

# Self-Attention with Relative Position Representations

Peter Shaw, Jakob Uszkoreit, Ashish Vaswani

발표자: 나 리 나

## 기존의 Transformer

- 'Attention is All You Need' 을 통해 소개
- Self-Attention 메커니즘
- **RNN, CNN과 달리 문서들의 상대적또는 절대적위치 정보가 명시되어 있지 않음**
  - > 때문에 위치 정보를 추가

## Self-Attention

$$z_i = \sum_{j=1}^n \alpha_{ij} (x_j W^V)$$

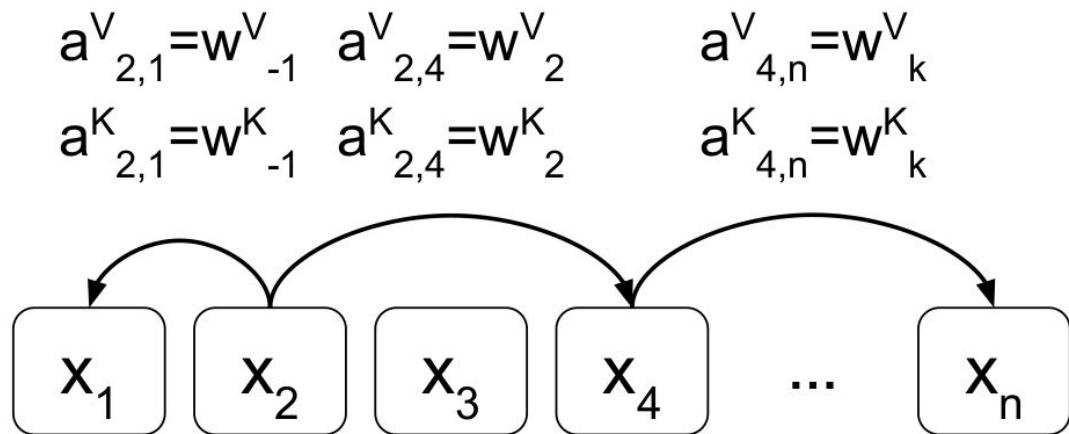
input and output

$$\alpha_{ij} = \frac{\exp e_{ij}}{\sum_{k=1}^n \exp e_{ik}}$$

softmax function

- input data가 Neural Network를 통과한 후, softmax 함수를 거치면 output

## Relation-aware Self-Attention



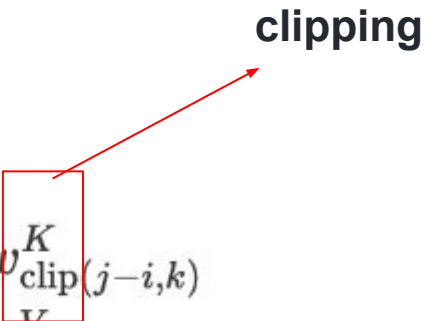
Relative Positional Encoding

상대적 위치 정보

$$e_{ij} = \frac{x_i W^Q (x_j W^K + a_{ij}^K)^T}{\sqrt{d_z}}$$

New attention score

## Relative Position Representations

$$\begin{aligned}a_{ij}^K &= w_{\text{clip}(j-i, k)}^K \\a_{ij}^V &= w_{\text{clip}(j-i, k)}^V \\ \text{clip}(x, k) &= \max(-k, \min(k, x))\end{aligned}$$


- 상대 위치에 대한 최대값:  $k$  지정 후 **clipping**
- 상대적인 위치에 대한 정확한 값은 특정 거리 이상에서는 중요하지 않다는 가정
- 모델은  $2k+1$ 의 고유한 **edge label**들만을 고려

## Efficient Implementation

- 모든 시퀀스 및 **attention head**에 대해 상대 위치 표현 공유

# Experiments

Model	Position Information	EN-DE BLEU	EN-FR BLEU
Transformer (base)	Absolute Position Representations	26.5	38.2
Transformer (base)	Relative Position Representations	<b>26.8</b>	<b>38.7</b>
Transformer (big)	Absolute Position Representations	27.9	41.2
Transformer (big)	Relative Position Representations	<b>29.2</b>	<b>41.5</b>

$k$	EN-DE BLEU
0	12.5
1	25.5
2	25.8
4	25.9
16	25.8
64	25.9
256	25.8