Vision

Introduction

This lecture is about vision, it will discuss how **simple cells** in **V1** are modelled and how their behaviour may be explained by sparseness.

The visual pathway

The visual system starts at the eye, where photons are detected and some denoising occurs; the optical nerve then carries the information to the thalamus, in the very centre of the brain, there it is further processed, compressed, and denoised before being relayed on to the visual cortex, at the very back of the cortex. It is processed in stages in the cortex with the information being passed forward, as objects are recognised the information fans out and is integrated with other signals, from memory, from other sensory modalities and other aspects of our cognition. The basic pathway is shown in a very old drawing in Fig. 1 and is summarised in Fig. 2. One notable aspect is that different sides of the brain deal with different sides of the visual field, so signals from the left sides of the retina of both eyes go to the right side of the brain and signals from the right sides so to the left side.

Light is detected at the retina. The retina is a surprising organ in that it is backwards compared to how you'd expect it to be organised; the layer with light detectors is at the back instead of the front. Leaving that aside though, light is detected in specialised cells called **photoreceptors**. These don't spike, but they do convert light into electrical activity. There are two types of photoreceptors: rods, which are important for vision in low light, and the cones, which are responsible for colour vision and important for vision in normal lighting conditions. The electrical activity of the photoreceptors is passed forward through **bipolar cells** to **ganglion cells**.

Ganglion cells aggregate activity from a number of photoreceptors, along with activity from some inhibitory cells in the intermediate layer and their axons form the optic nerve, carrying information to the thalamus. A sketch of the retina is given in Fig. 3 and the uneven distribution of cones and rods across the retina is illustrated in Fig. 4.

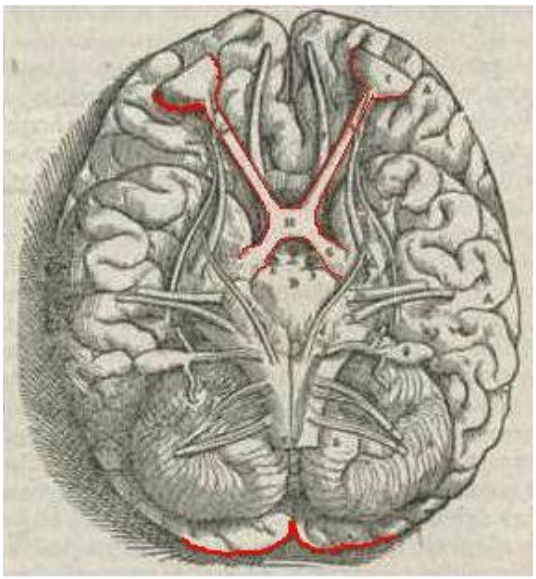


Figure 1: The visual pathway. This is an old drawing due to the sixteenth century Belgian anatomist Andreas Vesalius taken from his influential 1543 textbook *De Humani Corporis Fabrica*. In red are marked the retina, the optic nerves, the thalamus where they cross and the primary visual cortex (V1). [Image from Wikipedia].

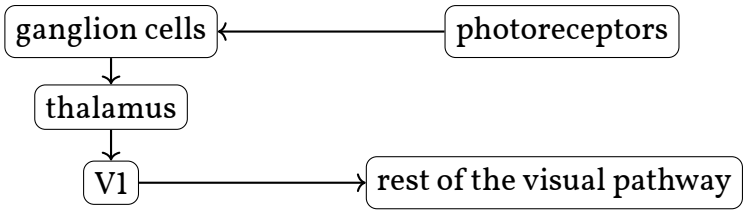


Figure 2: The visual pathway. This is a very rough diagram showing the visual pathway; V1 is the first visual area in the cortex.

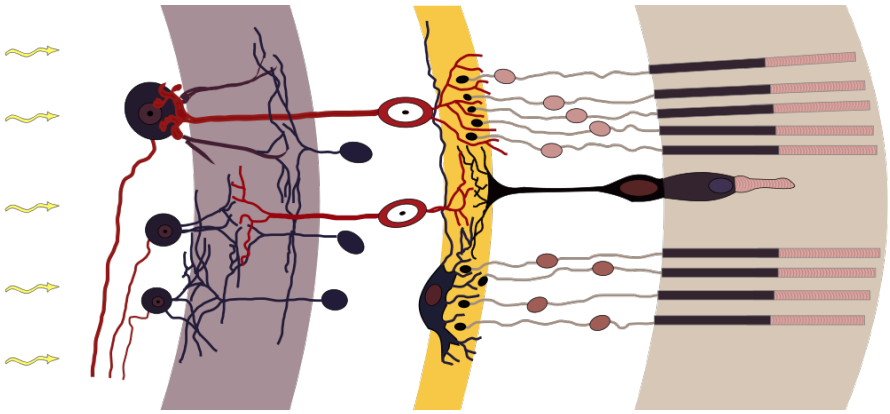


Figure 3: Rods, cones and nerve layers in the retina. The front of the eye is on the left. Light (from the left) passes through several transparent nerve layers to reach the rods and cones (far right). A chemical change in the rods and cones send a signal back to the nerves. The signal goes first to the bipolar and horizontal cells (middle yellow layer), then to the amacrine cells and ganglion cells (left-most purple layer), then to the optic nerve fibres. The signals are processed in these layers. First, the signals start as raw outputs of points in the rod and cone cells. Then the nerve layers identify simple shapes, such as bright points surrounded by dark points, edges, and movement. (Based on a drawing by Ramón y Cajal, 1911.) [Caption and drawing taken from Wikipedia: Cajal derivative work: Anka Friedrich via Wikimedia Commons]

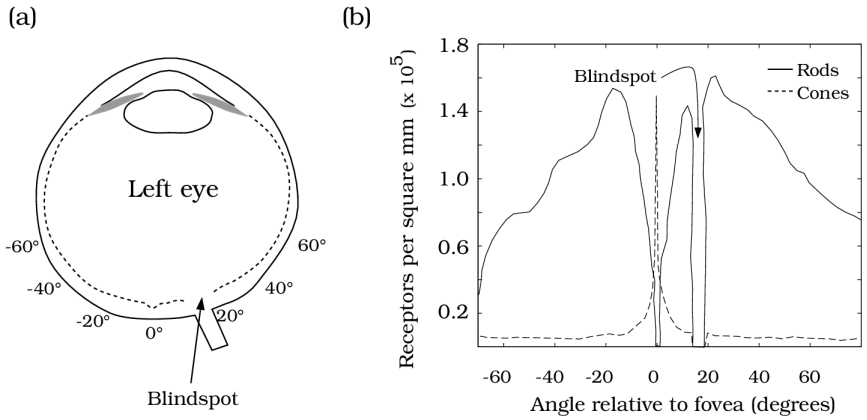


Figure 4: The distribution of rods and cones in the retina. [Image from [1]]

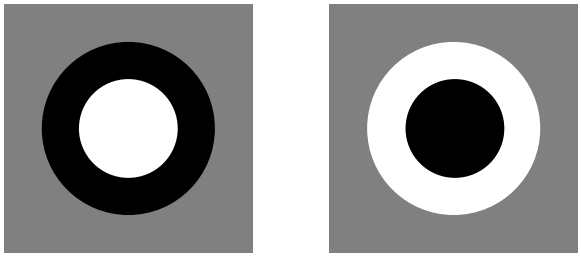
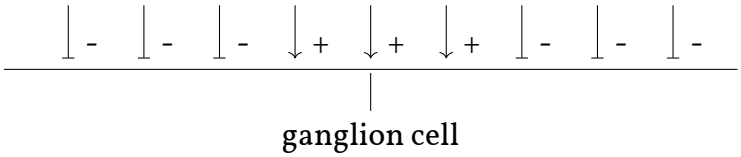


Figure 5: On and off cells respond to small contrast patches.

Receptive fields

Receptive fields are often described as the stimuli giving the largest response from a neuron. For ganglion and thalamic cells these are contrast patches, see Fig. 5, in **on-cells** small patches of the visual field where an illuminated region surrounded by an unilluminated one causes firing, different cells will respond to different locations. The width of the receptive fields vary from the size of full stop at reading distance in the center, to the size of a page near the periphery. In **off-cells** the contrast is reverse, the cell responds to an unilluminated region surrounded by an illuminated region. In practice these patterns are the result of excitatory and inhibitory synapses relaying information from these regions of the visual field.



In V1 there are cells called **simple cells** and cells called **complex cells**; we will concentrate on the simple cells, these have edge-like receptive fields; different cells respond to particular orientations in particular locations in the visual field.

The edge-like receptive fields in V1 were first discovered by Hubel and Wiesel [2]. They used an electrode to record from V1 neurons in anaesthetised cat; they moved an edge-like stimulus around until they found the position that caused the highest firing rate, they observed that the firing rate depended on orientation as well as position, see Fig. 7 and Fig. 8.

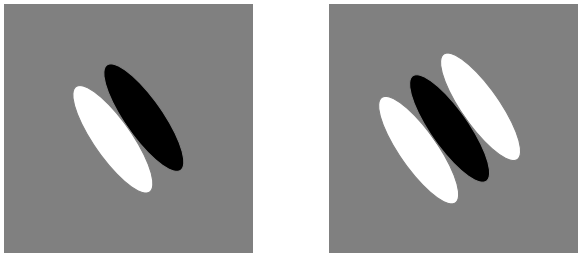


Figure 6: Simple cells in V1 respond to edges.

Linear models

One way to think about it is to imagine the entries in the receptive field are synapse strengths for inputs from cells responding to illumination at points in the visual field. To formalize this consider linear models of the neuron's activity. Let I_{ij} denote the illumination level at point (i, j) in the visual field, i and j are discrete coordinates, for simplicity we will treat everything discretely. Now, imagine a linear model of the activity of the neuron, with the firing rate depending linearly on the illuminations; leaving out any messing with the firing rate having to be positive, this means

$$\tilde{r} = r_0 + \sum w_{ij} I_{ij} \tag{1}$$

where r_0 is the background firing rate and w_{ij} give the receptive field. Of course the firing rate of a neuron doesn't satisfy a linear model but the idea is to choose the linear model which best approximates the neuron, that is, for example, to choose w_{ij} to minimize the average square error

$$\text{average square error} = \langle (r - \tilde{r})^2 \rangle \tag{2}$$

between r, the observed firing rate and $\tilde{r}$ is the estimated firing rate from the linear model.

As an example consider

$$[w_{ij}] = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & -1/8 & -1/8 & -1/8 & 0 \\ 0 & -1/8 & 1 & -1/8 & 0 \\ 0 & -1/8 & -1/8 & -1/8 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{3}$$

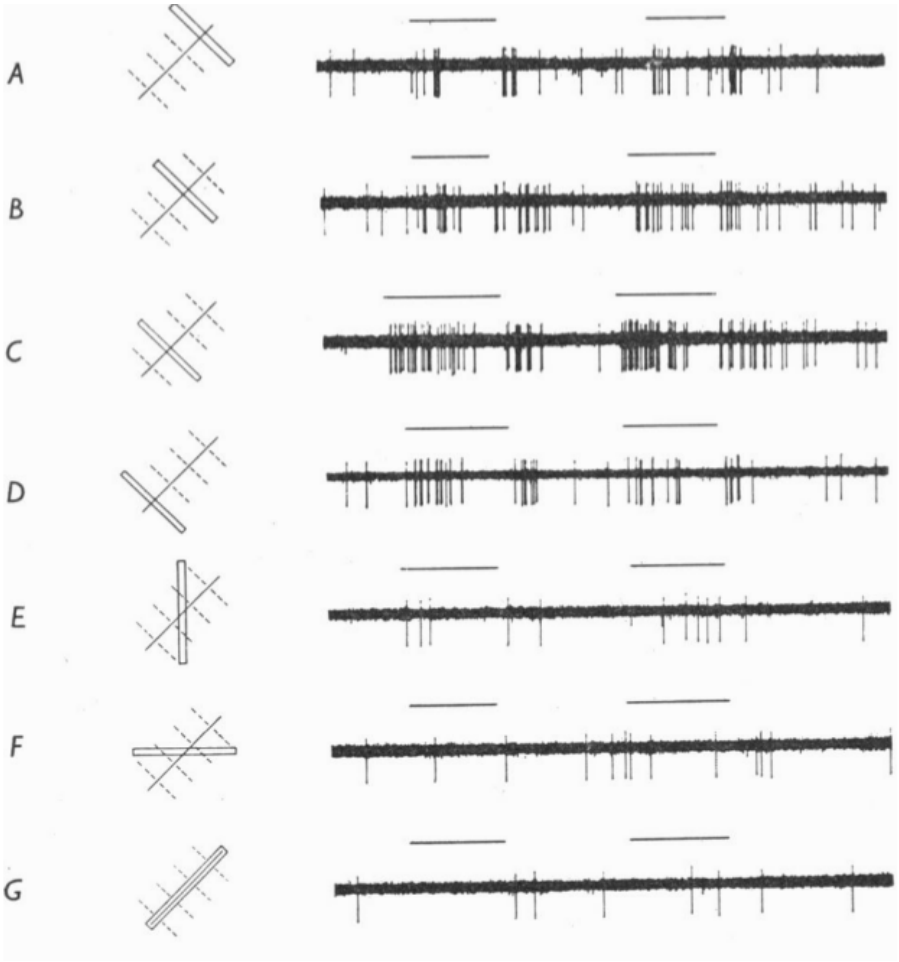
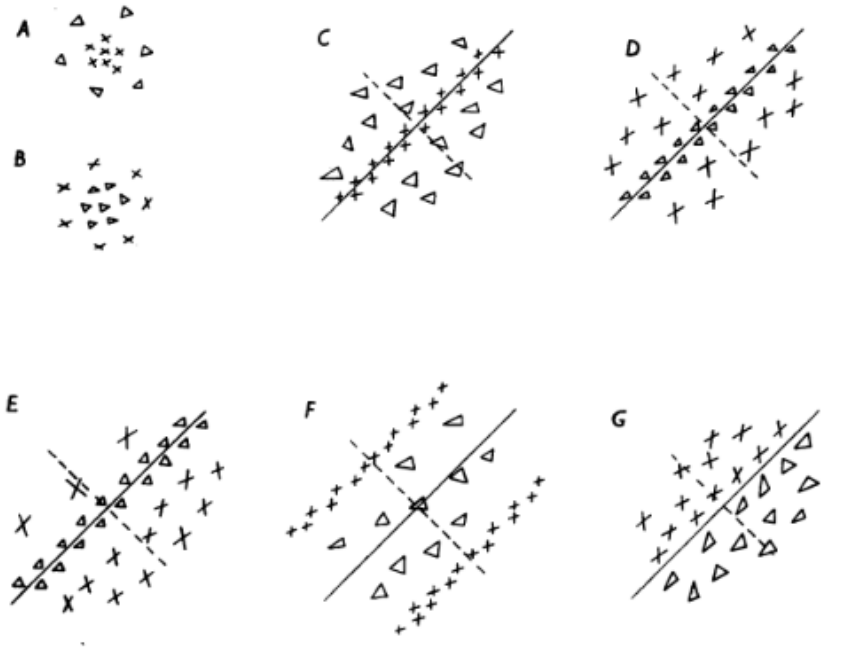


Figure 7: Experimental results from Hubel and Wiesel; the stimulus is a slit that allows light through from a souce, it is $0.125° \times 2.5°$ and is presented during the one-second period marked by the two bars over the plots. In the plots the vertical lines correspond to spikes. [Image from [2]].



Text-fig. 2. Common arrangements of lateral geniculate and cortical receptive fields. *A*. 'On'-centre geniculate receptive field. *B*. 'Off'-centre geniculate receptive field. *C–G*. Various arrangements of simple cortical receptive fields. ×, areas giving excitatory responses ('on' responses); △, areas giving inhibitory responses ('off' responses). Receptive-field axes are shown by continuous lines through field centres; in the figure these are all oblique, but each arrangement occurs in all orientations.

Figure 8: More experimental results from Hubel and Wiesel; here they have mapped out the excitatory (crosses) and inhibitory (triangles) areas for a number of neurons. [Image from [2]].

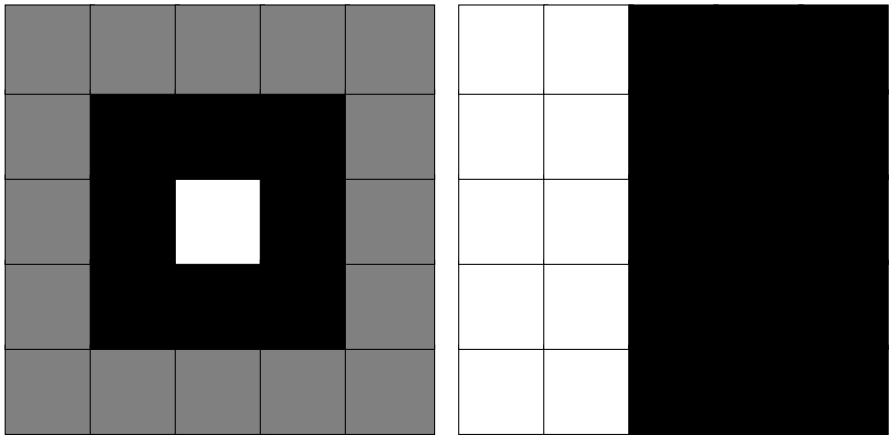


Figure 9: Receptive field and visual stimulus.

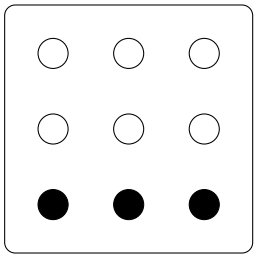
and

$$[I_{ij}] = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \end{pmatrix} \tag{4}$$

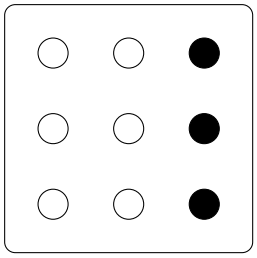
which is like a ganglion cell responding to an edge and is illustated in Fig. 9. If $r_0 = 2$ say then $\tilde{r} = 13/8$.

Features

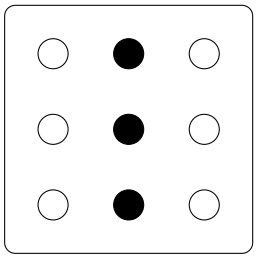
Knowing why V1 receptive fields have the particular stucture they do is likely to tell us something about what it is that the brain does to information in the sensory pathways. One idea is that it is related to feature extraction. To motivate this we will consider a fictitious world of simplified creatures; imagine we are one of these creatures and wish to decide how to react to other creatures we encounter. As in the real world, when we encounter a creature we need to decide between what are sometimes called the three Fs: fighting, fleeing and mating. Now imagine that the creatures all have a three by three pattern on their stomachs:



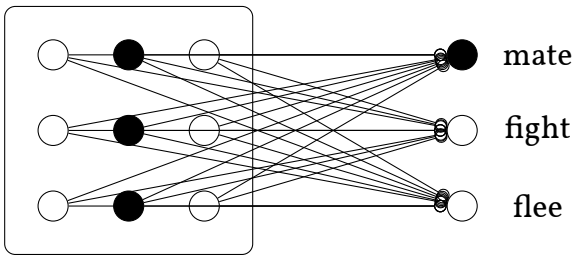
and that creatures to fight have a horizontal strip to the top or the bottom, as above, creatures to flee from, a vertical strip on the left or right, for example



and creatures to mate with, a central line, either horizontal or vertical like:



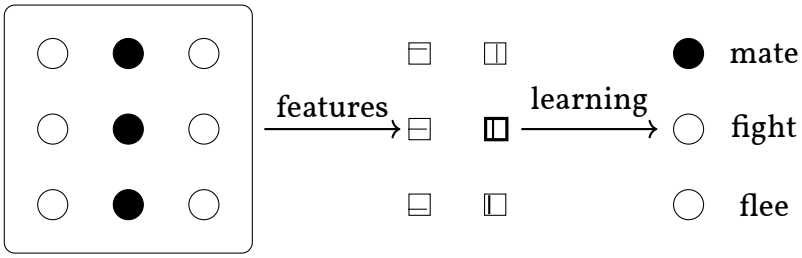
Now, imagine processing this information so as to rapidly decide what to do, the simplest neural network to process the patterns would look like



Clearly this would be very hard to learn; the connection from, say the bot-

tom middle node is on in two of the patterns above despite these patterns corresponding to different types of creature.

A far better strategy would be to first learn features and then learn the association between these features and the creature type. Here the features are clearly the horizontal and vertical bars



In this case the problem has been split in two; first the connections summarized as 'features' above are learned from the data, possibly using the sort of correllation structure learning provided for by STDP, the interpretation of these feature is then learned, this is clearly easier, the connections summarized as 'learning' have a far simpler task, which is good, since it is crucial to learn this sort of salient ecological information quickly.

Feature selection

Here we will consider what properties we would expect features to have. To do this lets imagine that there are neurons that correspond to the features and their activity represents the image. Say the feature=code neurons each have a receptive field and respond linearly and for simplicity we will leave out the background firing rate: for the sth feature neuron

$$a_s = \sum_{ij} w_{ij}^s I_{ij} \tag{5}$$

and conversely, the output can be represented by

$$I_{ij} = \sum_{s} a^s W_{ij}^s \tag{6}$$

Confusing aside - rate versus reconstruction

The slightly confusing thing here is that we are moving between the linear model

$$a_s = \sum_{ij} w_{ij}^s I_{ij} \tag{7}$$

with a_s the firing rate, we have changed to as rather than rs to be consistient with the papers this is based on, and the reconstruction

$$I_{ij} = \sum_s a_s W_{ij}^s \tag{8}$$

which estimates the image using the firing rates a_s.

We do this all the time with vectors:

$$\mathbf{v} = v_1\mathbf{i} + v_2\mathbf{j} + v_3\mathbf{k} \tag{9}$$

is the reconstruction where the corresponding project, for example

$$v_1 - \mathbf{v} \cdot \mathbf{i} \tag{10}$$

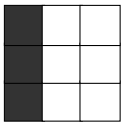
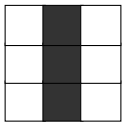
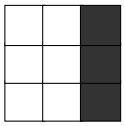
is like the linear model. However, the situation in this case is more straightforward, because the basis vectors are orthonormal

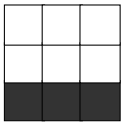
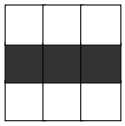
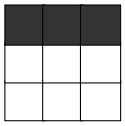
$$\mathbf{i} \cdot \mathbf{j} = \mathbf{j} \cdot \mathbf{k} = \mathbf{k} \cdot \mathbf{i} = 0 \tag{11}$$

and

$$\mathbf{i} \cdot \mathbf{i} = \mathbf{j} \cdot \mathbf{j} = \mathbf{k} \cdot \mathbf{k} = 1 \tag{12}$$

the same basis vector appears in the reconstruction and the projection: the coefficent v_1 of $\mathbf{i}$ in the reconstruction is the projection of $\mathbf{v}$ onto $\mathbf{i}$. However, in the case of vision the basis elements, the W_{ij}^s and w_{ij}^s are not orthonormal and therefore are not the same, working out the relationship between involves vectorizing the matrix indices i and j, so we won't go into it here, morally one is the inverse transpose of the other. In fact, here we will consider an example where the dimensions are different, where the image patches are 3×3 but there are only six features,

$W^1 =$  $W^2 =$  $W^3 =$ 

$W^4 =$  $W^5 =$  $W^6 =$ 

where the almost-black corresponds to one and white to zero, so put another way

$$[W^1_{ij}] = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \tag{13}$$

This means that the reconstructed image may not be equal to the original image and will just be an approximation to it.

Sparseness

The question now is what principle to use to select these features. In considering the mythical creatures we decided that features should be horizontal and vertical lines since that's the pattern that the creatures have. The question is how to find these features in general.

One idea, due to [3, 4], is to use **sparseness**. Continuing with the example with creatures with patterned bellies, looking at flee-from animal, for example, three dots are black so among neurons that code for individual dots three would be active, but among neurons coding for vertical or horizontal lines, only one would be active.

Of course this is a very artificial made-up example, nonetheless it is thought that 'sparseness' is a good way to define features [3, 4]. Very roughly, the fewer neurons needed to reconstruct an image, the more of the image each neuron is coding for; for this to work without having a vast number of neurons covering every possible combination of pixels, the neurons must code for features, pieces of image that occur regularly. The assumption, in short, is that all the images are mostly made of the same few building blocks.

The idea is as follows, let I be a image and $\tilde{I}$ an approximation to that image formed using features

$$\tilde{I}_{ij} = \sum_s a_s W^s_{ij} \tag{14}$$

Now W^s could be under-complete, like above, or over-complete, as it may be in the visual system, or the dimensions could be chosen to match. Either

way, even if the basis is not under-complete, $\tilde{I}$ is an approximation because the a_s are not just chosen to give an accurate reconstruction, but to do so in a sparse way; they are chosen to minimize

$$E = \sum_{ij}(I_{ij} - \tilde{I}_{ij})^2 + \beta \sum_s f(a_s) \tag{15}$$

This has two terms, the first measures the square error between I and $\tilde{I}$, the second is intended as a measure of sparseness, there are different choices possible, one example would be

$$f(x) = \log_2(1 + x^2) \tag{16}$$

Now, just looking at two dimensions $(1, 0)$ gives

$$\sum_s f(a_s) = \log_2(2) + \log_2(1) = 1 \tag{17}$$

whereas the less sparse $(1/\sqrt{2}, 1/\sqrt{2})$, which as a vector is the same length, gives

$$\sum_s f(a_s) = 2\log_2(3/2) = 1.17 \tag{18}$$

which is larger. The β here determines the trade-off between accuracy and sparseness, if β is small the square error is more important, if β is big the sparseness is.

The idea now is to take a corpus of image patches and find the best features, the ones that on average give the lowest values of E. The algorithm proceeds in two stages, for each image patch the a_s are chosen to minimise E basically by numerically solving the system of differential equations

$$\frac{\partial E}{\partial a_s} = 0 \tag{19}$$

Next, using the results of this calculation for all the images in the corpus, the features W_{ij}^s are adjusted

$$W_{ij}^s \rightarrow W_{ij}^s - \eta \frac{\partial \langle E \rangle}{\partial W_{ij}^s} \tag{20}$$

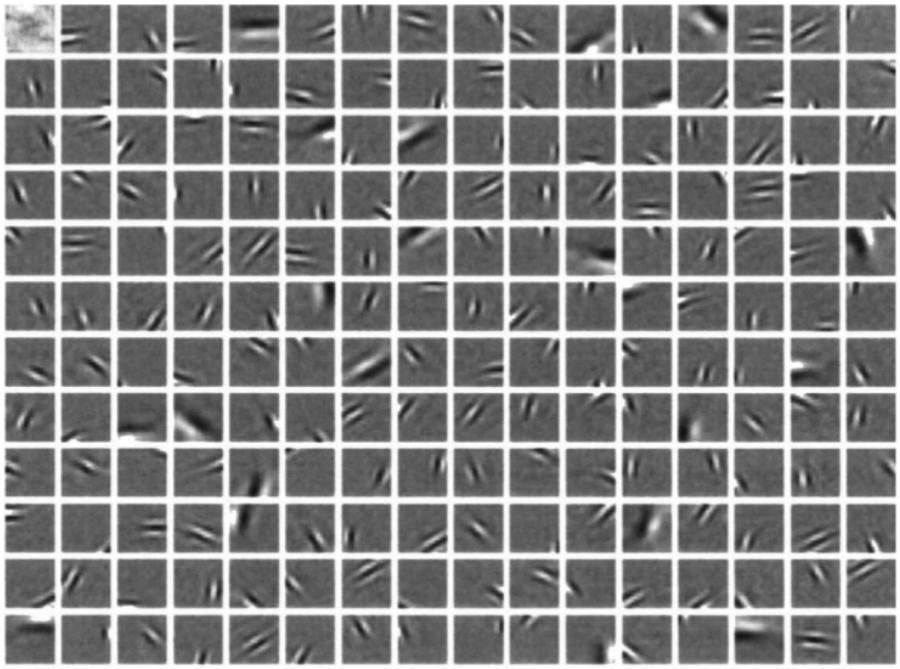


Figure 10: Sparse filters; the optimal features W_{ij}^s using 16×16 image patches cut from a corpus of pictures of the American northwest. [From [3]].

where η is a learning rate and $\langle E \rangle$ this average error. This should reduce the average error in the next run through the corpus. This is repeated until the best features are found. The details of how E is minimized over possible choices of the a_s and how to adjust the W_{ij}^s can be found in [3, 4]. The results are shown in Fig. 10 and, ignoring the complication that these are not actually the receptive fields, they do clearly resemble the receptive fields measured from V1.

As described, none of this seems very biological, the sparse filters were discovered using numerical optimisation routines. However, there are biologically plausible implementations using **Hebbian learning**, see for example [5]. There are other approaches which parallel sparseness as a way of distinguishing features, for example, Infomax, which examines the informativeness of putative features [6].

References

[1] Wandell, Brian A. Foundations of vision. (Sinauer Associates, 1995) foundationsofvision.stanford.edu/.

[2] Hubel DH, Wiesel TN. (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of Physiology 160: 106.

[3] Olshausen BA, Field DJ (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature, 381: 607–609.

[4] Olshausen BA, Field DJ (1997) Sparse coding with an overcomplete basis set: A strategy employed by V1?. Vision Research 37: 3311–3325.

[5] O'Reilly RC, Munakata Y (2000) Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain. MIT Press.

[6] Bell AJ, Sejnowski TJ (1995) An information-maximization approach to blind separation and blind deconvolution. Neural Computation 7: 1129–1159.