Attention Seeds: Extracting and Understanding Latent Attention Factors through Parameter-Reduction on Goal-Directed Attention Mechanisms

Candidate Number: QLDZ4[1]

MSc Machine Learning

Supervised by

Brad Love

Xiaoliang Luo

Submission date: 10th September 2022

Abstract

Top-down goal-directed attention tunes the visual system of neural networks, both artificial and biological, to focus on information most beneficial in achieving the current goal of the learner. Many implementations have transferred the performative enhancements of selective cognitive attention, found in categorisation models, to later neural network representations deep convolutional neural networks (DCNNs). However, this locks the number of attention parameters to equal that of the dimensions of the filter-space, leading to extremely high dimensional attention unlike what is found in categorisation models. Redundancies in the filter-space are therefore transferred to the attention parameters, decreasing explanatory power of learned attention weights. Motivated by the idea of aligning current attention mechanisms with cognitive attention and relieving such limitations, this dissertation develops a new parameter-reduction technique for top-down goal-directed attention mechanisms, dubbed *attention seeding*, that trains attention weights in a latent subspace of the filter-space. We postulate that by extracting low-dimensional latent attention factors we can better identify factors guiding the learning of attention weights and remove redundancies in full-dimensional attention mechanisms while maintaining perceptual performance, increasing our understanding of why attention weights are learned how they are.

Our main outcomes show that while operating at a 50% reduction in attention parameters the new mechanism maintained goal-directed perceptual boosts and model performance under a signal-detection analysis framework. Through representational similarity analysis, the new mechanism learned attention weightings that were found to be more sensibly explainable, with stronger correlations to underlying semantic, class-difficulty, and confusion-rate relationships between task-sets (goals) of the network. We uncovered that class-difficulty and inter-class confusion-rates were strongly influencing factors that drive the learning of attention weights within the low-dimensional subspace. Finally, we demonstrate that even under a parameter reduction and maintained performance, lower-dimensional attention weights retained large amounts of informational content found in the high-dimensional attention weights, strongly confirming the existence of low-dimensional latent attention factors driving high-dimensional attention weights.

Acknowledgements

The author pays his greatest acknowledgement to Xiaoliang Luo, who acted as a fantastic supervisor and was incredibly generous with his time in supporting the development of key ideas and research within the thesis; this project would not have been as great without him. The author would also like to pay great acknowledgment to Bradley Love for overall supervision of the project and for supplying the initial ideas that were foundational for completing the thesis presented. The author finally thanks both for an incredible educational experience while carrying out the project.

Code

For accompanying code to the thesis please see
`http://github.com/mazabdul7/msc_attention_project`

Contents

Chapter 1

Introduction

Machine Learning and Computational Neuroscience are two closely linked fields that have influenced each other time and time again within recent decades [Savage, 2019]. The emergence and rapid progress of deep learning (DL), a very successful subset of Machine Learning, can be attributed to model architectures being designed to mimic the processing of sensory information within the human brain. Artificial Neural Networks (ANNs) found in DL follow a feed-forward structure, with layers of nodes analogous to neurons and their integration and activation properties within the brain [Yamins and DiCarlo, 2016]. The emergence of successful ANNs can most definitely be attributed to the influence of neuroscience, but this influence is mutual. ANNs have proved very useful for studying the brain. ANNs were found to produce patterns of neural activity that resembled that recorded from the brain [Yamins et al., 2014]. ANNs can therefore make great models of the brain, further allowing researchers to examine and draw inferences about how different processes within the brain perform their biological functionalities [Savage, 2019]. This is just one of many successful developments drawn from the relationships between both fields. The potential of DL for neuroscience is exciting, and stretches many domains from better understanding of visual systems of the brain, to auditory and speech processing systems of the brain [Kell et al., 2018]. To date, most successful research outcomes have been associated with the visual system of the brain, a big focus of the work considered within this thesis.

1.1 Goal-Directed Attention & Our Contribution

The vision system within the human brain is known to accomplish many tasks; from object recognition to tracking, segmentation and more behavioural aspects such as obstacle avoidance and many more [DiCarlo et al., 2012]. Although very diverse in it's application, our view of the

vision system is that with a focus on performing object recognition. We define object recognition as the task of assigning labels to particular objects and these can be specific precise labels such as the name of the object in the input, or more general categorisation relating a grouping of objects of similar properties. The human vision system is remarkable in that such accurate classification of objects central to the visual receptive field occurs in as little as approximately $350ms$ [Rousselet et al., 2002, Thorpe et al., 1996]. But with so much going on in our world, how does the human brain localise learning to a single visual task, such as learning a new object, so quickly? Enter goal-directed attention and its crucial role in adapting human cognition - sifting through the vast amounts of information we consume to the attend to the features that align best with the current goals what we are trying to accomplish.

Human cognition models have been around for decades, with the aim to explain such learning by process of establishing a memory trace that improves the efficiency of assigning novel objects to contrasting groups (e.g new person is friend or foe). Popular models like ALCOVE [Kruschke, 1990] and SUSTAIN [Love et al., 2004] state that perceptual information, such as seeing an object, is translated to set of features organized along a set of dimensions. The role of goal-directed attention is to then reconfigure the visual system through attentional tunings of each feature dimension in order to adaptively determine the importance of features in achieving the current goal of the learner. This goal-directed pressure is exerted in a top-down fashion by the prefrontal cortex (PFC) on the vision system in order to favour goal-relevant information [Miller and Cohen, 2001]. This attention leads to greater inter-category similarity and greater between-category dissimilarity based on the categorical goal of the current task of the network. Take looking for your car keys in the kitchen. The attention focus may see attenuation that favours small and shiny metallic features. It may lead one to make a false alarm with a small spoon occluded by a chopping board, but in time increases the chances of finding the car keys (see Figure 1.1).

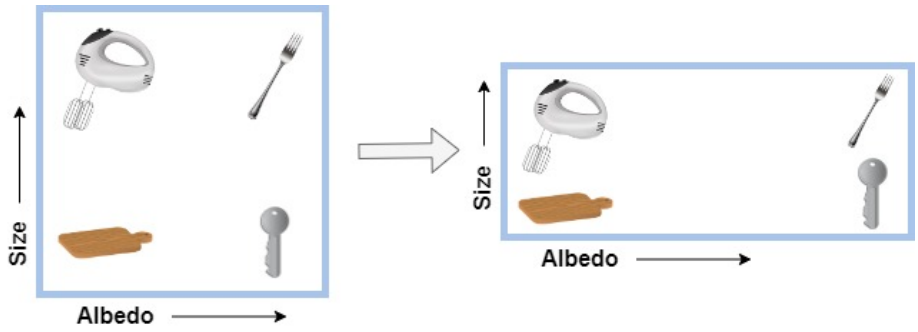


Figure 1.1: Attentions effect in altering the importance of feature dimensions. Four kitchen objects vary on two feature dimensions: *albedo* and *size*. When looking for ones car keys the attention mechanism attends to the albedo dimension (hence stretched) whereas attention to size is tuned downwards (hence compressed). As such the key becomes more similar to the shiny fork than to the plastic hand mixer or the chopping board. Figure adapted from [Luo et al., 2021].

Attention has seen great success across Machine Learning in key domains such as natural language processing (NLP) [Bahdanau et al., 2014, Luong et al., 2015, Vaswani et al., 2017] and image recognition [Dosovitskiy et al., 2020, Hu et al., 2018, Perez et al., 2017]. Typically, such applications do not exhibit a goal-directed agenda and are seen as bottom-up approaches since the mechanism is reactive to the context (e.g image or word) and not driven by a specific goal of the network. With a focus on neuroscience-informed attention mechanisms, we shift our attention to top-down goal-directed attention mechanisms. Recent years have seen great development in approaches to implementing such attention into deep hierarchical neural networks (DCNNs). DCNNs are a primary base to goal-directed attention mechanism due to the striking similarities in how DCNNs process naturalistic visual stimuli in comparison to the human ventral stream [Yamins and DiCarlo, 2016, Schrimpf et al., 2018, Wen et al., 2018]. Current state-of-the-art work in top-down goal-directed models have seen implementations of multiplicative [Luo et al., 2021] and additive [Lindsay and Miller, 2018] attention layers that modulate feature-map outputs of DCNN convolutional layers. Such approaches have also considered spatial attention in which different modulations are applied to different areas of the feature-map output based on the current goal of the network [Lindsay and Miller, 2018]. A goal of the network is typically embodied as the specialising of the network to a set of ImageNet classes, denoted the task-set, that the network is configured for. The most successful implementations were yielded from multiplicative, spatially-invariant attention mechanisms that modulate feature-maps in a one-to-one fashion [Luo et al., 2021, Lindsay and Miller, 2018]. For example, under direction of the current task (class) of the model, an image is passed through the network where it is encoded into a set of higher dimensional feature-maps output at a particular convolutional layer. The learned attention weights

that specialise to the current task are then used to modulate each feature-map in such a way that the model better localises the current task amongst all inputs to the network. Such implementations have exhibited great perceptual boosts and accuracy gains when predicting over samples of the task-set.

A glaring issue is that interfacing such top-down goal-directed attention mechanisms with later layers of a DCNN leads to very high-dimensional, and thus harder-to-understand attention weights. Attention weights found in aforementioned cognitive models are typically hand-encoded and of low-dimensions, therefore offering an intuitive and straightforward interpretation in that features corresponding to the current goal are modulated upwards, while those that confound the current goal are modulated downwards. Furthermore, with much research indicating the over-parameterisation and redundancies found in feature-map outputs of DCNNs [Xie, 2020, Qiu et al., 2021, Cai et al., 2022], not all attention weights are in fact required when achieving the current task of the network, and we wish to prove such redundancies are also existent within attention mechanisms. Many attention mechanisms that modulate feature-maps in a similar fashion are limited further in that the number of attention parameters trained must equal the number channels of the modulated feature-map.

Motivated by the idea of aligning current attention mechanisms with cognitive attention and relieving such limitations, we propose a new parameter-reduction technique for top-down goal-directed attention mechanisms that trains attention weights within a latent subspace of the full-dimensional filter-space. The lower-dimensional attention weights, dubbed *attention seeds*, are seen as low-dimensional latent factors that drive the high-dimensional attention weights. Through a projection mechanism, we project the attention seeds out of the latent attention subspace when modulating feature-maps in the original full-dimensional filter-space. The attention seeds projection utilises a non-negative step-up projection function that has clear interpretation in that the final attention weights are additive combinations of the latent factors represented by attention seeds. Aside from comparing the performance between standard goal-directed attention mechanisms and the proposed lower-dimensional attention seeds mechanism, we also attempt to localise which factors dictate the optimal convergence of attention weights during the learning process through a representational similarity analysis (RSA), a popular form of multivariate pattern analysis (MVPA) used within neuroimaging literature. By evaluating the similarities between the representational geometries learned by the low-dimensional attention seeds mechanism to underlying class-level factors such as the semantic entanglement between task-sets, task-set difficulty and also redundancies found in feature-map activations of task-sets, we can begin to quantify the explanatory power of low-dimensional attention weights in comparison to the

high-dimensional attention weights.
Concretely put, in this thesis we answer the following questions:

1. *Are redundancies found within the filter-space translated to redundancies within the attention parameters? If so can we develop a technique to perform a parameter-reduction on goal-directed attention to eliminate such redundancies, while maintaining all perceptual performance gains yielded from the attention mechanism?*

2. *Do these low-dimensional attention weights act as latent attention factors that drive the high-dimensional attention weights?*

3. *With less attention parameters to work with, are low-dimensional attention weights more greatly influenced by factors that govern similarities between learned attention weights? Does this therefore pose an increased likelihood in extracting such factors from representational analysis between learned low-dimensional attention weights?*

4. *Can we identify such factors that pose the greatest influence over the learning process of attention parameters?*

Answering these questions requires a two-staged research process which is maintained throughout this thesis. The first stage focuses on developing the attention seeds mechanism, a parameter-reduction technique for one-to-one modulating attention mechanisms. The second stage evaluates the efficacy of low-dimensional attention weights in yielding an increased likelihood of identifying such underlying factors influencing the learning of attention weights relative to the task-sets.

From a neuroscience standpoint, we hope this more faithful implementation of cognitive attention yields an increased explanatory power of learned attention weightings, and demonstrates that low-dimensional latent attentions factors exist, eliminating redundancies within full-dimensional attention weights. From a Machine Learning standpoint, with larger and larger neural networks seen everyday, model over-parameterisation and its negative impact on model efficiency and training times is becoming a growing concern. Our work provides a method for the elimination of redundant trainable attention parameters within models containing similar attention mechanisms. This offers great implications in increasing model efficiency and decreasing model training times, while maintaining performance integrity of the attention mechanism, and the model.

1.2 Brief Summary of Results

Within this thesis we successfully developed a new lower-dimensional top-down goal-directed attention mechanism as an extension of the work in [Luo et al., 2021]. The new mechanism, while operating at a 50% reduction in attention parameters, maintained goal-directed perceptual boosts and model performance under a signal-detection analysis framework in comparison to the standard goal-directed attention of [Luo et al., 2021]. The lower-dimensional seeds mechanism learned attention weightings that were found to capture underlying representational patterns more strongly between task-sets over the standard full-dimensional goal-directed attention mechanism. This is in line with our hypothesis that low-dimensional attention weights are warped more greatly by governing factors that influence similarities between attention weights, hence revealing these low-level factors. The lower-dimensional attention weights were also found to be more sensible, as we successfully correlate low-dimensional attention weights to underlying semantic relationships between task-set semantic embeddings. Our hypothesis that there is a low-dimensional set of latent factors driving the high-dimensional attention weights appears to be true, with lower-dimensional attention seed weights highly correlating with the full-dimensional standard attention mechanism weights across 21 distinct task-sets, consisting of 315 sets of attention weights. Even at a 94% reduction in trainable attention parameters, the low-dimensional attention weights were still strongly correlated to the full-dimensional attention weights, suggesting they are indeed low-dimensional latent factors. Class difficulty, and more importantly confusion rates amongst similar classes, were found to be strongly influencing factors governing the representational geometries underpinning the attention seeds learning process, providing more insight into understanding the how and why of learned attention weights. By demonstrating the success that pairing representational similarity analysis with dimensionality-reduction techniques had within our experiments, we champion that the same analysis can be used to investigate other governing factors influencing attention parameters, opening the doors to constructing a complete understanding of the role attention plays within DCNNs. In fact, we put forward that such a powerful pairing can be used across multiple domains of Machine Learning, leading to a fuller understanding of what is considered the unknowns of the training process of deep neural network parameters.

1.3 Overview of Thesis

This thesis is organised in the following parts. Section 2 presents a review of the literature related to top-down attention from a neuroscience perspective and a Machine Learning perspective, as

well as a review of techniques relating to dimensionality-reduction used within neural networks. Section 3 presents technical details and explanations pertaining to the new lower-dimensional goal-directed attention mechanism, dimensionality-reduction techniques, model training settings, chosen task-sets (goals) of the network, and finally, the model evaluation frameworks used which include the RSA. Section 4 contains the results of analyses, and evaluations of the results of the new attention seeds mechanism relative to the standard attention mechanism. Section 5 concludes the thesis with a general discussion of results and answers the four main questions we investigated. With the bulk of the thesis work dedicated to developing a working dimensionality-reduction technique for goal-directed top-down attention modulation, we also present suggestions for future work, with many promising avenues to extend research in furthering understanding of attention through the attention seeds mechanism.

Chapter 2

Related Work

2.1 Neuroscience-Based Models of Human Cognition

In understanding popular model architectures for DCNNs it is useful to review the neuroscience basis by which human categorical learning is modelled. Categorical learning is the process of establishing a memory trace that improves the efficiency of classifying novel objects to contrasting groups. Two popular models for such learning are known as ALCOVE [Kruschke, 1990] and SUSTAIN [Love et al., 2004]. These models state that perceptual information is translated to a set of features organised along a set of dimensions. These features are not directly generated from naturalistic inputs such as photographs, but are fixed discrete features that are hand-encoded based on the visual stimuli. For example a feature could be the size or colour of the supposed input. A discrete value is then assigned to rate the feature, again done manually and hence is a limitation to these models. ALCOVE is an extension of Nosofsky's generalised-context-model (GCM) [Nosofsky, 2011] and features a focus on determining directly which dimensions of latent representations are most relevant to the task at hand and assumes orthogonal feature representations. SUSTAIN implores a simple cluster structure that assumes feature representations to be correlated and aims to prove discontinuities between such representations during learning. Most notably, both models are formed under the idea that learning can be highly dependent on the current goals of the learner and as such we as humans optimise learning with such goals in mind. In both, the idea of feature modulation through attention is used to achieve such goal-directed learning, where upon observing new input stimuli, the model attenuates certain dimensions of the input representation in accordance with the overall goal in mind. Goals can be thought of as an external top-down influence catering to the demand of the human at the moment of processing any input stimuli. Already, both models provide a basis of constraints for which a cognitively

aligned attention mechanism should follow. We will refer back to these when considering the design of the new attention mechanism later on. The important idea portrayed by both cognitive models is that there is a low-dimensional set of attention modulations applied to the hand-encoded features of processed input stimuli, with each weighting corresponding to a clear feature of the input stimuli. This in-turn drives our idea that high-dimensional attention weights may contain redundancies or may simply attend to too many features making it harder to truly understand what the weightings mean.

2.2 Deep Learning and Convolutional Neural Networks

Within recent years, deep learning (DL) has seen tremendous success in obtaining outstanding performance across multiple distinct applicational domains, from natural language processing, to audio and speech processing, and most notably within the field of computer vision [Chai et al., 2021]. This is due to the relatively easy nature of feature extraction achieved through DL, specifically via libraries that make DL widely accessible such as Google's TensorFlow and Facebook's PyTorch [Abadi et al., 2015, Paszke et al., 2017]. Within computer vision, DL performs function approximation and feature extraction by utilising a multi-layer architecture in which earlier layers extract low-level features with later layers extracting high-level features. This leads to rich, high-dimensional representations of input data to neural networks which can then be processed and decoded differently dependent on the applicational task of the network. Neural networks are therefore parametric models with parameters within feature-extraction mechanisms. Parameters are then learned via a stochastic gradient descent update process [Amari, 1993] in order to minimise the objective loss between the training set and the predictions of the model. This leads to specialised models that have learned to intake data, encode and then extract underlying patterns from such encodings in order to achieve a variety of tasks such as image segmentation, classification and even object recognition.

Convolutional neural networks (CNN) are one of the most popular DL algorithms. CNNs encode a strong inductive bias of spatial and temporal equivariance through the translationally invariant convolutional process and shared weights within feature-extracting filters of layers. CNNs typically consist of consecutively stacked convolutional, pooling and densely-connected classification layers. Convolutional layers consist a set of filters, called kernels, with learned parameters that aim to minimise the overall loss of the network for the task at hand. Kernels convolve the input to a layer by sliding across the entire input region, resulting in a feature-map of activations. Activations are outputs of the convolutional process between filters and select areas

of the input and act as a measure of similarity between a particular filter within the convolutional kernel and the selected area of the input. The feature-map outputs are then passed onto the next layer, which performs the same feature-extraction process again using different convolutional kernels with different sizing, padding, and or strides, leading to rich encodings of input data to the network. These encodings can then be densely compressed to form output predictions conditioned on patterns extracted from the encodings. From a neuroscience perspective, activations are parallel to neuron responses within the ventral visual pathway to visual stimuli [Kalfas et al., 2018]. Typically in Machine Learning, input data to neural networks comprise a tensor of shape batch size$\times$input height$\times$input width$\times$input channels and after passing through the convolutional layer, the resulting feature-map is of shape batch size$\times$feature-map height$\times$feature-map width$\times$ feature-map channels. These feature-maps are equivalent to features extracted in categorisation models like ALCOVE and SUSTAIN, though automatic and formed from naturalistic inputs. Hence these feature-maps provide a basis for modulation by cognitive attention.

2.3 Hierarchical Deep Learning Networks as Models of the Brain

Under direction of neuroscience-informed attention mechanisms, like those found in cognitive models, our work is concerned with utilising accurate models of the human ventral stream, the pathway by which we view the world and consume visual stimuli to enable inference. Visual information is processed in two distinct pathways within the brain that are found to be influenced by an external task-relevant influence. The ventral visual pathway processes stimuli to answer 'what' a certain object we are viewing is and is typically invariant in it's representation to the demand of the task at hand. The second pathway is the dorsal pathway which is more adaptive and answers 'what we do' with the object - hence it is found to be more influenced by such a top-down external influence [Vaziri-Pashkam and Xu, 2017]. We focus more on models of the ventral stream due to great successes in the literature in linking neural network models to the ventral stream hierarchy with high correlations identified between internal representations within DCNNs to internal representations exhibited within different stages of the ventral stream. It can be seen that the ventral stream consists of a series of distinguishable but connected areas, going from earlier stages such as V1 and V2, to later, more complex areas such as V4 and the inferior temporal cortex (IT). Earlier stages are understood to encode the visual stimuli into patterns of neural activity where as later stages are known to perform decoding in which neural activity generates behaviour in the brain [Yamins and DiCarlo, 2016]. In fact, hierarchical DCNNs are known as

quantitatively accurate models for the encoding and decoding stages across multiple regions of the ventral visual stream, with accurate predictions of population aggregate responses in fMRI data from the ventral visual pathway [Yamins and DiCarlo, 2016]. Down to the basic mechanism that operates DCNNs, convolutions are known to be translation invariant, meaning that an object in image that is augmented via rotation or translation will elicit the same response, much like how the invariance in recognition of novel objects is a basic aspect of human vision [Han et al., 2020]. Due to consistent diameter increase of receptive field sizing across areas in the visual stream, the hierarchy lends itself to a nearest neighbour construction of the area input, akin to the nearest-neighbour computation of neighbouring pixels within the convolution process [Wilson and Wilkinson, 2015]. A very interesting result is that the authors of [Raghu et al., 2021] found that within DCNNs like ResNet [He et al., 2015], the effective receptive fields were highly local and increased in size gradually throughout the network, and as such this connection of DCNNs to the ventral visual pathway is strengthened. Mixing this with Gabor-like contour filter banks found within HCNNs, the feature extraction process itself becomes further invariant [Yamins and DiCarlo, 2016]. Successively training on augmented data that shows objects under many different conditions expected in the real-world leads the overall network to become largely invariant to many different types of augmentations such as illuminance, skewing, scaling and so forth. This is of critical importance as within the real-world, we view the many objects under varied unique conditions exhibiting identity-preserving image transformations, yet we are still quick to identify and label or categorise objects. Further studies measuring behavioural [Potter, 1976, Thorpe et al., 1996] and neuronal [Hung et al., 2005] evidence suggests this intolerance to such variance is well exhibited within the human visual stream. Furthermore, [Yamins and DiCarlo, 2016] states a good model for the ventral visual pathway in computational neuroscience is one that is image-computable; that is, it can generate responses for any arbitrary input, and that they are mappable to structures in the ventral pathway. Due to the maintenance of these many key characteristics, DCNNs are of great importance when building biologically-plausible models of the ventral visual pathway, and hence make perfect base models to drive our neurologically-plausible attention application on.

However, some of the literature identifies caveats to such early-to-early and late-to-late mappings of the ventral stream to DCNN stages. Authors in [Sexton and Love, 2022] directly spliced brain region activation's from fMRI data to their respective stages in DCNNs. This would prove that if there is indeed a relational mapping and preserved hierarchical nature to these models, then no significant performance drops should be observed and that this splicing is realisable. It was revealed that hierarchical representation may not be all that is at play. When interfacing the

model with fMRI brain activity from early and later regions, only later model layer interfacing was found to best correspond brain activity along the ventral stream. This lends a priori in that the positioning of goal-directed attention layers should be within later DCNN stages if we best want to emulate attentions role on neural activity within the human vision system.

Authors in [Schrimpf et al., 2018] give recent developed DCNNs in computer vision a 'brain-score', in which they evaluate how brain-like a model is with focus on parts of the brain known for object detection. They compared models on neural metrics and behavioural benchmarks such as comparing neural recordings from cortical areas and the similarity of internal signals between image-evoked feature activations in DCNNs to activations in different primate brain regions. A correlation between the ImageNet performance and neural data prediction was found to be weak for more recent models with higher Top-1 accuracy on the ImageNet test set, where as older models like DenseNet-169 and VGG-16 where found to have very high brain-scores, suggesting that they model the ventral visual pathway the best with all correlations found to be significant. From this result, and prior goal-directed attention work covered in the following section, we consider VGG-16 as a prime candidate for the base model to test our cognitively aligned top-down goal-directed attention mechanism on.

2.4 Attention Mechanisms within Deep Models

Attention is the core cognitive process of selectively focusing on key aspects of incoming information while tuning out other details. The human brain performs this by actively reformatting incoming sensory data to serve the host organisms current behavioural needs [Yamins and DiCarlo, 2016]. This property exhibited in neural attention models has seen tremendous success in state-of-the-art DL models within several application domains. Within NLP it has alleviated long-sequence dependencies in sequence-to-sequence models [Bahdanau et al., 2014, Luong et al., 2015], reducing memory constraints on processing such sequences enabling complex long-range tasks such as machine translation, sentiment analysis, and question-answering, among many other applications. More recently the emergence of more powerful self-attention mechanisms spawned an entirely new architecture based solely on attention known as transformers [Vaswani et al., 2017]. Transformers have revolutionised the field of NLP leading to such successes like giant pre-trained language models [Brown et al., 2020], and have contested state-of-the-art models in image recognition with vision transformers [Dosovitskiy et al., 2020].

Within the domain of DCNNs relative to our work, attention has also proved extremely valuable in goal-directed networks through feature-wise modulation of inter-layer activations.

The authors of [Lindsay and Miller, 2018] experimented with multiple forms of top-down goal-directed attention. Examples include feature similarity modulation to detect certain objects within adversarial images. They also evaluated how placement of the feature-wise attention modulation within the DCNN would affect the performance gains exhibited by the network. Such a top-down modulation is performed on the bottom-up representation of naturalistic input stimuli to the DCNN at later layers. The feature similarity gain model of attention stipulates that neural activity is positively modulated by attention proportional to a neuron's response to the attended features, a neuroscience-informed view of attention. The adversarial sets contained overlapped alpha-channel mixed images from different ImageNet classes, as well as array-like tiles of images from separate classes. For the new reader, ImageNet is a popular dataset comprising of over 15 million labeled naturalistic high-resolution images from around 22,000 categories [Deng et al., 2009] and is heavily used as a benchmark for image recognition models. They employed the use of a DCNN, specifically VGG-16, as their base model, however did not preserve the original 1000-way ImageNet-1k classification and developed 1000 binary classifiers that apply the feature gain modulation for their respective target-class in a one-vs-rest fashion leading to a goal-directed framework for each classifier. A baseline model was trained to evaluate the perceptual boost the attention mechanism provides on the target-class. The baseline model featured the same attention mechanism, however was trained on a normal 1000-way classification. With attention weightings learned through gradient descent, significant performance gains were exhibited when exhibited around the mid-layers, another hint for where we should place such attention influences. The criticality of the placement of the attention layer is no surprise as it is known within DCNNs there is non-uniform covariance of later layer activations to earlier layer activations, suggesting representational geometries are not maintained through propagation from early-to-late layers within DCNNs, and are in fact permuted heavily [Raghu et al., 2021]. As such any early attentional tunings may become redundant once reaching final classification layers [Lindsay and Miller, 2018].

Similarly, the authors of [Luo et al., 2021] evaluate a one-weight-per-channel feature-wise modulation of inter-layer activations, however, maintain the original 1000-way classification of VGG-16 and utilise an attention layer that is inserted within the pre-existing DCNN. Not altering the network structure is in line with how the PFC exerts a top-down goal-directed attention on the human vision system, acting as an external influence on the network as a whole [Miller and Cohen, 2001]. Again, a baseline model was trained with no specialisation to any specific target-classes. As such the resultant learned attention weights are thought to be task-generic. Task-specific models were trained by altering a custom cross-entropy loss to accommodate for

different intensities of attention, conveyed as an up-or-down-weighting of gradient updates to attention weights. Task-specific models are models with a goal in mind, simply put they focus on target-classes that the model specialises for, or in laymen terms, act as the goals of the network. This is found to be equivalent to training using a training set distribution that scales the percentage of target-class to non-target-class examples shown to the network per training cycle (epoch), proportional to the attention intensity. Through a signal-analysis framework, they evaluated the costs (false-alarm rates) and benefits (hit-rates) of the attention mechanism as well as considering the sensitivity and criterion of different attention intensities. Their primary goal was identifying the attention intensity which maximises the benefits of attention, while accordingly minimising the costs. The best attention intensity was found to be an intensity of 0.5, meaning the target-class comprised half the training set with the remaining half sampled uniformly at random each training epoch, leading to insight into preferred training distributions within our work. Both feature-wise top-down goal-directed attention mechanisms presented above are considered state-of-the-art implementations of attention found in cognitive models, and therefore provide great insight for the development of our new attention mechanism.

To summarise, placement of the attention mechanism appears to be most optimal within the middle layers of the network, and to maximise the performance gains yielded from attention, the training distribution should feature equal proportions of target-class and non-target-class samples during training. The implementation found in [Luo et al., 2021] is the most aligned to the goal-directed influences found within the brain, and as such we utilised it as a strong basis for the parameter-reduction technique developed within this thesis.

2.5 Dimensionality Reduction within Deep Models

Dimensionality-reduction techniques such as principal component analysis (PCA) [Maćkiewicz and Ratajczak, 1993] and non-negative matrix factorisation (NMF) [Dhillon and Sra, 2005]) have seen widespread usage within DCNNs within various applicational domains. Typically, they are associated with performing model compression through pruning of channel redundancies in intermediary outputs of network layers. Pruning is the act of removing redundant filters and has shown to successfully compress models, reducing their overall computational expense while preserving model accuracy [Li et al., 2016, Xie, 2020, Cai et al., 2022, Xie, 2020]. Channel redundancies are typically caused by high correlation and dependencies between layer neurons [Xie, 2020]. Such redundant channels are seen as a waste of computational resource as they often do not contribute to the final classification. Redundancies are also seen as a by-product

of the over-parameterisation of DCNNs. This leads the majority of convolutional layer weights to be very close to zero and hence these parameters hardly contribute to the classification [Qiu et al., 2021]. This connection is especially important given we decompose such convolutional kernels in our work, as such removing any redundancies will aid in down-stream projection of latent attention seeds described in later sections of this thesis. Singular Value Decomposition is typically involved in many pruning methods due to extensive use of PCA. PCA allows removal of dimensions within the original data that, once uncorrelated, contribute the least to the overall explained variance within the data, hence are the most redundant. SVD-based methods have also seen use across computer vision techniques, such as to boost inter-layer mixture of representations for multi-task learning [Gao et al., 2019], or to reduce redundancies in input training datasets for more efficient training [Xie, 2020].

Non-negative matrix factorisation, a matrix factorisation technique yielding non-negative factors, has enabled semi-supervised paradigms within applicational regions such as audio transcription through approximate labelling of weakly-labelled datasets [Chan et al., 2021]. The link between NMF's extracted basis components and K-Means clustering [Ding et al., 2005b] allows for this. Furthermore, we see extensive use within sound event detection by using NMF to pre-process audio and separate noise from real audio cues [Jang and Lee, 2020]. Basis' components are separately extracted on training sets of purely noise and purely speech data and then used as priors when compared to basis components extracted within target audio data. NMF can be useful when extracting recurring convolutional kernel components as it can bridge the link between understanding learned attention weightings to the convolutional process itself. Both techniques are well-researched with many complementary mathematical properties that align with constraints of the attention mechanisms, and therefore pose viable solutions when performing dimensionality-reduction within this project.

Chapter 3

Methods

3.1 Motivating Our Approach

The approaches by [Lindsay and Miller, 2018] and [Luo et al., 2021] in mimicking top-down goal-directed attention, exhibited by humans, in biologically-plausible HCNNs do not lead to understandable attention weightings like those found the cognitive models [Kruschke, 1990, Love et al., 2004]. Moreover, the attention weights are directly used to modulate the feature-map activations of the preceding layer, often leading to very high-dimensional attention modulations that uptake channel redundancies found within the original filter-space. Of course, it is clear to see the elimination of such redundancies when applying the attention techniques could extract low-level attention factors that in turn could lead to better interpretation of concepts developed during model training.

From a neuroscience point-of-view, the lack of sparsity in such intermediary representations can be a hindrance in relating activations of deeper network layers to later areas in the ventral visual pathway. Evidence in the literature has shown that there is a decrease of neuron densities along the ventral visual pathway, indicating a progressive reduction in the amount of information encoded at higher levels of the pathway [Charvet et al., 2013, Wilson and Wilkinson, 2015]. As such one perspective is that there is a projection of input stimuli from V1 into a low-dimensional subspace during propagation within the pathway, and this in-turn compounded with aspects of cognitive models, drives our interpretation of how the attention modulation should be implemented. Evidence to suggest this view was demonstrated in the work of [Lehky et al., 2014] in which authors showed via PCA that input stimuli with a high subspace dimensionality of $\sim$507 was effectively reduced to a subspace dimensionality of $\sim$92 dimensions within the later Inferior-Temporal cortical areas, unlike what is seen in DCNNs. An issue to consider with dimensionality-

reduction is that it often comes with a loss of information from the original full-dimensional data, however, effective dimensionality reduction is said to be achieved within the ventral pathway due to the highly correlative nature of the fundamental bases functions found in natural shapes within image components [Wilson and Wilkinson, 2015]. This means within our day-to-day vision we often consume visual stimuli with high redundancies due to these shared fundamental bases functions. When stimuli are highly correlated, effective dimensionality reduction can be achieved with minimal loss of information during propagation within the human ventral visual pathway. The literature suggests this is also the case in DCNNs, and therefore such reduction should be achievable in not only our DCNNs, but in the goal-directed attention mechanisms.

To address these problems with standard feature-wise top-down goal-directed attention, and with the motivation of faithfully implementing neurologically-plausible attention mechanisms in line with the neuroscience literature, we propose a new low-dimensional subspace based method of applying goal-directed attention, an extension to the method in [Luo et al., 2021]. This in turn will align standard feature-wise attention applications found in the literature with the idea of a low-dimensional sub-spacial projection of input stimuli when travelling from earlier to later areas of the ventral stream.

We call this method **attention seeding**, in which we hypothesise that there is a set of low-dimensional latent factors that drive the high-dimensional attention weights, which we define as **attention seeds**. If this were true, we should be able to perform a dimensionality-reduction in the trainable attention-related parameters while maintaining adequate perceptual boosts when re-configuring an existing network to specialise for the current task through a goal-directed attention mechanism. We also hypothesise that if such over-parameterisation and redundancies in the standard top-down attention mechanism were removed, then the learned lower-dimensional attention weights should pose an increased explanatory power, guiding us to more easily extract what class-and-model level factors drive the attention mechanism to achieve perceptual boosts under a goal-directed paradigm. We believe that with the lower degrees-of-freedom available with lower-dimensional attention weights, the model will exert a greater influence in warping attention weights under the direction of factors that drive the learning of attention weights. This is because attention weights are learned in a way that modulates features such that confounding classes are found to be less confusing to the model. Under a constant number of confounding classes to a target-class, decreasing degrees-of-freedom means decreasing the number of such modulations available, hence the factors that drive the learning of attention weights must act more greatly to achieve a similar level of de-obfuscation, thereby increasing the likelihood of identifying these factors during representational analysis. Uncovering these factors is paramount to reaching a

complete understanding of how attention's role inter-plays with every aspect of neural networks, and possibly within the ventral pathway.

Within the thesis, we attempt to localise which factors influence the learning of attention weights. We consider factors such as the semantic relationships between classes, class-difficulty, variance in filter activations, and the level of correlated data within filter activations elicited from grouped samples across multiple classes. We carefully selected these factors and our reasoning is presented in Section 3.4.1.

Below we outline methodology and implementation details pertaining to the newly proposed attention seeds mechanism.

3.2 Attention Model Formalism

This section first describes the general goal-directed attention layer that is used in [Luo et al., 2021]. Following this, a formalism of the process of attention seeding is given. The process performs dimensionality-reduction of filter kernels into a subspace, akin to how the ventral pathway processes stimuli, and thereafter trains and projects the attention seeds within this low-dimensional subspace. Characteristics of the subspace and a description of the dimensionality reduction techniques considered are also provided below.

We hope this more faithful application may lead to more explainable learned attention weightings by removing such channel redundancies while still enhancing performance of the DCNN in localising objects within hard-to-classify naturalistic scenes.

3.2.1 Base Hierarchical Convolutional Neural Network Model

Our contribution develops a plug-and-play parameter-flexible attention layer that can be incorporated into any DCNN with some constraints. Within this report, we use a pre-trained DCNN to incorporate our attention layer into. This is in line with the goal-directed influence the PFC plays on the ventral visual pathway when catering to a certain characteristic of the input visual stimuli in furthering ones goal. As such, this external influence does not warp the ventral visual pathway, that is, the pathway does not adapt to the current goal at hand and instead a top-down external influence is applied to the pathway via attention. Therefore, we choose to keep the pre-trained model completely frozen during experimentation, and instead use the attention layer in-place of this external influence. More on this can be found in Section 3.4.

We choose VGG-16 [Liu and Deng, 2015] as the base pre-trained DCNN to add our attention layer to. VGG-16 is a popular and relatively simple feed-forward DCNN that scores greatly within

benchmarks concerning predictivity of neural and behavioural responses along the ventral visual pathway when similar visual stimuli is shown to the network [Schrimpf et al., 2018] on neural datasets [Majaj et al., 2015, Rajalingham et al., 2018]. As such, VGG-16 makes a great base model to apply our external latent factor attention mechanism to.

VGG-16 is a 23 layer DCNN with 138,357,544 trainable parameters to be kept frozen. The model groups it's layers into five convolutional blocks that each downscale the 224x224 RGB input and extract complex features in order to sort stimuli into 1000 predefined classes as the input is propagated to the end of the network. Each convolutional block follows a stack of convolutional layers equipped with a ReLU non-linearity activation and downsized through max pooling at the end of consecutive blocks. The final two layers of the network feature fully-connected dense layers that project the final activations into the required 1x1000 logit prediction matrix. This logit matrix is then softmaxed via Equation 3.1 for the $K = 1000$ ImageNet-1k classes in order to convert the predictions into a probability vector. The maximally probable class is taken as the final prediction.

$$\sigma(\overrightarrow{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{3.1}$$

VGG-16 was pre-trained on a subset of the ImageNet dataset [Deng et al., 2009], a popular dataset comprising of over 15 million labeled naturalistic high-resolution images from around 22,000 categories. VGG-16 was submitted to ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) and as such was trained on the smaller, 1000 class ImageNet dataset comprising of around 1.3 million images and the held-out test-set of 48238 images to train and test our attention layer, respectively. We therefore use this same dataset when conducting our experimentation.

3.2.2 Standard Goal-Directed Feature-Wise Attention Layer

The standard goal-directed attention layer observed in the literature is inserted between two layers of the pre-trained DCNN in a similar fashion to Figure 3.1. The attention layer can be seen as a feature-wise modulation of the channels of its preceding layer and as such is connected in a one-to-one fashion to these channels. This means there is one weight modulation per channel along the filter axis, and this feature-wise attention is due to the assumption that the attention weight for a particular filter should be spatially invariant [Luo et al., 2021]. This modulation is defined as the Hadamard product, an element-wise multiplication between the preceding layer's activations and the attention weights.

Formally, we denote the pre-attention activation for a given image from layer $\mathbf{n}$ within a DCNN as $\mathbf{x_n}$, where $\mathbf{x_n} \in \mathbb{R}^{H \times W \times F}$ (H and W are the spatial dimensions of the representation

and F are the number of filters of the preceding layer). We denote the corresponding attention weights as $\mathbf{a} \in \mathbb{R}^{F}_{\geq 0}$. The attention mechanism modulation is mathematically defined as

$$\mathbf{x}^{*}_{\mathbf{n}} = \mathbf{x}_{\mathbf{n}} \odot \mathbf{a} \tag{3.2}$$

where $\mathbf{x}^{*}_{\mathbf{n}}$ are the post-attention activations. These attention weights were trained using gradient descent with a modified goal-directed cross-entropy loss [Luo et al., 2021], however, it was later found this is equivalent to proportioning the ratio of target-class to non-target-class samples within the training set. Our attention seeds mechanism is an extension to this standard attention layer, allowing flexibility in the number of trainable attention parameters by training attention parameters within a subspace of the full-dimensional filter-space.

3.2.3 Proposed Latent Space Attention Seed Layer

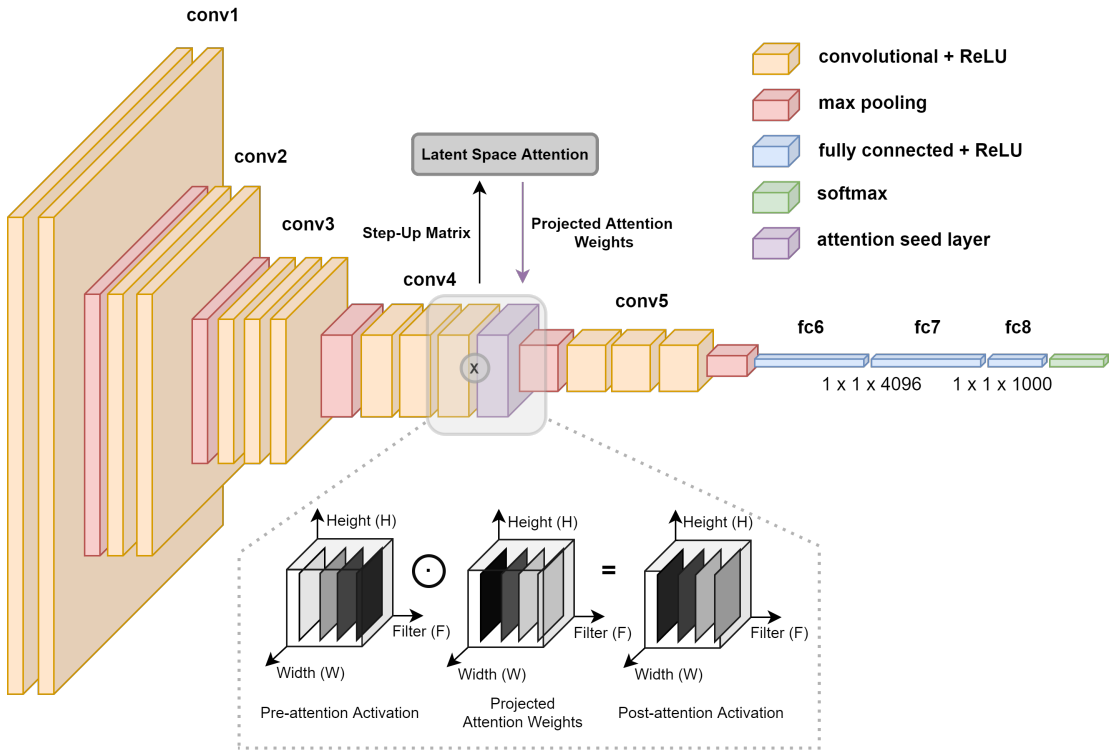


Figure 3.1: Integration of the attention seed layer with the VGG-16 model architecture. An image is presented to the input at conv1 and propagated from left to right where it is sorted into 1000 categories at the final fc8 layer. The layer preceding the attention layer outputs pre-attention activations with the shape of height×width×filters ($H \times W \times F$). The pre-attention layer has it's convolutional kernel, $\mathbf{K}$, flattened to shape $H \times W \times C$ rows and F columns, where C is the number of channels of the feature-map input to the preceding layer. The flattened kernel matrix is then factorised via matrix factorisation to obtain the dimensionality-reduced step-up matrix. The attention seed layer trains lower f-dimensional ($f << F$) latent space attention weights and utilises the step-up matrix to project the latent space attention back to the full F-dimensional filter-space, resulting in a single attention weight per filter of the pre-attention activations. The attention operation is carried out as the Hadamard product between pre-attention activations and projected attention weights. This re-weights the pre-attention activations using the corresponding projected attention weights as demonstrated in the bottom panel.

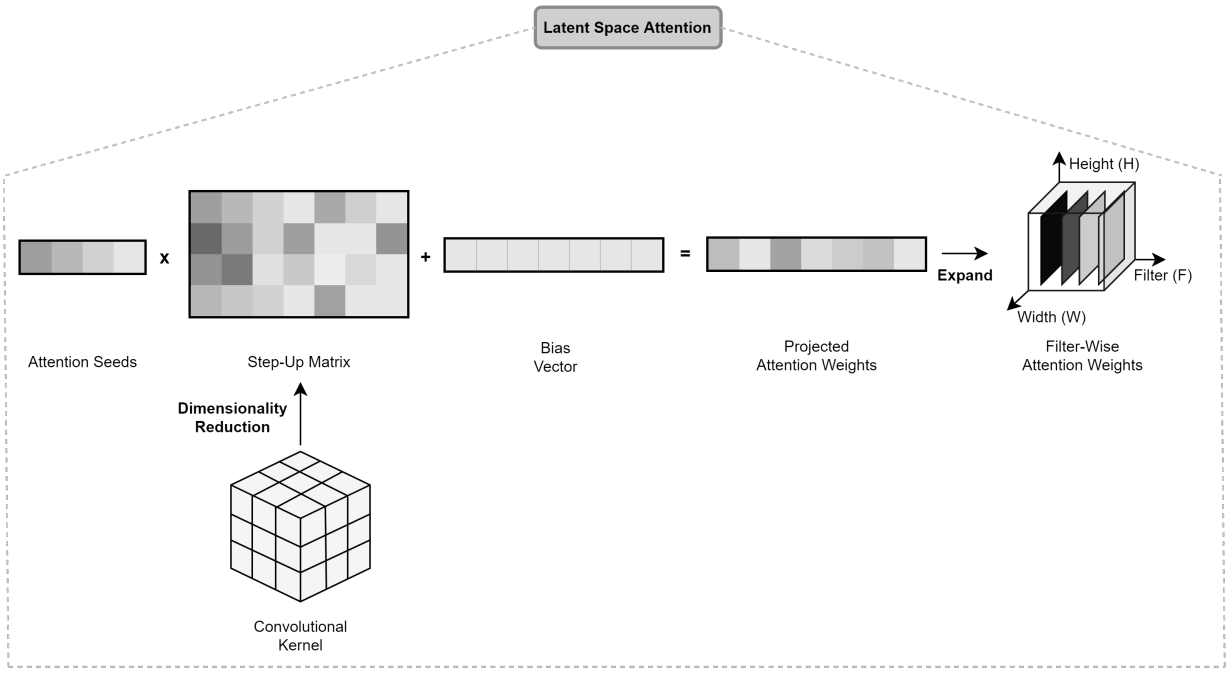


Figure 3.2: Attention seeds projection mechanism within the latent space. The low-dimensional attention seeds are the only trainable parameters within the architecture. The step-up matrix is fixed and therefore not trainable. The attention seeds, $\tilde{\mathbf{a}} \in \mathbb{R}^f$, are matrix multiplied with the step-up matrix, $\mathbf{P} \in \mathbb{R}^{f \times F}$, to yield projection attention weights, $\mathbf{a} \in \mathbb{R}^F$, of equal dimensionality to the preceding layer's activation channels. The expand function just indicates the projected attention weights map in a one-to-one fashion to each filter channel of the preceding layer.

We defined the *attention seeds* as a set of low-dimensional latent factors operating within a low-dimensional subspace that drive the high-dimensional attention weights towards the current classification goal of the network. Formally, the *attention seeds* are defined as a low-dimensional vector, $\tilde{\mathbf{a}}$, where $\tilde{\mathbf{a}} \in \mathbb{R}^f$ ($f << F$), where F is the number of sets of filters the preceding convolutional layer contains.

Instead of training the attention weights directly, attention seeds are trained within this principal subspace defined by the preceding layer kernel within our model. We define an affine transformation between attention seeds and attention weights, in which the model learns to adaptively and externally influence the output of a neural network based on the current categorical goal of the network via a projection of the latent subspace attention seeds back into the original filter dimensions of the preceding layer (see Figures 3.1 and 3.2 for more details). More formally, the transformation is given as

$$\mathbf{a} = (\tilde{\mathbf{a}}\mathbf{P})^T + \mathbf{b} \tag{3.3}$$

where $\mathbf{P}$ is a fixed linear transformation ($\mathbf{P} \in \mathbb{R}^{f \times F}$) which we denote as a step-up matrix

because it transforms low-dimensional attention seeds to high-dimensional attention weights. The step-up matrix is fixed and therefore not trainable. The bias term denoted $\mathbf{b} \in \mathbb{R}^F$, is a fixed vector initialised to 1.0 in order to facilitate training from the pre-trained state of the model. Furthermore, we place a non-negative constraint over the projected attention weightings, $\mathbf{a}$, in agreement with non-negative modulation from the neuroscience literature.

The step-up matrix is derived by applying dimensionality reduction on the kernels of the convolution layer preceding the attention layer (bias terms are not used). Formally, the convolution kernels are denoted $\mathbf{K} \in \mathbb{R}^{H \times W \times C \times F}$ where each kernel has size $H \times W \times C$, where C is equal to the input channels of the convolutional layer, and there are F sets of filters that are applied to each input feature-map. In order to achieve dimensionality reduction, we first must flatten $\mathbf{K}$ along the first three dimensions, such that $\mathbf{K}$ becomes a 2D matrix with $H \times W \times C$ rows and F columns. Generally, performing a dimensionality-reduction involves some sort of matrix factorisation on the data matrix. This is typically found to follow the following general structure, and as such performing this on $\mathbf{K}$ we obtain

$$\mathbf{K} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{3.4}$$

Preserving the top f principal components, we obtain a low-dimensional approximation of Equation 3.4:

$$\tilde{\mathbf{K}} = \tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}\mathbf{P}^T \tag{3.5}$$

where $\mathbf{P}$ is the step-up matrix that bridges attention seeds in a low-dimensional subspace to attention weights in a high-dimensional space, concluding the attention seeds layer formalism. More information on the meaning of dimensionality reduction and it's connection to the subspace projection can be found below in Section 3.3.

As seen in Figure 3.1, this attention layer can be inserted after any convolutional layer in order to give access to its kernel for extraction of the step-up matrix for the affine projection (see Figure 3.2). Here, we insert the attention seeds layer after the final convolutional layer of convolutional block four, within the middle region of the network. The middle region was chosen for placement due to successes of mid-level attention demonstrated in [Luo et al., 2021] and is in accord with neuroimaging results [Ahlheim and Love, 2018] and insights drawn from the literature. Furthermore, mid-level layers of DCNNs typically exhibit the most uniform representational similarity of centered layer kernel alignments with all other network layers, and as such performing the decomposition at the mid-level should lead to consistent results applicable to most network layers with little permutation of propagated representational stimuli [Raghu et al.,

2021]. This mechanism is particularly exciting as previously, top-down attention mechanisms were always constrained to be of equal dimensionality to the preceding convolutional layer's output channels. This is no longer a constraint and any number of seeds (trainable attention weights) can be permitted when training within the latent attention space.

3.3 Dimensionality Reduction

In this section we detail the different techniques considered in performing the dimensionality reduction of the convolutional kernel, $\mathbf{K}$, in order to obtain a valid step-up matrix, $\mathbf{P}$, for the attention seeds projection.

High-dimensional data, such as image data or feature-maps, often exhibit a convex hull that lies in a small subset of the original data space, that is they lie on a manifold within this high-dimensional space. Dimensionality reduction is a change of variables; moving from the original coordinate system of the high-dimensional space to a reduced coordinate system of the data manifold. Mathematically, we look for a low-dimensional approximation such that

$$\mathbf{A} \approx f(\mathbf{A}, \mathbf{h}, \theta) \tag{3.6}$$

where the reduction is a function of the original data $\mathbf{A}$, a low-dimensional hidden variable $\mathbf{h}$, and a set of parameters θ. A comparison of the two main techniques considered is presented in order to show a fair evaluation of how each technique's properties aid in solving our problem statement within constraints to the attention seeds formalism.

3.3.1 Problem Statement

We outline below the technical requirements during projection of the latent attention seeds into the original dimensional space. An affine transformation via an affine functional projection (Equation 3.3) is required to project the latent attention seeds in low dimension back into the original high-dimensionality subspace of filters. An affine function is a special composition of a linear function with a translation. For two vector spaces X and Y of dimensions M and N respectively, considering functions that map space X to space Y, $f : X \to Y$, then f is affine if $f(x) = \mathbf{A}\mathbf{x} + \mathbf{b}$ for some $N \times M$ matrix $\mathbf{A}$. Note, matrix multiplication of the form $f(x) = \mathbf{F}\mathbf{x}$ with constant matrix $\mathbf{F}$ and variable vector $\mathbf{x}$ constitutes a linear transformation on the vector $\mathbf{x}$. The multiplication property of a matrix transpose means that the transformation Equation 3.3 is indeed affine with projection from the low-dimensional subspace in place of X to the original

high-dimensional space in place of Y, and hence follows this constraint. Affine transformations preserve lines, parallelism and the ratios of lengths of parallel line segments as well as the dimension of any affine subspaces as it maps an affine space onto itself, with added degree of freedom of the translation component, hence making it an appropriate choice for projection of the attention seeds.

A second soft-constraint is that the step-up matrix is required to be non-negative. This is to aid in ensuring projected attention weights will be non-negative in line with non-negative neuron modulation within the brain, as well as the requirement that the decomposition of the convolutional kernel, and as such the step-up matrix output from such decomposition, should be an additive combination of the basis/principal components derived from the kernel. This is considered a soft-constraint as it does not prevent learned attention weightings from being non-negative. To overcome this, a non-negative constraint directly enforced on the learnable attention seeds as outlined later on, however it does aid in interpretability of the decomposition into the low-dimensional subspace.

3.3.2 Singular Value Decomposition (SVD)

At the heart of many dimensionality reduction techniques like PCA lies singular value decomposition (SVD). SVD is a factorisation of an $M \times N$ matrix $\mathbf{A}$, such that

$$\mathbf{A} = \mathbf{U}\mathbf{L}\mathbf{V}^{\mathbf{T}} \tag{3.7}$$

where $\mathbf{U}$ is a $M \times M$ orthogonal unitary matrix, $\mathbf{L}$ is a $M \times N$ diagonal matrix of decreasing singular values $\mathbf{s_i}$, where the number of singular values gives the rank of the original matrix, and $\mathbf{V}$ is an $N \times N$ orthogonal matrix of right singular vectors that dictate the principal directions of the original matrix $\mathbf{A}$ [Prince, 2012].

The cumulative effect of the transformations applied on data points by the decomposed matrices is straightforward. The matrix $\mathbf{V}^{\mathbf{T}}$ rotates and reflects the original points onto the principal directions of decreasing variance with scaling dictated by the magnitude of the corresponding singular value within the matrix $\mathbf{L}$. Finally, the matrix $\mathbf{U}$ rotates the resultant mapped transformation.

3.3.3 Principal Component Analysis (PCA)

Principal component analysis performs dimensionality reduction on a large set of features into a smaller set of features by applying a transformation in such a way that linearly correlated features are transformed into uncorrelated features. We utilise PCA extensively when performing experimentation on explained variance of task-set activations later on. Given p n-dimensional vectors $[x_1, ..., x_p]$ that form an $n \times p$ data matrix, $\mathbf{X}$, whose jth column is the vector $\mathbf{x}_j$ of samples of the jth variable, the decomposition extracts a linear combination of the columns of $\mathbf{X}$ with maximum variance given by $\sum_{j=1}^{p} a_j \mathbf{x}_j = \mathbf{Xa}$, where $\mathbf{a}$ is a vector of constants $a_1, a_2, ..., a_p$. The variance is given by $var(\mathbf{Xa}) = \mathbf{a^T Ca}$, where $\mathbf{C}$ denotes the dataset sample covariance matrix. As such maximising variance is equivalent to finding a p-dimensional vector $\mathbf{a}$ that maximises the quadratic form $\mathbf{a^T Ca}$. This can be expressed through a Lagrangian optimisation problem with unit norm constraint on $\mathbf{a}$, $\mathbf{a^T a} = \mathbf{1}$, and via induction leads to

$$\mathbf{Ca} = \lambda \mathbf{a} \tag{3.8}$$

an eigenvalue problem. To find the optimal vector, we compute the SVD

$$\mathbf{C} = \mathbf{VLV^T} \tag{3.9}$$

where $\mathbf{V} \in \mathbb{R}^{n \times n}$ is a column matrix of eigenvectors and $\mathbf{L} \in \mathbb{R}^{n \times p}$ is a diagonal matrix of eigenvalues, λ_i in decreasing order. These eigenvectors are the principal axes of the data, with projections of the data onto these principal axes given by $\mathbf{XV}$. The top f principal components, $\mathbf{a} = [a_1, ..., a_f]$, are then found by taking the first f columns of $\mathbf{V}$.

By centering the columns of $\mathbf{X}$, that is subtracting the mean of each column for each data point within that column, we obtain $\mathbf{X}^*$ such that for any element $x^*_{ij} = x_{ij} - \overline{x}_j$, where $\overline{x}_j$ denotes the column mean, we can extend PCA more generically by decomposing the data matrix $\mathbf{X}^*$ directly instead of its covariance matrix. This connection is possible as the covariance matrix is connected by

$$\mathbf{C} = \frac{(\mathbf{X}^*)^\mathbf{T} \mathbf{X}^*}{(n-1)} \tag{3.10}$$

As such for any arbitrary data matrix, $\mathbf{X}$, including our kernel matrix $\mathbf{K}$, PCA is a decomposition through SVD such that

$$\mathbf{X}^* = \mathbf{ULV^T} \tag{3.11}$$

then the columns of $\mathbf{V}$ are the principal axes and the columns of $\mathbf{UL}$ are the principal components

of the data.

When applied to our flattened kernel matrix, $\mathbf{K}$, to reduce the dimensionality from F to $f < F$, we take the first f columns of $\mathbf{U}$, the $f \times f$ upper-left square subset of $\mathbf{L}$ and the first f rows of $\mathbf{V^T}$. Reconstruction of the original matrix leads to a lower-rank approximation of the original data matrix

$$\mathbf{K_f^*} = \mathbf{U_f L_k V_f^T} \tag{3.12}$$

This means that we can take $\mathbf{V_f^T}$ as our step-up matrix of principal axes enabling projection of the attention seeds back into the original space, $\mathbf{P} = \mathbf{V_f^T}$. This orthogonal matrix factorisation technique, if used as the step-up matrix, most closely resembles an attention seeds mechanism under the ALCOVE cognition model from the neuroscience literature due to its assumption of orthogonal feature representations.

3.3.4 Semi Non-Negative Matrix Factorisation (SNMF)

Similarly to PCA, non-negative matrix factorisation (NMF) approximates an $p \times n$ non-negative data matrix $\mathbf{X}$ with a lower-rank approximation such that

$$\mathbf{X}_+ \approx \mathbf{F}_+ \mathbf{G}_+^\mathbf{T} \tag{3.13}$$

where $\mathbf{F} \in \mathbb{R}^{p \times f}$ is a column matrix of basis components found within the data, and $\mathbf{G} \in \mathbb{R}^{f \times n}$ is a column matrix of linear-weightings for each basis component and dictates how to reconstruct an approximation of $\mathbf{X}$ from the basis components in $\mathbf{F}$. NMF performs dimensionality reduction on p original dimensions to $f < p$ lower-rank dimensions.

NMF is especially useful on textual or image-based datasets due to the non-negative constraint. NMF yields non-negative factors and as such is very advantageous from the point of view of interpretability due to its connection to K-means clustering through the least-squares objective function of NMF [Ding et al., 2010, Ding et al., 2005b, Ding et al., 2005a]. As such each basis component can be thought of as a principal component or cluster centroid, in similar respect to PCA, and these are then scaled and distributed by membership indicators found in $\mathbf{G}$ when performing the reconstruction, making NMF a valuable decomposition technique due to the non-negative soft constraint placed on the step-up matrix $\mathbf{P}$.

An issue with NMF is that it can only decompose non-negative matrices, in which the convolutional kernel is strictly not. To address this, we use Semi-NMF (SNMF) [Ding et al., 2010, Wu and Wang, 2014], a variant of NMF in which the constraint on the data matrix $\mathbf{X}$ being non-negative is relaxed, meaning the factorisation still yields a non-negative matrix $\mathbf{G}$ of latent

weightings, however, $\mathbf{F}$ is no longer constrained to also be non-negative:

$$\mathbf{X}_{\pm} \approx \mathbf{F}_{\pm}\mathbf{G}_{+}^{\mathbf{T}} \tag{3.14}$$

Unlike PCA and SVD-based methods, NMF and SNMF are non-deterministic approximations, though NMF in particular assumes a deterministic framework, and is therefore usually very consistent in its factorisations. SNMF is trained on the data matrix under a K-means clustering like least-squares objective function:

$$O_{K-means} = \sum_{i=1}^{n}\sum_{k=1}^{K} g_{ik} \left\|\mathbf{x}_i - \mathbf{f}_k\right\|^2 = \left\|X - FG^T\right\|^2 \tag{3.15}$$

where $\|\mathbf{v}\|$ denotes the L_2-norm of an arbitrary vector $\mathbf{v}$, and $\|A\|$ denotes the Frobenius norm of an arbitrary matrix A. As can be seen the K-means objective function can be viewed as a matrix approximation objective function of the original data matrix as shown by the right-hand side of Equation 3.15. From this objective function, initialising the $\mathbf{F}$ and $\mathbf{G}$ matrices, one can derive an alternating update rule that iteratively approximates the data matrix under the required constraints. Initialising $\mathbf{G}$ through a K-means clustering yields cluster indicators where $\mathbf{G}_{ik} = 1$ if $\mathbf{x}_i$ belongs to cluster k else $\mathbf{G}_{ik} = 0$. A small constant is then added to all elements of $\mathbf{G}$. Then update $\mathbf{F}$ through

$$\mathbf{F} = \mathbf{X}\mathbf{G}(\mathbf{G}^{\mathbf{T}}\mathbf{G})^{-\mathbf{1}} \tag{3.16}$$

While fixing $\mathbf{F}$, then update $\mathbf{G}$ through

$$\mathbf{G}_{ik} \leftarrow \mathbf{G}_{ik}\sqrt{\frac{(\mathbf{X}^{\mathbf{T}}\mathbf{F})_{ik}^{+} + [\mathbf{G}(\mathbf{F}^{\mathbf{T}}\mathbf{F})^{-}]_{ik}}{(\mathbf{X}^{\mathbf{T}}\mathbf{F})_{ik}^{-} + [\mathbf{G}(\mathbf{F}^{\mathbf{T}}\mathbf{F})^{+}]_{ik}}} \tag{3.17}$$

where the superscript $+$ and $-$ indicate taking the positive and negative elements, respectively.

The factorisation of the data matrix into the $\mathbf{F}$ and $\mathbf{G}$ matrices, in the sense of SVD and PCA, is equivalent to obtaining the $\mathbf{U}$ and $\mathbf{V}$ matrices, however, with lack of an orthogonality constraint [Ding et al., 2006]. As such, like before, we can fix the step-up matrix for projection of the attention seeds back into the original space, as $\mathbf{P} = \mathbf{G}_{+}^{\mathbf{T}}$

Within this report, we use PyMF [Erler et al., 2018], a python matrix factorisation module with several constrained and unconstrained matrix factorisation methods, and more specifically, to perform the SNMF decomposition outlined above and in [Ding et al., 2010]. PyMF implements the iterative SGD-based update procedure outlined above to perform the SNMF decomposition under

random initialisation of matrices, which was found to be optimal in reducing the reconstruction error [Velten de Melo and Wainer, 2012]. This non-orthogonal matrix factorisation technique, if used as the step-up matrix, most closely resembles an attention seeds mechanism under the SUSTAIN cognition model from the neuroscience literature due to its assumption of correlated feature representations.

When using NMF it is advantageous to normalise the resultant decomposition. NMF outputs non-unique decompositions and as such different results can be obtained from each run leading to different interpretations of cluster assignments when referencing the K-means link. Normalisation of the resultant matrices enables comparability and alleviates this problem [Yang and Seoighe, 2016]. NMF is invariant with column scaling on $\mathbf{F}$ and row scaling on $\mathbf{G}$. If we consider a diagonal matrix $\mathbf{S}$ with size $f \times f$, then

$$\mathbf{F}_{\pm}\mathbf{G}_{+}^{\mathbf{T}} = \mathbf{F}_{\pm}\mathbf{S}^{-\mathbf{1}}\mathbf{S}\mathbf{G}_{+}^{\mathbf{T}} \tag{3.18}$$

Therefore we can use $\mathbf{S}$ to normalise columns of $\mathbf{F}$ through $\mathbf{F}' = \mathbf{F}_{\pm}\mathbf{S}^{-\mathbf{1}}$, and equivalently, normalise rows of $\mathbf{G}$ through $\mathbf{G}' = \mathbf{S}\mathbf{G}_{+}^{\mathbf{T}}$. [Yang and Seoighe, 2016] states the L-∞ norm yields the best normalisation result when seeking optimal cluster assignments. We therefore populate the diagonal matrix as $\mathbf{S}_{jj} = |\mathbf{F}_{ij}|_{\infty} \, \forall i$, meaning each diagonal component is equal to the max absolute value of its corresponding column in $\mathbf{F}$. The normalised version, $\mathbf{G}'$ is therefore used in all experiments henceforth.

3.3.5 Comparison of SNMF and PCA as Decomposition Techniques for the Projection Matrix

SNMF and PCA both pose viable solutions for dimensionality reduction of the convolutional kernel matrix and both give lead to slight differences in the projected attention weightings and there technical meanings.

In PCA, via SVD, we obtain the factorised $\mathbf{U}\mathbf{L}\mathbf{V}^{\mathbf{T}}$ matrices, in which the $\mathbf{U}\mathbf{L}$ matrices hold the principal components and $\mathbf{V}$ are the principal axes, however in SNMF the same projection matrix does not necessarily contain the same information - the rows are not principal axes' but instead are weightings of the fundamental basis elements found within the original data. A connection can be made between SVD and SNMF in that SNMF yields the SVD equivalent $\mathbf{U}$ and $\mathbf{V}$ matrices but does not give further insight into the eigenvalue composition for each basis element given by the singular value $\mathbf{L}$ matrix, and does not specifically give orthogonal basis elements. As such we lose information as to the overall importance of each basis component

reflected in the original data and can only interpret how each basis component contributes within each dimension separately.

Although the eigenvalue composition is lost, SNMF with its non-negative $\mathbf{G}_+$ weighting matrix gives better interpretability of the contribution each basis component gives in reconstruction, and more specifically its interpretability for learned attention seed weightings through positive linear combinations of the low-dimensional latent attention factors.

SNMF does suffer from potential initialisation error due to the non-deterministic update-rule based factorisation. It is known that the best initialisation for the $\mathbf{F}$ matrix that lead to the lowest reconstruction error are random column and random initialisation [Velten de Melo and Wainer, 2012] and this is also used within the PyMF library. Preliminary experiments decomposing an arbitrary matrix, $\mathbf{A}$, with 30 components via PCA and then SNMF showed the reconstruction error, as given by the residual sum of squares of errors

$$error_{rss} = \sum_{i=1}^{n} \sum_{j=1}^{p} (x_{ij} - \tilde{x}_{ij})^2 \tag{3.19}$$

where $\tilde{x}_{ij}$ denotes the reconstructed data matrix entry from either decomposition technique, near identical reconstruction errors at 93.6 and 94.8, respectively and as such we rule out initialisation error as a potential deciding factor.

When multiplying the attention seeds with the step-up matrix, under a PCA-derived regime this is equivalent to a projection of the attention seeds onto the top f largest variance principal axes with preserved uncorrelation of orthogonality across axes and therefore gives uncorrelated weightings in the original dimension space aiding in interpretability.

From an interpretability standpoint the step-up matrix is preferably constrained to being non-negative. Through an SNMF-derived regime, the step-up matrix represents how much each principal component found within the data is represented (or assigned membership) to each of the 512 original filters in the D-dimensional space. This leads to an interesting interpretation for the projection of the seeds: under an SNMF-based regime this would be equivalent to projecting the attention seeds in such a way that they are weighted based on how much each basis (principal) component was found in the original space (for each filter). With the connection between SNMF and K-means clustering, this would be equivalent to allocating seeds by modulating K-means cluster (or principal component) assignments within the original data; resulting in proportional projection into the 512-dimensional filter space with each filter owning its own independent set of uniquely weighted seeds for each cluster. Considering both factorisation techniques, we choose to use SNMF as the matrix factorisation technique powering our dimensionality-reduction within

the attention seeds mechanism as it best fits our technical constraints. This means our alignment of the standard top-down goal-directed attention mechanism is that most similar to the SUSTAIN cognitive model.

3.4 Attention Training

The base model, VGG-16, can be thought of as a function approximator, more specifically it emulates the underlying distributional function, f, that takes as input an image, $x \in \mathbb{R}^{224\times224\times3}$, and predicts via a max decision rule on the Softmax output the most likely ImageNet class, $c \in \{1, 2, ..., 1000\}$, the image belonged to

$$f(x) = \underset{c}{\text{argmax}}\ p(c|x) \tag{3.20}$$

Decomposing the overall VGG-16 function into two parts, we have the convolutional layers function, f_c, that transforms the input x to a latent representation, $z \in \mathbb{R}^{14\times14\times512}$. The final fully-connected layers, f_{fc}, map z to a probability vector through $p(c|x)$. Mathematically that is,

$$f_c(x) = \mathbf{z} \qquad f_{fc}(\mathbf{z}) = p(c|x)$$

The latent space attention is a multiplicative one-to-one modulation of the mid-level latent representations, z, of the preceding layer by the set of non-negative attention weights, $\mathbf{a} \in \mathbb{R}^{512}_{\geq0}$. This augments the VGG-16 function inducing a conditional dependency on the attention seeds themselves, $\tilde{\mathbf{a}}$, and the step-up matrix, $\mathbf{P}$, giving

$$\hat{f}(x, \tilde{\mathbf{a}}, \mathbf{P}) = p(c|x, \tilde{\mathbf{a}}, \mathbf{P}) = f_{fc}(\mathbf{a} \odot \mathbf{z}) \tag{3.21}$$

In particular we target the latent representation of the fourth convolutional block, $\mathbf{z} \in \mathbb{R}^{14\times14\times512}$, and compute the Hadamard product with the non-negative projected attention weights, $\mathbf{a}$, in accordance with Equation 3.3 giving,

$$(\mathbf{a} \odot \mathbf{z})_{i,j,k} = a_k z_{i,j,k} \tag{3.22}$$

Figure 3.2 shows the latent space attention layer mechanism in pictorial format.

3.4.1 Choosing Target Classes via NMF, PCA and Clustering

In order to facilitate a high-quality and interesting analysis of factors that influence the attention weights learning process, we aim to select target classes that cover a variety of model-level cues and class decomposition statistics such as variance and redundancies of response activations across class stimuli, difficulty of mastering the class and semantic entanglement of classes, all while also being mindful of higher-order class groupings, known as superordinates/super-classes. Superordinates are groupings of classes with consistent shared themes among the examples and are used such that chosen target classes have some sort of underlying basic relationship established which we hope leads to a higher standard of comparisons.

The factors used to select target-classes were considered as they each are derived from different domains of class and model interactions. We select semantic entanglement as it is a more superficial and external influencing factor. We select class difficulty as it is a model-level factor most closely tied with the perceptual effects of goal-directed attention [Smith et al., 2021]. Finally, we consider redundancies and variance of activations as a class-level factor that takes into account the uniqueness of responses from the convolutional kernel itself. If any of these factors are found to drive learned attention weights, future experiments can then focus on subsets of features most closely associated with domain that factor is derived from. We compound these factors by carefully selecting target-classes following two consecutive analyses outlines below.

Preliminaries: What are Target Classes?

Goal-directed attention means the model must be trained to specialise towards a specific task e.g trying to identify if a dog is present within an image. A target class, $\mathbb{T}$, is the set comprising all samples belonging to a single ImageNet class that the attention network specialises towards. Each target class therefore implies a different task of the network.

Preliminaries: Extracting Ground-Truth Semantic Label Embeddings

With semantic entanglement between target-classes being a factor we investigate, we must therefore extract semantic embeddings for the class labels to be used within the representational analysis later on. These semantic embeddings are only used to select the target-classes and within the representational similarity analysis as ground-truth relationships between the target-classes that we then look for in the low-dimensional attention weights.

Each ImageNet class name was converted into a embedding vector, $\mathbf{v} \in \mathbb{R}^{768}$, using BERT, a powerful language representation model that produces high-quality word embeddings through

training on large language corpus' [Devlin et al., 2018]. Firstly, class names are converted into tokens through the BERT tokenizer, a format of textual tokens that the model can understand. Then, the BERT encoder is used to convert the tokens into a latent representation in the form of a word embedding vector. For example two very semantically similar classes may exhibit highly correlated attention weightings in low-dimensionality. As such we choose our target class groupings in a way that is informed by the semantic and relative groupings of ImageNet classes. In order to facilitate a high quality comparison, choices for our target classes should encompass a wide variety in inter-group class decomposition and class statistics in such a way that the potential learned weightings can aid in interpretability within groupings. Previous research also suggests that classes should be chosen based off model-level variables such as difficulty in predicting classes as opposed to superficial visual cues [Smith et al., 2021]. We wish to also take into consideration the variation in filter activations output from the convolutional layer as we wish to investigate how this class-level variable will influence learned attention seed weightings.

Superordinate NMF-PCA Analysis

This experiment aims to extract the variance of filter activations elicited from superordinate samples passed through the network.

The provided superordinates list comprises 6 super-class groupings of a sub-sample of the ImageNet-1k class labellings. The superordinate classifications are as follows: *avian animals*, *canidae animals*, *felidae animals*, *kitchen objects* and *land-transport objects*. This experiment utilises NMF on the activations from each superordinate grouping. The intuition for this experiment is derived from the connection between NMF extracted basis components (topics) and K-means clustering centroids [Ding et al., 2005b], in which we effectively use NMF to extract themes (basis components or clusters) from activations and use PCA to compute the similarity of themes across all samples from the superordinate via explained variance of the first principal component. We wish to see how the variation of recurring themes extracted from these activations within superordinate groupings affect the overall mean superordinate class accuracy. If the same theme is found within many activations of a superordinate grouping, then the grouping is said to be highly correlated, and therefore contains less variance within activations. This experiment compounds analysis of not only the variance of activations, but also class difficulty and incorporates the underlying basic relationships of classes through their superordinate grouping. We then select two groupings that exhibit the greatest variety in mean class-difficulty and variance of correlations (themes) found within the activations.

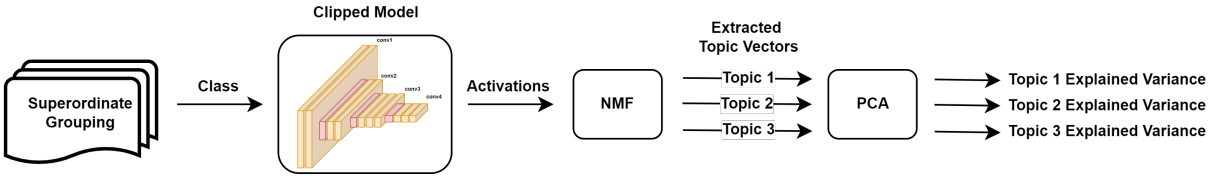


Figure 3.3: Flow-chart representation of the NMF-PCA experiment. This experiment used NMF on the superordinate activations output by the preceding layer to the attention seeds mechanism. The NMF extracts basis components (recurring themes) from the activations and then a one-eigenvalue PCA is used as a form of multi-dimensional cosine similarity on the covariance between all basis components, yielding the explained variance of each basis component (topic) across all classes in the superordinate. This will quantify how similar the activations are at a fundamental level across all classes. All class samples from the superordinate are passed through the (clipped) VGG-16 model to get the mid-level activations. We then use a three-topic NMF on the activations to extract the topic vectors found within the activations. Finally, a one-eigenvalue PCA extracts the variance found within the correlation matrix of the topic vectors.

The experiment is as follows: for each superordinate grouping, each class is fed through the network and the activational output from the end of the fourth convolutional block (the layer preceding the attention layer) is computed. The activations are of shape [class size$\times 28 \times 28 \times 512$] and are flattened to form a matrix of shape [class size, $28 \times 28 \times 512$] which NMF expects. We decompose this matrix into 3 basis components (or alternatively 3 K-means centroids) yielding topic vectors stored in the W matrix and the complementary activation/weighting matrix stored in the H matrix. Using 3 topics provided good separation of recurring basis components within the weighting matrix, that is basis components were found to be more distinct amongst each other, while not being too computationally expensive. We concatenate each class's topic vectors by topic, yielding 3 separate matrices i.e the first topic matrix holds all classes within the superordinate grouping's topic 1 vectors. The next step uses a one-eigenvalue PCA to compute the covariance of the respective topic vectors within the superordinate grouping to quantify their similarity. This is considered as a form of multi-dimensional cosine similarity between all classes topic vectors. PCA is therefore used on each of the three topic matrices to compute the explained variance of the first principal component $\in$ [0,1], where if it is close to 1 the topic vectors of all the classes within a superordinate are very correlated and therefore decomposed very similarly. We therefore end up with 3 explained variance values for each topic within the superordinate. This gives us a more robust measure of similarity in filter activations as the NMF groups activations into clusters based on similarity of basis components. The one-eigenvalue PCA then quantifies this as a single value. This is repeated for all groupings and each topic is plotted against the superordinate mean class accuracy in order to establish a relationship between variance of activations and the overall

superordinate difficulty. We then select two superordinate groupings that yield the greatest variety in both to train our attention models on. We only select two groupings in order to balance the computational expense of experimentation as many models are being trained for each target-class. Results for this experiment can be found in Section 4.1.

Semantically Entangled Classes through Community Detection Clustering

This experiment aims to identify which classes within the selected superordinate groupings are the most semantically similar. This therefore means chosen target-classes compound all three investigated factors together.

The superordinate groupings contain a large quantity of classes and we need to balance computational cost and time expenses by choosing a subset of target classes from both. We note the superordinates contain many classes that are not strongly semantically entangled. To overcome this we utilise community detection clustering [Blondel et al., 2008].

We convert the class labels and their respective word embedding similarity scores into a graph structure, $G = (V, E)$, where V is the set of vertices representing the nodes and E is the set of edges pairing nodes. The set of classes within a superordinate grouping, denoted s, and corresponding set of class-label semantic embeddings, denoted k, represent the nodes, $V = s$. The cosine similarity scores between embeddings of a class and all other remaining classes act as weights on edges, e, between the nodes. For two nodes, $\mathbf{x}$ and $\mathbf{y}$, and their corresponding semantic vectors, $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$, an edge, e, is defined as

$$e = xy = \frac{\tilde{\mathbf{x}} \cdot \tilde{\mathbf{y}}}{|\tilde{\mathbf{x}}||\tilde{\mathbf{y}}|}, \ \forall \tilde{\mathbf{x}}, \tilde{\mathbf{y}} \in k \ \text{ s.t } \ x \neq y \text{ and } x \leftrightarrow y$$

We then perform community detection clustering via optimal bipartite cuts made along the graph which maximise the modularity using Louvain heuristics [Blondel et al., 2008]. This is implemented through the NetworkX Python module [Hagberg et al., 2008]. Modularity measures the strength of division of a network into modules by the relative density of edges inside communities with respect to edges outside communities. It is measured on a scale between values of 0.5 (non-modular clustering) and 1 (fully modular clustering). More explicitly, the algorithm identifies the highest partition of the dendrogram generated by the Louvain algorithm. This extracts classes that are more highly semantically connected due to the edge weights equalling to the cosine similarity between class semantic embeddings. The cluster with the highest mean cosine similarity score forms the target-classes for that superordinate grouping. We therefore end up with two sets of classes from each superordinate that are highly semantically correlated

while also maintaining the variety in model-level cues such as class difficulty and variance of activations identified within the NMF-PCA experimentation. These two are merged together to form the target-class set, T, used henceforth.

Later on through the representational similarity analysis, we evaluate how sensible learned attention weights are by comparing dissimilarities between the attention weights to dissimilarities between the ground-truth semantic embeddings. We therefore attempt to extract a sensible semantic space from the attention weights. By selecting target-classes with a variety of the aforementioned factors we hope to link identified trends to the factors aiding us in uncovering which factors influence attention weights to be learned how they are and which factors do not.

3.4.2 Choosing Seed Dimensionality via PCA

In this experiment we look to identify the set of dimensions we will evaluate the attention seeds mechanism at.

The literature surrounding dimensionality reduction has revealed large redundancies in convolutional layer kernels through techniques like PCA by de-correlating high dimensional data into fewer independent principal components that wield large explanatory power. Motivated by this, the set of chosen dimensions, D, to analyse was decided through PCA on grouping activations. Each target class was fed through the model to obtain the latent output of the final layer of the fourth convolutional block, the layer preceding the attention layer. PCA was then run on the collection of samples consisting all target classes to identify the required dimensions for 50%, 75%, 85% and 95% explained variance within the latent representations. This method essentially reduces redundancies to identify the minimal required dimensions that maintain certain percentages of informational content of the representations. The identified dimensions were 17, 30, 35 and 42 dimensions, respectively. Furthermore, to consider the effect of higher dimensions and more degrees of freedom within learned weights we also consider choices of 90, 256 and the original 512 dimensions. The final chosen set, D, of dimensions to run training on consists of $[17, 42, 90, 256, 512]$. Choosing different dimensions allows us to analyse whether difficulty of classes or variation in latent activations affects the required number of seeds to achieve a perceptual boost via attention, as well as evaluate how dimensionality of the seeds mechanism affects performance and the representational geometries of attention weights.

3.4.3 Training Setup

Attention seeds are trained on a 1000-way classification while keeping the ImageNet pretrained VGG-16 model, f, as a fixed function, that is the attention seeds are the only trainable parameters within the attention model. The step-up matrix is a fixed matrix and therefore not trainable. A model trained using a fixed number of attention seeds for a target class at a specific attention intensity level is defined as an attention model. The training objective consists of minimising the multi-class cross-entropy (CE) loss between model predictions and the ground-truth dataset labels. The training set comprises 90% randomly sampled images from each class with the remaining 10% used as a validation set. The white-listed version of the ImageNet validation set is used as the test-set. The attention seeds are updated through the Adam optimiser [Kingma and Ba, 2014] at a learning rate of 0.05, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-7}$ with gradients derived on a mini-batch size of 128 examples. Images are pre-processed in line with standard VGG-16 pre-processing techniques [Liu and Deng, 2015]. Unlike [Luo et al., 2021, Smith et al., 2021] we do not use a maximum epoch count of 1000 with a stopping criterion based on the CE validation loss. Due to the nature of target-class training, the validation loss across the ImageNet validation set naturally increases as attention weights specialise towards the task-set of the model, therefore making the use of a stopping criterion redundant. For our evaluation case, we find it best to fix the epoch cycle count in order to enable a more fair comparison between learned weightings. Preliminary experimentation surrounding performance saturation led us to fix the epoch count to 25 epochs, with a learning rate scheduler that decays the learning rate beyond 15 epochs by $e^{-0.01}$ for each epoch thereafter. The use of a learning rate scheduler was found to aid in the consistency of convergence of attention seeds at a finer granularity offered by a lower learning rate.

We train attention models for each target class at each set dimensionality in D. Finally, task-generic baseline models were also trained at each dimensionality.

Task-Generic Baseline

The task-generic baseline models have their attention weights trained on examples from all 1000 ImageNet classes, and as such results in seed weights that are non task-specific and are optimised for ImageNet as a whole. We train the task-generic models for all dimensions to provide a reference point for performance and other evaluations relating to the attention seeds mechanism. For all models the initial seed weights are set to 0 with a bias of 1. This means through the attention projection mechanism all initial projected weights will be 1, meaning after the feature-wise modulation the model will continue training the network from the existent pre-trained weights of the standard VGG-16 model. We do this so that training issues are avoided and

network performance is not compromised given an fixed starting point for all target-class models to train from. To facilitate fair comparison, the task-generic model is also trained on precisely 2340 examples sampled uniformly at random. This is done to match the training set distribution characteristics of the task-specific attention models. More is explained below.

Top-down Attention

Task-specific attention models are trained on single target classes with examples from within and outside the target class set. [Luo et al., 2021] found that striking a balance between the performance benefits and costs of attention requires training the attention mechanism on equally sized task-set and non-task-set examples. This is equivalent to an attention intensity of 0.5 modulation. Each ImageNet class consist of 1170 examples and as such to maximise utility during training, the train set comprises all 1170 target class examples with the remaining 1170 examples sampled uniformly at random each epoch, giving 2340 examples in total per epoch. Due to limited sizing, and to ensure all classes are represented we enforce the class-level minimum example constraint in that the randomly sampled non-target task-set must show all remaining 999 ImageNet classes. That is, for example the non-target set would require 1170 examples from 999 classes meaning $\frac{1170}{999} = 1.17$ or 1 example per class minimum sampled randomly. In order to enable a more reliable comparison, 3 models are trained for each target class meaning 3 sets of learned seed weightings for each target class. This is helpful as we can then cross-validate model-performance, evaluate the variance and noise in per-class attention weightings, and finally perform a more robust representational similarity analysis.

3.4.4 Ensuring Non-Negativity of Projected Attention Weights

A requirement of the projected attention weights, $\mathbf{a}$, is that they must be non-negative due to non-negative attenuation of neural activity within the ventral visual pathway. With the step-up matrix, $\mathbf{P}$, being non-negative and the bias vector, $\mathbf{b}$, being fixed to 1.0 we allow attention seed weightings to be negative. Without the non-negative constraint being relaxed on the seeds, the final projected attention weights cannot attenuate below the value of 1.0. This opens the possibility of the attention seeds becoming too largely negative, resulting in negative projected attention weights. Therefore, we apply a layer weight constraint on the attention seeds such that the projected attention weight is always non-negative.

Mathematical Outline

Below outlines a simple constraining technique to ensure projected attention weights do not go below 0. The method works by clipping and proportionally scaling seed values through a constant scalar. This ensures no particular seed multiplication with any column within the step-up matrix goes below -1. Equation 3.3 shows the attention seed affine transformation. Since the bias vector, $\mathbf{b}$, is fixed to $\overrightarrow{1}$, we then have

$$\mathbf{a} = (\tilde{\mathbf{a}}\mathbf{P})^T + \overrightarrow{1} \tag{3.23}$$

We can subtract the bias vector and it is easy to see we just need to ensure the matrix multiplication, denoted $\mathbf{T}$, between the seeds in $\tilde{\mathbf{a}}$ and the step-up matrix $\mathbf{P}$ does not output any values below -1

$$\mathbf{T} = \tilde{\mathbf{a}}\mathbf{P} \geq -\overrightarrow{1} \tag{3.24}$$

We know each value within the matrix multiplication, $\mathbf{T}$, is a direct result of a seed vector multiplication with a column of the step-up matrix. More formally,

$$T_j = \sum_i \tilde{\mathbf{a}}_i \mathbf{P}_{ij} > -1 \tag{3.25}$$

The first step is to perform the matrix multiplication operation given by $\mathbf{T}$ to obtain the projection matrix. We then perform an element-wise clip of all values in $\mathbf{T}$ to ensure they are ≥ -1, giving $\tilde{T}$.

$$\mathbf{T} \overset{clip}{\rightarrow} \tilde{\mathbf{T}} \tag{3.26}$$

The clip operation sets all values in $\mathbf{T} < -1$ to -1. Next, we calculate the percentage difference between $\mathbf{T}$ and $\tilde{\mathbf{T}}$, which indicates the scalar multiplier required for each column, T_j, to satisfy Equation 3.25.

$$\delta\mathbf{T} = \frac{\tilde{\mathbf{T}}}{\mathbf{T} - \epsilon} \tag{3.27}$$

The ϵ is a small value, around 10^{-7}, that ensures no zero division errors and also that the percentage multiplier always scales the values to be slightly below -1 if they were clipped. Finally, we multiply the original seeds in $\tilde{\mathbf{a}}$ with the minimum scalar multiplier value in $\delta\mathbf{T}$.

$$\tilde{\mathbf{a}} \leftarrow \tilde{\mathbf{a}} \times min(\delta\mathbf{T}) \tag{3.28}$$

This is because the seeds are multiplied by each column in $\mathbf{P}$, therefore to ensure all values in $\mathbf{T}$ are greater than -1, we must scale the seeds for the biggest deviation found in $\delta\mathbf{T}$, indicated by

the smallest scalar multiplier. It helps to recognise all values in $\mathbf{T}$ that were not clipped, meaning they were not below -1, translate to a scalar multiplier of 1, therefore if no values were clipped the weights remain the same.

The inherent benefit of this method over clipping post-training is that the TensorFlow constraining operation performs the layer constraining after each gradient update. This means that after clipping, if any performance drop is observed the model will still continue to update layer weights in order to minimise the cross-entropy loss, which when done post-training could lead to irretrievable performance drops. Also this method means that the constraining is baked into the training process as opposed to requiring manual clipping post-training each time the model is loaded.

3.5 Evaluating Model Performance

Under stage 1 of the thesis, we wish to observe how the perceptual boost of the new low-dimensional top-down attention mechanism compares to the perceptual boost of the standard goal-directed attention mechanism. We also wish to identify how perceptual boosts vary with the number of seeds permitted to an attention seed layer, and more specifically, to the baseline models for each respective seed dimensionality. This analysis was carried out through a standard signal-detection framework [Macmillan and Creelman, 2009].

Evaluating stage 2 of the thesis requires more advanced analysis techniques. We use representational similarity analysis (RSA), a popular neuroscience technique that enables comparison of different multivariate data from different modalities by calculating the similarity between representations under different conditions from a common stimulus set [Popal et al., 2020]. This is done through mapping the different data sources to a shared representational space which will enable us to effectively quantify the ability to identify and extract patterns from the learned lower-dimensional attention weightings in comparison to the full-dimensional attention weightings. Through RSA, we also test the idea that we can extract a sensible semantic space from the attention weights trained within the lower-dimensionality subspace by comparing them to the ground-truth semantic relationships between class word embeddings. With target-classes chosen to yield a variety in model-and-class-level factors such as class-difficulty, variance in filter activations, and the level of correlated data within filter activations elicited from class samples, we compound the analysis to consider the influence such factors have over the learning of attention weights. Below we outline each analysis technique in more detail.

3.5.1 Signal-Detection Analysis Framework

The signal-detection framework analyses the performance of the trained attention models and baseline models through the hit and false alarm rates. The hit rate, H, is measured as the rate by which the model makes a correct classification (true positive), that is the model predicts the correct class as the most likely class. The false alarm rate, F, is measured as the rate by which the model mistakenly predicts the current model's target class as the the most likely class (false positive). For example if the attention model in question had the target class for Broom, and Broom was predicted the most likely label when no Broom is present in the example image then we consider this a false alarm. With these two metrics we may then compute the model sensitivity, d', and model criterion, c, as dimensionality changes. d' measures the ability of the model to distinguish the signal-present case from a signal-absent case amongst noise. It is given as

$$d' = \phi^{-1}(H) - \phi^{-1}(F) \tag{3.29}$$

where ϕ^{-1} is the inverse of the ϕ function that converts probabilities into z-scores within a standardised normal distribution. The higher the sensitivity the greater the difference in the signal-present and signal-absent distributions meaning better expected performance by the attention mechanism in distinguishing the target class. The criterion is a measure of bias of the model to respond with a hit in an ambiguous situation. It is the middle point between the signal-present and signal-absent distributions and is given as

$$c = -\frac{\phi^{-1}(H) - \phi^{-1}(F)}{2} \tag{3.30}$$

To facilitate a signal-detection analysis we use an equal number of target and non-target class examples when measuring performance. We use the entirety of the attention model's corresponding target class and uniformly at random sample an equal number of examples from the remaining 999 ImageNet classes to compute the aforementioned metrics. It is important to note while computing sensitivity and criterion using the inverse z-transform, ϕ^{-1}, problems arose when hit-rates or false-positive rates equate to 0 or 1. The inverse z-transform of such values are $-\infty$ and $+\infty$, respectively, and as such cause problems in signal-detection calculations. It is suggested to clip 0 values to $\frac{0.5}{n}$ and 1 values to $\frac{n-0.5}{n}$ [Macmillan and Kaplan, 1985], where n is the number of measurements within a sample, however this yields biased metrics as clipping to higher or lower values lead to different sensitivity and criterion values, therefore exact signal-detection metrics are harder to obtain [Stanislaw and Todorov, 1999]. Several other proposed methods look

to use non-parametric sensitivity estimates such as A' [Craig, 1979], which eliminate dependency on the ϕ^{-1} function. It is, however, satisfactory to perform clipping. Although inducing biases in measures, if performed consistently across all calculations, the trend across dimensionalities is maintained regardless of exact metric values, therefore inferences made from these trends are entirely valid.

Outside of this, we also evaluate the perceptual boost of the attention seeds mechanism on the model accuracy (class difficulty). Simply put, the ImageNet test-set images relating to the target class are retrieved and the accuracy of the task-generic baseline model for a particular dimensionality is compared to the task-specific specialised model. The differences between the two models indicates the impact of specialised task-specific seed training. Overall, this basic analysis quantifies the effect of low-dimensional subspace attention seed training, seed dimensionality and attention as a whole on model performance.

3.5.2 Representational Similarity Analysis

RSA is a form of multivariate pattern analysis (MVPA) used within neuroimaging literature as a solution to the limitations of mass-univariate statistical techniques [Haxby et al., 2001, Popal et al., 2020]. RSA looks to compare pattern responses of stimuli through a direct higher-order representational space of individual responses yielding a more suitable structure of representation across various data sources [Haxby et al., 2014, Kriegeskorte et al., 2008]. Hence, RSA made an appropriate tool to compare attention seed weightings across multiple target stimuli to their ground-truth underlying semantic relationships portrayed by the BERT-encoded class name embeddings. Relative to these ground-truth relationships, we can then begin to observe if the considered factors such as class-difficulty and variance in activations really play a role in influencing correlations to the semantic embeddings. For example, if one grouping correlates the attention weights to the semantic embeddings well but another does not, we can use these other factors to potentially explain why that second grouping did not, thereby identifying if that factor did indeed influence the learning of attention weights.

How to Perform a Representational Similarity Analysis

RSA is relatively easy to conduct and makes extensive use of representational dissimilarity matrices (RDMs). An RDM is a symmetrical matrix with entries populated by the dissimilarity between two stimuli at two specific indices, and a diagonal component of 0's as any stimuli cannot be dissimilar to itself. One can use any suitable distance measure for dissimilarity between sets of

stimuli, where dissimilarity is simply (1−similarity metric). A suitable similarity metric should be chosen when comparing the attention seeds and semantic vectors. Typical choices range from Pearson or Spearman correlation's, to Euclidean or Mahalanobis distance's and is dependent on the type of data compared. The attention seed weightings and semantic vectors are of the same order of magnitude meaning we do not require the magnitude-insensitive property of the correlative measures. Furthermore, since we require a more interpretable and direct comparison of these vectors within the vector space, we choose to use a variant of the Mahalanobis distance similarity measure. The Mahalanobis distance is more reliable than the standard Euclidean distance when different channels could possibly be correlated with each other due to in-built noise-normalisation [Kipnis, 2022]. Since each class has 3 models trained, meaning 3 learned seed weightings, the cross-validated Mahalanobis (Crossnobis) distance is used. The Crossnobis distance is an unbiased estimator of the square Mahalanobis distance [Walther et al., 2015, Berlot et al., 2020] as patterns are only ever multiplied across runs and not within runs. This means if the true distance is zero (two patterns differ only by noise) then the average estimated distance will be zero. Cross-validation also removes unequal noise biases that arise from single RDM use, thereby making inference more robust. We note that all similarity measures are found to be reliable among each other, so there is no one wrong choice [Walther et al., 2015]. The Crossnobis is given as

$$d_{i,j} = \frac{1}{M(M-1)} \sum_{m}^{M} \sum_{n \neq m}^{M} (\mathbf{b}_{i,m} - \mathbf{b}_{j,m})(\mathbf{b}_{i,n} - \mathbf{b}_{j,n})^T / C \tag{3.31}$$

where $\mathbf{b}$ is a $M \times C$ data matrix and $d_{i,j}$ is an entry in the resultant RDM matrix. Note that the distances are normalised by the number of channels, C. This allows comparison of distances across weightings with different number of dimensionalities. Multiple RDMs are then constructed at each set seed dimensionality between learned seed weightings within the superordinate groupings, for the baseline models, and finally for the ground-truth semantic relationships.

RDMs are then quantitatively compared through correlative measures like Pearson or Spearman correlation to reveal the extent to which the representational geometries in attention weights and semantic vectors are similar to each other. The Spearman correlation metric is recommended as rank-correlation distances are noise insensitive [Kriegeskorte et al., 2008]. Further to this, with each RDM differing in noise modality we also drop the assumption of a linear relationship when comparing RDMs and accordingly the Spearman correlation fits this purpose for RDM comparisons. The original form of the Spearman correlation is higher for predictions with tied ranks and can potentially introduce bias into analyses. Therefore, the expected Spearman's ρ, denoted ρ_a, under an unknown joint distribution regime between $\mathbf{x}$ and $\mathbf{y}$ with random tie-breaking

as an evaluation criterion is used. This is given by

$$\rho_a(\mathbf{x}, \mathbf{y}) = \mathbb{E}_{\substack{\tilde{\mathbf{a}}=\tilde{\mathbf{x}}-\frac{1}{n}\sum_{i=1}^{n} i,\, \tilde{\mathbf{x}}\sim Rae(\mathbf{x}) \\ \tilde{\mathbf{b}}=\tilde{\mathbf{y}}-\frac{1}{n}\sum_{i=1}^{n} i,\, \tilde{\mathbf{y}}\sim Rae(\mathbf{y})}} \left[\frac{\tilde{\mathbf{a}}^T\tilde{\mathbf{b}}}{\|\tilde{\mathbf{a}}\|_2 \left\|\tilde{\mathbf{b}}\right\|_2} \right]$$

$$= \frac{12}{n^3 - n}\mathbb{E}_{\tilde{\mathbf{a}}}[\tilde{\mathbf{a}}]^T\mathbb{E}_{\tilde{\mathbf{b}}}[\tilde{\mathbf{b}}]$$

$$= \frac{12\,\mathbf{x}^T\mathbf{y}}{n^3 - n} - \frac{3(n+1)}{n-1} \tag{3.32}$$

where n is the number of stimuli being evaluated, and the expectation is evaluated under the estimator of Spearman's ρ with ranks returned by the $Rae()$ function [Kipnis, 2022]. The Spearman correlation is computed between the flattened rank order of dissimilarities of the lower-triangle between two RDMs. The correlation metric measures how well the information in the candidate RDM is represented within the ground-truth RDM (or how predictive the ground-truth RDM is of the candidate RDM).

To conclude the RSA we fit models to the RDMs. With RDMs constructed at each dimensionality, it is beneficial to obtain a single RDM representative of all dimensions. An RDM interpolation model is built for each grouping's RDMs which interpolates the cross-validated RDMs across dimensions to achieve a single RDM representative of all dimensions. Interpolation models predict that the RDM is a linear interpolation between multiple consecutive RDMs. Interpolation models are primarily used to represent nonlinear effects of a single changed parameter [Kipnis, 2022], hence making the model suitable for taking into account dimensionality changes.

Noise Evaluation of the Representational Similarity Analysis

In computational neuroimaging it is beneficial to report model prediction performance with regard to an estimated noise ceiling [Lage-Castellanos et al., 2019]. This follows the idea that stimulus responses that are fixed often differ over repetitions only in noise. This assumption allows us to estimate how well the evaluated model's performance achieves the maximal possible performance when factoring noise levels in the data by estimating how much stimulus-related variance are in the data. Natural noise in the data will limit correlation performance when comparing model RDMs to the noiseless ground-truth RDMs. Since we train 3 models for each task-set across all dimensionalities, we can compute a noise ceiling by performing cross-validation over the RDMs, estimating the underlying true data-generating task-set model of seed weightings. The noise ceiling is plotted as a horizontal grey bar with an upper-bound and lower-bound estimate on the

group-average correlation with the RDM predicted by the true data-generating model. Performing a leave-one-out cross-validation over the RDMs for a superordinate grouping, for each split, the left-out RDM is compared to the average over the remaining RDMs to get the lower-bound on the noise ceiling. The left-out RDM is then compared to the average of all RDMs, including itself, to obtain the upper-bound estimate [Lage-Castellanos et al., 2019]. The average bounds over all splits form the final estimate for the noise ceiling. The lower-bound is an underestimate of the true noise ceiling as it does not utilise all RDMs when being computed. It has been proved that if we had the true data-generating model the best possible correlation achieved would fall within the noise ceiling range [Nili et al., 2014]. When fitting our models, the noise ceiling estimate is naturally computed using the cross-validation method [Kipnis, 2022].

Chapter 4

Results

4.1 Data Exploration Results

In this section we outline the results from experiments run to select the target-class set. Section 4.1.1 covers the results pertaining to selecting target-classes based on class-difficulty and variance of filter activations elicited from class samples. Two superordinate groupings are selected based on results from this section. Section 4.1.2 then selects target-classes that are the most semantically similar from the identified superordinate groupings through a community-detection clustering algorithm (see Section 3.4.1 for more details). We therefore end up with target-classes that are fundamentally connected through semantic relationships, and that are varied in factors such as class-difficulty and variance of filter activations in preparation for the RSA performed in Section 4.2.5.

4.1.1 PCA-NMF

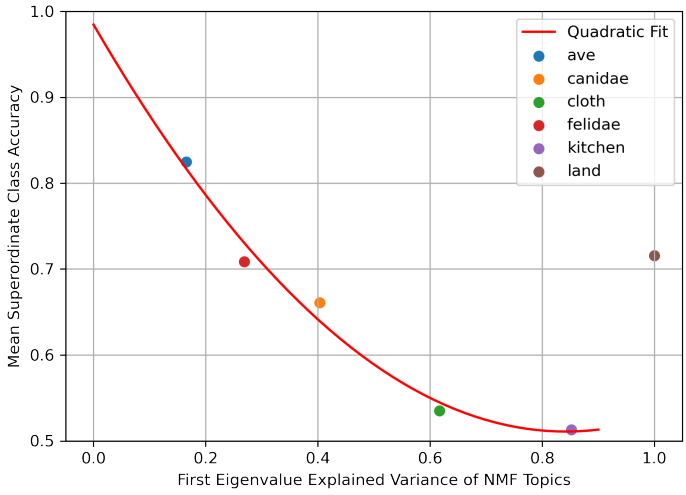


Figure 4.1: Plot of average superordinate grouping (baseline) model accuracies versus the explained variance of the first eigenvalue NMF-extracted topics across all superordinate grouping samples passed through the network. This plot is important in selecting target-classes from groupings that not only vary in overall class difficulty, but also vary in variance of activations from samples belonging to that grouping. Selecting target-classes in such a way allows us to take into account the effect such model-level variables such as mean grouping difficulty and variance of activations have on the learning process of attention weights. A polynomial regression fitting of the data indicated a quadratic fit minimised the mean-squared error between points and the predicted regression line. A trend indicates that increasing correlations in NMF-factorised grouping activations meant a decrease in overall grouping accuracies. The land grouping appears to be an outlier to this trend.

The overall trend seems to indicate that superordinate classes with the least correlation between extracted topics tended to have better class accuracies (less difficulty) overall. This makes sense, if the decomposition of activations from different classes within a superordinate are very similar, then the network will have more difficulty differentiating them when performing classification. One outlier from the trend is exhibited by the Land Transport (land) classes, which had perfect correlations within samples yet still performed well on mean superordinate accuracies. The Kitchen superordinate performed the worst at an accuracy of 51.2% and explained variance of 84% while the Avian (ave) superordinate performed the best at an accuracy of 82.7% and explained variance of 16.2%. With our goal of performing evaluations that encompass a variety of class-level performance and model-level cues such as variance of activations, the Avian and

Kitchen superordinates were chosen as superordinates for our target-class selection.

4.1.2 Community Detection Clustering

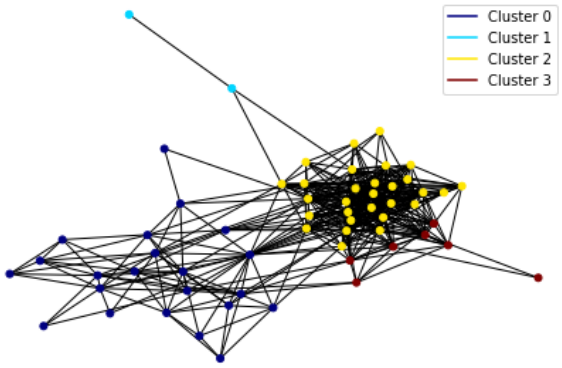


Figure 4.2: Visual plot of identified clusters of the Avian superordinate grouping through the community detection Louvain algorithm. The algorithm is used to find strongly semantically entangled classes within the superordinate grouping of classes. All semantic embeddings of classes belonging to the superordinate grouping occupy the nodes of the network, with edges between all nodes indicating the cosine similarity between two nodes semantic embeddings. The community detection algorithm then identifies strongly connected groupings based on the cosine similarity weighting between nodes. We require this to establish as much of a fundamental relationship between target-classes as possible, so enable easier comparison of the representational geometries between learned class seed weightings to their respective semantic embeddings.

Table 4.1: The mean cosine similarity scores between Avian classes within each identified cluster. As before choose the target-classes based on maximising the semantic similarity between target-classes to enable a better representational similarity analysis. The highlighted cluster was the chosen cluster. We ignore clusters with less than 5 members as this would not yield enough data to run a good quality evaluation.

Cluster	Mean Cosine Similarity	# of Cluster Members
0	0.57	22
1	0.71	2
2	0.63	29
3	**0.66**	7

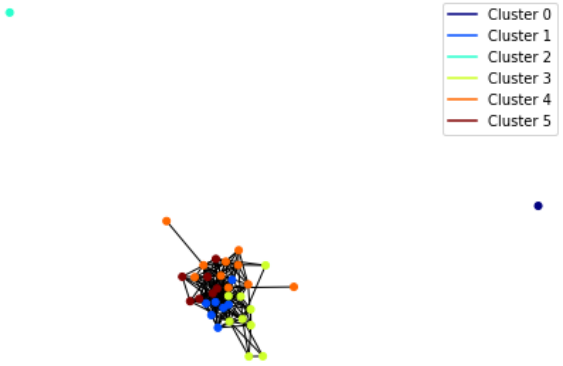


Figure 4.3: Visual plot of identified clusters of the Kitchen superordinate grouping through the community detection Louvain algorithm. As before, the algorithm is used to find strongly semantically entangled classes within the superordinate grouping of classes based on the cosine similarity between classes of the superordinate grouping.

Table 4.2: The mean cosine similarity scores between Kitchen classes within each identified cluster. We choose the target-classes based on maximising the semantic similarity between target-classes to enable a better representational similarity analysis. The highlighted cluster was the chosen cluster. We ignore clusters with less than 5 members as this would not yield enough data to run a good quality evaluation.

Cluster	**Mean Cosine Similarity**	**# of Cluster Members**
0	1.00	1
1	0.71	7
2	1.00	1
3	0.67	9
4	0.64	10
5	**0.72**	7

With superordinate groupings chosen to encompass a variety in class difficulties and activations of filters, each superordinate was then clustered based on their cosine similarity scores between class-name semantic embeddings. Figures 4.2 and 4.3 shows visually the clustering algorithm output for the Avian and Kitchen superordinates, respectively. The clusters identified in each superordinate with the highest cosine similarity, excluding clusters under the size of 5, were identified as cluster 3 in the Avian superordinate and cluster 5 in the Kitchen superordinate, with mean inter-cluster cosine similarities of 0.66 and 0.72, respectively. Both clusters contained 7 classes and these classes are merged to form the target-class set, T:

Avian: ['junco', 'kite', 'peacock', 'lorikeet', 'hummingbird', 'goose', 'crane']

Kitchen: ['paper towel', 'dishwasher', 'broom', 'saltshaker', 'mixing bowl', 'washbasin', 'dishrag']

As explained previously, selecting classes that are highly semantically entangled allows us to then analyse if learned attention seed weightings between such classes are also highly entangled, in order to gain a perspective at the underlying factors that govern why attention weights are learned how they are, outside of maximising model performance on the validation set. With task-sets (target-classes) selected, next experimentation makes use of these target-classes to evaluate the success of the attention seeds mechanism in performance, and in explanatory power of learned seed weightings. Please see Appendix A for a full list of the target-class set.

4.2 Main Results

In this section we detail and evaluate the main experimental results. The performance of the attention seeds mechanism is evaluated and analysed under a standard accuracy-based evaluation criterion as well as a signal-detection framework. The learned seed weightings are then analysed to obtain a deeper understanding into how dimensionality affects the distributions of seeds, as well as projected attention weightings, relative to the target-class training scheme. Finally, RSA is used to map seeds to a shared representational space where they are compared and contrasted to the ground-truth class label semantic embeddings, target-class factors investigated, dimensionality of the seeds mechanism, and finally to each other. This final analysis is crucial in understanding the underlying mechanistic processes underpinning learned weighting distributions and translating them to an understandable domain in which we can draw sensible inference from. All experiments evaluate results in comparison to the standard full-dimensional goal-directed attention mechanism found in [Luo et al., 2021] in order to investigate how the parameter-reduction technique fairs against the current state-of-the-art.

4.2.1 Attention Seeds Performance

Basic Analysis

This experiment examines the effect of dimensionality on the perceptual boosts achieved by the attention seeds mechanism. In quantifying this effect we can greatly examine the redundancies and over-parameterisation in utilising full-dimensional attention mechanisms. This experiment begins to answer the first two questions asked at the beginning of the thesis in Section 1. Attention

models were trained for the task-generic goal setting and task-specific goal setting. For each of the 14 target classes, 3 models were trained resulting in 42 learned attention weightings at each dimensionality for a total of 210 attention models. In order to infer the perceptual boost of the attention seeds mechanism for a task-set all samples of the task-set were retrieved. The difference in accuracy between the task-specific model relating to the task-set and the baseline model at the set dimensionality indicates the performance gain of the seeds mechanism.

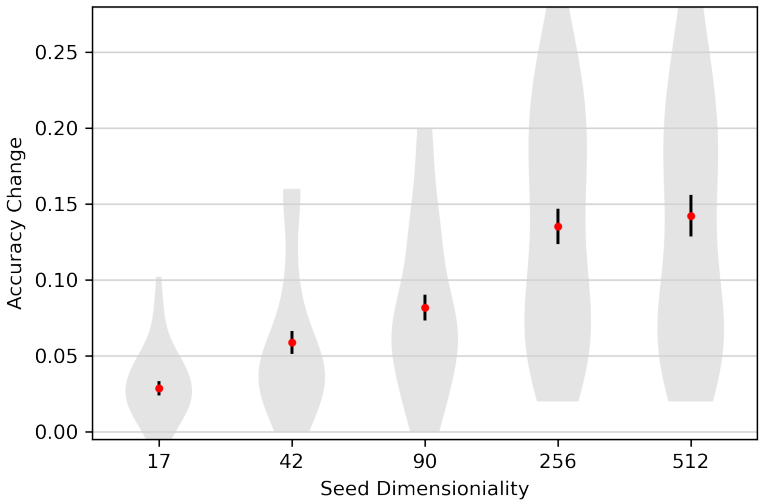


Figure 4.4: Violin plot of the perceptual boost produced by the 210 top-down attention seeds models on 14 visual tasks across increasing attention seeds dimensionalities. At a 50% reduction in trainable attention parameters(attention seeds), the 256 dimensional mechanism achieves perceptual boosts equal to that of the full 512 dimensional mechanism and speaks to the success of the parameter-reduction on standard goal-directed attention mechanisms. The perceptual boost indicates the success of the attention seeds mechanism at increasing the localising power of the specialised model in identifying the target-class among all test-set samples belonging to the target-class. The perceptual boost is computed as the accuracy change between the task-generic baseline attention seeds models to the task-specific attention seeds models at each dimensionality. The violin plots show the quartile ranges and estimated distributional densities of the perceptual boosts. The red dot indicates the mean perceptual boost at each dimensionality, with the black lines either side of the red dot indicating the standard error across all samples drawn from the 210 attention seeds models. With increasing seeds dimensionality meant an increase in perceptual boosts achieved by the attention seeds mechanism.

A perceptual boost was identified at all dimensionalities of the seeds mechanism. Figure 4.4 shows that perceptual boosts were found to increase with dimensionality from 17 to 512. Lower dimensionalities had significantly degraded performance gains as expected by the lower

degrees of freedom in projected attention weights. The highest perceptual boost was exhibited at dimensionality of 512, however, significant overlap between the violins of the 512 and 256 dimensionality models was observed, suggesting both dimensionalities operated similarly. This result could simply be an artifact of the stochastic training process resulting from a gradient-descent based update rule. A two-sided t-test between concurrently ranked perceptual boosts found the 512-256 result to not be significant ($t(82) = 0.39$, $p > 0.5$), suggesting that both 512 and 256 operate with equal performance and that the chance of the 512 model being better is greater than chance. This postulates that there are around 256 redundant dimensions within the original decomposition as the seed mechanism at 50% reduction in dimensionality performs as well as the full dimensional equivalent. The difference between the remaining 256-90, 90-42 and 42-17 paired dimensional models was found to be significant ($t(82) = 3.69$ $p < 0.002$, $t(82) = 2.02$ $p < 0.05$, and $t(82) = 3.4$ $p < 0.002$, respectively), suggesting optimal dimensionality for the mechanism is 256 dimensions.

The range in perceptual boosts within dimensions increased with dimensionality, resulting in increasing standard errors. Interestingly, it appears their is a floor and ceiling on perceptual boosts that grows in difference as dimensionality increases. The 17, 42 and 90 dimensional models have a lower bound close to 0 with increasing upper bounds. The 512 and 256 dimensional models have the greatest range, and consequently standard error, in perceptual boosts suggesting with more degrees of freedom there is a higher probability of not converging to the optimal seed weighting that leads to the greatest perceptual boost, as indicated by the spread out probability distributions of the violins. In particular, the specific shape of the violins at the 512 and 256 dimensions suggests a disconnect grows between lower perceptual boosts and higher perceptual boosts, meaning there is an inherent per-target-class cap in perceptual boost that is being achieved at these higher dimensions. The majority of target-classes exhibited a lower perceptual boost cap as indicated by a greater width of the bottom lobes of the violins.

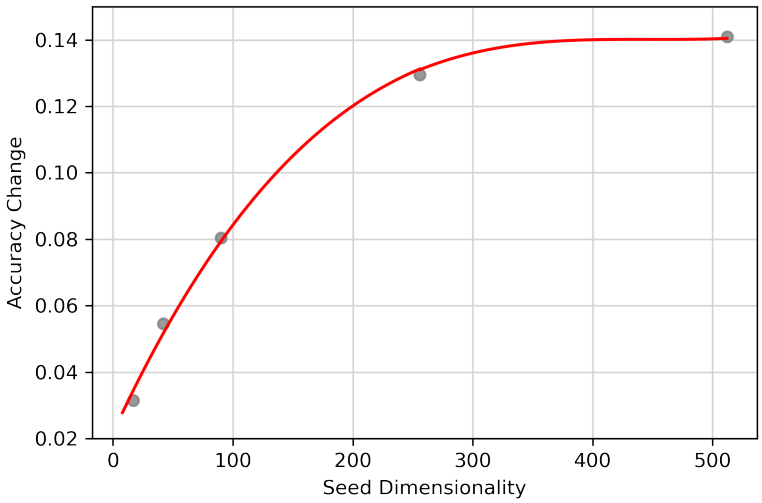


Figure 4.5: Scatter plot of the perceptual boost produced by the 210 top-down attention seeds models on 14 visual tasks. The trend (red line) seems to indicate a natural logarithmic relationship between increasing perceptual boosts and increasing dimensionalities. The perceptual boost increase plateau's at dimensionality 400 as seeds dimensionality increases between 256 dimensions to the full 512 dimensions. This speaks to the redundancies in utilising full-dimensional attention and shows indeed the over-parameterisation found in DCNNs is carried over to the modulating attention mechanisms. This is plotted on a continuous x-axis as to show the continuous relationship between the mean perceptual boosts (grey dots) attained with increasing dimensionalities.

Figure 4.5 shows that the mean perceptual boost smoothly increases with dimensionality with a natural logarithmic trend. Rapid increases in perceptual boost are observed at lower dimensionalities, but this change in perceptual boost decreases drastically and plateau's between the 256-512 dimensional range. Perceptual boost at all dimensions is apparent and provides evidence for the success of the dimensionality-reduced attention mechanism, especially the 512-256 reduction pair, again demonstrating the redundancy in the full-dimensional attention mechanism.

Unexpectedly, the overall mean accuracy at higher dimensions decreased slightly after applying the non-negative constraining technique to seed attention weights. The higher dimensional seed models tended to lead to negative projected attention weights suggesting that higher perceptual boosts could be attained if attention weights were allowed to be negative, however, this is in direct violation of the non-negative neuron modulation exhibited within the ventral visual pathway. We only utilised a simple re-scaling technique of seed weightings to ensure non-negativity of projected attention weights, however, it is clear more sophisticated techniques may be required

such as monotonic transformations of seed weightings in order to recover the small degradation in seed mechanism performance. Nonetheless, the perceptual boosts are apparent under the new lower-dimensional seeds mechanism, especially when comparing the consistent performance of the 256 dimensional mechanism to the full 512 dimensional mechanism.

Signal-detection Analysis

This experiment compares the performance of the attention seeds mechanism across increasing dimensionalities to the standard full-dimensional goal-directed attention mechanism, found in [Luo et al., 2021], under a signal-detection framework. This comparison is vital for relating the performance between both attention mechanisms, and more importantly, understanding the impact dimensionality-reduction has in relation to the standard attention mechanism. This more intricate analysis enables a comparison beyond standard accuracy-based metrics, compounding an examination on the costs and benefits of models [Luo et al., 2021], and the biases present within the attention mechanisms (see Section 3.5.1 for more details). Signal-detection metrics for the standard attention mechanism, trained in the same manner as the attention seeds mechanism, are denoted as '512-S' in figures.

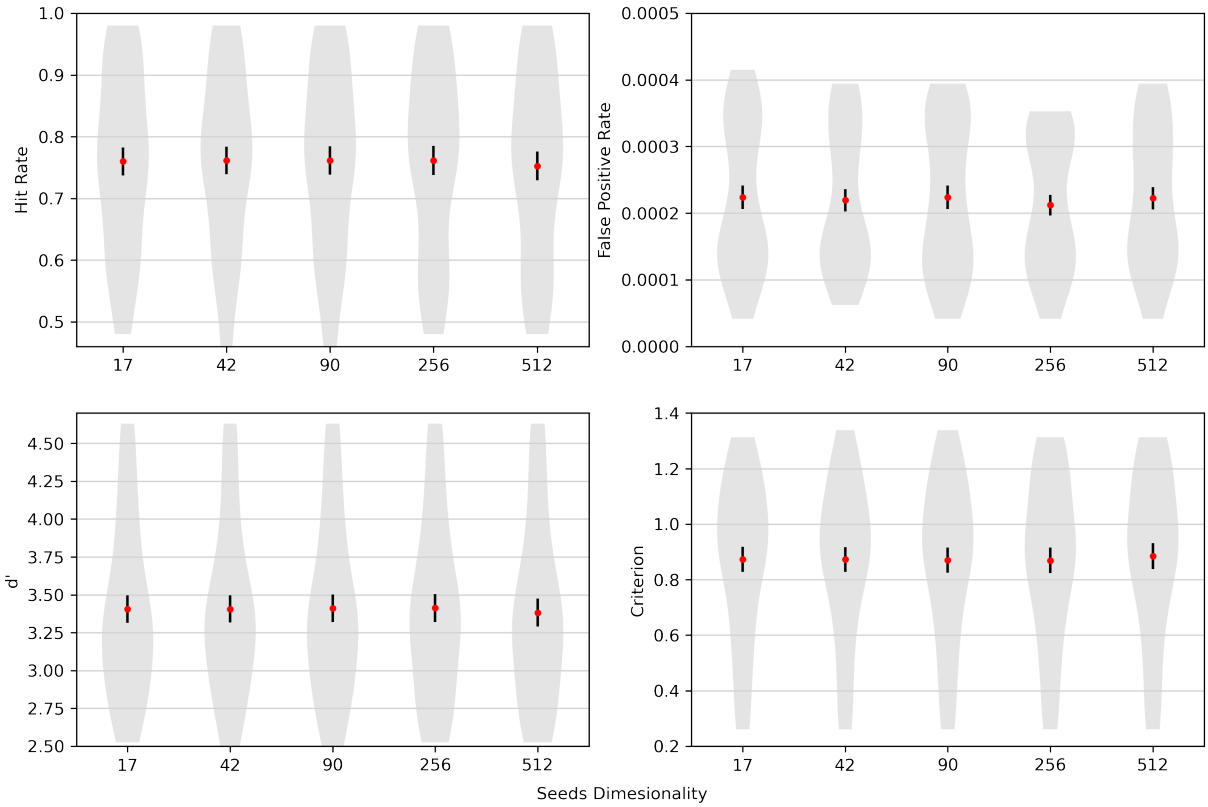


Figure 4.6: Violin plots of all signal-detection metrics for (baseline) attention seeds models at five increasing dimensionalities on all target-classes. The effect of seeds dimensionality is not apparent on the task-generic baseline models as the models specialise attention across all 1000 ImageNet classes. There is no inherent performance gain when training task-generic attention models as the pre-trained VGG-16 model they modulate has already reached optimal performance on across all ImageNet classes. This is in line with expectations and means the attention seeds mechanism performs the same as the full-dimensional attention mechanism when specialising generically to all classes in comparison to specialising to a single target-class. Dimensionality-reduced goal-directed attention with varying dimensionality of attention seeds across all ImageNet classes were tested. The red dots show the means of each violin with the standard error represented by the black lines either side of the mean point. The violin lobes show the estimated probability densities all samples generated at a specific dimensionality of the seeds mechanism.

For initial reference the signal-detection metrics were plotted for the task-generic baseline models at all seed dimensions and are depicted in Figure 4.6. Note that baseline models for the standard mechanism are not required as signal-detection metrics are calculated solely on the mechanism performance relative to itself. As expected, all metrics appear consistent across dimensions with no significant change apparent. With task-generic weightings optimising for all ImageNet classes, dimensionality will have no inherent impact on average model hit-rates,

as model hit-rates themselves do not increase over the standard pre-trained VGG-16 model for baseline models even when considering the full-dimensional models. Again, the 512 dimensional model appears to have slightly degraded performance across all metrics which could be due to over-parameterisation at such high dimensionalities or can simply be an artifact of the non-negative seeds constraining technique used.

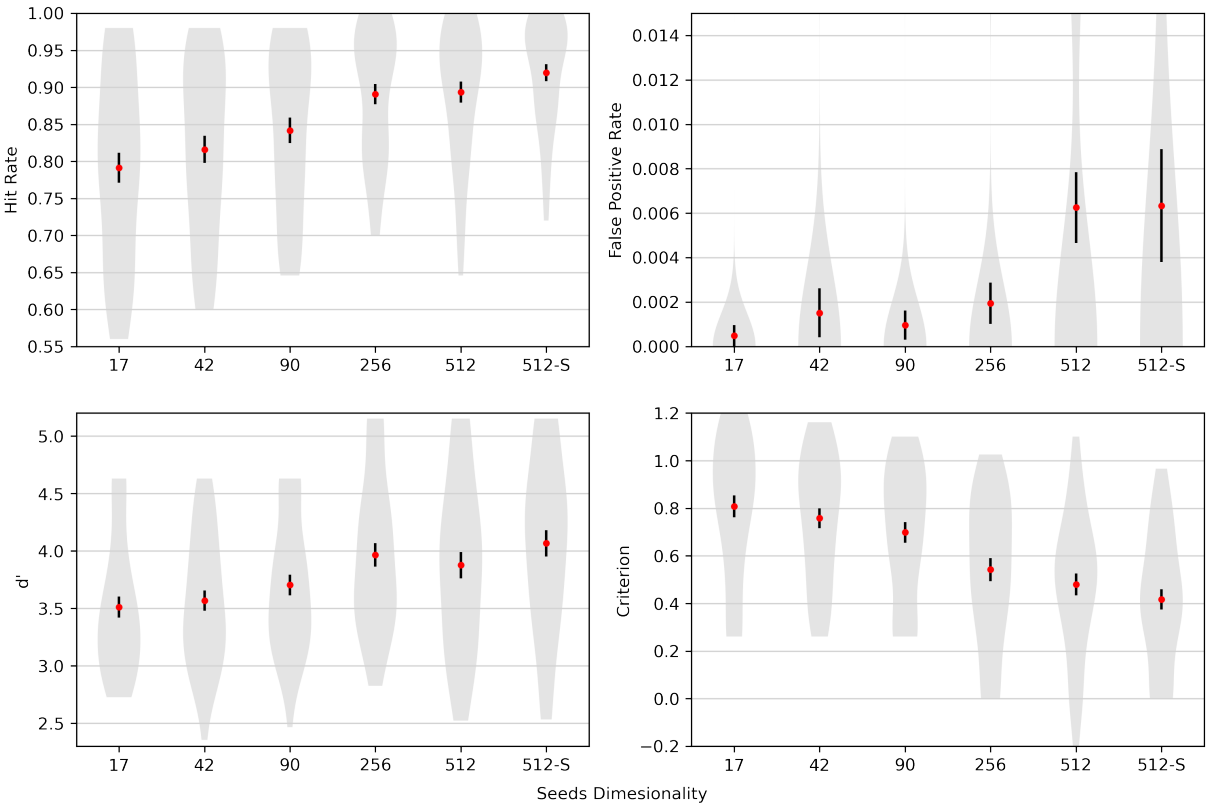


Figure 4.7: Violin plots of all signal-detection metrics for specialised attention seeds models. Dimensionality-reduced goal-directed attention at five increasing dimensionalities of attention seeds across all target-classes were tested. As attention seeds dimensionality increased, goal-directed attention had generally increasing benefits in higher hit-rates as well as generally increased false alarm (false-positive) rates. With increasing seeds dimensionality, model sensitivity (d') first increased to a maximum at 256 dimensions, and then decreased for the 512 dimensional seeds mechanism. The model was increasingly more biased towards making a false alarm (criterion decreased). The plot also shows the performance of the original standard goal-directed attention mechanism of [Luo et al., 2021]. At a 50% reduction in trainable attention parameters, the seeds mechanism performs just as well as the full-dimensional standard attention mechanism, while offering a much lower false alarm rate. Through a Student's t-test, model sensitivity was found to be equal between the lower 256 dimensional seeds mechanism to the standard full-dimensional attention mechanism. This plot therefore validates the success of the developed mechanism at reducing trainable attention parameters while not compromising attention performance. It also validates the idea that there exist lower-dimensional latent factors that drive the high-dimensional attention weights due to removed over-parameterisation. The red dots show the means of each violin with the standard error represented by the black lines either side of the mean point. The violin lobes show the estimated probability densities all samples generated at a specific dimensionality of the seeds mechanism.

Consistent with basic results presented above, Figure 4.7 shows that model hit-rates increased as seed dimensionality increased. The change appears to be linear between dimensions 17-256 but very similar mean hit-rates were observed between the 256-512 dimensions pair, again speaking to redundancies of the full-dimensional attention mechanism. It is apparent the 256 dimensional model was more reliable than the 512 dimensional variant as hit-rates were more densely distributed among higher hit-rate values leading to a smallest range in values of the violin. In fact, the standard error was found to be minimal at the 256 dimensional models at 0.013.

Model sensitivity peaked at the 256 dimensional seed mechanism, followed closely by the 512 dimensional model, further providing evidence in favour of the reduced-dimensionality attention mechanism. Following these, the next highest sensitivities are observed at the 90, 42 and finally 17 dimensional models. A student t-test between the decreasing pairs of sensitivities reveals the null hypothesis that there is no difference between subsequent mean sensitivities is under 50% for the 512-256, 256-90, 90-42 dimensional pairs, however, this probability is still fairly large. The 42-17 dimensional pair were found to be insignificant with $t(81) = 0.45\ p > 0.5$. Both the 512 and 256 dimensional models exhibit an equal upper-bound on sensitivity, however the 256 dimensional model has a smaller range in observed sensitivities and hence a smaller standard error. The upper lobe thickness of the violin plot is greater in the 512 dimensional model, suggesting more frequent higher sensitivities, yet, the lower-bound of the 512 dimensional model being much lower than the 256 dimensional variant limits this advantage.

Interestingly, the standard attention mechanism performs better than its 512-dimensional seeds counterpart, despite sharing equal trainable parameters and degrees of freedom. This result is a direct consequence of the performance degradation observed at higher dimensionality of seeds due to more severe constraining of seed weightings between gradient updates when projected attention weightings tended to be negative. The standard attention mechanism performs marginally better in sensitivity over the 256 dimensional seeds mechanism, even while containing double the number of trainable parameters. The sensitivity increase between the standard feature-wise attention mechanism and the 256 dimensional seeds mechanism was found be insignificant with $t(82) = 0.65$, $p > 0.5$. From a signal-detection perspective, this result is amazing and shows evidence to the success of the dimensionality-reduced attention mechanism, even while operating at 50% reduction in attention parameters and speaks to the over-parameterisation of the original filter-space. The standard attention mechanism exhibited the lowest criterion meaning it was the most biased to making a false alarm. As such the standard attention mechanism also exhibited the highest mean false-positive rate as expected by such a low criterion. It is clear the sweet-spot for dimensionality-reduction that minimises performance loss is the 256 dimensional seeds

mechanism.

Model false-positive (false alarm) rate fluctuated between lower dimensions, with a minimal mean false-positive rate observed at the lowest dimensionality mechanism. The 512 dimensional models exhibited a significantly higher false alarm rate at 0.006, 3 times higher than the second highest false-positive rate, meaning the 512 dimensionality mechanism tended to favour predicting the target-class even if the target-class was not present within the input stimuli. A decreasing model criterion suggests that as the seed dimensionality increased, the model was more biased in favour of a target class response, which was more likely to result in a false alarm, explaining why the 512 dimensional model had a high false-positive rate. However, the ratio difference in model criterion between the 256-512 dimensional pairs is significantly smaller than the ratio-difference in false-positive rates, despite having similar hit-rates and sensitivity. This suggests that the full-dimensional attention seed mechanism is over-parameterised, which at surface-level does not harm model performance entirely, however severely biases the model towards its respective task-set.

4.2.2 Summary of Key Findings: Performance Evaluation

The 256 dimensional attention seeds mechanism at a 50% reduction in trainable attention parameters performed near identically to the full-dimensional standard attention mechanism. The 256 dimensional seeds mechanism exhibited the same model sensitivity as the standard attention mechanism, and was found to be more consistent in achieving higher sensitivities over the standard attention mechanism. Across all dimensionalities, the seeds mechanism was less biased towards making a false-alarm than the standard attention mechanism. These results are amazing and prove strongly the success of the attention seeds mechanism in performing a parameter-reduction of state-of-the-art goal-directed attention mechanisms. Results here prove that redundancies found within the full-dimensional filters are indeed replicated within the attention mechanism that is modulating them, under this one-to-one multiplicative basis. This is pivotal as such attention mechanisms can now offer flexibility in choosing the number of trainable attention parameters without being restricted to the dimensionality of the filter-space! Furthermore, with increasingly wider neural networks this parameter-reduction can be extremely beneficial in performing increased model compression, improving model efficiency and lowering training times while maintaining performance gains of the attention mechanism. In fact we predict that if interfaced into higher-dimensional filter-spaces, say with deeper network layers, the mechanism could produce even greater reductions in trainable attention parameters while maintaining model performance, however this should be tested under future experimentation.

4.2.3 Attention Weights Analysis

In this experiment the distribution of learned attention seeds and projected attention weights are analysed across seed dimensions. All weightings are plotted using a histogram with bins normalised such that the area of the histogram integrates to 1, enabling a better comparison. Weightings are plotted for all task-specific models and task-generic baseline models. This experiment deeper analyses the effect dimensionality has on learned weightings, enabling an insight into how the underlying training process adapts for differences in degrees of freedom in order to achieve perceptual boosts on task-sets.

Seeds Distribution

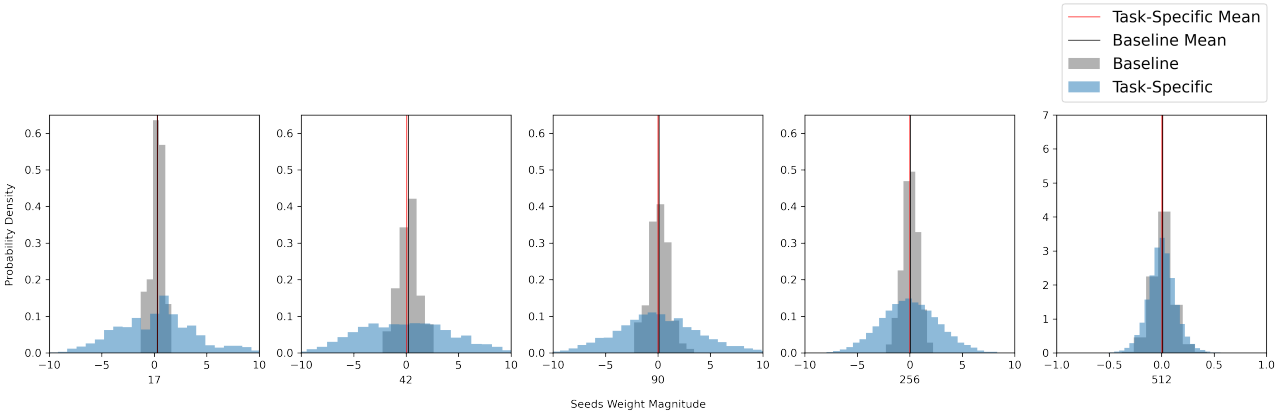


Figure 4.8: Comparing task-generic and task-specific attention seeds weights. With increasing dimensionality of seeds, task-generic weights maintained a similar distributional spread as expected due to the effect of dimensionality not influencing the performance of task-generic baseline models. Task-specific weights started off more diverse and broad but as dimensionality increased became more concentrated around the distributional mean. Lower number of trainable attention parameters means attention seeds must diversify more in order to cover ground lost in missing attention parameters in trying to achieve the optimal perceptual boost on the target-class. Due to the additive nature of attention seeds in projecting to the final attention weights, as attention dimensionality increases the seeds can more finely approximate the optimal final attention weights for a target-class. This translates to more smaller diversification in seed values as the summation over a larger number of attention parameters can take smaller steps in order to closely achieve the optimal attention weight for a filter. Note the final 512 dimensional plot has a smaller x-axis range of values as this distribution was extremely sharp around the zero-mean, hence to aid in visual interpretation, was scaled in. Distributional means were found to be the consistent across dimensionalities.

Earlier in the thesis we hypothesised that with the lower degrees-of-freedom available with lower-dimensional attention weights, the model would warp attention seeds more greatly under the influence of factors that drive the learning of attention weights. This means as seed dimensionality increases, the distributions in seed weights will become more sharply centered around 0. Compounding this with details of the projection mechanism, with more degrees of freedom to work with within the step-up matrix, the optimal attention weight for a specific filter within the original filter-space can more finely be achieved to a higher degree of precision. This means smaller adjustments, or less warping under influencing factors, are required to the principal component weightings found within the step-up matrix to achieve the optimal attention weight for the specific filter. This hypothesis also explains why attention seeds performance increases with increasing dimensionality towards the original 512 dimensions, simply put the projection mechanism can more closely approximate the optimal modulation weighting for a specific filter, maximising performance on the task-set.

Figure 4.8 shows task-generic seed weights stayed relatively consistent as dimensionality increased from 17 dimensions to 256 dimensions with increasing standard deviations within the 17-90 dimensional range, and subsequently decreasing standard deviations for dimensions beyond 90. All task-generic weightings feature a similar distributional shape with most weights close to the means which were found to be almost 0. Taking care of the x-axis limits, the 512 dimensional mechanism tended to center weights more sharply around the mean, with a standard deviation around 7 times smaller than the next smallest deviation.

Task-specific seed weights changed more drastically as seed dimensionality increased in-line with the hypothesis. At lower dimensions, the seed weights distributions are flatter with wider tails either side of the means and were significantly varied, with values ranging from -10 to 10. As seed dimensionality increased beyond 90 dimensions, the distributions began to reflect a more bell-shaped curve, with a higher concentration tending towards the means. Standard deviations decreased from 4.5 to 0.13 as seed dimensionality increased. The 512 dimensional seed models tended to have very similar distributions between both task-generic and task-specific seed weights which follows our hypothesis that more finer control is allowed and hence smaller seed weightings are needed regardless of the task-set in question.

Interestingly, the task-specific seed weightings concur with our hypothesis, yet the task-generic seed weightings do not entirely agree with the hypothesis as they consistently stay concentrated around the mean. The hypothesis stated above is closely linked with model performance changing with dimensionality. As seen previously, the task-generic models do not observe the same change in performance with dimensionality as the task-specific models, hence explaining this phenomena.

There is no benefit for the weightings to deviate too much from the zero-mean, as this projects to attention weights very close to 1.0. Understandably, with the task-generic models maintaining similar performance to the original pre-trained VGG-16 model, the attention modulation will favour weightings that do not modulate filters as performance in a task-generic sense was already maximised under the pre-training of the VGG-16 model.

Projected Attention Distribution

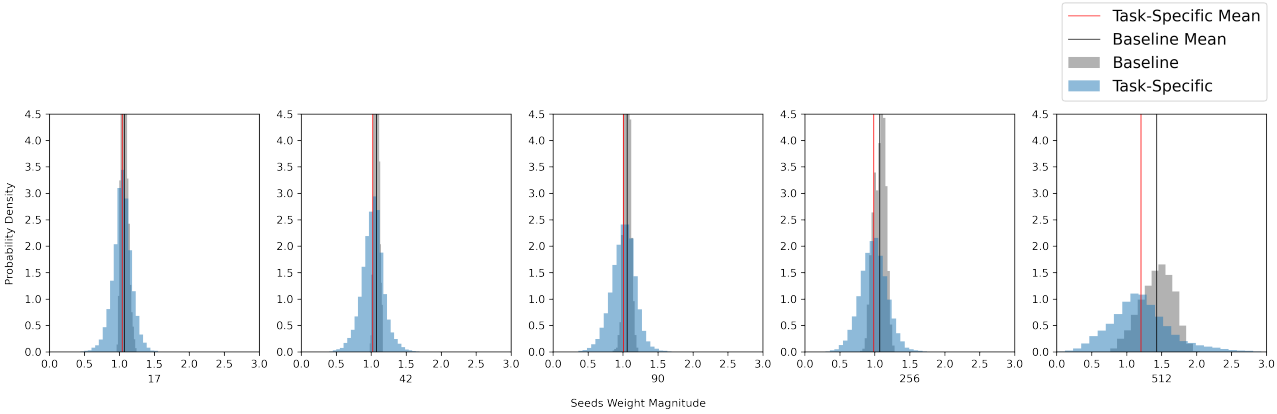


Figure 4.9: Comparing task-generic and task-specific projected attention weights. With increasing dimensionality of seeds, task-generic weights became more diverse and spread out with largely positively shifting distributional mean. Task-specific weights also exhibit the same spread in distributions with a smaller translation of the distributional mean with increasing dimensionality. Again, a lower number of attention seeds meant the attention seeds had to be of a larger scale and more diversified in order to achieve the optimal final attention weight. This meant the approximation to the optimal final attention weight would be more largely deviated the lower the number of attention seeds available, hence meaning slightly augmented distribution of final attention weights. The approximation to the optimal distribution (signal-detection metrics shows this was exhibited by the 256 dimensional model) is still quite good at the lower dimensional models as the range in distributional values is similar amongst all models. The same reasoning can be used to explain the shift in distributional means to the optimal mean. The 512 dimensional mechanism exhibited the greatest difference in distributions amongst both task-generic and task-specific weights to all other distributions, adding reasoning to why performance degradation was exhibited by the 512 dimensional models.

Figure 4.9 shows that projected attention weights not only changed in distributional spread across increased seed dimensions, but also translated means about the initialisation point of 1.0. At lower dimensions, both distributions have means around the initialisation point with a more concentrated distribution of projected attention weights. For all dimensionalities, the task-specific attention

weights have a higher standard deviation in values than the task-generic attention weights. As dimensionality increases, task-generic and task-specific distributional means begin to separate with the task-generic distribution translating at a more rapid rate. Both distributional tails increase with increasing dimensionality meaning more drastically different attention weights. Between dimensions 17-256 the task-specific means approaches the initialisation point of 1.0, while the 512 dimensional seed mechanism shifts to a mean of around 1.25. This is an interesting phenomena as it seems the seed weightings themselves contract around distributional means with increasing dimensionality, whereas, by contrast, the projected attention weights expand around the mean point with increasing frequency of filters turned off due to zero attention weights.

There is no clear explanation as to why 512 dimensional task-generic projected attention weights tended to favour up-scaling of features, with little projected attention weights attenuating below 1.0. In fact it is counter-intuitive as model performance was previously thought to be best maximised when filters are not modulated under the task-generic setting. Reflecting on the performance degradation seen in signal-detection metrics for the baseline models (Figure 4.6), this indeed could explain the performance degradation. With the projected task-specific weightings for the 512 dimensional mechanism having attenuations well below 1.0, the non-negative constraining operation may not be to blame for the performance degradation, but an unknown factor that influenced task-generic weightings to severely modulate features upwards.

Spearman correlations between increasing lower seed dimensionalities (17-42, 42-90, 90-256) were somewhat consistent yielding correlations between projected attention weights of 0.56, 0.55, and 0.41, respectively. A lower correlation was observed between the 256-512 pairing at 0.26, which is attributed to the larger difference in mechanism dimensionality.

It is important to note that in general, training the attention seeds weights, or more generally neural network parameters, requires a trade-off in features and underlying feature-extraction parameters during the learning process such that overall performance is maximised. Features within a particular filter can enable the network to better discriminate the class, while also being confounding when identifying other classes. The attention mechanisms modulate such features with weights determined by the filter contribution to overall network performance. Within the task-generic setting, this feature trade-off is more scrutinised, often requiring weightings optimal for all ImageNet classes - hence explaining the similarity in distributions across dimensions for task-generic weights. The feature trade-off is less stringent within task-specific models as the model is optimising features solely for its target class, often leading to a disconnect in learned weightings across different task-sets, with wider varying weights in comparison to the task-generic baseline models.

4.2.4 Summary of Key Findings: Attention Weights Analysis

Overall it was found that the effect of dimensionality strongly influences the distributional properties of task-specific attention seeds and projected attention weights. This influence is less so exerted on task-generic attention weights. The hypothesis that a dimensionality-reduction causes the model to more drastically warp attention seeds under the influence of factors that drive the learning of attention weights appears to be true. At lower dimensionality, task-specific attention seed distributions were more greatly diverse. As dimensionality increased to the full-dimensionality of the filter-space, attention seed distributions became less flat and more bell-shaped, with distributions concentrating more tightly around the zero-mean. In contrast, projected attention weights became slightly less concentrated around the distributional means as dimensionality increased. This result is interesting but is simply an artifact of projection mechanism in that higher dimensionality of attention seeds means more fine-grained control in the final projected attention weights more closely approximating the optimal attention weights.

4.2.5 Representational Similarity Analysis

Under stage two of the thesis, this experiment aims to identify whether similarities between learned seed weightings follow corresponding similarities between ground-truth semantic embeddings in an attempt to give an intuitive sense of the representational geometries underpinning the learning process of attention weights. This experiment also compounds analysis of the effects that class-difficulty and variance in activations different superordinate groupings has on the semantic similarities between classes, thereby possibly uncovering which factors are most closely linked with the warping of representational geometries of attention weights, explaining why attention weights are learned how they are. We first give a brief recap on the setup and motivations of this experiment:

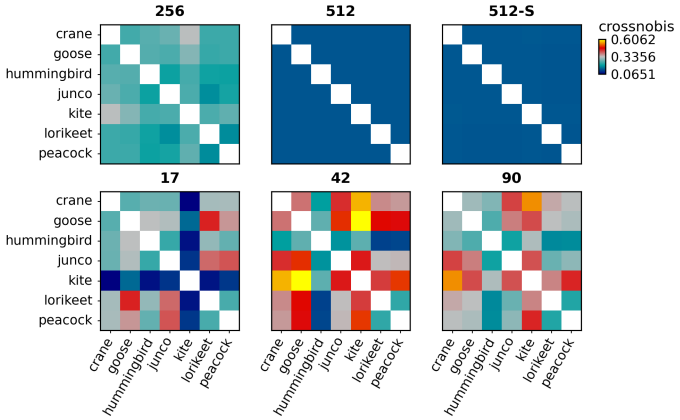
Recap on Experimental Details and Motivation

In this experiment, RDMs are constructed for learned seed weightings for each task-set at each set dimensionality, with RDMs cross-validated across trials of learned seed weightings for a particular task-set. RDMs were also constructed for class label semantic embeddings. The RDMs capture the dissimilarities between underlying factors of the attention seeds and semantic embeddings, respectively. These RDMs are then compared via a Spearman correlation measure. This comparison will allow us to better identify how closely matched the representational geometries of the attention seeds are to the the semantic embeddings, thereby allowing us to see how strong the influence of semantic entanglement is on the learning process of attention weights. With the target-classes chosen to produce variation in class-difficulty and variance of activations, we can then assign reasoning to why certain groups correlated highly and why did not, therefore we gain an understanding of how these latter factors also influence the learning process of attention weights.

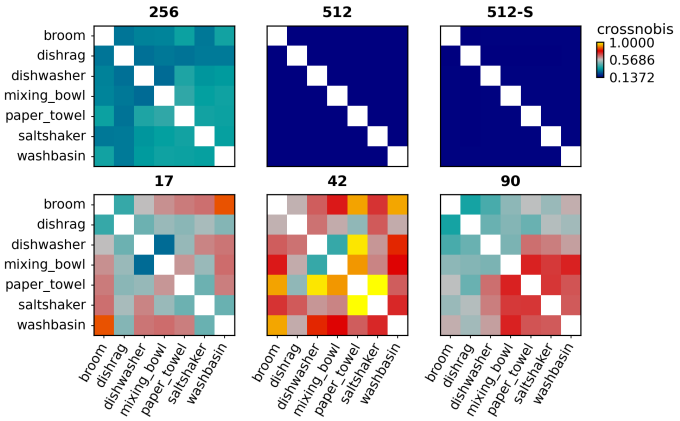
Understanding more than just model-level dynamics is helpful in building an overall picture of how top-down attention interplays with features of the DCNN when optimising for model performance. Utilising RSA we can identify if the attention seeds mechanism tends to better learn weightings that more easily allow us to spot patterns of representational similarities. As a reminder, if any of the previously mentioned factors were found to strongly influence the learning process of attention weights, then this will guide future experimentation by giving a priori on which types of factors are most at play within attention mechanisms. Our final goal is to fully understand the unknowns of the characteristics of attention within neural networks. Understanding this can revolutionise the field in building more specialised attention mechanisms better suited to different applicational tasks more easily.

During this experiment we found it beneficial to incorporate a new superordinate grouping of task-sets to lend a perspective more geared towards the factor of semantic similarity between class labels. Originally, target-classes were chosen based on factors such as explained variance in class activations and class difficulty. As such we introduce the Dogs (Felidae) superordinate, a grouping found to contain highly semantically and visually entangled target-classes. This gives better representation of the external factors class as described in Section 3.4.1, helping us uncover its influence more easily. Clustering in the same manner as outlined in Section 3.4.1 was used to identify an additional 7 target-classes to comprise the new task-sets. The new task-sets were trained in the same manner as the Avian and Kitchen task-sets (see Section 3.4.3.) The Dogs target-class set can be found in Appendix A.

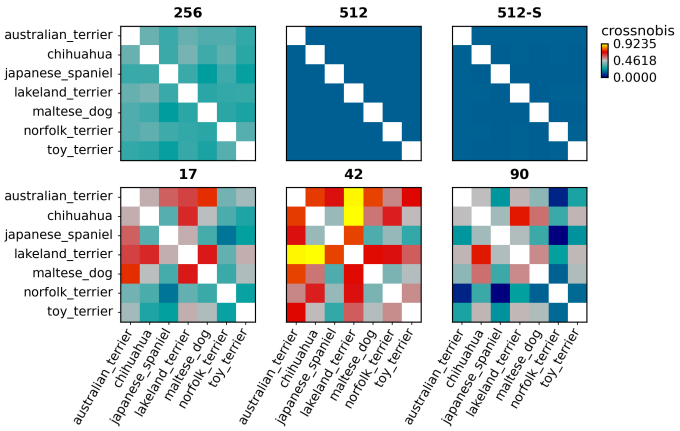
Visualising RDMs



(a) Avian grouping RDM



(b) Kitchen grouping RDM



(c) Dogs grouping RDM

Figure 4.10: (Continued on the following page.)

Figure 4.10: Visualising attention seeds RDMs for increasing attention seed dimensionalities, and for the standard goal-directed attention mechanism. RDMs are symmetrical matrices that capture the dissimilarities between a set of measured stimuli through a distance metric. As we attempt to link similarities between learned attention weights to similarities between their semantic relationships we compare RDMs between learned attention weights to RDMs between semantic embeddings of class labels. Dissimilarities were estimated through the cross-validated Mahalanobis distance. With each target-class having 3 attention models trained at each dimensionality, we cross-validate across all 3 learned seed weightings in order to reduce noise and give a more accurate dissimilarity measure between classes. Lower dimensional mechanisms (17, 42, and 90) capture the greatest dissimilarities between the attention seeds within a grouping. Higher dimensional models tended to learn attention weights that were much more similar, with the 512 dimensional models did not capture dissimilarities well with very low distances observed between classes, hence leading to a decreased explanatory power. We predict the lower-dimensional (42, 90, 256) attention mechanisms to better correlate the semantic embeddings. All RDM values were normalised to a global range $\in [0, 1]$ before plotting for convenience of comparison.

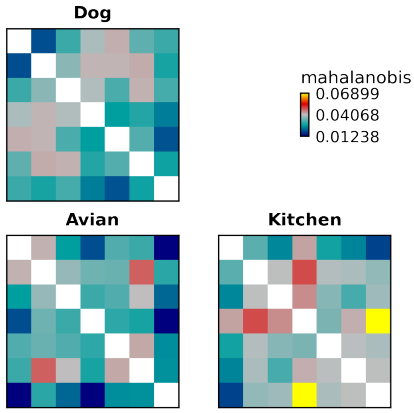


Figure 4.11: Visualising ground-truth semantic RDMs for each superordinate grouping. These RDMs capture dissimilarities between ground-truth semantic embeddings of the target-classes. As we attempt to link similarities between learned attention weights to similarities between their semantic relationships we compare RDMs between learned attention weights to RDMs between semantic embeddings of class labels. Dissimilarity was estimated using the normal Mahalanobis distance. Cross-validation was not required given the ground-truth semantic embeddings are noiseless. Dissimilarity magnitudes are of a much smaller scale relative to the dissimilarities of the attention weights. This is expected given target-classes were chosen to maximise semantic entanglement, hence semantic embeddings will be more similar. Dissimilarities are still present between semantic embeddings. The semantic RDMs most closely resemble the 90 and 256 dimensional attention seeds RDMs, hence we expect these lower-dimensional models to exhibit higher correlations, which would provide evidence for the lower-dimensional attention seeds mechanism in enabling better representational analysis in comparison to the full-dimensional attention mechanism. All RDM values were normalised to a global range $\in [0, 1]$ before plotting for convenience of comparison.

Figure 4.10 show the computed RDMs for the Avian, Kitchen, and Dog learned seed attention weightings, computed using the cross-validated Mahalanobis distance. Figure 4.11 shows the RDMs for each groupings semantic vectors, computed using the Mahalanobis distance. All seed RDMs show an overall decreasing mean dissimilarity for increasing seed dimensionalities, with the 42 dimensional seed weightings exhibiting the greatest dissimilarities for all superordinate groupings. Changing dimensions appears to be a significant factor in learned weighting dissimilarities. The 512 dimensional models appear to have near similar learned weightings across all groupings, with a minimal mean observed distance. This is most likely due to the same effect observed when analysing the distribution of attention seed weightings in that higher dimensions tended to concentrate seed weightings more tightly around the zero-mean, leading to vectors of higher similarity and magnitudes. This shows that the full dimensional attention mechanisms do not capture the dissimilarities in patterns between attention weights in comparison to the lower-dimensional attention weights. This suggests that the lower-dimensional attention weights can correlate better with the semantic embedding RDMs.

Immediately we can see that distance differences between the semantic vectors are of a small order of magnitude, suggesting the high semantic similarity within superordinate groupings is represented well with the Mahalanobis distance. Magnitude differences do not influence RDM comparisons as the Spearman correlation metric used is magnitude-insensitive. The Kitchen superordinate appears the greatest dissimilarities in not only the semantic RDMs but also the learned seed weighting RDMs. The Kitchen superordinate was chosen due to its high superordinate difficulty and explained variance of activations. These factors appear to influence the dissimilarities amongst semantic embeddings and seed weightings. Perhaps with an increasing superordinate difficulty, seed weightings tended to differ more in order to aid in discriminating such similar classes with high similarity in variance of activations. Overall, from visualisations we can already see the benefits of dimensionality-reduction as lower-dimensional attention weights exhibit an increased ability in capturing underlying patterns between the task-sets over the full-dimensional standard attention mechanism. We therefore predict that these lower-dimensional models will exhibit higher correlations to the semantic embeddings. If this is true, we can therefore sensibly extract a semantic space from the attention weights, meaning semantic entanglement can be a factor influencing the learning process of attention weights.

Comparing Individual Semantic Vector RDMs to Attention Seeds RDMs

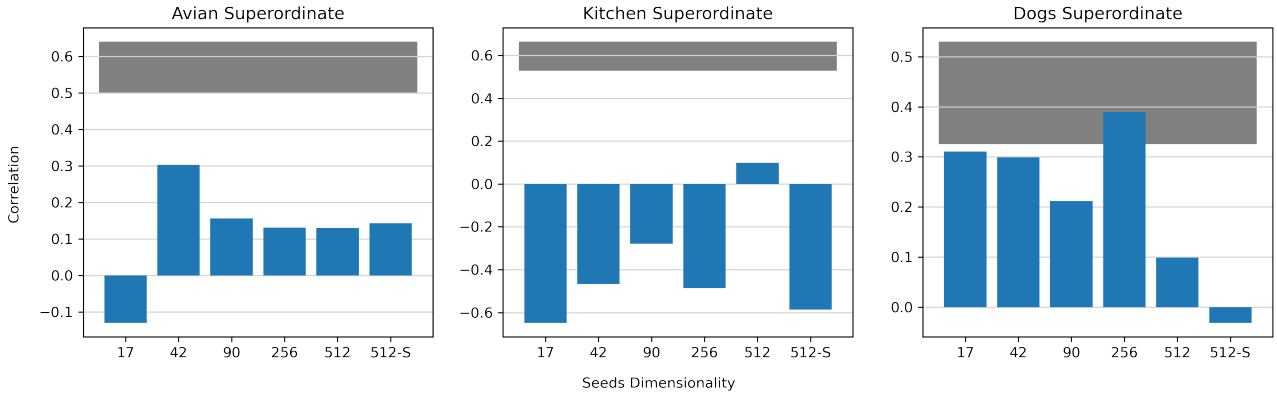


Figure 4.12: Spearman correlation comparisons between the candidate seeds RDMs and the respective grouping's ground-truth semantic embeddings RDMs at each trained seeds dimensionalities. The standard goal-directed attention mechanism RDM comparisons are denoted '512-S'. This allows evaluating if the dimensionality-reduced weights indeed have an increased explanatory power relative to the underlying semantic relationships between task-sets. Across all groupings the dimensionality-reduced attention weights correlated better than the full-dimensional attention weights providing evidence to the increased explanatory power of the lower-dimensional weights, inline with our hypothesis. The Avian and Dogs groupings were found to be positively correlated whereas, unexpectedly, the Kitchen grouping was found to be negatively correlated. The ordering of mean correlations across groupings did not align with orderings of mean class-difficulty or variance of activations, suggesting these factors may influence the attention seeds in opposing manners, or did not influence them at all. The Dogs superordinate learned weights with similarities that closely aligned to respective semantic similarities between classes, however the presence of such a large and low noise ceiling suggests the analysis contained a lot of noise which must be factored when concluding inference. A noise ceiling (the grey horizontal bar) is calculated to see how well a perfect model would perform in the presence of noise found in the data.

Table 4.3: Mean class accuracy (difficulty) of task-specific models for each superordinate grouping at each seeds mechanism dimensionality. The bottom row shows the mean accuracy for a particular dimensionality across all superordinates. The final column shows the mean accuracy for a particular superordinate across all dimensionalities of models. Orderings of mean accuracies between superordinates do not follow orderings of mean correlations seen between seed RDMs and semantic embeddings RDMs, nor do orderings of mean accuracies across dimensionalities. This suggests class-difficulty does not influence similarities of seeds to the semantic embeddings.

Grouping	**Dimensionality**					**Mean Accuracy**
	17	**42**	**90**	**256**	**512**	
Avian	0.8626	0.8715	0.8947	0.9332	**0.9450**	**0.9014**
Dog	0.6417	0.7307	0.8138	**0.8796**	0.8768	0.7885
Kitchen	0.7297	0.7730	0.8110	**0.8610**	0.8552	0.8060
Mean Accuracy	0.7447	0.7917	0.8398	0.8913	**0.8923**	-

Table 4.4: Explained variance of NMF-extracted basis components of all samples activations from each superordinate grouping of the pre-trained VGG-16 model. Activations were extracted from the same convolutional layer the attention seeds mechanism operates on, under the same technique outlined in 3.4.1. The furthest right column shows the mean explained variance of activations for a particular superordinate. A higher explained variance value indicates very high correlations within activations, hence indicates a high similarity between superordinate target-class samples. Orderings of mean explained variance between superordinates follows orderings of mean correlations seen between seed RDMs and semantic embeddings RDMs. This suggests that a low explained variance corresponds to attention weights that are more similar, hence correlating better with the semantic embeddings and gives much insight into the type of factors that guide the learning of attention weights.

Grouping	**Explained Variance of NMF Topics**			**Mean Explained Variance**
	Topic 1	**Topic 2**	**Topic 3**	
Avian	0.3404	**0.6753**	0.3235	0.4464
Dog	0.2751	**0.4332**	0.3139	0.3407
Kitchen	0.2593	**0.8041**	0.3618	**0.4750**

The correlation results for each superordinate grouping's seed weighting RDMs to their respective semantic embeddings RDMs are plotted in Figure 4.12. The standard attention mechanism correlation was also computed to provide a reference point. Seed attention weightings were found to best correlations of the standard top-down attention mechanism with the semantic embeddings. Even the Dogs superordinate, which successfully correlated all seed attention weightings with the

semantic embeddings, negatively correlated the full-dimensional standard attention mechanism weightings. This provides further evidence to the hypothesis that low-dimensional attention weights are more strongly warped under model influences driving the dissimilarities in learned attention weights, hence it was easier to correlate them. It can therefore be seen as favourable to utilise the seeds mechanism in increasing the likelihood of extracting meaningful patterns from learned attention weights relative to the underlying relationships exhibited between classes.

The Avian and Dog grouping seed weightings were found to be mostly positively correlated to their semantic embeddings, with the Dog grouping exhibiting the greatest Spearman correlation of 0.39 at dimensionality 256, and exhibited much stronger correlations on average than all other groupings. Surprisingly, the Kitchen grouping was found to be largely negatively correlated with its semantic embeddings. In each grouping a different lower-dimensional seed dimensionality was found to better correlate to the representational geometry of its respective semantic vectors, suggesting no one dimensionality can be recommended when attempting to identify the influence of factors.

Interestingly the 512 dimensional seed mechanism was found to be positively correlated with a similar magnitude across all groups. This could be due to the same effect identified with distribution of seed weightings in that weightings became more similar in magnitude as dimensionality increased. Naturally, this leads to increased similarity amongst weightings through the Mahalanobis distance as seen in the 512 dimensional weighting RDMs. Due to target classes being chosen to be semantically similar, the RDMs representing the semantic vectors will naturally be highly similar too. This means that there is a much higher probability of a positive correlation regardless of the learned seed weighting, suggesting the result at such high dimensionalities may be biased. However, the distance argument presented does not appear to entirely explain this phenomena as closer distanced dimensions, such as the 256 dimensional RDMs, did not always exhibit the highest correlations, and in the Kitchen case exhibited a significantly negative Spearman correlation meaning there may not be a bias generated at dimensions lower than the full filter-space dimensionality of 512.

The Effect of Class-Difficulty on Correlations

Class difficulty does not appear to be driving the correlations to semantic embeddings either. Table 4.3 shows that the lowest mean grouping accuracy was achieved at dimension 17, yet this dimensionality does not always exhibit the lowest correlations as would be expected given the lower degrees of freedom, and therefore less representational power. The highest mean superordinate accuracy is achieved by the 512 dimensional models, yet, again, the correlations

associated at this dimensionality are neither the highest nor the lowest. Given the ranking of similarities of the representational geometries within groupings, the most correlated was found to be the Dog grouping, followed by the Avian and Kitchen groupings, respectively, however this ordering does not coincide with the ordering of mean grouping accuracies. The Dog grouping appears to have the lowest mean grouping accuracy across dimensions at 0.7885 and the Avian grouping holding a significantly higher mean grouping accuracy at 0.9014. These results show that class difficulty is may not be driving correlations between learned weightings as previously thought when selecting target classes.

The Effect of Similarity of Activations on Correlations

The explained variance of filter activations does indeed provide an explanation for the ordering of correlations. The ordering of mean explained variance across fundamental basis components extracted from superordinate activations follows the ordering of mean correlations identified by the RSA results in Figure 4.12. The Dog superordinate exhibits a significantly lower explained variance at 0.34, followed by the Avian and Kitchen at 0.45 and 0.48, respectively. This suggests that a low explained variance, which means decompositions of activations amongst that superordinate were less correlated with each other, corresponds to attention weights that are more similar, hence correlating better with the semantic embeddings. An intuitive explanation for this is that a lower explained variance means different target-classes of a superordinate passed through the model were generally more dissimilar. This means the model would find them less confounding to each other (leading to less difficulty) as the higher dissimilarity allows the model to better discriminate the target-class. This explanation is supported further by the relationship explained variance had with class-difficulty in Figure 4.1 (see Section 4.1.1), in that lower explained variance meant lower mean grouping difficulties. Since they are less confounding, this leads to more similar attention weights as the attention mechanism does not need to warp activations as heavily, if the classes were confounding. This in turn means they correlate better to the semantic embeddings, as they are also similar to each other. This provides evidence that explained variance of class activations is an influencing factor driving the learning of attention weights.

Comparing RDMs from Interpolation Mixture Models across Dimensionalities

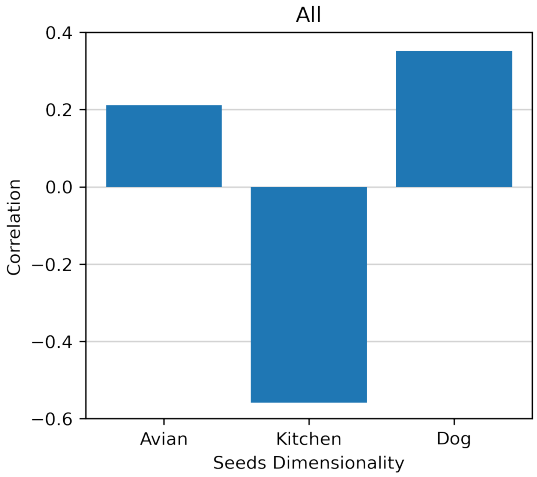


Figure 4.13: Spearman correlation comparisons between the interpolated model RDMs and the respective grouping's ground-truth semantic embeddings RDMs. The interpolation models interpolate a single RDM representative of RDMs across all seed dimensionalities for a particular superordinate grouping. This allows an overall evaluation irrespective of the dimensionality of seeds weights being considered within the comparison. Given the comparison involves all superordinate groupings and the interpolated dimensionality RDMs, it is much harder to estimate a noise ceiling that is accurate due to compounding of multiple noise sources from within the data, hence we omit a noise ceiling in this plot.

The interpolation models construct a single RDM representative of a mixture model of RDMs across all dimensionalities for a particular superordinate grouping. This allows a more overall comparison between the attention seeds RDMs to the semantic embeddings RDMs, summarising all trends in correlations into a single Spearman correlation metric. The Spearman correlation between the model outputs and semantic ground-truth RDMs are shown in Figure 4.13. As expected, the Dog class still retained the highest Spearman correlation at 0.35, followed by the Avian and Kitchen sets at 0.21 and -0.55 respectively. This ordering is again supported by the ordering of mean explained variance in filter activations (Table 4.4). What is strange is that the Avian groupings mean explained variance in activations is only slightly smaller than that of the Kitchen grouping, yet the Kitchen grouping semantic embeddings were significantly negatively correlated to the learned attention seed weightings. This suggests that although variance in activations could be an influencing factor driving attention weights, this influence is most likely at most moderate.

Analysis of Noise within Comparisons

Correlation values were found to be very similar between the cross-validated Mahalanobis distance RDMs and the normal Mahalanobis distance RDMs with decreasing effect of noise cancellation as dimensionality increased the optimal weighting was being achieved more consistently as dimensionality increased. The noise ceiling indicates a bound on the stimulus-related variance and indicates a range for the best performance achievable by the true data-generating model (optimal seed weightings) when fitted to the semantic embedding RDMs. We note that correlation values were not corrected for under the noise ceiling estimate as we found it more beneficial for simultaneously evaluating the fit and quality of the learned seed weightings. Relative to the noise ceilings, the Avian grouping appears to be weakly correlated, with a peak correlation of 0.3, approximately half of the maximum attainable correlation to the semantic embeddings. The Dog superordinate achieves correlations very close to the noise ceiling, with the 256 dimensional seed mechanism correlating within the noise ceiling. This suggests that the weightings were strongly correlated in the presence of noise with the semantic embeddings, despite not exhibiting correlations close to 1. Training a greater number of task-sets is known to reduce this noise amongst comparisons. Taking more precaution when identifying optimal training hyperparameters can also aid in reducing noise amongst folds of learned attention weights.

This comparison reveals that the optimal seed weightings learned vary greatly in capturing underlying semantic relationships between classes due to limitations placed by unknown factors influencing the representational geometries underpinning the attention seeds training process. Such factors may include the inherent training-set distribution between target-class examples a non-target-class examples, as this most closely underpins the training process for learned seed weights. For example, removing confounding class examples within the training set could remove any divergent influence on seed weights, causing them to converge to their true optimal task-set weighting that could most likely tie back to class-level similarities under an RSA.

Comparing All Semantic Vector RDMs to Attention Seeds RDMs

In order to provide the best chance at extracting a trend in correlations between similarly grouped classes, previous experimentation limited the scope of comparison to classes within a single grouping, hence considered only similar classes when correlating seeds to the semantic embeddings. In this experiment, we consider correlating RDMs by mixing all task-sets together. Correlations of this nature will consider not only the similarities between task-sets, but accounts for the dissimilarities amongst seed weightings that are not within the same semantically entangled group. A single RDM was constructed for all learned seed weightings for each dimensionality, and another RDM

constructed for all semantic embedding vectors. The ground-truth semantic RDM was compared to the seed RDMs for all dimensionalities using the same Spearman correlation metric.

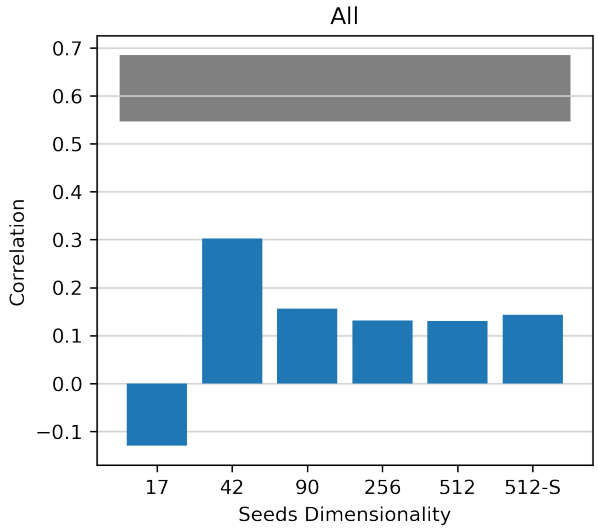


Figure 4.14: Spearman correlation comparisons between an RDM constructed with all target-classes across all groups and a ground-truth semantic embeddings RDM, also constructed with all target-classes across all groups, at each trained seeds dimensionalities. For reference to the standard goal-directed attention mechanism, RDM comparisons between the standard attention mechanism's learned weights and the ground-truth semantic embeddings for all target-classes is also shown, denoted '512-S'. This comparison effectively compounds all RDMs into a single RDM in order to gauge overall correlations across all learned attention weights to all semantic embeddings. Previous experiments considered correlations within a single grouping and hence contained only similarly entangled classes. This has the advantage of computing correlations that take into account the similarly grouped classes, and also their dissimilarities to other groupings, a more reliable measure of correlation is therefore rendered. This method also reduced noise amongst the RDM comparison process due to increased number of stimuli considered. A noise ceiling (the grey horizontal bar) is calculated to see how well a perfect model would perform in the presence of noise found in the data. Aside from the 17 dimensional model which was found to be highly noisy, lower-dimensional attention weights exhibited higher correlations to the semantic relationships between classes. This, again, provides evidence to the increased explanatory power of lower-dimensional attention weights, increased the likelihood of relating representational similarities to investigated factors thought to influence the learning of attention weights, hence proving our hypothesis. The noise ceiling is much smaller than previously suggesting less noise was present within the compared data due to increased number of stimuli considered. The 42 dimensional attention weights most closely replicated the semantic similarities and dissimilarities between semantic embeddings of class labels, suggesting this could be an optimal dimensionality for future experiments that look at different factors that influence the learning of attention weights.

Overall, Figure 4.14 a positive correlation was identified when comparing representational geometries of all learned attention seed weightings to their respective semantic vectors at each dimensionality. Most positive correlations were found to be weakly correlated with a moderate correlation of 0.3 exhibited by the 42 dimensional seed mechanism, less than half of the maximum attainable correlation to semantic embeddings, suggesting that semantic similarity is a weakly influence factor on the learning of attention weights.

Under consideration of all results evaluated thus far, no one dimensionality can be concluded to be the best for identifying the influence different model-related factors exert on the learning process of attention weights as a range of dimensionalities from 42 to 256 appear to correlate the most consistently. Most importantly, the seeds mechanism bested the standard attention mechanism, again, in correlations to the semantic embeddings between classes. The noise ceiling for this comparison was found to be moderate, at around 0.27 in height, however, given the mixture of superordinate groupings such a variance is expected. Nonetheless, a peak attainable correlation (upper-bound of noise ceiling) of 0.68 speaks to the limitations and or deficiencies rendered through a noisy training process. It is clear utilising a greater number of task-sets and training more seed weightings per task-set could yield a more robust analysis with less repercussions of the noise presented during training.

Correlations of Attention Weights Across Dimensionalities

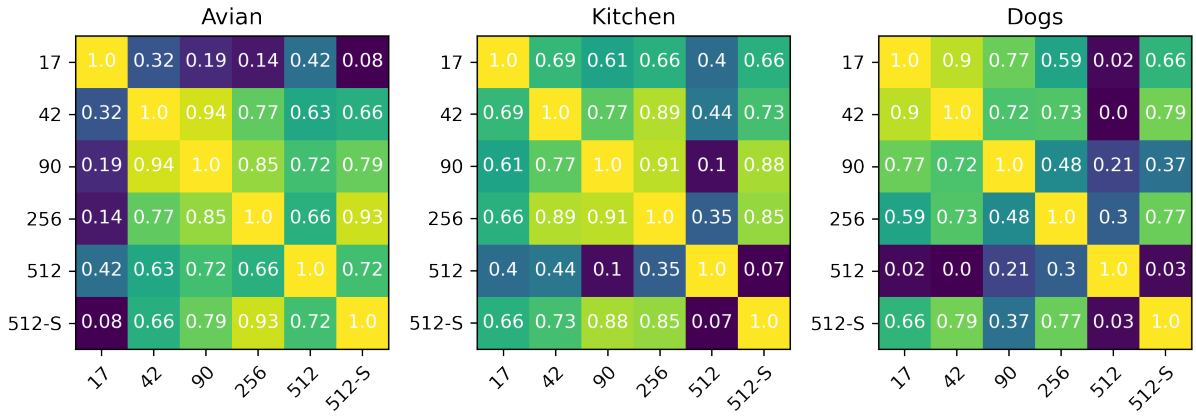


Figure 4.15: Visualising Spearman correlations between learned seeds weights RDMs and standard attention mechanism weights RDMs between each other for each superordinate grouping. In an attempt to explain the results above from a similarity-of-weights point-of-view this comparison was considered. It also allows us to identify how aligned lower-dimensional seeds weights are with the weights learned by the full-dimensional standard attention mechanism, denoted '512-S'. Most low-dimensional attention weights learned appear to be strongly correlated to the 512-S weights for each superordinate grouping, meaning high correlations between the standard attention mechanism and lower-dimensional seeds weights. High correlations mean low-dimensional attention weights indeed represent a high proportion of informational content found within the high-dimensional inline with our hypothesis that there exists a low-dimensional set of latent factors driving the high dimensional attention weights. The 512 dimensional seeds mechanism weights tended to exhibit low correlations suggesting overall they did not align representations to the standard attention mechanism, despite sharing the same number of attention parameters. This result does not matter as we care more about the dimensionality-reduced attention weights. Aside from the Avian grouping, the 17 dimensional model at a 97% reduction in attention parameters was very highly correlated to the full-dimensional attention weights, providing strong evidence for the hypothesis.

Figure 4.15 shows Spearman correlations between seed RDMs at all dimensionalities for each superordinate grouping, and for the standard attention mechanism. The lower-dimensional seeds weights appear generally strongly correlated with the full-dimensional standard attention mechanism weights. This provides strong evidence to support the hypothesis that there is a lower-dimensional set of latent factors driving the high-dimensional attention weights. In particular, we saw earlier that the 256 dimensional seeds mechanism, at a 50% reduction in trainable attention

parameters, maintained model performance on a variety of signal-detection metrics. When looking at the bottom row of Figure 4.15, the 256 dimensional seeds weights are consistently strongly correlated with the standard attention mechanism weights with a minimum and peak correlation of 0.77 and 0.93 exhibited by the Dogs and Avian grouping weights, respectively. Such strong correlations across all 21 target-classes between the lower-dimensional seeds and the standard attention mechanism weights. Even the 17 dimensional seeds mechanism at a 97% reduction in attention parameters exhibits very high correlations of 0.66 in both the Kitchen and Dogs groupings! If such a large amount of informational content is being retained by these lower-dimensional attention weights, these must be the low-dimensional latent factors driving the high-dimensional attention weights, hence the hypothesis appears mostly correct.

4.2.6 Summary of Key Findings: Representational Similarity Analysis

Overall, the lower-dimensional attention seeds mechanism consistently learned attention weights with representations more closely correlated to the underlying semantic embeddings of target-classes, with the best correlations exhibited by the 42, 90, and 256 dimensional mechanisms. Lower-dimensional attention RDMs captured the dissimilarities between attention weights well, while the full-dimensional attention weights did not. Attention seeds correlations to semantic embeddings were found to, at most, be moderate suggesting that semantic similarities between classes is not a largely influencing factor that drives similarities between attention weights, but is a factor nonetheless. Orderings of decreasing semantic correlations (Dog then Avian then Kitchen) aligned with orderings of increasing mean explained variance of class activations, meaning correlations between target-class activations are an influencing factor driving the learning of attention weights. This factor was found to be moderately influencing. Orderings of correlations did not follow orderings of class-difficulty (Avian then Kitchen then Dog when compared to the ground-truth semantic relationships between classes, suggesting it did not largely influence the semantic entanglement of learned attention weights. We noted that, with differences in mean correlations identified by different superordinate groupings under the same training procedure, we believe the investigated factors are compounding in warping the representational geometries of the attention weights, hence further experimentation through an ablation study design may be required to confirm strongly our conclusions presented here.

Low-dimensional seeds weights were found to be highly correlated to the full-dimensional standard attention mechanism weights, even at a 94% reduction in parameters for the 17 dimensional model, meaning low-dimensional attention weights indeed represent a high proportion of informational content found within the high-dimensional attention weights, providing strong

evidence for our hypothesis that there exists a low-dimensional set of latent factors driving the high dimensional attention weights and explains why performance maintenance was found in earlier results.

To conclude, there appears to be a more underlying factor that strongly influences learned seed weightings beyond seed dimensionality and model-level cues used to select target-classes that could not be entirely explained by experiments conducted thus far. This factor most likely closely relates the confounding nature of target-classes to the model being considered, and indeed such a factor most likely relates to the confusion between such target-classes as noted by our explanation of why explained variance in activations was a moderately influencing factor. It is entirely possible to speculate that there are too many factors at play compounding possible effects seen within the RSA, hence making spotting a trend much more difficult and leading to an overall lower quality of analysis. It is recommended that systematically considering all factors individually, say through an ablation study, can aid in identifying the most dominant factor that will allow for more guided attention seeds training that maximises similarity amongst similarly grouped classes.

4.2.7 Further Experiments Isolating Class Difficulty

This section details results of one further experiment performed that aims to isolate the factor of class-difficulty and its influence on the learned representational geometry of attention seeds. Previously, we concluded that the compounding class-level factors negatively impacted the quality of the RSA performed, hence made it harder to conclude strong inferences from the correlations. We showed that from the three considered factors, semantic similarity and similarities in inter-layer activations elicited between different target-classes were influencing factors that drove the learning of attention weights, however, no conclusion could be reached with class-difficulty. We now choose to isolate class difficulty, a factor previously found to be most closely tied with the effects of goal-directed attention [Smith et al., 2021], and examine its influence over attention seeds.

Experimental Setup

From previous experiments, we identified the most consistent seed dimensionalities for RSA were found to be the 42, 90, and 256 dimensional seed mechanisms. The 17 dimensional mechanism exhibited very high noise and variance amongst trials of learned weights under a single task-set. The 512 dimensional model was discarded as previously seen the 512 dimensional mechanism exhibited many unexplainable artifacts when analysing attention weights and correlations, that we

feel may skew the RSA performed here.

From our RSA results we hypothesised that the underlying factor thought to warp representational geometries is the confusion-rate between task-sets. We previously only considered classes that were similar to each other under a superordinate grouping. We postulate that attention may not be similar for similar classes as they tend to contain confounding items that cause confusions and errors, hence they may need to be separated from one another. We predict that two task-sets should have the same attention weights if they are hardly confused with each other, but exhibit the same confusions with other classes.

To test our hypothesis, we assemble a target-class set consisting of the top-20 most difficult and the top-20 easiest ImageNet-1k classes, meaning a target-class set of 40 distinct target-classes. Difficulty is given as the error rate of our task-generic baseline model on images from a target-class and was computed for all 1000 ImageNet-1k classes. For clarity of presentation, the top-20 hardest and top-20 easiest classes that formed our target-class set are listed in Appendix A. The target-class set used is ordered with the easiest classes in the first 20 indices, and the hardest classes in the last 20 indices. **This order is maintained throughout all experimental results and figures presented here**. Attention seeds models were then trained for each of the 40 target-classes in line with the same training procedure as all other experiments (see Section 3.4.3). Two confusion matrices are then constructed. We define confusion rates for a particular target-class as the unnormalised count by which each non-target-class was predicted as the most-likely class when the baseline model predicts on all examples from the target-class. This is not an issue in comparisons as measures are normalised before computing RDM distances. Furthermore as stated previously, the Spearman correlation is magnitude-insensitive. The first confusion matrix is a 40×40 matrix. The rows indicate the target-class with columns indicating confusion rates with the remaining 39 target-classes. Confusion rates were computed using the task-generic baseline model at each selected dimensionality. This matrix is used to demonstrate the relative confusion between the 40 chosen target-classes under the first condition of our hypothesis. The second confusion matrix is a 40×1000 matrix where each row indicates the confusion vector for one of the 40 target-classes. The confusion vector for a target-class is populated by the task-generic baseline confusion rates of the target-class with each of the remaining 999 ImageNet-1k classes.

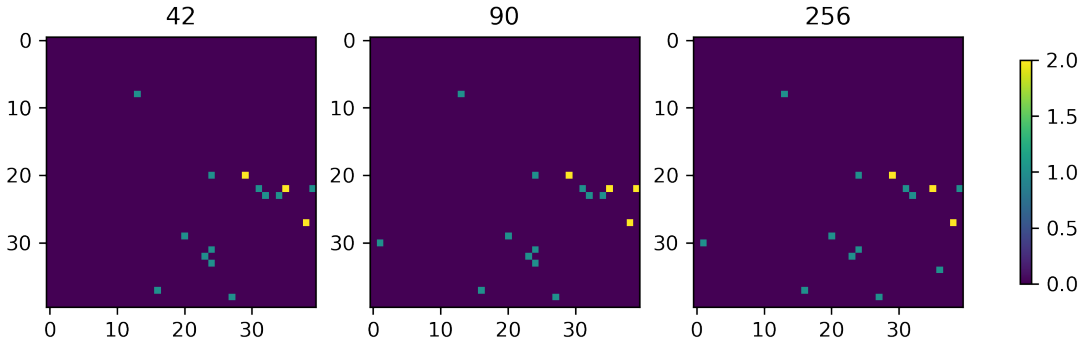


Figure 4.16: (Unnormalised) Confusion rates between all 40 target-classes to one-another. Each row represents the confusion rates between the corresponding target-class and all other target-classes indicated by the columns. The first 20 indices indicate the least difficult classes for the (baseline) model, with the remaining 20 indices indicating the most difficult classes for the (baseline) model. As expected, indices corresponding to the most difficult classes exhibit a greater proportion of confusions to the remaining classes. However, relative to the total number of predictions made by the model on the target-classes (47k), the maximum number of confusions seen between the target classes was 2. This means the target-classes chosen are hardly ever confused for each other. This satisfies the first requirement of our hypothesis on confusion rates.

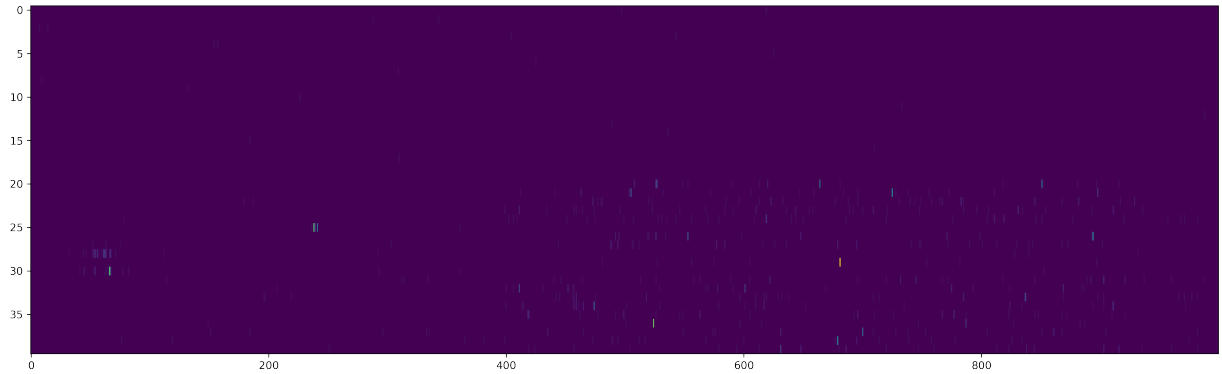


Figure 4.17: Confusion vectors for each of the 40 target-classes. Each row indicates a confusion vector for one of the 40 target-classes, and its (unnormalised) confusion rates with the remaining 999 ImageNet classes. The first 20 indices indicate the least difficult classes for the (baseline) model, with the remaining 20 indices indicating the most difficult classes for the (baseline) model. As expected, indices corresponding to the most difficult classes exhibit a greater proportion of confusions to the remaining classes. Under our hypothesis, target-classes with confusion vectors that are most similar will learn attention seeds weights that are most similar, and vice-versa for dissimilar confusion vectors. These confusion vectors are therefore converted into an RDM to capture dissimilarities, and compared to the RDM for attention seeds weights.

Figure 4.16 shows the confusion matrix between the 40 target-classes. Across all dimen-

sionalities, the total frequency of confusions between the target-classes was no more than 2.0. This indicates a low potential bias of later results, as the target-class set contains classes that are hardly confused with one another in comparison to the true-positives observed. Our target-class set therefore meets the first requirement of the hypothesis. Figure 4.17 shows the confusion vectors for each of the 40 target classes. The top-half of confusion matrix appears uniform as it corresponds to the 20 easiest classes in the target-class set, hence they are hardly confused for other classes. The bottom-half sees more confusions amongst the target-class and the remaining ImageNet classes, again expected as it corresponds to the 20 hardest target-classes.

With all the data collected we begin the RSA. An RDM was constructed for the second confusion matrix under the Crossnobis distance. This RDM will capture similarity and dissimilarity patterns amongst the confusion vectors for each of the 40 target-classes. Under our hypothesis, with all classes within the target-class set hardly being confused for each other, if a dissimilarity is found to be 0, then this corresponds to target-classes with confusion vectors that obey the hypothesis, and vice-versa else wise. If we then construct RDMs for the learned attention seeds for the target-class set, under the assumption that the hypothesis is true, we should examine very high Spearman correlations between the seed RDMs and the confusion vectors RDM. This comparison between the RDMs factors in the similarities and dissimilarities between confusion vectors and hence provides evidence for and against the hypothesis, respectively. The beauty of RSA is that we do not require confusion vectors that directly obey the hypothesis (dissimilarities of 0 in the RDM), instead, even if they obeyed the hypothesis slightly, this would result in dissimilarities slightly above 0. If we then compute the correlation between the patterns found between the confusion vectors and the patterns found between the attention seeds, a high correlation would indicate strong evidence for the hypothesis, again from both angles.

Visualising Attention Seeds and Confusion Vectors RDMs

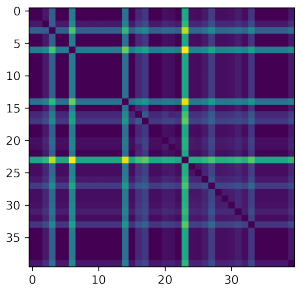


Figure 4.18: RDM capturing the dissimilarities between the 40 confusion vectors for all target-classes. Dissimilarities were estimated with the Mahalanobis distance. Characteristic plus-sign shaped patterns of higher dissimilarities were found to emanate from the diagonal. The frequency of these plus-signed patterns increases for the most difficult classes in comparison to the least difficult classes. Under our hypothesis, the classes that exhibit the lowest dissimilarities between confusion vectors should produce the the lowest dissimilarities between attention seeds weights, and vice-versa for the classes with the highest dissimilarities (plus-signed patterns). Therefore if the hypothesis is true, and confusion rates are a strongly influencing factor on the learning of attention weights, then these same patterns should be found in the attention seeds RDM.

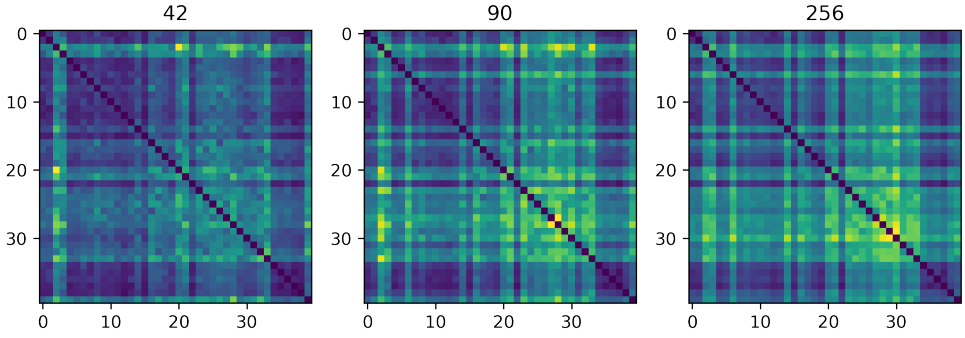


Figure 4.19: Target-class RDMs for attention seed weights across increasing attention seed dimensionalities. Dissimilarities were estimates using the cross-validated Mahalanobis distance. Astoundingly, the same plus-signed patterns of dissimilarities seen in the ground-truth confusion rates RDM is entirely replicated within the dissimilarities between attention seeds weights! It is replicated so clearly that the RDM for the attention seeds appears to be a more noisy version of the RDM for the confusion vectors. Greater dissimilarities are exhibited between the most difficult classes in comparison to the least difficult classes. This suggests that class-difficulty and confusion rates between classes are strongly influencing factors guiding the learning of attention weights. The dissimilarities become more apparent with increasing seeds dimensionalities.

Figures 4.18 shows the dissimilarities between the confusion vectors for the target-class set. Overall, dissimilarities remained low. Strips of highly dissimilar confusion vectors emanating from the diagonal are seen consistently throughout both easier and harder classes. As expected, harder classes tended to have higher dissimilarities between confusion vectors compared to easier classes due to greater differences amongst confusion vectors. Figure 4.19 shows the RDMs constructed between learned attention seeds for all target-classes at the set dimensionalities. Immediately it can be seen that more difficult classes (indicated by the bottom right quadrant) tended to lead to learned attention weights that were more dissimilar. Attention weights were also found to be more dissimilar when comparing attention seeds between less difficult classes and more difficult classes (indicated by the top-right and bottom-left quadrants). This effect is more observable as seed dimensionality increases, and provides evidence to the effect class difficulty plays on the representational geometries between learned seed weights. In fact, it appears to play a large role in influencing differences between the learned seed weights. What is most amazing is that across all RDMs, the same plus-sign shaped patterns emanating from the diagonal as seen between the confusion vectors is entirely replicated within the learned seed weights! The replication is clear to the point that the representational similarities between learned seed weights appear to be more noisy versions of the representational similarities between the confusion vectors. Examining on the visualisation of the RDMs, one may think the attention seeds RDMs exhibits the same plus-signed patterns but as dark strips (low dissimilarity) in comparison to the bright strips (high dissimilarity) in the confusion vectors RDM, however, it was verified the indices were not the same. For example the brightest plus-sign shaped pattern occurs at target-class number 24 in the confusion vectors RDM whereas the dark plus-sign shaped pattern occurs at target-class number 23 in the attention seeds RDM. This provides strong evidence to suggest the representational geometries of learned attention weights are highly dependent on how the confusing two particular target-classes are to one-another relative to all non-target-classes, and are less so driven by class-level factors such as semantic entanglement between classes. We now move to compare these quantitatively in order to verify inferences drawn from the visualisations of RDMs.

Comparing Attention Seeds RDMs to the Confusion Vectors RDM

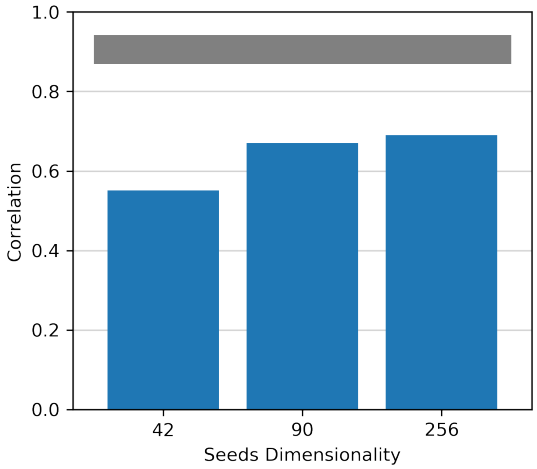


Figure 4.20: Spearman correlation comparisons between the learned seed weights RDMs and the ground-truth confusion vector RDM, at each trained seeds dimensionalities. A noise ceiling (the grey horizontal bar) is calculated to see how well a perfect model would perform in the presence of noise found in the data. All RDMs appear strongly correlated with increasing Spearman correlations with increasing dimensionality. The noise ceiling is close to 1 (perfect correlation) and is small in height suggesting a more reliable inference can be concluded.

Figure 4.20 shows Spearman correlations between the seed RDMs to the confusion vectors RDM at each dimensionality. As expected, Spearman correlations show that the learned seed weights under the 90 and 256 dimensional mechanisms are strongly correlated to the confusion vector RDM, providing strong quantitative evidence to support the hypothesis. A small noise ceiling shows the analysis is more reliable with less noise present amongst stimuli due to a larger number of samples present. The noise ceiling upper-bounds at 0.94, very close to 1.0, meaning the data was found to be of high-quality and largely consistent across trials and that conclusions can drawn are reliable.

At a high-level this relationship makes sense. In order to achieve perceptual boosts attention weights must be learned in such a way that classes which are more likely to be confused with the target-class are less confounding to the model. The target-class examples the model predicts incorrectly have activations modulated in such a way that deviates likelihoods away from the incorrectly predicted class and more towards the target-class. Within similar classes, this deviation must be the greatest as these classes will tend to be more confused with each other. The underlying factor that drives attention weights must then maximise the distances between confounding classes, and hence must be **driven most strongly by confusion-rates amongst classes**. This relationship

becomes more apparent as seed dimensionality increases, most likely due to the close link between class-difficulty and model performance. The 256 dimensional seeds mechanism performed the best during the performance evaluation, and evidently appears to exhibit the greatest correlation to the confusion vectors. This result is astounding - we have successfully localised a factor that appears to very strongly determine why attention weights are learned how they are! This is one giant step in understanding the true governing factors that lead attention weights to be how they are, really aiding in the interpretability of attention weights!

Perhaps in order to achieve a near perfect correlation, an experiment that directly optimises for choosing target-classes that follow the hypothesis classes that do not follow the hypothesis is required. We noted that some examples amongst the hard classes were still confused for one another. Even if this confusion frequency was extremely low, perhaps classes that do not confound each other at the slightest will lead to a noise ceiling more close to 1.0. In this experiment only 40 target-classes were used to verify the effect class-difficulty and confusion-rates have on representational geometries of learned attention weights. Further experiments can improve on this experiment by taking a more direct approach and plotting a confusion matrix between all ImageNet-1k classes, then selecting a greater number of target-classes directly in support and not in support of the hypothesis. This in-turn should provide a more full-proof evaluation of the hypothesis. With this all experiments and evaluations are complete.

Chapter 5

General Discussion & Suggestions for Further Work

5.0.1 General Discussion

This thesis was motivated by research in neuroscience and psychology pertaining to the role of attentional mechanisms found in category learning models in re-configuring DCNNs externally to specialise to a goal-directed task. More specifically, this thesis focuses on the evidence shown in the literature supporting the idea of a low-dimensional subspace projection of input stimuli from early-to-late cortical areas of the ventral stream and adapts state-of-the-art top-down attentional mechanisms to support such lower-dimensional projections. Naturally, many hypotheses can be drawn by analysing top-down attentional mechanisms that operate in low-dimensions, especially when considering the potential for significant increase in explainability of attention weights through removal of redundancies via dimensionality-reduction techniques. We proposed a new top-down attention mechanism that facilitates a goal-directed influence on the DCNN, yet operates at dimensionalities much lower than the dimensionality of the layer it is currently modulating; a limitation found in current state-of-the-art top-down attentional mechanisms. At the beginning of this thesis we asked four questions that we hoped to answer throughout the project. In turn, we revisit each question and briefly describe and explain the main outcomes from experiments that answer the question asked.

1. *Under the assumption that redundancies within the filter-space are translated to redundancies within the attention parameters, can we develop a technique to perform a parameter-reduction on goal-directed attention, while maintaining all perceptual performance gains yielded from the attention mechanism? - Yes!*

The developed attention seeds mechanism successfully performed a parameter-reduction while maintaining all perceptual performance gains and excelling in other areas of our performance evaluation criterion in comparison to a state-of-the-art goal-directed attention mechanism. We showed that at a 50% reduction in trainable attention weights, model performance in general, and more specifically model sensitivity, was maintained. Statistical hypotheses testing in the form of a Student's t-test verified the maintenance in model performance when utilising the lower-dimensional attention weights. The seeds mechanism also exhibited a lower model criterion meaning the mechanism was less biased towards making a false-alarm over the standard attention mechanism. This shows that redundancies found within the full-dimensional filters are indeed replicated within the attention mechanism that is modulating them, and that our proposed mechanism can eliminate them. The effect of seeds permitted to an attention layer (dimensionality) was found to significantly affect model performance and attention weight distributions. Lowering the seed count below the 50% reduction led to lower model performance, greater variance in attention weightings, and overall greater noise when conducting the seeds training process.

Our results are pivotal as similar attention mechanisms can now offer flexibility in choosing the number of trainable attention parameters, without being restricted to the dimensionality of the filter-space! Furthermore, with increasingly wider neural networks this parameter-reduction can be extremely beneficial in performing increased model compression, improving model efficiency and lowering training times while maintaining performance gains of the attention mechanism. SNMF-based dimensionality-reduction techniques are scarce in the literature, and have never been used in conjunction with attentional mechanisms in DCNNs. The success of utilising an SNMF-based dimensionality-reduction technique is apparent and lends interesting concepts that attention weights should be an additive combination of fundamental basis components found within the convolutional parameters owned by the model. Our results are intuitive; using SNMF on the convolutional kernel extracts fundamental basis components found to recur within the filter-space thereby removing redundancies. Using this factorisation in place of the step-up matrix naturally leads to attention seeds that do not exhibit the same redundancies found in the original filter-space.

2. *Is there a set of low-dimensional latent factors that drive the high-dimensional attention weights? - Yes!*

We find that attention seeds did indeed act as low-dimensional latent factors driving the high-dimensional attention weights. Evidence to suggest so was consistently produced throughout most experiments carried out within the thesis. Under successful maintenance of attention performance

gains while greatly reducing the number of attention parameters, this question was already answered. If model performance is maintained with fewer attention parameters then these fewer attention parameters must be the low-dimensional factors. This is because the dimensionality-reduction reduces attention weights and removes redundancies in over-parameterisations, meaning that if performance degradation is not observed, what's left must constitute fundamental factors that retain only the necessary attentional content required by the attention mechanism to achieve perceptual boosts. We also showed, via representational similarity analysis, highly strong correlations between learned lower-dimensional seeds weights and the full-dimensional standard attention mechanism weights across 21 distinct task-sets. Results showed that even at a 97% reduction in attention parameters, large amounts of informational content present within full-dimensional attention weights were retained within the low-dimensional weights, with Spearman correlations of 0.66 observed between the two sets of weights at such a great reduction!

3. *With less attention parameters to work with, are low-dimensional attention weights more greatly influenced by factors that govern similarities between learned attention weights? Does this therefore pose an increased likelihood in extracting such factors from representational analysis between learned low-dimensional attention weights? - Yes!*

4. *Can we identify factors that pose the greatest influence over the learning of attention parameters? - Yes, but requires an ablation study.*

The final two questions asked are compounded together as our findings answer yes to both from multiple experiments performed within the second stage of the thesis. Answering these questions required a more sophisticated analysis technique in the form of a representational similarity analysis (RSA) as used frequently in neuroscience literature. RSA enables comparison of representational geometries between attention weightings and the classes themselves by projecting both into shared representational space.

With less attention parameters to work with the low-dimensional attention weights indeed were more greatly influenced by factors governing the similarities between attention weights. From analyses on the distributional changes on attention seed weights that occurs with decreasing dimensionality, we found lower-dimensional seed weights were more greatly diverse, with standard deviations 34 times greater than their full-dimensional counterparts. This is because attention weights are learned in a way that modulates features such that confounding classes are found to be less confusing to the model. Under a constant number of confounding classes to a target-class, decreasing degrees-of-freedom means decreasing the number of such available modulations, hence the factors that drive the learning of attention weights must act more greatly to achieve a similar

level of de-obfuscation. Since the low-dimensional seeds are the only trainable parameters, this influence is more greatly exerted on the seeds, hence leading to the greater diversification of seed weights at lower dimensionalities, partly answering the first question.

To answer the second-half of the first question, and the second question more concretely, we use the RSA on learned seed weights. The first batch of RSA experiments performed this on classes that were semantically entangled i.e belonging to the same superordinate grouping allows an existing fundamental connection between task-sets that we can use to spot an underlying connection more easily. Classes were also chosen with a variety in class-and-model-factors such as variance in filters activated and class difficulty.

We found that overall, the lower-dimensional attention seeds mechanism consistently learned attention weights with representations more closely correlated to the underlying semantic embeddings of target-classes, with the best correlations exhibited by the 42, 90, and 256 dimensional mechanisms. To answer the second question, we found confusion rates to be a strongly influencing factor governing the similarities between learned attention weights. Confusion rates have a strong relationship to class difficulty, which is a factor influenced by the explained variance (correlations) of inter-layer activations between task-sets, and accordingly the most difficult classes tended to have more dissimilar attention weights when compared to the least difficult classes. Hence we concluded class-difficulty, and more specifically confusion rate similarities between different classes, to be a strongly influencing factor on the learning process of attention weights. Explained variance was, at most, moderately influencing but due to its close link to confusion rates it is indeed an influencing factor nonetheless. Other considered factors such as semantic similarities between classes were found to weakly influence the learning process of attention weights, with weak-to-moderate correlations at best seen in comparisons between representational geometries. We must note though that the RSA experiments related to the semantic relationships between classes were conducted on too few a number of task-sets, only 21 to be exact, and as such much noise was present during our analysis - an observed limitation to our findings. Through cross-validation over multiple trials of learned seed weights we were able to reduce noise in our experimentation, however the use of a greater number of task-sets would certainly be beneficial.

Across all RSA experiments the lower-dimensional attention seeds consistently bested the full-dimensional standard attention mechanism in producing correlations between the representational dissimilarities between the attention weights and the semantic embeddings, and again to representational dissimilarities between isolated factors such as class-difficulty and confusion rates. Therefore, they indeed pose an increased likelihood in extracting factors governing the learning of attention weights from representational analysis.

This finding is pivotal as it appears in performing a dimensionality-reduction of attention weights, by eliminating redundancies and over-parameterisations, we forced the model to more greatly exert influences that guide the learning process of attention weights, thereby allowing us to more easily identify and extract these influences from patterns within and between learned attention weights. This is extremely important as this opens up a realm of possible research in truly pinpointing the factors that most greatly influence attention and its role within the neural network when attenuating activations to perform the goal of the network. Understanding attention weights within attention mechanisms is traditionally a very hard task. Being able to understand what factors influence attention weights to be how they are brings us closer to understanding the weights themselves, and what they mean. To researchers, understanding how the attention mechanism interplays with every aspect of the network, down to its training distribution, will allow us to build more performant attention mechanisms dedicated to performing more specialised roles within the network, and more specifically can enable configurable attention mechanisms for different applicational tasks and domains of neural networks.

5.0.2 Suggestions For Further Work

With the bulk of the thesis work dedicated to developing a working dimensionality-reduction technique for goal-directed top-down attention modulation, this leaves many promising avenues to explore research in furthering interpretability of attention weights.

1. *Trial the use of other matrix decomposition techniques in extracting a step-up matrix*

Future work could evaluate the performance and explainability in utilising different matrix decomposition techniques, such as PCA or Convex-NMF, on the convolutional kernel to extract a step-up matrix. The SNMF technique satisfies a human-level explainability but did not lead to high explainability of learned seed weightings when analysed with RSA across all tested factors. SNMF was used as it was paticularly beneficial for our use-case, utilising this attention mechanism under different application areas may require different constraint on the step-up matrix, and hence different decomposition techniques.

2. *Run RSA on other fundamental relationships between target classes and model-level variables.*

In this thesis we only considered relating learned seed weightings to the semantic relationships amongst classes. We also only evaluated a subset of model-level variables such as class difficulty, confusion rates and variance in filter activations solicited by different superordinate groupings.

Further research can look to evaluate many other class-level and model-level relationships, under recommendation of an ablation study experimentation design, and how they influence the representational similarities of the learned seed weightings.

3. *Evaluate how altering the training set distribution between target-classes and non-target-classes affects learned seed weightings.*

A final recommended direction of research is evaluating the effect changing the distribution over the non-target-class set has on learned seed weightings. In this thesis the training set comprised all samples from the target-class, and an equal number of samples chosen uniformly at random from the remaining 999 ImageNet classes. One could manually curate different non-target sets with a non-uniform distribution over classes and see how that affects similarity between attention weights i.e using purely the most representational images from the remaining ImageNet classes during training. Further to this, one could investigate the extremities over this distribution. Instead of training under a 1000-way classification, one could train binary models as was seen in [Lindsay and Miller, 2018]. All present interesting research potential for understanding the factors at play when training attention seeds.

The success of the new seeds mechanism provides a great foundation for introducing dimensionality-reduction into the realm of goal-directed top-down attention mechanisms and allows for more flexibility in configuring the attention mechanism for the application use-case of the model. By demonstrating the success that pairing representational similarity analysis with dimensionality-reduction techniques had within our experiments, we champion that the same analysis can be used to investigate other governing factors influencing attention parameters. In fact, we put forward that such a powerful pairing can be used across multiple domains of Machine Learning, leading to a fuller understanding of what is considered the unknowns of the training process of deep neural network parameters. With many possible extensions to this method comes many directions for exciting future research surrounding attention, opening the doors to constructing a complete understanding of the role attention plays within DCNNs, and even within the human visual system.

Appendix A

List of ImageNet Target-Classes

A.1 Superordinate-Based Experimentation

Table A.1: ImageNet classes used within superordinate-based experimentation with their accompanying WordNet ID's.

Avian		**Dog**		**Kitchen**	
WordNet ID	**ImageNet Class Name**	**WordNet ID**	**ImageNet Class Name**	**WordNet ID**	**ImageNet Class Name**
n01534433	junco	n02085620	chihuahua	n03887697	paper_towel
n01608432	kite	n02085782	japanese_spaniel	n03207941	dishwasher
n01806143	peacock	n02085936	maltese_dog	n02906734	broom
n01820546	lorikeet	n02087046	toy_terrier	n04131690	saltshaker
n01833805	hummingbird	n02094114	norfolk_terrier	n03775546	mixing_bowl
n01855672	goose	n02095570	lakeland_terrier	n04553703	washbasin
n02012849	crane	n02096294	australian_terrier	n03207743	dishrag

A.2 Class-Difficulty-Based Experimentation

Table A.2: ImageNet classes used within class-difficulty-based experimentation with their accompanying WordNet ID's and baseline class accuracy.

20 Least Difficult			20 Most Difficult		
WordNet ID	**ImageNet Class Name**	**Baseline Accuracy**	**WordNet ID**	**ImageNet Class Name**	**Baseline Accuracy**
n03590841	jack-o'-lantern	0.960	n04152593	screen	0.104
n02130308	cheetah	0.960	n04560804	water_jug	0.160
n01820546	lorikeet	0.960	n04154565	screwdriver	0.191
n02917067	bullet_train	0.960	n04525038	velvet	0.200
n02112018	pomeranian	0.960	n04286575	spotlight	0.200
n03344393	fireboat	0.967	n02107908	appenzeller	0.200
n11879895	rapeseed	0.973	n03016953	chiffonier	0.213
n11939491	daisy	0.978	n03532672	hook	0.213
n02391049	zebra	0.979	n01740131	night_snake	0.220
n02006656	spoonbill	0.979	n03642806	laptop	0.224
n02111129	leonberg	0.979	n01756291	sidewinder	0.245
n02489166	proboscis_monkey	0.979	n03692522	loupe	0.250
n13044778	earthstar	0.980	n03866082	overskirt	0.260
n01518878	ostrich	0.980	n04355933	sunglass	0.260
n03393912	freight_car	0.980	n04370456	sweatshirt	0.260
n02116738	african_hunting_dog	0.980	n03658185	letter_opener	0.261
n02342885	hamster	0.980	n02895154	breastplate	0.265
n01534433	junco	0.980	n15075141	toilet_tissue	0.267
n04613696	yurt	1.000	n02999410	chain	0.271
n12057211	yellow_lady's_slipper	1.000	n03476991	hair_spray	0.277

Bibliography

[Abadi et al., 2015] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

[Ahlheim and Love, 2018] Ahlheim, C. and Love, B. C. (2018). Estimating the functional dimensionality of neural representations. *Neuroimage*, 179:51–62.

[Amari, 1993] Amari, S. (1993). Backpropagation and stochastic gradient descent method. *Neurocomputing*, 5(4):185–196.

[Bahdanau et al., 2014] Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate.

[Berlot et al., 2020] Berlot, E., Popp, N. J., and Diedrichsen, J. (2020). A critical re-evaluation of fMRI signatures of motor sequence learning. *Elife*, 9.

[Blondel et al., 2008] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

[Brown et al., 2020] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In Larochelle,

H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

[Cai et al., 2022] Cai, Y., Hua, W., Chen, H., Suh, G. E., De Sa, C., and Zhang, Z. (2022). Structured pruning is all you need for pruning cnns at initialization.

[Chai et al., 2021] Chai, J., Zeng, H., Li, A., and Ngai, E. W. (2021). Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Machine Learning with Applications*, 6:100134.

[Chan et al., 2021] Chan, T. K., Chin, C. S., and Li, Y. (2021). Semi-supervised nmf-cnn for sound event detection. *IEEE Access*, 9:130529–130542.

[Charvet et al., 2013] Charvet, C. J., Cahalane, D. J., and Finlay, B. L. (2013). Systematic, cross-cortex variation in neuron numbers in rodents and primates. *Cereb Cortex*, 25(1):147–160.

[Craig, 1979] Craig, A. (1979). Nonparametric measures of sensory efficiency for sustained monitoring tasks. *Hum Factors*, 21(1):69–77.

[Deng et al., 2009] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

[Devlin et al., 2018] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.

[Dhillon and Sra, 2005] Dhillon, I. S. and Sra, S. (2005). Generalized nonnegative matrix approximations with bregman divergences. In *Proceedings of the 18th International Conference on Neural Information Processing Systems*, NIPS'05, page 283–290, Cambridge, MA, USA. MIT Press.

[DiCarlo et al., 2012] DiCarlo, J. J., Zoccolan, D., and Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, 73(3):415–434.

[Ding et al., 2005a] Ding, C., He, X., and Simon, H. D. (2005a). Nonnegative lagrangian relaxation of k-means and spectral clustering. In Gama, J., Camacho, R., Brazdil, P. B., Jorge, A. M., and Torgo, L., editors, *Machine Learning: ECML 2005*, pages 530–538, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Ding et al., 2005b] Ding, C., He, X., and Simon, H. D. (2005b). *On the Equivalence of Nonnegative Matrix Factorization and Spectral Clustering*, pages 606–610.

[Ding et al., 2006] Ding, C., Li, T., Peng, W., and Park, H. (2006). Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, page 126–135, New York, NY, USA. Association for Computing Machinery.

[Ding et al., 2010] Ding, C. H., Li, T., and Jordan, M. I. (2010). Convex and semi-nonnegative matrix factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):45–55.

[Dosovitskiy et al., 2020] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale.

[Erler et al., 2018] Erler, J., Ramos-Ceja, M. E., Basu, K., and Bertoldi, F. (2018). Introducing constrained matched filters for improved separation of point sources from galaxy clusters. *ArXiv e-prints*.

[Gao et al., 2019] Gao, Y., Ma, J., Zhao, M., Liu, W., and Yuille, A. L. (2019). Nddr-cnn: Layerwise feature fusing in multi-task cnns by neural discriminative dimensionality reduction. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3200–3209.

[Hagberg et al., 2008] Hagberg, A., Swart, P., and S Chult, D. (2008). Exploring network structure, dynamics, and function using networkx.

[Han et al., 2020] Han, Y., Roig, G., Geiger, G., and Poggio, T. (2020). Scale and translation-invariance for novel objects in human vision. *Scientific Reports*, 10(1):1411.

[Haxby et al., 2014] Haxby, J. V., Connolly, A. C., and Guntupalli, J. S. (2014). Decoding neural representational spaces using multivariate pattern analysis. *Annual Review of Neuroscience*, 37(1):435–456. PMID: 25002277.

[Haxby et al., 2001] Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., and Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430.

[He et al., 2015] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition.

[Hu et al., 2018] Hu, J., Emile-Geay, J., Nusbaumer, J., and Noone, D. (2018). Impact of convective activity on precipitation 18o in isotope-enabled general circulation models. *J. Geophys. Res. Atmos.*, 123(23):13595–13610.

[Hung et al., 2005] Hung, C. P., Kreiman, G., Poggio, T., and DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science*, 310(5749):863–866.

[Jang and Lee, 2020] Jang, B. and Lee, S. (2020). Cnn based sound event detection method using nmf preprocessing in background noise environment. *The International Journal of Advanced Smart Convergence*, 9(2):20–27.

[Kalfas et al., 2018] Kalfas, I., Vinken, K., and Vogels, R. (2018). Representations of regular and irregular shapes by deep convolutional neural networks, monkey inferotemporal neurons and human judgments. *PLOS Computational Biology*, 14(10):1–26.

[Kell et al., 2018] Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., and McDermott, J. H. (2018). A Task-Optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644.e16.

[Kingma and Ba, 2014] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization.

[Kipnis, 2022] Kipnis, A. (2022). Python representational similarity analysis (rsatoolbox) toolbox - rsatoolbox 0.0.4 documentation.

[Kriegeskorte et al., 2008] Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Front Syst Neurosci*, 2:4.

[Kruschke, 1990] Kruschke, J. (1990). Alcove: A connectionist model of human category learning. In Lippmann, R., Moody, J., and Touretzky, D., editors, *Advances in Neural Information Processing Systems*, volume 3. Morgan-Kaufmann.

[Lage-Castellanos et al., 2019] Lage-Castellanos, A., Valente, G., Formisano, E., and De Martino, F. (2019). Methods for computing the maximum performance of computational models of fmri responses. *PLOS Computational Biology*, 15(3):1–25.

[Lehky et al., 2014] Lehky, S. R., Kiani, R., Esteky, H., and Tanaka, K. (2014). Dimensionality of object representations in monkey inferotemporal cortex. *Neural Comput*, 26(10):2135–2162.

[Li et al., 2016] Li, H., Kadav, A., Durdanovic, I., Samet, H., and Graf, H. P. (2016). Pruning filters for efficient convnets.

[Lindsay and Miller, 2018] Lindsay, G. W. and Miller, K. D. (2018). How biological attention mechanisms improve task performance in a large-scale visual system model. *eLife*, 7:e38105.

[Liu and Deng, 2015] Liu, S. and Deng, W. (2015). Very deep convolutional neural network based image classification using small training sample size. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 730–734.

[Love et al., 2004] Love, B. C., Medin, D. L., and Gureckis, T. M. (2004). SUSTAIN: a network model of category learning. *Psychol Rev*, 111(2):309–332.

[Luo et al., 2021] Luo, X., Roads, B. D., and Love, B. C. (2021). The costs and benefits of goal-directed attention in deep convolutional neural networks. *Computational Brain & Behavior*, 4(2):213–230.

[Luong et al., 2015] Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation.

[Macmillan and Creelman, 2009] Macmillan, N. A. and Creelman, C. D. (2009). *Detection theory: A user's guide*. Psychology Press.

[Macmillan and Kaplan, 1985] Macmillan, N. A. and Kaplan, H. L. (1985). Detection theory analysis of group data: estimating sensitivity from average hit and false-alarm rates. *Psychol Bull*, 98(1):185–199.

[Majaj et al., 2015] Majaj, N. J., Hong, H., Solomon, E. A., and DiCarlo, J. J. (2015). Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *Journal of Neuroscience*, 35(39):13402–13418.

[Maćkiewicz and Ratajczak, 1993] Maćkiewicz, A. and Ratajczak, W. (1993). Principal components analysis (pca). *Computers Geosciences*, 19(3):303–342.

[Miller and Cohen, 2001] Miller, E. K. and Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annu Rev Neurosci*, 24:167–202.

[Nili et al., 2014] Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., and Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLOS Computational Biology*, 10(4):1–11.

[Nosofsky, 2011] Nosofsky, R. M. (2011). *The generalized context model: an exemplar model of classification*, page 18–39. Cambridge University Press.

[Paszke et al., 2017] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in pytorch.

[Perez et al., 2017] Perez, E., Strub, F., de Vries, H., Dumoulin, V., and Courville, A. (2017). Film: Visual reasoning with a general conditioning layer.

[Popal et al., 2020] Popal, H., Wang, Y., and Olson, I. R. (2020). A Guide to Representational Similarity Analysis for Social Neuroscience. *Social Cognitive and Affective Neuroscience*, 14(11):1243–1253.

[Potter, 1976] Potter, M. C. (1976). Short-term conceptual memory for pictures.

[Prince, 2012] Prince, S. (2012). *Computer Vision: Models Learning and Inference*. Cambridge University Press.

[Qiu et al., 2021] Qiu, J., Chen, C., Liu, S., Zhang, H.-Y., and Zeng, B. (2021). Slimconv: Reducing channel redundancy in convolutional neural networks by features recombining. *Trans. Img. Proc.*, 30:6434–6445.

[Raghu et al., 2021] Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., and Dosovitskiy, A. (2021). Do vision transformers see like convolutional neural networks?

[Rajalingham et al., 2018] Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., and DiCarlo, J. J. (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. *Journal of Neuroscience*, 38(33):7255–7269.

[Rousselet et al., 2002] Rousselet, G. A., Fabre-Thorpe, M., and Thorpe, S. J. (2002). Parallel processing in high-level categorization of natural images. *Nature Neuroscience*, 5(7):629–630.

[Savage, 2019] Savage, N. (2019). How ai and neuroscience drive each other forwards. *Nature*, 571(7766):S15–S17.

[Schrimpf et al., 2018] Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Schmidt, K., Yamins, D. L. K., and DiCarlo, J. J. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*.

[Sexton and Love, 2022] Sexton, N. J. and Love, B. C. (2022). Directly interfacing brain and deep networks exposes non-hierarchical visual processing. *bioRxiv*.

[Smith et al., 2021] Smith, F. B., Roads, B. D., Luo, X., and Love, B. C. (2021). Understanding top-down attention using task-oriented ablation design.

[Stanislaw and Todorov, 1999] Stanislaw, H. and Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, 31(1):137–149.

[Thorpe et al., 1996] Thorpe, S., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, 381(6582):520–522.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

[Vaziri-Pashkam and Xu, 2017] Vaziri-Pashkam, M. and Xu, Y. (2017). Goal-Directed visual processing differentially impacts human ventral and dorsal visual representations. *J Neurosci*, 37(36):8767–8782.

[Velten de Melo and Wainer, 2012] Velten de Melo, E. and Wainer, J. (2012). Semi-nmf and weighted semi-nmf algorithms comparison.

[Walther et al., 2015] Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., and Diedrichsen, J. (2015). Reliability of dissimilarity measures for multi-voxel pattern analysis. *Neuroimage*, 137:188–200.

[Wen et al., 2018] Wen, H., Shi, J., Zhang, Y., Lu, K.-H., Cao, J., and Liu, Z. (2018). Neural encoding and decoding with deep learning for dynamic natural vision. *Cereb Cortex*, 28(12):4136–4160.

[Wilson and Wilkinson, 2015] Wilson, H. R. and Wilkinson, F. (2015). From orientations to objects: Configural processing in the ventral stream. *Journal of Vision*, 15(7):4–4.

[Wu and Wang, 2014] Wu, S. and Wang, J. (2014). Nonnegative matrix factorization: When data is not nonnegative. In *2014 7th International Conference on Biomedical Engineering and Informatics*, pages 227–231.

[Xie, 2020] Xie, G. (2020). Redundancy-aware pruning of convolutional neural networks. *Neural Computation*, 32(12):2532–2556.

[Yamins and DiCarlo, 2016] Yamins, D. L. K. and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365.

[Yamins et al., 2014] Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624.

[Yang and Seoighe, 2016] Yang, H. and Seoighe, C. (2016). Impact of the choice of normalization method on molecular cancer class discovery using nonnegative matrix factorization. *PLoS One*, 11(10):e0164880.