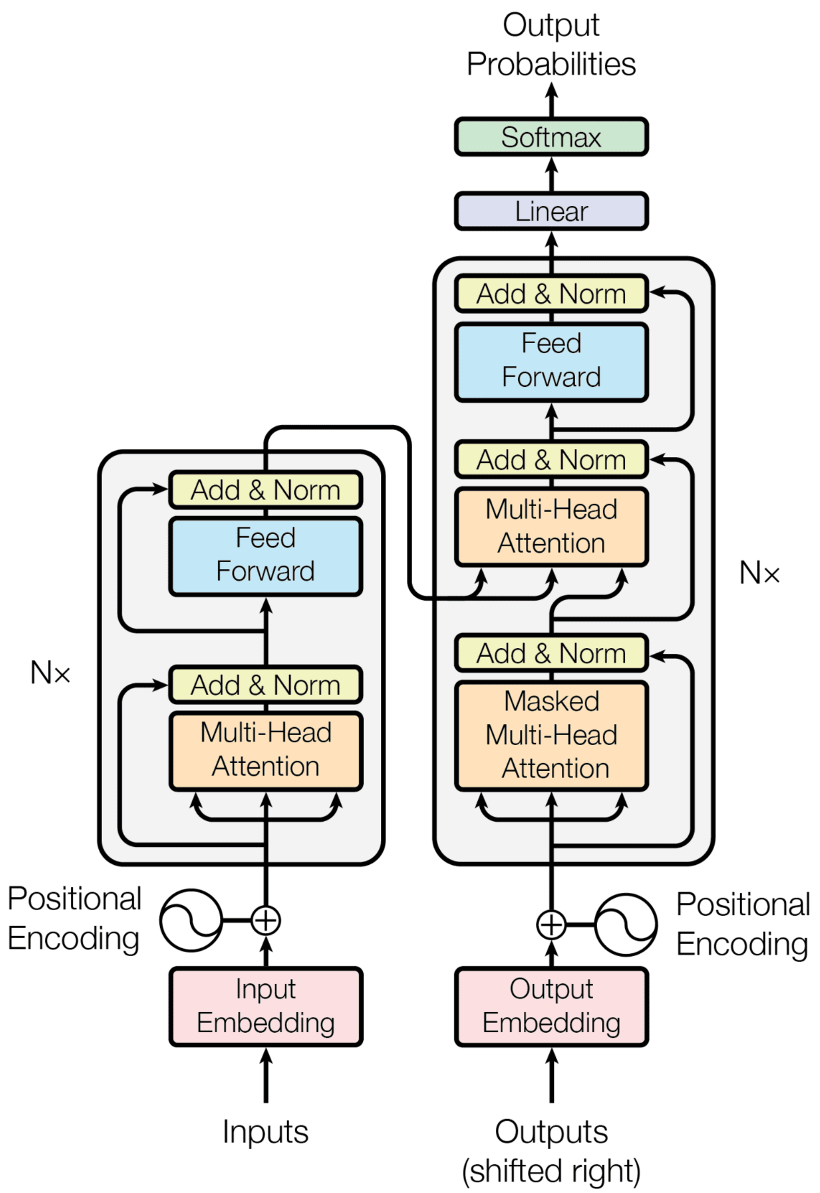
Introduction to Deep Learning
Transformers

Shikhar Agnihotri

Liangze Li

11-785, Fall 2023

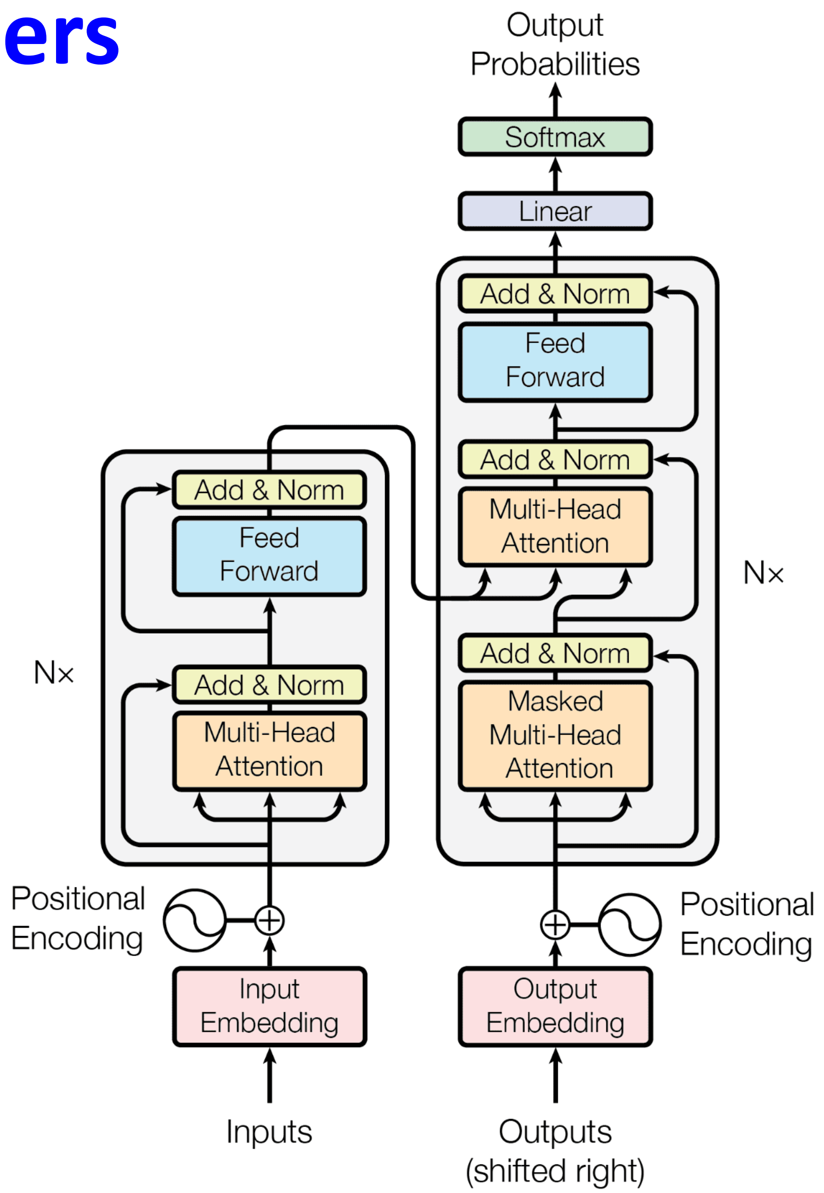
Transformers



Transformers

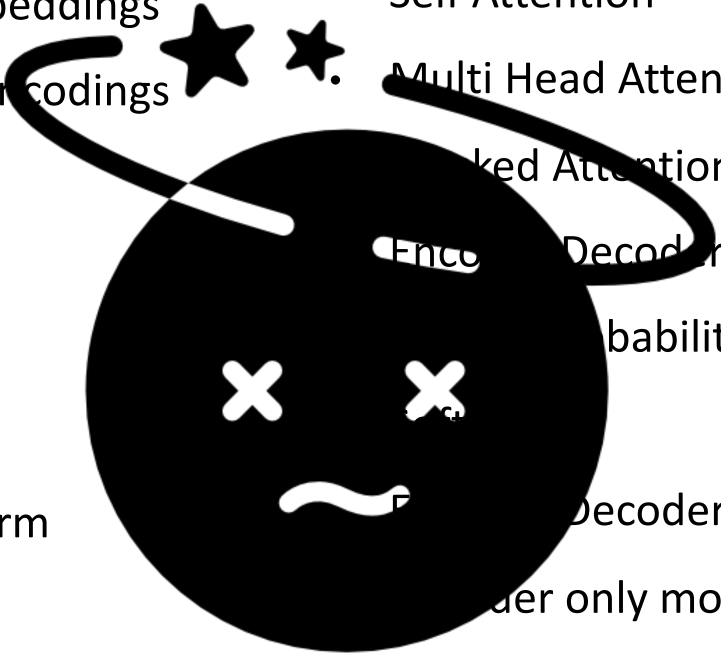
- Tokenizaton
- Input Embeddings
- Position Encodings
- Residuals
- Query
- Key
- Value
- Add & Norm
- Encoder
- Decoder

- Attention
- Self Attention
- Multi Head Attention
- Masked Attention
- Encoder Decoder Attention
- Output Probabilities / Logits
- Softmax
- Encoder-Decoder models
- Decoder only models

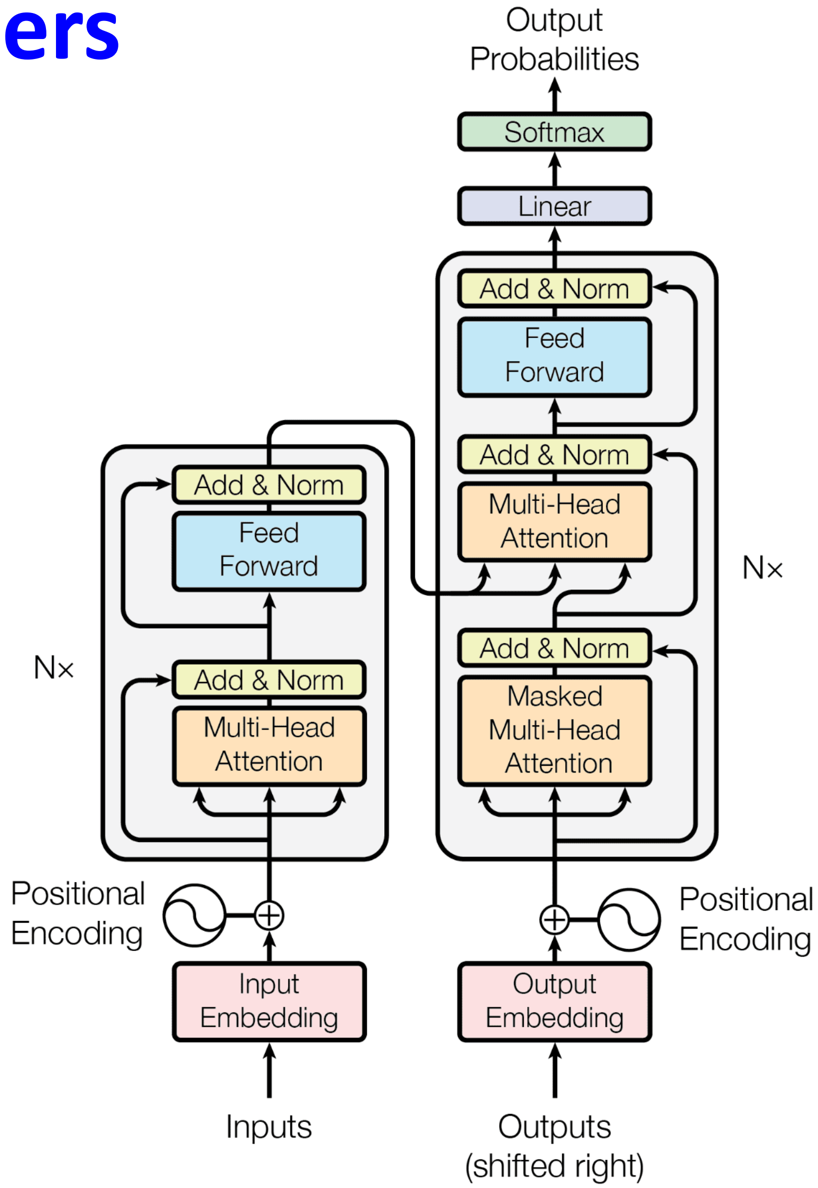


Transformers

- Tokenizaton
- Input Embeddings
- Position Encodings
- Residuals
- Query
- Key
- Value
- Add & Norm
- Encoder
- Decoder



- Attention
- Self Attention
- Multi Head Attention
- ked Attention
- Enc Decoder Attention
- babilities / Logits
- Decoder models
- der only models



Machine Translation

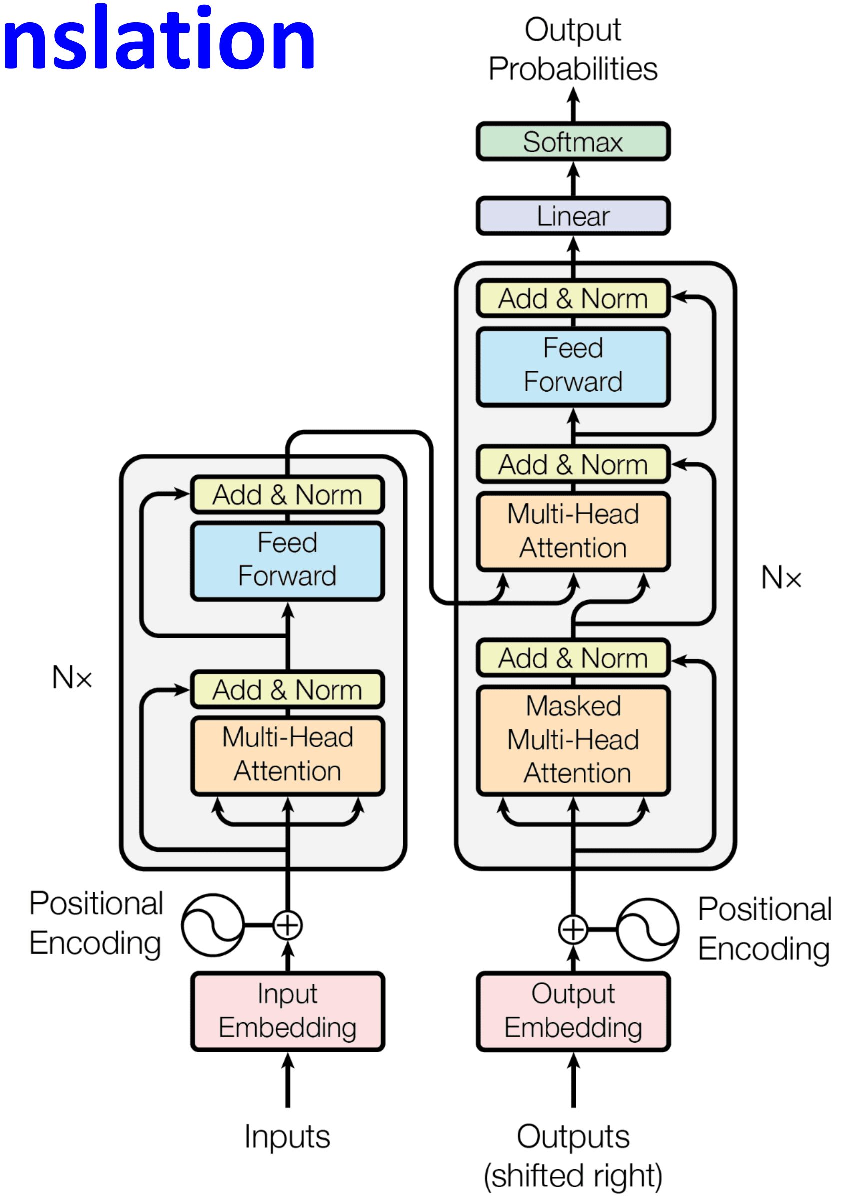
Targets

Ich have einen apfel gegessen

⬆

Inputs

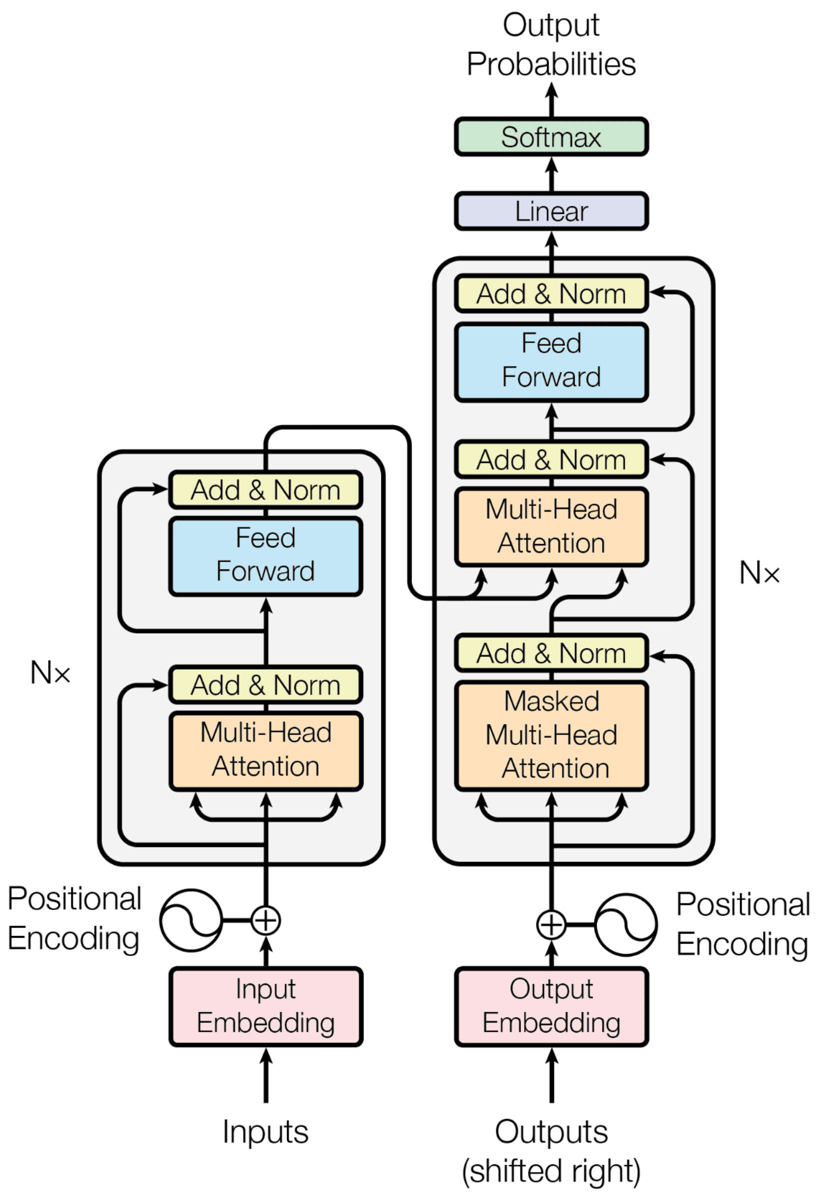
I ate an apple



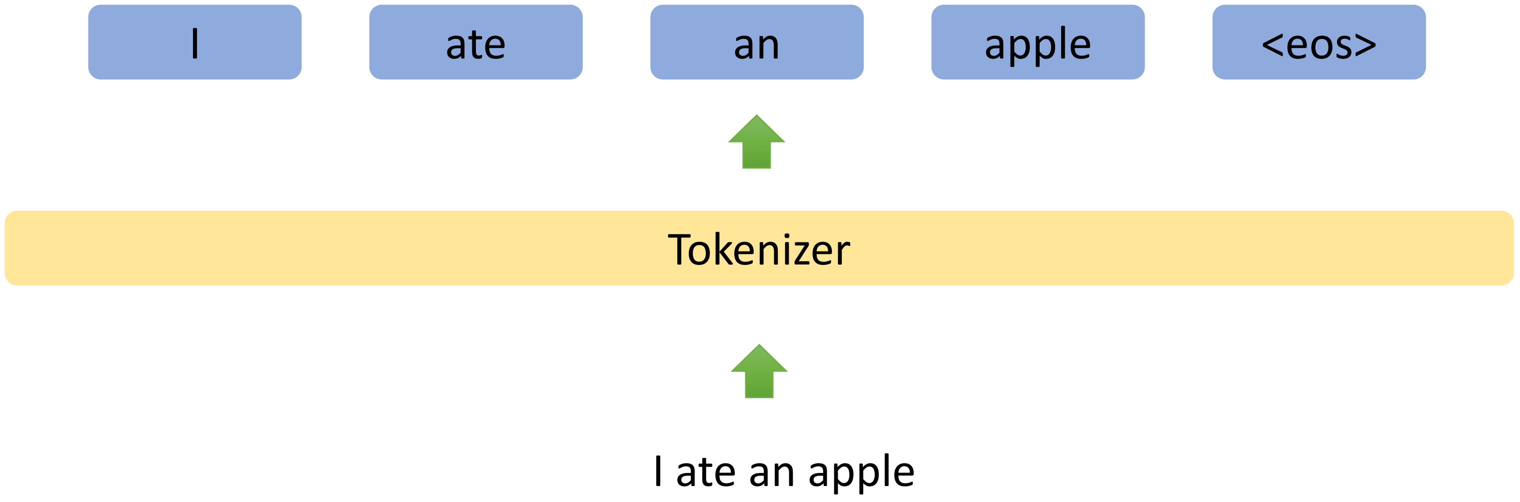
Inputs

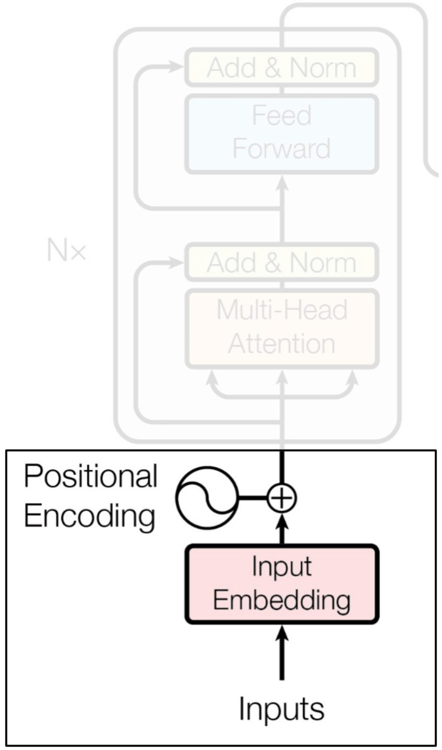
Processing Inputs

Inputs
I ate an apple



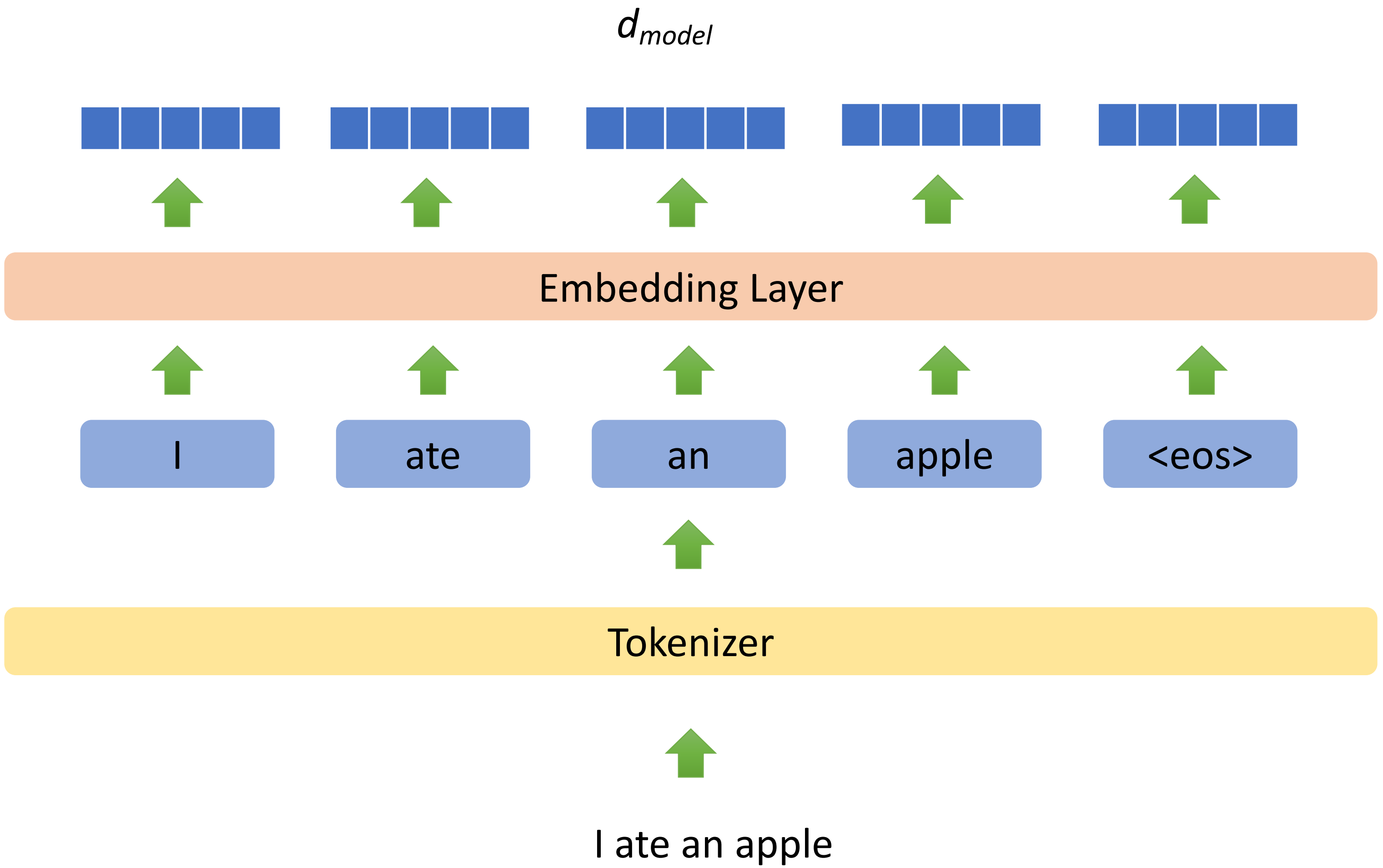
Inputs

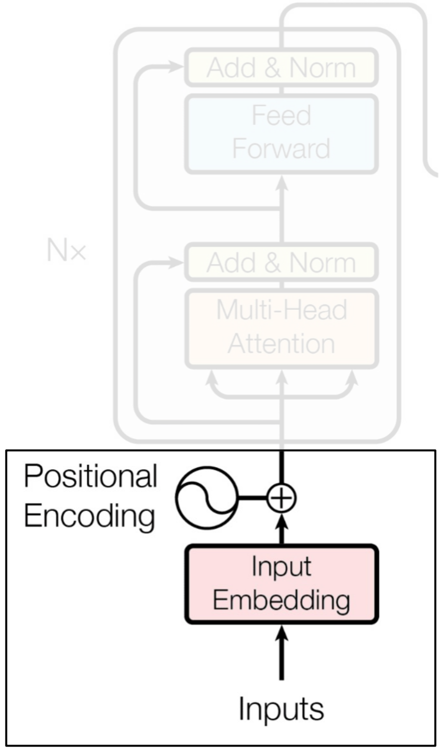




Generate Input Emebeddings

Inputs

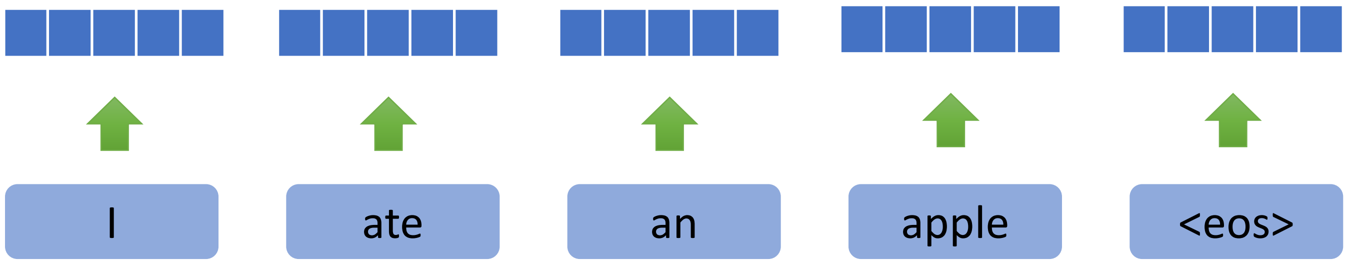


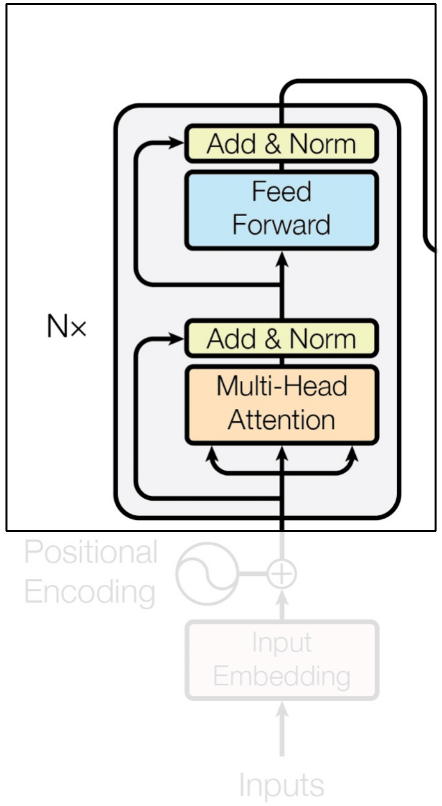


Generate Input Emebeddings

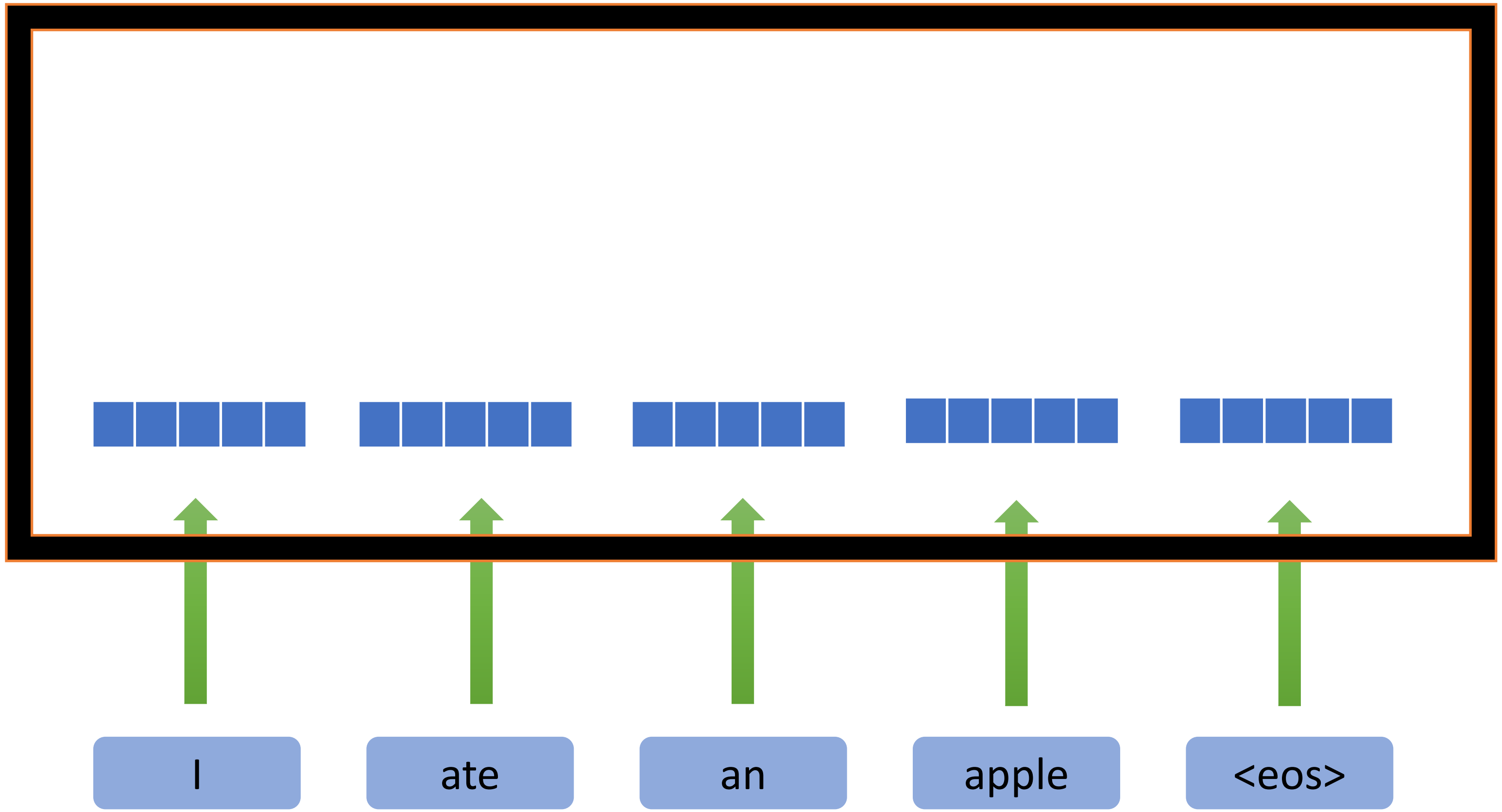
Encoder

WHERE IS THE CONTEXT ?

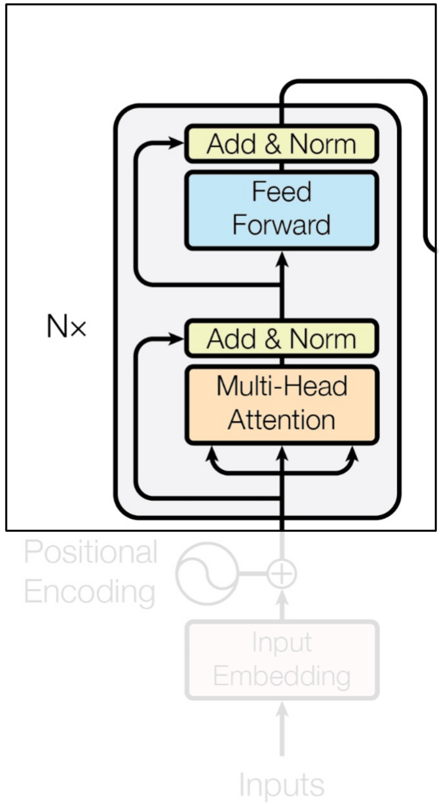




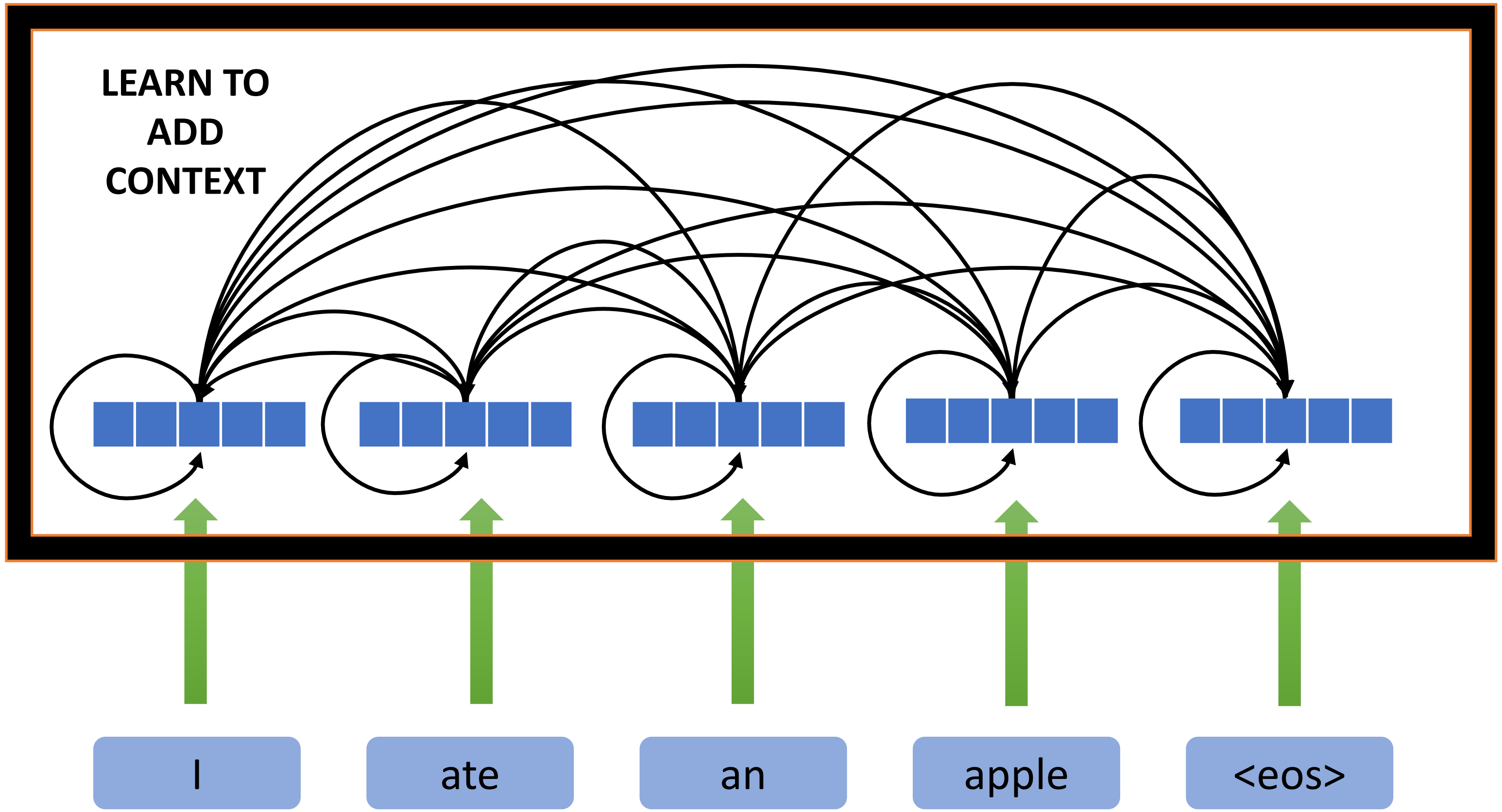
Encoder



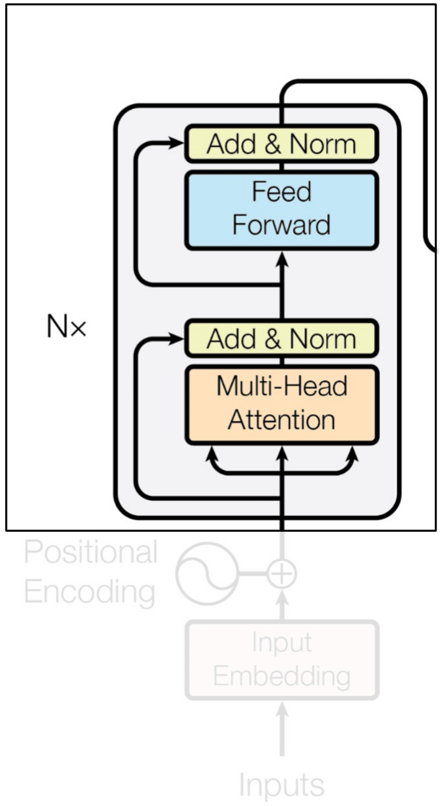
BLACK BOX OF SORTS



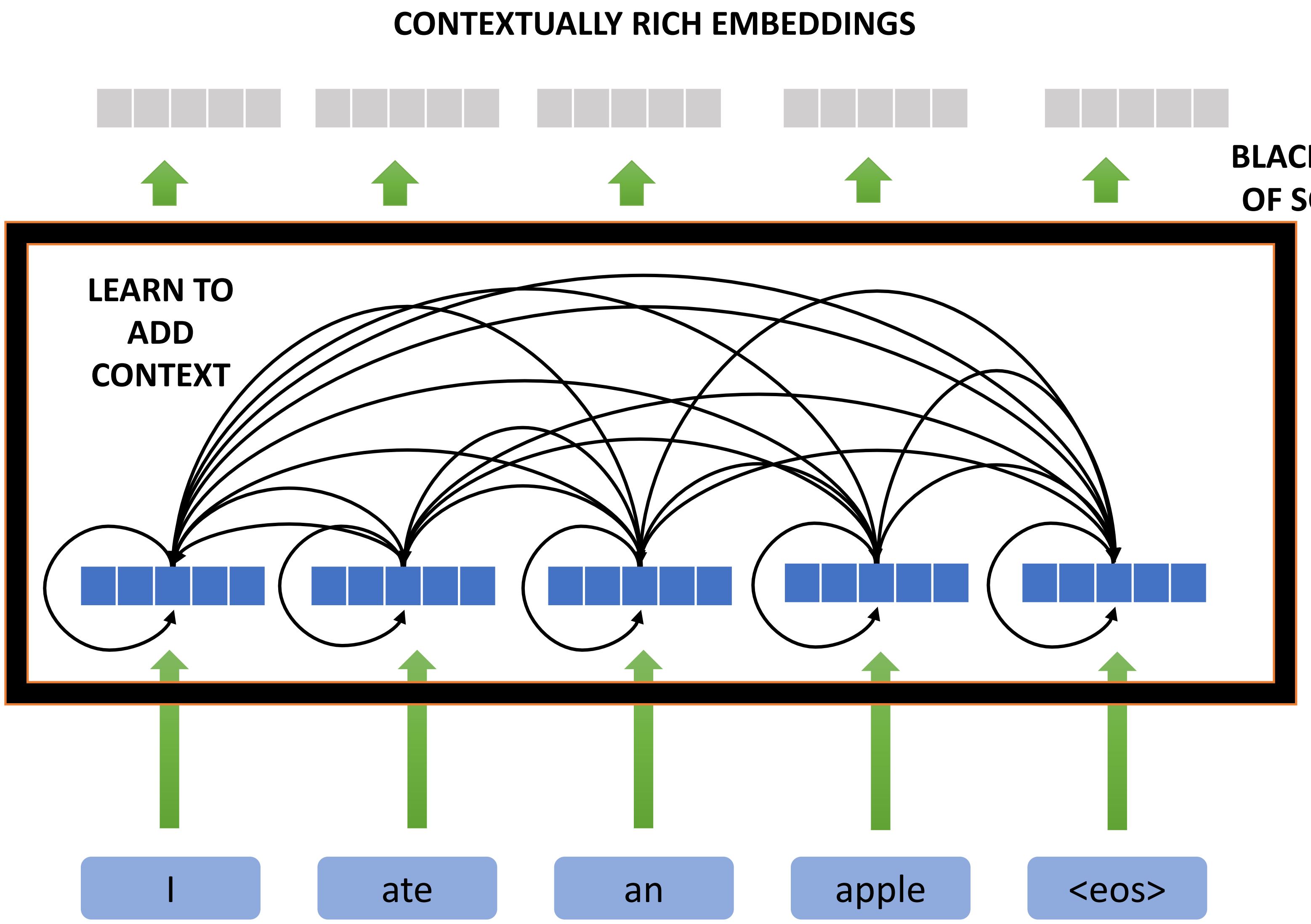
Encoder



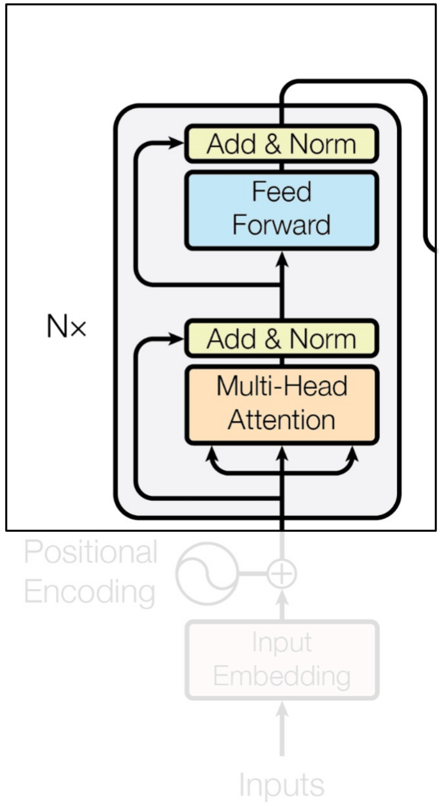
BLACK BOX
OF SORTS



Encoder



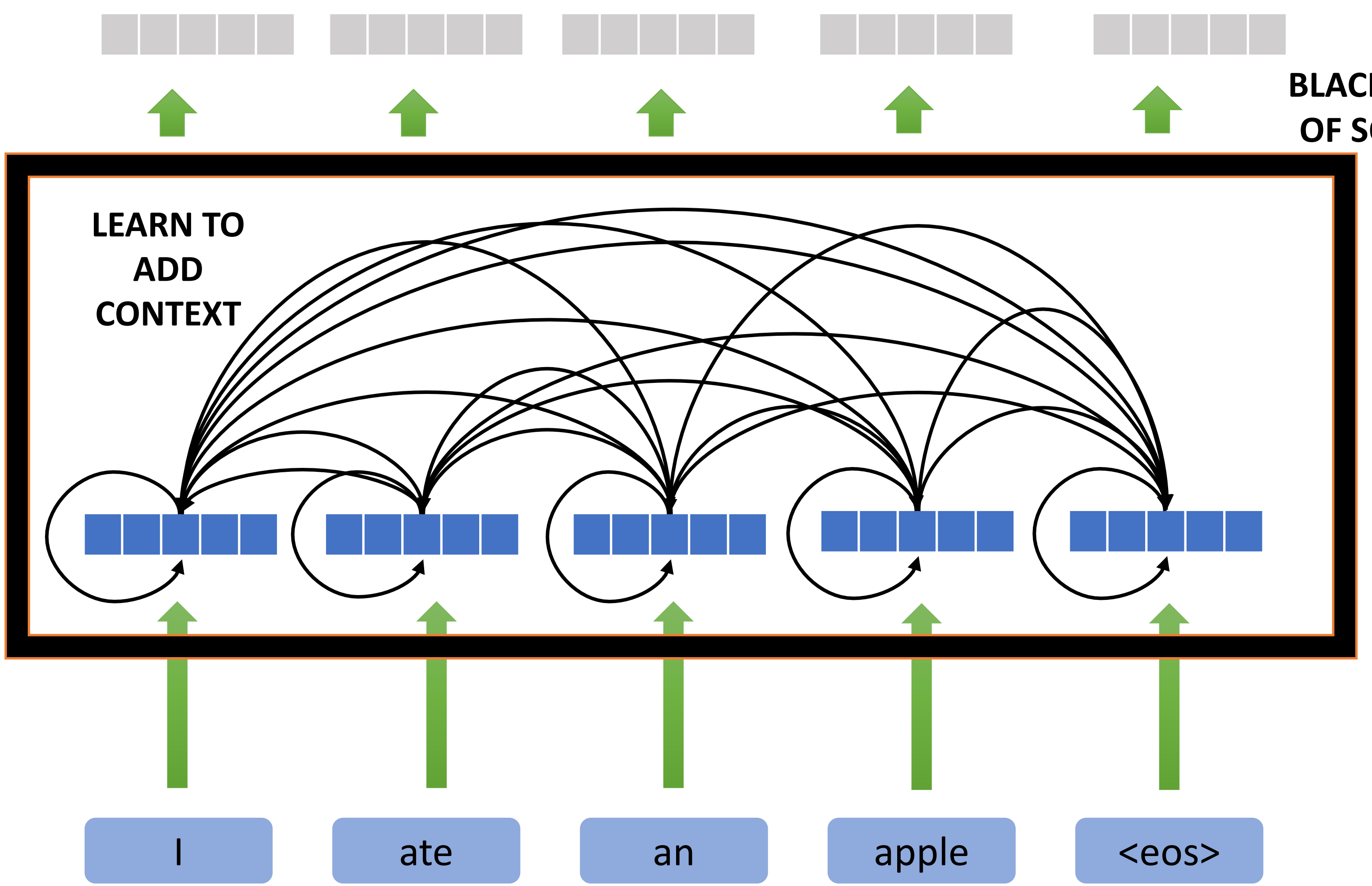
BLACK BOX
OF SORTS



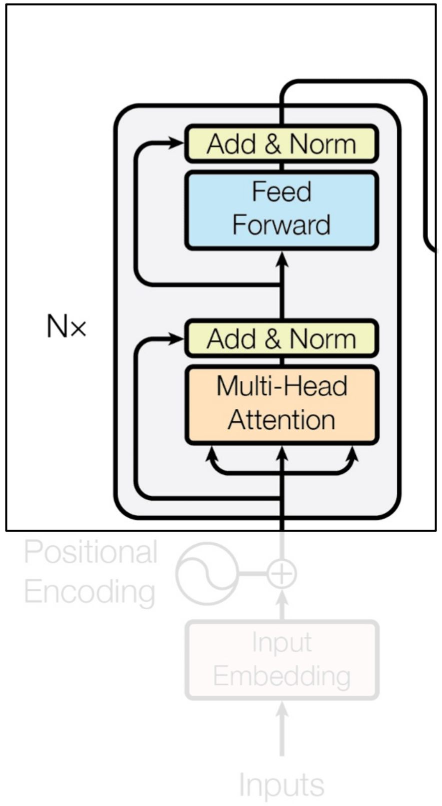
Encoder

$\alpha_{[\,i\,j\,]}$?

CONTEXTUALLY RICH EMBEDDINGS

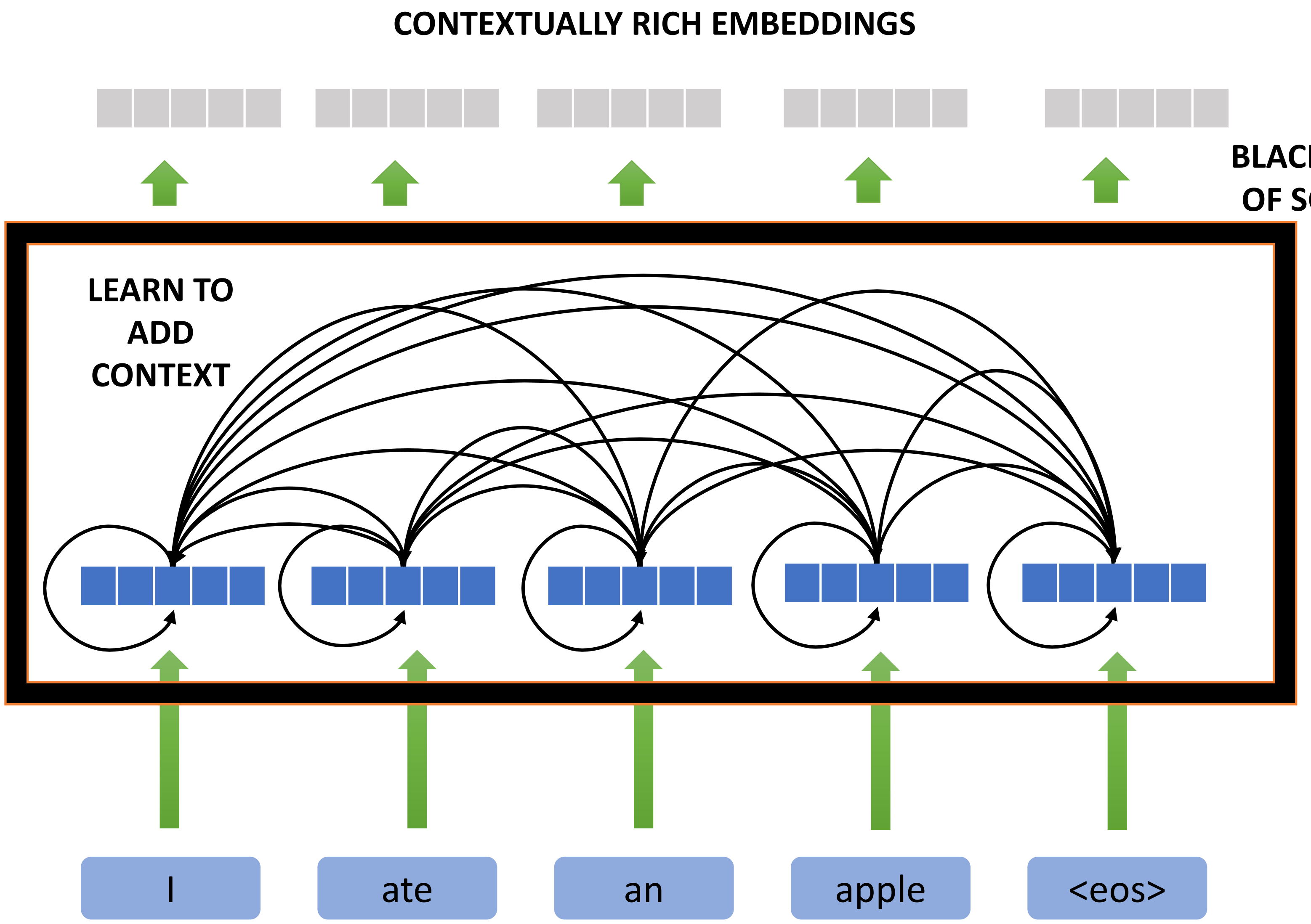


BLACK BOX
OF SORTS

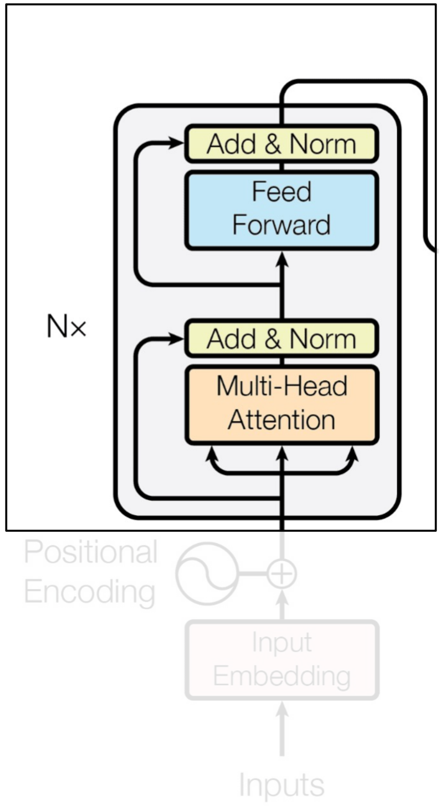


Encoder

$\alpha_{[\,i\,j\,]}$? Σ Π ?



BLACK BOX OF SORTS

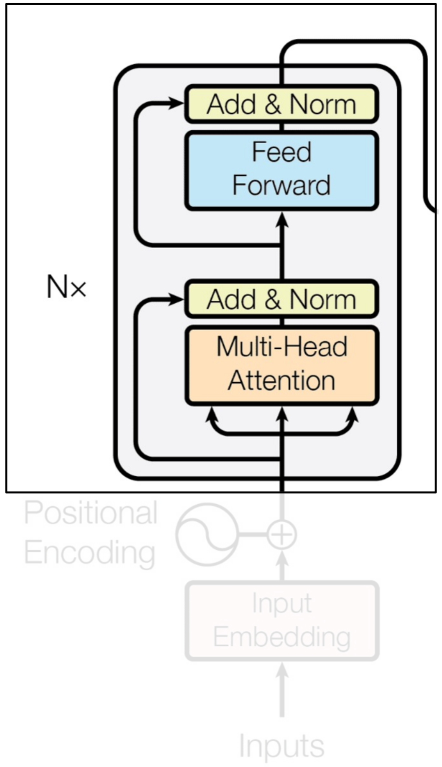


Attention

$\alpha_{[\,i\,j\,]}$?

From lecture 18:

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



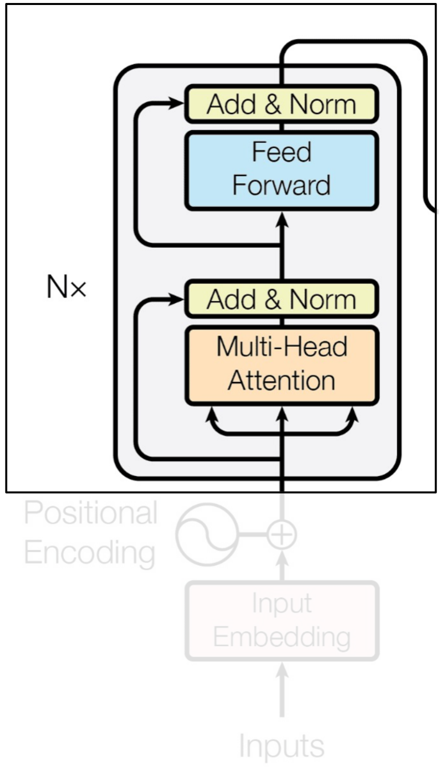
Attention

$\alpha_{[\,i\,j\,]}$?

From lecture 18:

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

- Query
- Key
- Value



Query, Key & Value

Database

{Key, Value store}


```
{"order_100": {"items":"a1", "delivery_date":"a2", ...}},
{"order_101": {"items":"b1", "delivery_date":"b2", ...}},
{"order_102": {"items":"c1", "delivery_date":"c2", ...}},
{"order_103": {"items":"d1", "delivery_date":"d2", ...}},
{"order_104": {"items":"e1", "delivery_date":"e2", ...}},
{"order_105": {"items":"f1", "delivery_date":"f2", ...}},
{"order_106": {"items":"g1", "delivery_date":"g2", ...}},
{"order_107": {"items":"h1", "delivery_date":"h2", ...}},
{"order_108": {"items":"i1", "delivery_date":"i2", ...}},
{"order_109": {"items":"j1", "delivery_date":"j2", ...}},
{"order_110": {"items":"k1", "delivery_date":"k2", ...}}
```

Query, Key & Value

{Query: “Order details of order_104”}

OR

{Query: “Order details of order_106”}

Database

{Key, Value store}

```
{"order_100": {"items":"a1", "delivery_date":"a2", ...}},
{"order_101": {"items":"b1", "delivery_date":"b2", ...}},
{"order_102": {"items":"c1", "delivery_date":"c2", ...}},
{"order_103": {"items":"d1", "delivery_date":"d2", ...}},
{"order_104": {"items":"e1", "delivery_date":"e2", ...}},
{"order_105": {"items":"f1", "delivery_date":"f2", ...}},
{"order_106": {"items":"g1", "delivery_date":"g2", ...}},
{"order_107": {"items":"h1", "delivery_date":"h2", ...}},
{"order_108": {"items":"i1", "delivery_date":"i2", ...}},
{"order_109": {"items":"j1", "delivery_date":"j2", ...}},
{"order_110": {"items":"k1", "delivery_date":"k2", ...}}
```

Query, Key & Value

{Query: "Order details of order_104"}

OR

{Query: "Order details of order_106"}

{Key, Value store}

```
{"order_100": {"items":"a1", "delivery_date":"a2", ...}},
{"order_101": {"items":"b1", "delivery_date":"b2", ...}},
{"order_102": {"items":"c1", "delivery_date":"c2", ...}},
{"order_103": {"items":"d1", "delivery_date":"d2", ...}},
{"order_104": {"items":"e1", "delivery_date":"e2", ...}},
{"order_105": {"items":"f1", "delivery_date":"f2", ...}},
{"order_106": {"items":"g1", "delivery_date":"g2", ...}},
{"order_107": {"items":"h1", "delivery_date":"h2", ...}},
{"order_108": {"items":"i1", "delivery_date":"i2", ...}},
{"order_109": {"items":"j1", "delivery_date":"j2", ...}},
{"order_110": {"items":"k1", "delivery_date":"k2", ...}}
```

Query, Key & Value

{Query: "Order details of order_104"}

OR

{Query: "Order details of order_106"}

{Key, Value store}

```
{"order_100": {"items":"a1", "delivery_date":"a2", ...}},
{"order_101": {"items":"b1", "delivery_date":"b2", ...}},
{"order_102": {"items":"c1", "delivery_date":"c2", ...}},
{"order_103": {"items":"d1", "delivery_date":"d2", ...}},
{"order_104": {"items":"e1", "delivery_date":"e2", ...}},
{"order_105": {"items":"f1", "delivery_date":"f2", ...}},
{"order_106": {"items":"g1", "delivery_date":"g2", ...}},
{"order_107": {"items":"h1", "delivery_date":"h2", ...}},
{"order_108": {"items":"i1", "delivery_date":"i2", ...}},
{"order_109": {"items":"j1", "delivery_date":"j2", ...}},
{"order_110": {"items":"k1", "delivery_date":"k2", ...}}
```

Query, Key & Value

{Query: "Order details of order_104"}

OR

{Query: "Order details of order_106"}

{Key, Value store}

```
{"order_100": {"items":"a1", "delivery_date":"a2", ...}},
{"order_101": {"items":"b1", "delivery_date":"b2", ...}},
{"order_102": {"items":"c1", "delivery_date":"c2", ...}},
{"order_103": {"items":"d1", "delivery_date":"d2", ...}},
{"order_104": {"items":"e1", "delivery_date":"e2", ...}},
{"order_105": {"items":"f1", "delivery_date":"f2", ...}},
{"order_106": {"items":"g1", "delivery_date":"g2", ...}},
{"order_107": {"items":"h1", "delivery_date":"h2", ...}},
{"order_108": {"items":"i1", "delivery_date":"i2", ...}},
{"order_109": {"items":"j1", "delivery_date":"j2", ...}},
{"order_110": {"items":"k1", "delivery_date":"k2", ...}}
```

Query, Key & Value

Done at the same time !!

{Query: "Order details of order_104"}

OR

{Query: "Order details of order_106"}

{Key, Value store}

```
{"order_100": {"items":"a1", "delivery_date":"a2", ...}},
{"order_101": {"items":"b1", "delivery_date":"b2", ...}},
{"order_102": {"items":"c1", "delivery_date":"c2", ...}},
{"order_103": {"items":"d1", "delivery_date":"d2", ...}},
{"order_104": {"items":"e1", "delivery_date":"e2", ...}},
{"order_105": {"items":"f1", "delivery_date":"f2", ...}},
{"order_106": {"items":"g1", "delivery_date":"g2", ...}},
{"order_107": {"items":"h1", "delivery_date":"h2", ...}},
{"order_108": {"items":"i1", "delivery_date":"i2", ...}},
{"order_109": {"items":"j1", "delivery_date":"j2", ...}},
{"order_110": {"items":"k1", "delivery_date":"k2", ...}}
```

Query, Key & Value

{Query: "Order details of order_104"}

OR

{Query: "Order details of order_106"}

```
{"order_100": {"items":"a1", "delivery_date":"a2", ...}},
{"order_101": {"items":"b1", "delivery_date":"b2", ...}},
{"order_102": {"items":"c1", "delivery_date":"c2", ...}},
{"order_103": {"items":"d1", "delivery_date":"d2", ...}},
{"order_104": {"items":"e1", "delivery_date":"e2", ...}},
{"order_105": {"items":"f1", "delivery_date":"f2", ...}},
{"order_106": {"items":"g1", "delivery_date":"g2", ...}},
{"order_107": {"items":"h1", "delivery_date":"h2", ...}},
{"order_108": {"items":"i1", "delivery_date":"i2", ...}},
{"order_109": {"items":"j1", "delivery_date":"j2", ...}},
{"order_110": {"items":"k1", "delivery_date":"k2", ...}}
```

Query

1. Search for info

Key

1. Interacts directly with Queries
2. Distinguishes one object from another
3. Identify which object is the most relevant and by how much

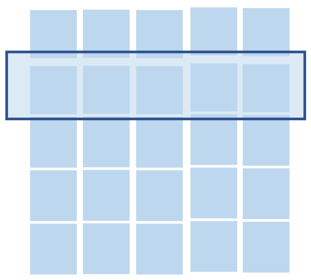
Value

1. Actual details of the object
2. More fine grained

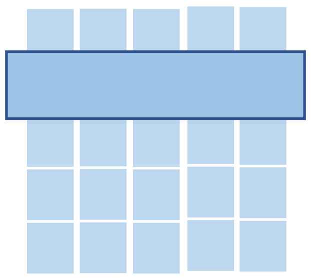
Attention

Query

Key Value Store

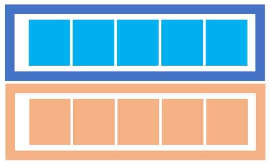


Key



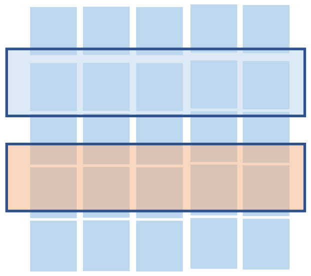
Value

Attention

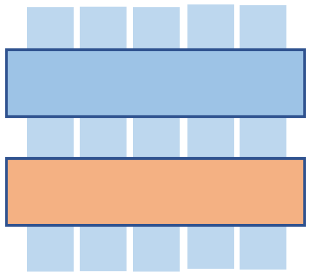


Query

Key Value Store



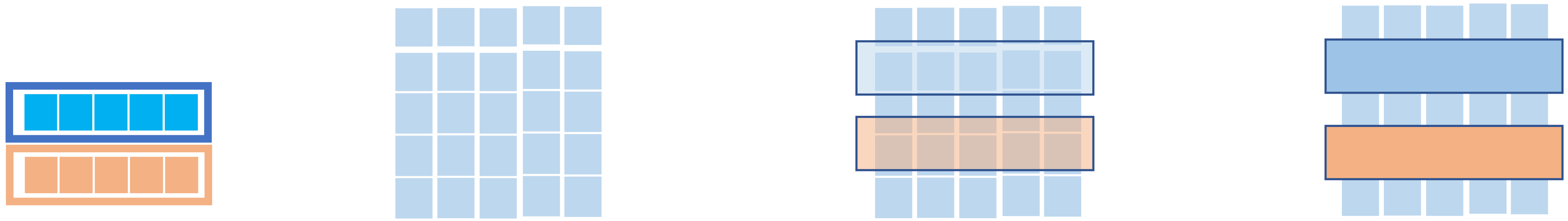
Key



Value

Attention

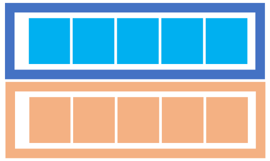
Done at the same time !!

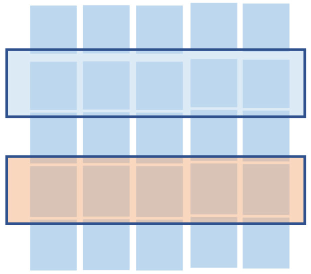


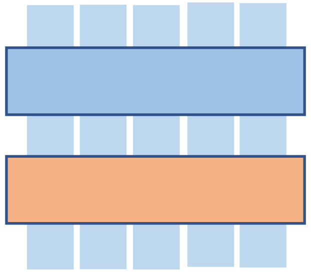
Query | Key Value Store | Key | Value

Attention

Parallelizable !!!







Query | Key Value Store | Key | Value

Q

QK^T

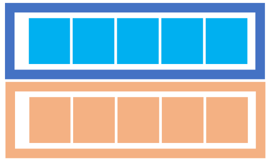
$softmax(\frac{QK^T}{\sqrt{d}})$

$softmax(\frac{QK^T}{\sqrt{d}})V$

Attention

Parallelizable !!!

Attention Filter

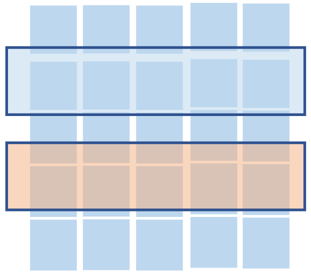


Query

Q

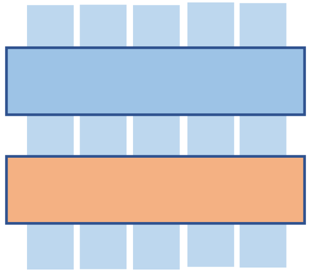
Key Value Store

QK^T



Key

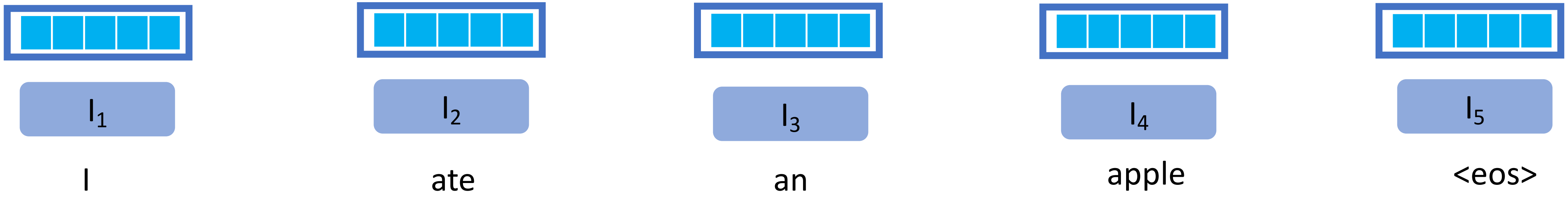
$softmax(\frac{QK^T}{\sqrt{d}})$



Value

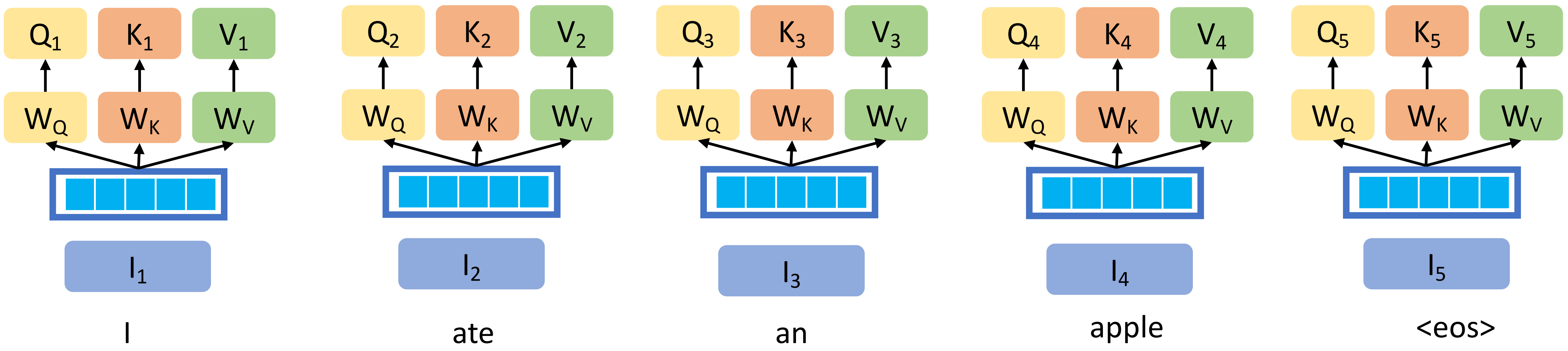
$softmax(\frac{QK^T}{\sqrt{d}})V$

Attention



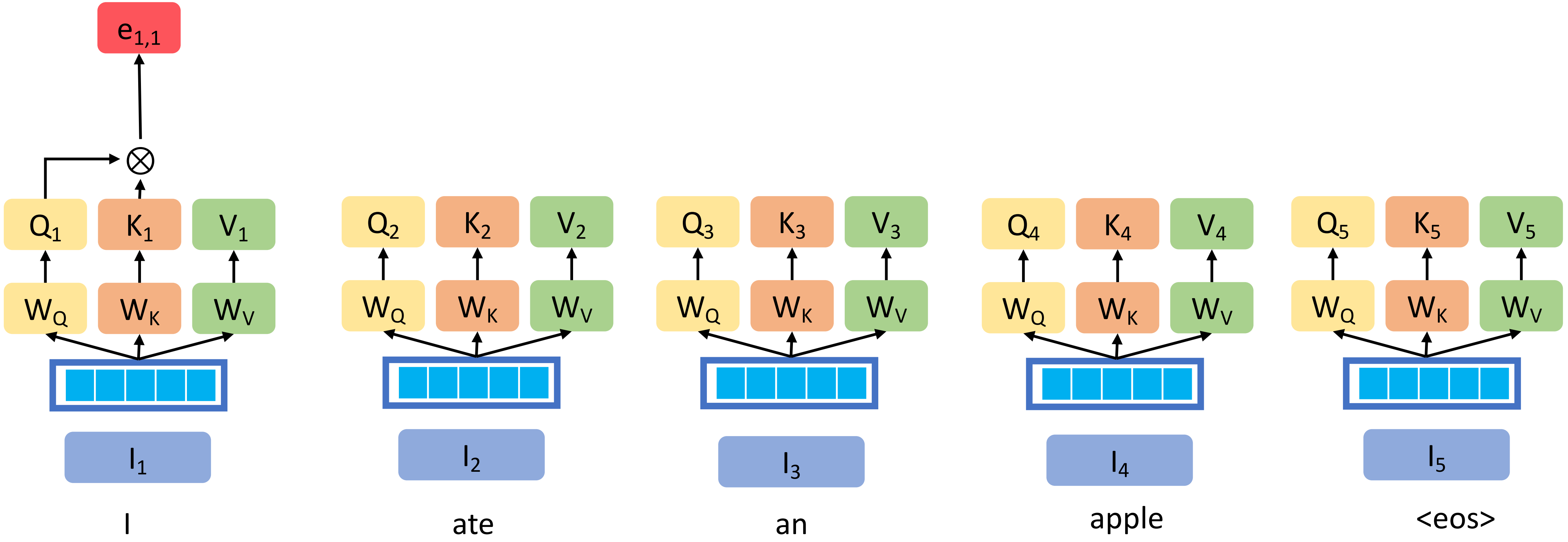
Dimensions across QKV have been dropped for brevity

Attention



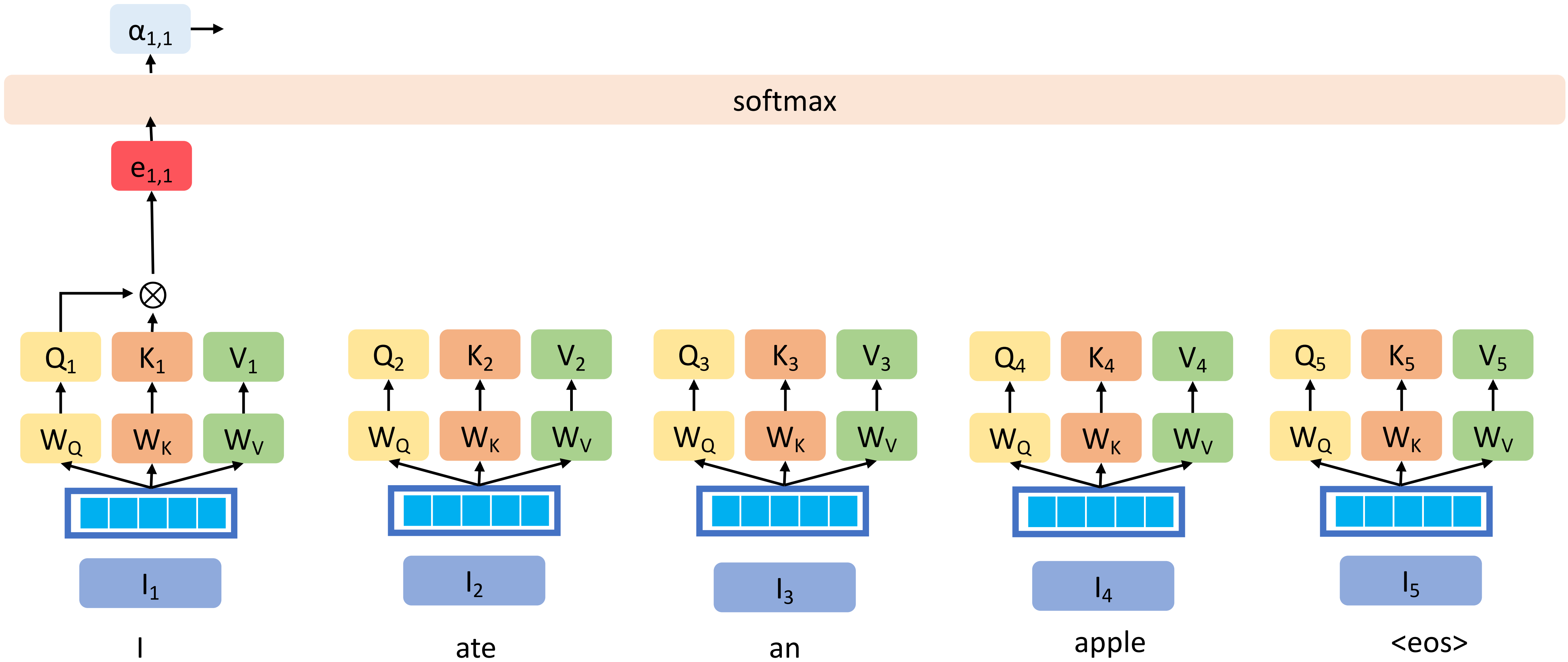
Attention

Dimensions across QKV have been dropped for brevity

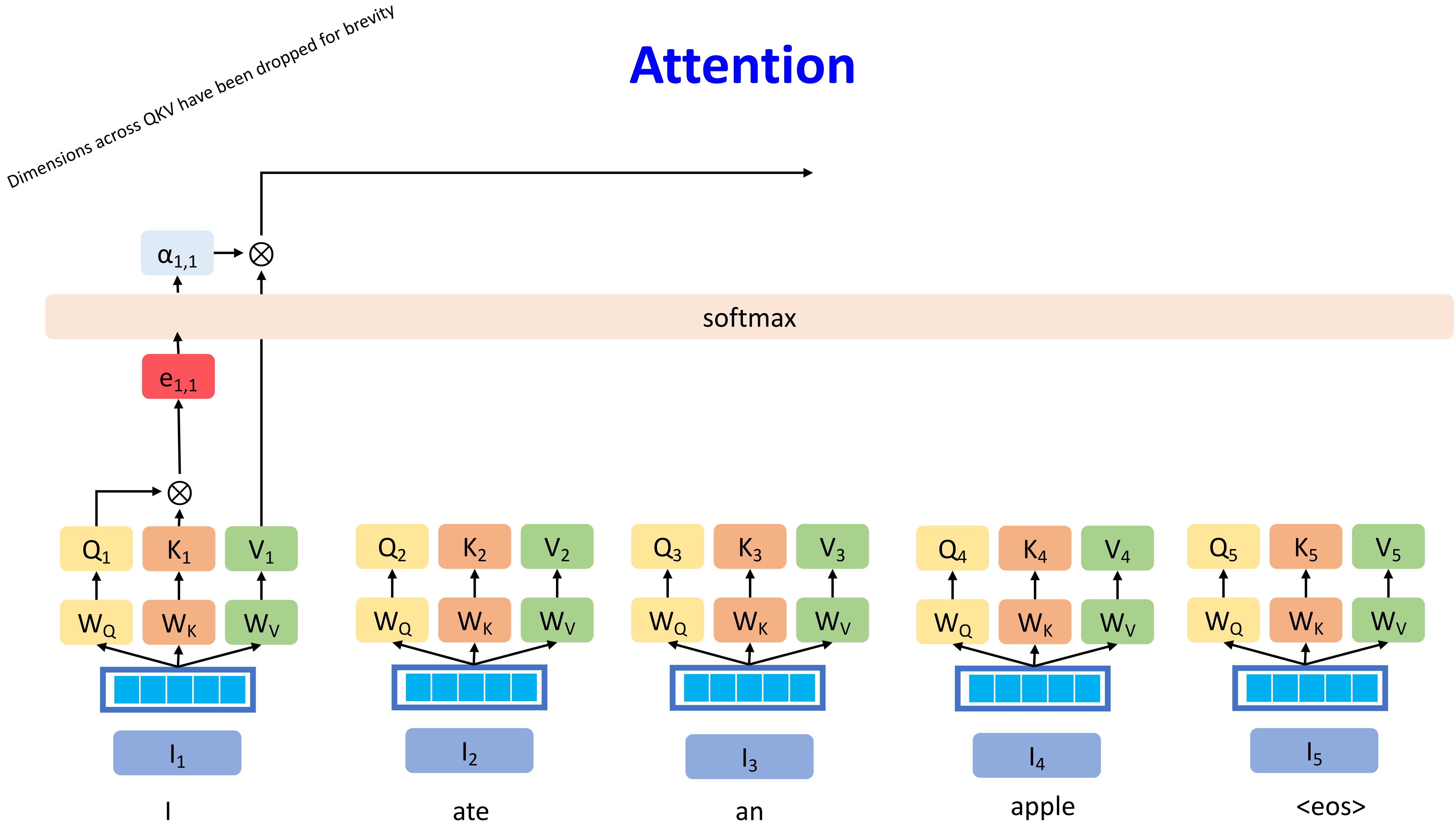


Attention

Dimensions across QKV have been dropped for brevity



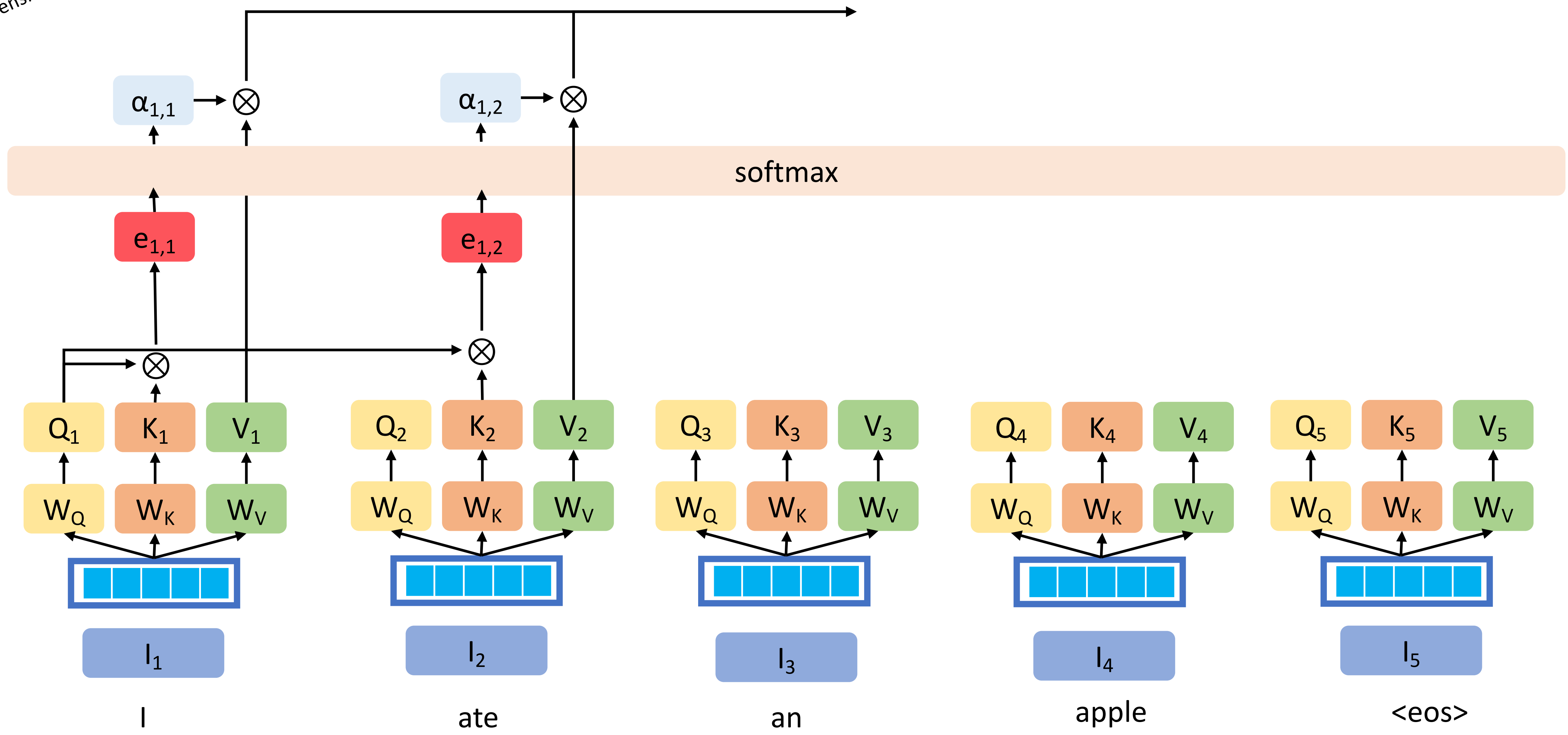
Attention



Dimensions across QKV have been dropped for brevity
α1,1
softmax
e1,1
Q1 K1 V1
Q2 K2 V2
Q3 K3 V3
Q4 K4 V4
Q5 K5 V5
WQ WK WV
I1
I2
I3
I4
I5
I
ate
an
apple
<eos>

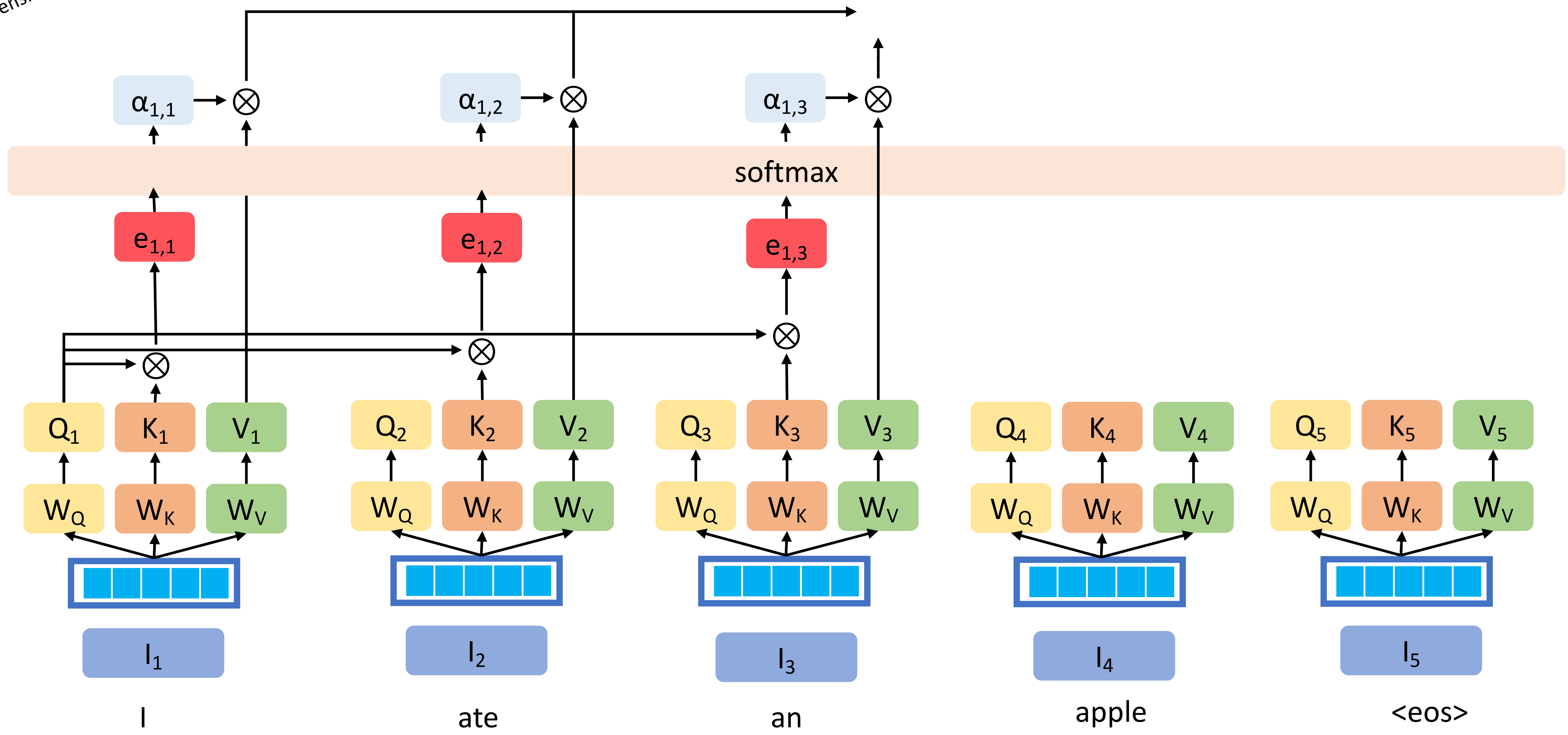
Attention

Dimensions across QKV have been dropped for brevity

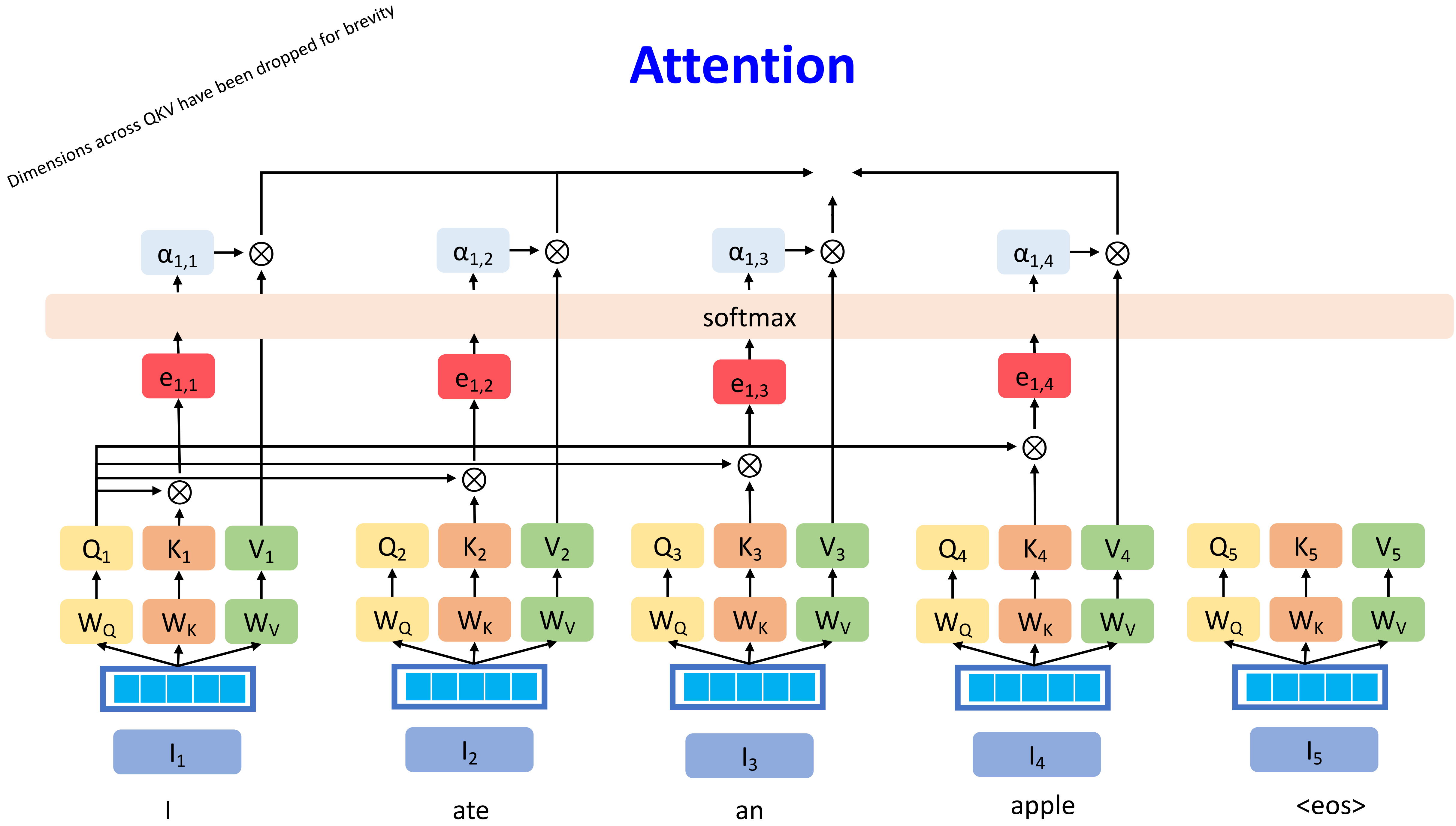


Attention

Dimensions across QKV have been dropped for brevity

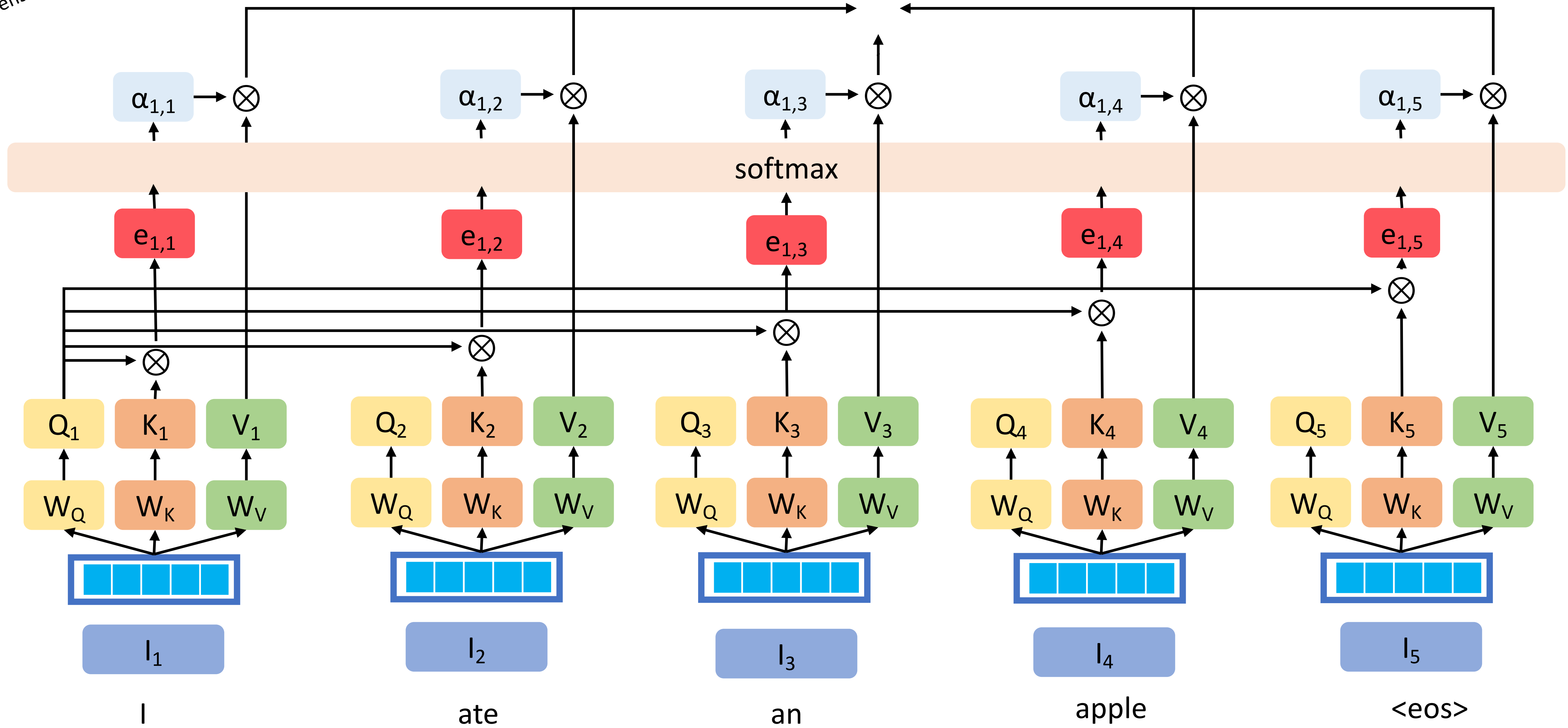


Attention

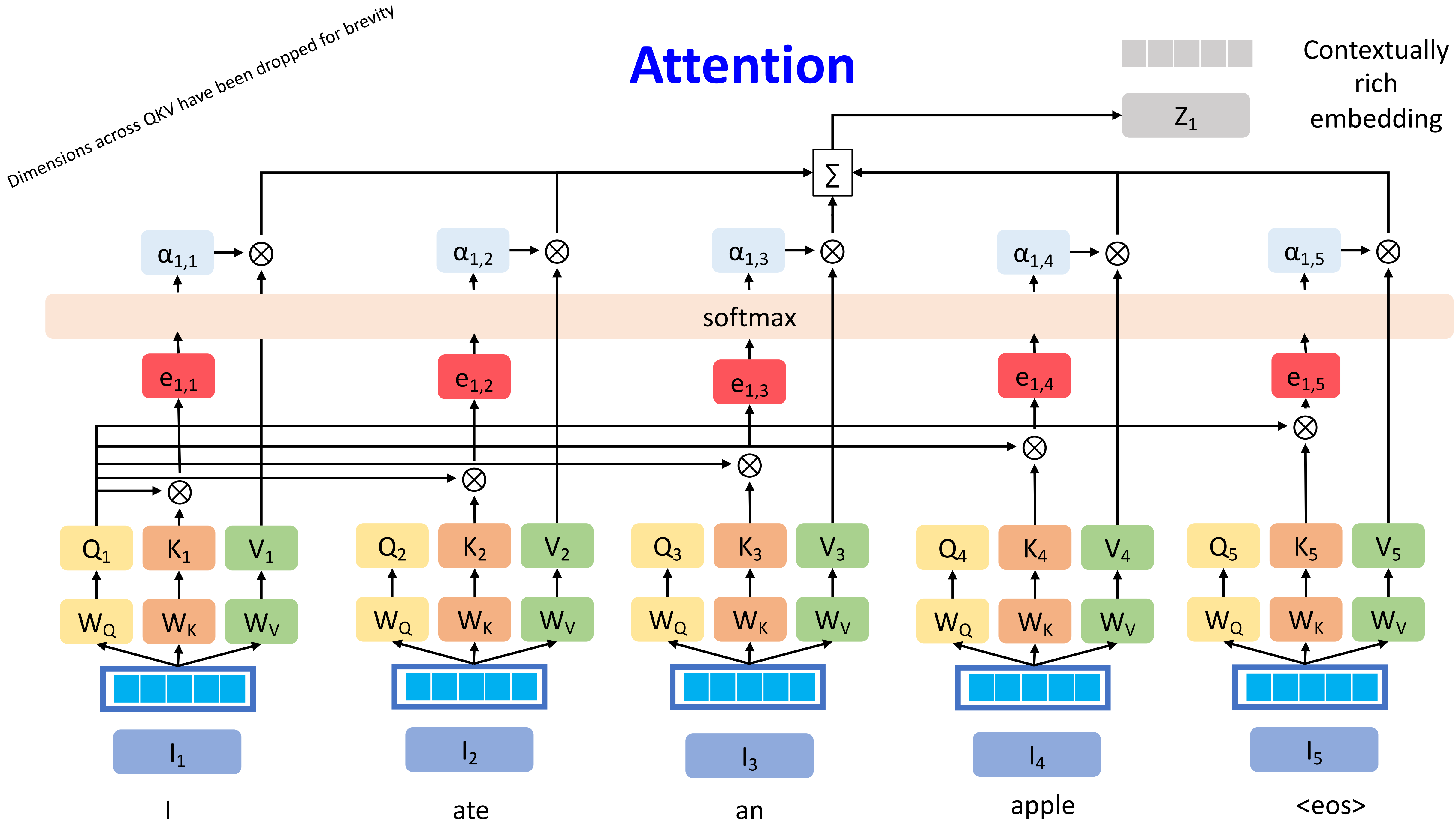

Dimensions across QKV have been dropped for brevity
softmax
$\alpha_{1,1}$
$\alpha_{1,2}$
$\alpha_{1,3}$
$\alpha_{1,4}$
$e_{1,1}$
$e_{1,2}$
$e_{1,3}$
$e_{1,4}$
Q_1 K_1 V_1
Q_2 K_2 V_2
Q_3 K_3 V_3
Q_4 K_4 V_4
Q_5 K_5 V_5
W_Q W_K W_V
I_1
I_2
I_3
I_4
I_5
I
ate
an
apple
<eos>

Attention

Dimensions across QKV have been dropped for brevity

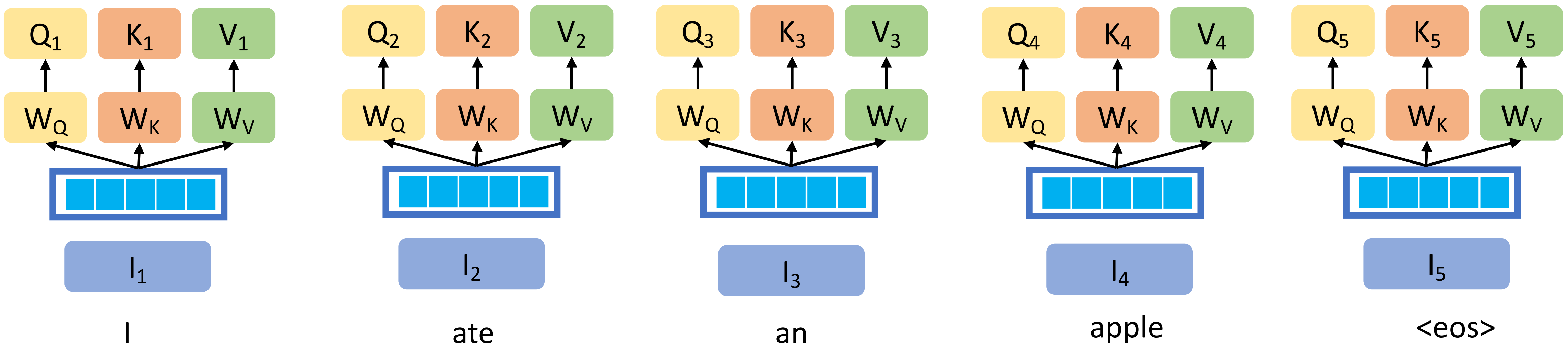


Attention

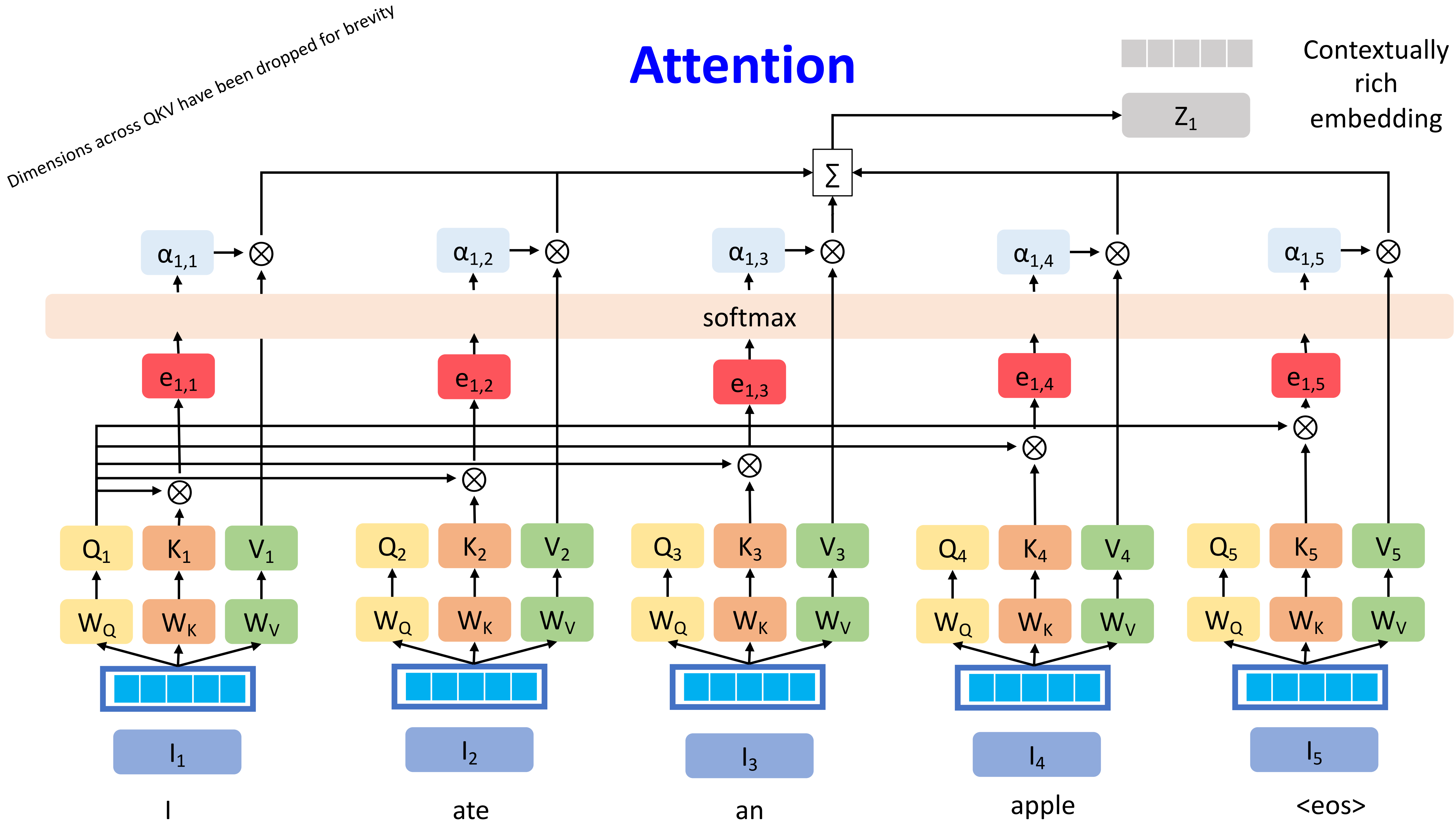

Dimensions across QKV have been dropped for brevity
Contextually rich embedding
Z1
∑
α1,1
α1,2
α1,3
α1,4
α1,5
softmax
e1,1
e1,2
e1,3
e1,4
e1,5
Q1 K1 V1
Q2 K2 V2
Q3 K3 V3
Q4 K4 V4
Q5 K5 V5
WQ WK WV
I1
I2
I3
I4
I5
I
ate
an
apple
<eos>

Dimensions across QKV have been dropped for brevity

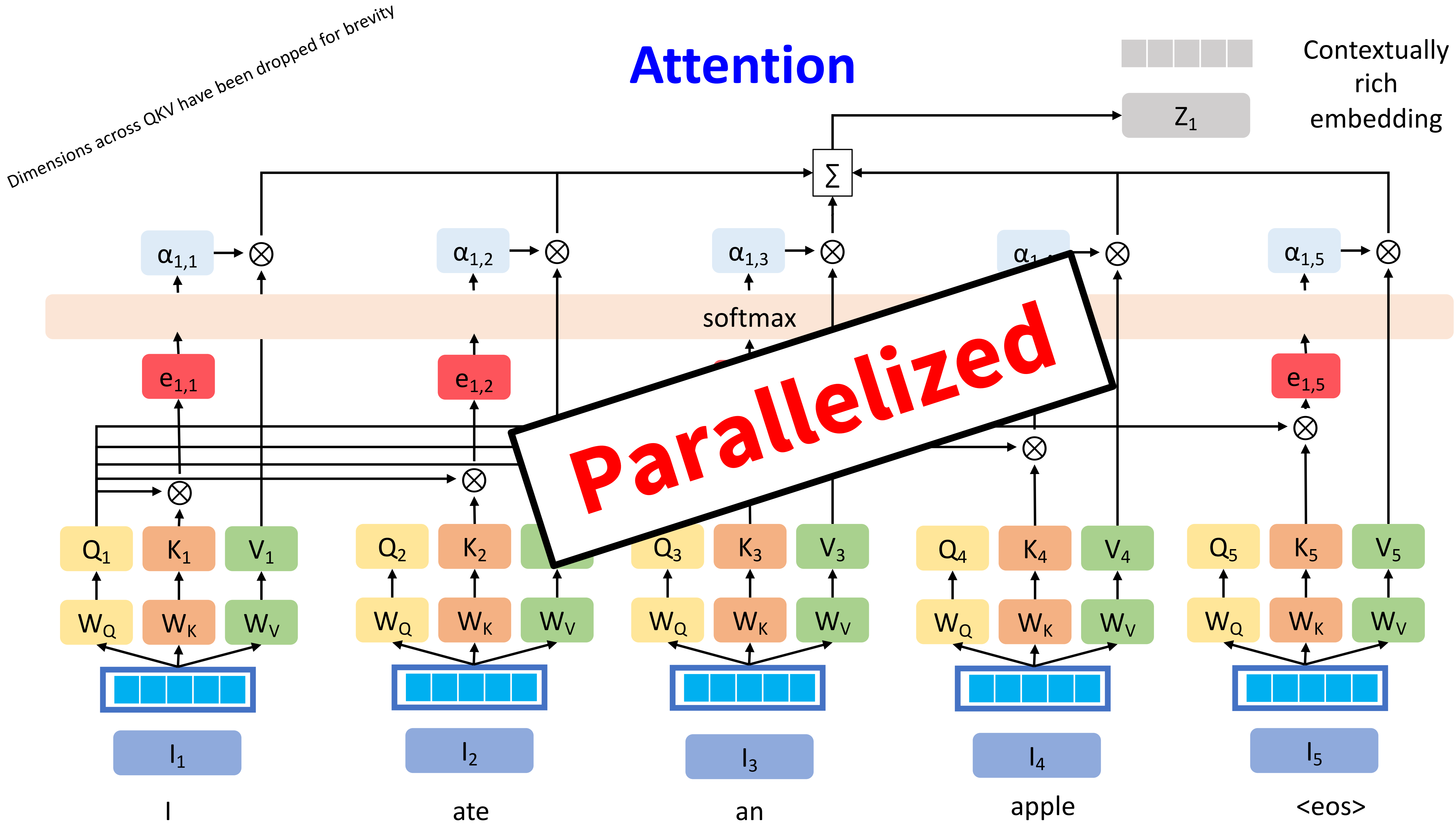
Attention



Attention


Dimensions across QKV have been dropped for brevity
Contextually rich embedding
Z_1
Σ
$\alpha_{1,1}$
$\alpha_{1,2}$
$\alpha_{1,3}$
$\alpha_{1,4}$
$\alpha_{1,5}$
softmax
$e_{1,1}$
$e_{1,2}$
$e_{1,3}$
$e_{1,4}$
$e_{1,5}$
Q_1 K_1 V_1
Q_2 K_2 V_2
Q_3 K_3 V_3
Q_4 K_4 V_4
Q_5 K_5 V_5
W_Q W_K W_V
I_1 I_2 I_3 I_4 I_5
I
ate
an
apple
<eos>

Attention



Dimensions across QKV have been dropped for brevity
Contextually rich embedding
Z_1
Σ
$\alpha_{1,1}$
$\alpha_{1,2}$
$\alpha_{1,3}$
$\alpha_{1,5}$
softmax
$e_{1,1}$
$e_{1,2}$
$e_{1,5}$
Parallelized
Q_1 K_1 V_1
Q_2 K_2
Q_3 K_3 V_3
Q_4 K_4 V_4
Q_5 K_5 V_5
W_Q W_K W_V
I_1 I_2 I_3 I_4 I_5
I
ate
an
apple
<eos>

Poll 1 @1296

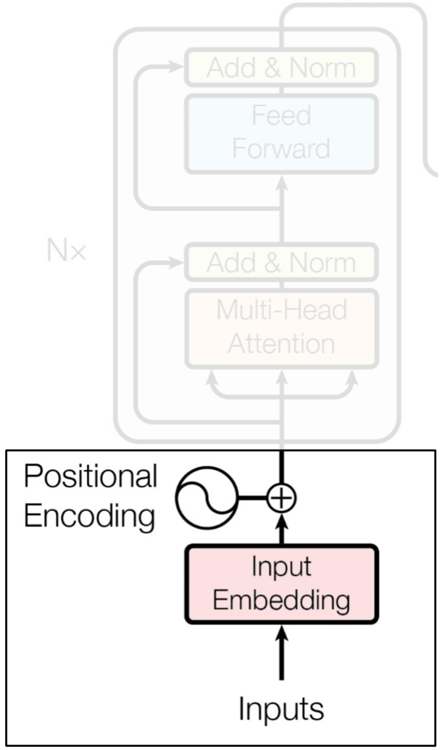
Which of the following are true about attention? (Select all that apply)

a. To calculate attention weights for input $\mathbf{I_2}$, you would use key $\mathbf{k_2}$, and all queries

b. To calculate attention weights for input $\mathbf{I_2}$, you would use query $\mathbf{q_2}$, and all keys

c. We scale the QK^T product to bring attention weights in the range of [0,1]

d. We scale the QK^T product to allow for numerical stability



Poll 1 @1296

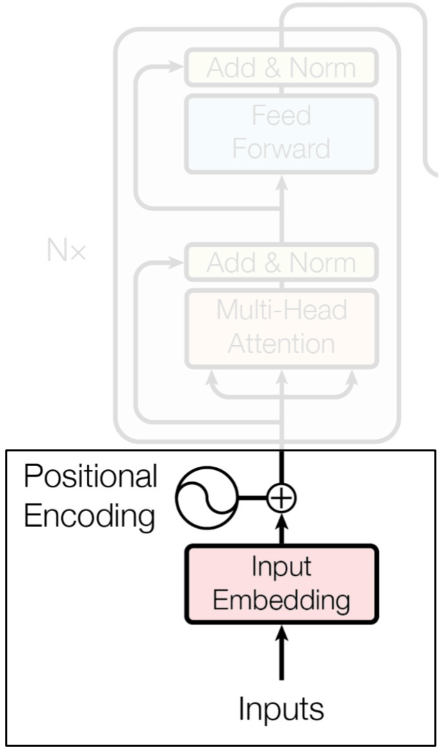
Which of the following are true about attention? (Select all that apply)

a. To calculate attention weights for input **I_2**, you would use key **k_2**, and all queries

b. To calculate attention weights for input I_2, you would use query q_2, and all keys

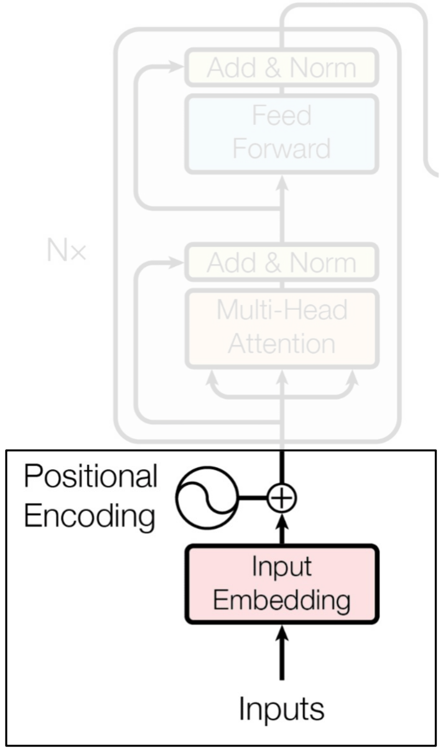
c. We scale the QK^T product to bring attention weights in the range of [0,1]

d. We scale the QK^T product to allow for numerical stability



Positional Encoding

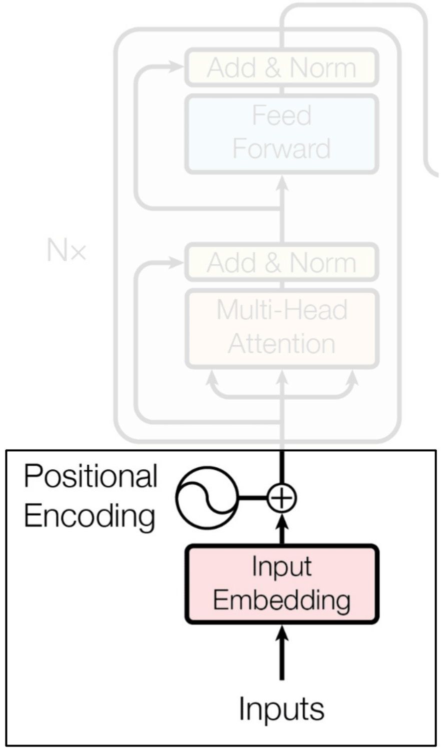
I | ate | an | apple | <eos>



Positional Encoding

I ate an apple <eos>

apple ate an I <eos>

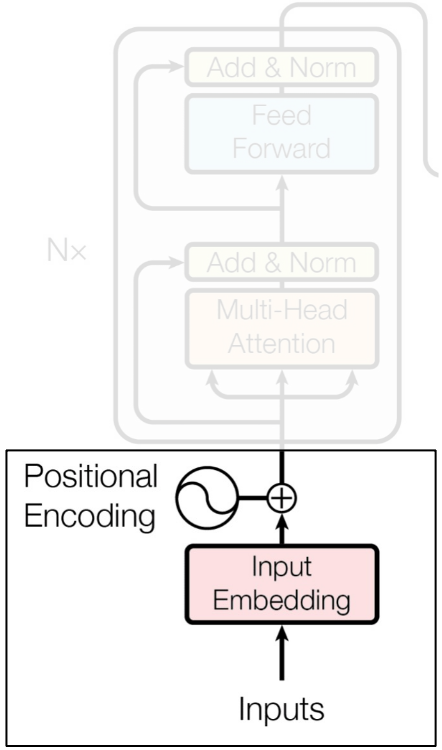


Positional Encoding

Positional Encoding

Requirements for Positional Encodings

- Some representation of time ? (like **seq2seq** ?)
- Should be unique for each position – not cyclic

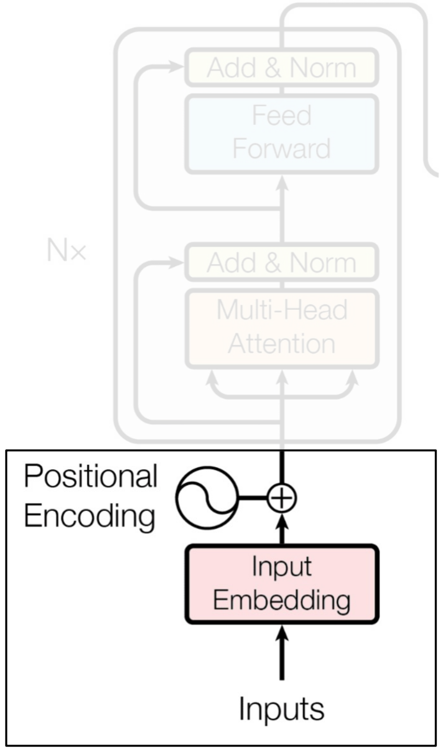


Positional Encoding

Positional Encoding

Requirements for Positional Encodings

- Some representation of time ? (like **seq2seq** ?)
- Should be unique for each position – not cyclic



Possible Candidates :

$$P_{t+1} = P_t + \Delta c$$

$$P_{t+1} = e^{P_{t_\Delta} c}$$

$$P_{t+1} = P_t^{\,t\Delta c}$$

Positional Encoding

Positional Encoding

Requirements for Positional Encodings

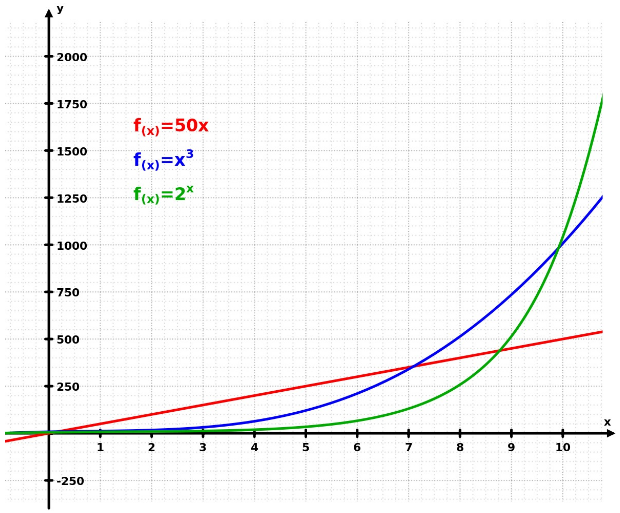
- Some representation of time ? (like **seq2seq** ?)
- Should be unique for each position – not cyclic

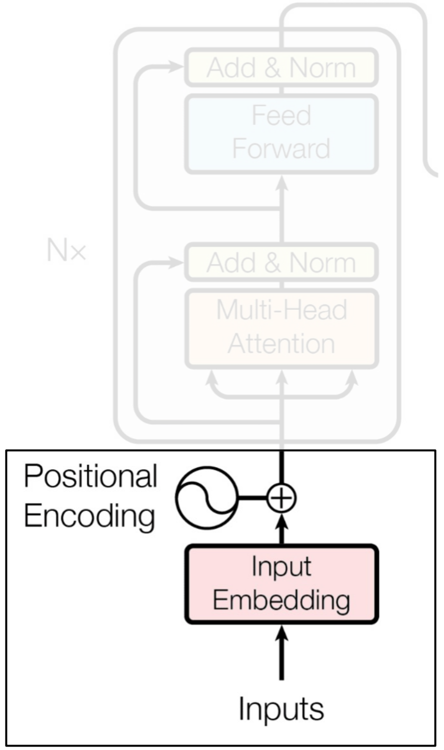
Possible Candidates :

$$P_{t+1} = P_t + \Delta c$$

$$P_{t+1} = e^{P_{t_{\Delta}} c}$$

$$P_{t+1} = P_t^{\,t\Delta c}$$





Positional Encoding

Positional Encoding

Requirements for Positional Encodings

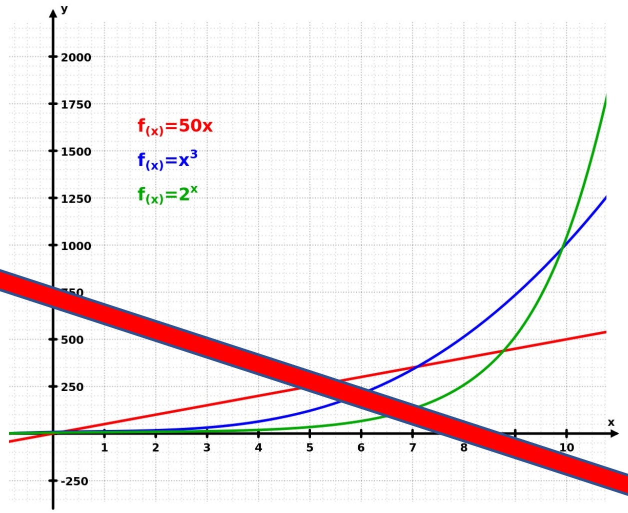
- Some representation of time ? (like **seq2seq** ?)
- Should be unique for each position – not cyclic
- **Bounded**

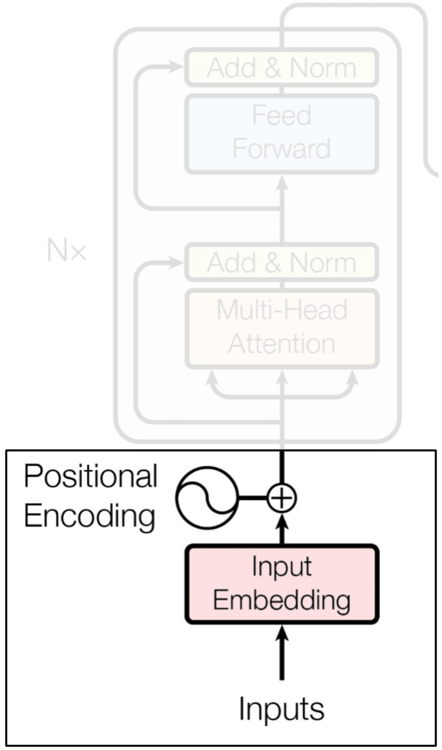
Possible Candidates :

$$P_{t+1} = P_t + \Delta c$$

$$P_{t+1} = e^{P_t \Delta c}$$

$$P_{t+1} = P_t^{\,t\Delta c}$$





Positional Encoding

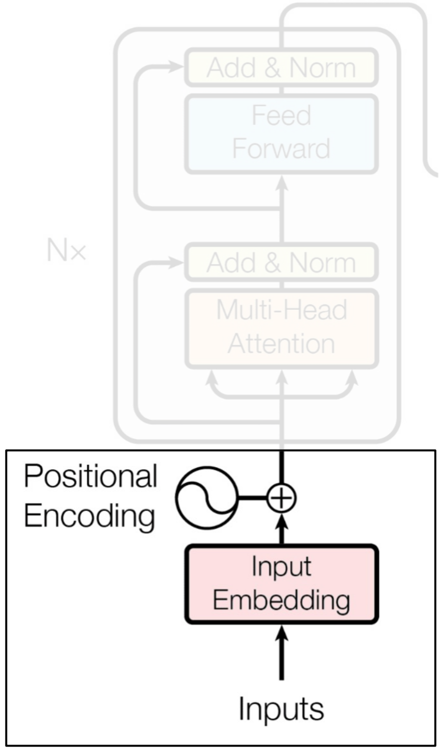
Positional Encoding

Requirements for Positional Encodings

- Some representation of time ? (like **seq2seq** ?)
- Should be unique for each position – not cyclic
- **Bounded**

Possible Candidates :

$$P(t + t') = M^{t'} \times P(t)$$



Positional Encoding

Positional Encoding

Requirements for Positional Encodings

- Some representation of time ? (like **seq2seq** ?)
- Should be unique for each position – not cyclic
- **Bounded**

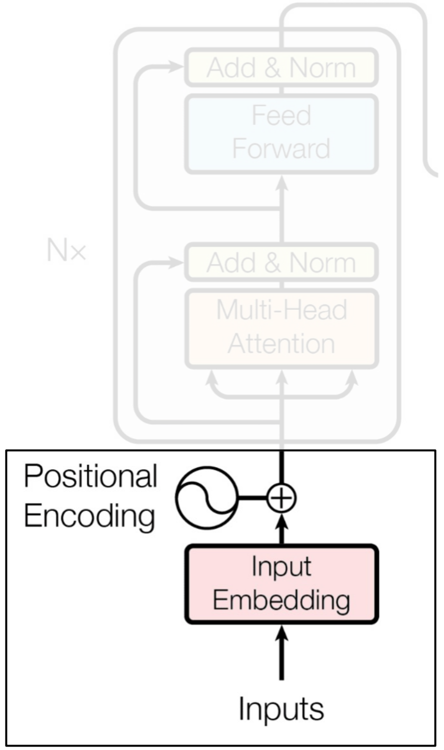
Possible Candidates :

$$P(t + t') = M^{t'} \times P(t)$$

M ?

1. **Should be a unitary matrix**
2. **Magnitudes of eigen value should be 1 -> norm preserving**

Positional Encoding



Positional Encoding

Requirements for Positional Encodings

- Some representation of time ? (like **seq2seq** ?)
- Should be unique for each position – not cyclic
- **Bounded**

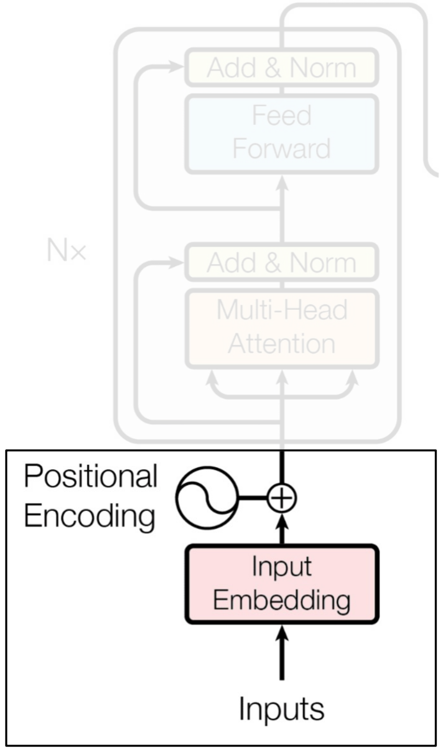
Possible Candidates :

$$\mathbf{P(t + t') = M^{t'} \times P(t)}$$

M

1. **The matrix can be learnt**
2. **Produces unique rotated embeddings each time**

Positional Encoding



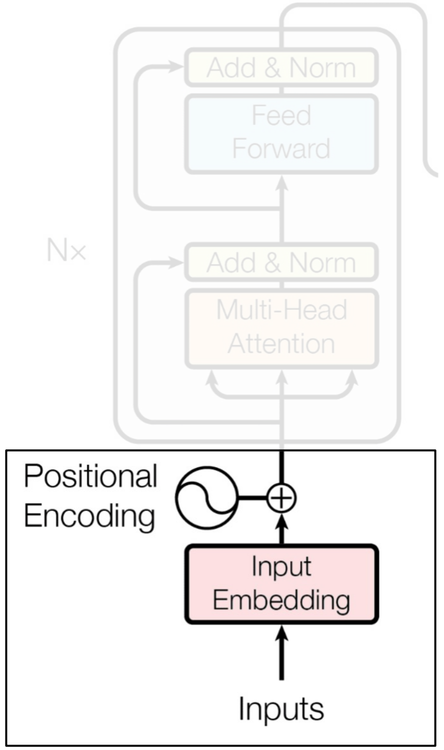
Rotary Positional Embedding

RoFormer: Enhanced Transformer with Rotary Position Embedding

$$f_{\{q,k\}}(\boldsymbol{x}_m, m) = \begin{pmatrix} \cos m\theta & -\sin m\theta \\ \sin m\theta & \cos m\theta \end{pmatrix} \begin{pmatrix} W^{(11)}_{\{q,k\}} & W^{(12)}_{\{q,k\}} \\ W^{(21)}_{\{q,k\}} & W^{(22)}_{\{q,k\}} \end{pmatrix} \begin{pmatrix} x^{(1)}_m \\ x^{(2)}_m \end{pmatrix}$$

Table 2: Comparing RoFormer and BERT by fine tuning on downstream GLEU tasks.

Model	MRPC	SST-2	QNLI	STS-B	QQP	MNLI(m/mm)
BERTDevlin et al. [2019]	88.9	93.5	90.5	85.8	71.2	84.6/83.4
RoFormer	**89.5**	90.7	88.0	**87.0**	**86.4**	80.2/79.8



REF: Rotary Positional Embeddings

Positional Encoding

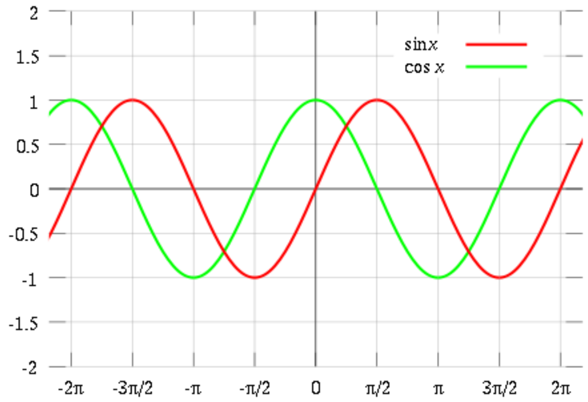
Requirements for Positional Encodings

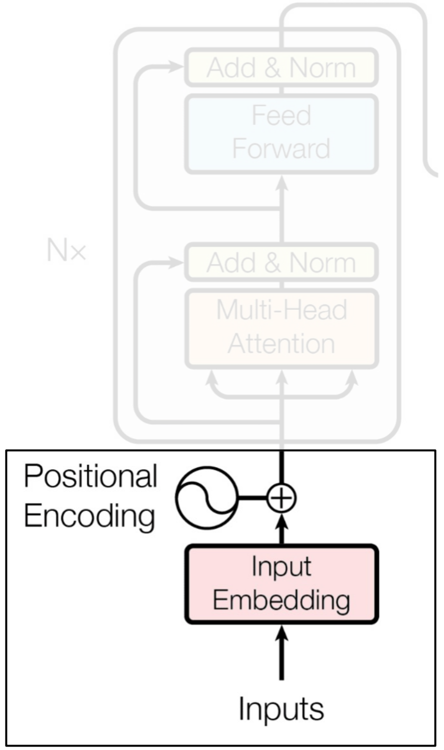
- Some representation of time ? (like **seq2seq** ?)
- Should be unique for each position – **not cyclic**
- Bounded

Actual Candidates :

*sine(**g(t)**)*

*cosine(**g(t)**)*





Positional Encoding

Positional Encoding

Requirements for g(t)

- *Must have same dimensions as input embeddings*
- *Must produce overall unique encodings*

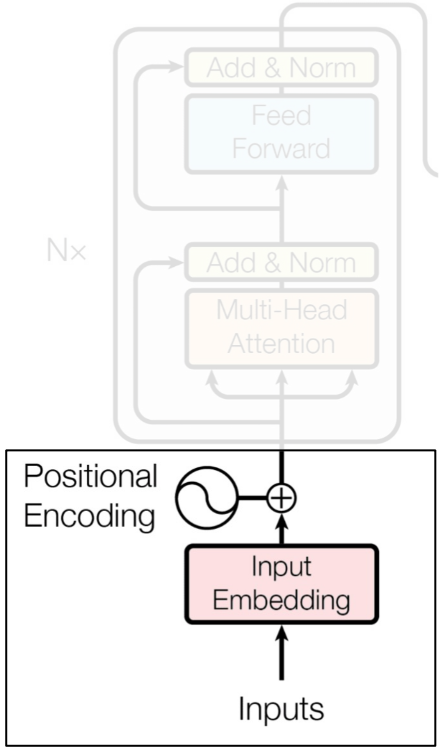
pos -> idx of the token in input sentence

i -> i^{th} dimension out of d

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{model}})$$

$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{model}})$$

Positional Encoding



Positional Encoding

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

Requirements for g(t)

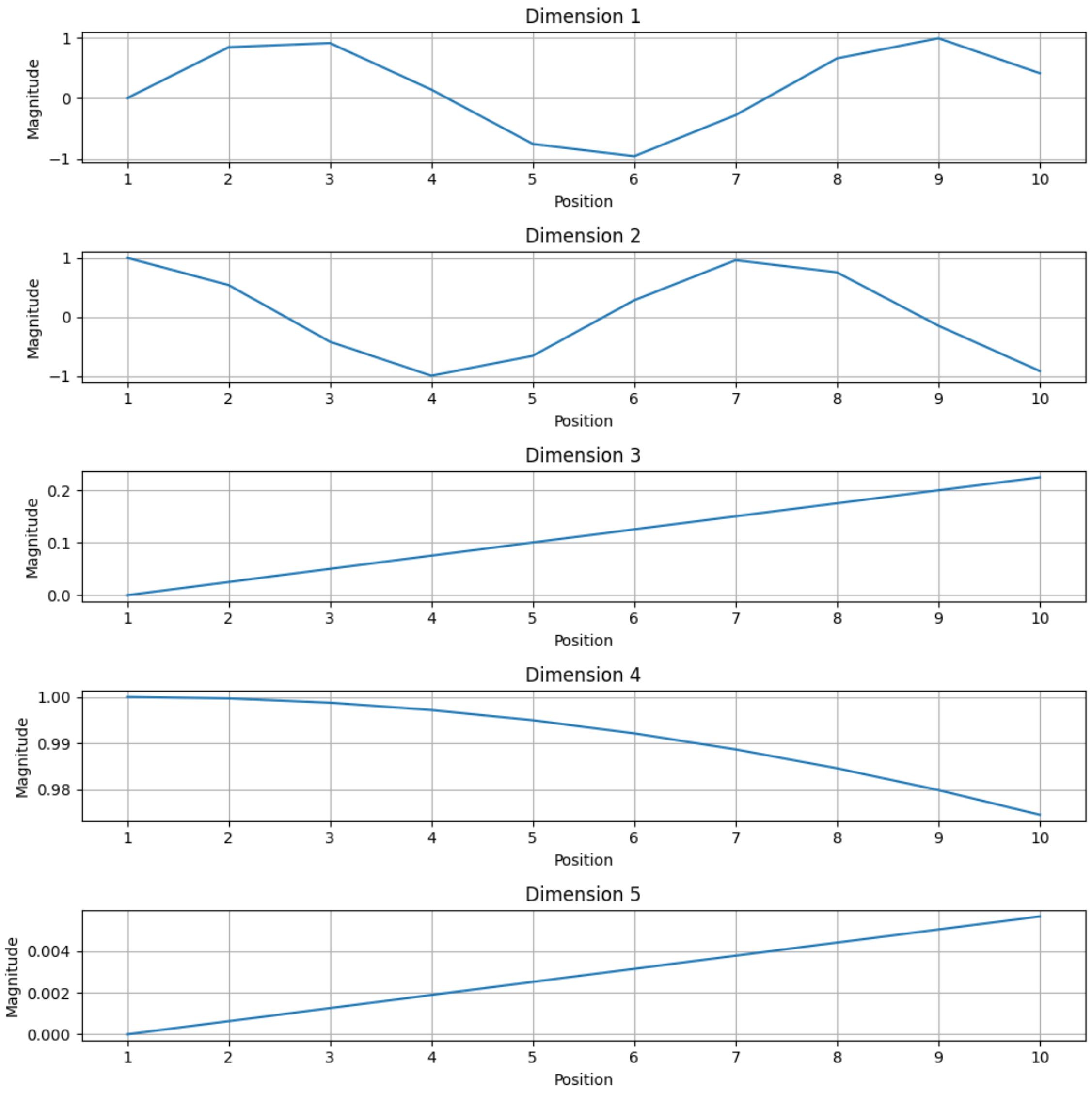
- *Must have same dimensions as input embeddings*
- *Must produce overall unique encodings*

pos -> idx of the token in input sentence

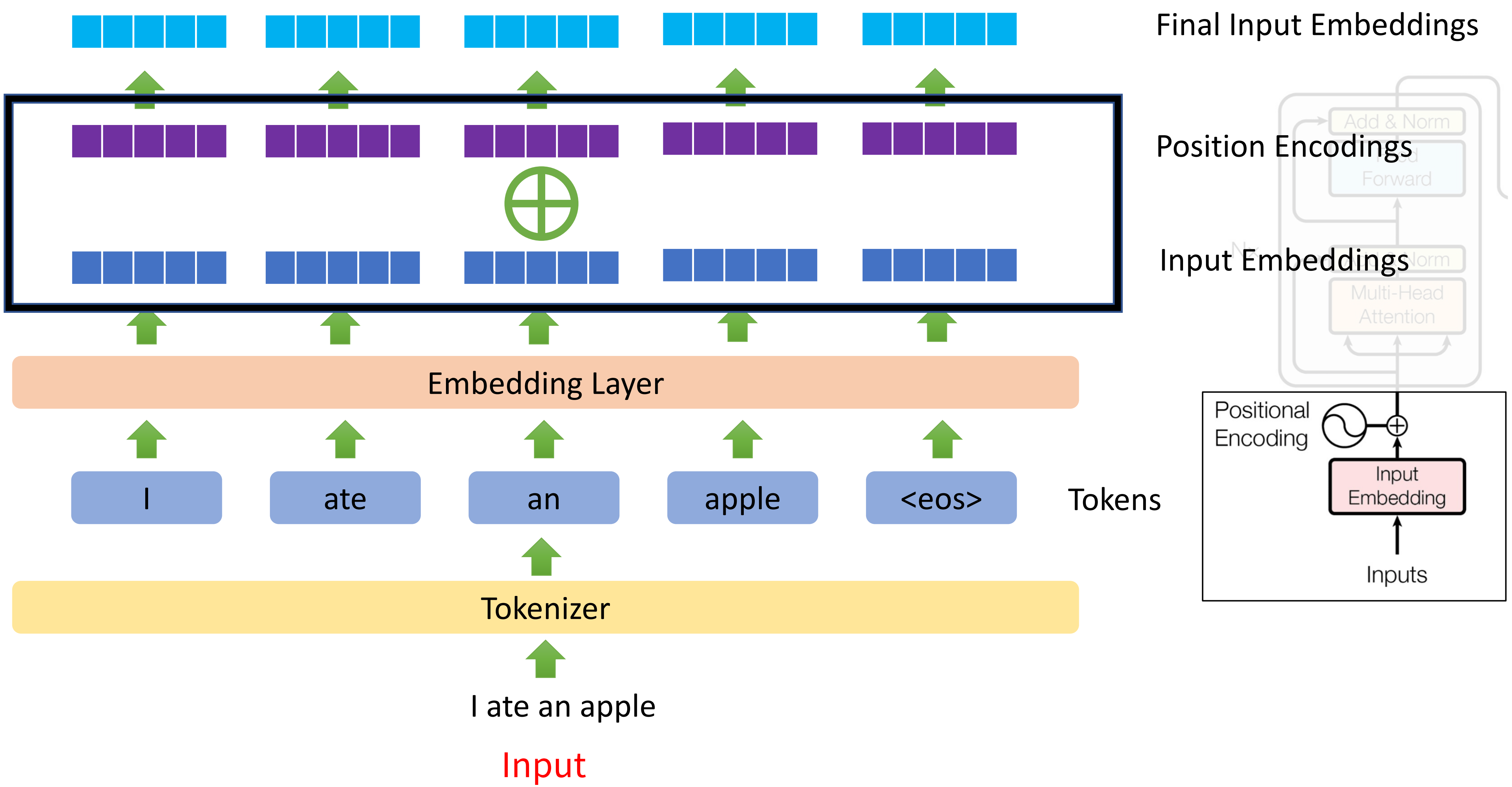
i -> i^{th} dimension out of d

Positional Encoding:

	0	1	2	3	4
Dim 1	0.000	0.841	0.909	0.141	-0.757
Dim 2	1.000	0.540	-0.416	-0.990	-0.654
Dim 3	0.000	0.025	0.050	0.075	0.100
Dim 4	1.000	1.000	0.999	0.997	0.995
Dim 5	0.000	0.001	0.001	0.002	0.003

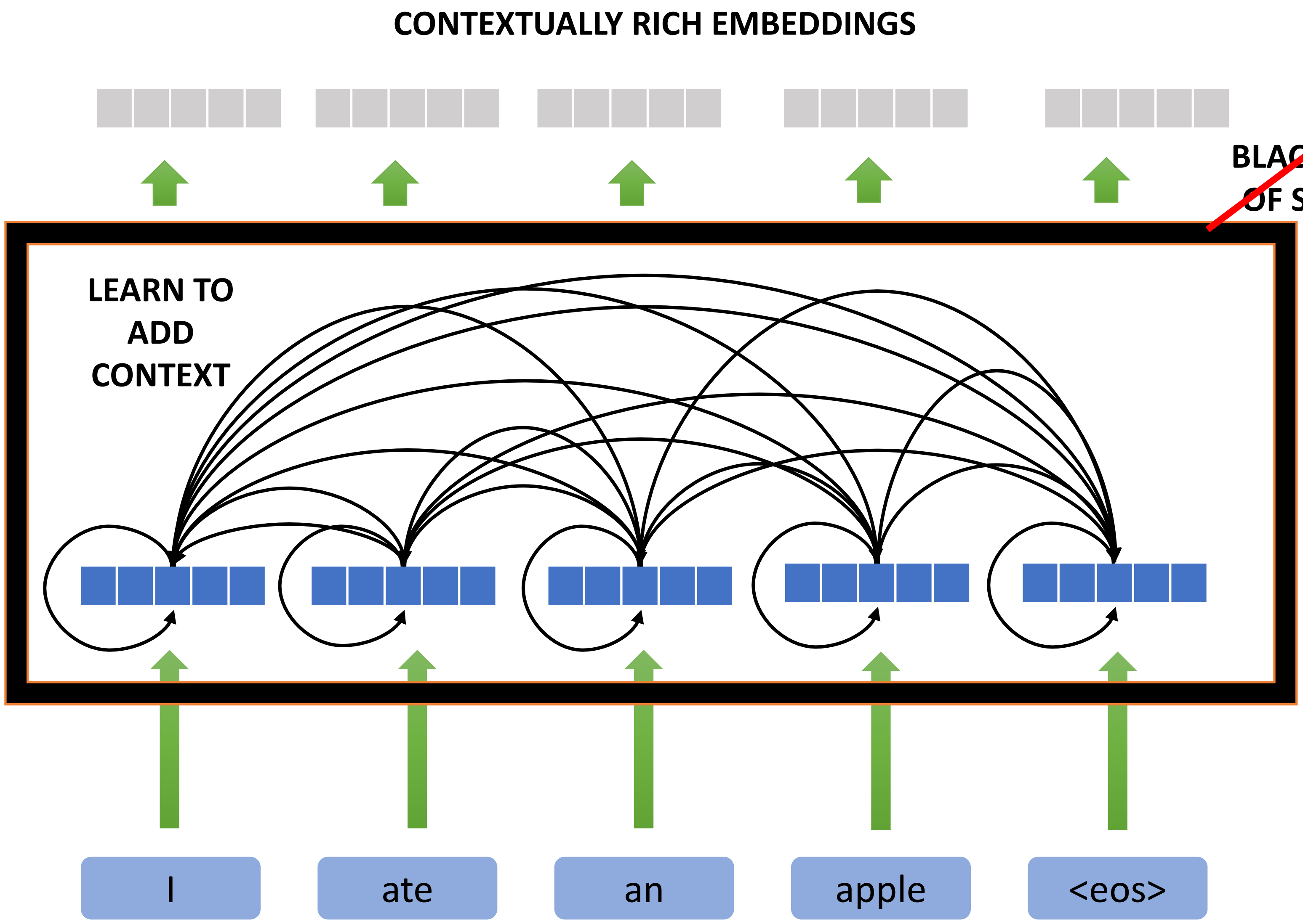


Positional Encoding

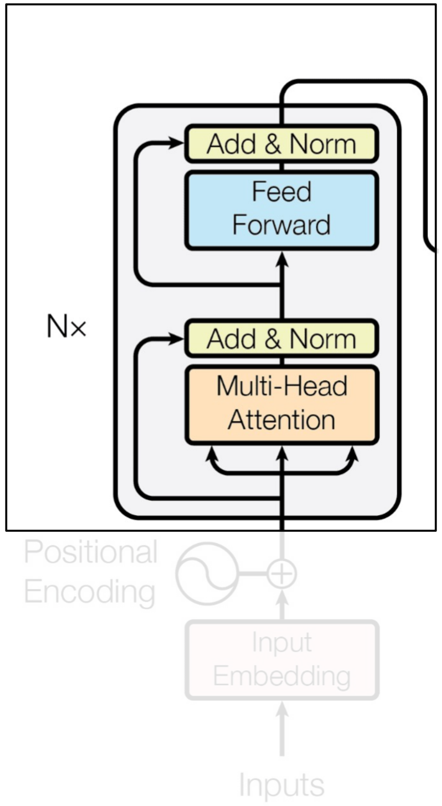


Encoder

$\alpha_{[\,i\,j\,]}$ Σ



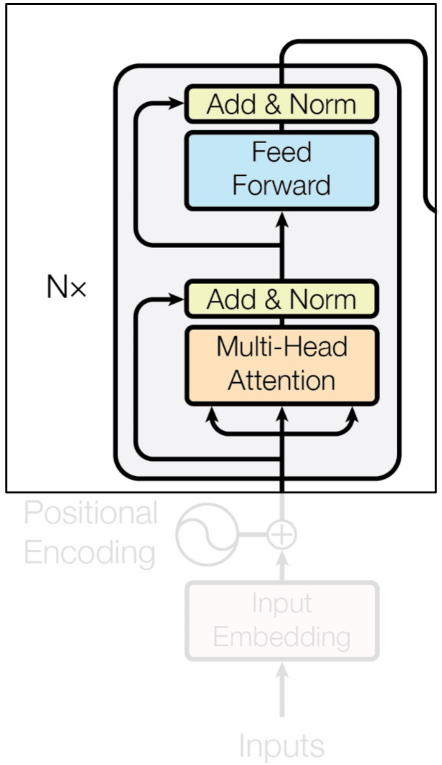
BLACK BOX OF SORTS



Self Attention

From lecture 18:

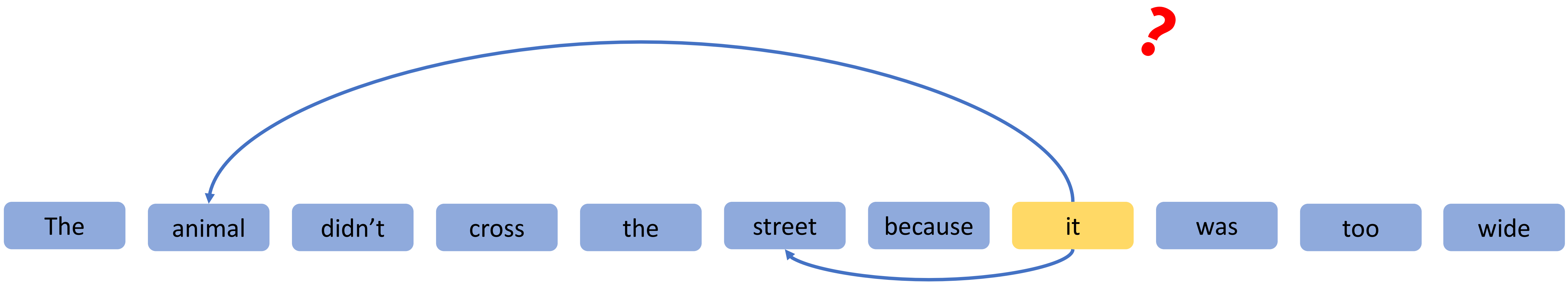
$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$



Self Attention

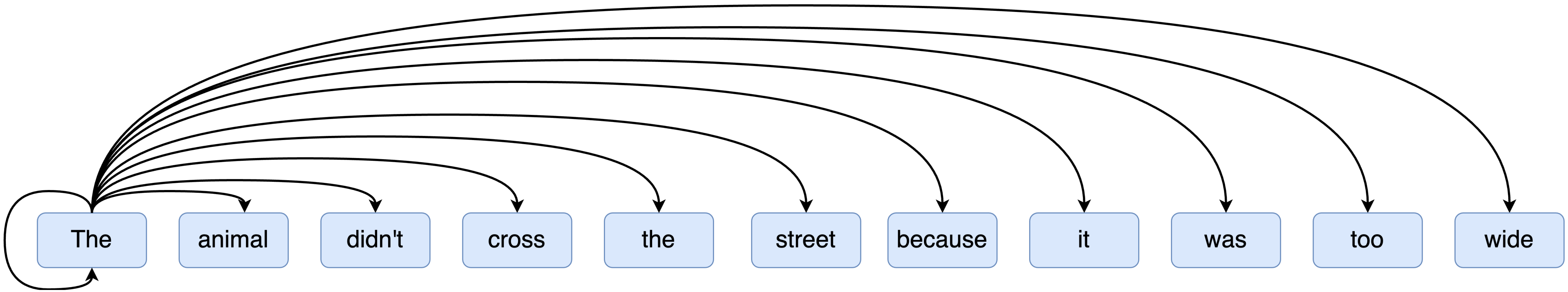
The animal didn’t cross the street because it was too wide

Self Attention

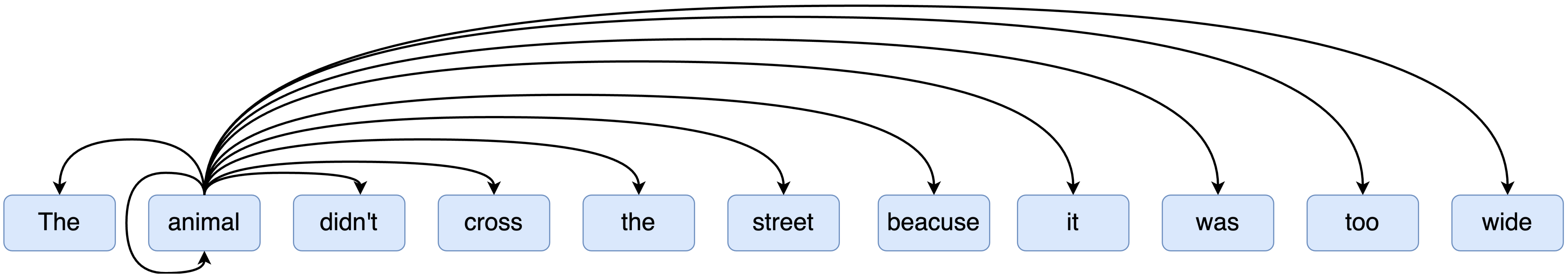


coreference resolution ?

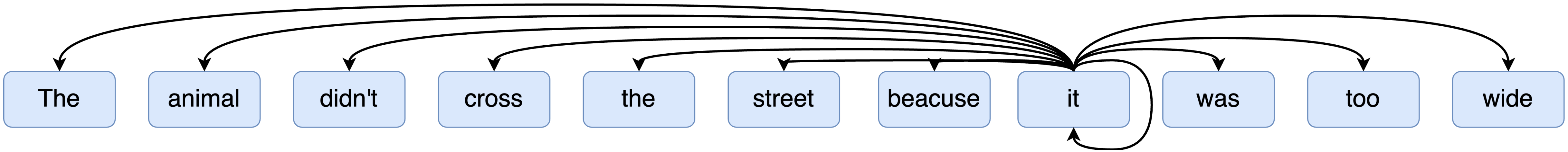
Self Attention



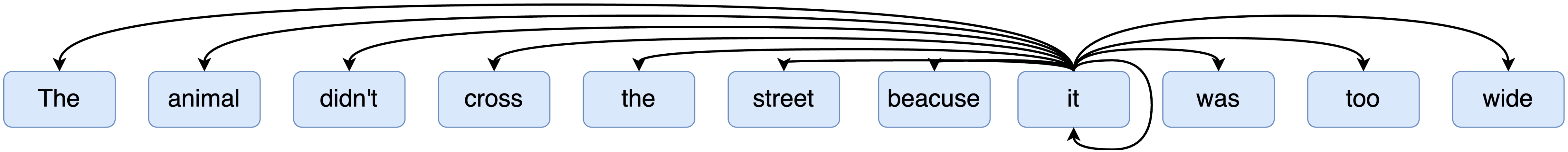
Self Attention



Self Attention



Self Attention



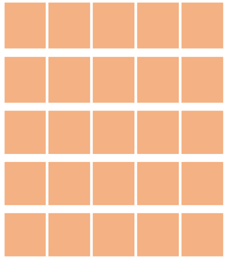
SELF

Query Inputs = Key Inputs = Value Inputs

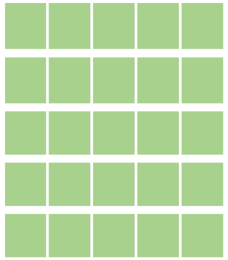
Self Attention

$R^{d_{model} \times d_{model}}$

W_Q

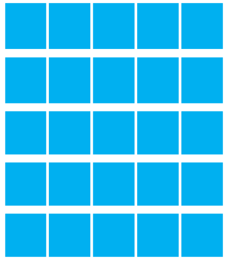


W_K

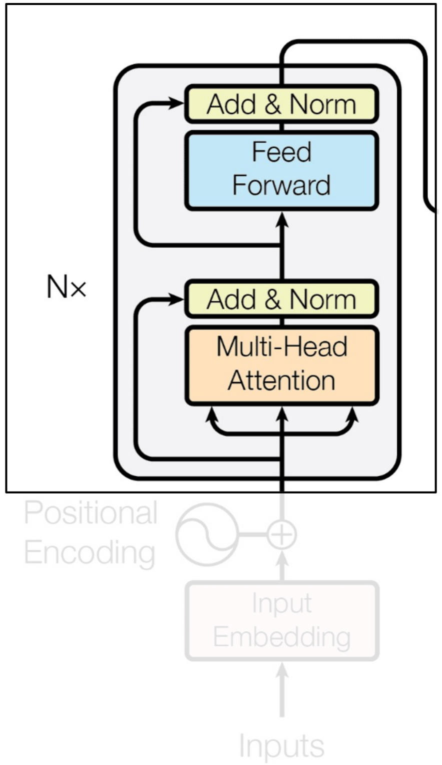


W_V

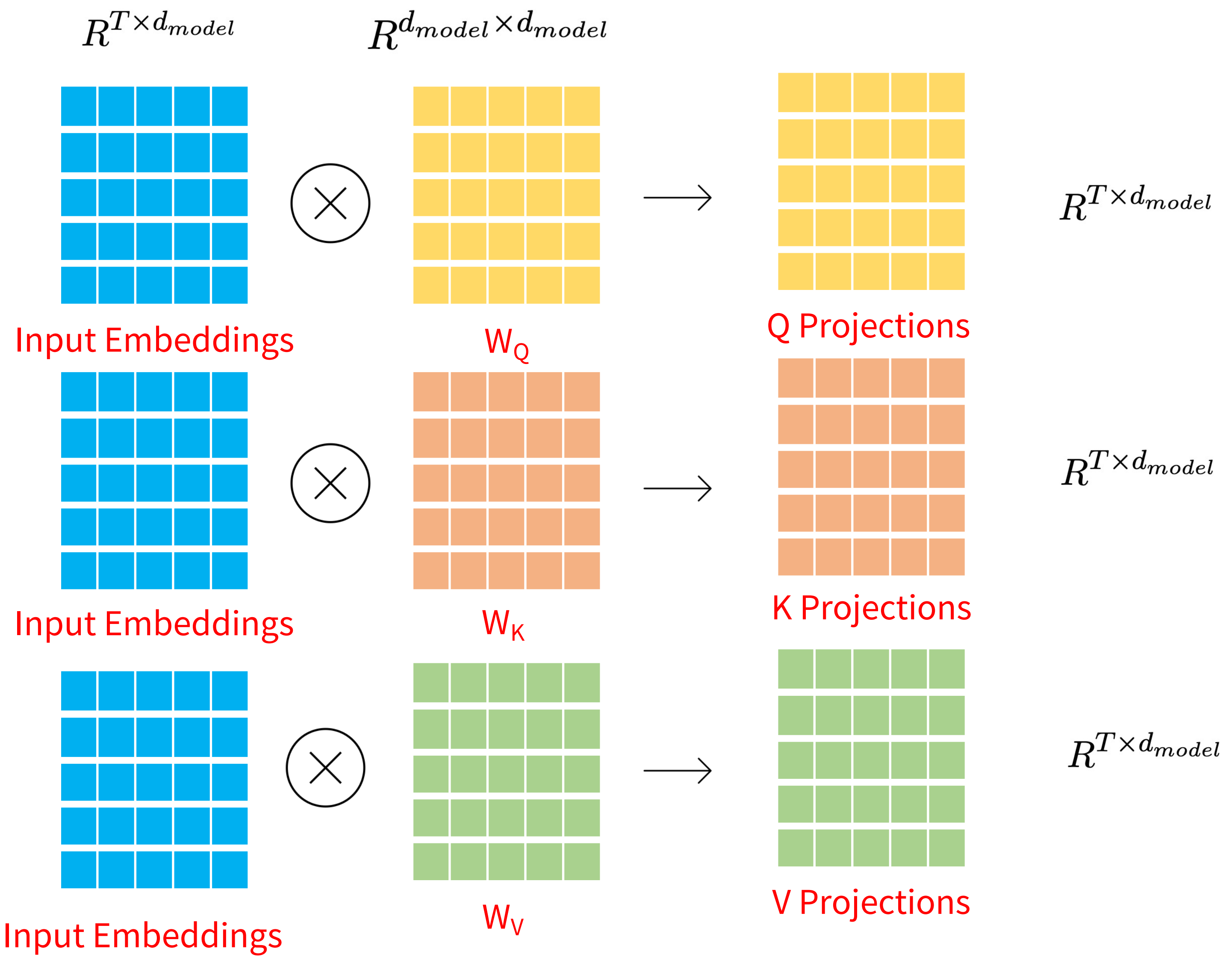
$R^{T \times d_{model}}$

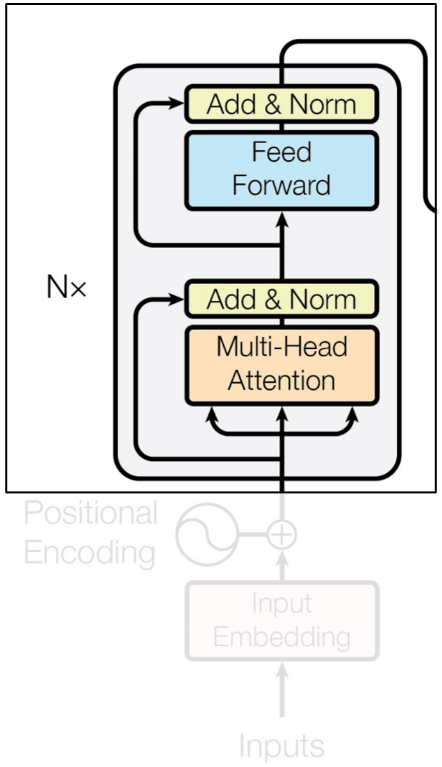


Input Embeddings



Self Attention



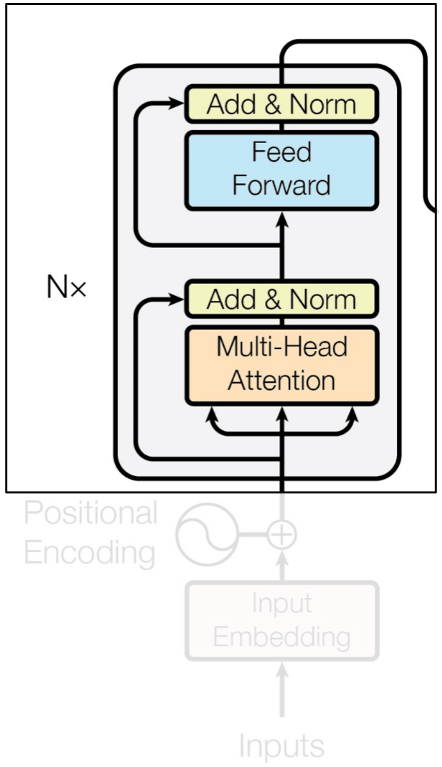


Self Attention

$R^{T \times d_{model}}$ $\quad$ $R^{d_{model} \times T}$

$$\text{softmax}\left(\frac{Q_{Projection} \otimes \left(K_{Projection}\right)^{T}}{\sqrt{d_{model}}} \right)$$

$R^{T \times T}$

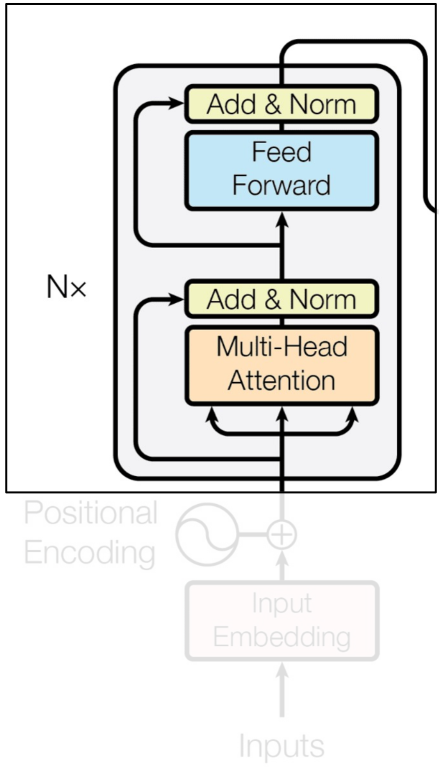


Self Attention

$\Theta(T^2 \times d_{model})$

$$\text{softmax}\left(\frac{\underset{R^{T \times d_{model}}}{Q_{Projection}} \otimes \underset{R^{d_{model} \times T}}{\left(K_{Projection}\right)^{T}}}{\sqrt{d_{model}}} \right)$$

$R^{T \times T}$

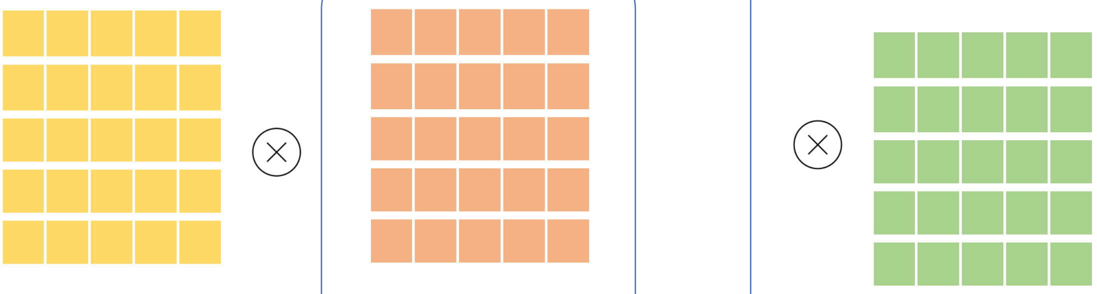


Self Attention

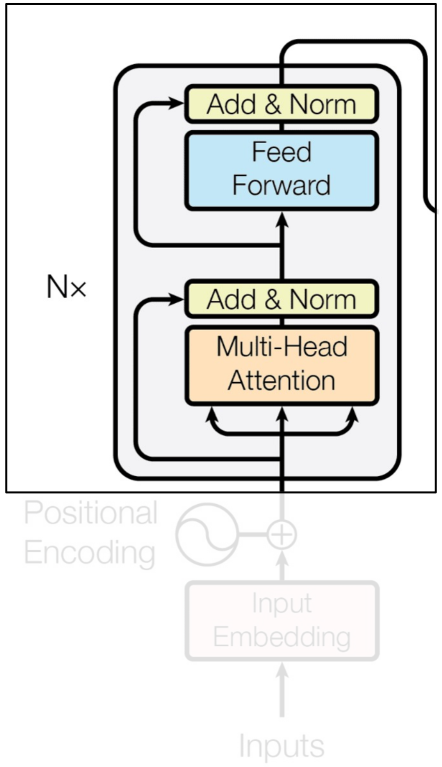
$\Theta(T^2 \times d_{model})$

softmax $\left(\dfrac{Q_{Projection} \otimes \left(K_{Projection}\right)^T}{\sqrt{d_{model}}} \right) \otimes V_{Projection}$

$R^{T\times d_{model}}$ $R^{d_{model}\times T}$ $R^{T\times d_{model}}$



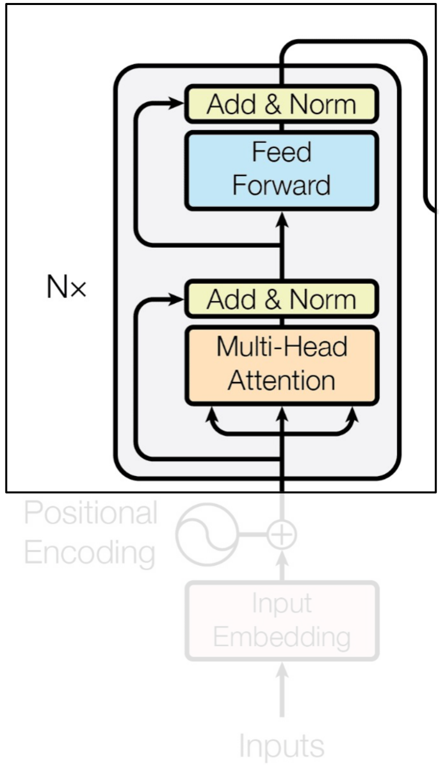
$R^{T\times T}$



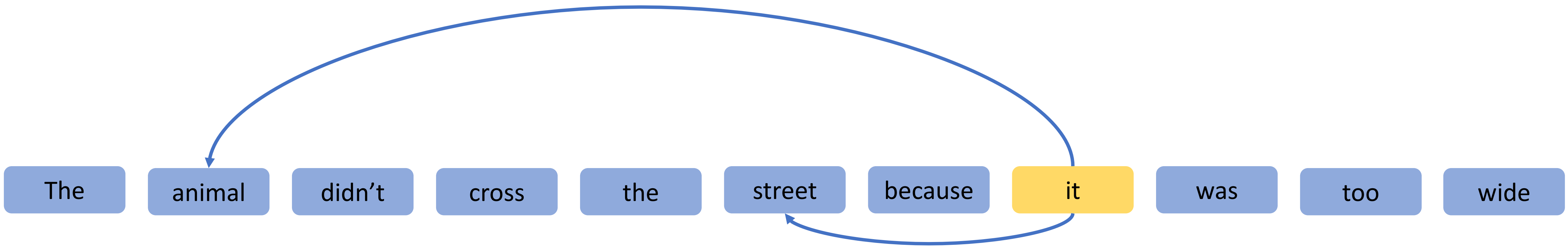
Self Attention

$R^{T \times d_{model}}$

Attention: Z

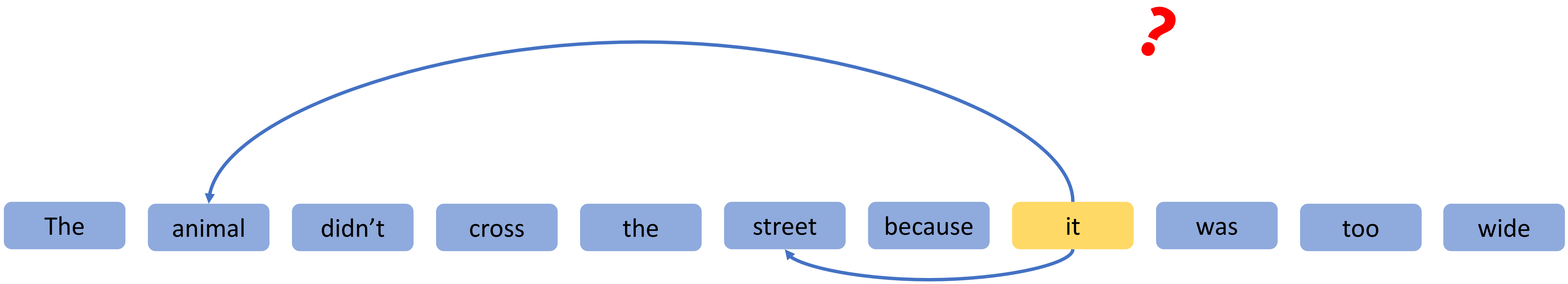


Self Attention



coreference resolution ✓

Self Attention



Sentence boundaries ?

coreference resolution ✓

Context ?

Semantic relationships ?

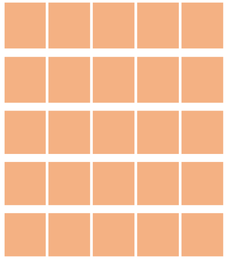
Part of Speech ?

Comparisons ?

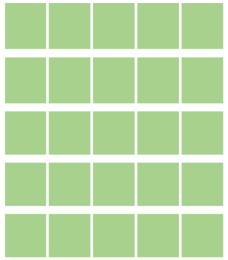
Self Attention

$$R^{d_{model} \times d_{model}}$$

W_Q

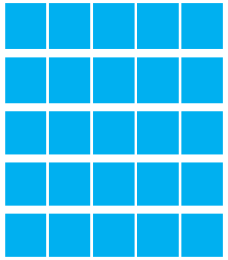


W_K

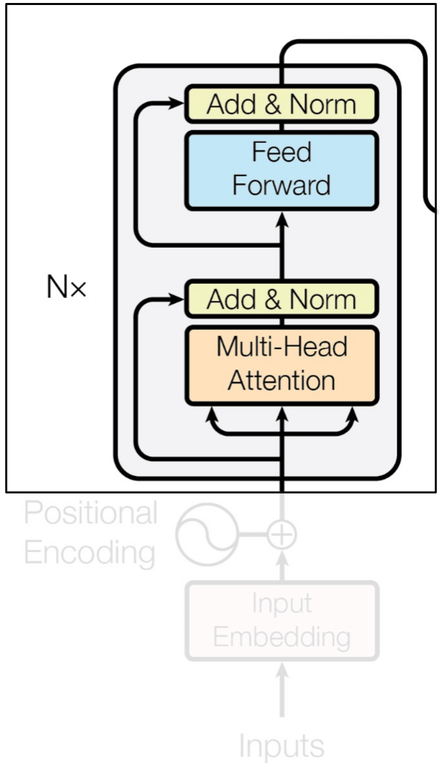


W_V

$$R^{T \times d_{model}}$$

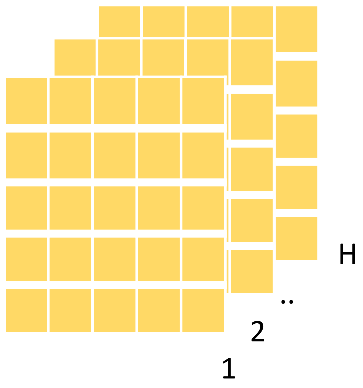


Input Embeddings

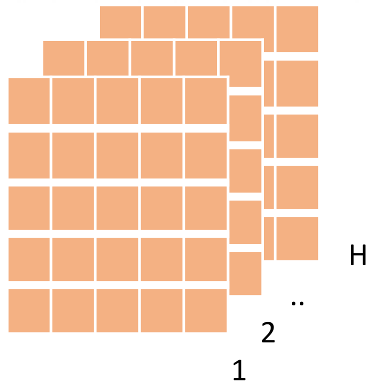


Multi-Head Attention

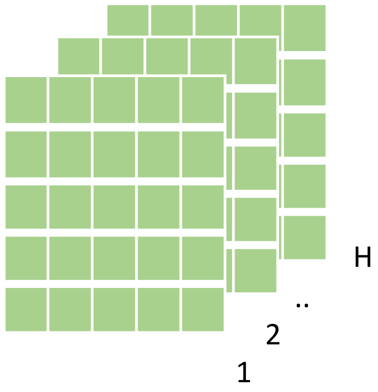
$R^{d_{model} \times d_h}$



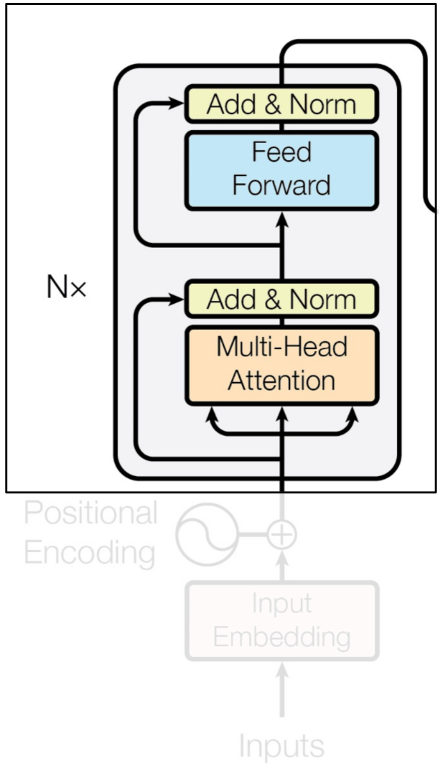
$W_{Q1}, W_{Q2}, \ldots W_{QH},$

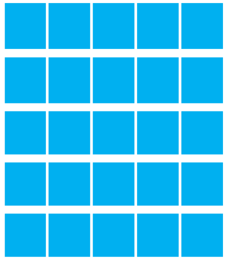


$W_{K1}, W_{K2}, \ldots W_{KH},$



$W_{V1}, W_{V2}, \ldots W_{VH},$



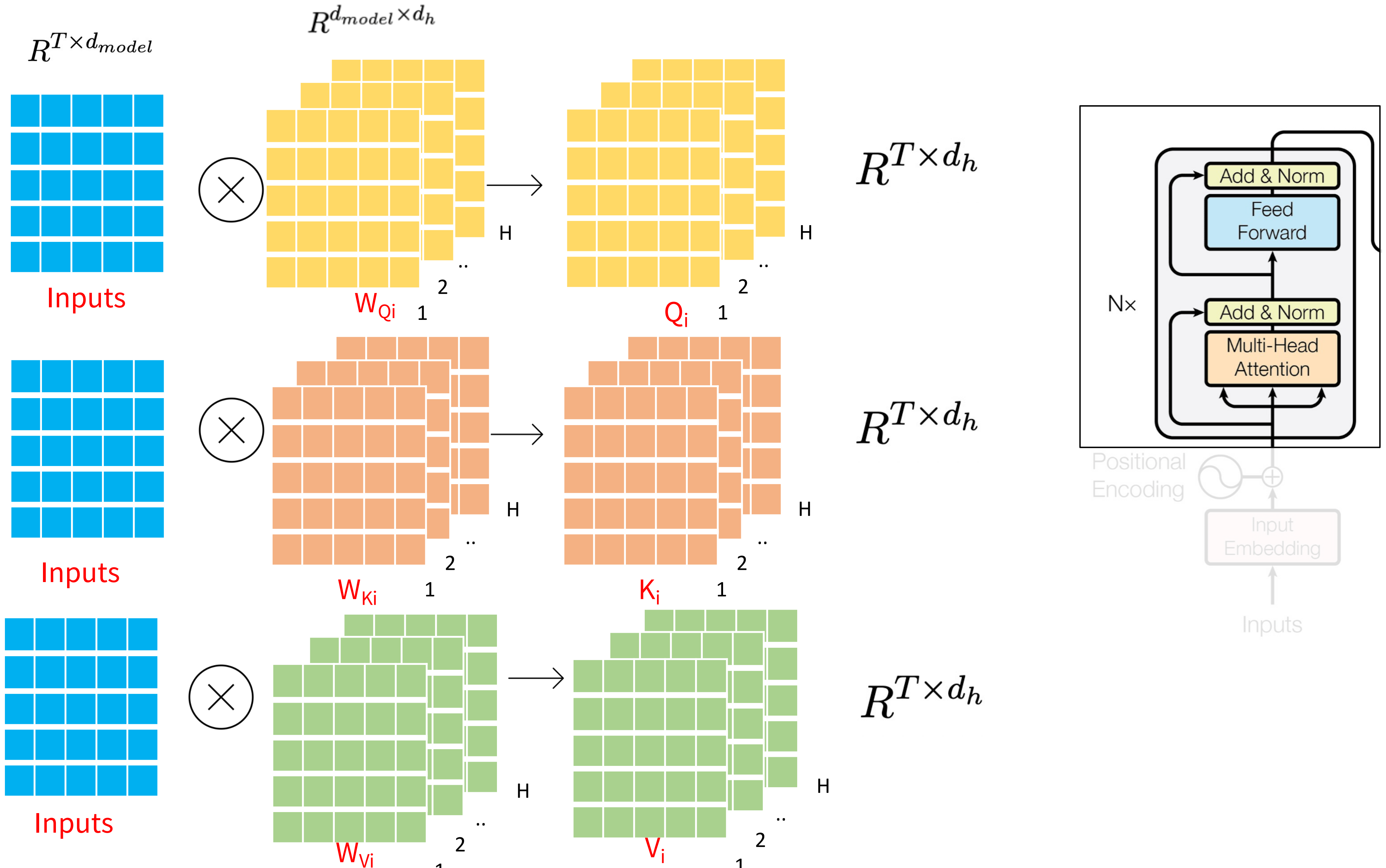


Input Embeddings

$R^{T \times d_{model}}$

$$d_h = \frac{d_{model}}{h}$$

Multi-Head Attention



Multi-Head Attention

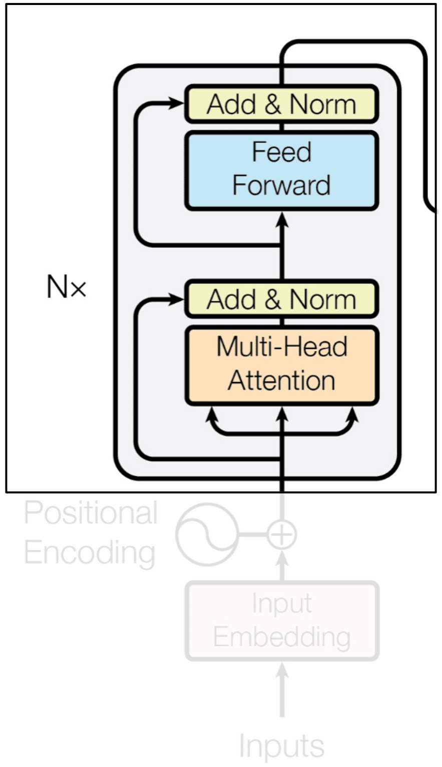
softmax

$R^{T\times d_h}$ $R^{d_h\times T}$ T $R^{T\times d_h}$

Q$_i$ ⊗ K$_i$ ⊗ V$_i$

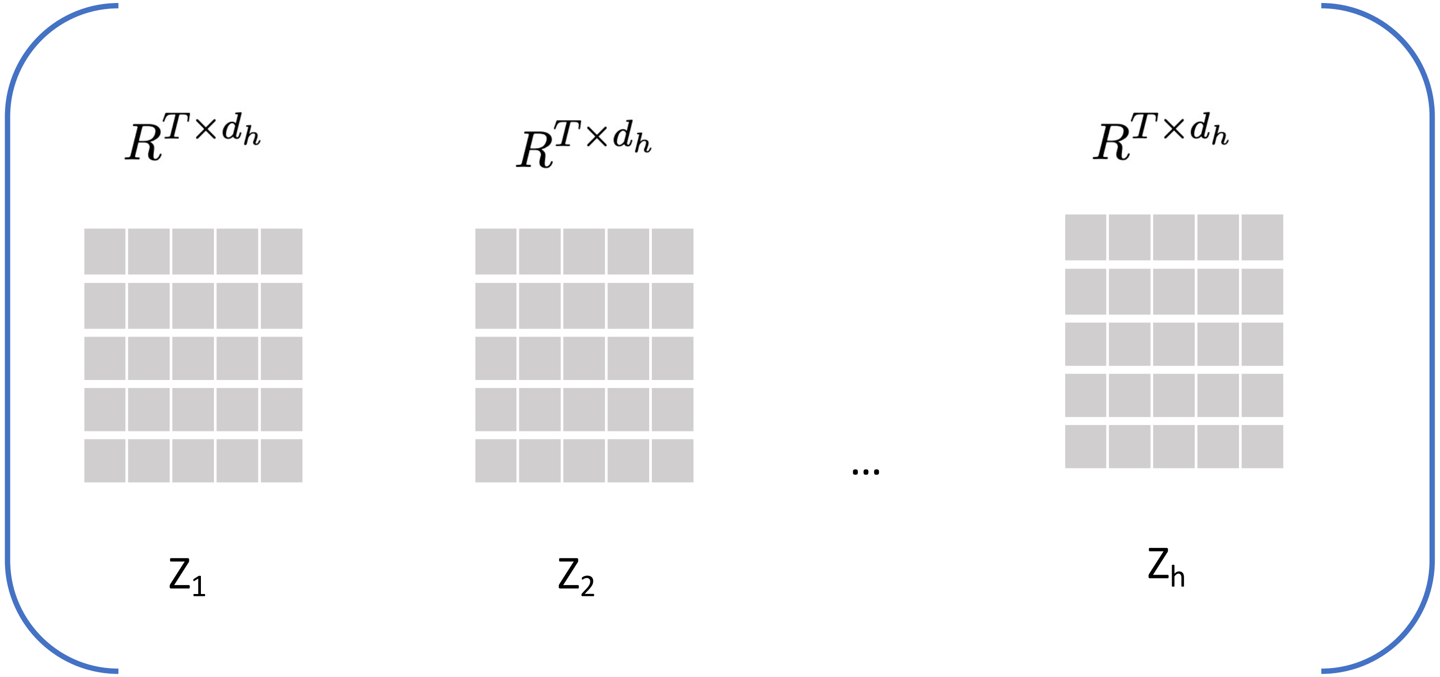
$\sqrt{d_{model}}$

$R^{T\times T}$

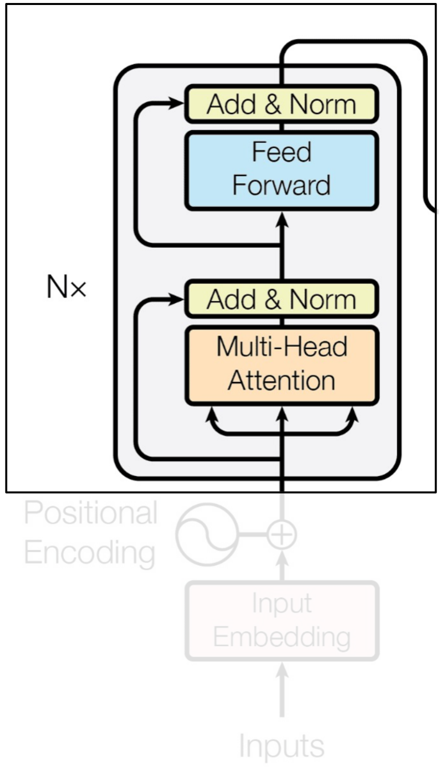


for all i ∈ [1, h]

Multi-Head Attention



CONCAT

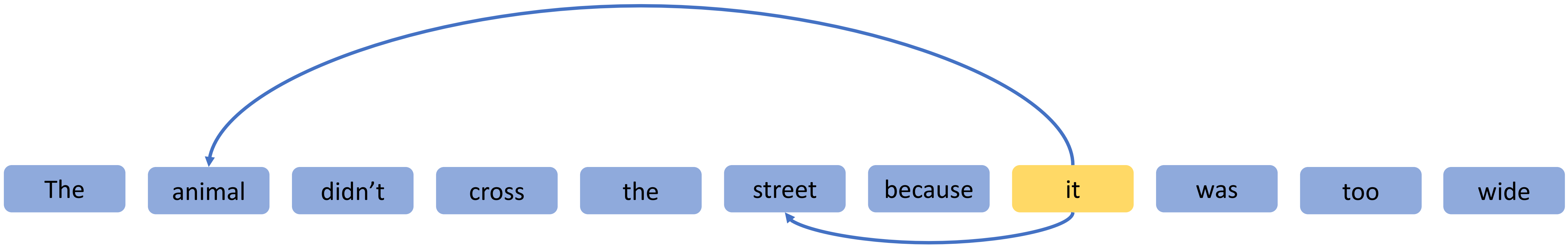


Multi Head Attention : Z

$$d_h = \frac{d_{model}}{h}$$

$$R^{T\times d_{model}}$$

Multi-Head Attention



Sentence boundaries ?

coreference resolution

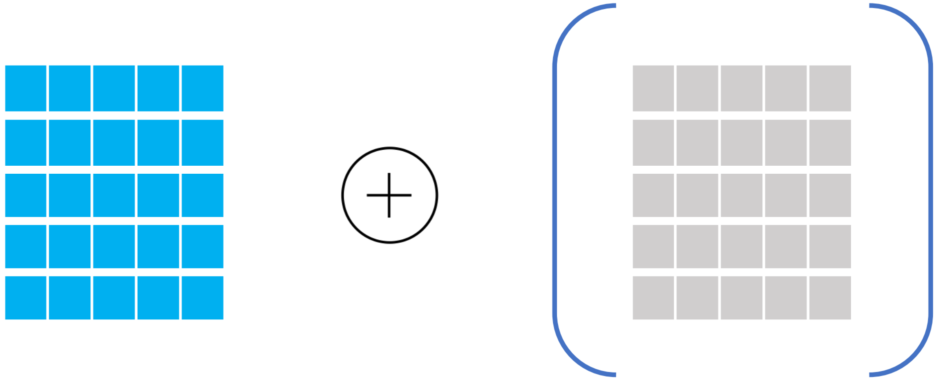
Context ?

Semantic relationships ?

Part of Speech ?

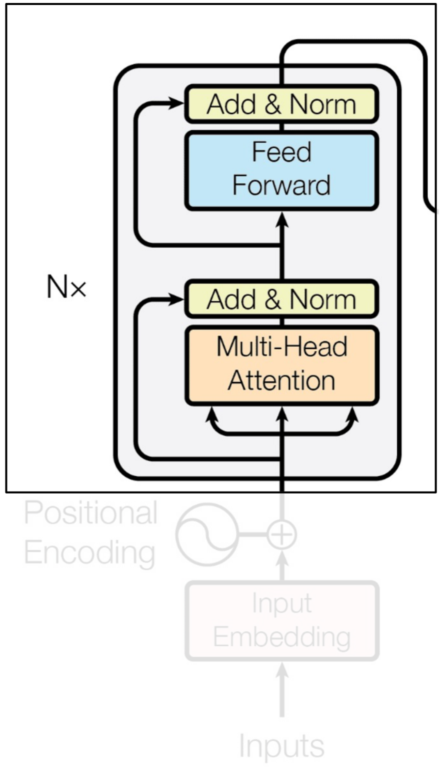
Comparisons ?

Add & Norm



Input

Norm(Z)



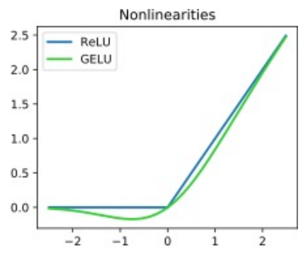
Normalization(Z)

- Mean 0, Std dev 1
- Stabilizes training
- Regularization effect

Add -> Residuals

- Avoid vanishing gradients
- Train deeper networks

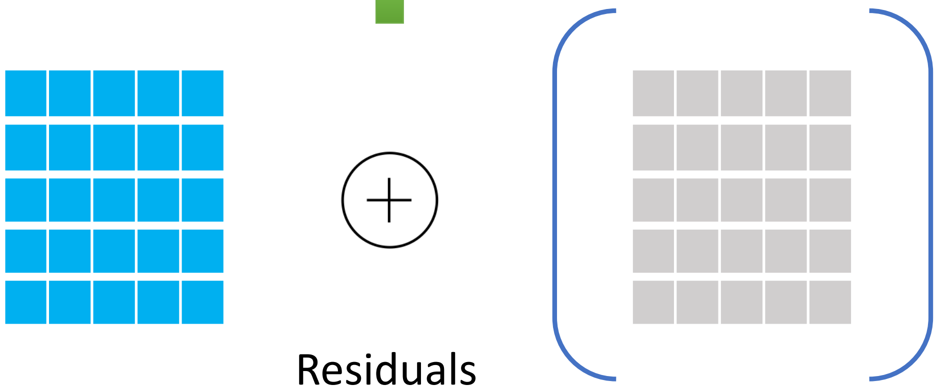
Feed Forward



Feed Forward

- Non Linearity
- Complex Relationships
- Learn from each other

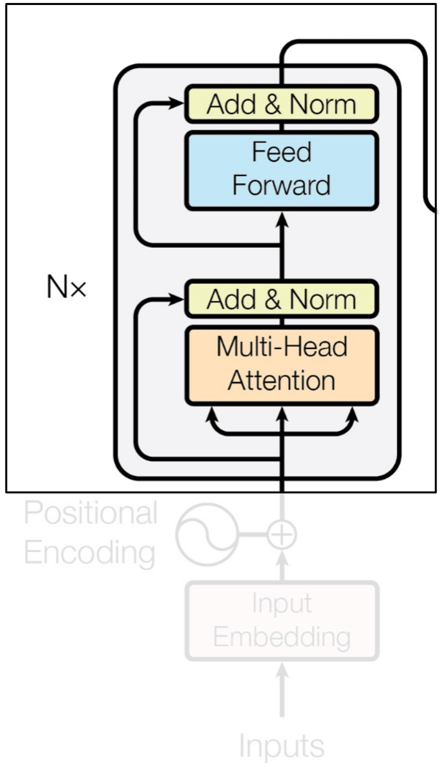
Feed Forward



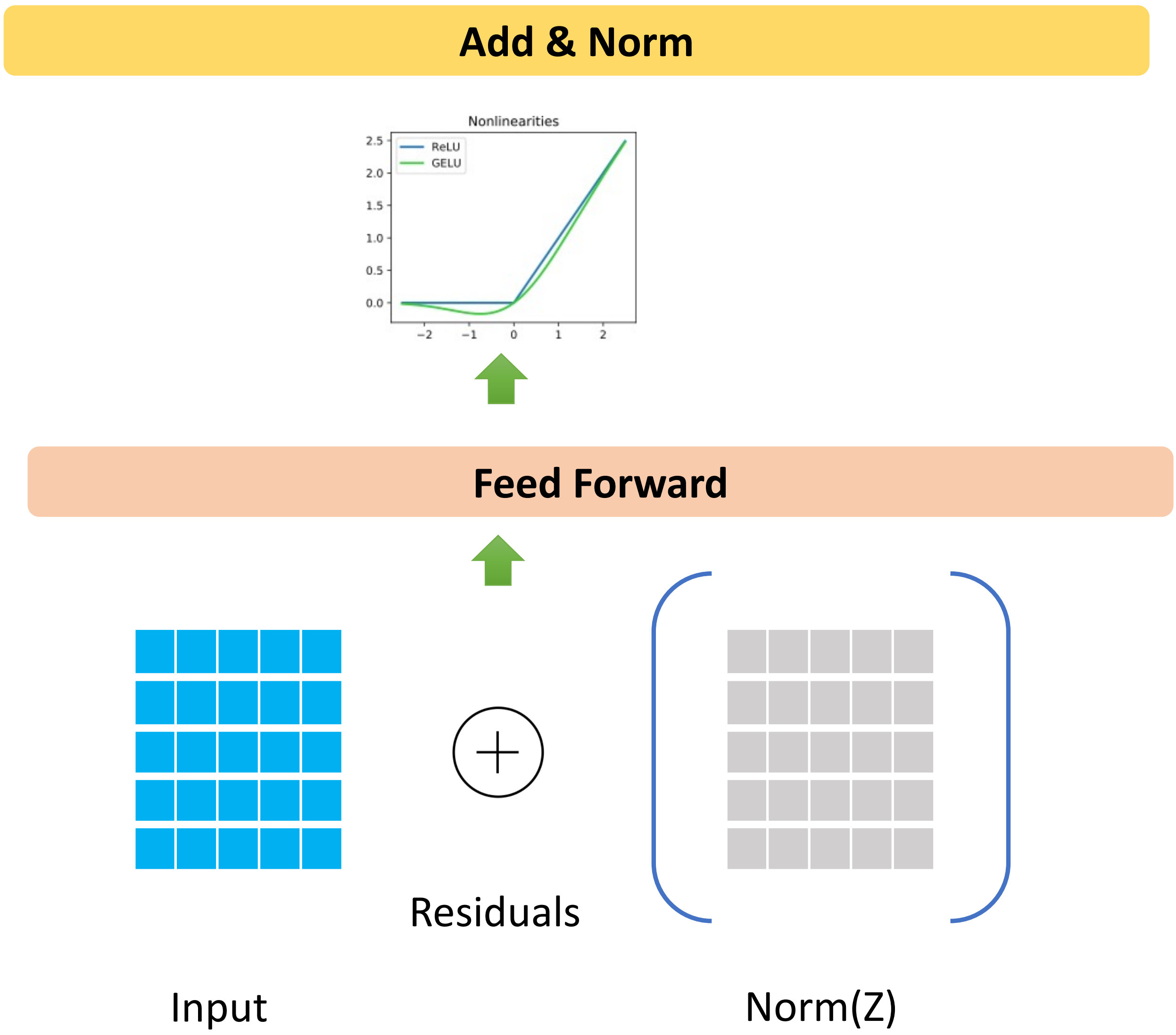
Residuals

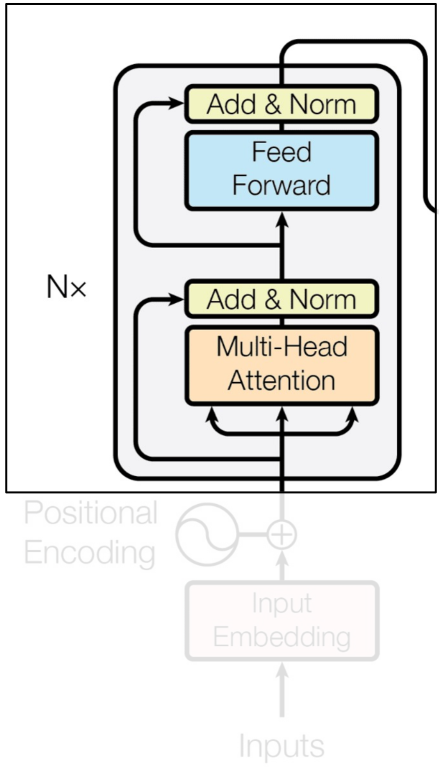
Input

Norm(Z)



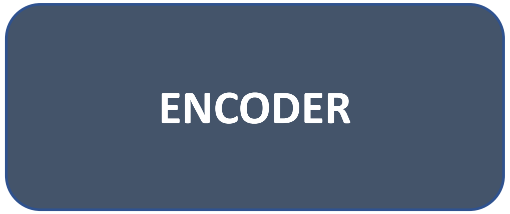
Add & Norm

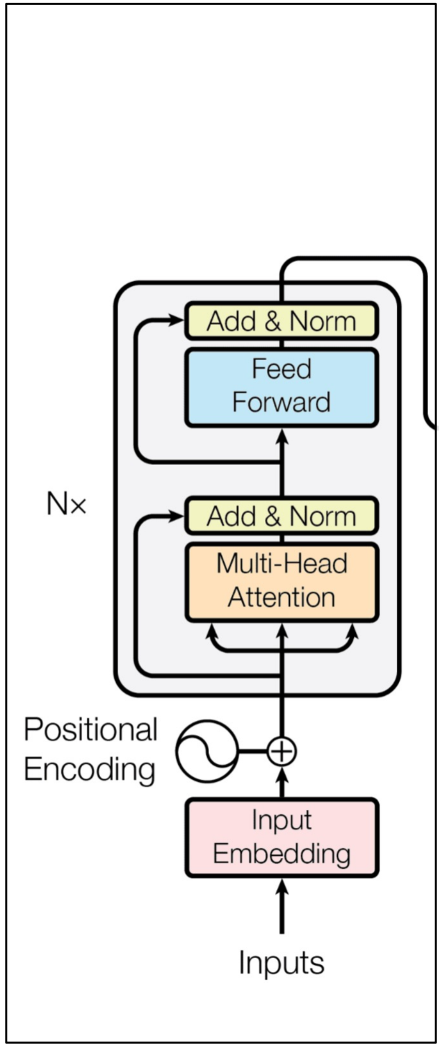

Add & Norm
Nonlinearities
ReLU
GELU
2.5
2.0
1.5
1.0
0.5
0.0
−2
−1
0
1
2
Feed Forward
Residuals
Input
Norm(Z)


Add & Norm
Feed
Forward
Nx
Add & Norm
Multi-Head
Attention
Positional
Encoding
Input
Embedding
Inputs

Encoders

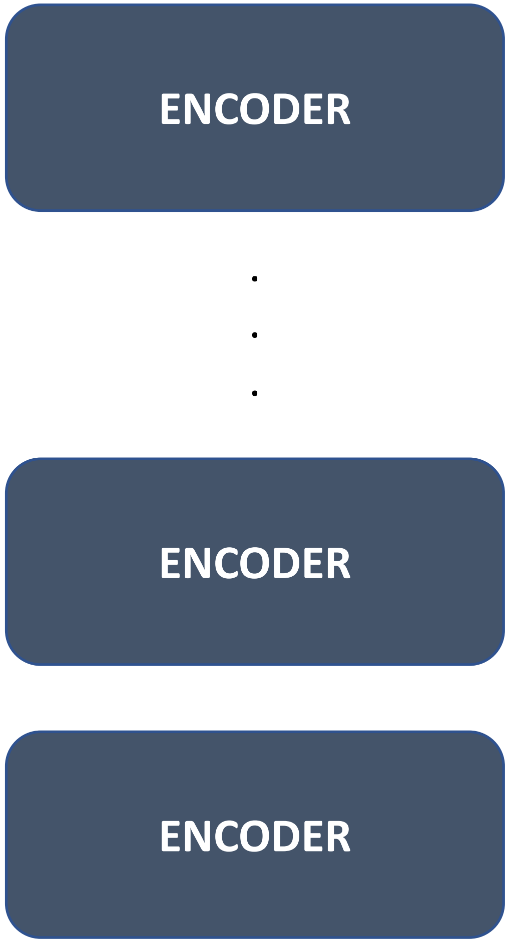
Encoder





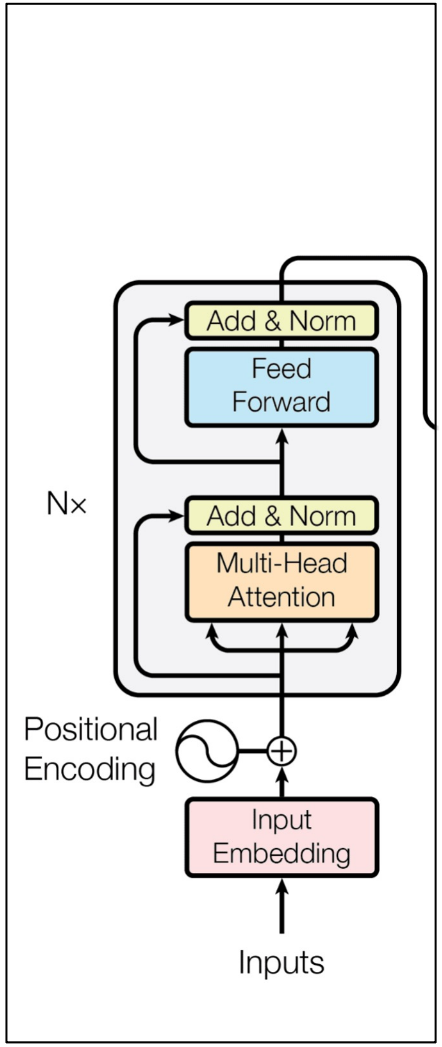
Encoders

Encoder



Input to Encoder_{i+1}

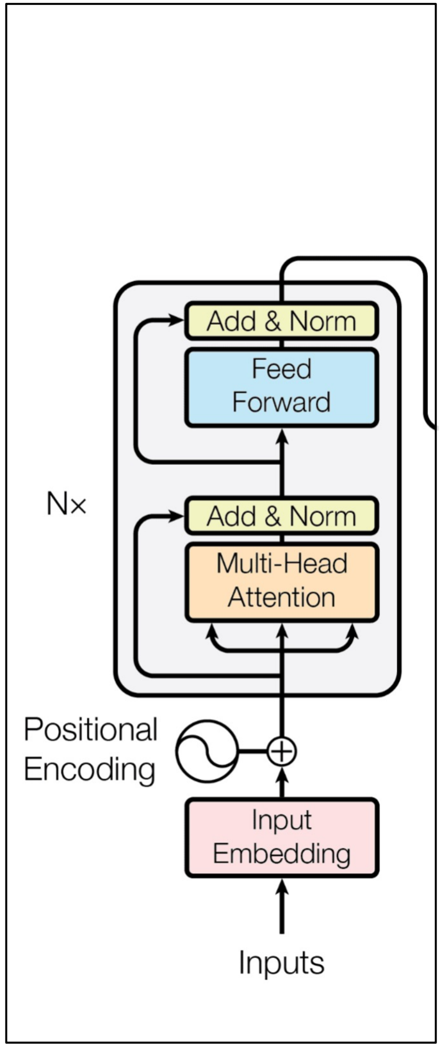
Output from Encoder_{i}



Transformers

- ✓ Tokenizaton
- ✓ Input Embeddings
- ✓ Position Encodings
- ✓ Residuals
- ✓ Query
- ✓ Key
- ✓ Value
- ✓ Add & Norm
- ✓ Encoder
- **Decoder**

- ✓ Attention
- ✓ Self Attention
- ✓ Multi Head Attention
- **Masked Attention**
- **Encoder Decoder Attention**
- **Output Probabilities / Logits**
- **Softmax**
- **Encoder-Decoder models**
- **Decoder only models**



Machine Translation

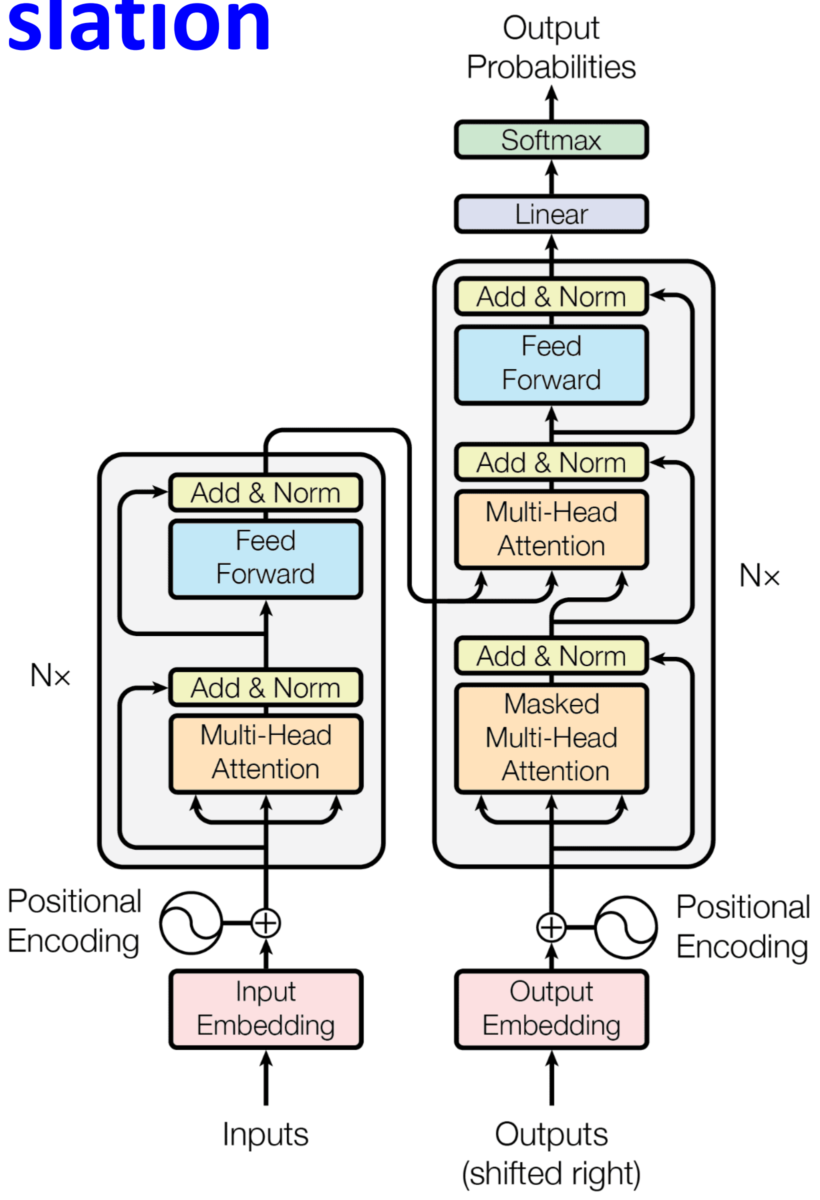
Targets

Ich have einen apfel gegessen

⬆

Inputs

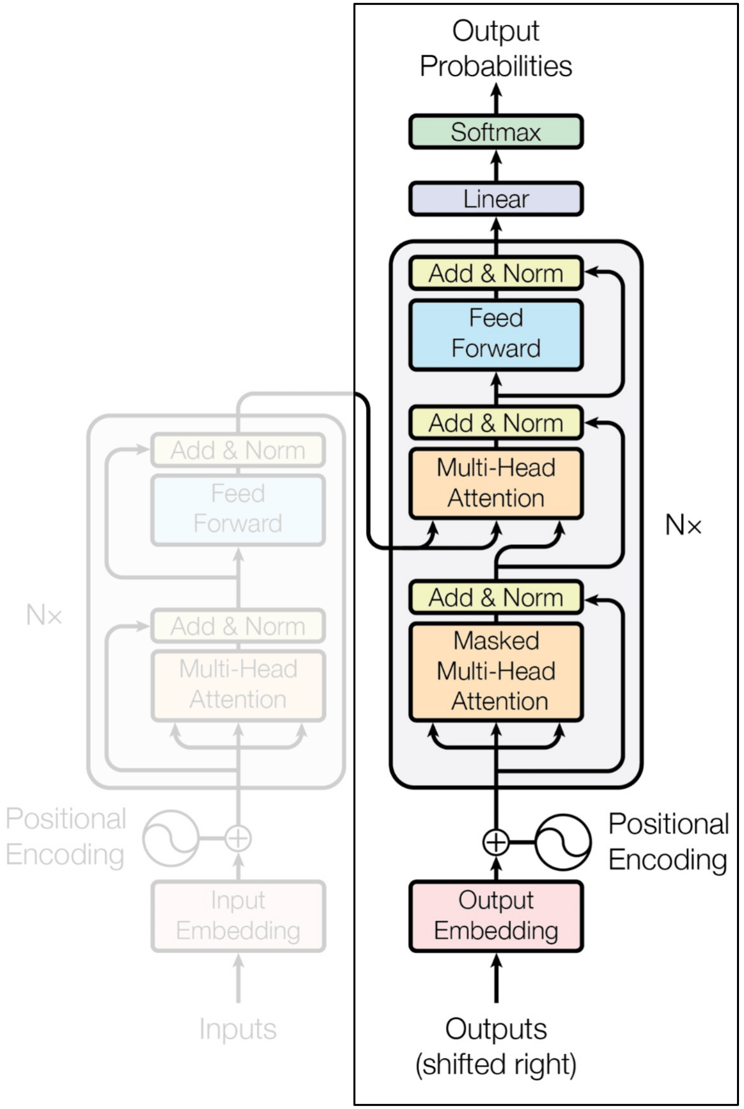
I ate an apple



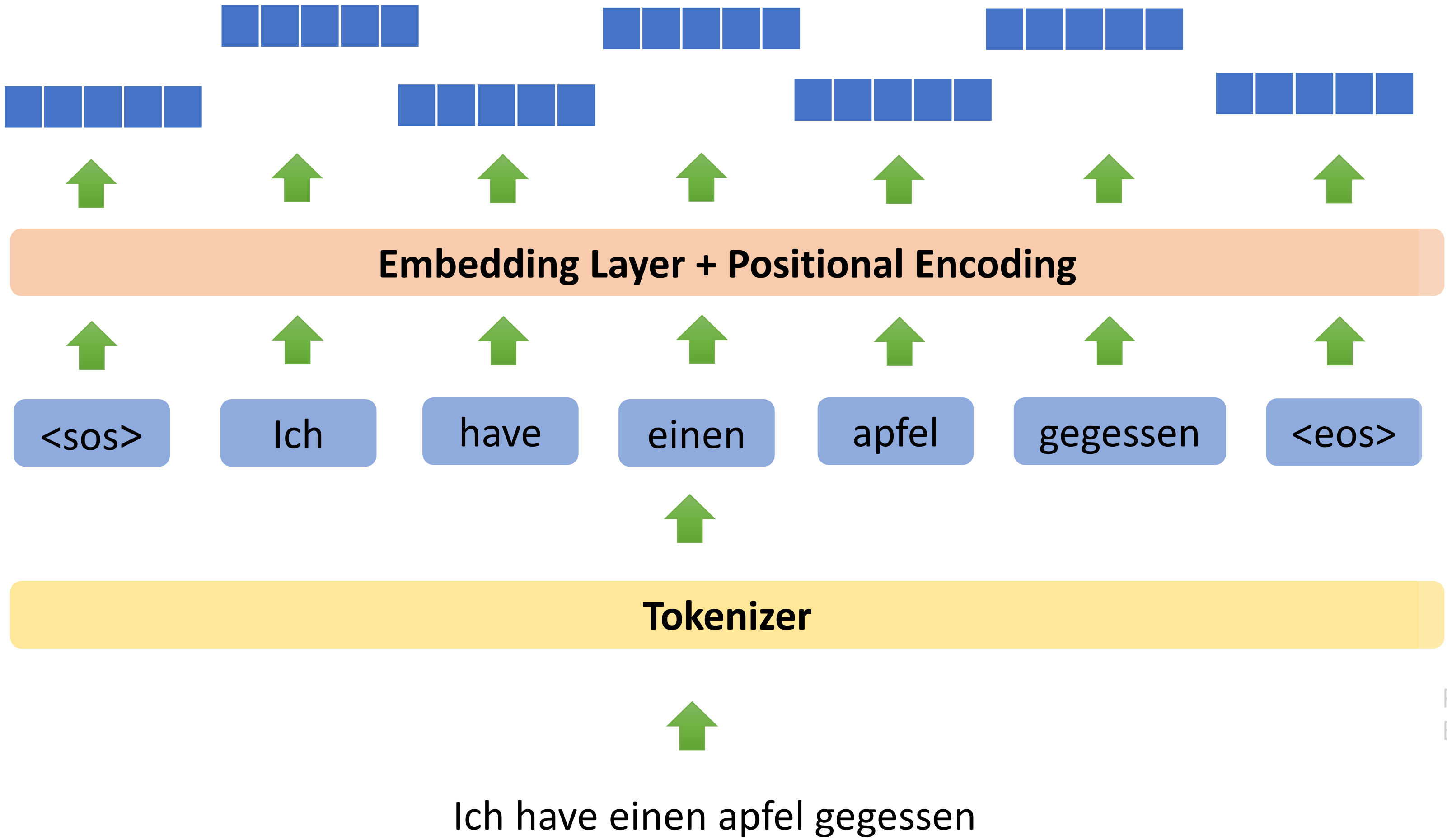
Targets

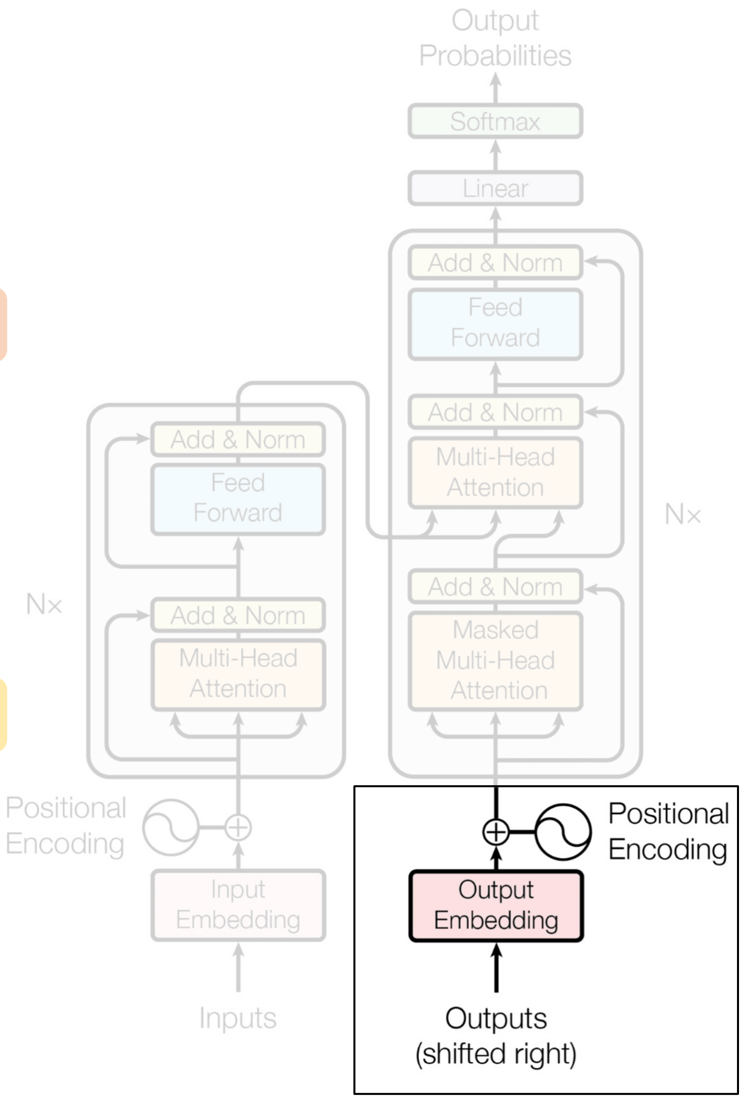
Targets

Ich have einen apfel gegessen



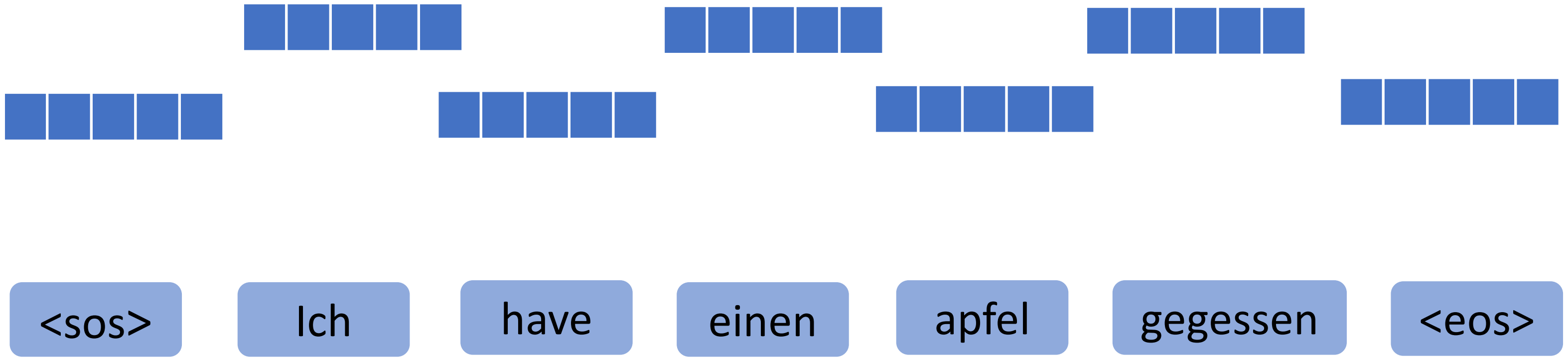
Targets

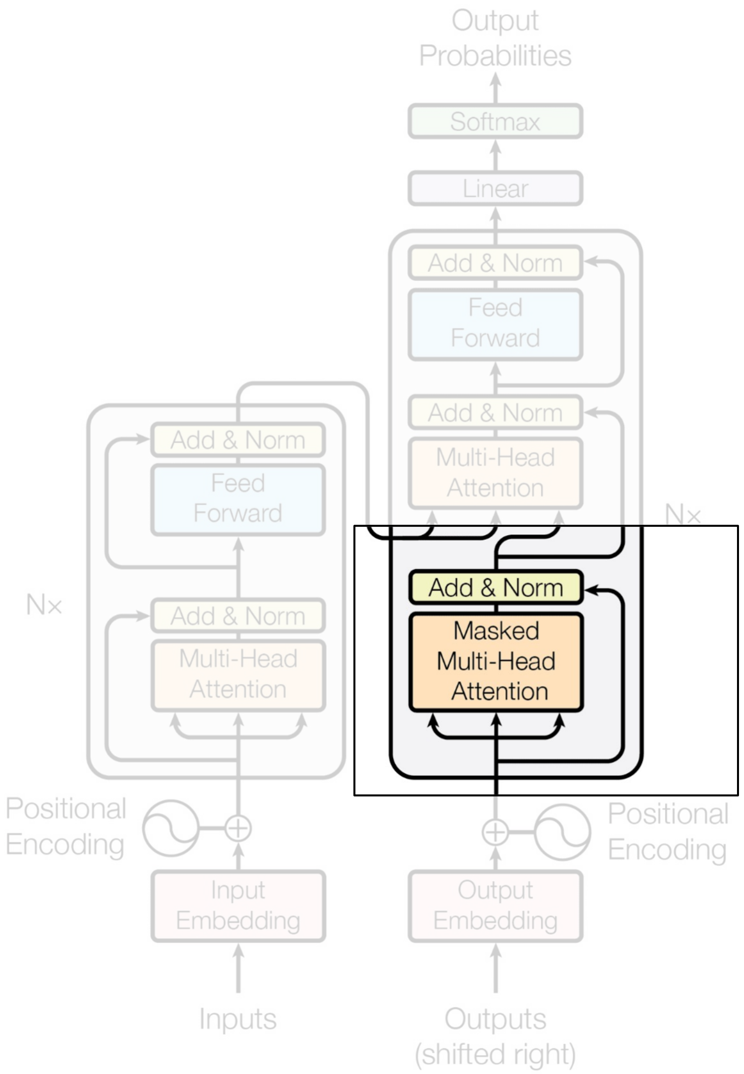




Generate Target Emebeddings

Masked Multi Head Attention



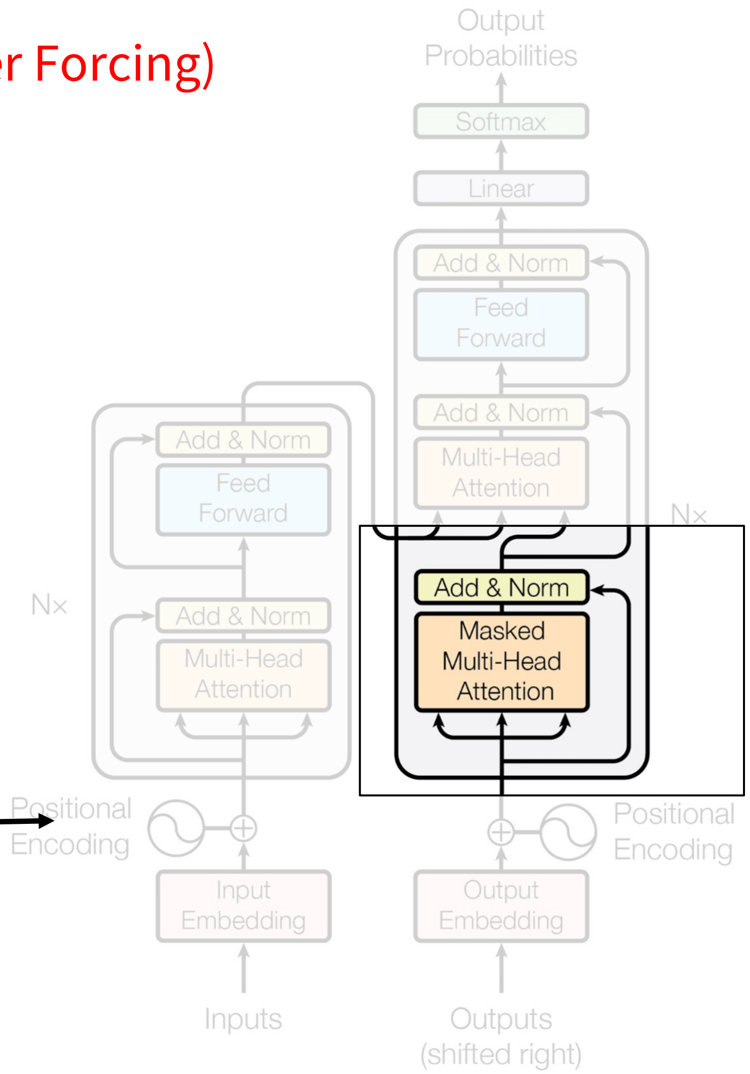


Masked Multi Head Attention

Decoding step by step (using Teacher Forcing)

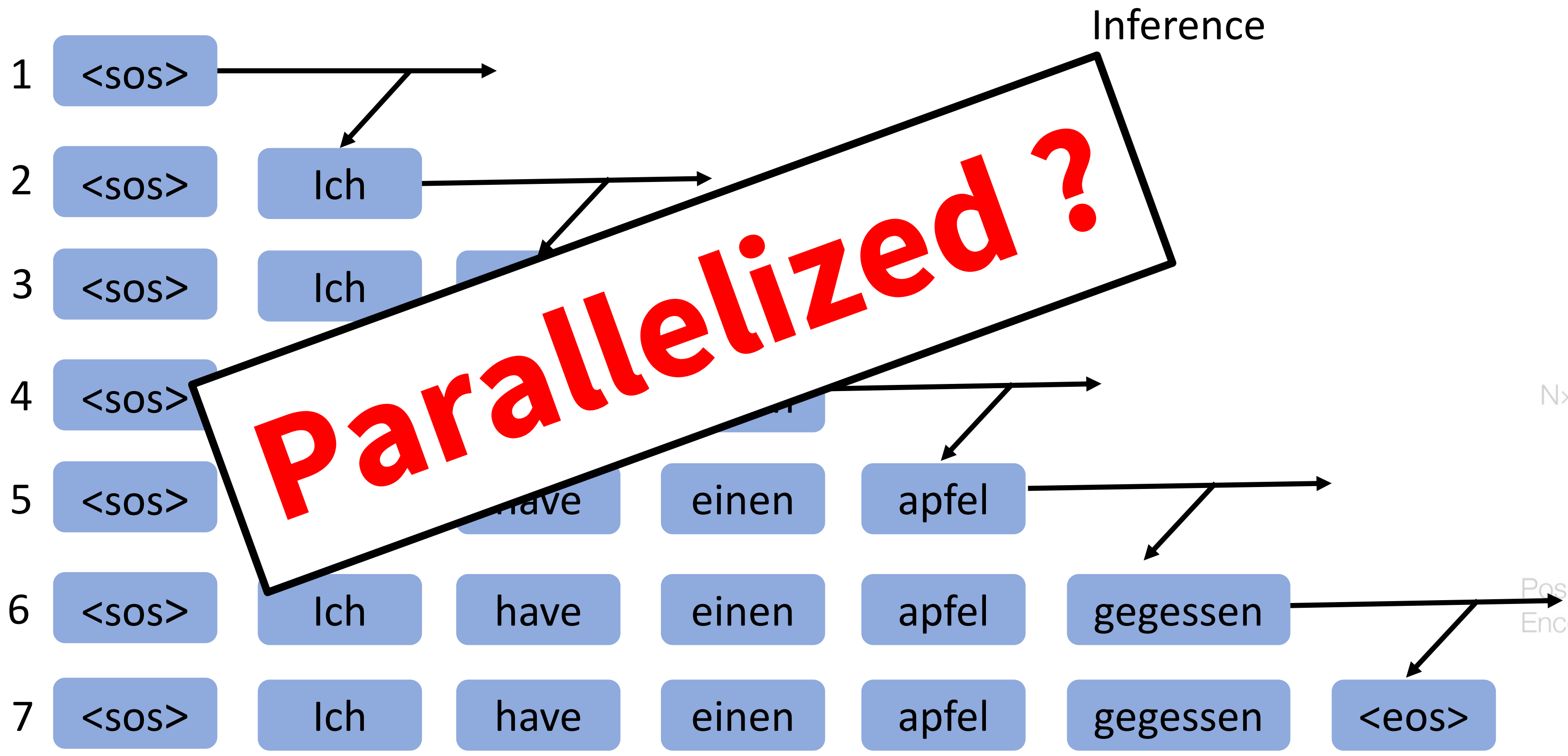
Inference

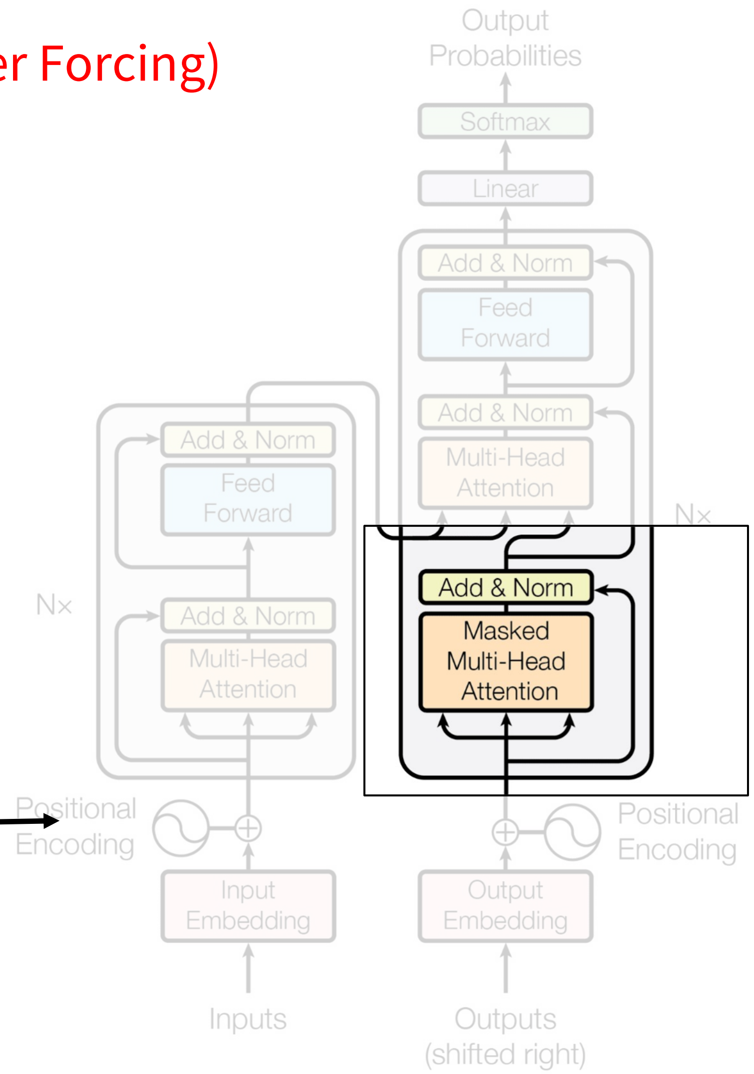
1 <sos>
2 <sos> Ich
3 <sos> Ich have
4 <sos> Ich have einen
5 <sos> Ich have einen apfel
6 <sos> Ich have einen apfel gegessen
7 <sos> Ich have einen apfel gegessen <eos>



Masked Multi Head Attention

Decoding step by step (using Teacher Forcing)



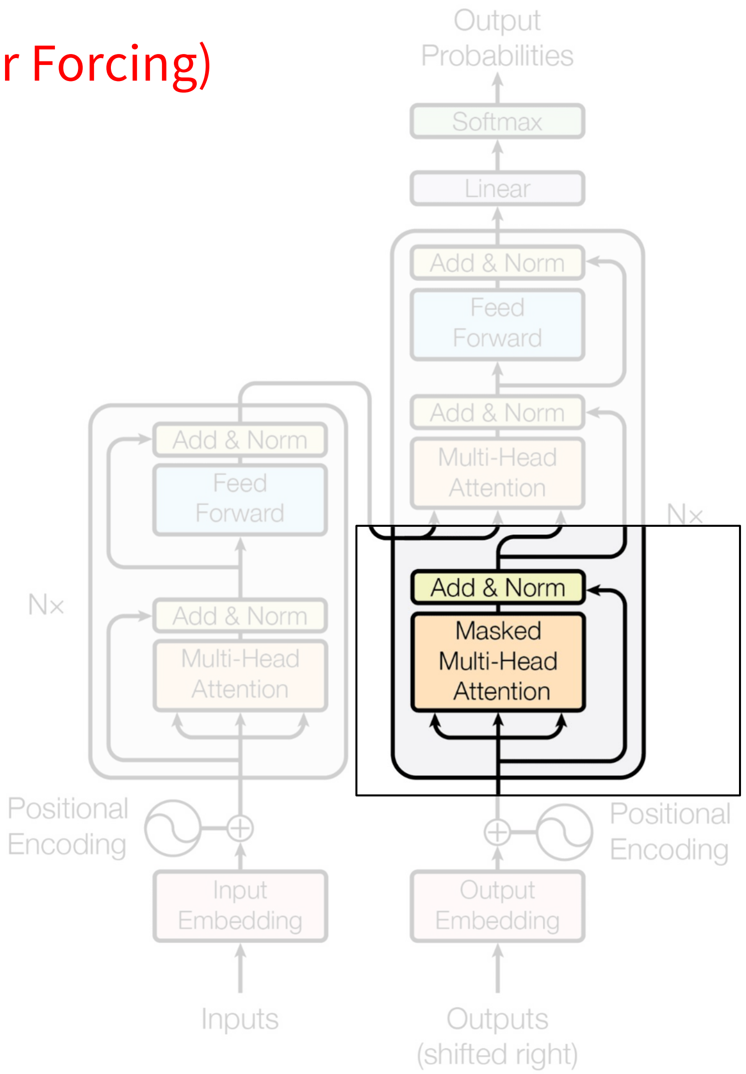


Masked Multi Head Attention

Decoding step by step (using Teacher Forcing)

Training

<sos> Ich have einen apfel gegessen <eos>



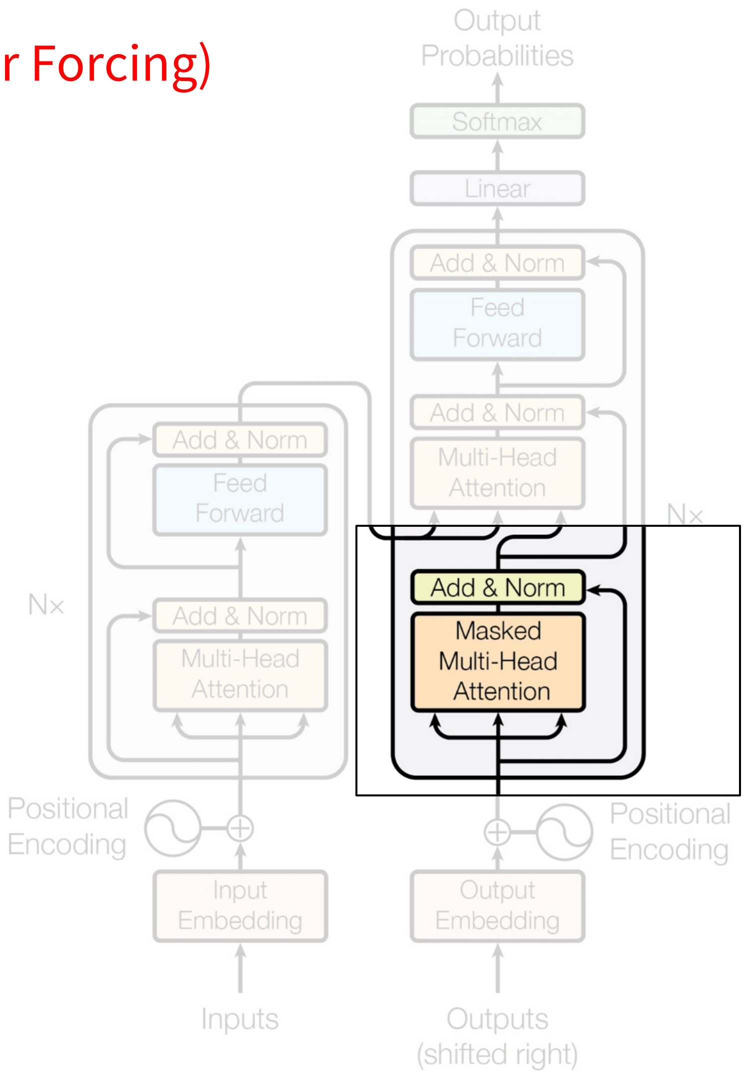
Masked Multi Head Attention

Decoding step by step (using Teacher Forcing)

Training

<sos> | Ich | have | einen | apfel | gegessen | <eos>

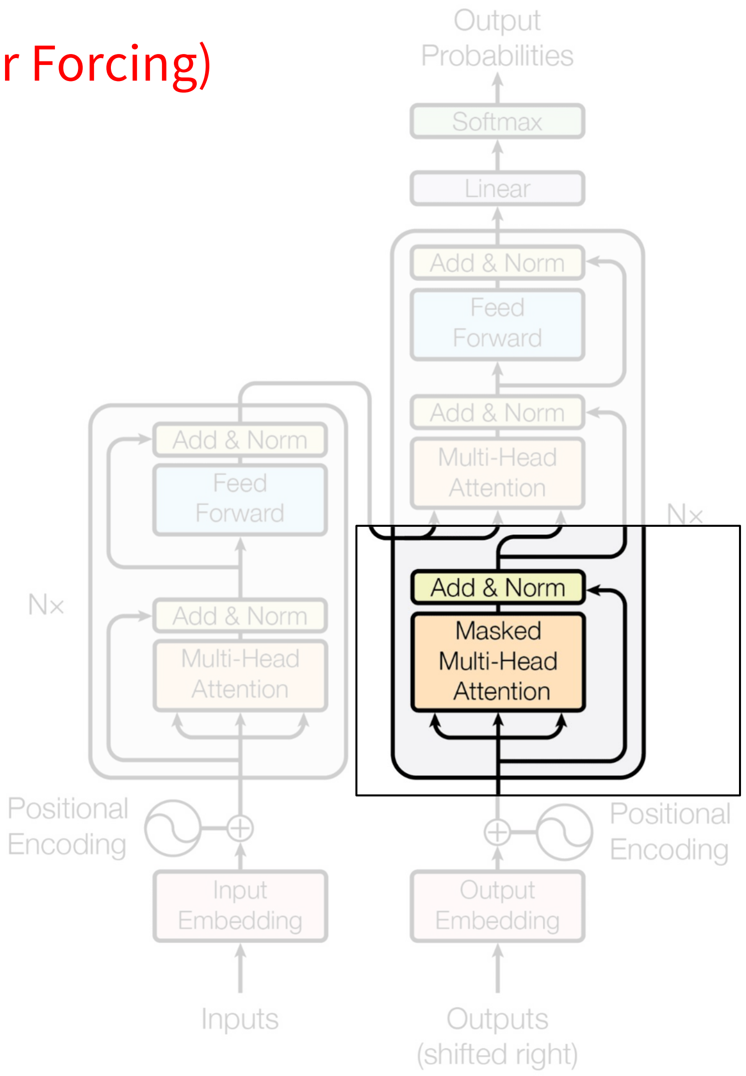
Outputs at time T should only pay attention to outputs until time T-1



Masked Multi Head Attention

Decoding step by step (using Teacher Forcing)

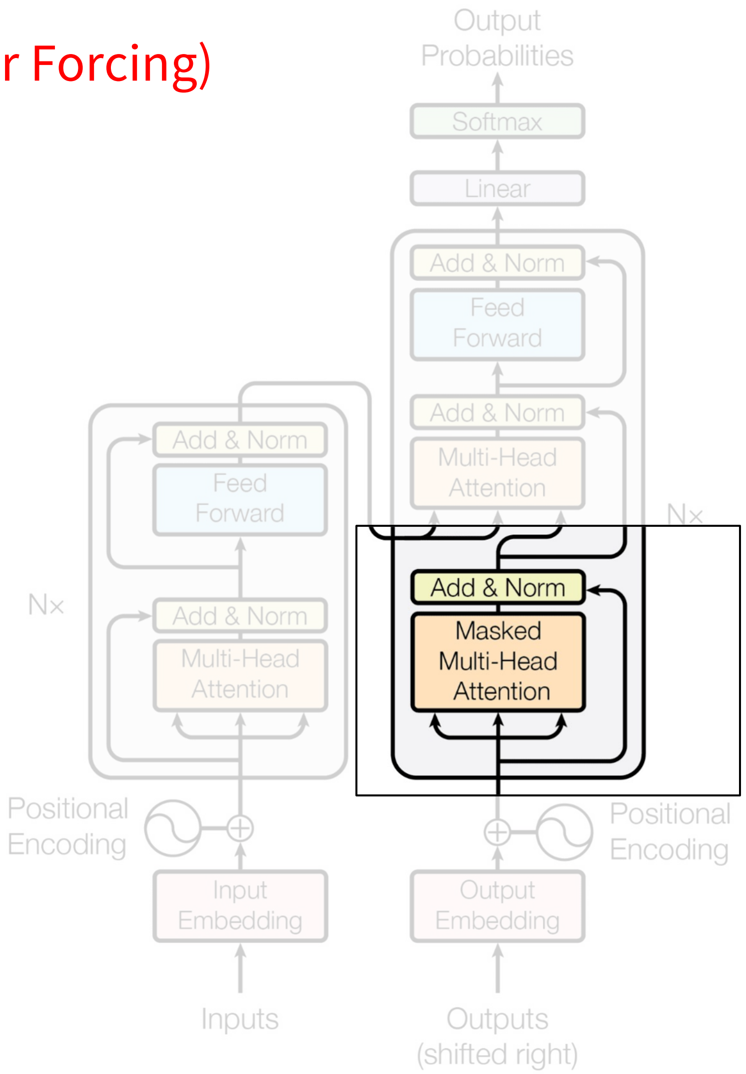
1	<sos>	Ich	have	einen	apfel	gegessen	<eos>
2	<sos>	Ich	have	einen	apfel	gegessen	<eos>
3	<sos>	Ich	have	einen	apfel	gegessen	<eos>
4	<sos>	Ich	have	einen	apfel	gegessen	<eos>
5	<sos>	Ich	have	einen	apfel	gegessen	<eos>
6	<sos>	Ich	have	einen	apfel	gegessen	<eos>
7	<sos>	Ich	have	einen	apfel	gegessen	<eos>



Masked Multi Head Attention

Decoding step by step (using Teacher Forcing)

1	<sos>	Ich	have	einen	apfel	gegessen	<eos>
2	<sos>	Ich	have	einen	apfel	gegessen	<eos>
3	<sos>	Ich	have	einen	apfel	gegessen	<eos>
4	<sos>	Ich	have	einen	apfel	gegessen	<eos>
5	<sos>	Ich	have	einen	apfel	gegessen	<eos>
6	<sos>	Ich	have	einen	apfel	gegessen	<eos>
7	<sos>	Ich	have	einen	apfel	gegessen	<eos>

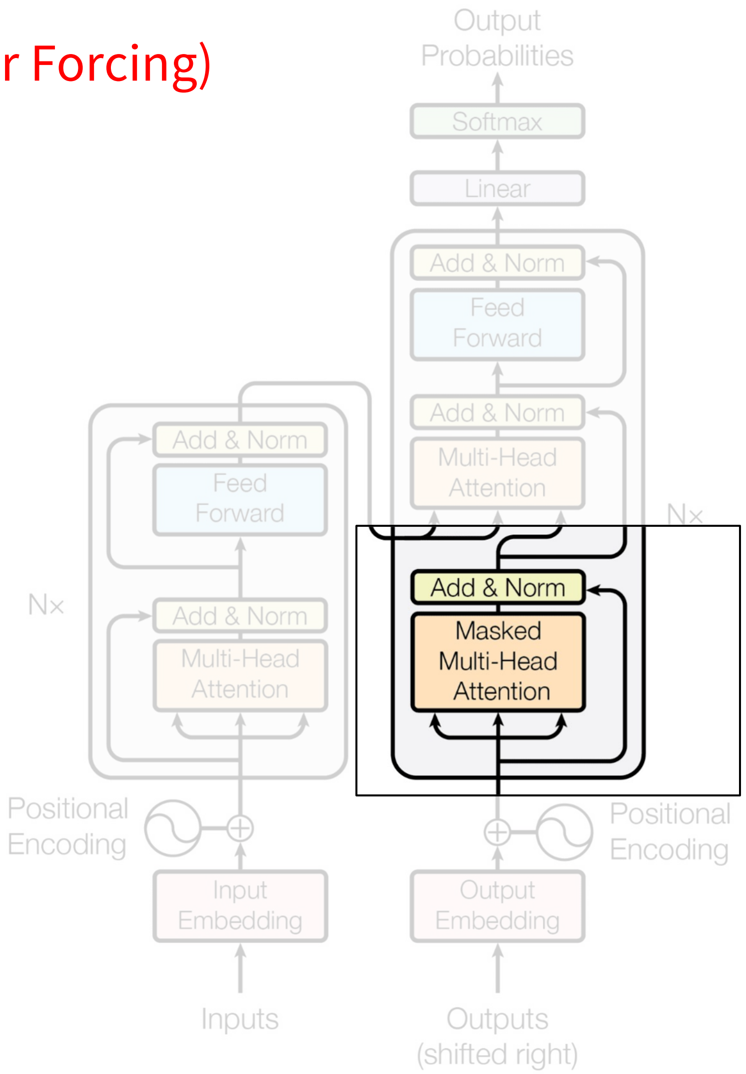


Mask the available attention values ?

Masked Multi Head Attention

Decoding step by step (using Teacher Forcing)

1	<sos>	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$
2	<sos>	Ich	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$
3	<sos>	Ich	have	$-\infty$	$-\infty$	$-\infty$	$-\infty$
4	<sos>	Ich	have	einen	$-\infty$	$-\infty$	$-\infty$
5	<sos>	Ich	have	einen	apfel	$-\infty$	$-\infty$
6	<sos>	Ich	have	einen	apfel	gegessen	$-\infty$
7	<sos>	Ich	have	einen	apfel	gegessen	<eos>

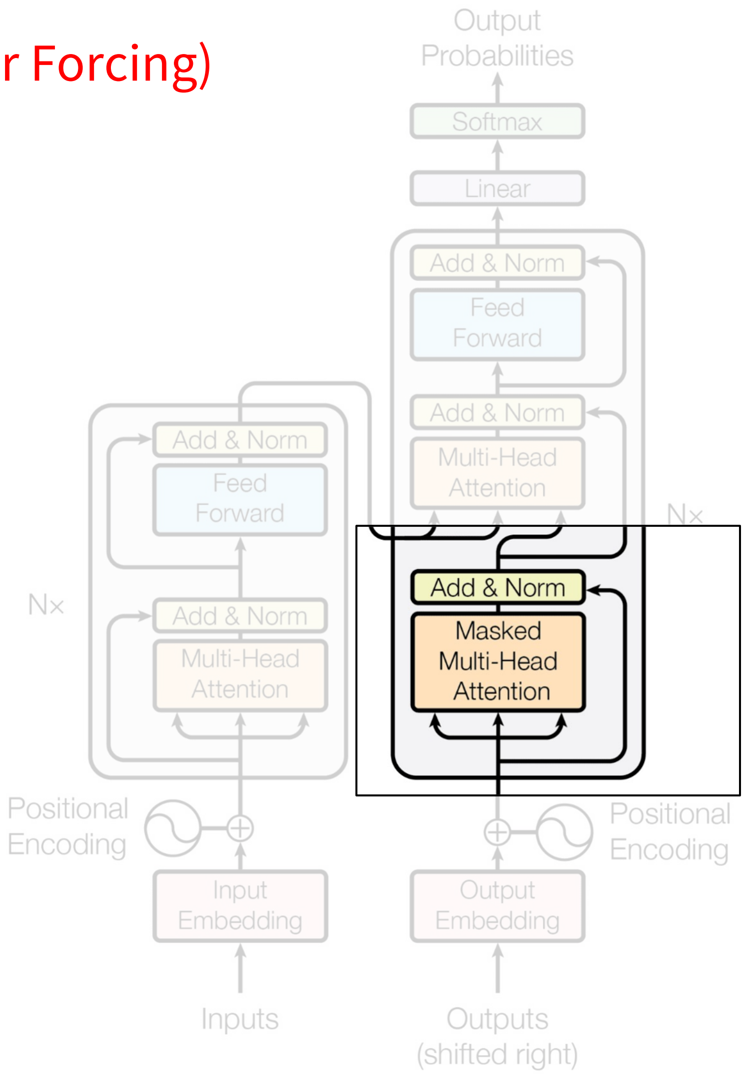


Masked Multi Head Attention

Decoding step by step (using Teacher Forcing)

1	<sos>	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$
2	<sos>	Ich	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$
3	<sos>	Ich	have	$-\infty$	$-\infty$	$-\infty$	$-\infty$
4	<sos>	Ich	have	einen	$-\infty$	$-\infty$	$-\infty$
5	<sos>	Ich	have	einen	apfel	$-\infty$	$-\infty$
6	<sos>	Ich	have	einen	apfel	gegessen	$-\infty$
7	<sos>	Ich	have	einen	apfel	gegessen	<eos>

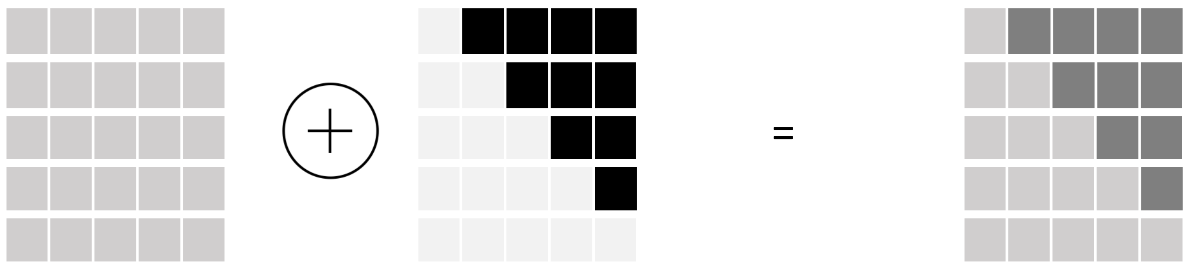
Softmax -> $-\infty$ -> 0



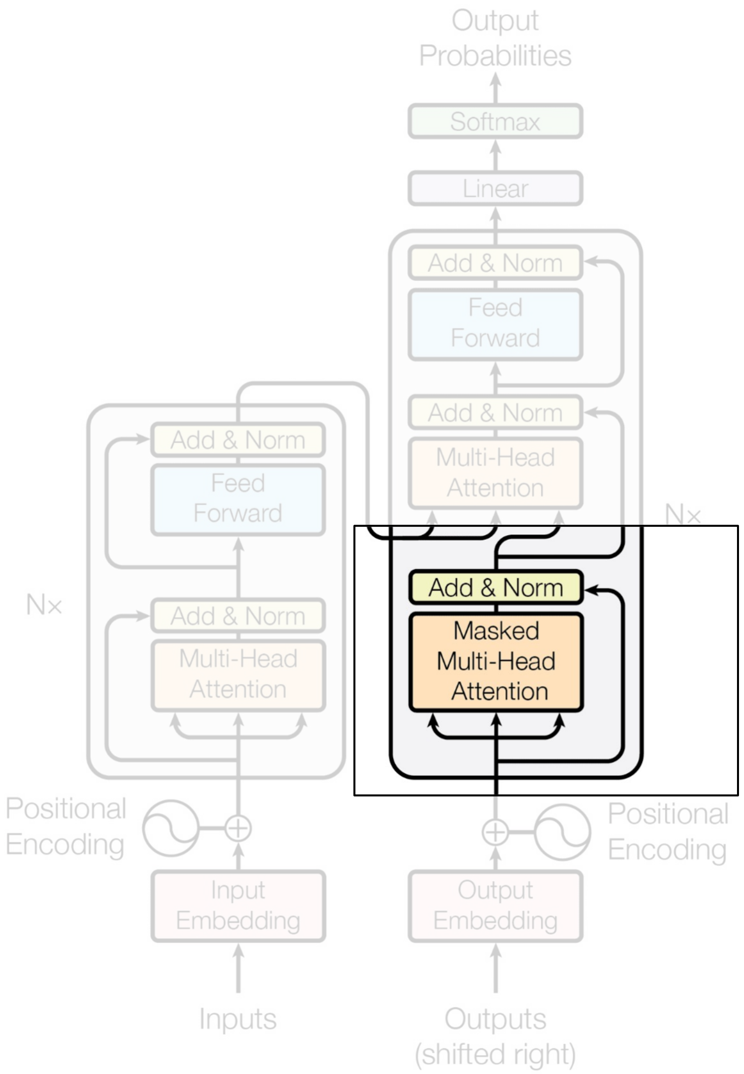
Masked Multi Head Attention

Masked Multi Head Attention

$R^{T \times T}$ $R^{T \times T}$



QK^T Attention Mask: M Masked Attention

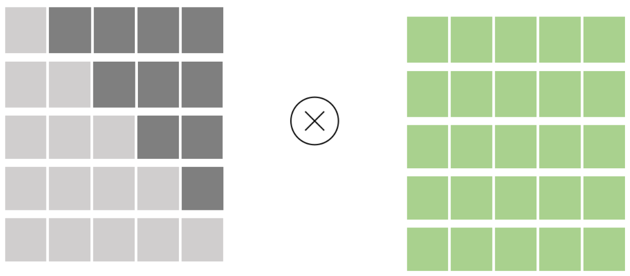


Masked Multi Head Attention : Z’

Masked Multi Head Attention

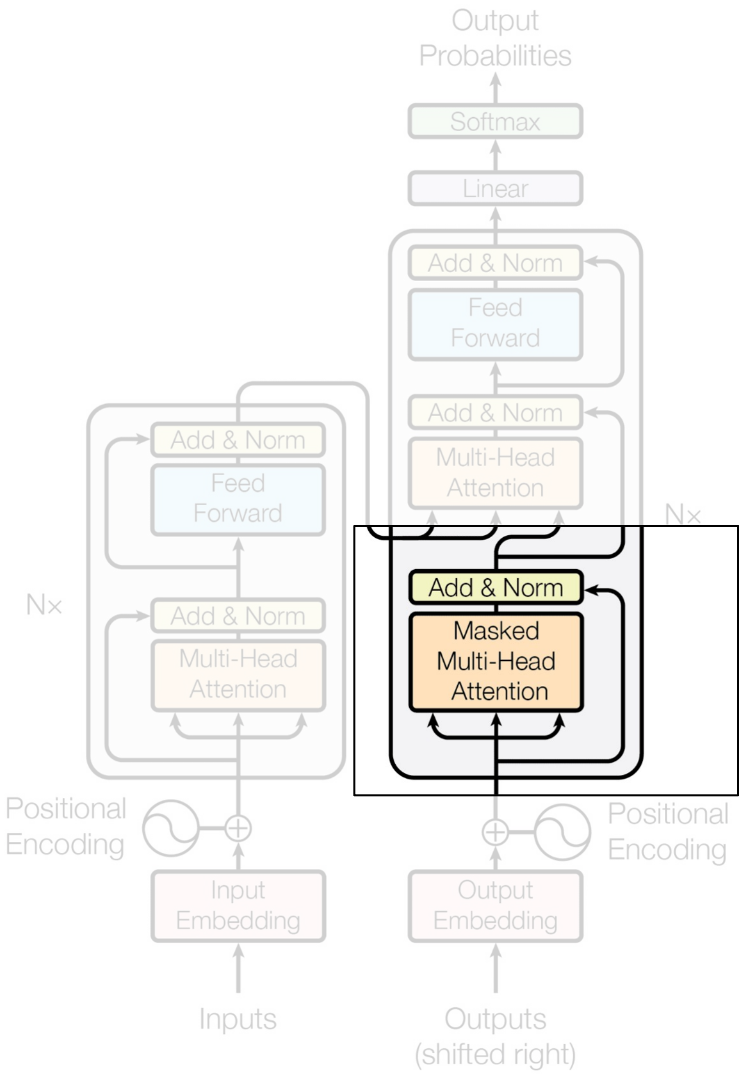
Masked Multi Head Attention

$R^{T \times T}$ $R^{T \times d_h}$

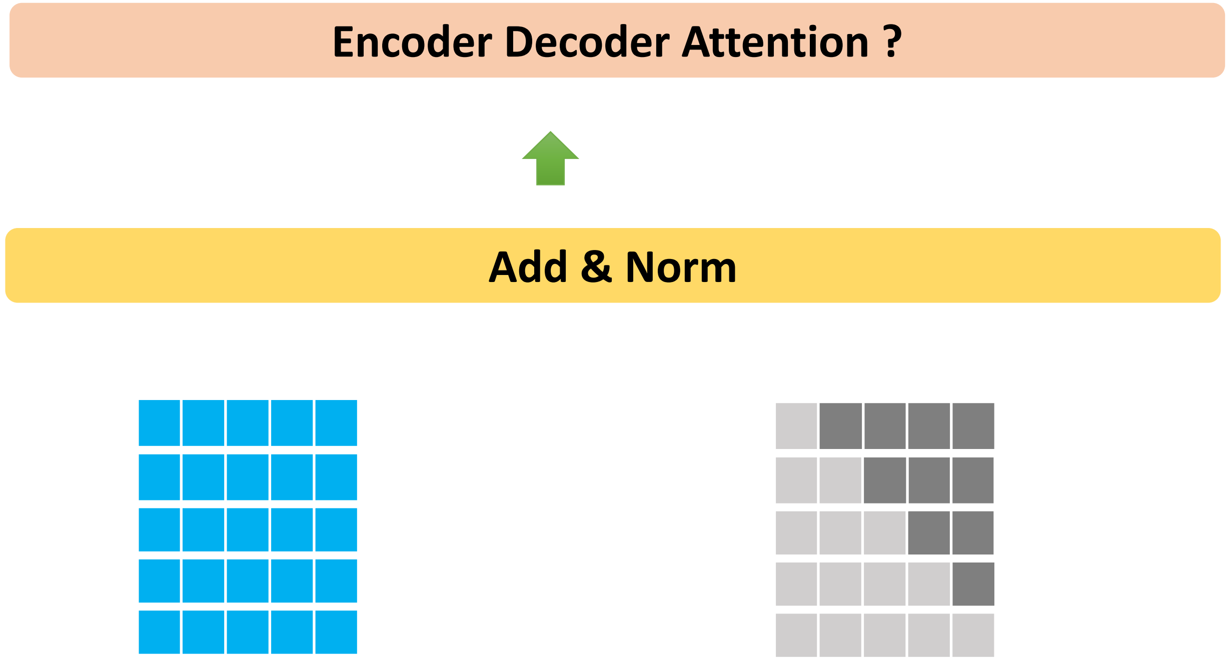


Masked Attention Values

Masked Multi Head Attention : Z’

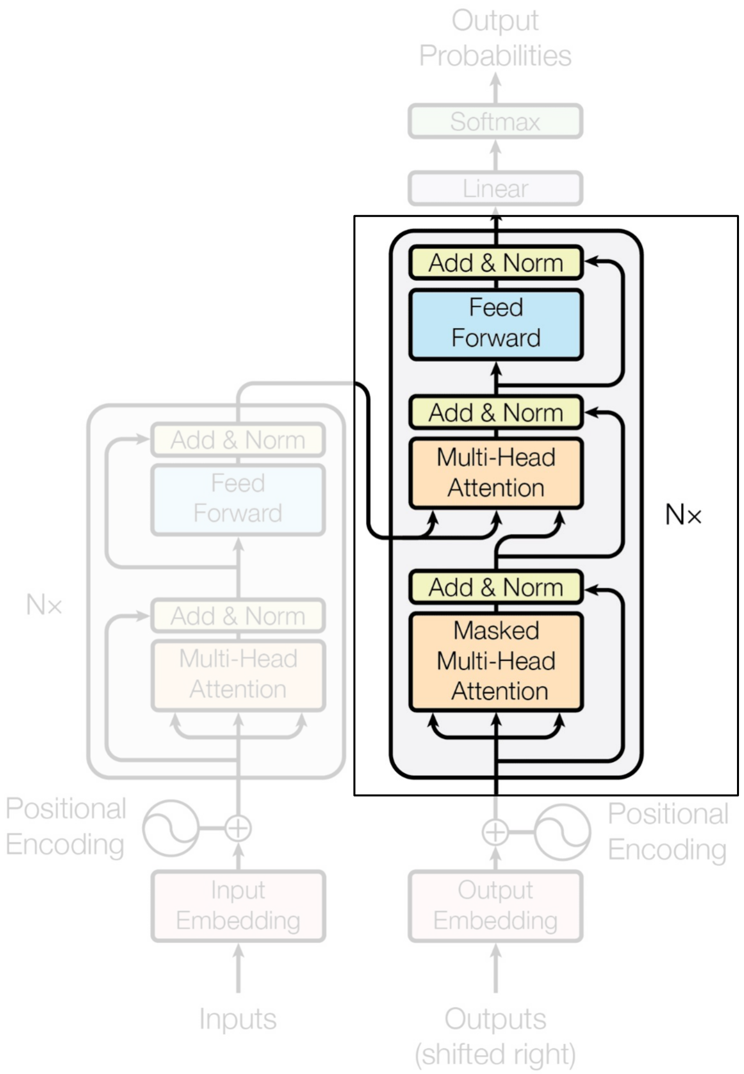


Encoder Decoder Attention



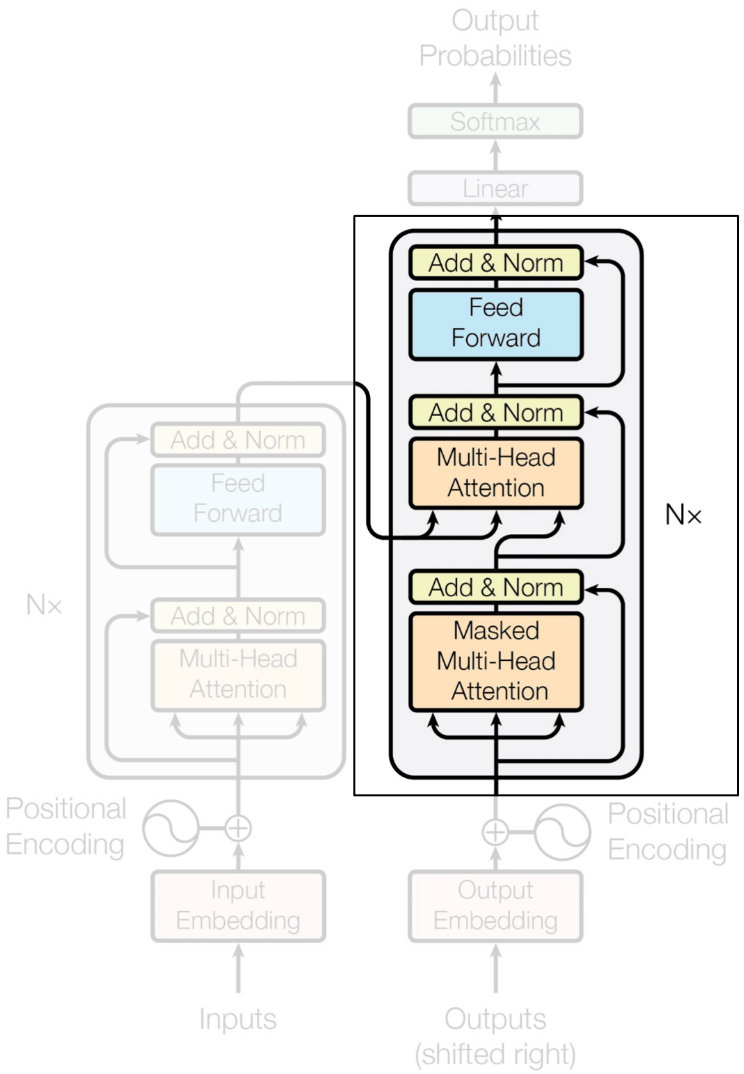
Input

Norm(Z’)



Encoder Decoder Attention

Encoder Decoder Attention ?



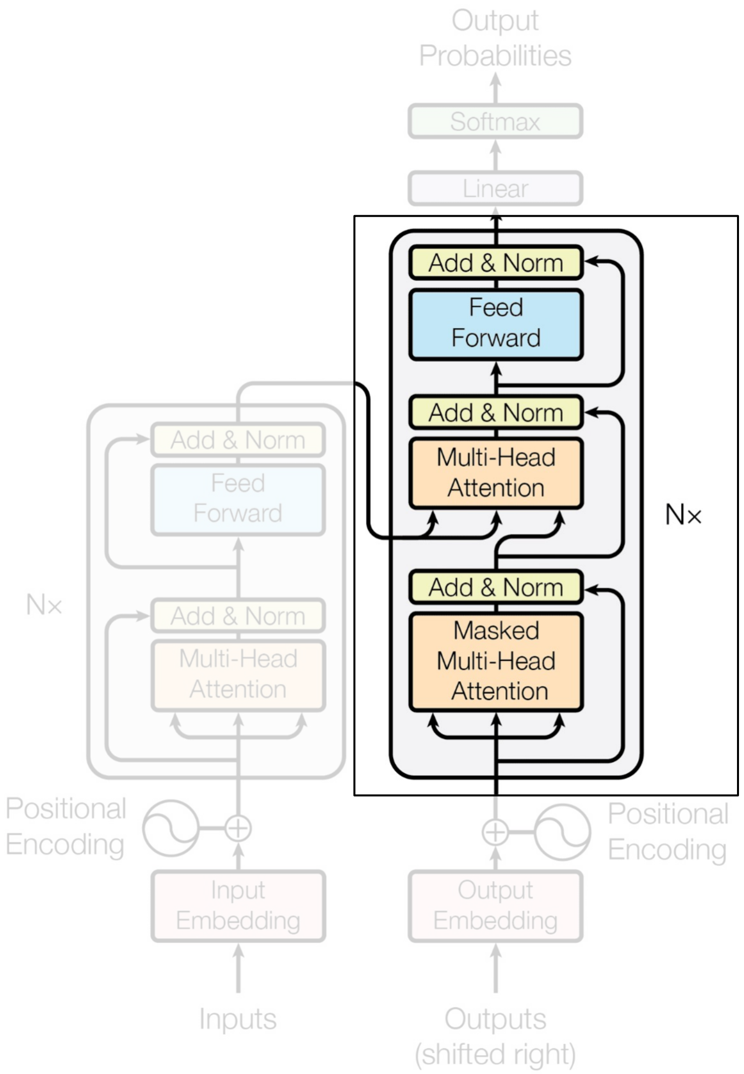
Encoder Decoder Attention

<u>Encoder **Self** Attention</u>

1. Queries from Encoder Inputs
2. Keys from Encoder Inputs
3. Values from Encoder Inputs

<u>Decoder **Masked Self** Attention</u>

1. Queries from Decoder Inputs
2. Keys from Decoder Inputs
3. Values from Decoder Inputs



Attention

{Query: “Order details of order_104”}

{Query: “Order details of order_106”}

{Key, Value store}

```
{"order_100": {"items":"a1", "delivery_date":"a2", ...}},
{"order_101": {"items":"b1", "delivery_date":"b2", ...}},
{"order_102": {"items":"c1", "delivery_date":"c2", ...}},
{"order_103": {"items":"d1", "delivery_date":"d2", ...}},
{"order_104": {"items":"e1", "delivery_date":"e2", ...}},
{"order_105": {"items":"f1", "delivery_date":"f2", ...}},
{"order_106": {"items":"g1", "delivery_date":"g2", ...}},
{"order_107": {"items":"h1", "delivery_date":"h2", ...}},
{"order_108": {"items":"i1", "delivery_date":"i2", ...}},
{"order_109": {"items":"j1", "delivery_date":"j2", ...}},
{"order_110": {"items":"k1", "delivery_date":"k2", ...}}
```

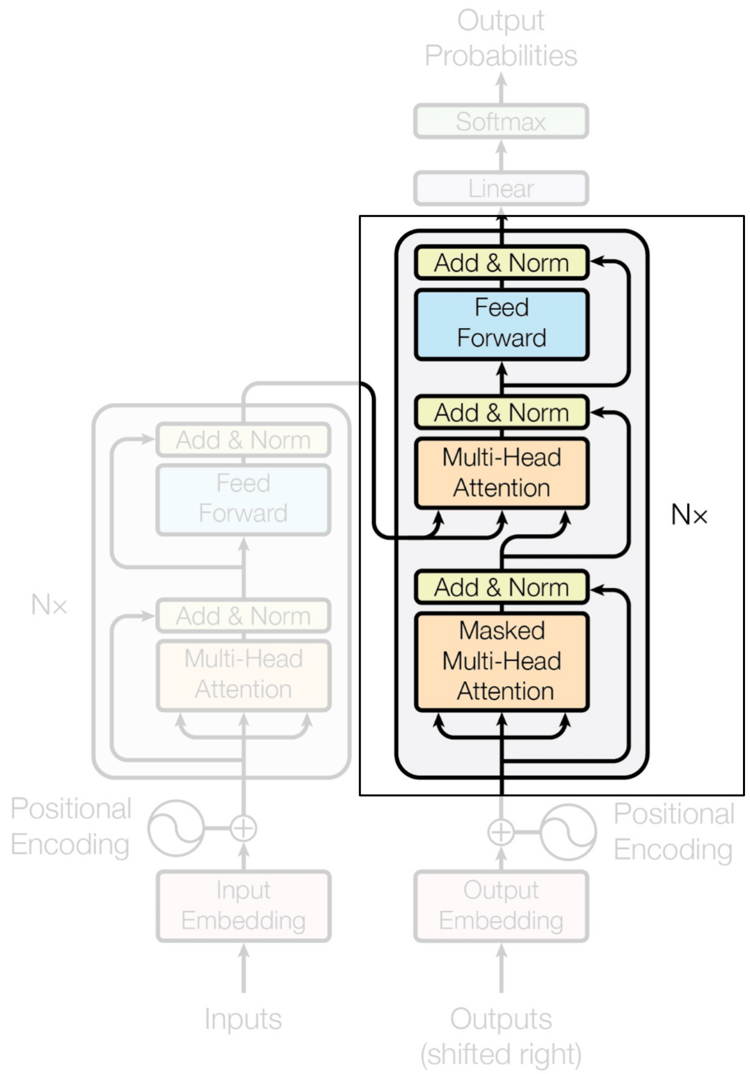
Encoder Decoder Attention

Encoder

Keys from **Encoder Outputs**
Values from **Encoder Outputs**

Decoder

Queries from **Decoder Inputs**



NOTE: Every decoder block receives the same FINAL encoder output

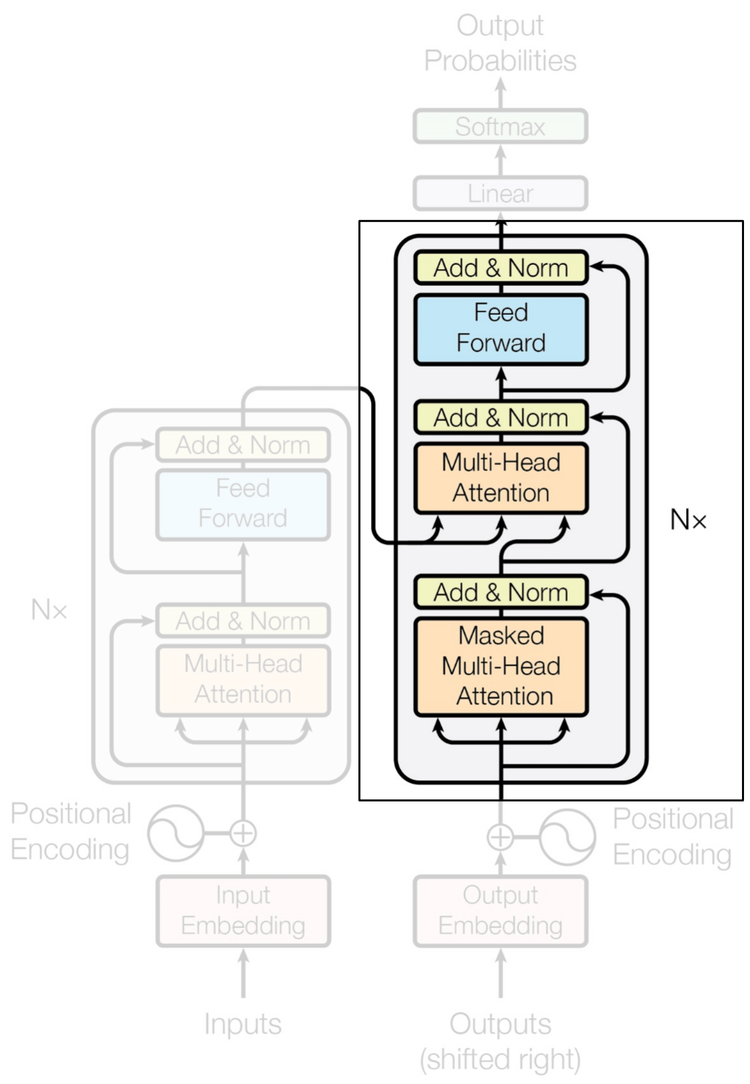
Encoder Decoder Attention

$R^{T_d \times d_{model}}$ Z''

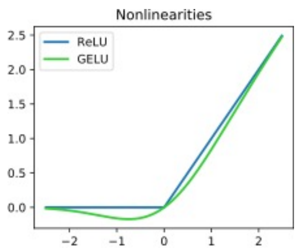
$R^{T_d \times T_e}$ $\text{softmax}(\frac{Q_d K_e{}^T}{\sqrt{d_{model}}})$. V_e $R^{T_e \times d_{model}}$

$R^{T_d \times T_e}$ $\text{softmax}(\frac{Q_d K_e{}^T}{\sqrt{d_{model}}})$

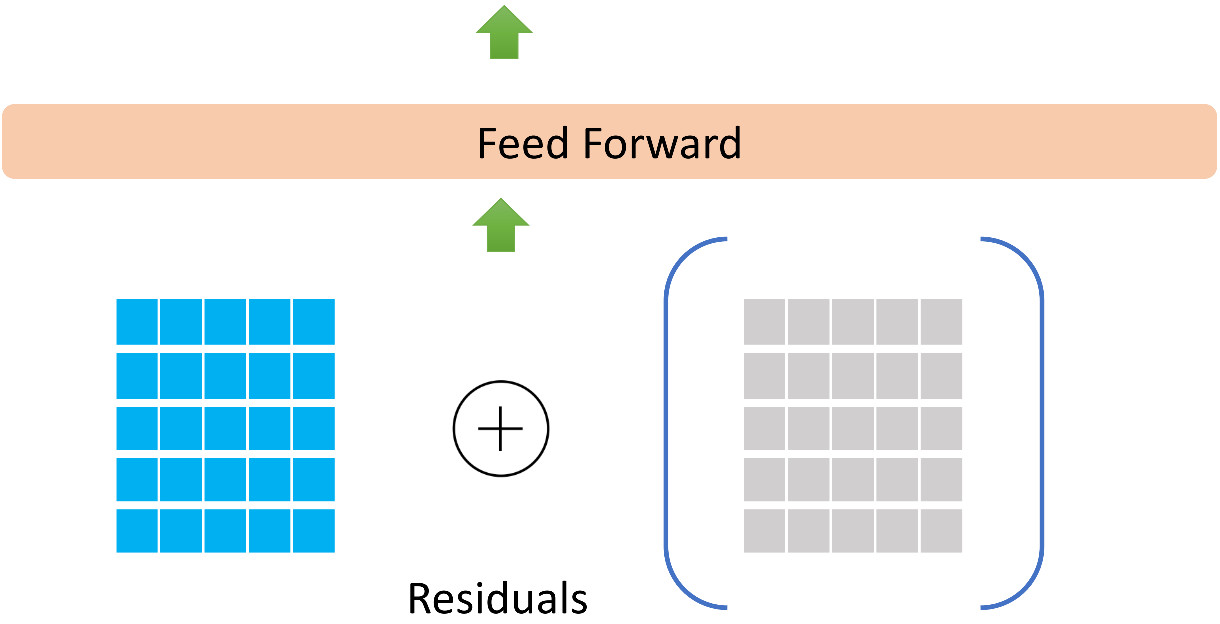
$R^{T_d \times d_{model}}$ Q_d K_e $R^{T_e \times d_{model}}$



Encoder Decoder Attention

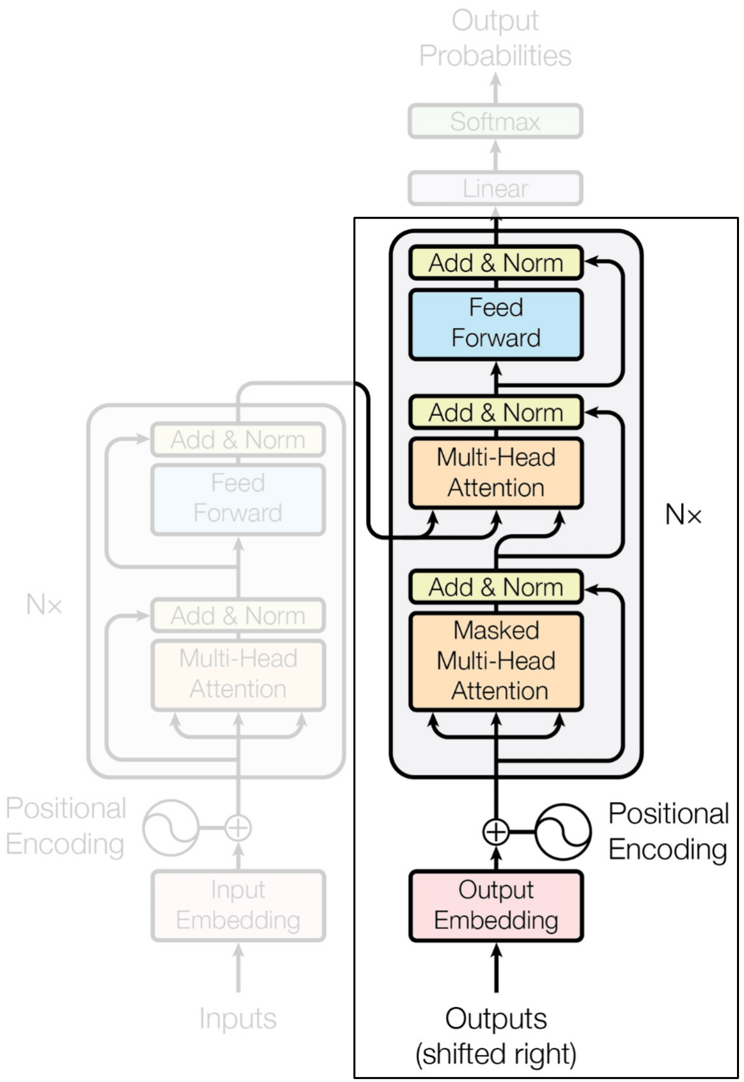


- Non Linearity
- Complex Relationships
- Learn from each other



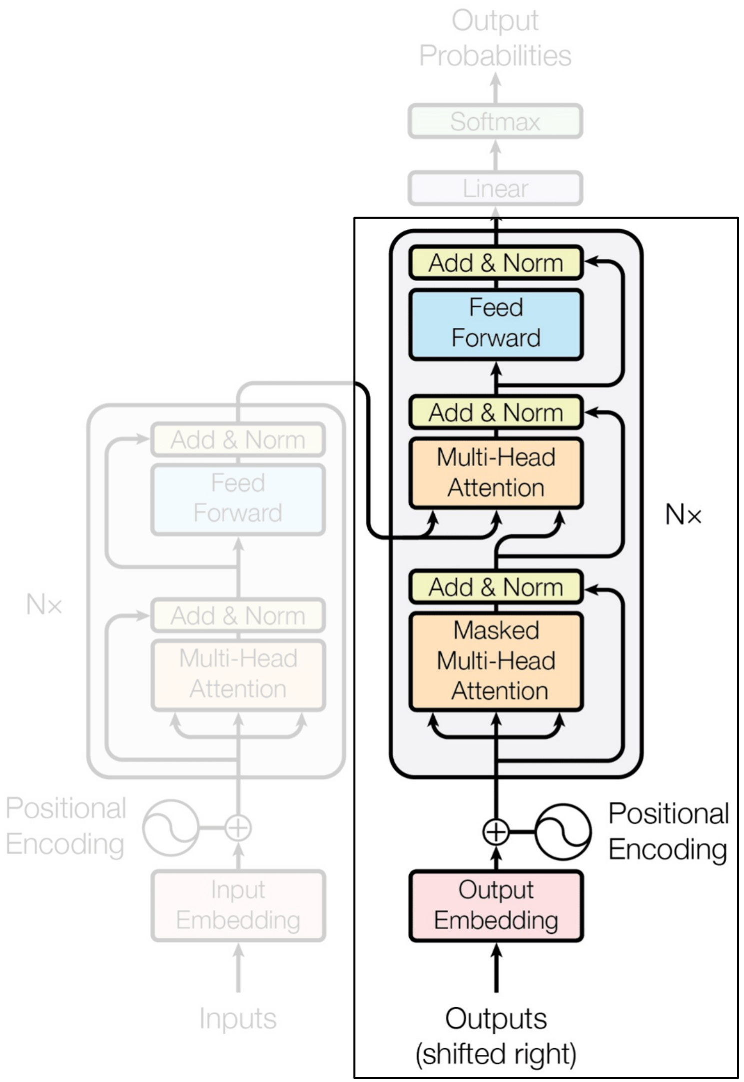
Add n Norm Decoder Self Attn

Norm(Z'')

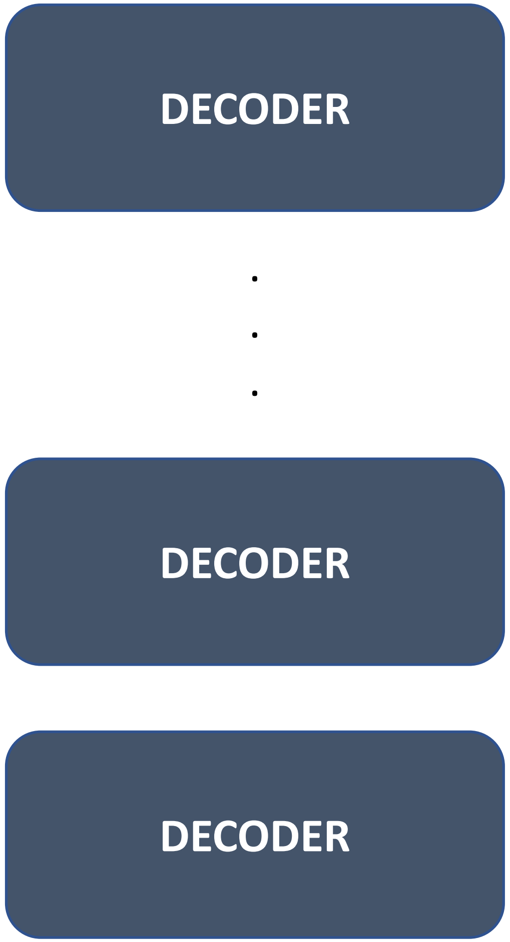


Decoder

DECODER

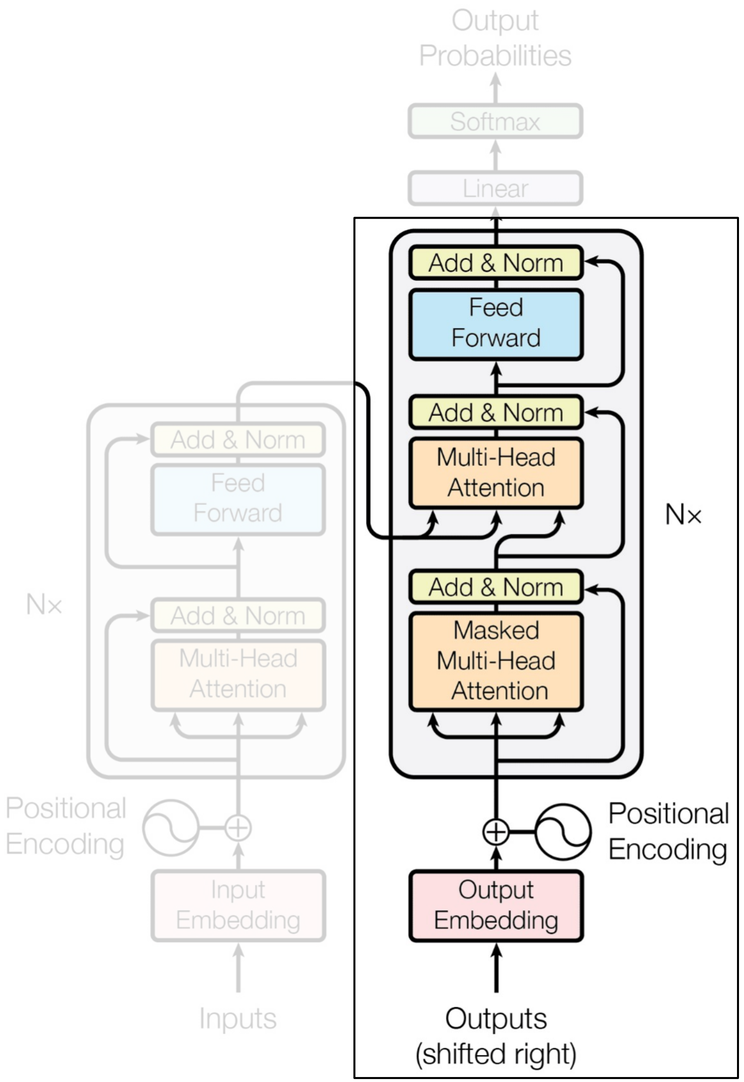


Decoder



$R^{T_d \times d_{model}}$

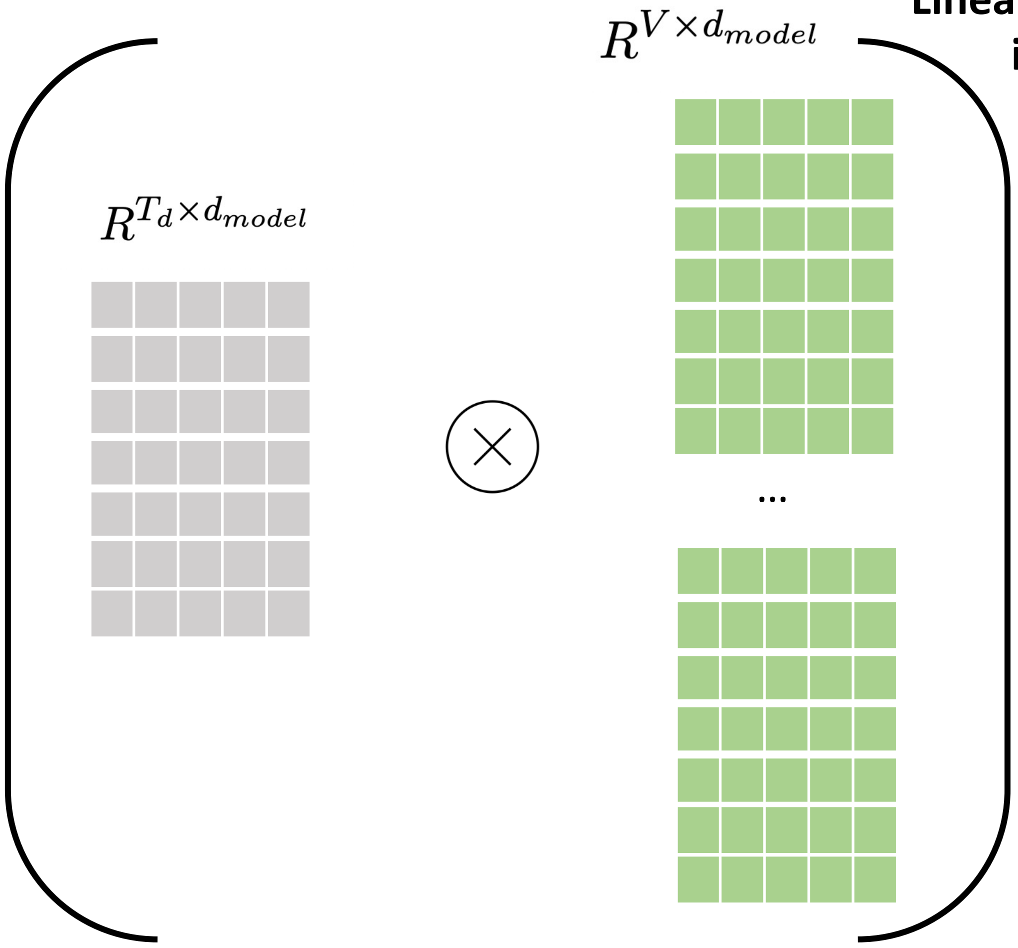
Decoder output



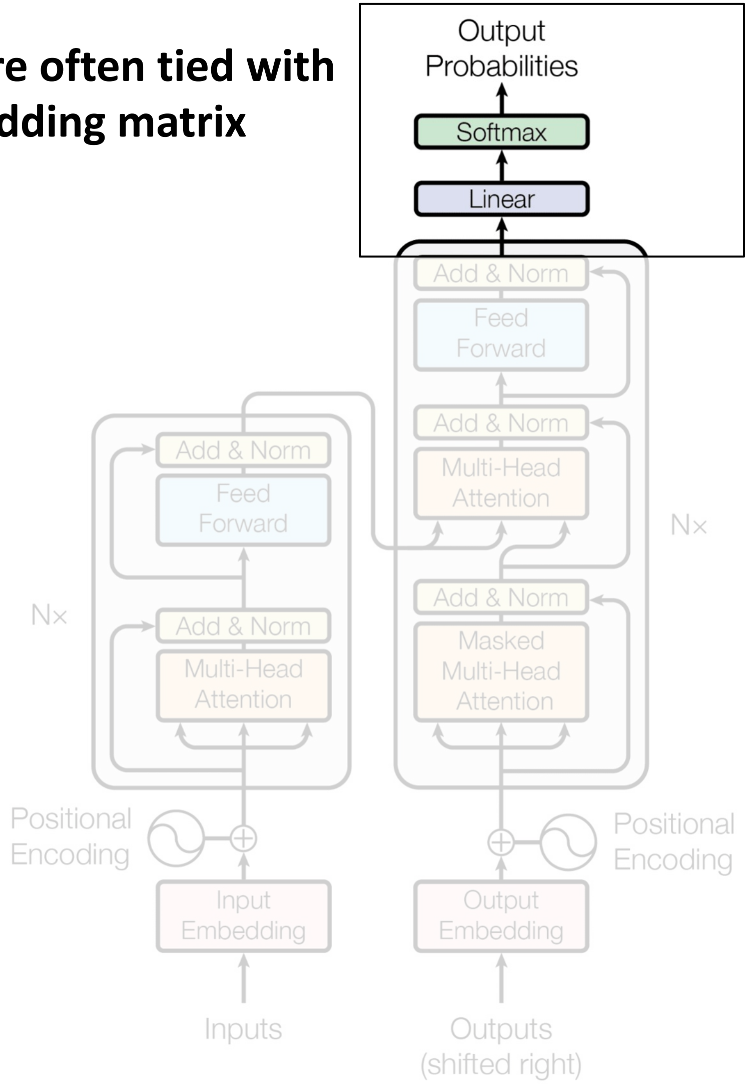
Linear

Linear weights are often tied with input embedding matrix

softmax

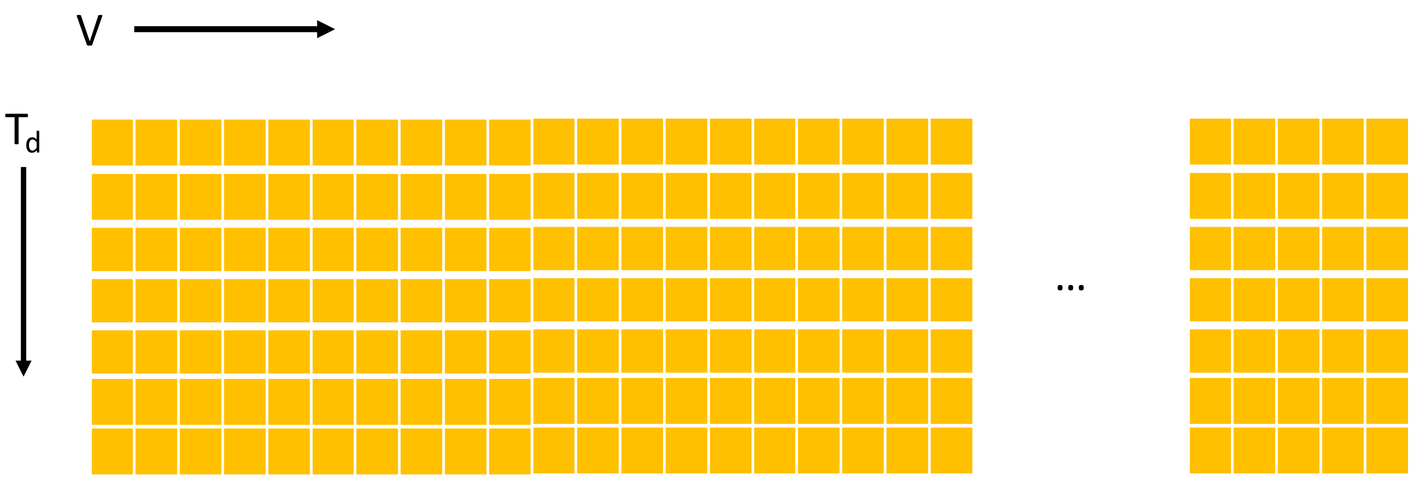


Final Decoder Output $\qquad R^{T_d \times V} \qquad$ Linear

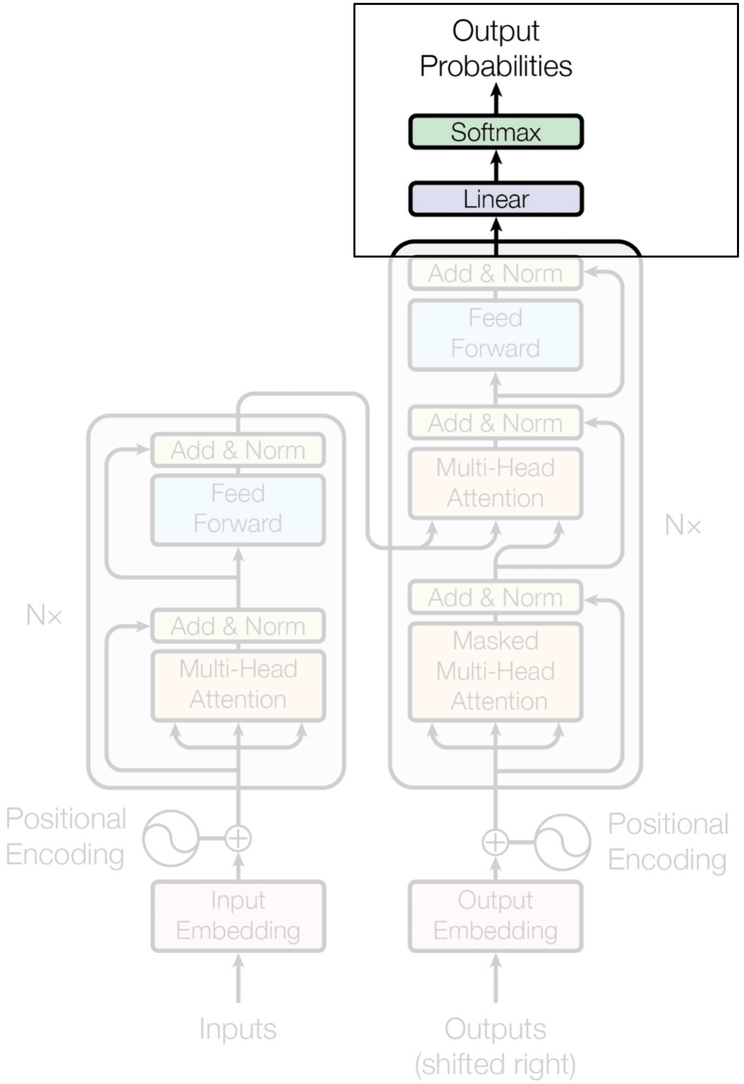


Softmax

Output Probabilities



$$R^{T_d \times V}$$



Poll 2 (@1297)

Which of the following are true about transformers?

a. Transformers can always be run in parallel
b. Transformer decoders can only be parallelized during training
c. Positional encodings help parallelize the transformer encoder
d. Queries, keys, and values are obtained by splitting the input into 3 equal segments
e. Multiheaded attention helps transformers find different kinds of relations between the tokens
f. During decoding, decoder outputs function as queries and keys while the values come from the encoder

Poll 2 (@1126)

Which of the following are true about transformers?

a. Transformers can always be run in parallel

b. Transformer decoders can only be parallelized during training

c. Positional encodings help parallelize the transformer encoder

d. Queries, keys, and values are obtained by splitting the input into 3 equal segments

e. Multiheaded attention helps transformers find different kinds of relations between the tokens

f. During decoding, decoder outputs function as queries and keys while the values come from the encoder

Transformers

Targets

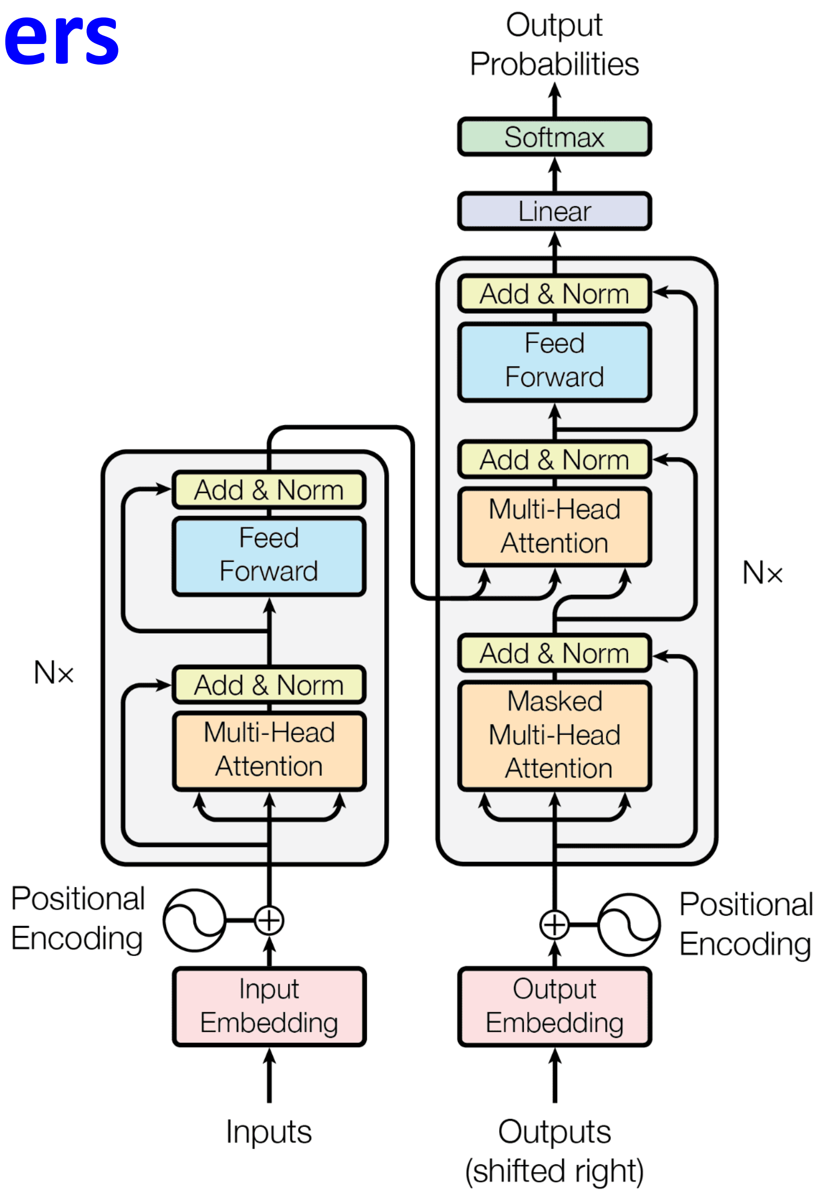
Ich have einen apfel gegessen

⬆

Inputs

I ate an apple

Machine Translation



Transformers

- ✓ Tokenizaton
- ✓ Input Embeddings
- ✓ Position Encodings
- ✓ Residuals
- ✓ Query
- ✓ Key
- ✓ Value
- ✓ Add & Norm
- ✓ Encoder
- ✓ Decoder

- ✓ Attention
- ✓ Self Attention
- ✓ Multi Head Attention
- ✓ Masked Attention
- ✓ Encoder Decoder Attention
- ✓ Output Probabilities / Logits
- ✓ Softmax
- **Encoder-Decoder models**
- **Decoder only models**