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

https://arxiv.org/pdf/1810.04805

0. Introduction

- 기존의 단어 임베딩(Word2Vec, GloVe)은 문맥 독립적이라는 한계가 있었음
- ELMo가 등장하며 문맥 의존적 표현이 가능해졌지만, 여전히 단방향성 제약 존재
- 이를 극복하기 위해 양방향 Transformer 기반 모델(BERT) 제안

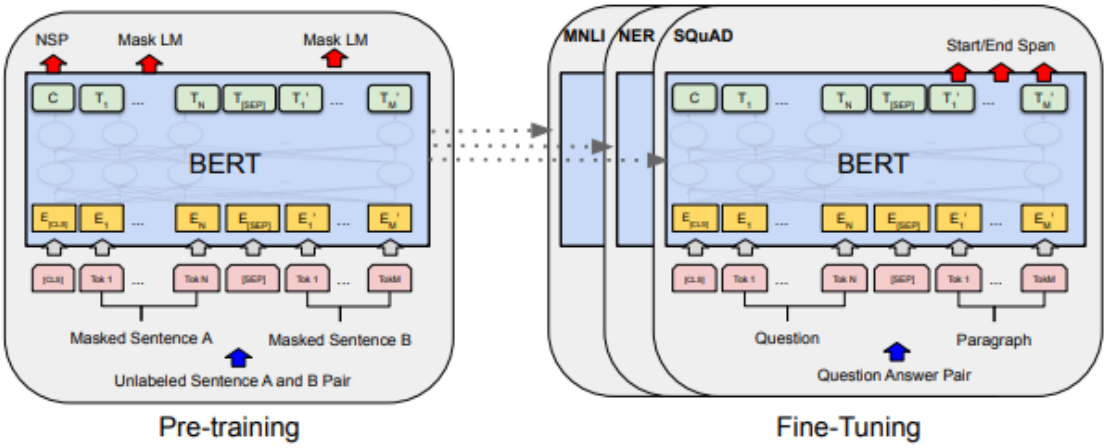
1. Overview

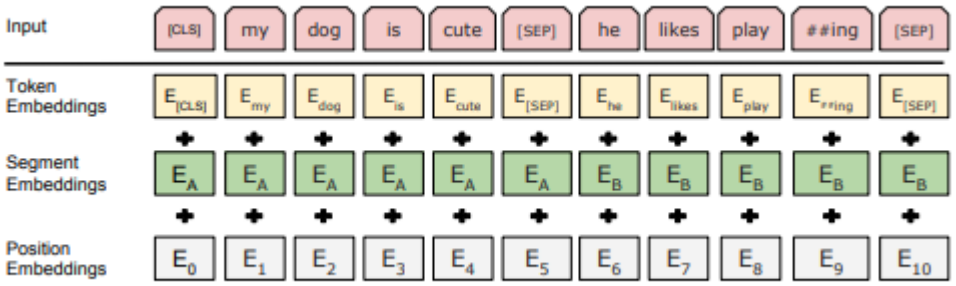
- BERT = Bidirectional Encoder Representations from Transformers
- Transformer의 Encoder 구조만 활용
- Pre-training 방식 :
  1. Masked Language Model (MLM) → 문장에서 일부 단어를 [MASK]로 가리고 예측
  2. Next Sentence Prediction (NSP) → 두 문장이 연속되는 문장인지 판별
- Fine-tuning: Downstream task(질문 답변, 자연어 추론, 분류 등)에 맞춰 가벼운 구조 추가

2. Challenges

- 기존 언어 모델은 왼쪽→오른쪽, 오른쪽→왼쪽 단방향만 가능 → 문맥 정보 제한
- 사전학습이 특정 task에 특화되어 범용성이 떨어짐
- 대규모 데이터와 자원이 필요하다는 점에서 학습 효율 문제

3. Method





- 모델 구조 : 12-layer Transformer Encoder (BERT-Base), 24-layer (BERT-Large)
- 학습 데이터 : BooksCorpus (800M words) + Wikipedia (2,500M words)
- 학습 절차 :
  - MLM : 15% 토큰 마스크 → 문맥 기반 예측
  - NSP : 두 문장이 실제 연속인지 여부를 이진 분류
- Fine-tuning : 사전학습된 BERT 위에 task-specific layer만 얹음

4. Experiments

- Benchmarks : GLUE, SQuAD, SWAG 등 다양한 NLP 과제

- 비교 대상 : ELMo, GPT, 기존 BiLSTM 기반 모델
- 학습 환경 : TPU 기반 대규모 학습, 수일 소요

5. Results

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

System	Dev EM	Dev F1	Test EM	Test F1
Top Leaderboard Systems (Dec 10th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
Published				
BiDAF+ELMo (Single)	-	85.6	-	85.8
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
$\text{BERT}_{\text{BASE}}$ (Single)	80.8	88.5	-	-
$\text{BERT}_{\text{LARGE}}$ (Single)	84.1	90.9	-	-
$\text{BERT}_{\text{LARGE}}$ (Ensemble)	85.8	91.8	-	-
$\text{BERT}_{\text{LARGE}}$ (Sgl.+TriviaQA)	**84.2**	**91.1**	**85.1**	**91.8**
$\text{BERT}_{\text{LARGE}}$ (Ens.+TriviaQA)	**86.2**	**92.2**	**87.4**	**93.2**

Table 2: SQuAD 1.1 results. The BERT ensemble is 7x systems which use different pre-training checkpoints and fine-tuning seeds.

System	Dev EM	Dev F1	Test EM	Test F1
Top Leaderboard Systems (Dec 10th, 2018)				
Human	86.3	89.0	86.9	89.5
#1 Single - MIR-MRC (F-Net)	-	-	74.8	78.0
#2 Single - nlnet	-	-	74.2	77.1
Published				
unet (Ensemble)	-	-	71.4	74.9
SLQA+ (Single)	-		71.4	74.4
Ours				
$\text{BERT}_{\text{LARGE}}$ (Single)	78.7	81.9	80.0	83.1

Table 3: SQuAD 2.0 results. We exclude entries that use BERT as one of their components.

- GLUE benchmark에서 SOTA 달성
- SQuAD (질문-답변)에서도 기존 모델 대비 큰 성능 향상
- 문맥 이해력과 문장 관계 이해 능력 모두 개선

6. Insight

- 혁신적 기여 : "Pre-training + Fine-tuning"의 패러다임 확립
- 양방향성 덕분에 문맥을 더 잘 반영할 수 있었음

- 한계 : NSP task의 필요성 논란, 엄청난 연산 자원 소모