NEW TYPES OF DEEP NEURAL NETWORK LEARNING FOR SPEECH RECOGNITION AND RELATED APPLICATIONS: AN OVERVIEW

Li Deng[1], Geoffrey Hinton[2], and Brian Kingsbury[3]
[1]Microsoft Research, Redmond, WA, USA
[2]University of Toronto, Ontario, Canada
[3]IBM T. J. Watson Research Center, Yorktown Heights, NY, USA

ABSTRACT

In this paper, we provide an overview of the invited and contributed papers presented at the special session at ICASSP-2013, entitled "New Types of Deep Neural Network Learning for Speech Recognition and Related Applications," as organized by the authors. We also describe the historical context in which acoustic models based on deep neural networks have been developed.

The technical overview of the papers presented in our special session is organized into five ways of improving deep learning methods: (1) better optimization; (2) better types of neural activation function and better network architectures; (3) better ways to determine the myriad hyper-parameters of deep neural networks; (4) more appropriate ways to preprocess speech for deep neural networks; and (5) ways of leveraging multiple languages or dialects that are more easily achieved with deep neural networks than with Gaussian mixture models.

Index Terms— deep neural network, convolutional neural network, recurrent neural network, optimization, spectrogram features, multitask, multilingual, speech recognition, music processing

1. INTRODUCTION

In recent years, the speech recognition community has seen a revival of interest in neural networks, which were popular during late 80's and early 90's but could not significantly outperform the very successful combination of HMMs with acoustic models based on Gaussian mixtures. Three main factors were responsible for the recent emergence of neural networks as high-quality acoustic models: (1) making the networks deeper makes them more powerful, hence deep neural networks (DNN); 2) initializing the weights sensibly [24][43][16][52] and using much faster hardware makes it possible to train deep neural networks effectively, and 3) using a larger number of (context-dependent) output units [8][10][48][49][53] greatly improves their performance.

The papers presented in this special session feature both the current state-of-the-art practice of DNNs and some promising new developments beyond the standard network architectures and learning methodologies. The new types of DNN models and learning techniques hold promise for creating better technology for future-generation speech recognition and possibly other applications.

To help readers understand and appreciate the material presented in our special session, we include an overview of the historical context in which DNN technology has been developed. The application areas covered include speech recognition, music processing, and language processing.

2. SPECIAL SESSION MOTIVATIONS

Deep learning has become increasingly popular [38] since the introduction of an effective new way of learning deep neural networks in 2006 [25][26]. It has proved very successful for acoustic modeling in speech recognition especially for large-scale tasks, and this success has been based largely on the use of the back-propagation algorithm with rather standard, feed-forward multi-layer neural networks; see a comprehensive review in [24] and reviews of earlier work in [6][44]. In addition to improved learning procedures, the main factors that have contributed to the recent successes of deep neural networks have been the availability of more computing power, the availability of more training data, and better software engineering. The initial breakthrough in acoustic modeling was triggered by the use of a generative, layer-by-layer pre-training method for initializing the weights sensibly before running the discriminative back-propagation learning procedure, but subsequent research has revealed that generative pre-training is unnecessary when there is a very large amount of labeled training data. Back-propagation can be started from random initial weights provided their scales are carefully determined to prevent the initial error derivatives from being very large or very small.

More than a year ago, four research groups (a group at Google plus the three groups represented by the current organizers) wrote an overview article [24] in which they presented their shared views on applying DNNs to acoustic modeling in speech recognition. Since then, the four groups and other speech or machine learning groups around the world have done a lot of new work developing new models and learning methods, and performing new evaluation experiments. The main aim of this special session is to highlight advances in the application of DNNs over the last year.

3. THE RECENT HISTORY OF DEEP NEURAL NETWORKS FOR ACOUSTIC MODELING

The DNNs that first showed big improvements over Gaussian Mixture Models (GMMs) for acoustic modeling all used minor variations of the same successful recipe, but training was so slow, even on GPUs, that it was impossible to perform the extensive experimentation required to establish which aspects of this recipe made it successful. What was important initially was to find any reasonable way of training DNNs that allowed them to outperform

GMMs and shallow neural nets. The particular recipe was just the first successful method to be developed and it was based on a lot of intuitive guesses without much evidence to support the individual decisions.

The successful recipe that was originally used in [42] presented at the NIPS-2009 Workshop [17] to train an acoustic model for speech recognition on the TIMIT database differed in several ways from previous attempts to use neural networks for acoustic modeling. The nets were much deeper and larger than previous attempts, having up to eight hidden layers with a few thousand hidden units per layer and full connectivity between adjacent layers. The final fine-tuning of the nets used the standard, discriminative back-propagation algorithm to compute gradients and stochastic gradient descent with momentum to update the weights, but before the fine-tuning started, the weights were initialized by using an unsupervised learning algorithm that had no knowledge of the labels used for fine-tuning. The unsupervised learning algorithm learned one hidden layer of binary stochastic features at a time with the aim of the learning being to model the statistical structure of the patterns of feature activations in the layer below (or in the MFCCs when learning the first hidden layer). Results were found to be only slightly superior to the then-best-performing single system, which was built on a deep/dynamic generative model called the hidden trajectory model (HTM) [14][15], in the literature and evaluated on the identical task. The error patterns produced by these two separate systems (DNN-HMM vs. HTM) were carefully analyzed at MSR and found to be very different, reflecting distinct core capabilities of the two approaches and motivating further studies on the DNN approach.

Over the following few years, researchers using DNNs for speech recognition discovered a lot of things about this recipe:

1) It works well for LVCSR and it works even better for LVCSR if the DNN's output units correspond to context dependent HMM states, and importantly, this choice keeps the decoding algorithm largely unchanged.

2) When there is a large amount of labeled data, the main effect of the pre-training is just to get the initial weights to be about the right scale so that back-propagation works well. But there are simpler ways of doing this.

3) Even if we use layer-by-layer pre-training, there are many alternatives to using Restricted Boltzmann Machines (RBMs) for pre-training each layer.

4) DNNs work significantly better on filterbank outputs than on MFCCs.

5) Speaker-dependent methods provide surprisingly little improvement over speaker-independent DNNs. While this was initially somewhat disappointing, using speaker-independent models reduces computational expense and latency when these models are used in applications.

6) DNNs work well for noisy speech.

7) Using full connectivity between the early layers is simple but not sensible. DNNs work much better for acoustic modeling if we use one or more convolutional layers that do weight-sharing across nearby frequencies and then pool the filter responses to similar frequencies thus giving some invariance to vocal tract differences between speakers.

8) Using standard logistic neurons is sensible but not optimal. DNNs learn much faster if we use rectified linear units. These also overfit faster but a new regularization method called "dropout" is very effective at controlling this overfitting.

9) The same methods can be used for applications other than acoustic modeling.

10) The DNN architecture can be used for multi-task learning in several different ways and DNNs are far more effective than GMMs at leveraging data from one task to improve performance on related tasks.

Some of these new discoveries about applying DNNs to speech recognition are described in the papers we selected for our special session and we describe those discoveries in more detail in the next section.

4. OVERVIEW OF THE SPECIAL SESSION PAPERS

Here we provide a technical overview of the five papers selected for the special session. The technical overview covers five promising ways of improving deep learning methods: (1) better optimization; (2) better types of neural activation function and better network architectures; (3) better ways to optimize the myriad hyper-parameters of DNNs; (4) more appropriate ways to pre-process speech for DNNs; and (5) ways of leveraging multiple languages or dialects that are more easily achieved with DNNs than with Gaussian mixture models.

Online, stochastic gradient descent has been the workhorse for neural network training, including deep learning, for over 25 years. This is not an accident, for stochastic gradient descent enjoys a number of advantages: it is very easy to implement; it makes extremely rapid progress per training sample processed [5], and well-implemented stochastic gradient descent (where care is taken in the randomization of training samples and choice of learning rates) frequently converges to better local optima than other algorithms.

The main problems with stochastic gradient descent have been the challenge of scaling to very large data sets and networks with many parameters and the challenge of learning very deep or recurrent neural network models. For scaling up deep learning, the most common recent solution has been the use of GPU hardware. The paper from Google [22] is notable because it features a distributed framework for deep learning that successfully uses a large compute cluster. The framework, called DistBelief [11][35], uses an asynchronous version of stochastic gradient descent that uses many different replicas of the neural net to compute gradients on different subsets of the training data in parallel. These gradients are communicated to a central parameter server that updates the shared weights and even though each replica will typically be computing gradients using slightly stale parameter values, stochastic gradient descent is robust to the slight errors this introduces. DistBelief also distributes the implementation of each replica across many cores which greatly increases the degree of parallelization.

Training very deep neural networks with stochastic gradient descent is difficult because the gradients tend to decrease as they are back-propagated through multiple levels of nonlinearity.

DistBelief deals with this problem by separately adapting the learning rate for each parameter. For recurrent neural networks, which are typically very deep in time, the "vanishing gradients" problem is even more severe. Recently developed versions of semi-online, second order optimization methods that use stochastic curvature estimates, such as Hessian-free optimization [39][40] have revitalized work on recurrent network models. Hessian-free training is used in the IBM paper [48] for sequence-discriminative training.

The paper from the University of Montreal [3] explores the training of recurrent models using modifications to stochastic gradient descent, and, following the work of [51], shows that these modifications can outperform Hessian-free baselines. Optimization ideas that are explored in this work include clipping the gradient if its norm exceeds a threshold and a new formulation of Nesterov accelerated gradient training.

The use of recurrent neural networks for acoustic modeling was pioneered by Tony Robinson [47] but they then fell out of favor because of the difficulty of training them. Recently, however, recurrent neural networks have achieved excellent results at language modeling [41] and the use of multiple hidden layers has allowed recurrent neural networks to outperform all other methods on TIMIT [21].

Related to optimization in deep learning are problems of regularization. Generative pre-training and standard methods such as weight decay (L2 regularization) are important, but beyond those are other useful ideas. The paper from the University of Toronto [9] describes "dropout," which is a regularization method that randomly omits some fraction of the units in each hidden layer during training. This procedure discourages brittle co-adaptations in which a hidden unit is useful only in the context of specific other hidden units. The dropout method is easily implemented and improves the performance of DNNs on a wide variety of standard benchmarks including TIMIT [23]. In the paper submitted to this special session, it is shown that dropout regularization can be combined with rectified linear hidden units to improve speech recognition on a 50-hour broadcast news task. Another regularization method that is proving its worth is the application of a sparsity penalty to hidden representations in a network, which is explored in [3].

A major barrier to the application of DNNs is that it currently requires considerable skill and experience to choose sensible values for hyper-parameters such as the learning rate schedule, the strength of the regularizer, the number of layers and the number of units per layer. Sensible values for one hyper-parameter may depend on the values chosen for other hyper-parameters and hyper-parameter tuning in DNNs is especially expensive because testing a single setting of the hyper-parameters is costly. Papers in this special session describe two methods for tackling this problem: paper [9] uses an off-the-shelf Bayesian optimization procedure [50], while paper [3] employs a sampling procedure [4] to avoid the expense of a full grid search.

Exploring different types of neuron activation function and different network architectures is a theme common to many papers in this session. Both papers [9] and [3] explore the use of rectified linear hidden units instead of logistic or tanh nonlinearities. Rectified linear units compute $y = \max(x, 0)$, and lead to sparser gradients, less diffusion of credit and blame in deep or recurrent networks, and faster training [54]. Paper [3] also proposes the use in recurrent networks of an explicit subset of leaky integrator units in the state-to-state map to better capture long-range dependencies, as well as the use of a powerful output probability model, the neural autoregressive distribution estimator [34].

Convolutional neural networks have been widely used in computer vision [36] where they have been very successful [32]. They showed early promise for acoustic modeling [33] but were later abandoned, probably because the convolution was done across time rather than across frequency. Temporal variation is already well-handled by the HMM so convolution across frequency is much more helpful because it provides partial invariance to changes in the properties of the vocal tract. In an important paper, Abdel-Hamid et. al. [1] demonstrated that convolution across frequency was very effective for TIMIT. More recent work described in the papers from Microsoft [12][2][13] shows that designing the convolution and pooling layers to properly trade-off between invariance to the vocal tract length and discrimination among speech sounds together with the "dropout" technique of regularization [27] leads to much better TIMIT phone recognition accuracy. This set of work also points to the direction of trading-off between trajectory discrimination and invariance expressed in the whole dynamic pattern of speech defined in mixed time and frequency domains using well designed weight sharing and pooling. The IBM paper [48] shows that convolutional neural networks are also useful for LVCSR and further demonstrates that multiple convolutional layers provide even more improvement when the convolutional layers use a large number of convolution kernels (i.e. feature maps).

In light of the powerful DNN learning architectures and algorithms developed recently, it is useful to re-examine some long-standing assumptions about the best ways to pre-process speech for acoustic modeling. Spectrograms contain rich information, but systems that use GMMs for acoustic modeling work best with transformed features such as MFCCs or PLPs whose elements are largely de-correlated. The powerful learning procedures for DNNs allow them to handle correlations between input features and also allow them to transform spectrograms in whatever way they want. This completely changes the conventional wisdom about the kind of pre-processing that is most helpful. The paper by Microsoft [12] analyzes the fundamental issue of what are effective features for use in the pattern recognition component of speech recognition. It reviews the use of spectrograms as the input features for deep auto-encoders to extract bottleneck higher-level features [18] and the extended work on multi-modal deep auto-encoder [45] (see the most recent work on audio-visual deep learning in [28]). It also presents the recent results on a DNN-based, large vocabulary speech recognizer with (Mel-scaled) spectrograms as the input features which outperforms the same recognizer but with MFCCs as the input features.

The final theme that emerges from the set of selected papers is the excellent performance of DNN acoustic models for multi-task learning [7]. Shallow models such as the GMMs used in the previous generation of acoustic models do not benefit nearly as much as DNNs from being trained on multiple languages simultaneously or from being trained on one language and then modified for another language (e.g. [37]). Both the Microsoft paper [12] (also [29]) and the Google paper [22] elaborate such a new capability, sharing the same example of multilingual speech recognition. In Figure 1, the multi-task learning accomplished by DNN is shown for two scenarios: a) with high practical value: learning joint representation for both 16k and 8k acoustic data for performing recognition for both wideband (e.g., high quality smart phone) voice search and narrowband telephony speech recognition, and b) multilingual or cross-lingual speech recognition that

effectively leverages acoustic training data across a wide range of languages.

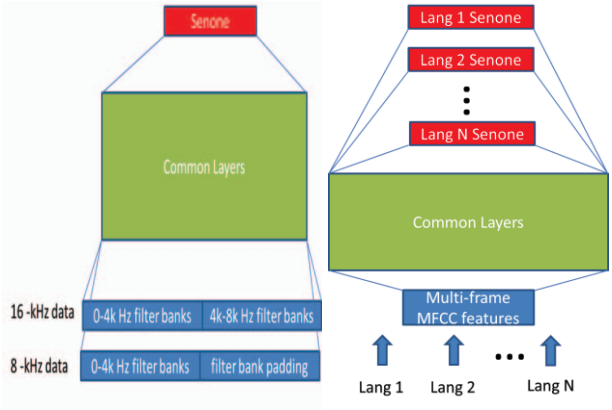

Figure 1: a) left: DNN training/testing with mixed-band acoustic data with16-kHz and 8-kHz sampling rates; b) right: Illustrative architecture for a multilingual DNN

5. CONCLUSIONS

In summary, the articles in our special session demonstrate that there continues to be rapid progress in acoustic models that use DNNs and that similar methods are also applicable in related domains such as music. The progress is occurring on many different fronts and is widening the already significant performance gap between acoustic models based on DNNs and those based on GMMs. We believe that the lessons we are learning in acoustic modeling are likely to be relevant to a wide range of other signal processing, language processing, machine learning, and artificial intelligence tasks.

6. ACKNOWLEDGEMENTS

We thank all authors for contributing their papers to our special session and thank anonymous reviewers who provided valuable feedback to us.

7. REFERENCES

[1] O. Abdel-Hamid, A. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition," ICASSP, 2012.

[2] O. Abdel-Hamid, L. Deng, and D. Yu. "Exploring convolutional neural network structures and optimization for speech recognition," Interspeech, 2013, submitted.

[3] Y. Bengio, N. Boulanger, and R. Pascanu. "Advances in optimizing recurrent networks," ICASSP, 2013.

[4] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," J. Machine Learning Research," Vol. 3, pp. 281-305, 2012.

[5] L. Bottou and Y. LeCun. "Large scale online learning," NIPS, 2004.

[6] H. Bourlard and N. Morgan, Connectionist Speech Recognition: A Hybrid Approach, Norwell, MA: Kluwer, 1993.

[7] R. Caruana, "Multitask Learning." Machine Learning, Vol. 28, pp. 41-75, Kluwer Academic Publishers, 1997.

[8] G. Dahl, D. Yu, L. Deng, and A. Acero, "Large vocabulary continuous speech recognition with context-dependent DBN-HMMs," ICASSP, 2011.

[9] G. Dahl, T. Sainath, and G. Hinton. "Improving DNNs for LVCSR using rectified linear units and dropout," ICASSP, 2013.

[10] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large vocabulary speech recognition." IEEE Trans. Speech and Audio Proc., vol. 20, no. 1, pp. 30 – 42, 2012.

[11] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, Q. Le, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, and A. Ng. "Large scale distributed deep networks," NIPS, 2012.

[12] L. Deng, J. Li, J. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams, Y. Gong, and A. Acero. "Recent advances of deep learning for speech research at Microsoft," ICASSP, 2013.

[13] L. Deng, O. Abdel-Hamid, and D. Yu. "Deep convolutional neural networks using heterogeneous pooling for trading-off acoustic invariance with phonetic confusion," ICASSP, 2013.

[14] L. Deng, D. Yu, and A. Acero. "Structured Speech Modeling," IEEE Trans. on Audio, Speech and Language Processing. Volume: 14 Issue: 5, Sep 2006. pp. 1492- 1504.

[15] L. Deng and D. Yu. "Use of differential cepstra as acoustic features in hidden trajectory modeling for phonetic recognition," ICASSP, 2007.

[16] L. Deng, D. Yu, and J. Platt. "Scalable stacking and learning for building deep architectures," ICASSP, 2012.

[17] L. Deng, D. Yu, and G. Hinton. "Deep Learning for Speech Recognition and Related Applications" NIPS Workshop, 2009 http://nips.cc/Conferences/2009/Program/event.php?ID=1512

[18] L. Deng, M. Seltzer, D. Yu, A. Acero, A. Mohamed, and G. Hinton, "Binary coding of speech spectrograms using a deep auto-encoder," Interspeech, 2010.

[19] L. Deng, G. Tur, X. He, and D. Hakkani-Tur, "Use of kernel deep convex networks and end-to-end learning for spoken language understanding," IEEE SLT, 2012.

[20] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," J. Machine Learning Research, 2011, pp. 2121-2159.

[21] A. Graves, A. Mohamed, and G. Hinton. "Speech recognition with deep recurrent neural networks," ICASSP, 2013.

[22] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean. "Multilingual acoustic models using distributed deep neural networks," ICASSP, 2013.

[23] G. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and Salakhutdinov. "Improving neural networks by preventing co-adaptation of feature detectors," 2012. http://arxiv.org/abs/1207.0580.

[24] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," IEEE Signal Processing Magazine, Vol. 29 (6), pp. 82-97, 2012.

[25] G. Hinton, S. Osindero, and Y. Teh. "A fast learning algorithm for deep belief nets. Neural Computation, Vol. 18, 1527-1554, 2006.
[26] G. Hinton and R. Salakhutdinov. "Reducing the dimensionality of data with neural networks. Science, Vol. 313. no. 5786, pp. 504 - 507, July 2006.
[27] G. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, & R. Salakhutdinov. "Improving neural networks by preventing co-adaptation of feature detectors," arXiv: 1207.0580v1, 2012.
[28] J. Huang and B. Kingsbury. "Audio-visual deep learning for noise robust speech recognition," ICASSP, 2013.
[29] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong. "Cross-language knowledge transfer using multilingual deep neural networks with shared hidden layers," ICASSP, 2013.
[30] N. Jaitly, P. Nguyen, and V. Vanhoucke, "Application of pre-trained deep neural networks to large vocabulary speech recognition," Interspeech, 2012.
[31] B. Kingsbury, T. N. Sainath, and H. Soltau, "Scalable minimum Bayes risk training of deep neural network acoustic models using distributed Hessian-free optimization," Interspeech, 2012.
[32] A. Krizhevsky, I. Sutskever, I. and G. Hinton. "ImageNet classification with deep convolutional neural Networks," NIPS 2012.
[33] K. Lang, A. Waibel, and G. Hinton. "A time-delay neural network architecture for isolated word recognition," Neural Networks, Vol. 3(1), pp. 23-43, 1990.
[34] H. Larochelle and I. Murray. "The neural autoregressive distributed estimator," ICML, 2011.
[35] Q. Le, M. Ranzato, R. Monga, M. Devin, K. Chen, G. Corrado, J. Dean, A. Ng, "Building high-level features using large scale unsupervised learning." ICML, 2012.
[36] Y LeCun, L Bottou, Y Bengio, P Haffner. "Gradient-based learning applied to document recognition," Proceedings of the IEEE 86 (11), 2278-2324.
[37] H. Lin, L. Deng, D. Yu, Y. Gong, A. Acero, and C-H Lee, "A study on multilingual acoustic modeling for large vocabulary ASR." ICASSP, 2009, pp. 4333–4336.
[38] J. Markoff. "Scientists See Promise in Deep-Learning Programs," New York Times, Nov 24, 2012.
[39] J. Martens. "Deep learning via Hessian-free optimization," ICML, 2010.
[40] J. Martens and I. Sutskever. "Learning Recurrent Neural Networks with Hessian-Free Optimization," ICML, 2011.
[41] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, and S. Khudanpur. "Recurrent neural network based language model," Interspeech, 2010.
[42] A. Mohamed, G. Dahl, and G. Hinton, "Deep belief networks for phone recognition," in Proc. NIPS Workshop Deep Learning for Speech Recognition and Related Applications, 2009.
[43] A. Mohamed, G. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," IEEE Trans. on Audio, Speech, and Language Processing, vol. 20, no. 1, pp. 14–22, 2012.
[44] N. Morgan. "Deep and Wide: Multiple Layers in Automatic Speech Recognition," IEEE Trans. on Audio, Speech, and Language Processing, vol. 20, no. 1, pp. 7-13, 2012.
[45] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, "Multimodal deep learning," ICML, 2011.
[46] T. N. Sainath, B. Kingsbury, B. Ramabhadran, P. Fousek, P. Novak, and A. Mohamed, "Making deep belief networks effective for large vocabulary continuous speech recognition", Proc. ASRU, pp. 30-35, 2011.
[47] A. Robinson, "An application to recurrent nets to phone probability estimation," IEEE Trans. Neural Networks, vol. 5, no. 2, pp. 298–305, 1994.
[48] T. Sainath, A. Mohamed, B. Kingsbury, B. Ramabhadran, "Convolutional neural networks for LVCSR," ICASSP, 2013.
[49] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," Interspeech, 2011, pp. 437-440, 2011.
[50] J. Snoek, H. Larochelle, and R. Adams, "Practical Bayesian optimization of machine learning algorithms," NIPS, 2012.
[51] I. Sutskever. "Training Recurrent Neural Networks," Ph.D. Thesis, University of Toronto, 2013.
[52] D. Yu, L. Deng, G. Li, and Seide F, "Discriminative pretraining of deep neural networks," U.S. Patent Filing, Nov. 2011.
[53] D. Yu, L. Deng, and G. Dahl, "Roles of pretraining and fine-tuning in context-dependent DNN-HMMs for real-world speech recognition," NIPS Workshop on Deep Learning and Unsupervised Feature Learning, Dec. 2010.
[54] M. Zeiler, M. Ranzato, R. Monga, M. Mao, K. Yang, Q. Le, P. Nguyen, A. Senior, V. Vanhoucke, J. Dean, and G. Hinton. "On rectified linear units for speech processing," ICASSP, 2013.