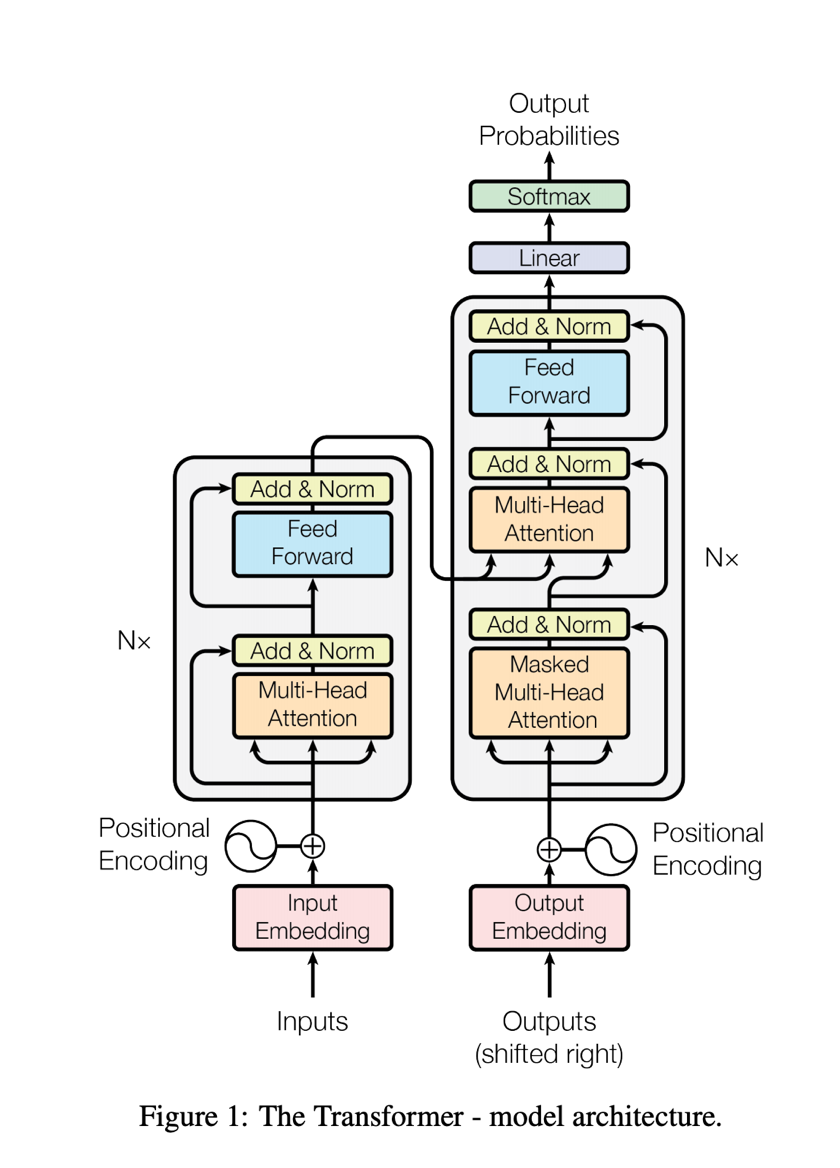
기존에 지배적이던 모델은 RNN이나 CNN에서 Attention Mechanism을 적용한 모델들이다. 구글이 이번 논문에서 발표한 Transformer는 Attention만으로 인코더와 디코더를 구성함으로써 보다 적은 계산량으로 높은 성능을 내는 모델이다. RNN모델은 문장이 길어질수록 메모리의 제약때문에 앞에 나온 sequence들을 반영하지 못한다는 단점이 있고, 기존 RNN을 이용하면 “The animal didn’t cross the street because it was too tired.” 라는 문장이 있을 때 it이 무엇인지 알아낼 수 없었다. 하지만 Transformer의 self-attention 기법을 사용하면서 그것이 가능해졌다.

**Model Architecture**

나이프이(가) 표시된 사진

자동 생성된 설명트랜스포머의 가장 주요한 특징은 벡터간 계산이 필요하던 것을 행렬로 바꾸어 병렬처리를 가능하게 했다는 것이다. Positional encoding으로 단어들의 위치정보를 계산하였으며, residual connection을 사용하여 레이어를 정규화 하면서 안정성을 더해주었다. 또한, 인코더의 input과 output의 차원이 같기 때문에 여러 인코더를 배치할 수 있다는 특징이 있다. 디코더 역시 인코더와 비슷하지만, 다른점은 마스킹을 적용한다는 것이다. 아직 나타나지 않은 단어에 대한 어텐션을 배재시키기 위함이다. 멀티헤드 어텐션에서는 key, value, query로 연산을 하고 피드포워드를 통해 행렬을 벡터로 변환시킨 후 리니어와 소프트맥스 레이어를 통해 벡터를 단어로 변환시킨다.

소프트맥스함수를 거친 이후에 label smoothing 기법이 사용되는데, 예를들어 thank you 라는 시퀀스가 한국어로 ‘고마워’ 와 ‘감사합니다’ 라는 라벨로 학습되어야 할 때, 고마워와 감사합니다는 전혀 상이한 벡터이기 때문에 학습에 지장이 생길 수 있음. 그러나 원핫벡터가 아닌 레이블 스무딩을 이용하게 되면 정답에 가까운 벡터값은 1에 가까운 실수로, 정답에 먼 벡터값은 0에 가까운 실수로 표현되기 때문에 어느정도 보완이 될 수 있음.