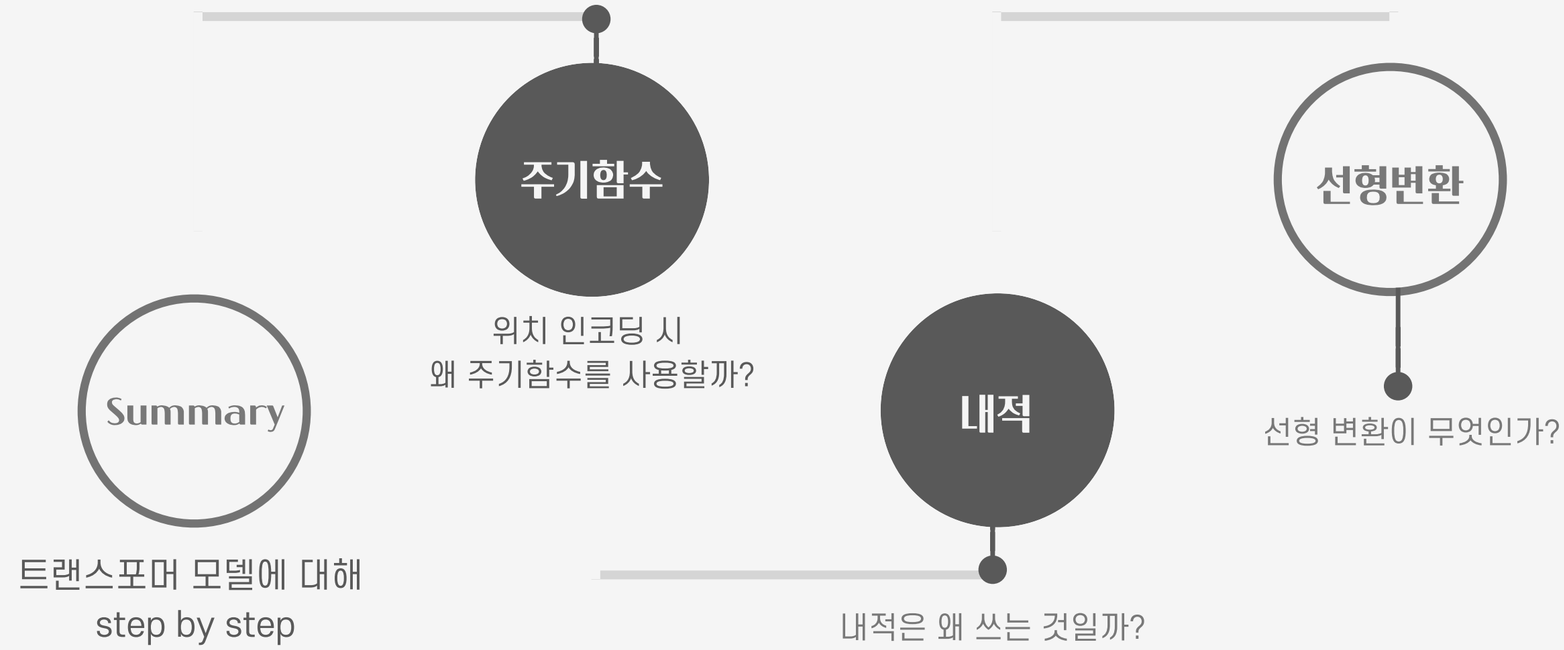
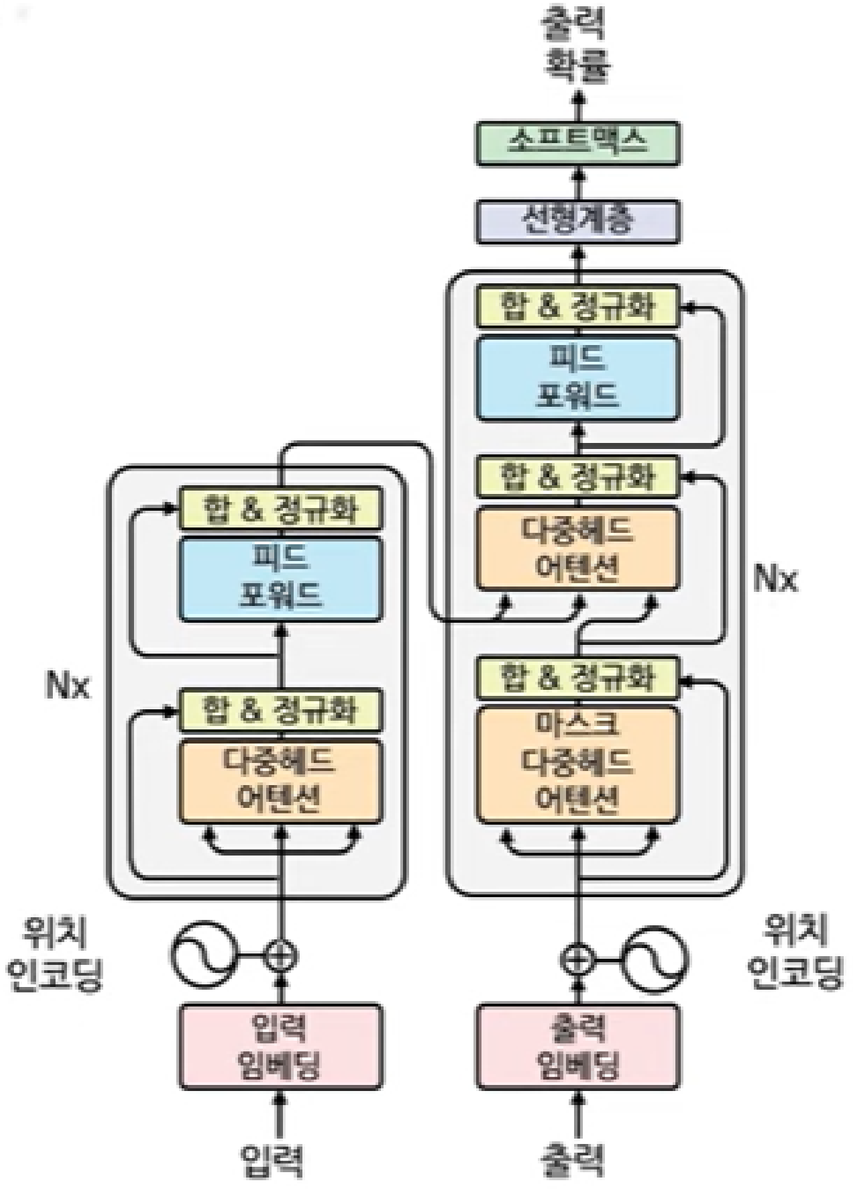
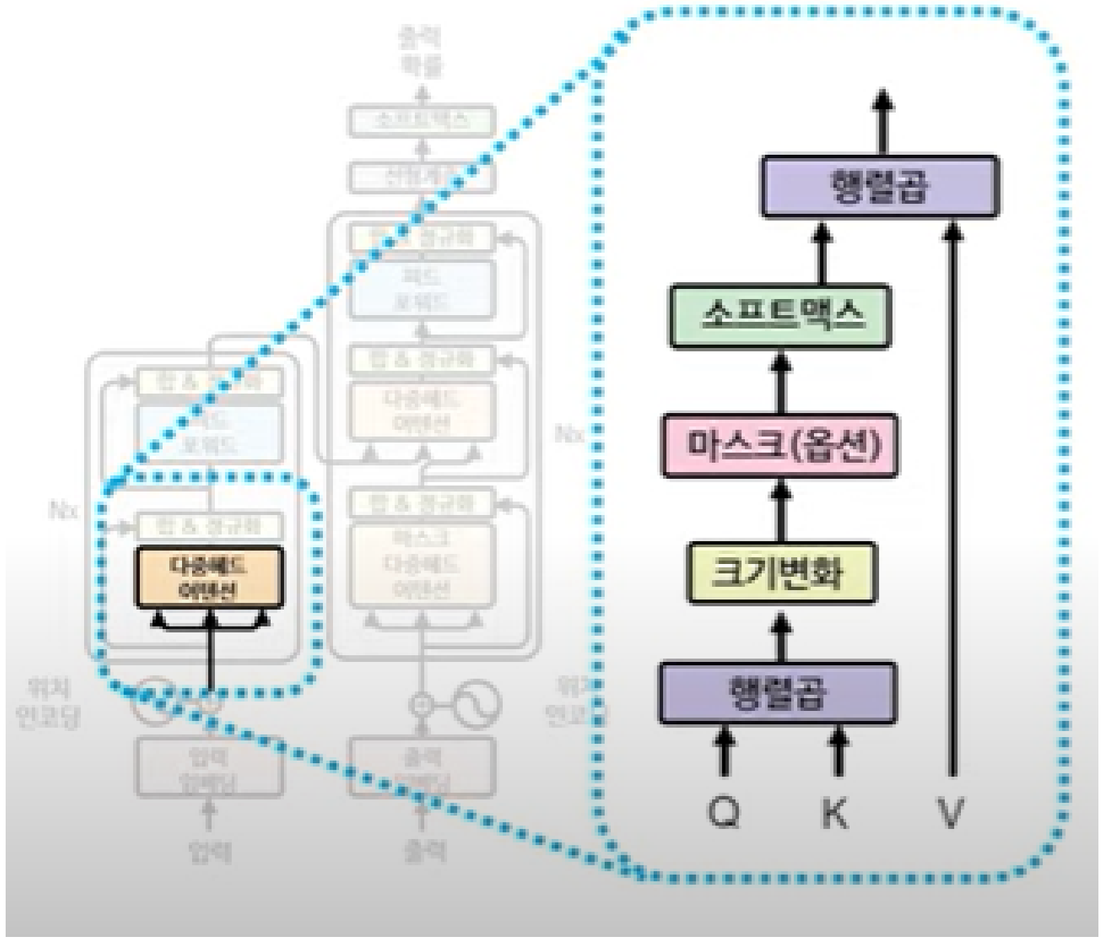
Transformer

을 위한 수학적 개념




출력
확률
소프트맥스
선형계층
합 & 정규화
피드
포워드
합 & 정규화
다중헤드
어텐션
합 & 정규화
마스크
다중헤드
어텐션
Nx
합 & 정규화
피드
포워드
합 & 정규화
다중헤드
어텐션
Nx
위치
인코딩
위치
인코딩
입력
임베딩
출력
임베딩
입력
출력


행렬곱
소프트맥스
마스크(옵션)
크기변화
행렬곱
Q
K
V

주기함수

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

1. 주기적 특성으로 인한 일반화:
주기함수는 시퀀스 길이에 관계없이 일관된 위치 정보를 제공합니다. 이는 모델이 다양한 길이의 시퀀스에 대해 잘 일반화할 수 있게 합니다.

2. 다양한 주기로 인한 고유성:
sin과 cos 함수는 서로 다른 주기로 변합니다. 이는 각 위치에 대해 고유한 인코딩 벡터를 제공하여, 모델이 위치 정보를 명확히 구분할 수 있게 합니다.

3. 상대적 위치 정보 제공:
사인과 코사인 함수의 주기적 특성은 위치 간의 상대적 거리를 유지합니다. 이는 모델이 위치 간의 관계를 더 잘 이해하고 학습할 수 있게 합니다.

내적

$$Attention(Q, K, V) = softmax\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) V$$

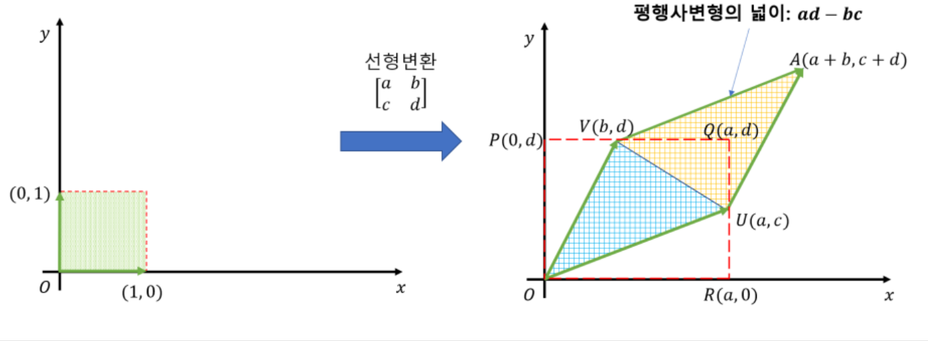
Attention의 목표는 value를 통해 가중치 합계를 계산하는 것.
--> 각 Value의 가중치는 주어진 Query와 Key가 얼마나 유사한가.

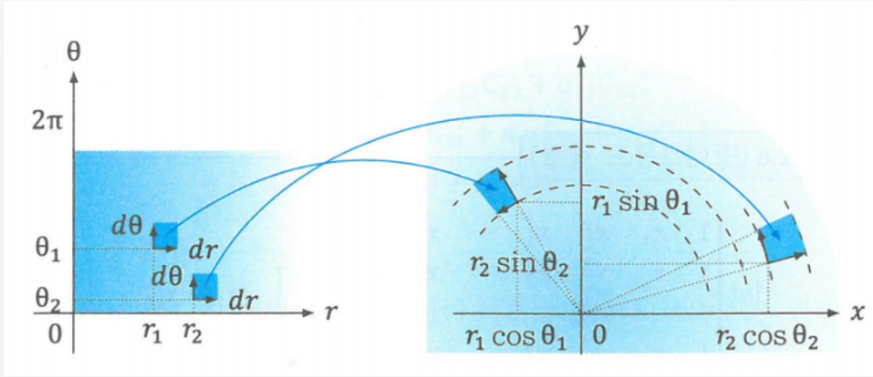
선형변환

$$FFN(x) = \underbrace{\max(0, xW_1 + b_1)}_{\text{ReLU}} \underbrace{W_2 + b_2}_{\text{linear transformation}}$$

linear transformation

<행렬식의 기하학적 의미>





선형 변환을 사용하는 이유는 모델의 표현력을 향상시키고, 입력 데이터의 다양한 특성을 효과적으로 학습할 수 있게 하며, 병렬 처리를 통해 효율성을 높일 수 있기 때문입니다. 단순히 계산하는 것보다 선형 변환을 통해 얻는 이점은 Transformer 모델이 복잡한 패턴과 관계를 더 잘 학습하고 처리할 수 있게 합니다.

Thank you!