ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

Deep Architectures for Speech Processing: Survey

Renu Karule, M. A. Potey

Student, Dept. of Computer Engineering, DYPCOE Akurdi, Pune,India

HOD, Dept. of Computer Engineering. DYPCOE Akurdi, Pune, India

ABSTRACT: The speech signal is the fastest and regular way of communication between humans. This fact has motivated researchers to think of speech as a fast and natural method of interaction between human and machine. However, for using most natural form of communication require that the machine should have the sufficient intelligence to recognize human voices. Speech processing research has been started two decades before and led to the improved level of different tasks like speech synthesis, speech recognition, speaker verification and speech separation. Recently, speech community is focusing on the use of deep architectures for the different speech processing task. Capability of deep architectures of representing acoustic features and training on large amount of data has motivated many researcher towards use of deep learning in speech processing. This report has studied the different deep architectures like deep neural network, deep convolution neural network, deep belief network, deep recurrent neural network and deep restricted Boltzmann machines and their use in speech processing.

KEYWORDS: Deep architecture; Neural Network; Speech processing; Speech recognition; Generative

I. INTRODUCTION

Speech processing is study of signals and methodology of how to deal with these signals. For speech processing signals are mostly represented in digital format. Speech processing gives practical and theoretical insights about how the human speech can be processed by the computers/machines. It involves speech recognition, speech synthesis, speech enhancement and spoken dialog system. There are some characteristics of speech signals like phonemes, prosody, IPA notation. Speech signals can contain message, speaker specific characteristics, emotions, language context. Basic parameters in speech processing involves pitch, SN ratio, voice intensity and quality. In the area of speech processing basic parameters are needed which play a vital role in good performance of the systems. Many algorithms have been developed to estimate those basic parameters from the speech signal, but their performance has not been explored sufficiently. Recently, to deal with effective representation of information from speech signals deep learning is state-of-art in speechprocessing area. It is machine learning methodology. Deep learning has turned out to be successful in tackling with many AI problems including speech information processing.

As research in AI is progressing in all areas, there is a need of model, which is capable of processing the complex input data and solving different kinds of complicated tasks. Deep architectures have been proved such kind of model. It is believed that deep architectures have good learning algorithms and excellent performance. Motivation behind the study of deep architectures for speech processing is the power of deep architectures for representation of features.There are several deep architectures, but CNNs and DBNs are the two milestones in the field of speech processing.

II. HISTORY OF DEEP ARCHITECTURE

Before use of deep learning, generative model has been used for speech processing tasks. Deep learning moved on the basic concept of hierarchical representation of features. In 1980s back-propagation algorithm has been popularized. In 90's segmentation and stationary models have been developed. In 2003-06, Structured Hidden Trajectory Models have been developed. Research in 2006 a class of deep generative models have been developed and in same year there are some breakthrough work happened in deep learning area like energy models and algorithms for deep belief network. In 2007-2009, some researcher studied various deep generative models. In 2010-2012 new work in which deep neural network with pre-trained followed by back-propagation fine-tuning. Gradually the research in deep learning moved

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

towards new deep architectures like deep convex net, hybrid architectures and discriminative architectures. Results from these architectures are varied depend on task for which they have been used.

III. DIFFERENT DEEP ARCHITECTURES

This paper is focusing on following deep architectures:

A) Convolutional Neural Network(CNN)
B) Recurrent Neural Network (RNN)
C) Restricted Boltzmann Machines(RBM)
D) Deep Belief Network(DBN)
E) Deep Convex Nets(DCN)
F) Deep Neural Networks(DNN)

Above deep architectures fall in category of generative, discriminative and hybrid deep architectures [12].

A) Convolutional Neural Network

Convolution neural network is discriminative type of architecture with each module consist of convolutional layer and pooling layer(Hamid et al.). Deep CNN model is formed by stacking up these modules, one on top of other or with DNN on top of it. CNN introduces a special network structure which consist of alternating convolution and pooling layers. The convolution layer shares many weights, and the output of convolution layer is subsampled by the pooling layer which reduces the data rate from the layer below. For CNN, the input data need to be organized as a number of feature maps which will feed into CNN. In case of speech processing, how to organize the speech feature vectors is important for processing of CNN. The input can be thought as spectrogram. In CNN one's the input feature maps are formed, the convolution and the pooling layer apply their respective operation to generate activations of the unit in that layer. Layer can also refer as ply in CNN. After several convolutions and pooling layers, finally fully connected layers are formed which results in high-level reasoning of neural network.

Below Figure shows the architecture of CNN. As shown in above figure, CNN architecture consist of three layers convolution, pooling and fully connected. In convolution layer, each neuron takes input from the previous layer, neurons sharing the same weight form feature maps. multiplying the local input with weight matrix W or localized filter [15].

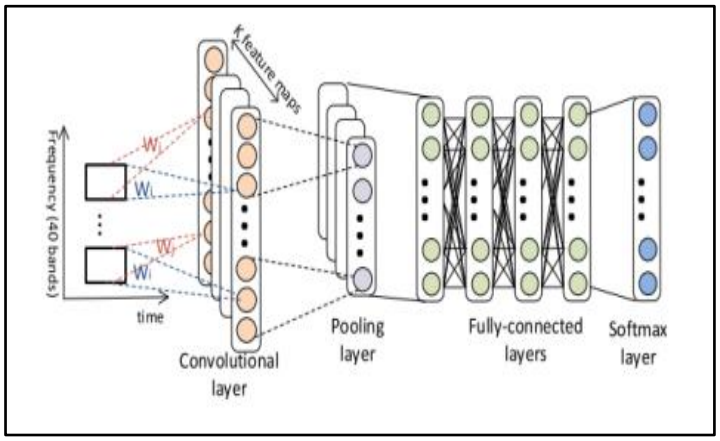

Fig: Convolution Neural Network [15]

Weight matrix W is replicated over all input and Convolution layer consist of many feature maps, generated with different localized filter to from different local pattern. For speech recognition task, input feature is a 2-D plane with frequency and time axis. After convolution layer, there is pooling layer which takes input from local region of the previous convolution layer and subsample it to produce a single output from that region. Pooling operator can be of type max or average pooling. Common pooling operator used by CNN is max pooling which selects the largest element from each sub-region. Subsampling not only reduce the computational complexity but also achieve degree of

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

robustness to modification in local pattern. Local behaviour of speech signals vary in low and high frequency regions. In past study, it has been proved that increase in the number of hidden units in convolution layer decreases the WER in ASR system.

B) Recurrent Neural Network

Recurrent neural network (RNNs) can be regarded as a class of deep generative architectures. RNNs are powerful in modeling sequential data of speech or text format. Depth of RNN can be as large as the length of input data sequence. Some past studies of RNN in speech recognition shows that instead of combining RNN with HMM , it is possible to train RNNs end-to-end for speech recognition. Depth of RNN can be expressed in terms of time, some researcher also studied whether it is beneficial to build RNN in depth of space. Depth in space can be obtained from stacking multiple recurrent hidden layers on top of each other. The structure of RNN is one in which each layer represents another step in time and that each time step gets one input and predicts one output. The transition function for each time step is constant.For a standard recurrent neural network following mathematical equations represent the prediction [5]

$$h_i = \sigma\,(W_{hh\,h_{i-1}} + W_{hx}x_i + b_h)$$

$$y_i = W_{yh}h_i$$

Where h_i is hidden layer at step i and x_i is the input layer at time step i, y_i is output layer at timestep i and the W are the weight matrix with bias b. Though theoretically RNNs are powerful, they are difficult for training. To improve the training capacity of RNN Hessian-free optimization have been applied. In addition to modifying the training algorithm, network architecture can also be modified to make it easier for training. Difficulty associated with training of RNN is mainly due to the highly volatile relationship between the parameters and the hidden states. Training of the RNN is performed using gradient methods.

C) Restricted Boltzmann Machines

Deep Boltzmann machine is generative model type [12]. Boltzmann machines are used to implement either generative or discriminative approach. Prior study showed that generative approach is good for speaker recognition while discriminative approach is good for speech recognition task of speech processing. RBM is variant of Boltzmann machines. RMB is stochastic neural network consisting of one layer of visible units, one layer of hidden units and bias unit. Each visible unit is connected to all hidden units and bias is connected to all hidden and visible units. In order to make learning of RBMs easier the network is restricted such as no two hidden units are connected and no two visible units are connected. RBMs have structure like bi-partite graph. RBM has fascinated many researcher due to the fact that it has been used as building block for the deep belief networks (DBN) model. RBM is a parameterized generative model representing the probability distribution. Learning a BM means if some training data is given, then adjusting the BM parameter such that the probability distribution represented by BM fits the training data. RBM is also particular kind of Markov Random Field (MRF).

A RBM is consists of m visible units V = $(V_1, \ldots V_m)$ to represent observable data and n hidden units H = $(H_1, \ldots H_n)$ to capture dependencies between observed variables. In binary RBM, the random variables (V, H) takes values (v, h) $\in \{0,1\}^{m+n}$ and the joint probability distribution under the model is given by as follow [19]:

$$p\,(v, h) = \frac{1}{z}e^{-E(v,h)}$$

with the energy function [19]

$$E\,(v, h) = -\sum_{i=1}^{n}.\sum_{j=1}^{m}.w_{i,j}\,v_j - \sum_{j=1}^{m}b_j\,v_j - \sum_{i=1}^{n}c_i\,h_i$$

For all i $\in$ {1…..,n} and j $\in$ (1……,m), $w_{i,j}$ is a real valued weight associated with the edge between units V_j and H_i and b_j *and* c_i are real valued bias terms associated with jth visible and ith hidden variable.

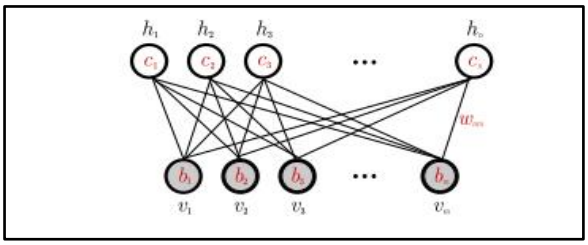

Fig: Restricted Boltzmann Machine model

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

D) Deep Belief Network

DBNs are formed by stacking up RBMs and trained them in greedy manner. DBNs are the graphical models which can learn deep hierarchical representation of training data. Previous studied proved that DBNs can be trained by principle ofgreedy layer-wise unsupervised training. For each layer, RBM is building block. After learning, DBN can be further trained in supervised fashion to perform classification. The process of training DBN in unsupervised way involve [12]:

1) Train the first layer as an RBM that models the raw input as its visible layer.
2) First layer can use for the representation of the input that will be use as data or second layer.
3) Data for the second layer can be obtained by sampling or by computing mean activation of hidden units.
4) Train the second layer as an RBM, taking the transformed data which, is obtained by sampling or mean activations as training data for the visible layer of that RBM.
5) Continue the above procedure for all desired layers. Either samples or mean values is propagating upward.

Fine tune all the parameters with respect to a supervised training criterion.

E) Deep Convex Nets

DBNs have been shown powerful in performing classification task including speech recognition. DBN training is difficult computationally. Some conventional techniques used for training DBNs involve stochastic gradient descent algorithm is difficult to parallelize on large scale of data. In 2011 Deng and Yu have proposed one deep architecture for solving the scalability problem.This new deep learning architecture is Deep Stacking Network (DSN)[12] also known as Deep Convex Nets (DCN). The basic idea behind this architecture is stacking, where simple modules of functions or classifiers are composed first and then they are stacked on top of each other to perform more complex function or classifiers. The name convex is due to the use of the convex optimization in learning the output network weights given the hidden units activities. Following fig shows two modules of DCN with their connections:

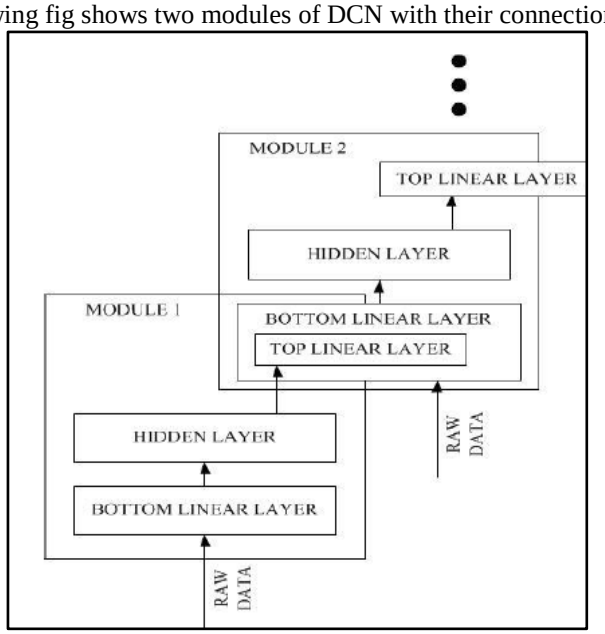

Fig: Deep Convex Network [17]

DCN consist of overlapping modules where each module includes three layers [17]:
-A first linear layer that includes set of linear input units equals to the dimensionality of the input features.
-A hidden layer that compromises set of non-linear units
-A second linear layer that compromises the plurality of linear output units (target classification classes).

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

If the DCN is utilized for the speech recognition task then a set of input units may corresponds to sample of speech waveform or the extracted features from speech waveform such as power spectra.

F) Deep Neural Network

A deep neural network is feed-forward network that has more than one layer of hidden units between its input and outputs. Each hidden unit uses some function to map its total input to scalar state that it sends to the next layer. Each hidden unit typically uses the logistic function to convert the total input to the scalar state that it sends to layer above.

$$y_j = logistic(x_j) = \frac{1}{1 + e^{-x_j}}$$

$$x_j = b_j + \sum_i y_i \, w_{i,j}$$

Where bj is bias of unit j and i is layer below current layer j and $w_{i,j}$ is the weight on a connection to unit j from unit i. Recently the advances in machine learning algorithms and computer hardware have led to more efficient methods for training DNNs that contain many layers of non-linear hidden units and a very large output layer. Some past studies have shown that DNNs can outperform GMMs at acoustic modeling for speech recognition on a variety ofdatasets including large datasets with large vocabularies.

IV. CHALLENGES OF DEEP ARCHITECTURES

Despite of dealing with complex task, feature extraction and representation power of deep architectures there are some training, optimization and scaling challenges present. In deep architectures if number of hidden layers are increasing then its will obviously going to increase depth and it seems that the accuracy will going to increase with that. But unfortunately this is not case.It is not necessary that every time if we increase the number of hidden layers in our deep architecture then accuracy will be increased. This is due to the fact that the speed of training or learning for each hidden layer is different.Many results of experiments on deep learning suggest that training deep networks involve difficult optimization. Like recurrent neural network involves difficulty in optimization due to local minima or ill-conditioning. Better optimization may also leads to problem in scaling computation.Another challenge is scaling computations. As deep learning uses the large modeling and huge datasets and capturing large amount of information, scaling is difficult.Inference and sampling is another challenge in deep architectures. As inference and sampling is the iterative process of learning it slows down the training of deep neural networks [8].

V. CONCLUSION

Deep Architectures are effective towards the feature representation and modeling technique.Basic working of deep architecture has been covered from speech processing task point of view. Speech signals are varying in nature and that's why there are many challenges in speech processing tasks. Deep learning has proved efficient in doing phoneme representation. DNNs, DBNs and CNNs are the prevalent for solving the mixed noise problems. Based on the way thesearchitectures learned features and how these architectures are intended for use, classificationhas been done in generative, discriminative and hybrid classes. Current challenges with deep architectures is scalability, complexity of network structure and dealing with more number of features.

REFERENCES

[1] L. Rabiner and B. Juang. "Fundamentals of Speech Recognition". Prentice-Hall, New Jersey,1993.
[2] Ben Gold and Nelson Morgan. "Speech and Audio Signal Processing ". Wiley, 2011.
[3] Jurafsky, Daniel, and James H. Martin. "Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition."
[4] Deng, Li, and Xiao Li. "Machine learning paradigms for speech recognition: An overview"Audio, Speech, and Language Processing, IEEE Transactions on 21.5 (2013): 1060-1089.
[5] Alex Graves, Abdel-rahman Mohamed, Geoffrey Hinton, "Speech Recognition with Deep Recurrent Neural Networks Department of Computer Science, University of Toronto.
[6] Deng, Li, et al. "Recent advances in deep learning for speech research at Microsoft." Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conferenceon. IEEE, 2013.
[7] Yu, Dong, and Li Deng. "Deep learning and its applications to signal and information processing [exploratory dsp]." Signal Processing Magazine, IEEE 28.1 (2011): 145-154.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

[8] Andrew Wagner, "Two Major Challenges with Speech-Recognition Technology, February27, 2013.
[9] Graves, Alan, Abdel-rahman Mohamed, and Geoffrey Hinton. "Speech recognition with deep recurrent neural networks." Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013.
[10] Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. "A fast learning algorithm for deep belief nets." Neural computation 18.7 (2006): 1527-1554.
[11] Ranzato, Marc Aurelio, et al. "On deep generative models with applications to recognition." Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011.
[12] Li Deng, "Three Classes of Deep Learning Architectures and Their Applications: A Tutorial Survey", Microsoft Research, Redmond, USA.
[13] Xue, Shaofei, et al. "Fast adaptation of deep neural network based on discriminant codes for speech recognition." Audio, Speech, and Language. Processing, IEEE/ACM Transactions on 22.12 (2014): 1713-1725.
[14] Siniscalchi, Sabato Marco, et al. "Speech recognition using long-span temporal patterns in a deep network model." Signal Processing Letters, IEEE 20.3 (2013): 201-204.
[15] Abdel-Hamid, Ossama, et al. "Convolutional neural networks for speech recognition." Audio, Speech, and Language Processing, IEEE/ACM Transactions on 22.10 (2014): 1533-1545.
[16] Deng, Li, et al. "Binary coding of speech spectrograms using a deep auto-encoder." Inter-speech. 2010.
[17] Deng, L. and Yu, D., "Deep Convex Network: A scalable architecture for Deep Learning,"Proc. Interspeech, 2010.
[18] Bengio, Yoshua, Aaron Courville, and Pierre Vincent. "Representation learning: A reviewand new perspectives." Pattern Analysis and Machine Intelligence, IEEE Transactions on35.8 (2013): 1798-1828.
[19] Jaitly, N.; Hinton, G., "Learning a better representation of speech soundwaves using restricted boltzmann machines," in Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on , vol., no., pp.5884-5887, 22-27 May 2011.
[20] Lee, Honglak, et al. "Unsupervised feature learning for audio classification using convolutional deep belief networks." Advances in neural information processing systems. 2009.
[21] Jaitly, Navdeep, and Geoffrey Hinton. "Learning a better representation of speech soundwaves using restricted boltzmann machines." Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. IEEE, 2011.
[22] Benigo Y., "Learning Deep Architectures for AI", Universite de Montreal C.P. 6128, Montreal, Qc, H3C 3J7, Canada.
[23] Bengio, Yoshua, et al. "Greedy layer-wise training of deep networks." Advances in neuralinformation processing systems 19 (2007): 153.
[24] T. Sainath, A. Mohamed, B. Kingsbury, B. Ramabhadran, "Deep convolutional neural networks for LVCSR",IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8614 - 8618, May 2013.
[25] Lu, Xugang, et al. "Speech enhancement based on deep denoising autoencoder." INTERSPEECH. 2013.

BIOGRAPHY

Renu Karuleis a Master'sStudent in the ComputerScience Department, D. Y. Patil College of Engineering Akurdi, Savitribai Phule Pune University. She received Bachelor of Technology degree in 2011 from GCOEA, Amravati. Her research interests are Machine Learning, Artificial Intelligence and Deep Learning.