ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 301-308

Unsupervised Models for Processing Visual Data

Darryl Charles and Colin Fyfe,

Applied Computational Intelligence Research Unit, The University of Paisley, Scotland.
Email: darryl.charles,colin.fyfe@paisley.ac.uk.

Abstract. We discuss three aspects of modelling information extraction from visual data. Firstly, we discuss pre-processing issues in the context of stability and biological plausibility. Secondly, we discuss the problem of extraction of depth information from stereo data. Finally, we discuss the extraction of (almost) independent features from a data set. We use these three aspects of processing visual data to illustrate some of the successes and issues involved in using unsupervised learning with artificial neural networks on such data sets.

1. Introduction

Probably the most studied sensory system is the visual system and so this system has provided researchers with fruitful inspiration on developing artificial neural networks. Often such networks attempt to also mimic the properties of the system which provided the inspiration. Sometimes the properties of the system precede the artificial neural network i.e. the network is developed in order to solve a particular purpose (perhaps to solve a problem which nature has already solved) and only later is biological relevance considered.

This paper will give a brief overview of the type of work which has been carried out with artificial neural networks in the context of vision processing. However, there has been such a volume of work carried out in this area that we will, of necessity, be very selective; a more comprehensive treatment is given in [1] in which even although a complete section of 30 papers is devoted to vision, the topic crops up in many other sections such as those dealing with biological networks or learning. That this is not a one-way street is shown in the book Vision Science [10] in which an appendix (widely referenced within the book) is devoted to connectionist modeling.

The attractive (to connectionist modelers) perspective that modern vision science offers is that vision is a computational process - one which takes place within living beings but also one which can be mimic-ed by video cameras and computers. This means that the process of vision is amenable to theoretical analysis and based on this analysis to empirical modelling. [10] gives a very

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 301-308

full account of this perspective from psychological, physical, biological and even historical viewpoints.

In this paper, we will introduce briefly a few of the topics which have exercised those interested in applying artificial neural networks to visual data. We will concentrate on unsupervised learning because of our belief that self-organisation has already been shown to be feasible in vivo and that connectionist modeling must aspire to similar levels of expertise. Perhaps, given how easy animals find the extraction of information from visual data, we might wonder why we are finding it so difficult to emulate animal expertise. At the core of this difficulty is the fact that this process is an inverse problem: we are attempting to construct a model of the world from perceived sensory images. However it is not possible simply to construct an inverse mapping to do so, since the process is essentially under-determined: for example, the image appearing on the retina may be caused by a line x cm long y cm from the retina or a line $2x$ cm long and $2y$ cm from the retina. Without other information (which may be historical, rather than information being captured currently) we have no way of determining the true situation.

In section 2, we consider pre-processing issues. Then, in subsequent sections, we consider textures, stereo visions and extraction of visual factors before completing with a short section mentioning a few applications.

2. Sampling and Pre-processing

When training an artificial neural network, we require a good data set. If wishing to investigate models of visual information processing, we have a potentially infinite data set to select from. Typically the data is first quantised, often 1 byte per pixel for greyscale processing meaning that each pixel takes values from 0 to 255. If we have colour processing, we often use 1 byte for each of three colours per pixel.

If our task is to identify features from the images (perhaps face identification) we must find and use the relevant parts of the image. If on the other hand, we are interested in early processing of visual data, we may possibly only sample from the images; for still images, we often use square patches taken randomly from the image. For video sequences, we will use cubes composed of one randomly chosen square patch taken forward in time.

However this range of values can lead to instability in the network or extremely long training times. Field [4]has suggested logarithmically preprocessed images. This can be justified for two reasons. Firstly, human physiology appears to be more linear in the logarithm of contrast as opposed to simple contrast. Secondly, log preprocessing may help alleviate problems of differing illumination. By taking logarithms, local ratios in image intensity are transformed into local differences of image intensity. Ratios of intensity should be more robust to changes in illumination than are absolute differences.

The use of square or rectangular sections of an image may in itself bias the results - edge effects may come into play. Therefore we may preprocess

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 301-308

images by passing them through radially decaying windows such as Gaussian windowing[5]. Recently based on the analytic work which has been carried out on kernels in a supervised setting [15], the use of particular pre-processing based on kernels has been investigated [12, 6]. This body of work provides a sound analytically derived foundation for nonlinear pre-processing, and has already been applied to image processing [3, 9]. Perhaps the most exciting aspect of this area is the possibility that we can design kernels to match particular invariances which we may require [13].

Finally we may consider transforming our data to the frequency domain. Apart from the fourier transform itself, the most popular method in this area involves wavelets[1] The ability of wavelets to be used in multi-resolution analysis [14] makes them the tool of choice for many involved in image restoration, denoising, compression etc.

3. Stereoscopic Vision

Becker and Hinton suggest that that they wish to constrain the learning problem by restricting the features of interest to those which are liable to be useful for later perceptual processing. In a general non-specific environment, there are regularities ("coherence") in that any part of the environment is very likely to be predictable from other close parts of the environment e.g. any object has a finite compact surface area and so there exists a set of points all physically close to one another which share visually similar features. Similarly there exists temporal coherence in our environment. Also there is a coherence across sensory modalities - we generally see and smell and feel an orange at a single instant in time. This suggests that we should use coherence to extract information from the input data; one objective that might be appropriate for a network would be the extraction of redundancy (which gives rise to coherence) in raw sensory data since we do not, for example, have to use sight, smell and touch of an orange in order to identify the orange.

We could perform error descent on the squared error of the difference between the outputs but one difficulty with this is that the network could simply learn to output a constant value at both neurons $\mathbf{x_1}$ and $\mathbf{x_2}$. So we need to force the neurons to extract as much information as possible but still ensure that they are agreeing. This suggests that the optimisation criterion should be to maximise the mutual information between the two neurons

$$I_{a,b} = H(a) + H(b) - H(a,b) \quad (1)$$
$$= H(a) - H(a|b) \quad (2)$$

Written this way we can see that by maximising the mutual information between the neurons, we are maximising the entropy (the expected information output) of each neuron while minimising the conditional entropy (the uncertainty left about each neuron's output) given the other's value. So we wish

[1] With this term, we include ridgelets, curvelets and any other -lets.

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 301-308

each neuron to be as informative as possible while also telling us as little as possible about the other neuron's outputs.

In a movie, the temporal sequence of images is correlated and additional information can be extracted by looking for temporal as well as spatial structure. For example, a movie of a rigidly moving object contains highly redundant information because the image of the object will appear in slightly different spatial locations on successive frames of the movie. Foldiak [?] showed how this translation invariance can be captured in simple feedforward network that used Hebbian synapses and an output layer of units with a short-term memory of previous inputs. The network was trained with moving lines and the response properties of neurons in the network were similar to those found in the visual cortex. This principle was generalized by Stone, who applied it to learning stereo disparity from dynamic stereograms.

We should also mention that [7, 8, 9] have developed both neural and kernel methods for this problem and have related their methods to Canonical Correlation Analysis. Canonical Correlation Analysis is a statistical technique used when we have two data sets which we believe have some underlying correlation. Consider two sets of input data; $\mathbf{x}_1$ and $\mathbf{x}_2$. Then in classical CCA, we attempt to find the linear combination of the variables which give us maximum correlation between the combinations. Let

$$\begin{aligned} \mathbf{y}_1 &= \mathbf{w}_1\mathbf{x}_1 = \sum_j w_{1j}x_{1j} \\ \mathbf{y}_2 &= \mathbf{w}_2\mathbf{x}_2 = \sum_j w_{2j}x_{2j} \end{aligned}$$

where we have used $\mathbf{x}_{ij}$ as the j^{th} element of $\mathbf{x}_1$. Then we wish to find those values of $\mathbf{w}_1$ and $\mathbf{w}_2$ which maximise the correlation between $\mathbf{y}_1$ and $\mathbf{y}_2$. A recent development [?] has shown that there are a family of networks which, by solving the generalised eigenproblem, can find stereo disparity: it may be shown [11] that an alternative method of finding the canonical correlation directions is to solve the generalised eigenvalue problem

$$\begin{bmatrix} 0 & \Sigma_{12} \\ \Sigma_{21} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{bmatrix} = \rho \begin{bmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{bmatrix} \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{bmatrix} \tag{3}$$

where ρ is the correlation coefficient. Taking $\mathbf{w} = [\mathbf{w}_1^T\mathbf{w}_2^T]^T$, we find the canonical correlation directions $\mathbf{w}_1$ and $\mathbf{w}_2$ using

$$\begin{aligned} \frac{d\mathbf{w}_1}{dt} &= \Sigma_{12}\mathbf{w}_2 - f(\mathbf{w}_1)\Sigma_{11}\mathbf{w}_1 \\ \frac{d\mathbf{w}_2}{dt} &= \Sigma_{21}\mathbf{w}_1 - f(\mathbf{w}_2)\Sigma_{22}\mathbf{w}_2 \end{aligned}$$

where the function $f()$ must satisfy some rather simple constraints. Using the facts that $\Sigma_{ij} = E(\mathbf{x}_i\mathbf{x}_j^T), i, j = 1, 2$, and that $y_1 = \mathbf{w}_1.\mathbf{x}_1$, we may propose the

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 301-308

instantaneous rules

$$\begin{aligned} \Delta \mathbf{w}_1 &= \mathbf{x}_1 y_2 - f(\mathbf{w}_1)\mathbf{x}_1 y_1 \\ \Delta \mathbf{w}_2 &= \mathbf{x}_2 y_1 - f(\mathbf{w}_2)\mathbf{x}_2 y_2 \end{aligned}$$

This method is closer to that of [2] which also appears in this volume.

4. Extraction of Visual Factors

To provide a reference point for our work, the experimental procedure and data used is similar to that outlined in one of the key papers in this area [Olshausen et al, 1996]. In these experiments (and ours), the data comprises sample patches from ten pre-processed images of natural scenes. The pre-processing method pre-whitens the images to level out the power spectrum, as low frequency aspects in images tend to dominate high frequency ones which tends to have much less power. Note, as the first experiment illustrates, that this particular method does not eliminate correlations between sample data pixels but evens out the variances per pixel of the data set - similar methods are often used to pre-process data before applying the statistical method Factor Analysis.

Figure 1 illustrates the considerable difference between PCA and PFA when applied to natural image patches. The PCA network only identifies global spatial frequency information, whereas the PFA implementation forms filters that are very localised. Note that these may both be implemented in very similar linear artificial neural architectures - the essential difference is that the PFA network has a half-wave rectification function on each of the output neurons. The rectification function may be thought of as being used, to some degree, for the identification of sparse structure in the data. The filters formed by the PFA network resemble the centre-surround Ganglion cells of the mammalian visual system. Now if we replace the rectification function in the PFA network by a much more 'sparseness inducing' function, such as the soft threshold function (described above) then the filters are much more oriented in nature (Figure 2). The filters formed by this implementation of the network now resemble the structure of simple cells in the V1 area of the visual cortex.

5. Conclusion

This short paper can only scratch the surface of some of the work which is being done in the area of artificial neural networks and visual data. We have only discussed two interesting features of the field and have not begun to discuss texture, colour, motion etc. Nevertheless, we consider that these two aspects are an important part of a major and developing field and one in which we believe that there is every chance of major success in the next few years. We look forward to these years with greedy anticipation.

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 301-308

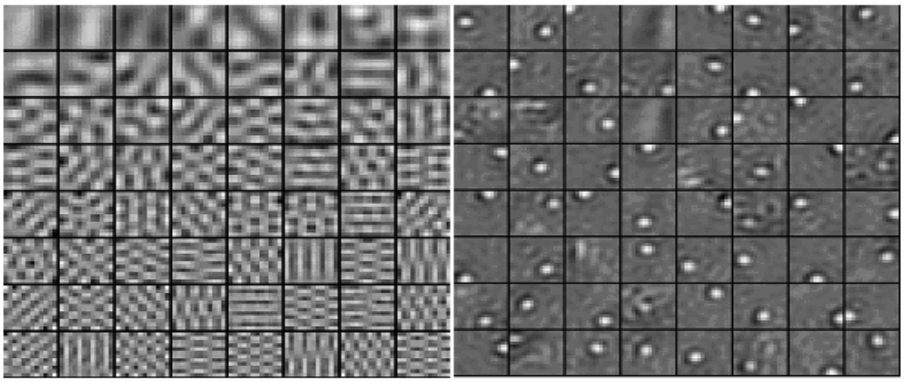

Figure 1: The results shown on the left image are of an unsupervised neural network used to perform PCA on the sample patches (12x12 pixels) from 10 natural images. The results shown on the right image are of an unsupervised neural network used to perform FA on the sample patches (16x16 pixels) from 10 natural images. The PCA method forms filters that resemble a Fourier basis which respond to that data in a global manner, whereas, the FA network forms filters that are much more local in nature - centre surround in nature.

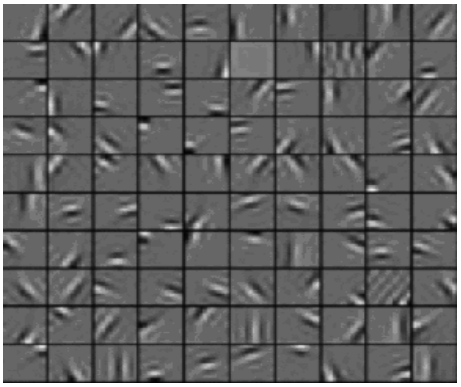

Figure 2: The results above illustrate the effect of making the function in our Factor Analysis network more 'sparseness inducing'. More complex filters are formed, resembling cell structure in the V1 region of the visual cortex. Data sample patches are 12x12 pixels in size.

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 301-308

References

[1] M. A. Arbib, editor. *The Handbook of Brain Theory and Neural Networks.* MIT Press, 1995.

[2] M. Borga and H. Knutsson. Canonical correlation analysis in early vision processing. In *Nineth European Symposium on Artificial Neural Networks, ESANN2001*, 2001.

[3] T. Evgeniou, M. Pontil, C. Papageorgiou, and T. Poggio. Image representations for object detection using kernel classifiers. In *ACCV*, 2000.

[4] David J. Field. What is the goal of sensory coding. *Neural Computation*, 6:559–601, 1994.

[5] C. Fyfe and R. Baddeley. Finding compact and sparse distributed representations of visual images. *Network: Computation in Neural Systems*, 6(3):1–12, 1995.

[6] C. Fyfe, D. MacDonald, P.L. Lai, R. Rosipal, and D Charles. *Recent Advances in Radial Basis Networks*, chapter Unsupervised Learning using Radial Kernels. 2000.

[7] P. L. Lai and C. Fyfe. A neural network implementation of canonical correlation analysis. *Neural Networks*, 12(10):1391–1397, Dec. 1999.

[8] P. L. Lai and C. Fyfe. A family of canonical correlation analysis networks. *Neural Processing Letters*, 2001. (Accepted for Publication).

[9] P.L. Lai and C. Fyfe. Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems*, 2001. (Accepted for publication).

[10] S. E. Palmer. *Vision Science, Photons to Phenomenology.* MIT Press, 1999.

[11] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis.* Springer, 1997.

[12] B. Scholkopf, A. Smola, and K.-R. Muller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

[13] A. J. Smola, S. Mika, B. Scholkopf, and R. C. Williamson. Regularized principal manifolds. *Machine Learning*, pages 1–28, 2000. (submitted).

[14] Murtagh F. Starck, J.-L. and A. Bijaoui. *Image Processing and Data Analysis, The Multiscale Approach.* Cambridge, 1998.

[15] V Vapnik. *The nature of statistical learning theory.* Springer Verlag, New York, 1995.