Hierarchy

Will Penny

24th March 2011

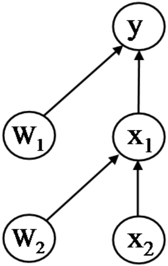
Hierarchical Model

Rao and Ballard (1999) presented a hierarchical model of visual cortex to show how classical and extra-classical Receptive Field (RF) effects could be explained by Bayesian inference in a cortical hierarchy.

$$
\begin{aligned}
y &= W_1 x_1 + e_1 \\
x_1 &= W_2 x_2 + e_2
\end{aligned}
$$

Joint Likelihood

The corresponding graphical model is



The joint likelihood of data and model parameters is

$$p(y, W_1, W_2, x_1, x_2) = p(y|W_1, x_1)p(W_1)p(x_1|x_2, W_2)p(W_2)p(x_2)$$

The first and second level prediction errors are assumed isotropic Gaussian with precisions λ_1 and λ_2. The priors over W_1 and W_2 are zero mean Gaussian, having isotropic covariances with precisions α_1 and α_2.

Joint Likelihood

All parameters W_1, W_2, x_1 and x_2 are learnt by gradient ascent of the relevant part of the the joint likelihood. For example, for x_1 we have

$$L(x_1) = \log[p(y|W_1, x_1)p(x_1|x_2, W_2)]$$

as all other terms do not depend on x_1. Maximising this function will implicitly maximise the posterior probability $p(x_1|y)$.

First Layer Activity

The joint log-likelihood as a function of first layer activity is

$$L(x_1) = -\frac{\lambda_1}{2}(y - \hat{y})^T(y - \hat{y}) - \frac{\lambda_2}{2}(x_1 - \hat{x}_1)^T(x_1 - \hat{x}_1)$$

with image predictions

$$\hat{y} = W_1 x_1$$

and predictions of first layer activity

$$\hat{x}_1 = W_2 x_2$$

$L(x_1)$ comprises precision weighted prediction error terms from both levels.

First Layer Activity

The activity is then updated by gradient ascent

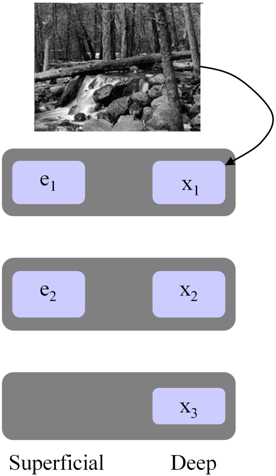
$$\tau_x \frac{dx_1}{dt} = \frac{dL(x_1)}{dx_1}$$

which gives

$$\tau_x \frac{dx_1}{dt} = \lambda_1 W_1^T (y - W_1 x_1) + \lambda_2 (\hat{x}_1 - x_1)$$

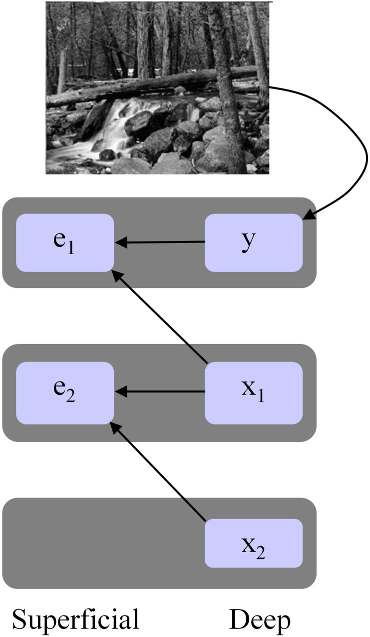
This is the same as online Bayesian learning for linear systems (last lecture). These updates have a simple predictive coding implementation.

Predictive Coding Architecture



Mumford (1992) ”I put forward a hypothesis on the role of the reciprocal, topographic pathways between two cortical areas, one often a ‘higher’ area dealing with more abstract information about the world, the other ’lower’ dealing with more concrete data. The higher area attempts to fit its abstractions to the data it receives from lower areas by sending back to them from its deep pyramidal cells a template reconstruction best fitting the lower level view. The lower area attempts to reconcile the reconstruction of its view that it receives from higher areas with what it knows, sending back from its superficial pyramidal cells the features in its data which are not predicted by the higher area. ”

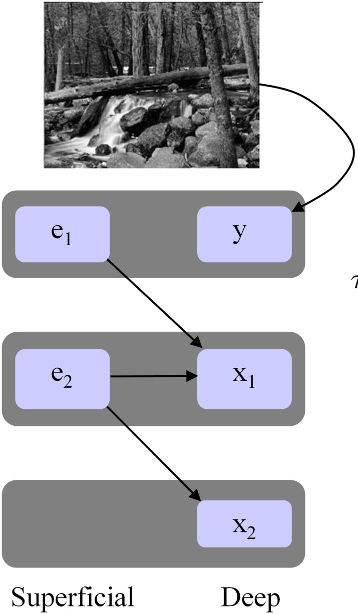
Update Error Units



$$e_1 = y - W_1 x_1$$

$$e_2 = x_1 - W_2 x_2$$

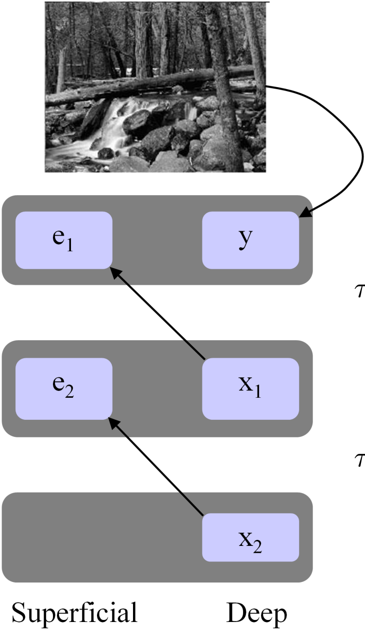
Update Causal Units



$$\tau_x \dot{x}_1 = \lambda_1 W_1^T e_1 + \lambda_2 e_2$$

$$\tau_x \dot{x}_2 = \lambda_2 W_2^T e_2$$

Update Connectivity

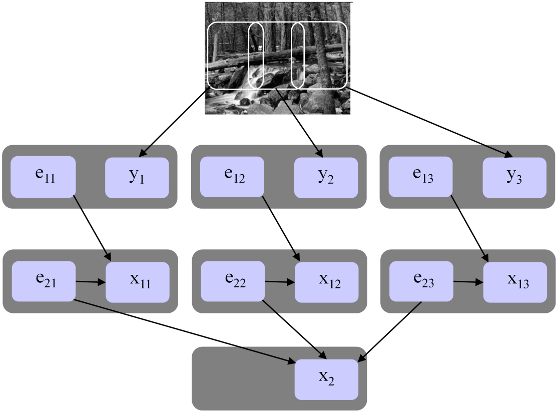


$$\tau_w \dot{W}_1 = \lambda_1 e_1 x_1^T - \alpha_1 W_1$$

$$\tau_w \dot{W}_2 = \lambda_2 e_2 x_2^T - \alpha_2 W_2$$

Increasing Receptive Field Size

In Rao and Ballard (1999) a Gaussian weighting was applied to the input of each first level node, so that it only sees a localised portion of the input image.



Each second level unit effectively sees the whole image.

Increasing Receptive Field Size

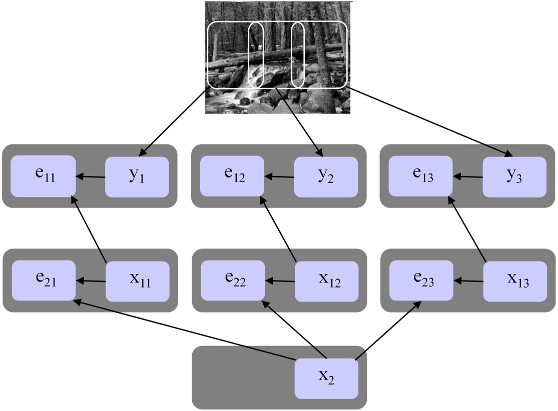
The images y (5 shown) are modelled with $j = 1..3$ first level modules with overlapping receptive fields.

$$
\begin{aligned}
y_j &= W_{1j} x_{1j} + e_1 \\
(256 \times 1) &= (256 \times 32)(32 \times 1) + (256 \times 1)
\end{aligned}
$$

Each module is a linear expansion of a basis set W, with coefficients x which are different for each image. Each module predicts activity in a 16×16 pixel patch. One can think of the ith row of W_{1j} as the projective field of the ith neuron in the jth module. And the ith entry in x_{1j} as the activity or firing rate of the ith neuron in the jth module. These coefficients are then constrained by a second level model

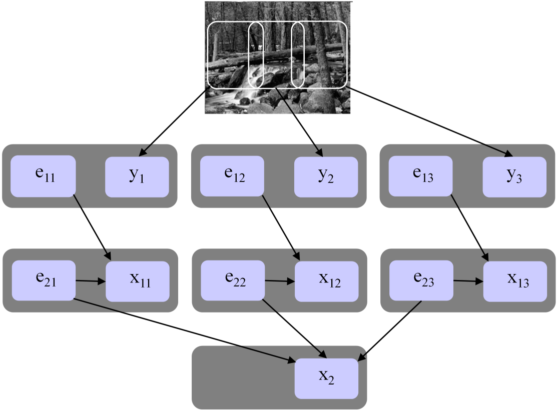
$$
\begin{aligned}
x_1 &= W_2 x_2 + e_2 \\
(96 \times 1) &= (96 \times 128)(128 \times 1) + (96 \times 1)
\end{aligned}
$$

Updating Error Units



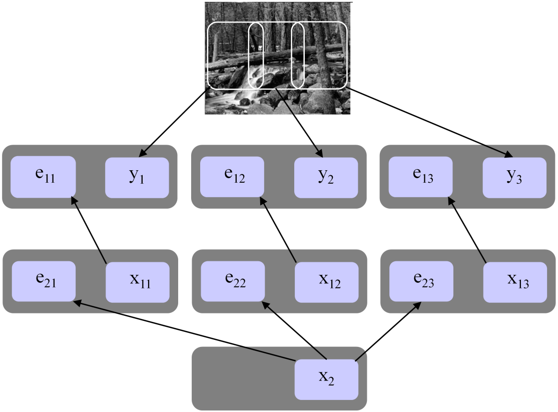
$$
\begin{aligned}
e_{1j} &= y_j - W_{1j} x_{1j} \\
e_{2j} &= x_{1j} - W_{2j} x_2(j)
\end{aligned}
$$

Updating Causal Units



$$\tau_x \dot{x}_{1j} = \lambda_1 W_{1j}^T e_{1j} + \lambda_2 e_{2j}$$

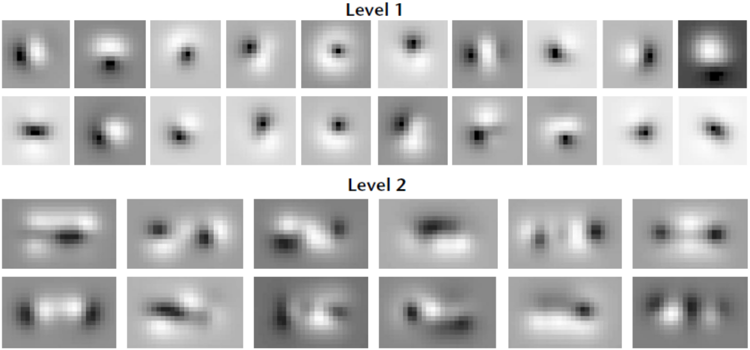
Updating Connectivity



$$\tau_w \dot{W}_{1j} = \lambda_1 e_{1j} x_{1j}^T - \alpha_1 W_{1j}$$

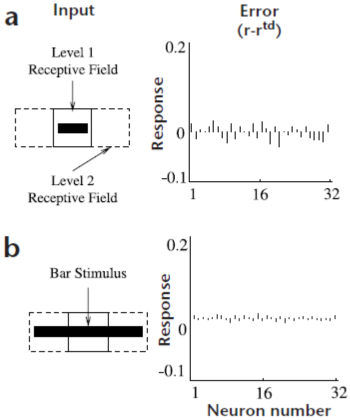
Receptive Fields

Level 1 receptive fields are reminiscent of Difference-of-Gaussian (DOG) filters that have been used to model simple-cell RFs in primary visual cortex.



Level 2 cells respond to more complex features.

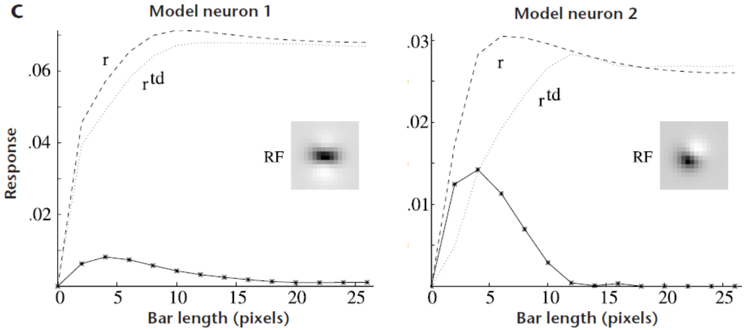
End-Stopping



The response of level-1 prediction error units diminishes for bars that go beyond the end of each level-1 receptive field. So-called *end stopping*.

End-Stopping

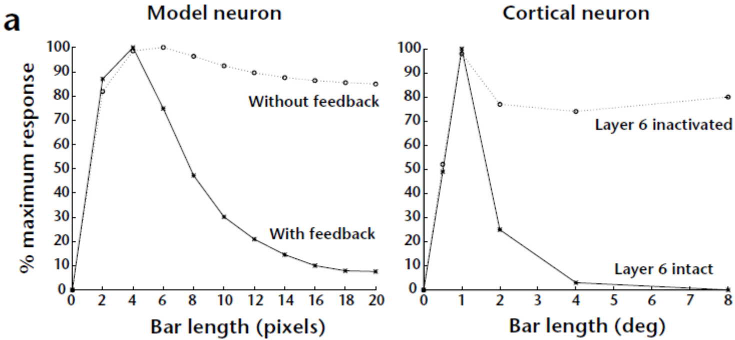
Error unit activity naturally arises from difference in bottom-up activity and top down predictions.



The network was trained on natural images for which short bars seldom occur in isolation. Short bars are generally part of longer bars.

End-Stopping

End Stopping disappears if feedback in the model is disabled or if layer 6 activity is inactivated in squirrel monkey (Sandell and Schiller, 1982).



Nonlinear models

Friston (2003) considers nonlinear hierarchical models of the form

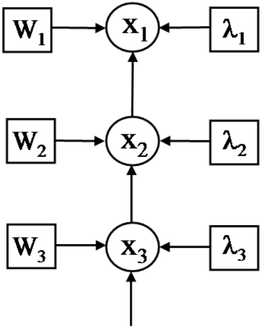
$$\begin{aligned} x_1 &= g(x_2, w_1) + e_1 \\ x_2 &= g(x_3, w_2) + e_2 \\ .. &= .. \\ x_{R-1} &= g(x_R, w_{R-1}) + e_{R-1} \end{aligned}$$

where $y = x_1$ is the observed data, $g(x_{i+1}, w_i)$ is some nonlinear function of hidden causes x_{i+1} and parameters w_i, and e_i is zero mean additive Gaussian noise with covariance C_i and i indexes the level in the hierarchy. C_i is parameterised by λ_i.

These equations embody 'structural priors'. The generative model is not dynamic. The recognition model is.

Generative Model

There are no priors over w or λ.



The joint probability of activities in all regions $r = 1..R$ is therefore

$$
\begin{aligned}
p(x|w,\lambda) &= \prod_{i=1}^{R} p(x_i|x_{i+1}, w_i, \lambda_i) \\
&= \prod_{i=1}^{R} \mathrm{N}(x_i; g(x_{i+1}, w_i), C_i)
\end{aligned}
$$

Joint Log Likelihood

The joint log likelihood is

$$L(x, w, \lambda) = \sum_{i=1}^{R} \log p(x_i | x_{i+1}, w_i, \lambda_i)$$

This can be written as (dropping constant terms)

$$L(x, w, \lambda) = \sum_{i=1}^{R} \left(-\frac{1}{2} e_i^T e_i - \frac{1}{2} \log |C_i| \right)$$

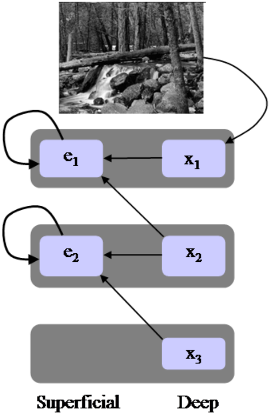
where the prediction errors are given by

$$e_i = C_i^{-1/2} [x_i - g_i(x_{i+1}, w_i)]$$

The hidden causes, parameters and variance components can be estimated using a gradient ascent scheme to find their MAP values.

Error Units

In the univariate case, if the error covariances have the form $C_i^{-1/2} = 1 + \lambda_i$ then the prediction errors can be written. The last term acts as a decay with time constant λ_i.



Hence

$$e_i = (1 + \lambda_i)^{-1} [x_i - g(x_{i+1}, w_i)]$$

Rearranging gives

$$e_i = [x_i - g(x_{i+1}, w_i)] - \lambda_i e_i$$

Hidden Units

The joint log-likelihood is

$$L(x) = \sum_i -\frac{1}{2} e_i^T e_i + \ldots$$

where

$$e_i = (1 + \lambda_i)^{-1} [x_i - g(x_{i+1}, w_i)]$$

Hence

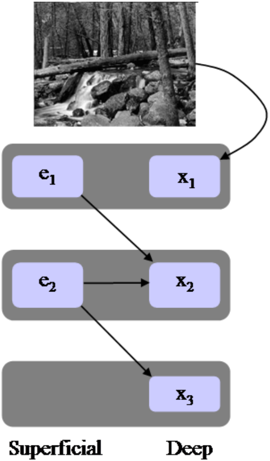
$$L(x_i) = -\frac{1}{2} e_{i-1}^T e_{i-1} - \frac{1}{2} e_i^T e_i$$

so

$$\frac{dL(x_i)}{x_i} = -\left(\frac{de_{i-1}}{dx_i}\right)^T e_{i-1} - \left(\frac{de_i}{dx_i}\right)^T e_i$$

Hidden Units

For the hidden causes we have

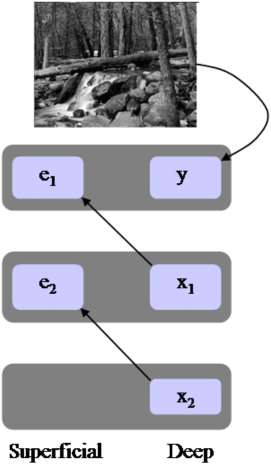


$$\tau_x \dot{x}_i = \frac{dL}{dx_i}$$
$$= -\left(\frac{de_{i-1}}{dx_i}\right)^T e_{i-1} - \left(\frac{de_i}{dx_i}\right)^T e_i$$

The first term is instantiated via forward recognition effects and the second term via lateral connections. These lateral connections embody the prior at each level. Connections are *reciprocal*.

Top-down synapses

For the top-down connections we have



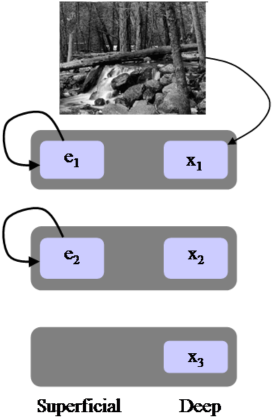
$$\tau_w \dot{w}_i = \frac{dL}{dw_i} = -\left(\frac{de_i}{dw_i}\right)^T e_i$$

This reduces to Hebbian learning for linear models

$\tau_w \dot{w}_i = (1 + \lambda_i) e_i x_{i+1}$

Recurrent synapses on error units

For the variance components



$$\tau_\lambda \dot{\lambda}_i = \frac{dL}{d\lambda_i} = -< \left(\frac{de_i}{d\lambda_i}\right)^T e_i > - \frac{1}{1+\lambda_i}$$

The self-connections whiten the errors

$\tau_\lambda \dot{\lambda}_i = (1+\lambda_i)^{-1}(e_i e_i^T - 1)$

References

K. Friston (2003) Learning and inference in the brain. Neural Networks 16, 1325-1352.

M. Mesulam (1998) From sensation to cognition. Brain (121), 1013-1052.

D. Mumford (1992) On the computational architecture of the neocortex II The role of cortico-cortical loops. Biological Cybernetics 66, 241-251.

R. Rao and D. Ballard (1999) Nature Neuroscience 2, 79-87.

G. Shepherd (2004). The Synaptic Organisation of the Brain. Oxford.