Transformer

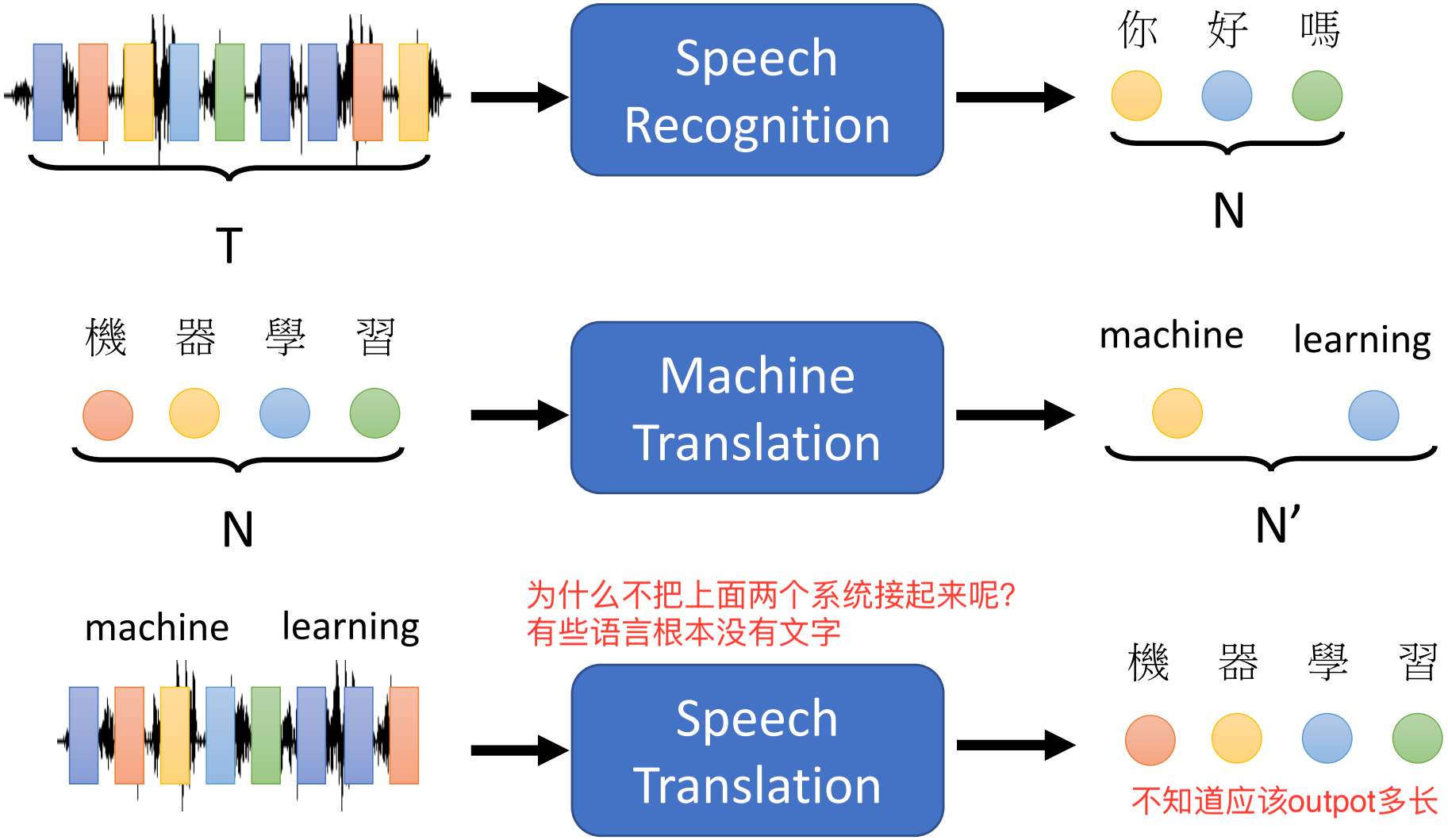
李宏毅

Hung-yi Lee

Sequence-to-sequence (Seq2seq)

Input a sequence, output a sequence　Transformer 其实就是一个 Seq2Seq的model

The output length is determined by model.



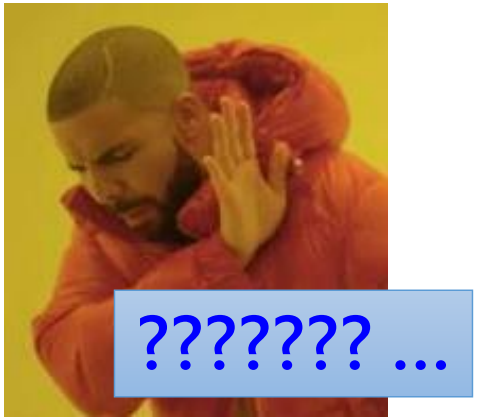
Language without text

Hokkien (閩南語、台語)

(Speech of Hokkien)

Speech Recognition

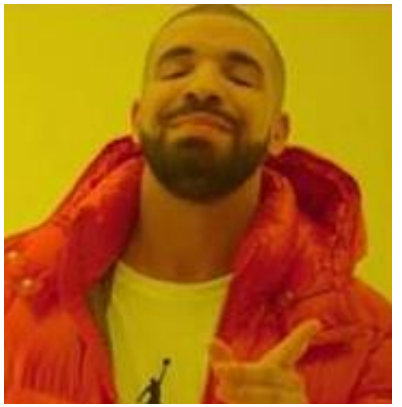
“母湯”



(Speech of Hokkien)

Speech Translation

“不行”



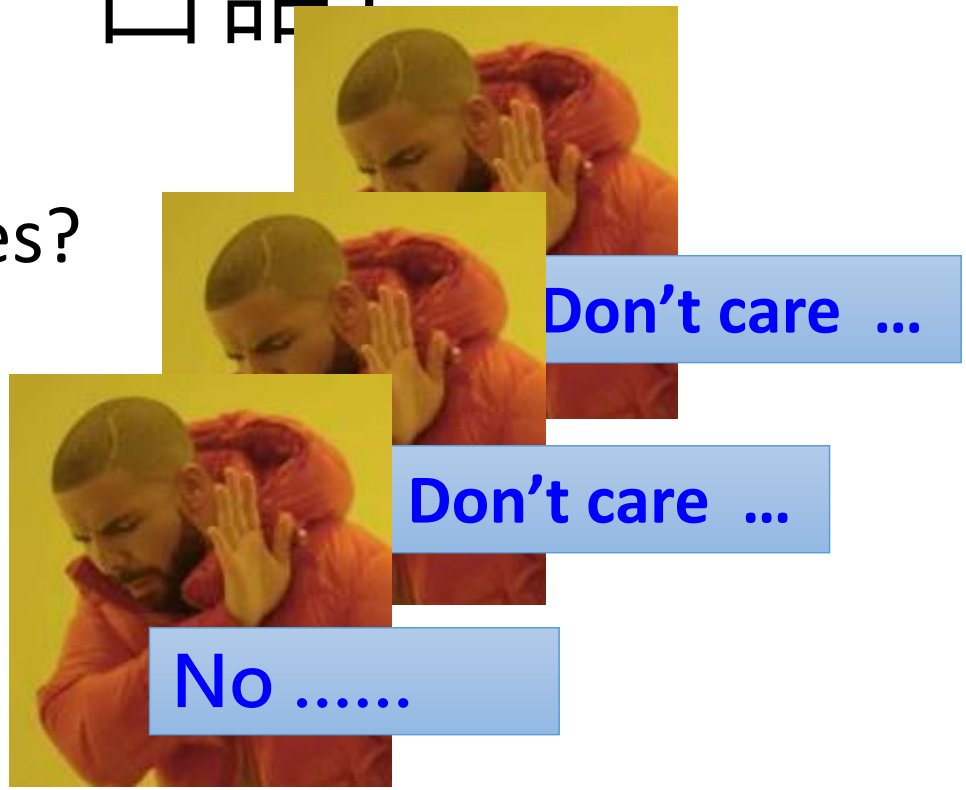
乡土剧，台语语音，中文字幕

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

Using 1500 hours of data for training

Hokkien (閩南語、台語)

- Background music & noises?
- Noisy transcriptions?
  字幕和语音没有对应上
- 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)

你 的 身 體 撐 不 住

沒 事 你 為 什 麼 要 請 假

要 生 了 嗎　Answer:不 會 膩 嗎

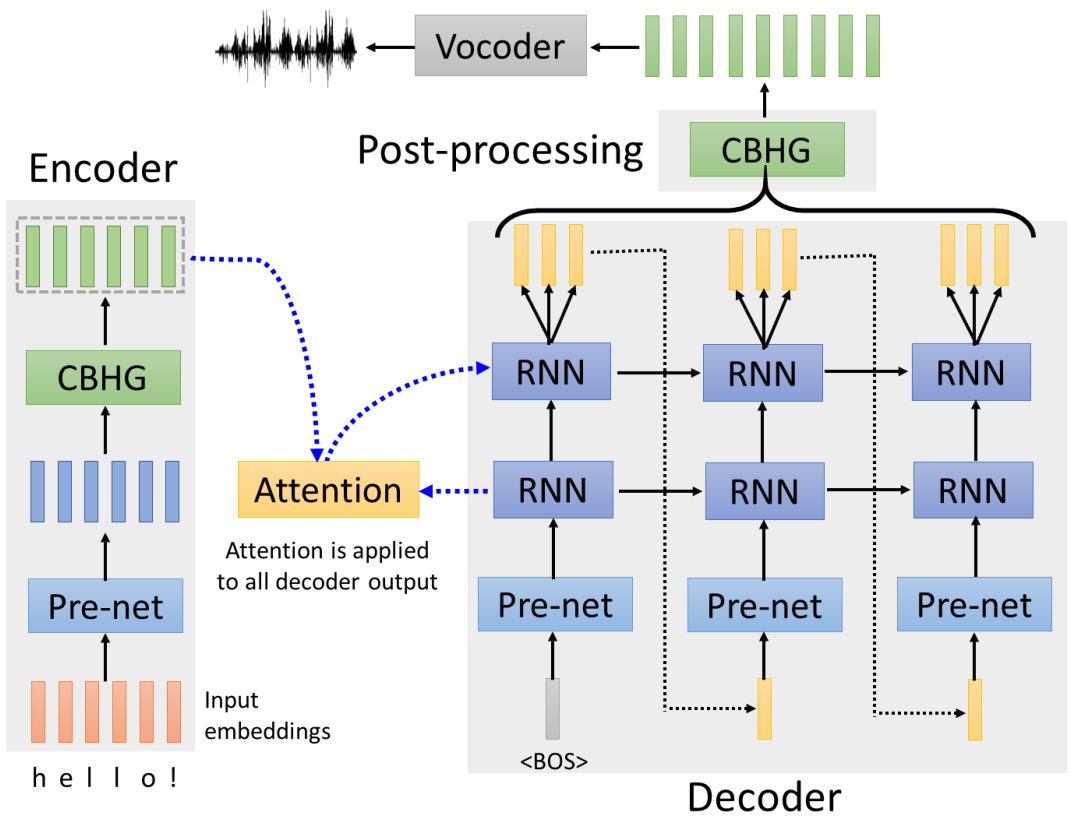
我 有 幫 廠 長 拜 託

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

感謝張凱為同學提供實驗結果



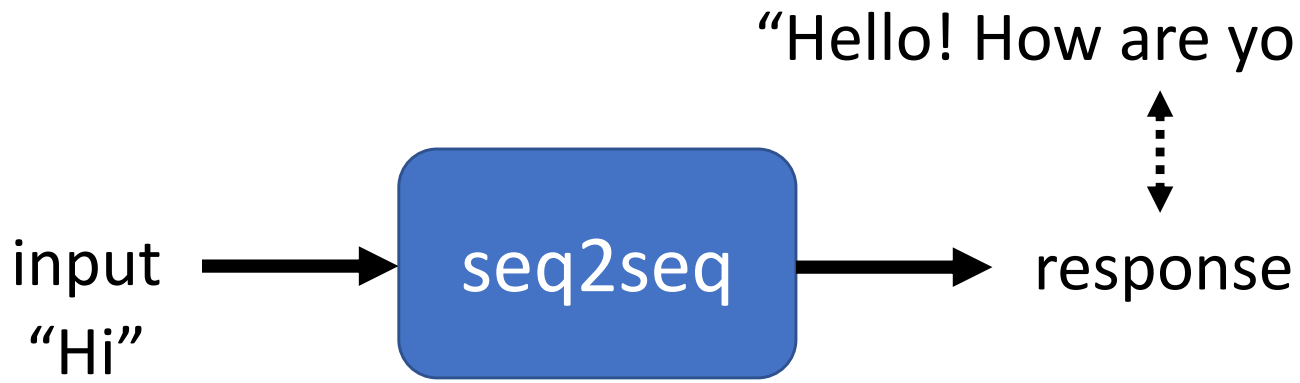
Taiwanese Speech Synthesis

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.

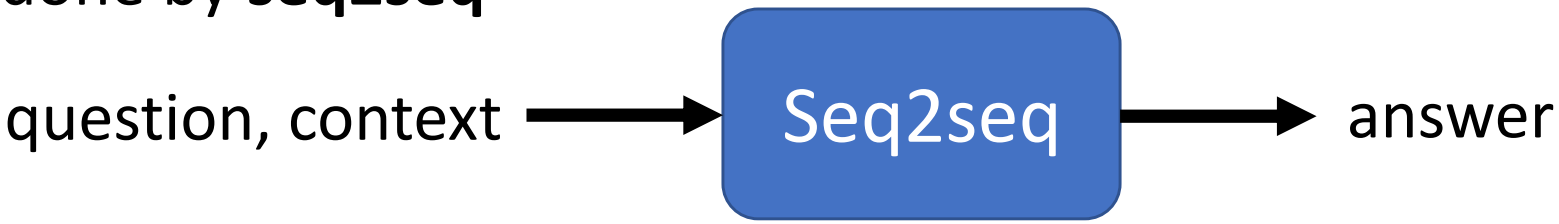
Most Natural Language Processing applications …

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

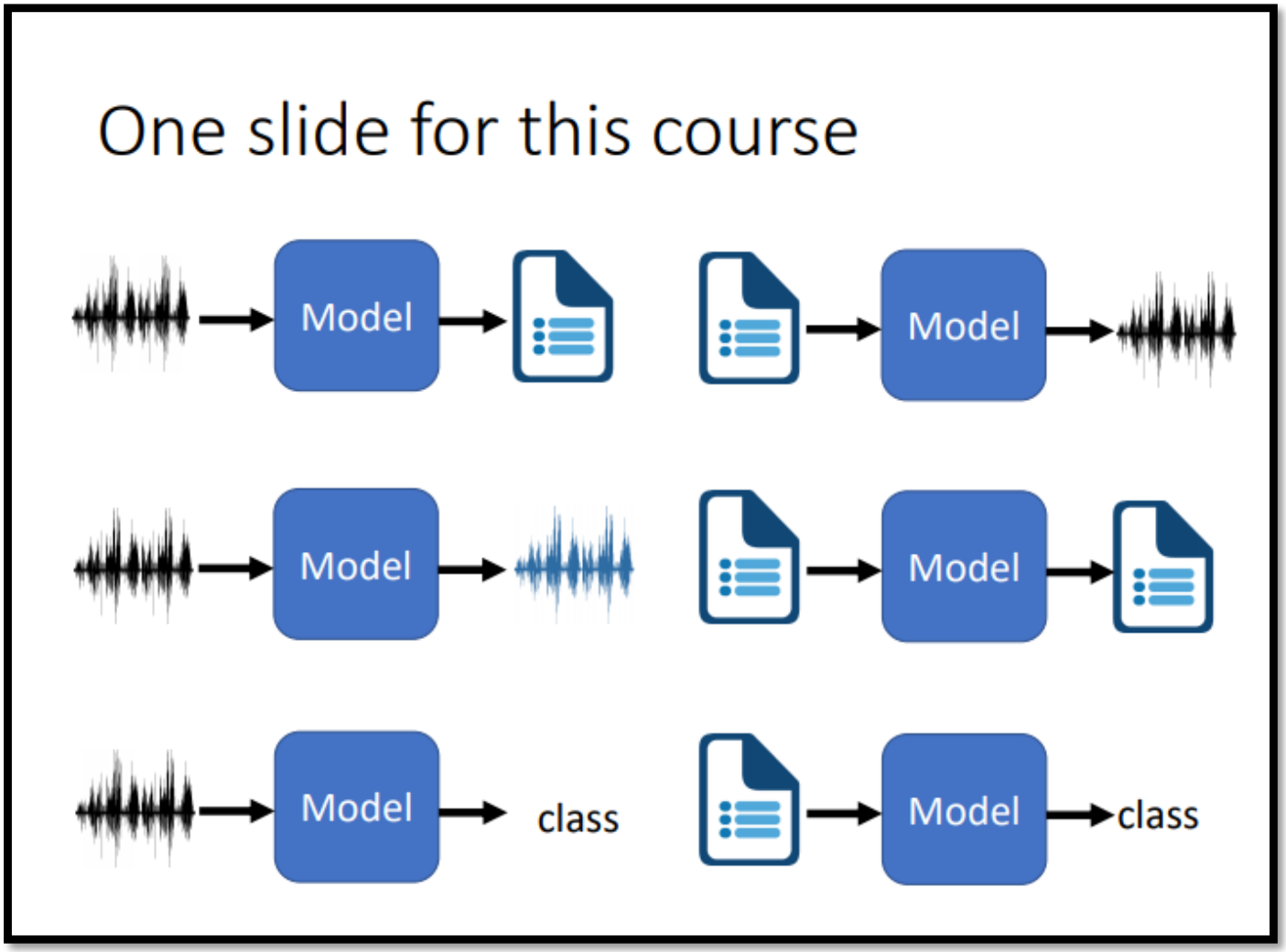
decaNLP

QA can be done by **seq2seq**



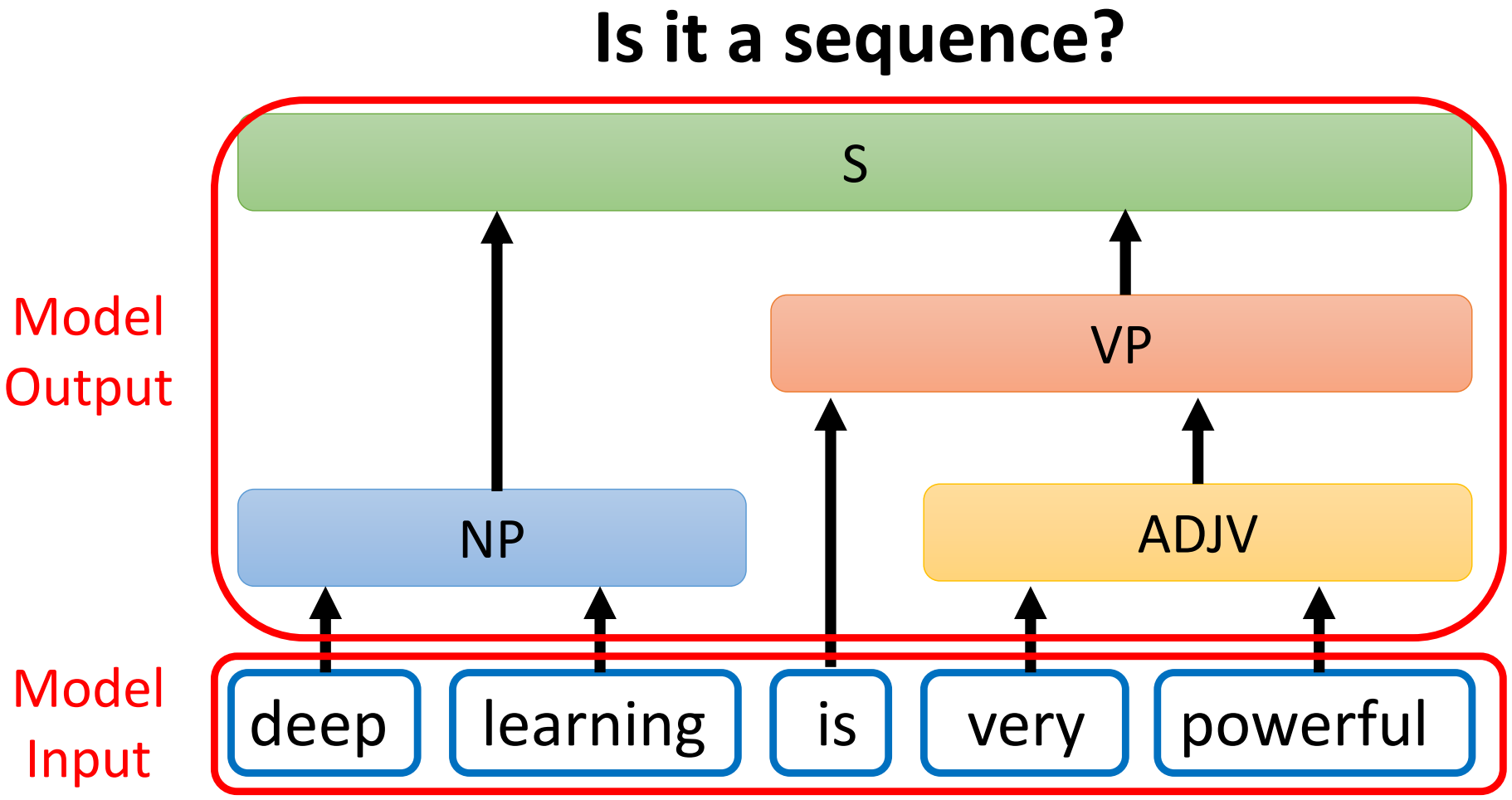
https://arxiv.org/abs/1806.08730
https://arxiv.org/abs/1909.03329

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

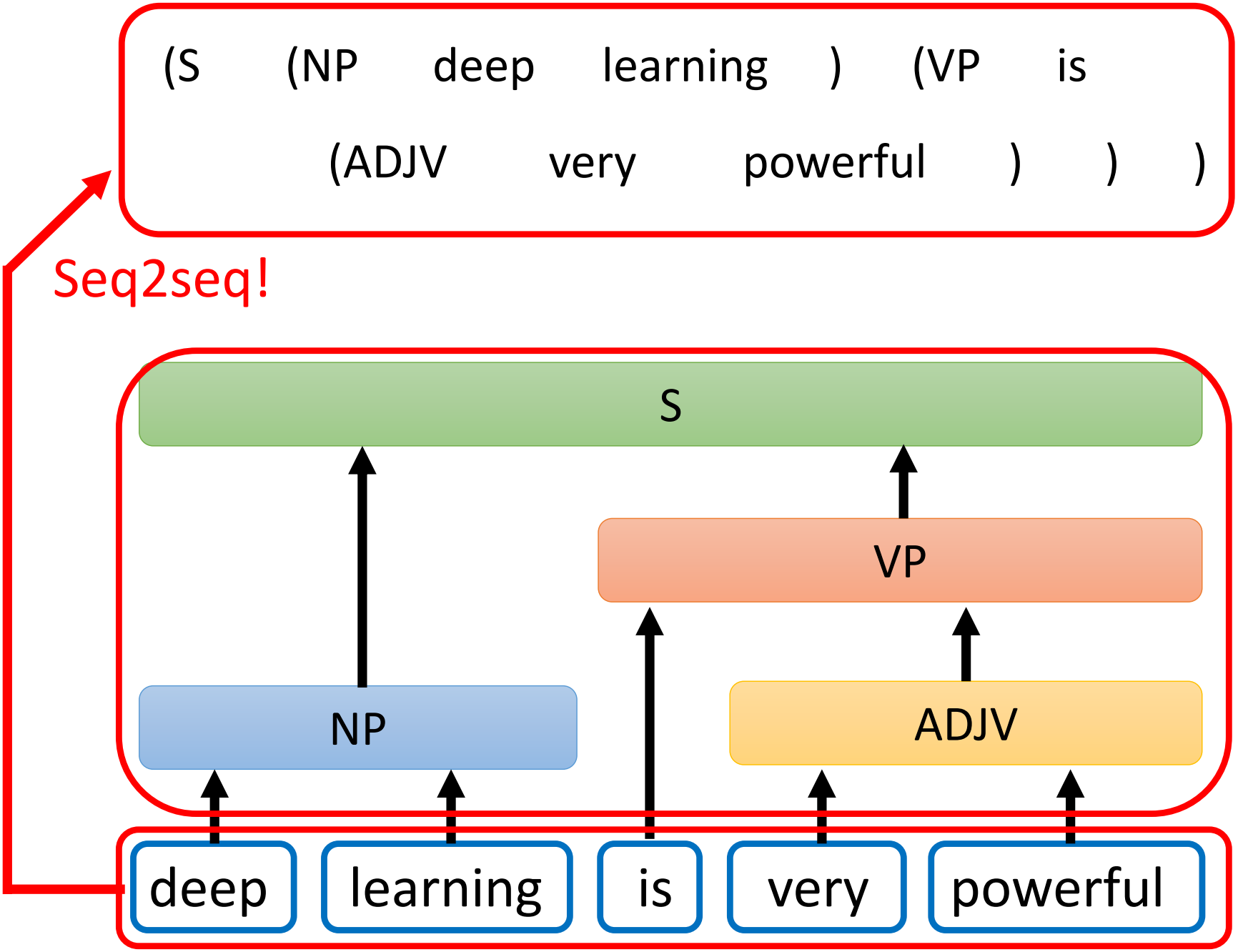


Source webpage: https://speech.ee.ntu.edu.tw/~hylee/dlhlp/2020-spring.html

Seq2seq for Syntactic Parsing



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

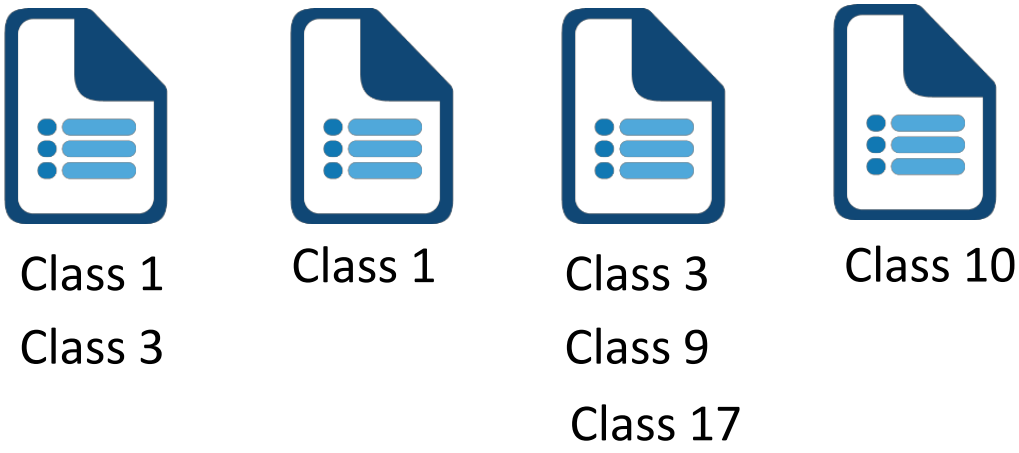
https://arxiv.org/abs/1412.7449

deep learning is very powerful

Seq2seq for Multi-label Classification

c.f. Multi-class Classification

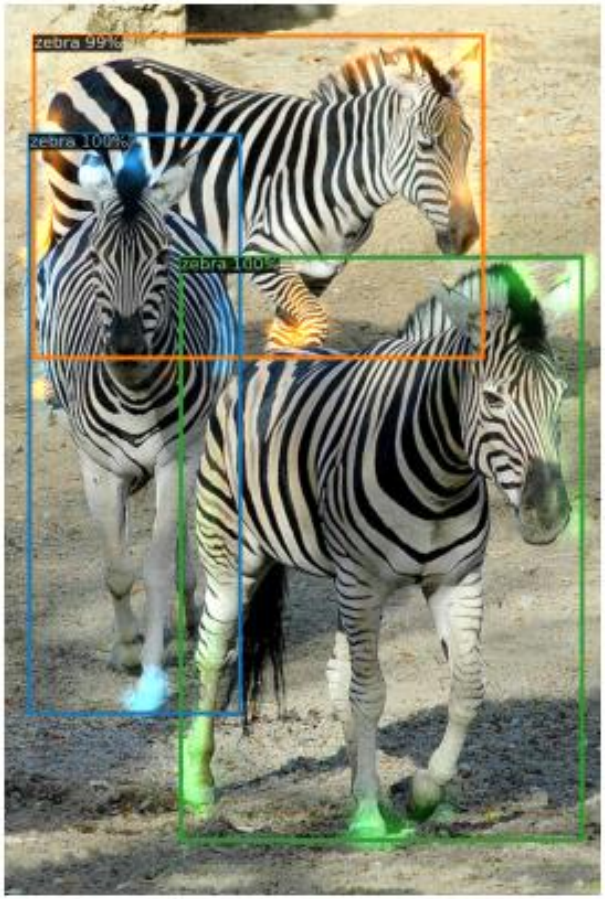
An object can belong to multiple classes.

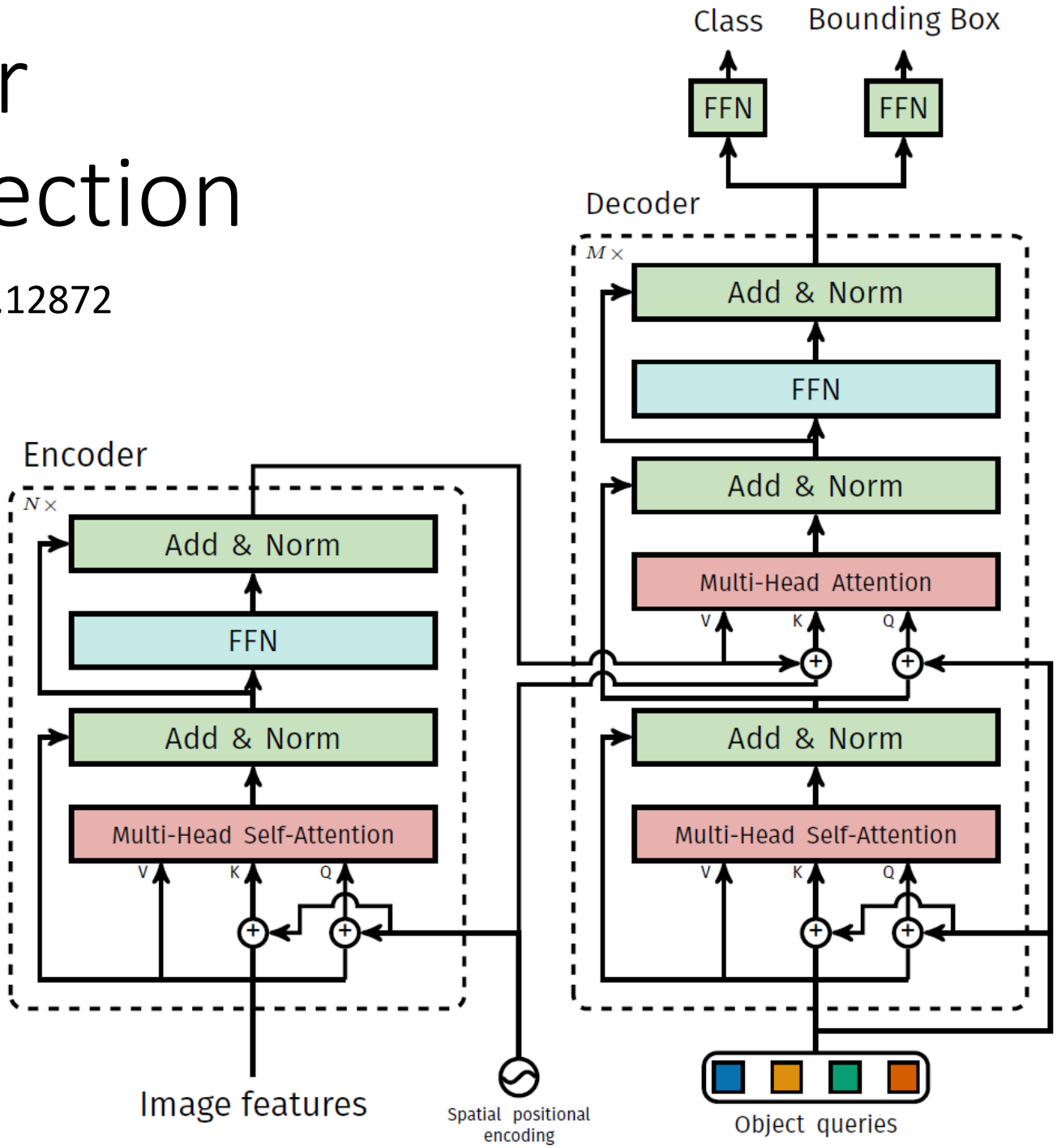


https://arxiv.org/abs/1909.03434
https://arxiv.org/abs/1707.05495

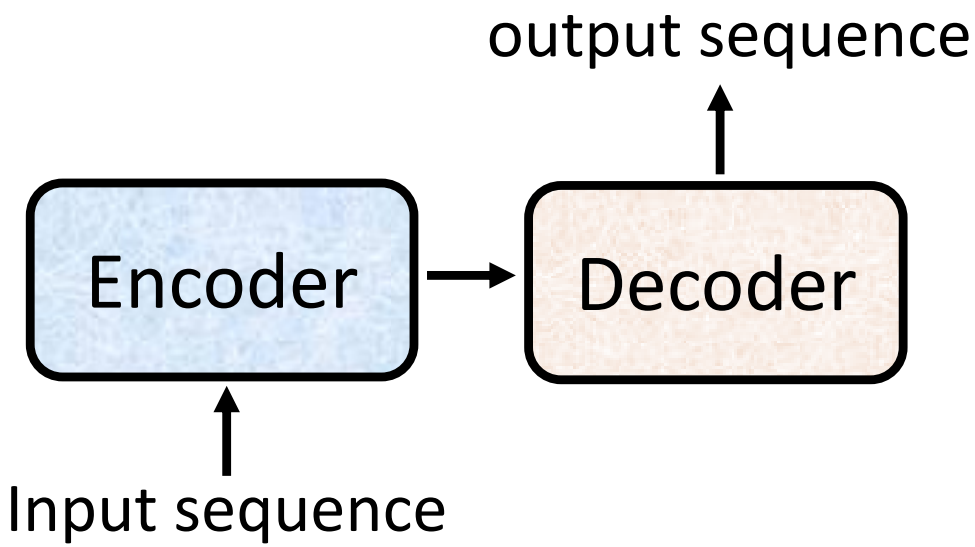
Seq2seq for Object Detection

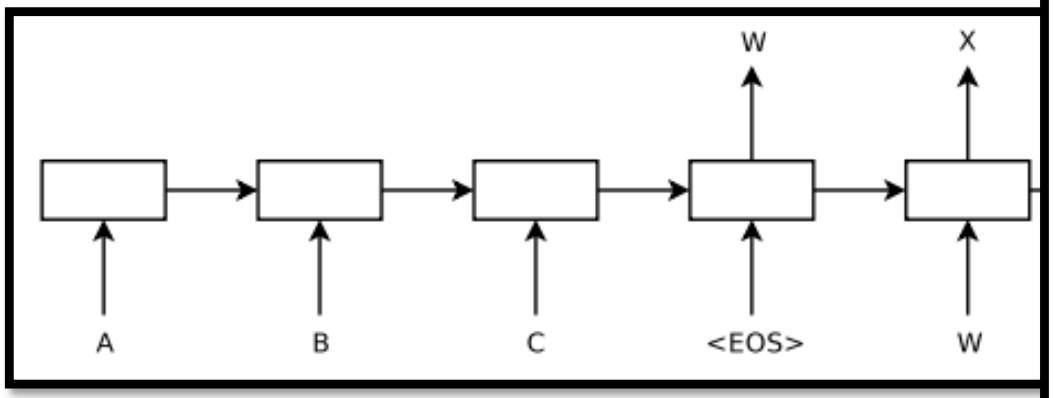
https://arxiv.org/abs/2005.12872





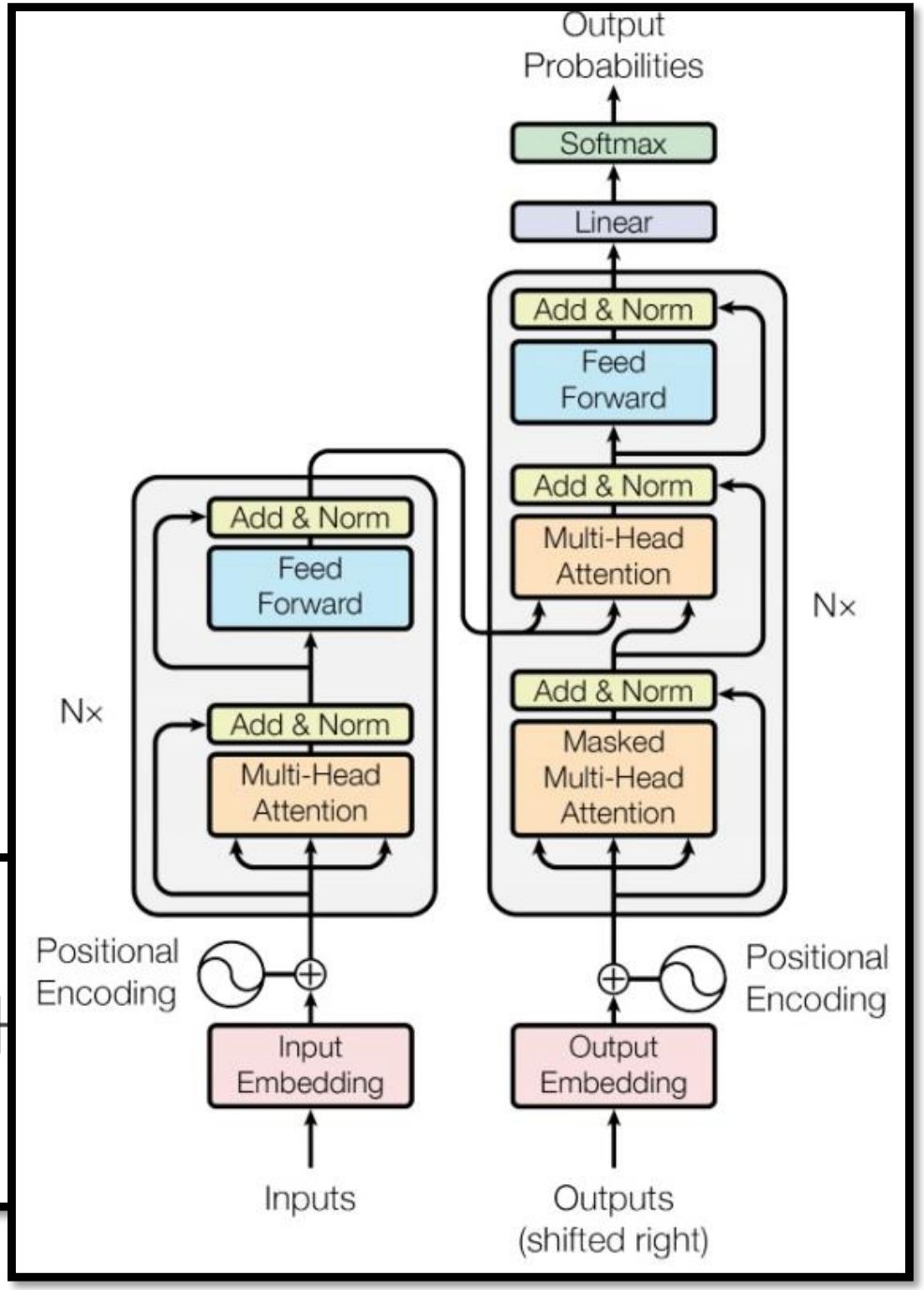
Seq2seq





Sequence to Sequence Learning with Neural Networks

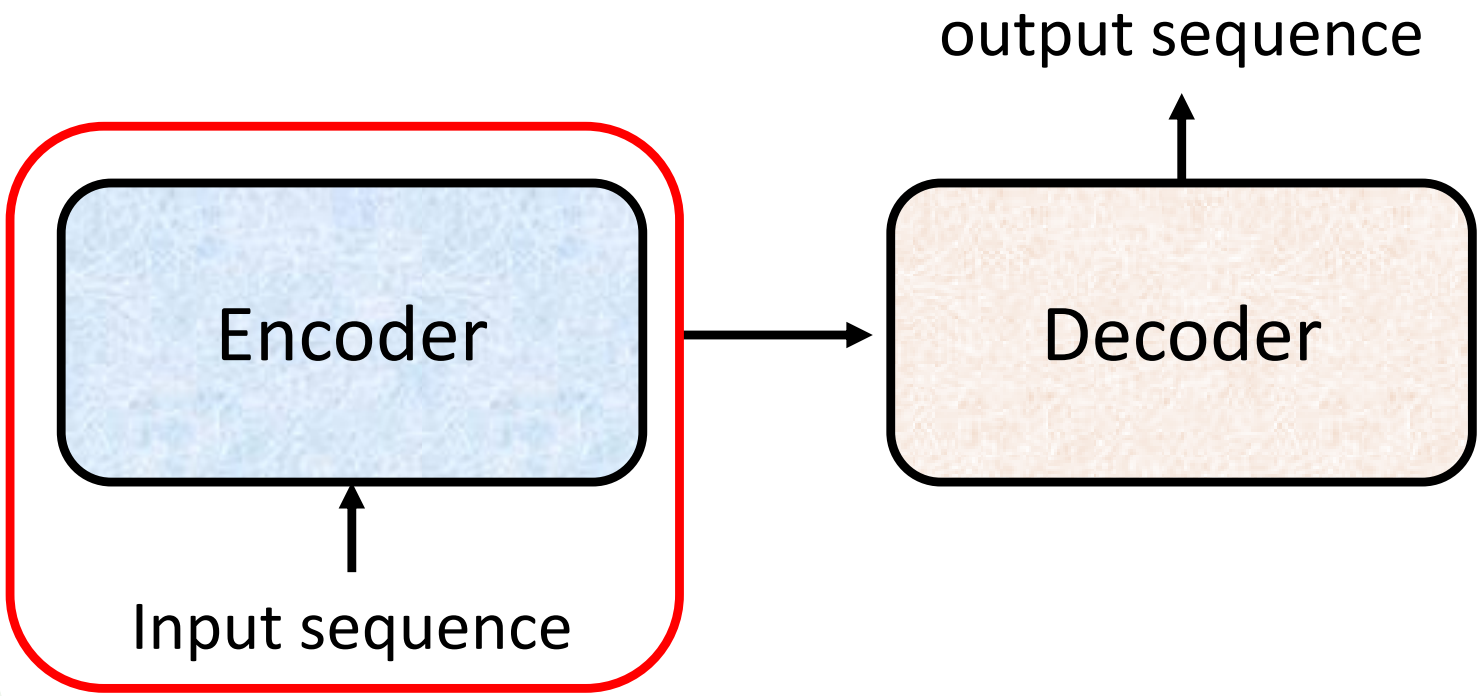
https://arxiv.org/abs/1409.3215



Transformer

https://arxiv.org/abs/1706.03762

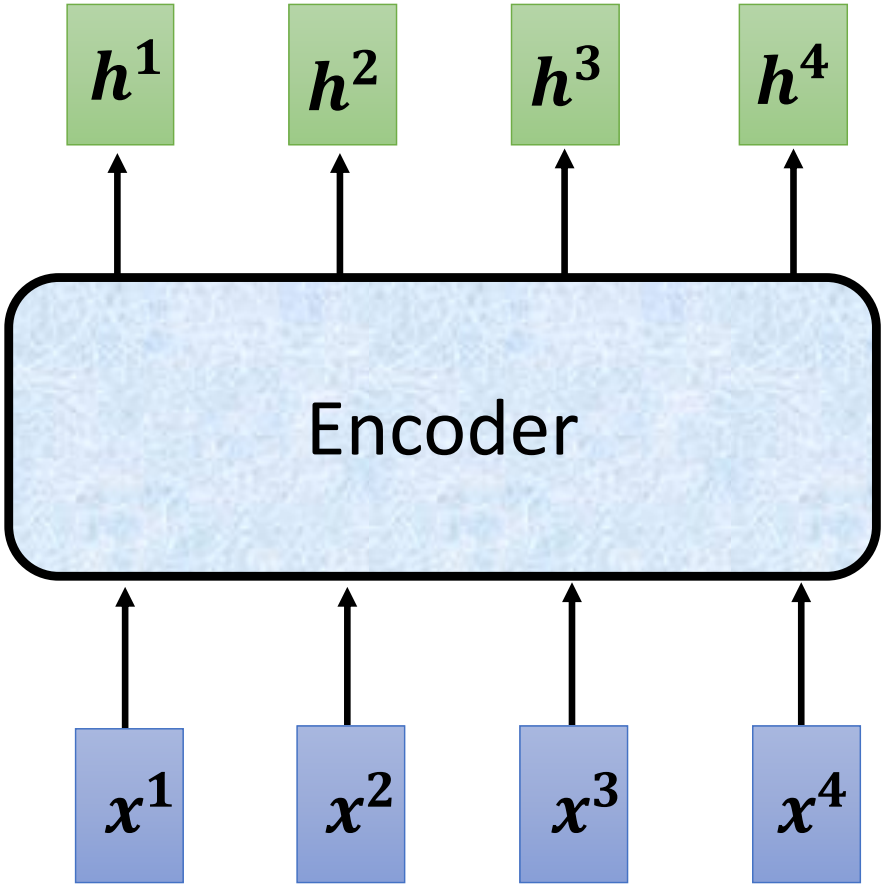
Encoder



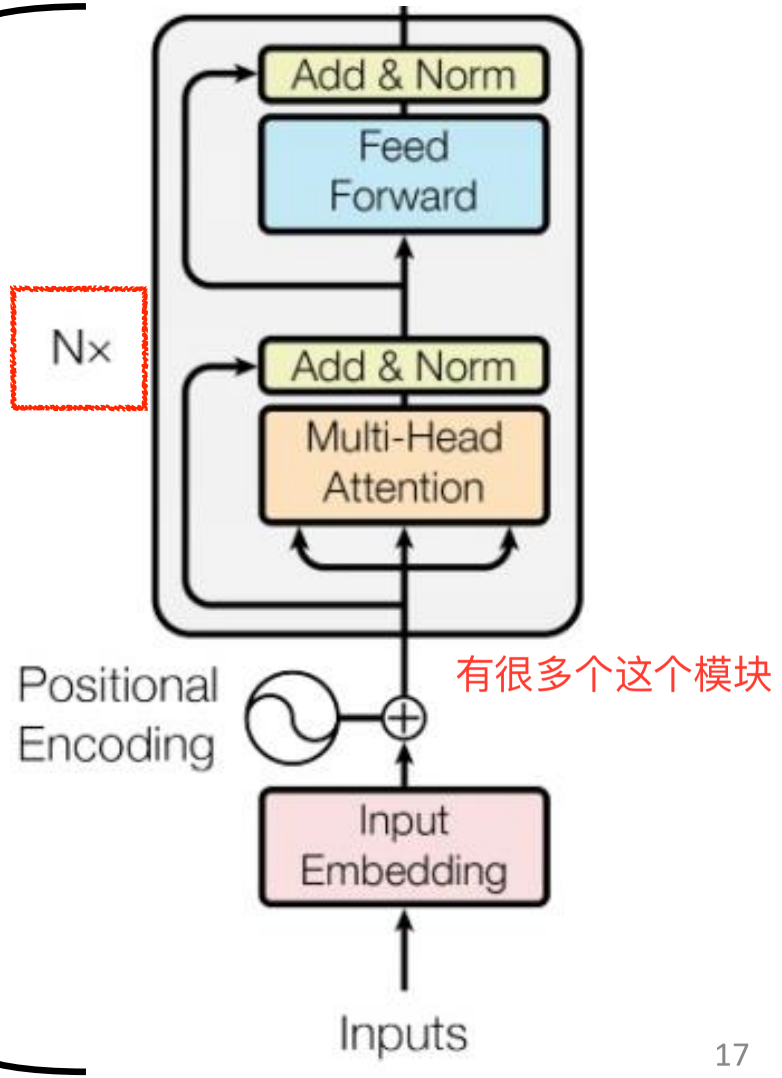
Encoder

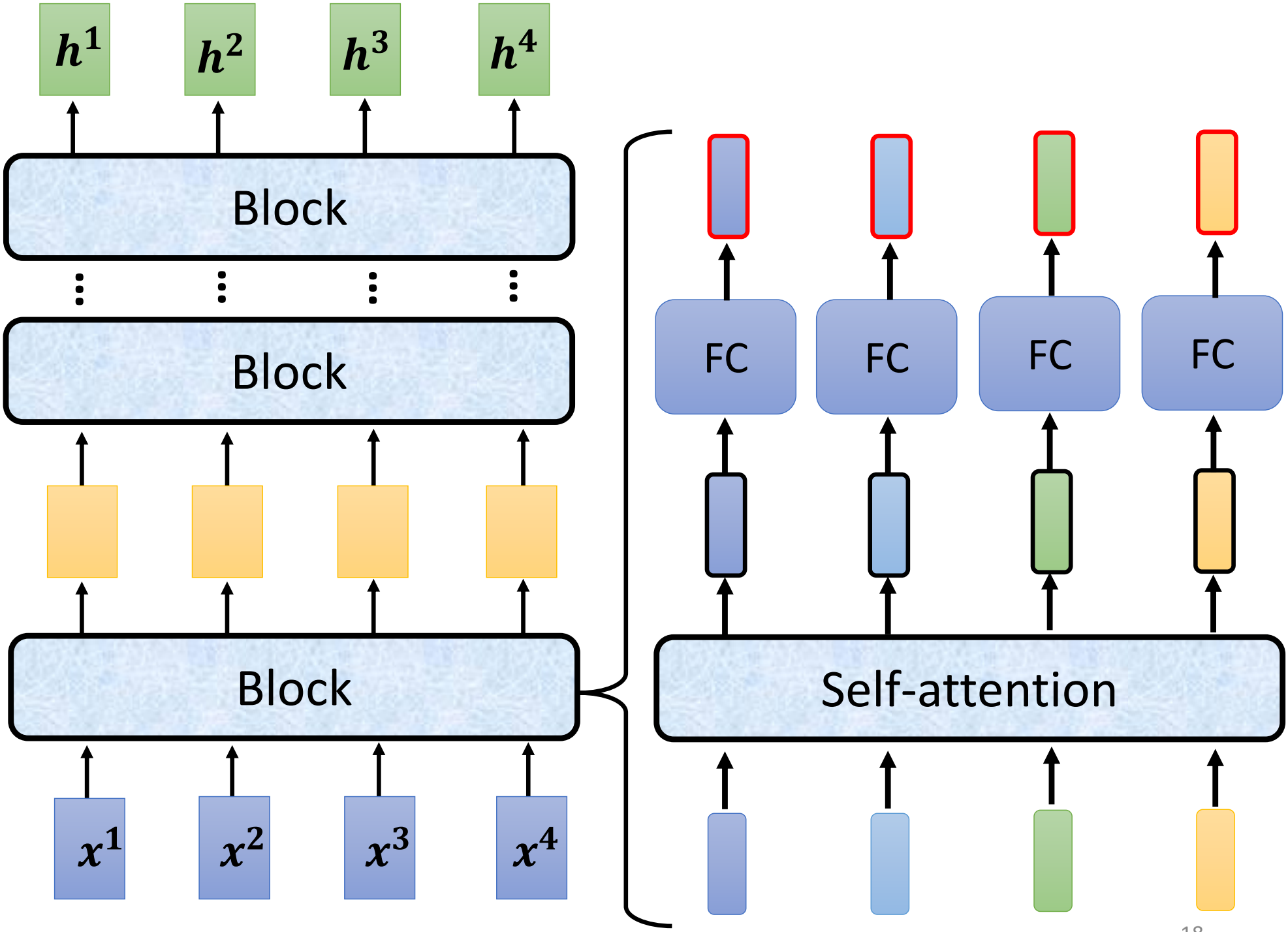
给一排向量，输出一排向量

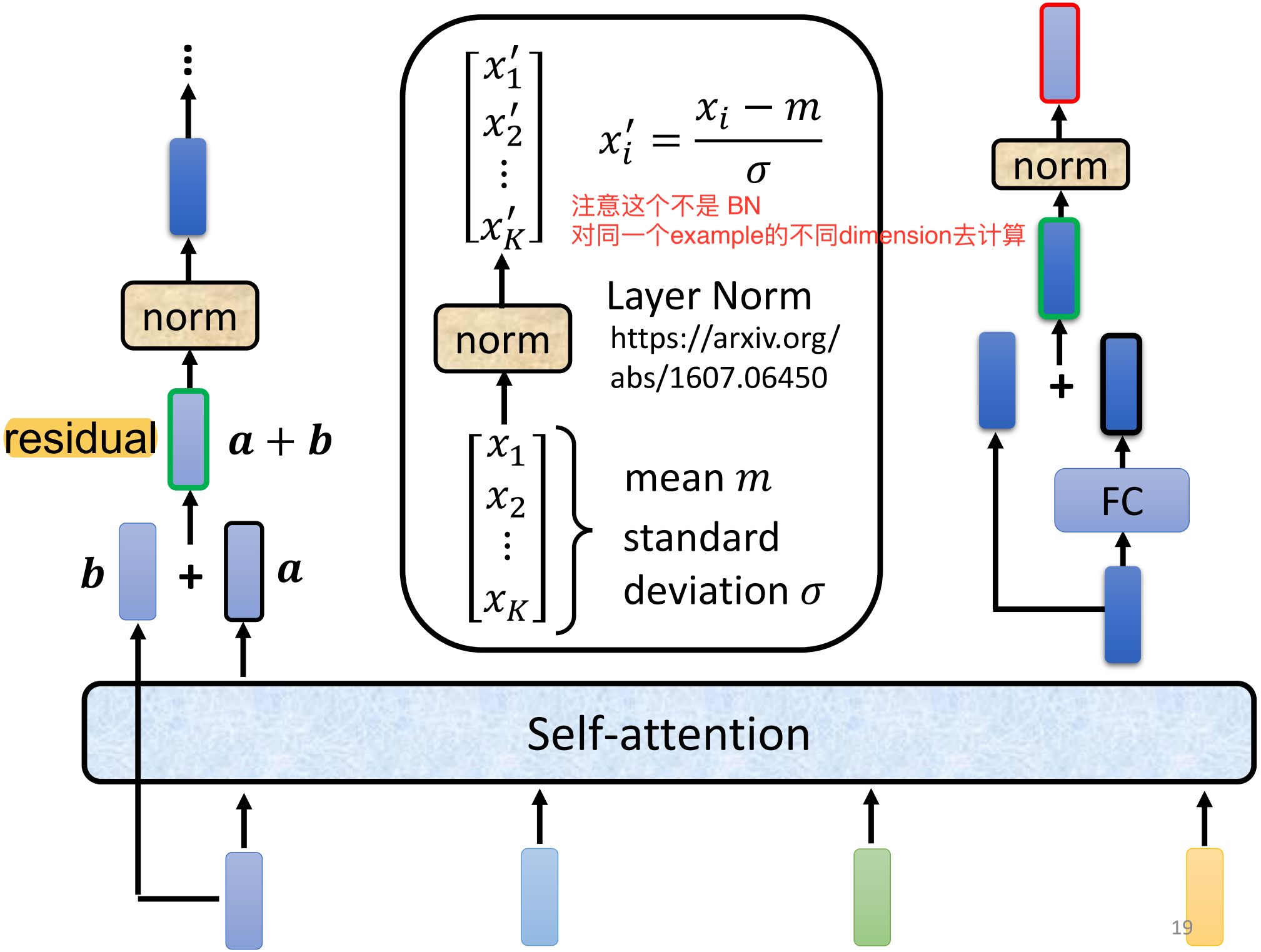
You can use **RNN** or **CNN**.



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
$x'_i = \frac{x_i - m}{\sigma}$
注意这个不是 BN
对同一个example的不同dimension去计算
norm
Layer Norm
https://arxiv.org/abs/1607.06450
mean m
standard deviation σ
norm
+
FC

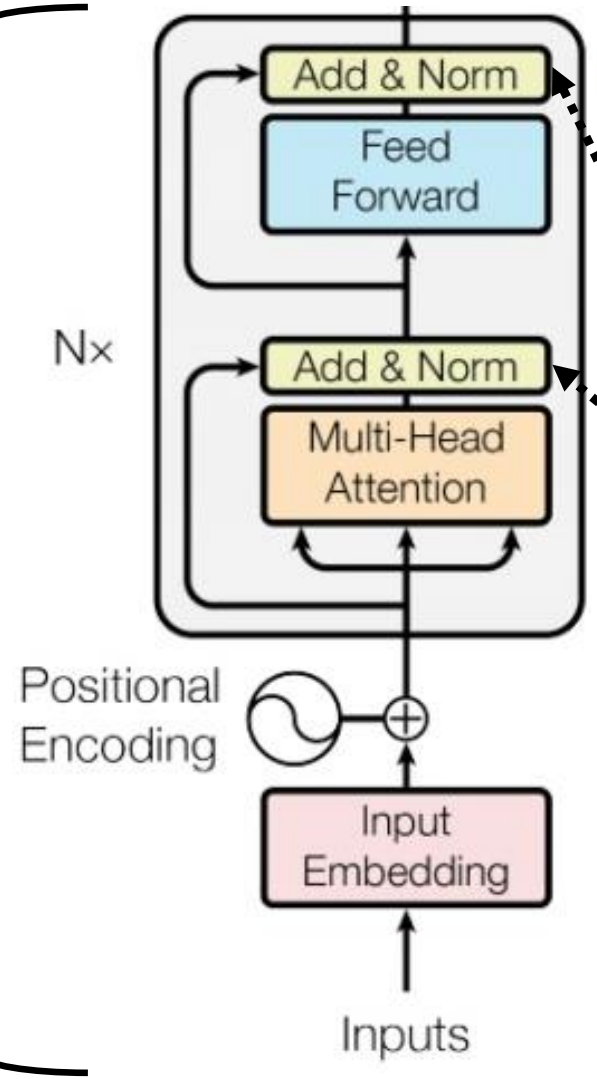
BERT

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

h^1 h^2 h^3 h^4

Encoder

x^1 x^2 x^3 x^4



Residual + Layer norm

To learn more ……

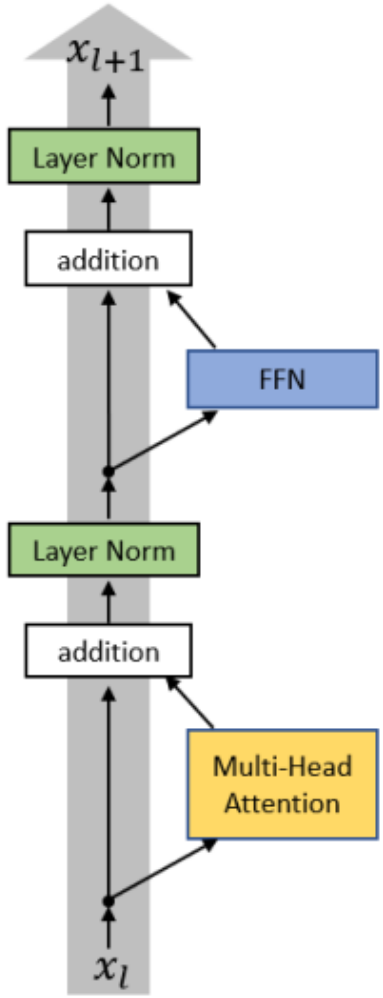
原论文的设计不一定最好

- 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

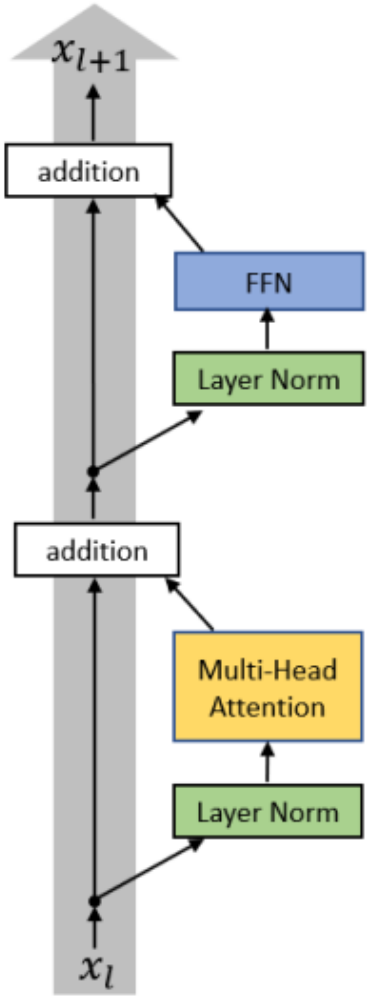
这个论文也很好，它首先讲了为什么在 Transformer中 BN 不如 LN

原论文的设计



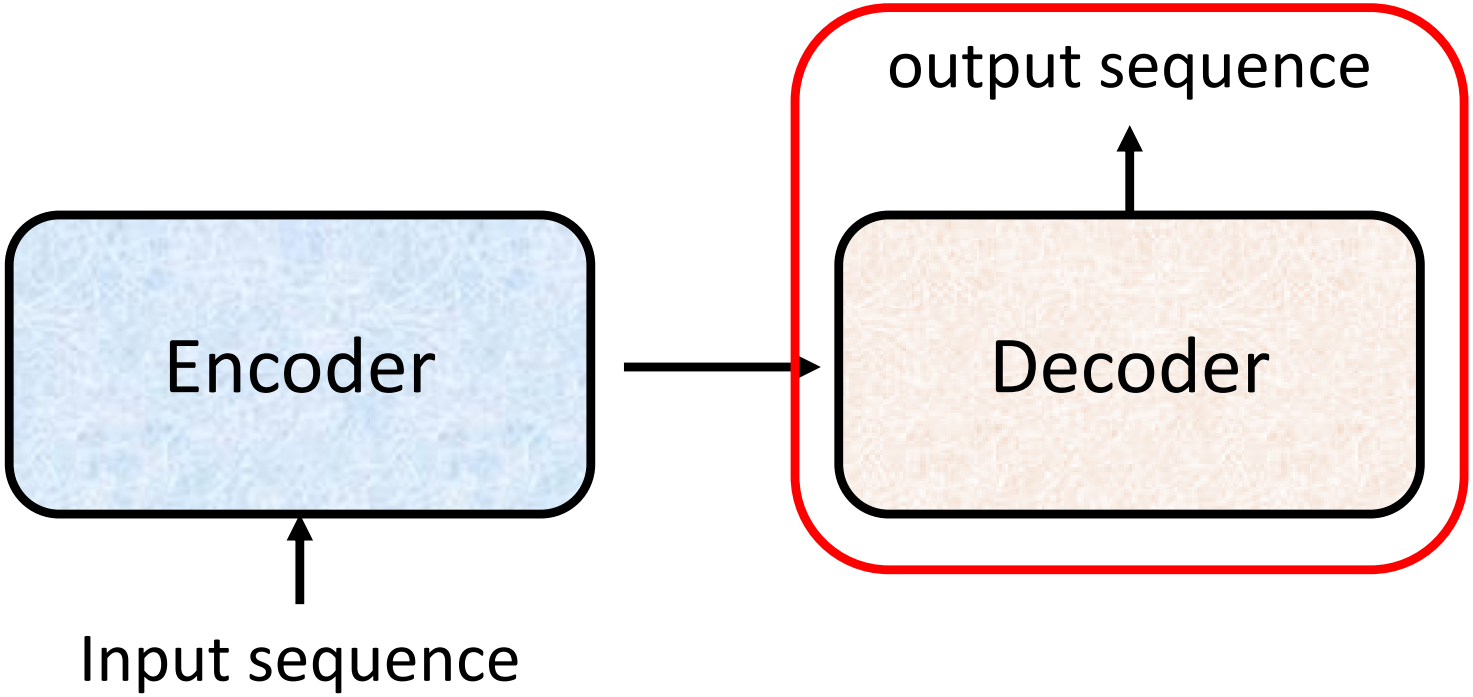
(a)

新的设计



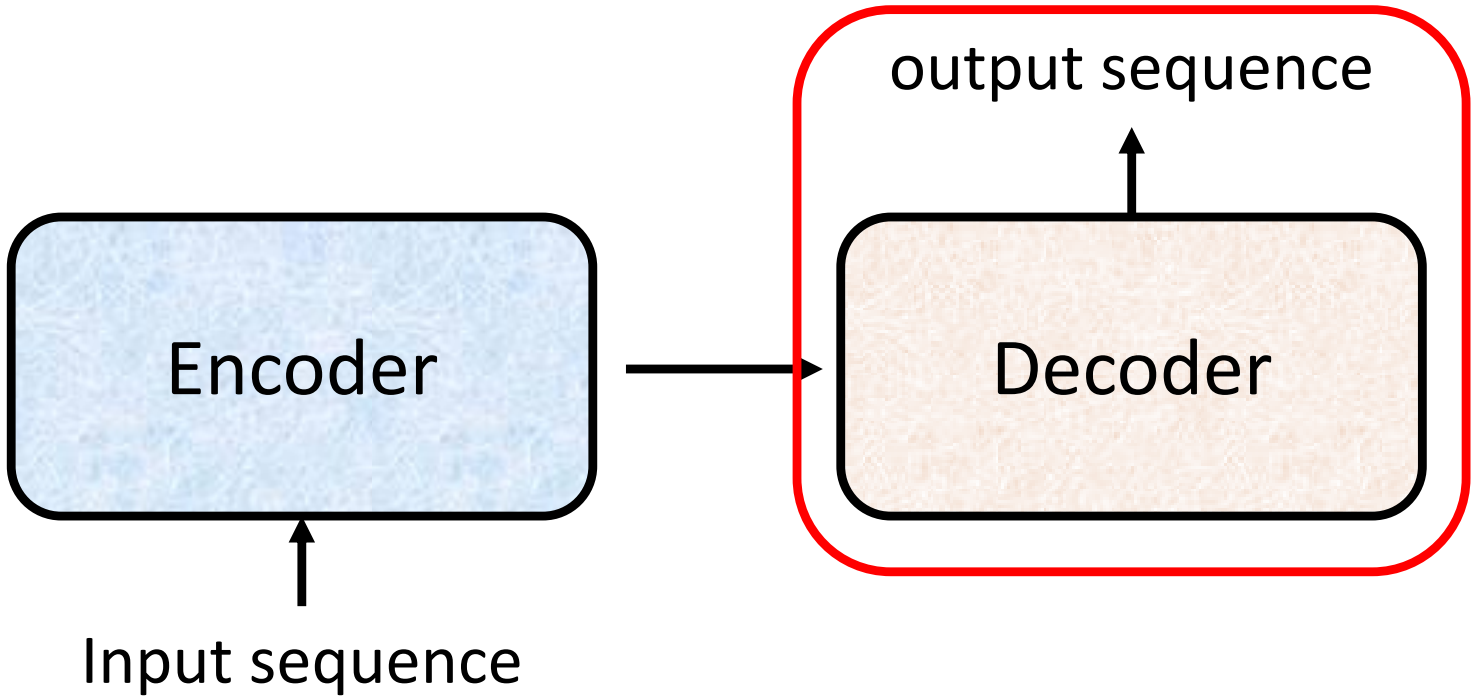
(b)

Decoder



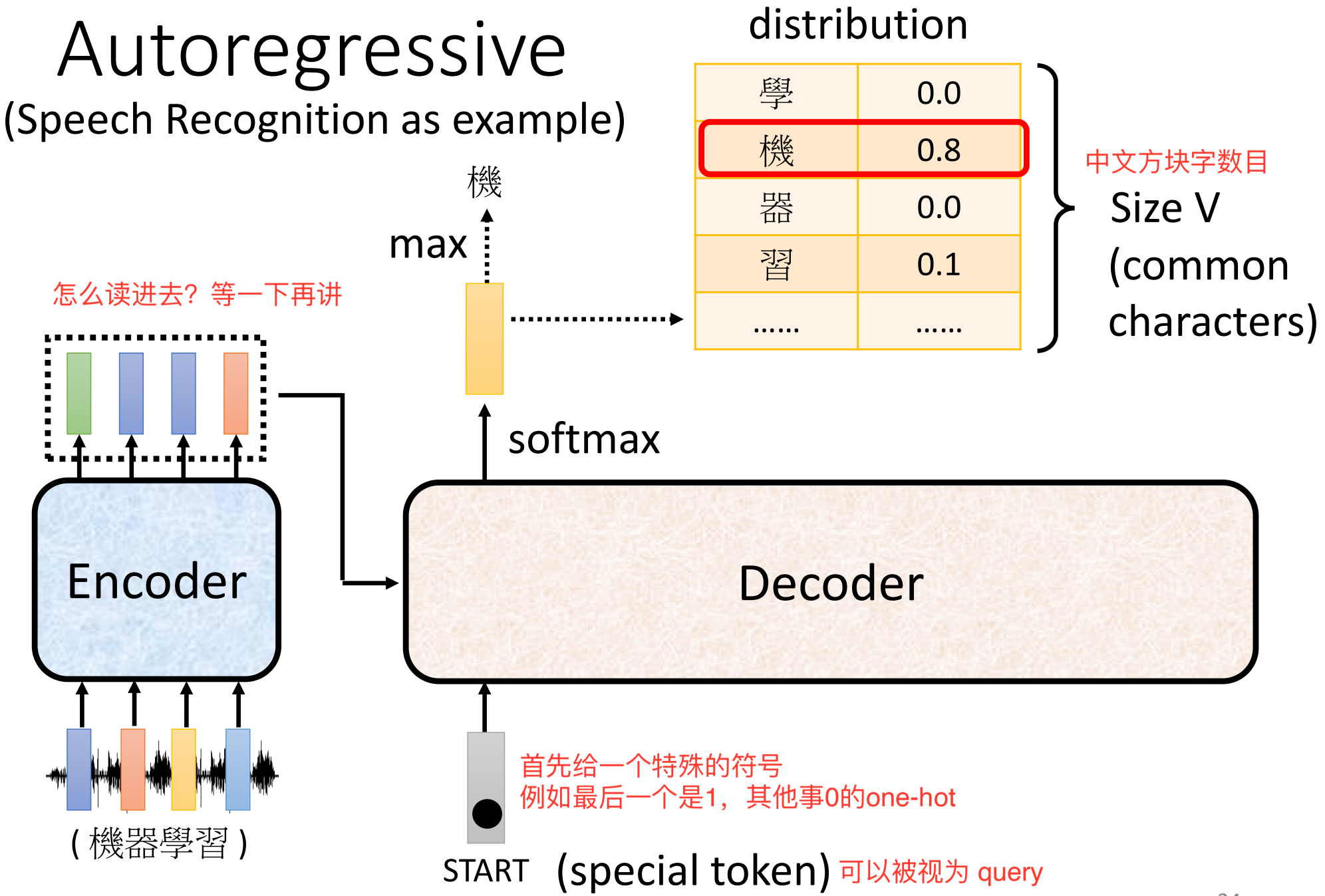
Decoder 有两种

Decoder – Autoregressive (AT)

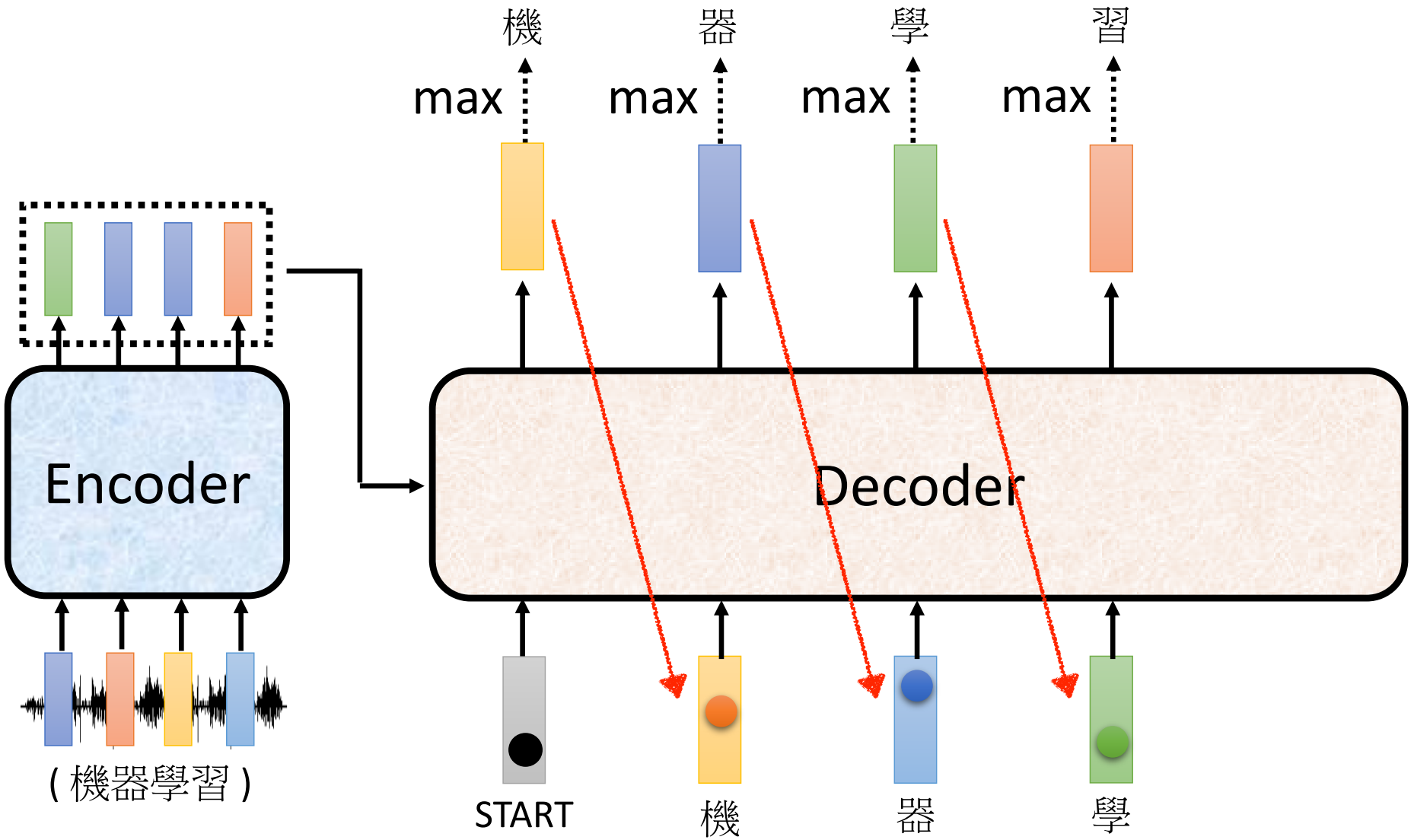


Autoregressive

(Speech Recognition as example)

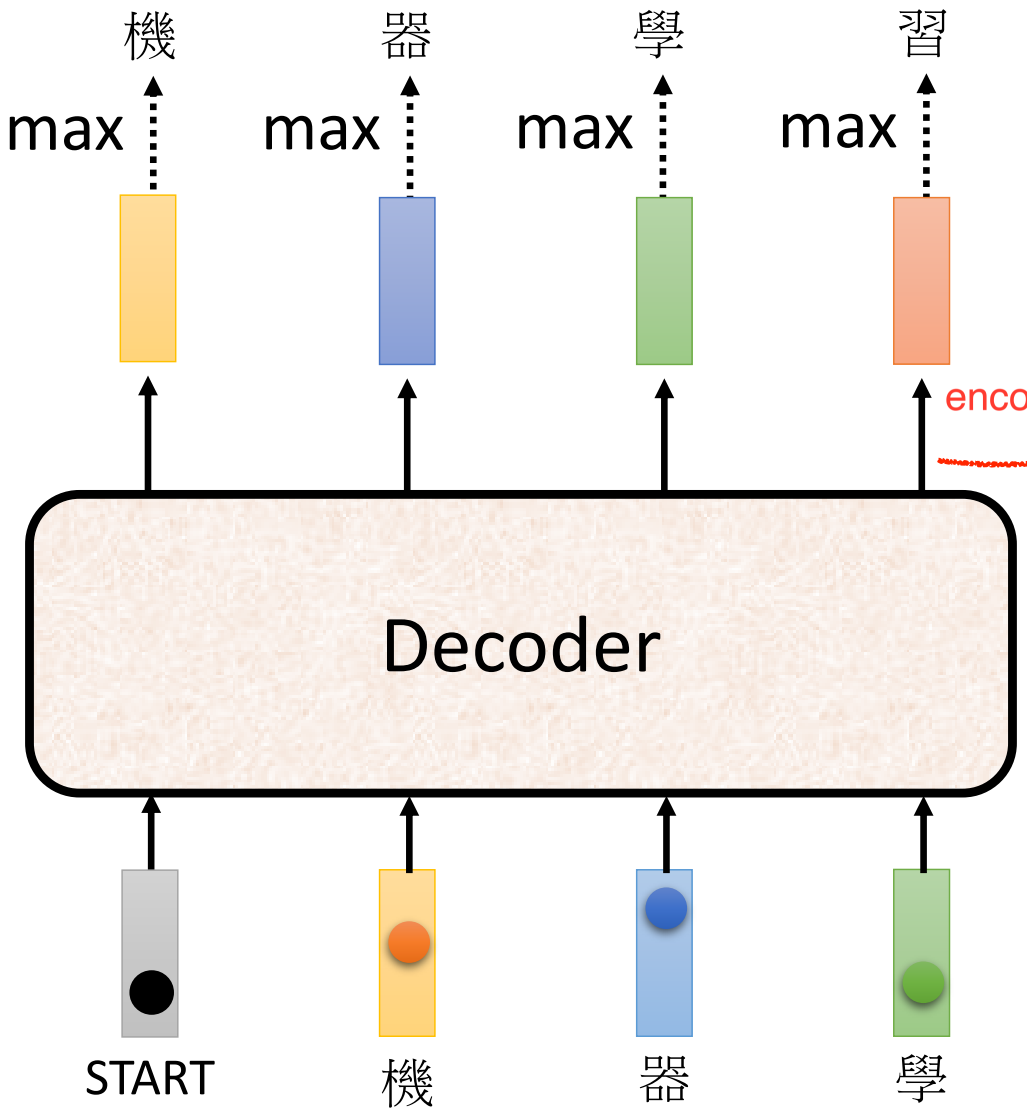


Autoregressive

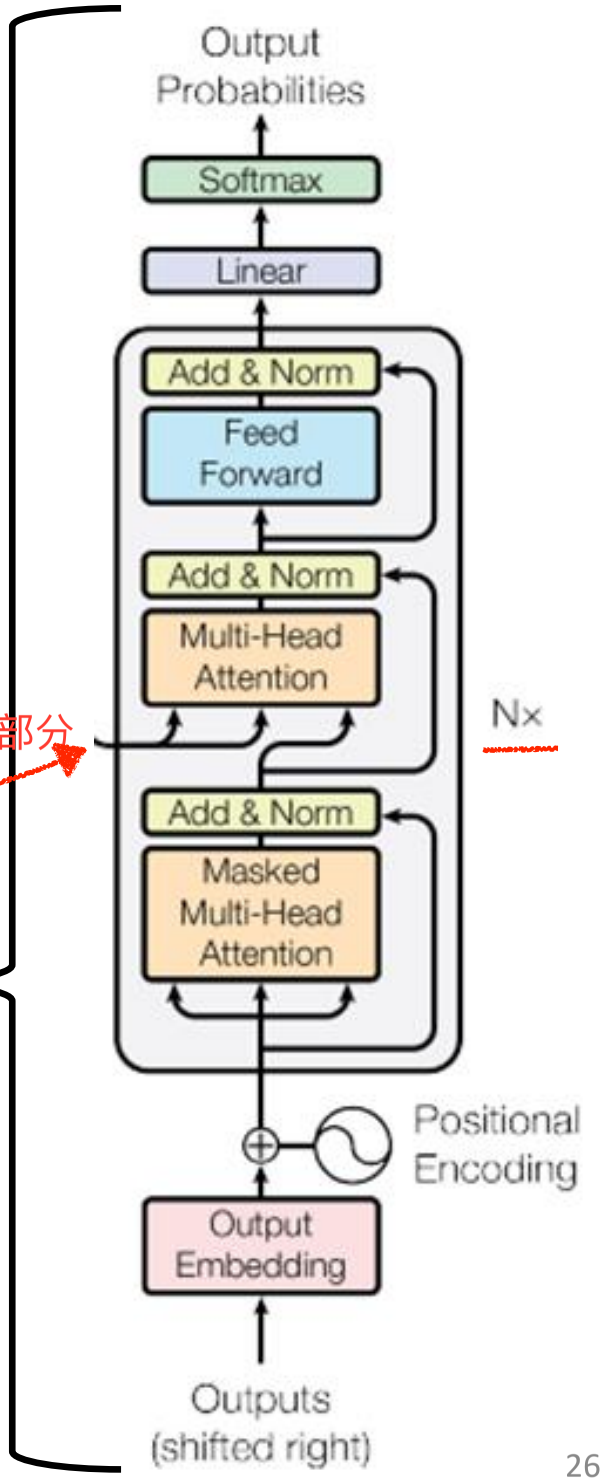


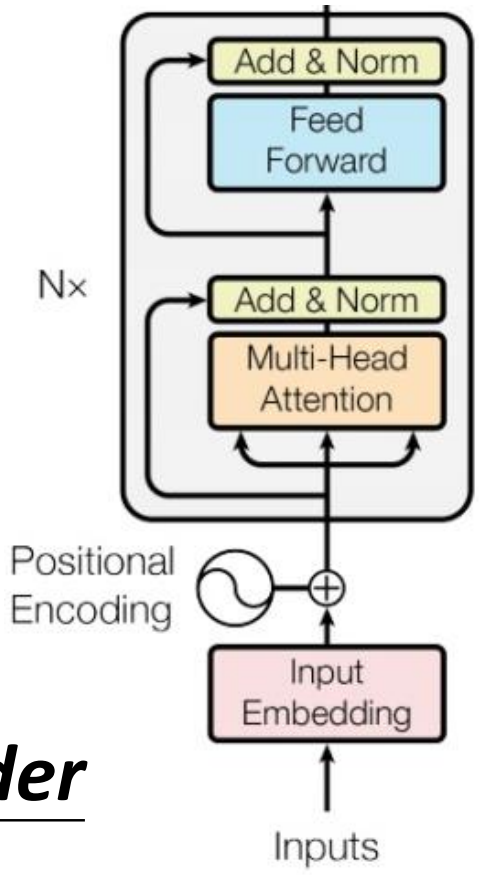
Decoder 会把自己的输出当作接下来的输入，所以有可能看到错误的东西，会不会一步错步步错呢?

ignore the input from the encoder here ☺

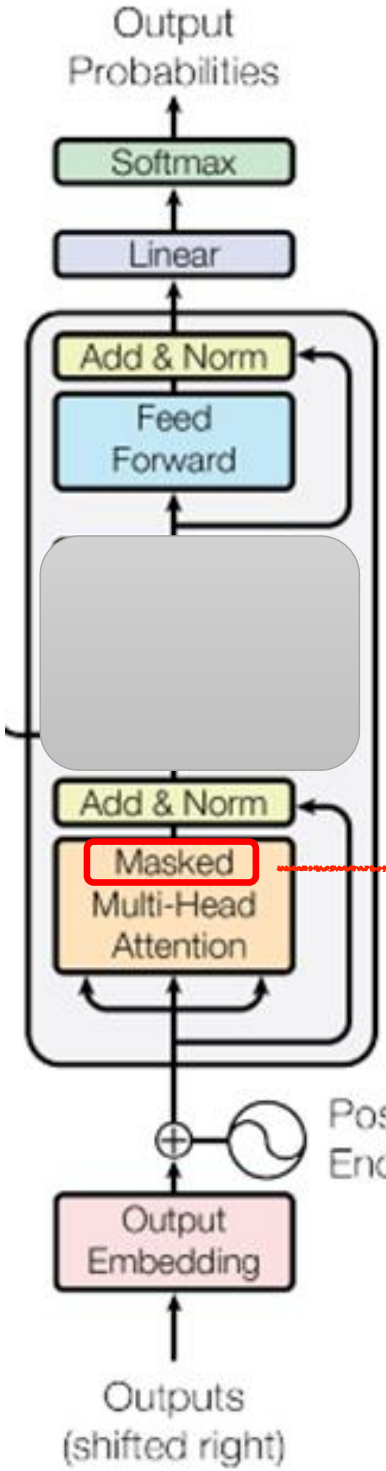


encoder的部分





Encoder



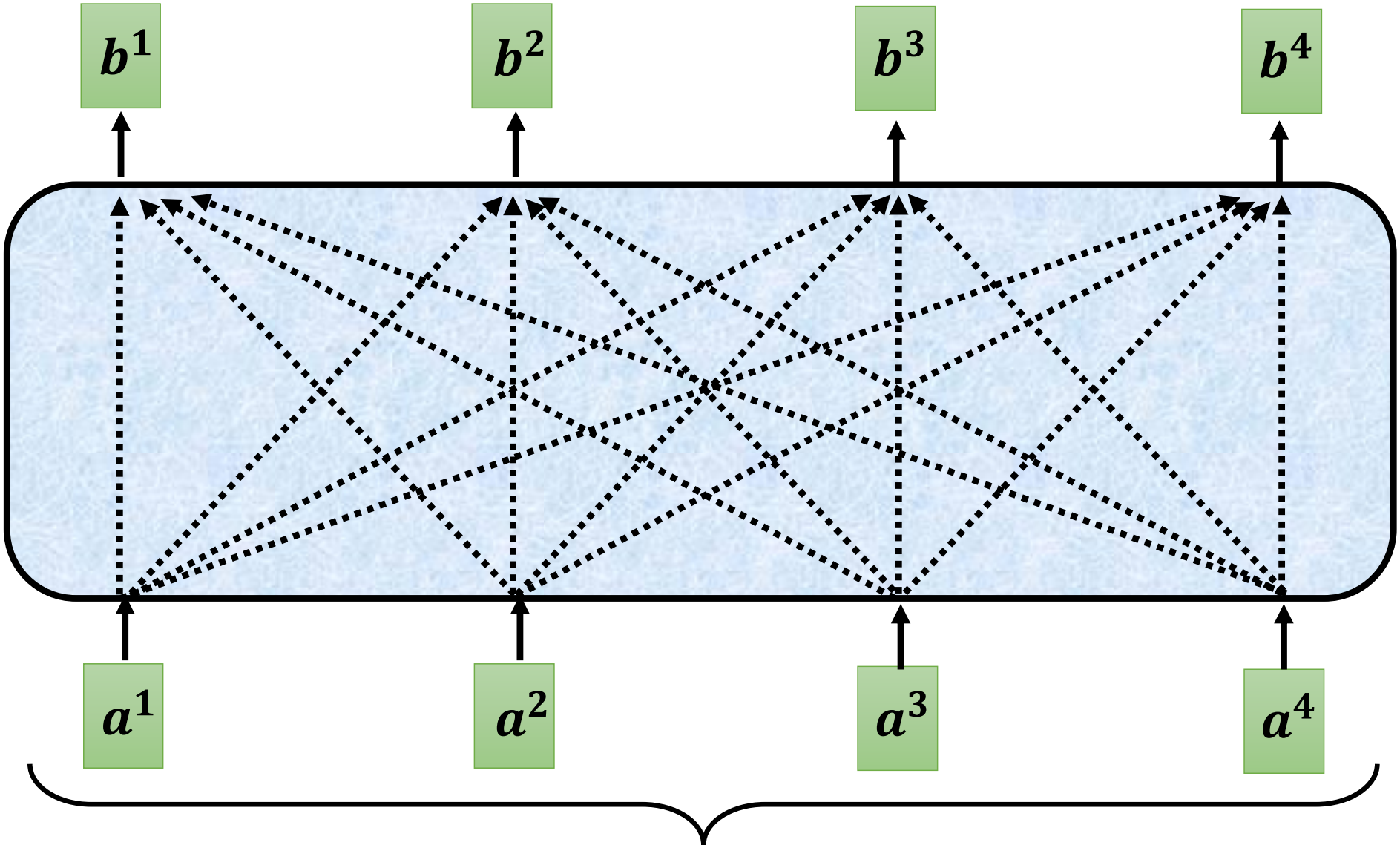
盖掉中间decoder和encoder基本一致

注意有个Masked

Decoder

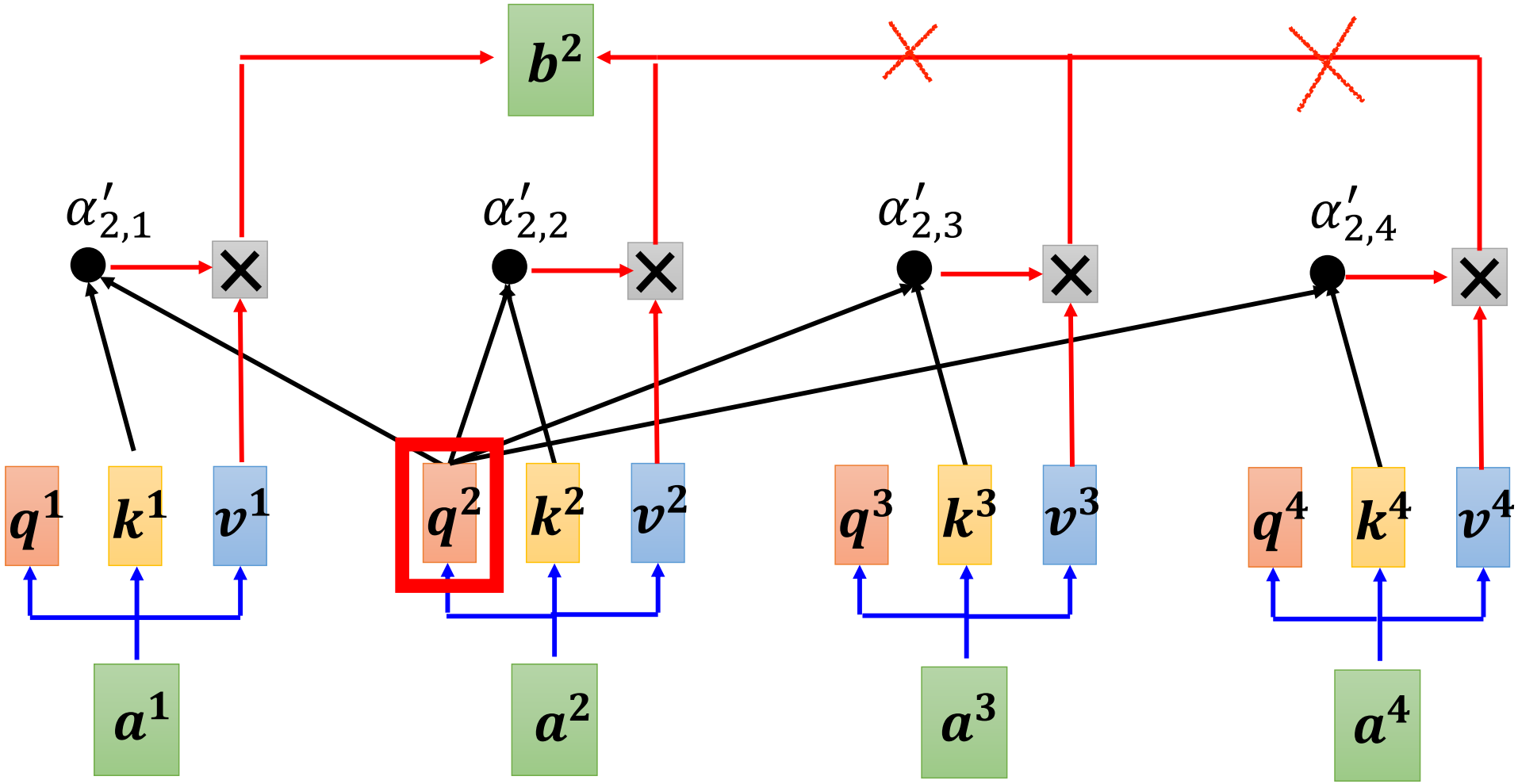
Self-attention → *Masked Self-attention*

现在不能再看右边的部分，产生b2的时候只能考虑a1和a2



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

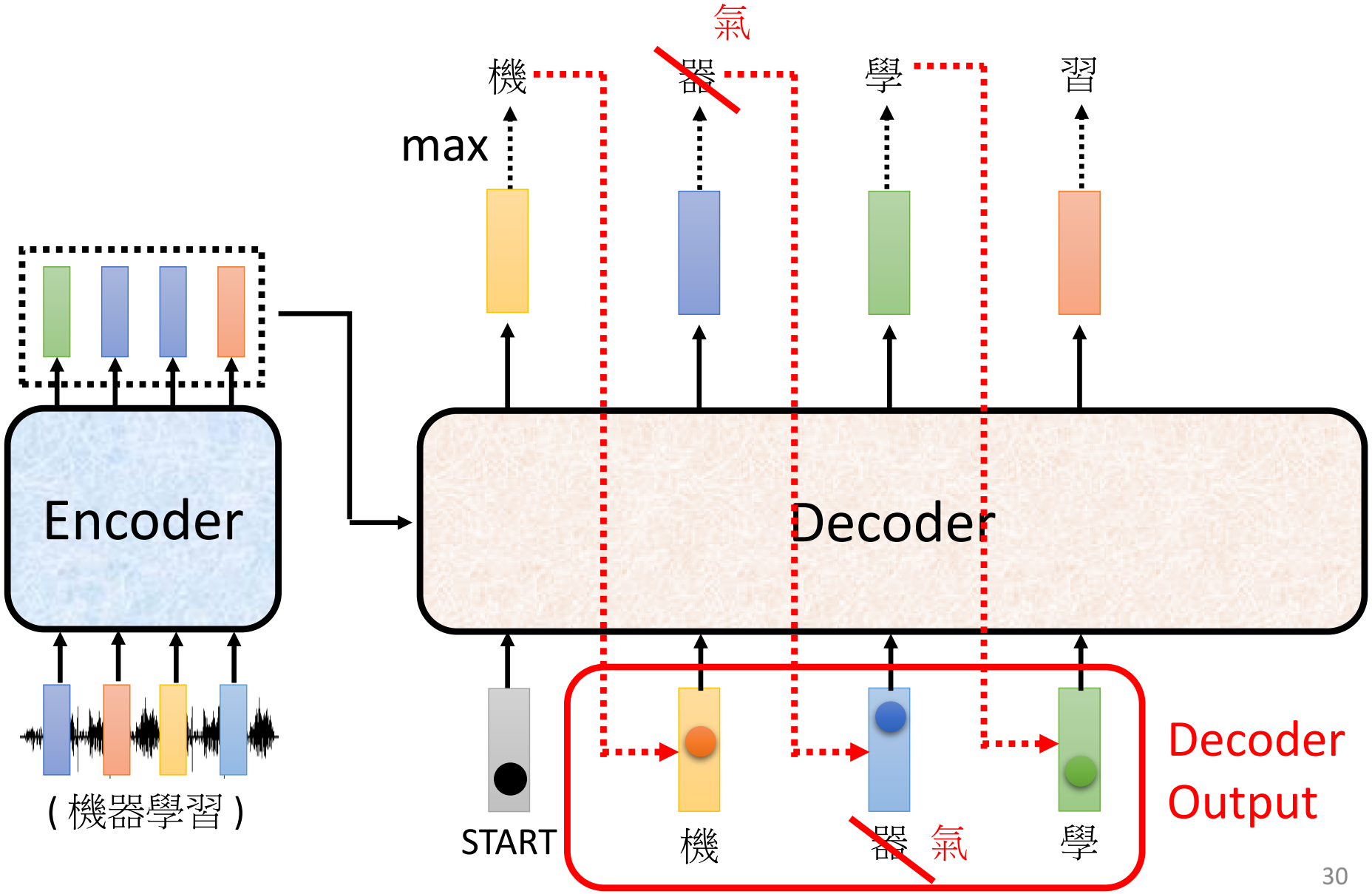
Self-attention → *Masked Self-attention*



Why masked? Consider how does decoder work

其实非常的直觉，输出是一个一个产生的，注意P14，物体识别就没有Mask了，因为输出不是一个一个产生的

Autoregressive

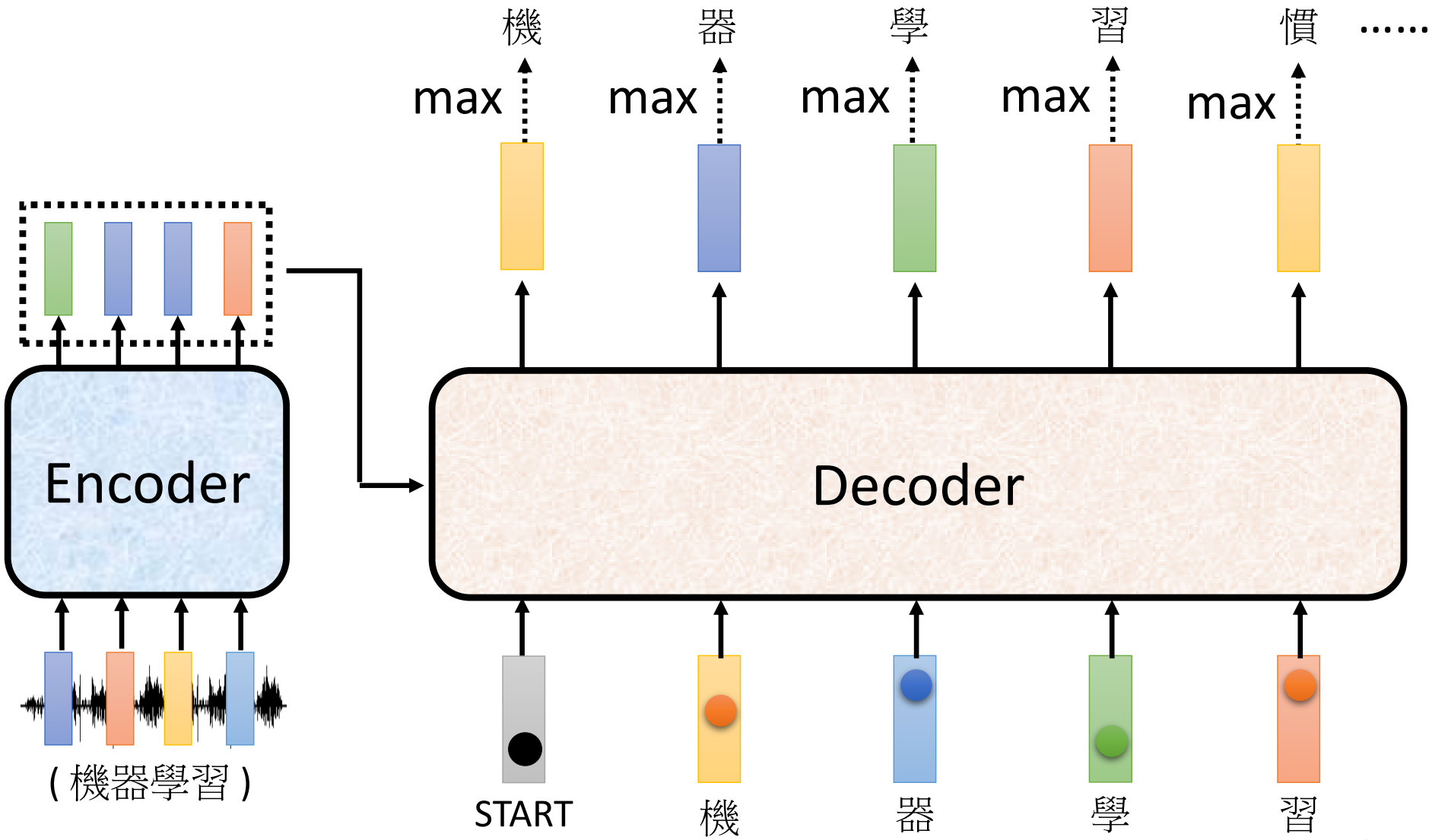


Autoregressive

We do not know the correct output length.

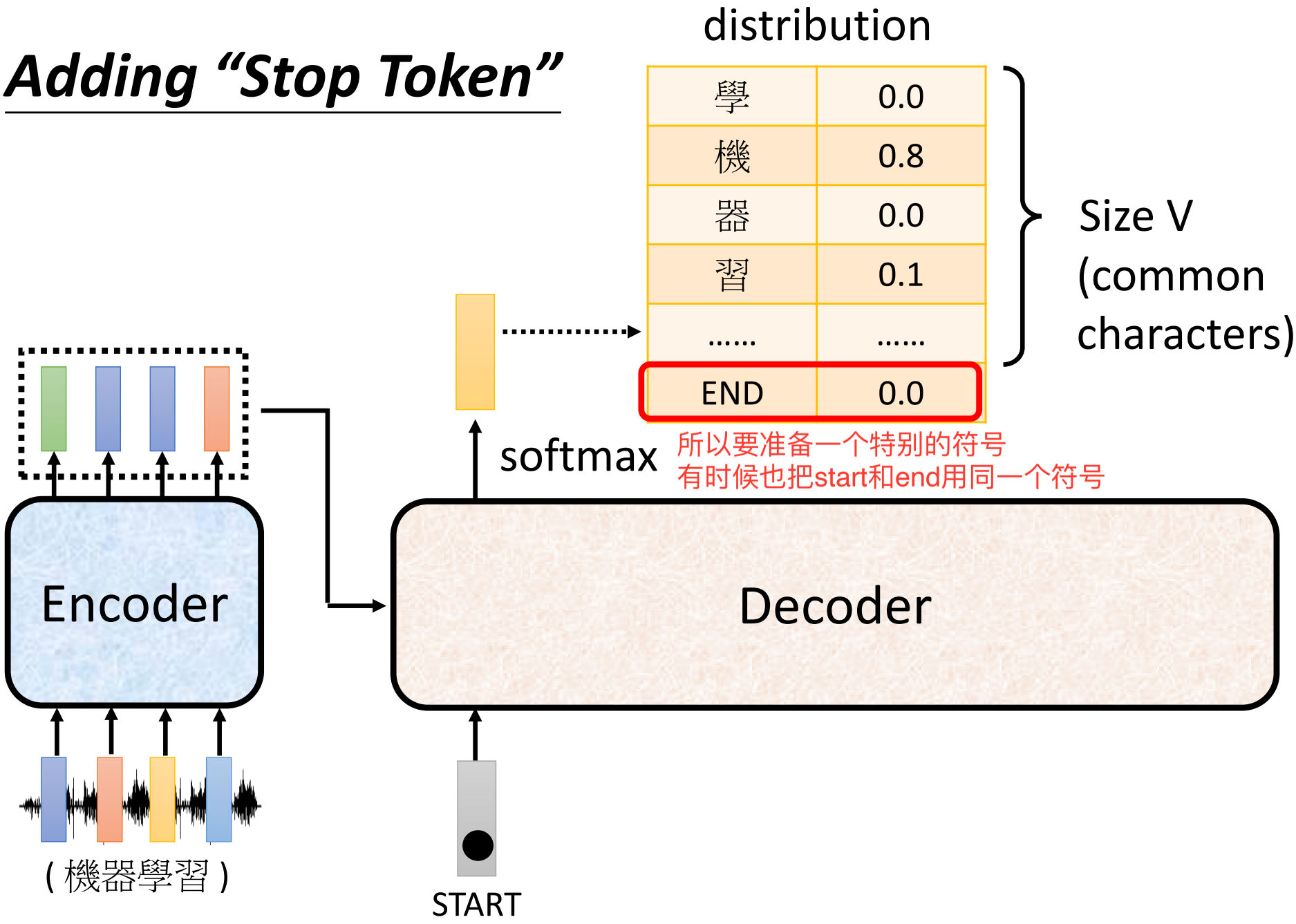
怎么停止呢?

Never stop!



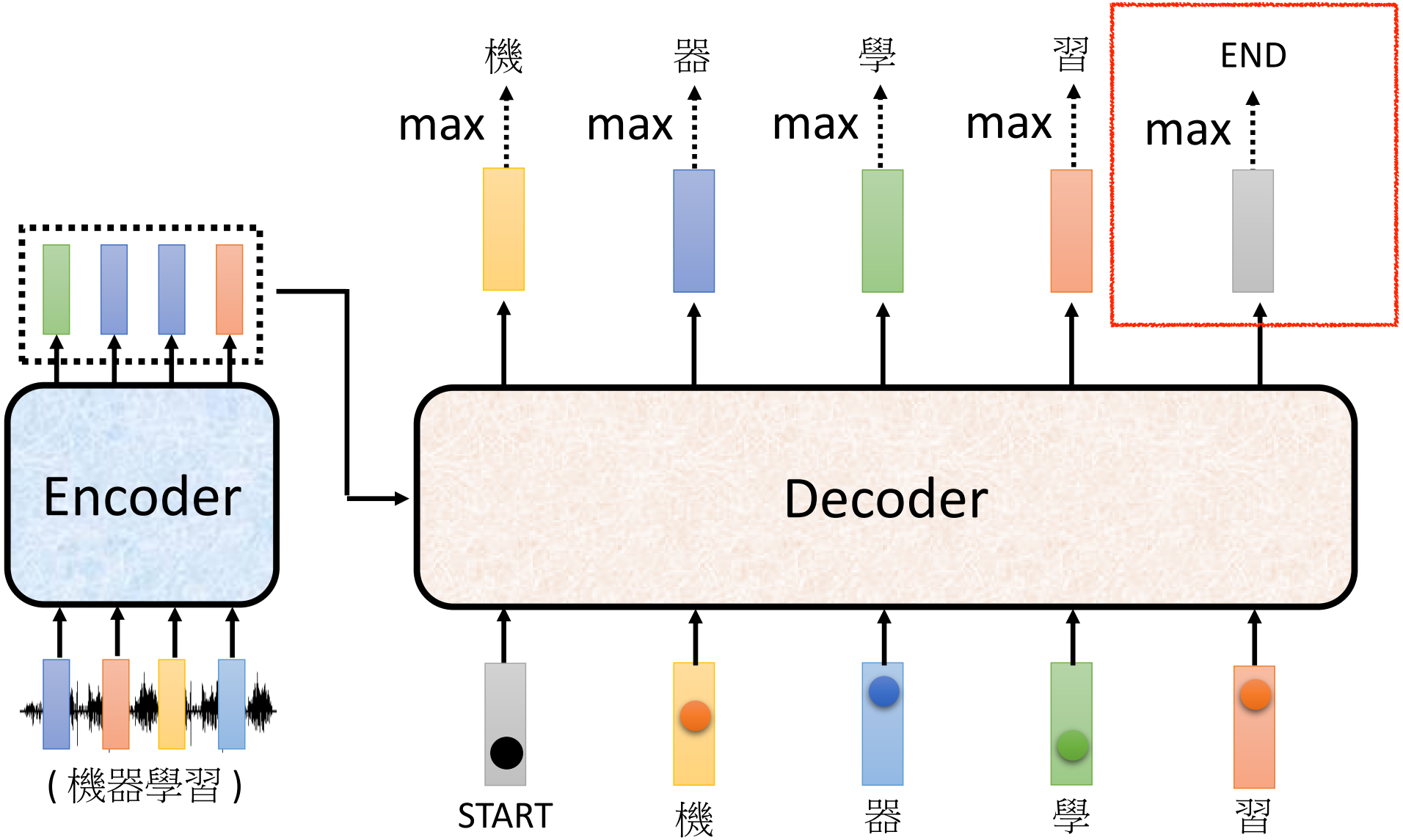
推文接龍 (Tweet Solitaire)

Adding "Stop Token"

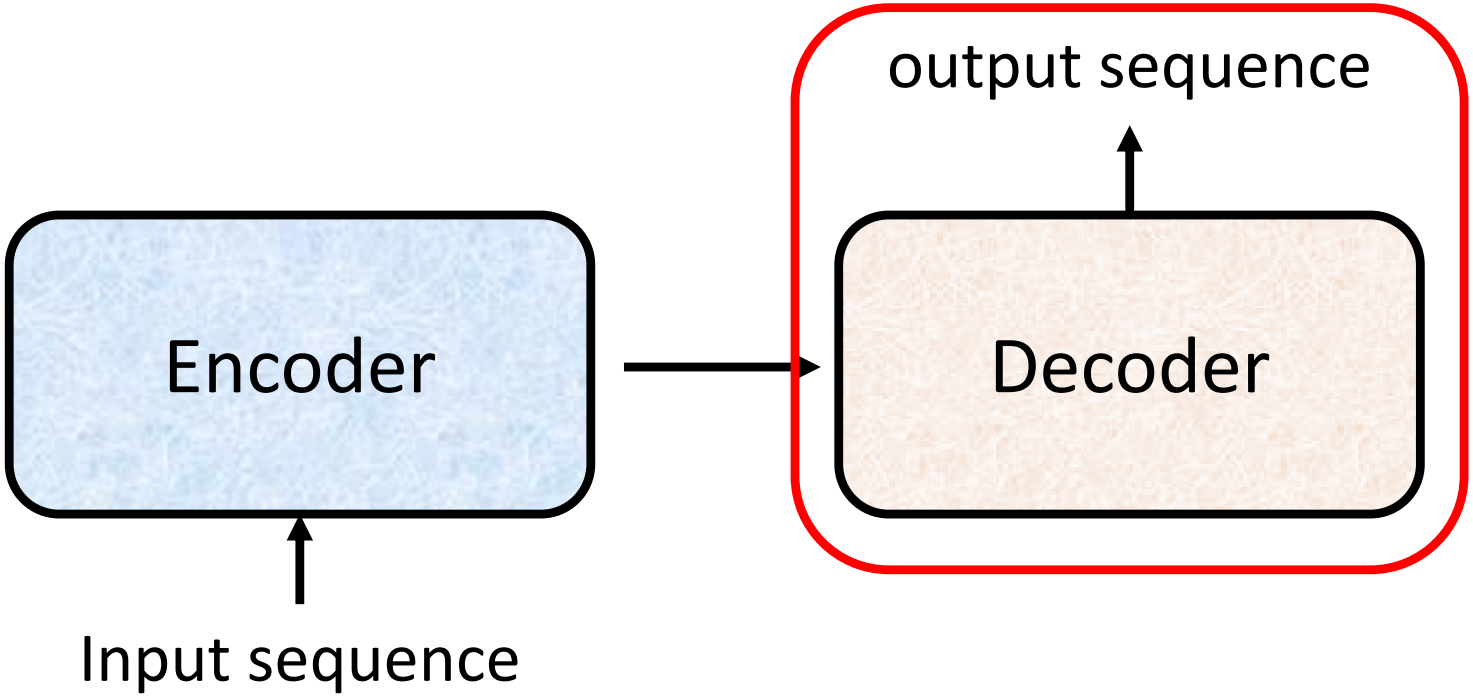


Autoregressive

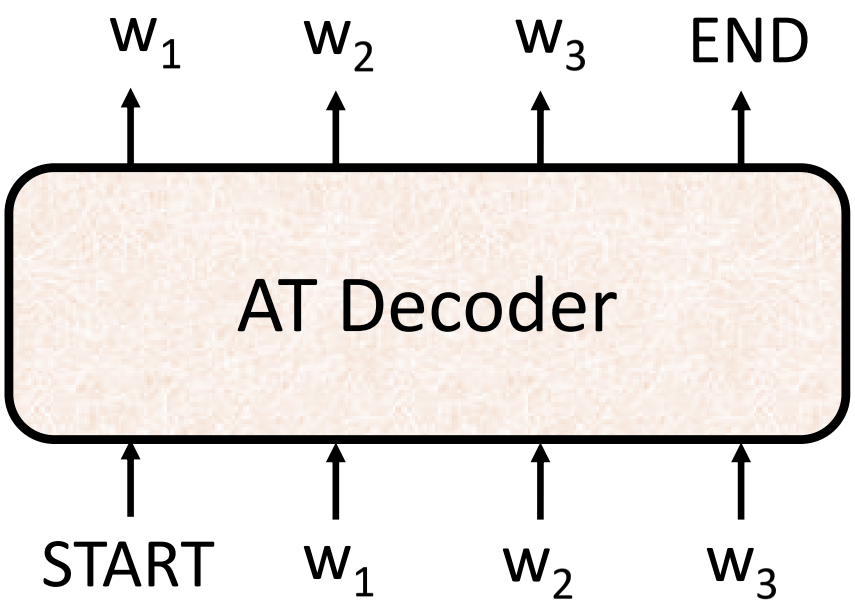
Stop at here!



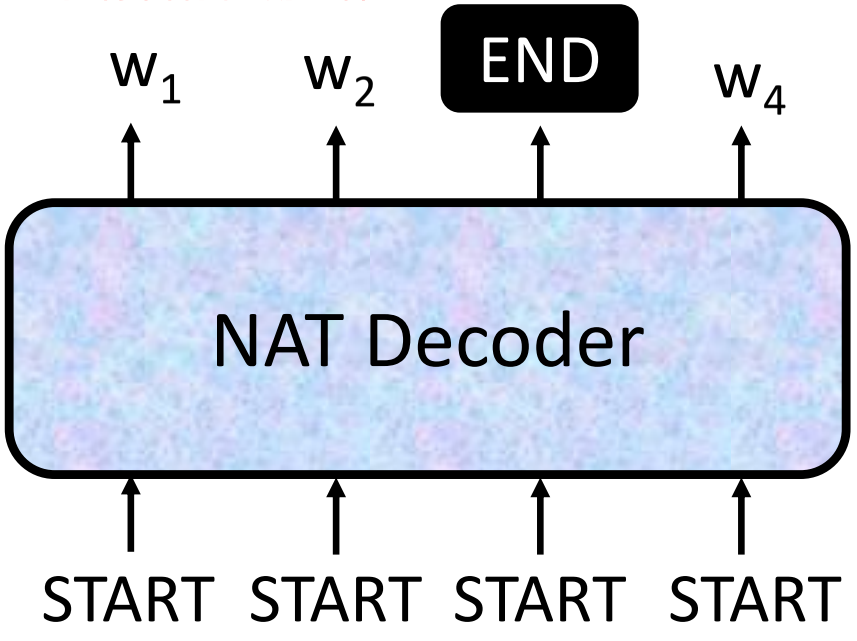
Decoder
– Non-autoregressive (NAT)



AT v.s. NAT



一次把整个句子都产生出来
就像目标检测一样



- How to decide the output length for NAT decoder?
  - Another predictor for output length 吃encoder的output进行预测
  - Output a very long sequence, ignore tokens after END 给一个上限

并行会很快！ 比较能够控制输出的长度，例如语音合成，将长度缩放一半就讲得快一点

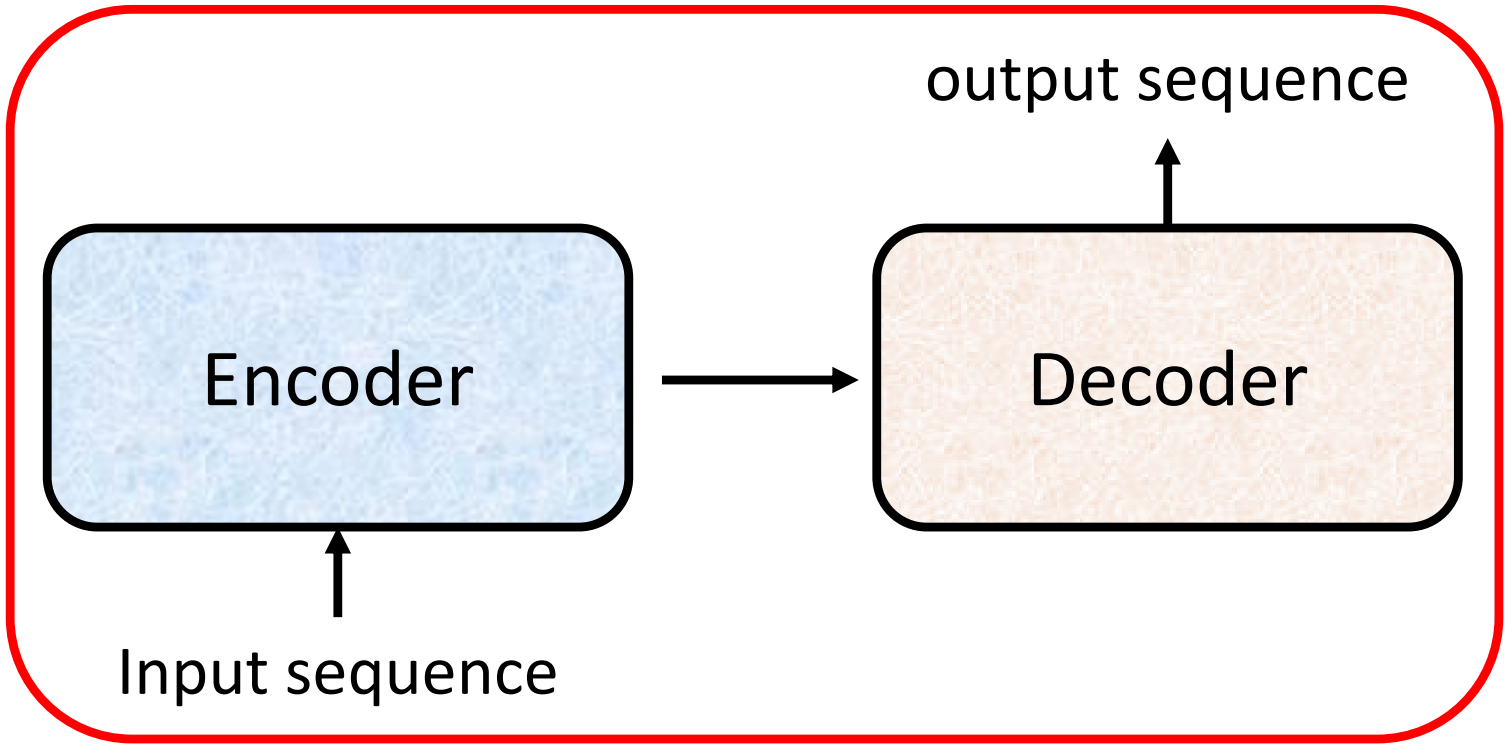
- Advantage: parallel, more stable generation (e.g., TTS)
- NAT is usually worse than AT (why? **Multi-modality**)

To learn more ……

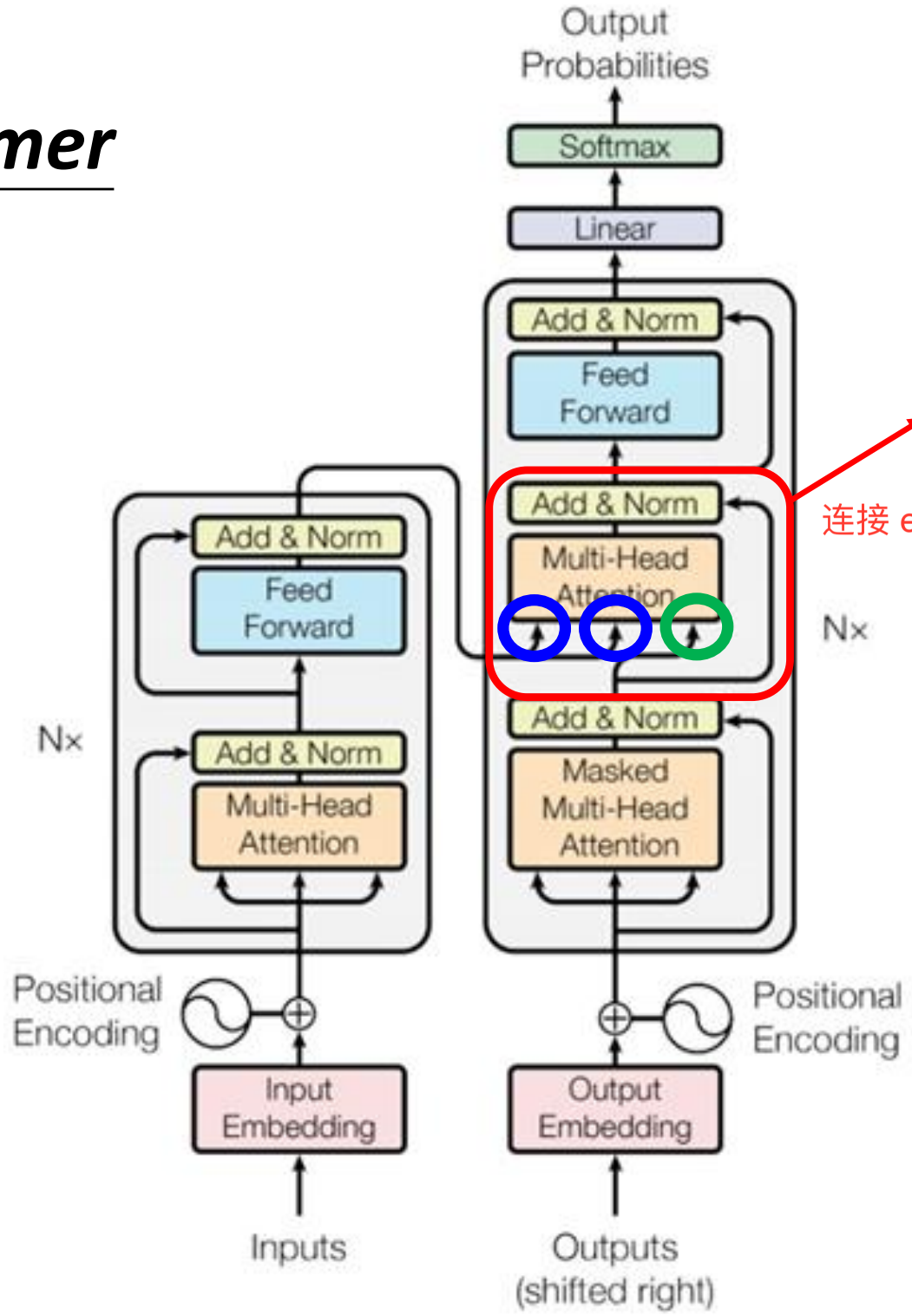
https://youtu.be/jvyKmU4OM3c

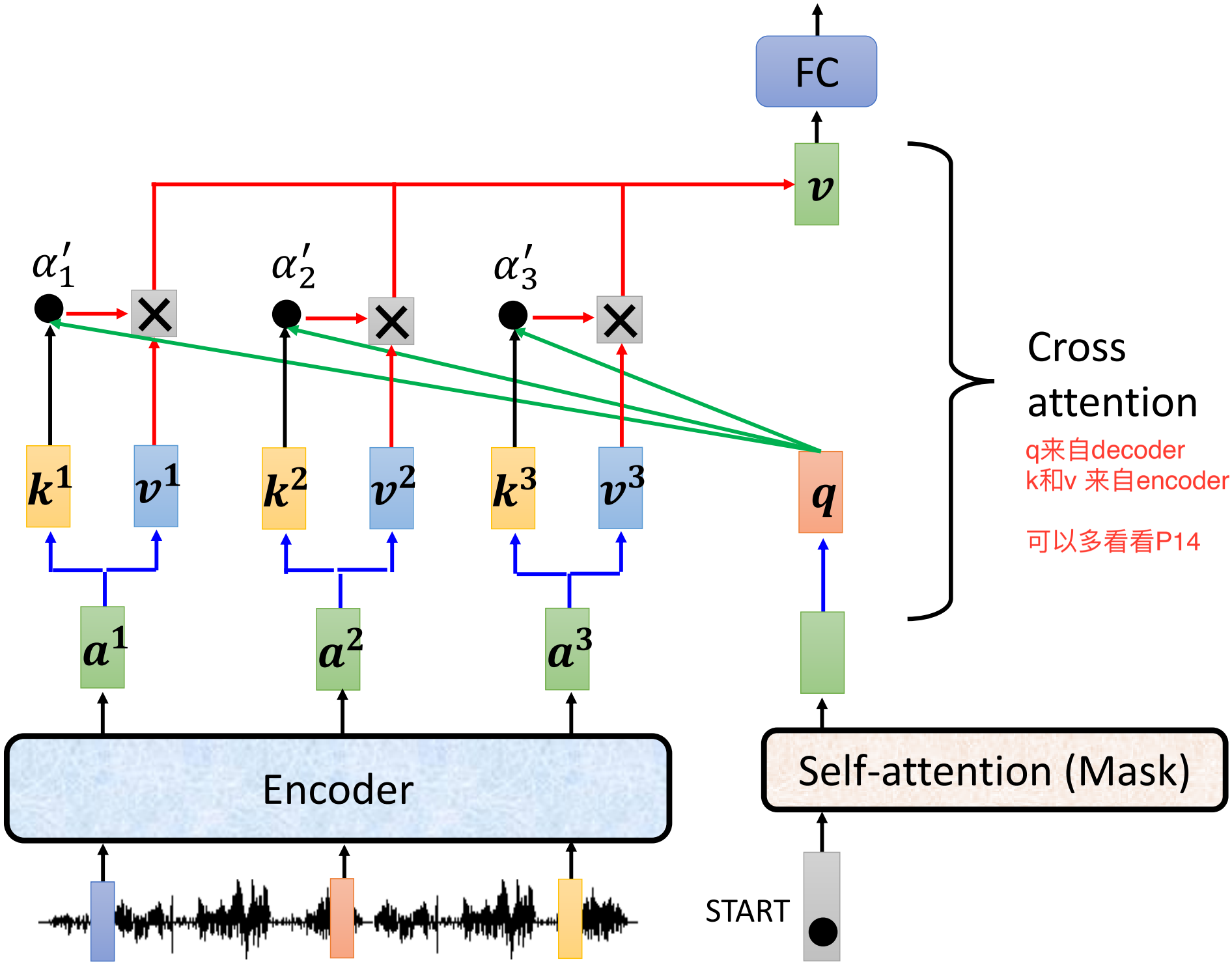
(in Mandarin)

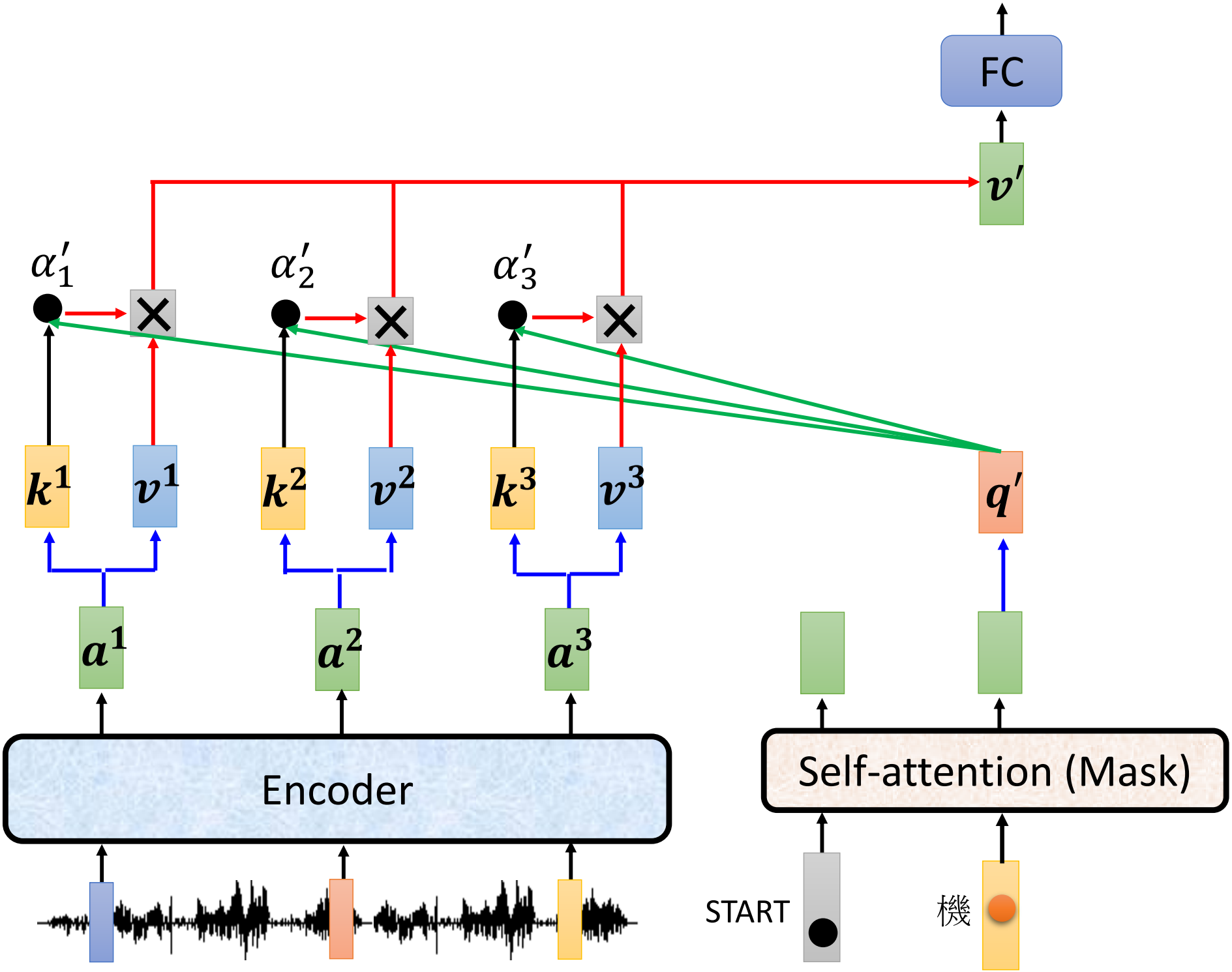
Encoder-Decoder



Transformer




FC
v
α'_1
α'_2
α'_3
k^1
v^1
k^2
v^2
k^3
v^3
q
Cross attention
q来自decoder
k和v 来自encoder
可以多看看P14
a^1
a^2
a^3
Encoder
Self-attention (Mask)
START

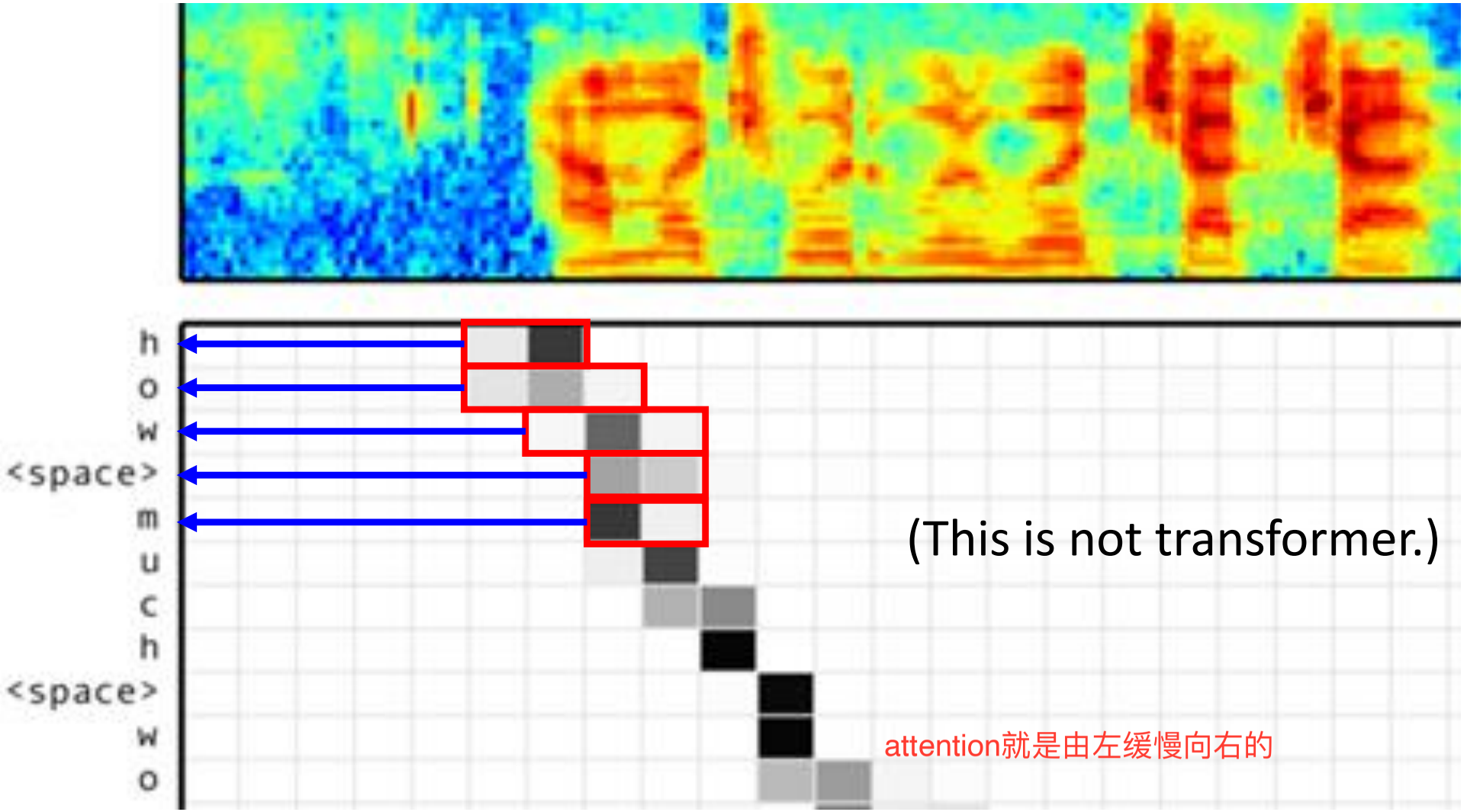

FC
v'
α'_1
α'_2
α'_3
k^1
v^1
k^2
v^2
k^3
v^3
q'
a^1
a^2
a^3
Encoder
Self-attention (Mask)
START
機

Cross Attention

2016年 Listen, attend and spell: A neural network for large vocabulary conversational speech recognition

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

每一列都是一个向量，横轴为时间



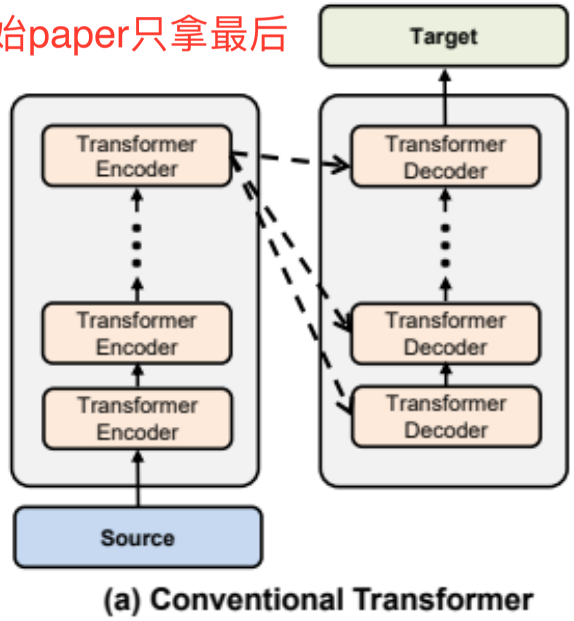
(This is not transformer.)

attention就是由左缓慢向右的

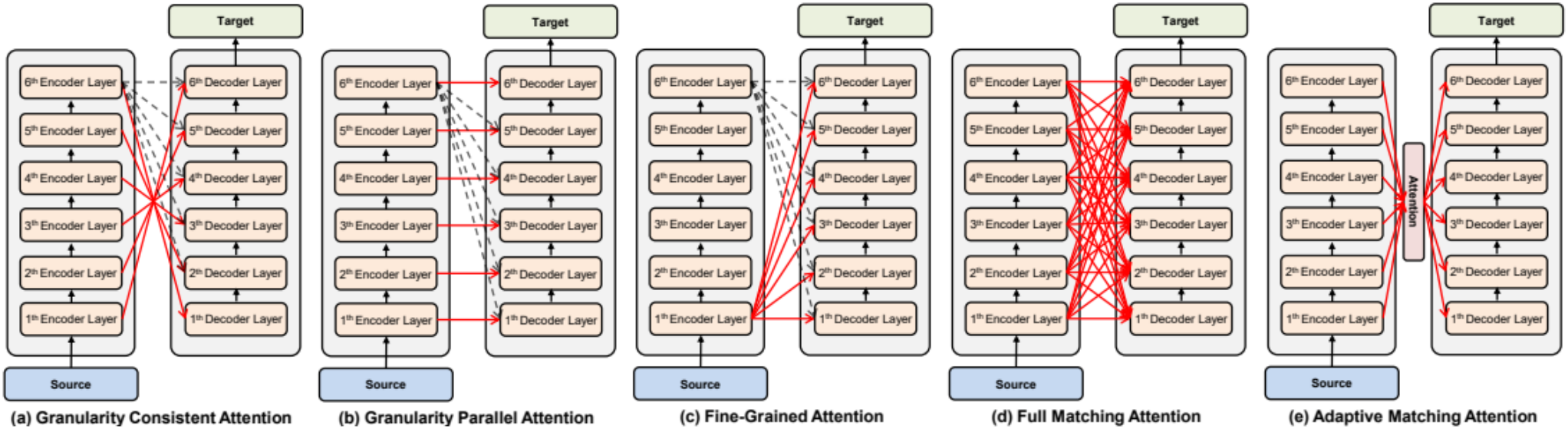
Cross Attention

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

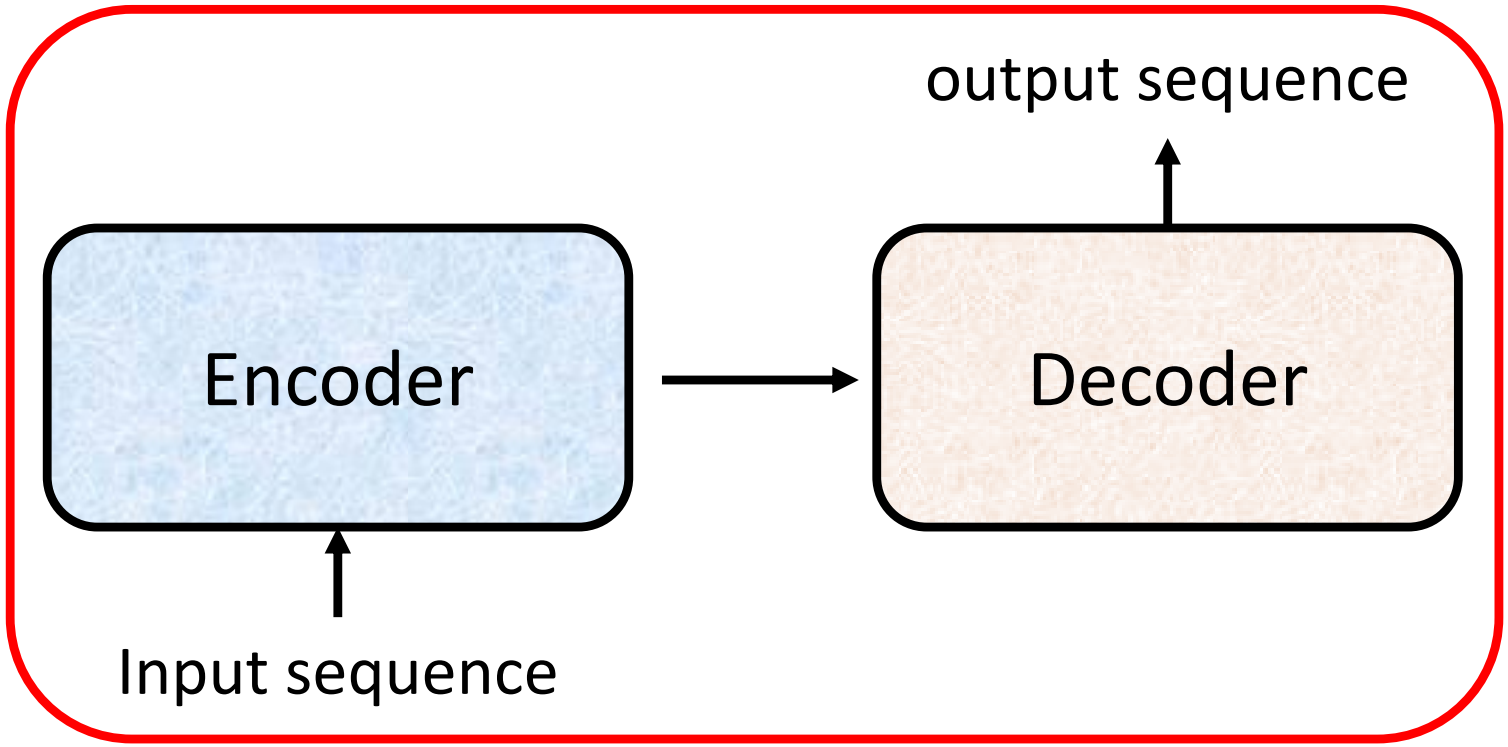
原始paper只拿最后

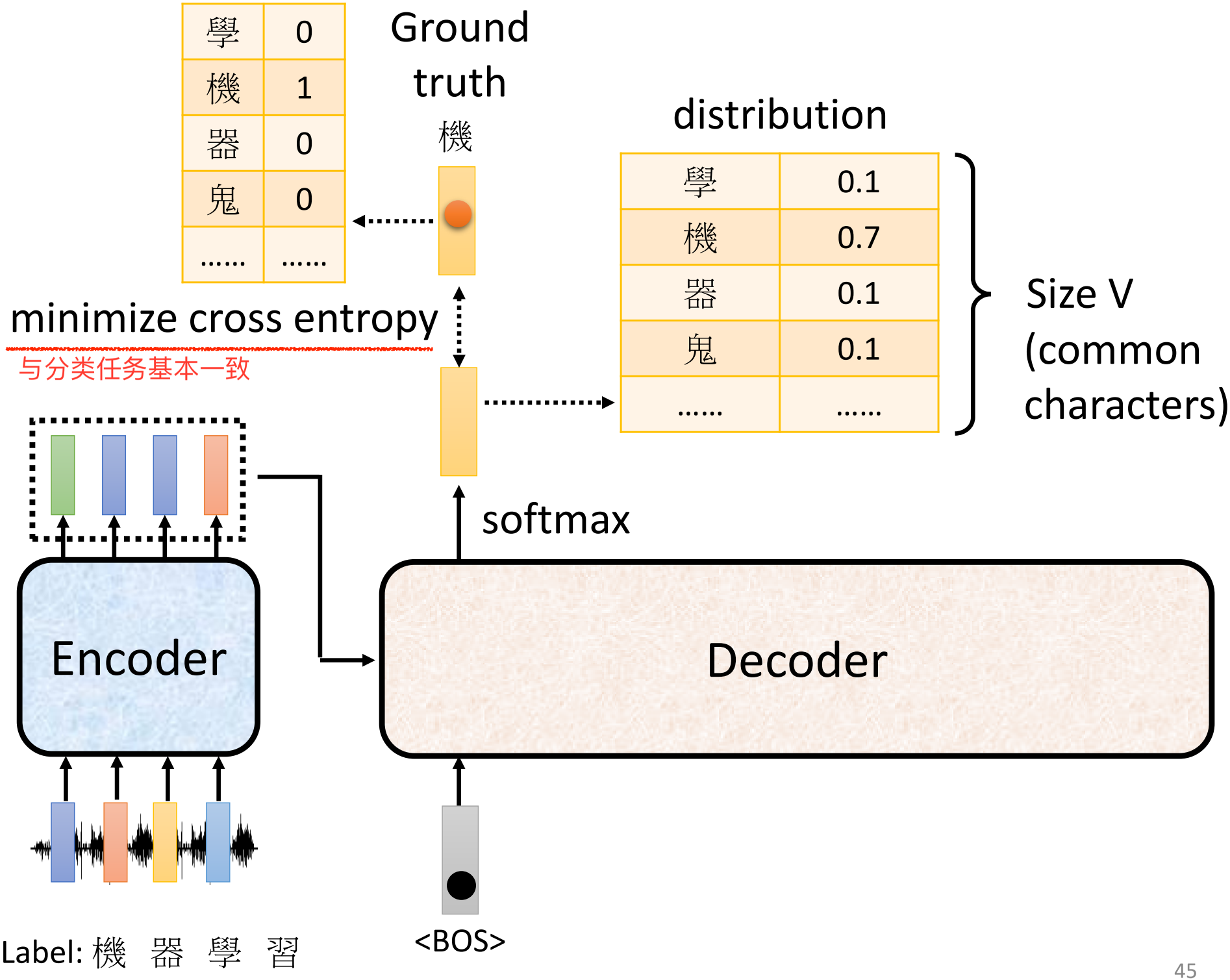


万事均可研究！！



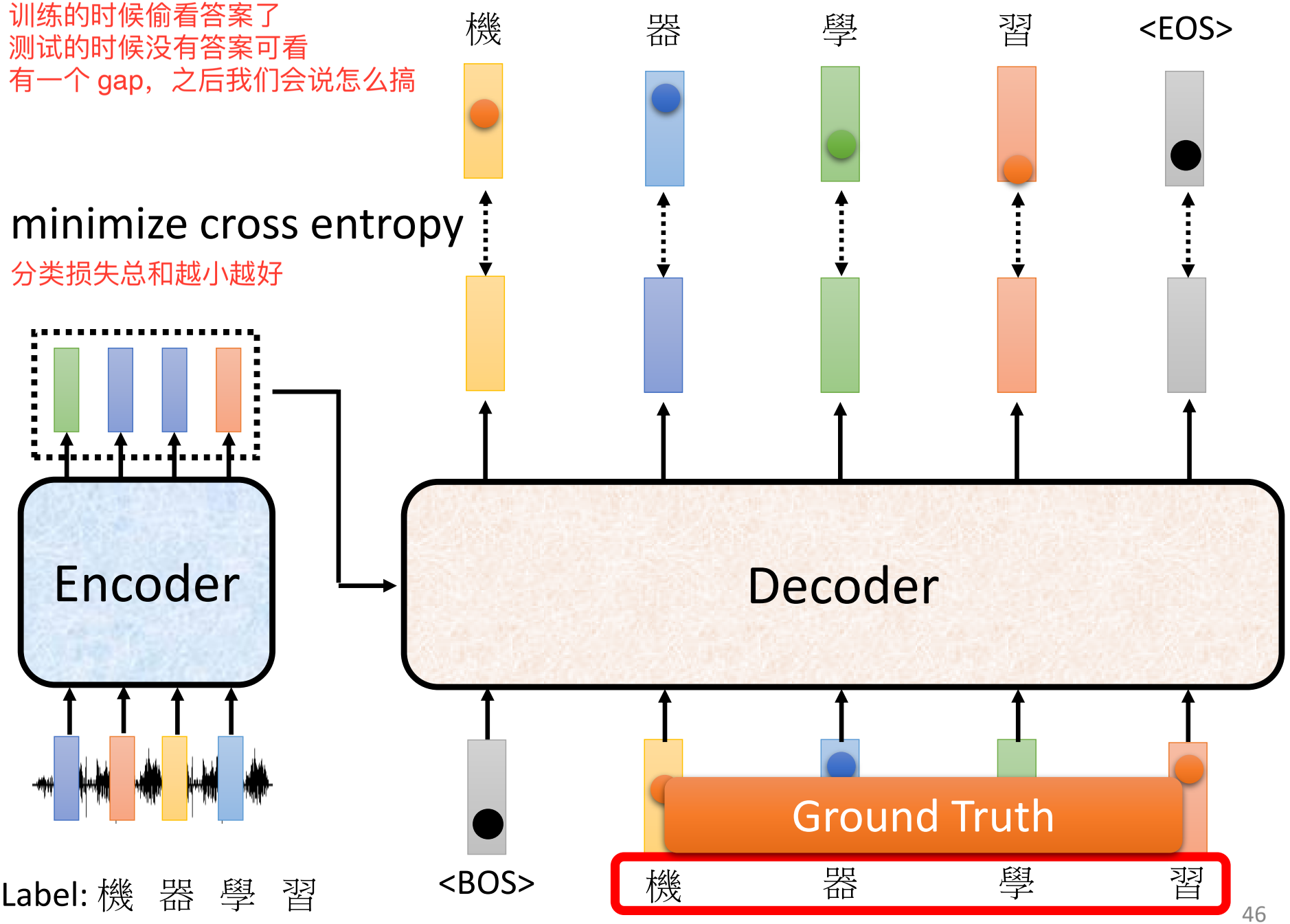
Training



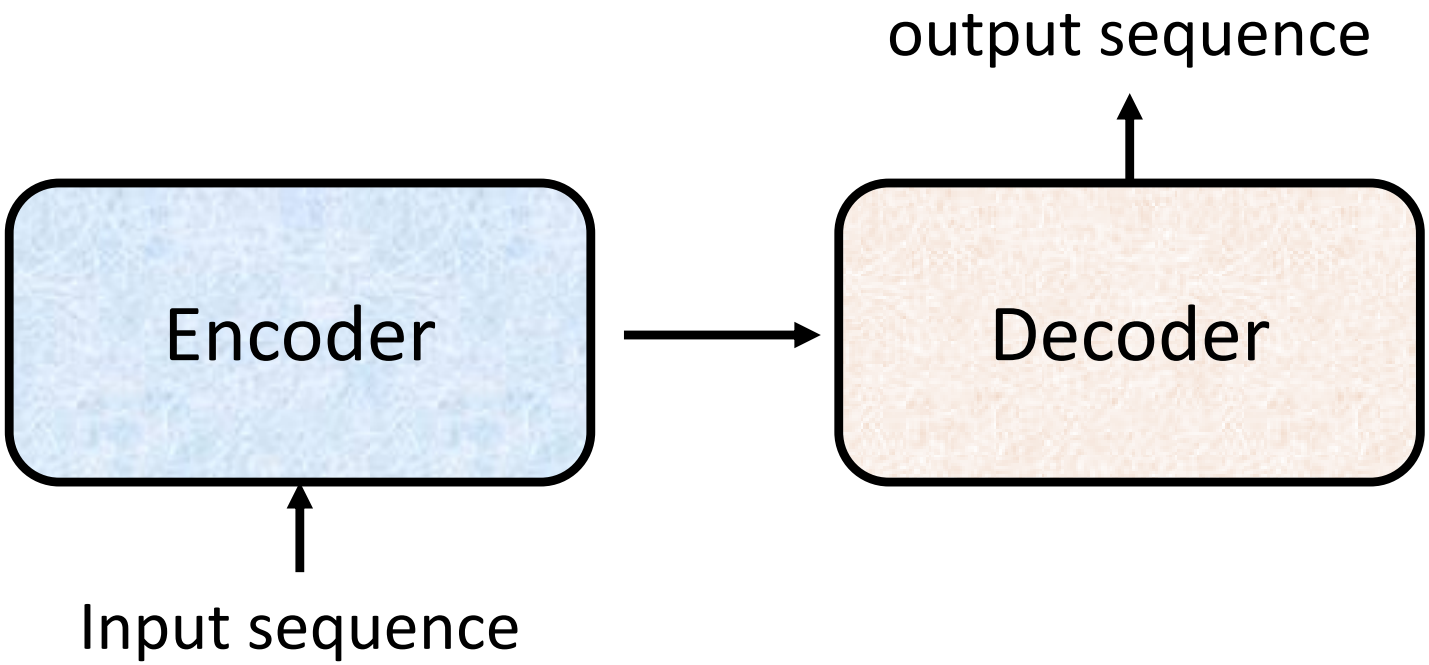

學 0
機 1
器 0
鬼 0
....
Ground truth
機
distribution
學 0.1
機 0.7
器 0.1
鬼 0.1
....
Size V (common characters)
minimize cross entropy
与分类任务基本一致
softmax
Encoder
Decoder
<BOS>
Label: 機 器 學 習

Teacher Forcing: using the ground truth as input.

训练的时候偷看答案了
测试的时候没有答案可看
有一个 gap，之后我们会说怎么搞



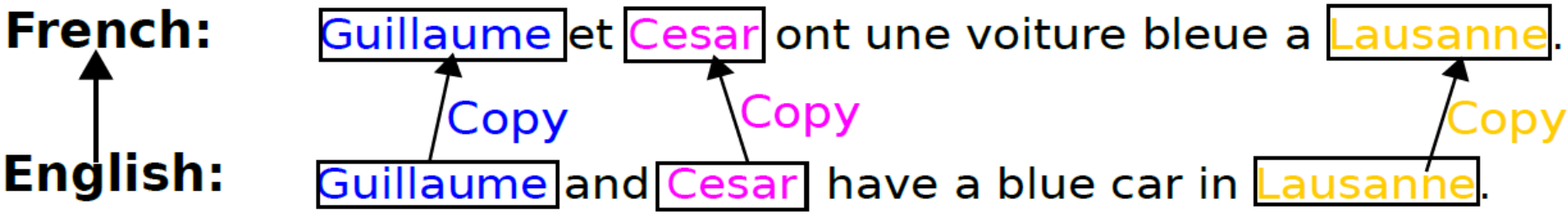
Tips



Copy Mechanism

copy 一部分就行了

Machine Translation



Chat-bot

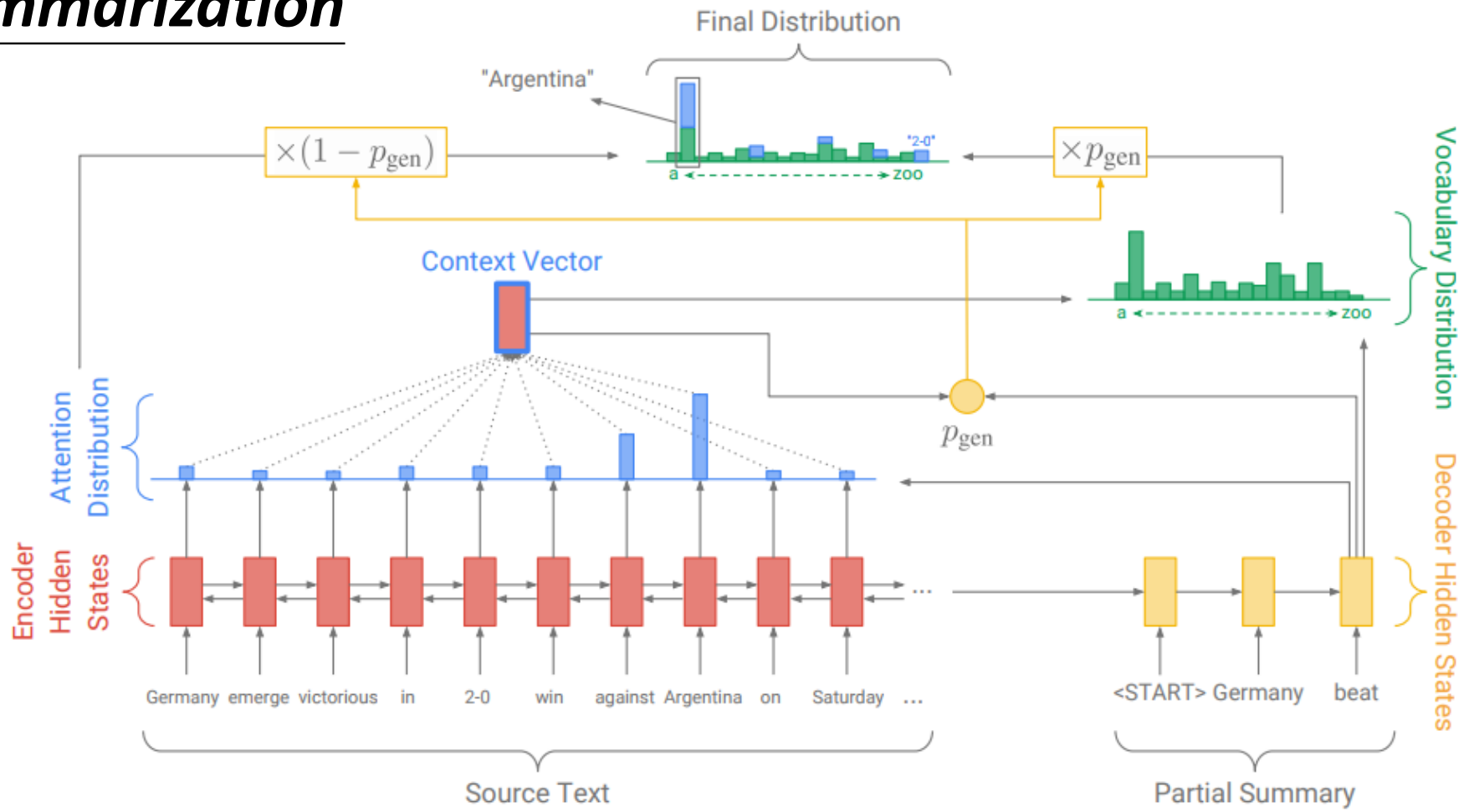
User: X寶你好，我是庫洛洛

Machine:庫洛洛你好，很高興認識你

Copy Mechanism

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

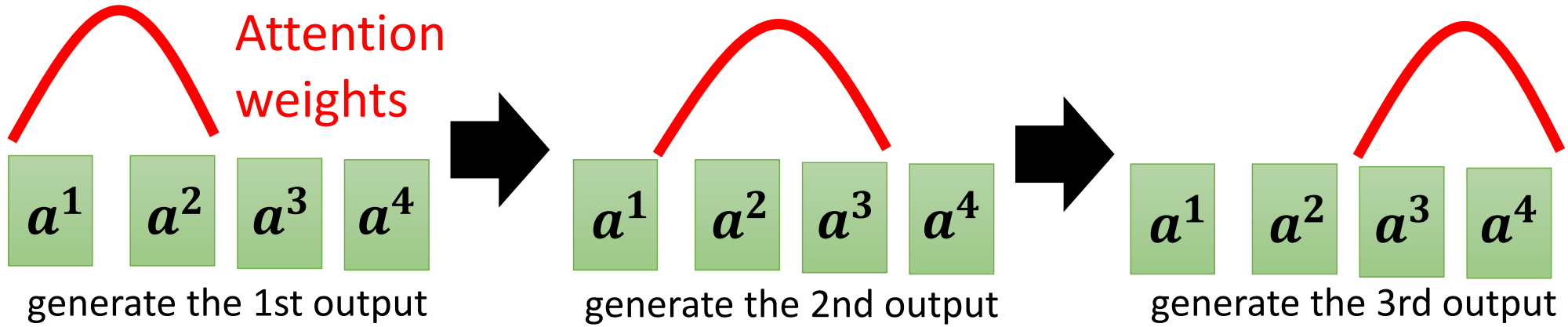
- 高雄發大財我現在要出征
- 發財發財發財發財
- 發財發財發財
- 發財發財
- 發財　　(Missing an input character!)

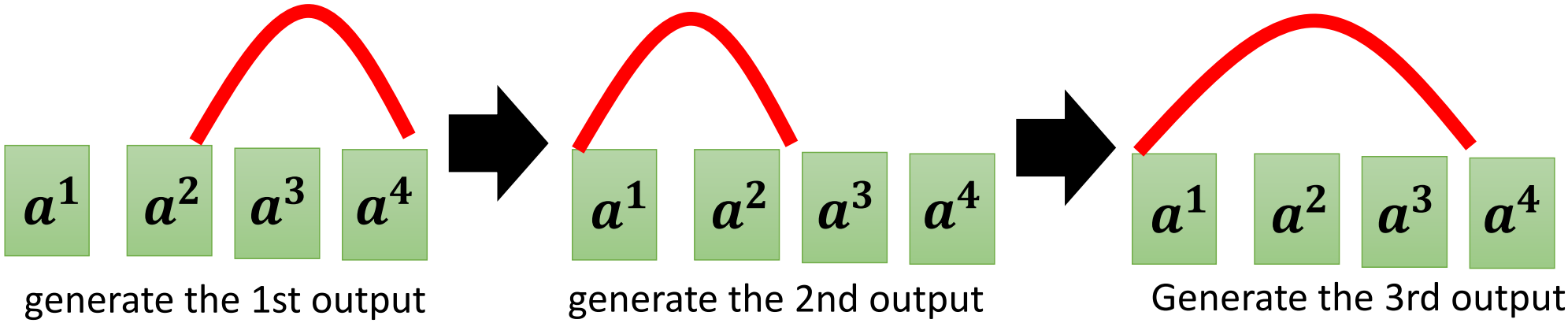
对于语音识别模型等
告诉模型attention只能从左到右，在训练中引入

Guided Attention

Monotonic Attention
Location-aware attention

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





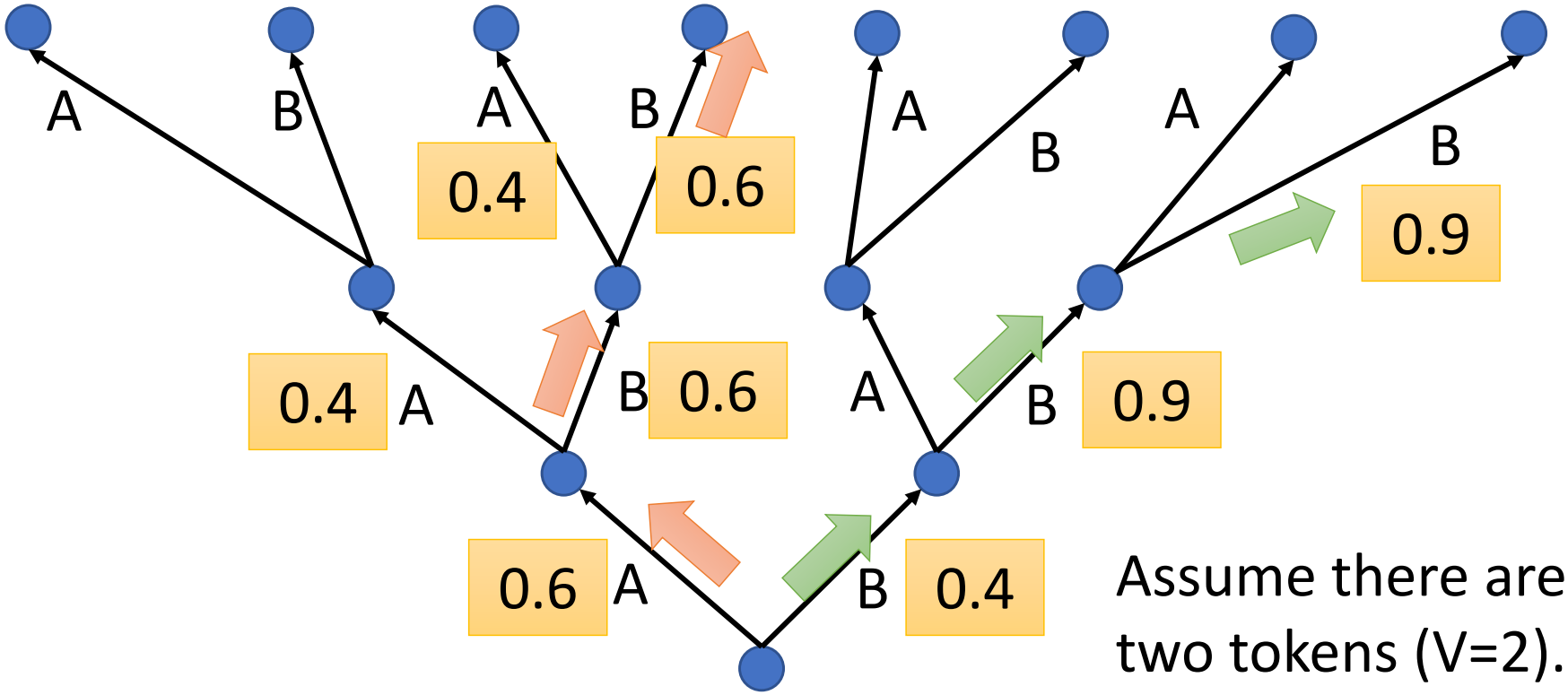
Something wrong!

Beam Search 有时候有用，有时候没有用

The red path is ***Greedy Decoding***. 局部最优而非全局最优

The green path is the best one.

Not possible to check all the paths … → Beam Search



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

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.

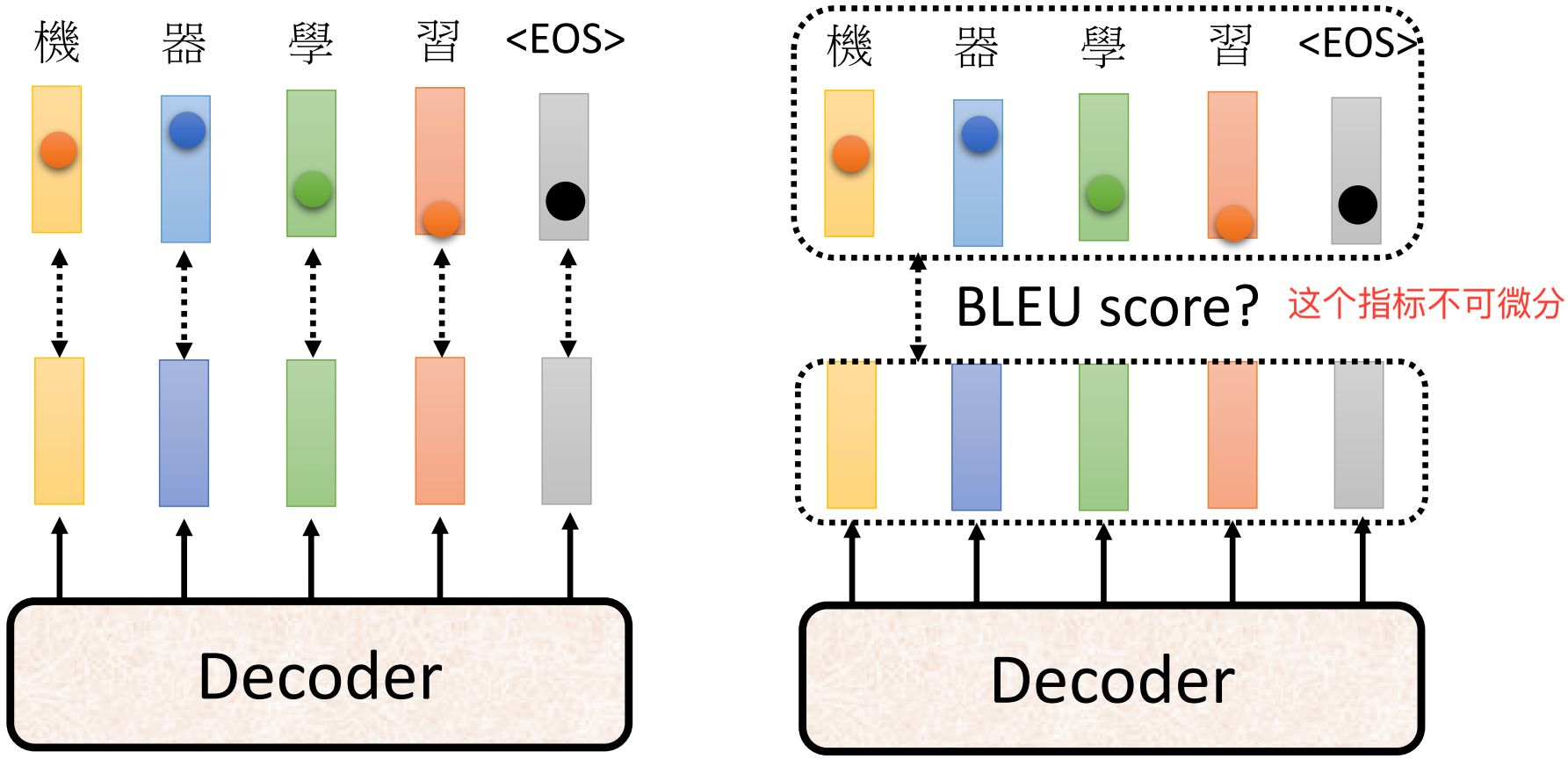
Beam Search, *b*=32:
"The study, published in the Proceedings of the National Academy of Sciences of the United States of America (PNAS), was conducted by researchers from the Universidad Nacional Autónoma de México (UNAM) and the Universidad Nacional Autónoma de México (UNAM/Universidad Nacional Autónoma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

Optimizing Evaluation Metrics?



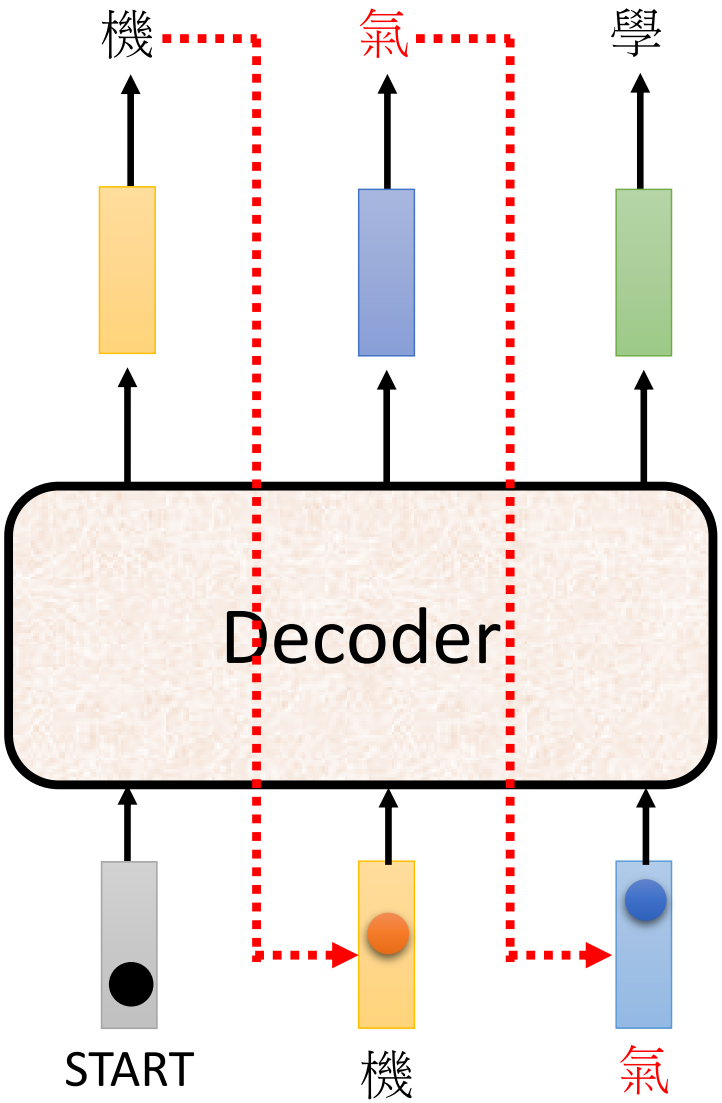
How to do the optimization?

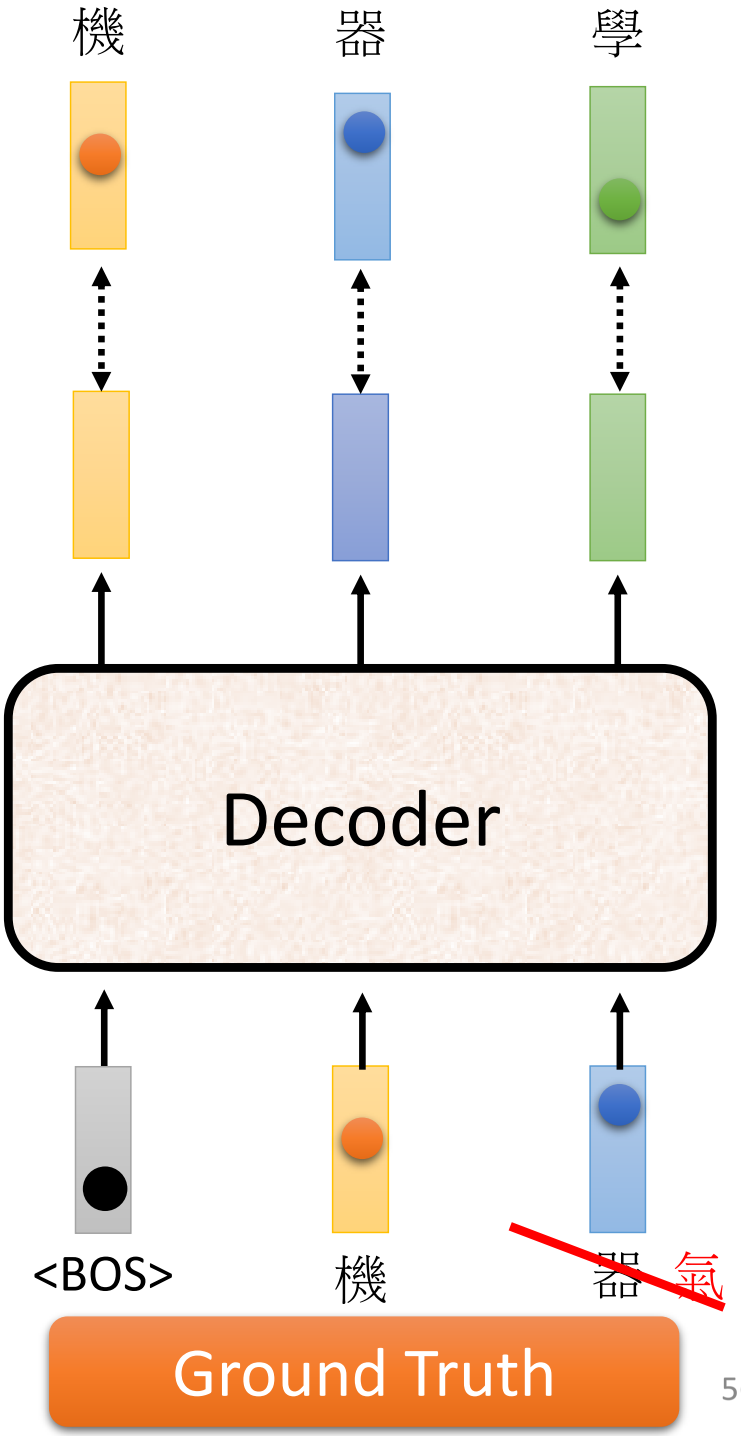
When you don't know how to optimize, just use reinforcement learning (RL)! https://arxiv.org/abs/1511.06732

一步错步步错，所以我们尝试在Decoder训练中加入一些错误的

There is a mismatch! ☹

exposure bias





Scheduled Sampling

- Original Scheduled Sampling

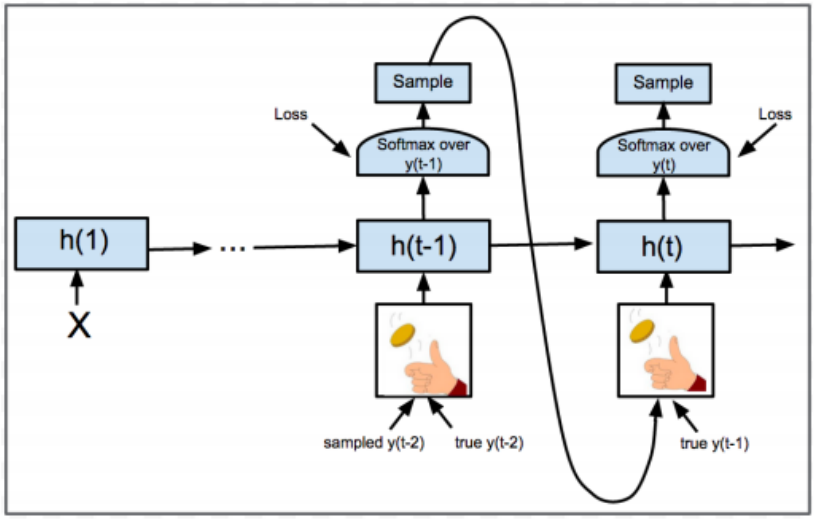
  https://arxiv.org/abs/1506.03099

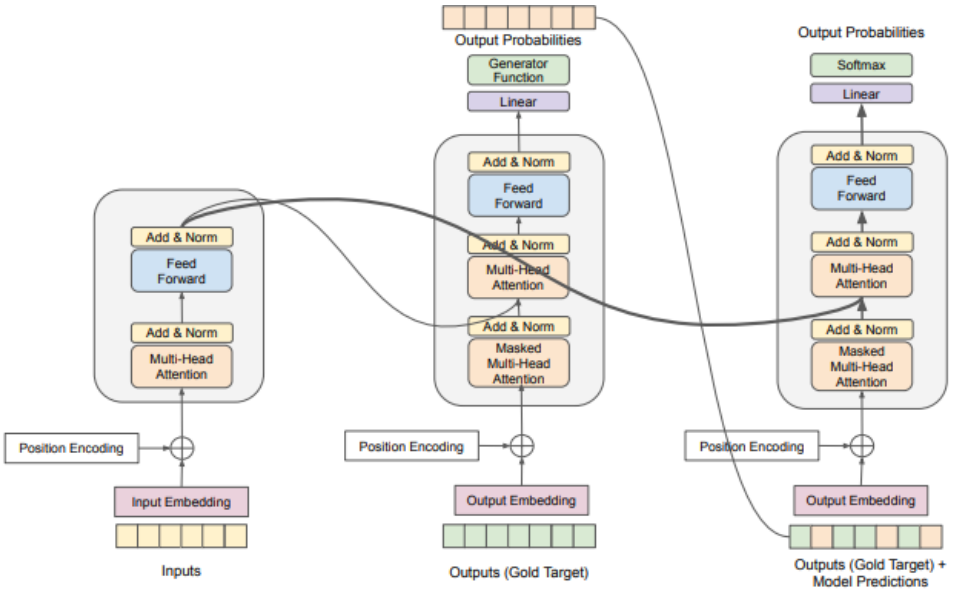
- Scheduled Sampling for Transformer

  https://arxiv.org/abs/1906.07651

- Parallel Scheduled Sampling

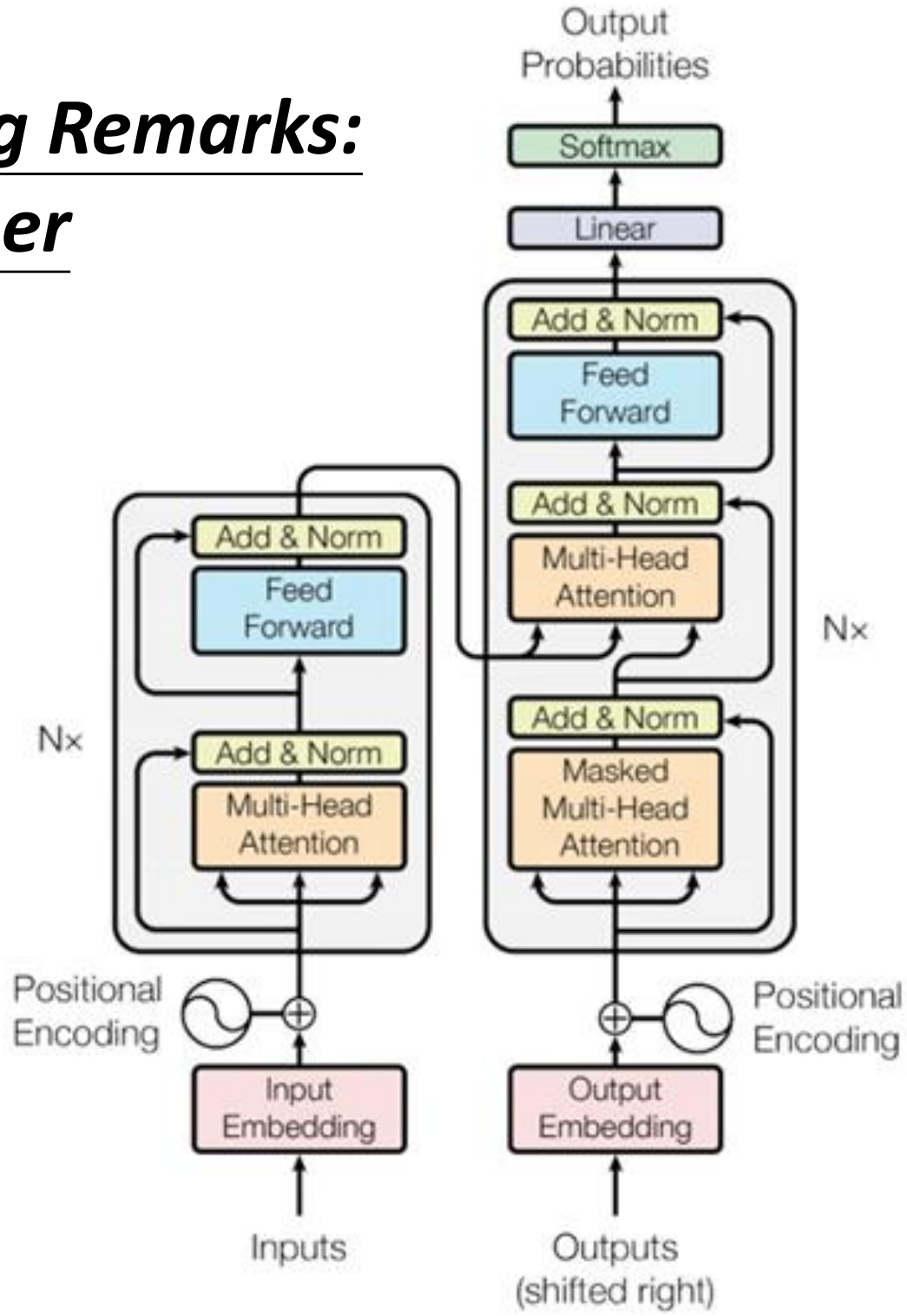
  https://arxiv.org/abs/1906.04331





Schedule Sampling

Concluding Remarks: Transformer



Q&A