Linear Filters:
How can they be learnt?
Two Principles:
Hebbian Learning and Sparsity

Alan Yuille

Why do the Receptive Fields have specific forms?

- The previous lectures described receptive field (weights) which had specific functional forms – e.g., Laplacian of a Gaussian (center-surround), Gabor functions.
- (1) Can biological mechanisms learn the weights of the receptive fields *without* any dependence on natural images? (Why? There is evidence that some receptive fields are learnt before birth). Or with limited assumptions about natural images?
- (2) Can we predict/learn the receptive fields from properties of natural images? Taking into account increasingly sophisticated properties.
- This will introduce some basic statistical properties of images, e.g., shift-invariance.

Two Principles of Neuroscience

- (1) *Hebbian Learning.*
- Hebbs rule is a neuroscience theory for how the synaptic strengths (weights of a neuron) are learnt. It predicts that the weights/strengths increases if the pre-synaptic input and the post-synaptic output are strongly correlated. Informally, “the neurons that fire together wire together”. Note: Hebb is a general principle and we will only describe one simple instantiation.
- (2) *Sparsity.*
- This is a general principle of neuroscience based on the observation that most neurons in the brain are inactive most of the time (Barlow). Informally, neurons fire sparsely. The concept of sparsity is studied by mathematicians and there is a rich and influential theory.

Statistical Properties of Natural Images

- Images are (statistically) shift-invariant. The correlation between the intensity I(.) at positions x and y (i.e. between I(x) and I(y)) depends only on the spatial displacement x-y between them. The correlation will get smaller as the displacement gets bigger.

- This means that learning rules will tend to favor certain types of receptive fields. In particular, there is a bias towards sinusoid functions modulated by Gaussians (the Gaussians impose spatial fall off – sometimes resulting from assumptions about the density of neurons). This gives “side-lobes”.
- Nonlinear models of neurons can be designed which would avoid these side-lobes. This seems to disagree with neuroscience experiments. But the analysis of these experiments also makes linear assumptions, so the side lobes may not really be there!

Others Ways of Learning the Receptive Fields

- More recently, researchers have taken Deep Nets trained on visual tasks – e.g., object recognition. This produces a hierarchy of receptive fields. In the first layer, the receptive fields are linear. In all higher layers the receptive fields are non-linear.
- Researchers compared the weights of the Deep Nets with the responses of the neurons (this was done for object recognition – e.g., J. DiCarlo & Yamins – and also for visual cortex Y. Zhang & T-S Lee).
- These findings suggest that many cells in V1 are better fit by non-linear models (Y. Zhang et al. 2018).
- We will return to this later in the course. Now we will return to linear models of neurons.