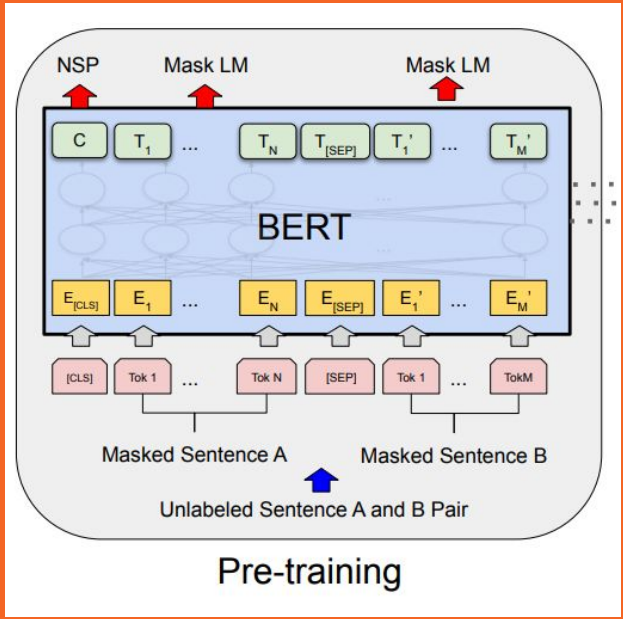
BERT



Bi-directional Encoder Representations from Transformers

어떻게 단어 임베딩에 더 효과적으로 의미를 담을 수 있을까?

- Semi-Supervised Learning
  - Unlabeled data를 활용해 단어 임베딩을 진행하는 unsupervised learning (language model)
  - Labeled data를 활용해 특정 task에 특화시켜 단어 임베딩을 진행하는 supervised learning

  →두 학습 방법의 결합으로 우수한 성능을 보일 수 있음

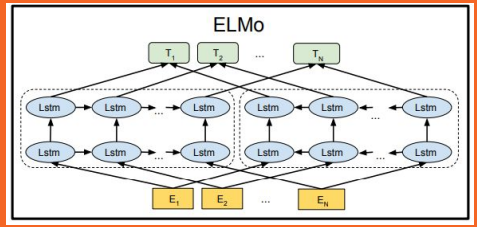
- BERT에서는 이 semi-supervised learning의 장점을 잘 활용할 수 있게 만들어진 학습 프레임워크이다

Bidirectional + Transformer = ?

BERT 이전 단어 임베딩 SOTA

i) Feature-based : ELMo

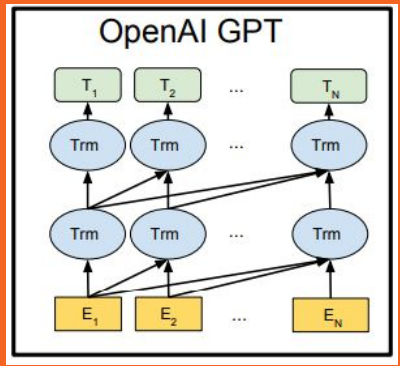
Shallow embedding and weights adjusted to be fine-tuned for specific tasks



→ 양방향을 사용해서 해당 단어의 앞뒤 문맥적 의미를 모두 임베딩 할 수 있다.
하지만 단순 LSTM 모델을 사용한다

ii) Fine-tuning : GPT

Embeddings fine-tuned with the same model architecture



→ 깊은 임베딩 모델인 트랜스포머 디코더 모델을 사용한다.
하지만, 단방향이기 때문에, 문맥적 의미가 해당 단어의 전에 등장한 단어만 임베딩된다.

Bidirectional + Transformer = BERT

트랜스포머의 인코더 블록과 디코더 블록 (복습)

- 디코더 블록
  - Input: sequence of embedded words (previous)
  - Output: the next word following the sequence
  - GPT는 트랜스포머 디코더를 language model에 적용해 단어 임베딩을 시도

- 인코더 블록
  - Input: sequence of embedded words (whole sentence)
  - Output: encoded vector of words (same size as the initial input)
  - BERT는 트랜스포머 인코더를 language model에 적용해 단어 임베딩을 시도

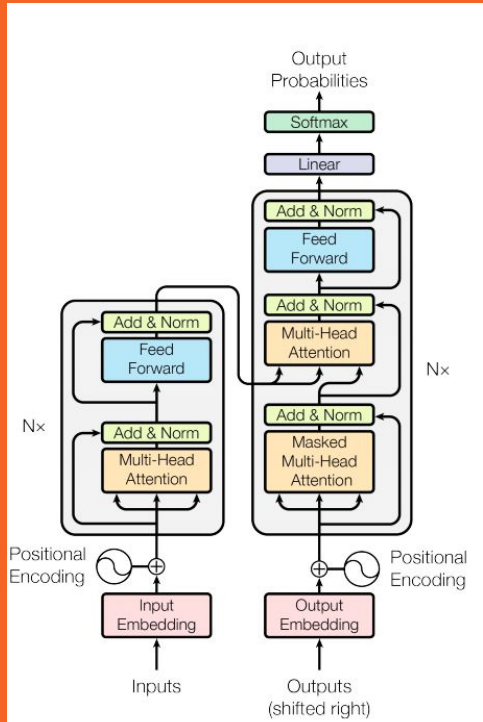


Figure 1: The Transformer - model architecture.

즉 모든 문맥적 의미를 담아내면서 transformer의 모델 복잡도를 적극 활용한다.

BERT의 2가지 foundation

i) Masked Language Model (MLM)

문장의 몇몇 (15%) 단어들을 가린다 ([MASK]) 그리고 이 가려진 단어를 예측하는 것이 목표
예: 'I was a little boy' → 'I was a [MASK] boy'

[MASK]는 test-set에서 나타나지 않기 때문에, 학습과 테스트 데이터의 mismatch가 생김

단어들 중 15%의 포지션을 랜덤으로 선정 (이 15%를 맞추는게 학습 목표)
선정된 단어는

1) 80%는 [MASK]
2) 10%는 다른 단어
3) 10%는 그대로

ii) Next Sentence Prediction (NSP)

2 문장이 주어졌을때, 2번째 문장이 1번째 문장의 다음 문장이 맞는지 아닌지 맞추는 classification task

예:

문장 1: He went to a grocery store
문장 2: He bought some eggs

문장 1: He went to a grocery store
문장 2: Eagles are fierce

Corpus에서 문장 1과 문장 2를 랜덤으로 샘플링

1) 50%는 문장 2가 문장 1의 다음 문장
2) 50%는 문장 2는 문장 1의 다음 문장이 아님

BERT의 2가지 foundation

BERT는 input으로 받는 "문장"을 임의의 길이의 연속적 단어의 조합으로 본다 (arbitrary span of contiguous words)

Input: [CLS] TOKENS ... [SEP] TOKENS ... [SEP]

i) [CLS]에서 나온 output은 classification output 이 됨 (classification token)
ii) TOKENS = MASK가 포함된 문장
iii) [SEP]는 input으로 2문장을 받을 시, 문장1과 문장2를 나눠주는 역할

예:
Data: [CLS] the man went to [MASK] store [SEP] he bought a gallon [MASK] milk [SEP]
Label: IsNext
Data: [CLS] the man [MASK] to the store [SEP] penguin [MASK] are flightless birds[SEP]
Label: NotNext

BERT - input representation

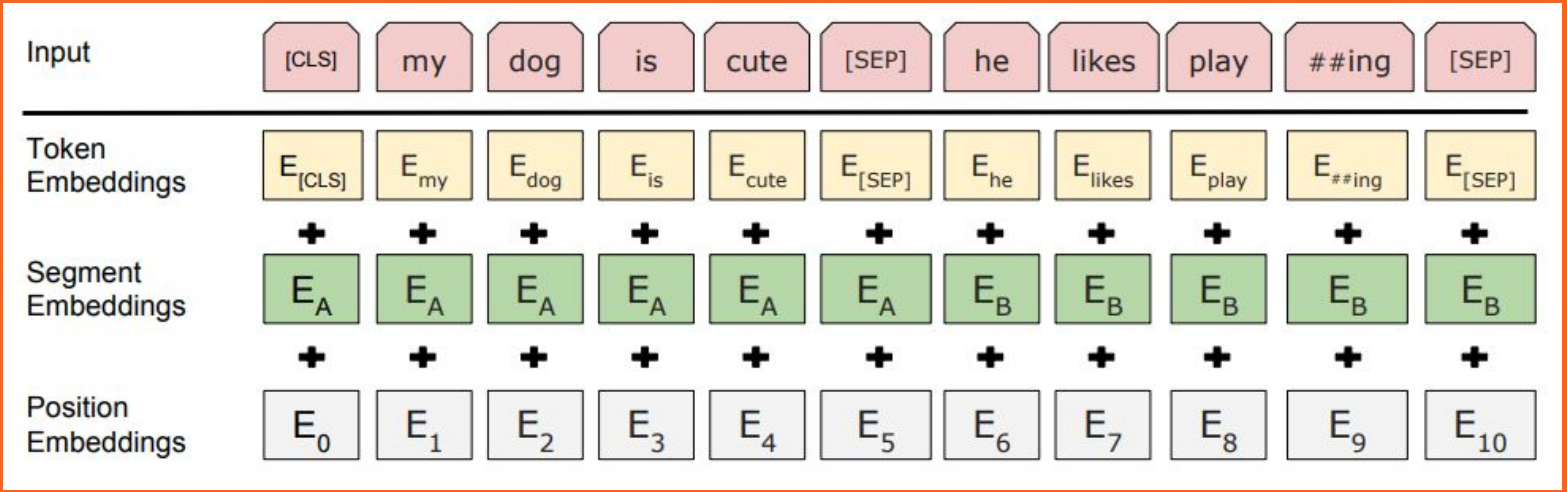
BERT의 input은 3가지 embedding값의 합으로 입력 된다.

1) 토큰 임베딩 (사전학습된 WordPiece - 30,000 단어량)

2) 해당 단어가 문장 A / 문장 B 중 어디에 속하는지 정보가 담긴 임베딩

3) Positional embedding

*WordPiece는 단어를 분리해서 학습시킴. (예: eating = eat, #ing)



BERT - 학습

- Pre-training and fine-tuning
  - Pre-train된 단어 임베딩 파라미터들은 적용될 task에 특화되어 다시 fine-tuning된다
    - Pre-train이랑 fine-tune의 모델 구조가 크게 바뀌지 않기에 큰 장점 (GPT랑 유사)

- BERT-base (GPT 비교용)
  - (L = 12, H = 768, A = 12, total parameter = 110M)
  - Seq_len = 512
  - Vocab size = 30,000 (WordPiece)

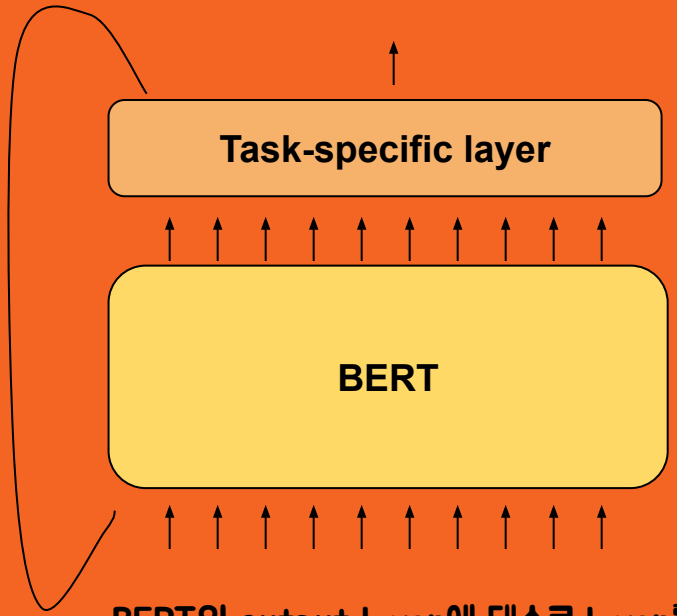
- BERT-large (메인)
  - (L = 24, H = 1024, A = 16, total parameter = 340M)
  - Seq_len = 512
  - Vocab size = 30,000 (WordPiece)

- Loss value = mean MLM likelihood + mean NSP likelihood

- 모델도 더 복잡하고 파라미터도 훨씬 더 많고, 데이터도 훨씬더 많고 TPU는 최고이고...

BERT - 활용

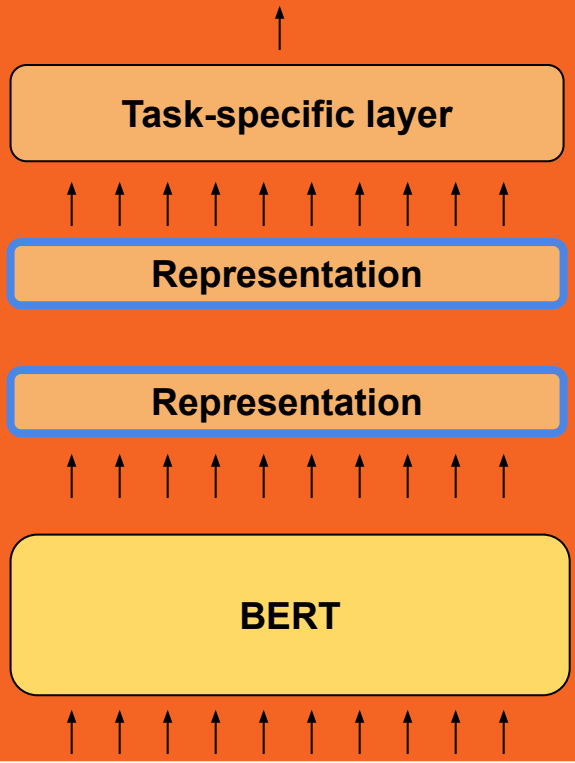
Fine-tuning



BERT의 output layer에 태스크 layer를 합쳐서 다 같이 학습

- 분류 task - [CLS]의 output 활용
- 기타 task - token output layer 활용

Feature-based



Token representation을 태스크 layer의 input값으로 입력

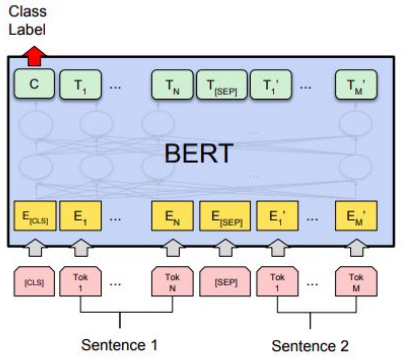
Token representation을 태스크 layer의 input값으로 입력

BERT의 output layer에서 나온 token representation을 추출

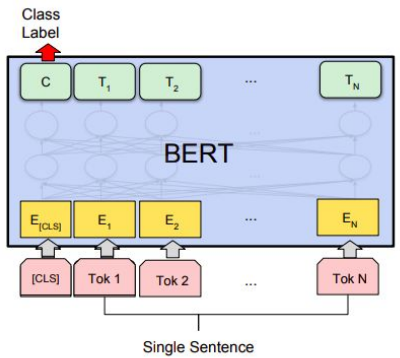
두 방법 다 좋은 성능을 발휘함 (거의 차이 없음)

BERT - 활용

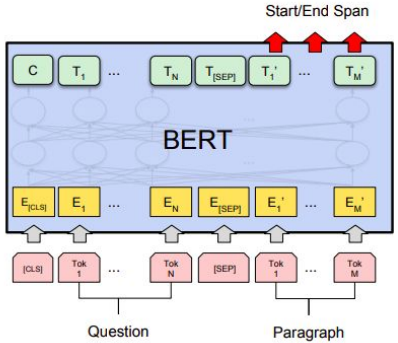
예시



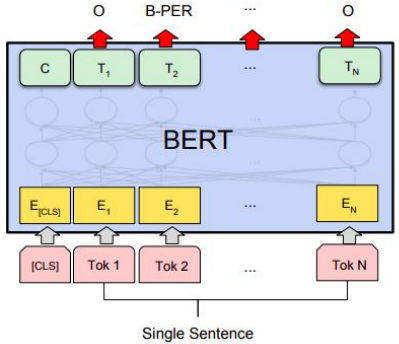
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



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



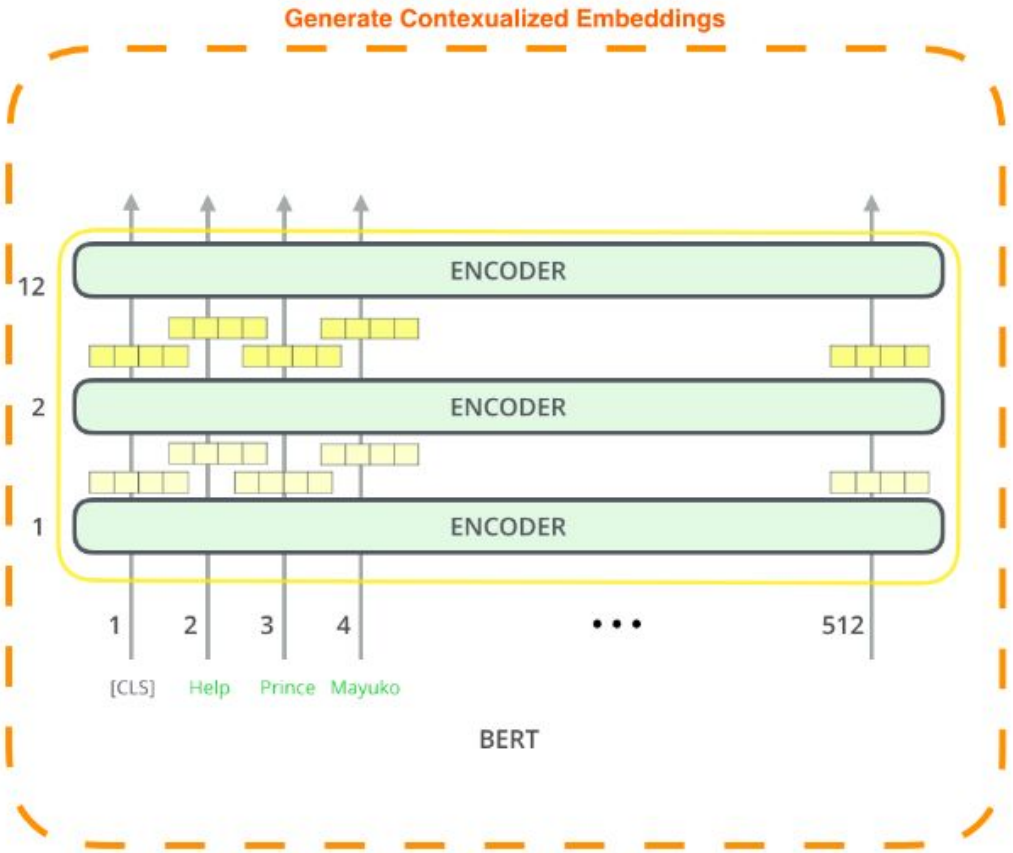
(c) Question Answering Tasks: SQuAD v1.1

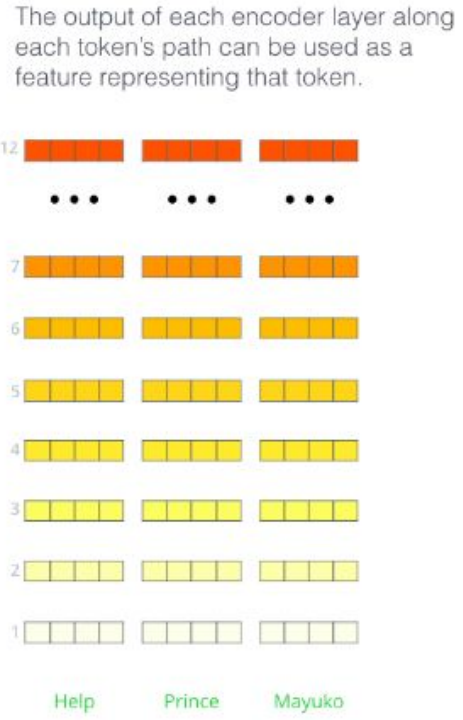


(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

BERT

예시





Best Result : Top 4 concatenation

BERT

문제점:

1) [MASK]는 fine-tuning task에 존재하지 않는다
2) [MASK]가 랜덤으로 선정되는데, 이 때 2개의 의미적으로 연결된 단어들이 함께 [MASK]된다면 이 관계를 포착하지 못한다

Further Studies

i) ALBERT https://arxiv.org/pdf/1909.11942.pdf
ii) XLNet https://arxiv.org/pdf/1906.08237.pdf
iii) RoBERTa https://arxiv.org/pdf/1907.11692.pdf
iv) ViLBERT https://arxiv.org/pdf/1908.02265.pdf
v) MT-DNN https://arxiv.org/pdf/1901.11504.pdf
vi) SenseBERT https://arxiv.org/pdf/1908.05646.pdf

인용 논문 및 기타 학습 자료

- **BERT - Pre-training of Deep Bidirectional Transformers for Language Understanding**
  https://arxiv.org/pdf/1810.04805.pdf
- **Attention is all you need**
  https://arxiv.org/pdf/1706.03762.pdf
- http://jalammar.github.io/illustrated-bert/
- https://lilianweng.github.io/lil-log/2019/01/31/generalized-language-models.html
- https://github.com/google-research/bert
- https://medium.com/@Ben_Obe/introduction-to-nlp-transformers-and-spacy-8ac9539f3bc1