12-16-2021 9:00 AM

Learning object representations in deep neural networks

Ehsan Tousi, *The University of Western Ontario*

Supervisor: Mur, Marieke, *The University of Western Ontario*
Co-Supervisor: Daley, Mark, *The University of Western Ontario*
A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Neuroscience

Abstract

Humans have the ability to learn visual representations of the surrounding environment with limited supervision. A major challenge in cognitive neuroscience is to understand the neural computations that give rise to this ability. Recent work has started modelling the neural computations implemented by the ventral visual system using deep convolutional neural networks (DCNNs). Despite their successes, DCNNs leave substantial amounts of variance in brain representations unexplained. We hypothesize that this may in part be due to the DNNs' sole reliance on supervision during representation learning. In this thesis, we investigate the role of training algorithms (supervised versus unsupervised) on the representational similarity between the computational models and brain data from human inferior temporal cortex. We show that one implementation of unsupervised contrastive learning yields more brain-like representations than the selected supervised learning method. Our findings suggest that human visual representations may in part arise from unsupervised learning during development.

Keywords: object recognition, human visual cortex, functional magnetic resonance imaging, representation learning, deep convolutional neural networks, unsupervised learning

Summary for Lay Audience

When we open our eyes, we instantly recognize the visual world around us. How does the brain so quickly make sense of the outside visual world? To address this question, we need to build computational models of the human visual system. Recent advances in deep learning have enabled the development of computational models that can perform real-life tasks such as object recognition. Like humans, the models need to 'develop' over a period of extensive learning. In this thesis, we examine the impact of learning goals on how the computational models learn to represent the outside visual world. We test whether certain learning goals give rise to more human-like object representations than others. We focus on one implementation of unsupervised learning - like a child discovering the world on their own - and one implementation of supervised learning - like a parent pointing at objects and naming them. We show that unsupervised learning gives rise to object representations that emphasize categories of behavioural relevance, including faces and animals. Furthermore, object representations learned through unsupervised learning show a closer match to human object representations than those learned through supervised learning. Our findings are consistent with the idea that unsupervised learning plays a role in object learning during human development.

Dedication

This thesis is dedicate to my late friend, Milad Ghasemi Ariani (1987- 2020).

On the 8th of January of 2020, Milad lost his life on his trip to Canada, when his plane was shot down by the Islamic Revolutionary Guard Corps.

Milad you were my companion from childhood, sharing every stage of our lives together. The memories we made will remain with me for the rest of my life.

Acknowledgements

I would like to express my most sincere gratitude to my supervisor, Dr. Marieke Mur, for her extraordinary supervision. Her outstanding support, especially during the most pressing times always went above and beyond in ensuring my well-being. Dr. Mur not only inspired me with her extensive knowledge and insight but also modelled the ideal scientist by always displaying a deep pleasure and passion for research. It was an absolute honour to work with Dr. Mur. I also want to thank Dr. Mark Daley for his continuous and unwavering support. His guidance was crucial in inspiring the direction of my thesis.

I would like to thank my advisory committee members for giving input on my project and providing me with their kind encouragement. I would also like to thank my program representative, Dr. Arthur Brown, for providing his encouragement and comments. He always made himself available whenever I needed support, helping me develop my scientific communication skills.

The completion of my thesis would not have been possible without the guidance and feedback of my beloved sister, Tahereh, my brothers, Hossein and Ali, and my dear friend and colleague Haider Al-Tahan. Our chats were invaluable in the development of my project and the resolution of challenges that I encountered.

I cannot express how critical my parents' support was in my success, nor the depth of my appreciation for their constant sacrifices for me. Throughout my life, our parents always prioritized my education while ensuring that all my needs were met. Despite studying overseas, I could always feel their love and encouragement through the difficulties of the last two years.

A special thanks to my friends and loved ones, Fatemeh Ahmadi, Ladan Shahshahani, David Mekhaiel, and Geetika Gupta. I could not have remained motivated, especially through the loneliness brought on by pandemic restrictions, without the pleasure of their company.

Contents

List of Figures

List of Tables

List of Abbreviations

- **DCNN** Deep Convolutional Neural Network
- **IT** Inferior Temporal
- **RSA** Representational Similarity Analysis
- **RDM** Representational Dissimilarity Matrix
- **BN** Batch Normalization
- **ReLU** Rectified Linear Units
- **fc** Fully Connected
- **MoCo** Momentum Contrast
- **ResNet** Residual Neural Network
- **VIF** Variance Inflation Factor

Chapter 1

Introduction

1.1 Motivation

When we open our eyes, waves of neural activity sweep through our visual system. We become aware of our surroundings and recognize the objects that populate our visual world. Somehow, the brain transforms the incoming visual signals into meaningful representations. Despite the ease with which we recognize objects, the computational task performed by the brain is far from trivial (Marr, 1982). Computer vision has only recently started to provide computational models that can recognize objects in real-world visual scenes with human-level performance (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014; He et al., 2016). These computational models are artificial neural networks that are inspired by the primate visual system and trained to perform real-life object recognition tasks. While the models are not explicitly trained to simulate brain information processing, i.e. they are only trained to perform a behavioural task, they predict brain activity across the primate visual system during object perception (Khaligh-Razavi and Kriegeskorte, 2014; Yamins et al., 2014; Güçlü and van Gerven, 2015). Importantly, they do better than previous computational models of the visual system (Khaligh-Razavi and Kriegeskorte, 2014; Yamins et al., 2014). However, the models still leave substantial parts of

the variance in brain responses unexplained (Storrs et al., 2020; Yamins et al., 2014; Schrimpf et al., 2018; Mell et al., 2021).

How can we close the gap between artificial neural networks and human brains? Artificial neural networks, when used as models of brain computation, abstract from real brains in substantial ways (Kriegeskorte and Douglas, 2018; Kietzmann et al., 2018). This implies that the researcher faces a range of design decisions when building artificial neural networks (Richards et al., 2019). Design components include learning goals, learning rules, network architecture, and training data (Kriegeskorte and Douglas, 2018; Richards et al., 2019). Learning goals express *what* the network is supposed to learn, learning rules specify *how* the network learns, the architecture determines how information can flow through the system, and the training data provide the network with 'experience'. Design choices influence the ability of the networks to predict human brain activity and behaviour, and there is room for improvement in each of the four design areas (Yamins and DiCarlo, 2016; Lillicrap et al., 2020; Kietzmann et al., 2019; Mehrer et al., 2021). Given the importance of learning goals in shaping the networks' internal representations (Marr, 1982; Grill-Spector and Weiner, 2014; Yamins and DiCarlo, 2016; Kriegeskorte and Douglas, 2018), the current thesis focuses on learning goals. The aim of the thesis is to test whether more human-like learning goals yield a better match between brain and models.

The current chapter starts with an introduction to the human ventral visual system, with a focus on the system's learning goals and internal object representations. Next, it will cover the conceptual context and mathematical intuition for deep convolutional neural networks (DCNNs). DCNNs are a specific type of artificial neural network that is commonly used for modeling the human ventral visual system (Schrimpf et al., 2018). The last section of the chapter introduces representational similarity analysis (RSA) (Kriegeskorte et al., 2008a), an experimental and data analytical framework that allows for quantitative comparison of object representations between humans and DCNNs.

1.2 The ventral visual system

Humans rely heavily on visual information for understanding and navigating the external world. For instance, we use vision to find our car keys in the morning, to decide when to cross a busy road, or to identify a friend in a crowded restaurant. All of the above everyday activities require accurate recognition of the objects in our environment. While object recognition seems effortless, it poses serious computational challenges. For example, we need to recognize an object independent of variability in viewpoint, illumination, or spatial position (Biederman, 1987; Serre et al., 2007; DiCarlo and Cox, 2007). Furthermore, we need to group objects that require similar behavioural responses together, even if those objects differ in visual appearance. In other words, we need to assign objects to behaviourally relevant natural categories, such as animate objects and faces (Rosch et al., 1976; Grill-Spector and Weiner, 2014). Categorization enables efficient selection of behavioural responses to objects we encounter in our environment. Categorization also allows for appropriate behavioural responses to novel objects provided the object is categorized successfully (Edelman, 1997).

In sum, one major learning goal of the human visual system is to successfully identify an individual object across viewing conditions. A second major learning goal is to generalize among objects from the same category without losing the ability to distinguish between individual objects. To achieve these goals, the system needs to abstract from variability in visual appearance that is uninformative for object identification and categorization, while keeping or even emphasizing variability that *is* informative. In doing so, it needs to strike the right balance between categorization and individuation.

Object recognition is supported by the ventral visual system (Figure 1.1) (Ungerleider and Mishkin, 1982; Goodale and Milner, 1992). The ventral visual system is a set of hierarchically organized brain regions in occipital and temporal cortex that process visual input. The system consists of cortical visual areas V1, V2, V4, and inferior temporal (IT) cortex. The

system's importance for object recognition was established by lesion studies in both nonhuman primates and humans, for reviews see Ungerleider and Mishkin (1982); Goodale and Milner (1992). These studies showed that lesions to more advanced processing stages of the ventral visual system result in significant impairments in visual discrimination and recognition of objects, without affecting the ability to locate or grasp objects (Gross, 1973; Goodale et al., 1991). A reverse pattern of behavioural impairments has been observed after lesions to posterior parietal cortex, which receives visual input from V1 via visual areas V2, V3, middle temporal area MT, and medial superior temporal area MST (Pohl, 1973; Perenin and Vighetto, 1988). This double dissociation further underscores the unique role of the ventral visual system in object recognition. It also forms the basis of the influential two-systems hypothesis (Ungerleider and Mishkin, 1982), which poses that the human brain contains separate visual pathways for perception and action (Goodale and Milner, 1992).

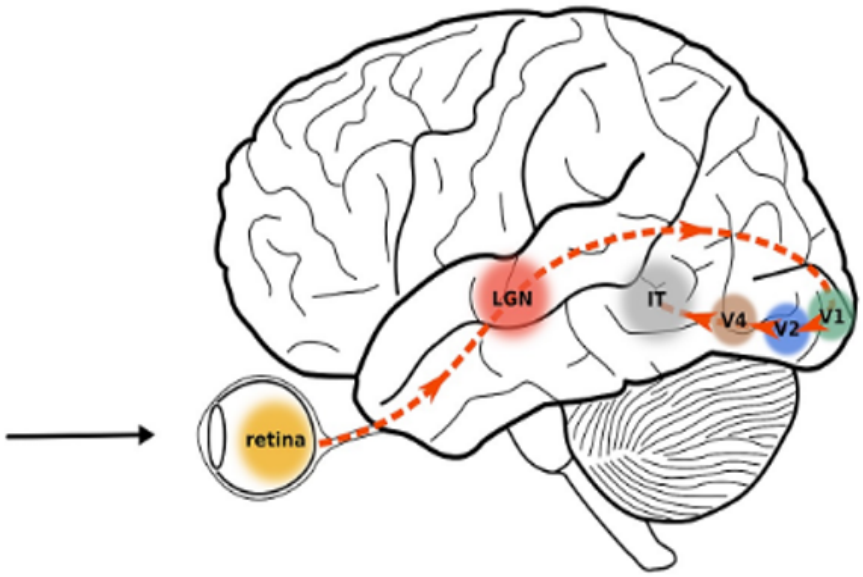

Figure 1.1: Schematic of the human ventral visual system. Light that hits the retina carries information about the external world. This information is processed by the ventral visual system, which consists of a set of hierarchically organized brain regions in occipital and temporal cortex. The system transforms the incoming visual signals into a meaningful representation of the external world. LGN = lateral geniculate nucleus, V = visual area, IT = inferior temporal cortex. Figure adapted from Kubilius (2017).

1.2.1 Functional properties of the ventral visual system

Over the decades surrounding and following the 'discovery' of the ventral visual system, a host of studies in both nonhuman primates and humans closely examined the functional properties of the ventral system in the healthy brain. The first of these studies were conducted by Hubel and Wiesel in the 1960s. They measured responses of neurons in macaque V1 to spots and patterns of light. They reported that V1 neurons preferentially respond to oriented bars of light or colours and that they have small receptive fields (<1 degree of visual angle) (Hubel and Wiesel, 1968). The receptive field of a visual neuron is defined as the area of the retina within which the action of light affects the neuron's firing. Subsequent electrophysiology studies in nonhuman primates found that the complexity of preferred stimuli increases when moving up the ventral cortical hierarchy, from gratings and combinations of orientations in V2 (Hegdé and Van Essen, 2000; Anzai et al., 2007), to texture and shape in V4 (Pasupathy and Connor, 2002; Kim et al., 2019) and object parts and categories in IT (Gross et al., 1972; Tanaka, 1996). Categorical response preferences appear to be strongest for categories of long-standing ecological relevance, including faces, body parts, and places (Desimone et al., 1984; Tsao et al., 2006; Bell et al., 2011). Furthermore, when moving up the ventral cortical hierarchy, receptive field sizes increase and neural responses to objects become more robust against changes in viewpoint, illumination, spatial position, and scale (Freeman and Simoncelli, 2011; Rust and DiCarlo, 2010; Tanaka, 1996; Hung et al., 2005).

The invention of functional magnetic resonance imaging (fMRI) in the early 1990s (Bandettini et al., 1992; Ogawa et al., 1992) enabled researchers to noninvasively examine the functional properties of the visual system at a considerably more extensive spatial scale. As predicted by Hubel and Wiesel in the 1960s, fMRI studies in humans revealed the co-existence of multiple functional maps in V1, including large-scale retinotopic maps and fine-grained orientation maps (Sereno et al., 1995; Yacoub et al., 2008). Furthermore, fMRI studies revealed the existence of functionally specialized regions at higher cortical stages of visual processing.

These regions include the lateral occipital complex, which as a whole responds more strongly to intact than scrambled objects (Malach et al., 1995; Grill-Spector et al., 1998) and a small number of category-selective regions, which respond more strongly to objects from their preferred than their non-preferred categories. The most well-known category-selective regions are the fusiform face area (FFA), the parahippocampal place area (PPA), and the extrastriate body area (EBA) (Kanwisher et al., 1997; Epstein and Kanwisher, 1998; Downing et al., 2001). The existence of functionally specialized regions for these categories is consistent with the electrophysiology findings previously reported in the nonhuman primate literature (Desimone et al., 1984; Tsao et al., 2006). Subsequent fMRI studies further confirmed the existence of category-selective regions in the nonhuman primate brain for faces, places, and body parts (Tsao et al., 2006; Rajimehr et al., 2011; Bell et al., 2011).

In the 2000s, studies in both humans and nonhuman primates showed that information about the category of a viewed object is also present outside of category-selective regions, in distributed patterns of activity across IT cortex (Haxby et al., 2001; Carlson et al., 2003; Kiani et al., 2007; Kriegeskorte et al., 2008b). These studies showed that activity patterns in IT cortex cluster according to natural categories of ecological relevance. The IT object representation shows a hierarchical category structure with a top-level division between animate and inanimate objects, and within the animates, a division between faces and bodies (Kiani et al., 2007; Kriegeskorte et al., 2008b). The distributed response patterns also contain information about object identity that is robust to changes in position, scale, and viewpoint (Hung et al., 2005; Eger et al., 2008; Anzellotti et al., 2013). The IT object representation matches between humans and nonhuman primates, both within and between categories (Kriegeskorte et al., 2008b). Furthermore, the IT object representation predicts perceived object similarity: object images that elicit similar activity patterns in IT cortex tend to be judged as similar by human observers (Mur et al., 2013). These findings suggest that IT cortex hosts an object representation that is at once categorical and continuous, that is shared between humans and nonhuman primates, and that may give rise to our conscious perceptual experience of visual objects. This object representation may arise from a low-dimensional feature map of object space laid out on the

cortical sheet, with clusters of similarly-tuned neurons corresponding to category-selective regions (Op de Beeck et al., 2008; Grill-Spector and Weiner, 2014; Bao et al., 2020).

1.2.2 Development of the ventral visual system

The studies reviewed so far suggest that the ventral visual system in healthy human adults successfully transforms images into meaningful object representations that can simultaneously support robust object identification and object categorization. This is exciting because it suggests that, at the adult stage of development, the ventral visual system has overcome the computational challenges associated with achieving its two major learning goals: robust object identification and categorization. This raises the question of how the IT object representation emerges during development.

Within days of birth, young infants preferentially look at faces (Livingstone et al., 2019). They also prefer moving stimuli over static stimuli (Livingstone et al., 2017). The early preference for faces remains present till at least four months of age, and then broadens to the entire category of animate objects by 10 months of age (Spriet et al., 2021). The first signs of robust object recognition have been reported as early as three months of age. Invariance to object rotation and size, as tested in habituation experiments, starts to develop around this time (Caron et al., 1979; Day and McKenzie, 1981). Invariance to viewpoint is slower to develop, and may take up to multiple years (Gliga and Dehaene-Lambertz, 2007; Nishimura et al., 2015). The first signs of object categorization have been reported as early as 18 months of age using sequential touching tasks (Bornstein and Arterberry, 2010). By three years of age, children are close to perfect at sorting objects into basic categories such as dogs, cats, cars, and trains (Rosch et al., 1976). By four years of age, they also start sorting objects into superordinate categories such as animals and vehicles (Rosch et al., 1976). These findings suggest that children develop the foundations for robust object identification and categorization within the first two years of life, and further develop these abilities over the years into young adulthood.

In IT cortex, preferential responses for faces over places have been reported as early as 1 month of age in nonhuman primates and 4-6 months of age in humans (Deen et al., 2017; Livingstone et al., 2017). These are comparable ages given the difference in life expectancy between macaques and humans. In both species, the spatial organization of these early face- and place-preferring regions is consistent with that seen in adolescence and adulthood (Deen et al., 2017; Livingstone et al., 2017). However, at this early age, responses are only weakly selective and the object representation in IT cortex is different from that in adulthood, showing little discrimination between categories other than faces and places (Deen et al., 2017; Livingstone et al., 2017). Stable and consistent selectivity for faces and clearly identifiable face-selective regions emerge around 6-7 months in macaques, which would correspond to approximately 2 years in humans (Livingstone et al., 2017). In both humans and nonhuman primates, the development of stronger category selectivity is mainly driven by a reduction in the response to objects from non-preferred categories, as opposed to an increase in the response to objects from the preferred category (Cantlon et al., 2010; Livingstone et al., 2017). This has been taken as evidence for pruning (Cantlon et al., 2010). Given that looking preferences precede the emergence of category-selective regions, it has been proposed that the former drives the latter. Evidence has been accumulating for the proposal that the visual system contains a retinotopic protomap at birth which is subsequently shaped by visual experience (Levy et al., 2001; Henriksson et al., 2015; Arcaro and Livingstone, 2017; Arcaro et al., 2017).

1.3 Deep convolutional neural networks

The studies reviewed in the previous section leave open how the ventral visual system solves the computational challenges associated with achieving robust object identification and categorization. Computational modelling work has demonstrated that the biological system accomplishes these tasks through a hierarchical cascade of linear and nonlinear image transformations that yield high-level representations of the visual environment (Serre et al., 2007; DiCarlo et al.,

2012; Yamins et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014). In this section we review DCNNs, which are currently considered the best computational models of human vision, with a focus on how we can leverage deep learning to test hypotheses about object representation learning in the brain.

1.3.1 Artificial neural networks

An artificial neural network is a group of interconnected artificial neurons that are conceptually inspired by biological neurons. Similar to the human visual cortex which consists of a set of hierarchically organized brain regions, the artificial neurons in neural networks are organized into multiple layers. A layer that receives input data is called the input layer while the output layer is the layer responsible for performing the end tasks, such as identifying objects in a set of images. Layers between input and output layers of neural networks are known as hidden layers. Once a neural network has more than one hidden layer, it is called a *deep* neural network (Kriegeskorte and Golan, 2019). Biological neurons can receive and transmit signals from and to other neurons via synapses, similarly, artificial neurons can receive and transmit signals from and to other artificial neurons via network connections. The inputs of artificial neurons can either be feature values extracted from the network's input data (for the neurons in the input layer), or they can be the outputs of other neurons (for the neurons in hidden or output layers). Furthermore, the most commonly used type of neural networks for computational modelling of human vision are *feedforward* neural networks, where information flows only in one direction, from input to output (Figure 1.2) (Goodfellow et al., 2016).

Each artificial neuron within a neural network computes its output (y_i) as the sum over the product between its inputs (x_i) and connections (through weights w_i) plus a bias term (b):

$$y_i = \sum_{i}^{n} x_i w_i + b \tag{1.1}$$

w_i determines the impact that the input exerts on the unit, thereby dictating the relative importance of each connection to the neuron.

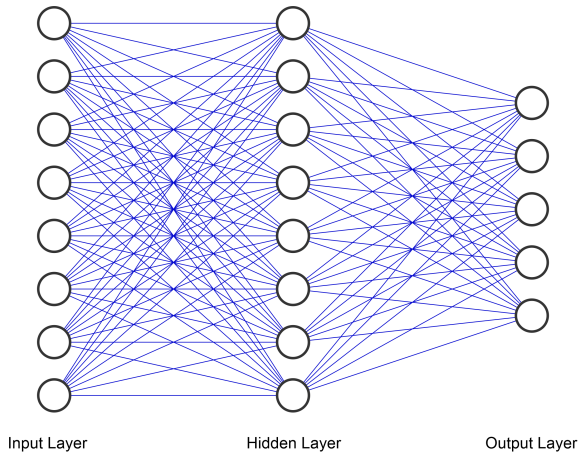

Figure 1.2: Illustration of feedforward neural network with one hidden layer

Current modelling efforts for understanding the human ventral visual system commonly use DCNNs, a particular type of feedforward artificial neural network (Kriegeskorte and Golan, 2019). DCNN model architectures are built from four basic layer types: (1) convolutional layers, (2) pooling layers, (3) activation layers and (4) fully connected layers. The following sections will describe these layer types in more detail.

Convolutional layers

Convolutional layers convolve the input image with filters and pass the result on to the next layer. The filters are weights that are learned by the DCNN during training on visual tasks. The filters can be thought of as visual feature detectors. The number of filters that a convolutional layer can learn are set by the experimenter. In each convolutional layer, artificial neurons are organized into feature maps. Artificial neurons within the same feature map learn the same filter. However, each artificial neuron applies that filter to a different part of the input image. That part of the image can be thought of as the artificial neuron's receptive field. Artificial neurons in convolutional layers thus simulate key properties of biological neurons in the visual system: they only respond to visual stimuli presented within their receptive field and they preferentially respond to specific visual features such as edges (Hubel and Wiesel, 1968). The convolutional layer is made up of three parts: input, kernels (also known as filters), and output (also known as feature maps). As the kernel moves over the input image, it performs an

element-wise multiplication with the part of the image that it covers at each step, and then sums the resulting values into a single output for that part of the image. Mathematically, convolution can be defined as follows:

$$(f * g)(i) = \sum_{m}^{M} g(m) \cdot f(i - m) \tag{1.2}$$

where f is a one-dimensional input, g is a one-dimensional kernel, and M is the size of the input. For every $i \in N$, where N is the size of input. The convolution operation can be generalized to higher dimensions, for example, a two-dimensional version of the operation can be obtained by convolving over two dimensions:

$$(f * g)(i, j) = \sum_{m}^{M} \sum_{k}^{K} g(m, n) \cdot f(i - m, j - k) \tag{1.3}$$

where the two-dimensional input has the size of $(M \times K)$. Deep neural networks handling images as input typically feature two-dimensional convolutional layers. Figure 1.3 illustrates the two dimensional operation, where the left most matrix is a 4×4 input and the middle matrix is the kernel of the convolution.

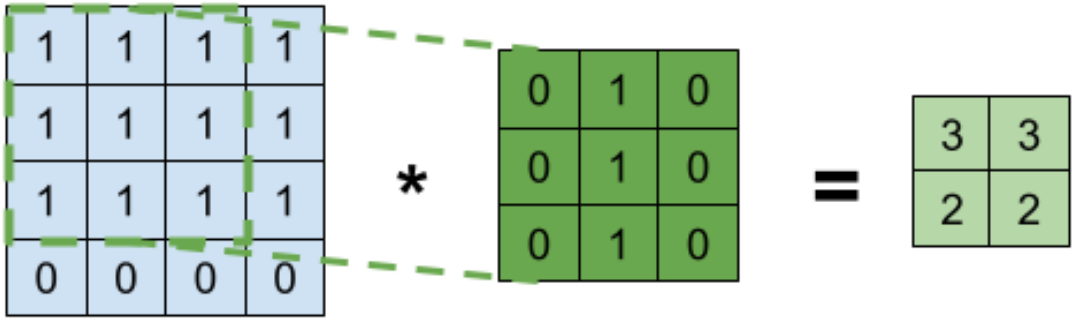

Figure 1.3: Schematic of two-dimensional convolution filter applied to an input matrix

Activation layers

Since the neuron's output is the weighted sum of the neuron's input, the output value can range from $-\infty$ to ∞. In most cases, it is desired to assign a binary activation state to the neurons so they can function as active or inactive units, similar to neurons in the brain. This can

be achieved by passing the weighted sum of the output through a non-linear activation function. Rectified Linear Units (ReLU) are commonly used for a variety of machine learning tasks and are one of the most popular activation functions (Goodfellow et al., 2016). ReLU apply the non-saturating activation function $f(x) = \max(0, x)$ to each neuron's activation. Non-linear activation functions give neural networks more expressive power, because they enable the networks to learn nonlinear input-output mappings (Kriegeskorte and Golan, 2019).

Pooling layers

To reduce the spatial dependency of features in a given layer, their values can be passed to a pooling layer. A pooling layer is essentially a downsampling operation that reduces the size of representations. Moreover, pooling reduces the number of computations in the network and prevents overfitting. Maximum pooling layers are the most common pooling layers used in the architecture of convolutional neural networks in computer vision (Goodfellow et al., 2016). A max pooling layer selects the maximum value from each chunk of input, where the chunk size is equal to the layer's kernel size (see Figure 1.4). The max pooling operation increases invariance to spatial position, which is also observed when moving up the ventral visual system (Hung et al., 2005; Serre et al., 2007; Rust and DiCarlo, 2010).

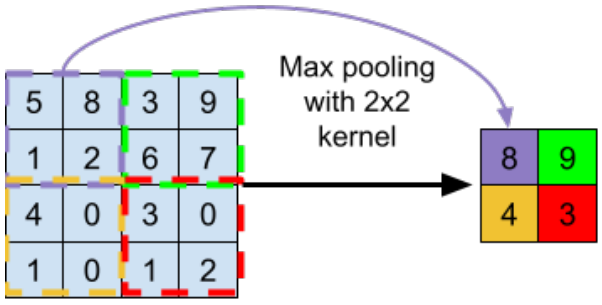

Figure 1.4: Schematic of two-dimensional maximum value pooling with a kernel size of two applied to an input matrix

Fully connected layers

When all neurons in a layer are connected to all neurons in the next layer, the layer is

referred to as fully connected. For most popular computer vision models, the final few layers are fully connected layers that collate the data extracted by previous layers to form the final output.

1.3.2 Training deep convolutional neural networks

DCNNs can be trained to perform real-life visual tasks, such as categorizing objects in images. During training the artificial neurons' weights and biases, also referred to as model parameters, are adjusted until the network achieves acceptable task performance. At the beginning of training, the model parameters are initialized with values drawn at random from a normal distribution (He et al., 2015). During training the network is iteratively presented with a set of inputs such as object images, and generates outputs such as category labels, given the current model parameters. In each iteration, the network's outputs are evaluated by a loss function, which is informed by the network's learning goals. The loss function measures the mismatch between the network's output and the desired output, which is referred to as the loss. The goal of training is to minimize the loss by adjusting the model parameters. One intuitive but inefficient way of minimizing the loss is to iteratively 'wiggle' the parameters and keep the adjustments that reduce the loss. One more efficient alternative is using gradient descent, which is an iterative optimization algorithm for finding a local minimum (Goodfellow et al., 2016).

Mathematically, gradient descent can be explained as follows. Let us assume that the multivariable function $\mathbf{F}(\mathbf{x})$ is defined and differentiable in the neighbourhood of point $\mathbf{a}$. Partial derivatives of $\mathbf{F}(\mathbf{x})$ are referred to as gradients of $\mathbf{F}(\mathbf{x})$. Gradient descent is built on the fact that at each point, $\mathbf{F}(\mathbf{x})$ will decrease the fastest in the direction of its negative gradients. Therefore, if in each iteration we set the next point, $\mathbf{a_{n+1}}$, to $\mathbf{a_n} - \alpha\nabla\mathbf{F}(\mathbf{a})$, then $\mathbf{F}(\mathbf{a_{n+1}})$ will be smaller than or equal to $\mathbf{F}(\mathbf{a_n})$ if we choose a small enough positive alpha. To put it simply, the negative term $\alpha\nabla\mathbf{F}(\mathbf{a})$ will move the point toward the local minimum. Conceptually, gradient descent can be explained by thinking of $\mathbf{F}(\mathbf{x})$ as a landscape of hills and valleys. The goal is to find the fastest way to the bottom of a nearby valley. Intuitively, the fastest way down corresponds to taking the

steepest descent.

In our case, the loss function is a multi-variable function of all model parameters and the training images. Hence, to apply gradient descent, we first need to calculate the partial derivatives of the loss function with respect to all the model parameters given all the training images. We then update the model parameters in each training step as described above: we multiply the gradients with a small positive number α and subtract the resulting terms from the current parameter values. Alpha is referred to as the learning rate, a hyperparameter that controls the strength of the gradients' influence on the update process.

DCNNs have millions of trainable parameters which require large training sets to achieve a desirable performance on their learning goal. Given currently available computing resources, the training sets are usually split up into subsets and parameter updates are being applied after each subset of training data rather than the whole training data. These subsets of the training data are called mini-batches. As a result, the optimization method is an approximation for gradient descent which is called stochastic gradient descent (SGD) (Goodfellow et al., 2016). Also, large DCNNs have too many parameters to feasibly compute the gradients of all parameters. Thus to implement SGD when training DCNNs, we use a method called *backpropagation* (Rumelhart et al., 1985). Backpropagation is based on the fact that in a DCNN, each parameter in a given layer is a function of the parameters of the preceding layers. Hence to avoid redundant computation, we can start computing the gradients at the last layer and then propagate them back down using the chain rule to derive the preceding layers' gradients. Once we compute all the gradients with backpropagation for a given mini-batch of the data, we can use SGD to update the parameters. We repeat this procedure until the algorithm converges to the minimum loss. It is important that we choose the learning rate to be large enough to not get stuck in local minimums, which will also result in a faster convergence. It is common practice to start with a larger learning rate, and then gradually decrease the learning rate during training (Goodfellow et al., 2016).

1.3.3 Supervised training

Deep learning algorithms commonly rely on a supervised approach to derive effective representations of the visual world. This approach can be considered analogous to the scenario of children learning about the visual world through explicit feedback from their parents. For example, a parent may correct a child's inaccurate categorization of an object in their environment. Similarly, supervised algorithms train the networks to learn a mapping between the input image and a category label. The underlying assumption with such an approach is that the learned latent representations in the hidden layers of the network carry effective representations for the designed tasks. Figure 1.5 illustrates a DCNN with an input image passing through multiple hidden layers which learn to transform the image into a representation meaningful to the task, for example a probability distribution across experimenter defined class labels such as "cat" or "dog". The class probabilities are computed by applying the softmax function to the final layer's activity pattern.

The most common loss function for supervised training is cross-entropy loss (Goodfellow et al., 2016). The cross-entropy loss measures the difference between the target probability distribution and the probability distribution predicted by the DCNN. In the case of binary classification, the cross-entropy loss can be formulated as $-(\mathbf{y}\log(\mathbf{p}) + (1-\mathbf{y})\log(1-\mathbf{p}))$ where y is the class indicator and p is the predicted class probability. The binary cross-entropy loss can be extended to multiple classes:

$$\mathcal{L}_{\mathbf{o}} = -\sum_{c=1}^{M} \mathbf{y}_{o,c} \log\left(\mathbf{p}_{o,c}\right) \tag{1.4}$$

where M is the number of classes, $\mathbf{y}$ is a binary indicator (0 *or* 1) that indicates whether class label c is the correct classification for observation o, and $\mathbf{p}_{o,c}$ is the predicted probability of classifying observation o in class c.

After training, DCNNs ideally correctly predict the category labels for images beyond the training data set. The most common benchmark for DCNNs in computer vision is object cat-

egorization performance on the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC), which has been held since 2010 (Deng et al., 2009; Russakovsky et al., 2015a). This competition uses a subset of ImageNet with approximately 1,000 images for each of the 1,000 predefined object categories. As a whole, there are about 1.2 million training images, 50,000 validation images, and 150,000 test images. Each image comes with a category label provided by human observers. With the aid of large human annotated datasets like ImageNet, DCNNs are now able to reach, and in some cases surpass, human-level performance at object categorization using purely supervised training (He et al., 2015; Schrimpf et al., 2018).

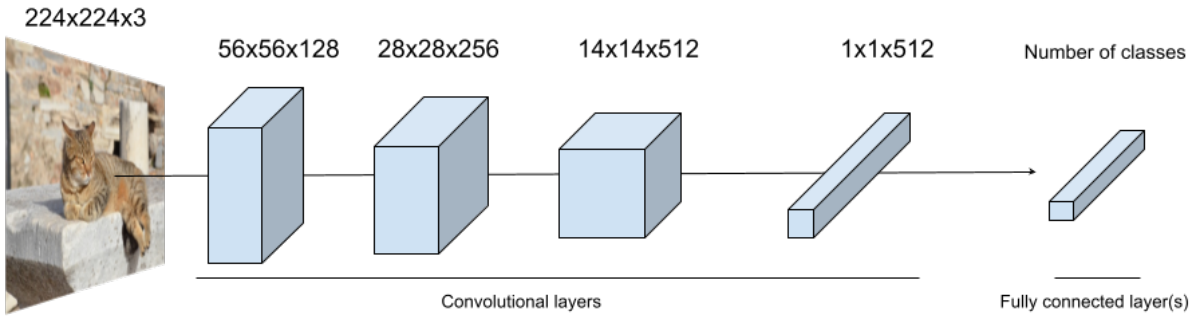

Figure 1.5: Schematic of an example deep convolutional neural network. Number of classes represents the number of object categories in the given data set.

1.3.4 Unsupervised training: contrastive learning

Supervised learning can work well, given a well-defined task and enough labelled input samples. However, good performance usually requires a large number of labelled samples, and collecting human-generated labels is challenging and expensive. As vast amounts of unlabelled data (e.g. text, images on the Internet) are available, it seems wasteful not to use them. Unlabelled data can be used for representation learning using unsupervised approaches. Unsupervised learning

is also thought to play an important role in human object learning (Zhuang et al., 2021; Konkle and Alvarez, 2021). Unsupervised approaches exploit the inherent structure in the data for learning visual representations. Some well-known approaches in this domain revolve around solving pretext tasks such as jigsaw puzzles (Noroozi and Favaro, 2016) and image colourization (Zhang et al., 2016) or generative modelling, for example using generative adversarial networks (Goodfellow et al., 2014). Recent work has started using self-supervised methods, which implement unsupervised learning via discriminative approaches, including contrastive learning methods (He et al., 2020; Chen et al., 2020). Contrastive learning methods are more successful at improving performance on visual recognition tasks than other unsupervised methods (He et al., 2020; Chen et al., 2020). Learning contrastive representations involves learning an embedding space in which similar input samples are represented close together and dissimilar ones are represented far apart.

Unsupervised contrastive learning aims to learn an embedding of the input space, where latent representations of two transformations of the same input sample (i.e. image) are close to each other while latent representations of two transformations of different input samples are far from each other. Examples of these transformations, which are called data augmentations, include randomly cropping a patch from the image and resizing to the original scale, randomly removing a patch from the image, introducing colour changes, introducing Gaussian noise, blurring the image, and rotating the image. Applying transformations such as colour-jittering, gray-scaling, horizontal flipping, and blurring on top of random cropping and resizing leads to better performance of unsupervised networks at visual recognition tasks (Chen et al., 2020).

Data augmentations are also used in supervised learning to increase invariance of the DCNNs to a small number of image variations, namely random cropping, resizing and horizontal flipping, but the augmentations are not essential to the training algorithm. Use of such augmentations relies on the assumption that a small distortion to an image should not make a difference to its semantic meaning. For example, if we take an image of a dog, flip it over the horizontal axis and grey-scale it, we expect the image to still represent a dog. A common

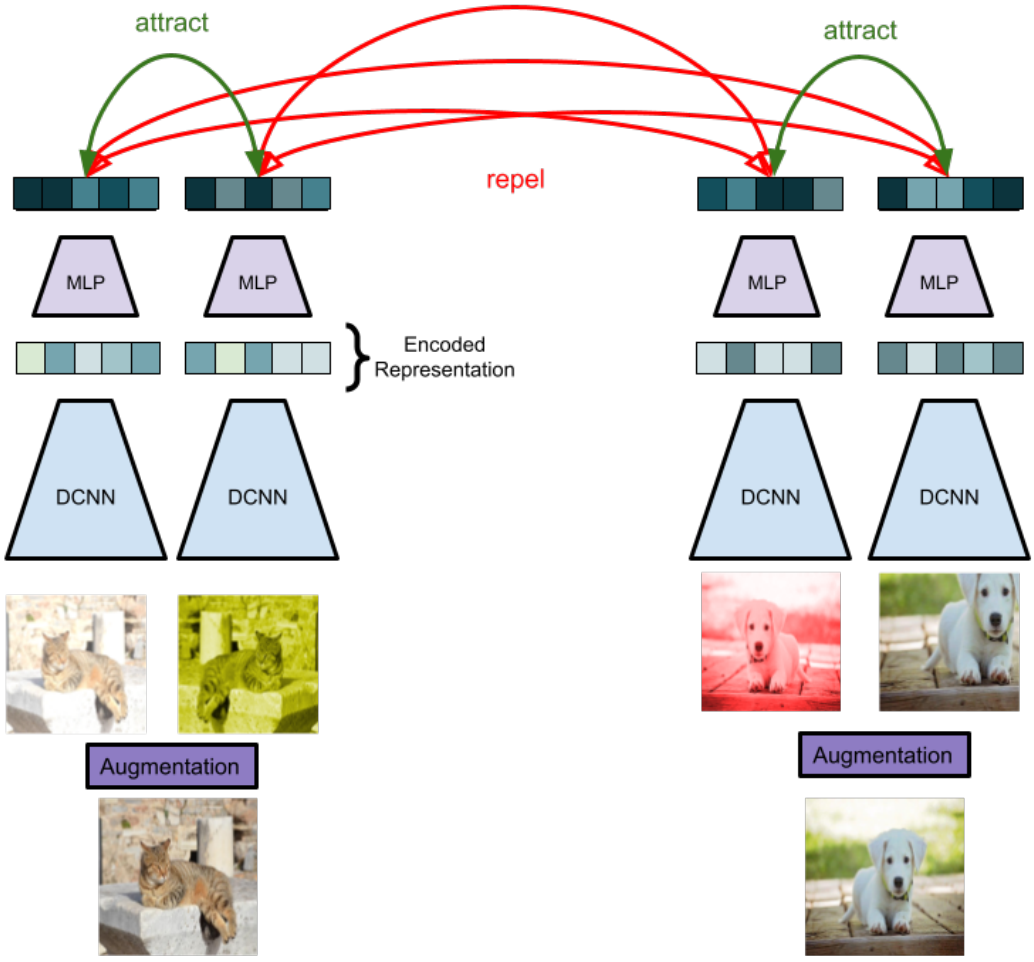

Figure 1.6: Schematic of the MoCo/SimCLR framework. Data augmentation is used to generate two different versions of an image. During training, DCNNs compute representations of the augmented images. The contrastive loss aims to minimize the representational distance between augmentations of the same image while maximizing the distance between augmentations of different images.

loss function for unsupervised contrastive learning is InfoNCE, which was introduced by van den Oord and colleagues as part of their Contrastive Predictive Coding (CPC) approach (Oord et al., 2018). Building on NCE (Gutmann and Hyvärinen, 2010), InfoNCE utilizes categorical cross-entropy loss to distinguish positive data points among a set of unrelated noise samples. This loss function is used in a number of unsupervised contrastive visual representation learning methods such as SimCLR, CPC, and MoCo (Chen et al., 2020; Oord et al., 2018; He et al., 2020). These methods, though they result from different motivations, can be viewed as dynamic dictionaries. Dictionary "keys" (tokens) are generated from samples of data (e.g. images) and represented through an encoder network. In an unsupervised fashion, encoders learn to perform dictionary look-up: an encoded "query" should be similar to its matching key and dissimilar to others. For MoCo, the infoNCE is defined as follows. $\mathbf{q}$ is the result of applying an encoder to an input sample, and a set of other encoded samples $\{k_0, k_1, k_2, ...\}$ are the keys of a dictionary. Let us assume that there exists a key in the dictionary that matches the query $\mathbf{q}$. The contrastive loss function measures how similar $\mathbf{q}$ is to its positive key $\mathbf{k}_+$ and dissimilar to all other keys (negative keys for $\mathbf{q}$). The similarity is measured by the InfoNCE loss:

$$\mathcal{L}_{\mathbf{q}} = -\log \frac{\exp(\mathbf{q} \cdot \mathbf{k}_+ / \tau)}{\sum_{i=0}^{K} \exp(\mathbf{q} \cdot \mathbf{k}_i / \tau)} \tag{1.5}$$

where τ is a temperature hyperparameter. The summation is performed over one positive and K negative samples. To generate the query and keys, MoCo uses two networks, an encoder network and a momentum encoder network. The two networks encode augmented versions of the same input images and the encoded representations are called queries and keys, respectively. The momentum encoder parameters θ_k are a moving average of the encoder parameters θ_q and are updated at each training step: $\theta_k \leftarrow m\theta_k + (1 - m)\theta_q$ where $m \in (0, 1)$ is the hyperparameter that dictates the degree of change. There is no gradient flowing through the momentum encoder. As depicted in Figure 1.7, the loss measures the log loss of a $(K + 1)$-way softmax-based classifier that aims to classify input query $\mathbf{q}$ as positive key $\mathbf{k}_+$.

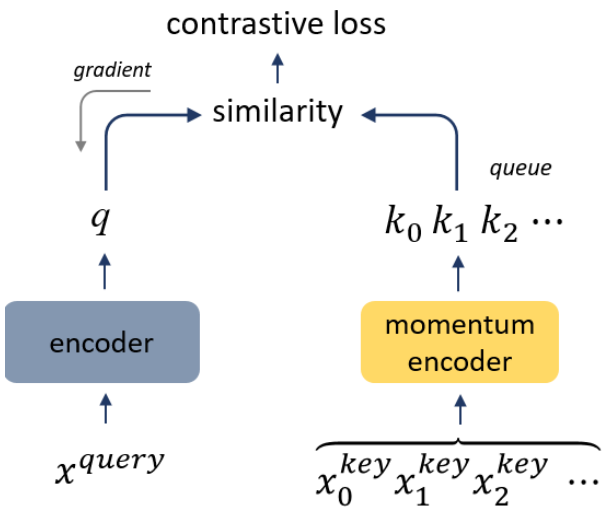

Figure 1.7: Momentum contrast learns a visual representation in an unsupervised manner by matching the encoded representation **q** of input query x^q to a dictionary of encoded keys $\{k_0, k_1, k_2, ...\}$ using a contrastive loss. The dictionary keys are generated while training on a set of input samples. Dictionary keys are encoded by a gradually progressing encoder, driven by a momentum update of the query encoder. They are built as a queue, with the current mini-batch enqueued and the oldest mini-batch dequeued. By using this method, a large and consistent dictionary can be utilized for learning visual representations.

1.3.5 Residual network (ResNet)

ResNets, or Residual Networks, were introduced in 2015 by He and colleagues (He et al., 2016). In that year, ResNets won first place in multiple large-scale object recognition challenges, including ILSVRC. Since then, ResNets have become the de facto standard computer vision models for object recognition tasks. Their pretrained ILSVRC versions have been incorporated as the backbone model for many other vision tasks including semantic segmentation and video action recognition. ResNets are also among the best models for predicting brain activity in the primate ventral visual system in response to images (Schrimpf et al., 2018) and feature prominently in contrastive learning applications (Chen et al., 2020; He et al., 2020). We therefore use a ResNet architecture in this thesis to simulate visual information processing in the human

brain.

The main contribution of the ResNet paper was the introduction of residual blocks to the standard feedforward architecture of DCNNs. Prior to the introduction of residual blocks, the performance of DCNNs at visual tasks would start to degrade after increasing the number of hidden layers beyond a certain depth (around 25-30 layers) (He et al., 2015, 2016). In order to resolve this problem, He and colleagues introduced residual blocks, which consist of a few convolutional layers and a skip-connection. The skip-connection adds the input arriving at the first layer of the block to the output of the last layer of the block. Skip-connections ensure that adding extra layers does not degrade the performance of the network since they enable the network to learn identity mappings (Figure 1.8). As a result, more layers can be added to the network which increases the overall performance of the model.

Another key component of the ResNet architecture is the batch normalization layer that follows each convolutional layer. In short, batch normalization layers normalize activations of hidden layers by changing the distribution of the activations to a normal distribution (i.e. mean of 0 and standard deviation of 1). This can be accomplished by subtracting the mean:

$$\begin{aligned} \mu &= \frac{1}{N}\sum_{i}^{N} x_i \\ x &\leftarrow x - \mu \end{aligned} \tag{1.6}$$

and normalizing by the variance:

$$\begin{aligned} \sigma^2 &= \frac{1}{N}\sum_{i}^{N}(x_i - \mu)^2 \\ x &\leftarrow \frac{x}{\sigma^2} \end{aligned} \tag{1.7}$$

where μ is the mean across the mini-batch, x is the activations from the preceding layers, and σ is the standard deviation across the mini-batch. Batch normalization preserves the range of activation values, thereby mitigating instances where a small number of features dominate the feature space by having large ranges in values (Ioffe and Szegedy, 2015). This in turn 'stabilizes' the weight updates during backpropagation as the weights do not require

modification to accommodate for different input distributions (Bjorck et al., 2018). Batch normalization has been shown to yield better model generalizability, performance, and faster training speed (Santurkar et al., 2018).

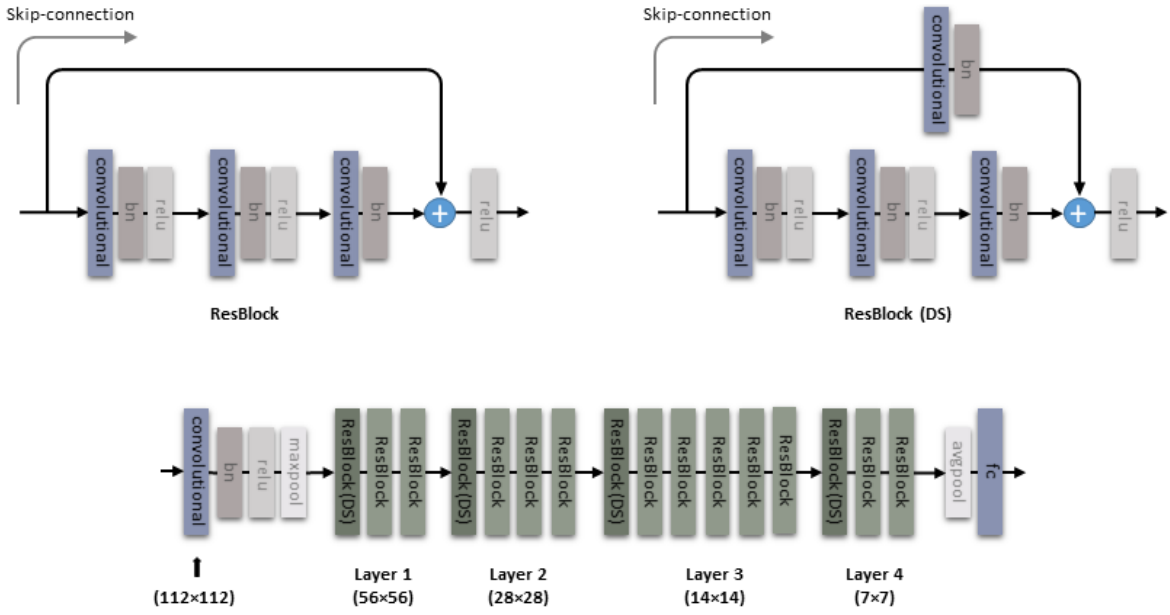

Figure 1.8: Illustration of the ResNet50 architecture. *ResBlock* shows the identity block which is the core building block of every ResNet architecture. The block consists of multiple repetitions of the following operations: convolution, batch normalization, nonlinear activation function (ReLU = rectified linear unit). The skip-connection adds the input of the the block to the output of the block. This essentially implements an identity mapping and allows the block to learn a residual function in reference to its input (He et al., 2016). *ResBlock (DS)* shows how down-sampling is implemented in a residual block to reduce the output size from one layer to the next. The full ResNet50 architecture is shown at the bottom, and indicates how blocks are combined into 'layers'. Images are presented to the first convolutional layer on the left, and the network's output is the activity pattern across the fully connected (fc) layer on the right. The output size of each layer is shown below the layers.

1.4 Representational similarity analysis

To test competing computational models of the human ventral visual system, we need to set benchmarks and develop tools to assess performance of the models at these benchmarks. Given that our aim is to develop better models of the ventral visual system, our most important benchmark is to predict brain activity in humans during object perception.

In this thesis, we use an existing fMRI data set as a benchmark (Kriegeskorte et al., 2008b; Mur et al., 2013). The data were acquired in healthy human adults while they were viewing coloured photographs of real-world object images from a range of ecologically relevant categories. fMRI is a noninvasive measurement technique that indirectly measures neural activity by measuring oxygen demand through the blood-oxygen-level-dependent (BOLD) signal (Bandettini et al., 1992). Although fMRI does not directly measure neural activity, the BOLD signal correlates with neural activity measured with in vivo cell recording techniques, especially local field potentials (Logothetis et al., 2001). fMRI has a spatial resolution in the millimetre range and a temporal resolution in the range of seconds (Goebel, 2007). The measurement units of fMRI are voxels, which are 3D pixels that usually are 2-3 mm cubic in size. fMRI has been used extensively to investigate the functional properties of the ventral visual system (Sereno et al., 1995; Malach et al., 1995; Kanwisher et al., 1997; Haxby et al., 2001; Kriegeskorte et al., 2008b). In this thesis, we will focus on predicting fMRI data from IT cortex, which can be considered the final processing stage of the ventral visual system.

To assess performance of the neural network models at predicting fMRI data from visual cortex, we need to relate the models to the brain data. This brings us to the correspondence problem: although the models are inspired by the brain, they are abstractions, and it is not clear how model units map onto fMRI voxels (Kriegeskorte et al., 2008a). In other words, we can present models and humans with the same images, we can measure the activity patterns elicited by these images across model units and brain voxels, but how can we determine whether the

activity patterns are similar? One solution is to abstract from the activity patterns and investigate dissimilarities between activity patterns instead. This obviates the need for correspondence between model units and voxels. Instead of relating the models and the brain at the level of activity patterns, we relate them at the level of representational geometry (Kriegeskorte et al., 2008a). This idea has a long history in psychology and neuroscience (Shepard, 1980; Edelman, 1998; Op de Beeck et al., 2001) and conceptualizes the activity patterns as points in a high-dimensional space spanned by the model units or voxels. Different images elicit different activity patterns and will 'live' at different locations in the high-dimensional response space. The distances between activity patterns reflect their dissimilarities and define a representational geometry that can be compared between models and the brain.

These ideas form the basis for representational similarity analysis (RSA) (Kriegeskorte et al., 2008a). RSA is an experimental and analytical framework for connecting models, brain activity data, and behaviour. In this thesis, we use RSA as a tool for assessing performance of neural network models at predicting fMRI data from visual cortex. We compute representational dissimilarity matrices (RDMs) for brain and models by extracting the activity patterns elicited by a set of object images and computing pairwise dissimilarities between the activity patterns. RDMs capture representational geometry, showing which stimulus information is emphasized and which is de-emphasized by a model or brain region. We can now quantitatively compare brain and model representations by correlating RDMs (Figure 1.9). Statistical inference on individual model performance or model comparisons is performed using random-effects analyses across subjects or randomization and bootstrap tests for small subject samples.

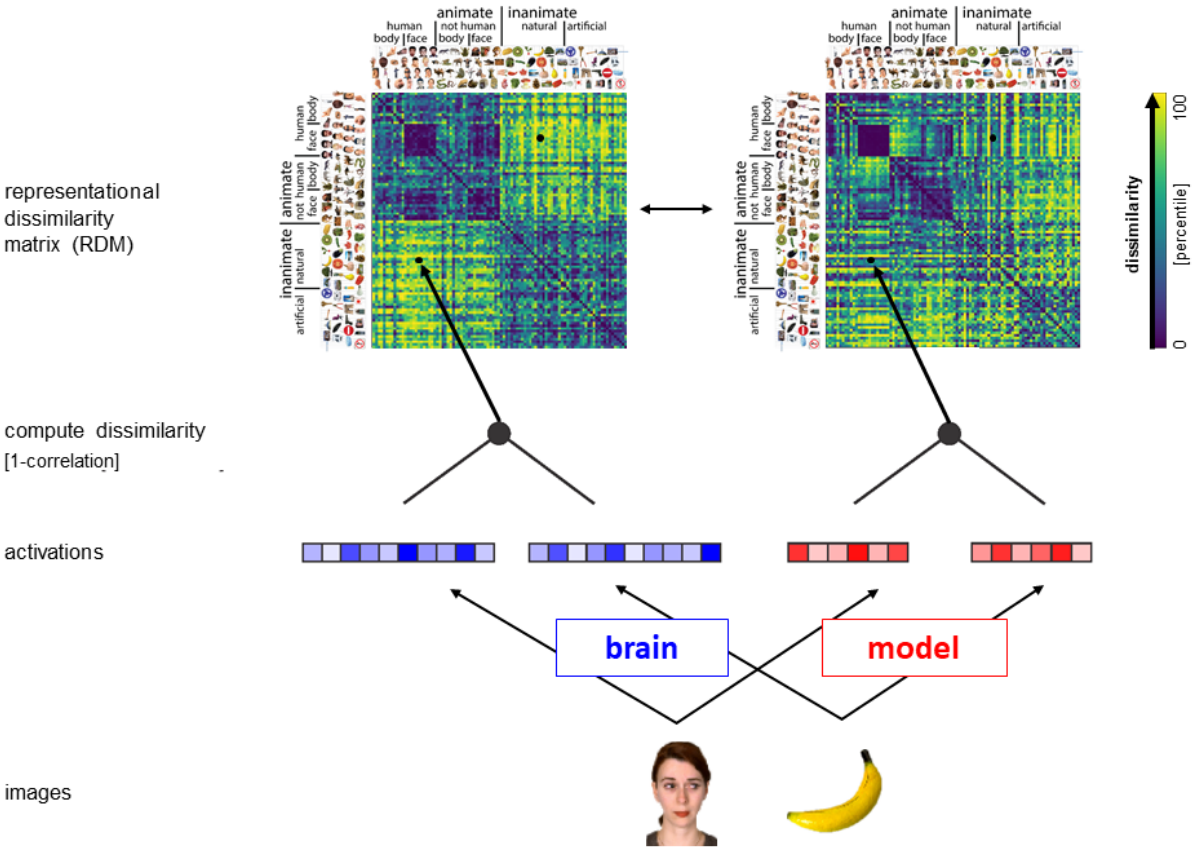

Figure 1.9: Illustration of the process of creating an RDM using brain or model activity patterns. For each pair of experimental stimuli, we extract activity patterns across a brain region or model layer of interest. We compute the dissimilarity between the activity patterns and store these in an RDM. The stimuli are shown on the axes and sorted by category. Each cell in the RDM represents the activity pattern dissimilarity between two stimuli. Dissimilarities are colour coded, using a spectrum of blue (lowest dissimilarity) to yellow (highest dissimilarity). We can now quantitatively compare brain and model representations.

1.5 Thesis overview

Chapter 1 serves as a general introduction to the human visual system for object recognition, the computational models used to simulate this system, and the analysis methods for assessing the goodness of fit of the models. This first chapter also elaborates on how DCNNs, currently considered to be the best computational models of the ventral visual system, are trained to learn object representations. Different training schemes impose different learning goals, some of which may be more prominent during human object learning than others. **Chapter 2** will examine the impact of learning goals on the ability of DCNNs to explain fMRI data from human visual cortex. While humans are likely subjected to both supervised and unsupervised learning goals during development, DCNNs are standardly trained with supervision only. We hypothesize that unsupervised learning gives rise to human-like object representations that emphasize natural categories and that better explain fMRI data from human visual cortex. We test this hypothesis by training ResNet50 on ImageNet using both supervised and unsupervised contrastive learning, and by assessing how well representations of object images match between the supervised and unsupervised ResNet50 versions and human IT cortex. Overall, our results confirm our hypotheses. Our findings indicate that learning goals are important for shaping visual representations and suggest that object representations in human visual cortex may arise, at least in part, from unsupervised learning during development. We discuss our results in more detail in **Chapter 3**.

Chapter 2

Unsupervised object learning explains natural category structure in human cortex

2.1 Introduction

The human brain is capable of interpreting the external visual world through identifying and localizing objects, including faces (Goodale and Milner, 1992; Ungerleider and Haxby, 1994; Kanwisher et al., 1997; Grill-Spector et al., 1998; Isik et al., 2014). This capability suggests that the brain transforms incoming visual signals into a meaningful representation of the outside world. While we understand visual scenes quickly and without apparent effort, the computational task performed by the brain is nontrivial. Recent years have seen major progress in our understanding of the neural computations that support object and scene recognition thanks to advances in deep learning (Richards et al., 2019). DCNNs reach human-level performance at object recognition and can also explain brain responses to images (Krizhevsky et al., 2012; Khaligh-Razavi and Kriegeskorte, 2014; Yamins et al., 2014; He et al., 2016). While these

results are very exciting and suggest that DCNNs are promising models of the human visual system, they still leave significant amounts of variance in human brain responses and behaviour unexplained (Geirhos et al., 2017; Schrimpf et al., 2018; Tang et al., 2018; Kar et al., 2019; Storrs et al., 2020).

An important difference between brains and DCNNs is the way in which they learn to recognize objects (Richards et al., 2019). DCNNs are traditionally trained using supervised learning, while humans also rely heavily on unsupervised learning (Tenenbaum et al., 2011). The main difference between the two types of learning is the learning objective. In supervised learning, the objective is to learn a mapping between input and output (Bzdok et al., 2018). For object recognition, the input is an object image and the output is a category label. In humans, the equivalent is a parent teaching a child to categorize objects by pointing at an object and providing the category label. In unsupervised learning, the objective is to learn structure present in the input data, without having access to labels (Barlow, 1989). For object recognition, structure may consist of clusters of objects that share similar visual features. In humans, the equivalent is a child learning that dogs are similar because they share a unique set of features which makes them different from other animals. Because parents only provide category labels every now and then, and because infants have limited language capacity, humans rely quite heavily on unsupervised learning during development (Tenenbaum et al., 2011).

Recent work indicates that DCNNs do not need category supervision to predict object representations in human visual cortex. Konkle and Alvarez (2021) and Zhuang et al. (2021) show that deep neural networks trained with unsupervised learning methods predict object representations along the primate ventral visual stream, and they do so at least as well as supervised networks. These results were reported for a variety of feedforward network architectures and unsupervised learning algorithms, and for both nonhuman primates and humans. These results are promising but previous studies did not closely examine the nature of object representations learned through unsupervised learning. In this study, we test what information an unsupervised learning training scheme can extract from natural images, and whether this information matches

the information extracted by humans. We hypothesize that unsupervised learning gives rise to human-like object representations that emphasize natural categories.

2.2 Methods

2.2.1 Evaluation images

To evaluate the match between object representations in DCNNs and object representations in high-level human visual cortex, we presented models and humans with the same set of 96 evaluation images. The images are coloured photographs of isolated real-world objects from a range of categories, including faces, animals, fruits and manmade objects (Figure 2.1) (Kriegeskorte et al., 2008b). Objects were displayed on a uniform grey background at an image size of 175 x 175 pixels.

Figure 2.1: To compare humans and networks we characterized how each of them represents these evaluation images. The images show objects from a wide range of categories, including faces, animate objects, and inanimate objects.

2.2.2 fMRI experiment and analyses

The fMRI experiment has been described in detail in (Kriegeskorte et al., 2008b). We therefore only describe the essential features here.

Participants. fMRI data were acquired in four healthy human participants (mean age = 35 years; two females) with normal or corrected-to-normal vision. Participants provided written informed consent before participating. The experiment was conducted in accordance with the Institutional Review Board of the National Institutes of Mental Health (Bethesda, Maryland).

Experimental design and task. We used a rapid event-related design to present the evaluation stimuli (stimulus duration = 300 ms, inter-stimulus-interval = 3700 ms). Participants performed a fixation-cross-colour detection task unrelated to the stimuli. Stimuli were displayed at fixation on a uniform grey background at a width of 2.9 degrees of visual angle. Each image was presented once per run in random order. Each run included 40 randomly interspersed baseline trials where no image was shown. Participants completed two sessions on separate days. We acquired six 9-minute runs in each session.

fMRI measurements. Blood oxygen level-dependent (BOLD) fMRI measurements were performed at standard spatial resolution (voxel volume: 1.95 x 1.95 x 2 mm^3), using a 3 T General Electric HDx MRI scanner, and a custom-made 16-channel head coil (Nova Medical). Single-shot gradient-recalled echo-planar imaging with sensitivity encoding (matrix size: 128 x 96, TR: 2 s, TE: 30 ms, 272 volumes per run) was used to acquire 25 axial slices that covered occipital and temporal cortex.

fMRI data preprocessing. fMRI data preprocessing was performed using BrainVoyager QX 1.8 (Brain Innovation). The first three data volumes of each run were discarded to allow the fMRI signal to reach a steady state. All functional runs were subjected to slice-scan-time correction and 3D motion correction. Data were converted to percent signal change. Analyses were performed in native subject space.

Defining the region of interest. We defined inferior temporal (IT) cortex on the basis of independent experimental data and restricted to a cortex mask manually drawn on each subject's fMRI slices. IT was defined by selecting the 316 most visually responsive voxels within the IT portion of the cortex mask. Visual responsiveness was assessed using the t map for the average response to the evaluation images. The t map was computed on the basis of one third of the runs within each session.

Estimating single-image activity patterns. We concatenated the runs within a session along the temporal dimension and estimated single-image activity patterns using univariate modelling. We constructed a linear model consisting of hemodynamic-response predictors for the evaluation images (one for each image) along with six head-motion parameter time courses, a linear-trend predictor, a six-predictor Fourier basis for nonlinear trends (sines and cosines of up to three cycles per run), and a confound-mean predictor. We fit the model to each voxel in the IT region of interest (ROI) to obtain a response-amplitude estimate for each of the evaluation images. We converted the estimates to t values. The pattern of t values across ROI voxels for one image is referred to as a single-image response pattern. We used these response patterns for further analysis.

Constructing the representational dissimilarity matrix. We computed representational dissimilarity matrices (RDMs) by computing pairwise dissimilarities (1 - Pearson r) between the single-image activity patterns. We computed an RDM for each participant and session, and then averaged the RDMs across sessions, resulting in one RDM per participant. The RDMs characterize which aspects of the images are emphasized in the IT ROI and which aspects are de-emphasized.

2.2.3 Neural network architecture and training

We used the ResNet50 architecture in this study. ResNet-50 is one of the most frequently used feedforward DCNN architectures in computer vision with good performance on large-

scale object recognition tasks (He et al., 2016). ResNet50 is also among the best DCNNs for explaining brain responses during object perception (Schrimpf et al., 2018). We trained ResNet50 using both supervised and unsupervised training schemes on the 2012 ImageNet Large Scale Visual Recognition Challenge data base (ILSVRC 2012) (Deng et al., 2009; Russakovsky et al., 2015b). The ILSVRC 2012 data base consists of 1.2M annotated photographs organized according to the WordNet hierarchy (Miller, 1995; Fellbaum, 1998). Each photograph is annotated with one of 1,000 object category labels. We used PyTorch (Paszke et al., 2019), a high-performance deep learning library, for network training.

For supervised training, we used a standard cross-entropy loss function for a single label classification task. See section 1.3.3 for the definition of the cross-entropy loss function. We applied the following standard image augmentations during supervised training: crop a random piece of the image and resize it to the network input size of 224×224 pixels followed by a random horizontal flipping (p = 0.5) (Figure 2.2). We set the initial learning rate to 0.05 for the supervised training.

For unsupervised training, we used a computationally-efficient implementation of contrastive learning, Momentum Contrast (MoCo v2) (He et al., 2020). MoCo is a discriminative approach which uses a contrastive loss function, InfoNCE, which trains the network to map two augmented views of each image of the data base into a lower-dimensional embedding space in which augmented views of the same image are clustered and augmented views of different images are separated. See section 1.3.4 for the definition of the InfoNCE loss function. MoCo v2 applies the following image augmentations during training: random crop and resize followed by colour jittering (p= 0.8), grey-scale (p= 0.2), Gaussian blur (p= 0.5), and random horizontal flipping (p = 0.5) (He et al., 2020) (Figure 2.2). To ensure that differences in training augmentations do not confound comparisons between supervised and unsupervised learning, we incorporated a supervised+ learning scheme which uses the supervised loss function with MoCo augmentations. We set the initial learning rate to 0.015 for the unsupervised training.

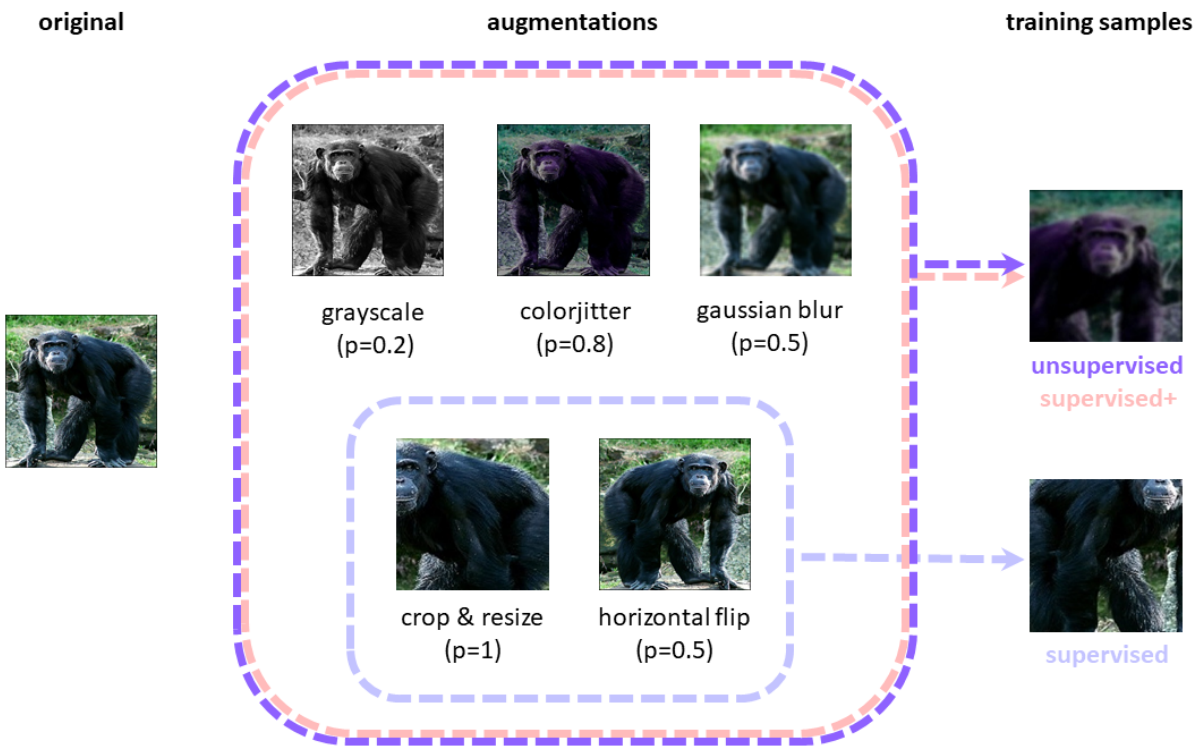

Figure 2.2: Training augmentations for the supervised, supervised+, and unsupervised training schemes. While supervised+ and unsupervised training schemes use all image augmentations, the supervised training scheme uses only a subset of image augmentations, consisting of crop & resize and horizontal flip. p is the probability that an augmentation is applied to a given image.

For a more accurate comparison of supervised and unsupervised learning, we did not apply the extra fully connected layers (multi-layer perception head) that MoCo v2 uses to improve classification accuracy. The supervised, supervised+, and unsupervised architectures are therefore identical except for the number of units in the last fully-connected layer. The last fully-connected layer consists of 1,000 units in the supervised and supervised+ architectures and 128 units in the unsupervised architecture. We trained all networks (supervised, supervised+, unsupervised) with multi-processing distributed data parallel training using nodes with four V100 GPUs. For all types of learning, we trained five instances of each network for 200 epochs, using a batch size of 128 images, and employing a cosine learning scheduler that gradually decreases the initial learning rate toward zero during training. We trained five network instances

to ensure that results are robust against individual differences among networks (Mehrer et al., 2020). Different network instances were created using different random weight initialization and different random assignment of data augmentations to images during training. The order of images shown to the networks over training epochs was identical across model instances.

Unlike supervised learning, unsupervised learning does not explicitly teach a network to classify objects in images. To compare ImageNet classification performance between supervised and unsupervised networks, we therefore first need to derive the object classification accuracy of the unsupervised network by applying a readout to the penultimate layer of the network. The readout transforms the representations in the penultimate layer into a probability distribution over classes, and needs to be learned after unsupervised training is completed. Since training the readout takes approximately 2 days, we ran it first for two randomly selected unsupervised model instances (out of five). For each model instance, we replaced the network's last 128-unit fully connected layer with a 1,000-unit fully connected layer as in the supervised network. We then froze the weights and biases of all other layers and trained the network on ImageNet for 100 epochs with supervised training. We implemented the training procedure for 10 training epochs, which were selected to span the network's learning trajectory (epochs 0, 1, 5, 10, 20, 40, 90, 120, 160, 200). We assessed the network's performance on the ImageNet validation images as for the supervised network. The resulting classification performance trajectories were very similar between the two unsupervised model instances. To make efficient use of limited computational resources, we therefore refrained from training the readout layer for the remaining three model instances. We averaged classification performance across the two tested model instances.

2.2.4 Characterizing network learning trajectories

To characterize how object representations in the networks developed over the course of training, we characterized the networks' internal representation of the evaluation images (Figure 2.1) as training progressed. After each training epoch, we provided the evaluation images as input to

the networks, and extracted the activity patterns elicited by the images from *layer 4*. During evaluation, we utilize the encoder network rather than momentum encoder network because the momentum encoder network parameters are a moving average of the encoder network. We focus our analyses on *layer 4* because it is located at a similar location in the visual processing hierarchy as IT. We computed RDMs by computing pairwise dissimilarities (1 - Pearson's r) between the activity patterns. This resulted in as many RDMs as training epochs for each network (supervised, supervised+, unsupervised). The RDMs characterize which aspects of the images get emphasized and which get de-emphasized by the networks as training progresses. The RDMs can be interpreted as the networks' learning trajectory.

2.2.5 Comparing object representations between brain and models

To test whether object representations in brains and models are related, we computed Spearman rank correlation coefficients between all possible pairs of a human subject RDM (IT) and a model instance RDM (layer 4). We computed the test statistic as the average correlation coefficient across all pairs. We used the Fisher transformation to ensure normality before averaging (Fisher, 1915). We used permutation tests for statistical inference (1,000 permutations). For each permutation, we randomly permuted the images and recomputed the RDMs, and then repeated the same procedure that we used for computing our test statistic. This created a null distribution of average correlation coefficients, which simulates the null hypothesis of no relationship between brain and model RDMs. If the actual average correlation coefficient falls within the top 5 percent of the null distribution, we reject the null hypothesis of unrelated RDMs, and conclude that object representations in brains and models are related. We performed the randomization test for each model (supervised, supervised+, unsupervised) and training epoch. We controlled the False Discovery Rate (FDR) (Benjamini et al., 2001) at 0.05 to correct for multiple comparisons over training epochs.

To test whether the unsupervised and supervised models differ in how well they mirror

the brain representations, we next examined differences in the average correlations coefficients between models. We bootstrap resampled the images for statistical inference (1,000 resamplings). We used a stratified approach, resampling images within the following categories: human faces, animal faces, human bodies, animal bodies, natural objects, and manmade objects. We applied the same bootstrap resamplings across subjects, model instances, and training epochs. For each resampling, we first computed the average Spearman rank correlation coefficient between pairs of brain and model RDMs as described above for the relatedness test. We then computed the difference in the average correlation coefficient between model pairs. Our bootstrap resampling procedure simulates the distribution of differences in average correlation coefficients between models that we would expect to observe if we repeated our analysis for different samples of images drawn from the same hypothetical population as our actual images, i.e. coloured photos of real-world objects. If 0 falls within the bottom or top 2.5 percent of the bootstrap distribution, we reject the null hypothesis of no difference between models, and conclude that one model is better than the other at mirroring the brain representations. We performed the bootstrap test for two model pairs (unsupervised and supervised; unsupervised and supervised+) and each training epoch. We controlled the FDR at 0.05 to correct for multiple comparisons over training epochs.

To evaluate how well the models are performing given the noise in the fMRI data, we calculated the noise ceiling of the data (Nili et al., 2014). The noise ceiling indicates the performance of the unknown true model given the noise in the data. To compute the lower bound of the noise ceiling, we calculated the Spearman rank correlation coefficient of each subject's RDM with the average RDM of the other three subjects and then averaged these values. To compute the upper bound of the noise ceiling, we calculated the Spearman rank correlation coefficient of each subject's RDM with the average RDM of all the subjects (including itself) and then averaged these values. If the best model does not reach the noise ceiling, this indicates that there is room for model improvement. If the best model reaches the noise ceiling, but the noise ceiling is low, this indicates that data quality is low and we need to invest in collecting better data.

To more closely investigate the match between brains and models, we compared the categorical structure of their object representations. We built a category model with predictors corresponding to natural categories of ecological relevance (Khaligh-Razavi and Kriegeskorte, 2014) (Figure 2.7). The model contained predictors for the following 10 categories: animate objects, inanimate objects, faces, human faces, animal faces, bodies, human bodies, animal bodies, natural objects, and manmade objects. We used binary predictors to model category clustering of response patterns (0 for within-category dissimilarities, 1 for between-category dissimilarities). We fit the category model, including an intercept, to the brain and model RDMs using non-negative least squares regression. We used non-negative least squares to ensure positive weights for our dissimilarity predictors (Khaligh-Razavi and Kriegeskorte, 2014; Jozwik et al., 2016). Before comparing predictor weights between the brain and the networks, we assessed multicollinearity of the predictors. We used the Variance Inflation Factor (VIF) to test for linear relationships between the predictors in the category model. The VIF gives an indication of how much more variable the predictor weights are than expected for a model without multicollinearity. Higher values indicate stronger multicollinearity. We fit the category model to each subject, model instance, and epoch. With this approach, we can study to what extent an object category emerges in the brain and model representations.

To test for differences between brain and models in category clustering, we used bootstrap resampling of the images (1,000 resamplings), using the same stratified approach as before. We applied the same bootstrap resamplings across subjects, model instances, and training epochs. For each resampling, we first fit the category model to all subjects and model instances and extracted the beta weights for the 10 category predictors. We then averaged the predictor weights across subjects and across model instances, and computed the difference in the average predictor weights between the brain and each model (unsupervised, supervised, supervised+). Our bootstrap resampling procedure simulates the distribution of differences in predictor weights between brain and models that we would expect to observe if we repeated our analysis for different samples of images drawn from the same hypothetical population as our actual images, i.e. coloured photos of real-world objects. If 0 falls within the bottom or top 2.5 percent of the

bootstrap distribution, we reject the null hypothesis of no difference between brain and models, and conclude that the strength of category clustering differs between the two. We performed the bootstrap test not only for the brain-model pairs but also for model-model pairs (unsupervised and supervised; unsupervised and supervised+), and for each category predictor and training epoch. We controlled the FDR at 0.05 to correct for multiple comparisons over training epochs.

2.3 Results

2.3.1 Object representations in human high-level visual cortex

Figure 2.3 visualizes the human IT object representation as measured with fMRI in four human subjects. The RDMs show a top-level category division between animate and inanimate objects, and within the animate objects, a cluster of faces. The two big blue squares in the RDMs show that activity patterns in response to objects cluster according to animacy: whenever a pair of objects consists of two animate or two inanimate objects, the two objects in the pair tend to elicit similar activity patterns. In contrast, as shown by the two big yellow squares, whenever a pair of objects consists of one animate and one inanimate object, the two objects in the pair tend to elicit dissimilar activity patterns. The four small blue squares within the animate cluster suggest that activity patterns in response to faces, whether they are human or animal, cluster together within the larger cluster of animate objects. The category structure is most clearly visible in the subject-average RDM, but is also evident in individual subjects. These observations suggest that object representations in human IT emphasize natural categories of long-standing ecological relevance.

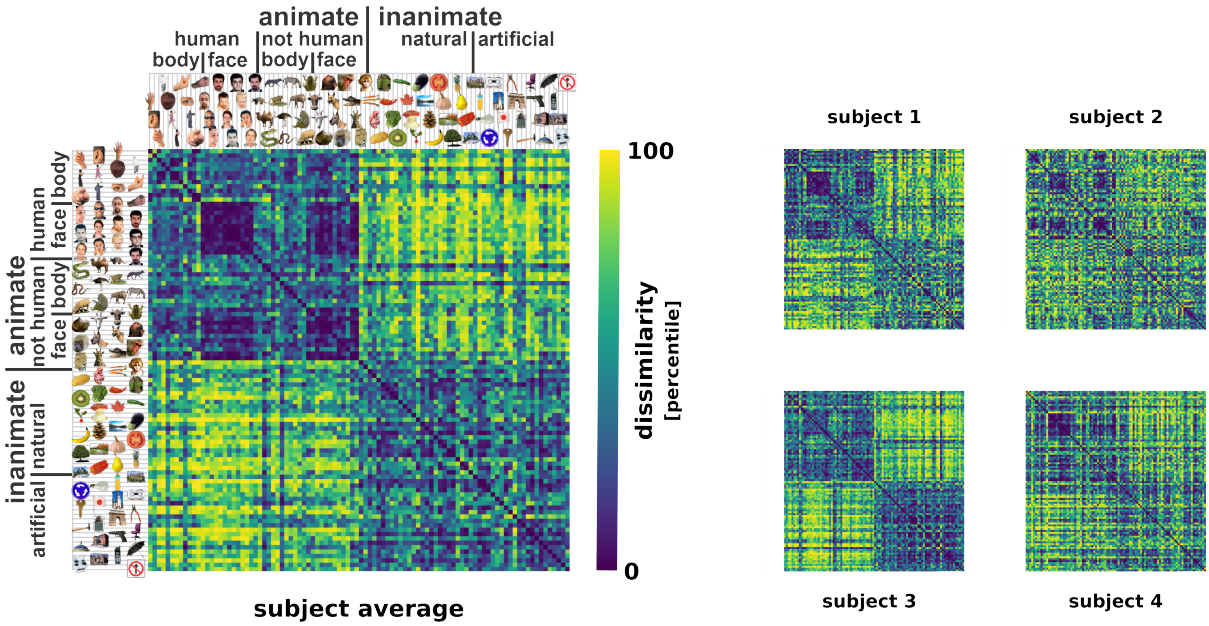

Figure 2.3: Representational dissimilarity matrices (RDMs) for human IT. The object representation in human IT emphasizes categories of ecological relevance, including animate objects, inanimate objects, and faces. Matrix entries represent response-pattern dissimilarities between pairs of object images ($1 - r$, where r is Pearson correlation coefficient; 316 most visually responsive bilateral IT voxels in each subject defined using independent data). The matrices were independently transformed into percentiles for easier visual comparison. The labels on the RDM axes indicate how object images are ordered along the axes.

2.3.2 Object representations in ResNet-50

Object classification performance

We first assessed object classification performance of the trained networks, which allows us to verify the effectiveness of the learned representations for the task. The task is classifying images into the 1,000 ILSVRC object categories, which is a standard computer vision benchmark for measuring object recognition performance. We measured ResNet-50's top-1 classification accuracy on the ILSVRC 2012 validation images after supervised and unsupervised training.

For the supervised model, the final output is a probability distribution across classes, from which we can directly derive the top-1 classification accuracy. For the unsupervised model, the final output is an object representation in a lower-dimensional embedding space. To estimate top-1 classification accuracy for the unsupervised model, we replaced the embedding layer with a linear classifier, froze the weights of the trained ResNet-50 architecture, and trained the linear classifier to map the object representations in the final convolutional layer to the 1,000 ILSVRC category labels. Models with effective representations should be able to learn a linear mapping from the representations to the class labels, and are expected to achieve a satisfactory top-1 classification accuracy.

As expected, Figure 2.4 shows that the supervised model, which was explicitly trained on the task, learned representations that are more effective for solving the task than the unsupervised model. The supervised top-1 classification accuracy reported here is comparable to prior work (He et al., 2016). This prior work also reported the top-5 classification accuracy, which was on par with human performance at the task (He et al., 2016, 2015; Russakovsky et al., 2015a). Impressively, the unsupervised model, which was trained using unsupervised contrastive learning and thus was not explicitly trained on the task, learned representations that yielded a classification accuracy within $\sim$ 10% of that of the supervised model. For both models, classification accuracy rose quickly early in training and stabilized later in training. To ensure that differences in classification accuracy cannot be attributed to differences in the choice of image augmentations between supervised and unsupervised training, we trained an additional supervised model. This model was trained with data augmentations that match those used in unsupervised training. We refer to this model as supervised+. Figure 2.4 shows that the choice of augmentations has little effect on top-1 accuracy with $< 1\%$ difference in performance between supervised and supervised+ models.

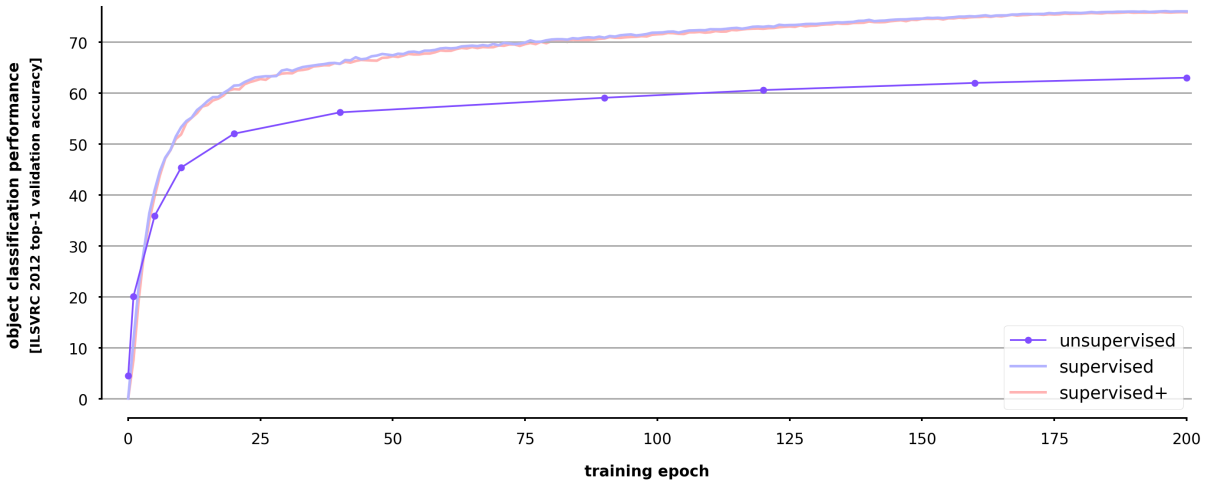

Figure 2.4: ImageNet (ILSVRC) top-1 classification accuracy for supervised and unsupervised networks. For supervised and supervised+ networks, performance was calculated and plotted for all 200 training epochs. For the unsupervised model, we evaluated performance for a subset of epochs, given the computational costs of estimating the performance for these models. Results for the selected epochs are denoted by purple ball markers and connected by line segments. Supervised networks outperform the unsupervised network at object classification.

Evolution of object representations over training

We next inspected the internal object representations of the networks as training progressed. After each training epoch, we presented the networks with the same object images as the human participants, and estimated the networks' internal object representations by computing RDMs. We focused on ResNet50 *layer 4*, which is located at a similar location in the visual processing hierarchy as IT. Figure 2.5 shows how the object representations evolve over unsupervised, supervised, and supervised+ training. The figure shows object representations for a subset of epochs, which were selected to cover the entire learning trajectory. The selected epochs more densely sample the earlier training epochs given that classification performance changes most quickly during this period.

Visual inspection of the RDMs suggests that the object representations are similar between networks before training begins (epoch 0). At this point, the networks have randomly initialized weights and their representations can be taken to reflect lower-level image similarity. In line with this, all networks show a cluster of human faces. After the first training epoch, the representations still appear quite similar across networks. However, after 10 training epochs, differences start to emerge between the unsupervised and supervised networks. The unsupervised version of the network starts to develop a face cluster (similar to human IT), while the supervised versions of the network start to develop a prominent human cluster. Furthermore, among the inanimate objects, the unsupervised network starts to show a cluster of artificial objects, while the supervised networks start to show a cluster of natural objects. After 40 training epochs, these differences are clearly visible. The two supervised models, which only differ in the data augmentations applied during training, show a highly similar developmental trajectory. This suggests that the additional augmentations applied during supervised+ training do not strongly affect the learned object representations.

These observations indicate that learning goals affect the object representations learned by the networks. Differences between supervised and unsupervised training emerge relatively early in training and stabilize after $\sim$ 40 epochs.

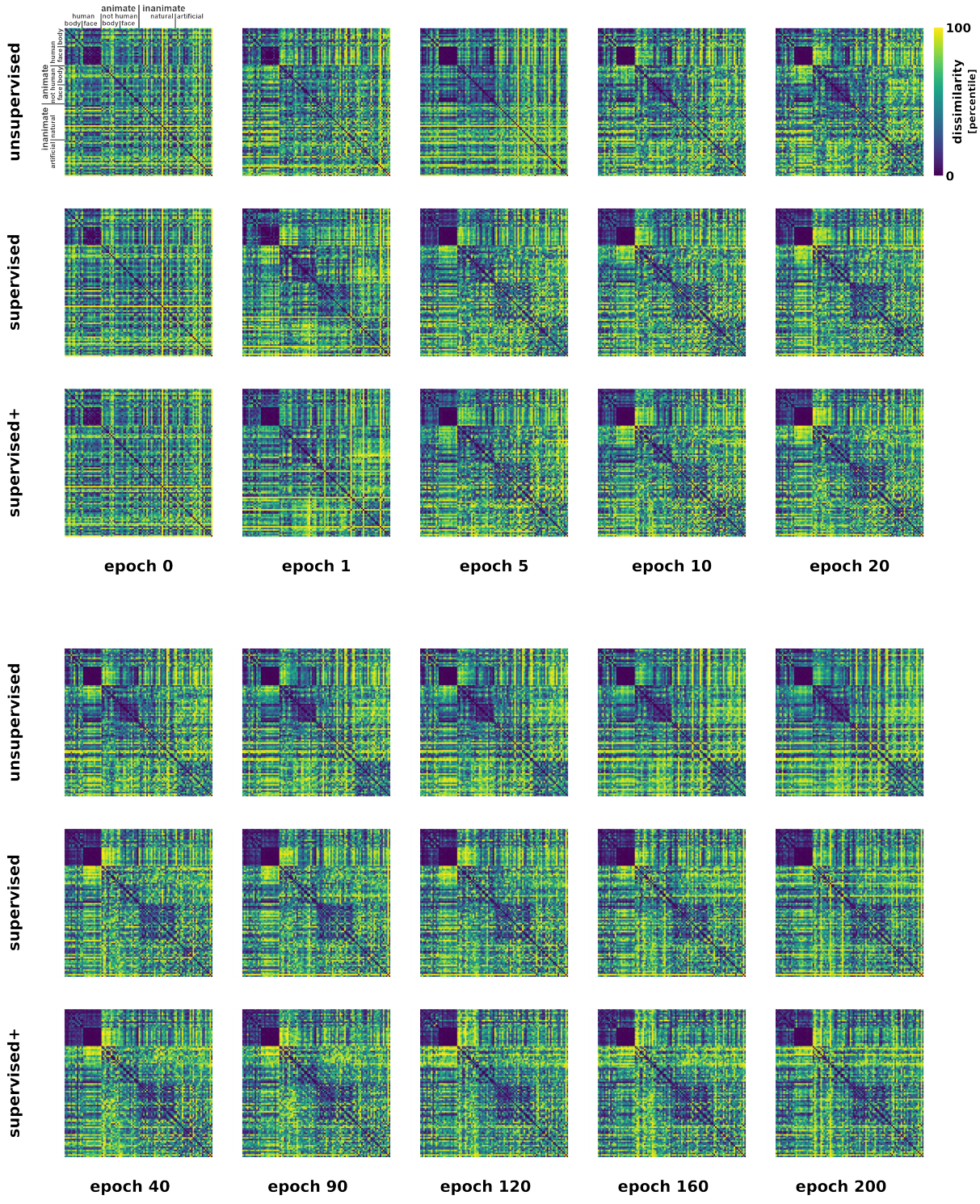

Figure 2.5: ResNet-50 learning trajectories for unsupervised, supervised, and supervised+ training. The trajectories consist of RDM snapshots showing the internal object representations in layer 4 of the networks as training progresses. RDM snapshots are displayed for a subset of

training epochs, chosen to cover the entire learning trajectory with denser sampling early in training. RDMs were independently transformed into percentiles for easier visual comparison. Results are shown for one model instance per training regime but are qualitatively similar across model instances within training regimes.

2.3.3 Comparing object representations between brain and models

To quantify the match between the object representations in brain and models, we correlated the human IT RDMs with the ResNet50 RDM learning trajectories. For each training epoch, we computed RDM correlations for all possible pairs of one human individual and one model instance, and then averaged correlations across pairs. Results are shown in Figure 2.6 for the three training regimes (unsupervised, supervised, and supervised+). Lines can be taken to reflect how closely the model representations reflect the human adult IT representation as training progresses. Object representations in all three networks are significantly correlated with the human IT object representation for all training epochs (permutation test, p-corrected < .05). This even holds for epoch 0, which is before training but after weight initialization. This suggests that a hierarchical cascade of linear and nonlinear transformations, using randomly drawn parameters, yields image representations that have some similarity to object representations in human IT, consistent with prior work (Yamins et al., 2014). For reference, Figure 2.6 shows performance of this null model using dotted lines. Model performance rises above the null performance soon after training commences.

Early in training, the supervised models outperform the unsupervised model at explaining the human IT object representation (bootstrap test, p-corrected < .05). However, this switches later in training: after epoch 60 the unsupervised model starts outperforming the supervised models (bootstrap test, p-corrected < .05). Supervised models reach their best performance at modelling object representations in human IT in the first few epochs. The unsupervised model instead reaches its best performance toward the end of training. The difference between

unsupervised and supervised models seems to increase over training. These results indicate that, while the unsupervised model is consistently worse at object classification than the supervised models, it is better at explaining brain representations than the supervised models as training progresses. Performance at explaining the IT object representation did not differ between the two supervised models. This suggests that the differences in data augmentations applied during training have little effect on performance of the supervised models. This result is consistent with the results on object classification, which also did not differ between the two supervised models. Importantly, none of the models reach the noise ceiling. The noise ceiling estimates performance of the true unknown model given the noise in the fMRI data (Nili et al., 2014) and is shown as a grey bar in Figure 2.6. This result suggests that there is room for improvement for all models.

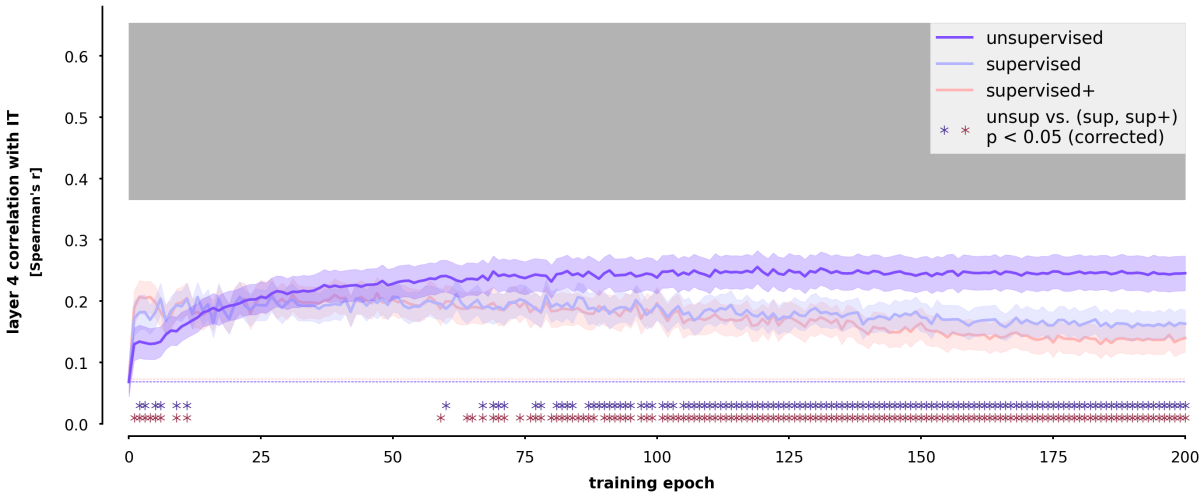

Figure 2.6: The unsupervised models outperform the supervised model in explaining the human IT object representation, except for the first few training epochs. Inferential results for model comparisons are based on bootstrap resampling of the evaluation images (1,000 resamplings). Solid lines show the correlation (Spearman r) between the network's learning trajectories of layer 4 and the human adult IT object representation. Shaded areas around the lines reflect standard error of the mean based on bootstrap resampling. The correlation with the human data at epoch 0 serves as a null model, whose performance is denoted by a dotted line. Asterisks indicate significant differences between the unsupervised model and the supervised model

(purple) and between the unsupervised model and the supervised+ model (red). The grey bar shows the noise ceiling.

Comparing the categorical nature of the representation

To better understand why the unsupervised network better explains the human brain data, we examined the categorical nature of the representations. For the object representations in brain and networks, we examined the strength of category clustering for a range of natural categories, including faces, animate objects, and inanimate objects. To do so, we fitted a category model to the brain and network RDMs using non-negative least-squares regression, and compared the strength of category clustering between human IT and each version of ResNet50 (unsupervised, supervised, supervised+). Figure 2.7 shows the category model, and the estimated weights for the category clustering predictors. We focus on results for predictors which had a significantly positive weight for human IT: faces, animate objects, and inanimate objects.

Multicollinearity of the predictors was assessed before comparing brain and network weights using the Variance Inflation Factor (VIF). Table 2.1 shows the VIF values for each category predictor. As a rule of thumb, VIF values above 5 are generally considered to indicate multicollinearity, although it should be noted that opinions differ slightly and cut-off values reported in the literature range between 2.5 and 10 (James et al., 2013; Allison, 2012). In our case, all VIF values are less than 2.5, which suggests that the category model does not suffer from multicollinearity and estimated predictor weights can be compared.

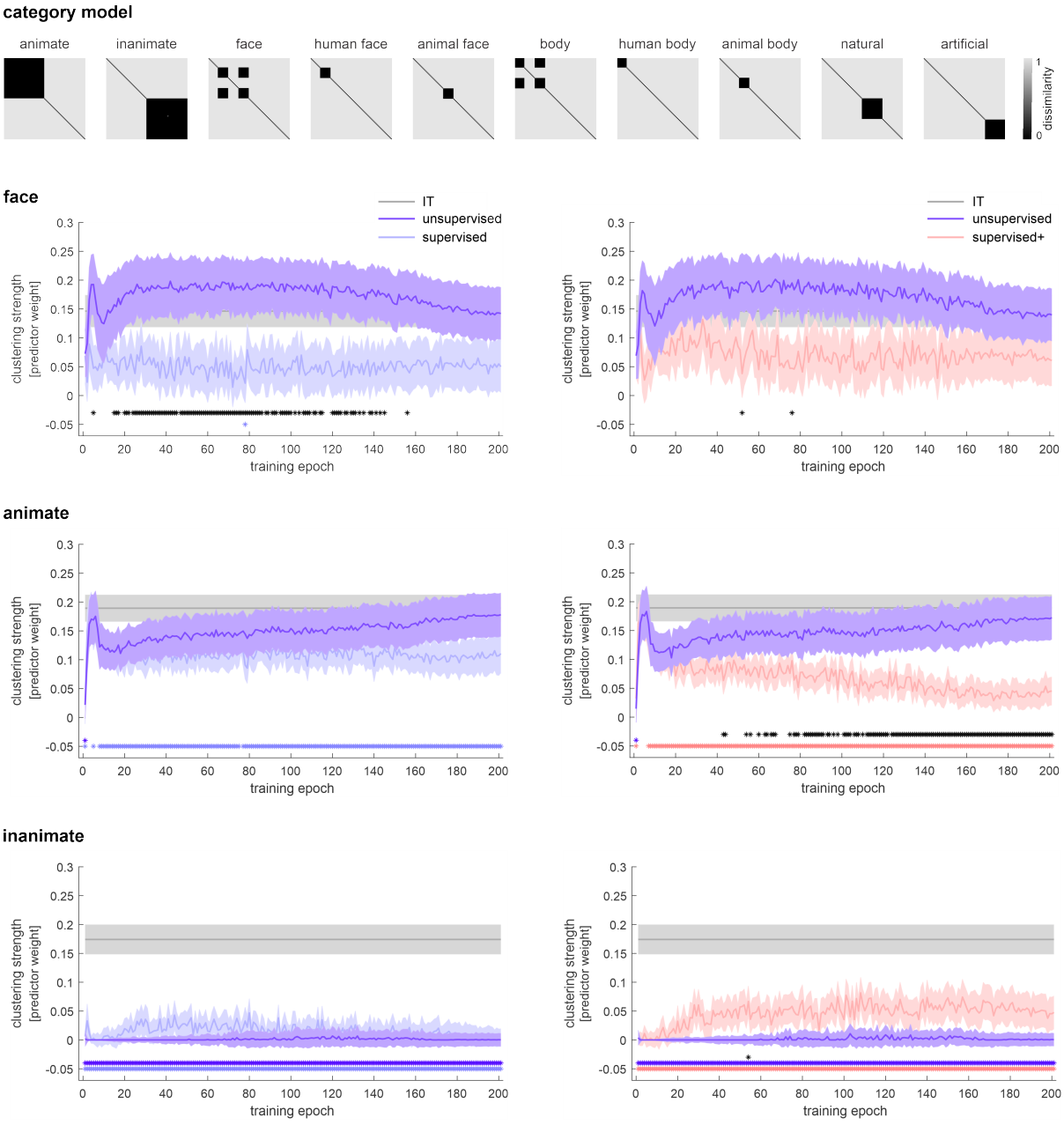

Figure 2.7: Category clustering of object representations in human IT and in *layer 4* of the unsupervised, supervised, and supervised+ networks. We assessed strength of category clustering by fitting a category model to all RDMs using non-negative least-squares regression. The top panel shows the predictors of the category model. The other panels show estimated weights for the face, animate, and inanimate category predictors. The left panels show results for human IT, unsupervised, and *supervised* networks. The right panels show results for human IT, unsupervised, and *supervised+* networks. Results suggests that the unsupervised network more closely resembles human IT than the supervised networks. In all plots the y-axis denotes

predictor weight and the x-axis shows training epoch. The grey line shows the weight for human IT, the dark purple for the unsupervised network, light purple for the supervised network, and light pink for the supervised+ network. Shaded regions reflect standard error based on bootstrap resampling of the evaluation images. Asterisks indicate significant differences between supervised and unsupervised models and humans. Results were corrected for multiple comparisons by controlling the false discovery rate at 0.05.

category	animate	inanimate	face	hum-face	ani-face	body	hum-body	ani-body	nat-obj	art-obj
VIF	1.8424	1.8424	2.2508	1.4372	1.4372	2.2508	1.4372	1.4372	1.3896	1.3896

Table 2.1: Variance Inflation Factors (VIF) for the category model. VIF indicates severity of multicollinearity in the category model used for the non-negative least-squares regression analysis. VIF values specify how much variance is increased in the respective estimated regression coefficient as a result of multicollinearity. All VIF values are smaller than 2.5 (our selected cut-off value).

Figure 2.7 shows that the strength of face clustering is similar between the unsupervised network and human IT. Consistent with this observation, strength of face clustering does not significantly differ between the unsupervised network and human IT (bootstrap test, p-corrected $< .05$). The strength of face clustering seems weaker in the supervised networks than in human IT. While this observation was not confirmed by statistical inference, the supervised network does show significantly weaker face clustering than the unsupervised network for most training epochs (bootstrap test, p-corrected $< .05$). A similar pattern of results was found for the strength of animate clustering, which does not differ significantly between the unsupervised network and human IT. In contrast, supervised models show weaker clustering of animate objects than human IT (bootstrap test, p-corrected $< .05$). Consistent with this result, animate clustering strength is significantly weaker in the supervised+ than the unsupervised network (bootstrap test, p-corrected $< .05$). All networks show significantly weaker clustering of inanimate objects than human IT. Taken together, these results suggest that the unsupervised contrastive learning scheme yields object representations that resemble those in human IT in terms of category

structure, and they do so more closely than object representations learned through supervised training.

2.4 Discussion

DCNNs are currently the best computational models of the human ventral visual system (Yamins et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014; Schrimpf et al., 2018). However, they cannot fully explain the categorical object representation in high-level human visual cortex (Storrs et al., 2020; Schrimpf et al., 2018). DCNNs are classically trained to recognize objects using supervised learning, which requires millions of labelled images (Russakovsky et al., 2015a). Humans, in contrast, rely heavily on unsupervised learning during development (Tenenbaum et al., 2011). Here, we examined how the type of learning affects a DNNs' internal representation of object images. We hypothesized that unsupervised learning gives rise to an object representation that emphasizes natural object categories and better explains human data than supervised learning.

Consistent with our hypothesis, our results show that the unsupervised contrastive learning scheme yields brain-like object representations that emphasize natural categories, including faces and animate objects. Furthermore, unsupervised versions of the networks better explain the human object representation than the supervised versions. This difference emerges relatively early in training and increases as learning progresses. The lower performance of supervised networks than unsupervised networks in modelling brain data might be explained by the granularity of the ImageNet class labels. ImageNet labels, which are used for supervised training, do not fully overlap with natural categories of ecological relevance (Mehrer et al., 2021; Chen et al., 2018). For example, there is no *face* category in ImageNet, whereas there are many distinct categories of animates, including diverse species of animals, birds, and plants (e.g. indigo finch, Afghan hound, Rhodesian ridgeback, etc). Thus, the unsupervised model's better performance may be explained by its contrastive learning scheme which is not constrained

by the ImageNet labels for training. This result is also consistent with recent studies showing higher performance of the contrastive unsupervised networks in transfer learning (Zhao et al., 2020; Kotar et al., 2021).

Our clustering strength analysis suggests that categorization of inanimate objects is weaker in supervised and unsupervised networks than in human IT. A weaker inanimate clustering in models could be explained by the training of models on static images, which could not capture one of the principal features distinguishing animate and inanimate objects, mobility (Haxby et al., 2020). Motion energy, especially if it is associated with biological motion, strongly drives responses in the ventral visual system (Russ and Leopold, 2015; Haxby et al., 2020). Without motion, it may be challenging to detect shared visual features between inanimate objects, which can be as varied as fruits and vegetables, natural scenes, buildings, and tools. Training the models on dynamic stimuli (videos) may improve their ability to represent inanimate objects as a category.

We also found that the unsupervised network explains face clustering and animate object clustering better than supervised and supervised+ networks, respectively. This suggests that the unsupervised network is able to 'discover' these categories through contrastive learning. While the supervised networks generally show weaker clustering of faces and animate objects, there are slight differences between the two supervised networks. The supervised network shows the weakest face clustering, while the supervised+ network shows the weakest animate clustering. These slight differences between the two networks must be due to the differences in image augmentations applied during training. The extra augmentations applied during supervised+ training consist of colour jittering, conversion to grey-scale, and Gaussian blurring. As long as these augmentations do not affect the most informative features for categorization, they may be beneficial because in that case they promote invariance to features that are not useful for categorization. This may explain why the supervised+ model is *better* at clustering faces than the supervised model: colour and high spatial frequency information are not the most relevant features for face categorization (Rajimehr et al., 2011; Henriksson et al., 2015). However,

colour has been shown to be informative for animate categorization (Kriegeskorte et al., 2008b), which may explain why the supervised+ model is *worse* at clustering animate objects than the supervised model.

Our findings underscore the potential of unsupervised learning for modelling human object learning, and show the importance of learning goals and 'visual diet' in shaping object representations. We anticipate these insights to contribute to building better computational models of the human visual system.

Chapter 3

Discussion

3.1 Overview

DCNNs are currently considered to be the best computational models of the human ventral visual system (Yamins et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014; Schrimpf et al., 2018). However, DCNNs fall short of properly explaining how human high-level visual cortex represents natural object categories (Storrs et al., 2020; Schrimpf et al., 2018). DCNNs are typically trained with category supervision. In contrast, humans learn through a combination of supervised and unsupervised strategies during development (Tenenbaum et al., 2011). In this thesis, we investigated how learning objectives affect the way a DCNN represents objects. We demonstrate that object representations learned by a feedfoward DCNN through contrastive unsupervised learning emphasize natural categories of ecological relevance, including faces and animate objects. Furthermore, the representations more closely match object representations in human IT than object representations learned through standard supervised training. Our findings suggest that learning objectives affect the nature of object representations learned by DCNNs. They also suggest that unsupervised contrastive learning yields more IT-like object representations than supervised learning. These findings are consistent with the idea that

unsupervised learning plays a role in human visual representation learning.

3.2 Conclusion

The lower performance of supervised than unsupervised DCNNs in modelling brain data might be explained by the fact that supervised DCNNs are constrained by the category labels provided during training. ImageNet labels, which are commonly used for supervised training (Russakovsky et al., 2015a), do not fully overlap with natural categories of ecological relevance (Mehrer et al., 2021; Chen et al., 2018). For example, there is no *face* class in ImageNet, whereas there are many fine-grained categories of animate objects, including diverse species of animals, birds, and plants (e.g. indigo finch, Afghan hound, Rhodesian ridgeback, etc). The unsupervised model's better performance may be explained by its contrastive learning scheme which is not constrained by the ImageNet labels for training.

According to our clustering strength analysis, categorization of inanimate objects is weaker in both supervised and unsupervised DCNNs than in human IT. Weaker inanimate clustering in models could be attributed to the fact that models are trained using static images and thus could not capture one of the the common features that differentiates animate and inanimate categories, mobility (Russ and Leopold, 2015; Haxby et al., 2020). Our findings also show that an unsupervised DCNN explains face clustering and animate object clustering in human IT better than a supervised DCNN. This can at least partly explain the better performance of the unsupervised model at explaining brain data: the human IT object representation also shows clusters of faces and animate objects. Finally, our findings indicate that the type of image augmentations used during supervised training, which affect the 'visual diet' of the DNNs, have no observable effect on object classification performance, but do have a small effect on the nature of the learned object representations.

3.3 Limitations and future work

3.3.1 Brain activity was measured in human adults only

While it is possible that many classes of models reach the same performance at explaining neural activity in human visual cortex, solely comparing the learning trajectory of candidate models with adult human data at one time point puts unnecessary constraints in the space of acceptable models of the visual system. In this thesis, we compared the networks' representational learning trajectories with the *adult* human IT object representation. However, ideally we want to compare the learning trajectories of different models to neural activities recorded at different ages, starting in infancy. Several studies have shown that major steps of categorical learning occur during infancy (Deen et al., 2017; Spriet et al., 2021). Currently, this kind of data is fairly rare because it is challenging to collect, but with a collective effort across labs this seems to be a plausible and exciting project for the near future.

3.3.2 Mismatch of categorical structure between training and evaluation images

All models used in this study were trained on ImageNet but model evaluations were performed using images from Kriegeskorte et al. (2008b). While this is generally considered to be a perfectly fine cross-validation scheme in the sense that we minimize the similarity of the training set and testing set, our focus on learning goals might raise a few concerns. First, the supervised models used in this study received supervision according to the category structure of the ImageNet data base which is not optimally designed to represent ecologically relevant categories. Therefore, one hypothesis is that if we train a model on an image set which consists of unbiased representative instances from ecologically relevant categories, then supervised models could potentially outperform self-supervised models. While we plan to address this

point by training supervised models on Ecoset (Mehrer et al., 2021) in the future, it is important to note that ultimately we would prefer models to explain human visual recognition with less supervision, to more closely simulate how humans learn (Tenenbaum et al., 2011). Training with limited supervision will also remove the constraints imposed by the category labels, which is expected to improve transfer learning and thus general visual task performance.

3.3.3 Going beyond an overall representational match

The current thesis focuses on object representations in an a priori selected layer of ResNet-50, which is located at a similar location in the visual hierarchy as human IT. This approach assumes a relatively strict correspondence between the visual hierarchy of the brain and the model. While early DCNN layers tend to generally map onto early visual cortex, and deeper DCNN layers onto higher-level visual cortex, this correspondence is not perfect (Güçlü and van Gerven, 2015). We therefore plan to combine the object representation across different DCNN layers to test whether a combined model better captures the representational space of the human brain. Outcomes may be insightful for designing new models of the ventral visual stream with hybrid brain-like properties of existing models.

The metric we use to compare object representations between brain and models, the correlation coefficient, is considered to be a reliable unbiased similarity metric that does not require any parameter fitting. The class of metrics based on correlational similarities, which includes RSA, has proven to be advantageous relative to linear mapping metrics (Kornblith et al., 2019). However, linear mapping (Yamins et al., 2014) makes it possible to fit different breakdowns of the model components (layers, units, combination of layers/units) to single neurons. This creates a more flexible model which may better fit the data and which would enable benchmarking based on classical effects at the single neuron level such as feature-specific tuning properties across different visual areas (gabors, curvatures, faces, etc) (Hubel and Wiesel, 1968; Gross, 1973; Anzai et al., 2007). However, this approach is also more prone to overfitting.

Furthermore, the 'ideal' model that this approach aims to build is slightly different: it does not necessarily have to closely match the hierarchy of the human visual system or the features represented in visual cortex, as long as linear combinations of model components and features can explain data acquired from visual cortex.

3.3.4 Training on (even) more human-like learning objectives

In our quest to examine which training scheme leads to a more brain-like representation along the learning trajectory, we used representative implementations of two major classes of training schemes: supervised (including supervised+) and contrastive unsupervised learning. Although the training schemes used in this thesis, for example MoCo for unsupervised learning, are considered to exhibit the main hallmarks of the large classes they represent, there are many other instances in each group that are worth investigating. For example, SimSiame (Chen et al., 2020) as an unsupervised contrastive model does not require any negative samples and does not require instant weight copying to another separate network. Both of these features are favourable in designing a biologically plausible neural network. Another interesting instance would be a hybrid of contrastive learning and supervision (Khosla et al., 2020; Majumder et al., 2021), which allows selecting positive samples for the contrastive cost function from the same 'class' instead of limiting them to be extracted from only the same 'image'. This feature allows learning a better embedding space.

3.3.5 Recurrent processing: the dynamics of object representations

Time-averaged data and feedforward computations are commonly used in human neuroscience and corresponding modelling studies (Khaligh-Razavi and Kriegeskorte, 2014; Yamins et al., 2014; Güçlü and van Gerven, 2015; Storrs et al., 2020). For this project, we also averaged brain activity over time in the fMRI experiment and used ResNet50, a feedforward DCNN,

to model object representations. However, the primate visual system contains a large number of lateral and feedback connections (Kravitz et al., 2013). These connections are thought to contribute to visual perception through recurrent interactions, which likely play a role in attentional processes and in integrating incoming visual signals with prior knowledge (Tang et al., 2018; Wyatte et al., 2014; Lamme and Roelfsema, 2000). Therefore, by measuring and modelling the rapid dynamics of representations across the ventral visual system, we can better understand the computational mechanisms of object recognition. In order to understand how the dynamics of object representations in the brain evolve over object learning, we need to collect brain data from infants and adults using high temporal resolution methods, such as electroencephalography (EEG) and magnetoencephalography (MEG) (Xie et al., 2021). The dynamics of object representation learning can be modelled with deep *recurrent* neural networks, which are also among the best models for explaining object representations in the brain (Schrimpf et al., 2018; Kubilius et al., 2018). We plan to incorporate these extensions in future work.

Bibliography

Allison, P. D. (2012). *Logistic regression using SAS: Theory and application*. SAS institute.

Anzai, A., Peng, X., and Van Essen, D. C. (2007). Neurons in monkey visual area v2 encode combinations of orientations. *Nature Neuroscience*, 10(10):1313–1321.

Anzellotti, S., Fairhall, S. L., and Caramazza, A. (2013). Decoding Representations of Face Identity That are Tolerant to Rotation. *Cerebral Cortex*, 24(8):1988–1995.

Arcaro, M. J. and Livingstone, M. S. (2017). A hierarchical, retinotopic proto-organization of the primate visual system at birth. *Elife*, 6:e26196.

Arcaro, M. J., Schade, P. F., Vincent, J. L., Ponce, C. R., and Livingstone, M. S. (2017). Seeing faces is necessary for face-domain formation. *Nature Neuroscience*, 20(10):1404–1412.

Bandettini, P. A., Wong, E. C., Hinks, R. S., Tikofsky, R. S., and Hyde, J. S. (1992). Time course epi of human brain function during task activation. *Magnetic resonance in medicine*, 25(2):390–397.

Bao, P., She, L., McGill, M., and Tsao, D. Y. (2020). A map of object space in primate inferotemporal cortex. *Nature*, 583:103–108.

Barlow, H. B. (1989). Unsupervised learning. *Neural Computation*, 1(3):295–311.

Bell, A. H., Malecek, N. J., Morin, E. L., Hadj-Bouziane, F., Tootell, R. B. H., and Ungerleider, L. G. (2011). Relationship between functional magnetic resonance imaging-identified regions and neuronal category selectivity. *Journal of Neuroscience*, 31(34):12229–12240.

Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N., and Golani, I. (2001). Controlling the false discovery rate in behavior genetics research. *Behavioural brain research*, 125(1-2):279–284.

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological Review*, 94(2):115.

Bjorck, J., Gomes, C., Selman, B., and Weinberger, K. Q. (2018). Understanding batch normalization. *arXiv preprint arXiv:1806.02375*.

Bornstein, M. H. and Arterberry, M. E. (2010). The development of object categorization in young children: Hierarchical inclusiveness, age, perceptual attribute, and group versus individual analyses. *Developmental Psychology*, 46(2):350–365.

Bzdok, D., Krzywinski, M., and Altman, N. (2018). Machine learning: supervised methods. *Nature Methods*, 15(5-6).

Cantlon, J. F., Pinel, P., Dehaene, S., and Pelphrey, K. A. (2010). Cortical Representations of Symbols, Objects, and Faces Are Pruned Back during Early Childhood. *Cerebral Cortex*, 21(1):191–199.

Carlson, T. A., Schrater, P., and He, S. (2003). Patterns of activity in the categorical representations of objects. *Journal of Cognitive Neuroscience*, 15(5):704–717.

Caron, A. J., Caron, R. F., and Carlson, V. R. (1979). Infant perception of the invariant shape of objects varying in slant. *Child Development*, 50(3):716–721.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Chen, Z., Ding, R., Chin, T.-W., and Marculescu, D. (2018). Understanding the impact of label granularity on cnn-based image classification. In *2018 IEEE international conference on data mining workshops (ICDMW)*, pages 895–904. IEEE.

Day, R. H. and McKenzie, B. E. (1981). Infant perception of the invariant size of approaching and receding objects. *Developmental Psychology*, 17(5):670–677.

Deen, B., Richardson, H., Dilks, D. D., Takahashi, A., Keil, B., Wald, L. L., Kanwisher, N., and Saxe, R. (2017). Organization of high-level visual cortex in human infants. *Nature Communications*, 8(1):1–10.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee.

Desimone, R., Albright, T., Gross, C., and Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *Journal of Neuroscience*, 4(8):2051–2062.

DiCarlo, J. J. and Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, 11(8):333–341.

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

Downing, P. E., Jiang, Y., Shuman, M., and Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science*, 293(5539):2470–2473.

Edelman, S. (1997). Computational theories of object recognition. *Trends in Cognitive Sciences*, 1(8):296–304.

Edelman, S. (1998). Representation is representation of similarities. *The Behavioral and brain sciences*, 21(4):449–498.

Eger, E., Ashburner, J., Haynes, J.-D., Dolan, R. J., and Rees, G. (2008). fMRI Activity Patterns in Human LOC Carry Information about Object Exemplars within Category. *Journal of Cognitive Neuroscience*, 20(2):356–370.

Epstein, R. and Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, 392(6676):598–601.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521.

Freeman, J. and Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature Neuroscience*, 14(9):1195–1201.

Geirhos, R., Janssen, D. H., Schütt, H. H., Rauber, J., Bethge, M., and Wichmann, F. A. (2017). Comparing deep neural networks against humans: object recognition when the signal gets weaker. *arXiv preprint arXiv:1706.06969*.

Gliga, T. and Dehaene-Lambertz, G. (2007). Development of a view-invariant representation of the human head. *Cognition*, 102(2):261–288.

Goebel, R. (2007). Localization of Brain Activity using Functional Magnetic Resonance Imaging. In Stippich, C., editor, *Clinical Functional MRI*, pages 9–51. Springer Berlin Heidelberg, Berlin, Heidelberg. Series Title: Medical Radiology.

Goodale, M. A. and Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25.

Goodale, M. A., Milner, A. D., Jakobson, L., and Carey, D. (1991). A neurological dissociation between perceiving objects and grasping them. *Nature*, 349(6305):154–156.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27.

Grill-Spector, K., Kushnir, T., Hendler, T., Edelman, S., Itzchak, Y., and Malach, R. (1998). A sequence of object-processing stages revealed by fmri in the human occipital lobe. *Human brain mapping*, 6(4):316–328.

Grill-Spector, K. and Weiner, K. S. (2014). The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, 15(8):536–548.

Gross, C. G. (1973). Inferotemporal cortex and vision. In Stellar, E. and Sprague, J., editors, *Progress in Physiological Psychology*, pages 77–123. Academic Press.

Gross, C. G., Rocha-Miranda, C. d., and Bender, D. (1972). Visual properties of neurons in inferotemporal cortex of the macaque. *Journal of Neurophysiology*, 35(1):96–111.

Güçlü, U. and van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014.

Gutmann, M. and Hyvärinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings.

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

Haxby, J. V., Gobbini, M. I., and Nastase, S. A. (2020). Naturalistic stimuli reveal a dominant role for agentic action in visual representation. *NeuroImage*, 216:116561.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9729–9738.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Hegdé, J. and Van Essen, D. C. (2000). Selectivity for complex shapes in primate visual area v2. *The Journal of Neuroscience*, 20(5):RC61.

Henriksson, L., Mur, M., and Kriegeskorte, N. (2015). Faciotopy—a face-feature map with face-like topology in the human occipital face area. *Cortex*, 72:156–167. The whole is greater than the sum of the parts.

Hubel, D. H. and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 195(1):215–243.

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

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR.

Isik, L., Meyers, E. M., Leibo, J. Z., and Poggio, T. (2014). The dynamics of invariant object recognition in the human visual system. *Journal of neurophysiology*, 111(1):91–102.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Jozwik, K. M., Kriegeskorte, N., and Mur, M. (2016). Visual features as stepping stones toward semantics: Explaining object similarity in it and perception with non-negative least squares. *Neuropsychologia*, 83:201–226.

Kanwisher, N., McDermott, J., and Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience*, 17(11):4302–4311.

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., and DiCarlo, J. 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.

Khaligh-Razavi, S.-M. and Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, 10(11):e1003915.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. (2020). Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*.

Kiani, R., Esteky, H., Mirpour, K., and Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of Neurophysiology*, 97(6):4296–4309.

Kietzmann, T. C., McClure, P., and Kriegeskorte, N. (2018). Deep neural networks in computational neuroscience. *BioRxiv*.

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K., Cichy, R. M., Hauk, O., and Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, 116(43):21854–21863.

Kim, T., Bair, W., and Pasupathy, A. (2019). Neural coding for shape and texture in macaque area v4. *The Journal of Neuroscience*, 39(24):4760–4774.

Konkle, T. and Alvarez, G. A. (2021). Beyond category-supervision: instance-level contrastive learning models predict human visual system responses to objects. *bioRxiv*.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. (2019). Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529. PMLR.

Kotar, K., Ilharco, G., Schmidt, L., Ehsani, K., and Mottaghi, R. (2021). Contrasting contrastive self-supervised representation learning pipelines. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9949–9959.

Kravitz, D. J., Saleem, K. S., Baker, C. I., Ungerleider, L. G., and Mishkin, M. (2013). The ventral visual pathway: an expanded neural framework for the processing of object quality. *Trends in cognitive sciences*, 17(1):26–49.

Kriegeskorte, N. and Douglas, P. K. (2018). Cognitive computational neuroscience. *Nature Neuroscience*, 21(9):1148–1160.

Kriegeskorte, N. and Golan, T. (2019). Neural network models and deep learning - a primer for biologists. *ArXiv*, abs/1902.04704.

Kriegeskorte, N., Mur, M., and Bandettini, P. A. (2008a). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4.

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

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25.

Kubilius, J. (2017). Ventral visual stream.

Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L., and DiCarlo, J. J. (2018). Cornet: Modeling the neural mechanisms of core object recognition. *BioRxiv*, page 408385.

Lamme, V. A. and Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in neurosciences*, 23(11):571–579.

Levy, I., Hasson, U., Avidan, G., Hendler, T., and Malach, R. (2001). Center-periphery organization of human object areas. *Nature Neuroscience*, 4(5):533–539.

Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., and Hinton, G. (2020). Backpropagation and the brain. *Nature Reviews Neuroscience*, 21(6):335–346.

Livingstone, M. S., Arcaro, M. J., and Schade, P. F. (2019). Cortex is cortex: Ubiquitous principles drive face-domain development. *Trends in Cognitive Sciences*, 23(1):3–4.

Livingstone, M. S., Vincent, J. L., Arcaro, M. J., Srihasam, K., Schade, P. F., and Savage, T. (2017). Development of the macaque face-patch system. *Nature Communications*, 8(1):14897.

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., and Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fmri signal. *nature*, 412(6843):150–157.

Majumder, O., Ravichandran, A., Maji, S., Polito, M., Bhotika, R., and Soatto, S. (2021). Revisiting contrastive learning for few-shot classification. *arXiv preprint arXiv:2101.11058*.

Malach, R., Reppas, J., Benson, R., Kwong, K., Jiang, H., Kennedy, W., Ledden, P., Brady, T., Rosen, B., and Tootell, R. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of the National Academy of Sciences*, 92(18):8135–8139.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman and Company.

Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N., and Kietzmann, T. C. (2021). An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, 118(8):e2011417118.

Mehrer, J., Spoerer, C. J., Kriegeskorte, N., and Kietzmann, T. C. (2020). Individual differences among deep neural network models. *Nature Communications*, 11:5725.

Mell, M. M., St-Yves, G., and Naselaris, T. (2021). Voxel-to-voxel predictive models reveal unexpected structure in unexplained variance. *NeuroImage*, 238:118266.

Miller, G. A. (1995). Wordnet: A lexical database for english. In *Communications of the ACM*, volume 38, pages 39–41.

Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P., and Kriegeskorte, N. (2013). Human object-similarity judgments reflect and transcend the primate-it object representation. *Frontiers in Psychology*, 4:128.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., and Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS computational biology*, 10(4):e1003553.

Nishimura, M., Scherf, K. S., Zachariou, V., Tarr, M. J., and Behrmann, M. (2015). Size Precedes View: Developmental Emergence of Invariant Object Representations in Lateral Occipital Complex. *Journal of Cognitive Neuroscience*, 27(3):474–491.

Noroozi, M. and Favaro, P. (2016). Unsupervised learning of visual representations by solving jigsaw puzzles. In *Proceedings of the European Conference on Computer Vision*, pages 69–84. Springer.

Ogawa, S., Tank, D. W., Menon, R., Ellermann, J. M., Kim, S. G., Merkle, H., and Ugurbil, K. (1992). Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proceedings of the National Academy of Sciences*, 89(13):5951–5955.

Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Op de Beeck, H., Haushofer, J., and Kanwisher, N. (2008). Interpreting fmri data: maps, modules and dimensions. *Nature Reviews Neuroscience*, 9(2):123–135.

Op de Beeck, H., Wagemans, J., and Vogels, R. (2001). Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nature neuroscience*, 4(12):1244–1252.

Pasupathy, A. and Connor, C. E. (2002). Population coding of shape in area v4. *Nature Neuroscience*, 5(12):1332–1338.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037.

Perenin, M. and Vighetto, A. (1988). Optic ataxia: a specific disruption in visuomotor mechanisms. i. different aspects of the deficit in reaching for objects. *Brain*, 111(3):643—-674.

Pohl, W. (1973). Dissociation of spatial discrimination deficits following frontal and parietal lesions in monkeys. *Journal of Comparative and Physiological Psychology*, 82(2):227—-239.

Rajimehr, R., Devaney, K. J., Bilenko, N. Y., Young, J. C., and Tootell, R. B. (2011). The "parahippocampal place area" responds preferentially to high spatial frequencies in humans and monkeys. *PLoS biology*, 9(4):e1000608.

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., et al. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, 22(11):1761–1770.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., and Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8(3):382–439.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1985). Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science.

Russ, B. E. and Leopold, D. A. (2015). Functional mri mapping of dynamic visual features during natural viewing in the macaque. *NeuroImage*, 109:84–94.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015a). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015b). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252.

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

Santurkar, S., Tsipras, D., Ilyas, A., and Mądry, A. (2018). How does batch normalization help optimization? In *Proceedings of the 32nd international conference on neural information processing systems*, pages 2488–2498.

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

Sereno, M. I., Dale, A., Reppas, J., Kwong, K., Belliveau, J., Brady, T., Rosen, B., and Tootell, R. (1995). Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science*, 268(5212):889–893.

Serre, T., Oliva, A., and Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences*, 104(15):6424–6429.

Shepard, R. N. (1980). Multidimensional scaling, tree fitting, and clustering. *Science*, 210(4468):390–398.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Spriet, C., Abassi, E., Hochmann, J.-R., and Papeo, L. (2021). Visual object categorization in infancy. *bioRxiv*.

Storrs, K. R., Khaligh-Razavi, S.-M., and Kriegeskorte, N. (2020). Noise ceiling on the cross-validated performance of reweighted models of representational dissimilarity: Addendum to Khaligh-Razavi & Kriegeskorte (2014). *bioRxiv*.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, 19(1):109–139.

Tang, H., Schrimpf, M., Lotter, W., Moerman, C., Paredes, A., Caro, J. O., Hardesty, W., Cox, D., and Kreiman, G. (2018). Recurrent computations for visual pattern completion. *Proceedings of the National Academy of Sciences*, 115(35):8835–8840.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285.

Tsao, D. Y., Freiwald, W. A., Tootell, R. B., and Livingstone, M. S. (2006). A cortical region consisting entirely of face-selective cells. *Science*, 311(5761):670–674.

Ungerleider, L. G. and Haxby, J. V. (1994). 'what'and 'where'in the human brain. *Current opinion in neurobiology*, 4(2):157–165.

Ungerleider, L. G. and Mishkin, M. (1982). Two cortical visual systems. In Ingle, D. J., Goodale, M. A., and Mansfield, R. J., editors, *Analysis of Visual Behavior*, pages 549–586. MIT Press.

Wyatte, D., Jilk, D. J., and O'Reilly, R. C. (2014). Early recurrent feedback facilitates visual object recognition under challenging conditions. *Frontiers in psychology*, 5:674.

Xie, S., Hoehl, S., Moeskops, M., Kayhan, E., Kliesch, C., Turtleton, B., Köster, M., and Cichy, R. M. (2021). Visual category representations in the infant brain. *bioRxiv*.

Yacoub, E., Harel, N., and Uğurbil, K. (2008). High-field fmri unveils orientation columns in humans. *Proceedings of the National Academy of Sciences*, 105(30):10607–10612.

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

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

Zhang, R., Isola, P., and Efros, A. A. (2016). Colorful image colorization. In *Proceedings of the European Conference on Computer Vision*, pages 649–666. Springer.

Zhao, N., Wu, Z., Lau, R. W., and Lin, S. (2020). What makes instance discrimination good for transfer learning? *arXiv preprint arXiv:2006.06606*.

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., and Yamins, D. L. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3).

Curriculum Vitae

Name:	Ehsan Tousi
Education:	MSc in Communication systems Engineering (2013-2016) Shahid Beheshti University, Tehran, Iran
	BSc in Electrical Engineering (2005-2010) Ferdowsi University of Mahhad, Mashhad, Iran
Awards:	Vision Science Society travel award (2021)
Related Work Experience:	Teaching Assistant (2019 - Present) The University of Western Ontario

Publications:

1. Tousi, E., Mur, M. (2021), Unsupervised object learning explains face but not animate category structure in human visual cortex. *Journal of Vision 21* (9), 2501-2501.

2. Toosi, T., Tousi, E., Esteky, H. (2017), Learning temporal context shapes prestimulus alpha oscillations and improves visual discrimination performance. *Journal of Neurophysiology* 118 (2), 771-777.