RoFormer: Enhanced Transformer with Rotary Position Embedding

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, Yunfeng Liu

발표자: 나 리 나

기존의 Positional Embeddings

- Absolute Positional Embeddings
    - 입력 시퀀스의 각 토큰에 고유한 위치 식별자를 제공

기존의 Positional Embeddings

- Relative Positional Embeddings
  - 토큰 간의 상대적 거리에 대한 정보를 인코딩

$$A_{i,j}^{abs} = \underbrace{E_{x_i}^T W_q^T W_k E_{x_j}}_{\text{(a) Content-to-Content}} + \underbrace{E_{x_i}^T W_q^T W_k U_j}_{\text{(b) Content-to-Position}} + \underbrace{U_i^T W_q^T W_k E_{x_j}}_{\text{(c) Position-to-Content}} + \underbrace{U_i^T W_q^T W_k U_j}_{\text{(d) Position-to-Position}}$$

$$A_{i,j}^{abs} = \left(W_q\left(E_{x_i} + U_i\right)\right)^T \left(W_k\left(E_{x_j} + U_j\right)\right)$$

Rotary Positional Embeddings(RoPE)

$$\langle f_q(\boldsymbol{x}_m, m), f_k(\boldsymbol{x}_n, n) \rangle = g(\boldsymbol{x}_m, \boldsymbol{x}_n, m-n).$$

- RoPE는 각 위치에 대한 회전 행렬을 사용하여 절대 위치를 인코딩하고, 상대적 위치 의존성을 명시적으로 셀프-어텐션 공식에 통합
- 토큰의 위치에 따라 인베딩이 회전함으로써 위치 정보를 효과적으로 모델링

Rotary Positional Embeddings(RoPE): A 2D Case

$$
\begin{aligned}
f_q(\boldsymbol{x}_m, m) &= (\boldsymbol{W}_q \boldsymbol{x}_m) e^{im\theta} \\
f_k(\boldsymbol{x}_n, n) &= (\boldsymbol{W}_k \boldsymbol{x}_n) e^{in\theta} \\
g(\boldsymbol{x}_m, \boldsymbol{x}_n, m-n) &= \mathrm{Re}[(\boldsymbol{W}_q \boldsymbol{x}_m)(\boldsymbol{W}_k \boldsymbol{x}_n)^* e^{i(m-n)\theta}]
\end{aligned}
$$

$$
f_{\{q,k\}}(\boldsymbol{x}_m, m) = \begin{pmatrix} \cos m\theta & -\sin m\theta \\ \sin m\theta & \cos m\theta \end{pmatrix} \begin{pmatrix} W^{(11)}_{\{q,k\}} & W^{(12)}_{\{q,k\}} \\ W^{(21)}_{\{q,k\}} & W^{(22)}_{\{q,k\}} \end{pmatrix} \begin{pmatrix} x^{(1)}_m \\ x^{(2)}_m \end{pmatrix}
$$

- 벡터가 2차원 평면에 있을 때의 처리 방식을 설명

Rotary Positional Embeddings(RoPE): General form

$$f_{\{q,k\}}(\boldsymbol{x}_m, m) = \boldsymbol{R}^d_{\Theta,m}\boldsymbol{W}_{\{q,k\}}\boldsymbol{x}_m$$

$$\boldsymbol{R}^d_{\Theta,m} = \begin{pmatrix} \cos m\theta_1 & -\sin m\theta_1 & 0 & 0 & \cdots & 0 & 0 \\ \sin m\theta_1 & \cos m\theta_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cos m\theta_2 & -\sin m\theta_2 & \cdots & 0 & 0 \\ 0 & 0 & \sin m\theta_2 & \cos m\theta_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \cos m\theta_{d/2} & -\sin m\theta_{d/2} \\ 0 & 0 & 0 & 0 & \cdots & \sin m\theta_{d/2} & \cos m\theta_{d/2} \end{pmatrix}$$

- General form은 2차원을 넘어서 더 높은 차원에서의 위치 정보를 처리할 수 있도록 확장

Rotary Positional Embeddings(RoPE)

- 이전 기법과의 비교
  - Additive vs Multiplicative
  - sinsusoid individually vs mix pairs of coordinates

$$\boldsymbol{q}_m^\intercal \boldsymbol{k}_n = (\boldsymbol{R}^d_{\Theta,m}\boldsymbol{W}_q\boldsymbol{x}_m)^\intercal(\boldsymbol{R}^d_{\Theta,n}\boldsymbol{W}_k\boldsymbol{x}_n) = \boldsymbol{x}^\intercal\boldsymbol{W}_q R^d_{\Theta,n-m}\boldsymbol{W}_k\boldsymbol{x}_n$$

Rotary Positional Embeddings(RoPE)

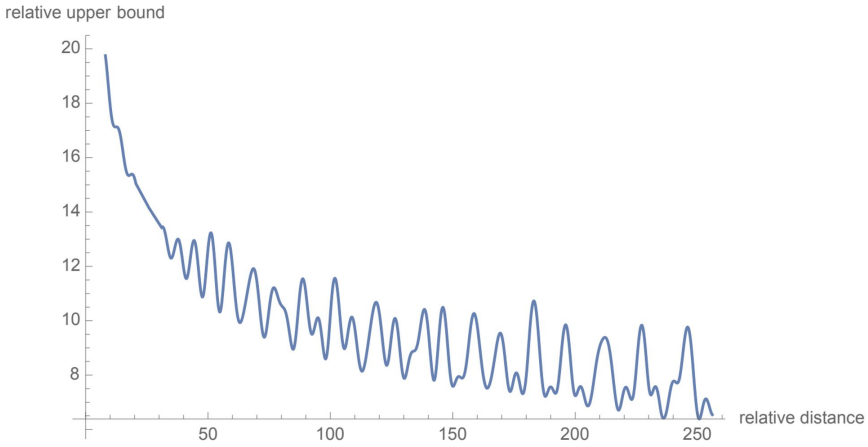

Figure 2: Long-term decay of RoPE.

- 장점
  - Long-term decay
    - 상대 위치가 멀어질수록 내적 값이 감소하는 효과
  - RoPE with linear attention
    - 시퀀스 길이 차원을 감소시키지 않고도 확장성을 유지