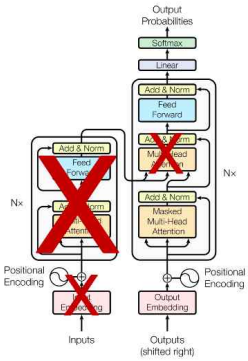
Transformer 공통과제

김예린, 김종진, 박정양, 이상헌

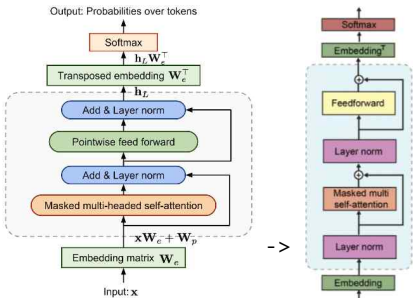
0. 선택한 논문

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
Language Models are Few-Shot Learners

1. 모델의 아키텍쳐



GPT와 BERT는 Transformer 구조를 기반으로 한 모델이다. 단, GPT는 Transformer의 디코더 모듈을 이용하고, BERT는 인코더 모듈을 이용한다는 차이점이 있다. 이 때 GPT는 인코더 모듈을 사용하지 않기 때문에, 디코더의 인코더-디코더 attention부분 역시 제외된다. 결론적으로 Masked Multi-Head Attention과 Feed Forward 두 가지 레이어로 이루어진 블록을 여러 개 쌓은 것이 GPT 모델의 기본적인 구조라고 할 수 있다. 이 디코더 블록이 많을수록 모델이 커지고 파라미터의 개수도 많아지지만, 동일한 정보를 더 많은 벡터로 표현할 수 있어 정확한 task 수행이 가능하다.



GPT의 여러가지 모델은 같은 구조와 학습 방법 및 작동원리를 공유하면서 모델의 크기, 학습 데이터셋의 크기 및 학습 횟수의 차이만 존재한다. 다만 디코더 블록의 구조의 경우 GPT1에서는 Layer normalization 과정이 각각의 하위 레이어 이후에 존재했지만, GPT2와 3는 하위 레이어의 앞쪽으로 위치가 변경됐다.

BERT는 Transformer의 인코더를 여러 개 쌓아 올린 구조이다. BERT-base 모델은 12개, BERT-large 모델은 24개를 쌓아 올린 구조로, 인코더를 쌓아 올려 생성한 모델을 pre-training을 진행한다. 그 과정에서 입력값을 Token Embedding, Segment Embedding, Position Embedding 층을 이용해 임베딩을 진행한다. Token Embedding은 WordPiece 임베딩 방식을 이용해서 실질적으로 단어를 토큰 형식으로 변경하는 과정이다. Segment

Embedding은 2개의 문장이 동시에 들어오는 task를 해결하기 위해 문장을 구분해주는 역할을 한다. Position Embedding은 Transformer에서 positional encoding과 같이 단어들의 위치 정보를 학습하는 역할을 하는 단계이다. 이 3가지 임베딩을 거쳐 만들어진 입력값으로 Masked LM(MLM)과 Next sentence prediction(NSP)을 이용한 pre-training을 진행한다. MLM이란, input token의 일부를 masking한 뒤 masking된 token을 예측하는 학습방식이다. NSP는 두 문장 사이의 관계를 이해하는 능력을 키우기 위해 진행하는 pre-training으로, 두 문장이 바로 붙어있는 문장인지 떨어져 있는 문장인지를 예측하는 학습방식이다.

2) 모델의 주로 사용 가능한 태스크
GPT와 BERT 모두 전반적인 NLP 태스크를 수행하는 데 폭넓게 사용할 수 있다. GPT는 기본 Language Modeling부터 텍스트 생성 및 요약과 분류, 번역, 프로그래밍 언어 작성, 질문에 대한 대답 등 우리가 흔히 ChatGPT를 사용해 처리할 수 있는 NLP와 관련된 여러 task 등에 이용할 수 있다. 또한 GPT4부터는 NLP뿐 아니라 이미지를 입력받아 이를 해석하는 등 멀티 모달의 영역까지 넓혀나가고 있으며, 최근에는 다른 언어 모델이 태스크를 잘 수행하는 지 여부를 판단하는 LLM-as-a-Judge의 기능도 수행하고 있다.
BERT 모델은 양방향으로 문맥을 고려할 수 있는 트랜스포머 기반 언어 모델로써, 문맥을 고려해 주어진 문서나 문장에서 질문에 대한 답변을 생성하는 QA task, 인간이 사용하는 자연어를 다른 언어로 변역하는 일인 기계 번역, 긴 텍스트를 간결하게 요약하는 등의 다양한 NLP task에 이용할 수 있다.

3) 모델의 의의와 한계점이 무엇인가? (0.5p 이내)
GPT 이전의 NLP 관련 모델들은 특정 task를 위해 수집된 labeled data로 모델을 fine tuning하는 방식으로 개발되었다. 하지만 레이블을 붙이는 데는 많은 시간과 비용이 소요되며, 특정 task를 위해 수집된 데이터로 학습이 이뤄졌기 때문에 높은 일반화 성능은 기대할 수 없다.
GPT 모델은 비지도 사전학습과 fine tuning 등을 활용한 준지도 학습방식으로 접근해 이러한 점들을 해결했다. 레이블이 지정되지 않은 대규모의 텍스트 데이터로 모델을 학습하고, 학습한 모델을 task에 맞게 소수의 labeled data만으로 fine tuning을 함으로써 general한 과제 수행 능력을 높였다. 더 나아가 GPT 2 모델부터는 In-context learning이 가능해지면서 일반화 성능을 끌어올렸다. Few-shot Learning 뿐만 아니라 zero-shot 세팅 하에서도 타 모델보다 높은 과제 수행 능력을 보였다.
GPT의 한계점 중 하나는 모델의 크기가 크다는 점이다. 모델과 학습 데이터의 양이 커질수록 모델의 성능 또한 향상되지만, GPT3는 1750억 개의 큰 파라미터를 가지고 있으며, 때문에 모델의 실질적인 활용성이 제한된다. 또한 앞서 서술했듯, GPT모델이 모든 NLP task에서 좋은 성능을 보인 것은 아니라는 한계점 또한 존재한다.
BERT 이전의 모델들은 한 방향으로만 학습을 진행했다. BERT는 이와 다르게 양방향 학습 방법을 사용하여 긴 context를 이해하고 대답하는 능력을 향상시켰다는 데에 의의가 있다. 다만, BERT는 일반적인 NLP task에는 잘 작동하지만, 사용 단어, 언어의 특성에 따라 Bio, Science, Finance 등 특정 분야의 task에는 잘 작동하지 않는다. 이러한 분야에 BERT를 적용하기 위해서는 그 분야에 대한 데이터셋을 구축하고, fine-tuning을 해야 한다는 한계점이 존재한다.