Lecture 8: Object categorization

Samuel Gershman

Harvard University

Roadmap

- High-level perception transforms the continuous world of sensory data into discrete categories like people, animals, and vehicles.

Roadmap

- High-level perception transforms the continuous world of sensory data into discrete categories like people, animals, and vehicles.
- In the visual system, this transformation happens along the "ventral stream" extending from primary visual cortex (V1) to inferotemporal (IT) cortex.

Roadmap

- High-level perception transforms the continuous world of sensory data into discrete categories like people, animals, and vehicles.
- In the visual system, this transformation happens along the "ventral stream" extending from primary visual cortex (V1) to inferotemporal (IT) cortex.
- At the end of this transformation, object categories can be read out from neural population activity by a linear decoder.

Roadmap

- High-level perception transforms the continuous world of sensory data into discrete categories like people, animals, and vehicles.
- In the visual system, this transformation happens along the "ventral stream" extending from primary visual cortex (V1) to inferotemporal (IT) 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. But there remain gaps between these models and human perception, suggesting flexible generative models of sensory data.

The computational problem

- The last lecture focused on two-alternative decisions, but many naturalistic decision problems involve far more alternatives.

The computational problem

- The last lecture focused on two-alternative decisions, but many naturalistic decision problems involve far more alternatives.
- Central function of high-level perception is categorizing objects present in sensory data, where the number of object categories is in the thousands.

The computational problem

- The last lecture focused on two-alternative decisions, but many naturalistic decision problems involve far more alternatives.
- Central function of high-level perception is categorizing objects present in sensory data, where the number of object categories is in the thousands.
- More than just a multi-alternative generalization of two-alternative problems, because it places greater demands on representation: category information is “entangled” at the level of low-level sensory cortex (e.g., V1)—it cannot be read out with a linear decoder of the sort that we used to model the evidence accumulator in LIP.

Linear decoders

- Linear decoder:

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

where s denotes the object category, 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.

Linear decoders

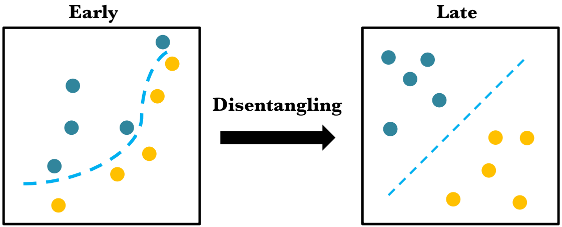
- Linear decoder:

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

  where s denotes the object category, 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.
- Early sensory representations need to be nonlinearly transformed such that category information is "disentangled"—i.e., linearly decodable.

Linear separability and disentanglement

Circles: exemplars (e.g., images); colors: category labels; axes: activity levels of different neurons (here just 2 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.



The ventral visual stream

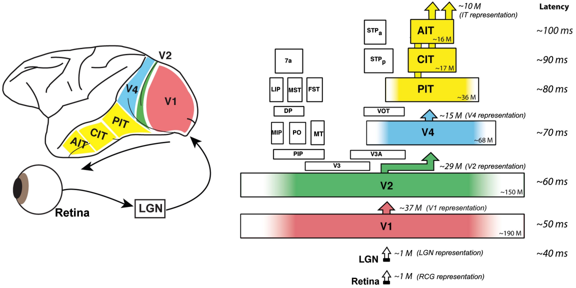
- It is widely believed that disentangling is achieved by a sequence of representational transformations along the ventral visual stream, extending from V1 to IT.

The ventral visual stream

- It is widely believed that disentangling is achieved by a sequence of representational transformations along the ventral visual stream, extending from V1 to IT.
- Several changes are apparent across the ventral stream: (1) loss of retinotopy; (2) increase in receptive field size; (3) transition from encoding low-level features to high-level semantic categories; (4) emergence of linear decodability.

The ventral visual stream

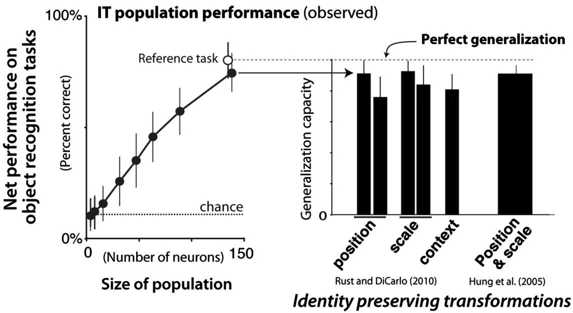
(Left) Anatomical organization in the primate brain. (Right) Each area's size is proportional to its cortical surface area, with the approximate number of neurons shown in the corner of each area. The approximate dimensionality of each representation (number of projection neurons) is shown above each area. Approximate median response latency is shown to the right of each area.



[DiCarlo et al 2012]

Classification performance from IT population activity

Performance generalizes across small position, scale, and context (background) variations.



[DiCarlo et al 2012]

Causal evidence that IT is important for object categorization

- 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].

Causal evidence that IT is important for object categorization

- 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 [Wiskrantz & Saunders 1984; Rajalingham & DiCarlo 2019].

Modeling the ventral stream with deep neural networks

- Deep convolutional neural networks (DCNNs) currently provide the most successful quantitative account of how the ventral stream achieves disentangling of object category information, though they are not without problems.

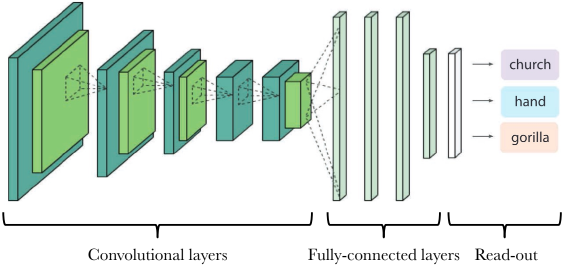
Modeling the ventral stream with deep neural networks

- Deep convolutional neural networks (DCNNs) currently provide the most successful quantitative account of how the ventral stream achieves disentangling of object category information, though they are not without problems.
- Consists of multiple layers, each containing multiple “units” (roughly corresponding to neurons or populations of neurons) that send outputs to the next layer. Units take a linear combination of inputs and then pass them through a non-linearity—thus implementing a linear decoder.

Modeling the ventral stream with deep neural networks

- Deep convolutional neural networks (DCNNs) currently provide the most successful quantitative account of how the ventral stream achieves disentangling of object category information, though they are not without problems.
- Consists of multiple layers, each containing multiple “units” (roughly corresponding to neurons or populations of neurons) that send outputs to the next layer. Units take a linear combination of inputs and then pass them through a non-linearity—thus implementing a linear decoder.
- In convolutional layers, all units share the same set of synaptic weights but apply these weights to different inputs (e.g., subregions of images). The shared weights function as learnable “filters” that are convolved with the input—i.e., applied uniformly to every part of the input.

Deep convolutional neural network for object categorization



[Bracci & Op de Beeck 2023]

Modeling the ventral stream with deep neural networks

- Training a DCNN involves adjusting the weights to optimize an objective function, which measures how well the network classifies training images.

Modeling the ventral stream with deep neural networks

- 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.

Modeling the ventral stream with deep neural networks

- 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 lecture, we will discuss how such an algorithm might be plausibly implemented in a neural circuit.

Correspondences between the ventral stream and DCNNs

- As noted, linear decoders of IT are able to match stimulus-specific confusions (i.e., misclassifications); the same is true for DCNNs [Rajalingham et al 2015].

Correspondences between the ventral stream and DCNNs

- As noted, linear decoders of IT are able to match stimulus-specific confusions (i.e., misclassifications); the same is true for DCNNs [Rajalingham et al 2015].
- Like human, DCNNs have greater difficulty discriminating mirror reflections of an image along the horizontal axis compared to reflections along the vertical axis [Jacob et al 2021.

Correspondences between the ventral stream and DCNNs

- As noted, linear decoders of IT are able to match stimulus-specific confusions (i.e., misclassifications); the same is true for DCNNs [Rajalingham et al 2015].
- Like human, DCNNs have greater difficulty discriminating mirror reflections of an image along the horizontal axis compared to reflections along the vertical axis [Jacob et al 2021.
- DCNNs exhibit greater similarity in late (putatively IT) activity patterns for horizontal reflections compared to vertical reflections. Measurements of IT show the same effect [Rollenhagen & Olson 2000].

Correspondences between the ventral stream and DCNNs

- Yamins et al [2014] fit a linear mapping from the final layer of the DCNN to IT population activity, and then evaluated this mapping on held-out data.

Correspondences between the ventral stream and DCNNs

- Yamins et al [2014] 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 better neural predictivity than any previous model, and that predictivity improved with categorization accuracy.

Correspondences between the ventral stream and DCNNs

- Yamins et al [2014] 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 better neural predictivity than any previous model, and that predictivity improved with categorization accuracy.
- Earlier layers provided good predictivity for upstream regions in the ventral stream (V1 and V4).

Correspondences between the ventral stream and DCNNs

- Yamins et al [2014] 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 better neural predictivity than any previous model, and that predictivity improved with categorization accuracy.
- Earlier layers provided good predictivity for upstream regions in the ventral stream (V1 and V4).
- Thus, DCNNs appeared to recapitulate the key transformational steps in the ventral stream.

Biological plausibility of DCNNs?

- 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 (except locally).

Biological plausibility of DCNNs?

- 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 (except locally).
- Not clear how convolution could be implemented—real neurons don't share weights.

Biological plausibility of DCNNs?

- 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 (except locally).
- Not clear how convolution could be implemented—real neurons don't share weights.
- DCNNs are (usually) feedforward, but the primate visual system has extensive feedback and lateral connections.

Recurrent models

- 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.

Recurrent models

- 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.
- The model with both lateral and feedback connections performed best under high levels of clutter.

Recurrent models

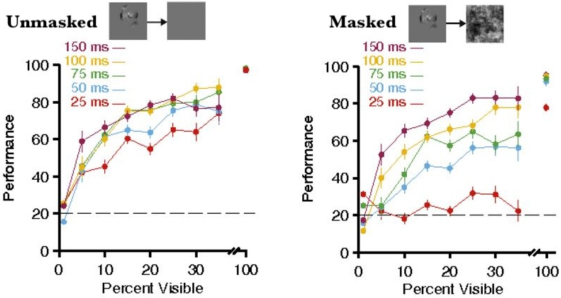
- Spoerer et al [2017] trained DCNNs on a digit categorization task under varying levels of clutter. They compared standard feedforward DCNNs with variants that also included lateral connections, feedback connections, or both.
- The model with both lateral and feedback connections performed best under high levels of clutter.
- The same model could capture bidirectional information flow between ventral stream regions [Kietzmann et al 2019].

Recurrent models

- 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.
- 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 and decoding of object information under occlusion.

Recurrent models

Human object categorization performance under masked and unmasked conditions. Each line corresponds to a different exposure time.



[Tang et al 2018]

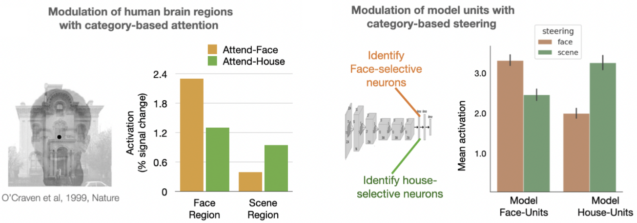
Recurrent models

- 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].

Recurrent models

- 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].
- Feedback connections can be used to “steer” DCNNs towards particular goal-directed representations [Konkle & 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].

Steering



[O'Craven et al 1999; Konkle & Alvarez 2023]

Connection to 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*.

Connection to 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*.
- Let s denote the object category label for sensory input x. Categorization can thus be framed as the problem of computing the posterior $p(s|x) \propto p(x|s)p(s)$.

Connection to 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*.
- 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.

Connection to 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*.
- 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.

Connection to Bayesian inference

- Cross-entropy loss function $L(q, s, x) = -\log q(s|x)$ that penalizes q for “betting” on the wrong label given the input x.

Connection to Bayesian inference

- Cross-entropy loss function $L(q, s, x) = -\log q(s|x)$ that 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

Connection to Bayesian inference

- Cross-entropy loss function $L(q, s, x) = -\log q(s|x)$ that 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
- Goal is to minimize the *expected* loss $\bar{L}(q) = \mathbb{E}[L(q, s, x)]$.

Connection to Bayesian inference

- Agent doesn't have access to the expected loss, but does have access to an empirical approximation based on a set of M training examples, $\{x_m, s_m\}_{m=1}^{M}$ sampled from $p(s, x)$:

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

Connection to Bayesian inference

- Agent doesn't have access to the expected loss, but does have access to an empirical approximation based on a set of M training examples, $\{x_m, s_m\}_{m=1}^{M}$ sampled from $p(s, x)$:

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

- Taking the expectation with respect to $p(s|x)$ under the cross-entropy loss yields:

$$\mathbb{E}[\hat{L}(q)|x] = \mathcal{D}[p(s|x)||q(s|x)] + \mathcal{H}[p(s|x)]$$

where $\mathcal{D}[p(s|x)||q(s|x)]$ is the KL divergence and $\mathcal{H}[p(s|x)]$ is the entropy.

Connection to Bayesian inference

- Agent doesn't have access to the expected loss, but does have access to an empirical approximation based on a set of M training examples, $\{x_m, s_m\}_{m=1}^{M}$ sampled from $p(s,x)$:

$$\bar{L}(q) \approx \hat{L}(q) = \frac{1}{M}\sum_m L(q, s_m, x_m)$$

- Taking the expectation with respect to $p(s|x)$ under the cross-entropy loss yields:

$$\mathbb{E}[\hat{L}(q)|x] = \mathcal{D}[p(s|x)||q(s|x)] + \mathcal{H}[p(s|x)]$$

  where $\mathcal{D}[p(s|x)||q(s|x)]$ is the KL divergence and $\mathcal{H}[p(s|x)]$ is the entropy.

- Second term does not depend on $q(s|x)$, so minimizing the expected cross-entropy loss is equivalent to minimizing the KL divergence, achieved when $p(s|x) = q(s|x)$.

Connection to Bayesian inference

- To summarize the argument: optimizing a generic 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.

Connection to Bayesian inference

- To summarize the argument: optimizing a generic 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.
- Bayesian inference is inevitable!

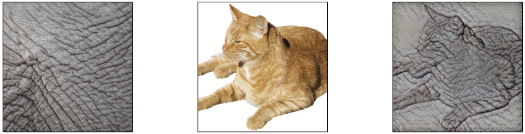
Challenges for deep neural networks: shape vs. texture

- Humans primarily rely on shape rather than other features like color or texture to categorize objects, but DCNNs seem to rely more on texture.

Challenges for deep neural networks: shape vs. texture

- Humans primarily rely on shape rather than other features like color or texture to categorize objects, but DCNNs seem to rely more on texture.
- 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].

Texture bias in DCNNs



(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

[Geirhos et al 2019]

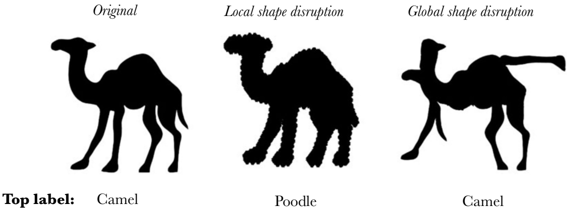
Shape sensitivity

- DCNNs are relatively more sensitive to local shape, which can produce other striking divergences from humans.

Shape sensitivity

- DCNNs are relatively more sensitive to local shape, which can produce other striking divergences from humans.
- 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.

Sensitivity to global vs. local shape



[Baker et al 2018]

Spatial relations

- 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 & Hummel 1996].

Spatial relations

- 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 & 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].

Study question

How would you design alternative neural network models that better capture human sensitivity to shape and relational structure?

Adversarial images

- DCNNs 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].

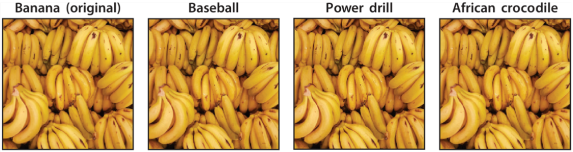
Adversarial images

- DCNNs 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].
- 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.

Adversarial images

- DCNNs 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].
- 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.
- 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.

Adversarial images



Banana (original)

Baseball

Power drill

African crocodile

[Wichmann & Geirhos 2023]

The richness of object perception

- Object perception is more than just categorization: we naturally perceive a wide range of material, physical, and spatial information.

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 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...
- 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.

The flattened manifold hypothesis

- Although the DCNNs are trained to do categorization, this doesn't mean that they *only* do categorization.

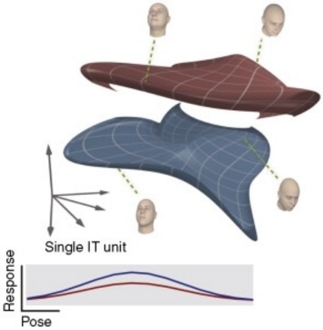
The flattened manifold hypothesis

- Although the DCNNs are trained to do categorization, this doesn't mean that they *only* do categorization.
- DCNNs (and IT) might represent objects on a "flattened" manifold [DiCarlo et al 2007]. Representations of objects from different categories are linearly separable, but they also vary smoothly along a low-dimensional manifold, such that IT responses are weakly predictive of object pose (and other properties).

The flattened manifold hypothesis

- Although the DCNNs are trained to do categorization, this doesn't mean that they *only* do categorization.
- DCNNs (and IT) might represent objects on a "flattened" manifold [DiCarlo et al 2007]. Representations of objects from different categories are linearly separable, 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.

The flattened manifold hypothesis



[DiCarlo et al 2007]

The flattened manifold hypothesis

- 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.

The flattened manifold hypothesis

- 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].

The flattened manifold hypothesis

- 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.

Testing the flattened manifold hypothesis

- 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).

Testing the flattened manifold hypothesis

- 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).
- Randomly selected subpopulations of around 700 neurons could achieve human-level accuracy on the property inference tasks.

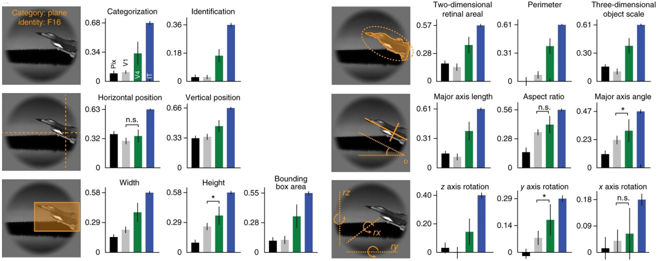
Testing the flattened manifold hypothesis

- 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).
- 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.

Testing the flattened manifold hypothesis

- 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).
- 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.

Decoding category-orthogonal object properties from the ventral stream



[Hong et al 2016]

Do DCNN models of the ventral stream account for the richness of object perception?

- Much like DCNNs, neurons in IT are susceptible to analogous adversarial attacks that imperceptibly perturb pixel values without altering global shape [Guo et al 2022].

Do DCNN models of the ventral stream account for the richness of object perception?

- Much like DCNNs, neurons in IT are susceptible to analogous adversarial attacks 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.

Do DCNN models of the ventral stream account for the richness of object perception?

- Much like DCNNs, neurons in IT are susceptible to analogous adversarial attacks 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 that 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.

What's missing?

- One idea is a division of labor between "graphics" in the ventral stream and "physics" in the dorsal (parietal) stream [Balaban & Ullman 2025].

What’s missing?

- One idea is a division of labor between “graphics” in the ventral stream and “physics” in the dorsal (parietal) stream [Balaban & 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.

What’s missing?

- One idea is a division of labor between “graphics” in the ventral stream and “physics” in the dorsal (parietal) stream [Balaban & 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.

What’s missing?

- One idea is a division of labor between “graphics” in the ventral stream and “physics” in the dorsal (parietal) stream [Balaban & 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); dorsal stream involvement in representation of stability and mass suggests that spatial representation (also important for physics) is only one component of its function.

Summary

- 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.

Summary

- 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.

Summary

- 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.