MAE 688: Machine Learning for Mechanical Engineers

Lecture 6-Transfomer

Dr. Bing Dong

Director, Built Environment Science and Technology (BEST) Lab
Professor
Mechanical and Aerospace Engineering

Syracuse University

Transformer

2017

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

Transformer

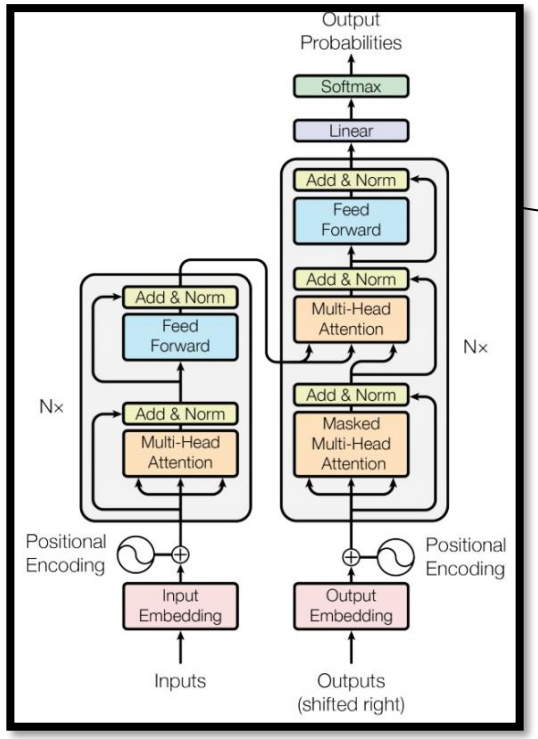
BERT 340M

GPT-2 (1.5B)

GPT-3 (175B)

GPT-4 (2T)

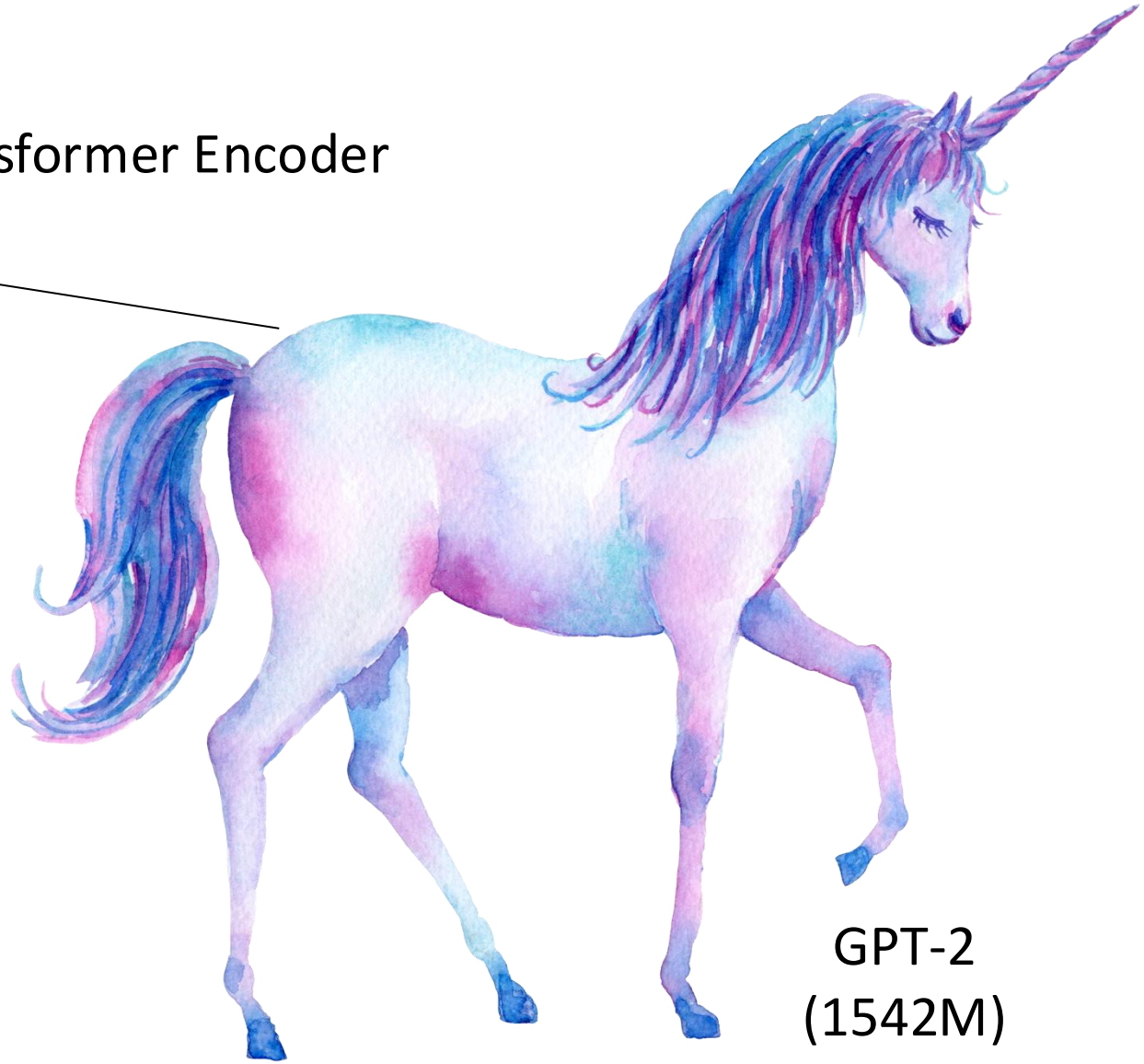
Generative Pre-Training (GPT)



Transformer Encoder

BERT (340M)

ELMO (94M)



GPT-2 (1542M)

Source of image: https://huaban.com/pins/1714071707/

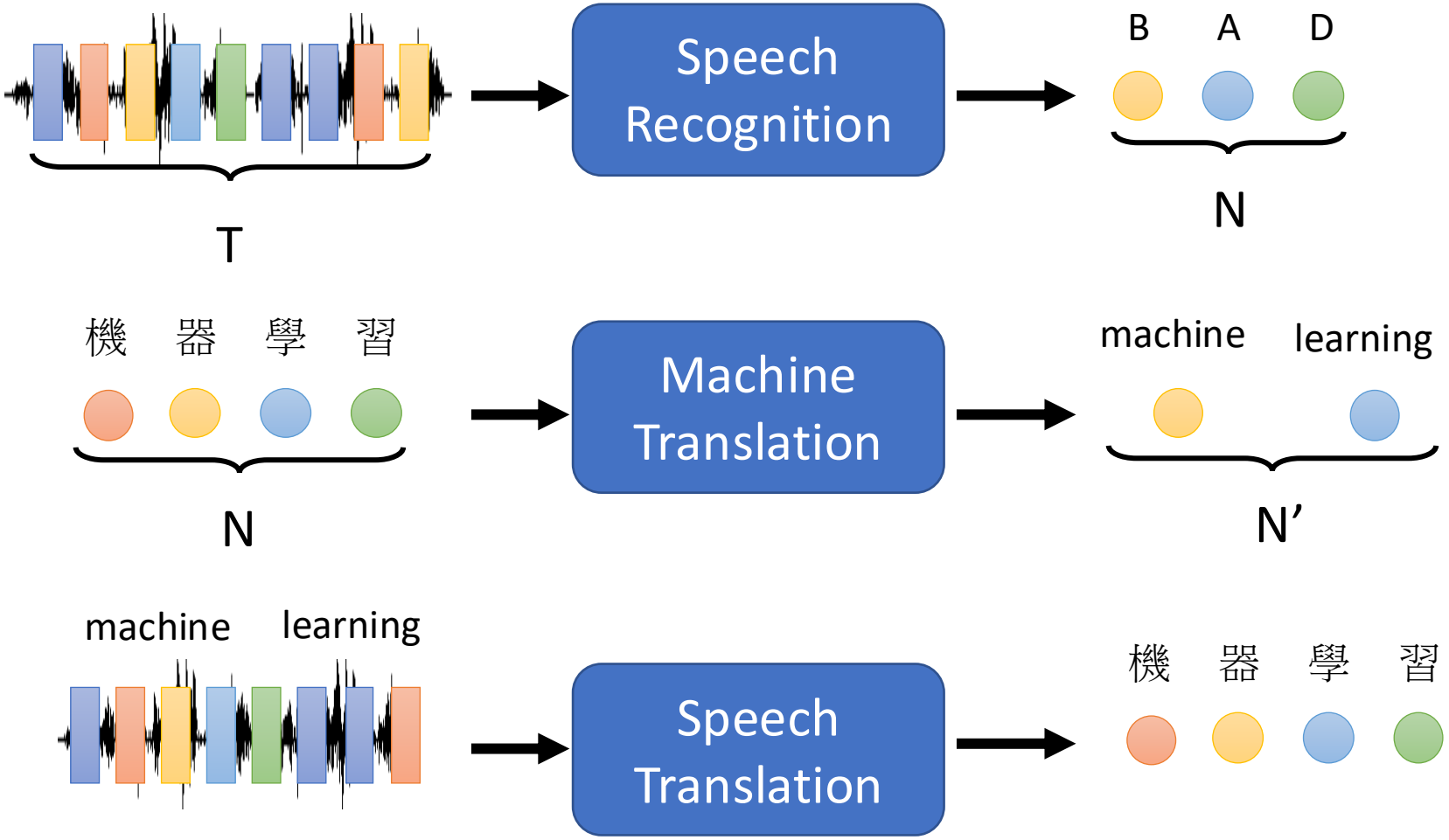
Why Transformer?

RNNs: slow sequential processing

LSTMs/GRUs: limited long-range memory (LLM has the same problem!!)

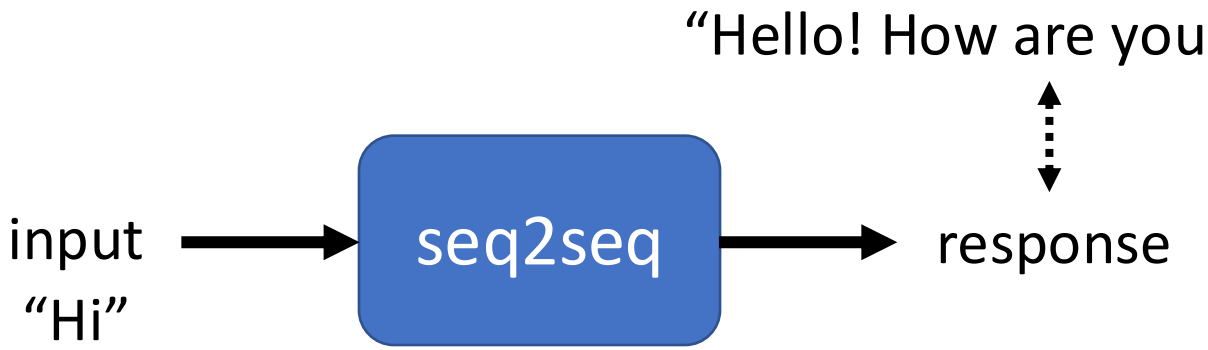
Sequence-to-sequence (Seq2seq)

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.

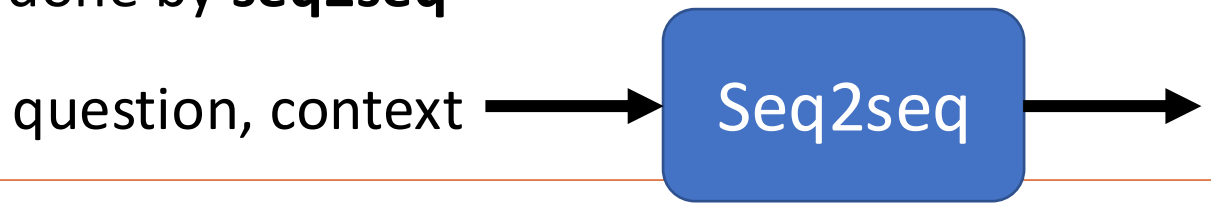
Most Natural Language Processing applications …

Question Answering (QA)

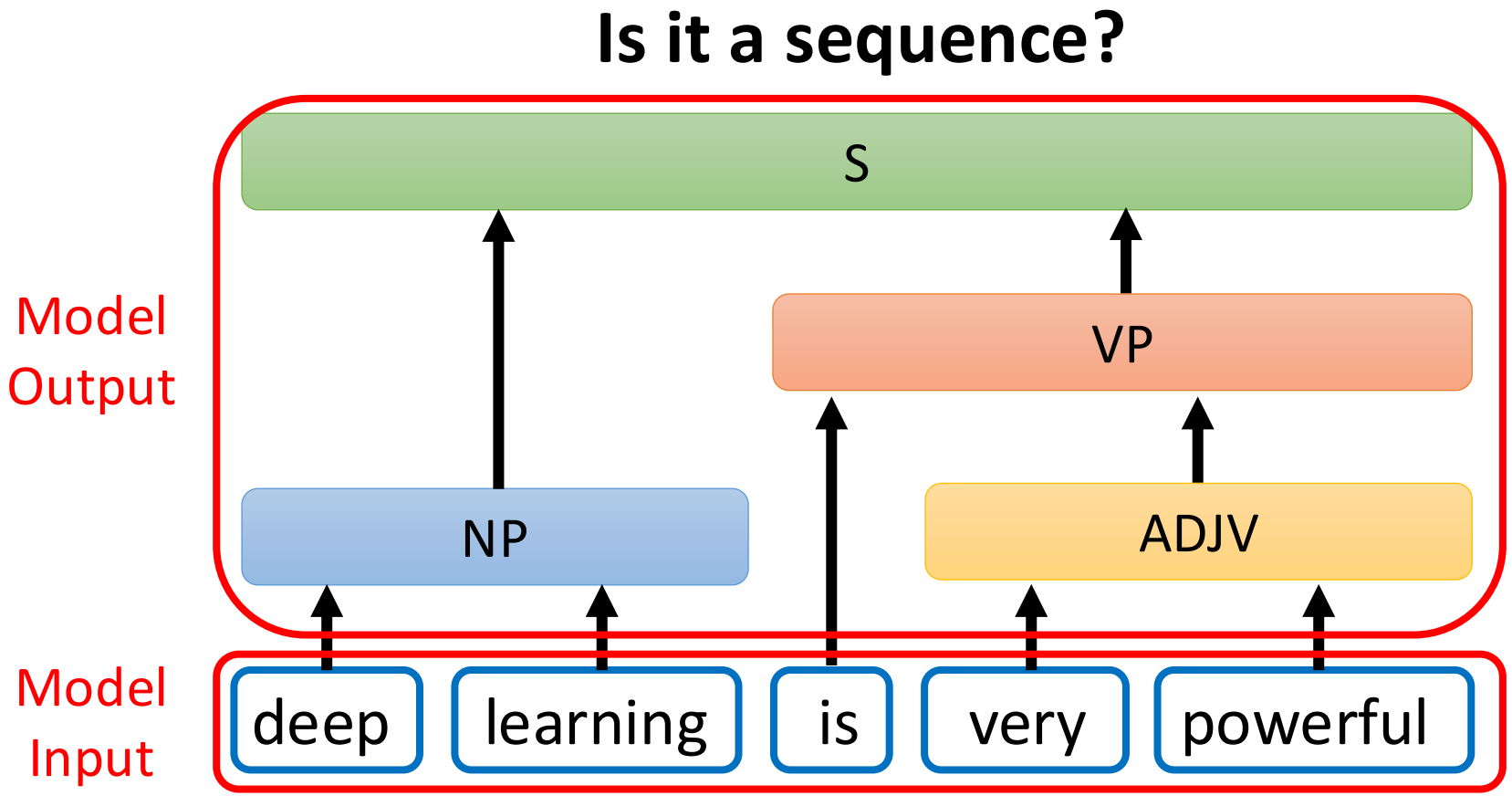
Question	Context	Answer
What is a major importance of Southern California in relation to California and the US?	...Southern California is a major economic center for the state of California and the US....	major economic center
What is the translation from English to German?	Most of the planet is ocean 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

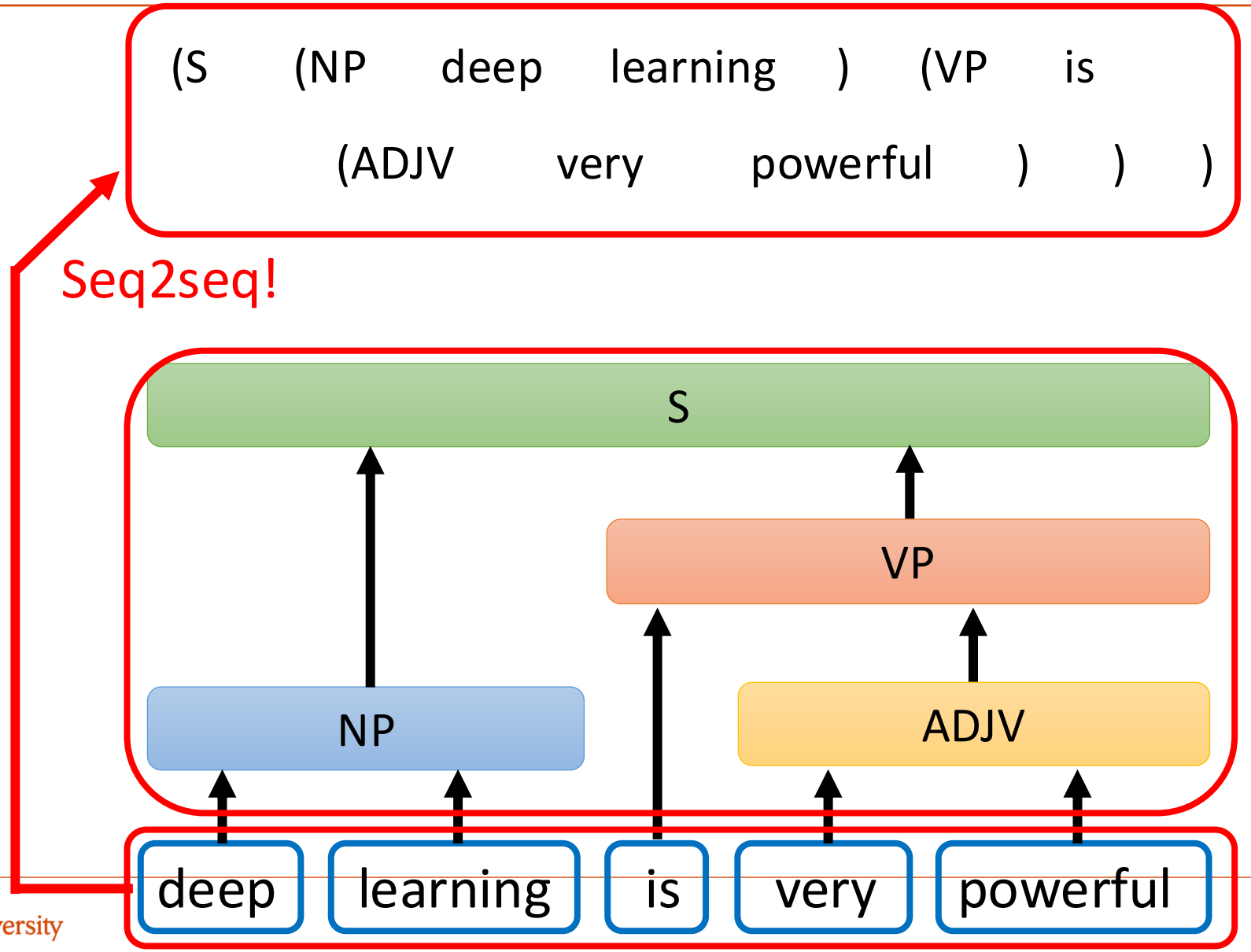
QA can be done by **seq2seq**



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

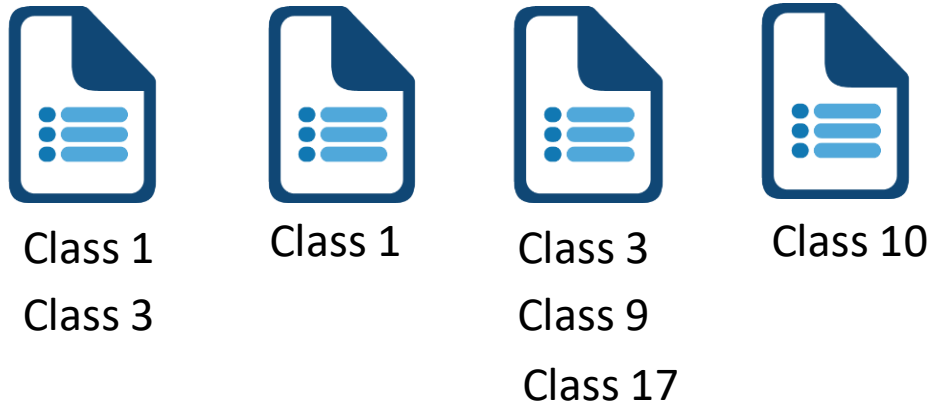
Geoffrey Hinton
Google
geoffhinton@google.com

https://arxiv.org/abs/1412.7449

deep learning is very powerful

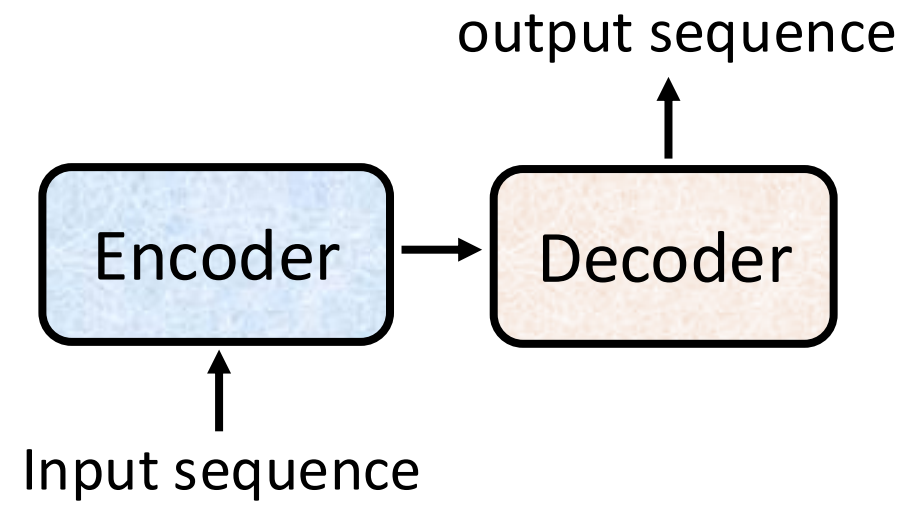
Seq2seq for Multi-Label Classification

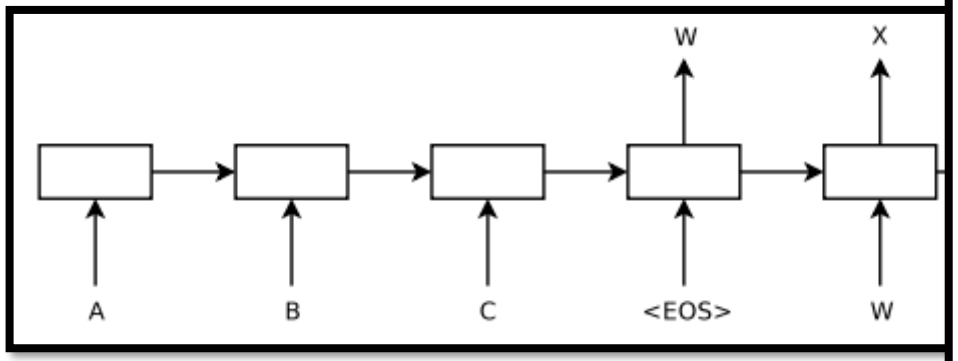
An object can belong to multiple classes.



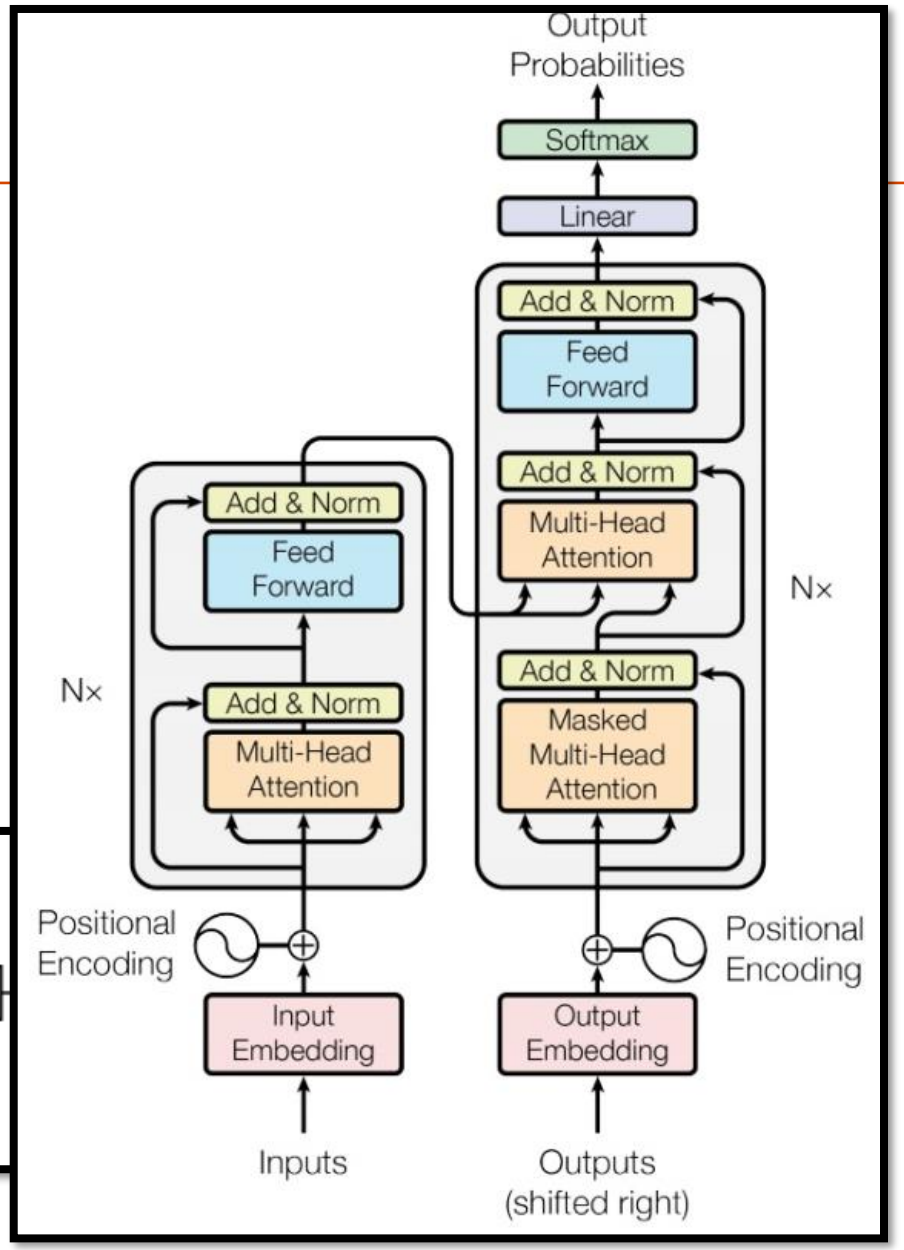
https://arxiv.org/abs/1909.03434
https://arxiv.org/abs/1707.05495

Seq2seq



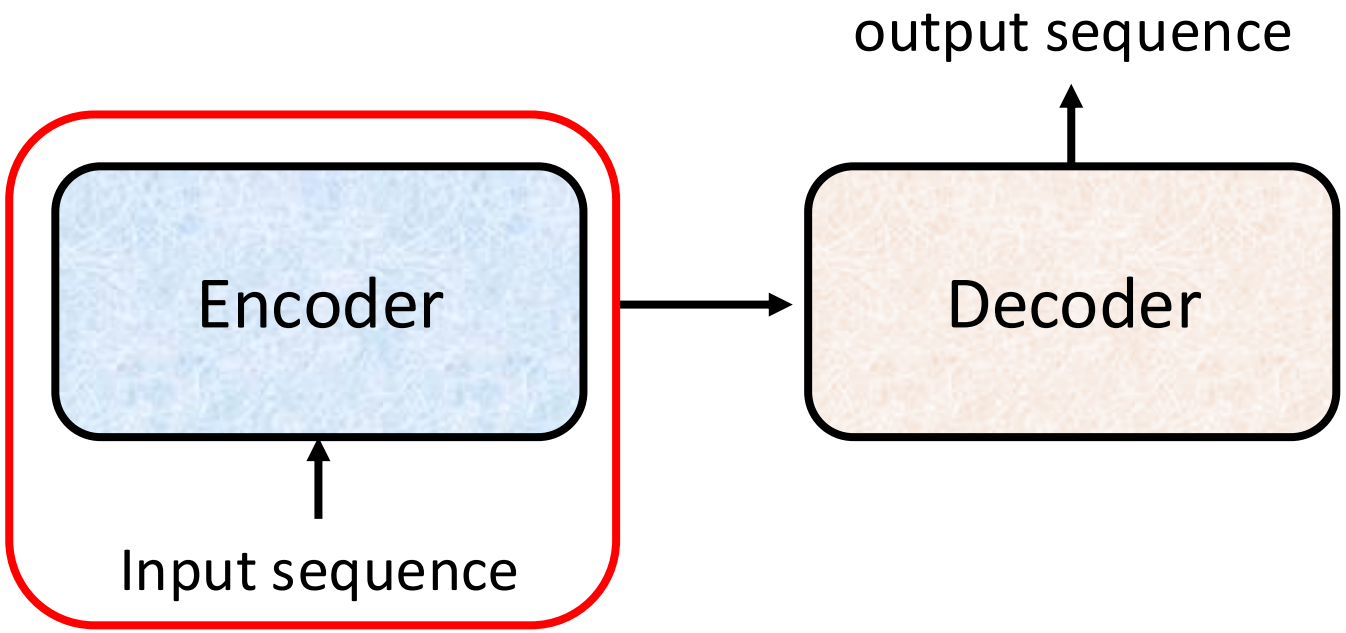


Sequence to Sequence Learning with Neural Networks
https://arxiv.org/abs/1409.3215



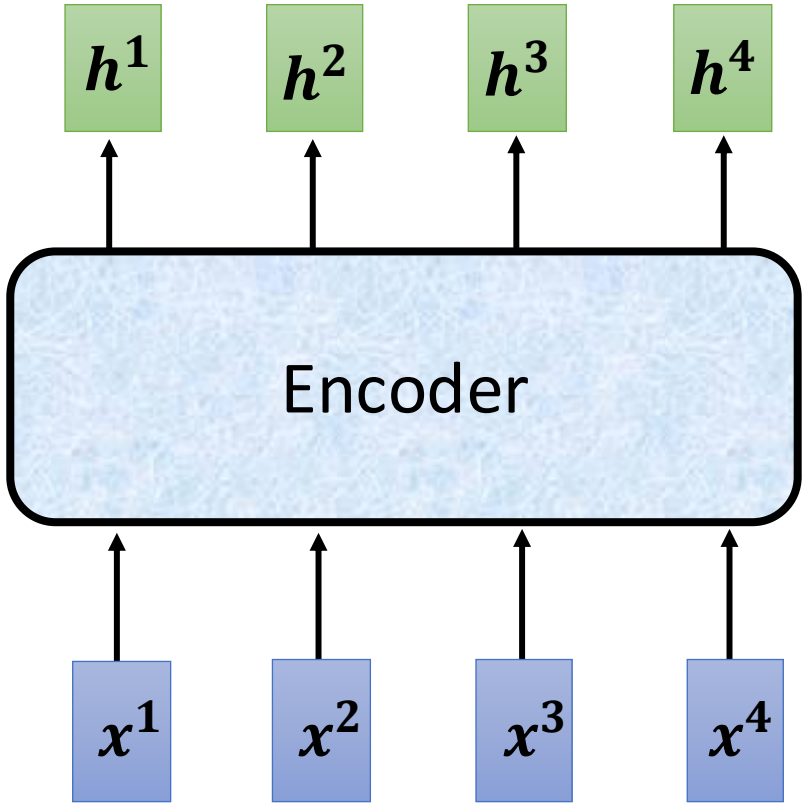
Transformer
https://arxiv.org/abs/1706.03762

Encoder

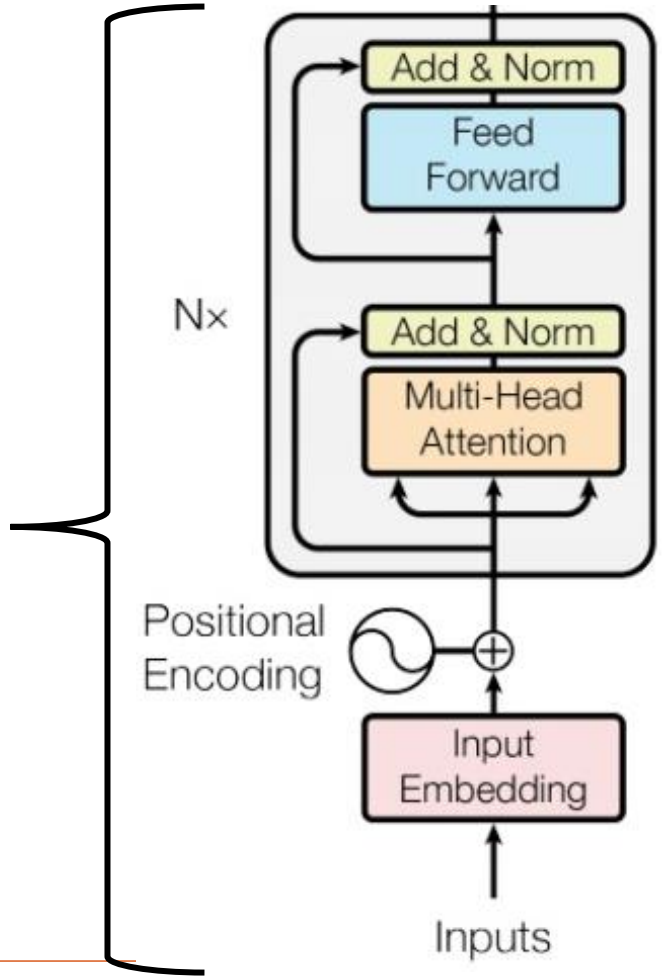


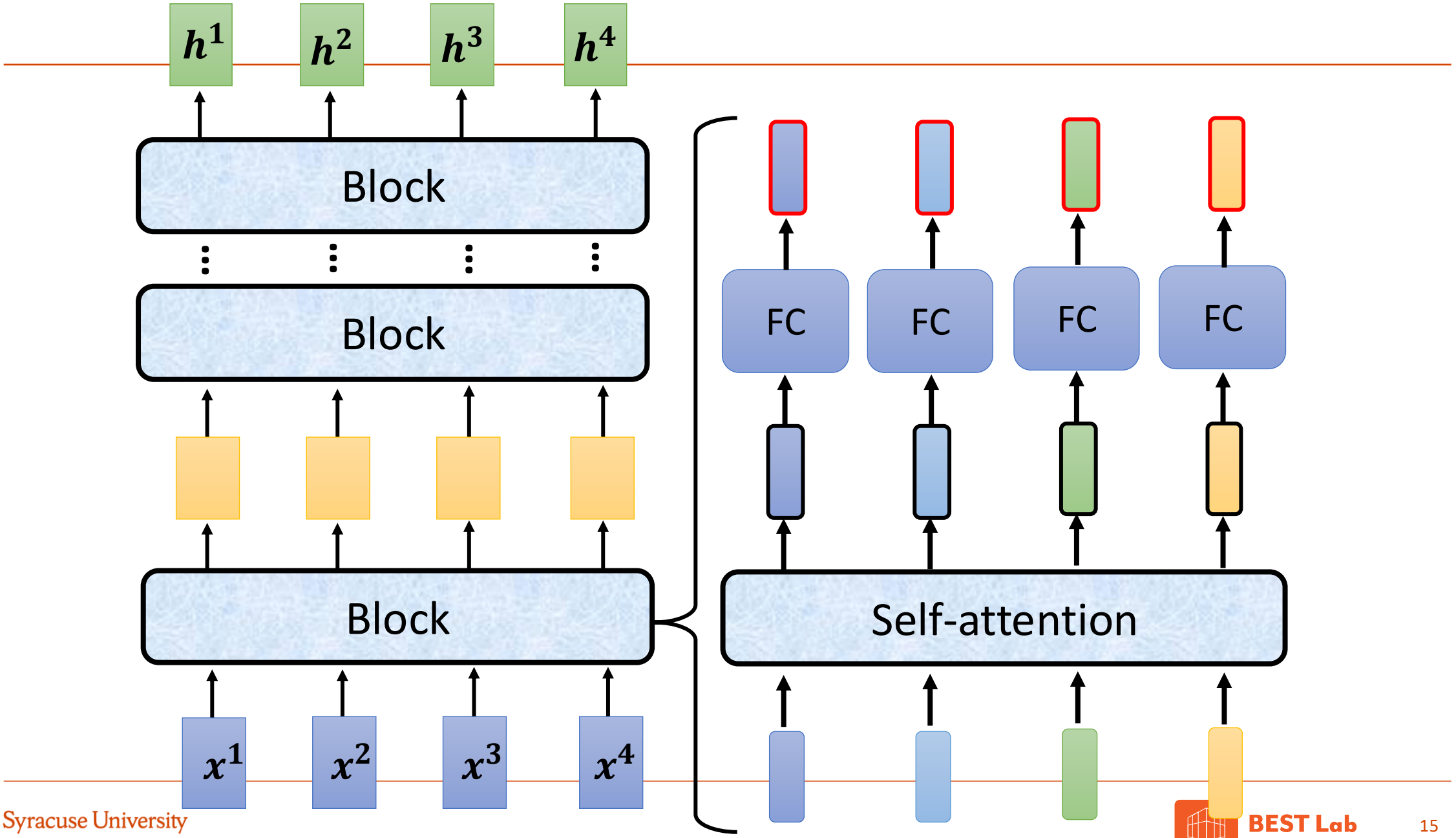
Encoder

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

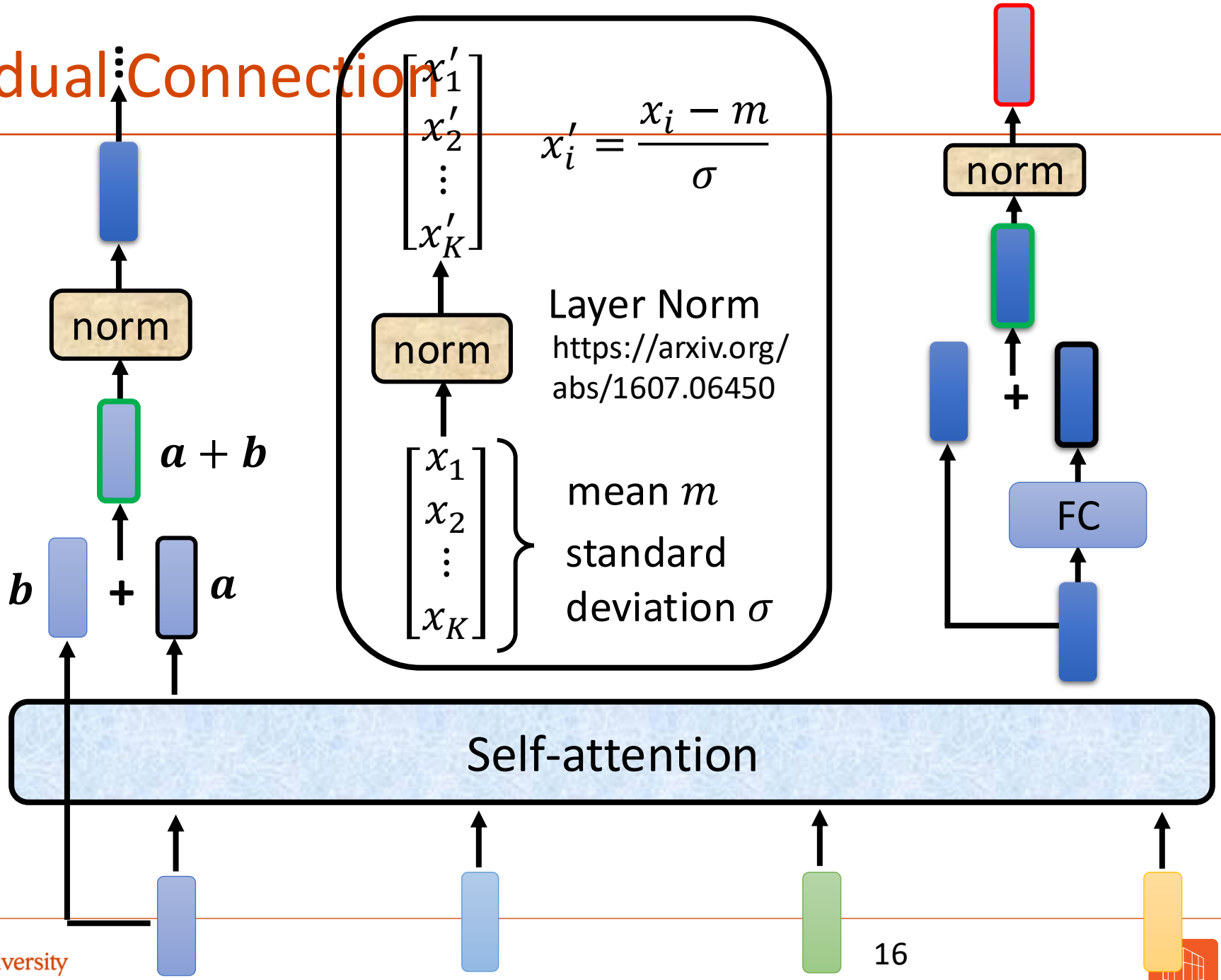


Transformer's Encoder

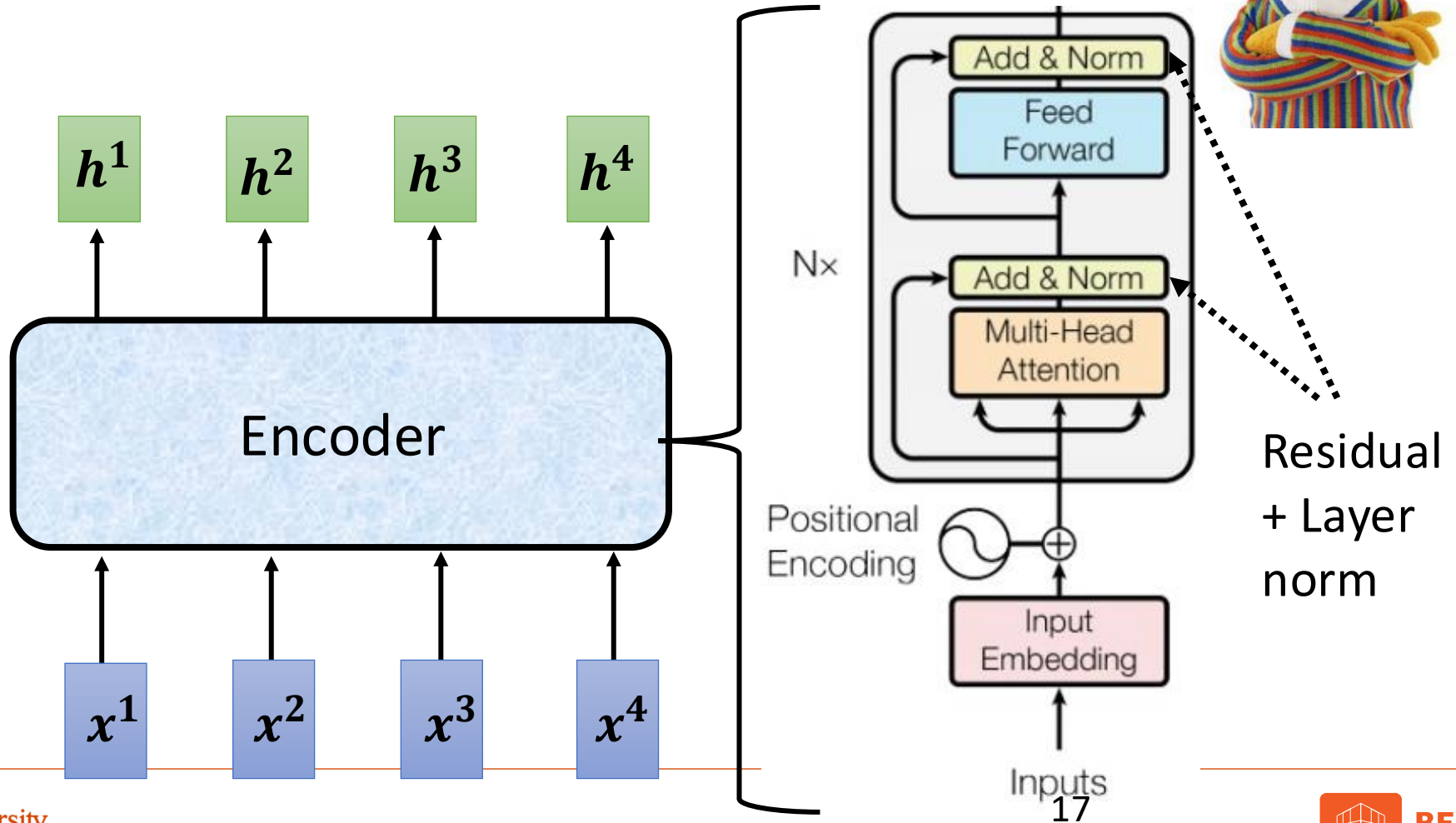



h^1
h^2
h^3
h^4
Block
Block
Block
FC
FC
FC
FC
Self-attention
x^1
x^2
x^3
x^4

Residual Connection



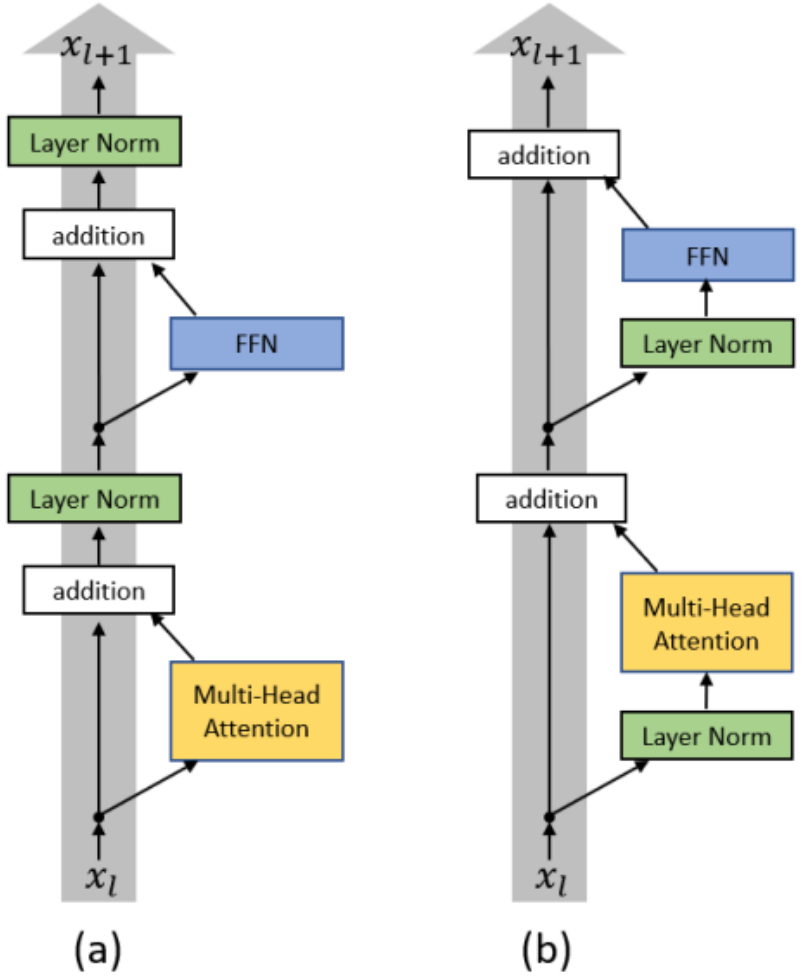
I use the **same** network architecture as **transformer encoder**.



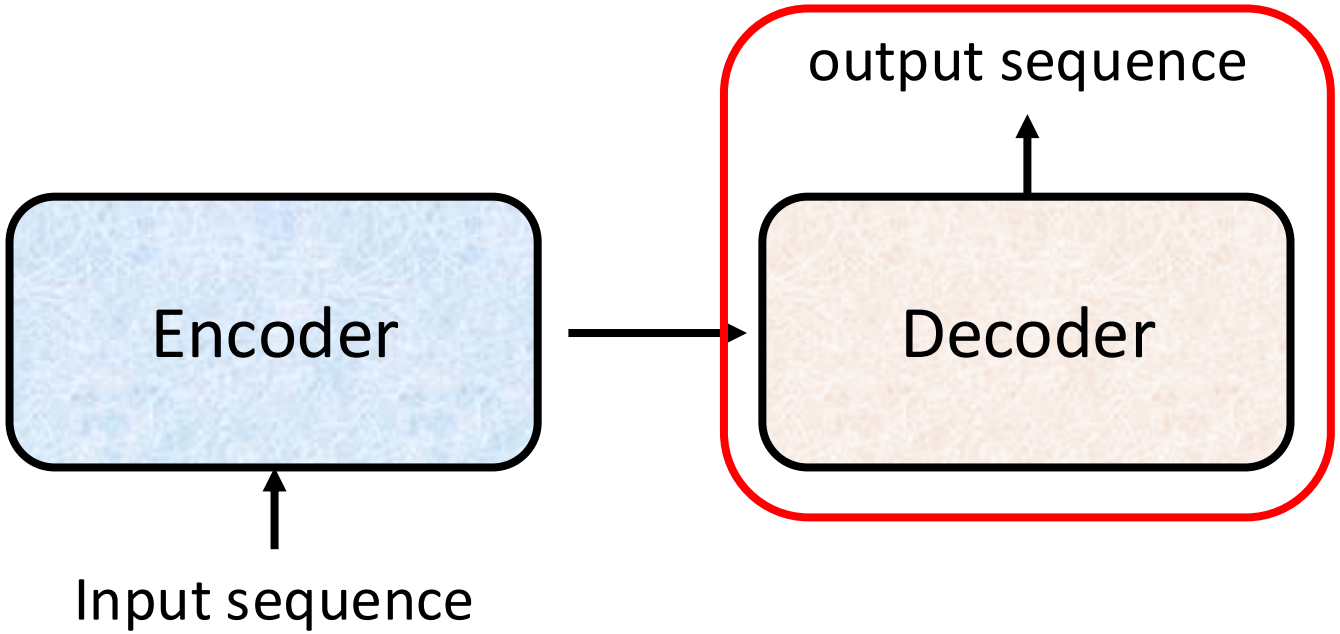
To learn more ……

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

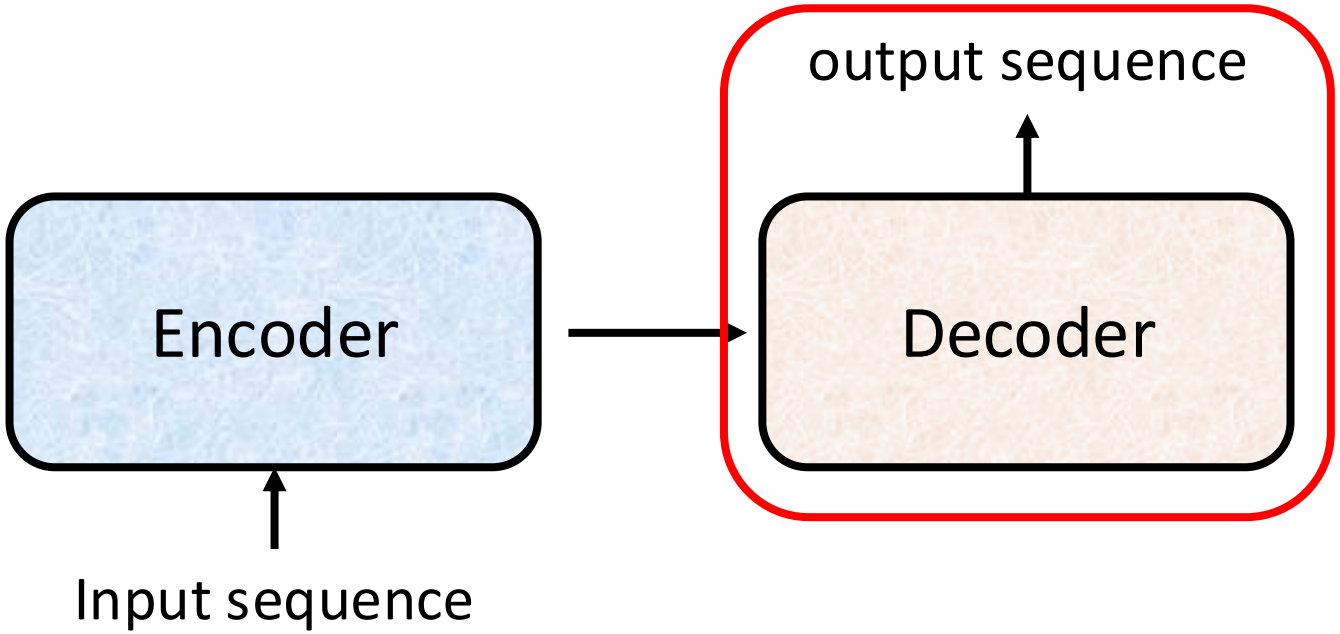
- PowerNorm: Rethinking Batch Normalization in Transformers
- https://arxiv.org/abs/2003.07845



Decoder

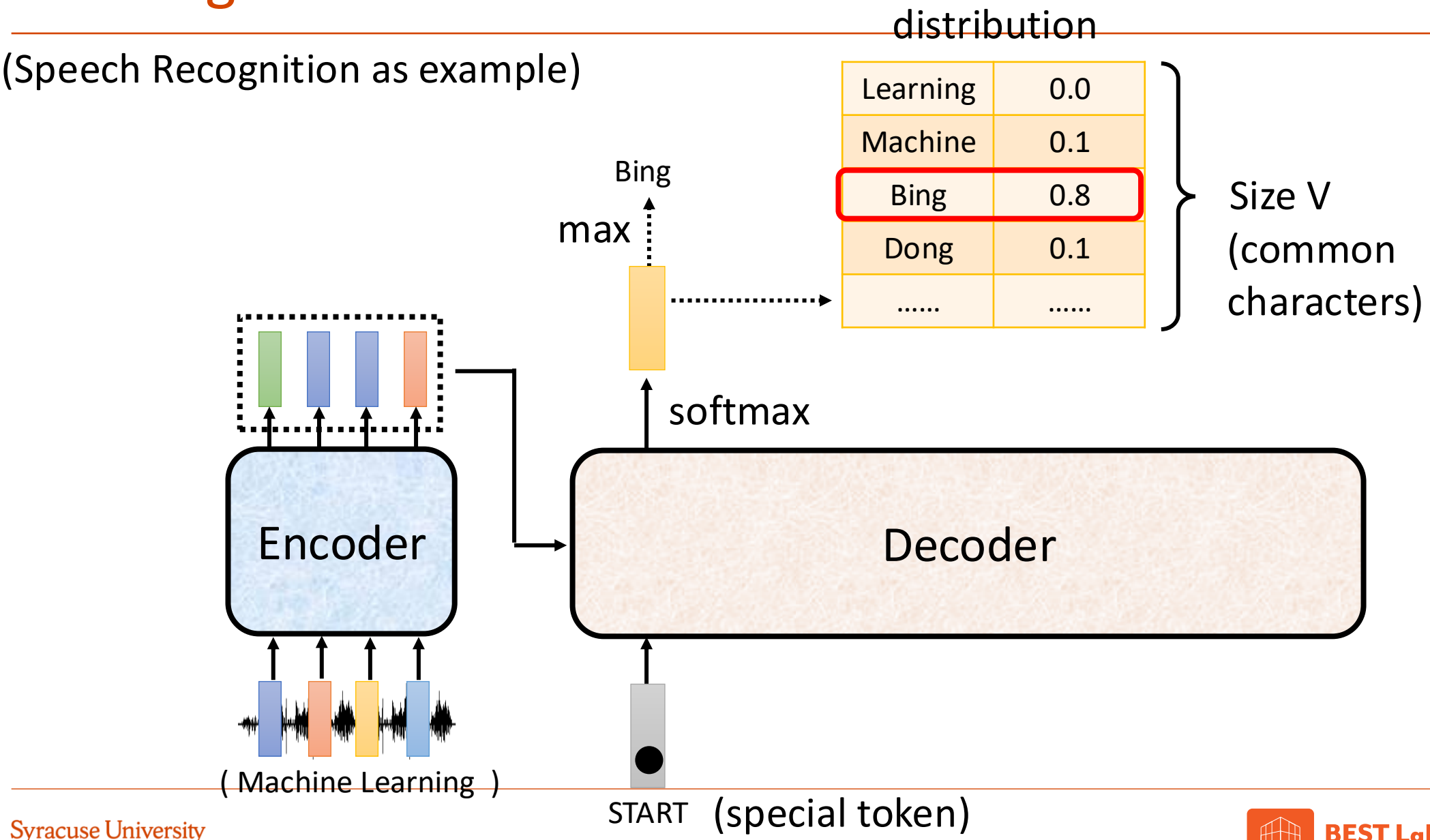


Decoder – Autoregressive (AT)

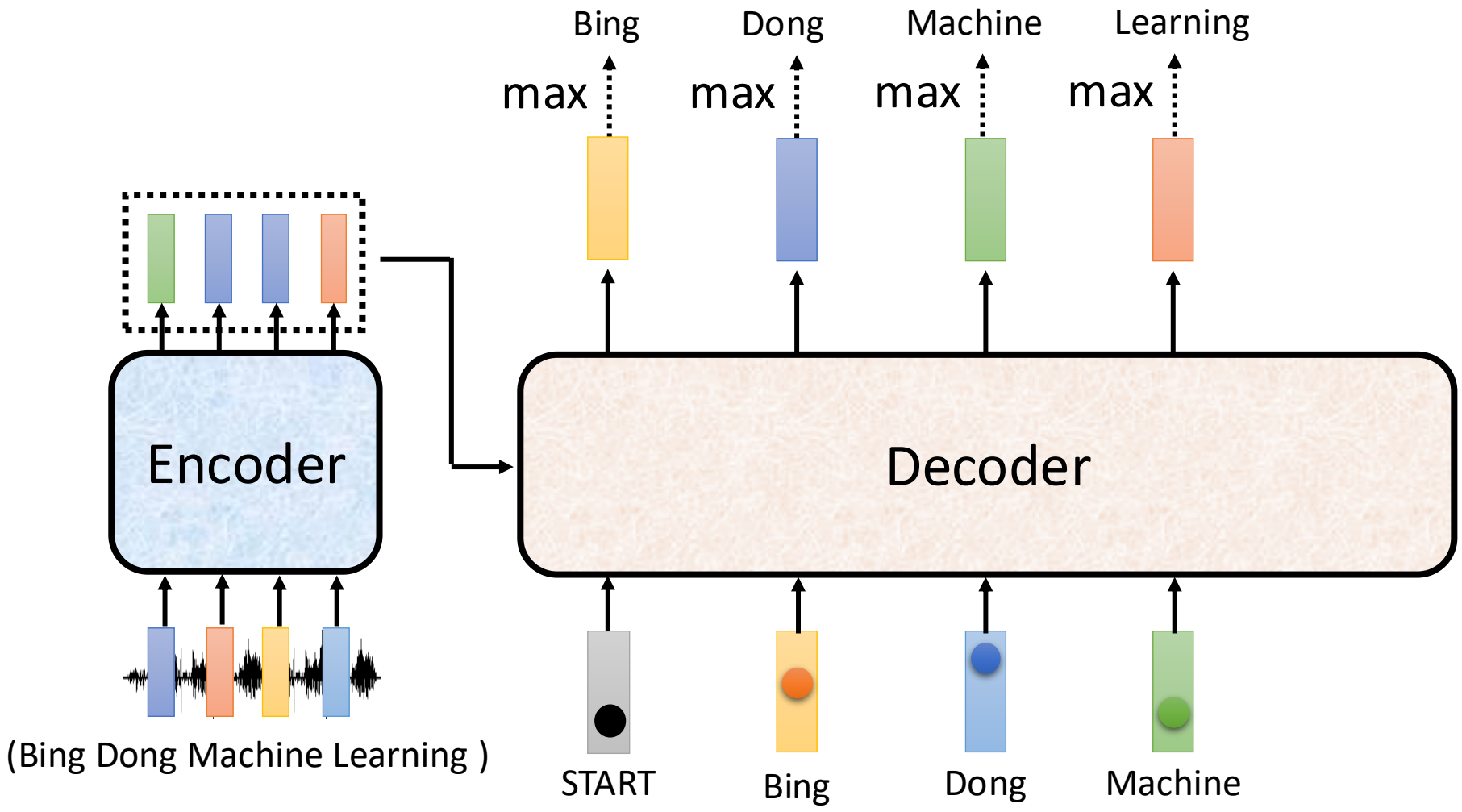


Autoregressive

(Speech Recognition as example)

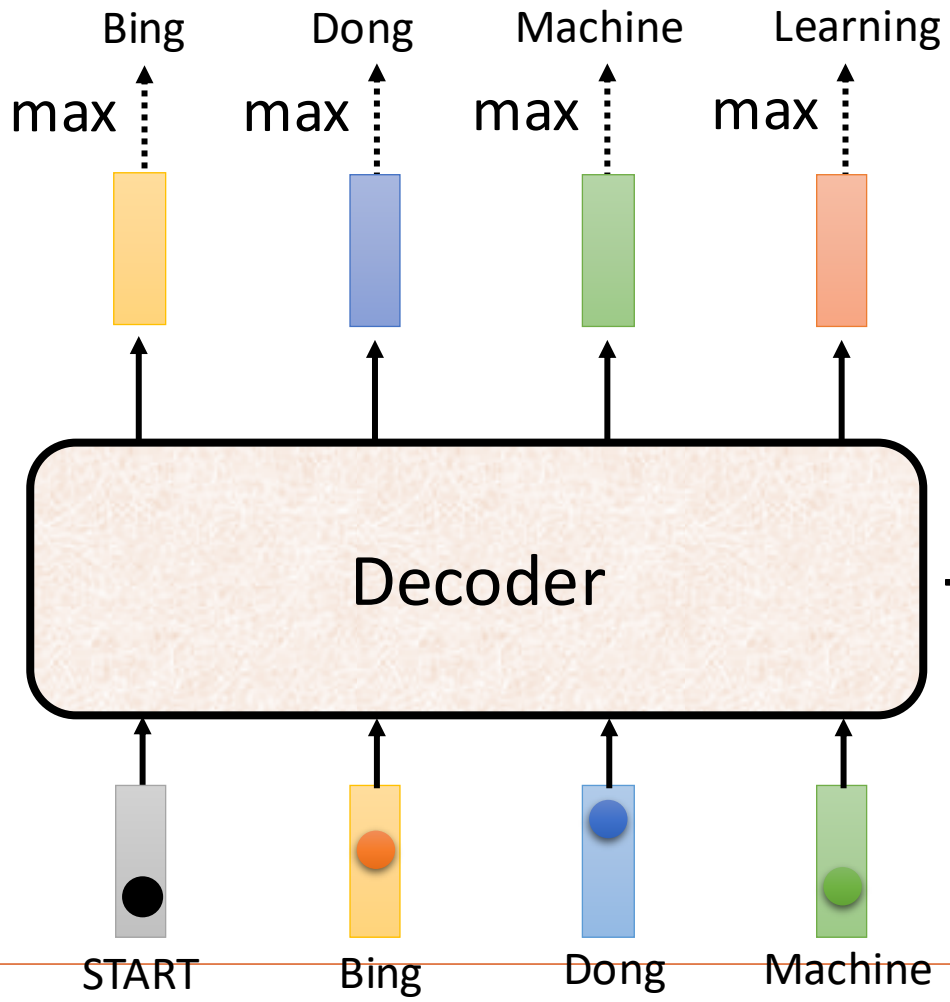


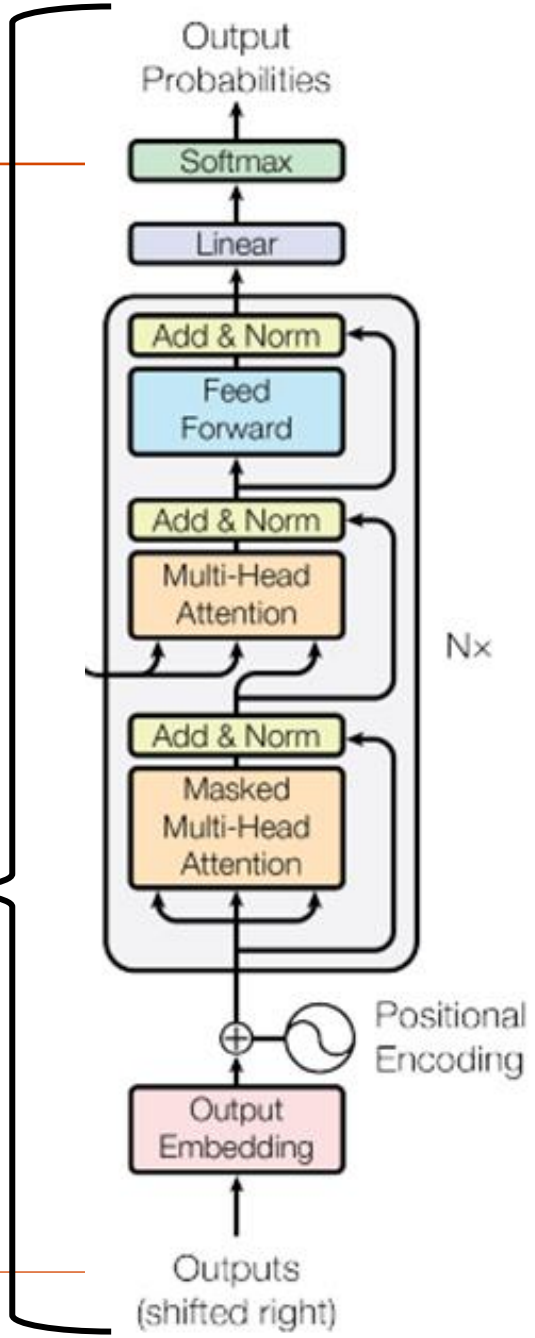
Autoregressive

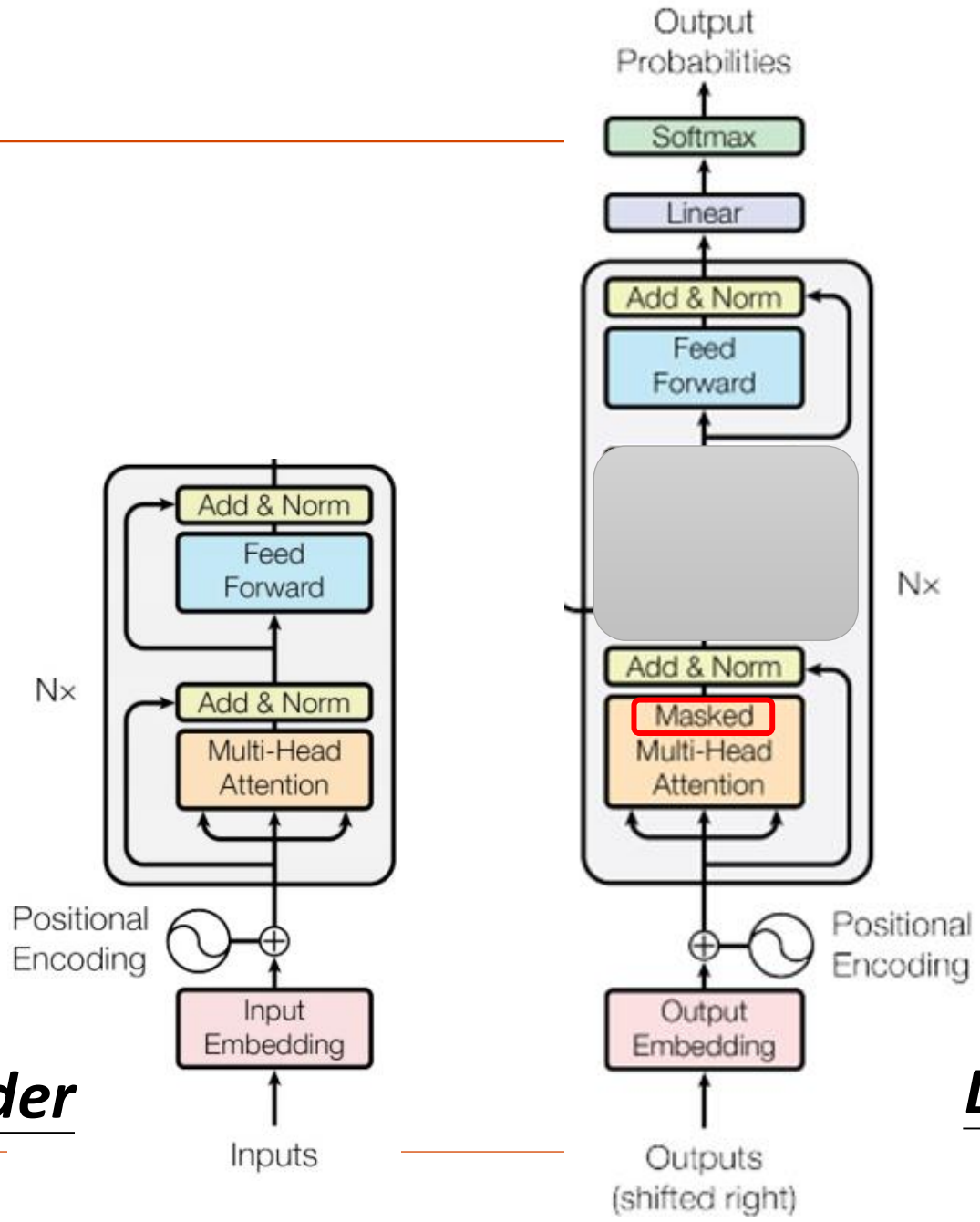


Autoregressive

ignore the input from the encoder here ☺





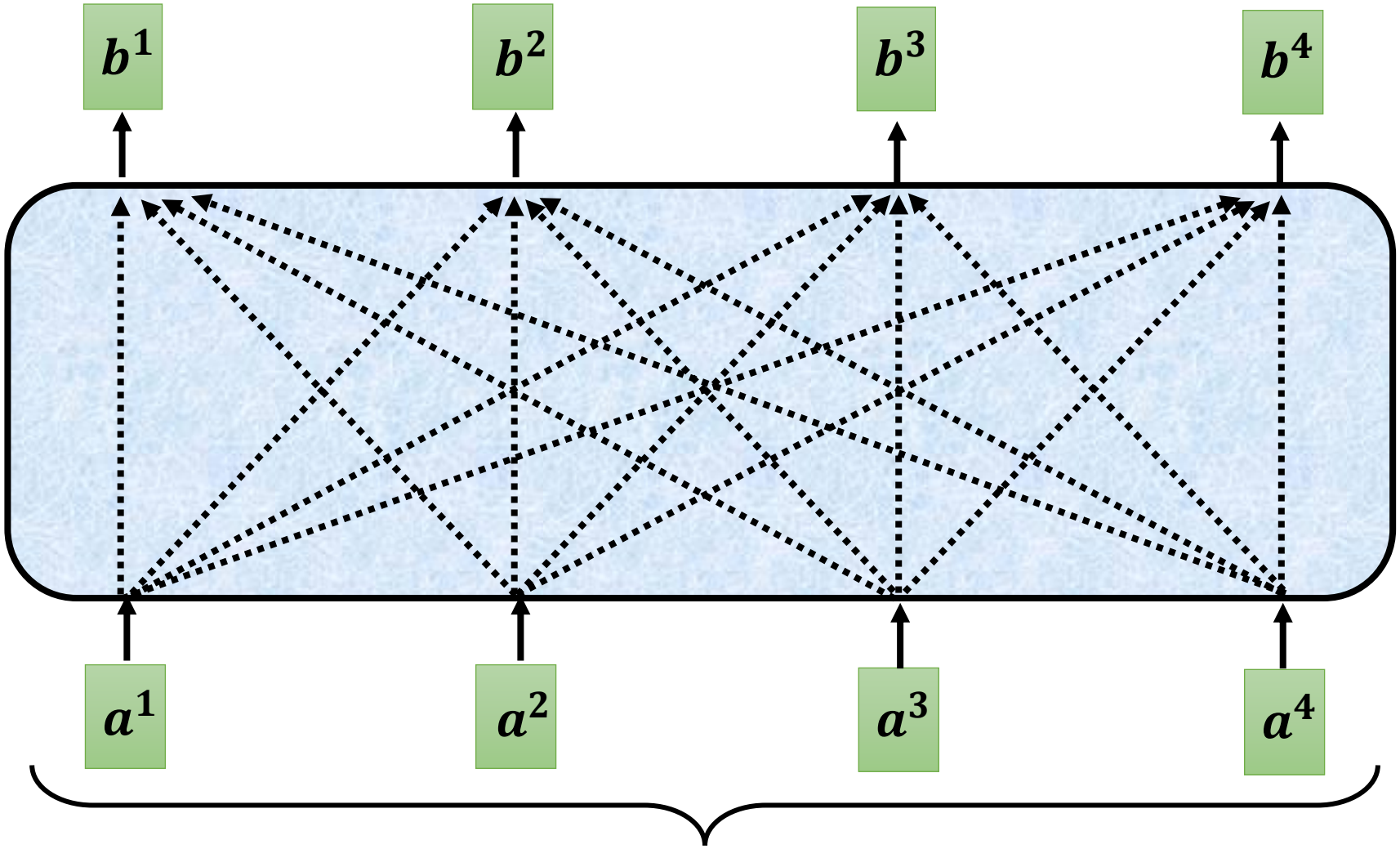


Encoder

Decoder

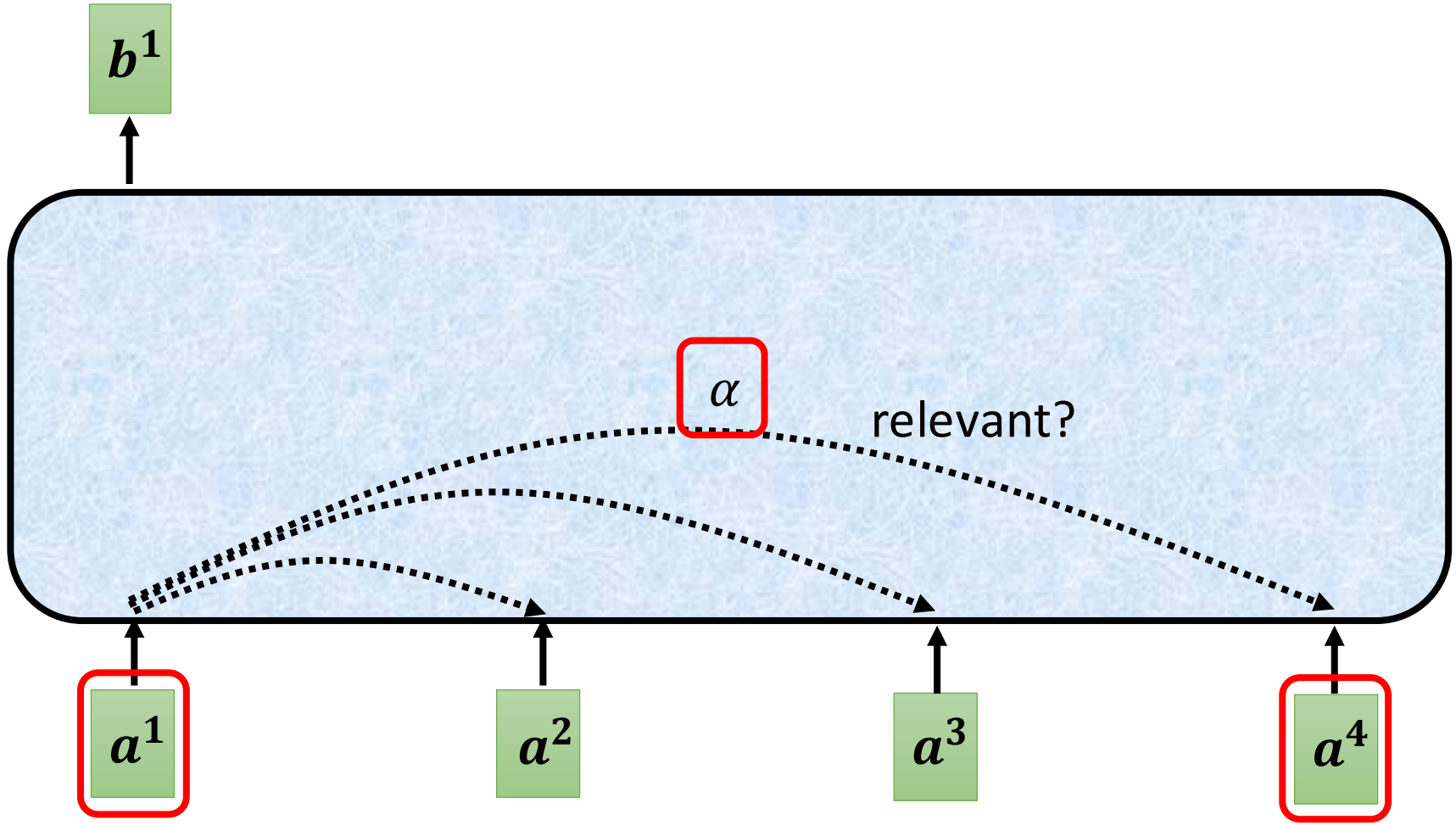
What’s inside Self-attention layer?

Self-attention



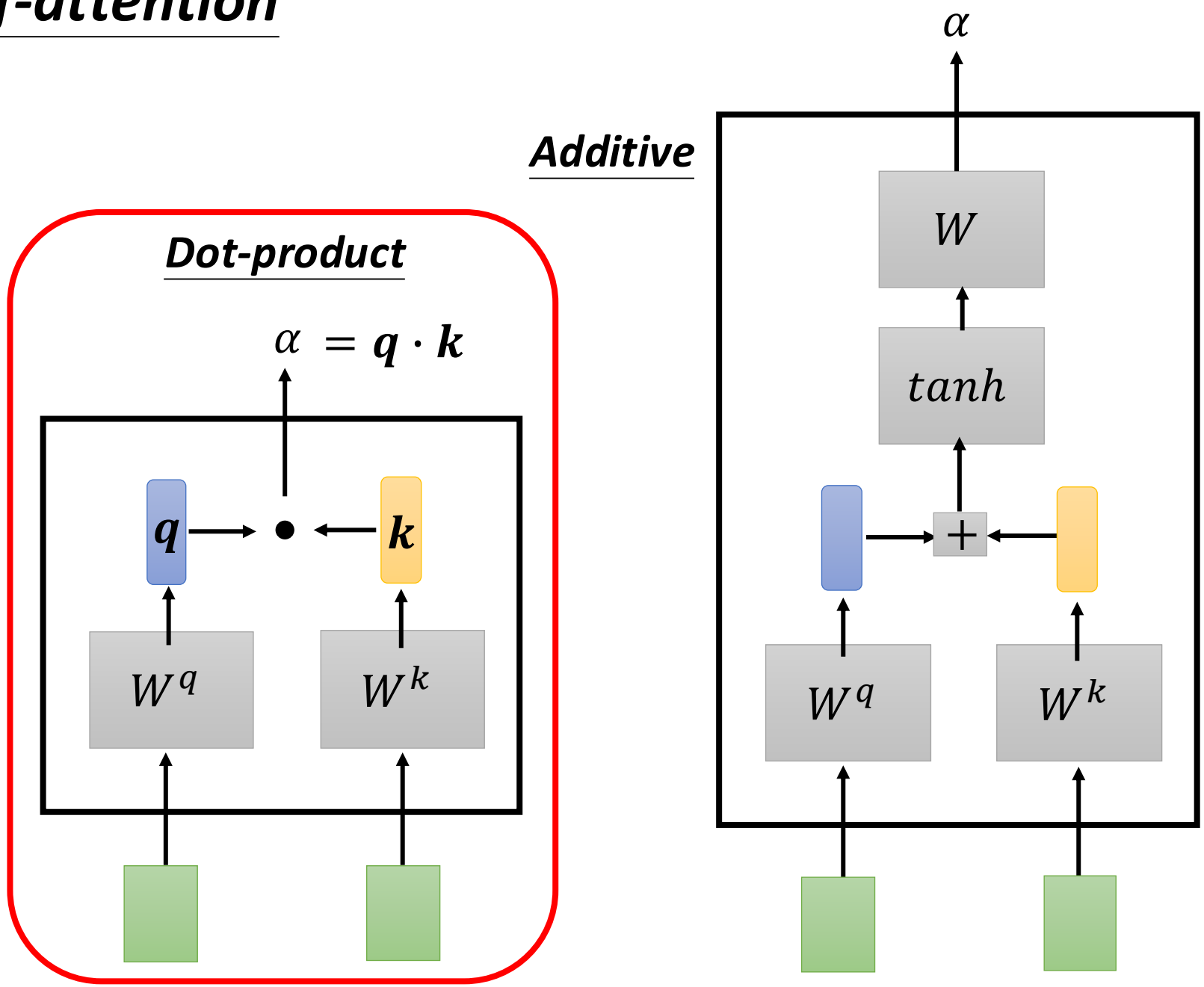
Can be either **input** or **a hidden layer**

Self-attention



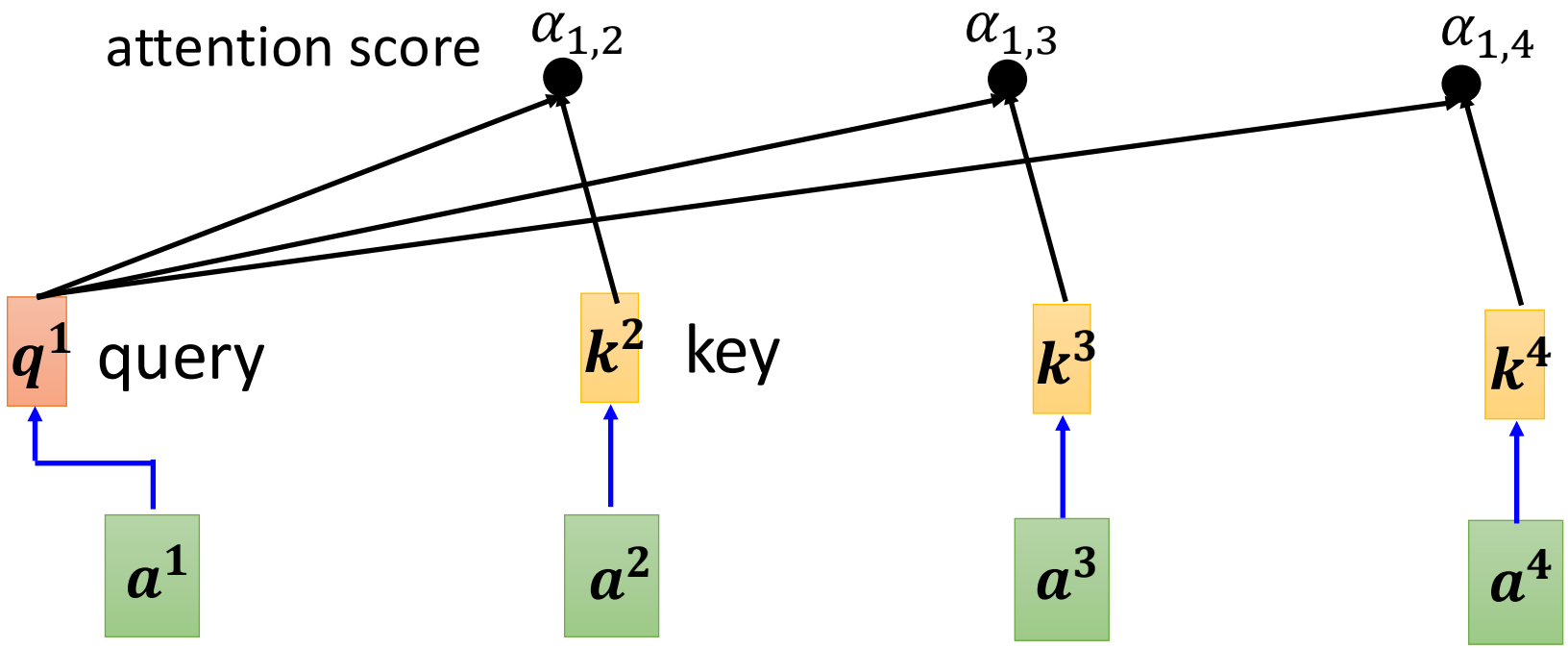
Find the relevant vectors in a sequence

Self-attention



Self-attention

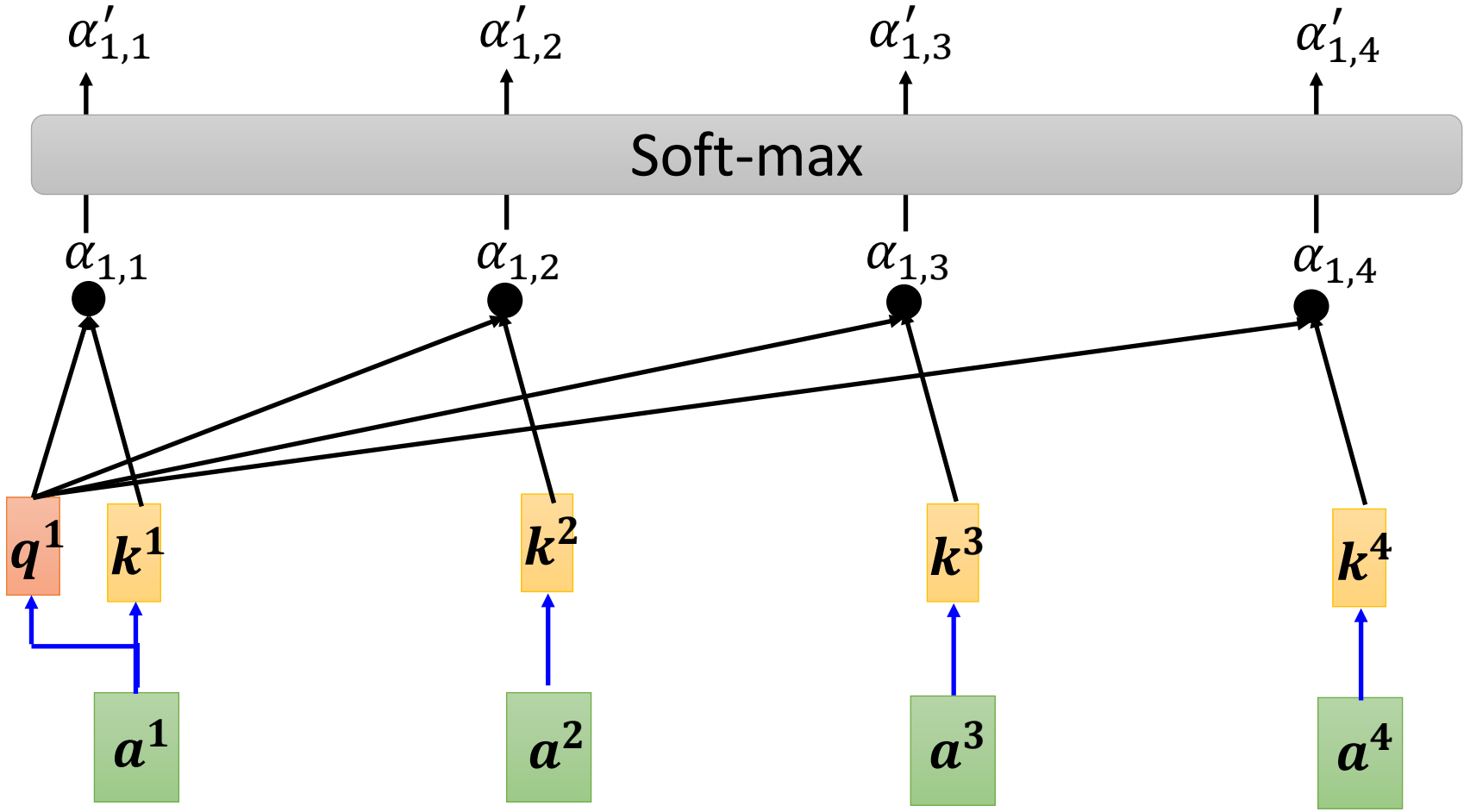
$$\alpha_{1,2} = \boldsymbol{q^1} \cdot \boldsymbol{k^2} \qquad \alpha_{1,3} = \boldsymbol{q^1} \cdot \boldsymbol{k^3} \qquad \alpha_{1,4} = \boldsymbol{q^1} \cdot \boldsymbol{k^4}$$



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

Self-attention

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



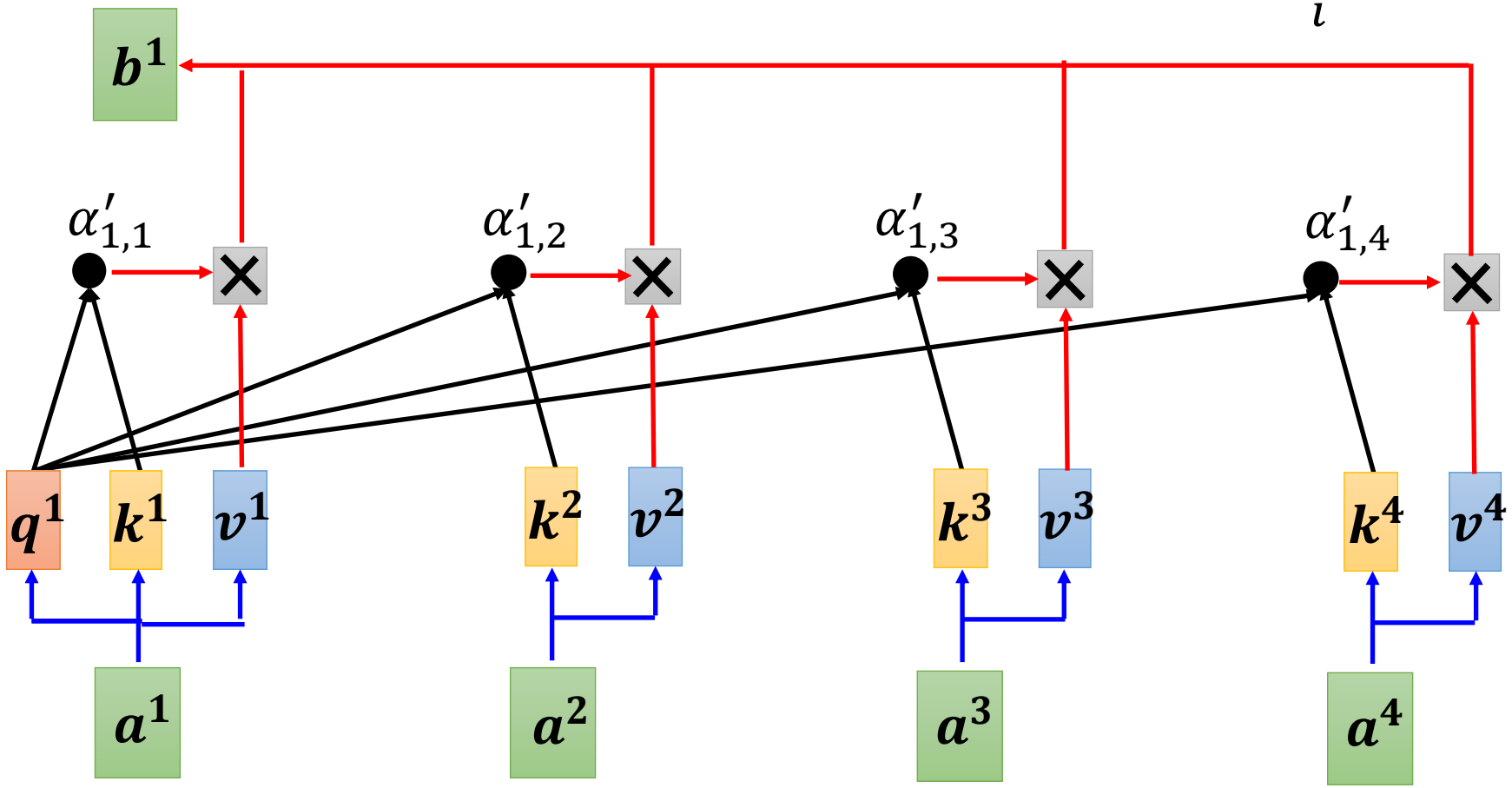
$q^1 = W^q a^1$ $k^2 = W^k a^2$ $k^3 = W^k a^3$ $k^4 = W^k a^4$

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

Self-attention

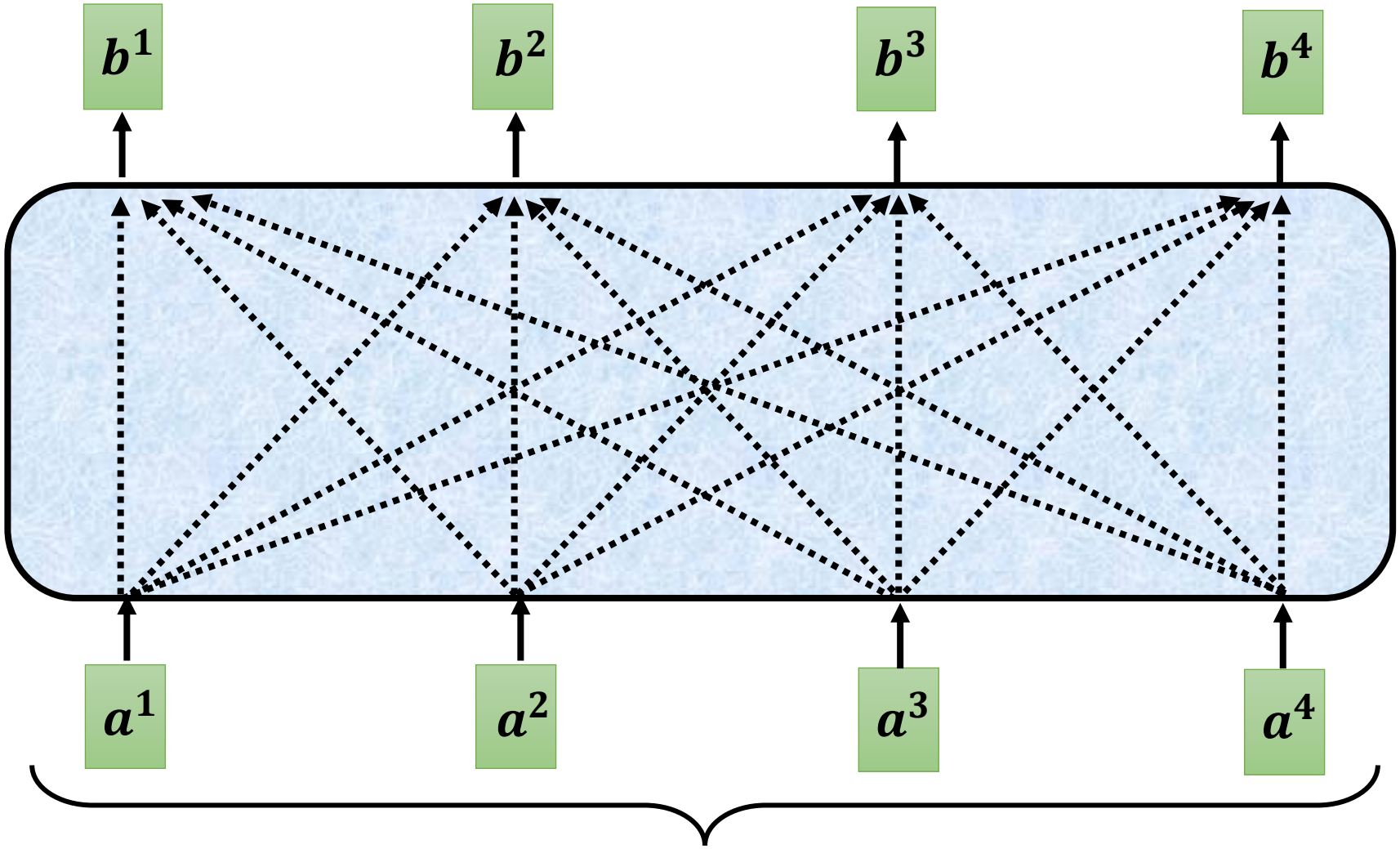
Extract information based on attention scores

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



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

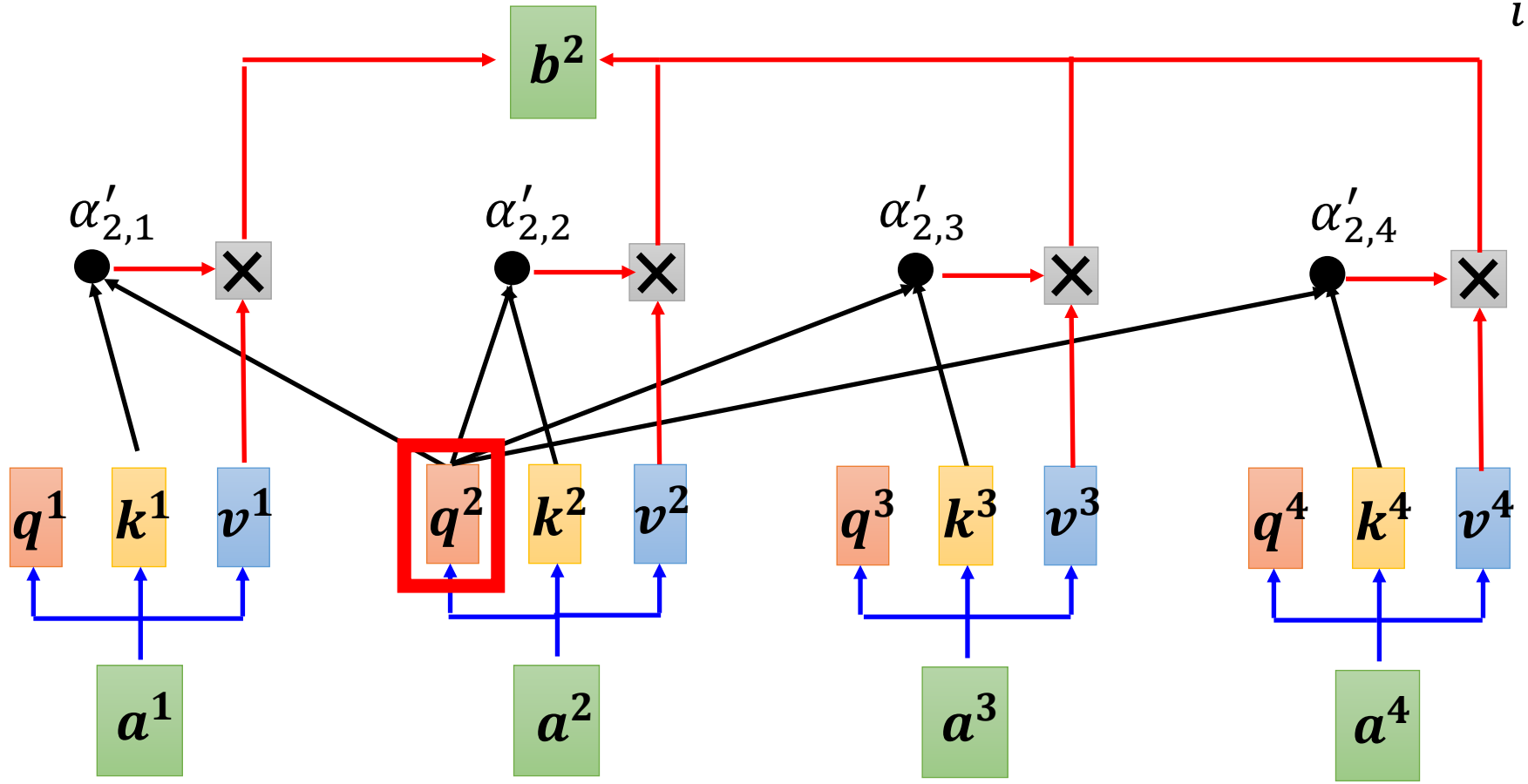
Self-attention



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

Self-attention → Masked Self-attention

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



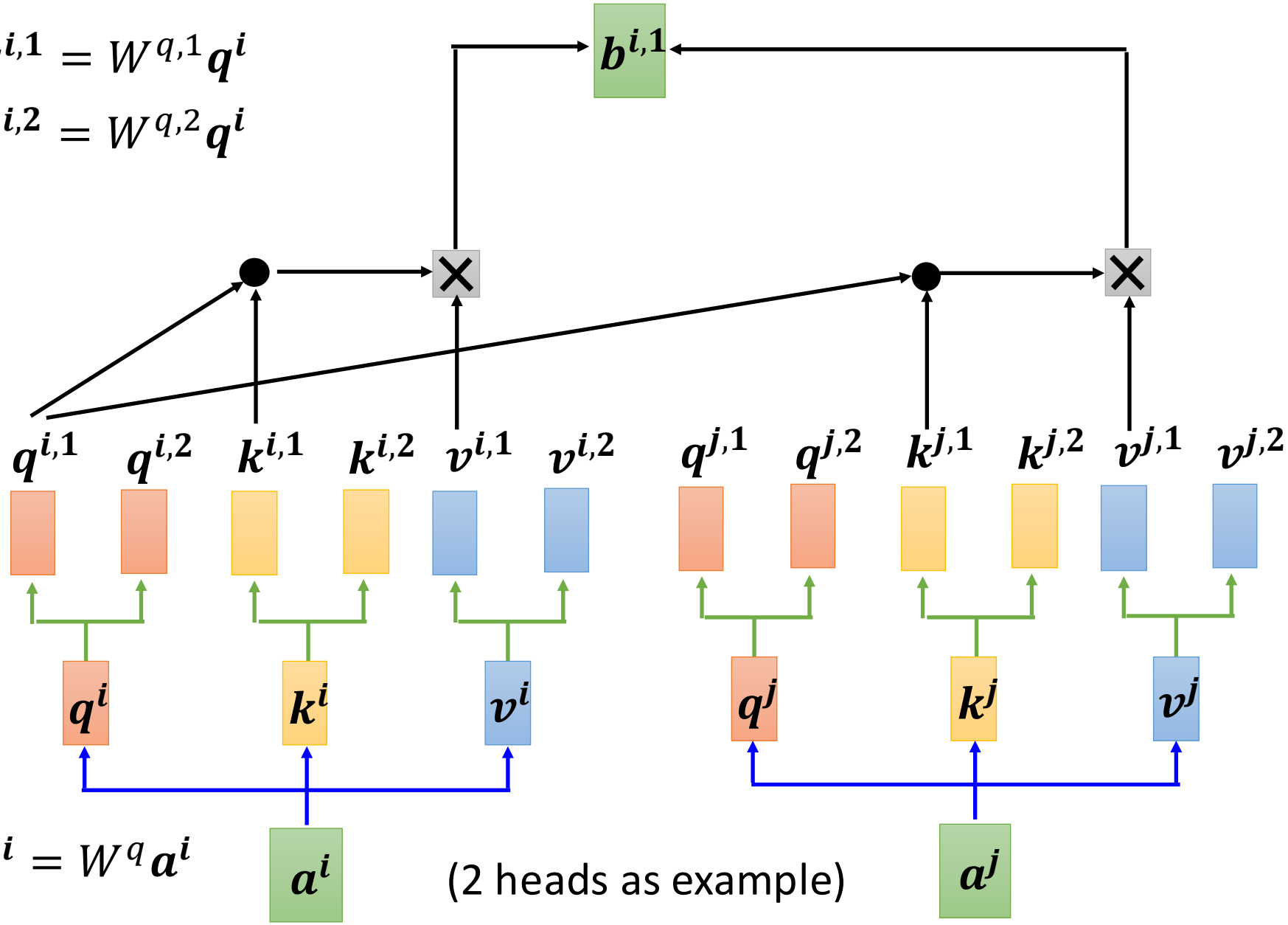
Why masked? Consider how does decoder work

Multi-head Self-attention

Different types of relevance

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

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



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

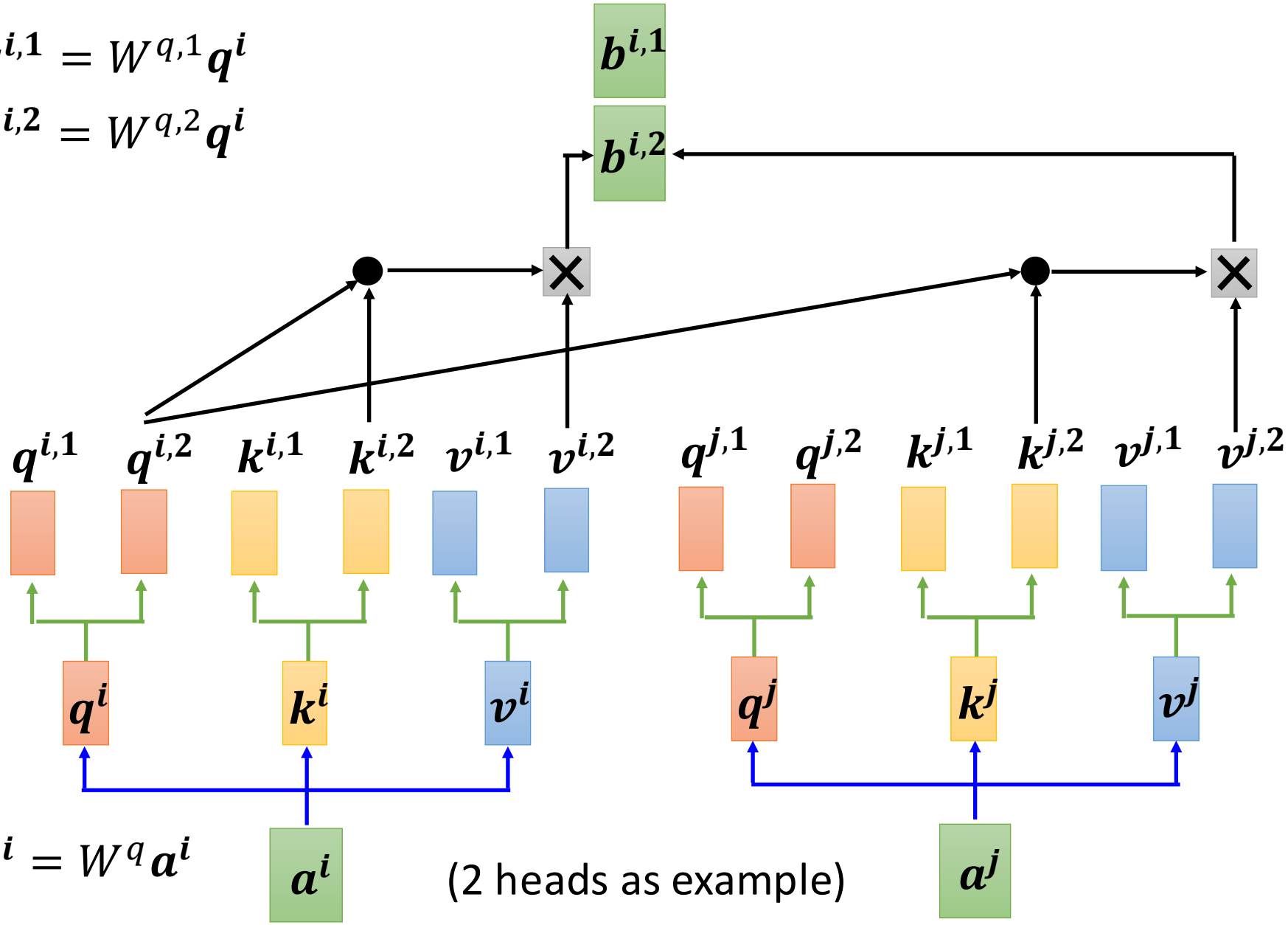
(2 heads as example)

Multi-head Self-attention

Different types of relevance

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

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

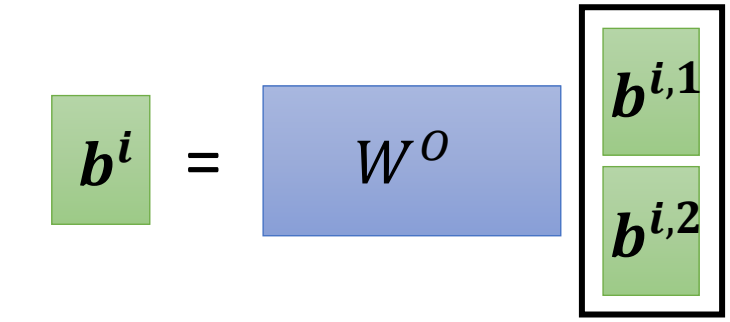


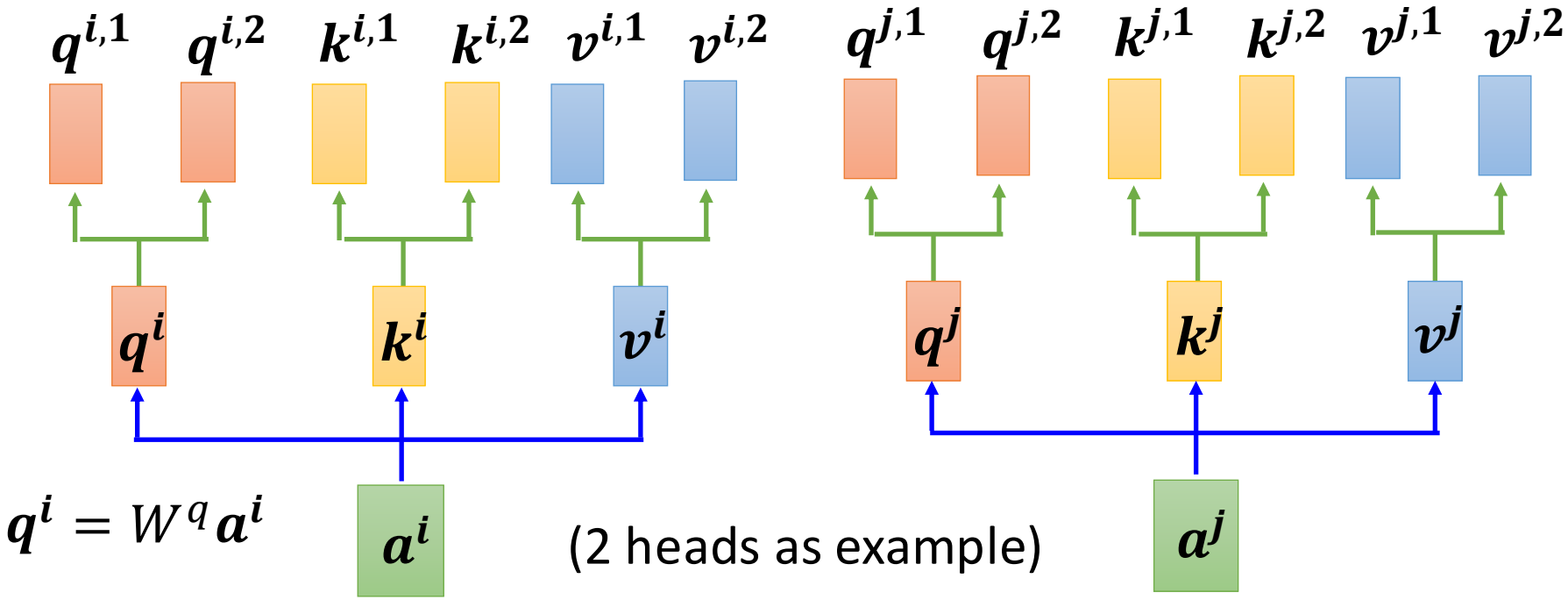
$$q^i = W^q a^i$$

(2 heads as example)

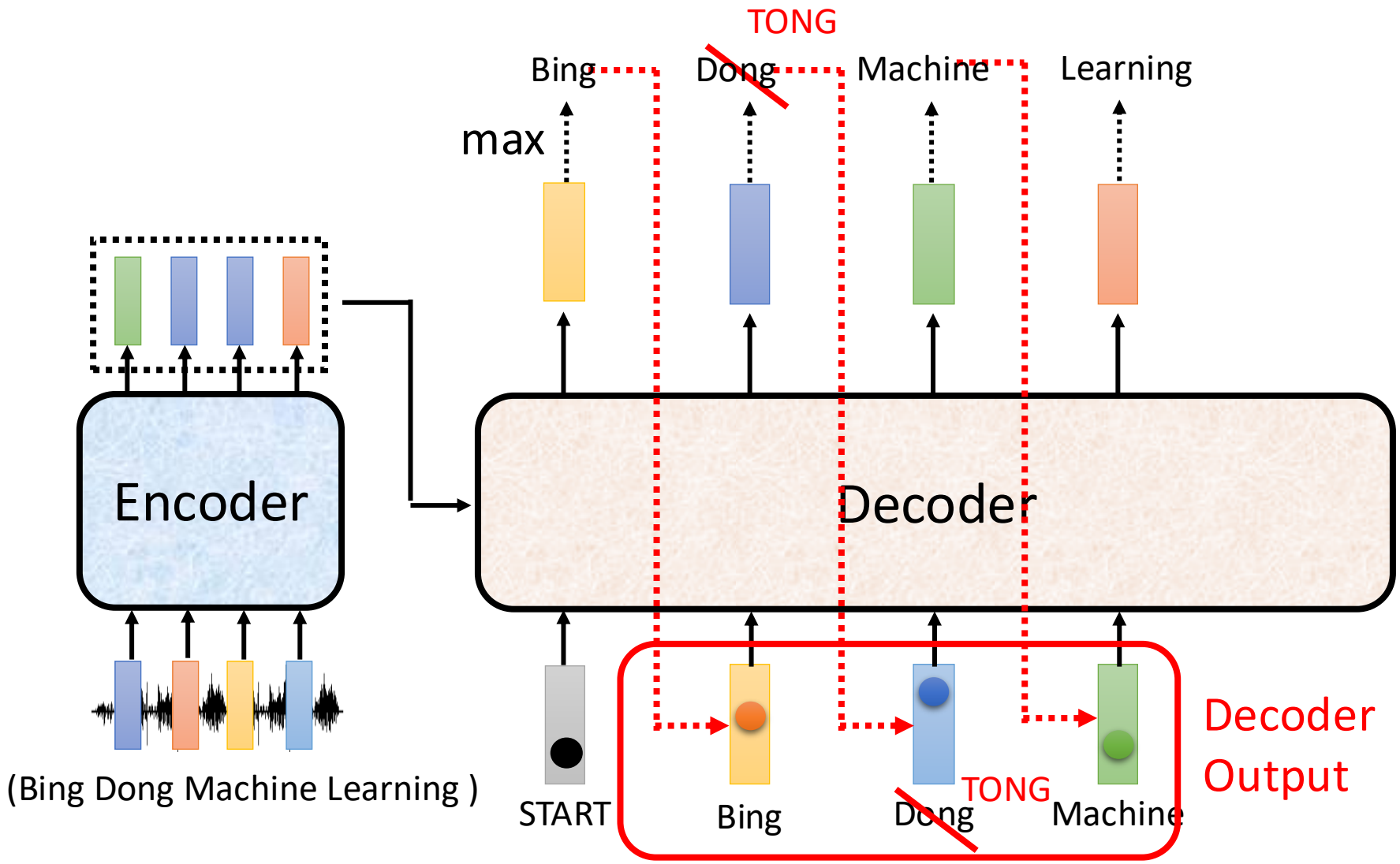
Multi-head Self-attention

Different types of relevance





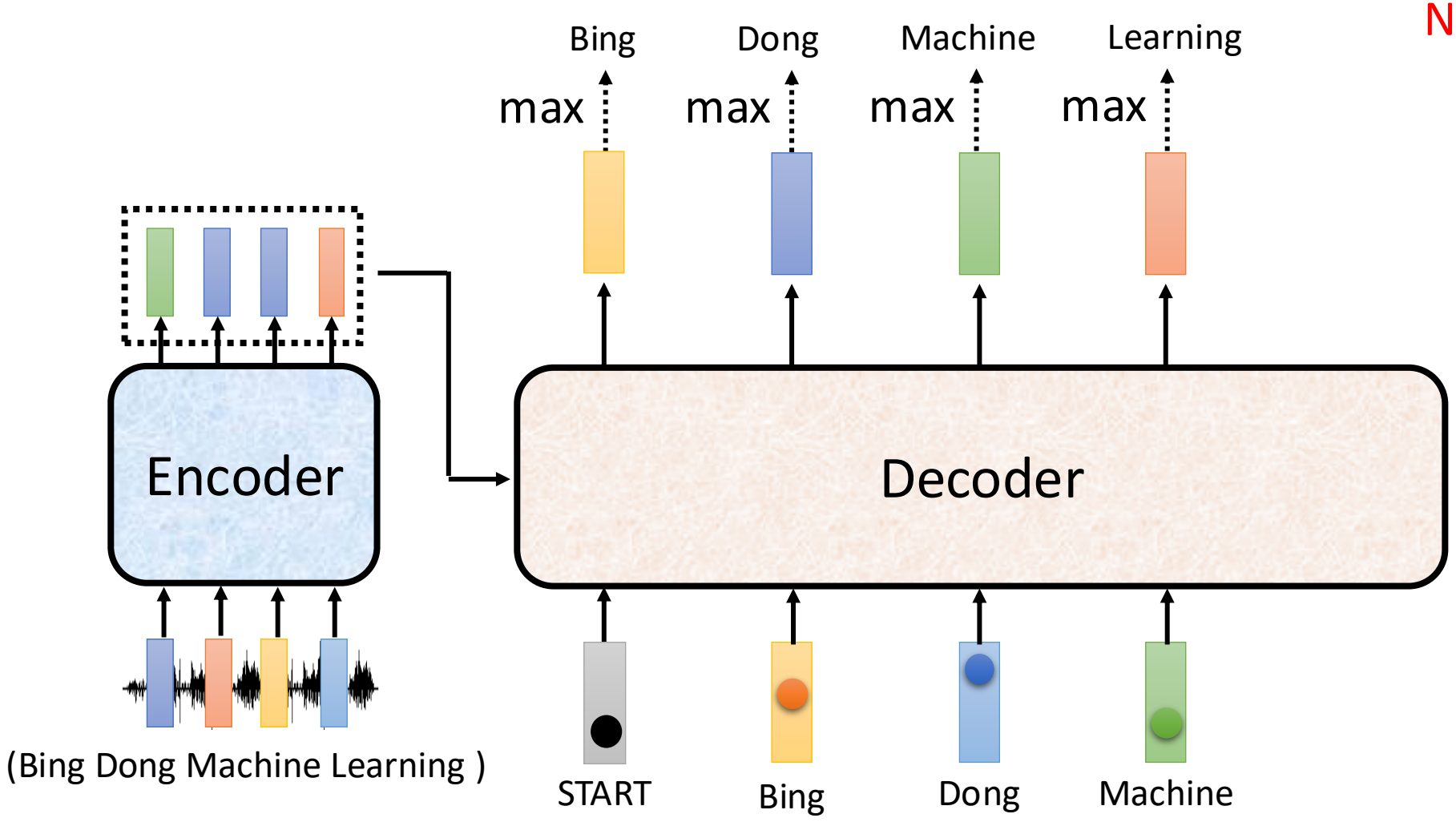
(2 heads as example)


TONG
Bing
Dong
Machine
Learning
max
Encoder
Decoder
(Bing Dong Machine Learning)
START
Bing
Dong
TONG
Machine
Decoder
Output

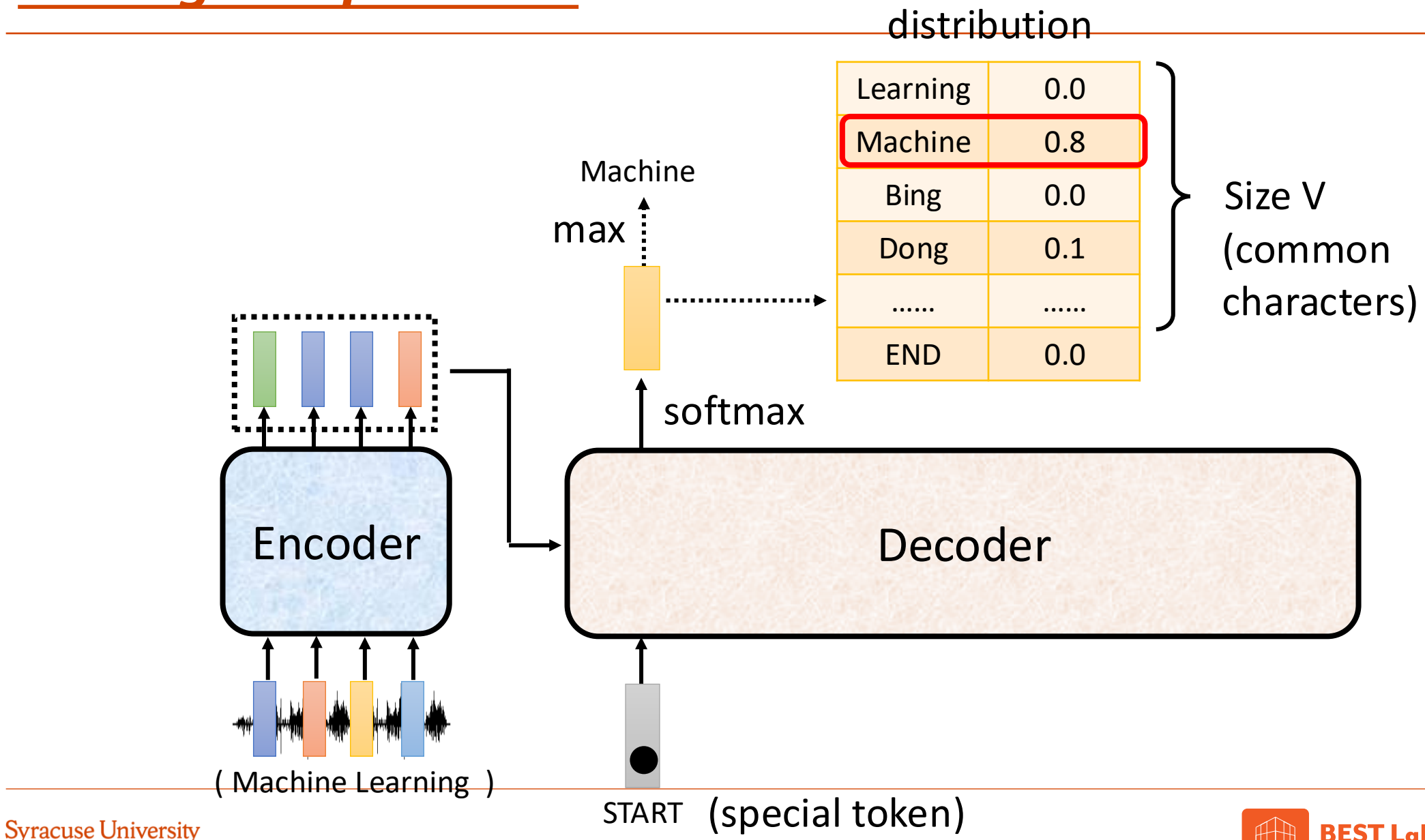
Autoregressive

We do not know the correct output length.

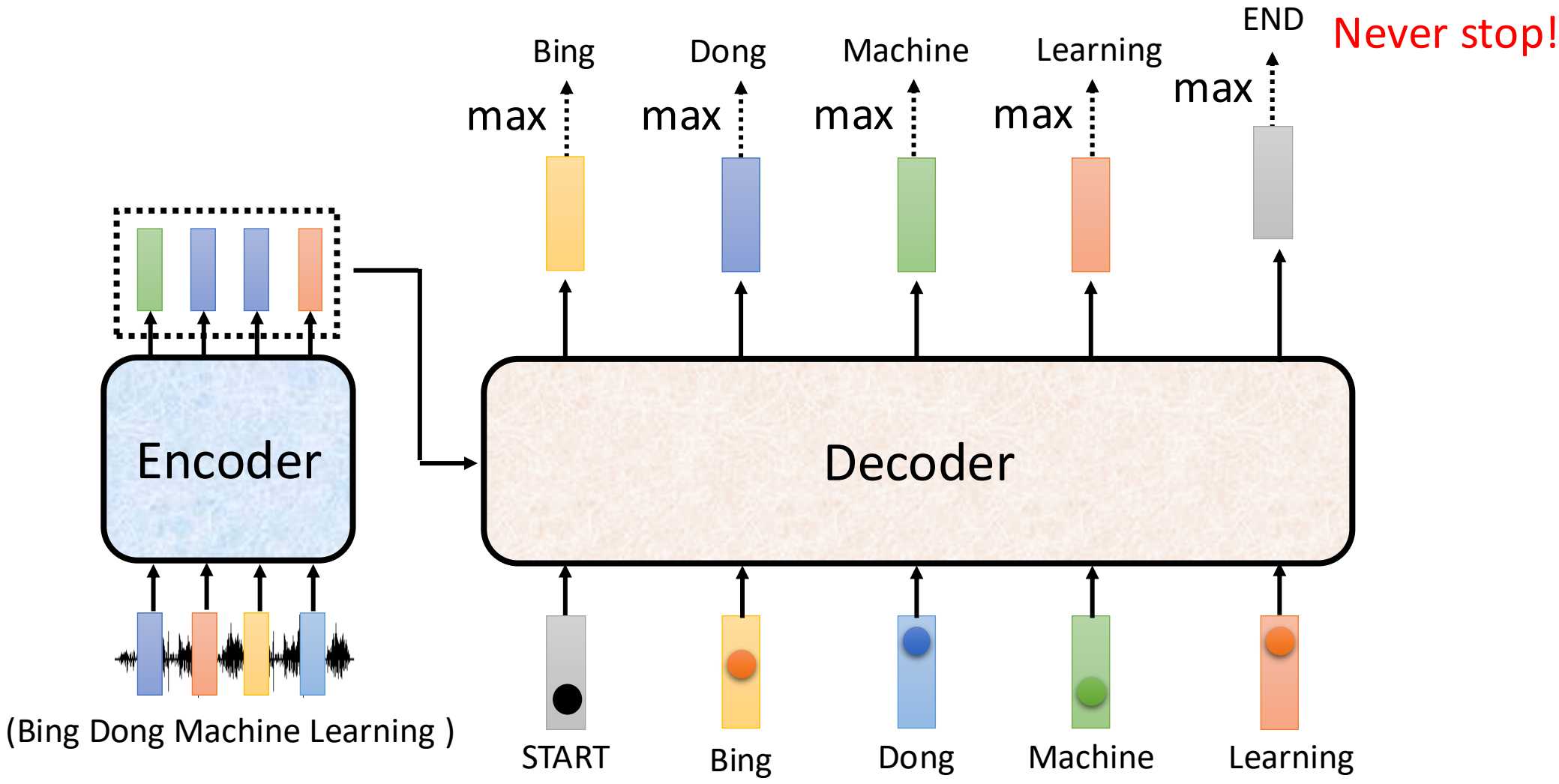
Never stop!



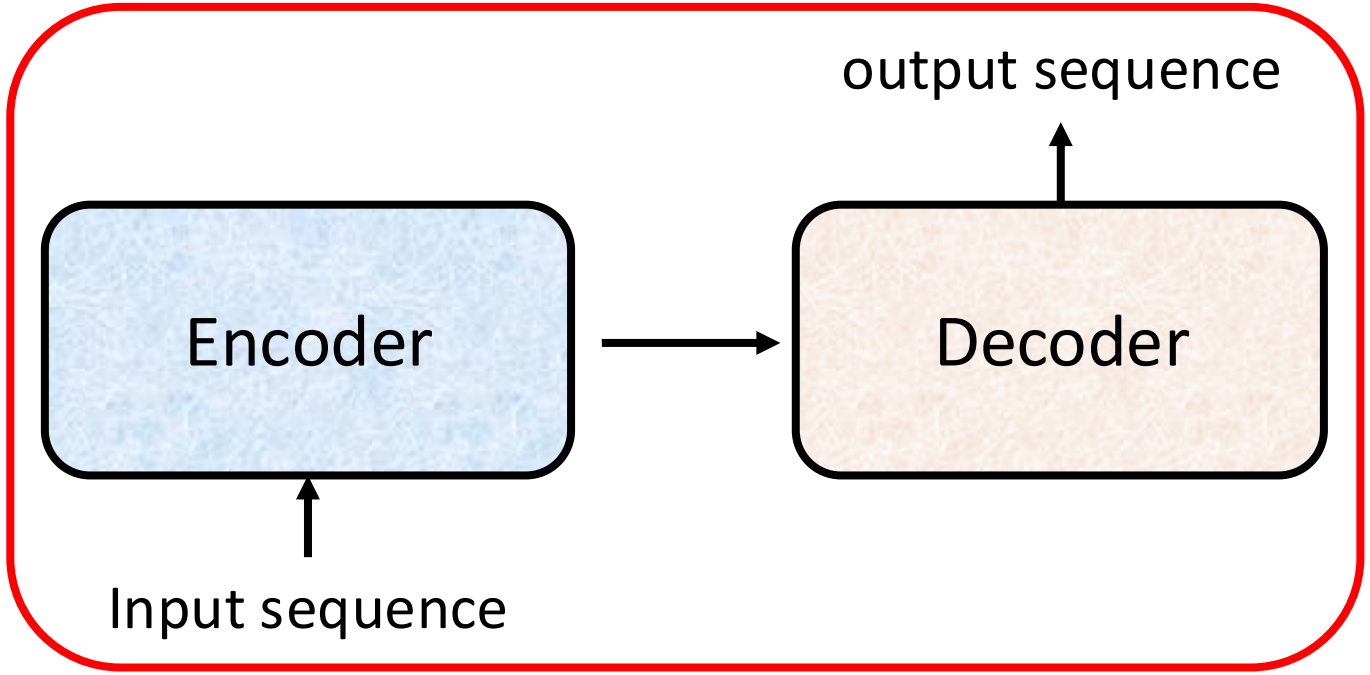
Adding "Stop Token"



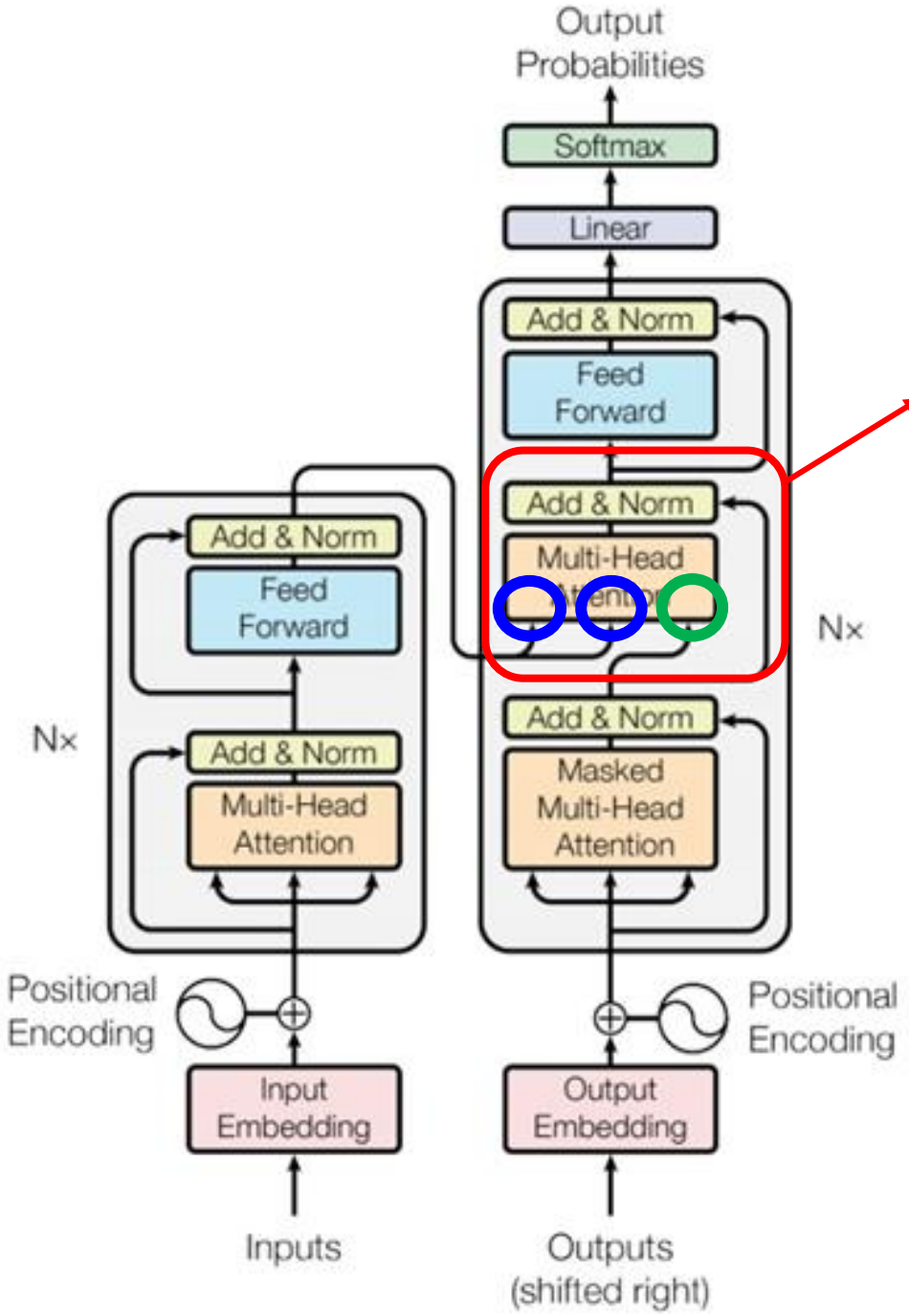
Autoregressive



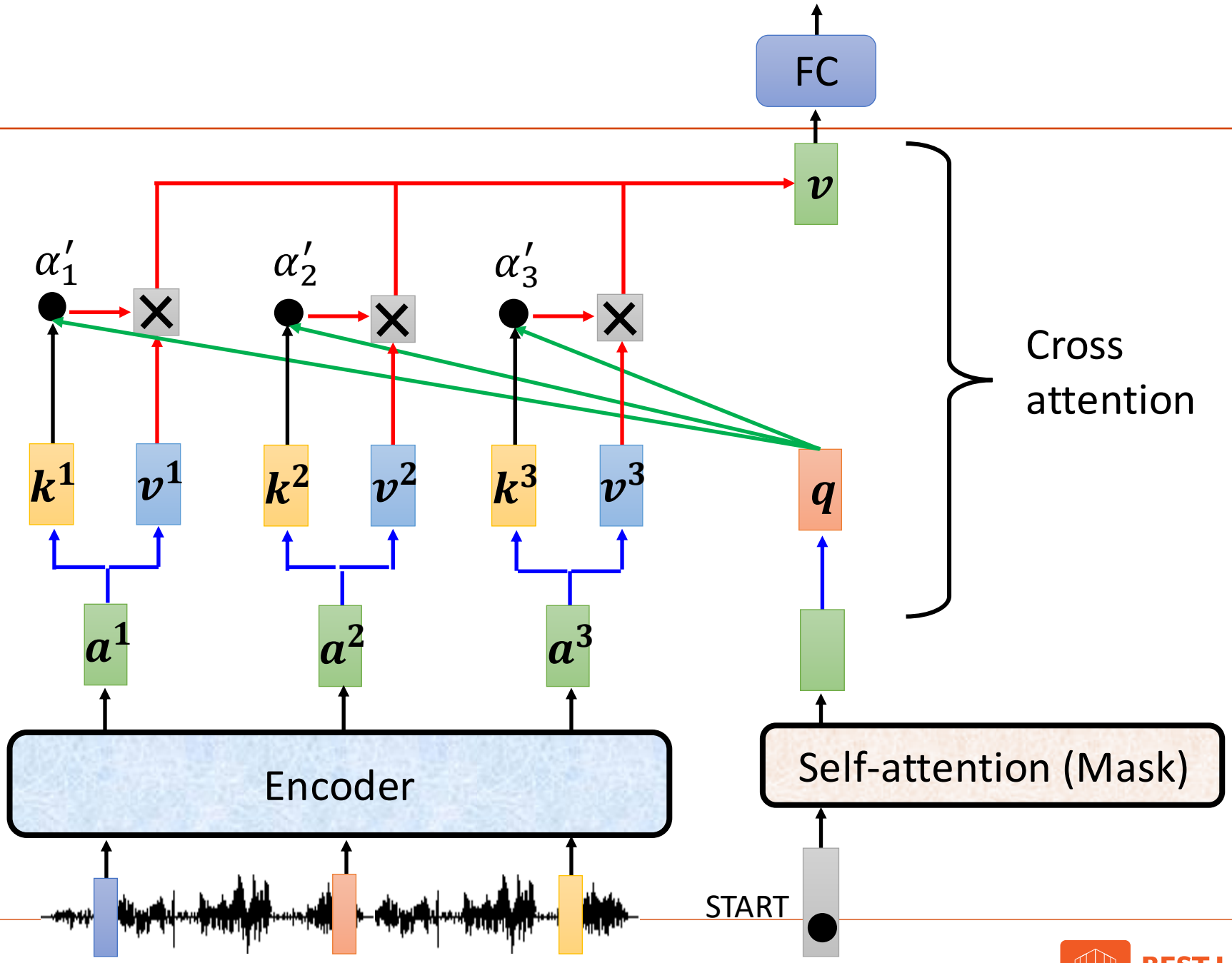
Encoder-Decoder

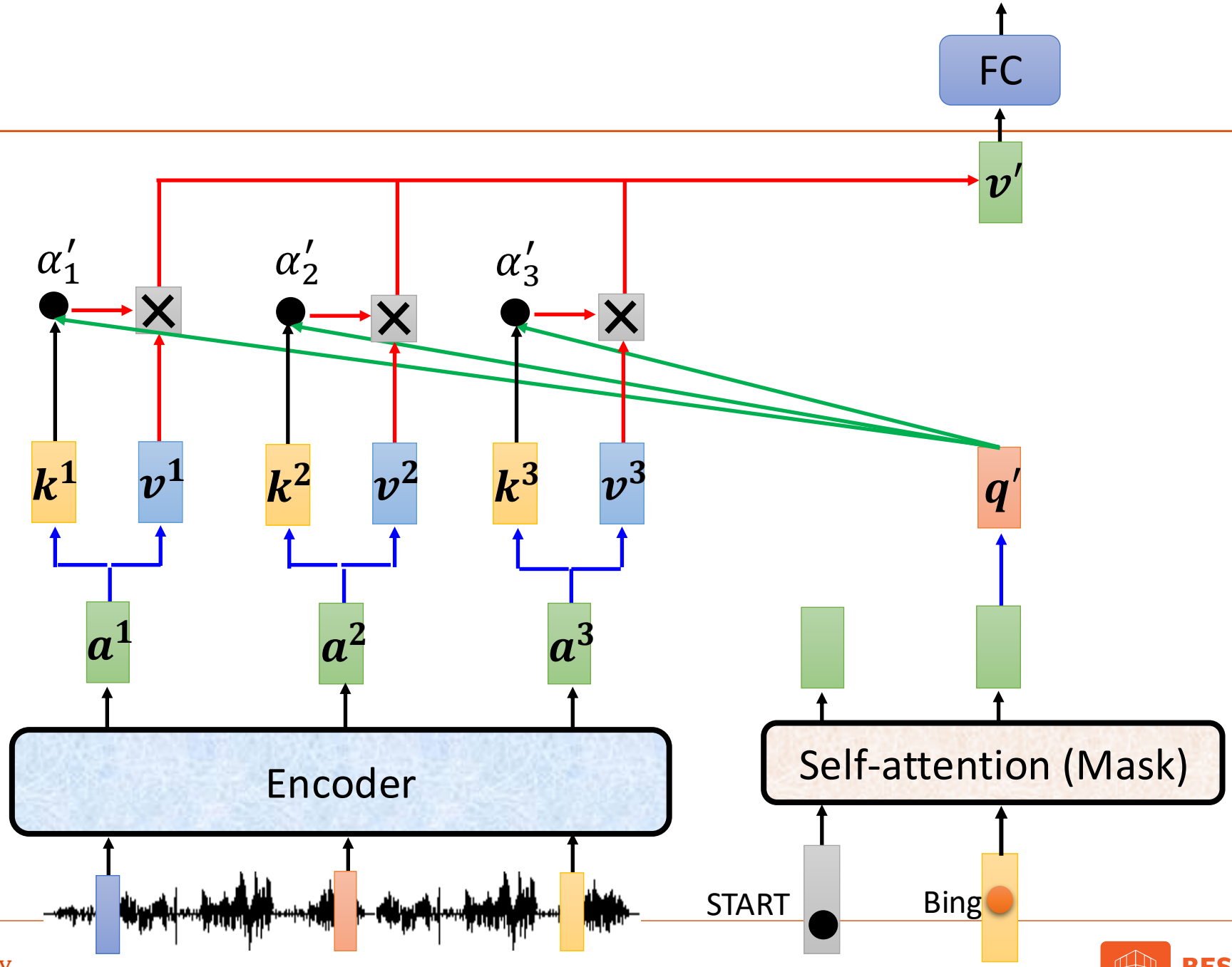


Transformer



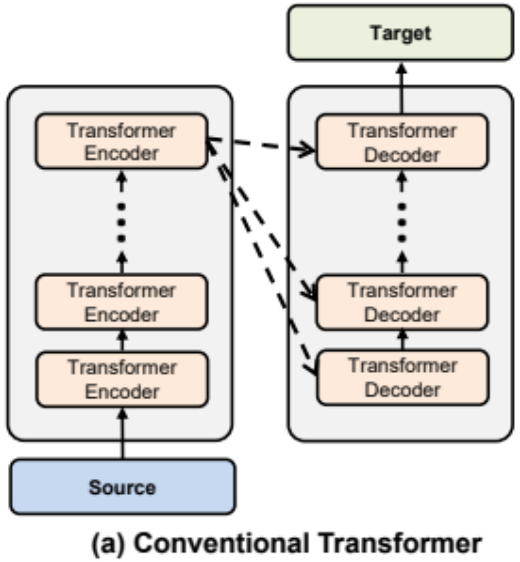
Cross attention

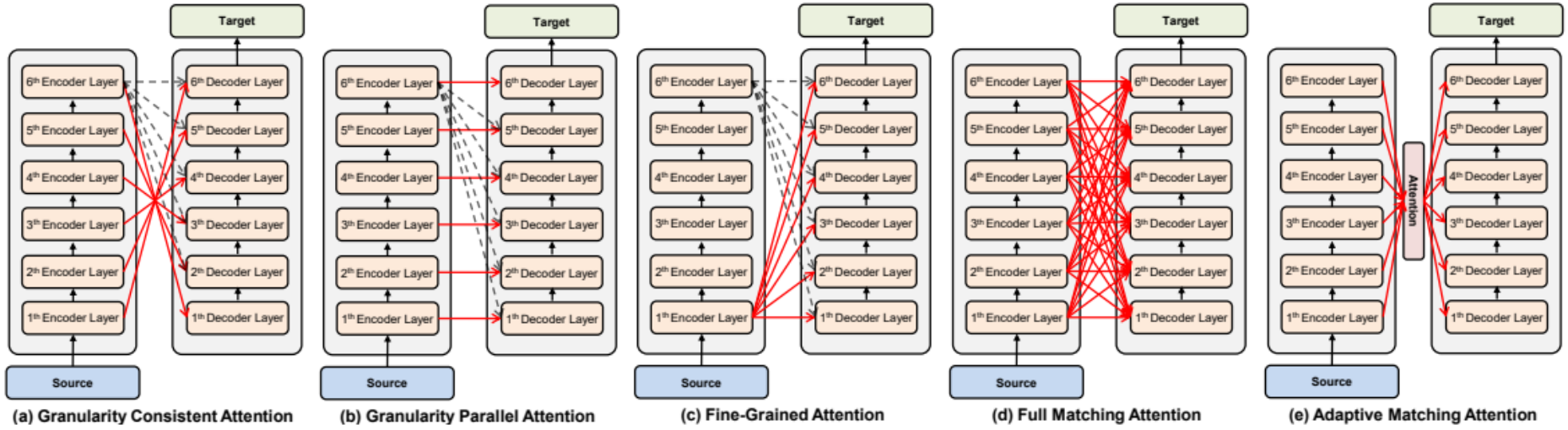

FC
v
α'_1
α'_2
α'_3
k^1
v^1
k^2
v^2
k^3
v^3
q
Cross attention
a^1
a^2
a^3
Encoder
Self-attention (Mask)
START


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
Bing

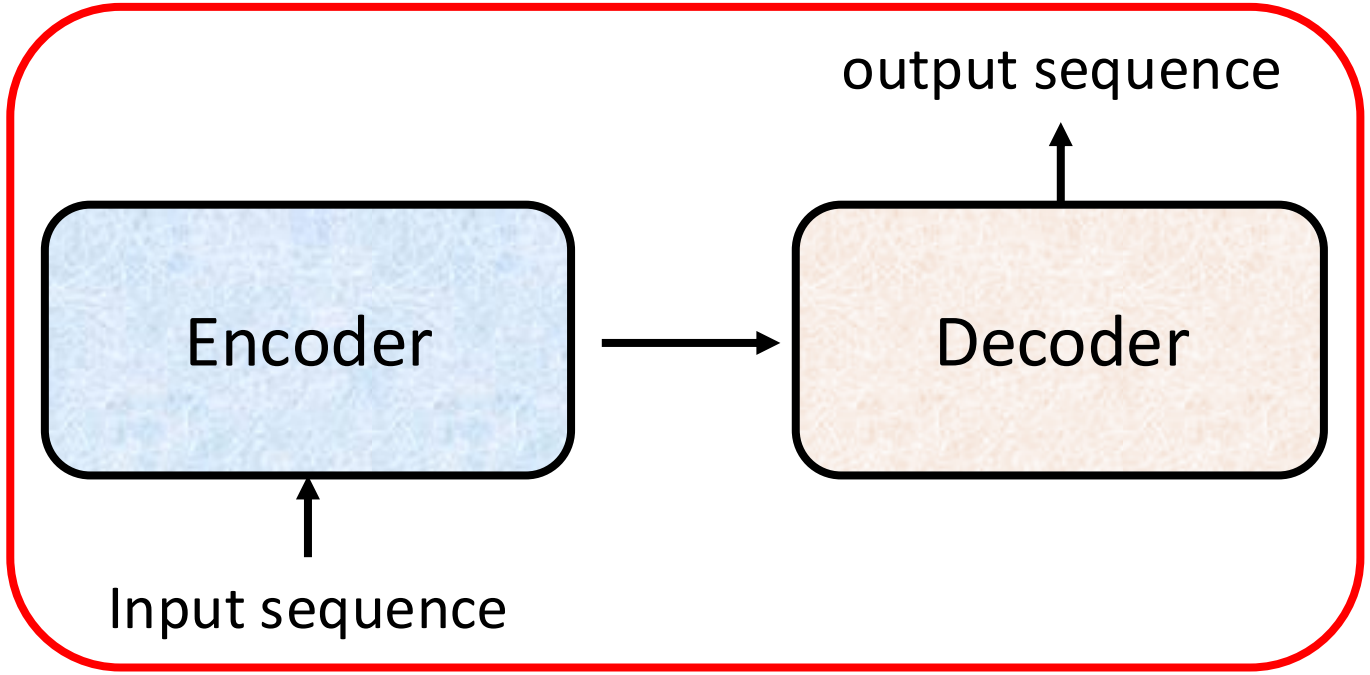
Cross Attention

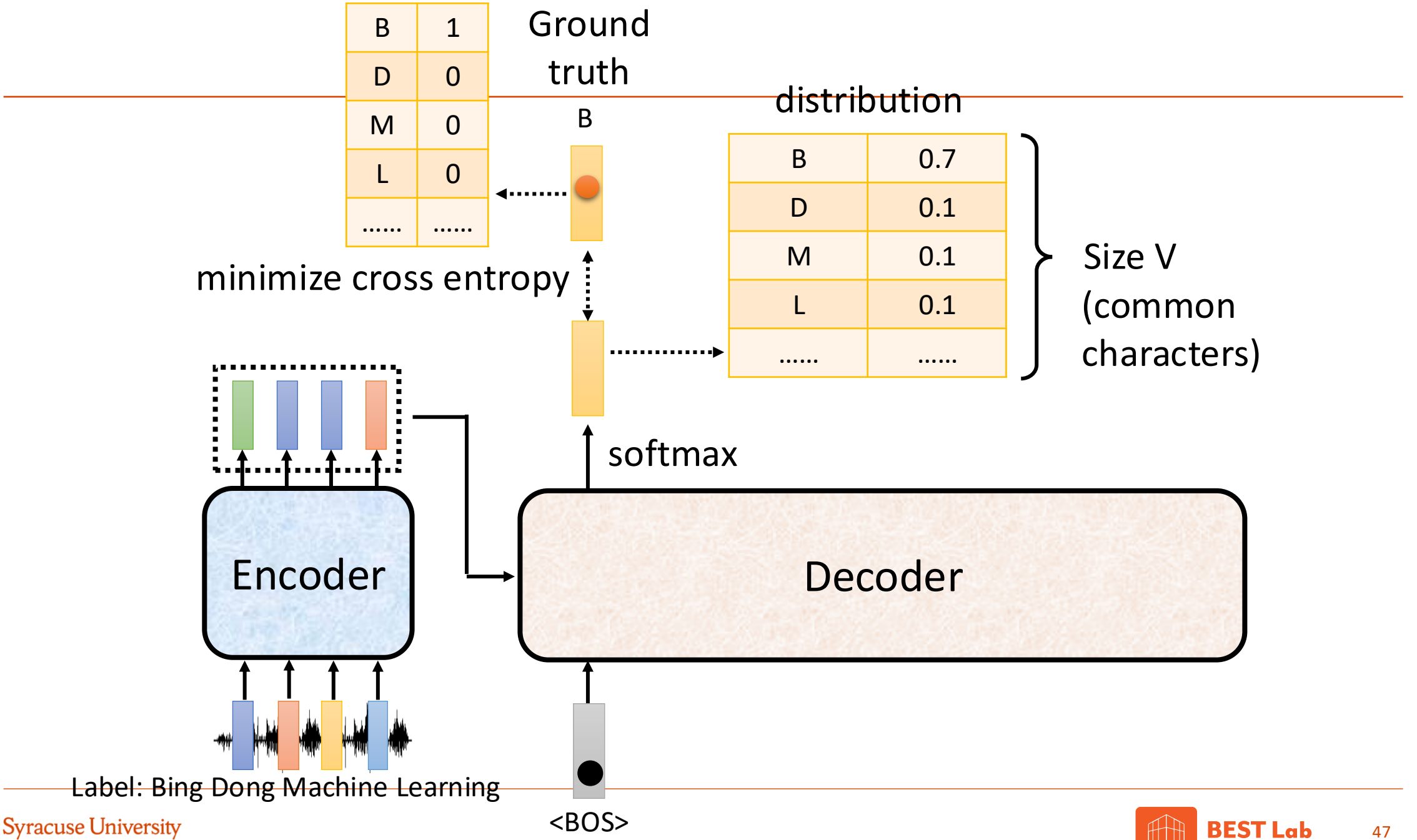
Source of image:
https://arxiv.org/abs/2005.08081



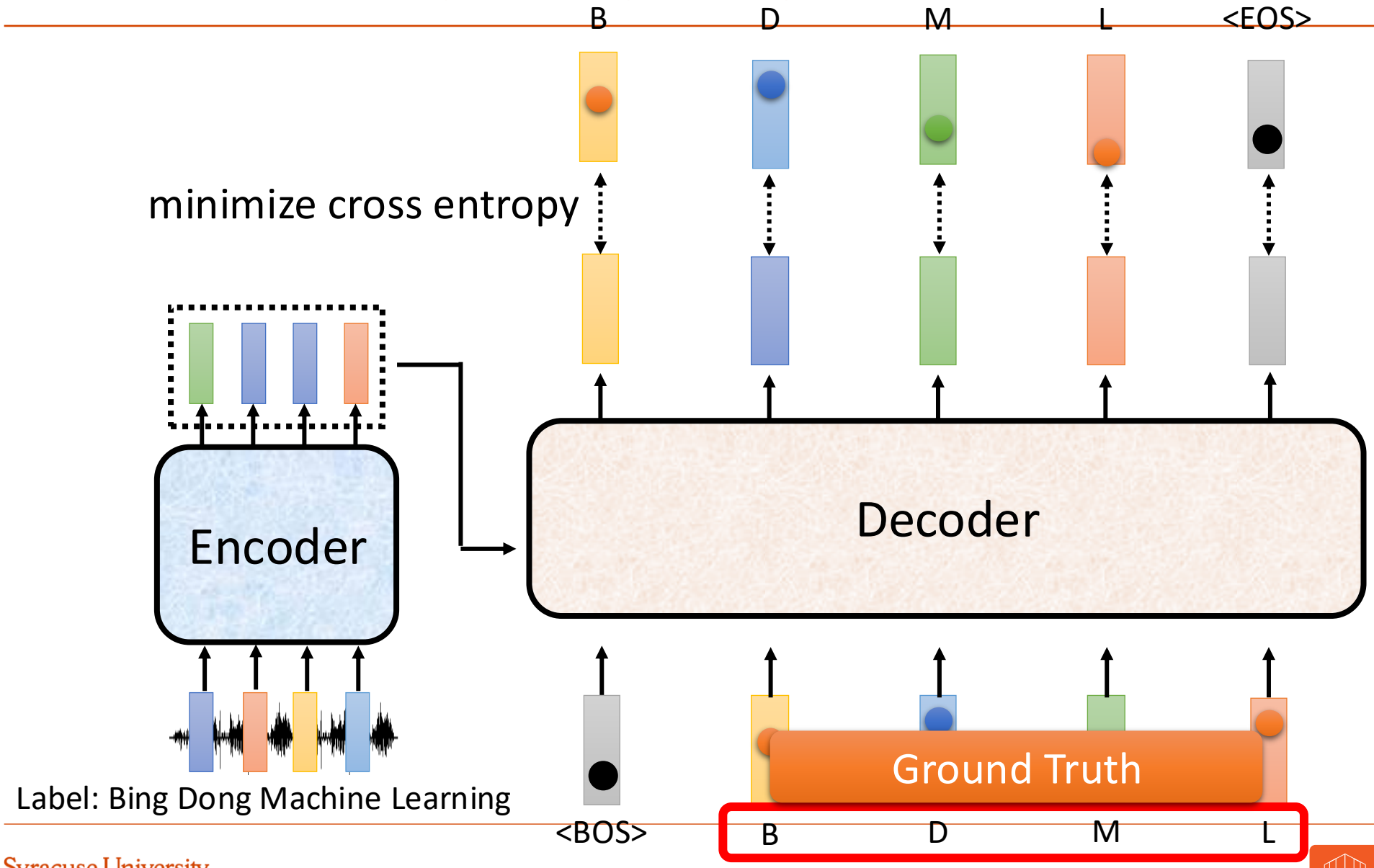


Training



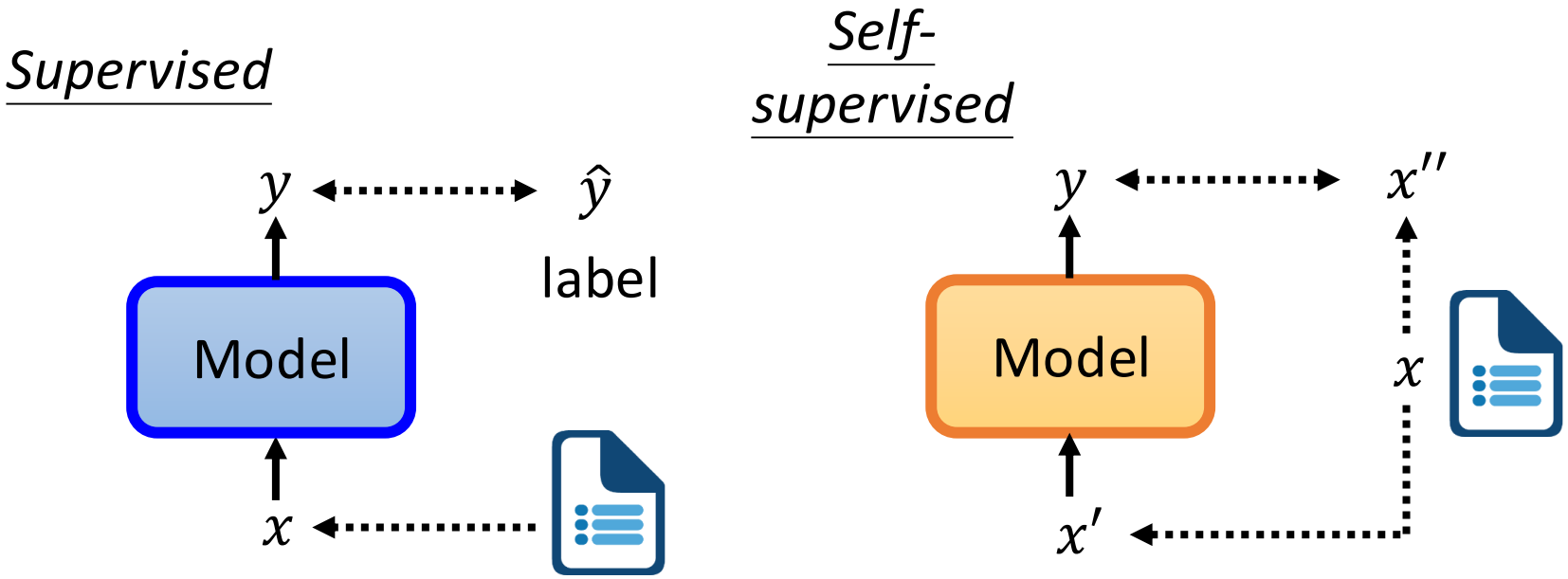

B 1
D 0
M 0
L 0
....
Ground truth
B
distribution
B 0.7
D 0.1
M 0.1
L 0.1
....
Size V (common characters)
minimize cross entropy
softmax
Encoder
Decoder
Label: Bing Dong Machine Learning
<BOS>

Teacher Forcing: using the ground truth as input.

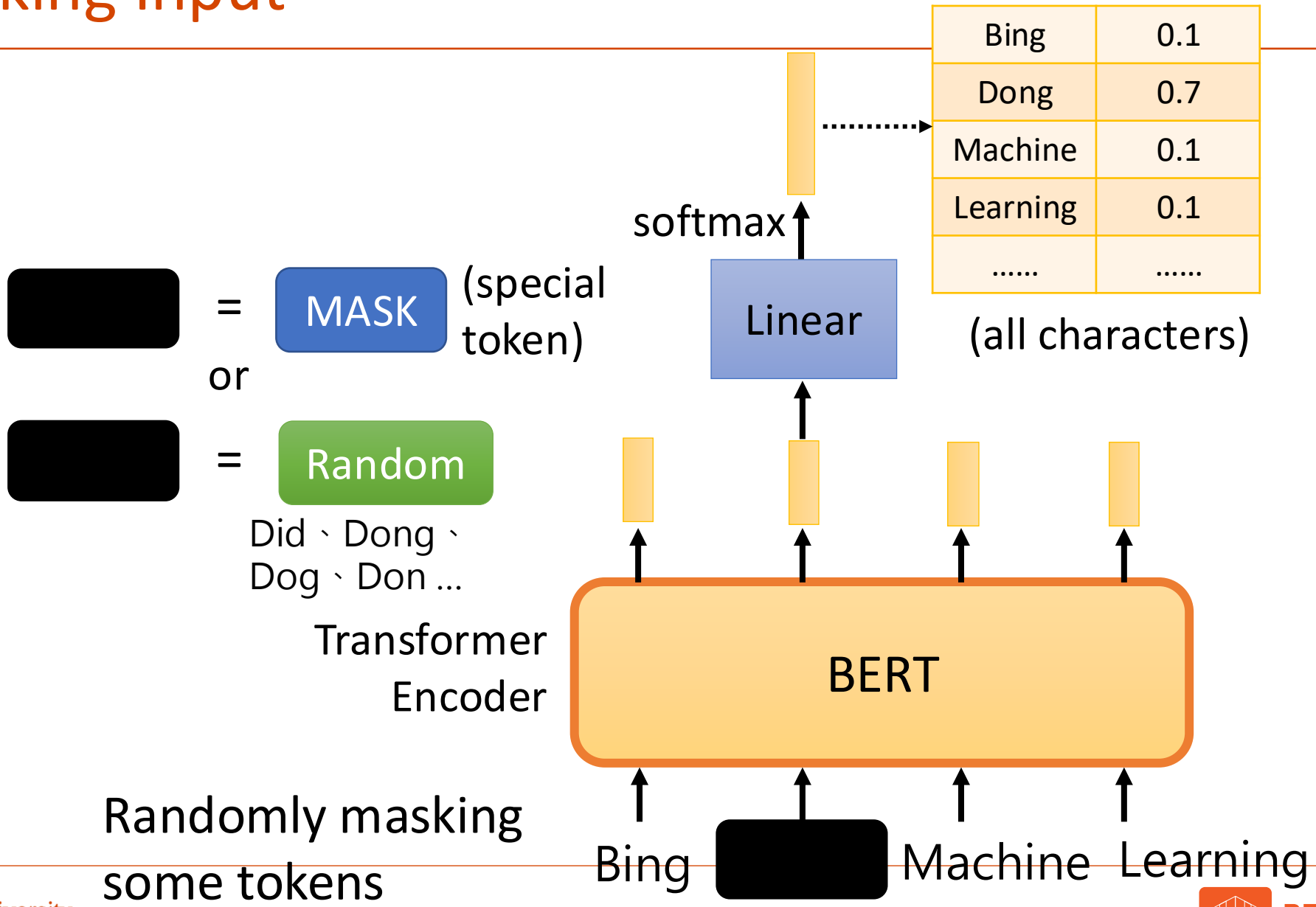


Large Language Model----A Self-Supervised Learned Transformer

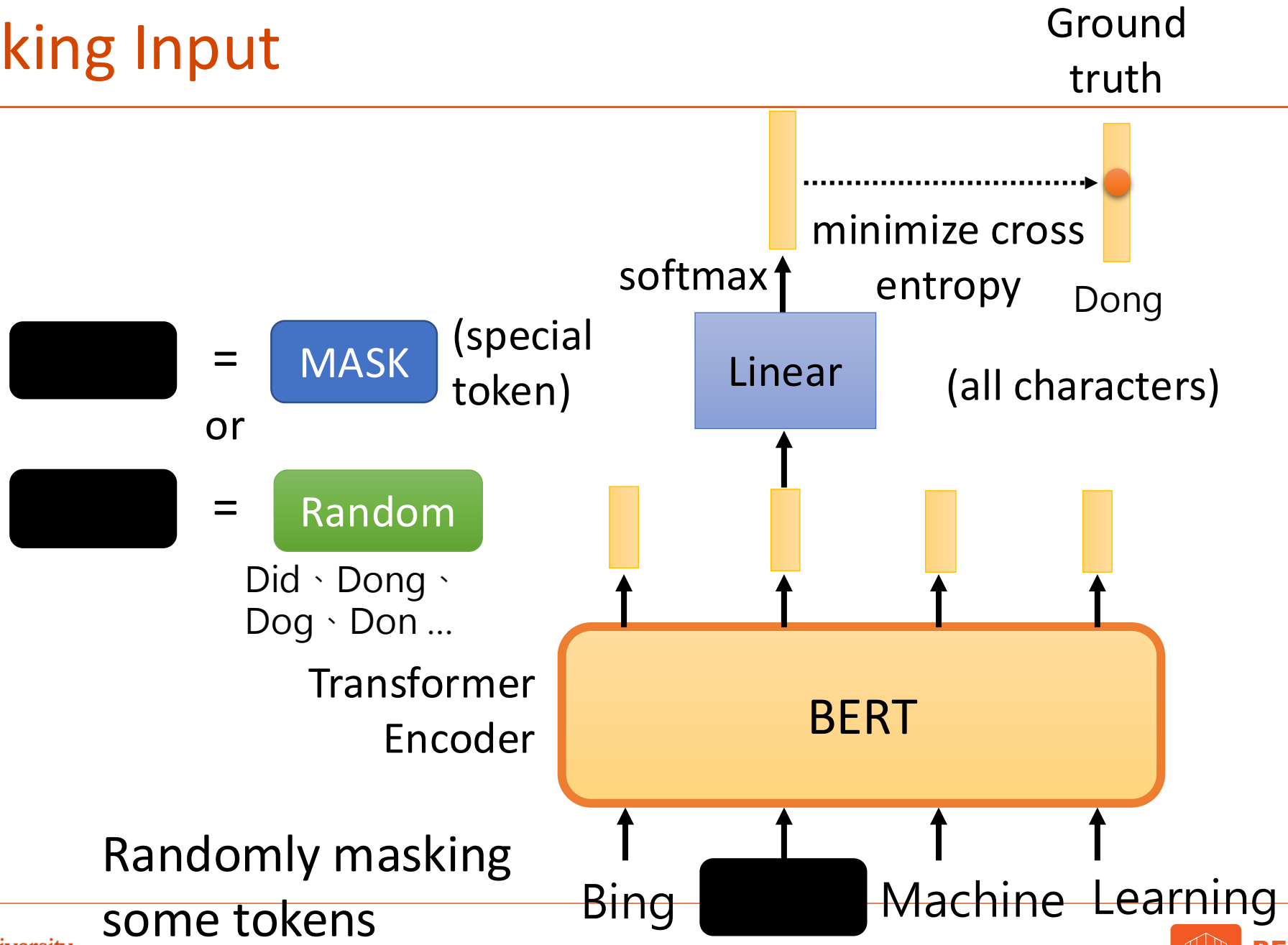
Self-Supervised Learning

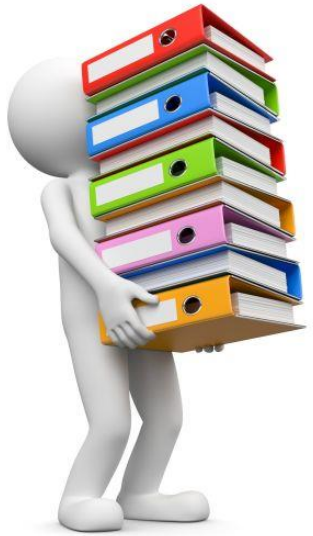


Masking Input



Masking Input





Self-supervised Learning

Pre-train

BERT

- Masked token prediction
- Next sentence prediction

Fine-tune

Model for Task 1

Model for Task 2

Model for Task 3

Downstream Tasks

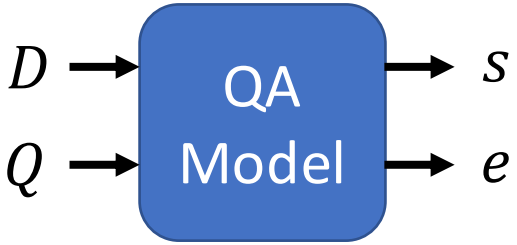
- The tasks we care
- We have a little bit labeled data.

How to use BERT

- Extraction-based Question Answering (QA)

Document: $D = \{d_1, d_2, \cdots, d_N\}$

Query: $Q = \{q_1, q_2, \cdots, q_M\}$



output: two integers (s, e)

Answer: $A = \{d_s, \cdots, d_e\}$

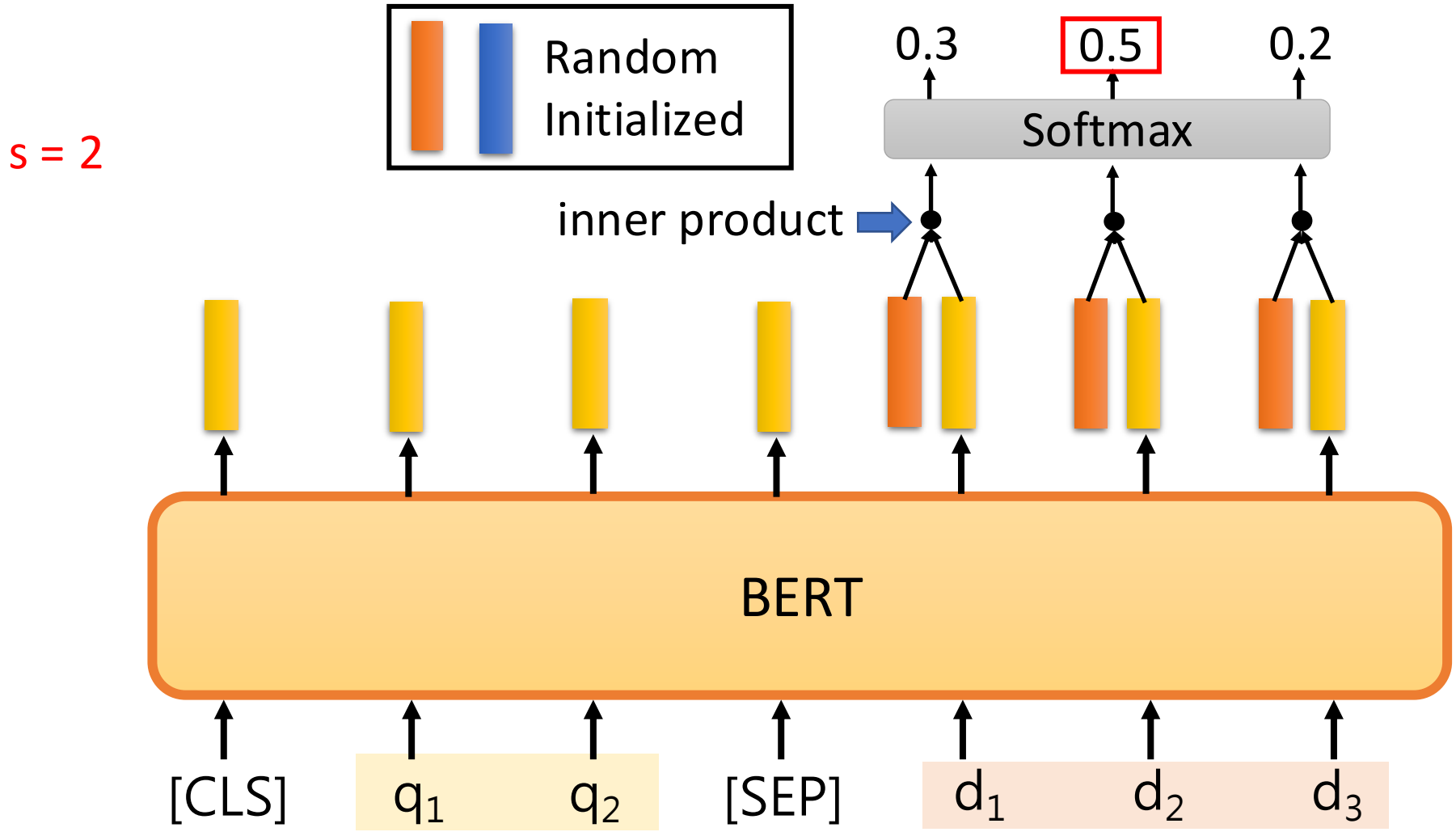
In meteorology, precipitation is any product of the condensation of 17 spheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain 77 atte 79 cations are called "showers".

What causes precipitation to fall?
gravity $s = 17, e = 17$

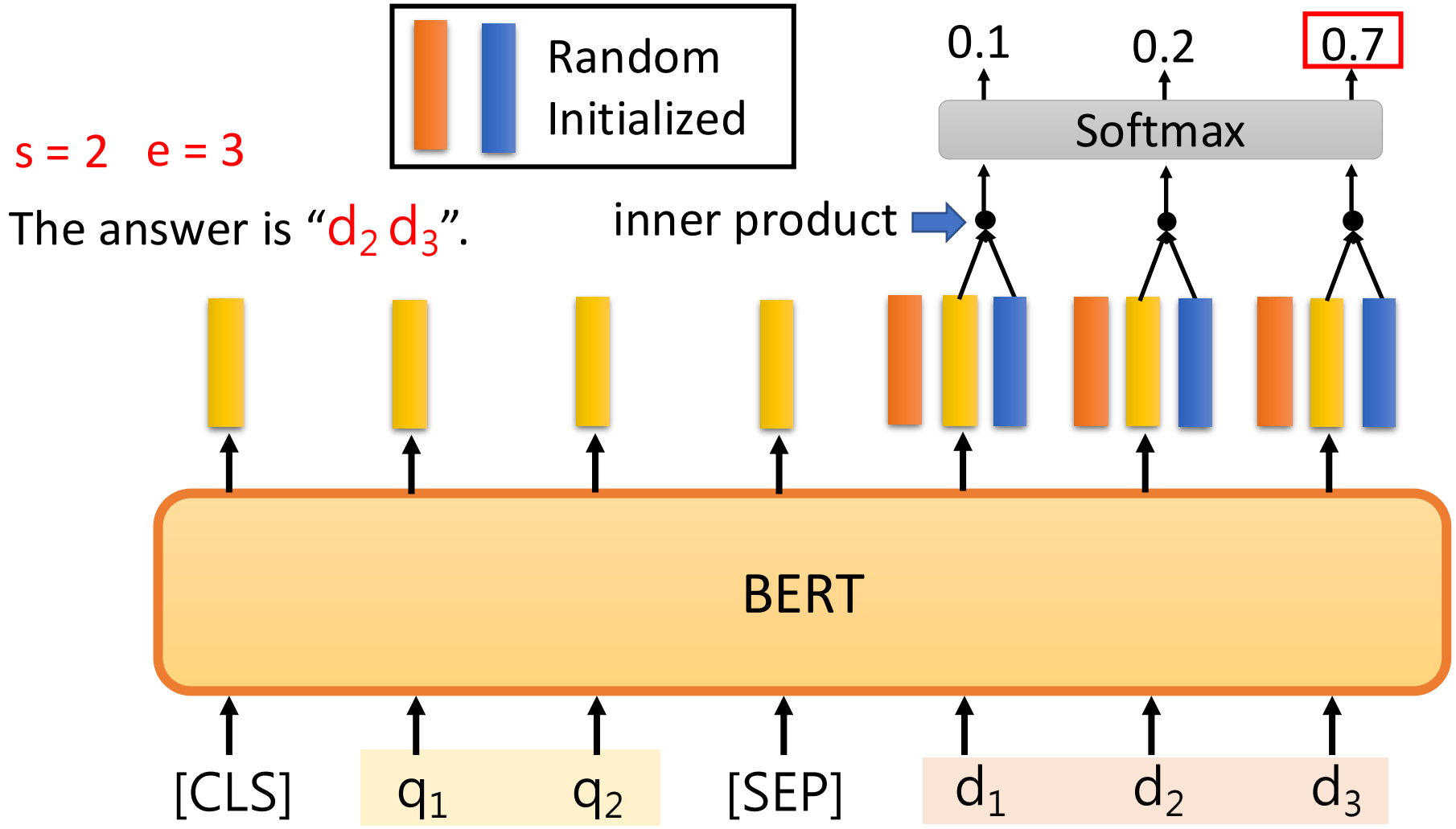
What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud $s = 77, e = 79$

How to use BERT



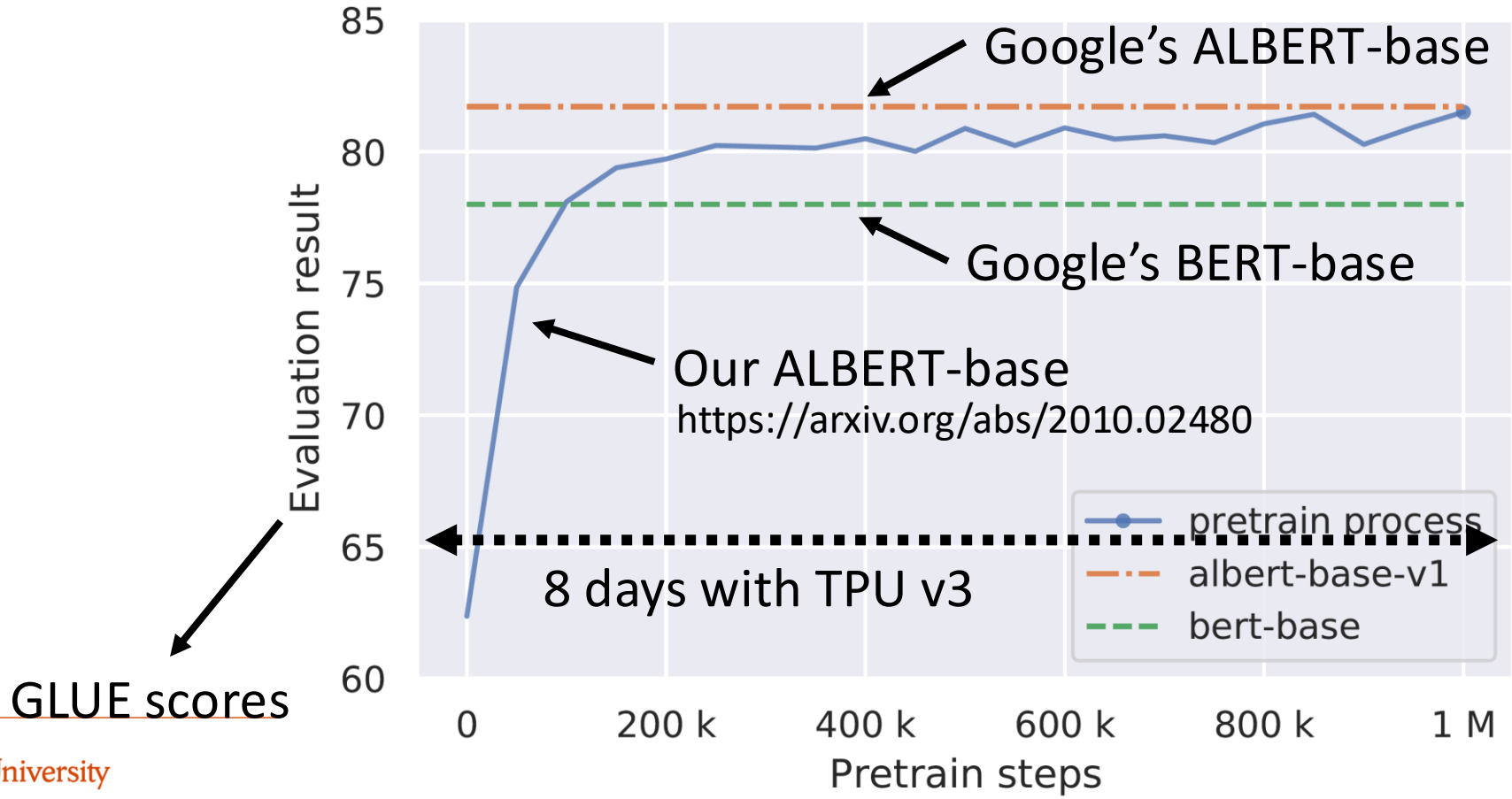
How to use BERT



Training BERT is challenging!

Training data has more than **3 billions** of words.

3000 times of **Harry Potter series**



Transformer

2017

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

Transformer

BERT 340M

GPT-2 (1.5B)

GPT-3 (175B)

GPT-4 (2T)

Any Questions?

bidong@syr.edu