Self-Attention with Relative Position Representations

모두연 NLP Bootcamp
김 성 운

Contents

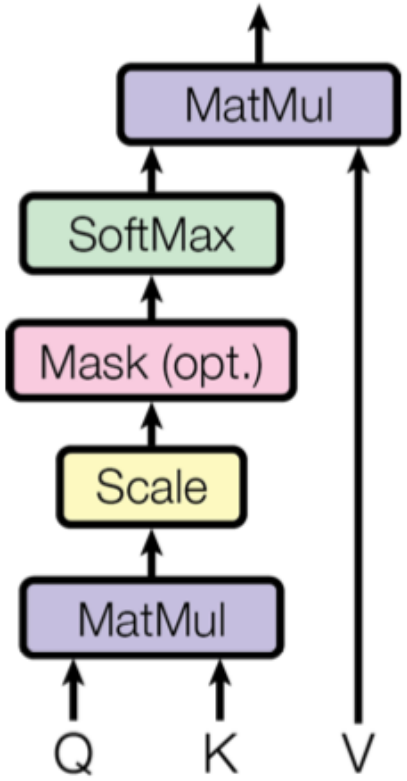
1. Abstract
2. Background
3. Proposed Architecture
4. Experiments and result

Abstract

- Transformer는 RNN, CNN과 달리 구조적으로 relative or absolute position information을 modeling 하지 않음
- Input에 adding representations of absolute positions를 추가 (Positional Encoding)
- Self attention에서 relative positions을 고려하는 방안을 제시
- English-to-German, English-to-French translation tasks에서 각각 1.3 BLEU, 0.3 BLEU 개선
- Combining relative and absolute position은 딱히…

Background

- Self-Attention



$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

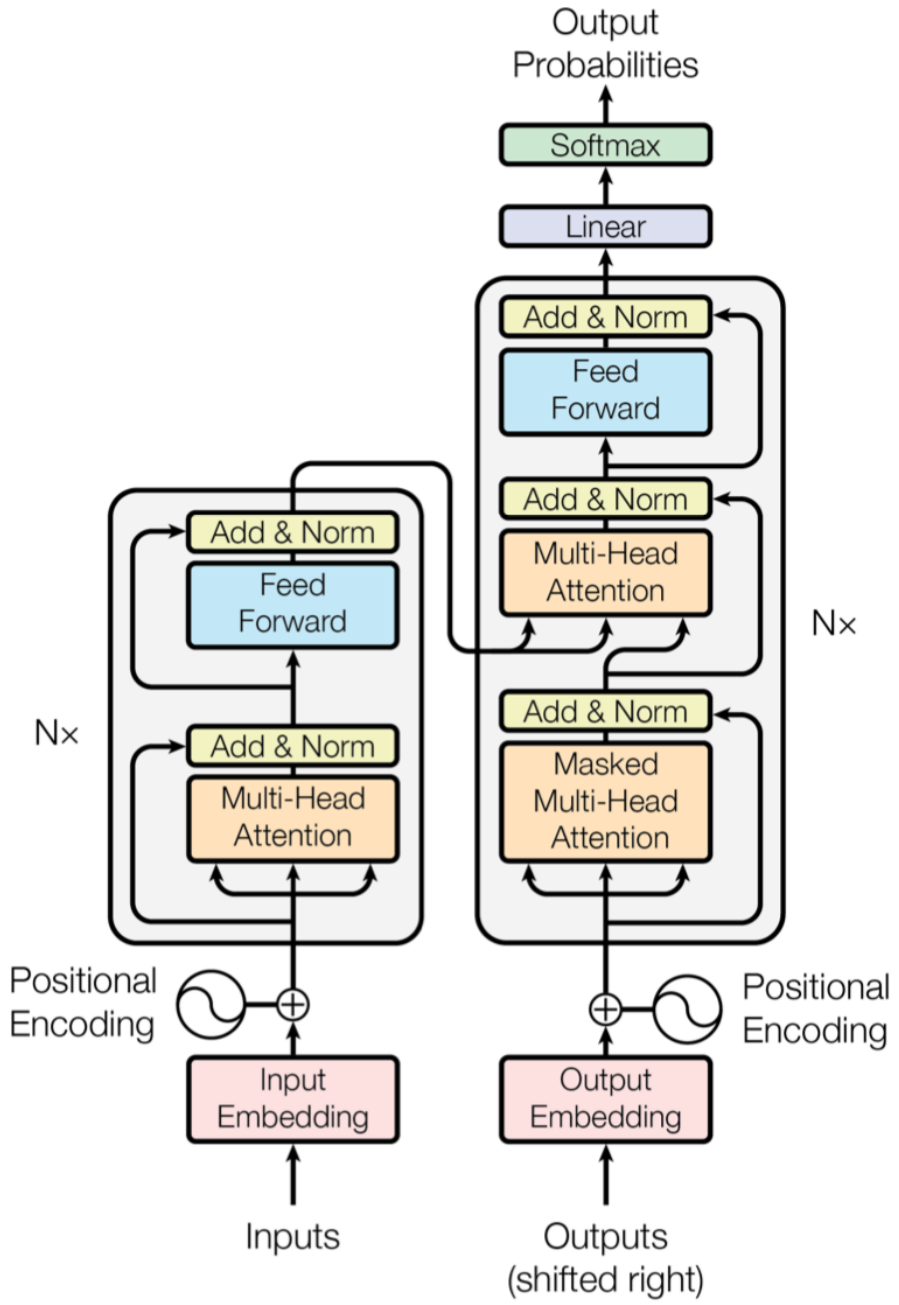


Figure 1: The Transformer - model architecture.

Proposed Architecture

- Relation-aware Self-Attention $\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$

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

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

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

Proposed Architecture

- Relative Position Representations and Efficient Implementation

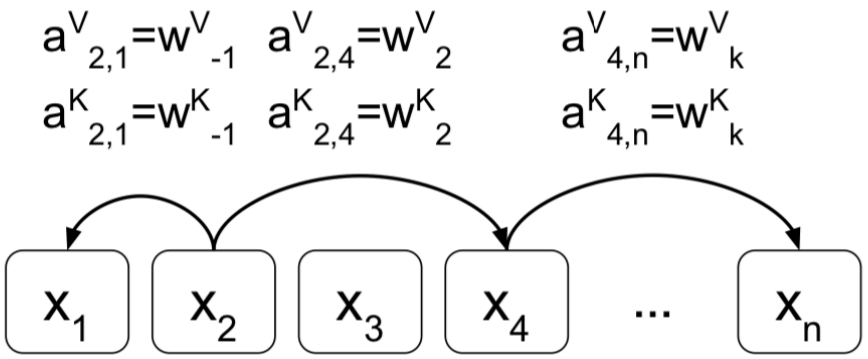


Figure 1: Example edges representing relative positions, or the distance between elements. We learn representations for each relative position within a clipping distance k. The figure assumes $2 <= k <= n - 4$. Note that not all edges are shown.

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

We then learn relative position representations $w^K = (w_{-k}^K, \ldots, w_k^K)$ and $w^V = (w_{-k}^V, \ldots, w_k^V)$ where $w_i^K, w_i^V \in \mathbb{R}^{d_a}$.

```
generate_relative_positions_matrix(7, 3)
```

```
array([[3, 4, 5, 6, 6, 6, 6],
       [2, 3, 4, 5, 6, 6, 6],
       [1, 2, 3, 4, 5, 6, 6],
       [0, 1, 2, 3, 4, 5, 6],
       [0, 0, 1, 2, 3, 4, 5],
       [0, 0, 0, 1, 2, 3, 4],
       [0, 0, 0, 0, 1, 2, 3]], dtype=int32)
```

Experiments and result

- WMT 2014 English-German dataset: 약 4.5M sentence pairs
- WMT 2014 English-French dataset: 약 36M sentence pairs
- 32,768 word-piece vocabulary

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

Table 1: Experimental results for WMT 2014 English-to-German (EN-DE) and English-to-French (EN-FR) translation tasks, using newstest2014 test set.

Experiments and result

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

Table 2: Experimental results for varying the clipping distance, k.

Notably, for $k \geq 2$, there does not appear to be much variation in BLEU scores. However, as we use multiple encoder layers, precise relative position information may be able to propagate beyond the clipping distance.

a_{ij}^V	a_{ij}^K	EN-DE BLEU
Yes	Yes	25.8
No	Yes	25.8
Yes	No	25.3
No	No	12.5

Table 3: Experimental results for ablating relative position representations a_{ij}^V and a_{ij}^K.

Including relative position representations solely when determining compatibility between elements may be sufficient, but further work is needed to deter- mine whether this is true for other tasks.