RECURRENT CONVOLUTIONAL NEURAL NETWORKS FOR SPEECH PROCESSING

Yue Zhao, Xingyu Jin

Department of Electronic Engineering, TNList,
Tsinghua University, Beijing, 100084, China

Xiaolin Hu

Department of Computer Science and Technology,
TNList, Tsinghua University, Beijing, 100084, China

MOTIVATION

- Existing CNN and RNN have specific disadvantages.
  - CNN has not exhibited significant improvement in speech processing.
  - RNN is expected to function well in modeling sequential, but is harder to train efficiently.
- A new architecture of Recurrent Convolutional Neural Network (RCNN) [1, 2] works well in object recognition and scene labeling.
- In view of the embedded RNN structure, RCNN is expected to function well in modeling speech, a typical temporally sequential data.

FORMULATION

- Conventional RNN: (neglecting bias term)

$$\mathbf{h}(t) = \sigma\left(W_{xh}\mathbf{x}(t) + W_{hh}\mathbf{h}(t-1)\right)$$

($\mathbf{x}(t)$: feed-forward input, $\mathbf{h}(t)$: hidden state at time t)

- Recurrent Convolutional Layer (RCL):

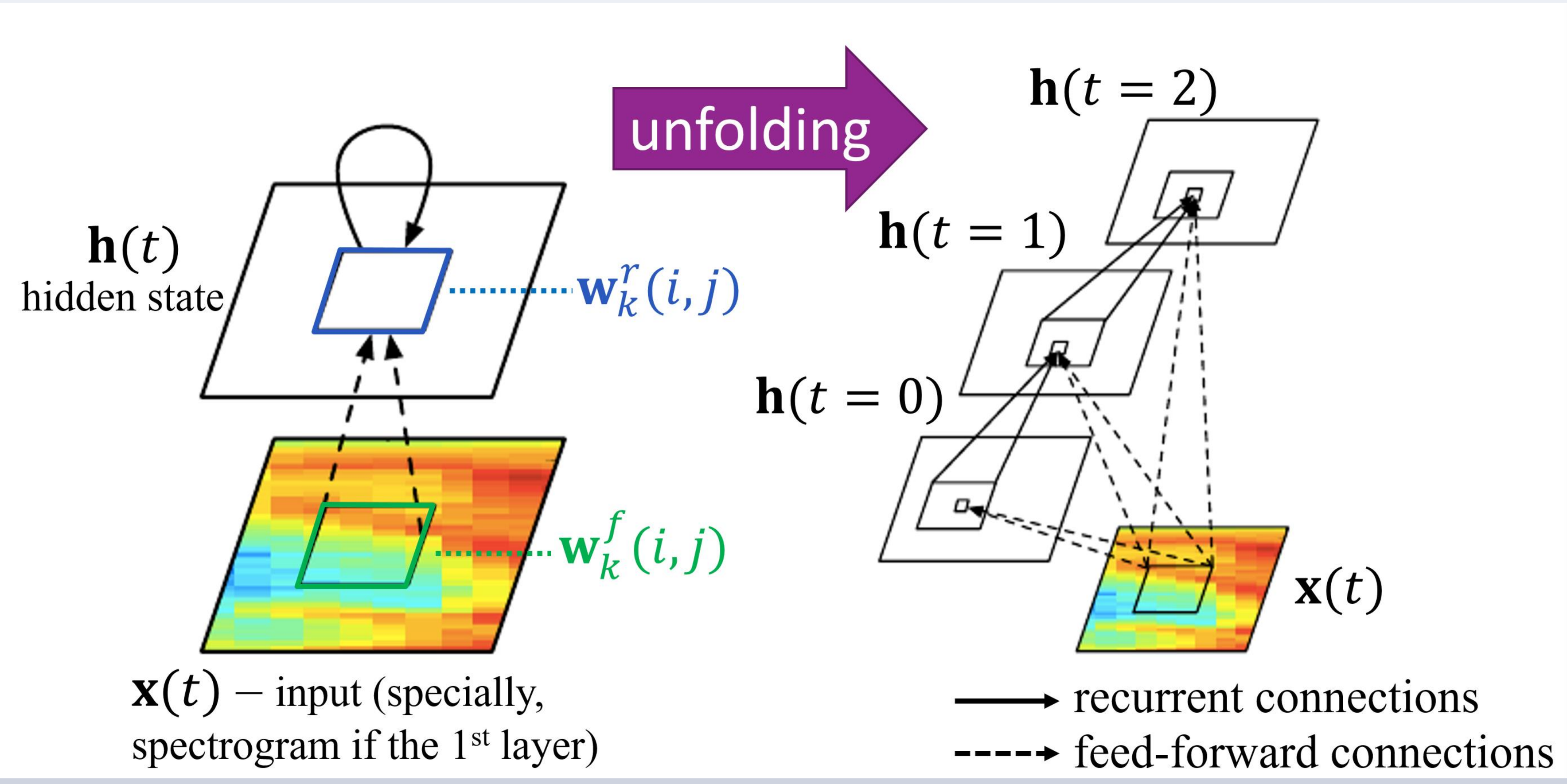
$$\mathbf{h}^{(t)}(i,j) = \sigma\left(\sum_{i'=-s}^{i'=s}\sum_{j'=-s}^{j'=s} W_k^f(i',j')\,\mathbf{x}^{(t)}(i-i',j-j') + \sum_{i'=-s'}^{i'=s'}\sum_{j'=-s'}^{j'=s'} W_k^r(i',j')\,\mathbf{h}^{(t-1)}(i-i',j-j')\right)$$

(W_k^f: k[th] feedforward kernel, W_k^r: k[th] recurrent kernel)

- Nonlinearity $\sigma(x) = f_{bn}(g(x))$ is realized by Rectified Linear (ReLU) $g(x) = \max(x, 0)$ and batch normalization $f_{bn}(x_i; \gamma, \beta)$.
- "time step" (t) in RCL: a RCL processes information from neighboring *time slots* and *frequency banks* at each *iteration*.

ILLUSTRATION

Illustration of a single RCL and its unfolded version with T=3.



RESULTS

- Phoneme recognition on TIMIT
  - Unfolding more times yields lower PER but there is a limit.
  - Outperform most ANN-HMM models. Competitive to existing methods. (More in paper)

	PER (dev set)	PER (core test set)
4-layer MLP	19.9%	22.0%
CL + pooling + 3-layer MLP	18.4%	20.0%
CL1 + CL2 + pooling + 3-layer MLP	19.2%	20.5%
RCL (T=1) + pooling + 3-layer MLP	18.3%	20.3%
RCL (T=2) + pooling + 3-layer MLP	17.3%	19.2%
RCL (T=2) + CL + 3-layer MLP	**17.0%**	**18.0%**
3-layer LSTM + HMM [3]*	17.7%	18.8%

- The speed of RCNN is faster than LSTM module, both when training and decoding.

	train	decode
RCNN	2012 samples per second	1.721 utterances per second
LSTM	275 samples per second	0.944 utterances per second

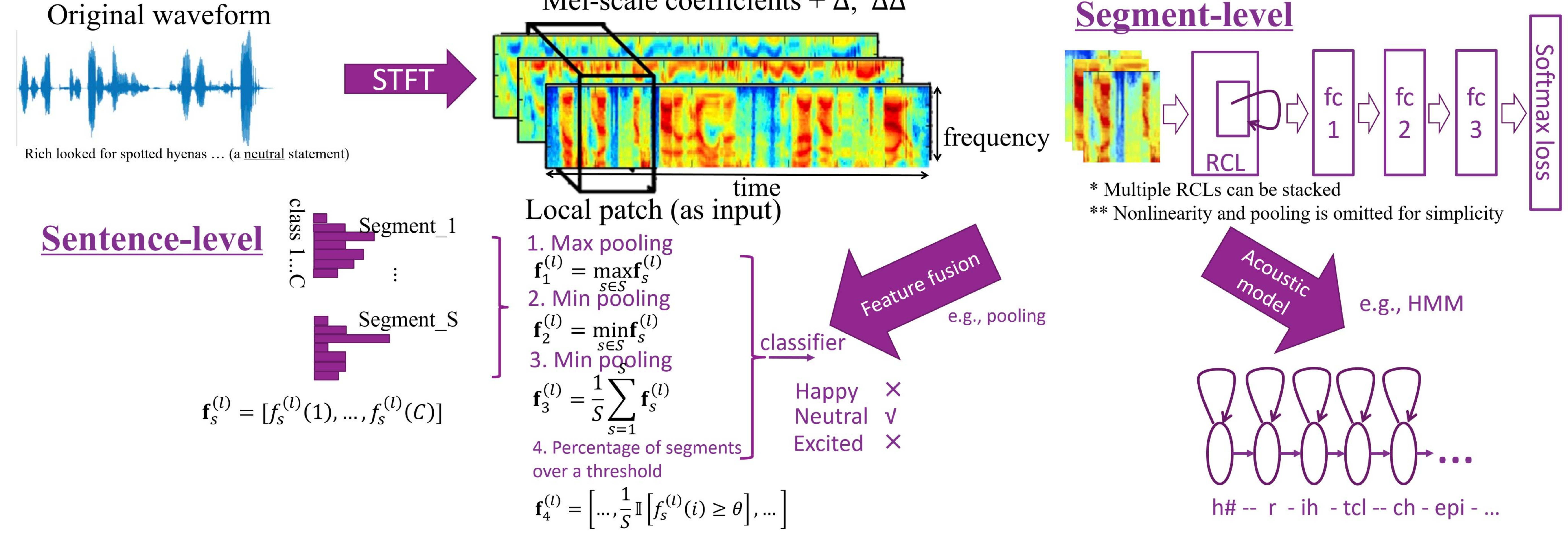
* Paper [3] works on another dataset. We borrow the structure for comparison of speed. Parameters may not be perfectly tuned.

RESULTS

- Emotion recognition on IEMOCAP
  - Spectral features, relatively lower-level feature than MFCC, can achieve competitive, even better results.

	Framewise accuracy	Weighted accuracy	Unweighted accuracy
3-layer MLP	41.4%	48.5%	39.9%
CL+2-layer MLP	43.1%	53.4%	41.6%
RCL+2-layer MLP	43.5%	**53.6%**	42.8%
(MFCC+pitch) MLP+SVM [4]	-	~50%	**~45%**
Log Spec(+PCA whitening)+CNN [5]	-	-	35.98% (40.02%)

PIPELINE



CONCLUSIONS

- Propose to use RCNN originally from computer vision to speech processing.
- RCNN achieves competitive results with existing models. Also, it runs faster than LSTM networks.
- Inspire more generic and efficient cross-modal deep learning models in the future.

REFERENCES

[1] M. Liang and X. Hu. Recurrent convolutional neural network for object recognition. CVPR 2015.

[2] M. Liang, X. Hu, and B. Zhang. Convolutional neural networks with intra-layer recurrent connections for scene labeling. NIPS 2015.

[3] Y. Zhang, G. Chen, D. Yu, K. Yaco, S. Khudanpur, and J. Glass. Highway long short-term memory RNNs for distant speech recognition. ICASSP 2016.

[4] K. Han, D. Yu, and I. Tashev. Speech emotion recognition using deep neural network and extreme learning machine. INTERSPEECH 2014.

[5] W. Zheng, J. Yu, and Y. Zou. An experimental study of speech emotion recognition based on deep convolutional neural networks. In International Conference on Affective Computing and Intelligent Interaction (ACII), pages 827–831. IEEE, 2015

SOURCE CODES

The source codes can be downloaded at: https://github.com/zhaoyue-zephyrus/RecurrentConvNet-for-Speech

Contact: Yue Zhao, thuzhaoyue@gmail.com