Deep learning and Boltzmann machines

KyunHyun Cho, Tapani Raiko, and Alexander Ilin

Deep learning has gained its popularity recently as a way of learning complex and large probabilistic models [1]. Especially, deep neural networks such as a deep belief network and a deep Boltzmann machine have been applied to various machine learning tasks with impressive improvements over conventional approaches.

Deep neural networks are characterized by the large number of layers of neurons and by using layer-wise unsupervised pretraining to learn a probabilistic model for the data. A deep neural network is typically constructed by stacking multiple restricted Boltzmann machines (RBM) so that the hidden layer of one RBM becomes the visible layer of another RBM. Layer-wise pretraining of RBMs then facilitates finding a more accurate model for the data. Various papers (see, e.g., [2], [1] and references therein) empirically confirmed that such multi-stage learning works better than conventional learning methods.

Unfortunately, even training a simple RBM which consists of only two layers of visible and hidden neurons is known to be difficult [7, 8]. This problem is often evidenced by the decreasing likelihood during learning. These failures have discouraged using RBMs and its extensions such as deep Boltzmann machines for more sophisticated and variety of machine learning tasks.

In our recent conference papers [4], we have proposed to use parallel tempering, an advanced Markov-chain Monte-Carlo sampling, as a replacement of a simple Gibbs sampling in obtaining samples from a model distribution defined by an RBM. It was shown that a better model with higher log-likelihood could be found using the stochastic gradient method based on PT compared to a widely-used method of minimizing contrastive divergence.

Additionally to the problem of using a simple Gibbs sampling we have determined other possible problems that discourage using an RBM as a building block for building a deep neural network. In [5] we identified density of training samples and learning hyper-parameters, such as a learning rate and an initialization of parameters, as two sources of difficulty in training RBMs. Furthermore, we also discovered that the conventional form of an energy function of Gaussian-Bernoulli RBM (GRBM) is defected in some sense that learning becomes easily unstable, in [3].

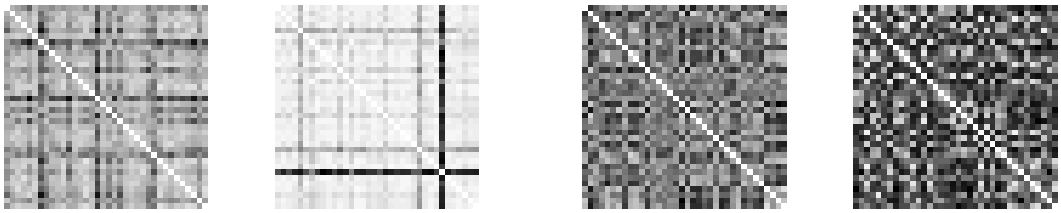

Figure 1: The angles between the update directions for the weights of an RBM with 36 hidden neurons. White pixels correspond to small angles, while black pixels correspond to orthogonal directions. From left to right: traditional gradient after 26 updates, traditional gradient after 364 updates, enhanced gradient after 26 updates, and enhanced gradient after 364 updates.

We have derived a new update direction for training RBMs, called enhanced gradient, in [5]:

$$w_{ij} \leftarrow w_{ij} + \eta_w \nabla_e \, w_{ij} \tag{1}$$
$$b_i \leftarrow b_i + \eta_b \nabla_e \, b_i \tag{2}$$
$$c_j \leftarrow c_j + \eta_c \nabla_e \, c_j, \tag{3}$$

where w_{ij}, b_i and c_j are weight between a visible neuron i and a hidden neuron j and biases for a visible neurons i and a hidden neuron j, respectively.

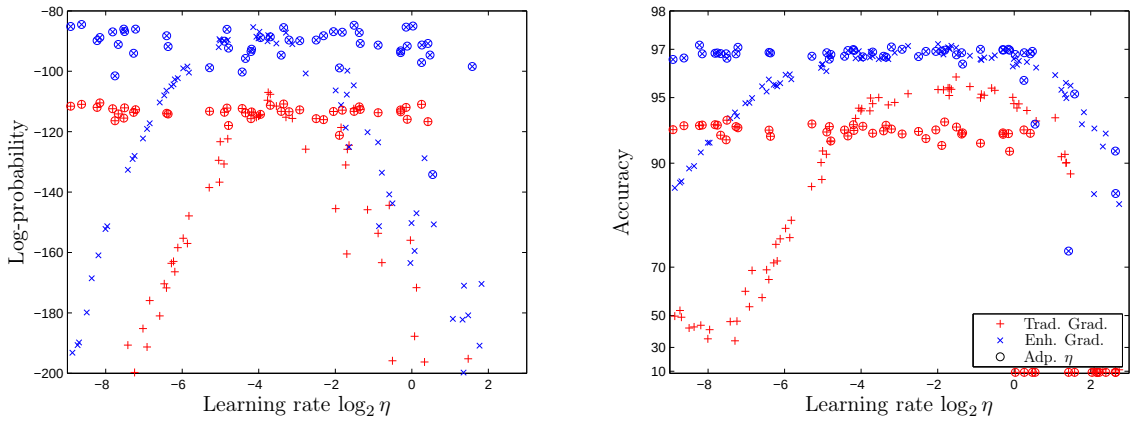

Figure 2: Log-probabilities and classificiation accuracies of test data for different initializations of the learning rate. The models were trained on MNIST using the stochastic gradient with parallel tempering.

The enhanced gradient makes learning based on the stochastic gradient invariant to the density of training samples as well as the sparsity of hidden neurons. It turned out that the enhanced gradient is more robust to the choice of learning hyper-parameters and makes the gradient per hidden neuron more orthogonal to each other as can be see in Figure 1. It was shown to help avoid a common degenerate case where most hidden neurons learn a bias.

Also in [5], we proposed a new adaptation mechanism, call adaptive learning rate, for choosing a learning rate on-the-fly. The adaptive learning rate greedily adapts the learning rate while learning parameters by maximizing the locally estimated log-likelihood. Together with the enhanced gradient, it shows in Figure 2 that more stable and better models can be trained.

All three approaches– parallel tempering, the enhanced gradient, and the adpative learning rate– have been shown to work with extensions of RBMs. In [3], we showed that these methods can be directly applied to a GRBM which replaces a binary visible neuron of an RBM with a Gaussian neuron. Furthermore, we showed that a hierarchical version of Boltzmann machines called deep Boltzmann machines (DBM) can readily use the proposed approaches in [6].

Additionally to studying Boltzmann machines for deep learning, a method of transforming a standard multi-layer perceptron by introducing linear shortcut connections and proposing transformations in non-linearities was proposed in [9]. It was shown in the paper that with the proposed transformations a faster convergence to a state-of-the-art performance can be achieved.

References

[1] Bengio, Y. (2009). Learning Deep Architectures for AI. *Foundations and Trends in Machine Learning*, 2:1–127.

[2] Salakhutdinov, R. (2009). *Learning Deep Generative Models*. PhD thesis, University of Toronto.

[3] Cho, K., Ilin, A., and Raiko, T. (2011a). Improved Learning of Gaussian-Bernoulli Restricted Boltzmann Machines. In *Proceedings of the Twentith International Conference on Artificial Neural Networks*, ICANN 2011.

[4] Cho, K., Raiko, T., and Ilin, A. (2010). Parallel Tempering is Efficient for Learning Restricted Boltzmann Machines. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2010)*, pages 3246 – 3253, Barcelona, Spain.

[5] Cho, K., Raiko, T., and Ilin, A. (2011b). Enhanced Gradient and Adaptive Learning Rate for Training Restricted Boltzmann Machines. In *Proceedings of the Twenty-seventh International Conference on Machine Learning*, ICML 2011.

[6] Cho, K., Raiko, T., and Ilin, A. (2011c). Gaussian-bernoulli deep boltzmann machine. In *NIPS 2011 Workshop on Deep Learning and Unsupervised Feature Learning*, Sierra Nevada, Spain.

[7] Fischer, A. and Igel, C. (2010). Empirical analysis of the divergence of Gibbs sampling based learning algorithms for restricted Boltzmann machines. In *Proceedings of the 20th international conference on Artificial neural networks: Part III*, ICANN'10, pages 208–217, Berlin, Heidelberg. Springer-Verlag.

[8] Schulz, H., Müller, A., and Behnke, S. (2010). Investigating Convergence of Restricted Boltzmann Machine Learning. In *NIPS 2010 Workshop on Deep Learning and Unsupervised Feature Learning*.

[9] Raiko, T., Valpola, H., and LeCun, Y. (2011). Deep Learning Made Easier by Linear Transformations in Perceptrons. In *NIPS 2011 Workshop on Deep Learning and Unsupervised Feature Learning*, Sierra Nevada, Spain.