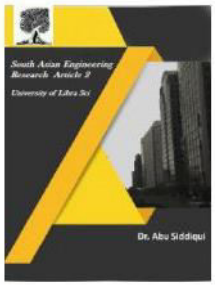
2581-4575
International Journal For Recent Developments in Science & Technology
Crossref
A Peer Reviewed Research Journal



FEATURE LEARNING IN DEEP NEURAL NETWORKS – STUDIES ON SPEECH RECOGNITION TASKS

T.SWATHI[1], B.AKHIL SAI[2], E.BHAVANA[3], DR. B. KRISHNA[4]

Student, Department of Computer Science and Engineering, CMR Technical Campus, Medchal, Hyderabad, Telangana, India[1,2,3]

Professor, Department of Computer Science and Engineering, CMR Technical Campus, Medchal, Hyderabad, Telangana, India[4]

Abstract :

Automatic speech recognition (ASR) has been an active research area for more than five decades. Recent studies have shown that deep neural networks (DNNs) perform significantly better than shallow networks and Gaussian mixture models (GMMs) on large vocabulary speech recognition tasks. In this paper, we argue that the improved accuracy achieved by the DNNs is the result of their ability to extract discriminative internal representations that are robust to the many sources of variability in speech signals. We show that these representations become increasingly insensitive to small perturbations in the input with increasing network depth, which leads to better speech recognition performance with deeper networks. We also show that DNNs cannot extrapolate to test samples that are substantially different from the training examples. If the training data are sufficiently representative, however, internal features learned by the DNN are relatively stable with respect to speaker differences, bandwidth differences, and environment distortion. This enables DNN-based recognizers to perform as well or better than state-of-the-art systems based on GMMs or shallow networks without the need for explicit model adaptation or feature normalization.

Keywords: Automatic speech recognition, GMM, HMM, Deep Neural Networks, Log-linear model.

1.Introduction:

In the recent years, in order to reduce human efforts and time consumed in communicating with a machine there has been an increasing interest in implementation of various algorithms to automate the automatic speech recognition with advanced deep learning techniques combined with image processing techniques. Automatic speech recognition (ASR) has been an active research area for more than five decades. However, the performance of ASR systems is still far from satisfactory and the gap between ASR and human speech recognition is still large on most tasks. One of the primary reasons' speech recognition is challenging is the high variability in speech signals. For example, speakers may have

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

different accents, dialects, or pronunciations, and speak in different styles, at different rates, and in different emotional states. The presence of environmental noise, reverberation, different microphones and recording devices results in additional variability. In this paper, we argue that the improved accuracy achieved by the DNNs is the result of their ability to extract discriminative internal representations that are robust to the many sources of variability in speech signals. We show that these representations become increasingly insensitive to small perturbations in the input with increasing network depth, which leads to better speech recognition performance with deeper networks. We also show that DNNs cannot extrapolate to test samples that are substantially different from the training examples. If the training data are sufficiently representative, however, internal features learned by the DNN are relatively stable with respect to speaker differences, bandwidth differences, and environment distortion. This enables DNN- based recognizers to perform as well or better than state-of-the-art systems based on GMMs or shallow networks without the need for explicit model adaptation or featurenormalization.

An automatic speech recognition system requires three main sources of knowledge: an acoustic model, a phonetic lexicon and a language model [3]. Acoustic model characterizes the sounds of the language, mainly the phonemes and extra sounds (pauses, breathing, background noise, etc). The phonetic lexicon contains the words that can be recognized by the system with their possible pronunciations. Language model provides knowledge about the word sequences that can be uttered. In the state-of-the-art approaches, statistical acoustic and language models, and to some extent lexicons, are estimated using huge audio and textcorpora.

2.Literature Review:

Review of literature on speech recognition systems genuinely demands the very first attention towards the discovery of Alexander Graham Bell about the process of converting sound waves into electrical impulses and the first speech recognition system developed by Davis et for recognizing telephone quality digits spoken at normal speech rate. This effort for automatic recognition of speech was basically centered on the building up of an electronic circuit for recognizing ten digits of telephone quality. Spoken utterances were analyzed to get a 2-dimensional plot of formant 1 vs formant 2. For pattern matching, a circuit was designed for determining the highest relative correlation coefficient between a set of new incoming data and each of the reference digit patterns. It was also observed that circuit adjustment helps the recognition system to perform well for the speech of different speakers. An indication circuit was built to display the recognized spoken digit. The approaches to speech recognition, evolved thereafter, had a major stress on finding speech sounds and providing appropriate labels to these sounds. Various approaches and types of speech recognition systems came into existence in

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

last five decades gradually. This evolution has lead to a remarkable impact on the development of speech recognition systems for various languages worldwide. Automatic speech recognition has been viewed as successive transformations of acoustic micro structure of speech signal into its implicit phonetic macro-structure. In other words, a speech recognition system is a speech-to-text conversion wherein the output of the system displays text corresponding to the recognized speech. Languages, on which so far automatic speech recognition systems have been developed, are just a fraction of total around 7300 existing languages. Russian, Portuguese, Chinese, Vietnamese, Japan, Spanish, Filipino, Arabic, English, Bengali, Tamil, Malayalam, Sinhala, Hindi are prominent among them. English is the language for which maximum work for recognition isdone.

Existing System :

Conventional speech recognizers use a hidden Markov model (HMM) in which each acoustic state is modeled by a Gaussian mixture model (GMM). The model parameters can be discriminatively trained using an objective function such as maximum mutual information (MMI) or minimum phone error rate (MPE) . Such systems are known to be susceptible to performance degradation when even mild mismatch between training and testing conditions is encountered. To combat this, a variety of techniques has been developed. For example, mismatch due to speaker differences can be reduced by Vocal Tract Length Normalization (VTLN) , which nonlinearly warps the input feature vectors to better match the acoustic model, or Maximum Likelihood Linear Regression (MLLR), which adapt the GMM parameters to be more representative of the test data. Other techniques such as Vector Taylor Series (VTS) adaptation are designed to address the mismatch caused by environmental noise and channel distortion. While these methods have been successful to somedegree, they add complexity and latency to the decoding process. Most require multiple iterations of decoding and some only perform well with ample adaptation data, making them unsuitable for systems that process short utterances, such as voice search. Recently, an alternative acoustic model based on deep neural networks (DNNs) has been proposed. In this model, a collection of Gaussian mixture models is replaced by a single context-dependent deep neural network (CD-DNN). A number of research groups have obtained strong results on a varietyof large-scale speech tasks using this approach .Because the temporal structure of the HMM is maintained, we refer to these models as CD-DNN-HMM acoustic models.

Proposed System:

In this paper, we analyze the performance of DNNs for speech recognition and in particular, examine their ability to learn representations that are robust to variability in the acoustic signal. To do so, we interpret the DNN as a joint model combining a nonlinear feature transformation and a loglinear classifier. Using this view, we show that the many layers of nonlinear

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

transforms in a DNN convert the raw features into a highly invariant and discriminative representation which can then be effectively classified using a log-linear model. These internal representations become increasingly insensitive to small perturbations in the input with increasing network depth. In addition, the classification accuracy improves with deeper networks, although the gain per layer diminishes. However, we also find that DNNs are unable to extrapolate to test samples that are substantially different from the training samples. A series of experiments demonstrates that if the training data are sufficiently representative, the DNN learns internal features that are relatively invariant to sources of variability common in speech recognition such as speaker differences and environmental distortions. This enables DNN- based speech recognizers to perform as well or better than state-of-the-art GMM-based systems without the need for explicit model adaptation or feature normalizationalgorithms.

3.System Requirements:

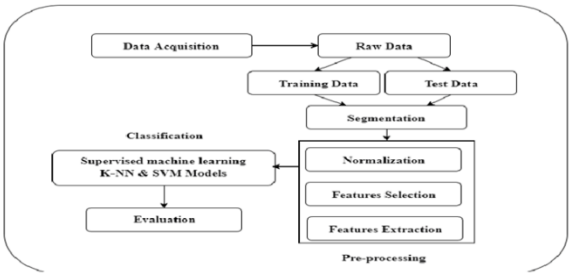
To be used efficiently, all computer software needs certain hardware components or other software resources to be present on a computer. These prerequisites are known as system requirements and are often used as aguidelines.

The hardware requirements used are I5 PROCESSOR 2.24 GHz system as it is available in multiple speeds, ranging from 1.90 GHz up to 3.80 GHz, and it features 3 MB, 4 MB or 6 MB of cache, 300GB Hard disk which is an electro-mechanical data storage device that uses magnetic storage to store and retrieve digital data using one or more rigid rapidly rotating platters coated with magnetic material and 16GB RAM.

The software requirements are Windows Operating System which serves as a base for computer programs to work off of. The OS controls the system's hardware, making it so internal components and peripherals work across all programs, Python as a front-end language which is a high level language used for GUI applications allows us to focus on core functionality of the application by taking care of common programming tasks and PostgreSQL for the backend which is a free and open- source relational database management system emphasizing extensibility and technical standards compliance. It is designed to handle a range of workloads, from single machines to data warehouses or Web services with many concurrentusers.

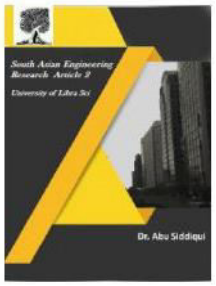
4.Architecture:

In the Machine learning architecture , the raw data goes through many steps before providing the output. The whole data is screened and inconsistencies and identical data is removed from the data set resulting in the output with complete accuracy.



Initially the data is received as a dataset is treated as a raw data .Then the data is

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal



divided into training data and test data. Usually we use 90% data for training and the remaining 10% is used for testing. Analysing the raw data into training data and test data is subjected to change based on the acceptance criteria for the accuracy of the model. Then the data goes through segmentation phase. It involves dividing the input into segments to simplify image analysis. Segments represent objects or parts of objects , after the segmentation phase the data is tested. This phase is optional based on the requirement.The next phase is Preprocessing phase which consists of three other sub phases .They are Normalization ,Feature selection and Feature extraction .

Normalization is a technique for training very deep neural networks that standardizes the inputs to a layer for each mini-batch. This has the effect of stabilizing the learning process and dramatically reducing the number of training epochs required to train deep networks. In Normalization we remove the repetitive data from the dataset which as it increases the accuracy of the model.Feature selection is one of the important aspects in data mining. Determining a subset of the initial features is called feature selection .Its necessity is felt due to very high dimensionality of data sets and growing computational methodologies of the target problems .Data mining aids in storing huge data and these data is full of noise i.e. redundant and irrelevant features .It is the pre -processing step where the noise is filtered, resulting in reducing the dimensionality of the data set and aids in creating computationally effective models with less time and cost. Feature extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and leading to better human interpretations. Feature extraction is related to dimensionality reduction. When the input data to an algorithm is too large to be processed and it is suspected to be redundant then it can be transformed into a reduced set of features.

Then we use the supervised machine learning algorithms like KNN and SVM models for classification using machine learning which is then used for evaluation of real time data.

5.Working :

Automatic speech recognition (ASR) has been an active research area for more than five decades. However, the performance of ASR systems is still far from satisfactory and the gap between ASR and human speech recognition is still large on most tasks. virtually impossible to avoid some degree of mismatch between the training and testing conditions. Conventional speech recognizers use a hidden Markov model (HMM) in which each acoustic state is modeled by a Gaussian mixture model (GMM). The model parameters can be discriminatively trained using an objective function such as maximum mutual information (MMI) [1] or minimum phone error rate (MPE) [2]. Such systems are known to be susceptible to performance degradation when even mild mismatch between training and testing conditions is

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

encountered. In this paper, we analyze the performance of DNNs for speech recognition and in particular, examine their ability to learn representations that are robust to variability in the acoustic signal. To do so, we interpret the DNN as a joint model combining a nonlinear feature transformation and a loglinear classifier. Using this view, we show that the many layers of nonlinear transforms in a DNN convert the raw features into a highly invariant and discriminative representation which can then be effectively classified using a log-linearmodel.

Deep Neural Networks :

A deep neural network (DNN) is conventional multi-layer perceptron (MLP) with many hidden layers (thus deep). If the input and output of the DNN are denoted as x and y, respectively, a DNN can be interpreted as a directed graphical model that approximates the posterior probability py|x(y = s|x) of a class s given an observation vector x, as a stack of (L + 1) layers of log-linear models. ThefirstLlayersmodeltheposteriorprobabilitiesofhiddenbinaryvectorshAgiveninputvectorsvA. If hAconsists of NAhidden units, each denoted as hA, the posterior probability can be expressed

$$p^A(h^A|v^A) = \frac{e^{z^A(v^A)\cdot h^A}}{\sum_{j=1}^{N^A} e^{z^A(v^A)\cdot 1} + e^{z^A(v^A)\cdot 0}}, \quad 0 \le A < L$$

Note that the equality betwe|en pyx(|y = sx) and pL(|y = svL) isvalid by making amean-field approximation [14] at each hidden layer.

DNNs learn more invariant features:

We have noticed that the biggest benefit of using DNNs over shallow models is that DNNs learn more invariant and discriminative features. This is because many layers of simple nonlinear processing can generate acomplicated nonlinear transform−.To show that this nonlinear transform is robust to small variations in the input features, let's assume the output of layer l 1, or equivalently the input to the layer l is changed from vAto vA+ δA, where δAis a small change. This change will cause the output of layer l, or equivalently the input to the layer A + 1 to change by

$$\delta^{A+1} = \sigma(z^A(v^A+\delta)^A) - \sigma(z(^Av^A)) \approx \mathrm{diag}\,\sigma'(z^I(^Av^A)^A)(w)\,\delta^{A\,TA}$$

Table 1: Effect of CD-DNN-HMM network depth on WER (%) on Hub5'00-SWB using the 309- hour Switchboard training set. DBN pretraining is applied.

$L \times N$	WER	$1 \times N$	WER
1×2k	24.2	–	–
2×2k	20.4	–	–
3×2k	18.4	–	–
4×2k	17.8	–	–
5×2k	17.2	1×3772	22.5
7×2k	17.1	1×4634	22.6
9×2k	17.0	–	–
5×3k	17.0	–	–
–	–	1×16k	22.1

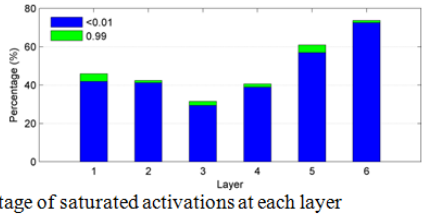


Figure 1: Percentage of saturated activations at each layer

The norm of the change δ^{A+1} is

$$\|\delta^{A+1}\| \approx \|\mathrm{diag}(\sigma'(z^A(v^A)))((w^A)^T\,\delta^A)\|$$
$$\le \|\mathrm{diag}(\sigma'(z^A(v^A)))(w^A)^T\|\|\delta^A\|$$
$$= \|\mathrm{diag}(v^{A+1}\circ(1-v^{A+1}))(w^A)^T\|\|\delta^A\|$$

where ◦ refers to an element-wiseproduct.

Learning by seeing:

We showed empirically that small perturbations in the input will be gradually shrunk as we move to the internal representation in the higher layers. In this

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

section, we point out that the above result is only applicable to small perturbations around the training samples. When the test samples deviate significantly from the training samples, DNNs cannot accurately classify them.In other words, DNNs must see examples of representative variations in the data during training in order to generalize to similar variations in the test data.

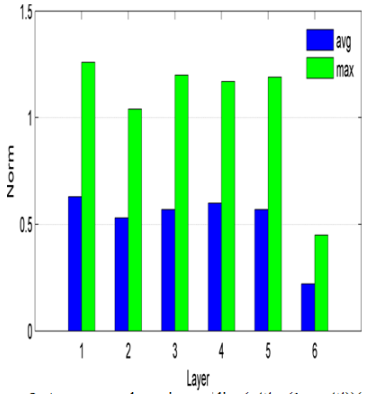


Figure 2: Average and maximum $\|\mathrm{diag}(v^{d+1} \circ (1 - v^{d+1}))(w^d)^T\|_2$ across layers on a 6×2k DNN.

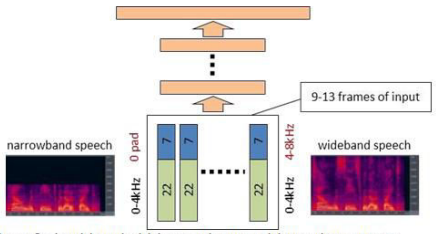


Figure 3: Illustration of mixed-bandwidth speech recognition using a DNN

Robustness to speaker variation:

A major source of variability is variation across speakers. Techniques for adapting a GMM-HMM to a speaker have been investigated for decades. Two important techniques are VTLN [3], and feature- space MLLR (fMLLR) [4]. Both VTLN and fMLLR operate on the features directly, making their application in the DNN context straightforward.

Robustness to environmental distortions:

In many speech recognition tasks, there are often cases where the despite the presence of variability in the training data, significant mismatch between training and test data persists. Environmental factors are common sources of such mismatch, e.g. ambient noise, reverberation, microphone type and capture device. The analysis in the previous sections suggests that DNNs have the ability to generate internal representations that are robust with respect to variability seen in the training data. In this section, we evaluate the extent to which this invariance can be obtained with respect to distortions caused by the environment

6.Conclusion:

In this paper we demonstrated through speech recognition experiments that DNNs can extract more invariant and discriminative features at the higher layers. In other words, the features learned by DNNs are less sensitive to small perturbations in the input features. This property enables DNNs to generalize better than shallow networks and enables CD-DNN-HMMs to perform speech recognition in a manner that is more robust to mismatches in speaker, environment, or bandwidth. On the other hand, DNNs cannot learn something from nothing. They require seeing representative samples to perform well. By using a multi-style training strategy and letting DNNs to generalize to similar patterns, we equaled the best result ever reported on the Aurora 4 noise robustness benchmark task without the need for

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

multiple recognition passes and model adaptation.

7.Acknowledgement:

The authors would like to acknowledge the support of the Chairman, Director and Head of the Department, Department of Computer Science and Engineering, CMR Technical Campus, Medchal, Hyderabad, Telangana, for their encouragement to the authors.

8.References:

[1] L. Bahl, P. Brown, P.V. De Souza, and R. Mercer, "Maximum mutual information estimation of hidden markov model parameters for speech recognition," in Proc. ICASSP, Apr, vol. 11, pp. 49–52.

[2] D. Povey and P. C. Woodland, "Minimum phone error and i-smoothing for improved discriminative training," in Proc. ICASSP,2002.

[3] P. Zhan et al., "Vocal tract length normalization for lvcsr," Tech. Rep. CMU-LTI-97-150, Carnegie Mellon Univ,1997.

[4] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," Computer Speech and Language, vol. 12, pp. 75–98,1998.

[5] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "HMM Adaptation Using Vector Taylor Series for Noisy Speech Recognition," in Proc. of ICSLP,2000.

[6] D. Yu, L. Deng, and G. Dahl, "Roles of pretraining and fine-tuning in context-dependent DBN-HMMs for real-world speech recognition," in Proc. NIPS Workshop on Deep Learning and Unsupervised Feature Learning,2010.

[7] G.E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pretrained deep neural networks for large vocabulary speech recognition," IEEE Trans. Audio, Speech, and Lang. Proc., vol. 20, no. 1, pp. 33–42, Jan.2012.

[8] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," in Proc. Interspeech,2011.

[9] F. Seide, G.Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in Proc. ASRU, 2011, pp.24–29.

[10] D. Yu, F. Seide, G.Li, and L. Deng, "Exploiting sparseness in deep neural networks for large vocabulary speech recognition," in Proc. ICASSP, 2012, pp.4409–4412.

[11] N. Jaitly, P. Nguyen, A. Senior, and V. Vanhoucke, "An application of pretrained deep neural networks to large vocabulary conversational speech recognition," Tech. Rep. Tech. Rep. 001, Department of Computer Science, University of Toronto,2012.

[12] T. N. Sainath, B. Kingsbury, B. Ramabhadran, P. Fousek, P. Novak, and A. r. Mohamed, "Making deep belief networks effective for large vocabulary continuous speech recognition," in Proc. ASRU, 2011, pp.30–35.

[13] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Large vocabulary continuous speech recognition with contextdependent dbn-hmms," in Proc. ICASSP, 2011, pp.4688–4691.