Towards Biologically Plausible Deep Neural Networks

by

Yena Han

B.S., Massachusetts Institute of Technology (2014)
M.Eng., Massachusetts Institute of Technology (2017)

Submitted to the Department of Electrical Engineering and Computer Science in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy at the MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2024

© 2024 Yena Han. All rights reserved.
The author hereby grants to MIT a nonexclusive, worldwide, irrevocable, royalty-free license to exercise any and all rights under copyright, including to reproduce, preserve, distribute and publicly display copies of the thesis, or release the thesis under an open-access license.

Authored by: Yena Han
Department of Electrical Engineering and Computer Science
December 7, 2023

Certified by: Tomaso A. Poggio
Eugene McDermott Professor of Brain and Cognitive Sciences
Thesis Supervisor

Accepted by: Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students
Towards Biologically Plausible Deep Neural Networks

by

Yena Han

Submitted to the Department of Electrical Engineering and Computer Science on December 7, 2023, in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Abstract

Human intelligence has long been a source of inspirations for developing artificial intelligence. Computational principles and components discovered in the human brain have been successfully applied to artificial systems. These artificial systems are not only useful for engineering tasks but also serve as computational models of the brain, connecting theories to empirical data. Conversely, artificial intelligence, deep neural networks in particular, has contributed to advancing the understanding of the brain. Deep neural networks when trained adequately can reproduce behavioral and neural data better than previously developed models. Here we present studies that contribute to this interplay between natural and artificial intelligence. We first investigate invariance, a key computational principle that enables robust visual recognition and efficient generalization to new visual concepts, in human vision. Based on the experimental results, we propose deep neural network architectures that support the observed human behavioral properties in invariant recognition tasks. Next, we introduce a comparison framework for deep neural networks, where ground-truth targets are known, such that interpretations from the comparison can be validated. We explore whether deep neural networks with high functional similarity measures can provide reliable insights into the architectural building blocks of the brain.

Thesis Supervisor: Tomaso A. Poggio
Title: Eugene McDermott Professor of Brain and Cognitive Sciences
Acknowledgments

Firstly, I would like to extend my deepest gratitude to my advisor, Tomaso Poggio, for all the support and guidance since I began as an undergraduate research assistant in the lab, continuing throughout my graduate studies. As a true pioneer at the intersection of machine learning and computational neuroscience, your vision has consistently inspired me to view intelligence, both artificial and biological, through a unified lens. Additionally, I am fortunate to have been a part of the CBMM, an initiative you led, which significantly contributed to my growth as a researcher.

I am sincerely grateful to my thesis committee members, Leslie Kaelbling and Bill Freeman. Your insightful questions and feedback have been invaluable. I also thank you for the courses you taught, which I took as an undergraduate and which greatly influenced my dive into artificial intelligence.

Throughout the years, I have had the privilege of meeting incredible mentors and collaborators — Brian Cheung, Gemma Roig, Gadi Geiger, Leyla Isik, Shimon Ullman, Daniel Harari, Pulkit Agrawal, and Luke Arend. I also want to express my appreciation to all past and current lab members for creating a welcoming and engaging environment in our lab, where we could discuss anything, whether it be research-related or not. Special thanks to Qianli Liao, Mengjia Xu, Andy Banburski-Fahey, Arturo Deza, Fernanda De La Torre, Akshay Rangamani, Tomer Galanti, Chiyuan Zhang, Andrea Tacchetti, Steve Voinea, Kathleen Sullivan, and Kris Brewer.

My heartfelt thanks go to my friends — Chae Won, Hyo Jeong, Jisoo, Hyun Ryong, Hyodong, Sue Yeon, Hyungsuk, Seojin, Hyunjung, Eun Young, Minkyung, Joonseok, and Jiewon — for the fun memories we made during my PhD journey.

Lastly, I would like to thank my family and husband for the endless support and encouragement. Your belief in my potential, even during moments of self-doubt, has been a constant source of motivation. I would not be where I am today without your unwavering faith and cheer.
# Contents

1 Introduction 19

1.1 Computational Models ............................................. 20
1.2 Invariance for a learning system .................................. 21
1.3 Functional similarity and architectures in deep neural networks . . . 22
1.4 Main Contributions .................................................. 23

2 Invariance in One-shot Learning 25

2.1 Abstract .............................................................. 25
2.2 Introduction .......................................................... 26
2.3 Results ................................................................. 28
   2.3.1 Experiment 1: Scale-invariance ............................... 29
   2.3.2 Experiment 2: Translation-invariance ....................... 31
   2.3.3 Do Deep Neural Networks capture the properties of invariant recognition? ................................................. 34
2.4 Discussion ........................................................... 40
2.5 Methods ............................................................... 44
   2.5.1 Psychophysical Experiments ................................. 44
   2.5.2 Model Experiments ............................................. 47
2.6 Supplementary Information .......................................... 50
   2.6.1 SI Experiment 2: Translation-invariance experimental results . 50
   2.6.2 Asymmetry between Central and Peripheral Learning ...... 54
   2.6.3 Supplementary Figures ....................................... 55
3 System Identification of Neural Systems

3.1 Abstract .................................................. 61
3.2 Introduction ........................................... 62
3.3 Related Work ........................................... 65
3.4 Methods ................................................. 67
  3.4.1 Linear Encoding Analysis .......................... 68
  3.4.2 Centered Kernel Alignment ......................... 69
  3.4.3 Identifiability Index ................................. 70
  3.4.4 Simulated Environment ............................... 70
3.5 Results .................................................. 71
  3.5.1 Different models with equivalent neural predictivity 71
  3.5.2 Identification of architectures in an ideal setting .... 72
  3.5.3 Effects of the stimulus distribution on identifiability 76
  3.5.4 Challenges of identifying key architectural motifs ... 78
3.6 Discussion ............................................... 80
3.7 Supplementary Information ............................. 81
  3.7.1 Model details for Section 2.4.1: Brain-Score ....... 81
  3.7.2 Model details for Section 2.4.4: Finding the key architectural 81
    motif .........................................................
  3.7.3 Coverage of neural recordings in target systems .... 83

4 Conclusion ................................................. 85
  4.1 Future Directions ....................................... 86
List of Figures

2-1  (A) Sample stimuli. *Top row:* shows target letters, and *Bottom row:* shows distractor letters paired with the target above. (B) Experimental design. *Top:* illustrates a sample trial of scale-invariance experiments, and *Bottom:* illustrates a sample trial of translation-invariance experiments. The test letter was either the same as the target or its pairing distractor letter. (C) Experimental procedure. Each target and test letters was presented for 33 msec after a fixation dot was presented for 1 sec at the center of the screen.

2-2  Scale-invariance experimental results. Target and test letters were always shown at the center of the screen, only their size was varied, and subjects responded same or different. Error bars represent standard error (Number of subjects $n = 10$ for non-Koreans and $n = 3$ for Koreans conditions).
2-3  Windows of invariance for different conditions. Recognition accuracy from translation-invariance experiments is shown in a color scale. The central window (top) indicates results for learning target letters at the center of the visual field and being tested at another position in the peripheral visual field. Recognition accuracy is shown at corresponding scales and eccentricities of test letters. The peripheral window (bottom left) is for the reverse order where target letters are learned in the peripheral visual field and tested at the center. For this condition, as the position of target letters is varied and test letters are fixed at the center, we plot recognition accuracy at the learned scales and eccentricities of target letters. The opposite window (bottom right) shows results for learning target letters at a position in the peripheral visual field and being tested at the same distance from the center but in the opposite side of the visual field. In all plots, the tested conditions are marked with circles and other data points are estimated using natural neighbor interpolation (Number of subjects n = 9 for 30° letter, n = 11 for 1° letter, and n = 10 for 2° letter size conditions).

2-4  Windows of visibility. They are computed in a similar way as the windows of invariance (Fig. 2-3), but obtained from testing Korean subjects who are familiar with the visual stimuli (Number of subjects n = 3).
(A) Sampling points of the early visual cortex in the plane of eccentricity and scale, both in visual degrees, reproduced from (Poggio et al., 2014). Each ball represents a neuron, and there is the same number of neurons at all scales. The neurons at a larger scale cover a larger eccentricity than those at a lower scale. (B) Multi-scaled centered crops of an input image. The figure shows 4 crops among 10 that are used as the input to Eccentricity-dependent Neural Network. From the left to the right, the scale of input crops becomes larger, which are seen by larger receptive fields. (C) Eccentricity-dependent Neural Network. The input to the model is simulated visual field sampled at multiple resolutions as shown in (B), and the model is composed of convolutional layers followed by spatial and scale pooling. For simplicity, we visualize a model with one convolutional and pooling layer.

Simulation results on scale-invariance. Scale-invariant representation is assessed by comparing the features of two Korean letters, unseen by models in the training phase, and classifying the letters the same or different, independent from their size. Both ENN and CNN are trained on MNIST handwritten digit dataset (LeCun) with data augmentation of various scales and shifts. The trained models are used to extract features for Korean letters. Two Korean letters are considered to have the same identity if their associated features have Pearson correlation higher than a threshold. For each condition, we select a threshold that maximizes the classification accuracy.
2-7 Windows of invariance for ENN and CNN. We test Korean letters in a same-different task, where the position of the letters varies. Top row: shows classification accuracy for ENN and bottom row: shows that for CNN in an interpolated color scale. For both models, the central window indicates results for using features of target letters placed at the center of the visual field and using features of test letters at a position in the peripheral visual field. The peripheral window shows results for the reverse order of testing. Conditions we tested are marked with circles.

2-8 Translation-invariance experimental results for non-Koreans. Three letter sizes (30', 1°, 2°) were tested and the range of translation was varied depending on the letter size. For each letter size and displacement, 5 different testing conditions, where target and test letters were shown at different positions, are plotted. Target and test letter positions are indicated below each bar. Opp represents testing at the opposite of the visual field in the same eccentricity as the target letter. Error bars represent standard error (Number of subjects n = 9 for 30' letter, n = 11 for 1° letter, and n = 10 for 2° letter size conditions. *p < 0.05, **p < 0.01, ***p < 0.001 two-tailed paired t-test).

2-9 Translation-invariance experimental results for Koreans. We tested the positions with the largest eccentricity among the conditions we tested non-Korean subjects for each size of letter (30' letters at D = 3°, 1° at D = 2.5°, and 2° at D = 7°). In addition, to investigate the range of visibility window, recognition accuracy for 30' letters at D = 5° and 7° is plotted. Error bars represent standard error (Number of subjects n = 3 for each condition).
2-10 Scale-invariance experimental results, showing $d'$. We computed $d' = Z(\text{Hit}) - Z(\text{False alarm})$, where $Z$ is the inverse of the cumulative Gaussian distribution. Hit and false alarm rates were the average across all subjects’ data ($n = 10$). To bound the values of $d'$, we added 0.5 to both hit and false alarm rates and 1 to both the number of signal trials and the number of noise trials Hautus (1995). The overall results showing robustness to scaling is consistent with Figure 2-2.

2-11 Windows of invariance, showing $d'$. As in Figure 2-3, we show $d'$ for three different conditions (central, peripheral and opposite learning). Consistent with Figure 2-3, we find that the range of invariance increases with scale for the central learning condition. Also, peripheral and opposite learning conditions result in a narrower range of invariance than the central learning condition.

2-12 Windows of visibility, showing $d'$. As in Figure 2-4, we show that the range of invariance is extended with experience. Overall results are consistent with Figure 2-4, except unlike in Figure 4, we observe a narrower range for $2^\circ$ letters than that for smaller letters for some contours of $d'$ values (e.g. $d' = 2$). However, for smaller $d'$ values, this is not the case. Also, the results are consistent with our main conclusion that the window of visibility is wider than the window of invariance.

2-13 Simulation results on scale-invariance using multi-scale features. Testing conditions are identical to those for Figure 2-6. Instead of using max-pooled features, we use features from all scale channels for target letters. See the results section.
2-14 Simulation results for translation-invariance. \((0 \rightarrow D)\) indicates the central learning conditions, when target letters are placed at the center of the visual field and test letters at an eccentricity \(D^\circ\) in the peripheral visual field. \((D \rightarrow 0)\) indicates the reverse order of testing which is the peripheral learning conditions. To reproduce asymmetry in recognition rates between central and peripheral learning ENN templates associated with target letters are from all scales. Each scale is matched with the features from test letters, which are pooled over scale channels, to find a scale that maximizes the correlation. A linear fit is plotted for ENN.

2-15 Translation-invariance simulation results from using features pooled over all scales and positions. This method has limitation to explain the asymmetric recognition accuracy between central and peripheral learning. In the main text, we instead use features from all scales to reproduce translation-invariance experimental results. See the results section.

2-16 Pearson correlation between same Korean letters at different positions. See the results section. Error bars represent standard deviation (Number of letters \(n = 27\)).

3-1 **Illustration of the research question:** If one of the candidate models matches the target neural network (brain) closely, can the current model evaluation methods accurately find that out?

3-2 Linear regression scores of deep neural networks for brain activations in the macaque visual cortex. Architecture list in Section 3.7.1. For V1, the top performing three models are in the VOneNet family (Dapello et al., 2020), which are explicitly designed to mimic the known properties of V1.
3-3  Linear regression scores as a function of target network layer. Different initialization seeds are used for source networks, except for one (MLP-Mixer-B/16*), which uses identical weights. Black dots \( \bullet \) indicate that the correct architecture does not outperform others with statistically significant differences \((p > 0.01)\). Red dots \( \bullet \) indicate that the median scores for those networks are higher than the median for the correct architecture. The ranking of source models is typically based on median scores (Schrimpf et al., 2020).

3-4  CKA for different source and target networks. The experimental setup is identical to Figure 3-3 besides using CKA instead of linear regression as the metric. As in Figure 3-3, when we test MLP-Mixer-B/16 as a target and the source network type matches the target, weights are identical. We show the results for MLP-Mixer-B/16 in bar plots with a pattern to indicate the difference from other targets.

3-5  **Top** Sample images of each stimulus image type. (a) Identifiability index using CKA and (b) linear regression for various types of stimulus images and target networks.

3-6  Different architectural variants (12 CNNs and 14 ViTs) are compared with two CNNs and one ViT target network. Each data point represents the maximum score of an architecture for the corresponding target layers. Solid markers indicate the mean score of the corresponding model class, and error bars display standard deviations. \( \circ \) indicates that the corresponding layer does not show a statistically significant difference between model classes.

3-7  CKA scores when all target units are tested. Experimental setups are identical to Figures 3-4, 3-5, and 3-6 otherwise. See Section 3.5.2 for discussion. For the target MLP-Mixer-B/16 in (a), the source MLP-Mixer-B/16 has identical weights with the target; thus CKA is trivially 1.
List of Tables

3.1 Model details: Number of layers included for each model . . . . . . . 82
Chapter 1

Introduction

The study of intelligence is a broad field encompassing both scientific and engineering disciplines. The scientific aspect of intelligence aims to understand the principles and mechanisms underlying intelligent behavior. On the other hand, the engineering aspect of intelligence focuses on replicating the intelligent behavior in artificial systems, often referred to as artificial intelligence. These two aspects have each made contributions to the progress of the other.

Human intelligence stands out as a unique type of intelligence. Therefore, there have been extensive efforts to take inspiration from the human brain and apply insights about how it works to the development of artificial intelligence. In the past, findings in neuroscience and cognitive science have led to some of the key contributions to engineering intelligent systems. For example, artificial neural networks, as the name implies, are loosely based on biological neurons in the brain. The hierarchical structure of deep learning models has a long history of adapting the motif from previous studies in neuroscience (Fukushima, 1980; Riesenhuber and Poggio, 1999; Serre et al., 2007). Moreover, reinforcement learning was inspired by the neural mechanisms of predictions and rewards during learning (Schultz et al., 1997).

Conversely, artificial intelligence has made strides in neuroscience as well. In particular, deep learning, which is a subfield of artificial intelligence, has become an important tool to study natural intelligence. Deep learning models are applied to decoding analyses of neural recordings (Schirrmeister et al., 2017; Glaser et al., 2020)
or to mapping neural connectomes (Lee et al., 2019).

Importantly, previous studies also support more explicit connections between artificial and natural intelligence. Task-optimized deep neural networks are shown to capture neural activities better than models previously developed in the neuroscience field (Yamins et al., 2014). The correspondence between deep neural networks and the brain has been found in modalities other than vision, such as audio or language (Kell et al., 2018; Schrimpf et al., 2021). A positive correlation between performance on engineering tasks and neural predictivity suggests that certain deep neural networks, despite not having biological relevance as their primary goal, can serve as good models of the brain.

In this thesis, we aim to contribute to the interplay between natural and artificial intelligence, with a particular focus on visual intelligence. Visual intelligence is a crucial part of human intelligence, as humans rely heavily on visual inputs to understand and interact with the external world. The ability to make visual judgments effortlessly, even in complex scenes, is a hallmark of human intelligence.

Artificial systems, whether developed based on neuroscientific findings or shown to have high neural predictivity through ad-hoc analyses, have demonstrated utility in engineering tasks. However, they are doubly important in systems and computational neuroscience fields, as they serve as computational models of the brain. The importance of these systems as brain models is further motivated in the following section.

1.1 Computational Models

Models for brains are descriptions of human cognitive functions, neural activations, or biochemical interactions at the synapse level. The various types of description can be more formalized in the framework of Marr-Poggio’s levels of analysis (Marr and Poggio, 1976). The Marr-Poggio’s hypothesis considers the brain as a system for information processing, and the system can be understood at three levels,

- Computational: describes what the system does and why.
• Algorithmic: describes how the system implements the computational goal. It involves step-by-step operations to have the desired representation.

• Physical: describes the physical realization of the algorithm. The focus is on the neural and biological activations in the brain.

The original idea of the tri-level analysis was to understand the system in three distinct levels, complementary to each other. However, modern models often span more than one level. In particular, deep neural networks (DNNs) are being proposed as models of the brain. While differing from the brain at the physical level, DNNs share computational and algorithmic similarities with the brain. For instance, they partially exhibit shared computational motifs, such as hierarchical processing and cascaded local filters and pooling. Furthermore, specifics of DNN models can be validated against empirical brain or behavioral data. Artificial and biological systems can process analogous inputs, and representational or behavioral comparisons allow evaluation of the proposed DNN models. This enables the validation of theories these models are built upon.

A model’s biological plausibility depends on various factors (Richards et al., 2019). For instance, previous studies have proposed more biologically plausible learning algorithms for DNNs (Xiao et al., 2018; Miconi, 2017). The training data can also be designed to be more naturalistic (Orhan et al., 2020). Among the different aspects, we focus on DNN architectures in this thesis, as they can capture computational and algorithmic understanding of the target biological system through deliberately designed components to achieve computational goals.

The studies presented in this thesis demonstrate how we can develop DNN architectures informed by findings from neuroscience. We also explore the reverse direction—specifically, whether current DNNs can reciprocally provide new insights into the brain. Each of these approaches is further motivated in Sections 1.2 and 1.3.
1.2 Invariance for a learning system

One way to improve computational models, as discussed earlier, is to incorporate principles deemed crucial for natural intelligence into artificial systems. In the first part of this thesis (Chapter 2), we consider invariance as a key principle underlying effortless visual recognition in complex, rapidly changing environments.

Invariance is a critical property that enables neural systems to achieve robust visual recognition despite various image transformations (Tacchetti et al., 2018; Anselmi et al., 2016). At a high level, having invariant representations for different transformations while maintaining specificity for distinct object categories leads to correct classification. The human visual system builds increasingly more invariant representation from lower visual areas to higher visual areas (Isik et al., 2013). Furthermore, invariant representations enable sample-efficient learning for a system by reducing the need for training on multiple views of the same instance (Poggio and Anselmi, 2016).

To develop models that mimic this generalization capability arising from invariant representation, in Chapter 2 we first characterize geometric invariance in human visual recognition empirically. Based on the observed invariance properties, we then design deep neural networks to replicate them.

1.3 Functional similarity and architectures in deep neural networks

Next, we explore how current DNNs developed for engineering purposes can reciprocally advance scientific understanding of natural intelligence. In Chapter 3, we begin by observing that various deep learning models, though not designed to mimic biology, turn out to capture the variance of brain activities well. Their architectures often lack clear biological counterparts in established neuroscience studies. Consequently, these neural networks present an opportunity to uncover new insights into the architectural components underlying biological vision.

To this end, we propose a framework to understand whether functional similar-
ity, which is a common metric for comparing DNNs to the brain, leads to accurate interpretations about model architectures. Using this framework, where we compare DNNs against another DNN of our choice, we investigate the capabilities and limitations of relying on functional similarity for system identification and explore ways to improve it.

1.4 Main Contributions

The main contributions of this thesis can be summarized as:

- The human visual system exhibits strong scale invariance in one-shot learning, maintaining recognition accuracy when the size of an object changes from 30 minutes of arc to 5 degrees of arc.

- Translation-invariance in one-shot learning is limited. The degree of translation-invariance increases with object size.

- Visual recognition is asymmetric between central and peripheral learning conditions.

- The range of translation-invariance expands with more visual experience.

- Eccentricity-dependent Neural Networks (ENN) with its built-in scale-invariance captures the invariance properties more closely than standard Convolutional Neural Networks.

- Various deep neural networks exhibit comparable functional similarity measures with the brain.

- Identifying the correct model architecture based on functional similarity scores is challenging, even under ideal conditions.

- The accuracy of system identification improves as a function of network depth.

- Stimulus images significantly impact system identification capability, with more naturalistic images enabling better identification.
• Increasing the number of neural recording units in a target system improves identifiability.
Chapter 2

Invariance in One-shot Learning

The contents of this chapter are adapted from Yena Han, Gemma Roig, Gad Geiger, and Tomaso Poggio, "Scale and translation-invariance for novel objects in human vision", Scientific Reports (2020).

2.1 Abstract

Though the range of invariance in recognition of novel objects is a basic aspect of human vision, its characterization has remained surprisingly elusive. Here we report tolerance to scale and position changes in one-shot learning by measuring recognition accuracy of Korean letters presented in a flash to non-Korean subjects who had no previous experience with Korean letters. We found that humans have significant scale-invariance after only a single exposure to a novel object. The range of translation-invariance is limited, depending on the size and position of presented objects. To understand the underlying brain computation associated with the invariance properties, we compared experimental data with computational modeling results. Our results suggest that to explain invariant recognition of objects by humans, neural network models should explicitly incorporate built-in scale-invariance, by encoding different scale channels as well as eccentricity-dependent representations captured by neurons’ receptive field sizes and sampling density that change with eccentricity. Our psychophysical experiments and related simulations strongly suggest
that the human visual system uses a computational strategy that differs in some key aspects from current deep learning architectures, being more data efficient and relying more critically on eye-movements.

2.2 Introduction

Invariance to geometric transformations can be a huge advantage for a visual recognition system. It is important to distinguish between invariance due to the underlying representation, which we refer to as intrinsic invariance, and example-based invariance for familiar objects that have been previously seen under several different viewpoints. The latter is computationally trivial and is available to any recognition system with sufficient memory and large training data. The first one, which may be hardwired or learned during a developmental period, provides a learning system the ability to learn to recognize objects with a much smaller sample complexity, that is with much smaller training sets (Anselmi et al., 2016; Poggio and Anselmi, 2016). This is not only a big advantage for any recognition system but it is also a key difference between today’s best deep learning networks and biological vision systems: the most obvious advantage of children versus deep networks is the ability to learn from a (labeled) training set that is several orders of magnitude smaller (Lake et al., 2017). The prototypical observation is that we can easily recognize a new object, such as a new face — seen only once— at a different scale.

Current deep neural networks exploit architectural priors for intrinsic invariance. For instance, Convolutional Neural Networks, which are widely used in computer vision, have an architecture hard-wired for some translation-invariance while they rely heavily on learning through extensive data or data augmentation for invariance to other transformations (Krizhevsky et al., 2012). Networks that incorporate a larger set of intrinsic invariances, such as rotation-invariance, have been proposed (Cohen and Welling, 2016; Dieleman et al., 2016; Marcos et al., 2017). Nevertheless, it is not clear which type of intrinsic invariance should be encoded in more biologically plausible models. As a consequence, it is important to characterize the degree
of invariance in human vision, starting from the simplest invariances—scale- and translation-invariance—and evaluate models that reproduce them.

Surprisingly, the available psychophysical results are often incomplete and inconclusive. Most experiments have targeted only translation-invariance, and a review (Kravitz et al., 2008) states that based on experimental data, the role of object position is not well understood and there is little evidence supporting the idea that human object recognition is invariant to position. Findings from previous studies range from “This result suggests that the visual system does not apply a global transposition transformation to the retinal image to compensate for translations.” (Nazir and O’Regan, 1990) to “For animal-like shapes, we found complete translation invariance,” (Dill and Edelman, 2001) and finally to “Our results demonstrate that position invariance, a widely acknowledged property of the human visual system, is limited to specific experimental conditions” (Dill and Fahle, 1998). Furthermore little research was conducted on scale-invariance with regard to unfamiliar stimuli (see Biederman and Cooper (1992); Furmanski and Engel (2000) for studies on scale-invariant recognition of familiar objects. Although a new set of objects different from those in the training phase was tested, the images are still of common objects (Furmanski and Engel, 2000)).

Physiological data on monkeys, on the other hand, give more consistent results on intrinsic invariance in the visual system. A few authors (Logothetis et al., 1995; Hung et al., 2005) reported that IT responses were invariant to scale and translation, once the monkeys learned a novel object under a single viewpoint. In humans, however, the extent of intrinsic invariant recognition is still unknown (see Ito et al. (1995); Rolls (2001); Rust and DiCarlo (2010) for studies on primate invariant recognition and Isik et al. (2013) for human invariant recognition of familiar objects).

In the experiments on translation-invariance, it is important to take into account that primate visual acuity depends strongly on eccentricity. Historically the eccentricity-dependence of visual acuity has been studied extensively (see Strasburger et al. (2011) for a review). In particular, previous studies using letter stimuli (Anstis, 1974; Strasburger et al., 1991) found that visual acuity decreases linearly with ec-
centricity. Therefore, if we consider the range of visual angle in which objects are recognizable for each size, we can define a window of visibility which lower bound is a linear relation between objects' size and position. The linear relation between recognizable scale and position of an object is also consistent with the physiological data that shows that the size of receptive fields in the primate visual cortex increases with eccentricity (Freeman and Simoncelli, 2011). The results imply that fine details, as required for instance to recognize letters at a distance, are visible only to the small receptive fields in the foveola, whereas coarser details, such as those associated with larger letters, are also visible to the larger receptive fields present at greater eccentricities.

The main questions of this paper can thus be phrased as follows. Does a window of invariance exist within the window of visibility? What is its geometry and size? In particular, for visibility there is a linear relation between scale and position. Is the same linear relation also valid for the window of invariance? We investigate these issues by examining human invariant recognition in the one-shot learning scheme, using previously unfamiliar visual stimuli. We also ask whether hierarchical Convolutional Neural Networks can account for the experimental data. In particular, we consider Eccentricity-dependent Neural Networks (ENN). ENNs – described more thoroughly later – implement the hypothesis that the human visual system has hardwired scale-invariance with the size of the receptive fields of the model neurons increasing with eccentricity (Poggio and Anselmi, 2016). These experiments, together with simulations, allow us to characterize invariant recognition arising from intrinsic brain representations.

2.3 Results

To study intrinsic invariance we analyzed results for recognition of unfamiliar letters in one-shot learning. For the one-shot learning task, we flashed a target Korean letter and then a test Korean letter, which was either the same as the target or a different distractor, to non-Korean subjects who were unfamiliar with Korean letters.
Figure 2-1: (A) Sample stimuli. Top row: shows target letters, and Bottom row: shows distractor letters paired with the target above. (B) Experimental design. Top: illustrates a sample trial of scale-invariance experiments, and Bottom: illustrates a sample trial of translation-invariance experiments. The test letter was either the same as the target or its pairing distractor letter. (C) Experimental procedure. Each target and test letters was presented for 33 msec after a fixation dot was presented for 1 sec at the center of the screen.

To investigate invariant recognition to transformations, we varied scale and position of the letters. When testing recognition in the peripheral visual field, we randomized to which side of the visual field letters were presented to prevent that subjects predict the letters’ position, fixate on the stimuli, and observe them with their foveal vision. We limited the presentation time to 33 ms to avoid eye movements. In Figure 3-1 we depict the experimental set-up and a set of Korean letters used.

2.3.1 Experiment 1: Scale-invariance

We tested scale-invariant recognition by flashing both target and test Korean letters at the fixation point in the center of the screen. First, we used 30' and 2° letter size. In Figure 2-2 we compare the three conditions when the size of target and test letters were (30', 30'), (30', 2°), and (2°, 30'), respectively, in which the first number
Figure 2-2: Scale-invariance experimental results. Target and test letters were always shown at the center of the screen, only their size was varied, and subjects responded same or different. Error bars represent standard error (Number of subjects n = 10 for non-Koreans and n = 3 for Koreans conditions).

of the pair refers to the size of the target letter and the second indicates the size of the test letter. Mean accuracy under all three conditions was higher than 0.85, which is significantly above chance (0.50). Changing the letter size did not have any statistically significant effect on performance ($F(2,18) = 0.94, p = 0.41$).

We performed a second set of experiments with a greater range of change in scale, in which we tested invariance of recognition with respect to 10-fold increase and decrease of letter size with 30° and 5° letters. Results were similar to those from the first setting. Mean accuracy was above 0.83, which is significantly higher than chance, and the difference in accuracy among the three presentation conditions was statistically non-significant ($F(2,18) = 0.80, p = 0.46$).

After observing that visual recognition is robust to scale change in one-shot learning, to compare the range of invariant recognition with that of recognition of familiar objects, we tested native Koreans under the same conditions. The results
confirmed that the task was not challenging to Koreans. Mean accuracy for all conditions was above 0.92 (Fig. 2-2 bottom). When these results were compared with non-Koreans’ data, we did not find any significant interaction between presentation conditions and whether the subjects were Koreans or not (combinations of 30’ and 2° letters: $F(2, 22) = 0.03, p = 0.97$; combinations of 30’ and 5° letters: $F(2, 22) = 0.23, p = 0.80$). We report results using another behavior performance metric $d'$ in Figure S3, which were consistent with the accuracy results.

2.3.2 Experiment 2: Translation-invariance

Next, we investigated translation-invariance by shifting the position of test letters from target letters. We divided the conditions into two categories: learning at the central visual field and learning at the peripheral visual field, based on the position where the target object is learned. We show recognition accuracy at different positions for each scale, which displays the relationship between scale, position, and degree of invariance in Figure 2-3 (performance $d'$ is reported in Fig. S4). More details on the experimental set-up are provided in the SI methods section. Recognition accuracy is shown as bar plots in Figure S1. We also performed similar analyses as for scale-invariance by comparing invariant recognition accuracy with baseline conditions (same position). Unlike scale-invariance experiments, this yielded statistically significant differences in some cases, which suggests limited translation-invariance. We report these results in Figure S1 and here we further analyze the properties of translation-invariance.

Since in a natural setting, humans are able to observe the unknown objects with their fovea, we first focus on analyzing the central learning condition (Fig. 2-3 top). For all scales, recognition accuracy was the highest at the center, when there was no displacement, and decreased with increasing distance from the center of the visual field. In addition, the range of translation-invariance increased with the scale of stimuli. While recognition accuracy was maintained high at a position as far as 7° in the periphery for 2° letters, it dropped significantly even at 1° for 30' letters. Considering the area where recognition accuracy is above a threshold (e.g. 0.85) as
Figure 2-3: Windows of invariance for different conditions. Recognition accuracy from translation-invariance experiments is shown in a color scale. The central window (top) indicates results for learning target letters at the center of the visual field and being tested at another position in the peripheral visual field. Recognition accuracy is shown at corresponding scales and eccentricities of test letters. The peripheral window (bottom left) is for the reverse order where target letters are learned in the peripheral visual field and tested at the center. For this condition, as the position of target letters is varied and test letters are fixed at the center, we plot recognition accuracy at the learned scales and eccentricities of target letters. The opposite window (bottom right) shows results for learning target letters at a position in the peripheral visual field and being tested at the same distance from the center but in the opposite side of the visual field. In all plots, the tested conditions are marked with circles and other data points are estimated using natural neighbor interpolation (Number of subjects $n = 9$ for $30'$ letter, $n = 11$ for $1^\circ$ letter, and $n = 10$ for $2^\circ$ letter size conditions).

the range of invariance, we observed a roughly V-shaped area. We found the same
tendency that recognition accuracy depends on eccentricity and scale in peripheral learning conditions.

Additionally, overall recognition accuracy was significantly lower under peripheral learning than under central learning, particularly when there was a change in resolution of test letters from that of target letters (Fig. 2-3 Peripheral window) i.e. translation-invariance was more limited under peripheral learning. In a related setting with peripheral learning, when target letters are learned in the peripheral visual field and test letters are presented at the same distance from the center but in the opposite side of the visual field, the range of invariance was less limited. Note that under this condition, the resolution of letters did not change and only their position was changed to the opposite side of the visual field. The corresponding window of invariance (Fig. 2-3 Opposite window) was still more limited than the results from central learning conditions.

**Does the range of invariance extend with experience?**

To compare the properties of intrinsic translation-invariance with those observed in subjects with experience, we tested native Korean subjects with the same experimental set-up as for the above experiments, displayed in Figure 3-1 (performance \(d'\) is reported in Fig. S5). For Korean subjects, we measured their recognition accuracy using the furthest position tested for each size among the conditions used for non-Korean subjects (30' letters at eccentricity D = 3°, 1° at D = 2.5°, and 2° at D = 7°). The mean accuracy performance for all three letter sizes was higher than 0.85 (Fig. S2), which confirms that the conditions for which we tested translation-invariance become trivial when the subjects have previous experience.

The above set of experiments suggest that the range of recognition is wider when the stimuli are familiar than in one-shot learning, and that recognition performance improved with experience and exposure to the stimuli at different positions. To further investigate the properties and tendency of visibility window, we tested eccentricities D = 5°, 7° for 30' letters. Compared to non-Koreans' results, we can confirm that overall recognition accuracy of Korean subjects is higher in Figure 2-4. In addition,
as in the case of testing non-Korean subjects, the range of visibility window was wider for central learning than for peripheral and opposite learning conditions.

2.3.3 Do Deep Neural Networks capture the properties of invariant recognition?

To understand the underlying brain computation that enables human invariant recognition characterized in psychophysical experiments, we compared the experimental data with computational modeling results. In particular, we investigated whether in-
Figure 2-5: (A) Sampling points of the early visual cortex in the plane of eccentricity and scale, both in visual degrees, reproduced from (Poggio et al., 2014). Each ball represents a neuron, and there is the same number of neurons at all scales. The neurons at a larger scale cover a larger eccentricity than those at a lower scale. (B) Multi-scaled centered crops of an input image. The figure shows 4 crops among 10 that are used as the input to Eccentricity-dependent Neural Network. From the left to the right, the scale of input crops becomes larger, which are seen by larger receptive fields. (C) Eccentricity-dependent Neural Network. The input to the model is simulated visual field sampled at multiple resolutions as shown in (B), and the model is composed of convolutional layers followed by spatial and scale pooling. For simplicity, we visualize a model with one convolutional and pooling layer.
variance properties observed in human one-shot learning can be learned by examples seen by the model or alternatively, requires an intrinsic architecture for them. We used Convolutional Neural Networks (CNN) to simulate the experimental results, as these models showed a significant success in explaining visual processing in the primate ventral stream (Yamins and DiCarlo, 2016; Hong et al., 2016; Khaligh-Razavi and Kriegeskorte, 2014; Cichy et al., 2016) and matching behavioral patterns of object recognition with humans (Kheradpisheh et al., 2016; Rajalingham et al., 2015). A trivial way to achieve invariant recognition, widely adopted in computer vision field, is to use data augmentation to train CNNs (Krizhevsky et al., 2012). Although models can reach human-level invariant recognition performance for familiar objects with this method, the strategies of CNNs in using diagnostic features were shown to be different from humans (Karimi-Rouzbahani et al., 2017). Moreover, it is unknown whether invariant recognition can be transferred to a new category of stimuli, unseen in the training phase. To show the limitation of this example-based invariance in one-shot learning, we compared CNNs with Eccentricity-dependent Neural Networks (ENN) (Poggio et al., 2014; Chen et al., 2017; Volokitin et al., 2017). ENNs, depicted in Figure 2-5, are modified from CNNs to have scale-invariance built into their architecture and have dependence of receptive field size on eccentricity, consistently with physiology data (Freeman and Simoncelli, 2011).

Both CNNs and ENNs were trained on MNIST handwritten digit dataset (LeCun) with data augmentation of various scales and positions. With this training, the networks should develop top-layer features capable of processing character-like stimuli. Those features are then used to evaluate the similarity of two Korean letters, as in the psychophysics experiments. Two Korean letters are considered to have the same identity if their associated features have Pearson correlation higher than a threshold. Here, we report results from applying a different threshold that maximizes accuracy for each condition. We also included distractor letters in testing so that we evaluate selectivity of the models.
Figure 2-6: Simulation results on scale-invariance. Scale-invariant representation is assessed by comparing the features of two Korean letters, unseen by models in the training phase, and classifying the letters the same or different, independent from their size. Both ENN and CNN are trained on MNIST handwritten digit dataset (LeCun) with data augmentation of various scales and shifts. The trained models are used to extract features for Korean letters. Two Korean letters are considered to have the same identity if their associated features have Pearson correlation higher than a threshold. For each condition, we select a threshold that maximizes the classification accuracy.

Simulation 1: Scale-invariance

As described earlier, the psychophysical experiments show that the human visual system is immediately invariant to scale change in one-shot learning. We first tested whether the results with ENNs, which are of course designed to be scale-invariant, fit the data. We evaluated the degree of scale-invariance for Korean letters, which the models did not see during training. As expected, accuracy when the target and test letters are of different size turned out to be significantly higher than chance (Fig. 2-6 left). Although classification accuracy for testing invariant conditions was lower than that for the baseline condition, when the letter size does not change, this was partly due to the difference between biological systems and computational models. In simulations, since there was no noise, the input images for target and test letters were exactly the same under the baseline condition, which resulted in 1.0 classification accuracy. Overall high accuracy in the testing shows that scale-invariance properties of ENN are consistent with the human data.
We then asked whether a model which has scale-invariance (example-based) for familiar objects shows intrinsic scale-invariance for a new set of objects. To test this hypothesis, we evaluated CNNs for Korean letters. Note that these models were scale-invariant for the trained MNIST dataset. The results obtained with the CNN model (Fig. 2-6 right) show that classification accuracy when the letter size changes was higher than chance but significantly lower than accuracy for ENN and psychophysical data. This limitation suggests that CNNs with data augmentation cannot account for scale-invariance in one-shot learning.

**Simulation 2: Translation-invariance**

In our psychophysical experiments, the degree of translation-invariance increases with letter size, both under central and peripheral learning. In our simulations (Fig. 2-7 bottom), CNNs were not able to replicate the property of limited translation-invariance. Accuracy for larger stimuli was higher than that for smaller stimuli, but it did not decrease with eccentricity. These results were expected due to translation-invariant model prior of the CNNs. For ENNs (Fig. 2-7 top), on the other hand, accuracy decreases with eccentricity while the range of invariant recognition increases with the size of letters, consistently with the psychophysical results. As in psychophysical results, if we choose a threshold classification accuracy and draw an accuracy contour, we can observe a V-shaped area of invariance. (We report raw data in Fig. S7. Window of invariance for ENN (Fig. 2-7 top) is based on the linear regression of the raw data.)

Furthermore, we investigated whether the models can reproduce the asymmetry in recognition rates between central and peripheral learning. The first idea we explored is that the one-shot learning stage stores templates obtained from processing the visual field at multiple scales. Thus, when target letters are presented at the center of the fovea, the associated templates contain all the full range of spatial frequencies. However, when target letters are shown at an eccentricity, since only the central visual field is sampled at high resolution, the templates are effectively low-pass versions of the foveal ones. Therefore, an explanation of the asymmetry between central and
Figure 2-7: Windows of invariance for ENN and CNN. We test Korean letters in a same-different task, where the position of the letters varies. Top row: shows classification accuracy for ENN and bottom row: shows that for CNN in an interpolated color scale. For both models, the central window indicates results for using features of target letters placed at the center of the visual field and using features of test letters at a position in the peripheral visual field. The peripheral window shows results for the reverse order of testing. Conditions we tested are marked with circles.
peripheral learning may start with the different range of resolutions available for templates memorized in the two situations of foveal vs. “peripheral” learning.

Since CNNs process images at one resolution only, they are unable to account for positional asymmetry in learning, as shown in Figure 2-7. In our simulations, features extracted by a CNN by removing the fully connected layer were used to compute Pearson correlation between target and test presentations to evaluate the identity of letters. Thus, there cannot be any asymmetry as a function of the order of the presentations. Within the class of ENN models, we use a network that uses features extracted from all scale channels to “learn” the representation for target letters. In Figure 5C, these features are the output of convolution and spatial pooling. For the test it uses pooling over the scale channel features. The comparison between stored template and new image is then carried out as follows. The scale channel of the stored letter that has the highest correlation with the test letter is chosen, and then a threshold is applied. The idea is that once we view a target letter with a specific cutoff in spatial frequencies, templates at multiple resolutions lower than that resolution in memory become available: each one can then be compared with the features of the test letter. Clearly the computation of the similarity between target and test is now asymmetric, unlike the original model of using features pooled over different scales and positions for the learning stage (Poggio et al., 2014) (Results from using the original model are provided in Figure S8. Scale-invariance simulations using the modified metric are shown in Figure S6; results are consistent with the conclusion from simulation 1). The results of the simulation showed that accuracy for peripheral learning conditions was lower than that for central learning conditions, similarly to the psychophysical data.

One may consider it may be the case that our method finds the optimal threshold that differentiates letters of the same or different identity, even when the underlying representation between translated objects are actually not being more dissimilar with eccentricity. We additionally confirmed that this is not the case by assessing the raw data, which is Pearson correlation between the same Korean letters at different positions (Fig. S9). The results verified that correlation between the representation
becomes lower with eccentricity.

2.4 Discussion

While it is widely agreed that humans are able to process complicated visual information invariant to transformation, so far it remained rather unclear whether this is possible because of previous exposure to the specific visual stimuli at different viewpoints or whether the visual system computes invariant representations for novel objects. To address the issue, we characterized the degree of invariance to transformation in one-shot learning, using stimuli for which the subject had no previous experience. We found that there is significant scale-invariance in recognition. We also found limited translation-invariance that increases with decreasing spatial frequency content of the stimuli, as expected (see for instance Anselmi et al. (2016)). Overall, as a function of eccentricity, the window of invariance is narrower than the window of visibility (i.e. acuity). Further, we observed an asymmetry between learning in the fovea and testing in the periphery with respect to the opposite sequence of training and testing.

Our experimental settings controlling familiarity to the objects as well as position and size of them clarify and extend previous studies on invariant recognition. Previous studies (Nazir and O’Regan, 1990; Dill and Edelman, 2001; Dill and Fahle, 1998) examined invariant recognition to translation when visual stimuli were first learned with the peripheral vision. However, unlike those experimental conditions, humans can freely observe unknown objects, and they mostly use foveal vision for learning target objects, since it is almost exclusively in a laboratory setting that peripheral training may happen. Therefore, our results on the asymmetry between central and peripheral one-shot training suggests a difference between natural and unnatural conditions (we refer as natural condition when the object is centered at the fixation point, and unnatural otherwise). While conclusions from previous studies on very limited position-invariance are drawn from peripheral training condition only, we observed stronger invariant representations in a more natural setting. Also, by testing two
subject groups who differed in familiarity to the visual stimuli, we confirmed that invariance depends on familiarity with the visual stimuli, consistent with Dill and Edelman (2001).

We compared the experimental results with computational models based on neural networks. One of our key contributions is that we conclude that standard CNNs cannot account for these experimental data on invariance, whereas a related class of neural networks, that we call ENNs, can. This suggests that ENNs might be better suited for computationally modeling the visual cortex than CNNs, which have been widely used for modeling the ventral stream (Yamins and DiCarlo, 2016; Hong et al., 2016; Khaligh-Razavi and Kriegeskorte, 2014; Cichy et al., 2016). Furthermore, our results suggest a rather different computational strategy from the one used in these models. In particular, the limited invariance to eccentrically located targets implies that several quite small “effective images” at different resolutions are available to later visual processing rather than a single large image at a fixed resolution (Poggio et al., 2014; Ullman et al., 2016). If objects are recognized at multiple resolutions in these effective images — i.e. they are not bound to any specific resolution— the models become scale-invariant.

The limitation of CNNs in contrast with ENNs in explaining scale-invariant recognition highlights the significance of an architectural prior (innate or developed during an early stage of visual experience, see Anselmi et al. (2016); Poggio and Anselmi (2016)). CNNs are designed under the assumption that objects should have the same features regardless of their position (assuming antialiasing is taken care of properly as convolutional architectures designed without considering the classical sampling theorem can also suffer from aliasing effect (Azulay and Weiss, 2018)). For other transformations, it is in principle possible that the models learn an invariant representation through rich training data. They would then be able to extract features invariant to transformations. A theory describing architectures capable of this feature was in fact developed in Anselmi et al. (2016). We found that, however, invariant recognition in CNNs is highly constrained to the exact type of dataset that are used for training, and there is very limited transfer invariance to other datasets, even when
they are similar. This suggests that CNNs mainly develop example-based invariance, limited to a memorized set of data. Our psychophysical results, on the other hand, indicate that human invariant recognition supports an alternative design choice which is consistent with neural networks that enforce scale-invariant representation, as in ENNs.

Although our results support built-in scale-invariance for computational models, the exact implementation details of the ENNs architectures tested here need to be further verified. In particular, pooling all scales at the last layer gives a high degree of scale-invariance, but this may well be different from the operations performed in the visual system. The dynamics of invariant recognition of familiar objects in the human visual system were studied in Isik et al. (2013), and the study suggested that the human visual system develops invariant representation in stages corresponding to different visual areas in the ventral stream. Thus, comparing neural recordings from the ventral stream with different layers in the models will be necessary for refining models that are fully consistent with the brain computation.

Additional future direction of the study would be investigating diagnostic critical spatial frequency in ENNs for object recognition. It was previously observed that critical bands of spatial frequency were scale-dependent except for face images (Oruç and Barton, 2010). The critical spatial frequency was measured by testing visual recognition of objects embedded in noise. Though the scale channel selected in ENNs depends on the object size, our results predict that critical frequency is scale-invariant since spatial frequency is normalized by the object size. Due to the different experimental setup, however, it is hard to directly compare our results with previous studies on spatial frequency. In particular, it is unclear how a background of noise would affect the scale channel selection in ENNs. Recognition of such images may involve multiple frequency channels to separate target objects from background. Therefore, analyzing the behavior of ENNs for more complex images will be relevant.

Our work on ENNs have implications for eye-movements. ENNs show greater positional invariance to low-resolution images, which suggests a particular strategy for driving saccades, from low to high frequency channels. Although for each fixation
only a small fragment of the input image is processed at high-resolution, information about the peripheral visual area extracted by low-resolution channels enables the models to plan the next saccade towards an informative position in the visual field. In this way an image can be efficiently processed without the need of processing the entire visual field at high-resolution (Gorodissky et al., 2018).

The computational strategy of ENNs also implies more robustness to clutter and attention to small parts of an image. It was shown (Volokitin et al., 2017) that a model similar to ENNs does not suffer from crowding at the fovea, regardless of background. On the contrary, CNNs fail to recognize the target if the background at testing is different than the background used at training. This suggests that ENNs in foveal learning condition are able to learn the target object independently of the background, and thus are more robust to clutter. In fact, for detection tasks, where localizing a small target in complex scenes is important, extracting features at multi-scale has proven particularly useful (Lin et al., 2017). Due to the nature of the detection networks which are biased to identify only familiar object categories (Ren et al., 2015; He et al., 2017; Lin et al., 2017), this class of models are not comparable with human psychophysical data obtained from one-shot learning (of course, another discrepancy is that these models do not have resolution decreasing with eccentricity). However, we expect that if the models are modified to learn new additional categories easily, those with explicit multi-scale sampling (Lin et al., 2017) would require fewer examples than uniform sampling to learn to detect a new object as in ENNs. Moreover, after some training period of the object, multi-scale channels open up the possibility of selecting the channel that is the most relevant to the contextual information, as suggested by the human behavioral studies by Eckstein et al. (2017).
2.5 Methods

2.5.1 Psychophysical Experiments

**Stimuli and Setup** To create the stimuli set we used 27 Korean letters as target objects, each of them paired with another Korean letter as distractor, depicted in Figure 1A. For each trial, a sequence of one of the 27 target letters was shown first as target, followed by the test letter, which is the same letter or its pairing distractor. The letters were black Arial presented in different positions and sizes on a uniform white background in a 60 Hz Dell U2412M monitor. We used the Psychophysics Toolbox (Brainard, 1997) for MATLAB (Guide, 1998) running on a Linux computer. Subjects were seated at a distance of 1.26 m with a chin rest for stable viewing.

**Experimental Design** The experimental protocol was approved by the Massachusetts Institute of Technology Committee on the Use of Humans as Experimental Subjects (COUHES), and all experiments were carried out in accordance with the approved guidelines and regulations. Subjects provided informed written consent before the experiment.

*Scale-invariance Experiment* To test scale-invariance, both target and test letters were presented at the center of the monitor, and the size of letters was varied. We pursued two blocks of experiments to test invariance to scale in recognition. In the first scale experiment block we tested letter sizes of $30'$ and $2^\circ$. Specifically, the combinations set of target and test letter sizes were $(30', 30')$, $(30', 2^\circ)$, and $(2^\circ, 30')$, in which the first element represents the target letter size, and the second the test letter size. Similarly, in the second scale experiment block we used letter sizes of $30'$ and $5^\circ$ with combinations of target and test sizes $(30', 30')$, $(30', 5^\circ)$, and $(5^\circ, 30')$, respectively. The same group of subjects participated in both blocks of scale experiments, with at least a day apart to ensure that the subjects did not remember the stimuli set.

*Translation-invariance Experiment* Translation-invariant recognition was evaluated by keeping the size of target and test letters constant and changing the position
of test letters with respect to the position of target letters. We divide the tested conditions into two categories:

1) Learning in central vision, where target letters were presented at the subject’s visual fixation point, which was in center of the monitor. In this condition, test letters were presented in the same position as the target (represented as \((0 \rightarrow 0)\)) or at the subject’s visual periphery. We indicate the latter as \((0 \rightarrow D)\), in which 0 is the target position at the center of the screen, and D indicates the eccentricity in visual degrees of the test letter position from the fixation point.

2) Learning in peripheral vision, where target letters were presented at the subject’s visual periphery. Then, the test letter appeared at the same eccentricity as the target letter (represented as \((D \rightarrow D)\)), at the center, \((D \rightarrow 0)\), or at the opposite side with the same eccentricity as the target letter, represented as \((D \rightarrow \text{Opp})\).

We tested both conditions of central and peripheral vision with: i) eccentricities \(D = 1, 2, 3^\circ\) with constant letter size of 30’, ii) eccentricities \(D = 2, 2.5^\circ\) with letter size of 1°, iii) eccentricities \(D = 2, 4, 5, 7^\circ\) with letter size of 2°. We tested larger letters for a wider range of displacement to reflect that the range of visibility increases linearly with the letter size (Anstis, 1974).

Since translation-invariance experiments had more conditions than scale-invariance experiments, and the same set of 27 Korean letters was used, the set was repeated in two separate sessions. First, subjects were tested on 27 trials and instructed to come back for the second session after taking a break of at least 40 minutes, to ensure that they did not remember the letters.

Also, we designed translation-invariance experiments such that the same group of subjects participated in two or three eccentricities of displacement for the same letter size, again with at least a day apart between two displacement conditions. The repetition was limited to three times to prevent subjects from developing familiarity with the stimuli, while enabling us to isolate the effect of displacement on the degree of invariance from subjects’ individual difference. Specifically, the same group of subjects participated in all conditions for 30’ letter size, and another group in all conditions for 1° letter size. For 2° letters, the same subject group was tested for D
= 2° and 7°, and another group for D = 4° and 5°. The subjects that participated in translation-invariance experiments were different from the group participated in the scale-invariance experiments.

**Participants** In order to examine the degree of invariance in a one-shot learning task, it is crucial that the stimuli were novel objects to subjects. We recruited participants in the experiments who were not familiar with Korean letters. All subjects had normal or corrected-to-normal vision. We tested 10 subjects for the scale-invariance experiments, and between 11 and 12 subjects for the translation-invariance experiments (for 30′ letter conditions: 12 subjects, 1° letter conditions: 11 subjects, 2° letter conditions for D={2°, 7°} and D={4°, 5°}: 12 and 11 subjects, respectively). If a subject performed worse than 0.6 accuracy performance for the trivial condition, where target and test letters were the same size presented at the center, (0→0), the subject was excluded from further analyses. Since the same group of subjects participated in two or three displacement conditions for comparison, if a subject performed below the baseline criteria for one displacement condition, the subject was excluded from other displacement conditions as well. After excluding the subjects below the baseline criteria, for scale-invariance experiments, 10 subjects were included. For translation-invariance experiments, 9 subjects per condition were included for 30′ letter conditions, 11 subjects per condition for 1° letter size, and 10 subjects per condition for 2° letter size.

We also tested 3 Korean subjects to confirm that the designed task is trivial and find the range of visibility window for subjects who have prior experience and memory of Korean letters. Note that for Koreans, we used the same experimental setup and task; yet, it was not testing invariant object recognition in one-shot learning, but visibility of the letters in different sizes and positions.

**General Experimental Procedure** Accuracy for recognizing letters was measured in a same-different task. Subjects were instructed to first fixate a black dot at the center of the screen. After 1 sec, the fixation dot disappeared and a target letter
was presented for 33 msec, followed by a white screen for 1 sec. Then, the fixation
dot reappeared for 1 sec, followed by a test letter for 33 msec, again followed by
a white screen for 1 sec. Finally, the question of the task appeared, in which the
subject was asked if the target and test letters displayed previously were the same
or different. In Figure 1C a sample sequence of letter presentations is shown. Every
trial was composed of new letter pairs, and randomly choosing if the test letter was
the same as the target or the distractor. The presentation time was limited to 33
msec to avoid eye movements, which ensured that the subjects would view the letters
at the designed eccentricity.

In both scale- and translation-invariance experiments, the order of stimuli was
randomized. The number of same and different trials as well as presentation on the
left and right visual field was balanced. Each condition had the same number of trials.

2.5.2 Model Experiments

To contrast the human behavioral data on invariance with computational modeling
results, we evaluate Eccentricity-dependent Neural Network (ENN), which was pro-
posed by Poggio et al. (2014) and previously studied in Volokitin et al. (2017); Chen
et al. (2017). In particular, we demonstrate that ENN is robust to change in scale, and
validate that it captures the major characteristics of translation-invariance observed
from human experimental data. We test a Convolutional Neural Network (CNN) as
a control to show that invariance properties of ENN, especially scale-invariant rep-
resentation of novel stimuli, are derived from the architectural design of the model
rather than a consequence of training with multiple scales and positions.

Models

*Eccentricity-dependent Neural Network (ENN)* ENN (depicted in Fig. 2-5) builds on
two key properties of retinal sampling (Poggio et al., 2014). One is that there are
receptive fields of different sizes for a specific position (McFarland et al., 2016), and the
other one is that the size of receptive fields for each position increases with eccentricity
(Freeman and Simoncelli, 2011). The model achieves invariance through weight-sharing and pooling across different positions and scale channels. As we hypothesized that the model captures invariant representations to transformations, we tested this model for the comparison with behavioral data on invariant object recognition.

On the implementation level, ENN is based on a CNN. The primary difference between ENN and CNN is that the input to ENN is multi-scaled centered crops of the input images. Figure 5B shows an example set of multi-scaled crops of input images. This way, the center of an image, which corresponds to the foveal region, is sampled at multiple resolutions. The peripheral part of an image is sampled only at a low resolution. Different scale channels have shared weights and in addition to spatial pooling, the model has pooling over different scales. For the results of simulations we partly used the implementation provided by Volokitin et al. (2017).

ENN that we tested has four layers and a fully connected layer at the end, resembling V1-V2-V4-IT-PFC in the human ventral stream. The size of stimuli or receptive fields are measured in pixels, so we introduced a hyperparameter for the conversion between number of pixels and visual angle, which is 450 pixels to 1°. With this correspondence, we could compare modeling results with human data more directly. For instance, to extract features of 30' letters, we placed letters of size 225 pixels in the simulated visual field for the model. As discussed previously, the input to the model is multi-scaled centered crops of images, and we use 10 crops, increased in size exponentially by a factor of 1.5. The entire visual field processed by the model is approximately 19°.

We tested different convolutional and pooling schemes over space and scale, and here we have reported the one that matched human behavioral data most closely. The first layer has a kernel size of $11 \times 11$ pixels convolution with a stride of 4 pixels and $5 \times 5$ pixels spatial pooling with a stride of 2 pixels. Other layers have a convolutional kernel size of $5 \times 5$ pixels with a stride of 1 pixel and a pooling kernel size of $5 \times 5$ pixels with a stride of 2 pixels. When scale-pooling was used on top of spatially pooled features i.e. to explain scale-invariance or to extract features of the test letters, 10 scale channels were max-pooled at the last layer.
When choosing parameters of the network, we confirmed that ENN and human psychophysical data empirically matched by comparing the window of visibility for digit recognition. For 30′ digits, it was measured that at around 10° from the center of the fovea, recognition accuracy was 67% for humans (Strasburger et al., 1991). If we do a linear interpolation for approximation, accuracy would be about 77% at around 7° for the same size of digits. Using our conversion ratio between pixels and visual angle, we observed accuracy of 72% for 30′ MNIST digits at 7° for ENN, roughly matching the human accuracy. This conversion ratio together with the parameters in the network are also consistent with the theoretically estimated size of the smallest receptive fields (Marr et al., 1980).

**Convolutional Neural Network (CNN)** The parameters used in CNN were the same as ENN, except that there was no multi-crop input channels or pooling over scales, since the model had only one scale channel. The resolution of the input to the model was chosen such that it matched that of the 5th scale channel in ENN, which is its mid-resolution.

**Statistical Analysis**

No statistical methods were used to predetermine sample sizes (number of subjects), but our sample sizes are similar to those reported in previous studies using similar experimental procedures (studies testing recognition of familiar letter stimuli (Strasburger et al., 1991; Watson and Ahumada, 2015; Anstis, 1974) and testing invariant recognition of objects (Furmanski and Engel, 2000; Dill and Edelman, 2001)). We analyzed the percentage of correct responses, combining both same and different trials. For all parametric tests, data distribution was assumed to be normal, but this was not formally tested. To analyze the difference in mean accuracy among three or more conditions, we computed analyses of variance (ANOVAs) or repeated measures ANOVAs, depending upon whether the data were acquired from different group of subjects or the same groups, respectively. Correlation between features in simulations was Pearson’s $r$. 

50
2.6 Supplementary Information

2.6.1 SI Experiment 2: Translation-invariance experimental results

We tested two categories for learning: learning at the central visual field and learning at the peripheral visual field, based on the position where the target object is learned. When we compared mean accuracy for a condition with another condition tested in the same subject group, we applied two-tailed paired t-tests. Results are shown in Figure 2-8. The same results are visualized in a color scale in Figure 2-3.

**Learning in the Central Visual Field** For the specific presentation conditions, we use the notation (0→D), where 0 represents that target letters are shown at 0°, which is the fixation point, and D indicates the eccentricity in the peripheral visual field at which test letters are presented. The range of D was varied depending on the letter size.

**Accuracy for (0→D)** When target letters were presented at the center and test letters at the periphery, accuracy for every combination of size and position was lower than the baseline condition (0→0). Statistically, this difference between (0→D) with (0→0) was significant for 30′ letter size presented at eccentricity D = {1°, 3°}, and 1° letters with D = 2° (t(8) = 2.31, p = 0.05; t(8) = 3.54, p = 0.01; t(10) = 3.32, p = 0.01). To test whether limited translation invariant recognition is mainly due to the lack of transferability between two positions or noisy representation of peripheral vision, we examined accuracy for (D→D), which represents the limit on acuity in one-shot learning. The difference between the conditions (0→D) and (D→D) was not significant in any of the conditions (all ps > 0.05). Moreover, we observed lower accuracy for the (0→D) condition for 30′ letter size than for 1° or 2° letters, where there was more limit on acuity, which indicated that imperfect translation-invariant recognition for (0→D) may be attributable to lower acuity. We consider the effect of lower acuity in one-shot learning and transferability to another position altogether as the factors of limited translation-invariance since we confirmed that Koreans did not
Figure 2-8: Translation-invariance experimental results for non-Koreans. Three letter sizes (30', 1°, 2°) were tested and the range of translation was varied depending on the letter size. For each letter size and displacement, 5 different testing conditions, where target and test letters were shown at different positions, are plotted. Target and test letter positions are indicated below each bar. Opp represents testing at the opposite of the visual field in the same eccentricity as the target letter. Error bars represent standard error (Number of subjects n = 9 for 30' letter, n = 11 for 1° letter, and n = 10 for 2° letter size conditions. *p < 0.05, **p < 0.01, ***p < 0.001 two-tailed paired t-test).
have limitation on both factors under the same testing conditions (Figure 2-9).

Learning at the Peripheral Visual Field

To further explore the extent of translation-invariance in the limited acuity conditions in one-shot learning, we next investigated the conditions when target letters are learned in the peripheral visual field. These include conditions (i) \((D\rightarrow0)\), in which the target letter is presented at an eccentricity of \(D^\circ\), and the test letter at the fixation point; (ii) \((D\rightarrow\text{Opp})\) in which the target letter is learned at eccentricity \(D^\circ\), and test letter is presented at the same eccentricity but in the opposite site of the visual field, and the baseline \((D\rightarrowD)\) in which both target and test letters are presented in the same position.

Accuracy for \((D\rightarrow0)\) We observed that recognizing test letters invariant to position was limited when target letters were learned at the periphery and test letters were presented at the fixation point. Accuracy for \((D\rightarrow0)\) condition was lower than that for the control condition \((D\rightarrowD)\) for \(1^\circ\) and \(2^\circ\) letter sizes. Statistical evaluation
also confirmed that the difference was statistically significant \((p < 0.05\) in all cases, except for \(2^\circ\) letters at \(D = 7^\circ\), where \(p < 0.06\); specifically, for \(1^\circ\) letters at \(D = \{1, 2^\circ\}\): \(t(10) = 3.00, 3.25, p = 0.01, 0.01\), respectively; for \(2^\circ\) letters at \(D = \{2, 4, 5, 7^\circ\}\): \(t(9) = 2.49, 2.32, 2.54, 2.20, p = 0.03, 0.05, 0.03, 0.06\), respectively). However, we did not observe the same effect for \(30^\prime\) letter size, for which the accuracy for \((D \rightarrow 0)\) condition was lower than \((D \rightarrow D)\) for \(D = \{1^\circ, 2^\circ, 3^\circ\}\), but only for \(D = 1^\circ\) was the difference statistically significant \((t(8) = 3.2, p = 0.01)\). This different behavior for small letters \((30^\prime)\) than larger letters \((1^\circ\) and \(2^\circ\)) might be due to the fact that larger letters have a mean recognition accuracy for the baseline condition, \((D \rightarrow D)\), higher than 0.80. On the other hand, the mean accuracy for \(30^\prime\) letter size for \((D \rightarrow D)\) with \(D = \{2^\circ, 3^\circ\}\) was much lower than 0.80 (both 0.71). This suggests that recognition for \(30^\prime\) letter size at the periphery in the same eccentricity for both target and test letters is limited by acuity in one-shot learning. The small difference with respect to \((D \rightarrow 0)\) might be due to such limited acuity. Thus, unless there was a pronounced limit on visual acuity at a particular position in one-shot learning, causing accuracy for \((D \rightarrow D)\) to be about or lower than 0.80, we observed significantly limited translation-invariant recognition.

Accuracy for \((D \rightarrow Opp)\) We next sought to test invariant recognition when the position of letters change, but not their resolution. To evaluate this condition, we tested the condition \((D \rightarrow Opp)\), where letters were translated to the opposite side of the visual field while preserving their distance from the central fixation point. Since both target and test letters were presented in the peripheral visual field for \((D \rightarrow Opp)\), accuracy could be limited by the acuity in one-shot learning, corresponding to the tested position. Thus, we compared accuracy for \((D \rightarrow Opp)\) with that for \((D \rightarrow D)\) to find whether there was a loss in recognition due to displacement in addition to the limit on acuity in one-shot learning. For \(30^\prime\) letter size, mean accuracy for \((D \rightarrow Opp)\) was slightly higher than \((D \rightarrow D)\) for \(D = \{1^\circ, 2^\circ\}\) and lower than \((D \rightarrow D)\) for \(D = 3^\circ\), yet none of them was statistically significantly different (all \(ps > 0.5\)). For \(1^\circ\) and \(2^\circ\) letter size, the mean accuracy for \((D \rightarrow Opp)\) was lower than the baseline condition \((D \rightarrow D)\). However, except for the setting of \(2^\circ\) letter size with \(D = 4^\circ\) \((t(9) = 2.49, p = 0.03)\),
none of them resulted in a statistically significant drop in accuracy (all $p > 0.05$). These results suggest that acuity has a more significant effect on recognizing 30’ letters than larger letters for (D→Opp) condition. Discriminating 1° or 2° letters had a marginal performance loss due to displacement to the opposite side of the visual field.

### 2.6.2 Asymmetry between Central and Peripheral Learning

One notable result is that the order for which condition letters were learned first had an effect in invariant recognition. For every size and position of letters, the mean accuracy for (0→D) was higher than that for (D→0). In statistical evaluation, 1° letters with D = 2° and 2.5°, and 2° letters with D = 7° had a statistically significant difference ($t(10) = 2.47, p = 0.03$; $t(10) = 2.97, p = 0.01$; $t(9) = 5.55, p = 0.0003$, respectively). As argued above, this might be because the acuity in one-shot learning limits the learnability to the level of accuracy at (D→D).
Figure 2-10: Scale-invariance experimental results, showing $d'$. We computed $d' = Z(\text{Hit}) - Z(\text{False alarm})$, where $Z$ is the inverse of the cumulative Gaussian distribution. Hit and false alarm rates were the average across all subjects’ data ($n = 10$). To bound the values of $d'$, we added 0.5 to both hit and false alarm rates and 1 to both the number of signal trials and the number of noise trials (Hautus 1995). The overall results showing robustness to scaling is consistent with Figure 2-2.
Figure 2-11: Windows of invariance, showing $d'$. As in Figure 2-3, we show $d'$ for three different conditions (central, peripheral and opposite learning). Consistent with Figure 2-3, we find that the range of invariance increases with scale for the central learning condition. Also, peripheral and opposite learning conditions result in a narrower range of invariance than the central learning condition.
Figure 2-12: Windows of visibility, showing $d'$. As in Figure 2-4, we show that the range of invariance is extended with experience. Overall results are consistent with Figure 2-4, except unlike in Figure 4, we observe a narrower range for $2^\circ$ letters than that for smaller letters for some contours of $d'$ values (e.g. $d' = 2$). However, for smaller $d'$ values, this is not the case. Also, the results are consistent with our main conclusion that the window of visibility is wider than the window of invariance.
SI: Scale-invariance supplementary results for ENN

Figure 2-13: Simulation results on scale-invariance using multi-scale features. Testing conditions are identical to those for Figure 2-6. Instead of using max-pooled features, we use features from all scale channels for target letters. See the results section.

SI: Translation-invariance supplementary results for ENN

Figure 2-14: Simulation results for translation-invariance. (0→D) indicates the central learning conditions, when target letters are placed at the center of the visual field and test letters at an eccentricity D° in the peripheral visual field. (D→0) indicates the reverse order of testing which is the peripheral learning conditions. To reproduce asymmetry in recognition rates between central and peripheral learning ENN templates associated with target letters are from all scales. Each scale is matched with the features from test letters, which are pooled over scale channels, to find a scale that maximizes the correlation. A linear fit is plotted for ENN.
Figure 2-15: Translation-invariance simulation results from using features pooled over all scales and positions. This method has limitation to explain the asymmetric recognition accuracy between central and peripheral learning. In the main text, we instead use features from all scales to reproduce translation-invariance experimental results. See the results section.

Figure 2-16: Pearson correlation between same Korean letters at different positions. See the results section. Error bars represent standard deviation (Number of letters n = 27).
Chapter 3

System Identification of Neural Systems

The contents of this chapter are adapted from Yena Han, Tomaso Poggio, and Brian Cheung, "System identification of neural systems: If we got it right, would we know?", *International Conference on Machine Learning*, PMLR (2023).

3.1 Abstract

Artificial neural networks are being proposed as models of parts of the brain. The networks are compared to recordings of biological neurons, and good performance in reproducing neural responses is considered to support the model’s validity. A key question is how much this system identification approach tells us about brain computation. Does it validate one model architecture over another? We evaluate the most commonly used comparison techniques, such as a linear encoding model and centered kernel alignment, to correctly identify a model by replacing brain recordings with known ground truth models. System identification performance is quite variable; it also depends significantly on factors independent of the ground truth architecture, such as stimuli images. In addition, we show the limitations of using functional similarity scores in identifying higher-level architectural motifs.
3.2 Introduction

Benchmarking in machine learning

Over the last two decades, the standard approach for machine learning engineers seeking better model performance has been to utilize predefined benchmarks to rank models from most to least relevant on specific tasks. This practice of model comparison through benchmarks has been instrumental in driving much of the rapid progress in the machine learning community. For example, in the field of computer vision, benchmark datasets used to evaluate models have evolved from relatively simple ones such as the MNIST dataset for handwritten digit recognition, to more complex datasets such as CIFAR-10 and ImageNet for object classification tasks. Similarly, benchmark datasets like the PASCAL VOC and Microsoft COCO have enabled standardized evaluation of models on additional computer vision tasks like object detection and segmentation.

The broad availability of standardized benchmarks has enabled validation and comparison of successful ideas at scale across the machine learning community. Models and techniques that improve benchmark performance become widely adopted. As a result, benchmarking has become a central paradigm in machine learning research and application development.

Benchmarking in neuroscience

Recently, the benchmarking paradigm has found its way into computational neuroscience as well with the advent of experimental frameworks like Brain-Score and Algonauts (Schrimpf et al., 2020; Cichy et al., 2021). These frameworks allow artificial models to be tested and compared on how closely they can predict neural recordings from human or animal brains, such as macaques.

For example, the Brain-Score framework provides a standardized way to assess how well computational models can predict neural responses in the primate visual system. Models are evaluated on their ability to predict recordings of real neurons or match behavioral data. Similarly, the Algonauts challenge involves predicting fMRI
measurements in humans in response to a dataset of natural scenes.

An open research question is whether and to what extent these engineering-style benchmarking approaches can be useful for driving progress in the natural sciences like neuroscience, as they have been in machine learning. There are important differences between the goals and constraints involved.

Limitations of benchmarks

For pure machine learning applications, benchmark tasks are often chosen such that they are immediately useful for real-world applications or closely related to downstream tasks where models will be deployed. Although there have been some studies suggesting limitations of model evaluation based solely on performance on standard datasets, or pointing out potential disconnects between benchmarks and real-world problems, they primarily argue for the need to improve benchmark design and mitigate overfitting, rather than questioning the overall utility of comparison frameworks and leaderboards (Barbu et al., 2019).

On the other hand, some of the benefits of benchmarking for computational neuroscience are less obvious or direct. Models that rank highly on predictive ability may be practically useful for applications, such as controlling neurons (Bashivan et al., 2019). However, developing models that closely match natural intelligence requires incorporating constraints from biology and neuroscience, like anatomical connectivity and biophysics. A model that accurately predicts responses but requires non-biological mechanisms would be considered falsified as an explanation of the brain, even if highly ranked in benchmarks.

Can function predict structure?

This raises the question of whether close functional similarity in a benchmarking scenario could serve as a reliable indicator of similarities in internal model structure and mechanisms. As an example, consider the conjecture that if two artificial model architectures, like convolutional neural networks and transformers, are trained on the same data and end up with similar representational abilities to predict brain
responses, we could conclude that the brain must be implementing computational motifs similar to one or both of those architectures.

For this to be true, the functional similarity over a large benchmark dataset would have to effectively constrain the architectural interpretation — the need for separate validation of biological constraints at the anatomical level would be less critical. We can examine this by testing whether standard similarity metrics can reliably distinguish between models with very different underlying motifs when presented with data generated from known architectures.

**Confounding factors in functional similarity**

When artificial models are compared on a common biological prediction benchmark (Yamins and DiCarlo, 2016), the typical interpretation is that models ranked higher have greater relevance as explanations of the brain. However, if multiple models with very different underlying mechanisms happen to predict neural responses equally well, the cause of the functional similarity remains unclear. It could reflect biologically relevant computational motifs, but could also result from similarities in the stimulus dataset or limitations of the similarity metric. This demonstrates there is not a one-to-one mapping between benchmark score and model interpretation. Our work indicates that multiple factors can give rise to representational similarity.

An interesting example is offered by Chang et al. (2021), which found that a 2D morphable face model performed best at predicting responses in inferotemporal cortex to faces. However, the operations required by this model, like correspondence matching and image vectorization, do not map clearly onto known biological neuronal mechanisms. This highlights that constraints and confounds beyond biological plausibility can affect predictive ability on a benchmark.
3.3 Related Work

Validation of deep neural networks as models of the brain

Previous studies have compared activations of deep neural networks, convolutional neural networks (CNNs) in particular, with neural data recorded from humans or animals (Yamins et al., 2014; Kriegeskorte et al., 2008). This correspondence between artificial and biological neural networks is not limited to the visual processing pathways responsible for object recognition. It also extends to neural circuits involved in action recognition, scene understanding, and other visual tasks (Tacchetti et al., 2017; Dwivedi et al., 2021).

Furthermore, similar comparisons between deep neural networks and biological neural data have been applied to sensory domains beyond vision. For example, in the auditory domain, Kell et al. (2018) showed that task-optimized neural networks can predict fMRI voxel responses in human auditory cortex and reproduce errors consistent with those observed in humans during natural sound processing. More recently, in the domain of human language processing, Schrimpf et al. (2021) showed that the activations of autoregressive language models have explained variance nearing 100% when compared to neural data recorded during naturalistic narrative listening or reading tasks in humans. This demonstrates that the parallels between artificial and biological neural networks are not limited to sensory systems, but extend to higher-level cognitive processing.

The functional alignment between deep neural networks and the brain has motivated research into driving neural activations. For example, Bashivan et al. (2019) showed that targeted neural populations in cortical area V4 in macaque could be activated beyond their natural firing rates using images synthesized based on artificial units in deep neural networks that mapped to the biological neurons.

Implications for Architecture

While the analogy between neural network models and the brain has been well validated, the extent of this correspondence across multiple levels (Marr and Poggio,
1976) has not been fully understood. Yet, previous works have often drawn implications about computational properties of the brain directly from functional correspondences at the representational level, making connections to implications at the architectural level of analysis. One previous study (Kar et al., 2019) argues for recurrent connections in the ventral visual stream, partly based on the higher neural predictivity of recurrent neural networks compared to shallower neural networks. Another study (Nonaka et al., 2021) observed that purely feedforward neural networks outperform those with branching or skip connections in terms of a similarity measure on neural activations. They concluded that the results indicate the significance of feedforward connections in the brain. Additionally, since fully-connected layers, which can integrate visual features spatially, better match certain brain areas, they reason that these areas may be responsible for spatial integration.

The unclear implications for brain architecture from functional similarity are discussed in St-Yves et al. (2023). By comparing various types of hierarchical connections of deep neural networks with fMRI data, they find that non-serial hierarchical connections can also yield high neural predictivity. In their case, based on well-established findings on the visual cortex showing hierarchical connections between visual areas, they can relatively safely attribute the equally high accuracy for multi-branch models as a disconnect between representational similarity and anatomical constraints. However, it is often the case that we have mixed or unclear findings about the biology. Especially going forward, in order to use computational models to expand our experimental works, it is important to understand the capabilities and limitations of drawing architectural implications from representational correspondence.

Potential methodological limitations in simultaneously evaluating models across all levels have complicated our understanding of the disconnect between representational space and biological constraints. Jonas and Kording (2017) investigated the robustness of standard analysis techniques in neuroscience with a microprocessor as a ground-truth model to determine the boundaries of what conclusions could be drawn about a known system. The presumption of correspondence could also be attributed to underappreciated variability from model hyperparameters (Schaeffer et al., 2022).
Prinz et al. (2004) showed that many different network configurations could lead to the same circuit behavior demonstrating a potential many-to-one nature for this correspondence. In a similar spirit to Jonas and Kording (2017); Lazebnik (2002), we evaluate system identification on a known ground-truth model to establish the boundaries of what architectural motifs can be reliably uncovered. We perform our analysis under favorable experimental conditions to establish an upper bound.

**Representational Similarity Measures in Machine Learning**

As modern neural network models have grown more prominent in unison with the corresponding resources to train these models, pre-trained reference models have become more widely available in research (Wightman, 2019). Consequently, the need to compare these references along different metrics has followed suit. Kornblith et al. (2019); Morcos et al. (2018) explored using different similarity measures between the layers of artificial neural network models. Kornblith et al. (2019) propose various properties a similarity measure should be invariant such as orthogonal transformations and isotropic scaling, while not invariant to invertible linear transformations. Kornblith et al. (2019) found centered kernel alignment (CKA), a method very similar to Representation Similarity Analysis (Kriegeskorte et al., 2008), to satisfy these requirements best. Ding et al. (2021) explored the sensitivity of methods like canonical correlation analysis, CKA, and orthogonal procrustes distance to changes in factors that do not impact the functional behavior of neural network models.

### 3.4 Methods

In computational neuroscience, we usually have a neural system (brain) that we are interested in modeling. We call this network a *target* and the proposed candidate model a *source*. Given the same visual stimuli to the target and source systems, we would like to evaluate how well the activation profiles of the source network match those of the target. The two predominant approaches for evaluating computational models of the brain use metrics based on linear encoding analysis for neural pre-
dictivity and population-level representation similarity. The first measures how well a model can predict the activations of individual units, whereas the second metric measures how correlated the variance of internal representations is. We study the following neural predictivity scores consistent with the typical approaches: Linear encoding analysis and Centered Kernel Alignment (CKA).

### 3.4.1 Linear Encoding Analysis

One common approach in linear encoding analysis assumes that different visual areas or layers have distinct feature spaces (Schrimpf et al., 2020). This suggests using source features from a single layer when comparing to target activations.

Another assumption the metric is based on is that we measure similarities between activations under linear transformation. The rationale is that one-to-one neuron mappings are unlikely, even across animals of the same species (Yamins and DiCarlo, 2016). Rather, the expectation is that feature vectors representing neural activations will share equivalent bases between the two systems.

On a related note, although linear encoding models are widely used, there has been discussion in the community about whether linear models are unnecessarily contrived, and non-linear models should be considered (Ivanova et al., 2022).

While linear encoding models are widely used, there has been some discussion on whether linear models are overly simplistic, and nonlinear approaches should be considered instead (Ivanova et al., 2022). However, since the experimental data we simulate involves mainly visual processing in the visual cortex, and not more complex cognitive functions, here our evaluation process uses linear models.

Based on these assumptions, we can formulate the setup as:

\[
Y = Xw
\]  

(3.1)

where \( Y \in \mathbb{R}^{n \times p_y} \) denotes the matrix representations of a target system over \( n \) stimulus images, \( X \in \mathbb{R}^{n \times p_x} \) is the matrix of representations with \( p_x \) features over the same set of images, and \( w \) is the linear regression weights.
For regularization, previous studies have applied variants of linear regression, such as partial least squares regression or performing principal component analysis first on the activations and then applying linear regression (Schrimpf et al., 2020; Yamins et al., 2014; Conwell et al., 2021; Kar et al., 2019; Mitchell et al., 2008). To reduce computational costs without sacrificing predictivity, we apply sparse random projection $S \in \mathbb{R}^{p_x \times q_x}$ for $q_x << p_x$, on the activations of the source model (Conwell et al., 2021). This projection reduces the dimensionality of the features to $q_x$ while still preserving relative distances between points (Li et al., 2006). Unlike principal component analysis, sparse random projection is a dataset-independent dimensionality reduction method. This removes any data-dependent confounds from our processing pipeline for linear regression and isolates dataset dependence to our variables of interest: the linear regression and candidate model.

We apply ridge regression on every layer of a source model to predict a target unit using these features. We use nested cross-validation in which the regularization parameter $\lambda$ is chosen in the inner loop and a linear model is fitted in the outer loop. The list of tested $\lambda$ values is $[0.01, 0.1, 1.0, 10.0, 100]$. We use 4-fold cross-validation for inner loops and 5-fold cross-validation for outer loops.

The final linear regression score for each target unit is the Pearson’s correlation coefficient $r(\cdot, \cdot)$ between the predicted responses of a source model and the ground-truth target responses to a set of stimulus images. That is,

$$\hat{\beta} = \arg\min_\beta \| Y - X S \beta \|_F^2 + \lambda \| \beta \|_F^2$$

$$LR(X, Y) = r(X S \hat{\beta}, Y)$$

As there are multiple target units, the median of Pearson’s correlation coefficients between predicted and true responses is the aggregate score for layer-wise comparison between source and target models. Note that a layer of a target model is usually assumed to correspond to a visual area, e.g., V1 or IT, in the visual cortex. Thus, for a layer-mapped model, we report the maximum linear regression scores across source layers for target layers.
3.4.2 Centered Kernel Alignment

Another widely used type of metric builds upon the idea of measuring the representational similarity between the activations of two neural networks for each pair of images. While variants of this metric abound, including Representational Similarity Analysis (RSA) or re-weighted RSA (Kriegeskorte et al., 2008; Khaligh-Razavi et al., 2017), we use CKA (Cortes et al., 2012) as Kornblith et al. (2019) showed strong correspondence between layers of models trained with different initializations, which we will further discuss as a validity test we perform. We consider linear CKA in this work:

\[
\text{CKA}(X, Y) = \frac{\| Y^T X \|^2_F}{\| X^T X \|_F \| Y^T Y \|_F} \quad (3.4)
\]

Kornblith et al. (2019) showed that the variance explained by (unregularized) linear regression accounts for the singular values of the source representation. In contrast, linear CKA depends on the singular values of both target and source representations. Recent work (Diedrichsen et al., 2020) notes that linear CKA is equivalent to a whitened representational dissimilarity matrix (RDM) in RSA under certain conditions. We also call CKA a neural predictivity score because a target network is observable, whereas a source network gives predicted responses.

3.4.3 Identifiability Index

To quantify how selective predictivity scores are when a source matches the target architecture compared to when the architecture differs between source and target networks, we define an identifiability index as:

\[
\text{Identifiability Index} = \frac{\text{score}(s = t) - \overline{\text{score}}(s \neq t)}{\text{score}(s = t) + \overline{\text{score}}(s \neq t)}
\]

where \( s \) is the source or candidate model, and \( t \) is the target model. In brief, it is a normalized difference between the score for the true positive and the mean score (\( \overline{\text{score}} \)) for the true negatives. Previous works (Dobs et al., 2022; Freiwald and Tsao,
Figure 3-1: **Illustration of the research question**: If one of the candidate models matches the target neural network (brain) closely, can the current model evaluation methods accurately find that out?

2010) defined selectivity indices in the same way in similar contexts, such as the selectivity of a neuron to specific tasks.

### 3.4.4 Simulated Environment

If a target network is a brain, it is essentially a black box, making it challenging to understand the properties or limitations of the comparison metrics. Therefore, to better analyze the metrics, we instead use artificial neural networks of our choice as target models for our experiments. The framework is illustrated in Figure 3-1.

We investigate the reliability of a metric to compare models, mainly to discriminate the underlying computations specified by the model’s architecture. We intentionally create favorable conditions for identifiability in a simulated environment where the ground truth model is a candidate among the source models. Taking these ideal conditions further, our target and source models are deterministic and do not include adverse conditions typically encountered in biological recordings, such as noise and temporal processing.
Figure 3-2: Linear regression scores of deep neural networks for brain activations in the macaque visual cortex. Architecture list in Section 3.7.1. For V1, the top performing three models are in the VOneNet family (Dapello et al., 2020), which are explicitly designed to mimic the known properties of V1.

3.5 Results

3.5.1 Different models with equivalent neural predictivity

We compare various artificial neural networks with publicly shared neural recordings in primates (Majaj et al., 2015; Freeman et al., 2013) via the Brain-Score framework (Schrimpf et al., 2020).

Our experiments show that the differences between markedly different neural network architectures are minimal after training (Figure 3-2), consistent with the previous work (Schrimpf et al., 2020; Kubilius et al., 2019; Conwell et al., 2021). Previous works focused on the relative ranking of models and investigated which model yields the highest score. However, if we take a closer look at the result, the performance difference is minimal, with the range of scores having a standard deviation < 0.03 (for V2=0.021, V4=0.023, IT=0.016) except for V1. For V1, VOneNets (Dapello et al., 2020), which explicitly build in properties observed from experimental works in neuroscience, significantly outperform other models. Notably, the models we consider have quite different architectures based on combinations of various components, such as convolutional layers, attention layers, and skip connections. This suggests
that architectures with different computational operations reach almost equivalent performance after training on the same large-scale dataset, i.e., ImageNet.

### 3.5.2 Identification of architectures in an ideal setting

One potential interpretation of the result would be that different neural network architectures are equally good (or bad) models of the visual cortex. An alternative explanation would be that the method we use to compare models with the brain has limitations in identifying precise computational operations. To test the hypothesis, we focus on where underlying target neural networks are known instead of being a black box, as with biological brains. Specifically, we replace the target neural recordings with artificial neural network activations. By examining whether the candidate source model with the highest predictivity is identical to the target model, we can evaluate to what extent we can identify architectures with the current approach.

We consider the following architectures:

**Convolutional:** AlexNet (Krizhevsky et al., 2012), VGG11 (Simonyan and Zisserman, 2014), ResNet18 (He et al., 2016)

**Recurrent:** CORnet-S (Kubilius et al., 2019)

**Transformer:** ViT-B/32 (Dosovitskiy et al., 2020)

**Mixer:** MLP-Mixer-B/16 (Tolstikhin et al., 2021)

These architectures are emblematic of the vision-based models used today. Each architecture has a distinct motif, making it unique from other models. For example, transformer networks use the soft-attention operation as a core motif, whereas convolutional networks use convolution. Recurrent networks implement feedback connections which may be critical in the visual cortex for object recognition (Kubilius et al., 2019; Kar et al., 2019). Moreover, mixer networks (Tolstikhin et al., 2021; Touvron et al., 2021; Melas-Kyriaizi, 2021) uniquely perform fully-connected operations over image patches, alternating between the feature and patch dimensions.
Figure 3-3: Linear regression scores as a function of target network layer. Different initialization seeds are used for source networks, except for one (MLP-Mixer-B/16*), which uses identical weights. Black dots • indicate that the correct architecture does not outperform others with statistically significant differences ($p > 0.01$). Red dots • indicate that the median scores for those networks are higher than the median for the correct architecture. The ranking of source models is typically based on median scores (Schrimpf et al., 2020).

**Linear Regression**

We first compare various source models with a target network, the same architecture as one of the source models and is trained on the same dataset but initialized with a different seed. We test a dataset composed of 3200 images of synthetic objects studied in (Majaj et al., 2015) to be consistent with the evaluation pipeline of Brain-Score. The ground-truth source model will yield a higher score than other models if the model comparison pipeline is reliable.

For most target layers, source networks with the highest median score are the correct network (Figure 3-3). However, strikingly, for several layers in VGG11, ResNet18, and CORnet-S the best-matched layers belong to a source model that is not the correct architecture. In other words, given the activations of ResNet18, for instance, and based on linear regression scores, we would make an incorrect prediction that the system’s underlying architecture is closest to a recurrent neural network CORnet-S. While it has been argued that recurrent neural networks essentially correspond to
Figure 3-4: CKA for different source and target networks. The experimental setup is identical to Figure 3-3 besides using CKA instead of linear regression as the metric. As in Figure 3-3, when we test MLP-Mixer-B/16 as a target and the source network type matches the target, weights are identical. We show the results for MLP-Mixer-B/16 in bar plots with a pattern to indicate the difference from other targets.

ResNets with weight shared across layers (Liao and Poggio, 2016), the ResNet18 architecture we consider here does not have such weight-sharing. The confusion between ResNet18 and CORnet-S is especially noteworthy in that the prediction leads to an incorrect inference about the presence of recurrent connections in the target network.

In addition, because of our ideal setting, where an identical network is one of the source models, we expect to see a significant difference between matching and non-matching models. However, for multiple target layers, linear regression scores for the non-identical architectures, when compared with those for the identical one, do not show a significant decrease in predictivity based on Welch’s t-test with $p < 0.01$ applied as a threshold (Figure 3-3). This result suggests that the identification of the underlying architectures of unknown neural systems is far from perfect.
Centered Kernel Alignment

Next, we examine another widely used metric, CKA, for comparing representations. Again, we compare different source models to a target model, also an artificial neural network. For the target models we tested, the ground-truth source models achieve the highest score (Figure 3-4). Still, some unmatched source networks lead to scores close to the matched networks, even for the target MLP-Mixer-B/16, where the source network of the same architecture type also has identical weights.

When applying CKA to compare representations, we subsample a set (3000) of target units to mimic the limited coverage of single-unit recordings. Assuming we can increase the coverage for future experiments with more advanced recording techniques, we test whether the identifiability improves if we include all target units. Additionally, methods similar to CKA, such as RSA, are often applied to interpret neural recordings, including fMRI, MEG, and EEG (Cichy and Oliva, 2020; Cichy and Pantazis, 2017), which can have full coverage of the entire brain or a specific region of interest. Therefore, we simulate such analyses by having all units in the targets. Overall, the ground-truth source models outperform the other source models by a significant margin (Figure 3-7 in Section 3.7). This suggests system identification can be more reliable with more recording units and complete target coverage.

3.5.3 Effects of the stimulus distribution on identifiability

A potentially significant variable overlooked in comparing computational models of the brain is the type of stimulus images. What types of stimulus images are suited for evaluating competing models? In Brain-Score, stimulus images for comparing models of the high-level visual areas, V4 and IT, are images of synthetic objects (Majaj et al., 2015). In contrast, those for the lower visual areas, V1 and V2, are images of texture and noise (Freeman et al., 2013). To examine the effect of using different stimulus images, we test images of synthetic objects (3200 images) and texture and noise (135 images), and additionally, ImageNet (3000 images) and web-crawled images LAION (3000 images) (Schuhmann et al., 2021) which are more natural images than the first
Figure 3-5: **Top** Sample images of each stimulus image type. (a) Identifiability index using CKA and (b) linear regression for various types of stimulus images and target networks.
two datasets.

In Figure 3-5, we analyze Identifiability Index for different stimulus images. More realistic stimulus images, i.e., synthetic objects, ImageNet, and LAION, show higher identifiability than texture and noise images for all target models. We observe identifiability increases with layer depth. Notably, even for early layers in target models, which would correspond to V1 and V2 in the visual cortex, texture and noise images fail to give higher identifiability.

Additionally, testing LAION images leads to Identifiability Indices as high as, or often higher than, those obtained using ImageNet images, on which all target systems are trained. This result indicates that the stimuli do not need to be exactly in-distribution with the training data for better identification and suggests that natural image statistics may be sufficient.

It is important to note that the images of texture and noise used in the experiment help characterize certain aspects of V1 and V2, as shown in previous work (Freeman et al., 2013). Specifically, the original study investigated the functional role of V2 in comparison to V1 and demonstrated that naturalistic structure modulates V2. While the image set plays an influential role as a variable in a carefully designed experiment to address specific hypotheses, it does not serve as a sufficient test set for any general hypothesis, such as evaluating different neural networks.

### 3.5.4 Challenges of identifying key architectural motifs

Interesting hypotheses for a more biologically plausible design principle of brain-like models often involve key high-level architectural motifs. For instance, potential questions are whether recurrent connections are crucial in visual processing or, with the recent success of transformer models in deep learning, whether the brain similarly implements computations like attention layers in transformers. The details beyond the key motif, such as the number of layers or the exact type of activation functions, may vary and be underdetermined within the scope of such research questions. Likewise, it is unlikely that candidate models proposed by scientists align with the brain at every level, from low-level specifics to high-level computation. Therefore, an ideal
Figure 3-6: Different architectural variants (12 CNNs and 14 ViTs) are compared with two CNNs and one ViT target network. Each data point represents the maximum score of an architecture for the corresponding target layers. Solid markers indicate the mean score of the corresponding model class, and error bars display standard deviations. ✗ indicates that the corresponding layer does not show a statistically significant difference between model classes.

methodology for comparing models should help separate the key properties of interest while being invariant to other confounds.

Considering it is a timely question, with the increased interest in transformers as models of the brain in different domains (Schrimpf et al., 2021; Berrios and Deza, 2022; Whittington et al., 2021), we focus on the problem of identifying convolution vs. attention. We test 12 Convolutional Networks and 14 Vision Transformers of different architectures (list in Section 3.7.2), and to maximize identifiability, we use ImageNet stimulus images. Note that an identical architecture with the target network is not included as a source network.

Figure 3-6 shows that for both CKA and regression, there is high inter-class variance for many target layers. For CKA, one layer in VGG13, 7 layers in ResNet34, and 7 layers in ViT-L/16, and for regression, three layers in VGG13, 6 layers in ResNet34, and one layer in ViT-L/16 do not show a statistically significant difference between the two model classes based on Welch’s t-test with $p < 0.01$ used as a threshold. The significant variance among source models suggests that model class identification can
be incorrect depending on the precise variation we choose, especially if we rely on a limited set of models.

### 3.6 Discussion

Under idealized settings, we tested the identifiability of various artificial neural networks with differing architectures. We present two contrasting interpretations of model identifiability based on our results, one optimistic (Glass half full) and one pessimistic (Glass half empty).

**Glass half full:** Despite the many factors that can lead to variable scores, linear regression and CKA give reasonable identification capability under unrealistically ideal conditions. Across all the architectures tested, identifiability improves as a function of depth.

**Glass half empty:** However, system identification is highly variable and dependent on the properties of the target architecture and the stimulus data used to probe the candidate models. For architecture-wide motifs, like convolution vs. attention, scores overlap significantly across almost all layers. This indicates that such distinct motifs do not play a significant role in the score.

Our results suggest two future directions for improving system identification with current approaches: 1) Using stimuli images that are more natural, i.e., closer to the inputs to the target network (brain) in a natural setting. 2) With more neurons recorded in the brain, neural predictivity scores can be more reliable in finding the underlying architecture.

On the other hand, it is worthwhile to note that we may have reached close to the ceiling using neural predictivity scores for system identification. As an example, when our source network is AlexNet, its regression scores against the brain (Figure 3-2) are on par with, or slightly higher than, the scores against another AlexNet (Figure 3-3). In other words, based on the current methods, AlexNet predicts the brain as well as, if not better than, predicting itself. This observation is not limited to AlexNet but applies to other target networks. This fundamental limitation of present evaluation
techniques, such as the linear encoding analysis used in isolation, emphasizes the need to develop new approaches beyond comparing functional similarities.

As we argued earlier, ranking models in terms of their agreement with neural recordings is the first step in verifying or falsifying a neuroscience model. Since several different models are very close in ranking, the next step – architectural validation – is the key. Furthermore, it may have to be done independently of functional validation and with little guidance from it, using standard experimental tools in neuroscience. A parallel direction is, however, to try to develop specially designed, critical stimuli to distinguish between different architectures instead of measuring the overall fit to data. As a simple example, it may be possible to discriminate between dense and local (e.g., CNN) network architectures by measuring the presence or absence of interactions between parts of a visual stimulus that are spatially separated.

3.7 Supplementary Information

3.7.1 Model details for Section 2.4.1: Brain-Score

Below is the full list of models tested on the benchmarks of Brain-Score as reported in Section 2.4.1. In addition to testing vision models pre-trained on ImageNet available from PyTorch’s torchvision model package version 0.12, we test VOneNets that are pre-trained on ImageNet and made publicly available by the authors (Dapello et al., 2020). VOneNets are also a family of CNNs.

**Convolutional Networks:** AlexNet, VGG11, VGG13, VGG19, ResNet18, ResNet34, ResNet50, ResNet101, VOneAlexNet, VOneResNet50, VOneCORnet-S

**Transformer Networks:** ViT-B/16, ViT-B/32, ViT-L/16, ViT-L/32

3.7.2 Model details for Section 2.4.4: Finding the key architectural motif

For each target network reported in Section 2.4.4, namely VGG13, ResNet34, and ViT-L/16, below is the full list of source models tested to compare two model classes,
CNN and transformer. For Tokens-to-token ViTs (T2T) (Yuan et al., 2021), we use models released by the authors. We use Twins Vision Transformers and a Visformer from timm library (Wightman, 2019). All other models are available from PyTorch’s torchvision model package version 0.12. All models are pre-trained on ImageNet.

**Convolutional Networks:** AlexNet, VGG11, VGG13, VGG16, VGG13_bn, ResNet18, ResNet34, ResNet50, Wide-ResNet50_2, SqueezeNet1_0, Densenet121, MobileNet_v2

**Transformer Networks:** ViT-B/16, ViT-B/32, ViT-L/16, ViT-L/32, T2T-ViT_t-14, T2T-ViT_t-19, T2T-ViT-7, T2T-ViT-10, Swin-B, Swin-S, Swin-T, Twins-PCPVT-Small, Twins-SVT-Small, Visformer-Small
### Table 3.1: Model details: Number of layers included for each model

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of Layers</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlexNet</td>
<td>10</td>
</tr>
<tr>
<td>CORnet-S</td>
<td>12</td>
</tr>
<tr>
<td>Densenet121</td>
<td>30</td>
</tr>
<tr>
<td>MLP-Mixer-B16-224</td>
<td>24</td>
</tr>
<tr>
<td>Mobilenet_v2</td>
<td>14</td>
</tr>
<tr>
<td>ResNet18</td>
<td>10</td>
</tr>
<tr>
<td>ResNet34</td>
<td>18</td>
</tr>
<tr>
<td>ResNet50</td>
<td>18</td>
</tr>
<tr>
<td>Squeezenet1_0</td>
<td>13</td>
</tr>
<tr>
<td>Swin-B</td>
<td>24</td>
</tr>
<tr>
<td>Swin-S</td>
<td>24</td>
</tr>
<tr>
<td>Swin-T</td>
<td>12</td>
</tr>
<tr>
<td>T2T-ViT-10</td>
<td>13</td>
</tr>
<tr>
<td>T2T-ViT-7</td>
<td>10</td>
</tr>
<tr>
<td>T2T-ViT_t-14</td>
<td>17</td>
</tr>
<tr>
<td>T2T-ViT_t-19</td>
<td>22</td>
</tr>
<tr>
<td>Twins-PCPVT-Small</td>
<td>16</td>
</tr>
<tr>
<td>Twins-SVT-Small</td>
<td>18</td>
</tr>
<tr>
<td>VGG11</td>
<td>10</td>
</tr>
<tr>
<td>VGG13</td>
<td>12</td>
</tr>
<tr>
<td>VGG13-BN</td>
<td>12</td>
</tr>
<tr>
<td>VGG16</td>
<td>15</td>
</tr>
<tr>
<td>Visformer-Small</td>
<td>16</td>
</tr>
<tr>
<td>ViT-B-16</td>
<td>12</td>
</tr>
<tr>
<td>ViT-B-32</td>
<td>12</td>
</tr>
<tr>
<td>ViT-L-16</td>
<td>24</td>
</tr>
<tr>
<td>ViT-L-32</td>
<td>24</td>
</tr>
<tr>
<td>Wide-ResNet502</td>
<td>18</td>
</tr>
</tbody>
</table>

### 3.7.3 Coverage of neural recordings in target systems

For single-unit neural recordings, we have a limited number of recording sites. In our experimental settings, this is analogous to subsampling units in a target system. Collins et al. (2010) estimates that there are approximately 35 million neurons in V1 in the primate visual cortex, with a decreasing number of neurons in higher visual areas, such as MT, which has 1.6 million neurons. When different visual areas are combined, single-neuron recording sites generally amount to a few hundred. To
maximize coverage, assuming approximately 100 sites for each of V1, V2, V4, and IT combined in one visual area, we can approximate the number of neural recording sites as 400 neurons for a back-of-the-envelope calculation. If we apply a similar coverage ratio from the brain to neural network models, the number of subsampled units would range from 1 to 200 for a single layer, considering that the number of units in models roughly ranges from 800K (early layers of ResNet) to 4096 (later layer of AlexNet). This range is smaller than our experimental condition of subsampling 3000 neurons per target layer.

While increasing the number of recording sites for better system identification may not be easily feasible currently, we cannot rule out the possibility that recording techniques will be substantially improved in the future. Furthermore, applying CKA as a comparison metric may be more analogous to using RSA on neuroimaging data, such as fMRI or MEG, which captures signals from the whole brain. Therefore, we consider the condition of measuring all target units (Figure 3-7).
Figure 3-7: CKA scores when all target units are tested. Experimental setups are identical to Figures 3-4, 3-5, and 3-6 otherwise. See Section 3.5.2 for discussion. For the target MLP-Mixer-B/16 in (a), the source MLP-Mixer-B/16 has identical weights with the target; thus CKA is trivially 1.
Chapter 4

Conclusion

In this thesis, we have investigated methods to develop deep neural networks that are more biologically relevant. In the first part (Chapter 2), we examined how neuroscience can inform artificial intelligence. Architectural priors for deep neural networks are typically chosen to optimize computation and performance. However, to design models that better align with the human visual system, our approach involved first characterizing intrinsic invariances, which are closely related to architectural priors.

Specifically, to examine intrinsic invariance, rather than invariance arising from prior experience, we studied recognition performance in one-shot learning tasks. We found significant scale-invariance, but limited translation-invariance that depends on object size and eccentricity. The window of invariance is smaller than the window of visibility. Moreover, incomplete translation-invariance is further supported by the asymmetry between learning at the fovea versus periphery.

We compared experimental data with Convolutional Neural Networks (CNNs) and Eccentricity-dependent Neural Networks (ENNs). ENNs better capture the properties of invariance, suggesting scale-invariant representations and eccentricity-dependent receptive fields are important for modeling human vision.

In a broader context, the first part is along the lines of studies developing neuro-inspired models, directly based on established findings on the brain. On the other hand, with the recent advent of highly-performant models in the engineering field, there have been studies comparing these models with the brain to find that they
better match neural activity than previously suggested models (Yamins et al., 2014). However, whether that implies better agreement in the internal mechanisms is less known. Since developing a better model architectures is one of the central problems in computational neuroscience, we focused on examining whether we can understand the architectural components better through these deep neural networks in the second part (Chapter 3).

We introduced a framework comparing candidate neural networks with a target artificial neural network. Our results show that even under an idealized setting, identifying the correct model architecture consistent with the target is limited. Deeper layers, corresponding to higher visual areas, tend to give better system identification. Our results suggest using more naturalistic images and increasing neural recording coverage can improve identifiability. Our findings emphasize the need for cautious interpretations when assessing the correspondence between deep neural networks and the brain.

4.1 Future Directions

System identification beyond architectural insights

In the system identification approaches introduced in Chapter 3, our primary focus is on uncovering the underlying architectural components of a target neural system. Since these architectural components serve as concrete theoretical hypotheses, we consider them as the objective for a first step. However, many neuroscience studies use neural network models to analyze other aspects of visual processing. For instance, objective functions or tasks represent an additional component. Specifically, some visual areas are considered to specialize in processing certain input elements, such as faces or scenes (Kanwisher and Yovel, 2006; Epstein and Kanwisher, 1998). In light of this, models trained on different tasks are compared with neural data from specific brain areas (Dwivedi et al., 2021; Oota et al., 2022). The area is then interpreted as developing representation responsible for the task for which the best-matching model is trained on. The framework we present in Chapter 3 can be readily adapted
to understand objectives beyond model architectures. Therefore, for future work, we anticipate that other identification tasks, such as task identification, can be evaluated within this proposed framework.

**Leveraging advancements in artificial intelligence**

While the gap between human and artificial intelligence has been significant in the past, it is now narrowing in terms of capabilities. Recent advancements in large language and multimodal generative models have led to the creation of artificial agents that match or even surpass natural intelligence in some aspects. These models are capable of generating texts and images that are often indistinguishable from those produced by humans.

These models offer a novel avenue for studying intelligence. Artificial systems, being more amenable to experimental manipulation and activation analysis, serve as ideal platforms for testing experimental paradigms and comparison metrics. These can then potentially be applied to biological systems. The research detailed in Chapter 3 exemplifies this approach. With the continuous evolution of more advanced models, we can further study hypotheses related to high-level cognitive processes.

It remains an open question how closely human brains and artificial neural networks parallel each other. A pivotal question for the scientific understanding of intelligence is whether a universal principle underlies both biological and artificial intelligence. A key future research direction involves exploring both forms of intelligence within a unified theoretical framework.
Bibliography


David Marr and Tomaso Poggio. From understanding computation to understanding neural circuitry. 1976.


Nicole C Rust and James J DiCarlo. Selectivity and tolerance (“invariance”) both increase as visual information propagates from cortical area v4 to it. Journal of Neuroscience, 30(39):12978–12995, 2010.

Rylan Schaeffer, Mikail Khona, and Ila Fiete. No free lunch from deep learning in neuroscience: A case study through models of the entorhinal-hippocampal circuit. bioRxiv, 2022.


