BERT (Bidirectional Encoder Representations from Transformers)

개요

BERT는 2018년 구글에서 개발한 사전 훈련된 언어 모델로, 양방향 트랜스포머 인코더를 기반으로 합니다. 기존의 단방향 언어 모델과 달리 문맥의 양쪽 방향을 모두 고려하여 단어의 의미를 파악할 수 있습니다.

주요 특징

1. 양방향성(Bidirectional)

- 기존 모델(ELMo, GPT 등)은 한쪽 방향으로만 문맥을 파악
- BERT는 Masked Language Model(MLM) 방식을 통해 양방향 문맥 학습 가능
- 문장 내 임의의 단어를 마스킹하고 이를 예측하는 방식으로 학습

2. 사전 훈련과 미세 조정(Pre-training and Fine-tuning)

- 대규모 코퍼스(위키피디아, BooksCorpus 등)로 사전 훈련
- 특정 태스크에 맞게 미세 조정하여 사용 가능
- 다양한 자연어 처리 태스크에 적용 가능한 범용성

3. 모델 구조

- 트랜스포머 인코더 층을 여러 개 쌓은 구조
- BERT-Base: 12개 층, 768 hidden units, 12 attention heads (110M 파라미터)
- BERT-Large: 24개 층, 1024 hidden units, 16 attention heads (340M 파라미터)

사전 훈련 방식

1. Masked Language Model (MLM)

- 입력 텍스트의 일부 토큰(약 15%)을 무작위로 마스킹
- 마스킹된 토큰을 예측하도록 학습

- 이 중 80%는 [MASK] 토큰으로 대체, 10%는 임의의 다른 단어로 대체, 10%는 원래 단어 유지

2. Next Sentence Prediction (NSP)

- 두 문장이 연속적인지 아닌지 예측하는 이진 분류 작업
- [CLS] 토큰을 사용하여 문장 쌍의 관계 학습
- 50%는 실제 연속된 문장, 50%는 무작위로 선택된 문장으로 구성

입력 표현

- 토큰 임베딩 + 세그먼트 임베딩 + 위치 임베딩의 합
- 토큰 임베딩: WordPiece 토크나이저로 분할된 단어 표현
- 세그먼트 임베딩: 문장 A와 B를 구분
- 위치 임베딩: 단어의 위치 정보 제공

BERT의 응용

- 텍스트 분류(감정 분석, 주제 분류 등)
- 개체명 인식(NER)
- 질의응답 시스템
- 문장 쌍 분류(자연어 추론, 의미 유사도 등)
- 텍스트 요약
- 기계 번역

BERT의 한계

- 생성 작업보다는 이해 작업에 더 적합
- 긴 시퀀스 처리에 제한(일반적으로 512 토큰 제한)
- 계산 비용이 높음
- 사전 훈련 후 지식 업데이트의 어려움

BERT 이후 발전 모델

- RoBERTa: 더 많은 데이터와 최적화된 학습 방식으로 BERT 개선
- DistilBERT: 경량화된 BERT 모델
- ALBERT: 파라미터 공유를 통해 효율성 증가
- ELECTRA: 더 효율적인 사전 훈련 방식 도입