Residual Bidirectional LSTM + Conditional Random Fields + Expanding Batch Sizes for Part of Speech Tagging

Tobi Akomolede
Master's in Computer Science
University of Southern California
akomoled@usc.edu

Abstract

This approach combines Residual Bidirectional LSTM Networks with Conditional Random Fields to for the part of speech tagging task in the CSCI 544 Deep Learning competition. Batch sizes are increased during training. This method achieved fourth place out of over 150 competitors.

1 Embedding

This approach trains word embeddings with a unit size of 300. Tokens, which are represented by their numerical index within the preset vocabulary, are mapped to an embedding using TensorFlow's "embedding_lookup" function.

2 Residual Bidirectional LSTM-CRF

As in Huang et al. [2015], this tagging approach uses a Bidirectional LSTM network. As in Wang et al. [2018], I use a two-layer deep stacked residual LSTM structure:

The word embeddings are input into the first BiLSTM. The output h_1 of the first BiLSTM is input into the second BiLSTM. The outputs of the BiLSTMS are combined under the following formulation:

$h_3 = LeakyRelu(h_1 + h2)$

The output h_3 is then input to a fully connected layer with dimensionality M, where M is the number of tags.

The number of hidden units for both BiLSTMs is 100.

3 Viterbi Decoding

Rather than simply Softmax-ing the output from the final layers to determine the part-of-speech tag, this approach uses CRF-decoding, as in Huang et al. [2015]. The conditional random field works at a sentence level rather than a token level. The probability of each tag is conditioned on the input token and the neighboring tags. The most likely sentence is chosen using the Viterbi decoding algorithm

4 Expanding Batch Sizes

Similar to Goyal et al. [2017], this approach uses a warmup procedure during training. However, rather than increase the learning rate, the learning rate is decreased and the batch size increased. The

Preprint. Work in progress.

first three epochs are trained at a learning rate of 0.005e-3 with batch size of 32. The remaining epochs are trained with a learning rate of 0.00071 with batch size of 128

References

Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991, 2015. URL `http://arxiv.org/abs/1508.01991`.

Zhongjing Wang, Bo Peng, and Xuejie Zhang. Using a stacked residual lstm model for sentiment intensity prediction. *Neurocomputing*, 322:93–101, 2018.

Priya Goyal, Piotr Dollár, Ross B. Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch SGD: training imagenet in 1 hour. *CoRR*, abs/1706.02677, 2017. URL `http://arxiv.org/abs/1706.02677`.