Transformer

李宏毅

Hung-yi Lee

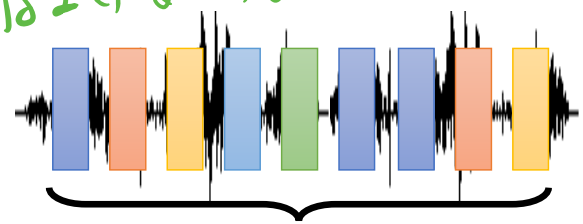
Sequence-to-sequence (Seq2seq)

transformer 就是 seq2seq model

Input a sequence, output a sequence

The output length is determined by model.

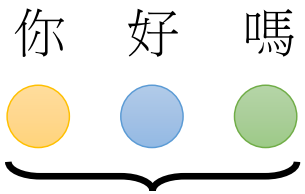
例子1 (語音的內容：你好嗎)



T 個向量

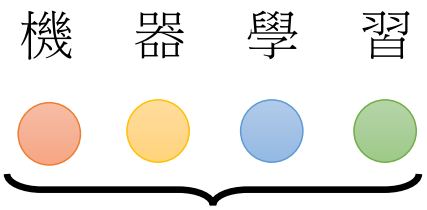
Speech Recognition

語音辨識



機器自己決定，輸出 → N 個向量

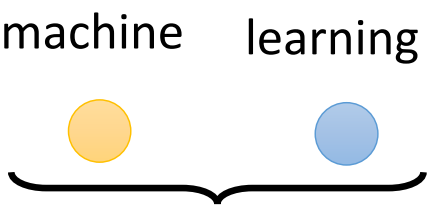
例子2



N

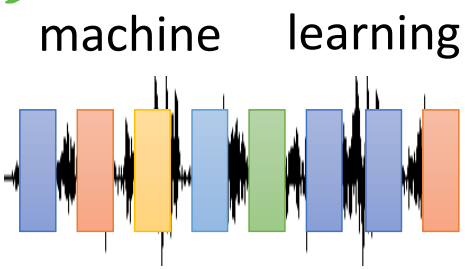
Machine Translation

機器翻譯



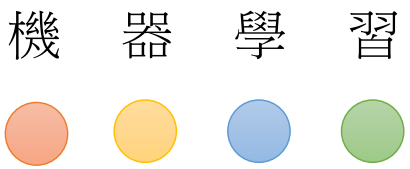
N'

例子3



語音翻譯

Speech Translation



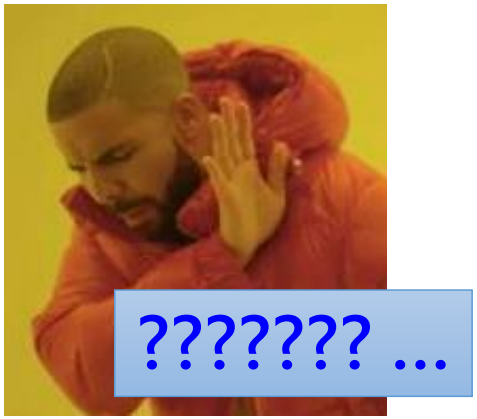
Language without text (e.g. 台語)

Hokkien (閩南語、台語)

(Speech of Hokkien)

Speech Recognition

“母湯”



(Speech of Hokkien)

Speech Translation

“不行”

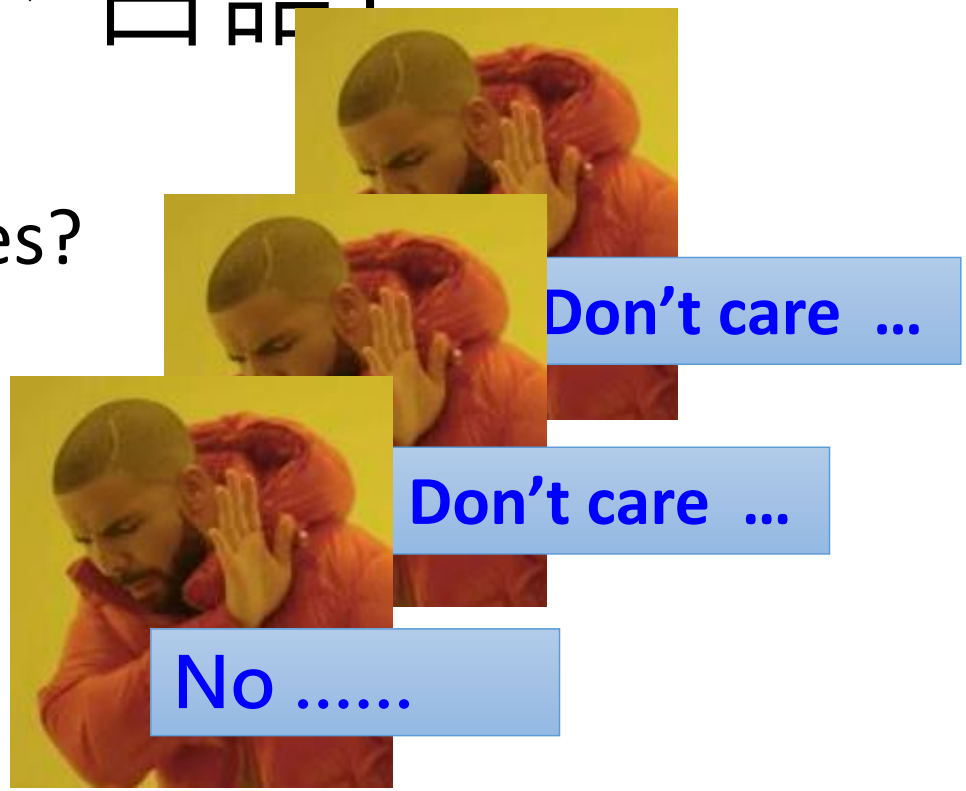
Local soap operas (鄉土劇) on YouTube
(Speech of Hokkien, Chinese subtitle)

Using 1500 hours of data for training

台語語音，中文字幕，太好了！印label好了！！

Hokkien (閩南語、台語)

- Background music & noises?
- Noisy transcriptions?
- Phonemes of Hokkien?



“硬train一發”
(Ying Train Yi Fa)

感謝 李仲翊 同學提供實驗結果

Hokkien (閩南語、台語)

成功 case

你 的 身 體 撐 不 住

沒 事 你 為 什 麼 要 請 假

失敗 case

要 生 了 嗎　Answer:不 會 膩 嗎

我 有 幫 廠 長 拜 託

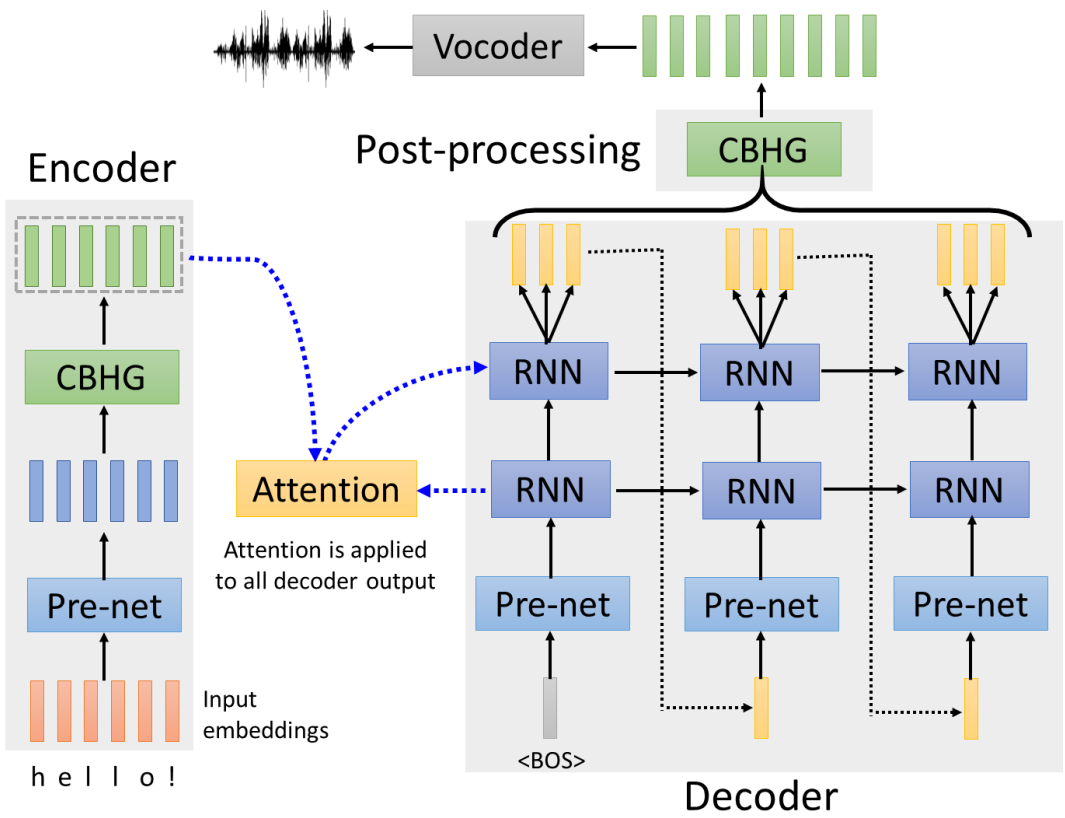
Answer: 我 拜 託 廠 長 了

To learn more: https://sites.google.com/speech.ntut.edu.tw/fsw/home/challenge-2020

Text-to-Speech (TTS) Synthesis

(輸入文字，輸出語音)
⇒ 語音合成

感謝張凱為同學提供實驗結果



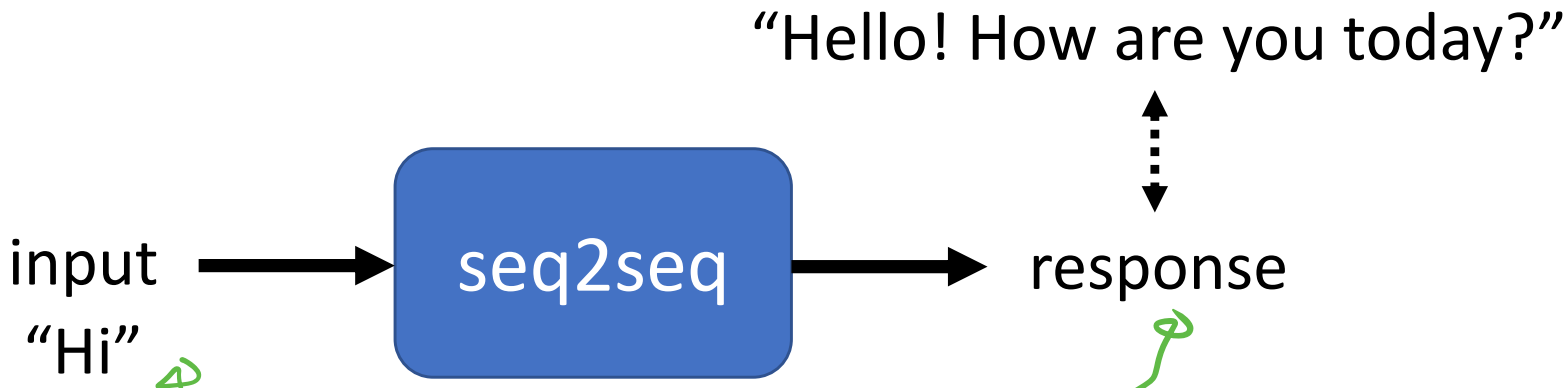
Taiwanese Speech Synthesis

Source of data: 台灣媠聲2.0

歡迎來到台大語音處理實驗室

最近肺炎真嚴重，要記得戴口罩、勤洗手，有病就要看醫生

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 ...

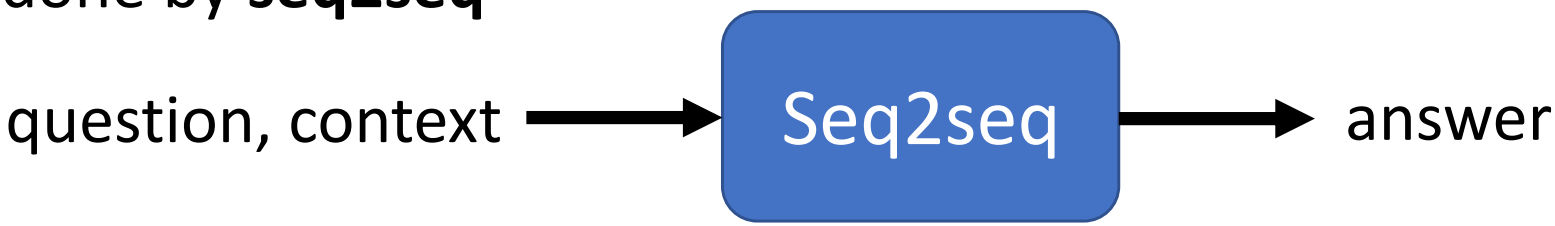
① 很多 NLP 的 task 都可以 formulate 成 ⑦ 的任務

Question Answering (QA)

③ Question	② Context (input)	④ Answer (output)
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
⑤ What is the translation from English to German?	Most of the planet is ocean water.	Der Großteil der Erde ist Meerwasser
⑥ 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
② 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

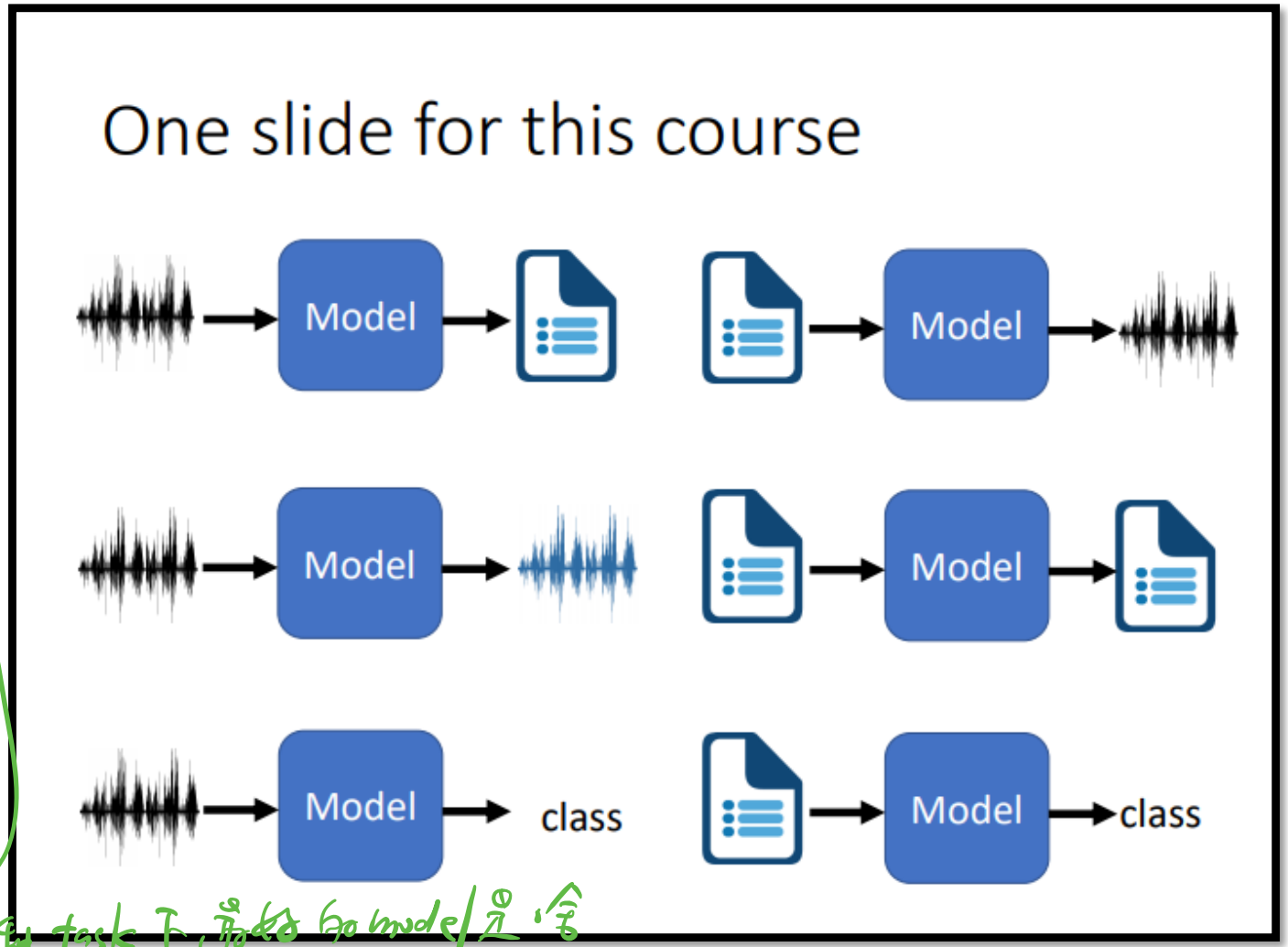
decaNLP

QA can be done by **seq2seq**



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

Deep Learning for Human Language Processing
深度學習與人類語言處理

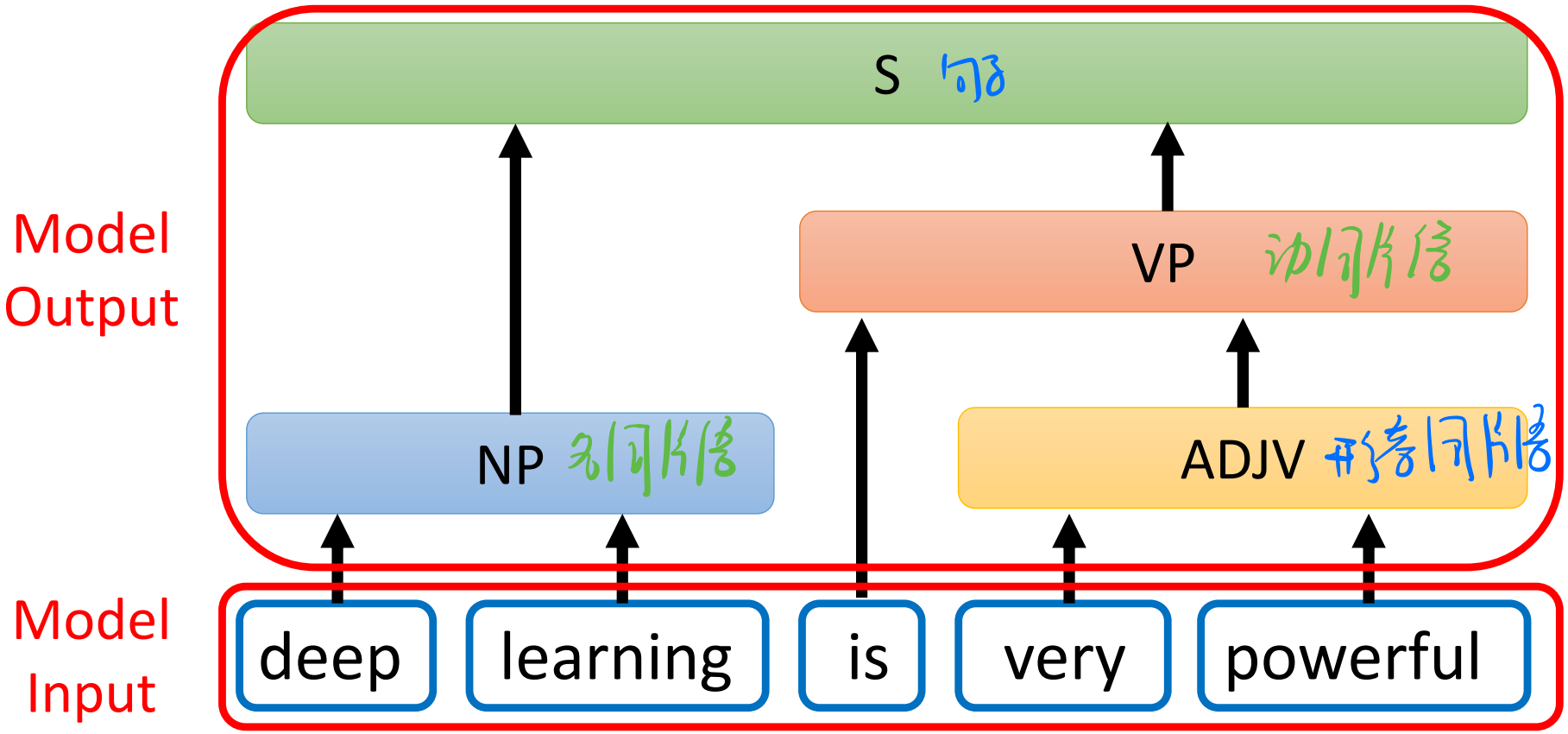


Source webpage: https://speech.ee.ntu.edu.tw/~hylee/dlhlp/2020-spring.html

Seq2seq for Syntactic Parsing

文法剖析 ⇒ 用 seq2seq 硬解

Is it a sequence?

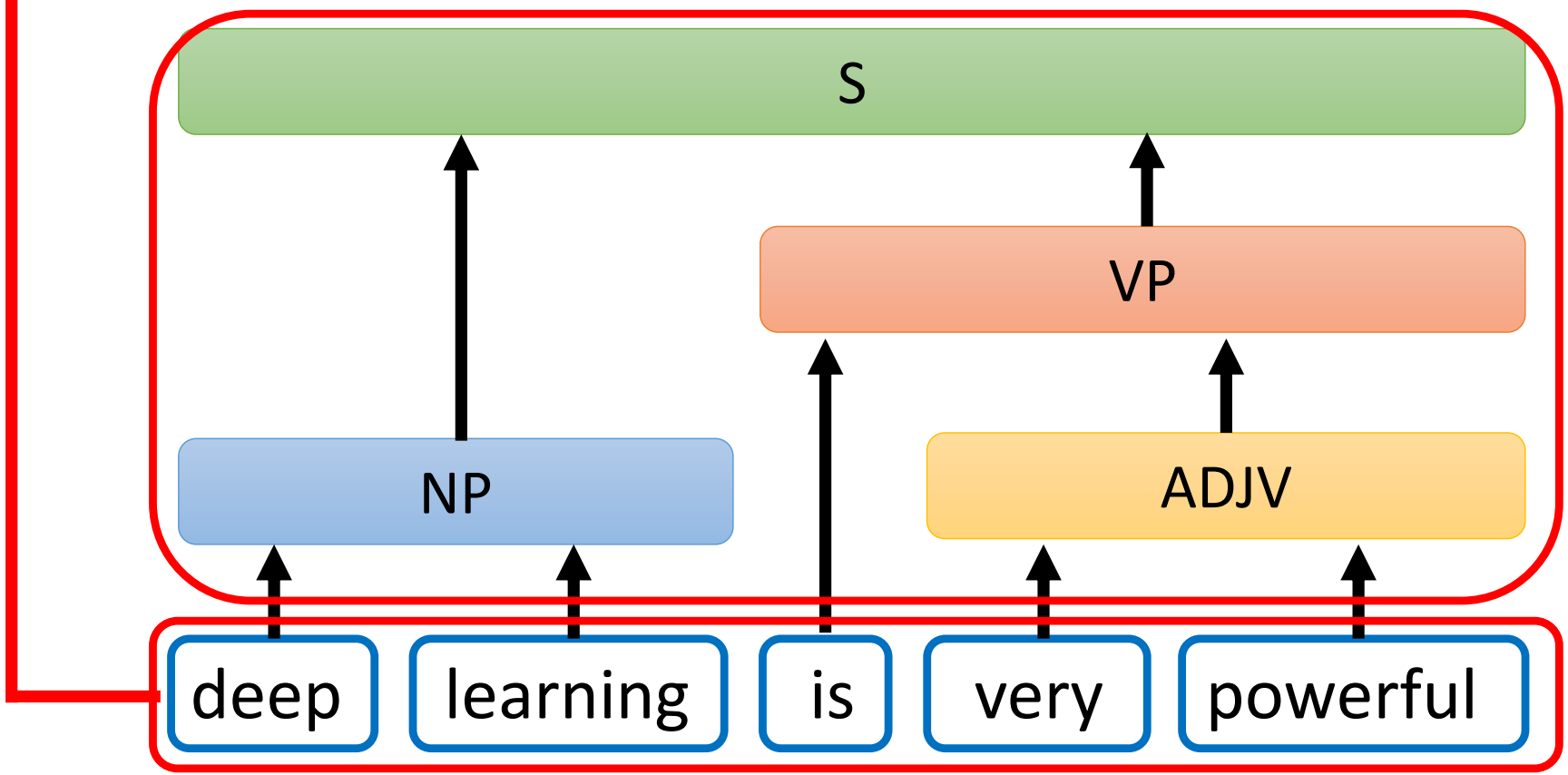


Seq2seq for Syntactic Parsing

↓ parsing 的 tree

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

Seq2seq!



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.

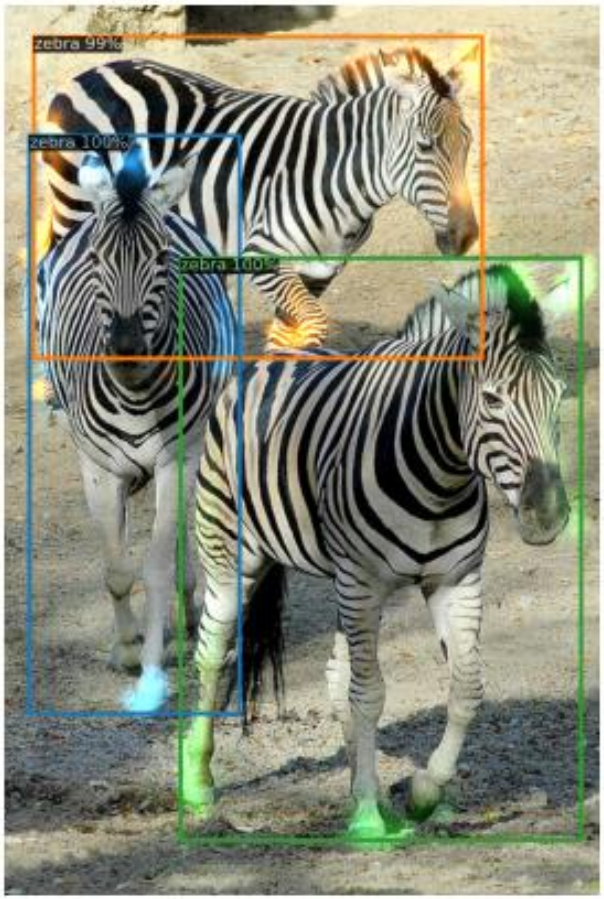
輸入一篇文章

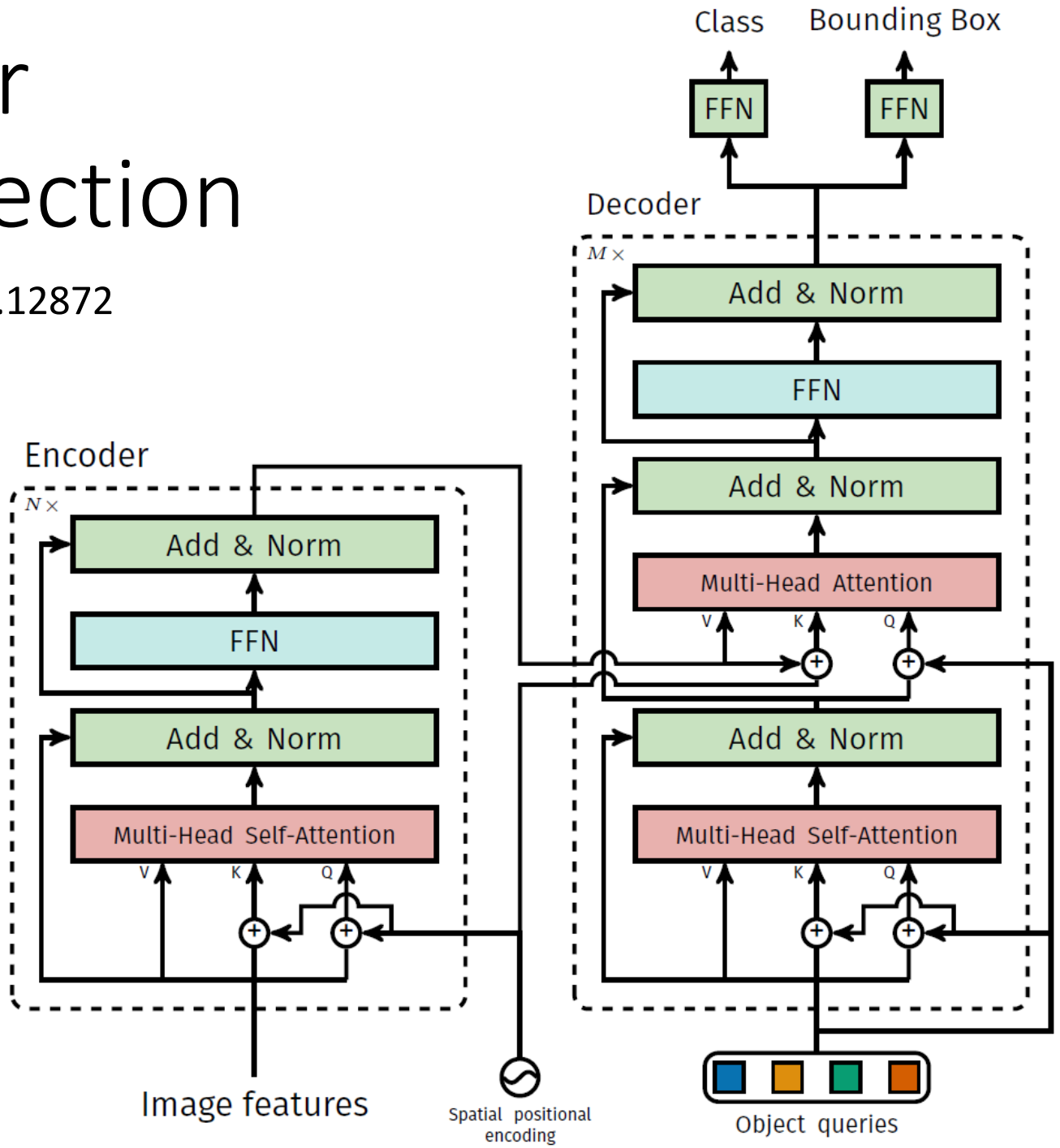
https://arxiv.org/abs/1909.03434
https://arxiv.org/abs/1707.05495

model 自己決定要output幾個class

Seq2seq for Object Detection

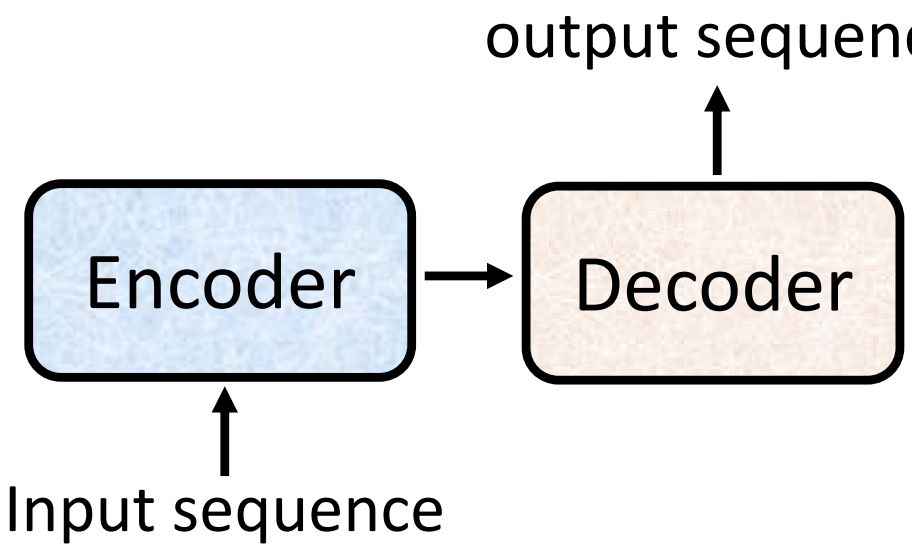
https://arxiv.org/abs/2005.12872

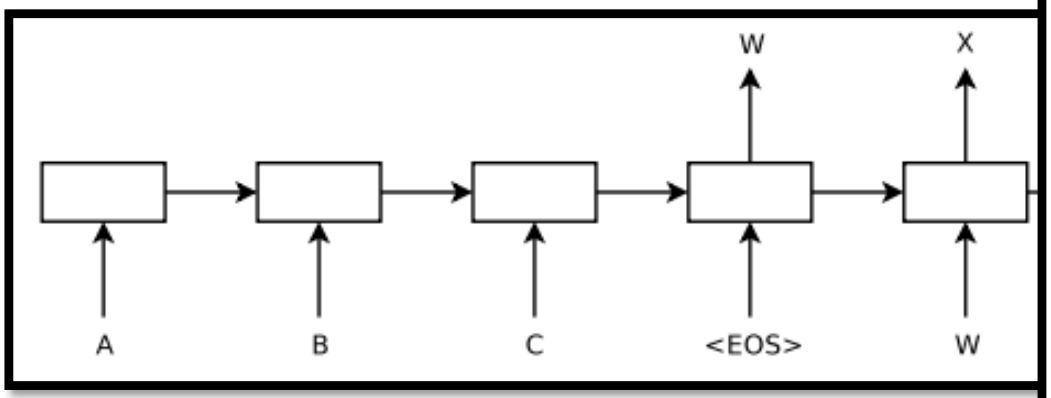




問格導如何做 seq2seq

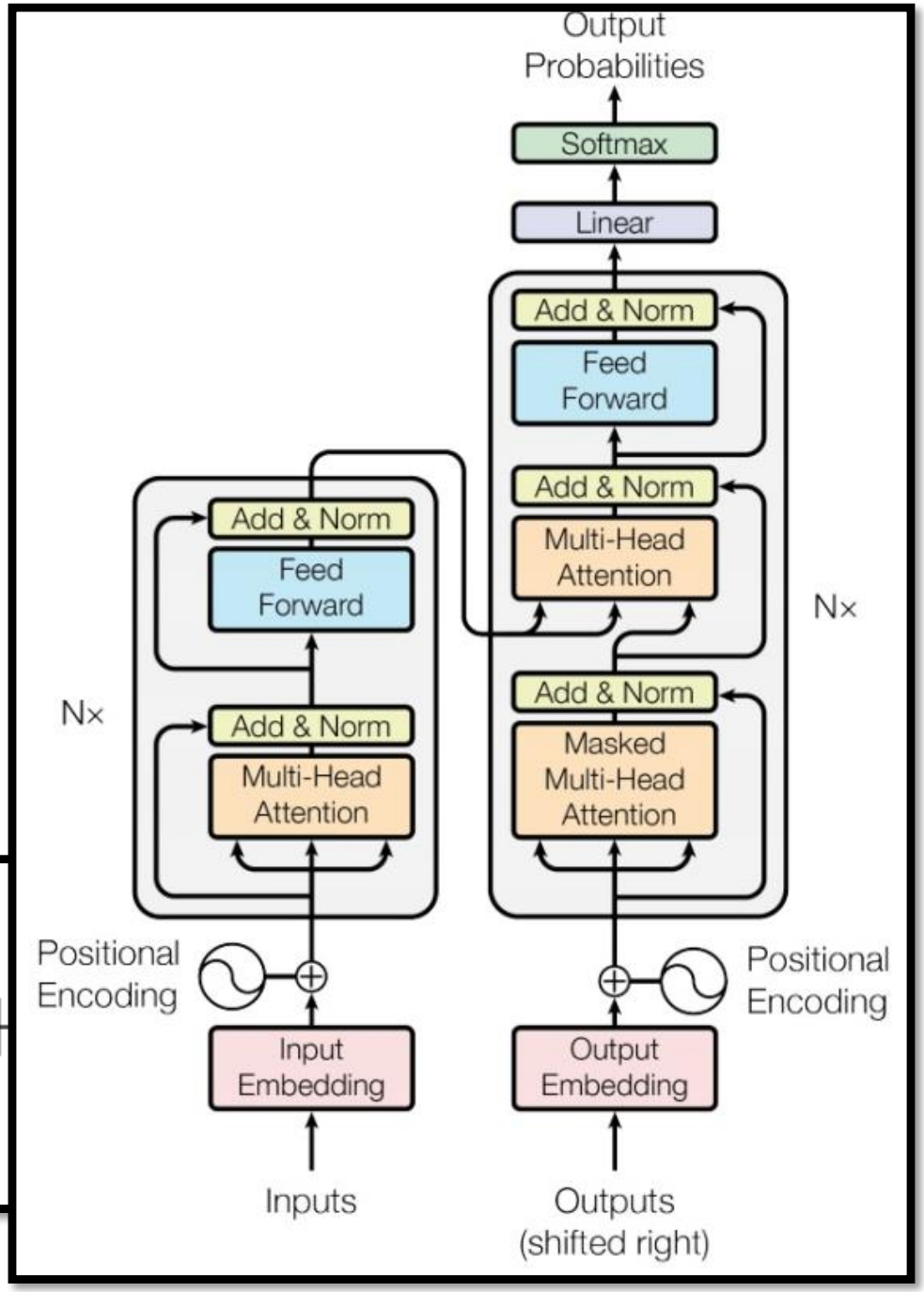
Seq2seq





Sequence to Sequence Learning with Neural Networks

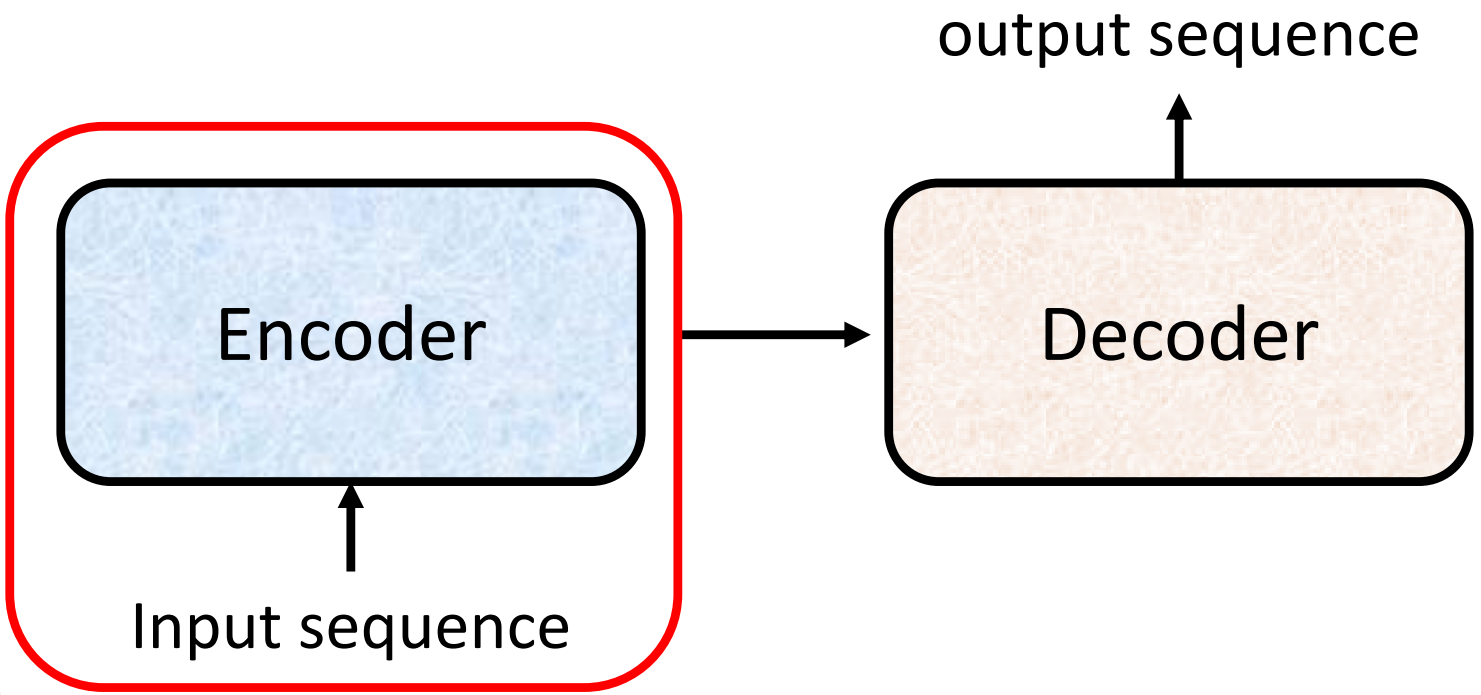
https://arxiv.org/abs/1409.3215



Transformer

https://arxiv.org/abs/1706.03762

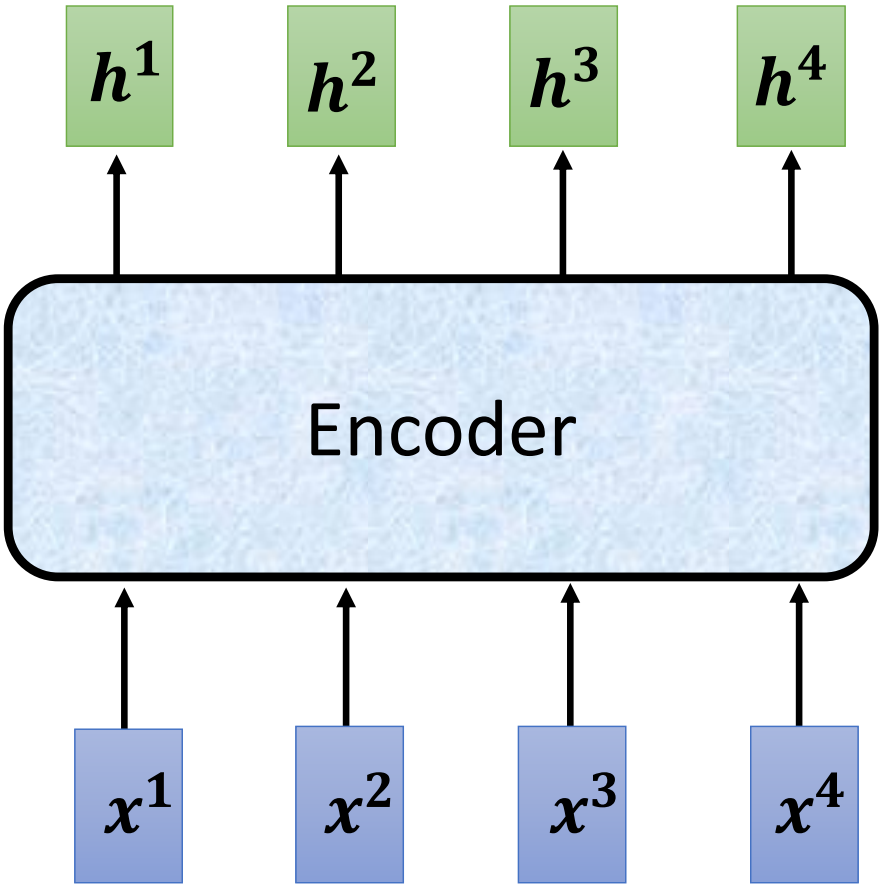
Encoder



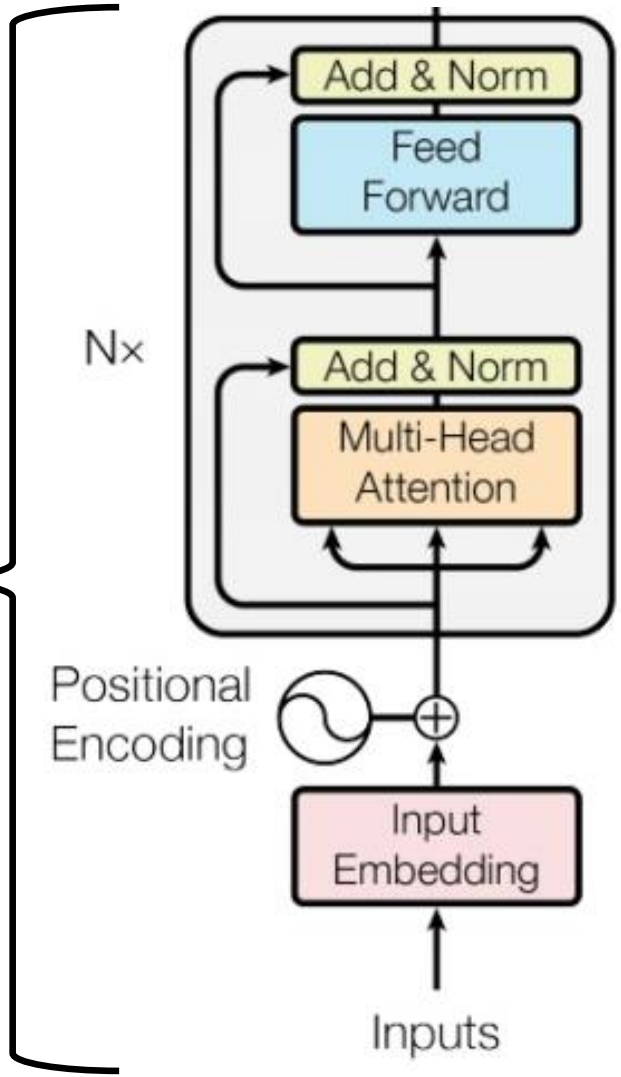
Encoder

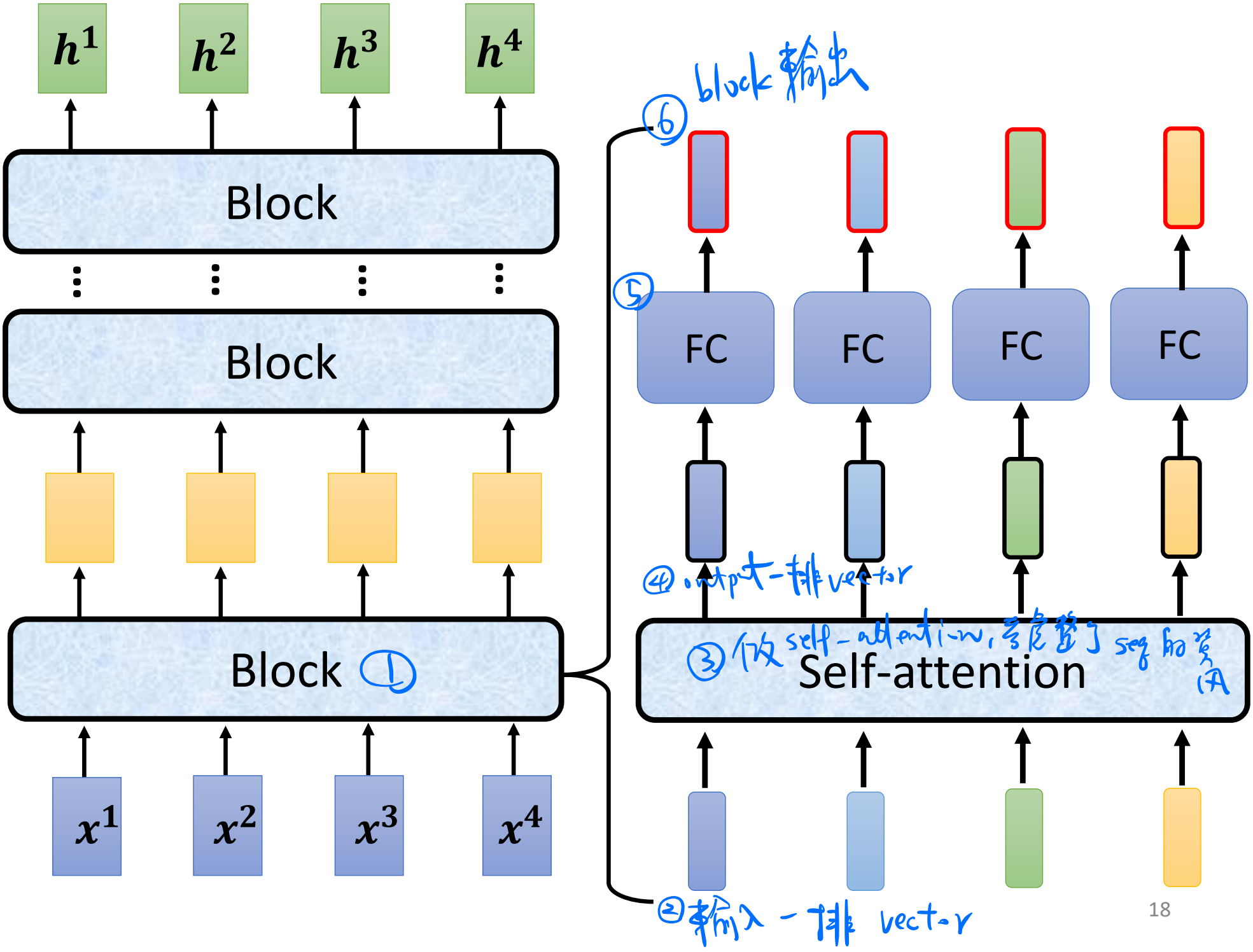
要做的事 => { input 一个向量 / output 一个向量 } => 要做到这了，可用 RNN or CNN or 今日之角

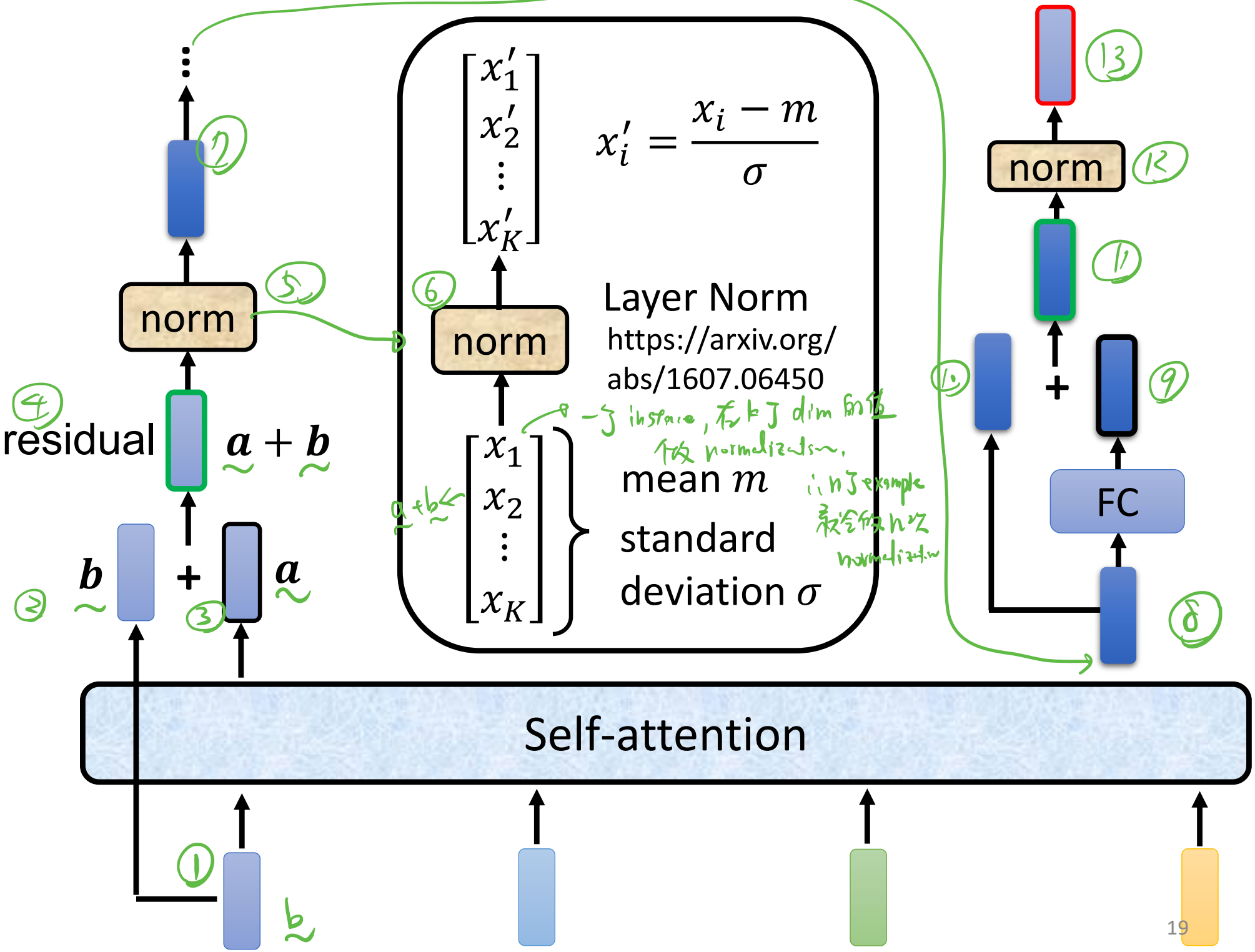
You can use **RNN** or **CNN**.



Transformer's Encoder

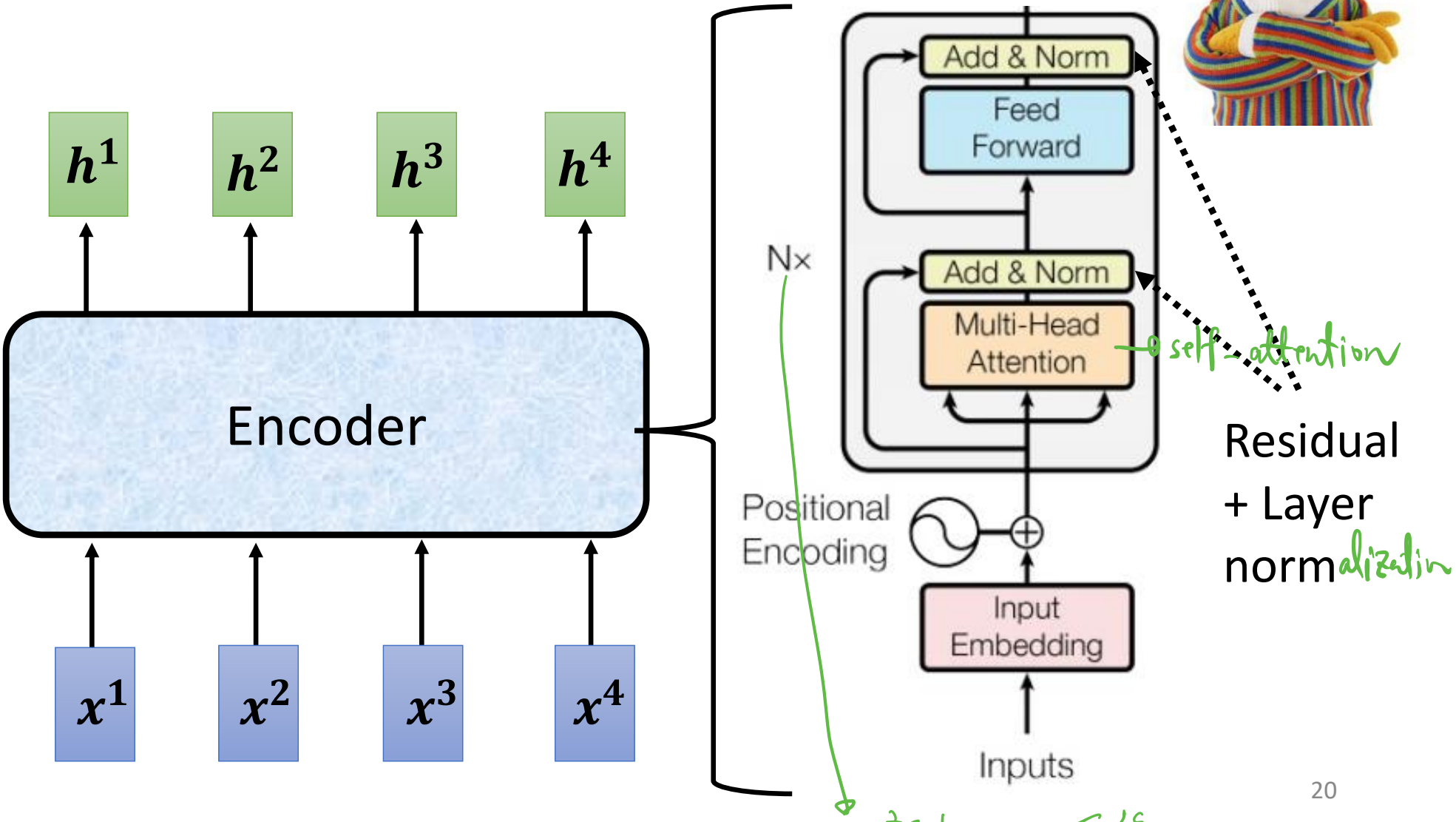



h^1
h^2
h^3
h^4
Block
Block
Block ①
x^1
x^2
x^3
x^4
⑥ block 输出
⑤
FC
FC
FC
FC
④ output 一排 vector
③ 做 self-attention，考虑整个 seq 的资讯
Self-attention
② 输入一排 vector


Self-attention
b
+
a
residual
a + b
norm
Layer Norm
https://arxiv.org/abs/1607.06450
mean m
standard deviation σ
$x'_i = \frac{x_i - m}{\sigma}$
FC
+
norm
1
2
3
4
5
6
7
8
9
10
11
12
13

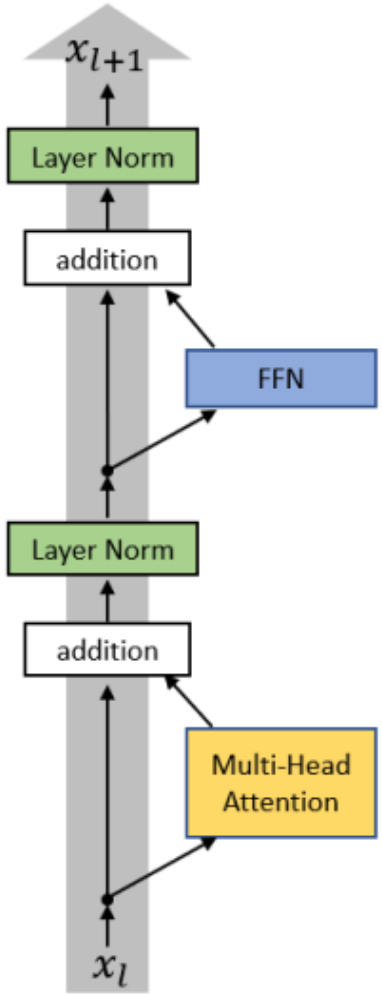
BERT

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

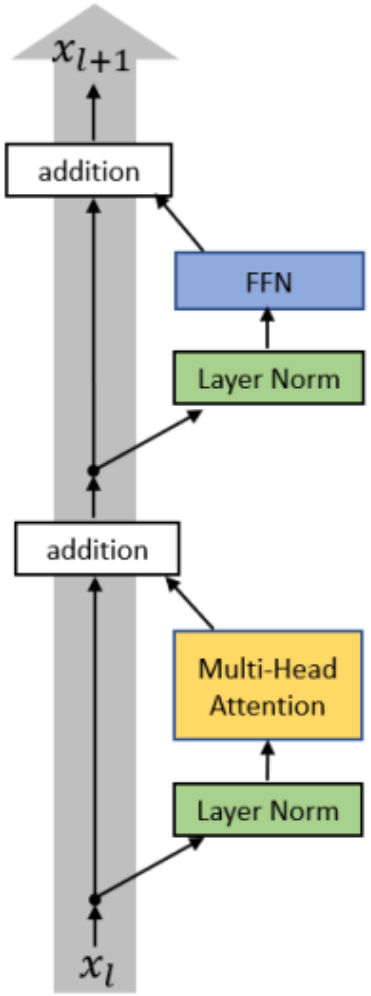


To learn more .....

- 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

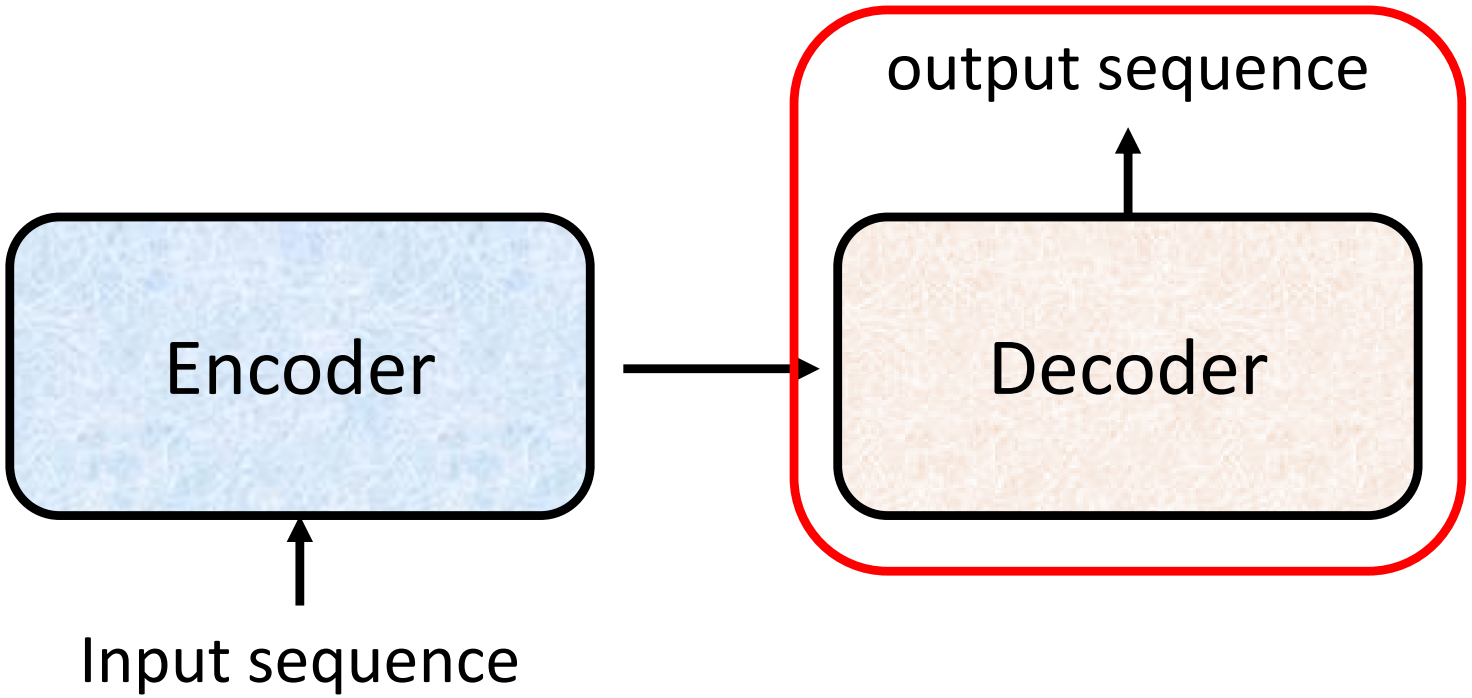


(a)



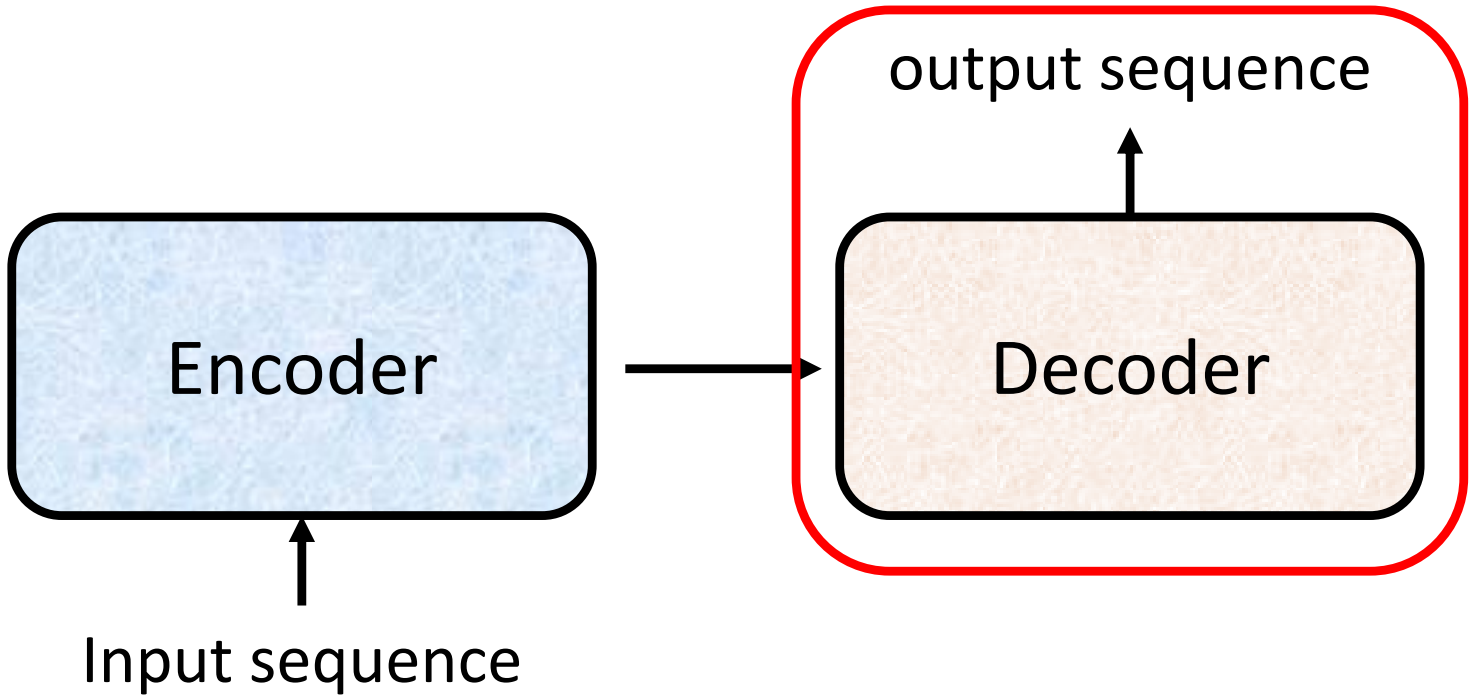
(b)

Decoder



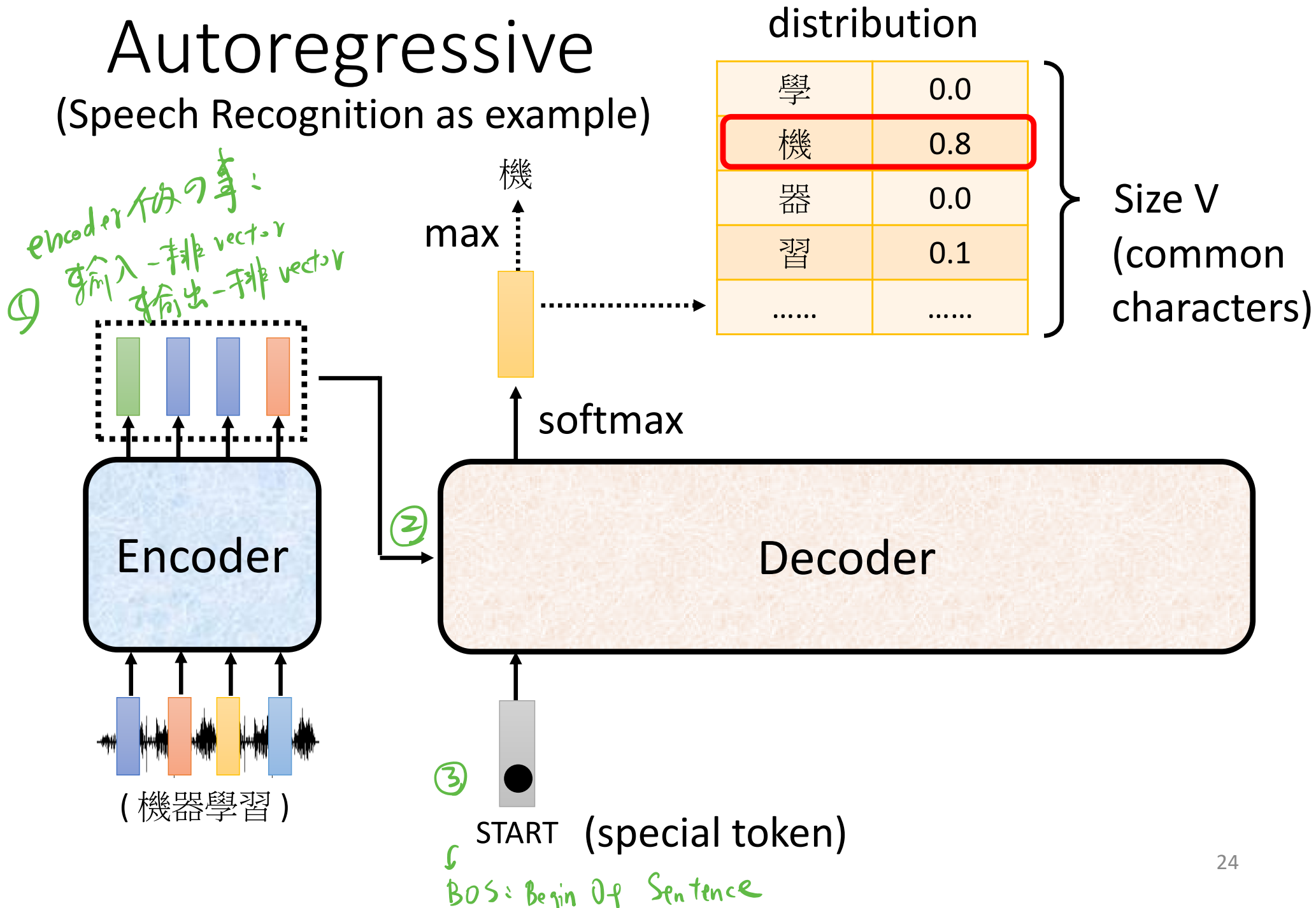
Decoder有2种 Autoregressive / Non-autoregressive

Decoder – Autoregressive (AT)

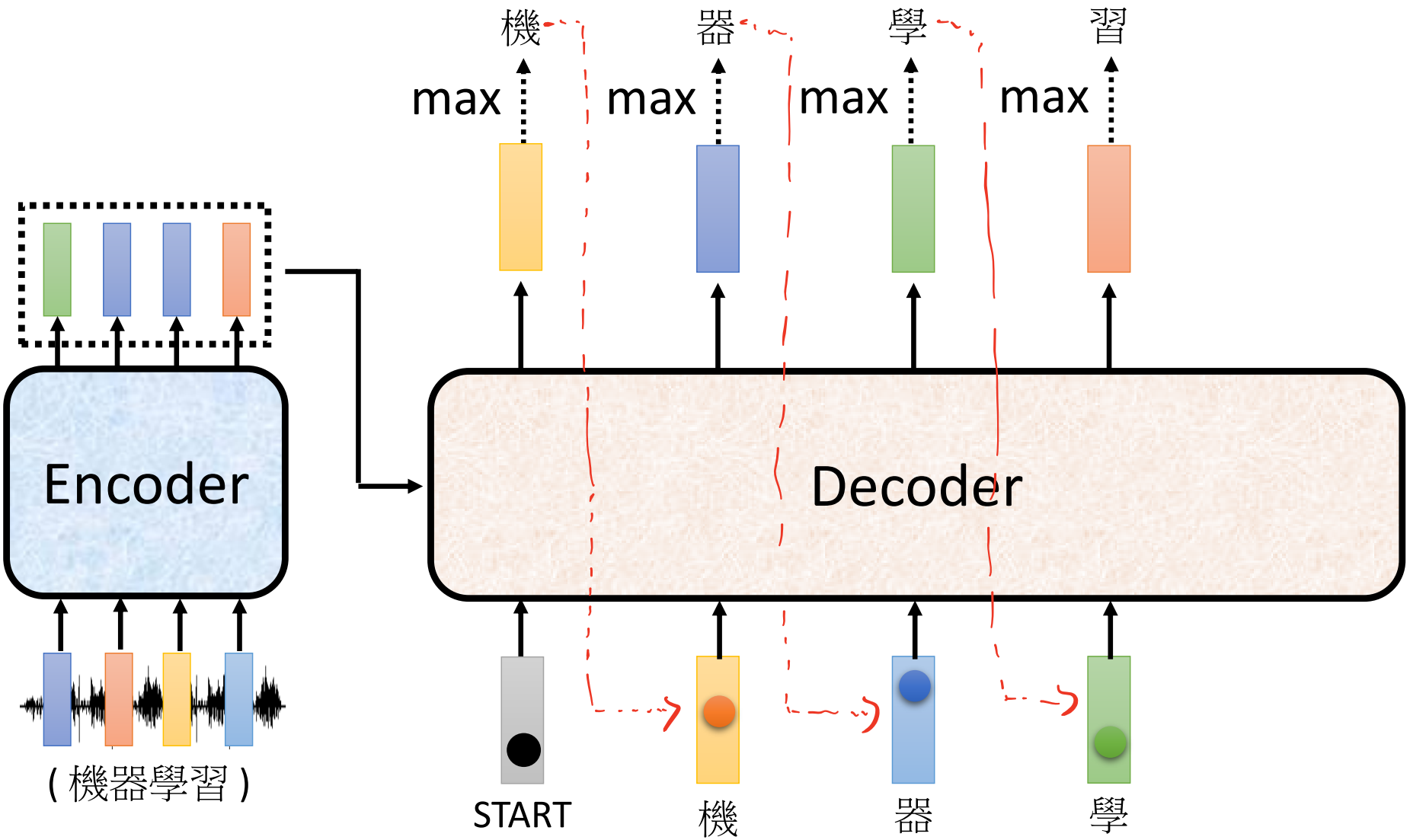


Autoregressive

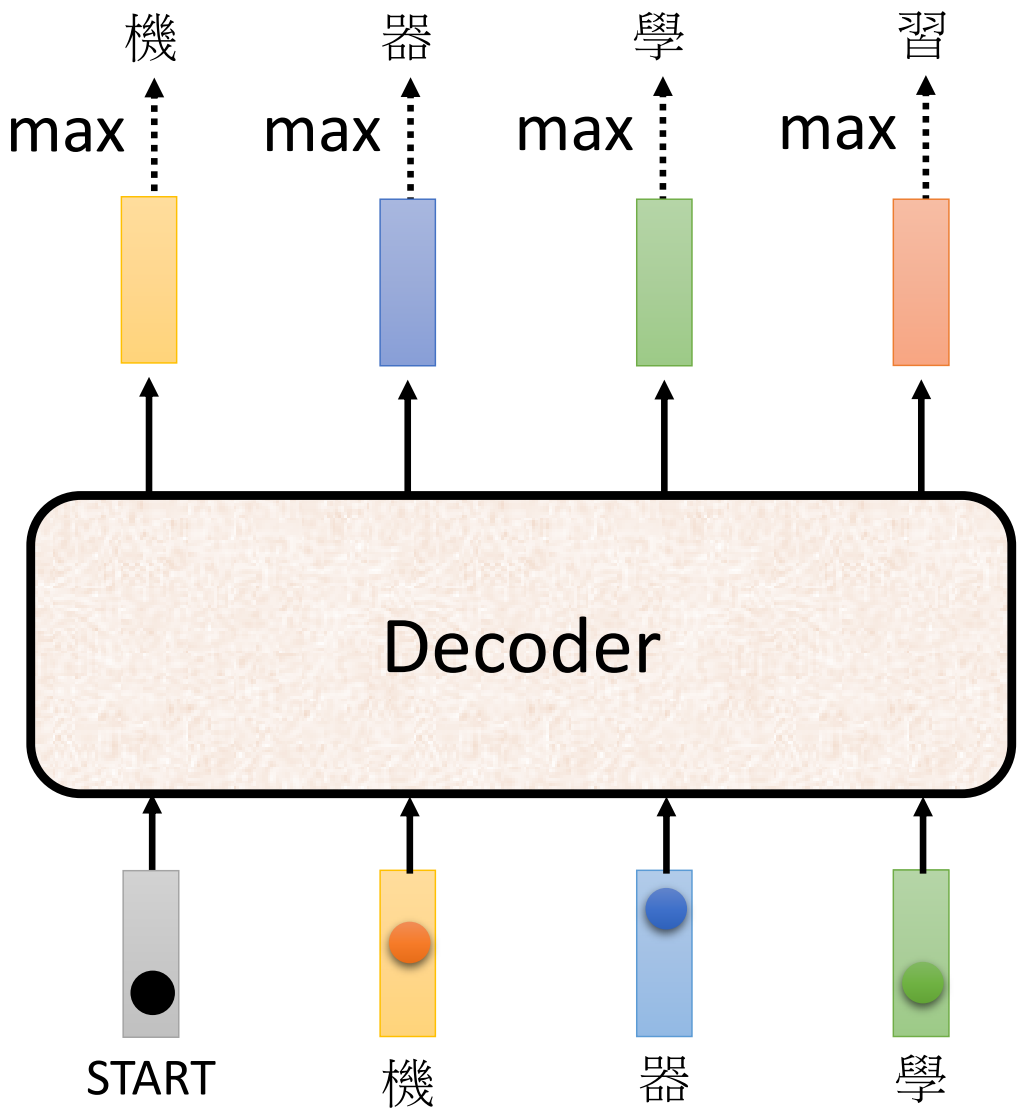
(Speech Recognition as example)

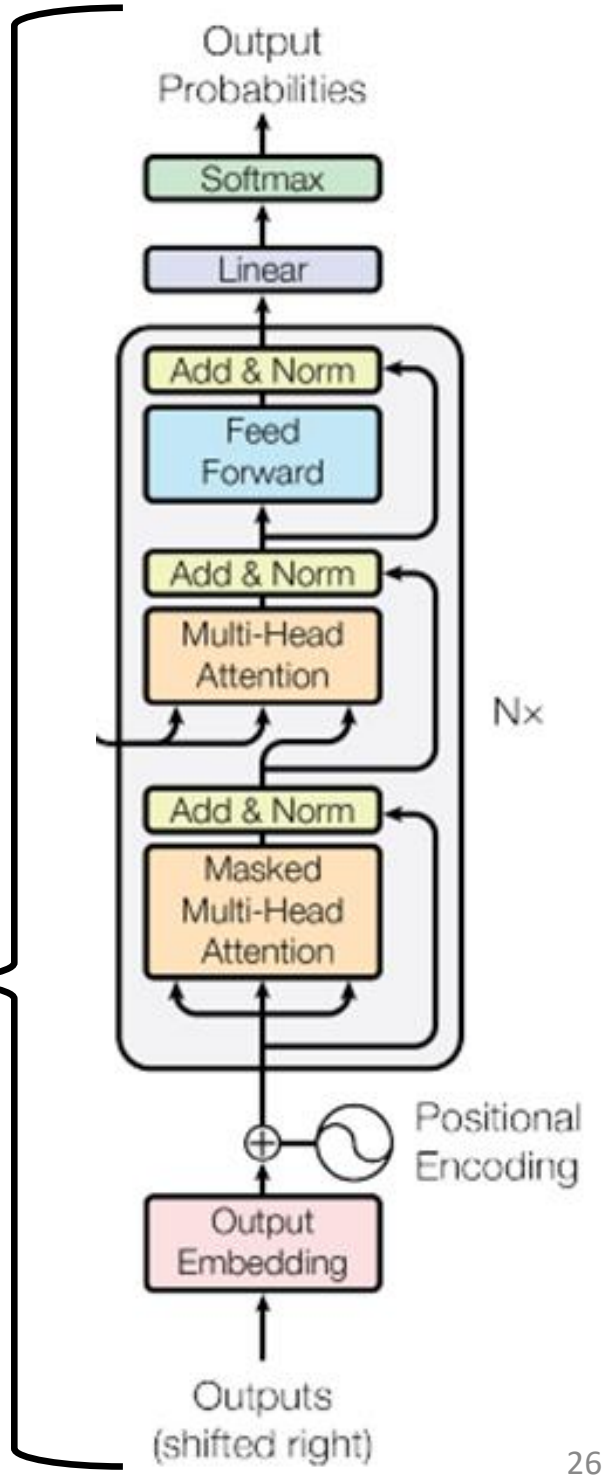


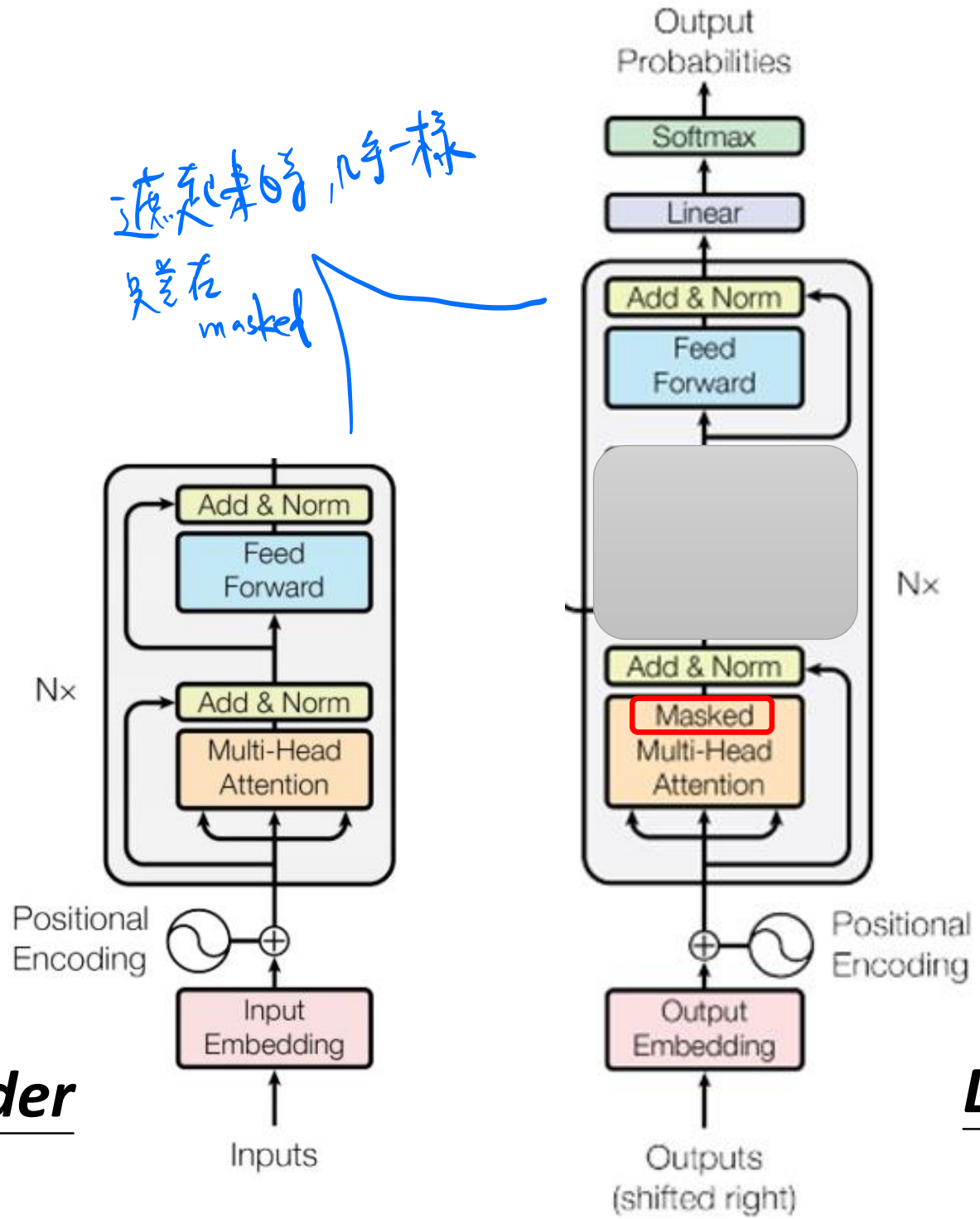
Autoregressive



ignore the input from the encoder here ☺



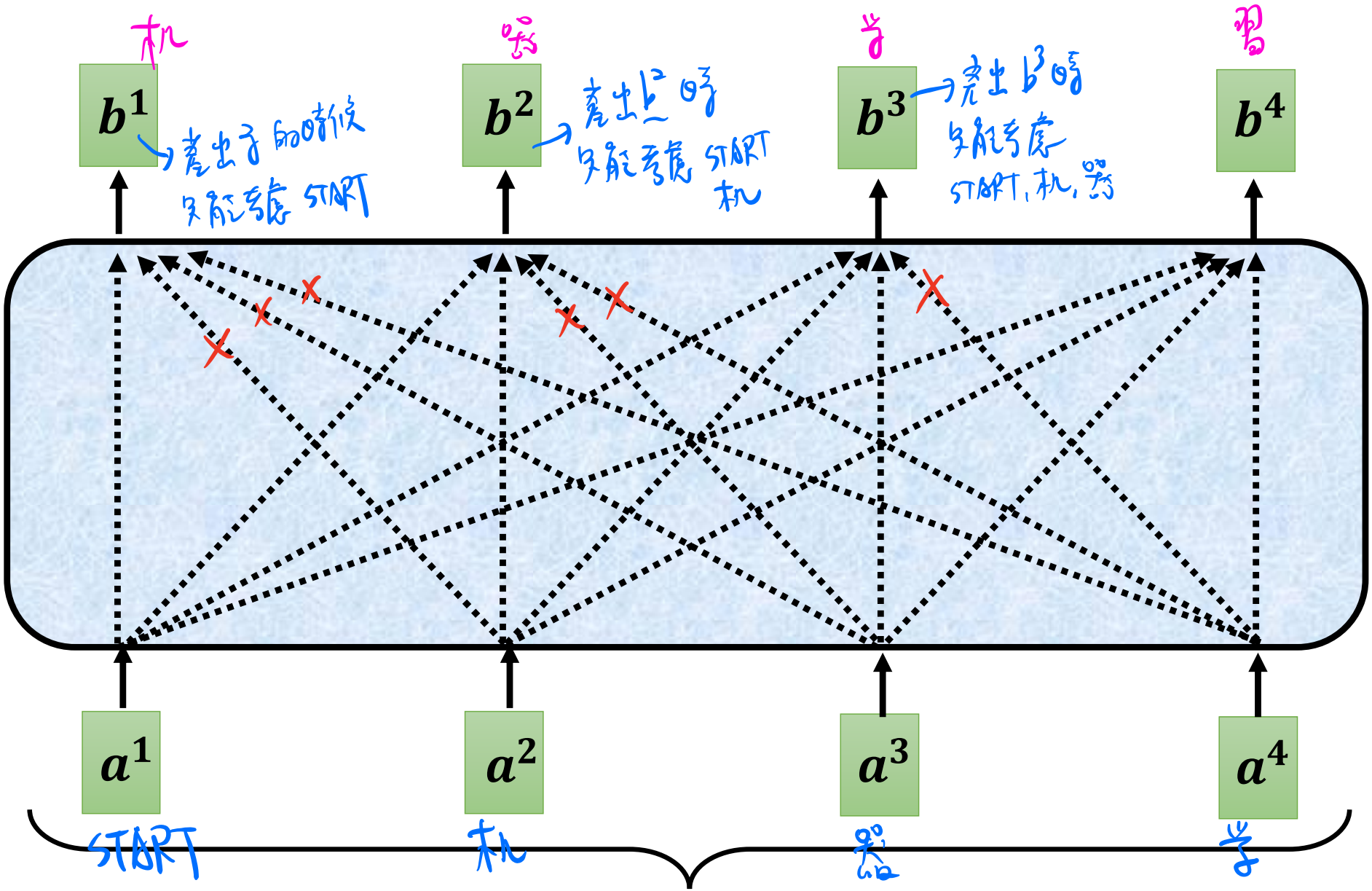




Encoder

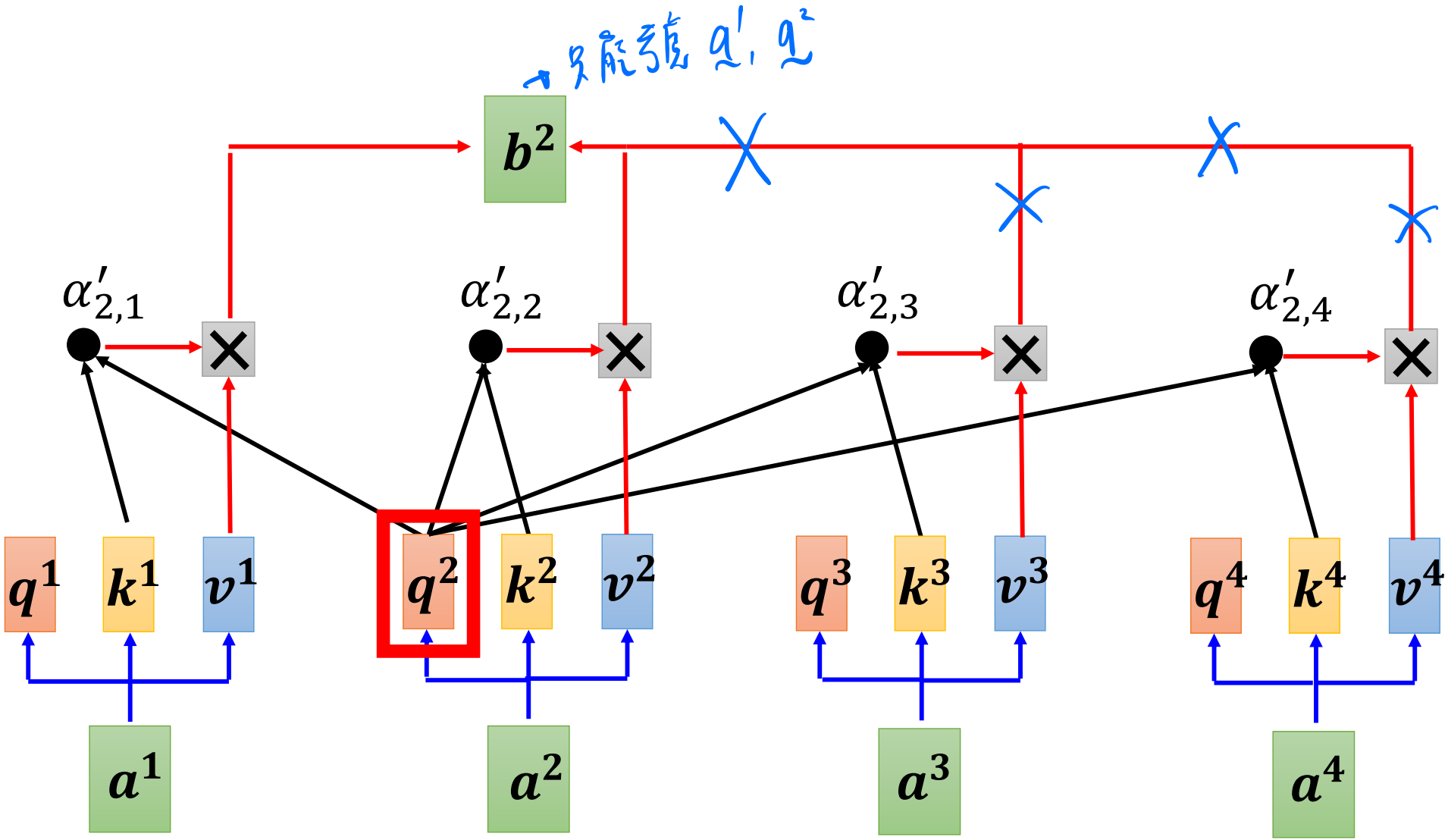
Decoder

Self-attention → Masked Self-attention



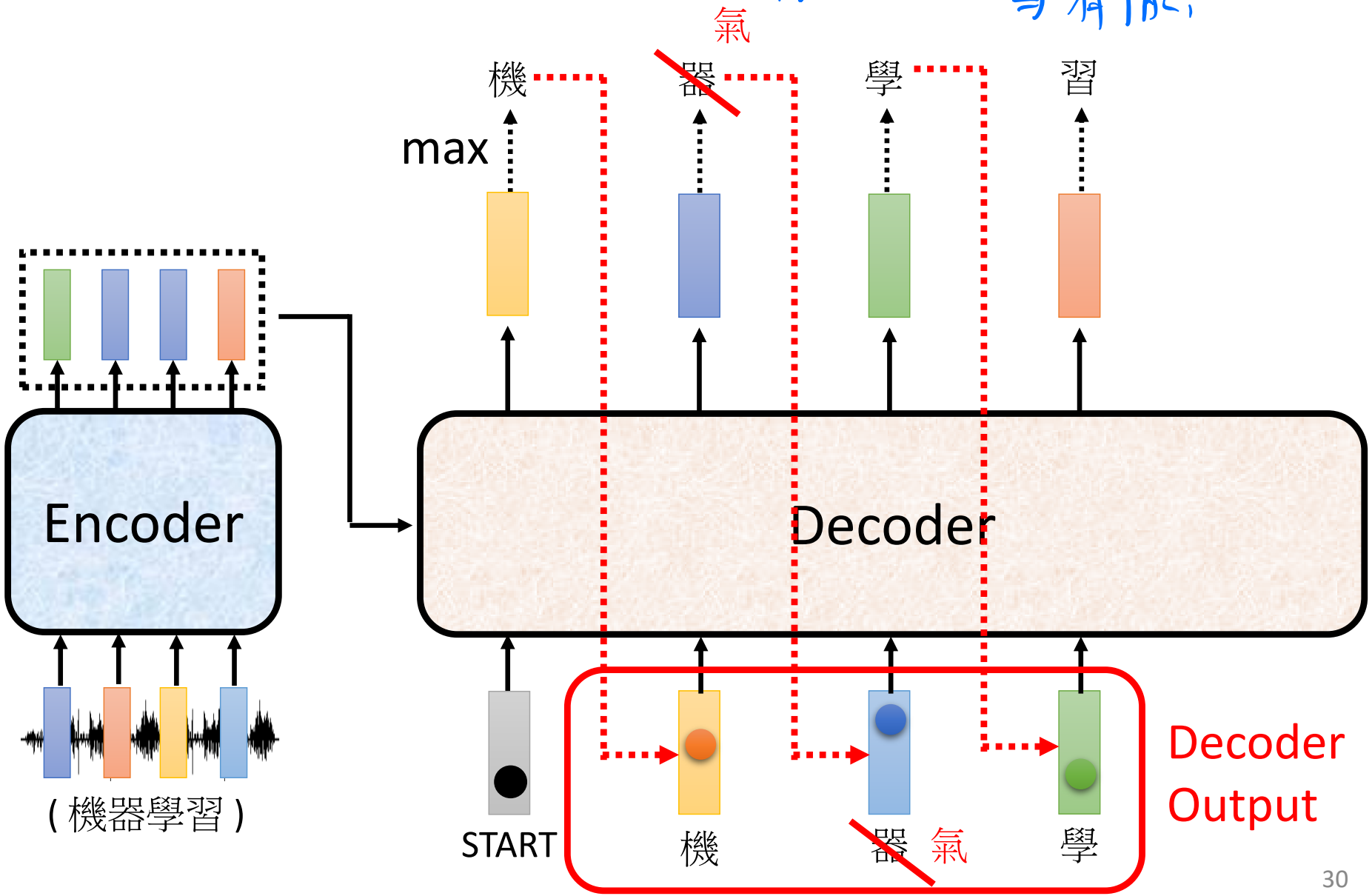
Can be either **input** or **a hidden layer**

Self-attention → Masked Self-attention



Why masked? Consider how does decoder work

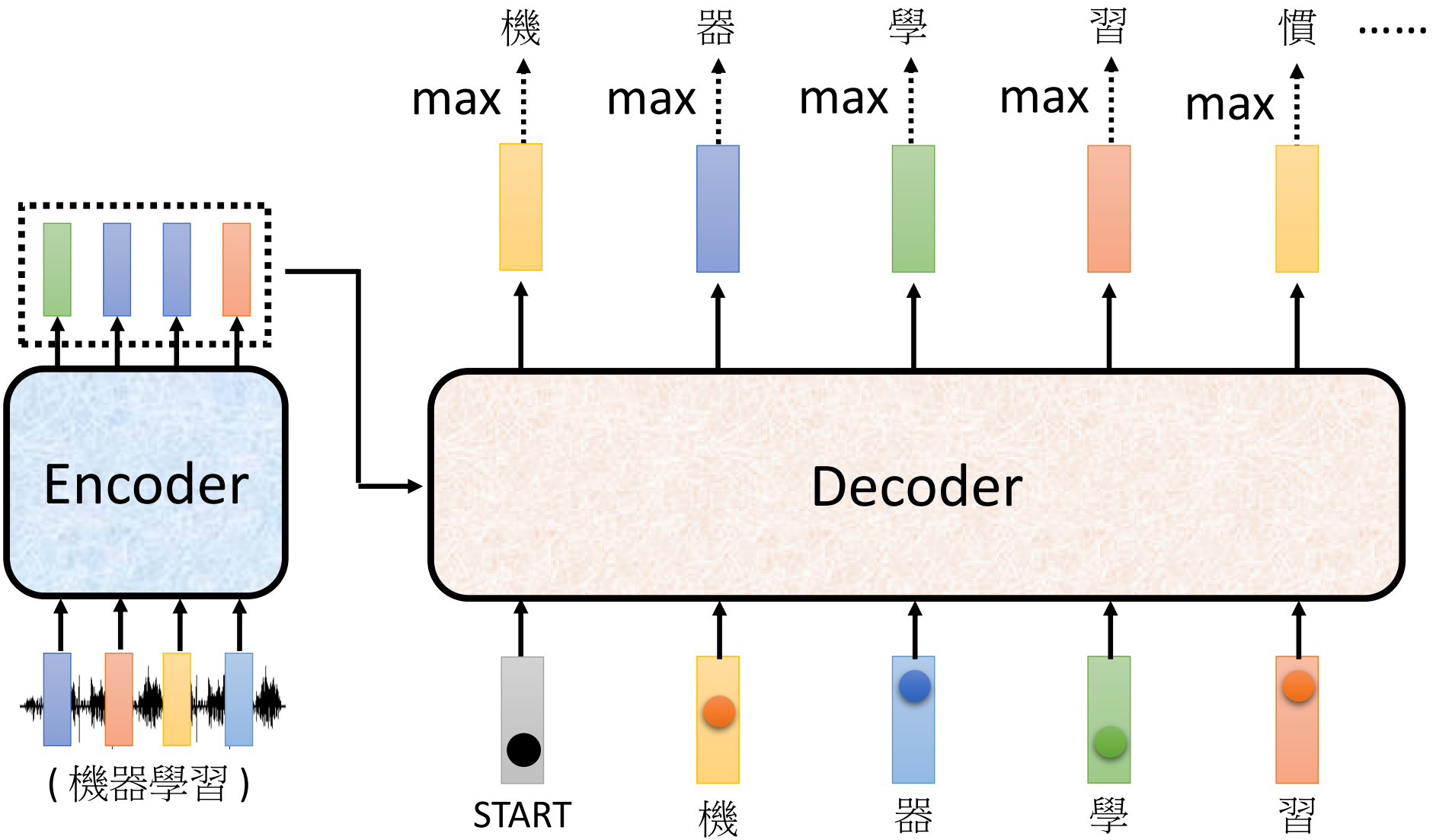
Autoregressive



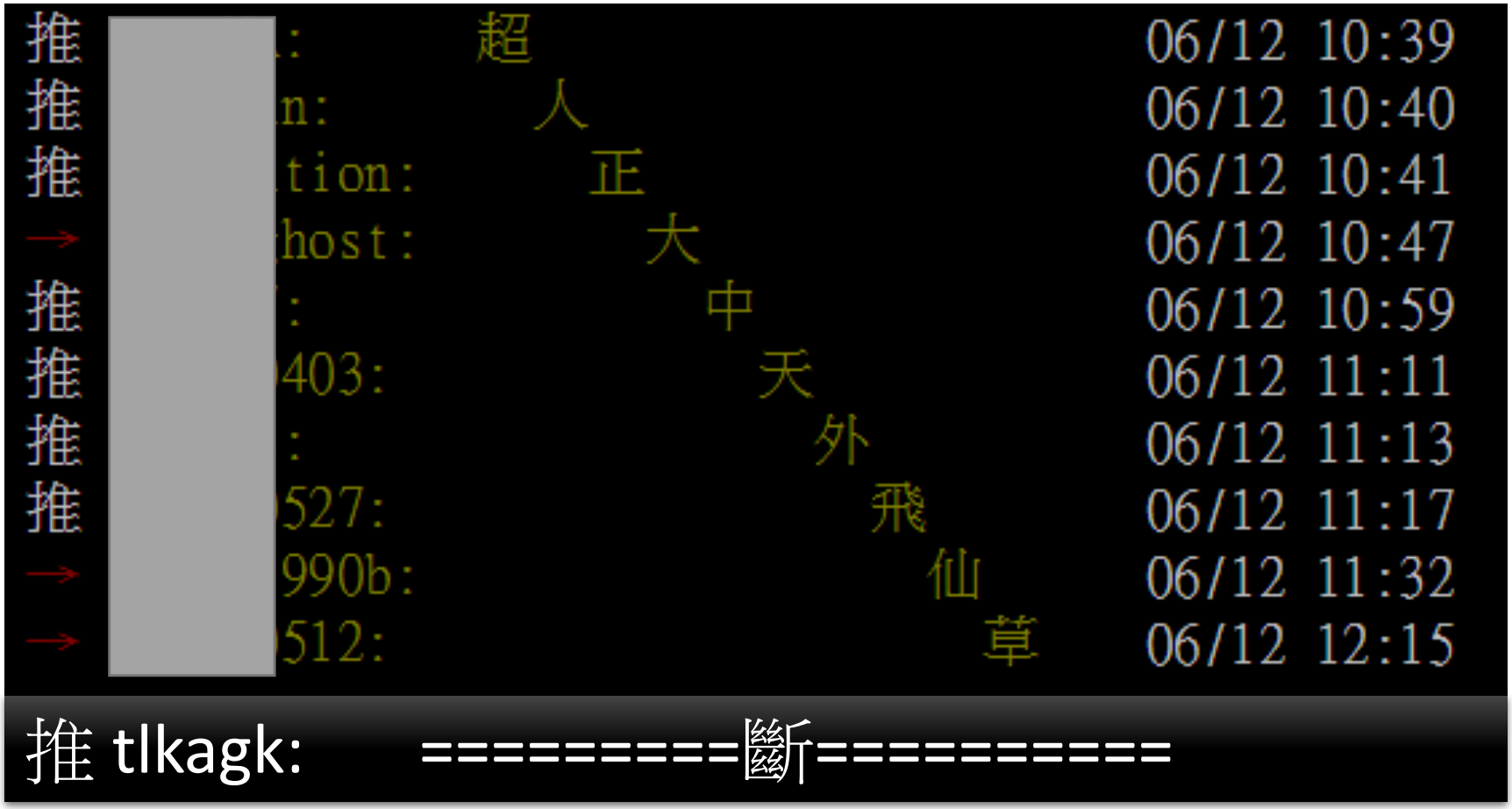
Autoregressive

We do not know the correct output length.

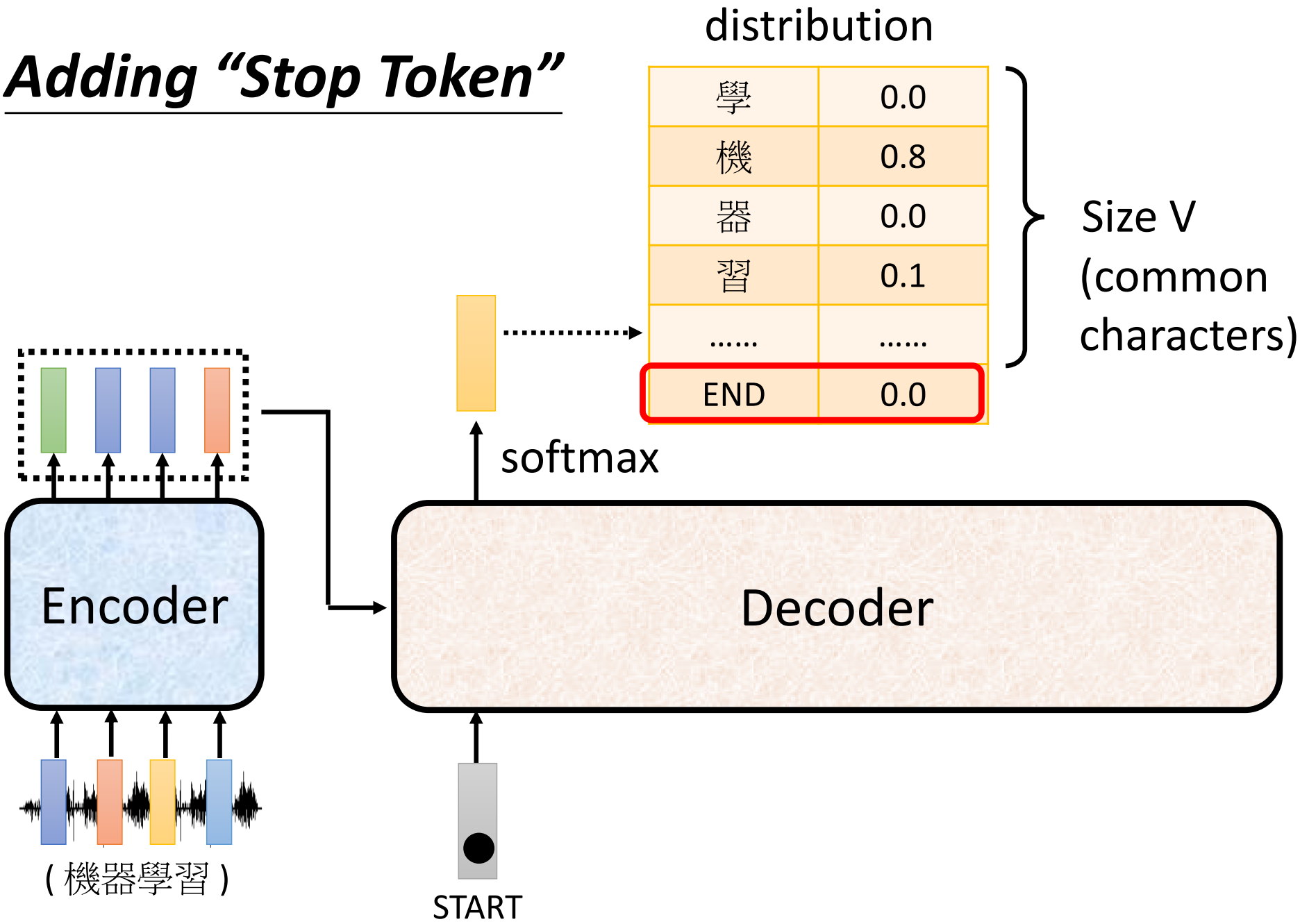
Never stop!



推文接龍 (Tweet Solitaire)

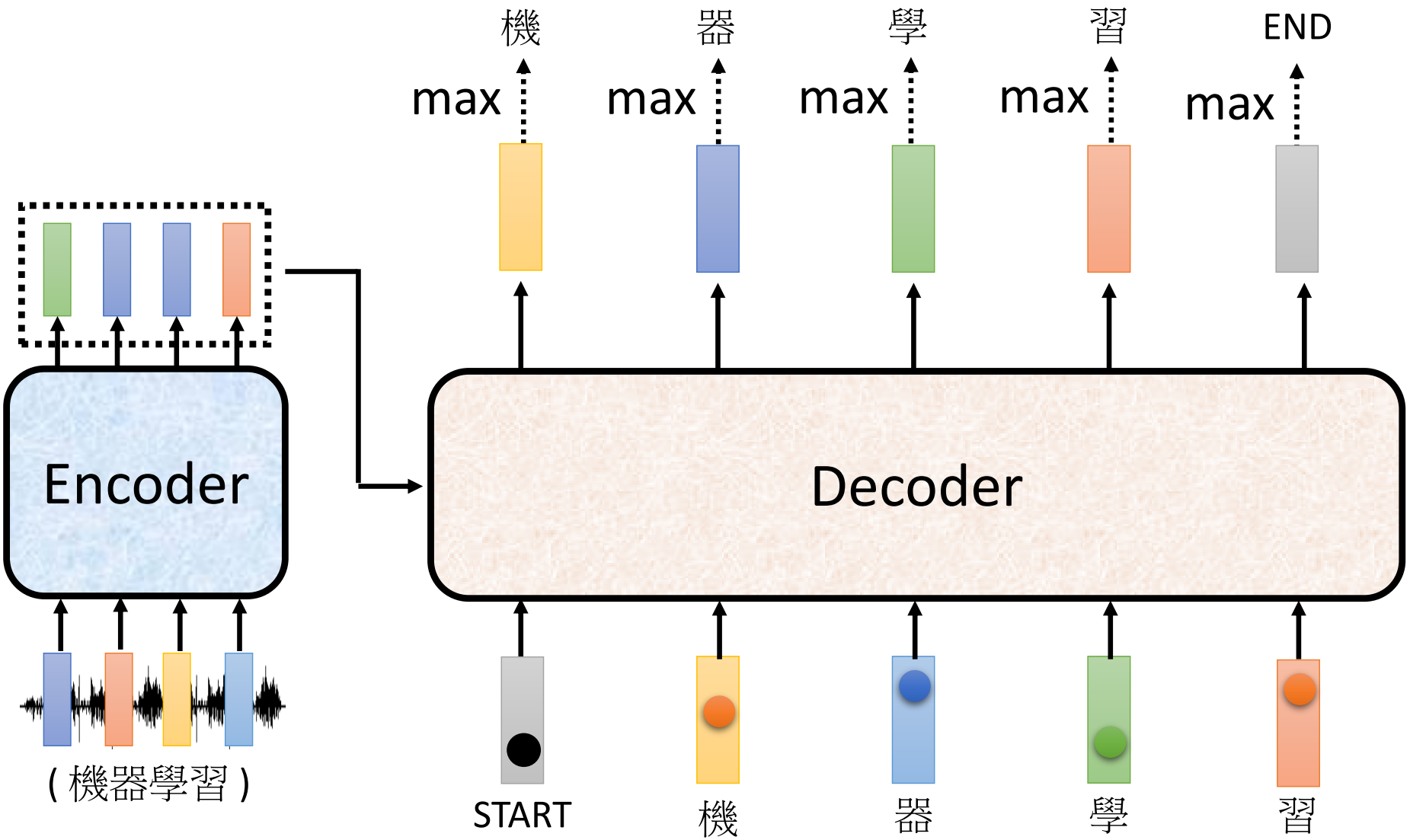


Adding "Stop Token"



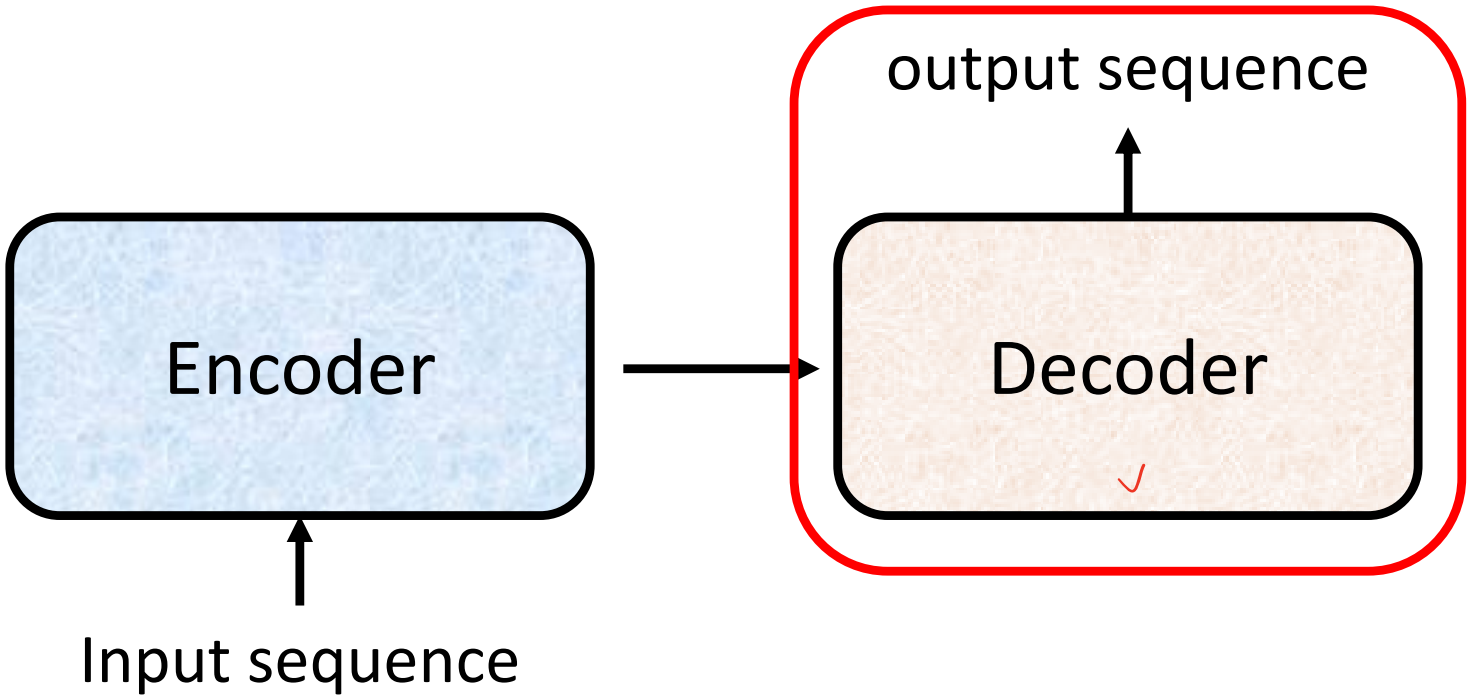
Autoregressive

Stop at here!

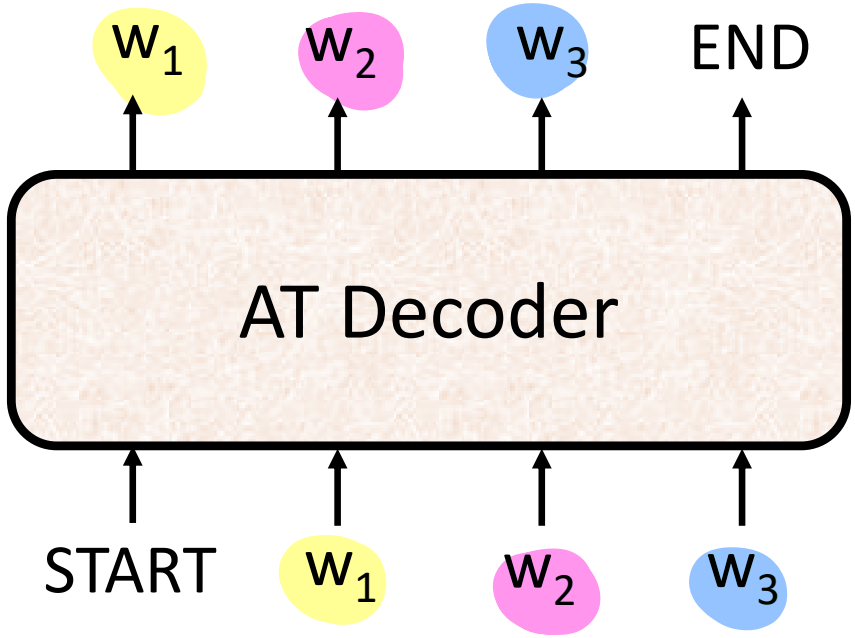


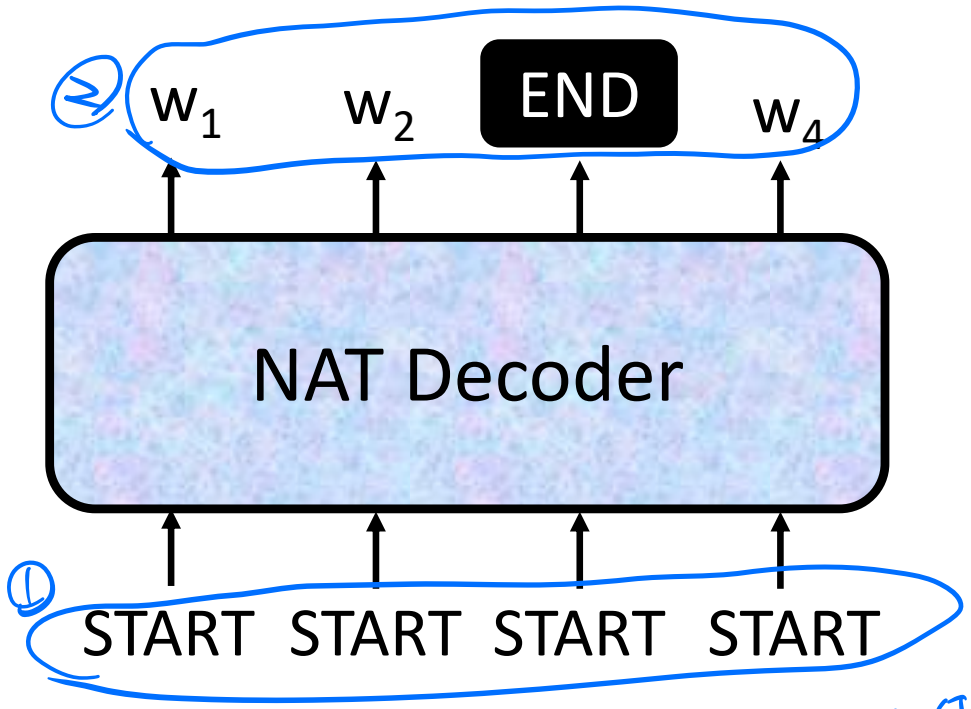
2页报告，介绍一下NAT

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**)

To learn more ……

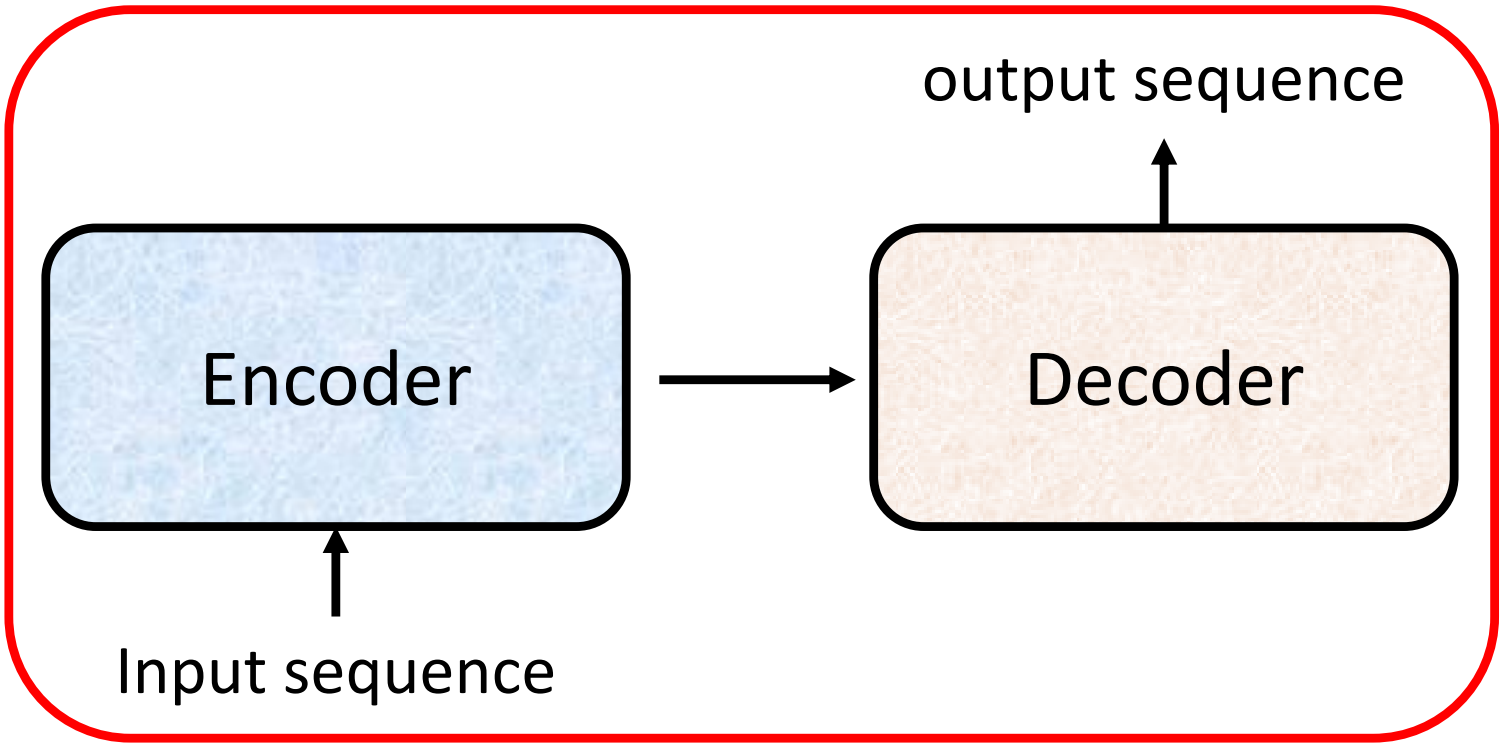
NATの助教課（大坑）

https://youtu.be/jvyKmU4OM3c

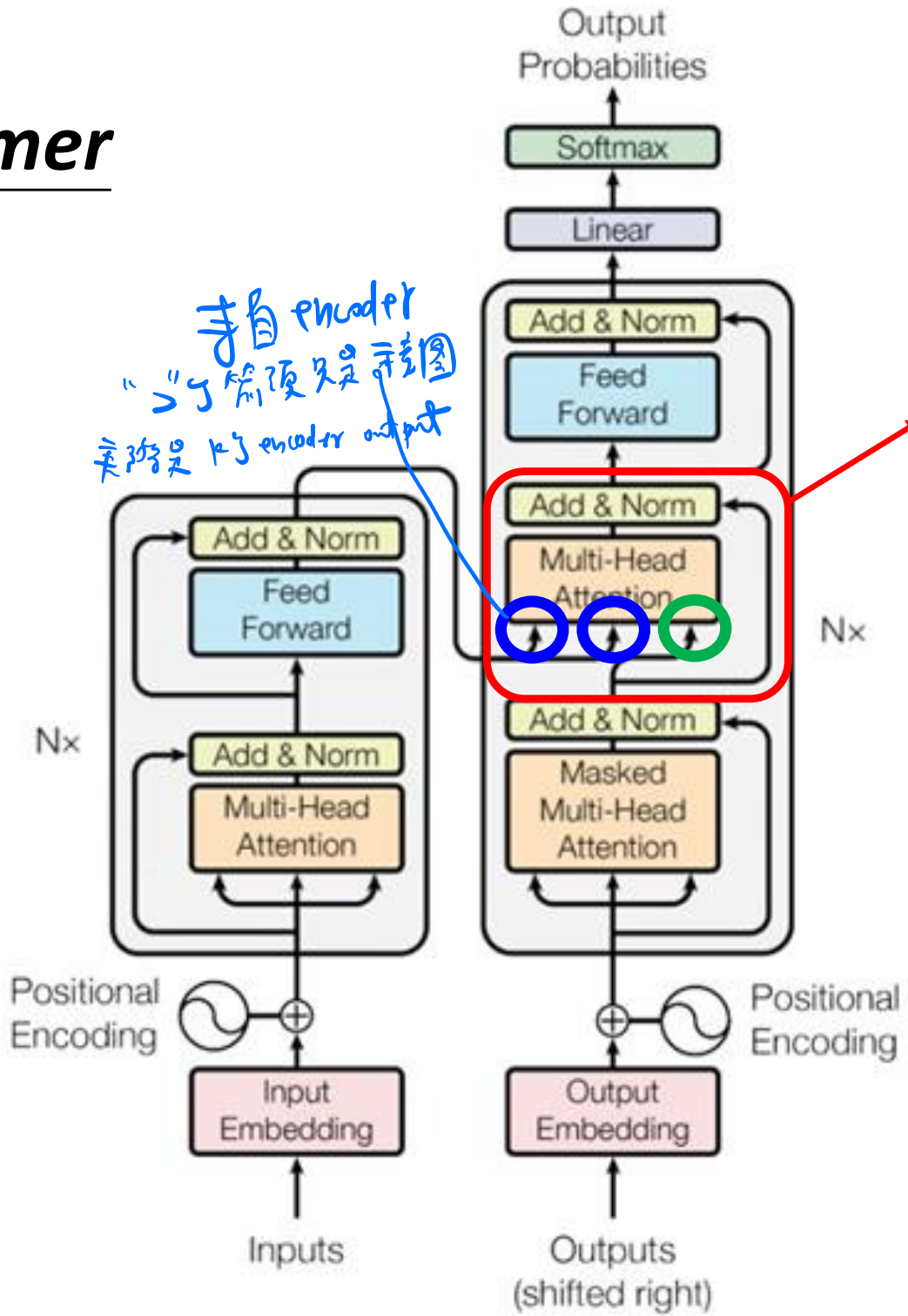
(in Mandarin)

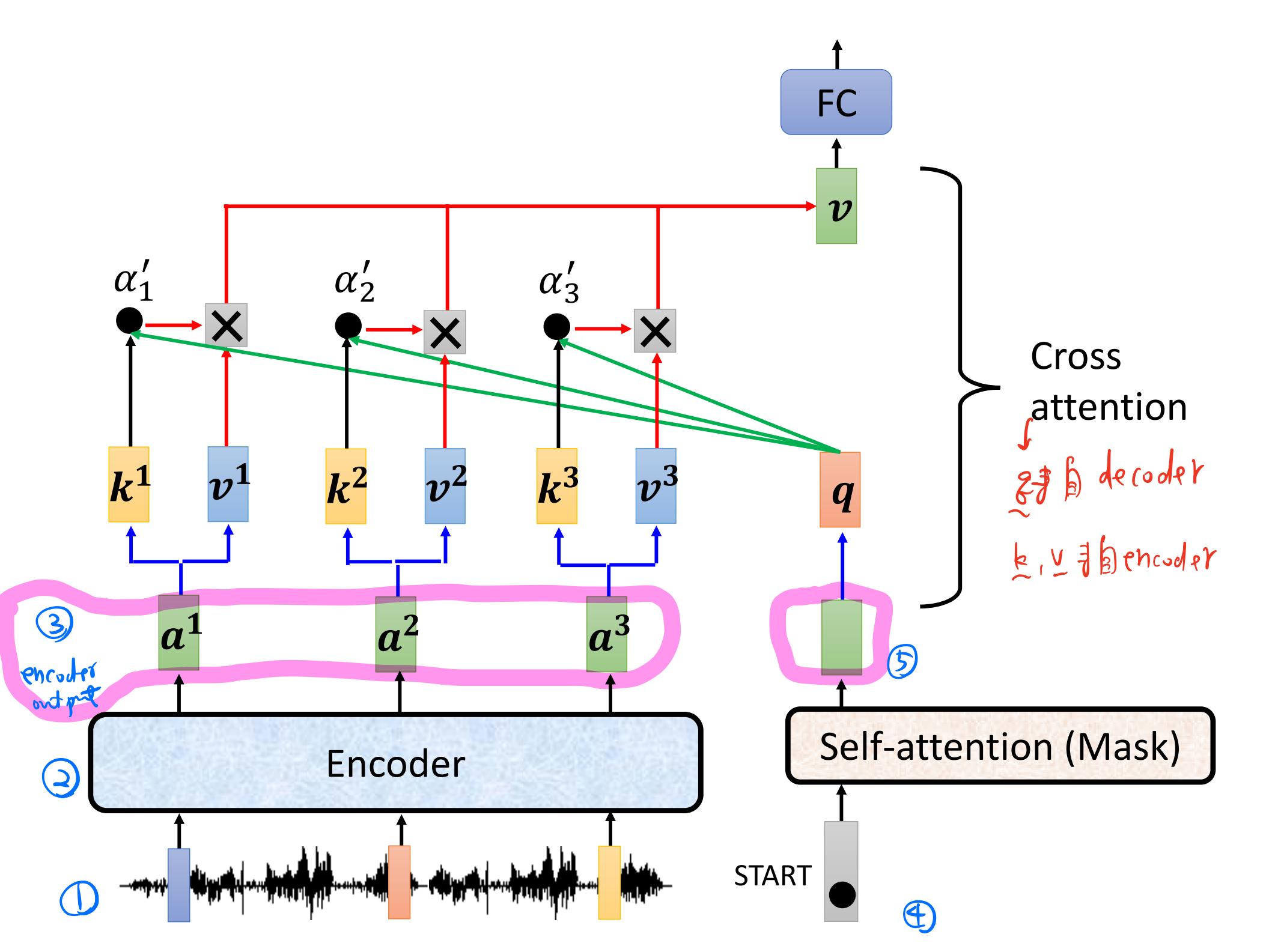
回頭看 decoder，用了被遮起來的地方 ⇒ cross-attention

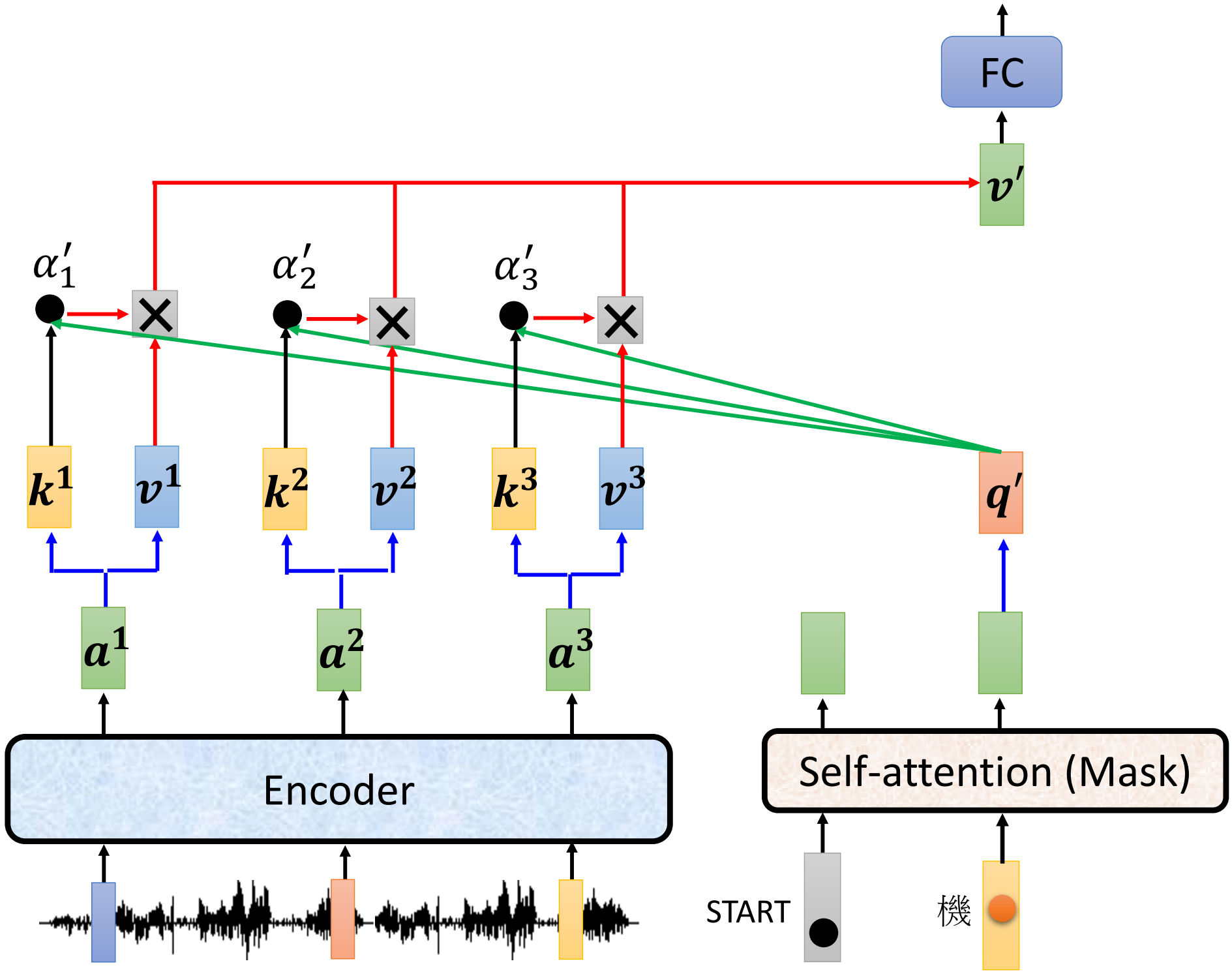
Encoder-Decoder



Transformer



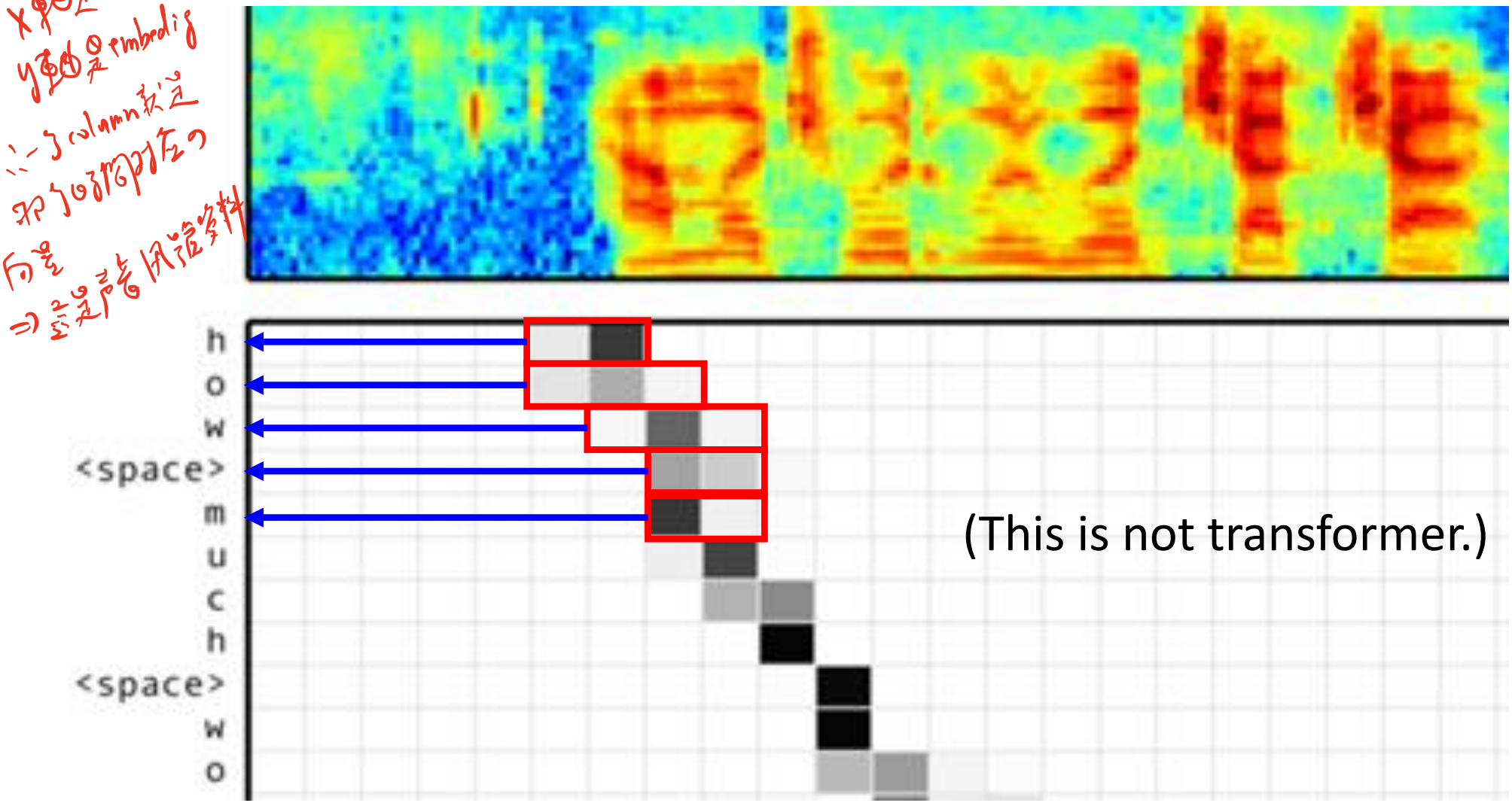

FC
v
α'_1
α'_2
α'_3
k^1
v^1
k^2
v^2
k^3
v^3
q
Cross attention
a^1
a^2
a^3
③
encoder output
⑤
Encoder
②
Self-attention (Mask)
①
START
④


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
機

Cross Attention

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

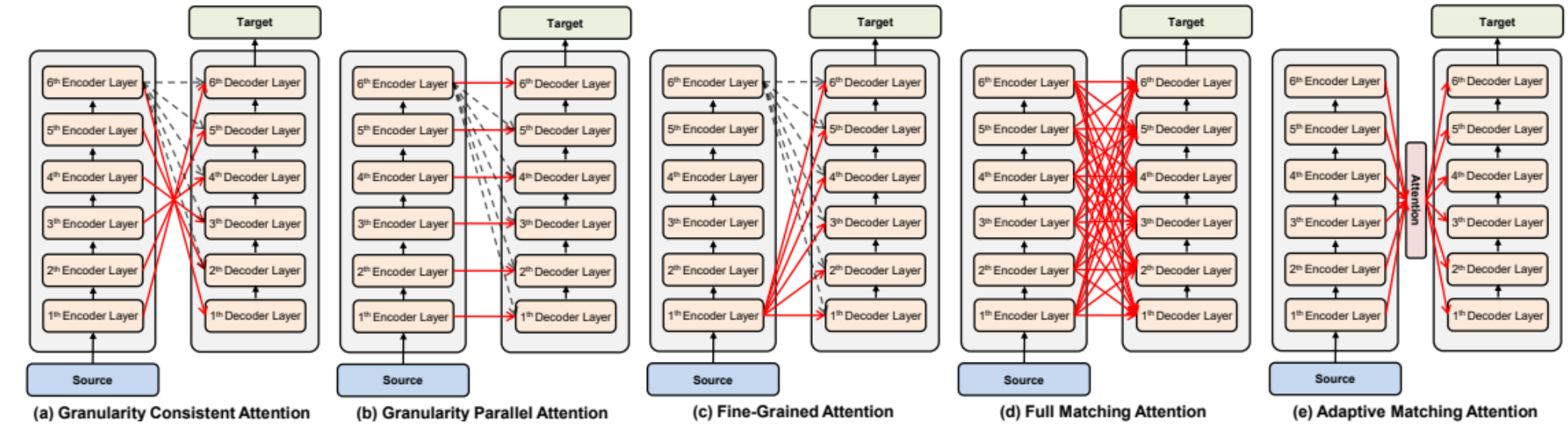
https://ieeexplore.ieee.org/document/7472621



(This is not transformer.)

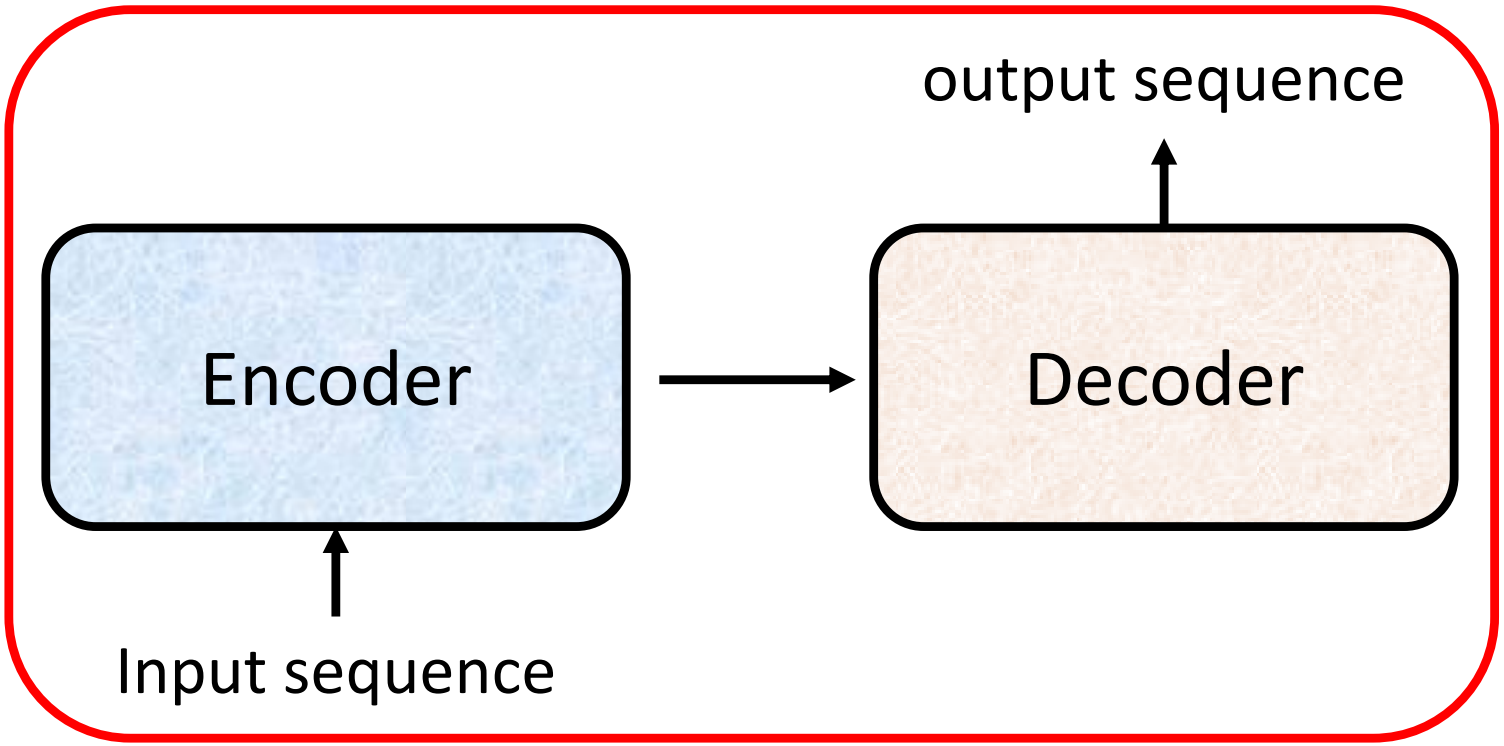
Cross Attention

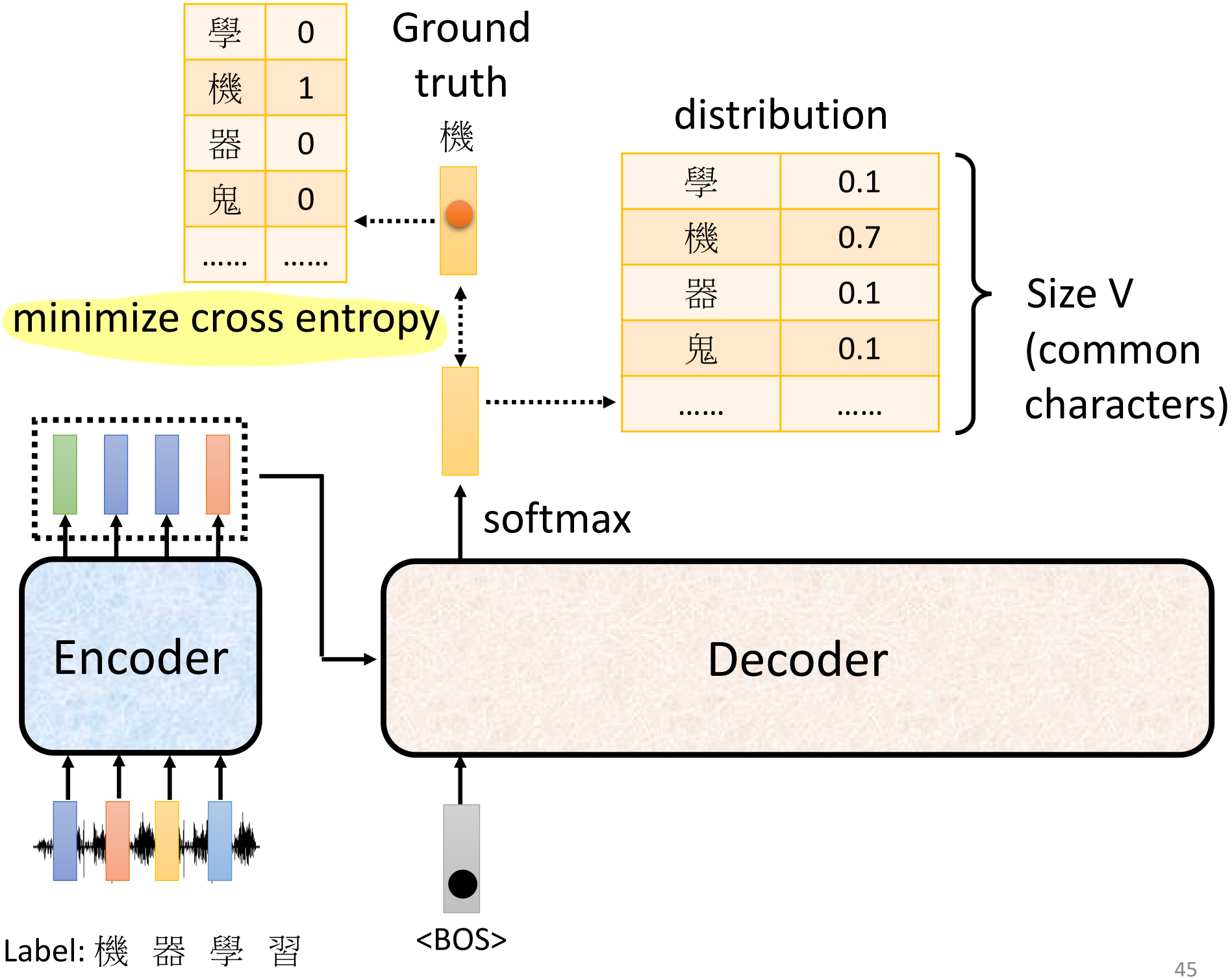
Source of image:
https://arxiv.org/abs/2005.08081



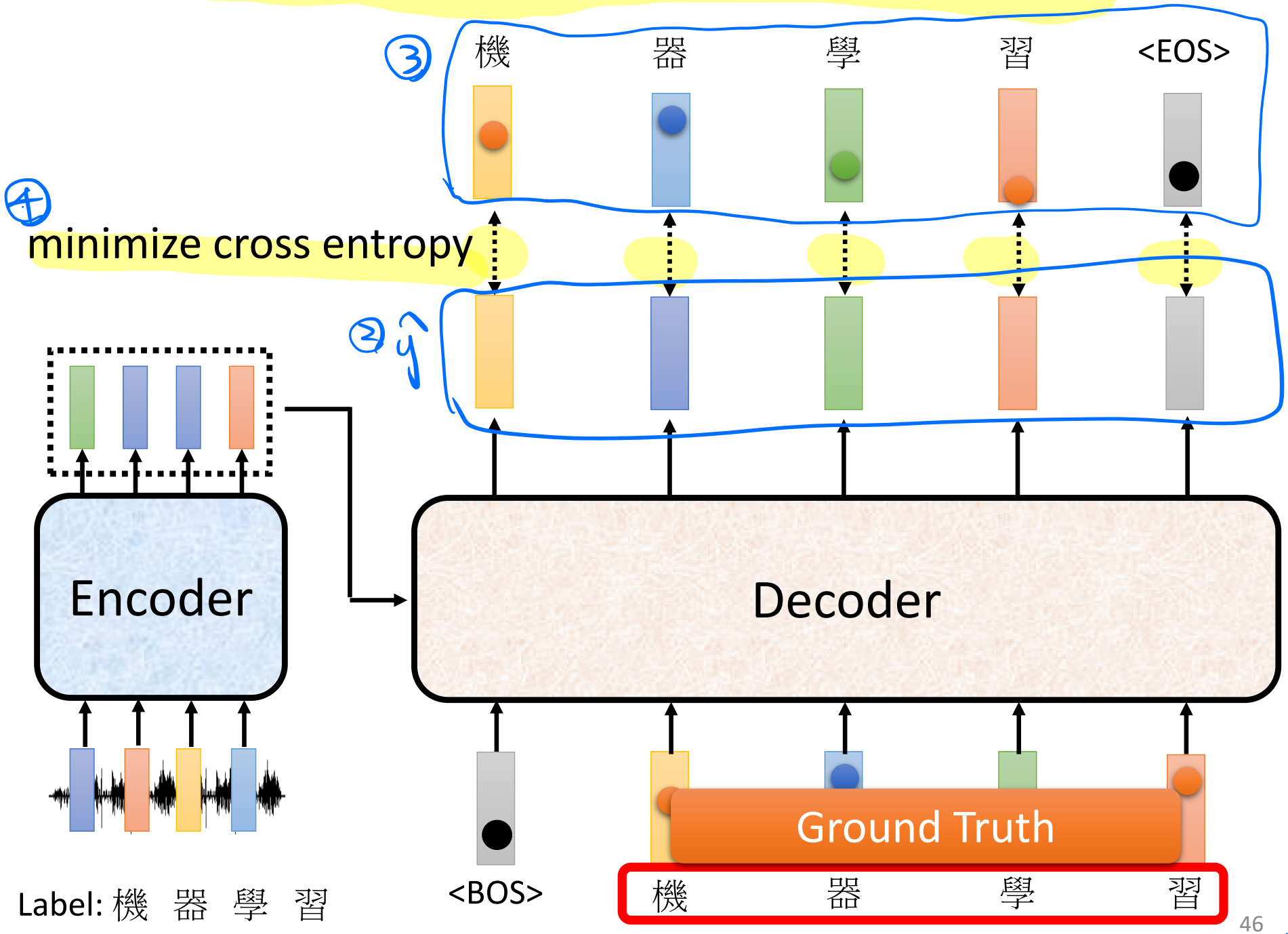
剛才都在講如何 inference，
現在來講怎麼 train

Training



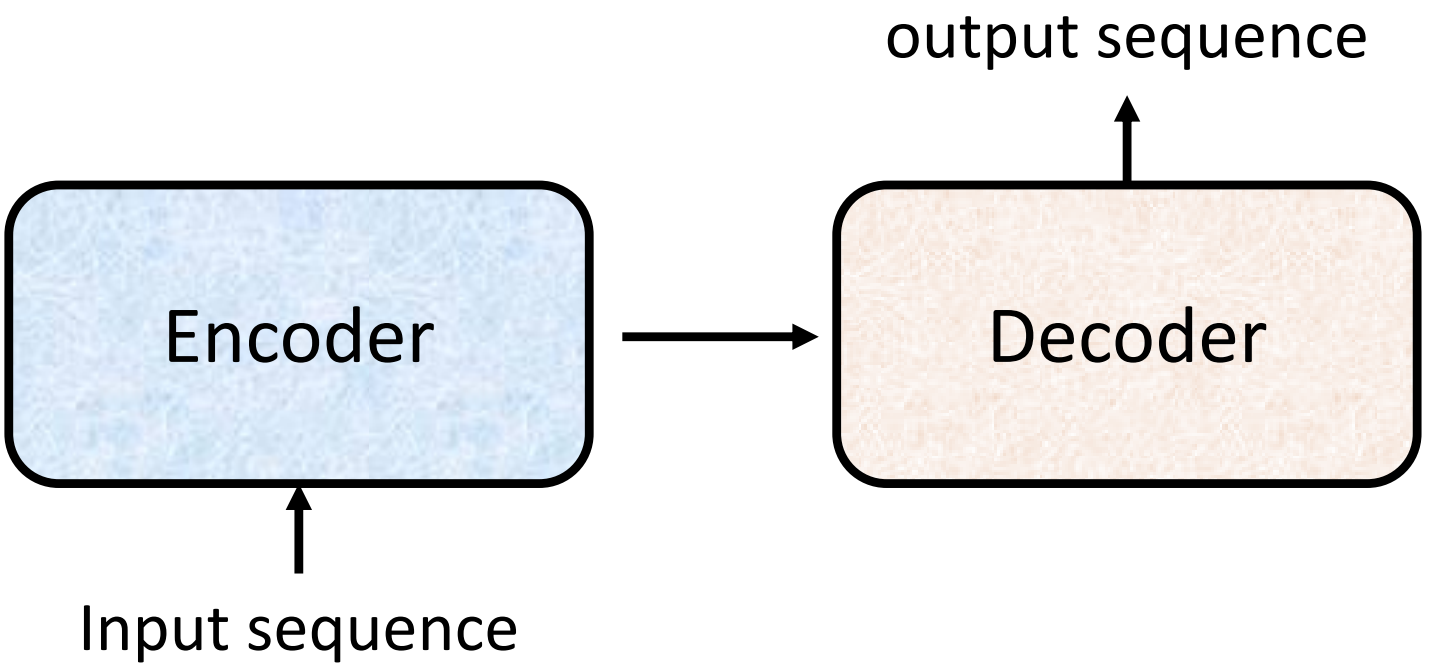

學 0
機 1
器 0
鬼 0
....
Ground truth
機
distribution
學 0.1
機 0.7
器 0.1
鬼 0.1
....
Size V (common characters)
minimize cross entropy
softmax
Encoder
Decoder
<BOS>
Label: 機 器 學 習

Teacher Forcing: using the ground truth as input.



① 把 $\hat{y}$ 丟入 decoder 的 input vector，和 ground truth

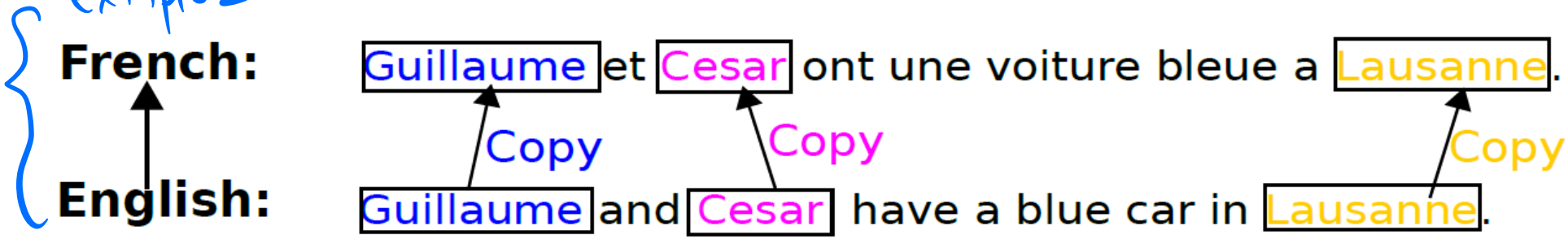
Tips



Copy Mechanism

Machine Translation

example 1



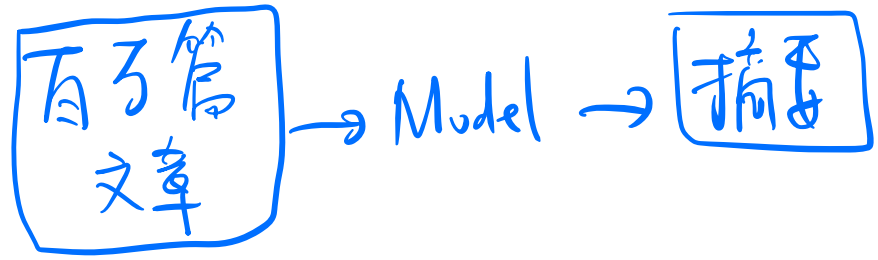
Chat-bot

example 2

User: X寶你好，我是庫洛洛

Machine:庫洛洛你好，很高興認識你

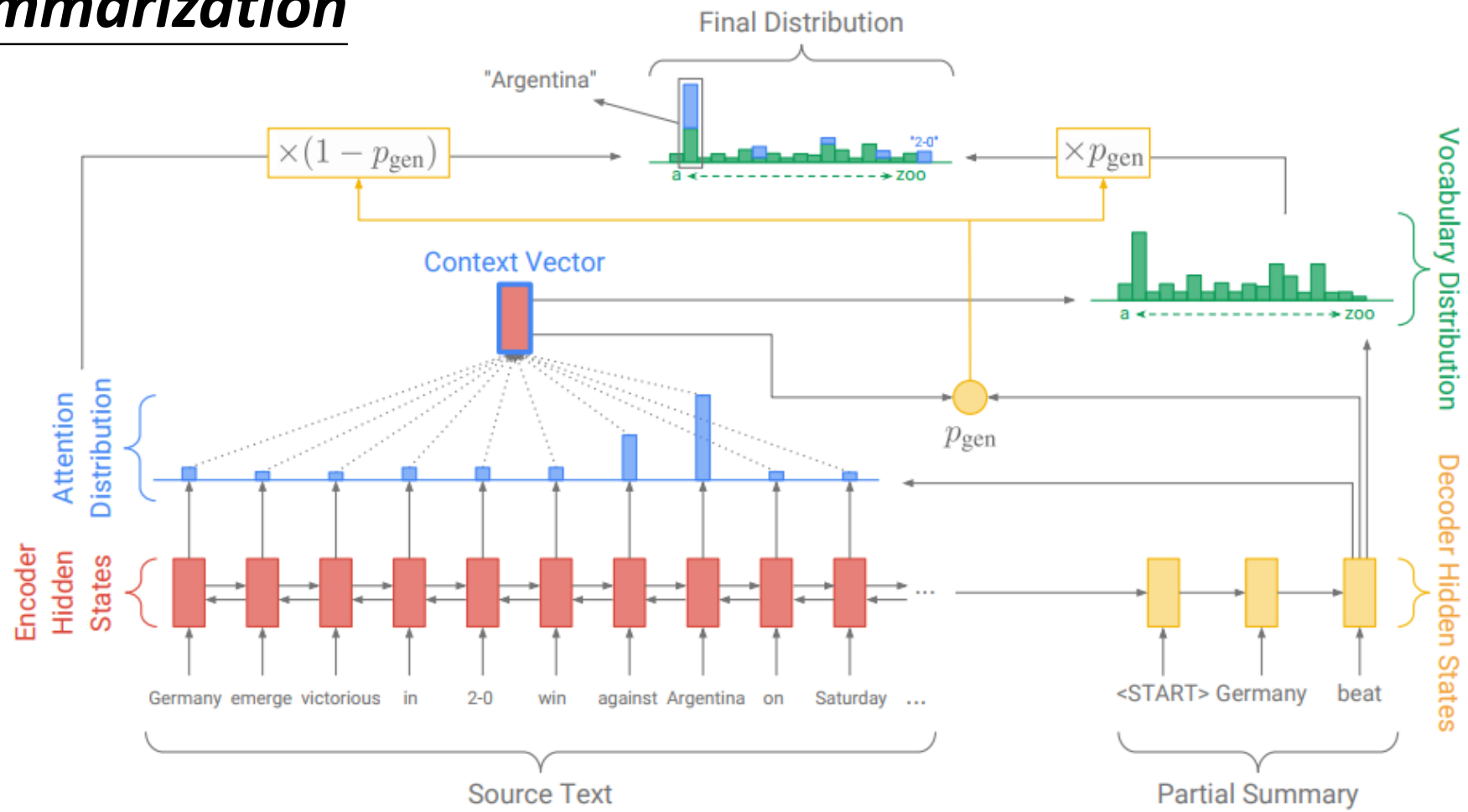
Copy Mechanism



example 3: 摘要 (常需要去copy原文些句子)

https://arxiv.org/abs/1704.04368

Summarization



Copy Mechanism

Pointer Network

https://youtu.be/VdOyqNQ9aww

Incorporating Copying Mechanism in Sequence-to-Sequence Learning

https://arxiv.org/abs/1603.06393

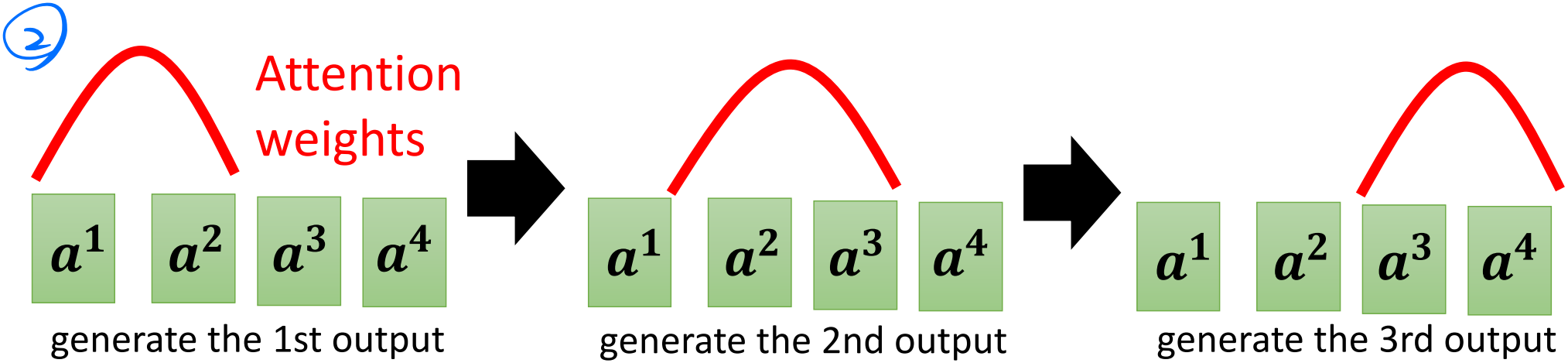
Guided Attention

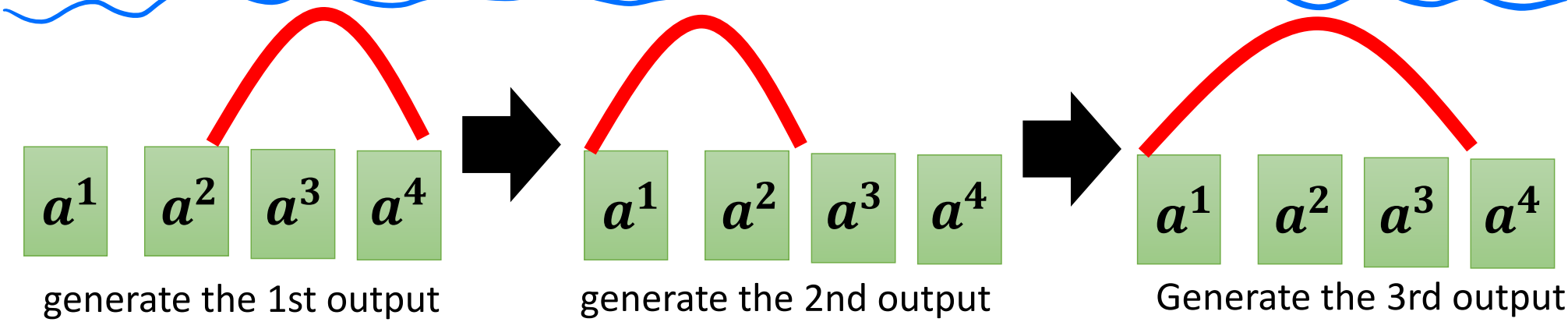
- 高雄發大財我現在要出征
- 發財發財發財發財
- 發財發財發財
- 發財發財
- 發財 (Missing an input character!)

Guided Attention

Monotonic Attention
Location-aware attention

In some tasks, input and output are monotonically aligned.
For example, speech recognition, TTS, etc.





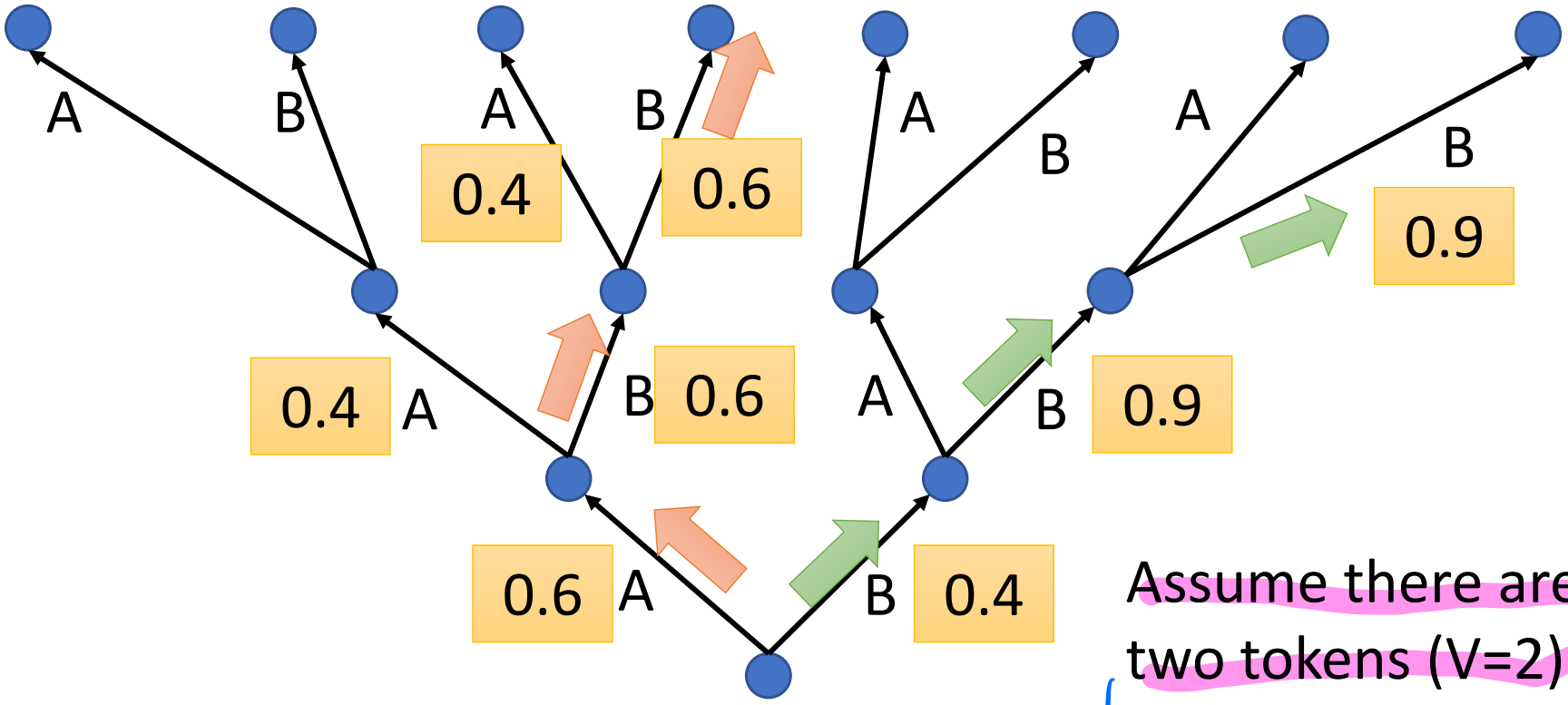
Something wrong!

Beam Search

② The red path is ***Greedy Decoding***.

③ The green path is the best one.

④ Not possible to check all the paths … → Beam Search



Assume there are only two tokens (V=2).

⑤

Beam search有时有用，有时没用，这篇文章告诉你 beam search 有用

Sampling

The Curious Case of Neural Text Degeneration

https://arxiv.org/abs/1904.09751

Context: In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

Beam Search, *b*=32:
"The study, published in the Proceedings of the National Academy of Sciences of the United States of America (PNAS), was conducted by researchers from the Universidad Nacional Autónoma de México (UNAM) and the Universidad Nacional Autónoma de México (UNAM/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de ..."

Pure Sampling:
They were cattle called Bolivian Cavalleros; they live in a remote desert uninterrupted by town, and they speak huge, beautiful, paradisiacal Bolivian linguistic thing. They say, 'Lunch, marge.' They don't tell what the lunch is," director Professor Chuperas Omwell told Sky News. "They've only been talking to scientists, like we're being interviewed by TV reporters. We don't even stick around to be interviewed by TV reporters. Maybe that's how they figured out that they're cosplaying as the Bolivian Cavalleros."

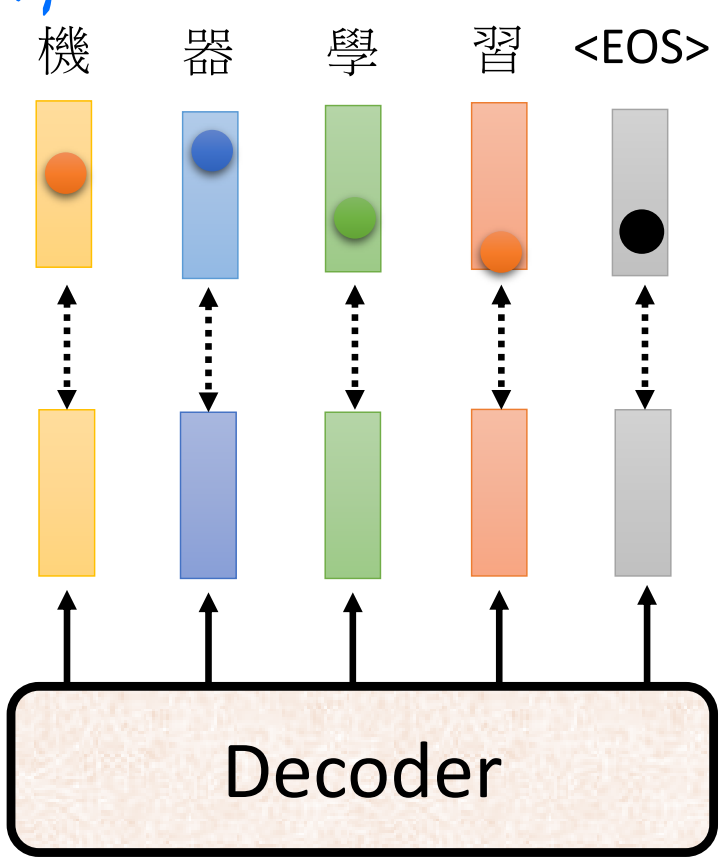
Randomness is needed for decoder when generating sequence in some tasks.

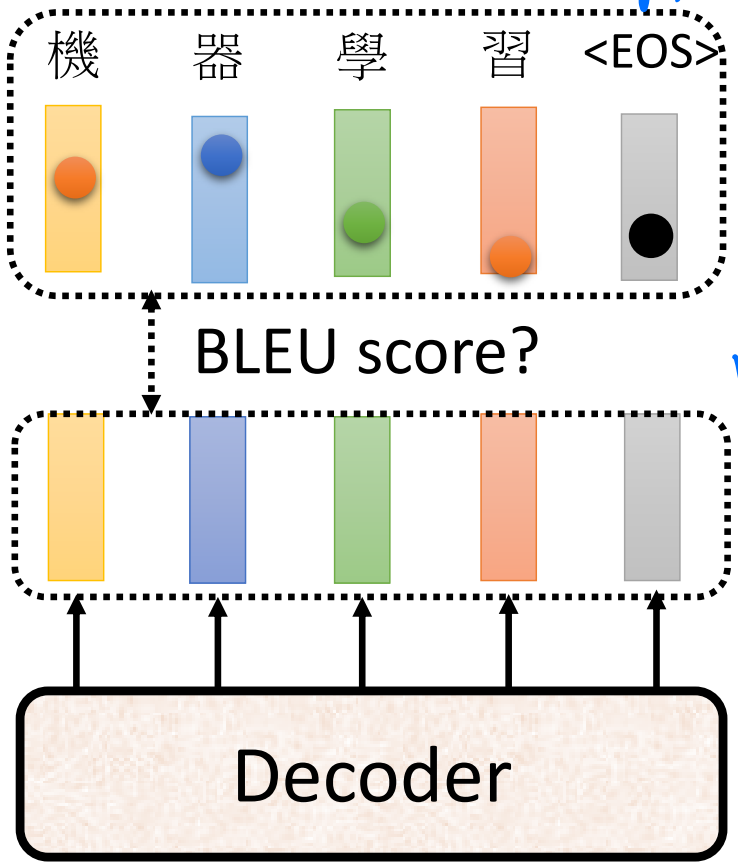
Accept that nothing is perfect. True beauty lies in the cracks of imperfection. ☺

真正的美，就在不完美之中

Optimizing Evaluation Metrics?

作業中，最終比較模型好壞的 metric 是 BLEU，但 training 卻用 cross-entropy，why？





BLEU不能算微分，不能做 backpropagation 只能當 validation set 的 metric 來 tune model，對於這種不能當 loss 來 optimize 的問題

How to do the optimization?

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

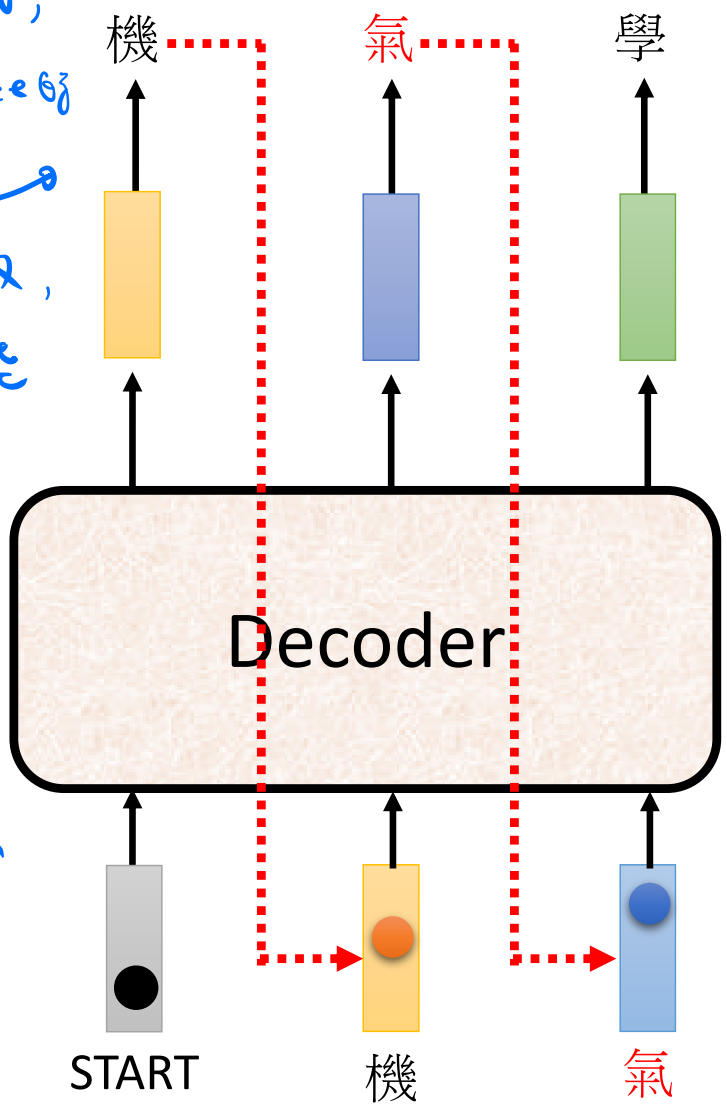
把 BLEU 當 reward，decoder 當 agent，硬 train 一發

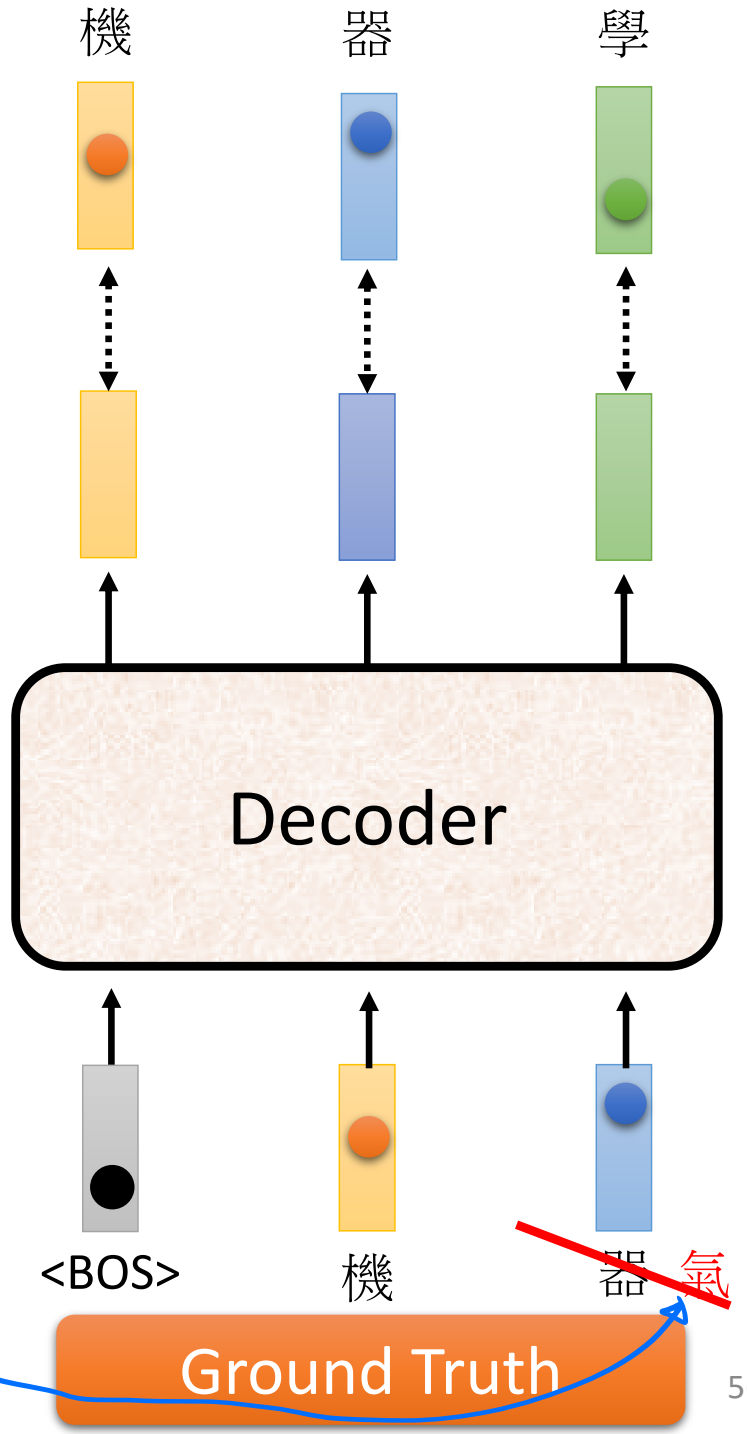
There is a mismatch! ☹

exposure bias

training 的，input 到 decoder 的都是 ground truth

但 inference 的是這樣做，那就可能一步錯步步錯





那，解決方法是，training 的時候故意加一些錯誤的 noise

Scheduled Sampling

- Original Scheduled Sampling

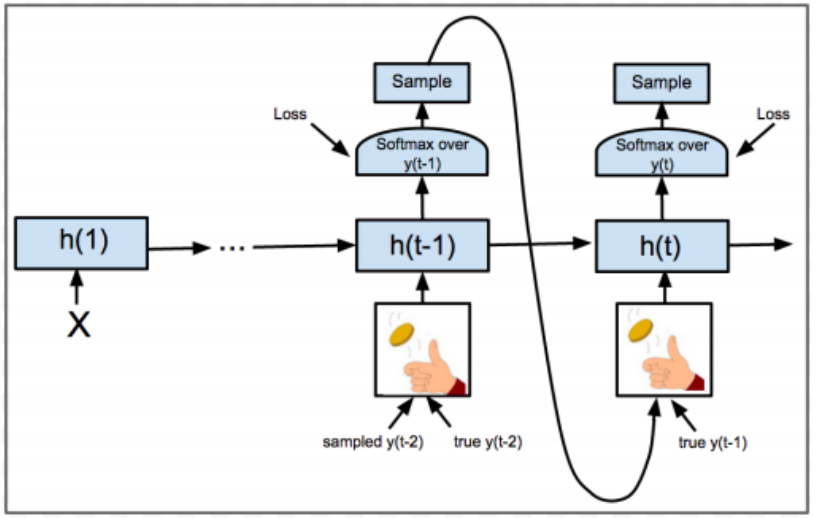
  https://arxiv.org/abs/1506.03099

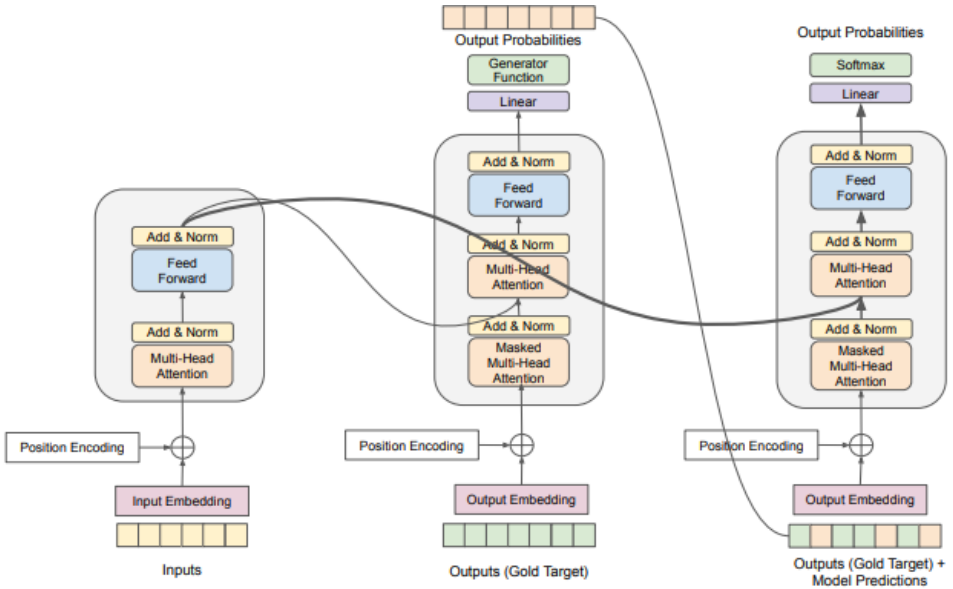
- Scheduled Sampling for Transformer

  https://arxiv.org/abs/1906.07651

- Parallel Scheduled Sampling

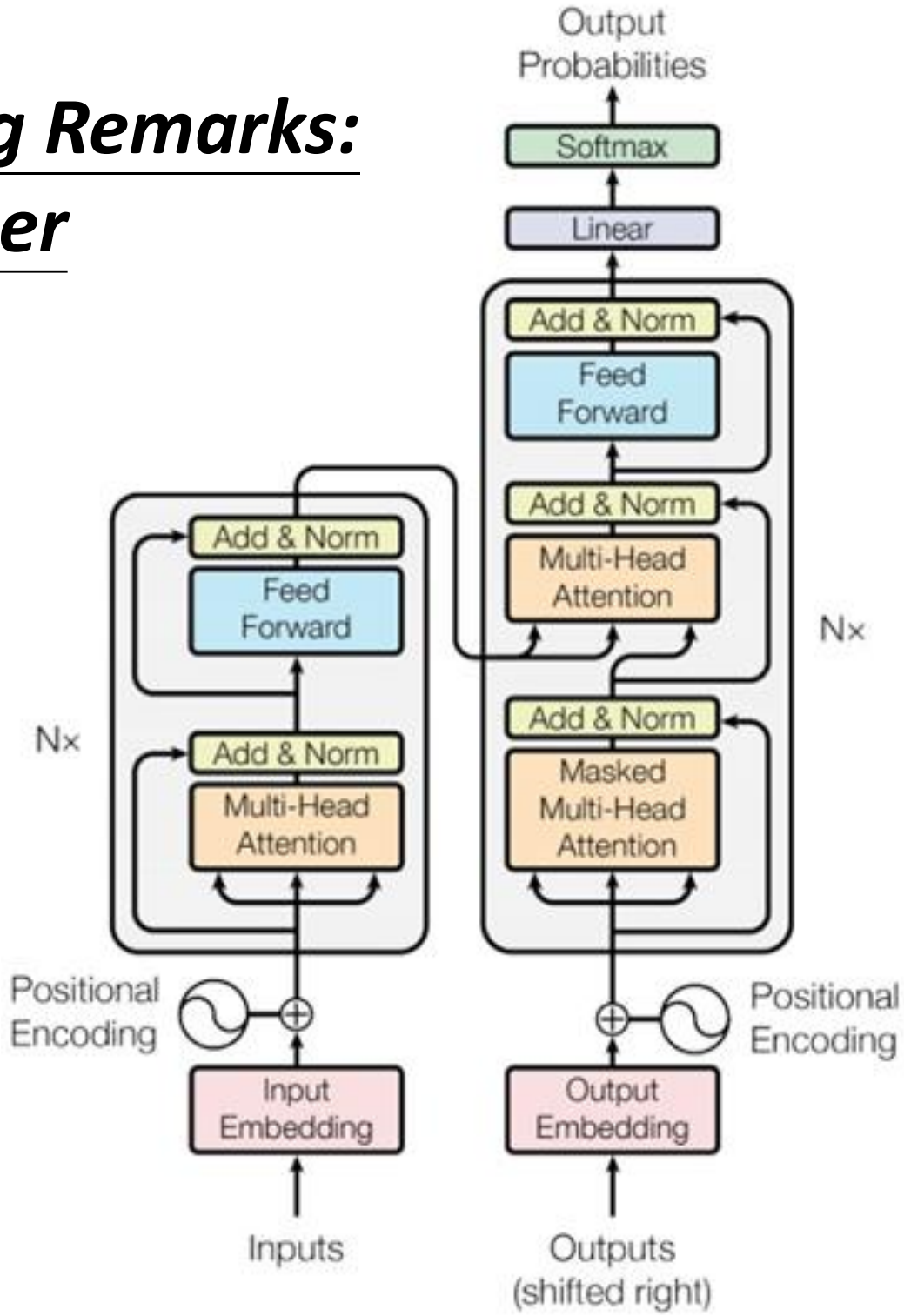
  https://arxiv.org/abs/1906.04331





Schedule Sampling

Concluding Remarks: Transformer




Q&A