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

Jacob Devlin, Ming-Wei Chang, Kenton Lee,
Kristina Toutanova

정영석

Contents

Sequence to Sequence

1. RNN Encoder-Decoder(최초의 신경망 기반 번역 모델)

- SMT 와 같은 통계기반 번역 모델의 한계를 극복함
- Rare word에 대한 대처, 문법적 표현이 통계기반 모델에 비해 부드러워짐
- 순환신경망 구조의 특성으로 긴 문장 처리에 한계가 존재

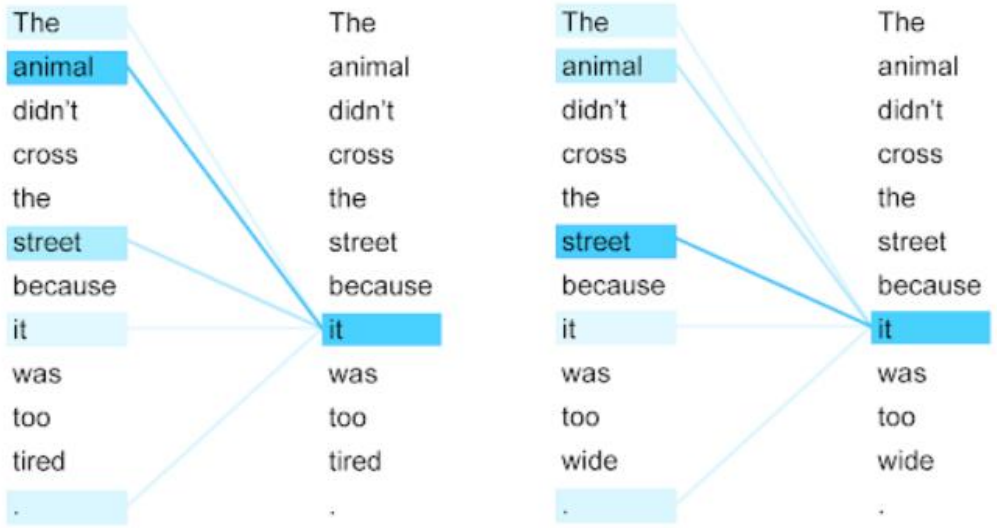
2. Attention based model (순환신경망 모델의 장기의존성 문제 개선)

- 매 시점 Decoder의 은닉상태와 Encoder 은닉상태의 Attention 연산을 통해 문장 전체의 의미를 참조하며 문장을 생성
- Attention 연산을 통해 순환신경망 구조의 장기 의존성 문제 개선

Sequence to Sequence

3. Transformer

- **Self-Attention**



- **Multi-Head Attention**

오는 7~8일 방한을 앞뒀던 마이크 폼페이오 미 국무장관이 일본·몽골·한국 아시아 3개 국 순방(4~8일) 일정을 재검토 중 이라고 밝혔다. 도널드 트럼프 미국 대통령이 신종 코로나바이러스 감염증(코로나19)에 확진된 데 따른 것이다.

ANSWER : 폼페이오 장관의 아시아 순방일정이 어떻게 되나요?

Introduction

▪ 연구 배경

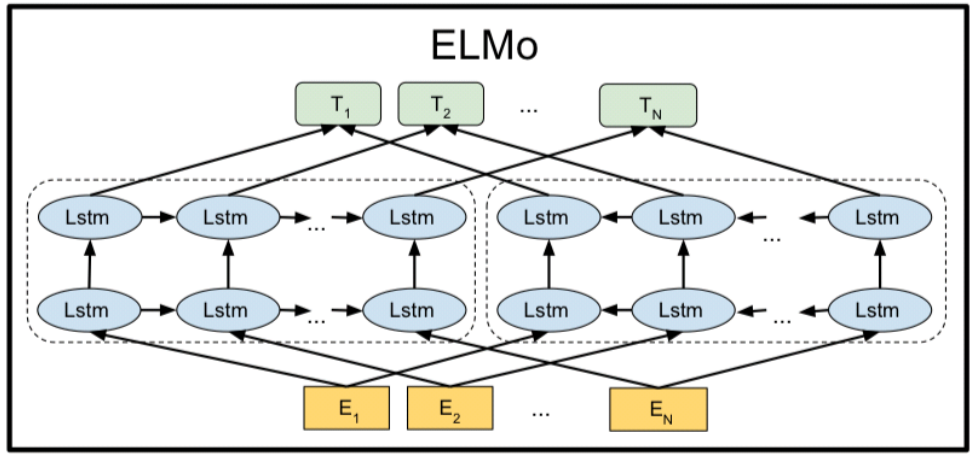
- 많은 NLP task에서 pretrained language model은 광범위하게 사용되고 있음
  - ✓ token-level : NER(name entity recognition), QA(question answering)
  - ✓ sentence-level : NLI(natural language inference), paraphrasing

- Feature-based & Fine-tuning Approaches
  - ✓ Feature-based : pre-trained feature를 모델의 additional features로 사용
    Task specific한 model architecture 필요
    (ELMo)
  - ✓ Fine-tuning : pre-training 후 전이 학습을 통해 가중치를 update함 (GPT-2)

Related work

- 관련연구

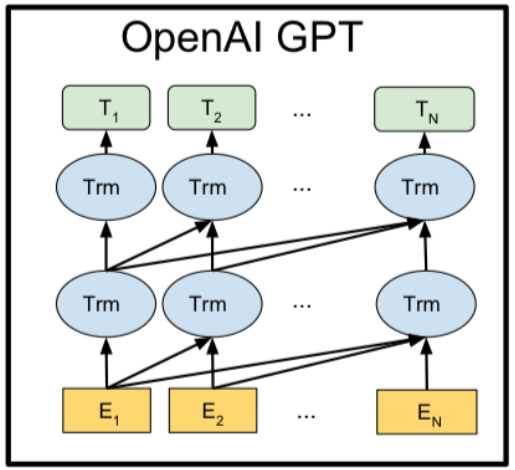
▪ Feature based Language Model(ELMo)



✓ task specific

$$\sum_{k=1}^{N} \left(\log p(t_k \mid t_1, \ldots, t_{k-1}; \Theta_x, \overrightarrow{\Theta}_{LSTM}, \Theta_s) \right.$$
$$\left. + \log p(t_k \mid t_{k+1}, \ldots, t_N; \Theta_x, \overleftarrow{\Theta}_{LSTM}, \Theta_s) \right).$$

▪ Fine-tuning Language Model(GPT-2)



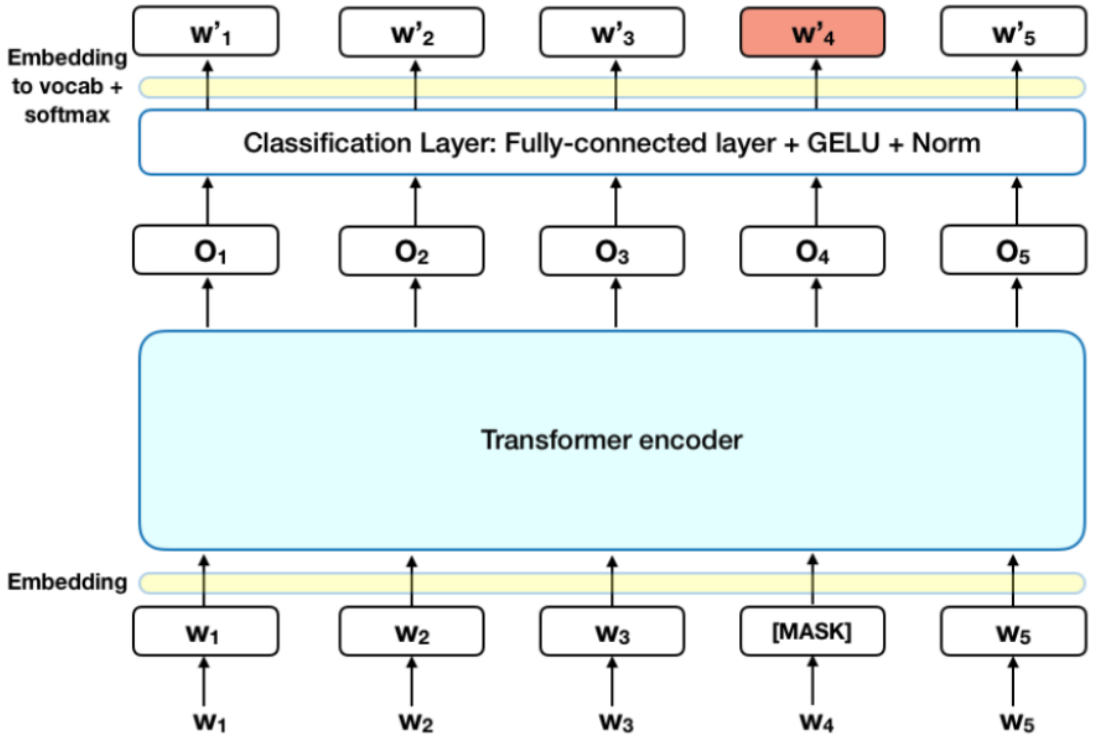
✓ Unidirectional Embedding

$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \theta)$$

Methodology

▪ Methodology

✓ MLM(Masked Language Model)



- Random하게 입력 토큰을 제거한 corpus 생성
- Masking 되는 토큰으로 선택된 문장내 토큰 비율 : 15%

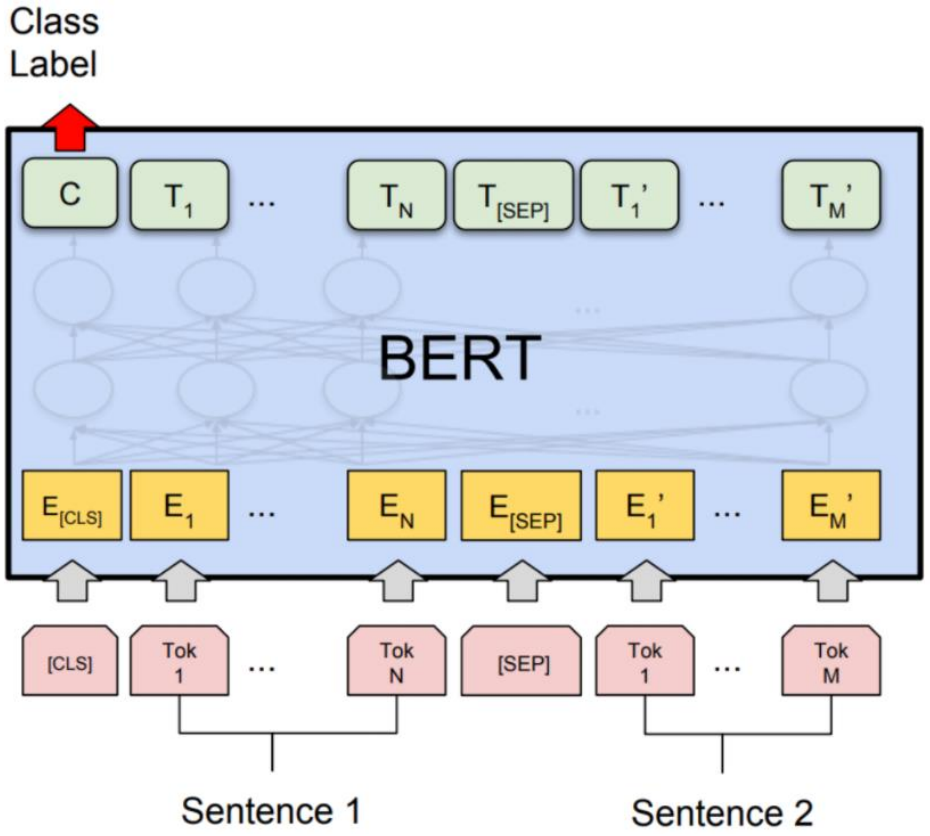
⬇

- 선택된 토큰의 80% : [Mask]
- 선택된 토큰의 10% : Random 토큰으로 변환
- 선택된 토큰의 10% : 변환하지 않음

"토큰별로 Bidirectional한 정보를 갖게 된다!!! "

▪ Methodology

✓ NSP(Next Sentence Prediction)



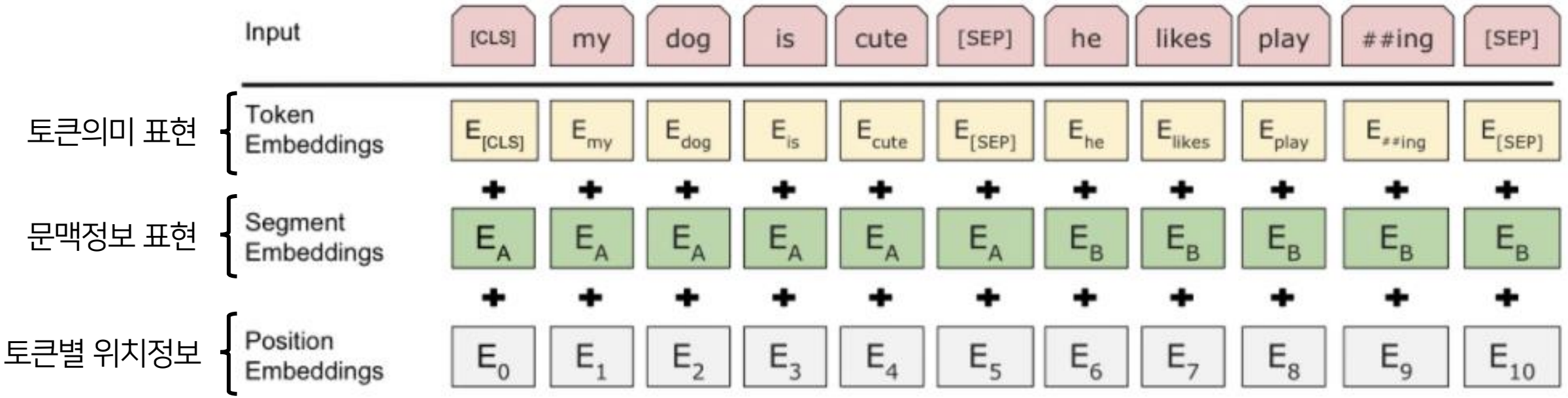
- 문장의 50% : 다음문장이 Random한 문장
- 문장의 50% : 정상적 context의 문장

"언어 모델이 문장사이의 관계를 이해하게 된다. "

Methodology

▪ Methodology(Fine-tuning)

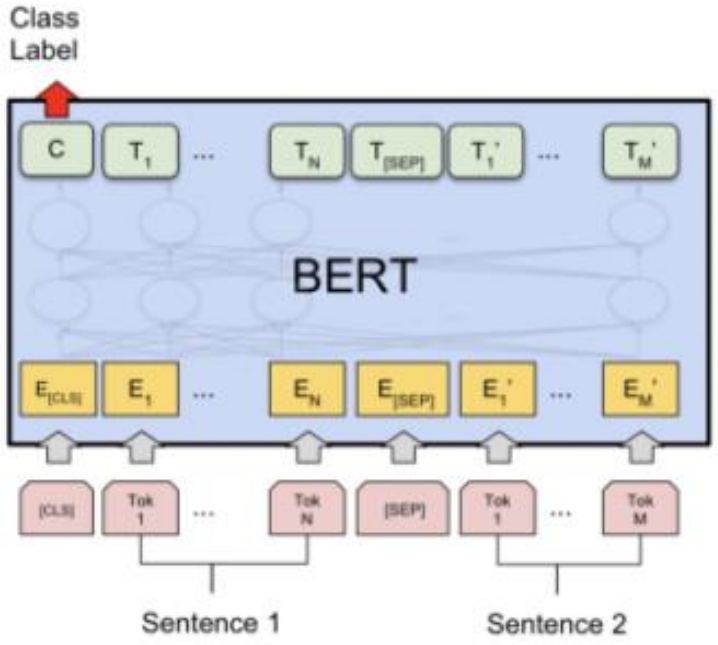
✓ BERT의 입력



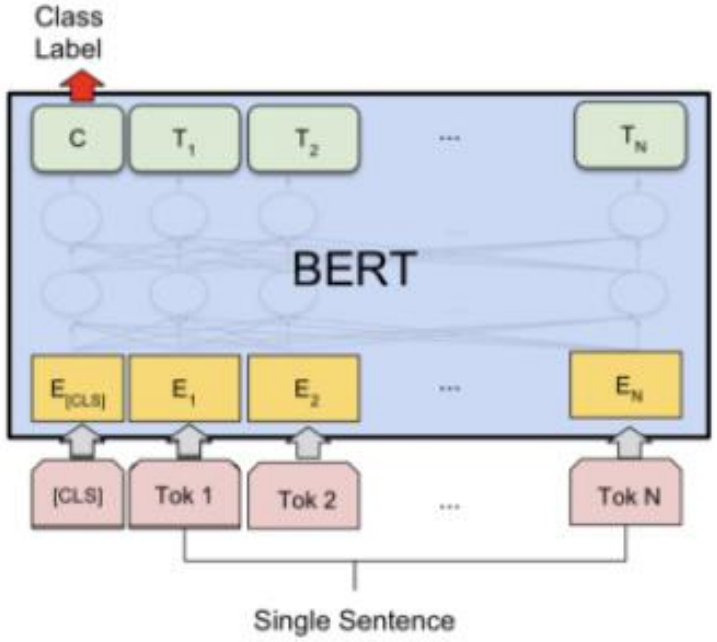
Experiments

- **Fine-tuning**

  ✓ GLUE(NLU task)



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



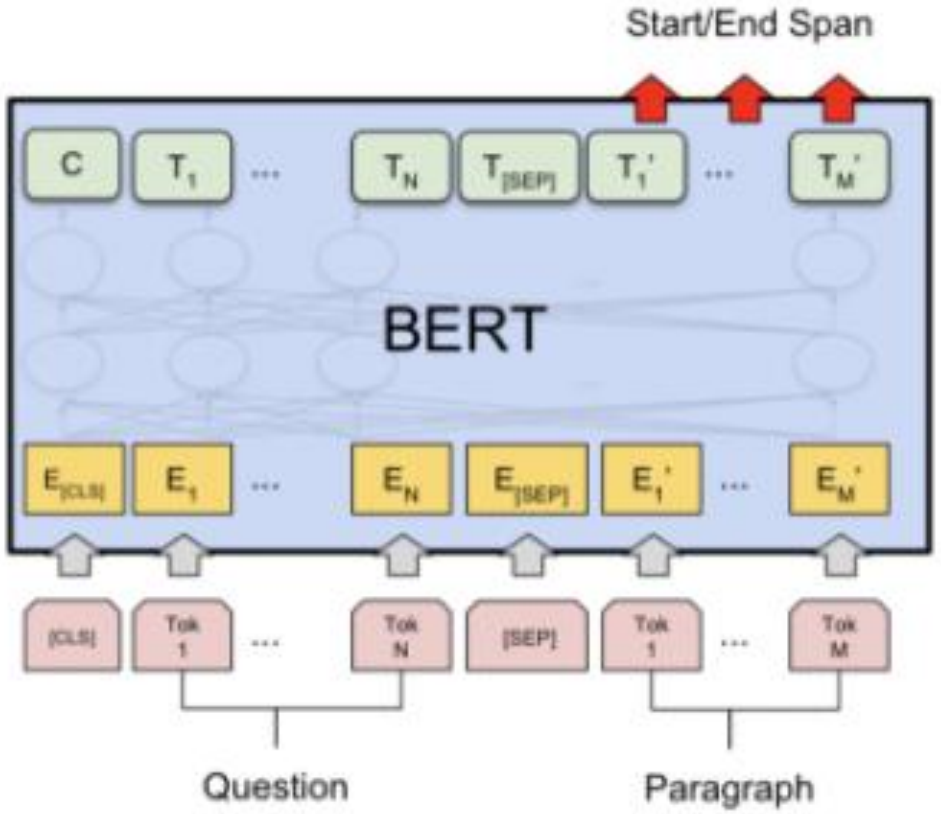
(b) Single Sentence Classification Tasks: SST-2, CoLA

✓ [CLS] 토큰을 추가하고 해당 토큰에서 얻어지는 값을 통해 결과 추론

Experiments

- **Fine-tuning**

  ✓ SQUAD



✓ Paragraph의 토큰 내 Span(질문에 대한 응답) 을 검출

$$\text{Maximize}(S \cdot T_i + E \cdot T_j),$$

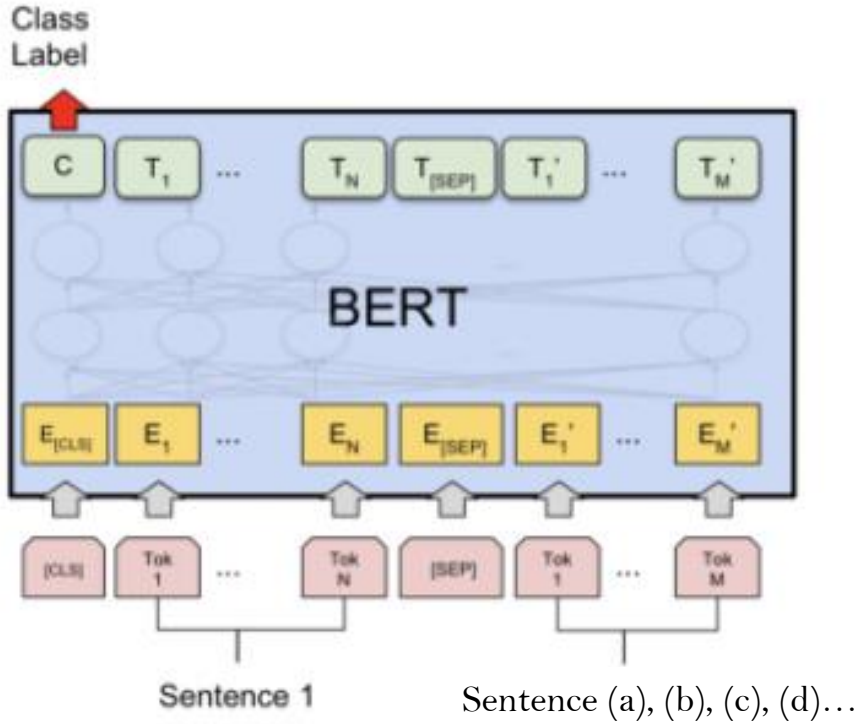
where S is start vector, E is end vector.

✓ 이때 Span의 시작점(S)과 끝 지점(E)은 softmax를 통해 찾음

Experiments

- **Fine-tuning**

  ✓ SWAG



✓ 다음 문장을 예측하는 Task

✓ [CLS] 토큰의 feature를 통해 결과를 예측함.

Experiments

▪ Ablation Studies

✓ Effect of Pre-training tasks

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

➢ No NSP 제거한 경우 NLI task에서 유독 성능이 많이 떨어짐

→ NSP가 문장간 관계분석에 도움이 된다.

➢ LTR & No NSP + BiLSTM 의 경우 Squad, MRPC task에서 성능이 떨어짐

→ BERT의 bidirectional 한 feature를 고려하는 특징이 문맥 분석에 도움이 된다.

Experiments

▪ Ablation Studies

✓ Feature-based Approach with BERT

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

✓ Feature-based Approach with BERT

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

THANK YOU