Attention

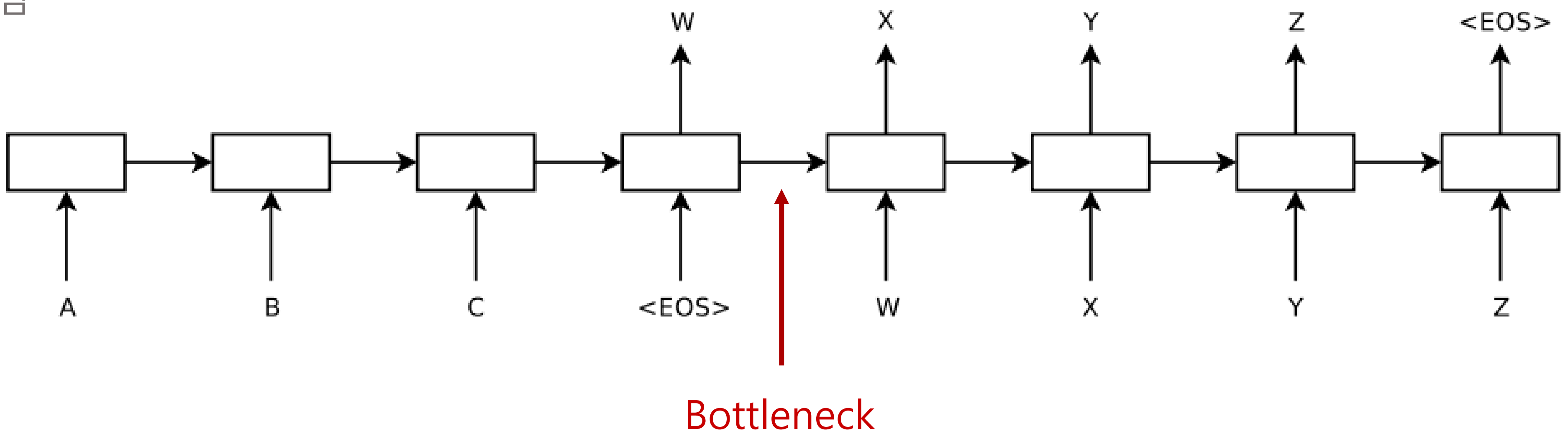
최원혁
20. 04. 13

index

1. Attention
2. Self-Attention

Attention

seq2seq 문제점



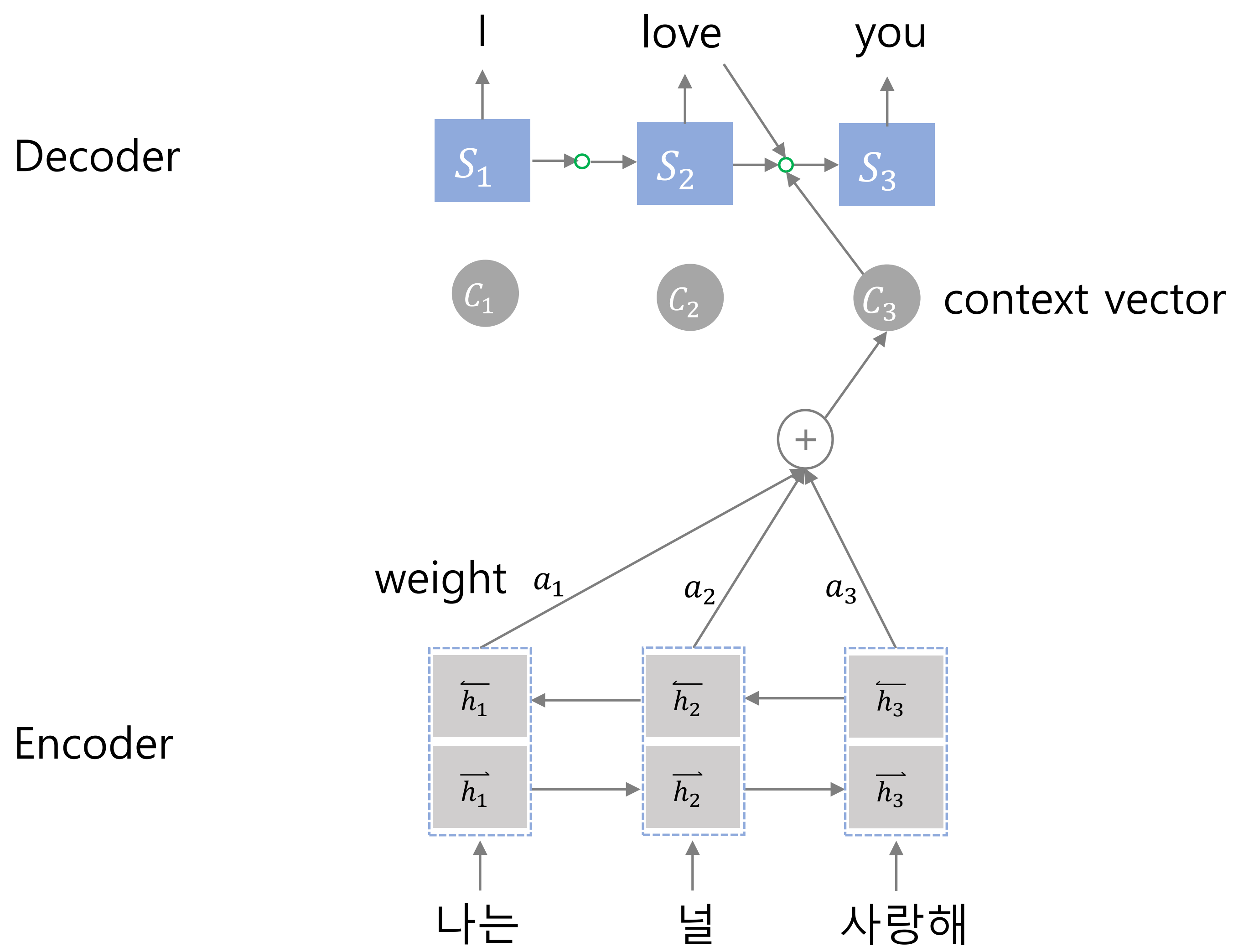
compress all the necessary information of a source sentence into a fixed-length vector

encode the input sentence into a sequence of vectors and chooses a subset of these vectors

Attention!!

Attention

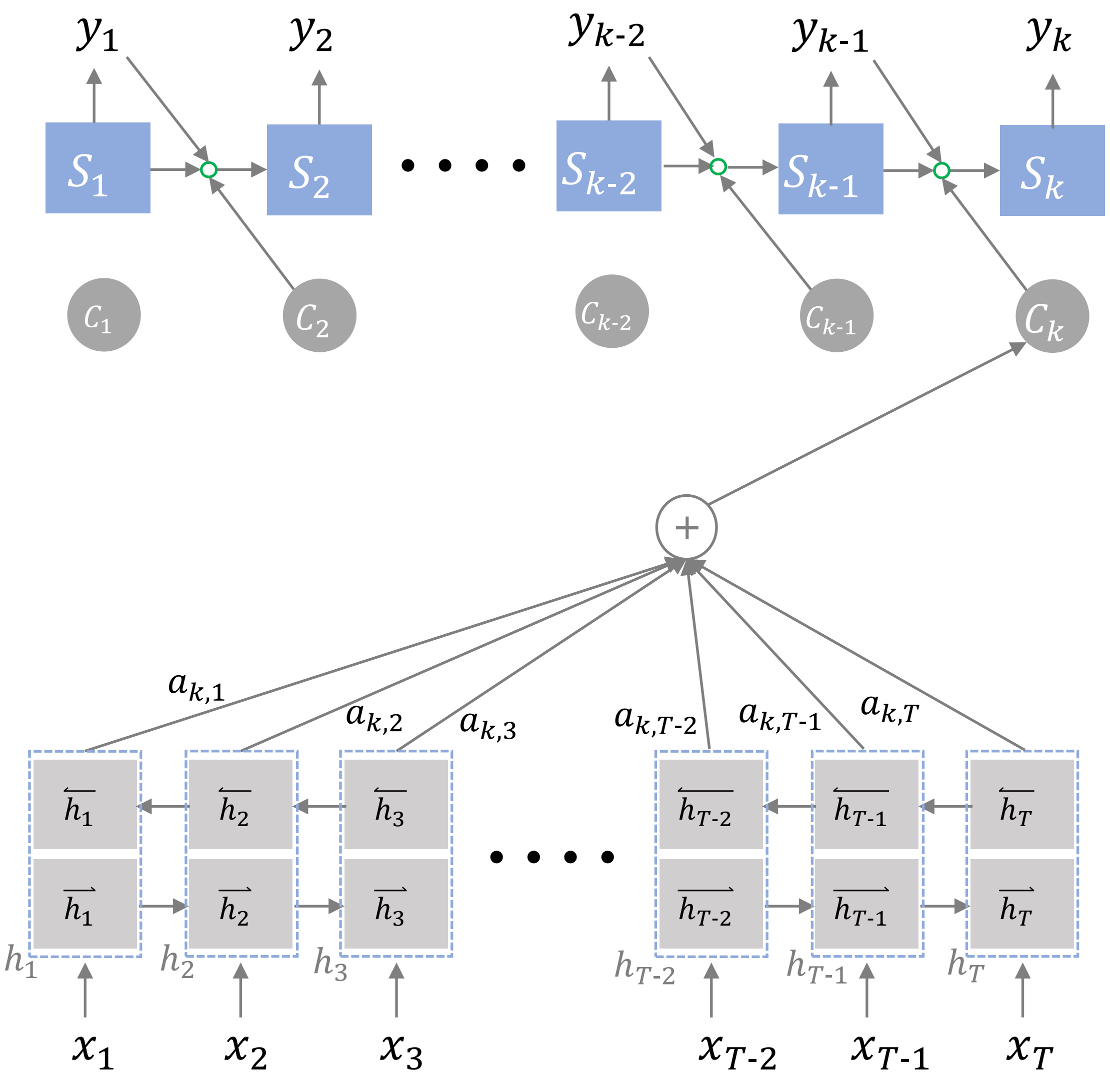
structure



Attention

structure

Decoder



Encoder

$y_k : Output$

$$S_k = f(S_{k\text{-}1}, y_{k\text{-}1}, C_k)$$

$$C_k = \sum_{j=1}^{T} \alpha_{kj} h_j$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{m=1}^{T} \exp(e_{im})}$$

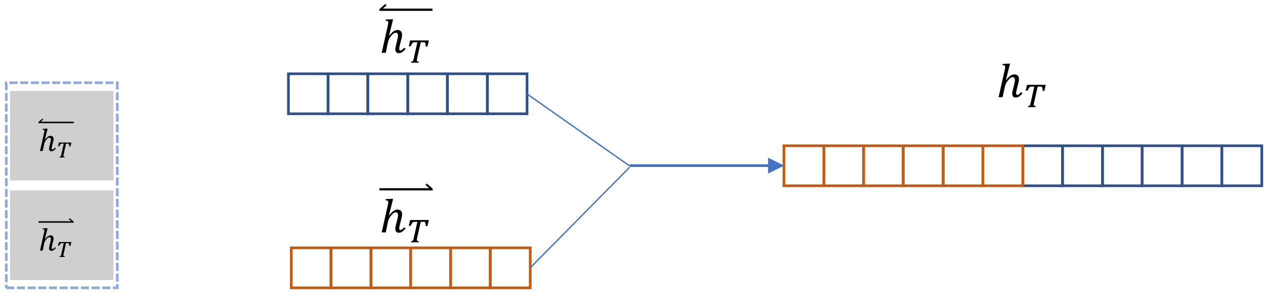
$$e_{ij} = a(S_{i-1}, h_j)$$

$$h_j = [\overrightarrow{h_j^T}; \overleftarrow{h_j^T}]^T$$

$x_r : Input$

Attention

structure



y_k: $Output$

$S_k = f(S_{k\text{-}1}, y_{k\text{-}1}, C_k)$

$$C_k = \sum_{j=1}^{T} \alpha_{kj} h_j$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{m=1}^{T} \exp(e_{im})}$$

$e_{ij} = a(S_{i-1}, h_j)$

$\boxed{h_j = [\overrightarrow{h_j^T}; \overleftarrow{h_j^T}]^T}$

x_r: $Input$

Attention

structure

e_{ij}는 scalar, $a(S_{i-1}, h_j)$에서 a는 alignment model

$$\boxed{e_{ij} = v_a^T \tanh(W_a S_{i-1} + U_a h_j)}$$

$$S_{i-1} \in \mathbb{R}^{2n},\ h_j \in \mathbb{R}^{2n},\ v_a \in \mathbb{R}^{n^*},\ W_a \in \mathbb{R}^{n^* \times n},\ U_a \in \mathbb{R}^{n^* \times 2n}$$

$y_k : Output$

$S_k = f(S_{k\text{-}1}, y_{k\text{-}1}, C_k)$

$$C_k = \sum_{j=1}^{T} \alpha_{kj} h_j$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{m=1}^{T} \exp(e_{im})}$$

$$\boxed{e_{ij} = a(S_{i-1}, h_j)}$$

$h_j = [\overrightarrow{h_j^T}; \overleftarrow{h_j^T}]^T$

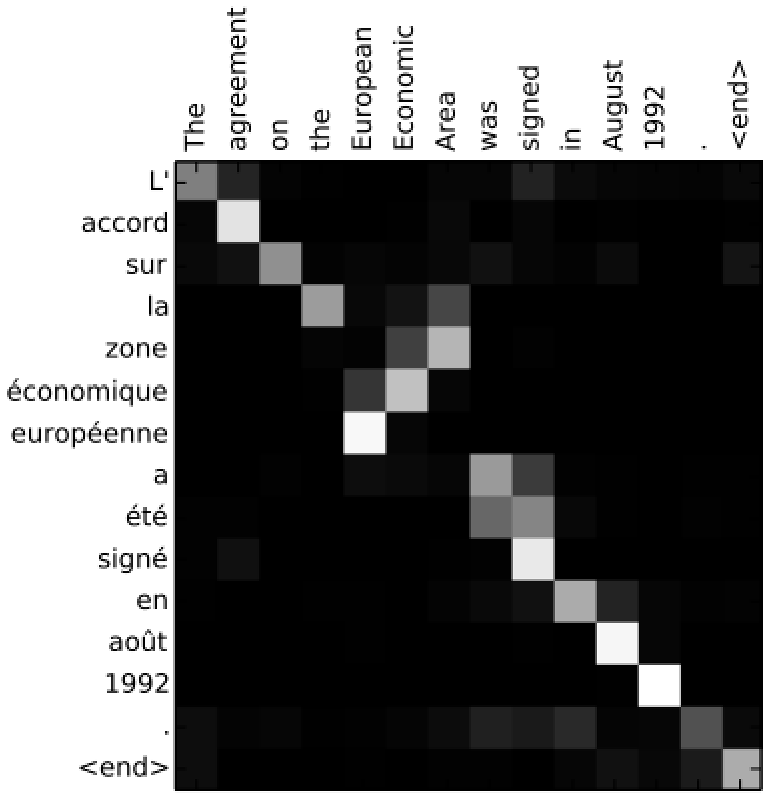
$x_r : Input$

Attention

structure

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{m=1}^{T} \exp(e_{im})}$$

alignment model로부터 계산된
S_{i-1}과 h_j의 관계를 softmax로 출력



$y_k: Output$

$S_k = f(S_{k\text{-}1}, y_{k\text{-}1}, C_k)$

$$C_k = \sum_{j=1}^{T} \alpha_{kj} h_j$$

$$\boxed{\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{m=1}^{T} \exp(e_{im})}}$$

$e_{ij} = a(S_{i-1}, h_j)$

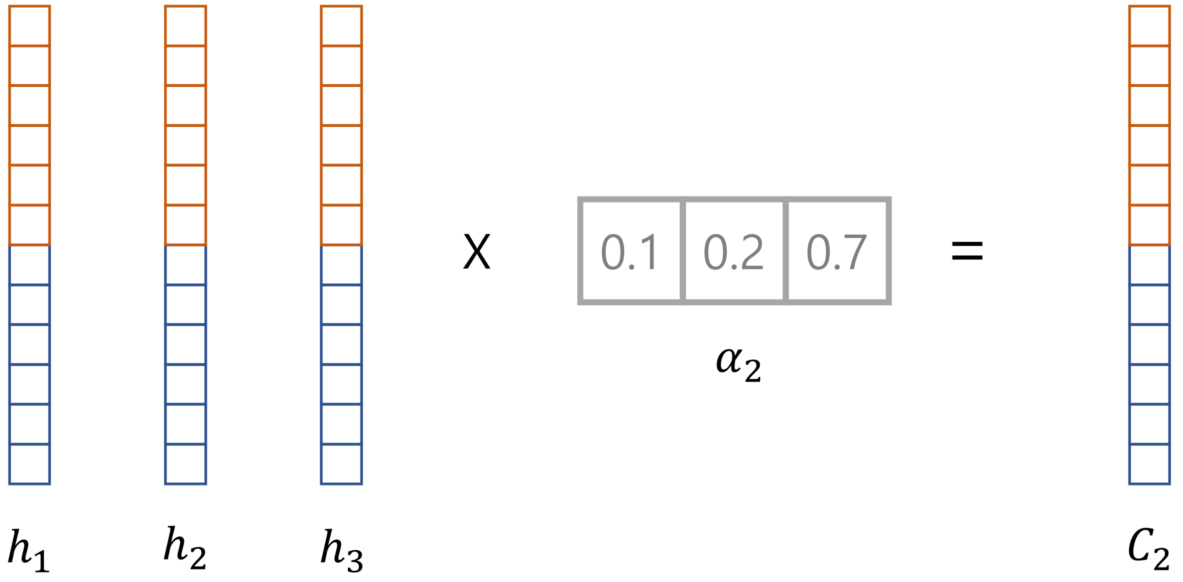
$h_j = [\overrightarrow{h_j^T}; \overleftarrow{h_j^T}]^T$

$x_r: Input$

Attention

structure

context vector C_k



y_k: $Output$

$S_k = f(S_{k\text{-}1}, y_{k\text{-}1}, C_k)$

$$\boxed{C_k = \sum_{j=1}^{T} \alpha_{kj} h_j}$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{m=1}^{T} \exp(e_{im})}$$

$e_{ij} = a(S_{i-1}, h_j)$

$h_j = [\overrightarrow{h_j^T}; \overleftarrow{h_j^T}]^T$

x_r: $Input$

Attention

structure

$S_k = f(S_{k\text{-}1}, y_{k\text{-}1}, C_k)$ f 는 RNN model function

$$s_i = (1 - z_i) \circ s_{i-1} + z_i \circ \tilde{s}_i,$$

$$\tilde{s}_i = \tanh\left(W E y_{i-1} + U\left[r_i \circ s_{i-1}\right] + C c_i\right)$$
$$z_i = \sigma\left(W_z E y_{i-1} + U_z s_{i-1} + C_z c_i\right)$$
$$r_i = \sigma\left(W_r E y_{i-1} + U_r s_{i-1} + C_r c_i\right)$$

원 논문에서는 GRU 사용

$y_k: Output$

$$\boxed{S_k = f(S_{k\text{-}1}, y_{k\text{-}1}, C_k)}$$

$$C_k = \sum_{j=1}^{T} \alpha_{kj} h_j$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{m=1}^{T} \exp(e_{im})}$$

$$e_{ij} = a(S_{i-1}, h_j)$$

$$h_j = [\overrightarrow{h_j^T}; \overleftarrow{h_j^T}]^T$$

$x_r: Input$

seq2seq context vector를 개선시키기 위해 제안되었지만, 현재는 다양한 딥러닝 모델링의 하나의 기술로 이용

Attention weight matrix 시각화를 통해 모델의 안전성을 점검하고 오류의 원인을 찾을 수 있음

Transformer, Bert로 이어지는 모델을 통해 자연어 처리 성능 향상의 기초가 되었음

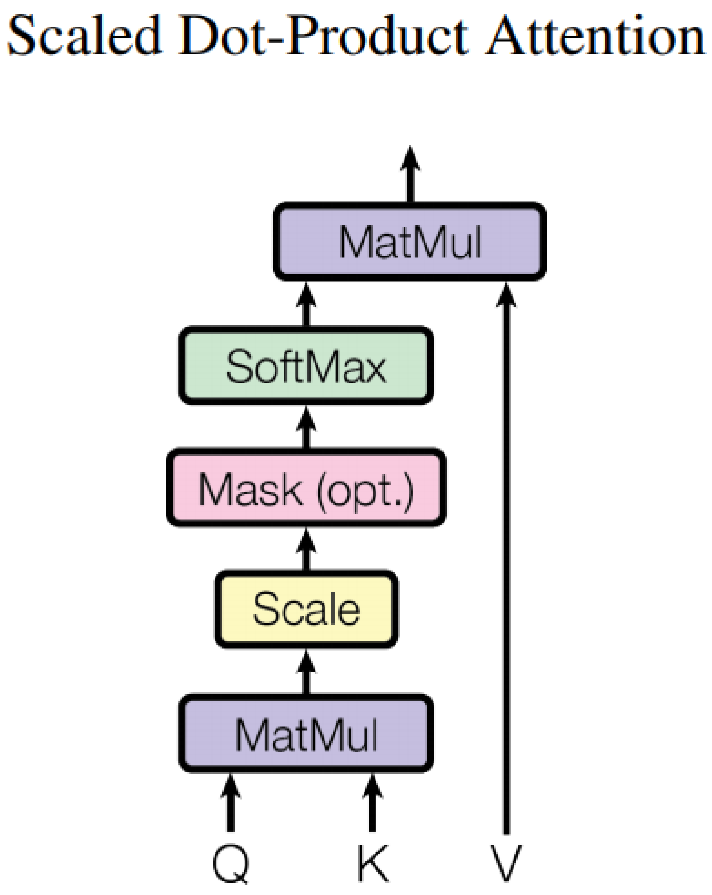
Attention과 차이점

Attention은 서로 다른 대상의 관계를 파악

Self-Attention은 자기 자신의 관계를 파악

Self-Attention

Scaled Dot-Product Attention



Q=K=V in Self-Attention

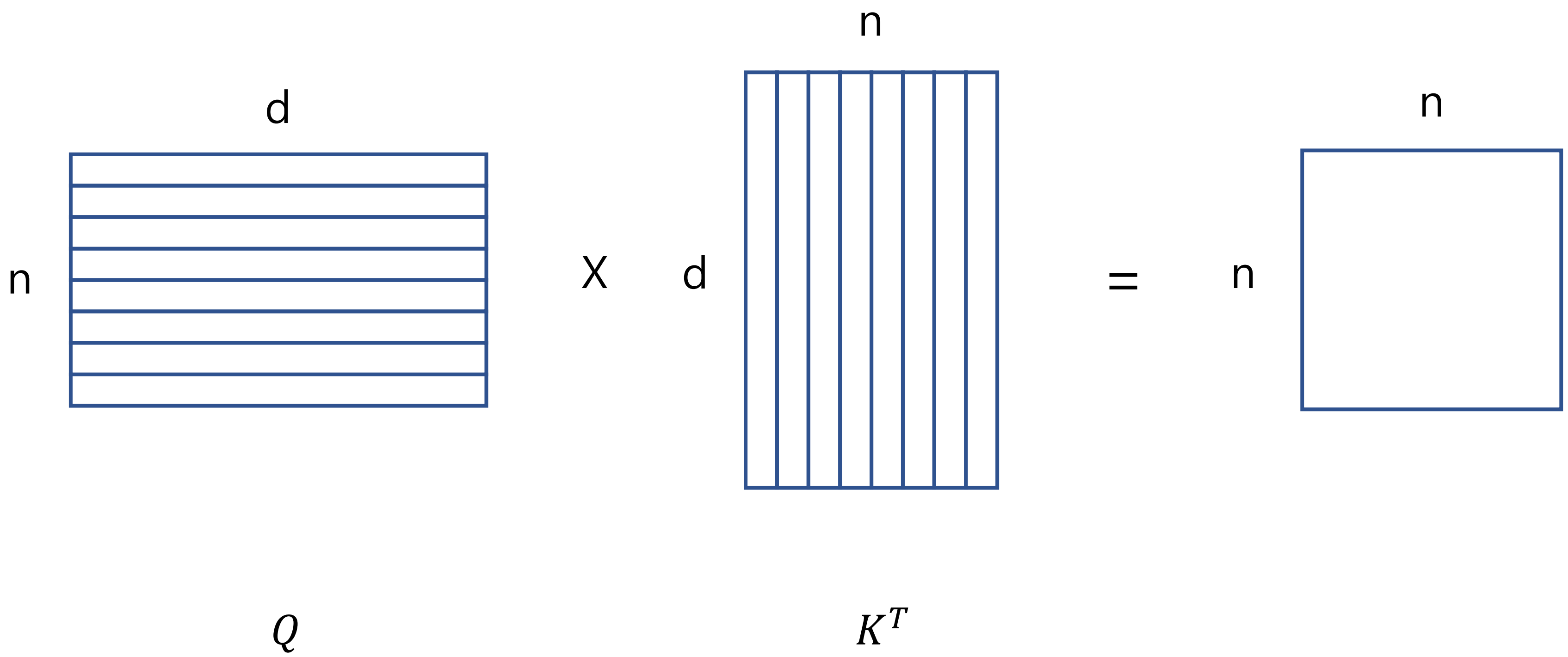
$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

$\sqrt{d_k}$는 QK^T의 값이 vector dimension에 따라서 큰 값을 가져서 보정

Self-Attention

QK^T

$$Q=K=V \in \mathbb{R}^{n\times d}$$



Self-Attention

QK^T

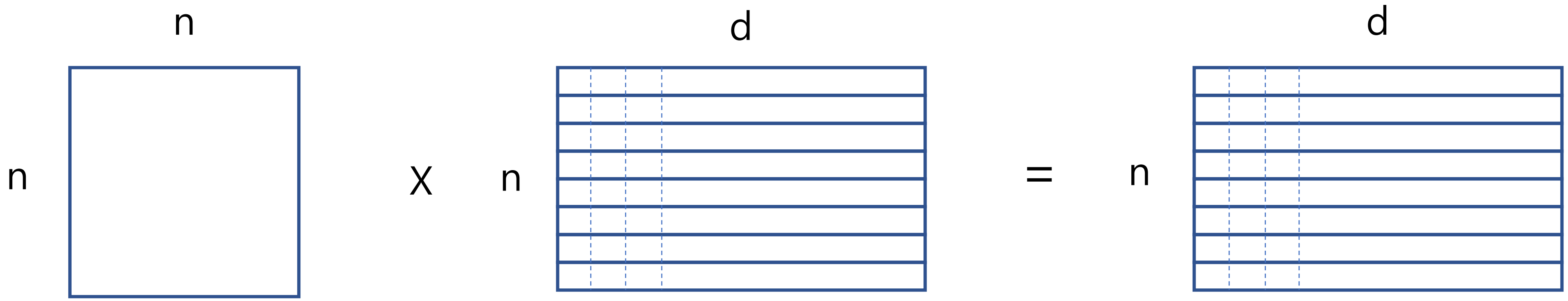
$$
\begin{array}{cc}
 & \begin{array}{cccccc} \quad Hello & \quad , & \quad how & \quad are & \quad you & \quad ? \end{array} \\
\begin{array}{c} Hello \\ , \\ how \\ are \\ you \\ ? \end{array} &
\begin{pmatrix}
78.49 & 43.29 & 1.2 & 41.74 & 91.43 & 74.47 \\
95.84 & 28.78 & 57.13 & 68.20 & -60.94 & 26.85 \\
-95.69 & -52.16 & 17.00 & 45.71 & 48.49 & 64.35 \\
-69.92 & 85.16 & 94.94 & 91.04 & -92.83 & 77.49 \\
65.85 & 55.85 & 62.54 & -97.46 & 76.38 & 13.20 \\
-30.05 & -4.52 & 76.02 & 42.35 & 15.29 & 63.61
\end{pmatrix}
\end{array}
$$

$$\Big\downarrow \text{softmax}$$

$$
\begin{array}{ccc}
 & \begin{array}{cccccc} \quad Hello & \quad , & \quad how & \quad are & \quad you & \quad ? \end{array} & \\
\begin{array}{c} Hello \\ , \\ how \\ are \\ you \\ ? \end{array} &
\begin{pmatrix}
72.40 * 10^{-06} & 1.23 * 10^{-21} & 6.51 * 10^{-40} & 2.62 * 10^{-22} & 9.99 * 10^{-01} & 4.30 * 10^{-08} \\
1.00 * 10^{+00} & 7.51 * 10^{-30} & 1.54 * 10^{-17} & 9.91 * 10^{-13} & 8.15 * 10^{-69} & 1.09 * 10^{-30} \\
3.12 * 10^{-70} & 2.51 * 10^{-51} & 2.72 * 10^{-21} & 8.03 * 10^{-09} & 1.29 * 10^{-07} & 9.99 * 10^{-01} \\
2.47 * 10^{-72} & 5.54 * 10^{-05} & 9.80 * 10^{-01} & 1.98 * 10^{-02} & 2.77 * 10^{-82} & 2.58 * 10^{-08} \\
2.67 * 10^{-05} & 1.21 * 10^{-09} & 9.75 * 10^{-07} & 3.17 * 10^{-76} & 9.99 * 10^{-01} & 3.64 * 10^{-28} \\
8.59 * 10^{-47} & 1.05 * 10^{-35} & 9.99 * 10^{-01} & 2.38 * 10^{-15} & 4.21 * 10^{-27} & 4.07 * 10^{-06}
\end{pmatrix}
& \begin{array}{c} = 1 \\ = 1 \\ = 1 \\ = 1 \\ = 1 \\ = 1 \end{array}
\end{array}
$$

Self-Attention

$softmax(QK^T)V$



$softmax(QK^T)$

V

Self-Attention

$softmax(QK^T)V$

$$
\begin{array}{c|cccccc}
 & Hello & , & how & are & you & ? \\
\hline
Hello & 0.1 & 0 & 0.06 & 0.1 & 0.6 & 0.14 \\
, & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
how & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
are & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
you & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
? & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots
\end{array}
\quad
\begin{array}{c}
d_v \\
\hline
v_{Hello} \\
v_{,} \\
v_{how} \\
v_{are} \\
v_{you} \\
v_{?}
\end{array}
=
$$

$softmax(QK^T)$ V

Self-Attention

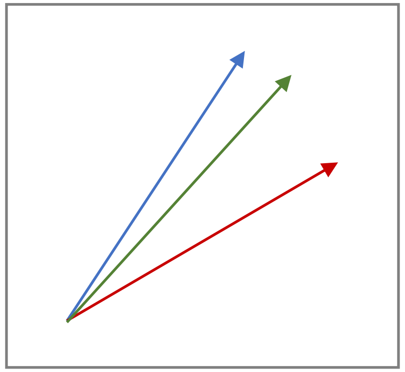
$softmax(QK^T)V$

$$\begin{array}{cc} & < - - - - - - - - - - - - - - - d_v - - - - - - - - - - - - - - - > \\ \begin{array}{c} Hello \\ , \\ how \\ are \\ you \\ ? \end{array} & \begin{pmatrix} 0.1v_{Hello} + 0v_{,} + 0.06v_{how} + 0.1v_{are} + 0.6v_{you} + 0.14v_{?} \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \end{pmatrix} \end{array}$$

관계가 가까운 단어의 vector를 더한다

Self-Attention

$softmax(QK^T)V$



Self-Attention

conclusion

Self-Attention은 자기 요소들 간의 관계를 알아낸다

End

Reference

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." *arXiv preprint arXiv:1409.0473* (2014).

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." *Advances in neural information processing systems*. 2014.

Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017.

Blog "Dissecting BERT Part 1: The Encoder https://medium.com/dissecting-bert/dissecting-bert-part-1-d3c3d495cdb3

Blog "Attention mechanism in NLP. From seq2seq + attention to BERT https://lovit.github.io/machine%20learning/2019/03/17/attention_in_nlp/

Youtube "십분딥러닝_12_어텐션(Attention Mechanism)" https://www.youtube.com/watch?v=6aouXD8WMVQ