Natural Language Inference Using Conditional Encoding

M. Salih Gerdan
Boğaziçi University

Abstract

The problem of natural language inference is important for many other sub-tasks in NLP. This paper attempts to recreate the neural model proposed by Rocktäschel et al. (2015) to solve this problem. It is an LSTM based model that reads both sentences, instead of encoding the sentences separately. The resulting model achieved inferior performance despite using a similar setup.

1 Introduction

The task of Natural Language Inference (NLI) is a vital task of Natural Language Understanding. It involves determining whether two sentences (the *premise,* and the *hypothesis*) entail/contradict/are neutral to each other.

This task had been done with hand-crafted systems, until the release of the Stanford Natural Language Inference (SNLI) dataset (Bowman et al., 2015), which allowed the training of neural network based systems. In more recent years, more modern systems such as BERT has been used. However this paper will deal with the more classic Bidirectional LSTM based system that has been proposed by Rocktäschel et al. (2015).

LSTM systems can be applied to this task in various ways. One of the common approaches has been to separately encode the two sentences and then use a classifier on the results. What made Rocktäschel et al. method stand out was that it let a single continuum of LSTMs read both sentences together, thus allowing information flow from the premise sentence to the hypothesis sentence. It also made use of *attention*, experimenting with both a simple attention layer and what is called a *word-by-word attention* layer.

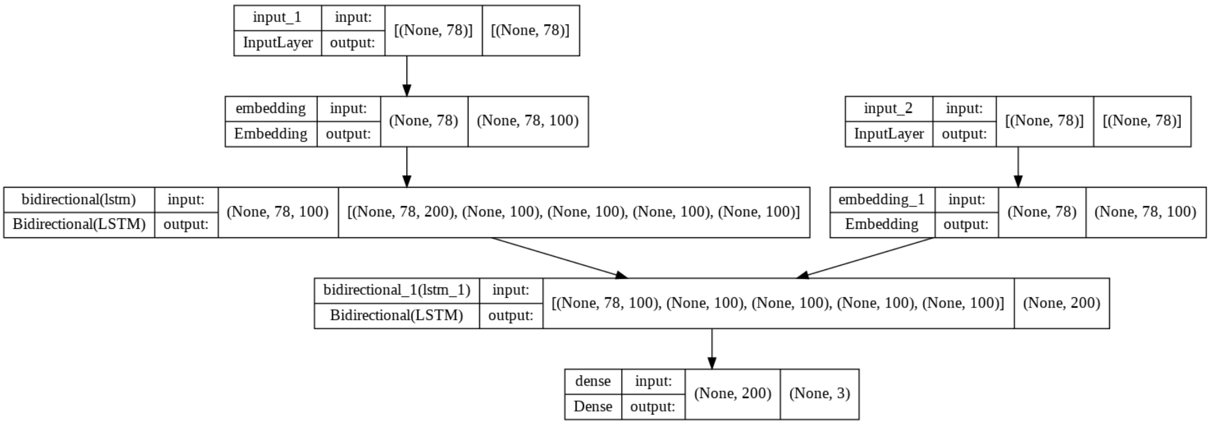
This idea of such a "conditional encoding" system has been utilized in the current paper. The attention system **was not** used.

The basic system using LSTMs with k=100 has performed with 72% accuracy on the SNLI dataset. This is very much inferior to the result achieved in the Rocktäschel et al. paper, which achieved 80.9% accuracy on the same dataset, without the attention layers. The possible causes of this is discussed further.

2 Methods

Two different biLSTM networks have been used for reading the premise and the hypothesis. When the first LSTM finishes reading the premise, the second LSTM with different parameters is initialized with the last state of the first LSTM. This achieves seamless flow of information between the networks, while being trained differently.

2.1 Bidirectional LSTMs



LSTMs (Hochreiter and Schmidhuber, 1997) have been preferred for their ease of use and proven performance.

The flow of information between the A Tensorflow graph of the model is given above. The arrow drawn from the first biLSTM into bidirectional_1(lstm_1) only shows the flow of LSTM state. The arrow from embedding_1 shows the actual input into the second biLSTM.

2.2 Embeddings

100-dimension GloVe vectors (Pennington et al.) has been used for word embeddings. This differs from the word2vec embeddings used in the original Rocktäschel et al. system.

GloVe has been preferred for its ease of use and availability.

3 Experiments

The Tensorflow library (Martin et al., 2015) is used to implement the system on Google Colaboratory.

Two systems have been constructed. One using k=100 for LSTMs and 100-dimension GloVe vectors. Another system with k=300 LSTMs with 300-dimension GloVe vectors have been trained.

However, the system with the larger dimensions took a much longer time to train, yet failed to provide any improvement over the 100 dimension system.

System	**LSTM size (k)**	**GloVe dimensions**	**Accuracy**
Baseline(Rocktäschel et al.)	100		80.9%
Baseline(Bowman et al.)	100		77.6%
basic	100	100	72.5%
basic_300	300	300	70.1%

Table 1: The systems

The 300-size system is ignored for the rest of the paper. The best result has been achieved with the 100-size system. The following figure shows its accuracy growth over training. We can see it achieved maturity at around epoch 12.

The cited accuracy rate for Rocktäschel et al. (2015) is only the system that does not incorporate attention. The best result in that paper is 83.5% using word-by-word attention.

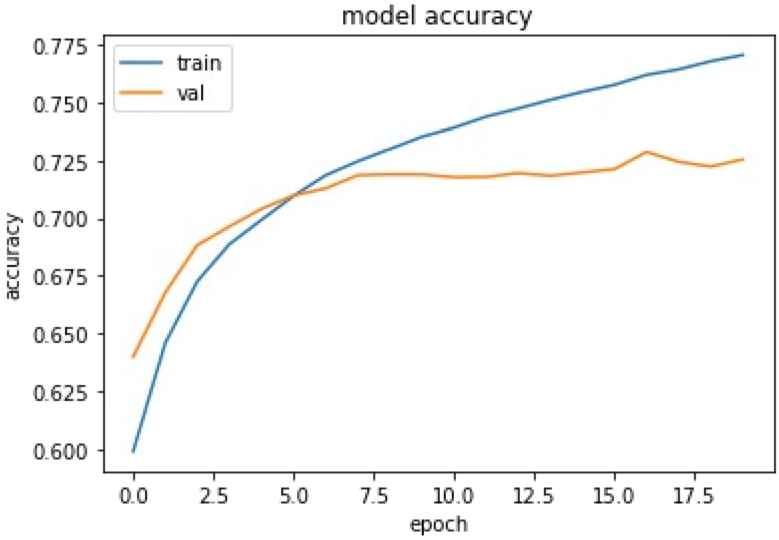


Figure 1: The training process of the 100-size system

4 Discussion

The failure of the 300-size system over the 100-size one shows that more dimensions does not always give improvements.

As for why the system failed to replicate the results of Rocktäschel et al. (2015) paper, one suspect is the use of different embeddings. The paper does mention having preferred word2vec over GloVe rather consciously, this may have been one of the reasons. However, as the current system underperforms Bowman et al. (2015) system which also used GloVe embeddings, this cannot be the blame alone.

The difference in systems likely lies in the difference in the details in the systems caused by different implementations.

Acknowledgements

Rocktäschel et al. for the basic idea of the system.

Reference

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Rafal Jozefowicz, Yangqing Jia, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Mike Schuster, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

Rocktäschel, T., Grefenstette, E., Hermann, K. M., Kočiský, T., & Blunsom, P. (2015). Reasoning about entailment with neural attention. *arXiv preprint arXiv:1509.06664*.