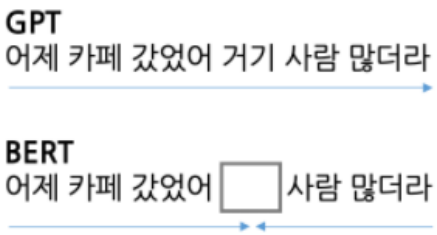
GPT(Generative Pre-trained Transformer)

1. GPT vs. BERT

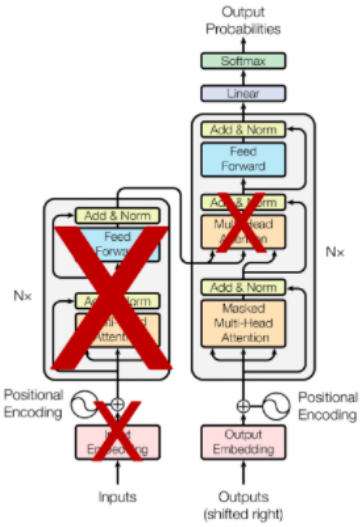


	GPT	BERT
모델	Pre-trained Language Model	Pre-trained Masked Language Model
구조	트랜스포머에서 **디코더**만 이용	트랜스포머에서 **인코더**만 이용
pre-training 방법	이전 단어들이 주어졌을 때 다음 단어를 예측하도록 pre-training	문장 중간에 빈칸을 만들고 해당 빈칸에 어떤 단어가 적절한지 예측하도록 pre-training
방향성	단방향(unidirectional)	양방향(bidirectional)

2. GPT의 구조

- 트랜스포머에서 인코더를 제외하고 디코더만 사용
- 인코더 쪽에서 보내오는 정보를 받는 모듈(Multi-Head Attention)도 제거

> GPT는 엄밀히 말하면 트랜스포머의 디코더 구조를 이용한다기 보다는 **Masked Multi-Head Attention**을 이용한다고 볼 수 있다.



2. GPT의 Embedding

1) 워드 임베딩

- GPT 역시 BERT와 마찬가지로 워드 임베딩을 위해 **서브워드 토크나이저**를 사용한다.
- GPT의 경우 길이가 50257인 단어 집합을 사용한다.

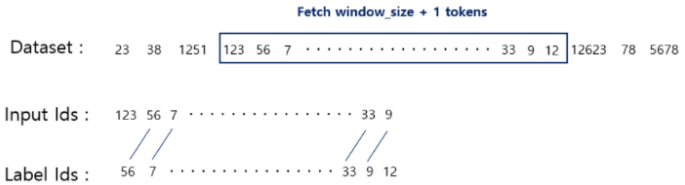
2) Position Embedding

- BERT와 마찬가지로 각 단어의 위치 정보를 포함해주기 위한 Position Embedding을 사용한다.
- 워드 임베딩과 더해서 사용하기 위해 임베딩 크기는 워드 임베딩과 같게 맞춰준다.

3. GPT의 Pre-training

1) Input & Label

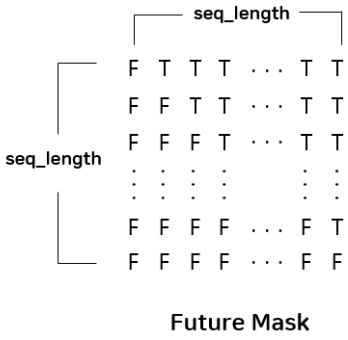
- GPT 의 Pre-training 은 다음에 올 토큰을 예측하는 방식으로 학습되어 별다른 라벨이 필요 없는 Unsupervised Learning이다.
- 따라서 input은 하나의 sequence가 되고 label은 각 단어의 다음에 올 단어들의 sequence가 된다.



https://ainote.tistory.com/17

2) Future Mask (Attention Mask)

- GPT는 Masked Multi-Head Attention 계층에서 현재 예측하고자 하는 단어의 앞부분만 예측에 활용할 수 있도록 마스킹 해준다.
- 때문에 이를 돕기 위한 Mask table이 Input으로 함께 주어진다.
- 아래의 그림에서 각 행이 하나의 시퀀스에 이용되는 마스크이며 F(False)는 마스크 되지 않는 경우, T(True)는 마스크 되는 경우를 나타낸다.

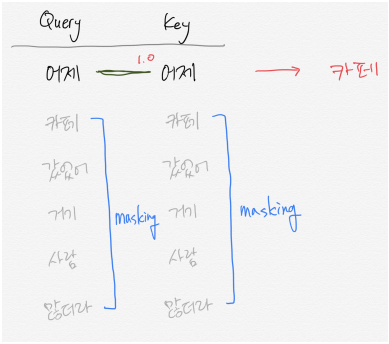


Future Mask

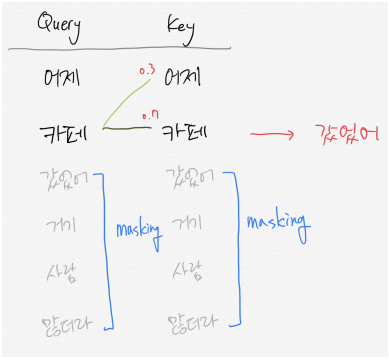
- 그 외에 모든 포워드 과정은 트랜스포머의 디코더 계층과 동일하다.

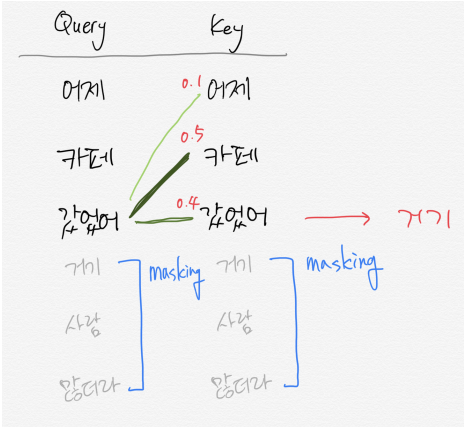
3) Fowarding 과정

- 아래와 같은 문장이 들어오면 GPT는 순차적으로 앞에서 부터 단어를 학습한다. 먼저 “어제” 라는 단어 하나만 보고 카페라는 단어를 예측한다.
- 이때 “어제” 를 제외한 모든 단어는 마스킹 되어 있어야 하며 이번 시점의 정답인 ”카페”에 해당하는 확률은 높이고 나머지 단어의 확률은 낮아지도록, 모델 전체를 업데이트 한다.



이 뒤도 같은 방식으로 진행한다.




Query
Key
어제
어제
0.1
0.5
카페
카페
0.4
갔었어
갔었어
거기
masking
masking
거기
사람
많더라
거기
사람
많더라


가 가격 … 갔었어 … 거기 … 많더라 … 사람 … 어제 … 카페…