Residual Neural Networks for Speech Recognition

Hari Krishna Vydana, Anil Kumar Vuppala
Speech and Vision Laboratory, LTRC, KCIS
International Institute of Information Technology, Hyderabad, India
hari.vydana@research.iiit.ac.in, anil.vuppala@iiit.ac.in

***Abstract*—Recent developments in deep learning methods have greatly influenced the performances of speech recognition systems. In a Hidden Markov model-Deep neural network (HMM-DNN) based speech recognition system, DNNs have been employed to model senones (context dependent states of HMM), where HMMs capture the temporal relations among senones. Due to the use of more deeper networks significant improvement in the performances has been observed and developing deep learning methods to train more deeper architectures has gained a lot of scientific interest. Optimizing a deeper network is more complex task than to optimize a less deeper network, but recently residual network have exhibited a capability to train a very deep neural network architectures and are not prone to vanishing/exploding gradient problems. In this work, the effectiveness of residual networks have been explored for of speech recognition. Along with the depth of the residual network, the criticality of width of the residual network has also been studied. It has been observed that at higher depth, width of the networks is also a crucial parameter for attaining significant improvements. A 14-hour subset of WSJ corpus is used for training the speech recognition systems, it has been observed that the residual networks have shown much ease in convergence even with a depth much higher than the deep neural network. In this work, using residual networks an absolute reduction of 0.4 in WER error rates (8% reduction in the relative error) is attained compared to the best performing deep neural network.**

I. Introduction

Traditional speech recognition systems has majorly relied on the paradigm of "beads of the string", which considers words to be composed of phones and phones are composed by sub-phone acoustic units (senones). In a Hidden Markov model-Gaussian Mixture Models (HMM-GMM) based speech recognition system, senones are modeled by GMMs as the states of HMM [1] and HMMs model the temporal relations among senones. Though GMMs have many advantageous properties such as faster convergence and capability to model any probability distribution with enough number of components, but GMMs cannot learn the data on the nonlinear manifolds [2]. To alleviate this problem, neural networks with single hidden layer are used to model the states of HMM (ie., senonoes) instead of a GMM. Recently DNNs are employed to model the senones has shown a significant improvement in speech recognition systems [2].

The development of deep learning methodologies have greatly influenced the performances of speech recognition systems. Deep learning methodologies majorly aim learning the feature hierarchies in which lower level features are composed to form a higher level representations. To learn better representations a more deeper network has to be trained. Superiority of networks with increased depth has been studied for various tasks such as speech recognition, language processing and AI are described in [3]. Training and optimizing a deeper network is more complicated than training and optimization of a shallow network [4]. The difficulties in training more deeper architectures with the sigmoid activation units and random initializations have been studied in [5]. Recent developments in deep learning methodologies can be majorly consolidated as development on learning methodologies, initializations and activation functions for training more deeper architectures. Though a significant amount of progress has been achieved in reducing the effect of exploding/vanishing gradients by the use of activation functions such as ReLU [6], PReLU [7] and normalized initializations [7] and normalizations like Batch-normalization [8], but optimizing a neural network with very deep architecture is an open problem and there have been many attempts to train a deep neural networks with plain stochastic gradient decent (SGD). Learning strategies like curriculum learning, continuation methods [9], mollifying networks [10] and use of noisy activation functions [11] have been studied to aid the optimization of a highly non convex objective functions.

Initial attempts to train the deep networks were studied in deep supervision [12]. In deep supervision, an auxiliary loss is forked in the intermediate layers, to provide a short path for back-propagating the gradients, the forked layers have two gradients i.e., from main loss and auxiliary loss. Despite the better performance of deep supervision, irrelevance of auxiliary loss at test time, mismatch between the training and testing objective functions is a major drawback. Recent architectures called Highway networks have been successful in training neural network architectures with arbitrary depth using SGD. Highway networks are characterized by pathways which allow unimpeded information flow across the layers of a network known as highways of information [4]. In a highway network a data-driven gated mechanism is employed to control the pathways of information and in a way they decide whether the layer should learn the mapping function or its residual counter part. Though the highway networks have provided capability to train an arbitrary depth architecture but improvements in the performance were not significantly high even for at 100 layer depth. Further studies have shown that the replacing the data driven gating mechanism of highway networks with an identity mapping has given rise to a new class of networks called residual networks (Resnets) [13]. The residual networks are enriched with advantageous properties such as capability to train networks of any depth with SGD.

ISBN 978-0-9928626-7-1

Unlike Highway networks strict use of the identity mapping across the information highways makes the residual network to learn only residual mappings and network of any depth is not prone to vanishing and exploding gradients [14]. The effectiveness of identity mapping for a residual network and their ease in training has been studied in [15]. Unlike the Deep supervision where axillary loss is forked at intermediate layer to reduce the effective depth while training the networks, in a residual network the identity mappings provide a better way of reducing the effective depth. Due to this mechanism of reducing the effective depth, a residual neural network has shown better ease in convergence and better generalization. This motivated us to study the influence of residual networks and their effectiveness for the task of speech recognition.

Apart from increasing the depth of the network, widening the networks have also shown better performance on image classification tasks [16]. In this study, the effectiveness of increasing width of the network along with with residual connections has been explored for the task of speech recognition. Unlike the other classification tasks speech recognition task has to handle many variabilities speech, speaker and emotion variabilities that are naturally expected to exist in speech data, so the classifier should be capable of exhibiting better generalization to these variabilities. Apart from the variabilities the number of classes in a speech recognition system is huge ie., it has to classify input frames to senones which are around 2000 even after state tieing [2]. During the study, the residual networks (Resnets) are employed to model the senones ie., probability of the present frame belonging to a senone is used as the emission probability of HMM.

The main focus of the study is presented below i.e.,

- Exploring the use of residual networks for the task of speech recognition
- Analyzing the effect of widening the residual blocks for the task of speech recognition
- Comparing the performance and convergence properties of residual and wide-residual networks for speech recognition.

The remaining paper is organized as follows: The database used and the architecture of residual network used in this study is described in section 2 and performance of the speech recognition systems developed using deep network (HMM-8-DNN) and a residual networks(HMM-Resnet) are compared in section 3. Section 4 gives the conclusion and future scope of the work.

II. Speech Recognition Framework used in this study

A. Database

Speech data from Wall Street Journal corpus (WSJ) [17] has been used during the study. In this study, a 14 hrs subset of WSJ corpus (si284-set) for training the speech recognition systems, eval-92 and dev-93 sets are used as test sets. Alignments for the training data are obtained form HMM-GMM based tri-phone speech recognition system, and these are used for training the deep neural networks.

B. Feature Extraction

Mel-frequency cepstral coefficients (MFCC) extracted form speech signal are spliced over 9 frames (± 4) in time to form a 117 dimensional feature vector. A linear discriminant analysis (LDA) is used to make this 117 dimensional input vector to a 40 dimensional vector and a feature-space maximum likelihood linear regression (fMLLR) transform is used for speaker variability normalization. The speaker normalized 40 dimensional vector is spliced in time over 11 frames (± 5) resulting a 440 dimensional feature vector. This entire feature extraction is performed using kaldi-pdnn toolkit [18].

C. Architecture

A deep neural network with six hidden layers comprising of ReLU units i.e., (440R-1024R-1024R-1024R-1024R-1024R-1024R-1991S) is used as a baseline system and this network is termed as 8-DNN in this study. The categorical entropy of the outputs is used as the loss function. ADADELTA [19] is used as an optimizer. The dropout of 0.1 is used for all the hidden layers [20]. During the work, a continuous increase in the validation loss for five successive epochs is considered as an early stopping criterion.

D. Residual Neural Network architecture

In this work, a HMM-Resnet architecture for speech recognition has been explored. The posterior probability obtained for a frame of speech using a residual network (resnet) is used as the emission probability of HMMs. If $H(x)$ is the mapping learned by feed forward deep neural network where x is input, then the network can also learn $H(x)-x$ mapping, but with a different ease of learning [13]. Thus the residual function ($F(x)$) thus becomes $F(x) := H(x)-x$, the residual network is implemented just as any deep neural network with a constraint $H(x) := F(x) + x$.

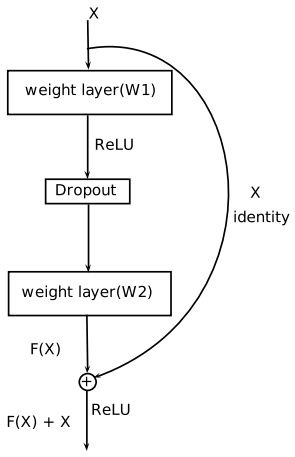

Fig. 2: Residual block used in this study.

In this work, the residual blocks presented in Fig. 2 is termed as Res. In the residual block (Res) the first weight layer contains W1-ReLU units and the second weight layer contains W2-ReLU units. During the course of study, the

ISBN 978-0-9928626-7-1 © EURASIP 2017

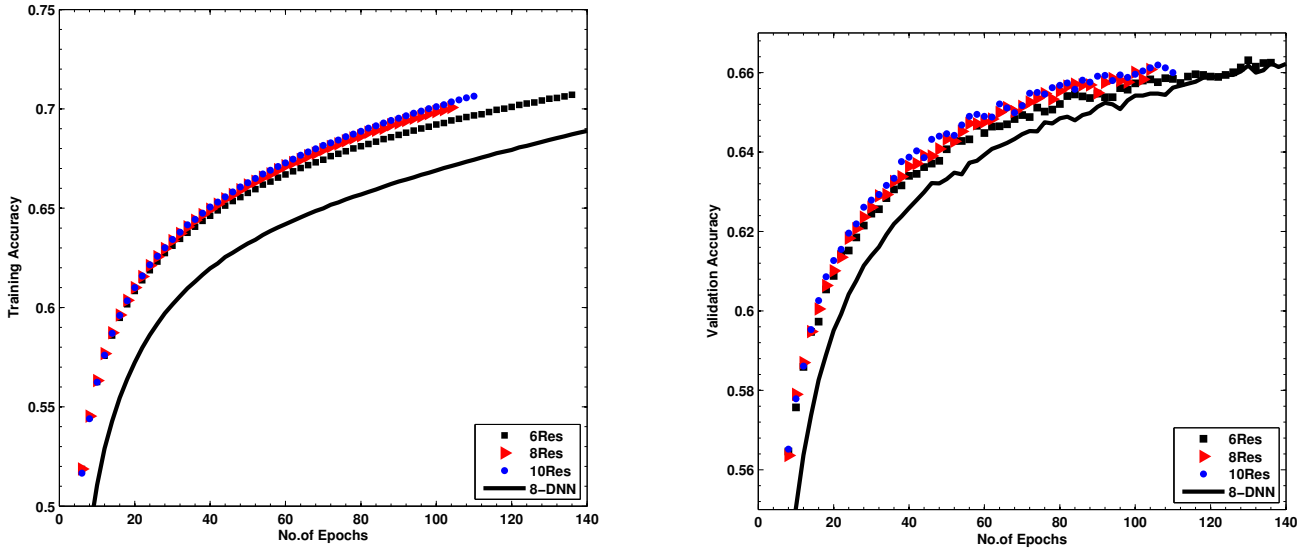

Fig. 1: Comparing the performance of speech recognition systems developed using 8-DNN and residual network in terms of frame error rate. Figures are generated from WSJ corpus mentioned in subsection II-A

second weight layer W2 always has the fixed number of units which is equal to the dimension of input (440), so that the output of weight layer W2 is directly added to input with out any zero-padding. A dropout regularization is used in-spite of Batch normalization, a dropout factor of 0.1 is used with all the residual blocks.

III. Residual Neural Network For Speech Recognition

In this work, HMM-Resnet hybrid systems have been explored for speech recognition, in which a residual network is employed for modeling the senones and the temporal sequence of senones is modeled using a HMM model. The residual network architectures are formed by stacking the residual blocks shown in Fig.2 with weight layers W1, W2 comprising of 440 units. During the study, 5% of the training set is held-out as validation set and frame error rates are computed over that set. During the study, residual network architecture formed by stacking 'n' residual blocks followed by a softmax is termed as nRes architecture. The performance of speech recognition systems developed using residual networks of varying depth (6Res, 8Res and 10Res) is presented in terms of frame error rate in Fig.1.

From Fig.1.(a), it can be observed that residual networks has shown good ease in training. The performance of deep neural network formed by stacking several fully connected layers (8-DNN) is shown by a solid line. From Fig.1.(b) the performance of residual networks is slightly better than the 8-DNN. The performance of speech recognition systems developed using residual networks in terms of word error rate (WER) is presented in Table.I. The WERs are reported on dev-93 and eval-92 sets. As the dev-93 set shares the same data environment, vocabulary size as training set a over-fit model appears to perform better, so in this study the performance on eval-92 is considered as a measure of networks generalization capability.

TABLE I: Performance of HMM-DNN and HMM-Resnet(6Res, 8Res and 10Res) speech recognition systems in terms of Word error rate (WER).

Test sets	stacked network	6Res	8Res	10Res
eval-92	5.19	5.07	**4.86**	5.14
dev-93	8.72	8.68	8.56	8.51

Row 1 of Table.I are the various network architectures and row 2, 3 are the WERs of various speech recognition systems on eval-92 and dev-93 sets respectively. As the depth of the network is increased from 6Res to 8Res an improvement in the performance can be observed from columns 2, 3 of Table.I. Network with 10Res has exhibited more over-fitting nature compared to the 6Res, 8Res architectures can be noted form Fig.1(a),(b) and the similar nature is also apparent in Table.I. Though the 8-DNN, residual networks (6Res, 8Res and 10Res) has exhibited similar performance on the validation set which can be seen from Figure .1, but the residual networks have exhibited better generalization than the 8-DNN which can be noted in Table.I.

In this study, the criticality of width of the residual layers has been explored and the networks formed by widening the residual layers are termed as wide-residual networks. In this study, the wide-residual networks are designed by stacking the residual blocks presented in Fig.I with weight layer W1 comprising of 1024 units and weight W2 comprises of 440 units. The performance of wide-residual networks in terms of frame error rate is presented in Fig.3.

From Fig.3, it can noted that the width of the networks

ISBN 978-0-9928626-7-1 © EURASIP 2017

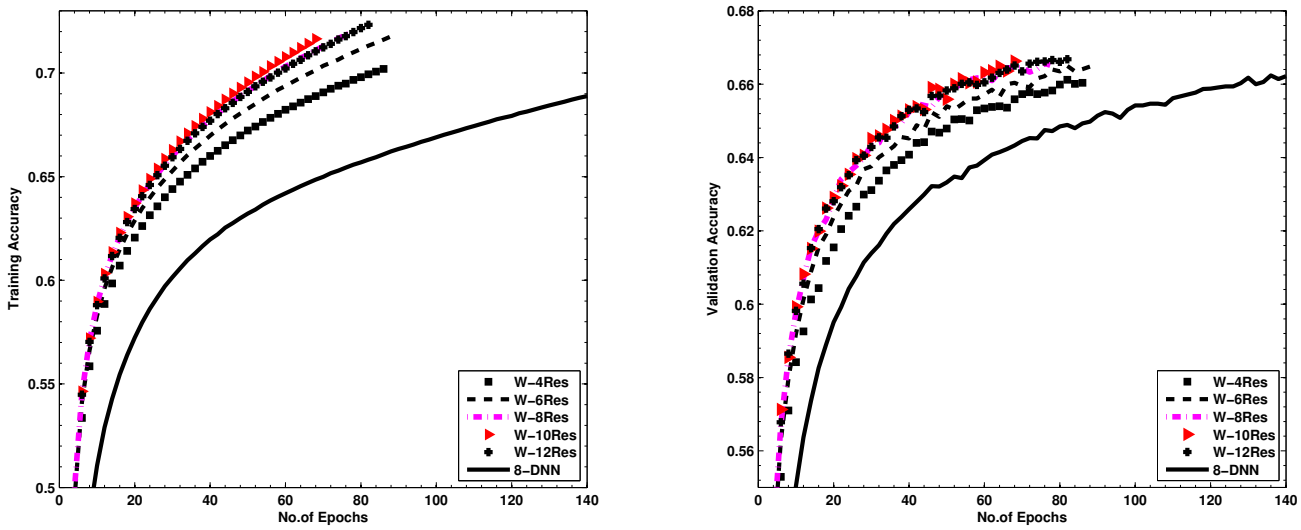

Fig. 3: Performance of speech recognition systems developed using 8-DNN and wide-residual networks in terms of frame error rate. Figures are generated from WSJ corpus mentioned in subsection II-A

has shown better ease in convergence. The width of the residual blocks is also a critical parameter along with the depth of the network, with the increase in the width the performance of wide residual network on validation dataset has significantly improved. The performance of wide-residual networks is significantly higher than residual networks and the 8-DNN. As the depth of wide-residual network is increased from 4 to 10 layers an increase in the performance is observed but as the depth of the network is further increased to 12 layers an over-fit in the model is observed. The performance of speech recognition developed using wide-residual networks is presented in terms of WERs in Table.II.

TABLE II: Performance of wide-residual networks for speech recognition.

Test sets	stacked network	W-4Res	W-6Res	W-8Res	W-10Res	W-12Res
eval-92	5.19	5.12	5.07	4.94	**4.77**	5.05
dev-93	8.72	8.68	8.43	8.60	8.62	8.44

Row 1 of Table.II are the various speech recognition systems developed by varying the depth of the wide-residual network. Row 2, 3 of Table.II are the WERs on eval-92, dev-93 sets respectively. Increasing the depth of the network from 4,10 layers the the performance speech recognition system has increased and at 12 layer depth the networks have exhibited an over-fitting nature and the same can be observed in terms of WER. Though wide-residual networks have shown less generalization capability at lower depths, at higher depths a significant improvements in the performance can be observed. At higher depths i.e., 10 layers the width of the network has shown a significant impact. The residual networks, wide residual networks have shown better generalization properties and an absolute reduction of 0.4% (8% reduction in the relative error) in WER is obtained.

IV. Conclusion & Future Scope

The recent developments in deep learning methodologies have enhanced the performance of speech recognition. In a HMM-DNN based speech recognition, where deep neural networks have been explored to model senones. The superiority of the networks with increased depth has been studied for multiple tasks. Recently there has been a many studies to train deeper networks like highway networks, residual networks. In this study, HMM-Resnet architectures for speech recognition have been explored. With the increased depth, the residual networks have exhibited better generalization and convergence properties. In this study, HMM-Resents have shown superior performance compared to HMM-DNN based speech recognition systems and an absolute reduction of 0.4% in WER or 8% reduction in the relative error is observed. Along with the depth of the network, the criticality of the width of the residual layers has been explored. Increase in width of the residual layers along with depth have aided the convergence. At higher depths increase in the width of the network has attained significant improvement in the performance of speech recognition systems. In the future studies, the other architectural aspects of residual neural networks would be explored for attaining improvements in speech recognition system.

References

[1] M. Ostendorf, "Moving beyond the beads-on-a-stringmodel of speech," in *IEEE Automatic Speech Recognition and Understanding Workshop*, 1999, pp. 79–84.

[2] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

ISBN 978-0-9928626-7-1 © EURASIP 2017

[3] Y. Bengio, "Learning deep architectures for AI," *Foundations and trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.
[4] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," *arXiv preprint arXiv:1505.00387*, 2015.
[5] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249–256.
[6] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks." in *International Conference on Artificial Intelligence and Statistics*, vol. 15, no. 106, 2011, p. 275.
[7] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *International Conference on Computer Vision*, 2015, pp. 1026–1034.
[8] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.
[9] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *International Conference on Machine Learning*. ACM, 2009, pp. 41–48.
[10] C. Gulcehre, M. Moczulski, F. Visin, and Y. Bengio, "Mollifying networks," *arXiv preprint arXiv:1608.04980*, 2016.
[11] C. Gulcehre, M. Moczulski, M. Denil, and Y. Bengio, "Noisy activation functions," *arXiv preprint arXiv:1603.00391*, 2016.
[12] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets." in *International Conference on Artificial Intelligence and Statistics*, vol. 2, no. 3, 2015, p. 6.
[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.
[14] ——, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.
[15] ——, "Identity mappings in deep residual networks," *arXiv preprint arXiv:1603.05027*, 2016.
[16] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.
[17] J. Garofalo, "Wall street journal-based continuous speech recognition (csr) corpus," *Linguistic Data Consortium, Philadelphia*, 1994.
[18] Y. Miao, "Kaldi+ pdnn: building dnn-based asr systems with kaldi and pdnn," *arXiv preprint arXiv:1401.6984*, 2014.
[19] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.
[20] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

ISBN 978-0-9928626-7-1