2-21-2023 4:00 PM

The Two Visual Processing Streams Through The Lens Of Deep Neural Networks

Aidasadat Mirebrahimi Tafreshi, *The University of Western Ontario*

Supervisor: Mur, Marieke, *The University of Western Ontario*
Co-Supervisor: Goodale, Melvyn A., *The University of Western Ontario*
Joint Supervisor: Maryam Vaziri Pahskam, *National Institute of Mental Health*
A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Computer Science

Abstract

Recent advances in computer vision have enabled machines to have high performance in labeling objects in natural scenes. However, object labeling constitutes only a small fraction of daily human activities. To move towards building machines that can function in natural environments, the usefulness of these models should be evaluated on a broad range of tasks beyond perception. Moving towards this goal, this thesis evaluates the internal representations of state-of-the-art deep convolutional neural networks in predicting a perception-based and an action-based behavior: object similarity judgment and visually guided grasping. To do so, a dataset of everyday objects was collected and used to obtain these two behaviors on the same set of stimuli. For the grasping task, participants' finger positions were recorded at the end of the object grasping movement. Additionally, for the similarity judgment task, an odd-one-out experiment was conducted to build a dissimilarity matrix based on participants' similarity judgments. A comparison of the two behaviors suggests that distinct features of objects are used for performing each task. I next explored if the features extracted in different layers of the state-of-the-art deep convolutional neural networks (DNNs) could be useful in deriving both outputs. The prediction accuracy of the similarity judgment behavior increased from low to higher layers of the networks, while that of the grasping behavior increased from low to mid-layers and drastically decreased further along the hierarchy. These results suggest that for building a system that could perform these two tasks, the processing hierarchy may need to be split starting at the middle layers. Overall, the results of this thesis could inform future models that can perform a broader set of tasks on natural images.

Keywords: Convolutional neural networks, grasping, object similarity judgment, multi-task CNN, Representational similarity analysis, ventral stream, the dorsal stream

Summary for Lay Audience

Our visual system enables us to recognize objects and people around us. It also enables us to move and interact with the world. Advances in the field of computer vision have given rise to models that can perform similarly to humans in recognizing objects and people. In this thesis, we study if the same models can also help us judge the similarity of objects or grasp them. Both these tasks require visual processing, but we show that they rely on different features of objects. Our results suggest that the simple models optimized for object recognition are not suitable for producing both similarity judgments and grasping behaviors and their architecture may need to be modified to allow for the human-like production of both behaviors. The results of this thesis shed light on the object properties that are relevant for perception and action and emphasize the importance of studying vision in the context of both perception and action.

Contents

List of Figures

Chapter 1

Introduction

Humans can exploit complex visual information to interact with the world in many different ways. Our visual system has evolved and specialized to our needs to help us move around our environment, recognize different objects and entities, understand their functionality, and guide our actions towards them. For instance, by only looking at a mug and a teapot we can recognize that they belong to two different categories. On the other hand, when we want to grasp these objects we would most likely grasp both the mug and the teapot from their handle. The visual system appears to emphasize different object features when the goal is to categorize objects (e.g., the shape of the objects) than when the goal is to grasp them (e.g., the orientation of the object and location of the handle).

In other words, the visual system is capable of producing two separate outputs from the same visual stimuli to serve two different goals. One goal is to retrieve previously learned categorical information about the objects we see. The other is to determine the posture of our hands and fingers to match the part of the object to be grasped. Even though these goals impose different computational requirements on the visual system, they are supported by our vision with seemingly no noticeable effort. However, the nature of the visual computations leading to these distinct outputs is not yet fully understood.

Historically many attempts have been made to computationally replicate visually guided grasping and object recognition behaviors in the fields of robotics and computer vision, respectively. Despite the remarkable achievements brought by artificial intelligence (AI) models to both fields, the present models fall short of the accuracy and robustness of the human visual system in both tasks. Therefore, given the unique advantage of human vision in solving these tasks, understanding the organization of the human visual system and the underlying mechanism of these behaviors could inform the design of better AI systems. These enhanced AI systems could in turn be used as models to form and test hypotheses about the information processing of the human visual system.

Since the dawn of artificial neural networks, the cross-talk between neuroscience and AI has been fruitful in both improvements of artificial visual systems and the understanding of the human visual system. However, computational modeling of vision has mostly been studied with respect to how the brain enables us to recognize objects and the ability of the current

models to support the grasping of objects has not yet been fully explored.

In this thesis, I first discuss the computational challenges that our visual system is up against when solving object recognition. I next review state-of-the-art AI models of vision to see how inspiration from neuroscience has had a crucial role in the development of these models. I then take a closer look at grasping as a visually guided action, including the computational challenges associated with grasping, and review how current robotic models perform object grasping. Finally, I will cover neuroscience studies to understand nature's solution to the computational challenges posed by object recognition and grasping and draw inspiration toward building better artificial networks.

1.1 Computational challenges of object recognition

As you are sitting at your desk during a normal working day, you can easily recognize the objects on your desk, your colleague who is walking by, and even the words you see on your computer screen. With every slight change in the viewing condition (e.g., changes in luminance, and viewpoint) or with every movement of your eyes and head, the image reflected in your eyes varies. Despite these substantial variations in appearance, we recognize the things we see in a fraction of a second [101], and with no apparent effort. This ability to rapidly recognize objects across different viewing conditions is called "core object recognition" [21] and it is so easily achieved by our visual system that we might underestimate the complexity of the computations required for this accomplishment.

1.1.1 Core Object Recognition Requires Invariance

In a computationally ideal world, each object would evoke a specific pattern of response on our retinas. In that case, each object would have a unique retinal activation pattern, and reading out the object's category name could easily be performed from that pattern of activations. In the real world, however, that is clearly not the case. Any identity-preserving changes to the properties of the object, the viewing environment, and the viewer reflect a unique image of the same object on our retina. Yet, this vast array of visual inputs needs to be assigned the same label. To solve this computational complexity, the visual system needs to develop different types of invariances [82].

For instance, the movements of the visual sensor (eye and head movements) project the target object onto different locations on the retina (position variability). The distance of the object from the viewer causes the object to appear at different sizes (scale variability) [79, 100]. Objects in the real world are usually three-dimensional and a single input of the visual system is a 2D reduction of the 3D object only from one angle [22]. Therefore, objects can be positioned at different angles from the viewer (pose variability).They can appear in different lighting conditions (luminance variability) and in different backgrounds and contexts (clutter variability). Besides, different variations of non-rigid objects or objects that can change shape (e.g., animals, and faces) need to be categorized into the same class (intra-class variability).

For instance, a sitting cat and a walking cat must both be categorized as "cats" despite the major differences in the 3D shape of these two objects [22].

1.1.2 Classification vs Identification

Another stumbling block that makes object recognition difficult is that it refers to different levels of discrimination between objects. On the one hand, the visual system should be able to differentiate a cat from a dog (classification). On the other hand, it must discriminate amongst dogs of different breeds (identification) [82]. To perform object classification, the visual system has to abstract from within-category visual differences and assign the same coarse label to a broader array of objects. Whereas for identification, those within-category differences play a crucial role such that more specific labels are assigned to objects based on intra-class visual differences [21].

Despite these computational complexities, the human visual system can recognize an abundance of objects [10], without cues about the location or properties of an object [22], across various transformations of its image [79, 100], and it does that in a blink of an eye [101, 128, 113]. Therefore, studying the human visual system provides us with an exceptional opportunity to understand the underlying computations, mechanisms, and architectural organizations that lead to successful object recognition. Unfolding the functionality of such a complex system requires hand-in-hand cooperation of different domains such as AI, computer vision, cognitive neuroscience, electrophysiology, and psychophysics. Over the years the border between these fields has been slowly fading as the discoveries of each field have motivated others in uncovering the way we see the world.

1.2 Solving object recognition: computational modeling

In computer vision, researchers have been on a mission to develop models that solve object classification directly from images, and some of these models were closely inspired and enhanced by findings in neuroscience. In this section, I review some of the groundbreaking developments in computer vision, brought about by discoveries in computer science and neuroscience.

1.2.1 The Perceptron: binary linear classification

In 1943, McCulloch and Pitts proposed the first mathematical model of the biological neuron [83]. Years later, in the early 1960s, psychologist Frank Rosenblatt incorporated ideas from this artificial neuron model to build the first machine that learns to perform image recognition [110]. The Perceptron was initially a single-layer linear binary classifier that learned to map the feature vector x to the output value $f(x)$ by updating the vector of weights W (Figure 1.1).

The input vector and vector of synaptic weights w, determine the value of the signal that enters the artificial neuron, and a Heaviside step function determines if the accumulated input signal is strong enough to make the neuron fire. In equation 1.1 $w \cdot x$ is the dot product $\sum_{i=1}^{m} w_i x_i$,

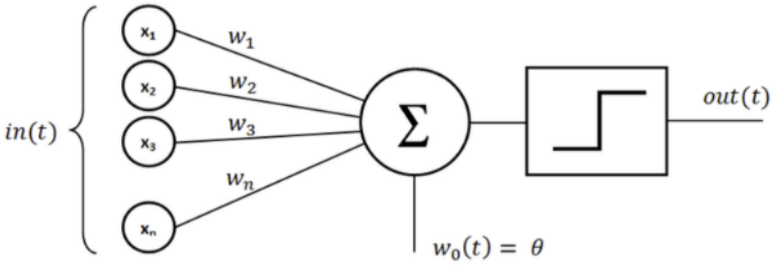

Figure 1.1: Schematic of the artificial neuron used in Perceptron

the number of inputs is denoted by **m**, and **b** represents the bias term that shifts the decision boundary away from the origin, independent of the input value.

$$f(x) = \begin{cases} 1 \text{ if} w.x + b > 0 \\ 0 \text{ otherwise} \end{cases} \tag{1.1}$$

1.2.2 Supervised Learning in the single-layer Perceptron

Supervised learning refers to a learning paradigm, where the model has access to pairs of input data and a ground truth or desired output value in the training phase. Before the learning phase begins, the values of the synaptic weights are initialized with random numbers. Through an iterative process, the model calculates an output for each training example and compares it to the desired output of that example. During training, the model uses a learning algorithm to improve its performance at replicating the desired output using. The learning algorithm of the single-layer Perceptron is as follows.

At every iteration, t, $y_j(t)$ is the output of the Perceptron to the j^{th} training input vector x_j. For each training example, $y_j(t)$ is calculated using equation 1.2, where $\mathbf{w}(\mathbf{t})$ is the weight vector in that iteration and $x_{j,i}$ is the value of the i^{th} feature of the j^{th} training input vector.

$$\begin{aligned} y_j(t) &= f[\mathbf{w}(t) \cdot \mathbf{x}_j] \\ &= f[w_0(t)x_{j,0} + w_1(t)x_{j,1} + w_2(t)x_{j,2} + \cdots + w_n(t)x_{j,n}] \end{aligned} \tag{1.2}$$

The output is then used in equation 1.3 to update the weight vector of the next iteration $t+1$. In this equation, which is called the delta rule [115], $\hat{y}_j$ is the ground truth or the desired output of the Perceptron for the input vector x_j, and α is the learning rate that controls the rate at which the algorithm updates the weights over iterations. It is worth noting that equation 1.3 shows the delta rule for a single neuron Perceptron. In case a Perceptron has more neurons the weights of each neuron are updated individually using the same rule.

$$w_i(t+1) = w_i(t) + \alpha \cdot (\hat{y}_j(t) - y_j(t))x_{j,i}, \text{ for all features } 0 \leq i \leq n. \tag{1.3}$$

This process is repeated for all the training samples in the training dataset. As seen in the above equations, the weights are updated immediately after the output of the Perceptron

is calculated for each training sample rather than waiting until the outputs of all samples are calculated.

1.2.3 Backpropagation: training multi-layer Perceptrons to approximate functions

To solve more difficult tasks, learning more complex representations is required. This can be partially achieved by building multiple-layer Perceptrons. By increasing the number of layers in the network, the number of interconnected weights also increases. In addition, the outputs of the middle layers are hidden, while the delta rule for updating the weights of a layer would depend on the output of those layers. Therefore, the above-mentioned learning algorithm is no longer effective in updating the weights as it is unclear which parameters need to be updated and how. Thus, for training multilayer Perceptrons a different algorithm, called backpropagation [36], is used.

Since there is no function for calculating a well-defined target output for the intermediate layers, we need to define an error function that calculates a loss value based on the difference between the desired output and the predicted output. Similar to the output function the error function E can also be parameterized by the weights of all the layers θ and the input data X. One of the classic loss functions that can be used is mean square error (Equation 1.4).

$$E(X,\theta) = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{1.4}$$

Equation 1.4 Mean Square Error, classic loss function for backpropagation, where y_i is the target value for input-output pair (x_i, y_i) and $\hat{y}_i$ is the computed output of the network on input x_i

The backpropagation algorithm [36] efficiently calculates the gradient of the loss function with regard to the parameters (weights) of each layer for each training example. To update the weights in the hidden layers, the backpropagation algorithm starts from the last layer, for which the output is available. It calculates the derivatives of the error function with respect to the weights entering the output layer from the penultimate layer. Equation 1.5 shows the calculation of these derivatives for each pair of input and output (x_d, y_d) and the weight parameter that enters the j_{th} neuron of the k_{th} layer from the i_{th} neuron in the prior layer. The error derivatives for the model neurons in the last layer are calculated. Further, these derivatives are combined using the chain rule to calculate error derivatives for the prior hidden layer weights all the way to the first layer. This backward chain of computation is called back-propagation since the calculated error in the last layer is propagated back to update all of the parameters in the network's computational graph. The value of these error derivatives (error gradients) determines the magnitude and direction of parameter updates in favor of minimizing the output error.

Since backpropagation calculates the gradients for all the weights at the same time it is far more efficient than directly computing the gradient for each weight separately, as seen in the previous learning algorithm.

$$\frac{\partial E(X,\theta)}{\partial w_{ij}^k} = \frac{1}{N}\sum_{d=1}^{N}\frac{\partial}{\partial w_{ij}^k}\left(\frac{1}{2}(\hat{y}_d - y_d)^2\right) = \frac{1}{N}\sum_{d=1}^{N}\frac{\partial E_d}{\partial w_{ij}^k} \tag{1.5}$$

Equation 1.5, Calculating the gradient of MSE with respect to the weights. $E(X,\theta)$ is the parametrized error function, N is the number of inputs, k denotes the layer number, w_{ij}^k is the weight parameter from the unit i in layer $k-1$ to unit j in layer k

After the error derivatives for each parameter are calculated an iterative algorithm called gradient descent [73] is used to update the weights of each layer using the calculated gradients with the goal of minimizing the loss (Equation 1.6).

$$\theta^{t+1} = \theta^t - \alpha\frac{\partial E(X,\theta^t)}{\partial \theta} \tag{1.6}$$

Gradient descent updating weights of at iteration $t+1$ based on the gradients from the previous iteration (t), where every updating step is called an iteration

Using backpropagation, the model learns to extract features from the training data to serve a specific task. These extracted features, therefore, are determined based on the nature of the input data and the model's loss function and the final weights contain useful knowledge that is specific to the data and task. This knowledge can be transferred to another model in a process called "transfer learning". In this process, the last layer of an already trained network is modified to match the number of classes in a new task. The model is subsequently re-trained on the new training set with a new loss function associated with the new task. Because the initial weights are not random, the model can incorporate some knowledge obtained from the first task in its training for the second one. This retraining phase, known as fine-tuning, can be applied to all, some, or just the weights of the last layer of the network depending on how much prior knowledge needs to be transferred to the next task. Not having to learn from scratch, transfer learning helps models learn a new task better and more efficiently. However, factors such as the similarity of the two datasets and the two tasks, determine how beneficial transfer learning is for different applications.

Overall, Perceptron's initial success in learning simple functions from input-output pairs generated considerable excitement. However, this excitement was eventually tempered as the Perceptron's limitations were discovered. One major downfall of this algorithm is that it only converges if the training dataset is linearly separable. Therefore, the Perceptron fails to solve more complex problems that require nonlinear decision boundaries. In addition, it incorporates a simplified version of a biological neuron and uses a linear algorithm to arrive at useful synaptic weights for classification. Due to this simplification, these models can not fully capture the behaviors seen in real neurons, and therefore fail to explain the complexity of neural behavior [15].

1.2.4 The discovery of the building blocks of the primary visual cortex

Hubel and Weisel were neurophysiologists who were interested in the processing of visual information in the primary visual cortex (V1) of cats [46]. They recorded the electrical activity of neurons exposed to various visual stimuli projected on a screen. During the experiment, they discovered two types of neurons in this brain region and proposed a model for how the cat's brain processes images. The first type of neurons, called simple cells, selectively responded to lines and edges at a particular orientation and spatial location. The second group, named complex cells, were less particular than simple cells in what type of stimuli they respond to. These cells were still selective for edges at a certain orientation. However, their response was equally strong to that oriented line appearing at a number of nearby locations within their receptive field. In other words, they had less spatial sensitivity compared to simple cells that were tuned for oriented lines appearing at a particular location in their receptive fields. Hubel and Weisel proposed that information is first transferred from the retina to simple cells, then the output of simple cells is fed into and aggregated by complex cells to allow for a slightly more abstract response to oriented lines. In other words, such a hierarchical mechanism would allow the system to develop invariances to certain features such as the location of oriented lines in the field of vision, which as discussed above, is essential for core object recognition. Hubel and Weisel's findings about the transmission of information through a network of neurons in V1 inspired computer scientists to build more biologically accurate models of the visual system. The neocognitron was one of the first computational models that incorporated these principles.

In 1980 a computer scientist called Fukushima adapted the idea of the Perceptron with inspiration from the findings of Hubel and Weisel [26]. This model consists of two types of cells. S-cells that are inspired by the simple cells in Hubel and Weisel's findings and C-cells that behave similarly to the complex cells in the cat's V1. S-cells capture basic features from the input by applying a 2-D weight grid on different locations of the input image. Each layer contains several "planes" of S-cells and the S-cells within a plane respond to the same preferred feature appearing at different locations. The responses of all the S-cells in a plane are aggregated into the C-cells using a nonlinear function. With this plane structure, S-cells can respond to specific low-level features in retinotopic layouts, meaning that they respect the topological distribution of stimuli on the retina.

The first layer of S-cells and C-cells is built to mimic the processes in the primary visual cortex. However, Fukushima repeats this organization of S-cells and C-cells several times to build a hierarchical neural network that aims to replicate the behavior of the entire ventral visual pathway. In this model, the response of the first layer of C-cells serves as the input of the next layer of S-cells. The neocognitron's hierarchical processing and use of simple and complex cells made the network less susceptible to shifts in the position of input patterns and able to recognize more complex image patterns. It is therefore considered the predecessor of today's convolutional neural networks (CNNs).

The neocognitron inspired many models of the visual system with similar hierarchical processing. HMAX is one of these models, which calculates the output of complex cells by applying a max operation on the output of simple cells of a plane [109]. Specifically, the out-

put of the simple cells with the same preference but with different receptive field locations are fed into a max pooling layer that extracts the maximum value among those units. This max pooling operation allows the complex cells to develop the same main preference as the simple cells connected to them, yet their activation becomes more invariant than simple cells to features such as spatial location. This developed invariance becomes more abstract as information flows deeper into the network leading to invariances to luminance, size, and viewpoint [39].

1.2.5 Deep convolutional neural networks: modern-day multi-layer perceptrons for computer vision

The neocognitron was the first CNN to introduce the building blocks of modern CNNs: convolutional layers (S-cells), and pooling layers (C-cells). In the convolutional layers, unit activations are calculated by sliding a 2-D grid of weights (filter) over the input matrix (image). The region in the image that a particular CNN unit receives input from is known as its 'receptive field'. For each CNN unit, the dot product of the filter and the image intensities is calculated, and the result is passed through an activation function. Doing this for all units creates a feature map. Units within a feature map tile the whole image and share the same filter. A filter can be thought of as a feature detector. In other words, activity across the units in the feature map (in response to an image) indicates at which location in the image a certain visual feature is present. Feature maps are similar to S-cell planes in the neocognitron (see Figure 1.2). The receptive field of each unit in a convolutional layer responds to a patch of the previous layer and the filter acts as the weight vector for that unit. The value of this filter is initialized randomly and then updated through learning algorithms such as backpropagation. Units that are responsive to the same feature at different locations can share filters.

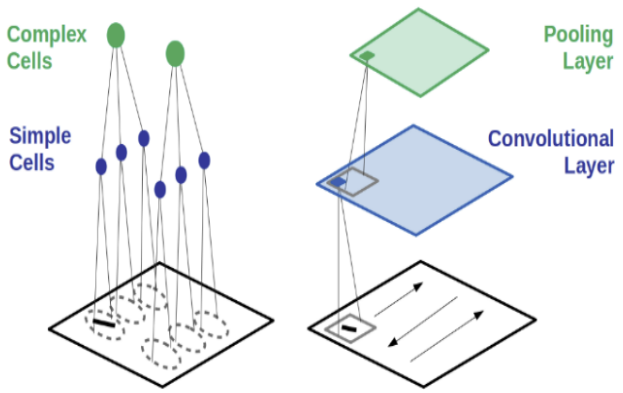

Figure 1.2: Comparison of convolutional and pooling layers in CNNs (Right) with simple cells and complex cells in the primary visual cortex (Left)

The pooling layers are downsampling layers with units that cover patches of the convolutional layers. Similar to C-cell layers in the recognition, this downsampling process makes the network's classification more robust to shifts in position. Higher pooling layers also show

susceptibility to changes in features such as illumination, size, and perspective[39]. The units in this layer can accumulate responses from the units in their receptive field by averaging or selecting the maximum.

In the CNN architecture, several pairs of convolutional and pooling layers are eventually followed by fully connected layers (similar to layers in a perceptron). The number of units in a fully connected layer is equal to the number of categories in the classification problem. The value of each unit in this layer indicates the probability that the input belongs to one of the output categories. Eventually, a softmax function is applied to these probability values to find the class with the highest probability and predict the final label for that input.

After each pair of convolutional and pooling layers, the image becomes abstracted into a new and smaller feature map. Therefore, further layers of the network capture more abstract features of the input. Also when applied to input data with grid-like topology (e.g., images) CNNs are able to consider the spatial relations between different features in their decision. Similar to simple cells and complex cells in V1 the receptive field of the processing units becomes smaller in the higher layers. Overall, the brain-inspired architecture of CNNs allowed them to outperform all previous models by capturing more complex features with fewer free parameters. Moreover, CNNs solved many of the problems associated with traditional neural network models by reducing the number of parameters. These problems included the problem of vanishing and exploding gradients seen during backpropagation [130, 6].

These models first gained popularity in 1989 when LeCun et al. trained a shallow CNN to accurately classify handwritten digits in a supervised manner [72]. Researchers later explored how these networks could be used to solve more complex and naturalistic problems. It was this motivation that led to the development of more naturalistic and complicated datasets such as ImageNet [19]. The ImageNet dataset consists of more than a million real-world images of objects. In 2010, the ImageNet challenge was introduced, which required categorizing ImageNet's test images into a thousand categories of objects. In 2012, a CNN with 8 layers called 'AlexNet' won the ImageNet challenge and demonstrated the power of CNNs in image recognition [68]. Based on AlexNet's success, it was demonstrated that integrating basic principles from neuroscience greatly improved the performance of computational models in visual object recognition.

Since AlexNet, different variations of the CNN architecture have been studied. The main changes to the architecture were the depth of the network [68], the number of units per layer, learning paradigms [43, 8], and the addition of skip connections between layers [19]. Although these models were originally inspired by neuroscientific findings, the development of new models in computer science parted ways with neuroscience. Developing new CNNs was no longer driven by correspondence with the brain, but rather to achieve better performance and efficiency on standard image benchmarks.

In 2012, Simonyan and Zisserman at Oxford [123] trained what they called a very deep neural network with 16-19 layers on a subset of ImageNet. Their model, VGG, achieved state-of-the-art performance on the ImageNet benchmark due to its significant depth compared to

previous models. Increasing the depth of the network results in an explosion in the number of parameters to be trained, and therefore, increases training time. To address this issue the authors fixed and reduced the filter size of all convolutional layers to 3×3 and those of pooling layers to 2×2. Although this idea reduces the number of parameters by a significant amount, it is no longer consistent with the biological fact that the size of neurons' receptive fields decreases from lower to higher areas in the visual processing hierarchy.

Ever since VGG, computer scientists have strived to increase the depth of DNNs while preserving a high performance on different benchmarks. This attempt is primarily motivated by the universal approximation theorem [45]. According to this theorem, a feedforward neural network with only a single layer can represent any function, given enough capacity and computational power. However, if a single-layer network wants to reproduce a complex function, that layer might need to have a massive number of neurons and it can be prone to overfitting. Therefore, researchers have been trying to compensate for a large single layer by building deeper networks.

Stacking more layers on top of each other to make deep neural networks faces two major difficulties. First, adding more layers to a network results in extremely long training times. The second issue is called the vanishing gradients problem, meaning that the gradient that is propagated back to the very first layers of the network becomes very small due to numerous multiplications, which can decrease the overall performance of the network (Fig 1.3).

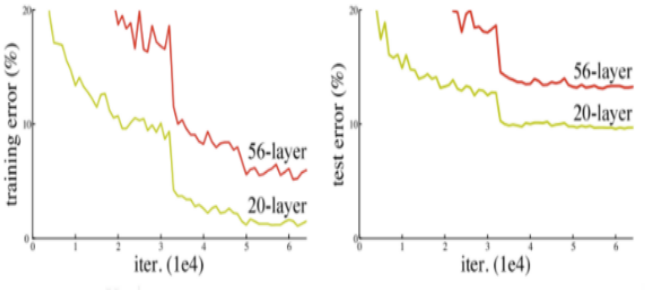

Figure 1.3: reduction of the train (Left) and test (Right) performance over iterations due to vanishing gradient problem in deeper non-residual neural networks. Figure adapted from [41]

Similar to the VGG model, ResNet uses convolutional kernels of fixed size, sacrificing biological consistency to greatly increase training efficiency. However, the most important innovation to enhance the efficiency and accuracy of this architecture is using what is called the "identity shortcut connection". This connection adds the input of a layer to the input of one or more layers ahead as in Fig 1.4. These direct connections from lower layers to higher layers ensure that the gradients from the higher layers at the end of these connections reach the lower layers without vanishing. A group of layers that are connected by this skip connection construct a "residual block" therefore the number of residual blocks equals the number of skip connections. Owing to the skip connections, a ResNet with N residual blocks is as prone to the vanishing gradients problem as a non-residual DNN with N layers. Therefore, ResNet

benefits from more complex deep residual representations without being more exposed than its shallower counterparts to performance decay. The relatively powerful representational ability of ResNet enhanced its performance on visual tasks beyond object recognition, such as face recognition.

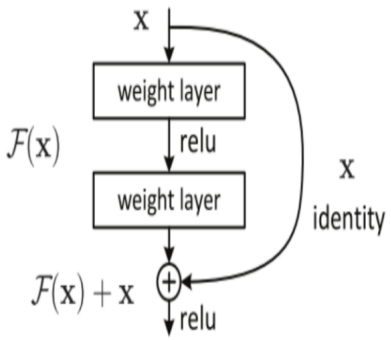

Figure 1.4: Schematic of a residual block in the ResNet architecture. Figure adapted from [41]

Recently, Kubilius et al. [70] attempted to build a model that preserves ResNet's performance on downstream tasks while increasing its consistency with the primate brain. Their model CORnet consists of 4 layers that are explicitly named after their corresponding primate brain regions that are responsible for object recognition: "V1", "V2", "V4", and "IT" (see Figure 1.5). The hierarchy of these regions and their functionality will be discussed in more detail in the next section.

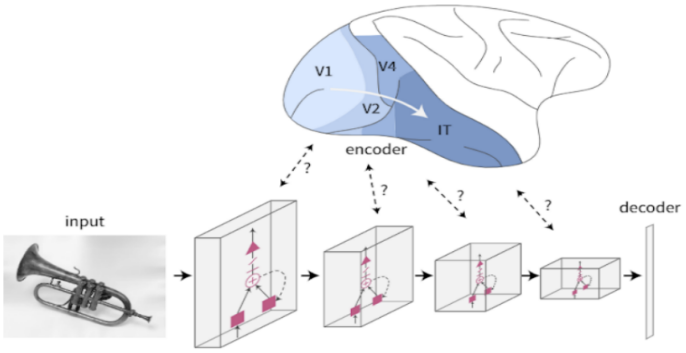

Figure 1.5: architecture of the 4-layer brain-inspired CNN (CORnet). The hierarchical and residual layers of this network (bottom) are directly inspired by the regions in the visual hierarchy in the macaque's brain (top). The figure was adapted from [70]

However, the most groundbreaking innovation in CORnet was adding recurrent connections to each layer, which enabled the relatively shallow network to learn complex tasks. CORnet further was shown to outperform popular state-of-the-art object recognition models on the

"brain score" benchmark [120], while achieving a higher object recognition performance on ImageNet (see Figure 1.6). Brain Score is a benchmark created to measure how brain-like a computational model is. The creation of such a benchmark motivates building better computer vision models without drastic deviations from the biological visual system. The outstanding performance of CORnet on Brain-score and ImageNet once again showed how inspiration from the brain can help build better-performing models.

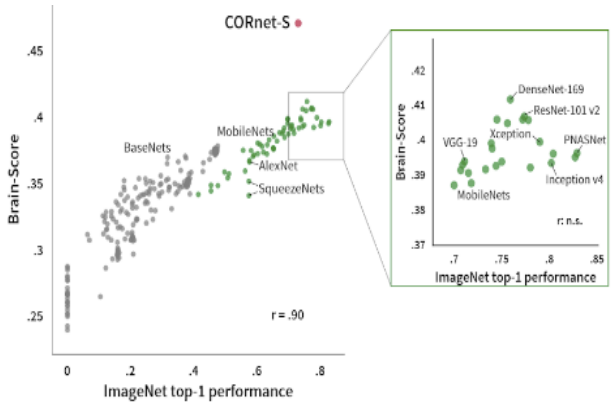

Figure 1.6: Comparison of CORnet (pink) with baseline (grey) and state-of-the-art CNNs (green) models on object recognition and brain-score benchmarks [70]

In sum, CNN models outperform all previous models in performing complex visual recognition tasks, mostly by getting inspired by the visual process in the brain. Their high performance on object recognition as well as the brain-score benchmark suggests that the crosstalk between neuroscience and computer vision could be promising in providing more insight into the computational solution to the problem of object recognition. However, human vision is, by no means, limited to object recognition. In fact, many of our activities in daily life rely on visual guidance. For instance, interacting with objects and grasping them require detailed and just-in-time visual processing.

1.3 Computational challenges of visually guided grasping

Grasping an object might seem like an easy task that humans accomplish effortlessly from early on in their lives. Solving this task, however, relies heavily on the quick processing of complex visual information. For instance, in order to pick up a cup of coffee, you first need to distinguish the cup from the other objects that you see, and the handle from the rest of the cup (if it has a handle). Furthermore, to reach for it, your brain needs to compute the cup's location with respect to your hand. On your way to grasping the object, your brain needs to adjust the shape of your hand to the shape, orientation, and size of the cup's handle right before your hand makes contact with it. Eventually, as your hand gets closer to the cup your fingers start to close in around the handle allowing you to successfully grasp and lift it without spilling any

coffee. To do all this, your brain has to process the visual information starting well before the hand reaches the object.

Now imagine performing the same action with your eyes closed. You might guess the location of the cup and its handle from your previous memory of the scene. However, if the cup is slightly moved, your memory is no longer helpful. In that case, your reach will no longer be as accurate. After a few reaching attempts you can find the cup, as your hand randomly touches it, and adjust the shape of your grasp based on the tactile feedback you get. Thereby, the precise completion of each component relies heavily on rapid visual processing before a movement is made. This visual information entering the visual system is constantly changing as our heads and eyes move. However, our visual system rapidly re-estimates this information from the new visual input until the task is accomplished.

The role of vision in the coordination of prehension has been investigated in several behavioral studies on humans. A study by Marc Jeannerod in 1986 shows a decomposition of prehension into two rather independent components. 1) The reaching movement toward a target object, which requires the visual system to determine the coordinates of a target point in a body-centered space. 2) The adjustment of fingers to the physical properties of the object (e.g., size, orientation, and shape) to form a stable grip, which reflects visual computation performed to calculate these object properties [51].

Although reaching for and grasping an object are usually performed together, studies show disparities between the visual process that underlies each component [50]. For instance, the reaching movement seems to rely on initial visual processing to determine the direction of the reaching; however, after the movement begins it becomes highly dependent on the visual feedback from the object with respect to the hand. Thereby eliminating that visual feedback (grasping in the dark shortly after seeing the target object) drastically affects the reach performance [49]. This is while the adjustment of the fingers to the object and the pre-shaping of the hand seem to rely mostly on that initial visual process prior to the movement onset. Accordingly, the shape of the hand for grasping does not appear to be directly affected by the lack of visual feedback, while it can be affected under difficult reaching conditions to compensate for inaccurate reaching [132]. Consistent with this, Winges et al. [133] show that visual feedback is not necessary for the gradual conformation of the hand to the object's shape despite its direct influence on reaching. Overall, these studies suggest that reaching and hand pre-shaping rely on different visual processes and therefore face different computational challenges. Although we will review some of these challenges for both components, the main focus of this thesis is on the hand pre-shaping component. Therefore, further mentions of grasping in the following chapters refer to this component unless reaching is explicitly mentioned.

1.3.1 Visually-guided estimation of 3D object information for grasping

To reach and grasp an object as accurately as we do, various 3D object properties need to be estimated solely from the 2D visual input. First, the target object has to be discriminated from the scene or other surrounding objects [16], because in grasping, one object is considered a target, and others are treated as obstacles that need to be avoided. Further, to reach for the

object, its exact location with respect to the actor's hand needs to be calculated [49, 132, 133]. When the location of the object, the actor, or the eyes move during the reaching movement, this information needs to be recalculated on the fly. Moreover, as the hand approaches the object it should change shape according to the features of the graspable part of the target object and the degrees of freedom of the hand. The shape and size of the object determine the organization of the fingers and the grip aperture. Additionally, the orientation of the object is required to determine the orientation of the wrist. Calculating each of these features from 2D input is a highly challenging task as they require an estimation of the 3D shape, depth, and pose (orientation with respect to the 3D space) of the objects. However, previous studies show that these features are accurately estimated by the visual system even before the reaching movement starts [49, 132, 133]. How the visual system solves most of these challenges is still an open question; however, it is evident that our eyes use the two images obtained by the left and right eye to calculate an approximate depth map [124]. Additionally, our brain can also use motion signals in dynamic data to estimate depth [58].

Grasp planning: hand pre-shaping. In addition to the complexity of estimating the 3D properties of objects, predicting the best grasp location on the 3D surface of objects from these properties is also quite challenging. An object can potentially be grasped in many different ways, most of which are not stable or appropriate for the goal of grasping (e.g., grasping a hammer to pick it up or to use it). Additionally, features such as the center of mass and weight that are not explicitly available in the input data are also deterministic in choosing grasp locations. Klein et al. [64] explain the complexity of this computation in further detail and suggest that information related to force closure (alignment of fingers with the object at contact locations), torque (distance from the center of mass), natural grasp axis (comfortable and preferred hand posture in humans, which can be influenced by the degree of freedom), grasp aperture (distance between the fingers), and visibility (viewpoint) can explain the human behavior in two-digit grasp estimation.

Adjustments after initiating the movement: Computational challenges involving grasping are not limited to the calculations in the planning phase. The environment in which we interact with objects is noisy and this noise can lead to miscalculations in the initial processing stage. Moreover, we live in a dynamic world, and the position of the target object or the actor might change during the movement. Changes like this will impose an additional challenge on the visual system as it has to rapidly adjust to the new conditions. As shown in previous studies, visual feedback often can be used to adjust reach and grasp errors [132]. Additionally, for adjusting the miscalculations in the hand pre-shaping, haptic feedback from contact with the object can be used as well as visual feedback [133].

1.3.2 Computational disparity between object recognition and vision-based grasping

To recognize an object our visual system has to be selective to features that are crucial in object identification and invariant to changes in other features. For instance, we can detect a pencil on a desk from features such as its bright color, and cylinder shape with a pointed tip. Now if

that pencil is put vertically in a pencil holder we are still able to identify it regardless of the changes in its orientation. Therefore our visual system has to be invariant to changes in features such as object orientation, its distance from the eye, luminance, and viewpoint to successfully recognize objects in different situations. In contrast, the orientation of the pencil plays a determining role in the shape and orientation of our hand when grasping the pencil from the desk rather than from the pencil holder. We would grasp a pen in the pencil holder with the same hand orientation as we would grasp a standing pencil. Therefore, features that are essential for identifying an object do not play as strong of a role as object shape, size, and orientation in our hand preshaping for grasp. In other words, to support grasping our brain needs to develop invariance to a different set of features than the invariance it develops to support object recognition.

In addition, for object perception the brain needs to understand the object properties in a "scene-based" manner, meaning that features such as the size and location of an object are contextually represented with respect to the size and location of other objects in the scene. However, to grasp an object our brain needs to predict the size of an object with respect to the actor's body, especially the hand and arm. Therefore, in grasping, the contextual reference of our brain for feature extraction is "actor-based" [16]. Additionally, in grasping, an object needs to be detected as the target and other objects in the scene are considered obstacles [3], while for object recognition all objects are taken into account as target. Another difference is that object features need to be perceived globally to enable object recognition, while for grasping local processing of object parts is required [27].

These differences suggest that our visual system is facing different computational challenges in object recognition compared to guiding object grasping. Still, it effortlessly supports both behaviors as needed. Therefore looking at how the brain can solve these computational problems can inspire building enhanced bio-inspired models of the visual systems.

1.4 Solving object grasping: computational modeling

One of the main goals in Robotics has been to model grasping, which has given rise to vision-based grasping models that overcome previously challenging tasks. Despite the advancements in robotic grasping, there is a wide gap between robotic and primate grasping, especially in generalizability to unseen and complex objects in real environments [78]. For instance, even when seeing an unknown object for the first time, humans can immediately and instinctively determine how to successfully grasp it. This feat is far from accomplished by the robotic grasp detection models. However, these models have come a long way in achieving this goal with the help of deep learning models.

Robotic grasping studies in the early 2000s mostly used 3D simulations to find appropriate grasps [9, 86, 85, 97, 96]. These models rely on the object's 3D model and prior physical information, while obtaining this prior information and complex 3D is by itself one of the challenges for the visual process. However, the use of powerful feature extractors such as CNNs eliminates the need for these priors as these models can find grasps solely from a single RGB-

D image of an object [76]. The depth images contain information about the 3D depth of each object in the scene which is captured using special depth cameras equipped with infrared sensors [116]. These visual models only require a single view of an object rather than a manually tuned 3D model of it and they can generalize to novel unseen objects with no need for complicated preprocessing and simulations [117, 104]. However, the calculation of depth from 2D information without tools such as infrared sensors is by itself a challenging task that the visual system has to rapidly solve. Additionally, the generalizability of these models has mostly been evaluated on objects that either have very simple shapes or that have a shape similar to the objects known by the model [76]. Human grasping, in contrast, can generalize far better to unusual and unknown shapes.

These data-driven and machine learning-based grasping models can be categorized as model-based and model-free approaches (Figure 1.7) depending on whether they directly predict grasp from visual input or go through an extra information extraction stage before predicting grasps [52].

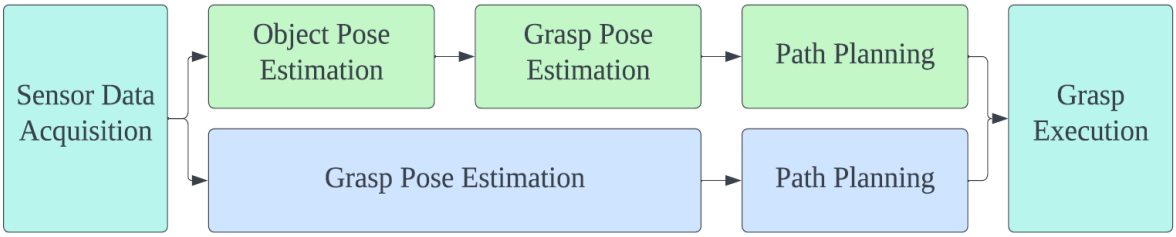

Figure 1.7: Pipelines for Model-based (top stream) and Model-free (bottom stream) approaches to robotic grasping

Model-based approaches: Model-based approaches to robotic grasping consist of two processing stages to predict grasp pose from the visual input: object pose estimation and grasp pose estimation [105, 129, 23]. During object pose estimation a computer vision model, usually a CNN, takes in input directly from the sensors and estimates the 6D pose of objects, which includes information about their 3D position and the 3D orientation [63]. More specifically, object pose estimation allows us to computationally formulate some of the challenges that are involved in visually guided grasping such as estimating the size, orientation, and location of the target. Further, the grasp pose estimation stage also uses deep networks to predict the best grasp pose based on the given object pose. Depending on the type of robotic gripper, the grasp pose can be represented in different ways. For instance, for parallel plate grippers, grasp pose is commonly parameterized using an oriented rectangle as in figure 1.8 where its height is the size of the grippers, its width is the distance between the grippers right before grasping, (x, y) denotes the coordination of its center, and *theta* is the orientation of this rectangle with respect to the horizontal axis. The grasp pose estimation in model-based approaches aims to solve the computational complexities of predicting grasp locations on the surface of objects based

on information about their physical properties. Thereby this processing stage faces the second group of computational challenges associated with grasping.

A major advantage of model-based approaches to grasping is that using object pose information allows the robot to precisely place the object rather than dropping it in a bin. More specifically, the model's access to extra information about the orientation and relative shape of the object informs the robot of how to accurately place grasped objects on a surface. On the downside, obtaining an accurate object pose heavily relies on the model's input diet. Therefore, these models have a hard time generalizing to objects with complex 3D shapes that are different from what they have seen in their training phase. Additionally, they require large amounts of training data that is only accessible through simulation, which makes their training less realistic and less robust to the challenges of the real world.

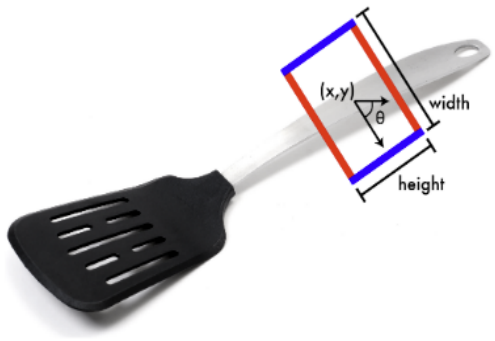

Figure 1.8: 6-dimensional grasp representation for parallel plate grippers. The figure was adapted from [61]

1.4.1 Model-free approaches

Model-free approaches directly estimate grasp pose from the input RGB or RGB-D images. Not having the object pose estimation stage is a blessing and a curse for these models. On the one hand, without the object pose estimation stage, they are trained in an end-to-end fashion and generalize better to novel objects [25]. On the other hand, without prior object-related information, they struggle to accurately place objects after grasping them [62]. These learning-based models use CNNs to either generate the grasp configurations (generative) [90, 74, 91, 106, 107, 125] or choose it from a number of sampled grasp candidates (discriminative) [80, 77, 81, 93, 114]. Current state-of-the-art robotic grasping performance on popular datasets [75, 20] belongs to a model-free approach [1].

In sum, CNNs have revolutionized robotic grasping capabilities in comparison to previous grasping models. They have enabled robots to grasp a diverse set of objects in a completely automatic way without human intervention even in cluttered and non-static environments. Despite these impressive enhancements, robotic grasping and manipulation are still in their infancy. Model-based approaches pick up and place objects accurately but fail to generalize to

novel shapes they have not seen before. Model-free approaches struggle with object placement, while they are better at generalizing to objects with novel 3D shapes. Therefore, none of these models achieve the dexterity and generalizability of human grasping. Even in cases where these models grasp objects successfully, they often grasp them in ways that would seem awkward to humans. These computational and behavioral differences between robotic and human grasping suggest that these models might be processing visual information or predicting grasp configurations differently than humans. Accordingly, to build better robotic grasping models it is important to compare the different parts of these models (visual and motor control processes) to the brain to understand the origin of these differences.

Unlike the rapidly increasing number of studies that compare object recognition models with the brain, robotic grasping has been progressing independently of grasping in the brain. This separation between the neuroscience of grasping and the development of computational models of grasping can eventually result in confusion on how to enhance robotic grasping models toward grasping objects as skillfully as humans. In a recent study, Michaels et al. [84] attempt to bridge the gap by building a model inspired by the modular structure of the anatomical grasping circuit in the primate brain. Their model is trained to predict the muscle dynamics used by primates in grasping objects from images of those objects. They further compared the internal representations of their model to neural data of regions in the macaque brain that control grasping movements to investigate the neural correspondence of their modules with these brain regions. Their model uses an object recognition CNN to extract visual features from RGB images (visual process). Further, a reduced version of the last CNN layer activations is fed into a 3-module grasp prediction recurrent neural network (RNN) that determines the length of 50 muscles in arms and hands (motor process), over time (figure 1.9). According to their results, the three modules can successfully explain the neural dynamics and inter-area differences across the three brain areas in the grasping circuit that they are inspired from (AIP, F5, M1). Therefore their model shows great success at imitating the primate brain in motor control of grasping and transferring visual features to muscle kinematics. However, it is unknown if their model can capture visual features that are required for grasping, similar to humans. This model as well as state-of-the-art robotic grasping models use CNNs that are optimized for object categorization. As object categorization and grasping follow different goals, the visual process required for each task is also different. In other words, to accomplish each of these tasks, the visual system needs to abstract from different sets of features and emphasize others. Therefore it becomes important to investigate whether the CNNs that have neural and behavioral similarities to humans in object recognition, can also capture the visual features important for object grasping. In other words, can CNNs trained in object recognition also emphasize the important features for grasping?

In summary, modeling of vision-based object recognition and robotic grasping has been pursued by the fields of computer science and robotics separately. However, it is unclear if state-of-the-art visual computation models for the perceptual task of object recognition can sufficiently guide robotic grasping. From a modeling perspective, a reliable model of the visual system must be able to replicate the behavior and functionality of the visual cortex. Accordingly, we need to evaluate the ability of current models in explaining visual behaviors beyond object recognition such as guiding object grasping.

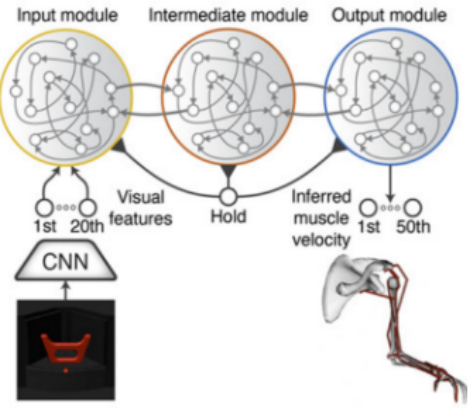

Figure 1.9: Schematic of the brain-inspired modular grasp prediction network by Michaels et al. [84]. RGB images are processed by an object recognition CNN, then the top 20 principal components of the last CNN layer representations are fed to the first RNN module as visual features. The output module predicts the length of 50 muscles over time.

1.5 Nature's solution to solving object recognition and grasping

As mentioned earlier, object perception and visually guided grasping are two separate goals achieved by our visual system. To serve each goal the visual system emphasizes and abstracts from different features, to enhance the information that is specific to that goal. The question is, how can the visual system process the same input in two different ways, in support of such different behaviors? On the one hand, it needs to develop invariance to recognize objects across different viewing conditions and group objects into categories despite differences in visual appearance between examples of the same category. On the other hand, the preshaping of the hand for grasping objects that share particular visual information can be similar, even if they belong to different categories. A large body of research has been conducted to understand how the visual system builds representations of the visual world in serving these separate behaviors.

1.5.1 Ventral and dorsal streams (What and Where/How)

In 1982, Leslie Ungerlareider and Mortimer Miskin [89] concluded from a series of electrophysiology experiments on monkeys and a range of other evidence that visual input is first processed in the primary visual cortex and then it is passed into two separate pathways in the cortex called the dorsal stream and ventral stream (Fig 1.10). Originally, the dorsal stream was thought to specialize in understanding spatial relationships between objects and visual guidance toward them, which is why it was also called the "where" pathway. On the other hand, the ventral stream is evidently involved in object recognition and perception, earning the name of the "what" pathway. A decade later Goodale and Milner [34] argued that dorsal stream functionality is not limited to understanding the location of objects but rather expands to guiding

our actions toward them, changing the name from "where" to "how" pathway. This was concluded from observed changes in the visual ability of primates with lesions to different parts of their visual cortex, which was also supported by single neuron recordings in non-human primates and neuroimaging studies in humans [37, 126, 47, 92].

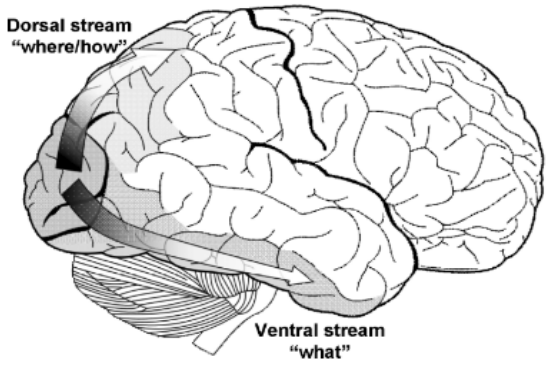

Figure 1.10: Dorsal and Ventral visual pathways. Figure adapted from [37]

Several studies on neurologically impaired patients have provided strong evidence in support of the dualism theory between the ventral and dorsal streams for object recognition and visually guided action.

Visual form agnosia: Visual agnosia is a condition caused by brain damage, wherein the patient loses the ability to recognize objects. A particular form of neural impairment to the ventral stream can cause a more specific condition that only affects the identification of shapes while preserving color coding, visual acuity, and performing actions towards objects such as efficiently grasping them [87, 108]. This impairment called "visual form agnosia" has been of particular interest in investigating the dualism theory of visual pathways since it affects the functionality of one pathway while the functionality of the other pathway remains intact.

Neural imaging on patient DF, who suffered from visual form agnosia, did not show any neural activation related to object recognition from shape cues. However, she was able to grasp objects seamlessly, and her brain activation in areas that are active during grasping was similar to that of healthy participants [48]. In 1991 Goodale and Milner, et al. performed a behavioral study to evaluate patient DF's ability in understanding object properties such as orientation and guiding hand and finger movements directed at the same object [32]. This task had two conditions called "matching" and "posting". In the first condition, the subject has to post a rectangular hand-held card into an oriented slot. On each trial, the orientation of the slot was changed. Therefore an adjustment in the hand orientation was required for accurate posting as the hand-held card approached the slot. In the matching condition, the subjects were asked to match the orientation of the card to that of the slot without reaching toward the slot. The matching condition requires the explicit perception of the orientation of the slot so that the card can be rotated in place to match it. Patient DF rotated her wrist before reaching the target to

perform the posting action accurately, while she performed randomly in the matching condition (figure 1.11). In other words, patient DF could effectively use visual information about the orientation of the slot to guide her actions towards an object, while failing to perceive that same object's property to make an explicit judgment.

These results show that the damage to the ventral stream in visual form agnosia patients is interfering with their object recognition skills and not affecting the visual abilities required for object grasping. This highlights the involvement of the ventral stream in object recognition and its dissociation from the role of the dorsal visual stream in guiding actions.

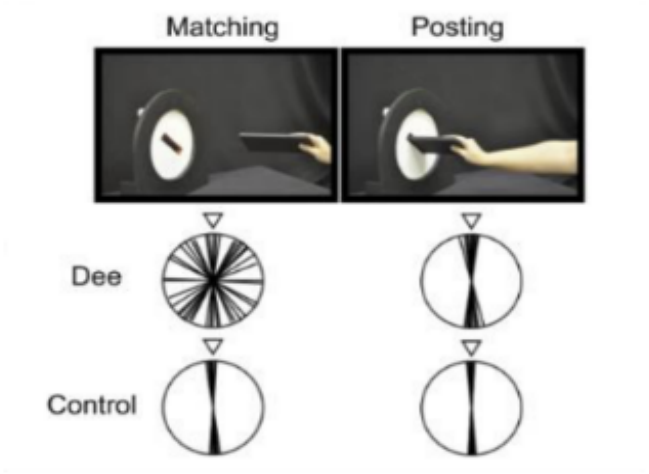

Figure 1.11: Performance of patient DF and control patients on matching and posting tasks (adapted from [37]). Lines show the orientation of the card with respect to the slot.

Optic ataxia: Another neurological disorder that has been insightful about the visual processing in the brain is known as optic ataxia and is caused by damage to the posterior parietal cortex where the dorsal stream regions are located. This condition affects the patient's ability to online visual guidance of action, while the visual perception of object properties (e.g., its location), as well as motor and somatosensory abilities and visual acuity, are preserved [28, 99, 31, 13]. Several studies attempted to characterize this deficit through reaching and grasping experiments.

In a similar experimental setup as the matching and posting task (figure 1.12), optic ataxia patients, who had dorsal-stream damage, had to pass their hand through an open slot in a rotatable disk. Despite accurately reporting the orientation of the slot on different trials, optic ataxia patients made both spatial errors – missing the open slot due to inaccurate reach – and rotation errors – approaching the slot with the wrong hand orientation – when they took action. In another reaching experiment, the authors show that these patients fail to reach toward an object although they can easily describe its location verbally [98]. Other experiments show that optic ataxia patients, unlike control participants, fail to adjust their reaching movement to avoid obstacles on the way to a target object, even though they have an accurate perception of the location of these obstacles [118]. Furthermore, these patients are able to detect sudden

changes in a target's position but cannot adjust their moving hand accordingly [112].

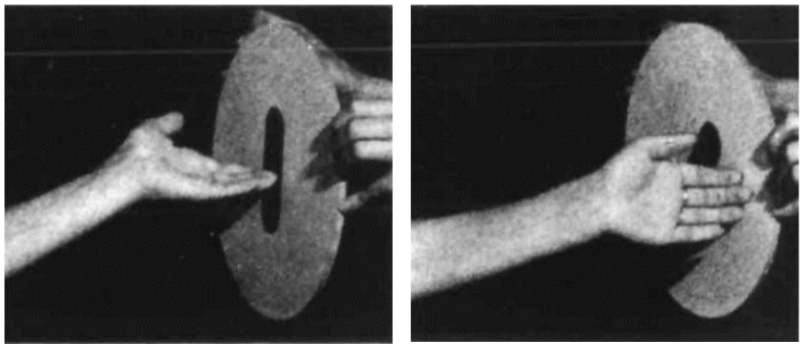

Figure 1.12: orientation (left) and spatial (right) errors in optic ataxia patients (adapted from [98])

Moreover, during grasping, optic ataxia patients show completely opposite deficits than visual form agnosia patients. They show no problem in perceiving the object's shape and size however fail to preshape their hands based on these features before grasping the object. They keep their hands wide open throughout reach and grasp, rather than reaching with an open hand, and gradually close it on the target object as they get close to it. Their reaching pattern and hand shape during grasping is rather similar to that of a blindfolded person grasping the object (see Figure 1.13) [37]. Therefore, these patients have difficulty grasping common objects, if not unreliably, although they are perfectly capable of recognizing and classifying objects [88] while visual form agnosia patients show the exact opposite pattern of deficits.

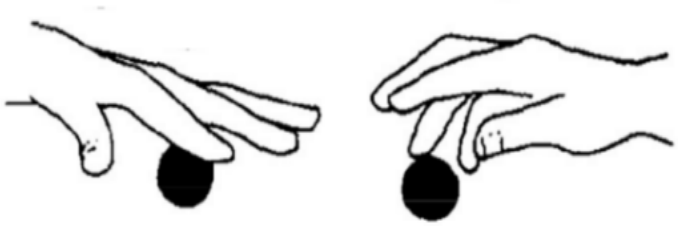

Figure 1.13: Optic ataxia patient failing to scale their grasp to the target object (left), hand preshaping of healthy participants during grasp (right)

Overall these results indicate that optic ataxia is not a problem in visual perception but it is rather a visuomotor deficit that affects the translation of visually perceived information into visually guided actions. Furthermore, the complementary abilities and deficits of visual form agnosia and optic ataxia patients show a double dissociation between visual perception and

visually guided action. Additionally, the fact that these disorders are caused by damage to the ventral stream and dorsal stream areas respectively, provides evidence for the hypothesis that visual perception and visually guided action might be processed through separate visual processing streams. In other words, the brain's solution to solving two different tasks of grasping and object recognition might lie in the divergence of the visual processing system into two separate pathways with separate goals.

Additionally, the result of behavioral studies using concurrent tasks and optical illusions highlight disparities in the underlying visual process required for action and perception [131, 134]. The ventral stream serves the goal of creating a 3D illusion of the world from 2D visual input. To achieve this, it needs to make assumptions about object properties, while these assumptions can, in fact, be different than their actual value in the physical world. For instance, we perceive a projection of a scene on the movie screen as we would perceive the actual scene. The dorsal stream, however, needs to estimate the absolute physical properties of objects in the 3D world with respect to our bodies, since it needs to carry out actions toward the physical world based on these estimations. Therefore, the dorsal stream process has been shown to be robust to optical illusions that affect our perception.

1.5.2 Comparison between humans and computer vision models in object recognition

Despite the great success and the rapid development of artificial neural networks that can recognize objects, these networks seem to be far behind humans in performing visual tasks efficiently and accurately. Many of these networks require substantial computational power to perform a task that our brain does with incomparably fewer resources. Even though many of these networks might show human-level accuracy in a specific task, they fail to generalize to even closely related tasks or perform the same task under different viewing conditions. This generalization is easily accomplished by humans. The human visual system has developed invariances to different visual conditions to overcome these computational challenges. In this sense, comparing these artificial vision models with the human visual system can be highly informative. It is through such comparisons on architectural, behavioral, and neural levels that we gain a better understanding of how visual tasks are accomplished. A number of studies have recently highlighted the similarities between CNNs and the visual system on these levels, validating them as reliable candidate models of this system.

Comparison to the brain on the Architectural level

The idea of simple and complex cells is not the only inspiration for CNNs from biological vision. Similar to preliminary computations done by the retina, the input images to these models are first normalized and separated into three color channels red, green, and blue. In addition, the hierarchy in these networks somewhat resembles the hierarchy in the visual system. Studies on the macaque and human brain suggest that a series of hierarchical regions along the ventral visual pathway contribute to the task of object recognition. According to this view, visual information from the primary visual cortex is passed through a series of regions in the

extrastriate cortex and forms complex categorical object representations in the inferior temporal cortex (IT) [66]. At each processing region along this hierarchy, the output of simpler feature detectors is aggregated into a more complex representation. The increasing size of the receptive fields of the later layers is also consistent with our understanding of the visual system.

These architectural similarities are explicitly incorporated into the design of CNNs. However, to determine whether these artificial networks process information similarly to the brain, we need to test whether they are able to replicate the behavior and neural activity of the visual system.

Comparisons of CNNs and the brain at the neural level

The hierarchical organization of feed-forward CNNs enables us to compare responses to neural network layers and regions in the visual system at different processing stages. In 2014, Yamins et al. recorded the extracellular activity of the macaque's brain while viewing images of complex objects. They further fed the same images to CNNs and extracted the output from the pooling layers of these networks. To validate if CNNs perform the same computations as regions in the brain, they fit linear regression models to predict the activity of brain regions from the activity of the CNN layers. They tested the model solutions on a held-out test set to validate the neural foresight of the model's layers [135]. They showed that CNNs with better object recognition accuracy could also predict neural activity more accurately. Furthermore, activation of the penultimate layer was the best predictor of mid-layers in the visual hierarchy (such as V4), while that of the last layer outperformed its predecessors in predicting activations of later layers in the visual hierarchy (such as IT). This correspondence of hierarchical processing between CNNs and the macaque brain was also found in human fMRI [38] and MEG [121] studies. According to the early to mid layers of object recognition, CNNs are better predictors than traditional V1 models of this region's neural activity [14].

Aside from regression, a method called representational similarity analysis (RSA) [67] has been widely used to measure the correspondence between the responses of populations of biological neurons and artificial model units. This method requires building a $n \times n$ representational dissimilarity matrix (RDM) for each population, while n is the number of stimuli. An RDM is a matrix that assembles the response pattern dissimilarities between all pairs of stimuli and can be used as a proxy for representing the properties that differentiate stimuli in that population. The correlation between the two RDMs shows how similar those properties are among the two populations. Khaligh-Razavi et al. [55] compared intermediate responses of AlexNet trained on ImageNet to regions in the human and monkey brain using RSA. They found a better correlation between AlexNet's last layer and several higher areas of the IT cortex than previous models.

A major benefit of RSA is that it is applicable to any type of response population. The output of different methods and models can be easily transformed into RDMs. Further, unlike regression analysis, which compares single neurons and units at a time, RSA compares the entire population at once. Therefore, The maximum similarity in RSA is achieved when the two compared response populations are entirely similar. However, the regression method allows

us to assign weights to the model's features based on how relevant they are in predicting the neural response. This weighting makes the regression analysis a more flexible method of comparing two populations compared to RSA [65, 127]. More specifically high similarity values can be achieved using regression analysis even if the entire populations are not similar and only a subset of the input population is similar to the entire output population.

One major accomplishment of CNNs has been their ability to explain neural data from higher areas in the ventral stream (e.g. V4 and IT). This is a particularly interesting ability of CNNs since the complexity of responses from these regions had made them extremely difficult to model using other methods compared to V1.

These results suggest that CNNs produce representations that resemble those measured along the macaque and human ventral visual stream. The uniquely high correspondence of CNNs with the neural activity of the brain has made these models popular among neuroscientists. A good computational model of the visual system can be used by neuroscientists in several ways. Using these networks, they are able to form new hypotheses and perform various kinds of experiments in a controlled and measured environment, allowing them to explore and validate different theories about how the visual system works. Recently CNNs have been used for these purposes as the best available models of the visual system.

Comparison of CNN output to human behavior

CNNs have shown comparable average accuracy to humans in object categorization. However, average object classification performance is not the only way to compare the behavior of these networks with humans. Specifically, since these models are explicitly optimized to increase this performance metric. Other performance metrics could be used to assess whether CNNs and humans consider similar features and characteristics for performing a task. With CNNs being able to classify images as accurately as humans, one can examine if these networks make the same mistakes as humans do.

For instance, Rajalingham et al. [103] took a closer look at the network's judgment by comparing its confusion matrix to that of animal behavior. For n classes, a confusion matrix is an $n \times n$ matrix indicating how often instances of one category have been misclassified as belonging to another class. In this study, it was found that CNN's confusion matrix matched that of animal behavior in the same task, showing that CNNs make similar mistakes as monkeys and humans in categorizing objects.

In other studies, human subjects were asked to rate the similarity between images of different objects instead of categorizing them [59, 53]. The output of these similarity judgment experiments resembles implicit information about the properties that are influential in the judgments. Jozwik et al. performed an RSA comparison between CNNs and conceptual models of human perception in predicting similarity judgments [53]. As a result, the last layer of deep networks successfully explained the category similarity judgment behavior outperforming the earlier layers. This is again consistent with the brain in the sense that category similarity judgment requires more abstract and high-level processing and is best explained by higher ar-

eas in the visual cortex such as IT. Another study [111] has explored the predictive ability of CNN representations on a more challenging similarity judgment task. Their results suggest that CNNs struggle to replicate human behavior when more complex elements of similarity are used. Moreover, visual psychophysics researchers have evaluated CNNs on a number of other behavioral principles such as typically [71], Gestalt [57], and animacy [11], suggesting that CNNs demonstrate similar behaviors to humans with respect to many of these components.

Despite these similarities, these models have shown behavioral differences from humans in a number of domains. For instance, some studies highlight that CNNs rely heavily on texture rather than shape when classifying images [4, 34]. This is while an older study argues that CNNs can be considered models of human shape sensitivity [69]. The fact that CNNs outperform humans in some tasks highlights another behavioral difference between deep networks and humans [56]. While this behavioral mismatch is desirable among computer scientists, it reduces the usefulness of such models for neuroscientists.

In line with these differences, Geirhos and Meding et al. [29] evaluated the state-of-the-art brain-like model CORnet to see if this model makes similar errors to errors made by humans. They compared the error consistency of CORnet and top-performing CNNs (e.g., ResNet) with humans. According to their results, humans show low inter-subject variability in their errors. Interestingly, the human error consistency of CORnet and ResNet almost completely overlap, meaning that they have similar behavioral strategies in object recognition that are quite different from humans. One side effect of this difference could be that these CNNs can be victims of adversarial attacks to which humans are completely robust. For instance by applying a certain type of noise to the input image, object recognition models fail to correctly recognize the objects, while humans do not notice any change in the input [2].

To address this issue Dapello et al. [17], created a hybrid CNN called VOneNet that consists of a neural network with fixed weights as its first hidden layer, and a CNN back-end. The first part, called the VOneBlock incorporates biological constraints of the primary visual cortex of primates using a classic neuroscientific model of this region. The VOneBlock can be followed by different CNNs such as AlexNet, CORNet, or Resnet. The motivation behind this work comes from the fact that CNNs that explained the neural data of the primary visual cortex (V1) demonstrated more robustness to adversarial attacks, suggesting that V1 incorporates strategies that are crucial to developing such robustness. They further showed that the addition of VOneBlock can significantly increase the robustness of any CNN to adversarial attacks without any explicit training against these attacks.

Humans make similar errors as humans (red), but recurrent CORnet-S (orange) makes almost exactly the same errors as feedforward ResNet-50 (blue): the two networks seem to implement a very similar strategy, but certainly not a “human-like” one according to error consistency analysis. (Note, however, that it really depends on the dataset and metric: CORnet-S shows promising results in capturing recurrent dynamics of biological object recognition, for example.) It seems that recurrent computations —which appear to be of particular importance in challenging tasks — are no silver bullet. While recurrence is often argued to be one of the key missing ingredients in standard CNNs towards a better account of biological vision,

a recurrent network does not necessarily lead to a different behavioral strategy compared to a purely feedforward CNN.

1.6 Two-stream computer vision models

Inspired by the two visual streams in the brain, only recently a number of computer vision studies started to investigate the use of dual pathway CNNs in support of multiple visual tasks. Scholte et al. [119] optimized a single CNN for two tasks simultaneously and investigated the contribution of each layer's processing units to each task. In one condition the objectives of the multi-task CNN were relevant to each other (ordinate and subordinate categorization) and in another condition, the CNN was trained on two unrelated tasks (object and text label classification). Relevant tasks were chosen such that they would require similar features to be extracted or ignored by the network. Further, they calculated the contribution of each unit in each layer of the related (RelNN) and the unrelated (UnrelNN) neural networks to each task. Interestingly they showed that in the UnrelNN the units in each layer become selectively responsive to one of the two tasks, whereas in the RelNN the units of all layers contribute to both tasks evenly. This divergence among the units of UnrelNN becomes more significant along the hierarchy of the network as more abstract features are extracted. Therefore, the network is able to build different invariances specific to each task. Overall their results suggest that as two outputs of a dual-task system get more distinct, the computational need for two separate processing pathways increases.

In a 2021 study, Bakhtiari et al. trained a CNN with two parallel pathways and a single self-supervised predictive loss function on videos [5]. They further compared their two-stream model with a single-stream model on two downstream tasks of motion discrimination and object categorization and evaluated the neural correspondence of their models with the brain responses of mice to these two tasks. Interestingly their results show that the activations of a two-stream architecture can better resemble neural responses to both tasks, while each stream is specialized to one of the tasks. However, the single stream network with the exact same objective could only explain neural data related to the recognition task. This study suggests that even with a single generic loss function, having two separate visual pathways enables the model to achieve better accuracy on both downstream tasks, while explaining the neural responses to both tasks.

1.7 This thesis: to what extent can one system solve both tasks?

Putting together the evidence from previous studies, this thesis follows a two-fold objective. Focusing on the two visual abilities of grasping and object perception, we first investigate the similarities or relatedness of these two tasks. Further, we see to what extent the hierarchical CNNs, widely used as computational models of vision in robotic grasping and object recognition, can explain human behavior on each task. The result of this thesis sheds light on the

reliability of these networks in guiding robotic grasping while enabling us to form concrete hypotheses about the organization and information processing of dorsal and ventral streams.

Chapter 2

Predicting Category Similarity and Grasping Behavior from CNN Layers

2.1 Introduction

Our daily activities rely heavily on our visual understanding of the world. As we look around, our visual system provides us with detailed information about our surroundings. We use this information to recognize the objects we see and to accurately guide our physical actions (e.g., grasping) toward them on the fly. Supporting a wide range of perceptual and active tasks is computationally challenging as for each task different object properties might be important. For instance, two objects from the same category may be grasped differently (e.g., grasping a cup with and one without a handle), while we might use similar grasping configurations of objects that belong to different categories (e.g., grasping a hammer vs a screwdriver). Therefore, the system has to extract different features of the same objects depending on the task. Some of these tasks are more computationally related to each other and, in these cases, it is more likely that the visual system supports both tasks by extracting the same set of features. As the tasks become more distinct the properties that are essential to solving each task become less overlapping. In other words, to support computationally distinct tasks the visual system needs to extract distinct features depending on the objective of the task. This raises the question of how objects are represented along the visual hierarchy to support a wide range of tasks.

Findings in the neuroscience literature bring insight into how the visual system generally solves this computational problem. There are extensive electrophysiology, neuroimaging, and behavioral evidence in humans and macaques that suggest the visual system is separated into two separate processing pathways, each serving different goals: the ventral pathway that is responsible for supporting perception-based behaviors (e.g. object recognition and similarity judgment) and the dorsal pathway that is involved in localizing objects and guiding actions towards them (e.g., grasping) [89, 34]. Despite this evidence, it is unclear how such distinct behaviors are represented along the dorsal and ventral streams and to what extent these representations overlap along the hierarchy of the streams.

Recent advances in Artificial Intelligence (AI) systems have made them valuable tools for running controlled experiments on models of the visual system and forming hypotheses about

the characteristics of the representations in the visual system [22, 53, 44]. As we reviewed in the previous chapter, deep convolutional neural networks are getting increasingly better at explaining the behavior and neural activation of the ventral visual pathway in humans and monkeys [70, 17, 120]. Additionally, their architecture allows researchers to access the internal representations that are formed in these systems. In particular, they have been useful in characterizing the representations leading to behaviors such as object recognition. Although most of the focus in these studies has been on object recognition, the ability of these models at explaining neural activity and behavior in primates extends to other visual perception tasks such as object similarity judgment. Object similarity judgment is well predicted by the representations in the higher regions in the ventral stream [40, 18, 24, 94]. Consistently, higher layers of Deep Convolutional Neural Networks (DCNNs) can predict human similarity judgment of real-world objects from a wide range of categories without being explicitly trained to do so [54].

Beyond perception tasks, DCNNs have been widely used in robotics to guide object grasping and outperform previous models of robotic grasping [52]. Additionally, these models can automatically capture features about the target object from depth images, eliminating the need for complex priors about the objects. This ability has further allowed these models to generalize to grasping unseen objects better than any previous models. Despite the impressive improvements made by these models to robotic grasping, they still fail to grasp objects as rapidly, accurately, and steadily as humans [16]. Moreover, they often rely on categorical information about the objects, which may not be necessary for accurate grasping in humans. Behavioral differences between robotic grasping models and object grasping in humans raise the question of whether the DCNN representations are able to support interactive behaviors such as grasping.

Studies show that representations related to several category-orthogonal attributes such as 3D pose, size, and orientation are explicitly present in the IT cortex and the higher layers of a hierarchical DCNN trained on object categorization, and this information increases along both the DCNN and the ventral stream hierarchy similar to object category representations [44]. Numerous neuroscience studies have emphasized the importance of these attributes in the pre-shaping of the hand before grasping objects. These results suggest a hypothesis that the higher-layer representations in a single DCNN architecture can extract sufficient information to support grasping. For this hypothesis to be true, the representations in higher layers of DCNNs should be good predictors of the grasping behavior. Meanwhile, an investigation on dual-task DCNNs, optimized for two tasks, suggests that as two outputs of the DCNN get more distinct, the computational need for two separate processing pathways increases [119]. These results suggest that the extent of overlap between the outputs of a system determines if a single computation suffices for serving both outputs. The amount of overlap between similarity judgment (perception-based) and object grasping (action-based) behaviors have not been thoroughly investigated. In general, it is unclear to what extent the representations captured by a single DCNN can explain action-based and perception-based behaviors beyond object categorization.

Inspired by these results, we investigate the ability of different DCNNs in explaining the

perception-based behavior of object similarity judgment and action-based behavior of grasping, without explicit optimization for either task. We collected these behaviors on the same set of objects and used state-of-the-art DCNNs to predict each behavior. For the grasping behavior, we asked human subjects to grasp a large set of 3D-printed everyday objects. Object similarity judgment behavior was collected using an odd-one-out experiment to obtain pairwise similarity judgment for pictures of the same 3D-printed objects. We measured the extent of the overlap of the two behaviors using a Representational similarity analysis (RSA). Next, we measured how well we could predict each behavior from the internal representations of DCNNs along the hierarchy.

2.2 Materials and Methods

Stimuli: The objects used for our behavioral experiments were chosen from a list of object categories in the THINGS dataset [42]. We eliminated non-graspable objects such as animals, foods, and large objects and chose object categories for which a 3D model was available on the www.thingiverse.com website. Using human annotation obtained from Amazon Mechanical Turk experiments we eliminated object models that were not recognizable to humans on Amazon Turk, leaving us with 58 object categories. We then adjusted the size of these objects to fit in a box of size 15cm and 15cm such that the objects could fit on a table inside an fMRI scanner for future neuroimaging experiments. Objects of these 58 categories were 3D printed using a white plastic material for the grasping experiment and they were further photographed for the similarity judgment experiment and the DNN analyses. 10 Images of natural objects in the chosen categories were used from the THINGS dataset of natural images for transfer learning in DNNs and categorization experiments in humans and models.

2.2.1 Object similarity judgment:

Participants: For the odd-one-out task, we used the Amazon Mechanical Turk platform to perform the online experiment. A total number of 653 subjects participated in our experiment. Participants gave informed consent prior to the experiments.

Procedure: Object similarity ratings were measured using an odd-one-out experimental design. In each trial images of three objects were presented to the participants asking them to choose the objects that are odd or most distinct among the three. Out of the 58 objects, all possible combinations of three were chosen for the trials and each trial was repeated twice to calculate a measure of reliability. Trials with extremely low (< 900ms) or extremely high (> 12000 ms) reaction times were cleaned from the data. The removed trials were repeated until two repeats of each trial were obtained. The experiment was conducted once using images of the 3D-printed objects (see Figure 2.1).

Analysis: To measure pair-wise similarity between object categories, we assigned a value of 0 to the similarity between the chosen object and the other two objects in each trial, while assigning a value of 1 to the similarity between the unselected objects. After repeating this procedure for all possible object triplets 112 similarity values were obtained for each pair

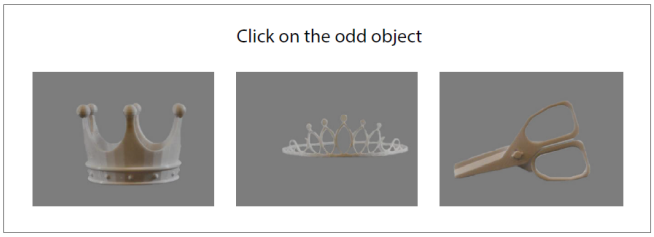

Figure 2.1: The participant's view of the odd-one-out task using images of 3D printed objects. Objects belong to the categories called "scissors", "tiara" and "crown" from left to right, respectively.

of objects. Further, we aggregated the similarity values across trials, resulting in a 58×58 dissimilarity matrix A, in which $A_{m,n}$ represents the probability that the object m and object n are chosen as a similar pair (see Figure 2.2).

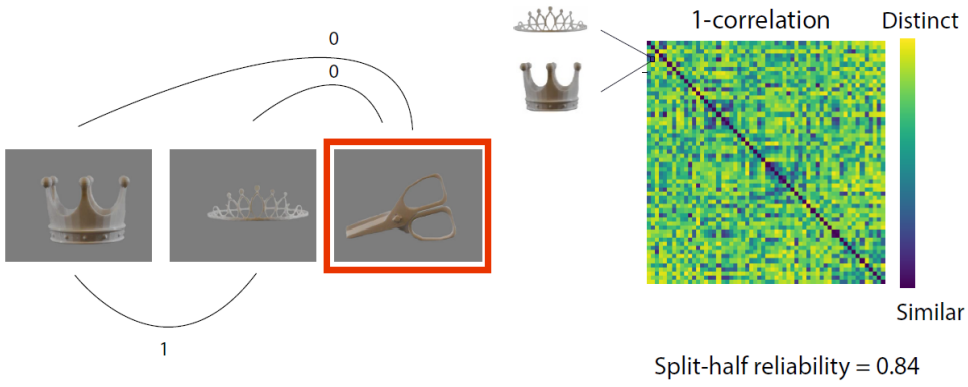

Figure 2.2: Building a representational dissimilarity matrix (right) from the odd-one-out experiment (left)

Reliability: To test the internal consistency of the collected data, split-half reliability was calculated by dividing the data into two subsets wherein each subset included the data for half of the participants. Data of these two halves were compared using the Spearman-rank correlation coefficient resulting in the reliability value.

2.2.2 Object categorization experiment

Participants: For the categorization task, we used the Amazon Mechanical Turk platform to perform the online experiment. A total number of 1152 subjects participated in our experiment.

Participants gave informed consent prior to the experiments.

Procedure: This experiment was designed as a 58-alternative forced choice experiment to obtain the categorization accuracy of human participants on images of naturalistic and 3D-printed objects. In each trial, subjects were presented with a target image belonging to one of 58 categories and a naturalistic image of all 58 categories as choices (Figure 2.3). The participants had to select the image that belongs to the same category as the target image. Target images were never presented in the choices. We repeated each trial 25 times. 12 percent of trials were removed that had extremely low (< 900ms) or extremely high (> 12000 ms) reaction times.

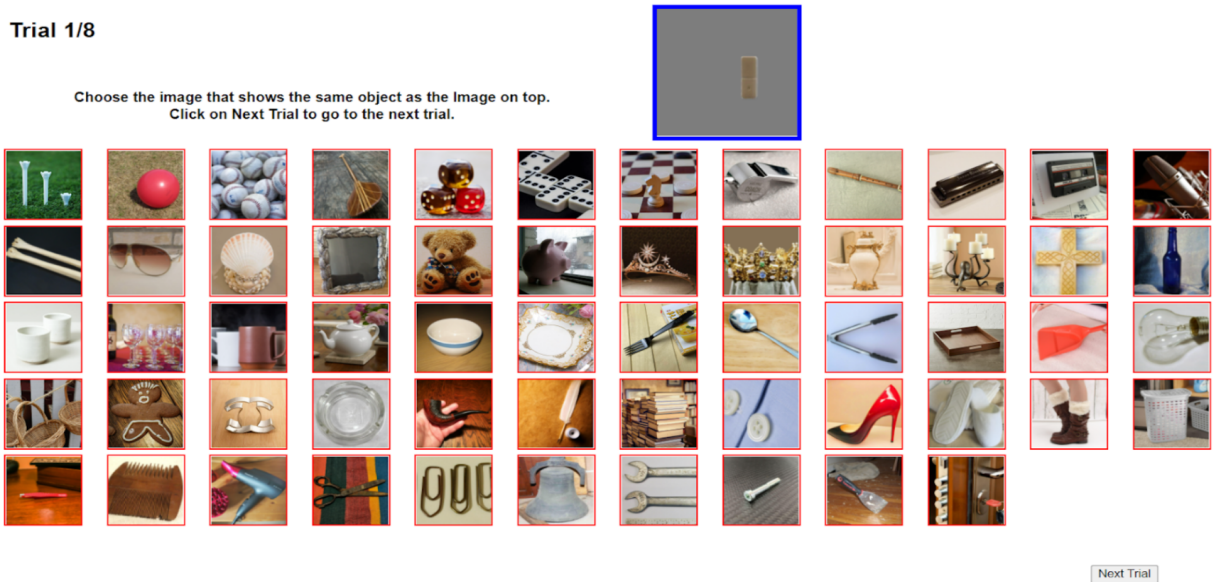

Figure 2.3: Participant view of the object categorization experiment on Amazon Mechanical Turk. The target object is in a blue frame and belongs to the category called “chess piece”. 58 objects in the red frame are the choices and each represents the category they belong to.

2.2.3 Grasp experiment

Participants: Fourteen adults (8 females, 6 males), ages 18 to 35, participated in the object grasping experiment. All participants were right-handed and had normal or correct-to-normal vision. All participants were in good health and had no history of psychiatric or neurological diseases. Participants gave informed consent prior to the experiments.

Procedure: Hand movements were tracked using a Polhemus Liberty electromagnetic position and orientation measuring system with an update rate of 240 Hz. 16 small position-tracking sensors were attached to the hand, fingers, and wrist (see Figure 2.4 for sensor placement). Prior to grasp, all objects were placed centrally on a plastic table that sat comfortably around and over the participant’s lap. Table height and distance were adjusted to ensure comfortable grasping positions for each participant. We used the 3d position of the 16 sensors

(concatenating the x, y, and z coordinates with respect to the starting position) to obtain vectors pertaining to the shape of the hand when grasping each object. The coordinates at the end of the grasping movement were considered for analysis since the fingers at the end of a grasping movement are formed to match the shape of the parts of objects that are used to grab onto the object. Accordingly, this measure can be used as a proxy for the object features relevant to performing grasp movements.

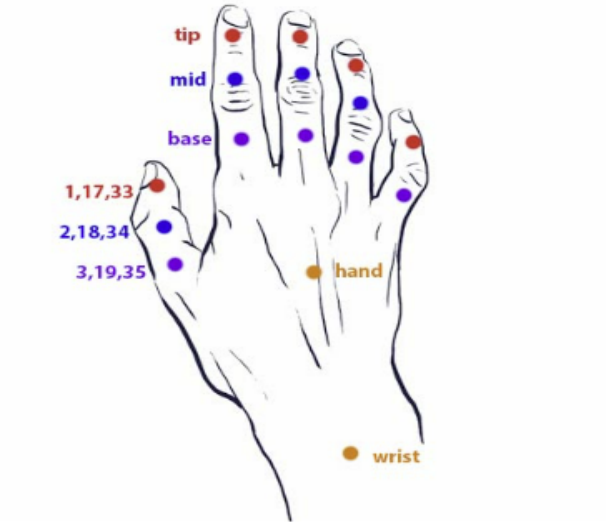

Figure 2.4: Sensor placement in the object grasping experiment

Analysis: A 48×58 matrix was obtained from this experiment, in which 48 denotes the final x, y, and z coordinates of the 16 sensors and 58 denotes the number of grasped objects, each from a different category. Further, based on the RSA method, Pearson's correlation coefficient between the response to every pair of stimuli was calculated to obtain a 58×58 dissimilarity matrix for the grasping behavior. Multi-dimensional scaling (MDS) [122] was used to reduce the dimensionality of the representational dissimilarity matrix (RDM). Using Pearson's correlation coefficient, sensor positions for different objects were centered before comparison, such that a high similarity value would be assigned to similar grasp postures at different positions

Reliability: To test the internal consistency of the collected data, split-half reliability was calculated between subjects. Data splits were compared using the Spearman-rank correlation coefficient resulting in the reliability value.

2.2.4 Hierarchical Clustering Analysis of the behaviors

After obtaining the RDMs for each behavior. Multi-dimensional scaling (MDS) was used to reduce the dimensionality of each RDM. The first n dimensions that explained at least 95 percent of the variance of each RDM were passed on as the input of an agglomerative hierarchical clustering algorithm to see how different objects are arranged in the embedding space of each

behavior. Hierarchical clustering is an unsupervised machine learning algorithm that takes an unlabeled dataset as input and groups the data points in that dataset into clusters based on a similarity metric [95]. One of these algorithms is called agglomerative hierarchical clustering which clusters a given input dataset in a bottom-up manner. Initially, each point in the dataset is considered as a single cluster and similar pairs of clusters are merged iteratively until only one cluster remains that contains all the points in the dataset. The output of this algorithm is a tree-shaped structure called a dendrogram that visualizes a hierarchy of clusters. In this study, we applied the hierarchical clustering analysis using the SciPy library with the cosine similarity measure for grasping and object similarity judgment behaviors.

Comparing the behaviors: RSA analysis was used to quantitatively compare the similarity judgment and grasping behaviors. In this method, representational dissimilarity matrices are calculated for each behavior. Further, the correlation of the off-diagonal values is computed as a measurement of the similarity between the behaviors.

2.2.5 Predicting behaviors using DCNNs:

Six different neural networks were chosen for this study including three classical DCNNs (AlexNet, VGG, ResNet50), two state-of-the-art brain-inspired deep neural networks (CORnet-S, VOneNet), and a novel vision transformer (Clip-ViT). Vision transformers were chosen since they are famous to be the best models to generalize over the unseen distributions and classes [102], and biologically inspired feedforward models were chosen because they are considered the best models of the ventral visual stream, which is responsible for human object recognition. All models were pre-trained on the ImageNet one thousand class object recognition dataset. However, since our object categories were not included in the ImageNet dataset, we fine-tuned the models on a set of naturalistic images of our chosen categories from the THINGS dataset. For testing, we fed the images of 58 3D-printed objects into the network and extracted the activation of each hidden layer. The max-pooling layers of the networks were chosen for the analysis. For some of the networks in which the pooling layer activations were not extractable, the output of each block of the network was used instead. We used activations from each layer to predict the two behaviors of object similarity judgment and grasp.

In order to predict the behaviors from the activations of the different layers of DCNNs, we used a stepwise Ridge Regression model with L2 regularization. We first used classical Multi-dimensional scaling (MDS) [12] to reduce the dimensionality of the two behavioral dissimilarity matrices. Next, we fitted the regression model to the output of each layer of DCNNs and the resulting reduced space for each behavior (6 dimensions for grasp and 15 dimensions for similarity judgment). Specifically for each DCNN layer and each behavior, we held out the model activation and behavioral data for some of the objects ($\frac{1}{5}$ of all the datasets). We then trained the multiple regression model to predict the reduced behavioral spaces from the DCNN layer on the remaining objects. Then we tested the regression model on the data for held-out objects. This process was repeated by holding out other objects until the reduced behavior spaces for all of the object categories were predicted. The regularization parameter was tuned for each regression model separately and the Spearman's rank correlation coefficient between the predicted and the MDS-reduced behavioral data was reported as the performance of that DCNN layer in predicting that behavior. For each behavior, the multiple regression

model predicted the top n dimensions of MDS that explain 95 percent of the variance of behavioral RDMs. In the future iteration of the study, the same analysis can be applied to predict all of the output dimensions of MDS, weighted by the variance explained by each dimension (eigenvalues).

2.3 Results

2.3.1 Behavioral results

To investigate whether a single visual process of DCNN can support both behaviors well, we need to first understand how the different objects are arranged in the output space of each behavior. For example, we expect objects that are semantically similar to be closer to each other in the object similarity judgment space as humans have been shown to use categorical information in making judgments about object similarities without being explicitly asked to do so [54]. Additionally, we expect objects that are grasped using similar hand configurations to fall close to each other in the grasping output space. To visualize the embedding space for each behavior we applied hierarchical clustering analysis to the MDS-reduced behavioral data. The result of the HCA highlights how participants differentiated objects for the two tasks of grasping them or similarity judgment.

Grasping output space: To evaluate the grasping behavioral data we calculated its split-half reliability, which was above 0.78, this suggests that the subjects' responses were consistent across the repeated trials. The results of the dimensionality reduction showed that 90 percent of the variance of the grasping behavior can be explained using the top 6 dimensions of the MDS, highlighting that the grasping output space is relatively low-dimensional. This is mainly due to the limited degree of freedom in our hands. More specifically, since our fingers can only move in restricted directions in 3D space, the grasping configuration for the 58 objects is represented using a small number of dimensions. The hierarchical clustering analysis on the reduced 6-D space showed that objects that are similar in size or orientation are more likely to be clustered together (see Figure 2.5). Additionally, objects of different sizes that have a similar commonly graspable part (e.g. a handle on a teapot and a mug) are also grasped similarly and fall closer to each other in the grasping embedding space.

Object similarity judgment output space: The split-half reliability for similarity judgment behavior was 0.84. The higher reliability of this behavior in comparison to the grasping behavior may be related to the differences in the nature of these behaviors: a single object can be stably grasped in different ways, while participants' judgment about the similarity of an object to the others is less susceptible to change over trial repetitions. It might also be related to the fact that the data from the similarity judgment is obtained from a larger pool of participants (326 participants in each split half, albeit each participant only saw a subset of objects) compared to the grasping experiment (7 participants in each split half). The output space of the similarity judgment behavior was found to be relatively higher in dimensions compared to the grasping behavior. The top 15 dimensions of the MDS explained 90 percent of the variance for this behavior. This reduced behavior served as the input to the hierarchical clustering analysis. Looking at the resulting dendrogram (Figure 2.6), it is evident that objects that are semantically and categorically similar are clustered early on despite their differences in other

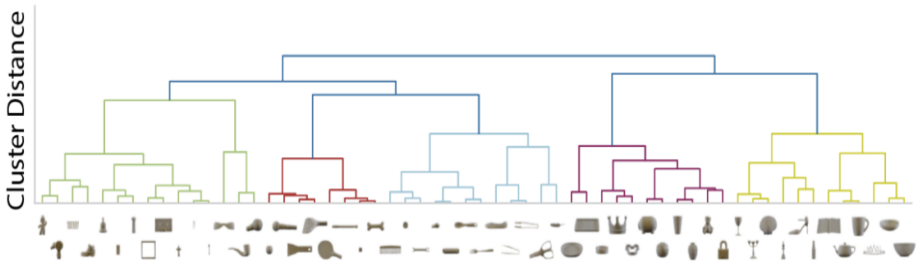

Figure 2.5: Dendrogram, the result of the HCA on grasping behavior. The dendrogram shows that objects are clustered based on their similarity in size and orientation of the graspable part as well as the shape of the grasp in this output space.

properties such as shape or size. For instance, we can see the early grouping of objects such as the crown and tiara, as well as the cup, mug, and wine glasses, which are all drink containers. In addition, shape similarities are also observable among objects of the same cluster.

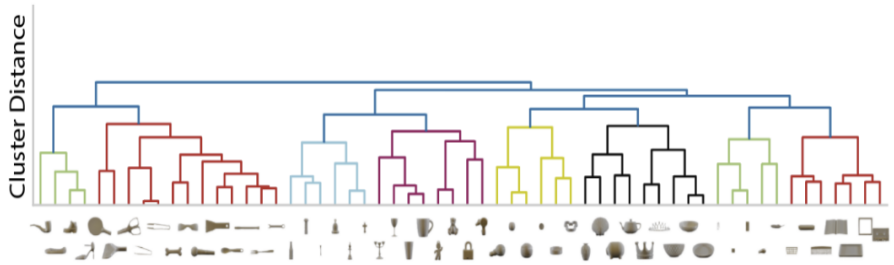

Figure 2.6: Dendrogram, the result of the HCA on object similarity judgment behavior. The dendrogram shows that objects are clustered based on their semantic and categorical as well as shape similarity in this behavioral output space.

After analyzing each behavior individually we want to investigate the extent to which the two behaviors overlap. In our comparison of the clustering dendrograms, size and orientation similarities seem to be the determining factors in the grasping output space. On the other hand, In the similarity judgment output space semantic and categorical similarities seemed to be the determining factors. Additionally, we noticed objects being clustered by shape similarities in both grasping and similarity judgments. These qualitative results suggest the differences in the strategies that participants use in performing each behavior.

To directly measure the relatedness of the behavioral spaces in a quantitative manner we used Representational Similarity Analysis (RSA). Figure 2.7 shows the results of this analysis. The correlation between the similarity judgments on 3D printed objects and the grasping behavior was 0.236. Although this value is significantly above zero, it is relatively low compared

to the noise ceiling of the data which is 0.78 and 0.84 for grasping and similarity judgment respectively. These results suggest that the object attributes that influenced participants' visually guided grasp are different from those affecting their similarity judgment. The low overlap between the two spaces raises the question of whether a single hierarchical visual process can support both tasks or not.

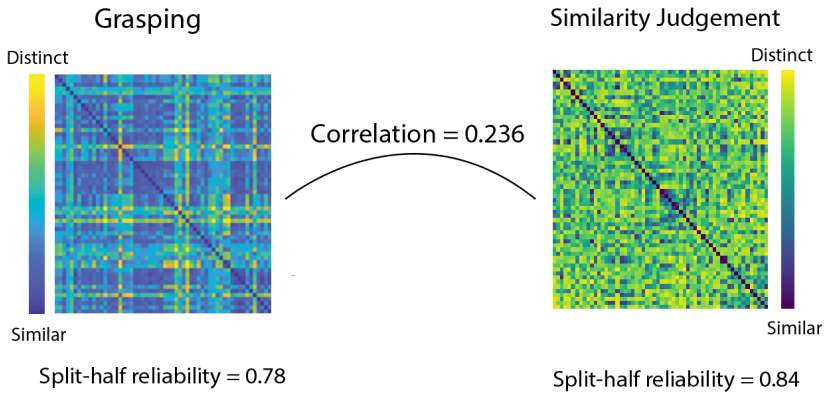

Figure 2.7: Comparison between the two behaviors on the 3D-printed objects

2.3.2 DCNN results

DCNNs have been shown to resemble the hierarchical process of the ventral stream since their lower layers show more correlation with earlier layers of the visual cortex and their higher layers are more correlated with high-level areas in the ventral visual stream. Besides, these models can obtain high performance in the object recognition task, making them a good candidate proxy of the ventral visual pathway which is suggested to be the neural substrate for object recognition. With this analogy, we can see to what extent can the brain-like representations of DCNNs trained on an object recognition task, predict these evidently distinct behaviors. This comparison can give us insight into whether the visual processing for these two behaviors needs to diverge in the brain or not.

Object Categorization: We first evaluated the performance of these models in categorizing images of 3D-printed objects. Since the models are trained on naturalistic images and the 3D-printed objects are different in color and texture from their naturalistic counterparts, we need to ensure the models can recognize these objects above the chance level before further analysis. Although we expect 3D-printed object categorization to be more challenging for the networks than recognizing objects in their natural color and texture; Particularly since previous studies have shown that these models tend to use the color and texture of objects in categorization tasks [30] and our 3d printed objects all have the same color and texture (white plastic). In addition, to evaluate the quality of the performed transfer learning from ImageNet categories to our chosen categories we tested the models on categorizing naturalistic

images of our objects. We calculated categorization performance for humans on both types of input (natural and 3D-printed objects). Figure 2.8 shows the categorization accuracy for the 3D-printed and natural images. All models successfully outperformed the baseline models on categorizing images of 3D-printed objects, however, all models except for Clip performed nearly 40 percent worse on 3D-printed objects compared to natural images. The performance of Clip was also lower in categorizing 3D printed objects compared to natural images, but this performance drop was smaller than the other models. This could be due to the unique training paradigm of Clip as it is trained contrastively and is exposed to more semantic input during training. Additionally, Clip is trained on a much larger dataset compared to the other models which can also result in differences in their performance [102]. All models performed at near human-level performance in natural image categorization. Human participants performed slightly worse when categorizing 3D-printed objects compared to natural objects. Overall these results suggest that 3D-printed objects are more difficult to categorize for humans and models. Also, removing the natural color and texture of objects has a more negative effect on supervised hierarchical neural networks trained on natural images.

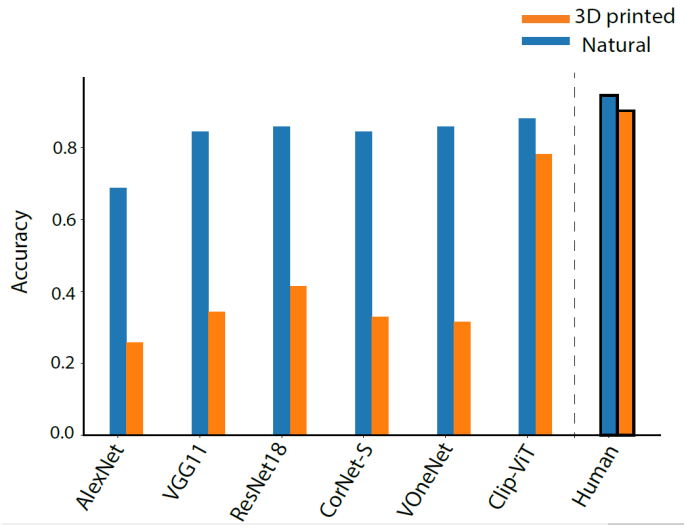

Figure 2.8: Categorization performance of DCNNs and Humans on images of naturalistic objects and 3-D printed objects

Regression Analysis results: Having established that the objects are recognizable to the models, we evaluated each neural network layer's ability to predict the behaviors. As mentioned both similarity judgment and grasping behaviors are the outputs of the visual system. Therefore, the representational spaces of the behaviors should be better predicted by later stages of the visual process that are closer to the output rather than the input. This is consistent with the results of previous studies that show that the IT cortex can explain similarity judgment and categorization behaviors better than its preceding visual regions. Therefore, by predicting each behavior from each layer of the models we can test if these models contain representations over their hierarchy that could support the two behaviors. As illustrated in Figure 2.9 (blue lines), the correlation between the layers of the networks and similarity judgment behavior increased as we moved from the first to the last layer. This increasing pattern of

correlations through the hierarchy along with the high peak correlation value in the later layers is consistent with what has been observed along the ventral stream in predicting similarity judgment from the brain responses [40, 18, 24, 94]. These results suggest parallels between the DCNN layers and the visual hierarchy underlying the similarity judgment behavior for the 3D-printed objects.

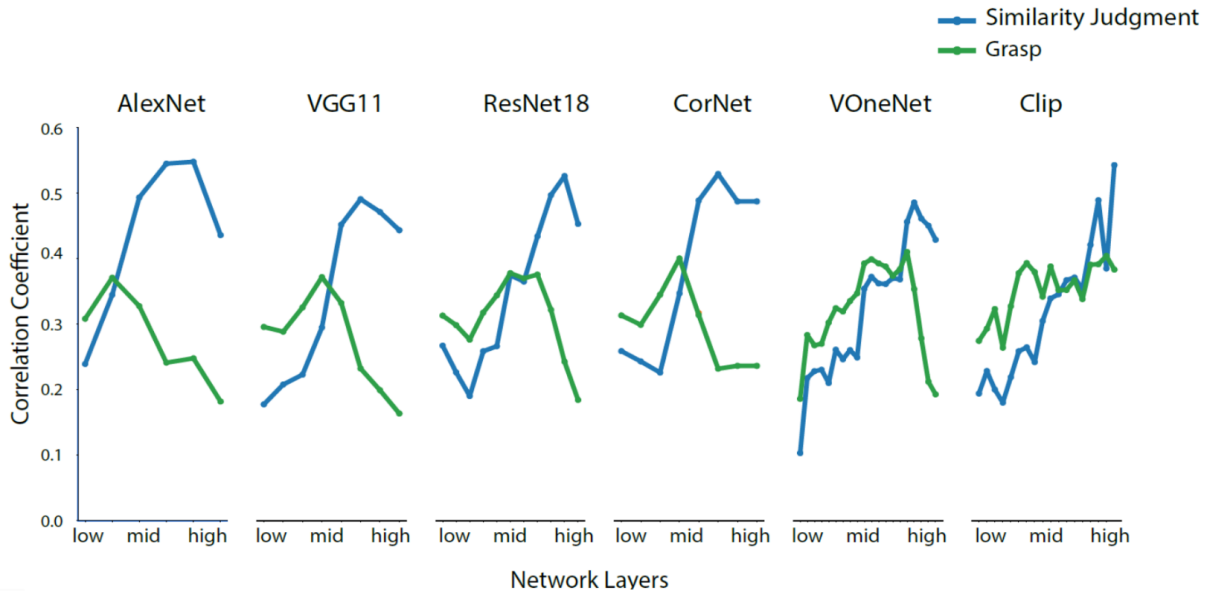

Figure 2.9: Results of the regression analysis on DCNNs and the behaviors. The *x* axis of each plot indicates the layers in the network hierarchy from the first to the last layer. Correlation values indicate the ability of each network layer at category similarity (blue) and grasping (green) behaviors.

We repeated the same analysis for the grasping behavior. As illustrated in figure 2.9 (green lines), the correlation between the layers of the network and the grasping behavior increased from the early to the middle layers. The highest correlation with grasping was observed in the middle layers. From middle to higher layers the correlation had a decreasing pattern, suggesting that the higher layers of the network fail to predict the arrangement of the objects in the grasping output space. The increasing pattern of prediction accuracy for the grasping behavior from the low to middle layers was similar to that for similarity judgments; Although the prediction accuracy for the similarity, judgments increased more rapidly than grasping. These results show that the early to middle layers contain features that could be useful in both grasping and similarity judgments. These features may be the same features that lead to some overlap between the two behaviors (indicated by the correlation of 0.24 between the two tasks). However, after progressing from the middle to later layers, the networks develop features that could be much more useful for the participants' similarity judgment than their grasping behavior.

Statistical Analysis: We used a permutation test to show that the correlation between the activation of the network and the similarity judgment behavior always peaks in later layers

compared to its correlation with grasping behavior and their difference is significant. To do so, we chose 100 different sets of held-out data for regression by shuffling the data before dividing it into train and test sets. Further, we calculated the correlation between two behaviors and the networks for each held-out data and obtained the index for the layer in which the correlation with the behavior peaked. The t-test result shows that for all networks the peak for correlation with similarity judgment occurred later than the peak correlation with the grasping behavior ($p < 0.05$). The result of this analysis shows that the CNN layer that best supports grasping behavior is always lower in the hierarchy than the layer that best supports similarity judgment behavior. In other words, these results emphasize that the middle layers are best at supporting grasping while the higher layers are better at supporting similarity judgment.

2.4 Discussion

In this thesis, we tackled the question of whether the internal representations of DCNNs can predict the two behaviors of object similarity judgment and grasping without explicit optimization for these tasks. We collected both behaviors on the same set of objects to investigate how these objects are differentiated from each other with respect to each behavior. Our behavioral results indicate a strong distinction between the two behaviors and suggest differences in the visual representations supporting each behavior. Further, we investigated if we could predict each behavior from the activity of DCNNs viewing the same objects. Based on our results, DCNN layers predicted similarity rating behavior increasingly better as the hierarchy progressed. However, the predictability of grasping behavior, which improved from the low to middle layers, rapidly declined after its peak at the middle layers. These results suggest that a single computational process is not enough to support these distinct behaviors.

Previous studies have investigated grasping and similarity judgment behaviors independently [54, 60, 3, 27]. This is the first attempt at studying these action-based and perception-based behaviors on the same set of everyday objects. Our unique stimulus set with a large set of 3D printed everyday objects, allows us to directly compare the representational spaces of the two behaviors. This is essential for understanding if the same underpinning visual processes can support both behaviors or not. Additionally, for the first time, this framework provides the opportunity to study the extent to which DCNN layers can predict the two behaviors. Such investigations combined with our knowledge of the human and artificial visual systems can help us uncover the computational requirements for producing the two behavioral outputs. They can also inspire new architectural designs for enhanced computer vision models that can generalize to tasks beyond object recognition.

Our behavioral results indicated that the participants naturally incorporate semantic and categorical features in judging which objects are similar, in the absence of any explicit instructions to focus on such features. The same objects are organized differently in the grasping behavioral space in which objects are organized mostly based on the size and orientation of the graspable part of the object. Although these properties have been shown to be incorporated in similarity judgment and grasping [REF], studies also mention shape as a property that is com-

monly important for both shaping our grasp and object similarity judgments [REF]. As such, the low correlation between grasp and similarity judgment might be surprising. We believe the reason for this low similarity is that our naturalistic objects have complex shapes and semantic meanings. In the presence of semantic and categorical information, shape properties may have a smaller effect on participants' similarity judgment behaviors. This reliance on categorical and semantic features probably allows the two behaviors to diverge more. This hypothesis could be tested with future experiments using artificial shapes that do not have semantic meanings.

In previous computer vision studies, DCNN model behavior has been extensively compared with human behavior in object recognition. A few studies have also compared their representational space with that obtained from similarity judgment. Despite the popularity of DCNNs in Robotic grasping, no studies have evaluated the extent to which the representations in these models can predict grasping behavior. Michaels et al, [84] suggested that the output from the last convolutional layer of a DCNN can be used to produce macaque grasping behavior. They also show that the same DCNN can also predict neural data in the macaque motor regions obtained during grasp movements. However, our results show that the higher layers of DNN fail to capture the relevant information for grasping. This distinction could be due to the fact that our objects are more complex than the simple shapes that were used in their model. Therefore the semantic meaning of our objects allows the high-level DNN activations to extract more semantic information and diverge from features that are relevant to grasp. Accordingly, we suggest that a promising architecture for future models could be a multi-task DCNN that is optimized for both action-based and perception-based tasks. Additionally, based on our results, this network can be split into two processing pathways from the middle layers to allow for better performance in the production of the two outputs.

Chapter 3

Discussion

Our visual system is capable of building a comprehensive representation of the 3D world that enables us to perform various tasks such as object recognition and similarity judgment. At the same time, it also enables us to interact with our surrounding objects in various ways and with a high level of precision. Numerous behavioral experiments suggest that visual perception and visually guided action are different in their nature as different object attributes might be considered for serving each objective [16, 131, 134, 3, 27]. These differences make supporting both objectives a complex computational challenge. Despite its complexity, our visual system seamlessly overcomes this challenge. Understanding how the human visual system achieves this feat can inspire building computational models that can perform a wide range of visual tasks beyond object recognition.

3.1 Object grasping and similarity judgments rely on distinct features

As a step toward this goal, in this thesis, I focus on object grasping and similarity judgment as examples of action-based and perception-based behaviors, respectively. To grasp an object in a stable manner, the visual system estimates different physical properties of the object to determine which part of the object should be grasped and how our hands should be shaped to match that object part. On the other hand, object similarity judgment is a complex perceptual task that requires a high-level and abstract understanding of the objects and their functionality [54, 60]. To gain more insight into these behaviors, experiments were conducted on the same set of stimuli, and the arrangement of objects in the representational space of each behavior was investigated using hierarchical clustering analysis. Results showed that features such as the size and orientation of the objects had a dominant effect on how participants shaped their hands for grasping them. This was consistent with previous neuroscience studies on grasping that have emphasized the direct effect of these features in the hand pre-shaping for grasp [133]. In the similarity judgment output space, objects with categorical and semantic similarity were grouped together, although no direct cues were given to the participants for considering semantic information in their judgments. This characteristic of similarity judgment behavior has been previously reported by Jozwik et al. [54], suggesting that the best predictors of object similarity judgments are category-based models rather than feature-based models. Moreover, RSA

comparison of the obtained behaviors showed that grasping and similarity judgment are mostly distinct with only a small overlap (correlation = 0.236). The low correlation of the behavioral spaces suggests that participants rely on different features for grasping compared to similarity judgment. In other words, the two behaviors impose distinct computational challenges on the visual system. Understanding and solving these challenges has been the driving motivation of numerous computer vision and robotics studies [105, 129, 23, 63].

3.2 DCNNs for object similarity judgment and grasping

DCNNs have shown remarkable performance at solving object recognition tasks. Additionally inspired by our knowledge of the ventral stream these networks are rapidly improving at explaining the neural activity of the ventral regions that are involved in object recognition 25, 26, 27. Besides, they have been shown to resemble the hierarchical processing along this visual pathway; Meaning that their earlier layers are better at predicting the early visual cortex, and later layers are the best predictor for higher levels of visual hierarchy in the IT cortex. Due to these abilities, DCNNs has become known as the best available models of the human visual system [70, 120] and therefore promising candidates for other visual tasks beyond object recognition such as robotic grasping. Despite their promising performance at improving robotic grasping [9, 86, 85, 97, 96], it is unclear whether these models can capture the representations required for this action-based behavior. Additionally, state-of-the-art robotic grasping models seem to fail either in the stable and accurate grasping of objects or generalization to grasping unseen and unknown objects [25, 62]. The significant behavioral distinctions between human and robotic grasping raise the question of whether the current DCNN architectures can extract relevant features for guiding object grasping behavior. Therefore, In this thesis, my goal was to investigate whether an action-based (grasping) and a perception-based (similarity judgment) behavior can be both predicted by the internal representations of these models without explicit training for either task.

Using a wide range of DCNNs from classic CNNs models (e.g. AlexNet, VGG11, ResNet-18), to brain-like models (CORnet-S, and VOneNet), and vision transformers (Clip-ViT), I evaluated the extent to which the internal representations of DCNNs explain these two behaviors. Results of the regression analysis in chapter 2 (Figure 2.9) showed that all models get increasingly better at explaining the similarity judgment behavior as the hierarchy progressed. In other words, as representations in the network became more abstract they get closer to the representational space of this perception-based behavior. Additionally, since these models were trained to do object categorization, the high-level representations in their hierarchy included information related to object categories. Therefore their pattern of correlation with similarity judgment is in line with previous evidence that humans naturally judge object similarities based on their categorical and semantic similarity, without any directions to do so [54, 60]. However, in cases where categorical similarities were not easily noticeable, participants may have relied on other characteristics of objects (e.g., size, shape) in their judgment. Meanwhile, these models showed increasingly better predictions of the grasping behavior from the early to middle layers. However, the similarity between layer representations and grasping rapidly decreased from the middle to high layers, after peaking in the middle. In most DCNNs, the

performance of the last layer in predicting grasp drastically declined such that it fell below the performance of the first layer, which extracted low-level features like edges. These results suggest that early on in the hierarchy DCNNs are capturing representations that are helpful for both similarity judgment and grasping. However from the middle layers onward the learned representations are specifically useful for similarity judgments and are most likely not needed for the visual process that determines participants' grasping configurations. More specifically, category-related features captured in the high layers of DCNNs fail to predict grasping. This is while the mid-level features captured by these models that are considered to be related to patterns and surface properties of objects [7] seem to explain grasping better than the features of other layers.

Overall this analysis suggests that a single hierarchical CNN is not able to learn the distinct features required for serving the two behaviors. And to explain these behaviors the computational architecture might need to be separately optimized for a different objective from the middle layers. These results are consistent with numerous neuroscientific findings that suggest visual information for grasping and object recognition is processed in separate visual streams: the dorsal pathway, and the ventral pathway, respectively. [34].

3.3 Interactions between dorsal and ventral pathways

Along with strong neuroscientific evidence that highlights the distinctions between the dorsal and ventral stream functionalities, a growing number of experimental evidence suggests a close interaction between the streams [33, 35]. For instance, behavioral studies on humans suggest that the physical properties of objects are not the only factors that determine grasping configurations and semantic attributes of the object retrieved from our perception of it are also influential [103]. Additionally based on the purpose of grasping (e.g., grasping a cup to drink from it or to put it in the dishwasher) the grasp pose can vary [101]. And to determine how to use an object we are first required to recognize it. In patients with ventral stream lesions grasping abilities remain mostly normal due to the intact dorsal stream processing. However, these patients struggle to grasp objects appropriately for different functions [129]. Moreover, patients with dorsal stream lesions elicit an opposite behavioral pattern as they can perform memory-guided grasping by processing the stored information that is shared with the ventral stream [12, 30]. Together these results suggest a tight integration between the two visual pathways. However, the extent of this interaction for different perception-based and action-based tasks is unclear.

3.4 Suggestions for future studies

The results of this thesis can be a stepping stone into a new line of investigations in both neuroscience and artificial intelligence. In general, visually guided action has been far less explored than visual perception in both fields. However, since vision is an active process, visual perception and interaction are equally influential in our general ability to function in the visual environment. Therefore a promising path toward understanding human vision and improving

computer vision might be studying perception and action in tandem. One possible future direction would be to use dual-pathway CNNs, where each pathway is optimized for one of these tasks. This architecture can be split into two pathways from various locations along the hierarchy and the resulting models can be compared on several levels: downstream performance on each task, explaining the representational space of human behavior, and neural correspondence with dorsal and ventral streams while performing the tasks. For each pair of tasks, this analysis can reveal if the two processing pathways are computationally required and to what extent the two streams should overlap. This approach can be applied to a variety of object properties. The need for two processing streams might differ across different properties as they have various levels of relatedness. These methods can be combined with neuroimaging techniques to achieve a broader understanding of the underlying computations of different behaviors in the visual system.

In sum, in this thesis, I demonstrated how behavioral experiments and computer modeling approaches can expand our understanding of certain behaviors and suggest better architectural designs for computer vision models. The results of this thesis shed light on the usefulness of using current object recognition DCNNs for active vision tasks. Moreover, they inform us about the possible architectural designs that would lead to enhanced robotic grasping and object perception models in future attempts. Such architectures can in turn be used as tools to characterize different kinds of action-based and perception-based behaviors and their extent of computational overlap. These investigations can result in a better understanding of the behaviors and how they might be represented along the visual streams in the brain. Therefore, they can inspire concrete and informed hypotheses about information processing in the visual pathways and their interaction, which can be evaluated in future neuroscience studies.

Bibliography

[1] Stefan Ainetter and Friedrich Fraundorfer. End-to-end trainable deep neural network for robotic grasp detection and semantic segmentation from RGB. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2021.

[2] Naveed Akhtar, Ajmal Mian, Navid Kardan, and Mubarak Shah. Advances in adversarial attacks and defenses in computer vision: A survey. *IEEE Access*, 9:155161–155196, 2021.

[3] Caterina Ansuini, Veronica Tognin, Luca Turella, and Umberto Castiello. Distractor objects affect fingers' angular distances but not fingers' shaping during grasping. *Experimental Brain Research*, 178(2):194–205, 2007.

[4] Nicholas Baker, Hongjing Lu, Gennady Erlikhman, and Philip J Kellman. Deep convolutional networks do not classify based on global object shape. *PLoS computational biology*, 14(12):e1006613, 2018.

[5] Shahab Bakhtiari, Patrick Mineault, Timothy Lillicrap, Christopher Pack, and Blake Richards. The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 25164–25178. Curran Associates, Inc., 2021.

[6] Valentina Emilia Balas, Raghvendra Kumar, and Rajshree Srivastava. *Recent trends and advances in artificial intelligence and internet of things*. Springer, 2020.

[7] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[8] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. *Advances in neural information processing systems*, 19, 2006.

[9] A. Bicchi and V. Kumar. Robotic grasping and contact: a review. In *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No.00CH37065)*. IEEE.

[10] Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115, 1987.

[11] Stefania Bracci, J Brendan Ritchie, Ioannis Kalfas, and Hans P Op de Beeck. The ventral visual pathway represents animal appearance over animacy, unlike human behavior and deep neural networks. *Journal of Neuroscience*, 39(33):6513–6525, 2019.

[12] Andreas Buja, Deborah F Swayne, Michael L Littman, Nathaniel Dean, Heike Hofmann, and Lisha Chen. Data visualization with multidimensional scaling. *Journal of Computational and Graphical Statistics*, 17(2):444–472, June 2008.

[13] Laurel J. Buxbaum and H. Branch Coslett. Subtypes of optic ataxia: Reframing the disconnection account. *Neurocase*, 3(3):159–166, May 1997.

[14] Santiago A Cadena, George H Denfield, Edgar Y Walker, Leon A Gatys, Andreas S Tolias, Matthias Bethge, and Alexander S Ecker. Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS computational biology*, 15(4):e1006897, 2019.

[15] Sydney Cash and Rafael Yuste. Linear summation of excitatory inputs by ca1 pyramidal neurons. *Neuron*, 22(2):383–394, 1999.

[16] ERIS CHINELLATO and ANGEL P. DEL POBIL. THE NEUROSCIENCE OF VISION-BASED GRASPING: A FUNCTIONAL REVIEW FOR COMPUTATIONAL MODELING AND BIO-INSPIRED ROBOTICS. *Journal of Integrative Neuroscience*, 08(02):223–254, June 2009.

[17] Joel Dapello, Tiago Marques, Martin Schrimpf, Franziska Geiger, David Cox, and James J DiCarlo. Simulating a primary visual cortex at the front of cnns improves robustness to image perturbations. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13073–13087. Curran Associates, Inc., 2020.

[18] H. P. Op de Beeck, K. Torfs, and J. Wagemans. Perceived shape similarity among unfamiliar objects and the organization of the human object vision pathway. *Journal of Neuroscience*, 28(40):10111–10123, October 2008.

[19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database, in 'cvpr09'. *IEEE Computer Society, Miami, Florida, USA*, pages 248–255, 2009.

[20] Amaury Depierre, Emmanuel Dellandrea, and Liming Chen. Jacquard: A large scale dataset for robotic grasp detection. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, October 2018.

[21] James J DiCarlo and David D Cox. Untangling invariant object recognition. *Trends in cognitive sciences*, 11(8):333–341, 2007.

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

[23] Zhikai Dong, Sicheng Liu, Tao Zhou, Hui Cheng, Long Zeng, Xingyao Yu, and Houde Liu. PPR-net:point-wise pose regression network for instance segmentation and 6d pose estimation in bin-picking scenarios. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, November 2019.

[24] D. M. Drucker and G. K. Aguirre. Different spatial scales of shape similarity representation in lateral and ventral LOC. *Cerebral Cortex*, 19(10):2269–2280, January 2009.

[25] Guoguang Du, Kai Wang, Shiguo Lian, and Kaiyong Zhao. Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: a review. *Artificial Intelligence Review*, 54(3):1677–1734, August 2020.

[26] Kunihiko Fukushima. Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural networks*, 1(2):119–130, 1988.

[27] Tzvi Ganel and Melvyn A Goodale. Visual control of action but not perception requires analytical processing of object shape. *Nature*, 426(6967):664–667, 2003.

[28] R Garcin, P Rondot, and J De Recondo. Optic ataxia localized in 2 left homonymous visual hemifields (clinical study with film presentation). *Revue neurologique*, 116(6):707–714, 1967.

[29] Robert Geirhos, Kristof Meding, and Felix A. Wichmann. Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13890–13902. Curran Associates, Inc., 2020.

[30] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness, 2018.

[31] Scott Glover. Optic ataxia as a deficit specific to the on-line control of actions. *Neuroscience &amp Biobehavioral Reviews*, 27(5):447–456, August 2003.

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

[33] Melvyn A Goodale and Angela M Haffenden. Interactions between the dorsal and ventral streams of visual processing. *Advances in neurology*, 93:249—267, 2003.

[34] Melvyn A Goodale and A David Milner. Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1):20–25, 1992.

[35] Melvyn A Goodale and David A Westwood. An evolving view of duplex vision: separate but interacting cortical pathways for perception and action. *Current Opinion in Neurobiology*, 14(2):203–211, April 2004.

[36] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[37] Charles G Gross. Representation of visual stimuli in inferior temporal cortex. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 335(1273):3–10, 1992.

[38] Umut Güçlü and Marcel AJ van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.

[39] Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245–258, 2017.

[40] Johannes Haushofer, Margaret S Livingstone, and Nancy Kanwisher. Multivariate patterns in object-selective cortex dissociate perceptual and physical shape similarity. *PLoS Biology*, 6(7):e187, July 2008.

[41] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[42] Martin N. Hebart, Adam H. Dickter, Alexis Kidder, Wan Y. Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I. Baker. THINGS: A database of 1, 854 object concepts and more than 26, 000 naturalistic object images. *PLOS ONE*, 14(10):e0223792, October 2019.

[43] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[44] Ha Hong, Daniel L K Yamins, Najib J Majaj, and James J DiCarlo. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature Neuroscience*, 19(4):613–622, February 2016.

[45] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[46] David H Hubel and Torsten N Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148(3):574, 1959.

[47] Juhani Hyvärinen and Antti Poranen. Function of the parietal associative area 7 as revealed from cellular discharges in alert monkeys. *Brain*, 97(4):673–692, 1974.

[48] Thomas W James, Jody Culham, G Keith Humphrey, A David Milner, and Melvyn A Goodale. Ventral occipital lesions impair object recognition but not object-directed grasping: an fmri study. *Brain*, 126(11):2463–2475, 2003.

[49] Marc Jeannerod. Intersegmental coodination during reaching at natural visual objects. *Attention and performance*, pages 153–169, 1981.

[50] Marc Jeannerod. The timing of natural prehension movements. *Journal of motor behavior*, 16(3):235–254, 1984.

[51] Marc Jeannerod. Mechanisms of visuomotor coordination: a study in normal and brain-damaged subjects. *Methods in Neuropsychology*, pages 41–78, 1986.

[52] Yun Jiang, Stephen Moseson, and Ashutosh Saxena. Efficient grasping from RGBD images: Learning using a new rectangle representation. In *2011 IEEE International Conference on Robotics and Automation*. IEEE, May 2011.

[53] Kamila M Jozwik, Nikolaus Kriegeskorte, Katherine R Storrs, and Marieke Mur. Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Frontiers in psychology*, 8:1726, 2017.

[54] Kamila M. Jozwik, Nikolaus Kriegeskorte, Katherine R. Storrs, and Marieke Mur. Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Frontiers in Psychology*, 8, October 2017.

[55] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11):e1003915, 2014.

[56] Saeed Reza Kheradpisheh, Masoud Ghodrati, Mohammad Ganjtabesh, and Timothée Masquelier. Deep networks can resemble human feed-forward vision in invariant object recognition. *Scientific reports*, 6(1):1–24, 2016.

[57] Been Kim, Emily Reif, Martin Wattenberg, Samy Bengio, and Michael C Mozer. Neural networks trained on natural scenes exhibit gestalt closure. *Computational Brain & Behavior*, 4(3):251–263, 2021.

[58] HyunGoo R Kim, Dora E Angelaki, and Gregory C DeAngelis. The neural basis of depth perception from motion parallax. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1697):20150256, 2016.

[59] Marcie L King, Iris IA Groen, Adam Steel, Dwight J Kravitz, and Chris I Baker. Similarity judgments and cortical visual responses reflect different properties of object and scene categories in naturalistic images. *NeuroImage*, 197:368–382, 2019.

[60] Marcie L. King, Iris I.A. Groen, Adam Steel, Dwight J. Kravitz, and Chris I. Baker. Similarity judgments and cortical visual responses reflect different properties of object and scene categories in naturalistic images. *NeuroImage*, 197:368–382, August 2019.

[61] Kilian Kleeberger, Richard Bormann, Werner Kraus, and Marco F. Huber. A survey on learning-based robotic grasping. *Current Robotics Reports*, 1(4):239–249, September 2020.

[62] Kilian Kleeberger, Richard Bormann, Werner Kraus, and Marco F. Huber. A survey on learning-based robotic grasping. *Current Robotics Reports*, 1(4):239–249, September 2020.

[63] Kilian Kleeberger and Marco F. Huber. Single shot 6d object pose estimation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2020.

[64] Lina K Klein, Guido Maiello, Vivian C Paulun, and Roland W Fleming. Predicting precision grip grasp locations on three-dimensional objects. *PLoS computational biology*, 16(8):e1008081, 2020.

[65] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529. PMLR, 2019.

[66] Dwight J Kravitz, Kadharbatcha S Saleem, Chris I Baker, Leslie G Ungerleider, and Mortimer Mishkin. The ventral visual pathway: an expanded neural framework for the processing of object quality. *Trends in cognitive sciences*, 17(1):26–49, 2013.

[67] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, page 4, 2008.

[68] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, May 2017.

[69] Jonas Kubilius, Stefania Bracci, and Hans P Op de Beeck. Deep neural networks as a computational model for human shape sensitivity. *PLoS computational biology*, 12(4):e1004896, 2016.

[70] Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Rishi Rajalingham, Ha Hong, Najib Majaj, Elias Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, Aran Nayebi, Daniel Bear, Daniel L Yamins, and James J DiCarlo. Brain-like object recognition with high-performing shallow recurrent anns. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[71] Brenden M Lake, Wojciech Zaremba, Rob Fergus, and Todd M Gureckis. Deep neural networks predict category typicality ratings for images. In *CogSci*, 2015.

[72] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[73] Claude Lemaréchal. Cauchy and the gradient method. *Doc Math Extra*, 251(254):10, 2012.

[74] Ian Lenz, Honglak Lee, and Ashutosh Saxena. Deep learning for detecting robotic grasps. *The International Journal of Robotics Research*, 34(4-5):705–724, March 2015.

[75] Ian Lenz, Honglak Lee, and Ashutosh Saxena. Deep learning for detecting robotic grasps. *The International Journal of Robotics Research*, 34(4-5):705–724, 2015.

[76] Beatriz León, Stefan Ulbrich, Rosen Diankov, Gustavo Puche, Markus Przybylski, Antonio Morales, Tamim Asfour, Sami Moisio, Jeannette Bohg, James Kuffner, and Rüdiger Dillmann. OpenGRASP: A toolkit for robot grasping simulation. In *Simulation, Modeling, and Programming for Autonomous Robots*, pages 109–120. Springer Berlin Heidelberg, 2010.

[77] Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International Journal of Robotics Research*, 37(4-5):421–436, June 2017.

[78] Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International journal of robotics research*, 37(4-5):421–436, 2018.

[79] Nikos K Logothetis and David L Sheinberg. Visual object recognition. *Annual review of neuroscience*, 19(1):577–621, 1996.

[80] Jeffrey Mahler, Jacky Liang, Sherdil Niyaz, Michael Laskey, Richard Doan, Xinyu Liu, Juan Aparicio Ojea, and Ken Goldberg. Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics, 2017.

[81] Jeffrey Mahler, Florian T. Pokorny, Brian Hou, Melrose Roderick, Michael Laskey, Mathieu Aubry, Kai Kohlhoff, Torsten Kroger, James Kuffner, and Ken Goldberg. Dex-net 1.0: A cloud-based network of 3d objects for robust grasp planning using a multi-armed bandit model with correlated rewards. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2016.

[82] David Marr. *Vision*. The MIT Press, 2010.

[83] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[84] Jonathan A. Michaels, Stefan Schaffelhofer, Andres Agudelo-Toro, and Hansjörg Scherberger. A goal-driven modular neural network predicts parietofrontal neural dynamics during grasping. *Proceedings of the National Academy of Sciences*, 117(50):32124–32135, November 2020.

[85] A.T. Miller and P.K. Allen. GraspIt! *IEEE Robotics &amp Automation Magazine*, 11(4):110–122, December 2004.

[86] A.T. Miller, S. Knoop, H.I. Christensen, and P.K. Allen. Automatic grasp planning using shape primitives. In *2003 IEEE International Conference on Robotics and Automation (Cat. No.03CH37422)*. IEEE.

[87] A David Milner, David I Perrett, Rhona S Johnston, Philip J Benson, Timothy R Jordan, David W Heeley, Diego Bettucci, Franco Mortara, Roberto Mutani, Emanuela Terazzi, et al. Perception and action in 'visual form agnosia'. *Brain*, 114(1):405–428, 1991.

[88] David Milner and Mel Goodale. *The Visual Brain in Action*. Oxford University Press, October 2006.

[89] Mortimer Mishkin, Leslie G Ungerleider, and Kathleen A Macko. Object vision and spatial vision: two cortical pathways. *Trends in neurosciences*, 6:414–417, 1983.

[90] Douglas Morrison, Peter Corke, and Jürgen Leitner. Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach, 2018.

[91] Douglas Morrison, Peter Corke, and Jürgen Leitner. Learning robust, real-time, reactive robotic grasping. *The International Journal of Robotics Research*, 39(2-3):183–201, June 2019.

[92] Vernon B Mountcastle, James C Lynch, Apostolos Georgopoulos, Hideaki Sakata, and C Acuna. Posterior parietal association cortex of the monkey: command functions for operations within extrapersonal space. *Journal of neurophysiology*, 38(4):871–908, 1975.

[93] Arsalan Mousavian, Clemens Eppner, and Dieter Fox. 6-dof graspnet: Variational grasp generation for object manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[94] Marieke Mur, Mirjam Meys, Jerzy Bodurka, Rainer Goebel, Peter A. Bandettini, and Nikolaus Kriegeskorte. Human object-similarity judgments reflect and transcend the primate-IT object representation. *Frontiers in Psychology*, 4, 2013.

[95] Frank Nielsen. Hierarchical clustering. In *Introduction to HPC with MPI for Data Science*, pages 195–211. Springer International Publishing, 2016.

[96] R. Pelossof, A. Miller, P. Allen, and T. Jebara. An SVM learning approach to robotic grasping. In *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA '04. 2004*. IEEE, 2004.

[97] Raphael Pelossof, Andrew Miller, Peter Allen, and Tony Jebara. An svm learning approach to robotic grasping. In *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004*, volume 4, pages 3512–3518. IEEE, 2004.

[98] M.-T. PERENIN and A. VIGHETTO. OPTIC ATAXIA: A SPECIFIC DISRUPTION IN VISUOMOTOR MECHANISMS. *Brain*, 111(3):643–674, 1988.

[99] M.-T. PERENIN and A. VIGHETTO. OPTIC ATAXIA: A SPECIFIC DISRUPTION IN VISUOMOTOR MECHANISMS. *Brain*, 111(3):643–674, 1988.

[100] Nicolas Pinto, N Majaj, Youssef Barhomi, E Solomon, and JJ DiCarlo. Human versus machine: comparing visual object recognition systems on a level playing field. *Cosyne Abstracts*, 2010.

[101] Mary C Potter. Short-term conceptual memory for pictures. *Journal of experimental psychology: human learning and memory*, 2(5):509, 1976.

[102] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.

[103] Rishi Rajalingham, Elias B Issa, Pouya Bashivan, Kohitij Kar, Kailyn Schmidt, and James J DiCarlo. Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, 38(33):7255–7269, 2018.

[104] D Rao, Q V Le, T Phoka, M Quigley, A Sudsang, and A Y Ng. Grasping novel objects with depth segmentation. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, October 2010.

[105] Joseph Redmon and Anelia Angelova. Real-time grasp detection using convolutional neural networks. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2015.

[106] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[107] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[108] Nichola J Rice, Robert D McIntosh, Igor Schindler, Mark Mon-Williams, Jean-Francois Demonet, and A David Milner. Intact automatic avoidance of obstacles in patients with visual form agnosia. *Experimental Brain Research*, 174(1):176–188, 2006.

[109] Maximilian Riesenhuber and Tomaso Poggio. Are cortical models really bound by the "binding problem"? *Neuron*, 24(1):87–93, 1999.

[110] F Rosenblatt. The perceptron, a perceiving and recognizing automaton:(project para). buffalo (ny): Cornell aeronautical laboratory. 1957.

[111] Amir Rosenfeld, Markus D Solbach, and John K Tsotsos. Totally looks like-how humans compare, compared to machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1961–1964, 2018.

[112] Yves Rossetti and Laure Pisella. Optic ataxia: beyond the dorsal stream cliché. In *Handbook of Clinical Neurology*, pages 225–247. Elsevier, 2018.

[113] Guillaume A Rousselet, Michèle Fabre-Thorpe, and Simon J Thorpe. Parallel processing in high-level categorization of natural images. *Nature neuroscience*, 5(7):629–630, 2002.

[114] Reuven Y. Rubinstein and Dirk P. Kroese. *The Cross-Entropy Method*. Springer New York, 2004.

[115] Ingrid Russell. The delta rule—university of hartford, 2014.

[116] Artur Saudabayev, Zhanibek Rysbek, Raykhan Khassenova, and Huseyin Atakan Varol. Human grasping database for activities of daily living with depth, color and kinematic data streams. *Scientific data*, 5(1):1–13, 2018.

[117] Ashutosh Saxena, Justin Driemeyer, and Andrew Y. Ng. Robotic grasping of novel objects using vision. *The International Journal of Robotics Research*, 27(2):157–173, February 2008.

[118] Igor Schindler, Nichola J Rice, Robert D McIntosh, Yves Rossetti, Alain Vighetto, and A David Milner. Automatic avoidance of obstacles is a dorsal stream function: evidence from optic ataxia. *Nature Neuroscience*, 7(7):779–784, June 2004.

[119] H. Steven Scholte, Max M. Losch, Kandan Ramakrishnan, Edward H.F. de Haan, and Sander M. Bohte. Visual pathways from the perspective of cost functions and multi-task deep neural networks. *Cortex*, 98:249–261, January 2018.

[120] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-score: Which artificial neural network for object recognition is most brain-like? September 2018.

[121] Katja Seeliger, Matthias Fritsche, Umut Güçlü, Sanne Schoenmakers, J-M Schoffelen, Sander E Bosch, and MAJ Van Gerven. Convolutional neural network-based encoding and decoding of visual object recognition in space and time. *NeuroImage*, 180:253–266, 2018.

[122] Roger N. Shepard. The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika*, 27(3):219–246, September 1962.

[123] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[124] Michael T Swanston and Walter C Gogel. Perceived size and motion in depth from optical expansion. *Perception & psychophysics*, 39(5):309–326, 1986.

[125] Christian Szegedy, Alexander Toshev, and Dumitru Erhan. Deep neural networks for object detection. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[126] Keiji Tanaka. Neuronal mechanisms of object recognition. *Science*, 262(5134):685–688, 1993.

[127] Jessica AF Thompson, Yoshua Bengio, Elia Formisano, and Marc Schönwiesner. How can deep learning advance computational modeling of sensory information processing? *arXiv preprint arXiv:1810.08651*, 2018.

[128] Simon Thorpe, Denis Fize, and Catherine Marlot. Speed of processing in the human visual system. *nature*, 381(6582):520–522, 1996.

[129] Jonathan Tremblay, Thang To, Balakumar Sundaralingam, Yu Xiang, Dieter Fox, and Stan Birchfield. Deep object pose estimation for semantic robotic grasping of household objects. In Aude Billard, Anca Dragan, Jan Peters, and Jun Morimoto, editors, *Proceedings of The 2nd Conference on Robot Learning*, volume 87 of *Proceedings of Machine Learning Research*, pages 306–316. PMLR, 29–31 Oct 2018.

[130] Ragav Venkatesan and Baoxin Li. *Convolutional neural networks in visual computing: a concise guide*. CRC Press, 2017.

[131] David A. Westwood and Melvyn A. Goodale. A haptic size-contrast illusion affects size perception but not grasping. *Experimental Brain Research*, 153(2):253–259, November 2003.

[132] Allan M Wing, Ailie Turton, and Carole Fraser. Grasp size and accuracy of approach in reaching. *Journal of motor behavior*, 18(3):245–260, 1986.

[133] Sara A Winges, Douglas J Weber, and Marco Santello. The role of vision on hand preshaping during reach to grasp. *Experimental Brain Research*, 152(4):489–498, 2003.

[134] A. Winkler, C. E. Wright, and C. Chubb. Dissociating the functions of visual pathways using equisalient stimuli. *Journal of Vision*, 5(8):362–362, March 2010.

[135] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

Curriculum Vitae

Name:	Aidasadat Mirebrahimi Tafreshi
Post-Secondary Education and Degrees:	Shariaty College, Technical and Vocational University (TVU) Tehran, Iran 2017 - 2021 B.Sc. Computer Software Engineering University of Western Ontario London, ON 2021 - 2023 M.Sc Computer Science (Artificial Intelligence)
Awards:	1^{st} place in Cognitive Neuroscience Competition for university Students (CNCS), Iran 2019 Ranked top 2% in Iran's national university entrance for B.Sc 2016
Related Work Experience:	Graduate Teaching Assistant Western University 2021 - 2022 Teaching Assistant Neuromatch Academy 2022

Publications:

Zoroufi, A., Mirebrahimi, A., Ungerleider, L., Baker, C. and Vaziri-Pashkam, M., 2022. Predicting Multiple behaviors from the activity of Deep Neural Networks: Is one visual hierarchy enough?. Journal of Vision, 22(14), pp.3530-3530.

Zoroufi, A., Mirebrahimi, A. and Ghafari, T., 2021. The Effect of Exogenous Temporal Attention on the Gradient of Spatial Attention. Journal of Vision, 21(9), pp.2730-2730.