**Problem**:

- Increase in model size comes with dependency on GPU/TPU, increase in training time.
- Increase in parameters doesn't necessarily leads to better performance (demonstrated through Bert-xLarge, where hidden layer units are 2048 as compared to 1048 of Bert-Large)

**Contribution/Proposal:**

- Parameter reduction technique to lower memory consumption/increase training speed.
- Self-supervised loss that focuses on modelling inter-sentence coherence: Sentence Order Prediction
- Embedding factorization: Removes embedding layer size dependency on Hidden layer size. Embeddings and hidden layer have now an addition matrix in between to project the embeddings from their space to hidden layer space. (|V| * E) X (E * H)

**Method**:

- Built on base architecture of BERT/Transformer with Gelu nonlinearities.
- Shared parameters: reuse parameters of layers of transformer instead of having unique weights for each layers. The layers that share weight could be:
    - Complete block
    - Only attention
    - Shared-FFN
- Next Sentence Prediction: Binary classification to tell current sentence comes after a previous sentence. (TRADITIONALLY USED)
- Sentence Order Prediction (SOP): Novel of ALBERT, swaps order of two sentences and predicts whether the order is correct or not.

**Results**:

- ALBERT with different parameters is compared with BERT base and Large : They beat SOTA with a very large model that doesn't speedup. They have ~2 accuracy dip with BERT base with training speed up and roughly 10% parameters.
- Shared parameters lead to a rough ~2 acc. Dip but over 80% model size reduction.
- More similar results
- Their training task SOP leads to better performance on downstream task as compared to BERT's training task of NSP.

**Comments/Feedback**:

- Code available publicly

- Parameter sharing is not that new, it has been used previously as compression technique before also.
- Embedding factorization is also not new. However, the extensive experiments they conduct with the above two factors, leads to a good empirical evidence of their hypothesis.
- They post hyper parameters in appendix, so I believe replicating results may not be an issue