Representation in Neural Networks

Adam Townsend

Thesis submitted for the Master of Philosophy degree (MPhil)

School of Humanities / Department of Philosophy

June 2023

Contents

Abstract

Artificial neural networks (ANNs) are computational systems that were inspired by biological neural networks in the brain. ANNs are trained to transform input into task appropriate output using learning algorithms rather than having all relevant aspects of the task explicitly encoded with symbolic rules. Despite the increasingly impressive performance and wide spread usage of ANNs in artificial intelligence (Krizhevsky et al.,2012., LeCun, et. al., 2015., Senjnowski, 2018., Floridi & Chiriatti, 2020), their operation remains somewhat mysterious. There is no widely accepted and comprehensive explanation of how these systems represent and process information (Bornstein, 2016., Habbis et. al., 2017, Schwartz-Ziv & Tishby, 2017). Approaches to explaining the operation of relatively simple neural network models have been discussed by philosophers since the inception of connectionist cognitive science. However, these discussions often relied on analysing the behaviour of a very small number of actual ANNs and there are important issues that still haven't been resolved. I address this by using empirical data from my own unique analysis of a broad range of novel ANNs to evaluate some key philosophical approaches to understanding and comparing neural network models. I focus on structural-resemblance approaches and there has been a shift towards using structural approaches in cognitive neuroscience (Williams & Colling 2018., Boone & Piccinini, 2015., Kriegeskorte et al., 2008., Raizada and Conolly, 2012). Structural-resemblance explanations of representation rely on the idea that structural relations between representations might systematically correspond to the structural organisation of relevant aspects of the represented domain. My empirical work begins by extending Laakso and Cottrell's (2000) method for assessing representation similarity in ANNs to explicitly compare the relevant structural relations between representations across distinct ANNs with diversely configured parameters. I apply this method to evaluate structural approaches to understanding representation in neural networks described by Churchland (1998,1989,1996,2007,2012) and O'Brien and Opie (2004,2006), along with other approaches including clustering (Shea, 2007) and mutual information (Azhar, 2017). My analysis of relatively simple facial recognition ANNs reveals that the structural relations between represented facial categories can vary between different ANNs and may reflect artificial relations rather than intuitive concepts of facial similarity. However, my analysis of a broad range of ANNs categorising various aspects of colour reveals that they develop robust and consistent task-dependent structural representations that do match the relational structure of corresponding human colour judgements. The task relevant structural arrangement of representations that are developed by these networks provides empirical support for the use of structural-resemblance approaches to explaining how ANNs represent and process information.

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.

I give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Adam Townsend 12/6/2023

Acknowledgements

Thank you to my supervisors Dr Jon Opie and Professor Gerard O'Brien for their input and support. I am particularly grateful to my primary supervisor Jon and really enjoyed our numerous and varied discussions over coffee in The Hub. Thanks also to my fellow postgraduate students for their support and engagement in a wide range of interesting philosophical discussions.

I would like to thank my parents for their encouragement and support in undertaking this project. I especially acknowledge my father who provided academic inspiration but sadly did not get to see the completed thesis.

Chapter 1 - Introduction

1.1 How do artificial neural networks operate?

Artificial neural networks (ANNs) are used to model aspects of cognition in connectionist cognitive science and are also widely used in a variety of artificial intelligence applications. Systems using deep convolutional ANNs can now beat human experts in complex games such as Go and Chess (Senjnowski, 2018). ANNs are also used in speech and image recognition, object detection and medical diagnosis with superior performance to traditional algorithmic computational approaches used in classical artificial intelligence (LeCun, et. al., 2015). ANNs are trained to transform input into task appropriate output using learning algorithms rather than having all relevant aspects of the task explicitly encoded in rules governing symbolic state transitions.

Despite the impressive performance of ANNs their operation remains somewhat mysterious and they have been described as inscrutable 'black boxes' (Schwartz-Ziv & Tishby, 2017). Determining what ANNs have actually learned and explaining how they transform their inputs into task appropriate outputs remains elusive. There is still no widely accepted and comprehensive explanation of how these systems represent and process information. The nature of the computational processes by which input is transformed to appropriate output and the types of internal representations that drive the behaviour of ANNs are not properly understood and cannot currently be adequately explained (Habbis et. al., 2017, p253).

In order for ANNs to provide comprehensive insight into aspects of cognition the process that facilitates the transformation of inputs into appropriate outputs needs to be clearly understood and explainable. ANNs used in artificial intelligence applications may be constrained by performance considerations rather than biological plausibility but if the computational processes of these systems is not properly understood it will not be possible to reliably predict their behaviour and the circumstances that lead them to fail. Therefore, a better understanding of ANNs is required in both cognitive science and machine learning.

There are various examples of machine learning systems providing erroneous and unexpected outputs on novel data. This can occur when the statistical regularities learned during training that facilitate accurate performance on the training data do not provide an appropriate conception of the task. An ANN may learn correlations present in the training data that are artefacts of limited or biased sampling rather than constitutive of the required task (Bornstein, 2016). If there is no suitable method to interpret the operation of ANNs then potentially erroneous performance may not be predicted before a critical failure. The performance of ANNs can also be compromised by the deliberate use of adversarial input data that has been specifically generated to be misclassified by the target network (Buckner, 2018).

A variety of methods have been developed to help understand the operation of ANNs. Zeiler and Fergus (2014) describe a method for generating visualisations of inputs associated with maximal activation of processing units in deep convolution neural networks. This provides some insight into the function of intermediate network layers and their representational states. Information theory has also been applied to help explain how ANNs learn and represent information. During the training process the mutual information between representations in intermediate layers and their corresponding input variables decreases

while the mutual information with the associated output increases (Tishby & Zaflavsky, 2015) (Shwartz-Ziv & Tishby, 2017). This shows that the ANNs learn to compress the input data with respect to the output. The transformation of input to output in ANNs can also be described mathematically using the Jacobian to show how the values of each specific output unit varies with respect to changes in the values of individual input units.

All of these methods provide some insight into important aspects of the transformation of input states across ANN layers. However, there is still no universally accepted description of the types of internal representational states these systems utilise and how those states govern the information processing capabilities. Explaining how neural network models operate has been a fundamental concern in the philosophy of connectionist cognitive science. Given the recent advances in artificial intelligence that rely on the application of increasingly complex ANNs, it is timely to review some of the foundational discussions of how relatively simple neural network models represent and process information.

Paul Churchland (1989,1996,1998,2007,2012) has been a prominent contributor to philosophical discussions of connectionist neural network modelling. He argues that the relations between representational states in a successfully trained ANN systematically reflect relevant relations between aspects of the task being performed. The structural relations between representations are similar to the structure of the target domain. The processing is sensitive to these structural relations and so the resulting behaviour of the system aligns with the task. Churchland's views are supported both theoretically and with a relatively limited range of ANNs. Further empirical analysis of this position is still required.

It is also prudent to review this type of approach to understanding ANNs as there has been a recent shift towards using similar methodology in cognitive neuroscience (Williams & Colling 2018). The recent application of mechanistic rather than functional approaches in cognitive neuroscience provides support for considering cognitive representations as having a structural similarity to their target rather than being syntactically composed symbol strings with an arbitrary physical relation to their representational content like a conventional digital computer (Williams & Colling 2018., Boone & Piccinini, 2016). Before describing how my novel empirical investigation will be used to help evaluate whether ANNs employ representations with inherent task relevant structural similarity relations a brief overview of ANNs will be provided.

1.2 Artificial neural networks

Artificial neural networks (ANNs) are abstract or idealized models of biological neural networks that are intended to model the computational properties of the brain understood as an information processing system. These computational systems are also referred to as connectionist networks and this type of computational approach was originally referred to as parallel distributed processing (Rumelhart & McClelland, 1986). ANNs operate in a fundamentally different way to conventional digital computers that provide an automated implementation of a formal rule based system that is generally implemented using binary states (high or low voltages) manipulated by logic gates. ANNs are usually simulated on digital computers, rather than physically implemented, although hardware based implementations are being developed.

ANNs consist of relatively simple processing units that are joined by weighted connections. This is inspired by the biological understanding that brains are composed of many neurons which are interconnected by synapses. ANNs can be simulated using a variety of programming languages and environments with varying objectives and biological realism (Randal et al., 2000). ANNs are used in connectionist cognitive science to model aspects of human cognition in an attempt to better understand actual brain functioning. However, ANNs are now also used extensively in common artificial intelligence applications, for example, responding to language-based queries and in games. This is because they tend to perform well on recognition, categorisation and constraint satisfaction problems which are difficult for traditional algorithmic computation to solve. The development of ANNs in machine learning applications generally focuses on achieving performance outcomes and computational optimisation rather than creating a model that can help explain aspects of human cognition.

The interconnected processing units in ANNs are typically arranged in layers. Each unit in the input layer is assigned an activation value correspond to a specific encoding of the particular input. The activation of units in subsequent layers is determined by summing the weighted activations from all units they are connected to (plus a bias value) and applying an activation (transfer) function such as sigmoid or ReLU (rectified linear unit) to facilitate non-linear processing. The activation of processing units can be described as modelling the average spiking frequency of a neuron or population of neurons and the connection weights as modelling the strength of synaptic connections.

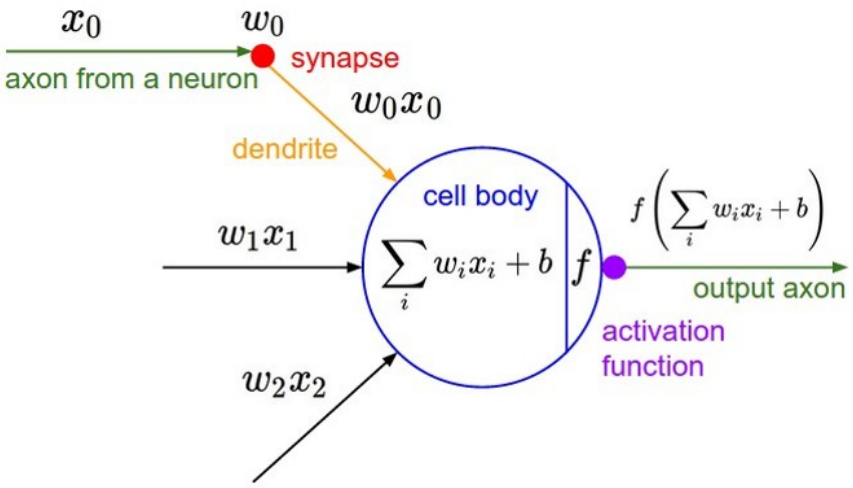


Figure 1.1. The operation of an individual processing unit or model neuron. (*URL=http://cs231n.github.io/convolutional-networks/*)

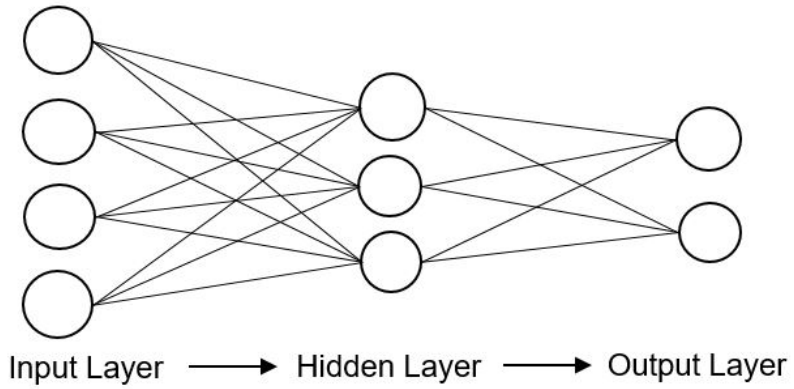


Figure 1.2. A fully connected three-layer feedforward ANN with four input units, three hidden units and two output units.

In order for ANNs to transform inputs into appropriate outputs they must have a configuration of connection weights (and biases) that is appropriate for their task or function. The weights are commonly determined by a supervised training method known as backpropagation (using gradient descent). This involves a training phase where the ANN is repeatedly presented with a set of input samples matched to the correct or desired output. The network's actual output is compared with the desired output and the connection weights are gradually adjusted to decrease the overall error based on their relative contribution to producing it. There are also unsupervised learning methods that do not require the output classes or targets to be known or available in advance. These methods aim to extract patterns or statistical regularities presents in the input data. Some unsupervised learning methods include auto-association (setting weights to recreate input patterns), self-organisation (SOMs) and Hebbian learning (strengthening connections between units that are active together).

Recent increases in computational power, including the use of GPUs to perform ANN training has facilitated the creation of much larger and more complex network architectures. For example, GPT-3 uses a transformer network with 175 billion parameters to generate language-based responses including writing essays or stories (Floridi & Chiriatti, 2020). A significant breakthrough in ANN performance was achieved in 2012 when a deep convolution neural network was used to win the ImageNet image classification competition (Krizhevsky et al.,2012). Convolutional neural networks are inspired by the visual processing system in the brain. They are often used for image classification and can be trained to recognise objects that may vary in location, size and orientation. Convolution layers act like small filters that move across the entire image space. They have localised connectivity with repeated patterns of weights rather than being connected to all units in the previous layer. There are usually many filters in a layer (there could be 100s) and there are also multiple convolution layers where increasingly abstract and complex aspects of the input images are extracted. Pooling layers are also used to reduce the overall dimensionality of the processed input.

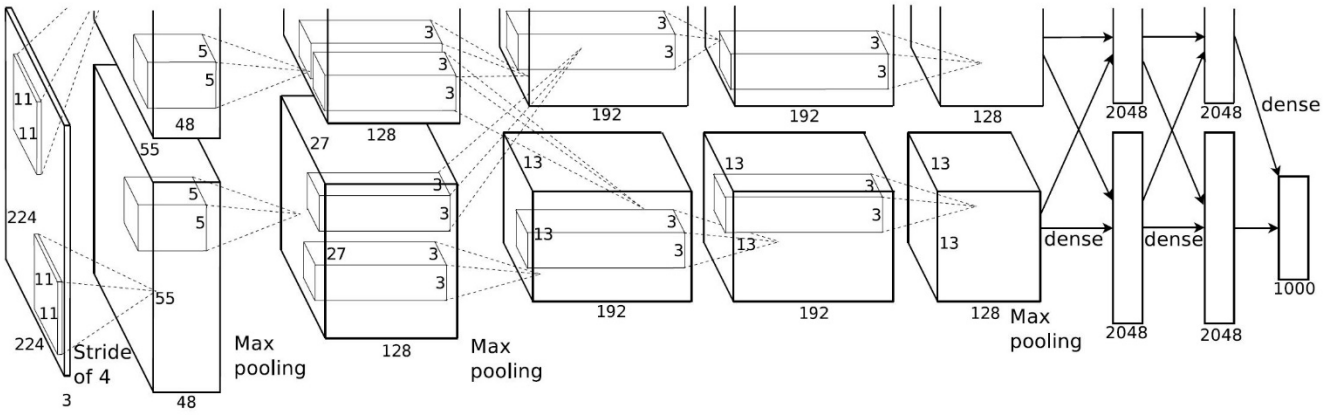


Figure 1.3. AlexNet processes a 150,528 dimensional input through five convolutional layers, three pooling layers and two fully connected layers to produce a 1000 dimensional output (Krizhevsky et al., 2012, p87).

There are some fundamental differences between the way ANNs and conventional digital computers represent and process information. Digital computers represent information using symbols that have an arbitrary or conventional relational to what they represent. The rules that govern the operation of the computer system are designed to respect the intended meaning of the symbols and facilitate processing in a manner that is consistent with the intended application or function.

This is quite different to ANNs where the information that is active during processing is characterised by the pattern of activation distributed across a layer or group of processing units. These transient activation patterns explicitly encode information that is currently active in the network and can be described as representing aspects of the task domain or objects in it. The persistent or long-term knowledge that an ANN employs to transform inputs into task-appropriate outputs is stored in the configuration of connection weights. This is distributed across the network and can be described as representing the implicit knowledge or information embedded in the network. The pattern of connections may represent abstract properties of the task domain that are relevant to correctly classifying or processing the input set. The activation pattern of a group of units in a network layer can be characterised as an activation vector or a point in activation space (an abstract mathematical space). Representational content is associated with the location of hidden layer activation points.

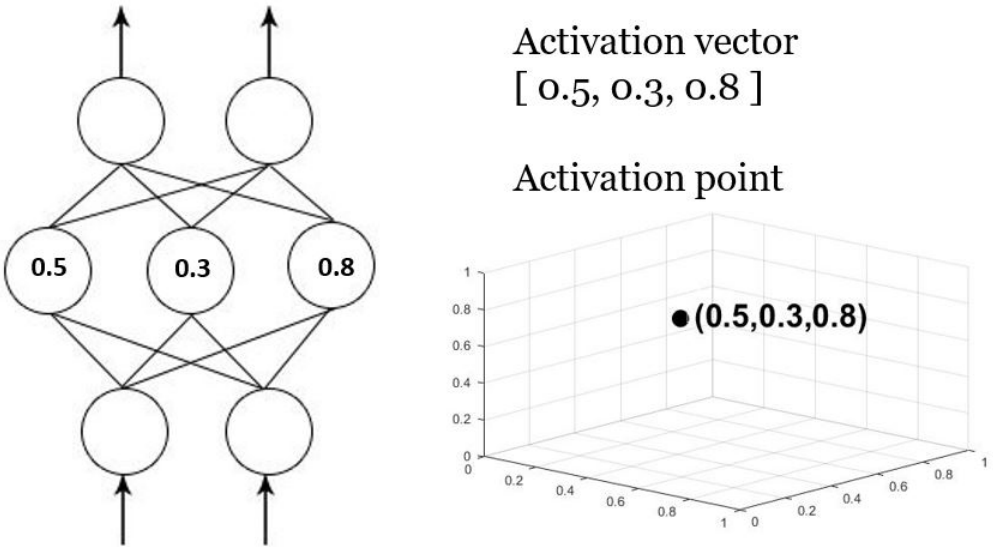


Figure 1.4. The activation values of each of three hidden units in a simple example ANN can be characterised as an activation vector or point in activation space.

In a trained ANN the hidden layer activation points are partitioned or grouped in a task relevant manner. For example, Figure 1.5 shows the collectively considered hidden layer activation values generated by a colour recognition network. The ANN was trained to categorise five colours from reflectance spectra input and has one hidden layer consisting of three hidden units. It is apparent that the activation points generated from each individual input colour are in a similar region and have relatively similar activation values.

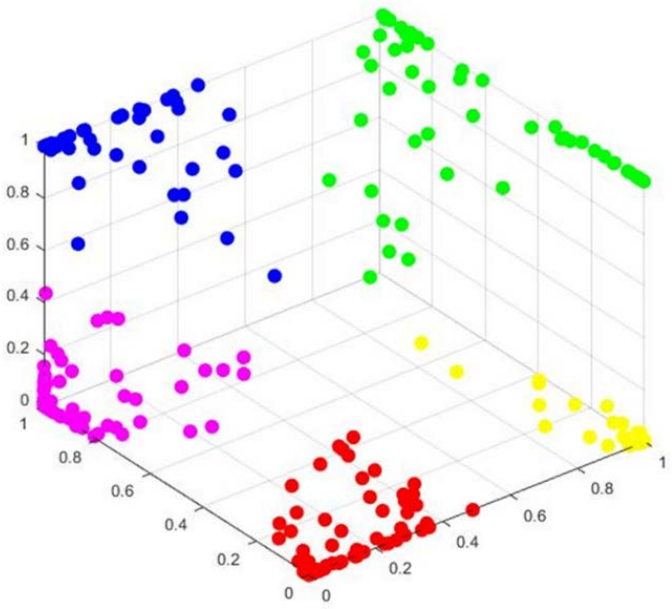


Figure 1.5. The distribution of hidden layer action points in a colour recognition ANN trained to categorise five colours.

1.3 Thesis overview

The use of deep artificial neural networks has facilitated groundbreaking performance in a range of artificial intelligence applications including natural language processing and complex games. However, the process by which these computational systems transform their input into task appropriate output is still not fully understood. Attempting to understand and explain how ANNs work is not a new problem. Approaches to explaining the operation of relatively simple neural network models have been discussed since the inception of connectionist cognitive science and there are important issues that have still not been fully resolved. Philosophers have sought to understand the types of representation and information processing that occurs in ANNs but they often relied on analysing the behaviour of a very small number of actual ANNs. Given the abundant utilisation and importance of ANNs in artificial intelligence it is timely to revisit some of the foundational and more contemporary discussions of neural network modelling in the philosophical literature and provide a more robust analysis of the claims using a far greater breadth of empirical data. It is also particularly prudent to consider structural-resemblance theories as there has been a shift towards using this type of methodology in contemporary cognitive neuroscience (Williams & Colling 2018).

In this thesis I will review and evaluate a range of approaches to understanding the operation of ANNs by using empirical data from my own unique analysis of a broad range of ANNs which were created for this research project. This analysis will provide insight into how representational content should be characterised and compared between ANNs. The discussion will focus on structural-resemblance approaches and extended empirical analysis using collections of relatively simple three-layer feedforward ANNs specifically generated for this purpose. The thesis is divided into four sections with the major arguments and novel empirical analysis provided in the second and third chapters.

Understanding how ANNs operate requires an explanation of how networks with vastly different architectures and patterns of connection weights can perform the same task. In connectionist cognitive science this is explained in terms of the processing of common internal representational states. In Chapter 2 I provide a review of the development of Paul Churchland's (1989,1996,1998,2007,2012) approach to explaining and assessing representational similarity in connectionist neural network models. Churchland has been a prominent contributor to foundational and contemporary discussions in connectionist cognitive science. He has used AANs to model aspects of biological cognition and also to provide insight into the qualitative feel of conscious experience.

Churchland (1998) explains that relative similarities and differences between collectively considered hidden layer activation patterns correspond to relevant similarities and differences between aspects of the represented domain. He used Laakso and Cotrell's (2000) empirical work to provide a quantitative measure for assessing representational similarity between different ANNs based on shared structural properties. I perform a new analysis of this method using empirical data from my novel facial image categorisation ANNs to show how the comparisons should be extended in order to explicitly compare the relevant structural properties of the representational spaces. This provides a more robust way to assess the representational similarity of distinct ANNs and is used in the subsequent analysis.

My results from the extended comparisons reveal that ANNs can develop robust structural relations between and within the facial categories being represented that were not explicitly trained for, but that this is not always required for accurate categorisation of facial images. I also compare the novel ANNs to the face recognition network originally discussed by Churchland (1996) to provide further insight into the structural relations between representations in different types of ANNs used to categorise facial images. My comparisons reveal that the structural relations formed between represented facial categories in ANNs may differ from our intuitive assessment of facial similarity.

In Chapter 3 I provide an exposition and comparison of three different approaches to determining representational content in neural network models including clustering (Shea, 2007), mutual information (Azhar, 2016) and structural-resemblance (O'Brien & Opie, 2004,2006). Shea (2007) claims that representational content is associated with clusters of activation values rather than individual patterns of activation. He explains that the content of these clusters is determined by the class or category with which they are most highly associated. Azhar (2016) argues that informational content should be associated with polytopes which characterise the potential activation patterns that map to specific output classifications. The representational content of a polytope is determined by the input class with which it shares the highest mutual information. O'Brien and Opie (2004,2006) explain that similarities and differences between the structural arrangement of collectively considered patterns of activation systematically reflect actual similarities and differences in the target domain. Representational content is determined by a structural resemblance between the activation patterns and task relevant aspects of the represented domain. This explanation appears to be consistent with Churchland's (1998,2007,2012) approach.

I compare the various theories of representational content determination and evaluate them using empirical data derived from my analysis of a wide range of novel colour categorisation ANNs. This includes a comparison of ANNs with varying configurations trained to perform three different aspects of colour categorisation from reflectance spectra input data. My analysis shows that distinct ANNs can develop a robust and consistent structure across sets of hidden layer activation patterns and also possess *unexploited* representational content. The representational structures that develop during training are task specific and match external characterisations of the structures of the corresponding task domains. The empirical results provide support for structural-resemblance approaches to explaining the operation and properties of these ANNs.

Chapter 4 begins with a discussion of O'Brien and Opie's (2006) structural-resemblance approach to understanding connection weight representation. I show that although this explanation may appear promising it is difficult to apply to all ANNs due to their unconstrained variety and complexity. The chapter concludes with a summary of the key issues and findings that my research has revealed along with some suggestions for future research.

Chapter 2 - Assessing representational similarity between artificial neural networks (ANNs)

Paul Churchland has been a prominent contributor to foundational and contemporary discussions regarding representation in connectionist neural network models. This chapter provides an exposition of Churchland's (1996,1998,2007) approach to characterising representational content in artificial neural networks (ANNs) and his application of Laakso and Cottrell's (2000) method for assessing representational similarity across different ANNs. This approach will be subjected to new and more rigorous analysis using empirical data from my novel facial image categorisation ANNs. My empirical investigation will highlight and compare the structural organisation of hidden layer activation spaces across a diverse range of facial image categorisation ANNs. The analysis will show how Laakso and Cottrell's method should be extended in order to explicitly compare the relevant structural properties of these representational spaces and provide additional insight into the relations between the representational states. Finally, I will discuss the use of structural methods for comparing representational content in cognitive neuroscience to provide further support for using this type of approach.

2.1 Representation in connectionist neural networks models

In order to provide an adequate explanation of how artificial neural networks (ANNs) operate it is necessary to understand how they can perform the same task with vastly different configurations. It seems likely that ANNs that perform the same input-output transformations with different architectures and connectivity must still have some commonality or shared properties that facilitate this capacity. In connectionist cognitive science the operation of neural networks is explained in terms of the processing of internal representational states. So, when different ANNs perform the same task it seems likely that they are manipulating representations with corresponding contents and the transformations they perform respect these contents.

Fodor and Lepore (1992,1996) challenged the connectionist approach and questioned how ANNs with varying configurations could have the same representational content instantiated by different patterns of activation. They argued that ANNs could not be plausible models for understanding aspects of cognition if there was no way to establish the identity or at least similarity of representational content between networks with different architectures, connections weights and patterns of activation. They explained that people can have thoughts with the same meaning and these thoughts are mental representations with the same content. However, it is extremely unlikely that when two people share a thought with the same content that they would have identical patterns of activity across the relevant neural locations because everyone has a distinct idiosyncratic configuration of synaptic connections and patterns of neural activity.

In order to address this issue methods for assessing representational similarity between different ANNs need to be explored. The simplest and most direct method for comparing the configuration of ANNs would be to calculate the similarity between corresponding connection

weights. However, connection weights can only be directly compared across networks of the same size and even then the results may be difficult to interpret. ANNs can perform the same task or function with similar (or identical) performance but still have vastly different connection weights due to the stochastic parameters used during training. The initial weight values and the order of presentation of input data usually have some random aspect and consequently the final weight values can vary considerably. There is currently no generally accepted method for comparing the informational or representational content of different patterns of connection weights (connection weight representation will also be discussed further in Chapter 4).

Due to the difficulty with establishing relevant similarities between patterns of connectivity in ANNs and in determining what (if anything) the weights actually represent, explanations have tended to focus on analysing the activation patterns that they generate. In relatively simple three-layer feedforward neural network models, analysis is focussed on the intermediate (hidden) layer because the specific encoding of content into activation patterns in both the input and output layers has been determined by an arbitrary or conventional specification of the model properties. The activation patterns generated at the hidden layer are not directly specified but determined by the training process.

The most direct approach for comparing transient activation states in a neural network would be to compare the actual activation vales from corresponding processing units (or neurons). Some of Churchland's (1989,1996) early examples showed the activation values of each hidden layer unit as a magnitude in a specific dimension of the corresponding activation space. The value along each axis of the activation space represents the degree to which a particular task relevant property is present in the input that generated it. The representational content of the layer is determined by the degree to which all the constituent properties or features are present. Two activation patterns have the same content if all of their elements have the same values and when characterised as activation points they occupy the same position in activation space. This is referred to as a microfeatural approach to representational content determination. The activation of individual processing units corresponds to the degree to which a constituent feature is present. This is a clear and relatively simple description that may have some intuitive appeal, however, Fodor and Lepore (1992,1996) argued that this type of explanation is unsatisfactory.

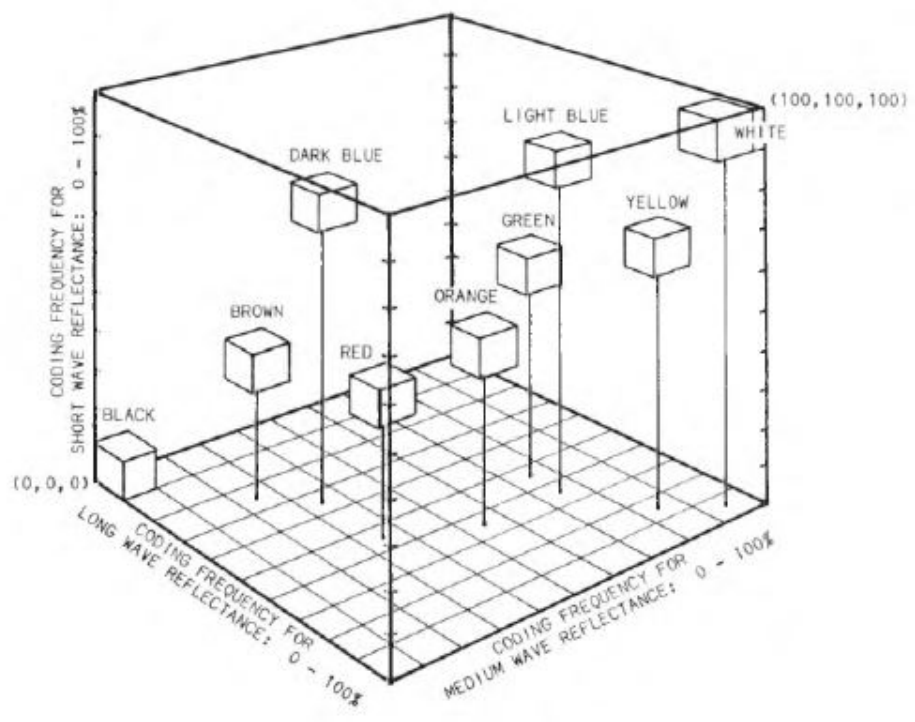


Figure 2.1. Example of a colour space with each axis coding for a particular wavelength range. Colour is determined by the position relative to the constituting axes (Churchland, 1989, p104).

If the representational content of activation patterns is determined by their location in activation space then assessing similarity between different ANNs would require a method for determining the specific property each processing unit was representing in order to match the axes across different activation spaces. Even if it were possible to determine the properties or features associated with each axis, a comparison could still only be applied to activation patterns with an identical number of elements. Additionally, representations with the same content may not necessarily have to be comprised of identical semantic features and may be realised across varying numbers of processing units (or neurons) (Fodor & Lepore 1992,1996). Shea (2007) also highlighted similar problems with using this type of microfeatural approach to representational content determination. He explained that this would be a maximally fine-grained individuation of content and result in networks being ascribed complex representations with content unlike familiar explanations. There is no reason that the semantic dimensions of the task domain would have to match the semantic dimensions of axes defined by activation values of individual nodes. Hidden layers with varying numbers of units could still have the same semantic dimension (Shea, 2007, p249).

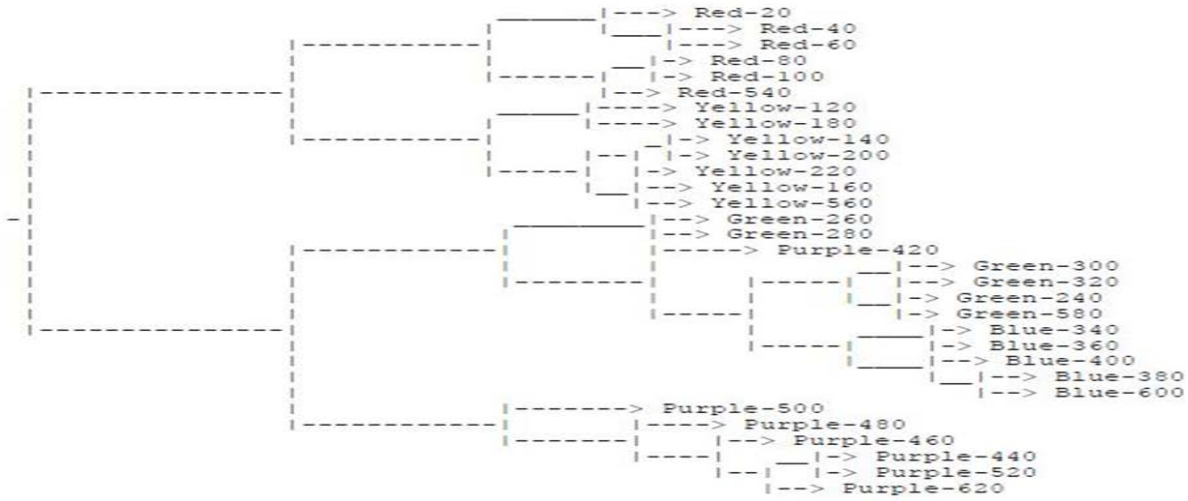
2.2 Characterising representation in neural network activation spaces

Churchland (1998) then provided a revised account and explained that representational content is associated with the relative locations of activation points rather than their absolute position as defined by the constituting axes. Activation points form part of a representational system and content is determined by the relative locations of all content bearing points in the

activation space rather than by specific attributes or properties assigned to the dimensions of the space. On this account, activation patterns are distributed representations rather than a simple concatenation of unit-level representations of individual properties or features of a target domain.

When the hidden layer activation patterns generated by an input dataset are considered collectively they reveal interesting properties about the operation of the ANN. In a successfully trained network the hidden layer activation points tend to be grouped or clustered together in a way that is consistent with the task being performed and facilitates correct output classification. Input patterns that are task-dependently similar generate patterns of activation in the hidden layer that are relatively similar and located within the same region of the corresponding activation space. Conversely, activation points associated with very different input patterns may be far apart and can be separated by a partitioning of the activation space. The task relevant grouping or clustering of points and related partitioning of activation space abstracts away from the absolute values and parameters of the network and facilitates comparison of content across diverse networks (Churchland, 1998).

The grouping of activation patterns can be visualised using dendrograms which show the hierarchical structure or clustering of points. This can be applied to ANN layers of any size as the structure of the space is presented independently of dimensionality. Figure 2.2 shows a comparison between the hidden layer activation spaces of two ANNs trained to perform the same five colour recognition task using the same dataset (Laakso & Cottrell, 2000). Activation points corresponding to the same colour are grouped more closely together and the larger colour clusters tend to be grouped near clusters of similar colours. This shows that the two networks have developed very similar groupings of hidden layer activation points. The dendrograms can be used to make a qualitative comparison of the structure of the activation spaces.



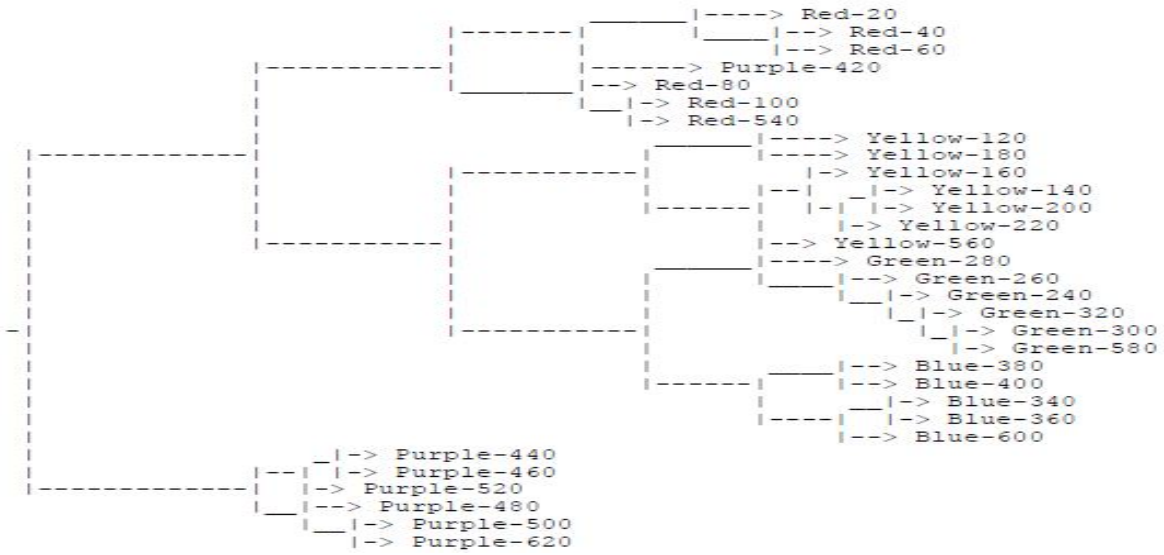


Figure 2.2. Dendrograms showing the grouping of hidden layer activation points in two colour recognition ANNs trained to perform the same task with the same input data (Laakso & Cottrell, 2000, p29).

Churchland (1989,1996,1998,2007) emphasised the representational significance of the relative locations of hidden layer prototypical activation points for each category or task relevant property that an ANN is trained to recognise. A prototypical hidden layer activation point is calculated by averaging the locations of all points in the corresponding region of activation space or determining the 'centre of gravity' that reveals the most characteristic point. The similarities and differences between categories, concepts or aspects of the task domain are systematically reflected by similarities and differences between the distances separating prototypical activation points for those categories. The deviation of an activation point from its corresponding prototypical point may also reflect how the representational content varies from typical examples of that particular category or type.

To help illustrate how activation spaces can be analysed consider a simple hypothetical ANN with two hidden units that is trained to categorise three different classes of input. The activation values of the two hidden units can be shown graphically as points in a two-dimensional activation space. Figure 2.3 shows how the activation points can be categorised and analysed. In this example the activation points are depicted using three different types of shapes to highlight the three different input classes that they correspond to. Variations within each of the three types of shapes are also used to depict variation between inputs belonging to the same categories.

The first image in Figure 2.3 shows how hidden layer activation points generated from three different input categories can be partitioned into three distinct regions of activation space by two linear separators. The second image shows the addition of prototypical points for each category which can be calculated by averaging the locations of all points belonging to the corresponding category. The third image shows the distances between the prototypical points and these distances can be used to characterise the relative locations of the points in activation space. In this example the prototypical activation points belonging to the

categories depicted by rectangles and triangles are relatively close together and both of these two points are relatively distant from points belonging to the third category depicted by red ovals. According to Churchland (1998) the distance relations between prototypical activation points reflect similarity and difference relations between the categories that are being represented by the ANN. In this example that would imply that the first two categories (depicted by rectangles and triangles) would be more similar to each other than they are to the third category (depicted by red ovals). The fourth image shows that the arrangement of activation points within each category can also be analysed. Churchland (1996,1998) claims that the distances between points within a category reflect relations between members of the category being represented. Activation points located near the prototypical point represent members of the category that are more prototypical of that category than points located further away from the prototypical point.

In general, for an ANN of any size the hidden layer activation values generated from the entire input set can be characterised as activation points and analysed to determine the structure of the corresponding activation space. In a trained ANN the activation points generated from each input class will be located in distinct regions of activation space. A prototypical point can be calculated for every input class and the distances between these points can be used to characterise their relative locations. The distances between activation points belonging to each specific category can also be calculate and analysed. Churchland (1996,1998) claims that the relative locations of prototypical points correspond to similarities and differences between the input classes or categories that they represent and the relative locations of points generated from each specific category correspond deviations between members of the category that they are representing.

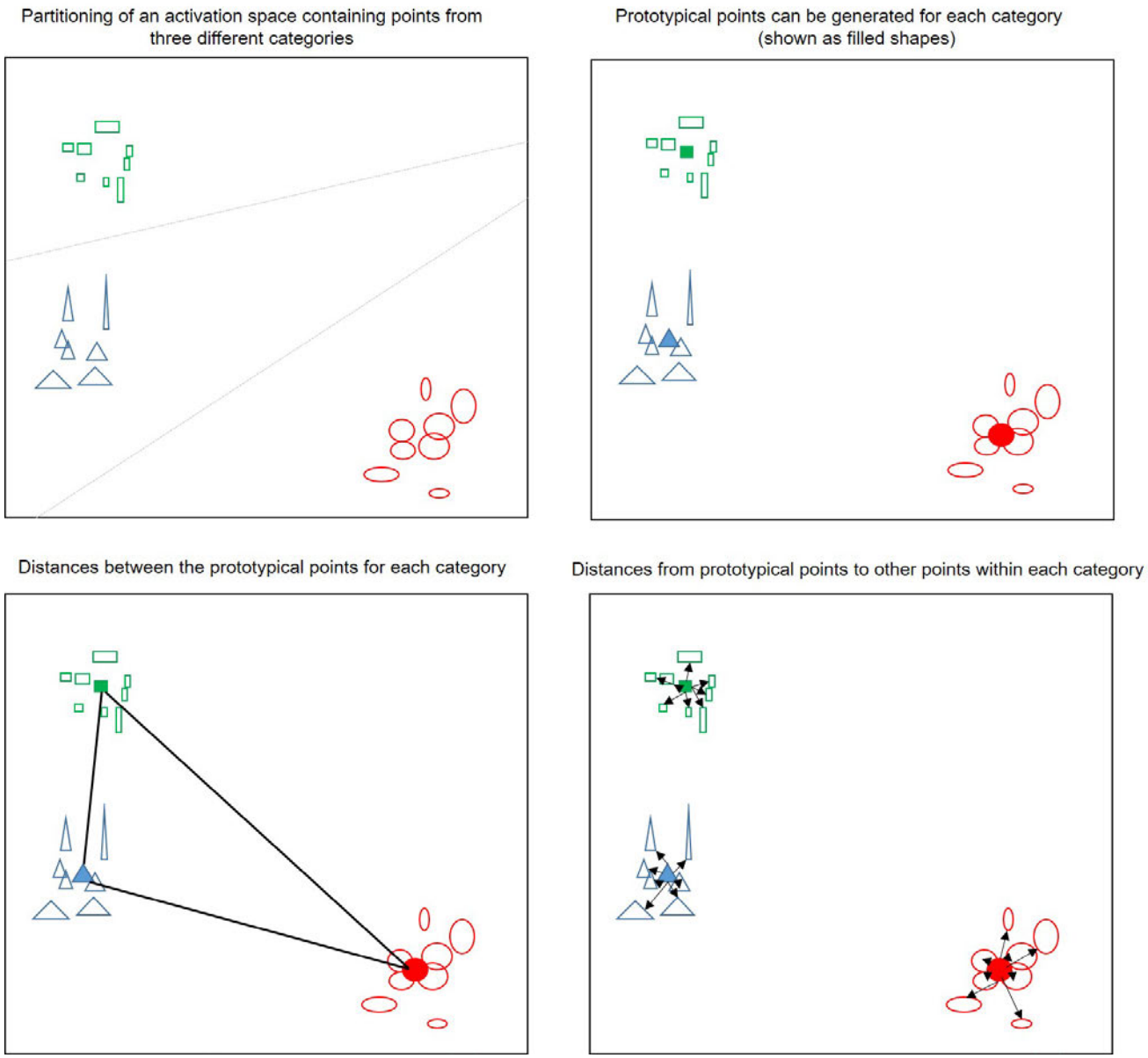


Figure 2.3. Analysis of a hypothetical activation space containing points from three different categories. Both the relations between categories and the relations within individual categories can be evaluated.

Churchland (1996) illustrated the partitioning of hidden layer activation space and location of prototypical points using a face-recognition ANN developed by Cottrell (Flemming & Cottrell, 1990). The three-layer feedforward network has an input layer comprised of a 64x64 unit grid with input activation values corresponding to the grey-scale luminance values from the 64x64 pixel facial images. The hidden layer is comprised of 256 units and the output layer has eight units. Five of the output units are designated to encode an arbitrary binary number which acts as a 'name' for each different face in the dataset, one unit indicates male gender, another unit indicates female gender and there is also one unit to indicate whether the input image is a face or not. The training set has 64 images of eleven different faces and 13 photos of non-face scenes (Churchland, 1996, pp38-55).

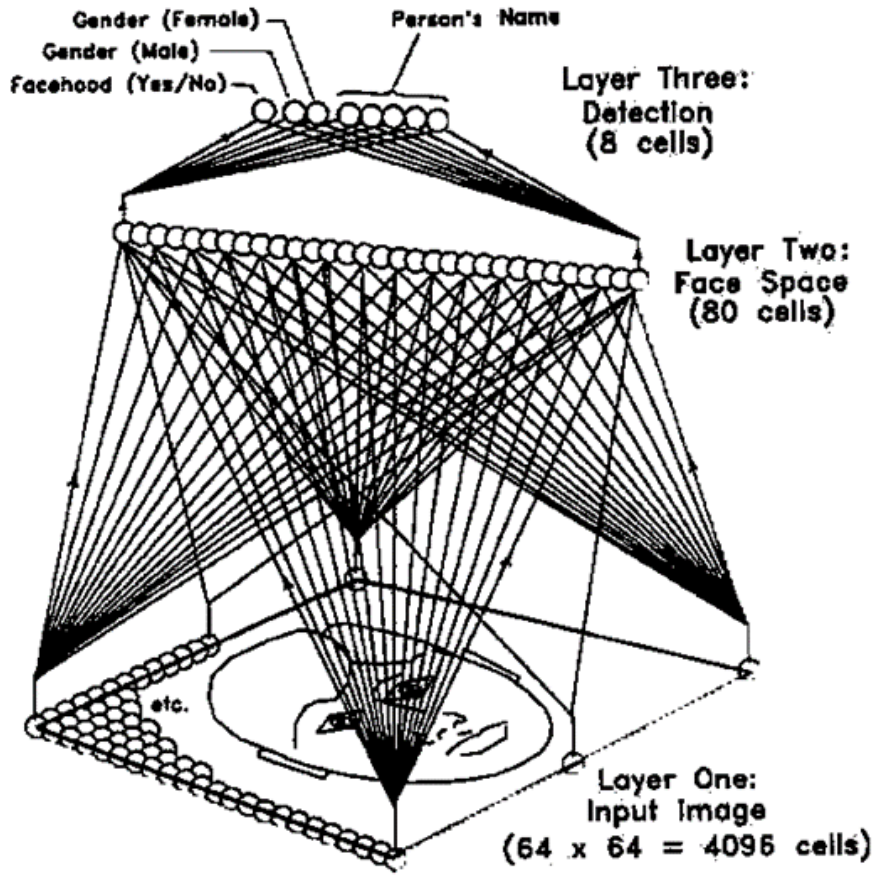


Figure 2.4. A connectionist neural network model for recognising faces (Churchland, 1996, p40).

The ANN was completely accurate in categorising all images in the training set and 98% accurate in identifying novel photos of the eleven different faces. When presented with entirely novel photos of scenes and facial images it achieved 100% accuracy in determining whether the input was a face or not and it was approximately 80% correct in determining gender. A three dimensional visualisation of the hidden layer activation space (Figure 2.5) reveals that the network partitions hidden layer activation space into face and non-face regions. The face region is partitioned into male and female regions which are further divided into regions corresponding to activation patterns associated with each subject. Prototypical activation points were determined for each gender subregion based on the locations of all the individual faces within these regions. The representational content of other points in the space varies according to their distances from the prototypes. For example, the activation points midway between the two prototypes that lie on the partition boundary correspond to gender-ambiguous faces (Churchland, 1996, pp40-53).

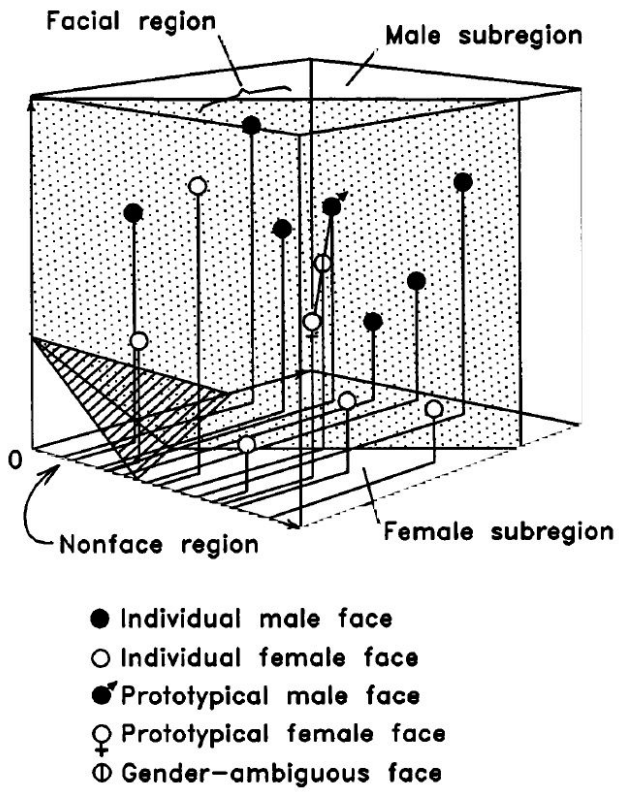


Figure 2.5. Hierarchy of partitions in hidden layer activation space (Churchland, 1996, p49).

Churchland (1996) also provided examples of the input images that would maximally activate individual hidden layer units (some example are shown in Figure 2.6). These images (referred to as holons) show that individual hidden units were not responding to individual or localised facial features but to holistic aspects of the facial images (Churchland, 1996, p48).

Figure 2.6. Holons generated to show the preferred stimulus of individual hidden layer units in the face recognition ANN (Churchland, 1996, p48).

Churchland (1998,2007) also described how the representational structure developed by a trained ANN could be described in terms of the geometric shape formed by connecting the prototypical points in hidden layer activation space with lines. The structure is a hypersolid which may be of very high dimensionality and the length of the edges connecting the vertices reflects the relative proximity of the corresponding prototypes. The internal representational structure of two ANNs would be deemed similar if these hypersolids could be closely matched through a process of linear transformations. Churchland (1998) explained that a common reference point would not necessarily be required because there would generally only be one unique way of matching the structures. Churchland (1998) illustrated this approach using two different activation spaces for hypothetical ANNs that categorise faces into one of four family groups (Hatfields, Andersons, Wilsons and McCoys). The hypersolids defined by the four prototypical activation points have different absolute location and orientation but have the same shape when considered relative to the constituting vertices (Churchland, 1998, p17).

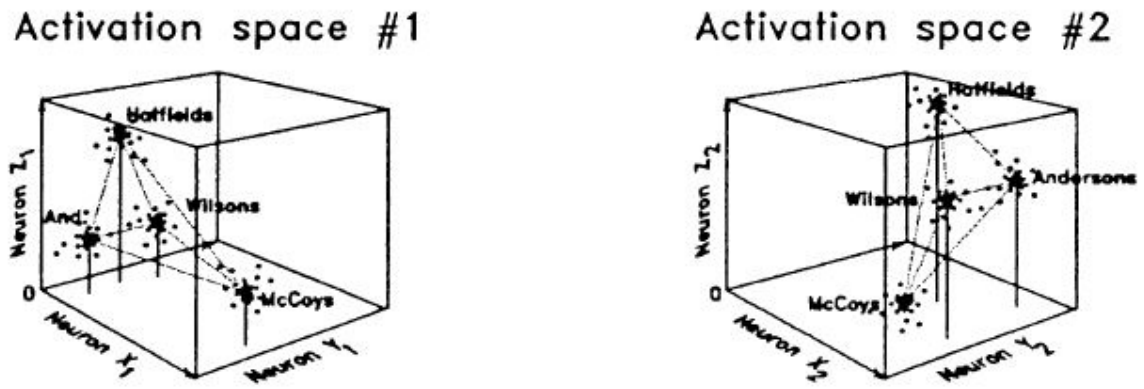


Figure 2.7. There is a similarity between the relative shapes of hypersolids defined by the prototypical activation points for hypothetical ANNs categorising faces into family groups. (Churchland, 1998, p17).

Churchland (1998) provided a method for quantifying the representational similarity between activation spaces in terms of these hypersolids which involves dividing the differences between the lengths of corresponding lines that define the hypersolids by the sum of the lengths and averaged for all edges that connect the prototypical points.

Similarity = 1 – Average [|AB – A'B'| / (AB + A'B')]

(where AB is the length of the line from point A to point B in network one and A'B' is the length of the corresponding line segment in network two and higher values indicate a higher degree of similarity.)

However, rather than applying this method of comparison, Churchland (1998) presented Laakso and Cottrell's (2000) method for assessing representational similarity which comes with supporting empirical results from a range of ANNs. For this reason, and based on my own evaluation of these two methods[1], I focus on Laakso and Cottrell's (2000) method from now on.

2.3 A quantitative method for assessing representational similarity

Churchland (1998) addressed Fodor and Lepore's (1992,1996) claim that it was not possible to assess the similarity of representational content between ANNs with different architectures and patterns of connection weights by reporting a method of comparison developed by Laakso and Cottrell (2000). This approach measures the similarity between the structures of corresponding collections of activation points across different activation spaces by calculating the pairwise distances between all sets of hidden layer activation points for the complete input set. This provides a set of distances that describe the relative locations of each point to every other point. The similarity between two ANNs with different

[1] These two measures were not directly compared by Churchland to establish whether they were consistent. In order to review and compare the validity of these two approaches I performed novel empirical comparisons using my own ANNs and new examples of simple structural arrangements. A detailed description of my empirical analysis and comparison of both methods is provided in Appendix 2. The results reveal that the varying approaches can produce inconsistent measures of similarity. Laakso and Cottrell's (2000) method was shown to be preferable because it is more sensitive to the relevant aspects of structural similarity.

activation values and varying numbers of processing units can then be assessed by comparing the corresponding sets of distances.

Laakso and Cottrell (2000) measure the representational similarity between ANNs by calculating the correlation coefficient (Pearson's r) between the corresponding sets of distances between activation points. This provides a measure of similarity ranging from minus one, indicating an inverse correlation, through to zero which indicates no correlation and up to one for complete correlation. High correlation values are associated with a high degree of representational similarity between the ANNs and low values indicate a lack of similarity.

Figure 2.8 shows how the similarity values obtained using this method align with the similarities and differences in the structural arrangement or relations between four points. For example, the relative locations of points in Structure 1 and Structure 4 are the same and they have a corresponding similarity value of one. The relative arrangement of these points is also fairly similar to Structure 2 and this is reflected by a similarity value of 0.7. However, these three structures all have a significantly different relative organisation of points to Structure 4 (for example, A is closest to D rather than furthest away) and so the similarity values are negative.

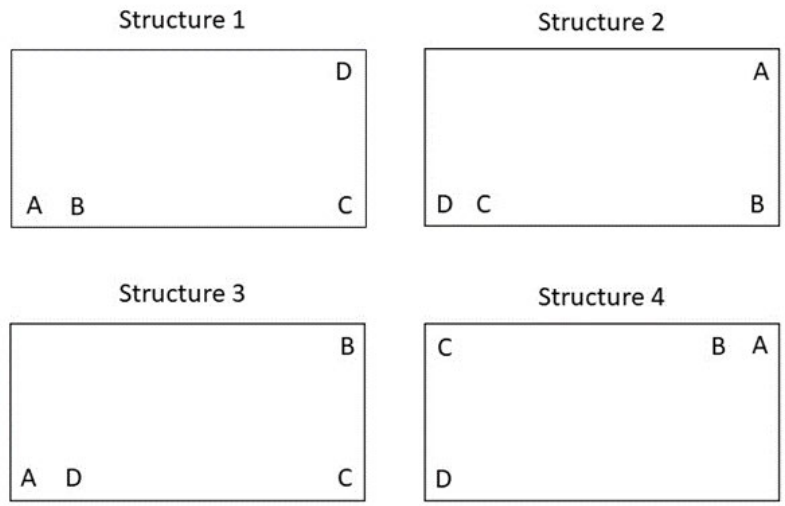


Figure 2.8. Comparison between the structural relations of four different arrangement of corresponding points.

Pairwise correlations between four structures				
Structure	**1**	**2**	**3**	**4**
1	1	0.7438	-0.3303	1
2	0.7438	1	-0.2759	0.7438
3	-0.3303	-0.2759	1	-0.2759
4	1	0.7438	-0.2759	1

Table 2.1. Quantitative similarity between the four structures in Figure 2.8 determined using Laakso and Cottrell's (2000) method.

Some advantages to using this method are that distances can be calculated in activation spaces of any dimensionality and corresponding sets of distances provide a description of the arrangement of points that is invariant to global translation, rotation and inversion. Calculating the correlation between two sets of distances provides a value that is also invariant to the global scale of the distances. This method can be used to determine the representational similarity between layers in different ANNs or to compare different layers in the same network.

Laakso and Cottrell (2000) applied their method of comparison to a wide range of ANNs that were trained to categorise colours from reflectance spectra. The first set of experiments involved training three-layer feedforward networks with three hidden units to distinguish between five colour categories. They used four different encodings of the reflectance spectra that ranged from 3 to 96 input units. The results showed a high correlation (greater than 0.9) between ANNs using the same input encodings and a significant correlation between all other networks except those using the sequential encoding scheme which had an assigned rather than direct relation to the reflectance spectra values. Additionally, the correlation between the set of input patterns and the hidden layer activation patterns was relatively low (approximately 0.2) which indicates that there was a genuine reconceptualisation of the input data during processing rather than just a direct recoding.

The second set of experiments compared ANNs with varying hidden layer dimensions that ranged from one to ten units and utilised two different sets of input encodings (real and sequential). The networks with three or more hidden units (the minimum required to achieve satisfactory performance) had highly correlated sets of hidden layer activation patterns (approximately 0.9) and low correlation with the input layer (approximately 0.3). Ten additional ANNs with five hidden units but different initial random weights were also trained for each of the two input encodings. Having different initial weights results in the networks developing substantially different connection weights during the training process. The hidden layers of the networks trained using the same input encoding were also highly correlated (0.93). These results show that ANNs with the same number of hidden units but *different* connection weights can develop similar representational schemes when trained on the same task. This supports the results from the first experiment and shows that ANNs with varying hidden layer dimensions and different connection weights can develop hidden layer activation spaces with similar relative locations of corresponding sets of activation points. In order to verify these results I recreated the full set of experiments presented by Laakso and Cottrell (2000) and obtained similar and consistent results but with subtle variations due to stochastic and unspecified training parameters.

Churchland (1998, p19) mentioned that including all activation points in the comparisons rather than just prototypical points would provide a more fine grained measure of similarity. However, results from the new empirical investigation that I have undertaken will show that the correlation values calculated using all points may not necessarily be consistent with comparisons using the prototypical points.

2.4 Extended assessment of representational similarity

Laakso and Cottrell's (2000) method for assessing representational similarity compared collections of activation patterns generated from entire input domains. Representational similarity was measured using the correlation between the pairwise distances between all points in corresponding activation spaces. However, the structure of activation spaces can also be analysed in terms of relations between prototypical activation points and the deviation of points from their corresponding prototypes. I will now investigate the relation between correlation values and the preservation of relevant structure by extending Laakso and Cottrell's (2000) original method for assessing representational similarity and applying this approach to compare a wider range of ANNs. I provide a novel empirical analysis of the structural organisation of hidden layer activation spaces that can be used to compare the representational frameworks developed by ANNs and inform theories of representational content individuation and determination.

I created and analysed a range of different feedforward ANNs with one hidden layer using Matlab with default pattern recognition network parameters. All of the ANNs were trained using supervised backpropogation to perform categorisation of facial images. The classification task involved identifying the particular individual that was depicted in an input image. The output layer of each network was comprised of units designated to represent each different individual present in the input dataset. The ANNs were trained to generate an activation value of one in the output unit assigned to the individual in the input image and zero in all the other output units.

I used two datasets including the California Facial Emotion (CAFE) dataset containing 8 images each of 10 different people (grey scale 240 X 380 pixels) and the Yale dataset containing 11 images each of 15 different people (grey scale 195 X 231 pixels). Firstly, I created a series of 20 ANNs with the same size hidden layer as Churchland's (1996) example (80 units) for each dataset. I then created a series of different ANNs ranging from 5 hidden units increasing sequentially up to 25 hidden units (21 different architectures) with 5 distinct networks of each architecture using different initial weights (21 X 5 = 105 ANNs in each series). A random selection of 75% of the input set was used for training, 15% for validation and 10% removed to test generalisation after training.

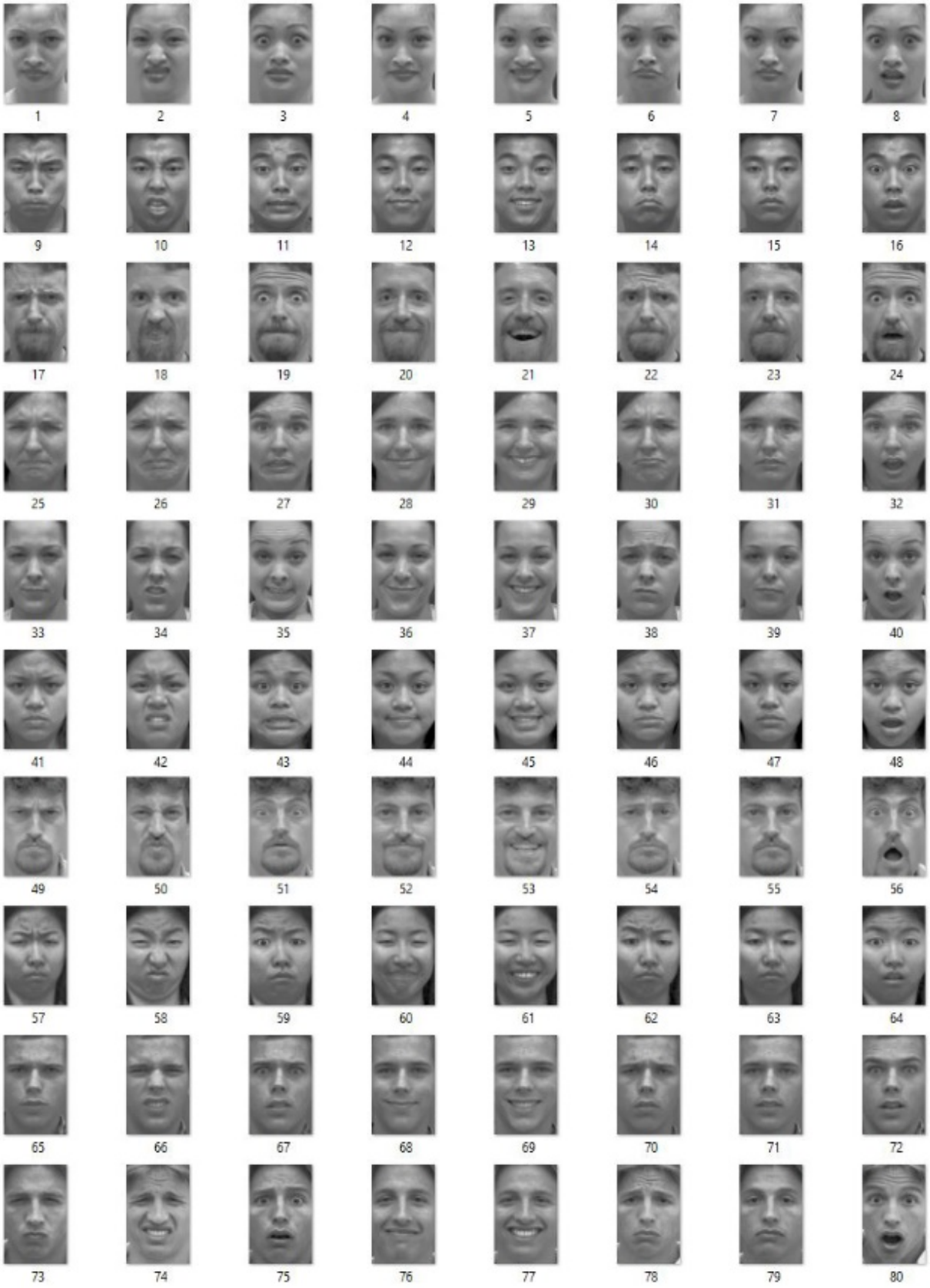


Figure 2.9. Images from the CAFE face dataset. The dataset contains eight photos each of ten individuals displaying the following emotions: anger, disgusted, fear, happy, happy with teeth showing, maudlin (sad), neutral, surprise.

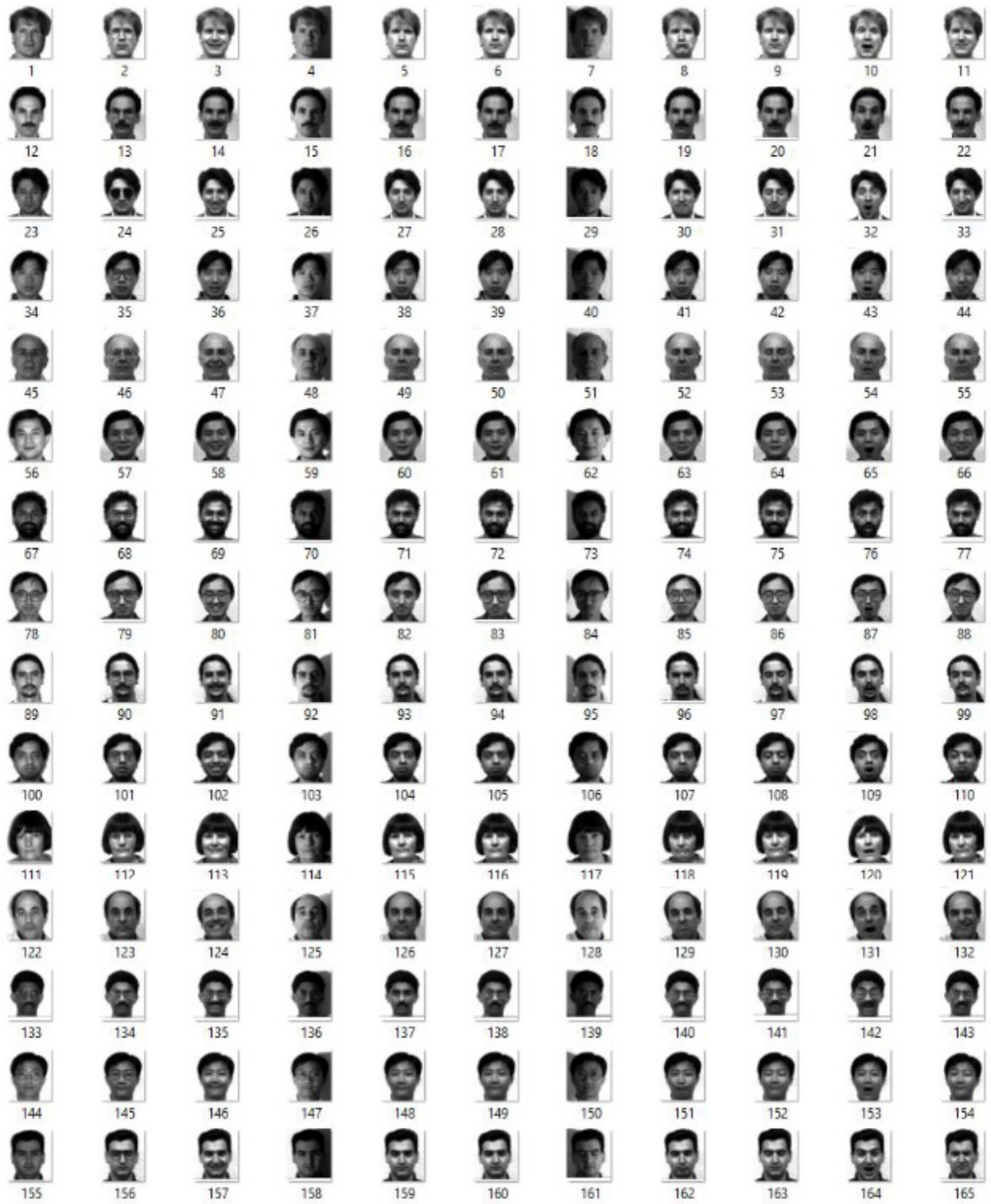


Figure 2.10. Images from the Yale Face Databaset. The dataset contains 11 images each of 15 individuals with the following emotion or context: centre light, with glasses, happy, left light, no glasses, normal, right light, sad, sleepy, surprised, wink.

Firstly, I calculated the representational similarity between the hidden layers of every pair of ANNs trained on the same input data using the original method described by Laakso and Cottrell (2000) and averaged the results. I then modified this method to directly compare the

prototypical activation points for each category (each of the different individuals pictured) that the ANNs had been trained to recognise. The prototypical points for each category were calculated by averaging the complete set of points belonging to that category. The correlation between pairwise sets of distances was then calculated using the prototypical points only rather than the full set of activation points. All pairwise ANN comparisons were calculated for each group of networks and then averaged. These comparisons provide a measure of the inter-category structural similarity developed across network layers.

I then modified the comparisons to determine the similarity between the relative locations of points belonging to the same categories. The full set of activation points belonging to each category that an ANN had been trained to recognise were identified and the respective sets of distances calculated for all points in each of the categories. For each pair of ANNs being compared, the correlation between corresponding sets of distances within each category was calculated and then averaged to determine an overall measure of similarity. This was repeated for all pairs of networks and the results were averaged to obtain an overall measure of intra-category structural similarity between the hidden layers.

2.5 Empirical results from facial image categorisation ANNs

The ANNs achieved close to complete accuracy on the training data (98.73-99.75%) and between 91 and 99 percent accuracy on the withheld test data. The series of networks trained to categorise faces from the CAFE image dataset using 80 hidden units had a very high average correlation between corresponding hidden layer distances when calculated using all activation points (0.94). There was also a high correlation between the arrangements of prototypical activation points (0.81) and a significant average correlation between the arrangements of points within each corresponding category (0.66).

The set of networks trained on the Yale face dataset provided fairly similar and consistent results to those obtained using the CAFE face dataset. The average correlation between the corresponding distances separating all hidden layer activation points was lower but still relatively high at 0.88. There was also a high correlation between the arrangement of prototypical activation points (0.82) and a high correlation between the arrangements of points within each corresponding category (0.87).

ANN training dataset	**CAFE Faces**	**Yale Faces**
Hidden Layer Size	80	80
HIDDEN LAYER COMPARISONS (Average correlations for all pairs of networks)		
Comparing all activation points	0.94	0.88
Comparing prototypical (inter-category) activation points	0.81	0.82
Comparing intra-category activation points	0.66	0.87

Table 2.2. Average correlation between the facial image categorisation ANNs with 80 unit hidden layers.

A summary of the average correlations between the relative locations of corresponding groups of hidden layer activation points for the ANNs with 80 hidden units is provided in

Table 2.2. These results show that although there is some variation between the similarity measures when calculated using all activation points compared to using the prototypical points they are still fairly similar and consistent. However, this is not necessarily always the case. My empirical analysis of ANNs with smaller hidden layers shows that they can be trained with a high level of accuracy and pairs of different networks can be assessed as having a high degree of representational similarity using Laakso and Cottrell's (2000) method but have virtually no structural similarity between corresponding prototypical activation points at all.

The ANNs with smaller hidden layers ranging between five and 25 hidden units had a fairly high degree of hidden layer similarity when calculated using Laakso and Cottrell's (2000) original method that includes all activation points. The average correlation between the corresponding distances separating all hidden layer activation points was 0.83 for the CAFE dataset and 0.74 for the Yale dataset. However, this was not the case using the new measures of inter-category and intra-category similarity that I have introduced. The average correlation between corresponding distances between prototypical hidden layer activation points was almost zero for both datasets (0.07). The average correlation between the distances between hidden layer points within corresponding categories was also very low (0.06 for the CAFE dataset and 0.23 for the Yale dataset).

ANN training dataset	**CAFE Faces**	**Yale Faces**
Hidden Layer Sizes	5 to 25	5 to 25
HIDDEN LAYER COMPARISONS (Average correlations for all pairs of networks)		
Comparing all activation points	0.83	0.74
Comparing prototypical (inter-category) activation points	0.07	0.07
Comparing intra-category activation points	0.08	0.23

Table 2.3. Average correlation between the facial image categorisation ANNs with 25 or less hidden units.

The similarity analysis of the facial image categorisation ANNs with 25 or less hidden units is summarised in Table 2.3. My new extended analysis indicates that although there was a high degree of hidden layer similarity between ANNs when calculated across the entire domain there was almost no similarity at the inter-category (prototypical) or intra-category levels. This shows that Laakso and Cottrell's (2000) original method does not always provide results that are consistent with the structural similarity of the prototypical activation points or intra-category activation points.

2.6 Analysis of structural representation in facial image categorisation ANNs

My empirical results show how extending the original method for assessing representational similarity across different ANNs to include inter-category and intra-category comparisons provides a more explicit description of the type of structural similarity that is preserved. The original method described by Laakso and Cottrell (2000) determined similarity values by the correlation between corresponding sets of distances between all points in the task domain.

However, my new results show that it can be useful to perform additional comparisons using task-related subsets of points in order to compare the task specific structural arrangement and relative locations of activation points. Including comparisons of the distances between corresponding sets of prototypical or average activation points for each category provides a measure of the inter-category similarity. Including comparisons of the distances between sets of points belonging to corresponding categories provides a measure of intra-category similarity. It may seem reasonable to expect that these values would be similar or at least consistent, however, my empirical results have shown that although they can be similar the values can also vary considerably in some cases. Analysis of facial categorisation ANNs with relatively small hidden layers showed that the similarity values calculated using the original Laakso and Cottrell (2000) method which compared the distances between all points were not consistent with the inter-category comparisons or the intra-category comparisons. However, there was much greater consistency with the ANNs that had larger 80-unit hidden layers.

Churchland (1998) explained that representational similarity can be assessed by comparing the relative locations or structure of prototypical activation points associated with category exemplars in corresponding activation spaces. This is a comparison of inter-category similarity. Additionally, the deviation of points from their corresponding prototypical point can reflect how their representational content differs from a prototypical example of the particular content type. The preservation of the structure of points within corresponding categories can be assessed by the intra-category similarity. Churchland (2012) has described activation spaces as conceptual landscapes where similarities and differences in the relative locations of activation points reflect actual similarities and differences in the represented domain.

O'Brien and Opie (2004) also describe how similarities and differences in the locations of hidden layer activation points reflect relevant similarities and differences in the objects or properties of the represented domain (an exposition of their position will be provided in Chapter 3). This suggests that the structural differences between activation points that define categorical distinctions, and between points within corresponding categories and across the collective set of points constituting a particular task domain, are important for determining their representational content. Explicit comparison of these relations in facial image categorisation ANNs has revealed interesting differences between Churchland's (1996) original face recognition ANN example, novel networks with the same number of units as the original example (80 units) and networks with a relatively small number of hidden units.

Both Churchland (1996) and O'Brien and Opie (2004) have discussed the representational properties of the simple three-layer feedforward face-recognition network developed by Cottrell (Flemming & Cottrell, 1990). Analysis of the hidden layer activation space shows partitioning between regions coding for male faces, female faces and for non-face inputs. Churchland (1996) explained that the average or most characteristic activation point for each region revealed prototypical examples of the categories and showed that a prototypical gender-neutral face would be determined by the point midway between the male and female prototypes.

The ANN used in this example had specific output units for determining gender in addition to the units for each facial category. If the network was specifically trained to make this distinction, then the activation space would need to be partitioned according to gender in

order to facilitate accurate performance. Different categorical distinctions could potentially be imposed as part of the output classification required for facial recognition ANNs. For example, instead of having output units to encode gender they could have been designated to code for other aspects of facial features such as eye, nose or mouth size and position or relations between them. If an ANN could be successfully trained using an alternative distinction it would develop partitioning that respects the chosen attribute and have prototype activation points corresponding to that distinction.

Churchland (1996) did not explain how the face recognition network that he described was trained but further investigation reveals that it was not just a standard application of supervised learning using backpropagation. This ANN was initially configured as an autoencoder that used unsupervised learning to reconstruct the entire facial image inputs at the output layer after compression through an 80-unit hidden layer. The hidden layer was subsequently connected to the final output classification layer that had specific units for gender and the individual facial categories. Supervised learning was then used to determine appropriate connection weights between the hidden and output layer (Flemming & Cottrell, 1990). The representational organisation developed at the hidden layer during the unsupervised learning process facilitated the categorisation tasks that were finally imposed. Autoencoders can be used to pre-train ANNs so that they develop some kind of conceptualisation of the task domain prior to supervised training. The new empirical data that I have provided is from ANNs that were trained using supervised learning with one output unit assigned for each individual person (facial category) and were not trained to explicitly categorise additional distinctions such as gender.

There was a high correlation between the sets of distances between corresponding hidden layer activation points for each group of facial image categorisation ANNs that were analysed. All the networks developed a partitioning of activation space corresponding to the separation of facial images belonging to the same person. However, in the ANNs with relatively small hidden layers (5-25 units) there was almost no correlation between the inter-category and the intra-category distances which indicates that the structure was not preserved at either of those levels. In this case it appears that the only structural similarity is the clustering or regional separation that provides the shared categorical distinctions.

In contrast to the ANNs with small hidden layers, the corresponding distances between prototypical activation points in networks with 80-unit hidden layers were highly correlated. This indicates that there is a robust preservation of structural relations between the hidden layer representations of different individuals in these larger networks. There was also a significant to high correlation between the arrangements of points within corresponding categories. This indicates that relations between hidden layer representations of different facial images belonging to the same individuals are being preserved across the networks. The ANNs developed a conceptualisation of the input domain that includes fairly robust and consistent relations both between and within categories that were not explicitly trained for. However, the grouping of category prototypes that correspond to relations between the facial images from different people do not necessarily reflect the separation by gender that was seen in Churchland's (1996) example where it was explicitly trained for.

It is not clear whether the relations between prototypes in these relatively simple and artificial facial image categorisation ANNs reflect typical similarity or difference relations that humans may use (or claim to use) to categorise or compare faces. For example, analysing one of

the ANNs with 80 hidden units trained using the CAFE dataset reveals that the closest prototype points are for person seven (images 49-56) and ten (images 73-80). The most distant prototype points are for person one (faces 1-8) and person eight (faces 56-64). A comparison of the neutral facial images for these individuals is provided in Figure 2.11 below. It is unclear whether people would agree that the two faces on the left look more similar to each other than the two faces on the right.

The same example ANN was also analysed to determine the similarity between person one and each of the other individuals. Figure 2.12 shows the ordering of similarity relations between person one and every other person based on the distances between their prototypical activation points. There may not be any general agreement with this ordering of facial similarity and it is not clear exactly what factors are contributing to it. However, the similarity relations are fairly consistent (but can still vary) across the different ANNs as indicated by the high correlation values between the corresponding prototype distances.

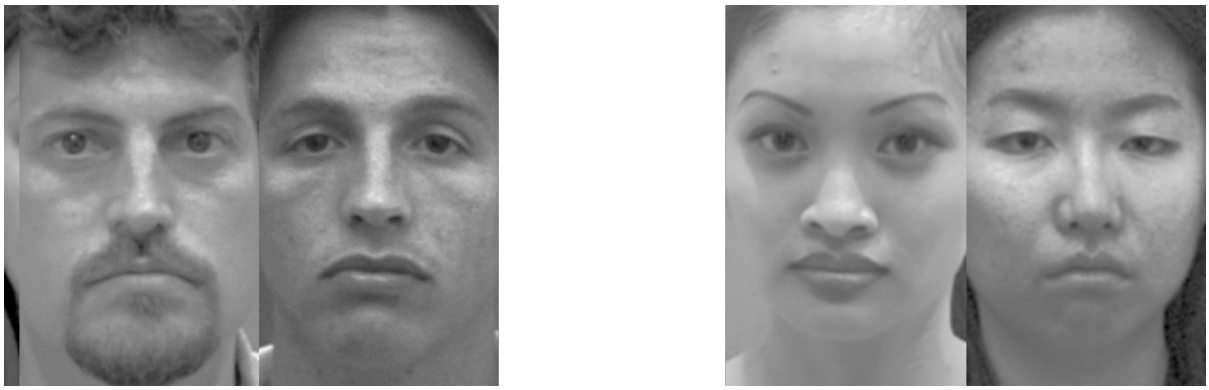


Figure 2.11. Examples of the most similar (left) and dissimilar facial images (right) from the CAFE dataset based on the distances between their corresponding hidden layer prototypes in an example ANN.

Figure 2.12. The left-most image is the first person in the CAFE dataset and each image to the right is ordered in terms of their degree of similarity to this person based on the distances between corresponding hidden layer prototypes from an example ANN. The further to the right the images are positioned the further away their corresponding prototypical activation points are to the prototype for the first image.

Similar comparisons were also performed for an ANN trained on the Yale facial image dataset. The results appear consistent with those obtained using the CAFE dataset. A comparison of the facial images from the individuals determined to be most similar and most

dissimilar are provide in Figure 2.13 below. There may be some agreement that the facial images determined to be most similar do appear more similar than those determined to be most dissimilar. However, it is unclear whether there would be any consensus regarding the ordering of similarity between a specific individual and all other individuals. For example, the ordering of similarity between person one and every other individual is shown in Figure 2.14.

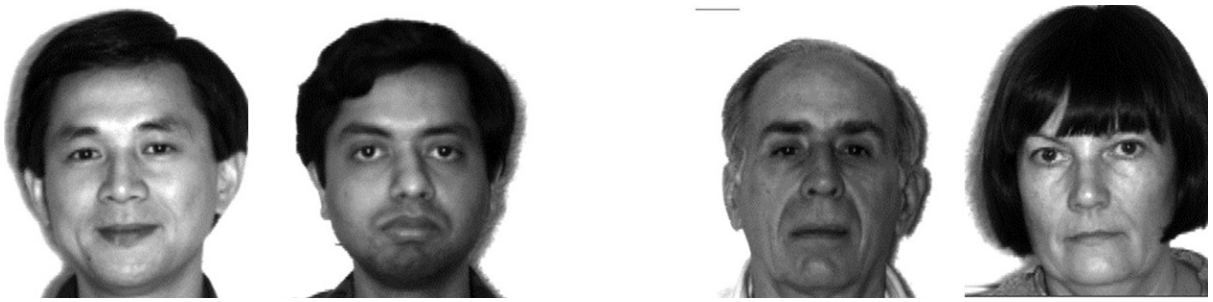


Figure 2.13. Examples of the most similar (left) and dissimilar facial images (right) from the Yale dataset based on the distances between their corresponding hidden layer prototypes in one of the ANNs.

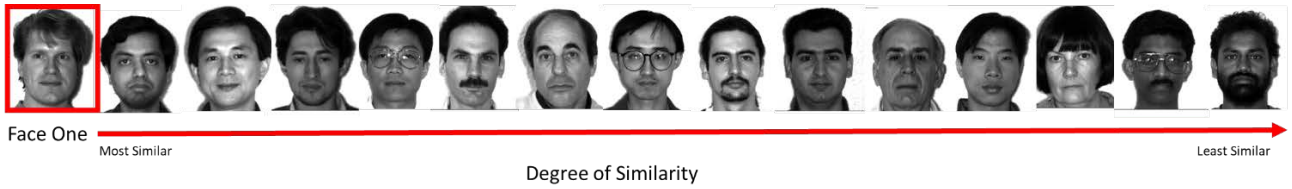


Figure 2.14. The left-most image is the first person from the Yale dataset and each image to the right is ordered in terms of their degree of similarity to this person based on the distances between corresponding hidden layer prototypes from an example ANN. The further to the right the images are positioned the further away their corresponding prototypical activation points are to the prototype for the first image.

The similarity relations that have been identified may reflect the types of statistical regularities that these ANNs learn in order to facilitate correct classification and could also provide insight into their cognitive plausibility. The analysis of these new ANNs has provided some insight into how the task relevant relations that ANNs developed during training may align with human judgements of similarity. Further empirical analysis will be provided in Chapter 3 that shows ANNs categorising various aspects of colour develop robust structural relations which do align with human colour judgements.

Churchland (1998) provided extremely valuable insight into the characterisation of representational in neural networks and the application of Laakso and Cottrell's (2000) method for quantitatively assessing representational similarity across diverse networks. The modification of this approach to include additional comparisons of inter-category and intra-category similarity has provided a more explicit and robust assessment of representational similarity between diversely configured groups of facial image categorisation ANNs. The extension of this method also facilitates comparisons that may overcome Garzón's (2003) objection to using Laakso and Cottrell's (2000) method for assessing representational

similarity between different ANNs. He stated that this method could only be applied to compare networks with identical input sets and target domains and argued this was very unrealistic.

Garzón (2003) explained that people can develop the same or similar concepts and categorical distinctions without being exposed to identical stimuli. The amount and quality of real examples for a particular concept vary considerably between individuals but they still learn concepts with common meanings. However, assessing representational similarity based on comparisons between relative locations of prototypical activation points rather than all activation points generated from the input domain facilitates the application to ANNs with non-identical input sets. Prototypical activations from ANNs classifying the same categories can be compared even if there are no common input samples in each category. This approach could also be used to compare the common prototypical structure across diversely configured groups of ANNs that have only partially overlapping task domains and output category distinctions. An example of how this approach can be applied to compare colour categorisation ANNs with partially overlapping input domains will be provided in Chapter 3.

2.7 Representational similarity analysis in cognitive neuroscience

Laakso and Cottrell's (2000) method for assessing representational similarity was introduced by Churchland (1998) to respond to claims that ANNs could not be biologically plausible models of cognition if diversely configured networks could not have the same (or similar) representational content. Recent work in cognitive neuroscience also supports the use of more abstract and relational methods of comparison. A similar type of approach to Laakso and Cottrell's (2000) method for assessing representational similarity is now also being used in cognitive neuroscience and is referred to as representational similarity analysis (Kriegeskorte et al, 2008) or representational geometry (Kriegeskorte & Kievit, 2013). This has been used to compare the representational content of actual brain states based on scans such as fMRI. Characterising representational content in terms of relations between brain states has the same advantages as the methods described for analysing ANNs. Comparisons of representational states do not require the determination of a specific correspondence between neurons, units, voxels, regions or whatever the relevant representational unit might be. Comparisons can also be made between activation patterns based on varying types of scans, with computational models and can be applied across different subjects and species.

The method involves determining all the pairwise relations between the patterns of brain activity (measured by haemodynamic response in the case of fMRI) associated with exposure to each stimulus category. The relations can be characterised in terms of either the similarity or dissimilarity between pairs of patterns and provides a measure of the difference between them. This difference is usually measured by spatial correlation but other metrics can be applied. Similarity is equated with the correlation value and the dissimilarity is calculated as one minus the correlation. A measure of dissimilarity may be preferred because it reflects the distance between the patterns being compared. All the pairwise difference values are stored in a dissimilarity (or similarity) matrix (RDM) which is symmetric around the main diagonal. The representational similarity between individuals or with computational models is determined by the correlation between corresponding RDMs

which indicate the difference values for corresponding sets of stimuli. This process follows a similar approach to Laakso and Cottrell's (2000) method which involves calculating all pairwise Euclidean distances between activation points and then determining representational similarity by the correlation between corresponding sets of distances.

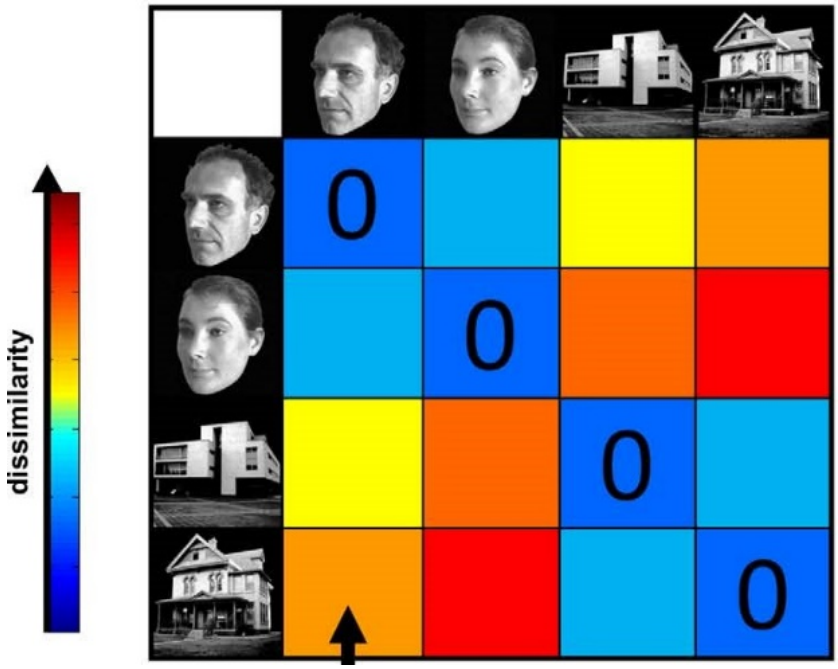


Figure 2.15. Dissimilarity matrix created from brain states corresponding to four different stimulus images (Kriegeskorte et al., 2008, p4)

Associating representational content with the relations between relevant brain states has also been successfully applied to perform across-subject decoding of brain states from fMRI data (Raizada & Conolly, 2012). The across-subject decoding requires labelled similarity matrices for other subjects that have been exposed to the same collection of stimulus categories. An unlabelled similarity matrix that only contains the similarity values between the patterns that require decoding is also constructed. Every permutation for labelling the patterns is then determined and used to produce a collection of matrices that covers all possible combinations of the stimulus categories being compared. The matrix that is most highly correlated with the similarity matrix constructed by averaging the real comparisons for each of the known subject decodings is determined to be the correct labelling. Experiments performed by Raizada and Conolly (2012) achieved a decoding accuracy of 91.7% across eight categories and six subjects (44 of 48 correctly labelled). This provides empirical evidence that supports the use of relational based approaches to comparing representational content in real cognitive systems. This type of approach could also be applied to ANNs but in this case the labels are generally already determined.

The methods for comparing or predicting representational content based on similarity relations were applied to compare patterns of activity generated from exposure to a range of different stimulus categories and focuses on the inter-category similarity. This is consistent

with Churchland's (1996,1998,2007) claim that representational similarity can be determined by comparing the structure or relative locations of prototypical activation points which are exemplars for the categories or relevant aspects of the task distinguished by a neural network model. Extending Laakso and Cottrell's (2000) method for assessing representational similarity to explicitly include comparisons of distances between the prototypical activation points that define inter-category similarity is more directly related to this method than the original approach which included all activation points. When analysing large and complex datasets of brain activity it may be useful to distinguish between different levels of similarity relations including comparing all individual members of the dataset, comparing categories or prototypes, and considering how similarities within particular categories may reflect deviations from prototypes. My novel analysis of ANNs has shown that applying these distinctions can reveal different aspects of the overall representational structure.

The use of relational methods for comparing representational content based on data from brain scans provides empirical evidence that assessing representational similarity by considering the correlation of differences in activation patterns can be successfully applied to biological cognition. This approach can also assist in determining the biological plausibility of neural network models. Comparing the inconsistent results obtained for the simple artificial face categorisation networks with data from actual biological cognitive processing of facial stimuli may reveal whether they are useful connectionist models at any relevant level of representational abstraction.

2.8 Summary

ANNs can perform the same task with different patterns of connectivity and distinct patterns of hidden-layer activation for corresponding inputs. Churchland (1996,1998) argued that the hidden layer activation space of appropriately trained ANNs is partitioned into task relevant distinctions with corresponding representational content. Prototypical activation points can be determined for each category and the representational content of activation points varies systematically with their distance from the prototypes. The relative locations or arrangement of collectively considered activation points reflects relevant relations in the represented domain.

Churchland (1998) also presented Laakso and Cottrell's (2000) quantitative method for assessing representational similarity between different ANNs by comparing the structural similarity between corresponding collective groups of activation points. I have shown that this approach should be extended to explicitly include comparisons between prototypes characterising categorical distinctions and comparisons within partitions corresponding to distinctions between members of the same category. These comparisons of structural similarity between corresponding activation spaces across diverse ANNs can be used to inform representational accounts of both artificial and biological neural networks and evaluate the biological plausibility of connectionist neural network models.

My empirical analysis of diversely configured collections of facial image categorisation ANNs showed that networks with relatively small hidden layers developed robust hidden layer partitioning that facilitated accurate performance but did not develop consistent relations

between or within the facial categories represented. However, ANNs with sufficiently sized hidden layers (80 units) developed robust and consistent structural relations between hidden layer activation patterns across different networks. There were similar structural relations preserved between the prototypical activations for each facial category and also between corresponding activation patterns within each facial category that were not explicitly trained for. However, these relations were not necessarily consistent with the gender distinction that was explicitly trained for in the example face recognition ANN described by Churchland (1996). It is unclear how closely the similarity relations that developed correspond with human judgements of facial similarity or whether they reflect artificial properties of the task domain. My empirical investigation is continued in Chapter 3 and reveals that ANNs categorising various aspects of colour develop robust structural relations which do align with human colour judgements.

Chapter 3 – Representation in neural network models

There are a variety of approaches to characterising the entities or vehicles of representation in neural networks and explaining how their content is determined. Shea (2007) claims that representational content is associated with clusters of activation patterns and the content is determined by the class or category with which they are most highly associated. Azhar (2016) argues that informational content should be associated with regions of activation space called polytopes and that their content is determined by the input class with which they share the highest mutual information. O'Brien and Opie (2004,2006) claim that representational content is determined by a resemblance between the structural arrangement of collectively considered activation patterns and the structure of relevant aspects of the task domain. This account appears to be consistent with Churchland's (1996,1998,2007,2012) approach that was described in Chapter 2. I begin this chapter by providing an exposition and comparison of these approaches. This is followed by a description of my original empirical investigation analysing a range of novel artificial neural networks (ANNs) categorising various aspects of colour. The results are used to evaluate the plausibility of the different accounts of representational content determination.

I trained diversely configured groups of ANNs to categorise three different aspects of colour from reflectance spectra data. The ANNs were trained to determine either the hue (colour), chroma (saturation), or value (lightness) associated with the reflectance spectra. I analysed the ANNs by applying the extension of Laakso and Cottrell's (2000) method for assessing representational similarity between ANNs that was developed and described in Chapter 2. To begin with I compared the similarity between the representational structures of ANNs performing the same categorisation tasks and then contrasted this with comparisons between networks performing different categorisation tasks using identical inputs. The comparisons included an analysis of the structures of the prototypical activation patterns for each category in the particular task domain and the relations between the structural arrangements of activation patterns within corresponding categories. I then also compared these groups of ANNs to corresponding groups of networks trained using a reflectance spectra dataset with higher resolution encoding, more reflectance spectra samples and more hue categories. Finally, I compared the representational structures of the ANN hidden layers with independent and objective external representations of the structures of the task domains using category prototypes to determine whether there is a structural resemblance between the representing and represented systems.

My analysis shows that colour categorisation ANNs performing the same task developed robust and consistent structures across their corresponding collective sets of hidden layer activation patterns and this included *unexploited* representational content. The representational structures developed during training were clearly task-dependent rather than being a simple compression of the input data. The structures were consistent across diversely configured groups of ANNs performing the same categorisation task even when they were trained using different datasets with varying but overlapping categories. The representational structures also matched independent characterisations of the structures of the corresponding task domains. The results show that structural-resemblance approaches provide a more comprehensive explanation of the operation of these ANNs than clustering or mutual information based theories.

3.1 Three approaches to characterising representational content

3.1.1 Content individuated by clusters of activation points

The operation of neural networks can be explained in terms of task appropriate transformations of internal representational states, however, the representational entities or vehicles cannot be directly tied to particular patterns of connectivity and activation that are characteristic of specific networks.[2] To understand how ANNs operate requires insight into the common properties that allow networks with diverse architectures, connection weights and patterns of activity to perform the same task and generalise to correctly categorise novel input samples. This includes determining the type of representational vehicles that are utilised and how their content is determined. Shea (2007) provides a number of desiderata for evaluating representational explanations of the operation of ANNs stating that they should '(i) capture some underlying property of a network's mechanism of operation by which it performs its task; (ii) abstract away from individual weight matrices and particular patterns of activation; (iii) be such that it may be shared by different networks trained on the same task; and (iv) form part of an explanation of the network's ability to project its correct performance to new samples outside the training set.' (Shea, 2007, p250)

Shea (2007) claims that the vehicles of representational content in connectionist neural network models are regions of activation space that correspond to clusters of activation patterns generated from correctly classified input samples. He explains that these clusters can be individuated by analysing the collection of hidden layer activation points for the task domain. The regions of the activation space associated with clusters of activation patterns can be determined by identifying collections of nearby points which are relatively distant from other points in the space. There are various methods available for assigning groupings based on relative proximity and Shea (2007) acknowledges that there are issues with determining the best approach and how boundaries should be characterised. However, he explains that a measure of proximity that is relative to the dimensions of the particular space allows points to be grouped together based on that proximity. These clusters or regions of activation space are the vehicles of representational content. When a neural network is presented with an input it activates a hidden layer cluster which in turn activates an output layer cluster (Shea, 2007, p252).

Shea (2007,p252) explains that different ANNs can have clusters with the same representational content because a cluster in one network can be activated by the same set of inputs as a cluster in another network. This does not depend on the number of hidden units, so activation spaces with varying dimensions can contain the same clusters. The clustering property can also be shared by ANNs with varying connection weights and activation values. Shea (2007,p253) claims that neural networks can have some clusters in common without requiring that all clusters in the corresponding activation spaces match. A cluster in one network may correspond to a cluster in another different network due to their shared activation by the same set of inputs without the networks having any other corresponding clusters. This suggests that the representational content of particular clusters

[2] For a discussion of this issue refer to Sections 1.2, 2.1, 2.2.

can be determined and compared independently of any relationship there may be between clusters in the same activation space.

Figure 3.1 shows Shea's (2007, p252) example of the clustering of activation points in an ANN with two hidden units. The activation level of Hidden Unit 1 is displayed on the horizontal axis and the activation level of Hidden Unit 2 on the vertical axis. Each point corresponds to the distributed activation across the two hidden layer units for a particular input. In this example the hidden layer points generated from the input set form two distinct clusters with a clear and substantial separation.

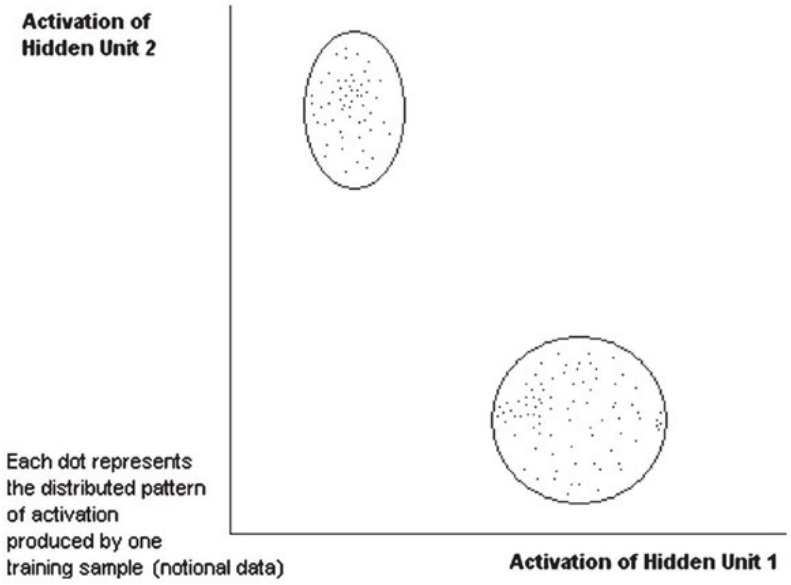


Figure 3.1. An example of the clustering of activation points in an ANN with two hidden units that determines two categorical distinctions (Shea, 2007, p252).

Shea (2007) explains that representational content should be ascribed to hidden layer clusters because they are the basis of the ability of trained ANNs to generalise correct performance to novel inputs. The input layer encoding of new data samples is different to any activation pattern encountered during training and may not even be linearly separable from other activation patterns that correspond to different output classifications. However, novel inputs can still produce hidden layer activation patterns that belong to the same clusters formed during training and in this sense the network can be described as performing the same operations on the new input data as on previously encountered samples from the training set. When a novel input generates hidden layer activation values that are within an existing cluster the network is representing the new sample as having properties in common with the training samples that are relevant to the output task. Activating the correct hidden layer cluster is a required intermediate step in transforming the input activation patterns into the correct output classifications. The hidden layer clusters represent task-dependent properties of the input samples that are relevant to the output classification. The hidden layer clusters may track a range of different properties that facilitate correct output classification including properties that the network is not explicitly trained to recognise at the

output layer (Shea, 2007, pp254-256). Shea (2007, p257) also states that this approach 'describes the content that should be ascribed to cluster, but does not attempt to capture the factors in virtue of which clusters have those contents'.

Shea (2018) also mentions neural network modelling in more recent discussions of representation in cognitive science but does not provide a more detailed or specific discussion of the clustering approach described here. There is a discussion of feedforward hierarchical processing (Shea, 2018, pp91-93) using the ALCOVE neural network model, however, this is a specific model that has some fundamental differences from the standard feedforward neural network models discussed in this chapter.

3.1.2 Content individuated by polytopes

Azhar (2016) claims that the vehicles of content in feedforward neural networks are geometric regions of hidden layer activation space referred to as polytopes. The regions are defined by specific output values they generate and the content is determined by the input class with which they share the highest Mutual Information. Azhar (2016) describes this approach using a simple hypothetical example ANN and then applies it to three groups of networks.

Azhar (2016) explains how polytopes can be determined in a three-layer feedforward ANN (depicted in Figure 3.2) consisting of an input layer, two hidden units and a single output unit that has been trained to distinguish between two input classes. The activation of the output unit (y) is determined by summing the weighted incoming hidden unit activations (h1*w1 + h2*w2) and the bias term (b) and applying a sigmoidal (tanh) activation function to generate an output value between -1 and +1. Input class one (x1) is indicated by positive output unit activation values (y) above a designated threshold value (T1) and input class two (x2) by negative values below a threshold value (T2).

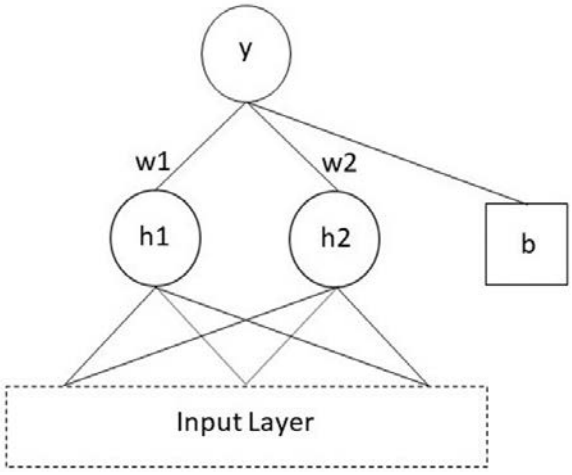


Figure 3.2. Azhar (2016) used a simple example ANN with two hidden units to explain how content is associated with regions in hidden layer activation space.

The activation level of the output unit and corresponding interpretation can be described by the following equations:

y = tanh(w1*h1 + w2*h2 + b) < T1 => Input belongs to input class one (x1)

y = tanh(w1*h1 + w2*h2 + b) >= T2 => Input belongs to input class two (x2)

The equations provide a partitioning of hidden layer activation space into decision regions that correspond with the output classifications. Output values between T1 and T2 indicate that the input does not belong to either input class but if T1 = T2 then the entire hidden layer activation space maps to a distinct output classification. An example of the separation of hidden layer activation space into two distinct decision regions in the case when T1 = T2 = 0 is provided in Figure 3.3. Azhar (2016) refers to these geometric spatial regions as polytopes. Polytopes are demarcated by the hyperplanes that divide activation space into decision regions that correspond to specific output classifications. ANNs with multiple output units would have overlapping sets of polytopes that segment the activation space into distinct regions which are also polytopes. 'In a sense, these decision regions are clusters of points in the activation space of the hidden layer that reliably signal the occurrence of a specific class of input into the network.' (Azhar, 2016, p702)

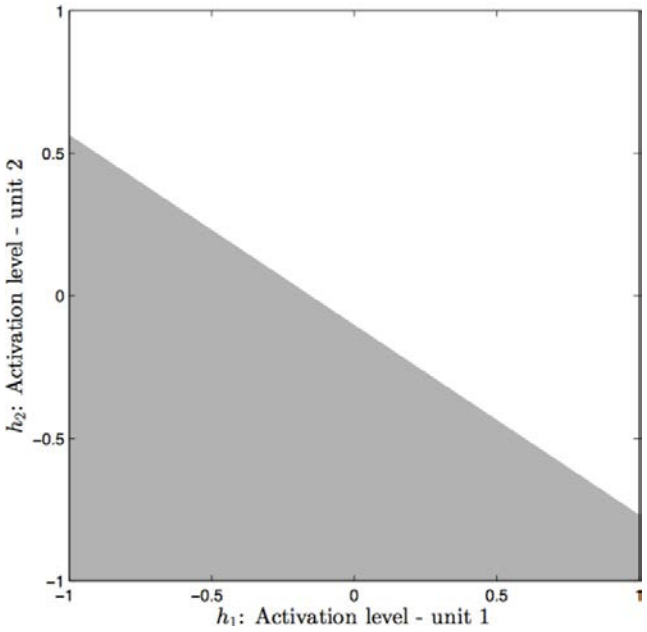


Figure 3.3. Decision regions in the hidden layer activation space of the example ANN where $w1$=2, $w2$=3, b=0.3 and T1 = T2 = 0. The equations defining the output activation and interpretation separate the space into two distinct regions corresponding to the input classes. The grey region is interpreted as being activated by class x1 and the white region by x2.

Azhar (2016) then introduces the information theoretic notion of mutual information in order to provide a robust and quantitate method for ascribing content to polytopes. This method can generalise to cases where the hidden layer activation points generated from a particular input class are not all linearly separable or located within the decision region that corresponds to their correct output classification. The approach is based on a statistical

method for content determination described by Usher (2001). Mutual information quantifies the dependency between two random variables and provides a measure of the amount of information that one variable provides about the other. For a network with N input classes (X = $x_1,x_2,\ldots,x_N$) and M polytopes (Y = $y_1,y_2,\ldots,y_M$, where M >=N) the mutual information between the input set (X) and the corresponding polytopes (Y) is defined as:

$$I(X;Y) = \sum_{n=1}^{N}\sum_{m=1}^{M} p(x_n, y_m)\log_2 \frac{p(x_n, y_m)}{p_X(x_n)p_Y(y_m)}$$

(Azhar, 2016, p702)

Azhar (2016) describes how point-wise mutual information can be used to quantify the amount of information that each individual polytope provides about the input classes and it is defined as:

$$\mathcal{M}(x_n, y_m) := p(x_n, y_m)\log_2 \frac{p(x_n, y_m)}{p_X(x_n)p_Y(y_m)}$$

(Azhar, 2016, p703)

The pointwise mutual information between a specific input class (x_n) and specific polytope (y_m) is equal to the joint probability of the input and polytope occurring together multiplied by the base two log of that joint probability divided by the product of the probability that the specific input occurs in the input set and the probability that the particular polytope is activated. The informational content of a polytope is determined by the input class with which it shares the highest pointwise mutual information. This is the input class with the highest contribution to the overall mutual information and indicates it is more strongly associated with the particular polytope than any of the other input classes.

Azhar (2016) applies this method to a hypothetical ANN that distinguishes between two different input classes from among ten specific inputs. The hidden layer activation space is separated into two decision regions corresponding to the output classifications as shown in Figure 3.4. The red triangles depict hidden layer activation points generated from class x1 and the green circles are points from x2. One of the inputs from x2 is located in the decision region that generates output values associated with inputs from x1 and so it will be misclassified.

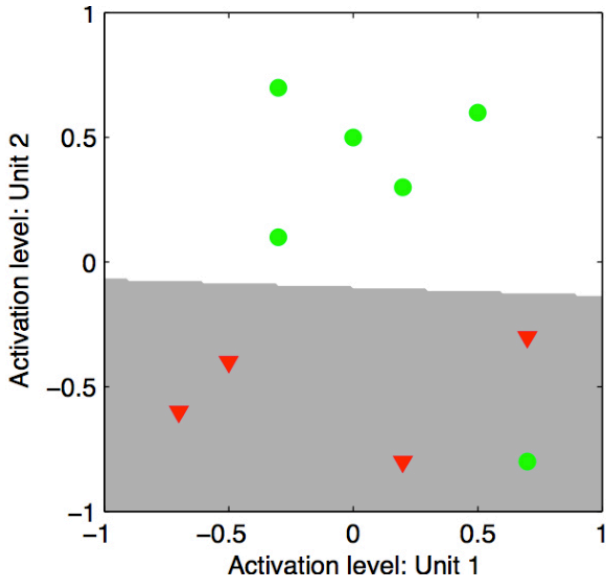


Figure 3.4. The hidden layer activation space of a hypothetical ANN trained to distinguish between two input classes (x_1 and x_2). The red triangles depict hidden layer activation points generated from class x_1 and the green circles are from x_2. The grey region is the polytope associated with input class x_1 and the white region with input class x_2.

In this case the polytopes do not cleanly separate the inputs from the two different classes. However, the informational content of the polytopes can be quantitatively determined by calculating the point-wise mutual information between all pairs of input classes and polytopes.

The joint probability distributions are:

$p(x_n, y_m)$	y_1	y_2
x_1	0.4	0
x_2	0.1	0.5

And the point-wise mutual information values are:

$\mathcal{M}(x_n, y_m)$	y_1	y_2
x_1	0.400	0
x_2	-0.158	0.368

Azhar (2016, p703) states that for any polytope y_m, its informational content is "input class x_n obtained in the network" where this input class is the one that has the maximal contribution to the mutual information. Calculating the point-wise mutual information for all pairs of polytopes and input classes shows that 'the grey polytope (y_1) has the content that red triangles (x_1) occurred in the network, while the white polytope (y_2) has the content that the green circles (x_2) occurred in the network' (Azhar, 2016, p704).

The content of a polytope is determined by the input class that it shares the highest mutual information with. Azhar (2016) applied this approach to determine the polytopes regions and their content in three distinct sets of ANNs implemented with varying architectures. This

showed that content could be robustly ascribed to networks with hidden layer activation points that are misclassified as well as networks where the input classes were neatly separated by the hidden layer polytopes.

3.1.3 Content determined by structural resemblance

O'Brien and Opie (2006) claim that the distributed patterns of hidden layer activation present during processing in connectionist neural networks model actual cognitive representational vehicles and when considered collectively they structurally resemble relevant aspects of the task domain. Additionally, they claim that the fixed patterns of connection weights which govern the information processing or computation in the networks is modelling neural connectivity that also has relevant structural resemblance relations to the task domain. This chapter focuses on analysing and comparing explanations of representation in the hidden layer activation space of ANNs. Connection weight representation will be discussed in Chapter 4.

Objects and collections of objects can resemblance each other in various different ways and to different degrees. The most obvious type of resemblance involves sharing one or more physical properties, for example colour, shape, size or weight. This is a first-order resemblance relationship. Distributed patterns of neural activation modelled by ANNs do not share these types of physical properties with the objects they represent. However, the relations between collective groups of activation values associated with a specific task domain may resemble or reflect a relevant set of relations between the objects that they represent. This can be described as a second order resemblance relationship and is an abstract and formal relation rather than a physical one (O'Brien & Opie, 2006, pp34-36).

O'Brien and Opie (2006, p36) explain that a 'system structurally resembles another when the physical relations among the objects that comprise the first preserve some aspects of the relational organisation of the objects that comprise the second'. For example, the spatial relations between points on a standard map systematically reflect spatial relations between locations in the world. Consider a map with three points that designate cities A, B and C. If cities A and B are closer together than cities A and C in the world then they will also be closer together on the map, and similarly, if city B is located between A and C in the world then it will also be located between A and C on the map. In this case the actual distance relations are preserved by distance relations on the map.

Additionally, because structural resemblance is a second order resemblance relation it does not depend on systems having the same type of physical relations between their objects. For example, geometric relations on a weather map can be used to represent relations between atmospheric pressure. Lines on a weather map referred to as isobars represent contiguous regions of equal atmospheric pressure and the spacing of these lines can be used to determine pressure gradients. The relative spacing and size of isobars corresponds to the relative size and direction of pressure gradients. So, there is a second order structural resemblance between the geometric layout of isobars on the map and the pressure gradients in the region being represented (O'Brien & Opie, 2004, p8).

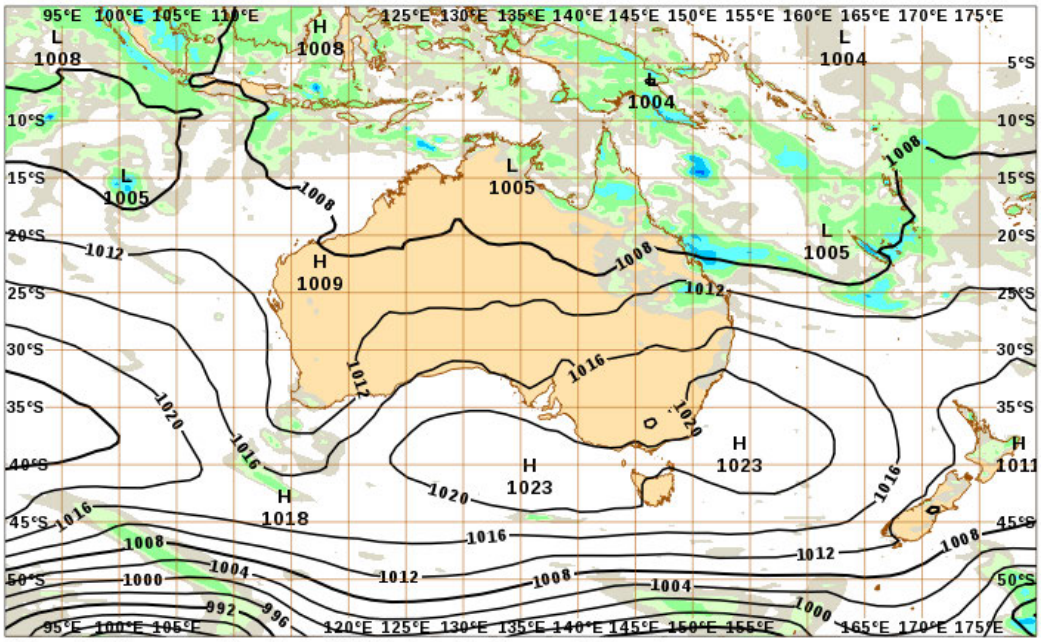


Figure 3.5. Lines referred to as isobars represent pressure gradients on a weather map.

O'Brien and Opie (2004, p8) also provide a more formal definition of second order resemblance using set theoretic notation by considering a system S containing V objects and R_V relations (they could be spatial, causal, structural or inferential relations) on those objects where $S_V = (V,R_V)$. 'There is a second order resemblance between two systems $S_V = (V,R_V)$ and $S_O = (O,R_O)$ if for at least some objects in V and relations in R_V there is a one-to-one mapping from V to O and from R_V to R_O such that when a relation R_V holds of objects in V, the corresponding relation in R_O holds of objects in O'. Structural resemblance is determined by the abstract relational organisation of the systems and does not depend on systems sharing the same first-order properties.

O'Brien and Opie (2004, p9) acknowledge that this is a relatively weak characterisation of second-order resemblance and can be strengthened. If every object in system V can be mapped onto an object in O while preserving all the relations defined on V the systems are said to be homomorphic. The resemblance is strengthened even further if there is a one-to-one mapping between objects in V and objects in O. This strong form of second order resemblance is referred to as isomorphism and describes systems with identical relational organisation. O'Brien and Opie (2004, p9) explain that discussions of second order resemblance in the literature often focus on isomorphism (for example Cummins, 1996, pp85-111) but suggest that this characterisation is too restrictive. The analysis of structural explanations of representation in neural networks provided in this chapter will focus on determining degrees of structural similarity rather than investigating or defining isomorphisms.

O'Brien and Opie (2004, p11) use the face-recognition ANN described by Churchland (1996) that was discussed in Chapter 2 (for details of the ANN refer to Section 2.2) as an example of how structural resemblance can be used to explain representation in connectionist neural network models. O'Brien and Opie (2004, p12) highlight the significance of the way the

network partitions hidden layer activation space into regions corresponding to the classification task. They claim that cluster analysis reveals there is a similarity between the structure of the activation space and the domain of faces.

O'Brien and Opie (2004) explain that different images of the same face are much more similar to each other than to any other facial images and they are represented by points in hidden layer activation space that are relatively close together. The similarities or differences between faces of different individuals are also reflected in the relative proximities of their respective hidden layer activation patterns. The hidden layer activation space is an abstract mathematical space used to characterise hidden layer activation. The activation patterns simulated in connectionist networks model actual physical patterns of neural activation. The distance relations in activation space model actual physical relations between patterns of neural activation. The activation patterns are representational vehicles and the physical relations between the vehicles mirror (some of) the relevant relations in the task domain. O'Brien and Opie (2004) suggest that the structural resemblance between the collective set of hidden layer activation patterns and the task domain may be what powers both the computational and behavioural capacities of neural networks. Neural networks are analogue computational systems and the intrinsic properties of their representational vehicles determine how they are processed. The semantic relations are determined by the physical relations between representational vehicles (O'Brien & Opie (2004, p12).

O'Brien and Opie (2006) also refer to examples of colour categorisation ANNs developed by Laakso and Cottrell (2000) and used by Churchland (1998) to show structural similarity in hidden layer activation spaces across ANNs with diverse configurations. O'Brien and Opie (2006, p34) trained a series of colour recognition ANNs on the same dataset and reported that they all partitioned their hidden layer activation space into linearly separable regions and activation points from particular colour categories were located in distinct regions. The activation space from one of the ANNs is shown in Figure 3.6. They explain that it is widely agreed that appropriate partitioning of activation space facilitates the correct categorisation of input data by trained networks but claim that when the hidden layer activation patterns are considered collectively they structurally resemble relevant aspects of the task domain. According to O'Brien and Opie (2006, p36) 'the set of hidden unit activation patterns generated across any trained-up connectionist network constitutes a system of representing vehicles whose physical relations sustain a second order resemblance relation with respect to the task domain over which the network has been trained'.

O'Brien and Opie (2004) claim that in an appropriately trained connectionist neural network model the hidden layer activation space is partitioned in a task relevant manner. Similarities and differences between activation patterns can be characterised mathematically using distance relations between activation points and these relations reflect actual similarities and differences between relevant aspects of the task domain. The collection of hidden layer activation patterns associated with the input domain of a connectionist neural network have a structural resemblance relation to the relevant properties or aspects of the corresponding task domain.

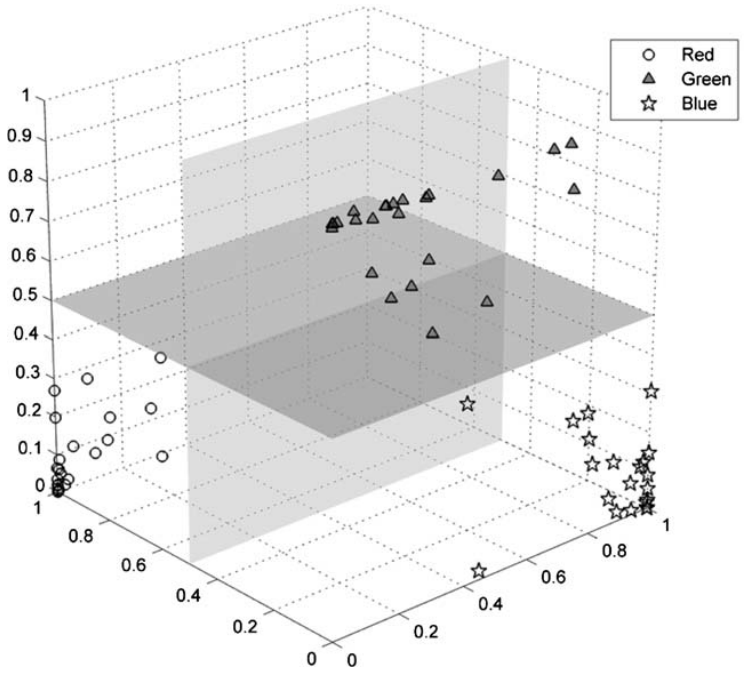


Figure 3.6. Partitioning of hidden layer activation space into linearly separable regions for a three colour categorisation task (O'Brien & Opie, 2006, p35).

O'Brien and Opie's (2006) explanation appears to be compatible with Churchland's (1996,1998,2007,2012) discussion of representation in connectionist neural networks that was described in Chapter 2. They use and build on some common examples (including Cottrell's facial recognition network and colour categorisation networks) but define the role of structural resemblance explicitly. Churchland has made similar claims that the hidden layer activation spaces of successfully trained neural networks are partitioned into regions that reflect relevant aspects of the task domain. Similarities and differences between neural activation patterns can be characterised by relative proximities in the activation space of connectionist models and correspond to similarities and differences between the content of the mental representations being instantiated.

Churchland (1996,1998,2007) also emphasises the significance of the relative locations of characteristic or prototypical activation points for each category or task relevant property that a neural network is trained to distinguish between. The relevant similarities and differences between categories, concepts or aspects of the task domain are systematically reflected by similarities and differences between the relative proximity of corresponding prototype activation points for those categories. Additionally, the deviation of activation points from their corresponding prototype may reflect the degree to which their representational content varies from typical examples of that particular category or type.

3.2 Differences between the various approaches to identifying and individuating content

I will now compare the various approaches to explaining representation in neural networks to identify the significant consistencies and differences. The positions described by Shea (2007), Azhar (2016), Churchland (1996,1998,2007,2012) and O'Brien and Opie (2004,2006) all consider how activation patterns are distributed in hidden layer activation space. The approaches vary but all these positions describe the hidden layer activation space in appropriately trained ANNs as having partitions or regions which are relevant to correct output classification and network performance.

The partitioning described by Azhar (2016) divides the space into polytopes that are quantitatively determined by the configuration of the ANN from the hidden layer to the output layer. The regions are directly determined from the equations governing the transformation of activation values to specific output classifications. O'Brien and Opie (2004) and Churchland (1996) describe how cluster analysis of the collective set of hidden layer activation patterns reveals a task relevant partitioning of activation space. The polytope regions proposed by Azhar appear to be generally consistent with the partitioning described by Churchland (1996) and O'Brien and Opie (2004) but the method provides exact boundaries that are defined by output classification. Shea (2007) proposes that the regions of interest are clusters of activation points corresponding to correctly classified inputs with boundaries determined by some sort of relevant proximity measures. In order for the points in these clusters to be correctly classified the cluster would need to be a sub-region of the corresponding polytope.

Shea (2007) ascribes representational content to entire clusters and each cluster has a specific content. Similarly, Azhar (2016) ascribes a specific content to each polytope region. He refers to informational content rather than representational content but compares his approach to other prominent theories of representational content including Shea (2007), Churchland (1989,1996,1998,2012) and O'Brien and Opie (2006). Both of these approaches may be problematic if representational content is claimed to be causally efficacious rather than merely explanatory because neural networks do not process or transform regions such as clusters or polytopes – they transform specific activation patterns.[3] However, this issue could be overcome by interpreting the ascription of content to a region as claiming that all activation points belonging to that region are representational vehicles that have exactly the same content. The networks are sensitive to regions in activation space in the sense that all activation points in a polytope or cluster are transformed in a way that leads to the same output classification. However, the descriptions provided by Shea (2007) and Azhar (2016) do state that content can be ascribed to the clusters or regions themselves rather than to all points within them.

The partitioning of hidden layer activation space described by Churchland (1996) and O'Brien and Opie (2004) is significantly different from the regional distinctions described by both Shea (2007) and Azhar (2016) because it allows for activation points in the same region to have (slightly) different contents. The relative proximities of activation points located in the same partition can correspond to variations in aspects or properties of the category being represented. Churchland (1996) explains that prototypical activation points can be

[3] Thanks to Dr Jon Opie for highlighting this point.

determined for regions that are exemplars of the corresponding categories and proximity to the prototypes reflects the relative similarity or difference to the category exemplar. Similarly, points on the boundary of a partition represent objects or properties of the task domain that are indeterminate or ambiguous with regard to the classifications corresponding to the regions that are adjacent to the boundary.

However, Churchland's (1996) description of how content varies across activation spaces is somewhat idealised. Training ANNs can lead to hidden layer activation points being pushed to the boundaries or vertices of the activation space because the network's performance is determined by the development of task appropriate partitioning at that layer. Laakso and Cottrell (2006,p132) observed that Churchland may depict prototypes as central points in corresponding regions but that this is not consistent with how hidden layer activation points are distributed in actual ANNs trained using backpropogation. They provided an example of the distribution of hidden layer activation points from an AAN distinguishing between two categories to highlight this issue (Figure 3.7).

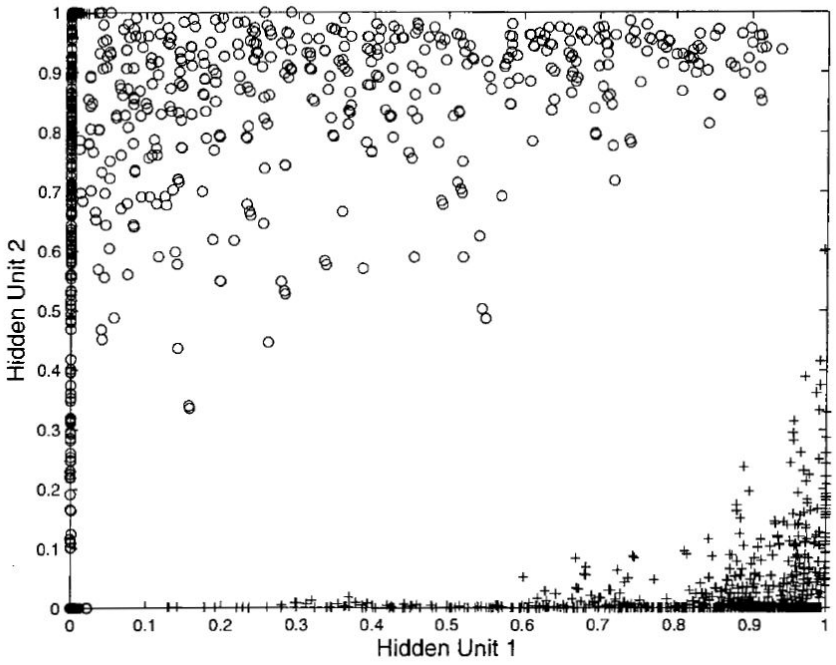


Figure 3.7. Laakso and Cottrell's (2006, p132) example of the distribution of hidden layer activation points in an ANN distinguishing between two categories.

The diagrammatic example of hidden layer clusters provided by Shea (2007, p252) is also idealised. Clusters may not always be easily defined by a proximity metric and there may not be a substantial separation between some points in neighbouring clusters. Figure 3.8 shows the actual distribution of activation points in four different ANNs each trained to categorise five colours (red, yellow, green, blue, purple). I used ANNs with two and three hidden units to allow visualisation of the hidden unit activations. The graphs show that although the activation space can be divided into regions that correspond to the specific categorical

distinctions being performed the activation values still tend to the corners or edges of activation space and are not uniformly distributed.

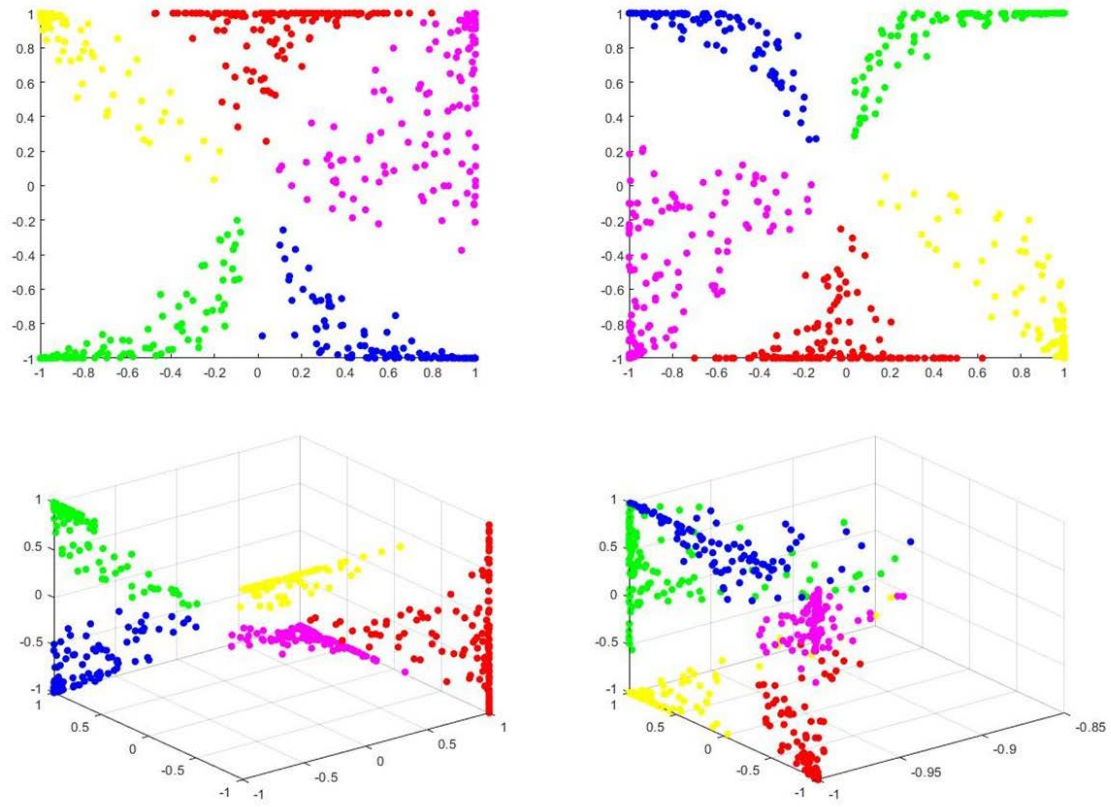


Figure 3.8. Distribution of hidden layer activation points from novel ANNs trained to classify 5 colours using 2 hidden units (2 networks) and 3 hidden units (2 networks).

The polytope regions described by Azhar (2016) are determined directly by the processing properties of the ANNs so they will always correspond to the categorical distinctions being performed. However, the hidden layer activation points generated from the input domain may only occupy a small sub region of their respective polytopes. This is evident from some of the example activation spaces provided by Azhar (2016) and is shown in Figure 3.9.

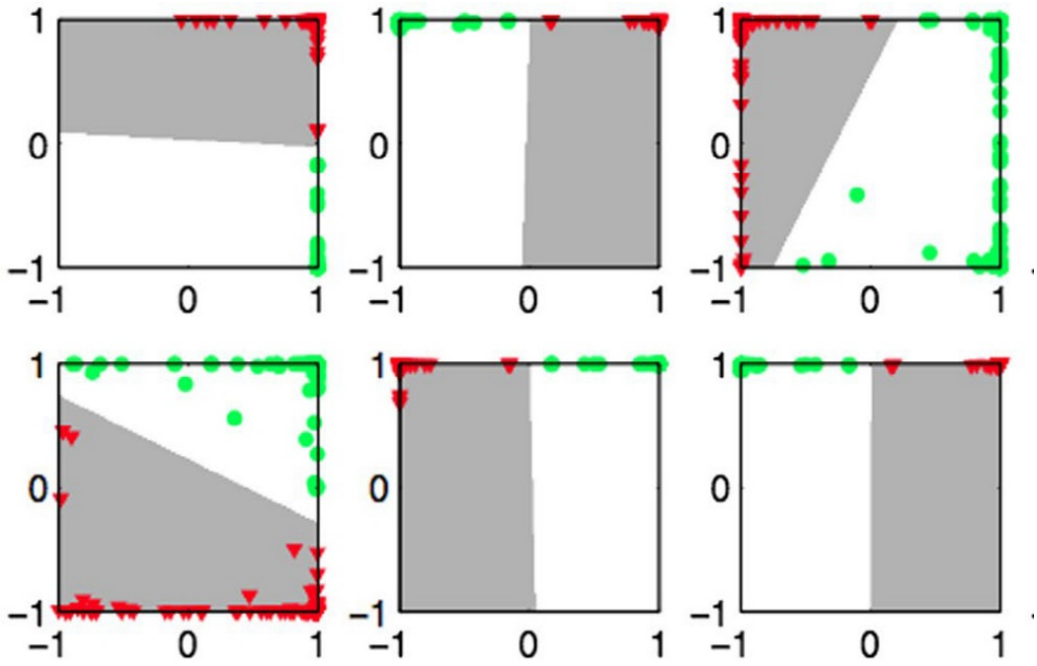


Figure 3.9. The example activation spaces provided by Azhar (2016, p705) also show that the activation points in a polytope may be located in small subregions of the region.

The application of different training methods or parameters may affect the resulting distributions of activation points. Further investigation is required to establish how the relations between activation states in biological neural networks compare to the activation spaces of ANNs and whether the training needs to be more tightly constrained. Artificial input layer activation values can be created that generate hidden layer activation points spanning the entire hidden layer activation space. The following examples in Figure 3.10 show the collections of hidden layer activation points corresponding to random inputs transformed by an ANN that was trained to classify reflectance spectra into five colours. The points do appear to span the entire activation space but they are more heavily concentrated at the boundary regions (note that a sigmoid activation function was used). However, the random inputs used in this example may not correspond to meaningful or actual input samples from any particular input domain.

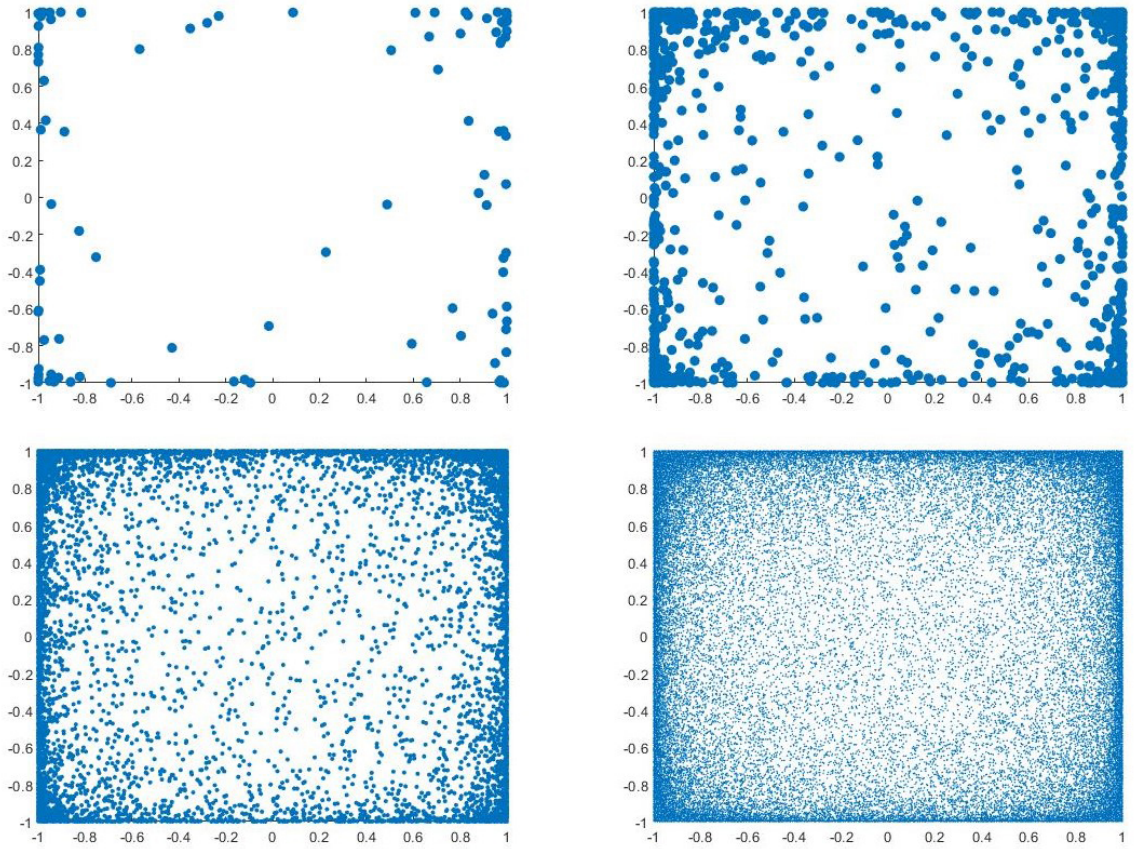


Figure 3.10. Examples of the distribution of hidden layer activation points generated by a network trained to categorise colours using two hidden units with random input sets of increasing size comprised of 100, 1000, 10000 and 100000 input patterns.

A key difference between the varying approaches to determining and individuating content in connectionist neural network models concerns the relations between regions or groups of activation points in hidden layer space. Churchland (1998) explains that the relative locations of prototypical activation points corresponding to category exemplars reflects how a network responds to similarities and differences between task relevant aspects of the categories being represented. The relative locations of prototype activation points reflect task relevant relations between the categories. The representational similarity between two ANNs can be assessed by comparing the structure or relative proximities of prototypical activation points in the hidden layer. O'Brien and Opie (2006, p36) claim that hidden unit activation patterns are representational vehicles and when considered collectively this system of vehicles structurally resembles aspects of the task domain.

The descriptions offered by Shea (2007) and Azhar (2016) do not take structural similarity to be a relevant factor in content determination. Shea (2007, p262) states that a cluster's relation to other clusters does not determine its content. It is possible for networks to have hidden layer clusters with the same content but with the clusters arranged differently. Shea (2007, p256) explains that hidden layer clusters only have content in virtue of the contents ascribed to the output layer. Azhar (2016) explains that the contents of polytope regions are

determined by their informational relation to the input classes. The relations between hidden layer activation points or regions are not relevant to content determination.

3.3 Empirical analysis of colour categorisation ANNs

I conducted an original empirical investigation to determine the plausibility of the various approaches to describing representation in neural networks. This included comparing the hidden layer activation spaces of a wide range of novel ANNs to determine whether they develop task relevant structural relations between activation patterns. This builds on the empirical investigation described in Chapter 2 and also includes comparisons between networks trained to perform different tasks using the same dataset. Details of the configuration and training of the ANNs are provided in Appendix 3.1.

I created six groups of three-layer feedforward ANNs using two different but related reflectance spectra datasets. Both datasets are comprised of reflectance spectra measured from colour chips classified using the Munsell colour system. This system aims to provide a structured and consistent method for describing human colour judgements. The Munsell colour system is a three-dimensional colour space that specifies colours based on hue (which is often referred to as colour), chroma (the saturation or intensity of the colour), and value (the lightness, brightness or tone of the colour). Figure 3.11 shows how colours are categorised using the Munsell colour system. Each colour is assigned one of ten possible hues (there are five primary and five intermediate hues) along with a lightness value between zero and ten and a saturation value between zero and twelve. The example shows a wedge of the colour space corresponding to colours with an intermediate hue between Purple and Blue, a lightness value of five and saturations extending from zero (completely unsaturated) to twelve. Figure 3.12 shows the distribution of colours over the three-dimensional Munsell colour space.

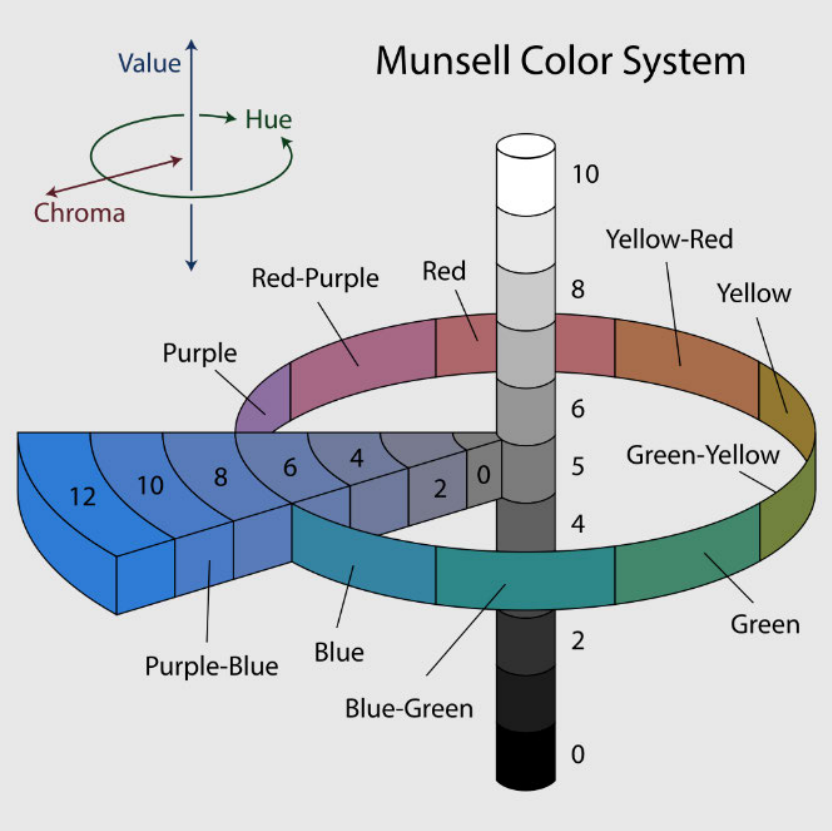


Figure 3.11. The Munsell colour system can be depicted as a three-dimensional colour space. Colours are specified by their hue, chroma and value. This example shows the purple-blue hue with an intermediate lightness value of five and increasing saturations from zero to twelve. (https://upload.wikimedia.org/wikipedia/commons/thumb/d/d5/Munsell-system.svg/1024px-Munsell-system.svg.png)

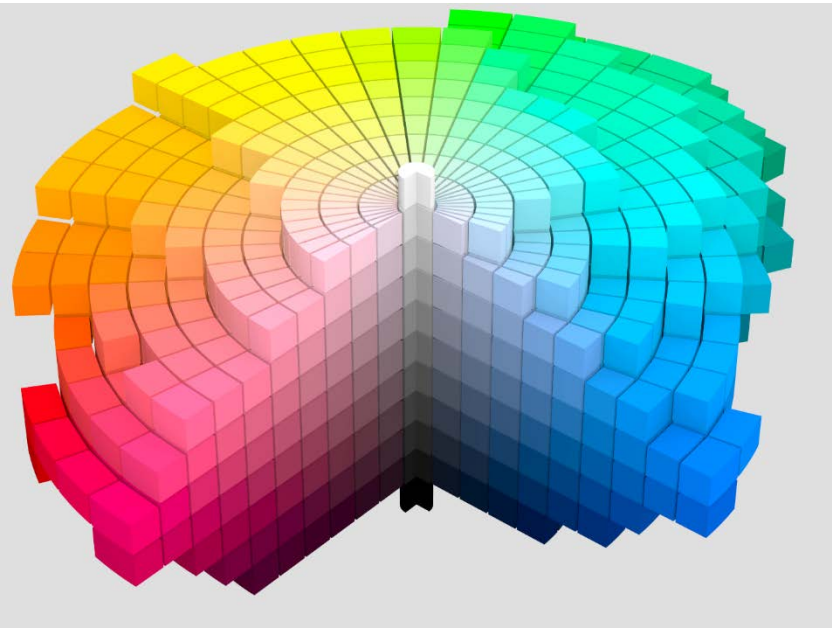


Figure 3.12. Arrangement of colours in the Munsell colour space. (https://en.wikipedia.org/wiki/Munsell_color_system#/media/File:Munsell_1943_color_solid_cylindrical_coordinates_gray.png)

The first dataset that I used to train three of the six groups of ANNs is the same Munsell colour dataset used by Laakso and Cottrell (2000). This dataset is comprised of reflectance spectra measured from 627 colour chips including the five primary hues (red, yellow, green, blue, purple) with varying chroma (1,2,4,6,8,10,12) and lightness values (2.5,3,4,5,6,7,8,9). The spectral intensity is measured from 400nm to 700nm at 5nm intervals. Each input sample is a 61-dimensional vector with the first component corresponding to the reflectance intensity at the wavelength 400nm, the second at 405nm and the final 61st component corresponding to the reflectance intensity at the wavelength 700nm. The original values of each component were between 0 and 4096 but they have been normalised to values ranging between zero and one. The reflectance spectra for each of the 627 input samples is represented as a 61-dimensional vector where each element has a value between zero and one with a precision of 12 decimal places. An example of the encoding of a red input sample is provided in Figure 3.13.

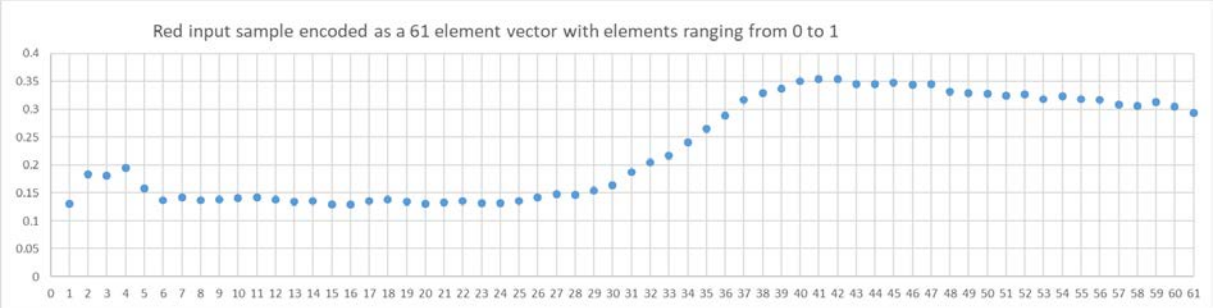


Figure 3.13. Encoding of a red input sample (with value = 5, chroma = 6) in the first dataset. The wavelength profile is encoded as a 61-element vector with values ranging from 0 to 1.

Laakso and Cottrell (2000) used this dataset to train a variety of three-layer feedforward ANNs to classify the five primary hues. Similarly, in this investigation a series of 120 ANNs were trained to classify the five hues. The series included ANNs ranging from two hidden units increasing sequentially up to 25 hidden units (24 different network architectures) with five networks using different initial parameters created for each architecture (24 X 5 = 120 ANNs in the group). The ANNs performed with very high accuracy after training and could correctly classify an average of 99.9% of the full dataset. The networks also performed almost as well when classifying the novel input samples withheld in the test set achieving an average accuracy of 99.5%.

To provide evidence of the development of task-dependent structured systems of representational vehicles in neural network models the simulations were extended to cover two additional task domains based on the same dataset. Another series of 120 ANNs were trained to classify the 627 reflectance spectra into the eight values that specify the lightness component of the colours. Like the hue categorisation ANNs, this series included networks ranging from two hidden units increasing sequentially up to 25 hidden units with five different networks created for each architecture (24 X 5 = 120 ANNs in the group). The networks

performed with very high accuracy after training and could correctly classify an average of 99.0% of samples from the full dataset and 98.2% from the test set.

Finally, a series of 105 ANNs were trained to classify the 627 reflectance spectra into the seven chroma that specify the saturation component of the colours. The series included ANNs ranging from five hidden units increasing sequentially up to 25 hidden units with five different networks created for each architecture (21 X 5 = 105 ANNs in this group). ANNs with two, three and four hidden units were not included for this task as it was not possible to train networks with adequate performance. This series of trained ANNs did not perform as accurately as the hue and value categorisation networks. They achieved an average accuracy of 89.5% on the full dataset and 80.9% on the novel test set.

The second dataset that I used also contained reflectance spectra measured from Munsell colour chips but included more samples with increased resolution. The dataset is comprised of reflectance spectra corresponding to the same five primary hues as the first dataset but also includes an additional five intermediate hues (red, red-purple, purple, purple-blue, blue, blue-green, green, green-yellow, yellow, yellow-red). The reflectance spectra covers the same range of chroma (1,2,4,6,8,10,12) and lightness values (2.5,3,4,5,6,7,8,9) as the first dataset. The spectral intensity was measured from 380nm to 800nm at 1nm intervals resulting in each input sample having 421 components rather than 61. The original values had been normalised and the reflectance spectra for each of the 1269 input samples is represented as a 421-dimensional vector with each element having a value between zero and one with a precision of four decimal places. An example of the encoding of a red input sample is provided in Figure 3.14.

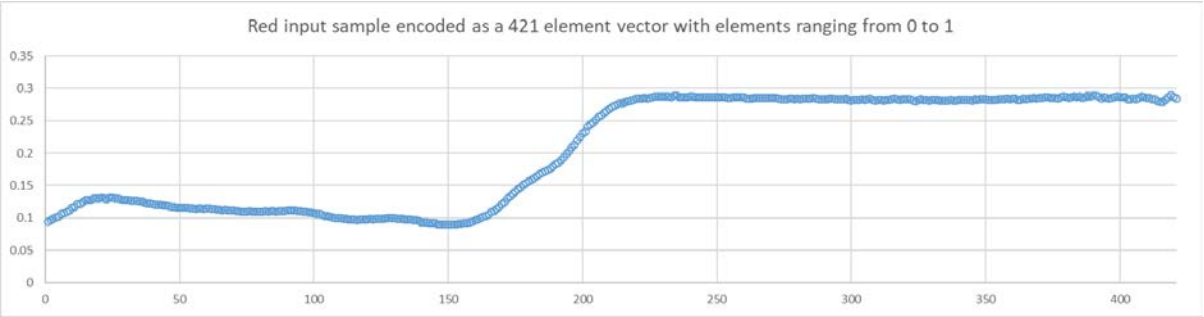


Figure 3.14. Encoding of a red input sample (with value = 5, chroma = 6) in the second dataset. The wavelength profile is encoded as a 421-element vector with values ranging from 0 to 1.

I trained another three groups of ANNs using the second dataset with the same three task domains that were applied to the first dataset. Each group of ANNs included a series of networks ranging sequentially from five to 25 hidden units with five networks of each architecture created using different initial parameters (21 X 5 = 105 ANNs). This produced 105 ANNs that determine the hue associated with the reflectance spectra using both primary and intermediate hues (10 hues), 105 ANNs that determine lightness values and 105 ANNs that determine chroma. The performance of these ANNs was very similar to the networks trained using the smaller and lower resolution reflectance spectra dataset. The six groups of colour categorisation ANNs developed for this investigation are summarised in Table 3.1.

ANN Groups	1	2	3
Input Data Set	Munsell Colour - 61 components, 627 samples		
Categorisation Task	5 Hues	8 Values	7 Chromas
Categories	(R,Y,G,B,P)	(2.5,3,4,5,6,7,8,9)	(1,2,4,6,8,10,12)
Input Size	61	61	61
Hidden Layer Size Range (sequential)	2 - 25	2 - 25	5 - 25
Output Layer Size	5	8	7
Number of ANNs	120	120	105
Average Total Performance (% correct)	99.91	98.40	89.54
Average Test Set Performance (% correct)	99.44	97.41	80.91
ANN Groups	4	5	6
Input Data Set	Munsell Colour - 421 componets, 1269 samples		
Categorisation Task	10 Hues	8 Values	7 Chromas
Categories	(R,RY,Y,YG,G,GB,B,BP,P,PR)	(2.5,3,4,5,6,7,8,9)	(1,2,4,6,8,10,12)
Input Size	421	421	421
Hidden Layer Size Range (sequential)	5 - 25	5 - 25	5 - 25
Output Layer Size	10	8	7
Number of ANNs	105	105	105
Average Total Performance (% correct)	98.52	99.95	92.21
Average Test Set Performance (% correct)	96.88	99.84	87.40

Table 3.1. Summary of the six groups of ANNs including their network architectures, datasets, task domains and accuracy.

In the next section (3.3) I will describe how the structural similarity between the distributions of hidden layer activation patterns both within and across the groups of ANNs was analysed using the comparison methods introduced and developed in Chapter 2. This included comparing the relative proximities of hidden layer activation points generated from the entire input domain, comparing the arrangement of the prototypical activation points corresponding to each categorical distinction and comparing the distribution of activation points within corresponding categories.

3.3.1 Structural similarity between input and hidden layer activation spaces

I began the empirical analysis by assessing the structural similarity between the input and hidden layers of each individual ANN to determine whether the information processing restructures the input in a task relevant manner rather than just performing some type of simple recoding or data compression. The similarity between the input and hidden layer activation spaces was determined by calculating the correlation between the relative proximities of corresponding sets of activation points. This provides a quantitative measure of the similarity between the structural arrangements of the collections of activation patterns. A value of one indicates complete correlation, values close to one indicate a high correlation and zero indicates there is no correlation. The correlation is interpreted as a quantitative measure of the degree of structural similarity. The correlation was calculated for each ANN and then averaged for each of the six groups of ANNs to provide a measure of similarity between the input and hidden layer for each group of networks with the same input and target domains.

The similarity between the structural arrangements of activation values were assessed at three levels. Laakso and Cottrell's (2000) original method was used to provide a comparison of the relative locations of all activation points in the entire input domain. This approach was then extended to explicitly compare prototypical activation points for each category to determine the structural organisation of the corresponding categories. The arrangement of activation values within each specific category were also compared to determine whether similarities and differences between members of the same category were preserved.[4]

The average correlation between the relative proximities of all input layer activation points with the corresponding hidden layer activation points for each of the ANNs performing the hue categorisation task was fairly low (0.20,0.30) but this increased for the chroma categorisation networks (0.44,0.45) and was reasonably high for the networks trained to categorise value (0.76,0.69). The average correlation between the relative proximities of the input layer prototypical activation points (calculated by averaging the activations belonging to the same categories) with the corresponding hidden layer prototype activation points for each ANN increased for all six groups of ANNs and was high for all three tasks (0.76-0.88). There was a moderate correlation between the relative proximities of the input layer activation points within a specific category and the corresponding hidden layer activation points for each ANN (averaging between 0.37-0.57 across the six groups of ANNs). The correlation values for each group of ANNs are summarised in Table 3.2.

ANN training dataset		61 components, 627 samples			421 components, 1269 samples		
Categorisation task		5 Hues	8 Values	7 Chromas	10 Hues	8 Values	7 Chromas
Layer comparison	**Structural comparison**	Average Correlations			Average Correlations		
Input - Hidden	All activation points	0.30	0.76	0.44	0.20	0.69	0.45
Input - Hidden	Category prototypes	0.76	0.87	0.88	0.81	0.86	0.80
Input - Hidden	Intra-categories average	0.54	0.57	0.61	0.37	0.55	0.57

Table 3.2. Summary of the average correlations between the structural similarity of input and hidden layer activation spaces for each group of ANNs.

The low correlation values for the hue and chroma categorisation tasks show that there was a substantial task-dependent reorganisation of the structural arrangement of hidden layer activation patterns. This was not apparent for the value categorisation task. This may be because value is a less abstract relation that is directly tied to overall intensity of the input activation patterns. However, there is a high degree of structural similarity between the prototypical input and prototypical hidden layer activation for each of the three categorisation tasks. At this level of analysis the comparison imposes the task distinction on the input layer activation patterns as the prototypes are calculated with respect to the specific task. These

[4] Laakso and Cotrell's (2000) method for assessing similarity and the differences between comparing the entire input domain, the relations between category prototypes and the arrangement of activation points within corresponding categories is discussed in detail in Chapter 2.

results provide support for the claim that ANNs transform input patterns in a task-dependent manner and can extract abstract properties of the input.

3.3.2 Structural similarity between hidden layer activation spaces of distinct ANNs

I then investigated whether ANNs produce hidden layer activation patterns with a task-dependent structure that is preserved across groups of ANNs with diverse configurations of weights and varying architectures. If this is confirmed and the structural similarity or resemblance relations reflect task relevant aspects of the target domain then this would provide support for structural approaches to describing the operation of ANNs and for using representational rather than strictly causal explanations in general. This would also provide some empirical support for Gladziejewski and Milkowski's (2017) distinction between describing neurons as detectors with values that reflect causal variations in the input stimulus and groups of neurons that realise causally efficacious structural representations that model aspects of an entire target domain.

I analysed the six groups of colour categorisation ANNs to determine whether the structures of collective sets of corresponding hidden layer activation points were preserved across different networks performing the same task. The hidden layer activation space of each ANN was compared to every other network in the same group and the correlation values were then averaged to provide a measure of similarity between all ANNs with the same input and task domains. For the groups with 105 ANNs, 5460 (105 x 104 / 2 = 5460) pairwise comparisons were performed and then averaged for both the entire input domain and for the category prototypes. The structural arrangements of activation points within corresponding categories were also compared for each group of ANNs with the correlation values averaged across all the networks and categories for each of the six groups. The average correlation values for each group of ANNs provides a measure of the preservation of structural similarity across the hidden layer activation spaces of distinct networks trained to perform the same task.

The ANNs trained to classify five hues (colours) from the 61 component reflectance spectra had highly correlated hidden layer activation spaces determined by the relations between all activation points in the domain (0.91). The structural relations between the prototypical activation points corresponding to inter-categorical distinctions were also highly correlated (0.92). The structural relations between each group of activation points belonging to the same input category were also compared and the correlation values averaged for each pair of networks in the group. The structural similarity was not as high as with the full set of activation points, or the prototype activation points but there was still a quite significant correlation (0.66). The group of ANNs trained to classify value (lightness) had very highly correlated hidden layer activation point structures (0.96) and prototype activation point structures (0.97). The average correlation between the structures of corresponding categorically determined groups of points was lower but still significant (0.60). The group of ANNs trained to classify chroma (saturation) had strongly correlated hidden layer activation point structures (0.81) but substantially less than the networks trained to classify both hue and value. However, the prototype activation structures were still very highly correlated (0.98) and the average correlation between the structures of corresponding categorically

determined groups of points was strong and substantially higher than the hue and value classification networks (0.81). The average correlations between the ANNs trained on the larger 421 component reflectance spectra dataset were also very similar and consistent with these results. The average correlation values for all six groups of ANNs are summarised in Table 3.3.

ANN training dataset		61 components, 627 samples			421 components, 1269 samples		
Categorisation task		5 Hues	8 Values	7 Chromas	10 Hues	8 Values	7 Chromas
Layer comparison	**Structural comparison**	Average Correlations			Average Correlations		
Hidden - Hidden	All activation points	0.90	0.96	0.81	0.94	0.96	0.78
Hidden - Hidden	Category prototypes	0.92	0.97	0.98	0.97	0.97	0.98
Hidden - Hidden	Intra-categories average	0.66	0.60	0.81	0.60	0.48	0.77

Table 3.3. Summary of the correlations between the structural similarity of corresponding hidden layer activation spaces in ANNs categorising aspects of colour.

In order to provide further evidence that ANNs develop task-dependent structures of hidden layer activation points the structural similarity comparisons were also applied to networks trained to perform different tasks using identical input sets. If the structures are task-dependent the similarity between networks performing different tasks should be low. There was a low to moderate (0.07 – 0.57) correlation between the structures of the hidden layer activation points in ANNs trained to perform different tasks using the same reflectance spectra input data. The comparisons were only performed using all points in the input domain because there are no corresponding prototypes and categories between different task domains. The much lower correlation values between ANNs performing different tasks shows that there is a strong task-dependent aspect to the representational structure developed at the hidden layer. The average correlation values for the comparisons between ANNs performing different tasks using the same input data are provided in Table 3.4.

ANN training dataset	61 components, 627 samples			421 components, 1269 samples		
Categorisation task	5 Hues	8 Values	7 Chromas	10 Hues	8 Values	7 Chromas
Comparison	Average Correlations			Average Correlations		
Hidden Layer Target1 - Hidden Layer Target2	0.17	0.17		0.07	0.07	
Hidden Layer Target1 - Hidden Layer Target3	0.57		0.57	0.40		0.40
Hidden Layer Target2 - Hidden Layer Target3		0.33	0.33		0.37	0.37

Table 3.4. Summary of the average correlations between the structural similarity of hidden layer activation spaces in networks trained to perform different tasks using the same input data.

The empirical results show that ANNs trained to perform the same task using the same dataset developed similarly structured hidden layer activation spaces. The correlation between the arrangements of points within categories was only moderate but the values calculated using the entire input domain and the category prototypes were both very high. In contrast, the correlations between networks trained to perform different colour categorisation tasks using the same data was moderate to very low. This shows that the colour

categorisation ANNs developed a task-dependent representational structure at the hidden layer.

3.3.3 Structural similarity between ANNs with different domains

I then compared the ANNs trained to classify hue, value and chroma from the 61 component reflectance spectra with 627 input samples to the networks trained on the 421 component reflectance spectra with 1269 samples. The aim was to determine whether structure was preserved within a subset of the full task domain and whether the input encoding or resolution affects the resulting representational structure at the hidden layer. The method for assessing structural similarity between ANNs is based on a comparison of the relative proximity of corresponding activation points. So, firstly I identified the corresponding 627 input samples from the 1269 sample dataset and ordered them in accordance with the smaller dataset. The hidden layer activation spaces of every ANN in each series trained on the 627 sample input set were compared to every network in the corresponding series trained on the 1269 sample input set and the similarity values averaged.

The group of ANNs trained to classify five hues (colours) from the 61 component reflectance spectra were compared with the networks trained to classify 10 hues from the 421 component reflectance. The relative proximities between the corresponding hidden layer activation patterns determined by the collective set of activation points in the 627 sample input were highly correlated (0.90) indicating a high degree of preserved structure. The relative proximities between the prototypical activations for the five categorical distinctions made by the ANNs distinguishing five hues were compared to the corresponding five prototypical activations selected from the ten prototypes determined by the categorical distinctions made by the networks classifying ten hues. The structures of the prototype activations in the different networks were also highly correlated (0.90). This shows a robust preservation of inter-category structure between the common categories. The structures of each group of activation points belonging to the same input category were also compared with the structures of their corresponding activation points and the correlation values averaged. The structural similarity was not as high as with the full set of activation points or the prototype activation points but there was still a significant correlation (0.65).

The groups of networks trained to classify value (lightness) from the 61 component reflectance spectra and the 421 component reflectance were also compared. The overall average correlation between the structures of hidden layer activation spaces determined using the full set of 627 activation points was very high (0.95). The structures of the prototypical activation points corresponding to the inter-category distinctions were also very highly correlated (0.97). The structures of activation points within corresponding categories were not as highly correlated but there was still a significant correlation (0.65).

Finally, the groups of ANNs trained to classify chroma (saturation) from the 61 component reflectance spectra and the 421 component reflectance were compared. The average correlation between the structures of the hidden layer activation spaces incorporating all 627 points was fairly high (0.75). The structures of the prototypical activation points were very highly correlated (0.97) indicating that the inter-category structures were very similar. The structures of the activation points within corresponding categories were also highly

correlated (0.78) indicating a high degree of structural similarity being preserved within categories across different networks. All the average correlation values are summarised in Table 3.5. These results show that ANNs trained using varying input encoding and domain size can still develop hidden layer activation spaces that preserve the structure of the partially overlapping set of activation points and their corresponding prototype points.

Categorisation task		Hue	Value	Chroma
Layer comparison	**Structural comparison**	Average Correlations		
Hidden - Hidden	All corresponding activation points	0.90	0.95	0.75
Hidden - Hidden	Corresponding category prototypes	0.90	0.97	0.97
Hidden - Hidden	Corresponding intra-categories average	0.65	0.48	0.78

Table 3.5. Summary of the structural similarity of activation spaces between ANNs trained to perform the same tasks using differently encoded and partially overlapping input datasets.

The colour categorisation ANNs trained on different but related reflectance spectra datasets still developed similarly structured hidden layer activation spaces. ANNs trained to make more categorical distinctions using a larger and higher resolution dataset still had a similar structural arrangement of corresponding activation points to the networks trained to recognise fewer categories from a smaller dataset. This shows that the preservation of structure was not dependent on the ANNs using identical input encodings and being trained to determine the same range of categorical distinctions.

3.3.4 Structural similarity between hidden layer activation space and the target domain

The ANNs trained to classify reflectance spectra according to the Munsell colour system develop hidden layer activation points with a robust structure that is preserved across diversely configured groups of networks performing the same task. The preservation of task specific structure provides support for structuralist approaches to explaining representation in connectionist neural networks and should be included in a comprehensive explanation of the operation of ANNs. The analysis of hidden layer activation points in ANNs suggests that neural networks employ a structured system of representational vehicles. However, structural explanations of representational content determination require a structural resemblance relation between the representing system and the represented domain rather than only requiring a similarity between the representational structures of neural networks operating within the same task domains.

O'Brien and Opie (2006) and Churchland (1996,1998) claim that the structural arrangement of collectively considered hidden layer activation patterns reflects relevant aspects of the structure of the domain that is being represented. Churchland (1996,1998) also emphasises the structural arrangement of prototypical activations which reflect relevant aspects of the structural arrangement of the categories being represented. They both discuss the structural organisation of activation spaces in some example ANNs, however, they do not provide an empirical evaluation that quantifies their similarity to the task domain. In order to evaluate

whether structural approaches are consistent with empirical evidence I expanded my investigation. I explicitly analysed whether the structures of the prototypical hidden layer activation patterns developed by ANNs categorising aspects of colour from reflectance spectra matched relevant aspects of the colour space in the corresponding task domains. This involved determining suitable objective or external characterisations of the task domains and then comparing them to the corresponding hidden layer activation point structures.

The Munsell colour system defines a colour space that varies systematically and can be objectively characterised. The colour space is defined by three components, hue, value and chroma. The colour space can be characterised as varying across three dimensions and orientated around the centre of a roughly spherical distribution. Lightness values vary from black at the bottom of the space through to white at the top. Chroma (saturation) increases gradually from the centre with deeply saturated colours on the periphery. The hue (colour) values wrap around the space perpendicular to the value axis. The components of the colour system are depicted in Figure 3.15.

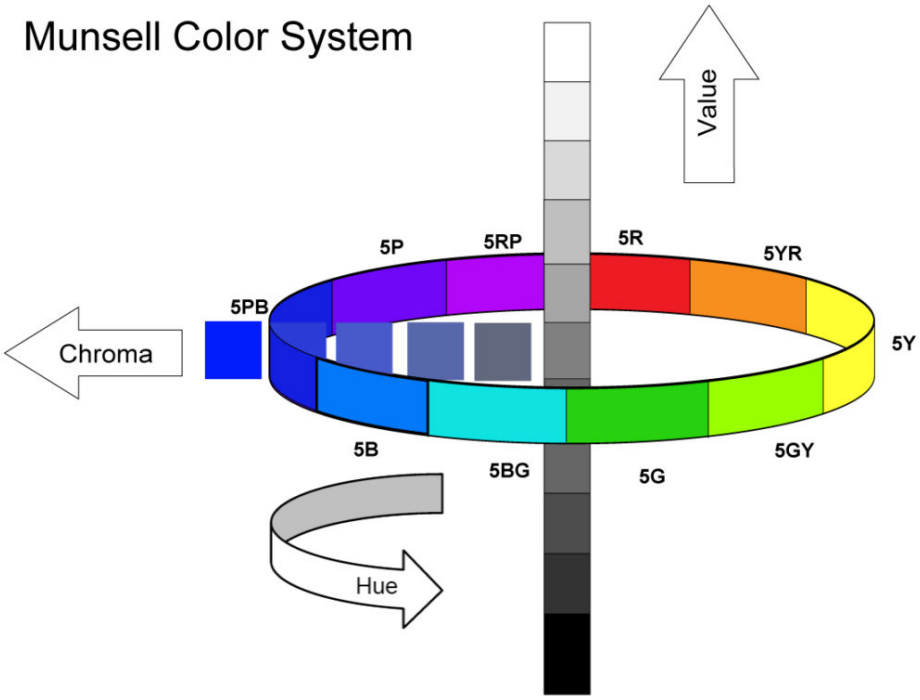


Figure 3.15. The colour space defined by the Munsell colour system varies across three dimensions (hue, value and chroma). (https://justpaint.org/wp-content/uploads/2017/06/Munsell-Illustration-w-Title-2.jpg)

I determined the structure of the component spaces by the relevant similarity and difference relations between each of their constituent properties. The relative similarities and differences correspond to objective aspects of human colour judgements. The relations between the constituents of each individual colour component can be depicted in two dimensions as shown in Figure 3.16. The hues form a circular colour wheel and the values lie along a line with chroma extending along perpendicular lines.

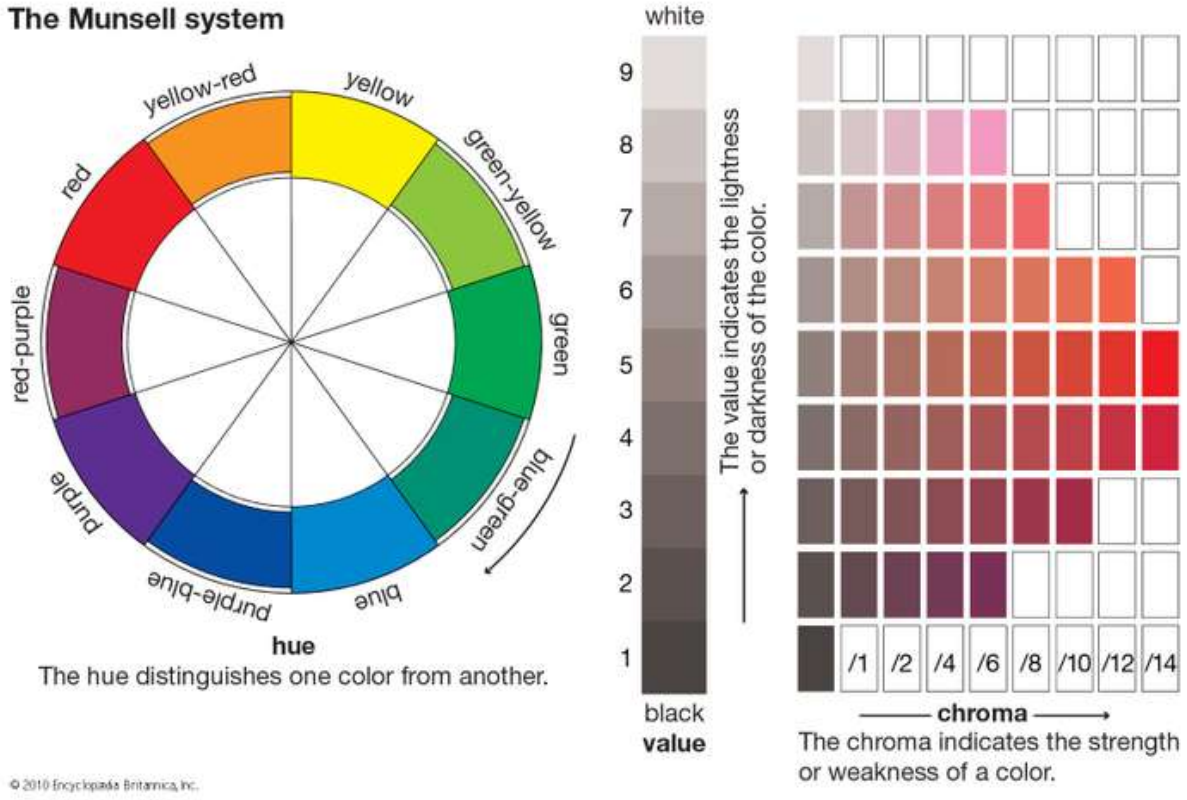


Figure 3.16. The three components of the Munsell colour system, hue, value and chroma. (https://www.britannica.com/science/Munsell-color-system)

I characterised the structure of the hues by representing them as points spaced equidistantly on the circumference of a circle following the order specified by the Munsell colour system (red, yellow, green, blue, purple). The set of distances between the points was used to determine a quantitative description of the structure based on the proximities of the corresponding hues in the task domain. The value and chroma components of the colour space already have quantitative characterisations of their constituents. The structure of the value domain was determined by calculating the distances between the range of lightness values present in the colour chips that the ANNs input reflectance data was sampled from. Similarly, the structure of the chroma domain was determined by calculating the distances between the range of chroma values present in the colour chips that the reflectance data was sampled from. The structural characterisations of the three colour components are shown in Figure 3.17.

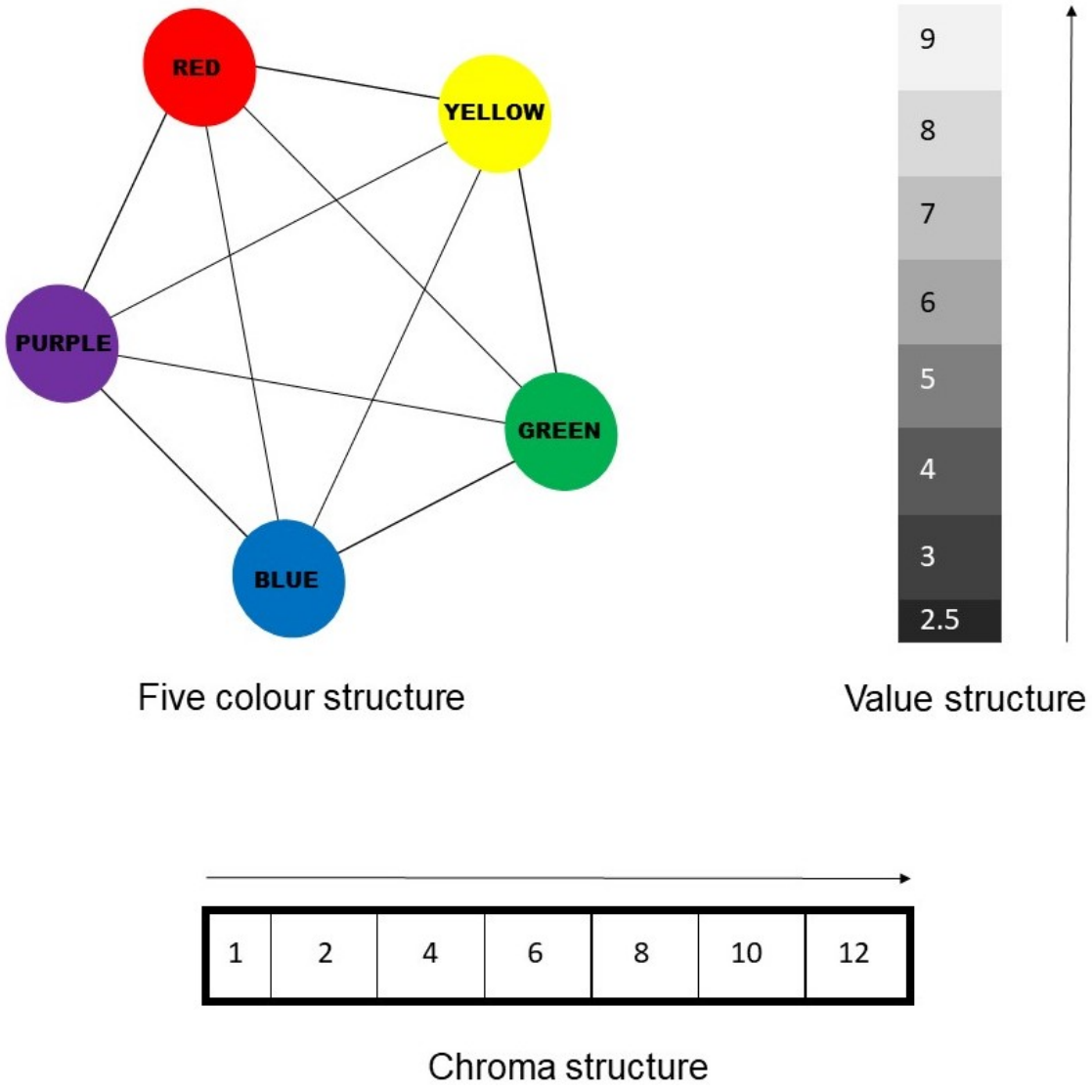


Figure 3.17. The structural relations between the five hues, the eight values and the seven chromas represented in the Munsell colour reflectance spectra datasets can be quantitatively characterised by proximity relations.

I compared the characterisations of the relations between the categories to the corresponding prototypical hidden layer activation structures determined by the categorical distinctions. Figure 3.18 shows the distribution of hidden layer activation points for two ANNs each with two hidden units and trained to categorise five colours (red, yellow, green, blue, purple). The prototype (average) activation points for each colour category are shown as larger circles containing the first letter of the colour category they represent. This provides a qualitative visualisation of the structural arrangement of category prototypes which appears consistent with the objective characterisation of hue relations that have been described. The relative locations of the prototypes in hidden layer activation space reflect the relative locations of the corresponding hues in the task domain.

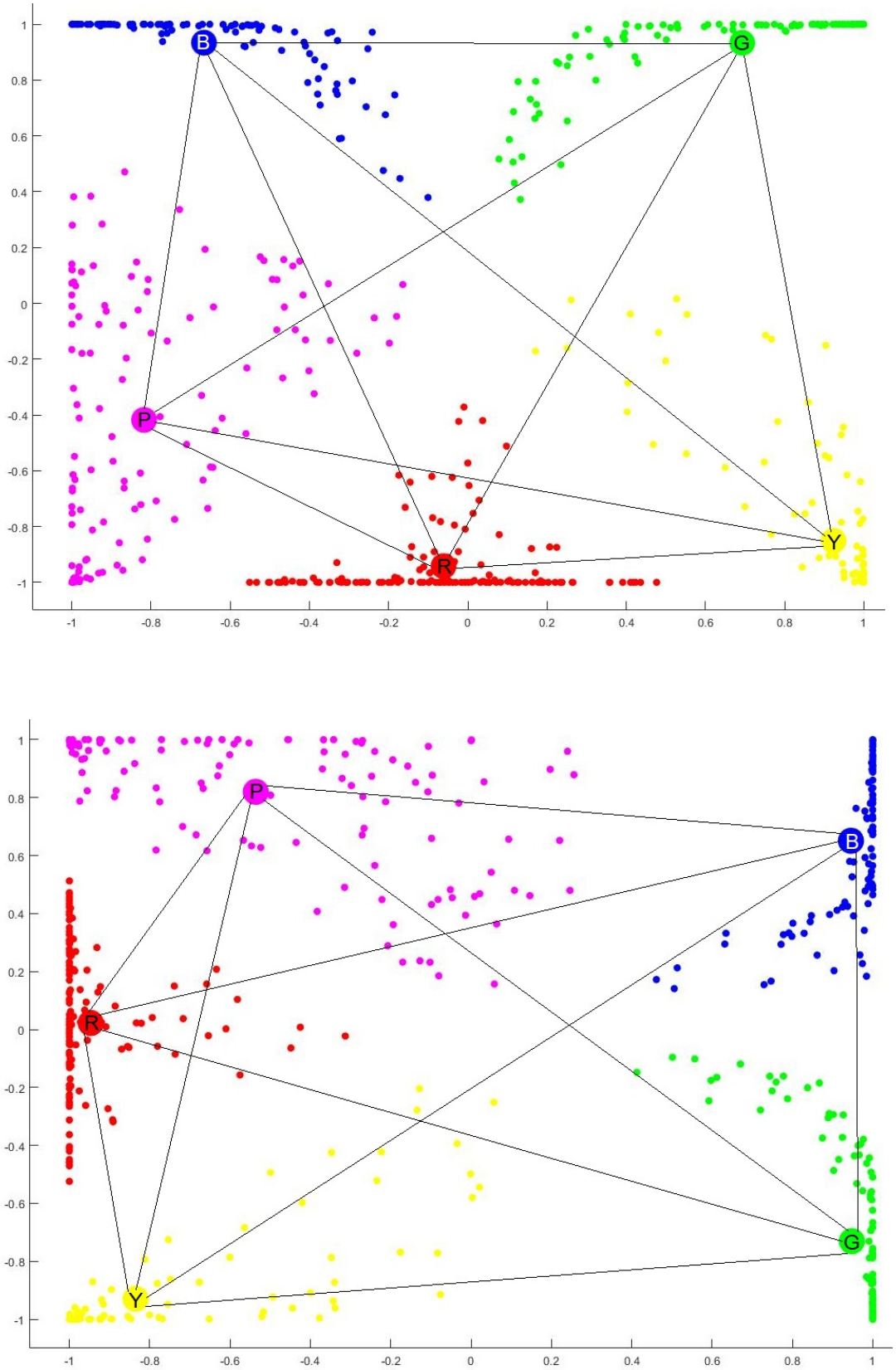


Figure 3.18. Examples of the hidden layer activation values from two different ANNs with two hidden units. The graphs show the collective set of hidden layer activation values for a five colour categorisation task (red, yellow, green, blue, purple) and the structure of the prototypical (average) activation points for each category.

I applied the quantitative method for determining the similarity between the structures of prototypical hidden layer activation points to compare each group of ANNs to their corresponding task structures in order to determine a measure of similarity. Firstly, the hidden layer prototype activation structures from the series of ANNs trained to classify five hues were compared to the objective characterisation of the structural relations between the five hues that were determined by their proximities. The ANNs trained to classify ten hues were compared to a similarly determined objective characterisation of the structural relations between the ten hues. Next, the hidden layer prototype activation structures of ANNs trained to classify eight lightness values were compared to an objective characterisation of the structural relations between the corresponding eight values. Finally, the hidden layer prototype activation structures of ANNs trained to classify seven chromas were compared to an objective characterisation of the structural relations between the corresponding seven chromas.

There was a very high correlation between the hidden layer prototype structures and the characterisations of the objective structure of the task domains for each of the three colour components (ranging from 0.89 to 0.97). Table 3.6 shows the average correlations for each of the six groups of ANNs. This shows that the networks developed robust structural relations between prototype activation points that correspond to the structure of the categories in the task domain. The structural relations between vehicles representing category prototypes reflects the structure of the represented domain. This supports structural approaches to explaining the operation of ANNs and shows that the relations between representational states corresponding to categorical distinctions mirror relations between the represented categories. [5]

ANN training dataset		61 components, 627 samples			421 components, 1269 samples		
Categorisation task		5 Hues	8 Values	7 Chromas	10 Hues	8 Values	7 Chromas
Layer comparison	**Structural comparison**	Average Correlations			Average Correlations		
Hidden Layer - Task Structure	Category prototypes	0.89	0.95	0.94	0.97	0.97	0.97

Table 3.6. The correlation between hidden layer prototype structures and an objective characterisation of the structures of the three colour components.

5

The structure of the Munsell colour space can also be characterised by the location of the component colours in three dimensional space. I developed an alternative quantification of the space by assigning three dimensional coordinates to each discrete colour. This approach allows comparisons including all activation values in the domain rather than just prototypes, however, the coordinates do not directly correspond to the structures of the three different task domains (hue, value, chroma) that the groups of ANNs were trained on. The results from these comparisons are more difficult to interpret because the task relevant structural properties are not distinctly categorised so the overview of this approach and the results are provided for reference in Appendix 3.2.

3.4 Empirical support for the varying explanations of representational content determination

I will now provide a summary of my empirical investigation and use the results to evaluate the validity of the theories of representational content determination and individuation described in Section 3.1. The three approaches that have been discussed in this chapter describe the operation of connectionist neural network models in terms of internal representational states based on either clustering, polytope regions or structural resemblance. These varying explanations will be compared with results from my empirical analysis of colour categorisation ANNs to determine their plausibility and explanatory insight.

The colour categorisation ANNs developed hidden layer activation spaces with structured groups of activation patterns which were preserved across distinct networks performing the same task. The ANNs were trained to make specific categorical distinctions but were not explicitly trained to recognise or encode relevant similarities and differences between the categories. The consistent encoding of these implicit relations is an interesting and significant property of these ANNs that needs to be addressed in a comprehensive explanation of their operation.

Structural relations between points in hidden layer activation space can be defined in varying ways and used to provide a measure of the representational or conceptual similarity between hidden layer activation spaces. The ANNs analysed in this empirical investigation were compared at three levels to be explicit about the structures concerned and their relation to the task domain.[6] The overall structure of the hidden layer space was determined using the complete group of activation points corresponding to the entire input domain. Inter-category structure was characterised using prototypical activation points corresponding to categorical distinctions and intra-category structure was determined using groups of points corresponding to specific categorical distinctions.

My empirical analysis showed that the colour categorisation ANNs developed robust structural relations between complete groups of hidden layer activation points and also between prototypical activation points that were very similar (shown by very high correlation values) across distinct networks performing the same task. In contrast, ANNs performing different classification tasks on the same datasets did not have similarly structured hidden layer activation spaces. These results show the ANNs developed task-dependent structural relations. The analysis of hidden layer activation spaces revealed distinct and task-dependent conceptualisations or representational structures that were not just a direct recoding or compression of the input data.

The task-dependent hidden layer structures and prototype structures revealed by my analysis were not explicitly present or defined in the input datasets. The ANNs were not explicitly trained to encode relations between the categories being classified but each group of networks developed consistent structural relations between hidden layer representations of corresponding categories. This is a robust and significant property of the colour categorisation ANNs. A comprehensive theory of how neural networks operate should be

[6] A detailed explanation of this approach is provided in Chapter 2.

able to explain the ability of ANNs to implicitly represent or conceptualise relations between categories.

Shea (2007) associates content with clusters of activation points which all represent the particular category associated with correct output classification. He claims that the relations between clusters are not important for determining or ascribing representational content to hidden layer clusters. However, this omits a significant and interesting property of the colour categorisation ANNs from an explanation of their operation. It has been shown that the colour categorisation ANNs develop robust and consistent structural relations between prototype activation points which correspond to the categorical distinctions. These results provide support for Churchland's (1996,1998) claim that there is a preservation of the structural relations (measured by relative proximities) between prototypical activation points that reflects relations between the corresponding categorical distinctions. Although the contents of polytope regions described by Azhar (2016) are not determined by relations between the hidden layer activation points or regions he does mention that this type of approach might be needed in order to compare the activation spaces of ANNs with different numbers of hidden units (Azhar, 2016, p714). Azhar (2016) suggests that comparing the relative proximities between the 'centers of mass of polytopes' could be used to determine a measure of similarity. This approach is very similar to comparing the structures of the prototype activation points because the 'centers of mass' of the polytopes are characteristic or prototypical points for the categorical distinctions that the ANNs make.

Analysis of the colour categorisation ANNs also showed a moderate to high within-category similarity for each of the categories compared across multiple networks. This shows that the ANNs were making intra-category distinctions with some consistency and were not representing every member of a category in an identical manner. The ANNs were not explicitly trained to make such fine grained distinctions and only needed to be able to separate the hidden layer activation points based on the targeted categorical distinctions in order to achieve accurate performance. However, my empirical results show that the arrangement of hidden layer activation points within the partitions or clusters of these networks was not random. This suggests that there may be more information encoded in the hidden layer than just the categorical distinctions classified at the output layer. The encoding of relations between categories and within categories are not explicitly used by the colour categorisation ANNs but carry information that could be extracted. This shows the presence of what Cummins (2006) refers to as unexploited representational content.

The presence of intra-category similarity is consistent with Churchland's (1996,1998) explanations of hidden layer representation. He describes how the representational content of hidden layer activation points varies systematically with their location in activation space. The arrangement of points within a partition corresponding to a particular categorical distinction reflects their degree of similarity to each other and the category exemplar or prototype. This is quite different to how Shea (2007) and Azhar (2016) ascribe and individuate representation content to regions of activation space. Shea (2007) claims that content should be ascribed to clusters and that all points within a cluster have the same content. However, the empirical results show there is some consistency between the structural arrangement of activation points within regions of activation space that are associated with the same output classifications or categories. The analysis provides some support for associating representational content with individual patterns of activity rather than specific regions of activation space. However, a more detailed analysis of the patterns of

distribution within regions and their relation to variations within the corresponding categories is required to confirm this.

The final stage of my empirical investigation involved comparing hidden layer prototype structures from the colour categorisation ANNs to an objective external categorisation of the structure of the task domains. This is an important additional step because structural approaches to content determination are claiming that there is a structural resemblance relation between systems of representing vehicles and the task domain. This also implies that ANNs trained on the same task should have very similar representational structures to each other and this is indeed the case with the colour categorisation ANNs. The very high correlation between proximity relations from corresponding groups of activation points across distinct ANNs performing the same task provides support for the claim that this structure relates to or reflects the task domain. However, the relation to the task domain still needed to be analysed explicitly to evaluate structural-resemblance theories.

The groups of ANNs classifying each of the three different components of colour (hue, value and chroma) had hidden layer prototype activation structures that were all highly correlated with the structures of their respective constituent categories. There was a robust similarity of structure between groups of hidden layer activation points and the corresponding task domain. The empirical analysis is consistent with the claim that activation patterns are representational vehicles and relations between the vehicles (modelled by proximity relations) mirror relevant relations between aspects or properties of the task domain. The preservation of structural relations between hidden layer activation patterns across different ANNs trained on the same task and the correspondence of the structural relations between category prototypes to the structure of the task domain are consistent with and support O'Brien and Opie's (2006) discussion of representation in connectionist neural networks. The results from this investigation provide empirical support for structural approaches to characterising representation in neural networks and explaining their operation.

Structural approaches to describing representation in ANNs provide a more comprehensive explanation of activation pattern representation than theories which disregard relations between groups of activation patterns. However, there is still more required to provide a complete understanding of the operation of ANNs. The patterns of activity generated at the hidden layer are determined by how the connection weights transform the input activity. A full explanation of the operation of neural networks also requires a robust understanding of how connection weights implicitly store or represent information. A brief discussion of connection weight representation is provided in Chapter 4.

3.5 Summary

The operation of neural networks can be explained in terms of the processing or transformation of internal representation states. I have reviewed a variety of approaches to characterising the representational vehicles (the representing entities) and determining their representational content. Shea (2007) claims that the representational vehicles are clusters of hidden layer activation points corresponding to correctly classified inputs. The content of the clusters is determined by their corresponding classification. He states that the relations between the clusters are not significant. Azhar (2016) claims that the representational

vehicles are geometric regions of hidden layer activation space called polytopes that are determined by the specific output classifications they generate. The content of a polytope is determined by the input class it shares the highest mutual information with.

O'Brien and Opie (2006) claim that the collectively considered distributed patterns of hidden layer activation present during processing in connectionist neural networks structurally resemble relevant aspects of the represented domain. Churchland (1996,1998,2007) emphasises the importance of the relations between prototypical activations for each categorical distinction a network is sensitive to and claims that they reflect relevant relations between the categories in the task domain. He also explains that the proximity of activation points to their corresponding prototype reflects their degree of deviation from the category exemplar. The vehicles of representation are individual hidden layer activation points with content determined by the relative proximities between the points and the structure of the represented domain.

To evaluate the validity of these approaches I conducted an original empirical investigation analysing a range of colour categorisation ANNs with varying configurations. The analysis revealed that the ANNs developed a robust task-dependent structural organisation of hidden layer activation patterns. The correlations between the relative locations of corresponding activation points for ANNs trained to perform the same task with the same dataset were very high. The structures of the prototype activations were also very highly correlated which indicates that the inter-category relations were preserved. There was also a moderate correlation between the structural arrangements of activation points from corresponding categories which indicates that the intra-category relations were not completely random but only partially preserved. These results were reinforced by comparing ANNs that were trained to perform the same task using different but related datasets that varied in both the input sample encoding, the number of inputs and the total number of output categories. The results were very similar and consistent with the comparisons between ANNs using the same datasets. The empirical results show that the colour categorisation ANNs developed very robust and consistent structural relations between prototype activations for corresponding colour categories.

The empirical results provide support for the claims of Churchland (1996,1998,2007) and O'Brien and Opie (2006). Structural similarity between corresponding ANNs is required for structural explanations of representation in neural networks but it is not sufficient. Structural explanations of representational content determination require a structural resemblance relation between the representing system and the represented domain. To directly evaluate the plausibility of this claim I expanded my empirical investigation to explicitly compare the structures of the hidden layer prototype activation points to independent and objective external characterisations of the structures of the corresponding task domains.

Analysis of the colour categorisation ANNs revealed that structures of the hidden layer prototype activation points were very similar to the structures of their corresponding task domains. This shows that the relations between the representational vehicles reflect relations between the categories that they represent. This is a structural resemblance relation between prototype activation points and the corresponding task domain. These empirical findings support Churchland's (1996,1998) claims that the structures of hidden layer prototype activations mirror relevant aspects of the task domain and also support O'Brien and Opie's (2006) claim that there is a structural resemblance between the

representing and represented domains. The results contradict Shea's (2007) claim that the relations between clusters of hidden layer activation points are not important and this issue is not directly addressed by Azhar's (2016) approach. Structural-resemblance approaches to characterising representation in neural networks are consistent with the empirical analysis of a range of diversely configured colour categorisation ANNs and provide a more comprehensive explanation of the operation of these networks.

Chapter 4 – Conclusion

This chapter begins with a discussion of O'Brien and Opie's (2006) structural-resemblance approach to understanding connection weight representation. I explain that although this explanation may appear promising it is difficult to apply to all ANNs due to their unconstrained variety and complexity. This is highlighted by inspecting the connection weights from a range of my colour categorisation ANNs. The chapter concludes with a summary of the key issues and findings that my research has revealed along with some suggestions for future research.

4.1 Connection weight representation

In Chapter 2 and 3 my discussion of representation in artificial neural networks (ANNs) focussed on the transient patterns of activation that are propagated through a network and the corresponding hidden layer activation spaces. This approach facilitates the identification of common representational properties of distinct ANNs and is a useful and important part of the explanation of how ANNs operate. However, the transient activation patterns are the result of processing that is driven by the connection weights. A comprehensive explanation of the operation of neural networks also needs to explain how the patterns of connectivity drive the behaviour of the system and what is implicitly represented by the connection weights.

Understanding how connection weights represent and process information is particularly challenging and may appear elusive because both the number and strength of connection weights can vary so greatly between ANNs, even when they are performing the same task. There is no direct method to compare the connection weight values of ANNs with different numbers of connections and even networks with the same number of weights often have vastly different absolute weight values due to the stochastic parameters used during training. The final weight configurations depend on the initial weight configurations as well as other parameters chosen during the training process.

In Chapter 3 I provided empirical support for O'Brien and Opie's (2006) explanation of how ANNs compute. They claim that hidden layer activation patterns are representing vehicles and when considered collectively they structurally resemble the task domain. O'Brien and Opie (2006, p37) acknowledge that a comprehensive understanding of how ANNs compute also requires an explanation of connection weight representation. They begin their investigation by calculating the pair-wise correlations between the hidden layer connection weight matrices of a group of simple three-layer feedforward colour categorisation networks. This analysis revealed that the correlation values were randomly distributed which shows that there is no simple, first-order relationship between the patterns of connectivity.

O'Brien and Opie (2006, p38) claim that the key players in neural network processing are what they refer to as 'fan-ins'. A fan-in is the vector of weight values describing the particular pattern of connectivity that feeds in to a specific processing unit. A fan-in at the hidden layer consists of the weight values connecting each input unit to a particular hidden layer unit. These fan-ins determine how a processing unit responds to a specific input

activation and are responsible for determining the structure of hidden layer activation space. This implies that the fan-ins play a crucial role in ANN processing.

O'Brien and Opie (2006, p38) propose that the hidden layer fan-ins of a trained neural network structurally resemble aspects of the task domain. They investigated this proposal by analysing ANNs with three hidden units that were trained to categorise three colour hues (red, green and blue) from a subset of the 61 element Munsell reflectance spectra dataset (a description of the dataset is provided in Section 3.3). They compared the hidden layer fan-ins with the mean reflectance spectra of the three colours the networks were trained to categorise. The mean reflectance spectra correspond to prototypical input activations for these networks. Interestingly, this revealed that each of the three fan-ins had a very similar structure to the mean reflectance spectrum of one of the three colours. This is highlighted in Figure 4.1 by a graphical comparison of the mean reflectance spectra with the fan-ins from one of the networks.

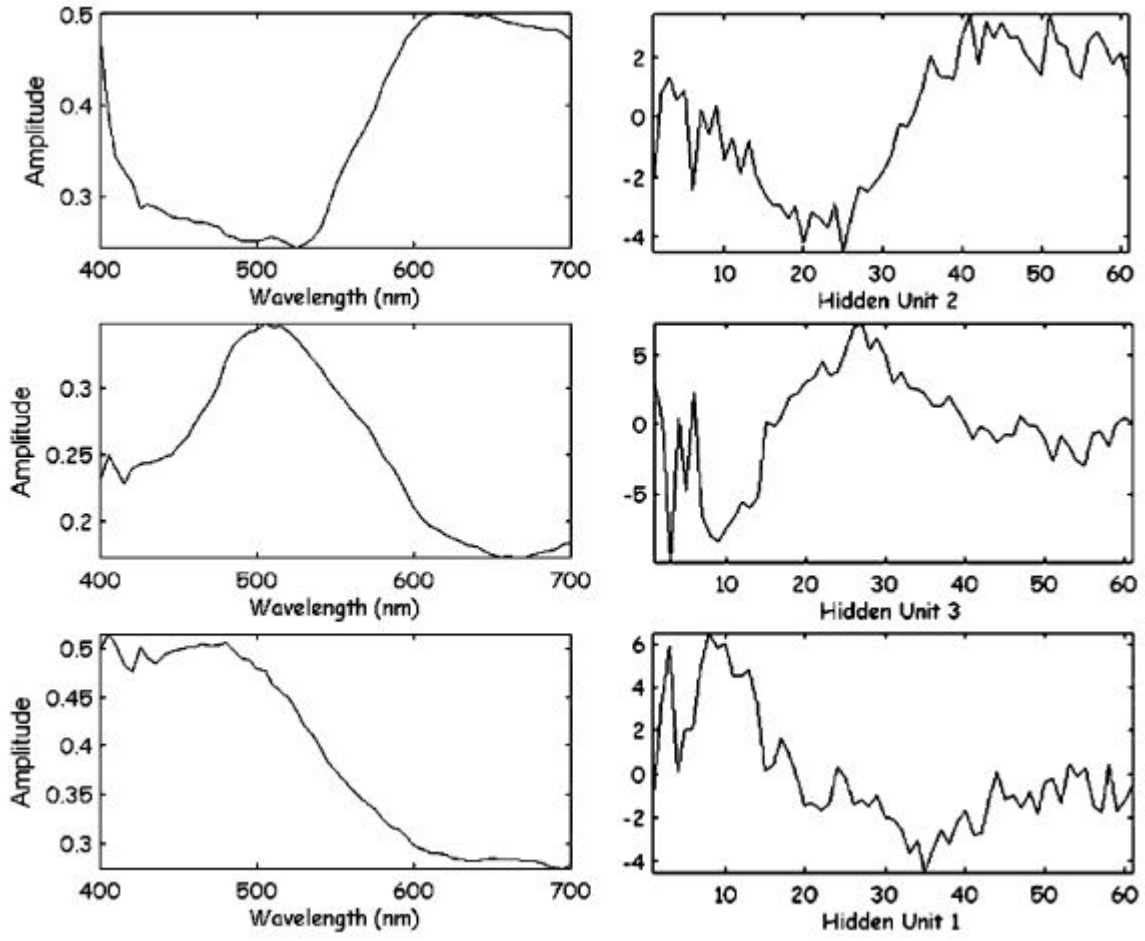


Figure 4.1. The mean reflectance spectra of the three input classes (red, green and blue) that an ANN was trained to categorise compared with the hidden unit fan-ins that the network developed (weight value on the y-axis, input index on the x-axis) (O'Brien & Opie, 2006, p40).

O'Brien and Opie (2006, p39) explain that the ''shape'' of fan-in vectors govern the activities of their respective hidden units by way of the dot product of connection weights and input activation. The dot product provides a measure of the similarity between the weight vectors and input vectors. O'Brien and Opie (2006, p39) claim that the fan-ins can be interpreted as filters that are sensitive to inputs with a particular shape. The dot product indicates how closely an input matches a particular fan-in filter and this dictates the activity of the corresponding hidden unit. Patterns of hidden layer activation reflect the degree of similarity between the input spectra and the fan-ins. The hidden layer activation space forms a map that allows filtered versions of the input spectra to be compared.

O'Brien and Opie (2006, p40) claim that connectionist neural networks are capable of performing certain tasks because they structurally resemble the task domain. The structural resemblance is sustained by both the collection of activation patterns that are produced across a network's hidden units in response to its various inputs and by the higher order structure of the network's hidden layer connection weights. This appears to be an appealing and comprehensive explanation of how ANNs operate. In Chapter 3 I provided empirical support for the first part of O'Brien and Opie's (2006) claim by analysing novel groups of diversely configured ANNs categorising aspects of colour from reflectance spectra. However, the second part of their claim requires additional empirical investigation.

O'Brien and Opie (2006, p40) only show one example of the resemblance between the fan-ins and the mean reflectance spectra for the colour categories that comprise the task domain. The ANN they describe also has exactly the same number of hidden units as the number of categories the network can distinguish between. I will show that it is difficult to provide more general empirical support for O'Brien and Opie's (2006) description of connection weight representation by analysing the fan-ins of a group of ANNs trained to categorise five colour hues (red, yellow, green, blue, purple) from the same 61 element Munsell reflectance spectra dataset that they used.

To begin with I analysed two ANNs that had the same number of hidden units as categories in the output domain. Figure 4.2 shows the shapes of the mean reflectance spectra of the five colours and the connection weight fan-ins of two distinct ANNs trained to categorise these five colours using five hidden units. This example shows that although there is some structural similarity between hidden unit fan-ins and the mean reflectance spectra this is not the case for all of the fan-ins and the degree of similarity varies. Extending O'Brien and Opie's (2006) approach to ANNs with varying numbers of hidden units is even more challenging. Figure 4.3 and 4.4 show the shapes of the mean reflectance spectra compared with the fan-ins of colour categorisation ANNs with two, three, four and twenty five (first five shown) hidden units. In these examples there is no longer any clear or obvious structural similarity between the fan-ins and the mean reflectance spectra. Some of the connection weight fan-ins also have quite erratic shapes which appear difficult to match to any aspects of the task being performed.

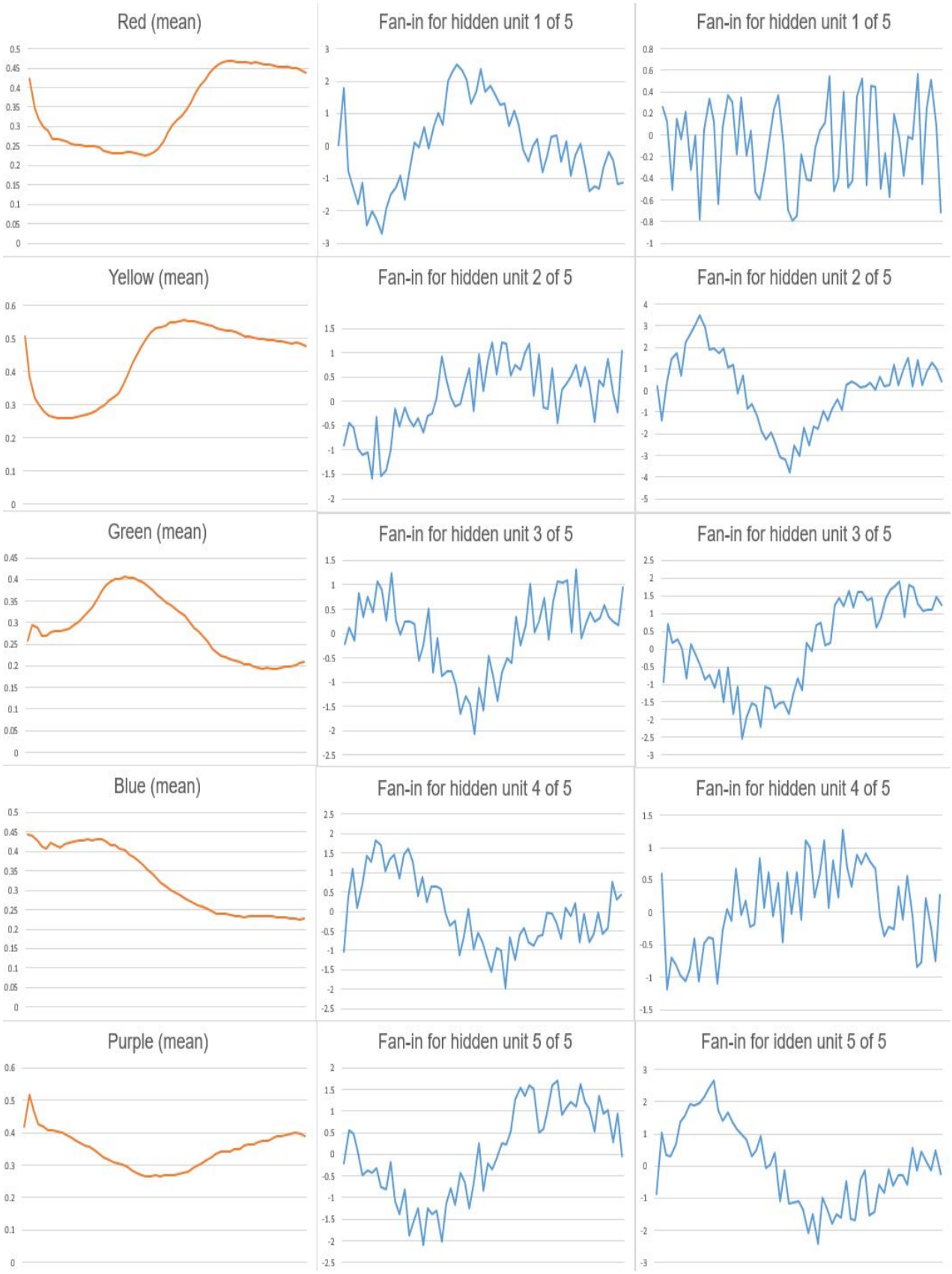


Figure 4.2. Shapes of the mean reflectance spectra for five colours from the Munsell dataset compared with the connection weight fan-ins of two distinct ANNs trained to categorise these five colours using five hidden units.

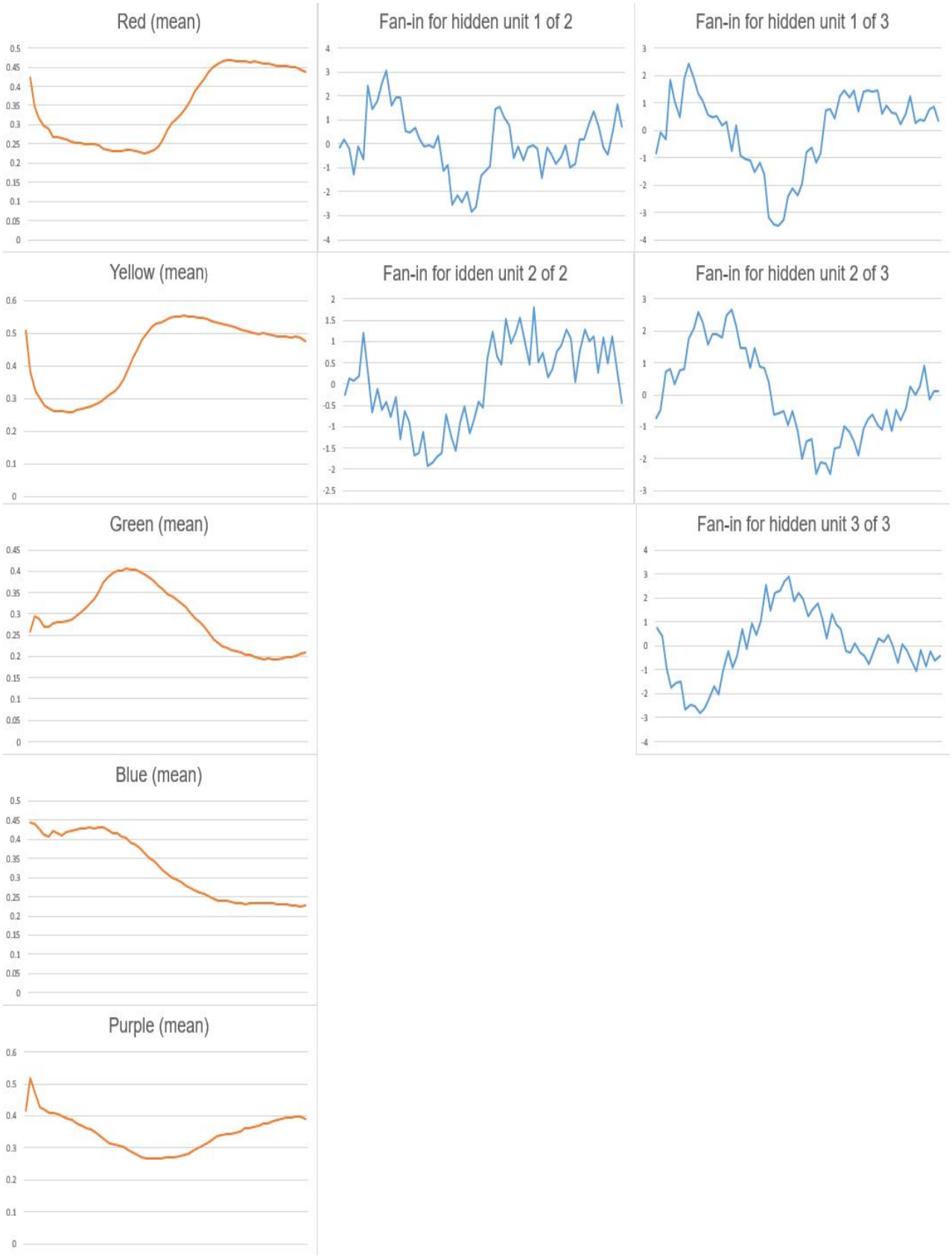


Figure 4.3. Shapes of the mean reflectance spectra for five colours from the Munsell dataset compared with the connection weight fan-ins of two distinct ANNs trained to categorise these five colours using two hidden units and three hidden units.

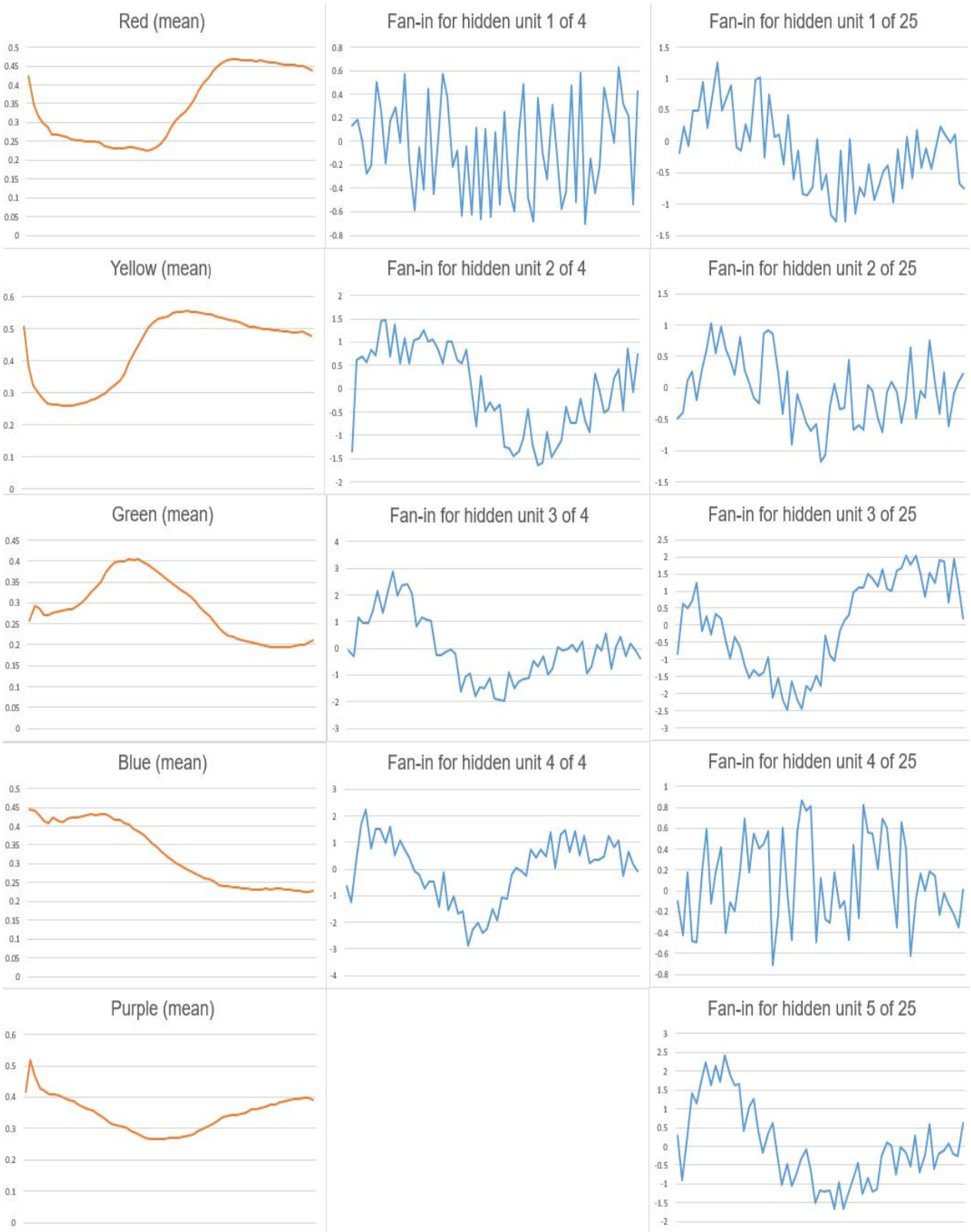


Figure 4.3 Shapes of the mean reflectance spectra for five colours from the Munsell dataset compared with the connection weight fan-ins of two distinct ANNs trained to categorise these five colours using four hidden units and 25 (5 shown) hidden units.

O'Brien and Opie's (2006) approach to describing connection weight representation has strong theoretical foundations and provides insight into some key aspects of connection weight processing. However, the hidden layer fan-ins cannot always by clearly matched to relevant aspects of the task domain. The way the weights filter for aspects of the task domain may involve prototypical input structures (such as the structures of the mean reflectance spectra) being spread across the hidden layer fan-ins in complex and convoluted combinations. It is unclear how to generalise O'Brien and Opie's (2006) approach so it can be applied to all ANNs. Understanding how connection weights drive the behaviour of ANNs requires further research in order to develop an explanation that can be applied to networks with any number of hidden units and any pattern of connectivity that facilitates correct performance.

4.2 Conclusion

Artificial neural networks (ANNs) are computational systems that were inspired by biological neural networks in the brain. ANNs are trained to transform input into task appropriate output using learning algorithms rather than having all relevant aspects of the task explicitly encoded with symbolic rules. Despite the increasingly impressive performance and wide spread usage of ANNs in artificial intelligence, their operation remains somewhat mysterious. Determining what ANNs actually learn and explaining how they transform their inputs into task appropriate outputs remains elusive. There is no widely accepted and comprehensive explanation of how these systems represent and process information.

Approaches to explaining the operation of relatively simple neural network models have been discussed by philosophers since the inception of connectionist cognitive science. They have sought to describe the information processing that occurs in terms of the transformation of internal representation states. However, these discussions often relied on analysing the behaviour of a very small number of actual ANNs and there are important issues that still haven't been resolved. In order to address this, I have reviewed and evaluated some key philosophical approaches to understanding the operation of ANNs by using empirical data from my own unique analysis of a broad range of novel ANNs.

In Chapter 2 I described a prominent position in foundational connectionist cognitive science developed by Paul Churchland (1989,1996,1998,2007,2012). He explains that analysing the hidden layer activation space of successfully trained ANNs reveals that they develop a task relevant partitioning. He claims that prototypical activation points can be determined for each category and the representational content of activation points varies systematically with their distance from the prototypes. The relative locations or arrangement of collectively considered activation points reflects relevant relations in the represented domain. Churchland (1998) used Laakso and Cottrell's (2000) quantitative method for assessing representational similarity to show that ANNs with different architectures and patterns of connection weights could be compared by determining the structural similarity between their hidden layer activation spaces.

I provided novel empirical results that show Laakso and Cottrell's (2000) method should be extended to explicitly include comparisons between the prototypes characterising categorical distinctions. The method was also extended to include explicit comparisons within partitions

corresponding to distinctions between members of the same category. I used this newly developed extended method to compare a broad range of novel facial recognition ANNs to determine whether the structures of their hidden layer activation spaces were consistent with Churchland's (1996,1998) claims.

My empirical analysis showed that ANNs with relatively small hidden layers (5-25 units) developed robust hidden layer partitioning that facilitated accurate performance but did not develop consistent relations between or within the facial categories represented. There was no consistency between the structural arrangements of prototypical activation points in this group of networks. In contrast, comparing distinct ANNs with 80 hidden units (the same number as the example ANN discussed by Churchland, 1996) showed that they developed robust and consistent relations between the structural arrangements of hidden layer activation patterns that were not explicitly trained for. There were similar structural relations preserved between the prototypical activations for each facial category and also between corresponding activation patterns within each facial category. However, it is unclear how closely these relations correspond to our intuitive assessment of facial similarity or whether they reflect artificial properties of the input data.

Laakso and Cottrell's (2000) original method was not sensitive to the structural differences between the hidden layer activation spaces of the facial categorisation ANNs with a relatively small number of hidden units (5-25) and those with a relatively large number (80). The extended method that I developed is required in order to provide an explicit assessment of the relevant structural similarities. This method has similarities to approaches used in neuroscience (for example, Kriegeskorte et al, 2008) and may assist in determining the biological plausibility of connectionist neural network models. My analysis showed that the structural relations between facial categories that developed in the ANNs with sufficiently large hidden layers may not align with or use the same relations as human judgements of facial similarity. However, my empirical investigation in Chapter 3 goes on to provide evidence that groups of distinct ANNs categorising various aspects of colour do in fact develop structural relations that directly align with objectively categorised human colour judgements.

I began Chapter 3 by reviewing the approaches to characterising representation in ANNs offered by Shea (2007), Azhar (2016) and O'Brien and Opie (2006). Shea (2007) claims that representational content is associated with clusters of hidden layer activation points corresponding to correctly classified inputs. He states that the relations between the clusters are not significant. Azhar (2016) claims that the representational vehicles are geometric regions of hidden layer activation space called polytopes that are determined by the specific output classifications they generate. The content of a polytope is determined by the input class with which it shares the highest mutual information. O'Brien and Opie (2006) claim that the collectively considered distributed patterns of hidden layer activation present during processing in connectionist neural networks structurally resemble relevant aspects of the represented domain. This appears consistent with Churchland's (1996,1998) approach although he emphasises the importance of the structural relations between prototypical activations for each categorical distinction the network is sensitive to. Structural approaches are also becoming more prominent in contemporary cognitive neuroscience (Williams & Colling 2018).

To evaluate the validity of these approaches I conducted an original empirical investigation analysing a range of ANNs with varying configurations that were trained to categorise three different aspects of colour from two distinct reflectance spectra datasets. The analysis revealed that the ANNs developed robust task-dependent structural organisation of hidden layer activation patterns. The structures of the prototypical activations for each category were very similar across each group of networks trained to perform the same task which indicates that the task-dependent inter-category relations were preserved. There was also moderate similarity between the structural arrangements of activation points within corresponding categories which indicates that the intra-category relations were also partially preserved. These results provide support for using structural approaches to explaining representation in ANNs.

I expanded my empirical investigation to explicitly compare the structures of the hidden layer prototype activation points to independent and objective external characterisations of the structures of the corresponding task domains. Analysis of the colour categorisation ANNs revealed that structural arrangements of hidden layer prototype activation points were very similar to the structures of their corresponding task domains. This shows that the relations between the representational vehicles structurally resemble the relations between the categories that they represent. These empirical findings support Churchland's (1996,1998) claims that the structures of hidden layer prototype activations mirror relevant aspects of the task domain and also support O'Brien and Opie's (2206) claim that there is a structural resemblance between the representing and represented domains. The results contradict Shea's (2007) claim that the relations between clusters of hidden layer activation points are not important and this significant property of ANNs is not directly addressed by Azhar's (2016) approach. Structural-resemblance approaches provide a more comprehensive explanation of the representational properties and operation of the colour categorisation ANNs.

My analysis of facial categorisation ANNs provided some support for using a structural-resemblance approach but also revealed some inconsistencies between different sized networks. The facial categorisation ANNs with 80 hidden units developed similar structural relations between prototypical activations but this was not the case for the networks with relatively small hidden layers (5-25 units). It may be that these ANNs were too small to have the representational capacity to learn or encode additional relations between categories that were not explicitly trained for. Perhaps it may also be argued that in these limited artificial cases the most practically relevant structural distinctions are just the separation of different categories. More research is required to understand why all the other groups of ANNs developed similar structural relations between categories without explicit training but the relatively small facial image categorisation networks did not.

A comprehensive explanation of the operation of neural networks also needs to explain how the patterns of connectivity drive the behaviour of the system and what is implicitly represented by the connection weights. In the previous section I provided an exposition of O'Brien and Opie's (2006) claim that the patterns of connection weights that fan-in to a hidden layer processing unit structurally resemble aspects of the task domain such as category prototypes. This is an appealing approach that could provide a comprehensive computational explanation of the operation of ANNs. However, I have shown that although this may work for simple or idealised cases it is difficult to apply to all ANNs due to the potential diversity of weight configurations. Developing a method for determining and

comparing the representational or transformational properties of connection weights that can be applied to any ANN is an important part of understanding ANNs that requires further research.

Appendices

Appendix 1 – Matlab code examples

My empirical analysis involved writing hundreds of lines of code in Matlab. An example of the scripts that were used to extend Laakso and Cottrell's (2000) method for assessing representational similarity are included below. Note that this code relies on my custom network object that includes the sets of distances between all activation points, the prototype activation points and the activation points within each category.

The function below determines the average correlation between corresponding sets of prototype distances in a group of ANNs.

```
function [averagecorrelation] = hlhlProtoCorr(networks)

% returns the average correlation between corresponding hidden layer
% prototype distances for all pairs of ANNs in the input group (networks)
% the input (networks) is an arbitrary length array of custom network structures
% each network structure includes a HiddenLayerPrototypeDistances field containing the set of distances

totalcorrelation = 0;
numberofcomparisons = 0;

% determine the cumulative correlation between corresponding sets of prototype distances for every pair of
ANNs in the input array
for netindex = 1 : size(networks,2)
    for comparisonindex = (netindex + 1) : size(networks,2)
        correlation = corrcoef((networks{netindex}.HiddenLayerPrototypeDistances),
                               (networks{comparisonindex}.HiddenLayerPrototypeDistances));
         totalcorrelation = totalcorrelation + correlation(1,2);
         numberofcomparisons = numberofcomparisons + 1;
   end;
end;

% calculate the average correlation between all corresponding hidden layer prototype distances based on the
number of comparisons performed
averagecorrelation = totalcorrelation / numberofcomparisons;
```

The function below determines the average correlation between corresponding sets of intra-cluster distances in a group of ANNs.

```
function [averagecorrelation] = hlhlIntraCorr(networks)

% returns the average correlation between corresponding hidden layer
% intra-category distances for all pairs of ANNs in the input group (networks)
% the input (networks) is an arbitrary length array of custom network structures
% each network structure includes an IntraCategoryDistances field containing the set of distances

totalcorrelation = 0;
numberofcomparisons = 0;

% determine the cumulative correlation between corresponding sets of hidden layer intra-cluster distances for
every pair of ANNs in the input array
for netindex = 1 : size(networks,2)
    for comparisonindex = (netindex + 1) : size(networks,2)
        for clusterindex = 1 : size((networks{1,netindex}.IntraCategoryDistances),1)
            correlation = corrcoef((networks{1,netindex}.IntraCategoryDistances{clusterindex,1}),
                          (networks{1,comparisonindex}.IntraCategoryDistances{clusterindex,1}));
           totalcorrelation = totalcorrelation + correlation(1,2);
          numberofcomparisons = numberofcomparisons + 1;
      end;
   end;
end;

% calculate the average correlation between all corresponding hidden layer intra-cluster distances based on the
number of comparisons performed
averagecorrelation = totalcorrelation / numberofcomparisons
```

Appendix 2 - An alternative method for assessing representational similarity

Churchland (1998) also described an alternative method for assessing representational similarity based on the relative locations of activation points but it was not empirically tested. I will briefly analyse this approach and compare it to Laakso and Cottrell's (2000) method in order to determine which type of structural variations the different methods are sensitive to. The alternative approach assesses representational similarity based on the differences between the lengths of corresponding lines that define the hypersolids characterised by the arrangement of prototypical activation points distributed across the hidden layers of a pair of ANNs. Each difference value is divided by the sum of the lengths of the two corresponding lines in order to provide some normalisation. The values are then averaged for the edges that connect the prototypical points. The similarity value is calculated by subtracting this average from one to produce a value between zero and one where higher values indicate a higher degree of similarity.

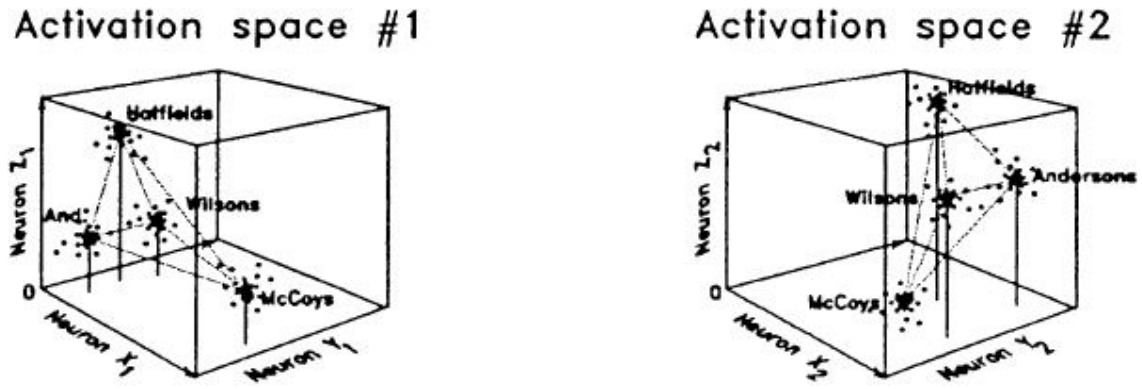


Figure A2. There is a similarity between the relative shapes of hypersolids defined by the prototypical activation points for hypothetical networks categorising faces into family groups. (Churchland, 1998, p17).

Similarity = 1 – Average [|AB – A'B'| / (AB + A'B')]

(where AB is the length of the line from point A to point B in network 1 and A'B' is the length of the corresponding line segment in network 2)

This simple similarity measure has the benefit of being invariant to translation, rotation and mirror inversion of the hypersolids defined by the structure of prototypical activation points. However, it is sensitive to the global scale of the hypersolids in activation space and so the inclusion of an additional term to adjust for this was also suggested. The scaling factor (labelled as 'c') is determined by summing the lengths of all edges in one space and dividing this by the sum of the lengths of all the corresponding edges in the comparison space. The scaling value is then multiplied by the second term in both the numerator and denominator to yield the final scale invariant similarity metric.

Similarity = 1 – Average [|AB – cA'B'| / (AB + cA'B')]

In order to compare this similarity metric with Laakso and Cottrell's (2000) method the similarity values were calculated for the same groups of ANNs (4 X 105 different networks) that were analysed previously. The method was also extended to include the same additional comparisons described in Chapter 2 and then applied to all activation points in the domain, the prototypical points and the distribution of points within their respective categories. Some of the results are consistent with Laakso and Cottrell's (2000) method, however, this is not always the case. The metric has a different output range (zero to one rather than minus one to one) which requires the interpretation to be adjusted but this does not account for the specific variation and inconsistency observed.

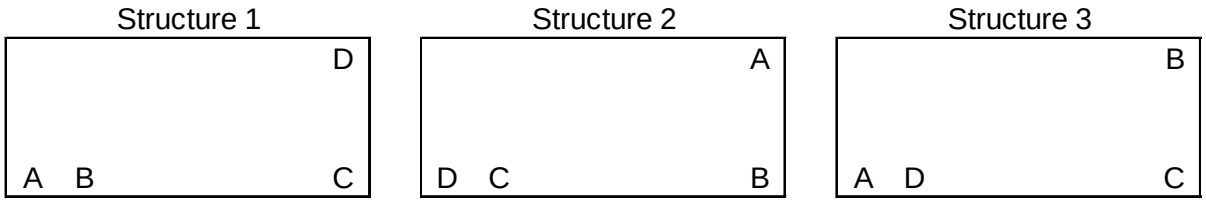
The similarity values calculated using this alternative method are consistent with the correlation values calculated using Laakso and Cottrell's (2000) method for the colour recognition networks. However, the results for the face recognition networks are not consistent with those calculated using Laakso and Cottrell's (2000) method (even allowing for the difference in output range). The similarity values for both series of 105 face categorisation networks calculated using the full set of activation points is high (0.89 and 0.90). The similarity values calculated based on the prototype points is even higher (0.94 and 0.93) and this indicates a high degree of structural similarity between the corresponding hypersolids and a high degree of representational similarity. This is in stark contrast to values calculated using Laakso and Cottrell's (2000) method which revealed there is almost no correlation (0.07 and 0.07) and consequently almost no representational or structural similarity between corresponding sets of prototypical points. The similarity values for the intra-cluster comparisons are quite low (0.51 and 0.53) and allowing for the reduced output range these values appear consistent with the correlations calculated using Laakso and Cottrell's (2000) method.

Colour categorisation networks similarity (alternative method)				
NETWORK DATASETS	Munsell Colour 61bin		Munsell Colour 421bin	
TARGETS	5 Colours		10 Colours	
COMPARISONS (Average correlations for all networks)	Average Correlations	Comparisons	Average Correlations	Comparisons
Input Layer - Hidden Layer	0.7449	105	0.7236	105
Input Layer Prototypes - Hidden Layer Prototypes	0.9141	105	0.8866	105
Input Layer IntraCluster - Hidden Layer IntraCluster	0.7453	525	0.6778	1050
Hidden Layer - Hidden Layer	0.9050	5460	0.9223	5460
Hidden Layer Prototypes - Hidden Layer Prototypes	0.9653	5460	0.9629	5460
Hidden Layer IntraCluster - Hidden Layer IntraCluster	0.7960	27300	0.7549	54600
Input Layer - Output Layer	0.6027	105	0.6890	105
Hidden Layer - Output Layer	0.7121	105	0.8107	105
Output Layer - Output Layer	0.8287	5460	0.9234	5460
Average Total Performance (% correct)	99.9332		98.5245	
Average Test Set Performance (% correct)	99.5314		96.8804	
Face categorisation networks similarity (alternative method)				
NETWORK DATASETS	CAFE Faces		Yale Faces	
TARGETS	10 Faces		15 Faces	
COMPARISONS (Average correlations for all networks)	Average Correlations	Comparisons	Average Correlations	Comparisons
Input Layer - Hidden Layer	0.8717	105	0.8706	105
Input Layer Prototypes - Hidden Layer Prototypes	0.9288	105	0.9095	105
Input Layer IntraCluster - Hidden Layer Intra-Cluster	0.6134	1050	0.6325	1575
Hidden Layer - Hidden Layer	0.8917	5460	0.8985	5460
Hidden Layer Prototypes - Hidden Layer Prototypes	0.9423	5460	0.9315	5460
Hidden Layer IntraCluster - Hidden Layer IntraCluster	0.5117	54600	0.5317	81900
Input Layer - Output Layer	0.8592	105	0.8607	105
Hidden Layer - Output Layer	0.8758	105	0.9036	105
Output Layer - Output Layer	0.9230	5460	0.9296	5460
Average Total Performance (% correct)	99.6071		98.9610	
Average Test Set Performance (% correct)	97.6190		94.9580	

Table A2. Results using the alternative method for assessing representational similarity that was described by Churchland (1998).

The alternative approach is a relatively simple and intuitive method of comparison. However, Laakso and Cottrell's (2000) method is more sensitive to the distribution of variation in the distances between activation points. The alternative method returned high similarity values when there were relatively small variations between distances. However, because the variations were averaged the results were not sensitive to subtle changes in the distribution of the variation which indicated differences in the relative locations of activation points. Similarity values determined using this approach did not reliably indicate the degree to which relevant structure was preserved in cases where relatively small differences in distances occurred in different locations and did not follow a similar pattern of variation.

The following simple example highlights how similarity values calculated using different comparison methods can vary significantly. The order of proximity between the points labelled A – D in Structure 1 is preserved in Structure 2 but the absolutes distances vary. The structures are similar and this is reflected by reasonably high similarity values from both methods. However, the points in Structure 3 have a significantly different arrangement. A is closest to D and most distant from B which is the opposite relation to Structure 1 and 2. Using Laakso and Cottrell's (2000) method reveals a substantial negative correlation between Structure 1 and 3 and a slight negative correlation between Structure 2 and 3. The values calculated using the alternative method do not obviously reflect the profound dissimilarity between structures and could be misinterpreted.



Laakso and Cottrell method (range -1 to +1)

	1	2	3
1	1	0.8011	-0.4788
2	0.8011	1	-0.0989
3	-0.4788	-0.0989	1

Alternative method (range 0 to +1)

	1	2	3
1	1	0.8084	0.6901
2	0.8084	1	0.6743
3	0.6901	0.6743	1

Analysing the similarity values determined using both approaches highlights the potential for different methods of comparison to provide inconsistent results. Laakso and Cottrell's (2000) method appears to be more sensitive to subtle patterns of variation in structure that are relevant to structural similarity comparisons. This provides a more robust comparison of the relation between the similarities and differences between the relative locations of points in one activation space with those in a corresponding activation space.

Appendix 3.1 - Further details of the colour categorisation ANNs configuration and analysis

My empirical investigation was conducted using three-layer feedforward ANNs comprised of an input layer that is fully connected to a hidden layer which in turn is fully connected to an output layer. The input layers have a unit corresponding to each element of the input dataset and the output layers have one unit for each possible output category or classification (which is standard for pattern recognition or classification networks). The ANNs were created and analysed using the Matlab software environment with the neural network toolbox installed. All of the ANNs are standard pattern recognition networks that include bias values for units in the hidden and output layers. The hyperbolic tangent sigmoid transfer (activation) function (tansig) was used at the hidden layer and provides output in the range -1 to +1. The softmax transfer function was used at the output layer and provides normalised activation values corresponding to the probability distribution over predicted output classes.

The ANNs were trained using scaled conjugate gradient backpropagation (trainscg). The ANN parameters were initialised using the Nguyen-Widrow initialization algorithm (initnw) which assigns initial weight values that are designed to distribute the active region of processing units approximately evenly across a network layers input space but it also includes some degree of randomness and so the initial connection weights of each network are unique. The cross-entropy performance function was used with a minimum performance gradient set to 1e-6. A random selection of 75% of the input set was used for training, 15% for validation and 10% removed to test generalisation after training. The maximum validation fails during training was set to 20 and the maximum total training epochs was 1000. The training of an ANN ceased when either the performance gradient was below 1e-6, the network performance on the validation set decreased more than 20 consecutive times, or the maximum of 1000 training epochs was reached.

Appendix 3.2 - Structural similarity compared using a three dimensional coordinate system

The structure of the Munsell colour space can also be characterised by the location of the component colours in three dimensional space. I developed an alternative quantification of the space by assigning three dimensional coordinates to each discrete colour. This approach allows comparisons including all activation values in the domain rather than just prototypes, however, the coordinates do not directly correspond to the structures of the three different task domains (hue, value, chroma) that the groups of ANNs were trained on. All of the ANNs trained on the reflectance datasets were also compared to this objective characterisation of the Munsell colour space. Each colour sample represented in the input set was designated coordinates based on the location relative to an origin in the centre of the space which corresponds to a colour with no determinate hue, median lightness value and no chroma). The distances between all colour points were calculated to determine a characterisation of the structure of the colour space based on proximity relations. This structure respects some aspects of the hue, value and chroma component domains but the overall colour space has significant differences from the individual domains. For example, a red hue with a value of five and chroma of 12 is maximally distant from a blue/green hue with a value of five and chroma of 12 in the colour space but in the chroma classification domain they would be relatively close together.

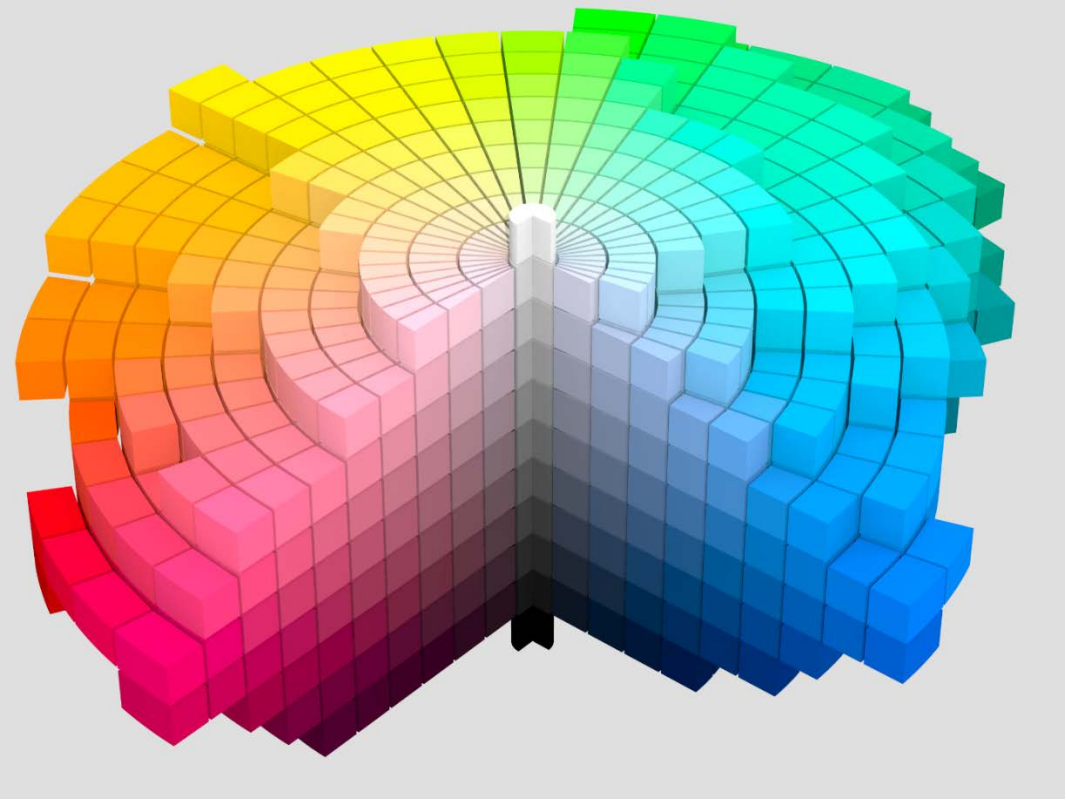


Figure A3.2. Colours in the Munsell colour space can be characterised by three dimensional coordinates.(https://en.wikipedia.org/wiki/Munsell_color_system#/media/File:Munsell_1943_color_solid_cylindrical_coordinates_gray.png).

I compared each colour categorisation ANNs to this objective characterisation of the structure of the Munsell colour space determined by proximity relations between the assigned three dimensional coordinates rather than the individual components of the colours they were trained to classify. The average correlation between the structures of the collective sets of hidden layer activation points and the structure of the corresponding colours in the colour space varied from moderate to fairly high. Prototypical points in the colour space were determined for each category by averaging the coordinates of the members of the category that were present in the reflectance spectra input samples. The prototype points are specific to the categorical distinctions in each of the three colour components and determined by these task relevant distinctions. The average correlation between the structure of the prototype points in the colour space and the corresponding prototypical hidden layer activation points for both hue and value were high to very high (0.85 to 0.95) and for chroma the correlation was moderate to high (0.67 to 0.79). The average correlation between the structure of the hidden layer activation points within each category and the structure of the constituents of the corresponding category in the colour space varied from moderate to high (0.43 to 0.78). However, these results are more difficult to interpret than the prototype comparisons provided in Chapter 3 because the task relevant structural properties are not distinctly categorised. The average correlation values for each group of ANNs are provided in Table A3.2.

ANN training dataset		61 components, 627 samples			421 components, 1269 samples		
Categorisation task		5 Hues	8 Values	7 Chromas	10 Hues	8 Values	7 Chromas
Layer comparison	**Structural comparison**	Average Correlations			Average Correlations		
Hidden Layer - Task Structure	All activation points	0.66	0.16	0.76	0.62	0.15	0.66
Hidden Layer - Task Structure	Category prototypes	0.85	0.95	0.79	0.94	0.95	0.67
Hidden Layer - Task Structure	Intra-categories average	0.54	0.43	0.78	0.49	0.48	0.72

Table A3.2. Average correlations between hidden layer activation structures and the structure of the colour space.

Bibliography

Azhar, F. 2016. Polytopes as vehicles of informational content in feedforward neural networks. *Philosophical Psychology,* 29**,** 697-716.

Boone, W. & Piccinini, G. 2016. The cognitive neuroscience revolution. *Synthese,* 193**,** 1509-1534.

Bornstein, A.M. 2016. Is artificial intelligence permanently inscrutable. *Nautilus*, *40*.

California Facial Expressions (CAFE) dataset, <http://www.cs.ucsd.edu/users/gary/CAFE/>.

Churchland, P. M. 1989. *A neurocomputational perspective: The nature of mind and the structure of science*, MIT press.

Churchland, P. M. 1996. *The engine of reason, the seat of the soul: A philosophical journey into the brain*, MIT Press.

Churchland, P. M. 1998. Conceptual similarity across sensory and neural diversity: The Fodor/Lepore challenge answered. *The Journal of Philosophy,* 95**,** 5-32.

Churchland, P. M. 2007. *Neurophilosophy at work*, Cambridge University Press.

Churchland, P. M. 2012. *Plato's camera: How the physical brain captures a landscape of abstract universals*, MIT Press.

Cottrell, G, Munro, P. & Zipscr, D. 1987. Learning internal representations of gray scale images: An example of extensional programming. In Proc. Ninth Annual Cognitive Science Society Conference, Seattle, Wa.

Cummins, R. 1996. *Representations, targets, and attitudes*, MIT press.

Cummins, R., Blackmon, J., Byrd, D., Lee, A. and Roth, M. 2006. Representation and unexploited content. *Teleosemantics*, pp.195-207.

Demuth, H. & Beale, M. 1992. Neural Network Toolbox. *For Use with MATLAB. The MathWorks Inc,* 2000.

Floridi, L. & Chiriatti, M. 2020. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, *30*, pp.681-694.

Fleming, M.K. and Cottrell, G.W. 1990, June. Categorization of faces using unsupervised feature extraction. In *1990 IJCNN International Joint Conference on Neural Networks* (pp. 65-70). IEEE.

Fodor, J. A. & Pylyshyn, Z. W. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition,* 28**,** 3-71.

Fodor, J. A. & Lepore, E. 1992. *Holism: A Shopper's Guide*, Blackwell.

Fodor, J. & Lepore, E. 1996. Churchland on state space semantics.

Garzón, F.C. 2000. State space semantics and conceptual similarity: reply to Churchland. *Philosophical Psychology*, *13*(1), pp.77-95.

Garzón, F. C. 2003. Connectionist semantics and the collateral information challenge. *Mind & language,* 18**,** 77-94.

Gładziejewski, P. & Miłkowski, M. 2017. Structural representations: causally relevant and different from detectors. *Biology & Philosophy,* 32**,** 337-355.

Hassabis, D., Kumaran, D., Summerfield, C. & Botvinick, M. 2017. Neuroscience-inspired artificial intelligence. *Neuron,* 95**,** 245-258.

Kriegeskorte, N., Mur, M. & Bandettini, P. A. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience,* 2**,** 4.

Kriegeskorte, N. & Kievit, R. A. 2013. Representational geometry: integrating cognition, computation, and the brain. *Trends in cognitive sciences,* 17**,** 401-412.

Krizhevsky, A., Sutskever, I. and Hinton, G.E. 2017. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), pp.84-90.

Laakso, A. & Cottrell, G. W. 2000. Content and cluster analysis: Assessing representational similarity in neural systems. *Philosophical Psychology,* 13**,** 47-76.

Laakso, A. & Cottrell, G. W. 2006. Churchland on connectionism. *Paul Churchland***,** 113.

LeCun, Y., Bengio, Y. and Hinton, G. 2015. Deep learning. *nature*, *521*(7553), pp.436-444.

McCauley, R. N., Churchland, P. S. & Churchland, P. M. 1996. *The Churchlands and their critics*, Blackwell Cambridge, MA.

McClelland, J. L. & Cleeremans, A. 2009. Connectionist models. In Bayne T., Cleeremans A. & Wilken P. (eds.), The Oxford Companion to Consciousness. Oxford University Press.

Munsell, A. H. 1992. Munsell book of color: Matte Finish Collection.

O'Brien, G. & Opie, J. 2004. Notes toward a structuralist theory of mental representation. *Representation in mind: New approaches to mental representation***,** 1-20.

O'Brien, G. & Opie, J. 2006. How do connectionist networks compute? *Cognitive Processing,* 7**,** 30-41.

O'Brien, G. & Opie, J. 2009. The role of representation in computation. *Cognitive processing,* 10**,** 53-62.

O'Brien, G. & Opie, J. 2011. Representation in analog computation. In Newen, A., Bartels, A., Jung, E. (eds.), Knowledge and Representation, CSLI Publications.

Raizada, R. D. & Connolly, A. C. 2012. What makes different people's representations alike: neural similarity space solves the problem of across-subject fMRI decoding. *Journal of cognitive neuroscience,* 24**,** 868-877.

Randall, C., O'Reilly, R.C. and Munakata, Y. 2000. *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. MIT press.

Rogers, T. T. & McClelland, J. L. 2014. Parallel distributed processing at 25: Further explorations in the microstructure of cognition. *Cognitive science,* 38**,** 1024-1077.

McClelland, J.L., Rumelhart, D.E. and PDP Research Group. 1986. *Parallel distributed processing* (Vol. 2, pp. 20-21). Cambridge, MA: MIT press.

McClelland, J.L., Rumelhart, D.E. and PDP Research Group. 1987. *Parallel Distributed Processing, Volume 2: Explorations in the Microstructure of Cognition: Psychological and Biological Models* (Vol. 2). MIT press.Sejnowski, T.J., 2018. *The deep learning revolution*. MIT press.

Shea, N. 2007. Content and its vehicles in connectionist systems. *Mind & Language,* 22**,** 246-269.

Shea, N. VI—Exploitable Isomorphism and Structural Representation. Proceedings of the Aristotelian Society, 2014. The Oxford University Press, 123-144.

Shea, N. 2018. *Representation in cognitive science* (p. 304). Oxford University Press.

Shwartz-Ziv, R. & Tishby, N. 2017. Opening the Black Box of Deep Neural Networks via Information. *arXiv preprint arXiv:1703.00810*.

Tishby, N. and Zaslavsky, N. 2015, April. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)* (pp. 1-5). IEEE.

University of Eastern Finland, S. C. R. G. Munsell colour databases, <https://www.uef.fi/web/spectral/>.

Usher, M. 2001. A statistical referential theory of content: Using information theory to account for misrepresentation. *Mind & Language*, *16*(3), pp.311-334.

Williams, D. & Colling, L. 2018. From symbols to icons: the return of resemblance in the cognitive neuroscience revolution. *Synthese,* 195**,** 1941-1967.

Yale Face Dataset, <http://vision.ucsd.edu/content/yale-face-database>.

Zeiler, M. D. & Fergus, R. 2014. Visualizing and understanding convolutional networks. European conference on computer vision. Springer, 818-833.