Hierarchical Sparse Coding

Ian Post

December 15, 2006

1 Introduction

A number of researchers have theorized that the brain may be employing some form of hierarchical model of features in visual processing. Nodes at the bottom of the hierarchy would represent local, spacially-oriented, specific features, while levels further up the hierarchy would detect increasingly complex, spatially-diffuse, and invariant features, with nodes in the uppermost layers corresponding to invariant representations of objects and concepts. For example, Mumford and Lee have outlined such a system employing hierarchical Bayesian inference to combine sensory input at the lowest levels with feedback from priors higher up [7].

Models have been developed based on the idea of sparse coding that seem to mimic many of the observed features of area V1 in the visual cortex—the lowest layer of the hierarchy. Specifically, we assume that natural images can be represented as a sparse linear combination of over-complete basis functions. Using unsupervised learning techniques and optimizing for sparseness, Olshausen and Field succeeded in generating such a set of bases that resemble the localized, oriented lines detected by simple cells in V1 [8]. Bell and Sejnowski used independent component analysis (ICA) and the infomax principle—maximizing the information preserved by the decomposition—to produce bases with similar characteristics [1].

These models are good as far as they go, but they cannot be readily extended to generate higher layers. In particular, we have assumed that the data is a linear combination of independent components, which limits the complexity of the structure that can be captured. Simply generating a new sparse code for the output of the first layer yields no new information.

2 Topographic ICA

Several related algorithms have been developed that attempt to extend Bell and Sejnowski's ICA to capture additional, non–linear structure by relaxing the independence assumption. I have primarily experimented with topographic ICA, a model proposed by Hyvärinen et al [5]. The basic idea is to group ICA bases into neighborhoods, such that components within a given neighborhood tend to be simultaneously active.

Let $\mathbf{x} = \mathbf{A}\mathbf{s}$, as in the usual ICA model, where $\mathbf{x}$ is the observed data and $\mathbf{s}$ the hidden, mixed sources. The s_i are optimized for independence, so little correlation exists between their actual values. However, we can capture the idea of simultaneous non-zero values through their energies s_i^2. Specifically, we want $cov(s_i^2.s_j^2) = E\{s_i^2 s_j^2\} - E\{s_i^2\}E\{s_j^2\} \neq 0$.

We slightly relax the independence assumption of ICA and assume that the correlated energies within a neighborhood are due to the influence of a further layer of independent latent variables $\mathbf{u}$

that determines the variances of $\mathbf{s}$. Hidden variables u_j are mixed with neighborhood weights h_{ij} and fed through a nonlinearity ϕ to determine the variance σ_i^2 of s_i:

$$\sigma_i = \phi(\sum_j h_{ij} u_j)$$

s_i is then generated as $s_i = \sigma_i z_i$, where the z_j are mutually independent variables with the same distribution as s_j with unit variance.

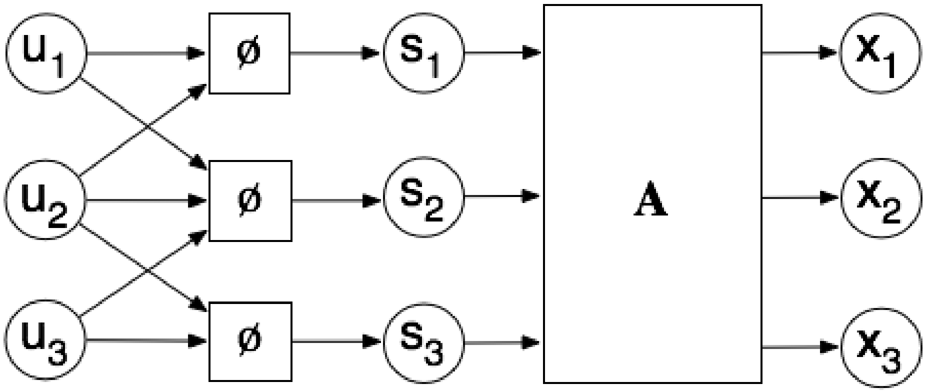

Figure 1: Topographic Independent Component Analysis. Independent sources $\mathbf{u}$ are mixed by neighborhood and run through a non-linearity ϕ which determines the variances of $\mathbf{s}$. The s_i are then mixed as in standard ICA.

The s_i are then conditionally independent given their variances, but are dependent due to relations between variances. Within neighborhoods components are uncorrelated

$$E\{s_i s_j\} = E\{z_i\}E\{z_j\}E\{\sigma_i\}E\{\sigma_j\} = 0$$

but will tend to have correlated energies since

$$E\{s_i^2 s_j^2\} - E\{s_i^2\}E\{s_j^2\} = E\{z_i^2\}E\{z_j^2\}[E\{\sigma_i^2 \sigma_j^2\} - E\{\sigma_i^2\}E\{\sigma_j^2\}]$$

The covariance of σ_i and σ_j is $\sum_k h_{ik} h_{jk} var u_k$ which is positive if s_i and s_j are in the same neighborhood. Constraining ϕ to be monotonic, then ϕ^2 is too, so applying ϕ^2 the covariance is still positive, so $cov(\sigma_i^2, \sigma_j^2)$ is positive, which implies the above equation [5].

Derivation of the learning rule is complex but is worked out in [5]. Learning is done using a gradient. For simplicity in the calculations ϕ is chosen to be $\phi(x) = x^{-\frac{1}{2}}$.

3 TICA Results

I experimented extensively with the neighborhood function and to a lesser extent with the non-linearity. As noted in [5], with sufficiently large, overlapping neighborhoods and enough bases,

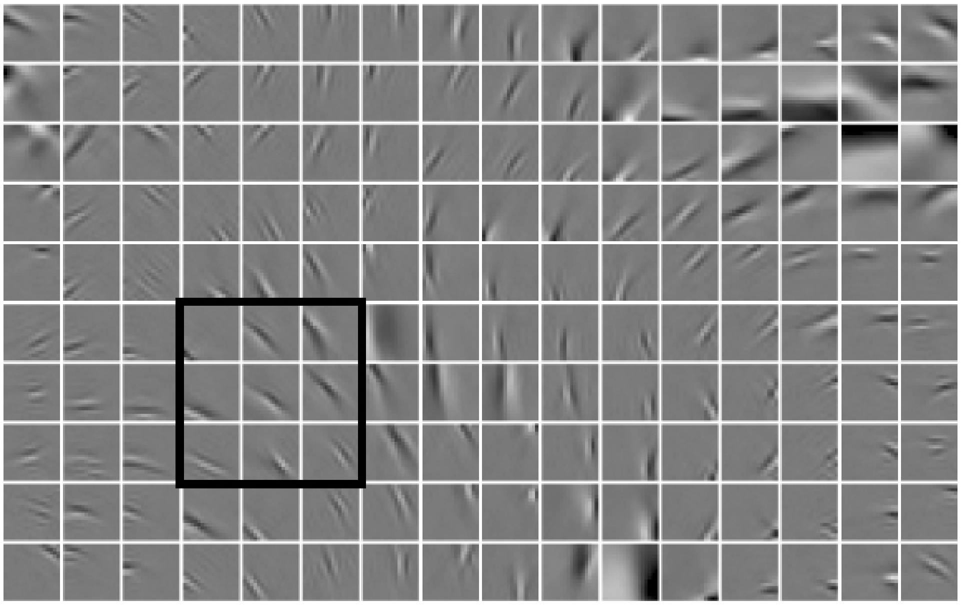

Figure 2: Topographic ICA with a 3 by 3 neighborhood. Neighborhoods resemble complex cells in that they exhibit phase invariance as well as limited in rotation and translation invariance.

components become group by orientation, and neighborhoods resemble the receptive fields of V1 complex cells. The exhibit phase invariance as well as limited translation and rotation invariance.

Large neighborhoods and more bases result in better defined neighborhoods, whereas smaller numbers of bases or tiny neighborhoods, such as the linear topography below, result in irregularities, fault lines of abrupt changes in basis orientation, etc. This appears to be due to limitations in the model, namely forcing components to fit the predefined topography or using too few bases to cover the image space, rather than any real correlations: such disjunctions become increasingly rare with more components and larger neighborhoods. Nor do they appear if the neighborhoods are disjoint, a special case of TICA known as independent subspace analysis [4].

More exotic neighborhoods including those with mixtures of positive and negative weights do not appear to add anything. They may even cause the bases to decay into gibberish with no discernible features (strange topography generally violates the assumptions of the algorithm). Adjustments in the nonlinearity also do not qualitatively change the results, although this function is highly constrained in its form due to technical issues in the learning rule. In general, the model is difficult to modify or generalize due to difficulties with intractable terms in learning.

I had hoped that creative fiddling and adjustments might allow learning of more interesting features, e.g. corners, but I think that would require an entirely new model, at least for the natural images I trained on. It is possible that artificial images with more distinct edges and corners would discover such structure. I experimented briefly with a further generalization of TICA proposed by Karklin and Lewicki [6] allowing much more general mixtures of the second layer of latent variables, but as hinted at in their paper, convergence can be tricky, and using their own code from Karklin's website, I was unable to reliably train the model. Even in the best of cases, results are extremely difficult to visualize.

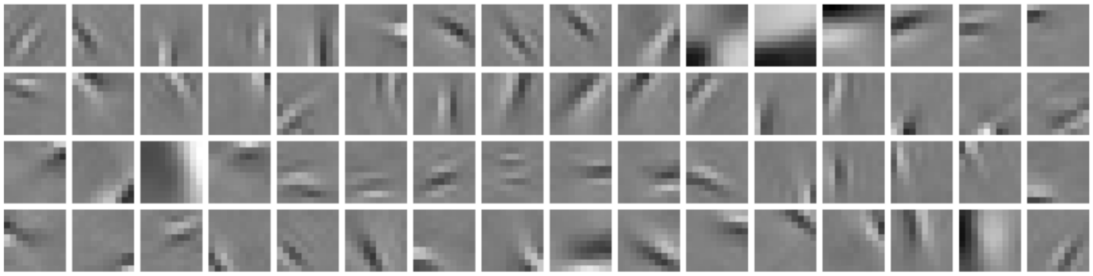

Figure 3: Topographic ICA with a linear 1 by 3 neighborhood, wrapping around to the next row. Some structure is discernible, but there is not enough overlap between neighborhoods to really bring it into relief.

4 Markov Random Fields and Future Work

One issue with TICA is that both the topography and weights must be specified *a priori*, forcing the user to guess what might be interesting and then forcing the data to conform to that model. This recently led me to begin experimenting with Markov random field models in the hope of learning optimal neighborhoods and weights. Learning is slow, but Hinton et al's contrastive divergence algorithm makes models of this scale and type practical [2, 3, 9]. [2] briefly describes an experiment that appears to produce results similar to TICA.

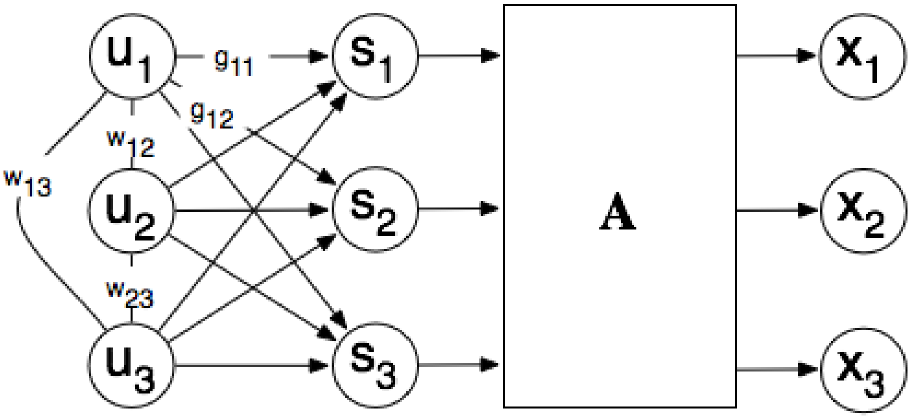

Figure 4: Markov Random Field model.

MRF models also offer huge advantages in extendibility over hierarchical ICA models. Whereas the second layer in TICA was laboriously constructed and is difficult to generalize, MRFs can, at least conceptually if not practically, be extended to arbitrary depths with only slight modifications. Given that oriented edges have emerged as bases with a number of different objective functions—

sparseness, mutual information, predictability, etc.—it is likely that similar components can be learned with an appropriate MRF model, yielding a unified framework for the entire hierarchy. As of now, I have completed only highly simplified proof of concept: treating ICA output as probabilities of binary MRF variables, I can reproduce some of the correlations revealed by TICA, although with far less sparsity in the connections. However, this model appears to offer much more promise.

References

[1] Anthony J. Bell and Terrence J. Sejnowski. The 'independent components' of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, 1997.

[2] G. Hinton, S. Osindero, and K. Bao. Learning causally linked markov random fields, 2005.

[3] G. E. Hinton, S. Osindero, and Y. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, pages 1527 – 1554, 2006.

[4] Aapo Hyvarinen and Patrik Hoyer. Emergence of Phase- and Shift-Invariant Features by Decomposition of Natural Images into Independent Feature Subspaces. *Neural Computation*, 12(7):1705–1720, 2000.

[5] Aapo Hyvarinen, Patrik O. Hoyer, and Mika Inki. Topographic Independent Component Analysis. *Neural Computation*, 13(7):1527–1558, 2001.

[6] Yan Karklin and Michael S. Lewicki. A Hierarchical Bayesian Model for Learning Nonlinear Statistical Regularities in Nonstationary Natural Signals. *Neural Computation*, 17(2):397–423, 2005.

[7] T. Lee and D. Mumford. Hierarchical bayesian inference in the visual cortex. *Journal of the Optical Society of America A*, 20:1434–1448, 2002.

[8] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[9] M. Welling and G. Hinton. A new learning algorithm for mean field boltzmann machines, 2002.