Chapter 8: Object categorization

High-level perception transforms the continuous world of sensory data into discrete categories like people, animals, and vehicles. In the visual system, evidence suggests that this transformation happens along the "ventral stream" extending from primary visual cortex to inferotemporal cortex. At the end of this transformation, object categories can be read out from neural population activity by a linear decoder. The most quantitatively successful models of this transformation are multi-layered neural networks trained for object categorization. Nonetheless, there remain gaps between these models and human perception, suggesting that the human brain has access to flexible generative models of sensory data.

The last chapter focused on two-alternative decisions, but many naturalistic decision problems involve far more alternatives. In particular, a central function of high-level perception is categorizing objects present in sensory data, where the number of object categories is in the thousands. Object categorization is, however, more than just a multi-alternative generalization of the decision problems studied in the last chapter, because it places greater demands on representation: category information is "entangled" at the level of low-level sensory cortex (e.g., V1; see Figure 1), in the sense that it cannot be read out with a linear decoder of the sort that we used to model the evidence accumulator in LIP (Pinto et al., 2008).

What we are calling "object categorization" is often called "object classification" or "object recognition."

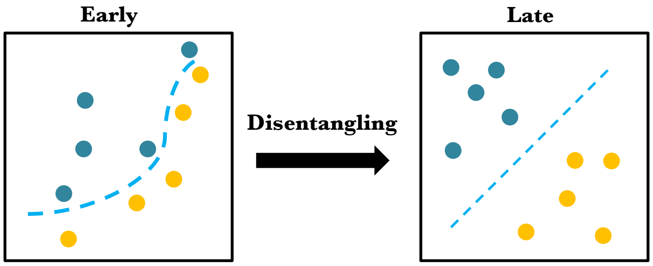


Figure 1: **Disentangling**. Each circle represents an exemplar (e.g., an image), color represents the category label, and the axes represent the activity levels of different neurons (here just two are shown for simplicity). Linear separability means that a line (or, more generally, a hyperplane in higher dimensions) can be constructed that perfectly separates the exemplars from each category. Early sensory representations are not linearly separable, but late representations (in IT cortex) are.

Specifically, a linear decoder for categorization takes the following form:

$$p(s|x) = f\left(\sum_d w_{ds} x_d\right), \tag{1}$$

where s denotes the object category (the state in this context), x_d is the firing rate of neuron d, w_{ds} the weight connecting neuron d to output neuron s, and $f(\cdot)$ is an output nonlinearity (e.g., softmax) that maps the outputs to probabilities. This means that early sensory representations need to be nonlinearly transformed such that category information is "disentangled"—i.e., linearly decodable. This chapter will focus on how this happens in the visual system, where it has been most extensively studied.

Even though $f(\cdot)$ is nonlinear, the decoder is still considered "linear" because the separating boundary between categories is linear, as illustrated in Figure 1.

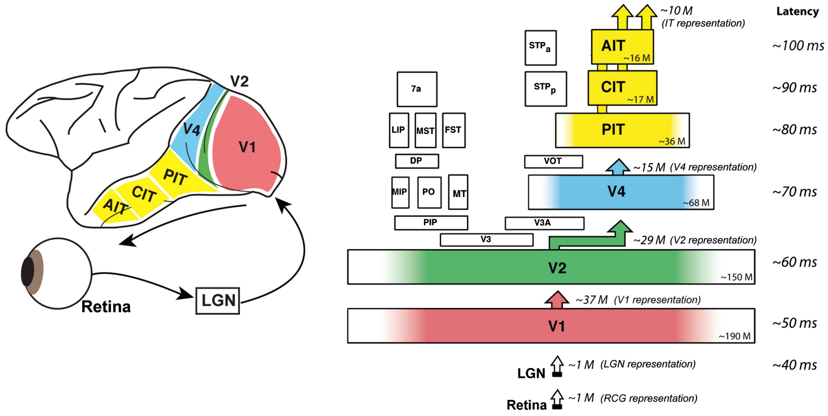


Figure 2: **The ventral visual stream**. (Left) Anatomical organization in the primate brain. (Right) Each area's size is proportional to its cortical surface area, with the approximate number of neurons shown in the corner of each area. The approximate dimensionality of each representation (number of projection neurons) is shown above each area. Approximate median response latency is shown to the right of each area. Reproduced from DiCarlo et al. (2012).

1 *The ventral visual stream*

It is widely believed that disentangling is achieved by a sequence of representational transformations along the ventral visual stream (Figure 2), extending from V1 to anterior inferotemporal (IT) cortex (DiCarlo et al., 2012). Several changes are apparent across the ventral stream. First, representations are initially retinotopically organized—neurons are tuned to specific retinal locations—and this retinotopy is eventually lost by the time visual information arrives in central/anterior IT. The loss of retinotopy is linked to the increase in receptive field size across the ventral visual stream; fields are sufficiently large in IT that retinotopy is no longer meaningful.

Second, stimulus selectivity changes qualitatively along the ventral stream, from orientation in V1, texture in V2, curvature in V4, and finally more semantically abstract categories in IT. These changes in selectivity are accompanied by increasing tolerance to low-level variations in scene parameters, such as lighting, size, position, and viewpoint (Zoccolan et al., 2007; Rust and DiCarlo, 2010). An example is shown in Figure 3.

Third, object categories can be linearly decoded from population activity in IT (but not areas earlier in the ventral stream), achieving high performance with only a few hundred neurons (Hung et al., 2005), as shown in Figure 4. Linear decoders of IT activity also quantitatively predict stimulus-specific error patterns exhibited by humans

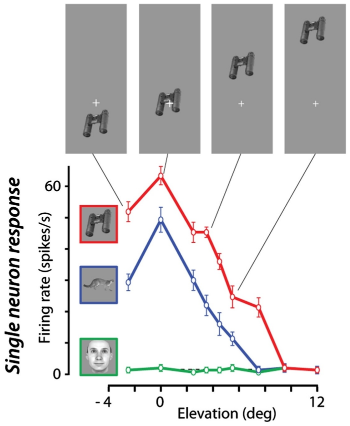


Figure 3: **Tolerance to position variation in an IT neuron**. Reproduced from DiCarlo et al. (2012), based on data from Zoccolan et al. (2007).

(Majaj et al., 2015).

Note that linearity of the decoder is only one way to constrain linking hypotheses between brain activity and computation. If the decoder is arbitrary, then one could in principle decode anything from the earliest sensory areas, rendering any claims about information coding in specific areas vacuous. Nonetheless, nonlinear decoding may occur in the brain (e.g., Pagan et al., 2016; Yang et al., 2021), so we should be cautious about relying too strongly on linearity.

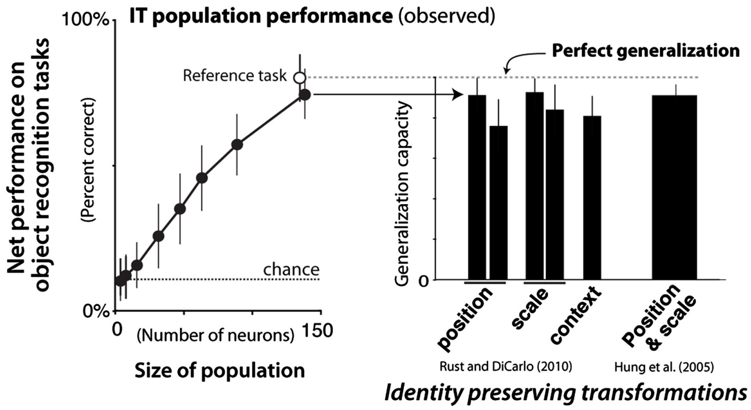


Figure 4: **Classification performance from IT population activity**. The right panel shows that performance generalizes across small position, scale, and context (background) variations. Reproduced from DiCarlo et al. (2012).

Causal evidence that IT is important for object categorization comes from stimulation, lesion, and inactivation studies. For example, stimulating (via localized electrical current) face-selective neurons in IT biases categorization judgments towards faces (Afraz et al., 2006), and alters change detection not only for faces but also for face-like stimuli (Moeller et al., 2017). Lesions and inactivations of IT produce selective impairments in object categorization (e.g., Weiskrantz and Saunders, 1984; Rajalingham and DiCarlo, 2019).

2 *Modeling the ventral stream with deep neural networks*

Deep convolutional neural networks (DCNNs; Figure 5) currently provide the most successful quantitative account of how the ventral stream achieves disentangling of object category information (see Kar and DiCarlo, 2024, for a review), though they are not without problems, as discussed later. Many variations of these networks have been studied, but here we will focus on some canonical motifs.

A DCNN consists of multiple layers, where each layer consists of multiple "units" (roughly corresponding to neurons or populations of neurons) that send outputs to the next layer. Units are perceptron-like (see Chapter 2), taking a linear combination of inputs and then passing them through a non-linearity—thus implementing the kind of linear decoder formalized in Eq. 1. In convolutional layers, all

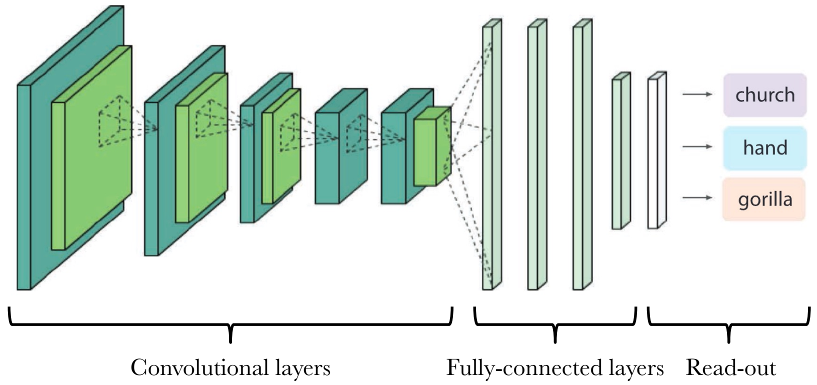


Figure 5: **Deep convolutional neural network for object categorization**. Adapted from Bracci and Op de Beeck (2023).

units share the same set of synaptic weights but apply these weights to different inputs (corresponding to subregions of images, in the case of visual processing). The shared weights thus function as learnable "filters" that are convolved with the input—i.e., applied uniformly to every part of the input. This is typically followed by some form of pooling (e.g., taking the max or mean over a subpopulation of units with similar receptive fields) and normalization (to bring the outputs into a standard range). Filtering and pooling are typically strided: the spatial resolution is reduced at each convolutional layer by retaining only a subset of outputs within each pool. Late layers (roughly corresponding to parts of IT) are fully connected and lack convolutional structure, since the spatial structure of receptive fields is diminished by repeated striding.

Training a DCNN involves adjusting the weights to optimize an objective function, which measures how well the network classifies training images. The most effective learning algorithms use some form of stochastic gradient descent, adjusting weights to follow the gradient of the objective function evaluated on small batches of labeled images. In the next chapter, we will discuss how such an algorithm might be plausibly implemented in a neural circuit.

2.1 *Behavioral and neural performance*

Despite decades of effort, no object categorization model was able to match human performance until DCNNs broke through the barrier with AlexNet (Krizhevsky et al., 2012). In fact, very similar DCNNs had already been around for several decades at that point (e.g.,. LeCun et al., 1989); other factors played a decisive role in the breakthrough. First, the ImageNet database provided a much larger and more diverse training set compared to previous ones. Second, computing power had increased dramatically since the advent of DCNNs, particularly through the use of graphics processing units, which enabled efficient parallelization of linear algebra.

In addition to matching overall human accuracy, DCNNs trained on object categorization can also match several more fine-grained aspects of human vision. As noted earlier, linear decoders of IT were able to match stimulus-specific confusions (i.e., misclassifications); the same is true for DCNNs (Rajalingham et al., 2015). Going beyond object categorization, Jacob et al. (2021) found that trained (and even sometimes untrained) DCNNs could qualitatively reproduce a range of phenomena in visual perception. For example, humans (as well as a number of other species) have greater difficulty discriminating mirror reflections of an image long the horizontal axis compared to reflections along the vertical axis (e.g., Sekuler and Houlihan, 1968, Figure 6). DCNNs likewise exhibited greater similarity in late (putatively IT) activity patterns for horizontal reflections compared to vertical reflections. This finding is particularly intriguing because measurements of IT show the same effect (Rollenhagen and Olson, 2000).

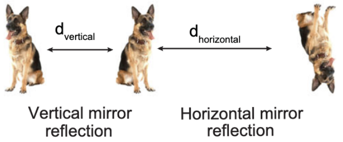


Figure 6: **Mirror reflections**. Neural representations are compared using Euclidean distance between activity vectors. Reproduced from Jacob et al. (2021).

Jacob et al. also identified a number of differences between human vision and DCNNs, which we will discuss below.

Yamins et al. (2014) undertook a more quantitative analysis of the match between DCNN internal representations and neural activity. They fit a linear mapping from the final layer of the DCNN to IT population activity, and then evaluated this mapping on held-out data. They found that DCNNs could achieve far better neural predictivity than any previous model, and that predictivity improved with categorization accuracy. Moreover, earlier layers provided good predictivity for upstream regions in the ventral stream (V1 and V4). Thus, to a first approximation, DCNNs appeared to recapitulate the key transformational steps in the ventral stream.

More recent work (Linsley et al., 2023) showed that the relationship between IT predictivity and accuracy breaks down for the best-performing models, due apparently to a reliance on different image features.

2.2 *Biological plausibility*

DCNNs are often taken to be the paradigmatic example of biological inspiration in artificial intelligence, which then fed back into neuroscience. An early precursor to modern DCNNs, the Neocognitron (Fukushima, 1980), was explicitly designed to mimic known receptive field properties of visual cortex. However, visual cortex is not really convolutional in the strict sense: unlike units in a convolutional layer, receptive fields in V1 are not simply shifted copies of one another (although this is a reasonable first-order approximation). The variability of receptive field shapes was already noted by Hubel and Wiesel (1959) in their pioneering physiology work:

> Some fields had long narrow central regions with extensive flanking areas: others had a large central area and concentrated slit-shaped flanks. In many fields the two flanking regions were asymmetrical, differing in

> size and shape; in these a given spot gave unequal responses in symmetrically corresponding regions. In some units only two regions could be found, one excitatory and the other inhibitory, lying side by side.
> (pp. 579-580)

A later study reported another violation of shift invariance: receptive field size increases with eccentricity away from the fovea (Hubel and Wiesel, 1974). It is possible to construct eccentricity-dependent deep neural networks (e.g., Chen et al., 2017; Deza and Konkle, 2020), but the important point is that these diverge from the central idea underlying convolutional networks.

Another problem with convolution is that it's not really clear how it could be implemented—real neurons don't share weights. Yamins and DiCarlo (2016) speculated that convolutional structure might emerge from experience-dependent plasticity even if it's not built into the architecture, due to the inherent shift invariance of visual input (i.e., things tend to look similar regardless of their location relative to the viewer). There is relatively little evidence that non-convolutional networks autonomously learn shift-invariant filters (though see Ingrosso and Goldt, 2022, for some specific conditions under which this works), and in any case we've already pointed out that representations in visual cortex aren't truly shift-invariant. Pogodin et al. (2021) studied a model with lateral connections between neurons within a layer that are updated using local learning rules (see next chapter). They showed that this model converges to a near-convolutional solution. Other ways to get a similar solution involve augmenting the training set with additional translations (Ott et al., 2020) or pruning/regularizing low-magnitude weights (Neyshabur, 2020; Pellegrini and Biroli, 2022).

Most DCNNs used in vision science have been purely feedforward. However, the primate visual system has extensive feedback and lateral connections. To some extent, work on object categorization has sidestepped this discrepancy by focusing on the "core" task of categorizing briefly presented and backward-masked stimuli (DiCarlo et al., 2012). This is thought to mainly rely on feedforward processing. Nonetheless, object categorization under naturalistic conditions will generally involve feedback and lateral processing (Kreiman and Serre, 2020). Several models have explored the implications of incorporating these processes (Figure 7).

Feedback and lateral connections are sometimes referred to collectively as *recurrent* connections.

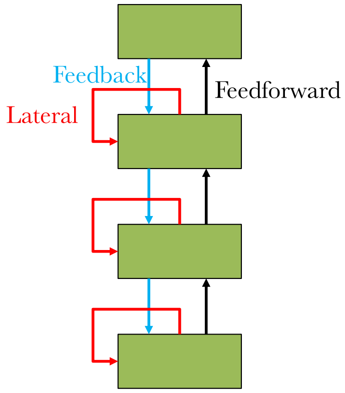


Figure 7: **An architecture with recurrence.**

In a computational study, Spoerer et al. (2017) trained DCNNs on a digit categorization task under varying levels of clutter. They compared standard feedforward DCNNs with variants that also included lateral connections, feedback connections, or both. Their main finding was that the model with both lateral and feedback connections performed best under high levels of clutter. The same model could capture bidirectional information flow between ventral stream regions (Kietzmann et al., 2019). Consistent with the hypothesis that recurrence is critical for categorizing degraded or occluded objects, backward masking (ostensibly attenuating feedback processes) significantly impairs both object categorization performance (Wyatte et al., 2012; Tang et al., 2018) and decoding of object information (Rajaei et al., 2019) specifically under occlusion (Figure 8). Further evidence comes from the finding that especially challenging images can only be decoded from IT at a behaviorally predictive level after a delay, as predicted by models with recurrence (Kar et al., 2019). Finally, feedback connections can be used to "steer" DCNNs towards particular goal-directed representations (Konkle and Alvarez, 2023), such as attending to one object in an image with several objects—similar to the way in which ventral stream representations are modulated by object-based attention (O'Craven et al., 1999).

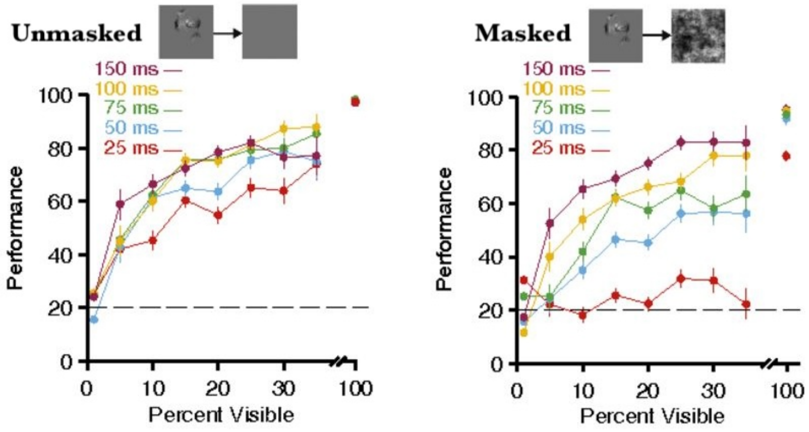


Figure 8: **Human object categorization performance under masked and unmasked conditions**. Each line corresponds to a different exposure time. Adapted from Tang et al. (2018).

We should be careful not to conclude from these studies that recurrence is *necessary* for successful performance (e.g., under occlusion), because technically a recurrent network can be "unfolded" into an equivalent feedforward network (see Wichmann and Geirhos, 2023, for further discussion). Nonetheless, the data support the hypothesis that recurrence is involved in visual processing, and that it is probably necessary for a complete account of object categorization in the brain.

3 *Theoretical perspectives*

Object categorization has been studied from several other theoretical perspectives. In this section, we show how DCNNs are connected to these perspectives.

3.1 Connection to Bayesian inference

A central organizing principle of this book is that the brain solves many computational problems using Bayesian inference. While on the surface the DCNNs reviewed above do not appear to be doing Bayesian inference *explicitly*, we can show that they are in fact doing it *implicitly*. The following derivation follows Xie (2025).

The analysis of human categorization as Bayesian inference goes back to Fried and Holyoak (1984).

As above, let s denote the object category label for sensory input x. Categorization can thus be framed as the problem of computing the posterior $p(s|x) \propto p(x|s)p(s)$. Rather than solve this problem directly, we will assume that the sensory input has been encoded into a neural representation $\phi(x)$, which is then decoded according to a distribution $q(s|\phi(x))$. We will use $q(s|x) = q(s|\phi(x))$ to denote the complete mapping from input to labels. Note that $q(s|x)$ is not required to be a Bayesian posterior; rather, both the neural representation and the decoder are chosen to optimize an objective function.

To derive the objective function, we start with a loss function $L(q, s, x)$, which penalizes q for "betting" on the wrong label given the input x. The loss is minimized when q places all its probability mass on s. A standard choice is the *cross-entropy loss*:

The cross-entropy loss is also sometimes known as the *log loss*.

$$L(q, s, x) = -\log q(s|x), \qquad (2)$$

where s here refers to the ground truth category label. The goal is to minimize the *expected* loss $\bar{L}(q) = \mathbb{E}[L(q, s, x)]$, also known as the *risk*, where the expectation is taken with respect to the joint distribution $p(s, x)$. In practice, an agent does not have access to the expected loss, but does have access to an empirical approximation (the empirical risk), $\hat{L}(q)$, based on a set of M training examples, $\{x_m, s_m\}_{m=1}^{M}$ sampled from $p(s, x)$:

Minimizing $\hat{L}(q)$ is known as *empirical risk minimization*. In general, it is not possible to guarantee good performance without placing some constraints on the function class from which q is drawn (Shalev-Shwartz and Ben-David, 2014).

$$\bar{L}(q) \approx \hat{L}(q) = \frac{1}{M} \sum_{m} L(q, s_m, x_m). \qquad (3)$$

The question is what happens to $q(s|x)$ in this setting. Taking the expectation of $\hat{L}(q)$ with respect to $p(s|x)$ under the cross-entropy loss yields:

$$\begin{aligned} \mathbb{E}[\hat{L}(q)|x] &= \mathbb{E}\left[-\sum_{s} p(s|x) \log q(s|x) \middle| x \right] \\ &= \mathcal{D}[p(s|x)||q(s|x)] + \mathcal{H}[p(s|x)], \end{aligned} \qquad (4)$$

where $\mathcal{D}[p(s|x)||q(s|x)]$ is the Kullback-Leibler (KL) divergence (see Chapter 3) and $\mathcal{H}[p(s|x)]$ is the entropy. Since the second term does not depend on $q(s|x)$, minimizing the expected cross-entropy loss is equivalent to minimizing the KL divergence. This minimum is achieved when $p(s|x) = q(s|x)$. In other words, optimizing a generic

Note that this equivalence requires that the network has enough representational capacity and training data such that $p(s|x) = q(s|x)$.

classifier in this way is equivalent to implicitly performing Bayesian inference, in the sense that the classifier will converge to the same probabilistic outputs as the posterior.

3.2 *Connection to exemplar models and kernel methods*

At first glance, DCNNs look very different from the most successful psychological account of categorization—*exemplar models* (Medin and Schaffer, 1978; Nosofsky, 1986; Kruschke, 1992). We will see, however, that they are closely related.

Exemplar models assume that a new exemplar (in this case, an image) is categorized by comparing it to other exemplars stored in memory. For concreteness, we will focus on the model developed by Kruschke (1992). The probability of assigning exemplar x to category s is computed by taking a weighted sum of similarities to exemplars $\{x_m\}_{m=1}^{M}$ stored in memory, and then applying a softmax (normalized exponential) transformation to a get a probability distribution:

$$q(s|x) \propto \exp\left[\sum_m \mu_{ms} k(x, x_m)\right], \tag{5}$$

where $k(x, x_m)$ is a similarity function and $\{\mu_{ms}\}$ is a set of exemplar weights trained to optimize accuracy on the categorization task.

This kind of model can be understood as a 3-layer neural network (Figure 9) where the input layer represents the exemplar, the hidden layer consists of units tuned to specific exemplar memories, and the output layer represents a distribution over category labels. The tuning of each hidden unit is defined by the similarity function.

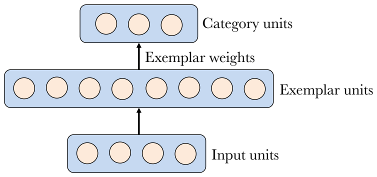


Figure 9: **Neural architecture for an exemplar model**.

A very similar idea has been used to model the categorization of 3D objects (Poggio and Edelman, 1990).

Exemplar models naturally explain why categorization accuracy is greater for exemplars repeated with high frequency compared to low frequency exemplars (Estes, 1986; Nosofsky, 1988). Intuitively, the summed similarity to the correct category is greater for these exemplars. Exemplar models also naturally explain why categorization accuracy for 3D objects monotonically declines with the distance between the viewpoints of an exemplar at training and test (e.g., Tarr, 1995): as distance increases, similarity (and hence the activation of the corresponding hidden units) declines.

While it appears that exemplar models involve completely different computational operations from DCNNs, we can connect them through the lens of kernel methods (Jäkel et al., 2009), where the similarity function is interpreted as a "kernel" function. To keep things simple, we will assume that the DCNN is fixed except for its decoder $q(s|\phi(x))$, which we will take to be a softmax function (though the key ideas extend beyond these assumptions):

$$q(s|\phi(x)) \propto \exp\left[\sum_d \mu_{ds}\phi_d(x)\right], \tag{6}$$

where w_s is a readout weight vector for category s, chosen to minimize a regularized loss function of the form $\hat{L}(w) + \Omega(w)$, where $\hat{L}(w)$ is the empirical risk (as defined in the previous section) and the regularizer $\Omega(\mu)$ penalizes large weights (more precisely, weight vectors with a large Euclidean norm). It can then be shown that the optimal decoder $q^*(s|\phi(x))$ can be written as a log-linear function of exemplar similarities:

This result is known as the *representer theorem* (Schölkopf et al., 2001).

$$q^*(s|\phi(x)) \propto \exp\left[\sum_m \mu_{ms}k(x, x_m)\right], \tag{7}$$

where the similarity function is computed by taking the dot product of the feature vectors, $k(x, x_m) = \phi(x) \cdot \phi(x_m)$. We have deliberately overloaded the notation for exemplar weights to highlight the correspondence between the two models: the optimal decoder for the DCNN has the same functional form as exemplar model described above.

The exemplar weights $\{\mu_{ms}\}$ depend on both the optimized readout weights and the similarity function.

An important aspect of this correspondence is that it changes the neural interpretation: rather than making the biologically questionable assumption that individual neurons are tuned to specific exemplars (which would require an unboundedly large number of neurons), we can adopt the more biologically defensible assumption that exemplars are labeled by the kind of feature-based transformation thought to happen in the ventral stream, with the assurance that these views are (under some conditions) equivalent.

4 *Challenges for deep neural networks*

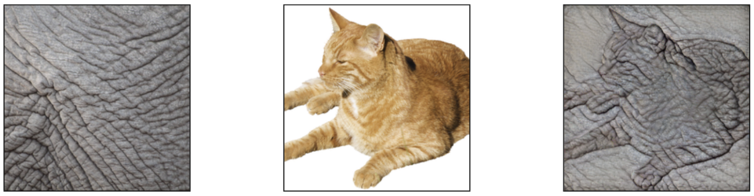
Despite their success as a model of object categorization in the ventral stream, DCNNs have been challenged on a number of fronts. Here we briefly review several points of divergence between brains and current DCNNs.

See Bowers et al. (2023) for a more comprehensive overview.

4.1 *Shape and relation sensitivity*

Humans primarily rely on shape rather than other features like color or texture to categorize objects. Studies have shown that line drawings (which lack color and texture) are recognized as quickly as color photographs (Biederman and Ju, 1988), whereas small structural changes to shape can have large effects on categorization (Biederman, 1987). For example, Biederman showed that deleting parts of a line drawing that are highly diagnostic of 3D structure could dramatically reduce categorization accuracy.

A particularly striking example of shape-sensitivity comes from a "style-transfer" task in which humans and DCNNs categorized images in which objects of one category were imprinted with textures from another category (Geirhos et al., 2019). Humans tend to ignore the transferred texture, whereas DCNNs are highly sensitive to texture (Figure 10). For example, cats rendered with elephant skin are labeled as elephants. Similar findings were reported by Baker et al. (2018). These findings are consistent with the finding that humans learn novel object categories primarily based on shape, even when non-shape features (e.g., color, position, size) are more diagnostic of the category, whereas DCNNs learn primarily based on non-shape features (Malhotra et al., 2022).



(a) Texture image
81.4% **Indian elephant**
10.3% indri
8.2% black swan

(b) Content image
71.1% **tabby cat**
17.3% grey fox
3.3% Siamese cat

(c) Texture-shape cue conflict
63.9% **Indian elephant**
26.4% indri
9.6% black swan

Figure 10: **Texture bias in DCNNs**. The numbers below each image show the top 3 DCNN outputs. Reproduced from Geirhos et al. (2019).

DCNNs are relatively more sensitive to local shape, which can produce other striking divergences from humans. Baker et al. (2018) showed that jittering contours has little effect on human categorization performance, but can dramatically change DCNN performance. In contrast, scrambling global shape dramatically reduces human performance but has relatively little effect on DCNN performance. An example is shown in Figure 11.

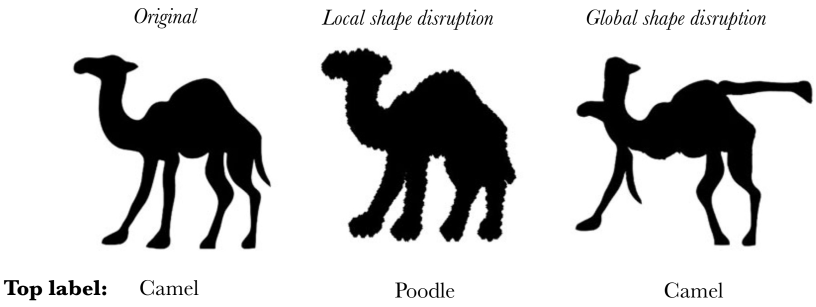


Figure 11: **Sensitivity to global vs. local shape in a DCNN**. Adapted from Baker et al. (2018).

Human categorization is also highly sensitive to spatial relations between parts: humans are much more likely to confuse object categories that share relational structure compared to those that have high pixel overlap without shared relational structure (Stankiewicz and Hummel, 1996). In contrast, DCNNs are not differentially sensitive to relational structure, even when trained on a distribution where relational structure is highly diagnostic of category membership (Malhotra et al., 2023).

4.2 *Adversarial images*

A remarkable discovery about DCNNs is that they are highly susceptible to *adversarial attacks*: an image can be distorted in such a way that it is has no effect on human category judgments (and is often imperceptible), but dramatically changes the category judgments of a DCNN (Szegedy et al., 2013). Some examples are shown in Figure 12. These images are constructed by starting with an image and its standard label (which a DCNN correctly identifies), and then searching for small pixel-level perturbations of the image such that the DCNN switches its category judgment to be strongly in favor of a different (wrong) label.

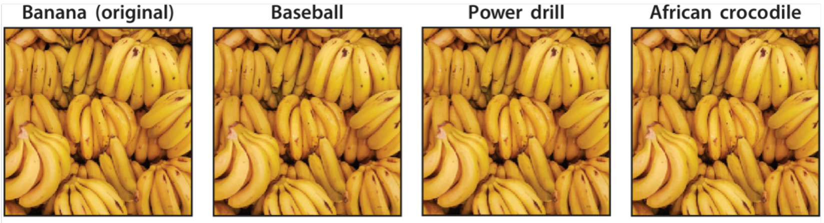


Figure 12: **Examples of adversarial images**. Reproduced from Wichmann and Geirhos (2023).

It has been claimed that humans are not susceptible to adversarial attacks (Wichmann and Geirhos, 2023), suggesting a fundamental difference between human and DCNN object categorization. Wichmann and Geirhos argue that the dependence of human object categorization on shape rather than texture means that small pixel-level perturbations will never be able to significantly alter human category judgments, since these perturbations have relatively little effect on shape. A similar argument can be made about relational structure.

5 *The richness of object perception*

Object perception is more than just categorization: we naturally perceive a wide range of material, physical, and spatial information. For example, we can easily report whether an object is soft, fluffy, smooth, elastic, heavy, fragile, large, far away, green, shiny... The list goes on and on. A complete model of object perception must be able to flexibly output all the same features that humans are able to report. No such model exists yet, but a few steps in this direction have been taken.

Although the DCNNs discussed in this chapter were trained to do categorization, this doesn't mean that they *only* do categorization. DiCarlo and Cox (2007) suggested that IT represents objects on a "flattened" manifold, as shown in Figure 13. Representations of objects from different categories are linearly separable (as dictated by the disentanglement hypothesis), but they also vary smoothly along a low-dimensional manifold, such that IT responses are weakly predictive of object pose (and other properties). Separability is ensured by flattening the manifold along the direction of the separating hyperplane. In other words, flattening means that category labels are separable while preserving smooth variation along pose or lighting dimensions.

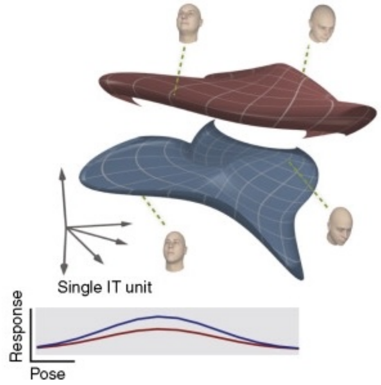


Figure 13: **Flattened manifold hypothesis for IT representations**. Reproduced from DiCarlo and Cox (2007).

Why should the ventral stream learn flattened manifolds? The fact that the encoding of category-orthogonal information in DCNNs increases over the course of training (Hong et al., 2016) suggests (somewhat paradoxically) that it must be useful for categorization. Theoretical arguments show why: the sample complexity of category learning (i.e., how many exemplars are needed to reach a target accuracy level) is lower for high-dimensional manifolds (Sorscher et al., 2022). This means that there is pressure from the training objective to prevent the manifold from completely collapsing all category-orthogonal dimensions.

An analysis of the relationship between neural geometry and generalization will be taken up in Chapter 15.

To test this hypothesis in the ventral stream, Hong et al. (2016) trained decoders for a wide range of object properties. They found that IT carried more information about many of these properties compared to earlier regions in the ventral stream (e.g., V4; Figure 14). Randomly selected subpopulations of around 700 neurons could achieve human-level accuracy on the property inference tasks. Like IT neurons, the deeper layers of a DCNN trained on object categorization could also be used to decode object properties. Thus, richer object representations may in part be an emergent property of training DCNNs to do object categorization.

Can arbitrarily rich object properties be read out from the ventral stream, and can this readout be explained by DCNNs? The answer to the first question remains to be fully worked out, but an affirmative answer is doubtful. It has been argued that the ventral stream is best understood as representing local image features, whereas the dorsal stream (extending along parietal cortex) represents global shape (Ayzenberg and Behrmann, 2022; Vaziri-Pashkam, 2024). Much like DCNNs, neurons in IT are susceptible to analogous adversarial at-

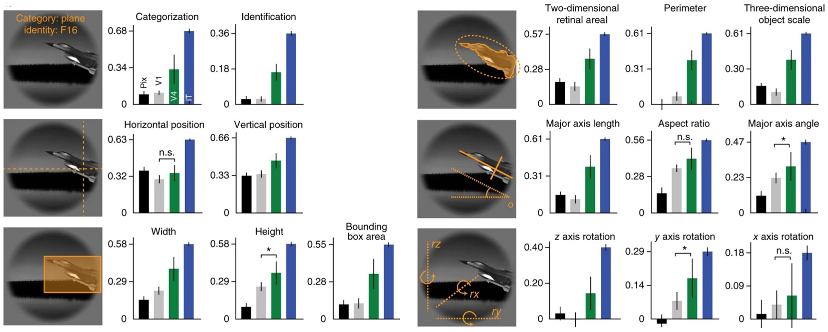


Figure 14: **Decoding category-orthogonal object properties from the ventral stream**. Reproduced from Hong et al. (2016).

tacks that imperceptibly perturb pixel values without altering global shape (Guo et al., 2022). Thus, DCNNs might be a good description of the ventral stream, but it is precisely for this reason that they are an incomplete model of object perception.

It's also unlikely that DCNNs trained on object categorization can support inferences about arbitrary category-orthogonal properties. For example, Pramod et al. (2022) showed physical stability cannot be reliably decoded from DCNN activity, mirroring the observation that physical stability could be decoded from dorsal stream areas but not from ventral stream areas. Again, this suggests that DCNNs are only part of the story about how the brain computes object properties.

Can we be more precise about what's missing? One idea is a division of labor between "graphics" in the ventral stream and "physics" in the dorsal stream (Balaban and Ullman, 2025). According to this dichotomy, the ventral stream is responsible for extracting image features, which are then used as data for reasoning about the underlying physical scene generating images. To evaluate a physical hypothesis, the dorsal stream can "render" the hypothesis into expected image features, which it can then compare with bottom-up signals along the ventral stream. This revises the classical view of the dorsal stream as a "where" pathway (computing spatial information about objects); evidence for dorsal stream involvement in representation of stability and mass suggests that spatial representation (also important for physics) is only one component of its function.

This is a particular way of formalizing the mantra that "vision is inverse graphics" (Kersten, 1997). A better formulation might be something like "vision is inverse graphics and forward physics."

6 *Conclusion*

An incredible convergence of artificial and natural intelligence is the invention of neural networks that both (i) achieve human-level object categorization performance, and (ii) quantitatively matching neural activity along the ventral stream. Nevertheless, these networks cannot explain all the relevant data on object perception, in large part because they are only really doing one part of object perception—extracting image features useful for categorization. This is likely a good description of the ventral stream, but other brain systems (e.g., a putative physics engine in the dorsal stream) are necessary to explain how we are able to extract rich inferences about the physical world from the impoverished 2D information arriving at the retina.

Study questions

1. Why is linear separability important for some models of object categorization, and what are the consequences if this assumption fails?
2. In what ways are deep convolutional neural networks biologically plausible, and in what ways do they diverge from biology?
3. How would you design alternative neural network models that better capture human sensitivity to shape and relational structure?

References

Afraz, S.-R., Kiani, R., and Esteky, H. (2006). Microstimulation of inferotemporal cortex influences face categorization. *Nature*, 442:692–695.

Ayzenberg, V. and Behrmann, M. (2022). Does the brain's ventral visual pathway compute object shape? *Trends in Cognitive Sciences*, 26:1119–1132.

Baker, N., Lu, H., Erlikhman, G., and Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLoS Computational Biology*, 14:e1006613.

Balaban, H. and Ullman, T. D. (2025). Physics versus graphics as an organizing dichotomy in cognition. *Trends in Cognitive Sciences*.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94(115-147).

Biederman, I. and Ju, G. (1988). Surface versus edge-based determinants of visual recognition. *Cognitive Psychology*, 20:38–64.

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., Puebla, G., Adolfi, F., Hummel, J. E., Heaton, R. F., et al. (2023). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, 46:e385.

Bracci, S. and Op de Beeck, H. P. (2023). Understanding human object vision: a picture is worth a thousand representations. *Annual Review of Psychology*, 74:113–135.

Chen, F. X., Roig, G., Isik, L., Boix, X., and Poggio, T. (2017). Eccentricity dependent deep neural networks: Modeling invariance in human vision. *AAAI Spring Symposium Series*.

Deza, A. and Konkle, T. (2020). Emergent properties of foveated perceptual systems. *arXiv preprint arXiv:2006.07991*.

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

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

Estes, W. (1986). Memory storage and retrieval processes in category learning. *Journal of Experimental psychology. General*, 115:155–174.

Fried, L. and Holyoak, K. (1984). Induction of category distributions: a framework for classification learning. *Journal of Experimental psychology. Learning, Memory, and Cognition*, 10:234–257.

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

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. (2019). Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations*.

Guo, C., Lee, M., Leclerc, G., Dapello, J., Rao, Y., Madry, A., and Dicarlo, J. (2022). Adversarially trained neural representations are already as robust as biological neural representations. In *International Conference on Machine Learning*, pages 8072–8081. PMLR.

Hong, H., Yamins, D. L., Majaj, N. J., and DiCarlo, J. J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature Neuroscience*, 19:613–622.

Hubel, D. H. and Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, 148:574–591.

Hubel, D. H. and Wiesel, T. N. (1974). Uniformity of monkey striate cortex: a parallel relationship between field size, scatter, and magnification factor. *Journal of Comparative Neurology*, 158:295–305.

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

Ingrosso, A. and Goldt, S. (2022). Data-driven emergence of convolutional structure in neural networks. *Proceedings of the National Academy of Sciences*, 119:e2201854119.

Jacob, G., Pramod, R., Katti, H., and Arun, S. (2021). Qualitative similarities and differences in visual object representations between brains and deep networks. *Nature Communications*, 12:1872.

Jäkel, F., Schölkopf, B., and Wichmann, F. A. (2009). Does cognitive science need kernels? *Trends in Cognitive Sciences*, 13:381–388.

Kar, K. and DiCarlo, J. J. (2024). The quest for an integrated set of neural mechanisms underlying object recognition in primates. *Annual Review of Vision Science*, 10.

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., and DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, 22:974–983.

Kersten, D. (1997). Inverse 3-D graphics: A metaphor for visual perception. *Behavior Research Methods, Instruments, & Computers*, 29:37–46.

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

Konkle, T. and Alvarez, G. (2023). Cognitive steering in deep neural networks via long-range modulatory feedback connections. *Advances in Neural Information Processing Systems*, 36:21613–21634.

Kreiman, G. and Serre, T. (2020). Beyond the feedforward sweep: feedback computations in the visual cortex. *Annals of the New York Academy of Sciences*, 1464:222–241.

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

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99:22–44.

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., and Jackel, L. (1989). Handwritten digit recognition with a back-propagation network. *Advances in Neural Information Processing Systems*, 2.

Linsley, D., Rodriguez Rodriguez, I. F., Fel, T., Arcaro, M., Sharma, S., Livingstone, M., and Serre, T. (2023). Performance-optimized deep neural networks are evolving into worse models of inferotemporal visual cortex. *Advances in Neural Information Processing Systems*, 36:28873–28891.

Majaj, N. J., Hong, H., Solomon, E. A., and DiCarlo, J. J. (2015). Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *Journal of Neuroscience*, 35:13402–13418.

Malhotra, G., Dujmović, M., and Bowers, J. S. (2022). Feature blindness: A challenge for understanding and modelling visual object recognition. *PLOS Computational Biology*, 18:e1009572.

Malhotra, G., Dujmović, M., Hummel, J., and Bowers, J. S. (2023). Human shape representations are not an emergent property of learning to classify objects. *Journal of Experimental Psychology: General*, 152:3380–3402.

Medin, D. L. and Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85:207–238.

Moeller, S., Crapse, T., Chang, L., and Tsao, D. Y. (2017). The effect of face patch microstimulation on perception of faces and objects. *Nature Neuroscience*, 20:743–752.

Neyshabur, B. (2020). Towards learning convolutions from scratch. *Advances in Neural Information Processing Systems*, 33:8078–8088.

Nosofsky, R. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of experimental psychology: General*, 115:39–61.

Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14:54–65.

O'Craven, K. M., Downing, P. E., and Kanwisher, N. (1999). fMRI evidence for objects as the units of attentional selection. *Nature*, 401:584–587.

Ott, J., Linstead, E., LaHaye, N., and Baldi, P. (2020). Learning in the machine: To share or not to share? *Neural Networks*, 126:235–249.

Pagan, M., Simoncelli, E. P., and Rust, N. C. (2016). Neural quadratic discriminant analysis: Nonlinear decoding with V1-like computation. *Neural Computation*, 28:2291–2319.

Pellegrini, F. and Biroli, G. (2022). Neural network pruning denoises the features and makes local connectivity emerge in visual tasks. In *International Conference on Machine Learning*, pages 17601–17626. PMLR.

Pinto, N., Cox, D. D., and DiCarlo, J. J. (2008). Why is real-world visual object recognition hard? *PLoS Computational Biology*, 4:e27.

Poggio, T. and Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature*, 343:263–266.

Pogodin, R., Mehta, Y., Lillicrap, T., and Latham, P. E. (2021). Towards biologically plausible convolutional networks. *Advances in Neural Information Processing Systems*, 34:13924–13936.

Pramod, R., Cohen, M. A., Tenenbaum, J. B., and Kanwisher, N. (2022). Invariant representation of physical stability in the human brain. *Elife*, 11:e71736.

Rajaei, K., Mohsenzadeh, Y., Ebrahimpour, R., and Khaligh-Razavi, S.-M. (2019). Beyond core object recognition: Recurrent processes account for object recognition under occlusion. *PLoS Computational Biology*, 15:e1007001.

Rajalingham, R. and DiCarlo, J. J. (2019). Reversible inactivation of different millimeter-scale regions of primate IT results in different patterns of core object recognition deficits. *Neuron*, 102:493–505.

Rajalingham, R., Schmidt, K., and DiCarlo, J. J. (2015). Comparison of object recognition behavior in human and monkey. *Journal of Neuroscience*, 35:12127–12136.

Rollenhagen, J. and Olson, C. (2000). Mirror-image confusion in single neurons of the macaque inferotemporal cortex. *Science*, 287:1506–1508.

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

Schölkopf, B., Herbrich, R., and Smola, A. J. (2001). A generalized representer theorem. In *International Conference on Computational Learning Theory*, pages 416–426. Springer.

Sekuler, R. W. and Houlihan, K. (1968). Discrimination of mirror-images: Choice time analysis of human adult performance. *The Quarterly Journal of Experimental Psychology*, 20:204–207.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.

Sorscher, B., Ganguli, S., and Sompolinsky, H. (2022). Neural representational geometry underlies few-shot concept learning. *Proceedings of the National Academy of Sciences*, 119:e2200800119.

Spoerer, C. J., McClure, P., and Kriegeskorte, N. (2017). Recurrent convolutional neural networks: a better model of biological object recognition. *Frontiers in Psychology*, 8:1551.

Stankiewicz, B. J. and Hummel, J. E. (1996). Categorical relations in shape perception. *Spatial Vision*, 10:201–236.

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

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

Tarr, M. J. (1995). Rotating objects to recognize them: A case study on the role of viewpoint dependency in the recognition of three-dimensional objects. *Psychonomic Bulletin & Review*, 2:55–82.

Vaziri-Pashkam, M. (2024). Two "what" networks in the human brain. *Journal of Cognitive Neuroscience*, 36:2584–2593.

Weiskrantz, L. and Saunders, R. (1984). Impairments of visual object transforms in monkeys. *Brain*, 107:1033–1072.

Wichmann, F. A. and Geirhos, R. (2023). Are deep neural networks adequate behavioral models of human visual perception? *Annual Review of Vision Science*, 9:501–524.

Wyatte, D., Curran, T., and O'Reilly, R. (2012). The limits of feedforward vision: Recurrent processing promotes robust object recognition when objects are degraded. *Journal of Cognitive Neuroscience*, 24:2248–2261.

Xie, Y. (2025). How do we interpret the outputs of a neural network trained on classification? In *ICLR Blogposts 2025*. https://d2jud02ci9yv69.cloudfront.net/2025-04-28-interpret-classification-11/blog/interpret-classification/.

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

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

Yang, Q., Walker, E., Cotton, R. J., Tolias, A. S., and Pitkow, X. (2021). Revealing nonlinear neural decoding by analyzing choices. *Nature Communications*, 12:6557.

Zoccolan, D., Kouh, M., Poggio, T., and DiCarlo, J. J. (2007). Trade-off between object selectivity and tolerance in monkey inferotemporal cortex. *Journal of Neuroscience*, 27:12292–12307.