A Survey of Deep Learning Techniques in Speech Recognition

Akshi Kumar, Sukriti Verma and Himanshu Mangla
Department of Computer Science and Engineering
Delhi Technological University
Delhi, India
akshikumar@dce.edu, sukriti_bt2k14@dce.ac.in, himanshumangla_2k14@dtu.ac.in

***Abstract*—The last decade has seen a rapid progress of deep learning techniques in various fields since the proposal of an efficient algorithm for training deep belief networks. Amongst these, speech recognition has seen significant breakthroughs with the application and integration of deep learning architectures. Deep Belief Networks have successfully outperformed GMM based acoustic models within the GMM-HMM hybrid that had dominated the field of speech recognition until now. Recent research with Convolutional Neural Networks and Recurrent Neural Networks for acoustic modeling has shown promising results. Both of these architectures are now being used to solve the problem end-to-end. With this article, a survey is provided on the application of three deep learning architectures in the field of speech recognition, namely, Deep Belief Networks, Convolutional Neural Networks and Recurrent Neural Networks.**

Keywords—Automatic Speech Recognition; Deep Learning Techniques; Deep Belief Networks; Convolutional Neural Networks; Recurrent Neural Networks

INTRODUCTION

Speech is the most natural and effective method of communication between human beings. Speech Recognition aims to transcribe speech to text. It is a standard classification problem where speech signals have to be mapped to or identified as words. However, it is not possible to work with speech documents if they are recorded as audio signals. Therefore, speech recognition has become a crucial area of research [1], [2].

There are many obstacles which make real-world speech recognition a challenging problem. Accents, speaking styles, different possible pronunciations, various languages, noise – are some of these obstacles. There's a substantial loss in accuracy when we move from a controlled experimental setups to real life situations. Despite this, automatic speech recognition has abundant usage in dictation, human-machine interfaces and control of machines among others.

Hidden Markov Models (HMMs) with an acoustic model based on Gaussian mixtures [3], also known as GMM-HMMs, have dominated the field of speech recognition for many years [4], [5], [6], [7], [8]. These speech recognition systems employ Hidden Markov Models to handle temporal variability and sequential structure of speech data and Gaussian mixture models (GMMs) to provide localized classifications. Over the years, GMMs have been shown to be statistically inefficient for modeling nonlinear relationships, such as the relationship between the acoustic features and human speech inputs. HMMs are also sensitive to mismatch between the training and testing data, particularly the mismatch introduced by environmental noise. Much effort has to be spent improving the robustness of the system to such distortions and ensuring that the system performs well often requires a large amount of training data. Moreover, traditional systems use heavily engineered processing stages, including noise removal, specialized input features and speech enhancement [9]. Traditional systems do not work on raw speech but on certain spectral features. The raw waveforms have to be processed to compute spectral features. Frequently used speech spectral features are Mel-frequency cepstral coefficients (MFCCs) and perceptual linear predictive coefficients (PLPs) [10], [11].

This is why, in the last decade, there has been an extensive application of neural networks and deep learning to perform speech recognition leading to significant novel results. The trend began two decades ago, when sophisticated results were achieved using hybrid ANN-HMM models. These models appropriated the use of Artificial Neural Networks (ANNs) with only one layer of hidden units having non-linear activation functions to predict probabilities over HMM states from short windows of acoustic coefficients [12]. ANNs are powerful models that can represent complex non-linear functions but at that time, neither the computation power nor the training algorithms that were available, were advanced enough for training ANNs with many hidden layers. So, hybrid ANN-HMM models could not replace the very successful combination of HMMs with acoustic models based on Gaussian mixtures.

In the last decade, this has been achieved. Through the application of deep learning, researchers have shown that Deep Neural Networks (DNNs) achieve better performance than GMMs for acoustic modeling in speech recognition systems, sometimes by a large margin [13]. The major breakthrough in deep learning was triggered by Hinton et al. [14] with the proposal of a novel deep learning architecture called a deep belief network (DBN) and the use of a generative, layer-by-layer pre-training method for initializing the weights and getting them to correct scales before the training procedure began. In addition to promising learning procedures, the main factors that have contributed to the

ISBN:978-1-5386-4119-4/18/$31.00 ©2018IEEE

recent success of deep learning is the rapid advancement of computing power, allowing researchers to make the neural networks deeper and more powerful, and availability of more training data. Acoustic models based on DNNs have now started replacing GMMs in traditional speech recognition systems [13], [15], [16].

Deep learning has emerged as a powerful promising technique in fields ranging from computer vision to natural language processing and even recommendation systems [16]. Deep neural networks (DNNs) hold the capacity to represent functions with higher complexity. Moreover, they can also work with raw data and learn rich representations as opposed to specialized input features. The purpose of this paper is to provide a survey as well as a brief tutorial on the application of deep learning in the field of speech recognition. The rest of the paper is organized as follows. In the first section, we provide a brief tutorial on Deep Belief Networks (DBNs) and discuss their application and integration in existing GMM-HMM speech recognition systems. In the second section, we move onto Convolutional Neural Networks (CNNs). In the third section, we discuss Recurrent Neural Networks (RNNs). In the last section, we bring attention to some very recent approaches that are experimenting with combinations of these deep architectures and trying to eliminate processing stages, solving the problem end-to-end.

I. DEEP BELIEF NETWORK

A. Restricted Boltzmann Machine

A restricted Boltzmann machine (RBM) is a particular type of neural network that has one layer of hidden units and one layer of visible units. As the name implies, RBMs are a special case of Boltzmann machines, having the restriction that their neurons must form a bipartite graph. Bipartite graph means that all visible units are connected to all hidden units, but there are no connections between units of the same layer i.e. there are no visible-visible or hidden-hidden connections. This restriction allows for efficient training algorithms compared to those that are available for the general class of Boltzmann machines. One such learning algorithm is the gradient-based contrastive divergence algorithm [17]. This algorithm and its variations are most often used to train an RBM.

RBMs can learn and model the structure in the input data. They can also be used for dimensionality reduction. In classification tasks, RBMs are used for feature learning. However, the features extracted by the RBMs via unsupervised learning may not be useful in the supervised learning tasks. Because they only have a single hidden layer, they are not as powerful.

B. Deep Belief Network

To explore the dependencies between neuron activations in the hidden and the visible layers in an RBM, Hinton et al. [14] stacked multiple RBMs together forming a multilayer, generative model called a DBN and marked the birth of deep learning. Every two adjacent layers of a DBN form an RBM. DBNs are more effective with a stronger learning capacity compared to an RBM, especially when applied to problems with unlabeled data. They also handle the problems of overfitting and underfitting [18], [16].

The training process of a DBN proposed by Hinton et al. [14] two stages: the pre-training stage and the fine-tuning stage. The RBMs within the DBN are pre-trained sequentially using a greedy layer-by-layer unsupervised learning algorithm. The visible layer of the lowest RBM is initialized with the input. The values in the visible layer are then transferred to the adjacent hidden layer where the activations of the hidden units are calculated. The representation obtained by this RBM is used as the training data for the next RBM and this training process continues until all the layers are traversed. With this method, one layer of mapped features is learnt at a time using the states of the features mapped by the previous layer as the training data. The pre-training stage serves as feature learning, after which, the weights that have been learned from the distribution of the input data serve as a much better starting point for the fine-tuning stage as opposed to random initialization as they have been set to the correct scales [19]. The fine-tuning stage is where the discriminative learning takes place. A final softmax layer, having one unit to represent each of the HMM states, is added to the DBN and a supervised algorithm with a cost function is used to train the whole DNN to predict the correct HMM state and the weights are adjusted further. One such algorithm is backpropagation that optimizes cross entropy between the target state and the state predicted from the DBN.

It is well noted that initializing the weights sensibly using the pre-training method strongly improves the performance of the DBN, allowing the fine-tuning stage to progress rapidly while also significantly reducing overfitting [20]. Compared with randomly initialized weights, these weights are closer to the global optima. In literature, these models are referred to by varied terms: DBN-HMM hybrid, DNN-HMM hybrid or fully-connected DNNs.

C. Acoustic Modeling using DBNs

Mohamed et al. [21] carried out the first successful experiment and demonstrated that a DNN-HMM hybrid model using an acoustic model based on a DBN in place of Gaussian mixtures outperformed state-of-the-art results on TIMIT dataset. In the next couple of years, various researchers employed the use of RBMs and DBNs to investigate the effects of these architectures and demonstrated the same [22], [23], [24], [25], [26].

In these experiments, DBNs are trained to classify individual frames of acoustic input. Therefore, DBNs are trained to optimize the cross entropy between the predicted HMM state and the correct HMM state for each frame. The correct prediction for each frame is decided upon by the HMM which uses a forced alignment between the speech data and transcription [12]. However, speech recognition is a sequential problem and this forced alignment is error-prone. Towards this goal, experiments were done by Mohamed et al. [27] using log conditional likelihood as the criteria to train a DBN at the sequence level. This technique also considered predictions scores from a bigram language model. With careful tuning of hyper parameters, it outperformed the frame-level training

using cross-entropy by almost 5% on the TIMIT phone recognition task. We discuss sequential approaches in more detail in the fourth section.

The initial successes of DBNs lead to research and experimentation with large vocabularies such as the Bing Voice Search task, English Broadcast News and Switchboard Dataset. Large vocabulary speech recognition tasks are tougher with a higher number of classes available for classification and less controlled conditions. Remarkable results on large vocabulary speech recognition tasks by using acoustic models based on DBNs were achieved subsequently [28], [29], [30], [31], [32], [33],[34]. The move from small to large vocabulary tasks was done by using context-dependent triphone HMMs having many thousands of tied states. A comprehensive review of these techniques was carried out by Hinton et al. [13].

Even though DBNs gave impressive results for large vocabulary speech recognition tasks, context-dependent HMMs were computationally demanding to train and suffered from the impending problem of scalability as vocabularies grew larger. To this end, Deng and Yu [35] proposed a deep architecture, referred to as deep convex networks (DCNs). An alternative method of pre-training was found to be effective for this architecture. It was called discriminative pre-training [13]. In this method, pre-training is accomplished directly by convex optimization. It begins with a neural network having only a single hidden layer and a softmax layer. This network is trained discriminatively. Then, a second hidden layer is added between the first hidden layer and the softmax output layer and the whole network is again discriminatively trained. This is done until the desired number of hidden layers is reached, after which fine-tuning is done.

The next phase of advancement in automatic speech recognition began with the investigation of convolutional neural networks (CNNs) and the introduction of convolutional RBMs where experiments were carried out by making the RBMs convolutional in time and frequency.

II. CONVOLUTIONAL NEURAL NETWORK

A. Introduction

A convolutional neural network (CNN) is a multi-layer neural network that consists of two different types of layers that alternate: the convolution layer and the pooling layer. The architecture of CNNs has been inspired from the structure of the animal visual cortex.

The convolution layer is used to extract features. It consists of a number of feature maps made up of neurons. Each neuron in the convolution layer processes data only for its receptive field which are features of a limited range and not the whole input. This receptive field is also called a filter and it strides over the input. Neurons in one feature map have the same weights connecting them to their inputs but receive different inputs. This concept is called weight sharing or parameter sharing. Weight sharing significantly reduces the number of different parameters to be learned while also granting the CNN with equivariance, so that whenever the input changes, the corresponding output also changes. The essence is that each neuron in a feature map extracts the same feature from the input, agnostic to the input region being considered. This also reduces the amount of memory needed while training a CNN. Hence, weight sharing greatly improves thelearning efficiency of a CNN [36]. The pooling layer comes after the convolution layer. It has the same number of feature maps as its preceding convolution layer. The difference is that the dimensions of the feature maps are smaller and hence, have a smaller number of neurons per feature map. This purpose of the pooling layer is to compute a lower resolution representation of the features that have been learned by the convolution layer through sub-sampling. One very common pooling function is the max pooling function, where each neuron simply computes the maximum value of the feature for its receptive field.

The convolution layer-pooling layer pairs are stacked up to obtain higher level features. On top of these layers, there is a standard fully connected layer, representing HMM states, that combines the effects of the features and is used for discriminative training of the network.

B. Acoustic Modeling using CNNss

After the success of DBN-HMM hybrid models for speech recognition, work was carried out using CNN based acoustic models. CNN for speech data with convolution along the time axis was first proposed by LeCun et al. [37], but no validation was carried out at the time. It was theorized that convolution along time will help obtain features robust to small temporal shifts.

This was confirmed by Lee et al. [38] and, Hau and Chen [39]. In these works, convolution was applied over windows of acoustic frames that overlap in time. This resulted in learning acoustic features that were relatively more stable with respect to variations arising from speakers and genders.

Abdel-Hamid et al. [40] achieved significant improvements by applying convolution and max-pooling along frequency axis rather than the time axis. Convolution along frequency axis was found to generate features robust to small frequency shifts, which often happens because of different speakers and even different moods. More researchers explored convolution over both time and frequency axes simultaneously [41], [42].

Results by Abdel-Hamid et al. [41] indicated that applying convolution along the time axis, while outperforming the DBN, gives significantly worse results than applying convolution along the frequency axis. Hence, further work by Abdel-Hamid et al. [43] went back to applying convolution only over frequency axis stating that HMMs do relatively well at handling temporal variability. This work also discusses the effects of using different speech spectral features as input and variations in CNN hyper parameters.

Above experiments demonstrated CNNs to outperform the fully connected DBN within the hybrid DNN-HMM model. This was because of the following two reasons. First, DBNs interpret the input in any order but speech spectral features are strongly correlated in frequency and time. Weight sharing allowed CNNs to capture these local correlations. And second,

weight sharing and pooling helps CNNs capture equivariance and imparts robustness and stability. For DBNs to capture this sort of invariance over small frequency and/or temporal shifts, a very high number of hyper parameters are required.

Experiments by Sainath et al. [44], [45], [46] showed that CNNs can achieve better performance than DBNs for large vocabulary tasks. These experiments involved careful parameter tuning, limited weight sharing and sequential training as opposed to frame-based. An empirical study on CNN based acoustic models for low resource languages performed by Chan and Lane [42] concluded that CNNs improve performance over DBNs in the low resource condition by providing robustness and better generalization.

C. Limited Weight Sharing and Pre-Training

A couple of changes from conventional CNNs that were widely explored and demonstrated improved performance were the concepts of limited weight sharing and RBM like pre-training [41], [43].

Instead of full weight sharing between neurons belonging to the same feature map, weight sharing is limited. Only those convolution layer units that are attached to the same pooling layer units share weights. This significantly increases the number of parameters to be learned but allows these convolution neurons to compute comparable features.

A convolutional RBM was proposed by [47]. The weights of a trained convolutional RBM serve as good initial values for training the convolution layer. The values learned by the convolution layer are then sub-sampled by the pooling layer. The outputs of the pooling layer are used as inputs to pre-train the next layer as discussed in case of DBNs. Lee et al. [38] demonstrated the convolutional RBM network to learn speech spectral features without supervision with convolution along time, finding it promising to work with complex, high-dimensional data.

Abdel-Hamid et al. [41], [43] investigated pre-training a convolutional RBM using limited weight sharing. It was found that this sort of pre-training improved performance only on large vocabulary speech recognition task. Improvements were obtained on the Bing Voice Search task which is a large vocabulary task but no improvements were obtained on the TIMIT phone recognition task.

III. RECURRENT NEURAL NETWORK

A. Introduction

We have discussed that acoustic modeling using deep feedforward networks such as DBNs and CNNs have led to dramatic improvements in the field of speech recognition in recent years. Unlike feedforward neural networks, recurrent neural networks (RNNs) are allowed to have connections that feed activations from units in a particular layer as input to units in the same or preceding layers. To make a decision for the current input, RNNs consider previous decisions, making them inherently deep in time. Hence, RNNs have indefinite temporal context compared to fixed context windows as are used in feedforward networks and it is only natural that they have been used for processing sequential data such as speech.

Training RNNs using backpropagation technique suffers from the vanishing gradient and exploding gradient problems [49]. Moreover, these problems limit the range for which RNNs can retain context. To address these problems, the Long Short-Term Memory (LSTM) architecture was proposed by Hochreiter and Schmidhuber [49]. LSTMs contain purpose built memory cells in the recurrent hidden layer to store information and are better at finding and exploiting long range context. Each memory cell contains an input gate that controls the flow of activations into the memory cell, an output gate that controls the flow of activations into rest of the network and a forget gate to allow resetting the memory cell.

B. Bidirectional RNN and LSTM

RNNs only retain and use previous context. To impart RNNs the ability to exploit future context, Bidirectional RNNs (BRNNs) were proposed Schuster and Paliwal [50]. BRNNs process data in both directions using two separate hidden layers in place of one. These two hidden layers feed forward to the same output layer. Bidirectional RNNs with LSTM cells in the recurrent hidden layers gives bidirectional LSTM network, which can access long-range context in both directions.

C. Acoustic Modeling using RNNs

Since RNNs can learn how much context they have to refer to, researchers have naturally experimented with HMM-RNN hybrid models in the past [51], [52]. But it has been difficult to bring the performance of RNN based acoustic models up to par with acoustic models based on DBNs and CNNs.

In recent years, much effort has been made in this direction. A very important work that has laid the groundwork for using RNNs to model sequential data, such as speech, was carried out by Graves et al. in 2006 [53]. This work introduced the Connectionist Temporal Classification (CTC) loss function, which allowed neural networks to learn alignments between a sequence of characters and unsegmented speech spectral features thereby obviating the need to use force alignments learned by Hidden Markov Models. CTC was shown to outperform HMM-RNN hybrids on TIMIT dataset in this work. Continued work by Graves et al. [54] demonstrated use of deep bidirectional LSTM for end-to-end speech recognition using CTC network combined with a language model and achieved state-of-the-art results on TIMIT dataset but found that it was difficult to integrate this with existing systems for larger vocabularies. Hence, Graves et al. [55] went back to exploring RNNs in combinations with HMMs in a hybrid setup. They used a deep bidirectional LSTM (DBLSTM) as an acoustic model within the standard DNN-HMM hybrid and obtained state-of-the-art results on the TIMIT dataset and a very small improvement over best published results on the Wall Street Journal speech corpus, concluding that the DBLSTM-HMM hybrid will need further investigation.

In recent years, RNN-HMM hybrid systems with deep bidirectional LSTM based acoustic models have improved significantly with the use of context-dependent phonetic units, context-dependent states for the LSTM output space and

distributed training methods to carry out large scale modeling [56], [57], [58]. In the next section we discuss end-to-end speech recognition which has been another crucial point of focus in recent years.

IV. END-TO-END SPEECH RECOGNITION

With DNN-HMM hybrids, DNNs are trained to predict HMM states for each frame. The frame-level targets are decided upon by the HMM by a forced alignment between the speech data and the transcription. Hence, the frame-wise cost function that is used to train the DNN does not capture the primary objective function which is obtaining the most accurate transcription possible. The repercussion is that sometimes an improvement in frame-level accuracy means no improvement or even a dip in the transcription accuracy. This is the inconsistency that end-to-end speech recognition seeks to avoid. Moreover, a large number of tuning parameters are needed with the frame-wise approach.

With end-to-end speech recognition the HMM is replaced by a neural network that learns alignments between a sequence of characters and unsegmented speech. One such model we have already mentioned is CTC proposed by Graves et al. [54]. End-to-end speech recognition aims to eliminate as much of the processing pipeline as possible and replace it with a single unified neural network.

RNNs have become a default method for end-to-end speech recognition. Foundational work by Graves et al. [56] combined a CTC network with a separate RNN that accounts for previous phoneme predictions while making the current prediction, thereby combining an acoustic model that is the CTC network and a language model that is the RNN transducer. Following this, much work has been done using RNN with CTC loss function to achieve end-to-end speech recognition. Hannun et al. [59] employed a 5-layer RNN with a bidirectional recurrent layer trained with CTC loss along with a language model to rectify phonetically plausible transcriptions. This technique outperformed the best results on the Switchboard dataset. They have also discussed many optimizations as their training system used multiple GPUs. Work by Amodei et al. [60] achieved a significant improvement over the 5-layer RNN proposed by Hannun et al. [59] by using a similar network but having 13 hidden layers and by applying convolution in some of the layers.

Other successful approaches for end-to-end speech recognition have been using attention based RNNs that are trained sequence-to-sequence, that is, to generate one sequence that is the transcribed text, given another sequence, that is speech input [61], [62], [63]. Hence, these are also called sequence-to-sequence models. Zhang et al. [64] experimented with very deep CNNs and convolutional LSTMs to capture complex non-linearities and showed an improvement over shallow sequence-to-sequence models. Further work by Zhang et al. (2017b) [65] went on to experiment with combining CNNs with CTC loss function instead of using the default RNN with CTC loss function combination stating that RNNs suffer from various difficulties in training. They achieved promising results on the small vocabulary task of the TIMIT dataset claiming their work to be a more efficient training technique.

V. CONCLUSION

The area of deep learning has seen rapid progress and lead to significant improvements in various fields. With this article, we have provided a brief tutorial and overview of deep learning techniques and architectures in the field of speech recognition. We have discussed acoustic models based on Deep Belief Networks, Convolutional Neural Networks and Recurrent Neural Networks. In recent years, acoustic models based on DBNs and CNNs have successfully replaced Gaussian mixtures and have been demonstrated to work quite well for large vocabulary tasks. Moreover, there has been the idea of eliminating processing stages, using one unified neural network to achieve end-to-end speech recognition. To this end, RNNs are now being experimented with but require much computation power for training. The usage of RNNs for acoustic modeling within a hybrid DNN-HMM system as compared to the usage of RNNs for end-to-end speech recognition using CTC loss function and a language model has had mixed reactions. Deep learning holds the power to work with raw inputs and learn rich representations while eliminating laborious processing stages. With rapid advancement of computational technologies, deep learning will only grow in the future.

REFERENCES

[1] Trentin, E., and Gori, M. (2001). A survey of hybrid ANN/HMM models for automatic speech recognition. Neurocomputing, 37(1-4), 91-126. doi: 10.1016/S0925-2312(00)00308-8

[2] Deng, L., and Li, X. (2013). Machine learning paradigms for speech recognition: An overview. IEEE Transactions on Audio, Speech, and Language Processing, 21(5), 1060-1089.

[3] Juang, B.-H., Levinson, S., and Sondhi, M. (1986). Maximum likelihood estimation for multi-variate mixture observations of markov chains (corresp.). IEEE Transactions on Information Theory, 32(2):307–309.

[4] Bahl, L., Brown, P., Souza, P. D., and Mercer, R. (1986). Maximum mutual information estimation of hidden markov model parameters for speech recognition. In Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'86., volume 11, pages 49–52.

[5] Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2):257–286.

[6] Deng, L., Lennig, M., Seitz, F., and Mermelstein, P. (1990). Large vocabulary word recognition using context-dependent allophonic hidden markov models. Computer Speech and Language, 4(4):345–357.

[7] Deng, L., Kenny, P., Lennig, M., Gupta, V., Seitz, F., and Mermelstein, P. (1991). Phonemic hidden markov models with continuous mixture output densities for large vocabulary word recognition. IEEE Transactions on Signal Processing, 39(7):1677–1681.

[8] Baker, J. M., Deng, L., Glass, J., Khudanpur, S., Lee, C.-H., Morgan, N., and O'Shaughnessy, D. (2009). Developments and directions in speech recognition and understanding, part 1 [dsp education]. IEEE Signal Processing Magazine, 26(3).

[9] Singh, S., Tripathy, M., and Anand, R. S. (2014). Subjective and objective analysis of speech enhancement algorithms for single channel speech patterns of indian and english languages. IETE Technical Review, 31(1):34–46.

[10] Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. IEEE Transactions on Acoustics, Speech, and Signal Processing, 29(2):254–272.

[11] Hermansky, H. (1990). Perceptual linear predictive (plp) analysis of speech. the Journal of the Acoustical Society of America, 87(4):1738–1752.

[12] Bourlard, H. A. and Morgan, N. (1993). Connectionist speech recognition: A hybrid approach.

[13] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine, 29(6):82–97.

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

[15] Deng, L., Li, J., Huang, J.-T., Yao, K., Yu, D., Seide, F., Seltzer, M., Zweig, G., He, X., Williams, J., et al. (2013). Recent advances in deep learning for speech research at microsoft. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pages 8604–8608.

[16] Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., and Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. Neurocomputing, 234:11–26.

[17] Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. Neural computation, 14(8):1771–1800.

[18] Arnold, L., Rebecchi, S., Chevallier, S., and Paugam-Moisy, H. (2011). An introduction to deep learning. In European Symposium on Artificial Neural Networks (ESANN).

[19] Hinton, G.E. and Salakhutdinov, R.R.(2006). Reducing the dimensionality of data with neural networks. science, 313(5786):504–507.

[20] Larochelle, H., Erhan, D., Courville, A., Bergstra, J., and Bengio, Y. (2007, June). An empirical evaluation of deep architectures on problems with many factors of variation. In Proceedings of the 24th international conference on Machine learning (pp. 473-480). ACM.

[21] Mohamed, A. R., Dahl, G., and Hinton, G. (2009). Deep belief networks for phone recognition. In Nips workshop on deep learning for speech recognition and related applications (volume 1, No. 9, page 39).

[22] Mohamed, A. R. and Hinton, G. (2010). Phone recognition using restricted Boltzmann machines. In Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, pages 4354–4357.

[23] Mohamed, A. R., Sainath, T. N., Dahl, G., Ramabhadran, B., Hinton, G. E., and Picheny, M. A. (2011). Deep belief networks using discriminative features for phone recognition. In Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, pages 5060–5063.

[24] Morgan, N. (2012). Deep and wide: Multiple layers in automatic speech recognition. IEEE Transactions on Audio, Speech, and Language Processing, 20(1):7–13.

[25] Mohamed, A. R., Dahl, G. E., and Hinton, G. (2012a). Acoustic modeling using deep belief networks. IEEE Transactions on Audio, Speech, and Language Processing, 20(1):14–22.

[26] Mohamed, A. R., Hinton, G., and Penn, G. (2012b). Understanding how deep belief networks perform acoustic modelling. In Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, pages 4273–4276.

[27] Mohamed, A. R., Yu, D., and Deng, L. (2010). Investigation of full-sequence training of deep belief networks for speech recognition. In Eleventh Annual Conference of the International Speech Communication Association.

[28] Yu, D., Deng, L., and Dahl, G. (2010). Roles of pre-training and fine-tuning in context-dependent dbn-hmms for real-world speech recognition. In Proc. NIPS Workshop on Deep Learning and Unsupervised Feature Learning.

[29] Sainath, T. N., Kingsbury, B., Ramabhadran, B., Fousek, P., Novak, P., and Mohamed, A. R. (2011). Making deep belief networks effective for large vocabulary continuous speech recognition. In Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on, pages 30–35.

[30] Seide, F., Li, G., Chen, X., and Yu, D. (2011a). Feature engineering in context-dependent deep neural networks for conversational speech transcription. In Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on, pages 24–29.

[31] Seide, F., Li, G., and Yu, D. (2011b). Conversational speech transcription using context-dependent deep neural networks. In Twelfth Annual Conference of the International Speech Communication Association.

[32] Dahl, G. E., Yu, D., Deng, L., and Acero, A. (2011). Large vocabulary continuous speech recognition with context-dependent dbn-hmms. In Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, pages 4688–4691.

[33] Dahl, G. E., Yu, D., Deng, L., and Acero, A. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Transactions on audio, speech, and language processing, 20(1):30–42.

[34] Pan, J., Liu, C., Wang, Z., Hu, Y., and Jiang, H. (2012). Investigation of deep neural networks (dnn) for large vocabulary continuous speech recognition: Why dnn surpasses gmms in acoustic mod- eling. In Chinese Spoken Language Processing (ISCSLP), 2012 8th International Symposium on, pages 301–305.

[35] Deng, L. and Yu, D. (2011). Deep convex net: A scalable architecture for speech pattern classification. In Twelfth Annual Conference of the International Speech Communication Association.

[36] Ian, G., Yoshua, B., and Aaron, C. (2016). Deep Learning. MIT Press.

[37] LeCun, Y., Bengio, Y., et al. (1995). Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks, 3361(10):1995.

[38] \bibitem{c39}Lee, H., Pham, P., Largman, Y., and Ng, A. Y. (2009b). Unsupervised feature learning for audio classification using convolutional deep belief networks. In Advances in neural information processing systems, pages 1096–1104.

[39] Hau, D. and Chen, K. (2011). Exploring hierarchical speech representations with a deep con- volutional neural network. In 11th UK workshop on computational intelligence (UKCI '11), page 37.

[40] bdel-Hamid, O., Mohamed, A. R., Jiang, H., and Penn, G. (2012). Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition. In Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, pages 4277–4280.

[41] Abdel-Hamid, O., Deng, L., and Yu, D. (2013). Exploring convolutional neural network structures and optimization techniques for speech recognition. In Interspeech, volume 2013, pages 1173–5.

[42] Chan, W. and Lane, I. (2015). Deep convolutional neural networks for acoustic modeling in low resource languages. In Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on, pages 2056–2060.

[43] Abdel-Hamid, O., Mohamed, A. R., Jiang, H., Deng, L., Penn, G., and Yu, D. (2014). Convolutional neural networks for speech recognition. IEEE/ACM Transactions on audio, speech, and language processing, 22(10):1533–1545.

[44] Sainath, T. N., Kingsbury, B., Mohamed, A. R., Dahl, G. E., Saon, G., Soltau, H., Beran, T., Aravkin, A. Y., and Ramabhadran, B. (2013a). Improvements to deep convolutional neural networks for lvcsr. In Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on, pages 315–320.

[45] Sainath, T. N., Mohamed, A. R., Kingsbury, B., and Ramabhadran, B. (2013b). Deep convolutional neural networks for lvcsr. In Acoustics, speech and signal processing (ICASSP), 2013 IEEE international conference on, pages 8614–8618.

[46] Sainath, T. N., Kingsbury, B., Saon, G., Soltau, H., Mohamed, A. R., Dahl, G., and Ram- abhadran, B. (2015). Deep convolutional neural networks for large-scale speech tasks. Neural Networks, 64:39–48.

[47] Lee, H., Grosse, R., Ranganath, R., and Ng, A. Y. (2009a). Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In Proceedings of the 26th annual international conference on machine learning, pages 609–616.

[48] Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. IEEE transactions on neural networks, 5(2), 157-166.

[49] Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[50] \bibitem{c51}Schuster, M., and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45(11), 2673-2681. doi: 10.1109/78.650093

[51] \bibitem{c52}Robinson, A. J. (1994). An application of recurrent nets to phone probability estimation. IEEE transactions on Neural Networks, 5(2), 298-305.

[52] Vinyals, O., Ravuri, S. V., and Povey, D. (2012, March). Revisiting recurrent neural networks for robust ASR. In Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on (pp. 4085-4088). IEEE.

[53] Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. (2006, June). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd international conference on Machine learning (pp. 369-376). ACM.

[54] Graves, A., Mohamed, A. R., and Hinton, G. (2013a, May). Speech recognition with deep recurrent neural networks. In Acoustics, speech and signal processing (icassp), 2013 ieee international conference on (pp. 6645-6649). IEEE.

[55] Graves, A., Jaitly, N., and Mohamed, A. R. (2013b, December). Hybrid speech recognition with deep bidirectional LSTM. In Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on (pp. 273-278). IEEE.

[56] Sak, H., Senior, A., and Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In Fifteenth annual conference of the international speech communication association.

[57] Geiger, J. T., Zhang, Z., Weninger, F., Schuller, B., and Rigoll, G. (2014). Robust speech recognition using long short-term memory recurrent neural networks for hybrid acoustic modelling. In Fifteenth annual conference of the international speech communication association.

[58] Sak, H., Senior, A., Rao, K., and Beaufays, F. (2015). Fast and accurate recurrent neural network acoustic models for speech recognition. arXiv preprint arXiv:1507.06947.

[59] Hannun, Awni, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger et al. "Deep speech: Scaling up end-to-end speech recognition." arXiv preprint arXiv:1412.5567 (2014).

[60] Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., and Chen, J. (2016, June). Deep speech 2: End-to-end speech recognition in english and mandarin. In International Conference on Machine Learning (pp. 173-182).

[61] Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., and Bengio, Y. (2015). Attention-based models for speech recognition. In Advances in neural information processing systems (pp. 577-585).

[62] Chan, W., Jaitly, N., Le, Q., and Vinyals, O. (2016, March). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on (pp. 4960-4964). IEEE.

[63] Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P., and Bengio, Y. (2016, March). End-to-end attention-based large vocabulary speech recognition. In Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on (pp. 4945-4949). IEEE.

[64] Zhang, Y., Chan, W., and Jaitly, N. (2017a, March). Very deep convolutional networks for end-to-end speech recognition. In Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on (pp. 4845-4849). IEEE.

[65] Zhang, Y., Pezeshki, M., Brakel, P., Zhang, S., Bengio, C. L. Y., and Courville, A. (2017b). Towards end-to-end speech recognition with deep convolutional neural networks. arXiv preprint arXiv:1701.02720. doi: 10.21437/Interspeech.2016-1446