Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 5, May 2015, pg.1017 – 1024

REVIEW ARTICLE

A Review: Speech Recognition with Deep Learning Methods

Rubi, Chhavi Rana

M.Tech, Software Engineering (2nd Sem), UIET, M.D. University Rohtak (Haryana), India

rubimalik95@gmail.com

ABSTRACT - Deep learning research has been successful beyond expectations in the last few years, both in terms of academic impact and industrial fallout. Deep learning is used in various fields for achieving multiple levels of abstraction like sound, text, images feature extraction etc. This paper discusses the concept of speech recognition with deep learning methods. Introduction of speech recognition, deep learning and deep learning methods is discussed in this review paper. Models of deep learning that are used in speech recognition is also described in this paper. This paper defines the related work on speech recognition using deep learning methods and about the sphinx, software allow the implementation of speech recognition in java language. The main motive of this review is to define the use of sphinx and eclipse to recognize speech.

KEYWORDS - Deep learning, Deep Belief Networks, Deep Convolutional Network, Restricted Boltzmann Machines, Sphinx, Eclipse, Speech Recognition.

I. INTRODUCTION

Deep Learning is a new area of Machine Learning research, which has been introduced with the objective of moving Machine Learning closer to one of its original goals: Artificial Intelligence. Deep Learning is about learning multiple levels of representation and abstraction that help to make sense of data such as images, sound, and text.

Deep learning has been interacted and closely related to the existing research fields such as neural networks, graphical models, feature learning, unsupervised learning, optimization, pattern recognition, and signal processing. This is also motivated by neuroscience, more similarities to our understanding of the brain's intelligence (learning from unlabeled data), and there are already number of applications in computer vision, speech recognition, natural language processing, hand writing recognition and so on. Deep learning is one of the progressive and promising areas in machine learning for the future tasks involved in machine learning especially in the area of neural network.

A. Deep learning algorithms

1) Deep Belief Networks
2) Deep Convolutional Network
3) Restricted Boltzmann Machines

1) Deep Belief Networks

Deep Belief Networks (DBNs) is basic kind of deep neural network architectures. This is a generative probabilistic model with one visible layer at the bottom and many hidden layers up to the output. Each hidden layer unit learns the statistical representation via the links to the lower layers. The more higher the layers, the more complex are the representations.

The main problem with the deep neural network architectures was the learning process, since the ordinary gradient descent algorithm does not work well and sometimes it makes the training quite impossible for a DBN. To circumvent this problem, a greedy layer wise unsupervised pre-training can be used. After the pre-training it is possible to do a successful supervised learning, done by a procedure called fine tuning, using the renowned gradient descent. The pre-training phase impacts on the choice of initial weights values for the actual supervised training stage. In practice it works magnificently better than the conventional random initialization of the weights, and causes to avoid local minima while using gradient descent in back propagation

DBNs are constructed by stacking many layers of restricted Boltzmann machines. An RBM comprises two layers, one is the visible and the other is hidden. Once we stack two RBMs on top of each other, the hidden layer of lower becomes visible to the top. The goal here is using the multiple layers of RBMs to model (represent) as close as possible to the reality, the distribution of the input. We are achieving this goal by multiple layers of non-linearity, which results in extraction the more accurate probabilistic representation for the input.

The number of layer and the number of units on each layer in the schema are only examples. So far we have seen that the pre-training phase for the DBN is done through a greedy bottom-up pass. One may wonder about the possible information hidden in the reverse pass (top-down). This is one of the shortcomings of the pre-training for DBNDNN architectures.

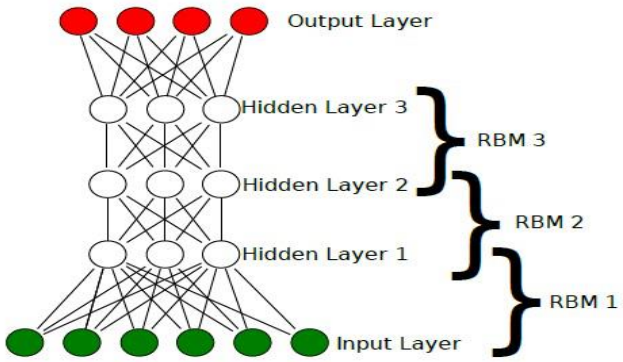

Fig- 1.2 The schema of a Deep Belief Network

2) Deep Boltzmann Machines

A Deep Boltzmann Machine (DBM) is a type of binary pair wise Markov random field (undirected probabilistic graphical models) with multiple layers of hidden random variables. It is a network of symmetrically coupled stochastic binary units. It comprises a set of visible units $V \in \{1, 0\}^D$ and a series of layers of hidden units $h^{(1)} \in \{1, 0\}^{F_1}$, $h^{(2)} \in \{1, 0\}^{F_2}$,……., $h^{(L)} \in \{1, 0\}^{F_L}$. There is no connection between the units of the same layer (like RBM).

For the DBM of the Figure 1.2, we can write the probability which is assigned to vector v as

$$p(v) = \frac{1}{Z}\sum_h e^{\sum_{ij} W_{ij}^{(1)} v_i h_j^1 + \sum_{jl} W_{jl}^{(2)} v_i h_j^1 h_l^2 + \sum_{lm} W_{lm}^{(3)} v_l h_m^4}$$

Where

$h = \{h^{(1)};h^{(2)};h^{(3)}$ are the set of hidden units

$\theta = \{W^{(1)};W^{(2)};W^{(3)}\}$ are the model parameters

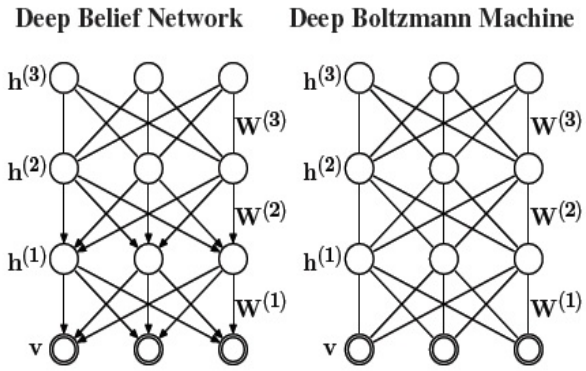

Fig- 1.3 A deep belief network (left) and deep Boltzmann machine (right).

There several reasons which motivate us to take advantage of deep Boltzmann machine architectures. Like DBNs they benefit from the ability of learning complex and abstract internal representations of the input in tasks such as object or speech recognition, with the use of limited number of labeled sensory data to fine-tune the representations which is built based on a large supply of unlabeled sensory input data. However, unlike DBNs and deep convolutional neural networks, they adopt the inference and training procedure in both directions, bottom-up and top-down pass, which enable the DBMs to better unveil the representations of the ambiguous and complex input structures.

3) Convolutional Neural Networks(CNNs)

CNNs belong to the feed forward network, but it combines three architectural ideas to ensure some degree of shift and distortion invariance: local receptive field, shared weights, and sometimes spatial or temporal sub sampling. The local receptive field is a small portion of the data, it has another name called the shift window; the shared weights makes all the feature maps at the same level be formed by the same parameterization; spatial and temporal sub sampling (usually the spatial subsuming) can shrink the feature maps by dimensions in the previous level, forming a new set of feature maps to which the convolution process can apply again.

The CNNs is comprised of a sequence of convolution process and sub sampling process the convolution process convolves an input with a trainable filter f_x and adds a trainable bias b_x to produce the convolution layer C_x, and the sub sampling process sums a neighborhood (four pixels), weights by scalar wx+1, adds the trainable bias bx+1, and passes through a sigmoid function to produce a smaller feature map Sx+1. Then, Cx is treated as the input data, converted to a set of smaller-dimension feature maps Sx+1 via the sub sampling process.

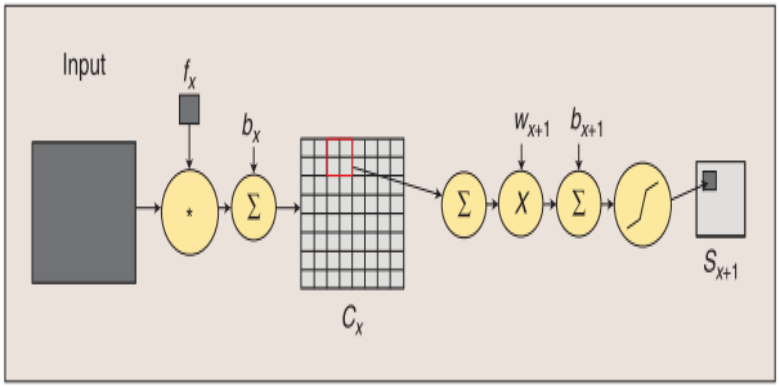

Fig- 1.4 The convolution process and sub sampling process

The input is converted into a convolution layer Cx via the convolution process. Then, Cx is treated as the input data, converted to a set of smaller-dimension feature maps Sx+1 via the sub sampling process. We can see that after each layer, the dimensions of the feature maps are decreased, which makes CNNs a useful model to reduce the number of parameters that must be learned and thus improves upon general feed-forward back-propagation training".

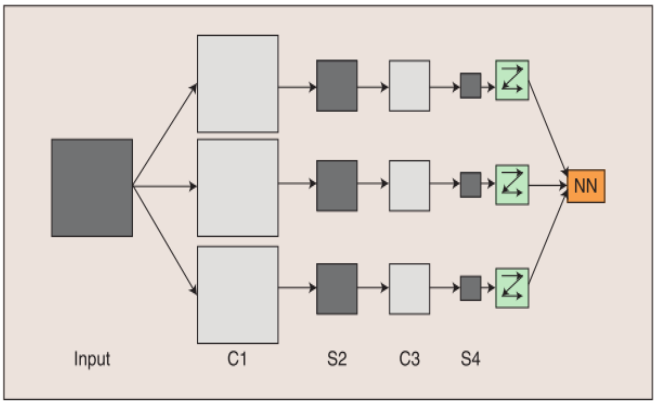

Fig- 1.5 A complete procedure of CNNs

B. Speech Recognition

Speech recognition means recognizing the speech and converting it into readable form or text. It is the ability of a machine or program to receive and interpret dictation, or to understand and carry out spoken commands. Speech recognition applications includes voice user interfaces such as voice dialing, call routine, search, simple data entry like entering credit card number etc.

Some requirements for speech recognition to implement in java are:

1) Sphinx 4
2) JSAPI
3) Eclipse 6.0
4) JDK

1) Sphinx 4

Sphinx 4 is a complete re-write of the Sphinx engine with the goal of providing a more flexible framework for research in speech recognition, written entirely in the 1ava programming language. Sun micro system supported the development of Sphinx 4 and contributed software engineering expertise to the project. Participants included individuals at merl, mit and cmu.

Current development goals include –

1. Developing a new acoustic model or trainer
2. implementing speaker adaptation
3. improving configuration management
4. Creating a graph based UI for graphical system design

2) JSAPI

The Java Speech API (JSAPI) is an application programming interface for cross-platform support of command and control recognizers, dictation system, and speech synthesizers.

It contains three packages:

1. javax.speech : Contains classes and interfaces for a generic speech engine.
2. Javax.speech.synthesis : Contains classes and interfaces for speech synthesis.
3. Javax.speech.recognition : Contains classes and interfaces for speech recognition.

3) Eclipse6.0

Eclipse is an integrated development environment. It contains a base workspace and extensible plug-in system for customizing the environment. Written mostly in java, it can be used to develop the applications. It can also be used to develop packages for software. Development environment include the Eclipse java development tool (JDT) for java.

4) JDK

The Java Development Kit contains tools need to develop java program. The tools includes compiler, java application launcher, applet viewer etc. Compiler converts java code into byte code. Application launcher open Java Runtime Environment (JRE), load classes and invoke main class. It work like a processer.

II. LITERATURE REVIEW

M.A.Anusuya and S.K.Katti [1, 3] presents a brief survey on Automatic Speech Recognition and discusses the major themes and advances made in the past 60 years of research, so as to provide a technological perspective and an appreciation of the fundamental progress that has been accomplished in this important area of speech communication. The design of Speech Recognition system requires careful attentions to the following issues: Definition of various types of speech classes, speech representation, feature extraction techniques, speech classifiers, data base and performance evaluation. The objective of this review paper is to summarize and compare some of the well known methods used in various stages of speech recognition system and identify research topic and applications which are at the forefront of this exciting and challenging field.

Santosh K.Gaikwad , Bharti W.Gawali and Pravin Yannawar [2] The Speech is most prominent & primary mode of Communication among of human being. The communication among human computer interaction is called human computer interface. Speech has potential of being important mode of interaction with computer .This paper gives an overview of major technological perspective and appreciation of the fundamental progress of speech recognition and also gives overview technique developed in each stage of speech recognition. This paper helps in choosing the technique along with their relative merits & demerits. A comparative study of different technique is done as per stages. This paper is concludes with the decision on feature direction for developing technique in human computer interface system using Marathi Language.

Shanthi Therese, Chelpa Lingam [4] Says that speech has evolved as a primary form of communication between humans. The advent of digital technology, gave us highly versatile digital processors with high speed, low cost and high power which enable researchers to transform the analog speech signals in to digital speech signals that can be scientifically studied. Achieving higher recognition accuracy, low word error rate and addressing the issues of sources of variability are the major considerations for developing an efficient Automatic Speech Recognition system. In speech recognition, feature extraction requires much attention because recognition performance depends heavily on this phase. In this paper, an effort has been made to highlight the progress made so far in the feature extraction phase of speech recognition system and an overview of technological perspective of an Automatic Speech Recognition system are discussed.

Sanjib Das [5] presents a brief survey on speech is the primary and the most convenient means of communication between people. The communication among human computer interaction is called human computer interface. Speech has potential of being important mode of interaction with computer. This paper gives an overview of major technological perspective and appreciation of the fundamental progress of speech recognition and also gives overview technique developed in each stage of speech recognition. This paper helps in choosing the technique along with their relative merits and demerits. A comparative study of different technique is done as per stages. This paper concludes with the decision on feature direction for developing technique in human computer interface system in different mother tongue and it also discusses the various techniques used in each step of a speech recognition process and attempts to analyze an approach for designing an efficient system for speech recognition. The objective of this review paper is to summarize and compare different speech recognition systems and identify research topics and applications which are at the forefront of this exciting and challenging field.

Nidhi Desai1, Prof. Kinnal Dhameliya, Prof. Vijayendra Desai[6,7] survey presents speech is the most natural form of human communication and speech processing has been one of the most inspiring expanses of signal processing. Speech recognition is the process of automatically recognizing the spoken words of person based on information in speech signal. Automatic Speech Recognition (ASR) system takes a human speech utterance as an input and requites a string of words as output. This paper introduce a brief survey on Automatic Speech Recognition and discuss the major subjects and improvements made in the past 60 years of research, that provides technological outlook and a respect of the fundamental achievement that has been accomplished in this important area of speech communication. Definition of various types of speech classes, feature extraction techniques, speech classifiers and performance evaluation are issues that require attention in designing of speech recognition system. The objective of this review paper is to summarize some of the well known methods used in several stage of speech recognition system.

Guillaume Gravier, Ashutosh Garg [8, 11] survey presents Visual speech information from the speaker's mouth region has been successfully shown to improve noise robustness of automatic speech recognizers, thus promising to extend their usability into the human computer interface. In this paper, we review the main components of audio-visual automatic speech recognition and present novel contributions in two main areas: First, the visual front end design, based on a cascade of linear image transforms of an appropriate video region-of-interest, and subsequently, audio-visual speech integration. On the later topic, we discuss new work on feature and decision fusion combination, the modeling of audio-visual speech asynchrony, and incorporating modality reliability estimates to the bimodal recognition process. We also briefly touch upon the issue of audiovisual speaker adaptation. We apply our algorithms to three multi-subject bimodal databases, ranging from small- to large vocabulary recognition tasks, recorded at both visually controlled and challenging environments. Our experiments demonstrate that the visual modality improves automatic speech recognition over all conditions and data considered, however less so for visually challenging environments and large vocabulary tasks.

Li Deng and John C. Platt [9] survey presents that deep learning systems have dramatically improved the accuracy of speech recognition, and various deep architectures and learning methods have been developed with distinct strengths and weaknesses in recent years. How can ensemble learning be applied to these varying deep learning systems to achieve greater recognition accuracy is the focus of this paper. We develop and report linear and log-linear stacking methods for ensemble learning with applications specifically to speech-class posterior probabilities as computed by the convoluional, recurrent, and fully-connected deep neural networks. Convex optimization problems are formulated and solved, with analytical formulas derived for training the ensemble-learning parameters. Experimental results demonstrate a significant increase in phone recognition accuracy after stacking the deep learning subsystems that use different mechanisms for computing high-level, hierarchical features from the raw acoustic signals in speech.

Li Deng, Jinyu Li, Jui-Ting Huang, Kaisheng Yao, Dong Yu, Frank Seide, Michael L. Seltzer, Geoff Zweig, Xiaodong He, Jason Williams, Yifan Gong, and Alex Acero Mic [10] survey describe that deep learning is becoming a mainstream technology for speech recognition at industrial scale. In this paper, we provide an overview of the work by Microsoft speech researchers since 2009 in this area, focusing on more recent advances which shed light to the basic capabilities and limitations of the current deep learning technology. We organize this overview along the feature-domain and model-domain dimensions according to the conventional approach to analyzing speech systems. Selected experimental results, including speech recognition and related applications such as spoken dialogue and language modeling, are presented to demonstrate and analyze the strengths and weaknesses of the techniques described in the paper. Potential improvement of these techniques and future research directions are discussed.

III. CHALLENGES

Speech Recognition is one of the challenging areas in computer science, a lot of pattern recognition methodology tried to resolve a good way and higher percentage of recognition. The challenges of deep learning are – making intelligent machines capable of not only hearing (speech) and seeing (vision), but also of thinking with a mind, i.e. reasoning and interface over complex, hierarchical relationships and knowledge sources that comprise a vast number of entities and semantics concept in the real world based in part on multi-sensory data from user. To this end, language and multimodal processing – joint exploitation and learning from text, speech (audio), and image (video) – is evolving into a new frontier of deep learning, beginning by mixture of research communities speech and spoken language processing, natural language processing, computer vision, machine learning.

Main challenges in speech recognition in deep learning methods are –

A. *The multi-model learning* – For advancing the ability of representation learning systems to discover semantic spaces that underlie multiple kinds of sensory input.

B. *The black-box learning* – For reducing the usefulness of having a human researcher working in the loop with the training algorithm.

IV. CONCLUSION

In this paper we have reviewed typical deep learning algorithms that are used in speech recognition and we also discussed about the software that are required to implement the speech recognition in java language.

BIBLIOGRAPHY

1. Li Deng, Jinyu Li, Jui-Ting Huang, Kaisheng Yao, Dong Yu, Frank SeideMichael L. Seltzer, Geoff Zweig, Xiaodong He, Jason Williams, Yifan Gong, and Alex Acero Microsoft Corporation, One Microsoft Way, Redmond, WA 98052, USA 2009

2. M.A.Anusuya and S.K.Katti ,Department of Computer Science and Engineering,Sri Jayachamarajendra College of Engineering, Mysore, India, (IJCSIS) International Journal of Computer Science and Inform4 ation Security,2009.

3. Shanthi Therese ,Chelpa Lingam, International Journal of Scientific Engineering and Technology , June 2013.,Review of Feature Extraction Techniques in Automatic Speech Recognition.

4. Speech Recognition Technique: A Review Sanjib Das Department of Computer Science, Sukanta Mahavidyalaya, (University of North Bengal), India, International Journal of Engineering Research and Applications (IJERA) May-Jun 2012.

5. Nidhi Desai1, Prof.Kinnal Dhameliya2, Prof.Vijayendra Desai3, International Journal of Emerging Technology and Advanced Engineering , December 2013, Feature Extraction and Classification Techniques for Speech Recognition: A Review.

6. Li Deng and John C. Platt, Microsoft Research, One Microsoft Way, Redmond, WA, USA, November 2010, Ensemble Deep Learning for Speech Recognition.

7. Santosh K.Gaikwad, Dr.Babasaheb Ambedkar Marathwada, Bharti W.Gawali, 2011, A Review on Speech Recognition Technique.

8. Samy Bengio and Georg Heigold, Google Inc, Mountain View, CA, USA, feb. 2007, Word Embeddings for Speech Recognition.

9. Audio-Visual Speech Gerasimos Potamianos, Member, IEEE, Chalapathy Neti, Member, IEEE, Guillaume Gravier,, Ashutosh Garg, Student Member, IEEE, and Andrew W. Senior, Member, IEEE 2006, Recent Advances in the Automatic Recognition.

10. Dandan Mo,December 4, 2012, A survey on deep learning: one small step toward AI.

11. Aalto University publication series, Foundations and Advances in Deep Learning, Kyunghyun Cho, 2014.

