TRANSFORMRT, BERT

1. Seq2Seq

Seq2Seq는 Encoder → Decoder 구조 모델이다.

- Encoder: 입력 문장 → 하나의 벡터(context vector)
- Decoder: 그 벡터를 보고 문장 생성

문제: 긴 문장일수록 context vector 하나에 정보가 압축됨 → 정보 손실 발생

그래서 등장한 것이 attention

2. Attention

Attention은 문장을 처리할 때 입력 문장의 모든 단어를 동일하게 사용하는 것이 아니라, 현재 필요한 정보와 더 관련이 높은 단어에 더 집중하도록 가중치를 부여하는 메커니즘이다.

- Attention score 계산

  현재 시점에서 필요한 정보(Query)를 기준으로 입력 단어(Key)들과의 관련도를 계산하여 각 단어의 중요도 점수를 만든다.

- Softmax 적용

  중요도 점수들을 확률 형태의 가중치로 변환하여, 어떤 단어를 얼마나 참고할지 결정한다.

- Attention value 생성

  각 단어의 정보(Value)에 가중치를 곱해 합산함으로써, 중요한 단어의 정보가 더 크게 반영된 최종 표현을 만든다.

3. Transformer

Transformer도 Encoder-Decoder 구조지만

- RNN/LSTM 사용 X
- Self-Attention만으로 문장 처리
- 모든 단어를 동시에 처리 (병렬 처리 가능)
- 긴 문장에서도 정보 손실 적음

Transformer - Encoder

(1) Self-Attention

입력 문장의 각 단어가 문장 내부의 다른 모든 단어를 참고하여 새로운 표현을 만드는 과정이다.

이를 통해 단어 간 관계(의존성)를 학습한다.

(2) Scaled Dot-Product Attention

- Query와 Key의 내적으로 단어 간 유사도를 계산
- 값이 너무 커지는 것을 방지하기 위해 √d 로 나누어 스케일링
- softmax를 통해 가중치를 만든 뒤 Value를 가중합

(3) Multi-Head Attention

attention을 여러 개 동시에 수행하는 구조

Transformer - Decoder

(1) Masked Multi-Head Attention

Self-attention은 원래 모든 위치가 서로를 볼 수 있기 때문에 모델이 정답을 미리 보는 것이 가능해진다. → 미래 단어를 보지 못하도록 마스킹을 적용한 self-attention

(2) Position-wise Feed Forward Networks

- attention 이후 각 단어 벡터를 개별적으로 한 번 더 변환하는 작은 신경망
- 모든 위치에 동일한 MLP를 독립적으로 적용

(3) Residual Connection & Layer Normalization

Residual Connection: 서브레이어(Self-attention, FFN 등)의 출력에 원래 입력을 그대로 더해주는 구조
Layer Normalization: residual로 더해진 결과를 정규화하여 값의 분포를 안정화

4. BERT

Transformer의 Encoder만 사용한 양방향 언어모델로 문장을 생성하기보다 문장 이해(Task) 에 사용된다.

- 좌우 문맥을 동시에 활용하는 bidirectional context 학습
- 사전학습(pretraining) 후 다양한 NLP task에 fine-tuning하여 사용

(1) Input Representation

BERT 입력은 다음 세 가지 임베딩을 더해 생성된다.

- Token Embedding: 단어 의미 정보
- Segment Embedding: 문장 A/B 구분 정보
- Position Embedding: 단어 위치 정보

또한 특수 토큰을 사용한다.

- `[CLS]`: 문장 전체 표현 (분류 task에 사용)
- `[SEP]`: 문장 구분

(2) Masked Language Model (MLM)

- 입력 문장의 일부 단어를 가리고(Mask) 주변 문맥을 이용해 해당 단어를 예측하도록 학습
- 양방향 문맥을 동시에 활용할 수 있도록 하는 학습 방식

(3) Next Sentence Prediction (NSP)

- 두 문장이 실제로 이어지는 문장인지 여부를 분류하는 task
- 문장 간 관계를 학습하기 위해 사용됨

(4) Fine-tuning

사전학습된 BERT 위에 간단한 출력층만 추가하여 다양한 NLP 문제를 해결한다.

예:

- Sentence Pair Classification
- Sentence Classification
- Question Answering
- Token Classification

5. BERT vs GPT vs BART

구분	**BERT**	**GPT**	**BART**
아키텍처	Encoder only	Decoder only	Encoder–Decoder
Attention 방향	양방향 (Bidirectional)	단방향 (Left → Right)	양방향+ 생성
학습 목표	MLM, NSP	Autoregressive LM (다음 토큰 예측)	Denoising Autoencoder
강점	문장 이해에 강함	문장 생성에 강함	이해 + 생성 모두 가능
대표 활용	NLI, QA, NER	텍스트 생성	요약, 번역