In: European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), Bruges, Belgium, 2010

Exploiting Local Structure in Stacked Boltzmann Machines

Hannes Schulz, Andreas Müller, Sven Behnke

University of Bonn – Computer Science VI, Autonomous Intelligent Systems Group
Römerstraße 164, 53117 Bonn, Germany

Abstract.
Restricted Boltzmann Machines (RBM) are well-studied generative models. For image data, however, standard RBMs are suboptimal, since they do not exploit the local nature of image statistics. We modify RBMs to focus on local structure by restricting visible-hidden interactions. We model long-range interactions using direct or indirect lateral interaction between hidden variables. While learning in our model is much faster, it retains generative and discriminative properties of RBMs of similar complexity.

1 Introduction

One of the main tasks in unsupervised learning is modelling the data distribution. Generative graphical models, such as Restricted Boltzmann Machines (RBM, [1]) are a popular choice for this purpose. RBMs model correlations of observed variables by introducing binary latent variables (features) which are assumed to be conditionally independent given the observed variables. This restriction is useful because, in contrast to general Boltzmann Machines, a fast learning algorithm exists (Contrastive Divergence [1]). RBMs are generic learning machines and have been applied to many domains, including text, speech, motion data, and images. In the most commonly used form, however, they do not take advantage of the topology of the input space. Especially when applied to image data, fully connected RBMs model long-range dependencies which are known to be weak in natural images [2].

One way to deal with this problem is to remove long-range parameters from the model. The advantage of this approach is two-fold: first, local connectivity serves as a prior that matches well to the properties of natural images and, second, the drastically reduced number of parameters makes learning in larger models feasible. The downside of local connectivity is that weaker long-distance interactions cannot be modelled at all. In this paper, we propose to compensate for this disadvantage by introducing direct or indirect lateral interactions between the local features.

While local receptive fields are well-established in discriminative learning, their counterpart in the generative case, which we call "impact area", is not well understood. In this paper, we investigate the capabilities of stacked RBMs and RBM-like graphical models with local impact area and lateral connections. We train our architecture on the well-known MNIST database of handwritten digits [3] and demonstrate the efficiency of learning. The hidden representations can then be used for classification. With a similar number of model parameters,

	RBM-51	RBM-392	LIRBM (11×11)	LIRBM (7×7)
log-prob.	-125.58	-101.69	-108.65	-109.95
#Parameters	$40'819$	$308'504$	$39'818$	$27'058$

Table 1: Test log-likelihood on MNIST for plain RBMs (RBM-51, RBM-392) and Local Impact RBMS with different impact area sizes, measured in nats.

we find that models which exploit image structure perform better for classification. We also show that models with local impact area can generate globally consistent images. Finally, the data probability under our model compares favourably with the data probability of a fully connected RBM.

2 Background on Boltzmann Machines (BM)

A BM is an undirected graphical model with binary observed variables $\mathbf{v} \in \{0,1\}^n$ (visible nodes) and latent variables $\mathbf{h} \in \{0,1\}^m$ (hidden nodes). The energy function of a BM is given by

$$E(\mathbf{v}, \mathbf{h}, \theta) = -\mathbf{v}^T W \mathbf{h} - \mathbf{v}^T I \mathbf{v} - \mathbf{h}^T L \mathbf{h} - \mathbf{b}^T \mathbf{v} - \mathbf{a}^T \mathbf{h},$$

where $\theta = (W, I, L, \mathbf{b}, \mathbf{a})$ are the model parameters, namely pairwise visible-hidden, visible-visible and hidden-hidden interaction weights, respectively, and $\mathbf{b}$, $\mathbf{a}$ are the biases of visible and hidden activation potentials. The diagonal elements of I and L are always zero. This yields a probability distribution $p(v)$

$$p(\mathbf{v}; \theta) = \frac{1}{Z(\theta)} p^*(\mathbf{v}; \theta) = \frac{1}{Z(\theta)} \sum_{\mathbf{h}} e^{-E(\mathbf{v}, \mathbf{h}, \theta)},$$

where $Z(\theta)$ is the normalizing constant (partition function) and $p^*(\cdot)$ denotes unnormalized probability.

In RBMs, I and L are set to zero. Consequently, the conditional distributions $p(\mathbf{v}|\mathbf{h})$ and $p(\mathbf{h}|\mathbf{v})$ factorize completely. This makes exact inference of the respective posteriors possible. Their expected values are given by $\langle \mathbf{v} \rangle_p = \sigma(W\mathbf{h} + \mathbf{b})$ and $\langle \mathbf{h} \rangle_p = \sigma(W\mathbf{v} + \mathbf{b})$, where σ denotes element-wise application of the sigmoid function. In practice, Contrastive Divergence (CD, [1]) is used to approximate the true parameter gradient $\partial \ln p(\mathbf{v}) / \partial w_{i,j} = \langle \mathbf{v}^T \mathbf{h} \rangle_+ - \langle \mathbf{v}^T \mathbf{h} \rangle_-$ by a Markov Chain Monte Carlo algorithm. Here $\langle \cdot \rangle_+$ and $\langle \cdot \rangle_-$ refer to the expected values with respect to the data distribution and model distribution, respectively. Approximating these quantities is called the positive and negative phase. Recently, Tieleman [4] proposed a faster alternative, called Persistent Contrastive Divergence (PCD), which employs a persistent Markov chain to approximate $\langle \cdot \rangle_-$. We use PCD throughout this paper.

When I is not zero, the model is called a Semi-Restricted Boltzmann Machine (SRBM, [5]). This model can be trained with a variant of CD by approximating $p(\mathbf{v}|\mathbf{h})$ using a few damped mean-field updates instead of many sequential rounds

of Gibbs sampling. We will refer to this model as SRBM$^-$, as the lateral connections only play a role in the negative phase. We will later introduce our model, the SRBM$^+$, where lateral connections influence both the positive and the negative phase.

RBMs can be stacked to build hierarchical models. The training of stacked models proceeds layer-wise by training the high-level models using the activations of the hidden nodes of the layer below as input.

3 Local Impact Semi-Restricted Boltzmann Machines $^+$

We now introduce two modifications to the architectures introduced in Section 2. Firstly, we restrict the impact area of each hidden node. To this end, we arrange the hidden nodes in multiple grids, each of which resembles the visible layer in its topology. As a result, each hidden node h_j can be assigned a position $x(h_j) \in \mathbb{N}^2$ in the input (pixel) coordinate system. This approach is similar to the common approach in convolutional neural networks [3]. We then allow w_{ij} to be non-zero only if $|x(v_i) - x(h_j)| < r$ for a small constant r, where $x(v_i)$ is the pixel coordinate of v_i. In contrast to the convolutional procedure, we do not require the weights to be equal for all hidden units within one grid. We call this modification of the RBM the Local Impact RBM or LIRBM.

Secondly, we allow l_{ij} to be non-zero if $i \neq j$. Learning in this model proceeds as follows. In the positive phase the visible activations are given by the input and the hidden activations are calculated using damped mean-field updates. The mean-field updates are determined by

$$\mathbf{h}^{(0)} = \sigma(W^T\mathbf{v} + \mathbf{b}), \qquad \mathbf{h}^{(k+1)} = \sigma(W^T\mathbf{v} + \mathbf{b} + L^T\mathbf{h}^{(k)}).$$

In the negative phase, we continue our persistent Markov chain from the current state $(\mathbf{v}_p^0, \mathbf{h}_p^0)$ by sampling $\mathbf{v}_p^1$ from $p(\mathbf{v}_p^0|\mathbf{h}_p^0) = \sigma(W\mathbf{h}_p^0 + \mathbf{a})$. We then generate a new hidden state by sampling from $p(\mathbf{h}^{k+1}|\mathbf{v}_p^1) = \sigma(W^T\mathbf{v}_p^1 + \mathbf{b} + L^T\mathbf{h}_p^k)$. We call this model SRBM$^+$ since, in contrast to SRBM$^-$, the lateral connections play a crucial role in the positive phase.

A common problem in training convolutional RBMs is that the overcomplete representation of the visible nodes by the hidden nodes enables the filters to learn a trivial identity [6, 7]. The proposed SRBM$^+$ does not suffer from this problem for two reasons: First, we use PCD for training. This means even if a training example can be perfectly reconstructed, there is still a non-zero learning signal. This signal stems from the dependency of the approximated gradient on the state of the persistent Markov chain. Second, due to not sharing weights, for a trivial solution to occur, all filters have to learn the identity separately.

4 Related Work

In the context of generative models of images, little work has been done to exploit local structure. A well known supervised learning approach that makes use of the local structure of images is the convolutional neural network by LeCun et al. [3].

Fig. 1: Visualization of direct and indirect lateral interaction. Left, center left and center right: activations of randomly selected hidden nodes in LIRBM, projected down from the first, second and third layer, respectively. Right: Visualizations of lateral interactions in LIRBM$^+$, see text in Sec. 5 for details.

Another approach to exploit local structure has been suggested by Behnke [8]. LeCun's ideas where transfered to the field of generative graphical models by Lee et al. [6] and Norouzi et al. [7]. Their models, which employ weight sharing and max-pooling, discard global image statistics. Our model does not suffer from this restriction. For example, when training landscapes, our model would be able to learn, even on the lowest layer, that there is always sky depicted in the upper half of the image.

To achieve globally consistent representations in spite of the local impact area, we make use of lateral connections between the latent variables. Such connections can be modelled indirectly using stacked RBMs as in Deep Belief Networks (DBN) [1] or Deep Boltzmann Machines (DBM) [9]. Stacks of more than two RBMs, however, are not guaranteed to improve the data likelihood. In fact, even stacks of two RBMs do not improve a lower likelihood bound empirically [9]. On the other hand, stacking locally connected RBMs yields larger effective impact area for higher layers and thus can enforce more global constraints.

A more direct way to model lateral interaction is to introduce pairwise potentials between observed or between latent variables. The former case, without restrictions to local impact areas, was studied by Osindero and Hinton [5]. Furthermore, Salakhutdinov and Hinton [10] trained general BMs with lateral interactions between both observed and latent variables on small problems. Due to long training times, this general approach seems to be infeasible for larger problem sizes. Furthermore, stacking of BMs would yield two kinds of lateral interactions in each layer.

In this work, we employ lateral interactions to implicitly extend the impact area of latent variables.

5 Experimental Results

In our experiments we analyze the effects of local impact areas and lateral interaction terms on RBMs. As a first step, we trained models with and without local impact areas on the MNIST database of handwritten digits [3].

The standard RBM model had 51 hidden units, so that it had slightly higher number of parameters than the locally connected model with a hidden layer consisting of two grids of size 14×14, and an impact area of size 12×11. Throughout the paper, we use 500 epochs for fully connected models and 80

epochs for locally connected models. Please note that the reduced number of epochs is not due to longer training times of locally connected models. On the contrary, fewer parameters have to be updated and the smaller parameter space results in faster convergence.

We then approximated the likelihood assigned to the test data under these models using Annealed Importance Sampling (AIS, [9]). The likelihood is measured in "nats", meaning the natural logarithm of the unit-less probabilities. The results are summarized in Table 1. The locally connected model had a test log-likelihood of -108.65 nats whereas the standard RBM had a test log-likelihood of -125.58 nats, showing a clear advantage of our model. For comparison, a fully connected RBM with as many hidden units as the local model achieves a test log-likelihood of -101.69 nats. However, this model has eight times more parameters. Further decreasing the number of parameters by reducing the size of the impact area by factor of 0.65 does not significantly reduce the test log-likelihood.

In a second experiment, we examine the suitability of the hidden representations, which were learned without supervision, for classification. For fair comparison, we employ a k-nearest neighbor classificator with $k = 3$. We find that features learned with LIRBM are more useful (3.12%) for classification than features learned by a plain RBM (5.24%) with a similar number of parameters. As above, we also trained an RBM with 392 hidden units (2.65%) for comparison. Furthermore, we observe that our LIRBM$^+$ model yields slightly better (although insignificantly so) results (3.07%) when compared to LIRBM. Surprisingly, our LIRBM$^-$ performs much worse than all other models. It could be that the whitening operation implicitly performed by the SRBM$^-$ [5] hurts classification performance while it is helpful for generative purposes.

Next, we evaluated the influence of direct and indirect lateral interactions in the generative context. To this end, we visualized filters and fantasies generated by LIRBM and LIRBM$^+$. The left three images in Figure 1 show weights in a LIRBM with three layers. In the first image, filters of the lowest hidden layer are projected down to the visible layer. We observe that small parts of lines and line-endings are learned. The second and third figure display filters from the second and third hidden layer, projected down to the visible layer. These filters are clearly more global. With their size, the size of the recognized structure increases as well. The fourth image visualizes lateral interactions in an LIRBM$^+$. Even columns show randomly selected filters h_j of the first hidden layer, while the patch to the left of a filter depicts a linear combination of all other filters h_i, weighted by their pairwise potential l_{ij}. Note that h_j does not contribute to this sum, since $l_{jj} = 0$. We observe that through lateral interaction the filter is not only replicated, it is even extended beyond its impact area to a more global feature.

Figure 2 shows fantasies generated by LIRBM$^+$ and a stack of LIRBMs. Markov chains for fantasies were started using binary noise in the visible layer. To show quick convergence to model distribution, less iterations were used on the deeper models. It is clear that fantasies produced by a single-layer LIRBM are only locally consistent and one can observe that stacking gradually improves

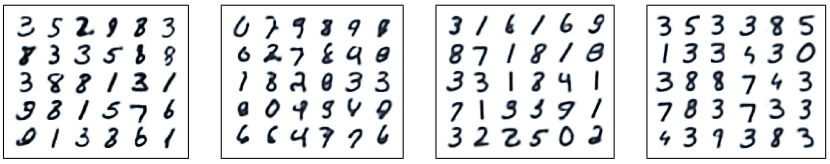

Fig. 2: Fantasies generated by our models from random noise. Left: LIRBM$^+$, one layer, 1000 steps in Markov chain. Center left: LIRBM, one layer, 1000 steps. Center right: LIRBM, two layers, 500 steps. Right: LIRBM, three layers, 250 steps. Lateral connections as well as stacking enforce global consistency.

global consistency. Lateral connections in the hidden layer significantly improve consistency even for a flat model.

These finds strongly support our initial claim that lateral interaction compensates for the negative effects of the local impact areas.

6 Conclusions

In this work, we present a novel variation of the Restricted Boltzmann Machine for image data, featuring only local interactions between visible and hidden nodes. While learning in this model is fast and few parameters yield comparably good data probabilities and classification performance, the model does not enforce global consistency constraints. We showed that this effect can be compensated for by adding lateral interactions between the latent variables, which we model directly by pairwise potentials or indirectly through stacking.

Due to the small number of parameters and its computational efficiency our architecture has the potential to model images of a much larger size than commonly used forms of RBMs.

References

[1] G.E. Hinton, S. Osindero, and W Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006.

[2] J Huang and D Mumford. Statistics of natural images and models. In *CVPR*, 1999.

[3] Y LeCun, L Bottou, Y Bengio, and P Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998.

[4] T Tieleman. Training restricted Boltzmann machines using approximations to the likelihood gradient. In *ICML*, 2008.

[5] S Osindero and G Hinton. Modeling image patches with a directed hierarchy of Markov random fields. *Advances in Neural Information Processing Systems*, 20:1121–1128, 2008.

[6] H Lee, R Grosse, R Ranganath, and Y Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *ICML*, 2009.

[7] M Norouzi, M Ranjbar, and G Mori. Stacks of Convolutional Restricted Boltzmann Machines for Shift-Invariant Feature Learning. In *CVPR*, 2009.

[8] S Behnke. *Hierarchical neural networks for image interpretation*. Springer-Verlag, 2003.

[9] R Salakhutdinov. *Learning Deep Generative Models*. PhD thesis, Univ. of Toronto, 2009.

[10] R Salakhutdinov. Learning and evaluating Boltzmann machines. Technical report, Univ. of Toronto, 2008.