0.1 Results - Test Split SNLI - Accuracy in %

Baseline: 35.44
Unidirectional LSTM: 32.37
BiLSTM: 32.37
BiLSTM with max-pooling: 32.37

0.2 Results - VAL Split SNLI - Accuracy in %

Baseline: 34.87
Unidirectional LSTM: 32.78
BiLSTM: 32.78
BiLSTM with max-pooling: 32.78

0.3 SentEval - MACRO Accuracy in %

Baseline: 53.79
Unidirectional LSTM: 53.79
BiLSTM: 53.79
BiLSTM with max-pooling: 53.79

0.4 SentEval - MICRO Accuracy in %

Baseline: 53.89
Unidirectional LSTM: 53.89
BiLSTM: 53.89
BiLSTM with max-pooling: 53.89

0.5 SentEval - Task Wise Accuracy in % - Baseline

Task: MR
Accuracy: 50.0
Dev samples: 10662
Task: CR
Accuracy: 63.76
Dev samples: 3775
Task: SUBJ
Accuracy: 50.0
Dev samples: 10000
Task: MPQA
Accuracy: 68.77
Dev samples: 10606
Task: SST2
Accuracy: 50.92
Dev samples: 872
Task: TREC
Accuracy: 22.93
Dev samples: 5452
Task: MRPC
Accuracy: 67.54
Dev samples: 4076

Task: SICKEntailment
Accuracy: 56.4
Dev samples: 500

0.6 SentEval - Task Wise Accuracy in % - Unidirectional LSTM

Task: MR
Accuracy: 50.0
Dev samples: 10662
Task: CR
Accuracy: 63.76
Dev samples: 3775
Task: SUBJ
Accuracy: 50.0
Dev samples: 10000
Task: MPQA
Accuracy: 68.77
Dev samples: 10606
Task: SST2
Accuracy: 50.92
Dev samples: 872
Task: TREC
Accuracy: 22.93
Dev samples: 5452
Task: MRPC
Accuracy: 67.54
Dev samples: 4076
Task: SICKEntailment
Accuracy: 56.4
Dev samples: 500

0.7 SentEval - Task Wise Accuracy in % - Bi-LSTM

Task: MR
Accuracy: 50.0
Dev samples: 10662
Task: CR
Accuracy: 63.76
Dev samples: 3775
Task: SUBJ
Accuracy: 50.0
Dev samples: 10000
Task: MPQA
Accuracy: 68.77
Dev samples: 10606
Task: SST2
Accuracy: 50.92
Dev samples: 872
Task: TREC

Accuracy: 22.93
Dev samples: 5452
Task: MRPC
Accuracy: 67.54
Dev samples: 4076
Task: SICKEntailment
Accuracy: 56.4
Dev samples: 500

0.8 SentEval - Task Wise Accuracy in % - Bi-LSTM with max-pooling

Task: MR
Accuracy: 50.0
Dev samples: 10662
Task: CR
Accuracy: 63.76
Dev samples: 3775
Task: SUBJ
Accuracy: 50.0
Dev samples: 10000
Task: MPQA
Accuracy: 68.77
Dev samples: 10606
Task: SST2
Accuracy: 50.92
Dev samples: 872
Task: TREC
Accuracy: 22.93
Dev samples: 5452
Task: MRPC
Accuracy: 67.54
Dev samples: 4076
Task: SICKEntailment
Accuracy: 56.4
Dev samples: 500

0.9 Comments

- All the sentence embeddings seem to be equally good in the transfer tasks, which result in them having similar MAX and MICRO accuracies.
- The baseline checkpoint seems to have the best accuracy compared to other models in both Test and Val splits of SNLI. Other models terminated early in few time-steps with the given hyper-parameters and convergence criteria as mentioned in the paper.

0.10 Notebook Analysis

- As highlighted, all the LSTM based models predicted contradiction, whereas baseline predicts entailment.

- In the first example, the models predicted Contradiction likely due to the contradictory nature of the phrase "Two men sitting" and "Nobody is sitting". The models missed the representation for sun and shade.

- In the second example, the models predicted Contradiction likely due to the contradictory nature of the phrase "A man is walking" and "No cat is outside". The models thought that walking is a thing usually done outside, but the hypothesis highlights that no cat is outside. Model failed to identify the difference between a dog being walked outside vs no cat present outside. Model assumes a negative correlation/contradiction, whereas in reality, the premise and hypothesis are completely independent of each other and hence neutral.