Pigeon Inspired Optimization for DNN-based ASR system

Shubhanshi Singhal

Assistant Professor, Department of Computer Engineering, TERii, Kurukshetra, India

Vishal Passricha

Assistant Professor, Department of Computer Engineering, National Institute of Technology, Kurukshetra

ABSTRACT—Neural Networks have been achieved great success in various fields like natural language processing, image recognition, computer vision task etc. Acoustic signals are also successfully modeled by the artificial neural networks in automatic speech recognition (ASR) systems. ASR is the task of mapping human speech into its corresponding text without any intervention of Human. Earlier, GMM/HMM-based acoustic models were used for the purpose of acoustic modeling. Deep Neural Networks have almost replaced the GMM/HMM-based model in the current era. In ASR systems,learning and validation are two phases that are performed to complete the task. The first phase is known as learning phase in which ASR system is trained. In this, DNN adjusts their weights to learn the signals. The second phase is known as validation phase. In this, the testing of ASR system is performed means how accurately it maps the acoustic signals into corresponding text. The recognition rate is used to measure the accuracy of the ASR systems means how accurately it maps the speech into text. It is directly related to its training. Stochastic Gradient Decent training is most popular and widely used training method for DNNs but it is more prone to overfitting. Overfitting means the weights of the network is so tuned that they cannot adjust themselves to reduce error. To overcome this problem, a pigeon inspired optimization (PIO) technique is applied to optimize the weight matrix of DNN. The PIO uses available heuristic to optimize the weight matrix andget more accurate results. The performance of ASR system that is optimized using PIO is evaluated for phoneme recognition on TIMIT dataset. The phoneme error rate (PER) achieved by it is 17.2% and relative improvement of 0.6% is achieved in PER over segmental recurrent neural networks. It also reduces the training time of DNN.

Keywords—Automatic Speech Recognition, Deep Neural Networks, Hidden Markov Model, Pigeon Inspired Optimization.

I. INTRODUCTION

Speech isa communication medium through which people interact with each other to share their information. Automatic Speech Recognition (ASR) is a computer-driven program whichtranslate human audio signal recorded by the microphone into the meaningful textual form. Transcription of the speech signals into its corresponding words with high accuracy is a challenging task due to variability present in speech signals like speaker accent, gender, age, and unwanted noise[1]. ASR system performs this in two stages. Initially, features are extracted from raw speech signals which are representations of short window power spectrum of frequency commonly derived using Fourier transformation of signal then take logs power for each frequency[2].Mel frequency cepstral coefficients (MFCCs) [3] or perceptual linear prediction[4] are popular feature extraction techniques.Extracted features represents the prior knowledge of acoustic speech production. Acoustic models relate the extracted feature into class conditional probability. Earlier, Gaussian Mixture Models (GMMs) [5]are widely used as an acoustic model but the emergence of deep learning models like Deep Neural Network (DNN)[6]have almost replaced the GMM-based acoustic models.GMM models were inefficient to classify the data which lie on or near to classification boundary of data surface line.DNNs have many hidden layers and a large output layer[6]. Large output layer benefits to accommodate many hidden Markov model(HMM) states which offer high discriminative power for phonemes recognition. The other benefit of the deep architecture is that it enables the ASR system to overcome the variation present in acoustic speech signal which is also given as an input to the first input layer. Many hidden layers and many neurons per layer of DNN make them more capable to create a complex and nonlinear relationship between acoustic inputs to an output. It has the capacity to handle large vocabulary dataset as training set and consequently reduces translational variances.In the second stage, the most likelywords areestimated using eithera statistical model or conditional model. For decoding purpose, an HMMis used to estimate the most likely sequence of utterance or phonemes[7]. This sequence is mapped with most likely stored words. The recent improvement in acoustic

models makes the system able to recognize the sequence of words with more accuracy and more robustness for large vocabulary recognition. DNN usesboth supervised and unsupervised learning mechanism for training from data and produces a discriminative function to map untrained data. Backpropagation techniques like stochastic gradient descent algorithm are used to adjustthe neuron's weight by calculating gradient or loss function. From few years, researchers of Microsoft, Google and IBM have beenachieved great achievement using DNNs as an acoustic model[8]. AlthoughDNN acoustic models are in leading role due to its ease of use and easy architecture. However, limited performance due to overfitting of training data is still an issue with DNN. The architecture modification is not so easy hence the performance can be risen only through other methods like generalization, optimized training etc. Pigeon inspired optimization(PIO) is a novel swarm intelligence algorithm used as optimization method[9]. To optimize the training process,PIOtechnique is usedthat updates the weight matrix using the heuristic available. By this method, the PER reduces to 17.2% on TIMIT dataset and a relative gain of 0.6% is achieved on segmental Recurrent Neural Network[10].

The rest of the paper is organized as follows. Section II provides related work. In section III, Pigeon inspired optimization algorithm is explained and applied on DNN/HMM based acoustic model. Section IVdescribes experimental setupused and the result afteroptimized training process is shown. We also compare these results with the existing models .Finally,the conclusionof the paper is given in section V.

II. RELATED WORK

Researchers began work on speech recognition in year 1950s with the help of a digital computer.Initially, they used analog to digital converters and frequency spectrogram for feature extraction from speech sound. HMM is a statistical model,described by Leonard E. Baum in the 1970s[11]. HMM is defined as a well-established approach for recognition application. Its parameters contain the characteristics of self-learning from its training data. It was first used automatic speech recognition in the 1990s to enhance the efficiency of speech recognition system.HMM has been dominant for ASR for atleast two decades. One of the critical parameters of HMM is the state observation probability distribution. In conventional HMM for ASR, GMM is used to model the state observation probabilities. The GMM/HMM are typically trained based on maximum likelihood criterion or other discriminative training strategies [12]. Various paradigms like neural network techniques, discriminative and connectionist approaches with HMM are also proposed.On the other hand, various generative models are explored such as context dependent HMM that uses Baum-Welch algorithm.GMM was firstly used in feature classification in the 1990s by Rose and Reynold [13]. This is a probabilistic modelsuccessfully used in ASR system based on expectation maximization algorithm. First hybrid model of NN-HMM was successfully used by J. Tebelskis in the year 1995 [14]. In thishybrid model, neural networks were used for acoustic modeling and HMM for decoding the most probable sequence of phonemes. Recently DNN has been replaced GMM from ASR task and computes state observation probabilities for all tied states in the HMM set [15, 16]. Earlier, neural networkshaving a few hidden layerswere used for classifyingthe phonemes from cepstral features. Recently, deep neural networks that have many hidden layers offer better result in acoustic modeling in ASR system. G. Hintonet al.[6]successfully implementedvarious deep architecture for acoustic modeling in speech recognition in2012. It has been reported that DNN/HMM has been achieved a large gain in many challenging ASR tasks [17, 18].

A. Hidden Markov Model

An HMM [19] is a stochasticmodel with related to Markov chain process that cannot observe directly but can be observed with the help of another stochastic process. HMM is used to model a word in a vocabulary where each hidden state represents a phoneme and calculates the most probable sequence of phonemes[20]. After the completion of training, HMM is used for decoding the sequence of words or pattern matching. HMM has three major works in speech recognition.

1) Evaluation problem: HMM model calculates the probability of a sequence of visible state v^T given θ model.

$$P\left(\mathrm{v}^{\bar{\mathrm{T}}}/_{\theta}\right) = \sum_{r=1}^{r_{max}} P(\mathrm{v}^{\bar{\mathrm{T}}} / \omega_r^T)\, P(\omega_r^T) \quad (1)$$

$$P(\omega_r^T) = \prod_{t=1}^{T} P\,(\omega(t)/\omega(t-1)) \quad (2)$$

$$P(v^{\bar{T}}/_{\omega_r^T}) = \prod_{t=1}^{T} P\,(v(t)/\omega(t)) \quad (3)$$

where T indicates a number of visible states r_{max}. is a number of the possible sequence of ω^t.

2) Decoding problem: A θ model of HMM calculate the most likely sequence or most probable sequence of the hidden state over which the machine has transition during generating a sequence of the visible state v^T.

3) Learning problem: HMM model is trained by supervised learning. HMM model trained the state transition probability (a_{ij}) and visible symbol emission probability b_{ik} using the backward algorithm. Backward algorithm calculates the probability that the machine will be in state (ω_i) at time instant t and will generate the remaining part of set visible symbol (v^T).

B. Deep Neural Network

The structure of DNN is a multi-layer perceptron (MLP). An$(L+1)$-layer MLP is used to model the posterior probability $P_{s|o}(s|o)$ of an HMM tied state s given an observation vector o.The first layers, $l = 0, \ldots, L-1$, are hidden layers that model posterior probabilities of hidden nodes h^i given input vector v^t from the previous layer while the top layer L is used to compute the posterior probability for all tied states using softmax:

$$P^l{}_{h_j|v}\left(h^l{}_j \middle| v^l\right) = \frac{1}{1+e^{-z^l{}_j(v^l)}} = \sigma\left(z_j^l(v^l)\right), \quad 0 \le l \le L \quad (4)$$

$$P^L_{s|v}(s|v^L) = \frac{e^{-z^L_j(v^L)}}{\sum_{s'} e^{-z^L_{s'}(v^L)}} = softmax_s(z^L(v^L)) \quad (5)$$

$$z^L(v^L) = (W^l)^T v^l + a^l \quad (6)$$

where W^l and a^l denote weight matrix and bias vectors for hidden layer l, and h^l_j and $z^l_j(v^l)$ denote the j[th] component of the hidden node h^l,and its activation $z^l_j(v^l)$ respectively.

C. PIO

Particle Swarm Optimization, Ant Colony Optimization, Artificial Bee Colony Optimization algorithms are popular optimization algorithms. Although these optimization algorithms have remarkable performance in solving optimization problems, still there is also the large space of improvement. In recent years, population-based swarm intelligence algorithms have been studied in depth and used in many areas to solve the optimization problem. The PIO algorithm is a novel swarm intelligence algorithm proposed by duan & Qiao in 2014[9]. In nature, pigeons find their destinations by relying onthe sun, magnetic field, and landmarks. The basic PIO has two operators which are map and compress operator and landmark operator. The map and compress operator is based on magnetic field and sun, and the landmark operator is based on landmarks.PIO has the capability of problem-solving can be used in various field of optimization like a shortest path in traveling salesman problems. PIO can also be applied to update the weight matrix of DNN to overcome the overfitting problem.

III. PROPOSED WORK

There are many generalization methods like dropout, drop-connect, Weight-tying etc. that solve the overfitting problem of DNN. However, the issue with these algorithms is that they re-initialize some weights to zero and create randomness. No doubt, most of the time, they improve the performance but sometimes they may deteriorate the performance of the system. In this section, we tried to resolve the overfitting issue of DNN. Pigeon optimization algorithm is inspired by bio-inspired optimization based on swarm behavior like fireflies, ant,and bee which is implemented for optimization problems. The leader of the pigeon flock initiates conversation and signal to another pigeon in the flock who acknowledge back by emulating the behavior of calling pigeon and manage side by side structure emerge in a flock of definite shape. The leader of the pigeon of the flock is chosen on the basis of the number of times calls to another pigeon in the flock. A fitness function $f(x)$ attach to every pigeon that count how many times a particular pigeon called to other pigeon in given population.From this functioning, PIO algorithm is applied tooptimize the weight matrix ofDNN/HMM model for speech recognition task. Here training of DNN has performed with the help of pigeon inspired optimization (PIO) technique.The original motive of this optimization approach is to achieve better accuracy and performance by optimizing thetraining processes of DNN/HMM model.PIO uses the available heuristic of the model to minimize mean square error. For each given inputs,the acoustic feature vector is estimated and given toDNN that calculatesposterior probabilities means state transition probability for each state.

DNN is generally trained with the help of backpropagating error derivative technique where the difference between actual output and expected output derivative is fed to the input node. There exist many hidden layers between input and outputs layer. Every hidden unit, j, uses a sigmoid function to map its total inputs from the previous layer, x_j, to scalar state, y_j which is sends to next upper layer unit.

$$y_j = logistic(x_j)\frac{1}{1+e^{-x_j}}; \quad x_j = b_j + \sum_i y_i w_{ij} \quad (7)$$

where b_j is bias of unit j, i is an index of the unit in the previous layer,w_{ij} is weight between below unit i to upper unit j. Unfortunately, in the case of overfitting, the adjustment of weights is not done i.e. no changes in weights. The PIO selects these kind of weights using the available heuristics and update them accordingly. By this, there is no randomness caused in network as caused by earlier methods. It offers a significant gain in the recognition rate.

Algorithm:

1. Let ∂_0 be (max) number of calls a pigeon p will make to group and search.
2. Let n be random pigeon population.
3. consider a population x_i with $iinrange : 1\ ton$
4. Select a flock of pigeon population having better searching time.
5. Let location of food be A
6. While food is located or $f(x_i) > \partial_0$do
   a. Evaluate vision radius of groupR_g and vision radius of pigeonR_i.
   Where $R = \sqrt{(lengthoflineofsight)^2 - (height)^2}$
   b. If food is not in $R_g$for a predefined time bound
   Redo the procedure after changing the route
   c. If food is in R_g for a predefined time bound
   Do until pigeon reaches food
   i. Evaluate the distance D for all pigeons in flock and food.
   $$D_{gA} = \sqrt{(x_g - x_A)^2 + (y_g - y_A)^2}$$
   ii. $min_D =$ minimum of D_{gA}
   iii. Return optimized path = min_D
   iv. $f(x_i) = f(x_i) + 1$ for pigeon with minimum distance from food
7. For optimized solution
   a. Evaluate $f(x)$ for pigeon p in the flock
   b. Maximum value of $f(x)$ is returned and it would be leader in flock

IV. RESULT AND DISCUSSIONS

A. Experimental Setup

Human acoustic speech observations are taken from the TIMIT corpus to evaluate the performance of the proposed optimization technique for DNN/HMM model. TIMIT consists of 6300 utterances from the 630 speakers. We used

183 target class labels (61 phones * 3 states/phone). For decoding purpose, a phone tri-gram model is used. After decoding, there 61 phone classes are mapped into 39 useful classes as in [21]. MFCC feature extraction technique is used for extracting the features from raw speech signals. The sliding window size is taken 25-ms with a fixed shift of 10-ms. 13 MFCC features + their first and second order derivatives + energy i.e. 40 observations are supplied as input feature vector. The proposed system is evaluated on MATLAB version 2017a for developing feature extractor module of ASR system. The acoustic module and decoding module have been developed using HTK 3.5 β-2 version toolkit.For neural network training, we used SGD training and PIO isused to optimize the weights of acoustic model.The objective of PIO is to minimize mean square error. The input layer on neural network includes a context window of 15 frames. The input of DNN is divided into40 bands. An experimentis performed on a high performance computer with Intel i7-8core processor; 8GB RAM and Windows 10 as operating system. We used NNSTART tool for neural network configuration on MATLAB environment. PIO is run for 100 iterations with size of 100.

B. Result

Table 1 shows the results for PIO algorithm in PER. The results clearly indicate a significant and persistent reduction in the PER is achieved as compared to SGD training. PIO is performed well and achievedPER of 17.2%. Result clearly indicates that training using PIO is better compared to back-propagationtraining methods for DNN/HMM model. The relative reduction of0.6% is achieved in PERon the same TIMIT dataset. Figure 1shows thetrade-off between PER and number of iterations, which indicates that in starting the PER reduces fast and later it becomes almost constant.

Table 1. Result of SGD and PIO training on TIMIT dataset

Technique	Phoneme Error Rate	
	Training	***Testing Set***
SGD	17.8	18.1
PIO	17	17.2

Table 2: Comparison of PIO trained DNN with Existing Models

Methods	Phone Error Rate
Monophone DBN-DNNs on fbank (8 layers)[22]	20.7%
Monophone mcRBM-DBN-DNNs on fbank (5 layers)[23]	20.5%
DNN with dropout [24]	19.7%
DNN+RNN [25]	18.8%
LSTM-RNN [26]	17.7%
Segmental RNN [27]	17.3%
PIO-trained DNN	17.2%

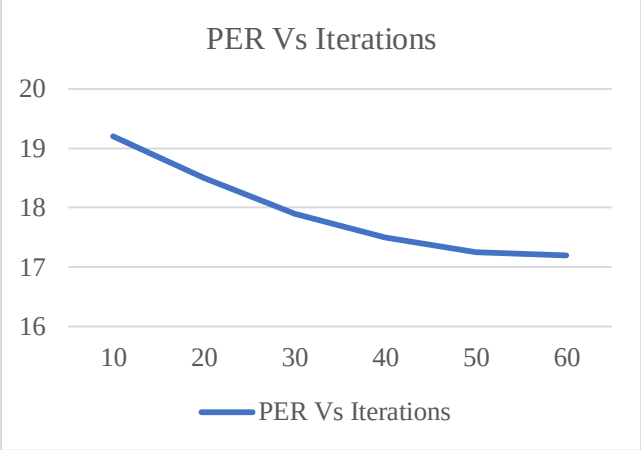

Fig. 1. PERvs.Number of Iterations

V. CONCLUSION

The primary aim of designing speech recognition systems is that it's should be high. There are various methods like improved feature extraction technique, better training algorithms, hybrid acoustic model etc. which have been used for it. Training by weight optimization technique is also growing area in the field of deep neural networks. In this paper, PIO-trained DNN is demonstrated for speech task. The proposed DNN gained 0.6% of relative improvement over segmental RNNs in phoneme recognition. This superiority is achieved because PIO-optimized the weight matrix of DNN and reduces the overfitting of training data.

REFERENCES

1. Pasricha, V. and R. Aggarwal. *Hybrid architecture for robust speech recognition system*. in *Recent Advances and Innovations in Engineering (ICRAIE), 2016 International Conference on*. 2016. IEEE.
2. Passricha, V. and R.K. Aggarwal, *Feature Extraction technique for Hindi speech recognition system.* International Journal of computing and application, 2018. **13**(2): p. 221-229.
3. Davis, S.B. and P. Mermelstein, *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences*, in *Readings in speech recognition*. 1990, Elsevier. p. 65-74.
4. Hermansky, H., *Perceptual linear predictive (PLP) analysis of speech.* the Journal of the Acoustical Society of America, 1990. **87**(4): p. 1738-1752.
5. Deng, L., *Computational models for speech production*, in *Computational models of speech pattern processing*. 1999, Springer. p. 199-213.
6. Hinton, G., et al., *Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups*. IEEE Signal Processing Magazine, 2012. **29**(6): p. 82-97.
7. Huang, X.D., Y. Ariki, and M.A. Jack, *Hidden Markov models for speech recognition*. 1990.

8. Deng, L., et al. *Recent advances in deep learning for speech research at Microsoft*. in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. 2013. IEEE.
9. Duan, H. and P. Qiao, *Pigeon-inspired optimization: a new swarm intelligence optimizer for air robot path planning.* International Journal of Intelligent Computing and Cybernetics, 2014. **7**(1): p. 24-37.
10. Lu, L., et al., *Segmental Recurrent Neural Networks for End-to-End Speech Recognition*, in *Interspeech 2016*. 2016, ISCA.
11. Baum, L.E., et al., *A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains.* The annals of mathematical statistics, 1970. **41**(1): p. 164-171.
12. Jiang, H., *Discriminative training of HMMs for automatic speech recognition: A survey*. Computer Speech & Language, 2010. **24**(4): p. 589-608.
13. Rose, R.C. and D.A. Reynolds. *Text independent speaker identification using automatic acoustic segmentation*. in *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*. 1990. IEEE.
14. Tebelskis, J., *Speech recognition using neural networks*. 1995, Carnegie Mellon University.
15. Yu, D., L. Deng, and G. Dahl. *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*. 2010.
16. Dahl, G.E., et al., *Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition.* IEEE Transactions on audio, speech, and language processing, 2012. **20**(1): p. 30-42.
17. Seide, F., G. Li, and D. Yu. *Conversational speech transcription using context-dependent deep neural networks*. in *Twelfth Annual Conference of the International Speech Communication Association*. 2011.
18. Seide, F., et al. *Feature engineering in context-dependent deep neural networks for conversational speech transcription*. in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. 2011. IEEE.
19. Rabiner, L.R., *A tutorial on hidden Markov models and selected applications in speech recognition.* Proceedings of the IEEE, 1989. **77**(2): p. 257-286.
20. Rabiner, L.R. and B.-H. Juang, *Fundamentals of speech recognition*. Vol. 14. 1993: PTR Prentice Hall Englewood Cliffs.
21. Lee, K.-F. and H.-W. Hon, *Speaker-independent phone recognition using hidden Markov models.* IEEE Transactions on Acoustics, Speech, and Signal Processing, 1989. **37**(11): p. 1641-1648.
22. Mohamed, A.R., G.E. Dahl, and G. Hinton, *Acoustic modeling using deep belief networks.* IEEE transactions on Audio, Speech and Language Processing, 2012. **20**(1): p. 14-22.
23. Dahl, G., A.R. Mohamed, and G.E. Hinton. *Phone recognition with the mean-covariance restricted Boltzmann machine*. in *Advances in Neural Information Processing Systems*. 2010.
24. Hinton, G.E., et al. *Improving neural networks by preventing co-adaptation of feature detectors*. in *arXiv preprint arXiv:1207.0580*. 2012.
25. Deng, L. and J. Chen. *Sequence classification using the high-level features extracted from deep neural networks*. in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2014. IEEE.
26. Graves, A., A.R. Mohamed, and G. Hinton. *Speech recognition with deep recurrent neural networks*. in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2013. IEEE.
27. Lu, L., et al. *Segmental recurrent neural networks for end-to-end speech recognition*. in *arXiv preprint arXiv:1603.00223*. 2016.