Transformer는 인코더-디코더 구조를 가진 모델에서 가장 일반적으로 사용되던 순환 레이어를 Multi-Head Attention로 대체한 최초의 어텐션 기반 시퀀스 변환 모델이다.

RNN, LSTM, GRU와 같은 신경망 기반 시퀀스 변환 모델은 일반적으로 인코더-디코더 구조를 가지고 있으며, 인풋과 아웃풋 시퀀스에 따라 순차적으로 문맥을 학습한다. 이러한 모델들은 데이터의 병렬 처리를 어렵게 만들고 시퀀스 길이가 길어질수록 학습이 어려워진다는 단점이 있었다.

Attention

Attention 메커니즘은 쿼리, 키, 값이라는 벡터를 사용하여 문맥에서 각 단어의 중요성을 고려하여 단어마다 다른 가중치를 부여한다.

Encoder/Decoder

인코더는 6개의 동일한 레어어로 구성되고, 각 레이어는 인풋 시퀀스 내의 위치 간 의존성을 계산하는 Multi-Head Attention과 피드포워드 네트워크로 이루어진다
디코더는 6개의 동일한 레이어로 구성되며, 각 레이어가 자기 자신에 대한 Self-Attention을 수행하고 인코더의 출력과의 Attention을 수행하는 서브 레이어가 추가된다.

Self-Attention과 Multi-Head Attention

Self-Attention은 입력된 단어들 간의 상호작용을 고려하여 각 단어가 다른 단어와 어떤 관계를 맺는지 파악하는 기법이다.
Multi-head Attention은 하나의 attention 헤드만을 사용하는 Self-Attention 방식을 확장시켜, 여러 개의 attention 헤드를 사용해 다양한 정보를 병렬로 처리한다. 이는 데이터를 여러 관점에서 분석할 수 있게 해준다.

Positional Encoding

Transformer는 순차적인 구조가 없기 때문에, 어떤 단어를 입력할 때 각각의 위치 정보를 모델에 추가해야 한다.
이를 위해 사용되는 Positional Encoding은 각 단어의 위치를 나타내는 정보를 벡터로 변환하여 입력 임베딩에 추가하는 방식으로 sin과 cos함수를 사용하여 단어의 위치 정보를 주기적으로 매핑해 문장 내의 어느 위치에 있는지 알 수 있도록 한다.

의의

Transformer는 기존의 순차적 처리 방식 대신 입력 시퀀스를 동시에 처리할 수 있어 훈련 속도가 빠르며, 긴 시퀀스 간의 의존 관계를 더 잘 학습할 수 있기 때문에 효율적이다.