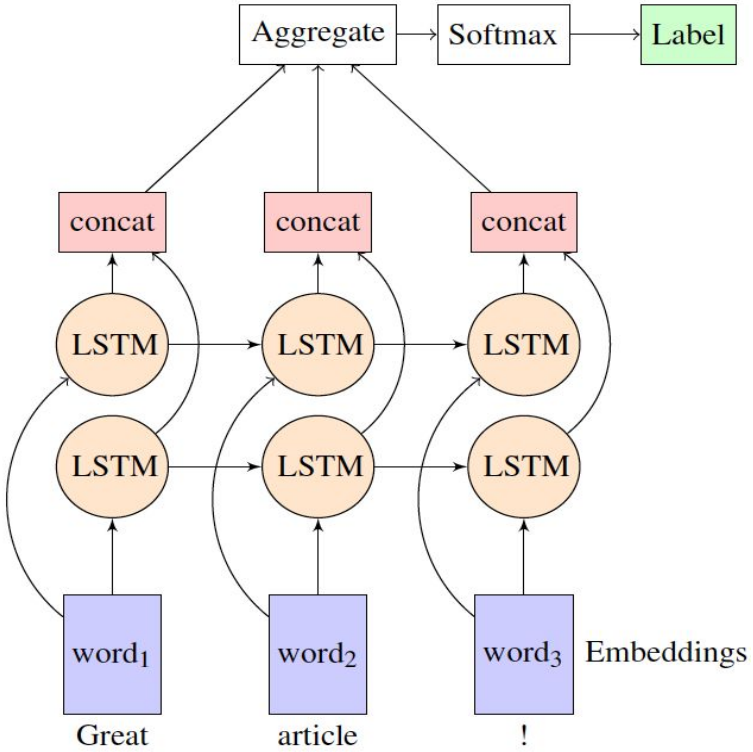
ADL Project-II

Mayank Gupta

Background



Bi-directional LSTM is used in a lot of tasks in NLP.

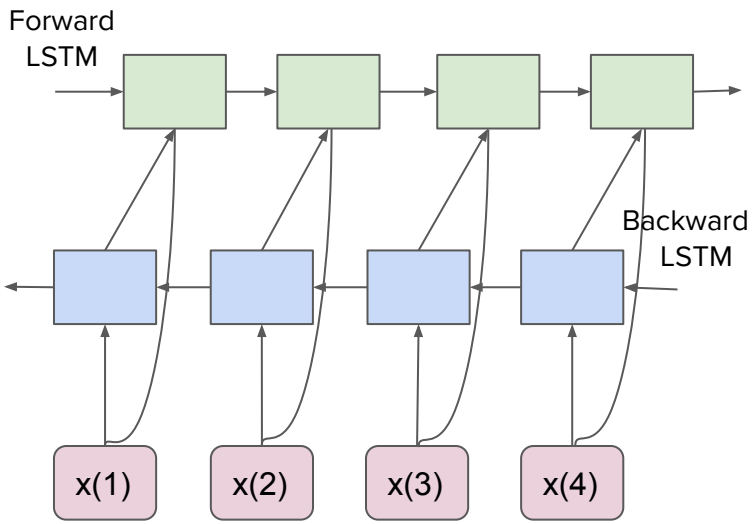
But in bi-directional LSTM both forward and backward LSTM run independently and final hidden state is just concatenation of forward and backward hidden states.

At each time-step LSTMs do not take into account future symbols while predicting current hidden state.

Proposed Approach

- In this work, I propose a new bi-directional LSTM network in which forward LSTM takes into account the complete input sequence.
- In the proposed architecture the forward and backward hidden states are not independent. This is achieved by feeding the hidden states of the backward LSTM as input to the forward LSTM along with the current input vector.
- I also implemented pre-training of new LSTM architecture using Masked Language modeling and Next Sentence Prediction so that it can be fine-tuned for other downstream tasks.

Technical Details



- In the new LSTM, hidden state of the backward LSTM is fed to the forward LSTM along with the current input vector, so that output hidden state of the forward LSTM depends on the complete input sequence.
- Pre-training of this model is done for MLM and NSP tasks. In MLM, we mask 15% of the tokens randomly and our task is to predict masked tokens. In NSP, we are given 2 sentences and have to classify if the second sentence is next to first sentence or not. I use vocabulary of 10,000 words for pre-training task.
- Details of token, segment embeddings and padding for pre-training task is similar to the BERT architecture.

Contributions(Novelty)

- Proposed a new type of LSTM architecture which better utilizes left and right context while predicting an output hidden state.
- Implemented pre-training of the new LSTM architecture using Masked Language modeling and Next Sentence Prediction tasks.

Results & Conclusion

Dataset used	New LSTM		LSTM	
	Loss	Accuracy	Loss	Accuracy
IMDB	0.351	85.71	0.368	84.3
SNLI	0.0023	77.8	0.0023	77.42
WikiText-2	5.803	-	5.837	-

Table 1. Comparison of models.

- 2-layer networks are used for experiments on IMDB and SNLI datasets, 1-layer networks are used for pre-training experiment on WikiText-2.
- New LSTM uses around 1.2x parameters than normal LSTM network in experiments.
- Loss of MLM task is reported in the table for WikiText-2 task, loss for NSP was similar for both the models. We can see from the table that performance of both models is similar.