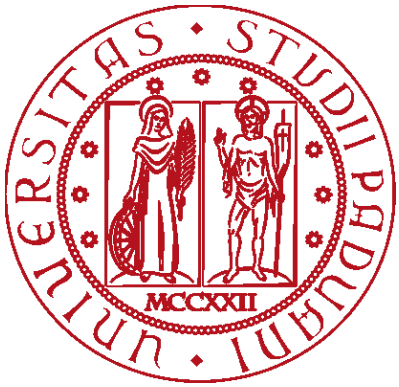

UNIVERSITY OF PADOVA

Department of General Psychology

Bachelor's Degree Course in Psychological Science

Final Dissertation

Comparing Primate's Ventral Visual Stream and the State-of-the-art Deep Convolutional Neural Networks for Core Object Recognition

Supervisor
Prof. Marco Zorzi

Candidate:
Charles Wani Victor Ladu
Student ID Number:
1222489

Academic Year
2022/2023

ABSTRACT

Our ability to recognize and categorize objects in our surroundings is a critical component of our cognitive processes. Despite the enormous variations in each object's appearance (Due to variations in object position, pose, scale, illumination, and the presence of visual clutter), primates are thought to be able to quickly and easily distinguish objects from among tens of thousands of possibilities. The primate's ventral visual stream is believed to support this view-invariant visual object recognition ability by untangling object identity manifolds. Convolutional Neural Networks (CNNs), inspired by the primate's visual system, have also shown remarkable performance in object recognition tasks. This review aims to explore and compare the mechanisms of object recognition in the primate's ventral visual stream and state-of-the-art deep CNNs. The research questions address the extent to which CNNs have approached human-level object recognition and how their performance compares to the primate ventral visual stream. The objectives include providing an overview of the literature on the ventral visual stream and CNNs, comparing their mechanisms, and identifying strengths and limitations for core object recognition. The review is structured to present the ventral visual stream's structure, visual representations, and the process of untangling object manifolds. It also covers the architecture of CNNs. The review also compared the two visual systems and the results showed that deep CNNs have shown remarkable performance and capability in certain aspects of object recognition, but there are still limitations in replicating the complexities of the primate visual system. Further research is needed to bridge the gap between computational models and the intricate neural mechanisms underlying human object recognition.

Keywords: Visual system, DNNs, object recognition, IT cortex

TABLE OF CONTENTS

ABSTRACT i

LIST OF FIGURES ii

INTRODUCTION 1

CHAPTER 1: VENTRAL VISUAL STREAM 4

1.1. Structure 4

1.2. Visual Representations 5

1.3. Untangling Object Manifolds 6

CHAPTER 2: DEEP CONVOLUTIONAL NEURAL NETWORKS 7

CHAPTER 3: COMPARING VENTRAL VISUAL STREAM AND DEEP CONVOLUTIONAL NEURAL NETWORKS 9

Methods: 9

Results: 9

CONCLUSION 21

BIBLIOGRAPHY 23

LIST OF FIGURES

FIGURE 1: NEURONAL POPULATIONS ALONG THE VENTRAL VISUAL STREAM (DICARLO & COX, 2007). 5
FIGURE 2: OBJECT UNTANGLING (DICARLO & COX, 2007). 7
FIGURE 3: CNN ARCHITECTURE (KRIZHEVSKY ET AL., 2012). 8
FIGURE 4: V4 NEURAL RESPONSE PREDICTIONS (YAMINS ET AL., 2014). 10
FIGURE 5: IT NEURAL RESPONSE PREDICTIONS (YAMINS ET AL., 2014). 10
FIGURE 6: KERNEL ANALYSIS CURVES OF NEURAL AND MODEL REPRESENTATIONS (CADIEU ET AL., 2014). 12
FIGURE 7: KERNEL ANALYSIS CURVES OF SAMPLE AND NOISE-MATCHED NEURAL AND MODEL REPRESENTATION (CADIEU ET AL., 2014). 13
FIGURE 8: EFFECT OF SAMPLING THE NEURAL AND NOISE-CORRECTED MODEL REPRESENTATIONS (CADIEU ET AL, 2014). 13
FIGURE 9: LINEAR-SVM GENERALIZATION PERFORMANCE OF NEURAL AND MODEL REPRESENTATIONS (CADIEU ET AL, 2014). 14
FIGURE 10: NEURAL AND MODEL REPRESENTATION PREDICTIONS OF IT MULTI-UNIT RESPONSES (CADIEU ET AL, 2014). 14
FIGURE 11: OBJECT-LEVEL REPRESENTATIONAL SIMILARITY BETWEEN MODEL AND NEURAL REPRESENTATIONS (CADIEU ET AL 2014). 15
FIGURE 12: OBJECT CATEGORIZATION (KUBILIUS ET AL., 2016). 17
FIGURE 13: MODEL PREFERENCE FOR SHAPE (KUBILIUS ET AL., 2016). 17
FIGURE 14: MODEL PREFERENCE FOR SHAPE (KUBILIUS ET AL., 2016). 18
FIGURE 15: EXAMPLES OF GEONS (KUBILIUS ET AL., 2016). 18
FIGURE 16: CATEGORICAL REPRESENTATIONS IN CNNS (KUBILIUS ET AL., 2016). 19
FIGURE 17: OBJECT LEVEL COMPARISON TO HUMAN BEHAVIOR (RAJALINGHAM ET AL., 2018). 20
FIGURE 18: IMAGE LEVEL COMPARISON TO HUMAN BEHAVIOR (RAJALINGHAM ET AL., 2018). 21

INTRODUCTION

Our ability to recognize and categorize objects in our surroundings is a critical component of our cognitive processes. Despite the enormous variations in each object's appearance (Due to variations in object position, pose, scale, illumination, and the presence of visual clutter), primates are thought to be able to quickly and easily distinguish objects from among tens of thousands of possibilities " (DiCarlo & Cox, 2007; DiCarlo et al., 2012). According to DiCarlo et al. (2012), this ability to quickly report the identity or category of an object after only a brief glimpse of visual input is called "core object recognition" which they considered to be at the heart of the brain's recognition system. Due to variations in object position, pose, scale, illumination, and the presence of visual clutter, any single object can produce an endless number of distinct images on the retina, making object recognition computationally challenging. According to Pinto et al. (2008), "This invariant problem is regarded as the computational crux of recognition". However, it is believed that the primate's ventral visual stream supports this view-invariant visual object recognition ability (Tanaka, 1996; DiCarlo et al., 2012). But, how the brain solves this problem is still unknown. DiCarlo and colleagues hypothesize that the ventral stream gradually "untangles" information about object identity (DiCarlo & Cox, 2007; DiCarlo et al, 2012).

The primate's ventral visual stream comprises a series of interconnected brain regions, starting from the primary visual cortex (V1), and ending in the inferotemporal cortex (IT). According to various researchers (Felleman & Van Essen, 1991; Riesenhuber & Poggio, 1999; DiCarlo & Cox, 2007; DiCarlo et al., 2012), the ventral visual stream processes visual information in a hierarchical manner, with each region specializing in detecting specific visual features, such as edges, and shapes, and producing increasingly complex representations of objects. V1 is the lowest area which processes simple features such as lines and edges, V2 intermediate, and V4 and IT are the highest areas which process high-level features such as object parts and objects, and support object recognition. Lesion studies in monkeys have demonstrated that damage to the IT cortex impairs object recognition (Holmes & Gross, 1984).

Deep convolutional neural networks (CNNs) are a type of artificial neural network that has been shown to be highly effective in object recognition tasks (Krizhevsky et al., 2012). These networks consist of multiple layers of artificial neurons that are trained to identify and classify objects based on their features. Convolutional neural networks (CNNs) are designed to mimic the processing of the primate ventral visual stream, with each layer of the network

processing increasingly complex visual features. They are inspired by the architecture of the primate's visual system, with multiple layers of processing units that extract features and categorize objects (Krizhevsky et al., 2012). They are composed of multiple layers of interconnected neurons, each layer is responsible for different levels of abstraction. The first layer detects simple features such as edges, while subsequent layers detect increasingly complex features such as object parts and textures. The final layer of a DCNN is responsible for categorizing the input image based on its features. DCNNs are trained using supervised learning on large datasets of labeled images. Recent studies have shown that deep CNNs can achieve impressive performance in object recognition tasks. For example, Krizhevsky et al. (2012) introduced the AlexNet architecture, which achieved top performance in the 2012 ImageNet Large-Scale Visual Recognition Challenge. Since then, numerous studies have built upon this architecture, achieving even better performance in object recognition tasks (He et al., 2016).

This review aims to explore and compare the mechanisms of object recognition in the primate's ventral visual stream and compare them to the state-of-the-art DCNNs. Comparing the mechanisms of object recognition in the primate's ventral visual stream and DCNNs can provide valuable insights into the cognitive processes involved in object recognition and the development of more advanced artificial intelligence systems.

The main research questions of this review are:

1- To what extent have recent state-of-the-art deep convolutional neural network models approached human-level object recognition?
2- How does the performance of deep CNNs compare to the performance of the primate ventral visual stream in core object recognition tasks?

To answer these questions, this review aims to achieve the following objectives:

1- To provide an overview of the literature on primates' ventral visual stream and deep convolutional neural networks.
2- To compare and contrast the mechanisms underlying primates' ventral visual stream and deep convolutional neural networks.
3- To identify the strengths and limitations of each approach for core object recognition.

This review is structured as follows: a brief overview of the primate's ventral visual stream, followed by a brief overview of the deep convolutional neural networks. Then a comparison between the primate's ventral visual stream and deep convolutional neural networks is presented, then followed by key conclusions.

CHAPTER 1: VENTRAL VISUAL STREAM

1.1. Structure

The primate's ventral visual stream is believed to house important circuits that underlie object recognition behavior. It's divided into different visual areas (V1, V2, V4 and IT) based on anatomical connectivity patterns, different anatomical structures and retinotopic mapping (Felleman & Van Essen, 1991). Most of the visual field for areas V1, V2, and V4 have complete retinotopic maps and each area can be seen as conveying a population-based re-representation of visually presented images (see Fig. 1). However, within the IT complex, crude retinotopy exists over the posterior portion (Boussaoud et al, 1991), but not in the central and anterior regions (Felleman & Van Essen, 1991). Stoerig and Cowey (1997) found that lesions in the posterior ventral stream can result in total blindness in that area of the visual field, and Holmes and Gross (1984) found that lesions or inactivation of anterior regions, particularly the inferior temporal cortex (IT) can cause a selective deficits in the ability to distinguish between complex objects. The ventral visual stream is thought of as a series of hierarchical processing stages that encode image content (such as object identity and category) increasingly explicitly in successive cortical areas (DiCarlo & Cox, 2007; DiCarlo et al., 2012; Felleman & Van Essen, 1991; Riesenhuber & Poggio, 1999). For instance, V1 neurons are defined as Gabor-like edge detectors that extract rough object outlines (Carandini et al, 2005), despite the fact V1 population does not exhibit robust tolerance to complex image transformations (DiCarlo et al, 2012). The inferior temporal (IT) cortex, the highest processing stage of the ventral stream, can, nevertheless, directly enable real-time, invariant object categorization (Hung et al., 2005; Rust & DiCarlo, 2010). The Midlevel ventral area (V4) consistently shows intermediate levels of object selectivity and variation tolerance (Rust & DiCarlo, 2010).

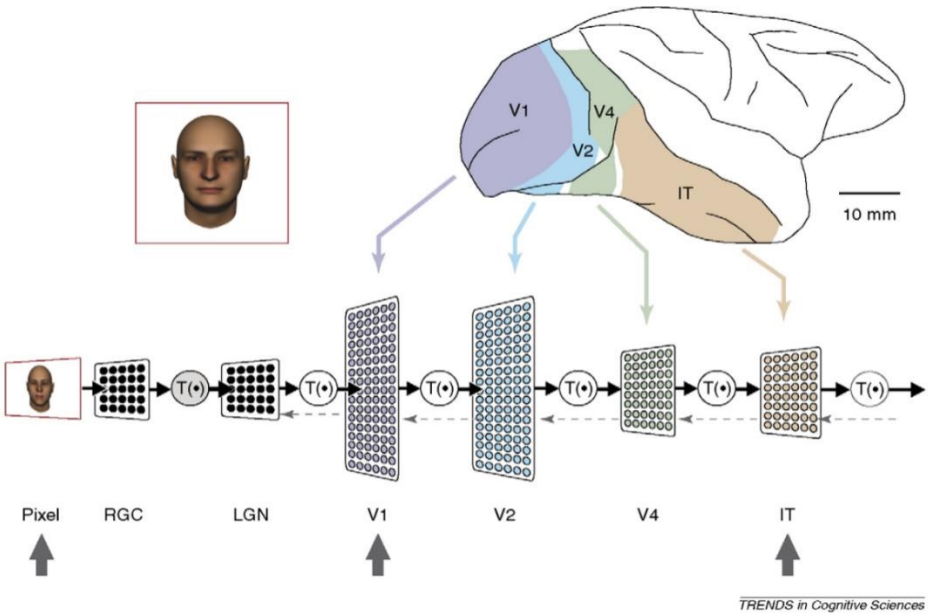

Figure 1: **Neuronal populations along the ventral visual stream (DiCarlo & Cox, 2007).** Here is shown a lateral schematic of a rhesus monkey brain.

1.2. Visual Representations

The inferior temporal cortex (IT), the highest level of the ventral visual stream in both humans and other primates, is assumed to be where visual representations are created (Tanaka, 1996; Logothetis & Sheinberg, 1996). Numerous individual IT neurons have been found to selectively respond to specific object classes, such as faces or other complex shapes, while also displaying some tolerance for changes in object size and position, pose, illumination, and low-level shape cues (Serre et al., 2007). Hung et al., (2005), focused on characterizing the initial wave of neuronal population images that are successively produced along the ventral visual stream as the retinal image is altered and re-represented on its way to IT (see Fig. 1) to better understand how the ventral stream represents visual information. They discovered that at least some core recognition can be supported by population representations in IT. In particular, despite changes in object position and size, simple linear classifiers can quickly (within <300 ms from image onset) and reliably determine an object's category by reading the firing rates of an IT population of ~200 neurons (Hung et al., 2005). It is useful to consider the elements that make up such a representation even though it is unclear how the ventral visual stream creates this powerful form of visual representation. In other words, a good representation is one that makes it easy to determine a presence of an object or a face by simply putting a linear decision function such as a hyperplane between the representations of two objects or faces (see Fig.

2b). A bad representation, on the other hand, is one where it is impossible to reliably distinguish between two representations using a linear decision function (see Fig. 2c).

1.3. Untangling Object Manifolds

According to DiCarlo et al., (2012), if we think of a population of neurons' response to a specific view of an object as a response vector in a space whose dimensionality is determined by the neurons in the population, an object undergoing an identity-preserving transformation results in a different pattern of population activity that corresponds to a different response vector. As a result, response vectors corresponding to all possible identity-preserving transformations create a low-dimension in a high-dimensional space which is defined as "an object identity manifold". To put it another way, a response of a population of neurons to an image can be thought of as a point in a high-dimensional space where each axis represents a response level of each neuron, and all the possible identity-preserving transformations of an object will form a low-dimensional manifold of points in the population vector space (see Fig. 2a). Therefore, each object identity manifold will be significantly curved for neurons with small receptive fields, such as retinal ganglion cells, and V1 cells (see Fig. 2c). Additionally, the manifolds corresponding to different objects will become tangled (see Fig. 2d). At Higher levels of visual processing, object identity manifolds are more flat, separated, or untangled as neurons tend to maintain their selectivity for objects despite changes in view (see Fig. 2b). Therefore, it is believed that object manifolds are gradually untangled by non-linear selectivity and the application of invariance computations at each stage of the ventral visual stream's processing (DiCarlo et al, 2012). As a result, a linear decision function, like a hyperplane, can be used to simply separate or differentiate one object's manifold from all other object manifolds.

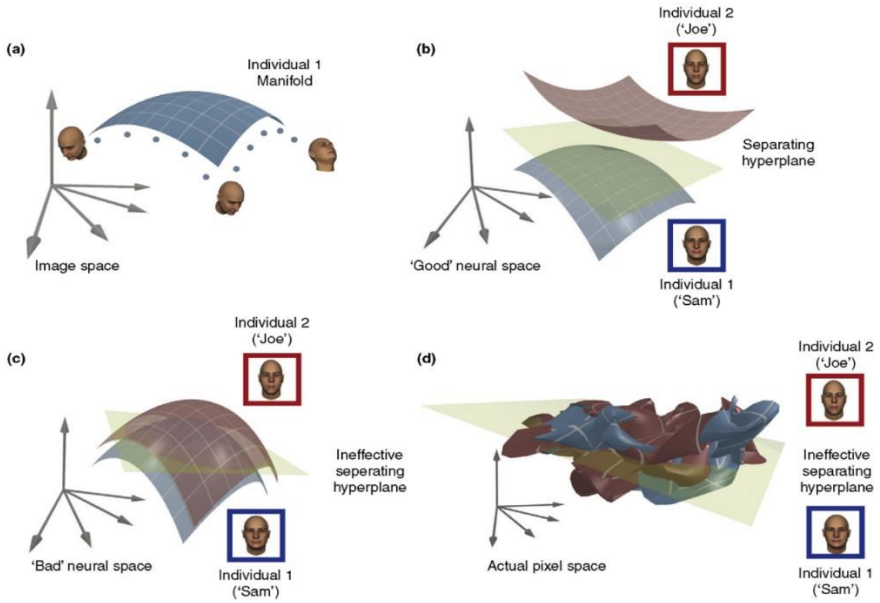

Figure 2: Object Untangling (DiCarlo & Cox, 2007). Here is shown object untangling. In a neuronal population space, each axis is one neuron's activity and the dimensionality of the space is equal to the number of neurons.

CHAPTER 2: DEEP CONVOLUTIONAL NEURAL NETWORKS

Convolutional Neural Networks are artificial neural networks that use convolutional layers to extract features from images and have become a powerful tool for image recognition and computer vision tasks. The first convolutional neural networks (CNNs) were inspired by the work of Hubel and Wiesel (Hubel & Wiesel, 1962; 1968), who studied the primary visual cortex of cats and found that simple cells respond to lines with a specific orientation, while complex cells have a larger receptive field and are more location invariant. This work led to the development of biologically inspired computational models for visual recognition, including the Neocognitron (Fukushima, 1980) and the HMAX (Riesenhuber & Poggio, 1999), which are early examples of CNNs. The Neocognitron consists of a cascade connection of modular structures that mimic the organization of the visual cortex and has a degree of location invariance and can recognize features in different positions. The HMAX model is a hierarchical feedforward architecture, it consists of alternating layers of simple and complex cells that detect increasingly complex features. The HMAX model has shown good accuracy in image recognition and can achieve close to human-level performance on rapid object recognition tasks (Chikkerur & Poggio, 2011). However, these early models consisted of only a few layers, which made them less effective at recognizing complex patterns in images.

The modern convolutional neural networks were also inspired by the structure and function of the primate ventral visual stream (Hubel & Wiesel, 1962; 1968), and they process information through a deep hierarchy of representations (typically 5 to 20 layers, hence the name deep Convolutional Neural Networks) to enable the recognition of object categories. The CNNs have become increasingly accurate over time, and are now capable of achieving close to human-level performance on some rapid object recognition tasks (Krizhevsky et al., 2012). The architecture of CNNs is composed of mainly three types of layers: convolutional, pooling, and fully-connected (see Fig. 3). The convolutional layer applies filters to input data to extract relevant features, while the pooling layer reduces the dimensionality of the features, and the fully-connected layer performs classification based on the extracted features. In CNNs, a Rectified Linear Unit (ReLU) is a commonly used non-linear activation function that introduces non-linearity in one or more layers. This helps the network to learn complex features and patterns from the input images. These networks are trained using a backpropagation algorithm, which enables them to learn from their errors and increase their accuracy over time. For instance, LeCun et al. (1989) demonstrated that Convolutional Neural Networks (CNNs) trained with supervised learning using the backpropagation algorithm could accurately classify handwritten digits. However, it wasn't until when an 8-layer CNN, called AlexNet (Krizhevsky et al., 2012), trained with backpropagation, surpassed the state-of-the-art performance on the ImageNet challenge, that CNNs gained widespread attention. Since then, many different CNN architectures have been explored, varying parameters such as network depth, pooling layer placement, number of feature maps, and training procedures (He et al., 2016).

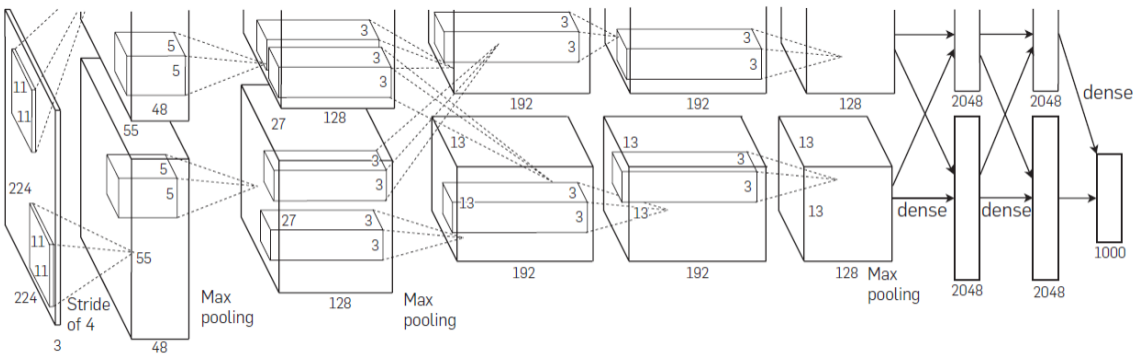

Figure 3: CNN Architecture (Krizhevsky et al., 2012). Here is shown the architecture of CNN with different types of layers.

CHAPTER 3: COMPARING VENTRAL VISUAL STREAM AND DEEP CONVOLUTIONAL NEURAL NETWORKS

Since convolutional neural networks were first developed, their performance and complexity have improved compared to earlier models of object recognition. But it's unclear whether these new models are close to achieving object recognition representation performance that is comparable to that shown in the IT cortex, or to put it another way, "is a question that has yet to be answered". Drawing from five studies (Yamins et al., 2014; Cadieu et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014; Kubilius et al., 2016; Rajalingham et al., 2018), here this review try to examine if these models are capable of achieving object recognition performance somewhat identical to the primates'. A detail account on all the findings in the five studies that compare the representational performance of IT cortex to convolutional neural networks are given below.

Methods:

A comprehensive literature search was conducted to identify relevant studies published between 2012-2022, with a focus on studies comparing the performance of deep CNNs and the primate ventral visual stream in object recognition tasks. The search was conducted through Scopus, Web of Science and Google Scholar. Five studies were selected that investigated the performance of deep CNNs and the primate ventral visual stream in object recognition tasks. The selected studies were critically evaluated based on their research design, methodology, and key findings.

Results:

Yamins et al., (2014), used high-throughput computational techniques to identify a neural network model that matches human performance on challenging object categorization tasks and found that a model's categorization performance is strongly correlated with its ability to predict individual neural unit responses and they identified a high-performing neural network model (HMO or hierarchical modular optimization) that matches human performance on object recognition. The authors also found that even though the model was not explicitly designed to match neural data, its output layer accurately predicts neural responses in the highest ventral visual area (IT cortex) (see Fig. 5) and its middle layers predict responses in the

midlevel visual area (V4 cortex)(see Fig. 4), suggesting that top-down performance constraints shape intermediate visual representations.

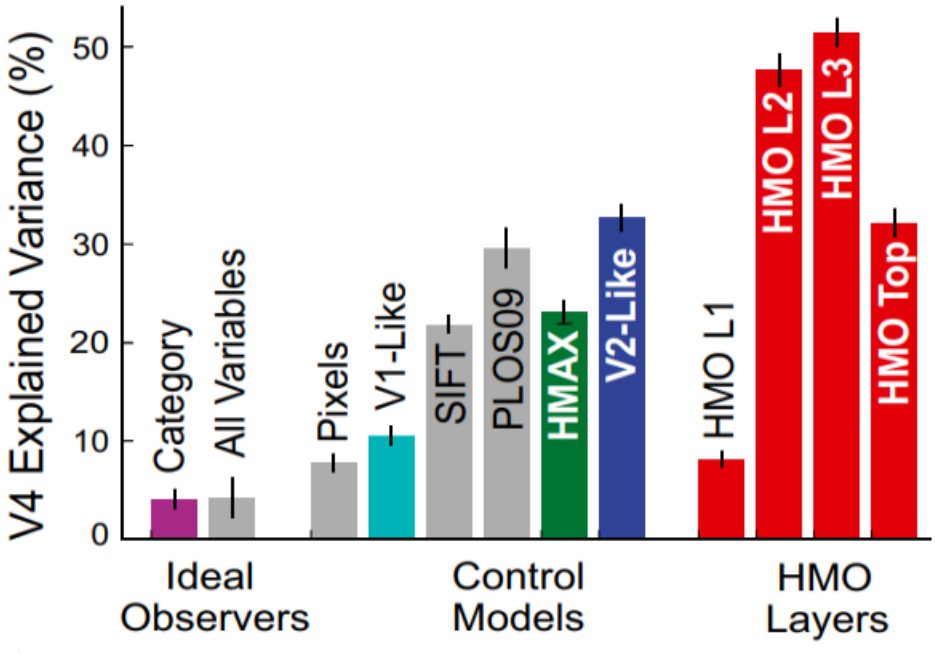

Figure 4: V4 neural response predictions (Yamins et al., 2014). The figure shows the HMO's middle-layer (L3) accurately predicts neural responses in V4.

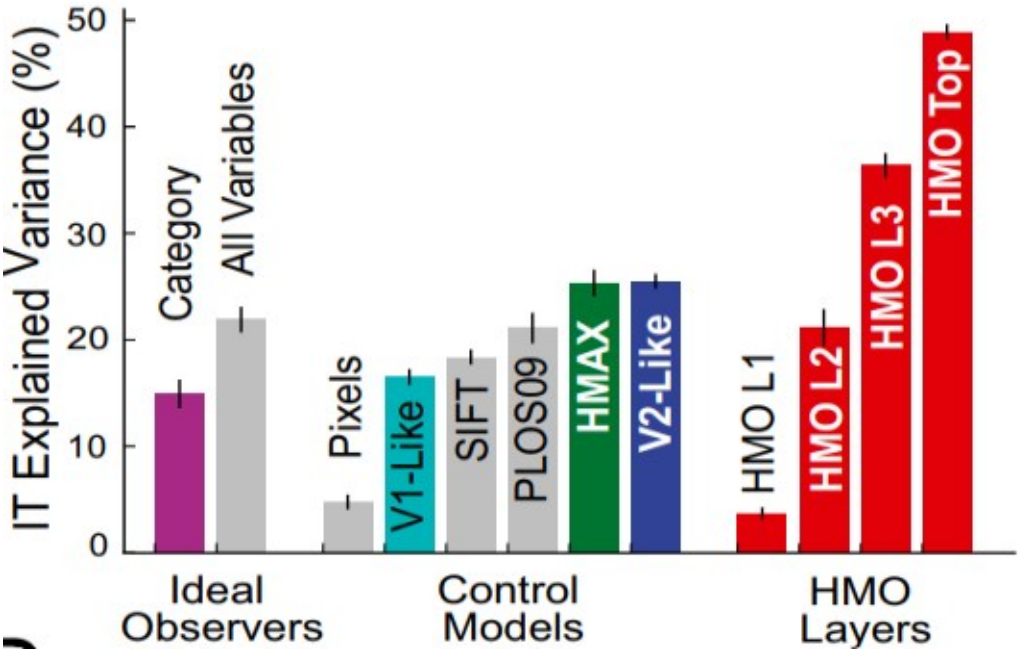

Figure 5: IT neural response predictions (Yamins et al., 2014). The figure shows the HMO's top-layer accurately predicts neural responses in IT.

Cadieu et al. (2014), used a novel extension of kernel analysis (Montavon; Braun; and Muller, 2011) that measures precision as a function of complexity and compared the performance of three deep convolutional neural networks: models described in Krizhevsky et al. 2012, Zeiler & Fergus 2013, and Yamins et al. 2014 or HMO, and three other biologically relevant representations including V1-like, V2-like, and HMAX models to the neural representations of the primate IT cortex. Before comparing the neural and model representations, the authors evaluated the absolute representational performance of various models on a computationally difficult task. They found that the V1-like and V2-like models performed poorly on the task, while the HMAX model performed slightly better. However, the three convolutional deep neural networks (DNNs) outperformed all other models. The Zeiler & Fergus 2013 model performed the best, followed by the Krizhevsky et al. 2012 model and the HMO model (see Fig. 6). To compare the performance of the IT neural representation to model representations, the authors took steps to ensure a fair comparison. They fixed the number of neural samples and model features and added noise to the model representations to correct for experimental noise. They estimated an experimental neural noise model based on the observation that spike counts of neurons are approximately Poisson and used this to produce noise-matched model representations for comparison using kernel analysis. They used kernel analysis curves to evaluate the precision of these representations and found that the IT neural representation performed better than the V4 neural representation, and was only matched by the top-performing DNN of Zeiler & Fergus 2013 after correcting for sampling and noise (see Fig. 7a). This suggests that the IT neural representation is quite competitive and effective for object recognition, despite the limitations of the recordings. The authors compared the performance of models and neural representations for single-unit neural recordings. Due to increased noise and fewer trials, the single-unit noise and sample-corrected model representations achieve lower precision than multi-unit noise and sample correction. The analysis shows that the single-unit IT representation performs better than the HMO representation, slightly worse than the Krizhevsky et al. 2012 representation, and is outperformed by the Zeiler & Fergus 2013 representation (see Fig. 7b). The authors computed the area under the curve from kernel analysis curves to obtain a summary number, omitting models with near-zero performance (i.e. V1-like, V2-like, and HMAX). They vary the number of multi-unit recording samples and features, and correct for neural noise by adding a matched neural noise level to the model representations and found that the Zeiler & Fergus 2013 deep neural network (DNN) performs similarly to the IT multi-unit representation, and both the

Krizhevsky et al. 2012 and Zeiler & Fergus 2013 DNNs outperform the IT single-unit representation after correcting for noise and sampling effects (see Fig. 8a). These findings are particularly interesting because they suggest that the DNNs can perform well even with limited training examples, where a simpler representation is necessary for generalization. The authors measured linear-SVM (support vector machine) generalization performance of neural and model representations and found that the Zeiler & Fergus 2013 representation achieved generalization comparable to the IT multi-unit neural sample for a simple linear decision boundary, which was consistent with the results obtained from kernel analysis (see Fig. 9). The authors measured the performance of the model representations as encoding models of the IT multi-unit responses and found that the Krizhevsky et al. 2012 and the Zeiler & Fergus 2013 DNNs achieved higher prediction accuracies than the HMO model (see Fig. 10). However, no model was able to fully account for the explainable variance in the IT multi-unit responses. The authors used representational similarity analysis (Kriegeskorte et al., 2008) to measure how similar the two representations are, they computed object-level representational dissimilarity matrices (Kriegeskorte et al., 2008) for model and neural representations and measured the Spearman rank correlations between the model-derived RDM and the IT multi-unit RDM. The results showed that there remains a gap between DNN models and IT representation when measured with object-level representational similarity (see Fig. 11a).

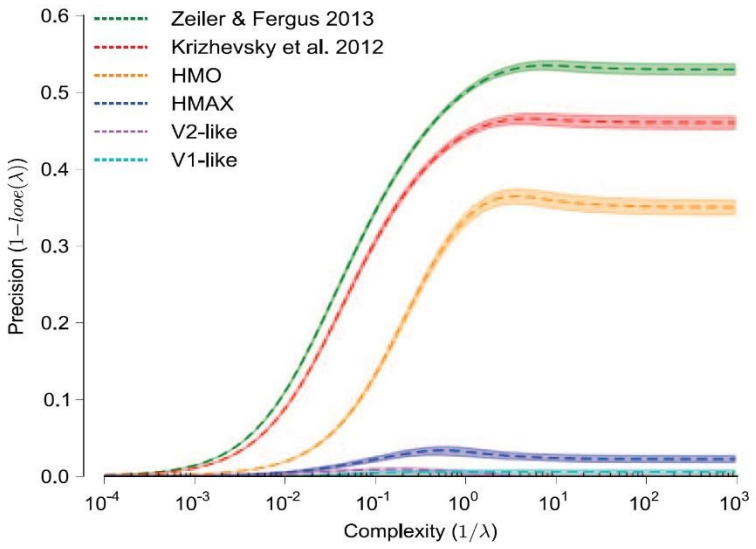

Figure 6: kernel analysis curves of neural and model representations (Cadieu et al., 2014). The figure shows precision of different models in performing object recognition tasks. The curves plot precision against complexity.

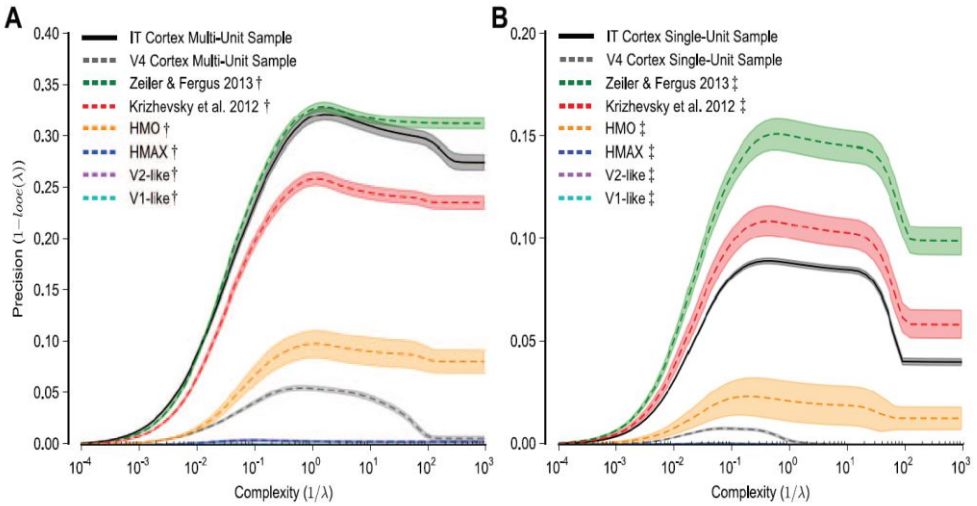

Figure 7: Kernel analysis curves of sample and noise-matched neural and model representation (Cadieu et al., 2014). The figure shows deep neural networks perform better than the V4 cortex sample, and some model representations rival or surpass the IT cortex representation in both multi-unit and single-unit analysis.

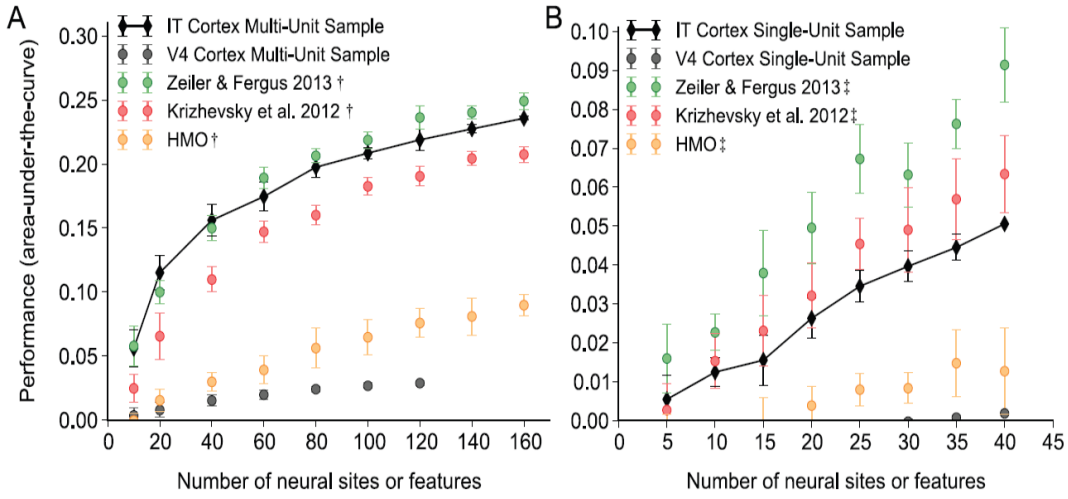

Figure 8: Effect of sampling the neural and noise-corrected model representations (Cadieu et al, 2014). The figure shows deep neural networks outperform the V4 cortex sample, and the Zeiler & Fergus 2013 representation outperforms the IT cortex representation in comparison to the measured sample for both multi-unit and single-unit analysis.

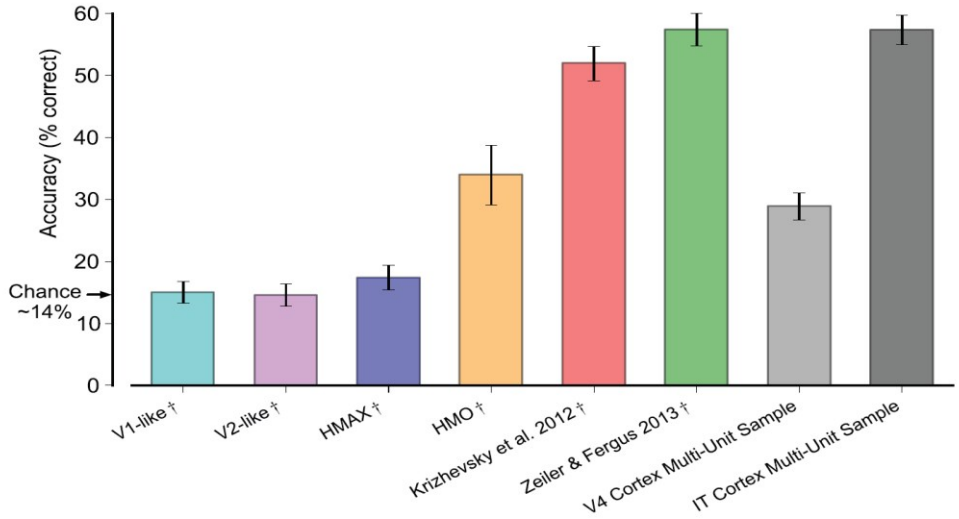

Figure 9: Linear-SVM generalization performance of neural and model representations (Cadieu et al, 2014). The figure shows the Zeiler & Fergus 2013 deep neural network (DNN) representation achieves comparable performance to the IT multi-unit neural representation, and both outperform the Krizhevsky et al. 2012 DNN representation.

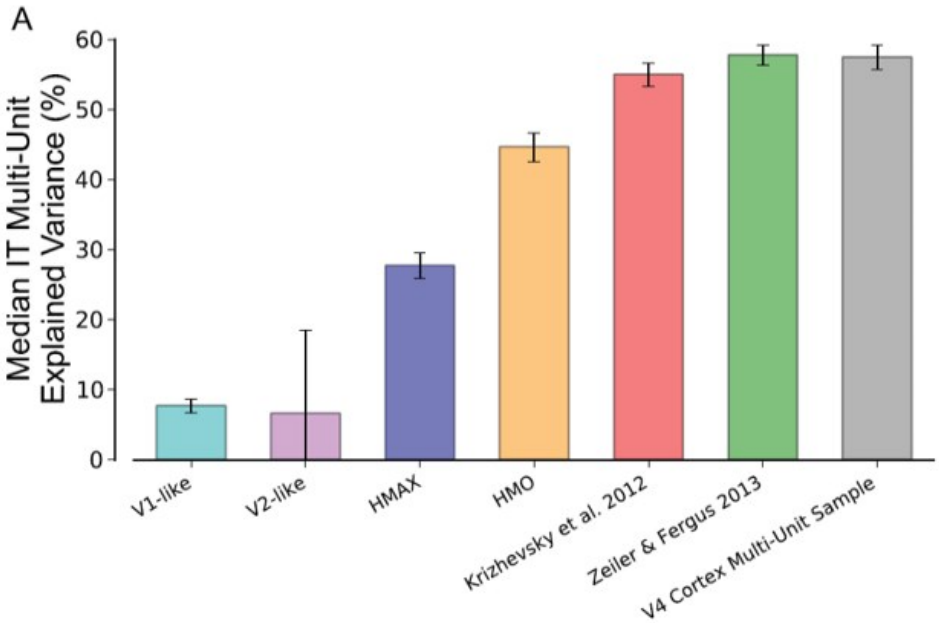

Figure 10: Neural and model representation predictions of IT multi-unit responses (Cadieu et al, 2014). The figure shows the Zeiler & Fergus 2013 DNN achieved comparable performance to the IT cortex multi-unit representation.

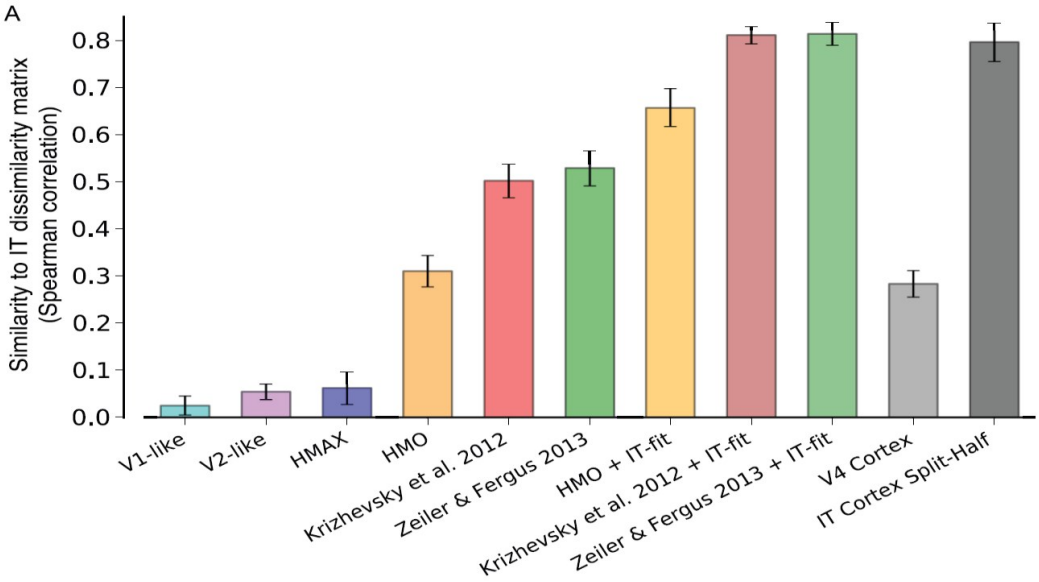

Figure 11: Object-level representational similarity between model and neural representations (Cadieu et al 2014). The figure shows the most recent DNNs offer compelling models of primate object recognition representations that rival with the representational performance of IT cortex.

Khaligh-Razavi and Kriegeskorte (2014), investigated how well computational models can explain the representation of visual objects in the inferior temporal (IT) cortex of humans and monkeys. They looked at the models' ability to categorize objects and how well their internal representations matched the IT cortex's representational geometry. They included some biologically inspired models such as HMAX, and several models from computer vision such as DCNNs. The authors found that better-performing models were more similar to IT representations in terms of their categorization performance and representational dissimilarities. They also found that models trained through supervised learning, such as deep convolutional neural networks, performed better in categorizing objects and explaining the IT representational geometry than unsupervised models.

Kubilius et al., (2016), conducted several experiments. In experiment 1, they evaluated the ability of convolutional neural networks (CNNs) to recognize objects based on their shape. The experiment uses a stimulus set of common objects presented in different formats, including color images, grayscale images, and silhouettes (see Fig. 12a), and compares the performance of CNNs and human observers in recognizing the objects. The results show that CNNs can maintain a robust and accurate performance even when all non-shape cues are removed (see Fig. 12b), demonstrating their ability to extract perceptually relevant shape dimensions for

object recognition. In experiment 2, the authors investigated whether convolutional neural networks (CNNs) develop representations that capture perceived shape dimensions rather than physical form. They used a stimulus set of novel shapes with orthogonal physical and perceived dimensions (see Fig. 13a) and found that deep models tended to cluster stimuli based on their perceived similarity (see Fig. 13b), whereas shallow models were better at capturing physical dissimilarity (see Fig. 13c). The correlation analysis showed that deep models captured perceived shape better than shallow models, and shape preference gradually increased throughout all CNNs layers (see Fig. 13d). The authors also tested whether deeper neural networks are better at capturing perceptual similarities between letters from different fonts (see Fig. 14a) and found that deep models were more sensitive to differences between fonts and captured perceived similarity of letters significantly better than shallow models (see Fig. 14d). The results suggest that deeper networks have a general property of forming perceptually-relevant representational spaces and are more robust than other models in certain scenarios. In experiment 3, the authors tested the sensitivity for non-accidental properties using a stimulus set of geon triplets (see Fig. 15a) based on the Recognition-by-Components (RBC) theory (Biederman et al., 1987) which suggests that object recognition is based on shape properties known as non-accidental, which remain invariant under natural variations such as lighting and viewpoint and the theory predicts that humans perceive changes in non-accidental properties more readily than equivalent changes in metric properties. Kubilius et al. (2016), found that deep neural networks showed a higher than-chance performance in discriminating between stimuli, with deeper networks performing slightly better than shallower ones (see Fig. 15b). In experiment 4, Kubilius et al., (2016), investigated the extent to which CNNs capture semantic human category judgments. They found that deep models captured some semantic structure in the stimuli, but the learned representations in CNNS are largely based on shape and not category (see Fig. 16c,d).

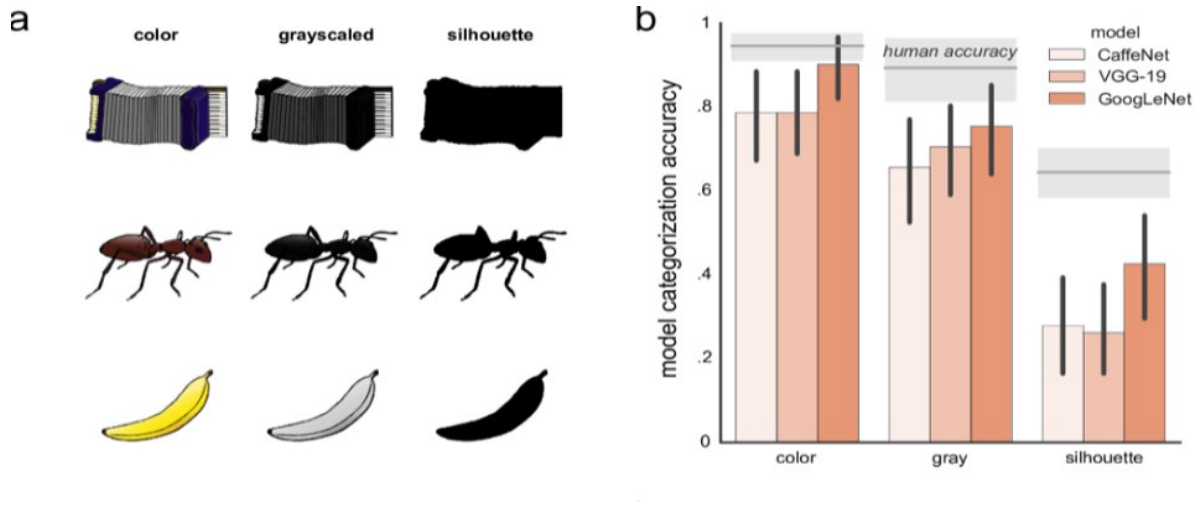

Figure 12: Object Categorization (Kubilius et al., 2016). a) shows examples of objects used in object categorization. b) shows the deep neural networks can represent perceived shape and achieve high accuracy in object recognition.

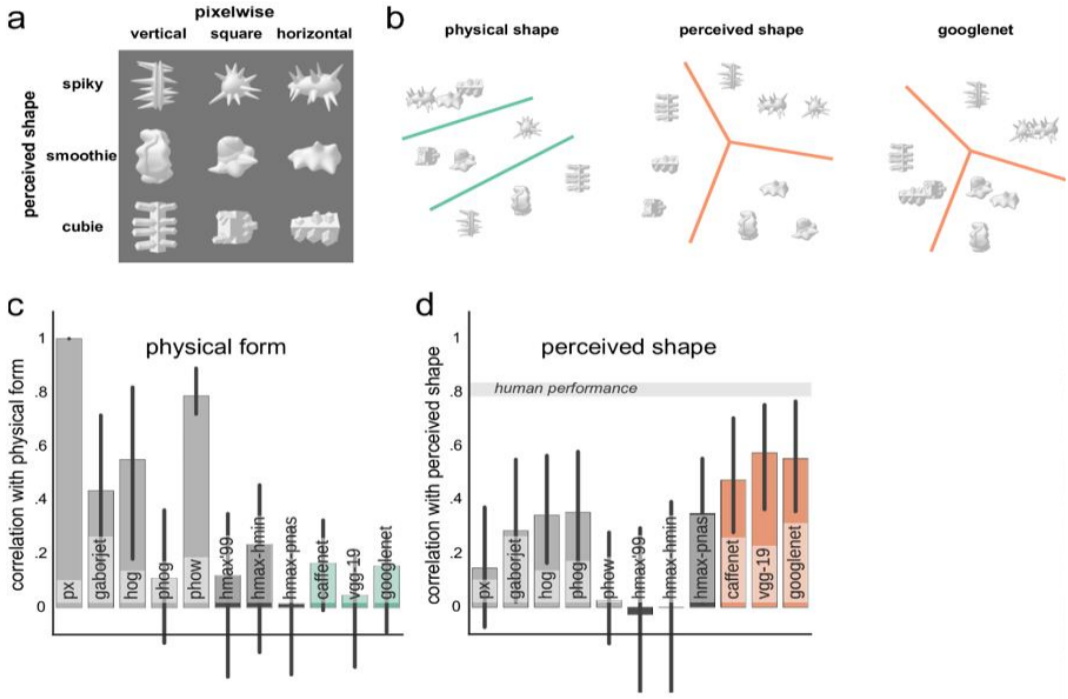

Figure 13: Model preference for shape (Kubilius et al., 2016). The figure shows (a) stimulus set of shapes with orthogonal physical and perceived dimensions, (b) the deep models tended to cluster stimuli based on their perceived shape rather than physical shape, while (c) shallow models were better at capturing physical similarity, and (d) the deep models captured perceived shape better than shallow models.

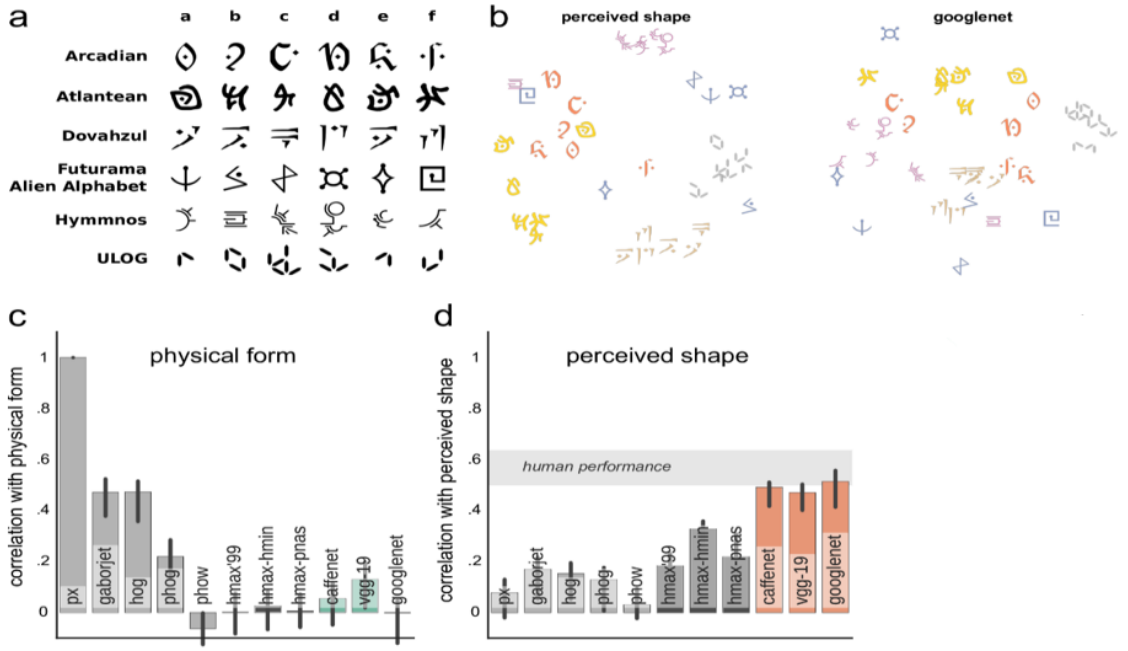

Figure 14: Model preference for shape (Kubilius et al., 2016). The figure shows a) stimulus set of six letters from six constructed fonts, b) comparison of the model outputs to human shape judgments, c) shallow models are better at capturing physical similarity, and d) deep models are better at capturing perceived shape similarity than shallow models.

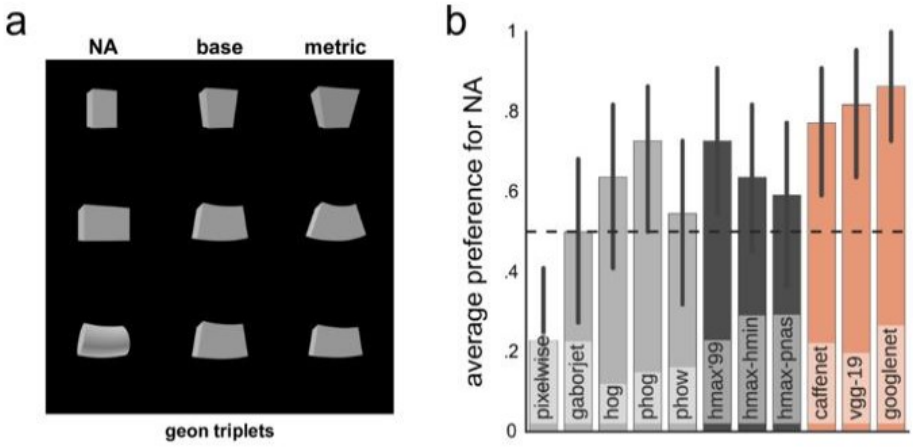

Figure 15: Examples of geons (Kubilius et al., 2016). The figure shows a) examples of geon triplets, and b) the model's performance in discriminating between stimuli with metric and non-accidental variants. Deeper networks tend to perform better than shallower ones.

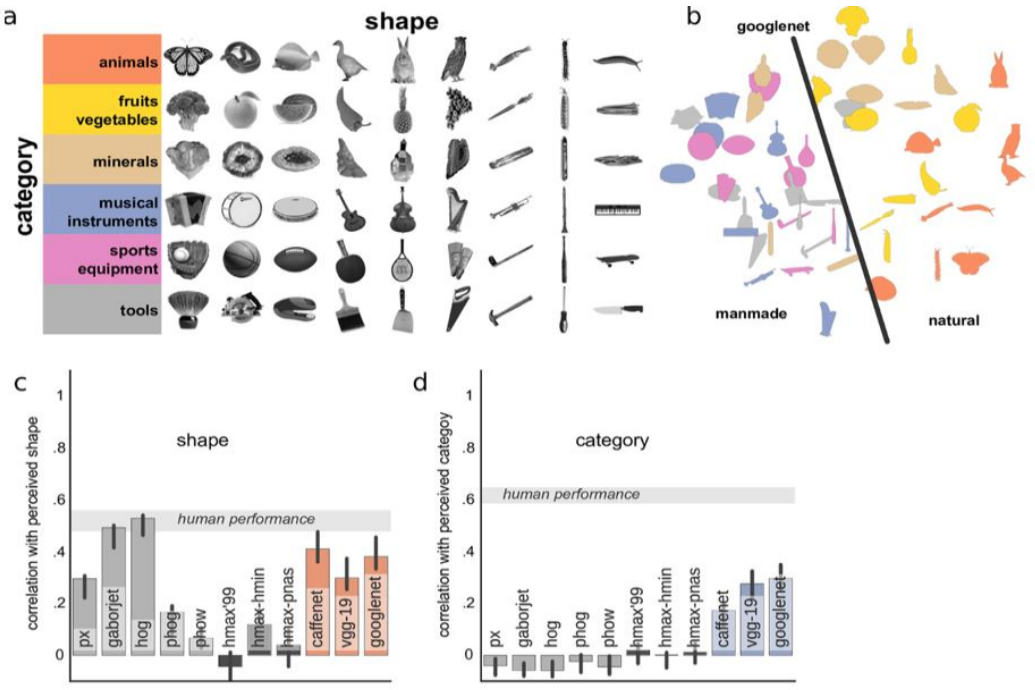

Figure 16: categorical representations in CNNs (Kubilius et al., 2016). The figure shows how convolutional neural networks (convnets) represent categorical information and shape in visual objects. a) stimulus set of objects with largely orthogonal shape and category information. b) shows convnets represent perceived shape and semantic category in parallel, with a clear separation between natural and man-made objects. (c-d) A correlation between model representations and human shape (c) and category (d) judgments.

Rajalingham et al., (2018) compared object recognition behavior of primates and artificial neural network models (DCNNs) using a series of behavioral metrics. Both biological and artificial visual systems were tested on the same set of images and characterized at multiple resolutions. The goal was to compare the behavioral signatures of the two systems and determine if they were statistically indistinguishable. At the object-level behavioral comparison, the authors found that deep convolutional neural network (DCNN) models had a higher similarity to human behavior than a baseline pixel model or a low-level V1 model (see Fig. 17). At the Image-level behavioral comparison, to determine the purely image-level behavioral variance, the authors computed the normalized image-level signature, which is obtained by subtracting the object-level discriminability from the one-versus-all image-level signature. The results showed a significant divergence between DCNNs and primates, with DCNN models being significantly less human-consistent than primates (see Fig. 18). The authors also tested modifications to the DCNN models to see if they could better match the primate visual system behavior. The authors tried changing the input to be more primate-like,

modifying the transformation from the internal model feature representation to the behavioral output, and fine-tuning the internal filter weights of the model. However, none of these modifications led to a significant improvement in the model's ability to accurately capture the image-level signatures of primates. Furthermore, Rajalingham et al., (2018) investigated whether certain image attributes, such as size, eccentricity, pose, and contrast, could explain the differences in performance between deep convolutional neural network (DCNN) models and primates in object recognition tasks and found that while the Inception-v3 (Szegedy et al., 2013) visual system model exhibited similar performance dependencies as primates, these image attributes could only explain a small portion of the variance in the DCNN models' residual signatures. Therefore, other factors likely contribute to the models' failure to capture primate image-level signatures.

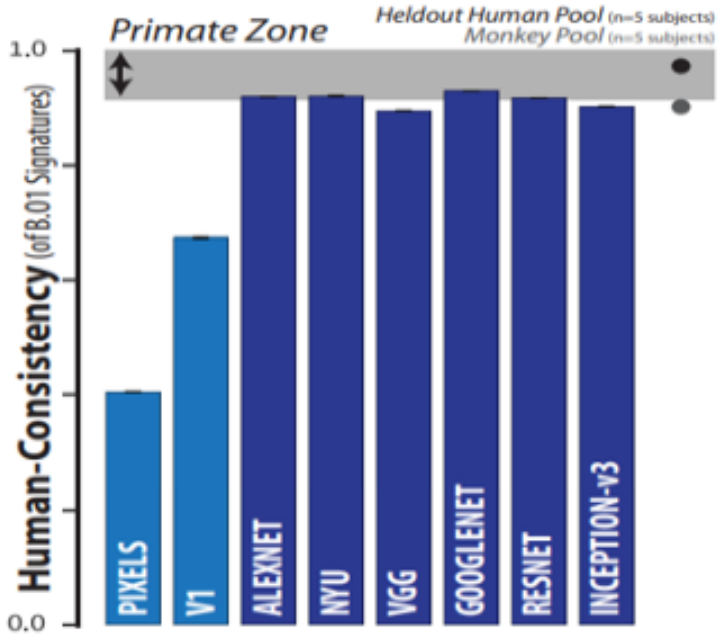

Figure 17: Object level comparison to human behavior (Rajalingham et al., 2018). The figure shows deep convolutional neural network (DCNN) models have higher similarity to human behavior than a baseline pixel model.

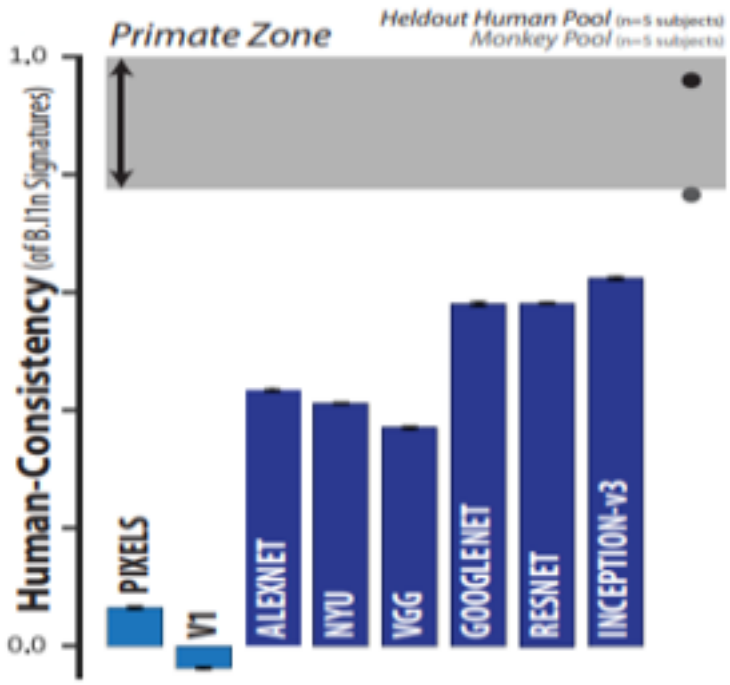

Figure 18: Image level comparison to human behavior (Rajalingham et al., 2018). The figure shows a significant divergence between DCNNs and primates, with DCNN models being significantly less human-consistent than primates.

CONCLUSION

In conclusion, the studies conducted by Yamins et al. (2014), Cadieu et al. (2014), Khaligh-Razavi and Kriegeskorte (2014), Kubilius et al. (2016), and Rajalingham et al. (2018) provide valuable insights into the performance and representational abilities of computational models and deep convolutional neural networks (CNNs) in object recognition tasks compared to the primate visual system.

Yamins et al. (2014) demonstrated that the performance of a neural network model in object categorization tasks strongly correlates with its ability to predict individual neural unit responses. They identified a high-performing neural network model (HMO) that matches human performance on object recognition, and found that its output layer accurately predicts neural responses in the IT cortex. The study suggests that top-down performance constraints shape intermediate visual representations.

Cadieu et al. (2014) compared the performance of different computational models, including deep CNNs, on challenging object recognition tasks. The deep CNNs outperformed other models, with the Zeiler & Fergus 2013 model performing the best. When comparing model representations to the IT neural representation, after correcting for noise and sampling

effects, the Zeiler & Fergus 2013 deep CNN performed similarly to the IT multi-unit representation. These findings highlight the competitive performance of deep CNNs and their ability to generalize with limited training examples.

Khaligh-Razavi and Kriegeskorte (2014) investigated the ability of computational models, including deep CNNs, to explain the representation of visual objects in the IT cortex. They found that better-performing models showed greater similarity to IT representations in terms of categorization performance and representational dissimilarities. Supervised learning models, such as deep CNNs, performed better in categorizing objects and explaining IT representational geometry compared to unsupervised models.

Kubilius et al. (2016) conducted experiments to evaluate the ability of CNNs to recognize objects based on shape and perceptual dimensions. Deep models demonstrated robust and accurate performance in object recognition tasks even when non-shape cues were removed. Deep CNNs captured perceived shape dimensions better than shallow models and showed sensitivity to differences between fonts. Additionally, deep models performed slightly better than shallow models in discriminating between stimuli with non-accidental properties. However, the learned representations in CNNs were largely based on shape rather than category.

Rajalingham et al. (2018) compared the object recognition behavior of primates and artificial neural network models, specifically deep CNNs. The results indicated that DCNN models had higher similarity to human behavior than baseline pixel models or low-level V1 models at the object-level behavioral comparison. However, at the image-level behavioral comparison, significant divergence was observed between DCNNs and primates, with DCNN models being less human-consistent. Modifications to the DCNN models did not lead to significant improvements in capturing the image-level signatures of primates. The study also found that certain image attributes could only explain a small portion of the differences in performance between DCNN models and primates.

Overall, these studies collectively emphasize the progress and challenges in developing computational models for understanding human object recognition. DNNs have shown remarkable performance and capability in certain aspects of object recognition, but there are still limitations in replicating the complexities of the primate visual system. Further research is needed to bridge the gap between computational models and the intricate neural mechanisms underlying human object recognition.

BIBLIOGRAPHY

Cadieu, C. F., Hong, H., Yamins, D. L., Pinto, N., Ardila, D., Solomon, E. A., ... & DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. PLoS computational biology, 10(12), e1003963.

Carandini, M., Demb, J. B., Mante, V., Tolhurst, D. J., Dan, Y., Olshausen, B. A., ... & Rust, N. C. (2005). Do we know what the early visual system does?. *Journal of Neuroscience*, *25*(46), 10577-10597.

Chikkerur, S., & Poggio, T. (2011). Approximations in the hmax model.

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. Trends in cognitive sciences, 11(8), 333-341..

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition?. Neuron, 73(3), 415-434..

Felleman, D.J., and Van Essen, D.C. (1991). Distributed hierarchical processing in the primate cerebral cortex. Cereb. Cortex 1, 1–47.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological cybernetics, 36(4), 193-202.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Holmes, E.J., and Gross, C.G. (1984). Effects of inferior temporal lesions on discrimination of stimuli differing in orientation. J. Neurosci. 4, 3063–3068.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of physiology, 160(1), 106.

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. The Journal of physiology, 195(1), 215-243.

Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. Science, 310(5749), 863-866.

Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS computational biology, 10(11), e1003915.

Kriegeskorte N, Mur M, Bandettini P (2008) Representational Similarity Analysis – Connecting the Branches of Systems Neuroscience. Frontiers in Systems Neuroscience 2.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., ... & Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron, 60(6), 1126-1141.

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

Kubilius, J., Bracci, S., & Op de Beeck, H. P. (2016). Deep neural networks as a computational model for human shape sensitivity. PLoS computational biology, 12(4), e1004896.

LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proceedings of the IEEE 86: 2278–2324.

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

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. Neural computation, 1(4), 541- 551.

Logothetis, N. K., & Sheinberg, D. L. (1996). Visual object recognition. Annual review of neuroscience, 19(1), 577-621.

Logothetis, N. K., Pauls, J., Bülthoff, H. H., & Poggio, T. (1994). View-dependent object recognition by monkeys. Current biology, 4(5), 401-414.

Montavon, G., Braun, M. L., & Müller, K. R. (2011). Kernel Analysis of Deep Networks. Journal of Machine Learning Research, 12(9).

Mutch, J., & Lowe, D. G. (2008). Object class recognition and localization using sparse features with limited receptive fields. International Journal of Computer Vision, 80, 45-57.

Pinto N, Barhomi Y, Cox DD, DiCarlo JJ (2011) Comparing state-of-the-art visual features on invariant object recognition tasks. IEEE Workshop on Applications of Computer Vision (WACV 2011): 463–470.

Pinto, N., Cox, D. D., & DiCarlo, J. J. (2008). Why is real-world visual object recognition hard?. PLoS computational biology, 4(1), e27.

Pinto, N., Doukhan, D., DiCarlo, J. J., & Cox, D. D. (2009). A high-throughput screening approach to discovering good forms of biologically inspired visual representation. PLoS computational biology, 5(11), e1000579.

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. Journal of Neuroscience, 38(33), 7255-7269.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature neuroscience*, *2*(11), 1019-1025.

Rust, N. C., & DiCarlo, J. J. (2010). Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area V4 to IT. Journal of Neuroscience, 30(39), 12978-12995.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. IEEE transactions on pattern analysis and machine intelligence, 29(3), 411-426.

Stoerig, P., & Cowey, A. (1997). Blindsight in man and monkey. *Brain: a journal of neurology*, *120*(3), 535-559.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.

Tanaka, K. (1996). Inferotemporal cortex and object vision. Annual review of neuroscience, 19(1), 109-139.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. Nature neuroscience, 19(3), 356-365.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proceedings of the national academy of sciences, 111(23), 8619-8624.

Zeiler, M. D. (2013). Fergus. R.: Visualizing and understanding convolutional networks. CoRR, abs/1311.2901.