Transformer

Attention is all you need. (2017.)
(Masked) Self-Attention

~~**ELMo (2018.)**~~
~~Bi-LSTM~~

GPT(2018.)
Self-Attention (AR)

BERT(2018.)
Self-Attention (Masked token)

XLNet(2019.)
PLM (AR+AE)

Attention is all you need

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

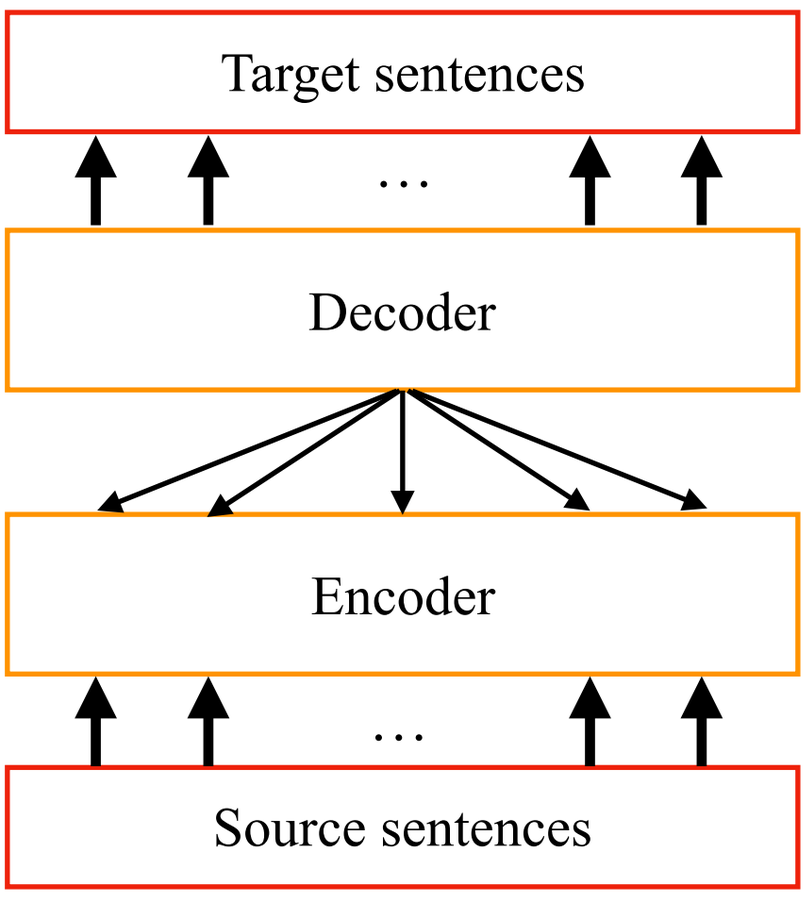
Publish: *Advances in Neural Information Processing Systems*
Pp: 5998 - 6008.

https://arxiv.org/abs/1706.03762

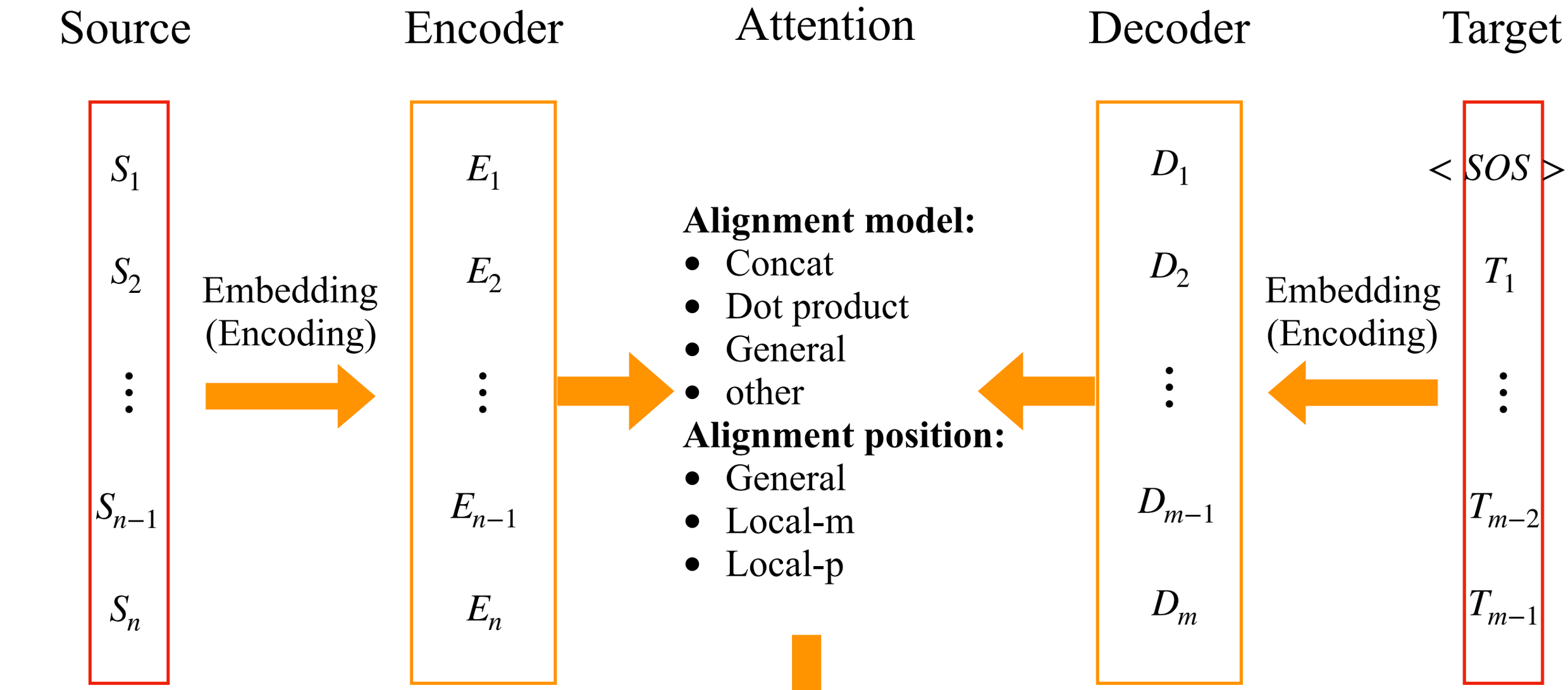
Introduction

Neural Machine Translation(NMT):

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



Related work



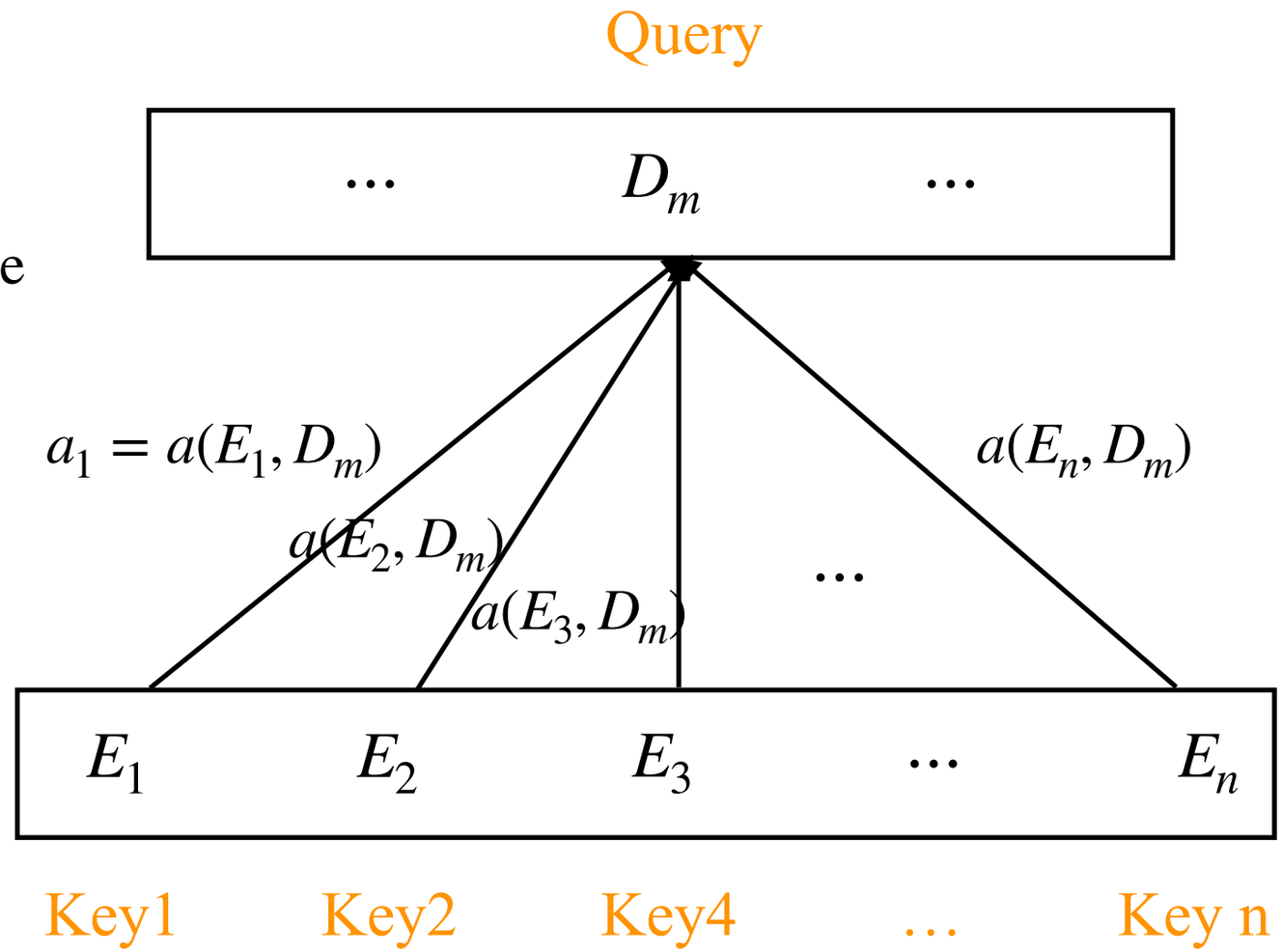
Related work

Decoder

D : Dncoder hidden state

Encoder

E : Encoder hidden state



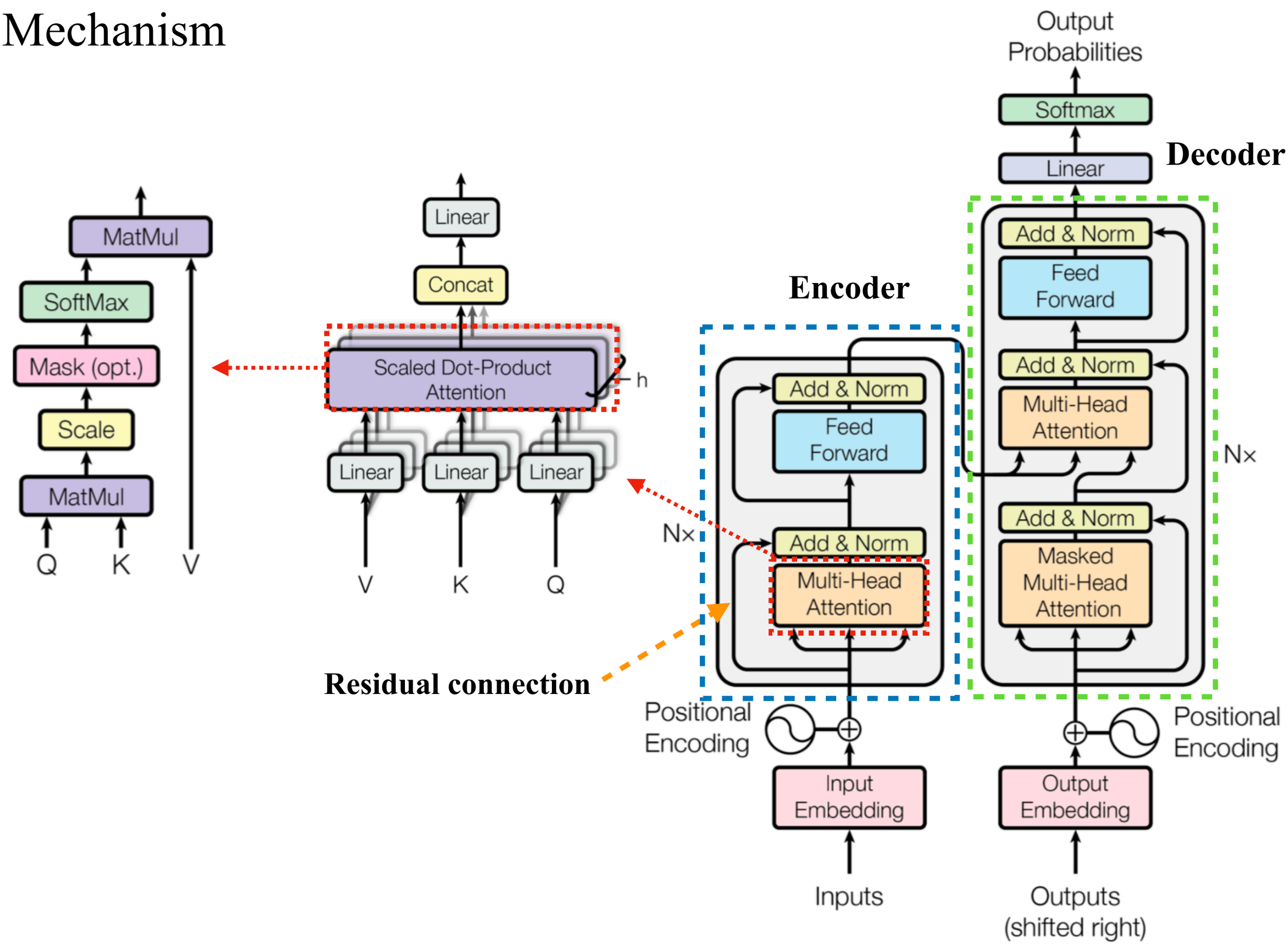
Context vector $c = a_1E_1 + a_2E_2 + a_3E_3 + \ldots + a_nE_n$

Value

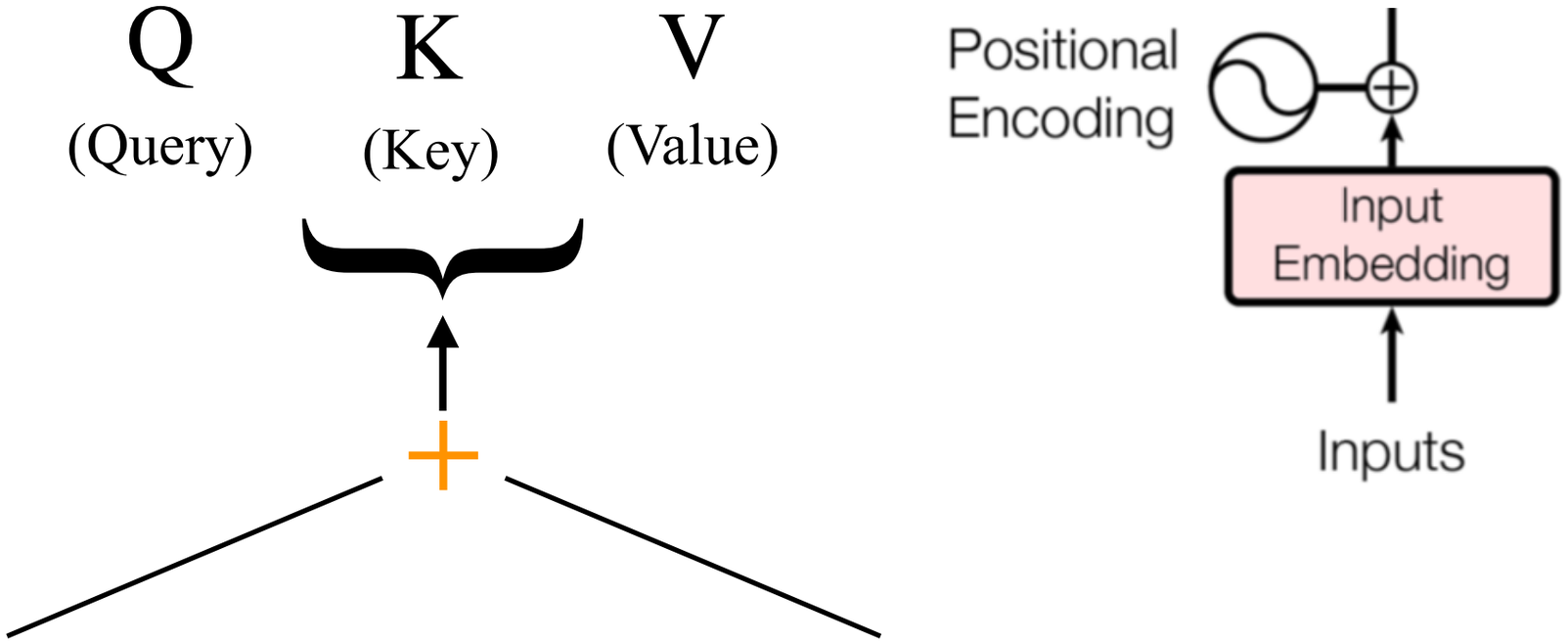
Transformer

Inner Product

Mechanism


MatMul
SoftMax
Mask (opt.)
Scale
MatMul
Q
K
V
Linear
Concat
Scaled Dot-Product Attention
h
Linear
Linear
Linear
V
K
Q
Output Probabilities
Softmax
Linear
Decoder
Encoder
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Nx
Add & Norm
Multi-Head Attention
Residual connection
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)

Mechanism



Positional Encoding

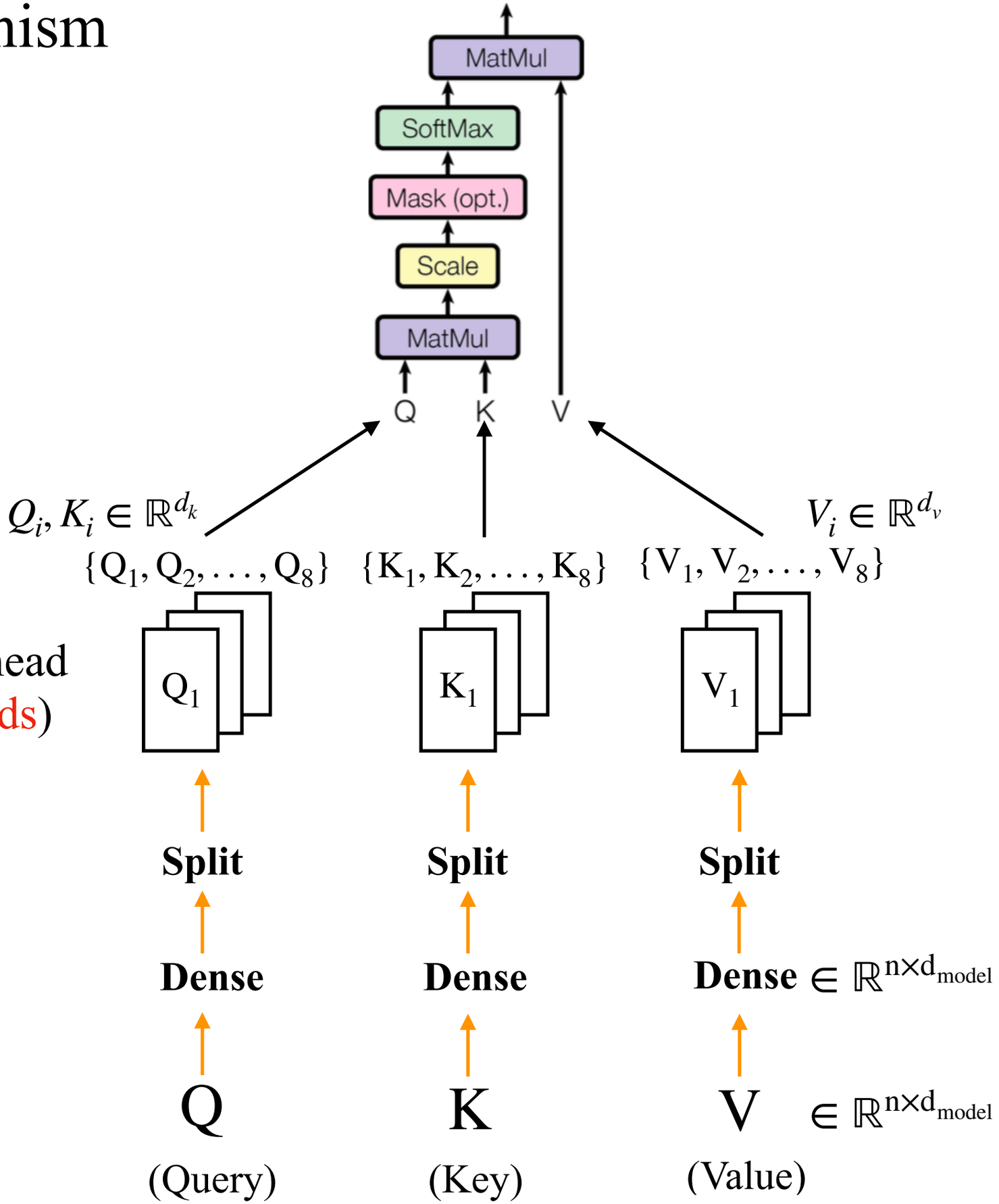
$$PE = \{p_0, p_1, p_2, \ldots, p_{n-1}\}$$

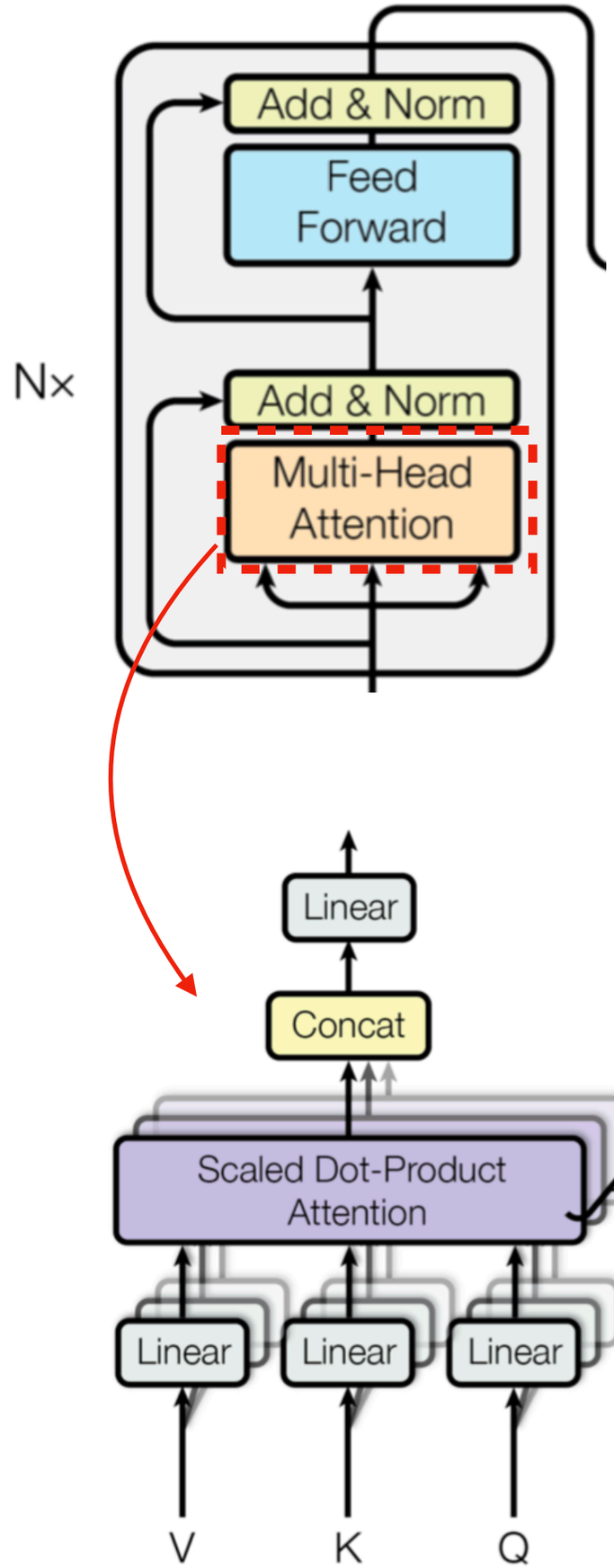
$$\begin{cases} PE_{(pos,2i)} = \sin(pos/10000^{\frac{2i}{d_{model}}}) \\ PE_{(pos,2i+1)} = \cos(pos/10000^{\frac{2i}{d_{model}}}) \end{cases}$$

Word Embedding

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

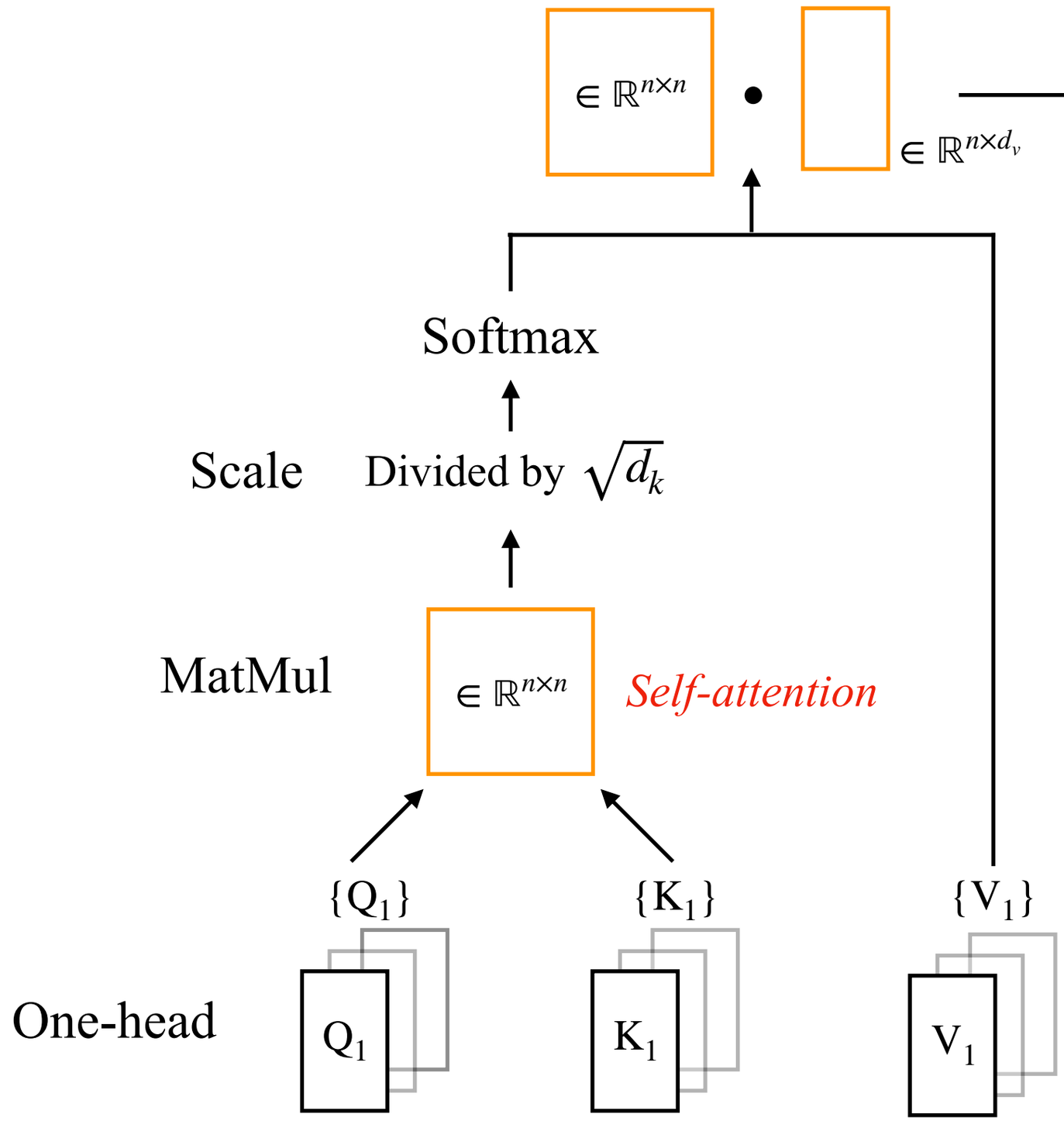
Mechanism


MatMul
SoftMax
Mask (opt.)
Scale
MatMul
Q
K
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
Dense
Dense
Dense $\in \mathbb{R}^{n \times d_{model}}$
Q
K
V $\in \mathbb{R}^{n \times d_{model}}$
(Query)
(Key)
(Value)

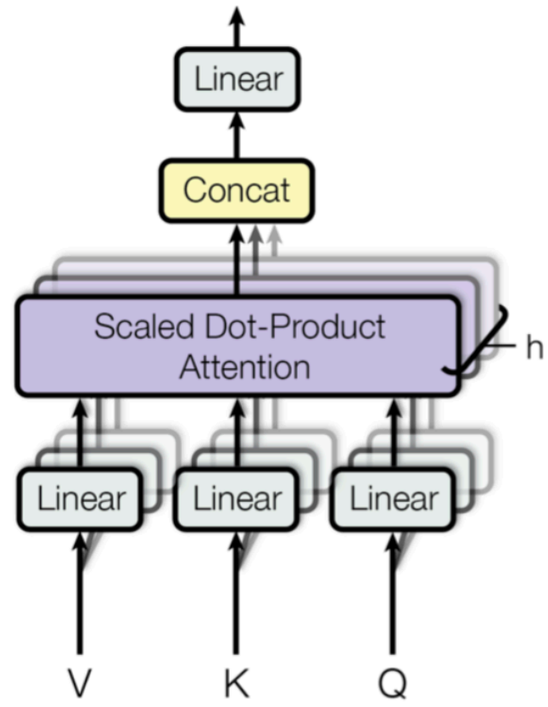

Nx
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Linear
Concat
Scaled Dot-Product
Attention
h
Linear
Linear
Linear
V
K
Q

Mechanism

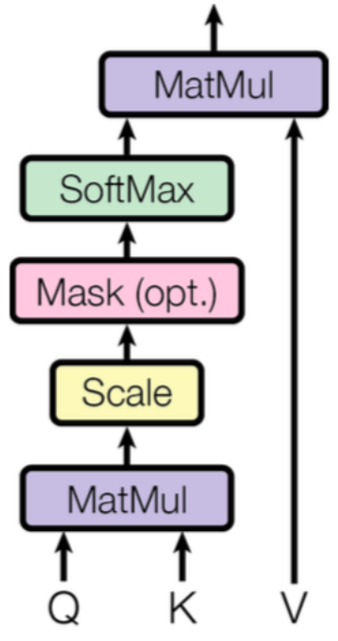
$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$



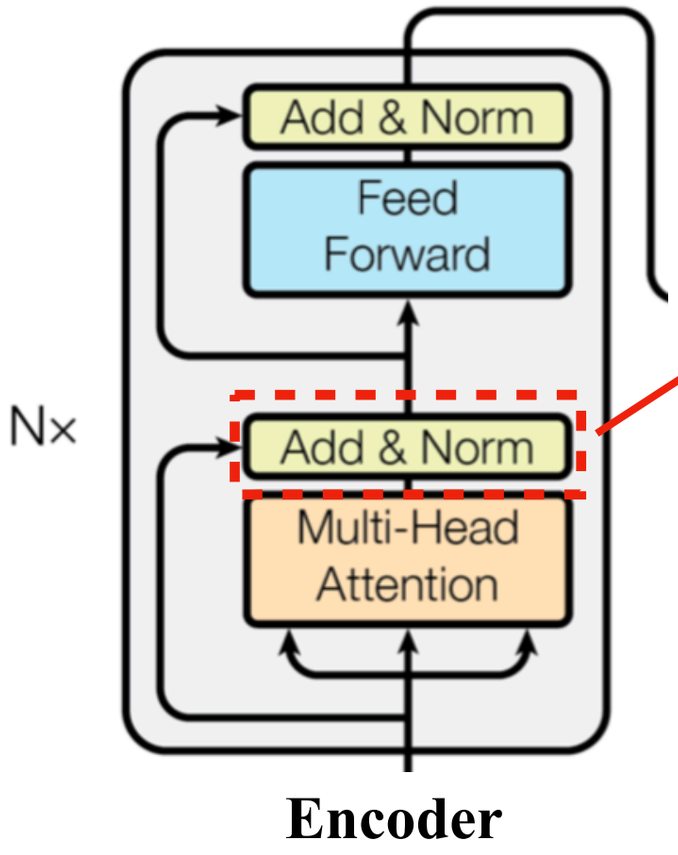
Multi-heads attention



Scaled Dot-Product Attention (Self-Attention)



Mechanism



Encoder

Residual connection

$$\text{Normalize}\Big(\underbrace{\square\square \cdots \square}_{\in \mathbb{R}^{n\times d_{model}}} + \underbrace{\text{Q (Query)}}_{\in \mathbb{R}^{n\times d_{model}}}\Big)$$

$$= \text{Normalize}\Big(\underbrace{\square}_{\in \mathbb{R}^{n\times d_{model}}}\Big)$$

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

$$\text{Mean} : \mu^l = \frac{1}{D}\sum_{i=1}^{D} z_i^l \qquad \text{Gains} : \alpha$$

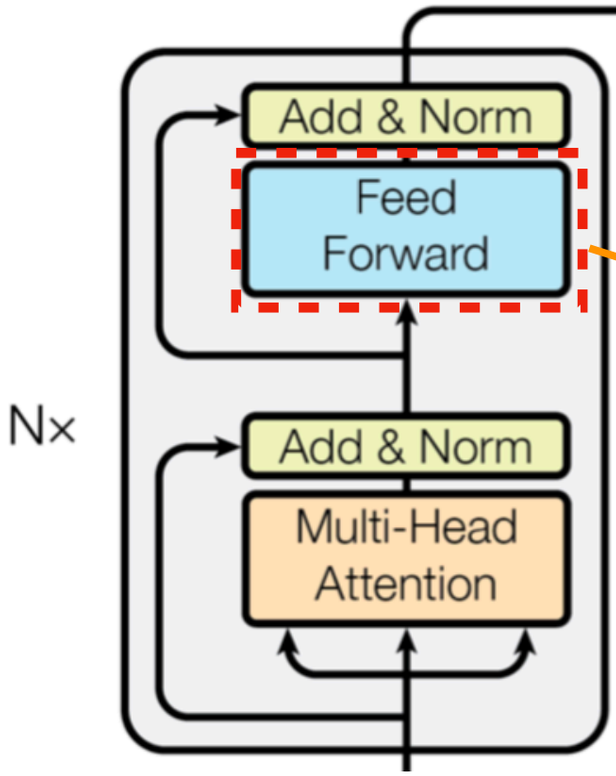
$$\text{Standard Deviation} : \sigma^l = \sqrt{\frac{1}{D}\sum_{i=1}^{D}(z_i^l - \mu^l)^2} \qquad \text{Biases} : \beta$$

Mechanism

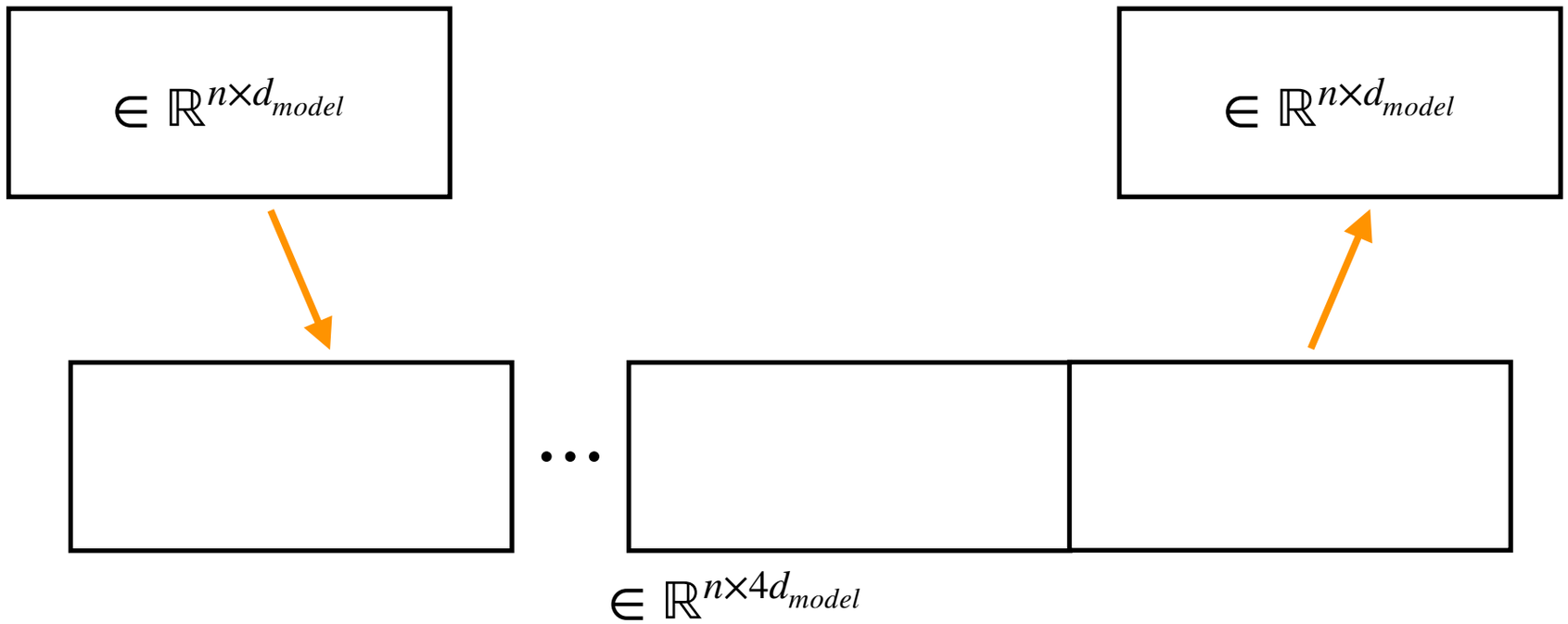
Feed Forward
(Dense、Conv1d)

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

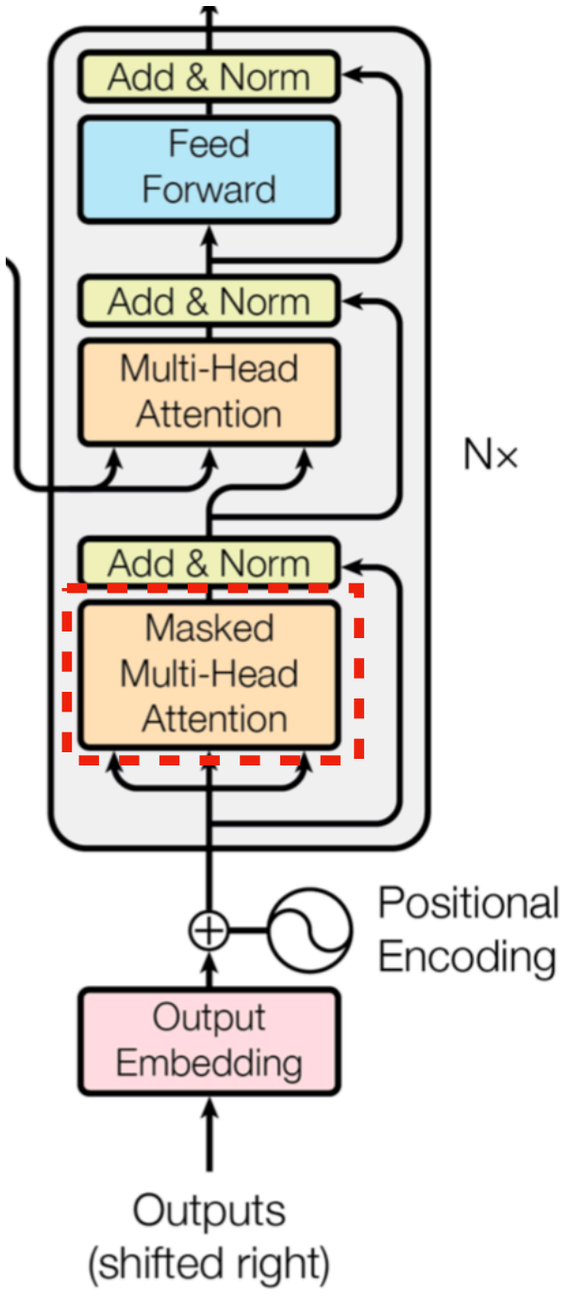
$$\mathrm{Dense}(d_{model}) \rightarrow \mathrm{Dense}\ (4 \times d_{model}) \rightarrow \mathrm{Dense}(d_{model})$$

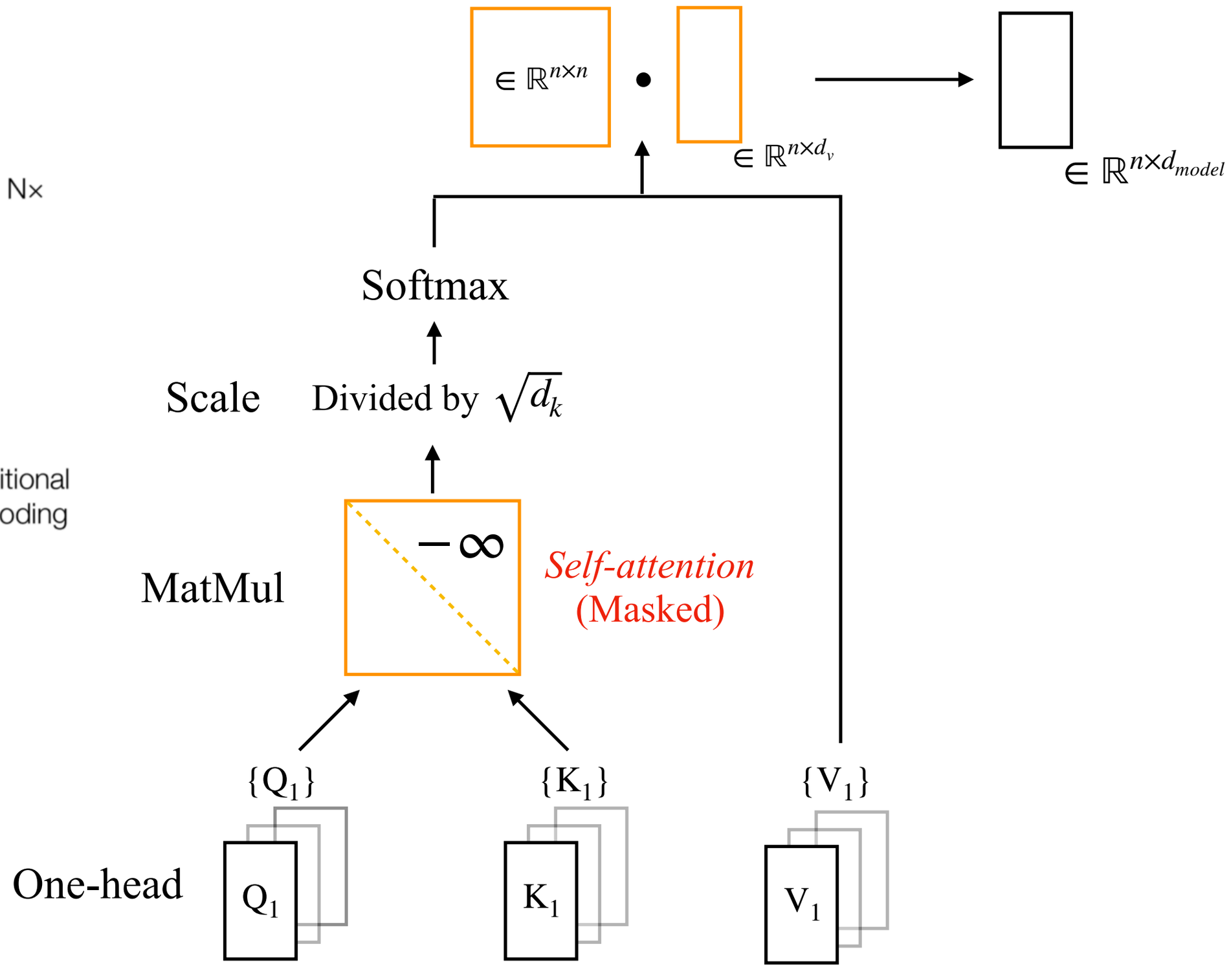


Encoder

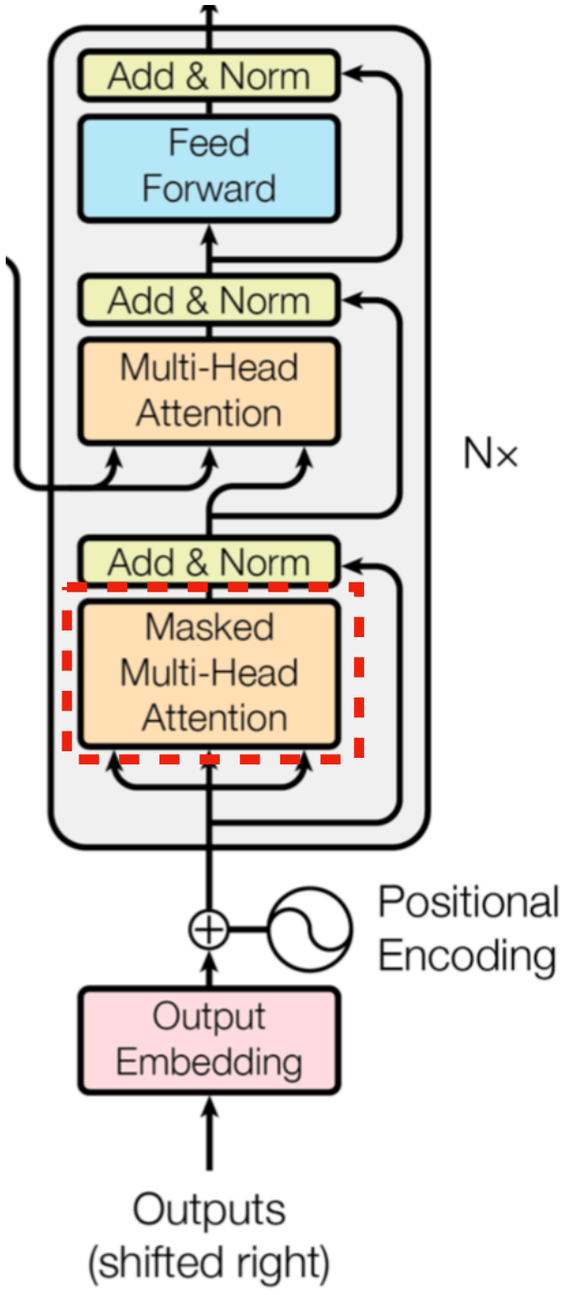


Mechanism


Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
N×
Positional Encoding
Output Embedding
Outputs (shifted right)
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

Mechanism

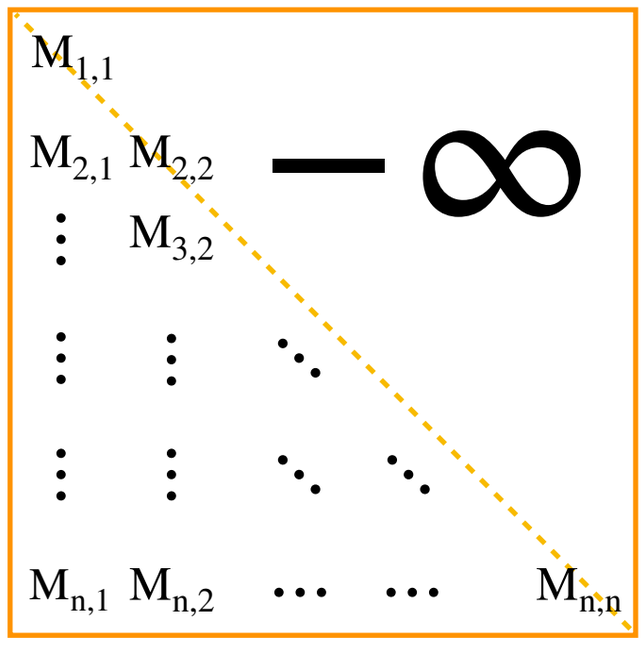


Softmax → 0

Scale Divided by $\sqrt{d_k}$

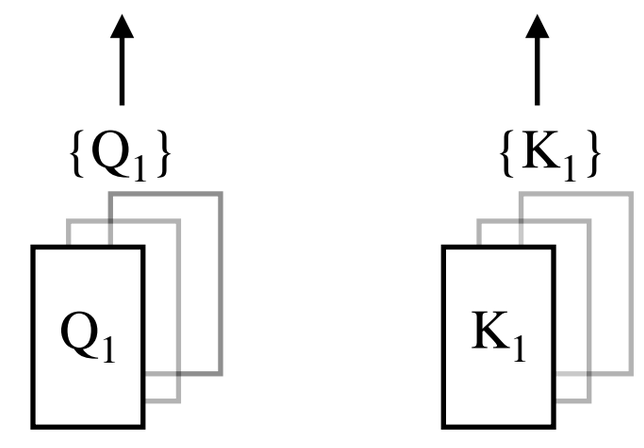
We need to prevent leftward information flow in the decoder

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

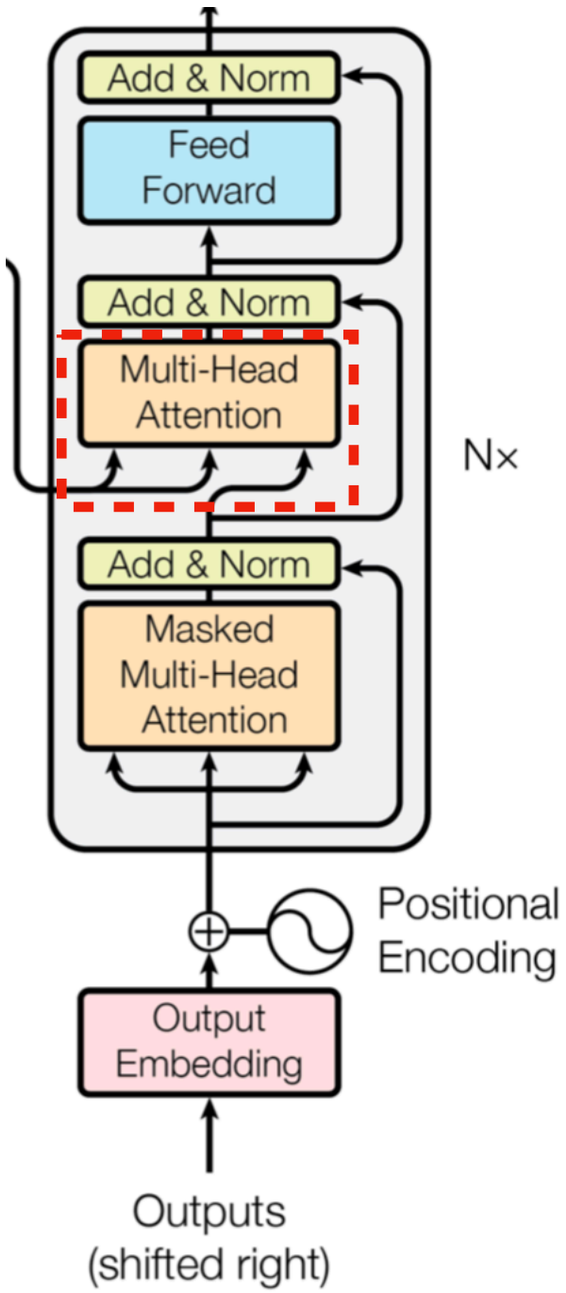


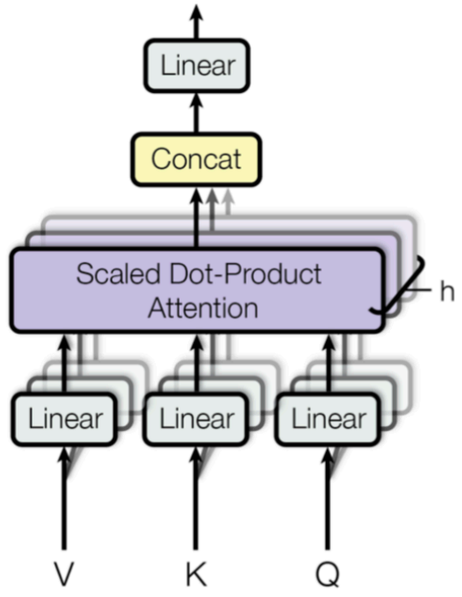
Self-attention (Masked)

One-head



Mechanism

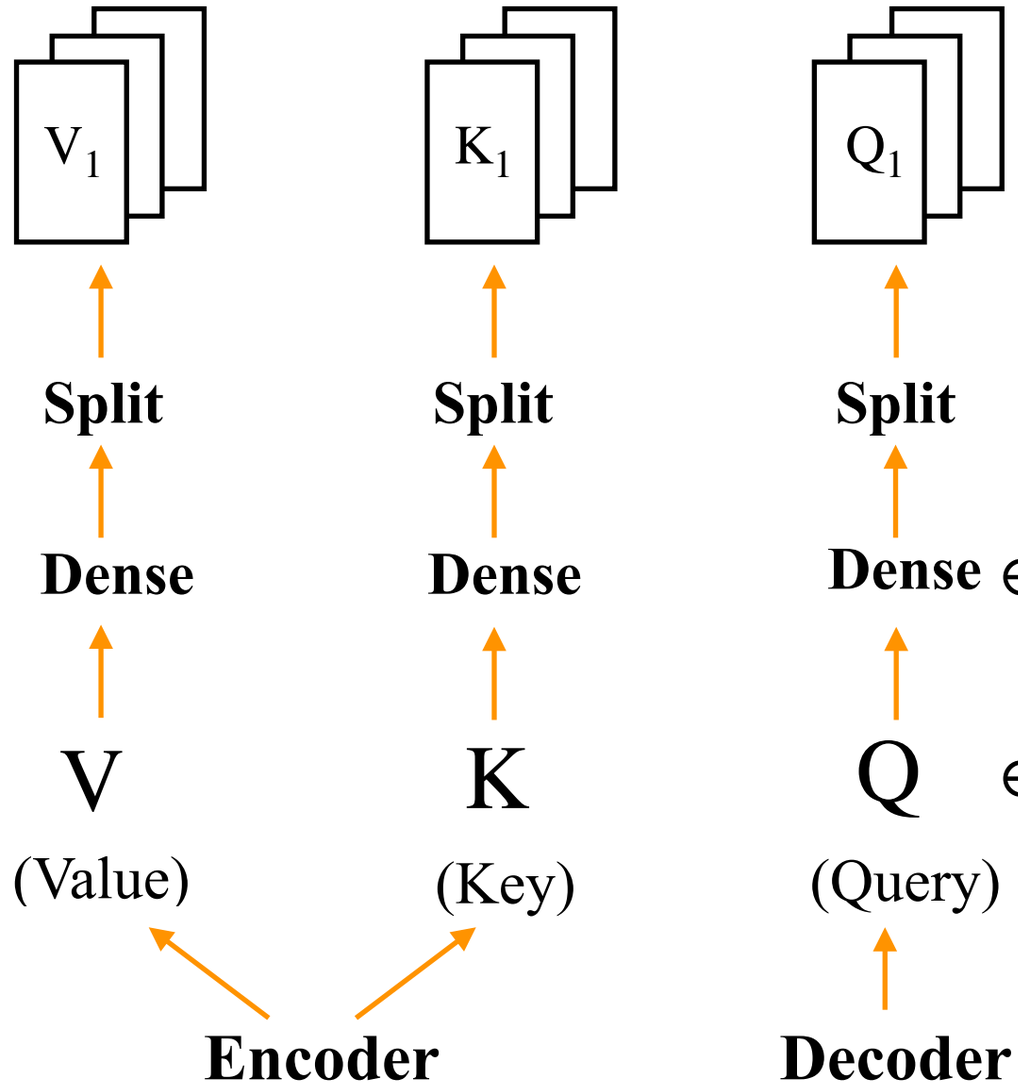




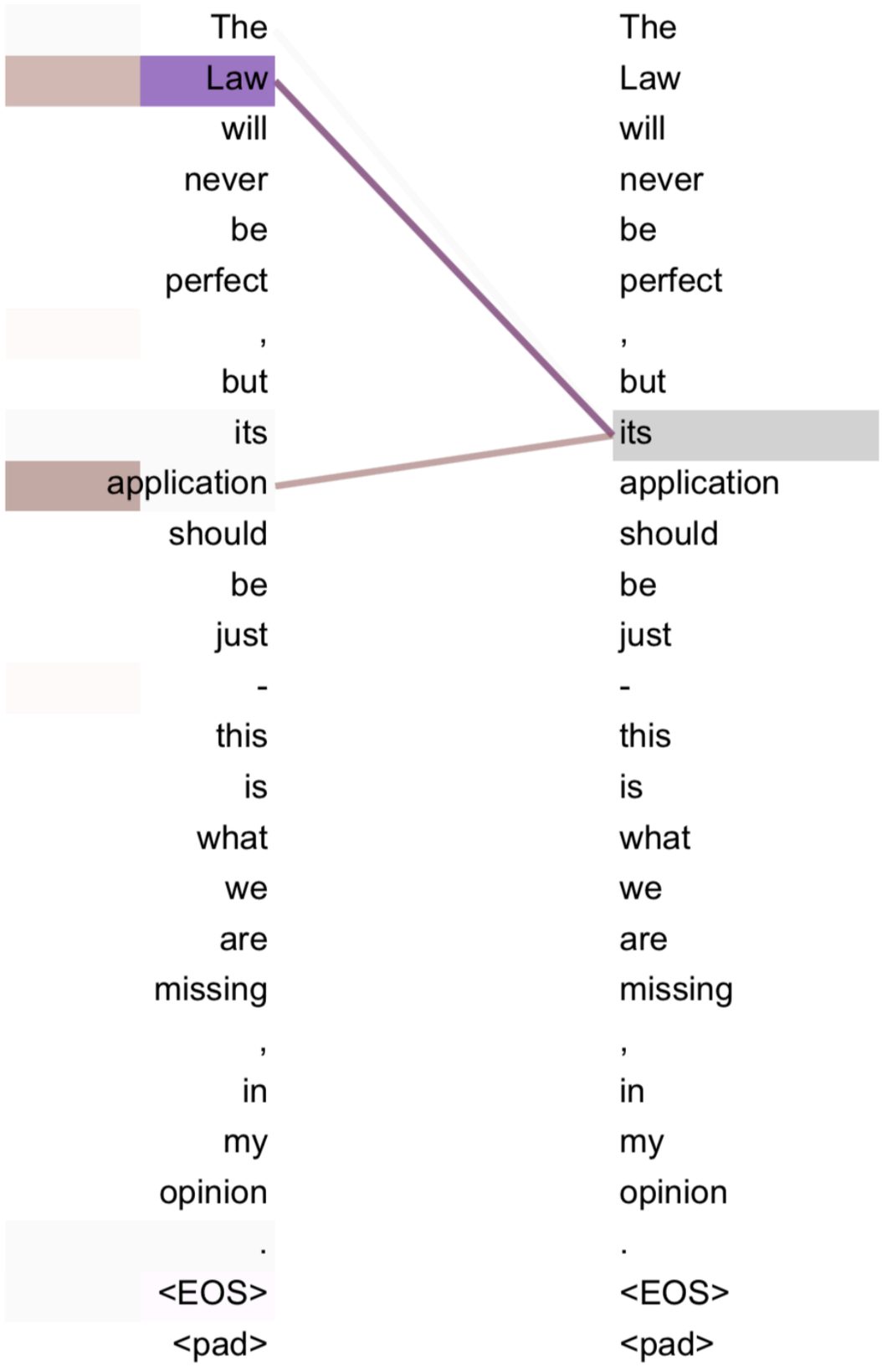
$V_i \in \mathbb{R}^{d_v}$

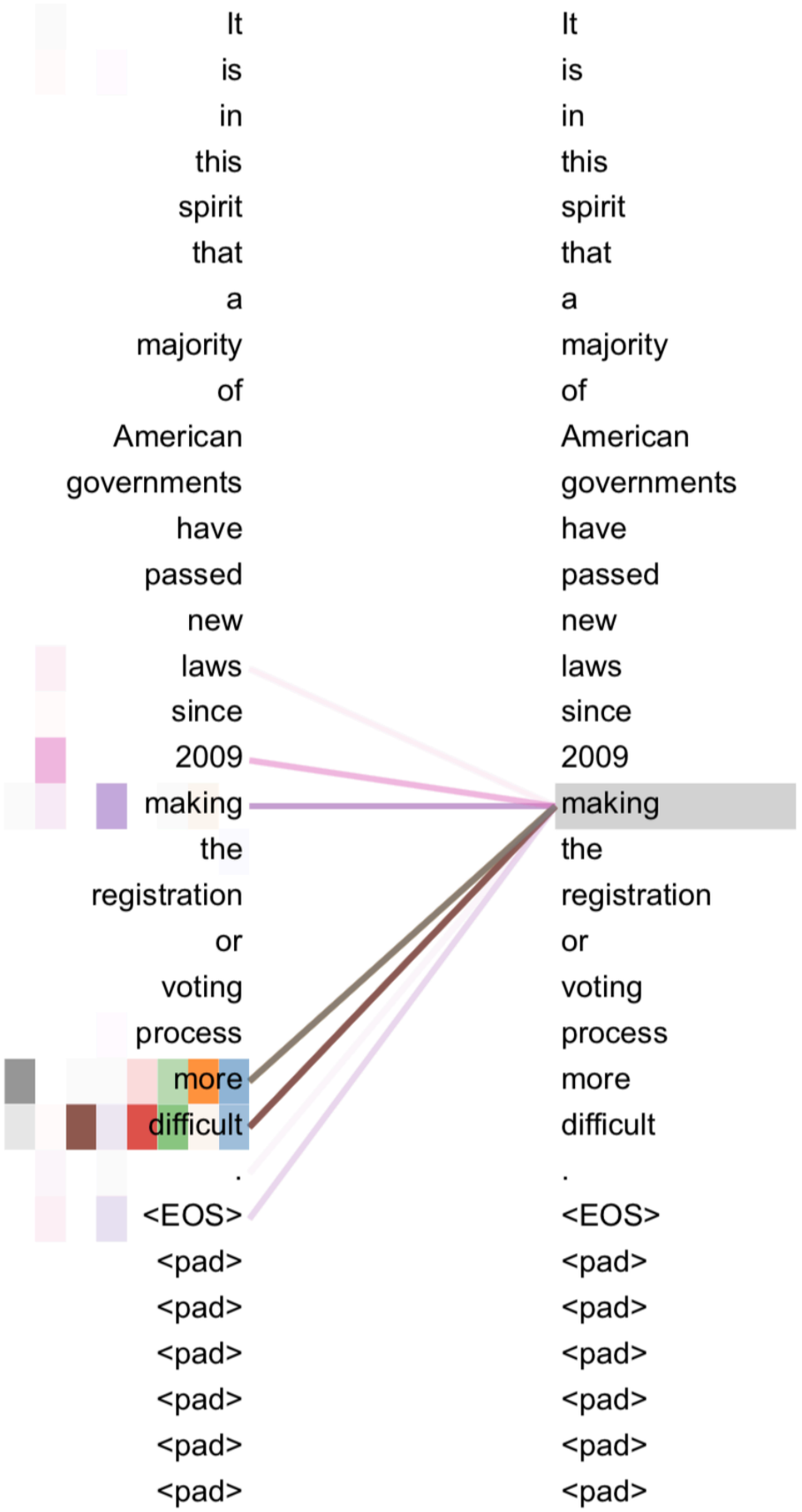
$Q_i, K_i \in \mathbb{R}^{d_k}$

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



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>

- **Attention mechanism:**

https://reurl.cc/NzNne

- **Attention is all you need:**

https://reurl.cc/nWlX1

- **BERT:**

https://reurl.cc/76bld

- **XLNet:**

https://reurl.cc/50baV