4. GPT, GPT-2

스터디자료 : https://docs.google.com/presentation/d/1DxngNiNGG1esu5MZWMFnws3ooRayAvCgIQVzVyb81Zw/edit?usp=sharing

· GPT(Generative pre-training)

- Transformer의 Decoder 부분만을 사용함. Why? 다음 단어를 예측하는데 있어서 정답을 알려주면 안되기 때문에 언어모델링을 위한 좋은 방법으로 판단함.
  - ex) '나는 밥을 먹었다.' 문장을 인공지능이 생성할 때, '밥' 단어를 예측하는데 인공지능이 생성하지 않은 단어 '먹다'를 알고있는건 말이 안되는 상황이기 때문.
- 단, Transformer의 Deocder만을 사용하기 때문에 Transformer의 Encoder-Decoder Attention을 사용하지 않음
- Decoder Layer를 12개 쌓아서 만들어 총 모델의 파라미터 개수는 117M (BERT 모델의 파라미터 개수 : 340M)

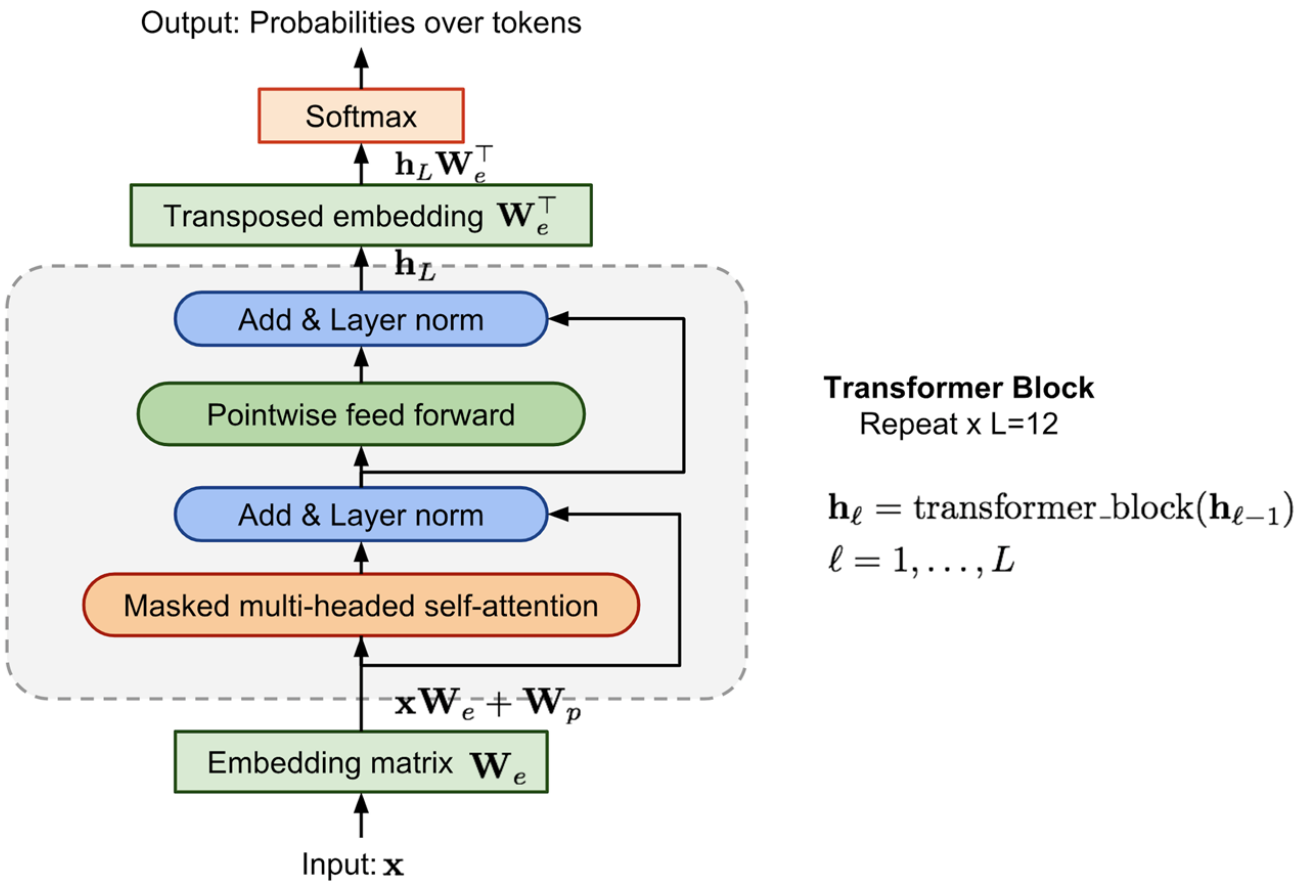
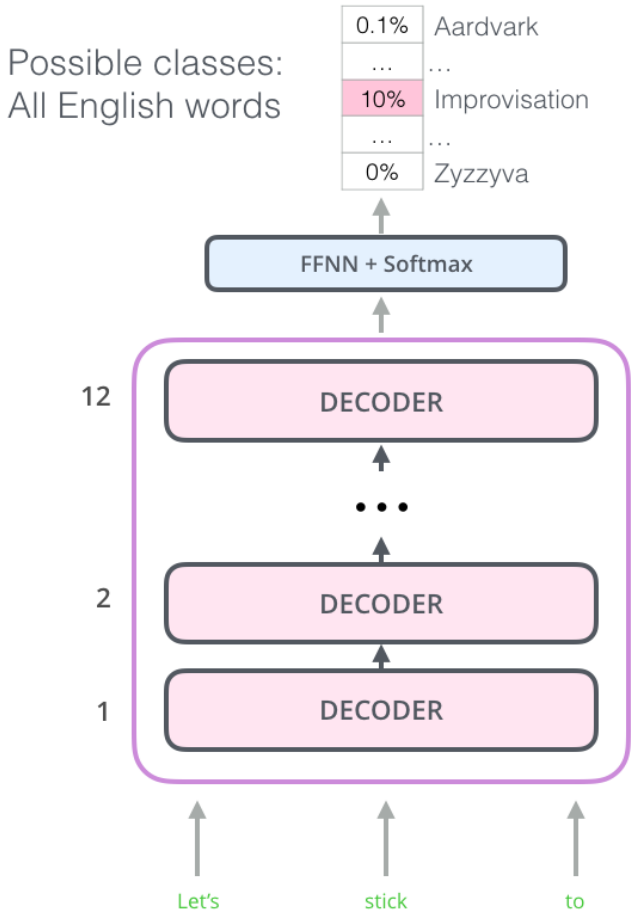


그림 1. GPT의 Decoder Layer 상세구조. Transformer Decoder에서 Encoder-Decoder Attention이 제거된 모습.



The OpenAI Transformer is made up of the decoder stack from the Transformer

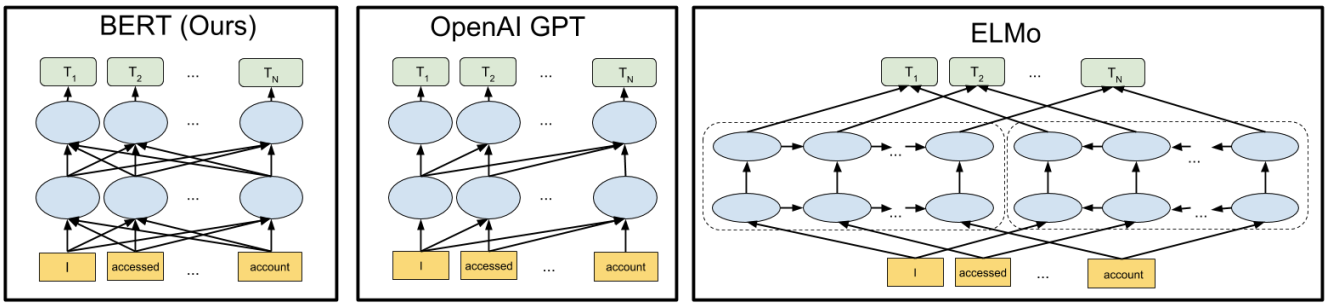
---

BERT와 GPT 비교

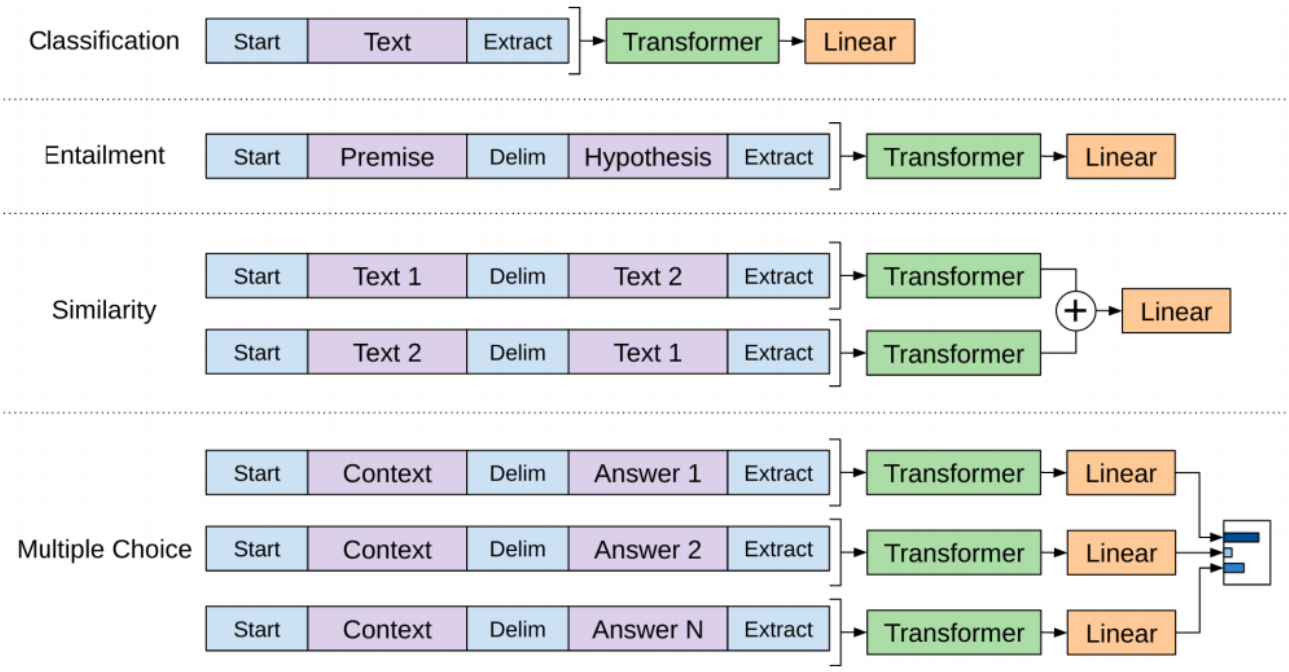
- 공통점
  - pre-train / fine-tuning 구조
  - Transformer 모델을 채용 (Layer 수도 동일)
- 차이점

  BERT는 Transformer의 인코더를 활용
  - 문장에 대해 양방향 Self Attention을 사용함 (미래의 문장에 대해 학습)

  GPT는 Transformer의 디코더를 활용
  - 문장에 대해 단방향 Self Attention을 사용함 (미래의 문장에 대해 학습하지 않음)



학습 및 활용 용도



GPT를 활용한 use case

GPT2(Generative pre-training-2)

- GPT와의 차이점
  - 모델의 크기 변경과 구조 수정
    - 데이터 크기, 어휘 크기, 임베딩 차원 크기, 모델 차원 수, 디코더 층 수를 모두 증가