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

모델의 아키텍처

트랜스포머 기반의 양방향 인코더로 구성되어 있다. 기존의 트랜스포머의 인코더 부분만 이용하여 구현한 모델로, 언어를 양방향으로 살펴볼 수 있게 설계되어 있다. input으로는 token embedding, segment embedding, position embedding을 수행한다. token embedding은 문장을 단어 단위로 끊어서 임베딩을 수행하게 된다. bert에서는 WordPiece embedding 방법으로 문장을 토큰 단위로 끊어준다. Segment embedding은 문장별 인덱스를 의미하는 것으로, 해당 모델의 경우 두 개의 문장을 쌍으로 하여 input으로 사용할 수 있기 때문에 이러한 문장들을 구별하기 위한 목적이다. Positional embedding은 토큰의 위치를 나타내는 것이다. Input 과정에서는 모든 sequence의 첫 토큰으로 [CLS]가 쓰인다. 이는 분류를 위한 스페셜 토큰으로, 문장 전체의 정보를 종합한 것이라고 할 수 있다. 두 문장을 입력받는 경우 [SEP] 토큰을 기준으로 문장을 구분하게 된다. 학습의 경우 Pre-training과 Fine-tuning 단계로 구성되는데, 가려진 토큰을 예측하는 방식으로 문맥을 이해하도록 모델을 pre-training하고, 이후 주어진 태스크를 수행하기 위해 fine-tuning하는 방식이다.

모델의 주로 사용 가능한 태스크

두 개의 문장이 주어졌을 때 해당 문장 사이의 관계가 어떤 범주인지 분류하는 sentence pair classification task를 수행할 수 있고, 한 문장이 주어졌을 때 그 문장이 어떤 범주에 속할지 분류하는 긍정부정의 감성 분석 또한 진행할 수 있다. 그리고 질문에 대한 올바른 답변인지 판단하는 태스크에도 활용할 수 있다.

모델의 의의와 한계

- BERT 모델의 의의
  BERT 모델의 맹점은 NLP 모델에서 Pre-training을 통해 모델의 성능을 대폭 향상시켰다는 것이다. 이를 통해 NLP 문제를 해결할 때 반드시 해당 문제와 관련된 데이터로만 모델을 훈련시켜야 하는 것이 아니라, 일반적인 NLP 데이터로 사전학습을 진행하고 이를 모델에 적용해도 된다는 점을 증명했다. BERT의 등장으로 많은 논문들이 사전학습 모델을 활용해서 다양한 문제를 풀어냈다. 또한 BERT를 통해 데이터 부족 문제를 일정 부분 해결할 수 있게 되었다. 해결하려는 문제와 관련된 데이터가 소량일 때 BERT를 이용하여 사전학습을 통해 모델을 훈련시킬 수 있기 때문이다.
- BERT 모델의 한계
  1) 계산 비용 및 메모리 요구량
     BERT는 3억개가 넘는 갯수의 파라미터를 가지고 있다. 따라서 필연저긍로 학습과 추

론에 상당한 계산 비용이 요구된다. 같은 이유로 많은 양의 메모리 또한 요구된다.

2) 대량의 학습 데이터 필요

BERT의 맹점은 Pre-training이다. 그러나 이를 위해서는 대량의 텍스트 데이터가 필요하다. 이로 인해 BERT는 개인이 쉽게 학습하기 어렵다는 한계가 존재한다.

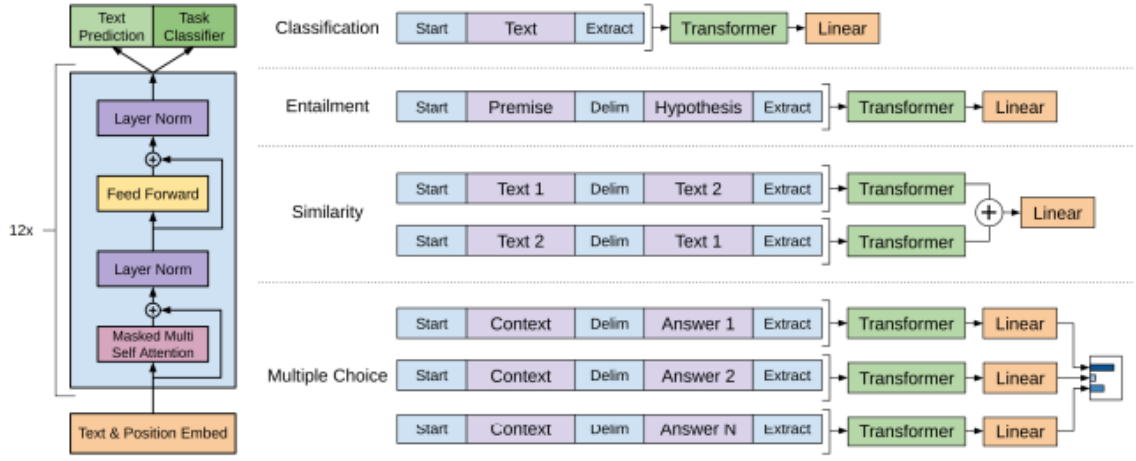
Improving Language Understanding by Generative Pre-Training (GPT1)

https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

모델의 아키텍쳐는 어떠한가?

GPT는 transformer의 디코더 아키텍처를 활용한다. Encoder에 집중한 BERT와 달리 Decoder 구조를 활용하여 token 생성에 더 집중하고 있음을 알 수 있다. 해당 논문은 기존에 사용하던 label data의 수가 적어 NLP에 어려움이 존재하여 unlabeled text로 학습을 활용하여 하나의 representation model을 생성하고 최소한의 수정을 통해 여러 task를 해결하고자 하였다.

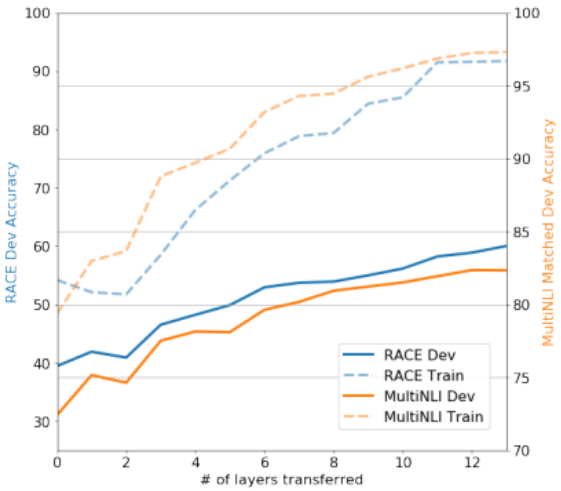
전이(transfer)할 때 task-specific input을 사용하고, input을 하나의 연속된 토큰 시퀀스로 처리하는 접근 방식을 선택하여 사전 학습 모델을 최소한으로 수정하고, 효과적으로 세밀 조정(fine-tuning)할 수 있도록 한다. 해당 논문은 준지도 학습 방식을 제안한다. 비지도 방식인 pre-training과 지도방식인 fine-tuning의 결합을 통해 훈련하는 것이다. 먼저, 미분류 데이터에서 언어 모델링을 사용함으로써 신경망 모델의 초기 매개 변수를 학습하는 pre-training 단계를 거친다. 그런 다음 학습된 파라미터에 대하여 labeled data를 이용해 특정 task에 적용하는 supervised fine-tuning 단계를 거친다.

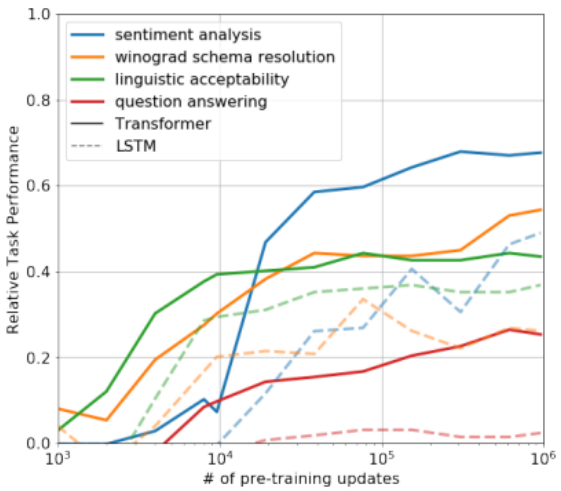


모델의 주로 사용 가능한 태스크는 무엇인가?

다양한 자연어 이해 작업에 사용할 수 있다. NLI에 활용되어 모순되거나 대등한 문장의 관계를 인식하고 분류할 수 있다. 해당 모델을 제안한 논문에서는 특히 NLI, 자연어 추론에서 여러 개의 문장과 언어 모호성 분야에서 좋은 성능을 보였음을 알려주었다. 질의 응답과 상식 추론에 활용될 수 있다. 또 두 문장이 의미적으로 동일한지를 추론하는 텍스트 유사도 평가에도 활용될 수 있다. 마지막으로 텍스트를 분류하는 작업에도 활용될 수 있다.

모델의 의의와 한계점이 무엇인가?





해당 모델은 기존의 RNN과 같은 구조를 탈피하여 transformer 구조를 사용했다는 것에 큰 의의가 있다. transferring embedding들이 layer마다 최대 9%의 성능 향상을 일으킨다는 것을 밝혔다. 이는 pre-trained model에서 각각의 레이어들이 유용한 함수를 포함하고 있음을 나타낸다. 또한 transformer의 attentional memory가 기존 LSTM들에 비해 전이 시에 도움이 되며, 학습 횟수에 따라 꾸준히 성능이 증가하는 것을 보였다. 또한 unlabeled data를 활용할 수 있는 pre-training 방식을 적용한 모델을 제안함으로써 여러 domain에 대한 일반적인 모델을 생성할 수 있다는 가능성을 제시해주었다. 그러나 fine-tuning 단계에서 input transformation과 label data가 필요하다는 점에서 완전한 비지도학습 모델이 아니라는 점을 한계로 들 수 있다.