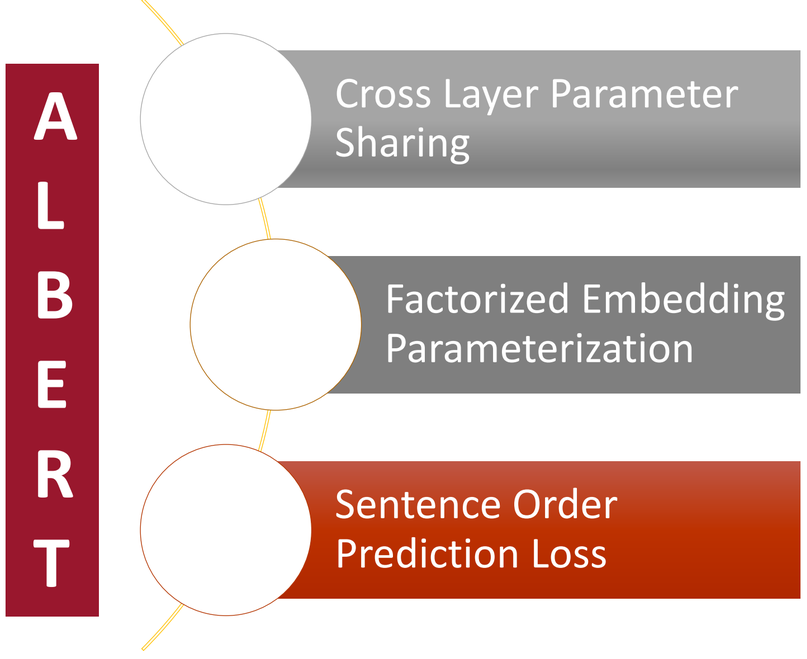
Sentence Order Prediction Variants and their evaluation on downstream tasks using BERT

Ritesh Kumar, Nikita Jaiman, Manisha Kumari Barnwal, Vasishtha Sriram Jayapati, Ajay Venkitaraman
University of Massachusetts Amherst

Goal: Implement and evaluate variants of Sentence Order Prediction Loss proposed in ALBERT



Motivation: The effect of SOP on the accuracy of ALBERT model for downstream tasks is not discussed in the paper.

1. Embedded **only SoP loss in BERT** to understand its impact on the performance over the GLUE dataset tasks.
2. **Implemented two variants** of the SoP loss and compared against the **BERT+ SoP** performance.

SoP Variants

SoP1: Up to 4 sentences

Extension of SoP from ALBERT

Relative position of any 2 sentences in the unlabeled wikitext-103 dataset.

SoP Classifier
[S_a, S_b , <rel. dist. b/w S_a, S_b >]

Example:
*{S_1,S_2, **1**} , {S_1,S_3, **2**}, {S_1,S_4, **3**}, {S_1,S_5, **4**}*

SoP2: Paragraph Context

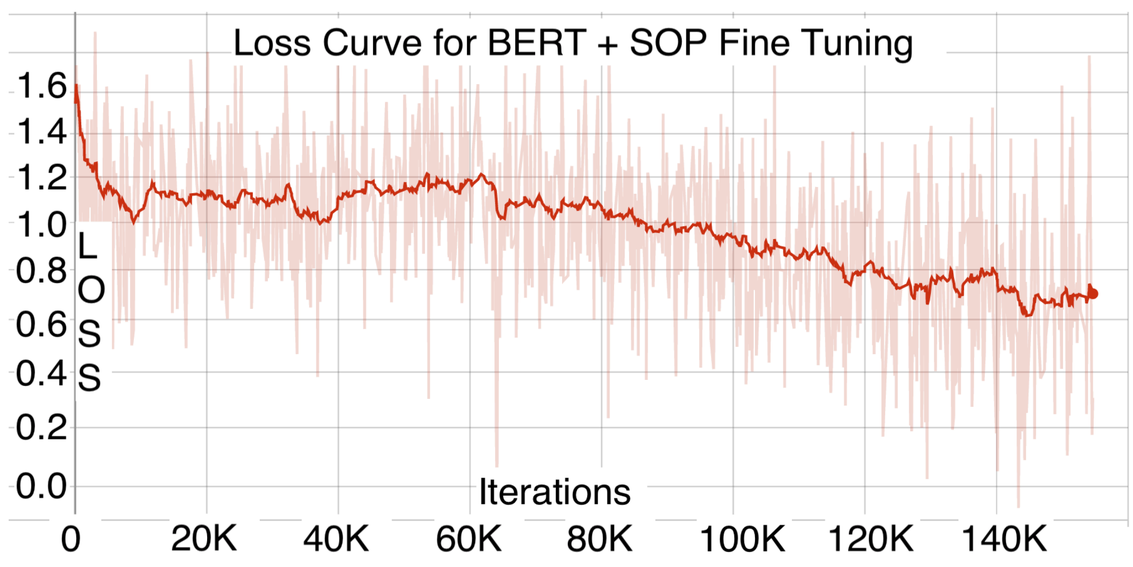
Extension of SoP from ALBERT

Trained positive and negative samples of <paragraph, sentence> pairs on unlabeled wikitext-103 dataset where the output denoted the inclusion of the sentence in that paragraph.

Example:
*{P_1 ,S_2, **1**} , {P_1 ,S_{33} , **0**}*

Results: Implemented and trained BERT with SoP variants from scratch as Fine-Tuning on top of BERT did not perform well. We considered BERT + SoP as the baseline for evaluation.

Dataset	Accuracy		
GLUE	*SoP*	*SoP1*	*SOP2*
MRPC	70.7	75.6	68.4
STSB	71.8	76.2	
QQP	74.4	76.3	63.2
MNLI(m)	67.2	**79.6**	32.7
QNLI	75.9	78.1	49.5
RTE	68.1	**77.8**	52.7
WNLI	69.3	**76.1**	56.3



BERT + SOP Fine tuning- Model doesn't converge

Conclusion:

1. SoP1 significantly improved the accuracies for all the tasks under GLUE dataset.
2. SoP2 gave comparable performance for the tasks under the GLUE dataset.

In Progress: Evaluate for text summarization task using CNN/Daily Mail dataset.