INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Exploiting Depth and Highway Connections in Convolutional Recurrent Deep Neural Networks for Speech Recognition

Wei-Ning Hsu, Yu Zhang, Ann Lee, James Glass

Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
{wnhsu,yzhang87,annlee,jrg}@csail.mit.edu

Abstract

Deep neural network models have achieved considerable success in a wide range of fields. Several architectures have been proposed to alleviate the vanishing gradient problem, and hence enable training of very deep networks. In the speech recognition area, convolutional neural networks, recurrent neural networks, and fully connected deep neural networks have been shown to be complimentary in their modeling capabilities. Combining all three components, called CLDNN, yields the best performance to date. In this paper, we extend the CLDNN model by introducing a highway connection between LSTM layers, which enables direct information flow from cells of lower layers to cells of upper layers. With this design, we are able to better exploit the advantages of a deeper structure. Experiments on the GALE Chinese Broadcast Conversation/News Speech dataset indicate that our model outperforms all previous models and achieves a new benchmark, which is 22.41% character error rate on the dataset.

Index Terms: speech recognition, recurrent neural network model, convolutional neural network model, highway connection, Mandarin/Chinese speech recognition

1. Introduction

In the past few years, neural network-based (NN) acoustic models have greatly improved automatic speech recognition (ASR) performance over traditional Gaussian mixture models (GMMs) on a variety of tasks [1, 2, 3, 4, 5, 6]. Further improvement has been achieved by using more advanced NN architectures that are specialized to model different aspects of the speech signal. For example, convolutional neural networks (CNNs) are designed to learn translational invariant features, and hence can address speaker normalization issues [7, 8, 9]. Recurrent neural networks (RNNs) are natural for sequence-to-sequence modeling and have demonstrated success in phonetic recognition [10], and ASR [11, 12, 13] due to their ability to learn temporal relationships. Very recently, a model combining the above two networks along with fully-connected deep neural networks (DNNs) called CLDNN has been proposed in [14], and achieved additional gains over any single model, or pairwise model combination.

In the computer vision literature, NN architectures on the order of 19 layers [15] have shown improvement over shallower NN models. However, the best NN architectures reported in the ASR literature are relatively shallow, typically on the order of 3 to 6 layers. The major difficulty with building a very deep NN is the vanishing/exploding gradient problem during training. The gradient issue was addressed with RNNs by the development of the long short-term memory (LSTM) block. Several architectures have also been proposed very recently to enable training of very deep networks [16, 17, 18, 19]. The idea behind these approaches is similar to the LSTM innovation: the introduction of gated linear dependence of memory cells between conventional adjacent layers in the NN model.

In this paper, we first extend the CLDNN model by adding highway connections to memory cells of adjacent LSTM layers, which introduces gated linear dependence. We refer to this model as a Highway CLDNN. The highway connections avoid the vanishing/exploding gradient flow along the NN layers, and hence enable training of deeper NNs. Subsequently, we exploit the power of depth by adding up to 8 LSTM layers, so that the entire NN architecture is in total 11 layers.

The study is conducted on the LDC GALE Chinese Broadcast Conversation/News Speech corpus. Initial experiments on a 120 hour dataset indicate that the Highway CLDNN model outperforms LSTM, Highway LSTM, and CLDNN models that all have 3 LSTM layers. Further improvement is achieved by adding more LSTM layers to the Highway CLDNN model. A similar trend is observed on a larger 500 hour GALE dataset. Compared with the same test set reported in [9], we achieve a new benchmark of 22.41% character error rate on this task.

The rest of paper is organized as follows. In Section 2, we introduce the building blocks of our model, and describe the architecture of the proposed Highway CLDNN. The experimental setup is summarized in Section 3, followed with results and discussion in Section 4. Finally, we conclude in Section 5.

2. Highway Convolutional Recurrent Deep Neural Network

In this section, we first give a brief introduction to the convolutional layers, recurrent layers, and highway connections, that are the essential building blocks in our work. Then we explain how to construct our proposed Highway CLDNN model for speech recognition with these components.

2.1. Convolutional Layers

Convolutional neural networks (CNNs), composed of at least one convolutional layer, have shown improvement over traditional fully-connected deep neural networks on many ASR tasks [7, 8, 9]. Unlike fully connected layers, convolutional layers take into account the input topology, and are designed to reduce translational variance by forcing weight sharing and applying mean/max pooling afterward.

Let input feature $\mathbf{x} \in \mathbb{R}^{T_{\mathbf{x}} \times F_{\mathbf{x}}}$ be a two dimensional matrix,

http://dx.doi.org/10.21437/Interspeech.2016-515

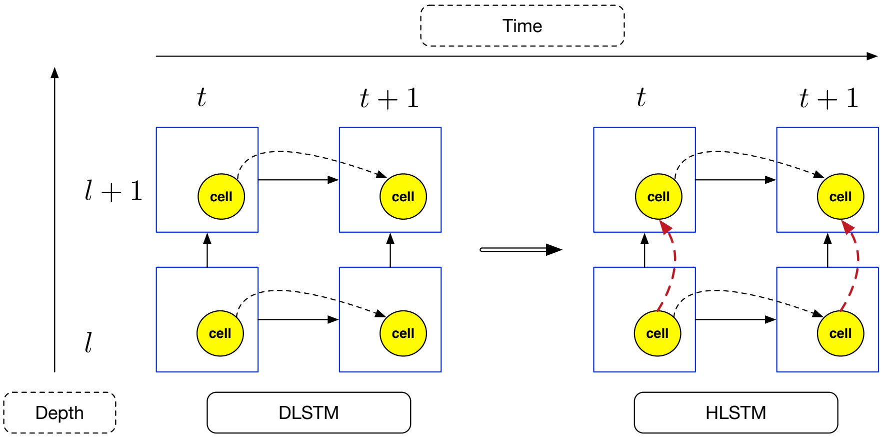

Figure 1: Comparison of Deep Long Short-Term Memory (DLSTM) RNN and Highway Long Short-Term Memory (HLSTM) RNN.

where $T_{\mathbf{x}}$ denotes the context window width and $F_{\mathbf{x}}$ denotes the number of frequency bands. Suppose there are K kernels with weight $\mathbf{W}_1, \mathbf{W}_2, \cdots, \mathbf{W}_K$ and bias $b_1, b_2, \cdots, b_K$. We use k to index kernels and the k-th kernel $\mathbf{W}_k \in \mathbb{R}^{T_k \times F_k}$. The activation (also called a feature map) of the k-th kernel centered at the (t, f)-position of the input feature is

$$h_{k_{t,f}} = \theta(\sum_{i=1}^{T_k}\sum_{j=1}^{F_k} x_{i+(t-\lceil \frac{T_k}{2} \rceil),j+(t-\lceil \frac{F_k}{2} \rceil)} W_{k_{i,j}} + b_k), \quad (1)$$

where θ is the activation function, which we set to be rectified linear units here. Note that we set $x_{i',j'} = 0$ if i', j' exceeds the boundary.

2.2. Recurrent Layers

Unlike feed-forward neural networks, recurrent neural networks (RNNs) contain feedback loops that feed activations not only to the next layer, but also to the current layer as input at the next time step. This architecture enables modeling of temporal relationships within a context window of dynamically changing size, which is highly desirable since contextual information plays an important role in acoustic modeling, and feed-forward NNs are limited to considering a fixed-size context window.

In practice, simple RNNs usually suffer from the vanishing/exploding gradient problem, when training with back-propagation through time (BPTT). To address this issue, long short-term memory (LSTM) blocks, as shown in the left part of Figure 1 were proposed in [20], which introduced a *gated linear dependence* (the black dashed arrows) between memory states of two consecutive time steps. An LSTM block is composed of an array of memory cells $\mathbf{c}$ as well as three gates: $\mathbf{i}$, $\mathbf{f}$, and $\mathbf{o}$, used to control information flow. They are defined as follows:

$$\mathbf{i}_t = \sigma(\mathbf{W}_{xi}\mathbf{x}_t + \mathbf{W}_{hi}\mathbf{h}_{t-1} + \mathbf{W}_{ci}\mathbf{c}_{t-1} + \mathbf{b}_i) \quad (2)$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_{xf}\mathbf{x}_t + \mathbf{W}_{hf}\mathbf{h}_{t-1} + \mathbf{W}_{cf}\mathbf{c}_{t-1} + \mathbf{b}_f) \quad (3)$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_{xo}\mathbf{x}_t + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{W}_{co}\mathbf{c}_{t-1} + \mathbf{b}_o) \quad (4)$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tanh(\mathbf{W}_{xc}\mathbf{x}_t + \mathbf{W}_{hc}\mathbf{h}_{t-1} + \mathbf{b}_c) \quad (5)$$

$$\mathbf{h}_t = \mathbf{W'}_{proj}(\mathbf{o}_t \odot \tanh(\mathbf{c}_t)) \quad (6)$$

where $\mathbf{x}_t$, $\mathbf{c}_t$, and $\mathbf{h}_t$ are input feature, cell state, and cell output respectively at time t. $\mathbf{W}_*$ and $\mathbf{b}_*$ are weight matrices and bias vectors connecting different gates. $\odot$ denotes an element-wise product. Note that the number of cells in each LSTM is set to one, and therefore $\mathbf{W}_c*$ are diagonal matrices. $\mathbf{W'}_{proj}$ is the projection matrix as proposed in [11].

2.3. Highway Connection

The Highway LSTM (HLSTM) RNN was proposed in [12]. As shown in Figure 1, it has a direct gated connection (the red dashed arrows) between the memory cells $\mathbf{c}_t^l$ in the lower layer l and the memory cells $\mathbf{c}_t^{l+1}$ in the upper layer $l+1$. The carry gate controls how much information can flow from the lower-layer cells directly to the upper-layer cells. The gate function at layer $l+1$ at time t is

$$\mathbf{d}_t^{l+1} = \sigma(\mathbf{W}_{xd}^{l+1}\mathbf{x}_t^{l+1} + \mathbf{W}_{cd}^{l+1}\mathbf{c}_{t-1}^{l+1} + \mathbf{W}_{ld}^{l+1}\mathbf{c}_t^l + \mathbf{b}_d^{l+1}), \quad (7)$$

where $\mathbf{b}_d^{l+1}$ is a bias term, $\mathbf{W}_{xd}^{l+1}$ is the weight matrix connecting the carry gate to the input of this layer. $\mathbf{W}_{cd}^{l+1}$ is a diagonal weight matrix from the past cell state to the carry gate in the current layer. $\mathbf{W}_{ld}^{l+1}$ is a diagonal weight matrix connecting the lower layer memory cell to the carry gate. $\mathbf{d}_t^{l+1}$ is the carry gate vector at layer $l+1$ at time t.

Using the carry gate, an HLSTM RNN computes the cell state at layer $l+1$ according to

$$\begin{aligned}\mathbf{c}_t^{l+1} &= \mathbf{d}_t^{l+1} \odot \mathbf{c}_t^l + \mathbf{f}_t^{l+1} \odot \mathbf{c}_{t-1}^{l+1} \\ &+ \mathbf{i}_t^{l+1} \odot \tanh(\mathbf{W}_{xc}^{l+1}\mathbf{x}_t^{l+1} + \mathbf{W}_{hc}^{l+1}\mathbf{m}_{t-1}^{l+1} + \mathbf{b}_c), \quad (8)\end{aligned}$$

while all other equations are the same as that in the standard LSTM RNNs as described in Eq. (2),(3),(4), and (6).

Thus, depending on the output of the carry gates, the highway connection can smoothly vary its behavior between that of a plain LSTM layer or simply pass on its cell memory from the previous layer. The highway connection between cells in different layers makes the influence from cells in one layer to the other more direct, and can alleviate the gradient vanishing problem when training deeper LSTM RNNs.

2.4. Highway CLDNN

Our Highway CLDNN structure followed the design in [14]. Figure 2 illustrates the architecture of our model. Since recurrent layers are able to capture temporal relationships, at each time step we pass frame x_t without context as input to the network. We use a filter bank feature along with a pitch feature to represent each frame x_t.

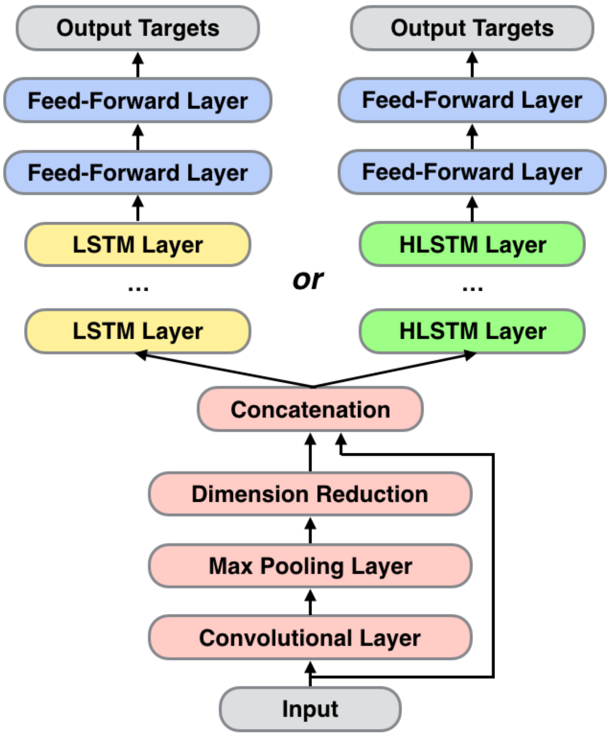

Figure 2: (H)CLDNN. The left path illustrates CLDNN, and the right path illustrates our proposed Highway CLDNN.

To address the speaker normalization issue, input features are first passed to one convolutional layer. This layer generates 256 feature maps, with kernels of consisting of a 1-by-8 receptive field. Non-overlapping max-pooling of both pooling window and stride 3 is applied immediately after this convolutional layer. As the output dimensionality is still very large (i.e. $\lceil 83/3 \rceil \times 256$), a projection layer to 256 dimensions is added on top of the max-pooling layer.

Subsequently, to model temporal relationships, output from the projection layer is then fed into a recurrent layer. In addition, following the suggestion in [14], we also pass the original input feature x_t to the recurrent layer, providing input representations from different levels.

Since speech signals possess information at different time scales [21], we would like the recurrent layers to capture temporal relationships at different scales as well, which deep RNNs have been argued to be able to learn [22]. Therefore, we use Highway LSTMs for recurrent layers in order to utilize the deep recurrent structure. Specifically, 5 layers of LSTM are applied with 1024 cells and a 512 dimensional projection at each layer.

Finally, we feed output from the last recurrent layer into fully-connected feed-forward layers, which provides better discrimination to the output targets. Two layers of 1024 hidden units each, with rectified linear activation functions, are applied.

3. Experiment Setup

3.1. Dataset

Our initial experiments to study the effect of highway connections and neural network depth are conducted on the GALE Phase 2 Chinese Broadcast Conversation Speech (LDC2013S04) corpus, which is approximately 120 hours. We randomly select 9 hours of speech as an evaluation set. Further experiments to verify the results are performed on a larger 500 hour corpus. In addition to the GALE Phase 2 Chinese Broadcast Conversation Speech, GALE Phase 3 Chinese Broadcast Conversation Speech Part 1 (LDC2014S09), Part 2 (LDC2015S06), and GALE Phase 2 Chinese Broadcast News Speech (LDC2013S08) are included. Here we use the same 3-hour evaluation set as in [9] for comparison. Evaluation sets in both experiments are held-out from their training sets.

3.2. Model Setup and Training

We use Kaldi [23] for feature extraction, decoding, and training of an initial HMM-GMM model. A maximum likelihood-criterion context-dependent speaker-adapted acoustic model is trained with a standard Kaldi recipe (`tri3b`). Forced alignment is performed to generate labels for neural network training.

The Computational Network Toolkit (CNTK) [24] is used for neural network training. As [14] suggests, we apply uniform random weight initialization for all layers without either generative or discriminative pretraining [1]. All neural networks are trained with a cross-entropy (CE) criterion, using truncated back-propagation-through-time (BPTT) [25] to optimize. Each BPTT segments contains 20 frames, and each mini-batch contains 40 utterances. No momentum is used for the first epoch and a momentum of 0.9 is used for subsequent epochs [26]. Ten percent of training data is held out as a validation set, which is used to control the learning rate. The learning rate is halved when no gain is observed after an epoch.

The input features for all models are computed every 10ms, and consist of 80 dimensional log Mel filterbank features, with an additional 3 dimensional pitch feature, as [9] suggests, since Chinese is a tonal language. The output targets are 4193 context-dependent states.

4. Results

4.1. Baseline Models

Here we consider three baseline models: (1) LSTM, (2) Highway LSTM, and (3) CLDNN. For the first two models, we chose the best architectures trained with the CE criterion reported in [12]: each hidden layer consists of 1024 memory cells together with a 512-node projection layer. A three-layer LSTM and Highway LSTM were shown to achieve the best performance. The architecture used for the CLDNN is similar to that proposed in [14]. Specifically, we adopt the design of a convolutional layer in [27], and choose LSTM parameters to be the same as the other two models.

The character error rate (CER) of all baseline models are reported in Table 1. Consistent with [14], the CLDNN achieves a 5% relative improvement over the LSTM. However, the Highway LSTM reaches the lowest CER among the three.

4.2. Deeper Neural Models

We begin with investigating the effect of increasing the depth of the network on all three baseline models. Table 1 summarizes the results of deeper models compared with the baselines. When increasing the number of layers, neural networks usually suffer from the vanishing gradient problem, as we can observe from the behavior of the LSTM model.

As stated in [12, 16], the highway connection between LSTM layers can alleviate the vanishing gradient problem, and hence enable the capability for training deeper models. Here a 1.6% relative improvement is observed with the Highway LSTM model, demonstrating that we can actually have more layers than the 3-layer model in [12].

Surprisingly, a deeper CLDNN model also achieves a slight improvement over the original CLDNN model. We attribute this phenomenon to not fully exploiting the advantage of the depth in the original design. In summary, we observe that deeper architectures help further reduce CER with Highway LSTM and CLDNN, and the relative performance among the three models remains consistent when adding layers.

Model	CER(%)	
	3 LSTM Layers	5 LSTM Layers
LSTM	31.42	31.46
CLDNN	29.79	29.48
Highway LSTM	29.70	**29.21**

Table 1: Character error rate comparison of deeper models and baselines trained on 120hr set.

4.3. Highway CLDNN

Next we examine the effect of adding highway connections between LSTM layers to the original CLDNN model, which is our proposed Highway CLDNN model. To separate the effect of highway connection and depth, only models of 3 LSTM layers are compared here.

Table 2 indicates that an additional 2% relative gain is observed by introducing gated linear dependence between LSTM layers, which demonstrates the effectiveness of our proposed model, and opens up the possibility of deeper convolutional recurrent deep neural networks. It is worth noticing that the relative gain from adding highway connections is smaller on the CLDNN model than on the LSTM model. We conjecture that highway connections in recurrent layers help more with input representations modeled at lower levels.

Model	CER(%)
	3 LSTM Layers
LSTM	31.42
CLDNN	29.79
Highway LSTM	29.70
Highway CLDNN	**29.23**

Table 2: Character error rate comparison of Highway CLDNN and baselines trained on 120hr set.

4.4. Deeper Highway CLDNN

The previous two sections demonstrated the advantage of deeper models and highway connections respectively. In this section, we exploit both strategies jointly and verify the result on both the 120 hour dataset, as well as the 500 hour dataset.

Table 3 shows the results on the 120 hour dataset. By combining both strategies, a 5-LSTM-layer Highway CLDNN achieves the best CER performance among all models. Note that it is possible to further reduce the error rate by adding more layers, but since the improvement from adding more layers is relatively marginal here, we will only discover this possibility on the larger data set.

Model	CER(%)	
	3 LSTM Layers	5 LSTM Layers
Highway LSTM	29.70	29.21
Highway CLDNN	29.23	**29.12**

Table 3: Character error rate comparison of highway models trained on 120hr set.

A similar trend can be observed in Table 4, which shows the CER results on the 500 hour dataset. The deeper Highway CLDNN also outperforms all other models. It is worth noting that the relative improvement of adding layers is larger in the 500 hour dataset than in 120 hour dataset. Specifically, the relative gain is 3% for Highway LSTM and 1% on Highway CLDNN for 500-hours, while 1.6% on Highway LSTM and 0.4% on Highway CLDNN for 120-hour's. This suggests that deeper models are suitable for larger datasets.

Model	CER(%)	
	3 LSTM Layers	5 LSTM Layers
Highway LSTM	23.33	22.63
Highway CLDNN	22.68	**22.45**

Table 4: Character error rate comparison of highway models trained on 500hr set.

As suggested previously, one would be interested in pushing the limit of model depth for ASR. Here we constructed a Highway CLDNN model with 8 LSTM layers, which is in total 11 hidden layers, and shows the results in Table 5. Instead of deteriorating the performance due to vanishing gradient problem that very deep models usually encounter, it nevertheless obtains a slight gain from increasing depth. This again proves the effectiveness of highway connections between layers, providing higher flexibility for choosing the number of layers in model.

Model	CER(%)		
	3 LSTM	5 LSTM	8 LSTM
Highway CLDNN	22.68	22.45	**22.41**

Table 5: Character error rate comparison of Highway CLDNN model of different numbers of layers trained on 500hr set.

5. Conclusion

In this paper, we present a comprehensive study of depth and highway connection in recurrent and convolutional recurrent deep neural networks. By exploiting the advantage of both strategies, we proposed a novel architecture called Highway CLDNN, which resolves the vanishing gradient problem when training very deep networks by introducing gated linear dependence of cells between layers. We tested our Highway CLDNN on different sized Chinese speech recognition tasks. Experimental results show that our Highway CLDNN outperforms all previous models and achieves a new benchmark on the GALE Phase 2 Broadcast Conversation Speech corpus. To the best of our knowledge, the best character error rate reported on the same evaluation set is 26.01% in [9] using a CNN model.

For future work, we plan to investigate the framework for alternative "highway block", such as residual networks [28] and grid LSTMs[17]. We also would like to look into much deeper models in a larger task.

6. References

[1] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE, 2011, pp. 24–29.

[2] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks." in *Interspeech*, 2011, pp. 437–440.

[3] 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," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[4] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7398–7402.

[5] S. M. Siniscalchi, D. Yu, L. Deng, and C.-H. Lee, "Speech recognition using long-span temporal patterns in a deep network model," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 201–204, 2013.

[6] B. Li and K. C. Sim, "Modeling long temporal contexts for robust DNN-based speech recognition," in *Interspeech*, 2014, pp. 353–357.

[7] T. N. Sainath, A.-r. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8614–8618.

[8] O. Abdel-Hamid, L. Deng, and D. Yu, "Exploring convolutional neural network structures and optimization techniques for speech recognition." in *Interspeech*, 2013, pp. 3366–3370.

[9] X. Hu, X. Lu, and C. Hori, "Mandarin speech recognition using convolution neural network with augmented tone features," in *Chinese Spoken Language Processing (ISCSLP), 2014 9th International Symposium on*. IEEE, 2014, pp. 15–18.

[10] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol. 18, no. 5, pp. 602–610, 2005.

[11] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling." in *Interspeech*, 2014, pp. 338–342.

[12] Y. Zhang, G. Chen, D. Yu, K. Yao, S. Khudanpur, and J. Glass, "Highway long short-term memory RNNs for distant speech recognition," *arXiv preprint arXiv:1510.08983*, 2015.

[13] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*. ACM, 2006, pp. 369–376.

[14] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4580–4584.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[16] K. Yao, T. Cohn, K. Vylomova, K. Duh, and C. Dyer, "Depth-gated LSTM," *arXiv preprint arXiv:1508.03790*, 2015.

[17] N. Kalchbrenner, I. Danihelka, and A. Graves, "Grid long short-term memory," *arXiv preprint arXiv:1507.01526*, 2015.

[18] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," *arXiv preprint arXiv:1505.00387*, 2015.

[19] ——, "Training very deep networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 2368–2376.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[21] S.-L. Wu, B. E. Kingsbury, N. Morgan, and S. Greenberg, "Incorporating information from syllable-length time scales into automatic speech recognition," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 2. IEEE, 1998, pp. 721–724.

[22] M. Hermans and B. Schrauwen, "Training and analysing deep recurrent neural networks," in *Advances in Neural Information Processing Systems*, 2013, pp. 190–198.

[23] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[24] D. Yu, A. Eversole, M. Seltzer, K. Yao, Z. Huang, B. Guenter, O. Kuchaiev, Y. Zhang, F. Seide, H. Wang *et al.*, "An introduction to computational networks and the computational network toolkit," Tech. Rep. MSR, Microsoft Research, 2014, http://codebox/cntk, Tech. Rep., 2014.

[25] R. J. Williams and J. Peng, "An efficient gradient-based algorithm for on-line training of recurrent network trajectories," *Neural computation*, vol. 2, no. 4, pp. 490–501, 1990.

[26] Y. Zhang, D. Yu, M. L. Seltzer, and J. Droppo, "Speech recognition with prediction-adaptation-correction recurrent neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5004–5008.

[27] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform cldnns," in *Interspeech*, 2015.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.