Memory The Slave Systems of Working Memory Individual Differences in Working Memory". In: *Science* 255.ii, pp. 556–559.

Paul, Douglas B. and Janet M. Baker (1992). "The design for the wall street journal-based CSR corpus". In: *Workshop on Speech and Natural Language (Association for Computational Linguistics)*, pp. 357–362. DOI: 10.3115/1075527.1075614.

Garofolo, John S. et al. (1993). *TIMIT: Acoustic-Phonetic Continuous Speech Corpus*.

Stein, B E and M A Meredith (1993). *The Merging of the Senses*. Cambridge, MA, USA: The MIT Press.

Diederich, Adele (1995). "Intersensory Facilitation of Reaction Time: Evaluation of Counter and Diffusion Coactivation Models". In: *Journal of Mathematical Psychology* 39.2, pp. 197–215.

Townsend, James T and Georgie Nozawa (1995). "Spatio-temporal Properties of Elementary Perception: An Investigation of Parallel, Serial, and Coactive Theories". In: *Journal of Mathematical Psychology* 39.4, pp. 321–359. ISSN: 10960880. DOI: 10.1006/jmps.1995.1033.

Van Rossum, Guido and Fred L Drake Jr (1995). *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam.

Munhall, K G et al. (1996). "Temporal constraints on the McGurk effect We thank L. Coady for testing subjects in Experiment 3 and 1. Jones for doing the video analysis in Experiment 2. Correspondence should be ad- dressed to". In: *Perception & Psychophysics* 58.3, pp. 351–362. URL: https://link.springer.com/content/pdf/10.3758/BF03206811.pdf%0Ahttp://download.springer.com/static/pdf/524/art%253A10.3758%252FBF03206811.pdf?originUrl=http%25253A%25252F%25252Flink.springer.com%25252Farticle%25252F10.3758%25252FBF03206811&token2=exp=1488.

Tanaka, Keiji (1996). "Inferotemporal Cortex and Object Vision". In: *Annual Review of Neuroscience* 19.1, pp. 109–139. DOI: 10.1146/annurev.ne.19.030196.000545. URL: https://doi.org/10.1146/annurev.ne.19.030196.000545.

Thorpe, Simon, Denis Fize, and Catherine Marlot (1996). "Speed of processing in the human visual system". In: *Nature* 381.6582, pp. 520–522. ISSN: 00280836. DOI: 10.1038/381520a0.

Hochreiter, Sepp and Jürgen Schmidhuber (Nov. 1997). "Long Short-Term Memory". In: *Neural Computation* 9.8, pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735. URL: http://dx.doi.org/10.1162/neco.1997.9.8.1735.

Dietterich, T. (1998). "Approximate statistical tests for comparing supervised classification learning algorithms." In: *Neural Computation* 10.7, pp. 1895–1924.

LeCun, Yann, Léon Bottou, et al. (1998). "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11, pp. 2278–2323. ISSN: 00189219. DOI: 10.1109/5.726791.

Perneger, Thomas V. (1998). "What's wrong with Bonferroni adjustment". In: *British Medical Journal* 316.7139, pp. 1236–1238. ISSN: 09598146. DOI: 10.1136/bmj.316.7139.1230.

Giard, M H and F Peronnet (1999). *Auditory-Visual Integration during Multimodal Object Recognition in Humans: A Behavioral and Electrophysiological Study*. Tech. rep.

Lamme, Victor A.F., Valia Rodriguez-Rodriguez, and Henk Spekreijse (1999). "Separate processing dynamics for texture elements, boundaries and surfaces in primary visual cortex of the Macaque monkey". In: *Cerebral Cortex* 9.4, pp. 406–413. ISSN: 10473211. DOI: 10.1093/cercor/9.4.406.

Riesenhuber, Maximilian and Tomaso Poggio (1999). "Hierarchical models of object recognition in cortex". In: *Nature Neuroscience* 2.11, pp. 1019–1025. ISSN: 10976256. DOI: 10.1038/14819.

Bradski, G (2000). "The OpenCV Library". In: *Dr. Dobb's Journal of Software Tools*.

Calvert, Gemma A., Ruth Campbell, and Michael J. Brammer (2000). "Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex". In: *Current Biology* 10.11, pp. 649–657. ISSN: 09609822. DOI: 10.1016/S0960-9822(00)00513-3.

Pasupathy, A and C E Connor (Nov. 2001). "Shape representation in area V4: position-specific tuning for boundary conformation." eng. In: *Journal of neurophysiology* 86.5, pp. 2505–2519. ISSN: 0022-3077 (Print). DOI: 10.1152/jn.2001.86.5.2505.

Slutsky, Daniel A. and Gregg H. Recanzone (2001). "Temporal and spatial dependency, of the ventriloquism effect". In: *NeuroReport* 12.1, pp. 7–10. ISSN: 09594965. DOI: 10.1097/00001756-200101220-00009.

Ernst, Marc O and Martin S Banks (2002). "Humans integrate visual and haptic information in a statistically optimal fashion". In: *Nature* 415.January, pp. 429–433.

Foxe, John J. et al. (2002). "Auditory-somatosensory multisensory processing in auditory association cortex: An fMRI study". In: *Journal of Neurophysiology* 88.1, pp. 540–543. ISSN: 00223077. DOI: 10.1152/jn.2002.88.1.540.

Hsing, Tailen, Sanju Attoor, and Edward R. Dougherty (2003). "Relation Between Permutation-Test P Values and Classifier Error Estimates". In: *Machine Learning* 52.1-2, pp. 11–30. ISSN: 00138703. DOI: 10.1023/A.

Lewald, Jörg and Rainer Guski (2003). "Cross-modal perceptual integration of spatially and temporally disparate auditory and visual stimuli". In: *Cognitive Brain Research* 16.3, pp. 468–478. ISSN: 09266410. DOI: 10.1016/S0926-6410(03)00074-0.

Macaluso, Emiliano, Jon Driver, and Chris D. Frith (2003). "Multimodal Spatial Representations Engaged in Human Parietal Cortex during Both Saccadic and Manual Spatial Orienting". In: *Current Biology* 13, pp. 990–999. DOI: 10.1016/S.

Wright, Tarra M. et al. (2003). "Polysensory interactions along lateral temporal regions evoked by audiovisual speech". In: *Cerebral Cortex* 13.10, pp. 1034–1043. ISSN: 10473211. DOI: 10.1093/cercor/13.10.1034.

Alais, David and David Burr (Feb. 2004). "The Ventriloquist Effect Results from Near-Optimal Bimodal Integration". In: *Current Biology* 14.3, pp. 257–262. ISSN: 09609822. DOI: 10.1016/j.cub.2004.01.029.

Barrouillet, Pierre, Sophie Bernardin, and Valérie Camos (2004). "Time Constraints and Resource Sharing in Adults' Working Memory Spans". In: *Journal of Experimental Psychology: General* 133.1, pp. 83–100. ISSN: 00963445. DOI: 10.1037/0096-3445.133.1.83.

Diederich, Adele and Hans Colonius (2004). "Bimodal and trimodal multisensory enhancement: Effects of stimulus onset and intensity on reaction time". In: *Perception and Psychophysics* 66.8, pp. 1388–1404. ISSN: 00315117. DOI: 10.3758/BF03195006.

Laurienti, Paul J. et al. (2004). "Semantic congruence is a critical factor in multisensory behavioral performance". In: *Experimental Brain Research* 158.4, pp. 405–414. ISSN: 00144819. DOI: 10.1007/s00221-004-1913-2.

Macaluso, E, N George, et al. (2004). "Spatial and temporal factors during processing of audiovisual speech: a PET study". In: *NeuroImage* 21.2, pp. 725–732. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2003.09.049. URL: https://www.sciencedirect.com/science/article/pii/S1053811903005998.

Miller, Jeff (2004). “Exaggerated redundancy gain in the split brain: A hemispheric coactivation account”. In: *Cognitive Psychology* 49.2, pp. 118–154. ISSN: 00100285. DOI: 10.1016/j.cogpsych.2003.12.003.

Molholm, Sophie et al. (2004). “Multisensory Visual-Auditory Object Recognition in Humans: A High-density Electrical Mapping Study”. In: *Cerebral Cortex* 14.4, pp. 452–465. ISSN: 10473211. DOI: 10.1093/cercor/bhh007.

Wallace, M. T. et al. (2004). “Unifying multisensory signals across time and space”. In: *Experimental Brain Research* 158.2, pp. 252–258. ISSN: 00144819. DOI: 10.1007/s00221-004-1899-9.

Amedi, A. et al. (2005). “Functional imaging of human crossmodal identification and object recognition”. In: *Experimental Brain Research* 166.3-4, pp. 559–571. ISSN: 00144819. DOI: 10.1007/s00221-005-2396-5.

Barraclough, Nick E. et al. (2005). “Integration of visual and auditory information by superior temporal sulcus neurons responsive to the sight of actions”. In: *Journal of Cognitive Neuroscience* 17.3, pp. 377–391. ISSN: 0898929X. DOI: 10.1162/0898929053279586.

Golland, Polina et al. (2005). “Permutation tests for classification”. In: *Annual Conference on Learning Theory*, pp. 501–515. ISBN: 9781538631591. DOI: 10.1109/PRNI.2017.7981495.

Johnson, Jeffrey S. and Bruno A. Olshausen (2005). “The recognition of partially visible natural objects in the presence and absence of their occluders”. In: *Vision Research* 45.25-26, pp. 3262–3276. ISSN: 00426989. DOI: 10.1016/j.visres.2005.06.007.

Quiroga, R Quian et al. (June 2005). “Invariant visual representation by single neurons in the human brain.” eng. In: *Nature* 435.7045, pp. 1102–1107. ISSN: 1476-4687 (Electronic). DOI: 10.1038/nature03687.

Stefanics, Gábor et al. (2005). “Cross-modal visual-auditory-somatosensory integration in a multimodal object recognition task in humans”. In: *International Congress Series* 1278, pp. 163–166. ISSN: 05315131. DOI: 10.1016/j.ics.2004.11.074.

Brungart, Douglas S. and Brian D. Simpson (2007). "Cocktail party listening in a dynamic multitalker environment". In: *Perception and Psychophysics* 69.1, pp. 79–91. ISSN: 00315117. DOI: 10.3758/BF03194455.

Craft, Edward et al. (June 2007). "A neural model of figure-ground organization." eng. In: *Journal of neurophysiology* 97.6, pp. 4310–4326. ISSN: 0022-3077 (Print). DOI: 10.1152/jn.00203.2007.

Hunter, J D (2007). "Matplotlib: A 2D graphics environment". In: *Computing in Science \& Engineering* 9.3, pp. 90–95. DOI: 10.1109/MCSE.2007.55.

Körding, Konrad P. et al. (2007). "Causal inference in multisensory perception". In: *PLoS ONE* 2.9, pp. 1–10. ISSN: 19326203. DOI: 10.1371/journal.pone.0000943.

Yuval-Greenberg, Shlomit and Leon Y. Deouell (2007). "What you see is not (always) what you hear: Induced gamma band responses reflect cross-modal interactions in familiar object recognition". In: *Journal of Neuroscience* 27.5, pp. 1090–1096. ISSN: 02706474. DOI: 10.1523/JNEUROSCI.4828-06.2007.

Doehrmann, Oliver and Marcus J. Naumer (2008). "Semantics and the multisensory brain: How meaning modulates processes of audio-visual integration". In: *Brain Research* 1242, pp. 136–150. ISSN: 00068993. DOI: 10.1016/j.brainres.2008.03.071. URL: http://dx.doi.org/10.1016/j.brainres.2008.03.071.

Wagemans, Johan et al. (2008). "Identification of everyday objects on the basis of silhouette and outline versions." eng. In: *Perception* 37.2, pp. 207–244. ISSN: 0301-0066 (Print). DOI: 10.1068/p5825.

Deng, Jia et al. (2009). "ImageNet: A large-scale hierarchical image database". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 248–255. ISBN: 9781424439911. DOI: 10.1109/cvprw.2009.5206848.

Parise, Cesare Valerio and Charles Spence (2009). "'When birds of a feather flock together': Synesthetic correspondences modulate audiovisual integration in non-synesthetes". In: *PLoS ONE* 4.5. ISSN: 19326203. DOI: 10.1371/journal.pone.0005664.

Stevenson, Ryan A. and Thomas W. James (Feb. 2009). "Audiovisual integration in human superior temporal sulcus: Inverse effectiveness and the neural processing of speech and object recognition". In: *NeuroImage* 44.3, pp. 1210–1223. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2008.09.034.

Van Rossum, Guido and Fred L Drake (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace. ISBN: 1441412697.

Atrey, Pradeep K. et al. (Nov. 2010). "Multimodal fusion for multimedia analysis: A survey". In: *Multimedia Systems* 16.6, pp. 345–379. ISSN: 09424962. DOI: 10.1007/s00530-010-0182-0.

Noppeney, Uta, Dirk Ostwald, and Sebastian Werner (May 2010). "Perceptual decisions formed by accumulation of audiovisual evidence in prefrontal cortex". In: *Journal of Neuroscience* 30.21, pp. 7434–7446. ISSN: 02706474. DOI: 10.1523/JNEUROSCI.0455-10.2010.

Ojala, Markus and Gemma C. Garriga (2010). "Permutation tests for studying classifier performance". In: *Journal of Machine Learning Research* 11, pp. 1833–1863. ISSN: 15324435.

Pan, Sinno Jialin and Qiang Yang (2010). "A Survey on Transfer Learning". In: *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*. Vol. 22. 10. IEEE, pp. 1345–1359. ISBN: 9789811559709. DOI: 10.1007/978-981-15-5971-6{_}83.

Shams, Ladan and Ulrik R. Beierholm (2010). "Causal inference in perception". In: *Trends in Cognitive Sciences* 14.9, pp. 425–432. ISSN: 13646613. DOI: 10.1016/j.tics.2010.07.001. URL: http://dx.doi.org/10.1016/j.tics.2010.07.001.

Werner, Sebastian and Uta Noppeney (2010). "Superadditive responses in superior temporal sulcus predict audiovisual benefits in object categorization". In: *Cerebral Cortex*. ISSN: 10473211. DOI: 10.1093/cercor/bhp248.

Bertin-Mahieux, Thierry et al. (2011). "The million song dataset". In: *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)* July, pp. 591–596.

Bursztein, Elie, Matthieu Martin, and John Mitchell (2011). "Text-Based CAPTCHA Strengths and Weaknesses". In: *Proceedings of the 18th ACM Conference on Computer and Communications Security*. CCS '11. New York, NY, USA: Association for Computing Machinery,

pp. 125–138. ISBN: 9781450309486. DOI: 10.1145/2046707.2046724. URL: https://doi.org/10.1145/2046707.2046724.

Fleuret, François et al. (2011). "Comparing machines and humans on a visual categorization test". In: *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* 108.43, pp. 17621–17625. ISSN: 00278424. DOI: 10.1073/pnas.1109168108.

Lee, Hwee Ling and Uta Noppeney (2011). "Long-term music training tunes how the brain temporally binds signals from multiple senses". In: *Proceedings of the National Academy of Sciences of the United States of America* 108.51. ISSN: 00278424. DOI: 10.1073/pnas.1115267108.

Ngiam, Jiquan et al. (2011). "Multimodal Deep Learning". In: *International Conference of Machine Learning (ICML)*.

Pedregosa, F et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Werner, Sebastian and Uta Noppeney (2011). "The contributions of transient and sustained response codes to audiovisual integration". In: *Cerebral Cortex* 21.4, pp. 920–931. ISSN: 10473211. DOI: 10.1093/cercor/bhq161.

Carandini, Matteo and David J. Heeger (2012). "Normalization as a canonical neural computation". In: *Nature Reviews Neuroscience* 13.1, pp. 51–62. ISSN: 1471003X. DOI: 10.1038/nrn3136.

Cireşan, Dan, Ueli Meier, and Jurgen Schmidhuber (2012). "Multi-column deep neural networks for image classification". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. February, pp. 3642–3649. ISBN: 9781467312264. DOI: 10.1109/CVPR.2012.6248110.

Crowston, Kevin (2012). "Amazon mechanical turk: A research tool for organizations and information systems scholars". In: *IFIP Advances in Information and Communication Technology* 389 AICT, pp. 210–221. ISSN: 18684238. DOI: 10.1007/978-3-642-35142-6{_}14.

DiCarlo, James J., Davide Zoccolan, and Nicole C. Rust (Feb. 2012). "How does the brain solve visual object recognition?" In: *Neuron* 73.3, pp. 415–434. ISSN: 08966273. DOI: 10.1016/j.neuron.2012.01.010.

Hinton, Geoffrey et al. (2012). "Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups". In: *IEEE Signal Processing Magazine* 29, pp. 82–97. ISSN: 10535888. DOI: 10.1109/MSP.2012.2209906.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems (NIPS)*. Nevada, pp. 1097–1105.

Stein, Barry E. (2012). *The New Handbook of Multisensory Processing*. 1st ed. The MIT Press. ISBN: 0262017121.

Wyatte, Dean, Tim Curran, and Randall O'Reilly (2012). "The Limits of Feedforward Vision: Recurrent Processing Promotes Robust Object Recognition when Objects Are Degraded". In: *Journal of Cognitive Neuroscience*, pp. 2248–2261. DOI: 10.1162/jocn. URL: https://www.apa.org/ptsd-guideline/ptsd.pdf%0Ahttps://www.apa.org/about/offices/directorates/guidelines/ptsd.pdf.

O'Reilly, Randall C. et al. (2013). "Recurrent processing during object recognition". In: *Frontiers in Psychology* 4.APR, pp. 1–14. ISSN: 16641078. DOI: 10.3389/fpsyg.2013.00124.

Wan, Li et al. (2013). "Regularization of Neural Networks using DropConnect". In: *International Conference on Machine Learning (ICML)*, pp. 1058–1066. DOI: 10.1109/TPAMI.2017.2703082.

Yamins, Daniel, Ha Hong, Charles Cadieu, and James J. Dicarlo (2013). "Hierarchical modular optimization of convolutional networks achieves representations similar to Macaque IT and human ventral stream". In: *Advances in Neural Information Processing Systems (NIPS)*, pp. 1–9.

Borji, Ali and Laurent Itti (2014). "Human vs. computer in scene and object recognition". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 113–120. URL: http://ilab.usc.edu/borji/.

Cadieu, Charles F. et al. (Dec. 2014). “Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition”. In: *PLoS Computational Biology* 10.12. ISSN: 15537358. DOI: 10.1371/journal.pcbi.1003963.

Cho, Kyunghyun, Bart Van Merrienboer, and Dzmitry Bahdanau (2014). “On the Properties of Neural Machine Translation: Encoder–Decoder Approaches”. In: *Conference on Empirical Methods in Natural Language Processing*. Vol. 1, pp. 103–111. ISBN: 9781601321916.

Goodfellow, Ian J, Jean Pouget-Abadie, et al. (2014). “Generative adversarial nets”. In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Goodfellow, Ian J., Yaroslav Bulatov, et al. (2014). “Multi-digit number recognition from street view imagery using deep convolutional neural networks”. In: *2nd International Conference on Learning Representations, ICLR 2014*. December.

Karpathy, Andrej et al. (2014). “Large-scale video classification with convolutional neural networks”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1725–1732. ISBN: 9781479951178. DOI: 10.1109/CVPR.2014.223.

Khaligh-Razavi, Seyed Mahdi and Nikolaus Kriegeskorte (2014). “Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation”. In: *PLoS Computational Biology* 10.11. ISSN: 15537358. DOI: 10.1371/journal.pcbi.1003915.

Lee, Hweeling and Uta Noppeney (2014). “Temporal prediction errors in visual and auditory cortices”. In: *Current Biology* 24.8, R309–R310. ISSN: 0960-9822. DOI: https://doi.org/10.1016/j.cub.2014.02.007. URL: https://www.sciencedirect.com/science/article/pii/S0960982214001456.

Leo, Fabrizio and Uta Noppeney (Sept. 2014). “Conditioned sounds enhance visual processing”. In: *PLoS ONE* 9.9. ISSN: 19326203. DOI: 10.1371/journal.pone.0106860.

Lin, Tsung Yi et al. (2014). “Microsoft COCO: Common objects in context”. In: *European Conference on Computer Vision (ECCV)*, pp. 740–755. DOI: 10.1007/978-3-319-10602-1{_}48.

Prolific (2014). URL: https://www.prolific.co/.

Sun, Yi, Yuheng Chen, et al. (2014). “Deep learning face representation by joint identification-verification”. In: *Advances in Neural Information Processing Systems* 3.January, pp. 1988–1996. ISSN: 10495258.

Taigman, Yaniv et al. (2014). “DeepFace: Closing the gap to human-level performance in face verification”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 1701–1708. ISBN: 9781479951178. DOI: 10.1109/CVPR.2014.220.

Wyatte, Dean, David J. Jilk, and Randall C. O’Reilly (2014). “Early recurrent feedback facilitates visual object recognition under challenging conditions”. In: *Frontiers in Psychology* 5.JUL, pp. 1–10. ISSN: 16641078. DOI: 10.3389/fpsyg.2014.00674.

Yamins, Daniel L.K., Ha Hong, Charles F. Cadieu, Ethan A. Solomon, et al. (2014). “Performance-optimized hierarchical models predict neural responses in higher visual cortex”. In: *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* 111.23, pp. 8619–8624. ISSN: 10916490. DOI: 10.1073/pnas.1403112111.

Zeiler, Matthew D. and Rob Fergus (2014). “Visualizing and understanding convolutional networks”. In: *European Conference on Computer Vision (ECCV)*, pp. 818–833. ISBN: 9783319105895. DOI: 10.1007/978-3-319-10590-1{_}53.

Abadi, Martín et al. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. URL: https://www.tensorflow.org/.

Bronkhorst, Adelbert W. (2015). “The cocktail-party problem revisited: early processing and selection of multi-talker speech”. In: *Attention, Perception, and Psychophysics* 77.5, pp. 1465–1487. ISSN: 1943393X. DOI: 10.3758/s13414-015-0882-9.

Donahue, Jeff et al. (2015). “Long-term Recurrent Convolutional Networks for Visual Recognition and Description”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2625–2634.

Güçlü, Umut and Marcel A J Van Gerven (2015). “Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream”. In: *Journal of Neuroscience*. DOI: https://doi.org/10.1523/JNEUROSCI.5023-14.2015. URL: http://dx.doi.org/10.6080/.

He, Kaiming et al. (2015). “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification”. In: *Proceedings of the IEEE International Conference on Computer Vision*. Vol. 2015 Inter, pp. 1026–1034. ISBN: 9781467383912. DOI: 10.1109/ICCV.2015.123.

Heilbron, Fabian Caba et al. (2015). “ActivityNet: A large-scale video benchmark for human activity understanding”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 961–970. ISBN: 9781467369640. DOI: 10.1109/CVPR.2015.7298698.

Ioffe, Sergey and Christian Szegedy (2015). “Batch normalization: Accelerating deep network training by reducing internal covariate shift”. In: *32nd International Conference on Machine Learning (ICML)*. Vol. 37, pp. 448–456. ISBN: 9781510810587.

Kingma, Diederik P. and Jimmy Lei Ba (2015). “Adam: A method for stochastic optimization”. In: *3rd International Conference on Learning Representations, ICLR*, pp. 1–15.

Kriegeskorte, Nikolaus (2015). “Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing”. In: *Annual Review of Vision Science* 1.1, pp. 417–446. ISSN: 2374-4642. DOI: 10.1146/annurev-vision-082114-035447.

Lecun, Yann, Yoshua Bengio, and Geoffrey Hinton (May 2015). “Deep learning”. In: *Nature* 521.7553, pp. 436–444. ISSN: 14764687. DOI: 10.1038/nature14539.

Liang, Ming and Xiaolin Hu (2015). “Recurrent convolutional neural network for object recognition”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 07-12-June. Figure 1, pp. 3367–3375. ISBN: 9781467369640. DOI: 10.1109/CVPR.2015.7298958.

McLoughlin, Ian et al. (2015). “Robust Sound Event Classification Using Deep Neural Networks”. In: *IEEE Transactions on Audio, Speech and Language Processing*. Vol. 23. 3. IEEE, pp. 229–245. DOI: 10.1201/9781003226277-9.

Mnih, Volodymyr et al. (Feb. 2015). “Human-level control through deep reinforcement learning”. In: *Nature* 518.7540, pp. 529–533. DOI: 10.1038/nature14236.

Russakovsky, Olga et al. (Dec. 2015). “ImageNet Large Scale Visual Recognition Challenge”. In: *International Journal of Computer Vision* 115.3, pp. 211–252. ISSN: 15731405. DOI: 10.1007/s11263-015-0816-y.

Shi, Xingjian et al. (2015). “Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting”. In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1–9. DOI: 10.1155/2018/6184713.

Simonyan, Karen and Andrew Zisserman (Sept. 2015). “Very Deep Convolutional Networks for Large-Scale Image Recognition”. In: *International Conference of Learning Representations (ICLR)*. URL: http://arxiv.org/abs/1409.1556.

Szegedy, Christian et al. (2015). “Going deeper with convolutions”. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 07-12-June. IEEE, pp. 1–9. ISBN: 9781467369640. DOI: 10.1109/CVPR.2015.7298594.

Zhang, Haomin, Ian McLoughlin, and Yan Song (2015). “ROBUST SOUND EVENT RECOGNITION USING CONVOLUTIONAL NEURAL NETWORKS”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 559–563. ISBN: 9781467369978. URL: https://ieeexplore.ieee.org/abstract/document/7178031/.

Abu-El-Haija, Sami et al. (2016). “YouTube-8M: A Large-Scale Video Classification Benchmark”.

Cichy, Radoslaw Martin, Aditya Khosla, et al. (2016). “Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence”. In: *Scientific Reports*. DOI: 10.1038/srep27755.

Dodge, Samuel and Lina Karam (Apr. 2016). “Understanding How Image Quality Affects Deep Neural Networks”. In: *Proceedings of the 2016 8th International Conference on Quality of Multimedia Experience (QoMEX’16)*. Pp. 1–6. URL: http://arxiv.org/abs/1604.04004.

Gatys, Leon A, Alexander S Ecker, and Matthias Bethge (2016). "Image Style Transfer Using Convolutional Neural Networks". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2414–2423.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). "6.5 Back-Propagation and Other Differentiation Algorithms". In: *Deep Learning*. MIT Press, pp. 200–220. ISBN: 9780262035613.

He, Kaiming et al. (Dec. 2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol. 2016-Decem. IEEE Computer Society, pp. 770–778. ISBN: 9781467388504. DOI: 10.1109/CVPR.2016.90.

Krugliak, Alexandra and Uta Noppeney (2016). "Synaesthetic interactions across vision and audition". In: *Neuropsychologia* 88, pp. 65–73. ISSN: 18733514. DOI: 10.1016/j.neuropsychologia.2015.09.027. URL: http://dx.doi.org/10.1016/j.neuropsychologia.2015.09.027.

Kubilius, Jonas, Stefania Bracci, and Hans P. Op de Beeck (2016). "Deep Neural Networks as a Computational Model for Human Shape Sensitivity". In: *PLoS Computational Biology* 12.4, pp. 1–26. ISSN: 15537358. DOI: 10.1371/journal.pcbi.1004896.

Laffitte, Pierre et al. (2016). "Deep neural networks for automatic detection of screams and shouted speech in subway trains". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 2016-May. IEEE, pp. 6460–6464. ISBN: 9781479999880. DOI: 10.1109/ICASSP.2016.7472921.

McFee, Brian (2016). "resampy: efficient sample rate conversion in Python". In: *Journal of Open Source Software* 1.8, p. 125. DOI: 10.21105/joss.00125. URL: https://doi.org/10.21105/joss.00125.

Miller, Jeff (2016). "Statistical facilitation and the redundant signals effect: What are race and coactivation models?" In: *Attention, Perception, and Psychophysics* 78.2, pp. 516–519. ISSN: 1943393X. DOI: 10.3758/s13414-015-1017-z.

Parascandolo, Giambattista, Heikki Huttunen, and Tuomas Virtanen (2016). "Recurrent neural networks for polyphonic sound event detection in real life recordings". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6440–6444. ISBN: 9781479999880. DOI: 10.1109/ICASSP.2016.7472917.

Phan, Huy et al. (2016). "Robust audio event recognition with 1-max pooling convolutional neural networks". In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 3653–3657. DOI: 10.21437/Interspeech.2016-123.

Seibert, Darren et al. (2016). "A performance-optimized model of neural responses across the ventral visual stream." In: *bioRxiv*. DOI: 10.1101/036475.

Stabinger, Sebastian, Antonio Rodríguez-Sánchez, and Justus Piater (2016). "25 years of CNNS: Can we compare to human abstraction capabilities?" In: *International Conference on Artificial Neural Networks*, pp. 380–387. ISBN: 9783319447803. DOI: 10.1007/978-3-319-44781-0{_}45.

Takahashi, Naoya et al. (2016). "Deep convolutional neural networks and data augmentation for acoustic event recognition". In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 2982–2986. DOI: 10.21437/Interspeech.2016-805.

Tsironi, Eleni, Pablo Barros, and Stefan Wermter (2016). "Gesture Recognition with a Convolutional Long Short-Term Memory Recurrent Neural Network". In: *ESANN 2016 - 24th European Symposium on Artificial Neural Networks*. ISBN: 9782875870278.

Ullman, Shimon et al. (2016). "Atoms of recognition in human and computer vision". In: *Proceedings of the National Academy of Sciences of the United States of America* 113.10, pp. 2744–2749. ISSN: 10916490. DOI: 10.1073/pnas.1513198113.

Wang, Cheng, Haojin Yang, and Christoph Meinel (2016). "Exploring multimodal video representation for action recognition". In: *Proceedings of the International Joint Conference on Neural Networks*. Vol. 2016-Octob. 1. IEEE, pp. 1924–1931. ISBN: 9781509006199. DOI: 10.1109/IJCNN.2016.7727435.

Arandjelovic, Relja and Andrew Zisserman (2017). “Look, Listen and Learn”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. ISBN: 9781538610329. DOI: 10.1109/ICCV.2017.73.

Çakır, Emre et al. (Feb. 2017). “Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection”. In: *IEEE/ACM Transactions on Audio, Speech and Language Processing* 25.6, pp. 1291–1303. DOI: 10.1109/TASLP.2017.2690575. URL: http://arxiv.org/abs/1702.06286%20http://dx.doi.org/10.1109/TASLP.2017.2690575.

Carreira, Joao and Andrew Zisserman (May 2017). “Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset”. In: URL: http://arxiv.org/abs/1705.07750.

Dodge, Samuel and Lina Karam (2017). “A study and comparison of human and deep learning recognition performance under visual distortions”. In: *26th International Conference on Computer Communications and Networks, (ICCCN)*. ISBN: 9781509029914. DOI: 10.1109/ICCCN.2017.8038465.

Geirhos, Robert, David H. J. Janssen, et al. (2017). “Comparing deep neural networks against humans: object recognition when the signal gets weaker”. In: *arXiv*.

Gemmeke, Jort F et al. (2017). “Audio Set: An ontology and human-labeled dataset for audio events”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. URL: http://en.wikipedia.org/wiki/Bird.

George, D et al. (2017). “A generative vision model that trains with high data efficiency and breaks text-based CAPTCHAs”. In: *Science* 2612.October, pp. 1–19.

Gibson, James J . (2017). “The Perception of the Visual World”. In: *Audio Visual Communication Review* 1.3, pp. 190–194.

Hershey, Shawn et al. (2017). “CNN architectures for large-scale audio classification”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings (ICASSP)*. ISBN: 9781509041176. DOI: 10.1109/ICASSP.2017.7952132.

Howard, Andrew G. et al. (2017). “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications”. In: *arXiv*. URL: http://arxiv.org/abs/1704.04861.

MoviePy (2017). URL: https://zulko.github.io/moviepy/ref/ref.html.

Ning, Guanghan et al. (2017). “Spatially Supervised Recurrent Convolutional Neural Networks for Visual Object Tracking”. In: *Proceedings - IEEE International Symposium on Circuits and Systems*. ISBN: 9781467368520. DOI: 10.1109/ISCAS.2017.8050867.

Petridis, Stavros et al. (2017). “End-to-End Audiovisual Fusion with LSTMs”. In: *The 14th International Conference on Auditory-Visual Speech Processing*, pp. 36–40.

Spoerer, Courtney J., Patrick McClure, and Nikolaus Kriegeskorte (2017). “Recurrent convolutional neural networks: A better model of biological object recognition”. In: *Frontiers in Psychology* 8.SEP, pp. 1–14. ISSN: 16641078. DOI: 10.3389/fpsyg.2017.01551.

Tao, Fei and Carlos Busso (2017). “Bimodal recurrent neural network for audiovisual voice activity detection”. In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Vol. 2017-Augus. International Speech Communication Association, pp. 1938–1942. DOI: 10.21437/Interspeech.2017-1573.

Vaswani, Ashish et al. (2017). “Attention Is All You Need”. In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1–11. DOI: 10.1109/2943.974352.

Wichmann, Felix A. et al. (2017). “Methods and measurements to compare men against machines”. In: *Electronic Imaging, Human Vision and Electronic Imaging*, pp. 36–45. DOI: 10.2352/ISSN.2470-1173.2017.14.HVEI-113.

Zhou, Yiren, Sibo Song, and Ngai Man Cheung (2017). “On classification of distorted images with deep convolutional neural networks”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings (ICASSP)*. January, pp. 1213–1217. ISBN: 9781509041176. DOI: 10.1109/ICASSP.2017.7952349.

Zhu, Zhuotun, Lingxi Xie, and Alan Yuille (2017). “Object recognition with and without objects”. In: *IJCAI International Joint Conference on Artificial Intelligence*, pp. 3609–3615. ISBN: 9780999241103. DOI: 10.24963/ijcai.2017/505.

Afouras, Triantafyllos, Joon Son Chung, and Andrew Zisserman (2018). “The conversation: Deep audio-visual speech enhancement”. In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 3244–3248. ISSN: 19909772. DOI: 10.21437/Interspeech.2018-1400.

Ben-Yosef, Guy, Liav Assif, and Shimon Ullman (2018). "Full interpretation of minimal images". In: *Cognition* 171.November 2017, pp. 65–84. ISSN: 18737838. DOI: 10.1016/j.cognition.2017.10.006. URL: https://doi.org/10.1016/j.cognition.2017.10.006.

Dodge, Samuel F. and Lina J. Karam (Nov. 2018). "Quality Robust Mixtures of Deep Neural Networks". In: *IEEE Transactions on Image Processing* 27.11, pp. 5553–5562. ISSN: 10577149. DOI: 10.1109/TIP.2018.2855966.

Ephrat, Ariel et al. (2018). "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation". In: *ACM Transactions on Graphics* 37.4. ISSN: 15577368. DOI: 10.1145/3197517.3201357.

Gabbay, Aviv et al. (2018). "Seeing through noise: Visually driven speaker separation and enhancement". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3051–3055. ISBN: 9781538646588. DOI: 10.1109/ICASSP.2018.8462527.

Geirhos, Robert, Carlos R M Temme, et al. (2018). "Generalisation in humans and deep neural networks". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Ed. by S Bengio et al. Vol. 31. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2018/file/0937fb5864ed06ffb59ae5f9b5ed67a9-Paper.pdf.

Gogate, Mandar et al. (July 2018). "DNN driven Speaker Independent Audio-Visual Mask Estimation for Speech Separation". In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 2723–2727. DOI: 10.21437/Interspeech.2018-2516. URL: http://arxiv.org/abs/1808.00060%20http://dx.doi.org/10.21437/Interspeech.2018-2516.

Gu, Chunhui et al. (2018). "AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6047–6056.

Kell, Alexander J E, Daniel L K Yamins, et al. (2018). "A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical

Processing Hierarchy". In: *Neuron*. DOI: 10.1016/j.neuron.2018.03.044. URL: https://doi.org/10.1016/j.neuron.2018.03.044.

Li, Yuanqing, Fangyi Wang, et al. (2018). "The effects of audiovisual inputs on solving the cocktail party problem in the human brain: An fMRI study". In: *Cerebral Cortex* 28.10, pp. 3623–3637. ISSN: 14602199. DOI: 10.1093/cercor/bhx235.

Liu, Kuan et al. (2018). "Learn to Combine Modalities in Multimodal Deep Learning". In: URL: http://arxiv.org/abs/1805.11730.

Noppeney, Uta, Samuel A. Jones, et al. (Nov. 2018). "See what you hear-How the brain forms representations across the senses". In: *Neuroforum* 24.4, pp. 237–246. ISSN: 23637013. DOI: 10.1515/nf-2017-A066.

Owens, Andrew and Alexei A Efros (2018). "Audio-Visual Scene Analysis with Self-Supervised Multisensory Features". In: *European Conference on Computer Vision (ECCV)*. URL: http://andrewowens.com/multisensory..

Rajalingham, Rishi et al. (2018). "Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks". In: *Journal of Neuroscience* 38.33, pp. 7255–7269. ISSN: 15292401. DOI: 10.1523/JNEUROSCI.0388-18.2018.

Robert, James, Marc Webbie, et al. (2018). *Pydub*. URL: http://pydub.com/.

Sandler, Mark et al. (2018). "MobileNetV2: Inverted Residuals and Linear Bottlenecks". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4510–4520. ISBN: 9781538664209. DOI: 10.1109/CVPR.2018.00474.

Tan, Chuanqi, Fuchun Sun, et al. (2018). "A survey on deep transfer learning". In: *27th International Conference on Artificial Neural Networks (ICANN)*. ISBN: 9783030014230. DOI: 10.1007/978-3-030-01424-7{_}27.

Zhou, Yipin, Zhaowen Wang, et al. (2018). "Visual to Sound: Generating Natural Sound for Videos in the Wild". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3550–3558.

Aller, Máté and Uta Noppeney (2019). "To integrate or not to integrate: Temporal dynamics of hierarchical Bayesian causal inference". In: *PLoS Biology* 17.4, pp. 1–31. ISSN: 15457885. DOI: 10.1371/journal.pbio.3000210.

Brendel, Wieland and Matthias Bethge (2019). "Approximating cnns with bag-of-local-features models works surprisingly well on Imagenet". In: *7th International Conference on Learning Representations, ICLR 2019*, pp. 1–15.

Cichy, Radoslaw M. and Daniel Kaiser (2019). "Deep Neural Networks as Scientific Models". In: *Trends in Cognitive Sciences* 23.4, pp. 305–317. ISSN: 1879307X. DOI: 10.1016/j.tics.2019.01.009.

Dodge, Samuel and Lina Karam (Mar. 2019). "Human and DNN classification performance on images with quality distortions: A comparative study". In: *ACM Transactions on Applied Perception* 16.2. ISSN: 15443965. DOI: 10.1145/3306241.

Geirhos, Robert, Claudio Michaelis, et al. (2019). "Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness". In: *7th International Conference on Learning Representations, (ICLR)*, pp. 1–22.

Jacobs, Robert A. and Christopher J. Bates (2019). "Comparing the Visual Representations and Performance of Humans and Deep Neural Networks". In: *Current Directions in Psychological Science (Association for Psychological Science)* 28.1, pp. 34–39. ISSN: 14678721. DOI: 10.1177/0963721418801342.

Kazakos, Evangelos et al. (2019). "EPIC-fusion: Audio-visual temporal binding for egocentric action recognition". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Vol. 2019-Octob, pp. 5491–5500. ISBN: 9781728148038. DOI: 10.1109/ICCV.2019.00559.

Kell, Alexander JE and Josh H. McDermott (Apr. 2019). "Deep neural network models of sensory systems: windows onto the role of task constraints". In: *Current Opinion in Neurobiology* 55, pp. 121–132. DOI: 10.1016/j.conb.2019.02.003.

Monfort, Mathew et al. (2019). “Moments in Time Dataset: One Million Videos for Event Understanding”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.2, pp. 502–508. DOI: 10.1109/TPAMI.2019.2901464.

Peirce, Jonathan et al. (2019). “PsychoPy2: Experiments in behavior made easy”. In: *Behavior Research Methods* 51.1, pp. 195–203. ISSN: 15543528. DOI: 10.3758/s13428-018-01193-y.

Sabir, Ekraam et al. (2019). “Recurrent Convolutional Strategies for Face Manipulation Detection in Videos”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) - Workshop*. URL: www.adobe.com/products/photoshopfamily.html.

Tan, Mingxing and Quoc V. Le (2019). “EfficientNet: Rethinking model scaling for convolutional neural networks”. In: *36th International Conference on Machine Learning, ICML 2019*. Vol. 2019-June, pp. 10691–10700. ISBN: 9781510886988.

Wang, Kafeng, Xitong Gao, et al. (2019). “Pay Attention to Features, Transfer Learn Faster CNNs”. In: *International Conference on Learning Representations (ICLR)*, pp. 1–14.

Yang, Ren et al. (2019). “Quality-gated convolutional LSTM for enhancing compressed video”. In: *IEEE International Conference on Multimedia and Expo*, pp. 532–537. ISBN: 9781538695524. DOI: 10.1109/ICME.2019.00098.

Zhang, Yuanyuan, Zi Rui Wang, and Jun Du (2019). “Deep Fusion: An Attention Guided Factorized Bilinear Pooling for Audio-video Emotion Recognition”. In: *Proceedings of the International Joint Conference on Neural Networks* 2019-July, pp. 1–9. DOI: 10.1109/IJCNN.2019.8851942.

Zhou, Pan, Wenwen Yang, et al. (2019). “Modality attention for end-to-end audio-visual speech recognition”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6565–6569. ISBN: 9781538646588.

Bergen, Ruben S. van and Nikolaus Kriegeskorte (2020). “Going in circles is the way forward: the role of recurrence in visual inference”. In: *Current Opinion in Neurobiology* 65, pp. 176–193. ISSN: 18736882. DOI: 10.1016/j.conb.2020.11.009. URL: https://doi.org/10.1016/j.conb.2020.11.009.

Cheng, Ying et al. (2020). “Look, Listen, and Attend: Co-Attention Network for Self-Supervised Audio-Visual Representation Learning”. In: *MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia*. October, pp. 3884–3892. ISBN: 9781450379885. DOI: 10.1145/3394171.3413869.

Dyck, Leonard E van and Walter Gruber (2020). “Seeing eye-to-eye? A comparison of object recognition performance in humans and deep convolutional neural networks under image manipulation”. In: July.

Funke, Christina M. et al. (Apr. 2020). “The Notorious Difficulty of Comparing Human and Machine Perception”. In: URL: http://arxiv.org/abs/2004.09406.

Gidon, Albert et al. (2020). “Dendritic action potentials and computation in human layer 2/3 cortical neurons”. In: *Science* 367.6473, pp. 83–87. ISSN: 10959203. DOI: 10.1126/science.aax6239.

Harris, Charles R et al. (2020). “Array programming with NumPy”. In: *Nature* 585, pp. 357–362. DOI: 10.1038/s41586-020-2649-2.

Khaki, Saeed, Lizhi Wang, and Sotirios V. Archontoulis (2020). “A CNN-RNN Framework for Crop Yield Prediction”. In: *Frontiers in Plant Science* 10.January, pp. 1–14. ISSN: 1664462X. DOI: 10.3389/fpls.2019.01750.

Li, Ang, Meghana Thotakuri, et al. (2020). “The AVA-Kinetics Localized Human Actions Video Dataset”. In: *arXiv preprint*, pp. 1–8. URL: http://arxiv.org/abs/2005.00214.

Mervitz, J. H. et al. (2020). “Comparison of early and late fusion techniques for movie trailer genre labelling”. In: *Proceedings of 2020 23rd International Conference on Information Fusion, FUSION 2020*. ISBN: 9780578647098. DOI: 10.23919/FUSION45008.2020.9190344.

Mihalik, Agoston and Uta Noppeney (2020). “Causal inference in audiovisual perception”. In: *Journal of Neuroscience* 40.34, pp. 6600–6612. ISSN: 15292401. DOI: 10.1523/JNEUROSCI.0051-20.2020.

Peirce, Jonathan et al. (2020). *Pavlovia*. URL: https://pavlovia.org/.

Plakal, Manoj and Dan Ellis (2020). *YAMNet*. URL: https://github.com/tensorflow/models/tree/master/research/audioset/yamnet.

Reddy, Chandan K.A., Vishak Gopal, et al. (2020). "The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Testing Framework, and Challenge Results". In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Vol. 2020-Octob, pp. 2492–2496. DOI: 10.21437/Interspeech.2020-3038.

Singer, Johannes, Katja Seeliger, and Martin N. Hebart (2020). "The Representation of Object Drawings and Sketches in Deep Convolutional Neural Networks". In: *2nd Workshop on Shared Visual Representations in Human and Machine Intelligence (SVRHM) Neural Information Processing Systems (NeurIPS)*.

Smaira, Lucas et al. (2020). "A Short Note on the Kinetics-700-2020 Human Action Dataset". In: *arXiv preprint*, pp. 1–5. URL: http://arxiv.org/abs/2010.10864.

Spoerer, Courtney, Tim Kietzmann, et al. (2020). "Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision". In: *PLoS Computational Biology*, p. 677237. DOI: 10.1101/677237.

Virtanen, Pauli et al. (2020). "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17, pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

Wang, Biao, Yaguo Lei, et al. (Feb. 2020). "Recurrent convolutional neural network: A new framework for remaining useful life prediction of machinery". In: *Neurocomputing* 379, pp. 117–129. ISSN: 18728286. DOI: 10.1016/j.neucom.2019.10.064.

Xiao, Fanyi, Yong Jae Lee, and Kristen Grauman (2020). "Audiovisual SlowFast Networks for Video Recognition".

Yu, Wentao, Steffen Zeiler, and Dorothea Kolossa (2020). "Multimodal integration for large-vocabulary audio-visual speech recognition". In: *European Signal Processing Conference (EUSIPCO)*, pp. 341–345. ISBN: 9789082797053. DOI: 10.23919/Eusipco47968.2020.9287841.

Akbari, Hassan et al. (2021). “VATT : Transformers for Multimodal Self-Supervised Learning from Raw Video , Audio and Text”. In: *Neural Information Processing Systems (NeurIPS)*. NeurIPS, pp. 1–20.

Aldeneh, Zakaria et al. (2021). “On the role of visual cues in audiovisual speech enhancement”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 2021-June, pp. 8423–8427. ISBN: 9781728176055. DOI: 10.1109/ICASSP39728.2021.9414263.

Fang, Huajian et al. (2021). “Variational Autoencoder For Speech Enhancement With a Noise-Aware Encoder”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 676–680. ISBN: 9781728176055.

Funke, Christina M. et al. (2021). “Five points to check when comparing visual perception in humans and machines”. In: *Journal of Vision* 21.3, pp. 1–23. ISSN: 15347362. DOI: 10.1167/jov.21.3.16.

Gupta, Gaurav et al. (2021). “Comparing recurrent convolutional neural networks for large scale bird species classification”. In: *Scientific Reports* 11.1, pp. 1–12. ISSN: 20452322. DOI: 10.1038/s41598-021-96446-w. URL: https://doi.org/10.1038/s41598-021-96446-w.

Heinke, Dietmar et al. (2021). “A failure to learn object shape geometry: Implications for convolutional neural networks as plausible models of biological vision”. In: *Vision Research* 189, pp. 81–92. ISSN: 18785646. DOI: 10.1016/j.visres.2021.09.004. URL: https://doi.org/10.1016/j.visres.2021.09.004.

Nagrani, Arsha et al. (2021). “Attention Bottlenecks for Multimodal Fusion”. In: *Neural Information Processing Systems (NeurIPS)*. NeurIPS, pp. 1–14.

Reddy, Chandan K.A., Harishchandra Dubey, et al. (2021). “ICASSP 2021 Deep Noise Suppression Challenge”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 2021-June, pp. 6623–6627. ISBN: 9781728176055. DOI: 10.1109/ICASSP39728.2021.9415105.

Rideaux, Reuben et al. (2021). “How multisensory neurons solve causal inference”. In: *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* 118.32.

Yuille, Alan L. and Chenxi Liu (2021). “Deep Nets: What have They Ever Done for Vision?” In: *International Journal of Computer Vision* 129.3, pp. 781–802. ISSN: 15731405. DOI: 10.1007/s11263-020-01405-z. URL: https://doi.org/10.1007/s11263-020-01405-z.

Zhang, Zhihui, Xiaoqi Li, et al. (2021). “Neural noise embedding for end-to-end speech enhancement with conditional layer normalization”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 2021-June, pp. 7113–7117. ISSN: 15206149. DOI: 10.1109/ICASSP39728.2021.9413931.

Cambridge Dictionary (2022). URL: https://dictionary.cambridge.org/dictionary/english/pouring.

Sun, Zehua, Qiuhong Ke, et al. (2022). “Human Action Recognition From Various Data Modalities: A Review”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–20. ISSN: 0162-8828. DOI: 10.1109/tpami.2022.3183112.