Attention is All You Need: From Words to Worlds

Muhan Zhang

April 26, 2024, Group Seminar

University of North Texas

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez*† University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* illia.polosukhin@gmail.com

Outline

- Background
- Highlights of Transformer Model
- Self-Attention and Multi-Head Attention
- Model Architecture
- Experimental Results
- Summary

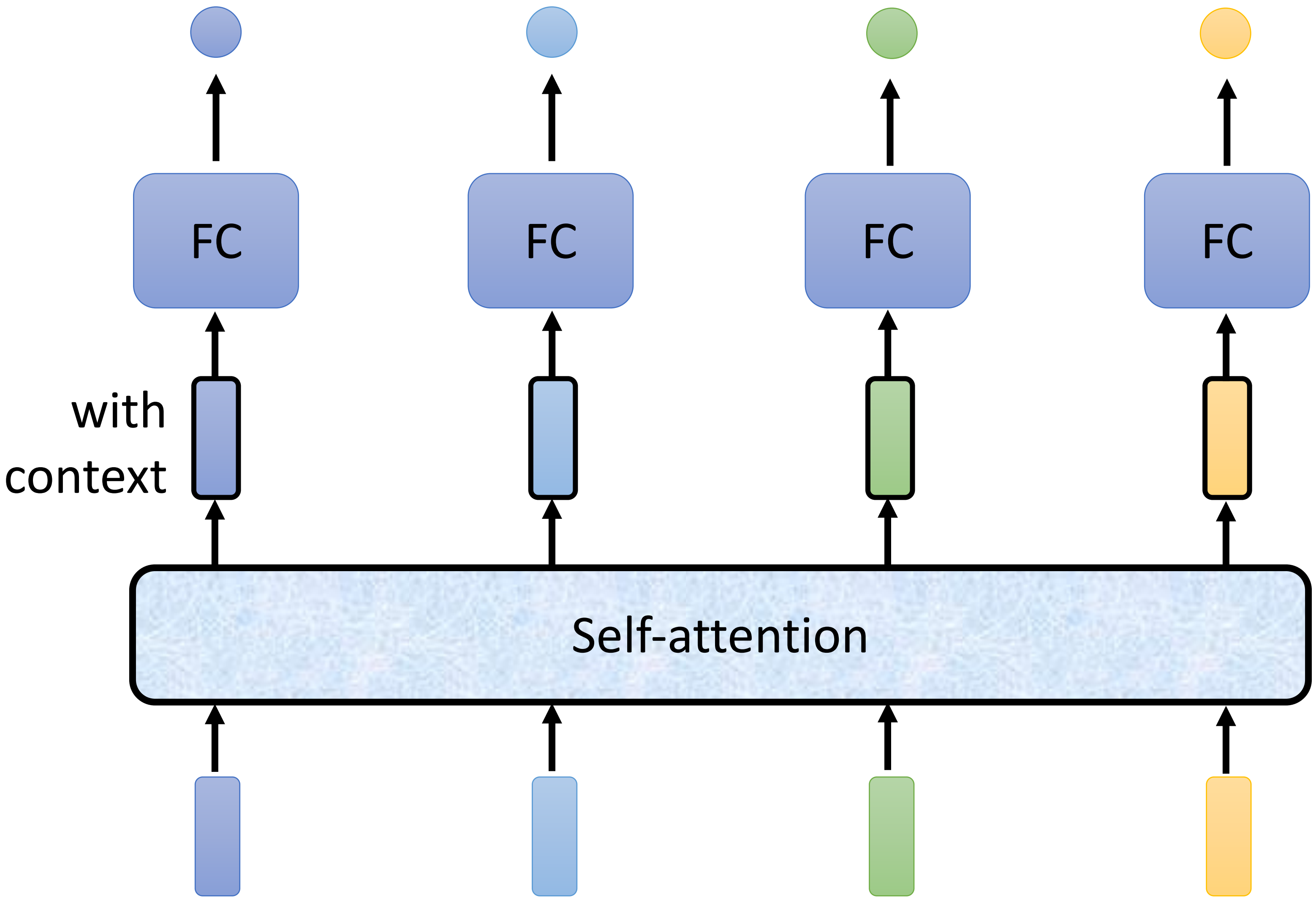
Background:

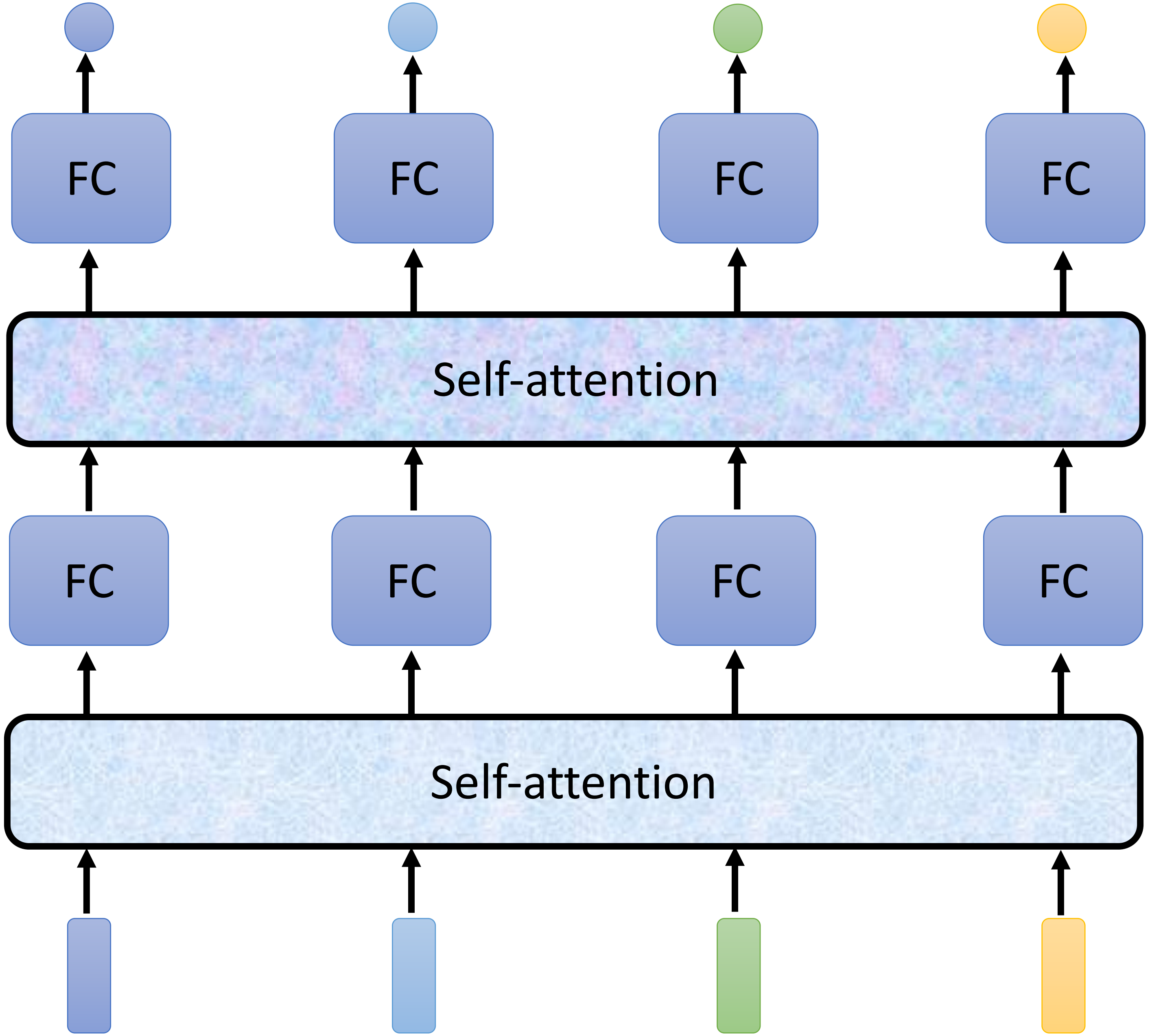
- Sequence transduction models have traditionally relied on complex CNN or RNN architectures
- CNNs have difficulty capturing long-range dependencies
- RNNs suffer from limited computational parallelism and slow training
- There is a need for more efficient and parallelizable models

Highlights of Transformer Model:

- Entirely based on attention mechanism, without using convolution or recurrence
- Highly parallelizable computation, allowing for very fast training
- Surpasses previous models (including ensembles) on machine translation tasks
- Generalizes well to other tasks such as English constituency parsing

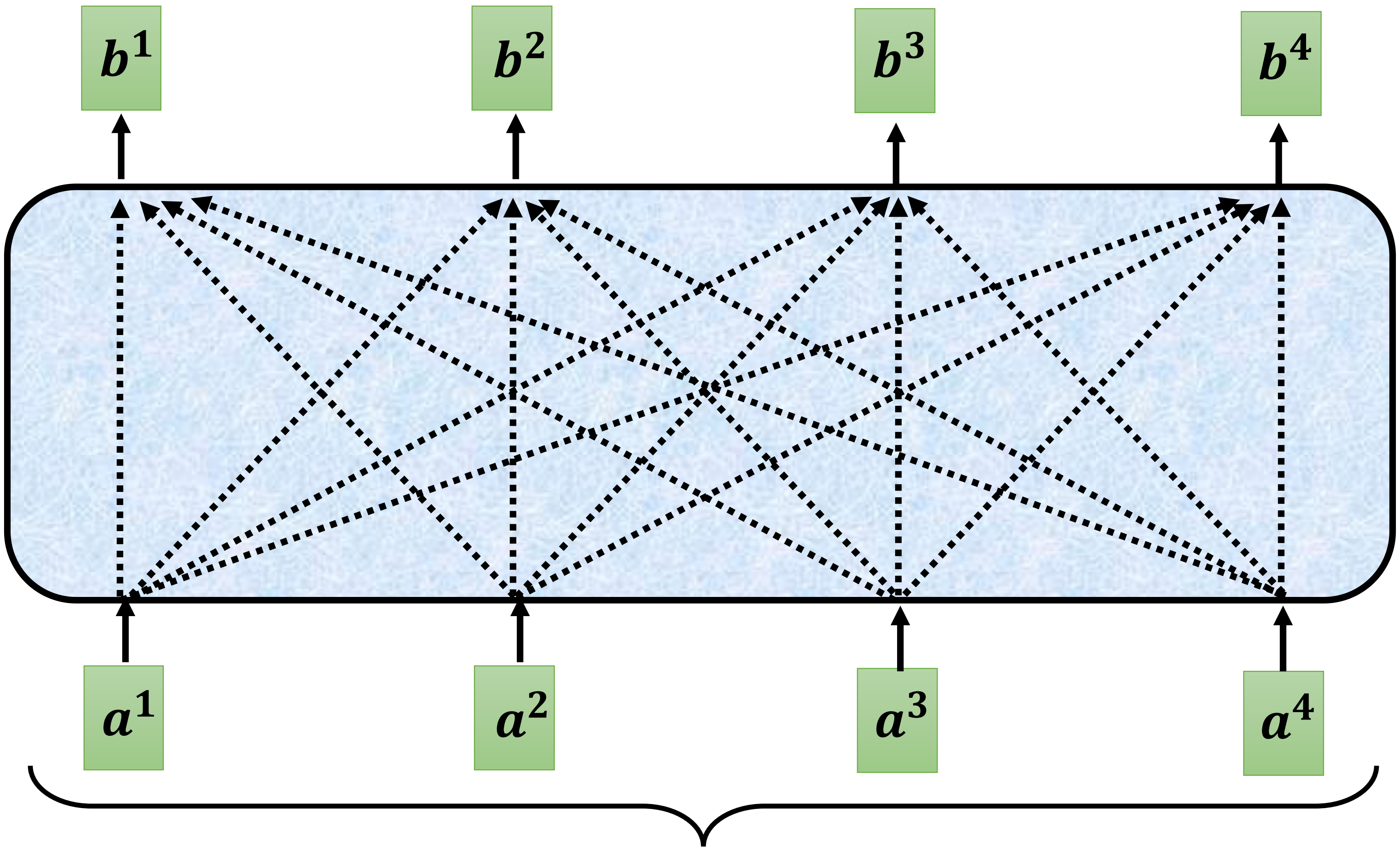
Self-Attention and Multi-Head Attention:





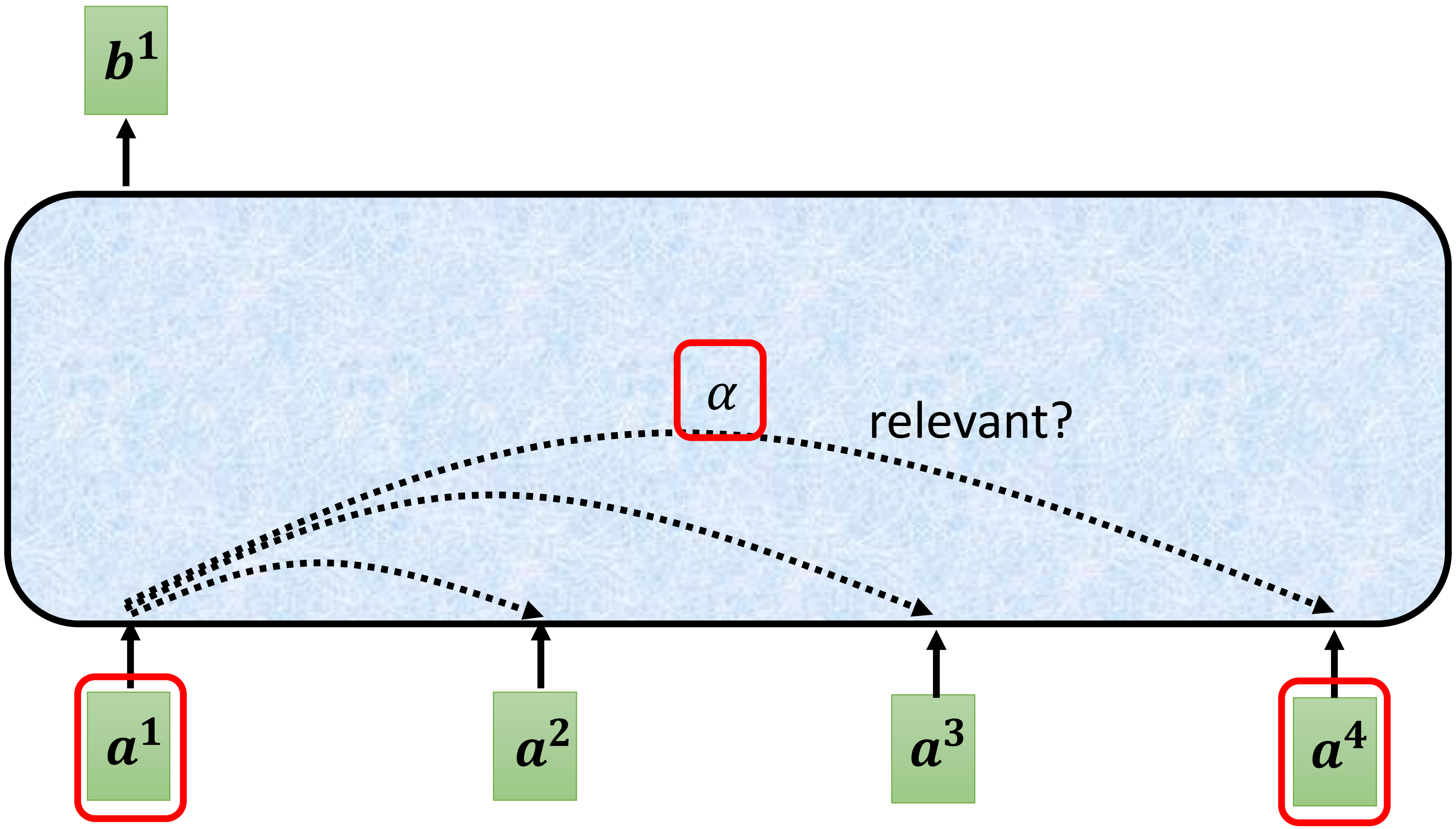
https://arxiv.org/abs/1706.03762

Self-attention



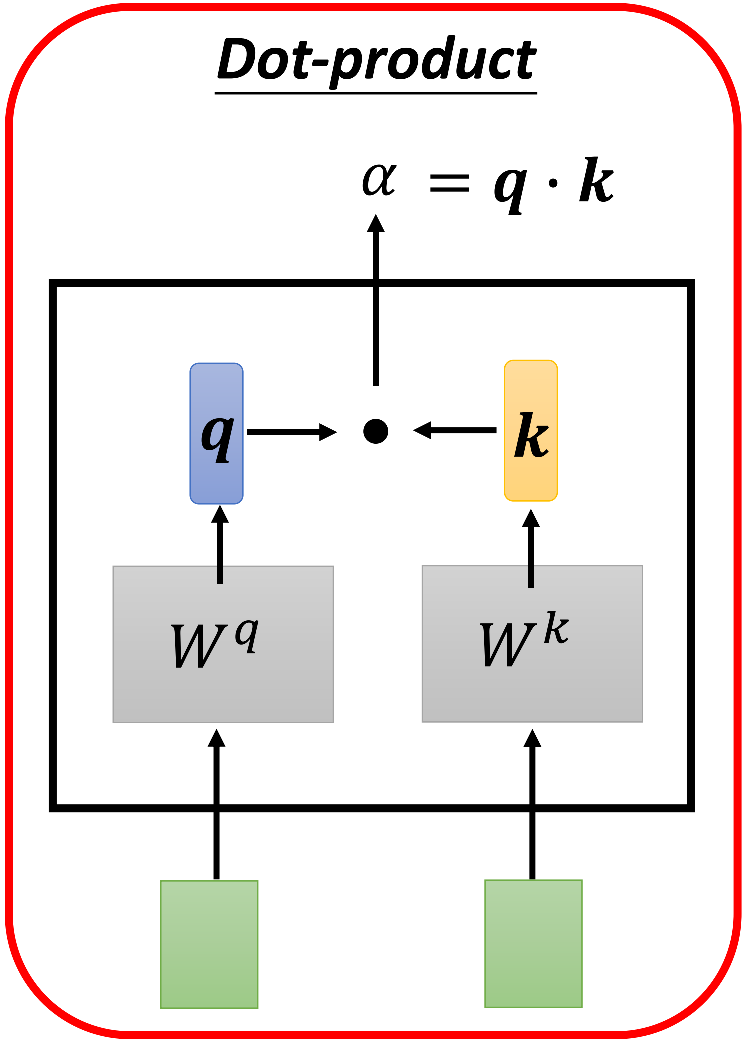
Can be either **input** or **a hidden layer**

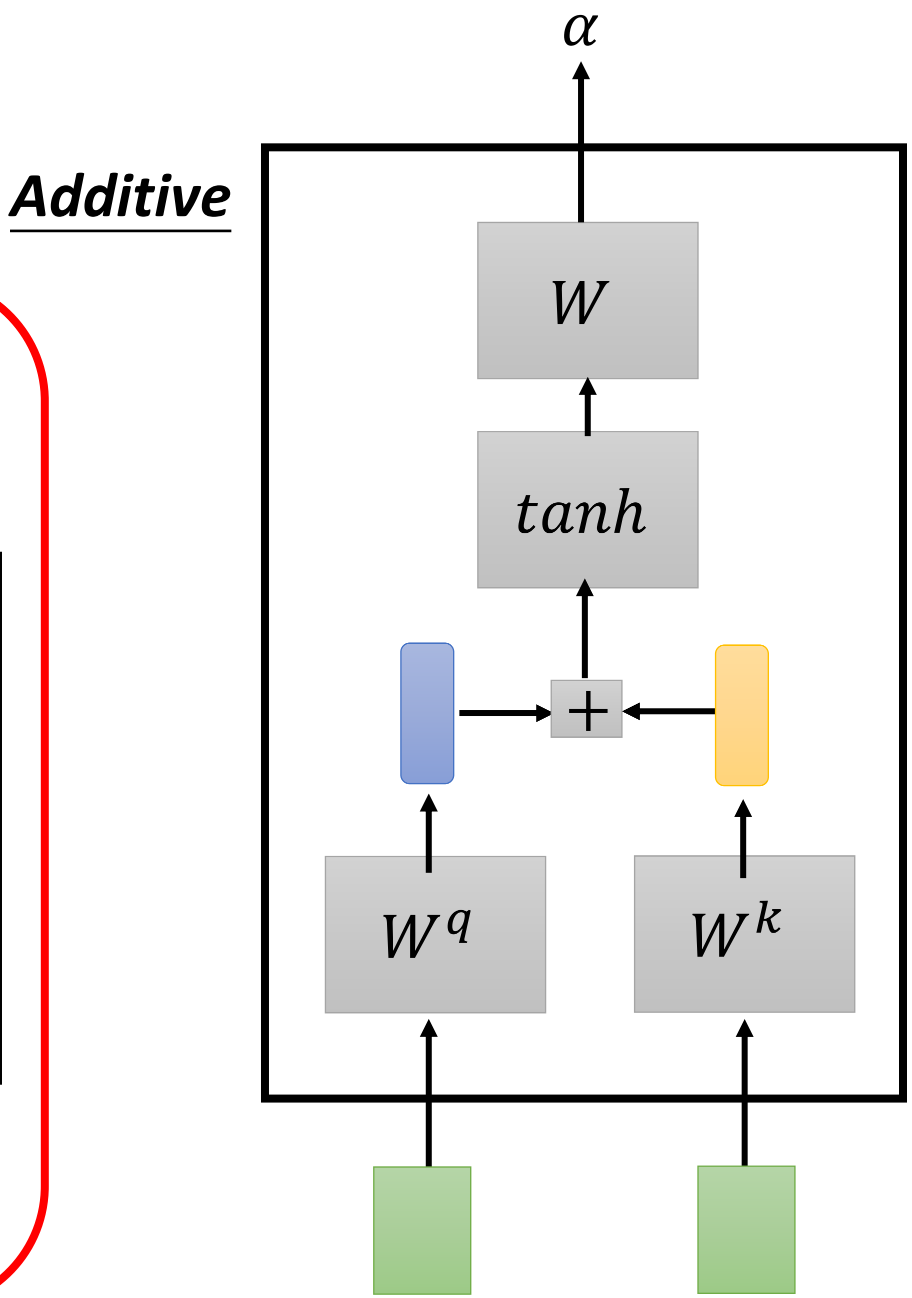
Self-attention



Find the relevant vectors in a sequence

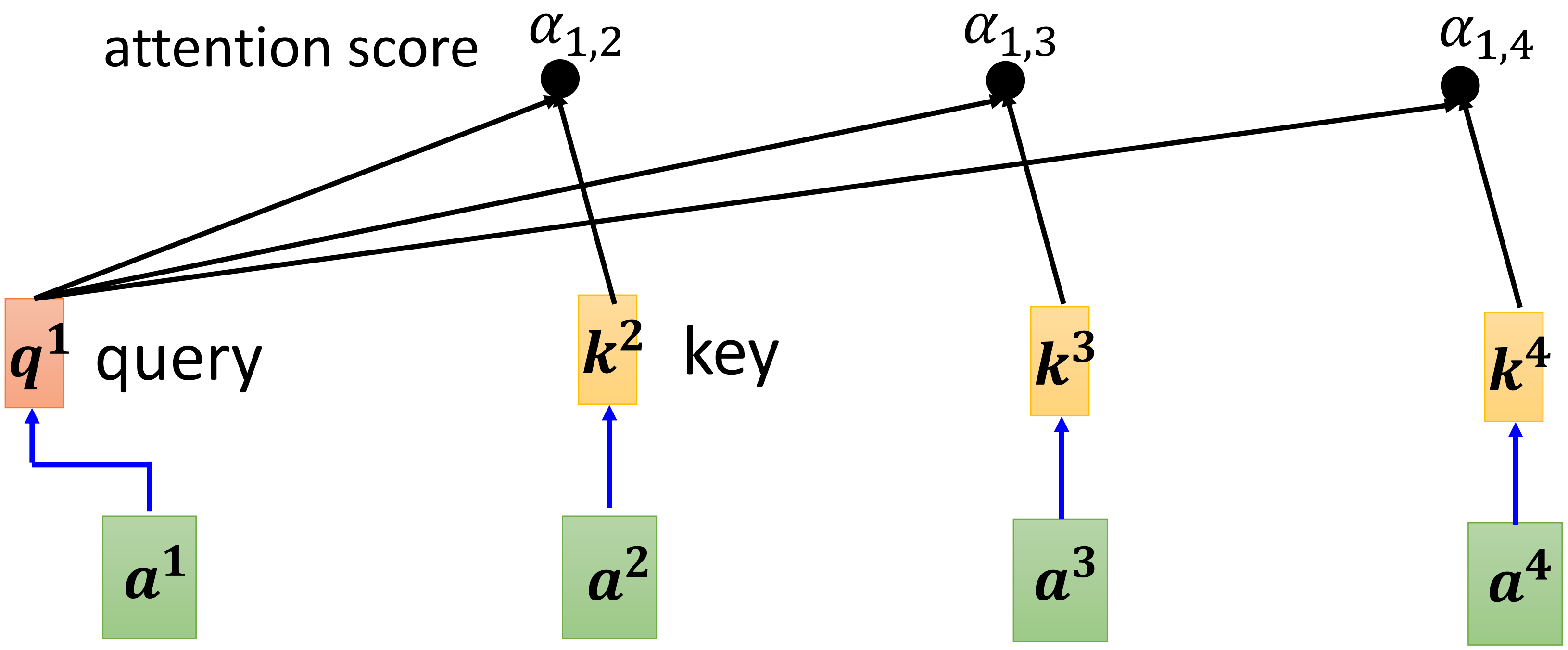
Self-attention





Self-attention

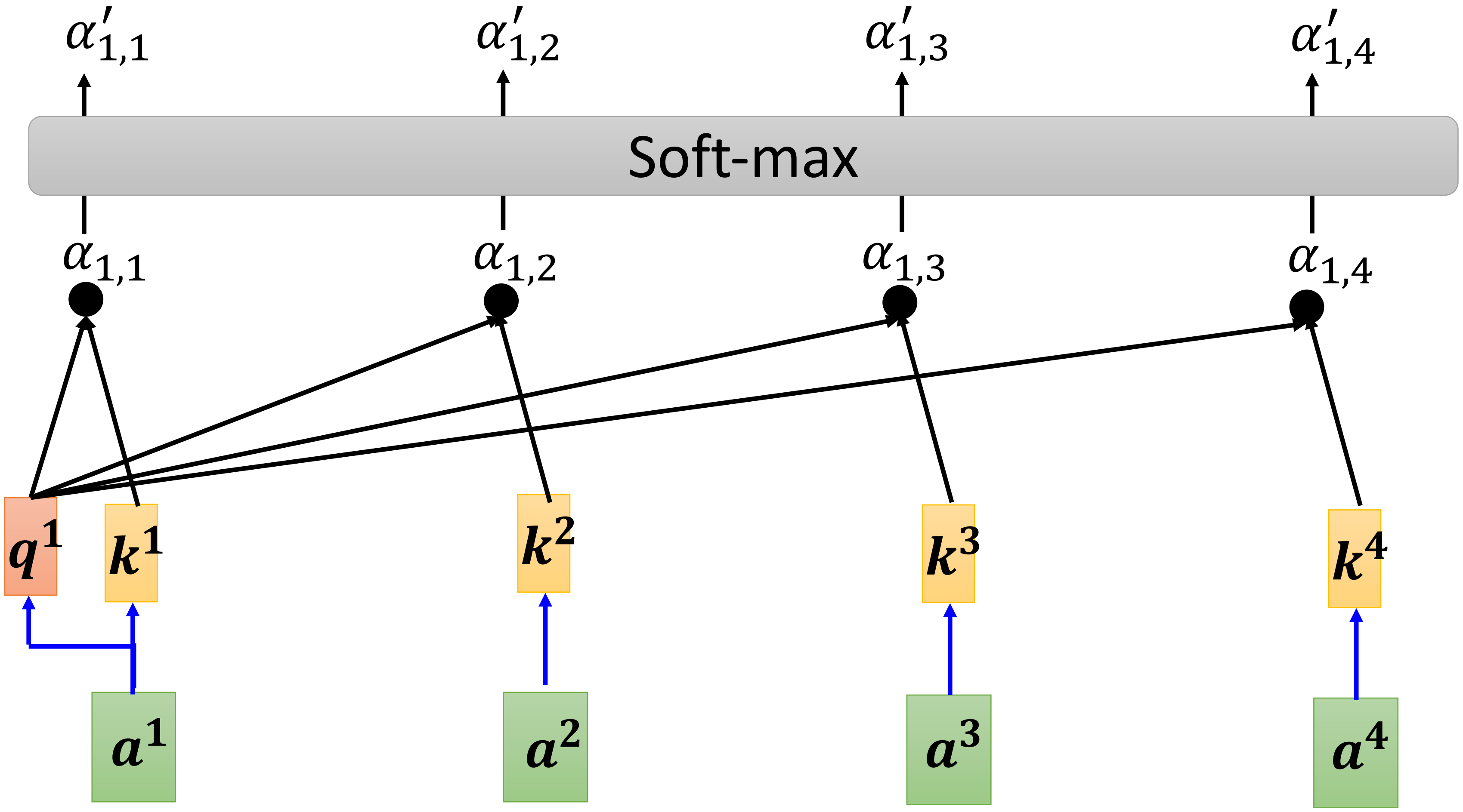
$\alpha_{1,2} = \boldsymbol{q}^1 \cdot \boldsymbol{k}^2$ $\quad$ $\alpha_{1,3} = \boldsymbol{q}^1 \cdot \boldsymbol{k}^3$ $\quad$ $\alpha_{1,4} = \boldsymbol{q}^1 \cdot \boldsymbol{k}^4$



$\boldsymbol{q}^1 = W^q \boldsymbol{a}^1$ $\quad$ $\boldsymbol{k}^2 = W^k \boldsymbol{a}^2$ $\quad$ $\boldsymbol{k}^3 = W^k \boldsymbol{a}^3$ $\quad$ $\boldsymbol{k}^4 = W^k \boldsymbol{a}^4$

Self-attention

$$\hat{\alpha}_{1,i} = exp(\alpha_{1,i})/\sum_j exp(\alpha_{1,j})$$



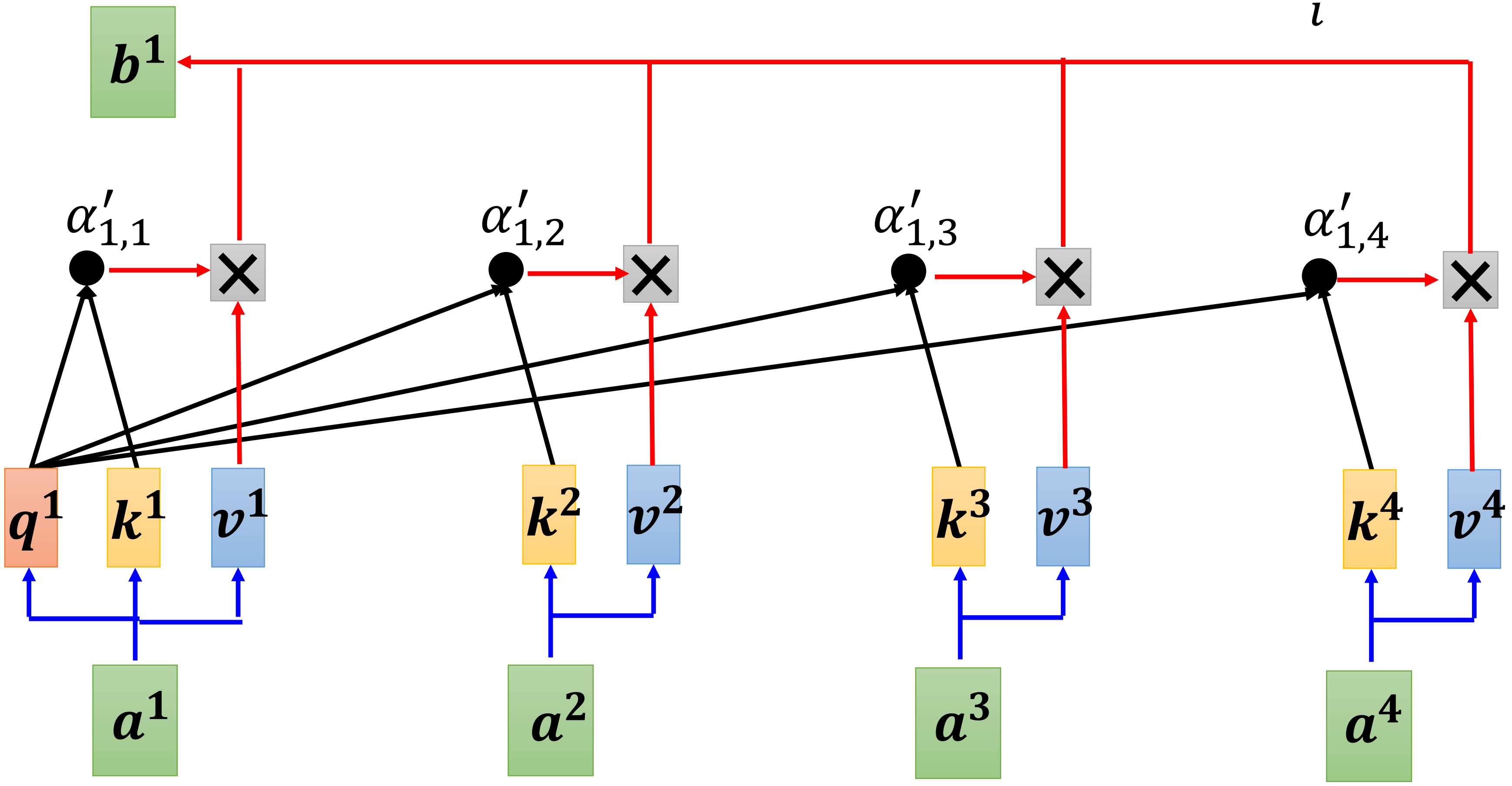
$\boldsymbol{q}^1 = W^q \boldsymbol{a}^1$ $\quad \boldsymbol{k}^2 = W^k \boldsymbol{a}^2$ $\quad \boldsymbol{k}^3 = W^k \boldsymbol{a}^3$ $\quad \boldsymbol{k}^4 = W^k \boldsymbol{a}^4$

$\boldsymbol{k}^1 = W^k \boldsymbol{a}^1$

Self-attention

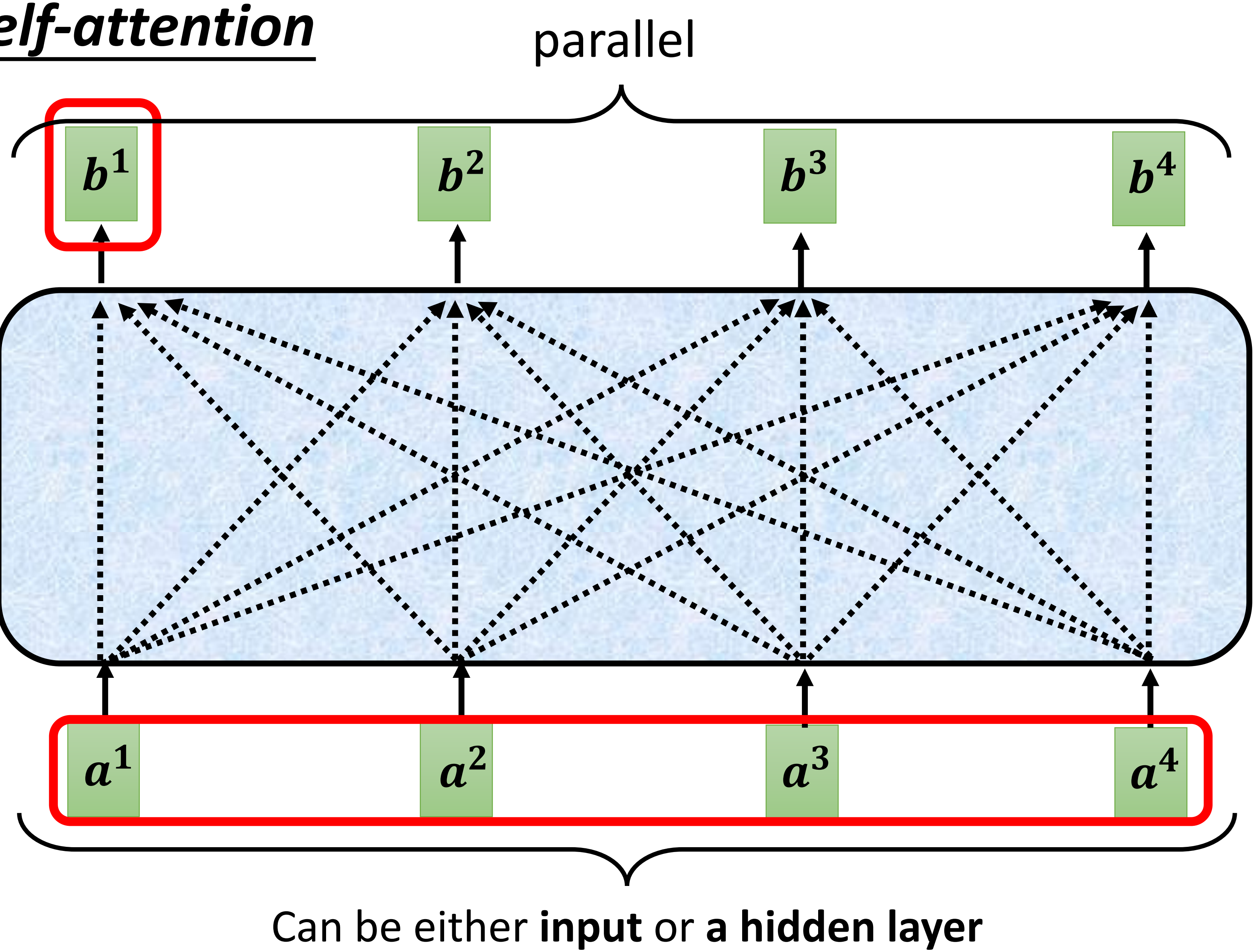
Extract information based on attention scores

$$\boldsymbol{b}^1 = \sum_i \alpha'_{1,i} \boldsymbol{v}^i$$



$\boldsymbol{v}^1 = W^v \boldsymbol{a}^1$ $\boldsymbol{v}^2 = W^v \boldsymbol{a}^2$ $\boldsymbol{v}^3 = W^v \boldsymbol{a}^3$ $\boldsymbol{v}^4 = W^v \boldsymbol{a}^4$

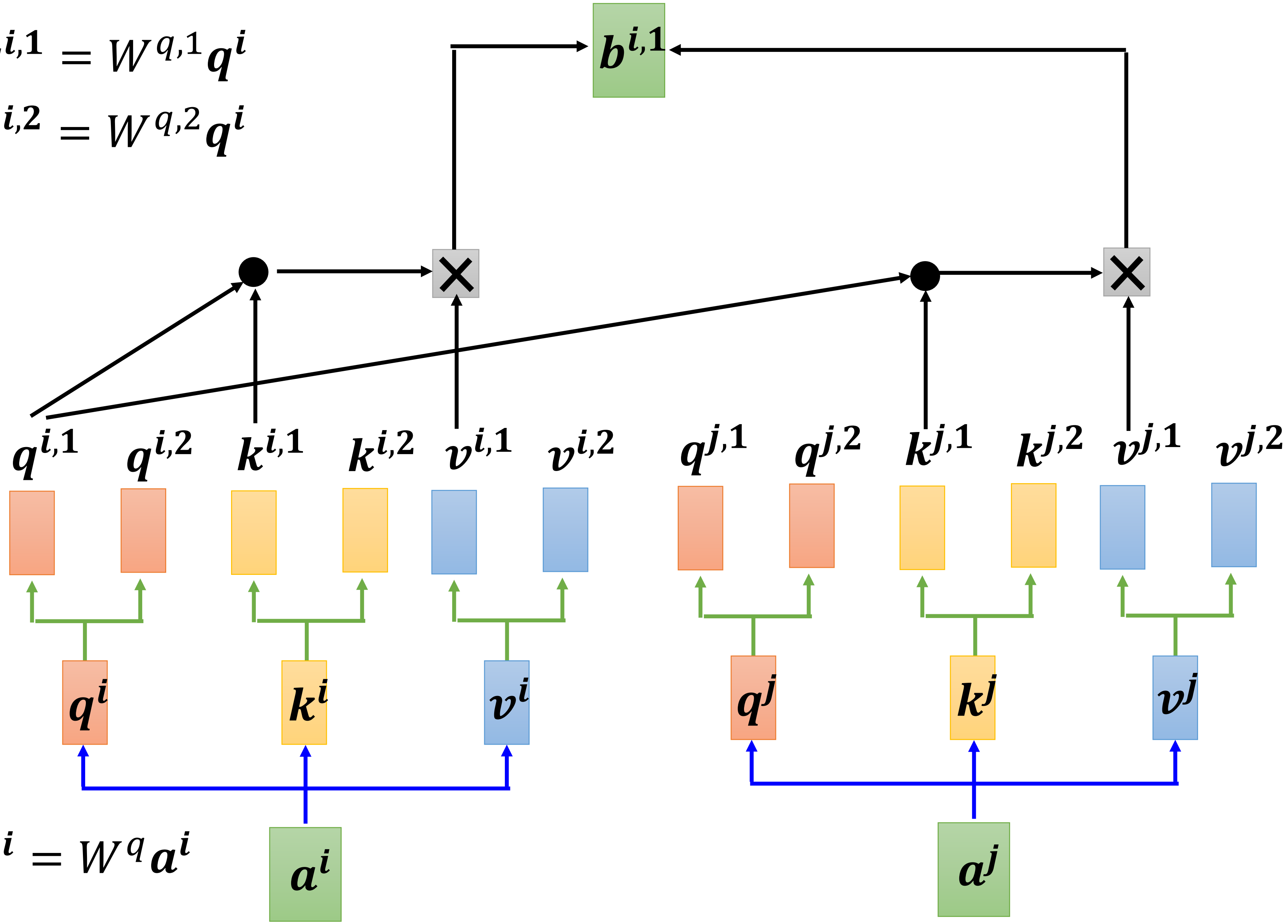
Self-attention



Multi-head Self-attention (2 heads as example)

$$q^{i,1} = W^{q,1} q^i$$

$$q^{i,2} = W^{q,2} q^i$$



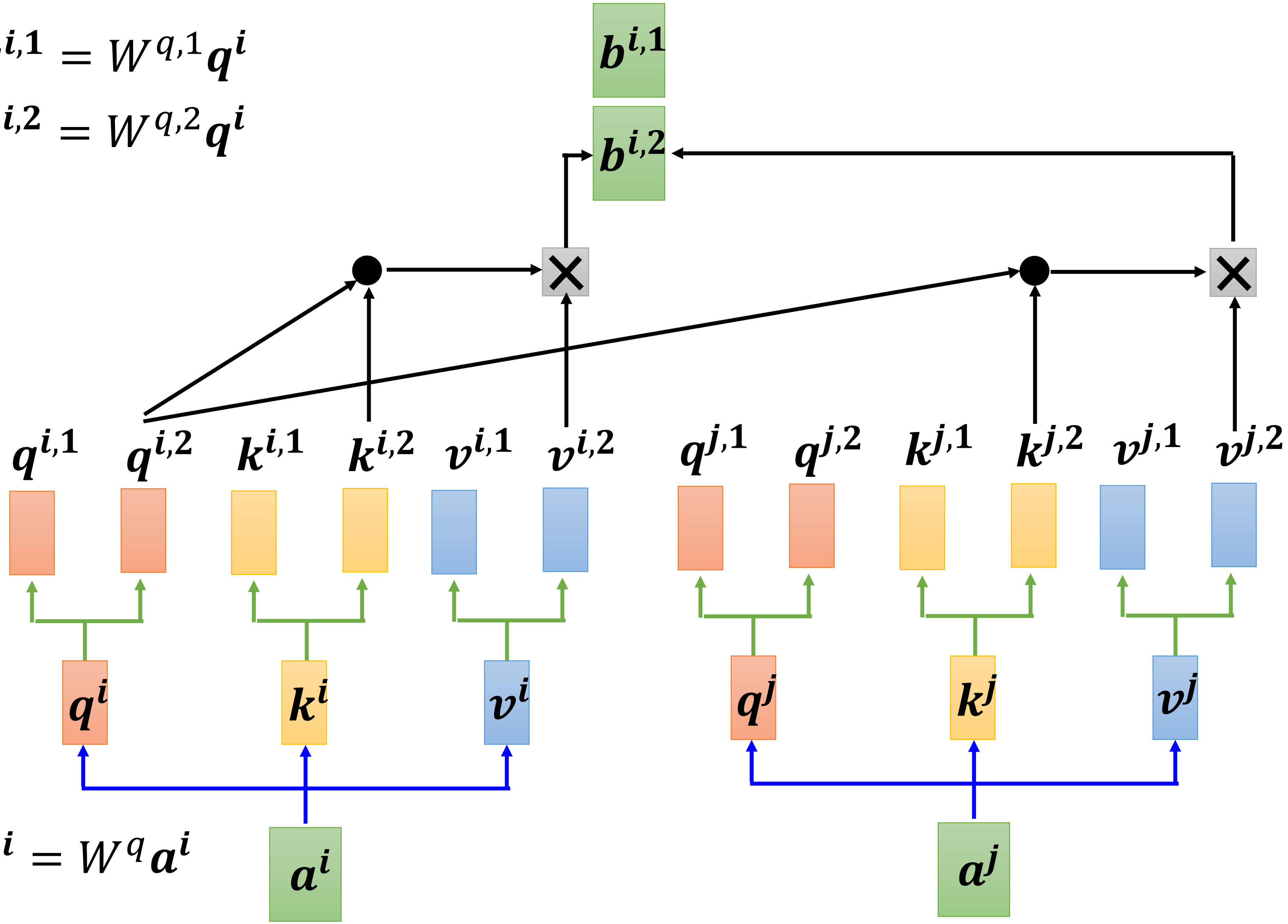
$$q^i = W^q a^i$$

Multi-head Self-attention

(2 heads as example)

$$q^{i,1} = W^{q,1} q^i$$

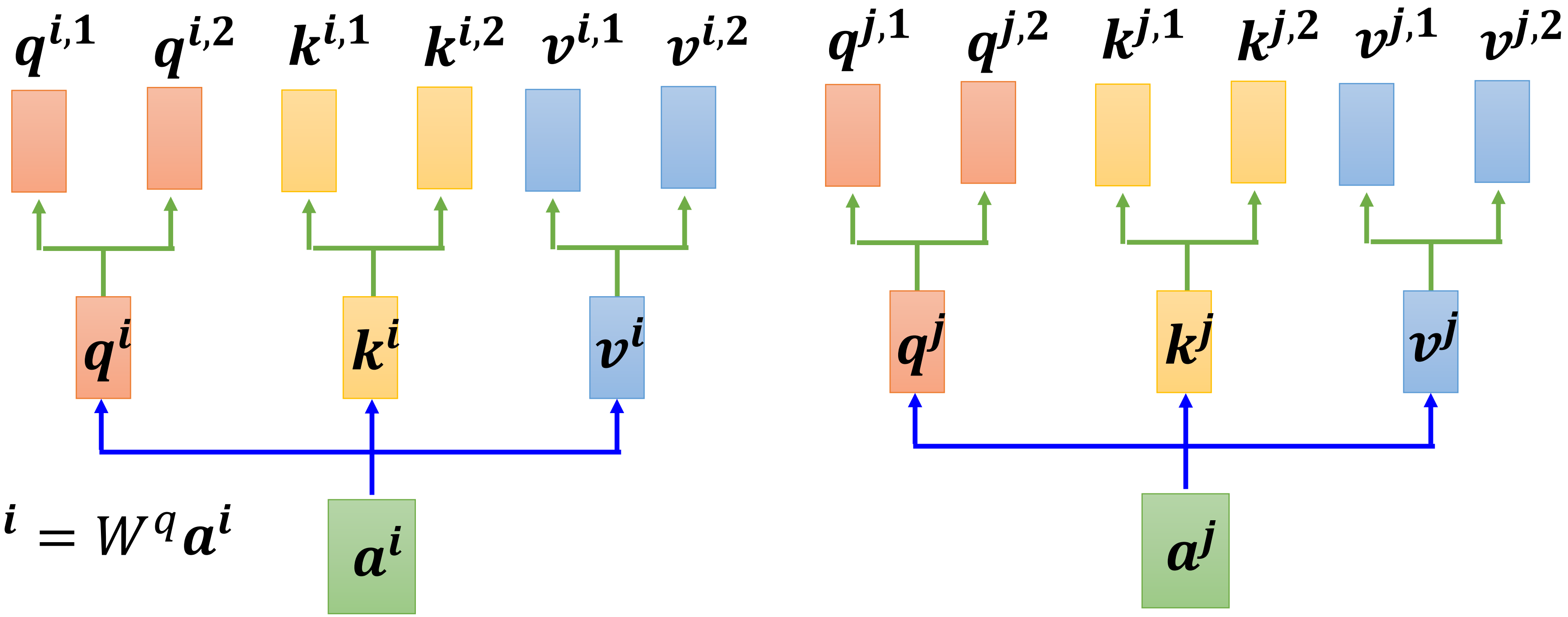
$$q^{i,2} = W^{q,2} q^i$$



$$q^i = W^q a^i$$

Multi-head Self-attention (2 heads as example)

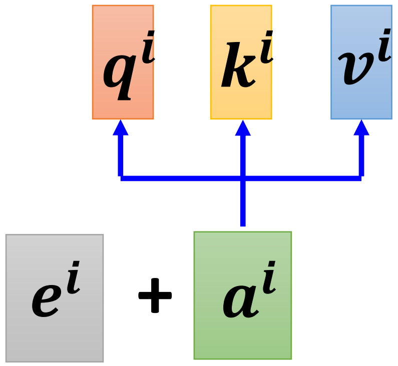
$$\boldsymbol{b}^i = W^O \begin{bmatrix} \boldsymbol{b}^{i,1} \\ \boldsymbol{b}^{i,2} \end{bmatrix}$$



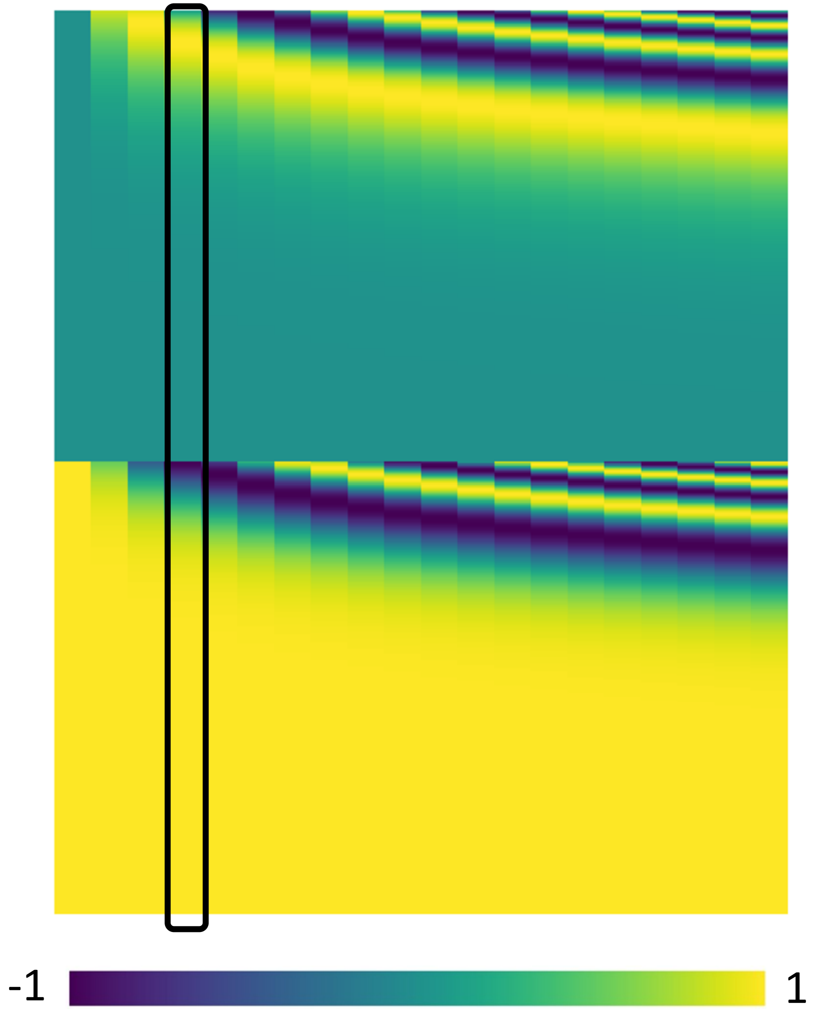
$q^i = W^q a^i$

Positional Encoding

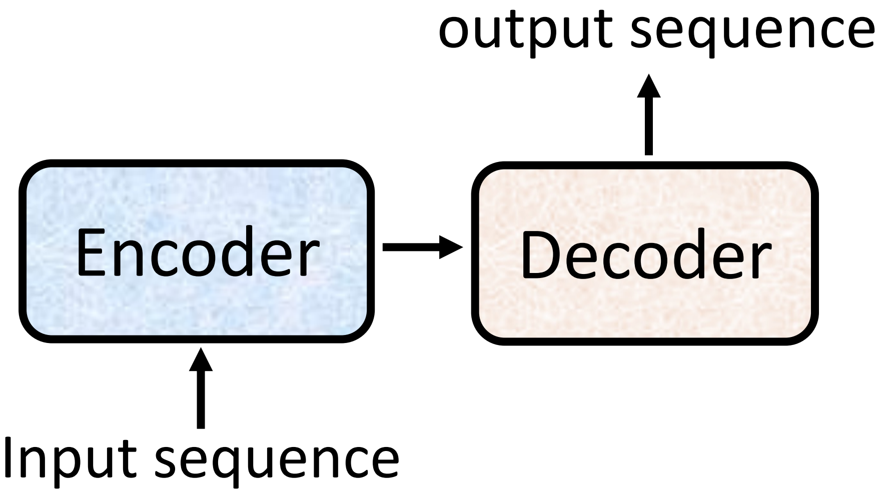
- No position information in self-attention.
- Each position has a unique positional vector e^i
- **hand-crafted**
- **learned from data**



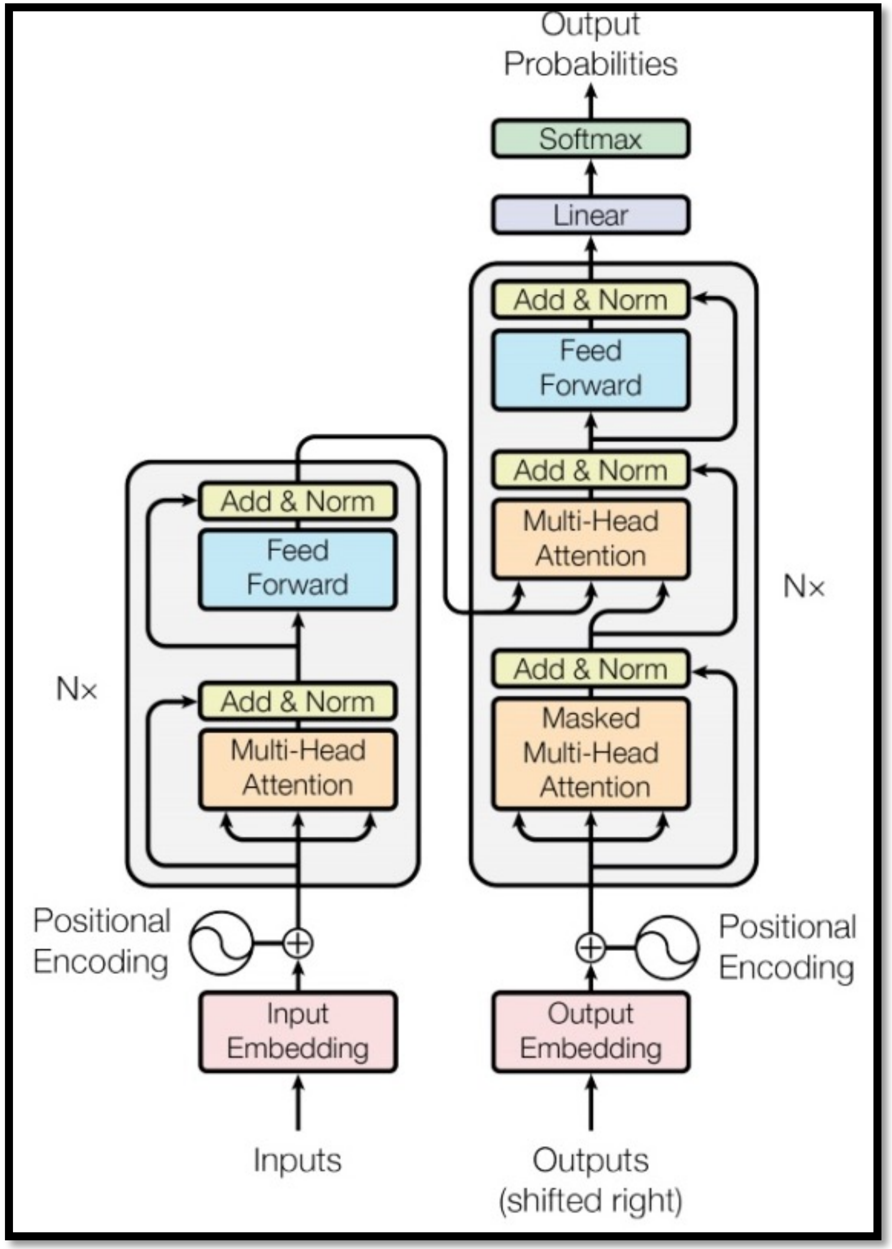
Each column represents a positional vector e^i



Model Architecture

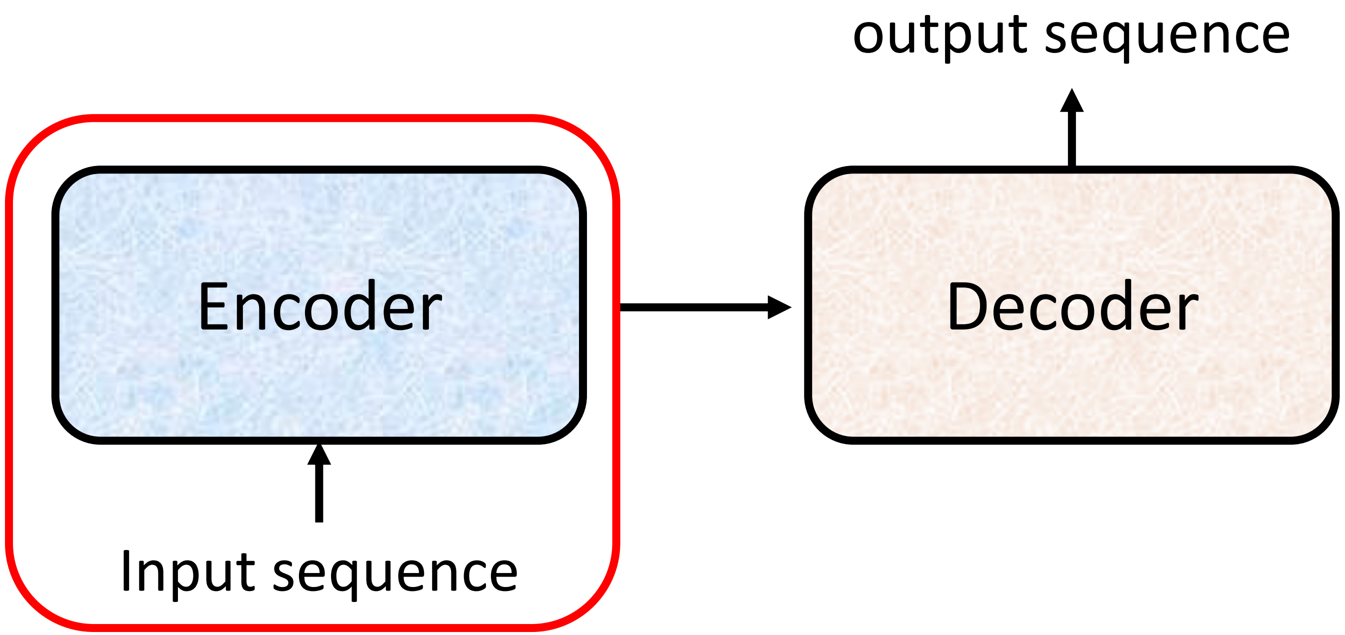


Sequence to Sequence Learning
with Neural Networks
https://arxiv.org/abs/1409.3215



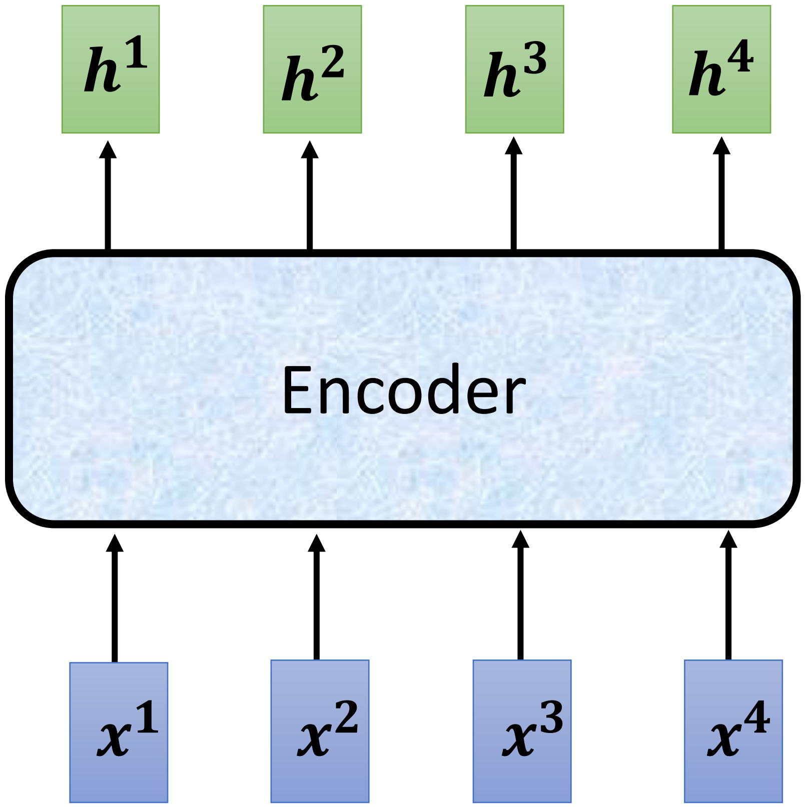
Transformer
https://arxiv.org/abs/1706.03762

Encoder

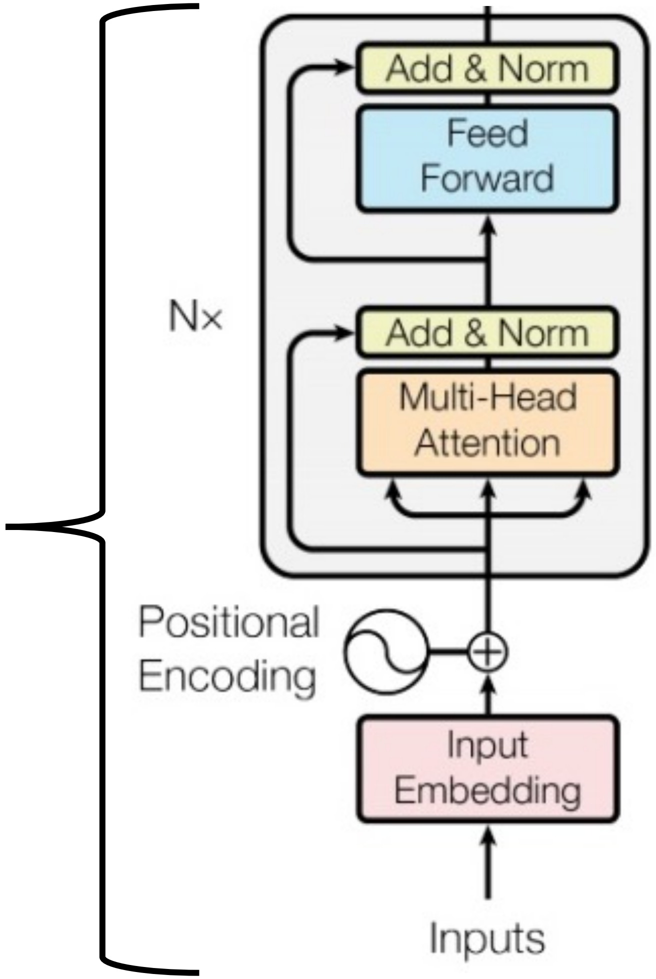


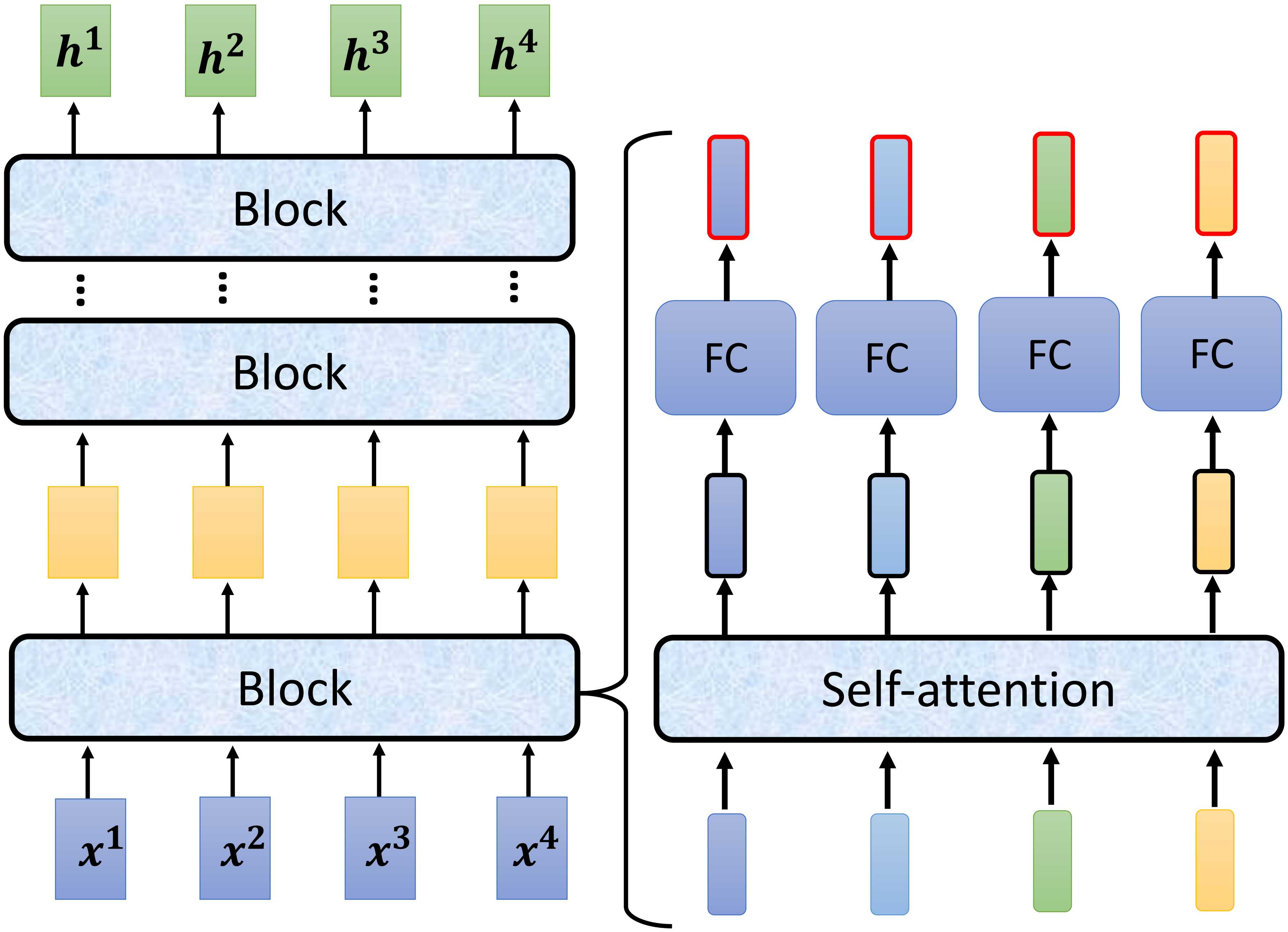
Encoder

You can use **RNN** or **CNN**.

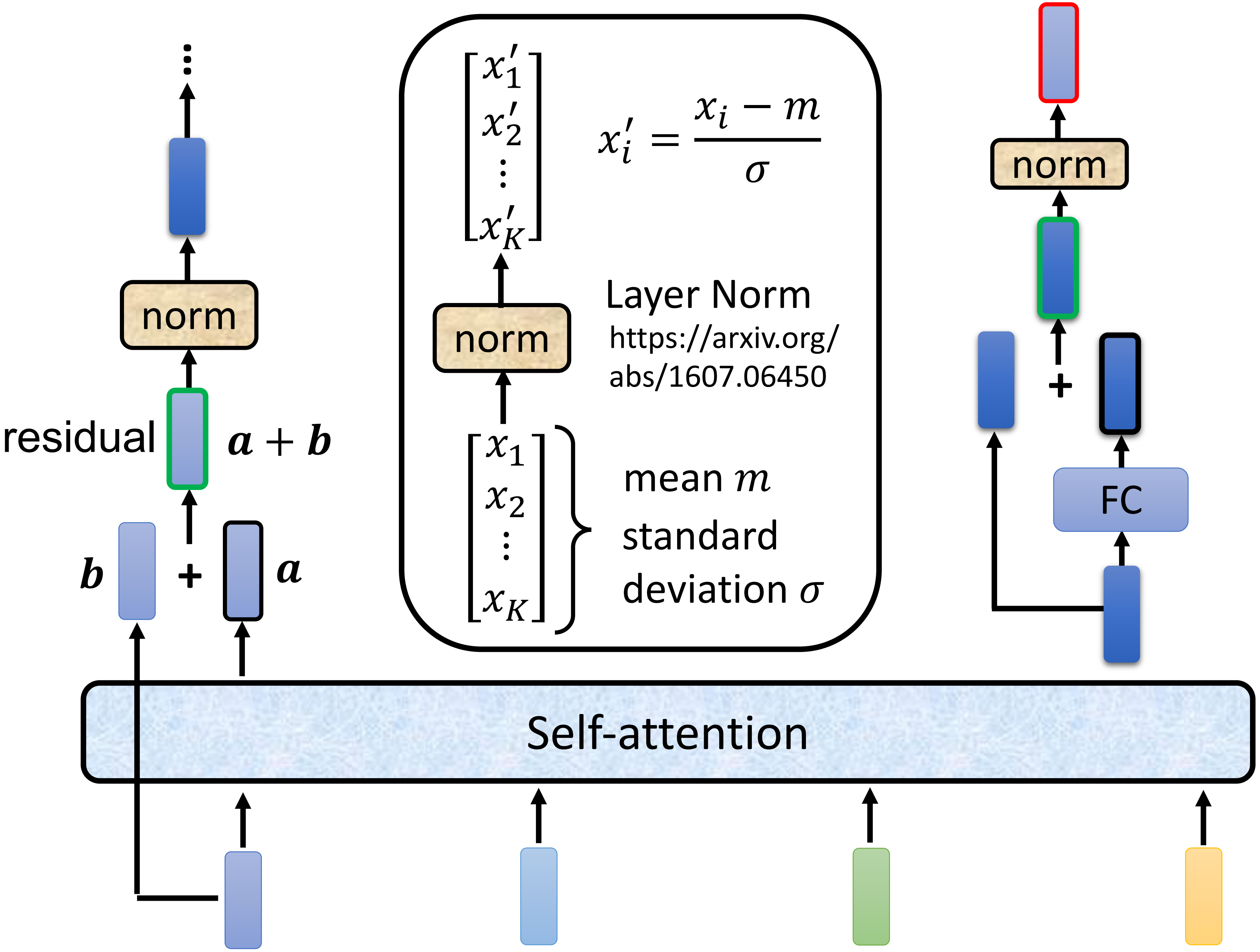


Transformer's Encoder

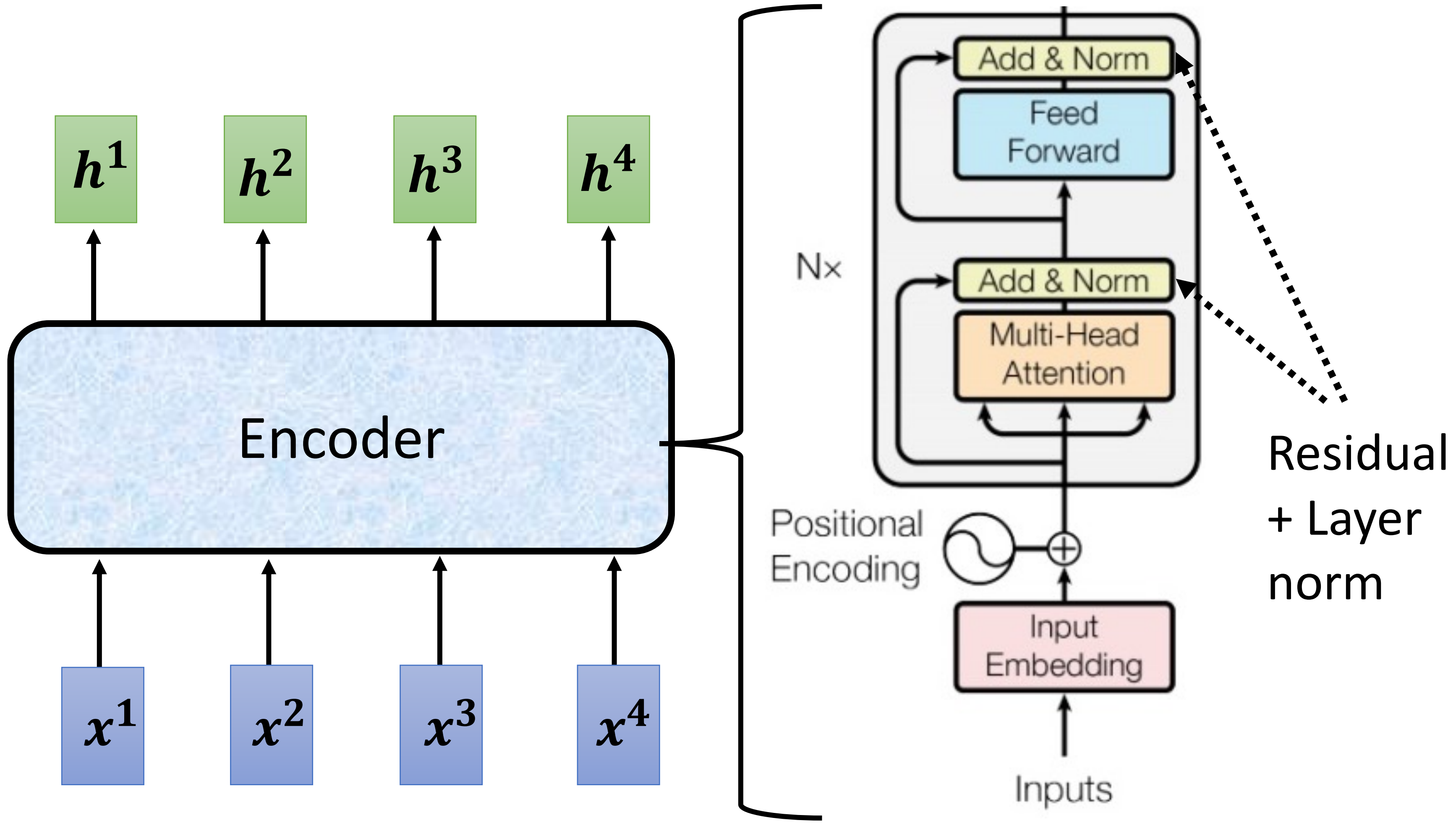




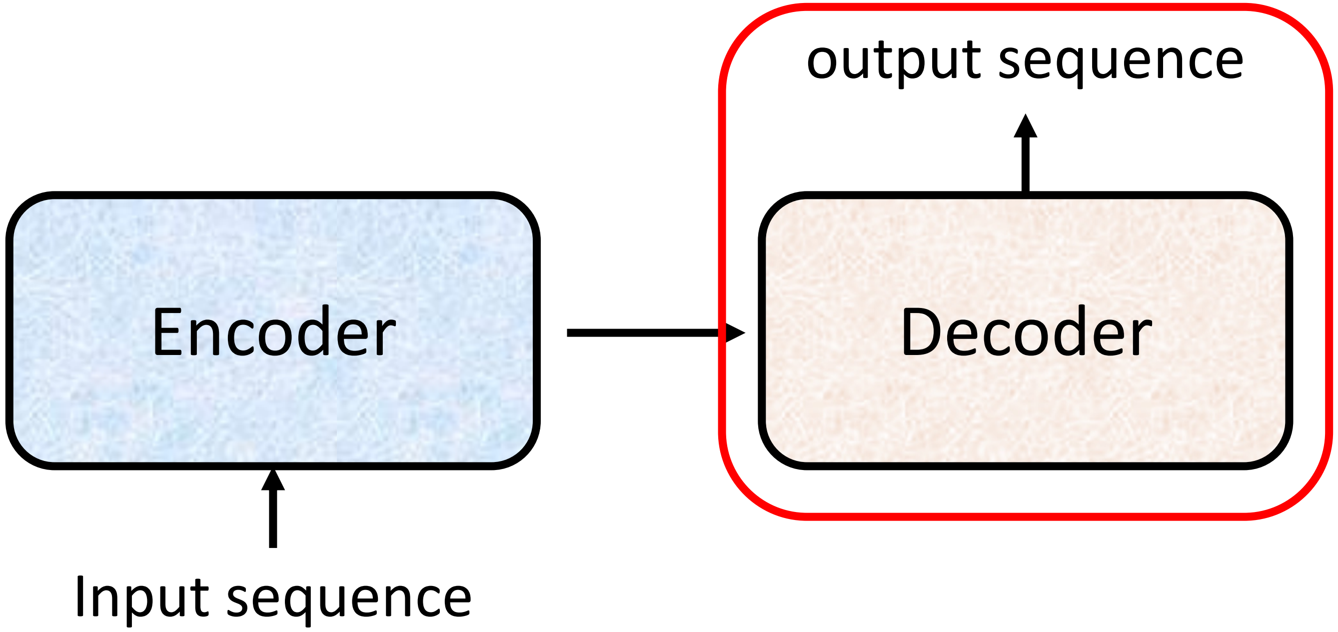
h^1 h^2 h^3 h^4
Block
Block
Block
x^1 x^2 x^3 x^4
FC FC FC FC
Self-attention



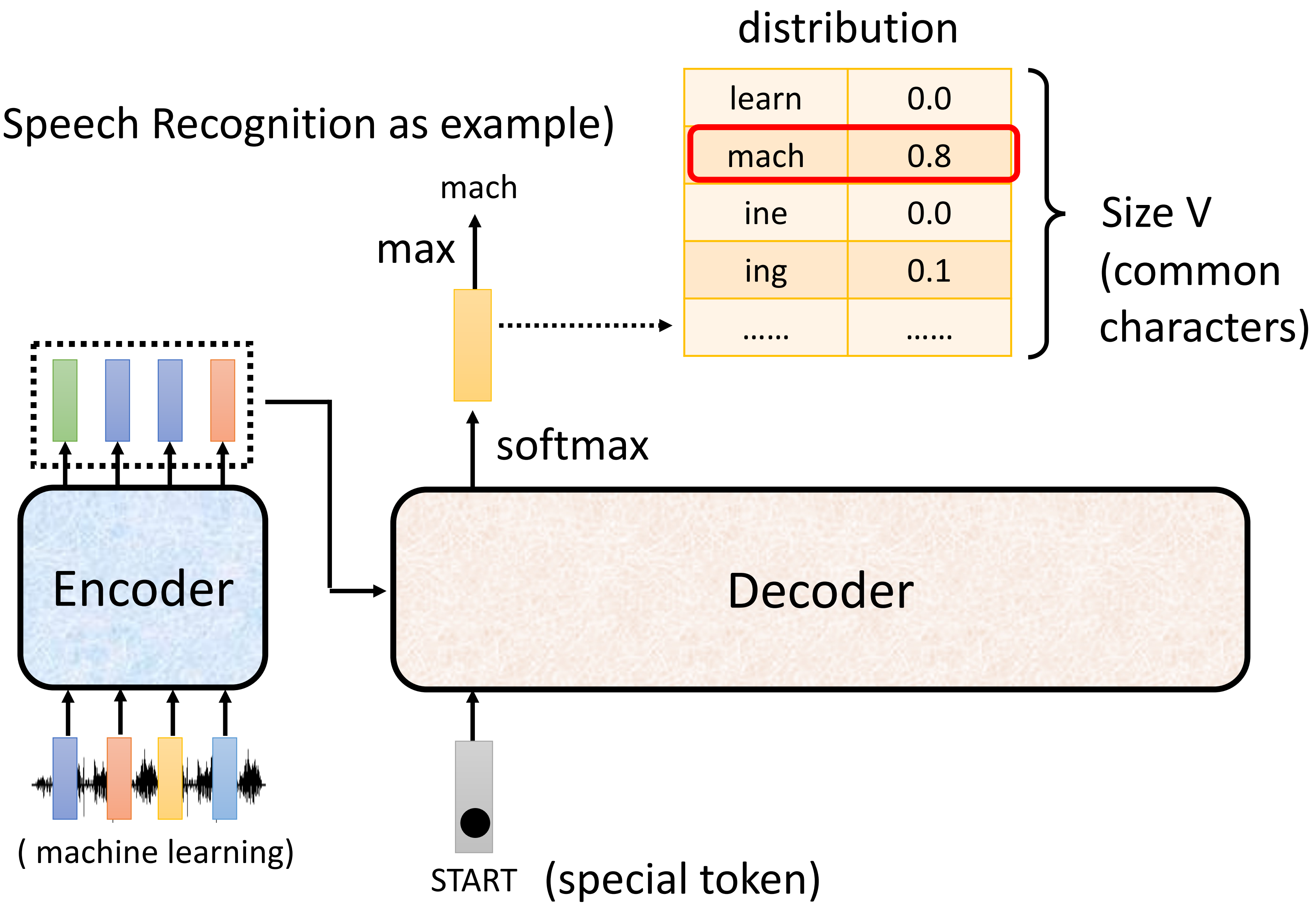
norm
residual
$a + b$
b
+
a
$\begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_K' \end{bmatrix}$
$x_i' = \frac{x_i - m}{\sigma}$
norm
Layer Norm
https://arxiv.org/abs/1607.06450
$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_K \end{bmatrix}$
mean m
standard deviation σ
norm
+
FC
Self-attention

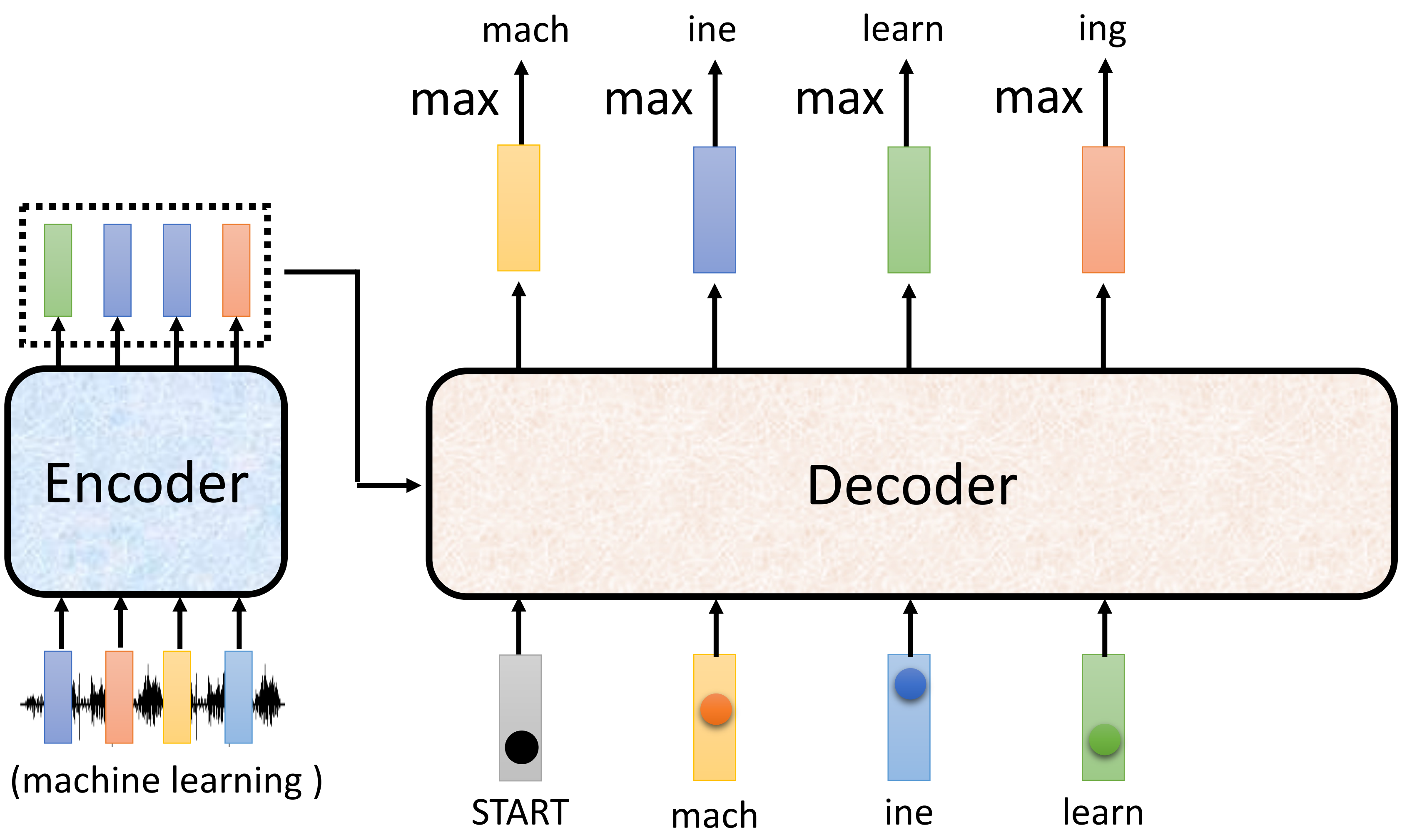

h^1
h^2
h^3
h^4
Encoder
x^1
x^2
x^3
x^4
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Positional Encoding
Input Embedding
Inputs
Residual + Layer norm

Decoder

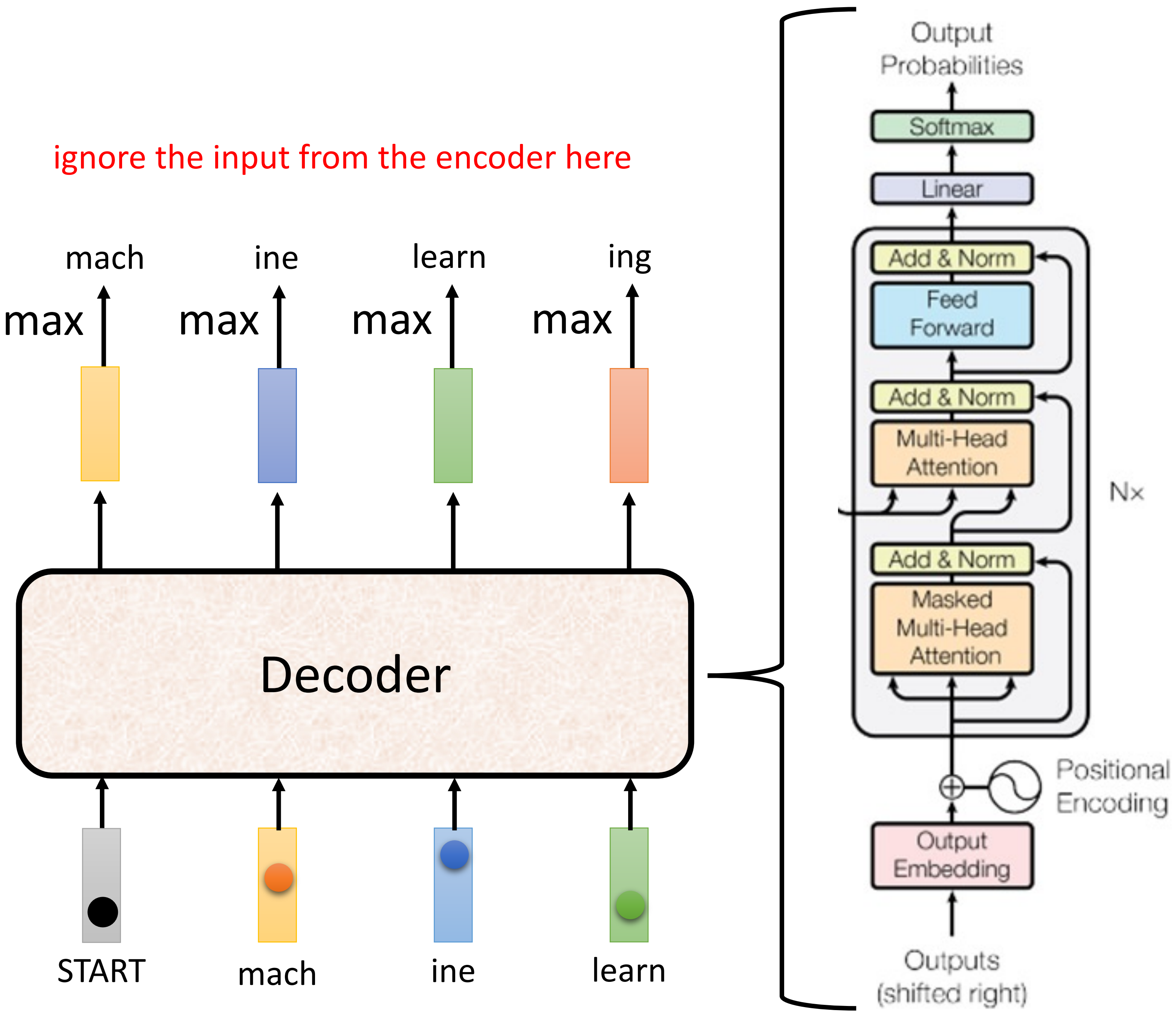


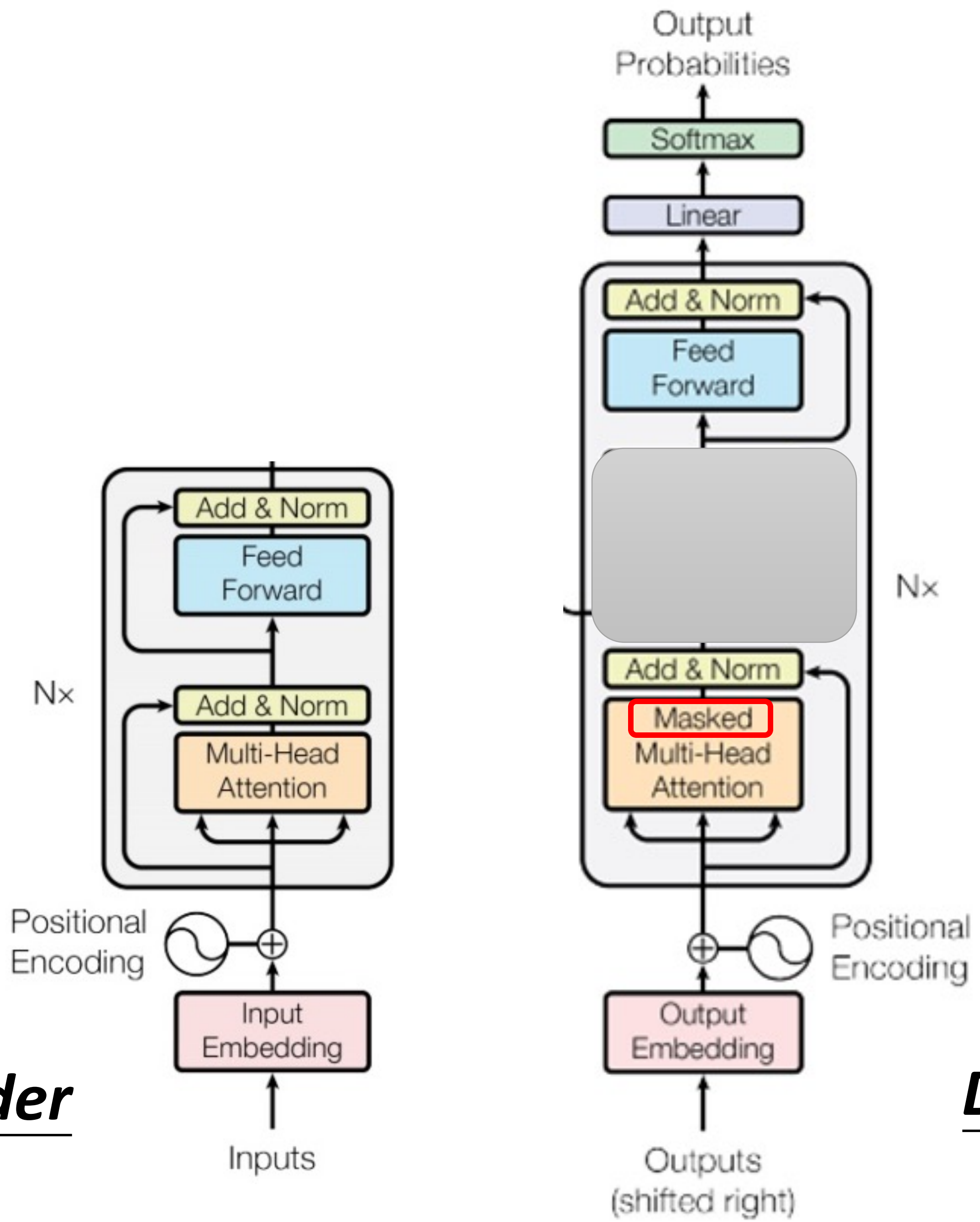
(Speech Recognition as example)




mach
ine
learn
ing
max
max
max
max
Encoder
Decoder
(machine learning)
START
mach
ine
learn

ignore the input from the encoder here

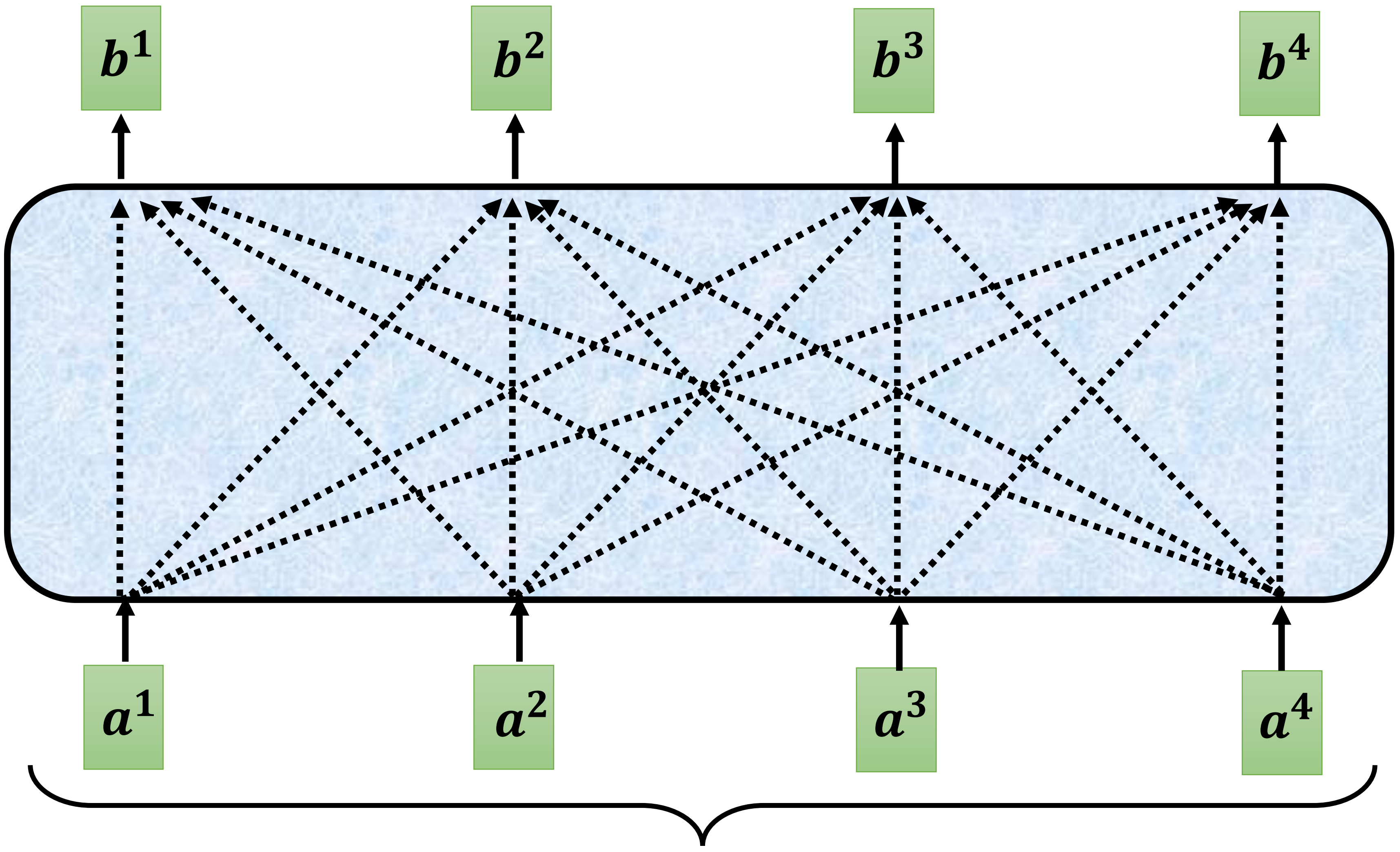




Encoder

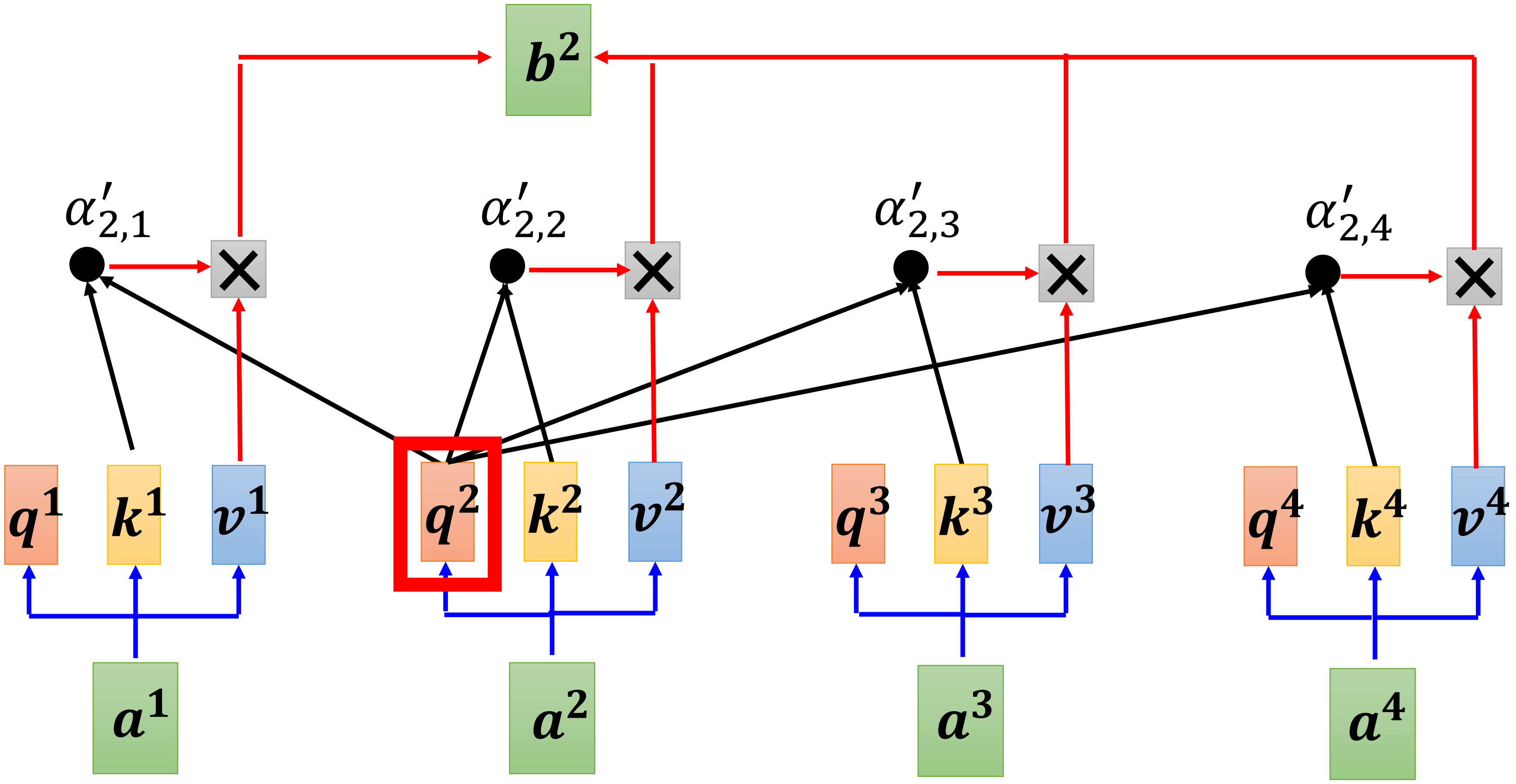
Decoder

Self-attention → Masked Self-attention

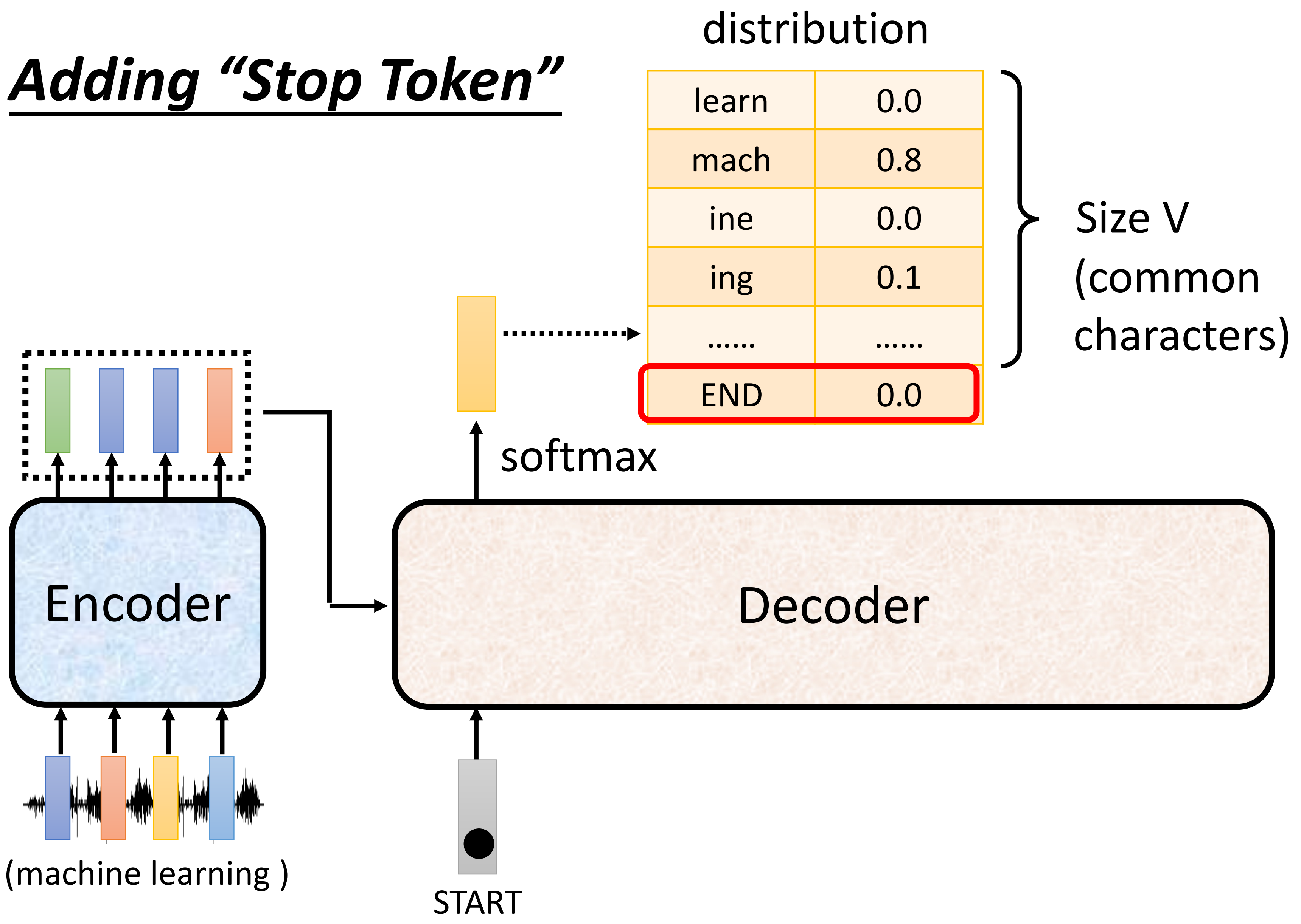


Can be either **input** or **a hidden layer**

Self-attention → Masked Self-attention

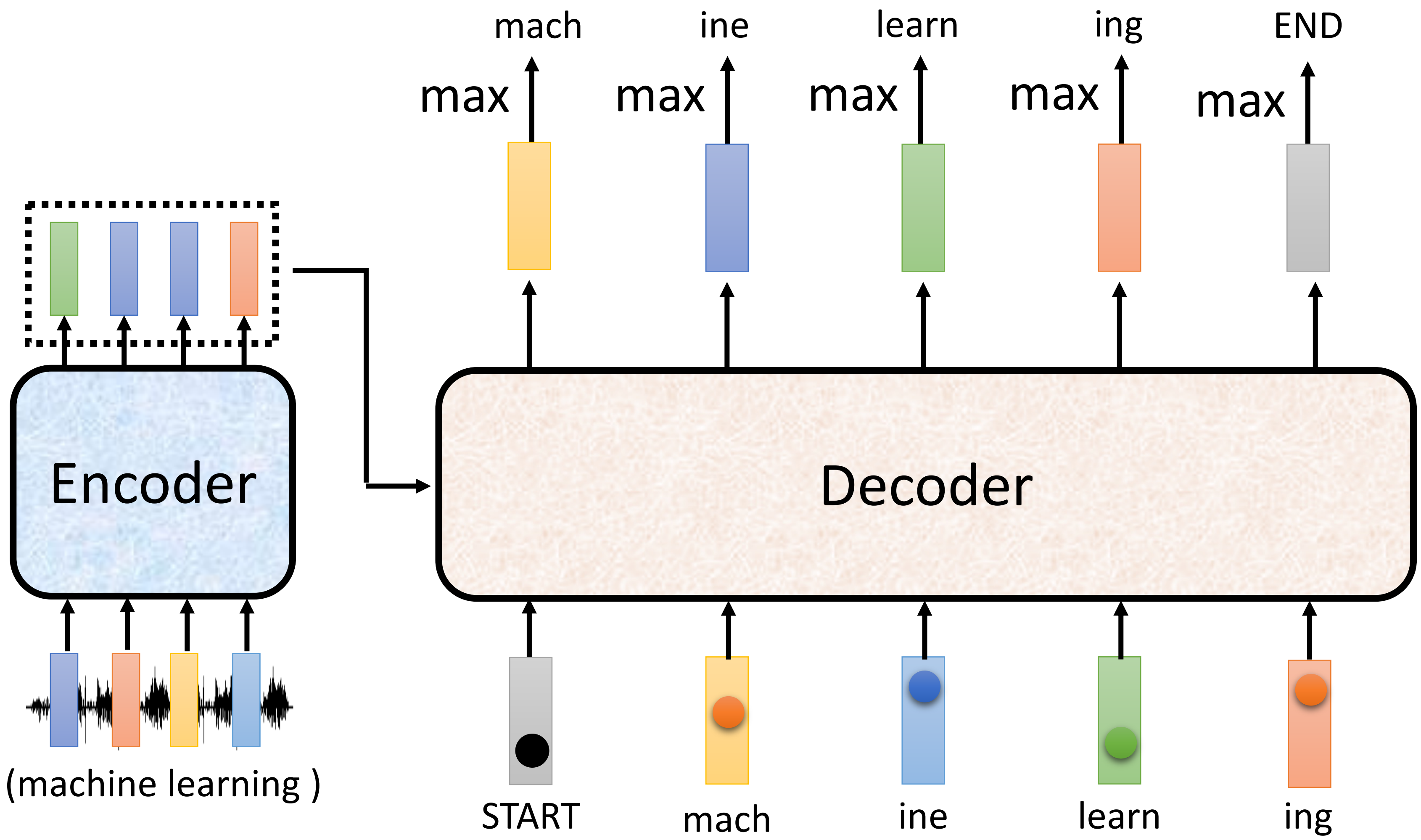


Adding "Stop Token"

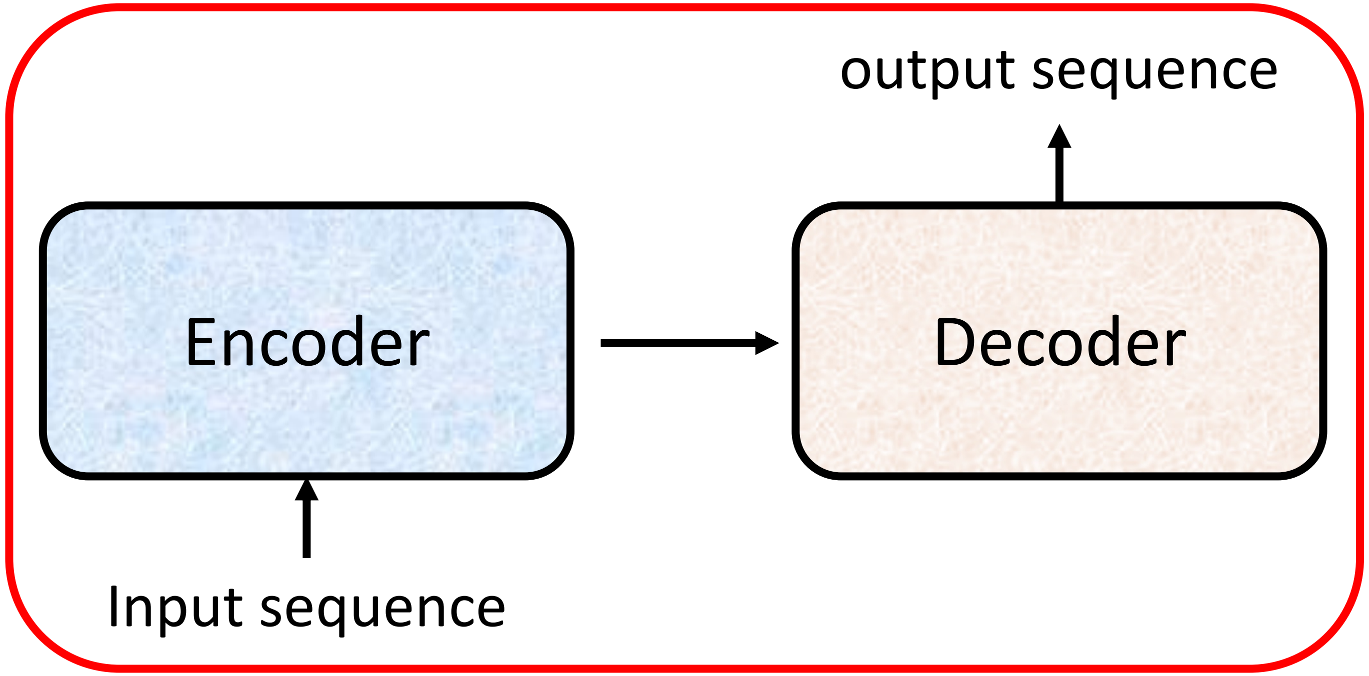


Autoregressive

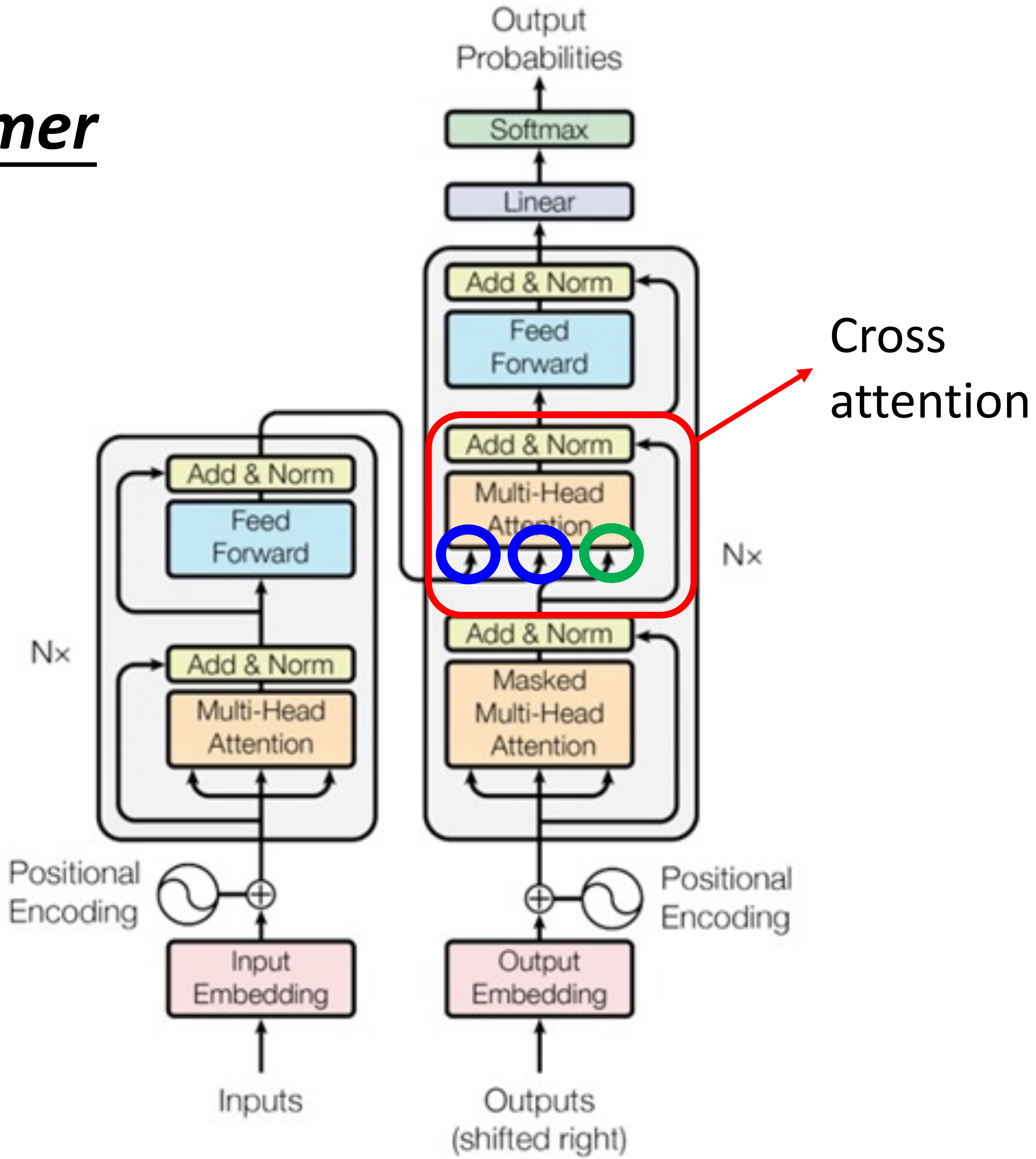
Stop at here!

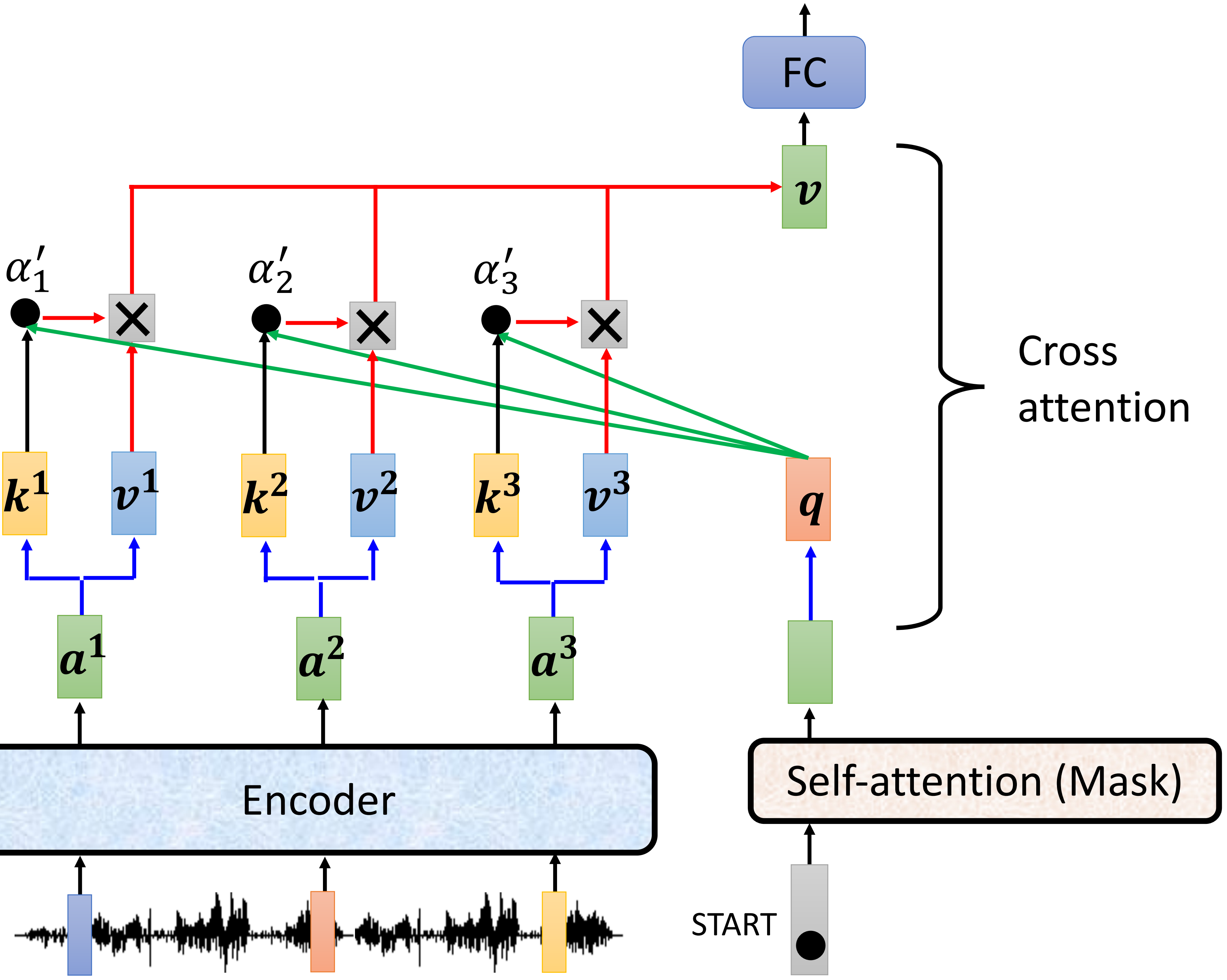


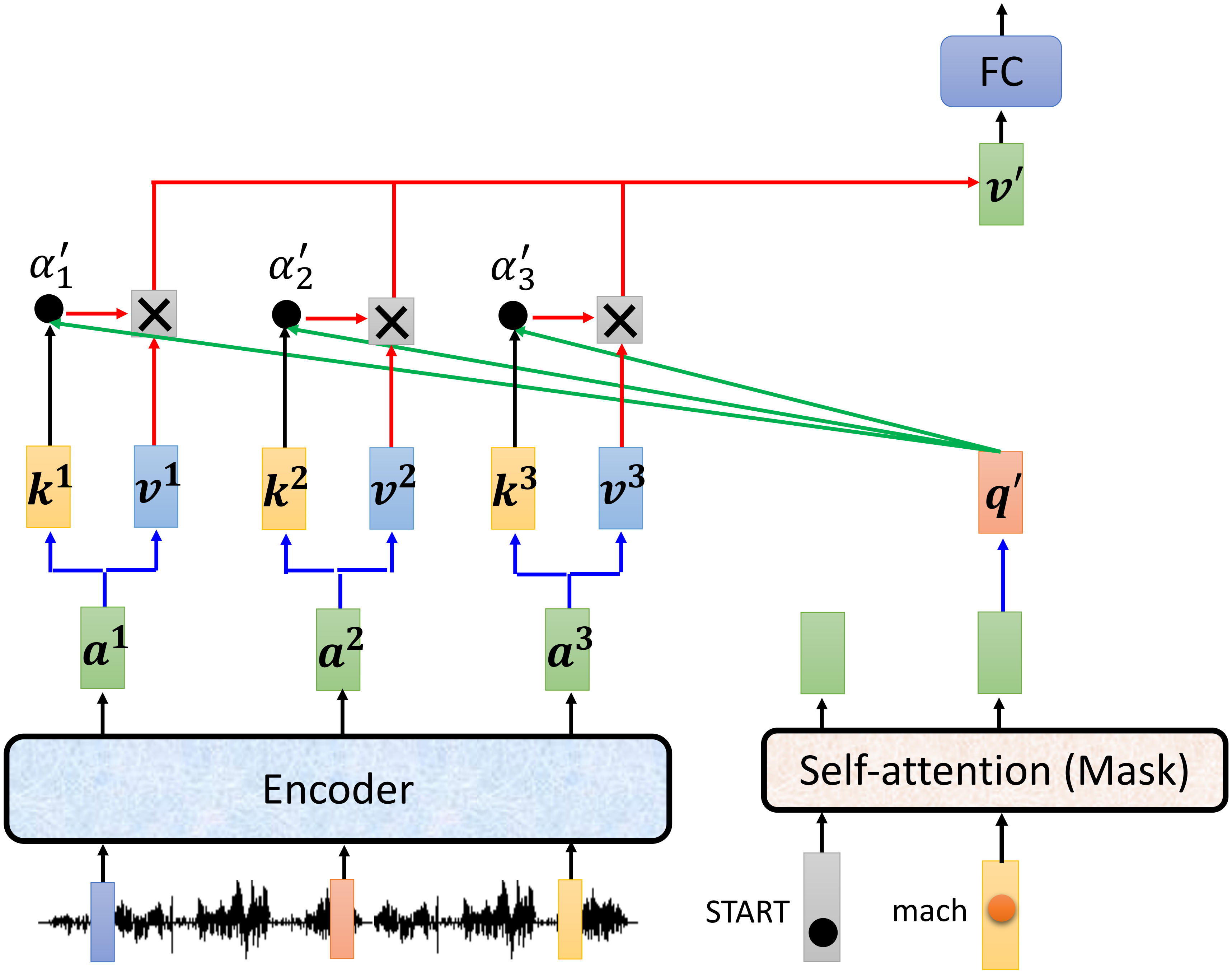
Encoder-Decoder



Transformer




FC
v
α'_1
α'_2
α'_3
Cross attention
k^1
v^1
k^2
v^2
k^3
v^3
q
a^1
a^2
a^3
Encoder
Self-attention (Mask)
START



FC
v'
α'_1
α'_2
α'_3
k^1
v^1
k^2
v^2
k^3
v^3
q'
a^1
a^2
a^3
Encoder
Self-attention (Mask)
START
mach

Experimental Results:

Machine Translation

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**	$2.3 \cdot 10^{19}$	

English Constituency Parsing

Parser	Training	WSJ 23 F1
Vinyals & Kaiser el al. (2014) [37]	WSJ only, discriminative	88.3
Petrov et al. (2006) [29]	WSJ only, discriminative	90.4
Zhu et al. (2013) [40]	WSJ only, discriminative	90.4
Dyer et al. (2016) [8]	WSJ only, discriminative	91.7
Transformer (4 layers)	WSJ only, discriminative	91.3
Zhu et al. (2013) [40]	semi-supervised	91.3
Huang & Harper (2009) [14]	semi-supervised	91.3
McClosky et al. (2006) [26]	semi-supervised	92.1
Vinyals & Kaiser el al. (2014) [37]	semi-supervised	92.1
Transformer (4 layers)	semi-supervised	92.7
Luong et al. (2015) [23]	multi-task	93.0
Dyer et al. (2016) [8]	generative	93.3

Summary:

- The Transformer is the first sequence transduction model based entirely on self-attention
- Strong parallel computing capability, extremely fast training speed
- Achieves SOTA results on tasks like machine translation, demonstrating the power of attention mechanisms
- Validates that attention is all you need to build efficient models