**COLLEGE OF ENGINEERING**

Preliminary Exam Report

**Signal Modeling Using Deep Neural Networks**

Seyedeh Saeedeh Khoshgoftar Ziyabari

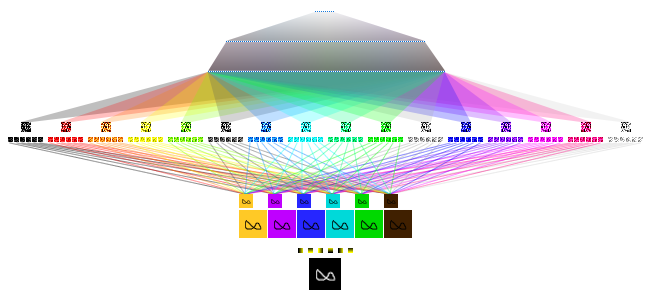
saeedeh@temple.edu

Institute for Signal and Information Processing,

Department of Electrical Engineering,

Temple University

February 2016



**Executive Summary**

Deep neural networks are becoming a fundamental component of high performance speech recognition systems. Performance of deep learning based systems (DNN) has recently exceeded that of their Gaussian mixture model (GMM) counterparts in acoustic modeling. In this talk, we will review three types of deep learning systems that are used for signal modeling: (1) a deep convolutional neural network (CNN) which is used to reduce spectral variation and model spectral correlations in signals; (2) a convolutional deep belief network (CDBN) that is a classic generative probabilistic model composed of convolutional restricted Boltzman machines (CRBM); (3) a recurrent neural network (RNN) that is used to model sequential data. Joint CNN/DNN models are used to achieve robust signal modeling. These systems give state of the art performance on Broadcast News (13.6% WER on 50 hrs of data, 12.7% WER on 400 hours of training data) and Switchboard (10.7% WER on 300 hours of training data). These results represent a substantial reduction in error rate over the previous best DNN results (2.7% and 2.4% absolute respectively on Broadcast News and 1.5% on Switchboard). We will review these results and discuss their implications for automated interpretation of EEGs.

In writing this review, one of my primary intentions is to find the most robust model for EEG signals using deep neural networks in order to detect seizure automatically with less error rate.

Table of Contents

[1 Introduction 1](#_Toc443381800)

[2 Deep convolutional neural network 1](#_Toc443381801)

[2.1 Applications 2](#_Toc443381802)

[3 Speech recognition models of ibm 4](#_Toc443381803)

[4 Convolutional deep belief network 6](#_Toc443381804)

[4.1 Application 7](#_Toc443381805)

[5 Conclusion and future work 8](#_Toc443381806)

[6 Reference 8](#_Toc443381807)

# Introduction

Most of typical speech recognition systems use hidden Markov models (HMMs) to deal with the temporal variability of speech and Gaussian mixture models to determine how well each state of each HMM fits a frame or a short window of frames of coefficients that represents the acoustic input. GMM is statistically inefficient for modeling data that lie on or near a non-linear manifold in the data space. The neural network work better than GMM for acoustic modeling of data that lie on or near nonlinear manifold .the more effective form of training comes from deep neural network which feed-forward, artificial neural network that has more than one layer of hidden units between its inputs and its outputs. Each hidden unit, j, typically uses the logistic function (Equation 1) to map its total input from the layer below, , to the scalar state, that it sends to the layer above (Equation 2) [1].

(1)

(2)

Deep neural networks (DNN) have diverse models that some of the robust ones are called Deep convolutional neural network (CNN), deep belief network (DBN) and recurrent neural network (RNN).

This report is organized into five sections. Section two is review of deep convolutional neural network (CNN) and its application. Section three is devoted to the speech recognition models of IBM and section four is focus on deep belief network (DBN) with its application in EEG signal and last chapter is conclusion and future work.

# Deep convolutional neural network

One of the promising deep neural networks is convolution neural network (CNN). It was historically used for hand writing and image processing. Recently, it has been used for speech recognition, also showing better performance over DNN [2]. There are some main properties in CNN that lead to better speech recognition performance. It used to reduce translational variance in signal and model spatial temporal correlation [3], [4]. As, the spectral representation of speech have interconnection in time and frequency, the CNN finds the local correlations and share them across the local region of input space [5]. CNN is more robust to slight formant shift due to different speaking style of speakers at low frequency region.

A typical convolutional neural network (CNN) is depicted in . The input feature with dimension t\*f (time \*frequency) is given to the input layer. The whole input will be convolved by the weight matrix (W) in order to find the local correlation in input signal. The overall convolutional operation produces n feature maps. The next layer is max-pooling that do the subsampling in order to reduce the time-frequency space. It helps to remove the variability in time-frequency space which is generated by speaking styles and channel distortions.



Figure 1:Typical convolutional neural network (CNN).

## Application

T.N. Sainath et al [6] proposed the joint CNN/DNN architecture which is illustrated in . They used 40 dimensional VTLN-warped mel-filtered bank+delta +double delta coefficients. A temporal context of 11 frames is used as input to train the CNN. The CNN and fully connected DNN have 1024 hidden unites per fully connected layer with sigmoid non-linearity. The output layer is softmax layer with 512 output targets. Both DNN and CNN are trained by optimizing the cross-entropy adjective function. The equation 3 shows the cross-entropy loss function. The denoted for the set of posterior probability for each target and is the reference targets which has a probability 1 for correct class and 0 for incorrect class.

(3)

The system uses the full weight sharing (FWS) with 2 convolutional layers of 256 hidden units followed by 4 fully connected layers of 1024 hidden united. The first and second convolutional layers use the 9\*9 and 4\*3 frequency-time filters respectively. Pooling size 3 was used for first layer and no polling for second layer. The system use the non-overlapping max pooling and pooling in frequency only. A pooling size was 3 in first layer and no pooling was done in second layer.

The various speaker adaption techniques are used to improve speech recognition performance like feature-space maximum likelihood linear regression features (fMLLR) and speaker identity vector (i-vector). Incorporating i-vector and fMLLR into CNN architecture are a bit challenging because CNN require features which obey a frequency and time locality property, while the i-vectore and fMLLR do not have it. Typically, fMLLR applied to linear discriminate analysis (LDA) features, which the locality in frequency are removed, therefore CNN unlike DNN cannot be used. As it is demonstrated in , feeding fully connected DNN layers with i-vector and fMLLR features and combining the output with one fully connected layer of the CNN, provide 10% improvement in WER over the baseline. The architecture of combining CNN and DNN is shown in .



Figure 2: Joint CNN/DNN architecture.



Table 1: WER with speaker-adopted feature.

Since the speech are sequence level task, the system should has the sequence training which leads to 10% -15% improvement over cross-entropy (CE) trained DNN[7],[8]. The second order hessian-free (HF) optimization method is critical for performance gains with sequence training [7]. The Rectified Linear Units (Relu) and dropout have been proposed as decent way to regularize large neural networks. It reduces about 5% WER in comparison to cross-entropy trained DNNs [9]. Therefore, ReLU +dropout during hessian-free sequence training strategy are used for CNN.

Joint CNN/DNN models are used to achieve robust signal modeling. These systems give state of the art performance on Broadcast News (13.6% WER on 50 hrs of data, 12.7% WER on 400 hours of training data) and Switchboard (10.7% WER on 300 hours of training data). These results represent a substantial reduction in error rate over the previous best DNN results (2.7% and 2.4% absolute respectively on Broadcast News and 1.5% on Switchboard).

# Speech recognition models of ibm

The IBM group presented improvement to their English Switchboard system that lowered the error rate substantially. They trained 2000 hr audio by joint (Recurrent Neural Network) RNN and CNN architecture with 32000 outputs. They trained larger acoustic model on large data. Furthermore, they used a neural network that have diverse architectures and operate on different input representations, so more accuracy gains from both feature and model of combination are achieved. The maxout nonlinearities and exponential and NN language models are used to reduced the error rate of the system. They proposed a joint model for modeling and training a CNN and DNN which is demonstrated in . The model is trained with 10-15 passes of cross-entropy on 2000 hours audio with 30 iteration of discriminative training using Hessian-free optimization. The DNNs operates on 11 spliced 40 dimensional fMLLR frames and 100-dimentional i-vector. It consists of 5 hidden sigmoid layers (four layers with 2048 units and one layer with 512). The CNN comprise of 2 convolutional layers with 128 and 256 filters respectively. The CNN input is 11 consecutive VTL-warped 40-dimentional logmel frames augmented with first and second derivatives with 9\*9 convolution window. The first layer of the model is specific layer and the remaining layers be shared. The output of specific layers feed the common layers. The output layer of CNN and DNN are combined together in order to speed up the computation. The performance of the joint model is 1.4 % better over 11.8% for CNN on Switchboard 300 hours.



Figure 3: Joint CNN/DNN model proposed by IBM group.

The IBM team proposed the joint RNN/CNN model for sequential data modeling. The first layer is recurrent and is followed by 4 non-recurrent hidden layers (3 layers with 2048 units and one layer with 512 units) and an output layer. Partially unfolded recurrent network operates on sliding window of 6 , 40-dimensional fMLLR frames from t to t+5 and 100 dimensional i-vectors. As the depicts, the joint model and sequence discriminative retraining improve performance about 0.5% and 0.1% on CallHome and Switchboards over score fusion of sequence training (ST) models. shows that the IBM system achieves an error rate of 8% on SWB by choosing n-gram language model and interpolating neural network language model (NNLM) to model M [10].



Table 2: Comparison of word error rates for CE and ST CNN, DNN, RNN and various score fusions on Hub5’00.



Table 3: Comparison of word error rates on Hub5’00 (SWB and CH) for existing systems.

# CONVOLUTIONAL DEEP BELIEF NETWORK

It is generative model that mixes direct and indirect connection between variables. It support efficient bottom to top and top to bottom probabilistic inference. It is composed of convolutional restricted Boltzmann machines (CRBMs). As it is depicted in , it consists of two hidden (H) and visible (V) layers. The visible unit has L channels and each channel has some visible units (Nv \*Nv). The hidden units K groups and each group has a NH\*NH hidden units. Convolutional filter size WS\*Ws for l-th channel correspond to k-th hidden group as and it shares between hidden units in k-th group. In order to have a more scalable CRBM the probabilistic max-pooling is proposed which it subsamples of higher layers. The equation 4 calculate theenergy function of the probabilistic pooling CRBM [11] . Its formula is as follow:

(4)

Subject to



Figure 4: Convolutional restricted Boltzmann machine (CRBM).

The CDBN which is illustrated in is stacked by several probabilistic max-pooling CRBMs. Its training is as deep belief network (DBN) which is accomplished by greedy layer-wise procedure. It means when a layer is trained, their parameters are frozen and its activations are served as the input of next layer.



Figure 5: Convolutional deep belief network [12].

## Application

In recent year deep learning has been successfully applied to image processing and speech recognition, while it is not been applied to the electroencephalographic (EEG) data. One of the significant properties of CDBN against DBN is addressing issue of the scaling model to realistic-size images. The weight is shared between the hidden and visible layers which reduce the large amount of variables and make the representations be invariance to small translations of input. The convolutional deep belief network (CDBN) is the most promising technique for EEG data because of using max pooling operation that reduced the computational burden while allowing full probabilistic inference. This property of CDBN makes it more proper for high-dimensionality and multichannel data. Furthermore, the CDBN is unsupervised learning that allow us to use large amount of unlabeled data [11].

As the EEG feature in frequency-domain is more obvious than time domain, it is chosen as an EEG feature. As number of channel is large to reduce the channel, the PCA is applied. The input data into CDBN is comprise of L (number of PCA components) channels of one dimensional vector. The two layers CDBN with 6\*6 filter and pooling ration 3 on first layer and filter 4\*4 and pooling ration 3 on layer 2 is proposed in this application. As it is depicted , the CDBN has the better performance (about 3%) over other feature extraction methods.



Table 4: Mean classification accuracy.

# Conclusion and future work

In this report, the different DNN architectures of acoustic modeling like deep convolutional neural network, recurrent neural network and joint models, and Convolutional deep belief network for EEG signal are presented. After investigation different deep neural architectures we found that the joint models produce the robust signal model. The joint CNN/DNN and RNN/CNN models are used for robust signal modeling. The joint architecture error rates are 10.7% and 9.4% respectively. The performance of the joint model on 50 hr English Broadcast News (BN), on 400 hr BN and on 300 hr Switchboard with 13.6%, 12.7% and 10.7 error rates respectively are states of are state-of-art. The error rate of CDBN on EEG signal is about 11.75% which is at least 2.55% better that other methods.

Our goal is to find the robust signal modeling for EEG signal to detect seizure automatically. As the EEG signals have their own properties like high dimensionality and comprising multiple channels, the most proper deep neural network should be chosen. Therefore, all the mentioned models will be applied on EEG signals and the performance of them will be compare together, the most proper architecture with robust signal modeling for automatic seizure detection will be chosen.

# Reference

[1] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, “Deep Neural Networks for Acoustic Modeling in Speech Recognition,” *IEEE Signal Process. Mag.*, no. November, pp. 82–97, 2012.

[2] O. Abdel-Hamid, A. R. Mohamed, H. Jiang, and G. Penn, “Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2012, pp. 4277–4280.

[3] Y. LeCun and Y. Bengio, “Convolutional Networks for Images, Speech, and Time-Series,” *Handb. brain theory neural networks*, p. 255, 1995.

[4] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2323, 1998.

[5] Y. LeCun, F. J. Huang, and L. Bottou, “Learning methods for generic object recognition with invariance to pose and lighting,” *Proc. 2004 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognition, 2004. CVPR 2004.*, vol. 2, pp. 97–104, 2004.

[6] T. N. Sainath, B. Kingsbury, G. Saon, H. Soltau, A. rahman Mohamed, G. Dahl, and B. Ramabhadran, “Deep Convolutional Neural Networks for Large-scale Speech Tasks,” *Neural Networks*, vol. 64, pp. 39–48, 2015.

[7] B. Kingsbury, T. Sainath, and H. Soltau, “Scalable minimum Bayes risk training of deep neural network acoustic models using distributed hessian-free optimization,” *Proc. Interspeech*, pp. 1–4, 2012.

[8] T. N. Sainath, B. Kingsbury, A. R. Mohamed, G. E. Dahl, G. Saon, H. Soltau, T. Beran, A. Y. Aravkin, and B. Ramabhadran, “Improvements to deep convolutional neural networks for LVCSR,” in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2013 - Proceedings*, 2013, pp. 315–320.

[9] G. E. Dahl, T. N. Sainath, and G. E. Hinton, “Improving deep neural networks for LVCSR using rectified linear units and dropout,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2013, pp. 8609–8613.

[10] G. Saon, H.-K. J. Kuo, S. Rennie, and M. Picheny, “The IBM 2015 English Conversational Telephone Speech Recognition System,” *Interspeech 2015*, pp. 3–7, 2015.

[11] Y. Ren and Y. Wu, “Convolutional deep belief networks for feature extraction of EEG signal,” *2014 Int. Jt. Conf. Neural Networks*, pp. 2850–2853, 2014.

[12] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, “Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations,” *Proc. 26th Annu. Int. Conf. Mach. Learn. ICML 09*, vol. 2008, pp. 1–8, 2009.