ATCS Assignment 1 - Error Analysis

Giulio Starace

A few observations can be made with regards to the learning curves shown in Fig. 1. First, from the validation accuracy curves we can see a clear difference in performance between the baseline model and the LSTM models, with the Max-pooled BiLSTM achieving the highest validation accuracy throughout. From these curves it is interesting to note the similarity in performance between the LSTM and the 'plain' BiLSTM, which actually does marginally worse. It is unclear why this occurs, especially since the BiLSTM contains the same information as the LSTM and could just learn to discard the additional concatenated information from the reversed LSTM. This may be explained by overfitting, with further evidence from the BiLSTM training accuracy being higher than the LSTM.

One may argue that the stopping criterion outlined in the paper was a bit too lenient and users may have benefited time-wise from more informed criterions. The behaviour of the validation loss curves in the LSTM-based models, with an initial dip followed by a slow rise may have been useful, particularly in combination with the validation accuracy curves. Perhaps the authors intuited that longer training could lead to better generalized sentence encoders at the expense of slightly worse NLI performance. This is similar reasoning as for why the same encoder is used for hypothesis and premise as opposed to using two separate specialized encoders.

For evaluation, we test our models both on the original SNLI task, as well as the SentEval sentence embedding evaluation suite. Table 1 shows the validation (a.k.a. "dev") and test accuracy of the four models implemented in this repository on the NLI task, as evaluated on the SNLI dataset. It also shows the micro and macro validation accuracy across the SentEval tasks outlined above. A few remarks can be made.

Table 1: Partial Replication of Table 3 of Conneau et al

	dim	**NLI**		**Transfer**	
		dev	**test**	**micro**	**macro**
Model					
Baseline	300	65.7	65.6	80.5	79.2
LSTM	2048	81.1	80.9	77.2	76.5
BiLSTM	4096	80.7	80.7	79.5	79.0
BiLSTM-Max	4096	**84.3**	**84.2**	**81.2**	**80.7**

Firstly, we see that the trends from validation accuracy on SNLI are mirrored in the test accuracy, so the discussion from the validation curves above holds. With regards to replication, for the LSTM and BiLSTM-Max model test accuracy we are somewhat close to the original results.

What is perhaps more interesting however is how the models perform on the SentEval transfer tasks. We see that while the BiLSTM-Max model still achieves the highest performance, the range across the various architectures is now much more compact. This seems to suggest that the underlying force dominating sentence-embedding performance is some aspect of the architecture that is shared across all variants. This theory may explain why the Baseline model, almost entirely based on the GloVe word-embeddings used in all models, performs so comparatively well in this case, coming second only to the best model.

If this were indeed the case, it would suggest that word-order is not properly learned in the LSTM models, despite the sequential nature of their learning. This is verified in fig. 2, where the embeddings of two sentences whose meanings are opposite but words are the same are compared. If word order matters, the resulting word-embeddings should differ, i.e. the difference should be non-zero.

While the difference is indeed non-zero, we do note that the embeddings are quite similar despite the sentences having opposite meanings. This can be interpreted as word order not fully being exploited.

This may be a limit of the models employed, which at best visit sentences sequentially in both directions, but cannot examine the sentences as a graph to construct inner representations akin to a parse tree, where word order becomes increasingly useful. At the expense of triteness, it would be interesting to extend the InferSent architecture with a transformer-based, BERT-like encoder for the sentences, to examine whether self-attention could be leveraged for improved SentEval performance.

Appendix

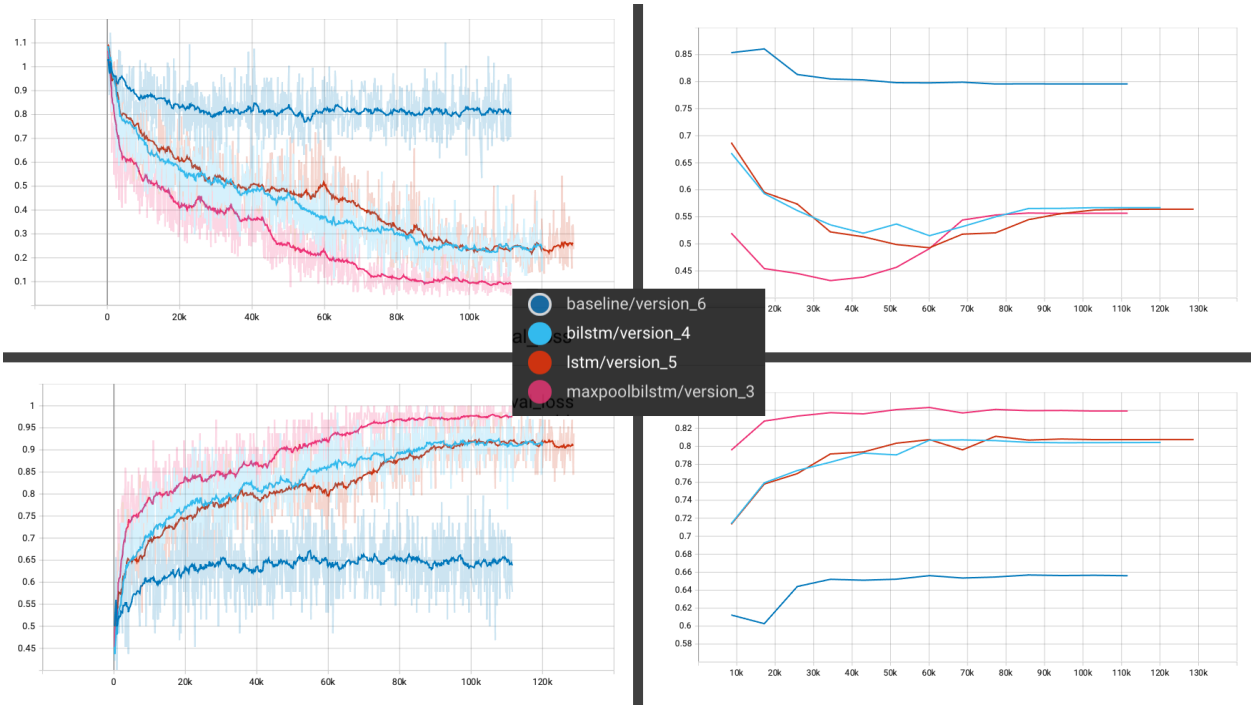

Figure 1: Training (left) and Validation (right) accuracy (bottom) and loss (top) curves of the four models implemented.

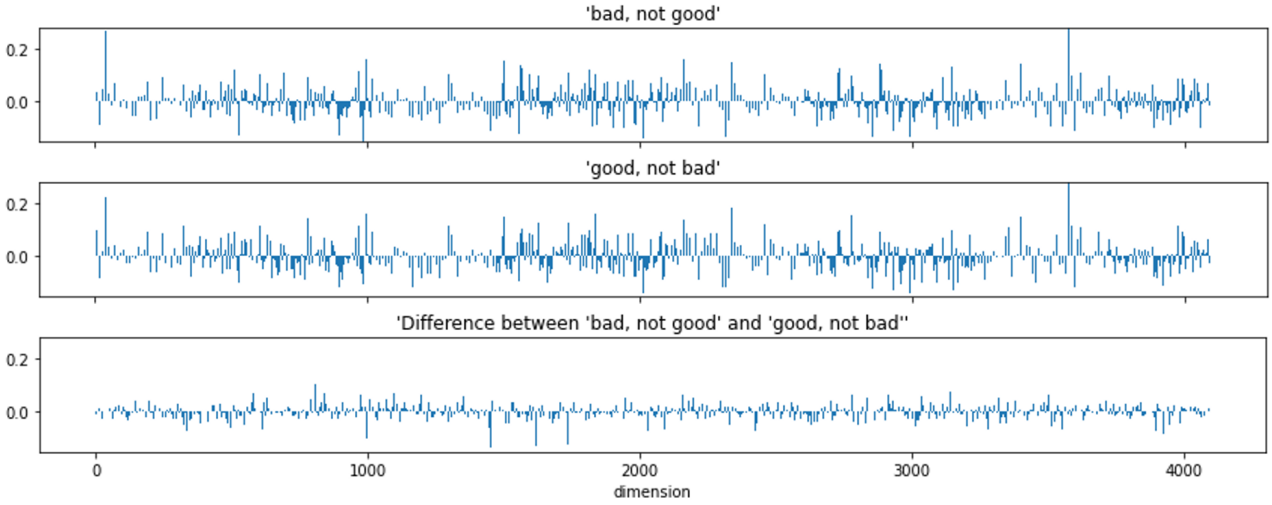

Figure 2: Visualization of sentence embedding of two sentences with opposite meanings using identical sets of words