A Hierarchical Generative Model of Recurrent Object-Based Attention in the Visual Cortex

David P. Reichert, Peggy Series, and Amos J. Storkey

School of Informatics, University of Edinburgh,
10 Crichton Street, Edinburgh, EH8 9AB, UK
{d.p.reichert@sms.,pseries@inf.,a.storkey@}ed.ac.uk

Abstract. In line with recent work exploring Deep Boltzmann Machines (DBMs) as models of cortical processing, we demonstrate the potential of DBMs as models of object-based attention, combining generative principles with attentional ones. We show: (1) How inference in DBMs can be related qualitatively to theories of attentional recurrent processing in the visual cortex; (2) that deepness and topographic receptive fields are important for realizing the attentional state; (3) how more explicit attentional suppressive mechanisms can be implemented, depending crucially on sparse representations being formed during learning.

1 Introduction

A Deep Boltzmann Machine (DBM) is a hierarchical, probabilistic, sampling based neural network that learns representations from which it generates or predicts the data it sees, utilizing recurrent processing. Though introduced in a machine learning context [1], these properties make the DBM an interesting model of processing in the cortex (cf. e.g. [2, 3]). In earlier work, we showed how the DBM can model homeostasis induced hallucinations [4]. Here, we demonstrate in a proof of concept how aspects of object-based attention can be modeled with a DBM as well – not in terms of saliency maps or eye movements, but in terms of what happens throughout the cortical hierarchy during the act of paying attention to an object in a visual scene. In that sense, this work can be understood as modeling in particular *covert* attention. It relates to approaches such as Selective Tuning [5] and others (e.g. [6, 7]), but is unique in capturing facets of attention in a framework implementing aforementioned general properties.

We qualitatively elucidate on the following aspects of theories of attentional processing in the cortex: First, the notion of a fast feed-forward (FF) sweep followed by subsequent recurrent processing, the latter being essential for perceiving objects when scenes are cluttered [8]; second, that, in directing attention to an individual object in a scene, an attractor state is assumed which binds together and emphasizes aspects of that object represented throughout the cortical hierarchy, suppressing representations of competing objects [9, 5]; third, the hypothesis that scene representations in the cortex are inherently such that higher stages represent primarily one object at a time, unlike lower stages such as V1 where the whole image is encoded in terms of low-level features [10].

Our main focus is the biological application, but on the technical side we show how deepness of the architecture and restricted receptive fields are important for realizing the attentional state, making the DBM robust against noise not seen in training. Finally, we explore additional suppressive attentional mechanisms to cope with problems beyond toy data, and argue that sparse representations could be critical to that end.

2 Setup

For brevity we only give a short overview of the model. See [1] on DBMs, and [4] on our specific setup, including additional neuroscientific motivation.

A DBM consists of several layers of stochastic neuronal units $\mathbf{x}$, usually with binary states, connected via symmetric weights $\mathbf{W}$, with no lateral connections within a layer to simplify computations. The lowest layer $\mathbf{x}^{(0)}$ contains the visible units representing the data the model is trained on, such as images. Higher layers $\mathbf{x}^{(k)}, k > 0$, consist of hidden units which learn to represent and generate the data. Together, these layers model the cortical stages of processing. The probability for a unit i to switch on is given by a sigmoid activation function,

$$P(x_i^{(k)} = 1 | \mathbf{x}^{(k-1)}, \mathbf{x}^{(k+1)}) = \frac{1}{1 + \exp(-B_i^{(k)} - T_i^{(k)})}, \tag{1}$$

with bottom-up input $B_i^{(k)} := \sum_l w_{li}^{(k)} x_l^{(k-1)} + b_i^{(k)}$ and top-down input $T_i^{(k)} := \sum_m w_{im}^{(k)} x_m^{(k+1)} + t_i^{(k)}$, which includes biases $b_i^{(k)}$ and $t_i^{(k)}$.[1]

The joint probability that the system assumes a state $\mathbf{x}$ is characterized by an energy function E,

$$P(\mathbf{x}) \propto \exp(-E(\mathbf{x})) \quad \text{with} \quad E(\mathbf{x}) = \sum_k -\mathbf{x}^{(k)T} \mathbf{W}^{(k)} \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)T} (\mathbf{b}^{(k)} + \mathbf{t}^{(k)}). \tag{2}$$

2.1 Data Sets and Plain DBM vs. RRF-DBM

Basic training works such that each hidden layer learns to generate the activities of the layer below, utilizing simple local Hebbian weight updates. We use the following data sets (Figure 1a-e): A toy dataset consisting of simple shapes at random image positions (*shapes*), and two variations thereof containing either multiple such shapes (*multi-shapes*) or clutter (*shapes+clutter*). And, the *MNIST* data set of handwritten digits, popular in machine learning, and a clutter variation (*MNIST+clutter*), using digits separated into 60,000 training and 10,000 test cases. We also compare two architectures: A plain DBM, and a more biologically inspired version where weights are restricted to be localized, realizing

[1] Two sets of biases are obtained when training the DBM layer-wise. We do not merge them as they contribute separately to bottom-up and top-down input in section 4.

receptive fields that increase in size in higher hidden layers (dubbed RRF-DBM for restrict. rec. field DBM).[2] Finally, a softmax label unit was attached to the top layer to allow for classification of the images [11].

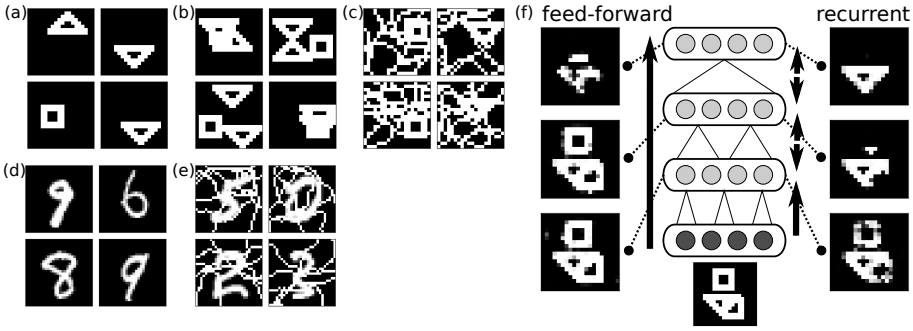


Fig. 1: (*a*-*e*) Data sets: (*a* and *d* for training). (*a*): *shapes* (squares, triangles and upside-down triangles). (*b*): *multi-shapes*. (*c*): *shapes+clutter*. (*d*): *MNIST*. (*e*): *MNIST+clutter*. (*f*): Projections of internal states across hidden layers in the RRF-DBM. The visibles are clamped to an image (bottom). After the initial bottom-up FF sweep (left), lower layers represent most of the image and the highest layer's state is unspecific. After 50 recurrent cycles (right), the highest layer has assumed an object specific state, and feed-back also biases the lower states toward the object.

3 Relation to Attentional Theories

Some theories pose that the aim of attentional processing is to form representations that are specific to one object at a time, especially in higher cortical areas [10, 5]. We thus trained the models on individual objects only (shapes or digits), but then tested them on the various cluttered data sets to see whether information about individual objects is retrieved in the highest hidden layer even when scenes are complex in ways not seen in training. To decode what is being represented in a hidden layer individually, we performed what we call top-down projections [4]: Given the layer's states, the activations of layers below are computed subsequently in a pure feed-back manner until a reconstructed image is obtained.[3] Using the reconstructed image from the top layer and its label

[2] Three hidden layers. The RRF-DBM had its number of units increased as necessary to compensate for the lower number of free parameters in the weights. No. of units: Shapes data sets: 500/500/500 (plain DBM), 26x26/26x26/26x26 (RRF-DBM). MNIST: 500/500/2000 (plain DBM), 28x28/28x28/43x43 (RRF-DBM). Receptive field sizes: 7x7/13x13/26x26. Pre-training: CD-1 for shapes, SAP (see [1]) for MNIST. No training of full DBM. Biases were initialized to -4, see section 4.

[3] This corresponds to generating from the top module in a Deep Belief Net, applied here in any hidden layer. Deterministic activations are used instead of samples.

unit, we analyzed whether the individual object was represented by computing classification and squared reconstruction errors with regards to that object.

When the model is run (performing Gibbs sampling on the joint probability), its state performs a random walk in the energy landscape along basins of attraction, which embody meaningful representations obtained during training. Because the latter are specific to individual objects by construction, the model assumes (stochastic) attractor states representing the objects being attended to [9, 5], as shown below.

Finally, the notions of a fast FF sweep and subsequent recurrent processing naturally fit into the DBM framework as well: During normal inference, processing in a DBM is recurrent in that each hidden layer is sampled taking as input the states of both adjacent layers (the top layer only receives input from below). Hidden layers can be sampled sequentially in cycles spanning the hierarchy. For the initialization however it makes sense to perform a pure bottom-up FF pass [1], ignoring respective higher layer states, as initial states there are meaningless. We found classification and reconstruction performance to be reasonable after just the initial FF pass on non-cluttered data sets. For cluttered images however, subsequent recurrent processing was important to achieve better object specific representations, in line with what is suggested for the cortex [8, 5].

3.1 Experiments: Inspection of the Hidden States

The plain DBM and the RRF-DBM were trained on the individual shapes or digits data sets, and then tested on the variations. To elucidate on what happens in the architecture during inference, an example case is displayed in Figure 1f. Here, the RRF-DBM had learned to represent individual shapes and is now run on an image of the *multi-shapes* set. Plotted are the decoded states of the three hidden layers both after the initial FF sweep and after 50 recurrent cycles. It becomes apparent that after the FF sweep, the hidden layer states are rather noisy, but the subsequent recurrent processing enables the top layer to form a clearer representation of an individual shape, allowing both for a localization of the object in image space and an improved classification.

We indeed find a shift from representing most of the scene in lower layers to representing the individual object in the highest layer. Representations are biased towards the attended object even in lower layers, but this results from feed-back from higher layers, as can be seen in the example by comparing the reconstructions of the first two hidden layers after the FF sweep and after recurrent processing. Only after the latter has taken place, involving feed-back from the topmost layer, are the representations biased toward the individual shape. In fact, when we removed the topmost layer of the RRF-DBM, no object specific state was assumed. This is partially because, due to the receptive field sizes, only the topmost layer has learned that training images only ever contained one shape. However, the deepness of the architecture plays a role in itself as well: We found that even for the plain DBM, a model with two hidden layers instead of three with the same total number of units performed worse (e.g. 43% vs. 22%

classification error on *multi-shapes*). We argue that, with higher layers being further removed from the data in terms of processing steps, there is more flexibility for the model to assume its preferred states when the data is noisy.

Finally, we point out that while the presented effect has some resemblance to how a Hopfield network can retrieve memories from noisy initializations, the DBM is a much richer model than a Hopfield net, both in a biological and a machine learning sense (see [4]). In particular, in the DBM, *latent*, hierarchical representations are retrieved from a continuously presented image, rather than memorized images from a noisy initialization.

3.2 Experiments: Quantitative Evaluation

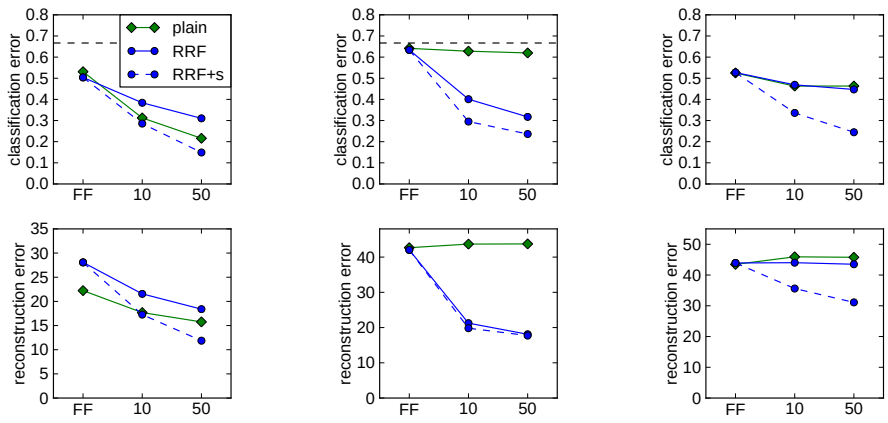
To evaluate the object specificity of the top layer states, classification and reconstruction errors were computed for the plain and RRF-DBM on the various cluttered data sets (Figure 2). For the *multi-shapes* set,[4] which is complex and novel relative to what the models had been trained on, the errors are rather high after the FF sweep, but drop profoundly after subsequent recurrent processing cycles (e.g. classification error drops from about 50% to about 20% for the plain DBM). This is true for both plain and RRF-DBM, the latter performing somewhat worse. For the noisy *shapes+clutter* set, performance is even worse after the FF sweep, with classification near chance. For the RRF-DBM, recurrent processing again helps greatly. Conversely, the plain DBM basically fails completely for this data set to retrieve the shape from the clutter. We thus conclude that at least for certain types of noise, restricted receptive fields make the DBM decidedly more robust (independently reported also in [12]).

On the other hand, for the *MNIST+clutter* set, which is based on somewhat more difficult data, recurrent processing barely improves the performance over the FF sweep. This will be addressed in the next section.

4 Top-down Suppression on Sparse Representations

Recurrent processing did not improve perception for *MNIST+clutter*. In addressing the underlying problem we can further clarify the issue of attentional processing in the architecture. Basically, the recurrent interactions in effect enable the higher layers to override image content according to what they prefer to represent. Having learned to represent individual objects, attentional selection can take place of for example one shape and suppression of others in the *multi-shapes* set. However, unlike simple toy shapes, the digits in MNIST vary much more in appearance. When presented with, for example, a digit 9 among clutter, the model should override the image representation in lower layers as to suppress the clutter. However, another way of reconciling the higher layers' expectation with the image could be to 'hallucinate' additional clutter to make the 9 into an 8. Suppression or imagination of image content are equally possible, and we find both when we decode the hidden states for *MNIST+clutter* (not shown).

[4] Errors were computed w.r.t. whichever of the three shapes was reconstructed best.



(a) *multi-shapes.* (b) *shapes+clutter.* (c) *MNIST+clutter.*

Fig. 2: Classification and reconstruction errors from top layer states for the three test sets. In each figure, scores are plotted for the plain DBM, RRF-DBM, and RRF-DBM with attentional suppression (section 4), taken after the FF sweep and after 10 or 50 subsequent recurrent cycles. Dashed lines denote chance classification error (0.9 for MNIST). (a)+(b): For the shapes sets, recurrent processing improves performance markedly, moreso with suppression. For the *shapes+clutter* set, the restricted receptive fields of the RRF-DBM are essential to retrieve the shape. (c): For *MNIST+clutter*, the additional suppressive mechanism is necessary to achieve improvement.

Thus, while the top-down influence in the DBM can be seen as implementing Hierarchical Bayesian Inference [2], for attentional top-down selection specifically we need mechanisms that increase the signal-to-noise ratio (signal being what is being attended to) without necessarily *changing* the content of the signal qualitatively, suppressing represented information related to the noise without 'hallucinating' additional content.

Two issues present in a standard DBM need to be overcome to that end: First, in a completely distributed representation, where the image is essentially encoded in the whole state vector $\mathbf{x}$, it is not clear how $\mathbf{x}$ is to be modified to achieve a suppression of image information localized to a certain part of image space. This is addressed by virtue of using the localized receptive fields, ensuring that units in lower layers only encode local information. The second issue is that switching an individual unit off (or on) does not necessarily correspond to suppressing information: For example, the unit could have inhibitory weights to the image. Indeed, we observed that for RRF-DBMs initialized with zero mean weights and biases, the learned representations are such that units tend to turn *off* when one of the shapes/digits is in their receptive fields.

To overcome the second issue, we initialized the unit biases to negative values at the beginning of training. This lead to a breaking of symmetry between units being on and off, and particularly to units being only sparsely activated

throughout training.[5] In essence they thus learned representations where they only turn on if something 'out of the ordinary' happens. In that sense, a unit conveys much more information by being on than by being off, and suppression of a unit can indeed be seen as effecting a suppression of represented information. With negatively initialized biases, units would indeed only turn on when some object (part) was in their receptive fields.[6]

With such sparse representations established, we explore a heuristic suppressive mechanism to enhance the attentional processing: Where top-down input $T_i^{(k)}$ to a unit i in hidden layer k is suppressive, i.e. $< 0.$, that input is multiplied by a factor $\zeta^{(k)}$ ($\geq 1.$). This effectively allows higher layers to suppress states in lower layers if they do not match their predictions. The modified top-down input $\tilde{T}_i^{(k)}$ is thus $= \zeta^{(k)} T_i^{(k)}$ if $T_i^{(k)} < 0$, $= T_i^{(k)}$ otherwise, so that the probability for a unit to switch on is now given as:

$$P(x_i^{(k)} = 1 | \mathbf{x}^{(k-1)}, \mathbf{x}^{(k+1)}) = \frac{1}{1 + \exp(-B_i^{(k)} - \tilde{T}_i^{(k)})}. \tag{3}$$

The RRF-DBM experiments were repeated with the suppressive mechanism active in the intermediate hidden layers[7] over 50 recurrent cycles (Figure 2). The performance increased in all cases. Particularly, for *MNIST+clutter*, recurrent processing with suppression now improved the scores markedly over the initial FF sweep.

4.1 Spatial vs. Object-Based Attention

So far we have modeled object-based attention, where higher layers can make use of learned patterns in the hidden states to emphasize object specific representations. However, the topographic sparse representations in the RRF-DBM also make it possible to apply suppressive spatial spotlights directly in the hidden layers, for example to control the internal state of the model to focus on selected objects in the *multi-shapes* set. Shortly, testing the RRF-DBM with Gaussian spotlights directed towards chosen shapes in the images, classification error computed w.r.t. the selected shapes was 18%, which is comparable to the scores reported in Figure 2 (plain DBM 22%, RRF-DBM + suppression 15%), where the models were free to select any shape in an image. Hence, spatial attention can be used to bias the internal state towards regions of the image.

[5] We found that this simple way of enforcing sparsity worked best for the problem at hand, compared to e.g. using a regularization on the gradient.

[6] Of course, this results from pixels being mostly off in the images. However, in the light of the argument, the images should themselves be understood as stand-ins for sparse representations of images, for instance the output of edge detectors, rather than as 'black and white' images.

[7] $\zeta^{(k)}$ adjusted manually for each data set and layer. Values ranged from 1 to 5.

5 Conclusion

We demonstrated in this proof of concept work how the DBM model, which uniquely embodies several properties of interest in the computational neuroscience community, can be related to theories of attentional recurrent processing in the cortex. We also elucidated on a special role of sparse representations for attentional information selection, which allowed us to explore novel mechanisms for suppressing irrelevant information. In the long run, cortical models will need to integrate sensory signals from multiple modalities with planning and motor control. We believe that accounting for attentional processing, which in the broader sense organizes information into relevant and irrelevant and routes it between cortical submodules in a task dependent fashion, will be crucial.

Acknowledgments We thank Y. Tang, N. Heess, J. Tsotsos, G. Hinton and the anon. reviewers for comments, and the EPSRC, MRC and BBSRC for funding.

References

1. Salakhutdinov, R., Hinton, G.: Deep Boltzmann machines. In: Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS). Volume 5. (2009) 448–455
2. Lee, T.S., Mumford, D.: Hierarchical Bayesian inference in the visual cortex. Journal of the Optical Society of America A **20**(7) (2003) 1434–1448
3. Fiser, J., Berkes, B., Orban, G., Lengyel, M.: Statistically optimal perception and learning: from behavior to neural representations. Trends in Cognitive Sciences **14** (2010) 119–130
4. Reichert, D.P., Series, P., Storkey, A.J.: Hallucinations in Charles Bonnet Syndrome induced by homeostasis: a Deep Boltzmann Machine model. Advances in Neural Information Processing Systems 23 **23** (2010) 2020–2028
5. Tsotsos, J.K., Rodriguez-Sanchez, A.J., Rothenstein, A.L., Simine, E.: The different stages of visual recognition need different attentional binding strategies. Brain Research **1225** (2008) 119–132
6. Deco, G., Rolls, E.T.: A neurodynamical cortical model of visual attention and invariant object recognition. Vision Research **44**(6) (2004) 621–642
7. Chikkerur, S., Serre, T., Tan, C., Poggio, T.: What and where: A bayesian inference theory of attention. Vision Research (2010) PMID: 20493206.
8. Lamme, V.A., Roelfsema, P.R.: The distinct modes of vision offered by feedforward and recurrent processing. Trends in Neurosciences **23**(11) (2000) 571–579 PMID: 11074267.
9. Serences, J.T., Yantis, S.: Selective visual attention and perceptual coherence. Trends in Cognitive Sciences **10**(1) (2006) 38–45 PMID: 16318922.
10. Rensink, R.A.: The dynamic representation of scenes. Visual Cognition **7**(1) (2000) 17
11. Hinton, G.E., Osindero, S., Teh, Y.: A fast learning algorithm for deep belief nets. Neural Computation **18**(7) (2006) 1527–1554
12. Tang, Y., Eliasmith, C.: Deep networks for robust visual recognition. In: Proceedings of the 27th Annual International Conference on Machine Learning, Haifa, Israel (2010) 1055–1062