Fast and Accurate Recurrent Neural Network Acoustic Models for Speech Recognition

Qilei Zhang

Jun 16 2018

Abstract

We have recently shown that deep Long Short-Term Memory (LSTM) recurrent neural networks (RNNs) outperform feed forward deep neural networks (DNNs) as acoustic models for speech recognition.

1. Introduction

While speech recognition systems using recurrent and feedforward neural networks have been around for more than two decades [1], it is only recently that they have displaced Gaussian mixture models (GMMs) as the state-of-the-art acoustic model [2]. More recently, it has been shown that recurrent neural networks can outperform feed-forward networks on large-scale speech recognition tasks [3].

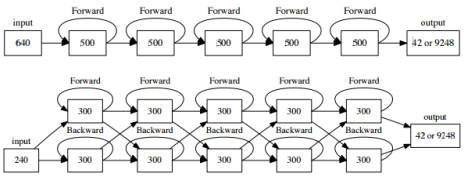

Figure 1. Layer connections in unidirectional (top) and bidirectional (bottom) 5-layer LSTM RNNs.

$$x'(t) = -V'(x) + A_0 cos(wt + o) + u(t) \qquad (1)$$

2. RNN Acoustic Modeling Techniques

In this work we focus on the LSTM RNN architecture which has shown good performance in our previous research, outperforming deep neural networks. [4].

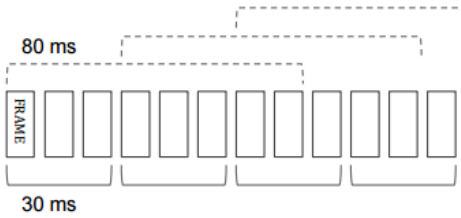

Figure 2. Stacking and subsampling of frames. Acoustic features are generated every 10ms, but are concatenated and downsampled for input to the network: 8 frames are stacked for unidirectional (top) and 3 for bidirectional models (bottom).

3. Experiments

We train and evaluate LSTM RNN acoustic models on handtranscribed, anonymized utterances taken from real 16kHz Google voice search traffic [5]. Our training set consists of 3 million utterances with average duration of about 4s [6]. To achieve robustness to background noise and reverberant environments we synthetically distort each utterance in a room simulator with a virtual noise source [7].

References

[1] C. Barat and C. Ducottet. String representations and distances in deep convolutional neural networks for image classification. *Pattern Recognition*, 54(1):104–115, 2016. 1

[2] C. D. Eiber, N. H. Lovell, and G. J. Suaning. Attaining higher resolution visual prosthetics: a review of the factors and limitations. *Journal of Neural Engineering*, 10(1):1102, 2013. 1

[3] A. Krogh. Neural network ensemble, cross validation and active learning. *Advances in Neural Information Processing Systems*, 7(10):231–238, 1995. 1

[4] J. Pustejovsky. The generative lexicon. *Computational Linguistics*, 17(4):409–441, 1998. 1

[5] T. D. Sanger. Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural Networks*, 2(6):459–473, 1989. 1

[6] H. H. Szu, B. A. Telfer, and S. L. Kadambe. Neural network adaptive wavelets for signal representation and classification. *Optical Engineering*, 31(9):1907–1916, 1992. 1

[7] M. C. Wittrock. Generative processes of comprehension. *Educational Psychologist*, 24(4):345–376, 1989. 1