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

Presenter: Baosong Yang

Motivation

- Fine-tuning Approaches: pre-train some model on a LM objective.

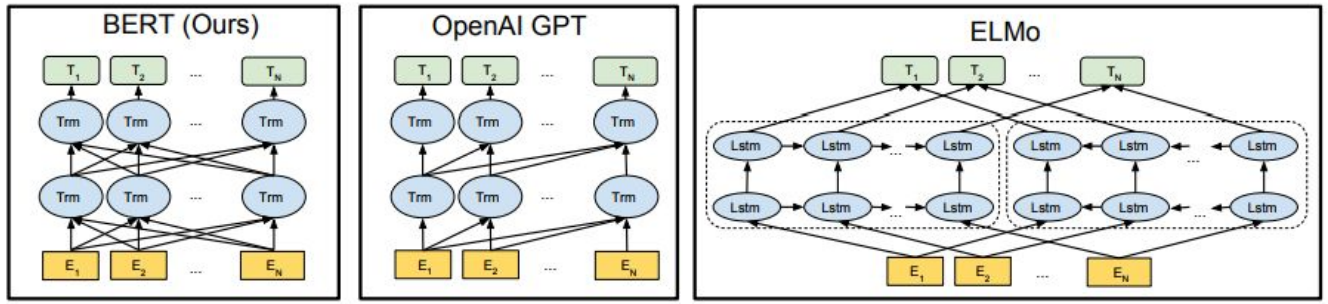


Figure 1: Differences in pre-training model architectures. BERT uses a bidirectional Transformer. OpenAI GPT uses a left-to-right Transformer. ELMo uses the concatenation of independently trained left-to-right and right-to-left LSTM to generate features for downstream tasks. Among three, only BERT representations are jointly conditioned on both left and right context in all layers.

- Considering forward and backward Information.
- ELMo: Bidirectional but independent $P(w_i|w_1, \ldots w_{i-1})$ $P(w_i|w_{i+1}, \ldots w_n)$

Model

- BERT-Base: L=12, H=768, A=12
- BERT-Big : L=24, H=1024, A=16

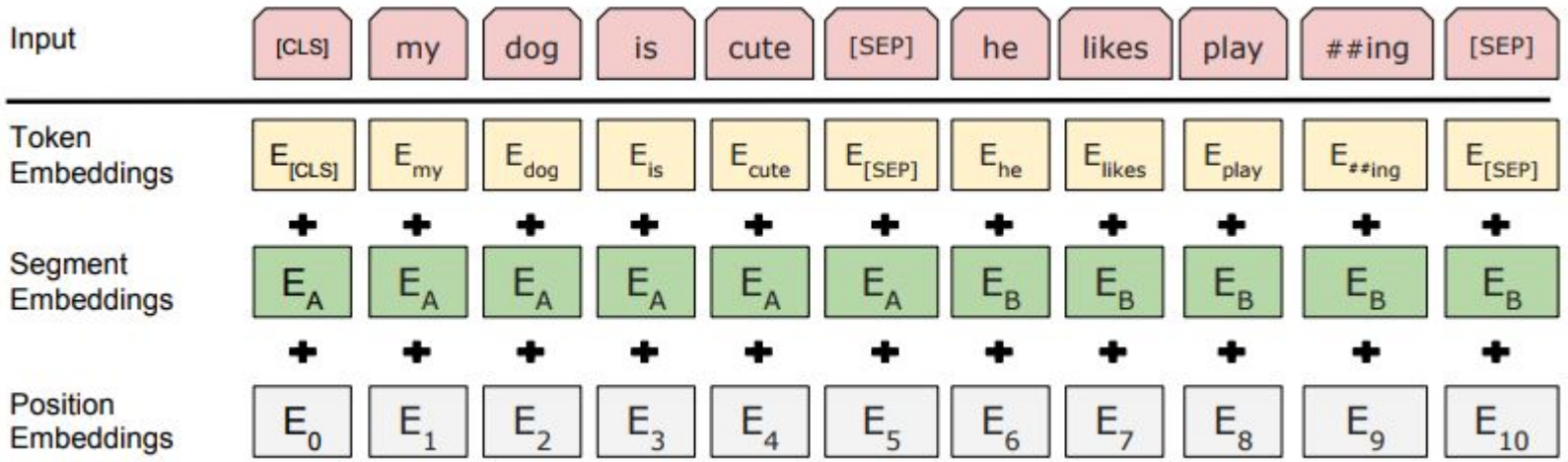
- Masked LM: predicting only those masked tokens
  - Previous approaches: can only be trained left-to-right or right-to-left, since bidirectional conditioning would allow each word to indirectly “see itself” in a multi-layered context.
  - Mask 15% tokens:
    - 80%: my dog is hairy → my dog is [MASK]
    - 10%: my dog is hairy → my dog is apple
    - 10%: my dog is hairy → my dog is hairy

Model

- Next Sentence Predction (QA and NLI):

Input = [CLS] the man went to [MASK] store [SEP]
he bought a gallon [MASK] milk [SEP]
Label = IsNext

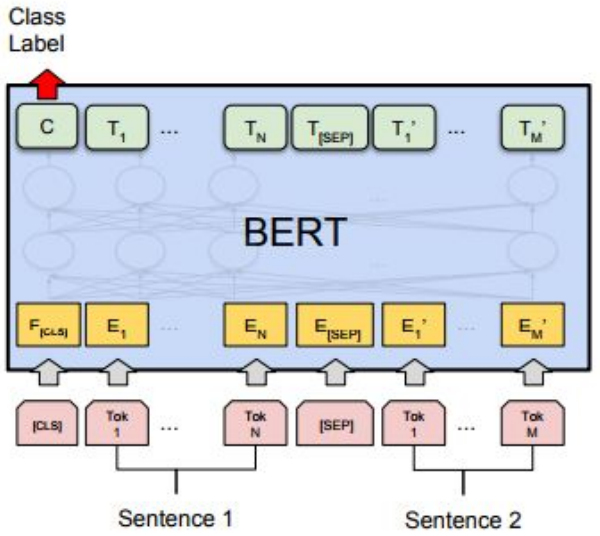
Input = [CLS] the man [MASK] to the store [SEP]
penguin [MASK] are flight ##less birds [SEP]
Label = NotNext



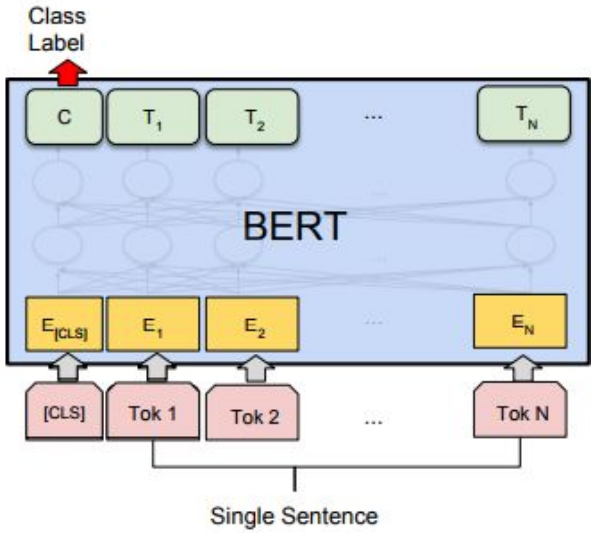
Model

- Fine-tuning:
  - QA: S and E are tranable parameters.

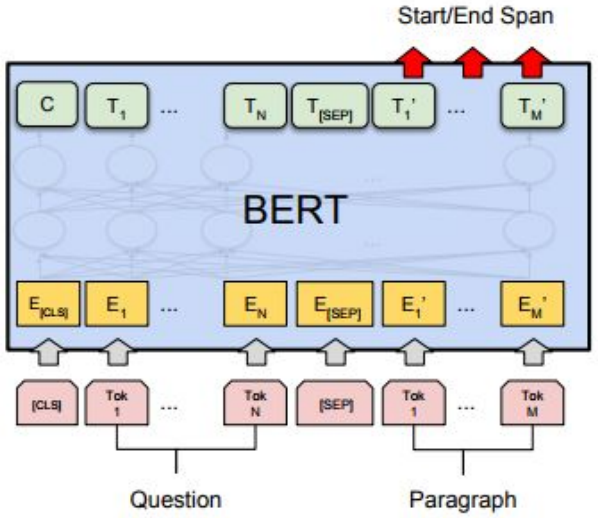
$$P_i = \frac{e^{S \cdot T_i}}{\sum_j e^{S \cdot T_j}}$$



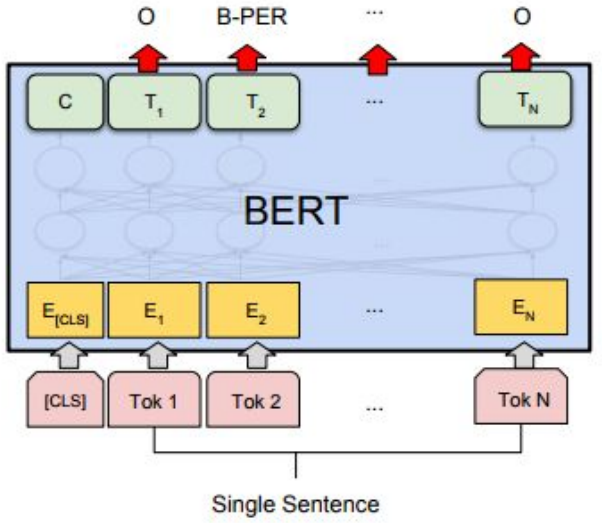
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



(b) Single Sentence Classification Tasks: SST-2, CoLA



(c) Question Answering Tasks: SQuAD v1.1



(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

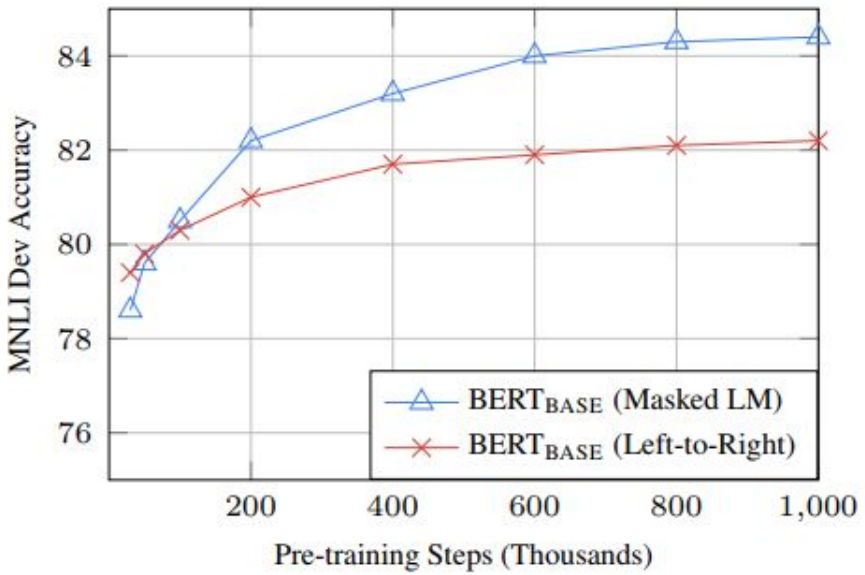
Experiments

- Pretraining:
  - Data: BooksCorpus (800M words) + Wikipedia (2500M words)
  - Batch size: 128,000 tokens/batch
  - Use gelu rather than relu
  - Vocabulary size??
- Fine-tuning:
  - Train the classification layer

Experiments

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

Hyperparams				Dev Set Accuracy		
#L	#H	#A	LM (ppl)	MNLI-m	MRPC	SST-2
3	768	12	5.84	77.9	79.8	88.4
6	768	3	5.24	80.6	82.2	90.7
6	768	12	4.68	81.9	84.8	91.3
12	768	12	3.99	84.4	86.7	92.9
12	1024	16	3.54	85.7	86.9	93.3
24	1024	16	3.23	86.6	87.8	93.7



Layers	Dev F1
Finetune All	96.4
First Layer (Embeddings)	91.0
Second-to-Last Hidden	95.6
Last Hidden	94.9
Sum Last Four Hidden	95.9
Concat Last Four Hidden	96.1
Sum All 12 Layers	95.5

Table 7: Ablation using BERT with a feature-based approach on CoNLL-2003 NER. The activations from the specified layers are combined and fed into a two-layer BiLSTM, without backpropagation to BERT.

Conclution

- Harder and harder and harder to reproduce SOTA.
- Waiting for the released codes and the data.

- Why performs better?
  - Large scale
  - A better object: $P(w_i|w_1,\ldots,w_{i-1},w_{i+1},\ldots,w_n)$

- Examined on classification tasks
  - Is it valuable to pretrain an reconstructor for MT.