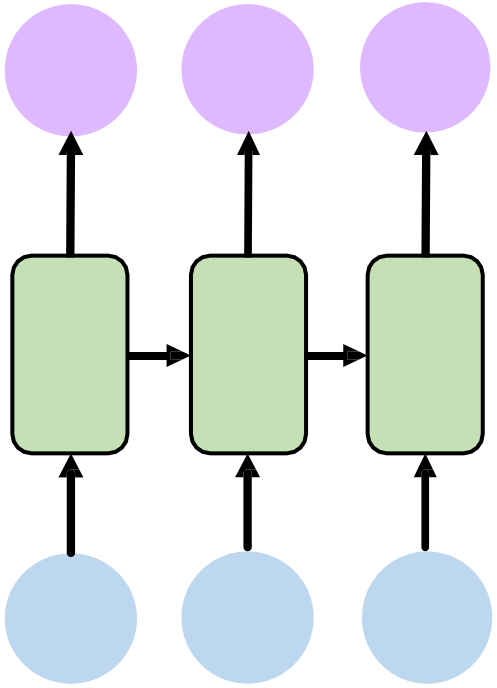
Attention and Transformers

Qiang Sun

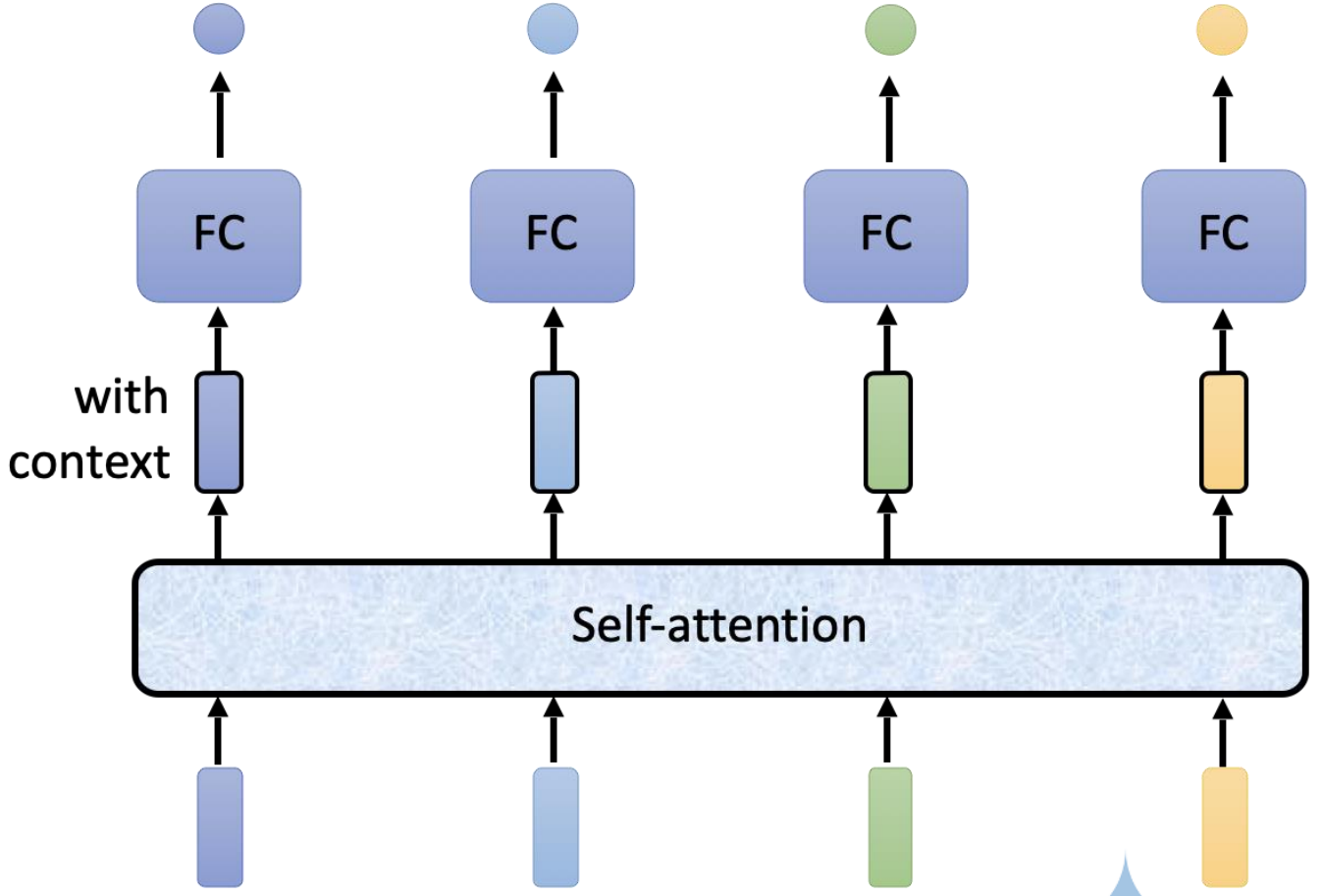
University of Toronto

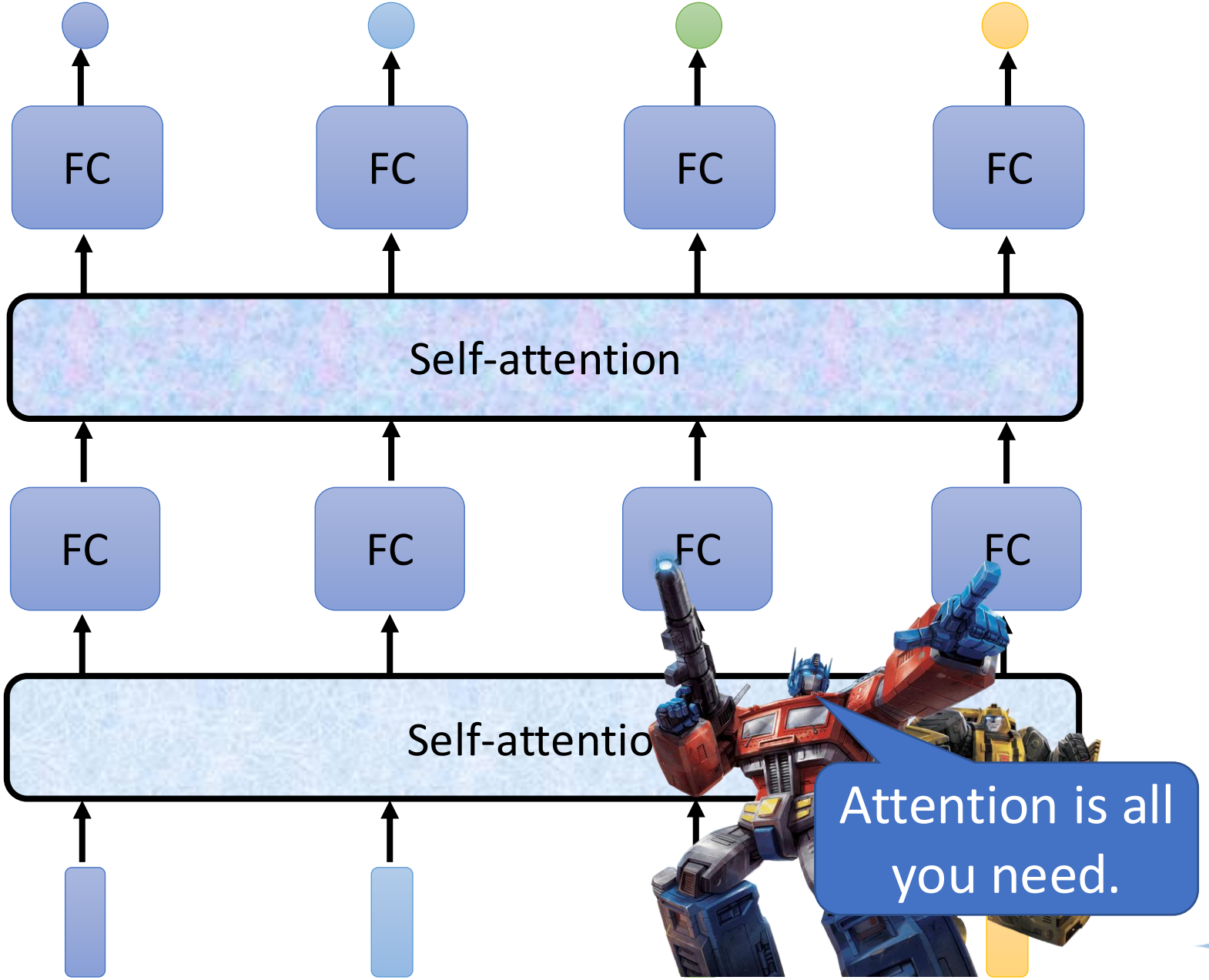
All credits go to Hung-yi Lee.

Sequence to Sequence Modeling



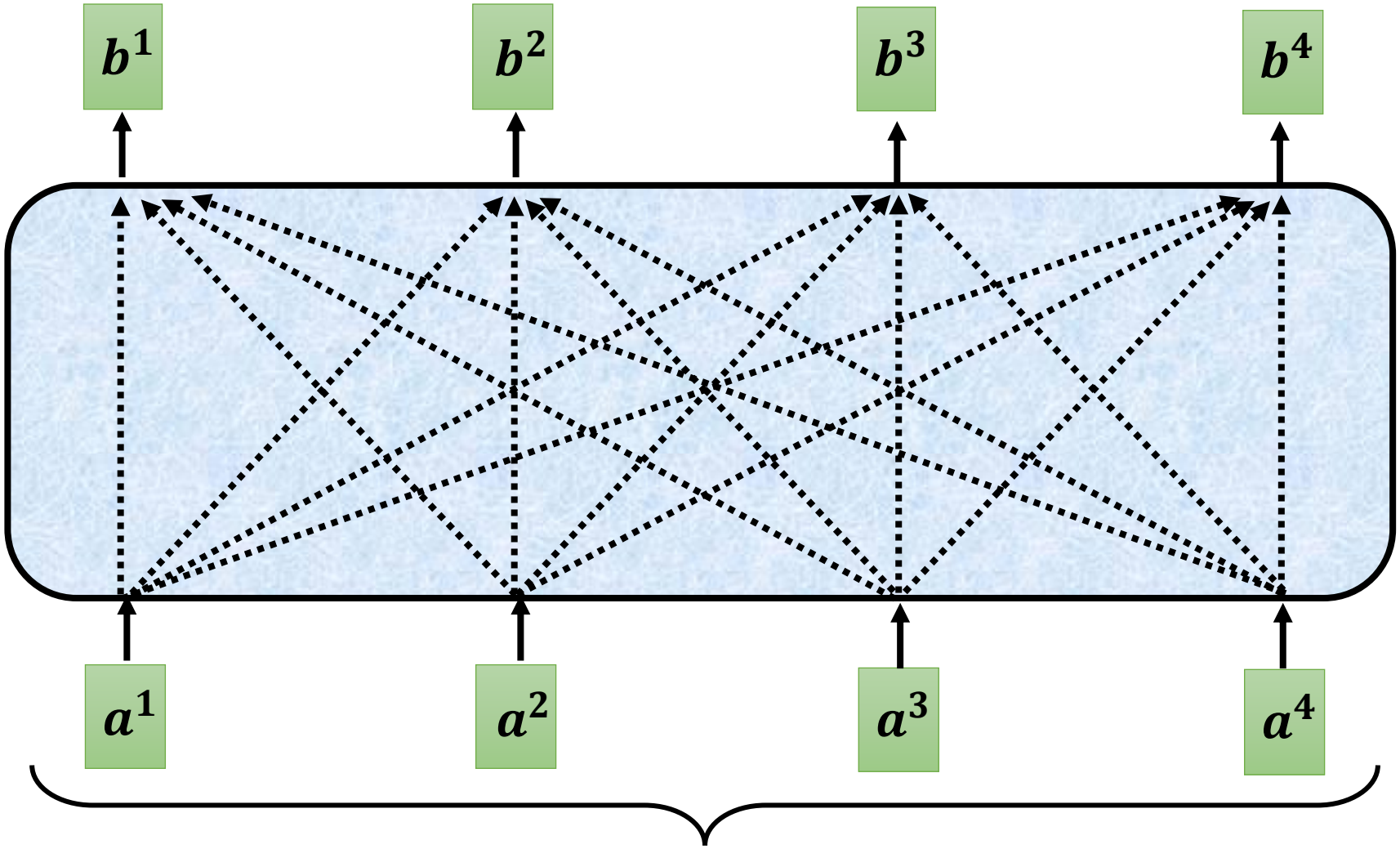
Seq to Seq
Machine Translation





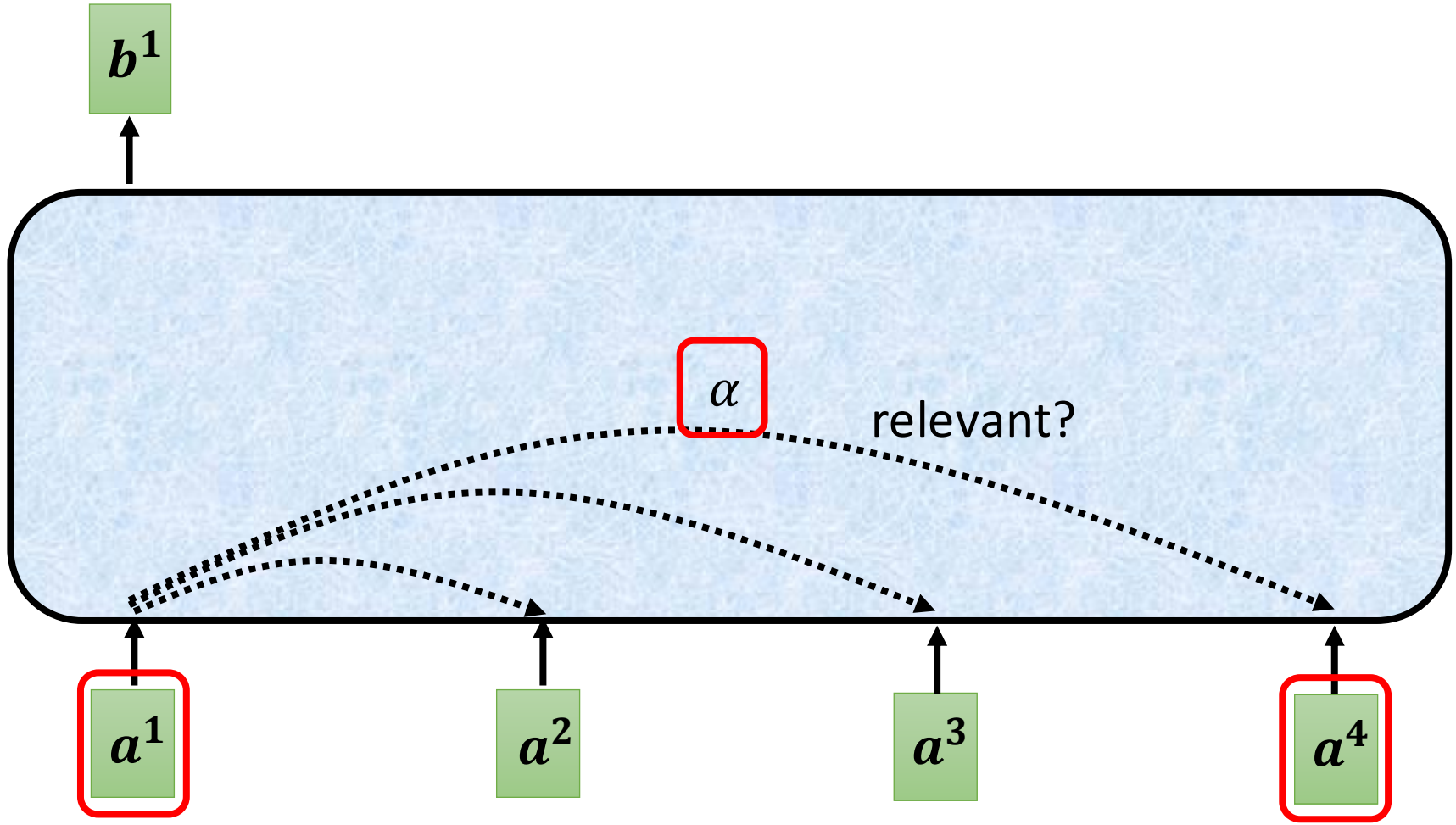
https://arxiv.org/abs/1706.03762

Self-attention



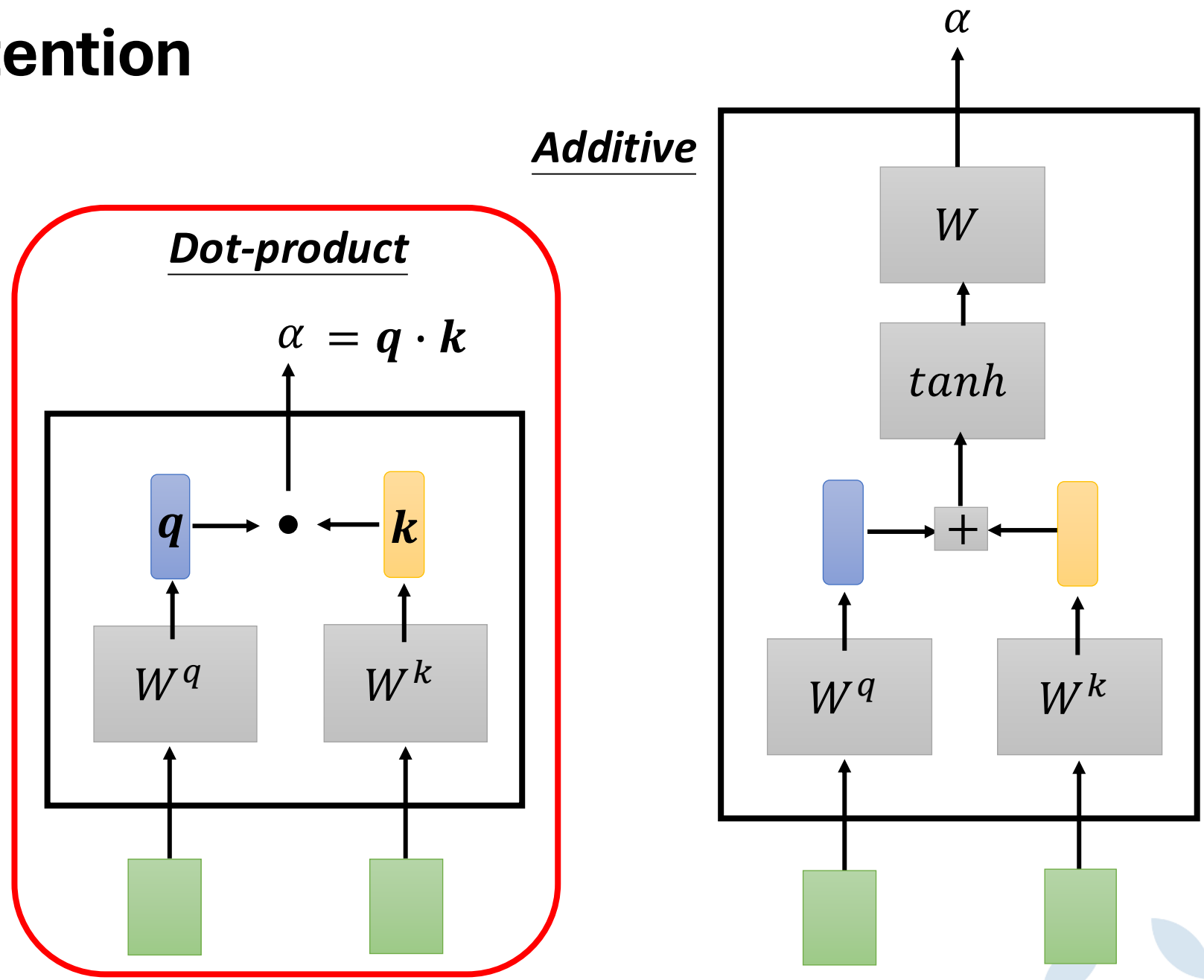
Can be either **input** or **a hidden layer**

Self-attention



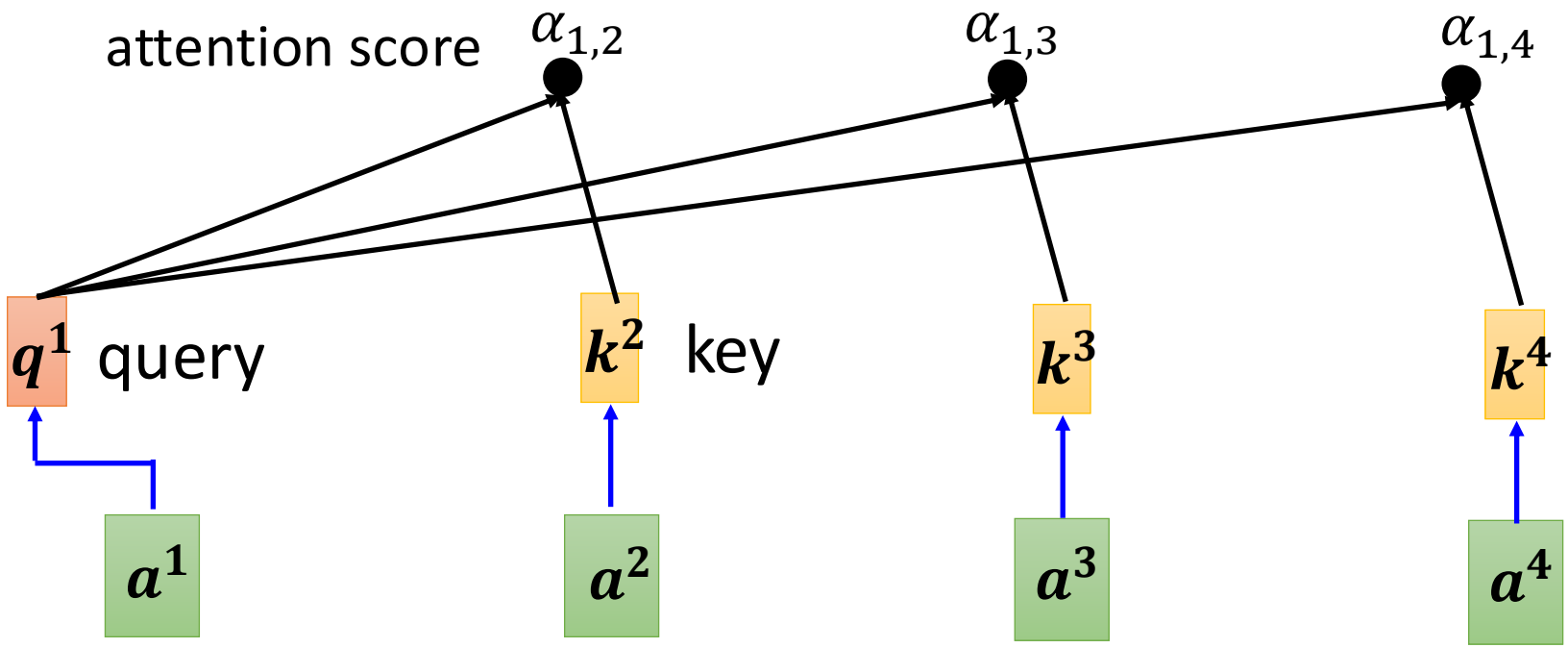
Find the relevant vectors in a sequence

Self-attention



Self-attention

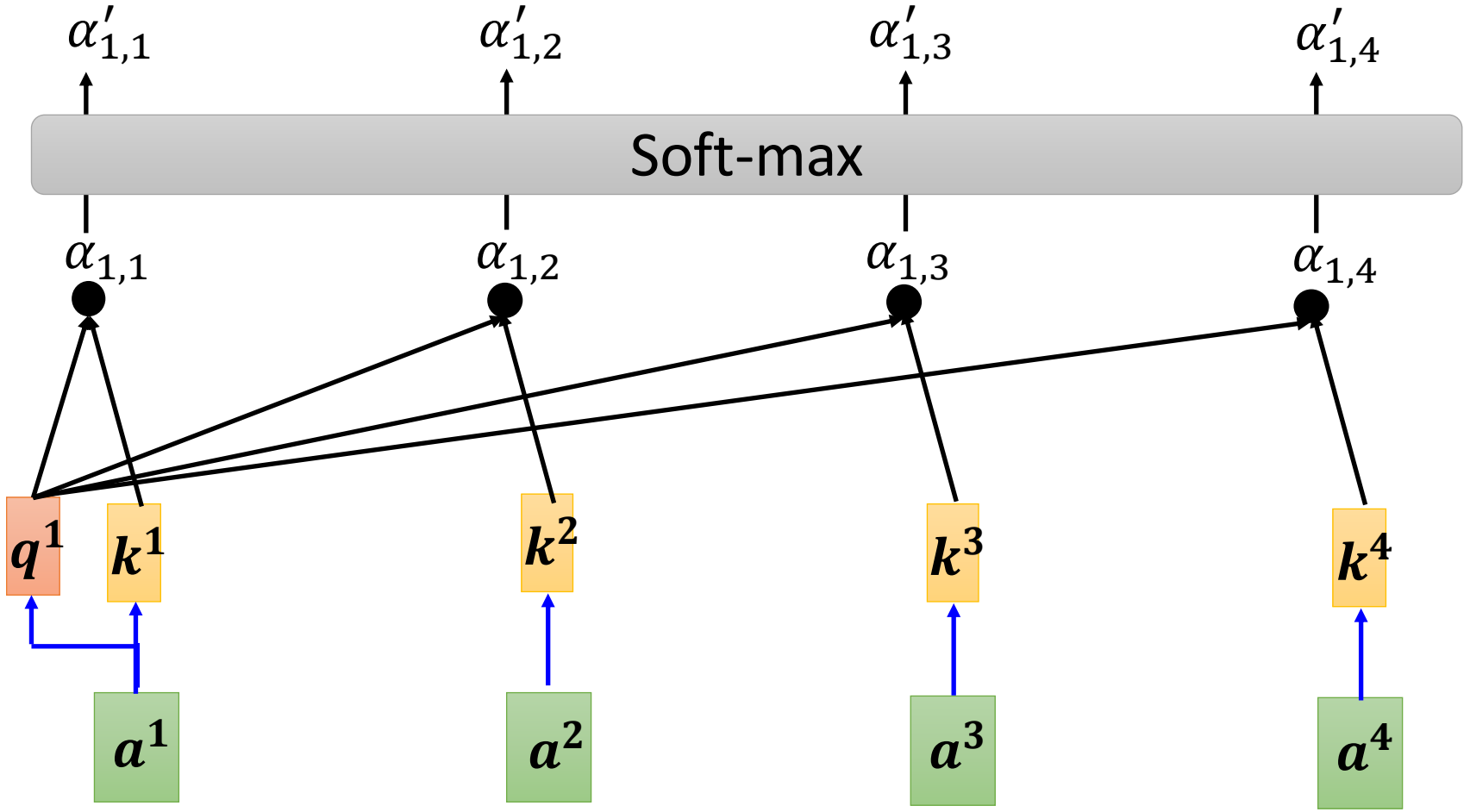
$$\alpha_{1,2} = \boldsymbol{q^1} \cdot \boldsymbol{k^2} \qquad \alpha_{1,3} = \boldsymbol{q^1} \cdot \boldsymbol{k^3} \qquad \alpha_{1,4} = \boldsymbol{q^1} \cdot \boldsymbol{k^4}$$



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

Self-attention

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



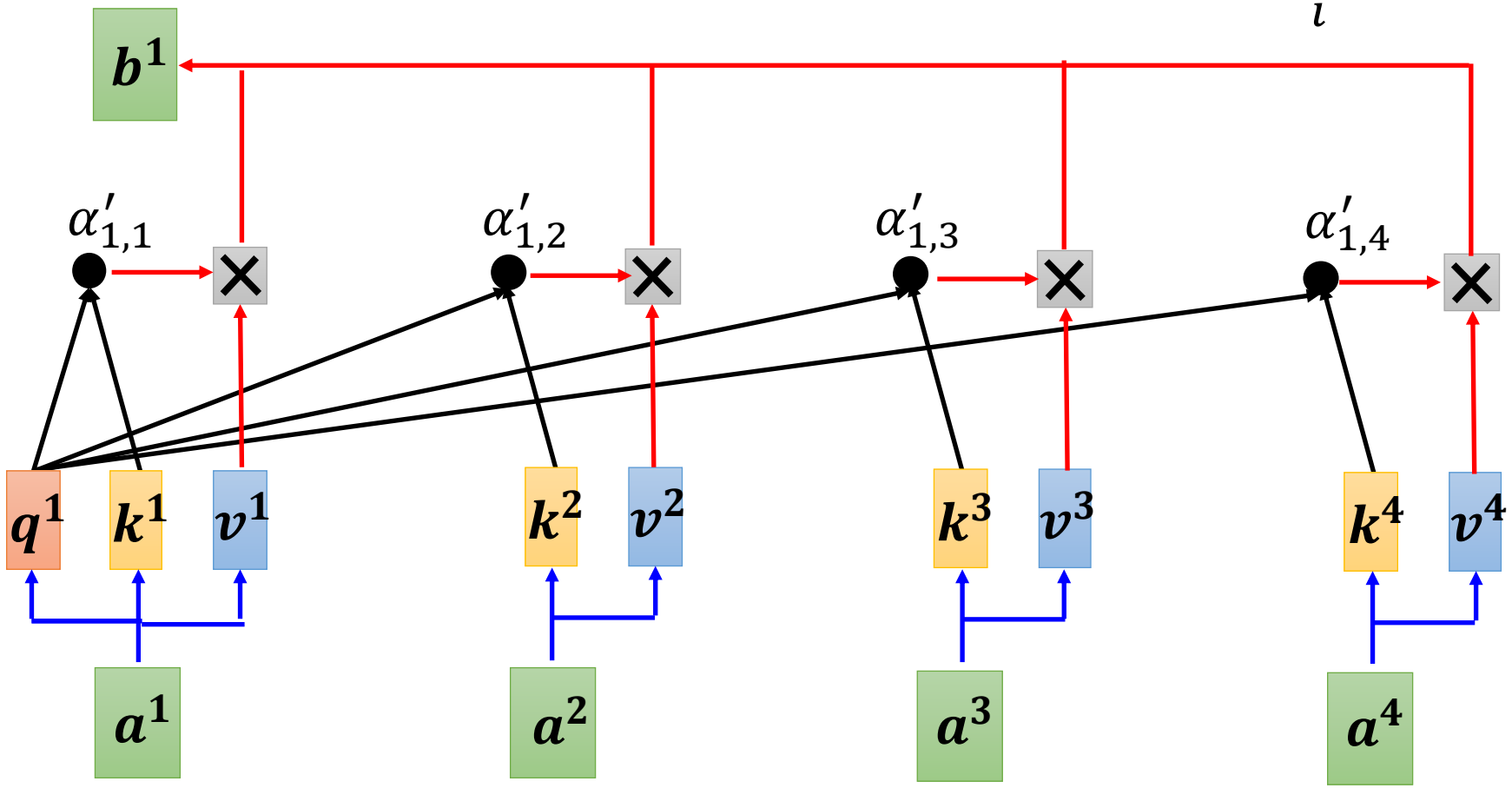
$q^1 = W^q a^1$ $k^2 = W^k a^2$ $k^3 = W^k a^3$ $k^4 = W^k a^4$

$k^1 = W^k a^1$

Self-attention

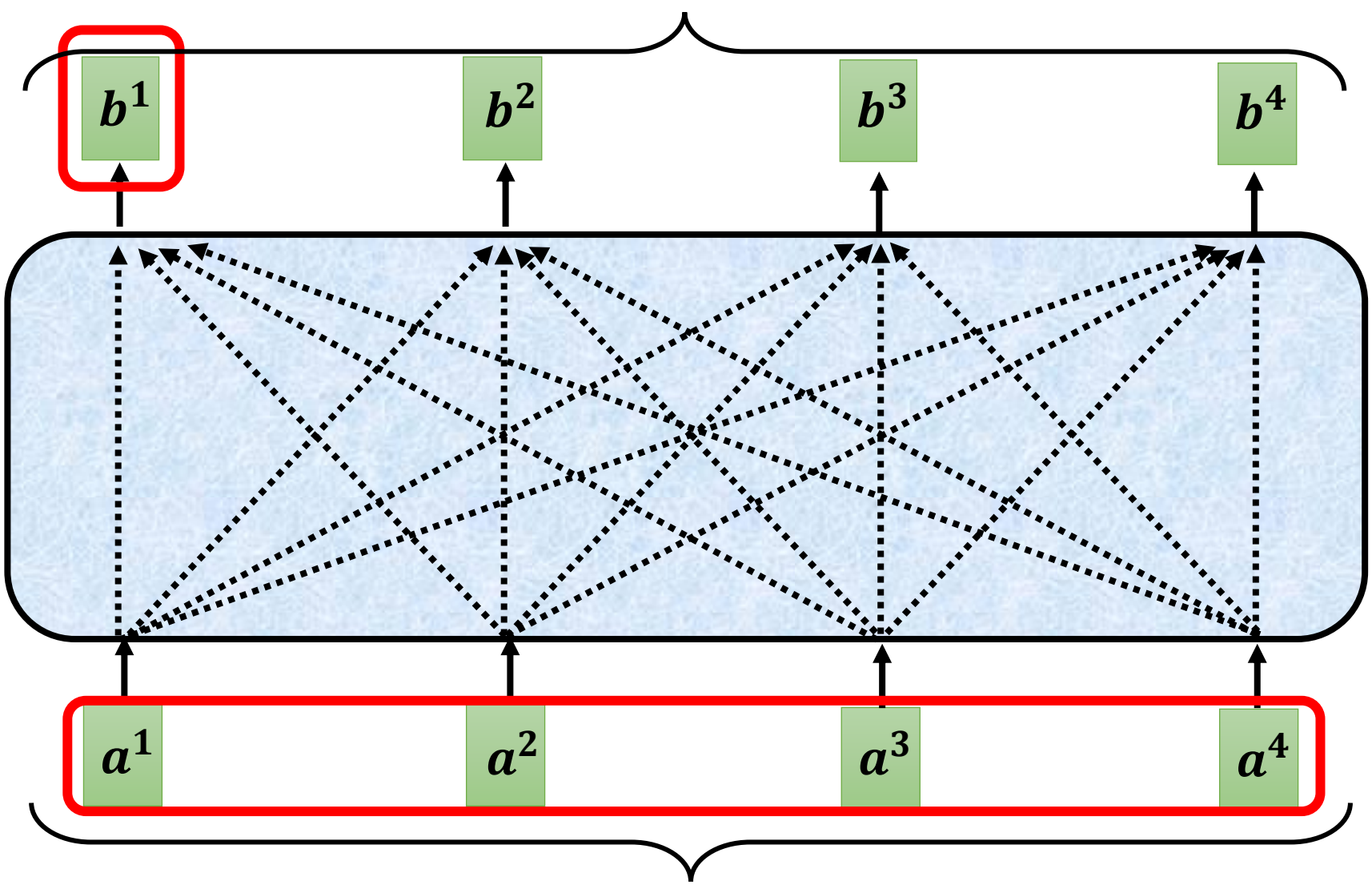
Extract information based on attention scores

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



$v^1 = W^v a^1$ $v^2 = W^v a^2$ $v^3 = W^v a^3$ $v^4 = W^v a^4$

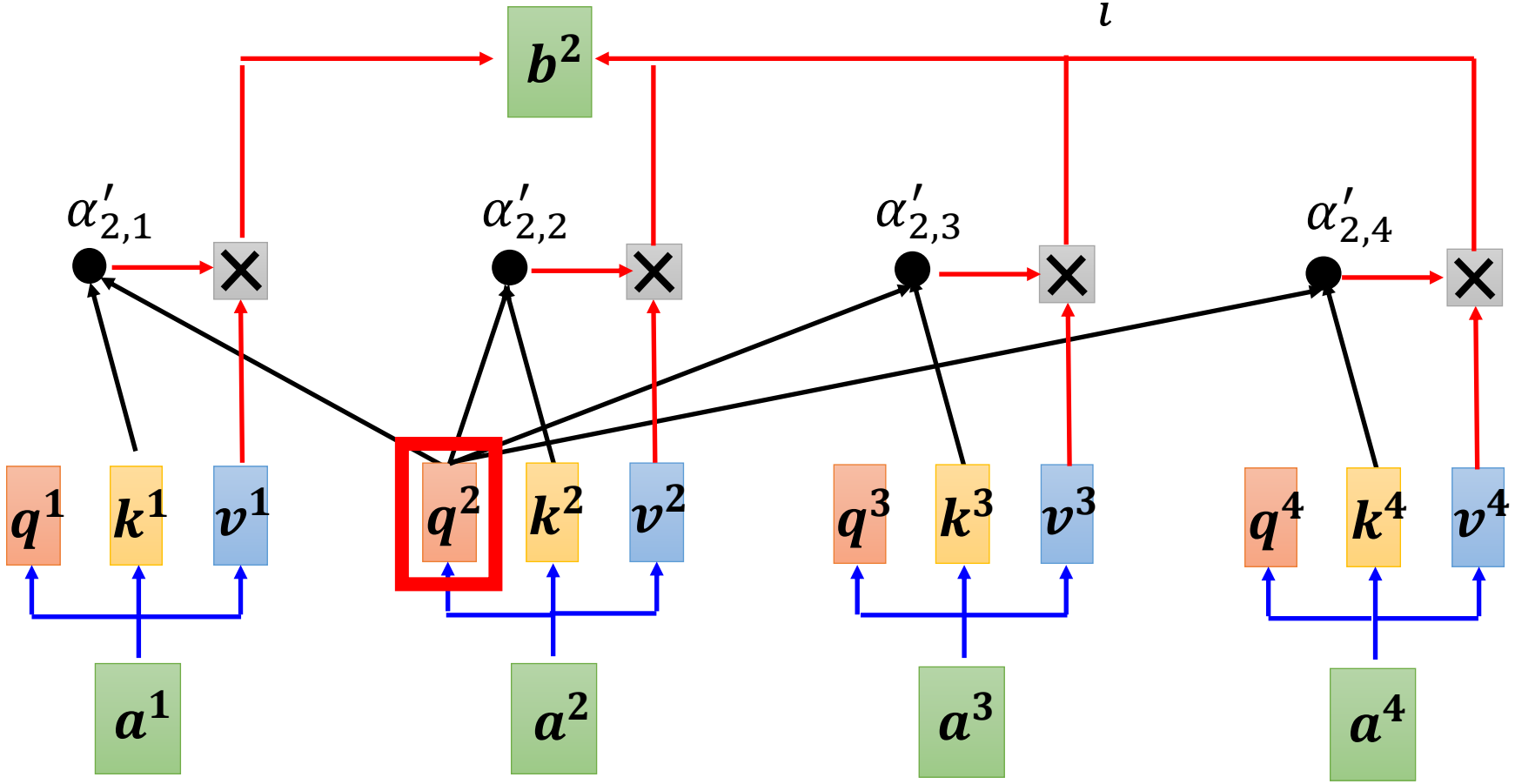
Self-attention



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

Self-attention

$$b^2 = \sum_i \alpha'_{2,i} v^i$$



Self-attention

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

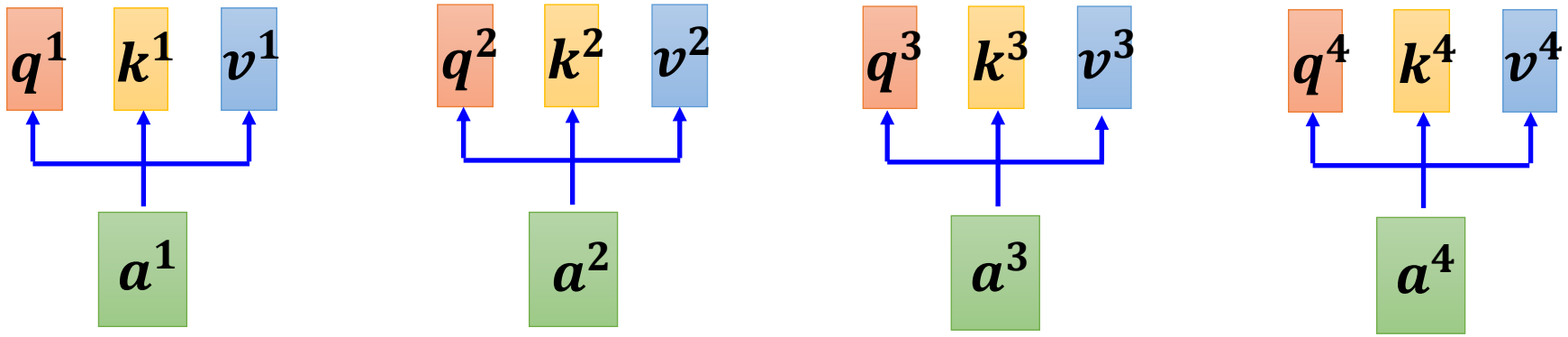
$$\begin{matrix} q^1 & q^2 & q^3 & q^4 \\ Q \end{matrix} = W^q \begin{matrix} a^1 & a^2 & a^3 & a^4 \\ I \end{matrix}$$

$$k^i = W^k a^i$$

$$\begin{matrix} k^1 & k^2 & k^3 & k^4 \\ K \end{matrix} = W^k \begin{matrix} a^1 & a^2 & a^3 & a^4 \\ I \end{matrix}$$

$$v^i = W^v a^i$$

$$\begin{matrix} v^1 & v^2 & v^3 & v^4 \\ V \end{matrix} = W^v \begin{matrix} a^1 & a^2 & a^3 & a^4 \\ I \end{matrix}$$

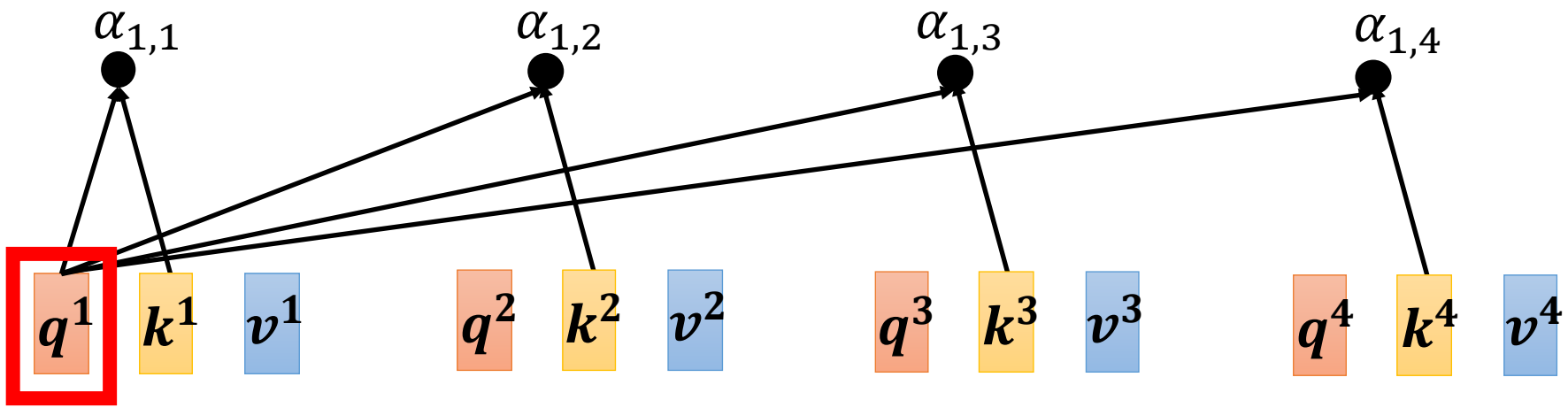


Self-attention

$$\alpha_{1,1} = \boldsymbol{k^1}\,\boldsymbol{q^1} \qquad \alpha_{1,2} = \boldsymbol{k^2}\,\boldsymbol{q^1}$$

$$\alpha_{1,3} = \boldsymbol{k^3}\,\boldsymbol{q^1} \qquad \alpha_{1,4} = \boldsymbol{k^4}\,\boldsymbol{q^1}$$

$$\begin{matrix}\alpha_{1,1}\\ \alpha_{1,2}\\ \alpha_{1,3}\\ \alpha_{1,4}\end{matrix} = \begin{matrix}\boldsymbol{k^1}\\ \boldsymbol{k^2}\\ \boldsymbol{k^3}\\ \boldsymbol{k^4}\end{matrix}\,\boldsymbol{q^1}$$

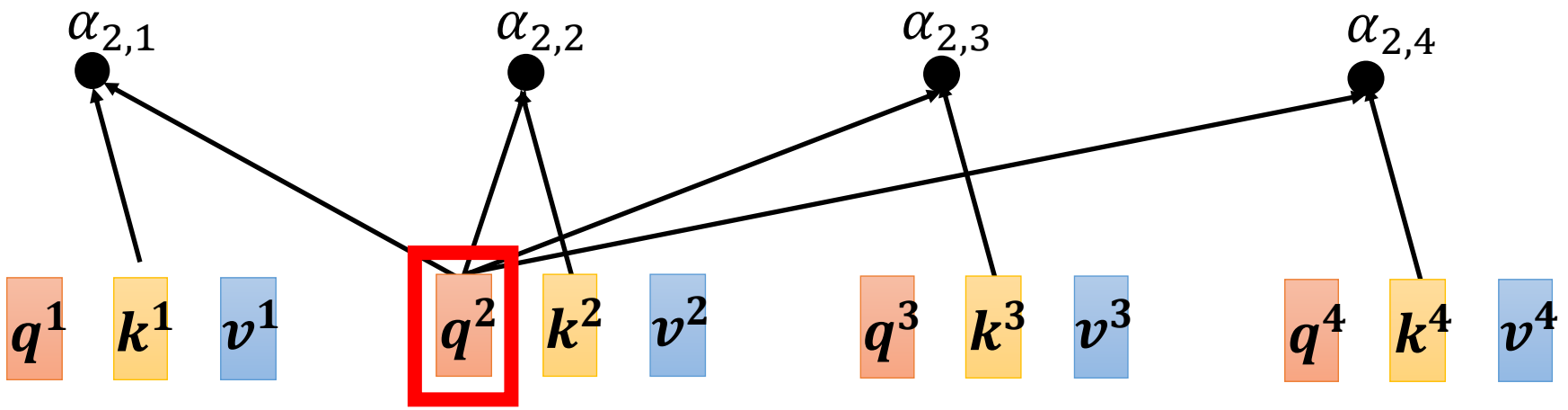


Self-attention

$\alpha_{1,1} = \boldsymbol{k^1}\ \boldsymbol{q^1}$ $\alpha_{1,2} = \boldsymbol{k^2}\ \boldsymbol{q^1}$

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

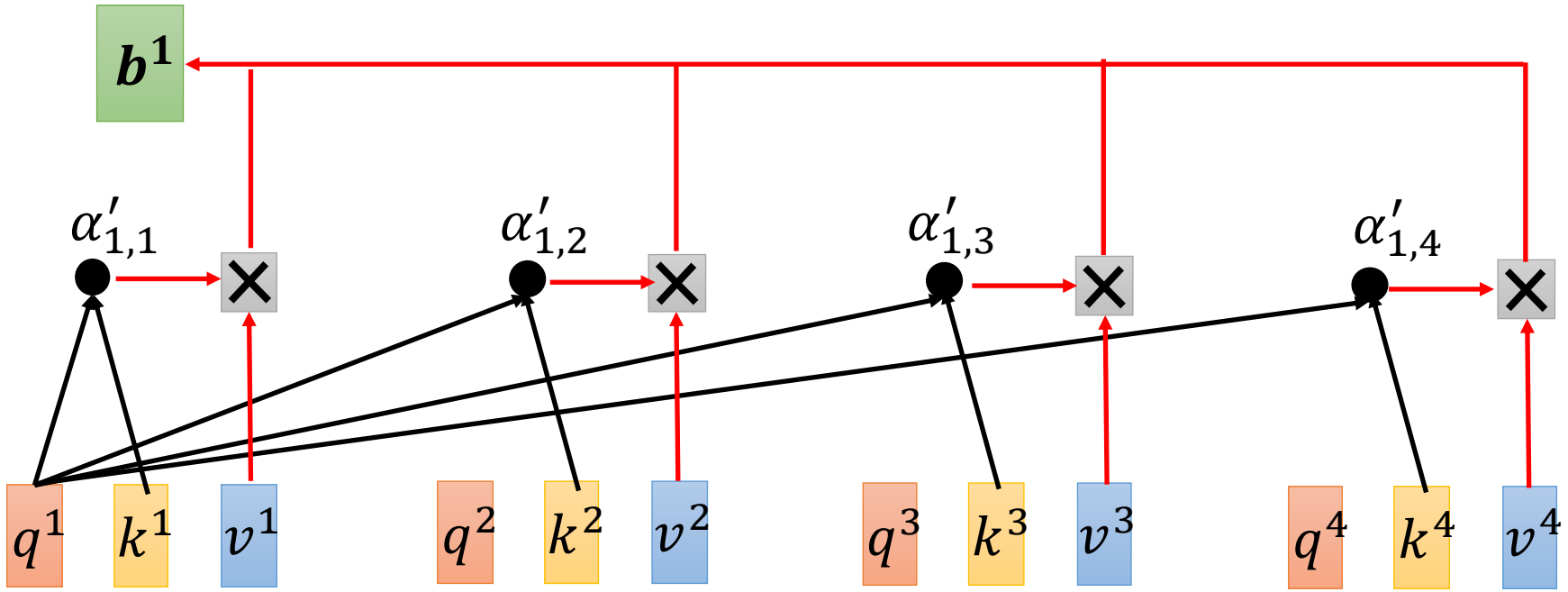
$$\begin{bmatrix}\alpha_{1,1}\\ \alpha_{1,2}\\ \alpha_{1,3}\\ \alpha_{1,4}\end{bmatrix} = \begin{bmatrix}\boldsymbol{k^1}\\ \boldsymbol{k^2}\\ \boldsymbol{k^3}\\ \boldsymbol{k^4}\end{bmatrix} \boldsymbol{q^1}$$



$$\begin{bmatrix}\alpha'_{1,1} & \alpha'_{2,1} & \alpha'_{3,1} & \alpha'_{4,1}\\ \alpha'_{1,2} & \alpha'_{2,2} & \alpha'_{3,2} & \alpha'_{4,2}\\ \alpha'_{1,3} & \alpha'_{2,3} & \alpha'_{3,3} & \alpha'_{4,3}\\ \alpha'_{1,4} & \alpha'_{2,4} & \alpha'_{3,4} & \alpha'_{4,4}\end{bmatrix} = A' \quad \text{softmax} \quad \begin{bmatrix}\alpha_{1,1} & \alpha_{2,1} & \alpha_{3,1} & \alpha_{4,1}\\ \alpha_{1,2} & \alpha_{2,2} & \alpha_{3,2} & \alpha_{4,2}\\ \alpha_{1,3} & \alpha_{2,3} & \alpha_{3,3} & \alpha_{4,3}\\ \alpha_{1,4} & \alpha_{2,4} & \alpha_{3,4} & \alpha_{4,4}\end{bmatrix} = A$$

$$\begin{bmatrix}\boldsymbol{k^1}\\ \boldsymbol{k^2}\\ \boldsymbol{k^3}\\ \boldsymbol{k^4}\end{bmatrix} = K^T \qquad \begin{bmatrix}\boldsymbol{q^1} & \boldsymbol{q^2} & \boldsymbol{q^3} & \boldsymbol{q^4}\end{bmatrix} = Q$$

Self-attention



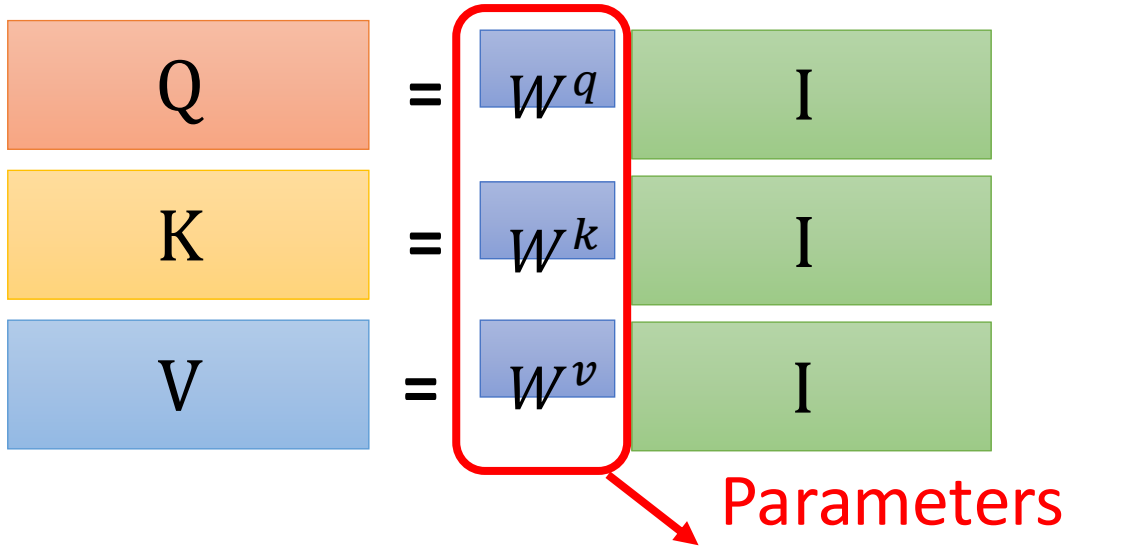
$b^1\ b^2\ b^3\ b^4$ = $v^1\ v^2\ v^3\ v^4$

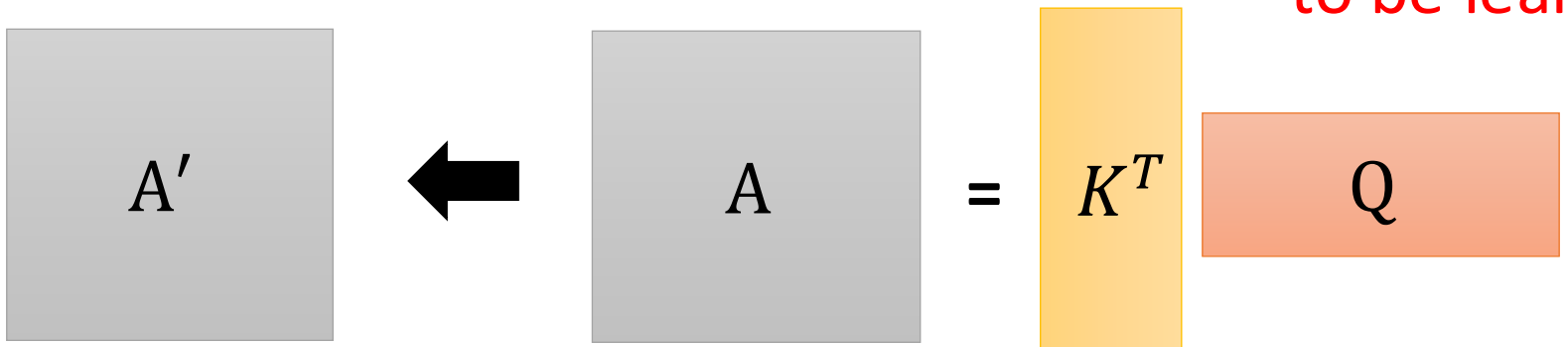
O V

$$\begin{matrix} \alpha'_{1,1} & \alpha'_{2,1} & \alpha'_{3,1} & \alpha'_{4,1} \\ \alpha'_{1,2} & \alpha'_{2,2} & \alpha'_{3,2} & \alpha'_{4,2} \\ \alpha'_{1,3} & \alpha'_{2,3} & \alpha'_{3,3} & \alpha'_{4,3} \\ \alpha'_{1,4} & \alpha'_{2,4} & \alpha'_{3,4} & \alpha'_{4,4} \end{matrix}$$

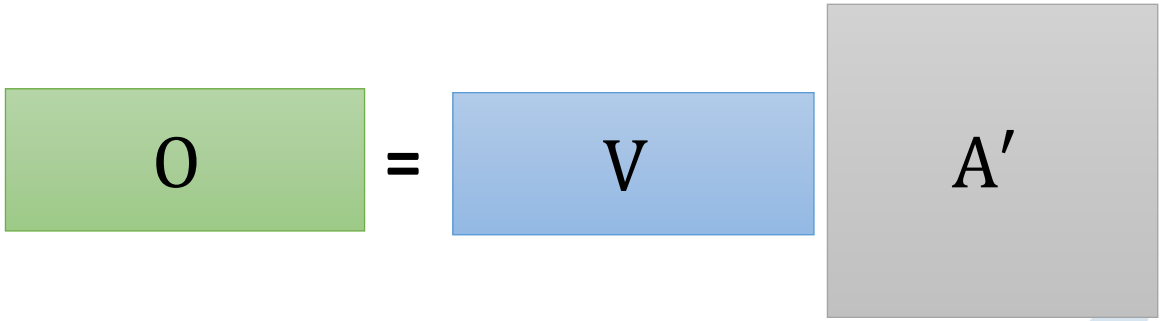
A'

Self-attention





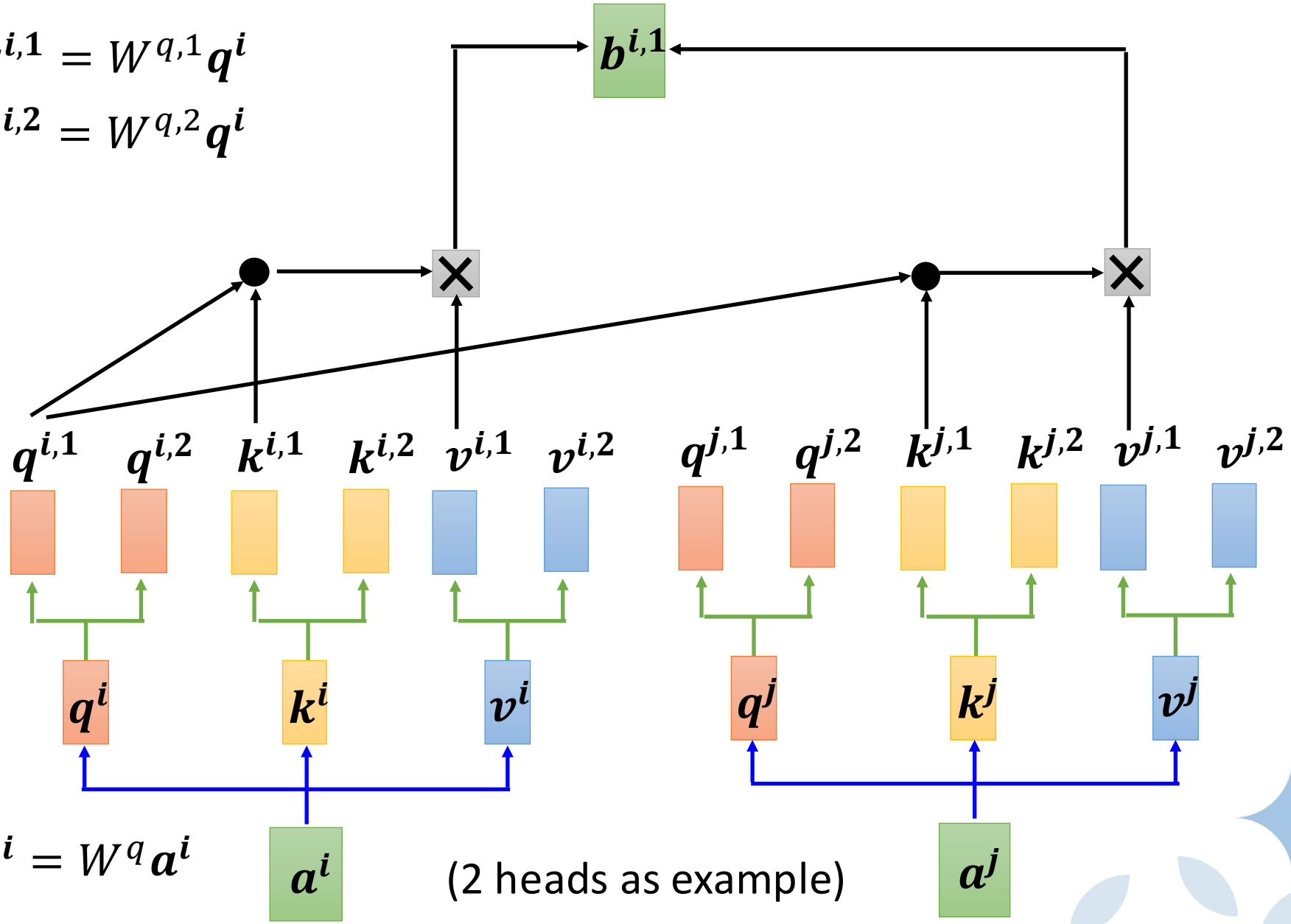
Attention Matrix



Multi-head Self-attention Different types of relevance

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

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



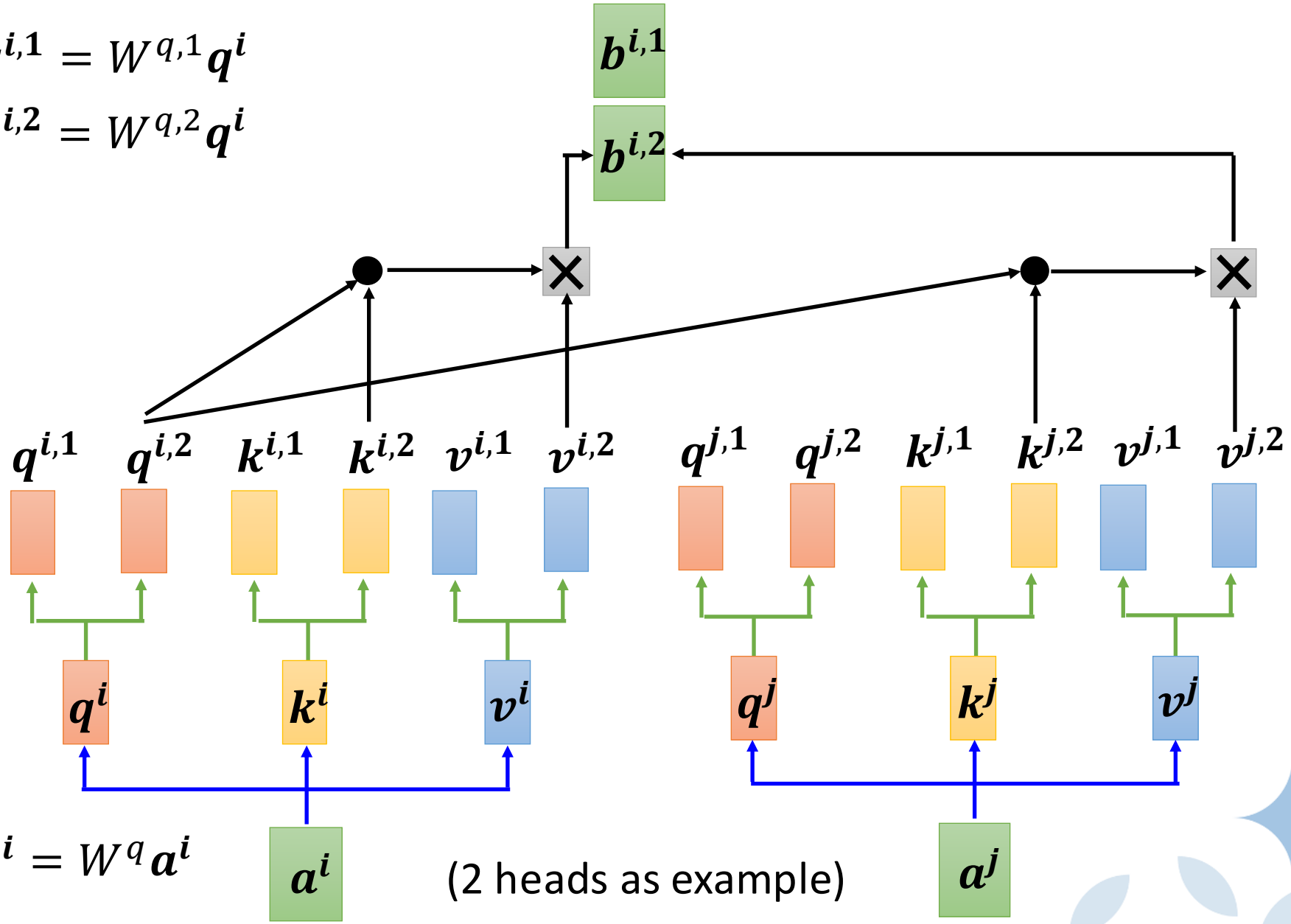
$$q^i = W^q a^i$$

(2 heads as example)

Multi-head Self-attention Different types of relevance

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

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

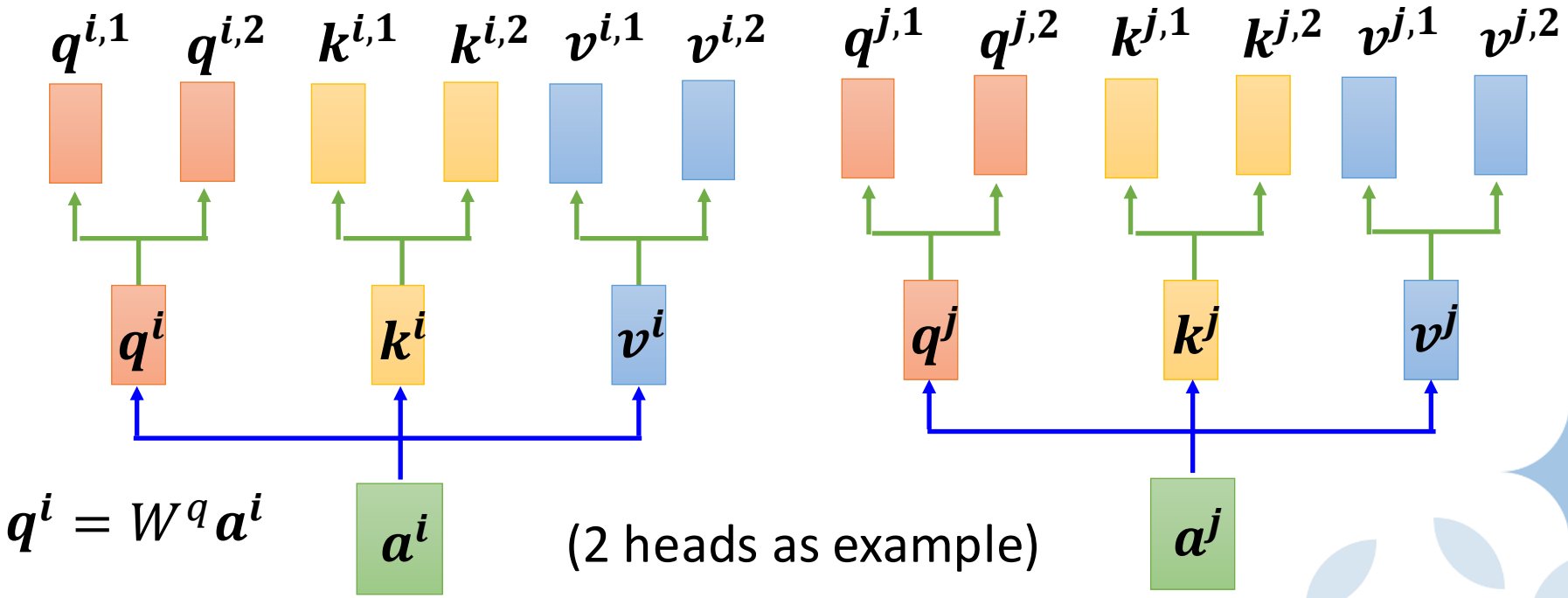


$$q^i = W^q a^i$$

(2 heads as example)

Multi-head Self-attention Different types of relevance

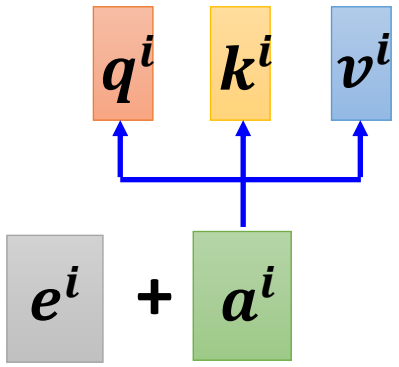
$$b^i = W^O \begin{bmatrix} b^{i,1} \\ b^{i,2} \end{bmatrix}$$



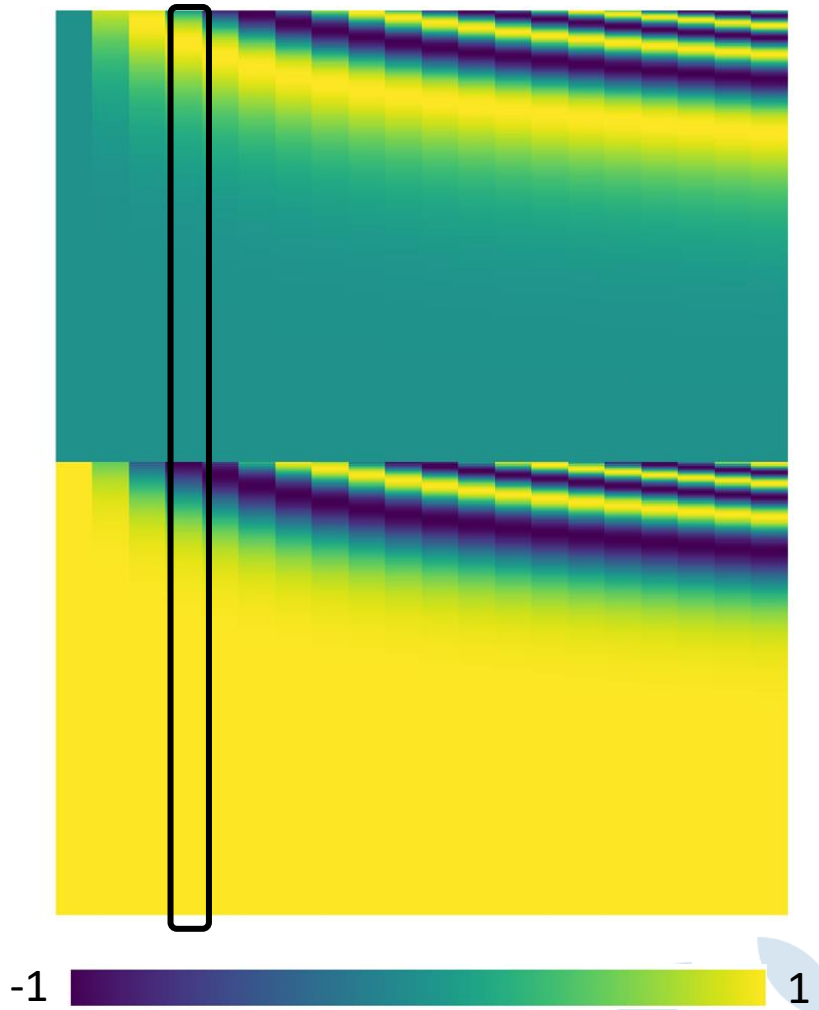
(2 heads as example)

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



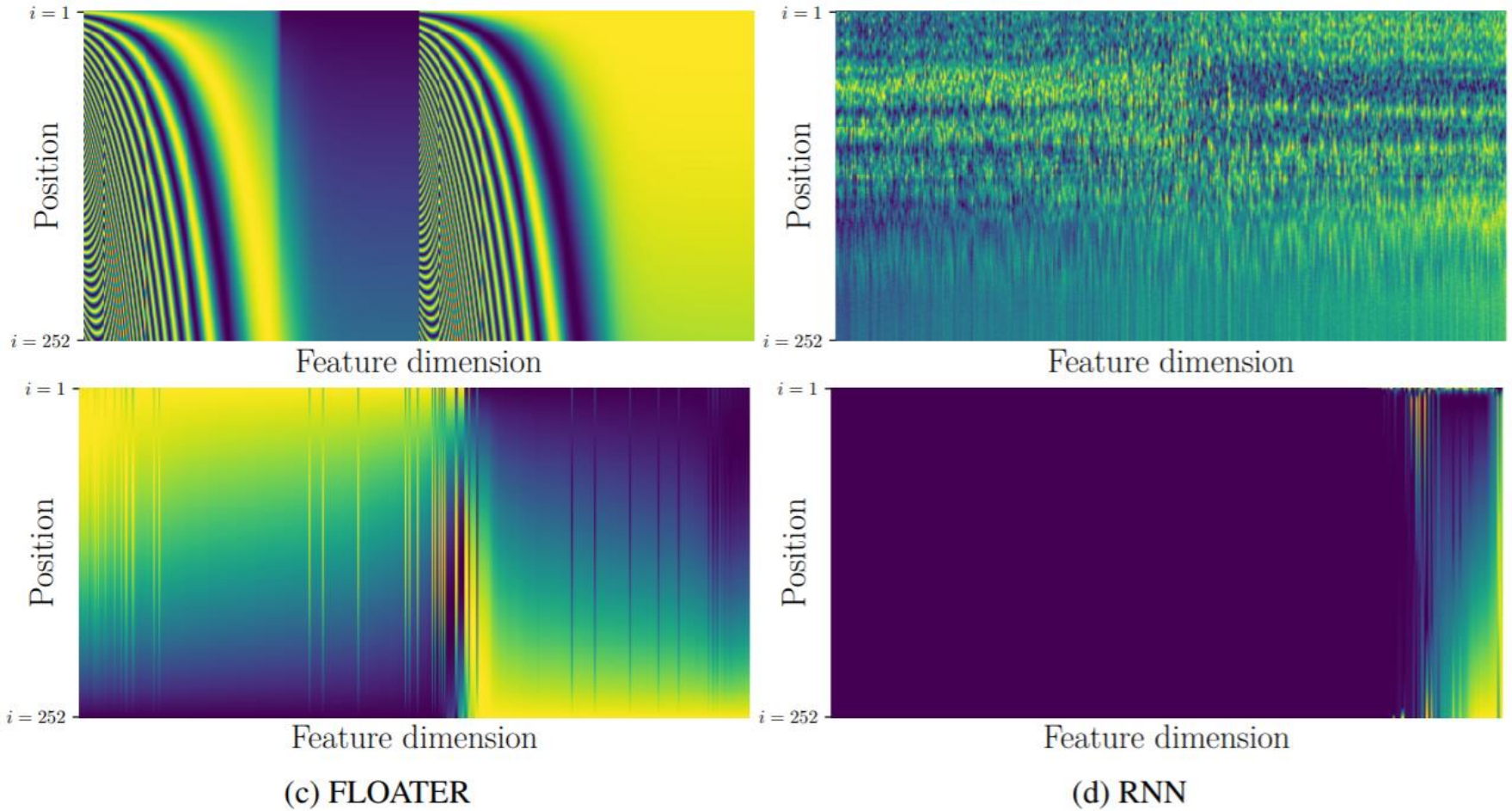
https://arxiv.org/abs/2003.09229

Table 1. Comparing position representation methods

Methods	Inductive	Data-Driven	Parameter Efficient
Sinusoidal (Vaswani et al., 2017)	✓	✗	✓
Embedding (Devlin et al., 2018)	✗	✓	✗
Relative (Shaw et al., 2018)	✗	✓	✓
This paper	✓	✓	✓

(a) Sinusoidal

(b) Position embedding



(c) FLOATER

(d) RNN

Many applications ...

Transformer

https://arxiv.org/abs/1706.03762

BERT

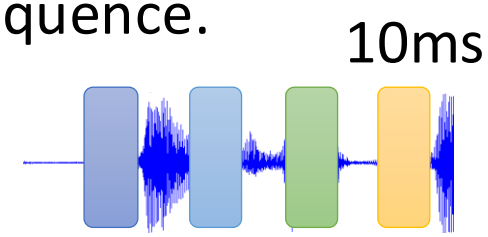
https://arxiv.org/abs/1810.04805

Widely used in Natural Langue Processing (NLP)!

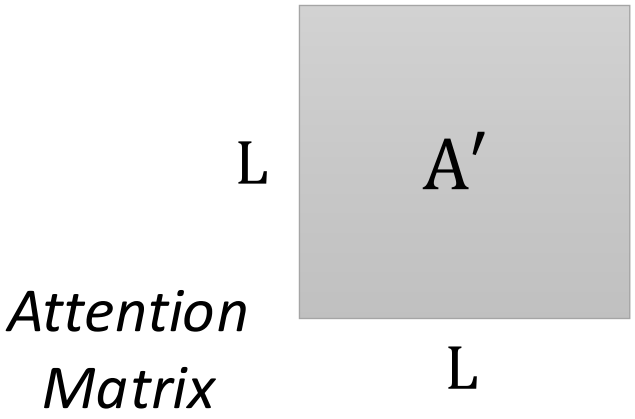
https://arxiv.org/abs/1910.12977

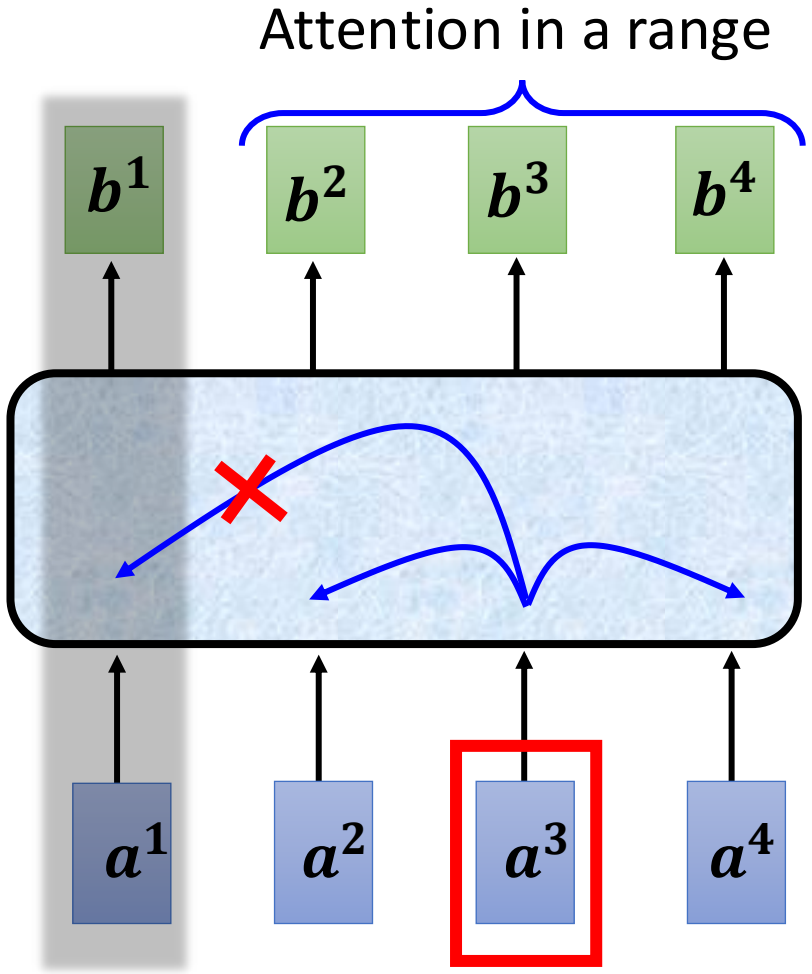
Self-attention for Speech

Speech is a very long vector sequence.



If input sequence is length L

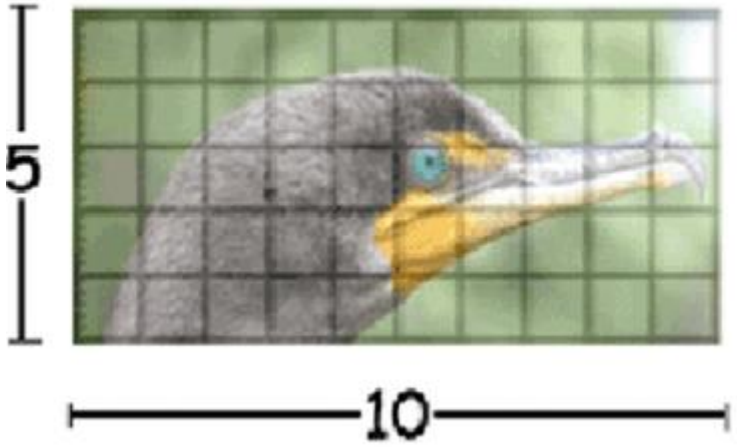


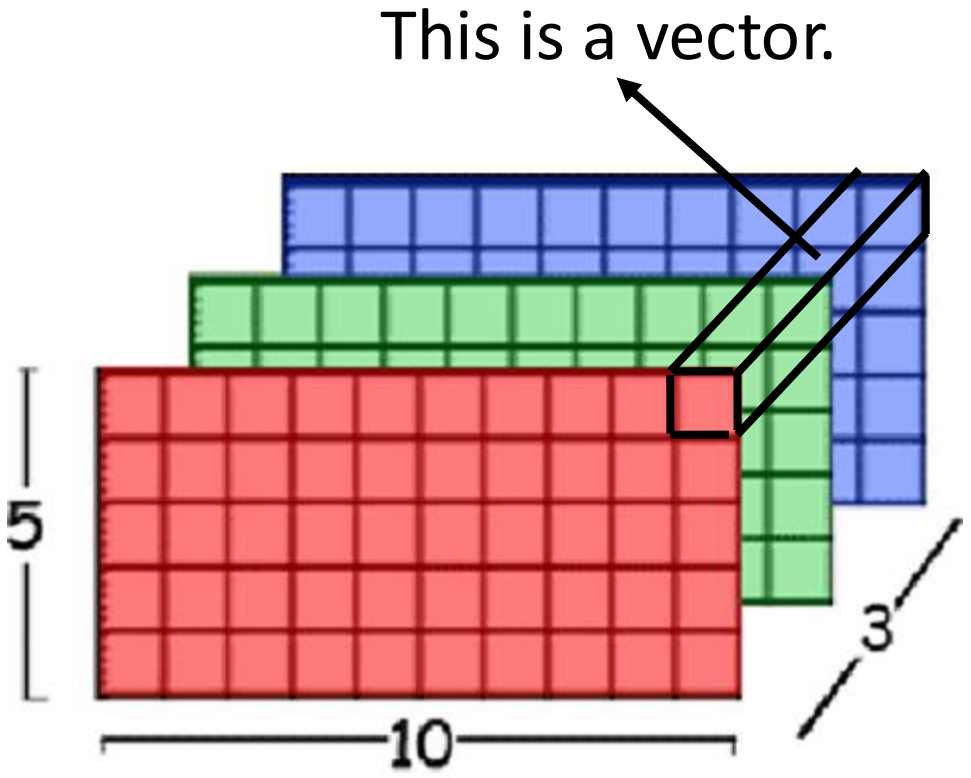


Truncated Self-attention

Self-attention for Image

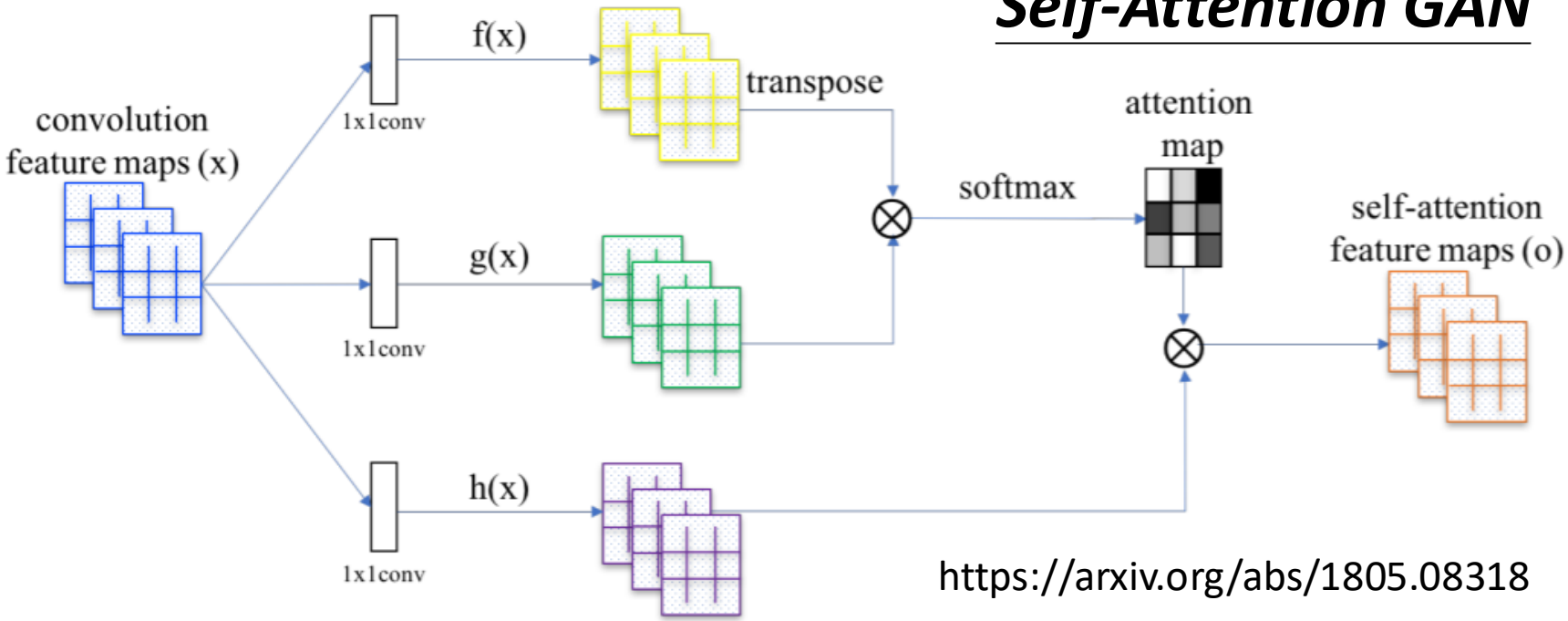
An **image** can also be considered as a **vector set**.





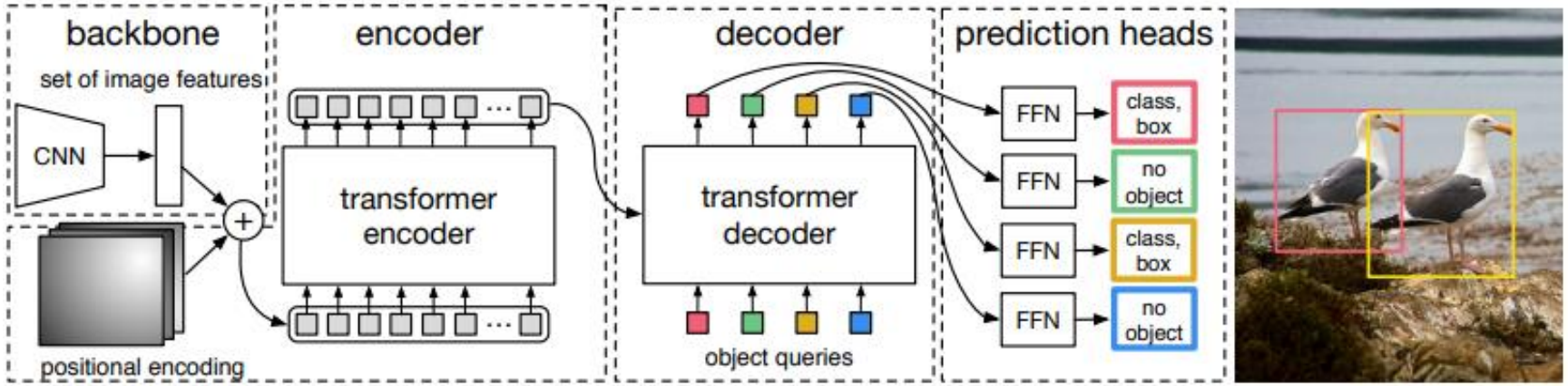
Source of image: https://www.researchgate.net/figure/Color-image-representation-and-RGB-matrix_fig15_282798184

Self-Attention GAN



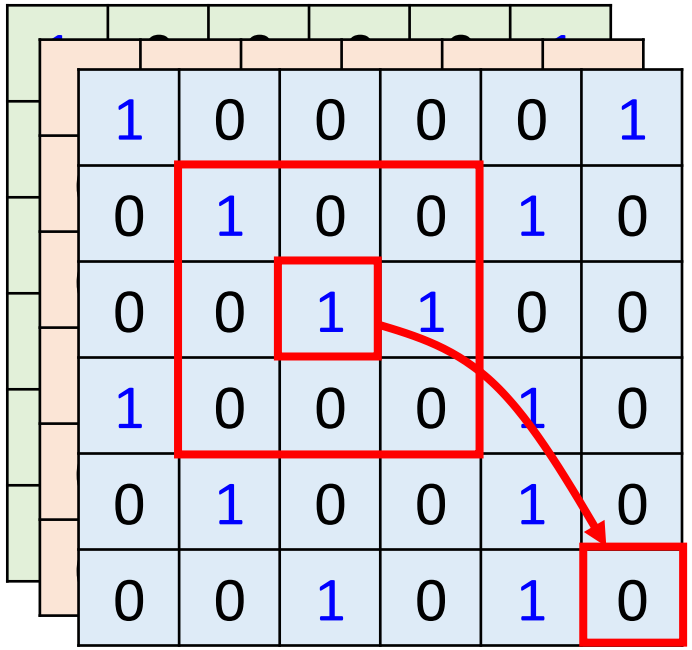
https://arxiv.org/abs/1805.08318

DEtection Transformer (DETR)



https://arxiv.org/abs/2005.12872

Self-attention v.s. CNN



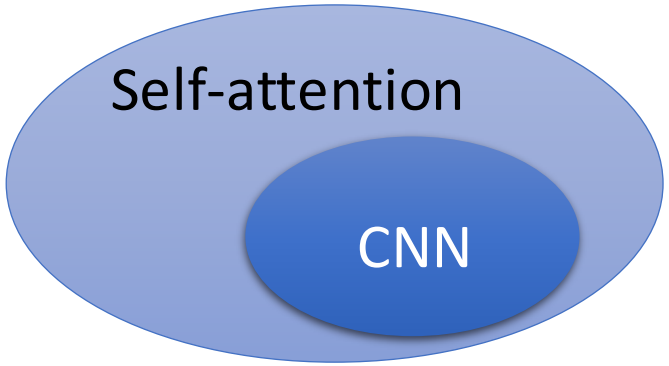
CNN: self-attention that can only attends in a receptive field

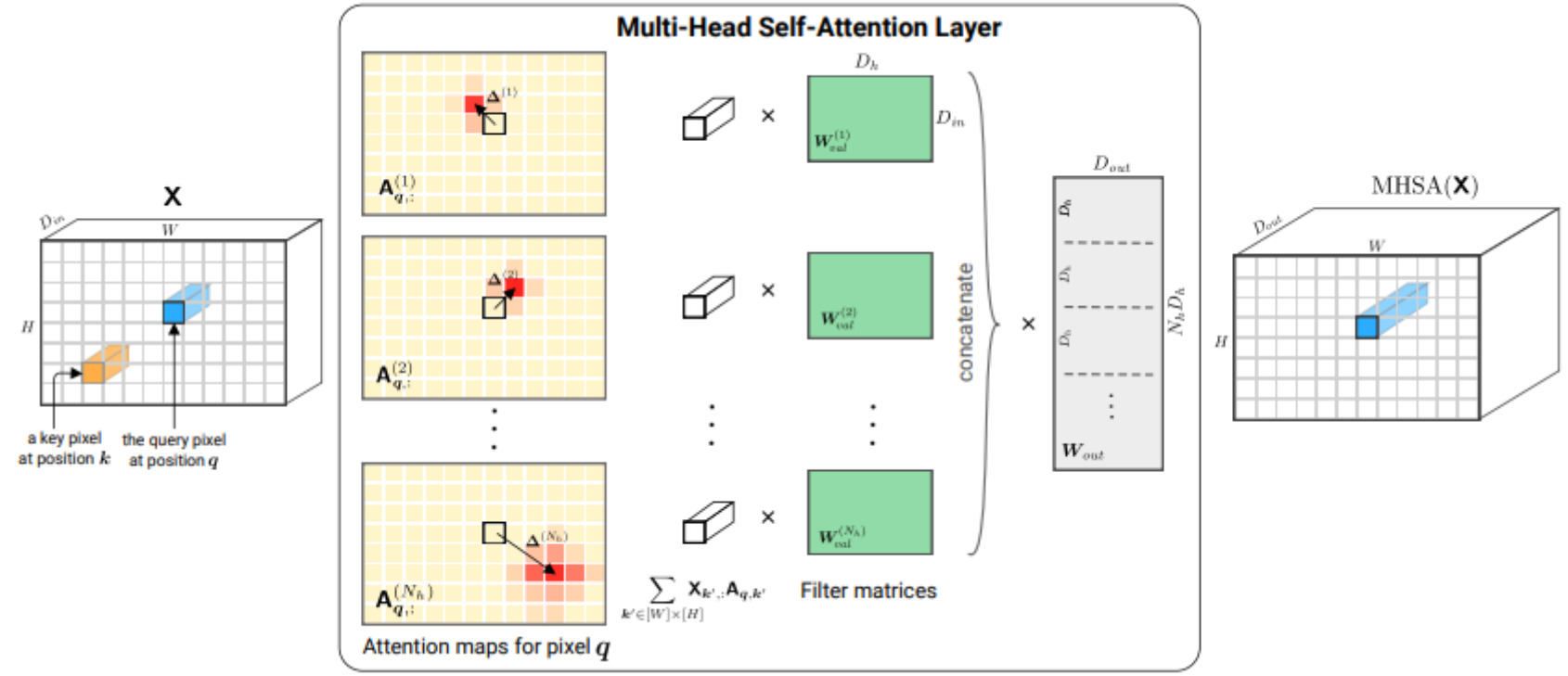
- CNN is simplified self-attention.

Self-attention: CNN with learnable receptive field

- Self-attention is the complex version of CNN.

Self-attention v.s. CNN





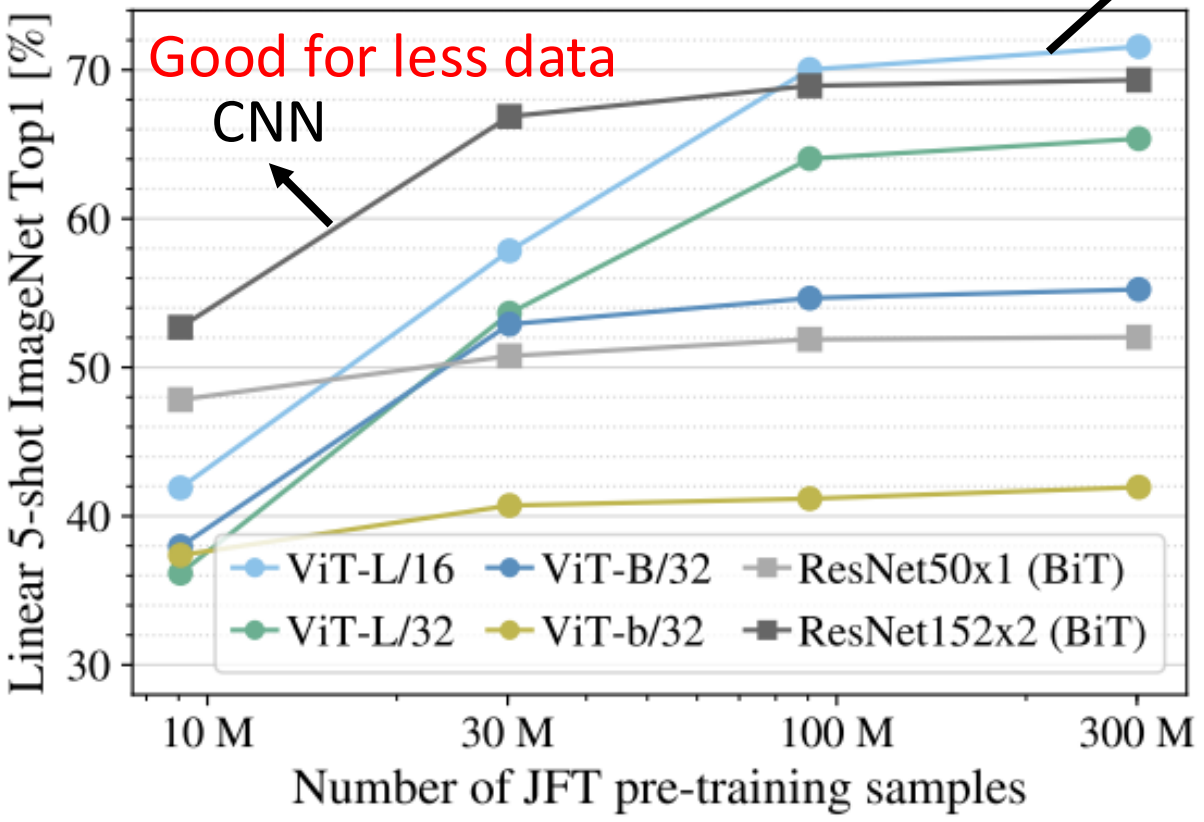
On the Relationship between Self-Attention and Convolutional Layers

https://arxiv.org/abs/1911.03584

Self-attention v.s. CNN

Good for more data

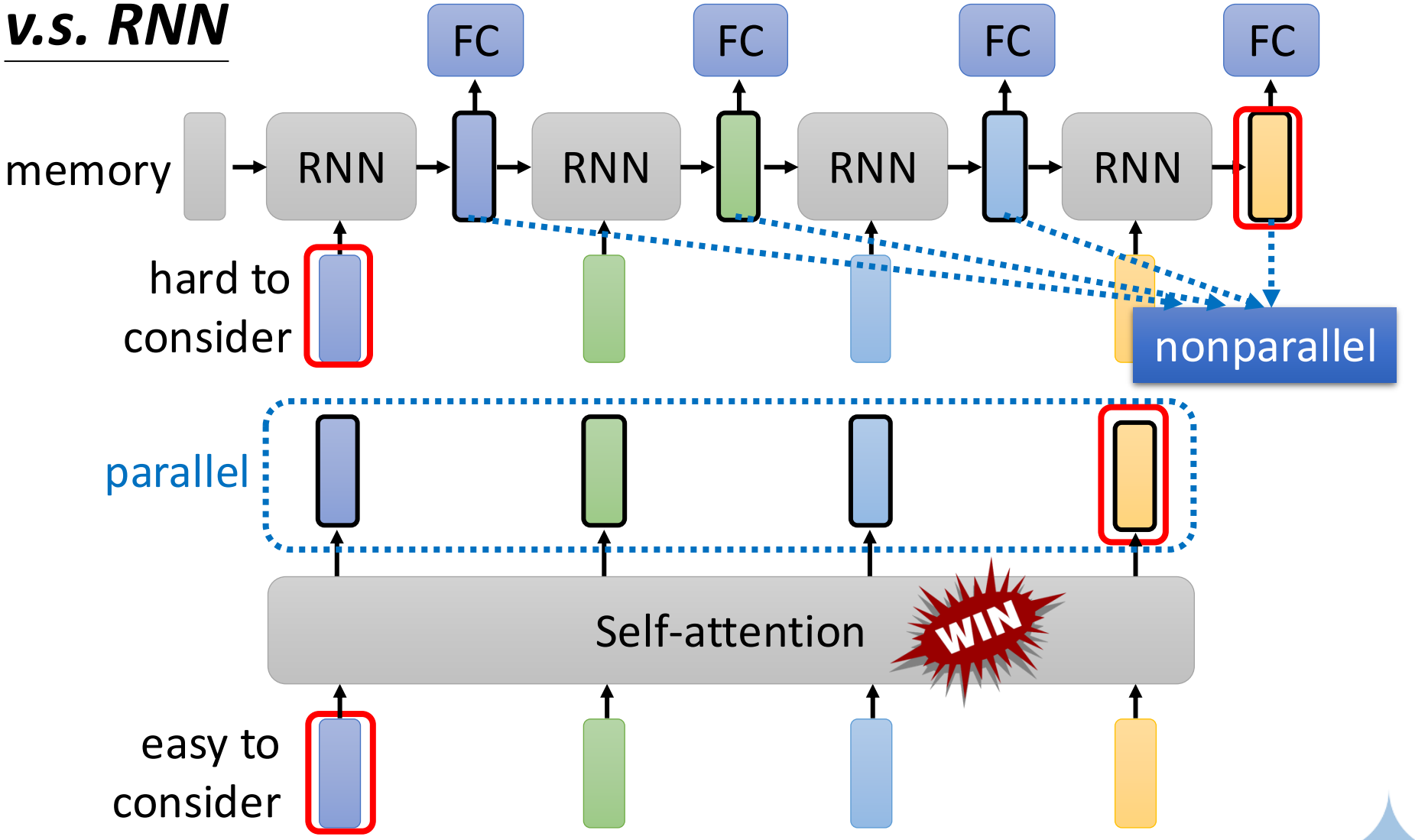
Self-attention



An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

https://arxiv.org/pdf/2010.11929.pdf

Self-attention v.s. RNN



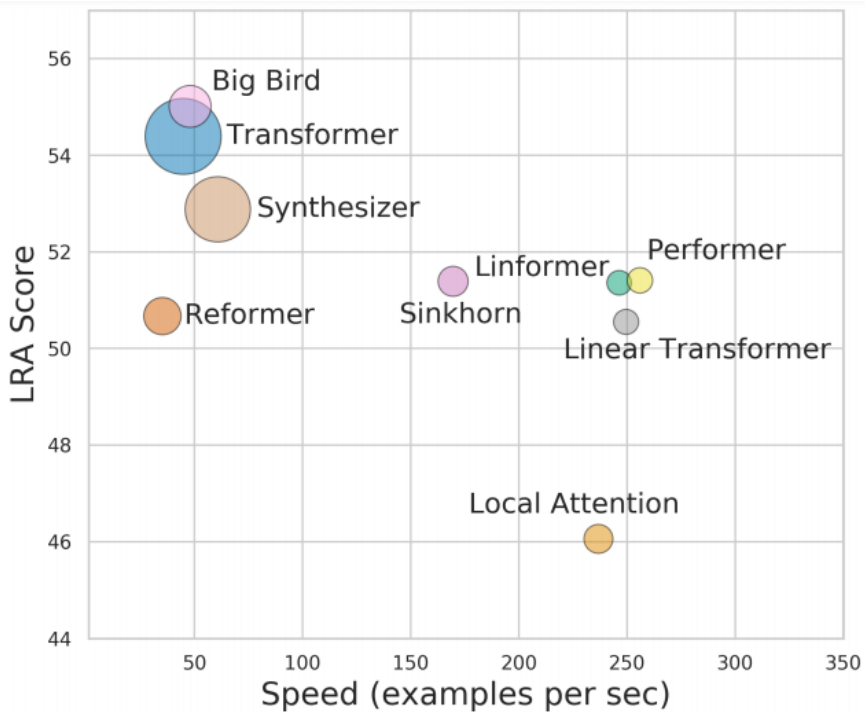
Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention

https://arxiv.org/abs/2006.16236

To Learn More …

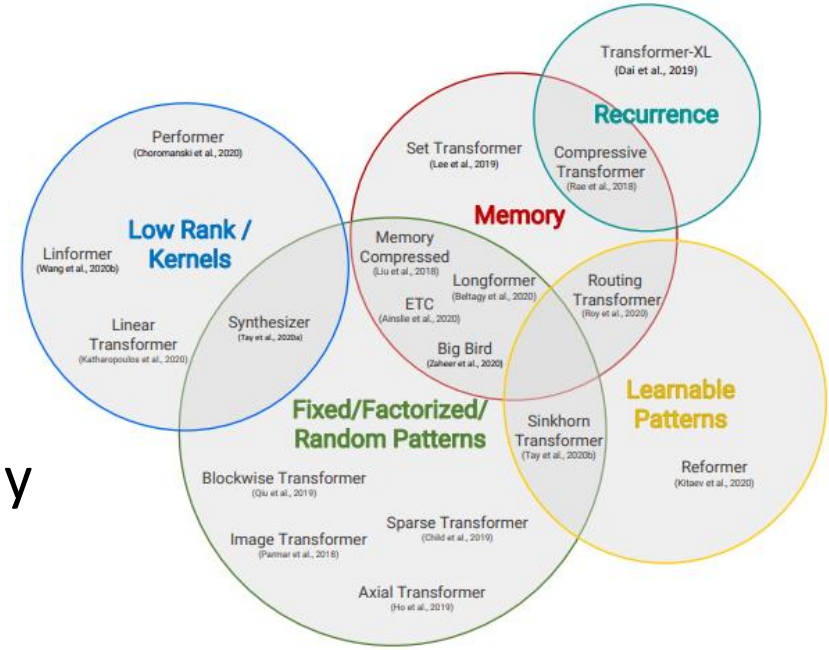
Long Range Arena: A Benchmark for Efficient Transformers

https://arxiv.org/abs/2011.04006



Efficient Transformers: A Survey

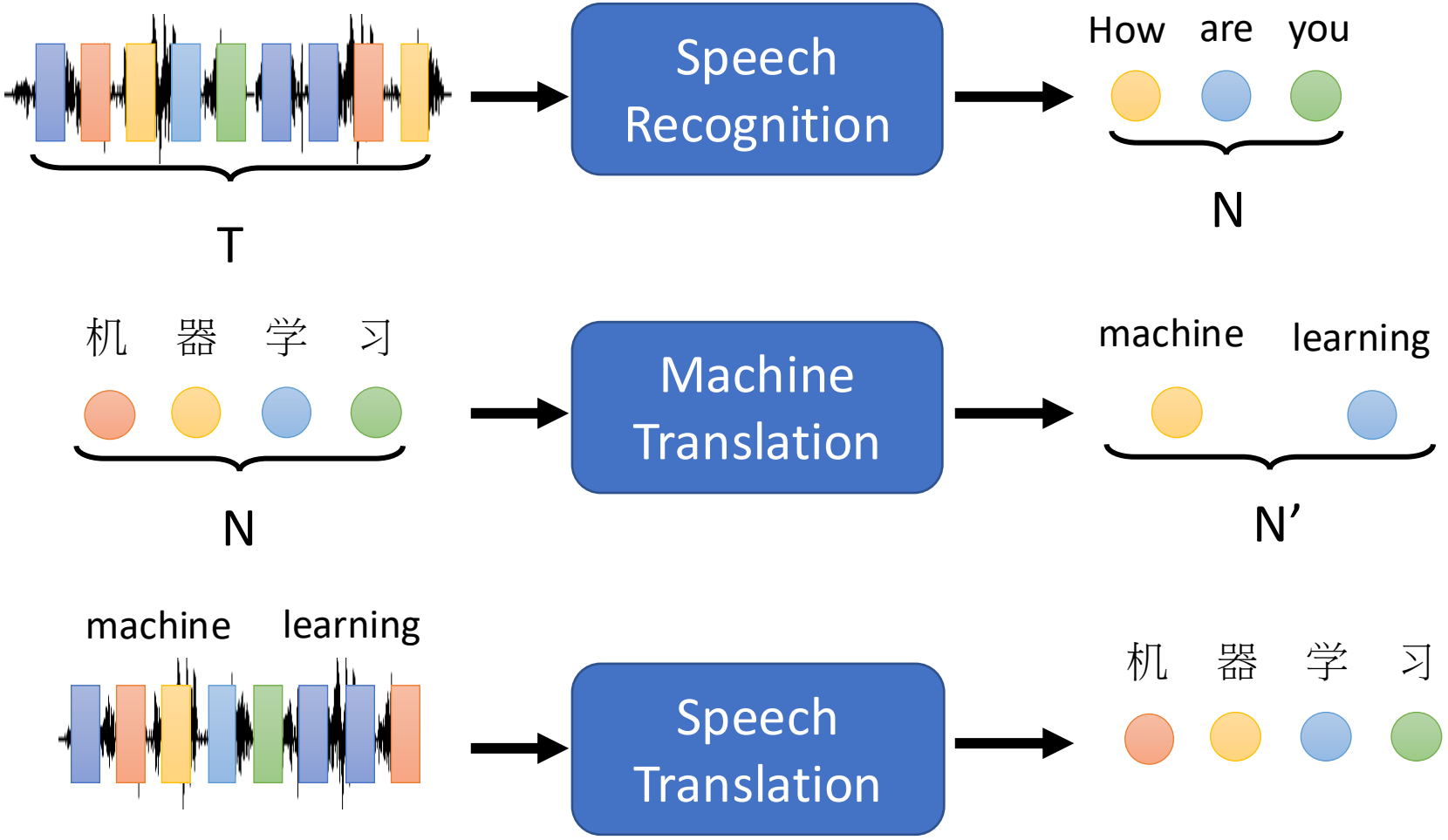
https://arxiv.org/abs/2009.06732



Sequence-to-sequence (Seq2seq)

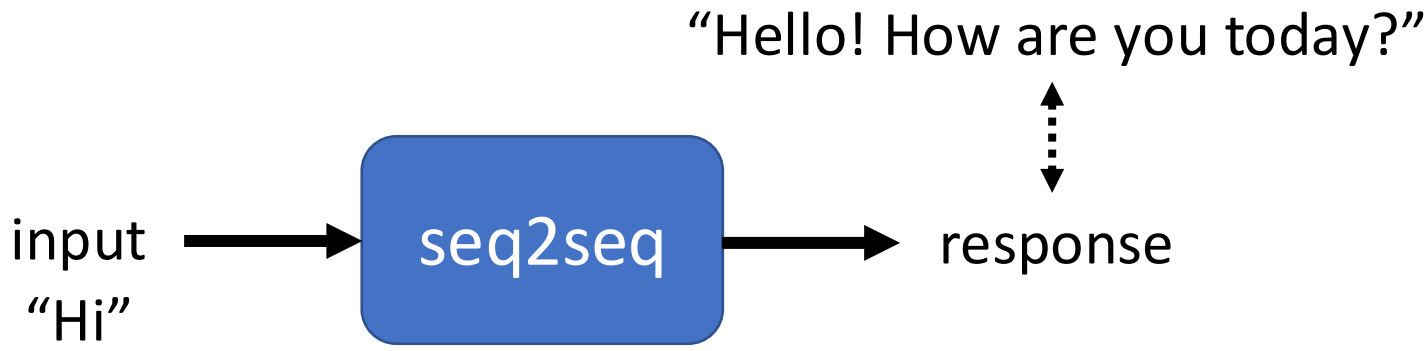
Input a sequence, output a sequence

The output length is determined by model.



Language without text

Seq2seq for Chatbot

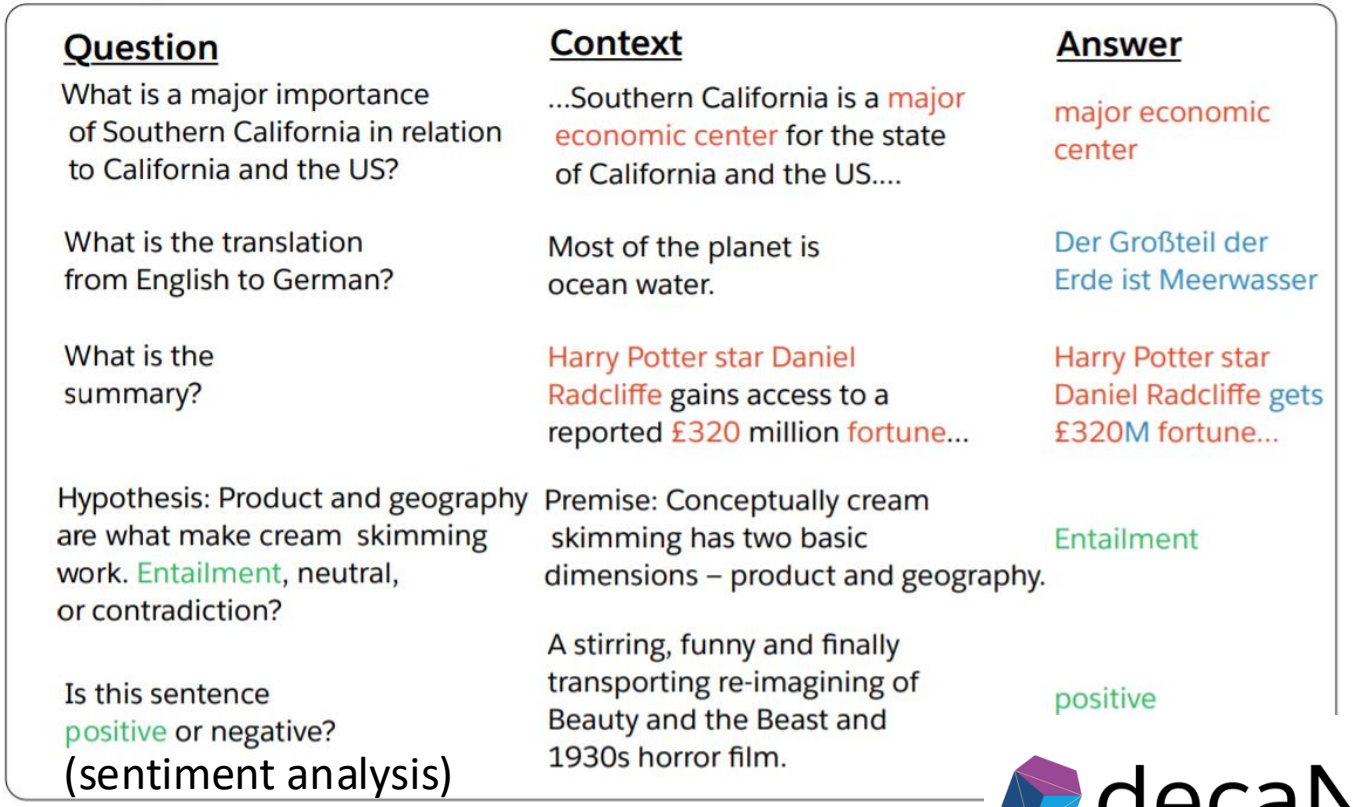


Training data:

[PERSON 1:] Hi
[PERSON 2:] Hello ! How are you today ?
[PERSON 1:] I am good thank you , how are you.
[PERSON 2:] Great, thanks ! My children and I were just about to watch Game of Thrones.
[PERSON 1:] Nice ! How old are your children?
[PERSON 2:] I have four that range in age from 10 to 21. You?
[PERSON 1:] I do not have children at the moment.
[PERSON 2:] That just means you get to keep all the popcorn for yourself.
[PERSON 1:] And Cheetos at the moment!
[PERSON 2:] Good choice. Do you watch Game of Thrones?
[PERSON 1:] No, I do not have much time for TV.
[PERSON 2:] I usually spend my time painting: but, I love the show.

Most Natural Language Processing applications …

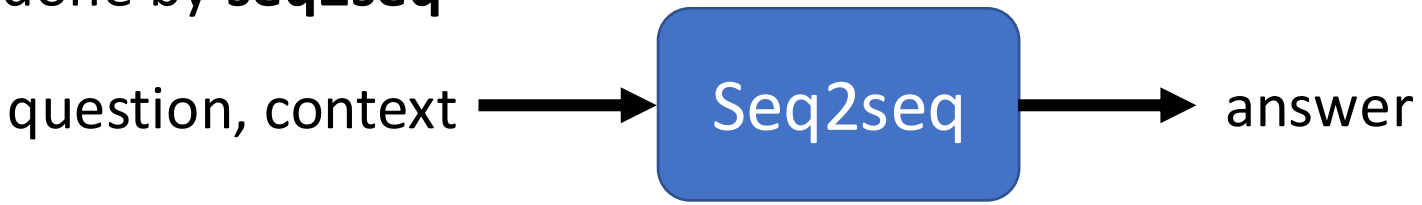
Question Answering (QA)



Question	Context	Answer
What is a major importance of Southern California in relation to California and the US?	...Southern California is a major economic center for the state of California and the US....	major economic center
What is the translation from English to German?	Most of the planet is ocean water.	Der Großteil der Erde ist Meerwasser
What is the summary?	Harry Potter star Daniel Radcliffe gains access to a reported £320 million fortune...	Harry Potter star Daniel Radcliffe gets £320M fortune...
Hypothesis: Product and geography are what make cream skimming work. Entailment, neutral, or contradiction?	Premise: Conceptually cream skimming has two basic dimensions – product and geography.	Entailment
Is this sentence positive or negative? (sentiment analysis)	A stirring, funny and finally transporting re-imagining of Beauty and the Beast and 1930s horror film.	positive

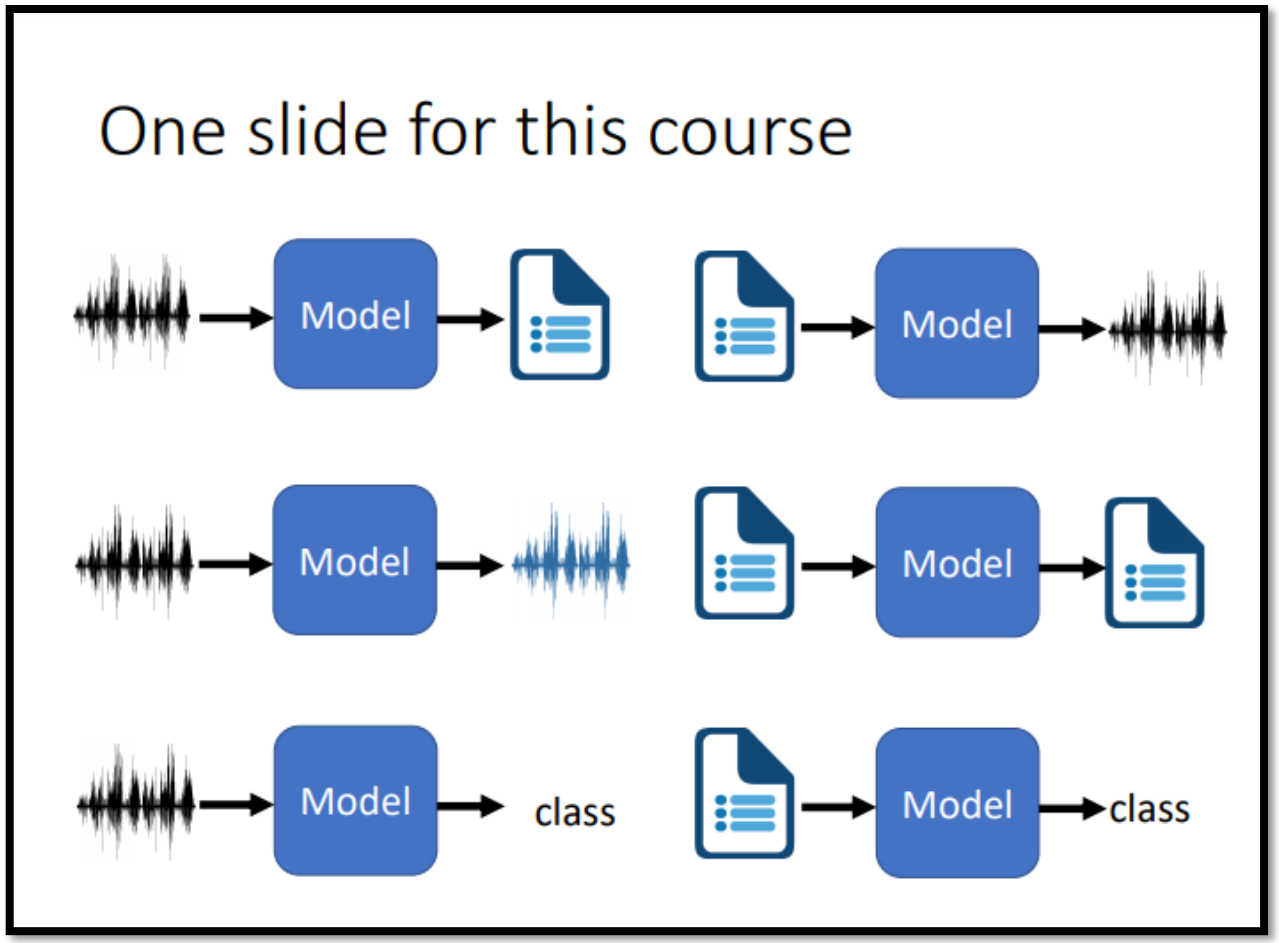
decaNLP

QA can be done by **seq2seq**



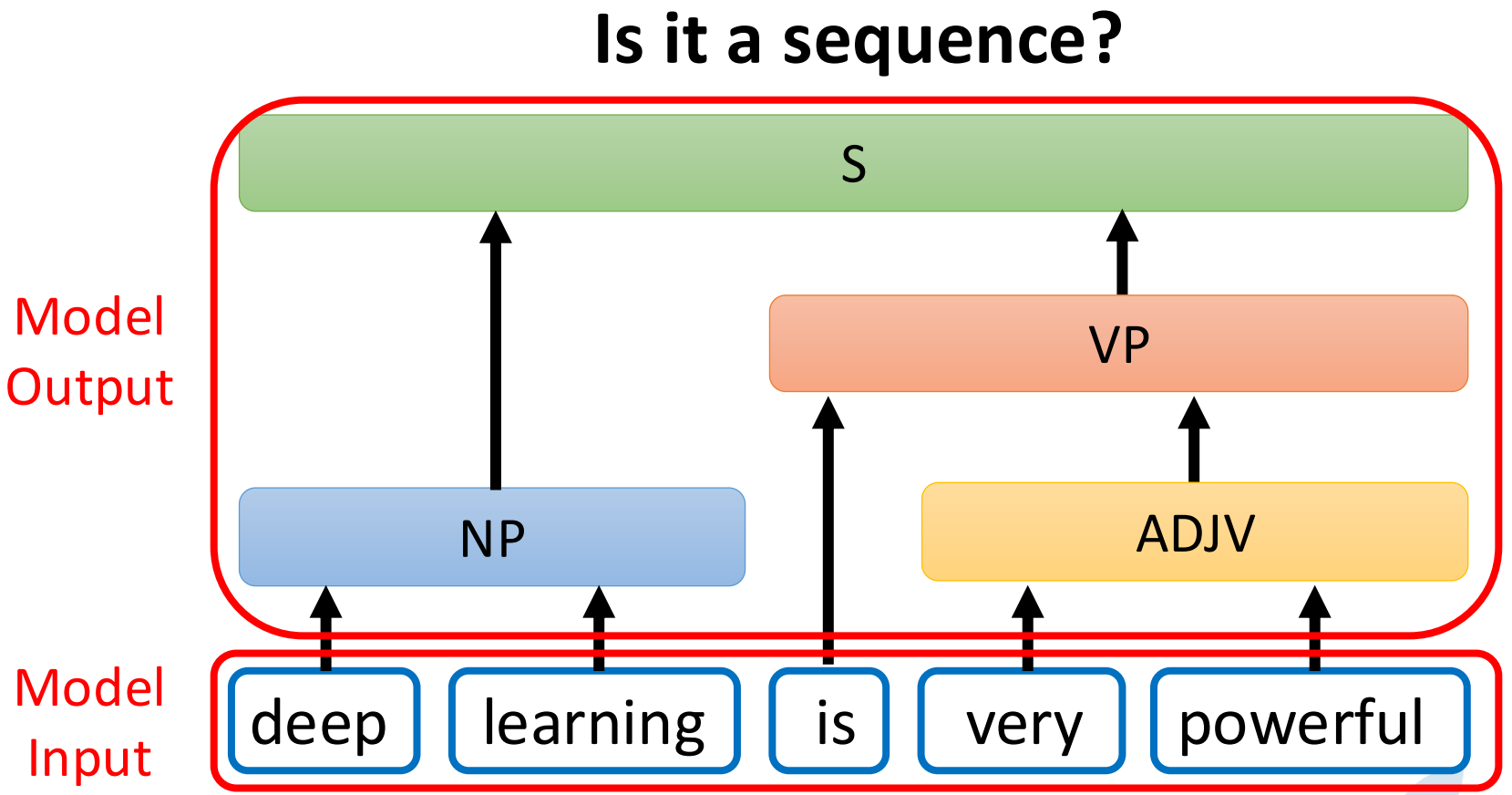
https://arxiv.org/abs/1806.08730
https://arxiv.org/abs/1909.03329

Deep Learning for Human Language Processing

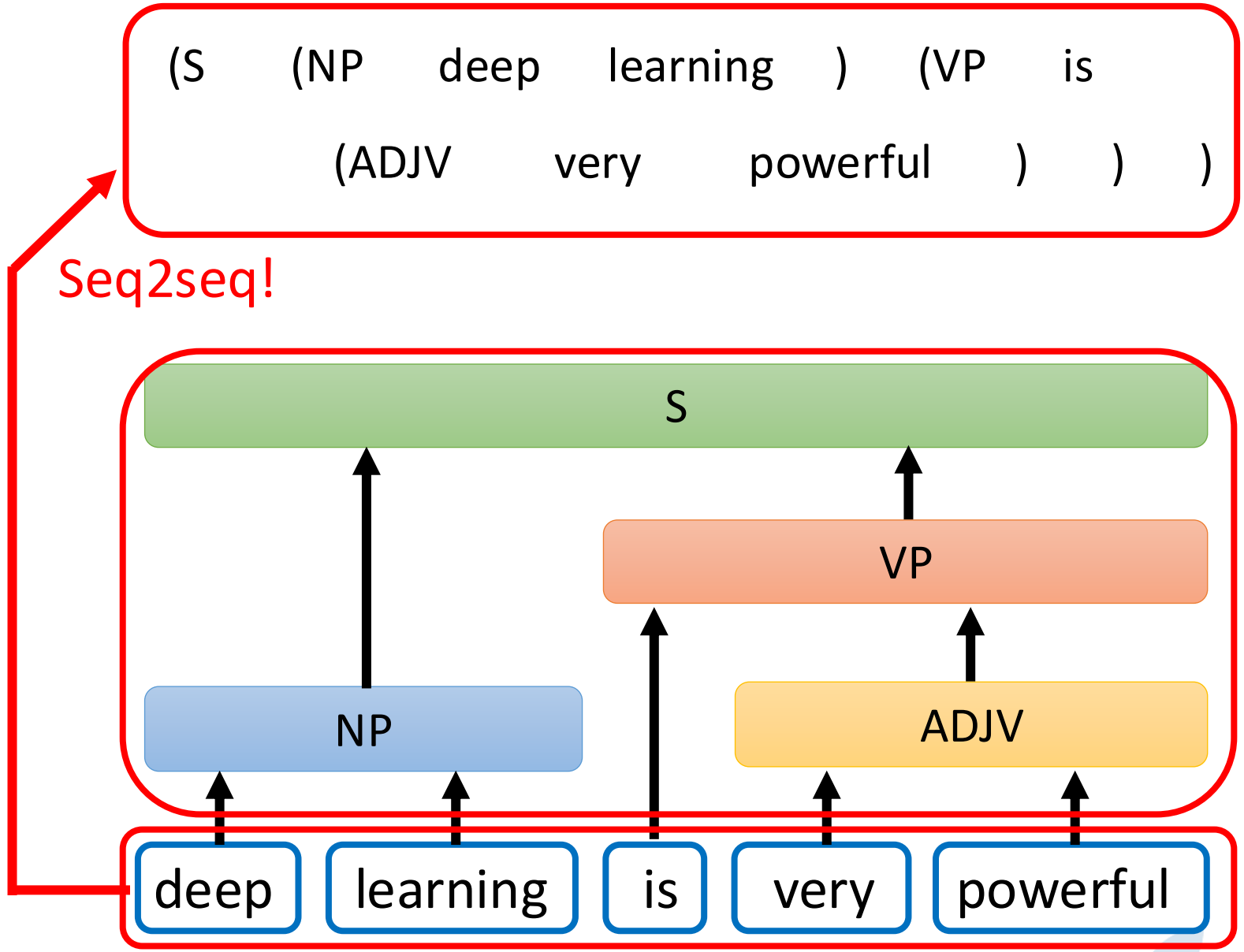


Source webpage: https://speech.ee.ntu.edu.tw/~hylee/dlhlp/2020-spring.html

Seq2seq for Syntactic Parsing



Seq2seq for Syntactic Parsing



Seq2seq for Syntactic Parsing

(S (NP deep learning) (VP is (ADJV very powerful)))

Grammar as a Foreign Language

Oriol Vinyals*
Google
vinyals@google.com

Lukasz Kaiser*
Google
lukaszkaiser@google.com

Terry Koo
Google
terrykoo@google.com

Slav Petrov
Google
slav@google.com

Ilya Sutskever
Google
ilyasu@google.com

Geoffrey Hinton
Google
geoffhinton@google.com

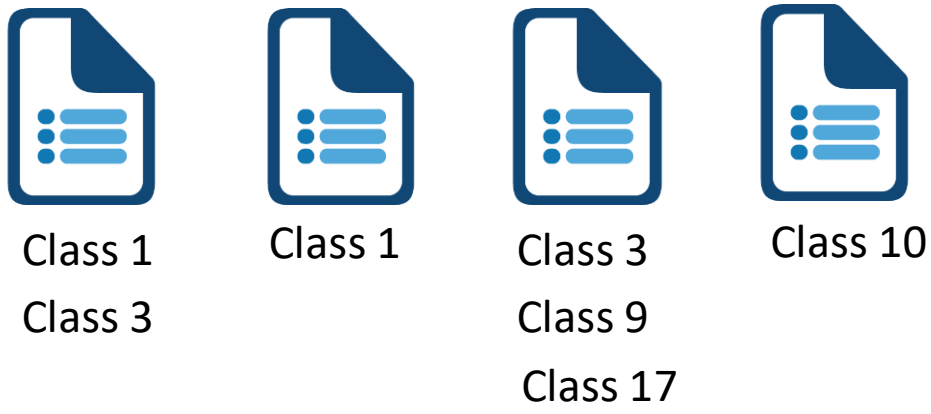
https://arxiv.org/abs/1412.7449

deep learning is very powerful

Seq2seq for Multi-label Classification

c.f. Multi-class Classification

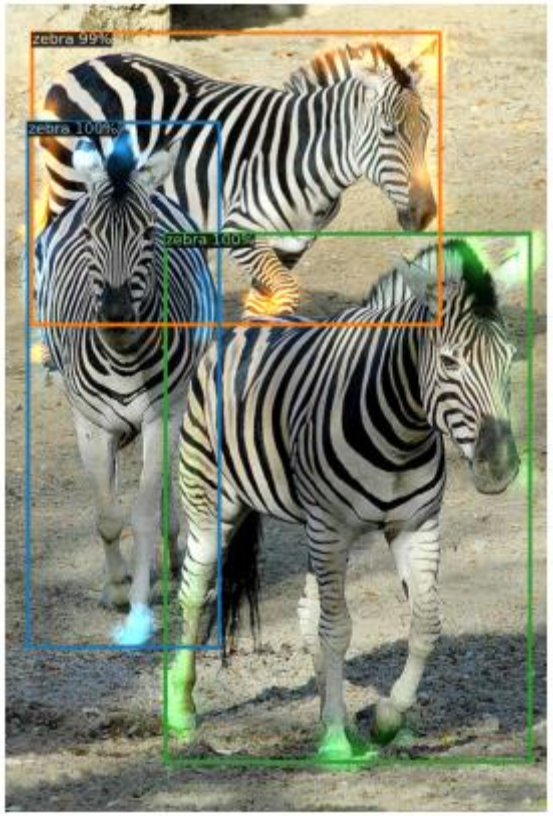
An object can belong to multiple classes.

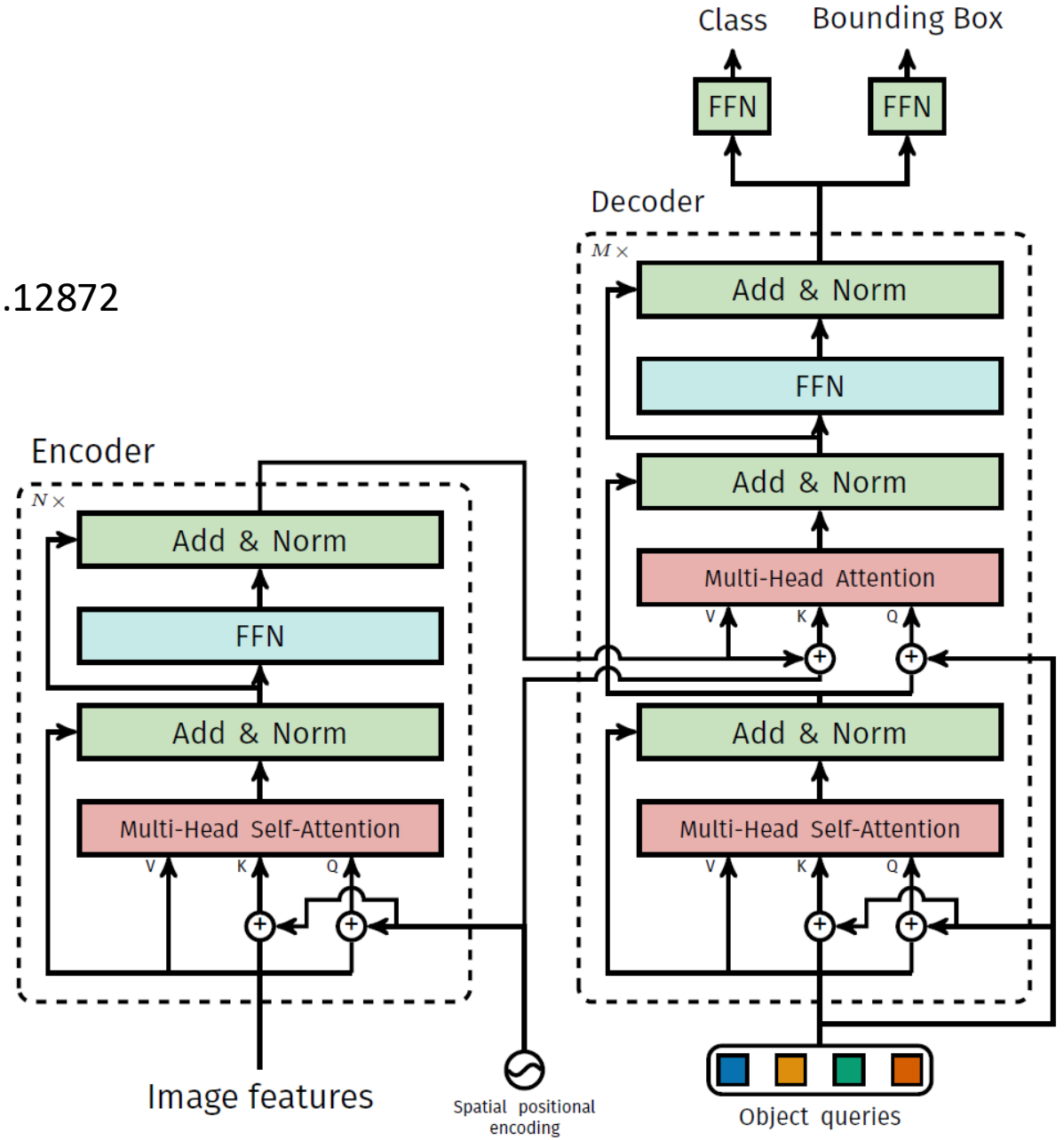


https://arxiv.org/abs/1909.03434
https://arxiv.org/abs/1707.05495

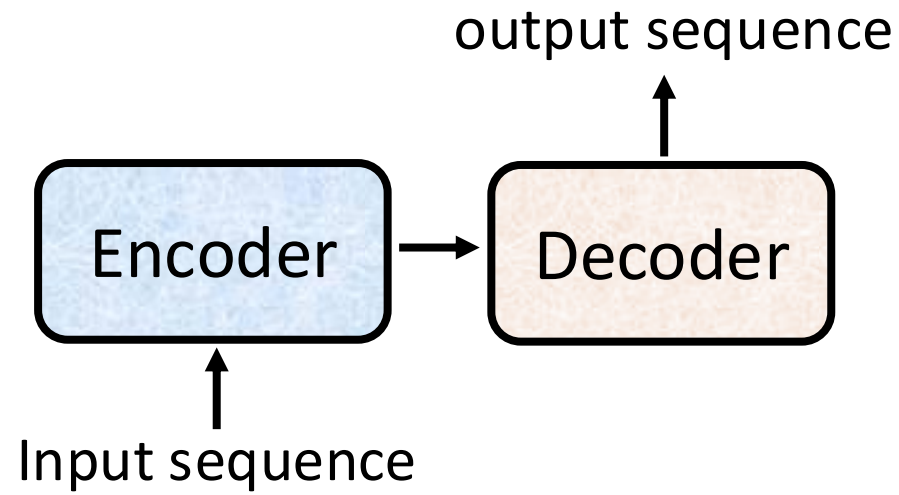
Seq2seq for Object Detection

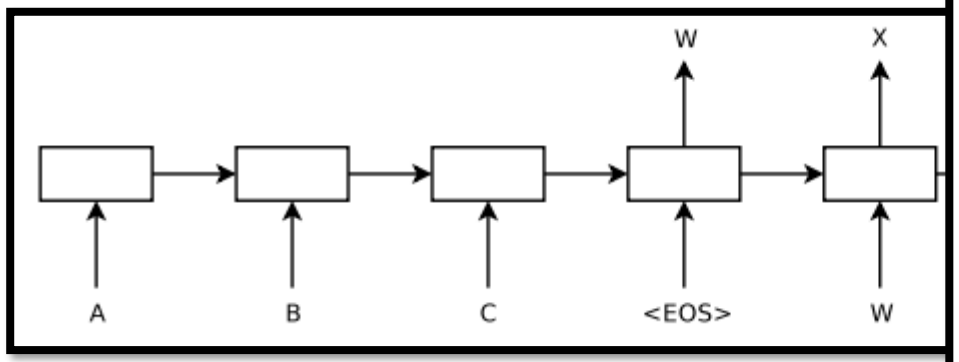
https://arxiv.org/abs/2005.12872





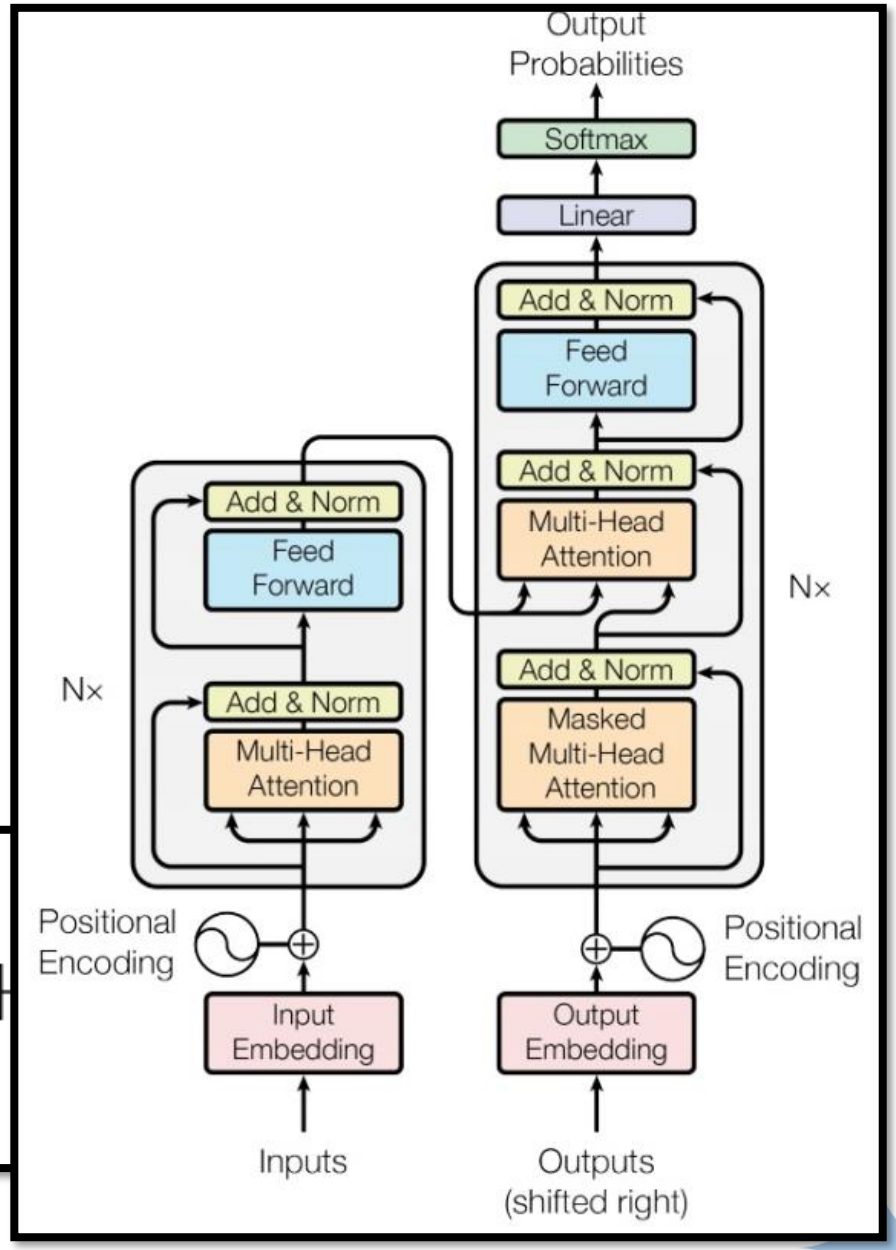
Seq2seq





Sequence to Sequence Learning with Neural Networks

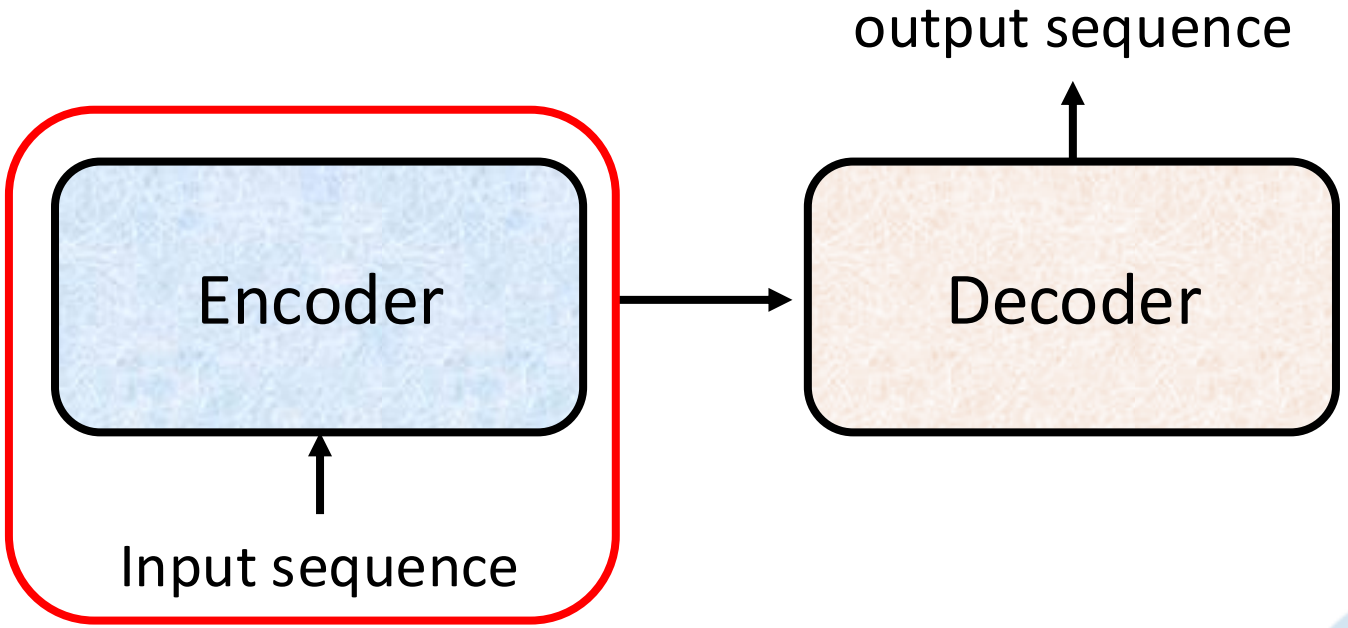
https://arxiv.org/abs/1409.3215



Transformer

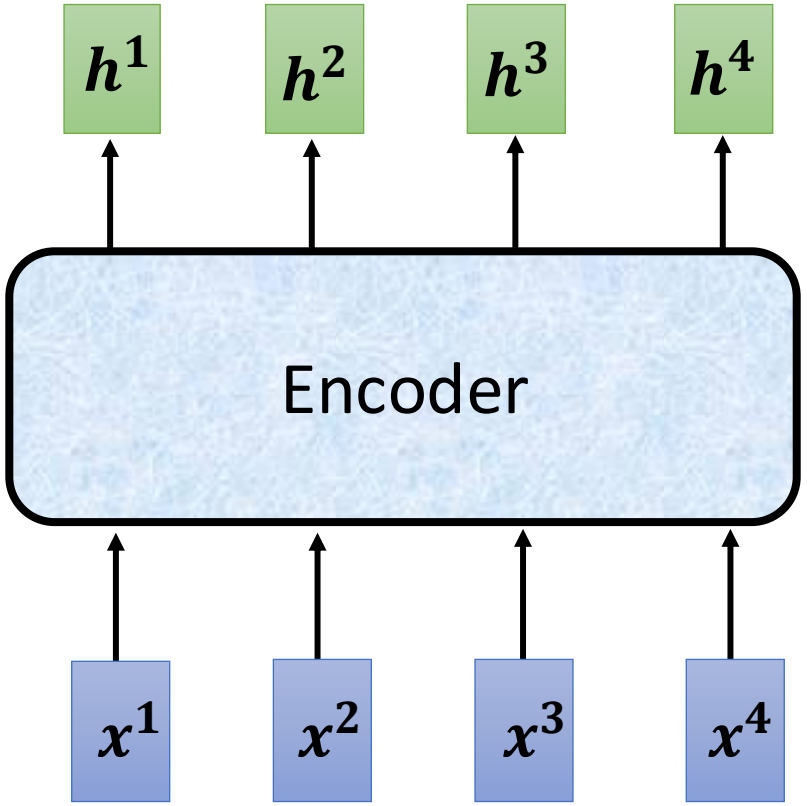
https://arxiv.org/abs/1706.03762

Encoder

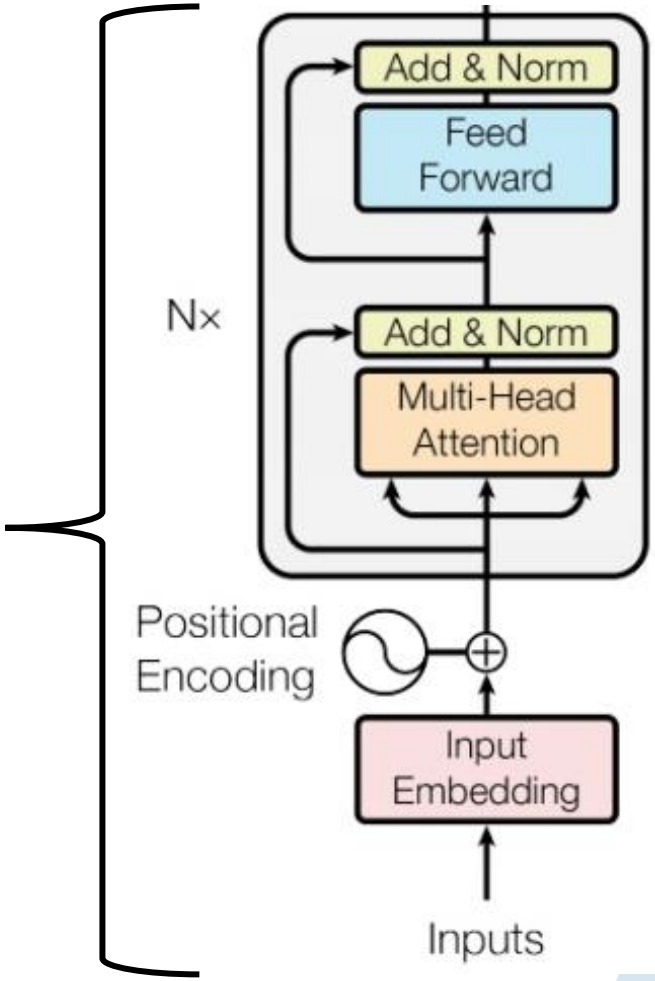


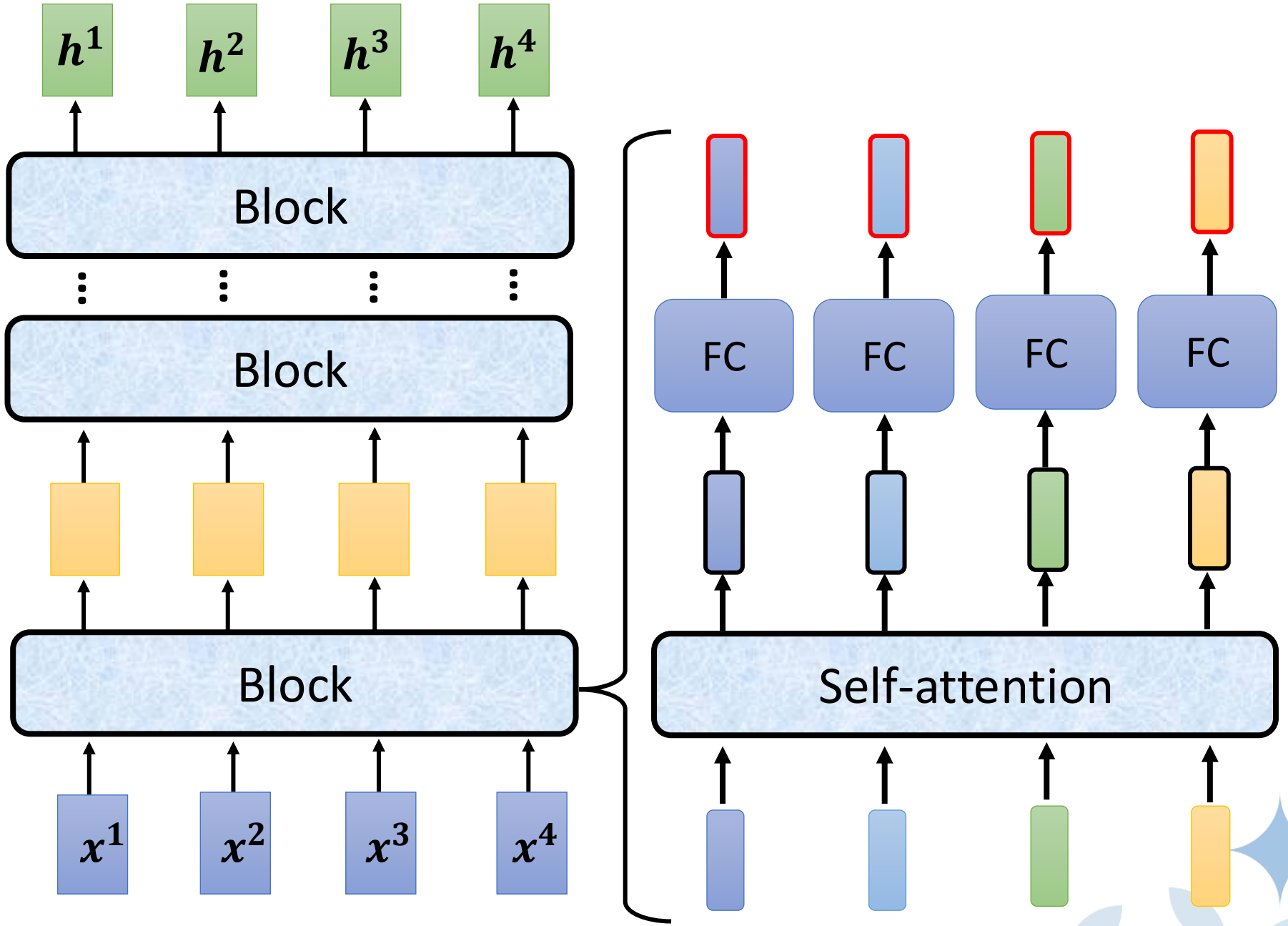
Encoder

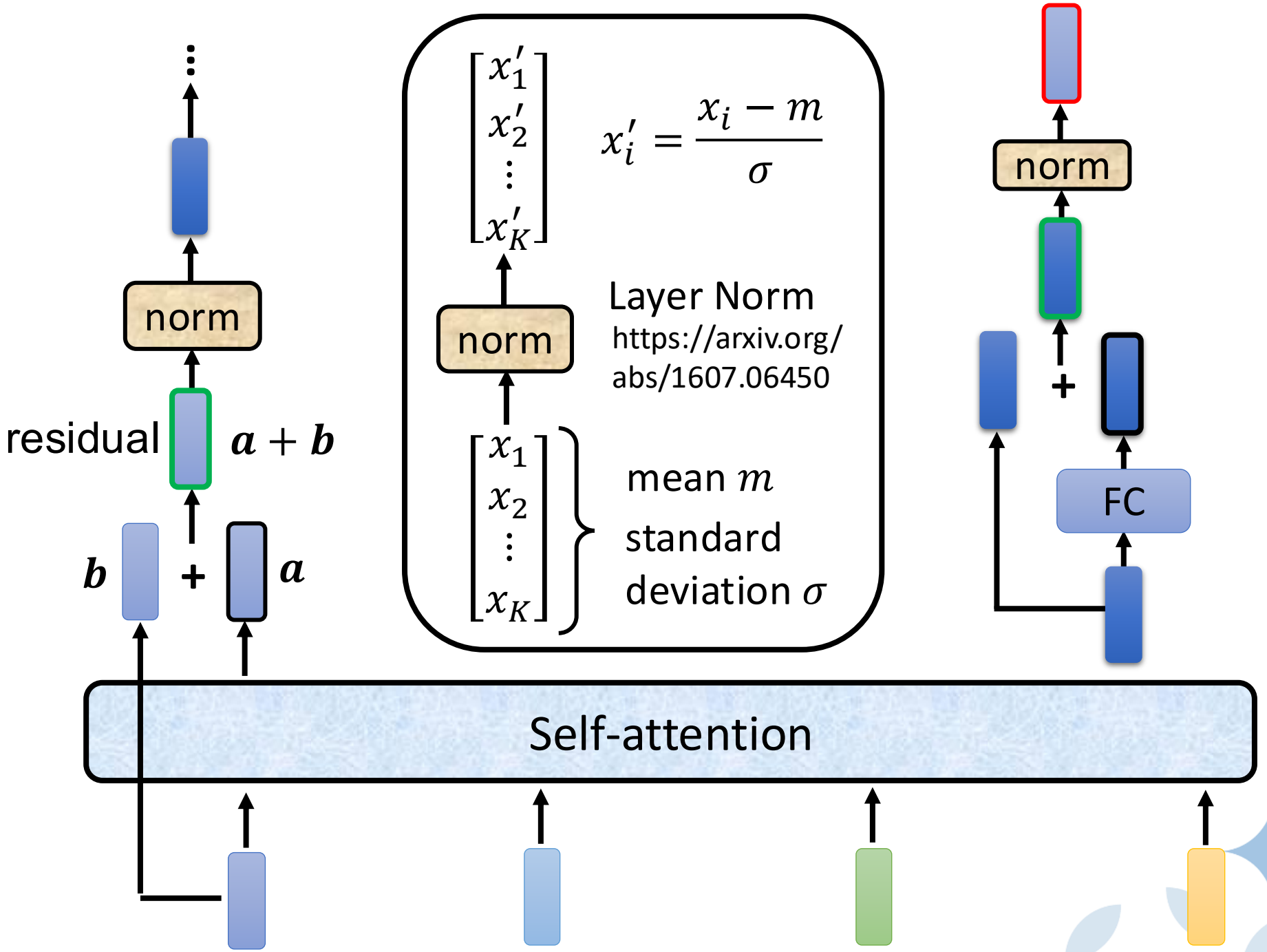
You can use **RNN** or **CNN**.



Transformer's Encoder

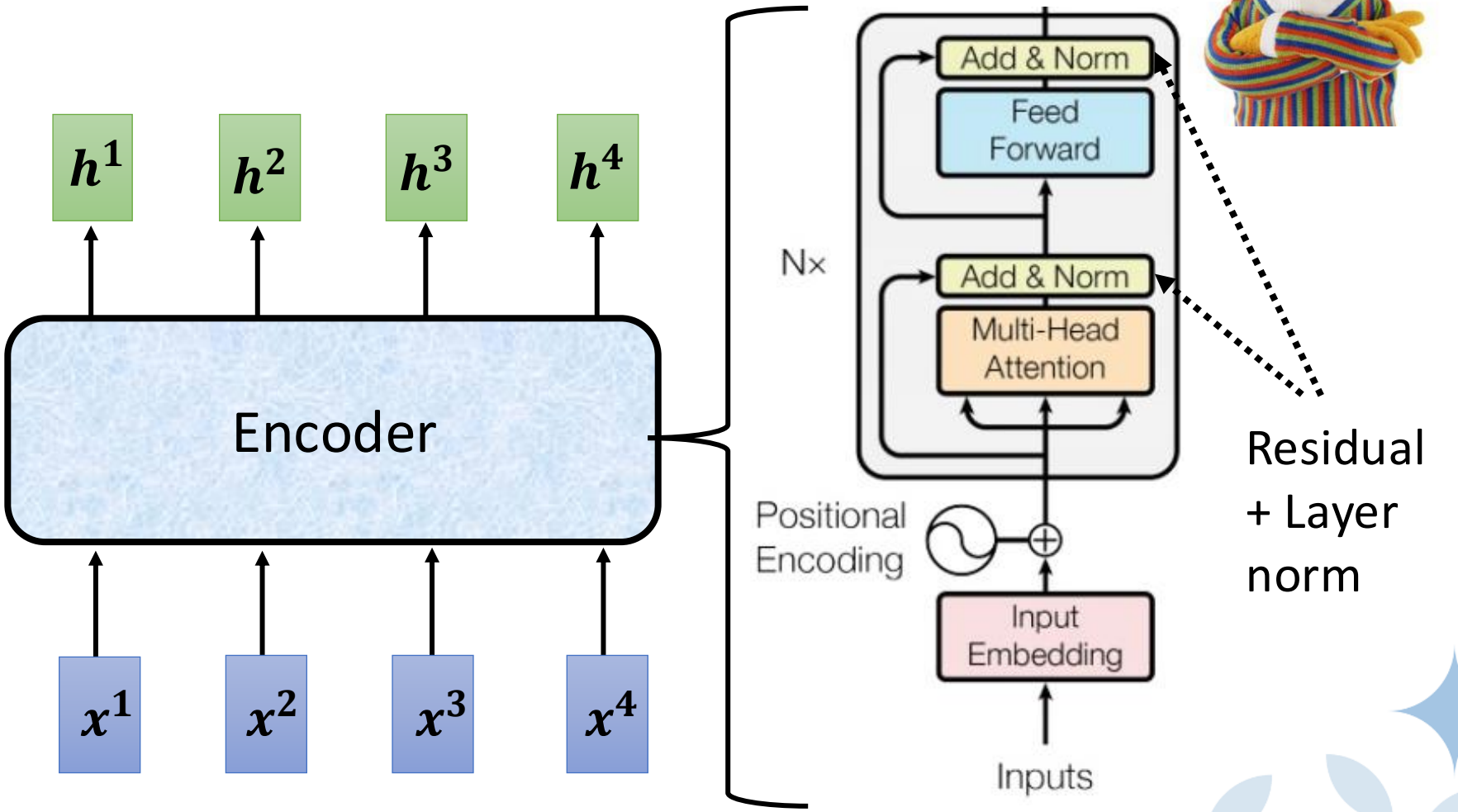



h^1
h^2
h^3
h^4
Block
Block
Block
x^1
x^2
x^3
x^4
FC
FC
FC
FC
Self-attention


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

BERT

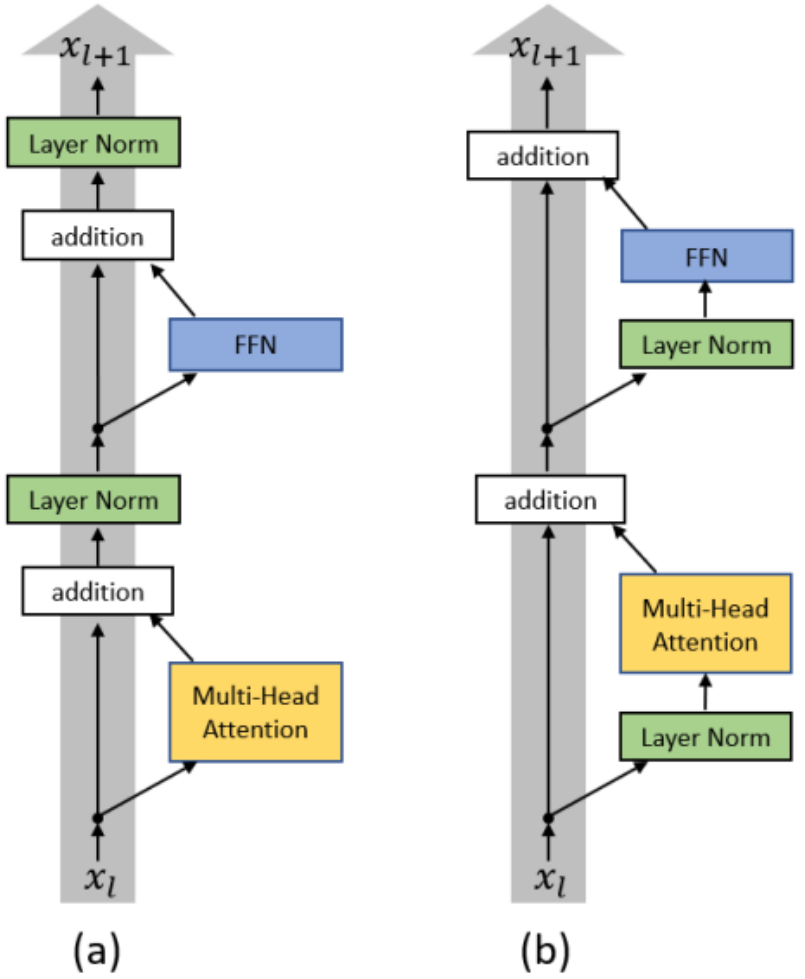
I use the **same** network architecture as **transformer encoder**.



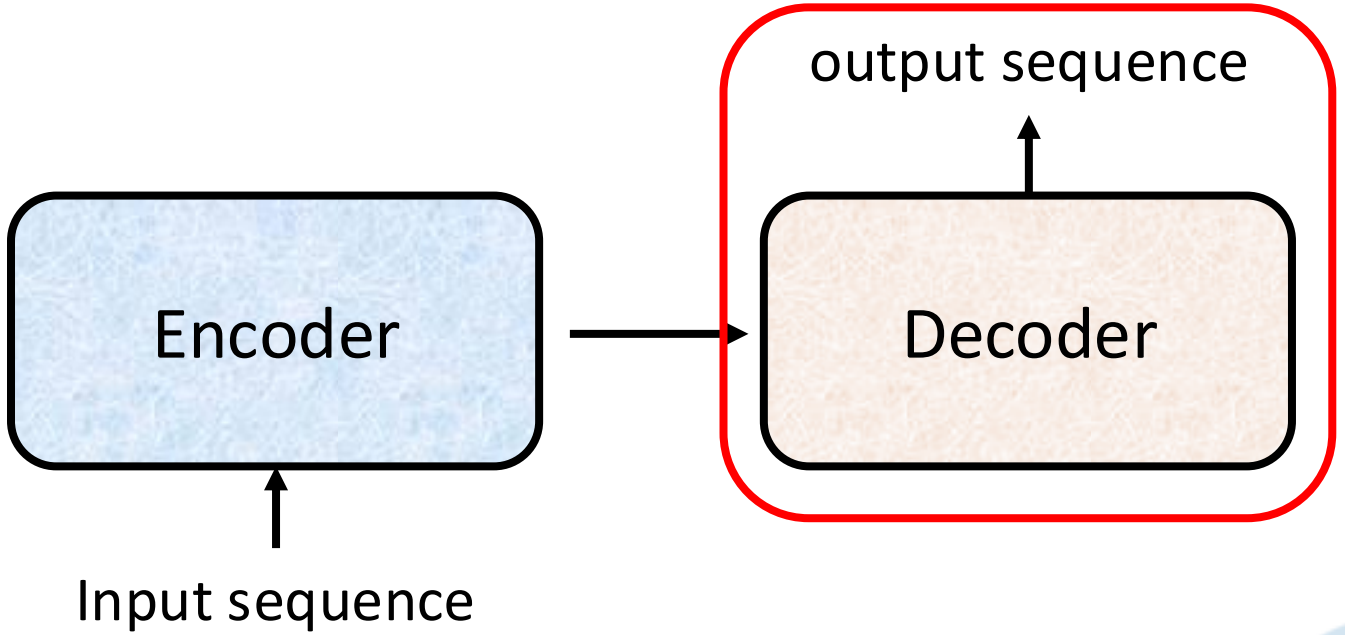
To learn more ……

- On Layer Normalization in the Transformer Architecture
- https://arxiv.org/abs/2002.04745

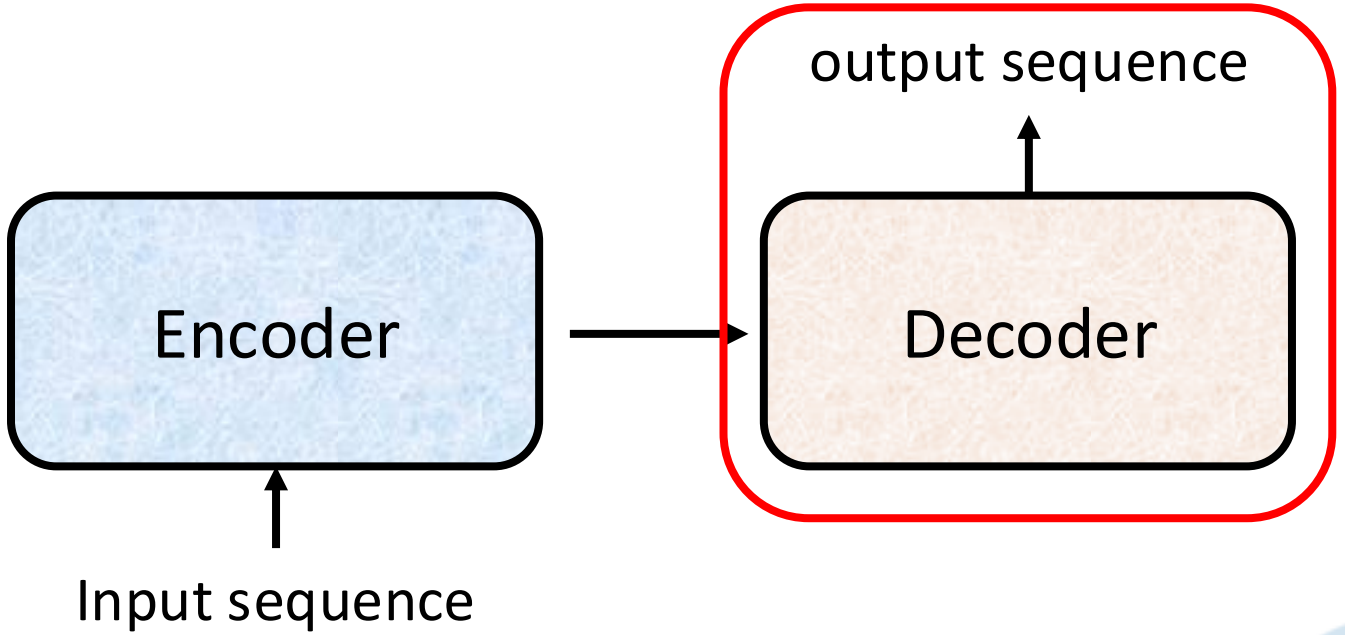
- PowerNorm: Rethinking Batch Normalization in Transformers
- https://arxiv.org/abs/2003.07845



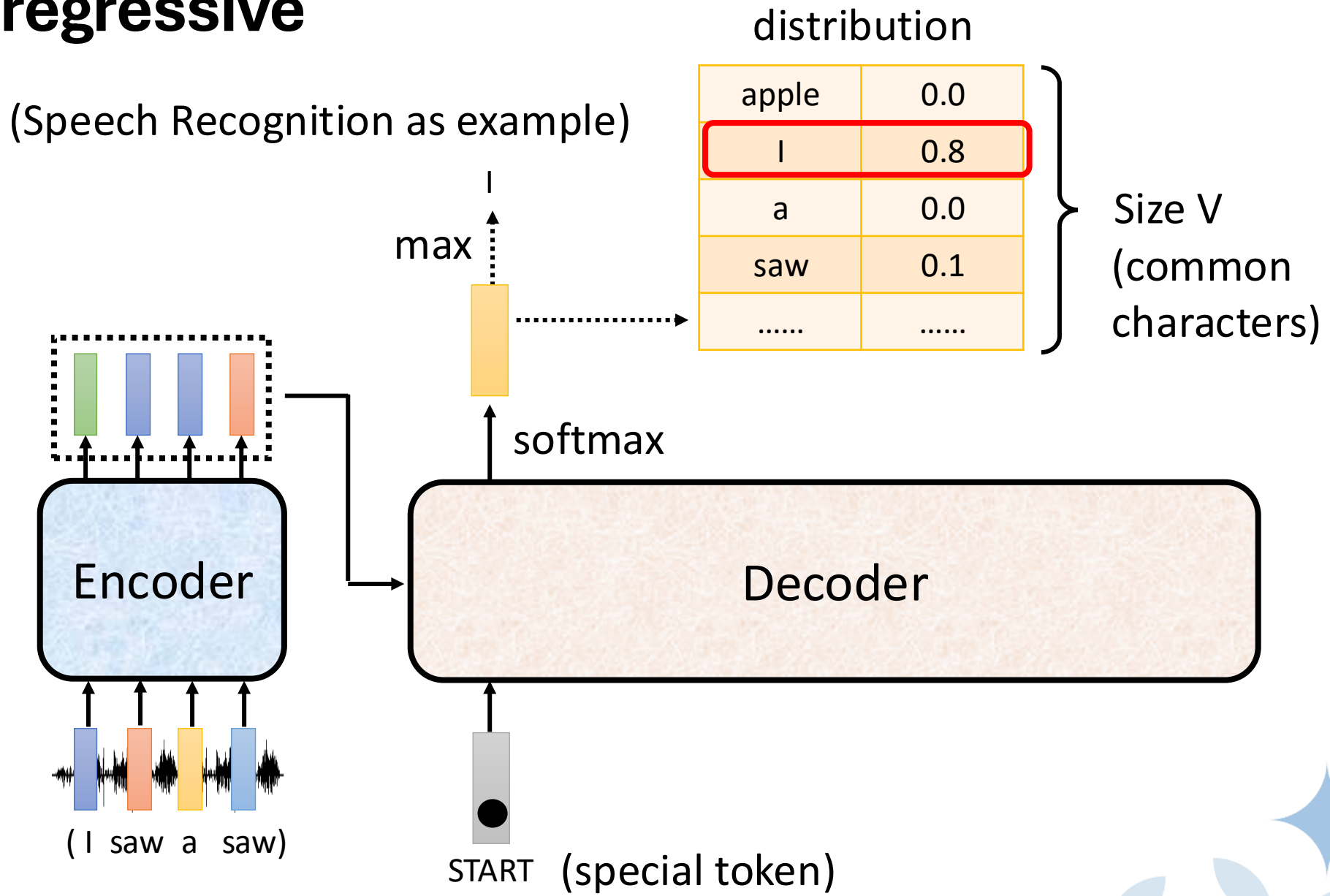
How to generate: Decoder



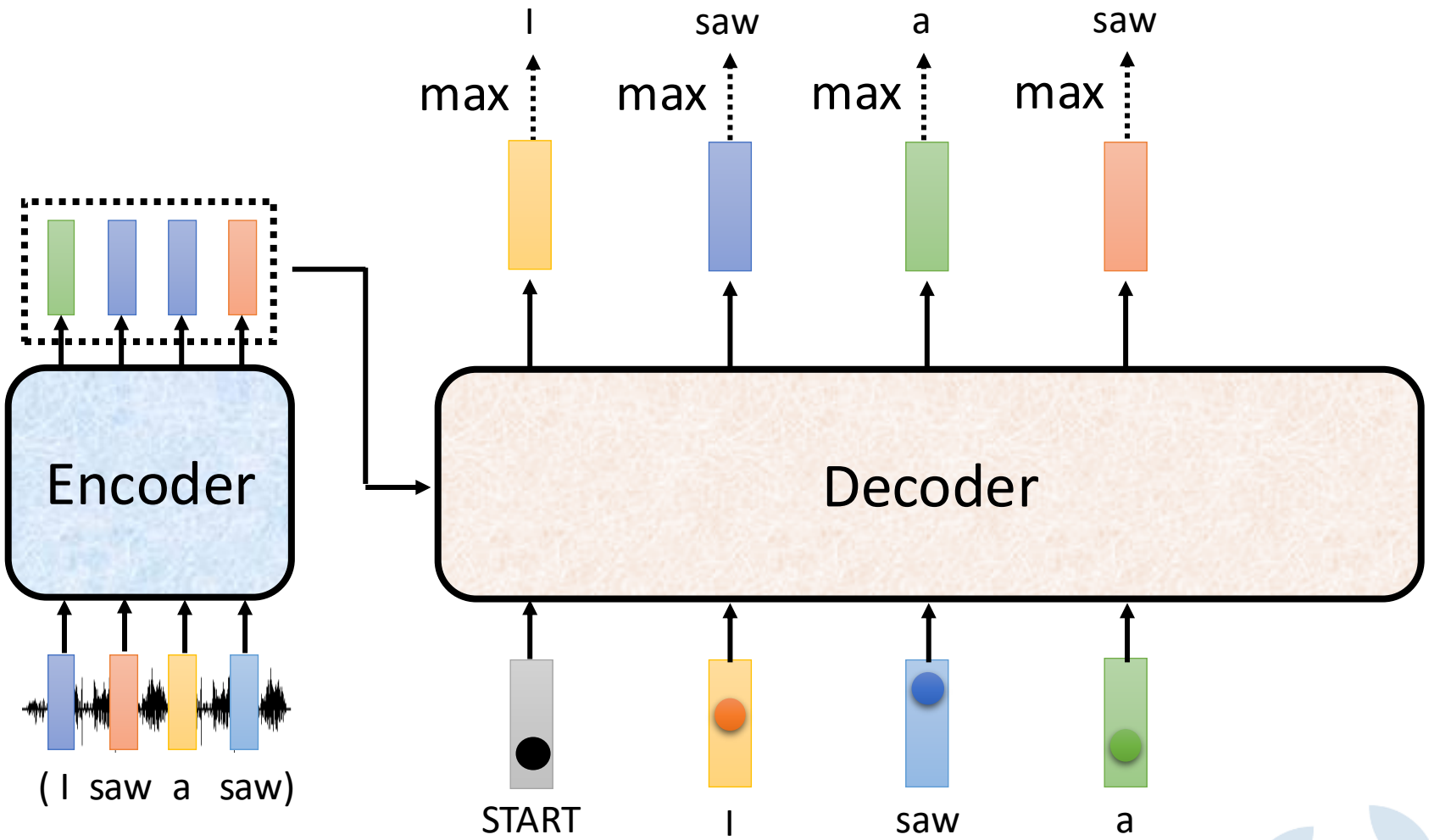
Decoder – Autoregressive (AT)



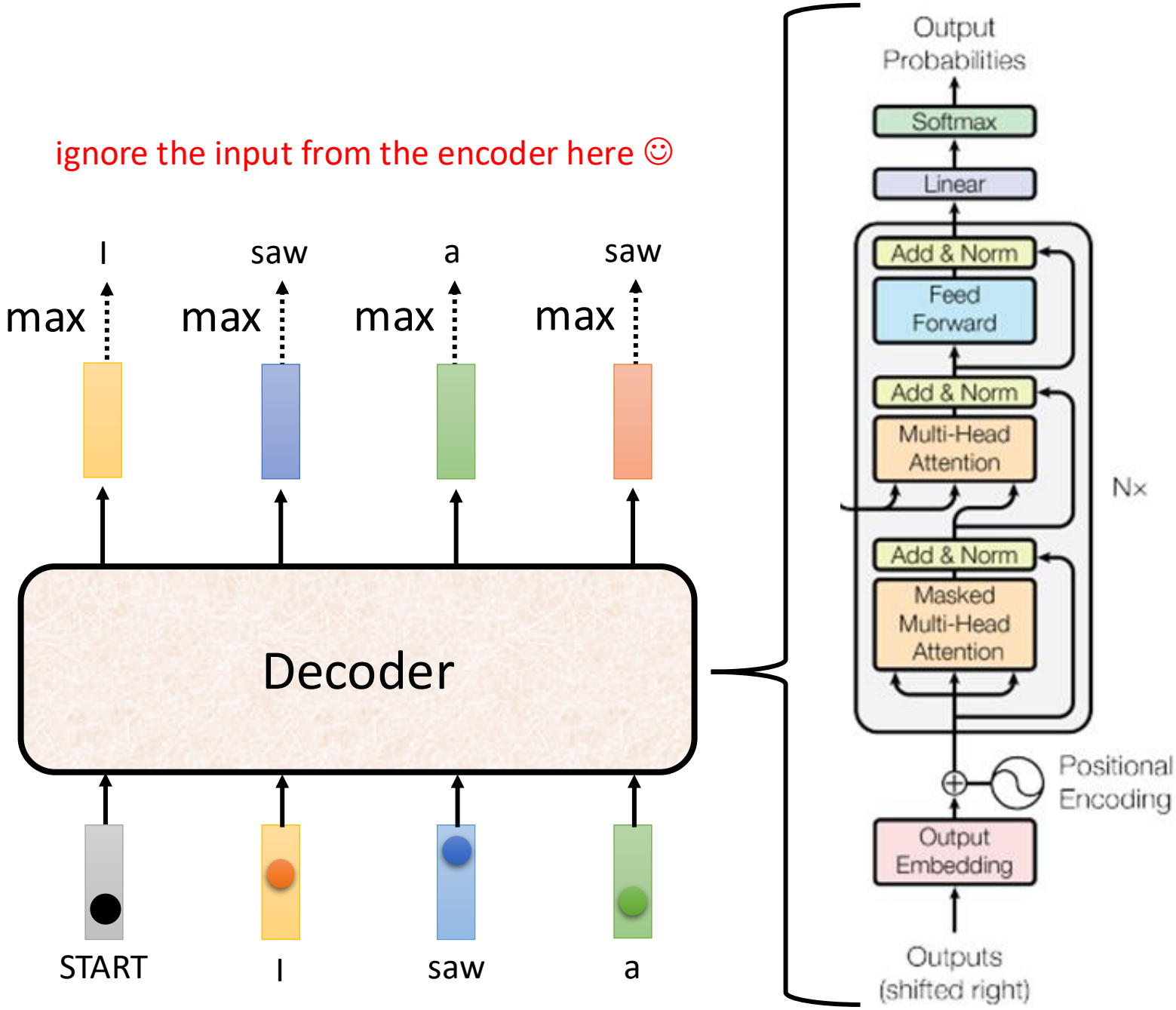
Autoregressive

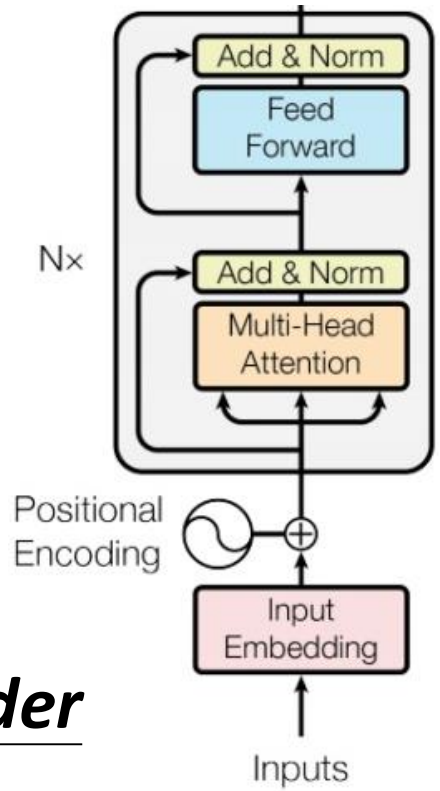


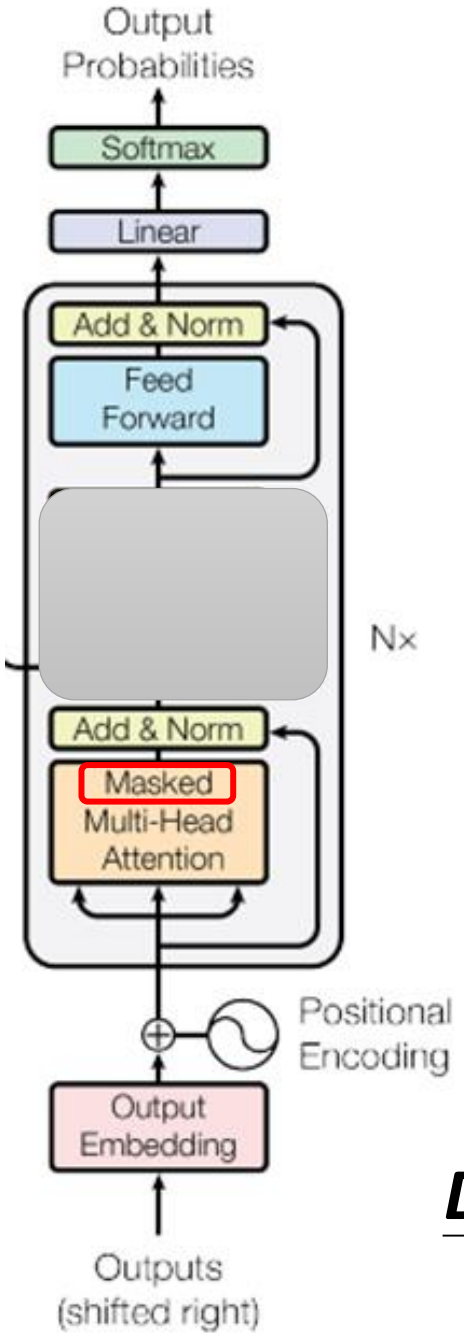
Autoregressive



ignore the input from the encoder here ☺



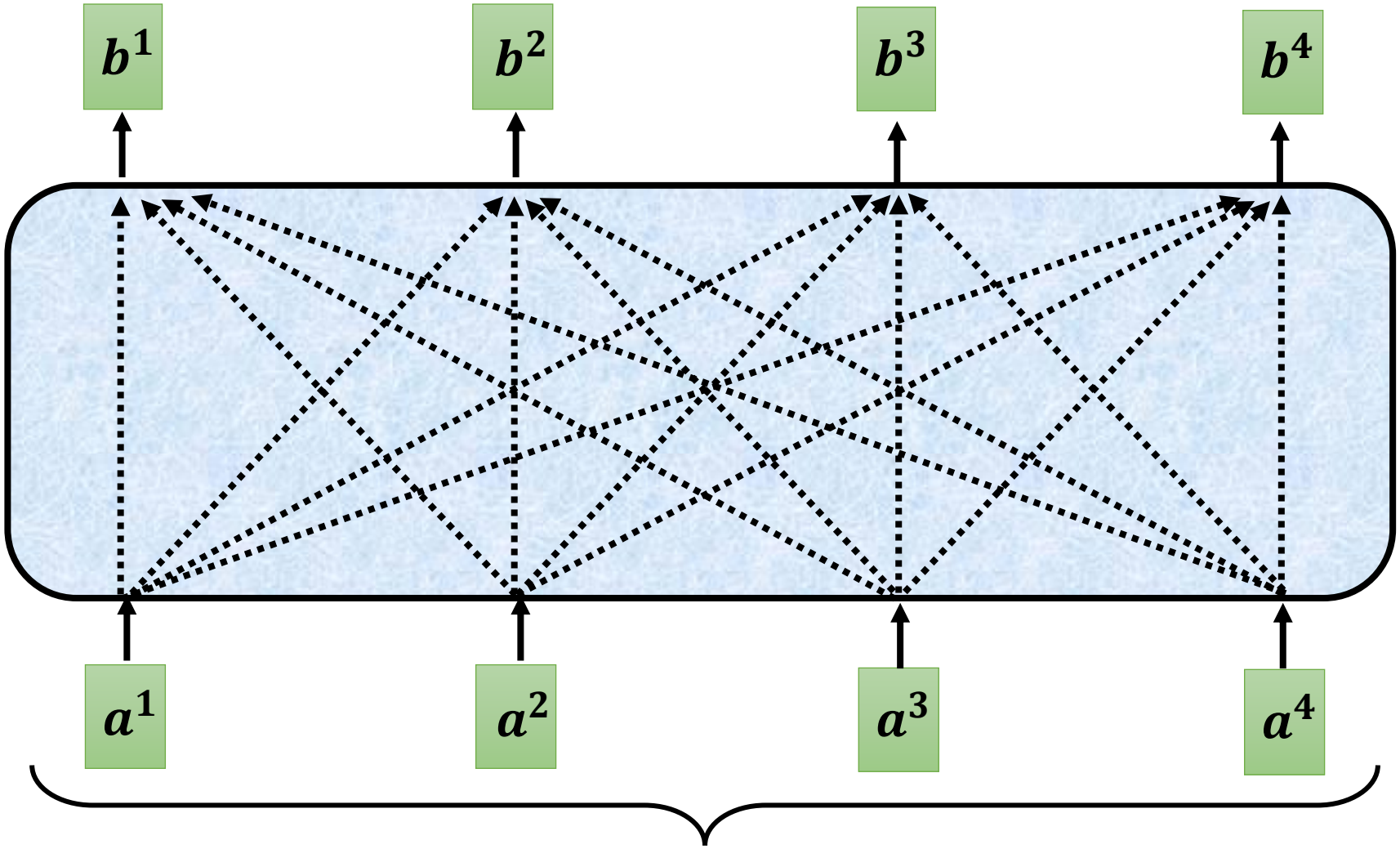




Encoder

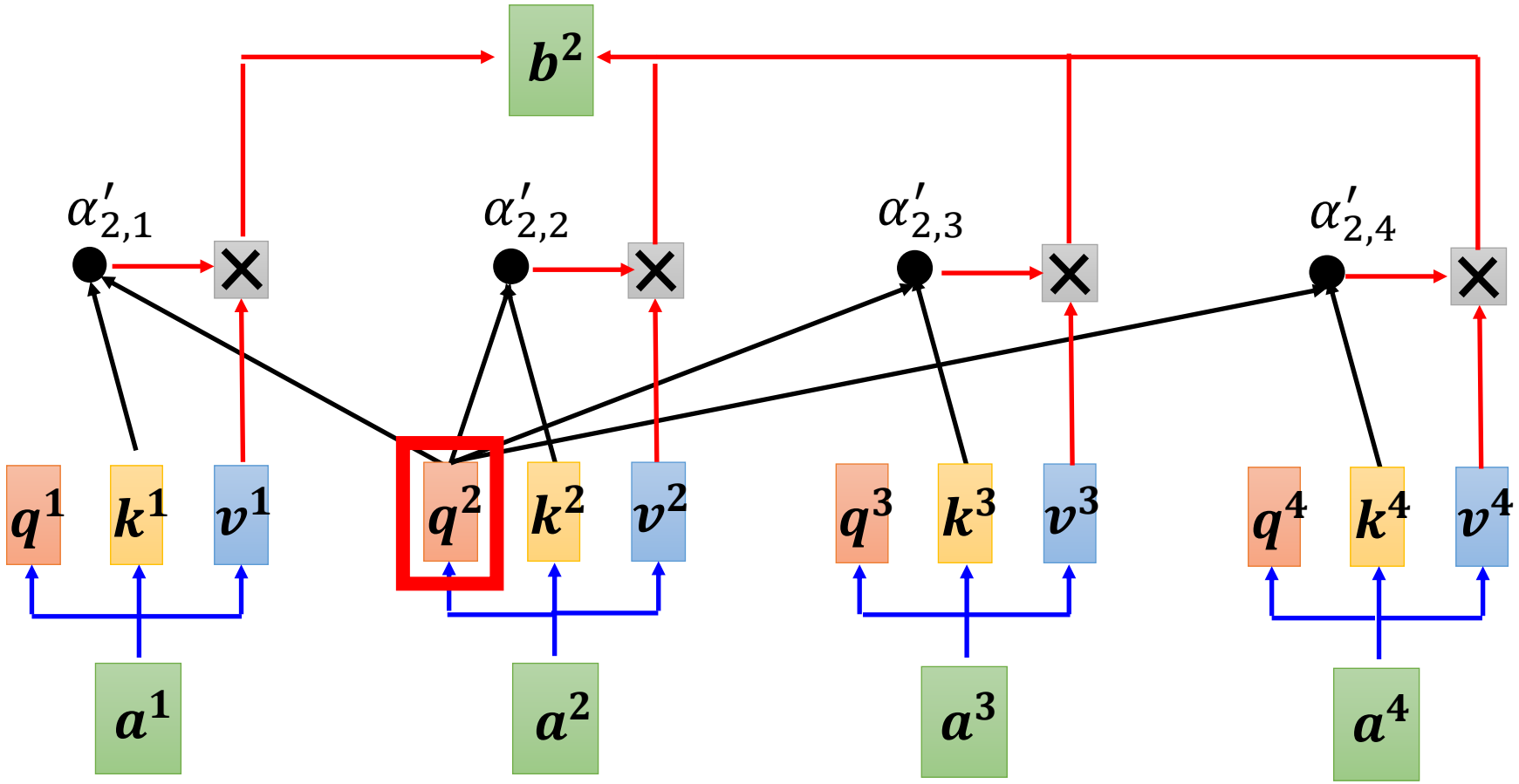
Decoder

Self-attention → Masked Self-attention



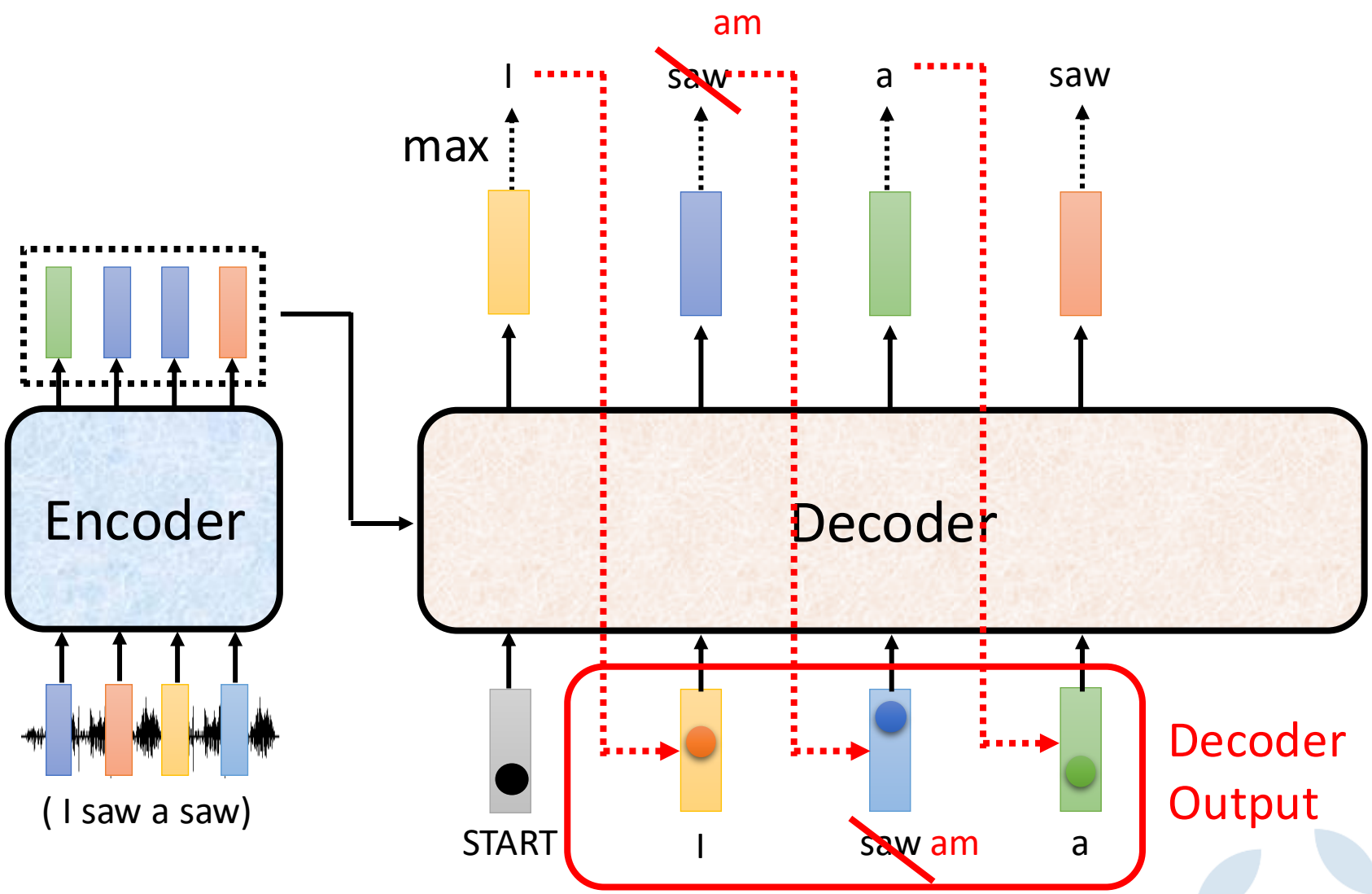
Can be either **input** or **a hidden layer**

Self-attention → Masked Self-attention



Why masked? Consider how does decoder work

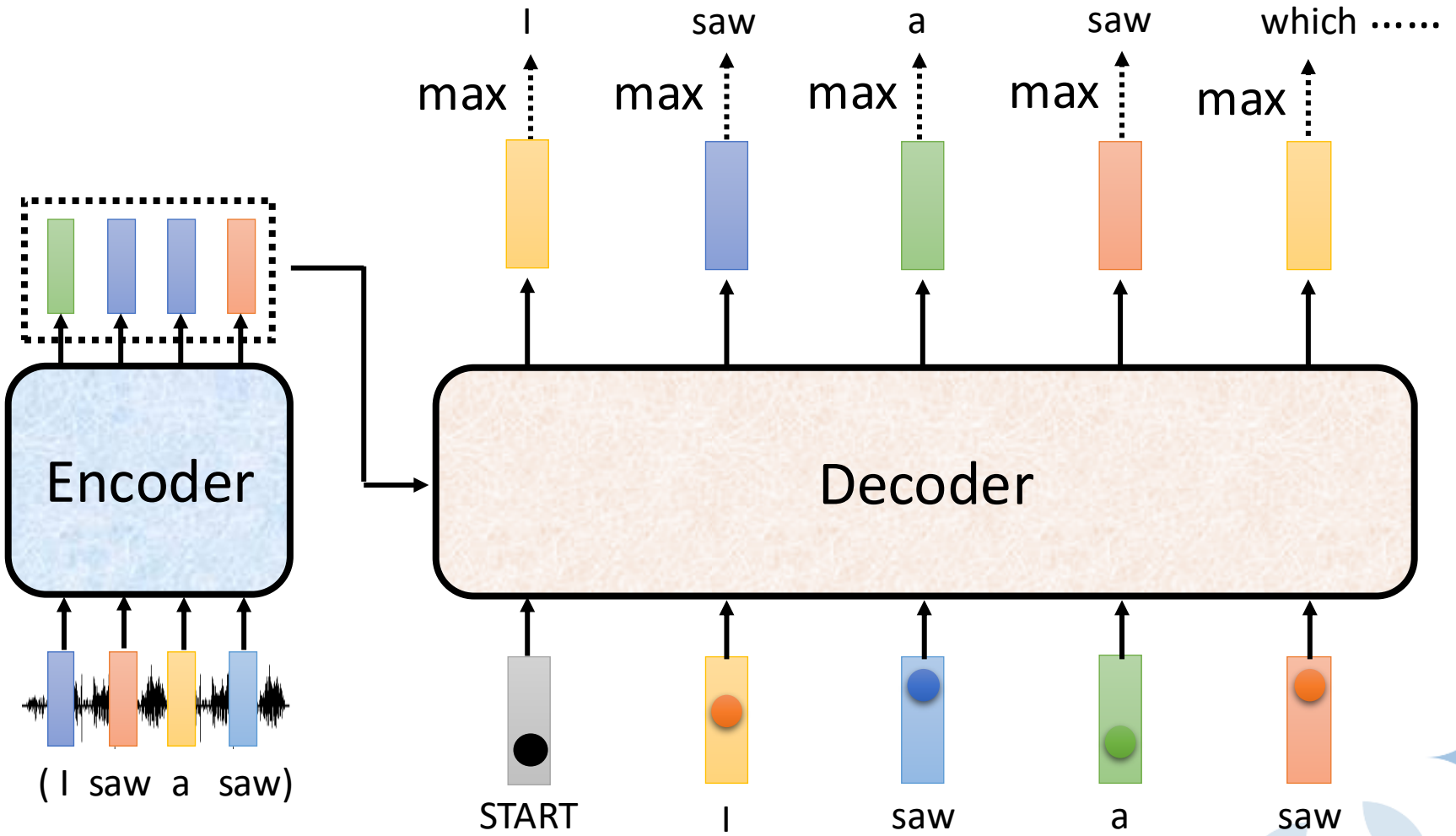
Autoregressive



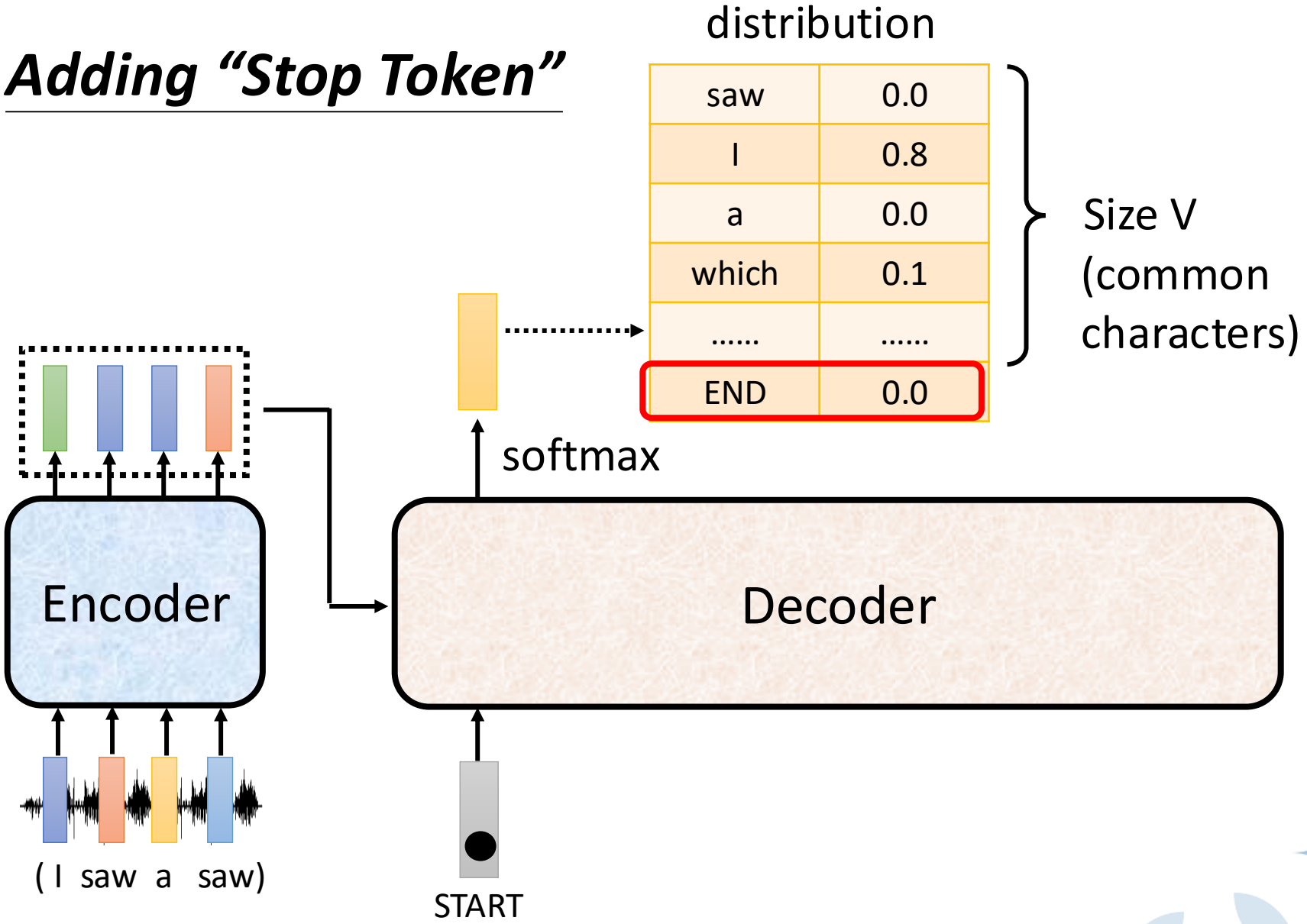
Autoregressive

We do not know the correct output length.

Never stop!

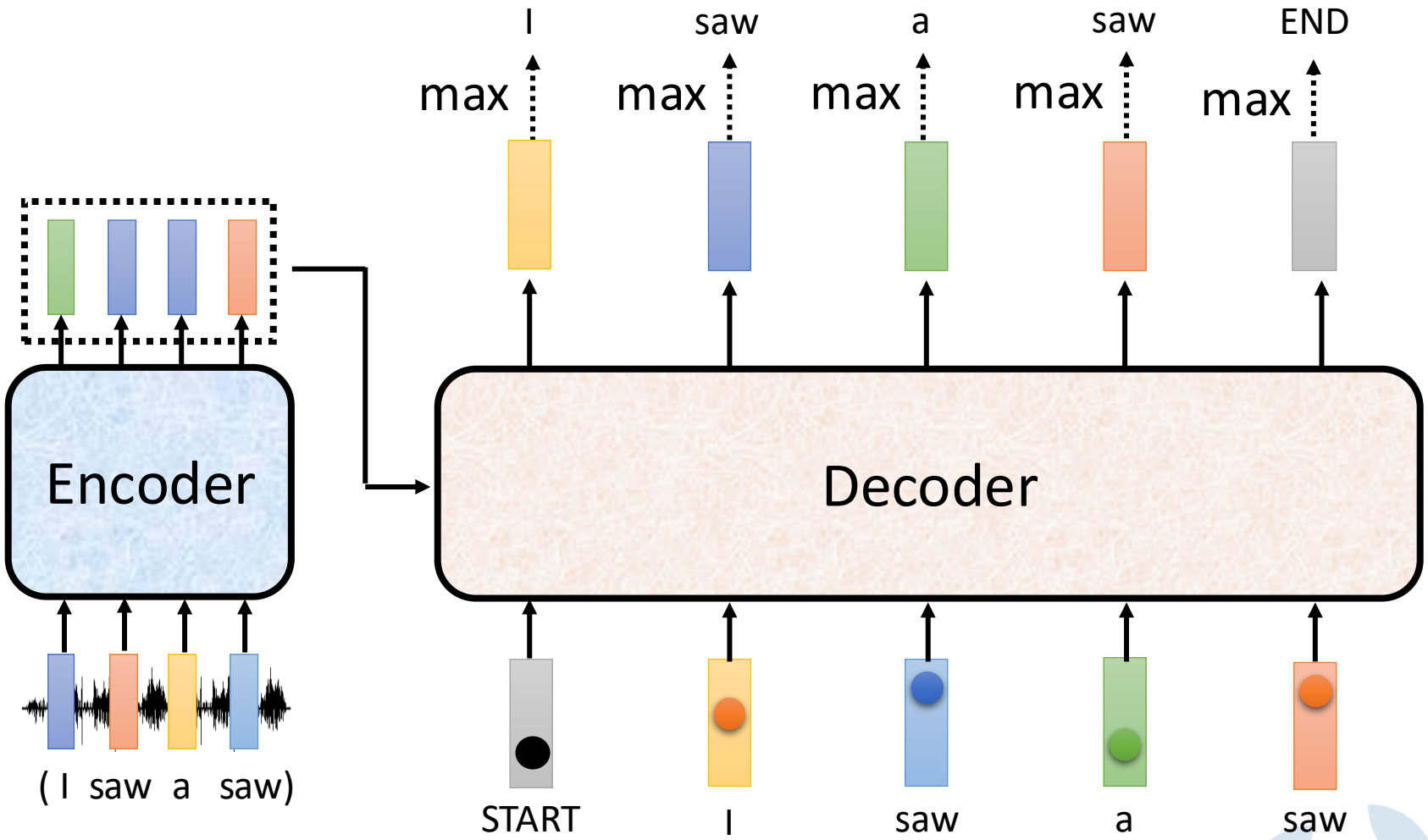


Adding "Stop Token"

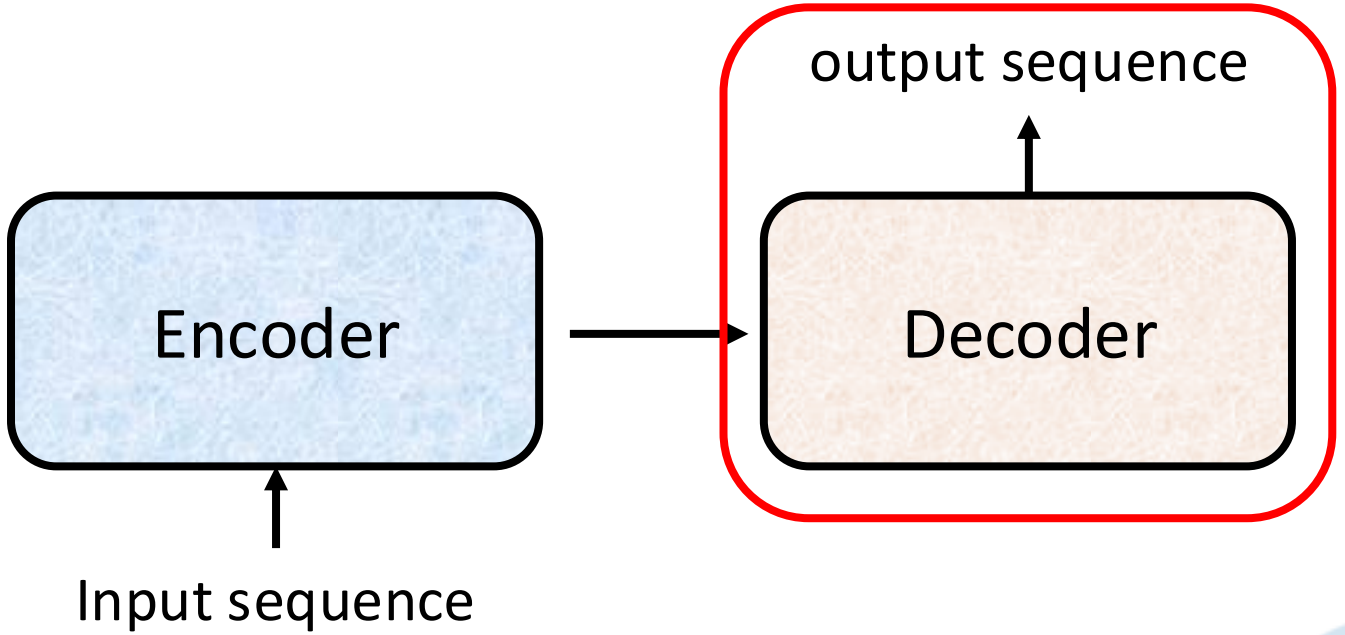


Autoregressive

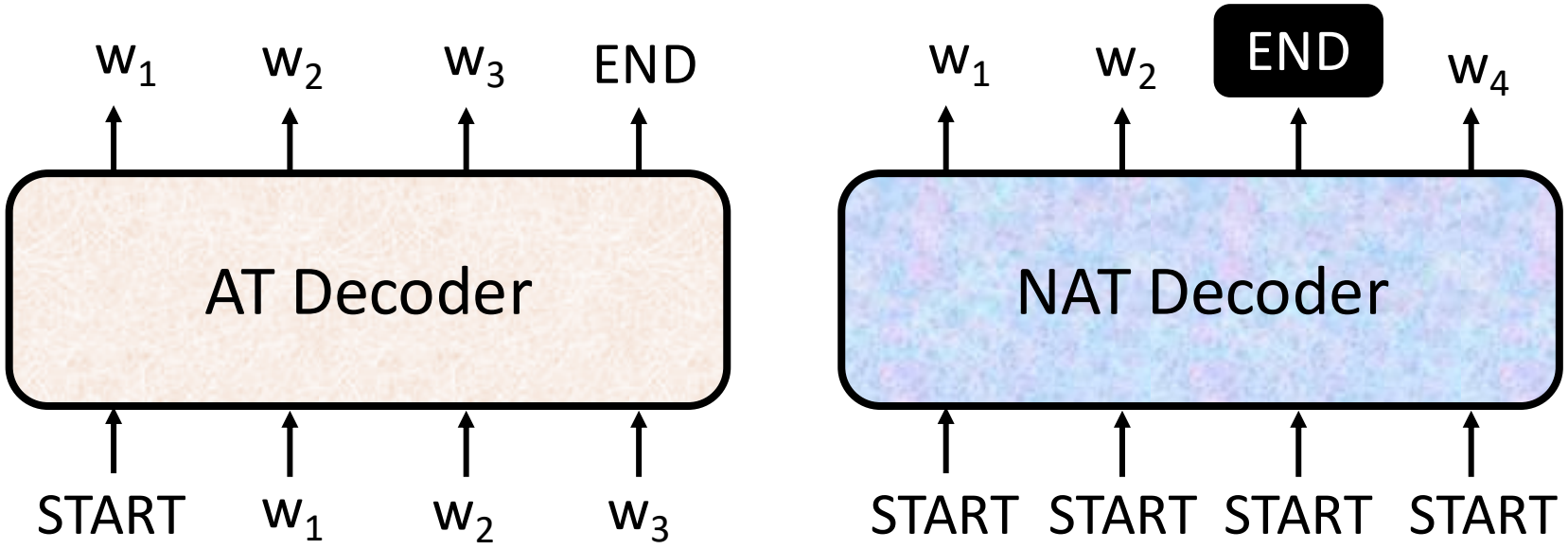
Stop at here!



Decoder
– Non-autoregressive (NAT)

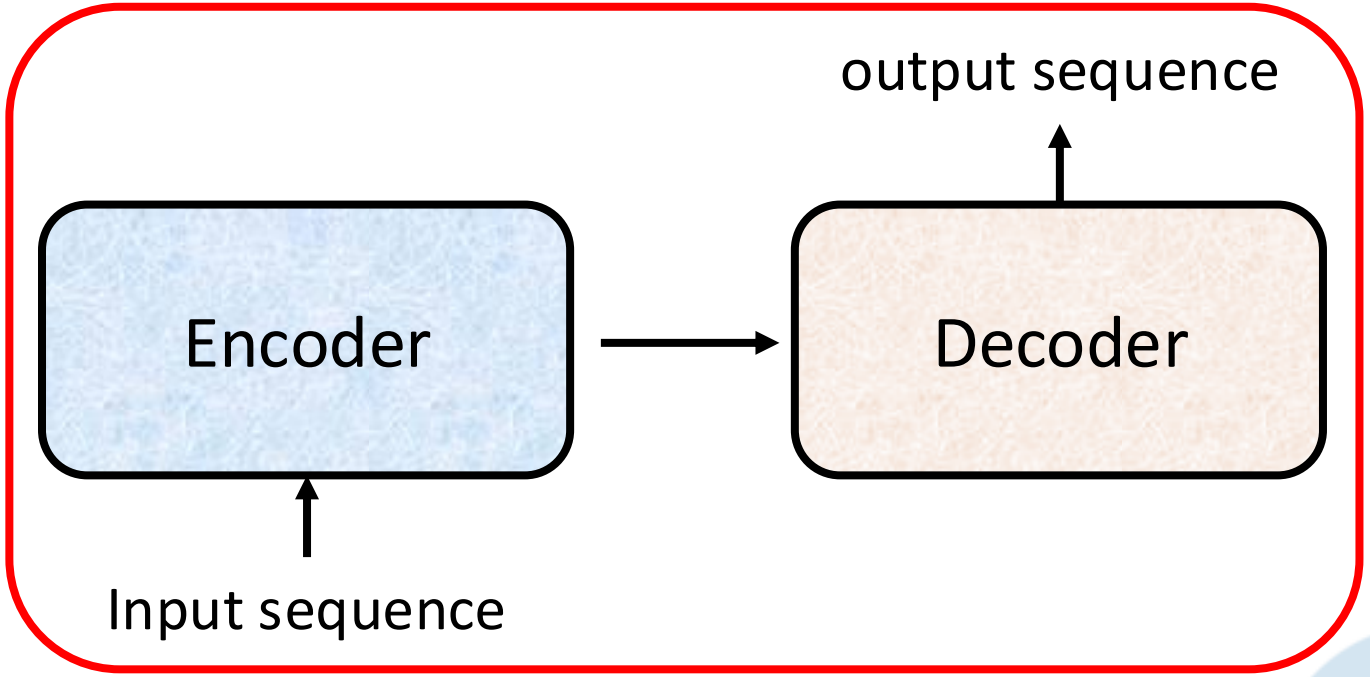


AT v.s. NAT

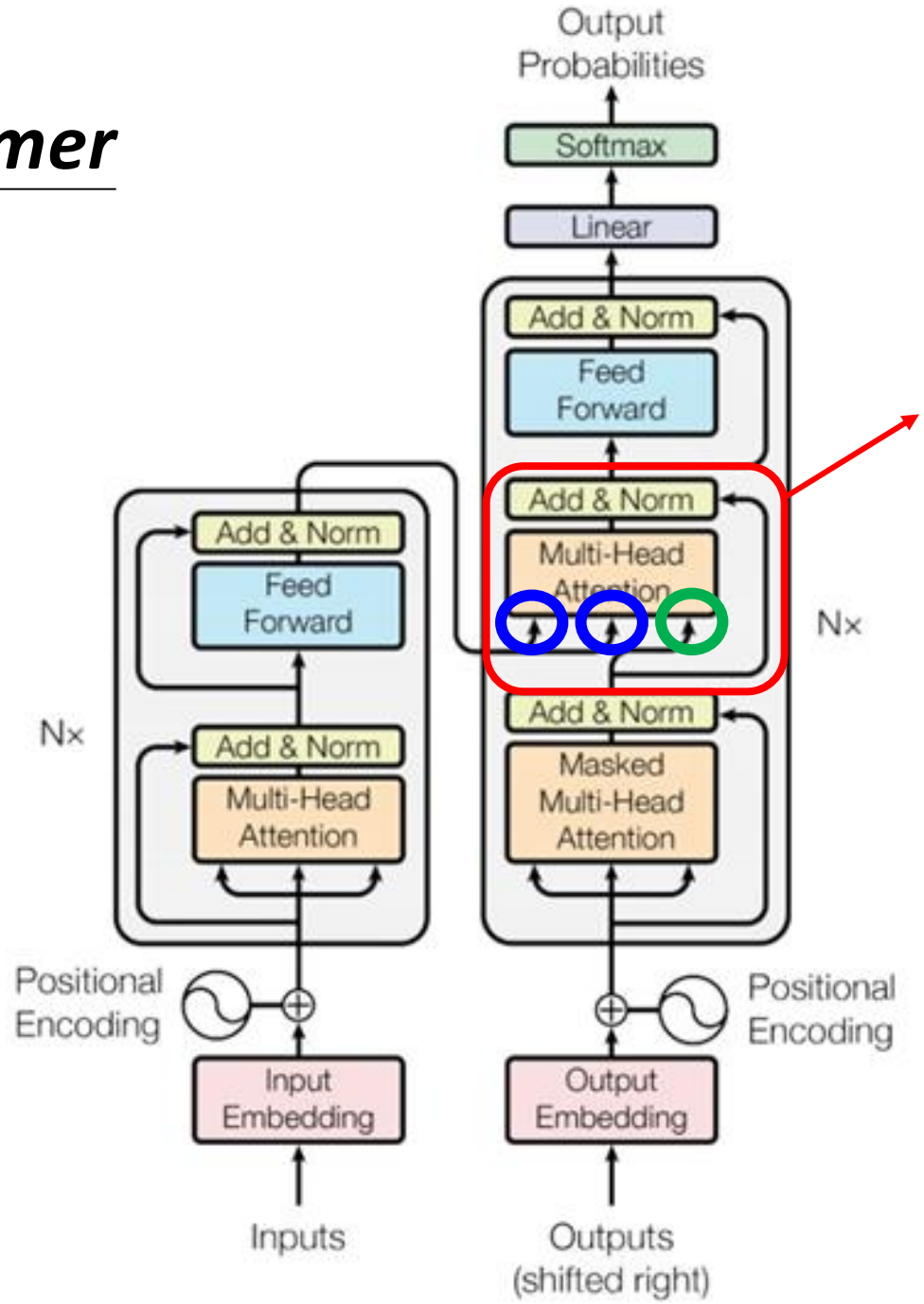


- How to decide the output length for NAT decoder?
  - Another predictor for output length
  - Output a very long sequence, ignore tokens after END
- Advantage: parallel, more stable generation (e.g., TTS)
- NAT is usually worse than AT (why? **Multi-modality**)

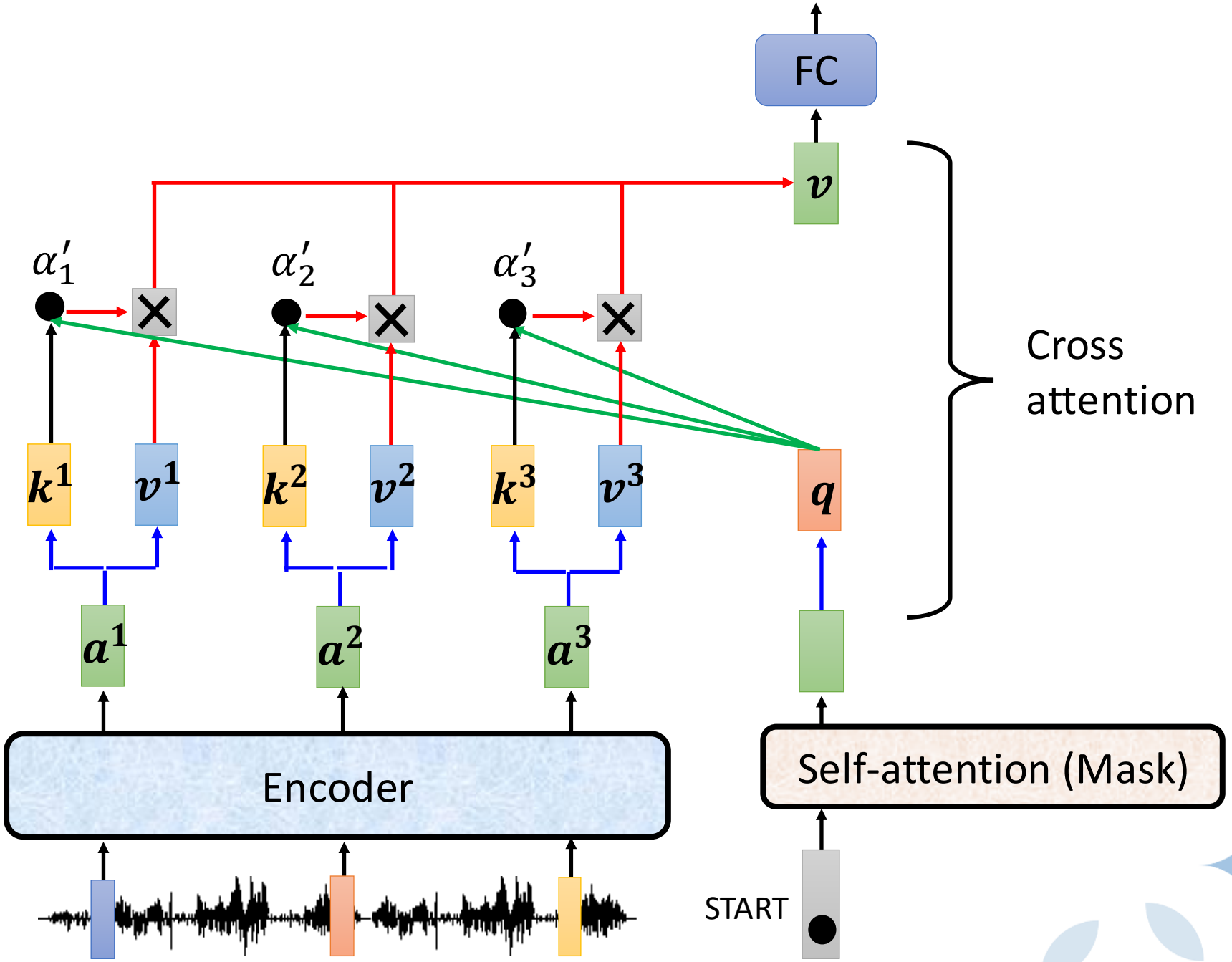
Encoder-Decoder

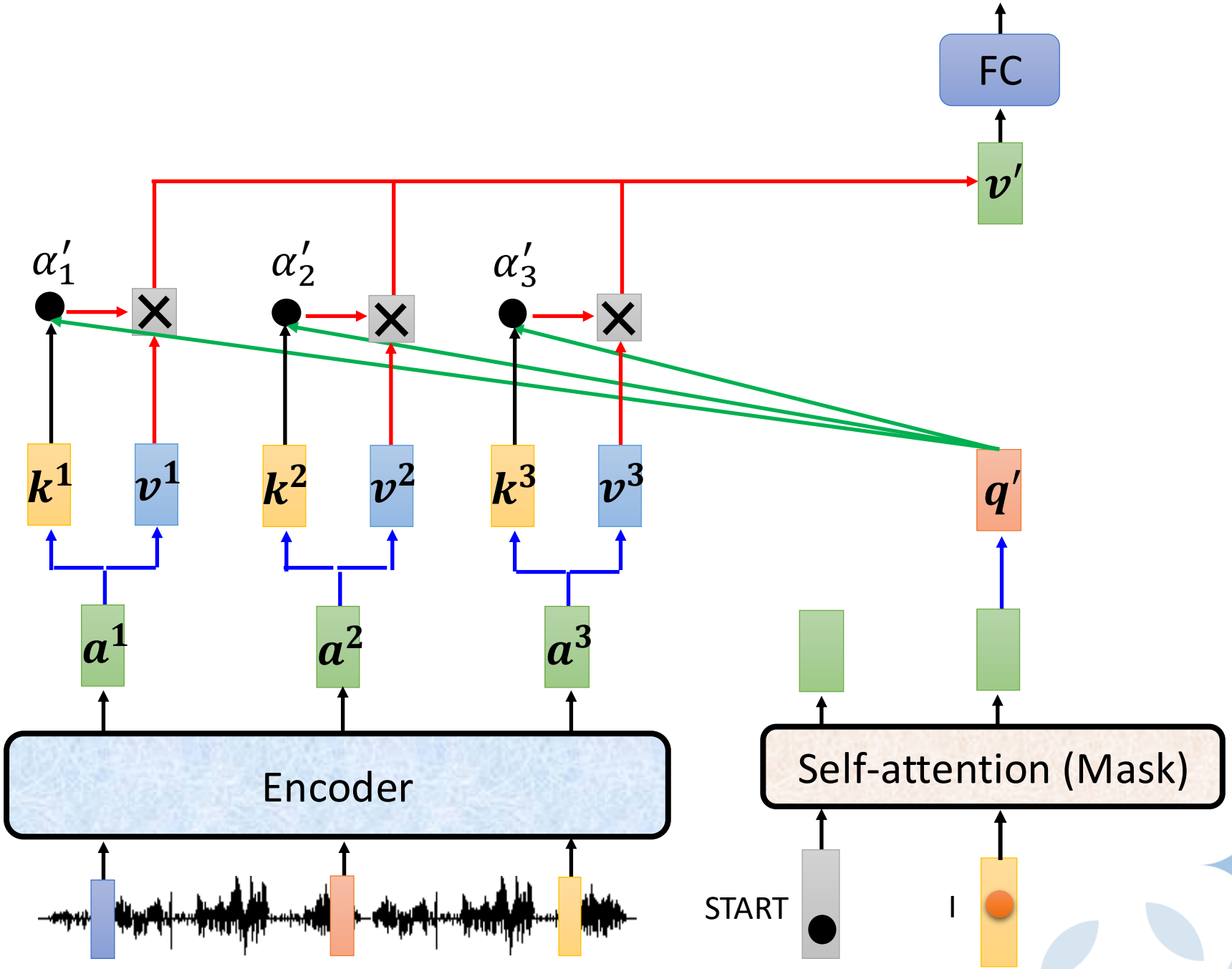


Transformer



Cross attention

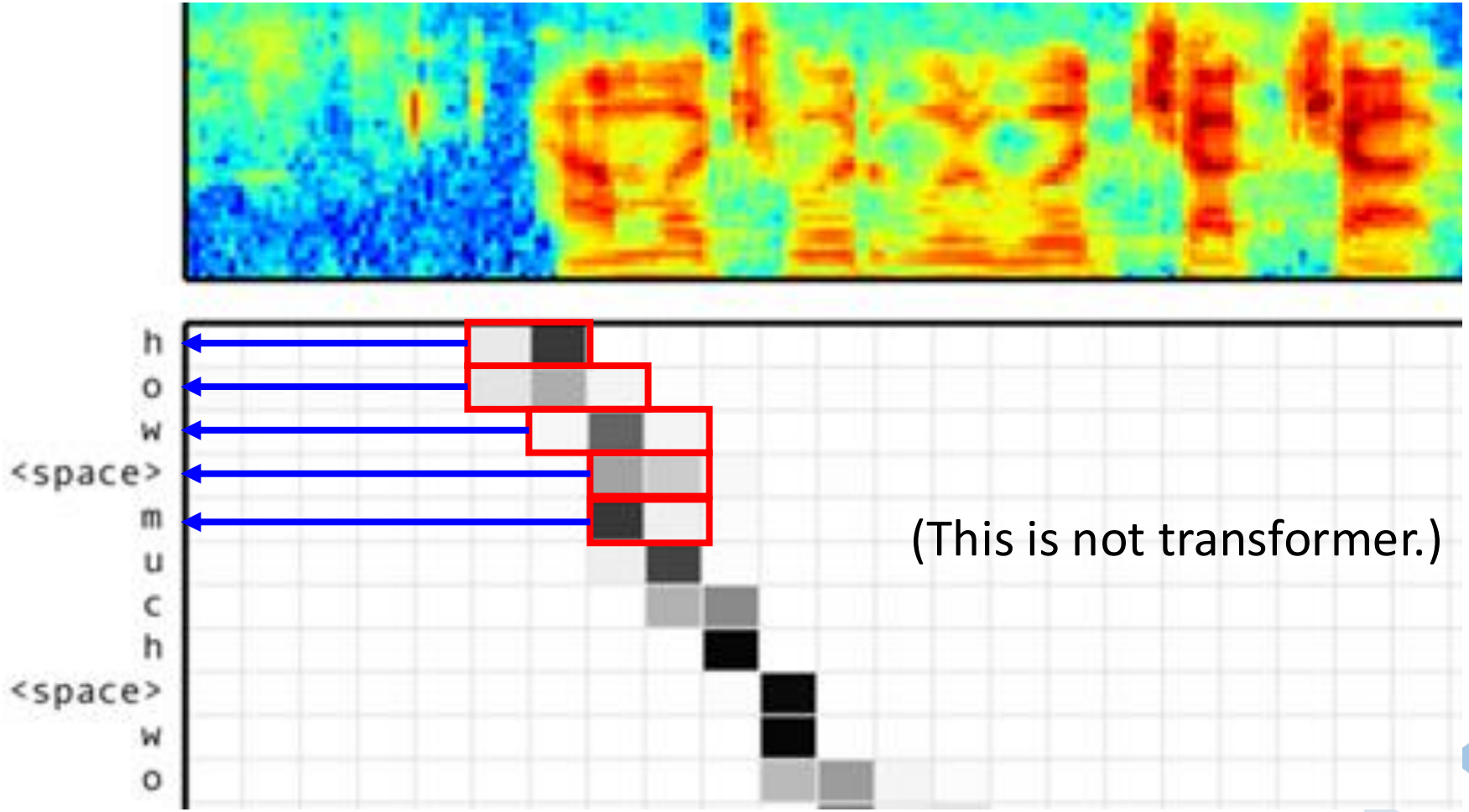

FC
v
α'_1
α'_2
α'_3
k^1
v^1
k^2
v^2
k^3
v^3
q
Cross attention
a^1
a^2
a^3
Encoder
Self-attention (Mask)
START


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
I

Cross Attention

Listen, attend and spell: A neural network for large vocabulary conversational speech recognition

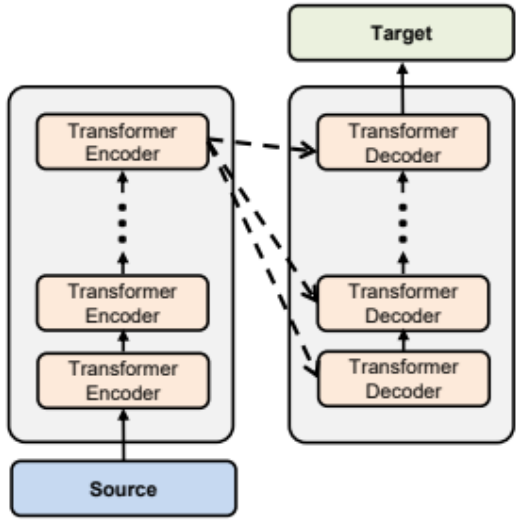
https://ieeexplore.ieee.org/document/7472621



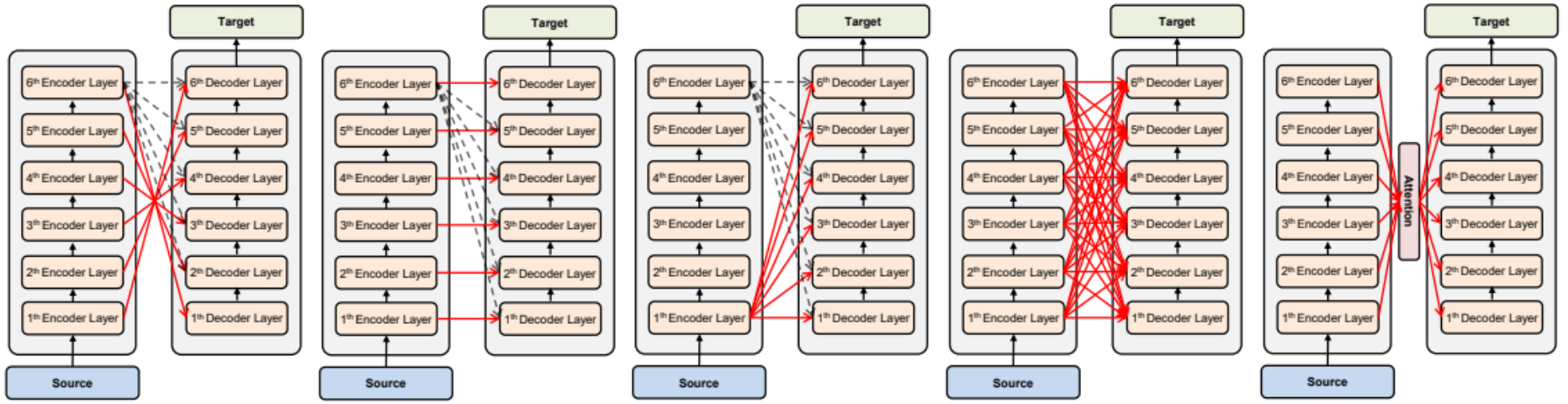
(This is not transformer.)

Cross Attention

Source of image:
https://arxiv.org/abs/2005.08081

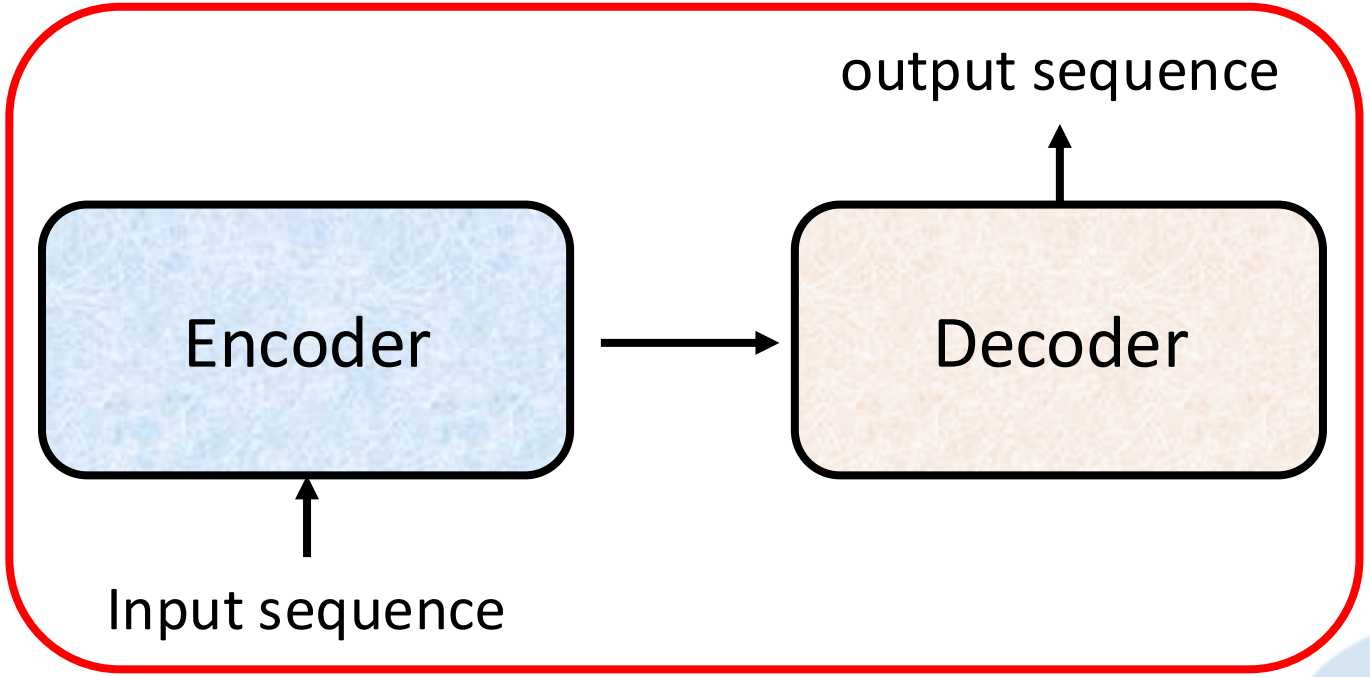


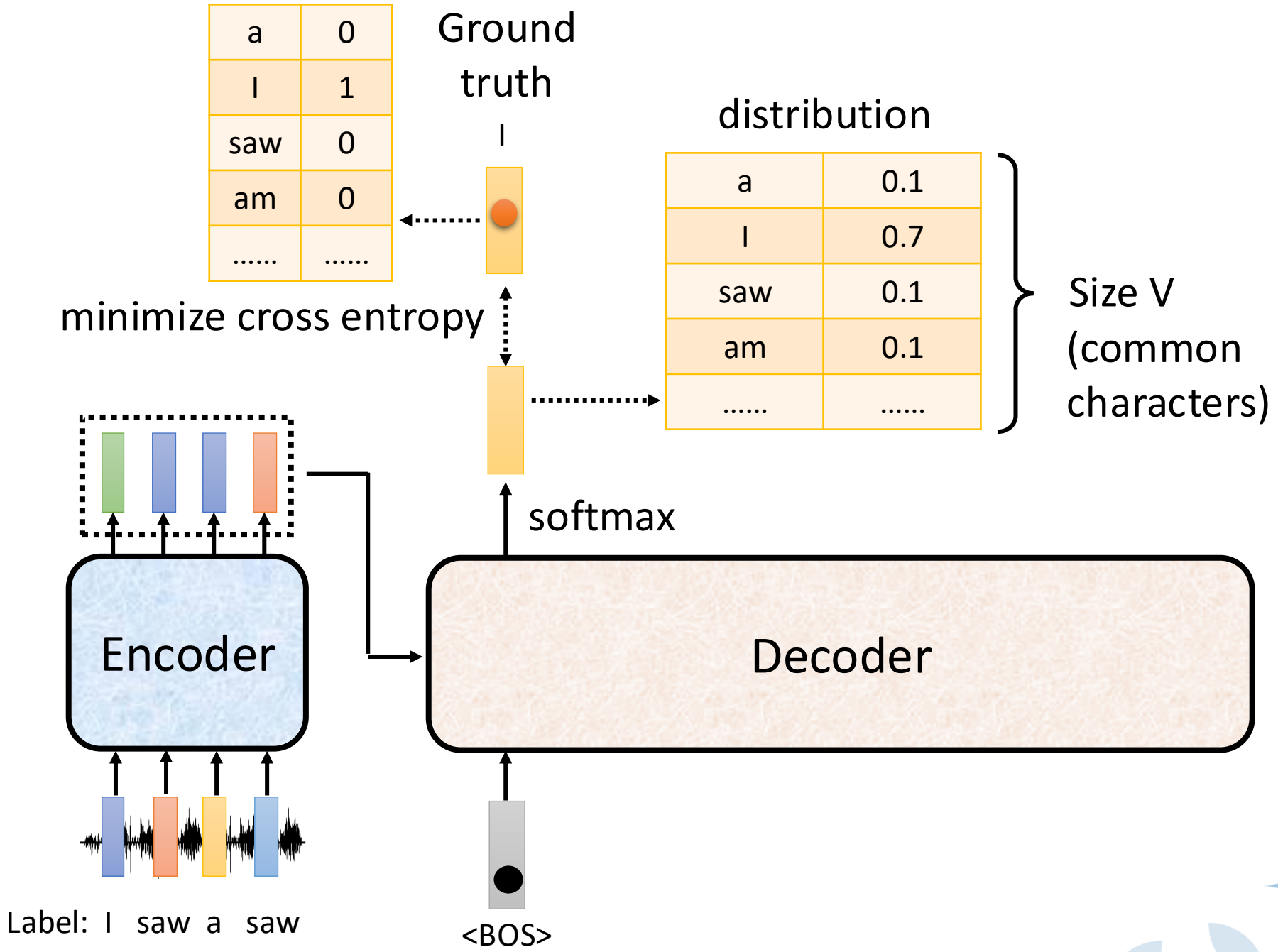
(a) Conventional Transformer



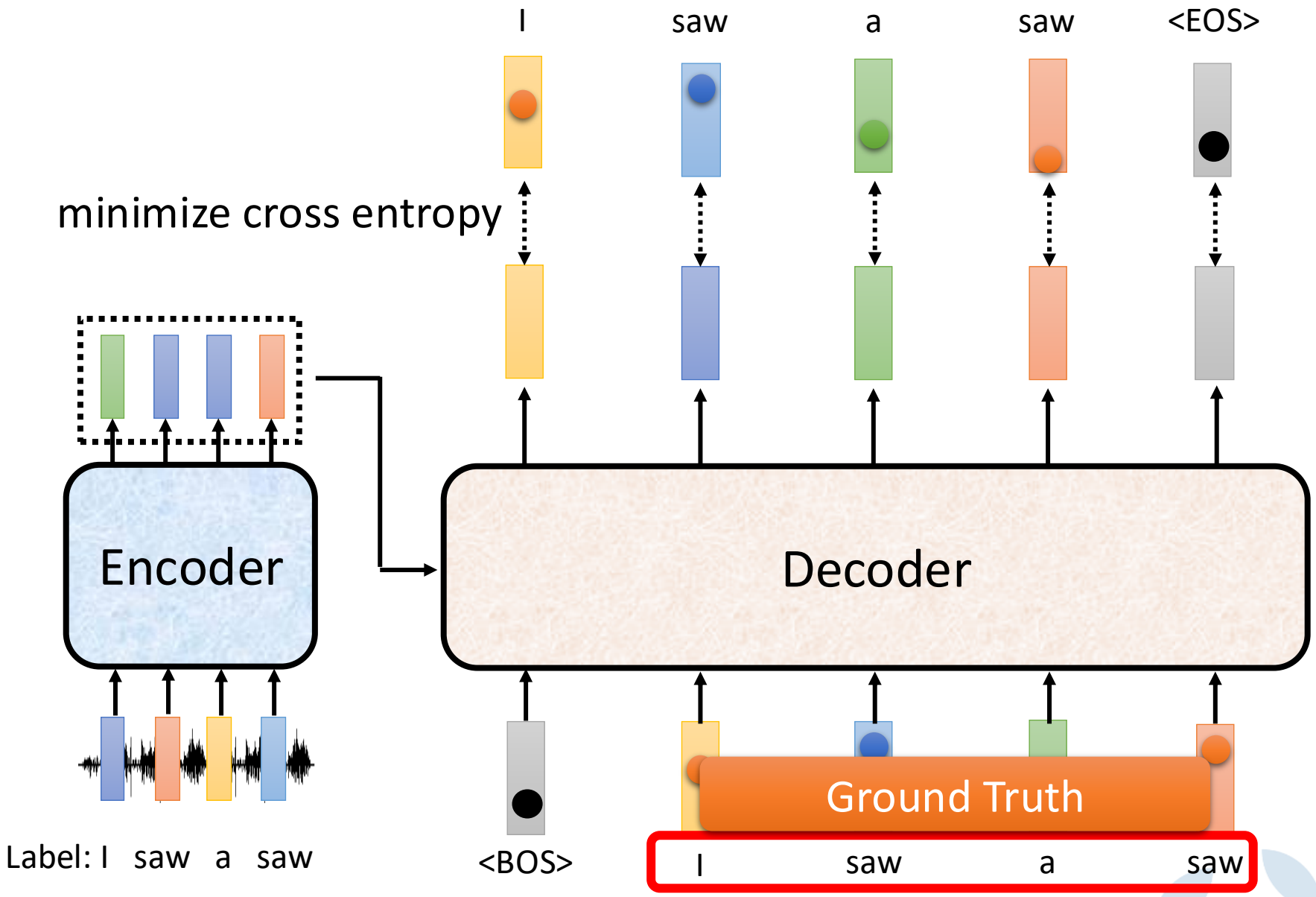
(a) Granularity Consistent Attention (b) Granularity Parallel Attention (c) Fine-Grained Attention (d) Full Matching Attention (e) Adaptive Matching Attention

Training

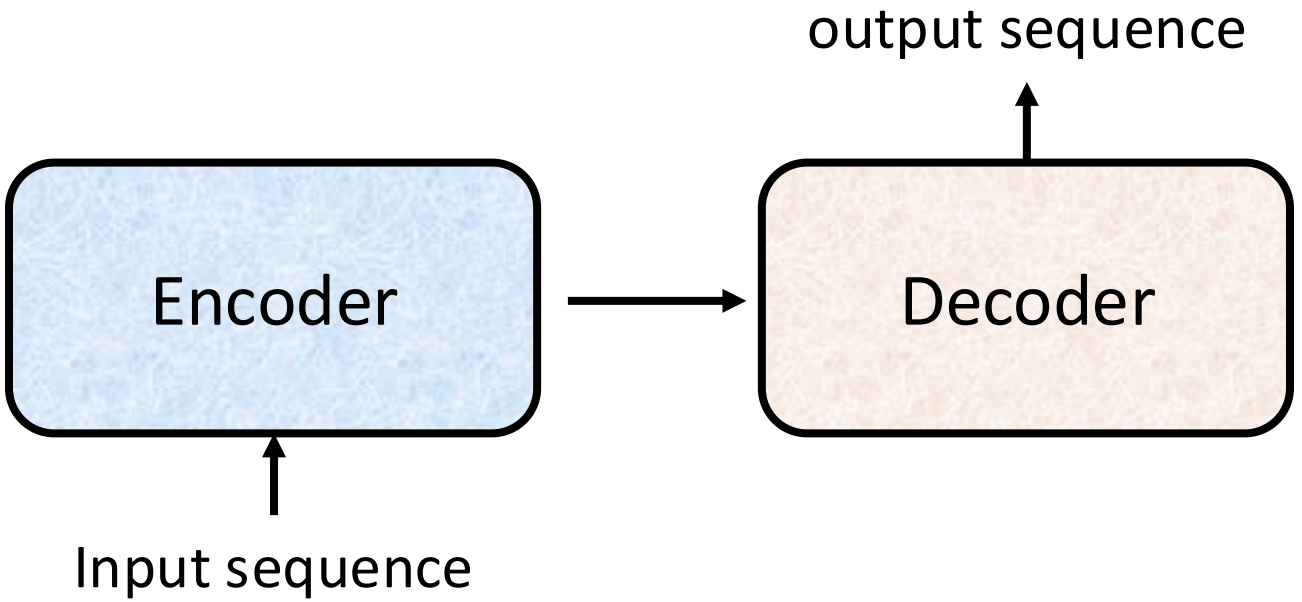



a 0
I 1
saw 0
am 0
…… ……
Ground truth
I
distribution
a 0.1
I 0.7
saw 0.1
am 0.1
…… ……
Size V (common characters)
minimize cross entropy
softmax
Encoder
Decoder
Label: I saw a saw
<BOS>

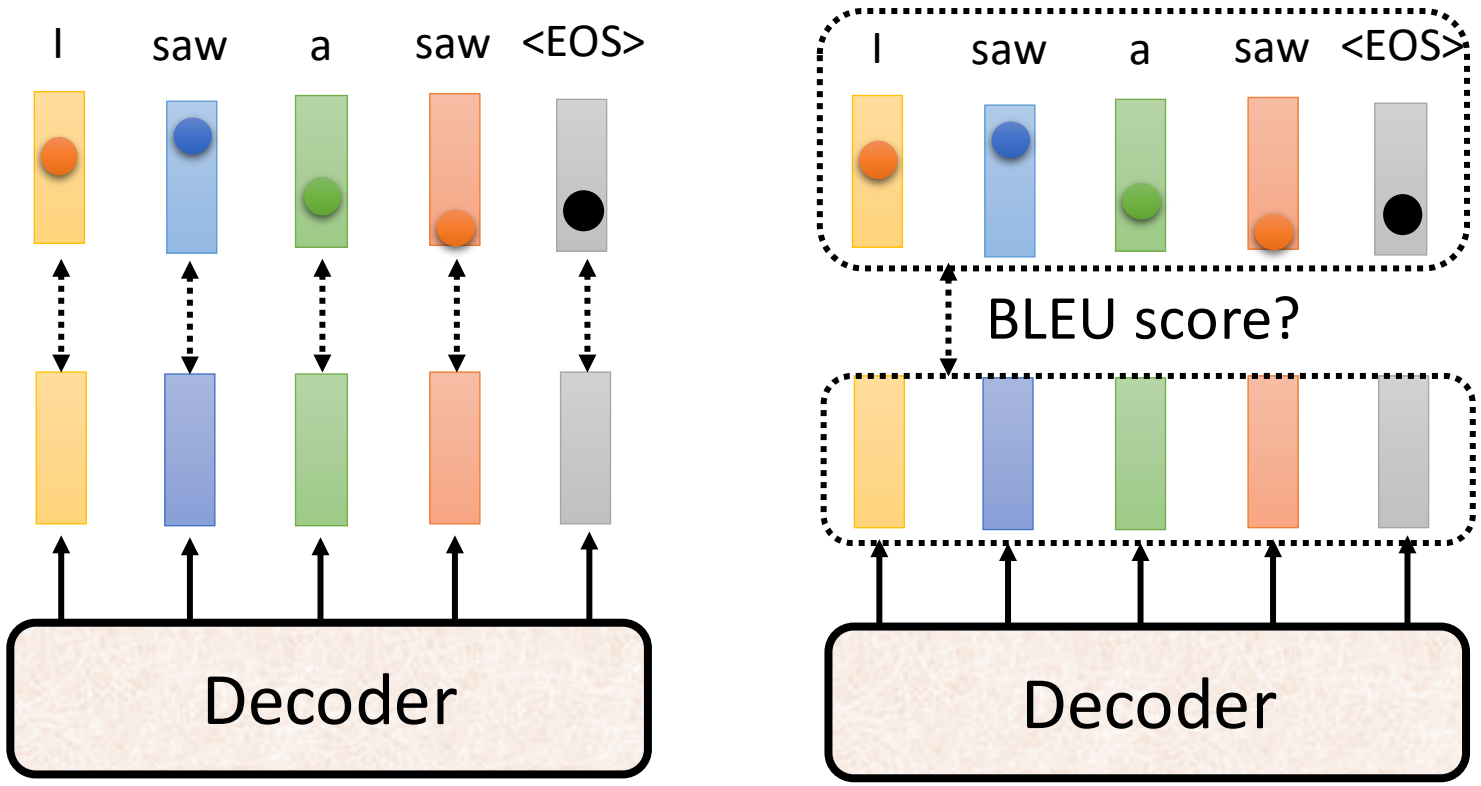
Teacher Forcing: using the ground truth as input.



Tips



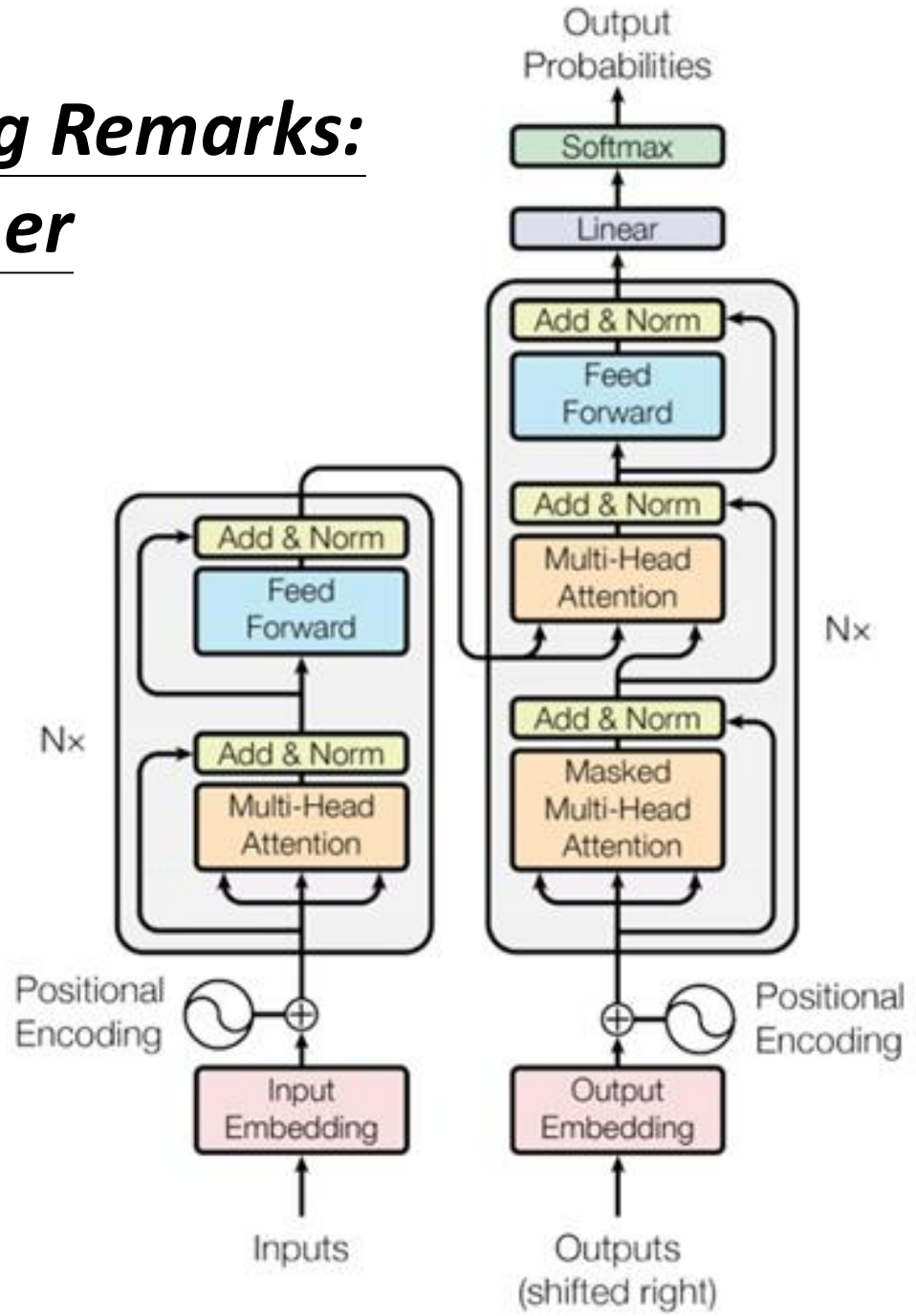
Optimizing Evaluation Metrics?



How to do the optimization?

When you don't know how to optimize, just use reinforcement learning (RL)! https://arxiv.org/abs/1511.06732

Concluding Remarks: Transformer



Questions?