ATCS practical-1

David Werkhoven 13909495

April 2025

1 Introduction

Understanding whether one sentence logically follows another requires reasoning, something humans do naturally. For AI models, however, reasoning is more challenging, as it involves understanding the full context of a given situation. To address this, [2] introduced a method using various sentence encoders for inference tasks. This work replicates their approach, focusing on four encoders: a mean encoder, unidirectional LSTM, bidirectional LSTM, and bidirectional LSTM with max pooling. After replication, results are gathered, and an error analysis is performed on the trained models.

2 Experimental Setup

The experimental setup follows that of [2]. All models are trained on the SNLI dataset [1], using stochastic gradient descent (SGD) with a learning rate of 0.1 and a decay rate of 0.99. After each epoch, the model is evaluated on the validation set. If the accuracy drops compared to the previous epoch, the learning rate is divided by 5 to help speed up convergence. The models are trained with a batch size of 64, and training stops early if the learning rate drops below 10^{-5}. The classifier is a one-layer MLP with a hidden size of 512. The input size for the LSTMs is set to 300 to match the GloVe embedding dimensions, and the hidden size is 512. A fixed random seed (1234) is used to ensure the results can be replicated.

The GloVe embeddings used are the 300-dimensional vectors trained on Common Crawl (840B tokens). These embeddings are kept fixed during training. To speed up training, only the words present in the SNLI dataset are included. This means any words not found in SNLI are ignored. While this choice helps reduce training time, using a larger vocabulary would likely improve performance and is recommended for addtional research.

3 Results and Analysis

Table 1 shows the results on the SNLI and SentEval datasets. The baseline mean model reaches 64.91% accuracy on SNLI, while LSTM models outperform it by about 20%, this shows the benefit of learning better. Interestingly, the mean model does better on SentEval than the LSTM and BiLSTM model without max pooling, likely due to the LSTMs limited vocabulary which was mentioned in section two. None of the models improved after 8 epochs, suggesting they converge quickly and that the learning rate adjustments and early stopping are effective.

Model	**Epoch**	**SNLI dev**	**SNLI test**	**SentEval micro**	**SentEval macro**
Mean	8	64.94	64.91	82.62	82.19
LSTM	8	83.05	82.70	80.25	79.99
BiLSTM	8	83.32	82.99	81.38	81.02
BiLSTM-Max	5	84.75	85.01	85.31	85.27

Table 1: The model results for the SNLI and SentEval datasets.

3.1 Error analysis

In Table 2, we see the examples used for the error analysis. All LSTM-based models performed reasonably well on these examples. However, the first example proved to be the most challenging. It contains lexical overlap between apple and fruit, and requires detecting the contradiction between a woman and no one. The models failed to recognize this contradiction, this implies they struggle with deeper semantic understanding and context-based reasoning.

Premise	**Hypothesis**
A woman is eating an apple	No one is eating fruit
A baby is crying in a crib	The baby is talking to its parents
The man was eating dinner when the phone rang	The man had already finished eating
Children are playing in a park	Kids are studying at school
A man is giving a presentation to a small audience	A man is speaking in front of a group

Table 2: Examples used in the error analysis

4 Conclusion

In conclusion, the replication of [2] was successful, with accuracy values close to those reported in the original work. While the models handle basic context and simple NLI tasks well, they struggle with deeper semantic understanding and complex reasoning. This is likely due to the simple classifier used. Adding more layers and non-linear activation functions could improve performance on more challenging inference cases.

References

[1] Samuel R Bowman et al. "A large annotated corpus for learning natural language inference". In: *arXiv preprint arXiv:1508.05326* (2015).

[2] Alexis Conneau et al. "Supervised learning of universal sentence representations from natural language inference data". In: *arXiv preprint arXiv:1705.02364* (2017).