Hybrid Speech Recognition with Deep Bidirectional LSTM

Alex Graves[1,2], **Navdeep Jaitly**[2] **and Abdel-rahman Mohamed**[2]

[1]DeepMind Technologies Ltd.

[2]University of Toronto

{graves, ndjaitly, asamir}@cs.toronto.edu

Introduction

Deep recurrent neural networks were recently shown to give state-of-the-art performance in phoneme recognition on the TIMIT database [1]. However these results relied on end-to-end training methods that are difficult to integrate into existing speech recognition systems. In particular, it is not straightforward to combine such methods with the pronunciation dictionaries and language models required for large vocabulary recognition. This work investigates the use of deep recurrent nets as acoustic classifiers in a traditional hybrid HMM-neural network system. The goal was to provide a straightforward alternative to deep feedforward nets that can be plugged into the standard large-vocabulary framework.

Network Architecture

A basic recurrent neural network (RNN) is defined by the following computation graph:

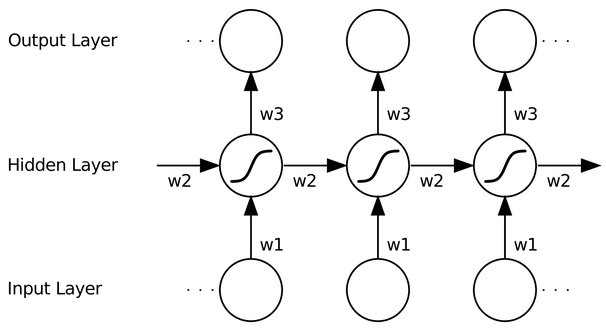

Figure 1: *Recurrent Neural Network*

One problem with basic RNNs is that they are only able to make use of previous context. In speech recognition, where whole utterances are transcribed at once, it is clearly beneficial to exploit future context as well. Bidirectional RNNs (BRNNs) [2] do this by processing the data in both directions with two separate hidden layers, which are then fed forwards to the same output layer.

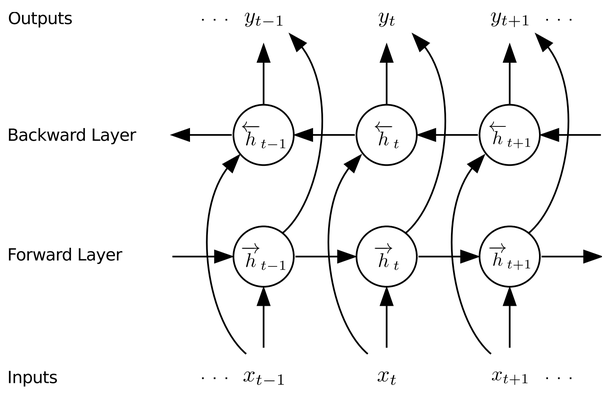

Figure 2: *Bidirectional Recurrent Neural Network*

A more subtle failing of RNNs is that they have trouble accessing long-range context. The problem is that the information stored in the hidden layer tends to decay over time as it repeatedly loops through the recurrent connections, and is also prone to being overwritten by new information arriving from the inputs. The Long Short-Term Memory network architecture (LSTM; [3]) addresses these issues by adding multiplicative gating units that allow the network to control the flow of information in and out of the hidden layer.

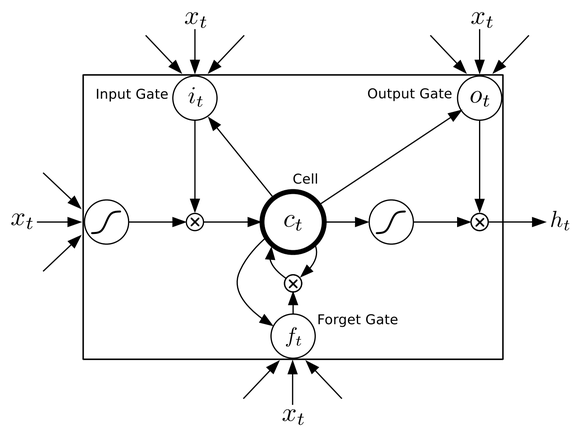

Figure 3: *Long Short-term Memory Cell*

A crucial element of the recent success of hybrid systems is the use of *deep* architectures, which are able to build up progressively higher level representations of acoustic data. *Deep RNNs* can be created by stacking multiple RNN hidden layers on top of each other, with the output sequence of one layer forming the input sequence for the next.

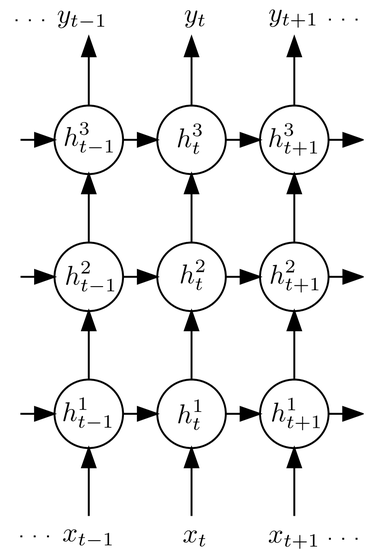

Figure 4: *Deep Recurrent Neural Network*

Putting all these ideas together gives deep bidirectional Long Short-Term Memory (DBLSTM), the architecture used in this paper:

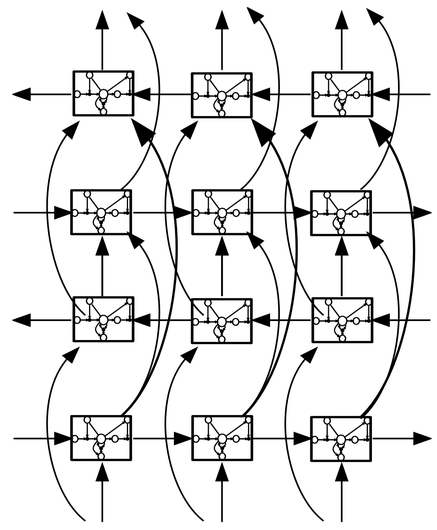

Figure 5: *Deep Bidirectional Long Short-Term Memory*

Network Training

Network training followed the standard HMM-neural network hybrid approach [4]: Frame-level state targets were provided by a forced alignment from a GMM-HMM system, and the network was then trained to classify the states using a softmax output layer. However there were two main differences from the way feedforward nets are usually trained:

- There was no context window on either side of the acoustic frame being classified (RNNs don't need this because they can transmit context through their internal state).
- The weight updates were calculated for entire utterances at once (reflecting the fact that every output may depend on every input). This differs from feedforward nets, where the updates are typically performed on mini-batches drawn randomly from the whole training set. One consequence is that RNN training is harder to parallelise.

Gaussian weight noise [5] was used to regularise the networks to prevent overfitting. We have found this method more effective for RNNs than better-known regularisers such as dropout or L1 or L2 weight penalties.

TIMIT Experiments

The first set of experiments were carried out on the TIMIT speech corpus. The aim was to see how hybrid training for deep bidirectional LSTM compared with the end-to-end training methods described in [1].

Table 1: *TIMIT Results with End-To-End Training.*

METHOD	DEV PER	TEST PER
CTC	19.05 ± 0.11	21.57 ± 0.25
CTC (NOISE)	16.34 ± 0.07	18.63 ± 0.16
TRANSDUCER	15.97 ± 0.28	$\mathbf{18.07 \pm 0.24}$

Table 2: *TIMIT Results with Hybrid Training.*

NETWORK	DEV PER TEST PER	DEV FER TEST FER	DEV CE TEST CE
DBRNN	19.91 ± 0.22 21.92 ± 0.35	30.82 ± 0.31 31.91 ± 0.47	1.07 ± 0.010 1.12 ± 0.014
DBLSTM	17.44 ± 0.156 19.34 ± 0.15	28.43 ± 0.14 29.55 ± 0.31	0.93 ± 0.011 0.98 ± 0.019
DBLSTM (NOISE)	16.11 ± 0.15 $\mathbf{17.99 \pm 0.13}$	26.64 ± 0.08 $\mathbf{27.88 \pm 0.16}$	0.88 ± 0.008 $\mathbf{0.93 \pm 0.004}$

Wall Street Journal Experiments

The second set of experiments were carried out on the Wall Street Journal (WSJ) speech corpus. The purpose was to gauge the suitability of hybrid DBLSTM-HMM for large vocabulary speech recognition, and in particular to compare with existing deep network and GMM benchmarks.

Table 3: *WSJ Results.*

SYSTEM	WER	FER	CE
DBLSTM	**11.7**	30.0	1.15
DBLSTM (NOISE)	12.0	**28.2**	**1.12**
DNN	12.3	44.6	1.68
SGMM [6]	13.1	–	–

Discussion

The experiments suggest that hybrid HMM-DBLSTM systems are as effective for phoneme recognition as end-to-end RNN training. They also suggest that DBLSTM can deliver a substantial advantage over deep feedforward networks as a frame-level acoustic classifier. However it is less clear whether improvements in classification are likely to lead to a significant improvement in word-level accuracy for large vocabulary systems. Indeed, we found that regularising DBLSTM led to a *decrease* in frame error rate and an *increase* in word error rate on the WSJ data. This highlights one of the fundamental problems with hybrid systems: the frame-level distribution the network is trained to optimise is significantly different from the sequence level distribution that is implicitly defined by the decoding lattice.

An obvious direction for future work would be to explore the use of full-sequence, large-vocabulary training for DBLSTM networks, either by extending the existing end-to-end methods for RNNs to incorporate pronunciation dictionaries and language models, or by adapting the discriminative, full-sequence methods currently in use for hybrid systems [7].

References

[1] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc ICASSP 2013*, Vancouver, Canada, May 2013.

[2] M. Schuster and K. K. Paliwal, "Bidirectional Recurrent Neural Networks," *IEEE Transactions on Signal Processing*, vol. 45, pp. 2673–2681, 1997.

[3] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[4] H.A. Bourlard and N. Morgan, *Connnectionist Speech Recognition: A Hybrid Approach*, Kluwer Academic Publishers, 1994.

[5] Kam-Chuen Jim, C.L. Giles, and B.G. Horne, "An analysis of noise in recurrent neural networks: convergence and generalization," *Neural Networks, IEEE Transactions on*, vol. 7, no. 6, pp. 1424 –1438, nov 1996.

[6] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. Dec. 2011, IEEE Signal Processing Society.

[7] Abdel rahman Mohamed, Dong Yu, and Li Deng, "Investigation of full-sequence training of deep belief networks for speech recognition," in *in Interspeech*, 2010.