Week 8 Report: Multi-Head Attention은 어떻게 다양한 관계를 학습하는가?

이번 주에는 Transformer의 핵심 모듈인 Multi-Head Attention이 왜 필요한지, 그리고 여러 head를 병렬로 운용함으로써 어떻게 다양한 패턴을 포착하는지를 수식적, 구조적, 그리고 엄밀성있는 관점에서 분석한다.

🔧 요약 답변

- W_i^Q, W_i^K, W_i^V는 각 head마다 다른 학습 가중치 파라미터(weight matrix)입니다.

- 각 head는 입력의 다른 부분, 혹은 다른 의미 관계에 집중하도록 설계되어 있습니다.

- 따라서 head는 "관계의 시야" 또는 "관심의 방향"을 분리해서 병렬로 보는 역할을 합니다.

🔍 정의 및 수식 구조

Head란?

입력 쿼리(Q), 키(K), 값(V)에 대해 독립적인 Attention 연산을 수행하는 단위

각 head는 다음과 같은 연산을 수행:

$\text{head}_i = \text{Attention}(Q \cdot W_i^Q, K \cdot W_i^K, V \cdot W_i^V)$

i번째 head는 쿼리 Q, 키 K, 값 V에 대해 각각 고유한 가중치 행렬 W_i^Q, W_i^K, W_i^V를 곱한 뒤, attention 연산을 수행한다.

W_i^Q, W_i^K, W_i^V: i번째 head만의 학습 가능한 가중치 행렬

W^o: 모든 head 결과를 결합한 후 최종 출력으로 투영하는 행렬

전체 구조:

$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_h) \cdot W^o$

모든 head의 출력을 순서대로 나란히 결합한 후, 출력 투영 행렬 W^o를 적용하여 최종 출력 값을 계산한다.

$Q, K, V \in \mathbb{R}^{n \times d\_model}$ (여기서 n은 토큰의 개수)

$W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d\_model \times d\_k}$

$W^o \in \mathbb{R}^{hd\_k \times d\_model}$

일반적으로 $d\_k = d\_model / h$

🎯 단일 Head vs Multi-Head 비교

단일 Head는 하나의 시야만 제공하지만, 다중 Head는 병렬적인 여러 시야를 통해 다양한 의미/구문 관계를 학습할 수 있다.

🧬 Head 간의 처리 관계

Head는 순차적으로 처리되는 것이 아니라 병렬적으로 동시 처리된다.

입력 Q, K, V는 동일하지만, 각 head는 자기만의 가중치 행렬을 사용하여 서로 다른 시야를 학습한다.

각 Head는 Q, K, V를 low-rank subspace로 투영하여 다른 종류의 관계를 추출한다.

수식적으로 엄밀한 해석

$Q \cdot W_i^Q$는 선형 변환에 의한 low-rank projection으로, 각 Head는 입력 공간의 서로 다른 부분공간에 대해 연산을 수행한다.

각 Attention은 bounded linear operator로 해석 가능:

$\text{head}_i(x) = \int_\Omega \alpha_i(x, y) \cdot V(y)\, d\mu(y)$

입력 x에 대한 head_i의 출력은, 확률 커널 $\alpha_i(x, y)$를 이용해 값 벡터 $V(y)$에 대한 가중합을 적분 형태로 계산한 것이다.

$\alpha_i(x, y) = \exp(Q(x) \cdot K(y)^T / \sqrt{d_k}) \,/\, \int_\Omega \exp(Q(x) \cdot K(z)^ \,/\, \sqrt{d_k})\, d\mu(z)$

→ 이 적분은 르벡 적분 관점에서 해석 가능하며, 각 위치에서의 softmax 가중치 계산이 측도적으로 정의된 공간에서 이뤄진다고 볼 수 있습니다.

**여기서 커널 $\alpha_i(x, y)$ 는 확률적 의미를 가지는 함수로서, attention 메커니즘에서 입력 x가 어느 위치 y에 얼마나 집중하는지를 나타내는 weight입니다. 이 커널은 softmax를 통해 정의되며, $\alpha_i(x, y)$를 만족하므로 확률 분포로 해석할 수 있습니다.

이는 Attention이 측도 공간 상의 확률 kernel operator임을 의미하며, 각 Head는 서로 다른 kernel로 작동한다.

✔ 정리

- 각 Head는 서로 다른 의미적 관계를 병렬적으로 파악하기 위해 고유한 투영 행렬을 가진다.

- Attention 연산은 함수 공간 상의 확률적 적분 연산으로 해석될 수 있다.

✔

항목	설명
W_i^Q, W_i^K, W_i^V	각 head의 개별 선형 투영 행렬
W^O	병렬 attention 출력 결합 후 투영 행렬
다중 Head	관계 표현의 다양성과 병렬성 확보
수학적 해석	저차원 투영 + softmax 커널로 인한 확률 가중 평균
측도 관점 해석	softmax는 가측성 및 정규화 가능한 르벡 커널로 표현 가능
커널 해석	$\alpha_i(x, y)$는 attention이 위치 에 얼마만큼 집중하는지를 나타내는 확률 커널