HYBRID LSTM-FSMN NETWORKS FOR ACOUSTIC MODELING

Asa Oines, Eugene Weinstein, Pedro Moreno

{asaj,weinstein,pedro}@google.com

ABSTRACT

This paper describes a series of experiments with neural networks containing long short-term memory (LSTM) [1] and feedforward sequential memory network (FSMN) [2, 3, 4] layers trained with the connectionist temporal classification (CTC) [5] criteria for acoustic modeling. We propose using a hybrid LSTM/FSMN (FLMN) architecture as an enhancement to conventional LSTM-only acoustic models. The addition of FSMN layers allows the network to model a fixed size representation of future context suitable for online speech recognition. Our experiments show that FLMN acoustic models significantly outperform conventional LSTM. We also compare the FLMN architecture with other methods of modeling future context. Finally, we present a modification of the FSMN architecture that improves performance by reducing the width of the FSMN output.

Index Terms— Long Short Term Memory, Feedforward Sequential Memory Networks, Connectionist Temporal Classification, future context, online speech recognition.

1. INTRODUCTION

Effective acoustic modeling for speech recognition requires taking into account the acoustic and phonemic context of the sound being recognized. This is due to the presence of co-articulation and other continuous-speech phenomena in human speech production. Much effort has been invested to train speech models that are sensitive to contextual information in the speech signal.

With the recent resurgence of deep neural networks (DNNs), these architectures are currently at the heart of most modern algorithms for acoustic modeling in automatic speech recognition. For ordinary feedforward neural networks, it is customary to provide a rudimentary measure of contextual information by "stacking" together consecutive frames of speech features. In recent years it has been demonstrated that cross entropy (CE) trained recurrent neural network (RNN) architectures such as long short term memory (LSTM) [1, 6, 7], which are context-aware due to their recurrence mechanism, have been shown to perform better than conventional feedforward neural networks. In the basic RNN approach, the network receives the sequence of feature vectors and attempts to label each feature vector with the correct phonetic label by using the information about the feature vector itself, as well as the recurrent "state" in the network's memory. This recurrent state allows the network to be sensitive to information seen at previous time steps in the feature stream. A modification called bidirectional RNN allows the network to consider both forward and the backward contextual information in its internal state representation. However, because of the need to process the feature sequence in reverse, this approach is not suitable for streaming applications such as voice search, where partial results are presented to the users as they are speaking. Additionally, CE trained networks have no concept of trying to learn the actual sequence of labels that must be produced to generate a correct transcript.

Even more recently, connectionist temporal classification (CTC) [5, 8] techniques have become adopted to remedy these limitations of CE training. CTC-trained models attempt to learn the sequence of labels that is required to produce the correct transcript, but do not attempt to model the label that should be given to a specific feature vector, nor do they attempt to output the labels in a way that is aligned temporally with the speech signal. As a result, they are able to use as much or as little future context as necessary, by choosing to defer outputting the label. This characteristic of CTC allows models trained with these techniques to outperform conventionally trained models (e.g., ones trained with the CE criterion). However, because the label sequence can be arbitrarily shifted from the input features corresponding to the labels, CTC models are difficult to use in a streaming setting. In such situations, modifications to the algorithm, such as enforcing a constraint of a maximum delay between features and corresponding output labels, are necessary [8] to make CTC models operate with acceptable latency characteristics.

Feedforward sequential memory networks (FSMN) [2, 3, 4] are a recently-proposed non-recurrent neural network topology that are able to model past and future contextual information explicitly through the use of memory blocks added to the network's hidden layers. FSMN models avoid the training-time complexity required to train RNNs (such as backward-propagation through time [9]), and have been shown to train faster than LSTMs models [4].

In this paper, we present our work with FSMN acoustic models. We conjecture that the learning ability of FSMN models is complementary to that of RNNs, and propose a new model topology that mixes LSTM and FSMN layers, a

combination we refer to as FLMN (Feedforward Long-short Memory Networks).

Starting with a very strong baseline based on clustered single-state triphone LSTM acoustic models trained with the CTC criterion, we demonstrate in thorough experiments on several languages in the context of a mobile voice search and dictation transcription task that FLMNs can outperform LSTMs of similar sizes.

The outline of the paper is as follows. In Section 2, we give a brief overview of FSMN networks and our motivation behind combining LSTM and FSMN. Next, in Section 3 we explain our experimental setup. Then in Section 4 we detail our experiments and present our results. Finally, in Section 6 we summarize our conclusions.

2. METHODOLOGY

2.1. Feedforward Sequential Memory Networks

Feedforward sequential memory networks (FSMN) [2, 3, 4] modify feedforward networks by allowing them to explicitly model past and future contextual information. These networks can be thought of as a high-order finite impulse response filter approximation [10] of the infinite impulse response filter of conventional RNNs [2]. FSMN models have been claimed to offer the temporal modeling power of RNN models while avoiding the additional complexity in training that comes with recurrent connections [2]. This enables FSMN models to be trained more simply and quickly than their RNN counterparts.

FSMN layers consist of two components; a traditional feedforward layer and a memory block. The memory block encodes the past N_1 and future N_2 activations of the feedforward layer into a fixed size representation. With vectorized FSMN the output of the memory block is obtained by element wise multiplication of those activations with a trainable matrix of encoding coefficients. This is described in Eq. 1 [2], where $\tilde{\mathbf{h}}_t^\ell$ represents the output of the memory block at time t, $\mathbf{h}^\ell$ represents the activations of the associated hidden layer, a^ℓ and c^ℓ represent trainable encoding coefficients, and $\odot$ denotes element-wise multiplication. This fixed size respresentation of contextual information is passed to the next layer with the current-frame activations.

$$\tilde{\mathbf{h}}_t^\ell = \sum_{i=0}^{N_1} a_i^\ell \odot \mathbf{h}_{t-i}^\ell + \sum_{j=1}^{N_2} c_j^\ell \odot \mathbf{h}_{t+j}^\ell \quad (1)$$

2.2. Hybrid LSTM/FSMN Networks

Our early experiments were to explore training FSMNs with the CTC criterion. We found that FSMNs could be trained effectively with CTC, and observed that the quality achieved with FSMN-CTC was roughly similar to that with the LSTM-CTC baseline. By design, FSMN layers model contextual information immediately surrounding the current acoustic frame while the LSTM layers are able to remember context indefinitely, allowing them to model more distant temporal dependencies. We thus hypothesized that the two topologies are able to compensate for each other's shortcomings when modeling contextual information about the temporal evolution of the speech signal. This led us to investigate hybrid LSTM/FSMN (FLMN) acoustic models, to explore whether the learning power of the two topologies could be complementary. Indeed, we observed that FLMN models were able to outperform both FSMN-only and LSTM-only acoustic models across all our evaluation metrics (see Section 4 for more information). The FLMN acoustic model architecture is shown in figure 1.

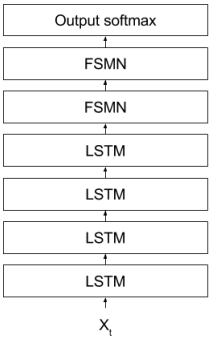

Fig. 1. The FLMN architecture

3. EXPERIMENTAL SETUP

3.1. Training setup

Our acoustic feature vector sequence follows previous work from our group on CTC acoustic models [8] and consists of 80-dimensional log-mel features computed with a 25ms window shifted every 10 ms. We then stack 8 consective frames keeping every third feature frame, resulting in a frame being processed every 30 ms. To improve performance on mixed bandwidth data we probabilistically downsample 20% of training examples from 16kHz to 8kHz. The features corresponding to the higher frequencies are subsequently zeroed out in accordance with [11]. In order to improve noise robustness we do multistyle training (MTR) [12] where training data is artifically distorted using a room simulator and by adding background noise with an SNR varying randomly between 5 and 25. All LSTM layers described in Section 4 couple the input and forget gates. This has been shown to reduce the number of parameters in the network without impacting quality [13].

3.2. Training methods

Models were trained with the CTC criterion [5] to convergence using asynchronous stochastic gradient descent [14].

Our targets were 8,191 clustered single-state triphones plus the blank symbol. Unless otherwise specified the set of search paths in the forward-backward algorithm is constrained to those paths for which the delay between the CTC labels and the "ground truth" alignment does not exceed 100ms as described in Section 1.

3.3. Data sets

Our training corpus consists of human transcribed voice search and dictation anonymized logs sent to our recognizers. The number of utterances and total length of our training corpus for each language is listed in Table 1. We withold 0.5% of these utterances to be used as a development set.

Language	Country of origin	Size	Length (hours)
Swedish	Sweden	3M	3.5k
English	India	11M	14.6k
Italian	Italy	10M	13.6k
French	France	16M	24.2k

Table 1: Training corpora

3.4. Evaluation

For each experiment we compare the quality of the overall system with the experimental acoustic model to the system with the baseline acoustic model described in Section 4.1. Word error rate (WER) results were measured on two types of test sets.

VS a set of voice search utterances, in the given language.

IME a set of dictation utterances, in the given language.

The size of these test sets typically include between 2k and 15k utterances, which corresponds to roughly between 3 and 20 hours of audio. All experiments made use of a standard WFST-based beam-search decoder, in conjunction with 5-gram LMs with ~15M n-grams and ~1M words, trained on data from the target language being decoded.

4. EXPERIMENTS AND RESULTS

4.1. Baselines

First we established baseline numbers for conventional LSTM models. Our baseline topology consisted of 5 fully connected LSTM layers of 768 units followed by a softmax layer with 8,192 outputs. In past work, it has been demonstrated that extending this topology beyond 5 layers does not yield improvements [15]. For completeness, we repeated these experiments and compared networks with 5-8 LSTM layers, coming to the same conclusion. The CTC label error rate (LER) on the single-state triphone clusters as measured on a held-out set for these experiments are detailed in Table 2.

Baselines were trained for all languages on which we ran experiments. WER and LER for these baselines are presented in Table 3.

Language	Layers	LER (%)
Swedish	5	27.5
	6	27.6
	7	27.5
	8	27.3

Table 2: Adding LSTM layers does not improve LER

Language	LER (%)	WER (%)	
		VS	IME
Swedish	27.5	20.4	17.4
English	27.7	22.0	19.2
Italian	20.5	12.7	7.4
French	24.0	14.2	10.2

Table 3: Baseline 5 layer LSTM LER and WER

4.2. Hybrid FLMN acoustic models

In this experiment we compared acoustic models trained with our FLMN topology against our LSTM baseline models. The FLMN topology consists of 4 fully connected LSTM layers of 768 units, followed by 2 fully connected FSMN layers of 768 units, followed by a softmax layer with 8,192 outputs. The FSMN layers use lookahead and lookback orders of 15 activations, translating to 450ms of past and future context. These lookahead and lookback orders were chosen to roughly correspond with the optimal orders found in [2]. WER and LER for FLMN models are shown in Table 4. We found that our FLMN models significantly outperformed the LSTM baselines in all but one language-test set pairs.

Language	VS WER (%)		IME WER (%)	
	FLMN	LSTM	FSMN	LSTM
Swedish	**19.6**	20.4	**16.5**	17.4
English	**20.5**	22.0	**17.9**	19.2
Italian	**12.0**	12.7	7.5	**7.4**
French	**13.3**	14.2	**10.1**	10.2

Table 4: FLMN results

4.3. Relaxing the CTC alignment constraint

One advantage our FLMN architecure has over the baseline is the amount of future context available to the model. Our goal in this experiment was to explore whether this advantage was instrumental in allowing the FLMNs to improve over the CTC baseline. Though CTC models are in general able to delay outputting labels indefinitely, our baselines were constrained

to limit the delay between acoustic features and their corresponding labels [8] to 100ms as described in Section 3.2. We thus trained CTC models relaxing this constraint by increasing it from 100ms to 550ms, making the total future context available to LSTM-CTC models equal to that in our FLMN architecture. The WER for these experiments and their respective baselines can be found in Table 5. We found that relaxing this constraint had had a small effect on WER when compared to modeling future context via FLMN.

Language	VS WER (%)		IME WER (%)	
	$\leq 550ms$	$\leq 100ms$	$\leq 550ms$	$\leq 100ms$
Swedish	20.4	20.4	17.4	17.4
English	**21.5**	22.0	**18.6**	19.2

Table 5: Relaxing the CTC alignment constraint on LSTM acoustic models

4.4. Varying FSMN context windows

In streaming applications such as voice search, it is important to have low latency between speech and the partial results presented. The lookahead order in FSMN layers sets a lower bound on this latency due to the amount of future context required to produce activations. To examine the tradeoff between latency and quality, we investigated how varying the lookahead and lookback orders of our FLMN architecture impacted WER. The results of these experiments are shown in Table 6. We found that we can reduce the order of the FSMN layers while maintaining LER gains over the baseline system but that these gains in LER did not translate into gains in WER.

Language	Context window		LER (%)	WER (%)	
	Activations	Time		VS	IME
French	15	450ms	18.9	**13.3**	**10.1**
	10	300ms	**18.6**	14.1	10.2
	5	150ms	20.1	14.0	10.2

Table 6: FSMN order in FLMN

4.5. Improving FSMN performance

In this experiment we investigated a way to improve the performance of the FSMN layers in our FLMN architecture. One disadvantage of FSMN layers is that the memory block encoding is concatenated with the activations of the feedforward layer. This results in an output that is twice as large as a feedforward or LSTM layer, doubling the number of parameters in the weight matrix of the following layer and thus the number of operations needed in the matrix multiplication. To resolve this we experimented with instead summing the encodings and the feedforward activations. This resulted in an increase in training speed of almost 50%. Additionally, this reduced the number of parameters in the model by 23% as shown in Table 7. Ongoing experiments presented in Figure 2 suggest that this modification does not affect acoustic model quality.

Topology	Parameters	Rel. difference (%)
LSTM	26.2MM	-
FLMN	29.3MM	+10.5
FLMN-sum	22.7MM	-15.4

Table 7: Number of trainable parameters

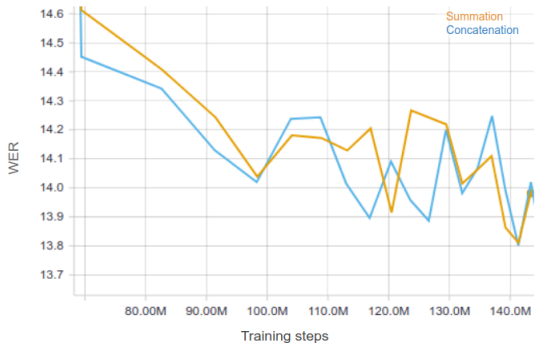

Fig. 2. Rolling WER eval of French FLMN models

5. DISCUSSION

The combination of FSMN and LSTM layers in the FLMN topology appears to have greater contextual modeling power than LSTM layers alone. The FSMN layers are able to explicitly model the context directly surrounding the current frame. This complements the implicit contextual modeling of LSTMs, which we believe are better suited towards modeling longer term context. This increase in contextual modeling power allows FLMN to outperform LSTM while using a similar number of parameters. Like LSTM, feedforward networks, and convolutional networks, FSMN are another useful "building block" in the acoustic modeling toolbox.

6. SUMMARY & CONCLUSIONS

We have described a new architecture for acoustic modeling combining LSTM and FSMN layers. Using this architecture, we have trained acoustic models with the CTC criterion for four languages and have demonstrated that these models consistently and reliably outperform their strong LSTM-CTC baselines. We hypothesized that the differences in these two topologies modeling of contextual information allows them to complement each other. Finally, we investigated techniques to improve the performance of these models, including methods that reduce latency and computational cost.

7. REFERENCES

[1] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[2] Shiliang Zhang, Cong Liu, Hui Jiang, Si Wei, Li-Rong Dai, and Yu Hu, "Feedforward sequential memory networks: A new structure to learn long-term dependency," *CoRR*, vol. abs/1512.08301, 2015.

[3] Shiliang Zhang, Hui Jiang, Shifu Xiong, Si Wei, and Li-Rong Dai, "Compact feedforward sequential memory networks for large vocabulary continuous speech recognition," in *Interspeech*, September 2016, pp. 3389–3393.

[4] S. Zhang, C. Liu, H. Jiang, S. Wei, L. Dai, and Y. Hu, "Nonrecurrent neural structure for long-term dependence," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 871–884, April 2017.

[5] Alex Graves, Santiago Fernández, and Faustino Gomez, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *In Proceedings of the International Conference on Machine Learning, ICML 2006*, 2006, pp. 369–376.

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

[7] Haşim Sak, Oriol Vinyals, Georg Heigold, Andrew Senior, Erik McDermott, Rajat Monga, and Mark Mao, "Sequence discriminative distributed training of long short-term memory recurrent neural networks," in *Interspeech*, 2014.

[8] Andrew, H. Sak, F. de Chaumont Quitry, T. Sainath, and K. Rao, "Acoustic modelling with cd-ctc-smbr lstm rnns," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec 2015, pp. 604–609.

[9] P. J. Werbos, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, Oct 1990.

[10] Alan V. Oppenheim and Ronald W. Schafer, *Discrete-Time Signal Processing*, Prentice Hall Press, Upper Saddle River, NJ, USA, 3rd edition, 2009.

[11] Jinyu Li, Dong Yu, Jui-Ting Huang, and Tifan Gong, "Improving wideband speech recognition using mixed-bandwidth training data in cd-dnn-hmm," in *IEEE Workshop on Spoken Language Technology*, January 2012.

[12] R. Lippmann, E. Martin, and D. Paul, "Multi-style training for robust isolated-word speech recognition," in *ICASSP '87. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Apr 1987, vol. 12, pp. 705–708.

[13] Klaus Greff, Rupesh Kumar Srivastava, Jan Koutník, Bas R. Steunebrink, and Jürgen Schmidhuber, "LSTM: A search space odyssey," *CoRR*, vol. abs/1503.04069, 2015.

[14] J. Dean, G.S. Corrado, R. Monga, K. Chen, M. Devin, Q.V. Le, M.Z. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, and A.Y. Ng, "Large Scale Distributed Deep Networks," in *Proc. NIPS*, 2012.

[15] Hasim Sak, Andrew W. Senior, and Françoise Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," *CoRR*, vol. abs/1402.1128, 2014.