Behaviourally meaningful representations from normalisation and context-guided denoising

Harri Valpola
Artificial Intelligence Laboratory
University of Zurich, Switzerland
valpola@ailab.ch

Abstract

Many existing independent component analysis algorithms include a preprocessing stage where the inputs are sphered. This amounts to normalising the data such that all correlations between the variables are removed. In this work, I show that sphering allows very weak contextual modulation to steer the development of meaningful features. Context-biased competition has been proposed as a model of covert attention and I propose that sphering-like normalisation also allows weaker top-down bias to guide attention.

1. Introduction

One of the longstanding questions in machine learning is how to extract from sensory inputs meaningful features which are most useful for motor control, predicting future rewards, etc. I propose a learning scheme and a model architecture which combine ideas from various machine learning techniques and models of neural information processing. The learning scheme is close to unsupervised learning but it can use supervisory signals to steer learning to provide representations that are most useful for the system. The model is designed to be an integral part of an autonomous robot but the basic principle seems useful for various other tasks, too.

The learning scheme is based on denoising source separation (DSS) which is a recently introduced framework for constructing source separation algorithms around a denoising procedure (Särelä and Valpola, 2004). Depending on the type of denoising, the learning scheme can range from fully unsupervised to mostly supervised. Here I consider a hierarchical, nonlinear model where increasingly abstract features are extracted from bottom-up inputs under the guidance of lateral, top-down and temporally delayed context. The context is used to implement the denoising required in DSS.

Interestingly, the present model, which was originally designed for feature extraction, turned out to be similar in many respects to a model of visual attention (Duncan and Humphreys, 1989) where attention emerges from top-down bias and local competition in a hierarchical network. In this article, I elaborate on the links between feature extraction and emergent attention. I propose that they can be viewed as similar processes operating on two different timescales. One of the principal ingredients of DSS is normalisation which allows very weak contextual influence to guide feature extraction. I propose that similar normalisation is also useful for attention mechanisms.

The rest of this article is structured as follows. Feature extraction and attention models which are the background for this work are introduced in Sec. 2. The new model and connection to models of attention are proposed in Sec. 3 and experiments using the new model are presented in Sec. 4. Finally, Sec. 5 discusses how the goals and value system of the robot are able to guide feature extraction and attention and what are the implications of the new model.

2. Background

This section first briefly introduces the feature extraction algorithms on which the proposed approach is based. Finally, models where attention emerges from biased competition are discussed.

2.1 Denoising source separation

Recently, blind source separation techniques have received a lot of attention in the signal processing community (Hyvärinen et al., 2001) and related models have been shown to bare resemblance to cortical feature extraction (Olshausen and Field, 1996). DSS is a recently developed framework for constructing source separation algorithms around a denoising procedure (Särelä and Valpola, 2004). The basic ingredients for a DSS-based algorithm are sphering, denoising and competitive, Hebbian-type learning.

Sphering refers to a normalisation scheme where signals are decorrelated and their variances are normalised. It is used by many blind source separation algorithms as preprocessing since it makes it possible to use simple correlation-based learning. With-

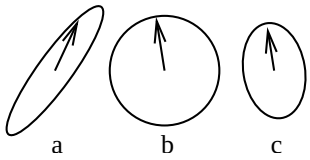

Figure 1: Original data set (a), after sphering (b) and after denoising (c). After these steps, the projection yielding the best signal-to-noise ratio, denoted by arrow, can be obtained by simple correlation-based learning.

out normalisation, learning would be biased towards the signals with the highest energy. These are often not the most useful signals. In DSS, sphering enables very weak and vague influence to guide signal extraction towards signals which are sought for. The processing stages of DSS are shown in Fig. 1.

2.2 Hierarchical nonlinear models

Slow feature analysis (SFA) is a technique for finding temporally-invariant features. With natural images as inputs, it has been shown to learn features baring similarity to those found at the early stages of visual system (Wiskott and Sejnowski, 2002). In SFA, the observations are first expanded nonlinearly and sphered. The expanded data is then high-pass filtered and projections minimising the variance are estimated. Due to the nonlinear expansion, it is possible to stack several layers of SFA on top of each others to extract higher-level, slowly-changing features, resulting in hierarchical SFA.

SFA is directly related to DSS. Instead of minimising the variance after high-pass filtering as in SFA, it is also possible to maximise the variance after low-pass filtering. SFA is thus equivalent to DSS with nonlinear data expansion and low-pass filtering as denoising. The structure of SFA is depicted in Fig. 2 from this viewpoint. SFA shares many features with earlier methods proposed for temporally-invariant feature extraction (Földiák, 1991, Kohonen et al., 1997, Parga and Rolls, 1998).

A common setup in large hierarchical models is that consequtive layers are sparsely connected. In neocognitron, for instance, the connections are topographic and localised (Fukushima, 1980). Each level compiles increasingly abstract features from local features at the previous level. At the highest levels, the features can be global and multi-modal.

2.3 Competition

Most feature extraction models which are learned in unsupervised manner have either an explicit or implicit competition mechanism. Explicit implementations include winner-take-all mechanisms and inhibitory lateral or top-down connections that decorrelate activations. Implicit mechanisms include orthogonalisation of bottom-up weights. The purpose of the competition mechanism, whether explicit or implicit, is to guarantee that a diverse set of features develops. Without competition, all features might develop to represent exactly the same thing.

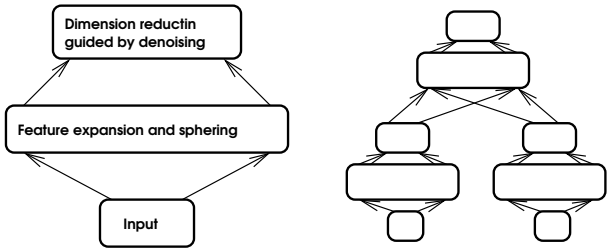

Figure 2: Left: One processing stage in SFA consists of feature expansion and sphering followed by denoising and dimension reduction. Denoising is implemented by low-pass filtering. Right: A two-level hierarchy of three SFA modules is shown on the right.

Local lateral inhibitory connections are often used in biologically motivated models. They need to be combined with localised bottom-up connections because otherwise the developed features would be redundant outside the range of the lateral inhibition.

2.4 Context-guided feature extraction

In hierarchical nonlinear models, such as a hierarchical SFA, each processing stage takes the features learned at the previous level and assembles new, more abstract features which are passed on to the next level. In such a system, feature extraction boils down to defining what is a useful abstraction. Raw input, such as camera images, often contains a vast amount of information and in order to be useful, the high-level representation needs to discard most of the information while retaining the meaningful parts.

Following several researchers, I argue that the mutual information between the features and context is an excellent criterion for developing useful features. One of the early developments in that direction was canonical correlation analysis (CCA), a statistical technique which is designed to find linearly correlated features from two data sets (Hotelling, 1936). Since then, the criterion has been extended from linear correlation to mutual information (Becker and Hinton, 1992) and from relation between two data sets to a more general concept of context. The context can include features at the higher levels (top-down context), features at distant neighbouring areas on the same level (lateral context) and features delayed in time (temporal context). For a recent review, see e.g. (Körding and König, 2001).

Figure 3 demonstrates how lateral and top-down

context can affect the interpretation of individual parts of images. In this case the context causes two identical low-level feature combinations to be interpreted differently while in some other cases the context can cause two different low-level feature combinations to be perceived the same. Context can thus aid at defining categories at perceptual timescale. When learning is activity-dependent, perceptual categorisation influences learning of categories. For example, word categories and other similar features have been shown to be learned based on similarities of the contexts in which words occur (Ritter and Kohonen, 1989, Honkela et al., 2003).

Figure 3: Depending on the context, it is possible to perceive either "RB" or "12 13" although those parts are identical in the two cases. This contextual influence is learned and the strength of the effect thus depends on such factors as the language and handwriting the observer is used to.

2.5 Emergent attention from biased competition

Contextual bias, predominantly top-down bias, combined with local lateral competition has been proposed as a model of covert attention in humans (Duncan and Humphreys, 1989). In simulations, such models have replicated many of the phenomena found in neurophysiological experiments (see, e.g., (Reynolds and Desimone, 1999, Reynolds et al., 1999, Deco and Schürmann, 2000, Deco and Rolls, 2004, Spratling and Johnson, 2004)). Attention can thus be seen as a dynamic process emerging from an interplay between long-range excitatory and local inhibitory connections. Different strengths of excitation and inhibition have been shown to give rise to several distinct functional regimes, covert attention being one of them (Szabo et al., 2004).

3. Proposed model

In this section, I introduce a feature extraction model which is based on the principles introduced in the previous section. I also suggest that context-guided feature extraction and emergent attention discussed in the previous section reflect the same principle but operate on different timescales.

3.1 Overall model structure

I propose that features can be extracted from sphered inputs via competitive learning which is biased by the context. Within DSS framework, context can be seen to provide information for denoising the activations. Sphering is important because it allows very weak contextual guidance to steer feature extraction.

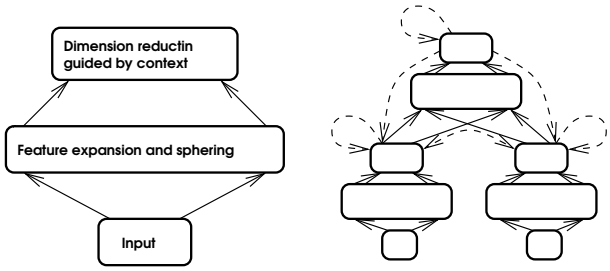

Figure 4: Left: One module consists of feature expansion and sphering followed by denoising and dimension reduction. Denoising is guided by context and is implemented by biased competition. Right: A two-level hierarchy of three modules is shown on the right. Dashed lines denote contextual inputs (top-down, lateral and delayed). Note the close resemblance with Fig. 2.

Slow feature analysis is closely related to the above scheme if the context of each feature is limited to the delayed value of the same feature. High mutual information is in practice usually achieved if the feature changes slowly. If the context includes more delayed values of the same feature or of other features, the feature no longer needs to change slowly. It is enough that the feature is predictable in the context. In speech, for instance, the target features could be phonemes. They can change faster than the input features, spectra, but phonemes are highly predictable given the context. Figure 4 highlights the similarity of SFA and the proposed scheme. The main difference is the criterion for denoising.

The basic ingredients of the model are the following:

- hierarchical architecture with a distinction between bottom-up and contextual inputs
- contex consisting of top-down, lateral and delayed inputs
- initial processing stage of feature expansion and sphering of bottom-up inputs
- subsequent processing stage integrating the sphered features with the context
- localised bottom-up connections and local competition
- neuron-like elements computing weighted sums of their inputs and activity-based Hebbian-like learning of the weights

There are several possible ways to implement the above model and it is unlikely that exact details are important.

The purpose of the initial processing stage is twofold. First, nonlinear feature expansion enables the model to implement a nonlinear mapping. Second, sphering renders the variance of all projections of the expanded feature space equal, allowing minimal contextual influence to guide feature extraction.

The second processing stage applies dimension reduction to the expanded feature space guided by the context. This is implemented by abstracted neurons that compute weighted sums of their inputs. The crucial point is that bottom-up inputs and context have separate sets of weights that are normalised independently and whose relative contribution is fixed. If the second processing stage is linear, essentially implementing principal component analysis (PCA), an infinitesimal contribution of the context is in principle able to bias dimension reduction.

An example of this is provided by CCA implemented within DSS framework. CCA can be be implemented by sphering the two given data sets and applying PCA for each data set such that the each principal component uses as its context the corresponding principal component from the other data set. CCA is the limiting case where the contribution of the context goes to zero. Figure 5 illustrates nonlinear CCA, which includes nonlinear feature expansion (Lai and Fyfe, 2000). In the proposed model, the contribution of the context is nonzero but small and the context is richer than in CCA.

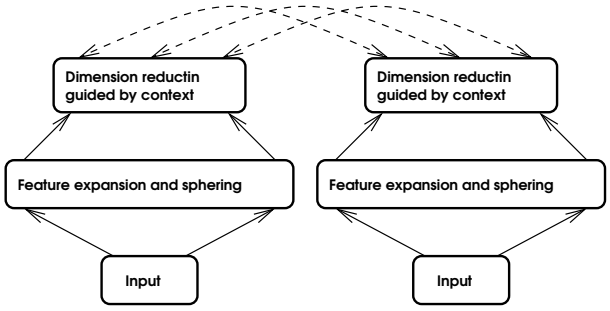

Figure 5: Nonlinear CCA can be implemented using the proposed model if the context of each PCA component is limited to the corresponding PCA component of the other data set and the contribution of the context is infinitesimal. If a delayed version of the data is used as the other data set, SFA follows (Friman et al., 2002).

Finally, note that while the bottom-up inputs need to be localised and match in extent the radius of local competition in the proposed model, there is no such restriction for contextual inputs. As long as the context only has a modulatory role and cannot activate the neurons without the bottom-up inputs, there is no risk of developing redundant features because outside the radius of the competition, the driving bottom-up inputs are different.

3.2 Detailed implementation

In order to keep things as simple as possible, batch learning is used, i.e. the parameters of the model are updated only after going through all the data. The advantage of batch learning is that parameters related to learning rates are not needed. In a real autonomous robot the model will obviously be implemented using online learning.

Instead of having a continuous field of neurons with localised inhibitory connections, discrete modules are used. Any bottom-up input is connected to all neurons in one module. In an on-line version, lateral inhibitory connections would be used, but now neurons are discouraged from developing redundant features by explicit orthogonalisation-like procedure. Orthogonalisation is not exact for feature expansion at the first stage because there may be more features than inputs and hence the basis cannot be orthogonal.

The feature expansion is implemented by a simple sparse coding algorithm. In matrix notation, the computation of activations $\mathbf{S}$ is as follows:

$$\mathbf{S} = \mathrm{orth}[\mathbf{f}(\mathbf{W}^T\mathbf{X})] \quad (1)$$

$$f(x) = [x - \tanh x]_+ \quad (2)$$

Here $\mathbf{X}$ denotes the inputs and each column vector is the vector of inputs, $\mathbf{W}$ is the weight matrix with the weights of each neuron as column vectors and $\mathbf{S}$ contains the activations resulting from different input vectors as column vectors. Approximate orthonormalisation is denoted by orth. The activation function $\mathbf{f}(\cdot)$ operates element-wise on the matrix and the function $[\cdot]_+$ denotes rectification, i.e. $[x]_+ = x$ if $x > 0$, otherwise $[x]_+ = 0$.

The activation function has a dual purpose. First, it implements a nonlinearity which makes it reasonable to build a hierarchy. Second, it plays a role in promoting the development of meaningful features. The adaptation algorithm can, in fact, be interpreted as DSS because the activation function $f(\cdot)$ takes small values towards zero (the gain is zero if the input is zero) but passes through greater values (gain approaches one as the input grows).

As usual in DSS, the weight matrix is updated simply as

$$\mathbf{W}_{\mathrm{tmp}} = \mathbf{X}\mathbf{S}^T \quad (3)$$

$$\mathbf{W}_{\mathrm{new}} = \mathrm{norm}(\mathbf{W}_{\mathrm{tmp}})\,, \quad (4)$$

where norm denotes normalisation. Since the data is sphered and $\mathbf{S}$ are approximately orthogonal, $\mathbf{W}$ will also be approximately orthogonal.

The first and second stage differ in their connectivity and dimension of representation. The feature expansion stage receives bottom-up inputs only, while the context-guided dimension reduction stage

receives the contextual inputs in addition to the expanded, sphered features. At this latter stage, I arbitrarily fixed the ratio of the sum of the weights from the context and from the sphered bottom-up features to be 1:9, i.e. 10 % of the input came from the context. The results are insensitive to this choice as long as the context only modulates but does not drive the activations.

3.3 Unified view

Acoustic source separation is often exemplified by the so-called cocktail-party problem. The task is to concentrate on one speaker in a room full of conversing people. The ability to select one target among many interfering ones is closely related to attention. Source separation is thus closely related to attention. As we have seen, however, source separation can also be used for feature extraction.

Competition and contextual bias can give rise to both emergent attention and development of features. There are obvious similarities and differences but I argue that the similarities go deeper than the surface and the differences can be traced back to difference in timescale. The similarites of the two phenomena not only have theoretical interest but can lead to genuine transfer of ideas. In particular, I propose that since normalisation, or sphering, has turned out to be very useful in feature extraction, it should be useful for attentional mechanisms as well.

The input for both feature extraction and attention is a set of activated bottom-up features. Context-guided denoising can then compile a new, more abstract representation, and competition inforces diversity guaranteeing that the representation remains rich. The main difference between feature extraction and attention is timescale. The development of new features is based on long-term statistics of the input features, i.e. there is integration over time, while attention can rapidly shift from one object to another. The input for attention is the active population of features on a short timescale and the end result is an object, or rather an event, defined by a new population of active features.

In feature extraction, the long-term statistics of the inputs guide slow development of synaptic weights. In attention, the active object is defined by fast-changing activations. It is also possible that the synapses have a fast-changing portion which serves binding and short-term memory (Triesch and von der Malsburg, 1996). Since the result of learning in feature extraction is stored in the synaptic weights, competition can be implemented by an orthogonalisation procedure operating on the weights. For attention this is not an option because the outcome is defined in terms of activations. A competition mechanism decorrelating the activations can serve both feature extraction and attention.

The main thesis of this paper is that sphering is useful for finding representations because it gives a pivotal role to contextual effects. Very weak contextual bias can thus steer the development of features and the same should hold for attention: normalisation should allow a much weaker bias to give rise to attention. In other words, the parameter regime corresponding to covert attention reported by (Szabo et al., 2004) would be expanded because much weaker top-down excitation would suffice.

Sphering normalises the statistics of of the inputs but as feature extraction and attention have different timescales, the timescales of normalisation should differ, too. The long-term covariance of the input features is sphered for feature extraction but the analogue of this for attention, covariance of input objects, is more elusive. The reason is that the set of features that constitutes an object can change in time. Normalisation should thus change in time and operate over the instantaneous populations of active features corresponding to different objects. Because the definition of an object is elusive, exact normalisation appears infeasible but this should not be a major problem. First, normalisation needs not be perfect in order to be useful, and second, the representations of objects can be normalised gradually towards higher levels of hierarchy. Imperfect and local normalisation schemes would therefore suffice.

Interestingly, some of the inhibitory contextual effects found on cortex might reflect instantaneous normalisation of objects. It is known that at low levels of bottom-up activation, contextual influence is predominantly excitatory but at higher levels, the influence gradually becomes suppressive (Angelucci et al., 2002). This is exactly what one would expect if two contextual mechanisms would operate in parallel: additive excitatory biasing and divisive inhibitory normalisation (Schwartz and Simoncelli, 2001).

It has been proposed that the neural activation levels (on V1 more specifically) relate directly to the saliency of objects (Zhaoping, 2002) and no separate "saliency map" is required. It seems plausible that the representation at higher levels reflects the activations at the lower levels. The simple definition of saliency could thus be: something able to activate a representation at higher levels. Perpectual phenomena such as saliency of closure or pop-out of anomalous features could be explained by excitatory biasing and inhibitory normalisation, respectively.

I propose that inhibitory normalisation serves attention by giving the weak biasing top-down influence a pivotal role in determining which representation wins local competition. What constitutes an object is learned gradually during development and then used on perceptual timescale to form and normalise objects by excitation and inhibition, respec-

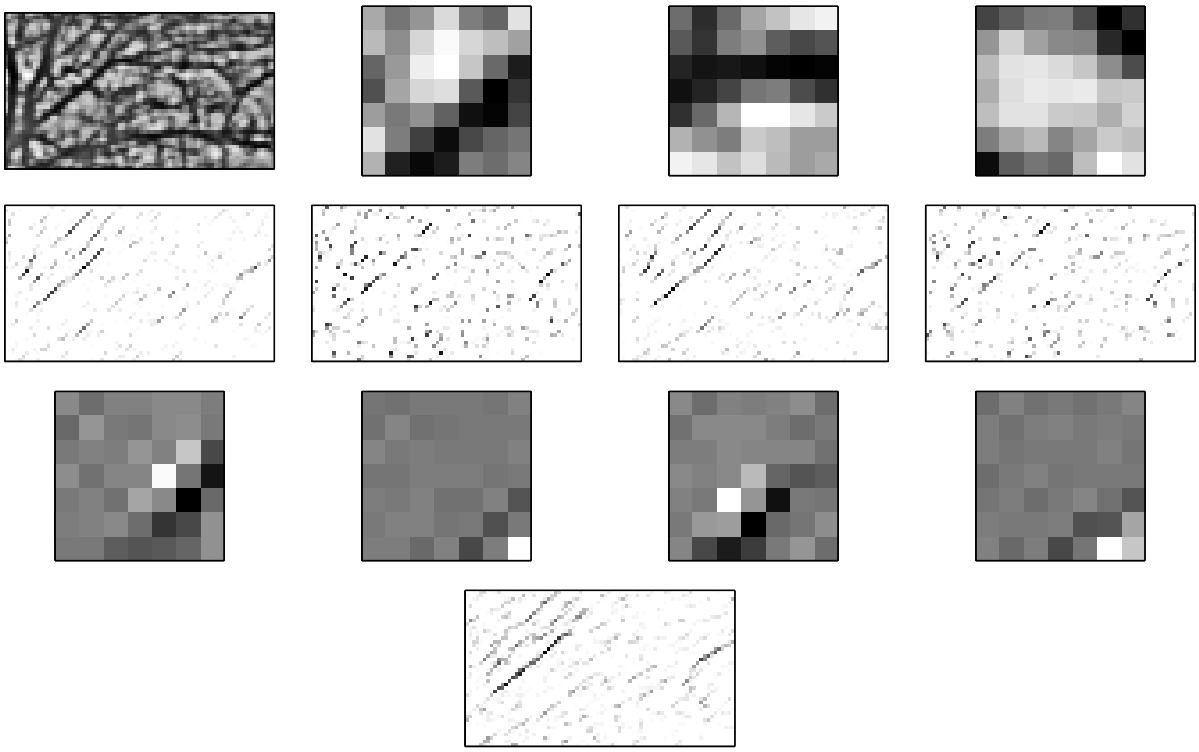

Figure 6: Top row: gray-scale image (left) and three 7 by 7 patches. Second and third row: activations at each image location and receptive fields of four neurons from the feature expansion layer. Bottom: output activations of a neuron pooling from the four features.

tively. The long-range connections are adapted based on correlations, i.e., by Hebbian-type learning rules. The long-range connections not only excite but also inhibit the targets via inhibitory interneurons. The net effect is additive excitatory biasing combined with divisive inhibitory normalisation.

4. Experiments

The purpose of these experiments is to demonstrate that sphering allows very weak contextual guidance to steer the development of meaningful features. I used sampled image patches as the data and will show that spatial context alone is able to give rise to features bearing similarity to translation invariant complex cells found on primary visual cortex. Complex cells have been shown to emerge from using temporal context (e.g. (Kohonen et al., 1997, Wiskott and Sejnowski, 2002)) but here I show that no temporal ordering of the data is required.

A small image of a tree was preprocessed by high-pass filtering and mild contrast normalisation. The data was separated onto on- and off-center channels. These steps resembles the early stages of visual processing on the retina and thalamus. The processed image is shown on the top left of Fig. 6.

First, 7 by 7 image patches were sampled (three out of 5,220 are shown in Fig. 6) and used for learning 100 features at the initial feature expansion stage. Many of the receptive fields that developed correspond to localised edge detectors. This is a typical result for sparse coding of images.

Second, 21 by 21 image patches divided into 9 subfields were sampled. They were used to train the second stage. On each subfield, dimension reduction from 100 to 20 features was guided by the context from the other subfields. The second stage leared to integrate together the responses of the edge detectors such that more invariant features developed. An example of such complex-cell-like feature which integrates four elementary edge detector features is shown in Fig. 6. The output feature is cleaner and more invariant than the constituent features at the feature expansion stage.

Note that most neurons at the feature expansion stage had even noisier outputs than the ones shown in the figure but they were not used as much to build the output features. The experiment thus shows that even with very limited data set (usually several natural images are used) and without using temporal context, it is possible to develop meaningful features by contextual guidance. Very weak contextual guidance is sufficient when the feature expansion stage provides sphered outputs.

5. Discussion

Using context to guide feature extraction and attention of an autonomous robot makes sense because context includes information about what the robot is doing, what its goals are, etc. The robot will thus develop such sensory features and guide attention in such a way that relevance to action and goals is maximised. For instance, if the robot practices grasping objects, its motor context (grasping) can guide the development of relevant visual features (graspable). Similarly, the robots attention would be guided to the act of grasping (motor attention) and the grasped object (visual attention). In this view, attention is the process by which a coherent sensorimotor percept and behaviour emerges. Cortical representation of action and reward has a decisive role in the development of sensory representations and categories. What one thinks and does determines what one perceives and learns.

It is well established that attention has a strong top-down component, i.e. attention is to a large extent active and task-driven. A likely source of this goal-oriented biasing is working memory located on prefrontal cortex. More specifically, basal ganglia have been proposed to gate thalamocortical loops which implement working memory (O'Reilly, 2003). The control of this active gating develops via reinforcement learning. The contents of working memory exerts top-down bias over the sensory representations and gives rise to movements, goal-oriented attention, etc., depending on the area of prefrontal cortex in question. Since the development of features is activity dependent, goal-oriented attention translated into behaviourally significant feature representations.

Note that this type of context-guided learning does not rule out a more unspecific role of value signals in learning. It has been proposed that values signals modulate the learning rate on cortical representations and thereby guide the development of feature representations (Sahani, 2004). However, modulation of learning rate only tells when to learn, not what to learn. The contents of working memory and context in general provide much more specific information about what exactly to learn. It is likely that the combination of information about what and when to learn is more useful than either one alone.

The experiments reported here concentrated on feature extraction and demonstrated only that sphering allows very weak contextual bias to steer the development of meaningful features. Using the same idea for models of context-guided attention seems to be a promising direction of future work. Normalisation of inputs, competition among local representation and biasing from context should be able to serve both feature extraction and attention and it would thus obviate the need for separate mechanisms for the two tasks (separate mechanisms were used by (Deco and Rolls, 2004)).

6. Conclusion

I combined two existing techniques, sphering and context-guided learning, and shown that sphering helps the development of meaningful features by allowing very weak contextual guidance to steer learning. The experiments only addressed feature extraction but I argued that the emergence of attention resembles feature extraction in many ways, the main difference being different timescale of operation. I therefore suggested that sphering-like normalisation should be useful for the models where attention emerges from local competition and top-down bias.

Acknowledgements

This work has been funded by European Commission project ADAPT (IST-2001-37173). I wish to thank Allard Kamphuisen for valuable discussions and for conducting experiments with an earlier version of the model, and Lars Schwabe and Martin Stetter useful discussions.

References

Angelucci, A., Levitt, J. B., Walton, E. J., Hupe, J. M., Bullier, J., and Lund, J. S. (2002). Circuits for local and global signal integration in primary visual cortex. *Journal of Neuroscience*, 22(19):8633–8646.

Becker, S. and Hinton, G. E. (1992). Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355:161–163.

Deco, G. and Rolls, E. T. (2004). A neurodynamical cortical model of visual attention and invariant object recognition. *Vision research*, 44:621–642.

Deco, G. and Schürmann, B. (2000). A hierarchical neural system with attentional top-down enhancement of the spatial resolution for object recognition. *Vision research*, 40:2845–2859.

Duncan, J. and Humphreys, G. (1989). Visual search and stimulus similarity. *Psychological Review*, 96:433–458.

Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, 3:194–200.

Friman, O., Borga, M., Lundberg, P., and Knutsson, H. (2002). Exploratory fMRI analysis by autocorrelation maximization. *NeuroImage*, 16(2):454–464.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202.

Honkela, T., Hyvärinen, A., and Väyrynen, J. (2003). Emergence of linguistic representations by independent component analysis. Technical Report A72, Lab. of Computer and Information Science, Helsinki University of Technology, Finland.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28:321–377.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent component analysis*. Wiley.

Kohonen, T., Kaski, S., and Lappalainen, H. (1997). Self-organized formation of various invariant-feature filters in the Adaptive-Subspace SOM. *Neural Computation*, 9(6):1321–1344.

Körding, K. P. and König, P. (2001). Neurons with two sites of synaptic integration learn invariant representations. *Neural Computation*, 13:2823–2849.

Lai, P. L. and Fyfe, C. (2000). Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems*, 10(5):365–377.

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

O'Reilly, R. C. (2003). Making working memory work: A computational model of learning in the prefrontal cortex and basal ganglia. Technical Report ICS Technical Report 03-03, University of Colorado Boulder.

Parga, N. and Rolls, E. T. (1998). Transform invariant recognition by association in a recurrent network. *Neural Computation*, 10(6):1507–1525.

Reynolds, J. H., Chelazzi, L., and Desimone, R. (1999). Competitive mechanisms subserve attention in macaque areas V2 and V4. *Journal of Neuroscience*, 19:1736–1753.

Reynolds, J. H. and Desimone, R. (1999). The role of neural mechanisms of attention in solving the binding proble. *Neuron*, 24:19–29.

Ritter, H. and Kohonen, T. (1989). Self-organizing semantic maps. *Biological Cybernetics*, 61:241–254.

Sahani, M. (2004). A biologically plausible algorithm for reinforcement-shaped representational learning. In Thrun, S., Saul, L., and Schölkopf, B., (Eds.), *Advances in Neural Information Processing 16 (Proc. NIPS'03)*. MIT Press. In press.

Schwartz, O. and Simoncelli, E. P. (2001). Natural signal statistics and sensory gain control. *Nature Neuroscience*, 4(8):819–825.

Spratling, M. W. and Johnson, M. H. (2004). A feedback model of visual attention. *Journal of Cognitive Neuroscience*, 16:219–237.

Szabo, M., Almeida, R., Deco, G., and Stetter, M. (2004). Cooperation and biased competition model can explain attentional filtering in the prefrontal cortex. Submitted.

Särelä, J. and Valpola, H. (2004). Denoising source separation. *Submitted to a journal*. Available at Cogprints `http://cogprints.ecs.soton.ac.uk/archive/00003493/`.

Triesch, J. and von der Malsburg, C. (1996). Binding — a proposed experiment and a model. In *Proc. Int. Conf. on Artificial Neural Networks (ICANN'96)*, pages 685–690, Bochum, Germany.

Wiskott, L. and Sejnowski, T. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural computation*, 14:715–770.

Zhaoping, L. (2002). A saliency map in primary visual cortex. *Trends in Cognitive Sciences*, 6(1):9–16.