Machine Learning for Human Vision and Language

Lecture 3:

Early (feedforward) visual processing

Ben Harvey

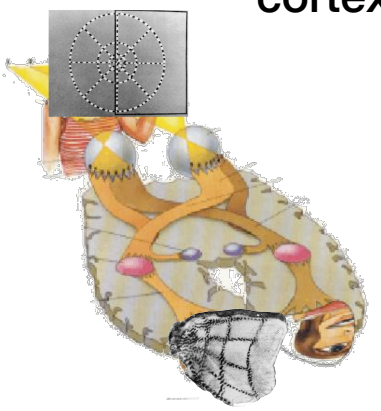
In the last class, we discussed how the computations performed by neurons in the brain represent and process information, and how these computations provides the inspiration for operations performed in deep convolutional machine learning networks.
-The principles involved apply equally well anywhere in the brain: the organisation of neurons is very similar everywhere.
-But when we start looking at the results and implications of this processing, it becomes very hard to talk in general terms, and it is very useful to follow a specific example

Why study early vision?

- Image processing is a common application of artificial deep networks
  - Particularly object recognition
- Early vision is very thoroughly investigated
  - Visual input is easy to control
  - Good animal models of human vision
  - Computational aspects are well understood
- Common application of DCNN as simulations of neural processing

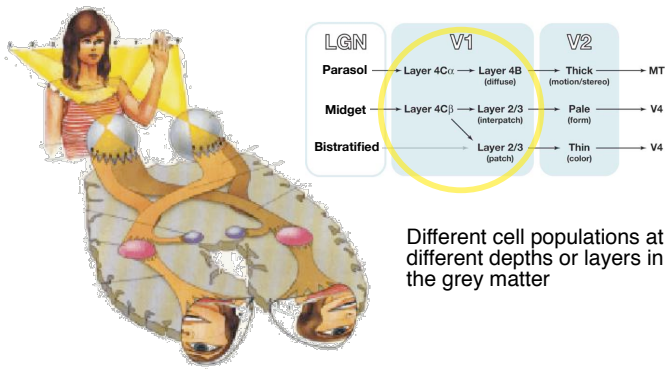
-We used the early human visual system as an example network, for various reasons.
-First, computer visual problems like object recognition are exceptionally difficult to solve using formal rules, and so have been an excellent application for artificial deep networks.
-Second, the early visual system is probably the best understood system in the human brain, at the computational level: we really understand HOW the brain works here. This is largely because it is very easy to precisely control visual input, and because we have good animal models of the visual system.
-Third, and related, early visual responses are a common application of DCNN as simulations of neural processing.
-However, we believe that very similar principles operate in other networks in the brain.
-We saw how the retina and primary visual cortex separate images into multiple feature maps for different colours and spatial frequencies.

Organisation of early visual cortex



On leaving the retina, ganglion cells feed in to the visual cortex via the thalamus's lateral geniculate nucleus, effectively a relay that doesn't perform transformations.
-In this process, the information from the two eyes is joined together. The left half of both retinas (the right half of the retinal image) goes to the primary visual cortex (V1) of the left hemisphere, and the other half to the right hemisphere.
-Here, sites responding to particular retinal image positions are stained on this cortical surface
-This reveals the cortical retinotopic map organization: the layout of visual field mirrored on cortical surface.
-As a result, spatial relationships from the image become spatial relationships in the cortex, as we discussed last time. This allows filtering and integration by dendritic tress with a limited cortical extent, a limited spatial extent in the feature maps.
-It also allow interactions over minimal distances: cells that wire together lie together for efficient interactions.

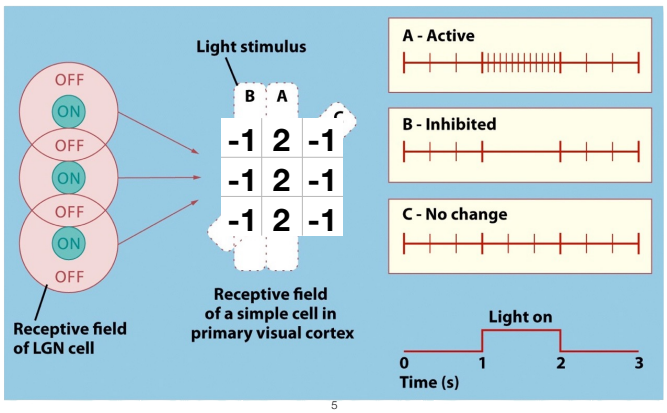
Organisation of early visual cortex



Different cell populations at different depths or layers in the grey matter

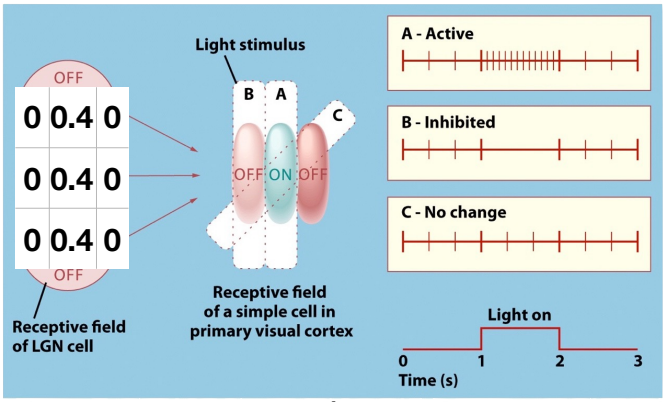
-However, there is some separation here between the representation of features carried by parasol, midget and bistratified cells, different cell groups carrying information from the retina.
-These are held in slightly different cell populations at different depths or layers in the gray matter.
-Gray matter layers are not equivalent to neural network layers, but here we do see several neural network layer-like stages, i.e. different numbers of synapse steps from the photoreceptor input, are sometimes in different layers of gray matter in the same cortical area.
-There is some spatial separation between these cell populations from the retina here, but they are also held in the same visual area at the V1 stage.
-The spatial relationships allow these neurons to interact most easily when they are carrying information on the same feature type, as these are closest together.
-But those for other feature types are still nearby, allowing analysis of relationships between different stimulus characteristics found in same location, using filters spanning all of these feature maps.

Orientation detection in primary visual cortex



Responses to specific orientations first emerge in V1. These cells respond when an edge has a specific orientation (the preferred orientation) and is shown in a specific position.
-Orientation selectivity can be built up from spatial interactions between different centre surround cells. This group of three all need to be active for this filter to respond, and the neighbours must be inactive.
-The corresponding convolutional filter might look something like this

Orientation detection in primary visual cortex



6

-But remember that the inputs to this filter already have suppressive surround organisation.
-So, expressed as transformations of the summation of these inputs, it might look more like this: all of these inputs need to be active to reach a threshold of one
-Even at this early stage, it is misleading to think of these filters operating on image inputs. The inputs are already outputs of the previous layer of filters.

Orientation detection in primary visual cortex

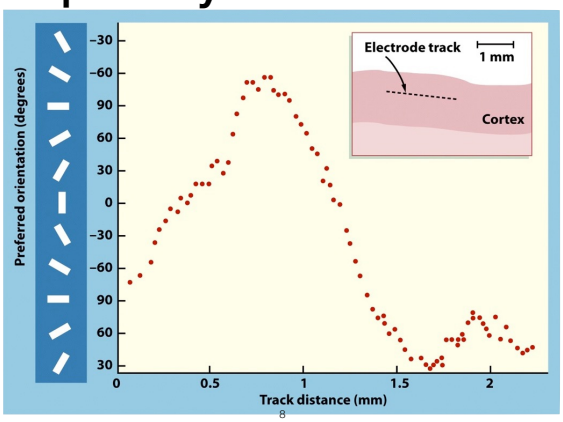
Image input

Feature map input

7

Unfortunately, humans are bad at thinking like this, in terms of transformations of representations rather than transformations of images.
-Conveniently, neurons and computers do not have this limitation: they transform whatever input they are given.
-At this early stage, we can easily make an image that gives an idea of where the contrast in the image is, the pattern of outputs in a feature map from the retinal ganglion cells.

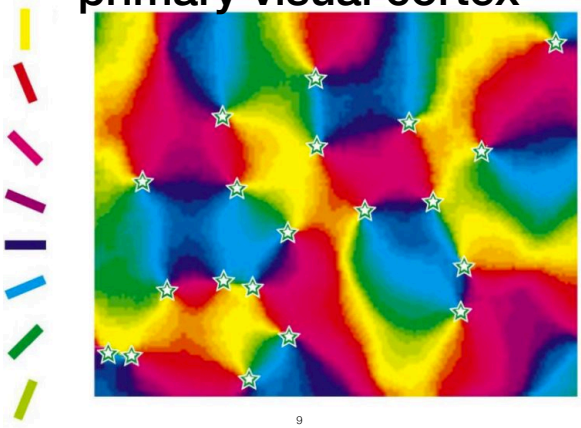
Orientation organisation in primary visual cortex



8

At a coarse scale, neurons processing the same part of the visual field are grouped,
-And neurons with similar orientation tuning are also grouped together. But this grouping is at a very fine scale, the scale of the cortical column. Each column contains neurons responding to the same position, but different orientation.
-So these different orientation columns form further feature maps

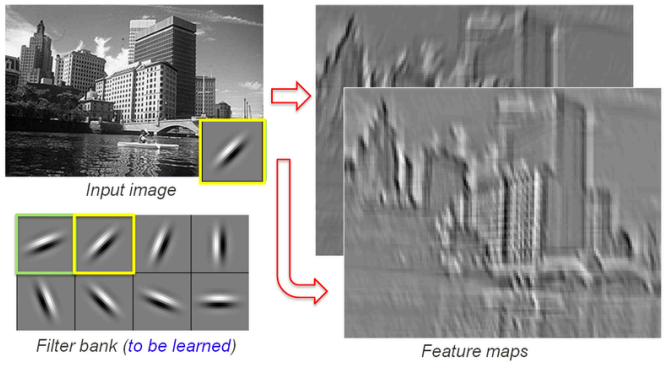
Orientation organisation in primary visual cortex



9

It has since been shown using 2D optical imaging techniques that orientation preferences change in two dimensions across the cortex. This description of 2D orientation 'pinwheels' has superseded the description of columns that is apparent when advancing an electrode along a 1-dimensional line

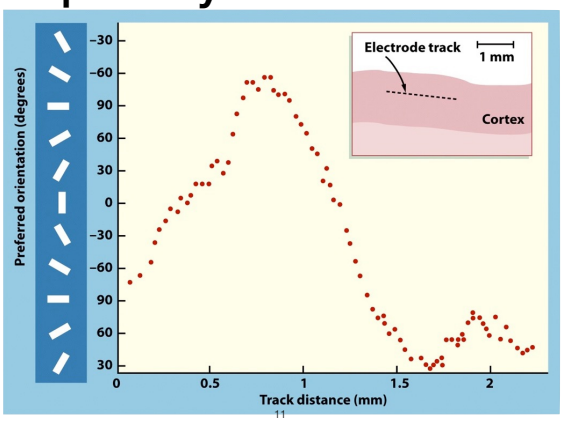
The filter/convolve operation



10

We have seen this separation into feature maps for different orientations before…

Orientation organisation in primary visual cortex

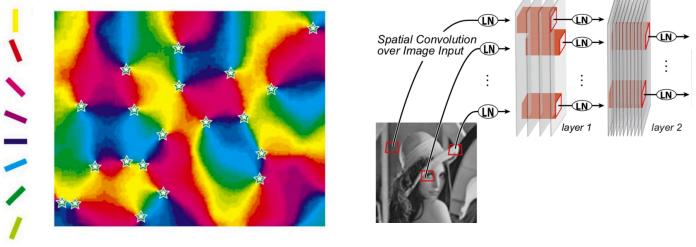


11

Again at this level, neurons that have similar responses are found together.
-This can be very computationally useful. If we have a filter that will accept a range of orientations, the neuron implementing this filter can synapse with a group of nearby neurons.
-However, in a biological network, these orientations can have any value: they are not limited to a small fixed number of discrete filters and feature maps.
-Instead, it's more like orientation is an entire third dimension of a feature space, with position forming the other two dimensions.
-So neighbours in this third dimension have similar responses, and filters may have a meaningful extent in this third dimension.
-This is not the case in artificial deep networks, where filters typically span all feature maps in a layer and can learn any relationship between different feature maps.
-However, if analysing relationships between nearby feature maps is most useful to task performance, filters can develop that largely ignore 'further' feature maps, which will have an effect of focussing on relationships between 'nearer' feature maps
-Constraining artificial network filter patterns to a neighbouring range of feature maps is rarely done, though it is certainly possible and may have some computationally benefits.

Feature map dimensions in a 2D cortex

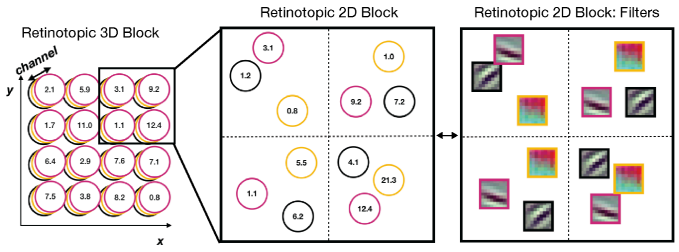
2D, intermixed feature maps **NOT** **3D, independent feature maps**



12

The gray matter in the brain is arranged in a folded 2-dimensional sheet, and does not access to a third spatial dimension
-However, in artificial DCNNs, filters span a third dimension across the multiple feature maps at each network level.
-Somehow, the brain creates an organised mixture of these different orientation components
-How does this organised, intermixed representation come about?

Feature map dimensions in a 2D cortex



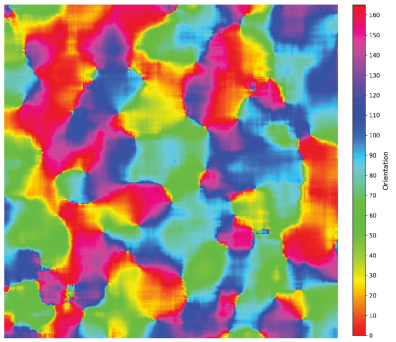
Fully connected model with connection from matrix to 2D space (2d matrix?)

Margalit, Lee, DiCarlo & Yamins (in preparation)

13

Super-recent work has simulated how this might happen in an artificial DCNN.
-Instead of a DCNN with independent channels, all units in the first layer are forced into the same 2d feature map.
-The convolutional filter leading to the next layer must sample within this single 2d feature map only.
-As part of the machine learning process, units within each x,y position can move around, and will tend to get closer if they respond similarly.

Feature map dimensions in a 2D cortex

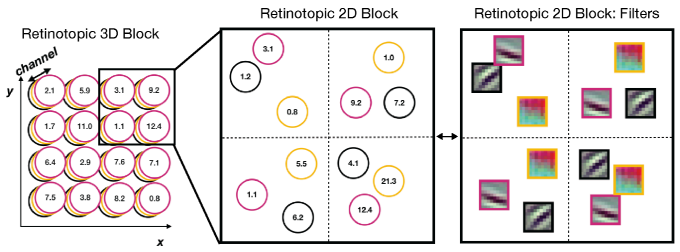


Margalit, Lee, DiCarlo & Yamins (in preparation)

14

If we look at where those units end up, units with similar orientation preferences (individual pixels in this image) come together, and self-organise into structures that closely resemble pinwheels.
-So by including spatial constraints in learning rules, it becomes clear that neurons with similar responses, which will tend to respond together, also lie together.

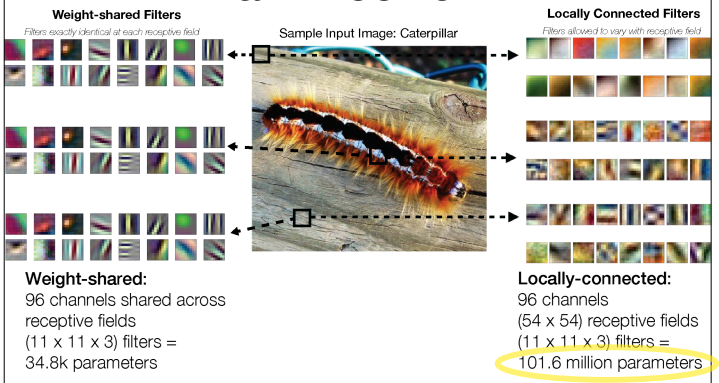
Feature map dimensions in a 2D cortex



Margalit, Lee, DiCarlo & Yamins (in preparation)

This study also found that the same organisation did not emerge if the same filter was used in all spatial positions, i.e. if weights were shared across the layer.
-In that case, a homogenous mix of orientation preferences emerged.

Feature map dimensions in a 2D cortex



Weight-shared:
96 channels shared across receptive fields
(11 x 11 x 3) filters =
34.8k parameters

Locally-connected:
96 channels
(54 x 54) receptive fields
(11 x 11 x 3) filters =
101.6 million parameters

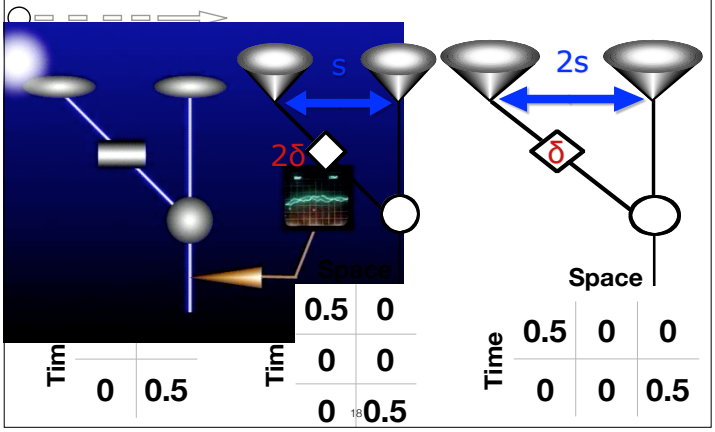
Margalit, Lee, DiCarlo & Yamins (in preparation)

-So they had to implement independent weights at each filter location.
-These filters differ across the image in some interesting ways, most notably responding to finer-scale features towards the centre of the input image.
-This is very computationally expensive, as a new set of weights must be learned for each of 54x54 image positions (in this example), and each of these has a large number of paramaters (11x11 pixels, 3 colour channels)
-As we discussed last time, this is no problem for biological neural networks, only for their artificial simulations.

Orientation selectivity

- Contrast is initially computed in an orientation-independent filter
  - In the retinal ganglion cell
  - Artificial DCNNs avoid this step, going directly from image to orientation
- Orientation-selective responses are computed in V1 by operations comparing these RGC outputs
- Orientation preferences gradually change across the cortex
  - At a much finer scale than the spatial visual field maps
- This squeezes multiple feature maps into the same 2D cortical surface
  - This may arise to optimise connection lengths
  - Only when filters are not shared across the network layer's image representation

Motion detection in primary visual cortex



18

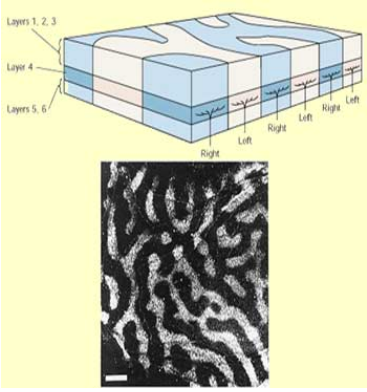
-It is rare for artificial deep networks to consider changes over time. They are generally given single images as inputs rather than movies, largely because of computational limitations. But biological visual systems inherently receive inputs that change over time.
-So biological systems have a mechanism to detect motion, its direction and its speed.
-This is a neuron with two spatially separated inputs, one with a delay. For the neuron to sum up to an action potential threshold, the inputs must arrive at the axon at the same time, so the the moving object must cross the span in a time corresponding to the delay.
-This can be expressed as a filter with a time dimension and, for simplicity, only one spatial dimension. Here the activation threshold would be 1.
-So this is specific for one speed, i.e. one ratio of the span to the delay. We can therefore change the speed that makes the response by changing the delay: increasing the delay gives a response to a slower speed.
-We can also change the speed by changing the span between the two inputs: increasing the span increases the speed

Elements in Scene Recognition

19

So we have separate responses to colour, orientation and motion in different neural populations.
We normally perceive all of these together.

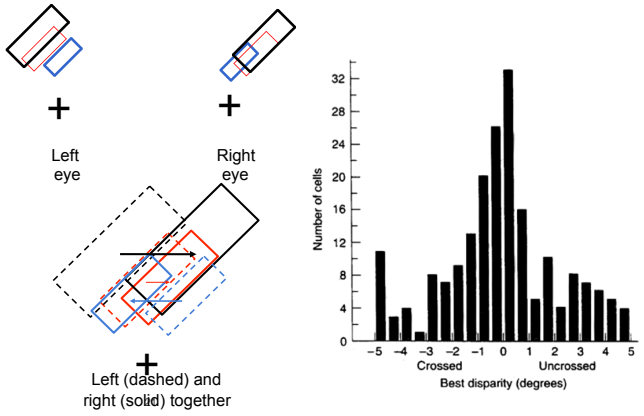
Binocular integration



20

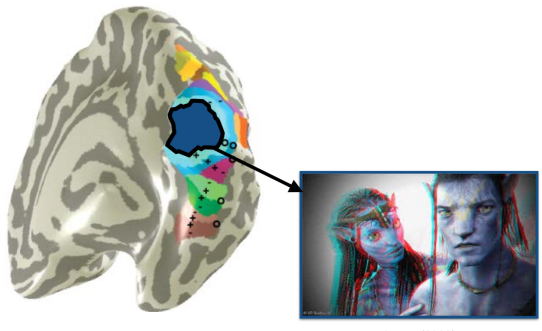
Primary visual cortex also receives separate inputs from the two eyes, doubling up the feature maps
-These are initially kept separate, as distinct cortical columns for the left and right eyes. Various methods can mark the parts of V1 receiving input from one eye, which forms a distinct pattern of interlaced columns, or feature maps from the two eyes.

Binocular integration



These feed in to neurons which integrate the responses from the two eyes, effectively with separate receptive fields for the two eyes.
-This means these cells respond to a particular position difference between the two eyes, which reflects the distance of the object from the viewer's fixation in depth.
-This underlies a large part of our depth perception.

Binocular disparity: and binocular integration



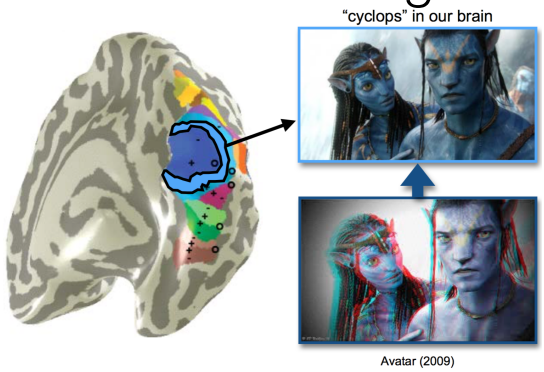
Avatar (2009)

Barendregt, Harvey, Rokers, Dumoulin (2015)

22

So if we look at the responses of V1, we see responses to the individual images in the two eyes

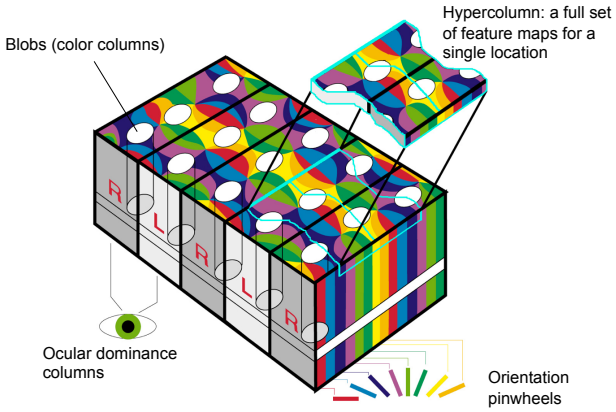
Binocular disparity: and binocular integration



Barendregt, Harvey, Rokers, Dumoulin (2015)

But after V1 (here in V2) we see responses that reflect the integrated perception from both eyes

Intermixed feature representations



These different elements effectively form several distinct feature maps that are all mixed together at the same cortical location.

-A full set of all feature maps for a single image location is called a hypercolumn, and takes up about 2x2mm of the cortical surface.

V1 feature maps

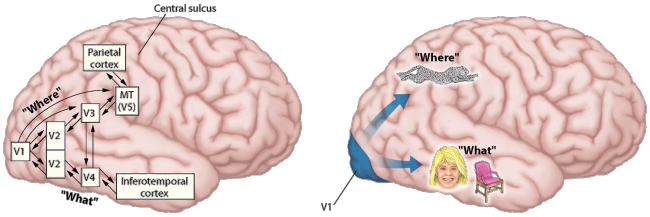
- Colour (red, green, blue yellow, light, dark)
- Eye (left, right)
- Spatial frequency (continuous)
- Orientation (continuous)
- Motion direction (continuous)
- Motion speed (continuous)

In V1, we therefore have a very large complex set of feature maps, with each feature represented at all spatial positions.

As a result, V1 is very large.

This set of feature representations forms the input into subsequent processing stages

Visual pathways beyond V1

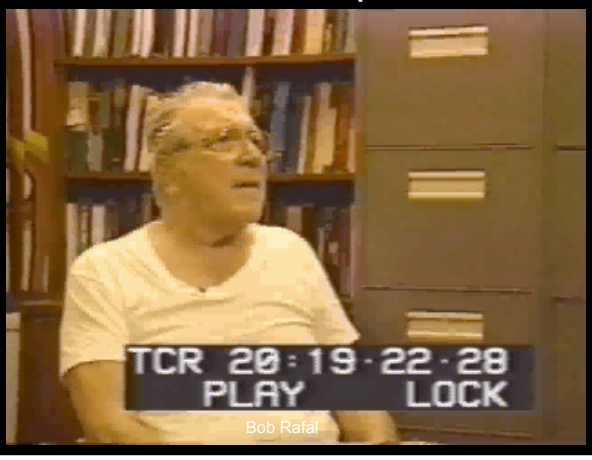


Midget-Parasol becomes What-Where
Form and color vs motion and space
Or Recognition vs Action
Or Ventral vs Dorsal
or Temporal vs Pariatal

26

-In V1, different features are processed in intermixed columns, and something similar happens in the next areas, V2 and V3.
-After V3, form and motion information are often processed separately in different areas
-One of these pathways specialises in motion and space processing, while the other is mainly involved in object recognition.
-These are effectively two deep networks, processing information relatively independently but simultaneously to achieve different goals.
-The network for object recognition is well simulated in artificial networks.
-The network for motion and space processing is not. It is relatively easy to give computer system good spatial information and make a good spatial model of its environment using a range of sensors and cameras. So spatial cognition is not an obvious target for deep learning in computer vision.
-For humans, we have to rely on the input from the two eyes and how that changes over time to make an internal 3D model of our environment.

Dorsal lesion: Optic ataxia

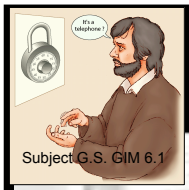


Bob Rafal

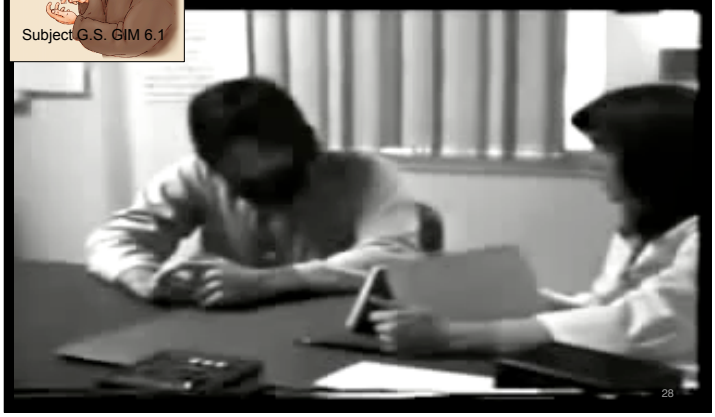
27

Damage to the brain areas in these two processing streams causes very different deficits.

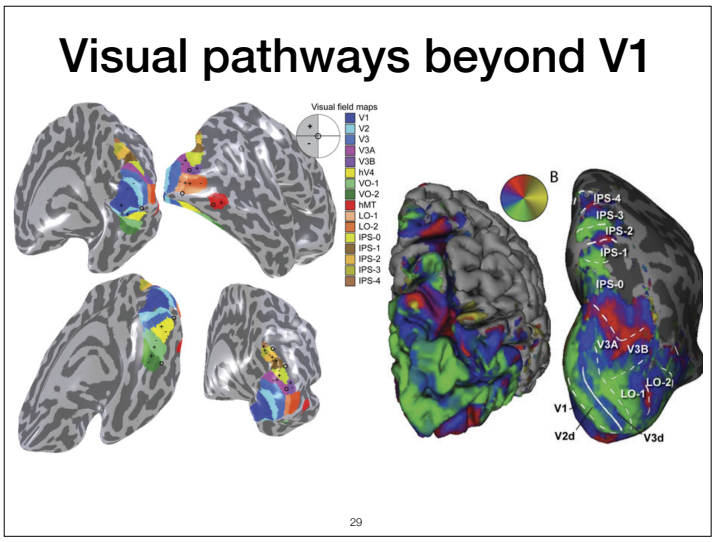
Ventral lesion: visual agnosia



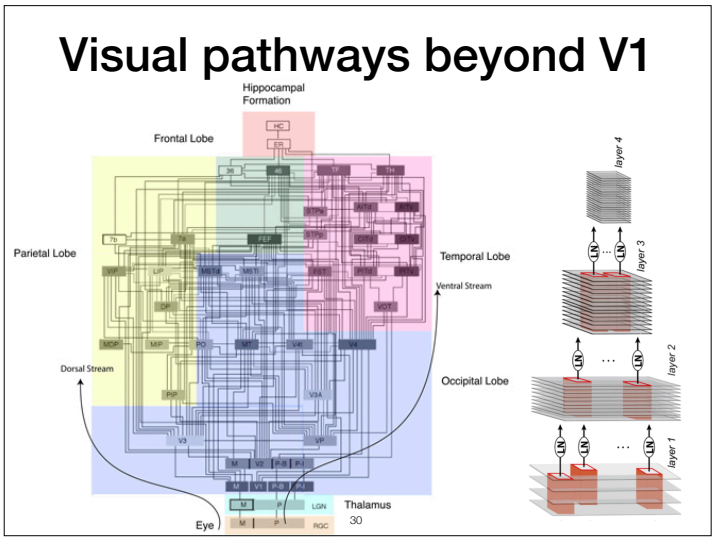
Subject G.S. GIM 6.1



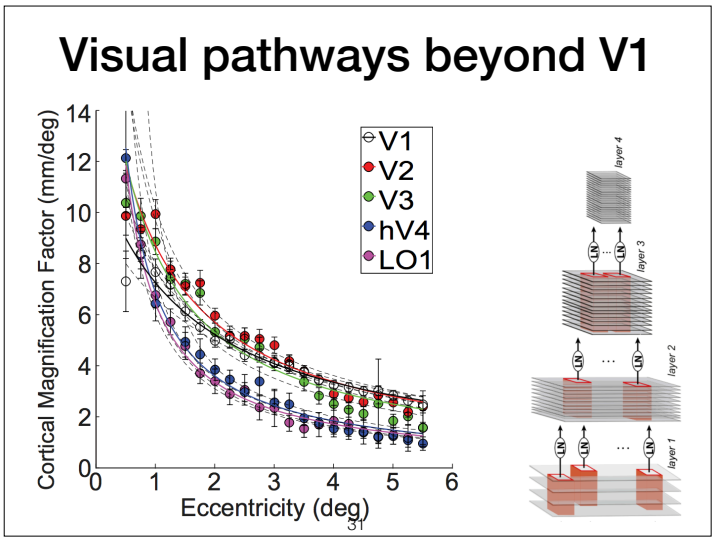
28



V1 is a visual field map, with a spatial layout.
-Beyond V1, we have a series of further visual field maps, at least 30 over the processing streams
-These may be best understood as layers of deep networks

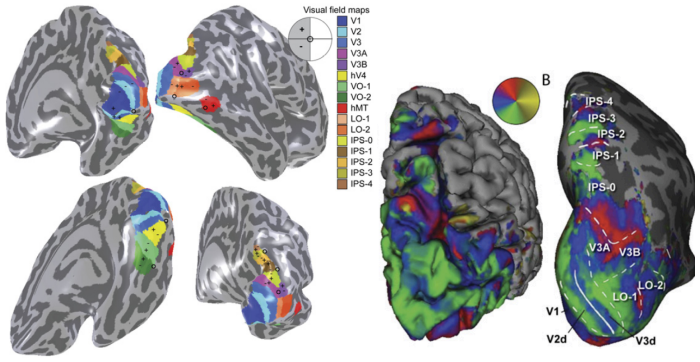


BUT the hierarchy is not as linear as in an artificial network. After a simple pathway from the eye to V1, lots of areas sample directly from V1, and everything really becomes a web of connections.
-This contrast sharply with the simple linear hierarchy that has typically been used for artificial networks.
-As we go through more stages, there become more of these areas
-This partially parallels the increasing number of feature maps through an artificial DCNN
-The difference here is that these later brain areas are physically separate, so it is not feasible for a convolutional filter to interact with many brain areas like it can interact with a stack of feature maps.
-Instead, different areas are forming different, separated pathways to achieve different tasks, for example object recognition and spatial perception.
-An artificial DCNN, as currently conceptualised, is trained to do a single task. A human can do lots of tasks with the same visual input, and in would likely cause conflicts if a single area had to be trained for many tasks with conflicting demands.



Going up this hierarchy (V1-V2-V3-V4-LO1 etc), all of the areas have a larger representation of the central visual field (more millimeters of cortex per degree of retina).
They also become smaller: the area under the curves decreases up the hierarchy.
This resembles the shrinking spatial dimension of an artificial network.

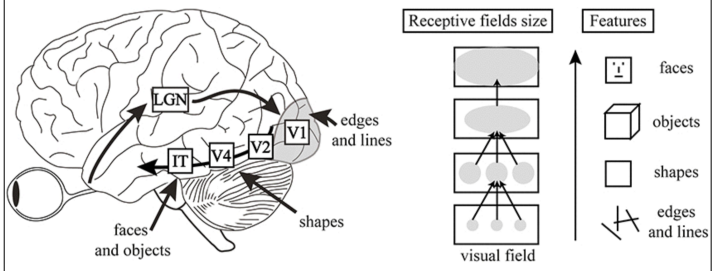
Visual pathways beyond V1



32

-The different visual field maps specialise in different things, but beyond a crude description of 'object analysis' and 'spatial analysis', it has been very hard to understand what each area does.
-Humans like to think in terms of an area having a specific function, but we are bad at thinking about a 'function' as being a deep network feature transformation, particularly for the higher layers.
-So humans are bad at thinking in deep network feature transformations, but the brain is very good at doing these feature transformations.'If the brain was simple enough for us to understand, we would be too simple to understand it'

Spatial integration up the hierarchy

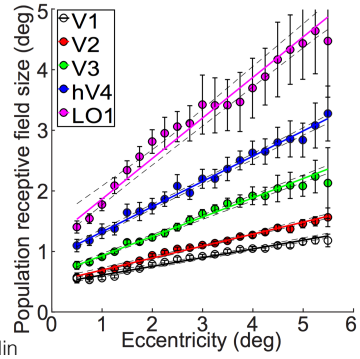


Herzog et al, 2014

33

Through these hierarchical stages, we find responses to increasing complex features over increasingly large areas of space.
-These spatial and feature transformations are occurring together: the integrating filters span both the the spatial representation and a range of features, just like the colour filters we saw in the retina.
-The feature transformations are a lot more complex than this, and don't make neat parts like this. We'll look at that later.
-The spatial integration is much simpler, so we'll look at that first.

Spatial integration up the hierarchy

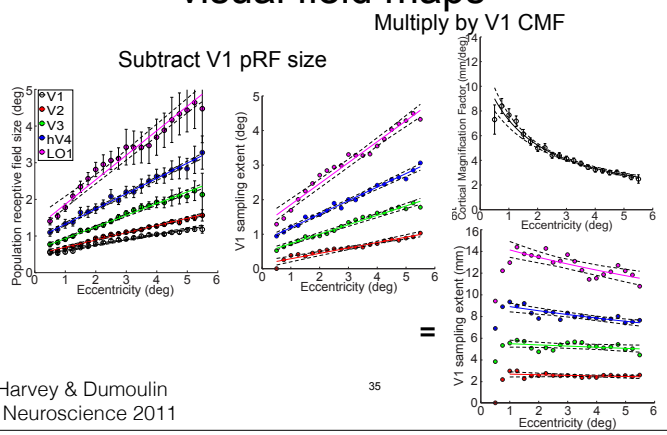


Harvey & Dumoulin
J Neuroscience 2011

34

As we already saw, receptive fields also get larger moving from the central to the peripheral visual field.
-As we move from early visual areas to mid-level shape processing, we see larger and larger receptive fields: the neurons respond to larger and larger areas of visual space following repeating convolution, which gives spatial integration.
-Later areas are no longer sampling from the retina or the visual image: their input is always coming through V1.
-Because a higher layer is inheriting much of its receptive field size from its inputs, any questions about spatial integration must take this into account. Again we must think how the inputs are transformed, not think in terms of the original image.

Spatial sampling between visual field maps

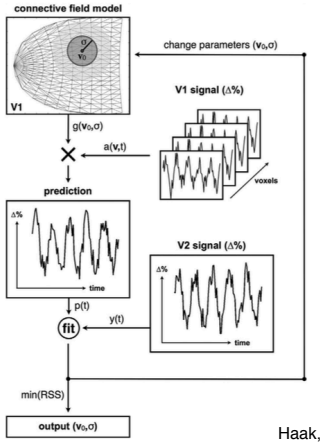


Harvey & Dumoulin
J Neuroscience 2011

35

So let's first subtract the receptive field sizes in V1 from those in other areas, reflecting the assumption that these later areas might sample primarily from V1.
Of course, this is an oversimplification: V3 gets some of its input from V2, but not all of it. V3 is also directly connected to V1. This is not a simple hierarchy.
If we subtract out V1's receptive field size, we see the change in receptive field size from V1, and how this changes from the central visual field to the periphery.
But V1 strongly over-represents the central visual field, which is quantified as the cortical magnification factor: how many millimeters of cortex process each degree of the visual image.
-If we multiply the change in receptive field size by this cortical magnification, we can convert the change in receptive field size, in visual image space, to an extent of spatial integration in cortical millimeters.
-This reveals that the spatial extent of the filter sampling from V1 to V2, V3 and maybe V4 is constant in cortical distance, though not in distance in the visual image.
-The spatial extent of V2's sampling of V1 is about 2mm, which corresponds closely to the size of a hypercolumn (a piece of cortex containing all feature maps for the same image location), so likely allows V2 to neurons to sample from the full range of V1 feature maps.
-So, an artificial network uses a common spatial extent of convolution filter across the whole feature map.
-But in a biological deep network, the feature map will be biased towards areas that are most useful to us, like the centre of vision. Within that biased representation, there also appears to be a common spatial extent of sampling.

Sampling between cortical areas



Haak, Winawer, Harvey
et al., (2012) *Neuroimage*

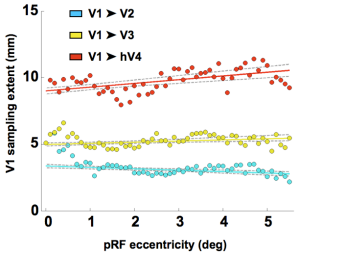
36

We can also ask about this sampling more directly.
If we have a group of measurements across the surface of V1, we can summarise a measurement in V2 as the sum of a group of spatially neighbouring V1 signals.
If we change the position and spatial extent of the group of V1 recording sites, we get different predictions of each measurement from V2. If we find the prediction that best fits the measured V2 signal, this tells us the location and size of the V1 area that the V2 recording site samples from.

Different papers same cortical extent of integration found

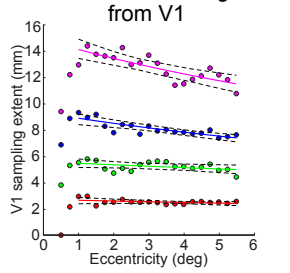
Sampling between cortical areas

Cortical extent of integration from V1



Haak, Winawer, Harvey
et al., (2012) *Neuroimage*

Cortical extent of integration from V1



Harvey & Dumoulin (2011)
J Neuroscience

- Receptive field: Sampling extent of the input image (in visual angle)
- Connective field: Sampling extent of the previous visual field map (in mm)

37

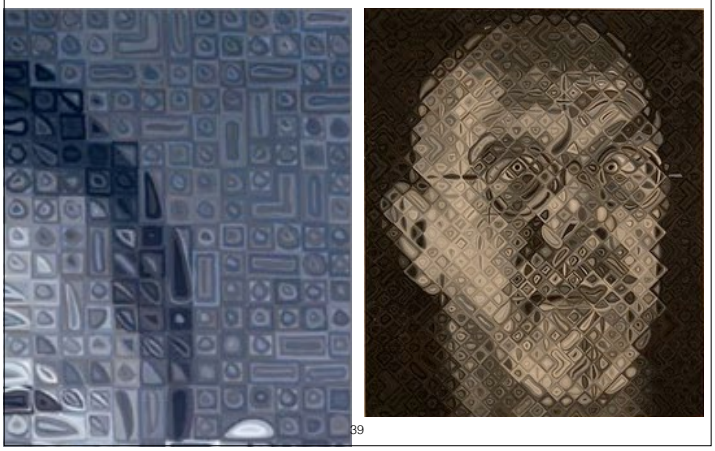
-This also shows that V2, V3 and V4 each sample from a constant cortical extent of V1 throughout the whole visual field map.
-Despite very different approaches, the sampling extents revealed are very similar.
-So this sampling extent is effectively the spatial extent of the filters operating between the feature maps in these different visual field maps.
-In artificial deep networks, this spatial extent of the filters between different layers is often called a 'receptive field'.
-But this is inaccurate when we think about the brain. A receptive field is a part of the visual image that feeds into a neuron. The part of the previous visual field map that feeds into a neuron is called a 'connective field'.
-In the brain, these are in very different units, degrees of the retina and millimeters of cortex respectively, so are quite different concepts.
-In artificial networks, they are also different, because the input image is in pixels while the hierarchical layers are in neural network units.

Spatial integration through the visual hierarchy

- Visual cortical areas beyond V1 also map the spatial visual field onto the cortical surface
  - Spatial relationships between image locations are maintained
- Multiple branching hierarchies of these visual field maps
  - Performing different tasks
- Main division is into ventral and dorsal streams
  - Focussed in temporal and parietal lobes respectively
  - Involved in object recognition and spatial perception/action planning
- Visual field maps remain biased towards central vision
  - Later visual field maps sample from approximately constant cortical areas of earlier visual field maps, regardless of visual position represented

38

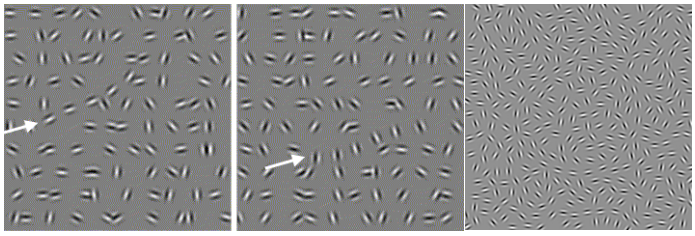
Feature transformations: orientation to objects



39

The other aspect of sampling filters is how they transform feature representations from simple features, like edges, to high-level features that reveal object identity
-This is the pattern within the filter rather than its spatial extent
-How do we get from this simple representation of oriented edges to a holistic impression of the face?
-Here we see a face broken down into its component edges. Using these, we can still see the face.
-The artist who made these pictures, Chuck Close, has a neurological problem in recognising faces (prosopagnosia), in common with around 2% of people. He relies on simpler features to recognise people, as his painting shows

Grouping edges to form objects

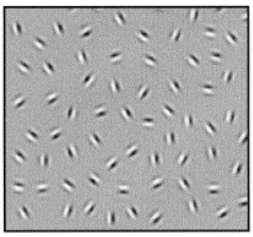


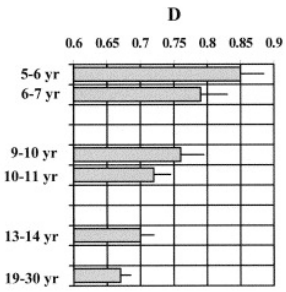
Linear 'snake' path | Linear 'ladder' path | Circular 'snake' path

40

The first important step between early visual processing and object processing is grouping edges.
-Here we see some possible groups, from simple lines of extended edges to more complex shapes like circles.

Grouping edges to form objects



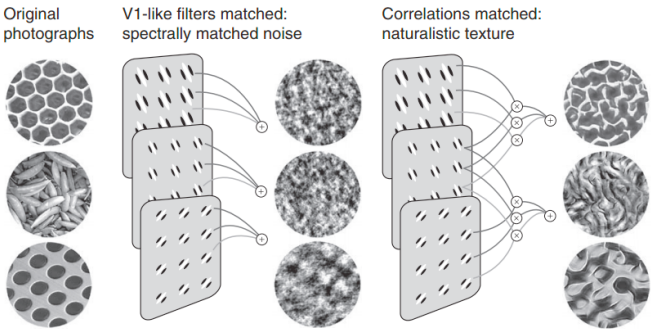


- Not present at birth
- Develops slowly as we learn statistical correlations present in visual images

41

Here we see that our ability to integrate and detect the groups of edges into a pattern increases with age until surprisingly late
-Here sensitivity increases into the 20's.
-We have already seen that colour opponency, surround suppression and orientation detection seem to be hard-wired.
-It is also likely that the spatial extent of filters is hard wired, as there are limits to how large a neuron's dendritic tree can grow to sample from its inputs.
-But beyond V1, it seems that filter **structure (weights)** must be learned.
-This learning does not depend on backpropogation of error. It depends instead on developing responses to patterns that we have seen before, though Hebbian learning.
-So this depends on the statistical structure of patterns of orientation that we have seen before.

What do later visual areas do? (2013 version)



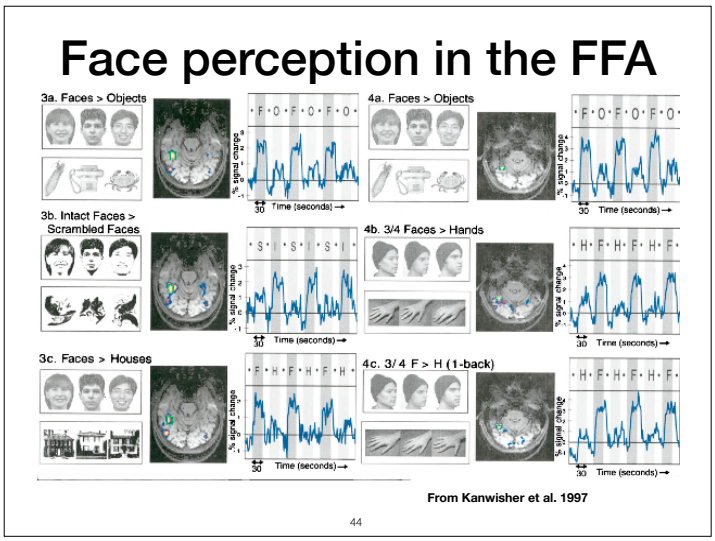
42

Beyond V1, it has even been hard to figure out what V2 does. Generally, it responds very similarly to V1 when presented with oriented edges.
-Because V1 only responds to the spatial frequencies and orientations in its input, we can make images with the same distribution of spatial frequency and orientation as a natural image, and V1 will respond equally well to both.
-However, V2 responds more strongly to natural images than these noise patterns.
-To make similar responses in V2, the image needs to have a similar pattern of local correlations of orientation as the natural image.
-This appears to be because such patterns of local correlations are common in natural images that have trained the filters linking V1 to V2
i.e. it responds when inputs have the correlations that are common in the real world. By V2, if it fires together, it wires together.
-Orientation-selective responses in V1 were first discovered in 1959. It took until 2013 to reveal how these were transformed by V2. Before 2013, researchers tried to explain V2's responses with reference to features in the input image, rather than patterns in V1's output.
-This highlights the limitations of human thinking about feature transformations: we have a bad habit of thinking in terms of the input image, rather than the pattern of activity in the previous layer.
-Beyond V2, we still don't have a good feeling for what drives responses, but we now understand it is likely to be a feature transformation from the outputs of previous network layers.
-It becomes very difficult for experimenters to think about neural representations and transformations beyond this.
-If we want to make a hypothesis about what later levels are doing computationally, the answer is likely to be in deep network terms.
-So deep networks become a very useful tool for guiding our hypotheses and interpreting experimental results.

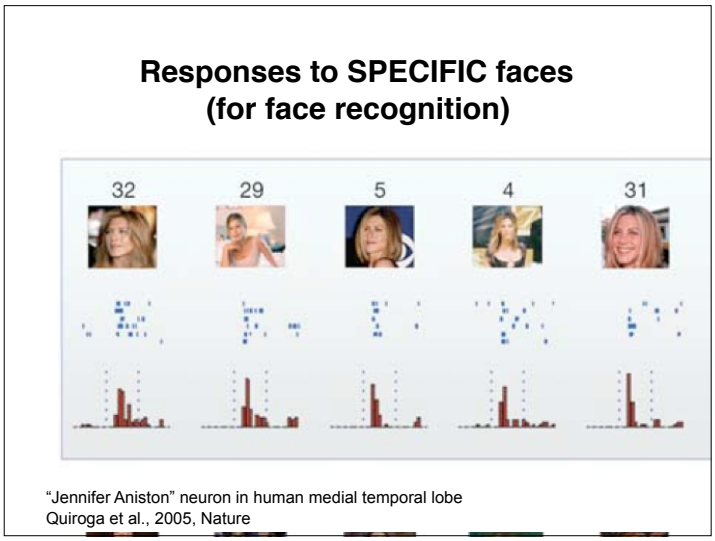
Object selective responses

43

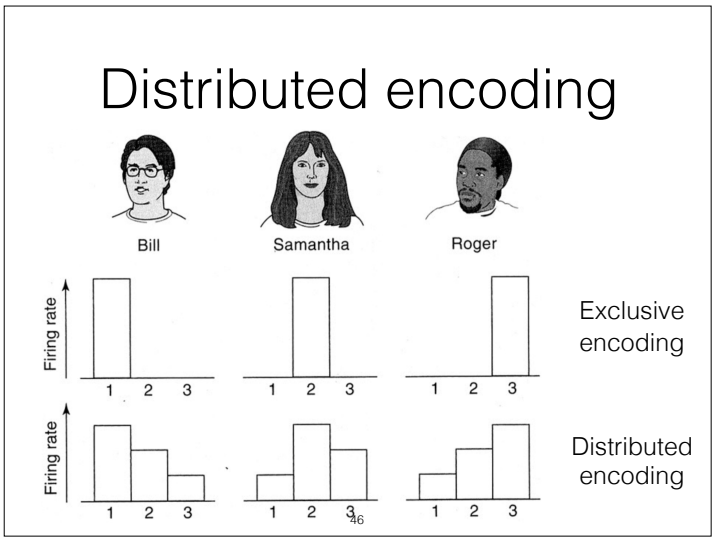
Much further up the processing hierarchy, we eventually find object-selective areas.
A common model for human object-selective responses is face processing



Here we see human face-selective cortex with fMRI.
It responds more to faces than various other object types



Looking at a whole brain area will group responses to ALL faces.
-But within this area, there are even cells that respond more strongly to specific faces, importantly regardless of the image used and its activation of the retina.
-This neuron responds strongly to Jennifer Aniston, from any angle, size and position.
-It responds less to Julia Roberts, even where she is at a similar angle, size and position as Jennifer Aniston.
So here the response is not driven by what the image looks like, but what it contains.



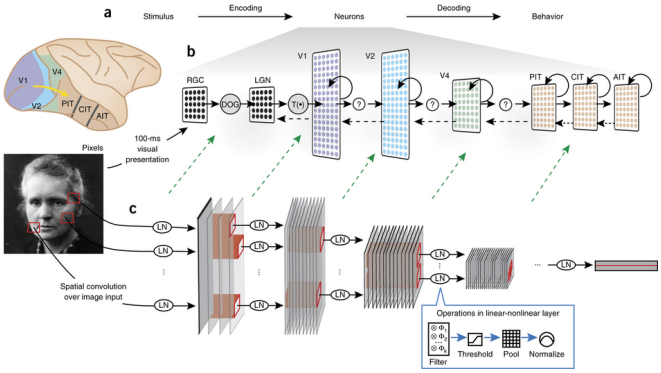
But, given the complex nature of feature map representations, perhaps we would not expect a single neuron to respond so cleanly.
-There is no need for information about a single object to be held by a single neuron
-Perhaps the pattern of activity across a larger group of neurons is used to represent the middle layers, and even the higher layers
-Then, the object identity is not straightforwardly reflected in the activity of a single neuron, but the pattern of activity in a larger population of neurons.

Advantages of distributed encoding

- Allows some cell death without representation failing (graceful degradation)
  - Most of the pattern remains
- Allows new patterns to be stored without new cells
  - New objects can be stored
  - A fixed group of cells can store a variable number of objects
- Consistent with measured cell properties
  - Rarely all-or-nothing responses
- Disadvantage: Harder for humans to understand

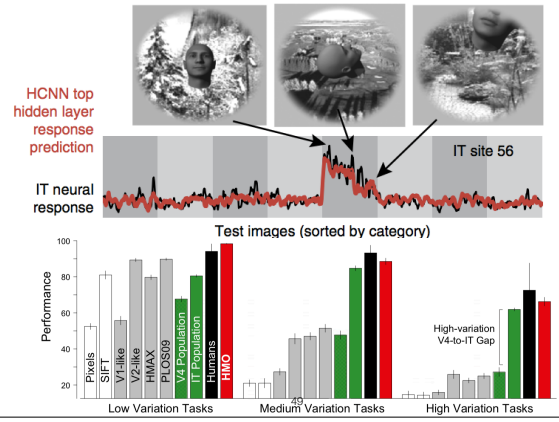
An important thing to note about this distributed representation is that the last layer of an artificial DCNN for object recognition is fully-connected.
So it makes a decision based on the pattern of responses across all of its top-layer units.
In a computation like this, a distributed pattern will allow the right decision just as easily as the response of a single cell.

Making object-selective responses with deep neural networks



Although we don't have a good feeling for what is happening in the middle layers, we believe that they are using deep network structure to transform visual image representations to object representations.
-If we compare the brain's object recognition network to a deep network, we see that there are many similarities in the spatial representation of features, and the spatial sampling between layers.
-Early layers do simple edge detections.
-Later layers just form associations between the output of earlier layers, based on previous experience of correlated response patterns.
-We understand the processing step at each level, just drawing associations in the activity of previous levels
-But it's hard to intuitively understand how this will work over many levels, which may be why it has taken us so long to understand.
-Despite that, we can simulate this process and see how the simulation performs.

Making object-selective responses with deep neural networks



Responses of later network layers closely resemble responses of face-selective cells.
-This is our current best theory of how object-selective responses arise.
-Importantly, we see a very similar response to faces even if the images vary greatly in position, size and viewing angle.
-In image sets with low variation of position, size and viewing angle, many simple computational models have high object recognition performance, comparable to human object classification abilities. V1-like models and V2-like models do well here.
-Once we introduce more variation in position, size and viewing angle, the population of IT neurons does much better than a population of V4 neurons. And a deep network (HMO, hierarchical modular organisation) does much better than the simple models in gray. Indeed, humans, IT neurons, and deep network models all do well here.

Face processing in DCNNs

Face processing is also an exciting new application of artificial DCNNs.

'Deepfakes' are videos where a DCNN is used to map one person's face onto another. The result is a very convincing video of the target face identity following the actions of the source face.

In this example, Freddie Mercury's face is mapped onto actor Remi Malek, who recently played Freddie in a movie using only a fake moustache and teeth.

It helps that Remi already looks a little like Freddie. Everything but the face remains unchanged, and these are Remi's ears and hair.

The deepfake process involves first showing the network a large training set of videos of the target face. The network then maps the features of the face and their movements, and matches those to the features of the source actor.

Feature transformations through the visual hierarchy

- The ability to group edges together into more complex patterns continues to develop into adulthood
  - It relies on learning, although earlier edge detection steps may be hard-wired
- Transformations find commonly-seen patterns in activity of earlier layers
  - Have been difficult for human experimenters to recognise
- Later stages are likely doing the same computation, but from more abstracted inputs (i.e. the outputs of earlier transformations)
  - But it's really hard to think about these representations and transformations
  - So we simulate them with artificial DCNNs
- Later in the ventral stream, face-selective neurons are found
  - Artificial DCNN simulations produce impressively similar results
  - Artificial DCNNs can convincingly manipulate facial identity

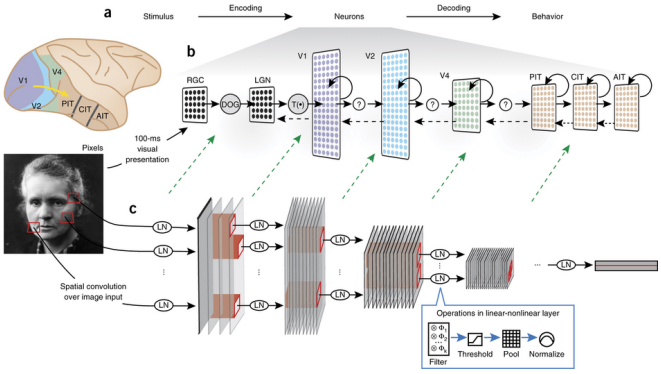
Now read this:

Kay KN, Naselaris T, Prenger RJ, Gallant JL (2008) Identifying natural images from human brain activity. Nature, 452 (7185): 352-355.

Now read this:

Yamins, D. L., H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert and J. J. DiCarlo (2014). "Performance-optimized hierarchical models predict neural responses in higher visual cortex." PNAS 111(23): 8619-8624.

Making object-selective responses with deep neural networks



Making object-selective responses with deep neural networks