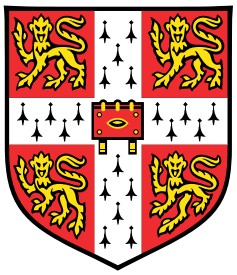
Computational models of the human visual cortex: on individual differences and ecologically valid input statistics



Johannes Mehrer

Supervisors: Tim Kietzmann, Nikolaus Kriegeskorte
MRC Cognition and Brain Sciences Unit
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Magdalene College December 2019

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Johannes Mehrer
December 2019

Acknowledgements

First, I would like to thank Tim C. Kietzmann and Nikolaus Kriegeskorte for guiding me through the journey of the last four years that culminated in this thesis. Your enthusiasm for the workings of the visual cortex are contagious and very inspiring. I am very grateful for your professional and personal support.

Further, I am very grateful for the opportunity to work with my lab colleagues Courtney J. Spoerer, Marieke Mur, Katherine Storrs, Patrick McClure, and Emer C. Jones. I will miss fun, witty and stimulating discussions, and great company. Thank you especially for the collaborations the work presented in chapter 2 (TCK, NK, CJS), chapter 3 (TCK, NK), and chapter 4 (TCK, NK, CJS, ECJ) is based on.

The MRC Cognition and Brain Sciences Unit (CBU) is a wonderful place for conducting a PhD in Neuroscience. My special thanks go to Fionnuala Murphy and Duncan Astle for their mentorship, and to the IT-team for providing an excellent computational infrastructure. For the financial support I would like to thank the Cambridge Trust, the Cambridge Philosophical Society, and the CBU.

The last four years have been a great opportunity to establish new friendships that I am certain will last far beyond this chapter of my life. I am very happy to have met you, Alex K., Gesa, Ibo, Julia, Kate, Matt, Paula, Pedro, Sneha, Shraddha, and Ture. I also want to thank Alex H., Cam, Dan, Helene, Jonas F., Jonas La., Jonas Le., Meret, Nils, Rick, and Tim, for being in my life.

Last, I would like to thank my family, Siegfried, Corinna, Lena, and Annabelle, for your love and support.

Abstract

Perception relies on cortical processes in response to sensory stimuli. Visual input entering the eyes ascends a cascade of processing steps from the retina to high-level regions of the cortex. Vision science investigates these transformations that give rise to high-level processing of visual objects, such as object recognition. In this thesis I investigate computational models of the human visual cortex with regard to their ability to predict cortical responses to visual objects. In particular, I describe two factors playing an important role in using deep neural networks (DNNs) to better understand cortical functioning: the initial weight state and ecologically more valid input statistics.

In Chapter 1 of this thesis I will introduce relevant literature pertaining to deep neural networks as a modeling framework for the visual cortex. Next, I will lay out the motivation for the research questions investigated in this thesis and described in detail in Chapters 2, 3, and 4.

Chapter 2 focuses on the impact of the initial weight state of a model on its ability to predict cortical representations. I describe work in which we demonstrate that two DNN instances identical in every aspect but their initial weights, yield very dissimilar representations. Relying on single network instances to predict cortical activation patterns in response to sensory stimuli poses a problem for computational neuroscience: depending on the initial set of weights the ability to mirror the cortical representations of these stimuli might vary. Thus, results based on single ("off-the-shelf") model instances - as commonly used in computational neuroscience - may not generalize. In contrast, using multiple DNN instances might alleviate this problem as they allow insights in the variability of a given model architecture to predict cortical representations. These individual differences between model instances suggest that to allow results to generalize more easily the model instances should be treated similar to human experimental participants.

In Chapter 3 I focus on ecologically more valid input statistics (in the form of training images) aiming to improve a model's ability to predict cortical representations. The most successful models of the human visual cortex to date are DNNs trained on object recognition tasks designed with machine learning goals in mind. However, the image sets used for training these DNNs are often not ecologically realistic. For example, training on the most-widely

used image set in computational neuroscience (ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012) requires the fine-grained distinction of 120 dog breeds, but does not contain visual object categories encountered frequently in everyday human life (e.g. woman, man, or child). This suggests that taking into account the human visual experience when training models of the human visual cortex on a categorization task might help to predict cortical representations. In this Chapter I describe the creation of a set of images aimed at mimicking the human visual diet: ecoset. Ecoset contains more than 1.5 million images from 565 basic level categories and is the largest image set specifically designed for computational neuroscience to date. Ecoset is freely available to allow the community to test their own hypotheses of models trained with input statistics matched to the human visual environment.

In Chapter 4 we build on the results from the previous two Chapters. Using multiple DNN instances I investigate whether a brain-inspired model architecture (vNet) trained on ecologically more valid input statistics (ecoset) might improve its ability to predict cortical representations. I first demonstrate that ecoset might improve an architecture's ability to mirror cortical representations. Furthermore, ecoset-trained vNet also outperforms state-of-the-art computer vision and computational neuroscience models in terms of mirroring cortical representations in the human brain. Thus, incorporating biological and ecological aspects, such as brain-inspired architectural features and ecologically more valid input statistics, into computational models may yield better predictions of response patterns in the human visual cortex.

Treating DNN instances similar to human experimental participants and considering ecological and biological factors for building these DNNs may be an important step towards better models of the human visual cortex. Such models might allow a better understanding of the cortical processes underlying high-level vision in the human brain.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

Vision plays an integral role in our lives. It enables us to perceive our proximal surroundings and to recognize objects that might be several kilometers away. On the basis of visual information our ancestors were able to identify predators or enemies allowing for preparation of conflict or escape. Further, we heavily rely on vision for identification of and communication with our fellow (human) beings on a daily basis. As such, vision is at the core of our everyday social interactions and is thus indispensable for a definition of what we are as a species.

But how exactly is vision implemented in the human brain? When light reflects off objects it enters the eye and the visual information travels along the visual processing hierarchy from the retina, through sub-cortical regions in the thalamus to the visual cortex. At each step the visual information undergoes transformations allowing the biological organism to extract information for high-level decision making and, finally, actions. The amount of cortex thought to underlie visual perception is - compared to the other senses - relatively large and may indicate the importance of vision to our system: the visual cortex (including visual-association areas) comprises more than 50% of the neo-cortex in humans and more than 60% in macaques (Felleman & Van Essen, 1991; Maunsell, 1987). Note, however, that the areas of the cortex typically dedicated to visual processing can also be involved in the processing of non-visual information. For example, in cases of disabled input to the visual system in the congenitally blind, “visual" cortex might respond to auditory and tactile stimuli (Lane et al., 2015). This demonstrates that the visual system specifically and the human brain in general are plastic and ever-changing systems whose exact workings we are only about to decipher.

To better understand how vision is implemented in the brain, vision science asks how sensory input to the eyes is transformed to give rise to high-level visual abilities and phenomena, such as object recognition, different forms of visual attention, or optical illusions (Baldauf & Desimone, 2014; Carbon, 2014; Posner, 1980; K. Tanaka, 1997). To investigate

these abilities and phenomena systematically, experiments are conducted at different stages of the visual processing hierarchy, as well as at different levels of explanation. Evidence may be gathered from the retina, sub-cortical or cortical regions. Different techniques allow to directly record (patch-clamp, single-cell recordings, multi-site arrray recordings; Gross et al., 1977; Hamill et al., 1981; Heuschkel et al., 2002; Neher and Sakmann, 1976; Sakmann and Neher, 1984, 1995; Thomas et al., 1972) or indirectly infer (ECoG, EEG, MEG, fMRI, PET; Engel et al., 1994; Haxby et al., 1994; Herrmann, 2001; J. Liu et al., 2002; Winawer et al., 2013) neural activation patterns in response to sensory stimuli. These recording techniques vary as to whether they are invasive and concerning their temporal and spatial resolution (for an overview see Fig 1.1 A).

To explain data from brain recordings, vision science has a long standing history of using models to better understand the workings of the human visual system. Every model depends on choices regarding the level of abstraction at which neural functioning should be reflected. For example, the groundbreaking model by Hodgkin and Huxley (1952) uses differential equations to describe the cellular mechanisms giving rise to an action potential in a single cell, but is not able to explain data on a systems or behavioral level (Hodgkin & Huxley, 1952). In contrast, state-of-the-art deep neural network (DNN) models may abstract away from basic neural features such as spikes, but are able to explain behavioral choices or higher-level cortical representations in humans and monkeys (Kietzmann, McClure, et al., 2019). For an overview of the constraints of some of the most important computational models in vision science, see Fig 1.1 B. As demonstrated in this sketch, modern DNNs show a relatively low level of biological detail. The constraints of the recording techniques on the one hand and of the models in visual neuroscience on the other hand (Fig 1.1) determine from which population data can be collected and in which way a research question can be investigated.

What constitutes a good model, is an open question and depends on the given research question and what follows are important criteria often used to determine the quality of a model. For example, parsimony is central to both scientific explanations in general and to (computational) models in neuroscience more specifically. A parsimonious model is as simple as possible and ideally provides an intuitive understanding of the complex system of interest and thus allows for effective communication of its explanatory merit. Another important criterion used to evaluate a given model is its ability to predict data not used for training. Trained on one part of the data the goodness of fit to left-out data demonstrates how well the model allows to generalize. For some models such a goodness-of-fit criterion can be easily combined with the aforementioned parsimony to create a single index reflecting the

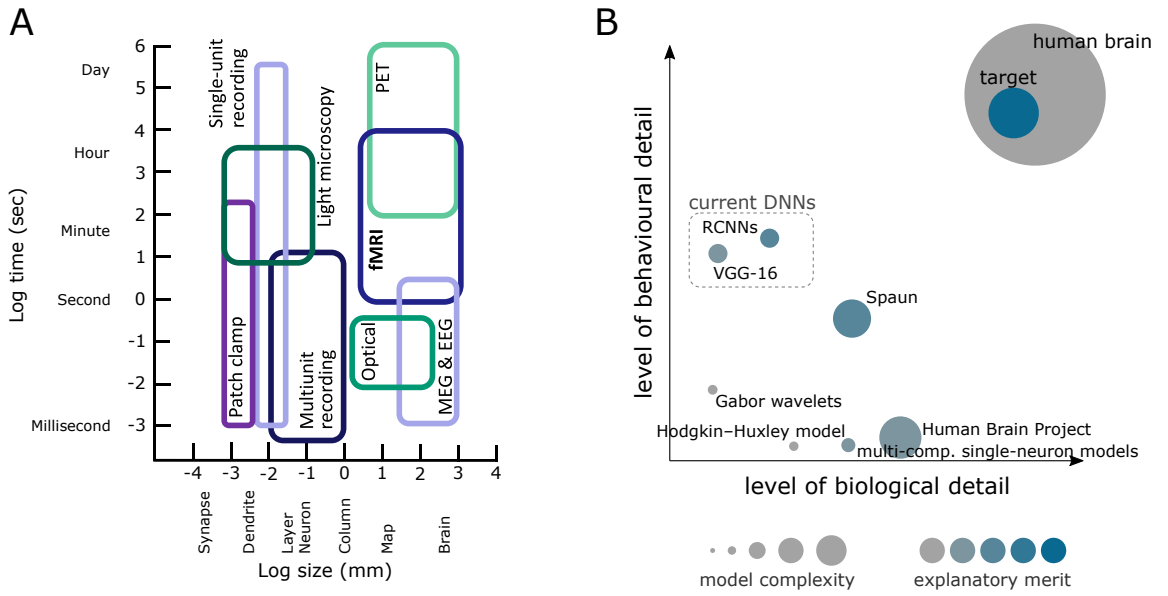


Fig. 1.1 **Constraints of recording techniques and of models used in visual neuroscience. A)** Approximate limits of spatial and temporal sensitivity of measurement techniques capturing cortical activations. fMRI: functional magnetic resonance imaging; PET: positron emission tomography; MEG: magneto-encephalography; EEG: electro-encephalography; Optical: e.g. functional near-infrared spectroscopy. Figure adapted from (Medaglia, 2017), originally distributed under CC-BY-4.0 license. **B)** Overview of constraints of computational models in vision science. Each model is a choice regarding the level of biological (x-axis) and behavioral (y-axis) detail. Two additional dimensions that concern these models are the model complexity, often estimated by the number of free parameters, and the explanatory merit. The latter relates to the insights into the workings of the brain rather than to its accuracy of biological detail. Figure reprinted with permission from (Kietzmann, McClure, et al., 2019).

complex relationship between these two factors in order to allow model selection (e.g. AIC, BIC; Kuha, 2004).

However, commonly used measures of parsimony, such as the number of free parameters, are challenged when dealing with the model class of deep neural networks (DNNs, see figure 1.1). First, the hierarchical structure of DNNs implies that parameters are not independent as units located in a given layer of an (exclusively feed-forward) network architecture depend on the activity in the previous layer. The exact degree of cross-unit dependence depends on many architectural features and its estimation is non-trivial. Thus the number of free parameters is at best a coarse estimate of a given DNN's complexity or its parsimony (Hodas & Stinis, 2018). In addition, whereas a classical notion of model complexity typically favor models with fewer trainable parameters for best generalization performance, over-parameterized

neural networks can reach higher testing performancs than more parsimonious ones (Novak et al., 2018; Sun & Nielsen, 2020).

Beyond model parsimony and goodness-of-fit, one factor important for models in computational neuroscience is its biological validity. Loosely inspired by the structures initially found in cat and monkey visual cortex (Hubel & Wiesel, 1959, 1962), deep neural networks bear some similarities with biological vision systems. However, the first DNNs excelling e.g. at the complex object recognition task ILSVRC 2012 (Deng et al., 2009; Russakovsky et al., 2015) lack important features of the human visual cortex, such as spikes or top-down and recurrent connections (Chatfield et al., 2014; He et al., 2014; A. Krizhevsky et al., 2012). It remains to be insvestigated which aspects of biological vision systems help to e.g. increase its ability to predict neural data.

Beyond explaining data recorded in humans and other animals, a desirable feature of a model is it's ability to generate novel predictions. For example, it has been shown that a DNN-based image synthesis method can be used to produce stimuli that can trigger spiking activity in specific neural sites at higher levels than occurring naturally (Bashivan et al., 2019). The ability to drive activity levels beyond natural levels implies that this model captures an important aspect of the system under investigation and thus reveals previously unknown functionality.

1.1 Initial weights and input statistics determine network internal representations

The initial wiring of the brain plays a pivotal role in how its embedding organism develops over its life span (Hagmann et al., 2010; Seckfort et al., 2008). In analogy to the brain, the structure and the initial set of weights of a given DNN, reflecting the strength of the connections across layers, determine the internal representations formed via task training. Using single instances of DNNs as models of the human visual cortex poses a problem for computational neuroscience as the initial set of weights of a given network instance might determine its fit to a set of neural data after completion of training. The second Chapter of this thesis is thus dedicated to investigating the strength of the impact of the initial set of weights on network internal representations.

In addition to the initial wiring of the brain, each organism is highly influenced by its experiences. Just as the input to the human visual system strongly influences its current state (Charest et al., 2014; Gauthier et al., 2000; Palmeri et al., 2004), DNN internal representations are shaped by the stimuli and the tasks they are trained on. In the third Chapter of this thesis

I will describe the creation of ecoset, an image set designed to approximate the human visual experience offering ecologically more valid DNN training on a complex object recognition task.

In the fourth Chapter I build on the results from Chapters 2 and 3. Here, multiple DNN instances per architecture (identical DNNs, only differing in the weight configuration at the beginning of training) are used to investigate whether a brain-inspired DNN architecture may better explain representations in the human visual cortex than i) the same architecture but trained on a computer vision image set instead, and ii) state-of-the-art computer vision models. Finally, in Chapter 5 I summarize the results of our investigations and discuss the implications following from the presented evidence.

Progress in (vision) science crucially depends on a converging operations approach. Bringing together findings from multiple levels of analysis and multiple recording techniques using various types of DNNs can provide a fuller understanding of the underlying perceptual and cognitive system. In general, in this thesis I investigate whether taking inspiration from ecology and biology helps to build better models of the human visual cortex. In other words, I investigate whether increasing the *ecological* and *biological* detail of modern DNNs may help to increase the level of *behavioral* detail.

1.2 A short history of neural networks used as models of the human visual cortex

The goal of vision science is to understand how input to the eyes is transformed along the visual stream to allow high-level visual perception to emerge. How, for example, is it possible that the activation patterns in the cortex allow humans to identify another person's face? To better understand the cortical processes underlying visual perception we require simplified versions of the system under investigation that abstract away from some building blocks, while keeping other functionalities of interest largely intact. To this end we need brain-computational models mimicking the cortical processes underlying a given task at a given level of abstraction (Kriegeskorte & Douglas, 2018). For vision science such models need to be able to process information from stable or moving visual stimuli and perform a task that is thought to rely on the processes in the visual stream.

"What I cannot create, I do not understand." is the powerful statement that Richard Feynman left us with shortly before his death. Applied to the field of computational visual neuroscience, this means that to understand the transformation from a pixel-like representation to a semantic one, eventually allowing complex decision making and motor responses,

we need to build a machine capable of performing each step along this way at a given level of abstraction.

It has been a long standing goal of artificial intelligence (AI) to build a machine capable of complex object recognition. AI and visual neuroscience have a long and intertwined history. The first computing machines were heavily influenced by structural and functional properties of the brain (Hassabis et al., 2017; Kriegeskorte, 2015). One of the most notable cases in which neuroscience findings had a strong and far-reaching influence on a subfield of AI, namely computer vision, is Hubel and Wiesel's discovery of single cell response profiles in the cat visual cortex in 1959 (Hubel & Wiesel, 1959, 1962). The authors differentiated between simple and complex cells responding to stimuli with varying degrees of complexity. Whereas simple cells can be excited using a bar stimulus presented at a specific orientation and location in the receptive field, complex cells offer some position invariance, and might be direction sensitive. This arrangement of cells in a hierarchy of increasingly complex response profiles in a biological vision system inspired an artificial neural network model that marked an important step in computer vision: the neocognitron (Fukushima, 1980). Trained in an unsupervised fashion, it exhibited some location invariance and thus allowed for a better recognition of single hand-written digits.

In addition to the position in the visual field, other factors, such as lighting, or viewpoint change the appearance of a given object, while its percept remains stable. How this perceptual constancy (contrasting with the variability of the input to the visual system) is implemented in the brain is one of the most fascinating questions in vision science (Biederman, 1987; Duhamel et al., 1997; Marr & Nishihara, 1978). For computer vision models to demonstrate various types of invariance and thus allowing object recognition on or even above the human performance level, more complex, deeper models than the neocognitron appeared promising. However, training neural networks with many hidden layers was a difficult task computer vision struggled with for a long time.

Combining architectural features from the neocognitron with computational advances in the form of "learning internal representation by error propagation" ("backpropagation", or short "backprop"; Rumelhart et al., 1986), LeCun (1989) was able to train a 4-layer network performing handwritten zip code recognition (LeCun et al., 1989). Despite this great achievement using a neural network with multiple hidden layers trained on a complex object recognition task using backprop, the machine learning community soon shifted its focus from neural networks to other techniques, such as support vector machines (SVMs). This re-orientation was motivated by better task performances on various classification tasks and because SVMs were easier to train than DNNs. However, a small number of research groups in machine learning and in visual neuroscience continued their efforts in investigating neural

networks as promising tools for complex classification problems and as models of the visual cortex (Kriegeskorte, 2015). From 1989 it took about two decades until DNNs trained with backpropagation were put back center stage of interest in the machine learning community.

Inspired by the neocognitron, Riesenhuber and Poggio (1999) set a milestone for visual computational neuroscience creating a network combining computer vision capabilities (position and scale invariant object recognition) with requirements from neuroscience (to mirror single cell activations in response to specific objects) in an architecture known as "HMAX" (Riesenhuber & Poggio, 1999; Serre, 2015). After adding important biological and ecological aspects (receptive field sizes matching those in macaques, and ecologically relevant input statistics in the form of a training set composed of natural images; Serre, Wolf, et al., 2007), HMAX was able to achieve very good performance levels on a group of object recognition tasks (Serre, Oliva, et al., 2007). Note that the two biological aspects that differentiate HMAX in its latest form from the original one - receptive field sizes and input statistics - play a central role in this thesis in the context of DNNs and will be discussed in-depth in Chapter 4. Notwithstanding its merits in both computer vision and visual neuroscience, HMAX was not able to solve more complex object recognition problems, e.g. the classification of millions of images from hundreds of categories. But only a few years later computer vision models capable of such feats of intelligence emerged and radically changed both computer vision and computational neuroscience.

Several parallel developments in computer vision culminated in 2012 in a DNN architecture which outperformed all other models on a complex object recognition task by a large margin: AlexNet (A. Krizhevsky et al., 2012; LeCun et al., 2015). In contrast to HMAX, backprop was used to train network internal features in the 7 layer architecture of AlexNet. In general, three main factors contributed to this revolution in computer vision, and with a short delay also in computational visual neuroscience.

First, large datasets offered a strong enough constraint to train millions of parameters in deep computer vision models with the goal to perform complex visual object recognition tasks. More specifically, the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), offering its challenge in 2010 for the first time, provided not only millions of images to train on, but also a stimulating competitive atmosphere motivating labs around the globe to submit their models to this benchmarking challenge (Deng et al., 2009; Russakovsky et al., 2015). I will return to the importance of image sets for training DNNs later in this introduction and investigate this topic at length in Chapters 3 and 4 of this thesis.

Second, the combination of several computational features enabled AlexNet to make a leap in performance of this complex object recognition task: using rectified-linear-units (ReLU) as activation function, regularizing the network through dropout, and weight-sharing

through convolutions, all of which are explained next. Shown to increase categorization performance in neural networks trained in an unsupervised fashion, AlexNet replaced commonly used sigmoidal activation functions with ReLU (Schmidhuber, 2015). Next, to avoid reliance of units on all input weights, generalization performance was increased through “dropout". For the dropout implementation used in AlexNet each unit was silenced - or “dropped out" - with a probability of 0.5 at each network update. In this way the network was forced to develop features able to deal with incomplete input (Srivastava et al., 2014). Last, while touching upon weight-sharing as already implemented in other forms in the neocognitron and the network presented by LeCun to recognize handwritten digits (LeCun et al., 1989), weight-sharing could unfold its full potential when applied to the DNN framework in the form of convolutional layers. In a feedforward convolutional DNN, such as AlexNet, the units of a given convolutional map of a layer share the set of connections to its preceding layer with all other units of the same map, thereby allowing detection of an object with a high degree of translation invariance. Conveniently, as the receptive fields of a given convolutional map share the same weights, convolutions also reduce the number of trainable parameters and thus the computational demands to train a model when compared to fully-connected layers.

The third main factor, contributing to the 2012 revolution in computer vision, was dedicated computer hardware for deep learning. Even with a reduced number of trainable parameters through weight sharing in convolutional layers the use of specialized processing units normally used for graphics applications (graphical processing units, GPUs) was necessary to allow a boost in task performance opening new horizons for both computer vision and computational neuroscience. It is an interesting and ironic twist to the story of computational neuroscience that the very same tools used to build the inverse of visual perception (graphics) allowed a quantum leap in the ability to perform complex object recognition tasks and to constitute the best to date models of the human visual cortex.

In sum, deep feedforward convolutional DNNs trained in an supervised manner using backprop - with AlexNet as its most prominent example - have revolutionized object recognition and other domains of machine learning, such as biological image segmentation, and face detection (LeCun et al., 2015). In 2012 AlexNet stunned the computer vision community by almost halving the error rate of the object recognition challenge of ILSVRC in comparison to the previous year. As human-level performance on a complex object recognition task had come within reach of the performance of an artificial neural network, the vision science community soon started testing whether a system yielding such high task performance levels might also be able to predict cortical responses recorded in humans and other primates. In 2013 and 2014 multiple labs showed independently that AlexNet and other DNN architec-

tures outperformed many other (much shallower) computer vision models with respect to their ability to mirror cortical representations in response to visual objects in the human and macaque ventral stream (Cadieu et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2013; Yamins et al., 2014). Note, however, that more recent work using even deeper architectures was unable to demonstrate an improved ability to predict cortical representations (Abbasi-Asl et al., 2018; Kalfas et al., 2017; Serre, 2019; Storrs & Kriegeskorte, 2019; Storrs et al., 2017).

One possible explanation for the relationship between successful object recognition in DNNs and their ability to predict cortical representations is that task learning yields (task-specific) features similar to those found in vision systems of humans and other primates (Kietzmann et al., 2009, 2008). Such features - similar to those found in biological vision systems, and thus able to span a similar representational space - can easily be visualized for lower stages of the ventral stream and shallow layers of DNNs. Recent visualization techniques for neural networks allow insights in the response properties of units in shallow and deep layers of DNNs (Fig 1.2).

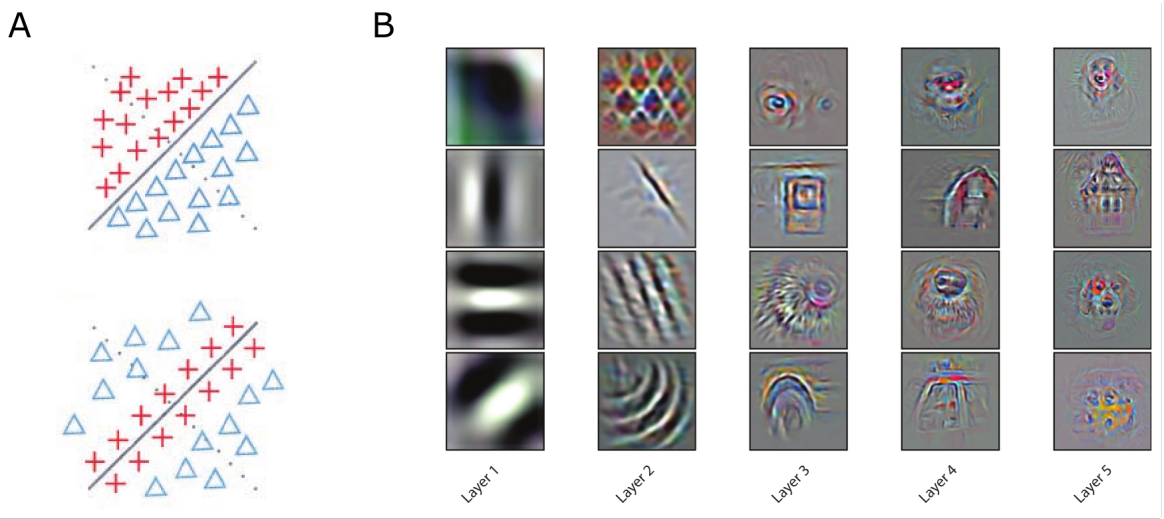


Fig. 1.2 **Receptive fields in biological vision systems are similar to those found in deep neural networks. A)** Cartoon of simple receptive field in the cat visual cortex. Figure adapted from Hubel and Wiesel, 1962 by Martinez and Alonso, 2003, reprinted here with permission from John Wiley and Sons (license nr. 4721300155889). **B)** Characterization of 4 stimuli (rows) strongly activating units from 5 layers (columns) in a DNN architecture from Chatfield et al., 2014. The stimuli in layer 1 resemble the receptive fields in the cat visual cortex shown in A. The stimuli were produced by Güçlü and van Gerven, 2015 using a visualization technique from Zeiler and Fergus, 2014. Figure adapted with permission from Güçlü and van Gerven, 2015.

Other studies compared the similarity of representations at specific depths of DNN models to those found at specific stages of the ventral stream, most of them using encoding or decoding techniques (Agrawal et al., 2014; Cichy et al., 2016; Devereux et al., 2018; Eickenberg et al., 2017; Güçlü & van Gerven, 2015, 2017; Hong et al., 2016; Horikawa & Kamitani, 2017a, 2017b). Not only do DNNs estimate cortical representations of visual information successfully, they are also the current best models of behavioral response patterns, and of similarity judgments in humans (Cichy & Kaiser, 2019; Kietzmann, McClure, et al., 2019; Kriegeskorte, 2015; Kriegeskorte & Douglas, 2018, 2019; Serre, 2019; Storrs & Kriegeskorte, 2019; Yamins & DiCarlo, 2016b). This evidence validates the status of DNNs as the best current models of (visual) cortical processes and the resulting phenomena and actions.

To conclude, deep feedforward convolutional networks have been very successfully used as models for the human visual stream and are thought to reflect well the cortical processes underlying core object recognition (DiCarlo et al., 2012). AlexNet marked the beginning of a new era of computational visual neuroscience in which end-to-end trainable machines outperform humans in challenging visual tasks, and are - at the same time - able to explain cortical representations at all stages of the visual cortex.

1.3 Deep neural network terminology

We have seen that deep feedforward convolutional networks have been very successfully used as models for the human visual stream. I will now explain some of the terminology to describe the architecture of these networks and how they can be trained to perform complex tasks at or even above human-level performance. First, the term "neural network" was originally used to describe a network of biological neurons, but is now used as the short form of *artificial* neural network Kriegeskorte, 2015. A deep neural network is conventionally referred to as "deep" when there is more than 1 hidden layer between the input and the output layer. Layers of a convolutional network are composed of maps, which all have the same number units. Units from the same map share the set of weights connecting them to units of maps in earlier layers. But units from different maps of the same layer might differ in the set of weights determinging their selectivity. Overall, the way in which network units (and thus maps and layers) are connected determines the network architecture.

To train a network, it needs to be identified how a change in a weight connecting two units affects the objective function to be optimized. To find solutions to this fundamental problem of credit assignment in deep neural networks the backpropagation is commonly used Rumelhart et al., 1986. Backpropagation is effective in that it allows to determine

the effect of a change in weight independent of its location in the network. It is efficient in that the effects from weight changes on the given objective function can be determined for multiple units and multiple input data points at the same time. Using the entire training set (standard gradient descent) or batches of input data (stochastic gradient descent) the weights are iteratively updated to find the (ideally global) minimum in the high-dimensional error-surface of the given objective function.

1.4 Individual differences between deep neural network instances

What unites the DNNs used as models for the visual stream is the reliance on a single DNN instance (Agrawal et al., 2014; Cichy et al., 2016; Devereux et al., 2018; Eickenberg et al., 2017; Güçlü & van Gerven, 2015, 2017; Hong et al., 2016; Horikawa & Kamitani, 2017a, 2017b; Kietzmann, McClure, et al., 2019; Storrs & Kriegeskorte, 2019; Yamins & DiCarlo, 2016b). With the exception of some of the recurrent models presented last, all aforementioned DNNs were created with the goal of excelling at a computer vision task, but were not constrained to match neural data or to directly mimic properties of the visual cortex. In other words, most of the models used by computational visual neuroscientists are single DNN instances pre-trained by and directly imported from the computer vision community. This might constitute a problem for computational neuroscience as the representations of a given architecture might vary depending on the initial set of weights of a given instance, whereby a single DNN instance does not suffice to capture this variability.

In machine learning many resources related to DNNs are dedicated to improving the accuracy on a given task. As task performance is indistinguishable across DNN instances (Li et al., 2016), the machine learning community is not usually concerned with training more than a single DNN instance per architecture (Chatfield et al., 2014; He et al., 2014; A. Krizhevsky et al., 2012). However, although the representations across DNN instances converge to the same similarity matrix in linear networks (Saxe et al., 2019, 2013), they may vary extensively in non-linear networks (Kornblith et al., 2019; Li et al., 2016; Lu et al., 2018; Morcos et al., 2018). Despite the machine learning focus of these studies, the results might have important implications for using DNNs as models for the visual cortex: the fit of a DNN instance to a given neural dataset might depend on the initial set of weights. To conclude, it remains to be investigated how strong such an effect might be and how multiple network instances of the same architecture may be able to describe a fuller picture of the ability of the given architecture to explain cortical representations.

The idea of multiple instances of a system performing a given task in an almost identical way, while relying on very different representations is closely related with the concept of “degeneracy". Degeneracy has its roots in molecular biology where it was first defined as “the ability of elements that are structurally different to perform the same function or yield the same output" (p. 13763, Edelman and Gally, 2001). In cognitive neuroscience, the term “degenerate" is used to describe the same or a very similar behavioral outcome while - at the same time - relying on different structures (e.g. due to a stroke) or on different task strategies (e.g. in response to top-down signals capturing task instructions; Noppeney et al., 2004; Noppeney et al., 2006; Price and Friston, 2002). It is in this sense that degeneracy applies to instances of the same DNN architecture: although performing a given task at an indistinguishable performance level, the network internal processes might rely on very different network internal representations. Individual differences play an important role when dealing with human subjects (Price & Friston, 2002). Similarly, individual differences might be an important feature when dealing with DNNs identical in every way, but differing in their initial set of weights: a possibly degenerate nature of DNNs would suggest that for results based on a specific DNN architecture to generalize, they might need to be based on multiple instances to allow for an estimation of the representational variability within this architecture. This raises the question how we can investigate the similarity of representations across differently initialized DNNs from the same architecture. Our attempt to answer this question is the content of Chapter 2 dealing with individual differences between DNN instances.

Although the question of the best weight initialization has been explored extensively in machine learning (Agrawal et al., 2014; Doersch et al., 2015; Glorot & Bengio, 2010; He et al., 2014; Mishkin & Matas, 2015; Saxe et al., 2013; Sutskever et al., 2013; D. Xie et al., 2017), these investigations have been conducted with task performance goals in mind, instead of relating to the usage of DNNs as models of the visual cortex. Similarly, the aforementioned studies investigating representational similarities across DNN instances (Kornblith et al., 2019; Li et al., 2016; Lu et al., 2018; Morcos et al., 2018) are not directed towards DNNs as models of biological vision systems. Accordingly, the techniques used in these studies to compare representational similarities do not include representational similarity analysis (RSA), one of the most widely-used tools in systems neuroscience to compare representations across species, individuals or between computational models and the brain (Kriegeskorte & Kievit, 2013; Kriegeskorte et al., 2008).

To address these issues, we use representational similarity analysis (RSA; Kriegeskorte and Kievit, 2013; Kriegeskorte et al., 2008; Nili et al., 2014) to compare the representational similarities across DNN instances. RSA is a multivariate analysis framework from systems neuroscience whose building block is the representational dissimilarity matrix (RDMs). Each

cell of an RDM describes the distance in activation space (e.g. cortical as estimated using fMRI or in-silico neurophysiology using deep neural networks), whereby large distances describe dissimilar representations and small distances describe similar representations. The complete RDM reflects the representational geometry of a given set of stimuli expressed in pairwise distances. When RSA is used to compare representations across network instances, the resulting representational consistency reflects the representational similarity. In Chapter 2 two different network architectures (VGG-753 - similar to VGG-S, (Chatfield et al., 2014) and All-CNN-C (Springenberg et al., 2015)) are used to demonstrate how representational consistency behaves as a function of network depth and how it depends on the distance measure used to compute representational distances within each DNN instance. In addition, to shed light on the differences, if any, we explore how the separability of classes in high-dimensional network activation space relates to the distribution of class instances around the class centroid. Further, to explore how representational similarity across DNNs can be recovered, we investigate the effect of a regularization technique applied to the network during training and testing.

1.5 Manipulating network internal representations

So far, I have described our plans to investigate individual differences between DNN instances and have outlined how our results might yield important implications for the usage of DNNs as models of the visual cortex of humans and other primates. Using groups of DNNs instead of single DNN instances to predict cortical representations might allow for a better generality of insights gained. As a next step, in Chapters 3 and 4 I turn to the impact of the input statistics (and the architecture) of a DNN on its internal representations and thus its ability to mirror representations in the cortex. Before I do, let me shortly describe which factors can - in general - be directly manipulated to alter network internal representations.

DNN internal representations of visual objects depend on the structure of the given architecture and the input information, e.g. in the form of training images. In addition, what determines how network internal representations are shaped depends on the the task itself and on how learning (realized through the change in weights) is implemented. In other words, there are four main factors influencing the internal representations of a neural network: functional objective, learning algorithm, network structure, and input statistics (Kietzmann, McClure, et al., 2019). I now explain how each of these factors impacts network internal representations.

Most of the DNNs used in computational neuroscience to predict cortical data are feedforward convolutional DNNs trained to minimize a classification error in a visual object

recognition task (Kietzmann, McClure, et al., 2019; Kriegeskorte, 2015; Richards et al., 2019; Serre, 2019). The large-scale supervised training of these models has been enabled by the emergence of large sets of labeled images freely available to the community (Deng et al., 2009; Russakovsky et al., 2015). This approach has shown to be very successful in computational neuroscience, but lacks ecological and biological realism insofar as other objectives and learning algorithms are thought to play a vital role in the brain, and go beyond the classification of visual objects. For example, it has been suggested that not one, but multiple objectives are being optimized by a given brain region at a single point in time (Marblestone et al., 2016). This notion relies on evidence suggesting that cortical representations may rely on overlapping maps (collectively forming distinct functional modules; Op de Beeck et al., 2008), implying that neural responses may encode different types of information at the same time (DiCarlo & Cox, 2007; Kietzmann, McClure, et al., 2019).

Instead of end-to-end and global training as performed in feedforward DNNs trained using backprop, much of the changes to neural connections in biological vision systems are thought to rely on unsupervised, local learning rules, such as Hebbian learning (Hebb, 1949; Song et al., 2000). This is in line with data suggesting that learning in the ventral stream appears independent of reward signals (N. Li & DiCarlo, 2012; Logothetis et al., 1995; Serre, 2019). However, where tested, supervised models outperform unsupervised counterparts with regard to predicting cortical representations, which renders them the preferred computational models of the human visual cortex (Khaligh-Razavi & Kriegeskorte, 2014).

Functional objectives and learning algorithms are only two of four main elements through which network internal representations can be directly manipulated. Let us thus now turn to the third element: network structure. A field within machine learning investigating architectural properties of DNNs is automated neural architecture search (NAS). This approach allows task performance optimization by systematic searches through hyper-parameter space (Elsken et al., 2019; Pham et al., 2018; S. Xie et al., 2019). By including architectural parameters in the search algorithm based on gradient descent, one can find high-performance convolutional architectures for large-scale image recognition tasks and recurrent architectures for language modeling (H. Liu et al., 2019). Until recently, this approach has only been used to optimize architectures for task performance. However, first evidence from computational neuroscience suggests that additionally including the fit to multiple sets of cortical data in the neural architecture search algorithm may help to find better models of the visual stream in both humans and other primates (Kietzmann, Spoerer, et al., 2019; Kubilius et al., 2018; Nayebi et al., 2018). To conclude, NAS can be used to algorithmically optimize a DNN's architecture with regard to its ability to predict cortical representations. In contrast,

knowledge about the structure of the human visual cortex can be directly implemented in structural elements of DNNs. We here take the latter approach by matching the receptive field sizes of a DNN architecture to the progressive increase in foveal receptive field sizes along multiple areas of the human ventral stream (for a more detailed account, see Chapter 4).

The fourth and last main element directly influencing the internal representations of a feedforward DNN are the input statistics, here in the form of images used for object recognition training. One of the two main changes from the first (1999) to the latest (2007) version of HMAX is the adjustment of the training set to natural scene statistics, to obtain a universal feature set enabling human-level performances on a range of complex object recognition tasks (Serre, Oliva, et al., 2007; Serre, Wolf, et al., 2007). Similarly, the input statistics are the main subject of Chapter 3 of this thesis, where I describe the creation of an image set designed with the goal to approximate the human visual experience.

1.6 Ecologically more valid input statistics

One question that applies to all four factors shaping network internal representations (functional objective, learning algorithm, network structure, and input statistics) is the one of ecological and biological plausibility. Although there are clear examples of highly effective engineering solutions showing little biological realism (planes, trains, cars), the success of deep neural networks in machine learning suggests that biological inspiration may help to create powerful engineering tools (Kriegeskorte, 2015). Furthermore, the unparalleled ability of DNNs to predict cortical representations, confirms the intuition that inspiration from the brain may help to build better models of the primate visual cortex. However, as discussed above (Fig 1.1), each model is the result of a choice regarding the level of *ecological* and *biological* on the one hand, and *behavioral* detail on the other one: some models may be able to explain neural activations on a cellular level, but fail to explain behavioral patterns of the host organism (Hodgkin & Huxley, 1952; Markram et al., 2011); other models, namely DNNs, are able to predict behavior, similarity judgments and high-level cortical representations (Kietzmann, McClure, et al., 2019), but operate via rate-coding and thus abstract away arguably important characteristics of neural processing, such as spiking. This raises the question whether DNNs with their relatively little ecological and biological detail may profit from additional ecological (visual input) and biological (architectural features) inspiration from the system to be modeled.

So far, most studies investigating DNNs as models for predicting cortical representations have been trained on the same image set: ILSVRC 2012 (Agrawal et al., 2014; Bankson

et al., 2018; Cadieu et al., 2014; Cichy et al., 2016; Devereux et al., 2018; Eickenberg et al., 2017; Güçlü & van Gerven, 2015, 2017; Hernández-García et al., 2019; Hong et al., 2016; Horikawa & Kamitani, 2017a, 2017b; Khaligh-Razavi & Kriegeskorte, 2014). This might pose problems for computational visual neuroscience. ILSVRC 2012 was designed for the machine learning community and thus exhibits only little ecological and biological realism. For example, in ILSVRC 2012, 120 of a total of 1,000 categories are different breeds of dogs, whereas it lacks categories important to humans, such as woman, man, or child. To more closely mimic the human visual experience we need an image set reflecting the objects we encounter in our daily lives. However, no publicly available dataset contains millions of images from multiple hundred categories necessary for training a modern DNN on an object recognition task. Hence, we created a novel set of images to help build better models of the human visual cortex: ecoset.

Deciding on a selection of categories for large image sets is complicated, which explains why for some image sets categories are selected in a subjective and non-principled way, e.g. by selecting categories "the authors deemed important" (p. 3, Kuznetsova et al., 2018). What renders this endeavor even more complex is that there is no general agreement on what constitutes a basic level category (Markman & Wisniewski, 1997; J. Tanaka & Taylor, 1991). In order to be comprehensive while not being redundant, the categories of an image set reflecting the human visual experience need to be mutually exclusive, and collectively exhaustive. In order to find such a set of categories, we asked what are the most important categories in the visual diet of humans?

Our attempt to answer this question and how we created ecoset is the content of Chapter 3. I describe how a list of objects covering the most important human visual input was used to download adequate images from multiple online sources. Further, I explain, which inclusion and exclusion criteria we used to obtain a set of 1.5 million images from 565 basic level categories - the largest to date image set designed for computational neuroscience. As ecoset will soon be freely available to the community, it allows researchers to test their own hypotheses with regard to the ecological plausibility of DNN input statistics.

1.7 An ecologically more valid visual diet for deep learning yields better models of human high-level visual cortex

After introducing the two main themes of this thesis, individual differences between DNNs (Chapter 2), and ecologically more valid input statistics (Chapter 3), in Chapter 4 I build

on the two previous Chapters and investigate whether ecological and biological inspiration helps build better models of the human visual cortex.

First, to create the DNN model used throughout Chapter 4, the receptive field sizes of a DNN architecture are matched to the progressive increase in foveal receptive field sizes along multiple areas of the human ventral stream: vNet. We then train multiple network instances of vNet on an ecologically more valid image set (ecoset) and compare their ability to predict cortical presentations with the same networks trained on a computer vision challenge instead. Brief, using multiple DNN instances we investigate whether the image set used to train a given DNN and its architecture matter with regard to its suitability as a model for the human visual cortex.

Initial findings suggested that models performing better at a given classification task are also better models of the human visual cortex (Kietzmann, McClure, et al., 2019; Kriegeskorte, 2015). Shallow vision models, performing worse than a DNN on multiple visual recognition tasks, were also outperformed with regard to their ability to predict cortical representations in human IT (Khaligh-Razavi & Kriegeskorte, 2014). In addition, randomly initialized DNNs (Yamins et al., 2014) confirmed this alleged relationship: the better its task performance, the higher the similarity of its internal representations to those found in primate IT. However, recent evidence suggests that engineering solutions might have diverged from biological vision systems as deeper and better performing models might not always be better able to predict cortical representations in human IT (Abbasi-Asl et al., 2018; Kalfas et al., 2017; Storrs & Kriegeskorte, 2019; Storrs et al., 2017). We thus also probe whether our brain-inspired architecture vNet trained on the ecologically more valid image set ecoset yields better prediction of cortical representations in human IT than state-of-art computer vision models, outperforming vNet at complex recognition tasks.

To conclude, in Chapter 4 I will first ask whether a DNN trained on ecologically relevant input statistics may better explain human visual cortex representations than when trained on a visual recognition task designed for machine learning. In a second step I will then test whether this brain-inspired DNN architecture may better explain cortical representations than state-of-the art models from computer vision and computational neuroscience.

1.8 Thesis overview

The following three Chapters will provide a thorough description of our investigations. In Chapter 2 I describe which impact the initial set of weights has on the representations of a given DNN. I also discuss the implications on the usage of single off-the-shelf network instances as models of the human visual cortex. Chapter 3 gives a detailed account of the

creation of an ecologically more valid image set designed for computational neuroscience. In Chapter 4 the themes of the two previous Chapters come together: here we investigate multiple instances per DNN architecture to elucidate whether further ecological and biological inspiration may yield better models of the human visual cortex. Finally, Chapter 5 concludes the thesis by summarizing the findings of each Chapter, relating it to ongoing investigations and giving an outlook to future experiments.

Chapter 2

Individual differences between deep neural network instances

Abstract

Deep neural networks (DNNs) excel at visual recognition tasks and are increasingly used as a modeling framework for neural computations in the visual system of the primate brain. In both engineering and computational neuroscience, analyses of DNNs usually rely on single network instances. However, each DNN instance, just like an individual brain, has a unique connectivity and unique representations. Here, we investigate DNN individual differences by training multiple network instances of the same architecture with the same training procedure, varying only the random initialization of the network weights. Using representational similarity analysis, a technique from systems neuroscience that characterizes representations in high-dimensional spaces, we demonstrate that this minimal change in initial conditions prior to training leads to substantial representational differences despite indistinguishable classification performance. These individual differences increase with network depth for a large range of distance measures, indicating shared lower-level, but diverging intermediate and higher-level representations across networks. As a possible explanation for these effects, we argue that the category objective used to train the networks, while optimizing for class separability, does not sufficiently constrain the arrangement of category clusters and instances in high-dimensional activation space. In line with this, category separability increases across layers while representational consistency decreases. We show that this decrease is due to differences in the alignment of category exemplars, rather than a misalignment of category centroids. Network regularization, in the form of Bernoulli sampling during training and test, increases the consistency of learned representations. Yet, considerable differences remain, suggesting that computational neuroscientists working with DNNs should base their inference on multiple network instances rather than single off-the-shelf networks. Characterizing individual differences can help machine learning researchers obtain a better understanding of successes and failures of DNN training and of the internal representations of DNN models.

2.1 Introduction

Deep neural networks have recently moved into the focus of the computational neuroscience community. Having revolutionized computer vision with unprecedented task performance, the corresponding networks were soon tested for their ability to explain information processing in the brain. To date, task-optimized deep neural networks constitute the best model class for predicting activity across multiple regions of the primate visual cortex (Cadieu et al., 2014; Güçlü & van Gerven, 2015; Khaligh-Razavi et al., 2014; Schrimpf et al., 2018; Yamins et al., 2014). Yet, the advent of computer vision models in computational neuroscience raises the question in how far network internal representations generalize, or whether network instances, just like experimental participants, exhibit individual differences. This would imply that the common practice of analyzing a single network instance is misguided and that groups of networks need to be analyzed to ensure the validity of insights gained.

Here we investigate individual differences among deep neural networks that arise from a minimal experimental intervention: changing the random seed of the network weights prior to training while keeping all other aspects identical. Our analyses of the network internal representations learned during training build on representational similarity analysis (RSA; Kriegeskorte et al., 2008), a multivariate analysis technique from systems neuroscience. RSA is based on the concept of a representational dissimilarity matrix (RDM), which characterizes a system's inner stimulus representations in terms of pairwise response differences. Together, the set of all possible pairwise comparisons provides an estimate of the geometric arrangement of the stimuli in high-dimensional activation space. The representations of two DNNs are considered similar if they emphasize the same distinctions among the stimuli, i.e. to the degree that their RDMs agree. Comparisons on the level of RDMs, which can be computed in source spaces of different dimensionality, thereby side-step the problem of defining a correspondence mapping between the units of the networks. To quantify RDM agreement across network instances, we define *representational consistency* as the shared variance between network RDMs (squared Pearson correlation of the upper triangle of the RDMs; Fig 2.1).

Based on this analysis approach, we visualize the internal network representations and test them for consistency. We then compare the size of the effects observed to differences between networks trained with different input statistics and test the reliability of the observations across multiple activity distance measures. Subsequently, we explore possible causes for these individual differences and investigate their interaction with network regularization.

2.2 Materials and methods

2.2.1 Deep neural network training

The main architecture used throughout all experiments presented here is All-CNN-C (Springenberg et al., 2015), a 9 layer fully convolutional network that exhibits state of the art performance on the CIFAR-10 dataset (Krizhevsky, 2009). To optimize architectural simplicity, the network uses only convolutional layers with a stride of 2 at layer 3 and 6 to replace max- or mean-pooling. We used the same number of feature maps [96, 96, 96, 192, 192, 192, 192, 192, 10] and kernel-sizes [3, 3, 3, 3, 3, 3, 3, 1, 1] as in the original paper (Fig 2.1 A).

To show that our results generalize beyond a single DNN architecture we trained an additional architecture reminiscent of VGG-S (Chatfield et al., 2014). In contrast to the original VGG-S architecture, we replaced the two deepest, fully-connected layers with convolutional layers to reduce the number of trainable parameters and thus the training duration by $\sim$80%. The number of feature maps used per layer was [96, 128, 256, 512, 512, 1024, 1024], and the kernel sizes were [7, 5, 3, 3, 3, 3, 3]. We used ReLU as the activation function at every layer. Mirroring the kernel sizes across layers, we refer to this architecture as "VGG-753".

All-CNN-C network instances were trained for 350 epochs using a Momentum term of 0.9 and a batch size of 128. All networks of the VGG-753 architecture were trained for 250 epochs using ADAM with an epsilon term of 0.1 and a batch size of 512. For both architectures, we used an initial learning rate of 0.01, the L2 coefficient was set to 10^{-5}, and we performed norm-clipping of the gradients at 500. Training of the main DNNs was performed on the full CIFAR-10 image set. CIFAR-10 consists of 10 categories of objects, each of which is represented by 5,000 training and 1,000 test images. Ten network instances were trained for the main analyses, all without dropout.

Network training was identical across all instances (same architecture, same dataset, same sequence of data points), with the exception of the random seed for the weight initialization. As a result, the networks only differ in the initial random weights, which are, however, sampled from the same distribution (He et al., 2014).

2.2.2 Comparing layer-internal representations across network instances

Representational similarity analysis and representational consistency

We characterize the internal representations of the trained networks based on representational similarity analysis (RSA; Kriegeskorte et al., 2008), a method used widely across systems neuroscience to gain insight into representations in high-dimensional spaces.

RSA builds upon the concept of a representational dissimilarity matrix (RDM), which stores all pairwise distances between the stimulus-driven pattern activations in response to a large set of input stimuli (Fig 2.1 A). Here we use 1,000 test stimuli, 100 from each of the 10 CIFAR-10 categories, such that the resulting RDMs have a size of 1000x1000 (Fig 2.1 B). The RDMs are symmetric around the diagonal and therefore contain 499,500 unique distance estimates. In the current set of experiments, pairwise distances (using correlation-, cosine-, and (unit length pattern-based) Euclidean-distance) are measured in the activation space of individual layers, where each unit corresponds to its own input dimension. The resulting matrix thereby characterizes the representational space spanned by the network units, as it depicts the geometric relations of all different input stimuli with respect to each other. This focus on relative distances renders RSA largely invariant to rotations of the input space (including random shuffling of input dimensions, but see Appendix A). It is therefore well suited for comparisons across deep neural network instances.

Because RDMs are distance matrices, they can be used as a basis for multidimensional scaling (MDS) to project the high-dimensional network activation patterns into 2D. While not a lossless operation, as high-dimensional distances can usually not be perfectly reproduced in 2D, MDS does nevertheless enable us to gain first insights into the internal organization by visualizing how network layers cluster the 1000 test images from the 10 different categories.

In addition to enabling 2D visualizations of network internal representations (or, put differently, the organization of test-images in high-dimensional layer activation space, Fig 2.3), RDMs themselves can be used as observations (each RDM is a point in the high-dimensional space of all possible RDMs) and thereby form the basis for computing "second-level" distance matrices. The resulting distance matrices can be used to compare representations across multiple network layers and network instances (rather than test-images as in first-level RDMs). Here, we compute a second level distance matrix based on the RDMs for all network layers and instances. Again, we use MDS to visualize the data points in 2D (Fig 2.4).

For a more quantitative comparison of network internal representations, characterized here in terms of RDMs, we define *representational consistency* as the shared variance across representational distances observed in high-dimensional network activation space. Representational consistency is computed as squared Pearson correlation between RDMs (Fig 2.1 C). If two network instances separate the test stimuli with similar geometry, the representational consistency will be high (max 1), whereas uncorrelated RDMs exhibit low representational consistency (min 0).

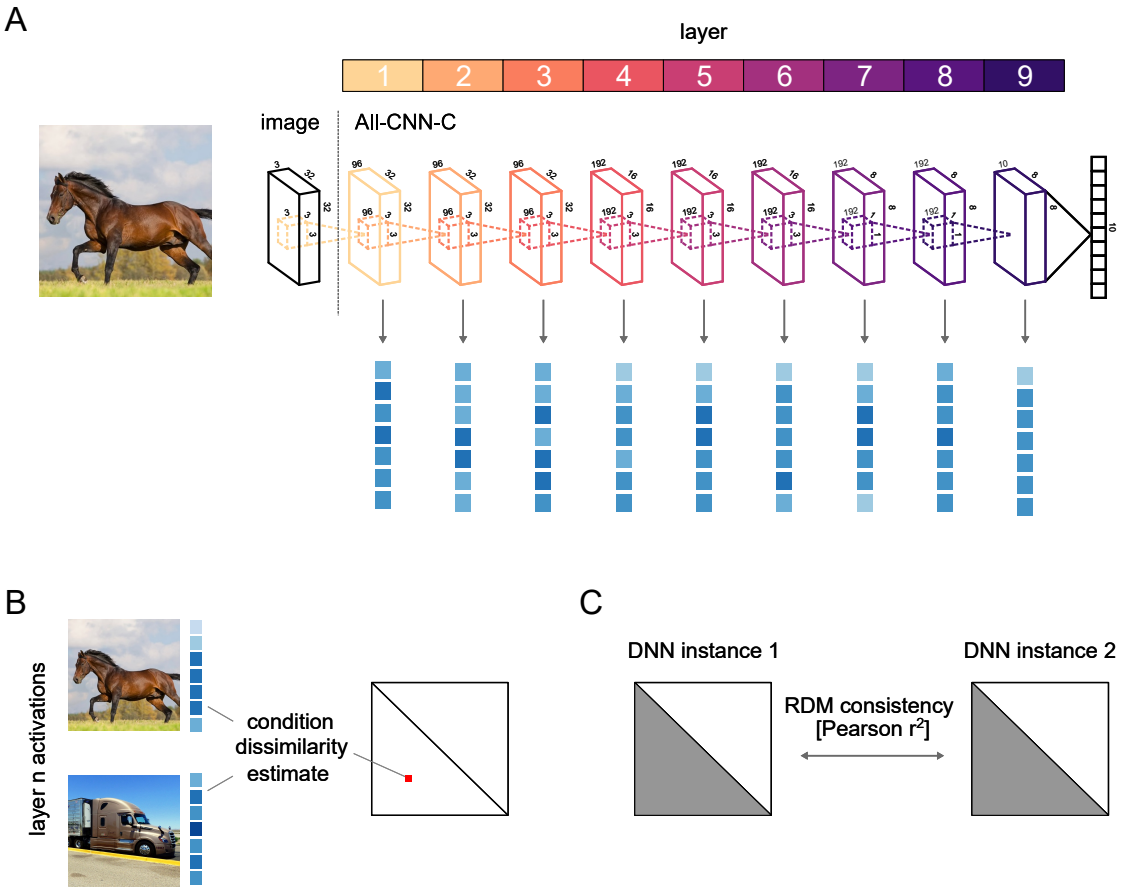


Fig. 2.1 **Characterizing network internal representations via representational similarity analysis and representational consistency.** **(A)** Our comparisons of network internal representations were based on their multivariate activation patterns, extracted from each layer of each network instance as it responded to each of 1000 test images. **(B)** These high-dimensional activation vectors were then used to perform a representational similarity analysis (RSA). The fundamental building blocks of RSA are representational dissimilarity matrices (RDMs), which store all pairwise distances between the network's responses to the set of test stimuli. Each test image elicits a multivariate population response in each of the network's layers, which corresponds to a point in the respective high-dimensional activation space. The geometry of these points, captured in the RDM, provides insight into the nature of the representation, as it indicates which stimuli are grouped together, and which are separated. **(C)** To compare pairs of network instances, we compute their representational consistency, defined as the shared variance between network RDMs.

Comparing the effect of weight initialization to differences in the input statistics

The main experimental manipulation in this work consists of using different random weights at the point of network initialization. To better understand the size of the effects on network internal representations, we compared the effects observed to differences that emerge from

using different images from the same categories (within-category split), or different categories altogether (across-category split).

To perform this control analysis, two subsets of CIFAR-10 were created. For the across-category division, we split the training and test sets on the level of categories. This resulted in two datasets with 5 categories each while preserving the number of images per category (5,000 training, 1,000 test images). For the within-category division, the dataset was split based on images rather than categories. This preserves the number of categories (10) but halves the number of training images per category. For an illustration of the splitting procedure that resulted in the within-category, and the across-category splits of CIFAR-10, see Fig 2.2.

In summary, the consistency of network instances resulting from different random weight initializations (different seeds, same categories, same images), was compared with (a) different images (same seed, same categories), and (b) different categories (same seed, different images; Fig 2.6). Five networks were trained for each half of the dataset for both splits (a, and b, resulting in $5 * 2 = 10$ network instances each). Representational consistency was computed using pairs of network instances with the same random seed (5 pairs for each split). Note that representational consistency was computed based on 1,000 test images from all 10 CIFAR-10 categories, independent of the image set used to train the networks.

Category clustering and its relation to representational consistency

To measure how well the layers of a network separate instances from different categories, we computed a category clustering index (CCI), which contrasts the distances of stimuli within the same category with the distances for stimuli originating from different categories. Based on the RDM computed for the 1000 test stimuli (100 stimuli per each of 10 categories), CCI contrasts distances of category exemplars within the category with distances across exemplars from different categories. It is defined as

$$CCI = \frac{(across - within)}{(across + within)}$$

and was computed for each layer of each network instance trained. CCI has a maximum of 1 (all categories cluster perfectly and are perfectly separable), and a minimum of 0 (no separability, same distances across and within categories).

In addition, we investigated the relationship between CCI and representational consistency. For each layer we computed the mean representational consistency across all 45 pairwise comparisons between 10 network instances and used Pearson correlation to demonstrate its relation to the mean class clustering indices (CCIs) across all 10 training seeds (Fig 2.8).

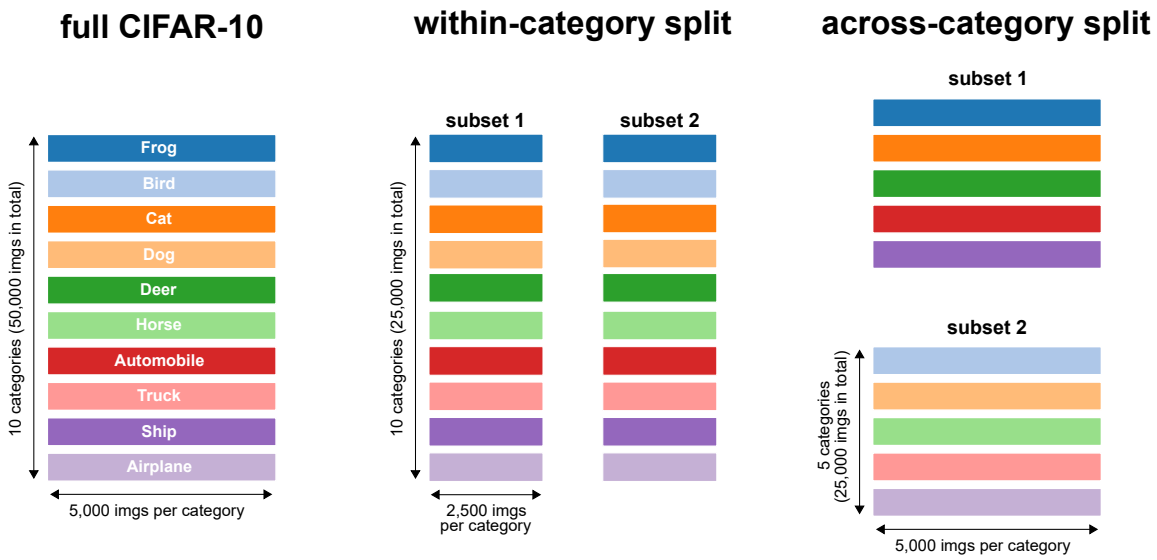


Fig. 2.2 **Visualization of the CIFAR-10 training sets used.** Different categories are shown as distinct colors. **Left panel:** The full CIFAR-10 training set consists of 10 categories with 5,000 images each, 50,000 images in total. **Center panel:** the within-category split dataset contains 10 categories with 2,500 images each, 25,000 images in total for each subset. **Right panel:** the across-category split dataset contains 5 categories with 5,000 images each, again 25,000 images in total for each subset. When splitting across categories, the number of animal- and vehicle-categories of the full CIFAR-10 set was equally distributed across the two subsets.

2.2.3 Investigating causes for decreasing representational consistency

To better understand the origins of changes in representational consistency, we compare (i) exemplar-based consistency, (ii) centroid-based consistency, (iii) consistency of within-category distances, and (iv) the effects of cocktail-blank normalization.

To understand whether a misalignment in the arrangement of individual category instances or the arrangement of entire classes is leading to decreased consistency, we computed the 10 class centroids and used their position in activation space to arrive at centroid-based representational consistency. This was compared with consistency based on all 1,000 stimuli (exemplar-based representational consistency), and consistency computed when only distances between instances of the same categories were considered (within-category consistency; Fig 2.9 A).

To rule out effects of changed RDM size in case of centroid-based RDMs (centroid RDMs contain 45 pairwise distances whereas the exemplar-based RDMs are composed of 499,500 entries), we computed a null distribution of RDM consistency based on centroids computed from randomly sampled classes (Fig 2.9 B).

Finally, to test in how far the distance measure used, rather than the representational geometries themselves, could be the source of individual differences (see Appendix A), we performed a cocktail blank normalization by subtracting the mean activation pattern across all images from each network unit, before computing the RDMs and representational consistency (Fig 2.11).

Experiments with regularization (Bernoulli dropout)

In an additional set of experiments, we explored how network regularization (here in the form of Bernoulli dropout) can affect network internal representations. Using the full CIFAR-10 set, we trained a set of 10 networks for each of 9 dropout levels (dropout probability ranging from 0 to 0.8, each of the resulting 90 DNNs was trained for 350 epochs). After training, we extracted network activations for a set of test images either by using no dropout at test time or by using multiple dropout samples for each test image. We obtained up to 10 samples extracted for each image while keeping the dropout mask identical across network instances and the dropout rate identical to training. We created one RDM per sample and then averaged up to 10 RDMs to obtain a single RDM representing the expected representational geometry upon dropout sampling (Fig 2.12).

2.3 Results

We here investigate the extent to which deep neural networks exhibit individual differences. We approach this question by training multiple instances of the All-CNN-C network architecture (Springenberg et al., 2015) and a custom architecture (VGG-753) on an object classification task (CIFAR-10), followed by an in-depth analysis of resulting network internal representations. Network instances varied only in the initial random assignment of weights, while all other aspects of network training were kept identical. All networks performed similarly in terms of classification accuracy (ranging between 84.4 - 85.9% and 77.6 - 79.0% top-1 accuracy for All-CNN-C, and VGG-753, respectively).

To study and compare network internal representations, we extracted network activation patterns for 1000 test images (100 for each of the CIFAR-10 categories, Fig 2.1 A) and characterized the underlying representations in terms of pairwise distances in the high-dimensional activation space (Fig 2.1 B). The reasoning of this approach is that if two images are processed similarly in a given layer, then the distance between their activation vectors will be low, whereas images that elicit distinct patterns will have a large activation distance. The matrix of all pairwise distances (size 1000x1000) thereby describes the representational

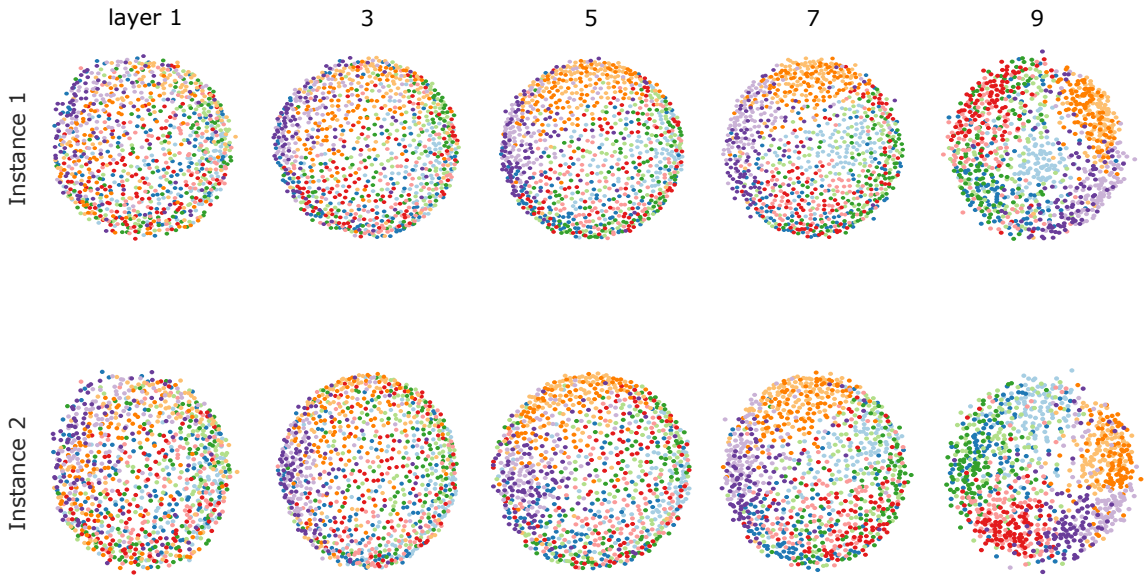


Fig. 2.3 **2D visualization of representational geometries in different depths of two network instances.** The internal representations of two network instances were characterized based on their representational geometries. We computed the pairwise distances (correlation distance) between activity patterns in response to 1000 test stimuli from 10 visual categories and visualized them in 2D via multidimensional scaling (MDS; metric stress criterion). With increasing depth, networks exhibit increased category clustering and emerging differences.

geometry of the test images, i.e. how instances of various object categories are grouped and separated by the units of a given layer (Kriegeskorte & Kievit, 2013).

2.3.1 Stronger category clustering and individual differences in later network layers

To visualize the representational geometries of different network instances and layers, we projected the data into 2D using multidimensional scaling (MDS, metric stress). As can be seen in Fig 2.3 for two exemplary cases of All-CNN-C, subsequent network layers increasingly separate out the different image categories, in line with the training objective.

Moving closer to the question of individual differences in network representations, we next investigated similarities on the level of RDMs. We again computed pairwise distances, but this time not based on activation patterns, but rather based on the network RDMs. Comparing patterns of representational distances has multiple benefits. For one, they offer a characterization of network internal representations that is largely invariant to rotations of the underlying high-dimensional space, including a random shuffle of network units ((see Appendix A for more details). Secondly, representational spaces of varying dimensionality

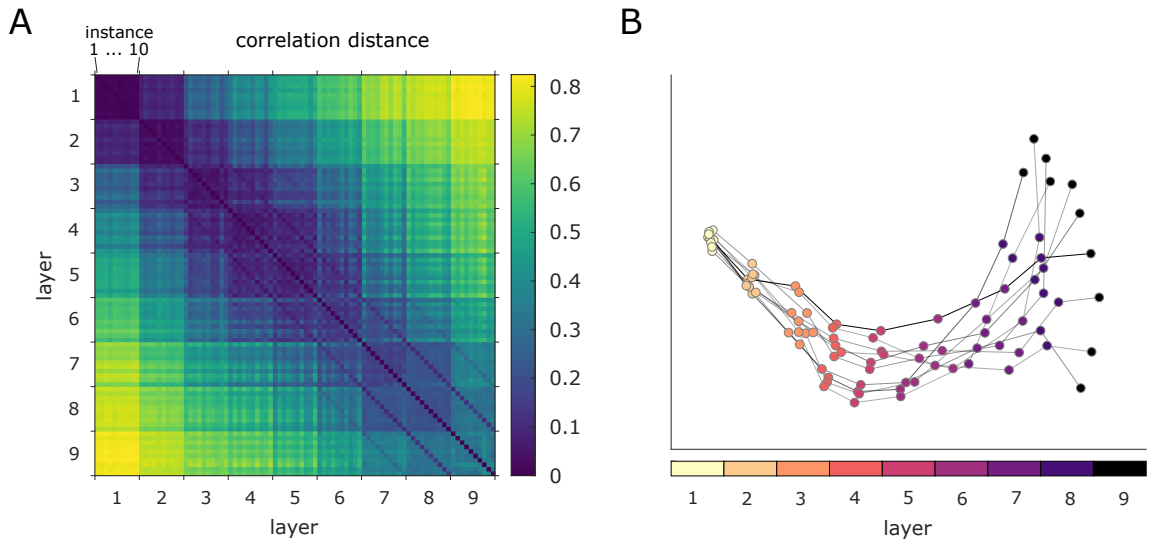


Fig. 2.4 **Network individual differences emerge with increasing network depth.** (**A**) We compare the representational geometries across all network instances (10) and layers (9 convolutional) for All-CNN-C by computing all pairwise distances between the corresponding RDMs. The dark blue stripes in the diagonal parallel and directly adjacent to the main diagonal indicate a higher similarity across adjacent layers within a given network instance compared to the similarities across instances within a given network layer (see figure Fig2, Appendix A). (**B**) We projected the data points in (**A**) (one for each layer and instance) into 2D via MDS. Layers of individual network instances are connected via grey lines. While early representational geometries are highly similar, individual differences emerge gradually with increasing network depth.

can be directly compared, as the dimensionality of the RDM is fixed by the number of test images used.

To compare representations across network layers and instances, we computed a second-level distance matrix ($n_{\text{nr. of network instances}}$ * $n_{\text{nr. of layers}}$ as rows and columns). Visualizing the respective distances in 2D (MDS, metric stress), we observe that representations diverge substantially with increasing network depth (Fig 2.4). While different network instances are highly similar in layer 1, indicating agreement in the underlying representations, subsequent layers diverge gradually with increasing network depth. Note that the blue stripes parallel to the main diagonal in Fig 2.4 A indicate higher similarity across layers within a given network instance compared to the similarities across instances and within a network layers.

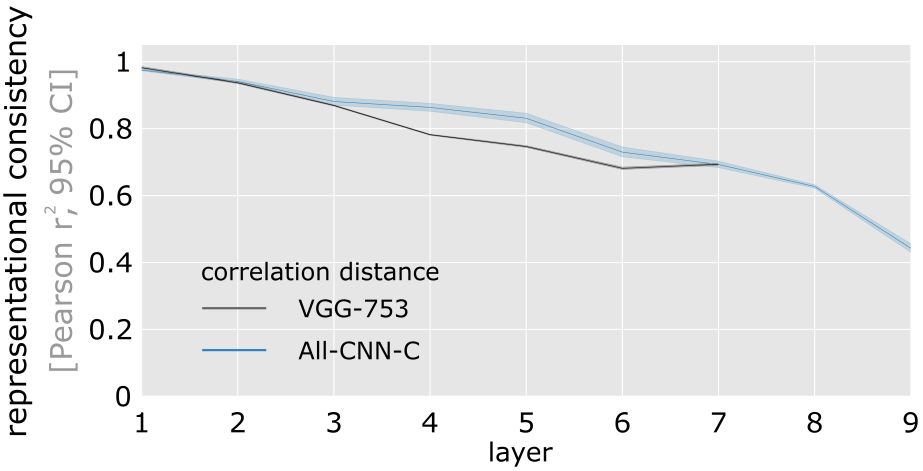


Fig. 2.5 **Representational consistency declines with increasing network depth.** Shown is the average representational consistency for each layer computed across all pairwise comparisons of network instances (45 comparisons for 10 instances, computed separately for two network architectures using correlation distance to compute RDMs). Error bars indicate 95% confidence intervals (bootstrapped).

2.3.2 Representational consistency decreases with increasing network depth

Following this initial qualitative assessment, we performed quantitative analyses for each network layer by testing how well the distribution of representational distances generalizes across network instances. This was accomplished by computing *representational consistency*, defined as the shared variance between the upper triangle of the respective RDMs (Fig 2.1 C, each triangle contains 499,500 distance estimates, results are obtained from 45 pairwise network comparisons for each respective layer and network architecture as 10 network instances are trained for each architecture). This measure of consistency is based on all pairwise distances between category exemplars (100 exemplars for 10 categories each). We therefore refer to this as *exemplar-based* consistency.

Two network architectures were tested (All-CNN-C, and VGG-753, see methods for details). Correlation distance was chosen as dissimilarity measure in computing RDMs, as it is currently the most frequently used distance measure in systems and computational neuroscience. As shown in Fig 2.5, representational consistency drops substantially with increasing network depth for both network architectures. To get better insights into the size of this effect, additional networks were trained (a) based on different images originating from the same categories, and (b) based on different categories (see methods for details). The observed drops in consistency for different weight initializations (to about 43% and 71%

for All-CNN-C and VGG-753, respectively), are comparable to training the networks with the same distribution of categories but completely separate image datasets (Fig 2.6, blue vs. orange).

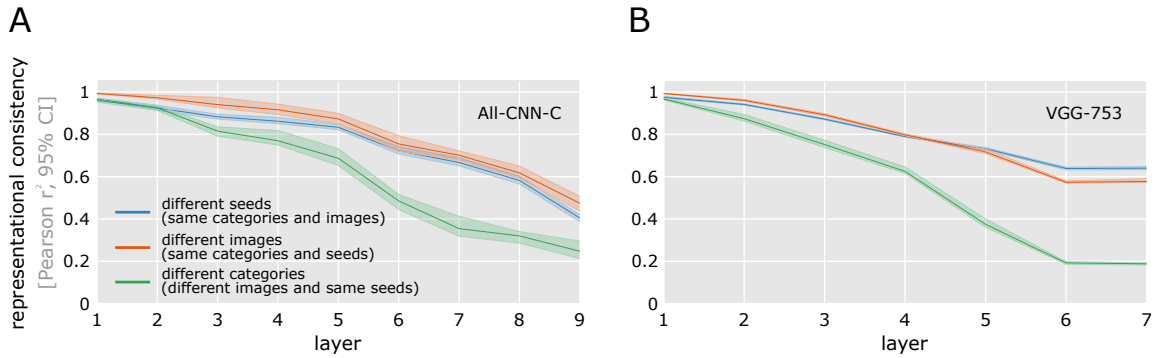


Fig. 2.6 **Representational consistency declines with increasing network depth when trained on separate image sets.** To better understand the size of the effect in Figure 4, we trained a separate set of networks based on (**A**) All-CNN-C, (**B**) VGG-753) while using different images from the same categories but the same seeds (orange), and different categories, different images, and same seeds (green). The minimal intervention of using a different seed for the random weight initialization (shown in blue, data equivalent to Fig 2.5) affects the internal representations about as much as using a completely different set of training images (10 categories per training set; orange). Please note that part of the larger drop in representational consistency for training with different categories (5 categories per training set; green) can be attributed to training only five categories while computing the RDMs with images from all 10 categories.

To ensure that the effects observed are not specific to correlation distance used in computing the RDMs further analyses were performed based on the following distance measures as well: cosine, (unit length pattern-based) Euclidean distance and norm difference (measuring the absolute difference in the norm activation vectors; Fig 2.7). In all cases, representational consistency was observed to drop considerably with increasing network depth. These results demonstrate that while different network instances reach very similar classification performance, they do so via distinct internal representations in the intermediate and higher network layers.

2.3.3 Causes of decreasing representational consistency

We have shown above that different network instances can exhibit substantial individual differences in their internal representations, comparable to networks trained with completely separate image sets. This finding has important implications for computational neuroscience, where single off-the-shelf (engineering) networks are often used as models of information

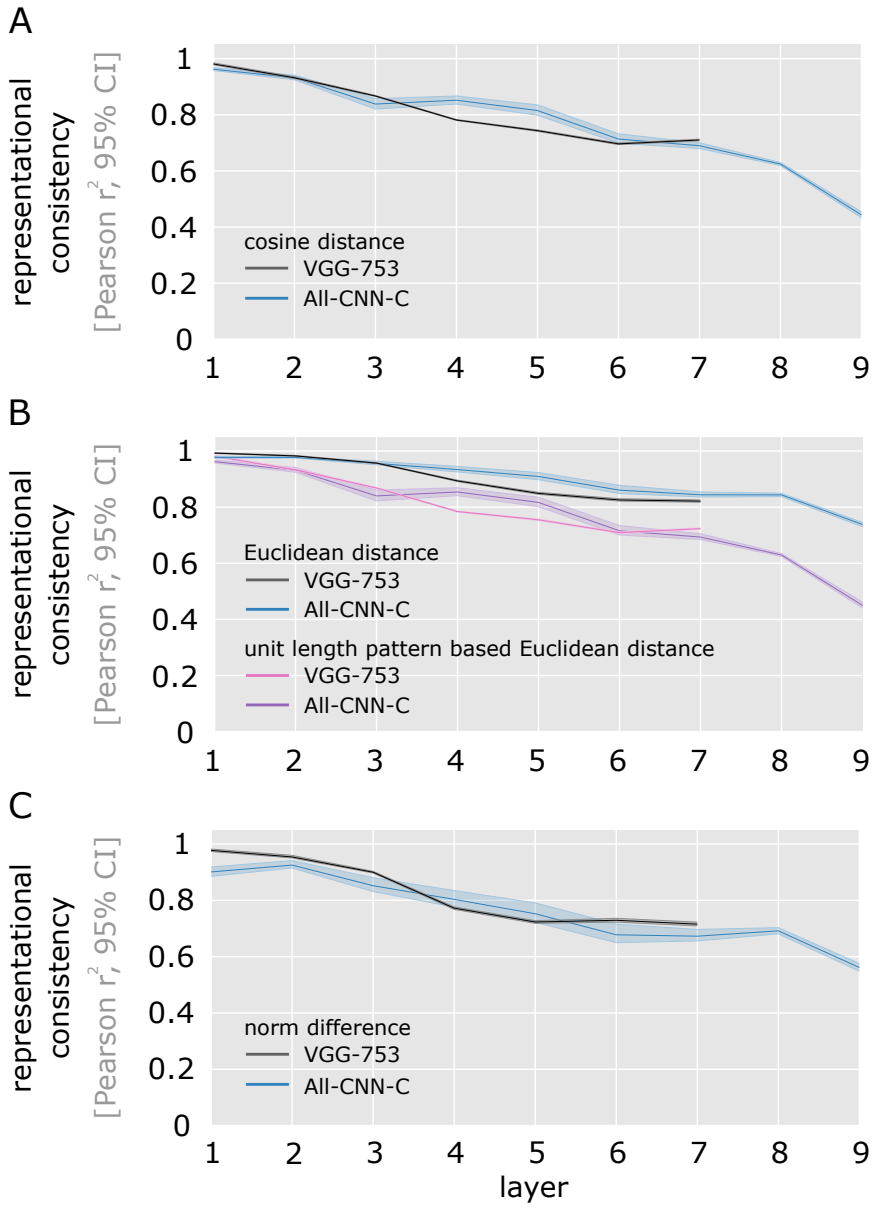


Fig. 2.7 **Representational consistency declines with increasing network depth irrespective of distance measure used to compute RDMs.** Representational consistency decreases with increasing layer depth for both tested DNN architectures, and across multiple different ways to measure distances in multivariate population responses (cosine (**A**), Euclidean distance and unit length pattern-based Euclidean distance (**B**), and differences in vector norm (**C**)). We show the average representational consistency for each layer, computed across all pairwise comparisons of network instances (45 comparisons for 10 instances), together with a 95% bootstrapped confidence interval.

processing in the brain. Next, we investigated why our network instances exhibit individual

differences despite reaching very similar classification accuracy and as a result of a minimal intervention.

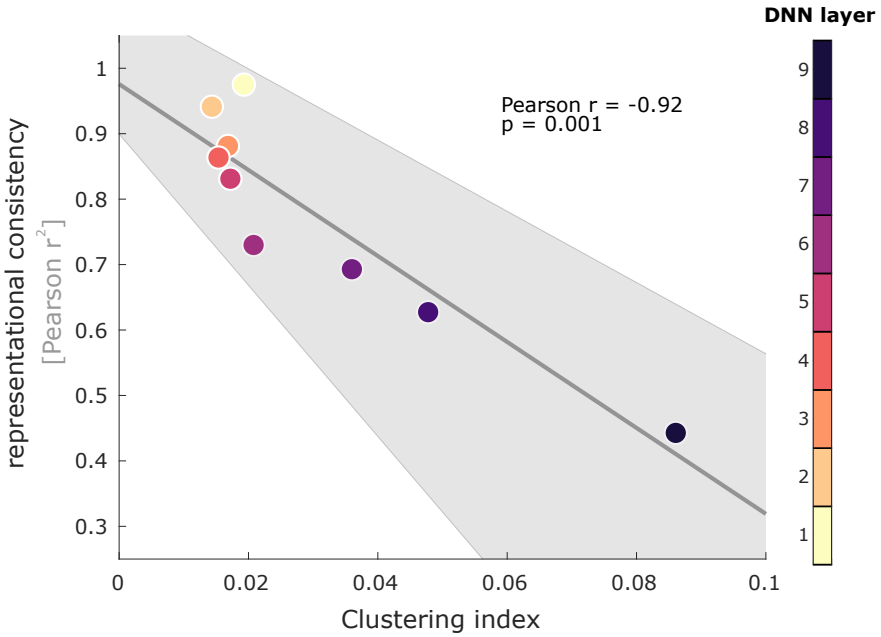


Fig. 2.8 **Representational consistency and category clustering are negatively correlated.** Optimized for categorization performance, deep neural networks aim to separate images from different categories in the network activation space. Here we show for all 10 network instances of All-CNN-C that increasing category separability across network layers (estimated here by a category clustering index) exhibits a negative relationship with mean representational consistency across trained network instances. Individual differences emerge while category clustering increases (95% bootstrapped CIs shown as grey area).

Our first analyses are based on the observation that the training goal of maximal category separability does not put a strong constraint on the relative positions of categories and category exemplars in high-dimensional activation space. To investigate this possibility, for the 10 network instances of All-CNN-C used in the previous section we first computed a category clustering index (CCI) for each network layer using the network responses to the set of 1000 test images (drawn from 10 categories). CCI is defined as the normalized difference in average distances for stimulus pairs from different categories (across) and stimulus pairs from the same category (within):

$$CCI = \frac{(across - within)}{(across + within)}$$

CCI approaches zero with no categorical organization and is positive if stimuli from the same category cluster together (maximum possible CCI = 1). We find a negative relationship between CCI and representational consistency (Pearson r = -0.92, p = 0.001; Pernet et al., 2013), indicating that network layers that separate categories better do exhibit stronger individual differences (Fig 2.8). (As shown in Fig 3 in Appendix A, we also investigated a possible association between representational consistency and accuracy, but found no evidence for such a relationship.)

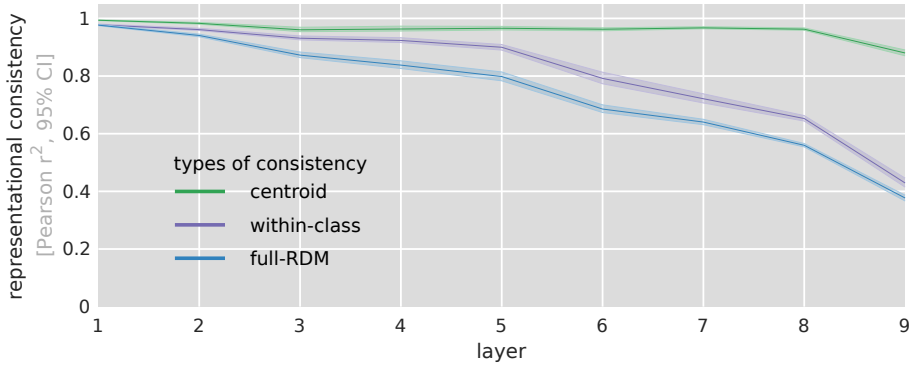


Fig. 2.9 **Category centroids are highly consistent across network instances.** Centroid-based representational consistency (green) remains comparably high throughout, whereas the consistency of within-category distances decreases significantly with increasing network depth. This indicates that differences in the arrangement of individual object instances, rather than large scale differences between class centroids are the main contributor to the observed individual differences. To allow for a comparison between centroid-based, full-RDM, and within-class consistency, we here computed consistency in spaces with the same dimensionality (for details, see main text).

This correlation between consistency and category clustering is consistent with two possible scenarios: networks can exhibit a different arrangement of the overall category clusters, or different arrangements of individual exemplars within the category clusters, as both are not constrained by the training objective to categorize. To investigate the variability in general cluster placement, we computed representational consistency based on the ten category centroids (RDMs computed from the pairwise distances of average response patterns for each category) and compared it with exemplar-based consistency.Note that centroid-based consistency relies on the 45 pairwise comparison between the 10 class centroids, whereas exemplar-based consistency relies on the 499,500 pairwise comparisons between the 1,000 exemplars. To allow for a comparison within the same dimensionality, for each pair of network instances we sampled 45 pairwise comparisons from the full RDM without replacement and averaged consistency across all 11,100 samples. The reliable arrangement of category centroids suggests that the main source of the observed individual differences

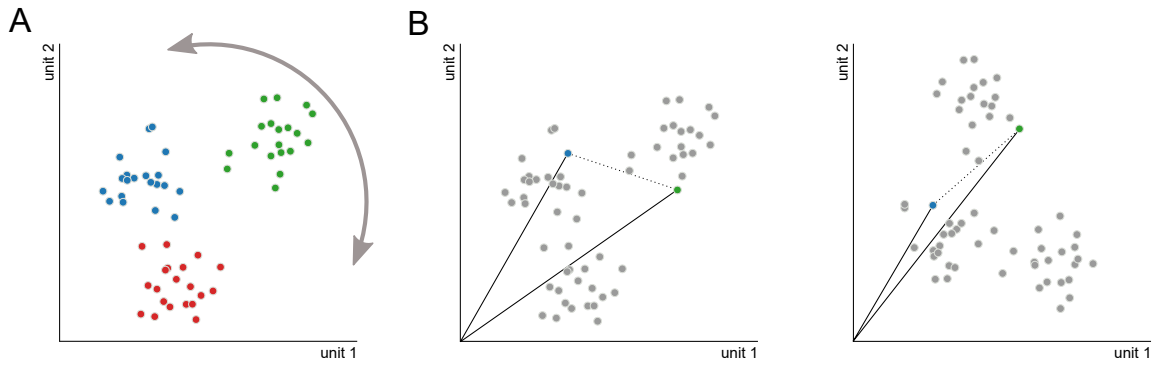


Fig. 2.10 **Rotation of ReLU activation space affects correlation- and cosine-distances.** (**A**) Three exemplary classes (blue, green, red) are rotated in the all-positive (post-ReLU) activation space, here shown as a 2D example. (**B**) When comparing the activation space before (left panel) and after the rotation, the angle between pairs of images can differ markedly, thereby leading to lower representational consistency despite an overall stable data arrangement. (see Appendix A for simulations using correlation distance).

lies in the arrangement of category exemplars within the category clusters. This view was corroborated by computing consistency on the within-category dissimilarities using the same sampling approach as for the low-dimensional full-RDM consistency. Here we observe a drop in consistency that is largely comparable to the decrease observed for low-dimensional (sampled) full-RDM consistency (Fig 2.9, blue) and to the decrease originally observed for full-RDM consistency without sampling (Fig 2.5).

In addition to an individual placement of category centroids and category instances, some properties of the underlying dissimilarity measures can be a source for lower representational consistency, especially in cases of a rotated representational space. Many commonly used DNNs use rectified linear units (ReLUs) as a nonlinear operation, resulting in unit activations ≥ 0. While overall rotations of this all-positive space will not affect classification performance, they can affect correlation and cosine distances (see Fig 2.10, and Appendix A (Fig 1) showing in addition that rotations around the origin affect correlation distances but not cosine distances).

To test the magnitude of this effect, we subtracted the mean activation pattern across all test images from the units of a given layer (cocktail blank normalization). As shown in Fig 2.11, this normalization leads to increases in representational consistency for RDMs computed using correlation or cosine distance. While the size of the effect is comparably small, these results indicate that a cocktail blank normalization can be of potential benefit when comparing correlation- or cosine-based RDMs of multiple DNNs or DNNs and brain data.

2.3.4 Network regularization (Bernoulli dropout) affects representational consistency

An explanation of individual differences via missing constraints imposed by the training objective raises the possibility that explicit regularization during network training can provide the missing representational constraints (McClure & Kriegeskorte, 2016b; Srivastava et al., 2014). We investigated this possibility experimentally by training networks at various levels of dropout regularization. We trained 10 network instances of All-CNN-C for each of 9 dropout levels (Bernoulli dropout probability ranging from 0 to 0.8; a total of 90 network instances trained) and subsequently tested the resulting representations for their ability to classify input, and for their representational consistency. To test for differences in task performance, we computed the top-1 categorization accuracy for the training- and test data. For the test data, we contrast network inference with and without dropout. In line with the literature (Srivastava et al., 2014), we find reduced training accuracy, but enhanced test accuracy at moderate dropout levels (Fig 2.12 A).

The effects of dropout training on representational consistency were again investigated using layer 9 of All-CNN-C, which exhibited the lowest consistency levels in our original analyses. These analyses revealed that dropout regularization yields increased representational consistency across network instances. When using no dropout at test time, a dropout probability of 0.6 during training provides the highest consistency level, reaching an average of 64.7% shared variance (rightmost column in Fig 2.12 B).

In analogy to our analyses of test accuracy when applying dropout at the time of inference, we investigated in how far this may affect representational consistency estimates. For each

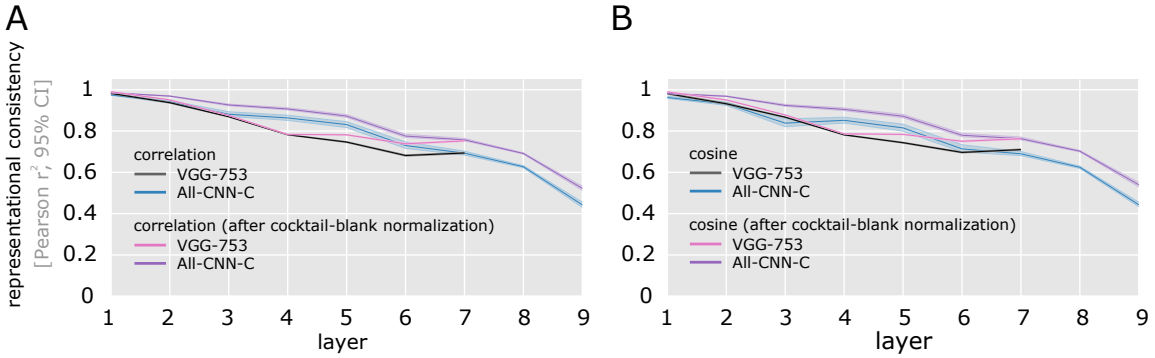


Fig. 2.11 **Cocktail blank normalization slightly increases consistency for correlation and cosine distance.** Centering the data via cocktail blank normalization increases representational consistency for correlation (**A**) and cosine distance (**B**). Euclidean distance measures are not affected, as the resulting representational geometries are rotationally invariant.

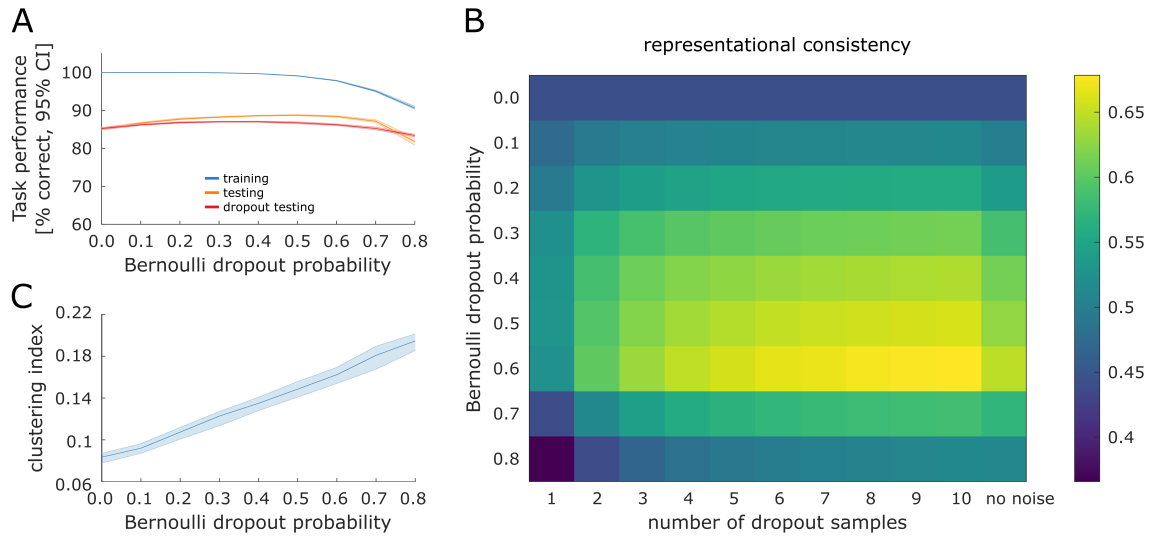


Fig. 2.12 **Effects of dropout regularization on task performance and representational consistency.** (**A**) Average task performance across all network instances (shown with 95% CI) for the training set (blue), test set (orange), and when using dropout sampling at inference time for the test set (red, 1 sample). For task performances of VGG-753, please see figure Fig 4, Appendix A. (**B**) Representational consistency in the final convolutional layer of All-CNN-C as a function of dropout probability during training and test (dropout probability at test time equal to training dropout). When using dropout at test time, multiple samples can be drawn for each stimulus in the test set (creating multiple RDMs). Consistency for network pairs was computed for the respective average RDM for each instance. Consistency was observed to be highest when 10 samples were obtained from a DNN trained and tested at a dropout rate of 60%. (**C**) For the penultimate layer of All-CNN-C the clustering index (for all layers also see Figure 7) increases with increasing Bernoulli dropout probability.

network instance, we computed 10 RDM samples while keeping the dropout mask identical across network instances and the rate identical to training. The average of a varying number of up to 10 RDM samples was subsequently used to compute representational consistency across network instances. We find that increasing the number of RDM samples led to increased representational consistency for all dropout levels. Maximum representational consistency was observed for 10 RDM samples at a dropout probability of 0.6, reaching an average of 67.8% shared variance across network instances. This suggests that dropout applied during training and test can increase the consistency of the representational distances across network instances.

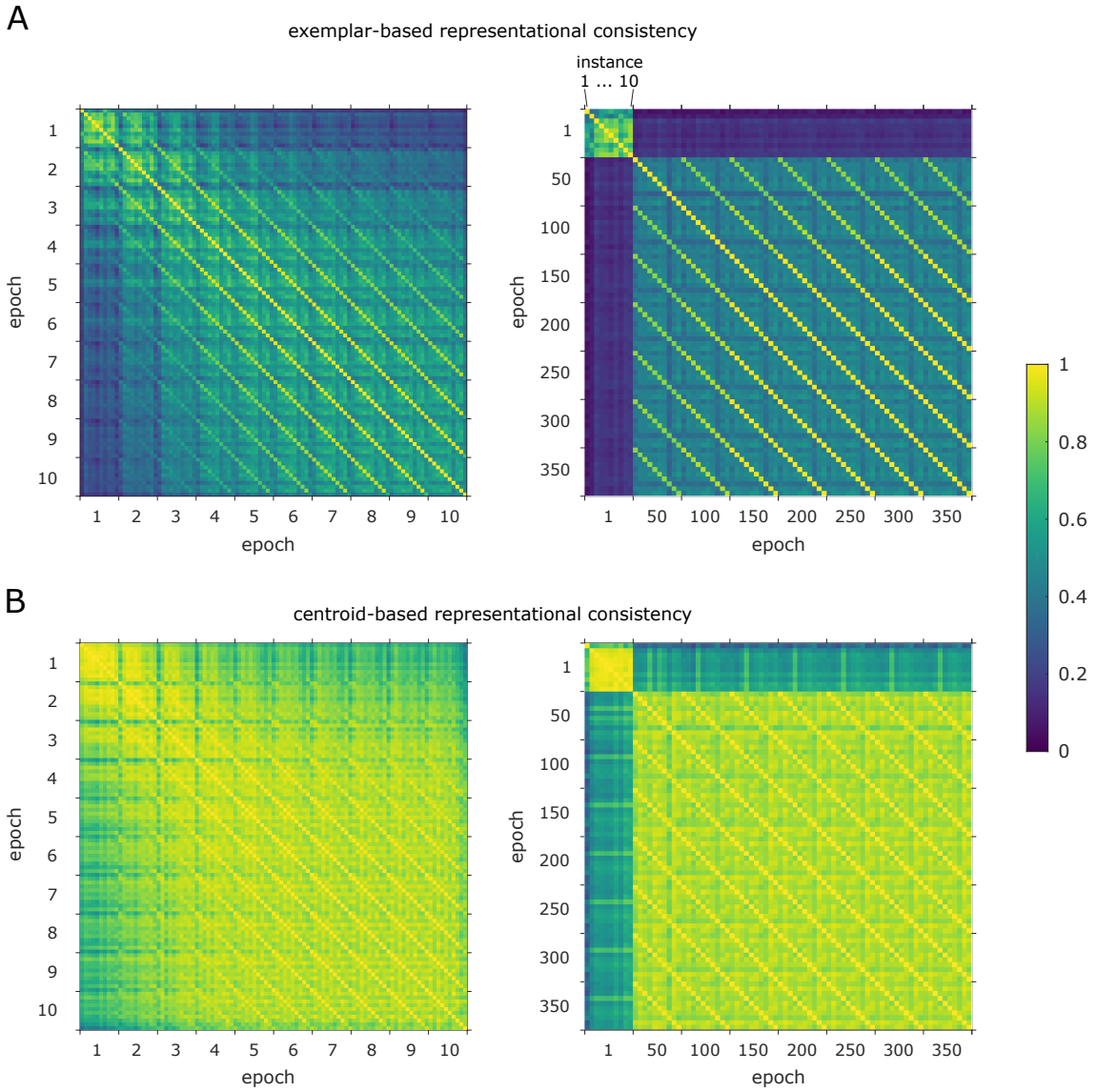


Fig. 2.13 **Penultimate-layer representational consistency across training consistency for RDMs based on individual images and on class centroids.** (**A**) Exemplar-based representational consistency across epochs [1, 2, 3, 4, 5, 6, 7, 8, 9, 10] (left) and across epochs [1, 50, 100, 150, 200, 250, 300, 350] (right). (**B**) Same as (**A**), but RDMs are based on class centroids instead of individual images. After the first epoch the representations across network instances show both image- and centroid-based consistency (left panel in both (**A**) and (**B**), respectively). But consistency decreases drastically when representations are compared with subsequent epochs, indicating that task training increases individual representational differences. From very few epochs onwards exemplar-based consistency within network instances and across epochs (traversing the RDM vertically or horizontally in steps of 10 cells) is remarkably stable and even saturates starting around epoch 150. The overall level of consistency across network instances is much higher for centroid- than for exemplar-based consistency ((**A**), vs. (**B**)).

As a possible explanation for how dropout could have affected representational consistency, we computed the category clustering index (CCI) for the penultimate layer of All-CNN-C and for different dropout levels. This is based on the idea that stronger clustering around the category centroids in the latest network layer will at the same time yield higher consistency, as the arrangement of category centroids is highly consistent. As shown in Fig 2.12 C, we observe a positive relationship between dropout probability and category clustering. However, while clustering is further enhanced for dropout levels >0.6, representational consistency starts decreasing. To further explore this effect, we re-computed centroid consistency for highest dropout level (0.8) and observed that centroid consistency is significantly decreased ($\mu_{dropout0.8} = 0.7422$, (95% bootstrapped CI = [0.6881, 0.7854]) compared to $\mu_{dropout0.0} = 0.8801$ (95% bootstrapped CI = [0.8700, 0.8905]) in the no dropout case). Thus, while denser clustering around centroids increases consistency in cases where the centroids themselves are consistent, high levels of dropout lead to less consistent centroids and therefore to an overall decrease in consistency.

2.3.5 Representational consistency across training trajectories

We have observed above that representational consistency across network instances is remarkably stable for category centroids. This raises the question as to whether this alignment is the result of task training, or whether category centroids are already well-aligned early during training. To investigate this, we computed representational consistency (exemplar-based and centroid based) across different network instances and training epochs. We extracted activation patterns from each network instance at different stages of training and subsequently computed pairwise representational consistency for the penultimate layer of All-CNN-C. Networks exhibit high consistency (computed across instances) after the first epoch, which decreases drastically from thereon, indicating that task training enhances individual differences. Yet, from very few epochs onwards networks exhibit remarkably stable representations with each network remaining on its own learning trajectory (Fig 2.13 A, multiple diagonal lines indicate stable representations across training compared to other network instances). Consistency seems to saturate from epoch 150 onwards, indicating minute changes in the network internal representations. Consistent with our earlier results, centroid-based consistency is overall higher across network instances even for the earliest epochs (Fig 2.13 B). Together, these results indicate that task training leads to decreased consistency, whereas learning trajectories of individual networks across time remain surprisingly robust.

2.4 Discussion

In a series of experiments, we here investigated how the minimal intervention of changing the initial set of weights in deep neural networks affects their internal representations. Operationalized as representational consistency, we demonstrated that significant individual differences emerge with increasing layer depth. This finding held true for various distance measures used to compute the RDMs (correlation distance, cosine distance, variants of Euclidean distance, and norm differences). RDMs computed from Euclidean distances showed the least differences. In part, this can be attributed to the fact that this distance measure is sensitive to differences in overall network activation magnitudes, which may overshadow more nuanced pattern dissimilarities, in line with the lower consistency observed for norm-standardizing Euclidean distances (unit length pattern-based Euclidean-distance).

We then explored multiple non-exclusive explanations for these network individual differences. Based on the hypothesis that the network training objective of optimizing for categorization performance may not sufficiently constrain the arrangement of categories and individual category instances, we analyzed category clustering, centroid arrangement, and within-category dissimilarities. All of these analyses point to a high consistency of category centroids, rendering differences between individual category instances the main contributor of the differences observed. As an additional source of variation, we identified an interaction between properties of the distance measures used and the ReLU nonlinearity in the DNNs. We showed that cocktail blank normalization in the DNN activation patterns can increase consistency for measures that are not robust to rotations that are not centered around zero (cosine distance) or general rotations (correlation distance). In addition to this, we showed that network regularization via dropout during training and test can enhance representational consistency estimates. As a partial explanation for this increase, we demonstrated that category centroids are highly consistent and that dropout enhances category clustering.

Our finding of considerable individual differences has important implications for computational neuroscience where single pre-trained computer vision networks are often used as models of information processing in the brain. Neglecting the potentially large variability in network representations will likely limit the generality of claims that can be derived from comparisons between DNNs and neural representations. While we here present multiple approaches that can increase consistency (cocktail-blank, dropout, and the choice of distance measure), significant differences remained. For computational neuroscience to take full advantage of the deep learning framework (Cichy & Kaiser, 2019; Kietzmann, McClure, et al., 2019; Kriegeskorte & Douglas, 2019; Richards et al., 2019), we therefore suggest that DNNs should be treated similarly to experimental participants, as analyses should be based on groups of network instances. Representational consistency as defined here will give

researchers a way to estimate the expected network variability for a given training scenario, and thereby enable them to better estimate how many networks are required to ensure that the insights drawn from them will generalize. In addition to the impact on computational neuroscience, we expect the concept of representational consistency, which can be applied across different network layers, architectures, or training epochs, to also benefit machine learning researchers in understanding differences among networks operating at different levels of task performance.

Chapter 3

Ecologically more valid input statistics for deep neural networks

Abstract

Deep learning, the key ingredient to today's high-performance computer vision, has recently found its way back into neuroscience where deep neural networks (DNNs) function as modeling framework for neural computations. The most commonly used DNNs are pre-trained on datasets originating from engineering challenges. They are therefore tuned to distributions of object categories that do not mirror the nature of the human visual experience. To alleviate this problem, we here introduce ecoset, the largest to date dataset designed specifically for computational neuroscience. Ecoset consists of >1.5 million unique images from 565 basic level categories that represent the most common, most concrete nouns of the English language. Most common to focus on important categories, and most concrete to include only concepts that can be visualized. Ecoset thereby closely matches the set of objects that humans frequently encounter and promises better computational models of the primate visual system. To allow for direct usability, ecoset will be freely available to the community for research and educational purposes.

3.1 Introduction

Deep neural networks (DNNs) have recently revolutionized computer vision and now dominate several areas of artificial intelligence. In computational neuroscience, too, DNNs are used increasingly as a powerful framework to instantiate and constrain neuroscience theories. Despite abundant differences in terms of missing biological details, DNNs provide the best currently available models for the computations along the primate visual system (Kietzmann et al., 2017; Kriegeskorte, 2015; Kriegeskorte & Golan, 2019; Marblestone et al., 2016; Richards et al., 2019; Serre, 2019; Storrs & Kriegeskorte, 2019).

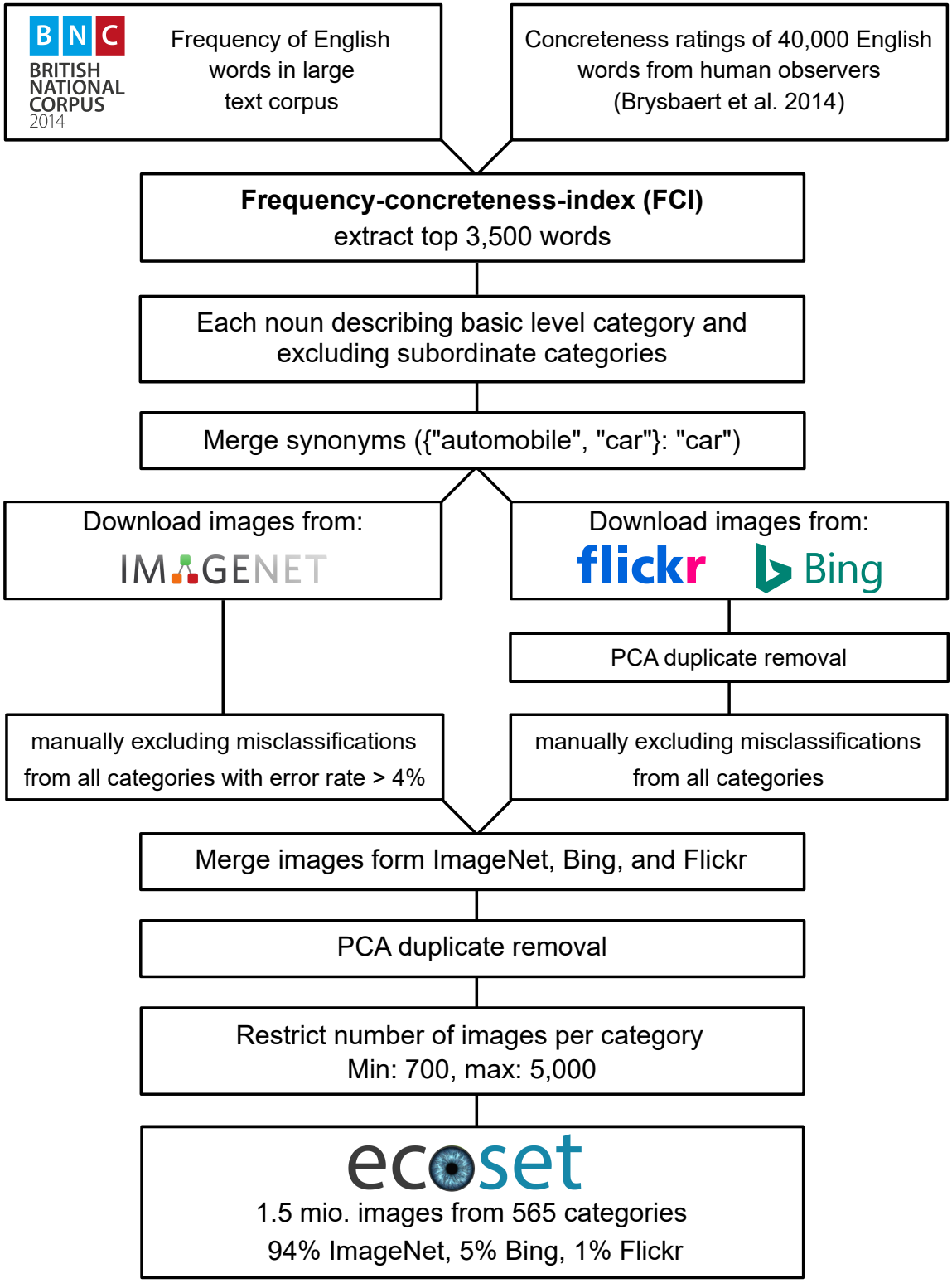


Fig. 3.1 **Selection process of ecoset categories and images.** Steps taken to create ecoset are shown schematically from top to bottom. For details, please see methods section.

Despite these promising early results, the most common approach in DNN-based computational neuroscience is to test networks that are pre-trained to excel at engineering challenges, such as the image classification task of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC 2012; A. Krizhevsky et al., 2012; Russakovsky et al., 2015). This computer vision dataset consists of 1,000 object categories differing in their origin and level of categorical specificity. For instance, next to basic level categories, the ILSVRC 2012 contains 120 different breeds of dogs while it lacks categories related to humans. This distribution of object categories is sensible from an engineering point of view, as it allows computer vision systems to demonstrate their versatile computational abilities. The human visual system, however, is known to be highly selective for its input statistics and important object categories, such as faces, bodies, tools, etc.. Successful models of the human visual system therefore need to be tuned to an appropriate set of object categories. In the past, such large-scale object datasets for deep learning in visual computational neuroscience were not available.

Here, we aim to alleviate this problem by introducing a new large scale image dataset suitable for deep learning in computational neuroscience: ecoset. Ecoset was created specifically to more closely approximate the human visual experience, and thereby to allow the field to train more accurate models of primate vision. Starting from a list of English nouns, the selection of object categories was based on linguistic frequency in the English language (British National Corpus, 2014; Leech et al., 2014), as well as concreteness ratings from human observers (M. Brysbaert et al., 2014). Linguistic frequency was used as a proxy for concept importance, whereas high concreteness ratings imply that the concept can be well approximated from images. The two parameters were subsequently joined to form a frequency-concreteness-index (FCI). Starting from the highest FCI, nouns were selected for inclusion if they described a basic level category. Category images were collected from ImageNet (93.5%), in addition to images obtained via Bing (5.1%, CC BY NC SA 2.0), and the image-hosting site Flickr (1.4%, CC BY NC SA 2.0). After multiple data quality assurance steps, ecoset now contains >1.5 million images originating from 565 non-overlapping basic level categories. Each category contains between 700 and 5,000 images. For an overview of the category and image selection process, see Fig 3.1 and the methods sections on "Dataset Statistics" and "Technical Validation". Examples of the ten categories with highest FCI are shown in Fig 3.2. As shown there, object categories for which strong neural selectivity is commonly found (e.g. faces, bodies, or tools) are included in ecoset.

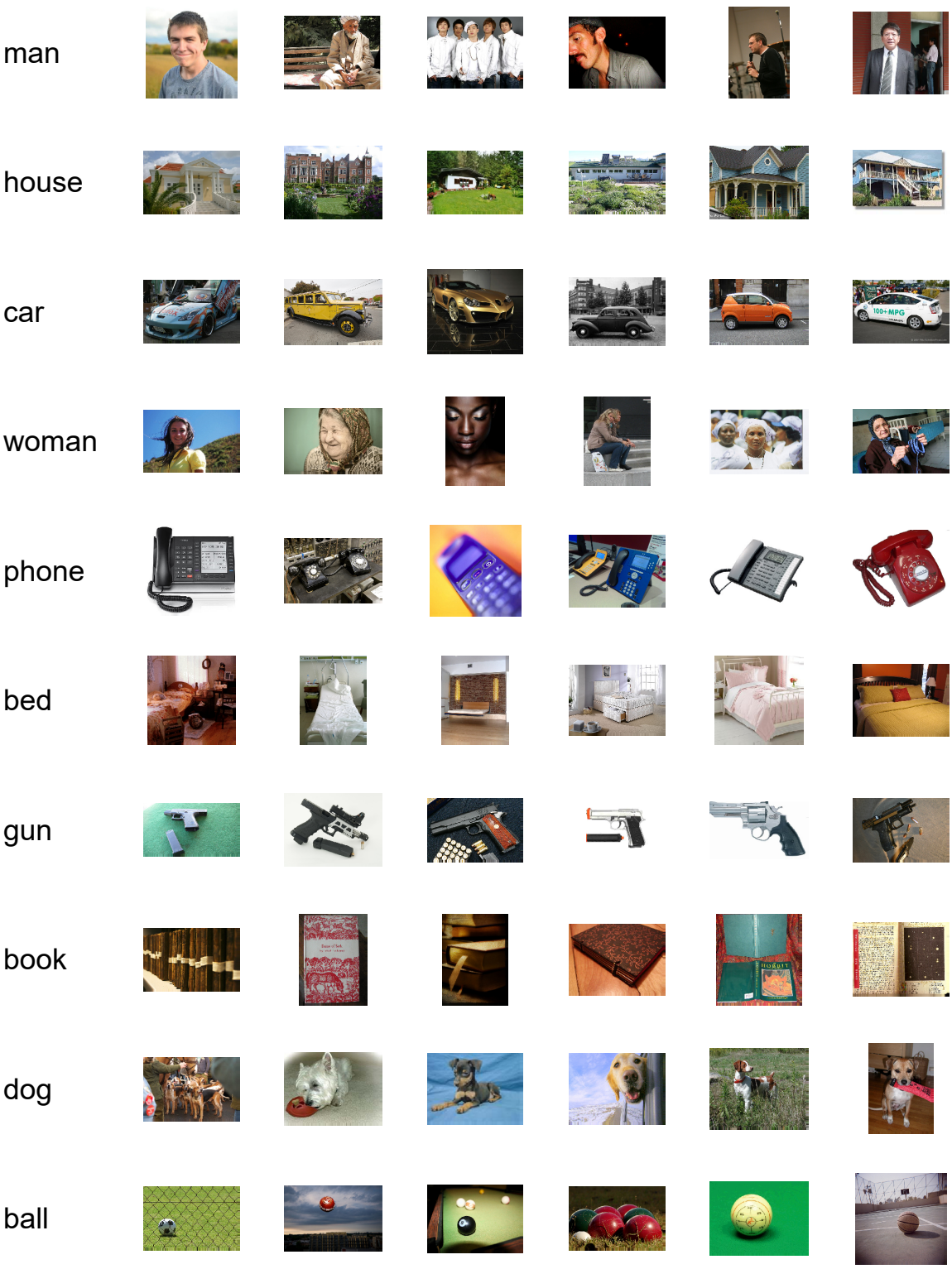


Fig. 3.2 **Example images from 10 ecoset categories.** Each row depicts images from one of the 10 (out of a total of 565) ecoset categories with the highest frequency-concreteness-index (FCI, 3.1) in descending order from top to bottom.

The following provides the specifics of how ecoset was created, and common challenges in the creation of large-scale training sets. Importantly, ecoset will be freely available for research and educational purposes at the online platform CodeOcean. Users are asked to refer to the provided license file for usage terms and conditions.

3.2 Methods

3.2.1 Selection of ecoset categories and category images

Each large-scale vision dataset is designed to focuses on specific aspects of vision, and the selection of categories to be included is ultimately guided by these computational goals. As mentioned above, many currently available vision datasets are constructed as engineering challenges and the respective design goals are therefore different from the aims of modeling in computational neuroscience (but see Hebart et al., 2019). In addition, the category selection process is oftentimes highly subjective. For instance, the Open Images Dataset v4 includes 600 object categories (out of a set of 19,794 candidates) that the authors "deemed important [and that showed] a clearly defined spatial extent" (p. 3, Kuznetsova et al., 2018). MS COCO included category candidates from a pre-existing dataset (PASCAL VOC; Everingham et al., 2010), plus a list of visually identifiable objects, and objects that are nameable by 4-8 year old children (Lin et al., 2014). The resulting list of 272 categories was subsequently reduced to 91 based on the authors' own commonness- and usefulness-ratings.

While a small level of subjectivity is inevitable, the selection process of categories for ecoset was specifically designed to follow an objective set of criteria. Central to this is the ranking of all nouns in the English language based on a frequency-concreteness-index (FCI).

Frequency-concreteness-index (FCI)

The aim of ecoset is to mirror basic level visual categories that are of high importance to human observers. This was accomplished by combining two parameters: the frequency with which a noun is used in the English language, and human ratings of the noun's concreteness. Linguistic frequency is used as a proxy for concept importance, whereas a focus on concrete nouns implies that they can be readily visualized. For instance, while the noun 'bird' has a concreteness rating of 5/5, the noun 'democracy' has a rating of 1.78/5. Inclusion in ecoset was dependent on a concreteness rating of ≥ 4.0. Joining these two parameters, we define a frequency-concreteness-index (FCI, formula 3.1), which allowed us to focus on the most common, most concrete nouns of the English language during the selection process.

Our frequency estimates are based on SUBTLEX US, a corpus of 51 million words appearing in American English film and TV subtitles (Brysbaert & New, 2009). The concreteness judgments are based on publicly available data from an online experiment in which 4,000 subjects rated 40,000 words with regard to their concreteness on a 5-level Likert-scale via the online platform Amazon Mechanical Turk (M. Brysbaert et al., 2014). Frequency estimates and concreteness ratings were each standardized, so the FCI has a meaningful range: "0" indicates minimum, and "1" maximum FCI (formula 3.1). We computed the FCI for all words intersecting ~~the BNC (BNC, 2014; Leech2014)~~ SUBTLEX US (Brysbaert & New, 2009) and the concreteness ratings mentioned above (M. Brysbaert et al., 2014) and further processed the 3,500 words with the highest FCI rating.

$$FCI = \frac{1}{2}\left(\frac{word\ frequency}{max(word\ frequency)}\right) + \frac{1}{2}\left(\frac{concreteness\ rating}{5}\right) \tag{3.1}$$

Inclusion and exclusion criteria for ecoset categories

All categories present in ecoset were classified as basic-level by our team. It should be noted that the definition of basic level categories is a matter of an ongoing scientific debate, and basic-level judgments can vary across individuals (Markman & Wisniewski, 1997; J. Tanaka & Taylor, 1991). Because of its inherently subjective nature, the classification of nouns that constitute basic-level categories was performed repeatedly to ensure consistency across the whole set. A list of all 565 ecoset categories is provided together with their ~~BNC~~ SUBTLEX US frequency, concreteness rating, FCI, and number of images (table 1, Appendix B).

Category selection was performed using the following criteria: First, nouns describing subordinate and superordinate categories were excluded (examples include "Terrier", or "vehicle"). Moreover, only single-word concepts were included as candidates, excluding separated compound nouns (e.g. "sail boat", "fire truck", etc.) as their own entities, because they are typically part of a basic level category (in the previous example "boat", and "truck", respectively). Third, we excluded nouns describing object parts (e.g. 'hand', 'roof', 'wheel'), as they co-occur in basic-level categories, rendering the image categories ambiguous. Fourth, synonyms were combined into a single category (e.g. "automobile" and "car" are summarized into a single "car" category). The resulting set of nouns describes basic level categories for which the resulting images can be ascribed to a single category (as required for many 1-hot encoded deep learning applications). Subsequent to the selection of candidate basic level categories, we used the corresponding nouns and their equivalents in other languages to download images from three different image-search or -hosting sites (ImageNet, Bing, Flickr).

Images from ImageNet

Most of the images of the final version of ecoset (∼94%) were downloaded from the ImageNet database containing about 14 million natural images (also including the 1.4 million images of ILSVRC 2012; Russakovsky et al., 2015). The category structure of ImageNet is based on the lexical WordNet hierarchy in which sets of words form a distinct semantic concept called a "synset" (Fellbaum, 2012; Miller, 1995). For each candidate ecoset category we used the ImageNet web interface to search for appropriate synsets to be included in ecoset.

The semantic relation between synsets containing a specific search term and ecoset candidate categories is *not* straight-forward. This is why assigning ImageNet synsets to candidate ecoset categories required category-specific choices by our team. For example, searching for "dog" revealed 59 synsets of which only 38 were included in the associated ecoset candidate category as they directly portraited canines. The other synsets found in response to "dog" were not included in this category as they described types of flowers (e.g. "American dog violet", or "dog fennel"), other types of animals (e.g. "dog flea", or "blacktail prairie dog", which is a rodent), and other non-canine categories including tools and types of mushrooms.

After downloading ImageNet images from adequate synsets for each ecoset candidate category, misclassified images were removed manually to ensure an expected error rate of less than 4%. Details are provided in the section about "Technical Validation". In addition to ImageNet, images were sourced via the search engine Bing and the image-hosting site Flickr.

Images from Bing and Flickr

To maximize the number of images per final ecoset category, we used multiple search terms per candidate category for both Bing and Flickr. Search terms included the original noun associated with the candidate category, as well as synonyms in English, and translations in 4 additional languages (French, Spanish, Italian, and German). For example, in order to search for images in the category “lightbulb”, we used the following list: “lightbulb” (original search term), “bulb” (synonym), “ampoule” (French), “bombilla” (Spanish), “lampadina” (Italian), “gluehbirne” (German). After downloading Bing and Flickr images, misclassified images were manually removed (see "Technical Validation" for details). Image search via Bing and Flickr was constrained to images under CC BY NC SA 2.0 license. For the Flickr API, we chose option 1 (NonCommercial shareAlike License), and for the Bing API we chose the option “share”, both referring to CC BY NC SA 2.0.

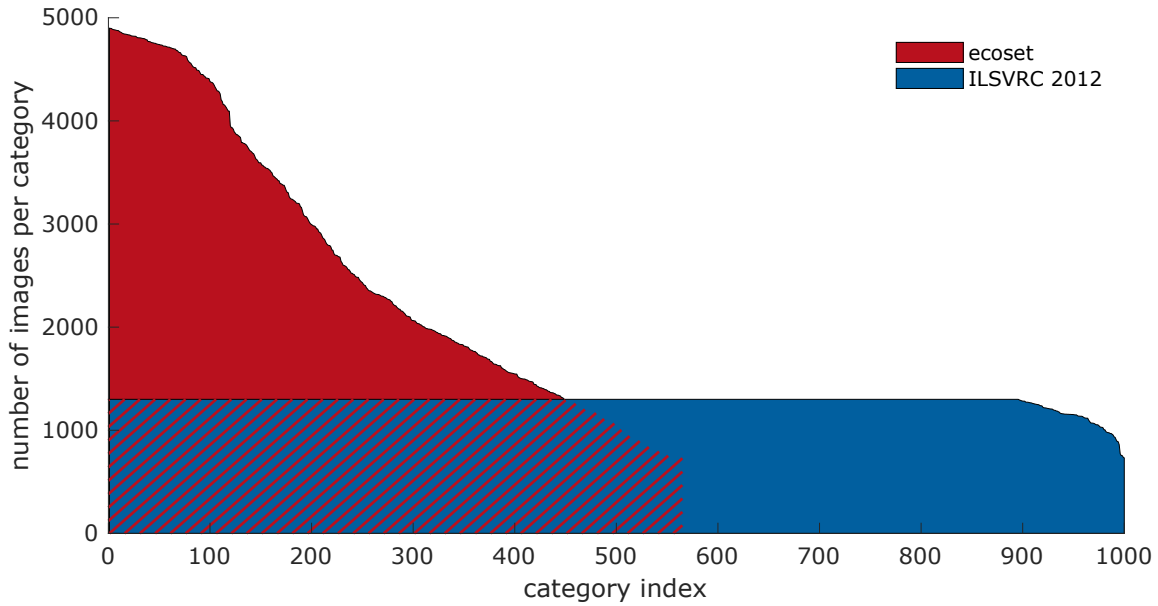


Fig. 3.3 **Ecoset image set statistics.** Ecoset (red) consists of 565 categories each containing between 600 and 4900 images, amounting to a total of 1,444,919 images in the training set. For comparison, there are 1,000 categories in the training set of ILSVRC 2012 object recognition task (blue), each containing between 732 and 1,300 images, amounting to a total of 1,281,167 images.

3.2.2 Dataset statistics

To maximize the number of images per category, we did not exclude small images (shorter side < 128 pixels) from ecoset. However, the number of such images in the final ecoset is small: in the training, validation, and test sets combined only 7.7% of all images have a shorter side smaller than 128 pixels. For an overview of the sizes of the images in ecoset, see Fig 3.4.

The final 565 ecoset categories can be associated with the following 5 super-ordinate binary category distinctions (Fig 3.5): natural (221 $\sim$ 40%) vs. artificial (334 $\sim$ 60%), animate (106 $\sim$ 19%) vs. inanimate (459 $\sim$ 81%), edibles (133 $\sim$ 24%) vs. non-edibles (432 $\sim$ 76%), tools (174 $\sim$ 31%) vs. non-tools (391 $\sim$ 69%), scenes (10 $\sim$ 2%) vs. non-scenes (555 $\sim$ 98%). For a list of all ecoset categories, please see table 1 (Appendix B). Some of the images of ecoset used to train the models described below might contain nudity. For the final version of ecoset to be freely available to the community, we identified and removed explicit content from ecoset using a NSFW-detection software from Yahoo (https://github.com/mdietrichstein/tensorflow-open_nsfw). Note that only 118 of a total of >1.5 million images had to be removed to yield a nudity-free version of ecoset.

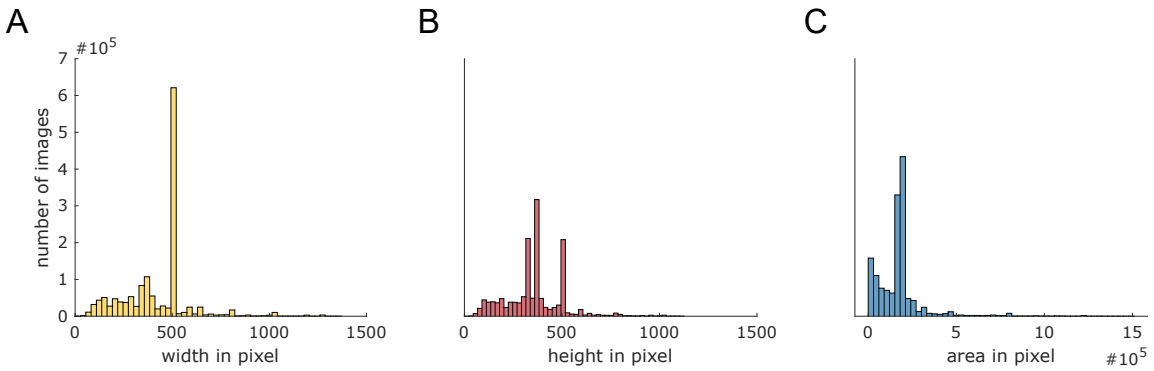


Fig. 3.4 **Distribution of ecoset image parameters.** Distribution of width (A), height (B), and area (C) across ecoset images. For illustrative purposes we discarded outliers ($x < \mu - 1.96\delta$ or $x > \mu + 1.96\delta$) w.r.t. width, height, or area.

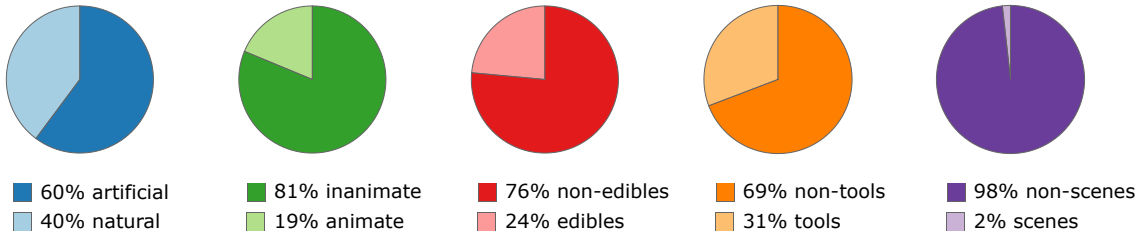


Fig. 3.5 **Membership of ecoset categories in super-ordinate categories.** The 565 basic level ecoset categories can be associated with 5 binary super-ordinate categories: natural (221 $\sim$ 40%) vs. artificial (334 $\sim$ 60%), animate (106 $\sim$ 19%) vs. inanimate (459 $\sim$ 81%), edibles (133 $\sim$ 24%) vs. non-edibles (432 $\sim$ 76%), tools (174 $\sim$ 31%) vs. non-tools (391 $\sim$ 69%), scenes (10 $\sim$ 2%) vs. non-scenes (555 $\sim$ 98%).

3.3 Data Records

All our data will be accessible as a CodeOcean capsule. The data repository contains three main directories in which the complete ecoset image set resides. The ecoset directory contains separate subdirectories for training, validation, and testing that are in turn comprised of one subdirectory per category. All images that constitute ecoset are available in jpg-format, and the entire ecoset is available in a single zip file from which single categories can be easily extracted without extracting the entire image set.

3.4 Technical Validation

One problem of obtaining images from multiple sources (here ImageNet, Bing, and Flickr) is that in addition to some images being plain duplicates, an image might re-occur in the set of a candidate category with a slightly different aspect-ratio, in a different resolution, or it might include a frame, etc. In short, the very same image might have been manipulated in some minor ways and duplicates might have thus found their way into a given ecoset candidate category. To eliminate all duplicates in ecoset, we applied the following PCA-based duplicate removal technique. As a result, the >1.5 million images in ecoset are unique.

Note that we used this PCA-based duplicate removal technique at two different stages of the image selection process (Fig 3.1): first, for all images downloaded from Bing and Flickr to reduce the number of images to be manually inspected with regard to misclassifications. The PCA-based cleaning step had to be used a second time after images from ImageNet, Bing, and Flickr were merged, as we could not exclude that the same image had been retrieved from multiple sources. The following steps were performed for each category separately.

First, we cropped the center square of each image, resized it to 128x128 pixels, and performed a PCA preserving 90% of the variance across all images of that category. To estimate image similarity we computed the pairwise correlation between images projected into PC space. On the basis of 10 exemplary categories, we established a cut-off value above which a pair of images was labeled as duplicates (Pearson $r > 0.975$). If multiple duplicates per category instance existed, we only included the one with the largest resolution in the ecoset candidate category and discarded all others.

A visual inspection of image samples taken from groups of ImageNet synsets belonging to the same ecoset candidate category revealed that misclassification rates for some categories were higher than 10%. This can severely limit the generalization performance of deep neural networks, as random labels increase the chances that the networks memorize rather than generalize (Morcos et al., 2018). We therefore performed a manual sampling and cleaning step to ensure that the expected error in the category label across all categories is $< 4\%$. For the images from ImageNet (93.5% of all images) we visually inspected 100 randomly sampled images from each candidate ecoset category. If more than 4 of those 100 images were found to be misclassifications, the whole category was cleaned manually. Otherwise, all images were included in the associated ecoset category. All images downloaded via Bing and Flickr (5.1% or 76,819 images, and 1.4% or 20,560 images, respectively) were visually inspected and misclassified exemplars were removed.

3.5 Limitations of ecoset

Ecoset was created with the goal in mind to reflect human visual experiences. As a proxy for visual importance we used word frequency in the English language (estimated using American television and film subtitles, SUBTLEX US Brysbaert and New, 2009) and combined it with concreteness ratings from human observers to guide the category selection process. This approach resulted in a set of 565 categories which can be easily visualized and which appear often in spoken English.

However, the selection process of ecoset categories and images could be honed, e.g. by estimating visual importance in another way. For example, to use a very similar approach to ours, frequency estimates from additional spoken or written text corpora from different languages could be combined with more comprehensive concreteness ratings. To go beyond relying on text corpora, video footage obtained from cameras mounted on young human participants might provide a better estimate of what is visually important to humans. Labeling the objects occurring in single video frames might reflect more directly what input the human visual system relies on during phases of visual object category learning.

3.6 Usage Notes

We created ecoset to provide a large-scale image set specifically designed for computational neuroscience. To facilitate the usage of this ecologically more valid image set, ecoset will be made freely available for research and educational purposes at CodeOcean. Users are asked to refer to the provided license file for usage terms and conditions.

Ecoset comes pre-split into training, validation and testing data. Each one of the three subsets contains one folder per ecoset category. This setup is the most common format for various deep learning frameworks and therefore allows for a quick integration into existing pipelines. Details on the characteristics of ecoset categories and images are provided in the dataset statistics section.

After the introduction of ecoset in this Chapter, in the next Chapter I will investigate whether training DNNs on ecoset instead of ILSVRC 2012 may help to better explain cortical representations in human IT.

Chapter 4

A brain-inspired DNN (vNet) and an ecologically more valid visual diet for deep learning (ecoset) yields better models of human high-level visual cortex

Abstract

Deep neural networks have revolutionized computer vision applications and represent the current best models of visual information processing in the primate brain. For the latter, computational neuroscientists commonly rely on pre-trained networks whose architectures were engineered for high performance on computer vision datasets. Moving beyond this common practice, we here report progress in modeling human higher level visual cortex by using biologically inspired network architectures and ecologically more realistic input statistics. As a major stepping stone, we use *ecoset*, a database of >1.5 million images from 565 image categories specifically selected to better capture the distribution of ecologically relevant object categories. Our experiments are based on a deep neural network architecture, *vNet*, which closely mimics the progressive increase in receptive field sizes along the human ventral stream, as estimated by human population receptive field mapping. We show that training vNet on ecoset leads to significant improvements in predicting representations in human inferotemporal cortex (IT). The trained networks improve upon the previous state of the art (Alexnet, VGG-19, and Densenet-169) while being considerably less complex. This is shown for two separate fMRI datasets covering a large variety of 1292 visual stimuli. Together, these results indicate that computational visual neuroscience will benefit from moving beyond computer vision models. Importantly, *ecoset* and *vNet* are both freely available for usage and further developments within the community.

4.1 Introduction

Recently, deep neural networks (DNNs) have revolutionized computer vision and are currently the best models of the visual cortex (Kietzmann, McClure, et al., 2019; Kriegeskorte, 2015; Richards et al., 2019; Serre, 2019; Yamins & DiCarlo, 2016a). To allow DNNs to closely mirror cortical representations, their features need to be shaped through training on a complex object recognition task. However, training DNNs on such a task is non-trivial and requires large computational costs. Hence, most studies using DNNs as models for the primate visual cortex have relied on single, off-the-shelf instances of the same DNN

architecture (mostly AlexNet; A. Krizhevsky et al., 2012) trained on one specific task, namely the object recognition task of the ImageNet Large Scale Visual Recognition Competition (ILSVRC) 2012 (Abbasi-Asl et al., 2018; Agrawal et al., 2014; Cadieu et al., 2014; Cichy et al., 2016; Devereux et al., 2018; Eickenberg et al., 2017; Güçlü & van Gerven, 2015; Hong et al., 2016; Horikawa & Kamitani, 2017a, 2017b; Kalfas et al., 2017; Khaligh-Razavi & Kriegeskorte, 2014; Russakovsky et al., 2015). This poses two main problems in computational neuroscience both suggesting a possible lack of generality of insights gained: 1. Relying on single network instances, and 2. training on an object recognition task designed for machine learning applications.

First, the ability of a single DNN instance to predict cortical representations might depend on its initial set of weights (Chapter 2). Thus, using a single DNN instance might not be able to reveal the full picture portraying a given architecture's ability to predict cortical representations. Instead, multiple instances of this DNN architecture might be necessary to reveal the variability of its fit to a given set of data. This is why we base our results presented here on groups of networks of the same architecture.

Second, most of these DNN models are trained on the very same image set, namely ILSVRC 2012, designed for machine learning purposes. Training a DNN on a task dissimilar to the human visual experience, such as the recognition of 120 dog breeds among a total of 1,000 categories, yields networks capable to mimic representations in the human cortex surprisingly closely. In this light, however, it appears plausible that training a DNN to perform a task more closely related to the human visual experience, such as recognizing objects most frequently encountered in daily human life, might yield network internal representations even better able to predict cortical representations. Hence, we tested whether training DNNs on ecoset, the first large-scale image set created with computational neuroscience goals in mind, yields network internal representations better able to predict cortical representations.

In general, the internal representations of DNNs can be altered through network structure, input statistics, functional objective, and learning algorithm (Kietzmann, McClure, et al., 2019). We here target the the first two factors to directly manipulate network internal representations. We consider the network architecture by creating a brain-inspired 10-layer DNN architecture, vNet, mimicking the progressive increase in foveal receptive field sizes along multiple areas of the human ventral stream (V1, V2, V3, hV4, LO, TO, pFUS, and mFUS), as estimated via population receptive field mapping (Grill-Spector et al., 2017; Wandell and Winawer, 2016; Fig 4.3). Further, considering the second factor, input statistics, we train vNet on an ecologically relevant set of images, ecoset, containing >1.5 million images from 565 basic level categories and thus constituting the first set of images specifically designed for computational neuroscience. The selection of categories and images of ecoset

was aimed at approximating the human visual experience by including only those categories that i) appeared often in the English language, ii) were perceived as concrete by human observers (M. Brysbaert et al., 2014), and iii) reached a minimum number of images per category of 700 (for details, see Chapter 3).

To probe the ability of a given network to mirror cortical representations we use two independent fMRI experiments containing varying numbers of stimuli (92 vs. 1,200) and varying numbers of participants (15 vs. 5; Cichy et al., 2014; Horikawa and Kamitani, 2017a). To compare representations of visual objects between DNNs and humans, we use representational similarity analysis (RSA; Kriegeskorte et al., 2008). In DNNs we record the activation patterns elicited by the same set of stimuli used in the fMRI experiments. In both DNNs and human subjects we use pairwise comparisons of stimuli to create representational distance matrices (RDMs) reflecting the representational geometry of a given set of stimuli. As RDMs abstract away from the input modality they allow for a comparison of a given set of stimuli between patterns from DNN units and from fMRI voxels.

In short, we investigate how a brain-inspired neural network architecture (*vNet*) and ecologically plausible input statistics (*ecoset*) may help to better explain representations in the human visual cortex. We first ask whether training vNet on the ecologically relevant ecoset improves its ability to predict cortical representations. For this we compare ecoset-trained vNet instances with the same models trained on ILSVRC 2012 instead. Next, we compare ecoset-trained vNet to other DNN architectures representing state-of-the-art in computer vision, and in computational neuroscience: Alexnet, VGG-19, and Densenet-169 (Huang et al., 2016; A. Krizhevsky et al., 2012; Simonyan & Zisserman, 2015).

4.2 Methods

We here investigate how i) the input statistics of a DNN and ii) the network structure contribute to its ability to mirror cortical representations. We hypothesize that drawing inspiration from ecological and biological aspects of vision systems in primates might help to build better models of the human visual cortex. In this section I first present the image sets used to train the networks presented in this study: ecoset and ILSVRC 2012. Next, I describe the creation of the brain-inspired network vNet. In the remainder of this section I lay out how RSA is used to assess representational similarity between models and cortical data.

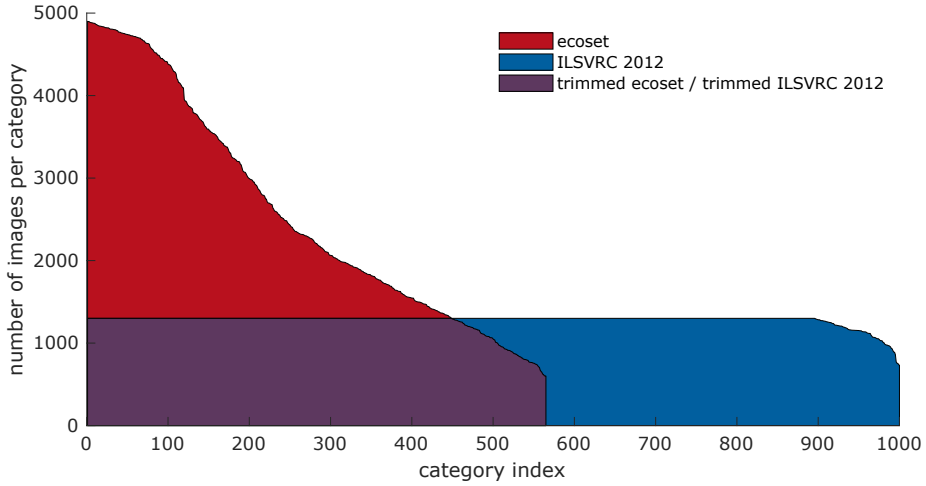


Fig. 4.1 **Ecoset image set statistics.** ecoset (red) consists of 565 categories each containing between 600 and 4900 images, amounting to a total of 1,444,919 images in the training set. For comparison, there are 1,000 categories in the training set of ILSVRC 2012 object recognition task (blue), each containing between 732 and 1,300 images, amounting to a total of 1,281,167 images. For a fair comparison between ecoset and ILSVRC 2012 we created trimmed versions of both image sets (purple) that are identical in the number of categories (565) and are matched regarding category sizes (600 - 1,300).

4.2.1 Image sets for training DNNs

Full ecoset vs. *full* ILSVRC 2012

ILSVRC 2012 was created to allow machine learning researchers to test and compare which DNN architecture best categorizes objects from a set of fine-grained classes. The image set contains a total of 1.3 million images from 1,000 categories, and category size of the training set range from 732-1,300 images (see "ILSVRC 2012" in Fig 4.1). Most DNNs currently used as models for the primate visual cortex have been trained on ILSVRC 2012, highlighting its importance for computational neuroscience. The categories and images of ILSVRC 2012 were selected with machine learning goals in mind, which explains why 120 of the total 1,000 categories are dog breeds. As such, ILSVRC 2012 differs drastically from the human visual experience, which raises the question of whether training on more ecologically relevant stimuli might yield better models of the human visual cortex.

In contrast to ILSVRC 2012, ecoset is an image set created specifically for computational neuroscience as the selection of categories and images has been performed with the goal to closely approximate the human visual experience. As described in detail in Chapter 3 categories and images of ecoset were selected with the help of a frequency-concreteness-index. Word frequency in the English language was used as a proxy for importance in the

(visual) human experience, whereas concreteness ratings by human observers secured that objects can be easily visualized. Ecoset consists of >1.5 million images from 565 basic level categories, and category sizes of the training set ranges from 600-4,900 images.

Trimmed ecoset vs. *trimmed* ILSVRC 2012

ILSVRC 2012 and ecoset differ in the number of categories and the category sizes. In a comparison of networks trained on the full versions of these image sets, the varying number of categories and of category sizes might confound their ability to predict neural data. For a better comparison, we thus created trimmed versions of ILSVRC 2012 and ecoset that are equal with regard to both number of categories (565) and the category sizes (see trimmed image sets in purple in Fig4.1). For this, we selected all 565 categories from ecoset and a set of 565 randomly chosen categories from ILSVRC 2012. Next, to hold the number of images per category equal across trimmed image sets, we selected images in the following way. The 565 categories of trimmed ecoset and of trimmed ILSVRC 2012 were ordered according to category size and were paired across images sets. For each category from either ecoset or ILSVRC 2012 that contained more images than its counterpart from the other image set, we randomly selected a number of images equal to the number of images in the smaller category. In this way both *trimmed* ecoset and *trimmed* ILSVRC contain 565 categories and follow the same distribution of category sizes with minimally 600 to maximally 1,300 images per category (see purple area in Fig4.1).

4.2.2 DNN architecture

Throughout the current study we used a ten-layer architecture, "vNet", created with the goal to match the foveal receptive field sizes as found in areas of the human ventral stream, as estimated via population receptive field mapping (Grill-Spector et al., 2017; Wandell & Winawer, 2016). The first 8 Layers of vNet are matched with the following cortical regions: L1 - V1, L2 - V2, L3 - V3, L4 - hV4, L5 - LO, L6 - TO, L7 - pFUS, and L8 - mFUS; for layer 9 and 10 the receptive field sizes were linearly extrapolated from the previous layers. The number of feature maps used per layer are [128, 128, 256, 256, 512, 512, 1024, 1024, 2048, 2048], the kernel sizes are [7, 7, 5, 5, 3, 3, 3, 3, 1, 1], and the width and height dimensions of layers are [128, 128, 64, 32, 32, 32, 16, 8, 4, 2] (Fig 4.2). The following transformations were applied to each layer in the network: 2x2 max-pooling (optional), convolution, dropout (with a dropout probability of 0.2), ReLU and group normalization (Wu & He, 2018). These operations were consistently applied to all layers, with the exception of max-pooling. Max-pooling was applied before all convolutions with the exception of layers

[1, 2, 5, 6]. After the penultimate layer we applied global average pooling followed by a fully-connected linear readout.

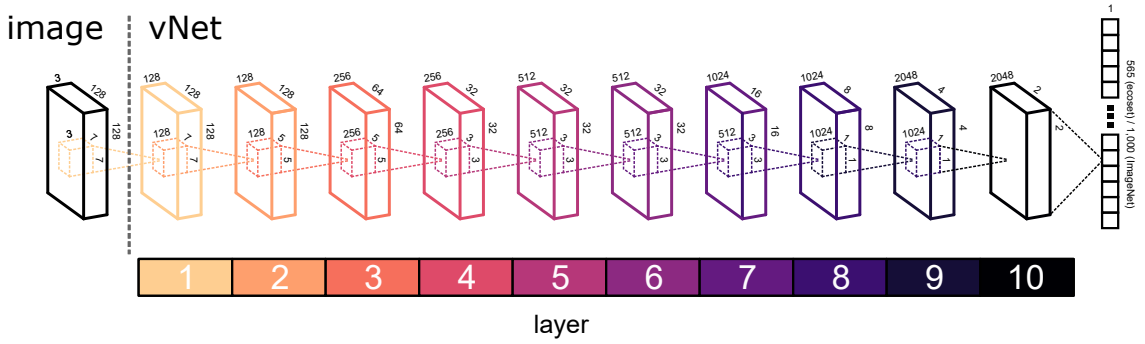


Fig. 4.2 **vNet architecture.** vNet is a 10-layer architecture created with the goal in mind to match the foveal receptive field sizes as found in areas of the human ventral stream (for details see Fig 4.3). All 10 layer contain the following building blocks: 2x2 max-pooling, convolution, dropout (with a dropout probability of 0.2), ReLU (with the exception of the input and layers [1, 2, 5, 6], where no max-pooling was applied).

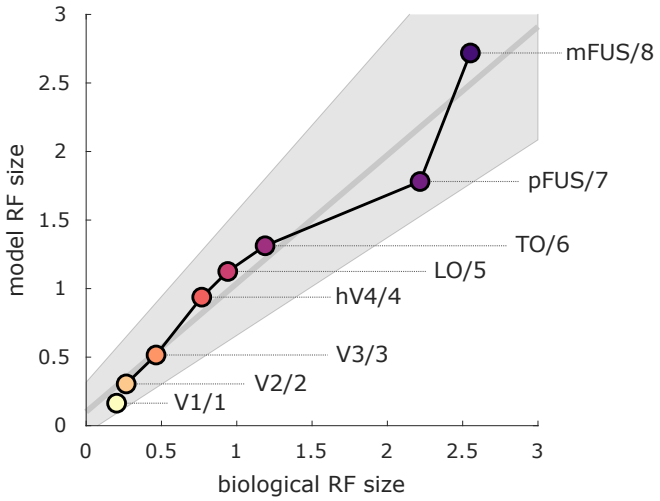


Fig. 4.3 **Receptive field sizes of vNet adjusted to mimic primate visual cortex.** vNet is created with the goal to mimic the progressive increase in receptive field sizes along the human ventral stream, as estimated by human population receptive field mapping. The first 8 layers of vNet were matched with the available biological receptive field sizes (L1 - V1, L2 - V2, L3 - V3, L4 - hV4, L5 - LO, L6 - TO, L7 - pFUS, and L8 - mFUS; Grill-Spector et al., 2017; Wandell and Winawer, 2016), whereas for layer 9-10 the target receptive field sizes were determined by linear extrapolation based on the previous layers.

4.2.3 DNN training

We trained groups of vNet instances only differing in their initial set of weights for both ecoset and ILSVRC 2012. First, we trained 10 instances of vNet on *full* ecoset and, additionally, 10 instances on *full* ILSVRC 2012 that were initialized with the same 10 sets of weights. In other words, we obtained 10 pairs of DNN instances whose members each start training using the exact same initial set of weights, but were trained on different image sets (*full* ecoset vs. *full* ILSVRC 2012). Second, we initialized another 20 DNN instances in the same way, but trained 10 on *trimmed* ecoset and their 10 identically initialized counterparts on *trimmed* ILSVRC 2012.

Image preprocessing

When DNN architectures used as models for the visual cortex are trained on ILSVRC 2012 an input image size of around 224x224 pixels is commonly used (Agrawal et al., 2014; Cadieu et al., 2014; Cichy et al., 2016; Devereux et al., 2018; Eickenberg et al., 2017; Güçlü & van Gerven, 2015, 2017; Hong et al., 2016; Horikawa & Kamitani, 2017a, 2017b; Khaligh-Razavi & Kriegeskorte, 2014). However, our preliminary investigations suggested that reducing the input image size to e.g. 128x128 neither significantly decreases task performance, nor does it impair the DNNs' ability to predict neural data. Thus during preprocessing we cropped the squared center area and down- or upsampled images to a size of 128x128 pixels. Utilizing this relatively small image size means that the overall scale of the DNNs and the time required to train them is decreased.

Training parameters

For each network (10 instances for each of the 4 training sets: (*full* / *trimmed*) ecoset, (*full* / *trimmed*) ILSVRC2012) we initialized the weights using the MSRA scheme (He et al., 2014), and set additional parameters as follows: initial learning rate: 0.02 (using Adam to update the learning rate throughout training; Kingma and Ba, 2014), weight decay: 10^{-5}, dropout probabality: 0.2 (Srivastava et al., 2014). Each of the models was trained with a batch size of 256 and for 80 epochs. Image preprocessing, and training and validation of the models was implemented in TensorFlow 1.10.0 (Abadi et al., 2016).

Accounting for a skewed distribution of category sizes

An imbalanced training set, i.e. varying numbers of images per category, can be problematic with regard to DNN training. More specifically, when the number of images vary across

categories (ILSVRC 2012: [732, 1300], ecoset: [600, 4900], trimmed ecoset or trimmed ILSVRC 2012: [600, 1300]), the size of a category might determine its impact on the formation of DNN features. To account for the varying category sizes in (trimmed) ecoset and ILSVRC 2012 we considered two main strategies. First, when oversampling images until each category contains the same number of images as the largest category, all but the largest category contain duplicates. Performing DNN training using this oversampling approach to account for unequal category sizes, only yielded very low validation accuracies.

We thus used another way of compensating for the effect of the varying category size: a weighted loss function. Here a category-size-specific factor is applied to each image when the overall loss of a mini-batch is computed. Specifically, we multiplied the loss of an image with the inverse of the number of images of the category of said image. In order to obviate the need to adjust learning rate schemes across models training with and without a weighted loss function, we normalized the weights to have a mean of 1. We applied the weighted loss function described above to compute the error of each mini-batch during training.

Ecoset, vNet, and code to extract network activations are freely available

To allow researchers to test their own hypotheses about ecologically more valid input statistics, in addition to the ecoset dataset we will soon also provide the DNNs presented here (implemented in TensorFlow 1.10.0; Abadi et al., 2016) at the online platform CodeOcean. For a rapid adoption by the community, we further provide code that can readily be run online at CodeOcean to obtain DNN activation patterns from all layers in response to arbitrary input images without knowledge of deep learning and without the need to install software on a local machine.

4.2.4 fMRI data sets

We investigated the effect of the image set used for training DNNs (ecoset vs. ILSVRC 2012) on their ability to predict object representations in the human visual cortex in two independent fMRI data sets (Cichy et al., 2014; Horikawa & Kamitani, 2017a).

fMRI data set 1 - 92 stimuli, 15 subjects, (Cichy et al., 2014). The first fMRI data set used in our experiment is from a combined fMRI-MEG study (3T, TR 2 sec., voxel size of functional data 2 mm isotropic) investigating object recognition in 15 healthy subjects (10 female, age: mean $\pm$ std = 25.87 $\pm$ 5.38; Cichy et al., 2014). We did not use the MEG data, but only the fMRI data for our analyses.

In the fMRI experiment each subject performed 10-14 runs, each lasting 384 seconds, during which the 92 stimuli were presented once in random order at 2.9° visual angle. Stimuli

fMRI data set	**Cichy et al. 2014**	**Horikawa et al. 2018**
Number of subjects	15	5
Number of stimuli	92	1,200
Number of presentations per stimulus	10-14	1
Stimulus presentation size (° visual angle)	2.9	12
Magnetic field strength (Tesla)	3	3
Voxel size of functional data (mm isotropic)	2	3
TR (sec.)	2	3

Table 4.1 Main characteristics and recording parameters of the two fMRI data sets analyzed.

were randomly interspersed with 30 baseline trials during which subjects had to indicate with a button press a change in color of the fixation cross to maintain the subjects' attention.

The region of interest IT in dataset 1 was delineated using masks based on "WFU Pickatlas", and "IBASPM116 Atlas" to include bilateral fusiform and inferior temporal cortex. For a more detailed account of the MRI acquisition parameters and the experimental set up, please see Cichy et al., 2014.

fMRI data set 2 - 1,200 stimuli, 5 subjects, (Horikawa & Kamitani, 2017a). The second fMRI data set used for the current experiment stems from an fMRI study (3T, TR 3 sec., voxel size of functional data 3 mm isotropic) investigating perception and imagery of everyday objects in 5 healthy subjects (1 female, age range 23-38 years) in two separate (perception/imagery) experiments (Horikawa & Kamitani, 2017a). We did *not* investigate any data from the imagery experiment. Instead, we exclusively used the data from the perception experiment which consisted of a training and a testing session, whereby we only used data from the training session. The experimental stimuli were taken from ImageNet 2011 (fall release) of which the authors selected 200 object categories.

During the training sessions each of 1,200 stimuli from 150 object categories (8 images per category) was presented once. Each image block lasted for 9 seconds during which the stimulus was presented at a rate of 2Hz and at 12° visual angle. To maintain attention subjects performed a one-back repetition task, whereby repetitions occurred pseudo-randomly and during every 11th stimulus block on average. For a more detailed description of the fMRI acquisition parameters and the experimental setup, please see Horikawa and Kamitani, 2017a.

What we refer to as the cortical region "IT" in the current study, is referred to as higher visual cortex or "HVC" in the original paper (Horikawa & Kamitani, 2017a). HVC (and thus our IT) is defined as a region manually delineated on the flattened surface comprising LOC, FFA, and PPA, after each of these regions had been identified using separate localizer experiments. For a more detailed account of the ROI definition please see Horikawa and

Kamitani, 2017a and for a comparison of the most important parameters of the two datasets, please see table 4.1.

4.2.5 Predicting representations of visual objects in human IT

To compare object representations between DNNs and human IT, we are using representational similarity analysis (RSA). At the core of the RSA framework stands the representational dissimilarity matrix (RDM) which reflects the representational geometry of a set of objects, i.e. how instances of various object categories are grouped and separated by the units of a given layer (Kriegeskorte & Kievit, 2013). Each cell of an RDM describes the similarity between the activation patterns elicited in response to a pair of stimuli. In this way RSA avoids the correspondency problem and thus allows for comparisons between individuals or species, between DNNs, and between DNNs and individuals (Kriegeskorte et al., 2008).

First, for both fMRI studies described above (Cichy et al., 2014: 92 stimuli, 15 subjects; Horikawa and Kamitani, 2017a: 1,200 stimuli, 5 subjects), we created RDMs based on the cortical activation patterns in inferior temporal cortex (IT) to form a single RDM for each subject for the ROI "IT". To allow a comparison between representations in DNNs and the human visual cortex, we extracted network activation patterns to the same set of images used as stimuli in the fMRI experiments. To obtain an RDM reflecting the representational geometry of the set of stimuli in the DNN, pairwise dissimilarities were estimated using correlation distance. We performed these steps for each DNN-instance and -layer separately (see Fig 4.4, left panel). Note that no additional fitting procedure was applied to map the network RDMs to human RDMs (e.g. reweighting and remixing as performed in Khaligh-Razavi and Kriegeskorte, 2014 or in Storrs et al., 2020). This is done to probe the unaltered alignment between DNNs and human IT: the best model of IT will not only capture the mere presence of certain features but will instead mirror the full distribution of feature selectivity as found in the brain.

To compare the representational geometry between DNNs and human IT, we first correlated each subject-specific fMRI-RDM with DNN-RDMs for each DNN-instance and layer separately using Spearman's ρ and then averaged across subjects. In this way, we obtain a distinct similarity estimate per layer of each DNN-instance trained on ecoset or ILSVRC 2012 while including subject-specific information in our analysis. For an illustration of the main steps of the analysis, please see Fig 4.4. For each layer, we thus obtain 10 similarity estimates from ecoset-trained vNet instances, and 10 from their ILSVRC 2012-trained counterparts. After training we analyze the 10 pairwise differences between networks trained with ecoset and ILSVRC 2012 (one pair per random seed determining the initial weights). These pairwise differences indicate whether training a DNN-architecture with an identical initial

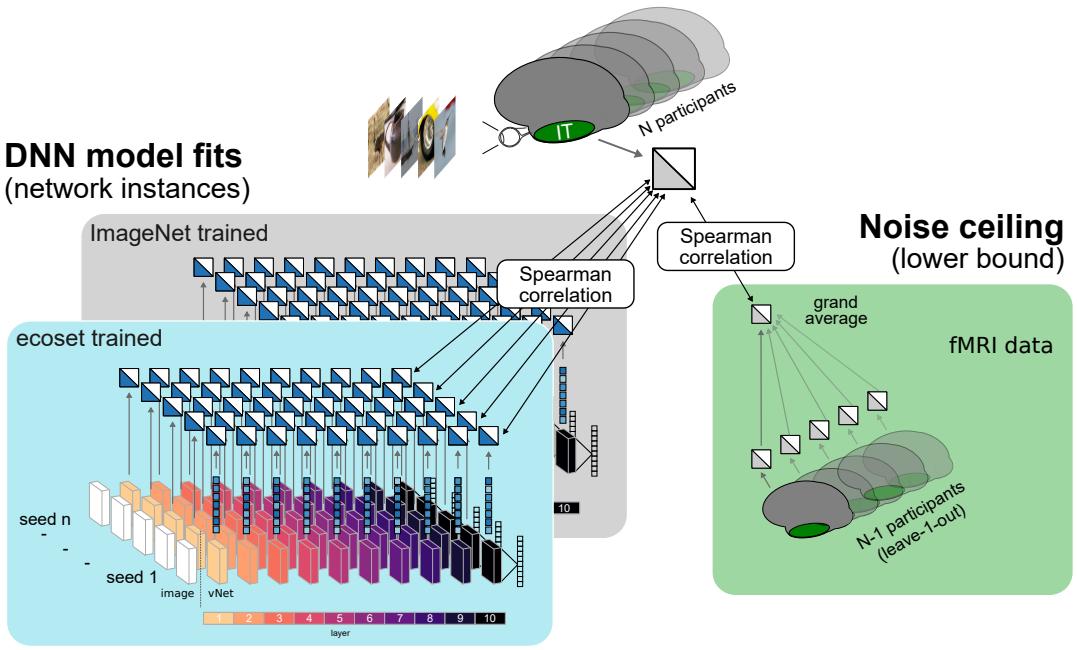


Fig. 4.4 **Predict representations of visual objects in human IT. Right panel:** fMRI activation patterns were recorded in response to a set of stimuli from one region of interest, inferior temporal cortex (IT, dark green), in two different fMRI studies (Cichy et al., 2014, or Horikawa and Kamitani, 2017a). For the ROI (IT, dark green) we computed pairwise correlation distances between activation patterns in response to single images to form a fMRI-based representational dissimilarity distance matrix (RDM) and averaged the RDMs across subjects. **Left panel:** Similarly, for each of 10 DNN instances trained on an image set (ecoset, ILSVRC 2012) we extracted activations to the same sets of stimuli used in the fMRI experiments (Cichy et al., 2014; Horikawa & Kamitani, 2017a). As for the fMRI data, we used correlation distance to compute the pairwise distances between activation pattern in response to single stimuli to create a DNN-based RDM and performed this procedure for each layer separately. Finally, we correlated the fMRI-RDM of a single subject (central RDM above the light blue, grey and green boxes) with the RDM of a DNN-instance and -layer (left panel) to obtain a distribution of correlation values for each layer and subject. Next we averaged across subject to obtain a distribution of correlation values for each layer across DNN instances. To test for the difference in the ability to explain cortical representations between training vNet on ecoset vs. ILSVRC 2012, we independently performed the sequence of steps shown on the left for groups of 10 instances trained on the respective image set.

set of weights on ecoset better explains object representations in human IT than when the same architecture and initialization is trained on ILSVRC 2012.

Next, we test whether the distribution of differences of these similarity estimates between pairs of DNN instances is significantly different from zero, using Wilcoxon signed-rank test.

In order to correct for multiple comparison introduced by testing for differences in each of the 10 layers of vNet separately, we take a conservative approach using Bonferroni correction.

Last, to compare ecoset-trained vNet with state-of-the-art models from computer vision and computational neuroscience (AlexNet, VGG19, and DenseNet169), we extracted RDMs from these models based on the stimuli used in both fMRI experiments using correlation distance to compute pairwise comparisons. From each of these networks, and for each fMRI datset separately, we used the layer with the best prediction for IT representations to compare the fit to the one achieved by the best layer of the 10 instances of ecoset-trained vNet.

4.3 Results

4.3.1 vNet task performance across epochs

All vNet instances were trained (on both *full* ecoset and *full* ILSVRC 2012) for 80 epochs. Final mean performances on *full* ecoset are at 80.7 % during training and 64.7 % during validation, and on *full* ILSVRC 2012 at 84.3 % during training and 59.3 % during validation (Fig 4.5). The variance of validation and also of training performances across training seeds lies within a few percent points (e.g. at epoch 80: $\sigma^2_{ecosetTrain} = 0.029; \sigma^2_{ecosetTest} = 0.059; \sigma^2_{ILSVRC2012Train} = 0.023; \sigma^2_{ILSVRC2012Test} = 0.056$) indicating that the initial set of weights affects final task performance only minimally. In addition we trained vNet on trimmed versions of ecoset and ILSVRC 2012. Final mean performances on *trimmed* ecoset are at 78.0 % during training and 61.8 % during validation, and on *trimmed* ILSVRC 2012 at 82.5 % during training and 67.9 % during validation (Fig 4.6).

When task performances are compared across DNNs, this is only sensible when the same test is used for all networks. For example, the comparison of test (or validation) performances of two instances of vNet is valid when the same images from the same categories are used for testing (or validation). As the number of categories and the number of images per category are unequal between full ecoset and full ILSVRC 2012, differences in test (or validation) performance can obviously not be interpreted with regard to the ability of the networks to predict cortical representations. For the trimmed image sets the case is less clear as both sets have the same number of categories and are even identical in the distribution of category sizes, rendering them equal with regard to a coarse measure of classification difficulty not considering the identity of classes or images. However, as class and image identities of the test (or validation) set need to be identical to allow for a valid comparison of classification performances, differences in testing (or validation) performance between trimmed ecoset and

trimmed ILSVRC 2012 can also not be interpreted with regard to their potentially different ability to predict cortical representations.

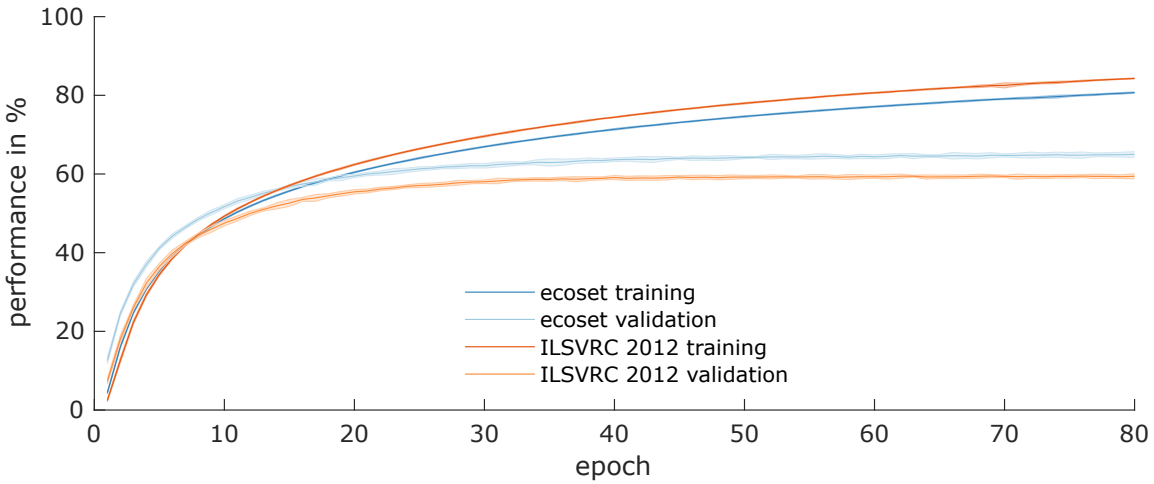


Fig. 4.5 **vNet performance on *full* ecoset and *full* ILSVRC 2012.** Training (ecoset: dark blue, ILSVRC 2012: dark orange) and validation performance (ecoset: light blue, ILSVRC 2012: light orange) across all 80 epochs. Final mean performances on *full* ecoset are at 80.7 % during training and 64.7 % during validation, and on *full* ILSVRC 2012 at 84.3 % during training and 59.3 % during validation. The shaded areas indicate $\mu \pm 1.96 * \sigma$ across training seeds.

4.3.2 *Full* ecoset vs. *full* ILSVRC 2012 trained vNet

We first tested for a difference in the ability of vNet instances trained on full ecoset in comparison to vNet instances with identical initializations, but trained on full ILSVRC 2012 instead. When aiming at explaining representations of 92 stimuli in IT of 15 subjects (fMRI data set 1), we found the 10 DNN-instances trained on ecoset to be significantly better than their 10 counterparts trained on ILSVRC 2012 in the penultimate layer (Wilcoxon, $p_{layer:10} = 0.04$, Bonferroni corrected for all 10 layers; Fig 4.7 A).

Second, we repeated the same test with the same 2 groups of 10 networks, this time aiming at explaining representations of the 1,200 stimuli in IT from an independent fMRI dataset of 5 subjects (fMRI data set 2; Horikawa and Kamitani, 2017a). Here we found that the 10 DNN-instances trained on ecoset were significantly better at explaining IT representations in all 3 of the deepest layers (Wilcoxon, $p_{layer:\{8,9,10\}} = \{0.003, 0.001, 0.001\}$, Bonferroni corrected for all 10 layers, Fig 4.7 B). To summarize our results regarding late stages of the visual stream, training vNet on ecoset might increase its ability to predict cortical representations in human IT.

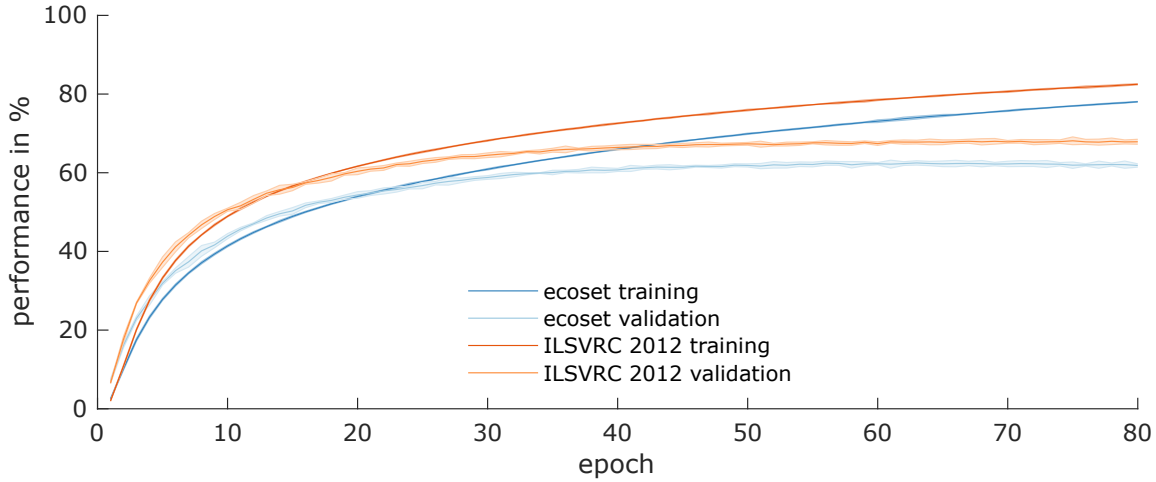


Fig. 4.6 **vNet performance on *trimmed* ecoset and *trimmed* ILSVRC 2012.** Training (ecoset: dark blue, ILSVRC 2012: dark orange) and validation performance (ecoset: light blue, ILSVRC 2012: light orange) across all 80 epochs. Final mean performances on *trimmed* ecoset are at 78.0 % during training and 61.8 % during validation, and on *trimmed* ILSVRC 2012 at 82.5 % during training and 67.9 % during validation. The shaded areas indicate $\mu \pm 1.96 * \sigma$ across training seeds.

4.3.3 *Trimmed* ecoset vs. *trimmed* ILSVRC 2012 trained vNet

The results presented so far are all based on a slightly unequal comparison, because ecoset and ILSVRC 2012 vary in the number of categories and the category sizes. The ability of the associated groups of DNN-instances (ecoset vs. ILSVRC 2012) to explain cortical representations could hence be attributed to these factors rather than to the fact that one image set is targeted towards task-performance benchmarking in machine learning, and the other one towards computational modeling in neuroscience. This is why we repeated the analysis described above for DNN-instances not trained on the *full* versions of ecoset and ILSVRC 2012, but on the *trimmed* variants of these image sets that are identical in the number of categories and the distribution of category size.

In order to exclude the number of categories or the category sizes as confounding factors explaining the difference in the ability of a DNN to predict cortical representations, we trained a group of 10 DNNs on *trimmed* ecoset and on *trimmed* ILSVRC 2012 that are identical with regard to these factors. For training vNet instances on the *trimmed* image sets we used the same hyperparameters as previously described for training on the *full* image sets.

We first tested for a difference in the two groups of DNN-instances in their ability to predict cortical representations of 92 visual objects in 15 subjects (fMRI dataset 1). For this fMRI dataset, we found that ecoset trained DNN-instances are significantly better at explaining IT

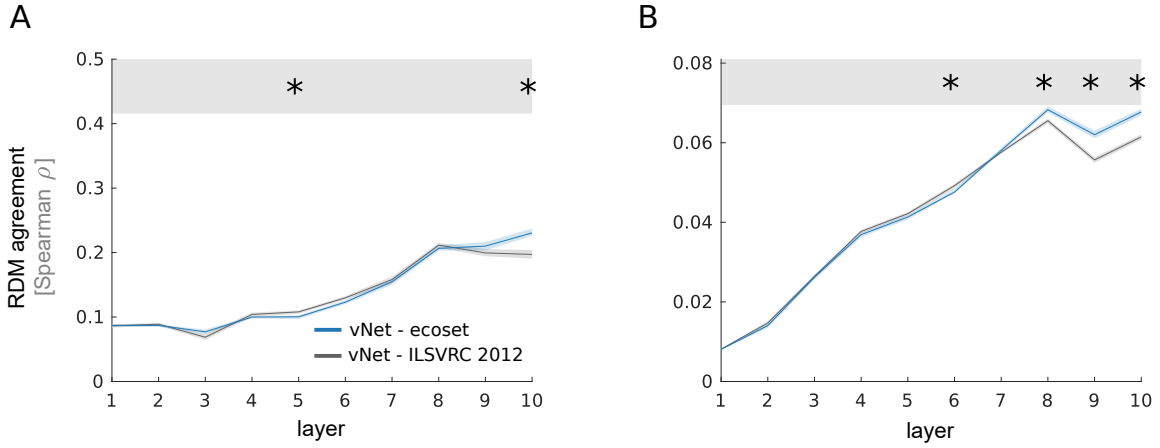


Fig. 4.7 **vNet trained on *full* ecoset and *full* ILSVRC 2012 explains human IT.** (**A**) Cichy et al., 2014: IT representations of 92 visual objects in 15 subjects compared to a group of 10 instances of the vNet-architecture trained on *full* ecoset (blue) or on *full* ILSVRC 2012 (grey). Stars indicate significant differences ($\alpha = 0.5$) in the ability of a group of DNN instances to predict cortical representations after Bonferroni correction for multiple testing for all layers has been applied. In the penultimate layer (10) DNN instances trained on *full* ecoset are significantly better able to explain human IT representations than when the 10 DNN instances with the same sets of initial weights are trained on *full* ILSVRC 2012. (**B**) Same as A, but DNNs explain a different fMRI data set, namely Horikawa and Kamitani, 2017a: IT representations of 1,200 visual objects in 5 subjects. In the 3 deepest layers DNN instances trained on *full* ecoset are significantly better able to explain human IT representations than when the 10 DNN instances with the same sets of initial weights are trained on *full* ILSVRC 2012.

representations in layers 7 and 10. (Wilcoxon, $p_{layer:\{7,8,9,10\}} = \{0.001, 0.001, 0.001, 0.001\}$, Bonferroni corrected for all 10 layers; Fig 4.8 A).

Next, we investigated the same question, this time explaining representations of 1,200 visual object in 5 subjects (fMRI dataset 2). In layers 8, 9, and 10 we found that ecoset-trained vNet instances showed a significantly better ability to predict cortical representations in human IT than their ILSVRC 2012-trained counterparts (Wilcoxon, $p_{layer:\{7,8,9,10\}} = \{0.003, 0.001, 0.002, 0.001\}$, Bonferroni corrected for all 10 layers; Fig 4.8 B).

After excluding the confounding factors "number of categories" and "category size", ecoset-trained vNet still yields significantly better predictions of representations in human IT than their ILSVRC 2012-trained counterparts. This confirms our previous results based on *full* image sets and suggests that ecoset, created with the goal to mimic the human visual experience, might be better suited for training DNNs to predict cortical representations than relying on the computer vision image set ILSVRC 2012. Relating representational consistency (as discussed in 2 for networks trained on CIFAR10 instead of ILSVRC 2012 or

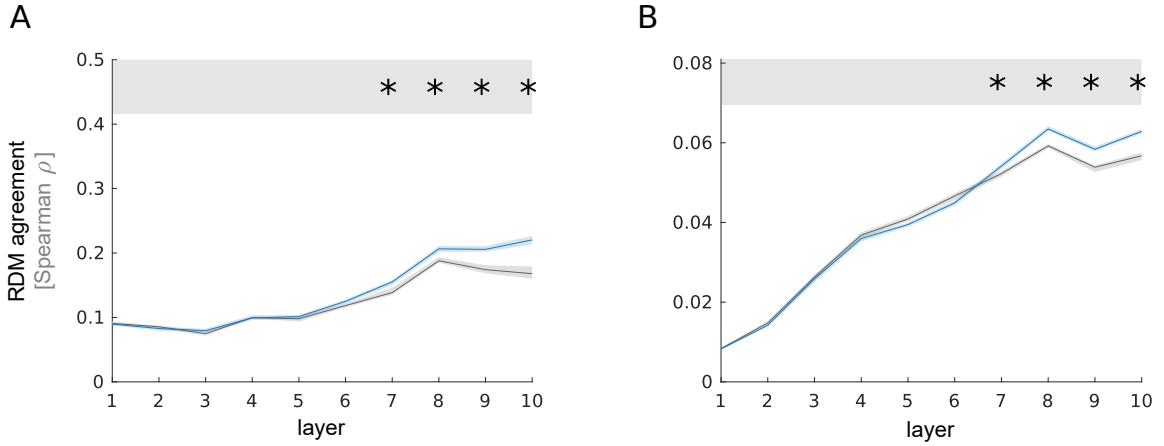


Fig. 4.8 **vNet trained on *trimmed* ecoset and *trimmed* ILSVRC 2012 explains human IT.** (**A**) Cichy et al., 2014: IT representations of 92 visual objects in 15 subjects compared to a group of 10 instances of the vNet-architecture trained on *trimmed* ecoset (blue) or on *trimmed* ILSVRC 2012 (grey). Stars indicate significant differences ($\alpha = 0.5$) in the ability of a group of DNN instances to predict cortical representations after Bonferroni correction for multiple testing for all layers has been applied. In layer 7 and in the penultimate layer (10) DNN instances trained on *trimmed* ecoset are significantly better able to explain human IT representations than when the 10 DNN instances with the same sets of initial weights are trained on *trimmed* ILSVRC 2012. (**B**) Same as A, but DNNs explain a different fMRI data set, namely Horikawa and Kamitani, 2017a: IT representations of 1,200 visual objects in 5 subjects. In the 3 deepest layers DNN instances trained on *trimmed* ecoset are significantly better able to explain human IT representations than when the 10 DNN instances with the same sets of initial weights are trained on *trimmed* ILSVRC 2012.

ecoset) to the ability of a network instance to predict cortical representations, we found no strong evidence for such a relationship (for details, see Fig 5 in Appendix C).

4.3.4 *Full* ecoset trained vNet vs. state-of-the-art computer vision and computational neuroscience models

Our findings so far suggested ecoset to be a better suited image set for training DNNs used as models in computational neuroscience. However, the results are insofar restricted as we have not investigated the ability to predict cortical representations of ecoset-trained vNet in comparison with other architectures. To elucidate whether ecoset-trained vNet might also be able to better predict cortical function than state-of-the-art models in computer vision and computational neuroscience, we tested it against AlexNet, VGG-19, and DenseNet-169 (Huang et al., 2016; A. Krizhevsky et al., 2012; Simonyan & Zisserman, 2015). For this, we took the same approach as in the previous section and extracted activations from these three

models in response to the experimental stimuli of fMRI dataset 1 and 2 to compute RDMs for each layer, reflecting their representational geometry at different depths of the network. When only the layers best predicting the IT representations were compared across DNN architectures, we found that that ecoset-trained vNet may be better able to predict cortical representations than any of the other tested models (not included in bootstrapped confidence interval, $p < 0.05$; Fig 4.9). To obtain an impression of the similarity of RDMs between models and the fMRI data, please see (Fig 4.10). These findings are especially interesting as vNet has fewer parameter than AlexNet and VGG-19, and performs worse on ILSVRC 2012 than VGG-19 and DenseNet-169. In other words, neither the number of trainable parameters, nor the testing (or validation) performance on a commonly used complex object recognition task can reliably indicate whether a DNN architecture might be a good model of cortical representations.

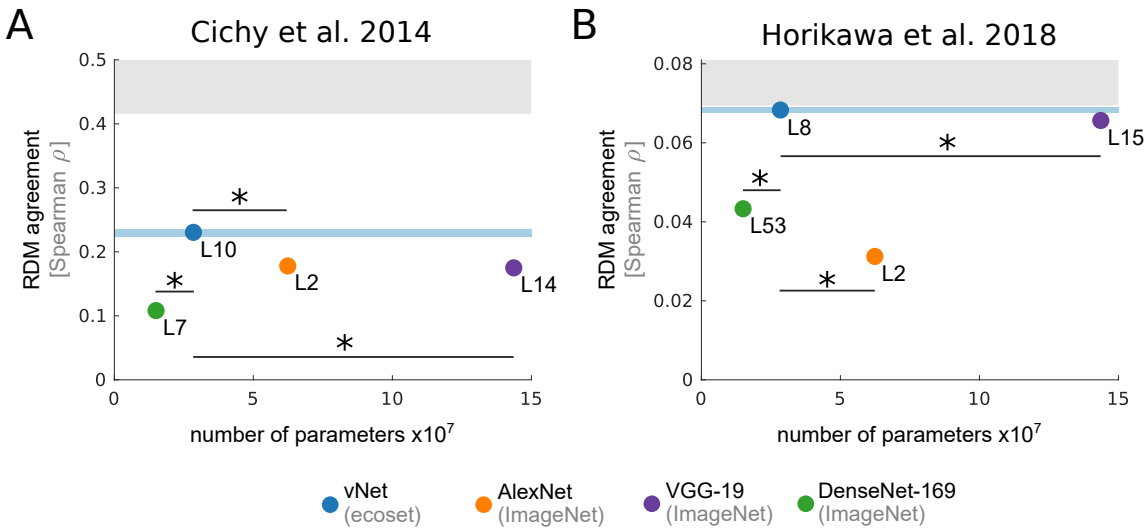


Fig. 4.9 **IT predictability: best layer of ecoset-trained vNet in comparison to best layers of state-of-the-art computer vision and computational neuroscience models.** Across data from two independent fMRI experiments (Cichy et al., 2014; Horikawa & Kamitani, 2017a), ecoset-trained vNet (blue) explains human IT representations significantly better than the best performing layer of VGG-19 (purple), AlexNet (green), or DenseNet-169 (orange). Blue bars represent the 95% CI for vNet, bootstrapped from 10 vNet instances. Stars indicate a significant difference at $\alpha = 0.05$.

4.4 Discussion and conclusion

We investigated whether using a brain-inspired DNN architecture (vNet) trained on ecologically valid input statistics (ecoset) may increase its ability to explain cortical representations.

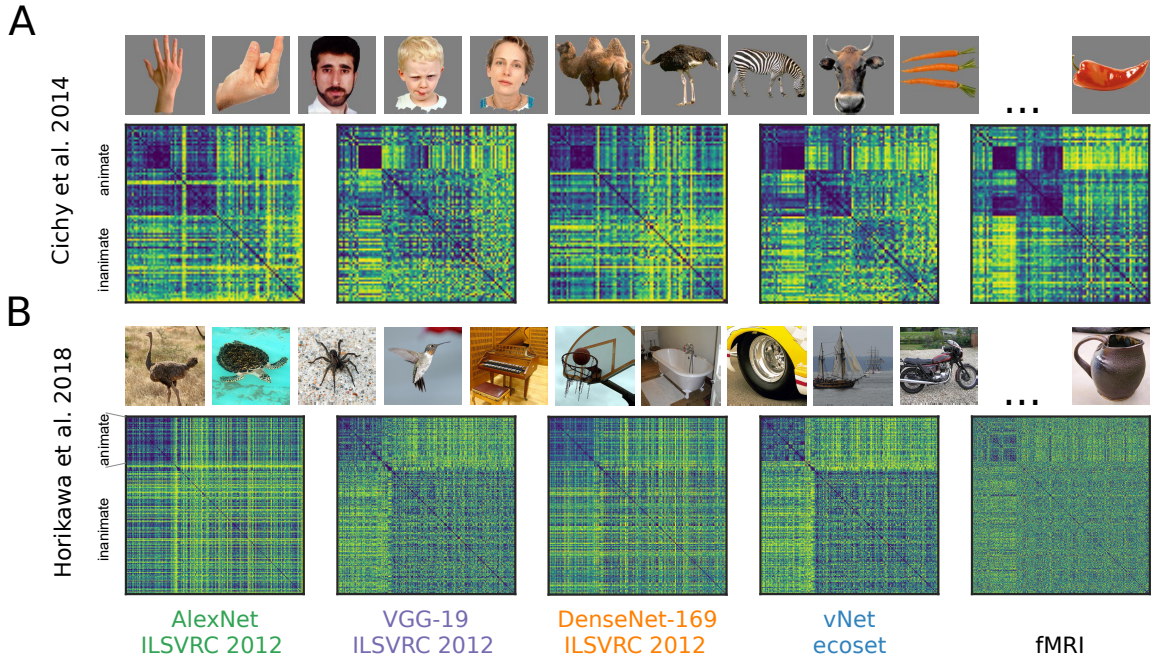


Fig. 4.10 **Comparison of investigated models on the level of representational dissimilarity matrices (RDMs).** Representational dissimilarity matrices (RDMs) for the layers of several deep neural network architectures (column 1-4) that best predict human IT, together with the target RDM computed from fMRI data (column 5).

For this, we trained groups of identical DNNs on the full version of ecoset and ILSVRC 2012, and - to exclude explanations based on differences in the number of categories or distribution of category sizes - also on trimmed versions of these image sets. To quantify the representational alignment between DNNs and the brain, we used RSA (Kriegeskorte et al., 2008). Importantly, we did not perform any reweighting (Khaligh-Razavi & Kriegeskorte, 2014) or linear readout of the DNN activation profiles (Schrimpf et al., 2018) in order to probe the unaltered alignment between the two systems: the best model of IT will not only capture the mere presence of certain features but will instead mirror the full distribution of feature selectivity as found in the brain.

Our first results from both *full* and *trimmed* training sets suggest that training on a more realistic set of images might help to better explain cortical representations in two independent sets of fMRI data (Fig 4.7, Fig 4.8). Second, using the same fMRI data sets, we showed that ecoset-trained vNets may also better explain IT cortical representations than DNNs that are state-of-the-art in computer vision or computational neuroscience and far exceed vNet in the number of layers, number of trainable parameters or their testing (or validation) performance on a complex object recognition task (Fig 4.9).

We have suggested earlier to acknowledge idiosyncrasies of DNNs differing in their initial sets of weights and to treat them similar to experimental subjects (Chapter 2). Any analysis relying on a single DNN instance might yield different results depending on the specific set of weights before training. We thus base our results on multiple instances of vNet to allow for an estimation of the variability of the architecture's ability to predict cortical representations across instances. This is in contrast with most studies using DNNs as models of the human visual cortex, where results are based on a single DNN instance (Agrawal et al., 2014; Cadieu et al., 2014; Cichy et al., 2016; Devereux et al., 2018; Eickenberg et al., 2017; Güçlü & van Gerven, 2015, 2017; Hong et al., 2016; Horikawa & Kamitani, 2017a, 2017b; Khaligh-Razavi & Kriegeskorte, 2014). Here we could demonstrate that by using a group of identical DNNs, only differing in the initial set of weights allows more reliable results with regard to the ability of a specific DNN architecture to predict cortical representations.

Moving beyond this proof of concept, future work will investigate which ecological and biological details are required to further improve DNN predictive performance as all models investigated here including vNet are feedforward models, the most widely used class of models in cognitive computational neuroscience (Kietzmann, McClure, et al., 2019). Among others, these biological details include recurrent (Kar et al., 2019; Spoerer et al., 2019) and skip connections (Huang et al., 2016; Miikkulainen et al., 2017), drawing inspiration from magno-, and parvo-cellular pathways (Mahdisoltani et al., 2018; Mei & Singh, 2018), foveation (Wu & He, 2018; Zhang et al., 2018), other training objectives beyond categorization (McClure & Kriegeskorte, 2016a), and more biologically realistic learning rules (Kriegeskorte & Douglas, 2018; Serre, 2019).

Chapter 5

General discussion

The previous three Chapters describe multiple experiments and the creation of a large-scale image set conducted and created with the goal to investigate the representations of visual objects in the human brain. After exploring representational differences between deep neural network (DNN) instances, and creating an image set enabling DNN training on ecologically more valid input statistics, I have built on these results by investigating how additional ecological and biological inspiration might help to better explain cortical representations of visual objects in the human brain. In the remainder of this thesis I will first summarize the findings of each Chapter separately before discussing what can be learned from the overall results in order to build better models of the human visual cortex.

5.1 Summary of results

5.1.1 Individual differences between deep neural network instances

In a series of experiments we demonstrated how the internal representations across deep neural network instances are affected by the minimal intervention of changing the initial set of weights. To compare representations across DNNs we used representational similarity analysis (RSA), a widely used technique from neuroscience, to compare representations of visual objects across species and between a given computational model and the brain. We defined representational consistency as the shared variance between two representational dissimilarity matrices (RDMs), each reflecting the representational geometry of a set of objects in a given artificial neural network or a biological vision system.

The training on a classification task requires linearly separable classes in high-dimensional activation space at least in the penultimate layer of a given network. In line with this, we found that category-clustering increases across layers. Importantly, this increasing category-

clustering is accompanied by a decreasing representational consistency across layers - a result that generalizes across network architectures and distance measures used to compute the underlying RDMs.

One explanation for why representational consistency decreases while category-clustering increases across DNN layers is that the classification objective does not sufficiently constrain the configuration of category instances in high-dimensional activation space. While classes are largely linearly separable, the configuration of single class exemplars might differ across DNN instances. Confirming this line of thought, our analyses revealed a high consistency of category centroids, suggesting that differences between individual category instances may be the main contributor of the differences observed. Additionally, we could show that an interaction between the non-linearity used at each layer of a given DNN (ReLU) and the distance measure used to compute RDMs might be a reason for the decreasing representational consistency across layers. A simple standardization of the activations before computing pairwise distances for the overall representational geometry may increase consistency across DNNs when cosine and correlation distance are used to compute the underlying RDM. Finally, we demonstrated that regularization through dropout used during training and testing might allow to partially recover consistency across networks.

The individual differences observed across network instances are of importance, because it is common practice in computational neuroscience to use single instances of off-the-shelf DNNs to predict cortical representations in the human visual cortex. As task-performance is indistinguishable across instances and DNN training requires considerable computational costs, multiple instances per DNN architecture are generally not available. However, in the light of our results, we suggest to use multiple instances per DNN architecture to allow for improved generality of the conclusions drawn from comparisons between cortical and DNN representations. The number of networks required for such analyses can be judged by computing consistency across network instances.

5.1.2 Ecologically more valid input statistics for deep neural networks

With the goal to design an ecologically more valid image set for training DNNs used as models of the human visual cortex, we created ecoset, a set of >1.5 million images from 565 basic level categories approximating the human visual experience. As a proxy for visual importance we first created an index using the frequency of nouns in the English language and concreteness ratings by human observers and created a list of candidate categories. After downloading images from various online sources for each category, we applied rigorous cleaning procedures aiming to obtain non-overlapping categories covering those objects that are most frequent in the human visual diet.

The most commonly used image set for training DNNs used as models of the human visual stream is ILSVRC 2012. This image set is designed with machine goals in mind, namely to allow for models to excel at fine-grained categorical divisions, such as the recognition of 120 different dog breeds. We believe that computational models trained to predict representations in the human visual stream should be trained on input statistics similar to the human visual diet. To allow the community to test their own hypotheses about ecologically more valid input statistics ecoset will be freely available to the community for research and educational purposes.

5.1.3 A brain-inspired DNN (vNet) and an ecologically more valid visual diet for deep learning (ecoset) yields better models of human high-level visual cortex

Combining the results from Chapter 2 and 3, in Chapter 4 we investigated whether brain-inspired DNNs (vNet) trained on ecologically relevant input statistics (ecoset) are able to better predict cortical representations in human IT than i) the same brain-inspired DNN architecture, but trained on ILSVRC 2012 instead, and ii) state of the art computer vision models. We could show that training vNet instances on ecoset improves its ability to predict cortical representations when compared to the same networks trained on ILSVRC 2012 instead. Additionally, our analyses further revealed that ecoset-trained vNet may better predict cortical representations than stat-of-the-art computer vision models that have more layers or more free parameters and that reach a higher classification performance on ILSVRC 2012 than vNet.

What unites all the investigations presented in this thesis and what thus constitutes an overarching theme is that additional ecological and biological inspiration might help to build better DNN models of the human visual cortex. First, acknowledging the representational difference between networks with identical architecture and indistinguishable task performance suggests that DNNs should be treated similar to human experimental subjects. Second, using ecologically relevant input statistics for training DNNs might improve their ability to predict representations found in biological vision systems. And last, using a brain-inspired DNN trained on this ecologically plausible set of images, highlights the importance of questions regarding input statistics and architecture for building models of the human visual cortex. In the remainder of this thesis I will discuss how drawing further inspiration from ecology and biology may help build better models of the human visual cortex in the future. Further, I will explore how computational visual neuroscience can profit from a free and open science

infrastructure that encourages transparency and a culture of sharing tests, tasks, models, and data.

5.2 Future models in vision science: more biological plausibility?

Evolution has yielded biological systems that successfully survived and reproduced throughout millennia. The structures and functions underlying these systems have always fascinated humankind in general and scientists more specifically (Bar-Cohen, 2005; Dickinson, 1999). For example, the skin of dolphins allowing them to move at high speeds under water, echolocation by ultrasound in bats, or strong and flexible spider web fibers have all attracted much attention by different scientific disciplines and allowed influential bio-inspired technological achievements (Popescu, 1999).

Learning from biological systems is referred to as *bionics* (Steele, 1960) or *biomimetics* (Schmitt, 1969). Sometimes used synonymously (Vincent, 2001), today the two terms are mostly described as opposites. Whereas bionics is mostly associated with technology used to rehabilitate or even enhance human capabilities (prosthetics, etc.; Dickinson, 1999), biomimetics describes the art of incorporating ideas from biological systems into technological inventions (Bhushan, 2009; Vincent et al., 2006).

The first neural networks were inspired by the structure of the primate visual cortex and thus demonstrate how biological inspiration can help building powerful computer vision applications (Fukushima, 1980; LeCun et al., 1989; Riesenhuber & Poggio, 1999). On the other hand DNNs demonstrate that core object recognition can be performed by abstracting away from biological detail so far that they can work without arguably important features, such as "spikes, ion channels, dendritic nonlinearities, complex microcircuits, neuromodulation, or a host of other physiological phenomena" (p. 115, Tripp, 2018). It thus remains elusive whether additional ecological and biological inspiration from some of the features and function of the human cortex may improve the performance of i) DNNs in computer vision, and ii) DNNs used as models of the human visual cortex.

5.2.1 Biological inspiration for computer vision models

One of the most important features of the brain is its recurrence. In computer vision specifically and in machine learning more generally, the introduction of recurrence in the form of long-short-term-memory (LSTM) or gated recurrent units (GRU) allowed great achievements especially in tasks requiring sequence modeling, such as language translation,

or speech synthesis (Cho et al., 2014; Chung et al., 2014; Donahue et al., 2015; Hochreiter & Schmidhuber, 1997). However, the exact implementation of recurrence in this form only bears little biological plausibility as networks using LSTMs and GRUs do not commonly include lateral and top-down connections that are ubiquitous in biological vision systems. However, first studies combining networks composed of LSTM units and additional top-down connections may, for example, help to improve action recognition (Shi et al., 2017).

In addition, attention is known to play an important role in the human visual system. Although the exact mechanisms may remain far behind the richness and diversity of the processes used in biological vision systems (Serre, 2019), the implementation of attentional features in DNNs yield important results. Using inspiration from our knowledge about attention in the human brain, it has been shown that DNNs are able to perform sensible image captioning (Xu et al., 2015), object localization (Biparva & Tsotsos, 2018), or challenging object detection tasks using feature-based attention (Lindsay & Miller, 2018). Combining attentional features with recurrent connections to improve a network's localization performance (Ba et al., 2014) or gave rise to a distinction between "what"- and "where"-information (Mott et al., 2019), reminiscent of the "what"- and "where"-pathways of the primate visual system (Ungerleider & Haxby, 1994). These examples demonstrate how additional biological inspiration has helped to make progress in computer vision.

To conclude, DNNs used in machine learning and computer vision may profit from biological inspiration. But does drawing additional biological inspiration also help to build better models of the human brain? I will now discuss which factors shaping network internal representations (functional objective, learning algorithm, network structure, and input statistics; Kietzmann, McClure, et al., 2019) may help to mimick cortical representations more closely.

5.2.2 Ecological and biological inspiration for computational neuroscience models

One of the first intuitions about DNNs used in computational modeling of the visual cortex was that the model that better performs a given classification task may also be better able to predict cortical representations (Kriegeskorte, 2015). As deeper feedforward models have been shown to reach higher object recognition performances, this suggested that deeper models might also be better models of the human visual cortex. However, it has been suggested that current state-of-the-art computer vision models with hundreds of layers may exceed the number of stages of the human visual system (Serre, 2019). In line with this, recent results suggest that deeper architectures than AlexNet (e.g. VGG; Simonyan and

Zisserman, 2015) are unable to better predict cortical representations than when e.g. AlexNet with its 7 layers is used (Abbasi-Asl et al., 2018; Kalfas et al., 2017; Storrs & Kriegeskorte, 2019; Storrs et al., 2017). In line with these results, our results from Chapter 4 suggest that models that are deeper or perform better at ILSVRC 2012 than vNet, did not outperform ecoset-trained vNet with regard to their ability to predict cortical representations.

Beyond architecture depth, connectivity matters. More specifically, a biologically more plausible implementation of recurrence than LSTMs or GRUs uses bottom-up, lateral and top-down connections allowing performance increases under difficult conditions (Spoerer et al., 2017), and on two complex object recognition tasks when compared to feedforward models with the same number of trainable parameters (Spoerer et al., 2019). It is this biologically more plausible implementation of recurrence that allowed to explain cortical dynamics during high-level object recognition across multiple regions of the ventral stream (Kietzmann, Spoerer, et al., 2019). In line with this, allowing bottom-up and top-down long-range connections, rDNNs have been shown to outperform feedforward nets with regard to matching the dynamics of the primate visual system (Nayebi et al., 2018). Another study allowing long-range skipping and recurrent connections yielded the best score combining the fit to multiple sets of neural data (Kubilius et al., 2018). And last, a study investigating recurrent DNNs set to identify challenging images revealed that late IT responses were best explained by very deep DNNs or shallow recurrent DNNs, suggesting a functional equivalence between recurrence and additional non-linear transformations as performed in deeper networks (Kar et al., 2019). In sum, these studies demonstrate the importance of recurrence for DNNs used as models for the human visual stream.

Another structural feature important in both biological and artificial vision systems is the receptive field size (Hubel & Wiesel, 1962; Serre, Wolf, et al., 2007). The receptive field size does not only determine which cells or units in one visual region or layer feed into the next one, but also defines the overlap with adjacent cells or units and thus plays a pivotal role in how information is passed through the respective vision system. We explored how adapting receptive field sizes in a DNN to those found in primate visual system influences network internal representations. Our results in Chapter 4 revealed that this inspiration from a biological vision system might improve the ability of a given network to explain cortical representations as our brain-inspired architectures outperformed state-of-the-art computer vision models that either have a larger number of trainable parameters, more layers, and/or reached a higher classification performance on an important computer vision benchmark (ILSVRC 2012).

In addition to this structural inspiration from a biological vision system, we asked how the input statistics influence a network's internal representations and its ability to predict

cortical representations. Our results revealed that mimicking the visual experience of humans might help to build better models of the human visual cortex. This finding is also related to our results from Chapter 2 where we suggest to treat DNNs similar to human experimental subjects: using multiple DNN instances allows for a better generality of results based on a given DNN architecture.

5.3 A way ahead in computational visual neuroscience

We have started to explore which network structures and input statistics may help to explain cortical representations, but it might be equally important to test which objective functions and learning algorithms might help to build better models of the human visual cortex. For this to be a fruitful endeavor, we need to strengthen collaborations across labs, but also across disciplinary boundaries.

ILSVRC offered not only free access to millions of images for DNN training, but also provided the stimulating atmosphere motivating labs around the globe to compete against each other in an object recognition task. This lead to the creation of AlexNet and other architectures that now constitute the most widely used models to predict cortical responses in human IT. To make use of the full potential of the deep learning framework for computational modeling of the human visual cortex, we need to continue our efforts to share tasks, models, data and tests (Kriegeskorte & Douglas, 2018). Just as ILSVRC is freely available to the community, we have created an ecologically more valid image set for DNN training also freely available to the community. In addition, all instances of our brain-inspired architecture vNet trained on ecoset or ILSVRC 2012 will be shared online, so that scientists interested in testing their own hypothesis pertaining to ecoset or vNet, can directly use this image set and our DNNs. Importantly, the availability of multiple instances per architecture, offering to assess the representational variability between networks, allows for a better generalization of the results. As not only training DNNs, but also extracting activations can be non-trivial, we provide code ready to be executed at an online platform (CodeOcean) to extract activations in response from all vNet instances to an arbitrary set of images. In this way we hope to facilitate access to both our task (ecoset) and our models (vNet) for rapid adoption by the community.

What is missing to allow the computational neuroscience community to collaborate faster and more easily is the accessibility of data and tests. To assess the fit of a given model, behavioral or cortical data should be openly available. However, there are many ways in which the fit can be assessed or inference about a given architecture can be drawn. The discussion about the way in which computational models of the humans visual cortex should

be assessed is at the core of cognitive computational modeling and subject to an important current debate (Lage-castellanos et al., 2019).

However, it also appears as a fruitful endeavor to agree on a specific way of assessment allowing to compare many different architectures on the exact same test. First steps have been taken in this direction by BrainScore (Schrimpf et al., 2018) and the Algonaut's project (Cichy et al., 2019). In both projects the competitor is challenged to predict cortical response patterns from multiple datasets from humans and other primates. Both projects demonstrate how sharing data and tests may offer an infrastructure and atmosphere inviting labs around the globe to compete against each other in bechnmarking their models not with regard to task performance, but to their ability to explain cortical function.

By making tests, tasks, models, and data freely available and thus stimulating collaborations across labs und disciplinary boundaries, the field of computational visual neuroscience will be able to develop better models of the human visual cortex. The findings discussed in this thesis - treating DNNs similar to human participants to allow for greater generalbility of the results, and using brain-inspired architectures and ecologically relevant input statistics - may be a step toward a better understanding of the cortical processes underlying vision.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., ... Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16)*, arXiv 1605.08695. https://doi.org/10.1016/0076-6879(83)01039-3

Abbasi-Asl, R., Chen, Y., Bloniarz, A., Oliver, M., Willmore, B., Gallant, J., & Yu, B. (2018). The DeepTune framework for modeling and characterizing neurons in visual cortex area V4. *arXiv*, 1–46.

Agrawal, P., Stansbury, D., Malik, J., & Gallant, J. (2014). Pixels to Voxels: Modeling Visual Representation in the Human Brain, arXiv 1407.5104, 1–15. http://arxiv.org/abs/1407.5104

Ba, J., Mnih, V., & Kavukcuoglu, K. (2014). Multiple Object Recognition with Visual Attention, arXiv 1412.7755, 1–10. http://arxiv.org/abs/1412.7755

Baldauf, D., & Desimone, R. (2014). Neural Mechanisms of Object-Based Attention. *Science*, *1268*(April), 424–428.

Bankson, B. B., Hebart, M. N., Groen, I. I., & Baker, C. I. (2018). The temporal evolution of conceptual object representations revealed through models of behavior, semantics and deep neural networks. *NeuroImage*, *178*(May), 172–182. https://doi.org/10.1016/j.neuroimage.2018.05.037

Bar-Cohen, Y. (2005). Biomimetics: mimicking and inspired-by biology. *Yoseph Bar-Cohen, "Biomimetics: mimicking and inspired-by biology," Proc. SPIE 5759, Smart Structures and Materials 2005: Electroactive Polymer Actuators and Devices (EAPAD), (6 May 2005)*, *5759*(May 2005), 1. https://doi.org/10.1117/12.597436

Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, *364*(6439). https://doi.org/10.1126/science.aav9436

Bhushan, B. (2009). Biomimetics: Lessons from Nature - an overview. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *367*(1893), 1445–1486. https://doi.org/10.1098/rsta.2009.0011

Biederman, I. (1987). Recognition-by-Components: A Theory of Human Image Understanding. *Psychological Review*, *94*(2), 115–147. https://doi.org/10.1037/0033-295X.94.2.115

Biparva, M., & Tsotsos, J. (2018). STNet: Selective tuning of convolutional networks for object localization. *Proceedings - 2017 IEEE International Conference on Computer Vision Workshops, ICCVW 2017*, *2018-Janua*, 2715–2723. https://doi.org/10.1109/ICCVW.2017.319

Brysbaert, M., Warriner, A., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Brysbaert, & New. (2009). Moving beyond Kucera and Francis : A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Cadieu, D., Hong, H., Yamins, D., Pinto, N., Ardila, D., Solomon, E. a., Majaj, N. J., & DiCarlo, J. J. (2014). Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLoS computational biology*, *10*(12), arXiv 1406.3284, 35. https://doi.org/10.1371/journal.pcbi.1003963

Carbon, C. C. (2014). Understanding human perception by human-made illusions. *Frontiers in Human Neuroscience*, *8*(JULY), 1–6. https://doi.org/10.3389/fnhum.2014.00566

Charest, I., Kievit, R. a., Schmitz, T. W., Deca, D., & Kriegeskorte, N. (2014). Unique semantic space in the brain of each beholder predicts perceived similarity. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(40), 14565–70. https://doi.org/10.1073/pnas.1402594111

Chatfield, K., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Return of the Devil in the Details: Delving Deep into Convolutional Nets, arXiv 1405.3531, 1–11. https://doi.org/10.5244/C.28.6

Cho, K., van Merrienboer, B., Bahdanau, D., & Bengio, Y. (2014). On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. *arXiv*, arXiv arXiv:1409.1259v2.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv*, arXiv arXiv:1412.3555v1, 1–9.

Cichy, R., & Kaiser, D. (2019). Deep Neural Networks as Scientific Models. *Trends in Cognitive Sciences*, *23*(4), 305–317. https://doi.org/10.1016/j.tics.2019.01.009

Cichy, R., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Deep Neural Networks predict Hierarchical Spatio-temporal Cortical Dynamics of Human Visual Object Recognition. *arXiv*, arXiv 1601.02970, 15. https://doi.org/10.1038/srep27755

Cichy, R., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature neuroscience*, *17*(3), 455–462. https://doi.org/10.1038/nn.3635

Cichy, R., Roig, G., Andonian, A., Dwivedi, K., Lahner, B., Lascelles, A., Mohsenzadeh, Y., Ramakrishnan, K., & Oliva, A. (2019). The Algonauts Project: A Platform for Communication between the Sciences of Biological and Artificial Intelligence. *arXiv*, *2*arXiv arXiv:1905.05675v1. https://doi.org/10.32470/ccn.2019.1018-0

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Li, F. F. (2009). ImageNet: a Large-Scale Hierarchical Image Database, In *Ieee computer society conference on computer vision and pattern recognition*. https://doi.org/10.1109/CVPR.2009.5206848

Devereux, B. J., Clarke, A., & Tyler, L. K. (2018). Integrated deep visual and semantic attractor neural networks predict fMRI pattern-information along the ventral object processing pathway. *Scientific Reports*, *8*(1), 1–12. https://doi.org/10.1038/s41598-018-28865-1

DiCarlo, J., & Cox, D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, *11*(8), 333–341. https://doi.org/10.1016/j.tics.2007.06.010

DiCarlo, J., Zoccolan, D., & Rust, N. (2012). How does the brain solve visual object recognition? https://doi.org/10.1016/j.neuron.2012.01.010

Dickinson, M. H. (1999). Bionics: Biological insight into mechanical design. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(25), 14208–14209. https://doi.org/10.1073/pnas.96.25.14208

Doersch, C., Gupta, A., & Efros, A. a. (2015). Unsupervised Visual Representation Learning by Context Prediction. *arXiv preprint*, arXiv arXiv:1505.05192v1, 1422–1430. https://doi.org/10.1109/ICCV.2015.167

Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T., Austin, U. T., Lowell, U., & Berkeley, U. C. (2015). Long-term Recurrent Convolutional Networks for Visual Recognition and Description. *CVPR*.

Duhamel, J. R., Bremmer, F., BenHamed, S., & Graf, W. (1997). Spatial invariance of visual receptive fields in parietal cortex neurons. *Nature*, *389*(6653), 845–848. https://doi.org/10.1038/39865

Edelman, G. M., & Gally, J. A. (2001). Degeneracy and complexity in biological systems. *Proceedings of the National Academy of Sciences*, *98*(24), 13763–13768. https://doi.org/10.1073/pnas.231499798

Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, *152*(January 2016), 184–194. https://doi.org/10.1016/j.neuroimage.2016.10.001

Elsken, T., Metzen, J. H., & Hutter, F. (2019). Neural Architecture Search. *Journal of Machine Learning Research*, *20*arXiv arXiv:1808.05377v3, 63–77. https://doi.org/10.1007/978-3-030-05318-5_3

Engel, S., Rumelhart, D., Wandell, B., Lee, A., Glover, G., Chichilnisky, E., & Shadlen, M. (1994). fMRI of human visual cortex. *Nature*, *369*(June), 525.

Everingham, M., Van Gool, L., Williams, C., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, *88*(2), 303–338. https://doi.org/10.1007/s11263-009-0275-4

Fellbaum, C. (2012). WordNet, In *The encyclopedia of applied linguistics*. https://doi.org/10.1002/9781405198431.wbeal1285

Felleman, D., & Van Essen, D. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, *1*, 1–47. http://www.cogsci.ucsd.edu/%7B~%7Dsereno/201/readings/04.03-MacaqueAreas.pdf

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4), arXiv arXiv:1011.1669v3, 193–202. https://doi.org/10.1007/BF00344251

Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Science*, *3*(2), 191–197.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *PMLR*, *9*arXiv arXiv:1011.1669v3, 249–256. https://doi.org/10.1.1.207.2059

Grill-Spector, K., Weiner, K., Kay, K., & Gomez, J. (2017). The Functional Neuroanatomy of Human Face Perception. *Annual Review of Vision Science*, *3*(1), 167–196. https://doi.org/10.1146/annurev-vision-102016-061214

Gross, G. W., Rieske, E., Kreutzberg, G. W., & Meyer, A. (1977). A new fixed-array multi-microelectrode system designed for long-term monitoring of extracellular single unit neuronal activity in vitro A NEW FIXED-ARRAY MULTI-MICROELECTRODE SYSTEM DESIGNED FOR LONG-TERM MONITORING OF EXTRACELLULAR SINGLE UNIT NEURONAL A. *Neuroscience Letters*, *6*, 101–105.

Güçlü, U., & van Gerven, M. (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *Journal of Neuroscience*, *35*(27), arXiv 1411.6422, 10005–10014. https://doi.org/10.1523/JNEUROSCI.5023-14.2015

Güçlü, U., & van Gerven, M. (2017). Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *NeuroImage*, *145*, 329–336. https://doi.org/10.1016/j.neuroimage.2015.12.036

Hagmann, P., Sporns, O., Madan, N., Cammoun, L., Pienaar, R., Wedeen, V. J., Meuli, R., Thiran, J. P., & Grant, P. E. (2010). White matter maturation reshapes structural connectivity in the late developing human brain. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(44), 19067–19072. https://doi.org/10.1073/pnas.1009073107

Hamill, O. P., Marty, A., Neher, E., Sakmann, B., & Sigworth, F. J. (1981). Improved Patch-Clamp Techniques for High-Resolution Current Recording from Cells and Cell-Free Membrane Patches. *Pfluegers Archiv*, *291*(1), 85–100.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-Inspired Artificial Intelligence. *Neuron*, *95*(2), 245–258. https://doi.org/10.1016/j.neuron.2017.06.011

Haxby, J., Horwitz, B., Ungerleider, L. G., Maisog, J. M., Pietro, P., & Grady, C. L. (1994). The Functional Organization of Human Extrastriate Cortex.pdf. *The Journal of Neuroscience*, *14*(11), 6336–6353.

He, K., Zhang, X., Ren, S., & Sun, J. (2014). Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classificatio. *ICCV conference paper*, arXiv {\{}1502.01852v1{\}}. https://doi.org/10.1109/ICCV.2015.123

Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Van Wicklin, C., & Baker, C. (2019). THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *bioRxiv preprint*, *53*(95), arXiv arXiv:1011.1669v3, 45–52. https://doi.org/http://dx.doi.org/10.1101/545954

Hebb, D. O. (1949). *The Organization of Behavior: A Neuropsychological Theory* (Vol. 63). John Wiley & Sons. https://doi.org/10.2307/1418888

Hernández-García, A., Mehrer, J., Kriegeskorte, N., König, P., & Kietzmann, T. (2019). Deep neural networks trained with heavier data augmentation learn features closer to representations in hIT. https://doi.org/10.32470/ccn.2018.1046-0

Herrmann, C. S. (2001). Human EEG responses to 1-100 Hz flicker: Resonance phenomena in visual cortex and their potential correlation to cognitive phenomena. *Experimental Brain Research*, *137*(3-4), 346–353. https://doi.org/10.1007/s002210100682

Heuschkel, M. O., Fejtl, M., Raggenbass, M., Bertrand, D., & Renaud, P. (2002). A three-dimensional multi-electrode array for multi-site stimulation and recording in acute brain slices. *Journal of Neuroscience Methods*, *114*(2), 135–148. https://doi.org/10.1016/S0165-0270(01)00514-3

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hodas, N. O., & Stinis, P. (2018). Doing the Impossible : Why Neural Networks Can Be Trained at All, *9*(July), 1–7. https://doi.org/10.3389/fpsyg.2018.01185

Hodgkin, A. L., & Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *Bulletin of Mathematical Biology*, *52*(1-2), 25–71. https://doi.org/10.1007/BF02459568

Hong, H., Yamins, D. L. K., Majaj, N. J., & DiCarlo, J. J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature Neuroscience*, *19*(4), 613–622. https://doi.org/10.1038/nn.4247

Horikawa, T., & Kamitani, Y. (2017a). Generic decoding of seen and imagined objects using hierarchical visual features. *Nature communications*, *8*(May), 15037. https://doi.org/10.1038/ncomms15037

Horikawa, T., & Kamitani, Y. (2017b). Hierarchical neural representation of dreamed objects revealed by brain decoding with deep neural network features. *Frontiers in Computational Neuroscience*, *11*(January), 1–11. https://doi.org/10.3389/fncom.2017.00004

Huang, G., Maaten, L. V. D., & Weinberger, K. Q. (2016). Densely Connected Convolutional Networks. *CVPR*.

Hubel, D., & Wiesel, T. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, *148*(3), 574–591. https://doi.org/10.1113/jphysiol.1959.sp006308

Hubel, D., & Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 106–154. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1359523/

Kalfas, I., Kumar, S., & Vogels, R. (2017). Shape Selectivity of Middle Superior Temporal Sulcus Body Patch Neurons. *eNeuro*, *4*(June), 1–21.

Kar, K., Kubilius, J., Schmidt, K., Issa, E., & DiCarlo, J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, *22*(6), 974–983. https://doi.org/10.1038/s41593-019-0392-5

Khaligh-Razavi, S., Henriksson, L., Kay, K., & Kriegeskorte, N. (2014). Explaining the hierarchy of visual representational geometries by remixing of features from many computational vision models. *bioRxiv*, 1–35. https://doi.org/10.1101/009936

Khaligh-Razavi, S., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS computational biology*, *10*(11). https://doi.org/10.1371/journal.pcbi.1003915

Kietzmann, T., Lange, S., & Riedmiller, M. (2009). Computational object recognition: A biologically motivated approach. *Biological Cybernetics*, *100*(1), 59–79. https://doi.org/10.1007/s00422-008-0281-6

Kietzmann, T., Lange, S., & Riedmiller, M. (2008). Incremental GRLVQ: Learning relevant features for 3D object recognition. *Neurocomputing*, *71*(13-15), 2868–2879. https://doi.org/10.1016/j.neucom.2007.08.018

Kietzmann, T., McClure, P., & Kriegeskorte, N. (2017). Deep Neural Networks in Computational Neuroscience. *Bioarxiv*. https://doi.org/10.1101/133504

Kietzmann, T., McClure, P., & Kriegeskorte, N. (2019). Deep Neural Networks in Computational Neuroscience. *Oxford Research Encyclopedia of Neuroscience*, (August), 1–28. https://doi.org/10.1093/acrefore/9780190264086.013.46

Kietzmann, T., Spoerer, C., Sörensen, L., Cichy, R., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, 201905544. https://doi.org/10.1073/pnas.1905544116

Kingma, D., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. *arXiv*, arXiv 1412.6980, 1–15. https://doi.org/http://doi.acm.org.ezproxy.lib.ucf.edu/10.1145/1830483.1830503

Kornblith, S., Norouzi, M., Lee, H., & Hinton, G. (2019). Similarity of Neural Network Representations Revisited, arXiv 1905.00414. http://arxiv.org/abs/1905.00414

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, *1*(1), 417–446. https://doi.org/10.1146/annurev-vision-082114-035447

Kriegeskorte, N., & Douglas, P. (2018). Cognitive computational neuroscience. *Nature Neuroscience*, *21*(9), 1148–1160. https://doi.org/10.1038/s41593-018-0210-5

Kriegeskorte, N., & Douglas, P. (2019). Interpreting encoding and decoding models. *Current Opinion in Neurobiology*, *55*, 167–179. https://doi.org/10.1016/j.conb.2019.04.002

Kriegeskorte, N., & Golan, T. (2019). Neural network models and deep learning. *Current Biology*, *29*(7), R231–R236. https://doi.org/10.1016/j.cub.2019.02.034

Kriegeskorte, N., & Kievit, R. (2013). Representational geometry: integrating cognition, computation, and the brain. *Trends in cognitive sciences*, *17*(8), 401–12. https://doi.org/10.1016/j.tics.2013.06.007

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, *2*(November), 4. https://doi.org/10.3389/neuro.06.004.2008

Krizhevsky. (2009). Learning Multiple Layers of Features from Tiny Images. . . . *Science Department, University of Toronto, Tech. . . .*, arXiv arXiv:1011.1669v3, 1–60. https://doi.org/10.1.1.222.9220

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances In Neural Information Processing Systems*, arXiv 1102.0183, 1–9.

Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L. K., & DiCarlo, J. J. (2018). CORnet: Modeling the Neural Mechanisms of Core Object Recognition. *bioRxiv*, 408385. https://doi.org/10.1101/408385

Kuha, J. (2004). AIC and BIC Comparisons of Assumptions and Performance. *Sociological Methods and Research*, *33*, 188–229. https://doi.org/10.1177/0049124103262065

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Duerig, T., & Ferrari, V. (2018). The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale, arXiv 1811.00982, 1–20. https://doi.org/arXiv:1811.00982v1

Lage-castellanos, A., Valente, G., Formisano, E., & De Martino, F. (2019). Methods for computing the maximum performance of computational models of fMRI responses. *PLoS Comput Biol*, 1–25. https://doi.org/10.5281/zenodo.1489531.

Lane, C., Kanjlia, S., Omaki, A., & Bedny, M. (2015). "Visual" cortex of congenitally blind adults responds to syntactic movement. *Journal of Neuroscience*, *35*(37), 12859–12868. https://doi.org/10.1523/JNEUROSCI.1256-15.2015

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, *1*(4), 541–551. https://doi.org/10.1162/neco.1989.1.4.541

Leech, G., Raysen, P., & Wilson, A. (2014). *Word frequencies in written and spoken English*. Harlow: Longman. https://doi.org/10.4324/9781315840161
Leech G, Rayson, P, Wilson, A (2001) Word frequencies in written and spoken English. Harlow: Longman #

Li, N., & DiCarlo, J. (2012). Neuronal learning of invariant object representation in the ventral visual stream is not dependent on reward. *Journal of Neuroscience*, *32*(19), 6611–6620. https://doi.org/10.1523/JNEUROSCI.3786-11.2012

Li, Yosinski, J., Clune, J., Hod, L., & Hopcroft, J. (2016). Convergent Learning: Do different neural networks learn the same representations? *ICLR*, (2014), arXiv arXiv:1511.07543v3, 1–21.

Lin, T., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, L., & Dollar, P. (2014). Microsoft COCO: Common Objects in Context. *arXiv*, arXiv 1405.0312v3, 3686–3693. https://doi.org/10.1109/CVPR.2014.471

Lindsay, G., & Miller, K. (2018). How biological attention mechanisms improve task performance in a large-scale visual system model. *eLife*, *7*, 1–29. https://doi.org/10.7554/eLife.38105

Liu, H., Simonyan, K., & Yang, Y. (2019). DARTS: Differentiable Architecture Search. *ICLR 2019*, arXiv 1806.09055, 1–13. http://arxiv.org/abs/1806.09055

Liu, J., Harris, A., & Kanwisher, N. (2002). Stages of processing in face perception: an MEG study. *Nature Neuroscience*, *5*(september), 910–916. https://doi.org/10.1038/nn909

Logothetis, N. K., Pauls, J., & Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, *5*(5), 552–563. https://doi.org/10.1016/S0960-9822(95)00108-4

Lu, Q., Chen, P.-H., Pillow, J. W., Ramadge, P. J., Norman, K. A., & Hasson, U. (2018). Shared Representational Geometry Across Neural Networks, (Nips), arXiv 1811.11684, 1–7. https://doi.org/arXiv:1811.11684v1

Mahdisoltani, F., Berger, G., Memisevic, R., & Fleet, D. (2018). The more fine-grained , the better for transfer learning, (Nips), 1–8.

Marblestone, A., Wayne, G., & Kording, K. (2016). Towards an integration of deep learning and neuroscience, *10*(September), arXiv 1606.03813, 1–41. https://doi.org/10.3389/fncom.2016.00094

Markman, A., & Wisniewski, E. (1997). Similar and different: The differentiation of basic-level categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(1), 54–70. https://doi.org/10.1037/0278-7393.23.1.54

Markram, H., Meier, K., Lippert, T., Grillner, S., Frackowiak, R., Dehaene, S., Knoll, A., Sompolinsky, H., Verstreken, K., DeFelipe, J., Grant, S., Changeux, J. P., & Sariam, A. (2011). Introducing the Human Brain Project. *Procedia Computer Science*, *7*, 39–42. https://doi.org/10.1016/j.procs.2011.12.015

Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London. Series B, Containing papers of a Biological character. Royal Society (Great Britain)*, *200*(1140), 269–294. https://doi.org/10.1098/rspb.1978.0020

Martinez, L. M., & Alonso, J. M. (2003). Complex receptive fields in primary visual cortex. *Neuroscientist*, *9*(5), 317–331. https://doi.org/10.1177/1073858403252732

Maunsell, J. (1987). Visual Processing In Monkey Extrastriate Cortex. *Annual Review of Neuroscience*, *10*(1), 363–401. https://doi.org/10.1146/annurev.neuro.10.1.363

McClure, P., & Kriegeskorte, N. (2016a). Representational Distance Learning for Deep Neural Networks. *ICLR 2016*, arXiv 1511.03979v5.

McClure, P., & Kriegeskorte, N. (2016b). Representing inferential uncertainty in deep neural networks through sampling, (2000), arXiv 1611.01639, 1–14. https://arxiv.org/pdf/1611.01639.pdf

Medaglia, J. D. (2017). Functional neuroimaging in traumatic brain injury: From nodes to networks. *Frontiers in Neurology*, *8*(AUG), 1–18. https://doi.org/10.3389/fneur.2017.00407

Mei, J., & Singh, T. (2018). Intra-thalamic and thalamocortical connectivity. *2018 IEEE/ACM 1st International Workshop on Software Engineering for Cognitive Services (SE4COG)*, 9–14. https://doi.org/10.1145/3195555.3195556

Miikkulainen, R., Liang, J., Meyerson, E., Rawal, A., Fink, D., Raju, B., Shahrzad, H., Navruzyan, A., Du, N., & Hodjat, B. (2017). Evolving Deep Neural Networks. *Arxiv preprint*, arXiv arXiv:1703.00548v2.

Miller, G. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39–41. https://doi.org/10.1145/219717.219748

Mishkin, D., & Matas, J. (2015). All you need is a good init. *arXiv preprint*, arXiv 1511.06422, 1–13. http://arxiv.org/abs/1511.06422

Morcos, A., Raghu, M., & Bengio, S. (2018). Insights on representational similarity in neural networks with canonical correlation, (Nips), arXiv 1806.05759. https://doi.org/arXiv:1806.05759v3

Mott, A., Zoran, D., Chrzanowski, M., Wierstra, D., & Rezende, D. J. (2019). Towards Interpretable Reinforcement Learning Using Attention Augmented Agents, arXiv 1906.02500. http://arxiv.org/abs/1906.02500

Nayebi, A., Bear, D., Kubilius, J., Kar, K., Ganguli, S., Sussillo, D., DiCarlo, J. J., & Yamins, D. L. (2018). Task-driven convolutional recurrent models of the visual system. *Advances in Neural Information Processing Systems*, *2018-Decem*(NeurIPS), 5290–5301.

Neher, E., & Sakmann, B. (1976). Single-challe currents recorded from membrane of denerated frog muscle fibres. *Nature*, *260*, 799–801.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A Toolbox for Representational Similarity Analysis, *10*(4). https://doi.org/10.1371/journal.pcbi.1003553

Noppeney, U., Friston, K. J., & Price, C. J. (2004). Degenerate neuronal systems sustaining cognitive functions. *Journal of Anatomy*, *205*(6), 433–442. https://doi.org/10.1111/j.0021-8782.2004.00343.x

Noppeney, U., Penny, W. D., Price, C. J., Flandin, G., & Friston, K. J. (2006). Identification of degenerate neuronal systems based on intersubject variability. *NeuroImage*, *30*(3), 885–890. https://doi.org/10.1016/j.neuroimage.2005.10.010

Novak, R., Bahri, Y., Abolafia, D. A., Pennington, J., & Sohl-dickstein, J. (2018). Sensitivity and Generalization in Neural Networks: an Empirical Study. *ICLR*, arXiv arXiv:1802.08760v3, 1–21.

Op de Beeck, H., Haushofer, J., & Kanwisher, N. (2008). Interpreting fMRI data: maps, modules and dimensions. *Nature Neuroscience Review*, *9*(4), 341. https://doi.org/10.1038/nrn2314.Interpreting

Palmeri, T. J., Wong, A. C.-n., & Gauthier, I. (2004). Computational approaches to the development of perceptual expertise, *8*(8). https://doi.org/10.1016/j.tics.2004.06.001

Pernet, C. R., Wilcox, R., & Rousselet, G. A. (2013). Robust correlation analyses: False positive and power validation using a new open source matlab toolbox. *Frontiers in Psychology*, *3*(JAN), arXiv 9605103, 1–18. https://doi.org/10.3389/fpsyg.2012.00606

Pham, H., Guan, M. Y., Zoph, B., Le, Q. V., & Dean, J. (2018). Efficient Neural Architecture Search via parameter Sharing. *35th International Conference on Machine Learning, ICML 2018*, *9*arXiv arXiv:1802.03268v2, 6522–6531.

Popescu, A. I. (1999). Bionics, Biological systems and teh principle of optimal design. *Acta Biotheoretica*, *46*, 299–310.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychologyt*, (32), 3–25.

Price, C. J., & Friston, K. J. (2002). Degeneracy and cognitive anatomy. *Trends in Cognitive Sciences*, *6*(10), 416–421. https://doi.org/10.1016/S1364-6613(02)01976-9

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., … Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, *22*(11), 1761–1770. https://doi.org/10.1038/s41593-019-0520-2

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025. https://doi.org/10.1038/14819

Rumelhart, D., Hinton, G., & Mcclelland, J. L. (1986). Learning Internal Representationa by Error Propagation. *Nature*, (323), 533–536. https://web.stanford.edu/class/psych209a/ReadingsByDate/02%7B%5C_%7D06/PDPVolIChapter8.pdf

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), arXiv 1409.0575, 211–252. https://doi.org/10.1007/s11263-015-0816-y

Sakmann, B., & Neher, E. (1984). Patch Clamp Techniques for Studying Ionic Channels in Excitable Membranes. *Annual Review of Physiology*, *46*(1), 455–472. https://doi.org/10.1146/annurev.physiol.46.1.455

Sakmann, B., & Neher, E. (1995). *Single-Channel Recordning* (Vol. 8). https://doi.org/10.1007/s10330-009-0147-y

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, *166*(23), 11537–11546. https://doi.org/10.1073/pnas.1820226116

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2013). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks, arXiv 1312.6120, 1–22. http://arxiv.org/abs/1312.6120

Schmidhuber, J. (2015). Deep Learning in neural networks: An overview. *Neural Networks*, *61*arXiv arXiv:1404.7828, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Schmitt, O. H. (1969). Some interesting and useful biomimetic transforms, In *Third int. biophysics congress*.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Schmidt, K., Yamins, D. L. K., & DiCarlo, J. J. (2018). Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? *bioRxiv*, 407007. https://doi.org/10.1101/407007

Seckfort, D. L., Paul, R., Grieve, S. M., Vandenberg, B., Bryant, R. A., Williams, L. M., Clark, C. R., Cohen, R. A., Bruce, S., & Gordon, E. (2008). Early life stress on brain structure and function across the lifespan: A preliminary study. *Brain Imaging and Behavior*, *2*(1), 49–58. https://doi.org/10.1007/s11682-007-9015-y

Serre, T. (2019). Deep Learning: The Good, the Bad, and the Ugly. *Annual Review of Vision Science*, *5*(1). https://doi.org/10.1146/annurev-vision-091718-014951

Serre, T. (2015). Hierarchical Models of the Visual System (D. Jaeger & R. Jung, Eds.). In D. Jaeger & R. Jung (Eds.), *Encyclopedia of computational neuroscience*.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences*, *104*(15), 6424–6429. https://doi.org/10.1073/pnas.0700622104

Serre, T., Wolf, L., Bilschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust Object Recognition with Cortex-Like Mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(3), 102–104.

Shi, Y., Tian, Y., Wang, Y., Zeng, W., & Huang, T. (2017). Learning Long-Term Dependencies for Action Recognition with a Biologically-Inspired Deep Network. *Proceedings of the IEEE International Conference on Computer Vision*, *2017-Octob*, 716–725. https://doi.org/10.1109/ICCV.2017.84

Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv*, arXiv 1409.1556, 1–14. http://arxiv.org/abs/1409.1556

Song, S., Miller, K., & Abbot, L. (2000). Competitive Hebbian Learning Through Spiking-Timing Dependent Plasiticity. *Nature*, *3*, 919–926. https://doi.org/10.16953/deusbed.74839

Spoerer, C., Kietzmann, T., & Kriegeskorte, N. (2019). Recurrent networks can recycle neural resources to flexibly trade speed for accuracy in visual recognition. *bioRxiv*, 1–22. https://doi.org/10.1101/677237

Spoerer, C., McClure, P., & Kriegeskorte, N. (2017). Recurrent convolutional neural networks: A better model of biological object recognition. *Frontiers in Psychology*, *8*(SEP), 1–14. https://doi.org/10.3389/fpsyg.2017.01551

Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2015). Striving for Simplicity: The All Convolutional Net. *Iclr*, arXiv 1412.6806, 1–14. http://arxiv.org/abs/1412.6806

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, *15*arXiv 1102.4807, 1929–1958. https://doi.org/10.1214/12-AOS1000

Steele, J. E. (1960). How Do We Get There?, In *Bionics symposium: Living prototypes–the key to new technology*. WADD Technical Report 60-600, Wright Air Development Division, Wright˜...

Storrs, K., Kietzmann, T. C., Walther, A., Mehrer, J., & Kriegeskorte, N. (2020). Diverse deep neural networks all predict human IT well , after training and fitting. *Bioarxiv*. https://doi.org/https://doi.org/10.1101/2020.05.07.082743

Storrs, K., & Kriegeskorte, N. (2019). Deep Learning for Cognitive Neuroscience. *arXiv*, arXiv 1903.01458, 1–26. http://arxiv.org/abs/1903.01458

Storrs, K., Mehrer, J., Walther, A., & Kriegeskorte, N. (2017). Domain-specialised CNNs of realistic depth best explain FFA and PPA representations, In *Presentation at cosyne 2017, salt lake city, us*.

Sun, K., & Nielsen, F. (2020). Lightlike Neuromanifolds, Occam's Razor and Deep Learning. *arXiv preprint*, (1), arXiv arXiv:1905.11027v2, 1–19.

Sutskever, I., Martens, J., Dahl, G., & Hinton, G. (2013). On the importance of initialization and momentum in deep learning. *Proceedings of the International Conference on Machine Learning*, *61*(2), 20–22.

Tanaka, J., & Taylor, M. (1991). Object categories and expertise: Is the basic-level in the eye of the beholder? *Cognitive Psychology*, *23*, 457–482. https://doi.org/10.1016/0010-0285(91)90016-H

Tanaka, K. (1997). Mechanisms of visual object recognition: Monkey and human studies. *Current Opinion in Neurobiology*, *7*(4), 523–529. https://doi.org/10.1016/S0959-4388(97)80032-3

Thomas, C. A., Springer, P. A., Loeb, G. E., Berwald-Netter, Y., & Okun, L. M. (1972). A miniature microelectrode array to monitor the bioelectric activity of cultured cells. *Experimental Cell Research*, *74*(1), 61–66. https://doi.org/10.1016/0014-4827(72)90481-8

Tripp, B. (2018). A deeper understanding of the brain. *NeuroImage*, *180*, 114–116. https://doi.org/10.1016/j.neuroimage.2017.12.079

Ungerleider, L. G., & Haxby, J. V. (1994). 'What' and 'where' in the human brain Introduction. *Current Opinion in Neurobiology*, *4*, 157–165. http://psych.colorado.edu/%7B~%7Dkimlab/ungerleider%7B%5C_%7Dhaxby.94.pdf

Vincent, J. (2001). Stealing ideas from nature. https://doi.org/10.1007/978-3-7091-2584-7

Vincent, J., Bogatyreva, O., Bogatyrev, N., Bowyer, A., & Pahl, A. (2006). Biomimetics: Its practice and theory. *Journal of the Royal Society Interface*, *3*(9), 471–482. https://doi.org/10.1098/rsif.2006.0127

Wandell, B., & Winawer, J. (2016). Computational neuroimaging and population receptive fields Understanding sensory circuits, *19*(6), 349–357. https://doi.org/10.1016/j.tics.2015.03.009

Winawer, J., Kay, K., Foster, B., Rauschecker, A., Parvizi, J., & Wandell, B. (2013). Asynchronous broadband signals are the principal source of the bold response in human visual cortex. *Current Biology*, *23*(13), 1145–1153. https://doi.org/10.1016/j.cub.2013.05.001

Wu, Y., & He, K. (2018). Group normalization. *arXiv*, *11217 LNCS*arXiv arXiv:1803.08494v3, 3–19. https://doi.org/10.1007/978-3-030-01261-8_1

Xie, D., Xiong, J., & Pu, S. (2017). All you need is beyond a good init: Exploring better solution for training extremely deep convolutional neural networks with orthonormality and modulation. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, *2017-Janua*arXiv arXiv:1703.01827v3, 5075–5084. https://doi.org/10.1109/CVPR.2017.539

Xie, S., Zheng, H., Liu, C., & Lin, L. (2019). SNAS: Stochastic Neural Architecture Search. *ICLR 2019*, arXiv 1812.09926, 1–17. http://arxiv.org/abs/1812.09926

Xu, K., Ba, J. L., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R. S., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. *32nd International Conference on Machine Learning, ICML 2015*, *3*arXiv arXiv:1502.03044v3, 2048–2057.

Yamins, D., & DiCarlo, J. (2016a). Eight open questions in the computational modeling of higher sensory cortex. *Current Opinion in Neurobiology*, *37*, 114–120. https://doi.org/10.1016/j.conb.2016.02.001

Yamins, D., & DiCarlo, J. (2016b). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), arXiv science.aab3050, 356–365. https://doi.org/10.1038/nn.4244

Yamins, D., Hong, H., & Cadieu, C. (2013). Hierarchical Modular Optimization of Convolutional Networks Achieves Representations Similar to Macaque IT and Human Ventral Stream. *Advances in neural information processing systems*, (October), 1–9. http://machinelearning.wustl.edu/mlpapers/papers/NIPS2013%7B%5C_%7D4991%7B%5C%%7D5Cnpapers3://publication/uuid/E90976F4-5E4C-482D-B785-561E5A45B9D2

Yamins, D., Hong, H., Cadieu, C., Solomon, E., Seibert, D., & DiCarlo, J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111

Zeiler, M., & Fergus, R. (2014). Visualizing and understanding convolutional networks. *Computer Vision–ECCV 2014*, *8689*, 818–833. http://link.springer.com/chapter/10.1007/978-3-319-10590-1%7B%5C_%7D53

Zhang, M., Ma, K. T., Lim, J. H., & Zhao, Q. (2018). Foveated neural network: Gaze prediction on egocentric videos. *Proceedings - International Conference on Image Processing, ICIP*, *2017-Septe*, 3720–3724. https://doi.org/10.1109/ICIP.2017.8296977

A | Rotation sensitivity of correlation distance and representational consistency within vs. across layers

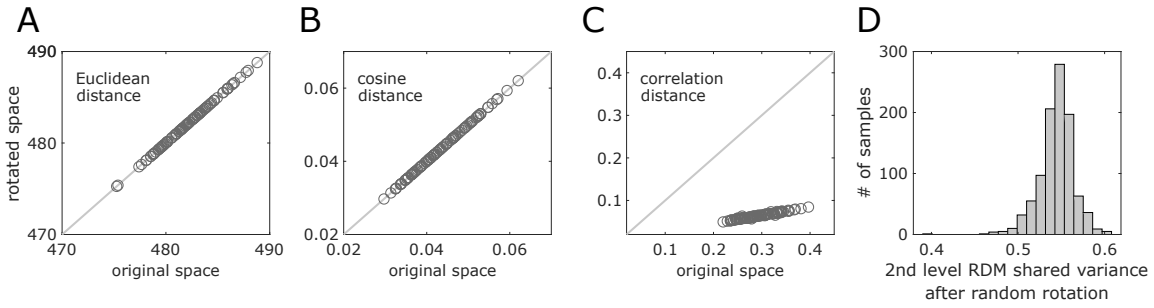


Fig. 1 **Appendix A: Rotation sensitivity of correlation distance.** We computed the distance between two random vectors before and after both vectors were randomly rotated around the origin using the same rotation matrix. This procedure was performed for 100 vector pairs in the above simulation. Rotating both vector pairs does not have an effect when Euclidean or cosine distance is used to compute the vector pair distances (**A, B**). However, when correlation distance is used, rotations around the origin lead to decreased overall distances, and an imperfect correlation (**C**). Computing a distance involves a projection of two vectors a plane cutting through the origin that is orthogonal to the all-1 vector. This projection differs if the original vectors are rotated. Accordingly, when RDMs are based on correlation distance (here based on 10 example responses), rotations around the origin lead to decreased representational consistency, despite the fact that the relative arrangement of datapoints remained identical after the rotation (**D**).

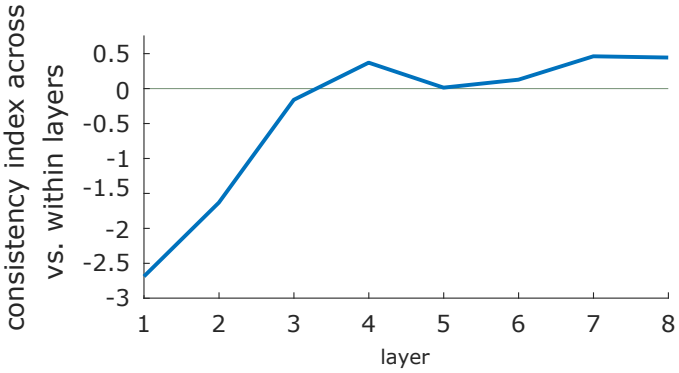


Fig. 2 **Appendix A: Consistency index across vs. within layers.** We computed consistency across instances (and within layers, e.g. off-diagonal elements in Fig2.4 A, $cell_{layer_4,layer_4}$) and subtracted its mean from consistency computed within instances (and across layers, e.g. diagonal elements in Fig2.4 A in $cell_{layer_4,layer_5}$), standardized by the overall mean. This indicates that starting at layer 4 network instances are more consistent across adjacent layers than instances within layers.

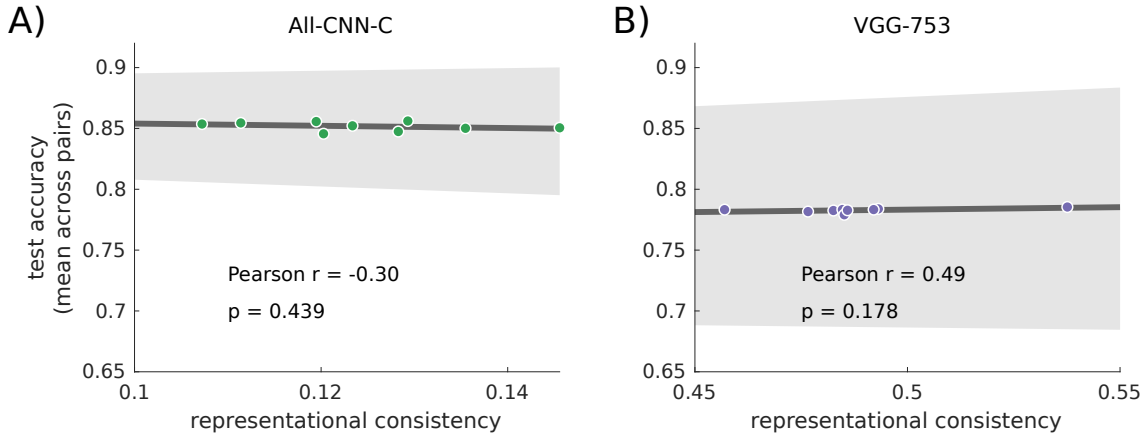


Fig. 3 **Appendix A: Relating test accuracy representational consistency**. We here investigated whether test accuracy is related to representational consistency. For example, high test accuracy of two network instances might allow for less variability in the network internal representations than when two instances with relatively low test accuracy are compared. Note that test accuracy is associated with a single network instance, whereas consistency reflects the representational similarity of a pair of instances. To address this issue, we computed the mean test accuracy of adjacent pairs of networks after sorting them according to their test accuracy. In this way we obtained 9 mean test accuracy values from 10 network instances. We then computed consistency for the same 9 pairs of networks (at the penultimate layer) and used Pearson correlation to relate mean test accuracy and consistency. For both architectures, All-CNN-C (A) and VGG-753 (B), we find correlations between test accuracy and representational consistency not to be significant (p = 0.439 and p = 0.178, respectively).

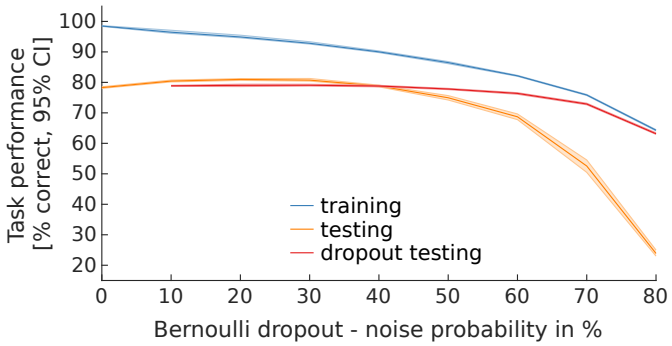


Fig. 4 **Appendix A: VGG-753 task performance across noise levels.** Average task performance across all network instances (shown with 95% CI) for the training set (blue), test set (orange), and when using dropout sampling at inference time for the test set (red, 1 sample).

B | List of ecoset categories

Table 1 **List of ecoset categories.** A list of all 565 ecoset categories sorted according to the frequency-concreteness-index (FCI) used during category selection (for details on the FCI, see formula 3.1). As described in the section on in- and exclusion criteria for ecoset categories "category name" describes single word concepts. "SUBTLEX US frequency" describes the number of times the word appears in the corpus SUBTLEX US (Brysbaert & New, 2009). "concreteness" refers to concreteness ratings based on an online experiment based on 4,000 subjects M. Brysbaert et al., 2014, and "# of images" refers to the total number of images per category.

category index	category name	SUBTLEX US freq.	concreteness	FCI	# of images
001	man	94133	4.79	0.979000	4620
002	house	26214	5.00	0.639239	4905
003	car	24636	4.89	0.619857	4988
004	woman	22166	4.46	0.563738	4898
005	phone	13756	4.86	0.559067	4732
006	bed	9543	5.00	0.550689	4821
007	gun	10873	4.83	0.540753	4862
008	book	9026	4.90	0.537943	4928
009	dog	9835	4.85	0.537240	4904
010	ball	5353	5.00	0.528433	3640
011	fire	10990	4.68	0.526375	4957
012	horse	4737	5.00	0.525161	4997
013	city	8624	4.79	0.524808	1416
014	fish	4258	5.00	0.522617	1392
015	child	8040	4.78	0.520706	4391
016	boat	4885	4.93	0.518947	1464
017	table	5387	4.90	0.518614	4983
018	tree	3315	5.00	0.517608	2471
019	clock	2990	5.00	0.515882	4942

Continuation of Table 1					
category index	category name	SUBTLEX US freq.	concreteness	FCI	# of images
020	bag	4796	4.90	0.515475	4981
021	camera	2907	5.00	0.515441	4948
022	cup	2634	5.00	0.513991	4781
023	ship	5043	4.87	0.513787	4955
024	kitchen	2974	4.97	0.512797	1954
025	key	4430	4.89	0.512531	3713
026	bird	2318	5.00	0.512312	3325
027	pig	1996	5.00	0.510602	4304
028	bus	3783	4.90	0.510094	4053
029	bridge	2331	4.97	0.509381	4836
030	pizza	1709	5.00	0.509078	4870
031	computer	3011	4.93	0.508993	4591
032	church	3553	4.90	0.508872	4921
033	doll	1263	5.00	0.506709	3880
034	bell	2006	4.96	0.506655	4868
035	stairs	1212	5.00	0.506438	2081
036	apple	1207	5.00	0.506411	3542
037	flower	1161	5.00	0.506167	4550
038	ring	4730	4.81	0.506124	1792
039	snake	1140	5.00	0.506055	4906
040	mountain	1805	4.96	0.505587	4996
041	road	5709	4.75	0.505324	1597
042	wall	3605	4.86	0.505148	4837
043	tiger	945	5.00	0.505019	4887
044	toilet	1474	4.97	0.504829	3733
045	train	4848	4.79	0.504751	4768
046	bottle	2588	4.91	0.504747	4510
047	turtle	869	5.00	0.504616	4933
048	cookie	852	5.00	0.504526	950
049	egg	1328	4.97	0.504054	3781
050	river	2829	4.89	0.504027	2532
051	cat	3383	4.86	0.503969	4985
052	truck	3716	4.84	0.503738	4976
053	basket	672	5.00	0.503569	3490

Continuation of Table 1					
category index	category name	SUBTLEX US freq.	concreteness	FCI	# of images
054	bear	2928	4.88	0.503552	3670
055	moon	2548	4.90	0.503534	853
056	milk	2169	4.92	0.503521	2586
057	blanket	662	5.00	0.503516	1806
058	lemon	613	5.00	0.503256	2656
059	frog	603	5.00	0.503203	3597
060	pillow	581	5.00	0.503086	1988
061	elephant	580	5.00	0.503081	4869
062	cow	1301	4.96	0.502910	4834
063	banana	547	5.00	0.502905	3656
064	goat	537	5.00	0.502852	4782
065	knife	2387	4.90	0.502679	4737
066	ladder	472	5.00	0.502507	865
067	popcorn	465	5.00	0.502470	2136
068	refrigerator	427	5.00	0.502268	3476
069	jar	424	5.00	0.502252	3026
070	jail	3602	4.83	0.502133	2199
071	hamburger	397	5.00	0.502109	3693
072	toast	1707	4.93	0.502067	1907
073	umbrella	382	5.00	0.502029	1955
074	bean	349	5.00	0.501854	4741
075	castle	1099	4.96	0.501837	1379
076	flashlight	302	5.00	0.501604	3807
077	tomato	301	5.00	0.501599	4800
078	strawberry	282	5.00	0.501498	2405
079	leopard	276	5.00	0.501466	4290
080	donkey	273	5.00	0.501450	3350
081	axe	249	5.00	0.501323	4381
082	mailbox	212	5.00	0.501126	2892
083	grape	204	5.00	0.501084	4845
084	vase	196	5.00	0.501041	2799
085	carrot	195	5.00	0.501036	2316
086	tractor	190	5.00	0.501009	2452
087	cupcake	167	5.00	0.500887	1070

Continuation of Table 1					
category index	category name	SUBTLEX US freq.	concreteness	FCI	# of images
088	fern	163	5.00	0.500866	4941
089	bagel	156	5.00	0.500829	1832
090	telescope	150	5.00	0.500797	4516
091	cactus	148	5.00	0.500786	4571
092	tent	892	4.96	0.500738	4858
093	microscope	129	5.00	0.500685	4390
094	kite	117	5.00	0.500621	1272
095	lantern	103	5.00	0.500547	2010
096	octopus	99	5.00	0.500526	1054
097	lamp	657	4.97	0.500490	3300
098	blender	85	5.00	0.500451	1295
099	burrito	84	5.00	0.500446	2256
100	mango	84	5.00	0.500446	1874
101	binoculars	80	5.00	0.500425	1704
102	steak	828	4.96	0.500398	2257
103	gravel	73	5.00	0.500388	1456
104	escalator	66	5.00	0.500351	746
105	walrus	57	5.00	0.500303	1179
106	horseshoe	52	5.00	0.500276	1355
107	antelope	50	5.00	0.500266	4428
108	tongs	40	5.00	0.500212	2123
109	porcupine	33	5.00	0.500175	1646
110	camcorder	32	5.00	0.500170	1312
111	mousetrap	30	5.00	0.500159	843
112	lion	783	4.96	0.500159	4644
113	cauliflower	28	5.00	0.500149	2395
114	shower	2097	4.89	0.500138	1877
115	hotdog	20	5.00	0.500106	3407
116	warthog	16	5.00	0.500085	1622
117	thimble	14	5.00	0.500074	1052
118	guardrail	14	5.00	0.500074	1012
119	dustpan	13	5.00	0.500069	1293
120	crawfish	13	5.00	0.500069	1881
121	eyedropper	7	5.00	0.500037	932

Continuation of Table 1					
category index	category name	SUBTLEX US freq.	concreteness	FCI	# of images
122	nectarine	3	5.00	0.500016	1574
123	flyswatter	2	5.00	0.500011	979
124	lollypop	2	5.00	0.500011	919
125	lightbulb	0	5.00	0.500000	2071
126	corn	725	4.96	0.499851	4212
127	cave	713	4.96	0.499787	4537
128	pie	1466	4.92	0.499787	900
129	spider	515	4.97	0.499735	4934
130	bread	1445	4.92	0.499675	4802
131	motorcycle	455	4.97	0.499417	3890
132	monkey	1709	4.90	0.499078	4664
133	whale	574	4.96	0.499049	4895
134	airplane	557	4.96	0.498959	4965
135	shovel	349	4.97	0.498854	3087
136	bucket	511	4.96	0.498714	2308
137	rabbit	1068	4.93	0.498673	2958
138	necklace	497	4.96	0.498640	2386
139	moose	282	4.97	0.498498	1688
140	glass	3096	4.82	0.498445	4841
141	drum	432	4.96	0.498295	4921
142	mop	211	4.97	0.498121	2698
143	bracelet	398	4.96	0.498114	3404
144	spoon	388	4.96	0.498061	3095
145	stove	387	4.96	0.498056	4722
146	lettuce	173	4.97	0.497919	4551
147	ashtray	166	4.97	0.497882	1367
148	lake	1836	4.88	0.497752	1435
149	noodles	309	4.96	0.497641	3890
150	walnut	100	4.97	0.497531	1512
151	pastry	98	4.97	0.497521	2348
152	ferret	83	4.97	0.497441	1929
153	fig	62	4.97	0.497329	1991
154	eggplant	56	4.97	0.497297	1971
155	violin	242	4.96	0.497285	4690

Continuation of Table 1					
category index	category name	SUBTLEX US freq.	concreteness	FCI	# of images
156	chipmunk	42	4.97	0.497223	2813
157	milkshake	42	4.97	0.497223	901
158	blackberry	38	4.97	0.497202	1397
159	sushi	222	4.96	0.497179	1744
160	apricot	32	4.97	0.497170	1666
161	drawers	218	4.96	0.497158	4411
162	doughnut	215	4.96	0.497142	1658
163	lawnmower	24	4.97	0.497127	4465
164	snowplow	20	4.97	0.497106	1092
165	chalkboard	12	4.97	0.497064	1931
166	backpack	186	4.96	0.496988	2622
167	alligator	178	4.96	0.496945	3190
168	cockroach	174	4.96	0.496924	3172
169	lime	168	4.96	0.496892	1681
170	lighthouse	157	4.96	0.496834	2693
171	dolphin	141	4.96	0.496749	4853
172	piano	1268	4.90	0.496735	4825
173	earring	138	4.96	0.496733	2561
174	chicken	3148	4.80	0.496721	4825
175	blueberry	131	4.96	0.496696	3318
176	grapefruit	126	4.96	0.496669	1544
177	thermometer	112	4.96	0.496595	3814
178	cranberry	99	4.96	0.496526	2724
179	iceberg	92	4.96	0.496489	1165
180	wasp	73	4.96	0.496388	3671
181	tweezers	52	4.96	0.496276	1339
182	asparagus	50	4.96	0.496266	2430
183	croissant	45	4.96	0.496239	1602
184	teapot	44	4.96	0.496234	2164
185	needle	608	4.93	0.496229	1854
186	acorn	37	4.96	0.496197	1258
187	bumblebee	33	4.96	0.496175	1727
188	anvil	32	4.96	0.496170	971
189	seashell	16	4.96	0.496085	4887

Continuation of Table 1					
category index	category name	SUBTLEX US freq.	concreteness	FCI	# of images
190	birdhouse	13	4.96	0.496069	1691
191	thyme	10	4.96	0.496053	4256
192	thumbtack	9	4.96	0.496048	1107
193	saltshaker	6	4.96	0.496032	1466
194	paperclip	2	4.96	0.496011	1209
195	matchstick	1	4.96	0.496005	897
196	envelope	513	4.93	0.495725	3787
197	pineapple	130	4.94	0.494691	2079
198	wheelchair	316	4.93	0.494678	2417
199	butterfly	281	4.93	0.494493	4978
200	camel	256	4.93	0.494360	4588
201	balloon	442	4.92	0.494348	3606
202	beach	2888	4.79	0.494340	2583
203	guitar	795	4.90	0.494223	4853
204	crate	209	4.93	0.494110	1187
205	wrench	202	4.93	0.494073	3080
206	eggroll	1	4.94	0.494005	1318
207	taco	158	4.93	0.493839	1183
208	rat	1663	4.85	0.493833	3956
209	meatball	132	4.93	0.493701	2795
210	emerald	131	4.93	0.493696	1386
211	omelet	121	4.93	0.493643	2878
212	sheep	685	4.90	0.493638	4807
213	jukebox	116	4.93	0.493616	1289
214	desert	1427	4.86	0.493580	720
215	raspberry	96	4.93	0.493510	3568
216	snail	90	4.93	0.493478	4763
217	pistachio	77	4.93	0.493409	1443
218	groundhog	76	4.93	0.493404	3521
219	videotape	264	4.92	0.493402	868
220	jellyfish	74	4.93	0.493393	2674
221	carousel	73	4.93	0.493388	2049
222	hippo	72	4.93	0.493382	2001
223	pear	68	4.93	0.493361	3832

Continuation of Table 1					
category index	category name	SUBTLEX US freq.	concreteness	FCI	# of images
224	bullet	1950	4.83	0.493358	1654
225	pliers	59	4.93	0.493313	2990
226	toothpick	52	4.93	0.493276	1142
227	stagecoach	51	4.93	0.493271	1180
228	hourglass	49	4.93	0.493260	1612
229	coliseum	44	4.93	0.493234	2104
230	blowtorch	43	4.93	0.493228	1288
231	treadmill	42	4.93	0.493223	2506
232	cake	2298	4.81	0.493206	3865
233	bobcat	20	4.93	0.493106	1571
234	cumin	16	4.93	0.493085	1162
235	hedgehog	15	4.93	0.493080	1730
236	opossum	4	4.93	0.493021	2426
237	pumpkin	553	4.90	0.492937	3535
238	bowl	1094	4.87	0.492811	4230
239	worm	516	4.90	0.492741	4550
240	candy	1825	4.83	0.492694	4245
241	clarinet	80	4.92	0.492425	2326
242	peanut	630	4.89	0.492346	1656
243	gondola	37	4.92	0.492197	1418
244	padlock	35	4.92	0.492186	2539
245	television	1729	4.83	0.492184	4796
246	barnacle	32	4.92	0.492170	2376
247	leek	15	4.92	0.492080	964
248	cashew	11	4.92	0.492058	2063
249	streetlamp	2	4.92	0.492011	1830
250	razor	351	4.90	0.491864	3552
251	peach	324	4.90	0.491721	2081
252	pudding	314	4.90	0.491668	3726
253	coin	497	4.89	0.491640	1377
254	grasshopper	47	4.91	0.491250	4815
255	pea	199	4.90	0.491057	4033
256	toaster	198	4.90	0.491052	1566
257	chalk	183	4.90	0.490972	858

Continuation of Table 1					
category index	category name	SUBTLEX US freq.	concreteness	FCI	# of images
258	bee	528	4.88	0.490805	4848
259	bicycle	337	4.89	0.490790	4945
260	screwdriver	128	4.90	0.490680	1803
261	pencil	503	4.88	0.490672	3492
262	jaguar	121	4.90	0.490643	1545
263	garlic	306	4.89	0.490625	1767
264	rhino	106	4.90	0.490563	4615
265	wheat	293	4.89	0.490556	1462
266	waterfall	95	4.90	0.490505	1012
267	squirrel	279	4.89	0.490482	4939
268	bulldozer	66	4.90	0.490351	1996
269	closet	1381	4.83	0.490335	2283
270	broom	243	4.89	0.490291	3561
271	starfish	38	4.90	0.490202	2019
272	ladle	38	4.90	0.490202	2481
273	scoreboard	31	4.90	0.490165	1652
274	rice	769	4.86	0.490085	2906
275	crouton	13	4.90	0.490069	971
276	lasagna	183	4.89	0.489972	2352
277	flea	169	4.89	0.489898	893
278	towel	722	4.86	0.489835	4729
279	bench	493	4.87	0.489619	4923
280	hammock	71	4.89	0.489377	757
281	windmill	65	4.89	0.489345	3637
282	pan	627	4.86	0.489330	2966
283	avocado	62	4.89	0.489329	2121
284	guacamole	56	4.89	0.489297	1726
285	cane	425	4.87	0.489257	2022
286	sprinkler	47	4.89	0.489250	2132
287	microphone	232	4.88	0.489232	4475
288	outhouse	38	4.89	0.489202	1494
289	anthill	29	4.89	0.489154	1030
290	tortilla	16	4.89	0.489085	1493
291	eggbeater	12	4.89	0.489064	1378

Continuation of Table 1					
category index	category name	SUBTLEX US freq.	concreteness	FCI	# of images
292	hare	195	4.88	0.489036	3177
293	breadbox	6	4.89	0.489032	1043
294	nest	566	4.86	0.489006	701
295	rifle	743	4.85	0.488947	4842
296	skunk	166	4.88	0.488882	1810
297	barrel	542	4.86	0.488879	4761
298	typewriter	161	4.88	0.488855	4482
299	bun	147	4.88	0.488781	3991
300	hay	325	4.87	0.488726	2687
301	mosquito	93	4.88	0.488494	3694
302	coffin	461	4.86	0.488449	2078
303	newspaper	1208	4.82	0.488416	4797
304	condom	263	4.87	0.488397	702
305	deer	444	4.86	0.488358	4911
306	mouse	975	4.83	0.488179	4515
307	candle	409	4.86	0.488172	3296
308	kumquat	24	4.88	0.488127	2201
309	cymbals	24	4.88	0.488127	995
310	mall	964	4.83	0.488120	1478
311	potato	576	4.85	0.488060	1211
312	candelabra	6	4.88	0.488032	3342
313	knot	188	4.87	0.487999	3799
314	lobster	374	4.86	0.487987	3966
315	crib	316	4.86	0.487678	2503
316	broccoli	116	4.87	0.487616	2416
317	fireworks	287	4.86	0.487524	1507
318	ant	273	4.86	0.487450	4857
319	crowbar	66	4.87	0.487351	703
320	thermostat	58	4.87	0.487308	934
321	caterpillar	57	4.87	0.487303	4617
322	papaya	56	4.87	0.487297	1075
323	ceiling	426	4.85	0.487263	2274
324	zucchini	49	4.87	0.48726	1453
325	pecan	48	4.87	0.487255	1000

Continuation of Table 1					
category index	category name	SUBTLEX US freq.	concreteness	FCI	# of images
326	wallet	1163	4.81	0.487177	2164
327	radish	31	4.87	0.487165	2947
328	onion	216	4.86	0.487147	3939
329	crayon	21	4.87	0.487112	1572
330	pancake	202	4.86	0.487073	1165
331	canoe	182	4.86	0.486967	3873
332	casket	162	4.86	0.486860	1820
333	tunnel	912	4.82	0.486844	872
334	scissors	341	4.85	0.486811	4815
335	cork	146	4.86	0.486775	1936
336	tofu	137	4.86	0.486728	1276
337	zebra	128	4.86	0.486680	4449
338	kangaroo	118	4.86	0.486627	4034
339	hamster	109	4.86	0.486579	3071
340	missile	670	4.83	0.486559	2593
341	dishwasher	103	4.86	0.486547	1025
342	bamboo	80	4.86	0.486425	2549
343	altar	259	4.85	0.486376	3340
344	otter	69	4.86	0.486367	4661
345	nacho	67	4.86	0.486356	1267
346	calculator	66	4.86	0.486351	4195
347	fence	819	4.82	0.486350	4975
348	rhubarb	55	4.86	0.486292	1974
349	stethoscope	48	4.86	0.486255	1516
350	library	1170	4.80	0.486215	3303
351	tadpole	30	4.86	0.486159	1432
352	dollhouse	22	4.86	0.486117	1894
353	trashcan	19	4.86	0.486101	1607
354	guava	13	4.86	0.486069	2038
355	pomegranate	13	4.86	0.486069	2657
356	tamale	12	4.86	0.486064	1389
357	anteater	11	4.86	0.486058	3981
358	plum	174	4.85	0.485924	1486
359	oyster	156	4.85	0.485829	2797

Continuation of Table 1					
category index	category name	SUBTLEX US freq.	concreteness	FCI	# of images
360	cinnamon	152	4.85	0.485807	2086
361	bandage	146	4.85	0.485775	3133
362	elevator	1245	4.79	0.485613	1066
363	mistletoe	98	4.85	0.485521	2274
364	marshmallow	77	4.85	0.485409	1056
365	custard	64	4.85	0.485340	3760
366	wristwatch	59	4.85	0.485313	1660
367	corkscrew	57	4.85	0.485303	1512
368	kazoo	31	4.85	0.485165	904
369	kebab	31	4.85	0.485165	1758
370	granola	30	4.85	0.485159	948
371	gargoyle	27	4.85	0.485143	1911
372	scone	25	4.85	0.485133	1867
373	mantis	23	4.85	0.485122	2641
374	parsnip	4	4.85	0.485021	856
375	curtain	525	4.82	0.484789	2418
376	scorpion	136	4.84	0.484722	1300
377	crown	698	4.81	0.484708	2206
378	lemonade	281	4.83	0.484493	1505
379	wolf	1034	4.79	0.484492	3966
380	bugle	88	4.84	0.484467	4817
381	graveyard	272	4.83	0.484445	2622
382	tumbleweed	22	4.84	0.484117	1302
383	plate	1308	4.77	0.483948	2173
384	dragonfly	145	4.83	0.483770	2111
385	flag	892	4.79	0.483738	1598
386	crocodile	115	4.83	0.483611	3261
387	mushroom	109	4.83	0.483579	4923
388	beetle	105	4.83	0.483558	4961
389	cucumber	101	4.83	0.483536	1741
390	sloth	74	4.83	0.483393	2738
391	dough	810	4.79	0.483302	2307
392	sphinx	52	4.83	0.483276	1864
393	canyon	418	4.81	0.483220	2102

Continuation of Table 1					
category index	category name	SUBTLEX US freq.	concreteness	FCI	# of images
394	iguana	40	4.83	0.483212	3018
395	chalice	32	4.83	0.483170	1524
396	doormat	28	4.83	0.483149	1110
397	hairpin	18	4.83	0.483096	2134
398	aloe	15	4.83	0.483080	4531
399	scaffolding	12	4.83	0.483064	1726
400	platypus	3	4.83	0.483016	1022
401	brownie	180	4.82	0.482956	825
402	casino	1039	4.77	0.482519	1527
403	shrimp	444	4.80	0.482358	3364
404	grate	52	4.82	0.482276	2412
405	loudspeaker	47	4.82	0.482250	4906
406	tower	1165	4.76	0.482188	4899
407	submarine	362	4.80	0.481923	2891
408	rug	531	4.79	0.481820	4677
409	ape	493	4.79	0.481619	2516
410	banner	302	4.80	0.481604	1273
411	syringe	99	4.81	0.481526	2228
412	hanger	69	4.81	0.481367	2212
413	cannon	444	4.79	0.481358	3642
414	kale	28	4.81	0.481149	2205
415	pothole	25	4.81	0.481133	964
416	chili	382	4.79	0.481029	2160
417	waterspout	5	4.81	0.481027	890
418	stadium	312	4.79	0.480657	3653
419	spacecraft	115	4.80	0.480611	1597
420	boar	111	4.80	0.480590	2068
421	cheese	1991	4.70	0.480575	4826
422	celery	95	4.80	0.480505	1215
423	hammer	636	4.77	0.480378	3694
424	matchbook	51	4.80	0.480271	1234
425	coconut	234	4.79	0.480243	2438
426	beet	14	4.80	0.480074	3154
427	stegosaurus	5	4.80	0.480027	1010

Continuation of Table 1					
category index	category name	SUBTLEX US freq.	concreteness	FCI	# of images
428	chocolate	1499	4.72	0.479962	3949
429	muffin	297	4.78	0.479578	3628
430	turnip	88	4.79	0.479467	2058
431	wire	1403	4.72	0.479452	3476
432	chandelier	72	4.79	0.479382	1968
433	forklift	49	4.79	0.479260	1789
434	fondue	47	4.79	0.479250	3425
435	gazebo	46	4.79	0.479244	1864
436	wheelbarrow	31	4.79	0.479165	1645
437	melon	218	4.78	0.479158	4512
438	earpiece	27	4.79	0.479143	2420
439	paintbrush	27	4.79	0.479143	1300
440	bib	25	4.79	0.479133	2013
441	strongbox	16	4.79	0.479085	3333
442	steamroller	12	4.79	0.479064	1310
443	breadfruit	6	4.79	0.479032	1542
444	dishrag	5	4.79	0.479027	1649
445	anchor	378	4.77	0.479008	1945
446	fountain	352	4.77	0.478870	2613
447	parachute	162	4.78	0.478860	3052
448	sink	863	4.74	0.478584	3101
449	radiator	103	4.78	0.478547	2782
450	burner	93	4.78	0.478494	2292
451	llama	72	4.78	0.478382	4258
452	playground	260	4.77	0.478381	3848
453	okra	26	4.78	0.478138	2786
454	gramophone	16	4.78	0.478085	2846
455	burlap	11	4.78	0.478058	2451
456	earwig	5	4.78	0.478027	2037
457	calipers	1	4.78	0.478005	2896
458	spinach	130	4.77	0.477691	1672
459	couch	1197	4.71	0.477358	4592
460	parsley	43	4.77	0.477228	2701
461	koala	31	4.77	0.477165	2457

Continuation of Table 1					
category index	category name	SUBTLEX US freq.	concreteness	FCI	# of images
462	bobsleigh	12	4.77	0.477064	1768
463	highlighter	6	4.77	0.477032	1822
464	silverfish	4	4.77	0.477021	1138
465	greenhouse	112	4.76	0.476595	2400
466	blimp	53	4.76	0.476282	2092
467	tray	410	4.74	0.476178	1915
468	mandolin	24	4.76	0.476127	2779
469	spareribs	22	4.76	0.476117	2035
470	gecko	19	4.76	0.476101	2238
471	volcano	170	4.75	0.475903	1600
472	cabbage	148	4.75	0.475786	4473
473	kettle	143	4.75	0.475760	2249
474	antenna	122	4.75	0.475648	4809
475	panda	108	4.75	0.475574	2015
476	microchip	74	4.75	0.475393	3011
477	toolbox	64	4.75	0.475340	1194
478	weasel	250	4.74	0.475328	1465
479	hairbrush	37	4.75	0.475197	1788
480	squeegee	14	4.75	0.475074	866
481	flashbulb	3	4.75	0.475016	1337
482	pretzel	102	4.74	0.474542	2161
483	sawmill	21	4.74	0.474112	1354
484	joystick	15	4.74	0.474080	2388
485	persimmon	4	4.74	0.474021	3484
486	wand	157	4.73	0.473834	1373
487	graffiti	103	4.73	0.473547	2019
488	giraffe	76	4.73	0.473404	838
489	chinchilla	27	4.72	0.472143	2365
490	sundial	23	4.72	0.472122	1947
491	beaker	22	4.72	0.472117	1500
492	honeycomb	20	4.72	0.472106	1583
493	fishnet	11	4.72	0.472058	938
494	odometer	11	4.72	0.472058	1932
495	scallion	8	4.72	0.472042	1441

Continuation of Table 1					
category index	category name	SUBTLEX US freq.	concreteness	FCI	# of images
496	chess	380	4.70	0.472018	4018
497	photocopier	2	4.72	0.472011	3268
498	chair	2511	4.58	0.471338	4901
499	locker	815	4.67	0.471329	1150
500	newsstand	53	4.71	0.471282	1204
501	reef	204	4.70	0.471084	3453
502	cheetah	117	4.70	0.470621	2047
503	salsa	109	4.70	0.470579	1103
504	auditorium	80	4.70	0.470425	2610
505	extinguisher	78	4.70	0.470414	1364
506	battery	633	4.67	0.470362	3227
507	dildo	68	4.70	0.470361	723
508	chestnut	65	4.70	0.470345	3170
509	manhole	38	4.70	0.470202	2923
510	gooseberry	14	4.70	0.470074	1267
511	salamander	8	4.70	0.470042	4195
512	spearmint	6	4.70	0.470032	1409
513	hotplate	5	4.70	0.470027	864
514	moth	116	4.69	0.469616	4920
515	lizard	247	4.68	0.469312	4936
516	beaver	246	4.68	0.469307	2367
517	nutmeg	36	4.69	0.469191	1360
518	cannabis	23	4.69	0.469122	1181
519	cogwheel	2	4.69	0.469011	1219
520	bolt	351	4.67	0.468864	2443
521	hut	674	4.65	0.468580	4732
522	robot	621	4.65	0.468299	3519
523	shield	418	4.66	0.468220	1933
524	chameleon	37	4.68	0.468197	3301
525	defibrillator	24	4.68	0.468127	845
526	bison	17	4.68	0.46809	3621
527	geyser	11	4.68	0.468058	2147
528	raccoon	73	4.67	0.467388	3115
529	coleslaw	63	4.67	0.467335	1847

Continuation of Table 1					
category index	category name	SUBTLEX US freq.	concreteness	FCI	# of images
530	thermos	57	4.67	0.467303	1477
531	macadamia	31	4.67	0.467165	812
532	compass	207	4.66	0.467100	2848
533	ruler	162	4.66	0.466860	1740
534	helicopter	806	4.62	0.466281	3081
535	manatee	10	4.66	0.466053	1645
536	pinwheel	7	4.66	0.466037	1820
537	cherry	693	4.62	0.465681	4362
538	pickle	235	4.64	0.465248	1357
539	projector	62	4.64	0.464329	4628
540	guillotine	42	4.64	0.464223	920
541	tapioca	31	4.64	0.464165	1167
542	birdcage	24	4.64	0.464127	1320
543	coffeepot	21	4.64	0.464112	3052
544	sharpener	17	4.64	0.464090	1589
545	baguette	14	4.64	0.464074	1440
546	gearshift	11	4.64	0.464058	1702
547	pier	334	4.62	0.463774	1576
548	drain	440	4.61	0.463337	1258
549	artichoke	28	4.63	0.463149	2371
550	oscilloscope	4	4.63	0.463021	1908
551	cube	152	4.62	0.462807	2400
552	stapler	44	4.62	0.462234	1737
553	bubble	408	4.60	0.462167	3882
554	ukulele	29	4.62	0.462154	1811
555	tyrannosaurus	20	4.62	0.462106	1255
556	winterberry	1	4.62	0.462005	805
557	cauldron	24	4.61	0.461127	1028
558	cassette	93	4.60	0.460494	2583
559	meteorite	41	4.60	0.460218	990
560	urinal	39	4.60	0.460207	770
561	hazelnut	9	4.60	0.460048	907
562	rainbow	407	4.57	0.459162	2085
563	sieve	27	4.59	0.459143	1916

Continuation of Table 1					
category index	category name	SUBTLEX US freq.	concreteness	FCI	# of images
564	shredder	17	4.59	0.459090	1007
565	hovercraft	26	4.58	0.458138	1590

C | Relating representational consistency and IT prediction

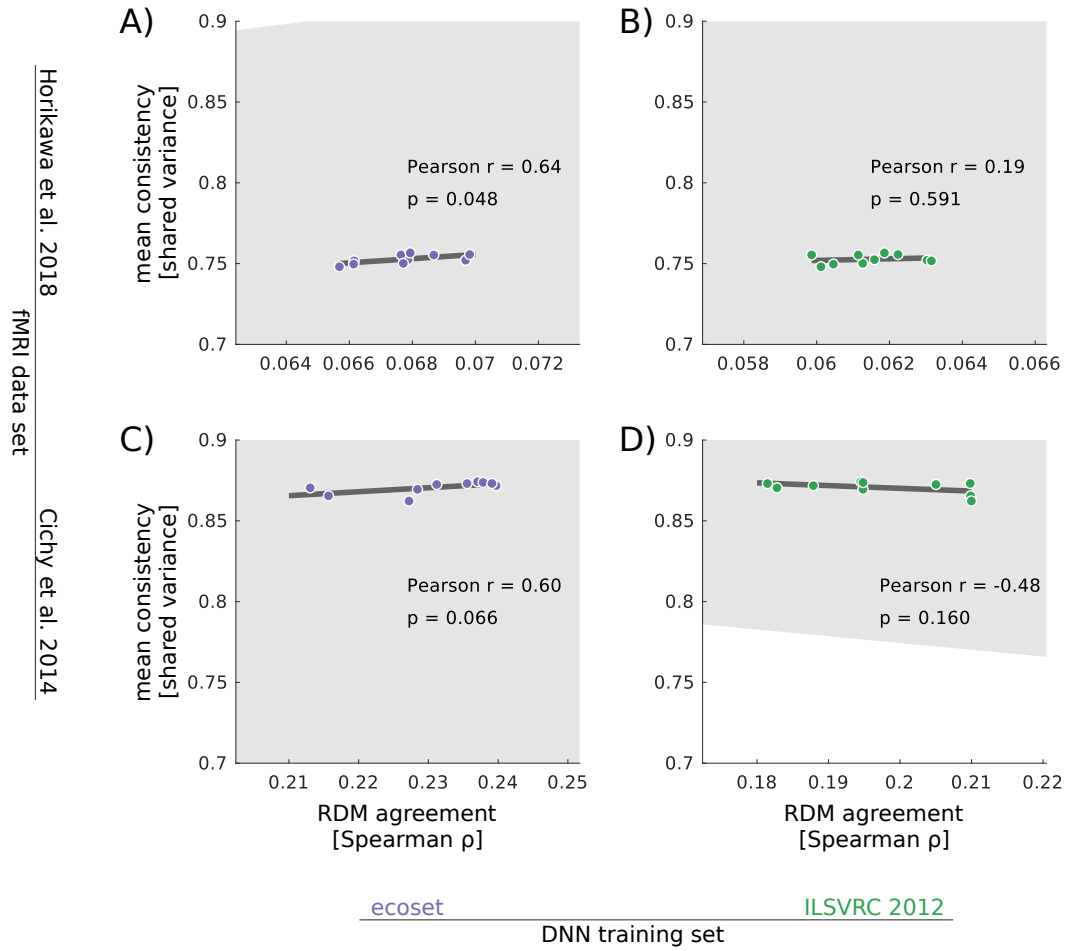


Fig. 5 **Appendix C: Representational consistency and IT prediction.** We here investigated the relationship between representational consistency (see chapter 2) and the ability of models to predict IT representations. We ask whether a DNN's representational consistency with other instances may be able to predict its ability to match cortical representations in human IT. For each DNN instance we computed 1.) the consistency to all other other instances using the squared Pearson correlation as in (link to chapter 2 where consistency is defined) and averaged across all 9 pairwise comparisons, and 2.) the IT-prediction between this instance and all subjects using Spearman's and averaged across all 15 (data set 1, Horikawa et al. 2018) or 5 (data set 2, Cichy et al. 2014) subjects. Next, we correlated the mean consistency and mean IT-prediction values using Pearson correlation (95% bootstrapped confidence interval in grey). We performed this analysis on both fMRI data sets (data set 1 - upper row, data set 2 - lower row) and for DNNs trained on either ecoset or ILSVRC 2012 (left and right column, respectively). We found a significant correlation only for data set 1 when DNNs were trained on ecoset (**A**). In the three other cases (**B, C, D**) no significant correlation was found. To conclude, we did not find strong evidence for a relationship between representational consistency and a DNN's ability to predict cortical representations.