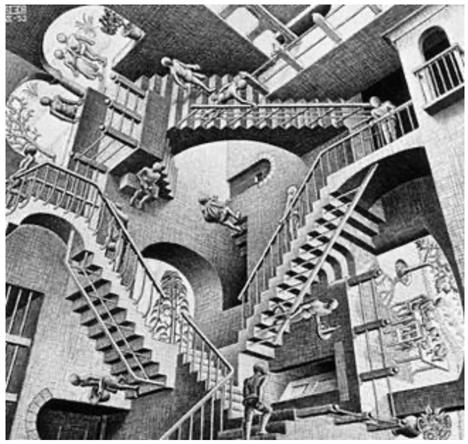
Recognizing objects: Part II



Trying to understand perception by studying only neurons is like trying to understand bird flight by studying only feathers; it just cannot be done. To understand bird flight, you need to understand aerodynamics, only then can one make sense of the structure of feathers and the shape of wings. Similarly, you can't reach an understanding of why neurons in the visual system behave the way they do, just by studying their anatomy and physiology.

David Marr

Map of the next few classes

- Today: Recognizing Objects part 2 + discussion
  - Readings: Bowers et al., 2023, *Neural Networks Made Easy*
- Tuesday, Jan 28: What is so Special About Faces?
  - Reading: Harada et al., 2020: first experimental paper
  - **Tuesday Jan 28: Deadline** to submit your first MC exam question for 1% of your grade (don't forget to indicate correct answer)
- Thursday, Jan 30: Review
- Tuesday, Feb 4: Midterm 1

- Appreciate our amazing capacity for object recognition
- Describe the roles of two visual cortex systems in object recognition
- Discuss the contribution of classic fMRI research to our understanding of object processing stages in humans
- Compare challenges posed to human and computer vision

As a reminder, here are the learning objectives for the lectures on object recognitionLast lecture we covered the first two objectives. Let’s review key information from the last lecture and finish building a conceptual foundation for moving on.

Review question

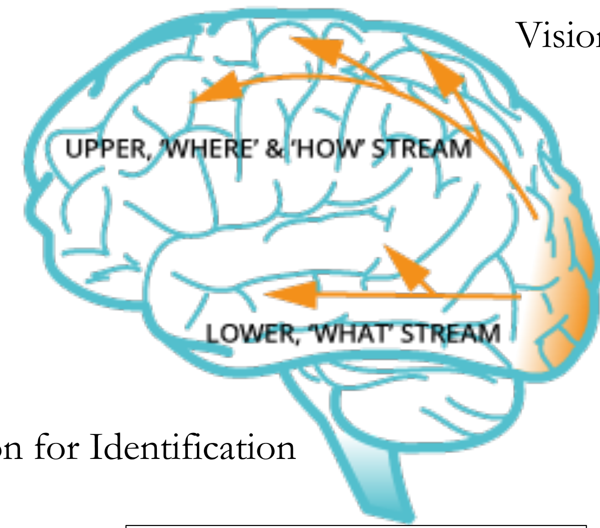
What Information does Prof Barnes NOT say V1 cells outside of lower layer IV are sensitive to?

a. Orientation
b. Length
c. Width
d. Motion Direction
e. None of the above

e. They are sensitive to all of those things.

Review: Dorsal and Ventral Streams

Vision for Guided Action

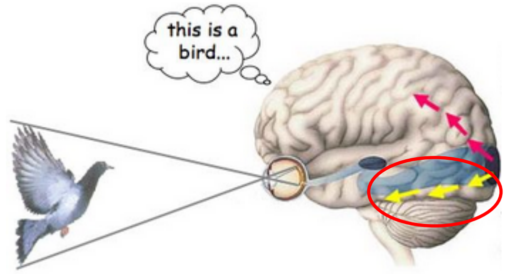


Vision for Identification

Last lecture we talked about the impoverished information that arrives in V1 from the retina – the upside down, distorted retinotopic map that is sensitive to all those features listed in the last slide. And we introduced the dorsal and ventral visual streams: Visual information is passed from the retina to the primary visual cortex. In two feedforward sweeps, it then travels to the parietal lobe through the dorsal stream, which gives us vision for directed action. And at the same time it travels into the temporal lobe through the ventral stream giving us vision for identification. Because we're focusing on object perception we're going to talk more about the ventral stream from this point in the lecture on.

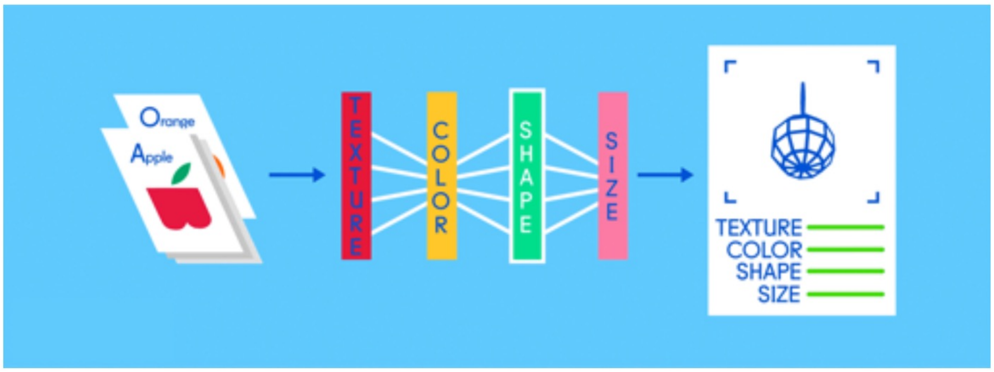
How do we get high-level object representations?

- In temporal lobe regions further along the ventral visual stream ("downstream")
- Neuron receptive fields the whole size of the visual field
- Sensitive to image category



How do we get high level representations from the impoverished information V1 represents? How do we know we're looking at a bird that is partly hidden by a branch and about 20 feet away from us? 1) In high level areas of the ventral stream, visual signal also gets integrated with all kinds of semantic knowledge you have about the meaning of the object that comes from a whole lifetime of experiences with birds, as well as all the things you've learned and read about them… To review, V1 receives an upside down and distorted 2D map of the scene that you're looking at. And each neuron's receptive field- what it responds to, is just a tiny point on the map of the visual world that's transmitted from the retina. 1) But by the time you get to the end of the dorsal stream a 3) neuron's receptive field can take in whole objects and more.2) Beyond that, neurons are sensitive to what category of object you are looking at. **Q: ANY QUESTIONS ABOUT THIS?**

CNNs mimic ventral stream processes



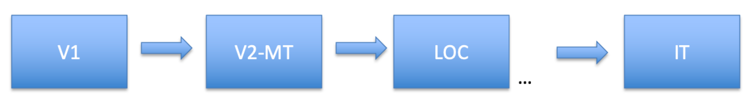
8

Questions? So as it happens, some convolutional neural networks used in machine visionare
actually are designed to mimic what we know about ventral stream processes. You'll
recognize this image from the Neural Networks Made Easy article assigned for this
class. I quote: "Here, we feed the neural network vast amounts of training data,
labeled by humans so that a neural network can essentially fact-check itself as it's
learning. As pictures are fed in, the network breaks them down into their most basic
components, i.e. edges, textures and shapes so each "filter" acts like specific types of
neuron in primary visual cortex. As the picture propagates through the network,
these basic components are combined to form more abstract concepts, i.e. curves
and different colors which, when combined further, start to look like a stem, an

entire
orange, or both green and red apples." End quote. So this is very much the
kind of
thing the ventral stream is doing and as the Blog on How AI helps us
understand
human vision,, which we read for last class points out, architects of CNNs
have built
effective ones based on understanding of how the ventral stream works.

- “Feedforward” because information gets passed forward from one area to the next from V1 to inferotemporal cortex (IT).
- ”Bottom-up” because information moves in stages from earlier to later and from simple and concrete to complex and more abstract
- “Feature-based” because the whole object is thought to be
- pieced together out of its elements or *features* (colour, orientation, motion)

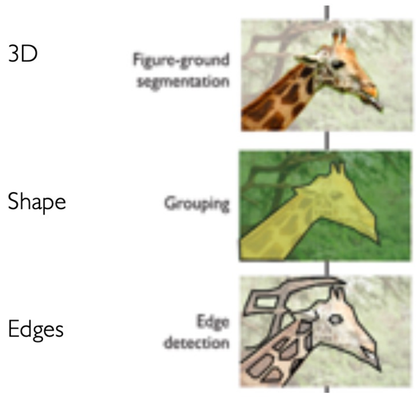
VENTRAL STREAM



1) At the bottom here the boxes represent stages of the ventral stream, starting at V1, or the bottom, and ending up in inferotemporal cortex, or IT, at the top. 2) We call this process a "bottom-up" process, because ...Just as in a CNN, part of what the human ventral visual cortex does is move information in stages to represent increasingly complete and abstract elements of the visual world. Bottom – the bottom is v1 – or even the retina. Information in V1 is the foundation. The building blocks of the more complex elements in higher order cortx.. 3) Remember in the last lecture we talked about information being relayed and I introduced you to a bucket brigade? This is the feedforward sweep. IT is the end of the line. So to summarize, information moves along the Ventral Stream from V1 (primary visual cortex) to the next area of visual cortex (V2), to lateral occipital cortex, which you'll hear

more
about soon, being passed from area to area along until it gets to IT and beyond.

Feature Detection Models of Human Vision: Mapping the steps from fragment to whole



- Information from V1 is passed forward along the ventral stream in a *feedforward sweep*
- Bits glued together at increasing levels of abstraction
- Object matched to amental template

1/23/25 PSYC 365: Class 5 10

I now want to introduce the idea of A FEATURE DETECTION MODEL? Take note as this is going to be important later on. These are models of human vision that map quite well onto what these neural networks are doing. IT's A MODEL of human visual processing that describes object recognition as the result of a series of steps from fragment to whole. Feature detection. means that we detect features of an object and glue them together into a whole. So in a feature detection model, 1) -- we just covered that. 2) as it moves along bits are glued together at increasing levels of abstraction. So for example, first edges may be detected as in the bottom rectangle. That allows grouping into the outline of objects. Then foreground figures or objects can be segmented out from the background... Finally, 3) the object is matched

to a
mental template, in this case of a giraffe. With this, we now have the background to
move on to our third learning objective in the next video.

Learning Objectives (object processing cont)

- Appreciate our amazing capacity for object recognition
- Describe the roles of two visual cortex systems in object recognition
- **Discuss the contribution of classic and more recent fMRI research to our understanding of object processing stages in humans**
- Compare challenges posed to human and computer vision

Now we're ready for Objective #3.

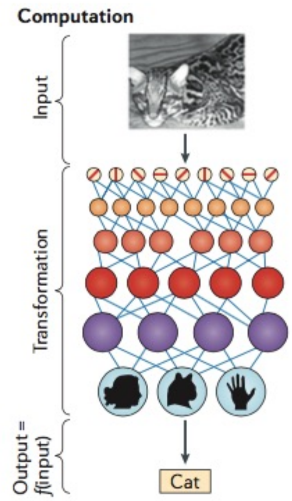
Encoding Approach: Mapping the ventral stream with fMRI

Kalanit Grill-Spector

For this, we’re going to go into some detail on a couple of studies by Kalanit Grill-Spector that are described briefly in Passingham tChapter 2. This is a great example of an encoding approach resulting in a functional brain map of how ventral visual stream in humans.

Mapping the Brain: What we knew in 1998

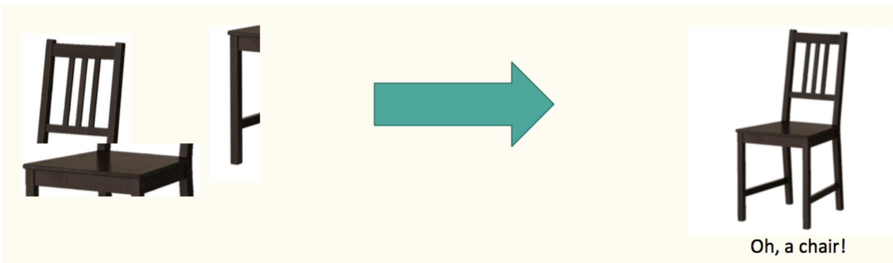
- 2D upside down retinotopic map to category selective regions
- Increase in receptive field size until sensitive to whole objects
- What are the steps in between?



1) So at the time Dr. Grill-Spector did the studies described in the textobook, we knew about the retinotopic maps, and that V1 neurons did points and edges, as in the top row of dots. 2) We knew that neurons increased in their receptive field size as we moved forward in the visual stream as illustrated in these middle rows. And and we knew that areas toward the end of the ventral stream were sensitive to whole objects and even categories of obects as shown at the bottom. 3)But what wasn't known back in 1998 was what happened in between. Again this is very similar to problems facing us with neural network models. We know what we put in and the categories they are able to detect at the output layers, but we actually don't know what's going on in all the "hidden layers" in between."

Grill-Spector et al., 1998

- In what stages does the visual cortex move from responding from parts to wholes?
- What are the stages in between?
- LO or LOC – Lateral Occipital Complex

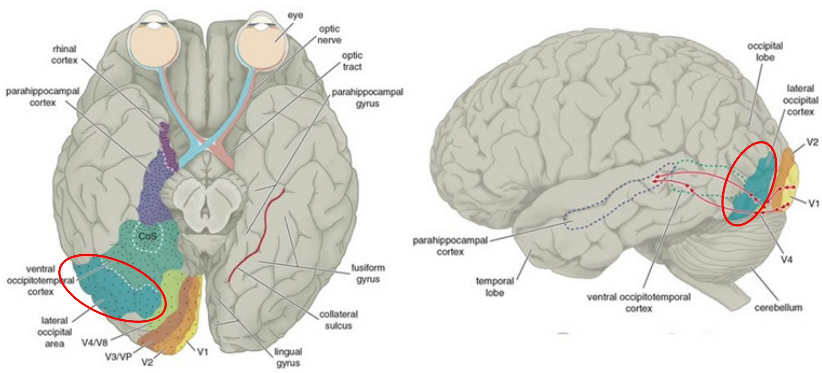


So what Grill-Spector and her colleagues wanted to know was 1) and 2) Was there a sudden jump from neurons or voxels responding to bits to objects? Or did the ventral stream build whole objects gradually out of pieces? 3) In particular what happened in a region called LO or LOC: Lateral occipital cortex. At the time there were competing computational models of what was going on in LOC. One predicted that higher level object areas in LOC would respond to shapes/entire objects, another that they would respond to object parts.

Introducing Lateral Occipital Complex (LOC)

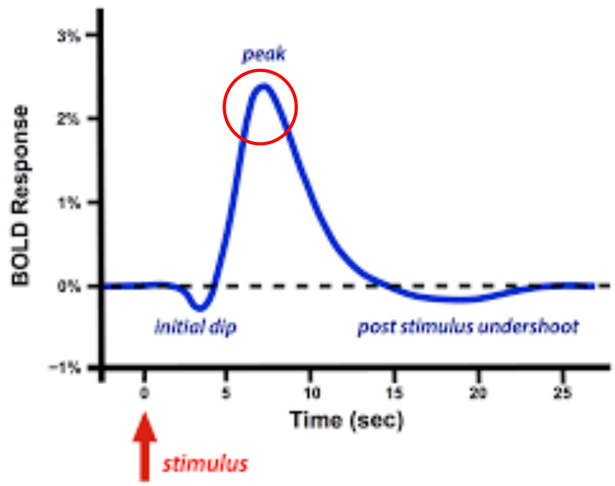
Important for object recognition

- Relies on shape



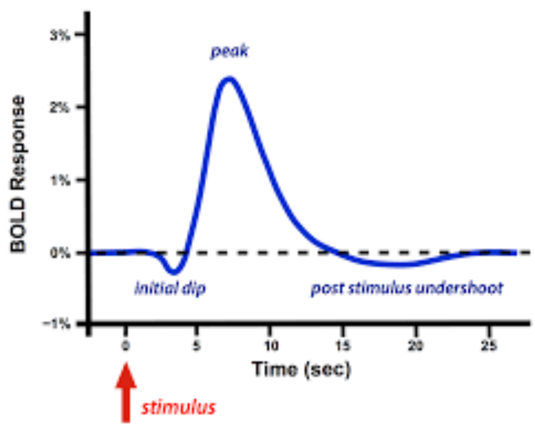
23/1/25 | PSYC 365: Class 6 | 15

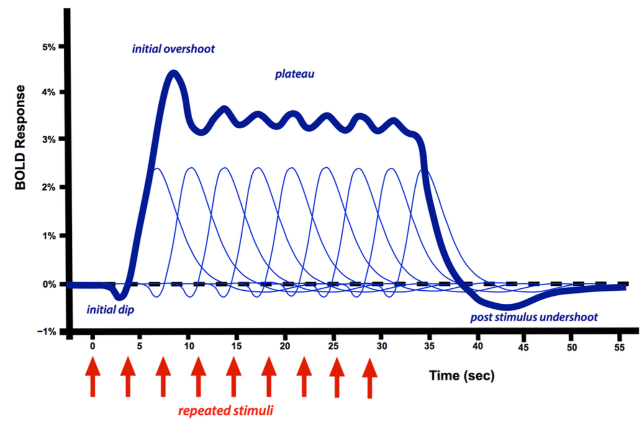
Since we're going to focus on LOC let me show you where it lies in the ventral stream of the visual cortex. Looking at the brain from below the ventral stream is in colour, with different regions of the ventral stream in different colours. Right at the back, you can see primary and secondary visual cortex and then (1) here is LOC on the outside (lateral) in a turquoisey colour along with other regions of the ventral visual stream. The LOC plays a key role in detecting whole objects as things in themselves. So (as Passingham says) in 1998 we knew it was important for object recognition but we didn't know how that happens. The goal was to understand the process by which bits are put together into wholes. And to understand that we need to know that the processing occurs in **stages.** Just like in a CNN, earlier stages process smaller bits, and later stages integrate them.



Let's review what we look at in an fMRI study as the dependent variable. Our dependent variable is the BOLD response in a voxel, and that is a measure of the ratio of oxygenated to de-oxygenated blood, which is on the Y axis. RF The squiggly line is the hemodynamic response function, which DESCRIBES the change over time in the ratio of oxygenated to de-oxygenated blood after an event -- in this case after a visual image is presented on a computer stream. You can see it going up and then back down.. The peak is a good 4-7 seconds after the event it’s responding to occurred. As the needs of the neuronal activity are met, blood flow returns to previous levels. So this is what happens if you do something – say show a picture to a participant and then wait for half a minute. But what if you want your experiment to go faster than showing a picture every half minute? What if you want to show a lot of pictures quickly?

HRF for fast repeated stimuli

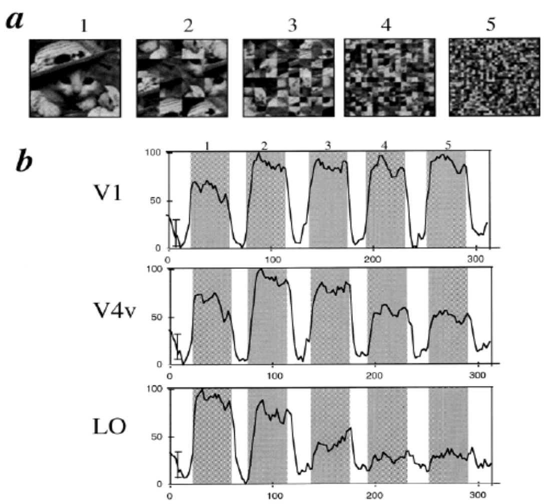




23/1/25 | PSYC 365: Class 6 | 17

So here is the HRF if you show a single picture. But what if you show pictures one after the other fast enought that nstead of a single peak you get a plateau created by a whole bunch of peaks one after the other. This is called a block design, where you would show a whole bunch of pictures of one category and measure the height of the plateau created by all the pictures over a period of many seconds -- again that's the BOLD response or ratio of oxygenated to deoxygenated blood on the Y axis. Then you'd show a whole bunch of examples of a different category of pictures and measure the plateau height for that Ok so hold that in mind for a minute until we get to the actual data from the Grill-Spector study.

What lies in between bits and whole objects?



- Tested different degrees of scrambling
- V1 likes scramble
- V4v partial objects (simpler more localized features)
- LOC: Prefers whole objects (mostly)

Grill-Spector et al., 1998

This is the first Grill-Spector study mentioned in Passingham Chap 2. 1) They tested different degrees of scrambling. They took a bunch of pictures and scrambled them to different degrees and measured BOLD activity along the ventral steream. That's categories 1-5 at the top– tiny fragments to bigger squares to full pics. So there would be a block of a whole bunch of whole pics, another block of pics scrambled to the second level and so on up to level 5. Participants had to say the name of the picture in their heads, even the scrambliest ones.
b) here is the data -- the plateaus of HRFs. A taller hat means the region was more responsive to that kind of image. 2) V1 really doesn't like the cat it likes all the bits with a lot of edges. A region of secondary visual cortex,3) V4v there likes intermediate levels of scrambling – it likes object shapes. 4) And LOC likes whole objects best, then parts, and then it doesn't like them at all. **Meaning LOC is most responsive to whole images.** Note there is a big jump between V4 and LO**.** But this was true of only some of the voxels. A lot of LO voxels actually looked like V4, suggesting the landscape is divided in a more subtle way.

Summary

- Regions along the ventral stream progress from processing bits to part images to wholes, *with simple shapes and partial objects in the middle*
- Authors: Consistent with a "stage-wise hierarchical scheme of object processing"
- Also consistent with increasingly large receptive fields

So that gives us some of the answer to HOW it's done. It builds complex object representation from different elements.

Things we have learned from fMRI

- Encoding Approach: Responses to object parts and then whole objects moving along occipital cortex from EV to LOC
- Encoding Approach using adaptation: More view and size invariant processing as we move along ventral stream.

The visual system gets smarter and smarter about recognizing objects with all kinds of tricky transformations as we move down the ventral stream.
QUESTIONS?

Reading question

According to the blog on machine learning, what distinguishes the first *convolutional* layer of a neural network from earlier AI methods?

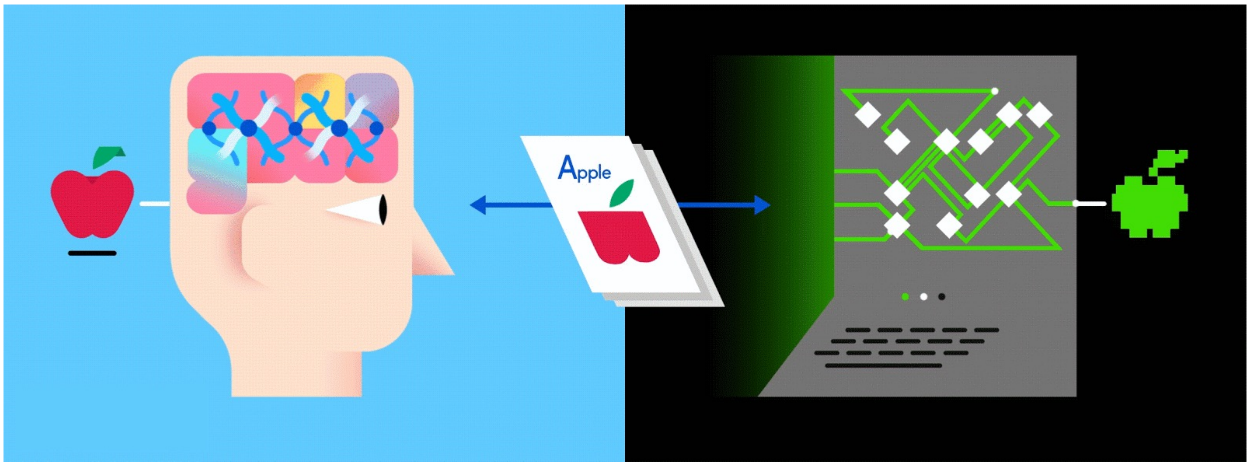
a. The filters are hand designed to pick out specific features

b. Specific algorithms piece together parts into wholes

c. They learn just by looking at data

Ans. C. What's particularly powerful is that unlike earlier AI methods these filters aren't hand designed; they learn and refine themselves purely by looking at data. We don't tell them what features to look for. They figure it out. There is not a set number of features they can pick out. Talk about how we don't tell these networks what features it will be useful to break the image down into. It figures it out for itself based on how the filters break down the image and keeps what's useful. AS it turns out, what's useful to the network is often - but not always - what's useful to the human brain (edges, textures, object parts).

Neural Networks use a similar approach



22

Ok so now we know more about how the ventral stream processes objects from part to wholes. Neural networks do it in a very similar way: Again, *As pictures are fed in to the input layer, the network breaks them down into their most basic components: e.g., edges textures and shapes. As these pictures propagate through the network, these basic components are combined to form more abstract concepts, e.g., curves and different colors which, when combined further, start to look like a stem, an entire orange, or both green and red apples.*

Convolutional layers create maps

23

In CNNs you then have convolutional layers that create maps. Here you multiply or convolve the image with different filters. This process pulls out features – including edges, as V1 does, and colour as V4 does. By this process of filtering, the convolution layers create maps — different, broken-down versions of the picture based on these different elements. Each map is dedicated to a different filter and these maps indicate where its "neurons" see an instance (however partial) of the color red, stems, curves and the various other elements of, say, an apple. There can be multiple convolutional layers in a CNN. Here is an illustration of a first layer that puts together edges and blobs, another that puts together textures. A later convolutional layer puts together object parts, and it can do that in different ways. After that an activation layer goes back and highlights what might be particularly useful – the blog compares it to Marie Kondo going back and finding the things that spark joy.

Reading question

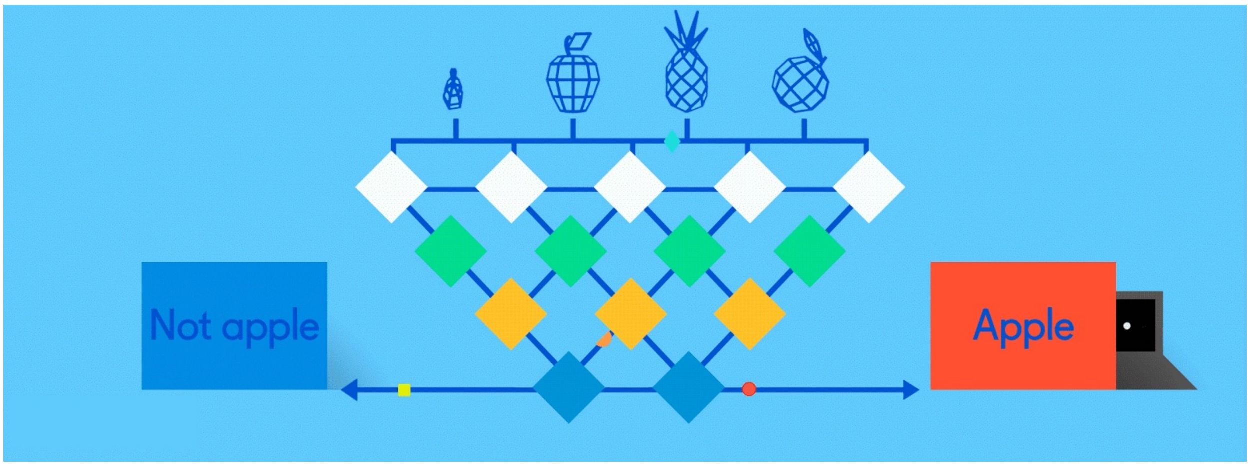
An example of what the pooling layer of a neural network does is:

a. Edits features such as redness or curviness into the best examples
b. Reduces amount of data so it's manageable.
c. Compiles the features into entire objects
d. All of the above

24

Ans. D. After all the convolving there are too many maps – too much data to be manageable. So then you need to reduce it down. So we pick out the best examples of of each item to keep and get rid of the rest. It edits down each feature map into a *Reader's Digest* version of itself, so that only the best examples of redness, stem-ness or curviness are featured. And then it compiles the bits into an entire object.

Fully Connected layer: Guess the Object!



25

Then in the fully connected layer the pooled maps are connected to output and each node (in this case apple/orange) votes on each feature map. It then looks at how far off the answers are. (e.g., do 20% vote for orange maps when the object is an apple)? It then goes back and refines maps in previous layers. At first, its predictions will appear as random guesses, as no real learning has taken place yet. If the input image is an apple, but “orange” is predicted, the network’s inner layers will need to be adjusted.

That article was written before the era of L>MS. LLMs are trained on a lot of data, they have extra features that allow them to compress information and detect relationships between words an concepts, and they make very good predictions. And GPT 4 can work with images as well as text. But they aren’t perfect by any means!   This example is from a blog called AI Weirdness.

Here's the prompt I gave ChatGPT4: "Please generate a set of mammals on a plain white background, each mammal species clearly labeled."*However, ChatGPT4 is a text-generating model, so it doesn't have the ability to generate its own images. Instead, it's designed to pass my prompt to another model called DALL-E3. But not before modifying my prompt for "better results". The prompt it actually sent to DALL-E3 was*: "Photo of a collection of mammals neatly arranged on a plain white background. The mammals include a lion, an elephant, a kangaroo, a panda, a dolphin, and a bat. Each mammal has a clear, legible label beneath it indicating its species name."

Then DALL-E3 generated the image above, and now we know which model to blame for the labels. Note that ChatGPT4 did not specifically ask for a giraffe, but DALL-E3 did two of them anyways.

Q for CS/AI –savvy students: The neural networks made easy paper described a supervised learning process. Are LLMs supervised or unsupervised? Selfsupervised?

Bowers et al: Machines are NOT just like us!

- DNNs as good as or better at classifying (categorizing) objects
- But claims they are similar to human vision ignore psychology literature on human vision
  - They don't account for psych findings about "core properties" of human vision!
- DNNs predict behavioural or brain responses (e.g. activation in IT or object classification performance)-- but they don't tell us HOW or WHY

Justification of DNNs as best models based on *prediction-based experiments*:

1. DNNs are most accurate at classifying images
2. DNNs are best at predicting human errors
3. DNNs are best at predicting brain responses

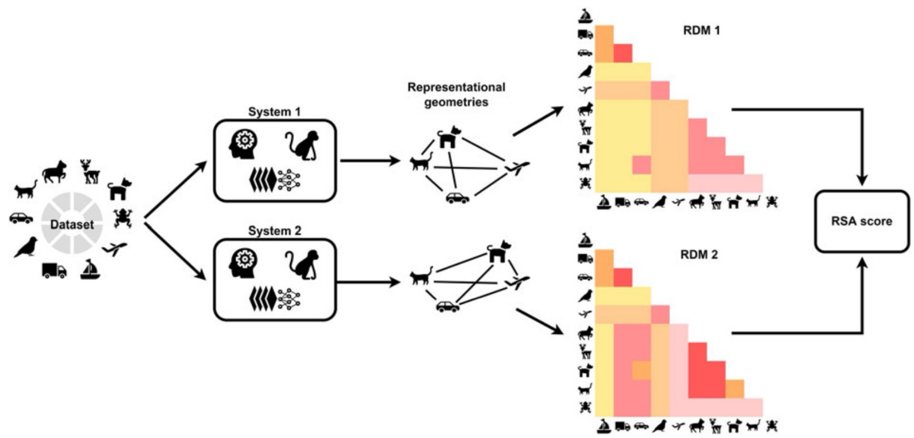
Bowers et al., 2023, Deep problems with neural network models of human vision , *Behavioral and Brain Sciences*

CNN = Convolutional Neural Network
DNN = Deep Neural Network

27

The authors of the Bowers paper take issue with the claim from the Neural Networks Made Easy blog that machines are just like us. Note BBS is a very high impact cog sci publication where researchers will post a theory or opinion and experts in the field are invited to respond to it. **Q:** What is the difference between a *prediction-based experiment* and a *controlled experiment*? **Prediction models** predict experimental results overall – after averaging across prediction. **Controlled experiments** systematically vary features of stimuli or other conditions to understand questions about WHAT information is being used to recognize objects

Using RSA to compare brains & DNNs



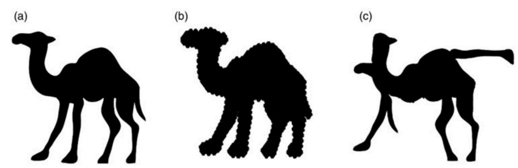
28

One example they use of how human or monkey brains and DNN performance is compared is using RSA.. A series of stimuli from a set of categories (or conditions) are used as inputs to two different systems (e.g., a brain and a DNN). The corresponding neural or unit activity for each stimulus is recorded and pairwise distances in the activations within each system are calculated to get the representational geometry of each system. This representational geometry is expressed as a representational dissimilarity matrix (RDM) for each system. Finally, an RSA score is determined by computing the correlation between the two RDM. BUT

CORRELATION IS NOT CAUSATION, AND THIS CAN BE DRIVEN BY CONFOUNDS

Some inconsistencies between humans & machines

- DNNs use texture and humans use shape to categorize images
- DNNs use local shape and humans use global shape
- DNNs are bad at identifying degraded or deformed images
- DNNs can't distinguish boundaries from surfaces
- And more…



29

2) Illustration of (a) a silhouette image of a camel, (b) and image of a camel in which local shape features were removed by including jittered contours, and (c) and image of a camel in which global shape was disrupted. The DNNs had more difficulty under conditions (b) than (c). Images taken from Baker et al. (2018b). 4) In human vision boundaries and surfaces of objects are processed separately and then combined early in the visual processing stream to perceive colored and textured objects. This separation is observed in V1 with neurons in the “interblobs” system coding for line orientations independent of color and contrast and neu- rons in a

“blob” system coding for color in a way that is less dependent on orientation

Size & Viewpoint Invariance

Size invariance

- The ability to know an object is the same when it is a different size on the retina

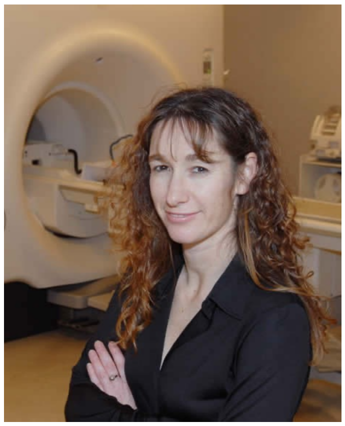
Viewpoint invariance

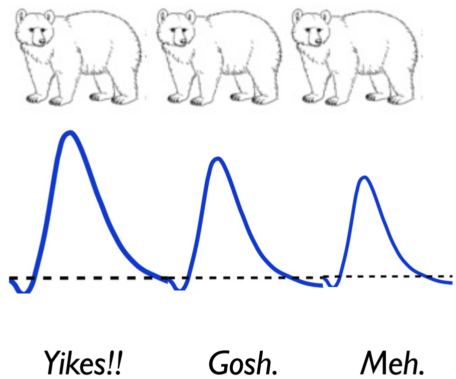
- The ability to recognize objects from different viewpoints

30

The next Grill-Spector study we're going to talk about focuses on size and viewpoint invariance. To recognize an object, we need to be able to recognize it as the same under different conditions. 1. ....even though the size on the retina is different 2. Even though the image on our retina has no information about what the object looks like from other views. Our retina is really literal, right? Nothing about the literal map that it gets gives you information about perspective. **So How does the brain do it? WHERE in the visual stream do neurons or groups of neurons understand that a shoe is a shoe whether, because of distance, it's huge or tiny.**

fMRI Adaptive Suppression

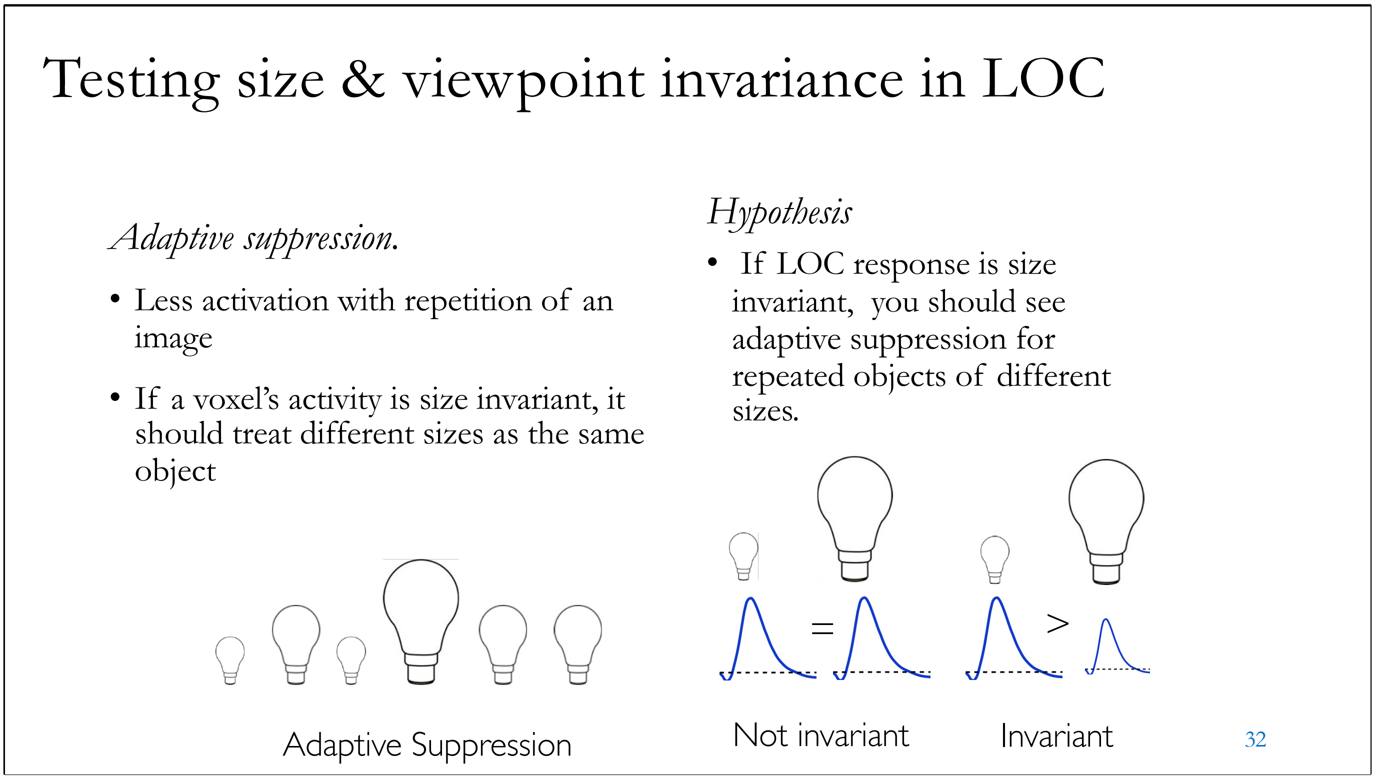




Grill-Spector et al., 1999

31

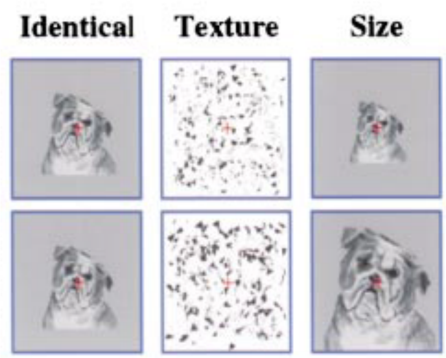
To get at this problem, Grill Spector used a clever approach that took advantage of a phenomenon called adaptive suppression. That is, when you see an image repeated, the hemodymamic response is smaller with repetition. The visual system treats new objects as requiring more brain resources than familiar ones. The idea is that when an image is familiar the neurons don't have to work as hard or they become more efficient at recognizing the image.   So adaptive suppression is when the HDR is reduced with repetition. This is an illustration for single hemodynamic responses to single pictures. It's like seeing a bear in your walnut tree. The first time, yikes! You have a strong reaction. The second time you sure sit up and pay attention. The third time you think, oh there's the bear again.



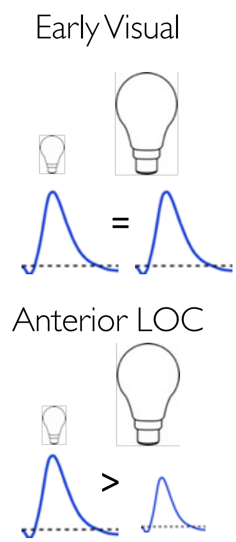
Active suppression is sometimes also called adaptation. That's another case of terms that are used interchangeably. This is the logic of this Grill-Spector study: The idea was that since you get less activation with repetition of an image, and you show a bunch of images of different sizes, then if a given voxel's activity is size invariant -- if it recognizes an object no matter what size it is -- then it should treat a single object shown in different sizes as the same object. So if you show objects in a whole range of sizes (or viewpoints) like so (x 6). One of two things will happen (7). If neurons in a visual cortex region are NOT size invariant then the voxels in that region will only show adaptation for objects that are the exact same size. They will treat objects of different sizes as different objects and they will not show adaptation to them. So when you test adaptation by showing a small and a large object, the hemodynamic response -- again this is one hemodynamic response function to presentation of a single picture on screen -- it will be the same size because it won't shrink. (click) BUT if the neurons ARE size invariant they will treat every lightbulb as as the same object regardless of size, just as you would, and when you test with 2 different sizes there will be a smaller HDR for the repeated object. So to summarize, for the LOC,, they hypothesized (click):... They expected that neurons in early visual cortex, which they used as a control region, would NOT show adaptive suppression to objects of different sizes.

What regions show size invariance?

Stimuli



Results: Size



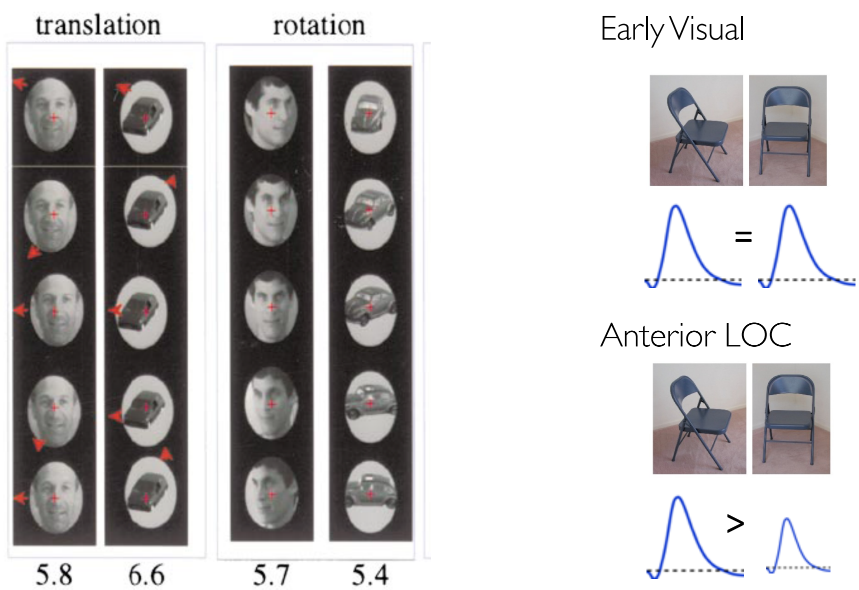
Anterior LOC had size invariant voxels

Earlier visual regions did not

33

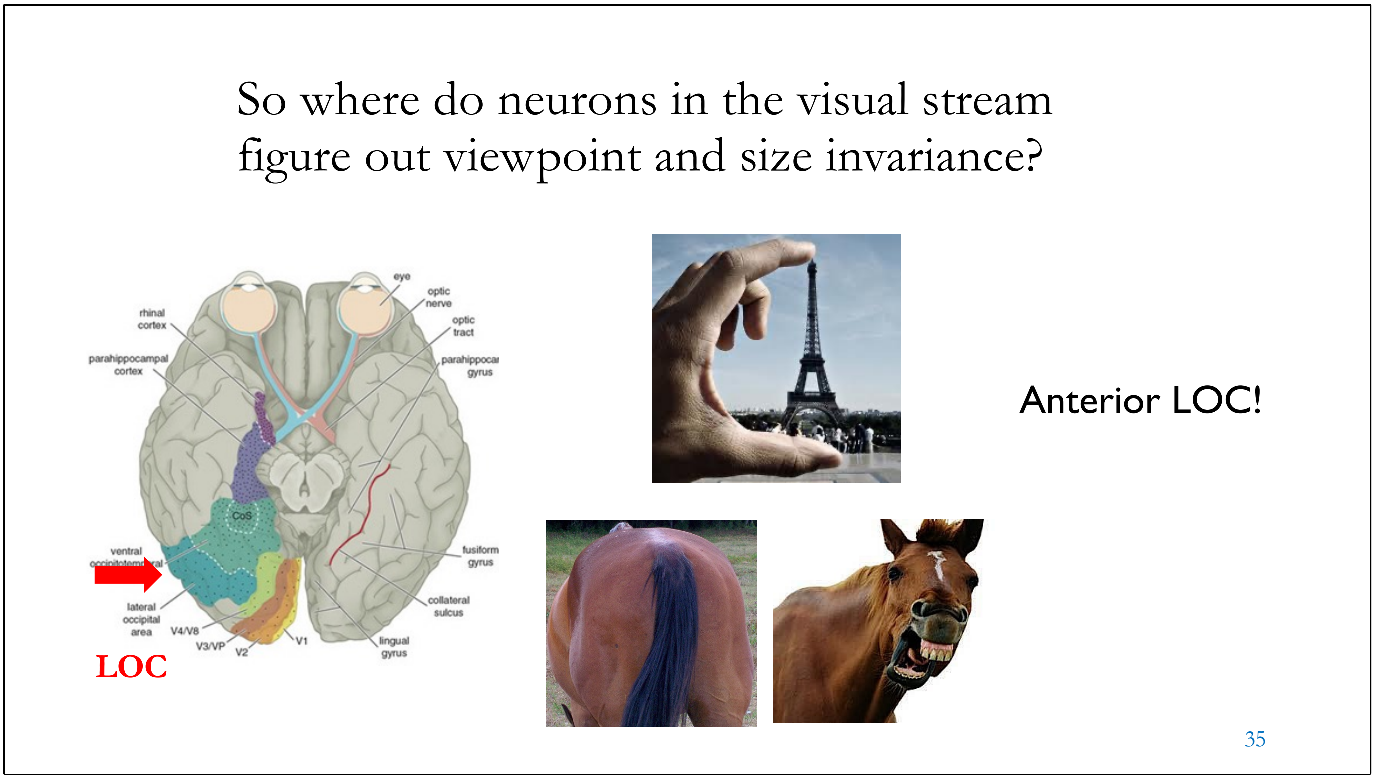
They had several different conditions of different types of images they showed in the scanner. In one they showed **Identical pictures** of the same image – all the same size– illustrated by the dogs in the left column.. Do you think the HDR would be reduced or the same size for this? They also showed a condition where there were different **textures,** as in the middle column. We're primarily interested in LOC so texture would be a control condition **Q. But what region would be sensitive to texture?** Finally the condition of interest was one where they showed the same picture in different sizes as on the left. What did they find?
First let's look at EV, the control region. [1] Here the HDR was the same size for a large version of an image as for a small version of the same image. So the early visual cortex is treating them as different objects. This makes sense if you think about how literal the retinotopic maps in EV are. Now let's look at LOC [2] Some voxels in the anterior (front part) of the LOC showed a smaller HDR with the repetition of an image of a different size, indicating that the neurons in that voxel treated it as the same object -- in other words, their response was size invariant. So to summarize [3]

What regions show viewpoint invariance?



34

The next question concerned viewpoint invariance. So they showed a bunch of faces and other objects that were just moved to a different part of the screen (translation) or were rotated to a different angle. Where did voxels treat an object as the same even when shown from different viewpoints? Once again, as expected, EV voxels show the same size HDR for the same object shown at different angles, indicating it sees them as different objects. And once again voxels in anterior LOC -- further along the ventral stream --showed adaptive suppression indicating the neurons in those voxels were viewpoint invariant. But voxels in the posterior LOC which is earlier in the ventral stream, right after EV, did not show adaptive suppression indicating viewpoint invariance. So you start to see voxels with viewpoint invariance somewhere midway along LOC.



Focused on Loc [click]– that patch on the outside or lateral section of the back of the brain The question of where in the visual cortex you start to see view invariance is the question Grill-Spector set out to answer by looking at adaptive suppression. And again the focus was on the LOC which really likes shapes. And it turns out that the anterior part of the LOC, which is the part furthest along the ventral stream, is where you start to see reliable viewpoint and size invariance. QUESTIONS?

Summary

- EV and earlier LOC only "recognize" one size or viewpoint
- *In general*, later LOC regions "recognize" objects from any angle or any size
- Later LOC regions do this *in part* by integrating information from earlier regions…

36

[click] by "recognize" I mean show adaptation effects. [click] From this mix of results the authors conclude that LOC is functionally varied: Called a complex for a reason!

Things we have learned from fMRI

- Encoding Approach: Responses to object parts and then whole objects moving along occipital cortex from EV to LOC
- Encoding Approach using adaptation: More view and size invariant processing as we move along ventral stream.

In a future class on classifying objects we'll look at what we've learned from later studies using a decoding approach.

Learning Objectives (object processing cont)

- Appreciate our amazing capacity for object recognition
- Describe the roles of two visual cortex systems in object recognition
- Discuss the contribution of classic fMRI research to our understanding of object processing stages in humans
- Compare challenges posed to human and computer vision

Discussion time!

For next class

- Read Harada et al., 2020