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

Future Context Attention for Unidirectional LSTM Based Acoustic Model

Jian Tang[1], *Shiliang Zhang*[1], *Si Wei*[2], *Li-Rong Dai*[1]

[1]National Engineering Laboratory for Speech and Language Information Processing
University of Science and Technology of China, Hefei, Anhui, P. R. China
[2]iFLYTEK Research, iFLYTEK Co., Ltd.

{enjtang,zsl2008}@mail.ustc.edu.cn, siwei@iflytek.com, lrdai@ustc.edu.cn

Abstract

Recently, feedforward sequential memory networks (FSMN) has shown strong ability to model past and future long-term dependency in speech signals without using recurrent feedback, and has achieved better performance than BLSTM in acoustic modeling. However, the encoding coefficients in FSMN is context-independent while context-dependent weights are commonly supposed to be more reasonable in acoustic modeling. In this paper, we propose a novel architecture called attention-based LSTM, which employs context-dependent scores or context-dependent weights to encode temporal future context information with the help of a kind of attention mechanism for unidirectional LSTM based acoustic model. Preliminary experimental results on TIMIT corpus have shown that the proposed attention-based LSTM achieves a phone error rate (PER) of 20.8% while PER is 20.1% for BLSTM. We have also presented a lot of experiments to evaluate different context attention methods.

Index Terms: attention, LSTM, future context, speech recognition

1. Introduction

While it has been around for more than two decades that speech recognition systems employed recurrent and feedforward neural networks, it is only recently that they have displaced Gaussian mixture models (GMMs) as the state-of-the-art acoustic model[1, 2, 3, 4]. More recently, it has been widely reported that recurrent neural network (RNN) architectures, especially those with long short-term memory (LSTM), outperform feedforward deep neural networks (DNN) on large-scale speech recognition tasks[5, 6, 7].

Equipped with self-connected memory cells and three multiplicative gates to store and control the flow of information[8, 9], forward unidirectional LSTM (referred to LSTM thereafter) is powerful in learning past long time-dependencies or context, which is why it is recognized to outperform DNN[5, 6, 10, 11]. Bidirectional LSTM (BLSTM) learns past and future context by processing the input sequence in both the forward and backward directions. Generally, BLSTM can achieve better performance in speech recognition compared with LSTM[10] because it learns past and future context. However, BLSTM is built to operate on an entire sequence or sentence that leads to high time latency, so it faces difficulty for online speech recognition. In contrast to BLSTM, LSTM has no time latency shortcoming. Therefore, it is desired to combine future context to LSTM to make LSTM perform as well as BLSTM with low time latency.

Some studies have addressed the above problem. Target-delay is the simplest way of using future context. It is easy to implement, but its speech recognition performance is not as satisfactory as expected[10]. Feedforward Sequential Memory Networks (FSMNs) are a type of DNN-based model that employs a bidirectional FIR-like structure to learn past and future long-term context[12, 13]. FSMN shows a state-of-the-art speech recognition performance comparable with BLSTM and exhibits the advantage of being trained more easily because it remains a pure feedforward structure[13]. However, the coefficients of FIRs or the encoding weight matrix is fixed after the model is trained and is context-independent. In [14], to address the time latency problem, row convolution is proposed to combine the future context by inserting a special layer between LSTM layers, which also applies context-independent weight matrix to encode future context. From the above discussion, the weight matrix for encoding future context in both FSMN and row convolution are context-independent, while context-dependent (CD) weights are commonly supposed to be more reasonable in acoustic modeling.

In this paper, to address the time latency problem, a novel architecture called attention-based LSTM is proposed, which employs CD scores or CD weights to encode temporal future context information with the help of a type of attention mechanism for a unidirectional LSTM-based acoustic model. The expectation is that the proposed attention-based LSTM performs similarly to BLSTM. To evaluate the future context modeling ability of the proposed attention-based LSTM, preliminary experiments are conducted on TIMIT speech recognition tasks. The experimental results show that the proposed attention-based LSTM achieves a phone error rate (PER) of 20.8%, while PER is 20.1% for BLSTM.

2. Related Work

2.1. FSMN architecture for acoustic model

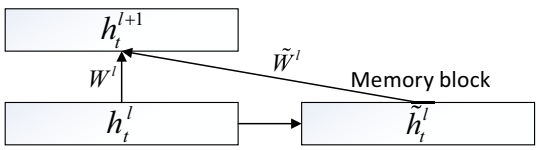

Figure 1: Illustration of an N_1-order feedforward sequential memory network (FSMN) in the l-th layer

FSMN is essentially a standard feedforward fully connected neural network with some memory blocks appended to the hidden layers[13]. In unidirectional FSMN, the memory block, as shown in Figure 1, is used to encode N_1 past activities of the hidden layer into a fixed-size representation.

http://dx.doi.org/10.21437/Interspeech.2016-185

The outputs of the l-th hidden layer for the entire sequence are represented as $\boldsymbol{H^l}$=($\boldsymbol{h_1^l}, \boldsymbol{h_2^l}, ..., \boldsymbol{h_T^l}$), and the encoding weight is denoted as $\boldsymbol{A^l}$=($\boldsymbol{a_1^l}, \boldsymbol{a_2^l}, ..., \boldsymbol{a_{N_1}^l}$). Depending on whether $\boldsymbol{a_i^l}$ is vector or not, FSMNs have the following two variants: i) scalar FSMN (sFSMN); ii) vectorized FSMN (vFSMN). In sFSMNs, all hidden outputs share the same encoding coefficients at the same time-step, while vFSMNs adopt different encoding coefficients for different hidden outputs. This work only introduces the sFSMN process.

In unidirectional sFSMNs, the operations of encoding the past context and calculations of the activation of the units in the next hidden layer are shown in eq.(1) and eq.(2), respectively.

$$\widetilde{\boldsymbol{h}}_{\boldsymbol{t}}^{\boldsymbol{l}} = \sum_{i=1}^{N_1} a_i^l \cdot \boldsymbol{h_{t-i}^l} \tag{1}$$

$$\boldsymbol{h_t^{l+1}} = f(\boldsymbol{W^l h_t^l} + \widetilde{\boldsymbol{W}}^{\boldsymbol{l}} \widetilde{\boldsymbol{h}}_{\boldsymbol{t}}^{\boldsymbol{l}} + \boldsymbol{b^l}) \tag{2}$$

where $\boldsymbol{W^l}$ and $\boldsymbol{b^l}$ represent the standard weight matrix and bias vector for layer l, respectively, and $\widetilde{\boldsymbol{W}}^{\boldsymbol{l}}$ denotes the weight matrix between the memory block and the next layer.

Unidirectional FSMNs only consider the past information in a sequence. By integrating both the context in the past as well as certain future context within a look-ahead window from the current location of the sequence, unidirectional FSMNs can be extended to the following bidirectional versions:

$$\widetilde{\boldsymbol{h}}_{\boldsymbol{t}}^{\boldsymbol{l}} = \sum_{i=0}^{N_1} a_i^l \cdot \boldsymbol{h_{t-i}^l} + \sum_{j=1}^{N_2} c_j^l \cdot \boldsymbol{h_{t+j}^l} \tag{3}$$

where N_1 and N_2 are the window size of past and future context, respectively.

2.2. Attention mechanism for acoustic modeling

An attention-based recurrent sequence generator(ARSG) is a recurrent neural network that stochastically generates an output sequence $\boldsymbol{Y} = (\boldsymbol{y_1}, \boldsymbol{y_2}, ...\boldsymbol{y_T})$ from an input sequence $\boldsymbol{X} = (\boldsymbol{x_1}, \boldsymbol{x_2}, ...\boldsymbol{x_{T_1}})$. In [15, 16, 17, 18, 19], $\boldsymbol{x_t}$ is processed by an encoder that outputs a sequential hidden representation $\boldsymbol{H} = (\boldsymbol{h_1}, \boldsymbol{h_2}, ...\boldsymbol{h_{T_1}})$. The attention mechanism is typically based on end-to-end architecture, and it cooperates with the encoder-decoder architecture. An encoder and downsampling operation can increase the variance of input sequence vector[17, 19], which helps to achieve alignment information easier.

At the t-th time-step an ARSG generates an output $\boldsymbol{y(t)}$ by focusing on the relevant elements of $\boldsymbol{H}$:

$$\boldsymbol{\alpha_t} = softmax(energy(\boldsymbol{s_{t-1}}, \boldsymbol{\alpha_{t-1}}, \boldsymbol{H})) \tag{4}$$

$$\boldsymbol{g_t} = \sum_{j=1}^{L} \alpha_{tj} \boldsymbol{h_j} \tag{5}$$

$$\boldsymbol{y_t} = Generator(\boldsymbol{s_{t-1}}, \boldsymbol{g_t}) \tag{6}$$

$$\boldsymbol{s_t} = Recurent(\boldsymbol{s_{t-1}}, \boldsymbol{g_t}, \boldsymbol{y_t}) \tag{7}$$

where the $Generator$ is an MLP with softmax outputs[15, 19], and $\boldsymbol{s_{t-1}}$ is the (t-1)-th state of an additional recurrent neural network. $\boldsymbol{\alpha_t}$ is a vector of the attention score at time t, and L is the length of the attention vector. Using the terminology from [15, 20], we call $\boldsymbol{g_t}$ a glimpse. At each time-step t, the energy function in eq.(4) computes the scalar energy e_{tj} for each time-step j, using vector $\boldsymbol{h_j} \in \boldsymbol{H}$ and s_{t-1}. The scalar energy e_{tj} is converted into a probability distribution (α_t) over time using a softmax function[19].

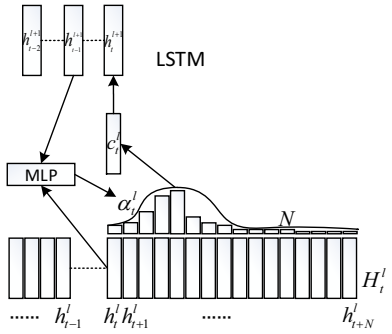

Figure 2: Attention-based LSTM architecture with future context size of N.

3. Attention-based LSTM architecture for speech recognition

Considering the coarticulation effect in a speech utterance, speech features of each phone are mostly influenced by several phones before and after it. Therefore, it may not be necessary to model the entire utterance as in BLSTM[21, 22]. This insight motivates us to propose the following architecture.

Attention-based LSTM is a LSTM (forward unidirectional) with attention mechanism combined into one or more hidden layers. For instance, Figure 2 shows this architecture with attention added into its l-th layer. As illustrated in the figure, the structure of the attention-based LSTM consists of an attention part and a LSTM part. The former offers context-dependent score to weight or select future context, which helps the latter to find future information in a context-dependent manner.

Given a feature vector sequence which consists of future N contexts $\boldsymbol{H_t^l} = (\boldsymbol{h_t^l}, \boldsymbol{h_{t+1}^l}, ..., \boldsymbol{h_{t+N}^l})$, each $\boldsymbol{h_t^l} \in R^{d \times 1}$ (d is the size of hidden node). The equations of the attention mechanism for the proposed attention-based LSTM are as follows:

$$\boldsymbol{\alpha_t^l} = softmax(energy(\boldsymbol{h_{t-1}^{l+1}}, \boldsymbol{H_t^l})) \tag{8}$$

$$\boldsymbol{c_t^l} = \sum_{j=0}^{N} \alpha_{tj}^l \boldsymbol{h_{t+j}^l} \tag{9}$$

$$\boldsymbol{h_t^{l+1}} = lstm(\boldsymbol{c_t^l}) \tag{10}$$

where $energy$ is an MLP network, and $lstm$ represents one or more LSTM layers. In this architecture, a simpler attention mechanism is employed. From eqs.(8)(9)(10), it can be observed that only the previous LSTM cell output vector $\boldsymbol{h_{t-1}^{l+1}}$ rather than the previous state $\boldsymbol{s_{t-1}}$ of an additional recurrent neural network (see eq.(8) and eqs.(4)(7)) is adopted as feedback information. The previous alignment α_{t-1}^l is also omitted from the eq.(8). Similar to FSMN, the attention score a_{tj}^l also can be scalar or vector. Here, the attention-based LSTM with scalar attention score (denotes as ALSTM) is the focus.

In this paper, three types of energy functions have been attempted, and the equations are as follows:

$$e_{tj}^l = \boldsymbol{w^T} \tanh(\boldsymbol{V h_{t+j}^l} + \boldsymbol{W h_{t-1}^{l+1}} + \boldsymbol{b_a}) \tag{11}$$

$$\boldsymbol{e_t^l} = \tanh(\boldsymbol{U h_{t-1}^{l+1}} + \boldsymbol{b_a}) \tag{12}$$

$$e_{tj}^l = < \frac{\boldsymbol{V h_{t+j}^l}}{\|\boldsymbol{V h_{t+j}^l}\|}, \frac{\boldsymbol{W h_{t-1}^{l+1}}}{\|\boldsymbol{W h_{t-1}^{l+1}}\|} > \tag{13}$$

where $\boldsymbol{V}$, $\boldsymbol{W}$ and $\boldsymbol{U}$ are matrix, $\boldsymbol{w}$, $\boldsymbol{b_a}$ are vectors. e_{tj}^l represents the energy value for the $(t+j)$-th frame at time-step t.

Table 1: *Comparison (model size, recognition performance and parameters) of various acoustic models. sFSMN and vFSMN denote bidirectional scalar and vectorized FSMNs, respectively; the memory blocks are added in the 2^{nd}, 4^{th} and 6^{th} layers. Row convolution represents adding row convolution in each LSTM layer. Target-delay denotes the LSTM with 5 frame delay.*

model	model size	PER	Parameters(M)
LSTM	3×512	23.1	2.93
	3×256	23.2	1.49
BLSTM		20.1	2.93
Row convolution		21.2	2.14
Target-delay		21.7	1.49
sFSMN	6×1024	21.6	8.85
vFSMN		20.0	8.90

The proposed architecture has some differences when compared with previous works[15, 17, 18, 19]. First, because of the alignment information in the CE criterion, the previous alignment α_{t-1}^{l} is considered unnecessary in computing the attention score α_{t}^{l} at the current time step. Second, outputs from the higher hidden layer of the ALSTM itself are used as the feedback information to compute the attention score of the current layer without an additional neural network like the usual attention mechanism (see eq.(11)). This simplifies the attention mechanism. Third, as the attention mechanism is simplified, the proposed ALSTM architecture can be applied to one or more hidden LSTM layers more easily and flexibly.

4. Experiments

In this section, the attention-based LSTM mechanism discussed in section 3 is evaluated on TIMIT phone recognition task. In this experiments, 123-dimensional FBK features are extracted, and 183 target class labels are labeled, the other experimental settings are similar to [23]. The networks were implemented using Theano library[24].

The initial experiments seek the performance of LSTM, BLSTM, target delay, row convolution, bidirectional sFSMN and vFSMN on TIMIT task. Table 1 shows the results. First, by comparing the results listed in Table 1 with a larger LSTM baseline which has a 23.1%, it can be concluded that the performance gain of the other models mainly comes from using future contexts rather than more model parameters. Second, bidirectional vFSMNs get best result that even outperforms BLSTM , indicating that utilization of temporal context within a limited window is enough to get future and past information. In the following experiments, all ALSTM models are trained with the size of 3×256 (3 hidden layers×256 nodes for each layer).

4.1. Experimental study on the necessity of past context modeling for ALSTM

Although the motivation of the proposed attention-based LSTM architecture is to model the future context for LSTM, experimental results are required to confirm whether the past context modeling is dispensable for ALSTM or whether the LSTM in ALSTM has already well modeled the past context. For this purpose, the bidirectional ALSTM architecture is setup in a similar way as the bidirectional FSMN, as shown in eq.(3):

$$\boldsymbol{c_t^l} = \sum_{i=1}^{N_1} a_{ti}^{l} \boldsymbol{h_{t-i}^l} + \sum_{j=1}^{N_2} c_{tj}^{l} \boldsymbol{h_{t+j}^l} + a_{t0}^{l} \boldsymbol{h_t^l} \qquad (14)$$

Table 2: *PERs (in %) for using I-ALSTM(Input layer attention-based LSTM) with different window sizes. bi-direction denotes bidirectional I-ALSTM, future and previous denote I-ALSTM using past and future content, respectively.*

Window size	6	7	8	9	10	11
bi-direction	23.1	23.1	22.4	22.3	22.1	22.9
future	22.8	23.0	22.2	22.5	22.2	22.6
past	23.9	23.5	22.8	23.4	23.4	23.4

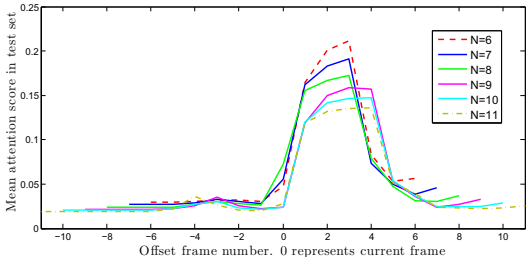

Figure 3: The attention score using I-BALSTM with different window sizes N (N_1 = N_2 from 6 to 11).

To reduce experimental efforts, the bidirectional ALSTM architecture is applied to the first hidden layer, denoted as the input layer bidirectional attention-based LSTM (I-BALSTM). Figure 4 a) shows the model structure of I-BALSTM. The energy function of eq.(12) (demonstrated as the best energy function in later experiments) is employed in this experiment. The results, as illustrated in the second row and denoted as bi-direction in Table 2, show that I-BALSTM with the window size of 10 (N_1=N_2=10) yields the best performance. However, this performance is worse than row convolution. Moreover, from Figure 3, the proposed attention mechanism pays more attention to the context at the position similar to target-delay regardless the window size. The score is higher in the range within future five frames offset from the current frame and is much lower in all past context. Therefore, the past context modeling is dispensable for ALSTM, or the LSTM in ALSTM has already been well modeled the past context.

In order to further confirm the above conclusion, speech recognition experiments are implemented at the I-BALSTM setup using either future part (second and third term of right side of eq.(14)) or past part (first and third term of right side of eq.(14)). The results are shown in the last two rows of Table 2. It can be observed that the recognition performance is mainly contributed by the future part. Therefore, the experiments further demonstrate that past context modeling is dispensable for ALSTM. The best performance is achieved with a window size of 10 for future context modeling; thus the window size N_1, N_2 is set to be 0, 10 separately in the following experiments.

4.2. Layer-wise ALSTM

Compared with the performance of row convolution (21.2%) and sFSMN (21.6%), I-BALSTM doesn't bring improvements. Therefore, ALSTM architecture is applied to each hidden layer, denoted as the layer-wise attention-based LSTM (L-ALSTM). Figure 4 b) shows the model structure of L-ALSTM. From Figure 4, it can be observed that attention at the each hidden layer gets feedback information from the current layer output. In this experiment, the three energy functions are also compared,

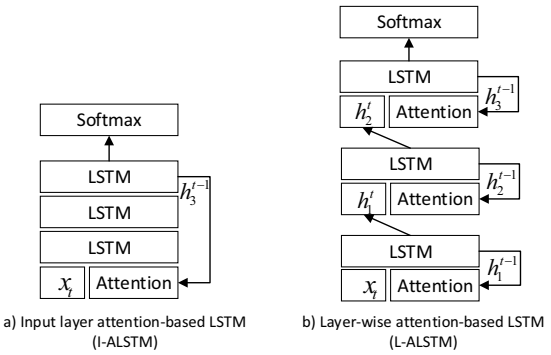

Figure 4: Structures of the input layer attention-based LSTM (I-ALSTM) and layer-wise attention-based LSTM (L-ALSTM).

and the L-ALSTMs with energy functions eqs.(11)(12)(13) are represented as L-ALSTM-1, LALSTM-2, L-ALSTM-3, respectively.

Figure 5 shows the attention scores for each layer. From Figure 5, it is observed that the attention scores in each layer have a similar envelope shape with the different attention spans of the future context. In layers 1 and 2, the attention mainly focuses on the future 2-3 frames. In layer 3, the attention to future context has wider future time span (3-5 future frames), which is very similar to the attention behavior of I-BALSTM (Figure 3). The feedback information for attention of the 3^{rd} layer in L-ALSTM and the 1st layer in I-BALSTM is both from the last hidden layer, so this may suggests that the wider attention span of the future context is due to the highest level feedback information. Moreover, the up-tail phenomenon is observed in the last fewer future frame in Figure 5. By analyzing the attention score of one sentence in the test set illustrated in Figure 6, the up-tail phenomenon may be generated by the up-tail attention scores of the silent frames.

The performances for L-ALSTM are reported in Table 3. Table 3 shows that L-ALSTM can achieve a significant performance improvement, which indicates that adding the attention mechanism in each layer is a better choice. Moreover, L-ALSTM-2 achieves the best performance, demonstrating that the energy function in eq.(12) is best (20.8%) in terms of performance among the 3 energy functions. All L-ALSTMs perform better than target-delay (see Table 1). The proposed L-ALSTM only employs scalar scores, and all L-ALSTMs perform better than the comparable model sFSMN, thus, it can be concluded that the context-dependent encoding weights work better than context-independent ones for acoustic modeling. However, the performance of L-ALSTM is still worse than BLSTM(20.1%), vFSMN(20.0%). Thus, extension of the proposed ALSTM architecture should be extended to the vector version to cope with the different context dependencies in the different dimensions of the representation or feature sequence.

5. Conclusions and Future Work

In summary, a novel architecture is proposed in this work called attention-based LSTM, which employs CD scores or CD weights to encode temporal future context information with the help of a type of attention mechanism for a unidirectional LSTM-based acoustic model. The experiments in TIMIT corpus show that when equipped with an attention mechanism in each layer, ALSTM achieves a 10.3% relative reduction in PER compared with LSTM and a 13.4% relative reduction for BLSTM. Moreover, ALSTM can achieve better performance than the comparable sFSMN, illustrating that CD weights work better than context-independent weights in acoustic modeling.

From the results in Table 3, it is clear that the improvement is significant when extending the encoding coefficient from scalar FSMNs to the vector one. Therefore, the proposed ALSTM architecture should be extended to the weight future context in both the temporal and node dimensions. And implementation the proposed ALSTM on larger corpus such as SWB task is also planned in future work.

Table 3: *PERs and PERRs (in %) for ALSTM with different types of energy function at N = 10.*

model	PER	PERR
LSTM	23.2	-
sFSMN	21.6	6.9
L-ALSTM-1	21.3	8.2
L-ALSTM-2	20.8	10.3
L-ALSTM-3	21.5	7.3
BLSTM	20.1	13.4
vFSMN	20.0	13.7

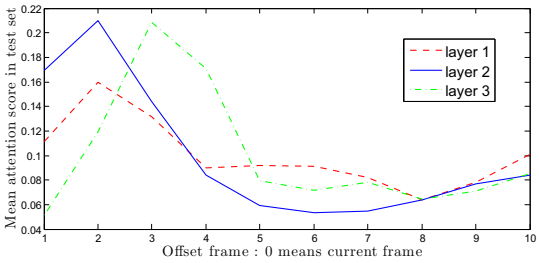

Figure 5: The attention score for L-ALSTM which uses future context at N=10.

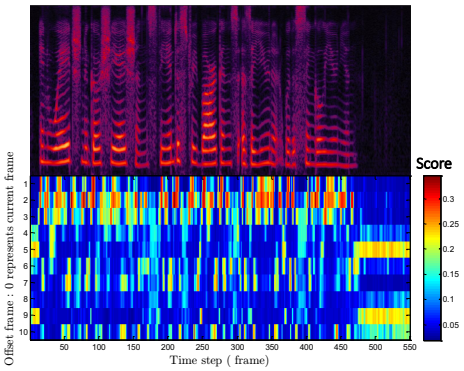

Figure 6: The attention score (future context L-ALSTM) for sentence SI1386 in the test set. The top figure is spectrogram.

6. ACKNOWLEDGMENT

We acknowledge the support of the following organizations or programs for research fundingNational Nature Science Foundation of China (Grant No.61273264), Science and Technology Department of Anhui Province (Grant No.15CZZ02007), Chinese Academy of Sciences (Grant No.XDB02070006),National Key Technology Support Program(2014BAK15B05).

7. References

[1] T. Robinson, M. Hochberg, and S. Renals, "The use of recurrent neural networks in continuous speech recognition," in *Automatic speech and speaker recognition*. Springer, 1996, pp. 233–258.

[2] H. A. Bourlard and N. Morgan, *Connectionist speech recognition: a hybrid approach*. Springer Science & Business Media, 2012, vol. 247.

[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] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6645–6649.

[5] H. Sak, O. Vinyals, G. Heigold, A. Senior, E. McDermott, R. Monga, and M. Mao, "Sequence discriminative distributed training of long short-term memory recurrent neural networks," *entropy*, vol. 15, no. 16, pp. 17–18, 2014.

[6] 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.

[7] H. Sak, A. Senior, K. Rao, and F. Beaufays, "Fast and accurate recurrent neural network acoustic models for speech recognition," *arXiv preprint arXiv:1507.06947*, 2015.

[8] F. Gers, "Long short-term memory in recurrent neural networks," Ph.D. dissertation, Universität Hannover, 2001.

[9] A. Graves, *Supervised sequence labelling*. Springer, 2012.

[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] A. Graves, S. Fernández, and J. Schmidhuber, "Bidirectional lstm networks for improved phoneme classification and recognition," in *Artificial Neural Networks: Formal Models and Their Applications–ICANN 2005*. Springer, 2005, pp. 799–804.

[12] S. Zhang, H. Jiang, S. Wei, and L. Dai, "Feedforward sequential memory neural networks without recurrent feedback," *arXiv preprint arXiv:1510.02693*, 2015.

[13] S. Zhang, C. Liu, H. Jiang, S. Wei, L. Dai, and Y. Hu, "Feedforward sequential memory networks: A new structure to learn long-term dependency," *arXiv preprint arXiv:1512.08301*, 2015.

[14] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," *arXiv preprint arXiv:1512.02595*, 2015.

[15] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in Neural Information Processing Systems*, 2015, pp. 577–585.

[16] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[17] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," *arXiv preprint arXiv:1508.04395*, 2015.

[18] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, "End-to-end continuous speech recognition using attention-based recurrent nn: first results," *arXiv preprint arXiv:1412.1602*, 2014.

[19] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell," *arXiv preprint arXiv:1508.01211*, 2015.

[20] V. Mnih, N. Heess, A. Graves *et al.*, "Recurrent models of visual attention," in *Advances in Neural Information Processing Systems*, 2014, pp. 2204–2212.

[21] K. Chen, Z.-J. Yan, and Q. Huo, "Training deep bidirectional lstm acoustic model for lvcsr by a context-sensitive-chunk bptt approach," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[22] Q. Huo and C. Chan, "A study on the use of bi-directional contextual dependence in markov random field-based acoustic modelling for speech recognition," *Computer Speech & Language*, vol. 10, no. 2, pp. 95–105, 1996.

[23] Z. Huang, J. Tang, S. Xue, and L. Dai, "Speaker adaptation of RNN-BLSTM for speech recognition based on speaker code," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[24] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, "Theano: a cpu and gpu math expression compiler," in *Proceedings of the Python for scientific computing conference (SciPy)*, vol. 4. Austin, TX, 2010, p. 3.