Cognition and Neuroscience (Module 2)

Last update: 19 May 2024

Academic Year 2023 – 2024

Alma Mater Studiorum · University of Bologna

Contents

1 Object recognition

Vision Process that, from images of the external world, produces a description without irrelevant information (i.e. interference) useful to the viewer. Vision

This description includes information such as what is in the world and where it is.

Remark. Vision is the most important sense in primates (i.e. in case of conflicts between senses, vision is usually prioritized).

Remark. Vision is also involved in memory and thinking.

Remark. The two main tasks for vision are:

- Object recognition.
- Guiding movement.

These two functions are mediated by (at least) two pathways that interact with each other.

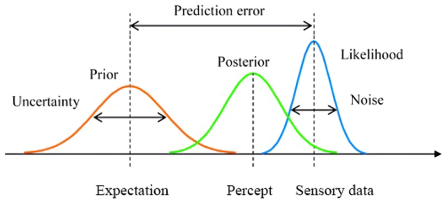
Vision Bayesian modeling Vision can be modeled using Bayesian theory. Bayesian modeling

Given an image I and a stimulus S, an ideal observer uses some prior knowledge (expectation) of the stimulus ($\mathcal{P}(S)$) and input sensory data ($\mathcal{P}(I|S)$) to infer the most probable interpretation of the stimulus in the image:

$$\mathcal{P}(S|I) = \frac{\mathcal{P}(I|S)\mathcal{P}(S)}{\mathcal{P}(I)}$$

Remark. Prior knowledge is learned from experience. It could be related to the shape of the object, the direction of the light or the fact that objects cannot overlap.

Remark. If the image is highly ambiguous, prior knowledge contributes more to disambiguate it.



Feed-forward processing Involves the likelihood $\mathcal{P}(I|S)$.

Feed-back processing Involves the prior $\mathcal{P}(S)$.

Remark. Perception integrates both feed-forward and feed-back processing.

Vision levels A visual scene is analyzed at three levels: Vision levels

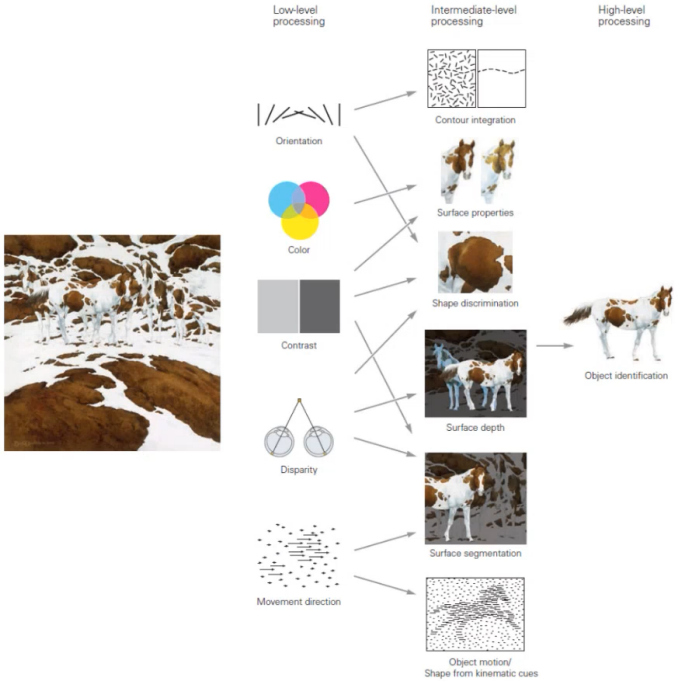
Low level Processes simple visual attributes captured by the retina such as local contrast, orientation, color, depth and motion.

Intermediate level Low-level features are used to parse the visual scene (i.e. local features are integrated into the global image). This level is responsible for identifying boundaries and surfaces belonging to the same object and discriminating between foreground and background objects.

High level Responsible for object recognition.

Once the objects have been recognized, they can be associated with memories of shapes and meaning.

Case study (Agnosia)**.** Patients with agnosia have their last level of vision damaged. They can see (e.g. avoid obstacles) but cannot recognize object or get easily confused.

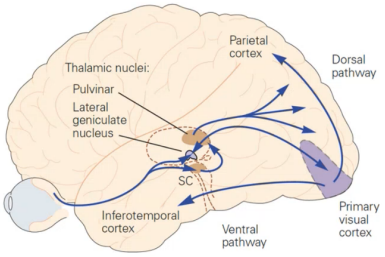


1.1 Pathways

Retino-geniculo-striate pathway Responsible for visual processing. It includes the:

Retino-geniculo-striate pathway

- Retina.
- Lateral geniculate nucleus (LGN) of the thalamus.
- Primary visual cortex (V1) or striate cortex.
- Extrastriate visual areas (i.e. beyond the area V1).



Ventral pathway Responsible for object recognition. It extends from the area V1 to the temporal lobe (feed-forward processing). Ventral pathway

Remark. This pathway emphasizes color.

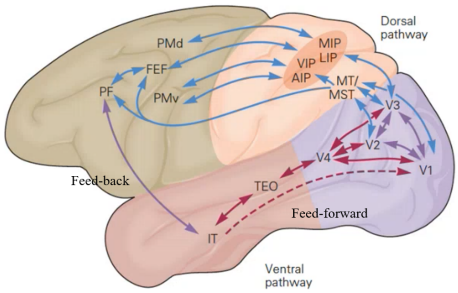
Remark. The connection from the frontal lobe encodes prior knowledge (feed-back processing).

Dorsal pathway Responsible for movement guiding. It connects the V1 area with the parietal lobe and then with the frontal lobe. Dorsal pathway

Remark. This pathway is colorblind.

Remark. The ventral and dorsal pathways are highly connected and share information.

Remark. All connections in the ventral and dorsal pathways are reciprocal (i.e. bidirectional).



1.2 Neuron receptive field

Single-cell recording Technique to record the firing rate of neurons. A fine-tipped electrode is inserted into the animal's brain to record the action potential of a single neuron. Single-cell recording

This method is highly invasive but allows to obtain high spatial and temporal readings of the neuron firing rate while distinguishing excitation and inhibition.

Remark. On a theoretical level, neurons can fire a maximum of 1000 times per second. This may actually happen in exceptional cases.

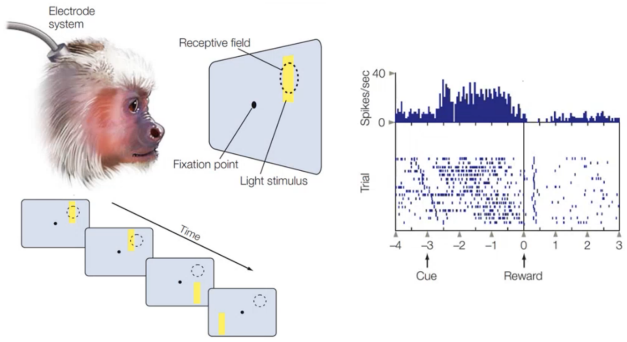
Receptive field Region of the visual scene at which a particular neuron will respond if a stimulus falls within it. Receptive field

Remark. The receptive field of a neuron can be described through a Gaussian.

Case study. A monkey is trained to maintain fixation at a point on a screen. Then, stimuli are presented in various positions of the visual field.

It has been seen that a particular neuron fires vigorously only when a stimulus is presented in a particular area of the screen.

The response is the strongest at the center of the receptive field and gradually declines when the stimulus moves away from the center.



Remark. Neurons might only react to a specific type of stimuli in the receptive field (e.g. color, direction, ...).

Case study. It has been seen that a neuron fires only if a stimulus is presented in its receptive field while moving upwards.

Retinotopy Mapping of visual inputs from the retina (i.e. receptive field) to the neurons. Retinotopy

There is a non-random relationship between the position of the neurons in the visual areas (V1, V2, V4): their receptive fields form a 2D map of the visual field in such a way that neighboring regions in the visual image are represented by adjacent regions of the visual cortical area (i.e. the receptive fields are spatially ordered).

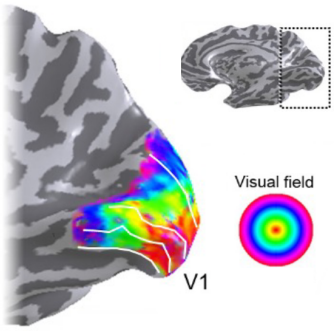


Figure 1.1: Mapping from the visual field to the neurons in the primary visual cortex (V1)

Eccentricity The diameter of the receptive field is proportional to the wideness of the visual angle. Eccentricity

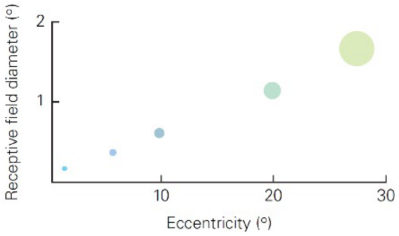


Figure 1.2: Relationship between visual angle and receptive field diameter

Cortical magnification The neurons (retinal ganglion cells, RGCs) responsible for Cortical magnification

the center of the visual field (fovea) have a visual angle of about 0.1° while the neurons at the visual periphery reach up to 1° of visual angle.

Accordingly, more cortical space is dedicated to the central part of the visual field. This densely packed amount of smaller receptive fields allows for the highest spatial resolution at the center of the visual field.

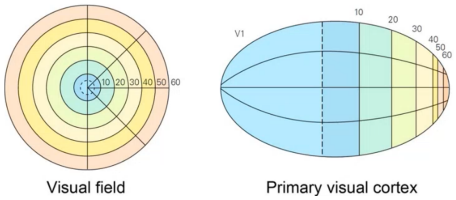


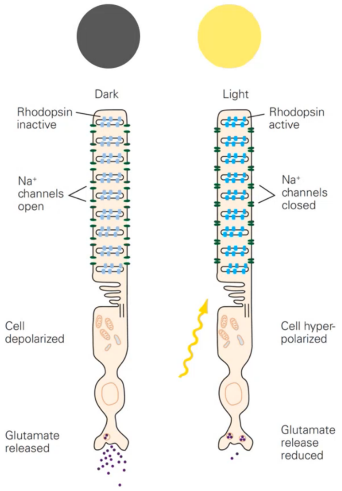
Figure 1.3: Cortical magnification in V1

Remark. The brain creates the illusion that the center and the periphery of vision are equal. In reality, the periphery has less resolution and is colorblind.

Hierarchical model of receptive field The processing of some information in the visual image is done through an incremental convergence of information in a hierarchy of receptive fields of neurons. Along the hierarchy the size of the receptive field increases.

1.3 Retina cells

Photoreceptor Specialized neurons that are hyperpolarized in bright regions and depolarized in dark regions. Photoreceptor



Remark. They are the first layer of neurons in the retina.

Retinal ganglion cell (RGC) Neurons of the visual cortex with a circular receptive field. They are categorized into: Retinal ganglion cell (RGC)

ON-center RGCs that are activated in response to a bright stimulus in the center of the receptive field. ON-center cells

OFF-center RGCs that are activated in response to a dark stimulus in the center of the receptive field. OFF-center cells

Remark. They are the last layer of neurons in the retina, before entering LGN of the thalamus.

The receptive field of RGCs is composed of two concentric areas. The inner one acts according to the type of the cell (ON-center/OFF-center) while the outer circle acts antagonistically to the inner area.

Remark. A uniform stimulus that covers the entire receptive field of an RGC produces a weak or no response.

Remark. RGCs are not responsible for distinguishing orientation or lines.

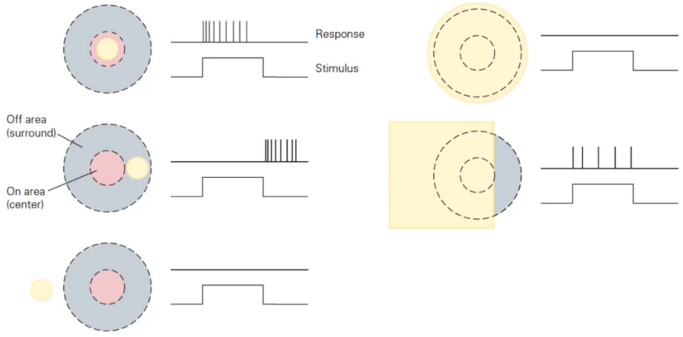
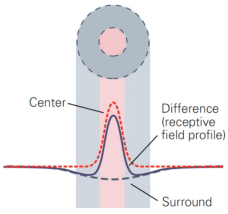
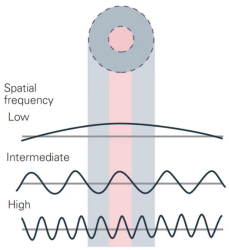


Figure 1.4: Responses of an ON-center RGC

Remark. The response of RGCs can be described by two Gaussians: one is positive with a narrow peak and represents the response at the center while the other is negative with a wide base and covers both the inner and outer circles. Their difference represents the response of the cell (receptive field profile).



Band-pass behavior The visual system of humans has a band-pass behavior: it only responds to a narrow band of intermediate frequencies and is unable to respond to spatial frequencies that are too high or too low (as they get canceled by the two antagonistic circles). Band-pass behavior

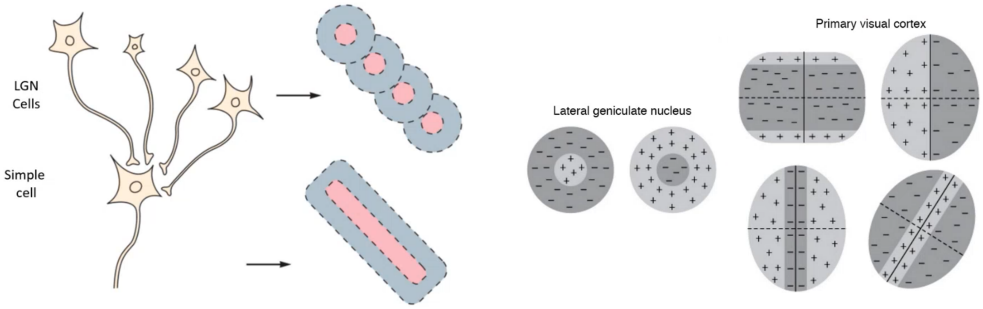


1.4 Area V1 cells

1.4.1 Simple cells

Simple cells

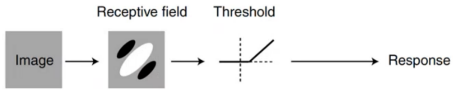
Neurons that respond to a narrow range of orientations and spatial frequencies.
This is the result of the alignment of different circular receptive fields of the LGN cells (which in turn receive their input from RGCs).



Case study. In monkeys, the neurons of the LGN have non-oriented circular receptive fields. Still, simple cells are able to perceive specific orientations.

Simple cells model The stages of computation in simple cells are:

1. Linear filtering through a weighted sum of the image intensities done by the receptive field (i.e. convolutions).
2. Rectification (i.e. thresholding with non-linearity) to determine if the neuron has to fire.

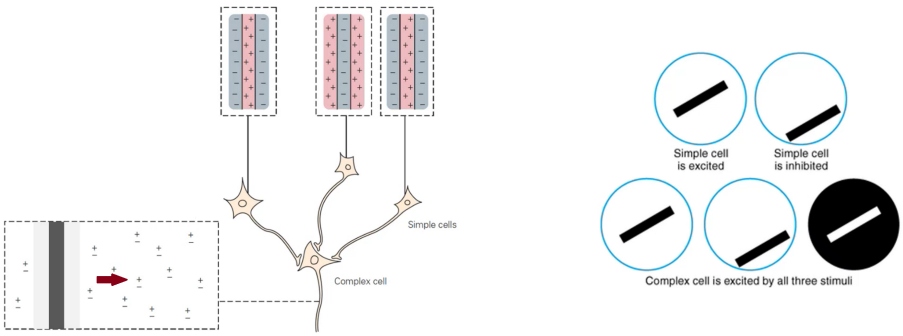


1.4.2 Complex cells

Complex cells

Neurons with a rectangular receptive field larger than simple cells. They respond to linear stimuli with a specific orientation and move in a particular direction (position invariance).

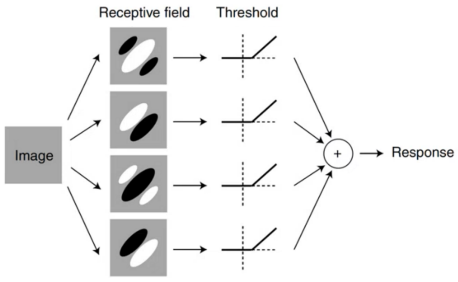
Remark. At this stage, the position of the stimulus is not relevant anymore as the ON and OFF zones of the previous cells are mixed.



Complex cell model The stages of computation in complex cells are:

1. Linear filtering of multiple receptive fields.

2. Rectification for each receptive field.
3. Summation of the response.



1.4.3 End-stopped (hypercomplex) cells

End-stopped cells

Neurons whose receptive field has an excitatory region surrounded by one or more inhibitory regions all with the same preferred orientation.
This type of cell responds to short segments, long curved lines (as the tail of the curve that ends up in the inhibition region is not the preferred orientation) or to angles.

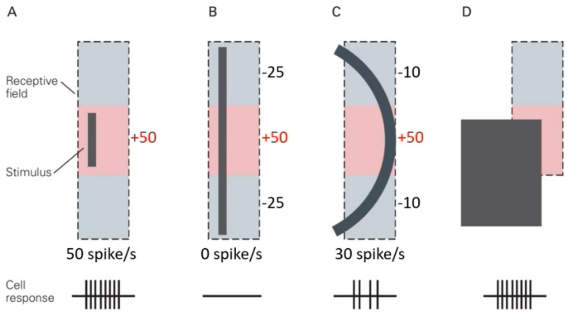
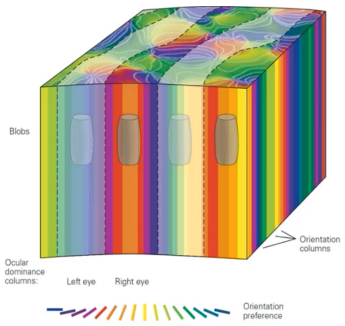


Figure 1.7: End-stopped cell with a vertical preferred orientation

1.4.4 Ice cube model

Each 1 mm of the visual cortex can be modeled through an ice cube module that has all the neurons for decoding all the information (e.g. color, direction, ...) in a specific location of the visual scene (i.e. each cube is a block of filters).



1.5 Extrastriate visual areas

Extrastriate visual areas Areas outside the primary visual cortex (V1). They are responsible for the actual object recognition task.

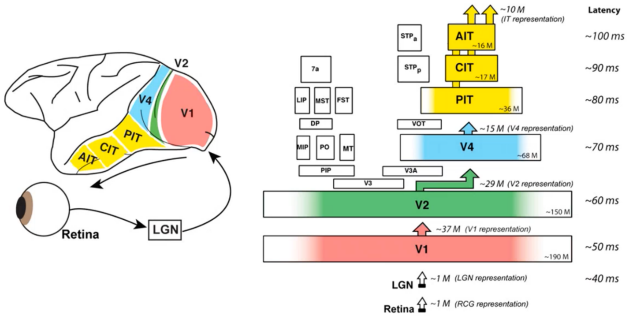


Figure 1.8: Ventral pathway

Visual object Set of visual features (e.g. color, direction, orientation, ...) perceptually grouped into discrete units. Visual object

Visual recognition Ability to assign a verbal label to objects in the visual scene.

Identification Recognize the object at its individual level. Identification

Categorization Recognize the object as part of a more general category. Categorization

Remark. In humans, categorization is easier than identification. Stimuli originating from distinct objects are usually treated as the same on the basis of past experience.

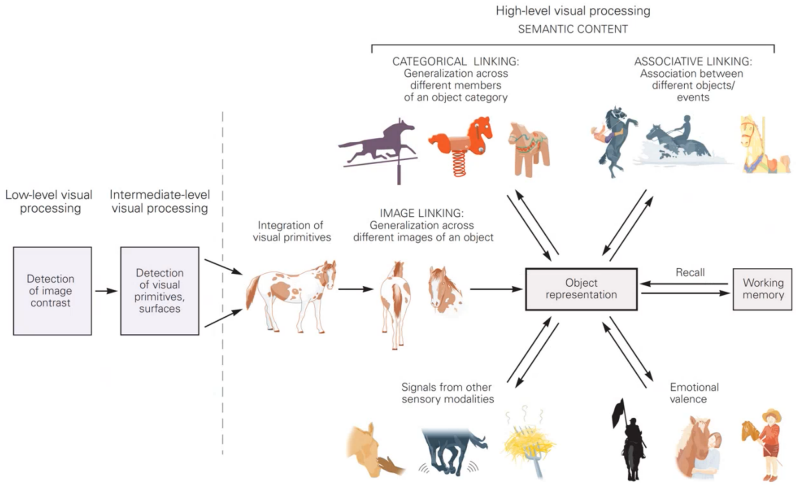


Figure 1.9: Processes that start after an object has been recognized

Remark. For survival reasons, after an object has been recognized, its emotional valence is usually the first thing that is retrieved to determine if the object is dangerous.

Object recognition requires both the following competing properties:

Selectivity Different responses to distinct specific objects. Selectivity

Consistency Similar responses to transformations (e.g. rotation) of the same object (generalization). Consistency

Core object recognition Ability to rapidly (< 200 ms) discriminate a given visual object from all the other possible objects. Core object recognition

Remark. Primates perform this task exceptionally well even if the object is transformed.

Remark. 200 ms is the time required to move the eyes. Experiments on core object recognition don't want candidates to move their eyes. Moreover, it prevents feedback processing from starting.

1.5.1 Area V4

Area V4

Intermediate cortical area responsible for visual object recognition and visual attention. It facilitates figure-ground segmentation of the visual scene enabling both bottom-up and top-down visual processes.

1.5.2 Inferior temporal cortex (IT)

Inferior temporal cortex (IT)

Responsible for object perception and recognition. It is divided into three areas:

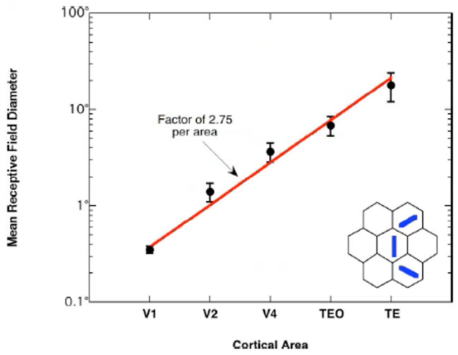
- Anterior IT (AIT).
- Central IT (CIT).
- Posterior IT (PIT).

Remark. It takes approximately 100 ms for the signal to arrive from the retina to the IT.

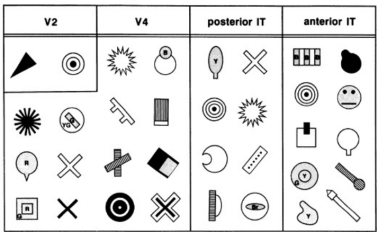
Remark. The number of neurons decreases from the retina to the IT. V1 can be seen as the area that sees everything and decides what to further process.

Remark. Central IT and anterior IT do not show clear retinotopy. Posterior IT shows some sort of pattern.

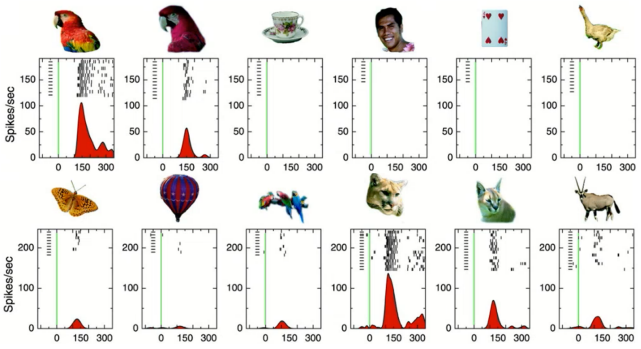
Remark. Receptive field scales by a factor of ~ 3 after passing through each cortical area.



Remark. It is difficult to determine which stimuli trigger the neurons in the IT and what actual stimuli trigger the IT is still unclear.
Generally, neurons in this area respond to complex stimuli, often biologically relevant objects (e.g. faces, hands, animals, ...).

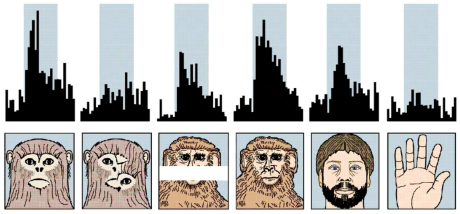


(a) Stimuli that trigger specific neurons of the ventral pathway

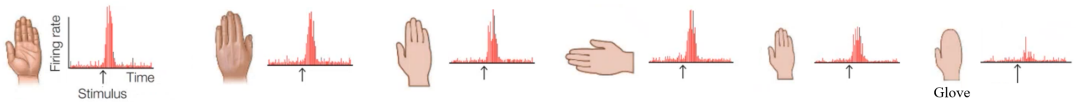


(b) Responses of a specific IT neuron to different stimuli

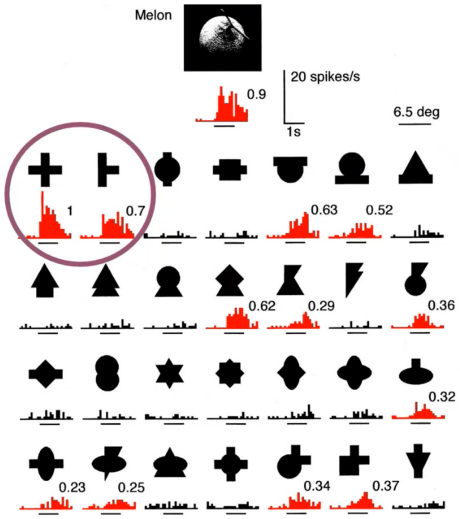
Case study (IT neurons in monkeys)**.** Several researchers observed that a group of IT neurons in monkeys respond selectively to faces. The response is stronger when the full face is visible and gets weaker if it is incomplete or malformed.



It also has been observed that an IT neuron responds to hands presented at various perspectives and orientations. A decrease in response is visible when the hand gets smaller and it is clearly visible when a glove is presented.



Case study (IT neuron response to a melon)**.** An IT neuron responds to a complex image of a melon. However, it has been shown that it also responds to simpler stimuli that represent the visual elements of the melon.

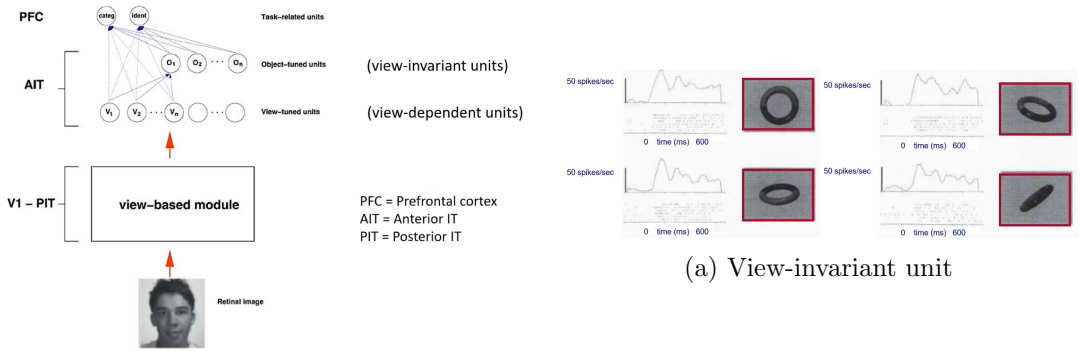


View-dependent unit The majority of IT neurons are view-dependent and respond only to objects at specific points of view.

View-dependent unit

View-invariant unit 10% of IT neurons are view-invariant and respond regardless of the position of the observer.

View-invariant unit

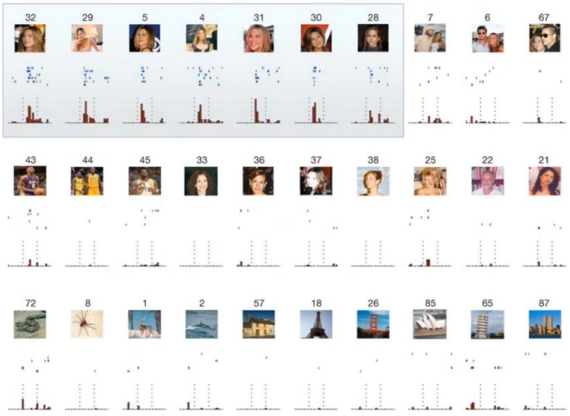


(a) View-invariant unit

Gnostic unit Neuron in the object detection hierarchy that gets activated by complex stimuli (i.e. objects with a meaning).

Gnostic unit

Case study (Jennifer Aniston cell)**.** An IT neuron of a human patient only responded to pictures of Jennifer Aniston or to its written name.



1.5.3 Local vs distributed coding

Local coding hypothesis IT neurons are gnostic units that are activated only when a particular object is recognized.

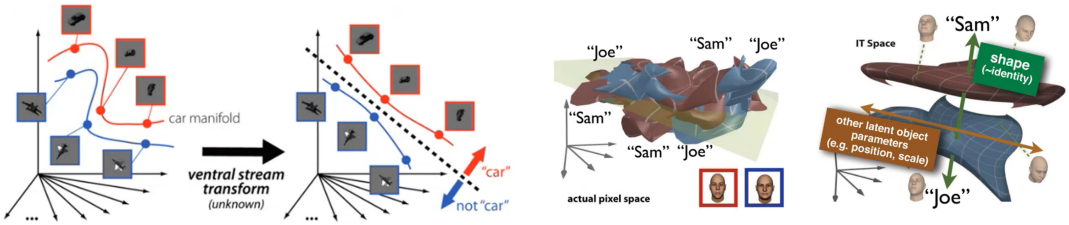
Local coding hypothesis

Distributed coding hypothesis Recognition is due to the activation of multiple IT neurons.

Distributed coding hypothesis

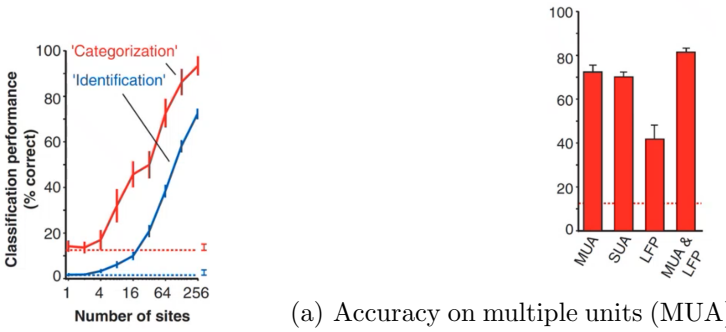
Remark. This is the most plausible hypothesis.

Case study (Neurons in vector space)**.** The response of a population of neurons can be represented in a vector space. It is expected that transformations of the same object produce representations that lie on the same manifold.
In the first stages of vision processing, various manifolds are tangled. Object recognition through the visual cortex aims to untangle the representations of the objects.



Case study (Classifier from monkey neurons [1])**.** An animal maintains fixation at the center of a screen on which images of different categories are presented very quickly (100 ms + 100 ms pause) at different scales and positions.
The responses of IT neurons are taken with some offset after the stimulus (to give them time to reach the IT) and converted into vector form to train one binary classifier (SVM) for each category (one-vs-all).
Once trained, testing was done on new stimuli. Results show that the performance increases linearly with the logarithm of the number of sites (measured neurons). It can be concluded that:

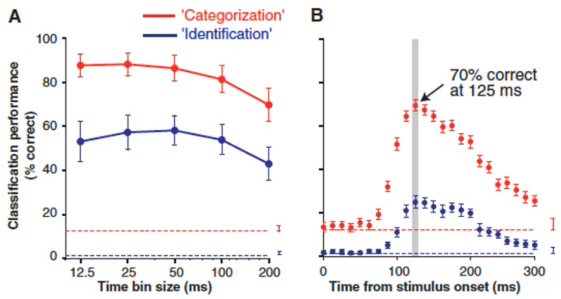
- Categorization is easier than identification.
- The distributed coding hypothesis is more likely.



(a) Accuracy on multiple units (MUA), a single unit (SUA) and readings made on the cortex, not inside (LFP)

Time-wise, it has been observed that:

- Performance gets worse if the measurement of the neurons spans for too long (no explanation was given in the original paper, probably noise is added up to the signal for longer measurements).
- The best offset from the stimulus onset at which the measures of the IT neurons should be taken is 125 ms.



It has also been observed that the visual ventral pathway, which is responsible for object recognition, also encodes information on the size of the objects. This is not strictly useful for recognition, but a machine learning algorithm is able to extract this information from the neural readings. This hints at the fact that the ventral pathway also contributes to identifying the location and size of the objects.

Case study (Artificial neural network to predict neuronal activity [6])**.** Different neural networks are independently trained on the task of image recognition. Then, the resulting networks are compared to the neuronal activity of the brain.
The network should have the following properties:

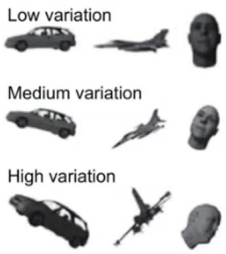
- Provide information useful to support behavioral tasks (i.e. act as the IT neurons).
- Layers of the network should have a corresponding area on the ventral pathway (mappable).
- It should be able to predict the activation of single and groups of biological neurons (neurally predictive).

Dataset A set of images is divided into two sets:

Train set To train the neural networks.

Test set To collect neuronal data and evaluate the neural networks.

Images have different levels of difficulty (low to high variation) and are presented on random backgrounds.

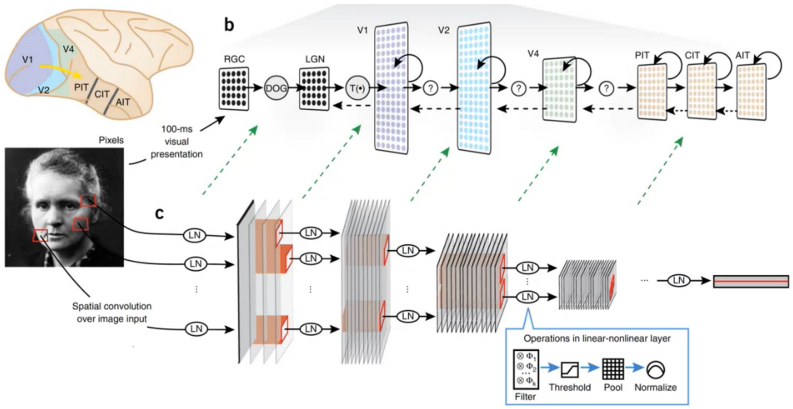


Neuronal data Neuronal data are collected from the area V4 and IT in two macaque monkeys. They are tasked to maintain fixation at the center of a screen on which images are presented for 100 ms followed by a 100 ms blank screen.

For each stimulus, the firing rate is obtained as the average of the number of spikes in the interval 70 ms - 170 ms after the stimulus onset.

Neural network training Hierarchical convolutional neural networks (HCNN) are used for the experiments. They are composed of linear-nonlinear layers that do the following:

1. Filtering through linear operations of the input stimulus (i.e. convolutions).
2. Activation through a rectified linear threshold or sigmoid.
3. Mean or maximum pooling as nonlinear aggregation operation.
4. Divisive normalization to output a standard range.



The HCNNs have a depth of 3 or fewer layers and are trained independently from the neuronal measurements. For evaluation, models are divided into groups following three different criteria:

- Random sampling.
- Selection of models with the highest performance on the high-variation images.
- Selection of models with the highest IT neural predictivity.

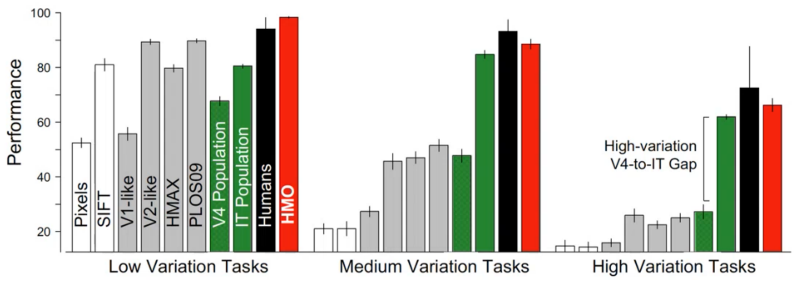
Resulting HCNNs are also used to create a new high-performance architecture through hierarchical modular optimization (HMO) by selecting the best-performing modules from the trained networks (as each layer is modular).

Evaluation method Evaluation is done using the following approach:

- Object recognition performances are assessed using SVM classifiers:
  - For neural networks, the output features of a stimulus are obtained from the activations at the top layers.
  - For neuronal readings, the output features of a stimulus are obtained by converting the firing rates into vector form.
- To measure the ability of a neural network to predict the activity of a neuron, a partial least squares regression model is used to find a combination of weights at the top layers of the network that best fits, using as metric the coefficient of determination (R^2), the activity of the neuron on a random subset of test images.
- An ideal observer is used as a baseline. It has all the categorical information to make correct predictions but it does not use a layered approach.

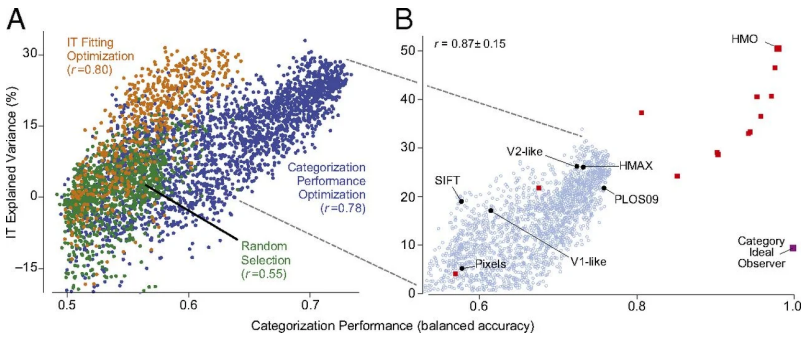
Results It has been observed that:

- The HMO model has human-like performances.

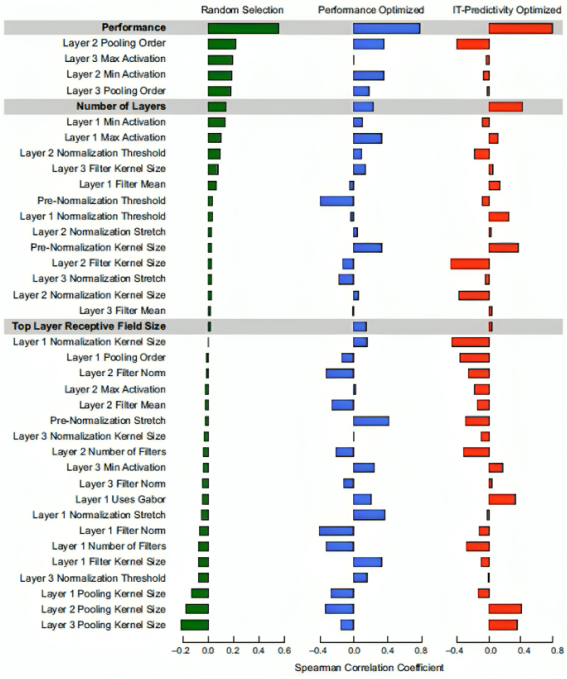


- The higher the categorization accuracy, the better the model can explain the IT.

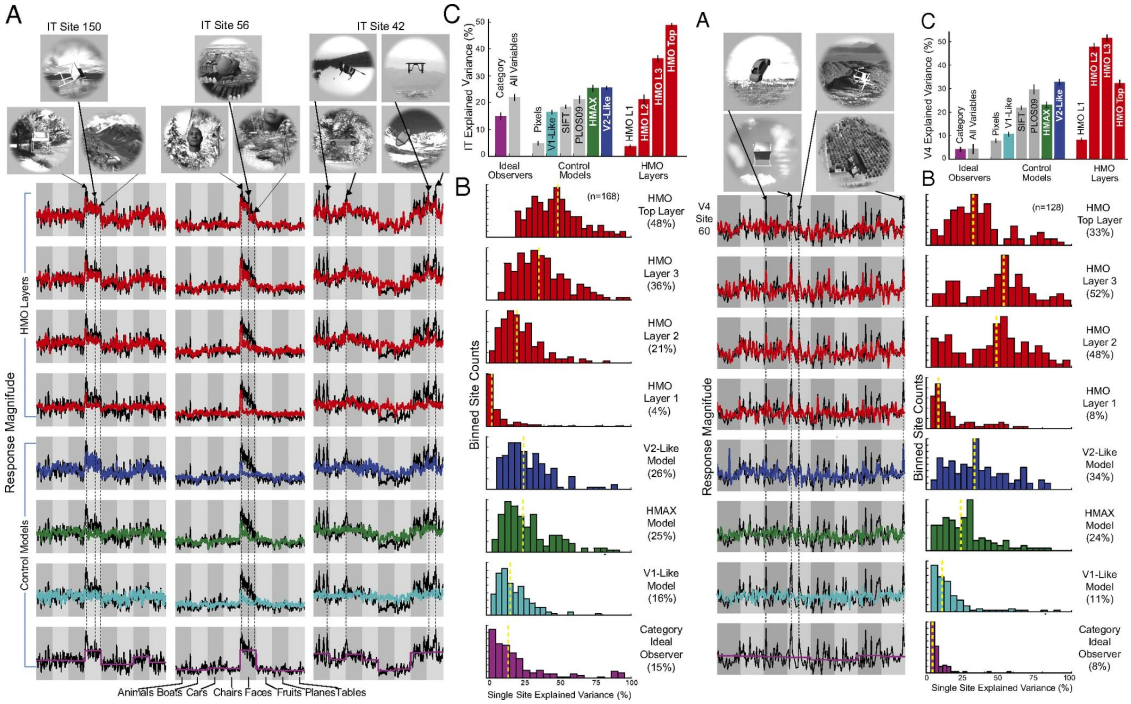
  Moreover, forcefully fitting a network to predict IT as the main task predicts the neuronal activity worse than using a model with high categorization accuracy.



- None of the parameters of the neural networks can independently predict the IT better than performance (i.e. the network as a whole).



- Higher levels of the HMO model yield good prediction capabilities of IT and V4 neurons. More specifically:
  - The fourth (last) layer of the HMO model predicts well the IT.
  - The third layer of the HMO model predicts well the area V4.



(a) Results on IT

(b) Results on V4

Figure 1.14: (A) Actual neuronal activity (black) and predicted activity (colored). (B) R^2 value over the population of single IT neurons. (C) Median R^2 over the population of IT neurons.

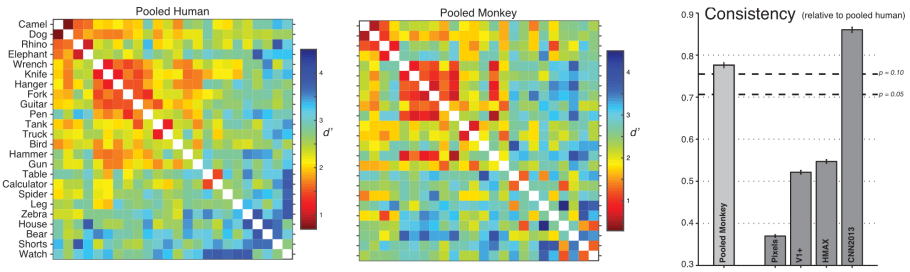
2 Object recognition emulation through neural networks

2.1 Convolutional neural networks

Deep convolutional neural networks (DCNNs) show an internal feature representation similar to the representation of the ventral pathway (primate ventral visual stream). Moreover, object confusion in DCNNs is similar to the behavioral patterns in primates.
However, on a higher resolution level (i.e. not object but image level), the performance of DCNNs diverges drastically from human behavior.

Remark. Studies using HCNN have also been presented in the previous chapter.

Case study (Humans and monkeys object confusion [3])**.** It has been seen that monkeys show a confusion pattern correlated to that of humans on the task of object recognition. Convolutional neural networks also show this correlation while low-level visual representations (V1 or pixels, a baseline computed from the pixels of the image) correlate poorly.



Case study (Primates and DCNNs object recognition divergence [4])**.** Humans, monkeys and DCNNs are trained for the task of object recognition.
To enforce an invariance recognition behavior, each image has an object with a random transformation (position, rotation, size) and has a random natural background.

- For humans, a trial starts with fixation. Then, an image is displayed for 100 ms followed by a binary choice. The human has to make its choice in 1000 ms.
- For monkeys, a trial starts with fixation. Then, an image is displayed for 100 ms followed by a binary choice. The monkey has up to 1500 ms to freely view the response images and has to maintain fixation on its choice for 700 ms.
- DCNNs are trained as usual.

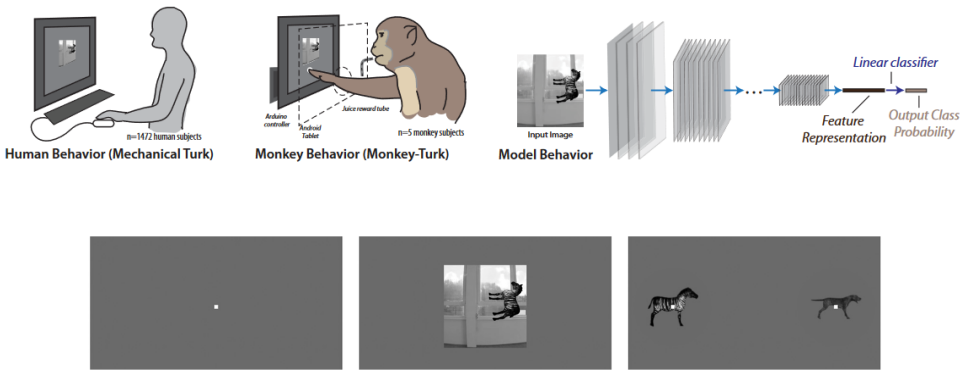


Figure 2.1: Steps of a trial

Performance is measured using behavioral metrics. Results show that:

Object-level Object-level measurements are obtained as an average across all images of that object.

Recognition confusion of primates and DCNNs are mostly correlated.

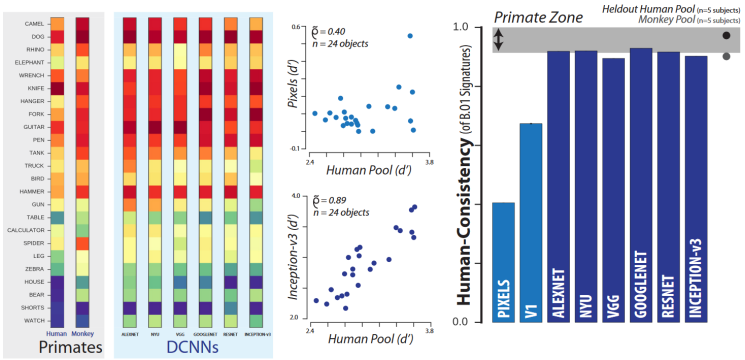
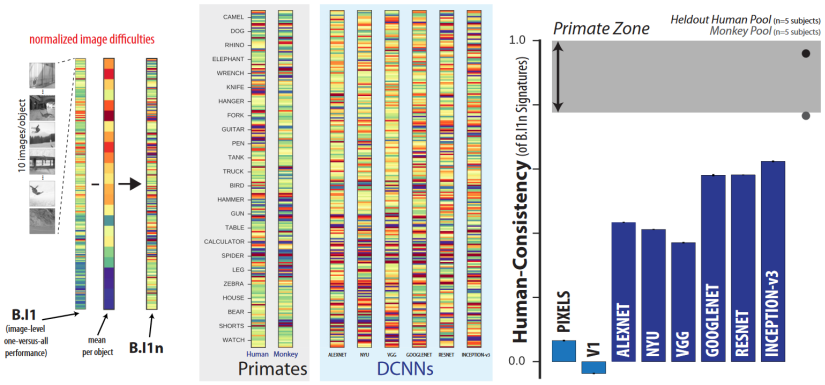


Figure 2.2: Object-level results. In the first part, warmer colors indicate a better classification.

Image-level Image-level measurements are obtained by normalizing the raw classification results.

All DCNNs fail to replicate the behavioral signatures of primates. This hints at the fact that the architecture and/or the training process is limiting the capability of the models.



2.2 Recurrent neural networks

2.2.1 Object recognition

The short duration for which candidates of the previous experiments were exposed to an image suggests that recurrent computation is not relevant for core object recognition. However, the following points are in contrast with this hypothesis:

- DCNNs fail to predict primate behavior in many cases.
- Specific image instances (e.g. blurred, cluttered, occluded) are easy for primates but difficult for DCNNs.

This hints at the fact that recurrent computation might be involved, maybe at later stages of the recognition process.

Case study (Primates recognition reaction time [2])**.**

Recognition training and evaluation Humans, macaques and DCNNs are trained for the task of object recognition on images with two levels of difficulty:

Control images Easier to recognize.

Challenge images Harder to recognize.

Results show that primates outperform DCNNs on challenge images.

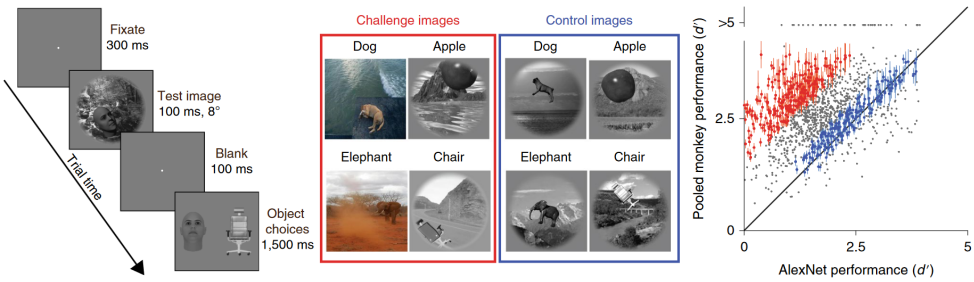
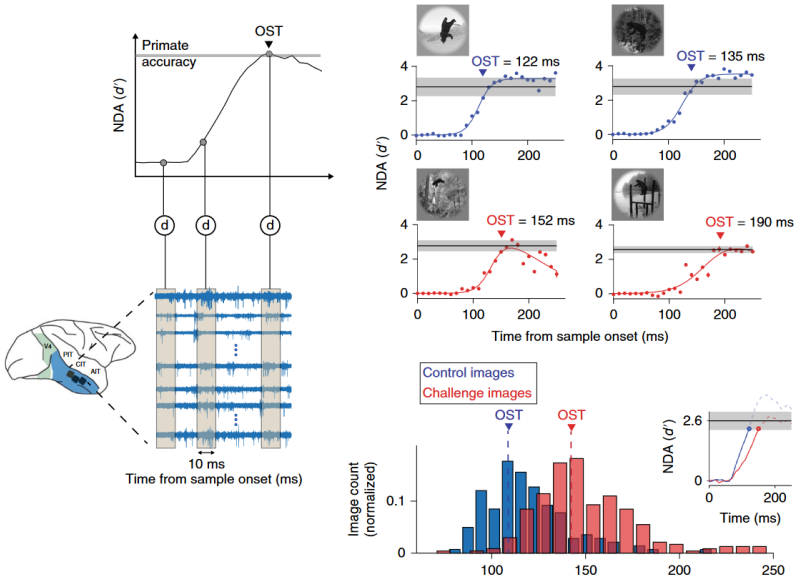


Figure 2.3: Trial steps, example images and behavioral comparison between monkeys and DCNNs. Red and blue points in the graph are challenge and control images, respectively.

Reaction time It also has been observed that the reaction time of both humans and monkeys for challenge images is significantly higher than the reaction for control images (ΔRT = 11.9 ms for monkeys and ΔRT = 25 ms for humans).

To determine the time at which the identity of an object is formed in the IT cortex, the neural activity is measured every 10 ms after the stimulus onset and a linear classifier (decoder) is trained to determine the **neural decode accuracy (NDA)** (i.e. the best accuracy that the classifier can achieve with the information in that time slice). We refer with **object solution time (OST)** the time at which the NDA reached the primate accuracy (i.e. high enough).

It has been observed that challenge images have a slightly higher OST ($\sim$ 30 ms) whether the animal was actively performing the task or passively viewing the image.

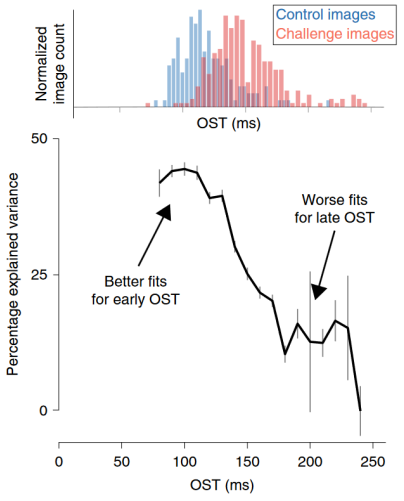


DCNN IT prediction The IT neuronal response for a subset of challenge and control images has been measured across 10 ms bins to obtain two sets R^{train} and R^{test} (50/50).

During training, the activation F^{train} of a layer of the DCNN is used to predict R^{train} through partial least square regression (i.e. a linear combination of F^{train}).

During testing, the activation of the same layer of the DCNN is transformed using the found parameters and compared to R^{test}.

Results show a higher predictivity for early responses (which are mainly feed-forward) and a significant drop over time. The drop coincides with the OST of challenge images, hinting at the fact that later phases of the IT might involve recurrence.



CORnet IT prediction The previous experiment has also been done using deeper CNNs that showed better predictivity. This can be explained by the fact that deeper networks simulate the unrolling of a recurrent network and are therefore an approximation of them.

Deeper networks are also able to solve some of the challenge images but those that remained unsolved are those with the longest OSTs among the challenge images.

CORnet, a four-layer recurrent neural network, has also been experimented. Results show that the first layers of CORnet are good predictors of the early IT phases while the last layers are good at predicting the late phases of IT.

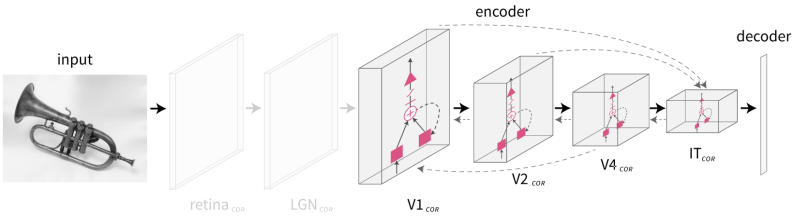
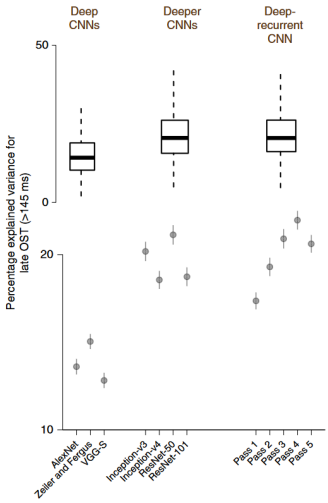
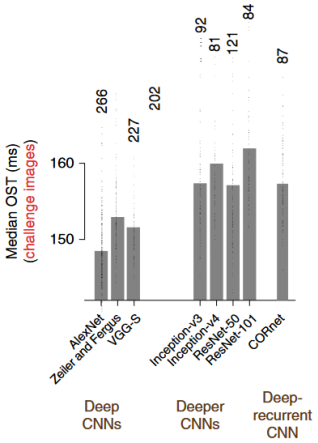


Figure 2.4: Architecture of CORnet



(a) Predictivity for deep, deeper and recurrent CNNs



(b) Number (on top) and median OST (bars) of the unsolved images for each model

Remark. Recurrence can be seen as additional non-linear transformations in addition to those of the feed-forward phase.

2.2.2 Visual pattern completion

Pattern completion Ability to recognize poorly visible or occluded objects. Pattern completion

Remark. The visual system is able to infer an object even if only 10-20% of it is visible.

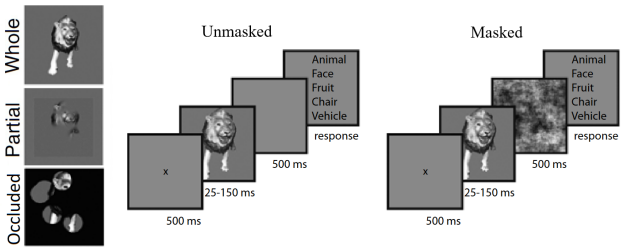
It is hypothesized that recurrent computation is involved.

Case study (Human and RNN pattern completion [5])**.**

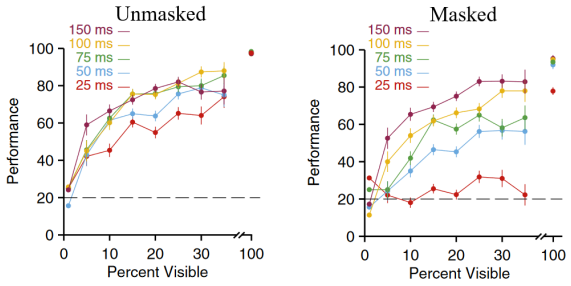
Trial structure Whole and partial images are presented to humans through two types of trials:

Unmasked After fixation, an image is displayed for a short time followed by a blank screen. Then, a response is required from the candidate.

Backward masking After fixation, an image is displayed for a short time followed by another image. Then, a response is required from the candidate. The second image aims to interrupt the processing of the first one (i.e. interrupt recurrent processing).



Human results Results show that subjects are able to robustly recognize whole and partial objects in the unmasked case. In the masked case, performances are instead worse.



Moreover, measurements show that the neural response to partially visible objects is delayed compared to whole images, hinting at the fact that additional computation is needed.

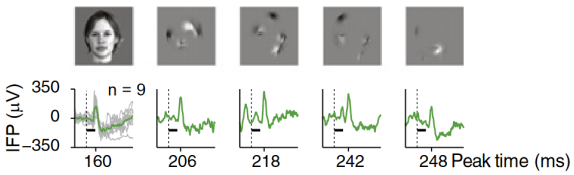
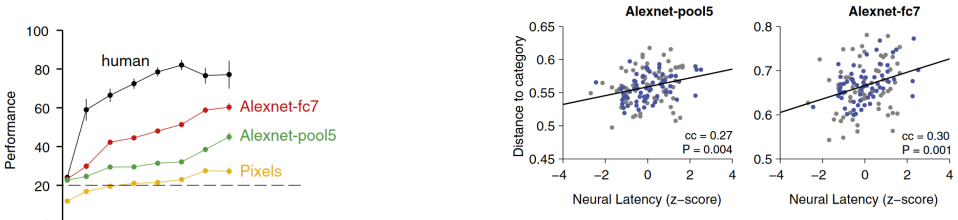


Figure 2.6: Activity (IFP) of a neuron that responds to faces

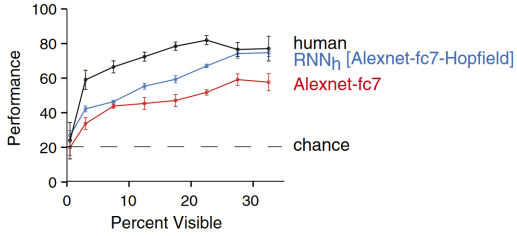
CNN results Feed-forward CNNs have also been trained on the task of object recognition.

- Performances are comparable to humans for whole images but decline for partial images.
- There is a slight correlation between the latency of humans' neural response and the distance of the internal representation in the CNNs of each partial object to its whole image.

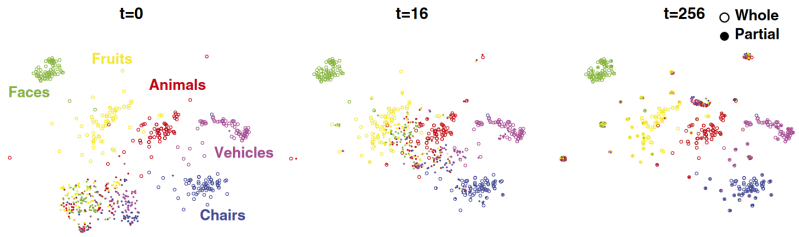


(a) Representation and latency correlation. The color of the dots depends on the electrode that measured the latency.

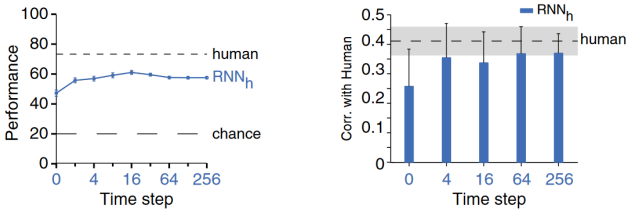
RNN results Recurrent neural networks have also been tested by using existing CNNs enhanced through attractor networks[1] (Hopfield network, RNNh). Results show that RNNh has higher performance in pattern completion.



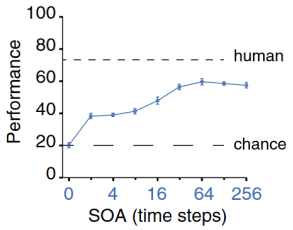
Moreover, by plotting the temporal evolution of the internal representation of partial objects, it can be seen that, at the beginning, partial images are more similar among themselves than their corresponding attractor point, but, over time, their representation approaches the correct cluster.



Time-wise, RNNh performance and correlation with humans increase over the time steps and saturates at around 10-20 steps. This is consistent with the physiological delays of the human ventral visual stream.



By backward masking the input of the RNNh (i.e. present the image for a few time steps and then change it), performance drops from $58 \pm 2\%$ to $37 \pm 2\%$.



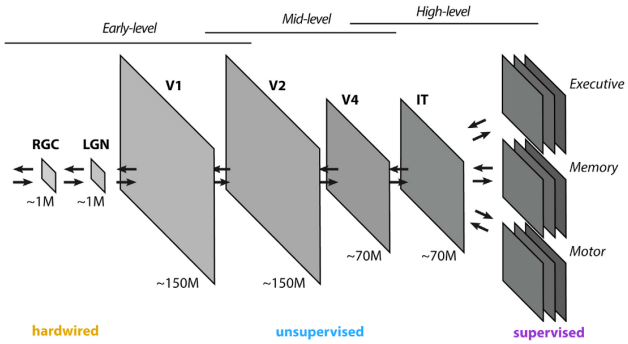
[1]Recurrent network with multiple attractor points each representing a whole image. By processing the same partial image for multiple time steps, its representation should converge to an attractor point.

2.3 Unsupervised neural networks

Most of the models to simulate the visual cortex are trained on supervised datasets of millions of images. Such supervision is not able to explain how primates learn to recognize objects as processing a huge amount of category labels during development is highly improbable. Possible hypotheses are:

- Humans might rely on different inductive biases for a more efficient learning.
- Humans might augment their initial dataset by combining known instances.

Unsupervised learning might explain what happens in between the representations at low-level visual areas (i.e. the retina), which are mostly hardcoded from evolution, and the representations learned at higher levels.



Case study (Unsupervised embedding [7])**.** Different unsupervised embedding methods are used to create a representation for a dataset of images that are then assessed on various tasks.

Contrastive embedding Unsupervised embedding method that uses a DCNN (which simulates low-level visual areas) to create the representation of an image in a low dimensional space and then optimize it by pushing each embedding closer to its close neighbors and far from its background neighbors.

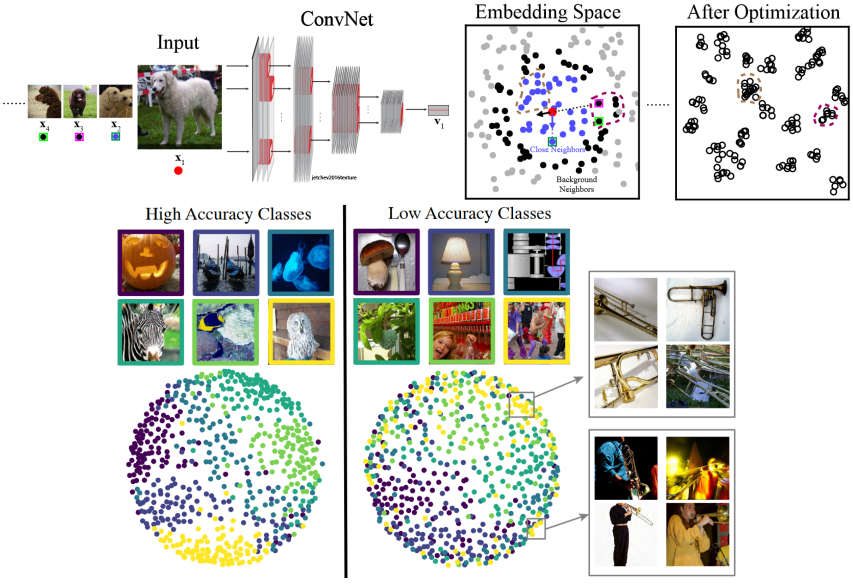


Figure 2.8: Workflow and visualization of the local aggregation algorithm

Results on object recognition tasks To solve the tasks, unsupervised embeddings are used in conjunction with a linear classifier. A supervised DCNN is also used as a baseline.

Results show that:

- Among all the unsupervised methods, contrastive embeddings have the best performances.
- Unsupervised methods equaled or outperformed the DCNN on tasks such as object position and size estimation.
- The DCNN outperforms unsupervised models on categorization tasks.

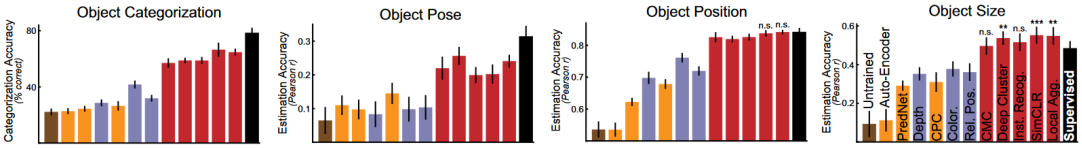


Figure 2.9: Evaluation accuracy of an untrained model (brown), predictive encoding methods (orange), self-supervised methods (blue), contrastive embeddings (red) and a supervised DCNN (black).

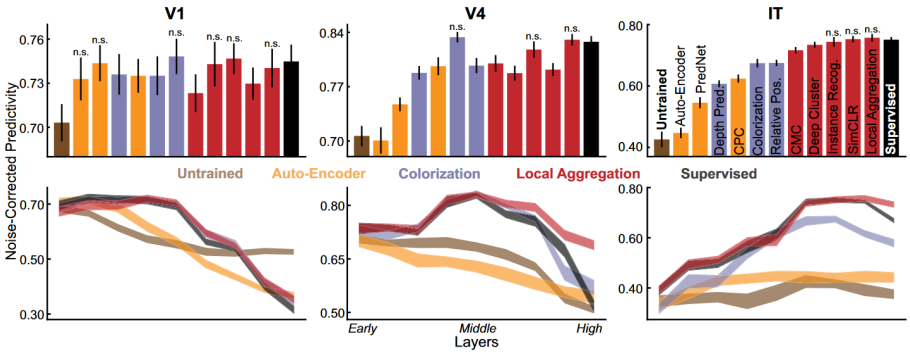
Results on neural data Techniques to map the responses of an artificial network to real neural responses have been used to evaluate unsupervised methods.

Results show that:

Area V1 None of the unsupervised methods are statistically better than the DCNN.

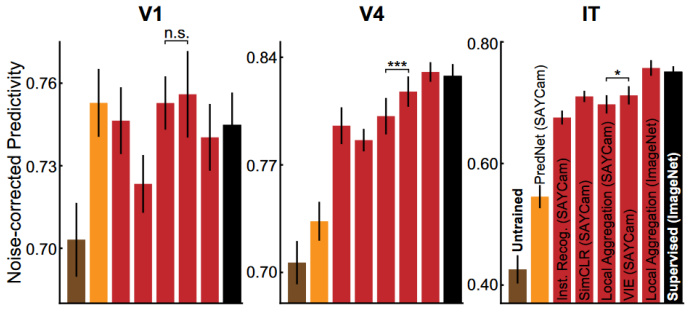
Area V4 A subset of methods equaled the DCNN.

Area IT Only contrastive embeddings equaled the DCNN.



Results on video data As training on single distinct images (ImageNet) is significantly different from real biological data streams, a dataset containing videos (SAYCam) has been experimented with. A contrastive embedding, the VIE algorithm, has been employed to predict neural activity.

Results show that embeddings learned from videos are comparable to those learned from only images.



Semi-supervised learning Semi-supervised embedding aims to find a representation using a small subset of labeled data points and a large amount of unlabeled data.

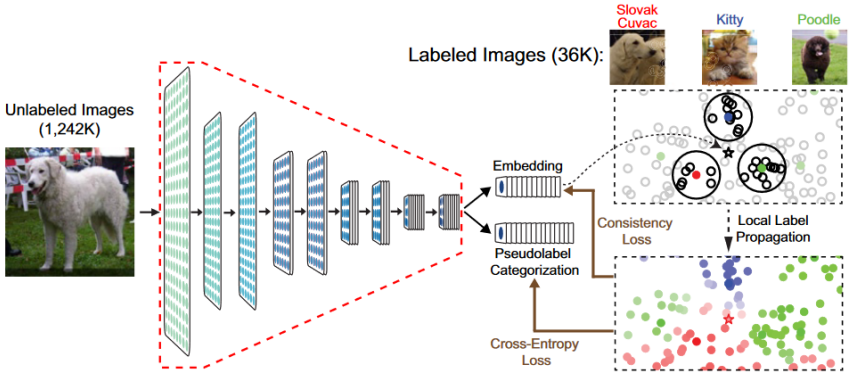
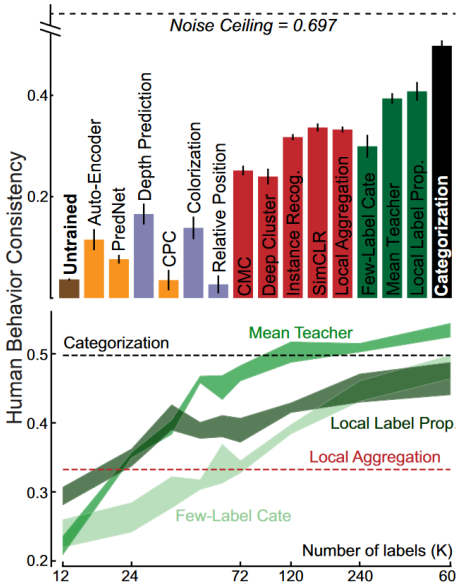


Figure 2.10: Workflow of the local label propagation algorithm

Results show that semi-supervised embeddings with only a 3% of supervision are substantially more consistent than purely unsupervised methods. Although, the gap between them and the DCNN still remains.

Nevertheless, a significant gap is also present between the results of all the models and the noise ceiling of the data, indicating that there still are inconsistencies between artificial networks and the human visual system.



Bibliography

[1] Chou P. Hung et al. "Fast Readout of Object Identity from Macaque Inferior Temporal Cortex". In: *Science* 310.5749 (2005), pp. 863–866. DOI: 10.1126/science.1117593.

[2] Kohitij Kar et al. "Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior". In: *Nature Neuroscience* 22.6 (2019), pp. 974–983. DOI: 10.1038/s41593-019-0392-5.

[3] Rishi Rajalingham, Kailyn Schmidt, and James J. DiCarlo. "Comparison of Object Recognition Behavior in Human and Monkey". In: *Journal of Neuroscience* 35.35 (2015), pp. 12127–12136. DOI: 10.1523/JNEUROSCI.0573-15.2015.

[4] Rishi Rajalingham et al. "Large-Scale, High-Resolution Comparison of the Core Visual Object Recognition Behavior of Humans, Monkeys, and State-of-the-Art Deep Artificial Neural Networks". In: *Journal of Neuroscience* 38.33 (2018), pp. 7255–7269. DOI: 10.1523/JNEUROSCI.0388-18.2018.

[5] Hanlin Tang et al. "Recurrent computations for visual pattern completion". In: *Proceedings of the National Academy of Sciences* 115.35 (2018), pp. 8835–8840. DOI: 10.1073/pnas.1719397115.

[6] Daniel L. K. Yamins et al. "Performance-optimized hierarchical models predict neural responses in higher visual cortex". In: *Proceedings of the National Academy of Sciences* 111.23 (2014), pp. 8619–8624. DOI: 10.1073/pnas.1403112111.

[7] Chengxu Zhuang et al. "Unsupervised neural network models of the ventral visual stream". In: *Proceedings of the National Academy of Sciences* 118.3 (2021), e2014196118. DOI: 10.1073/pnas.2014196118.