BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Daniel Crockett

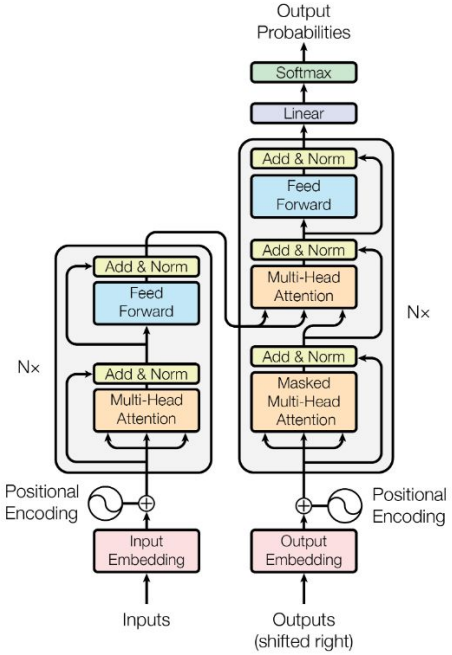
CONS

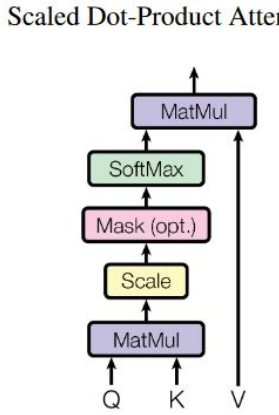
Lack of Model Description

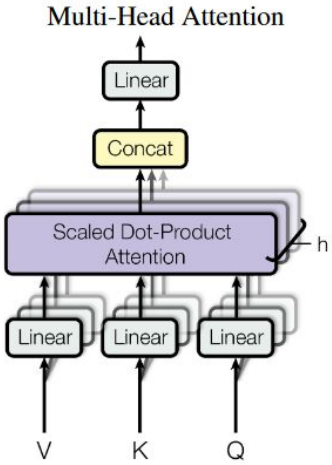
Essentially tell you to read “Attention Is All You Need” by Aswani, et al.

What does effectively identical mean?

Because the use of Transformers has become ubiquitous recently and our implementation is effectively identical to the original, we will omit an exhaustive background description of the model architecture and refer readers to Vaswani et al. (2017)







Unexplained Decisions

- WordPiece Embeddings
- Learned Positional Embeddings
- Adam Parameters for Pretraining
- GeLU for BERTlarge

Advertisement

Training of $\text{BERT}_{\text{BASE}}$ was performed on 4 Cloud TPUs in Pod configuration (16 TPU chips total).[5] Training of $\text{BERT}_{\text{LARGE}}$ was performed on 16 Cloud TPUs (64 TPU chips total). Each pre-training took 4 days to complete.

3.5 Fine-tuning Procedure

For sequence-level classification tasks, BERT fine-tuning is straightforward. In order to obtain a fixed-dimensional pooled representation of the input sequence, we take the final hidden state (i.e., the output of the Transformer) for the first token

[5] https://cloudplatform.googleblog.com/2018/06/Cloud-TPU-now-offers-preemptible-pricing-and-global-availability.html

Unfair Advantage

No comparison between BERT trained only on BooksCorpus and GPT

- Wikipedia - 2,500 million words
- BooksCorpus- 800 million words

Tasks	Dev Set MNLI-m (Acc)	QNLI (Acc)	MRPC (Acc)	SST-2 (Acc)	SQuAD (F1)
$\text{BERT}_{\text{BASE}}$	84.4	88.4	86.7	92.7	88.5
No NSP	83.9	84.9	86.5	92.6	87.9
LTR & No NSP	82.1	84.3	77.5	92.1	77.8
+ BiLSTM	82.1	84.1	75.7	91.6	84.9

Table 5: Ablation over the pre-training tasks using the $\text{BERT}_{\text{BASE}}$ architecture. "No NSP" is trained without the next sentence prediction task. "LTR & No NSP" is trained as a left-to-right LM without the next sentence prediction, like OpenAI GPT. "+ BiLSTM" adds a randomly initialized BiLSTM on top of the "LTR + No NSP" model during fine-tuning.

Lack of Future Work

“... important future work is to investigate the linguistic phenomena that may or may not be captured by BERT”