Learning and Evaluaing Deep Bolztmann Machines

Ruslan Salakhutdinov and Geoffrey Hinton
Department of Computer Science, University of Toronto
rsalakhu,hinton@cs.toronto.edu

1 Abstract

Building intelligent systems that are capable of extracting high-level representations from high-dimensional sensory data lies at the core of solving many AI related tasks, including object recognition, speech perception, and language understanding. Theoretical and biological arguments strongly suggest that building such systems requires deep architectures that involve many layers of nonlinear processing. Hinton et.al. [2] introduced a fast, greedy learning algorithm for Deep Belief Networks that would learn one layer of features at a time. This new learning algorithm has generated substantial interest in academia and many variants of it have been successfully applied in many application domains. However, a crucial disadvantage of these deep probabilistic models is that the approximate inference is very limited, because it is performed in a single bottom-up pass, and will fail to adequately account for uncertainty when interpreting ambiguous sensory inputs.

In this work, we present a new learning algorithms for a different type of hierarchical probabilistic model: a deep Boltzmann machine (DBM). Unlike deep belief networks, a DBM is a type of Markov random field, or undirected graphical model, where all connections between layers are undirected. Deep Boltzmann machines are interesting for several reasons. First, like deep belief networks, DBM's have the potential of learning internal representations that become increasingly complex, which is considered to be a promising way of solving object and speech recognition problems. High-level representations can be built from a large supply of unlabeled sensory inputs and the very limited labeled data can then be used to only slightly fine-tune the model for a specific task at hand. Second, unlike existing models with deep architectures, the approximate inference procedure, in addition to bottom-up pass, can incorporate top-down feedback, allowing deep Boltzmann machines to propagate uncertainty better, and to deal more robustly with ambiguous inputs. This is perhaps the most important distinguishing characteristic of this model. Third, these deep hierarchical models can be adapted to semi-supervised learning and learning with structured outputs. Finally, in the presence of enormous amounts of sensory data, the entire model can be trained online, processing one example at a time.

Exact inference and maximum likelihood learning of DBM's are intractable. The original learning algorithm for general Boltzmann machines used randomly initialized Markov chains in order to approximate gradients of the likelihood function [3]. This learning procedure, however, was too slow to be practical. Recent advances in the machine learning, statistics, and optimization communities allow us to develop fast learning algorithms for deep multi-layer Boltzmann machines. Approximate inference can be performed using variational approaches, such as mean-field. Learning can then be carried out by applying a stochastic approximation procedure (SAP), that uses Markov chain Monte Carlo to approximate the model's expectations ([6, 7, 8, 5]). SAP belongs to the class of well-studied stochastic approximation algorithms of the Robbins-Monro type ([6, 4]) and provides nice asymptotic convergence guarantees. The idea behind these methods is straightforward. Let θ_t and X^t be the current parameters and the state. Then X^t and θ_t are updated sequentially as follows:

- Given X^t, a new state X^{t+1} is sampled from a transition operator $T_{\theta_t}(X^{t+1}; X^t)$ that leaves p_{θ_t} invariant.
- A new parameter θ_{t+1} is then obtained by replacing the intractable model's expectation by the expectation with respect to X^{t+1}.

This unusual combination of variational methods and MCMC is essential for creating a fast learning algorithm for DBM's. Furthermore, the learning procedure can be easily applied to undirected graphical models that generalize Boltzmann machines to exponential family distributions.

Figure 1 shows the architecture and the samples generated the two-layer DBM, trained on the MNIST dataset. Certainly, all samples look like the real handwritten digits. After discriminative fine-tuning, the 2-layer BM achieves an error rate of 0.95% on the full MNIST test set. This is, to our knowledge, the best published result on the permutation-invariant version of the MNIST task. This is compared to 1.4% achieved by SVM's, 1.6% achieved by randomly initialized backprop, and 1.2% achieved by the deep belief network.

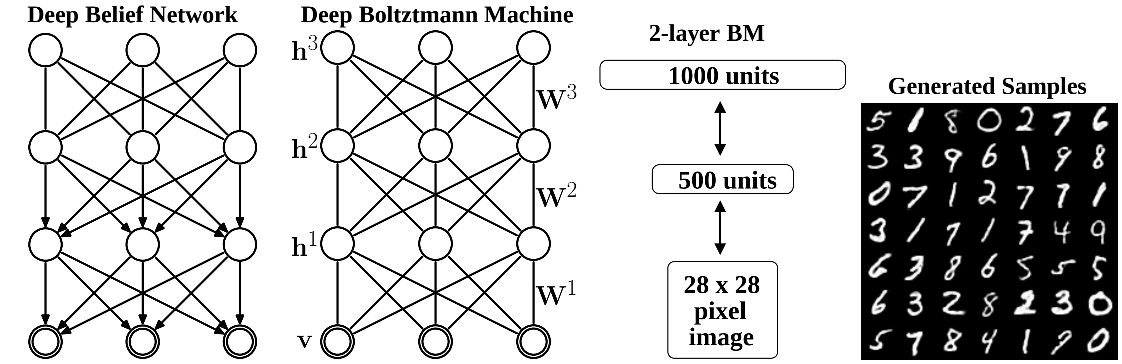

Figure 1: **Left:** Deep Belief Network vs. Deep Boltzmann Machine. **Right:** The architecture of two-layer deep Boltzmann machine used for MNIST, along with random samples generated from this DBM by running the Gibbs sampler for 100,000 steps. The images shown are the *probabilities* of the binary visible units given the binary states of the hidden units.

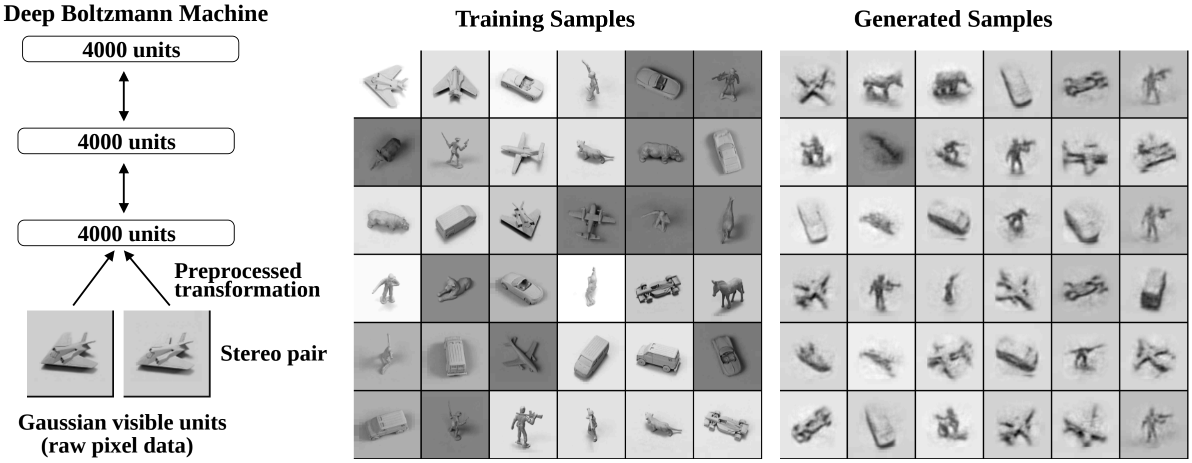

Figure 2: **Left:** The architecture of deep Boltzmann machine used for NORB. **Right:** Random samples from the training set, and samples generated from the deep Boltzmann machine by running the Gibbs sampler for 10,000 steps.

When trained on NORB dataset, the DBM achieves a misclassification error rate of 10.8% on the full test set. This is compared to 11.6% achieved by SVM's [1], 22.5% achieved by logistic regression, and 18.4% achieved by the K-nearest neighbours. To show that DBM's can benefit from additional *unlabeled* training data, we augmented the training data with unlabeled data by applying simple pixel translations, creating a total of 1,166,400 training instances. After learning a good generative model, the discriminative fine-tuning (using only the 24300 labeled training examples without any translation) reduces the misclassification error down to 7.2%. Figure 2 shows samples generated from the model. Note that the model was able to capture a lot of regularities in this high-dimensional highly-structured data, including different object classes, various viewpoints and lighting conditions.

References

[1] Y. Bengio and Y. LeCun. Scaling learning algorithms towards AI. *Large-Scale Kernel Machines*, 2007.

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

[3] G. Hinton and T. Sejnowski. Optimal perceptual inference. In *IEEE conference on Computer Vision and Pattern Recognition*, 1983.

[4] H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Stat.*, 22:400–407, 1951.

[5] T. Tieleman. Training restricted Boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the International Conference on Machine Learning*, 2008.

[6] L. Younes. Parameter inference for imperfectly observed Gibbsian fields. *Probability Theory Rel. Fields*, 82:625–645, 1989.

[7] L. Younes. On the convergence of Markovian stochastic algorithms with rapidly decreasing ergodicity rates, March 17 2000.

[8] A. L. Yuille. The convergence of contrastive divergences. In *NIPS*, 2004.