HYBRID LSTM-FSMN NETWORKS FOR ACOUSTIC MODELING

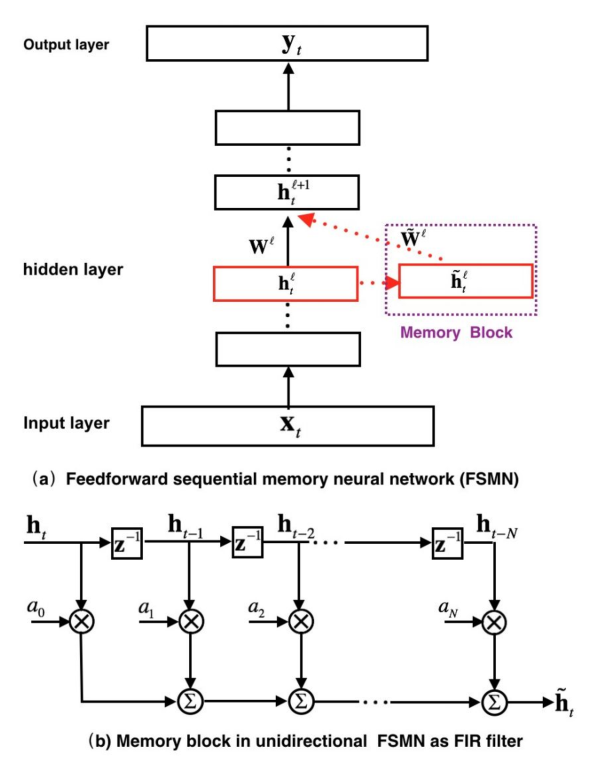
Asa Oines, Eugene Weinstein, Pedro Moreno
Google Inc., USA

Introduction

- Contextual information is key to training acoustic models.
- Recursive neural networks (RNNs) such as Long short term memory (LSTMs) are context-aware due to their recurrence mechanism, and usually outperform conventional neural networks.
- Connectionist temporal classification (CTC) allows neural networks to be trained to output the desired sequence of feature labels, but without the requirement to label specific feature vectors with specific labels (alignment).

FSMN

- Feedforward sequential memory networks (FSMN): Recently-proposed non-recurrent neural network topology which models past and future context through the use of memory blocks.
- Shown to be competitive with and even outperform LSTMs, and are faster to train ["Feedforward Sequential Memory Networks: A New Structure to Learn Long-term Dependency", S. Zhang, C. Liu, H. Jiang, S. Wei, L. Dai, Y. Hu, TASLP, VOL. 25, NO. 4, April 2017].
- FSMN structure: Architecture diagram borrowed from above paper (see also other FSMN paper in this session):



- Feedforward layers are combined with memory blocks which encode the past N_1 and future N_2 activations of a feedforward layer.

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

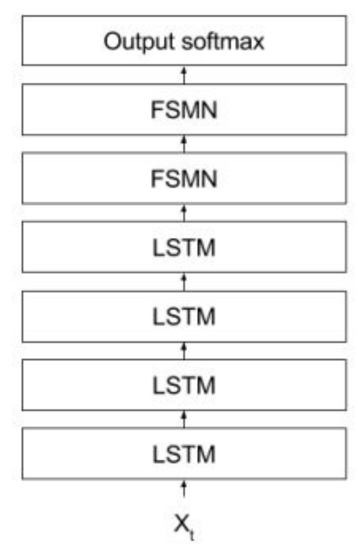
- $\tilde{\mathbf{h}}_t^\ell$: activations of the hidden feedforward layer at time t
- $\mathbf{h}^\ell$: output of the memory block
- c^ℓ, a^ℓ : trainable encoding coefficients
- $\odot$: element-wise multiplication

Combining FSMN + LSTM

- In our early experiments, FSMN and LSTM performed comparably when trained with the CTC objective.
- Based on the hypothesis that their modeling power can be complementary, we decided to experiment with combining the two layer types in one network.

Hybrid LSTM/FSMN (FLMN)

- LSTM and FSMN layers are combined in one network.



Training Setup

- 80-dimensional log-mel features
- 25ms-window frames computed every 10ms
- Process every third frame (every 30ms)
- Mixed-bandwidth training (16kHz data, 20% downsampled to 8kHz, with features zero-padded)
- Artificially distort data with room simulation, added background noise (multistyle training - MTR)
- Models trained with CTC criterion using asynchronous stochastic gradient descent (ASGD)
- FSMNs have 450ms of future+past context (15 frames)

Data Sets

- Human-transcribed voice search and dictation training corpora:

Language	Country	# utterances	# hours
Swedish	Sweden	3M	3.5K
English	India	11M	14.6K
Italian	Italy	10M	13.6K
French	France	16M	24.2K

- Test sets: human-transcribed test data
- VS (voice search); IME (dictation)
- Between 2K and 15K utterances (3-20 hours of audio)

Baseline Experiments

- Adding LSTM layers does not improve LER (Swedish):

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

- Baseline 5 layer LSTM LER (label error) and WER:

		WER (%)	
Language	LER (%)	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

Hybrid FLMN Models

- 4 fully connected LSTM layers + 2 fully connected FSMN layers (768 units per layer).
- Softmax with 8192 outputs (context-dependent phones).

Language	VS WER (%)		IME WER (%)	
	FLMN	LSTM	FLMN	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

Relaxing Real-time Requirements

- Our CTC models are ordinarily trained to output the label at most 100ms after it was spoken (for real-time ASR considerations).
- To match the context window sizes, we relaxed this to 550ms; however, quality does not improve.
- This suggests FLMN is better equipped to model short-distance context.

Language	LSTM VS WER (%)		LSTM IME WER (%)	
	≤ 550ms	≤ 100ms	≤ 550ms	≤ 100ms
Swedish	20.4	20.4	17.4	17.4
English	**21.5**	22.0	**18.6**	19.2

Varying FSMN Context Window

- We conducted experiments to evaluate the effect of varying the context window size in FLMNs.
- Smaller context windows degrade in WER (French).

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

Reducing Model Size

- FSMN layer concatenates outputs of feedforward layer and memory block.
- This results in a doubling of the size of weight matrix of following layer.
- We experimented with summing instead of concatenating to reduce model size.
- Experiments suggest that this does not affect performance.

Conclusions

- Combining FSMN and LSTM layers (FLMN) yields greater contextual modeling power than LSTM alone, in models that have similar numbers of parameters.
- This is likely due to the fact that FSMN focuses on context surrounding the current frame, while LSTMs are better at modeling longer-term context.