Attention is all you need

Author: Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I..

Publish: *Advances in Neural Information Processing Systems*

Pp: 5998 - 6008.

Presenter: WENWEI KANG

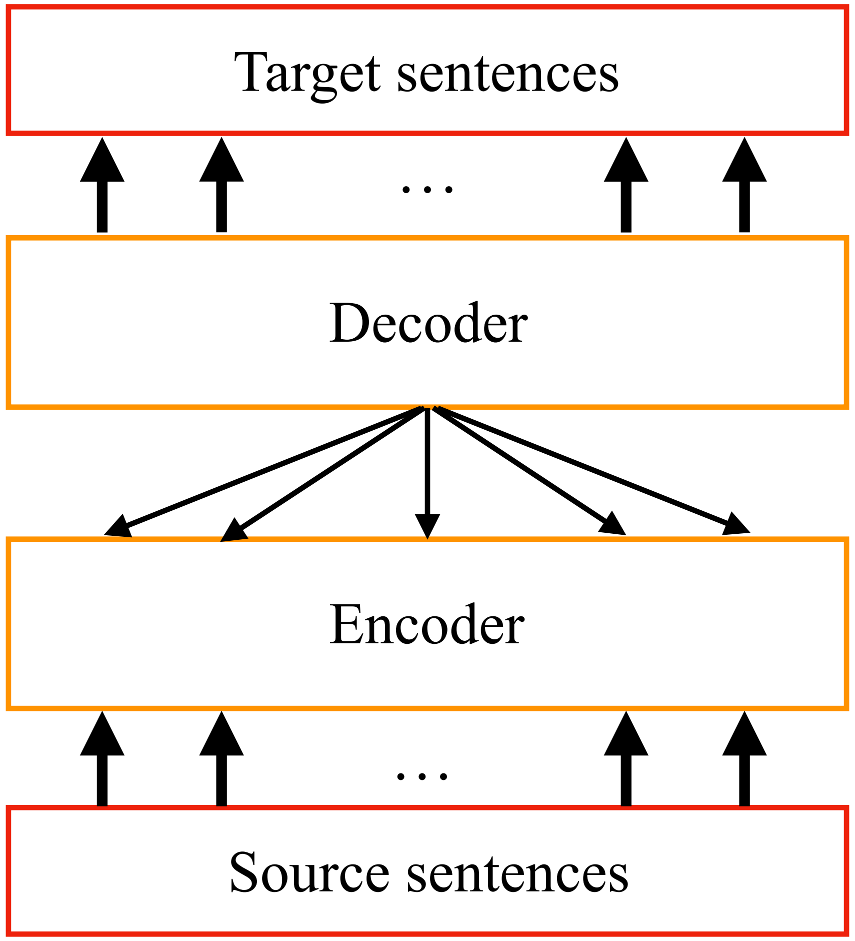
https://arxiv.org/abs/1706.03762

- Introduction
- Encoder - Decoder
- Transformer
- Evaluation

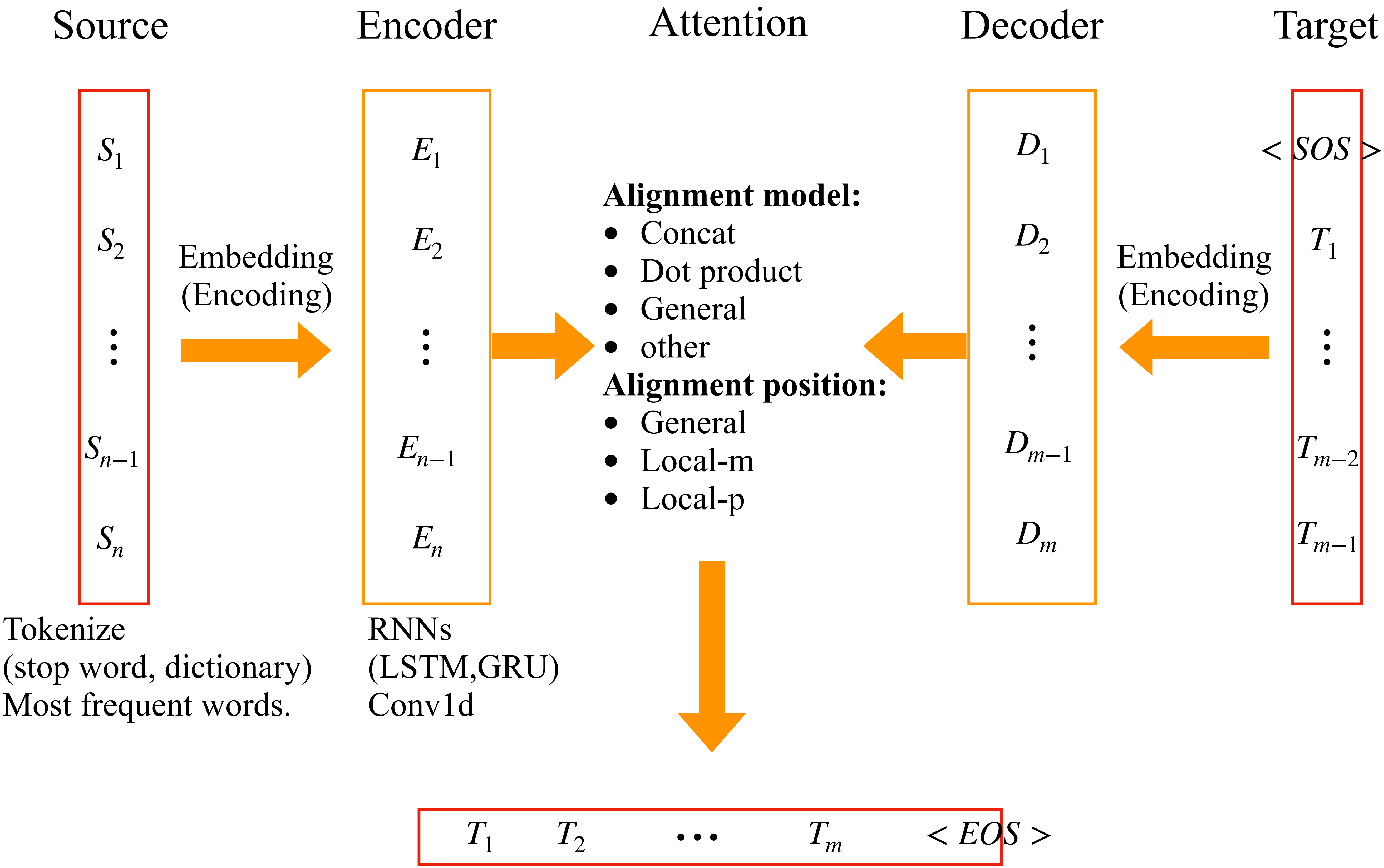
Introduction

Neural Machine Translation(NMT):

- **Statistical based:** Phrase-based + large LM (Moses)
- **NN based:** Encoder - Decoder (Seq2seq, ConvS2S, ensemble …)



Encoder - Decoder



Source

S_1

S_2

$\vdots$

S_{n-1}

S_n

Tokenize
(stop word, dictionary)
Most frequent words.

Embedding
(Encoding)

Encoder

E_1

E_2

$\vdots$

E_{n-1}

E_n

RNNs
(LSTM,GRU)
Conv1d

Attention

Alignment model:

- Concat
- Dot product
- General
- other

Alignment position:

- General
- Local-m
- Local-p

Decoder

D_1

D_2

$\vdots$

D_{m-1}

D_m

Embedding
(Encoding)

Target

$< SOS >$

T_1

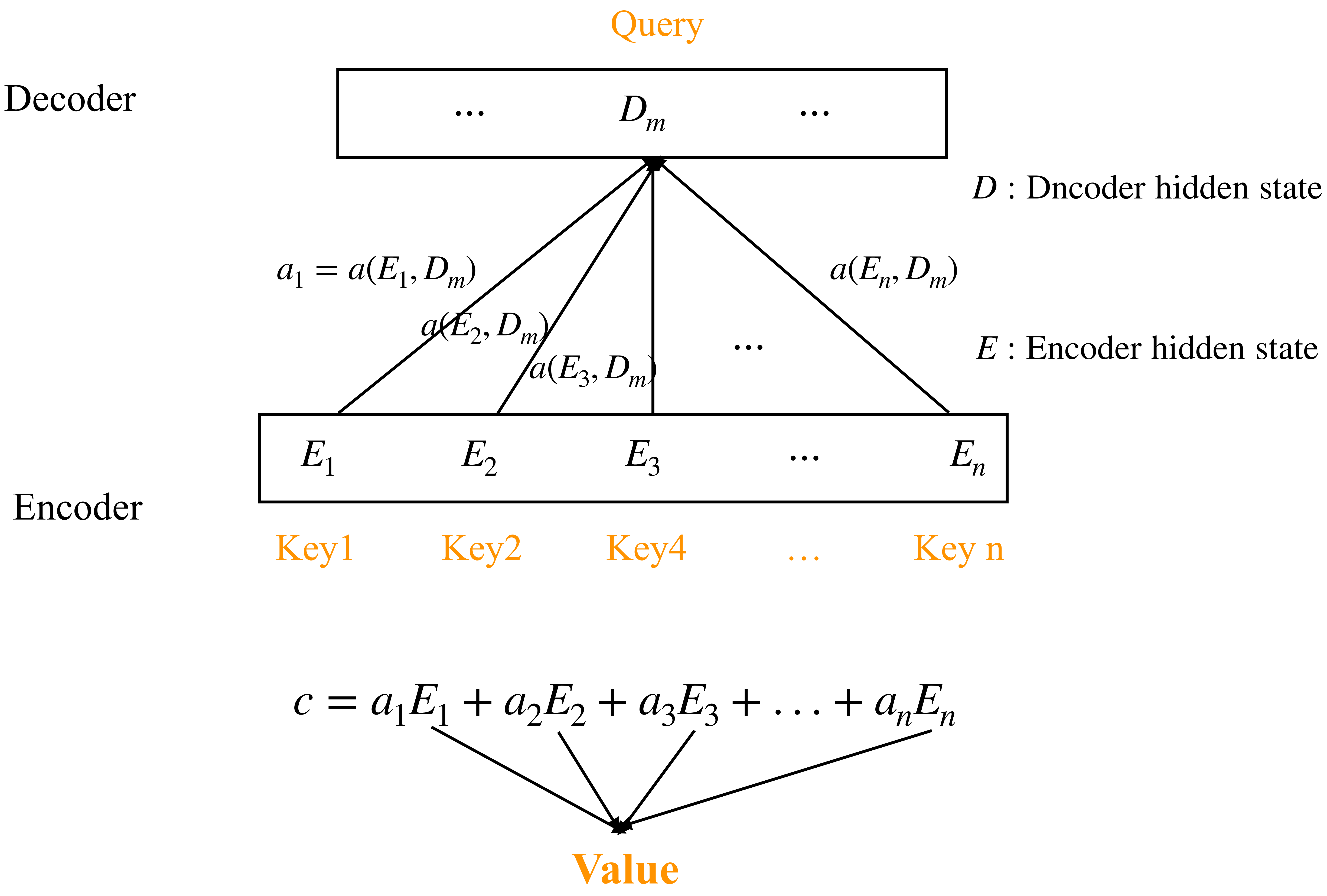
$\vdots$

T_{m-2}

T_{m-1}

$T_1 \quad T_2 \quad \cdots \quad T_m \quad < EOS >$

Encoder - Decoder



Query

Decoder

… D_m …

D : Dncoder hidden state

$a_1 = a(E_1, D_m)$

$a(E_2, D_m)$

$a(E_3, D_m)$

…

$a(E_n, D_m)$

E : Encoder hidden state

E_1 E_2 E_3 … E_n

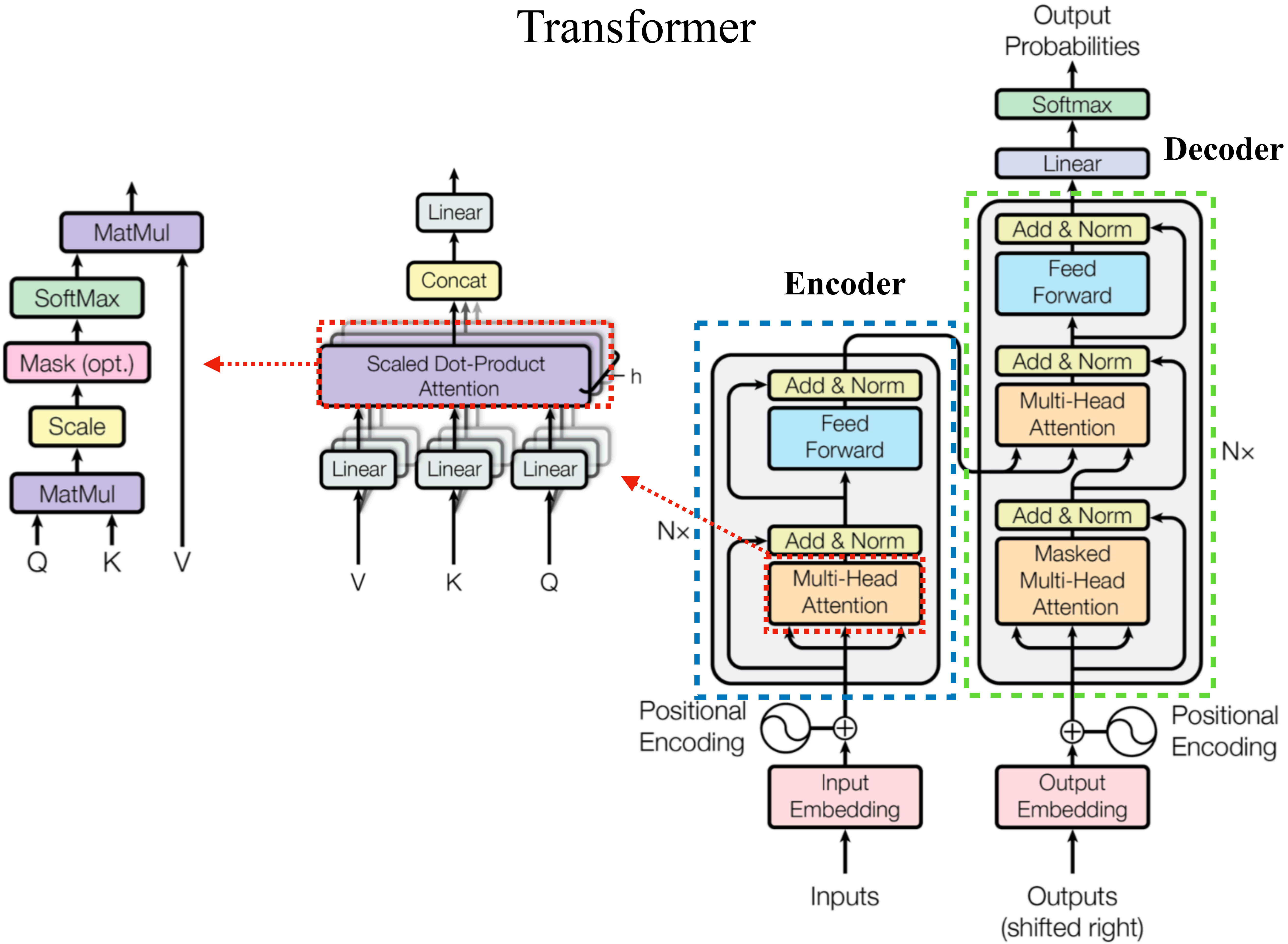
Encoder

Key1 Key2 Key4 … Key n

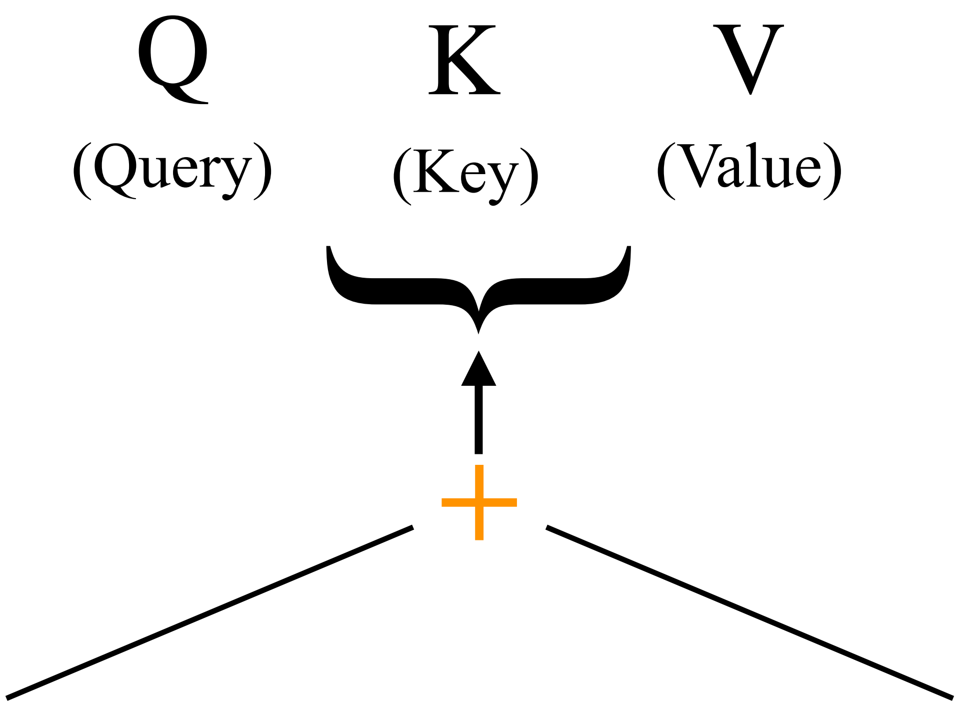
$$c = a_1E_1 + a_2E_2 + a_3E_3 + \ldots + a_nE_n$$

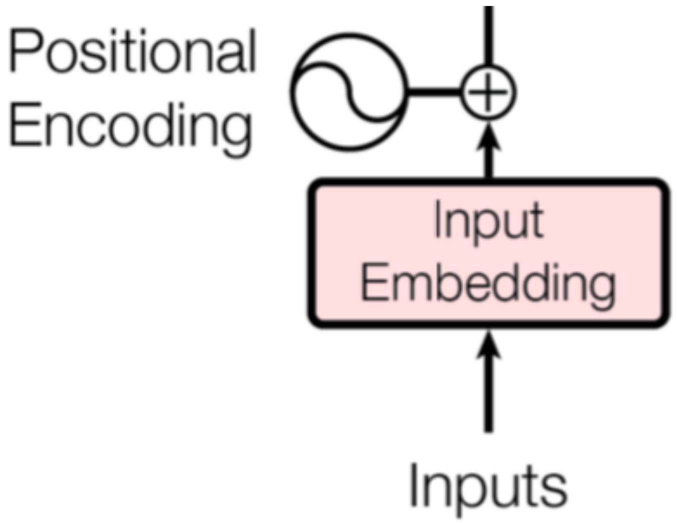
Value

Transformer



Transformer





Positional Encoding

$$PE = \{p_0, p_1, p_2, \ldots, p_{n-1}\}$$

$$\begin{cases} PE_{(pos,2i)} = \sin(pos/10000^{\frac{2i}{d_{model}}}) \\ PE_{(pos,2i+1)} = \cos(pos/10000^{\frac{2i}{d_{model}}}) \end{cases}$$

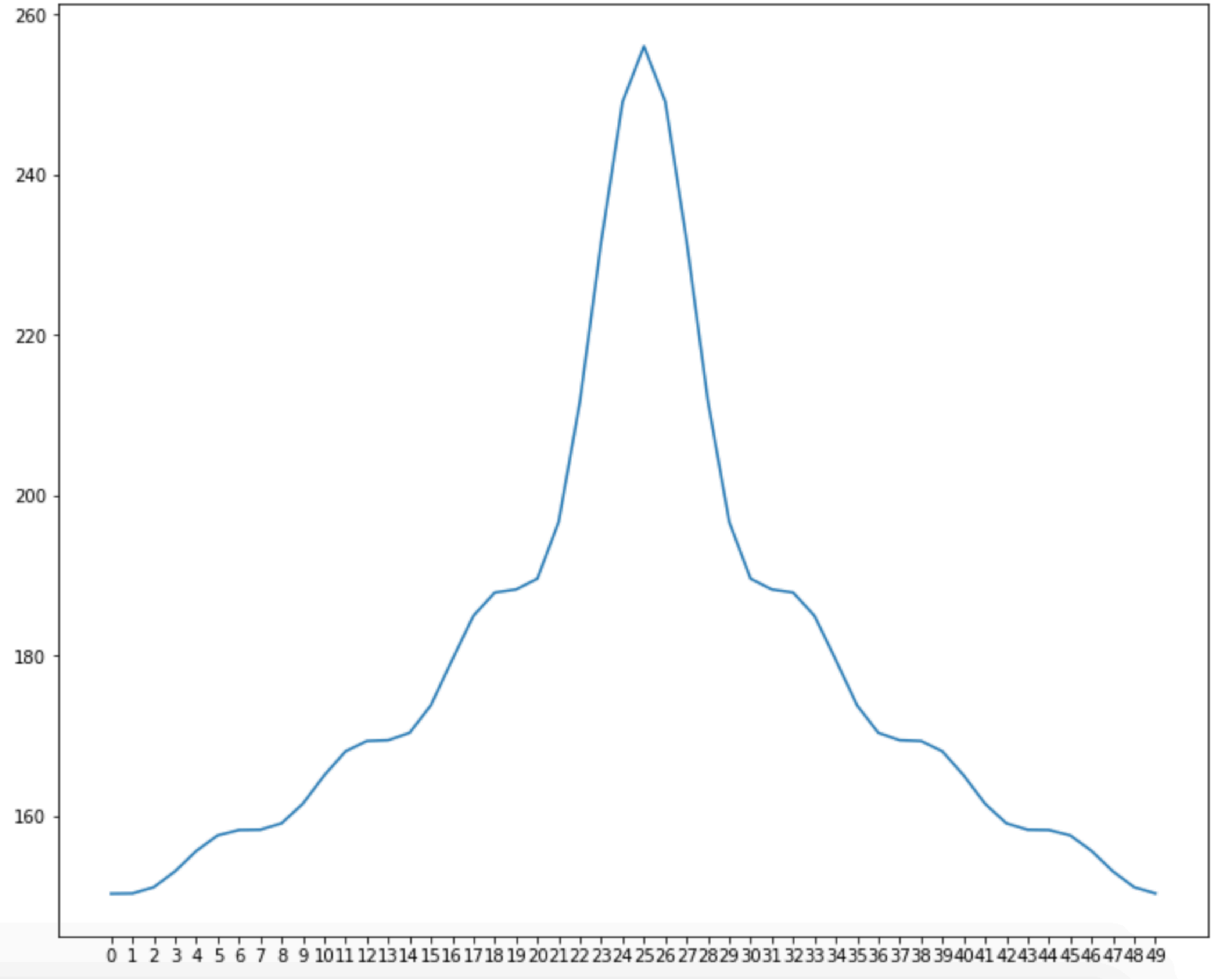
Word Embedding

$$W = \{w_1, w_2, \ldots, w_n\},\ w_i \in d_{model}$$

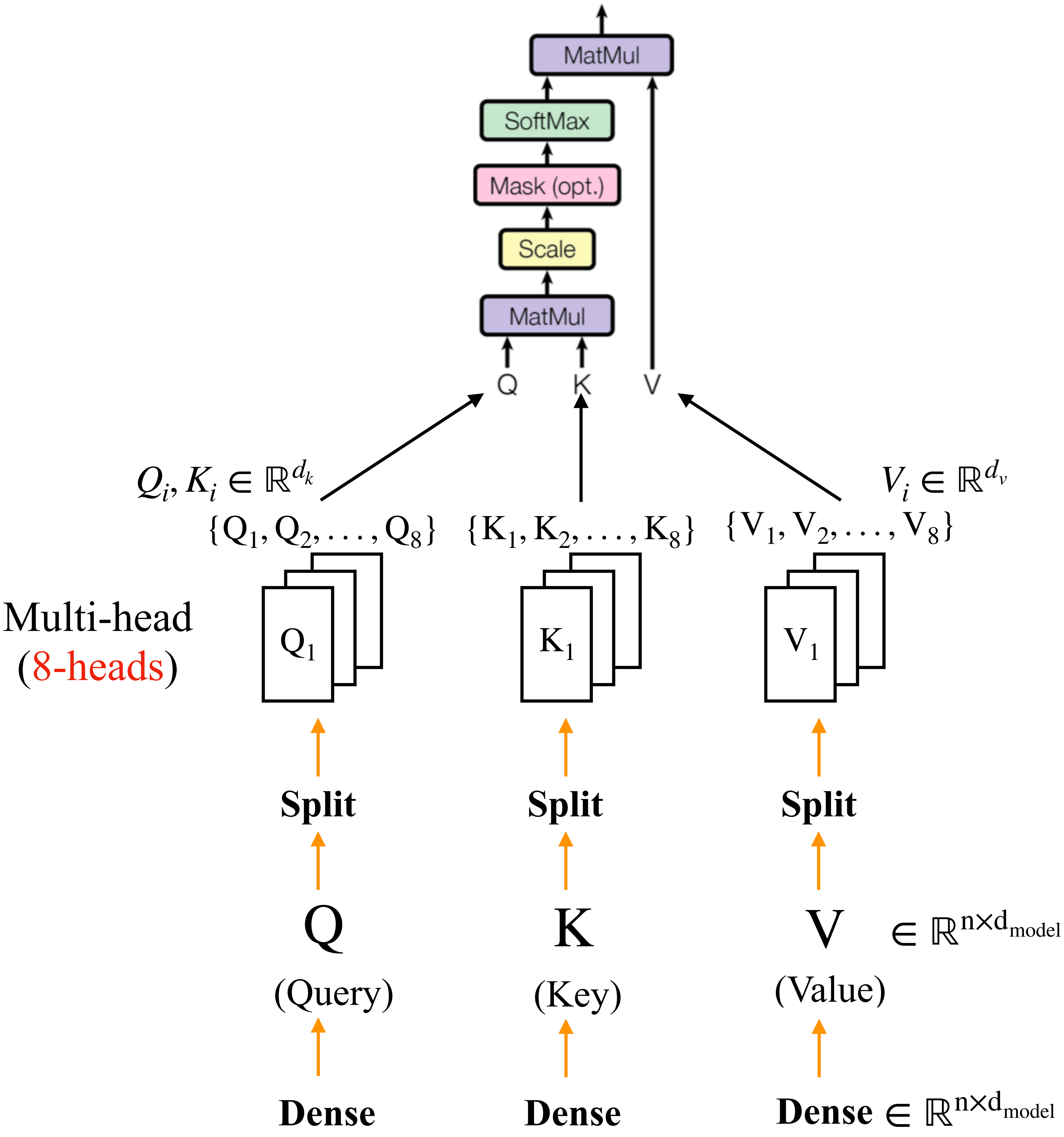
Positional Encoding

Positional Encoding: $\{p_0, p_1, ..., p_{49}\}$, $p_i \in d_{model}$

1. Calculates the inner product p_{24} to the others p_0 *to* p_{49}.
2. Get product $V_0, V_1, ..., V_{49}$.



V



$Q_i, K_i \in \mathbb{R}^{d_k}$

$V_i \in \mathbb{R}^{d_v}$

$\{Q_1, Q_2, \ldots, Q_8\}$ $\{K_1, K_2, \ldots, K_8\}$ $\{V_1, V_2, \ldots, V_8\}$

Multi-head (8-heads)

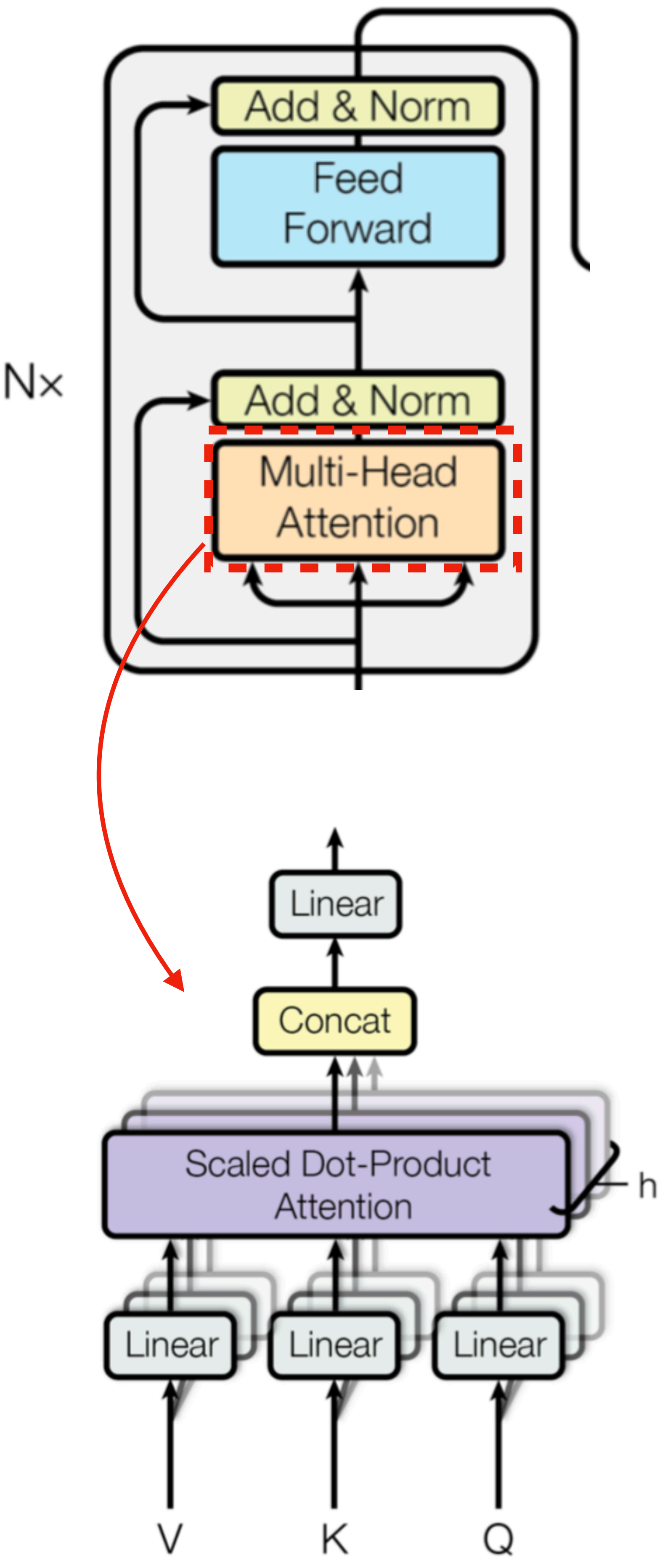
Q_1 K_1 V_1

Split **Split** **Split**

Q K V $\in \mathbb{R}^{n \times d_{model}}$

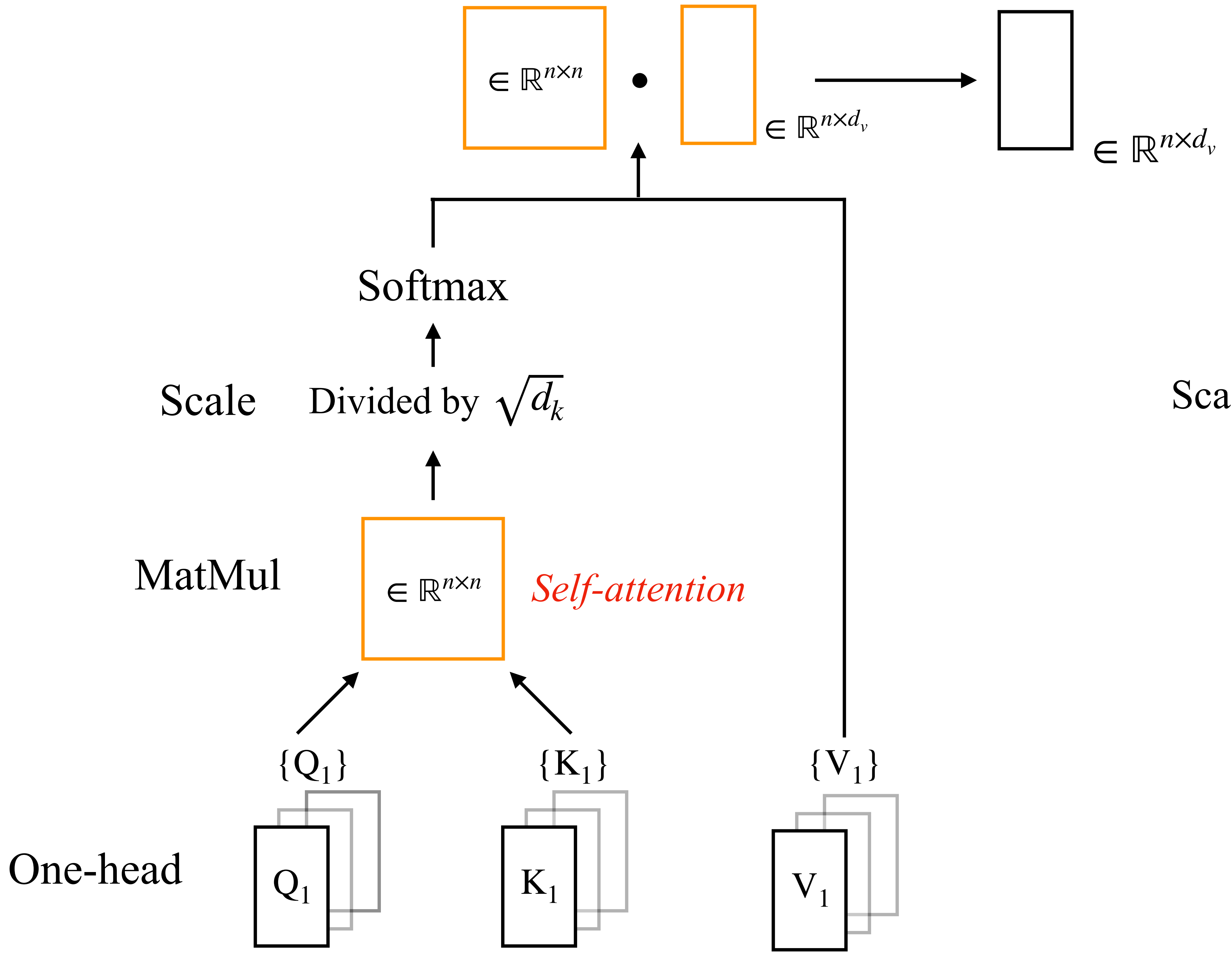
(Query) (Key) (Value)

Dense **Dense** **Dense** $\in \mathbb{R}^{n \times d_{model}}$



Transformer

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



$\in \mathbb{R}^{n\times n}$ • $\in \mathbb{R}^{n\times d_v}$ → $\in \mathbb{R}^{n\times d_v}$

Softmax

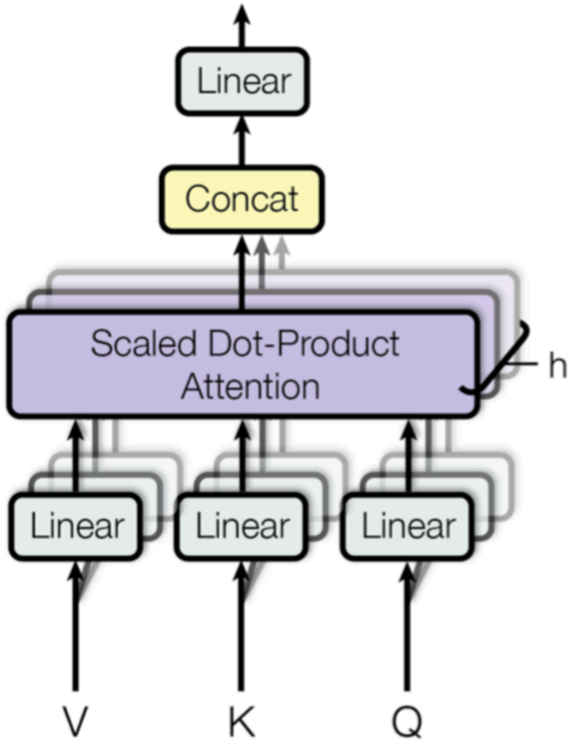
Scale Divided by $\sqrt{d_k}$

MatMul $\in \mathbb{R}^{n\times n}$ *Self-attention*

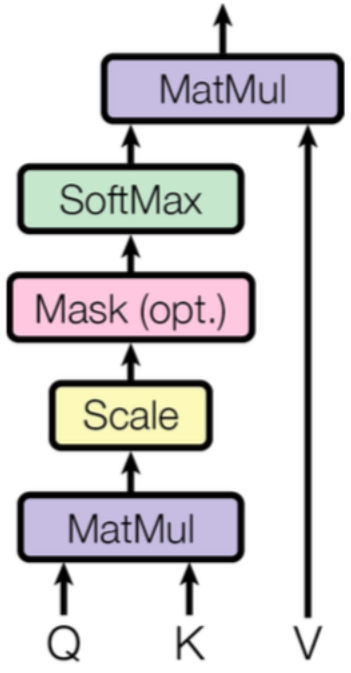
$\{Q_1\}$ $\{K_1\}$ $\{V_1\}$

One-head Q_1 K_1 V_1

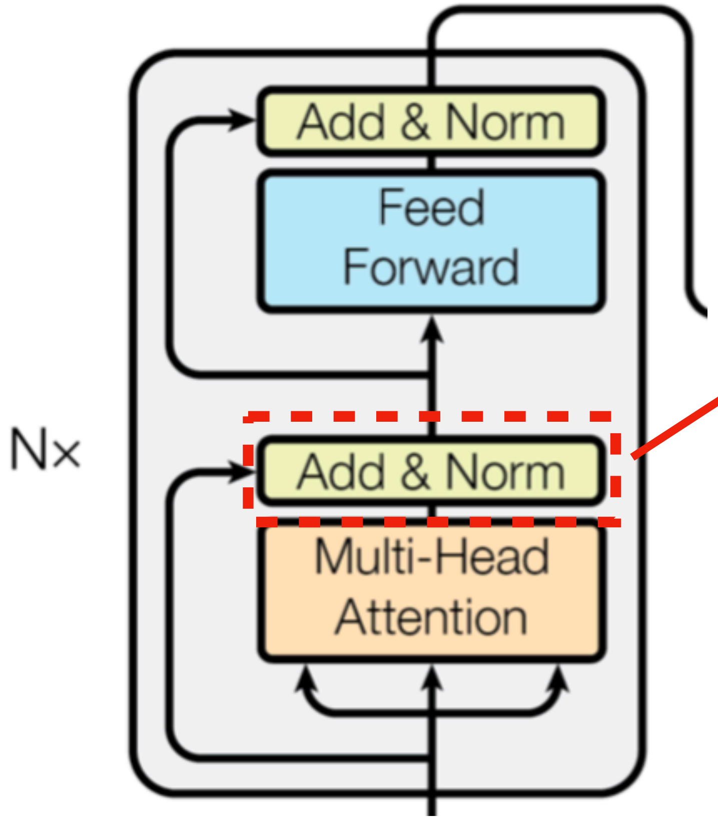
Multi-heads attention



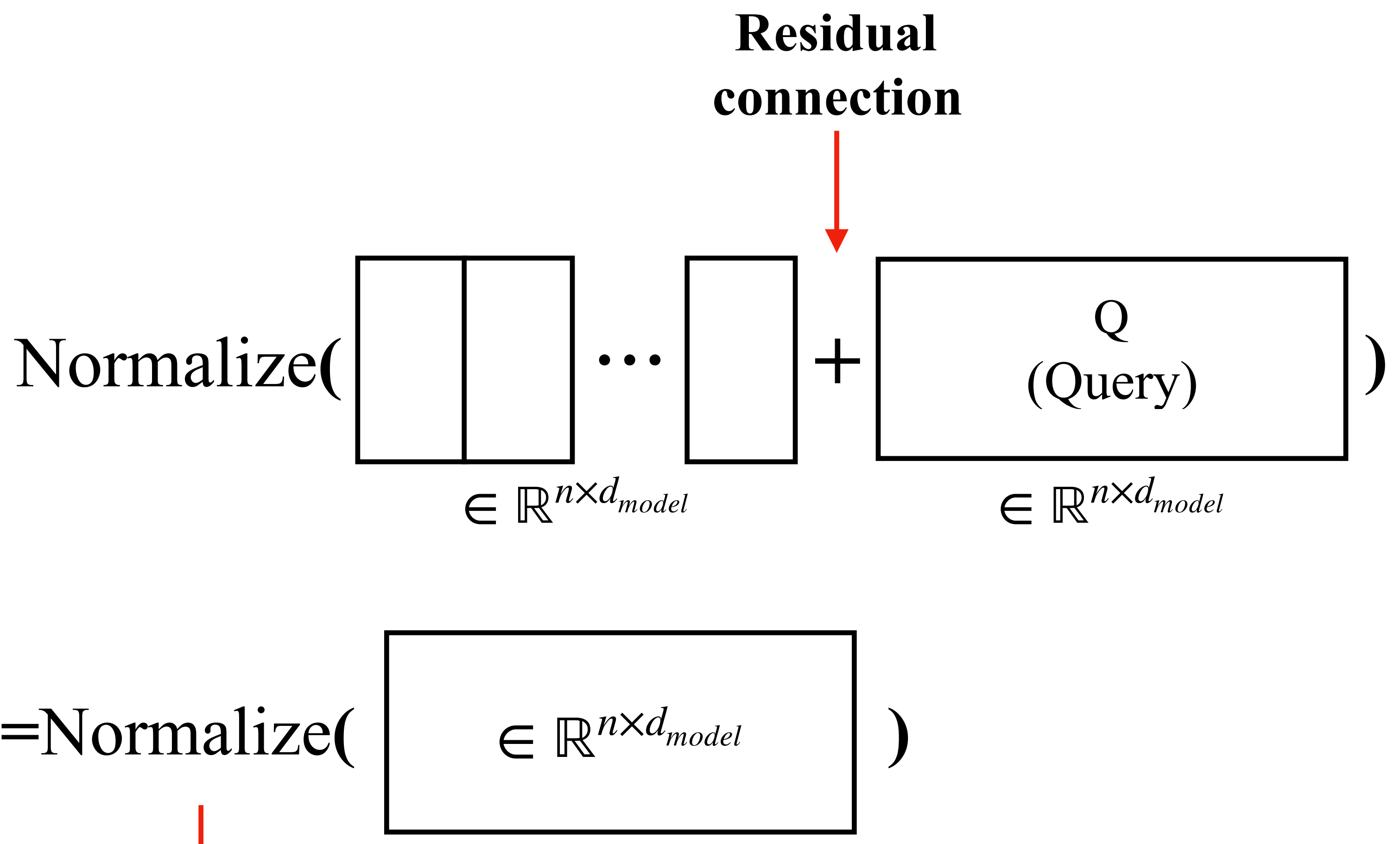
Scaled Dot-Product Attention (Self-Attention)



Transformer



Encoder



Residual connection

Normalize($\cdots$ + Q (Query))

$\in \mathbb{R}^{n \times d_{model}}$ $\in \mathbb{R}^{n \times d_{model}}$

=Normalize($\in \mathbb{R}^{n \times d_{model}}$)

Layer Normalize: $LN(z; \alpha, \beta) = \dfrac{(z - \mu)}{\sigma} \odot \alpha + \beta$

Mean : $\mu^l = \dfrac{1}{D} \sum_{i=1}^{D} z_i^l$

Gains : α

Biases : β

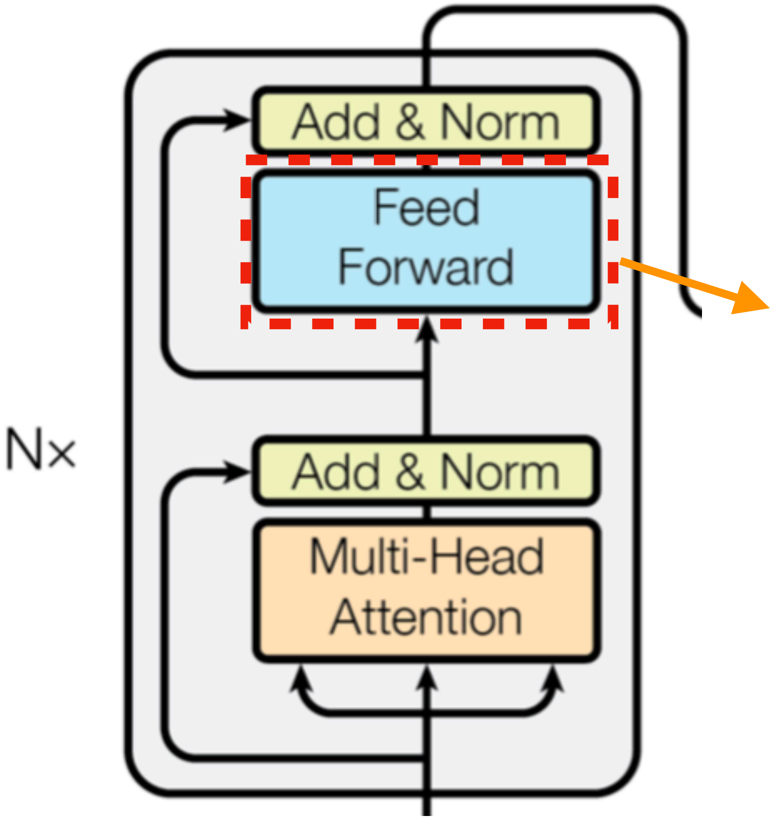
Standard Deviation : $\sigma^l = \sqrt{\dfrac{1}{D} \sum_{i=1}^{D} (z_i^l - \mu^l)^2}$

Transformer

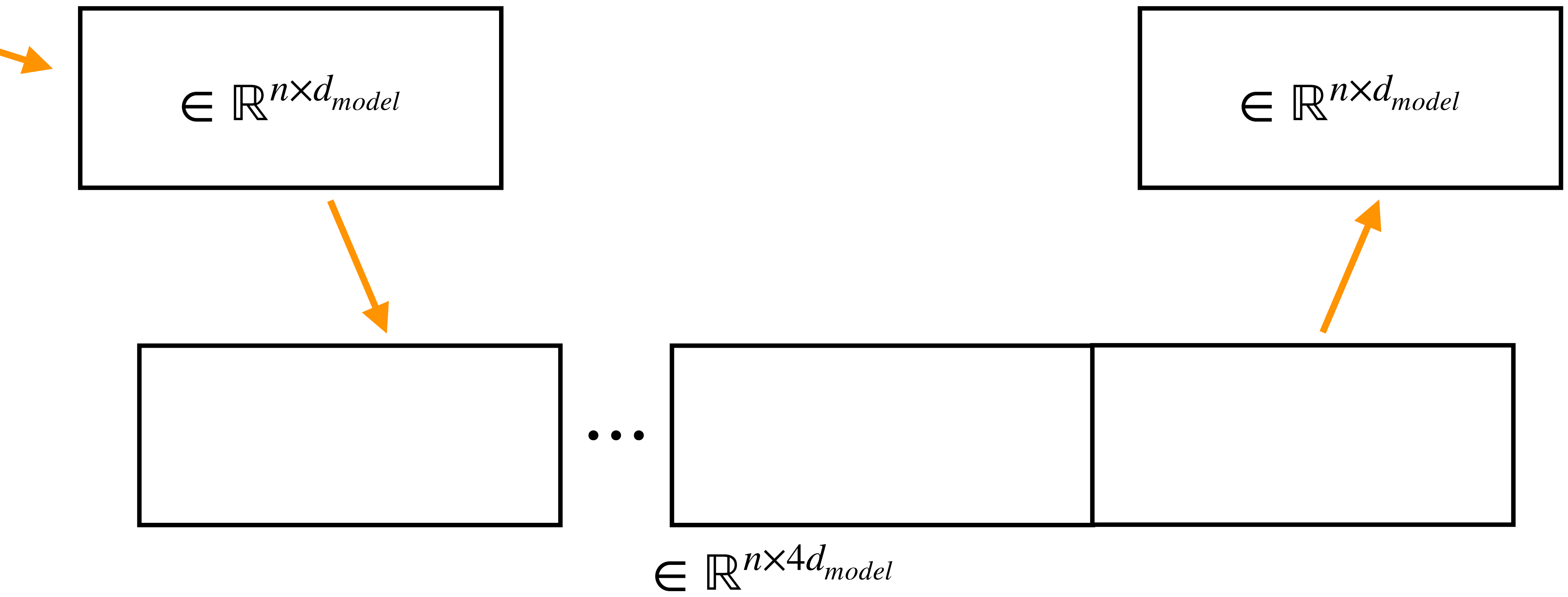
Feed Forward
(Dense、Conv1d)

$$\mathrm{FFN}(x) = max(0, xW_1 + b_1)W_2 + b_2$$

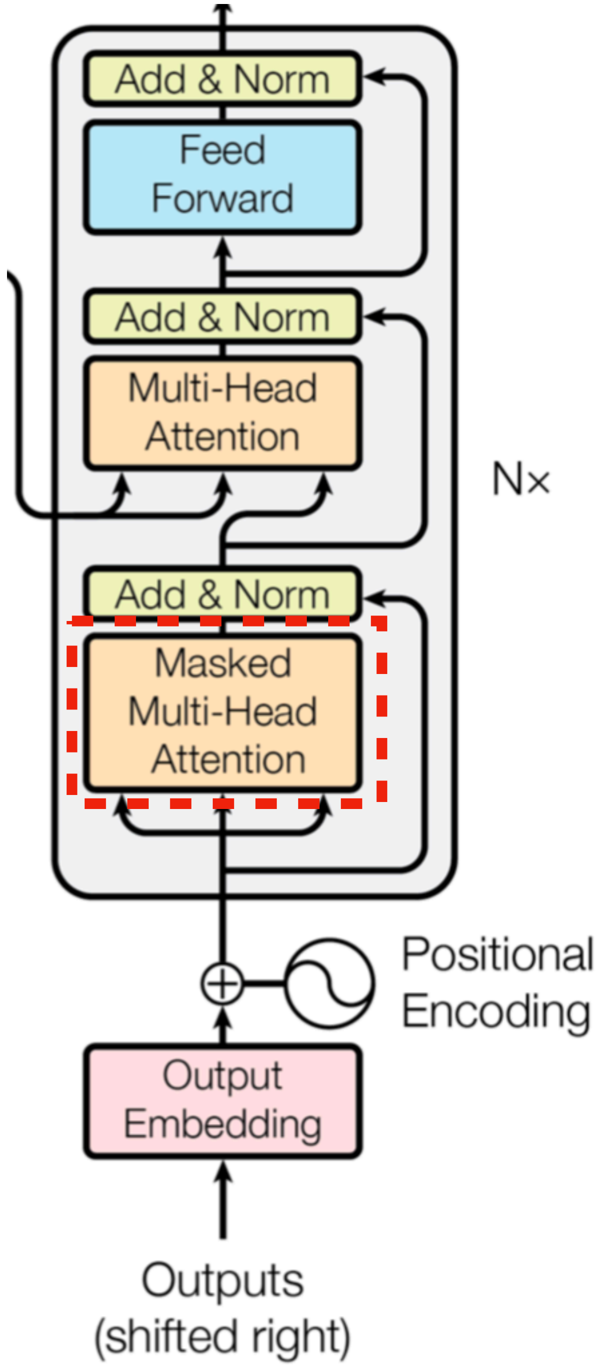
$$\mathrm{Dense}(d_{model}) \rightarrow \mathrm{Dense}\ (4 \times d_{model}) \rightarrow \mathrm{Dense}(d_{model})$$



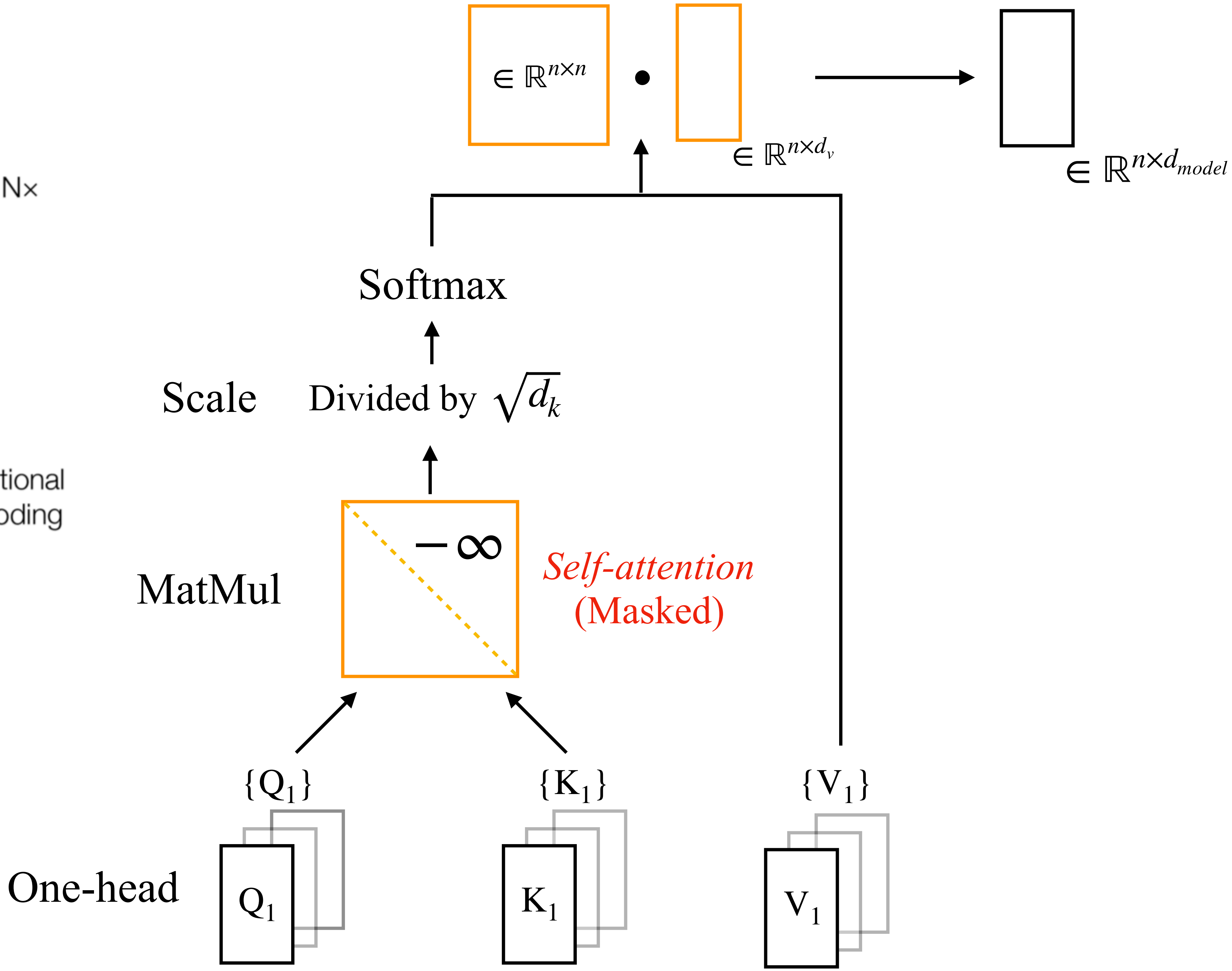
Encoder


$\in \mathbb{R}^{n \times d_{model}}$
$\in \mathbb{R}^{n \times d_{model}}$
…
$\in \mathbb{R}^{n \times 4d_{model}}$

Transformer



Decoder



$\in \mathbb{R}^{n \times n}$ • $\in \mathbb{R}^{n \times d_v}$ → $\in \mathbb{R}^{n \times d_{model}}$

Softmax

Scale Divided by $\sqrt{d_k}$

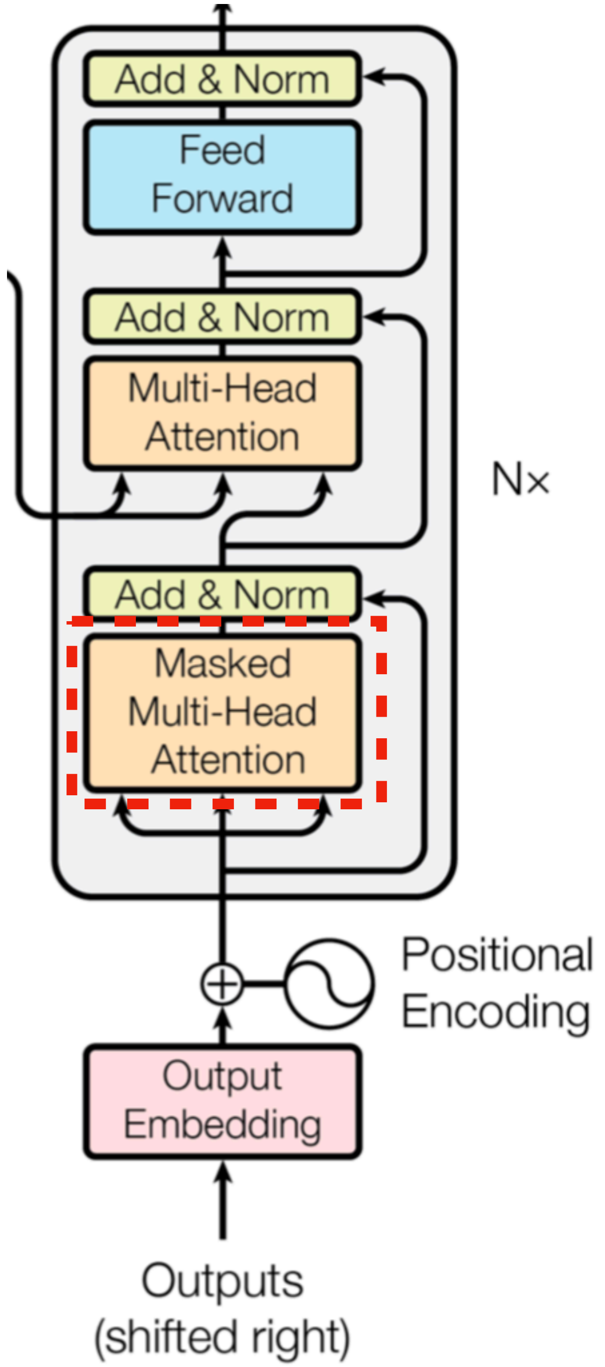
MatMul $-\infty$

Self-attention
(Masked)

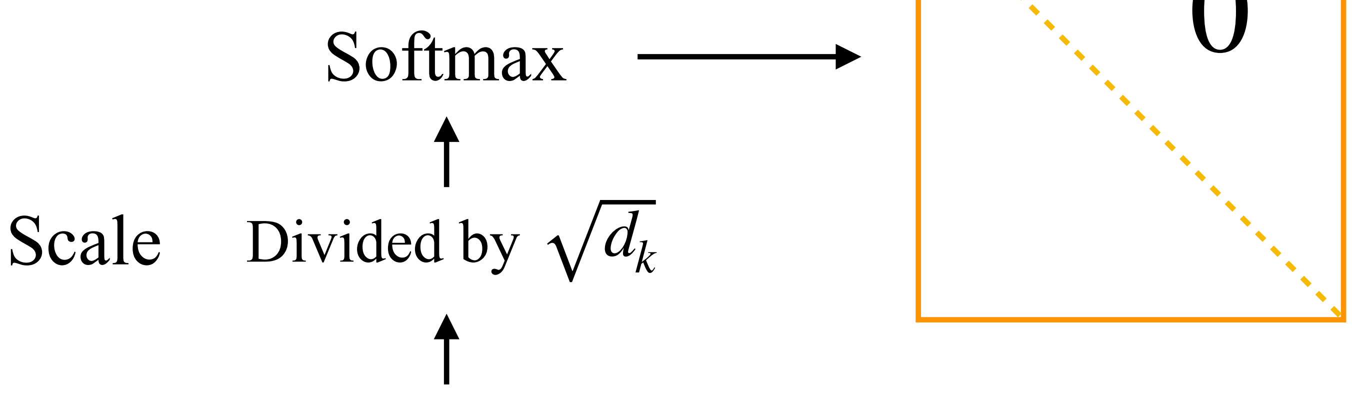
$\{Q_1\}$ $\{K_1\}$ $\{V_1\}$

One-head Q_1 K_1 V_1

Transformer



Decoder



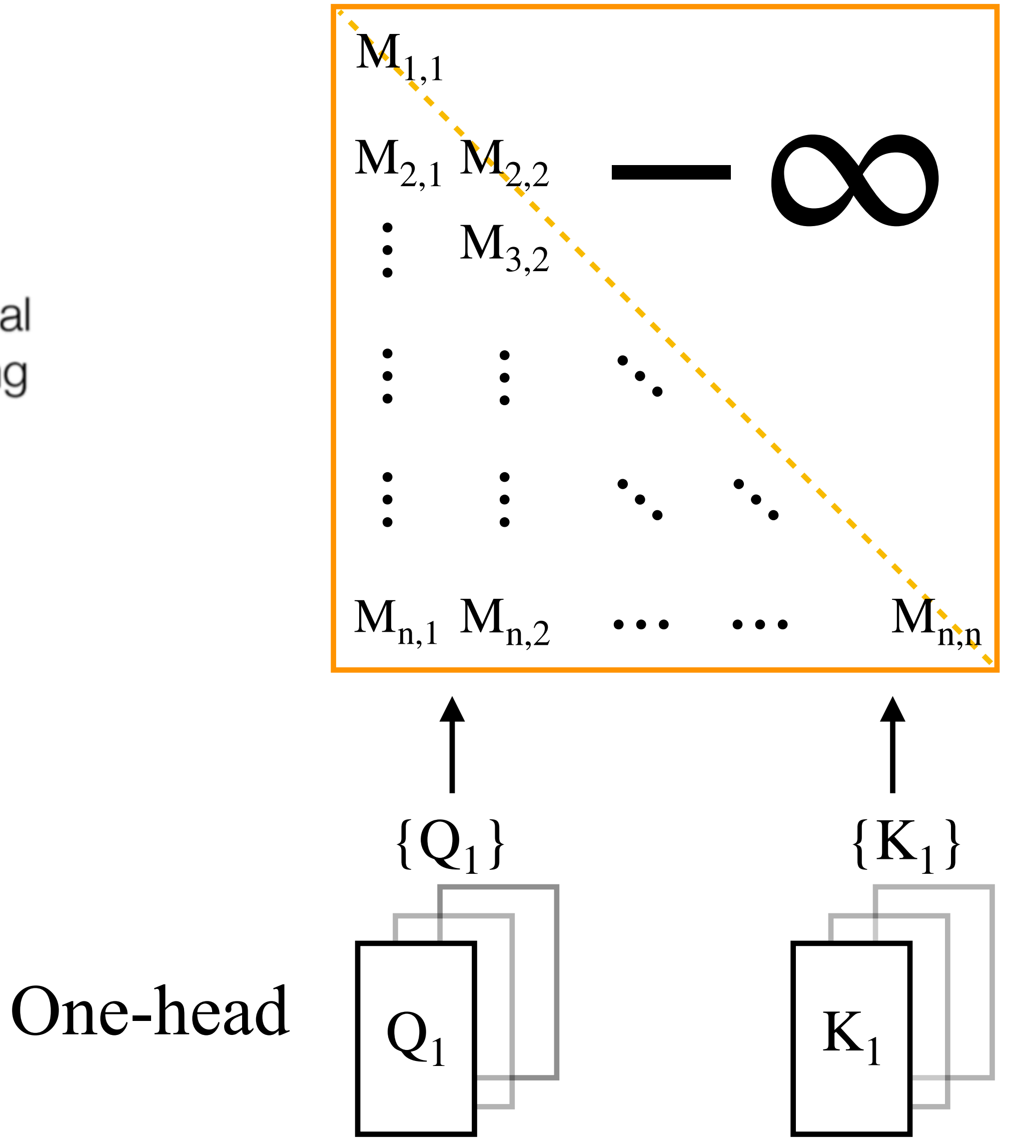
Softmax

Scale Divided by $\sqrt{d_k}$

0

We need to prevent leftward information flow in the decoder

$M_{1,1}$: The similarity between Q_1 and K_1



$M_{1,1}$

$M_{2,1}$ $M_{2,2}$

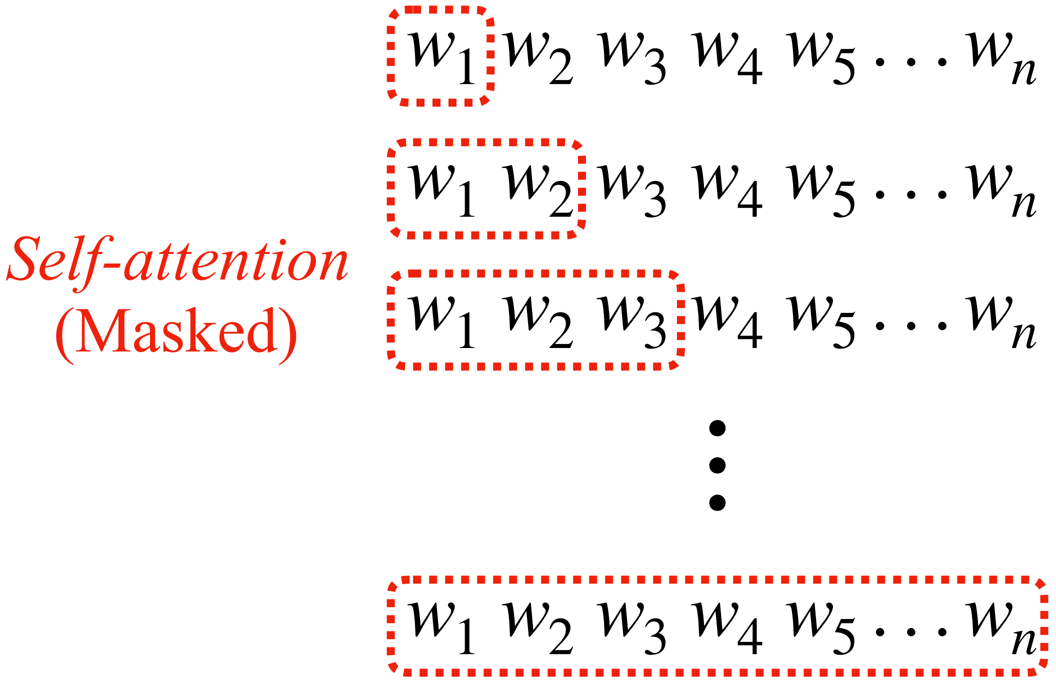
$M_{3,2}$

$M_{n,1}$ $M_{n,2}$ … … $M_{n,n}$

$-\infty$

$\{Q_1\}$ $\{K_1\}$

One-head Q_1 K_1



Self-attention
(Masked)

$w_1\ w_2\ w_3\ w_4\ w_5 \ldots w_n$

$w_1\ w_2\ w_3\ w_4\ w_5 \ldots w_n$

$w_1\ w_2\ w_3\ w_4\ w_5 \ldots w_n$

⋮

$w_1\ w_2\ w_3\ w_4\ w_5 \ldots w_n$

Evaluation

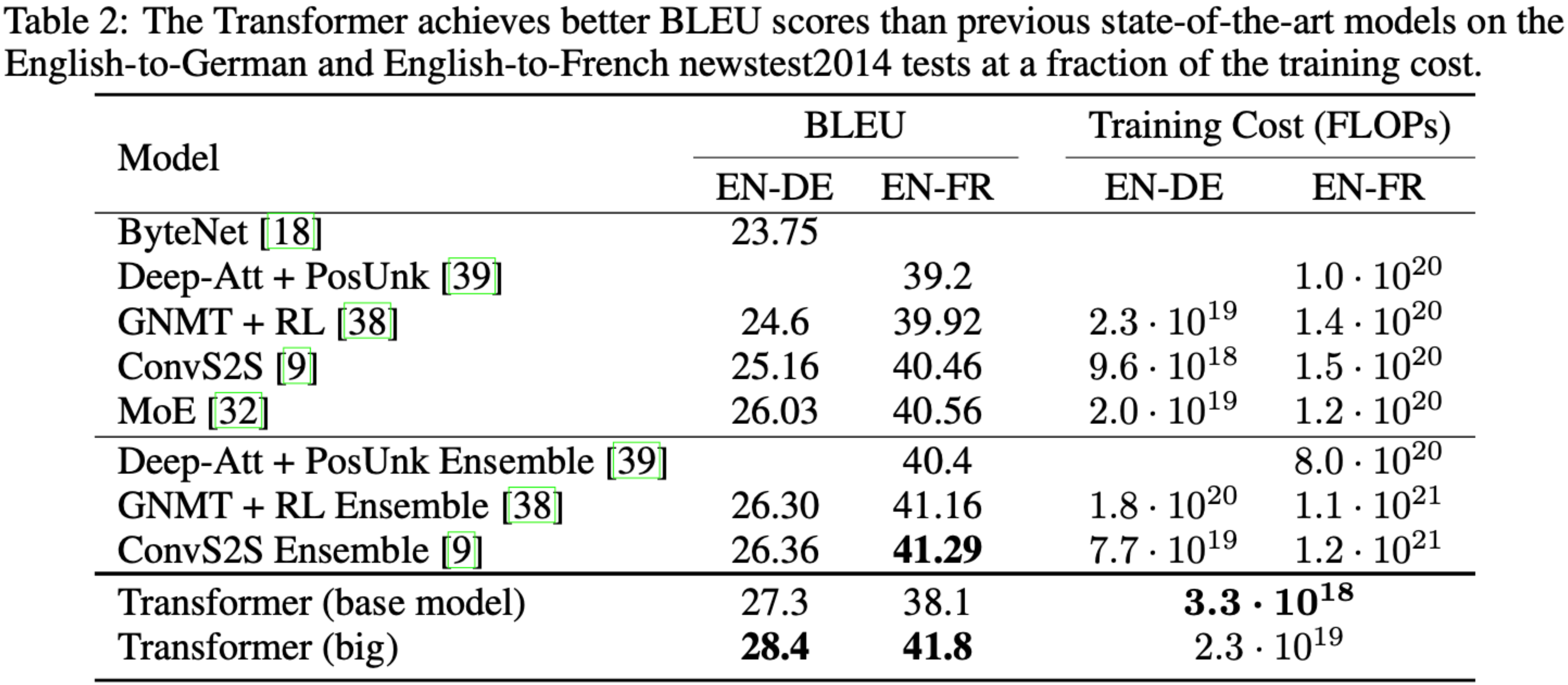
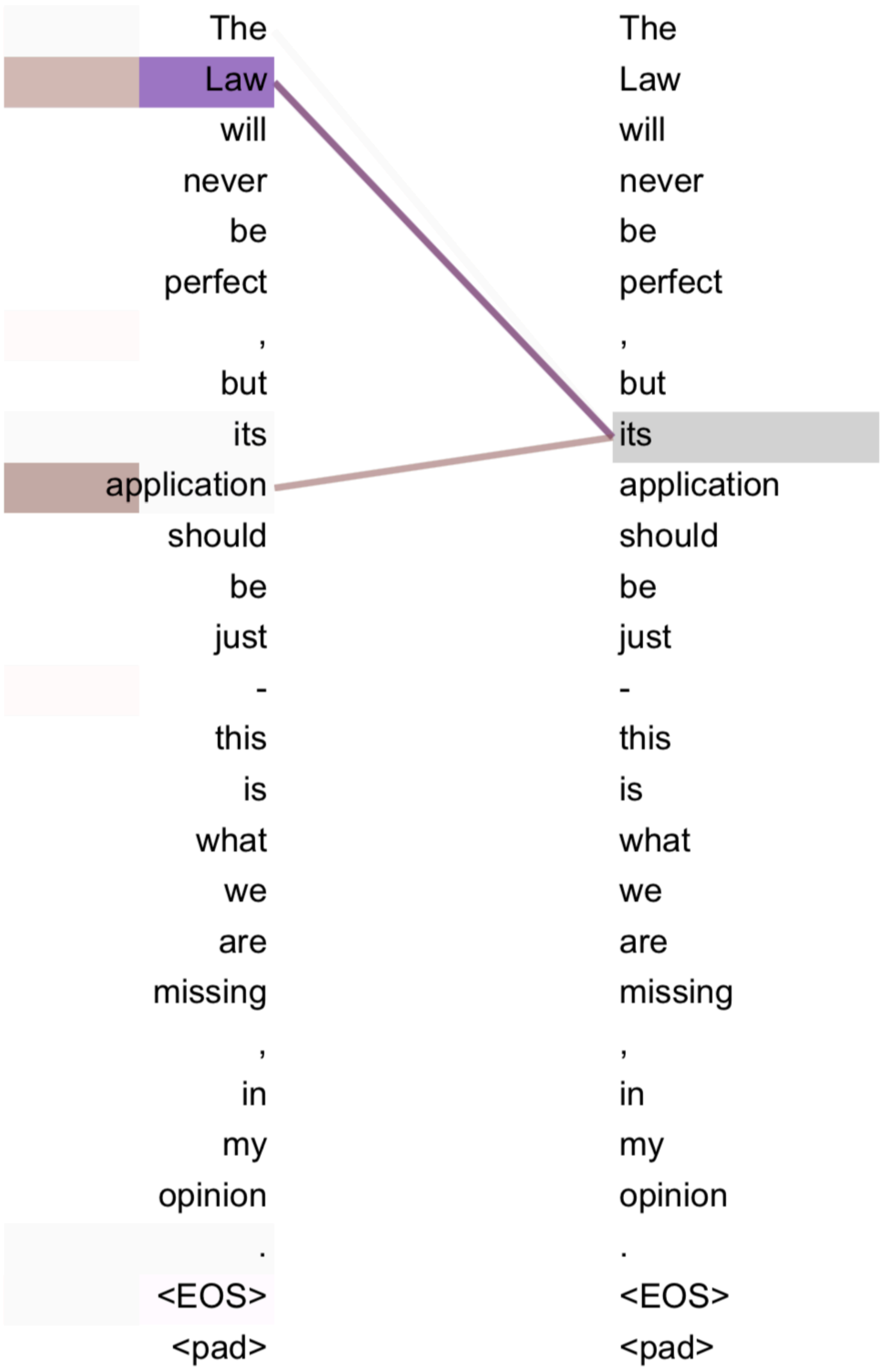


Table 2: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	$\mathbf{3.3 \cdot 10^{18}}$	
Transformer (big)	**28.4**	**41.8**	$2.3 \cdot 10^{19}$	

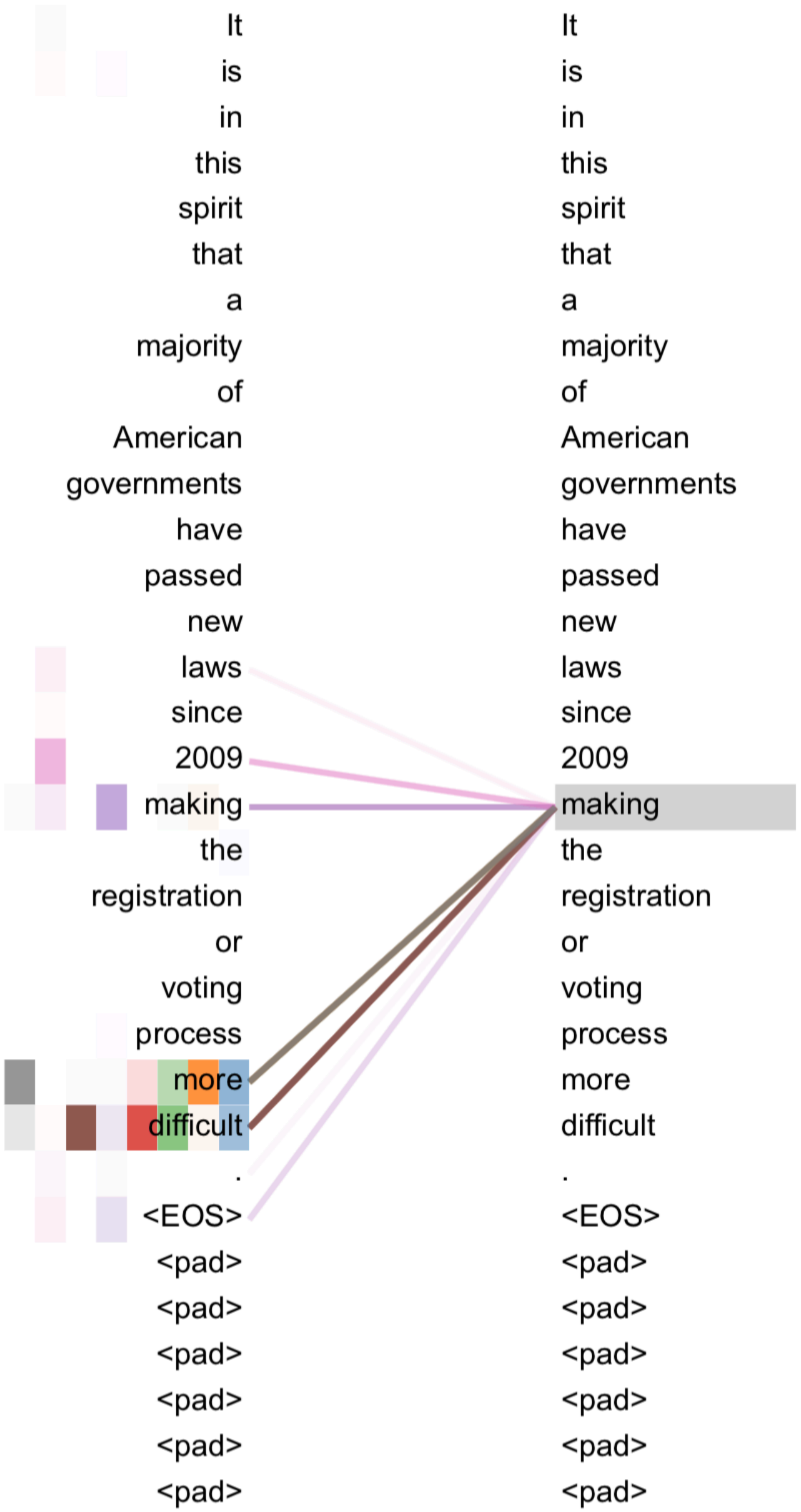
- **ByteNet:** 2 convolution layers
- **Deep-Att + PosUnk:** 2 Bi-LSTM layers(Encoder) + 1 LSTM layer (Decoder)
- **GNMT + RL:** 7 LSTM layers + 1 Bi-LSTM layer(Encoder) + 8 LSTM layers(Decoder)
- **Transformer(base):** training 100,000 steps(12 hours), 0.4 seconds per steps
- **Transformer(big):** training 300,000 steps(3.5 days), 1.0 seconds per steps

Evaluation



The Law will never be perfect , but its application should be just - this is what we are missing , in my opinion . <EOS> <pad>

The Law will never be perfect , but its application should be just - this is what we are missing , in my opinion . <EOS> <pad>

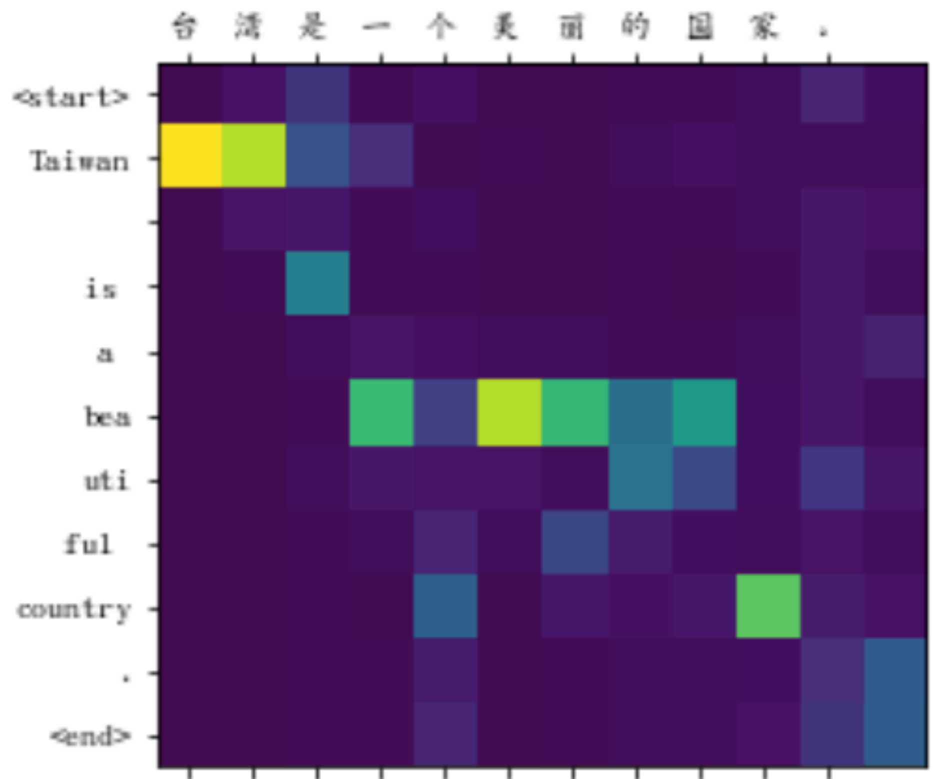


It is in this spirit that a majority of American governments have passed new laws since 2009 making the registration or voting process more difficult . <EOS> <pad> <pad> <pad> <pad> <pad> <pad>

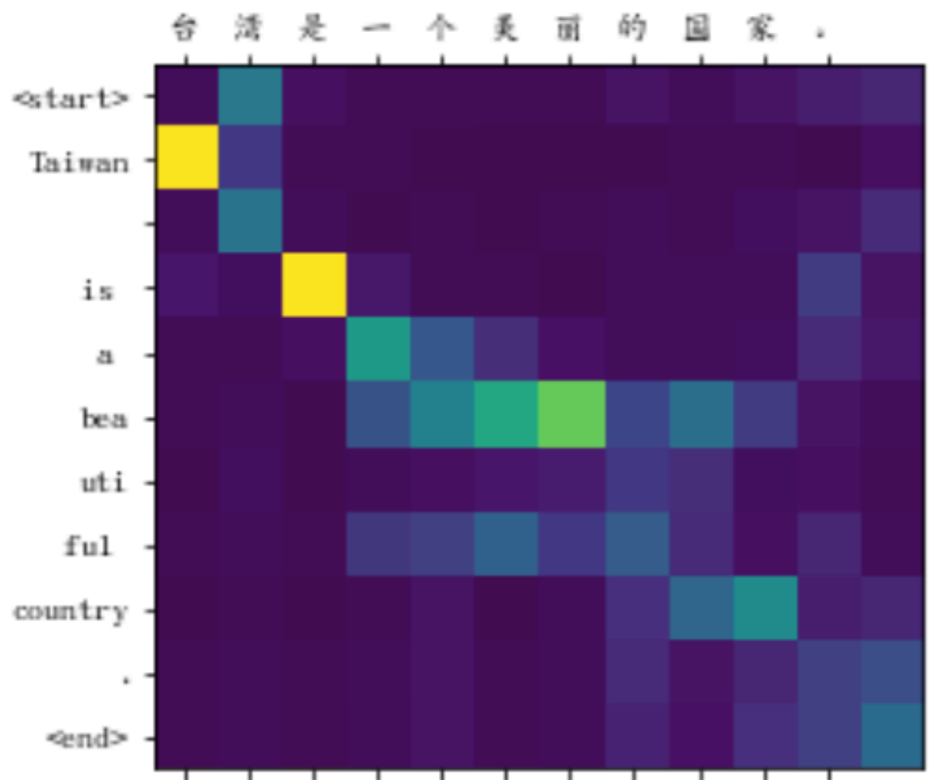
It is in this spirit that a majority of American governments have passed new laws since 2009 making the registration or voting process more difficult . <EOS> <pad> <pad> <pad> <pad> <pad> <pad>

Evaluation

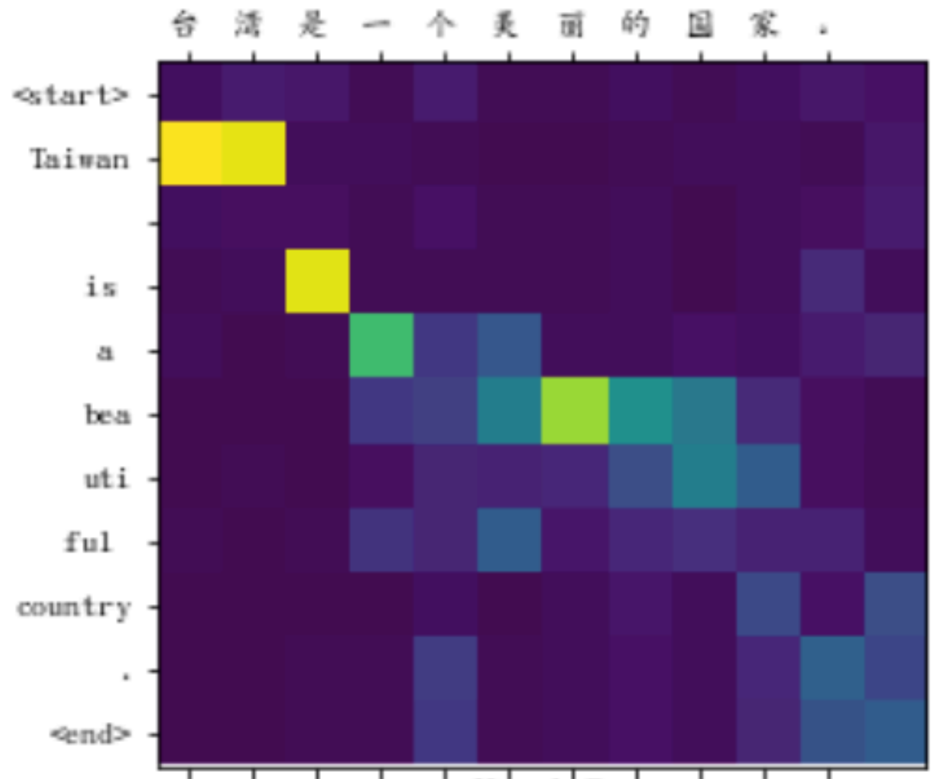
```
Source sentence:
 Taiwan is a beautiful country.

Predict sentence:
 台湾是一个美丽的国家。
```



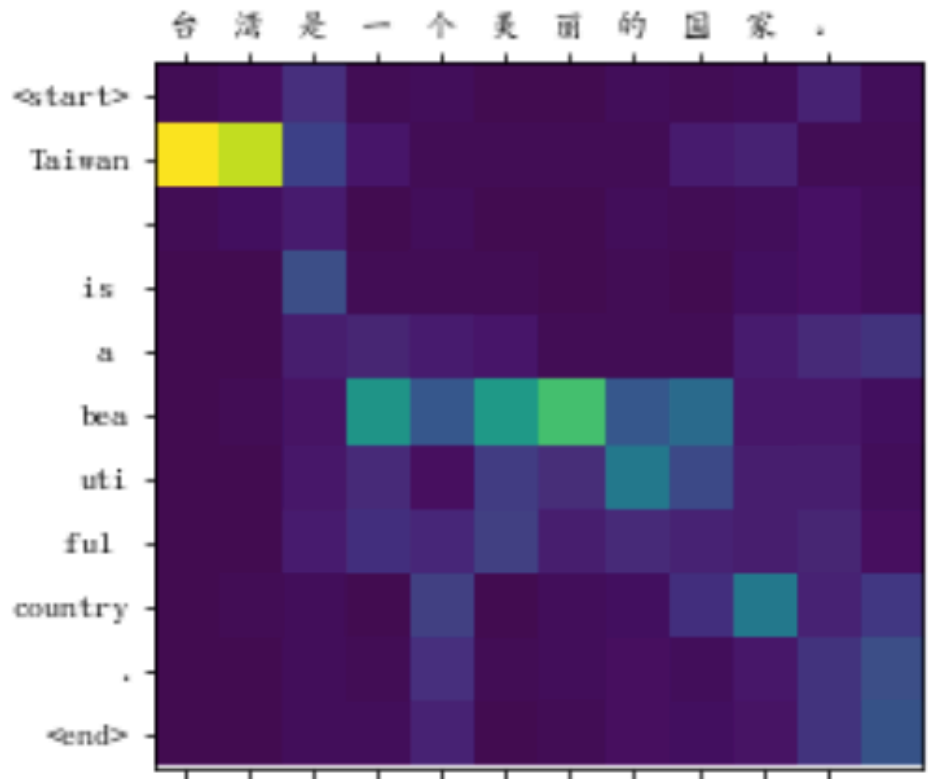
Head 3



Head 4



Head 5



Head 6