Transformer

李宏毅

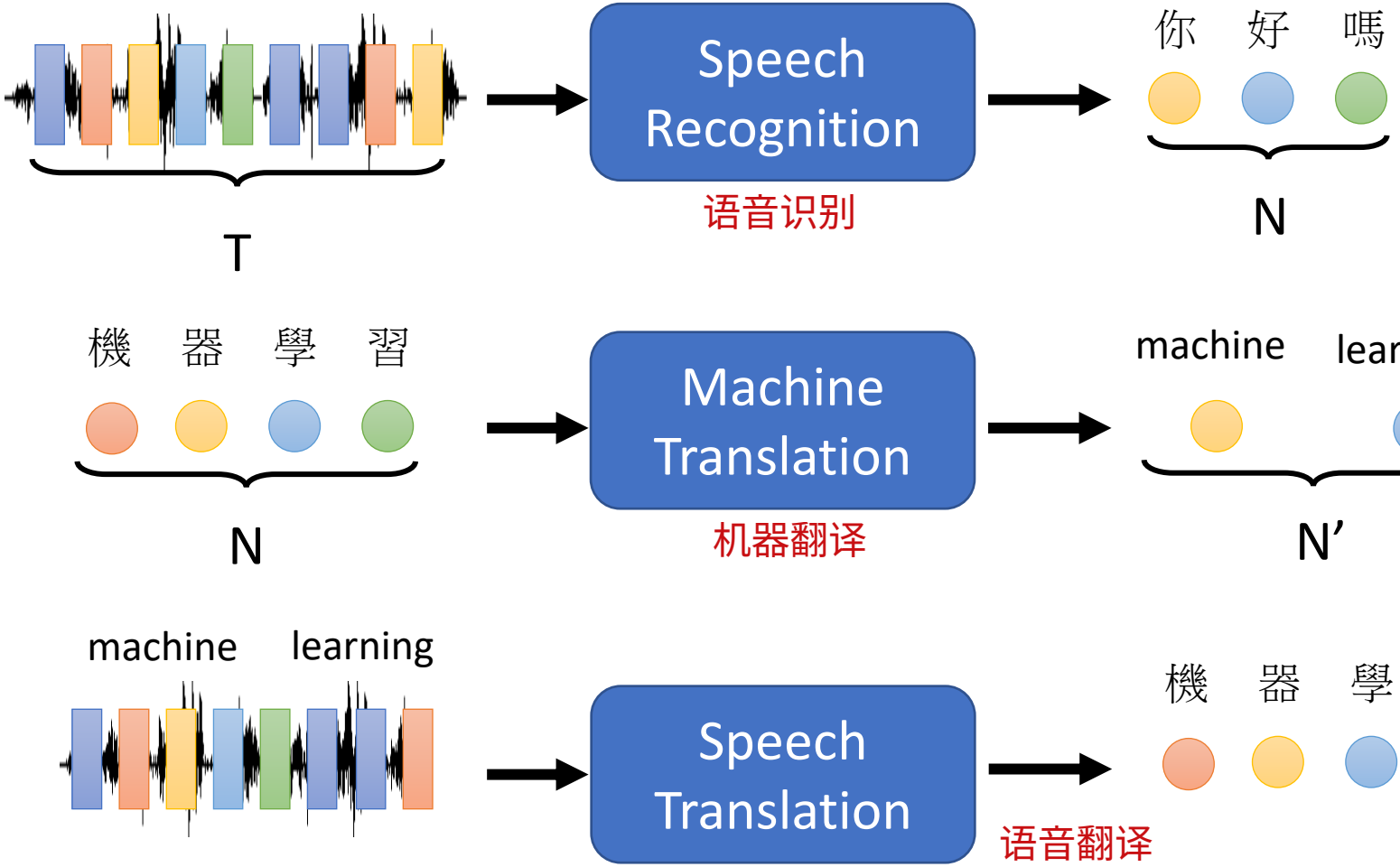
Hung-yi Lee

Transformer和
Bert很有关系

Sequence-to-sequence (Seq2seq) 应用

Input a sequence, output a sequence

The output length is determined by model.



Language without text

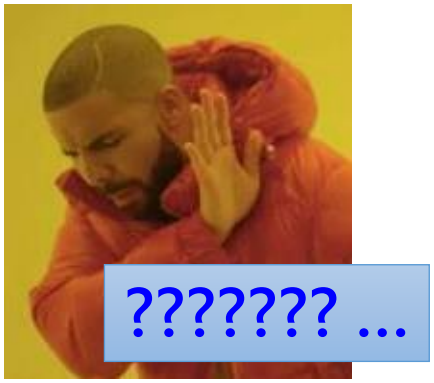
（世界上有很多语言没有文字）

Hokkien (閩南語、台語)

(Speech of Hokkien)

Speech Recognition

“母湯”



(Speech of Hokkien)

Speech Translation

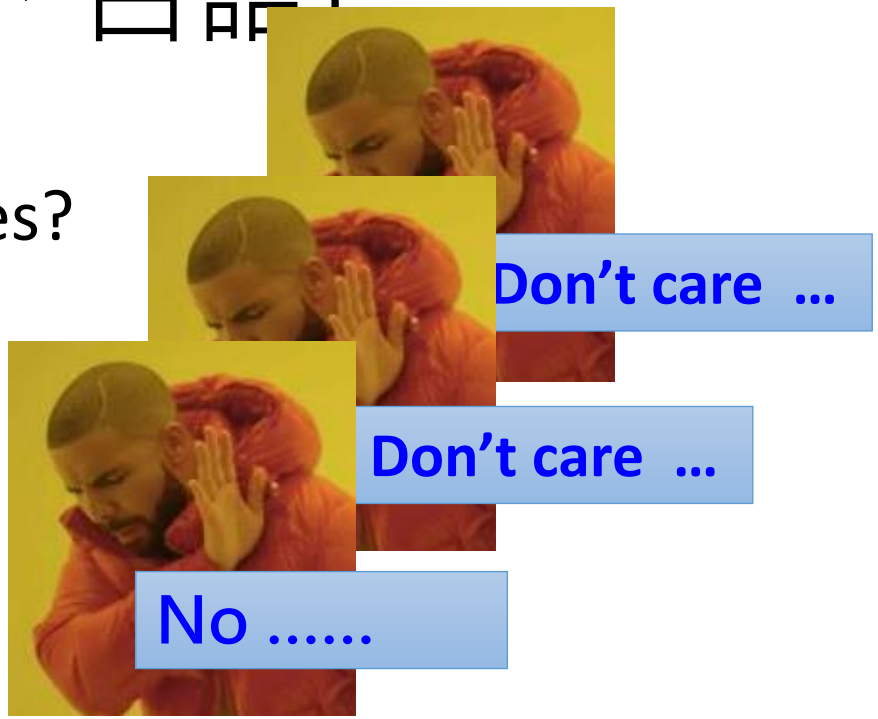
“不行”

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

Using 1500 hours of data for training

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)

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

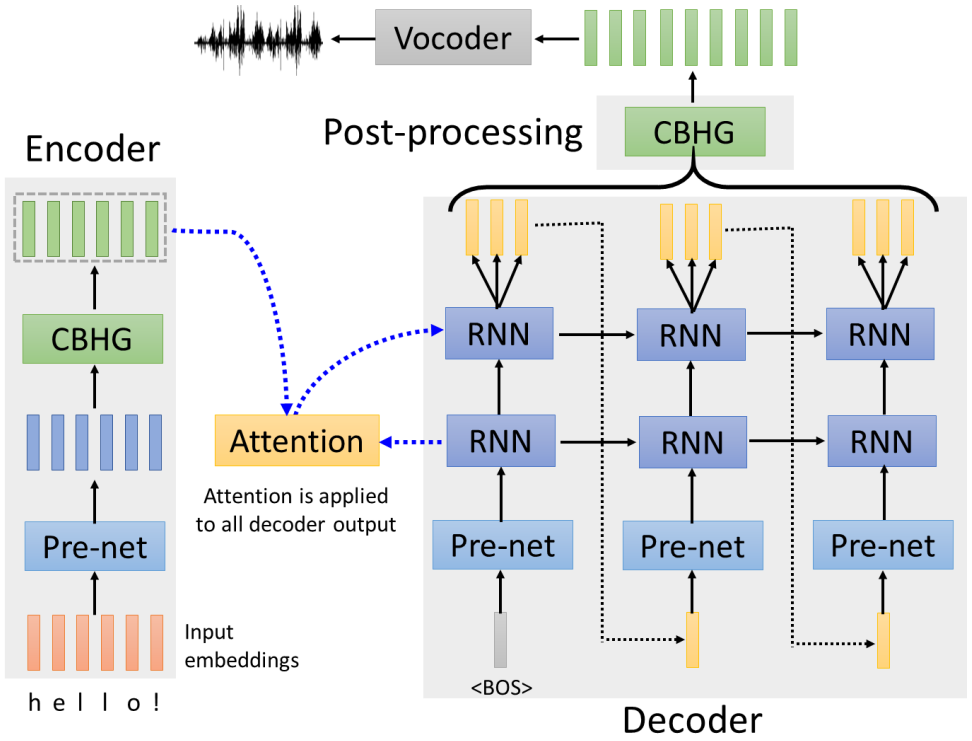
我 有 幫 廠 長 拜 託

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



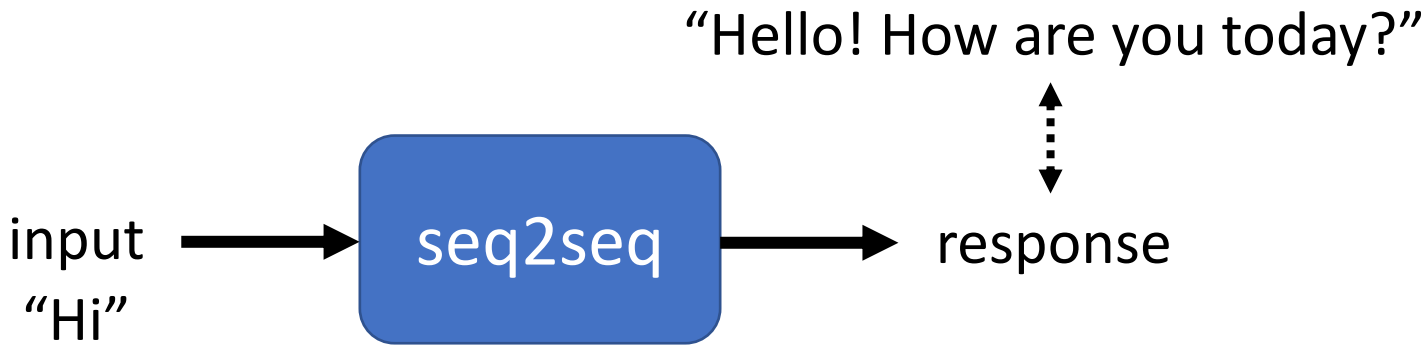
Taiwanese Speech Synthesis

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.

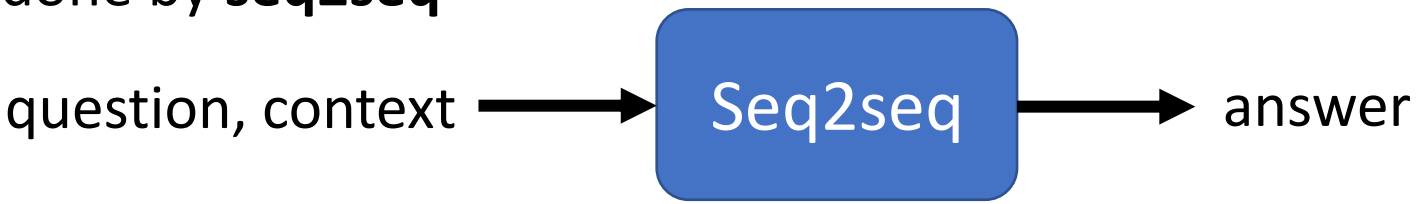
Most Natural Language Processing applications …

Question Answering (QA)

NLP上的任务都可以被看作是QA的问题

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? (Translation)	Most of the planet is ocean water.	Der Großteil der Erde ist Meerwasser
What is the summary? (summarization)	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? (Classification)	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

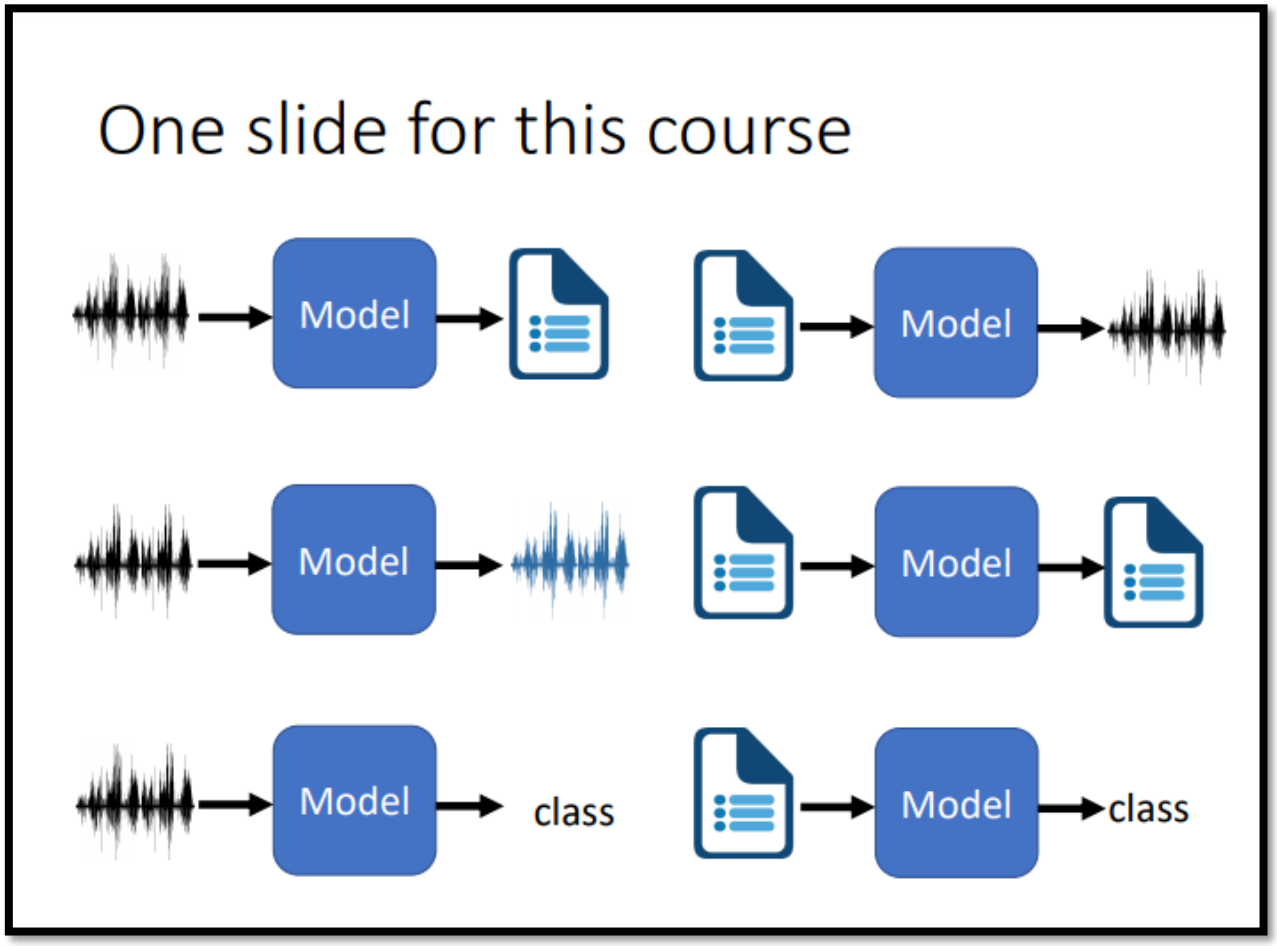
QA can be done by **seq2seq**



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

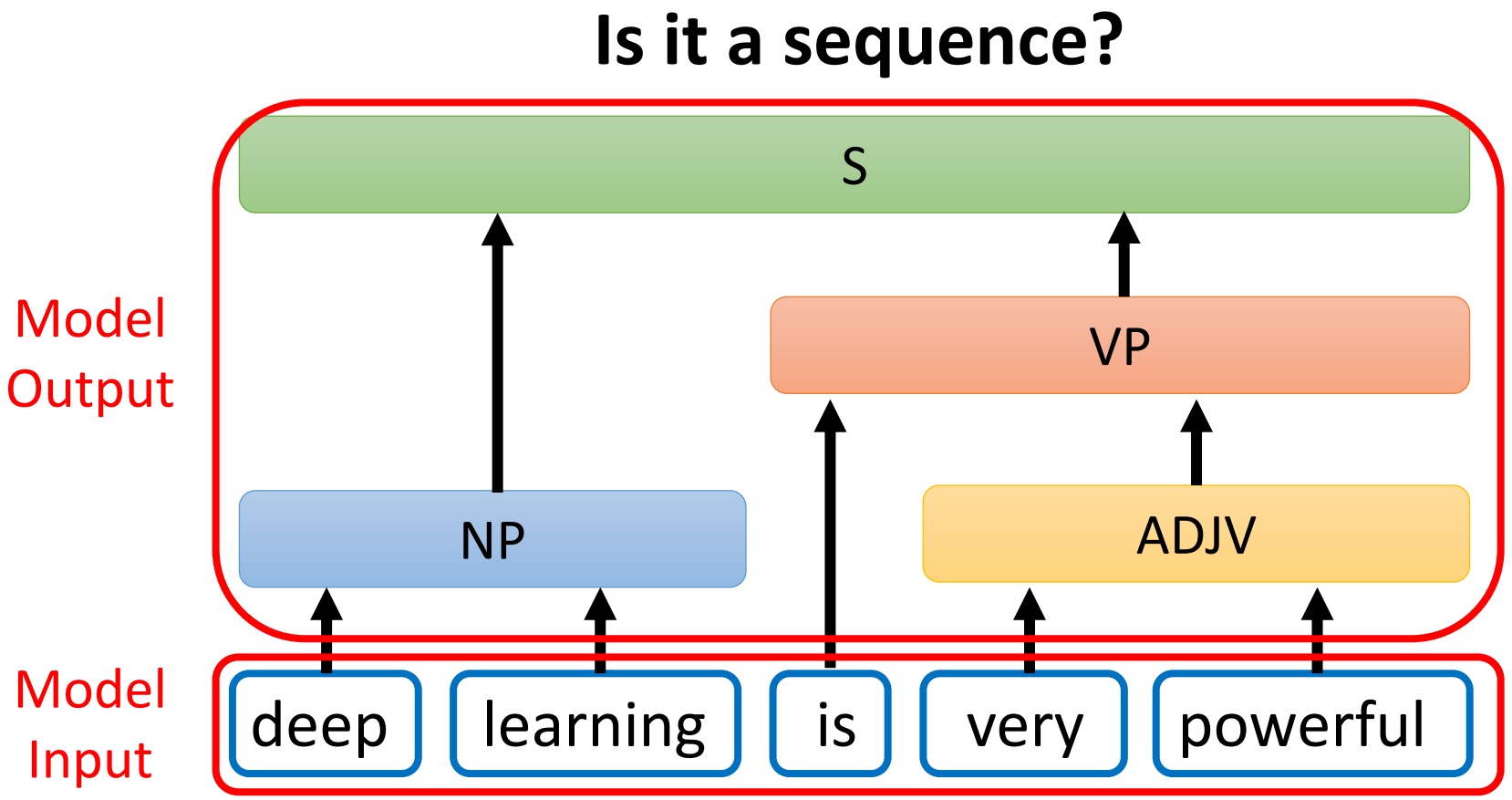
Deep Learning for Human Language Processing
深度學習與人類語言處理

针对NLP上的任务，客制化模型的话，往往会得到更好的效果。直接丢进Seq2seq的话，相当于用瑞士军刀。

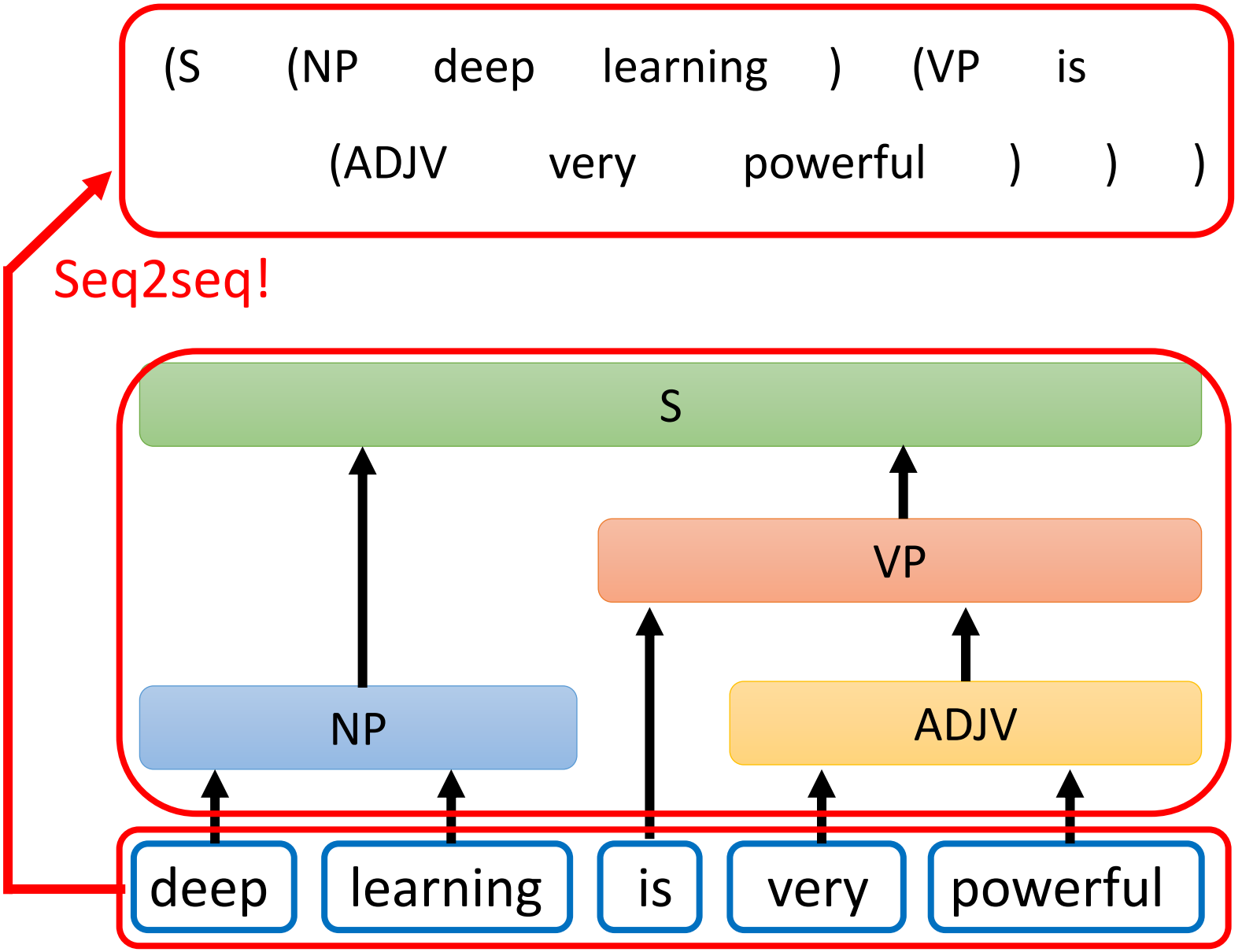


Source webpage: https://speech.ee.ntu.edu.tw/~hylee/dlhlp/2020-spring.html

Seq2seq for Syntactic Parsing 文法剖析



Seq2seq for Syntactic Parsing



Seq2seq for Syntactic Parsing

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

Grammar as a Foreign Language

没有tips，Adam都没用

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

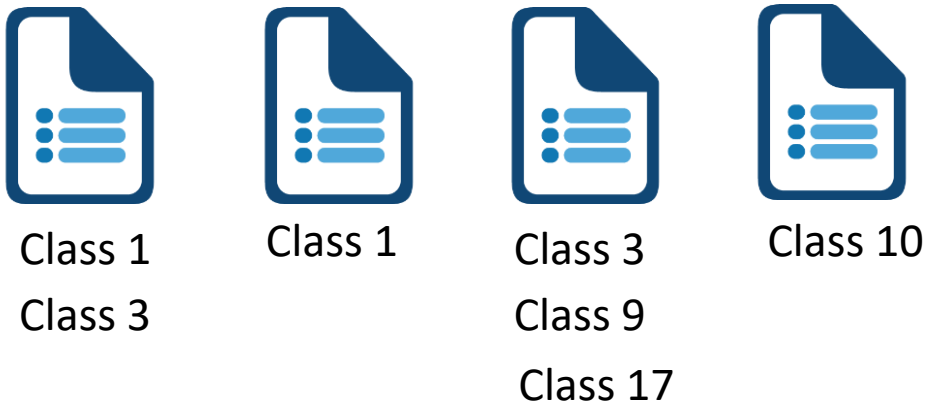
上古神兽

deep learning is very powerful

Seq2seq for Multi-label Classification

c.f. Multi-class Classification
多元分类。
在数个class中选某个class

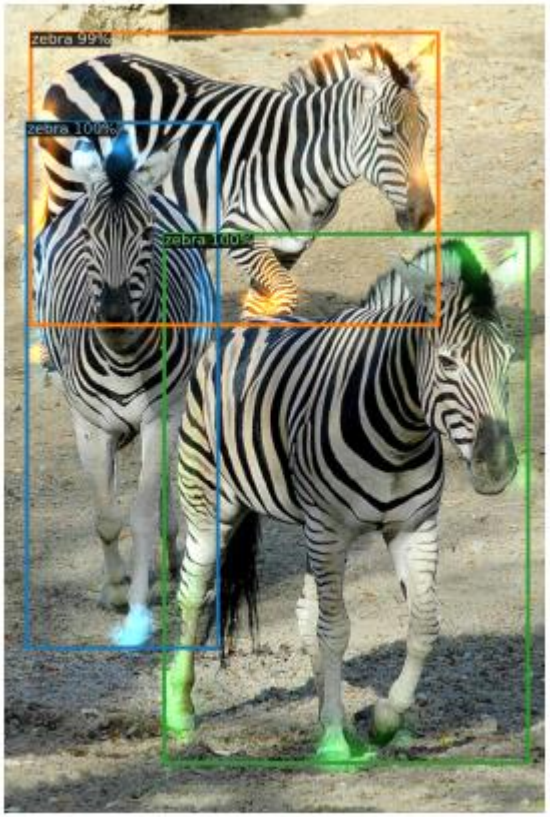
An object can belong to multiple classes.

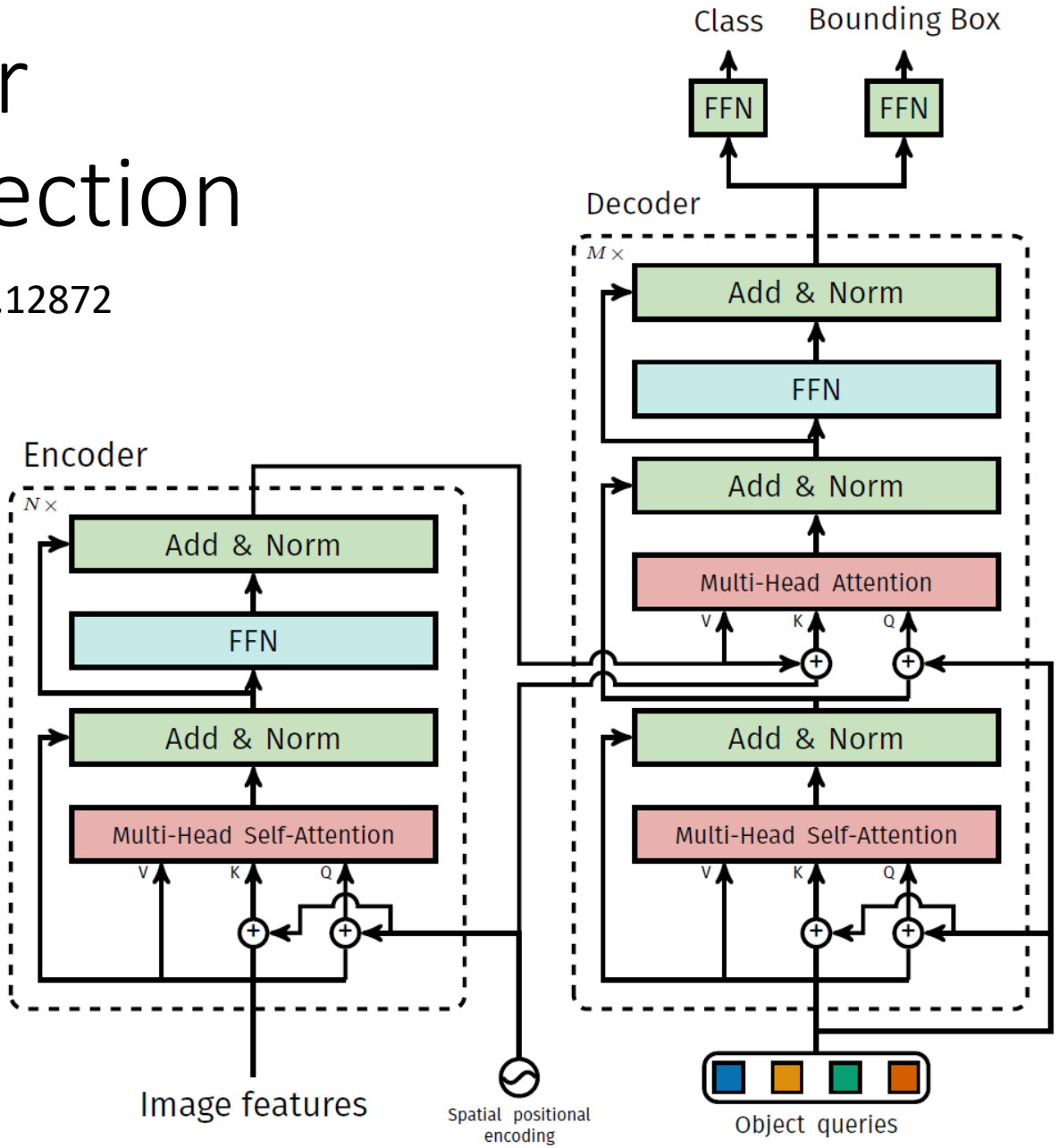


https://arxiv.org/abs/1909.03434
https://arxiv.org/abs/1707.05495

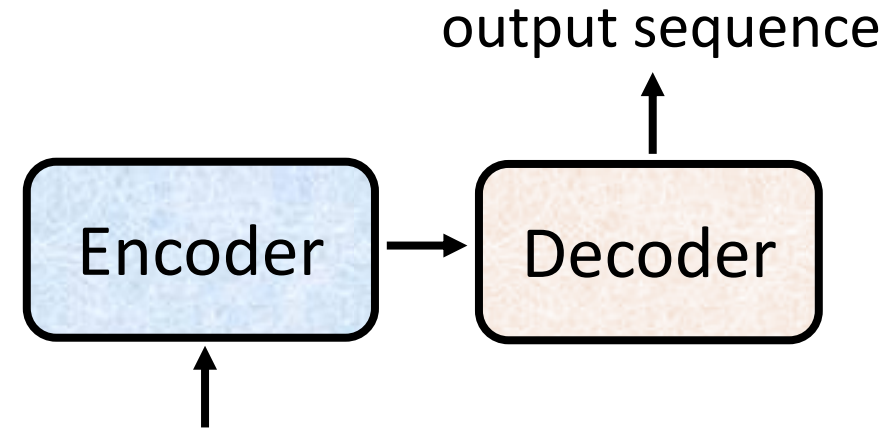
Seq2seq for Object Detection

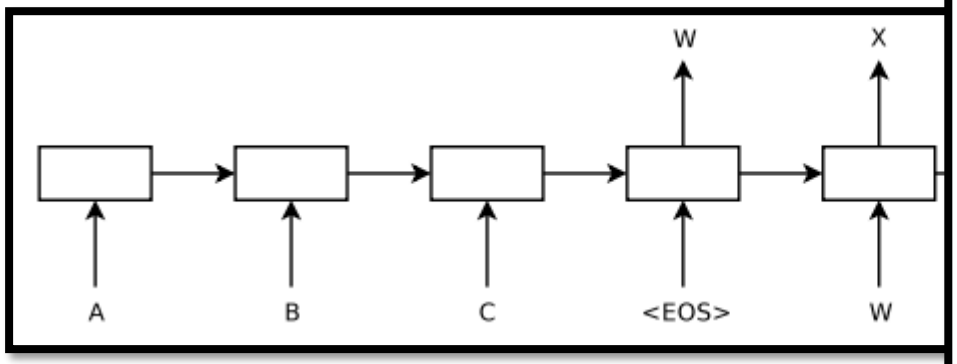
https://arxiv.org/abs/2005.12872





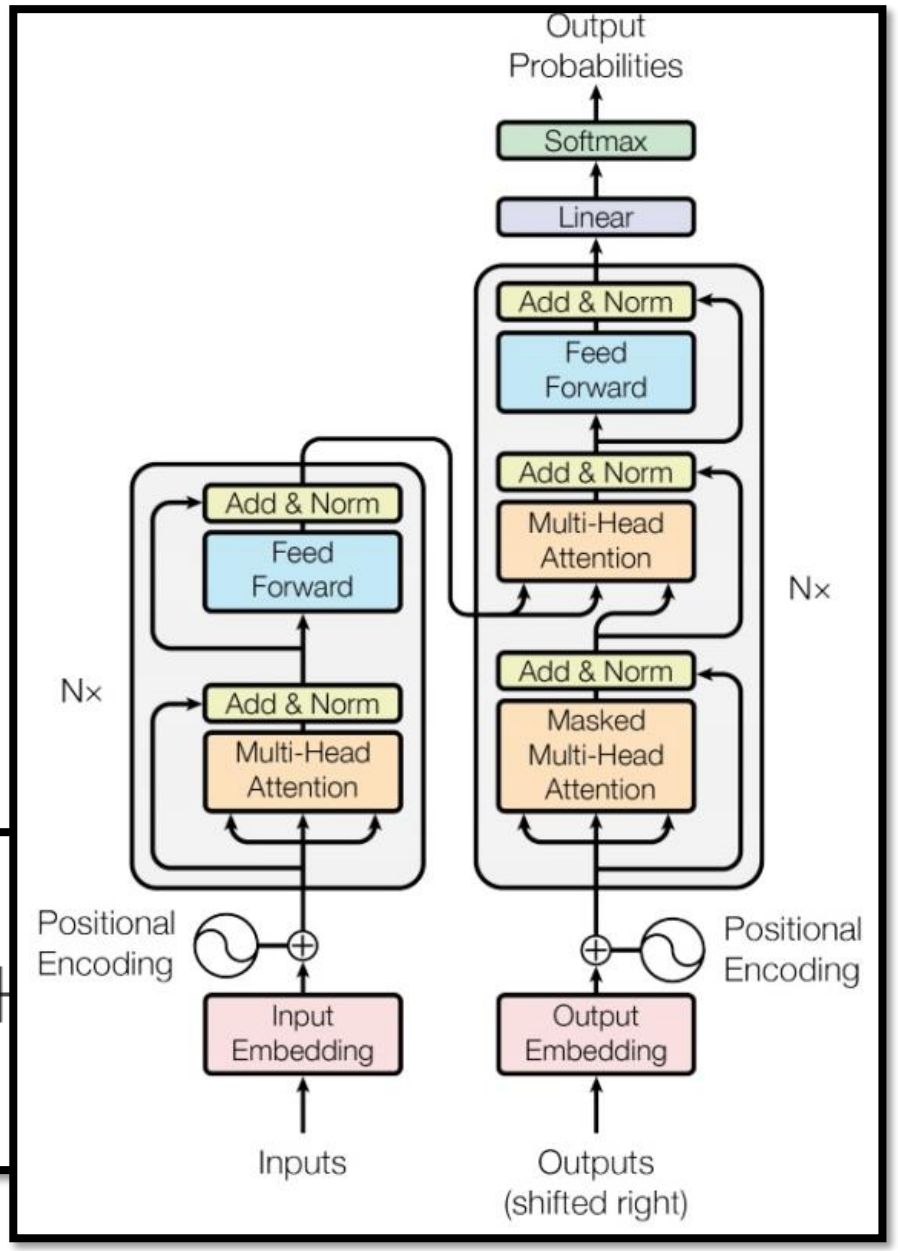
Seq2seq





Sequence to Sequence Learning with Neural Networks

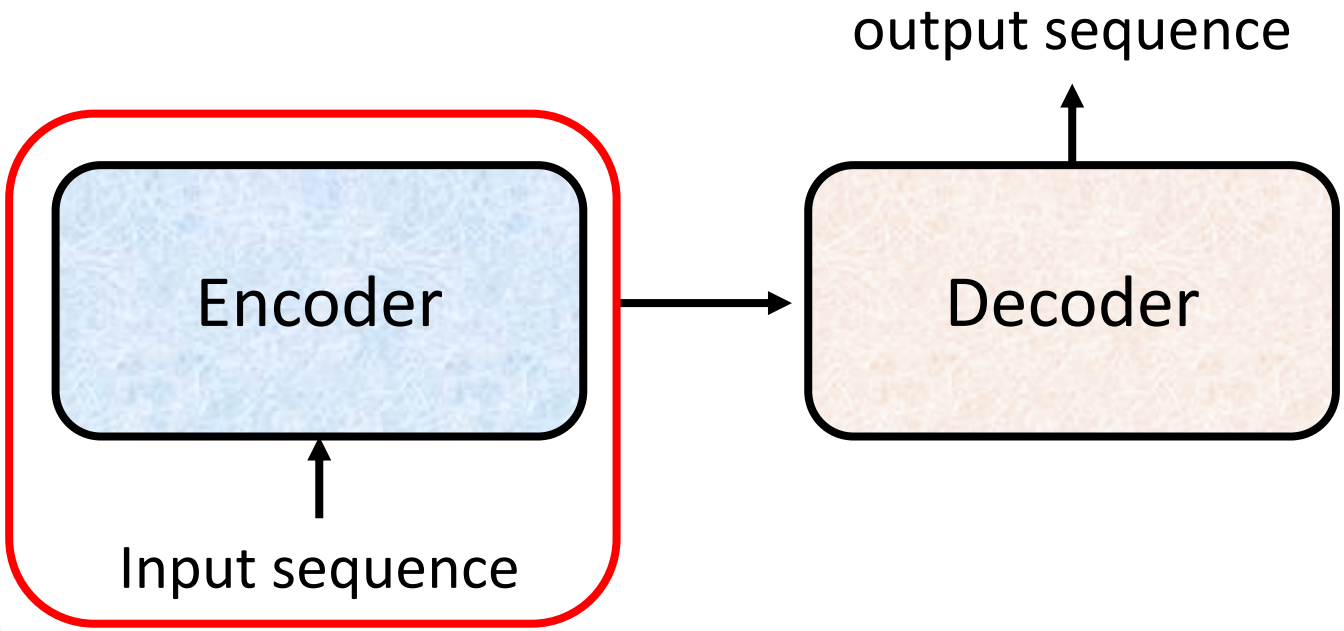
https://arxiv.org/abs/1409.3215



Transformer

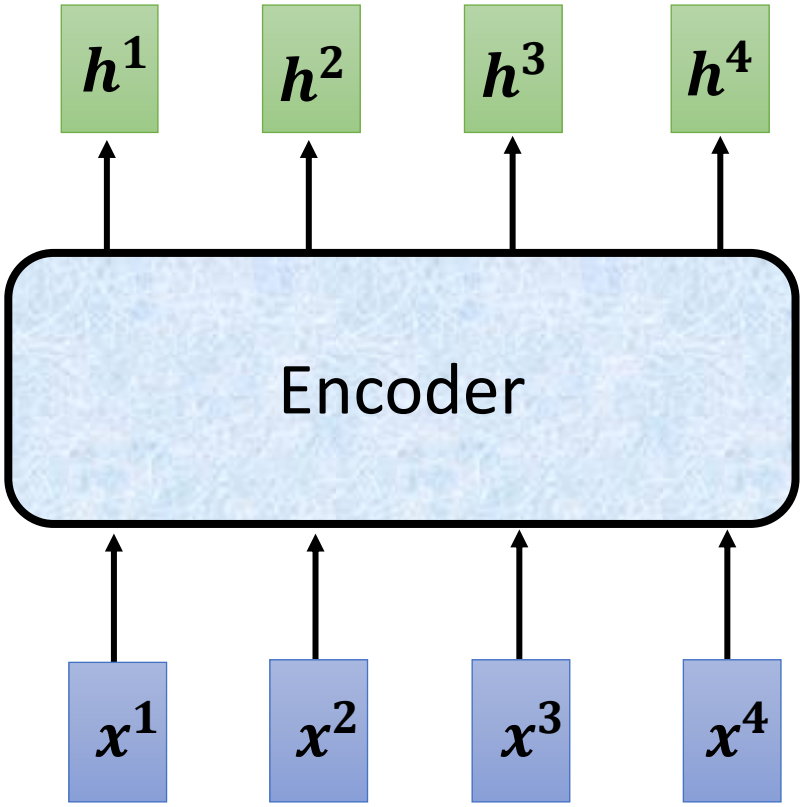
https://arxiv.org/abs/1706.03762

Encoder

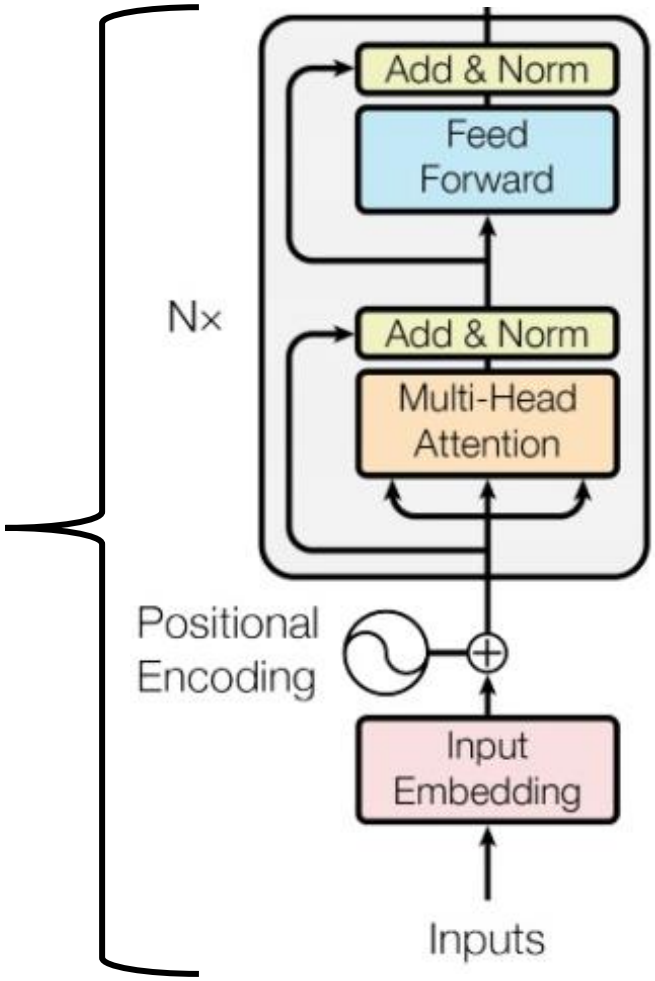


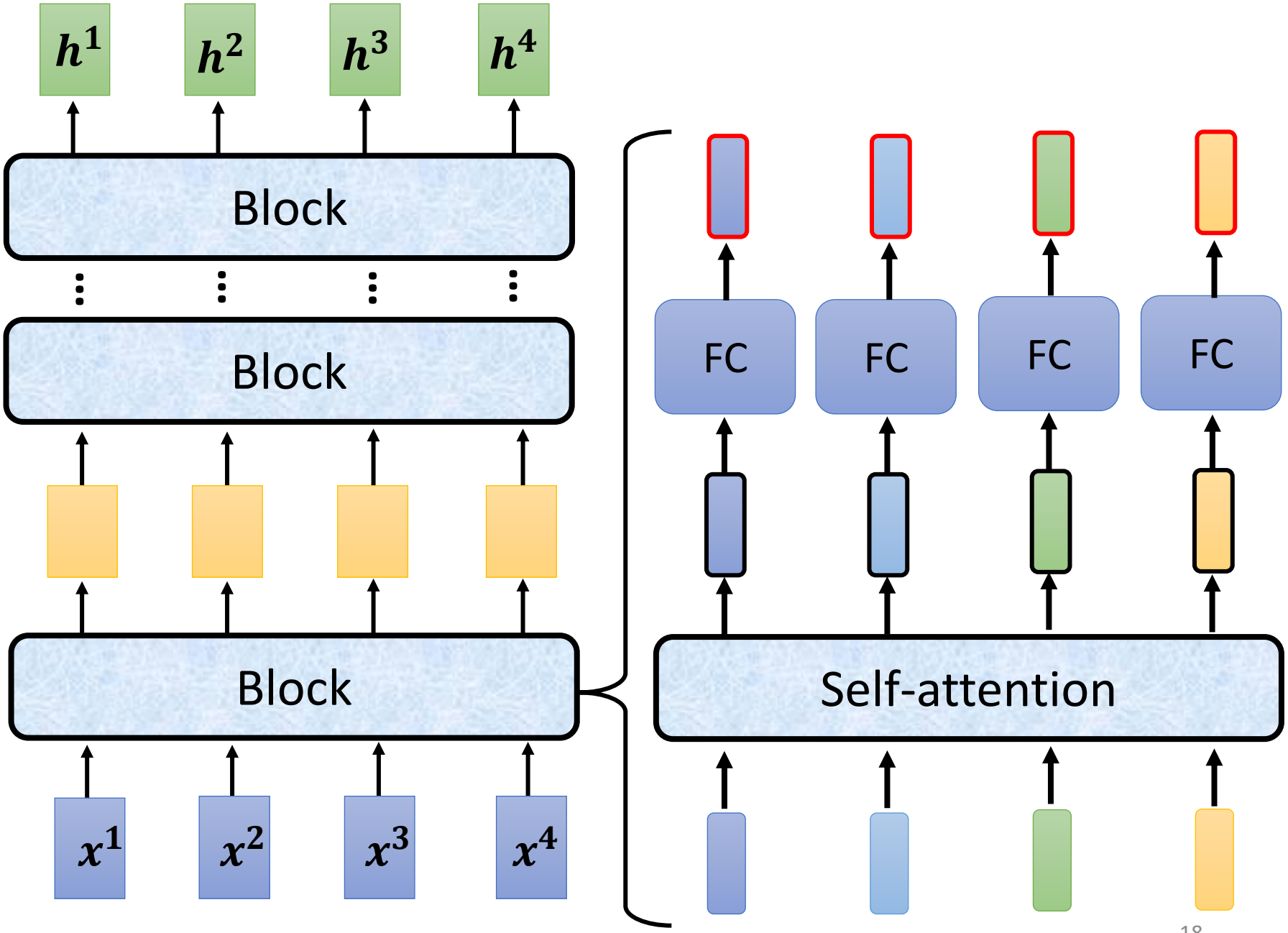
Encoder

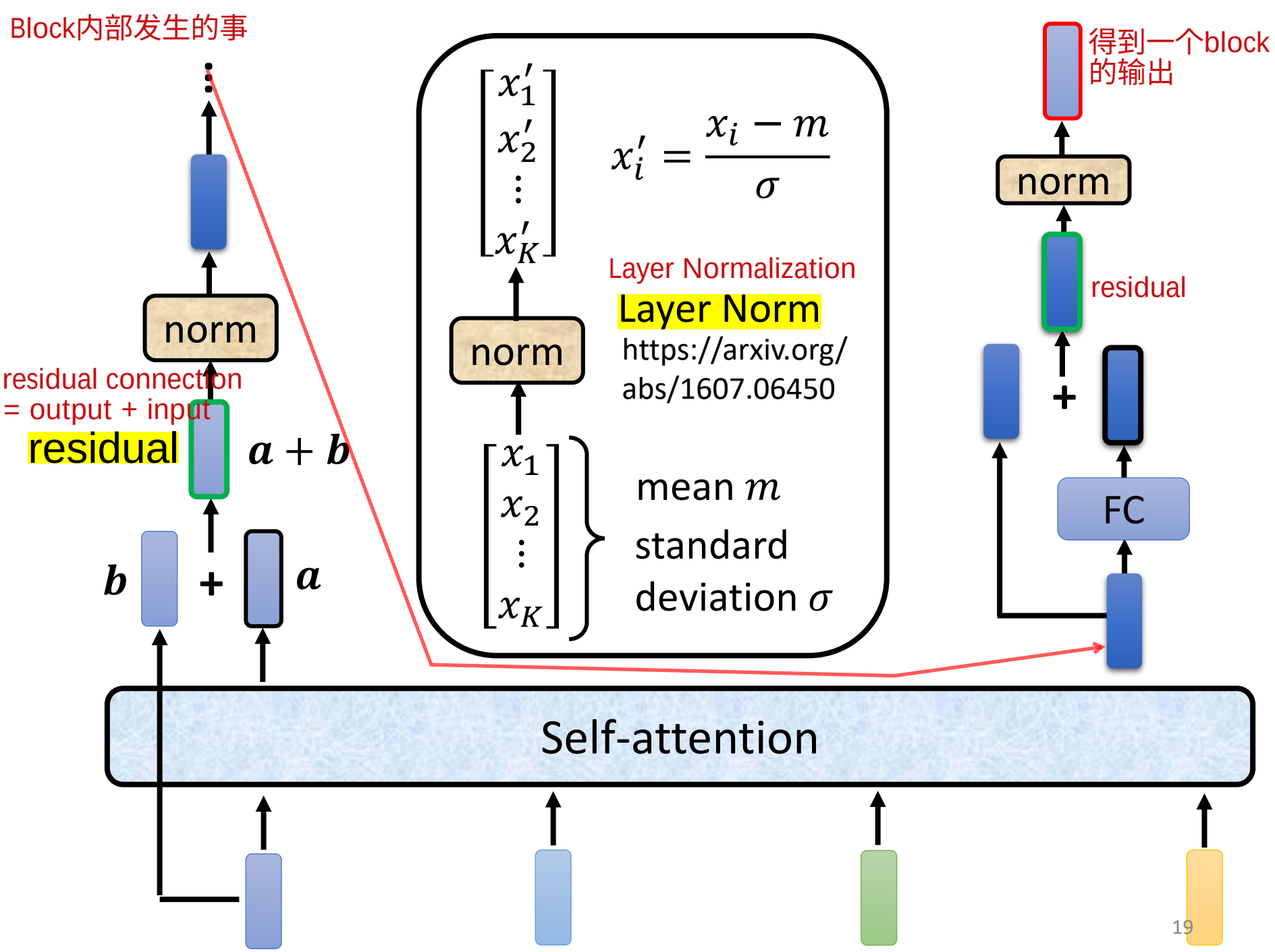
You can use **RNN** or **CNN**.



Transformer's Encoder

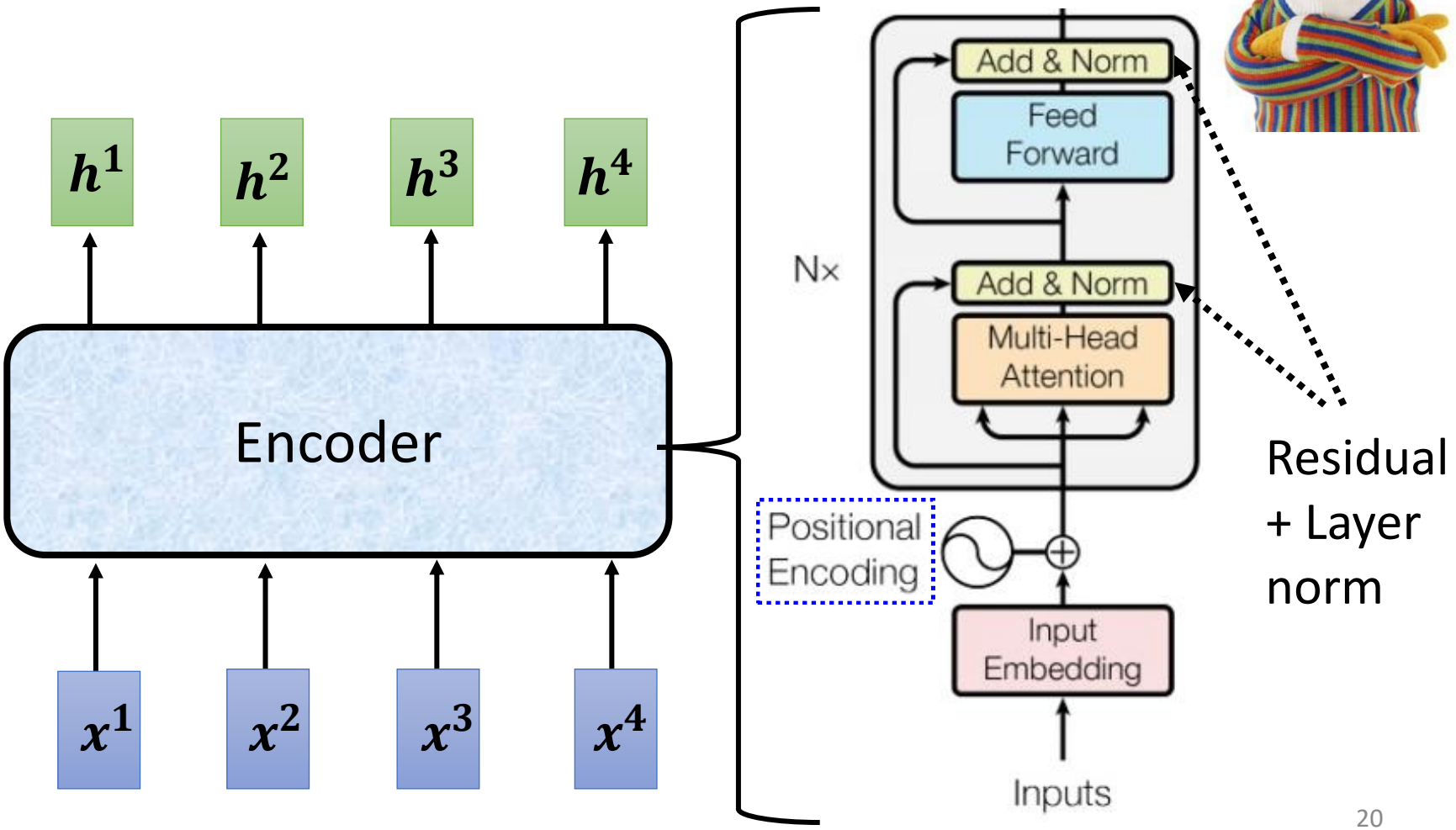



h^1
h^2
h^3
h^4
Block
Block
Block
x^1
x^2
x^3
x^4
FC
FC
FC
FC
Self-attention


Block内部发生的事
得到一个block的输出
norm
residual
FC
residual connection
= output + input
residual
$a+b$
b
a
$x_1' \; x_2' \; \vdots \; x_K'$
$x_i' = \frac{x_i - m}{\sigma}$
Layer Normalization
Layer Norm
https://arxiv.org/abs/1607.06450
$x_1 \; x_2 \; \vdots \; x_K$
mean m
standard deviation σ
Self-attention

BERT

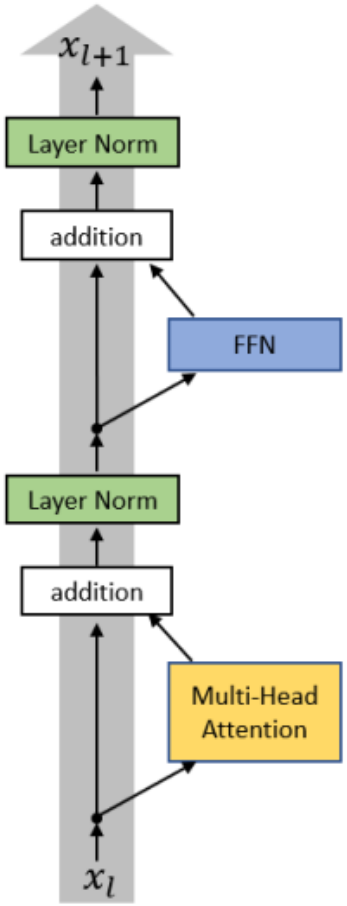
I use the **same** network architecture as **transformer encoder**.



To learn more ……

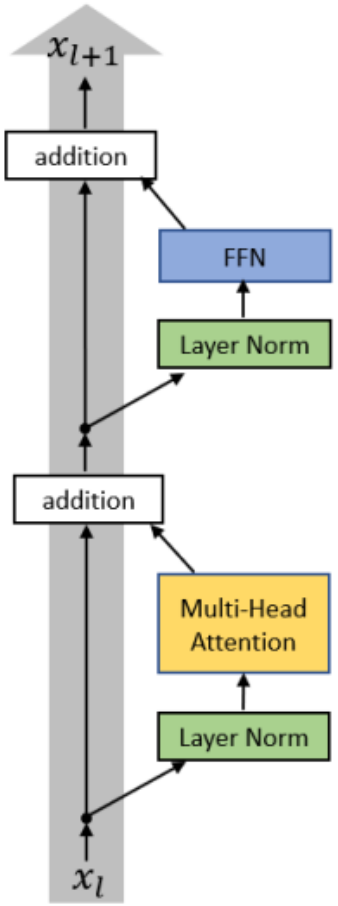
- On Layer Normalization in the Transformer Architecture
- https://arxiv.org/abs/2002.04745

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



(a)

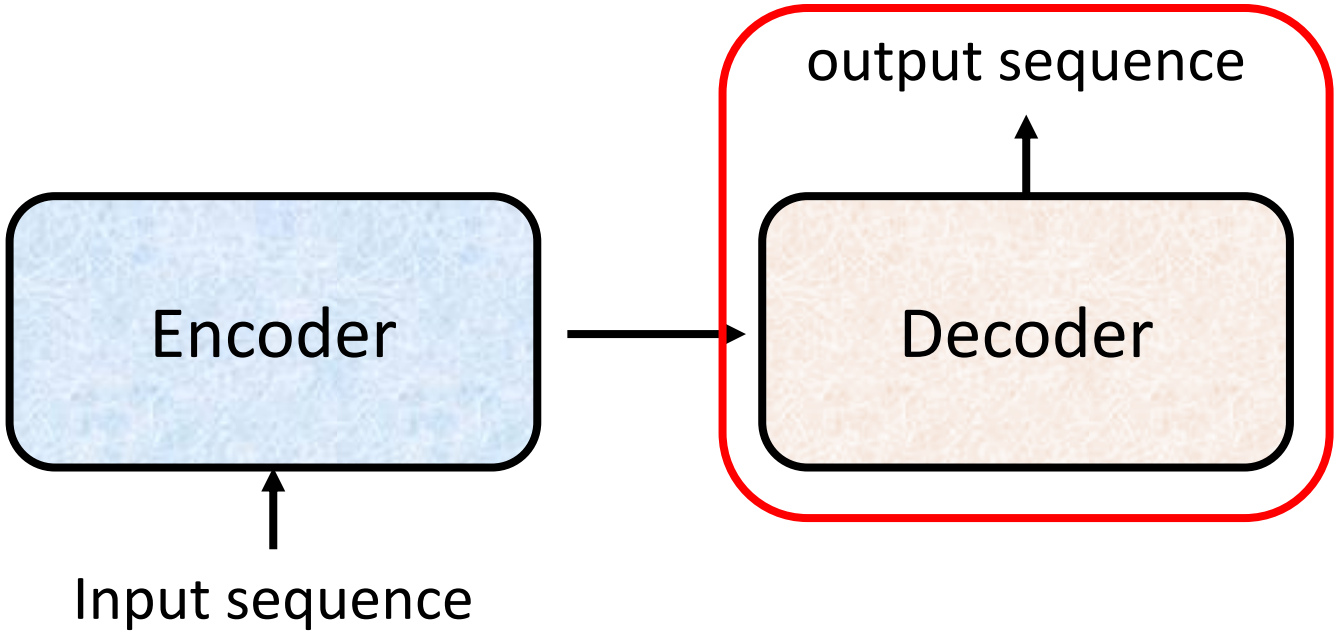
原始transformer



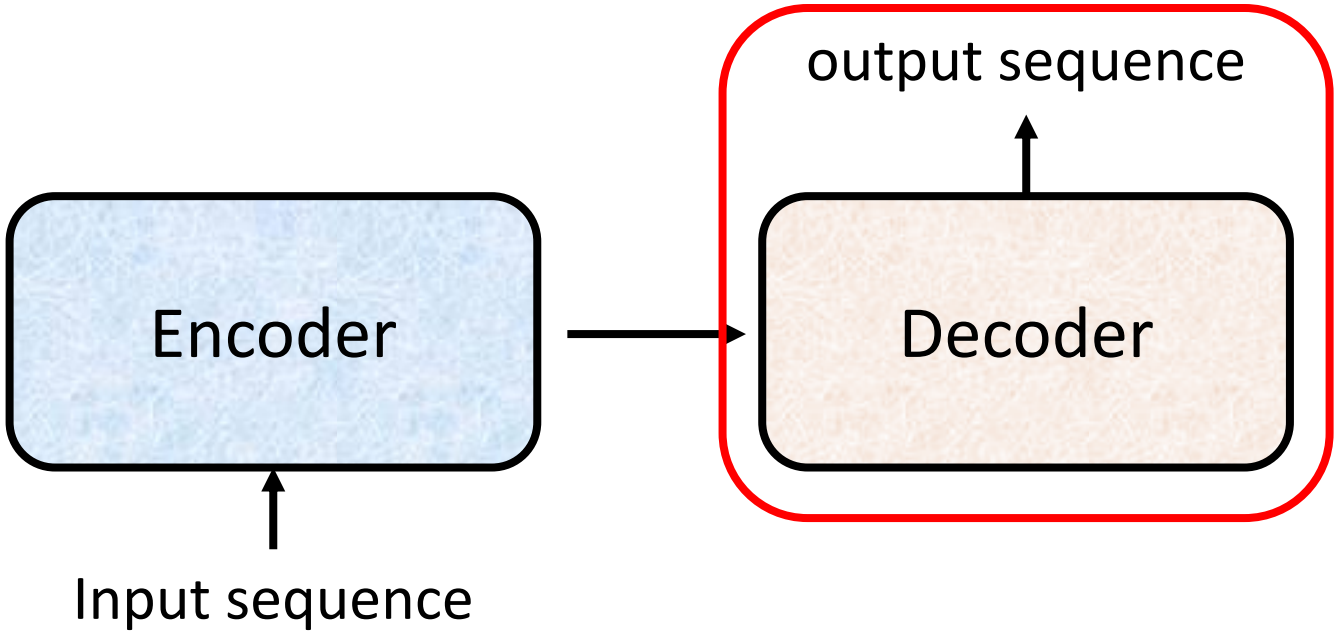
(b)

改良后：residual和norm的位置换一下

Decoder

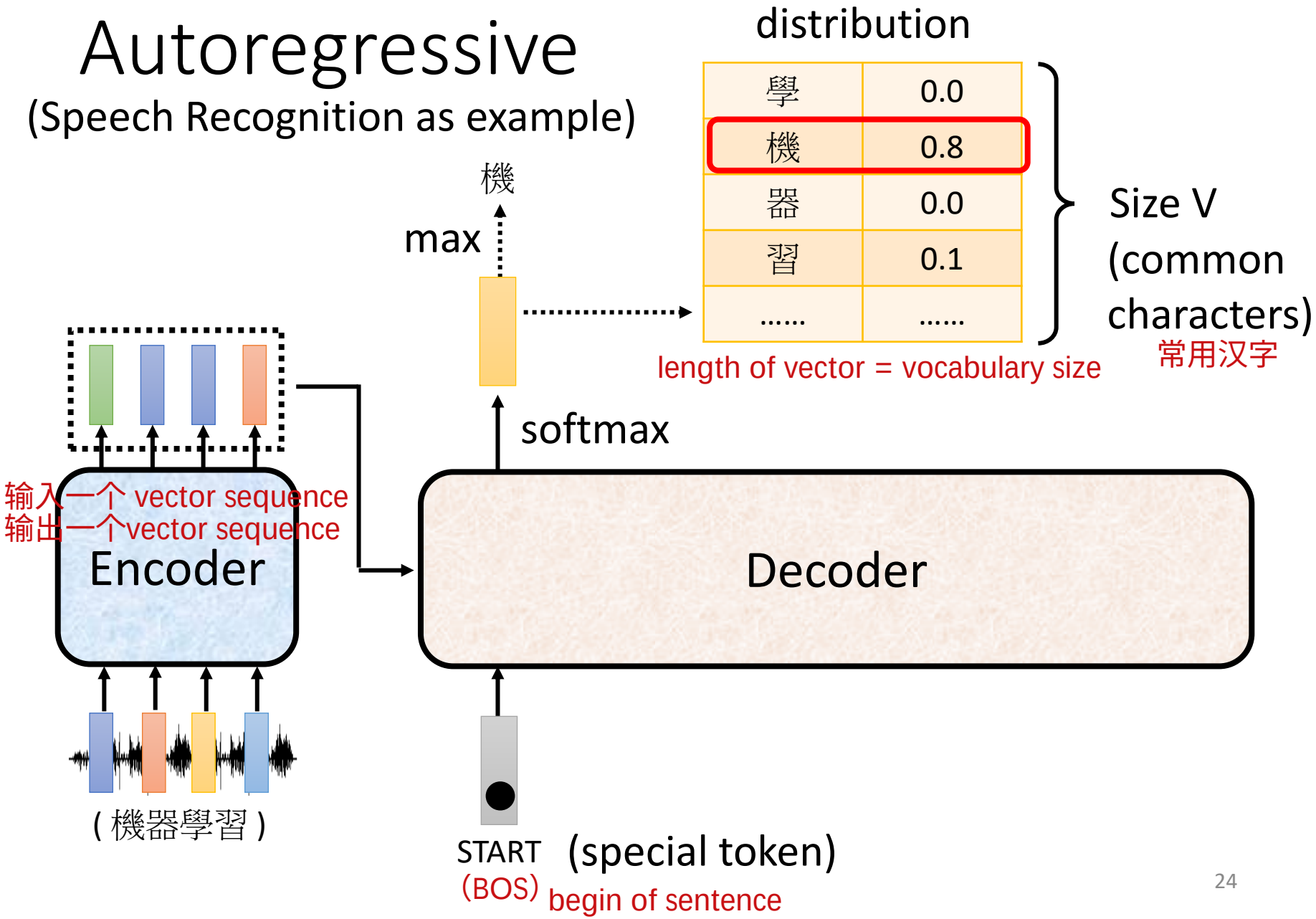


Decoder
– Autoregressive (AT)

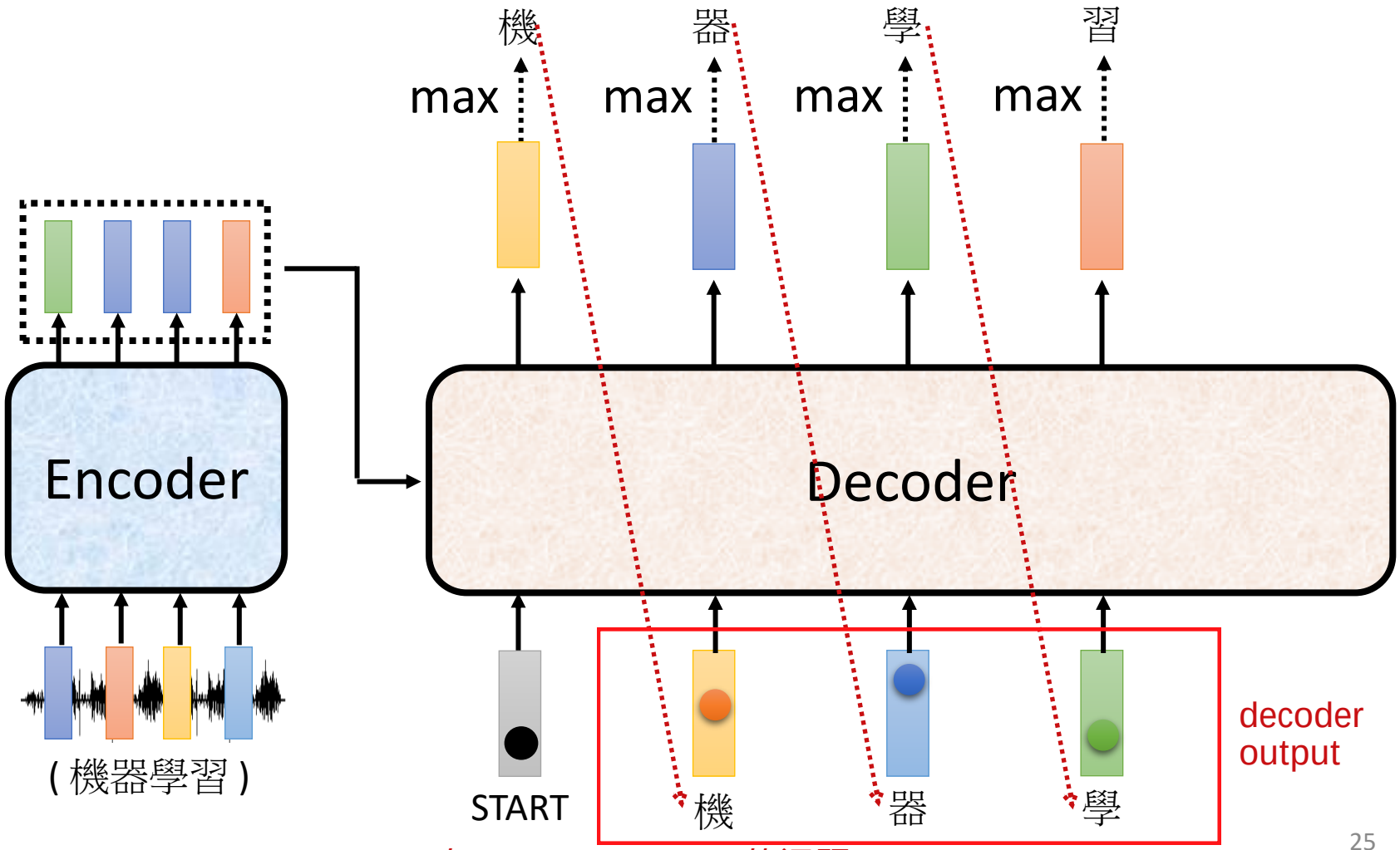


Autoregressive

(Speech Recognition as example)

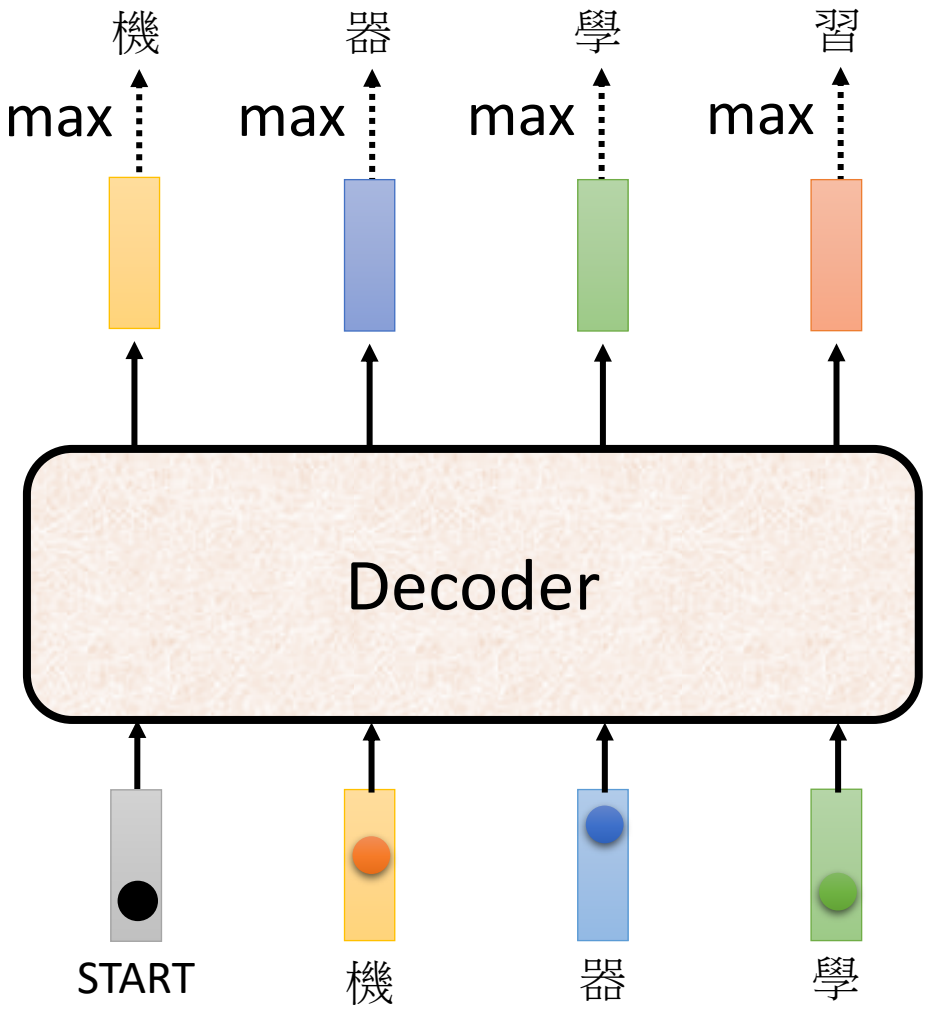


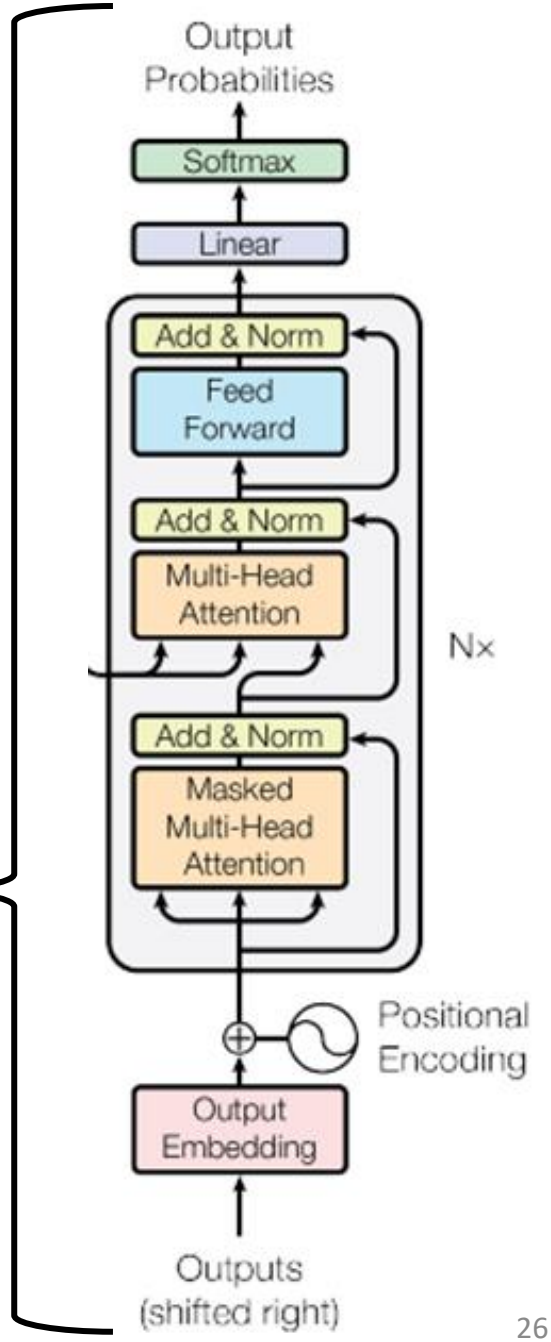
Autoregressive



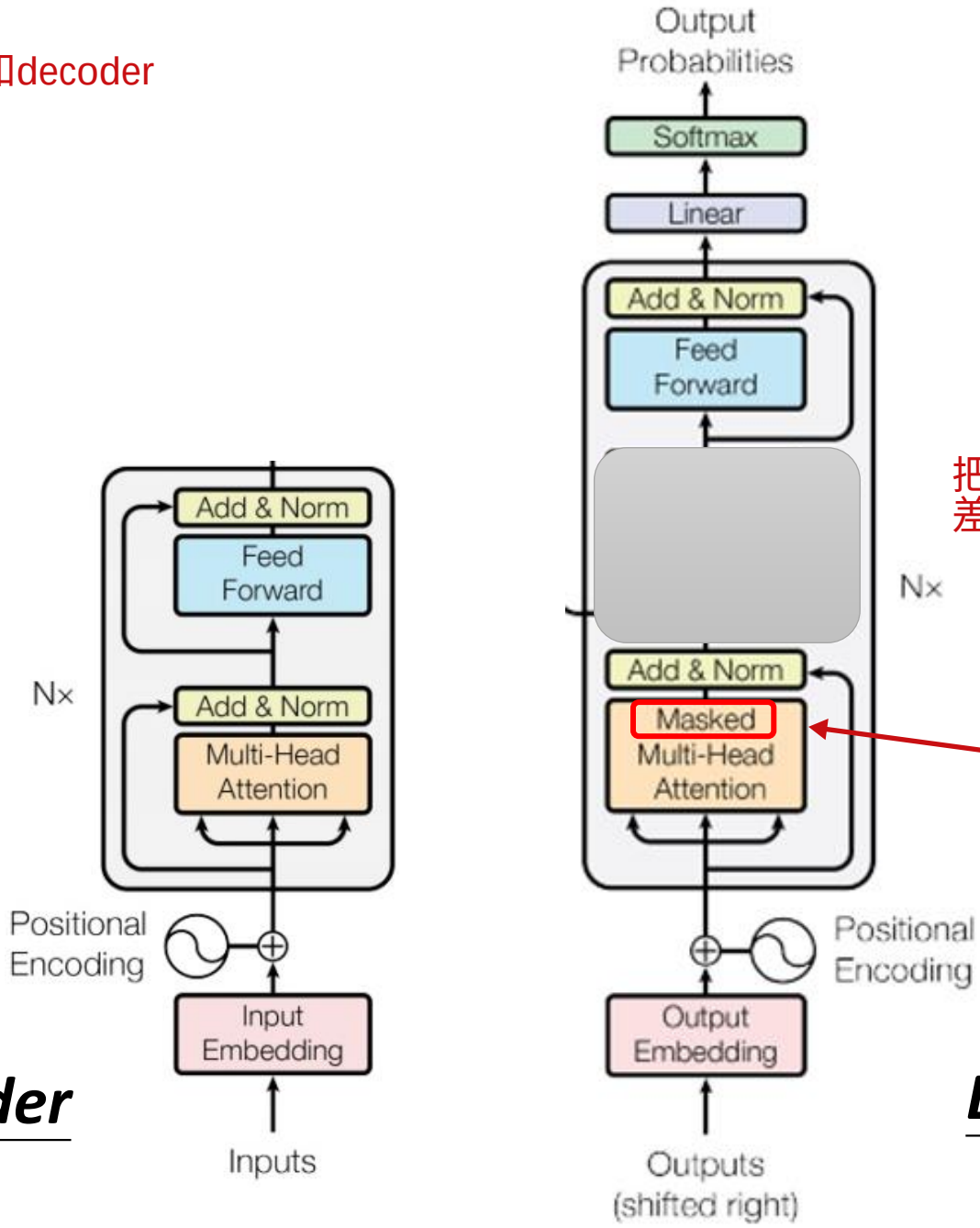
有error propagation的问题：use uncertain inputs to calculate

ignore the input from the encoder here ☺





比较一下encoder和decoder

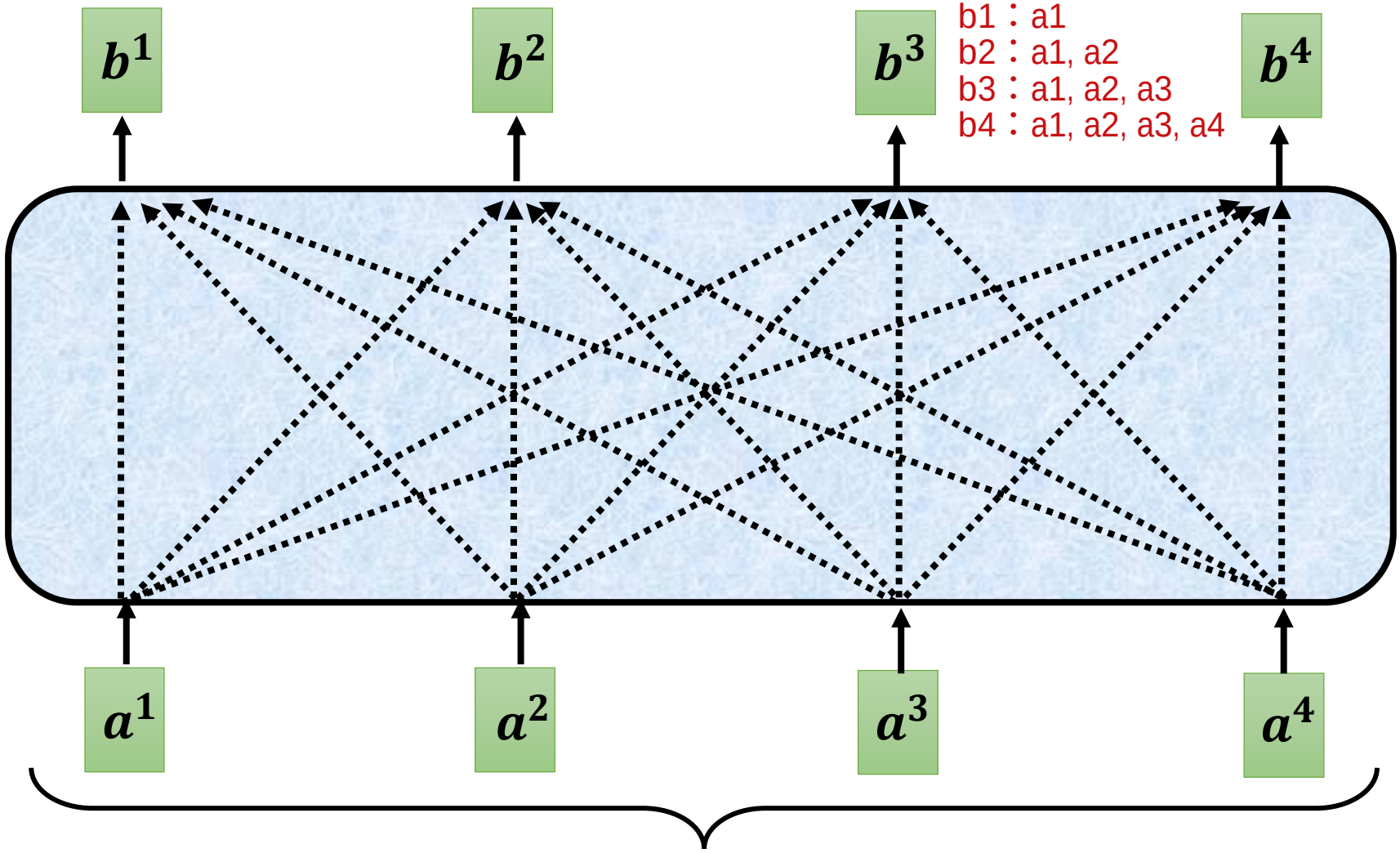


把这块遮起来之后，差别不大

masked attention

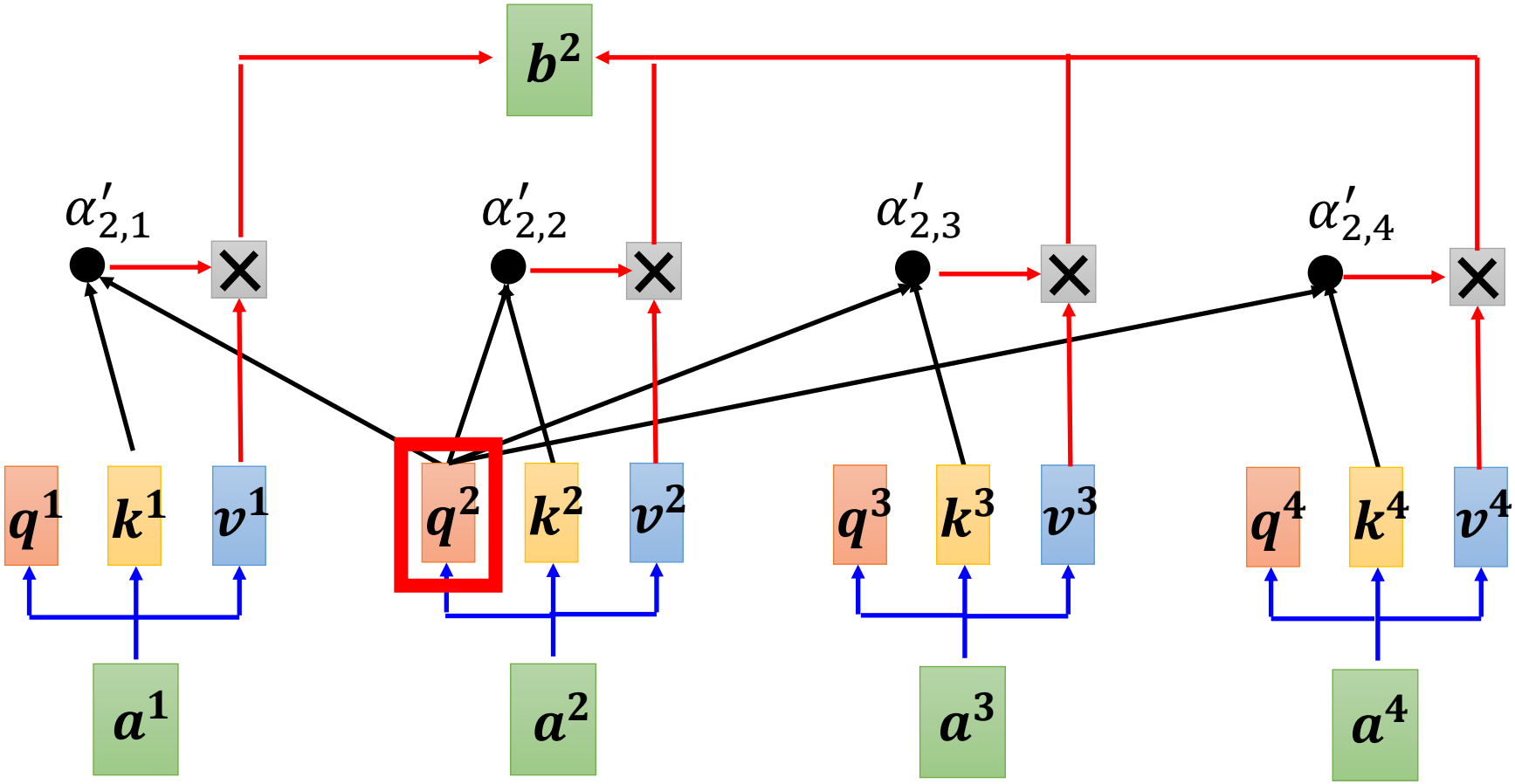
Self-attention → *Masked Self-attention*

只考虑左边的，不考虑右边的



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

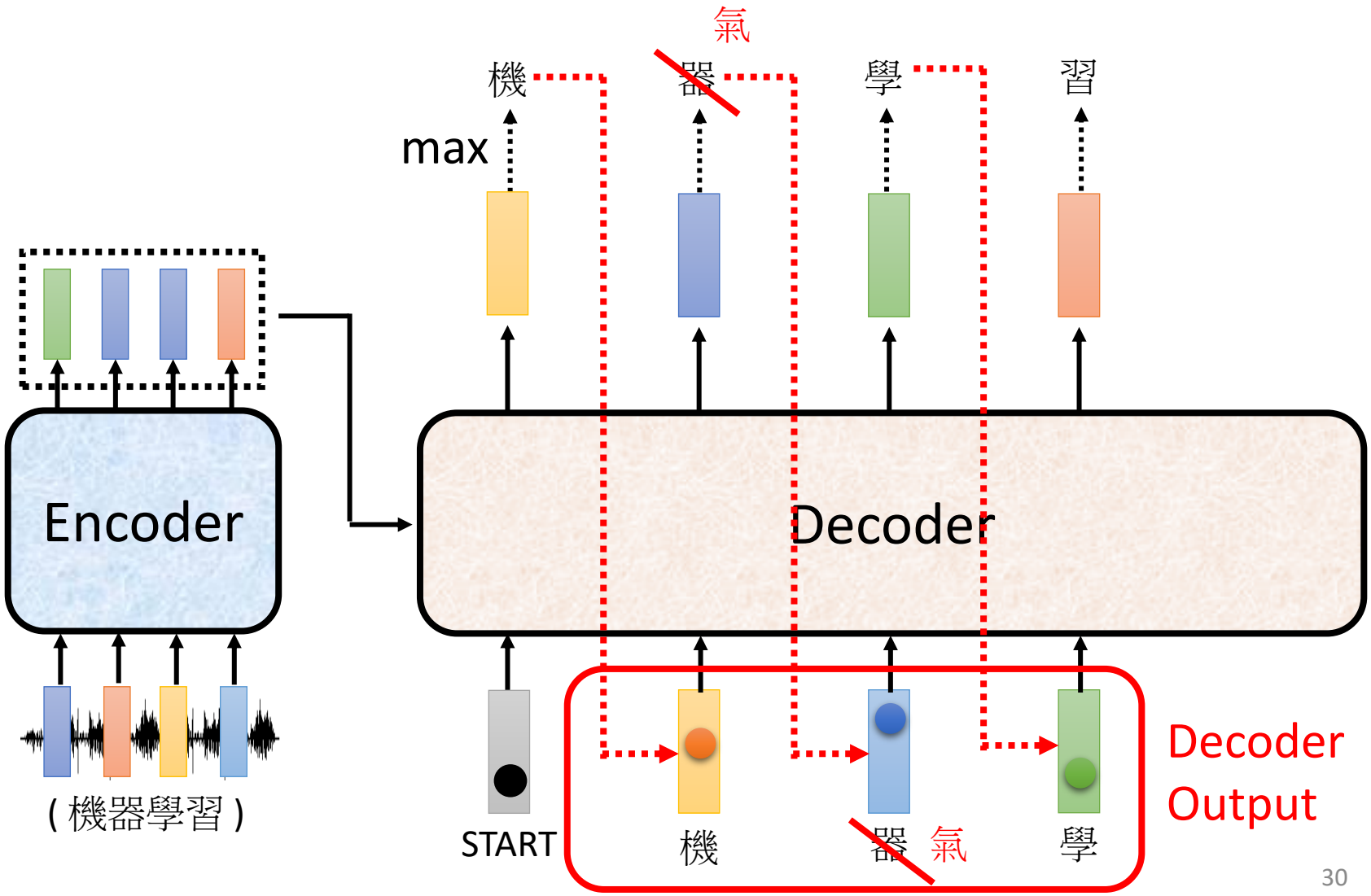
Self-attention → Masked Self-attention



Why masked? Consider how does decoder work

encoder的输入是一个一个产生的
所以decoder要一个一个出结果

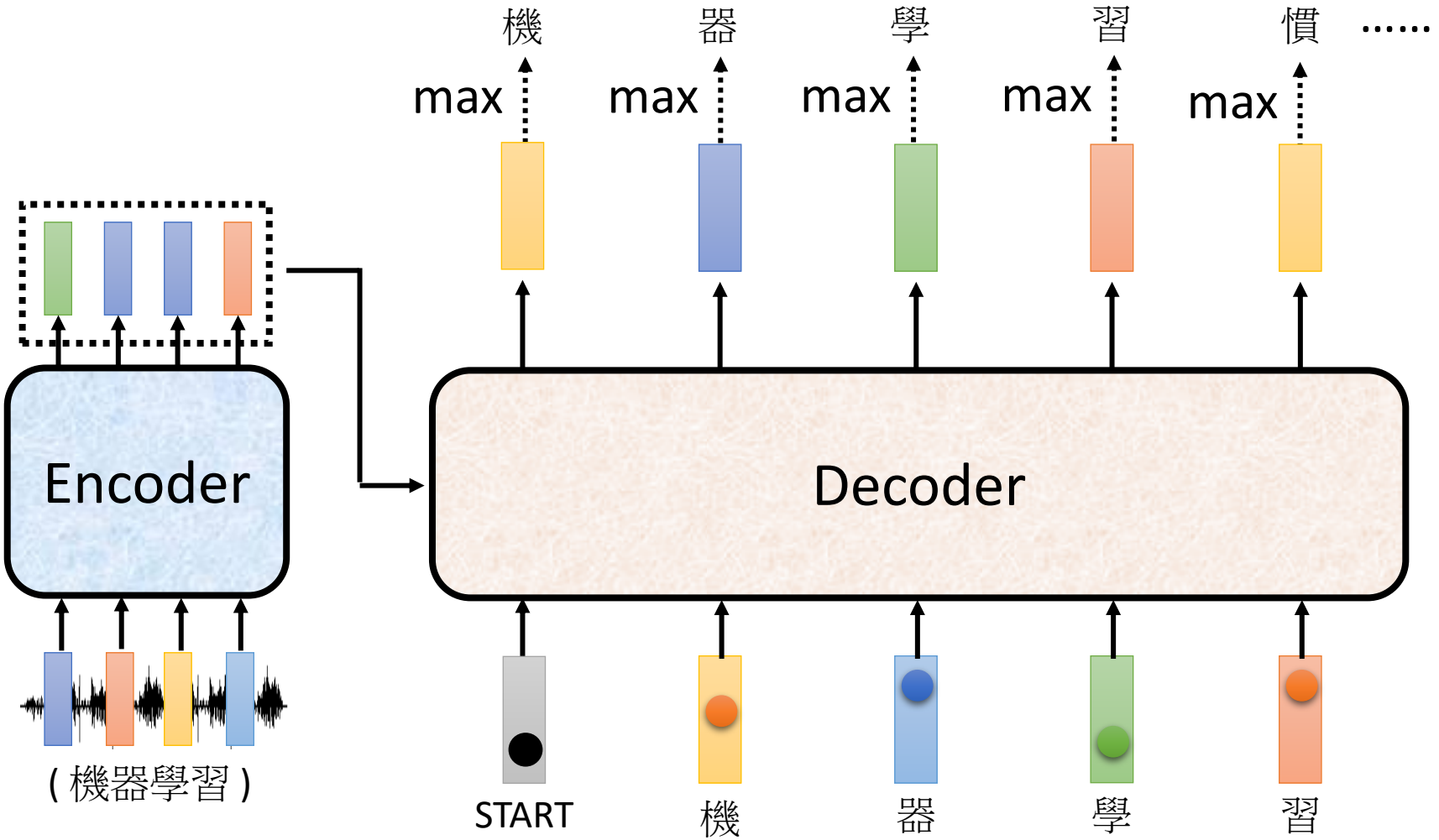
Autoregressive



Autoregressive

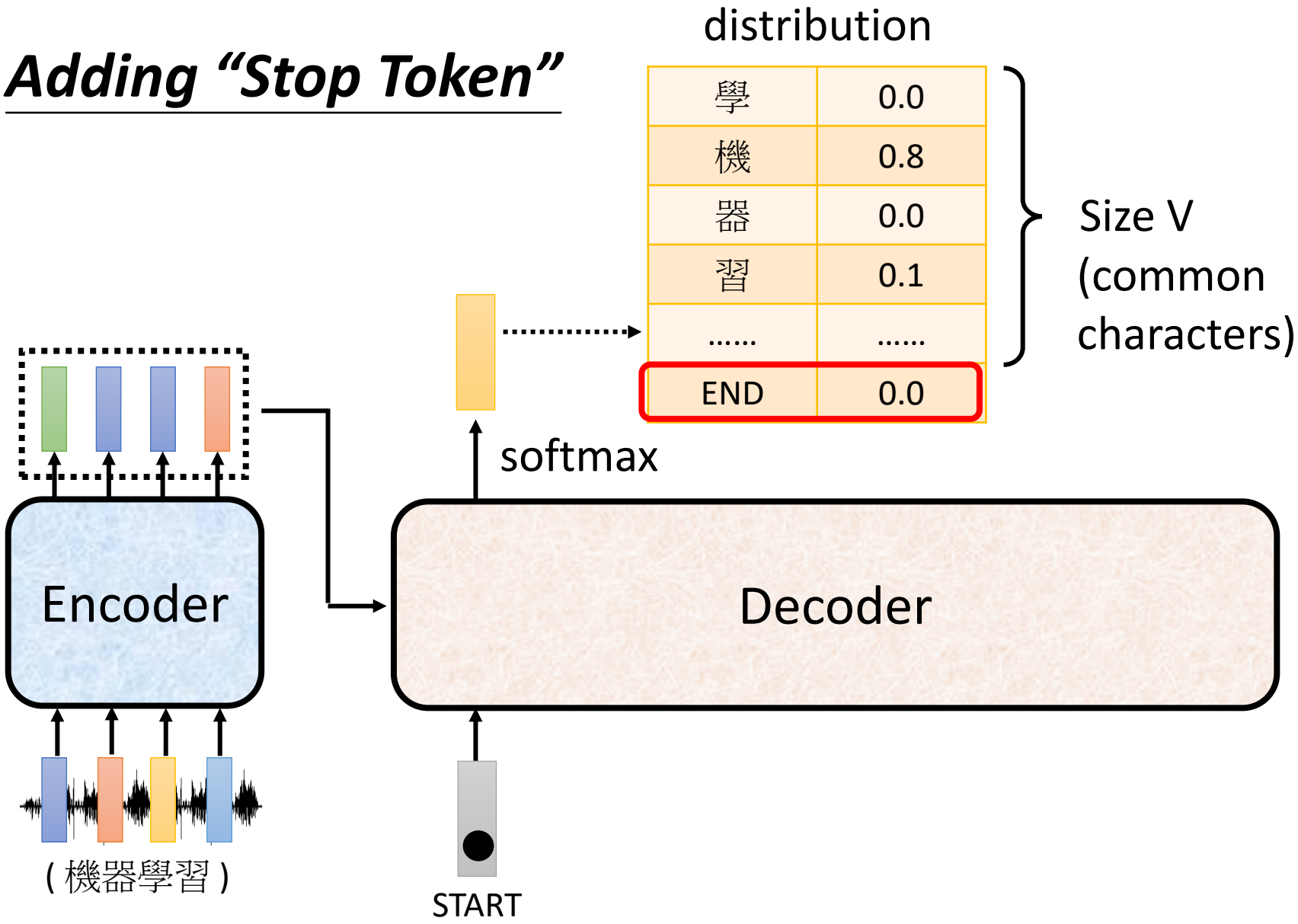
We do not know the correct output length.

Never stop!



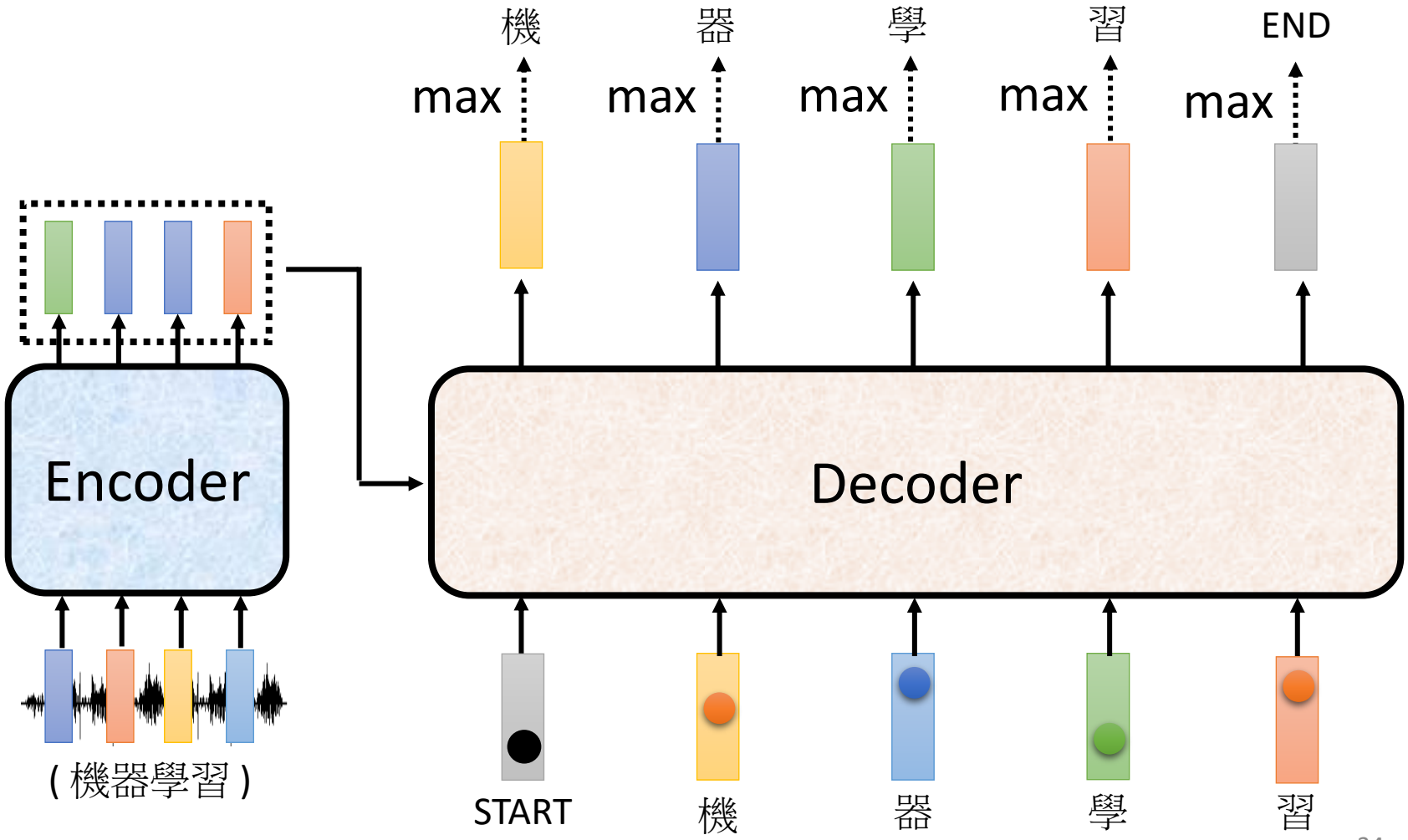
推文接龍 (Tweet Solitaire)

Adding "Stop Token"

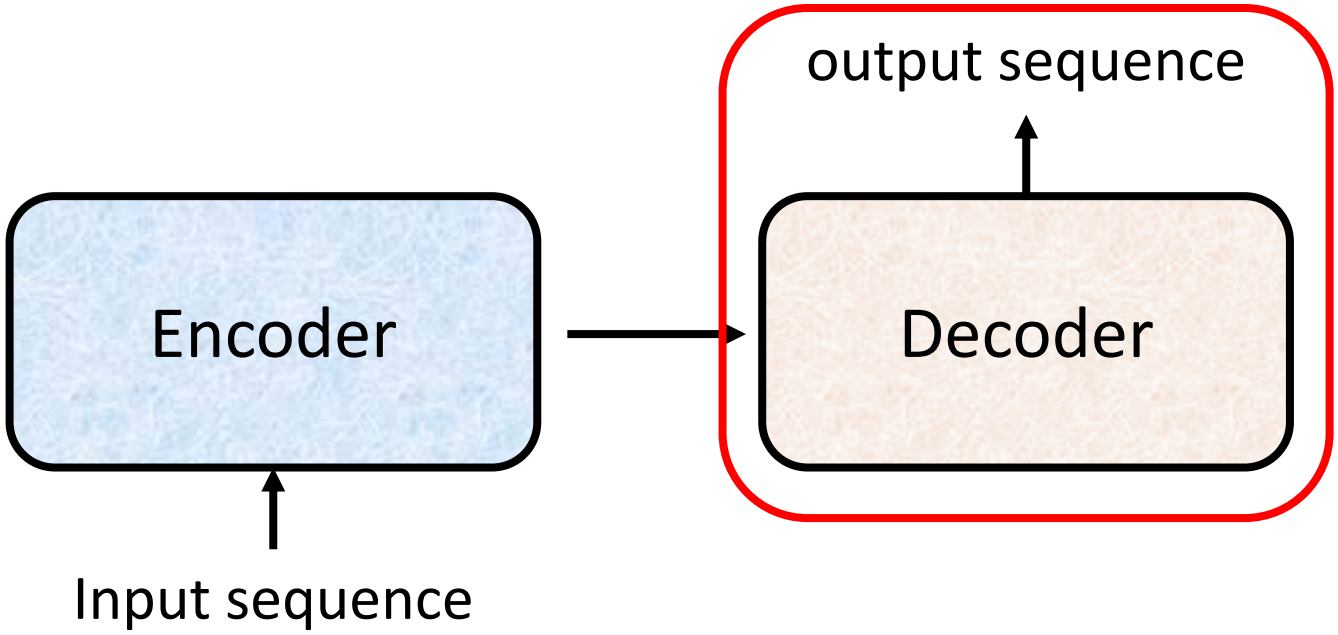


Autoregressive

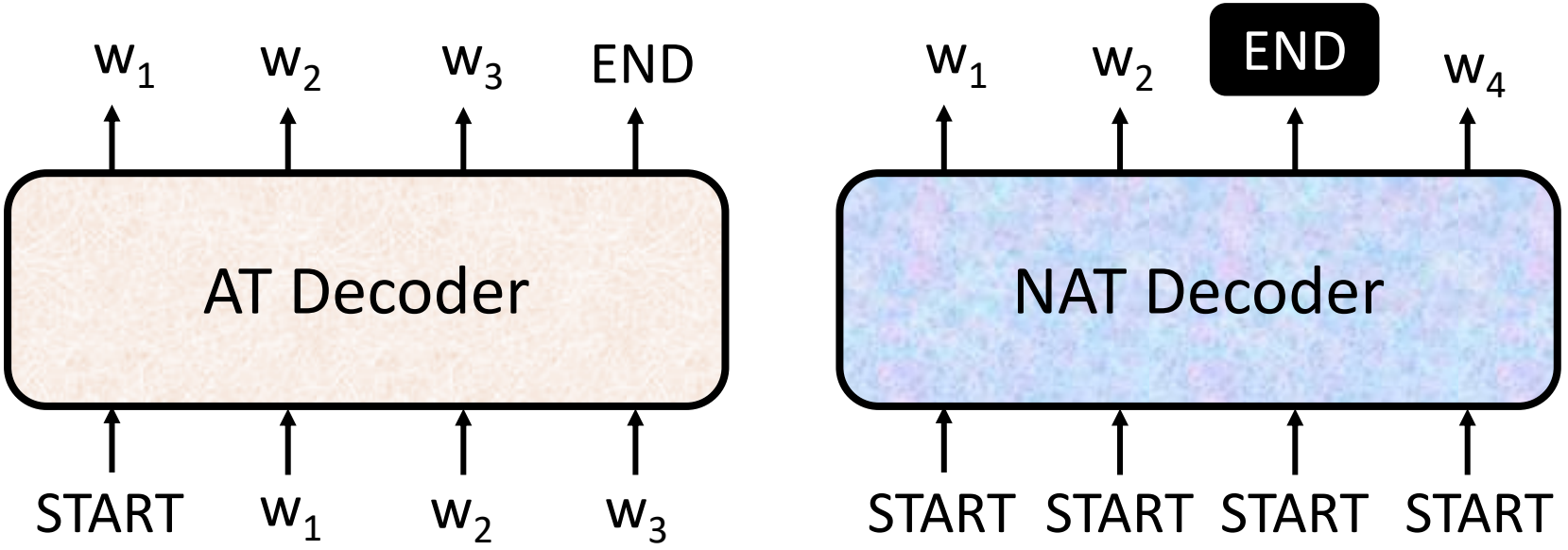
Stop at here!



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

NAT更快 Controllable Output Length

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

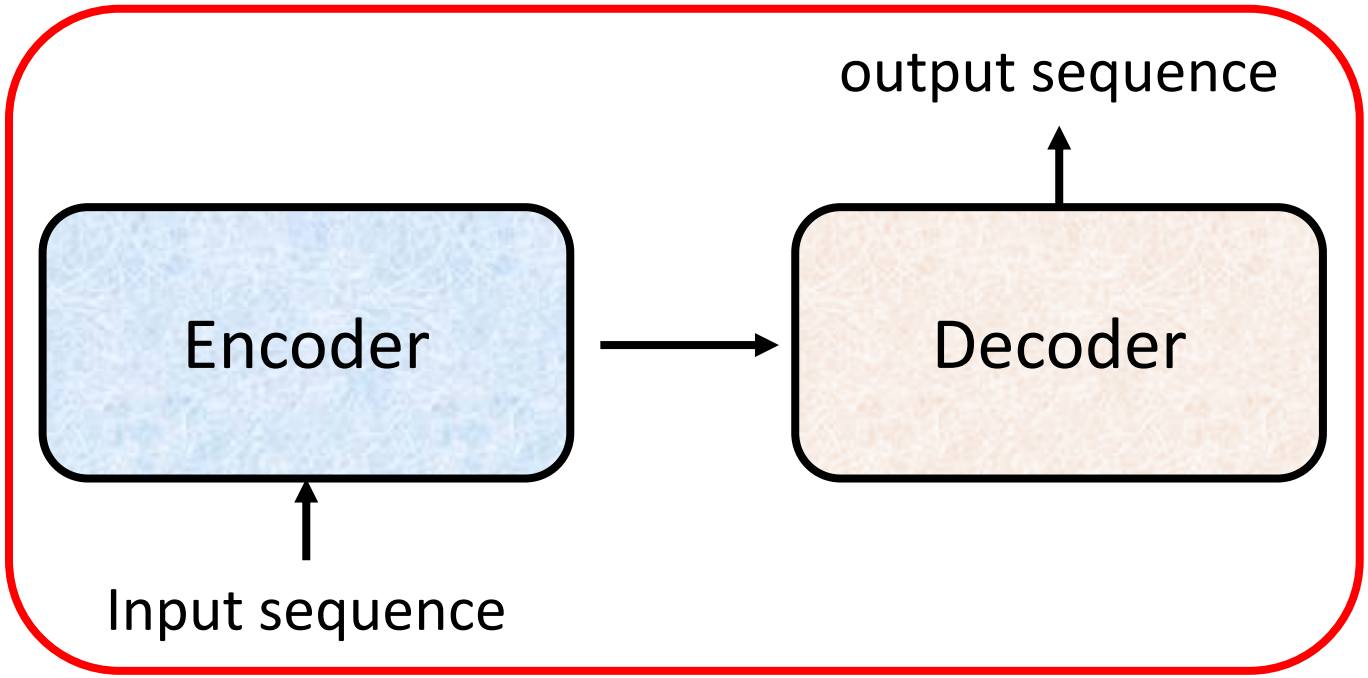
To learn more ……

助教课：NAT

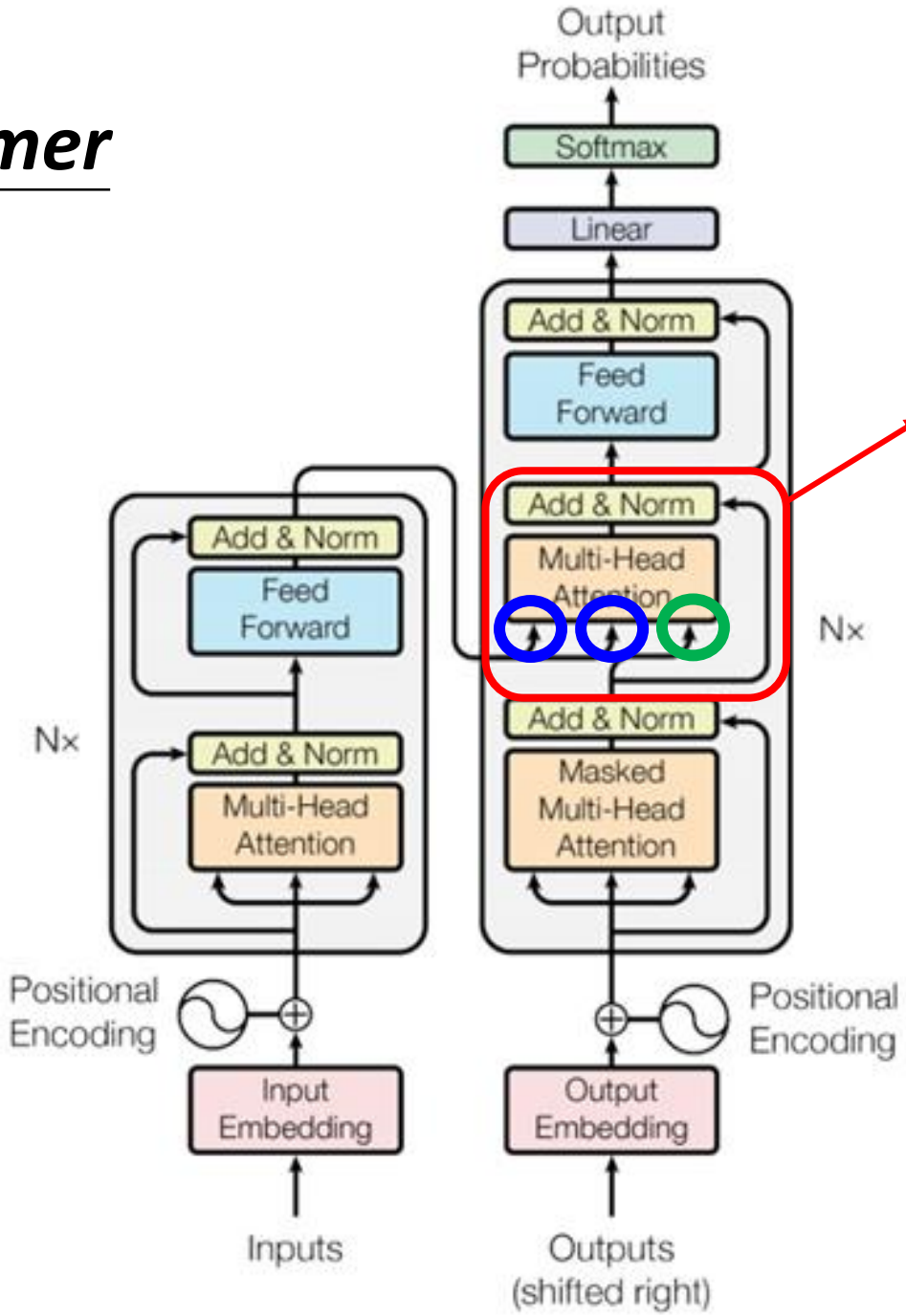
https://youtu.be/jvyKmU4OM3c

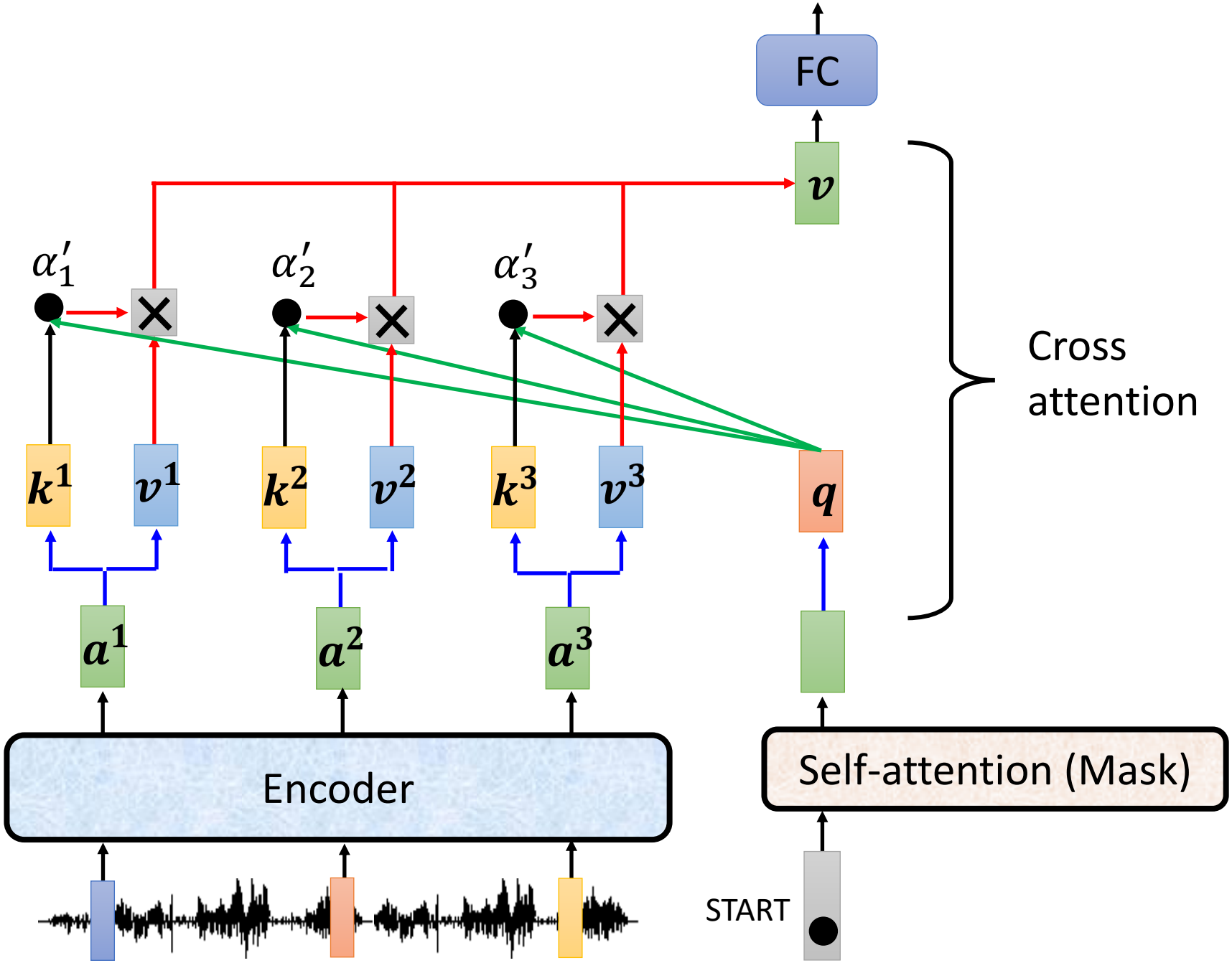
(in Mandarin)

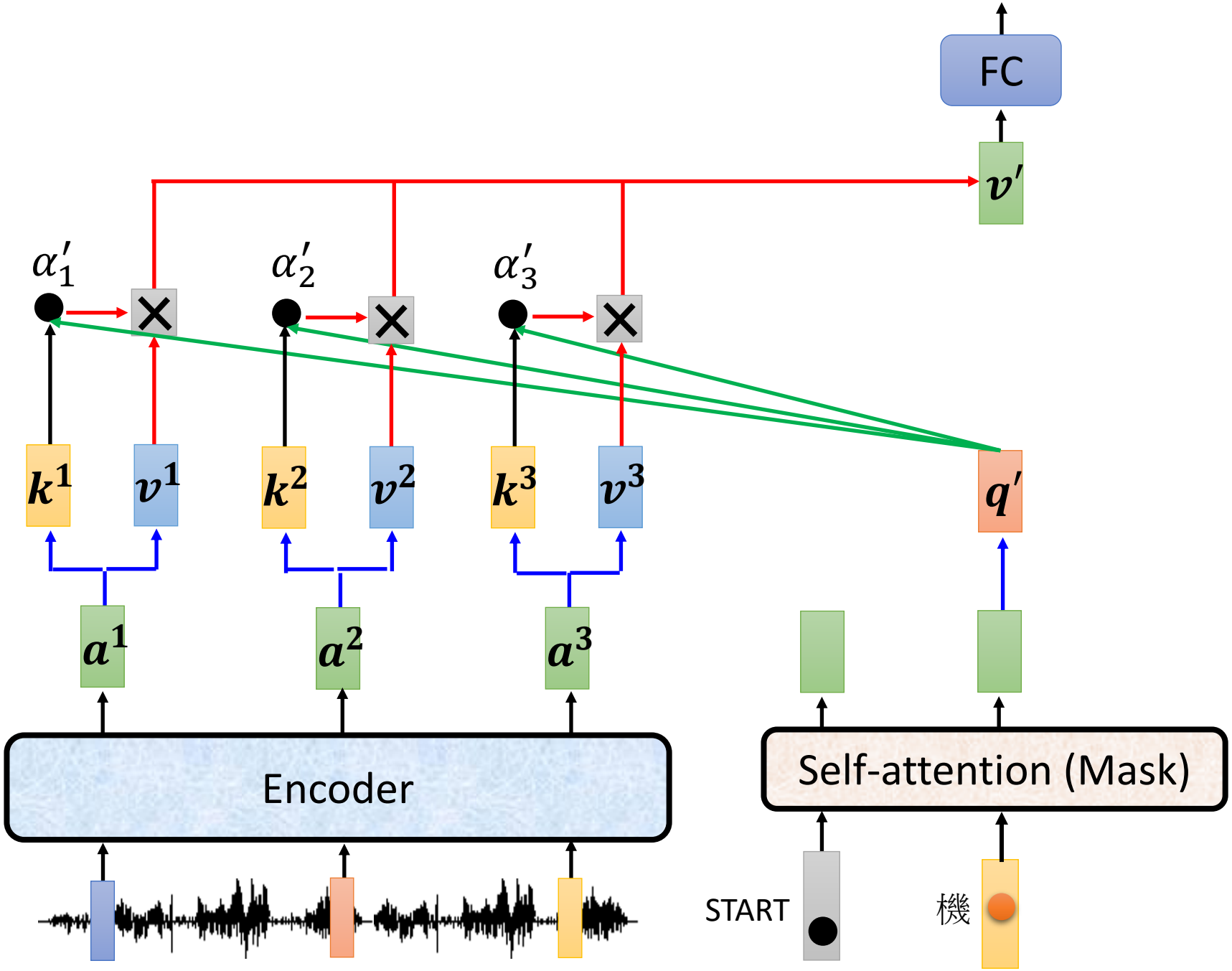
Encoder-Decoder



Transformer



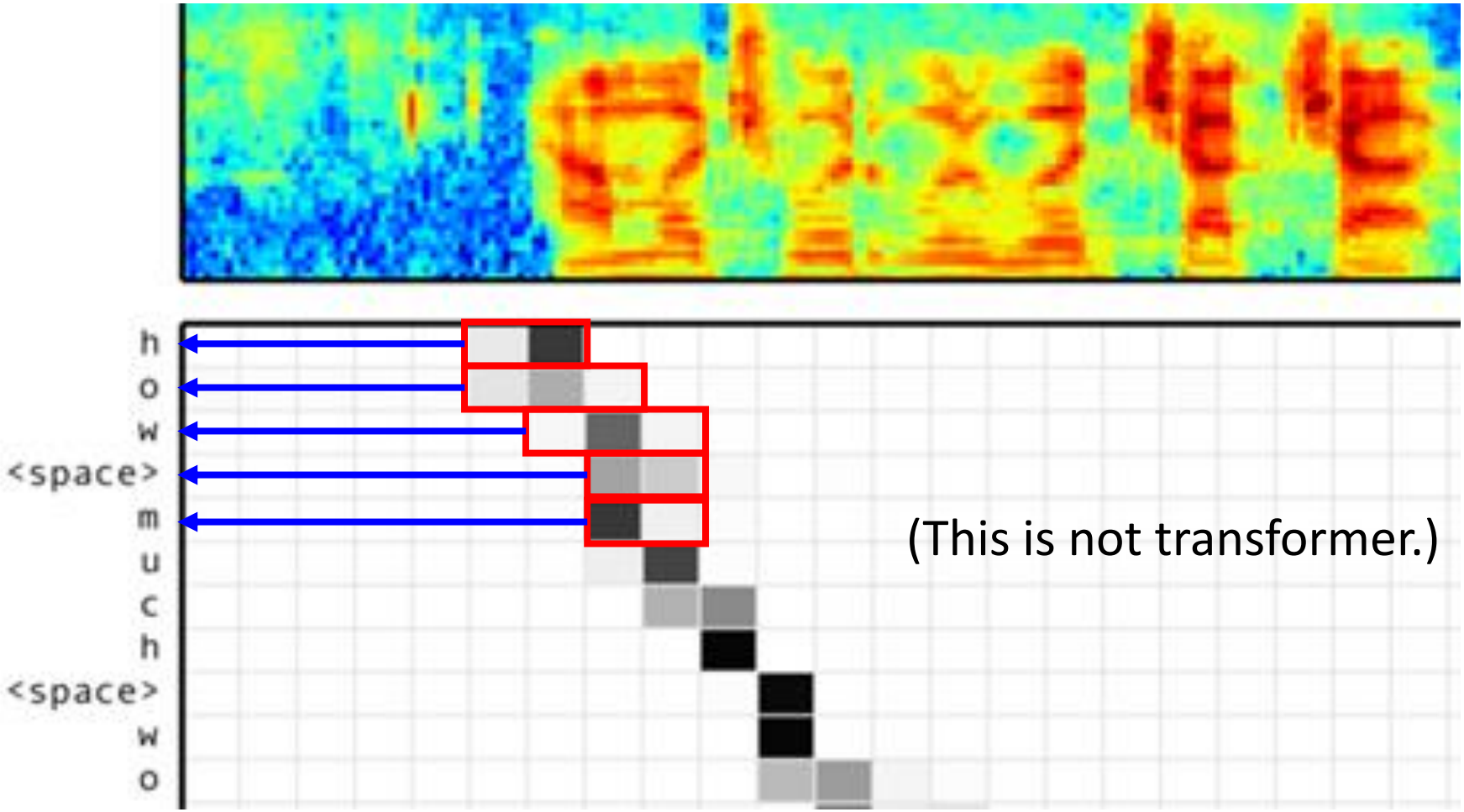

FC
v
α'_1
α'_2
α'_3
Cross attention
k^1
v^1
k^2
v^2
k^3
v^3
q
a^1
a^2
a^3
Encoder
Self-attention (Mask)
START


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

Cross Attention

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

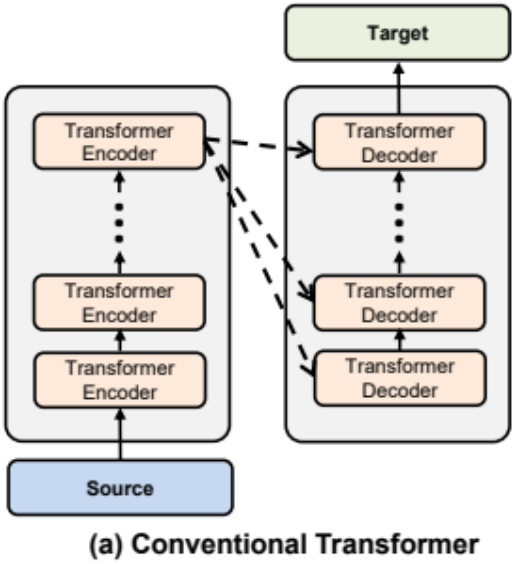
https://ieeexplore.ieee.org/document/7472621



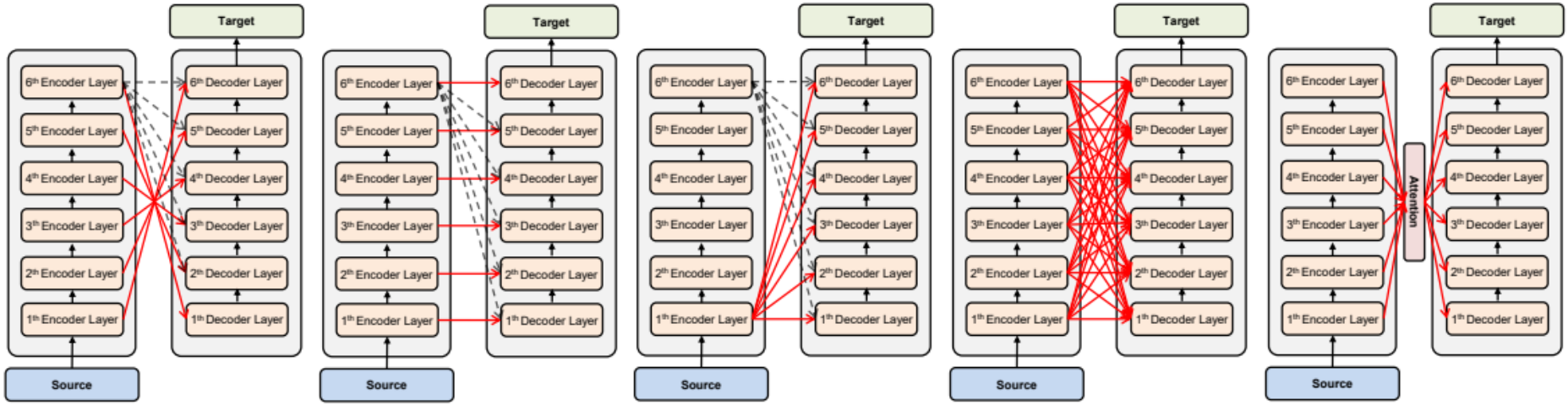
(This is not transformer.)

Cross Attention

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

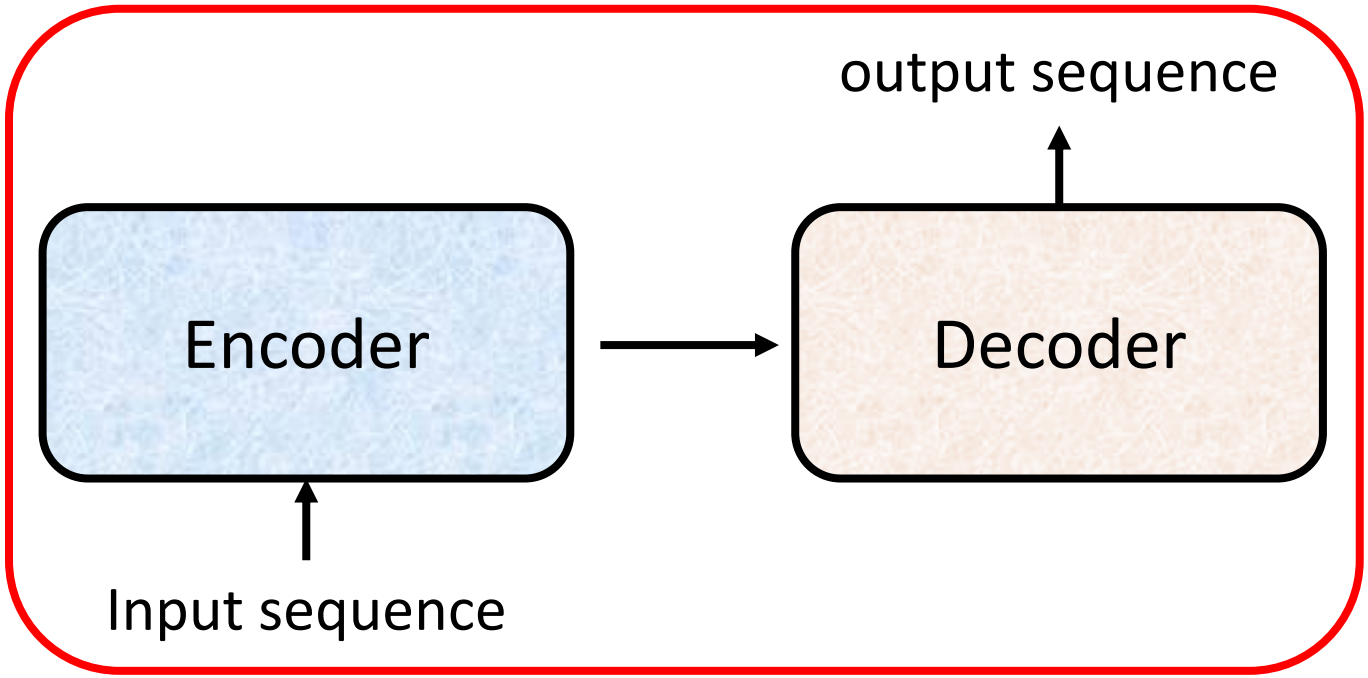


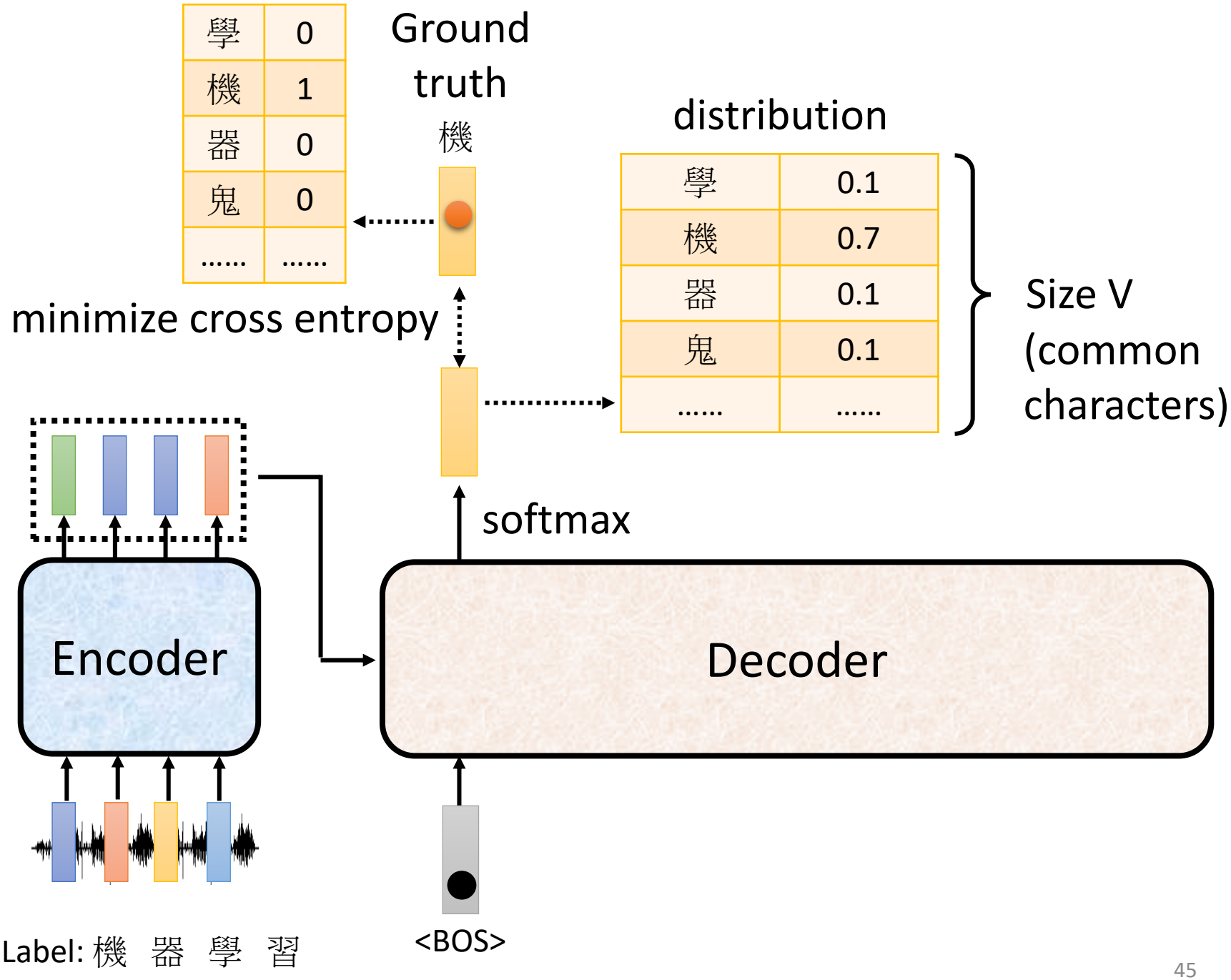
(a) Conventional Transformer



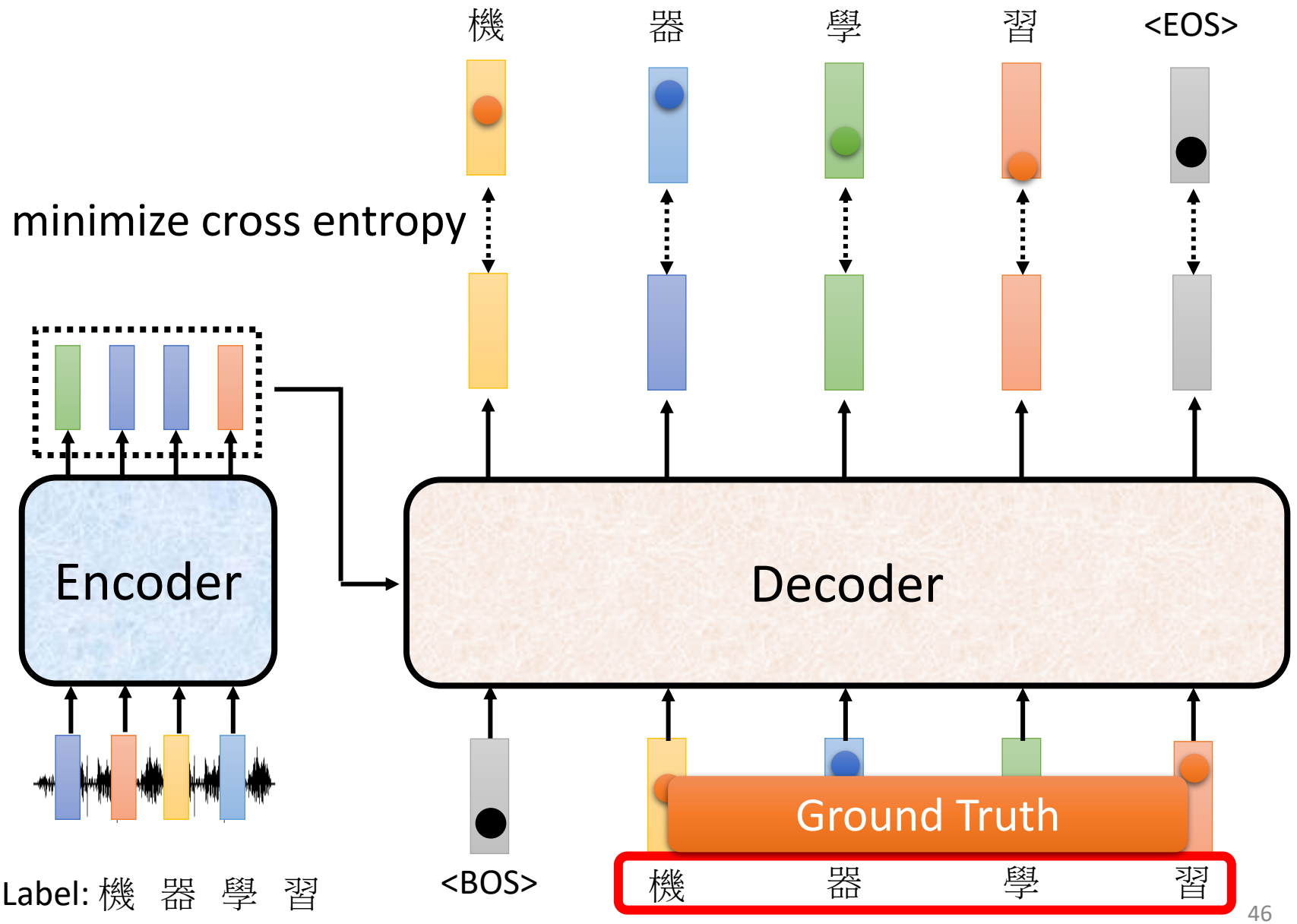
(a) Granularity Consistent Attention (b) Granularity Parallel Attention (c) Fine-Grained Attention (d) Full Matching Attention (e) Adaptive Matching Attention

Training

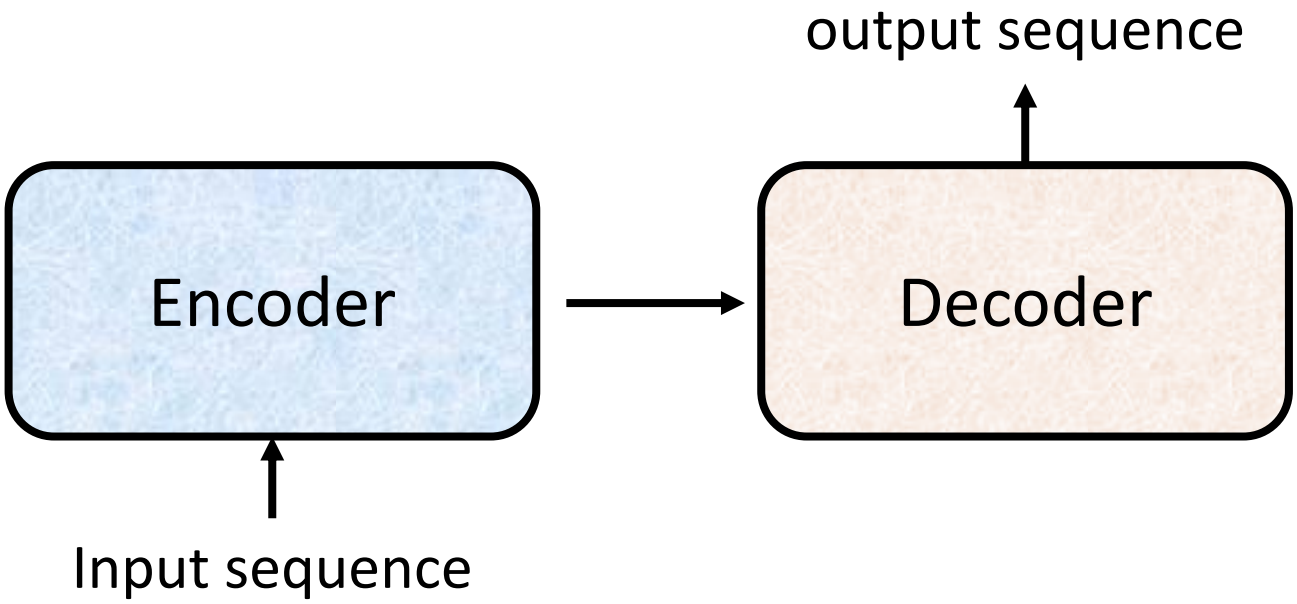



學 0
機 1
器 0
鬼 0
....
Ground truth
機
distribution
學 0.1
機 0.7
器 0.1
鬼 0.1
....
Size V (common characters)
minimize cross entropy
softmax
Encoder
Decoder
<BOS>
Label: 機 器 學 習

Teacher Forcing: using the ground truth as input.

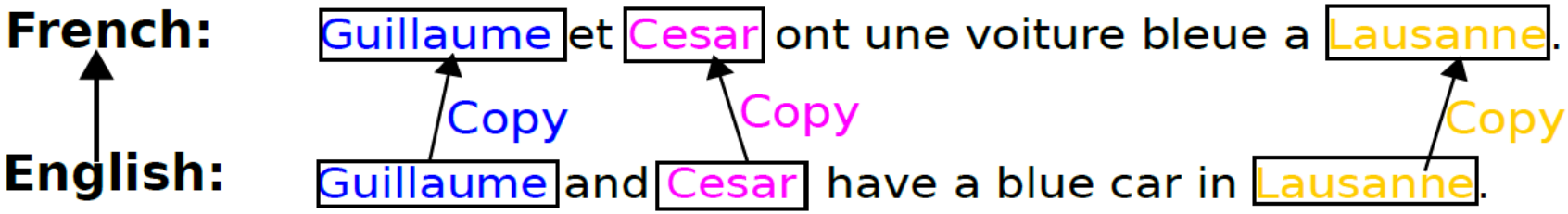


Tips



Copy Mechanism

Machine Translation



Chat-bot

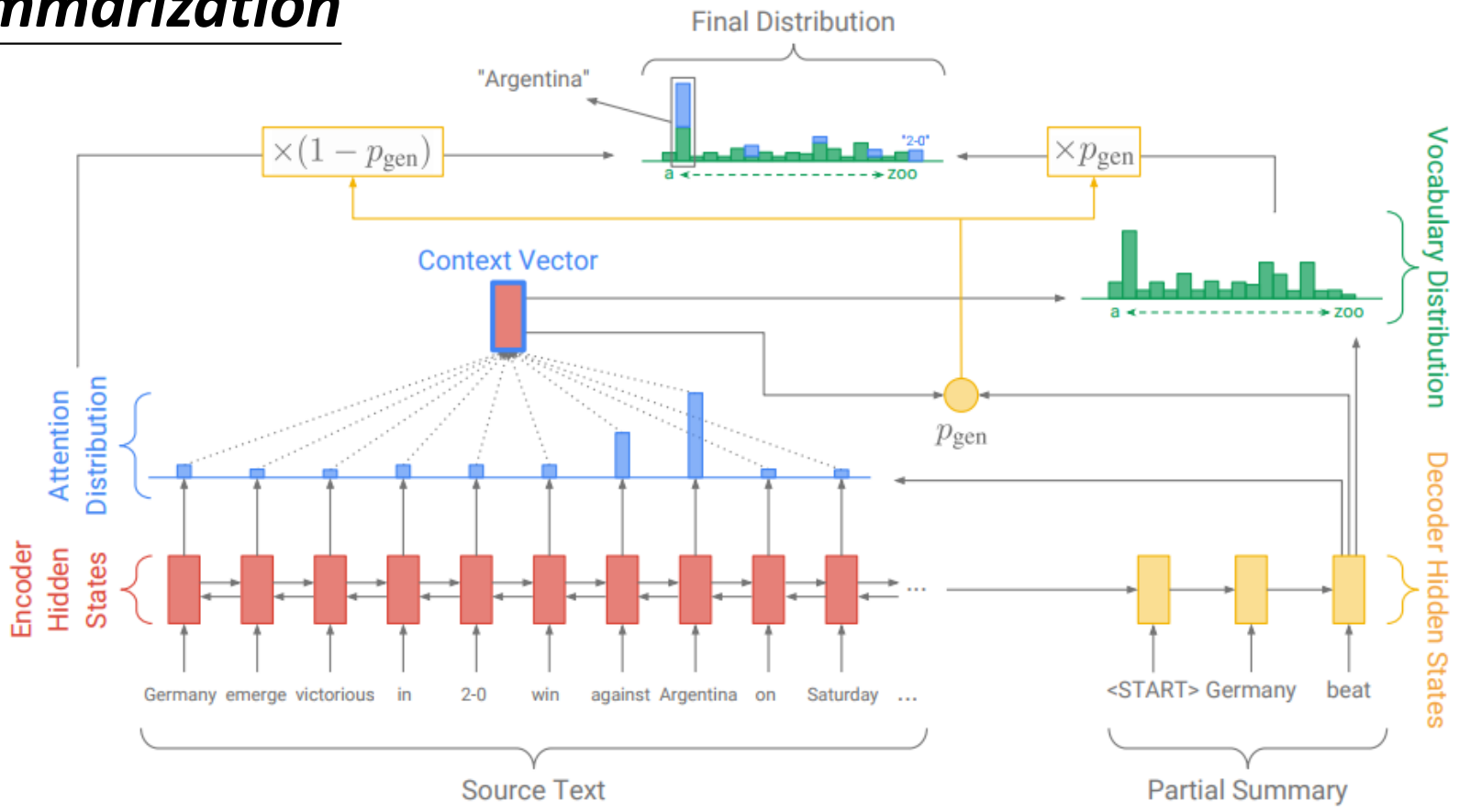
User: X寶你好，我是庫洛洛

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

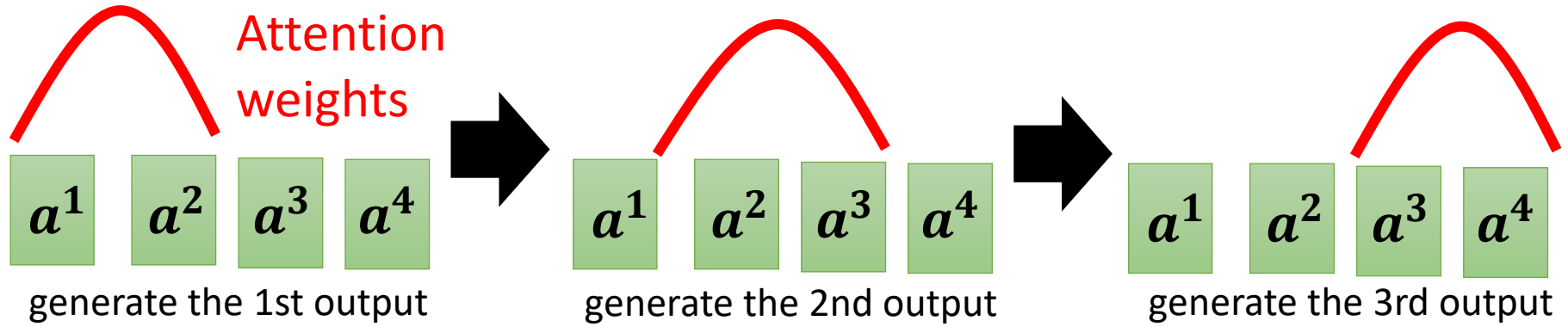
Guided Attention

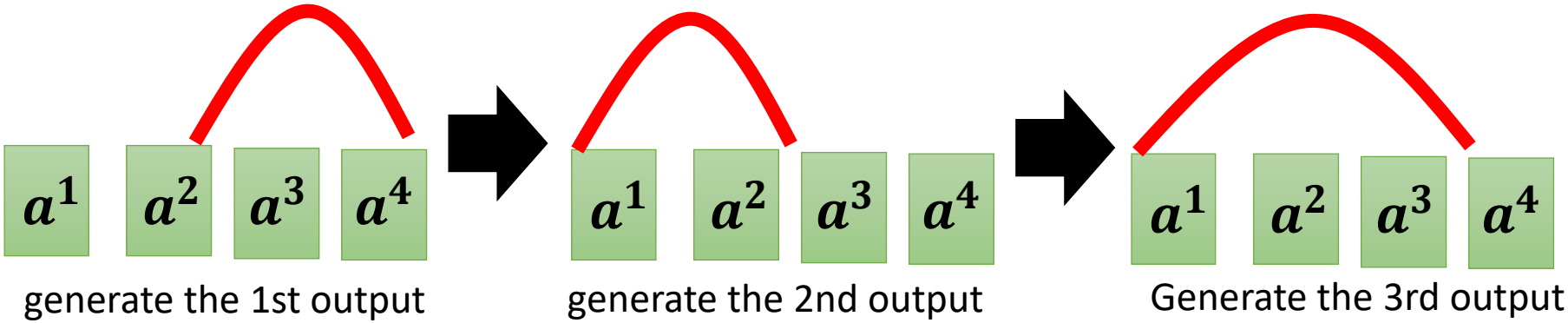
- 高雄發大財我現在要出征
- 發財發財發財發財
- 發財發財發財
- 發財發財
- 發財　　(Missing an input character!)

Guided Attention

Monotonic Attention
Location-aware attention

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





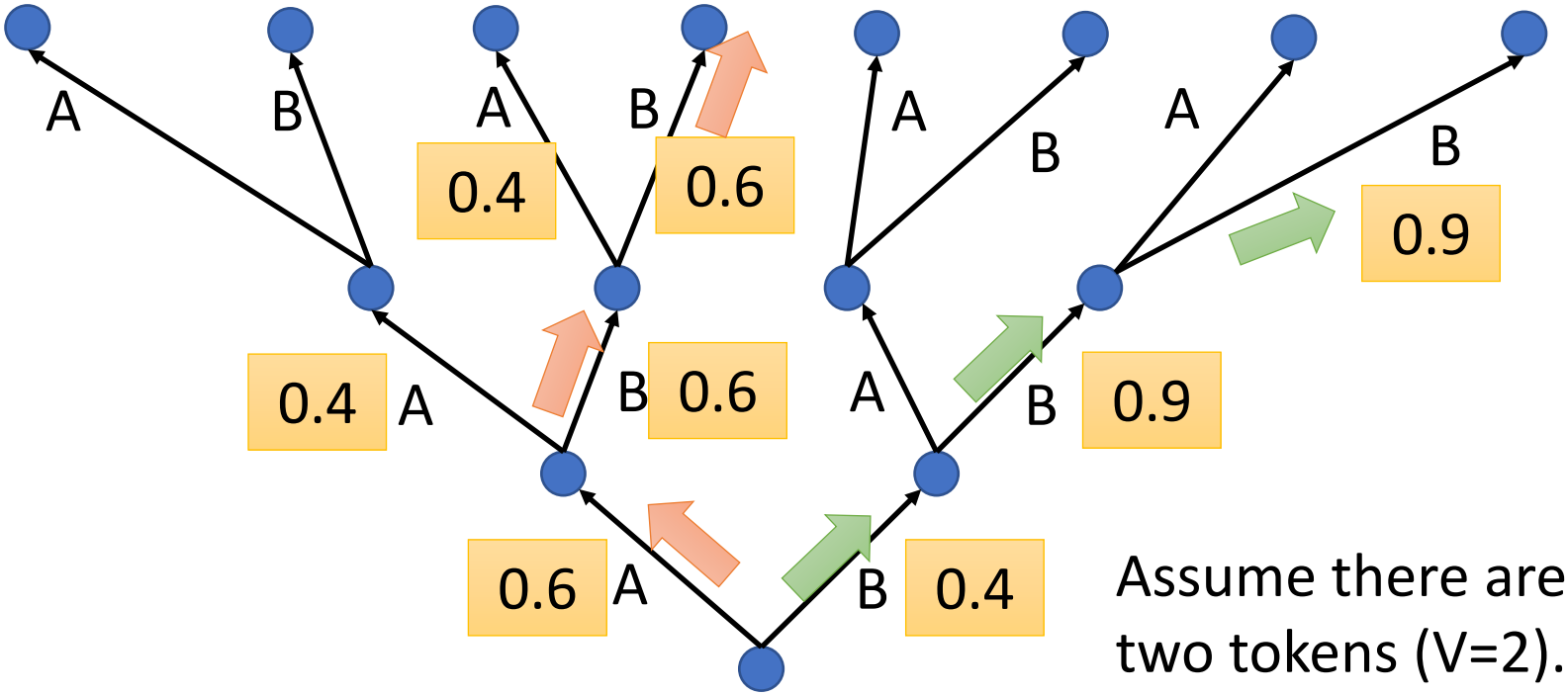
Something wrong!

Beam Search

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

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.

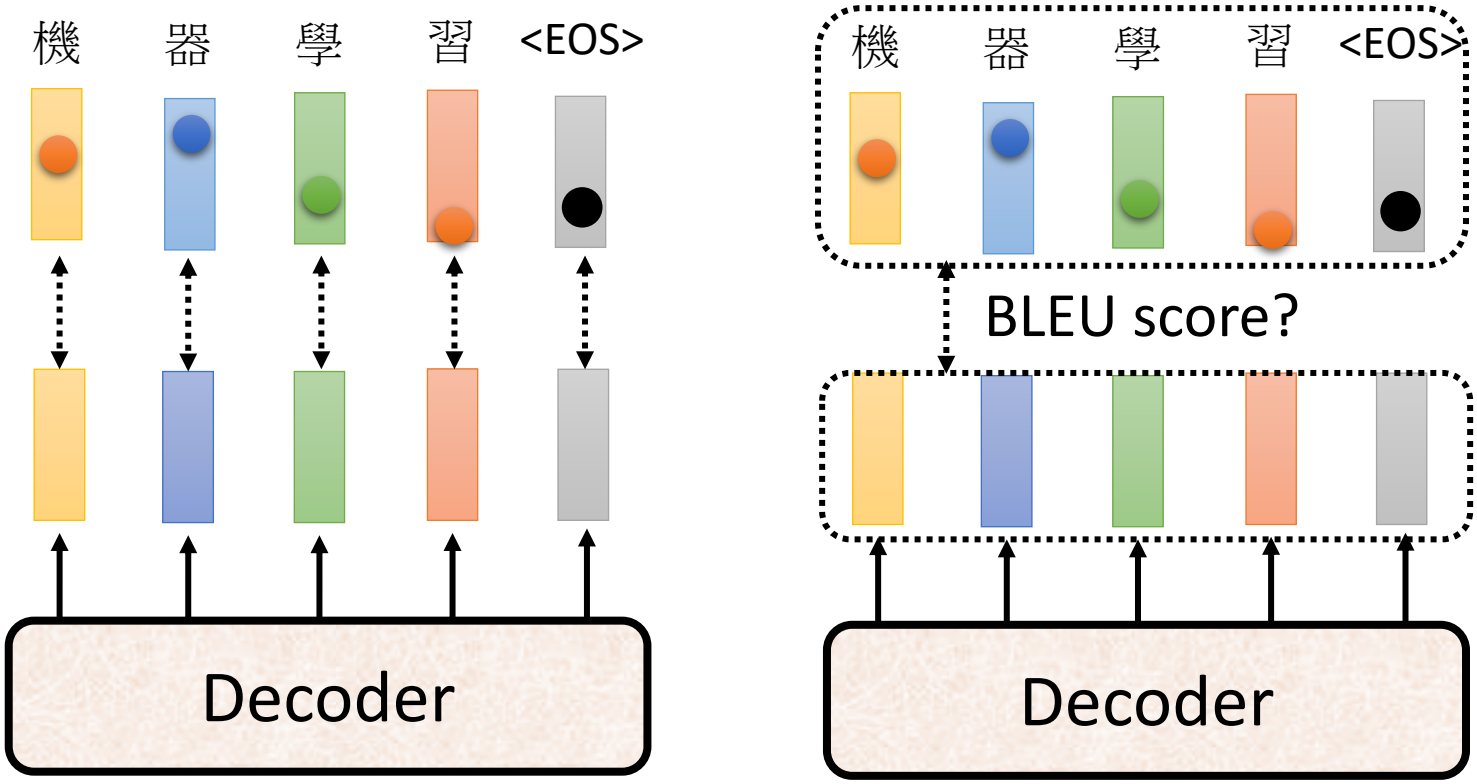
Beam Search, *b*=32:
"The study, published in the Proceedings of the National Academy of Sciences of the United States of America (PNAS), was conducted by researchers from the Universidad Nacional Autónoma de México (UNAM) and the Universidad Nacional Autónoma de México (UNAM/Universidad Nacional Autónoma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

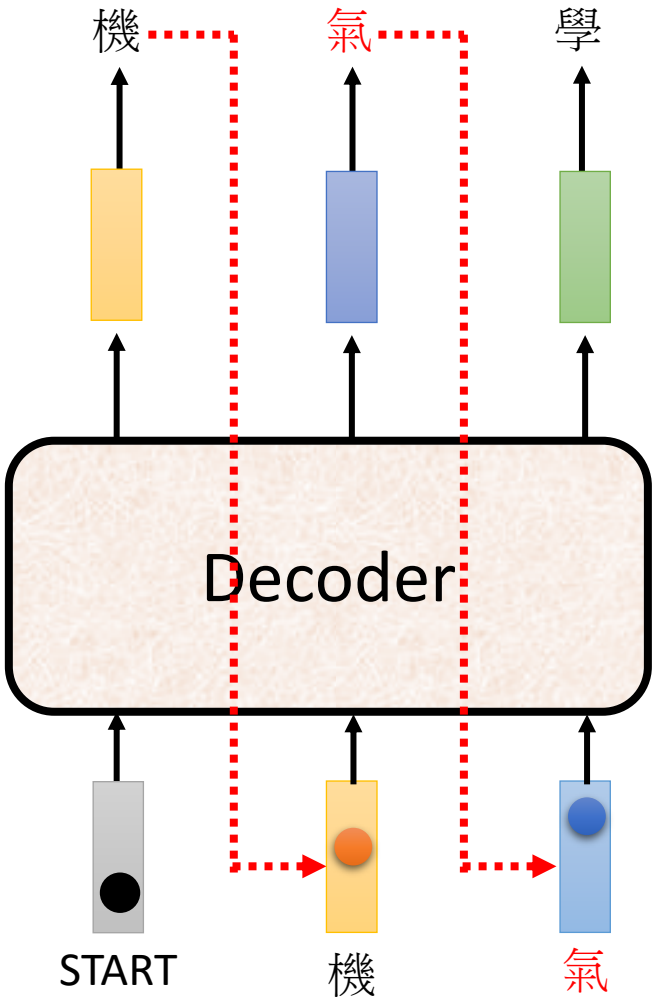


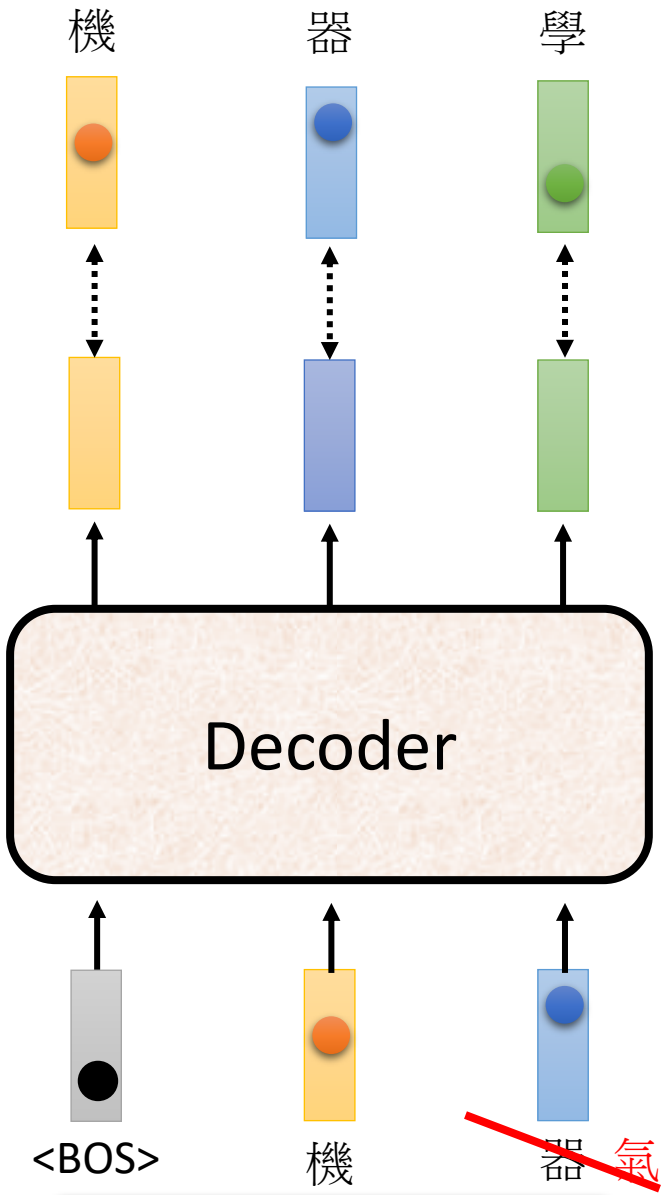
How to do the optimization?

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

There is a mismatch! ☹

exposure bias





Ground Truth

Scheduled Sampling

- Original Scheduled Sampling

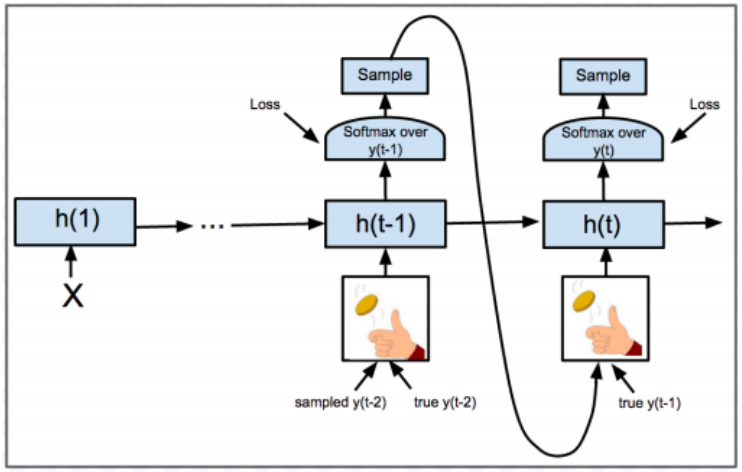
  https://arxiv.org/abs/1506.03099

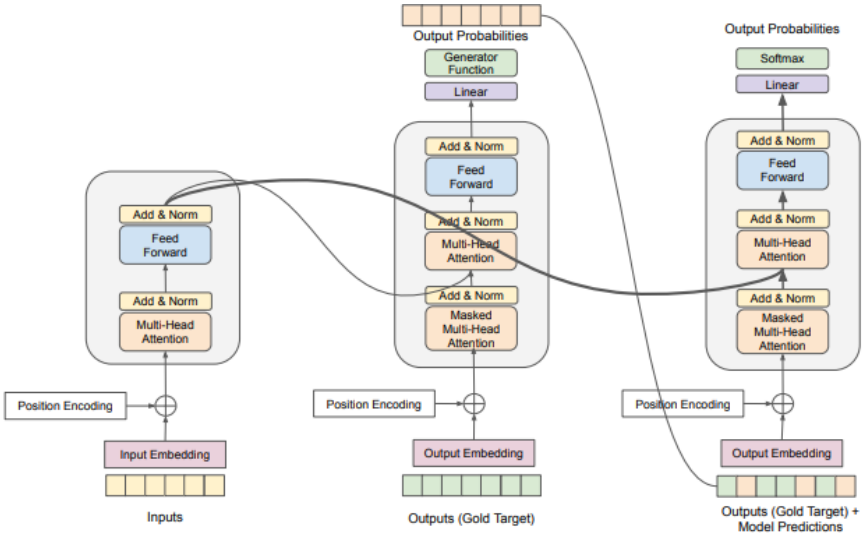
- Scheduled Sampling for Transformer

  https://arxiv.org/abs/1906.07651

- Parallel Scheduled Sampling

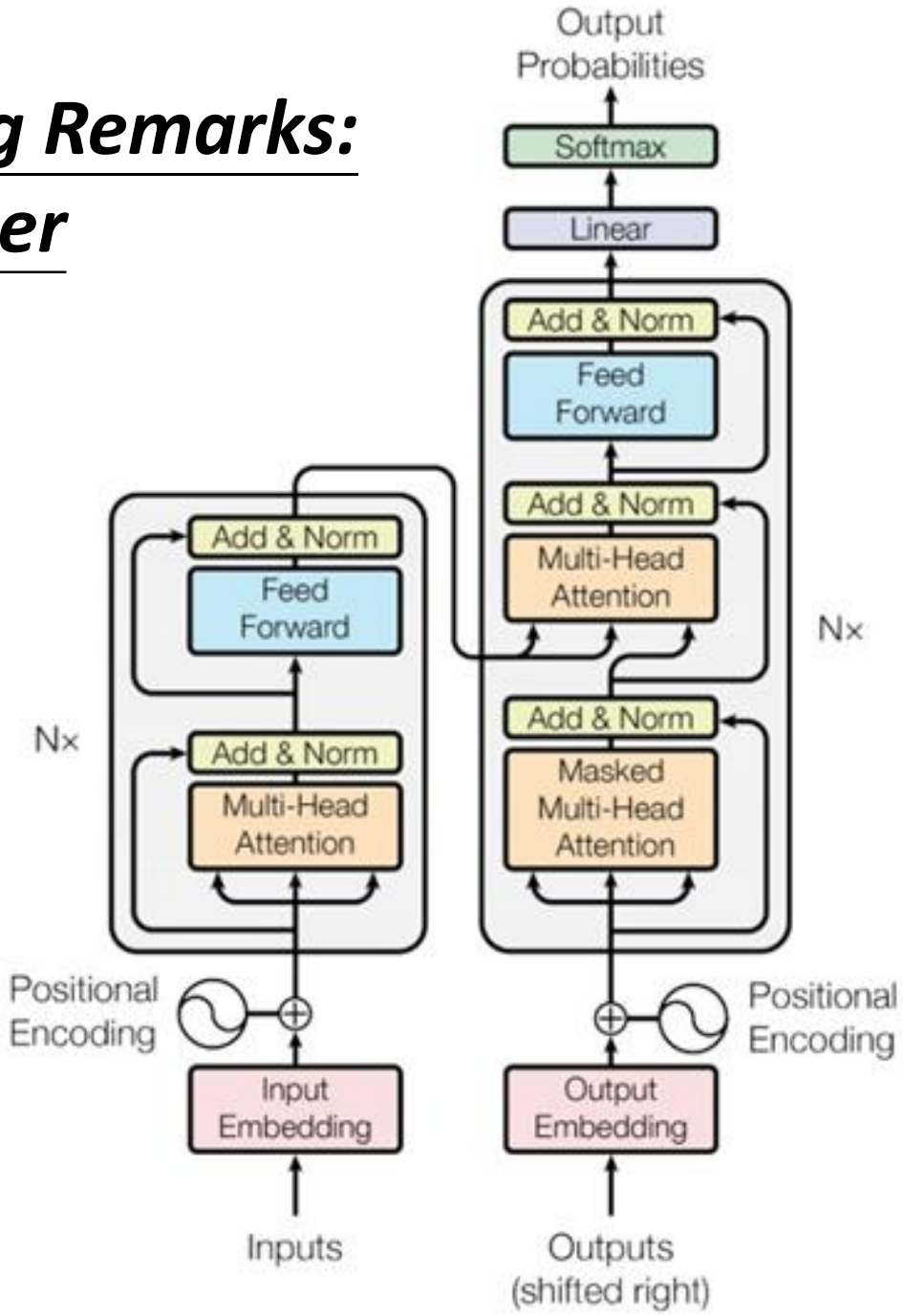
  https://arxiv.org/abs/1906.04331





Schedule Sampling

Concluding Remarks: Transformer




Q&A