Deep Convex Network: Architectures and Parallelizable Learning

Li Deng and Dong Yu

Microsoft Research, Redmond, WA

deng@microsoft.com, dongyu@microsoft.com

Abstract

We recently developed a DBN-HMM (Deep Belief Net/Hidden Markov Model) architecture for large-scale speech recognition with three steps in learning and decoding: 1) Stacked RBM pre-training; 2) Stochastic gradient descent back-propagation in fine tuning of DBN; and 3) Tying DBN weights in the phone-bound temporal dimension and using dynamic programming to arrive at the decision rule for speech recognition (Dahl, Yu, and Deng, ICASSP 2011 and IEEE Trans. Audio, Speech and Language Proc., 2011).

While achieving remarkable success with this approach compared with the state-of-the-art in discriminatively trained HMM, we are faced with the seemingly insurmountable problem of scalability in dealing with virtually unlimited amount of speech training data in all sorts of acoustic environments. The main difficulty comes from the stochastic gradient learning algorithm in the fine tuning step of DBN. This step of the DBN-HMM learning algorithm is extremely difficulty to parallelize over many machines with practical advantages, after some detailed exploration.

To overcome the scalability challenge, we have recently developed a novel deep learning architecture, which we call deep convex network (DCN). The learning algorithm in the DCN is a convex one in each layer of the DCN. And the learning algorithm is batch-based instead of stochastic, naturally lending it amenable for parallelization.

The construction of the DCN is related to our earlier work on building the deep hidden CRF (Yu, Wang, Deng, IEEE J. Selected Topics in Signal Proc., Dec 2010), but the learning algorithm is much simpler in each layer, enabling the use of hundreds of layers in the overall DCN architecture. We have empirically found that without doing global error fitting over all layers, the DCN can already achieve excellent discrimination results as long as a very deep architecture is built. One version of the algorithm contains a module that makes use of nonlinear random projection inspired by the work of extreme learning machine (Huang, 2006). While this module gives reasonably good results after it is embedded in the deep structure, we found that replacing it with the restricted Boltzmann machine (RBM) contributes to over 30% of error reduction in the MNIST experiment. We will present detailed experimental results in the workshop.

We are grateful to many helpful discussions with and suggestions and encouragements of this work by G.-B. Huang, John Platt, Geoff Hinton, Dave Wecker, and Alex Acero. The code for nonlinear random projection was provided and explained by Dr. G.-B. Huang.

References:

- George Dahl, Dong Yu, Li Deng, and Alex Acero, Large Vocabulary Continuous Speech Recognition With Context-Dependent DBN-HMMS, in *Proc. IEEE-ICASSP, Prague*, May 2011
- Dong Yu, Shizhen Wang, and Li Deng, Sequential Labeling Using Deep-Structured Conditional Random Fields, in *IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING*, IEEE, December 2010
- Dong Yu, Li Deng, and George E. Dahl, Roles of Pre-Training and Fine-Tuning in Context-Dependent DBN-HMMs for Real-World Speech Recognition, in *NIPS 2010 workshop on Deep Learning and Unsupervised Feature Learning*, December 2010
- Abdel-rahman Mohamed, Dong Yu, and Li Deng, Investigation of Full-Sequence Training of Deep Belief Networks for Speech Recognition, in *Interspeech 2010*, September 2010
- Li Deng, Mike Seltzer, Dong Yu, Alex Acero, Abdel-rahman Mohamed, and Geoff Hinton, Binary Coding of Speech Spectrograms Using a Deep Auto-encoder, in *Interspeech 2010*, September 2010
- G.-B. Huang, L. Chen and C.-K. Siew, “Universal Approximation Using Incremental Constructive Feedforward Networks with Random Hidden Nodes”, *IEEE Transactions on Neural Networks*, vol. 17, no. 4, pp. 879-892, 2006.