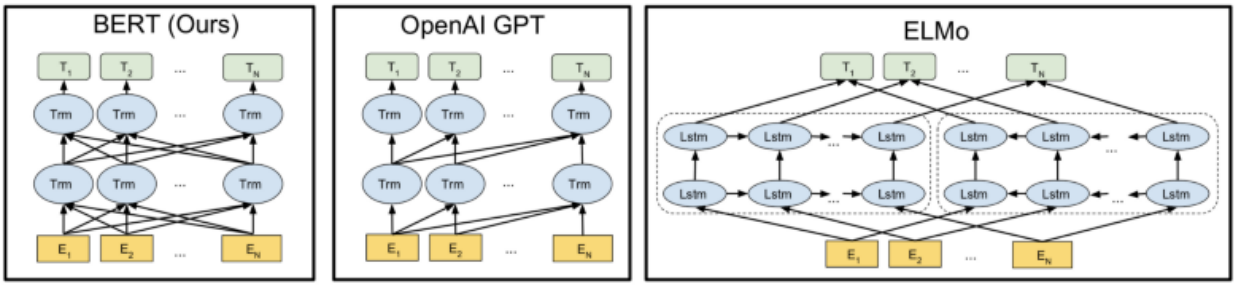
Overview of BERT's Smaller Sized Variants: ALBERT vs DistilBERT vs TinyBert.

Introduction

The world of NLP encoder based tasks have not been the same since the introduction of BERT. While BERT broke boundaries in Sequence-to-sequence tasks and Natural Language Understanding tasks such as QnA, Natural language inference and Sentiment classification. They have been relatively large and difficult to productize without deep pockets, cost to users or even harm to the environment. BERT as a model needs about 450 MB of RAM to be hosted, not considering the network I/O bottleneck. This problem ushered the need for smaller models like ALBERT, DistilBERT and TinyBERT. The goal of this review is to review these attempts to miniaturize BERT and the challenges they face at doing that.

Overview of BERT

BERT: Bidirectional Encoder Representations from Transformers came out in 2018 showing that a Pre-trained Language Model can achieve state of the art results in various NLP tasks. BERT was able to overcome its predecessors by replacing the unidirectional behavior of Left to Right by randomly masks some of the tokens from the input using the "masked language model" (MLM) to predict the original approach based on its context. And Secondly introducing a concept of NSP Next Sentence Prediction which essentially means that given the current sentence, can we predict the next sentence.



As from the image above we can see BERT being bidirectional, OpenAI gPT being Left-to-Right and ELMO a concatenation of left and right LSTMs not a fusion of Transformers like BERT.

Why are the Models so Large

BERT was trained on Wikipedia and BooksCorpus (800M words) (Zhu et al., 2015) and English Wikipedia (2,500M words) and the final model has about 110M parameters and BERT large, a variant, has 340M parameters. To simplify, a floating point is worth 4 bytes in memory and 110M*4Bytes would require around 440MB storage needed in RAM to run a BERT model.

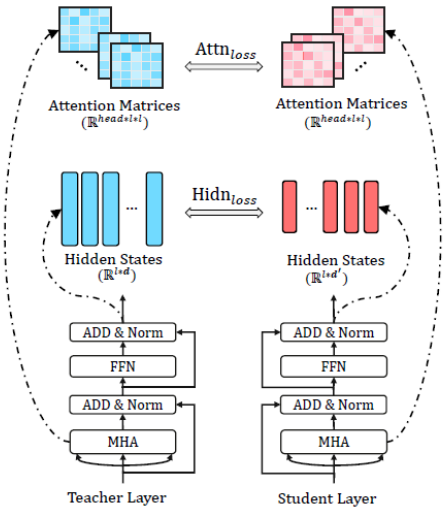
DistilBERT

Knowledge distillation was introduced to reduce the size of BERT. In the paper by Sanh et al, titled, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". Knowledge Distillation, KD is a compression technique in which a compact model - the student S - is trained to reproduce the behavior of a larger model - the teacher T - or an ensemble of models.
The goal is to emulate the output distribution of the teacher's model using the Kullback-Leibler divergence as a compression technique similar to what we covered in class..

$$KL(p||q) = \mathbb{E}_p(log(\frac{p}{q})) = \sum_i p_i * log(p_i) - \sum_i p_i * log(q_i)$$

From the experimentation, the number of layers was a major determinant of inference time, more than the hidden size. This huge reduction in size from 110M parameters to 66M (40% smaller) caused an increase in inference time to be 60% faster with only a drop to 97% of BERT performance(Sanh, et al).

TinyBERT



TinyBERT also used the Knowledge distillation approach but does it somewhat differently. Jiao et al in their paper use general distillation in pre-training and the task-specific distillation as it was observed that actual BERT is over-parameterized. The third approach was Data Augmentation as word level replacement was done using pre-trained BERT and GloVE (Pennington et al).
It was able to achieve more than 96.8% the performance of teacher BERT on General Language Understanding Evaluation (GLUE)(Wang et al) tasks with a penalty of performance. TinyBERT was able to achieve this using 14.5M parameters(~60MB of RAM space).

ALBERT:

At around the same time as DistilBERT and for the same reason, Zhenzhong et al introduced "*ALBERT: A Lite BERT for Self-supervised Learning of Language Representations"* . It does a few things different from BERT: projecting the embedding matrix to a lower dimensional dense matrix, replacement of the NSP with modeling inter-sentence coherence called Sentence-Order prediction, removing dropout, addition of 10X more data and cross-layer parameter sharing. All without the penalty of performance. In fact, ALBERT xlarge(60M parameters) outperforms BERT in SQuAD1.1, SQuAD2.0, MNLI, SST-2, and RACE benchmarks. Base ALBERT has about 12M parameters which is about 48MB of RAM space needed.

What are the Compromises?

Next Sentence Prediction: For DistilBERT, the *pooler* which is used for Next Sentence Prediction task is removed and the number of layers is reduced by half. Hence reducing what DistilBERT is capable of to MLM based tasks.
Inference Speed: ALBERT on memory is about 50 MB which is about the same size as TinyBERT. However, TinyBERT is about 10x faster than ALBERT(same speed as BERT).
Poor Performance: The Knowledge Distillation approaches – TinyBERT and DistilBERT – all performed poorly in general compared to actual BERT.

Discussion and Conclusion

We can see that different attempts to reduce current models have some sort of penalty. Either with reduction of capabilities, poorer performance or speed. Bringing NLP to be generally accessible remains a problem. The broad theme here among all the approaches also is transferred learning for NLU tasks is better than classical approaches. Which is good and bad. Good that we have it and bad that current State of the Art approaches can only be done by large corporations with such resources to undertake it. Hence, pushing groundbreaking deep learning away from the hands of most.

In conclusion, we have seen the various approaches at reducing BERT's memory and network footprint. With the cost of memory and advancement in computing this may not be much of an issue in the near term. However, with global warming and new SOTA models being about 500+ Billion parameters which translates to about 2TB RAM requirement, cutting the size of these models would soon be the focus of NLP as Moore's Law may be upon us if it isn't already.

References

A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

J. Pennington, R. Socher, and C. D. Manning. 2014. Glove: Global vectors for word representation. In EMNLP

Sanh, Victor et al. “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.” ArXiv abs/1910.01108 (2019): n. pag.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. *TinyBERT: Distilling BERT for Natural Language Understanding*. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 4163–4174, Online. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. *Aligning books and movies: Towards story-like visual explanations by watching movies and reading books*. In Proceedings of the IEEE international conference on computer vision, pages 19–27.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations.* In ICLR