Transformer

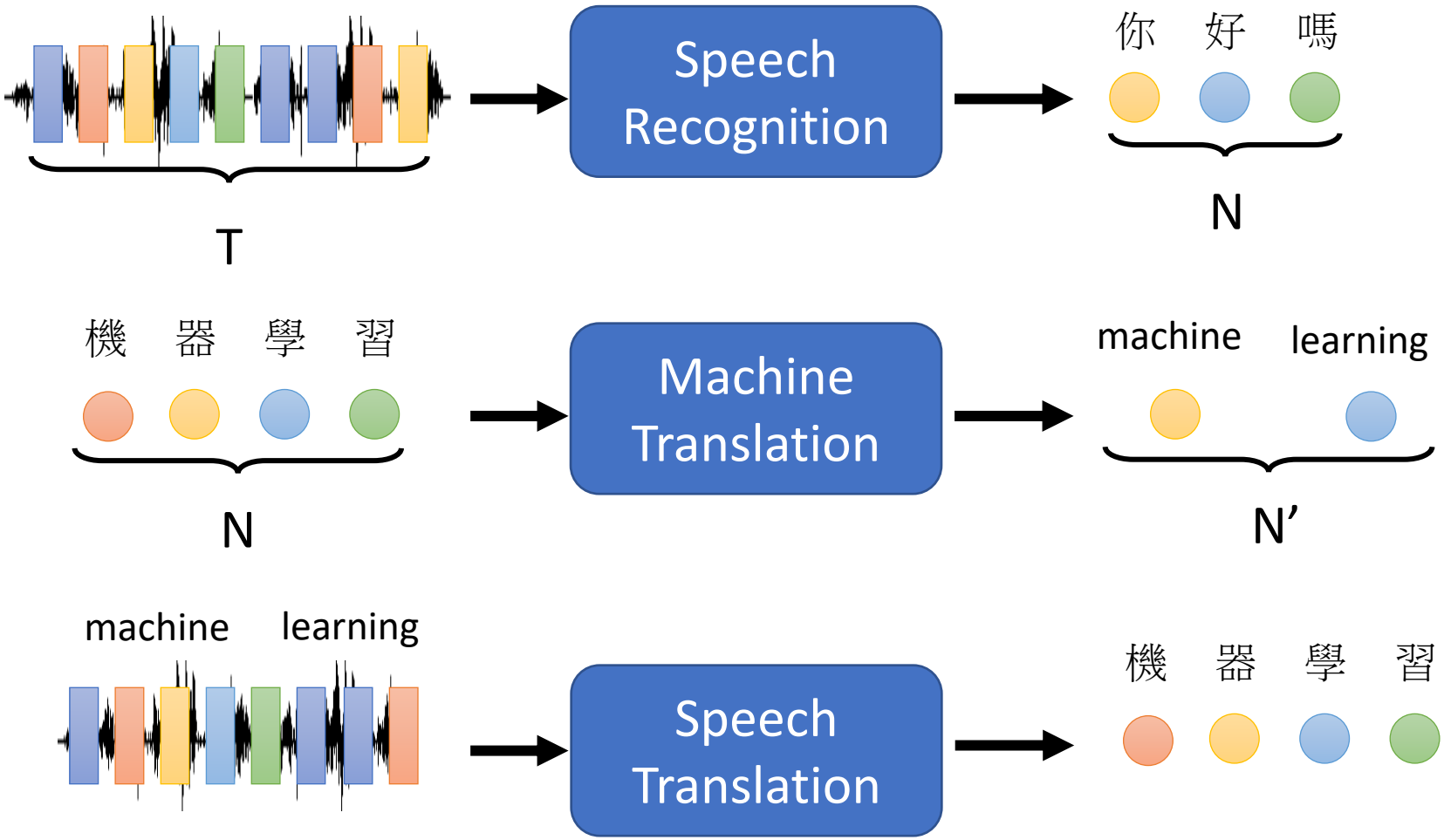
李宏毅

Hung-yi Lee

Sequence-to-sequence (Seq2seq)

Input a sequence, output a sequence

The output length is determined by model.



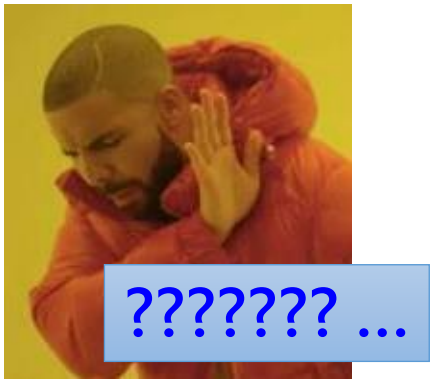
Language without text

Hokkien (閩南語、台語)

(Speech of Hokkien)

Speech Recognition

“母湯”



(Speech of Hokkien)

Speech Translation

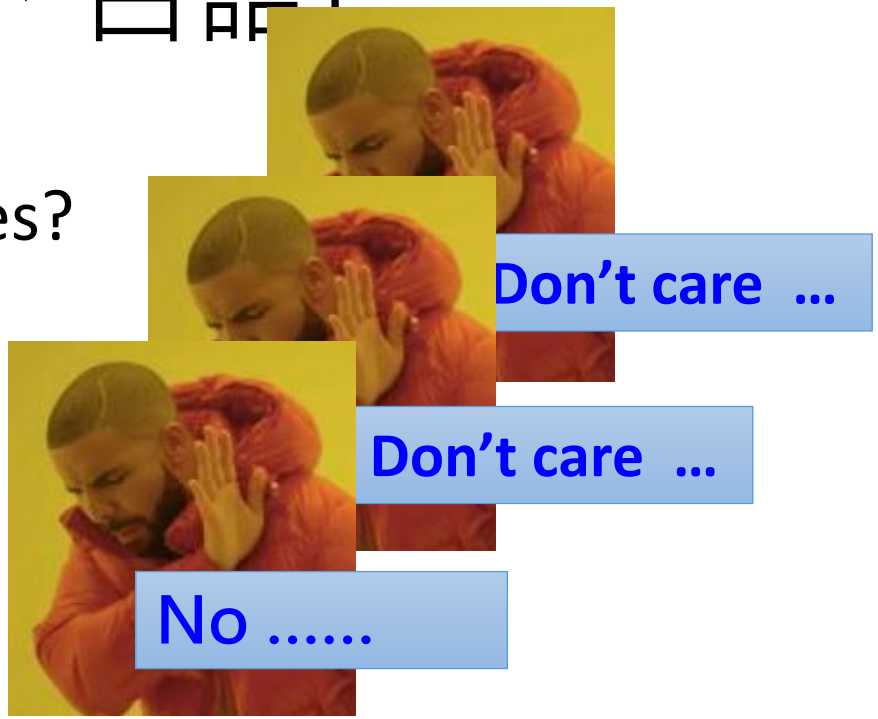
“不行”

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

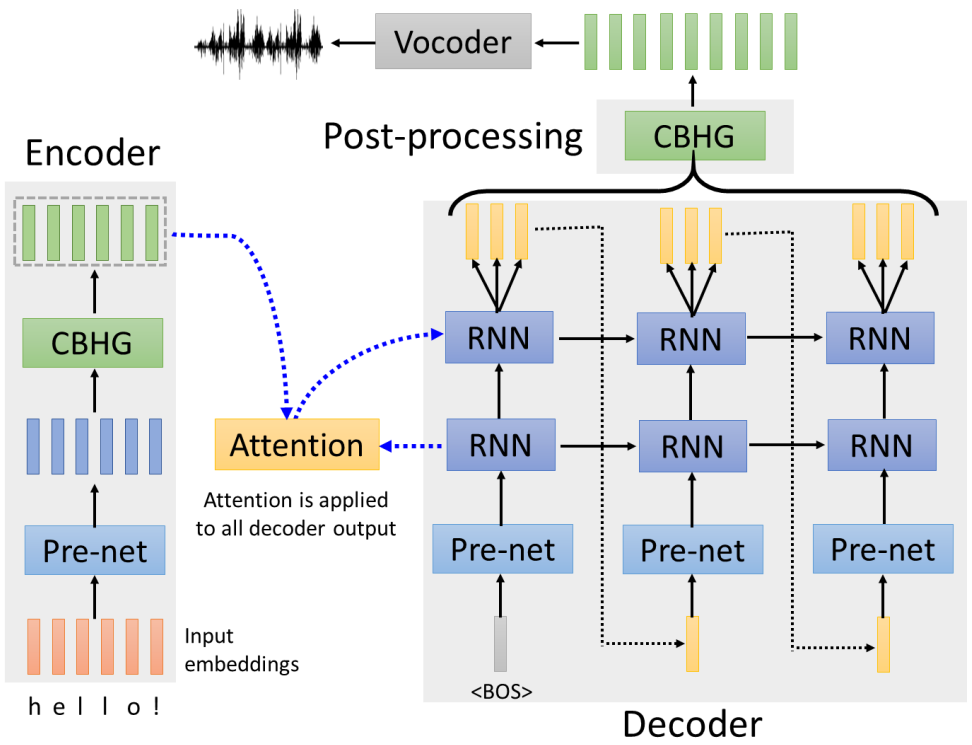
我 有 幫 廠 長 拜 託

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



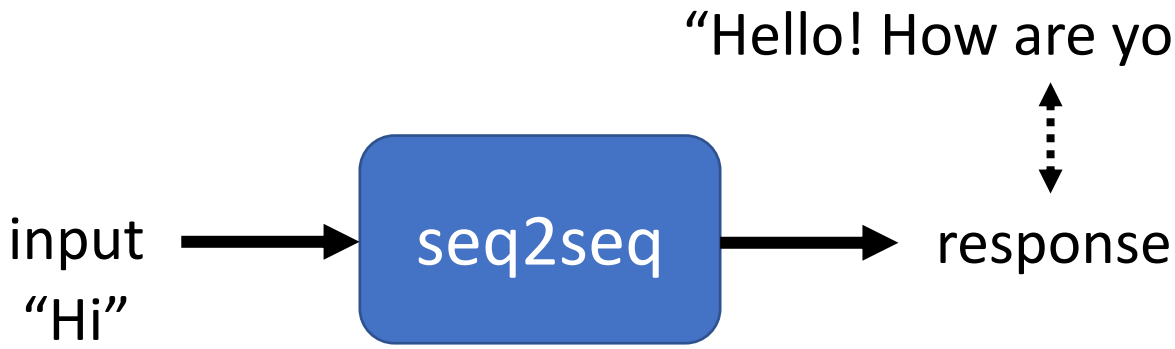
Taiwanese Speech Synthesis

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:

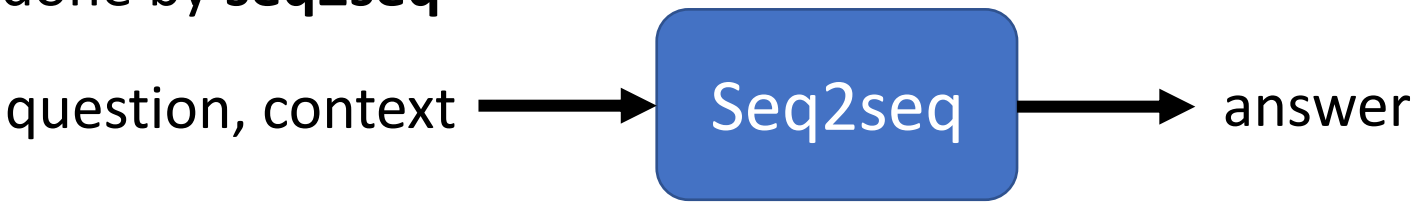
[PERSON 1:] Hi
[PERSON 2:] Hello ! How are you today ?
[PERSON 1:] I am good thank you , how are you.
[PERSON 2:] Great, thanks ! My children and I were just about to watch Game of Thrones.
[PERSON 1:] Nice ! How old are your children?
[PERSON 2:] I have four that range in age from 10 to 21. You?
[PERSON 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)

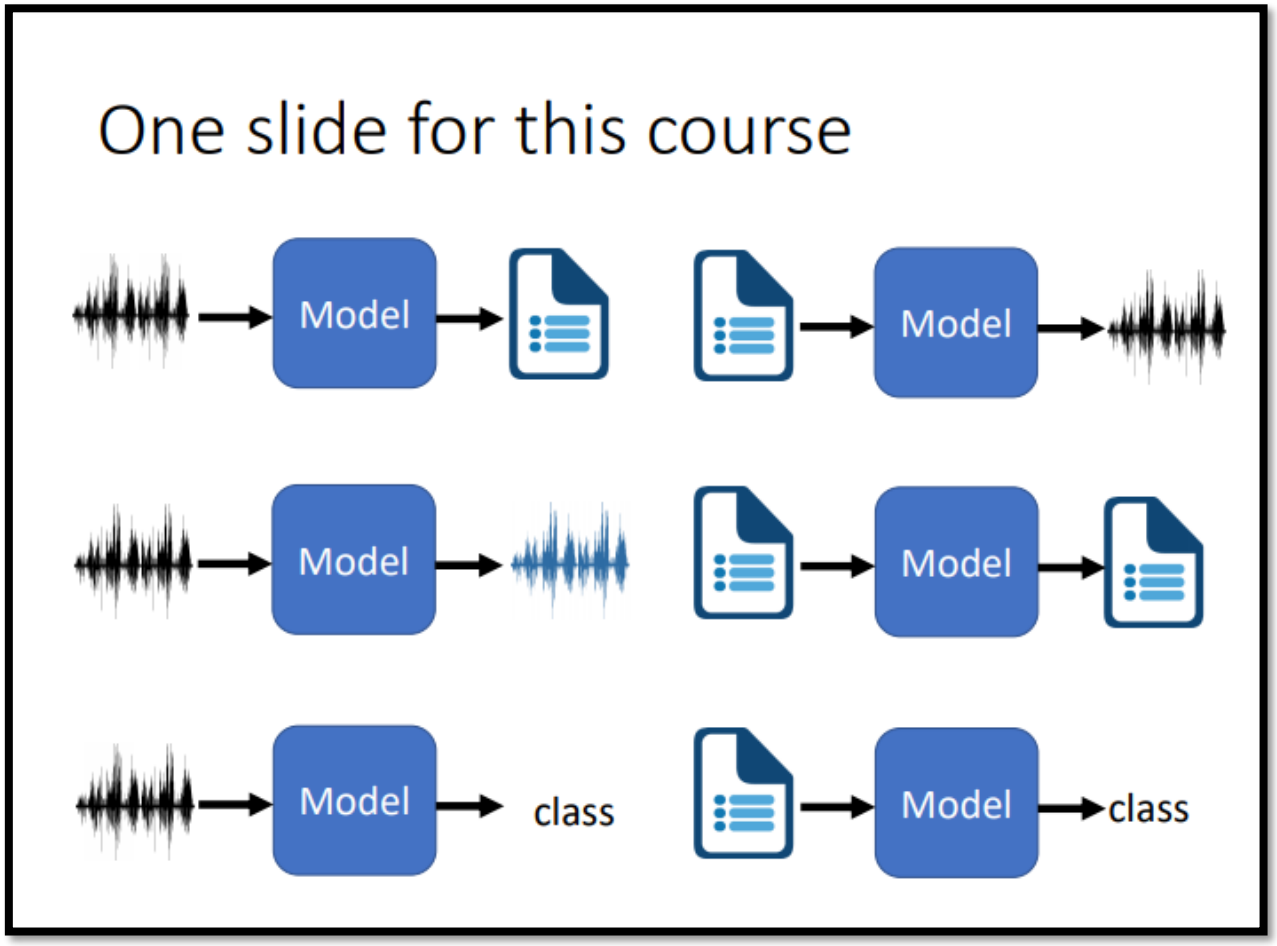
Question	Context	Answer
What is a major importance of Southern California in relation to California and the US?	...Southern California is a major economic center for the state of California and the US....	major economic center
What is the translation from English to German?	Most of the planet is ocean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

QA can be done by **seq2seq**



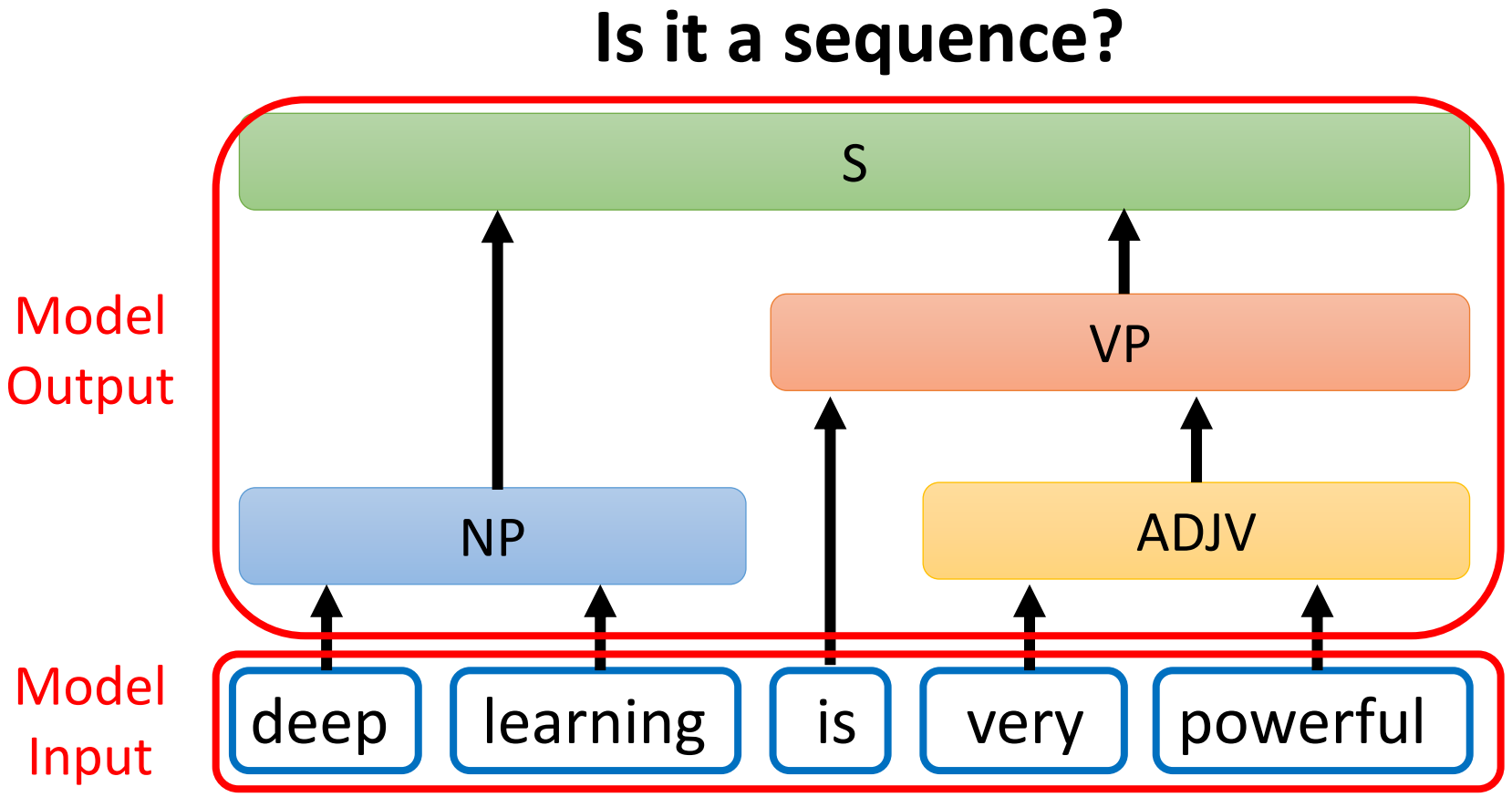
https://arxiv.org/abs/1806.08730
https://arxiv.org/abs/1909.03329

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

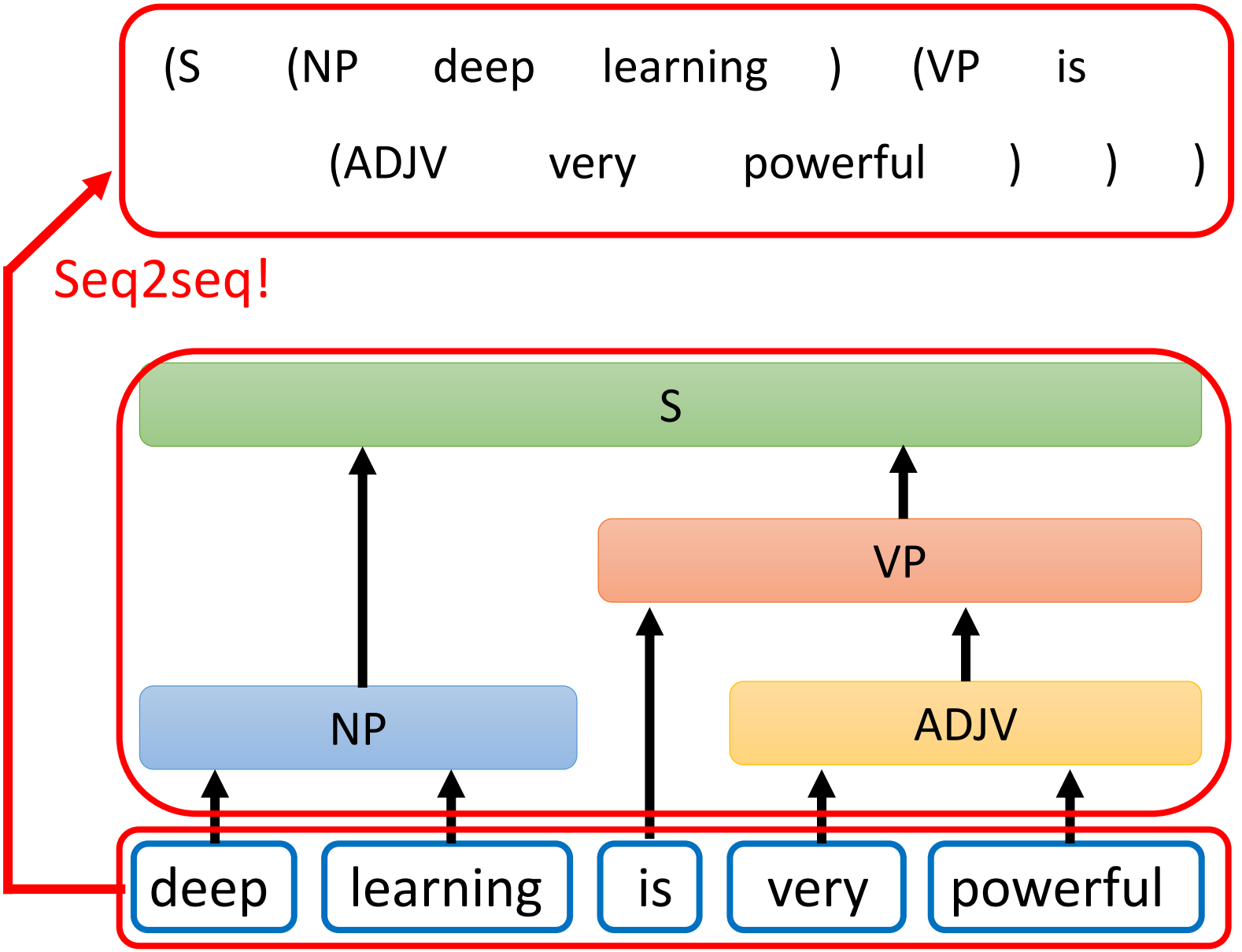


Source webpage: https://speech.ee.ntu.edu.tw/~hylee/dlhlp/2020-spring.html

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

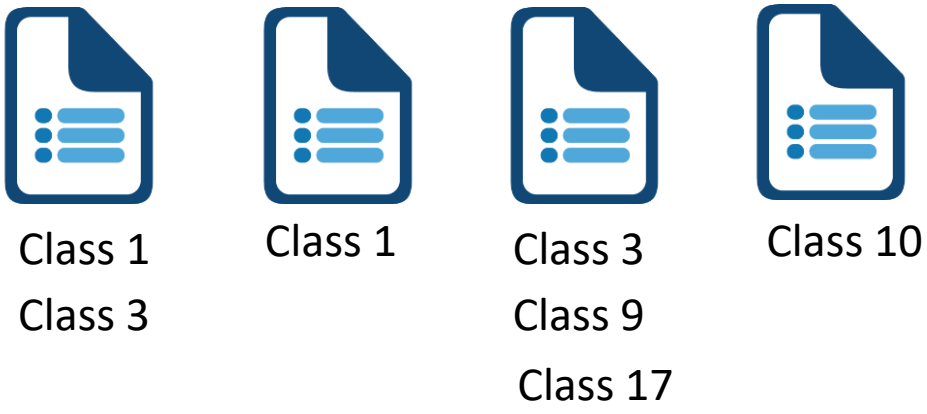
https://arxiv.org/abs/1412.7449

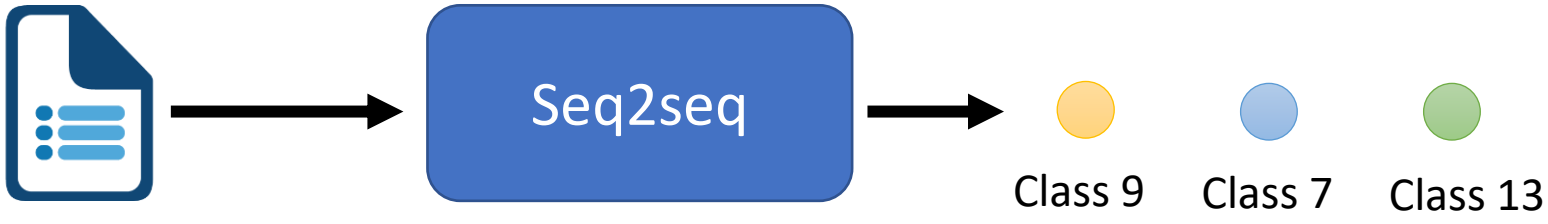
deep learning is very powerful

Seq2seq for Multi-label Classification

c.f. Multi-class Classification

An object can belong to multiple classes.

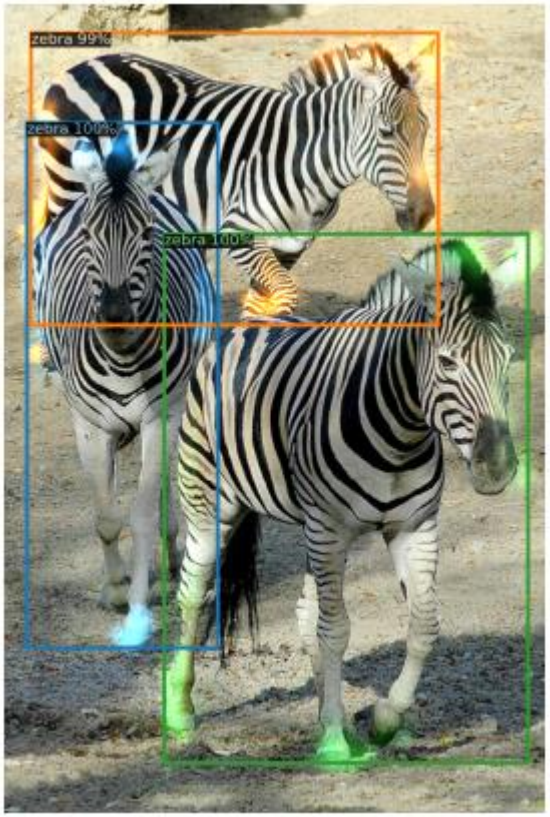


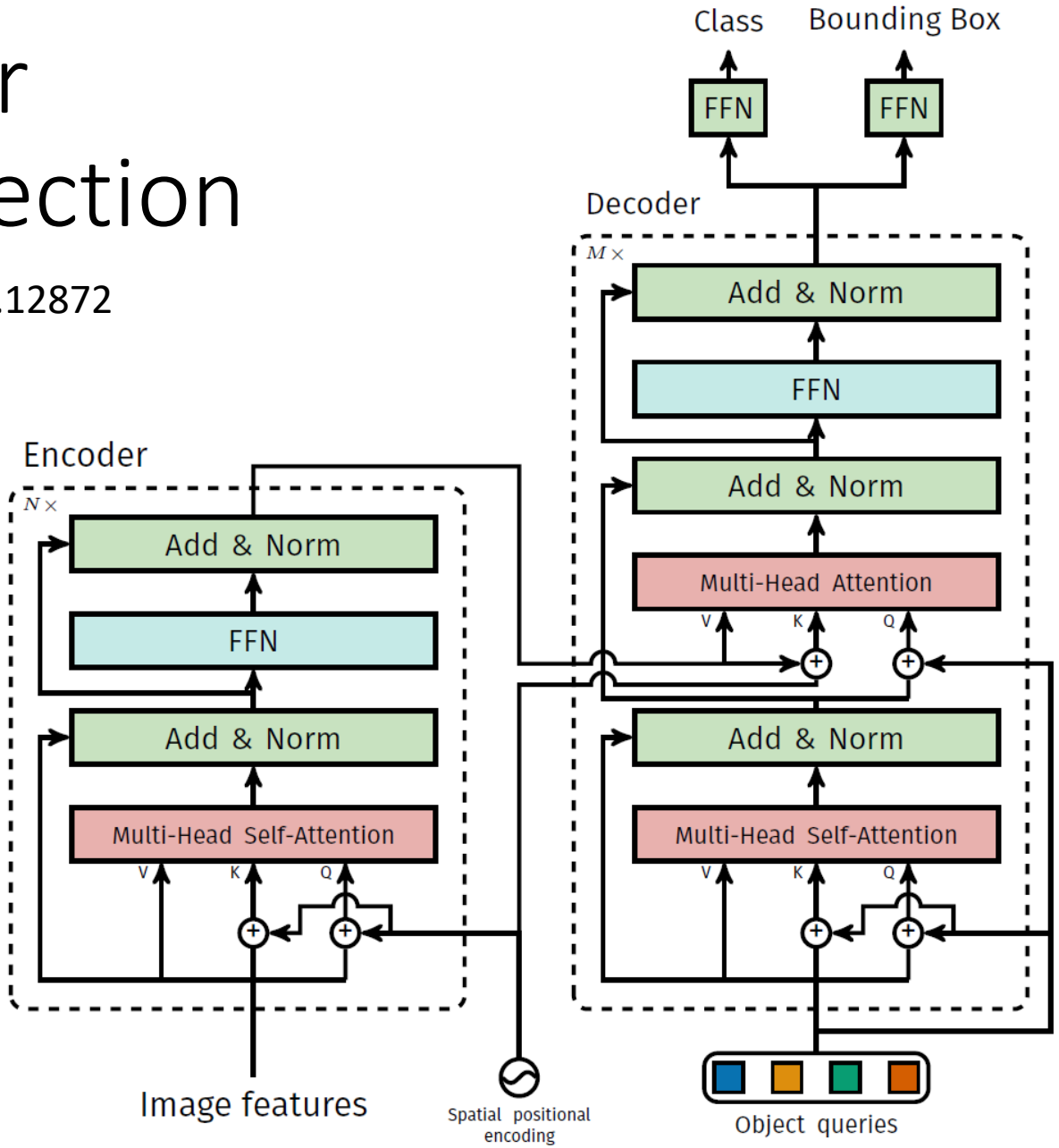


https://arxiv.org/abs/1909.03434
https://arxiv.org/abs/1707.05495

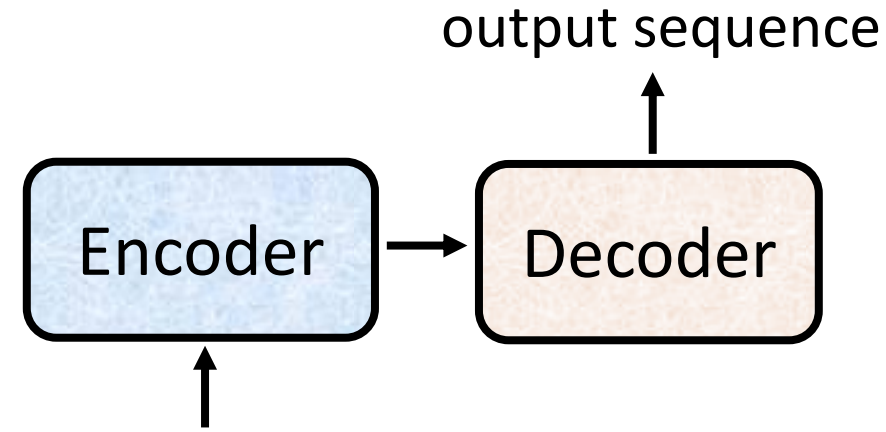
Seq2seq for Object Detection

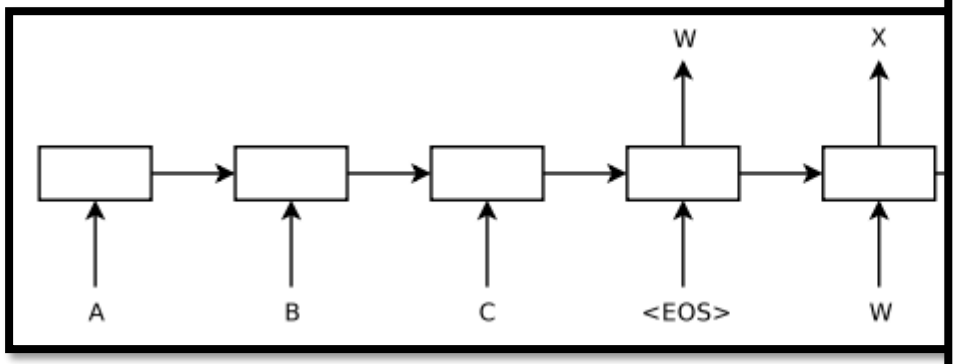
https://arxiv.org/abs/2005.12872



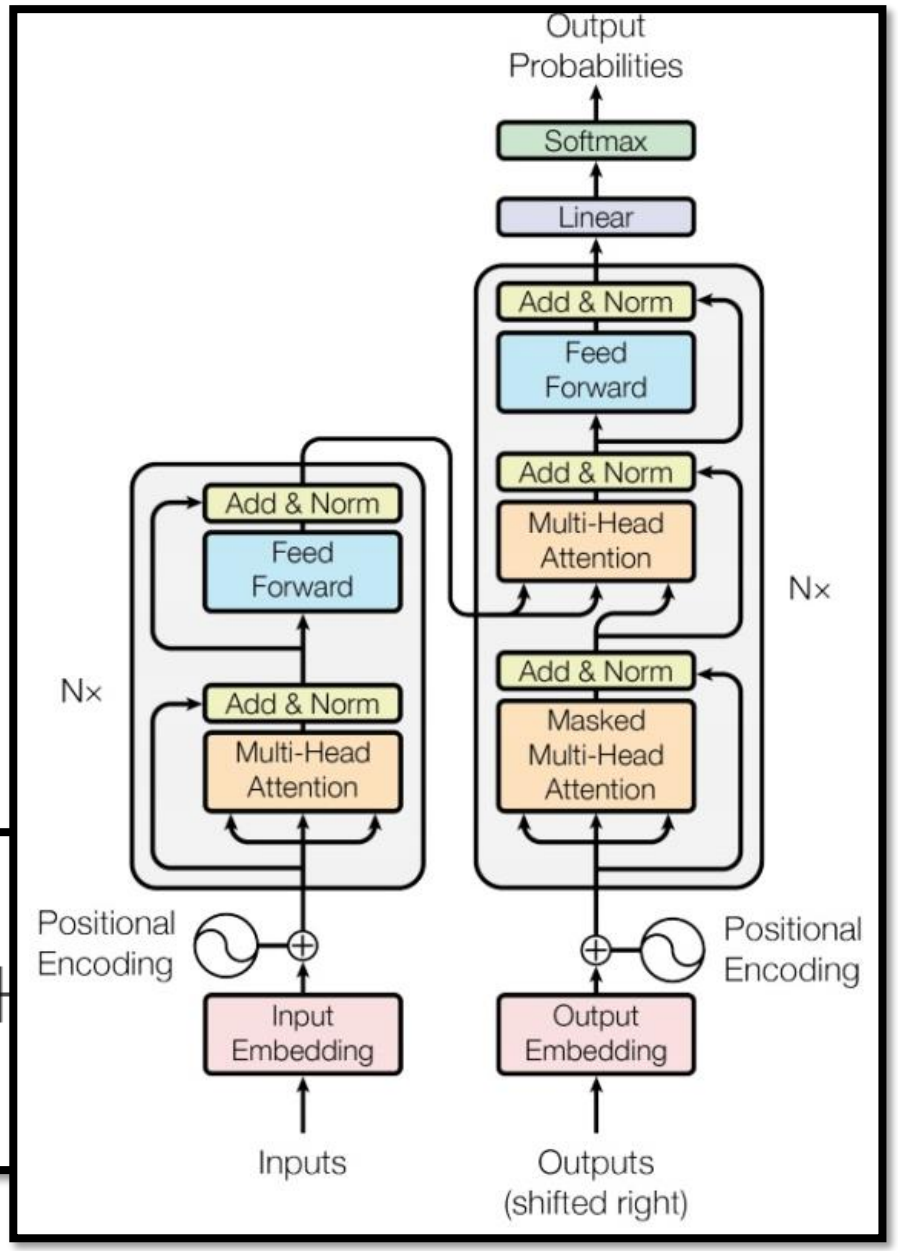


Seq2seq



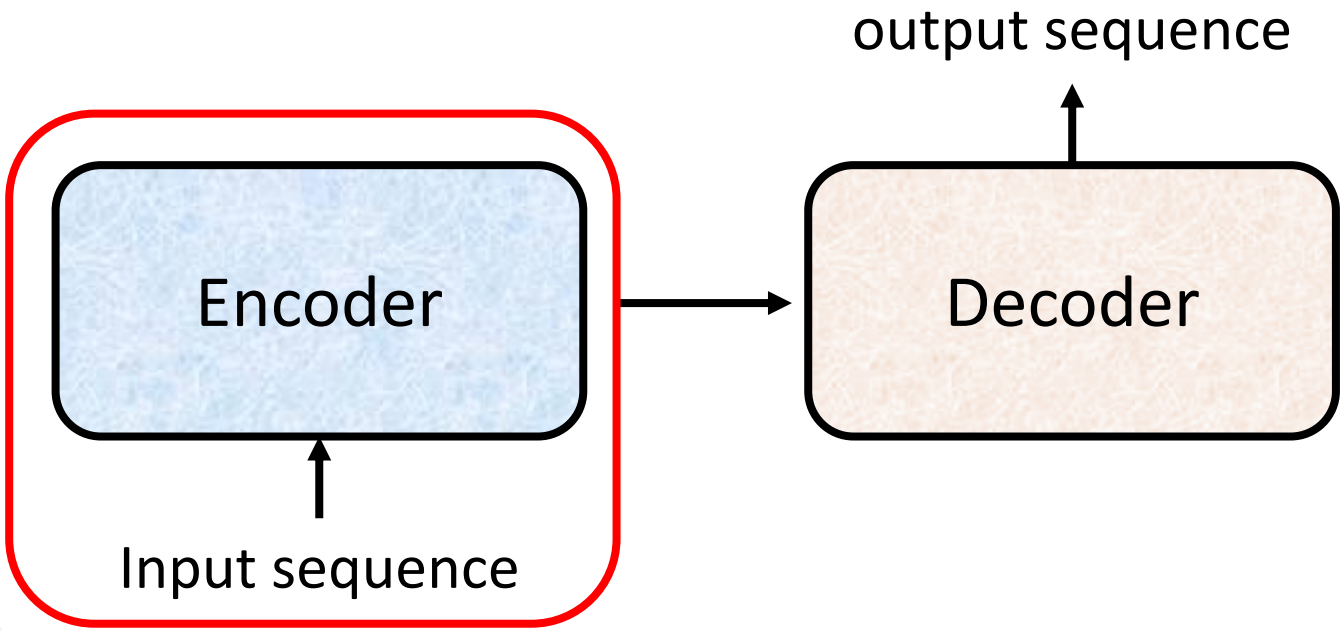


Sequence to Sequence Learning with Neural Networks
https://arxiv.org/abs/1409.3215



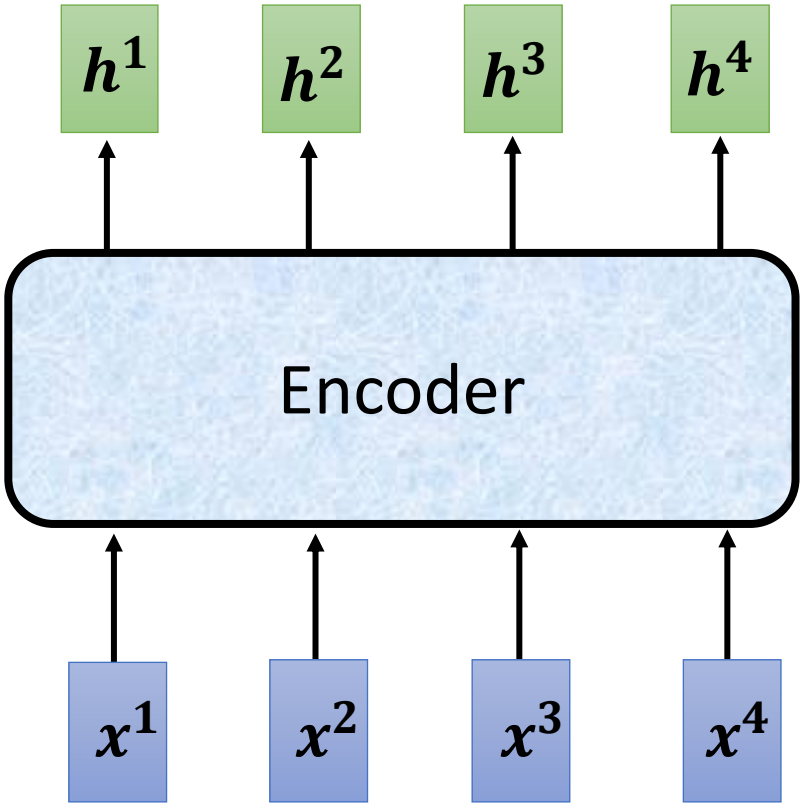
Transformer
https://arxiv.org/abs/1706.03762

Encoder

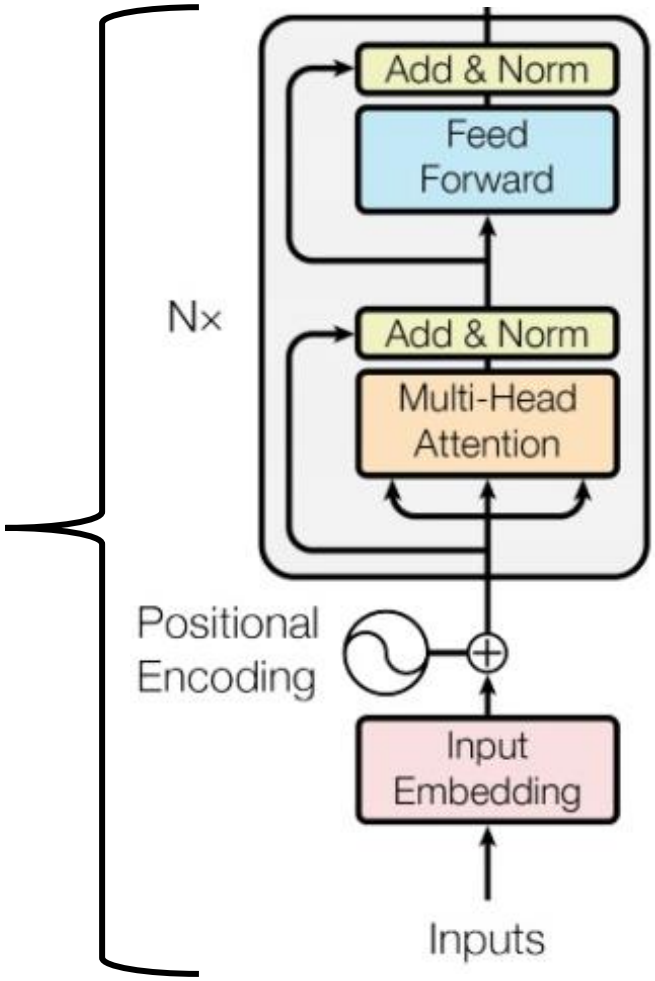


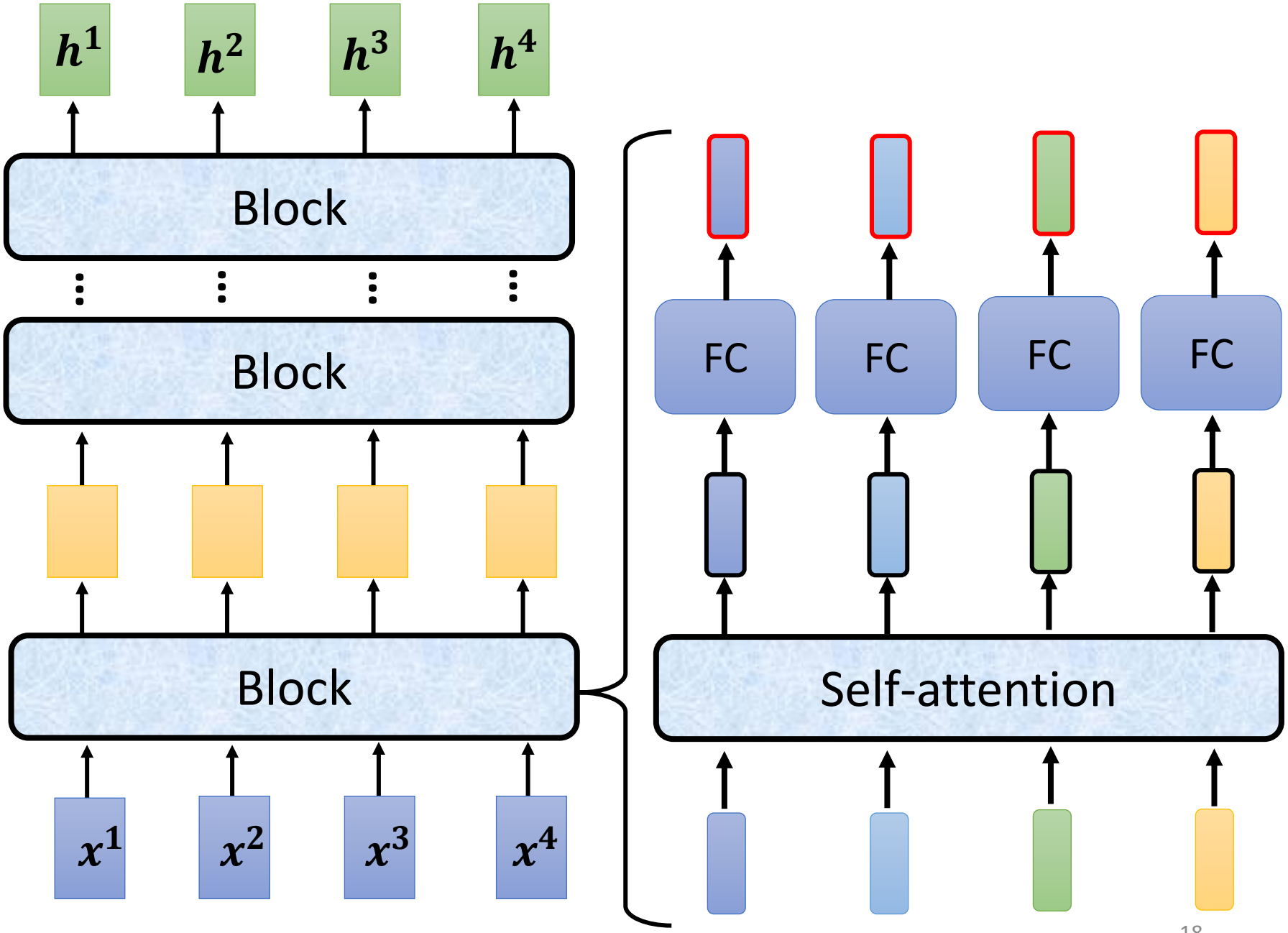
Encoder

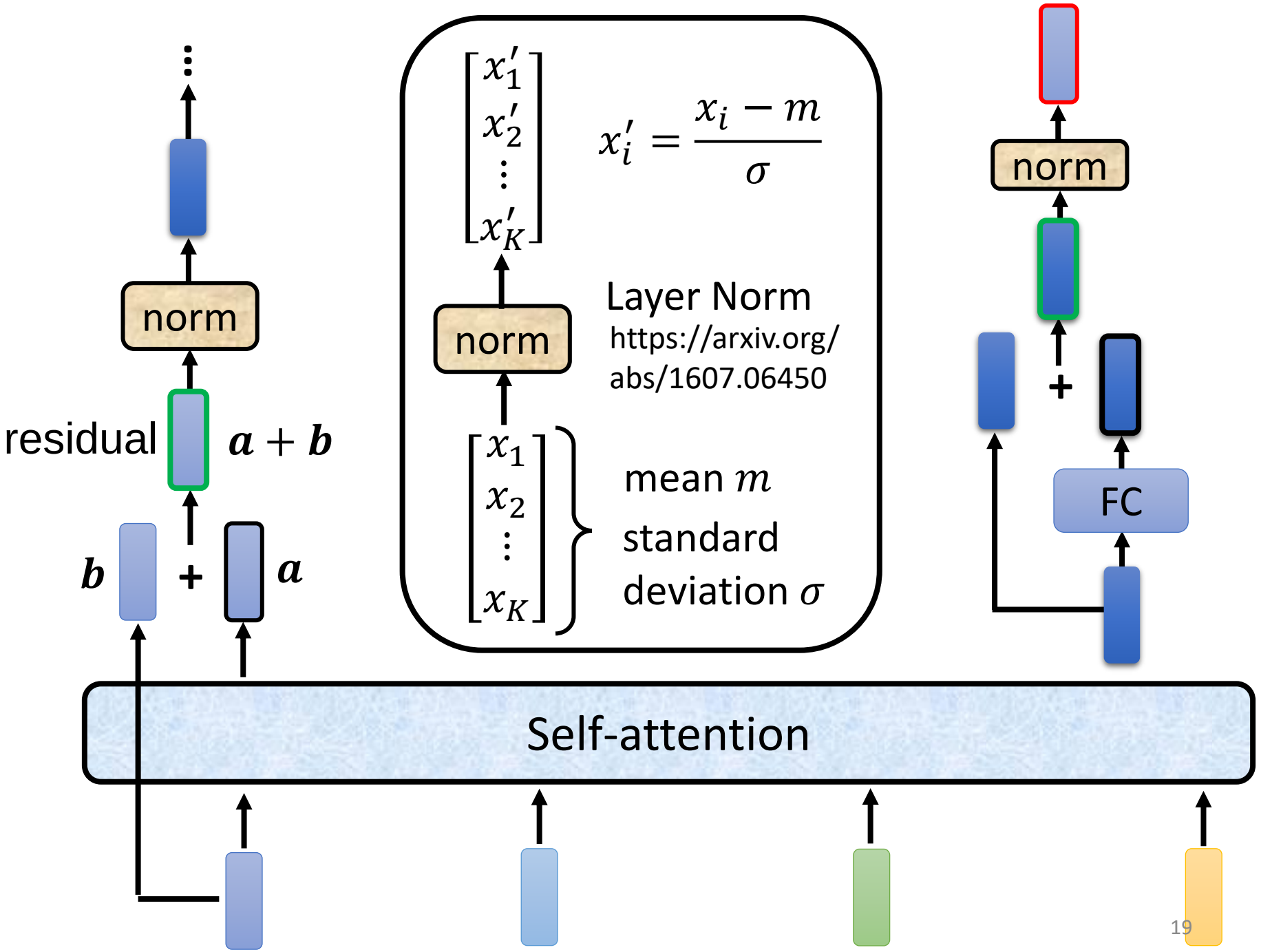
You can use **RNN** or **CNN**.



Transformer's Encoder

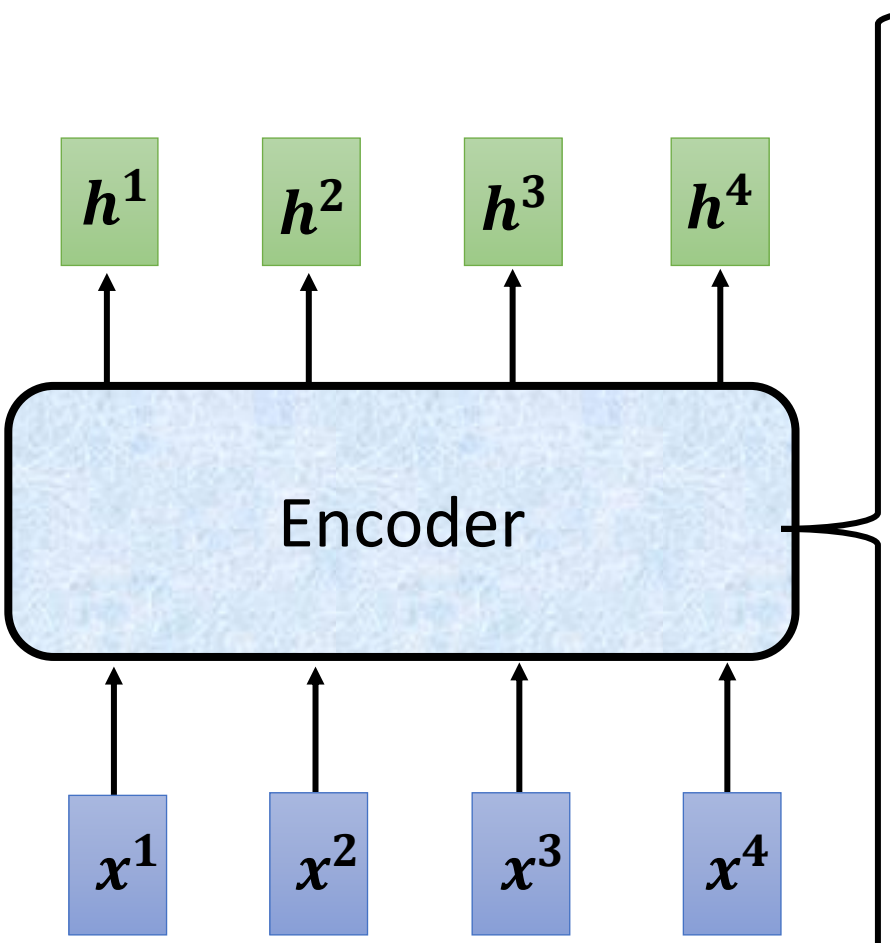


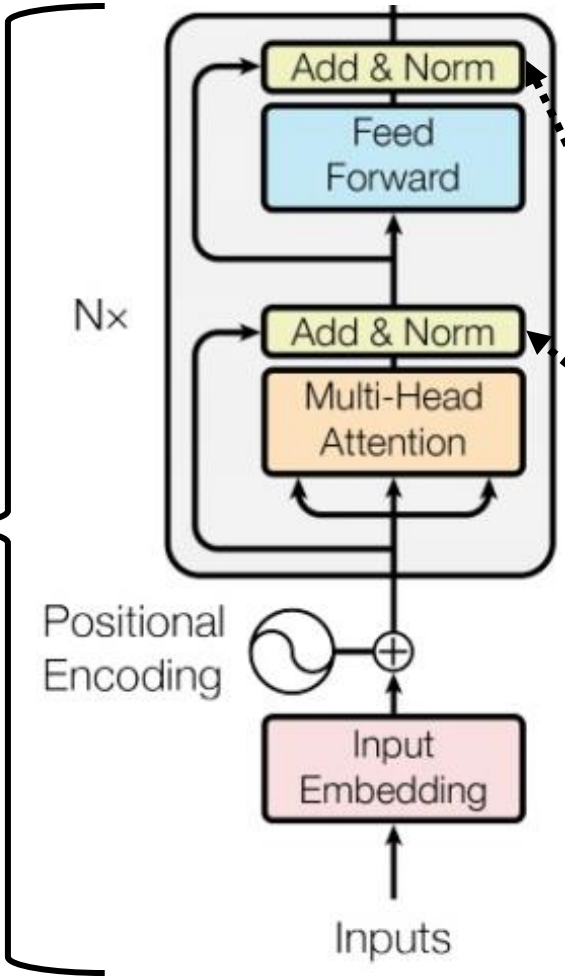

h^1
h^2
h^3
h^4
Block
Block
Block
x^1
x^2
x^3
x^4
FC
FC
FC
FC
Self-attention


norm
residual
$a + b$
b
+
a
$\begin{bmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_K \end{bmatrix}$
$x'_i = \frac{x_i - m}{\sigma}$
norm
Layer Norm
https://arxiv.org/abs/1607.06450
$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_K \end{bmatrix}$
mean m
standard deviation σ
norm
+
FC
Self-attention

BERT

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



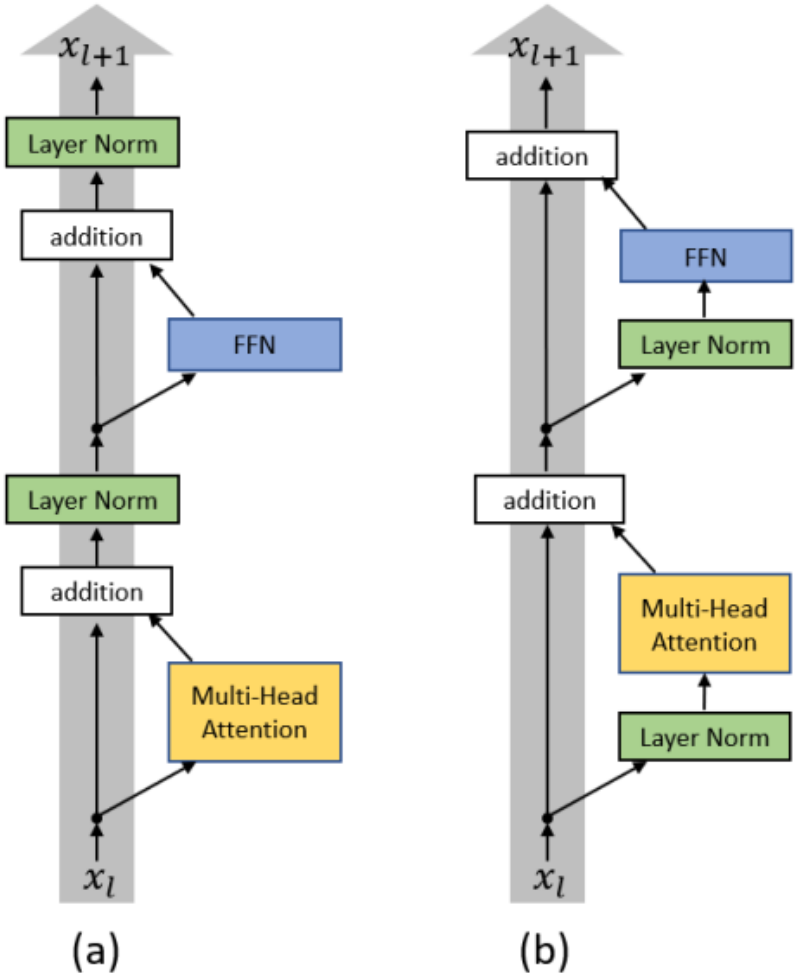


Residual + Layer norm

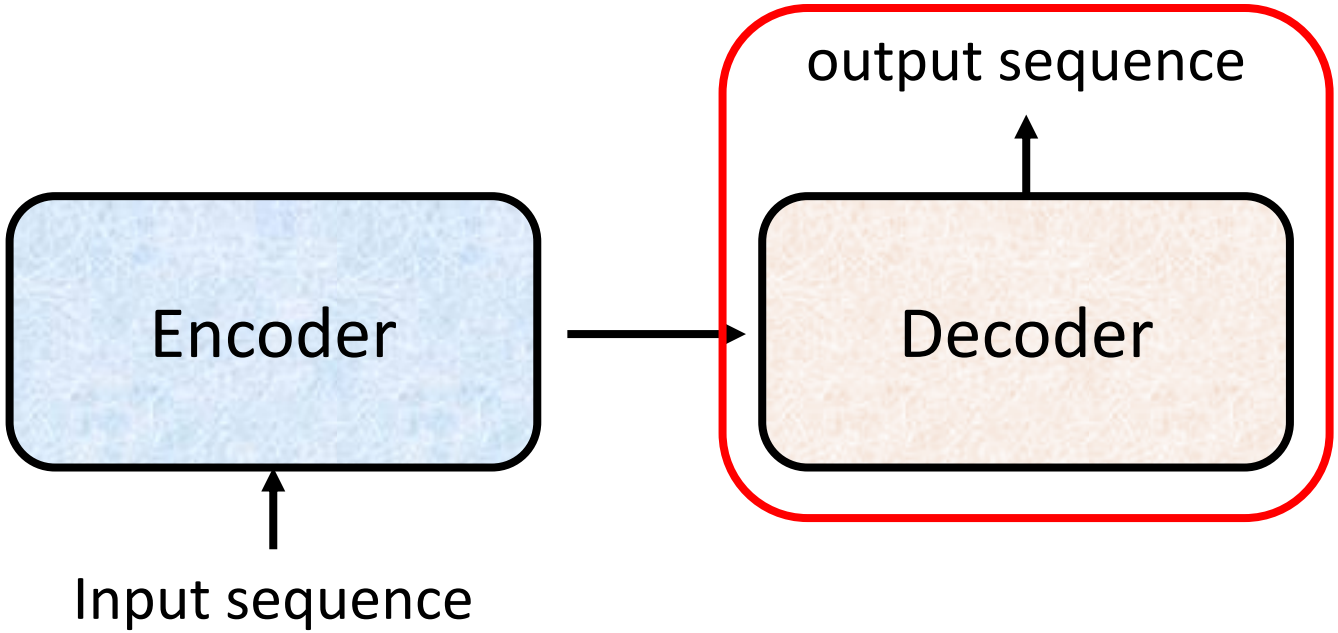
To learn more ……

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

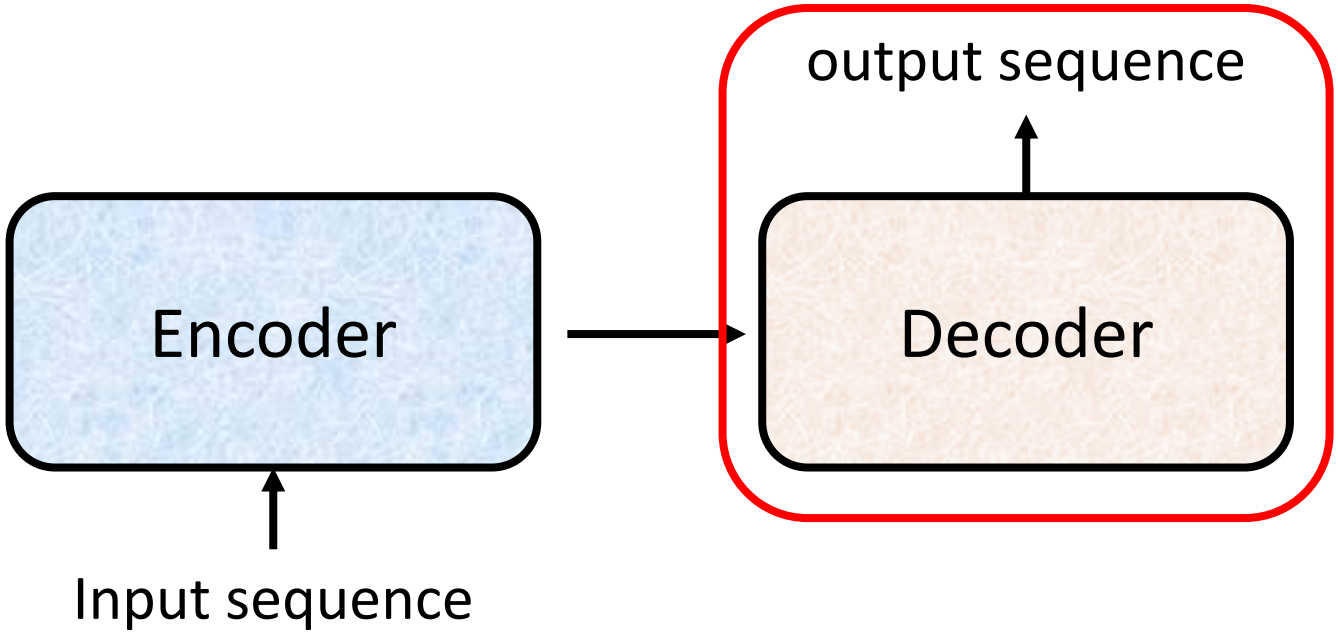
- PowerNorm: Rethinking Batch Normalization in Transformers
- https://arxiv.org/abs/2003.07845



Decoder

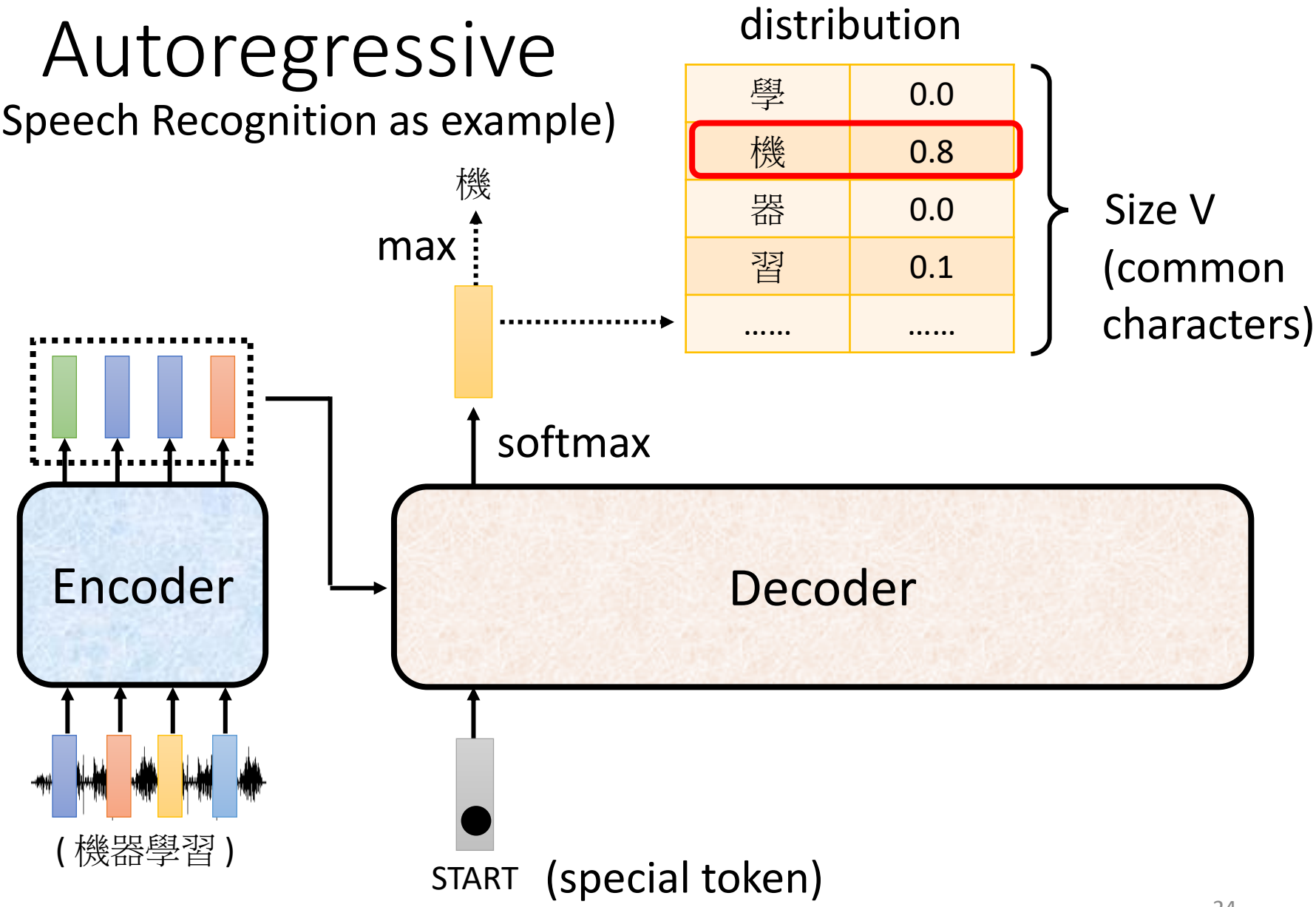


Decoder – Autoregressive (AT)

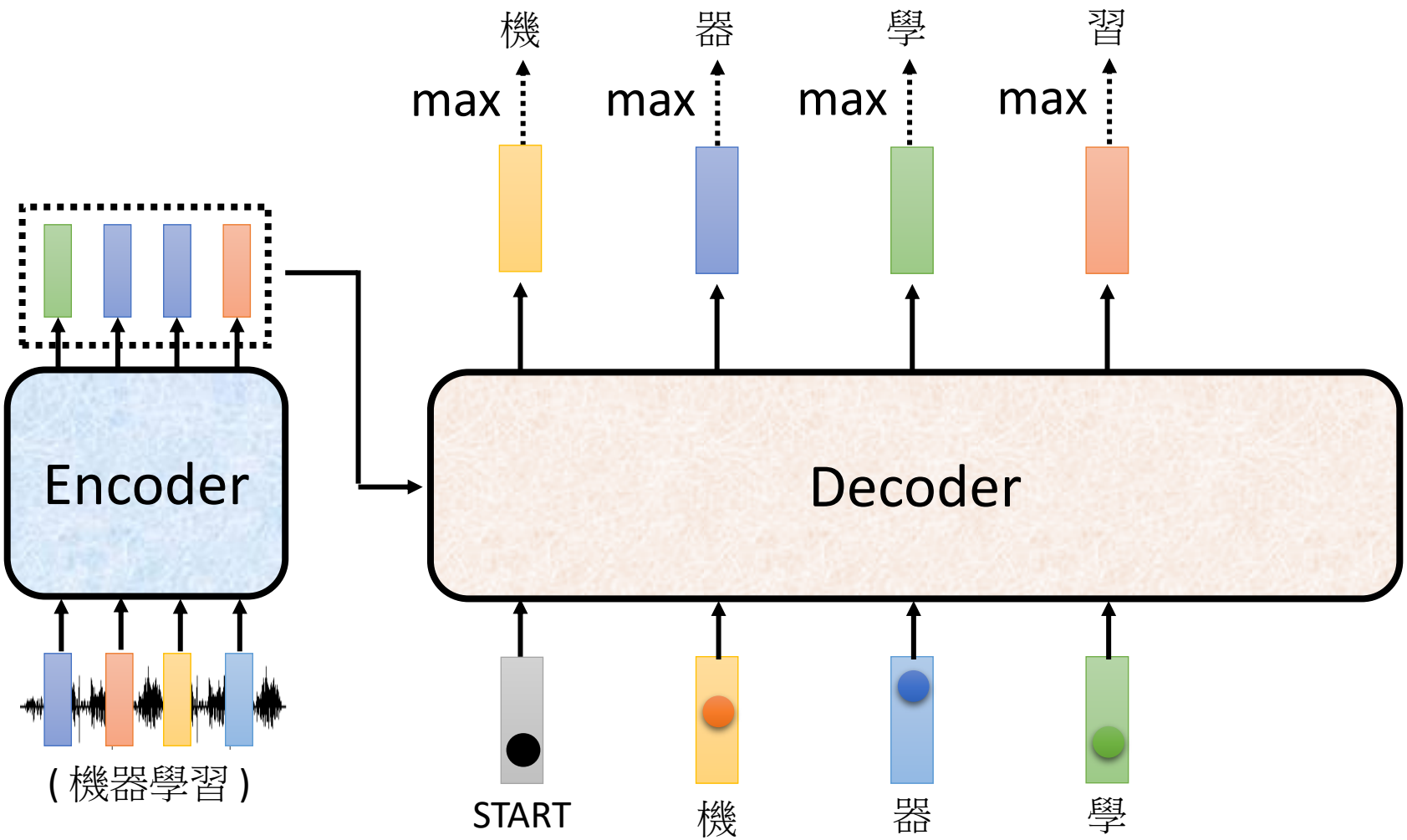


Autoregressive

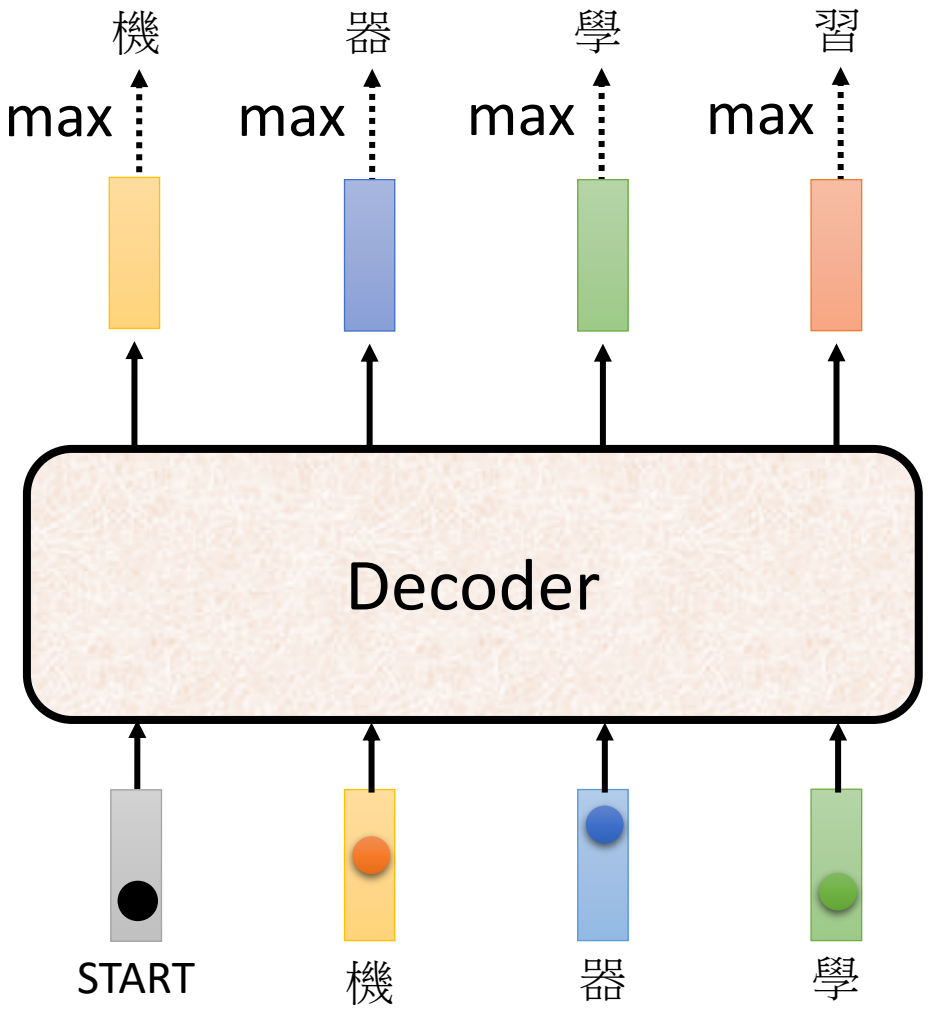
(Speech Recognition as example)

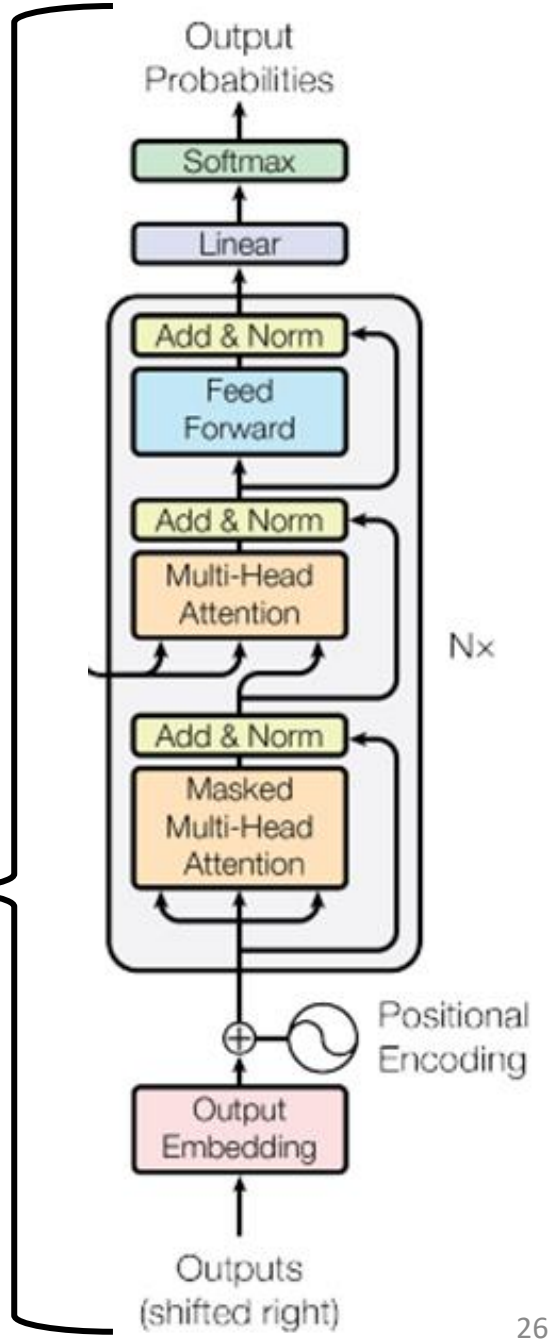


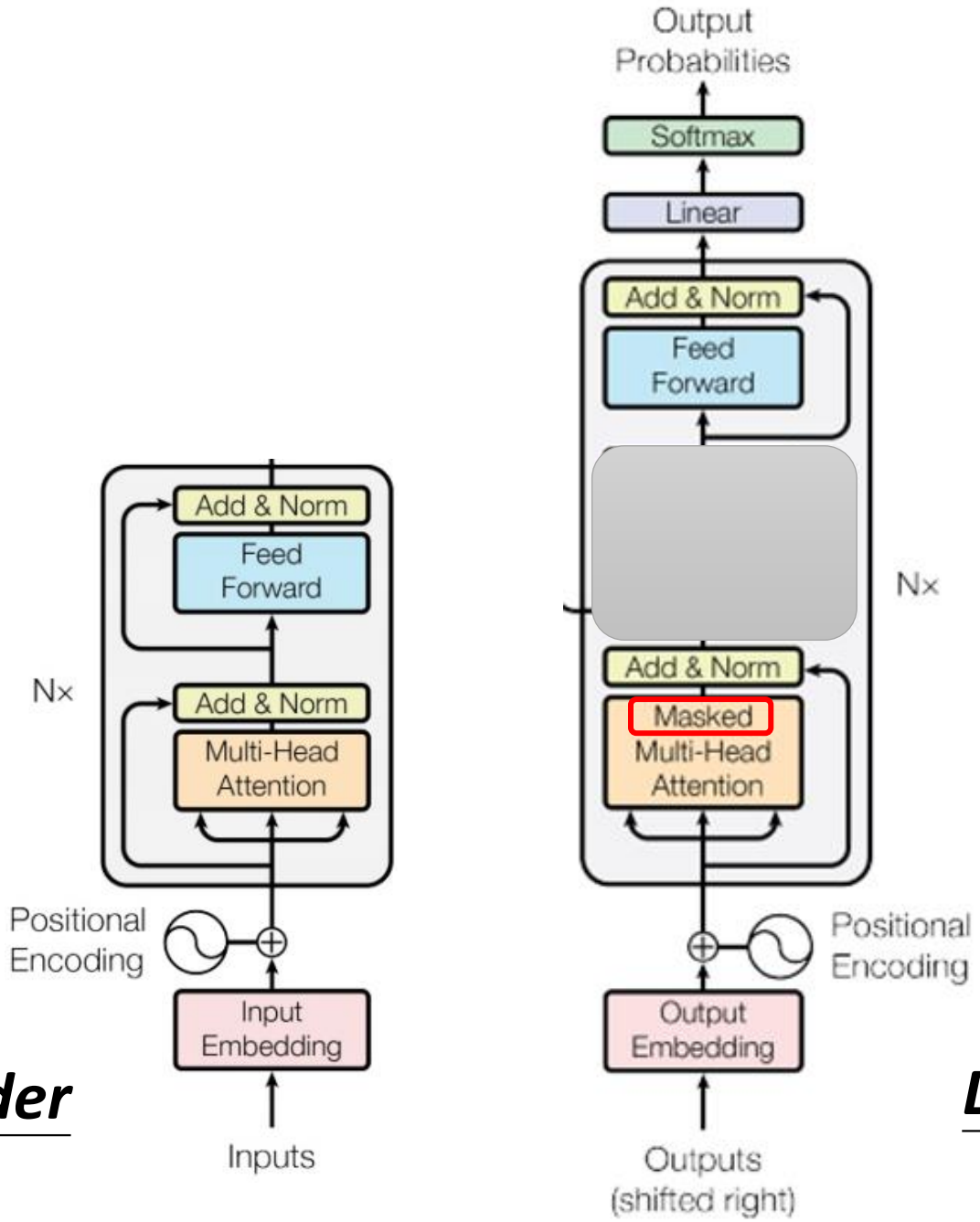
Autoregressive



ignore the input from the encoder here ☺



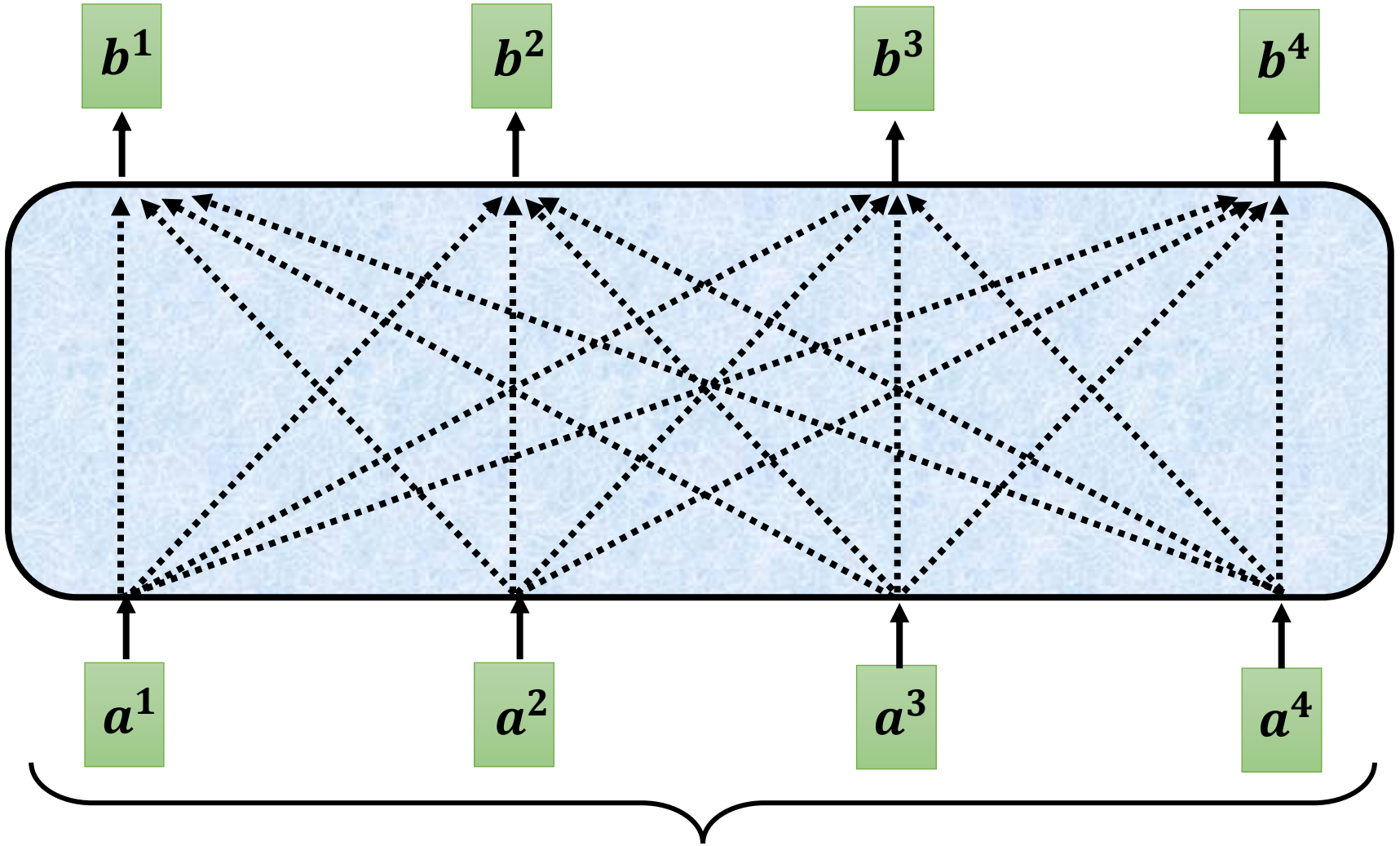




Encoder

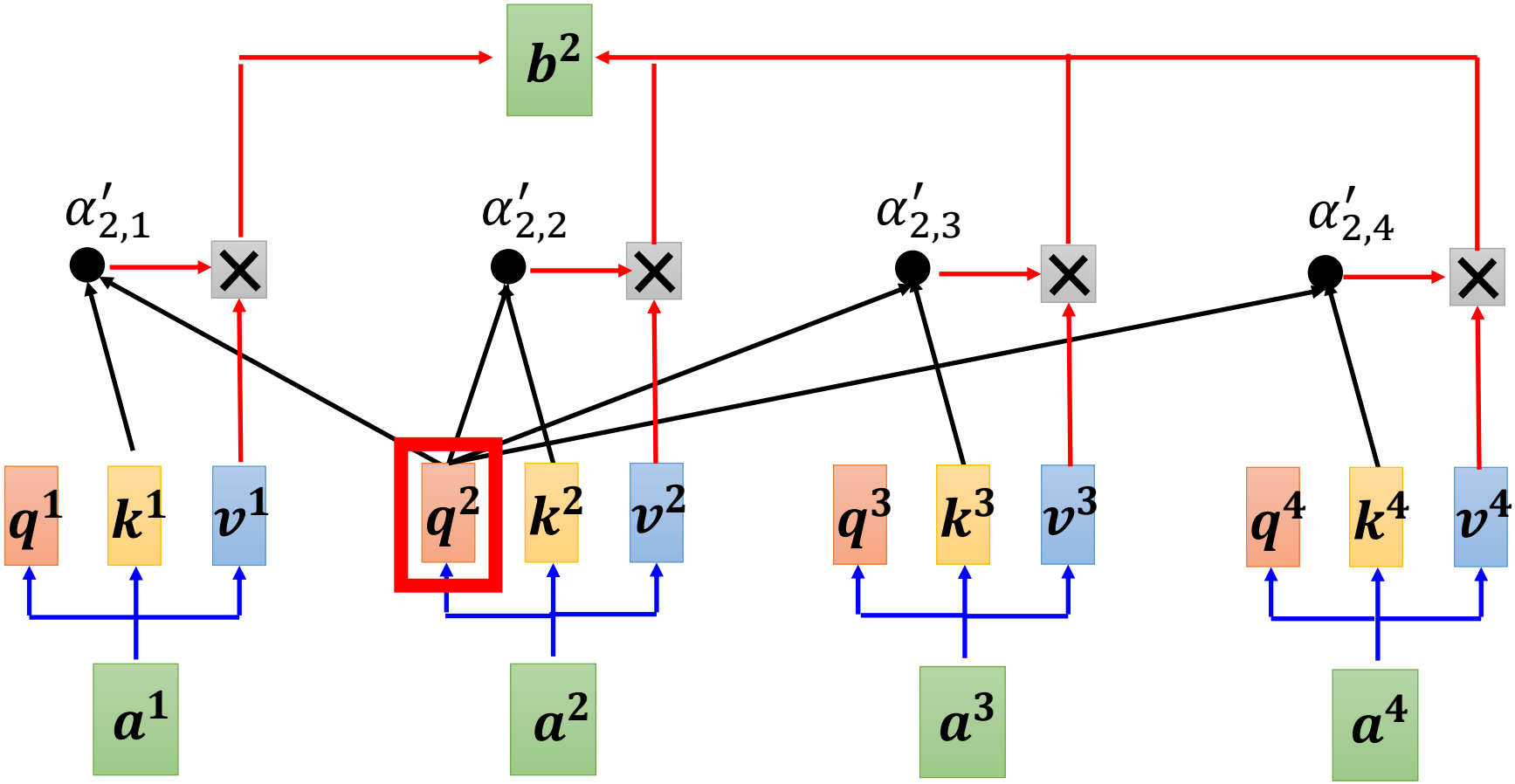
Decoder

Self-attention → *Masked Self-attention*



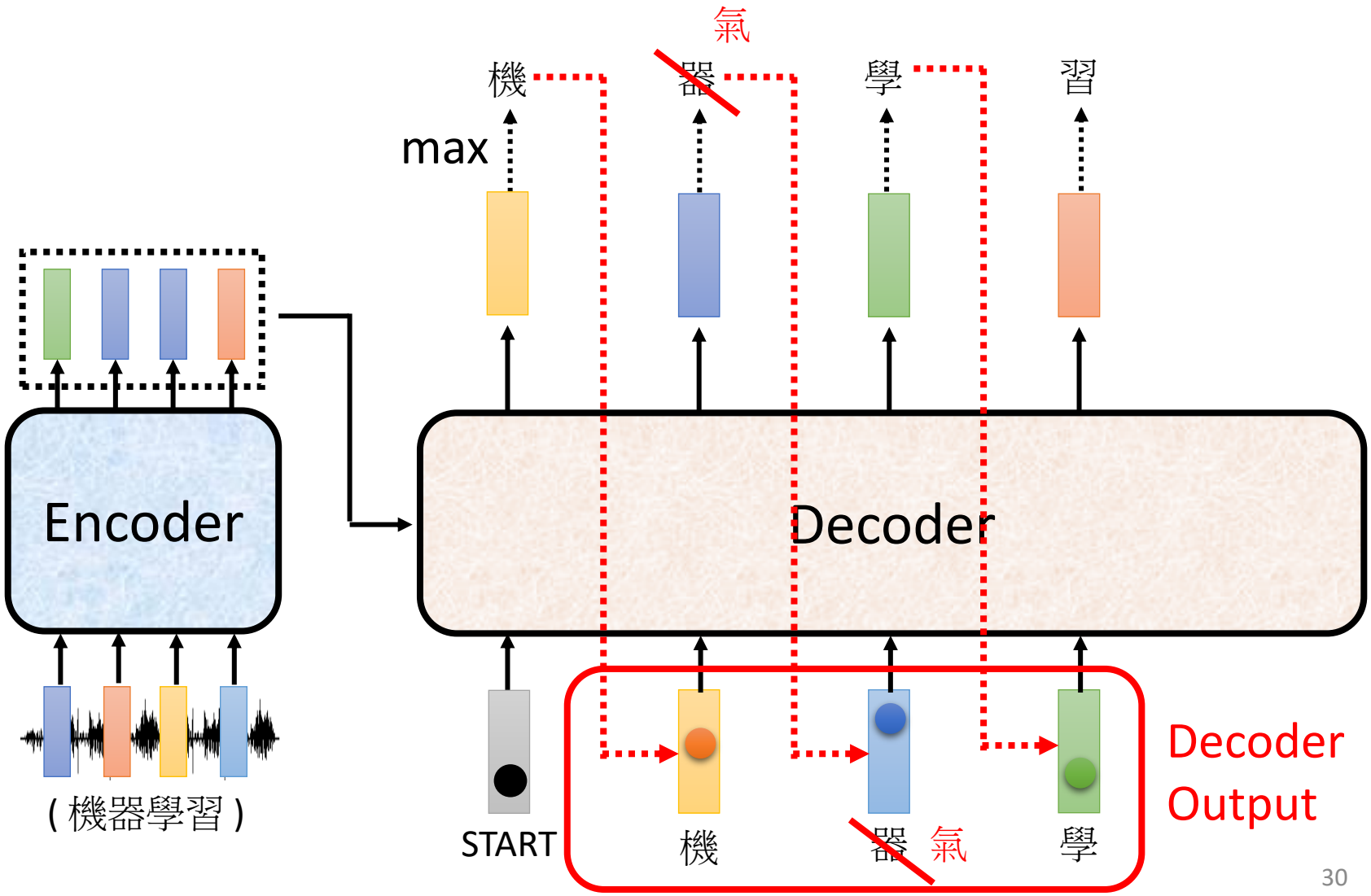
Can be either **input** or **a hidden layer**

Self-attention → *Masked Self-attention*



Why masked? Consider how does decoder work

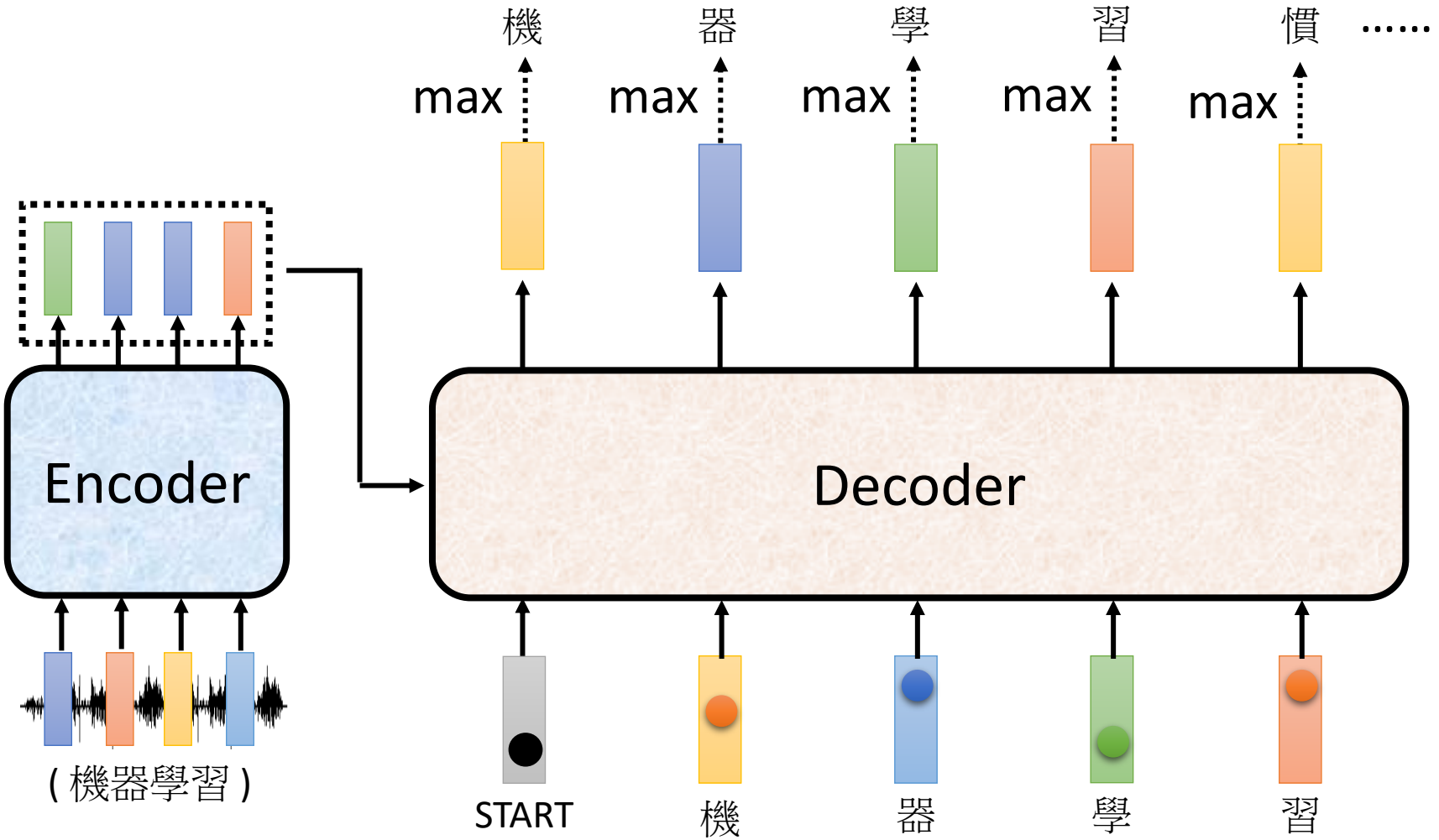
Autoregressive



Autoregressive

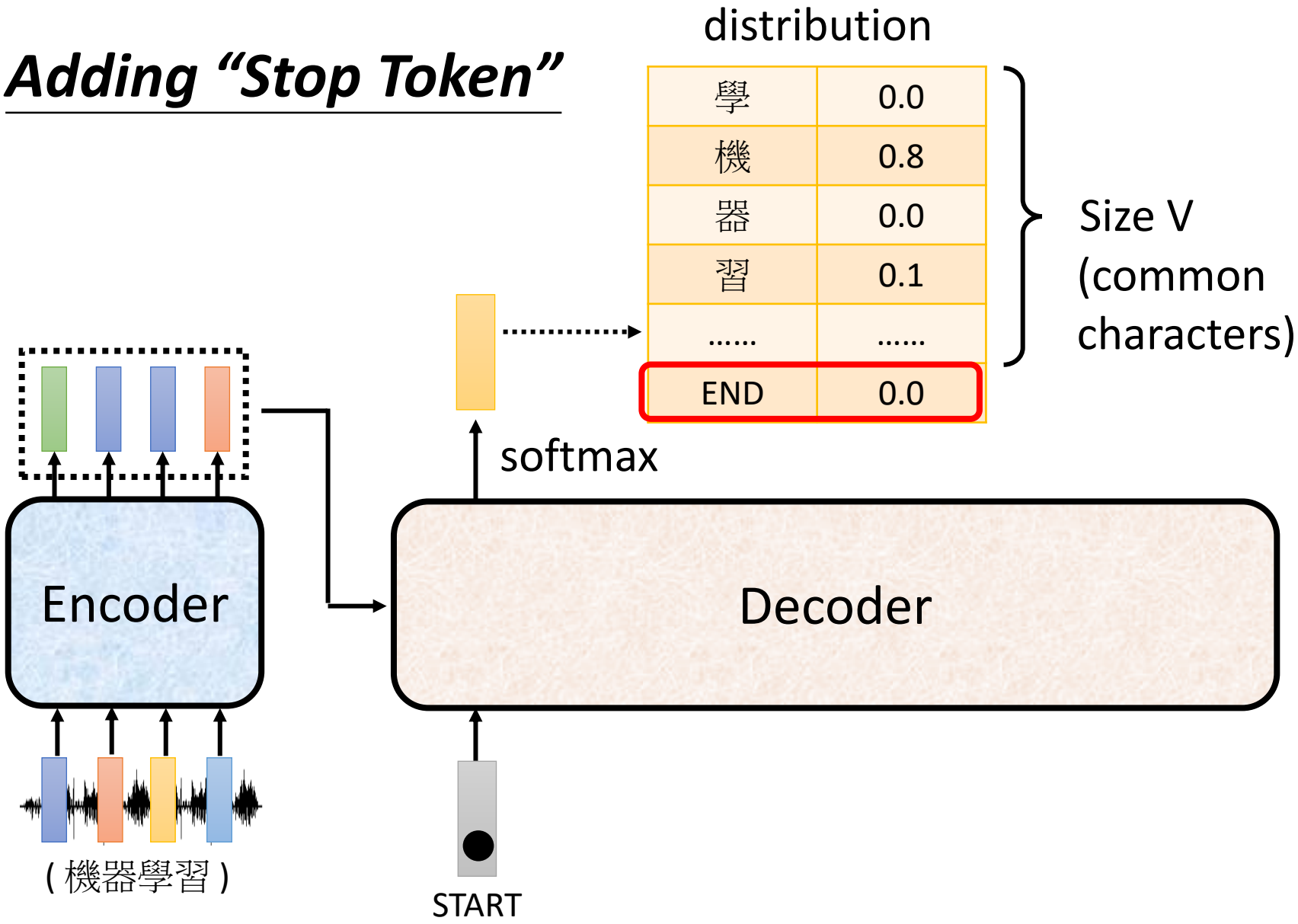
We do not know the correct output length.

Never stop!



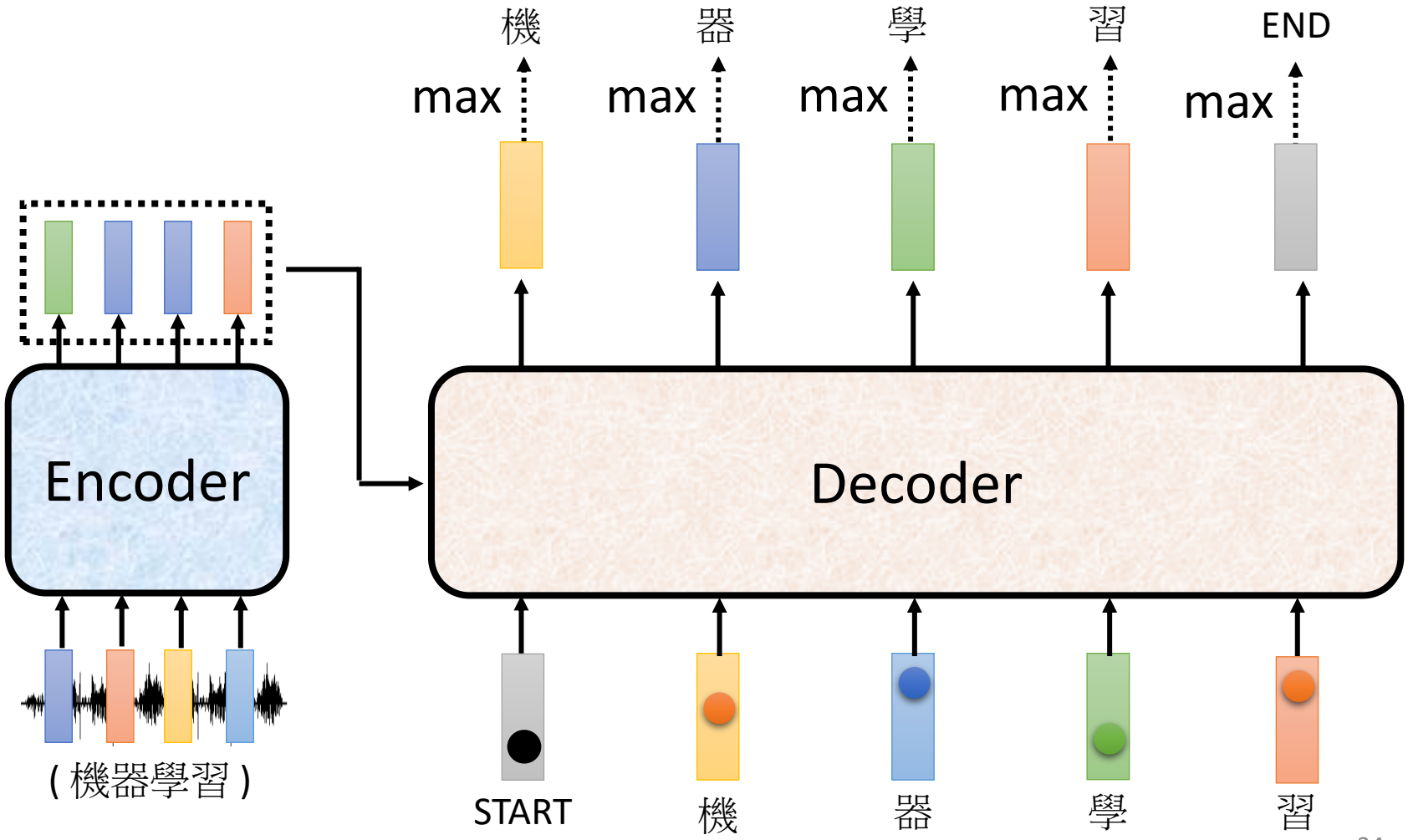
推文接龍 (Tweet Solitaire)

Adding "Stop Token"

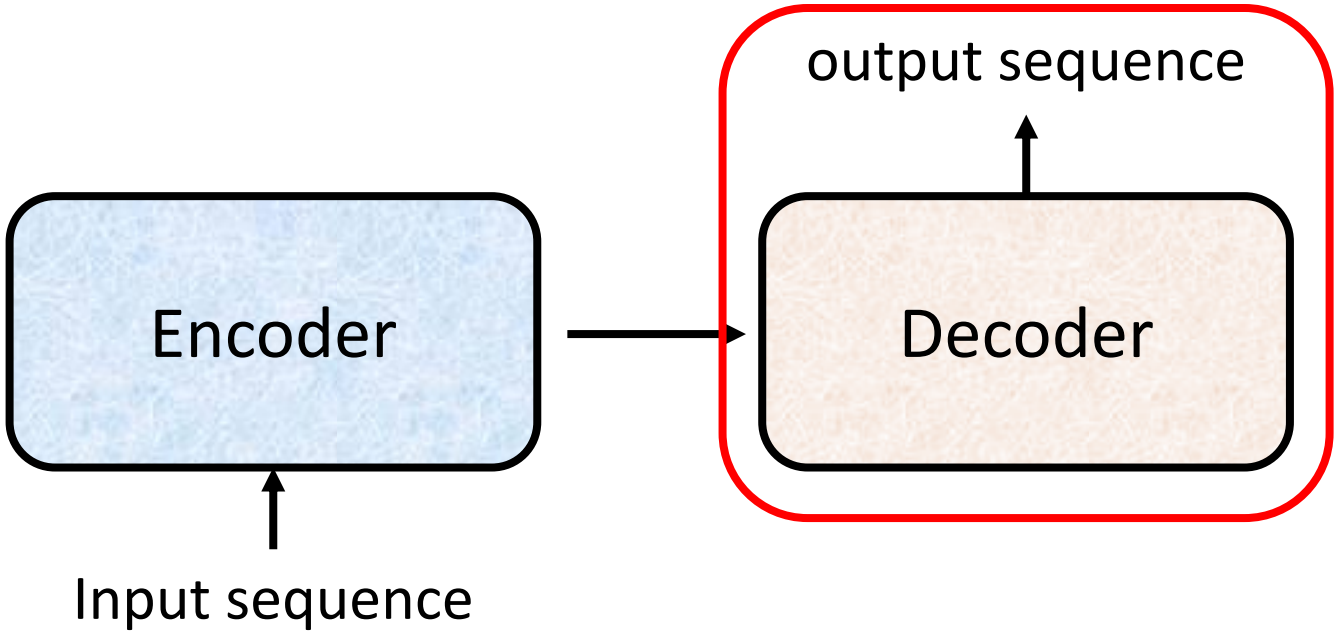


Autoregressive

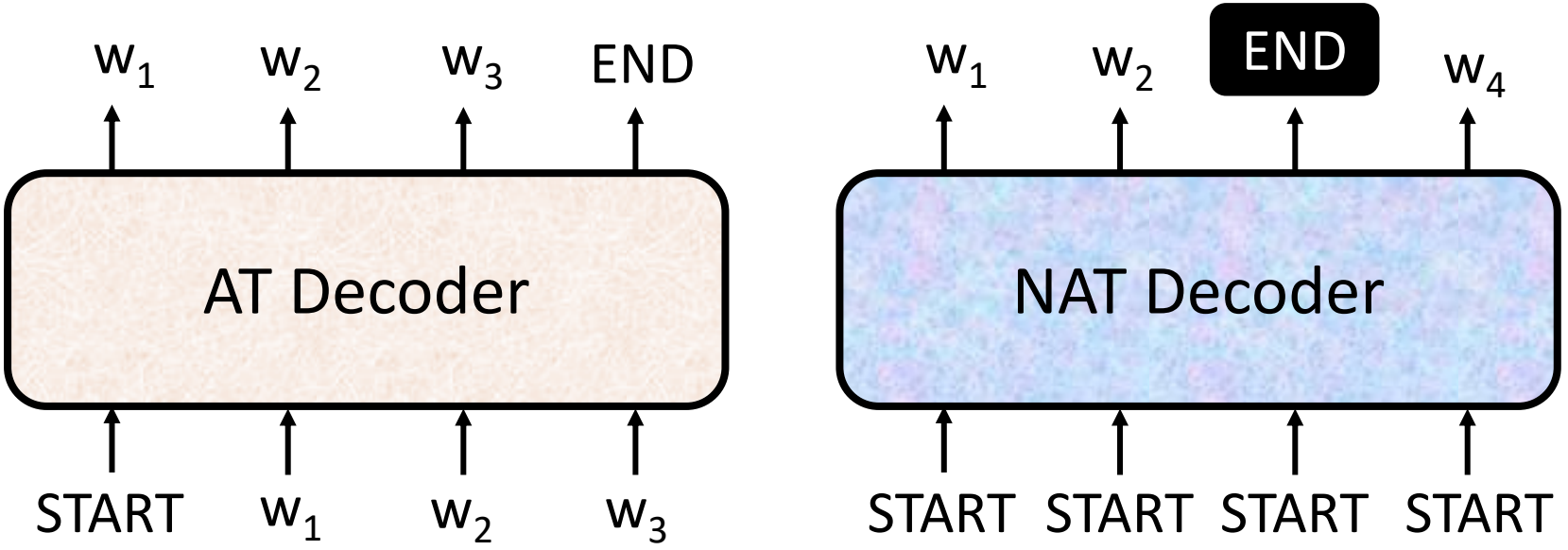
Stop at here!



Decoder
– Non-autoregressive (NAT)



AT v.s. NAT



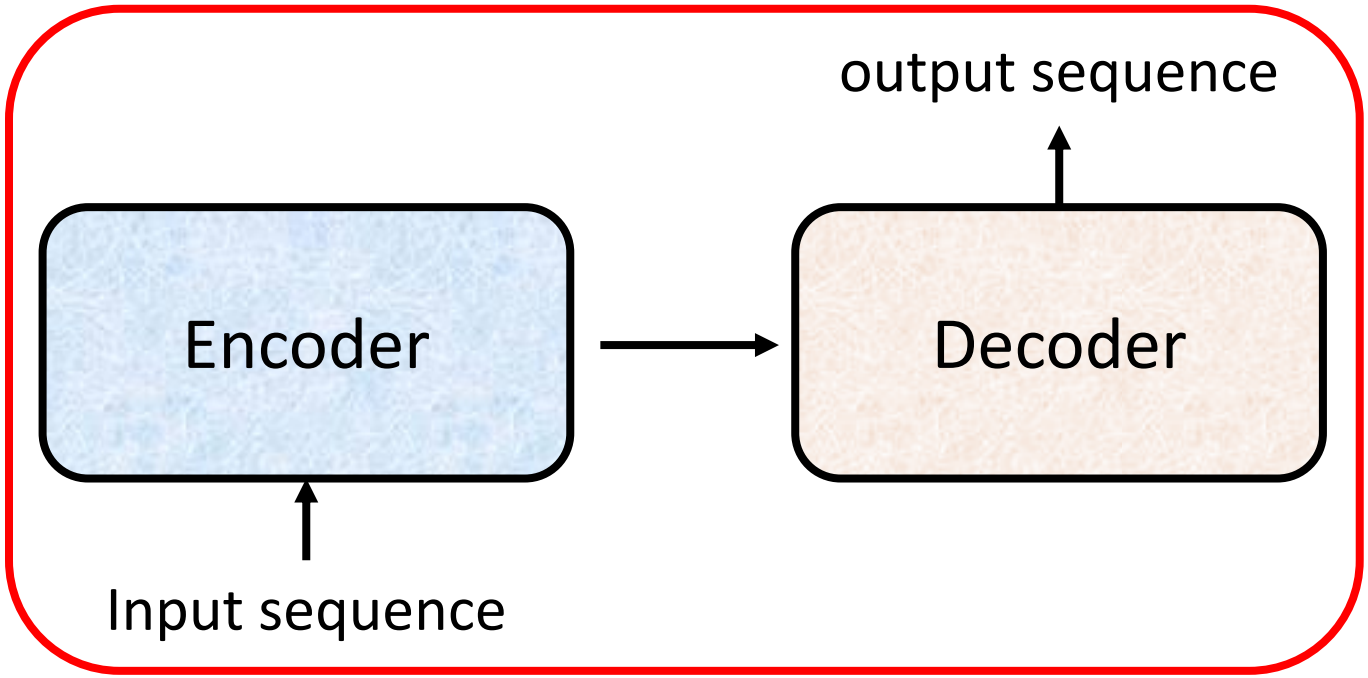
- How to decide the output length for NAT decoder?
  - Another predictor for output length
  - Output a very long sequence, ignore tokens after END
- Advantage: parallel, more stable generation (e.g., TTS)
- NAT is usually worse than AT (why? **Multi-modality**)

To learn more ……

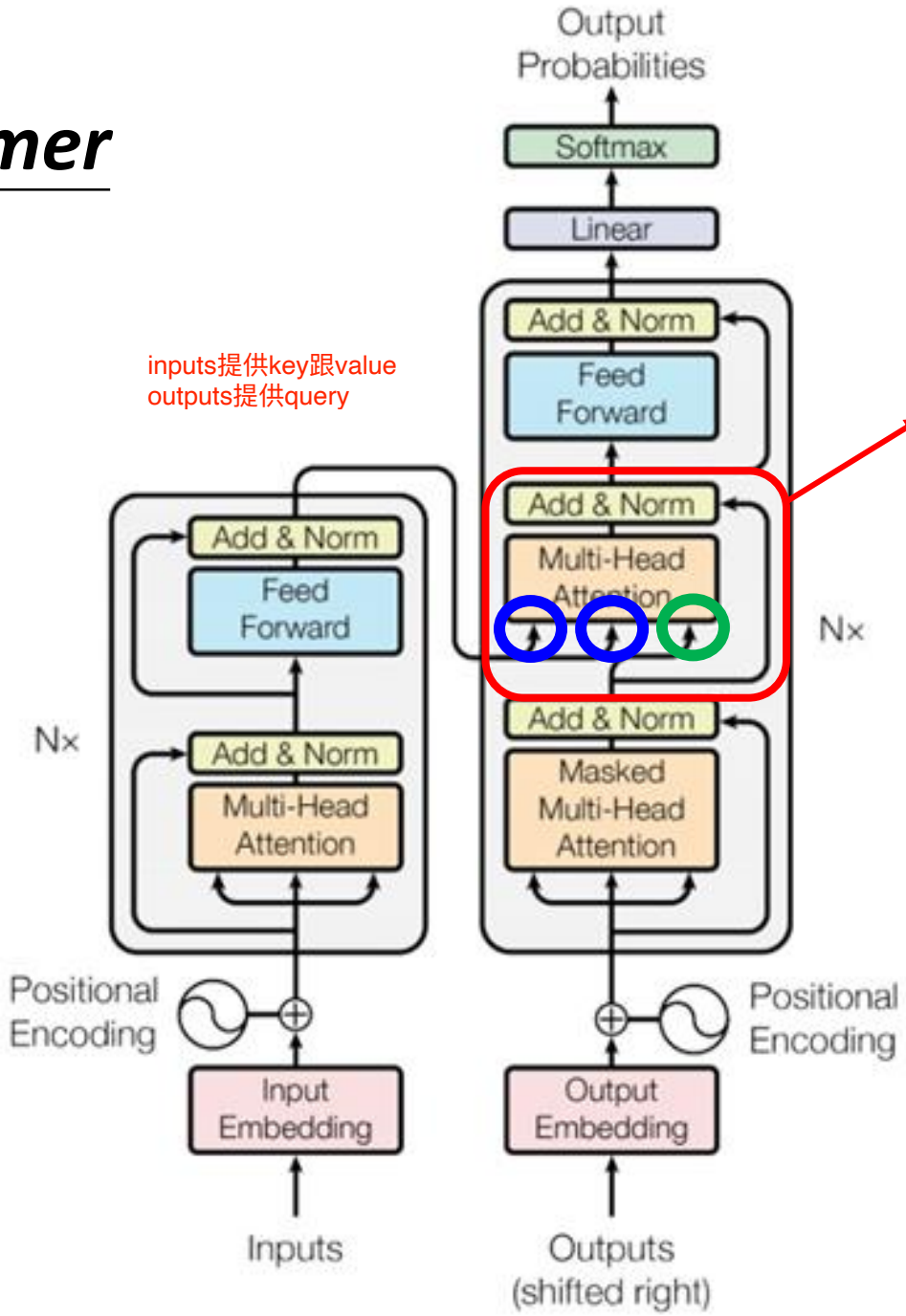
https://youtu.be/jvyKmU4OM3c

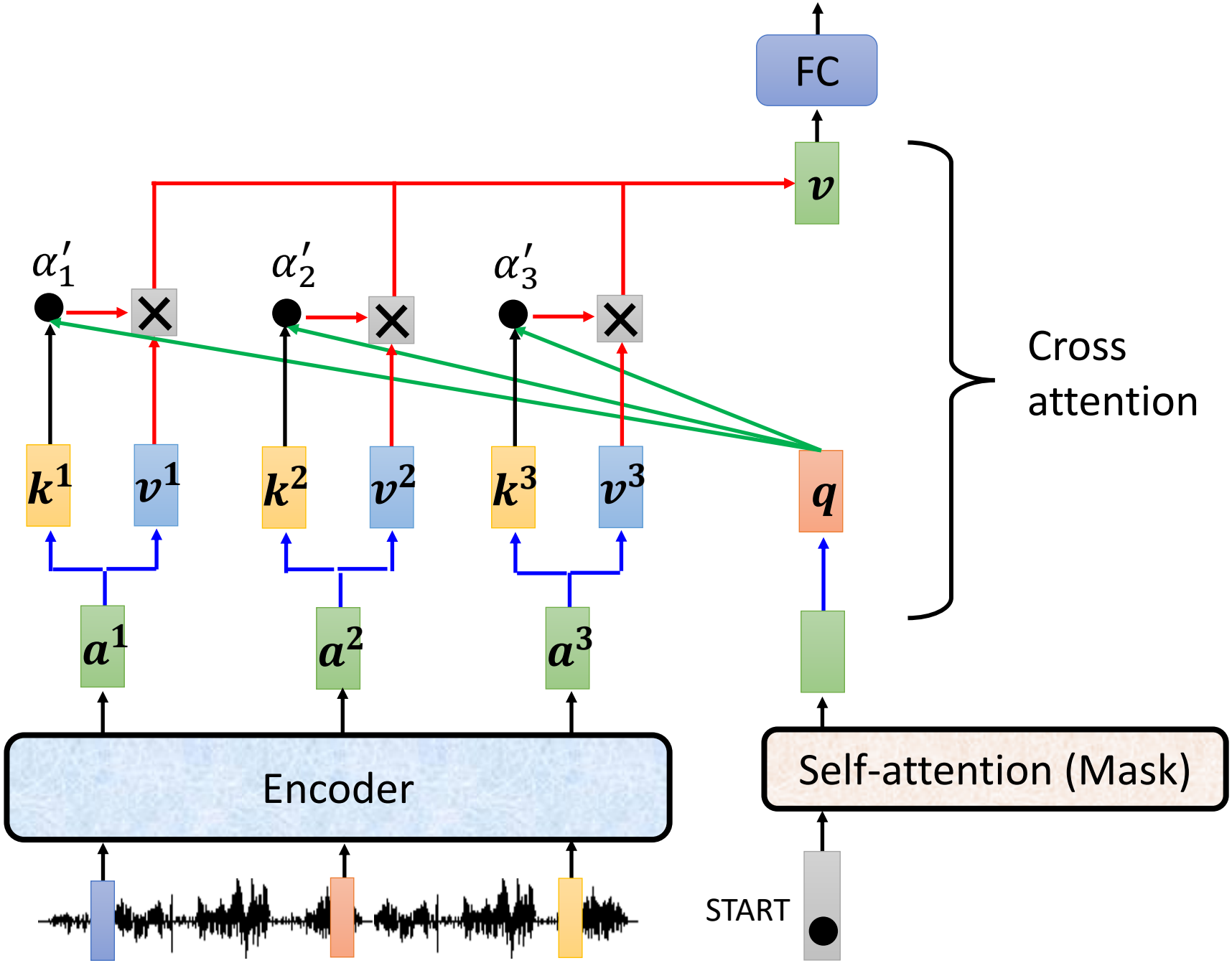
(in Mandarin)

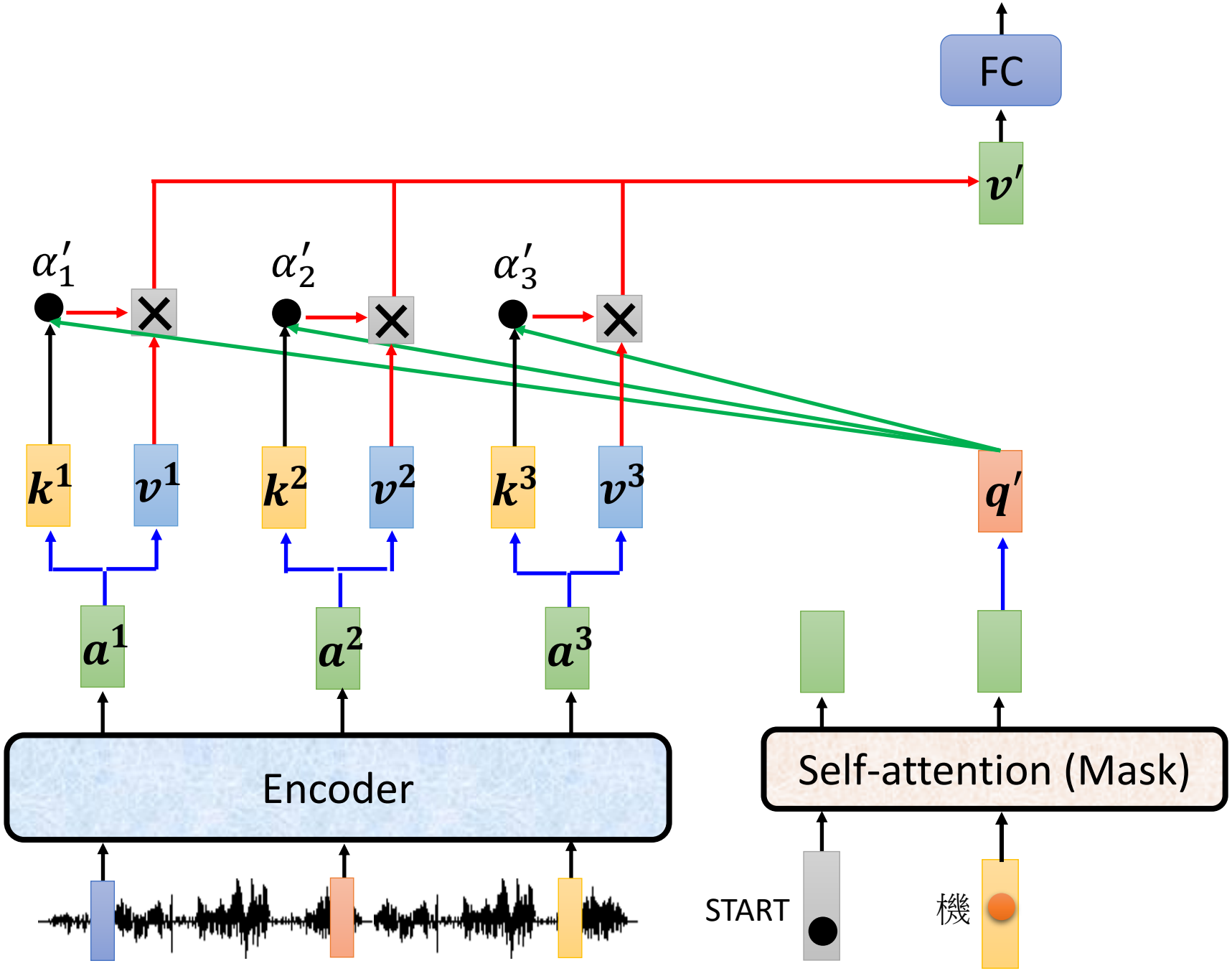
Encoder-Decoder



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

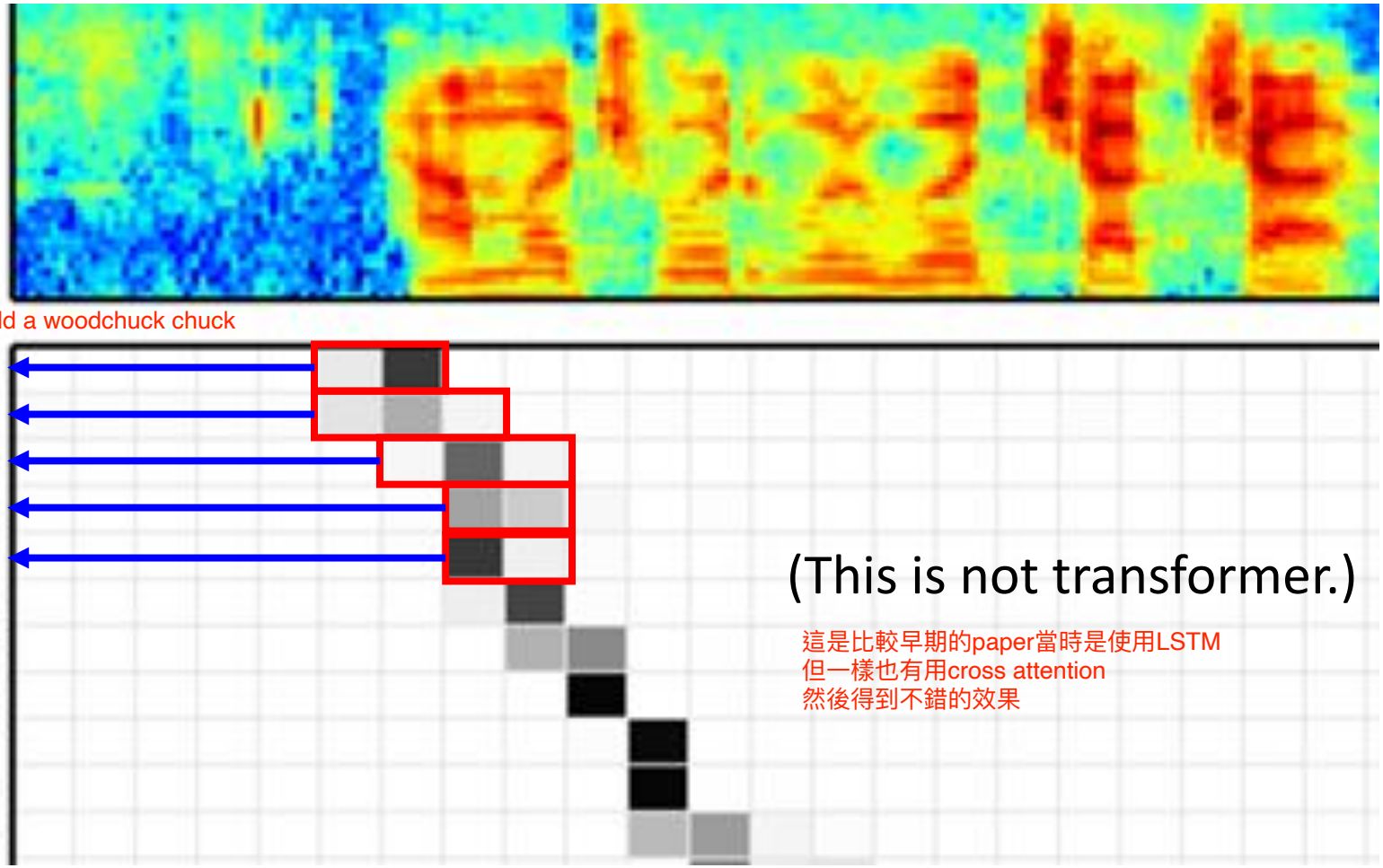

FC
v'
α'_1
α'_2
α'_3
k^1
v^1
k^2
v^2
k^3
v^3
q'
a^1
a^2
a^3
Encoder
Self-attention (Mask)
START
機

Cross Attention

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

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

聲音訊號
實際上內容是
How much wood would a woodchuck chuck



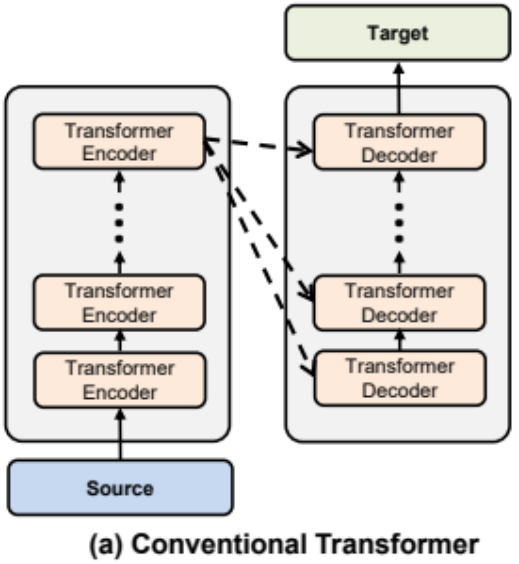
(This is not transformer.)

這是比較早期的paper當時是使用LSTM
但一樣也有用cross attention
然後得到不錯的效果

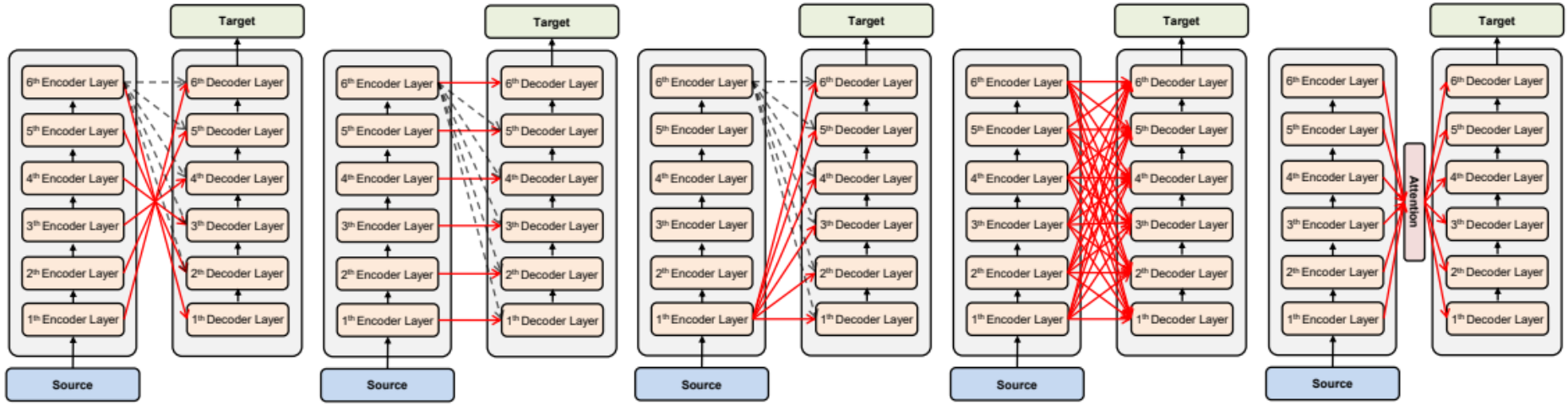
原本paper的cross attention僅有參考encoder最後一層的輸出
但這篇paper探討其他種cross attention的可能性

Cross Attention

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

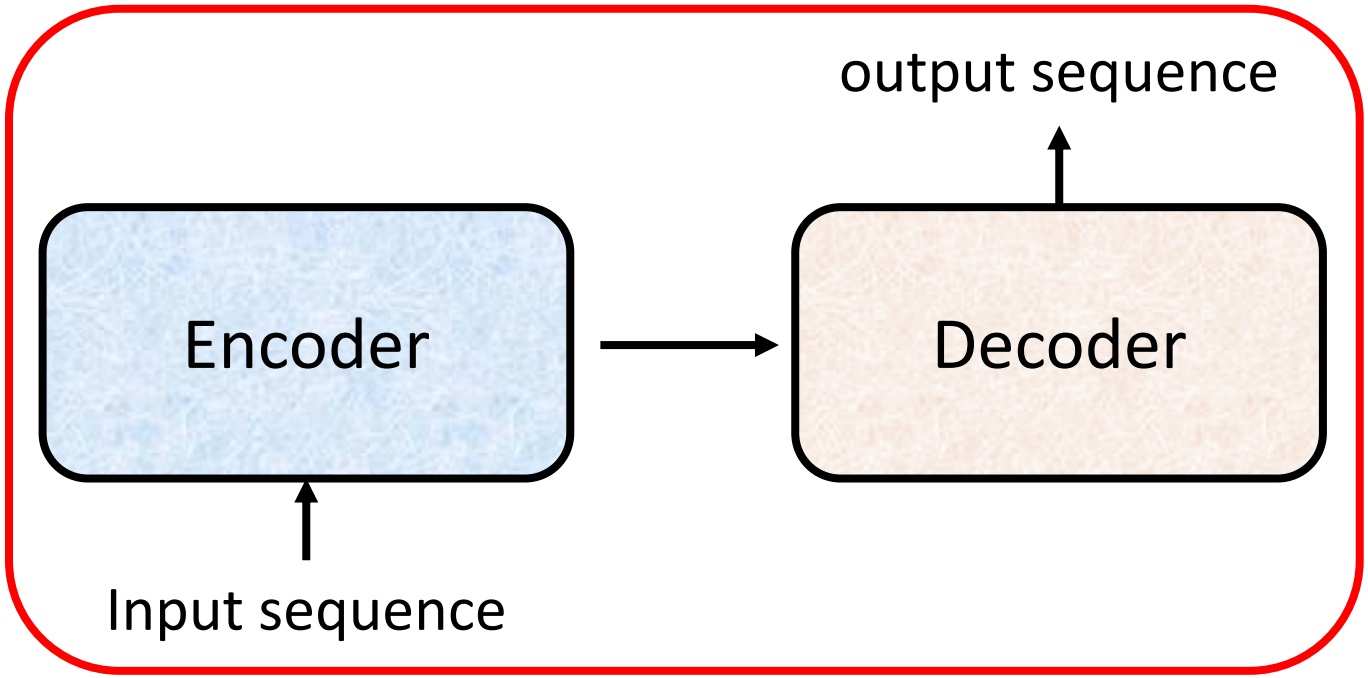


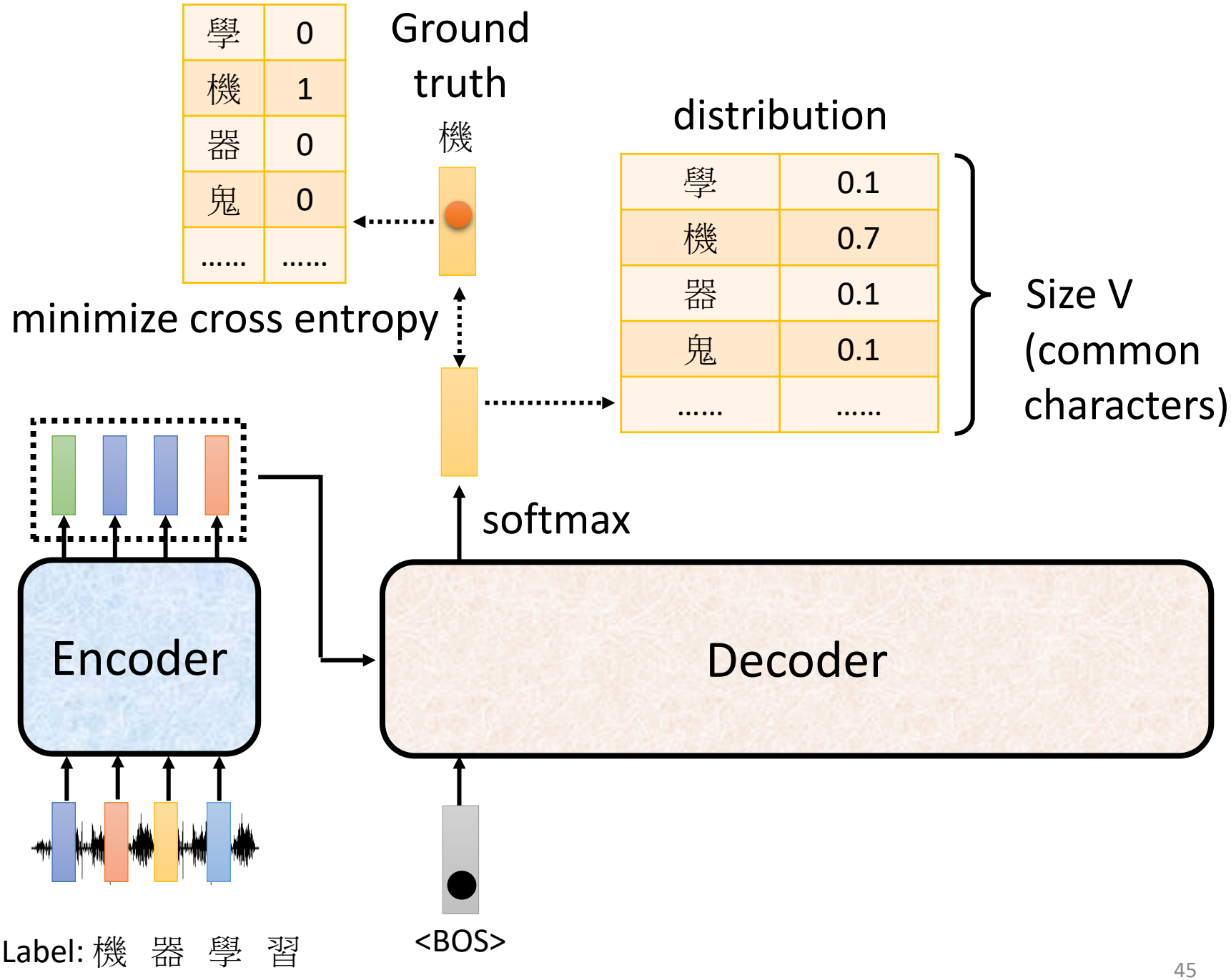
(a) Conventional Transformer



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

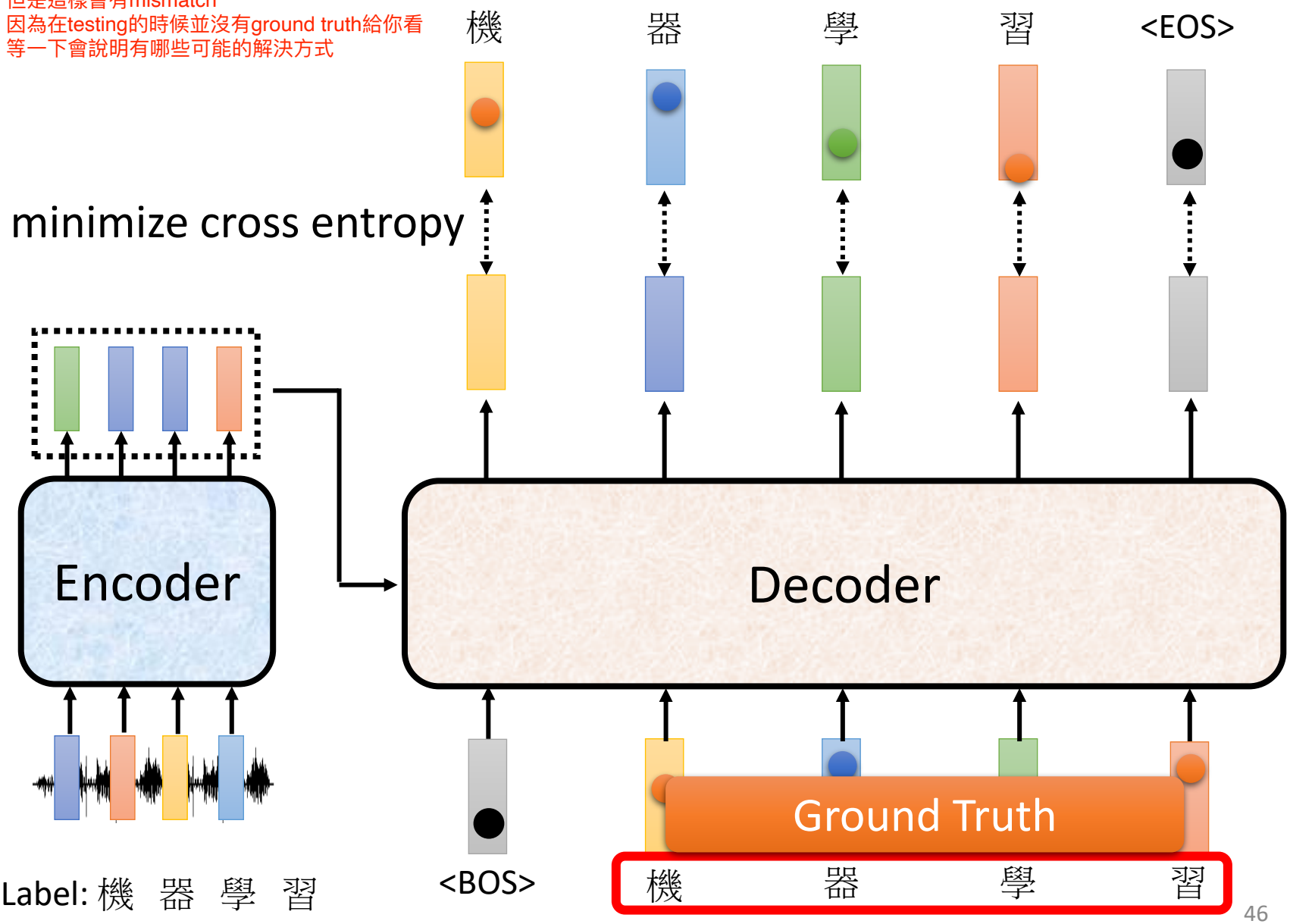
Training



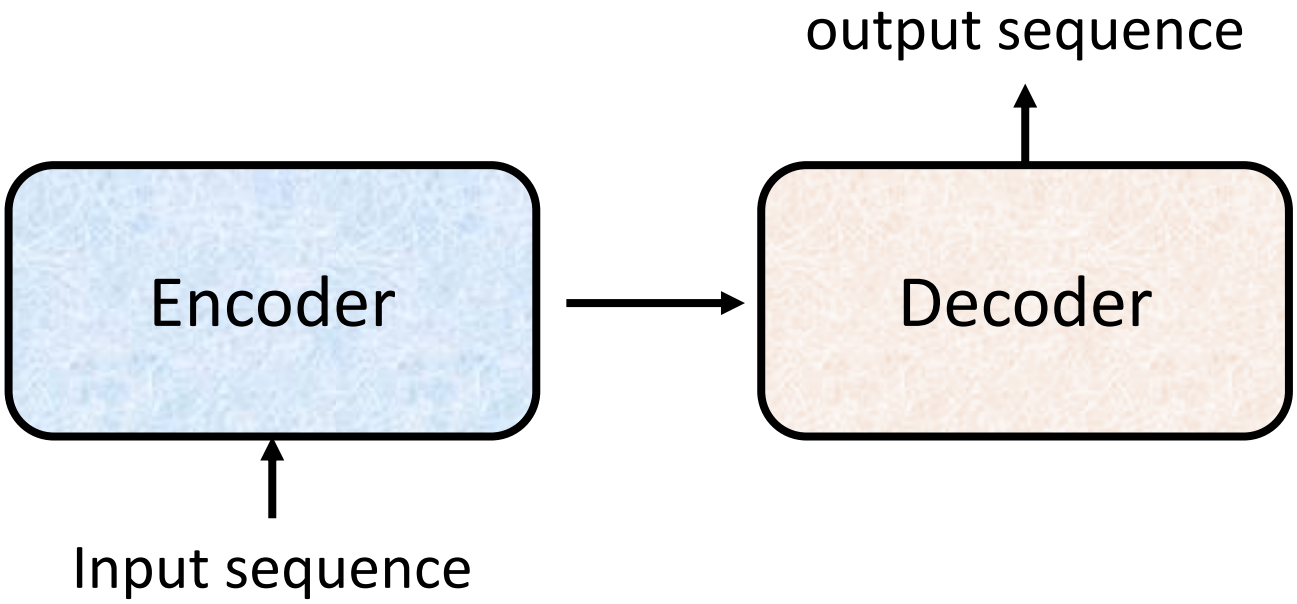

學 0
機 1
器 0
鬼 0
....
Ground truth
機
distribution
學 0.1
機 0.7
器 0.1
鬼 0.1
....
Size V (common characters)
minimize cross entropy
softmax
Encoder
Decoder
<BOS>
Label: 機 器 學 習

Teacher Forcing: using the ground truth as input.

但是這樣會有mismatch
因為在testing的時候並沒有ground truth給你看
等一下會說明有哪些可能的解決方式



Tips

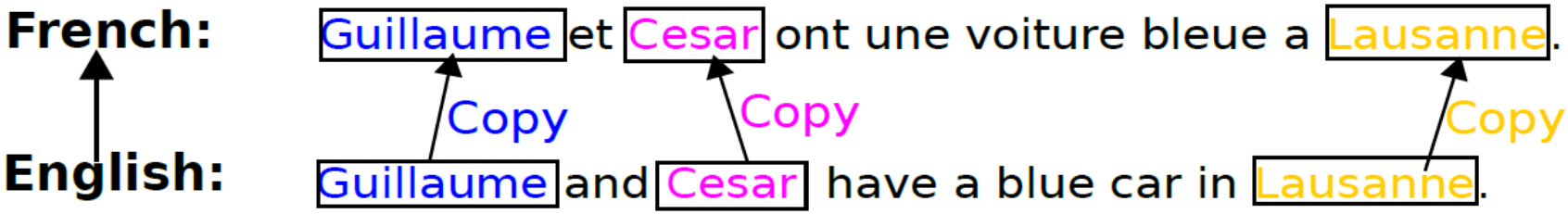


Copy Mechanism

很多時候機器並不需要自己寫出某個詞彙
因為那可能是專有名詞（根本沒有看過）
或是其實直接照抄input中某一段落就好

e.g. Chat-bot
User: 小傑不能用念能力了！
Machine: 不能用念能力是什麼意思？

Machine Translation



Chat-bot

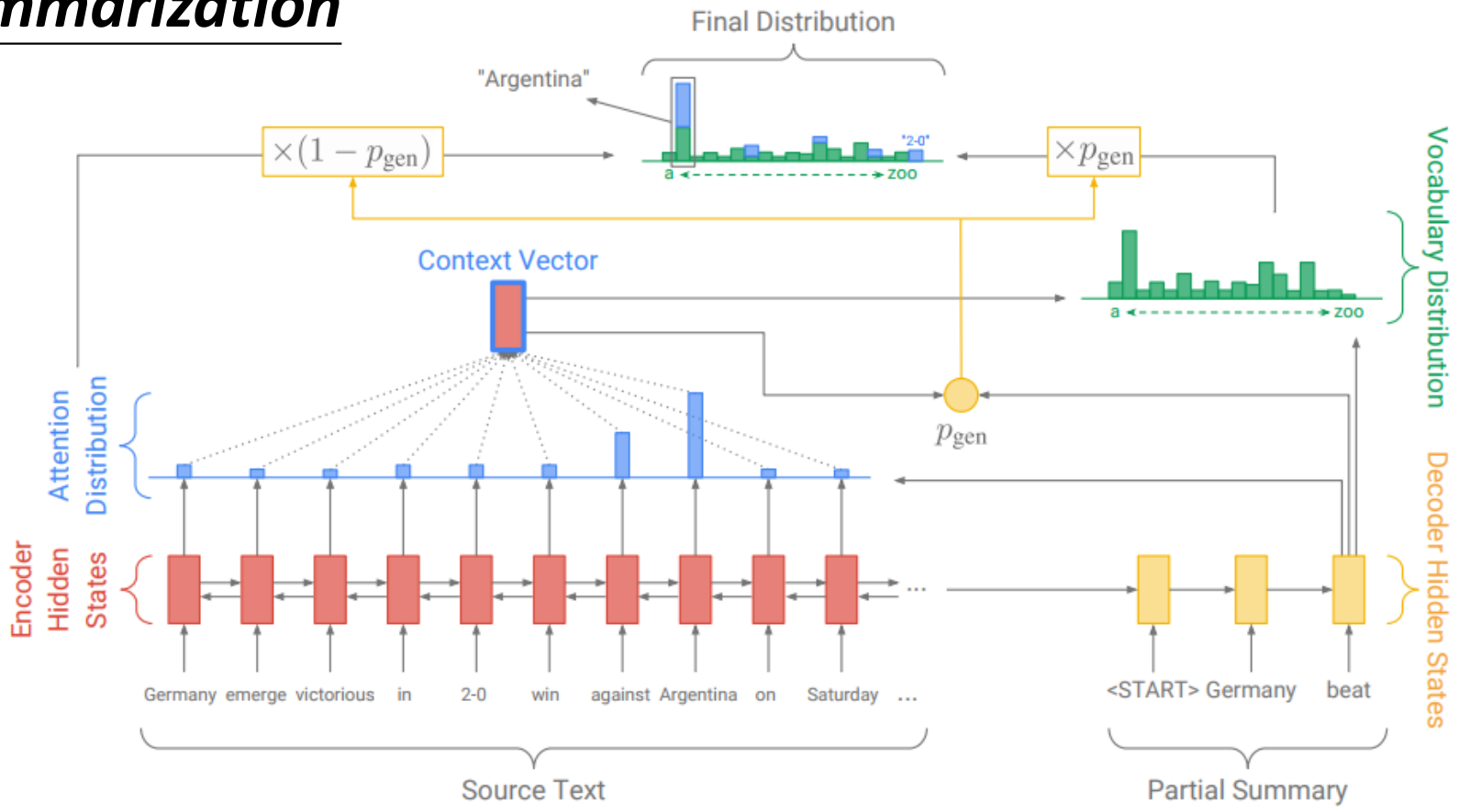
User: X寶你好，我是庫洛洛

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

這是在講copy mechanism的影片

Incorporating Copying Mechanism in Sequence-to-Sequence Learning

https://arxiv.org/abs/1603.06393

Guided Attention

TTS as exmple

而guided attention就是強迫machine要把所有的input都看過一次（避免漏看的問題）
但是也只有在語音合成或是語音辨識的時候會比較在乎
例如chatbot有沒有真的把input看完其實我們也不知道


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

前面四個都講得很好
但是最後一個卻只有講出「發」這個音
可能是因為training set中短句子的data很少
但是這種情況其實不常發生


發財 (Missing an input character!)

這兩個是關於這項技術的關鍵詞彙

Monotonic Attention

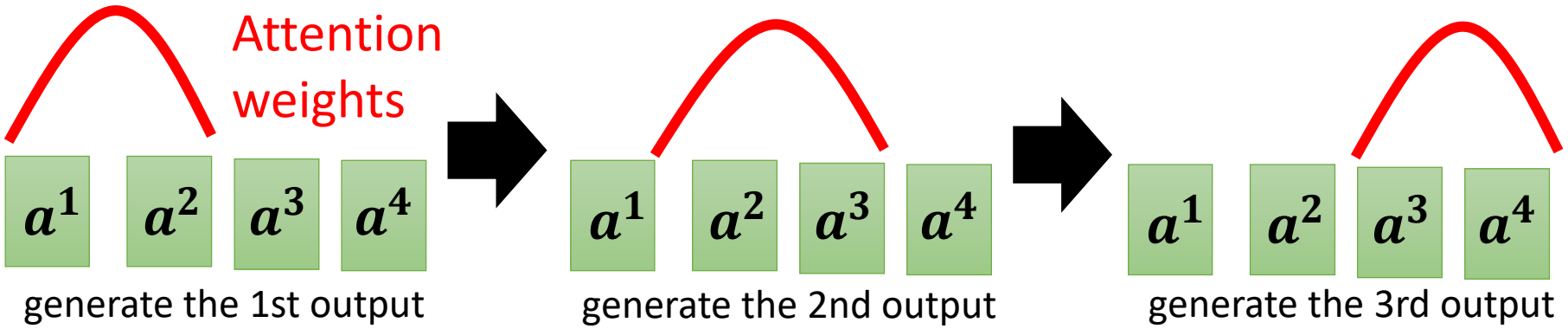
Location-aware attention

Guided Attention

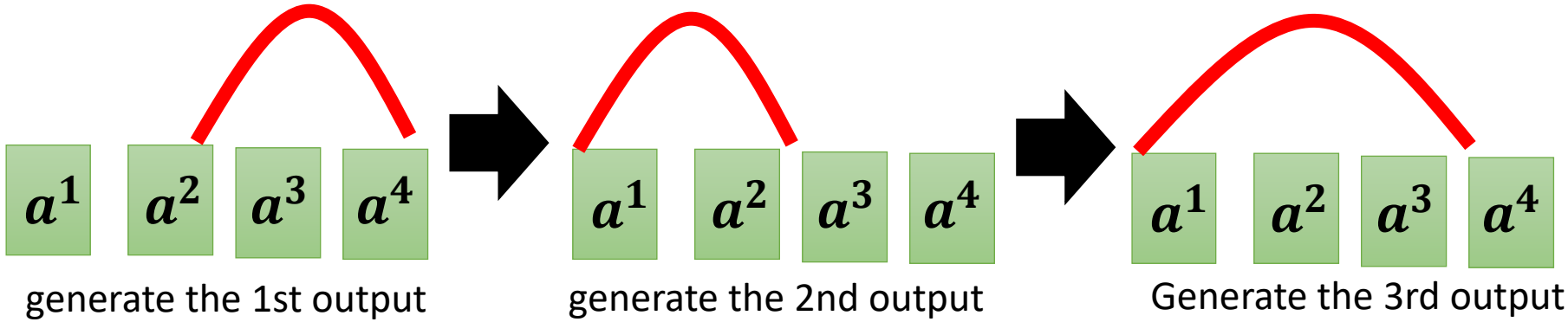
In some tasks, input and output are monotonically aligned.

For example, speech recognition, TTS, etc.

例如在做語音辨識的時候，因該要從左看到右



但若在語音辨識時，亂看一通，就很可能是有哪裡出錯了



Something wrong!

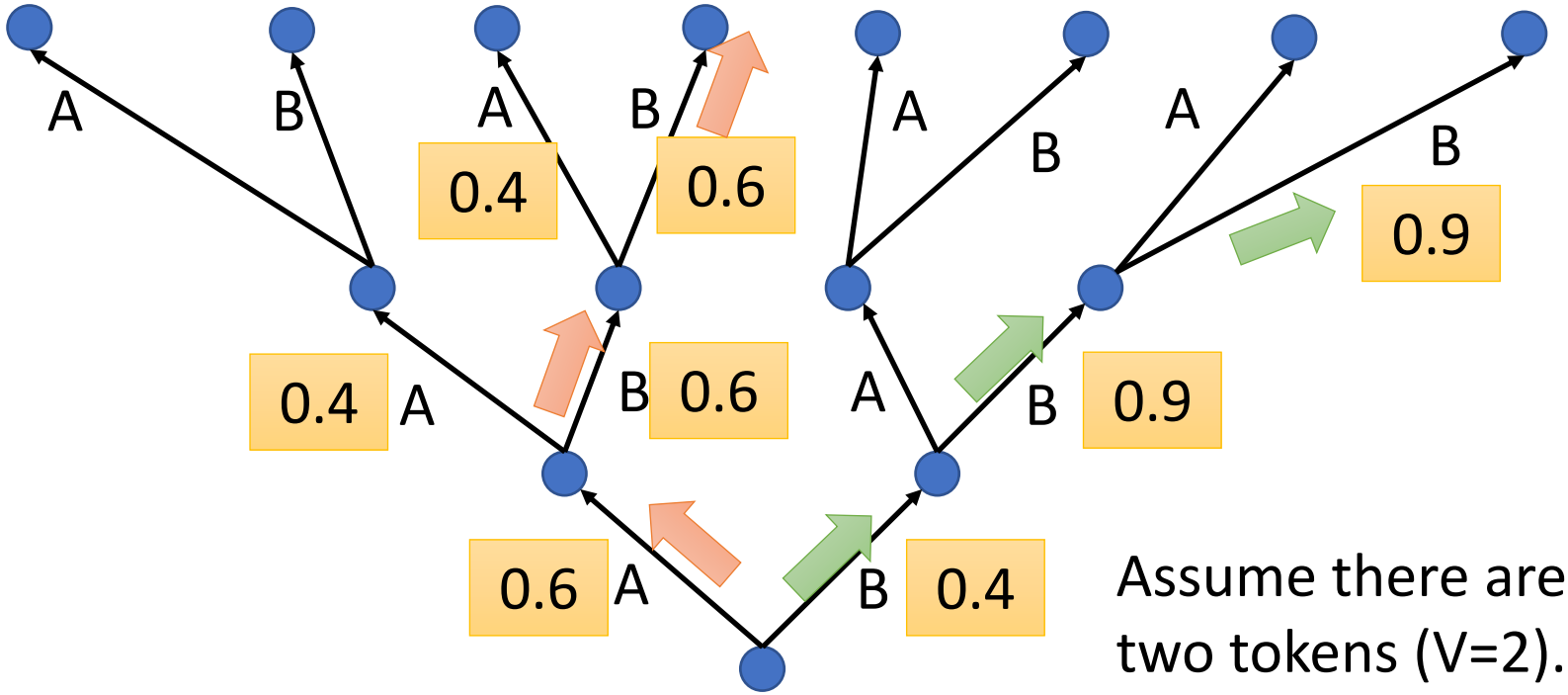
在做decoding的時後，每一次都選最好的那一個詞彙並不一定是最好的演算法
而beam search就是一個在找「整體輸出」比較好的演算法
但其實beam search有時候有用，有時候沒用

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

通常有固定答案的
beam search就比較有幫助
e.g. 語音辨識

但若沒有固定答案的
beam search就比較沒有幫助
e.g. chat bot

這篇paper在說beam search其實不一定好

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.

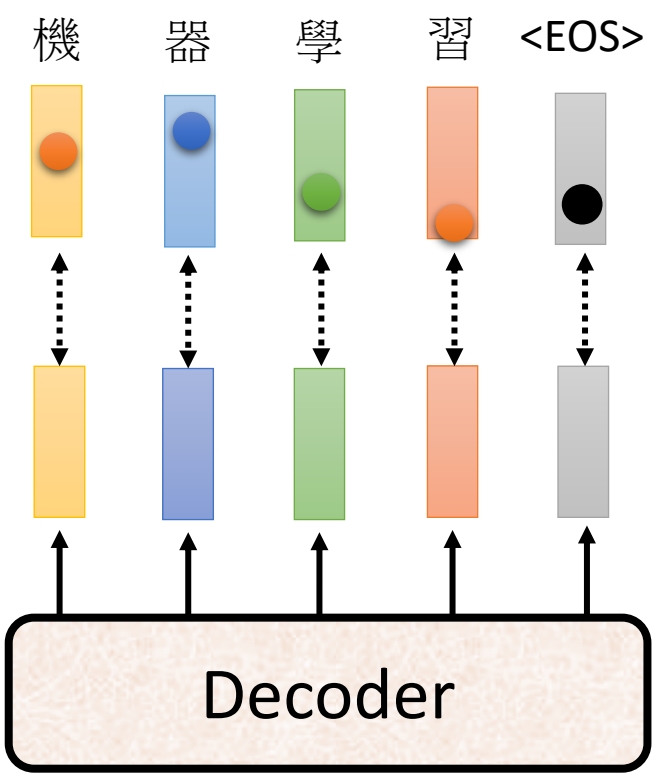
其實語音合成也是屬於沒有固定答案的

有一個googler說
語音合成的model在做testing的時候
也要加入一些雜訊，這樣結果才會好
（機器覺得最好的，人類並不一定覺得是最好的）

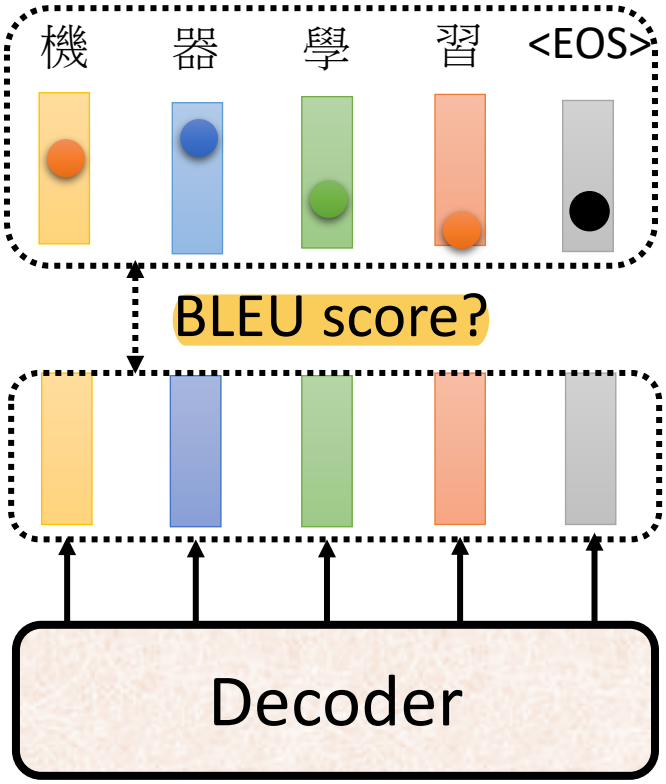
Accept that nothing is perfect. True beauty lies in the cracks of imperfection. ☺

Optimizing Evaluation Metrics?

訓練的時候其實是在minimize cross entropy



但我們在testing上的loss是用BLEU score



但minimize cross entropy並不代表BLEU score也會跟著降低
validation set應該使用BLEU score來挑model

那為什麼不用BLEU score來進行訓練？因為不能微分

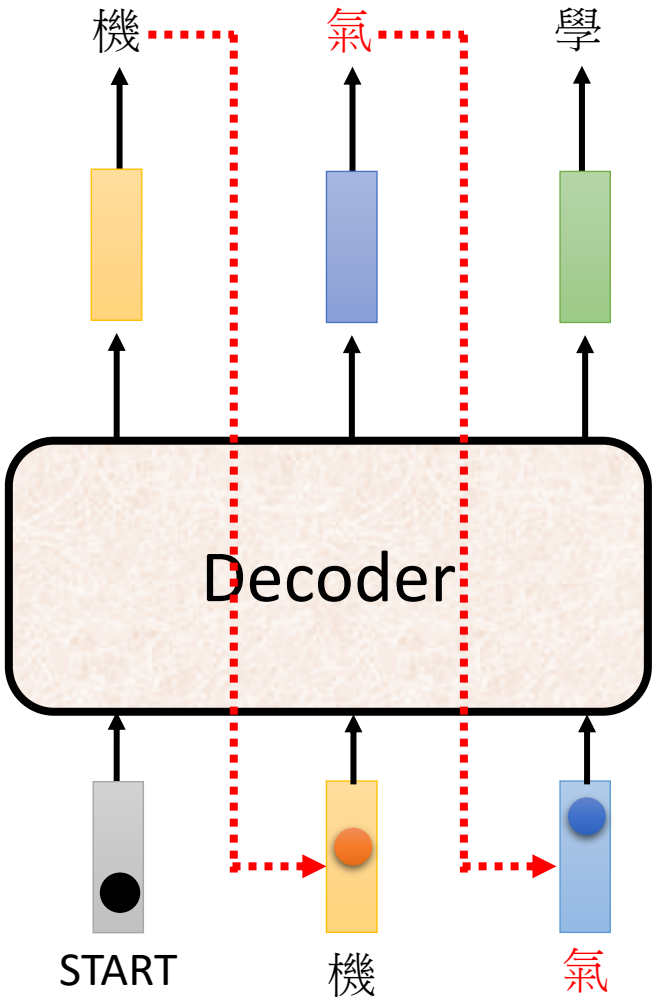
How to do the optimization?

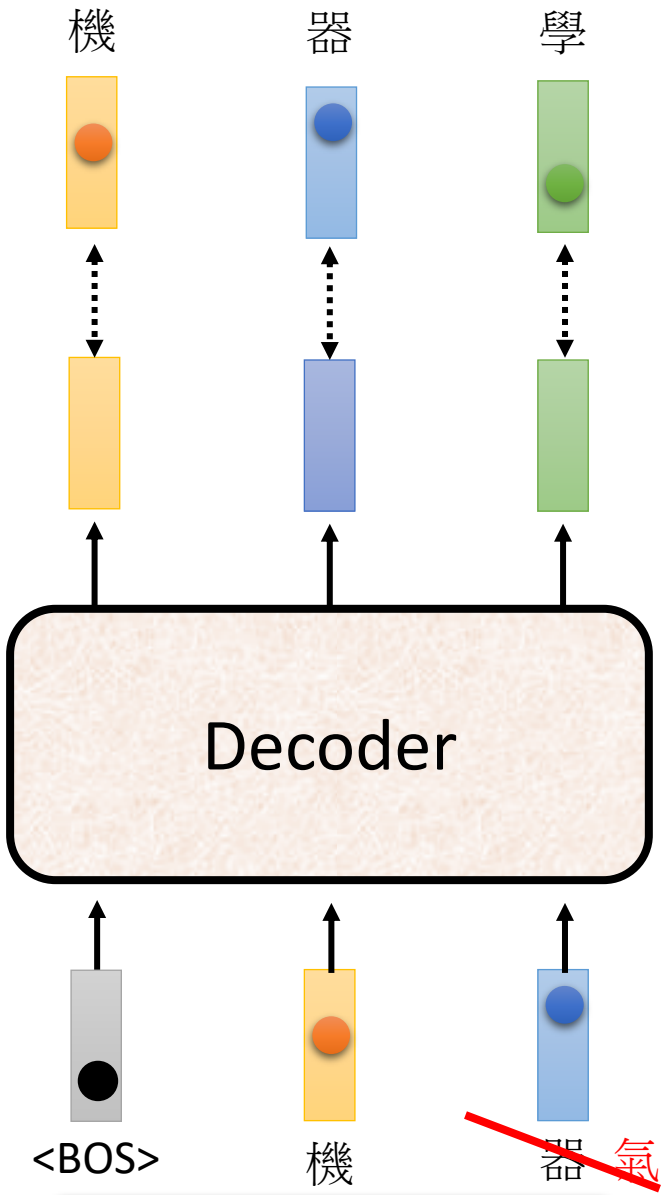
When you don't know how to optimize, just use reinforcement learning (RL)! https://arxiv.org/abs/1511.06732

有人直接用RL硬train

There is a mismatch! ☹

exposure bias





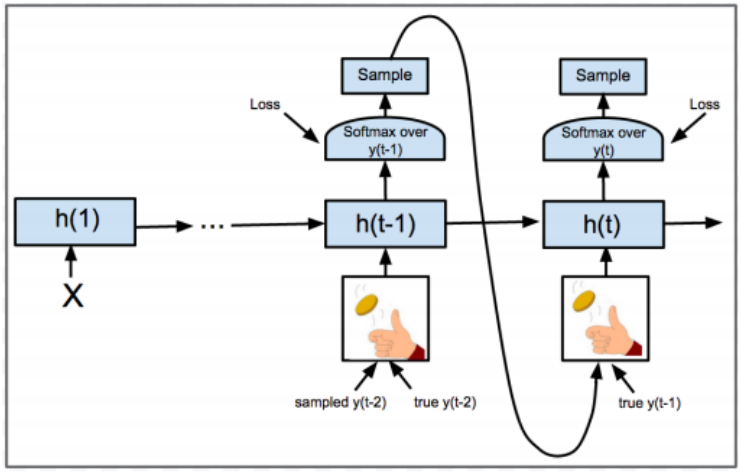
Ground Truth

在訓練的過程中，不要只給機器看正確的答案再去預測
偶爾也給他看自己預測出來的答案（錯誤的答案）
這樣可以訓練得更好，這個技術叫做schedule sampling

Scheduled Sampling

- Original Scheduled Sampling LSTM時期的schedule sampling

  https://arxiv.org/abs/1506.03099



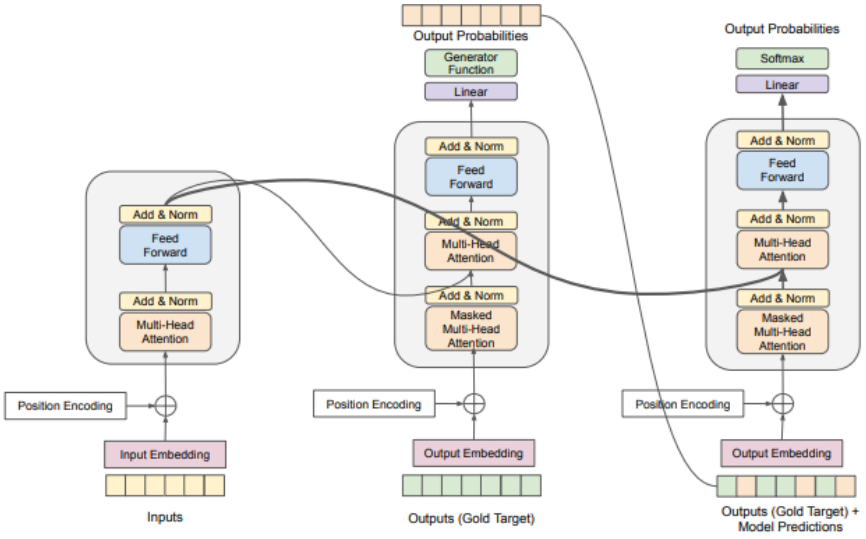
- Scheduled Sampling for Transformer

  https://arxiv.org/abs/1906.07651

  LSTM的schedue samping會遇到平行化問題
  這篇幫他做點優化

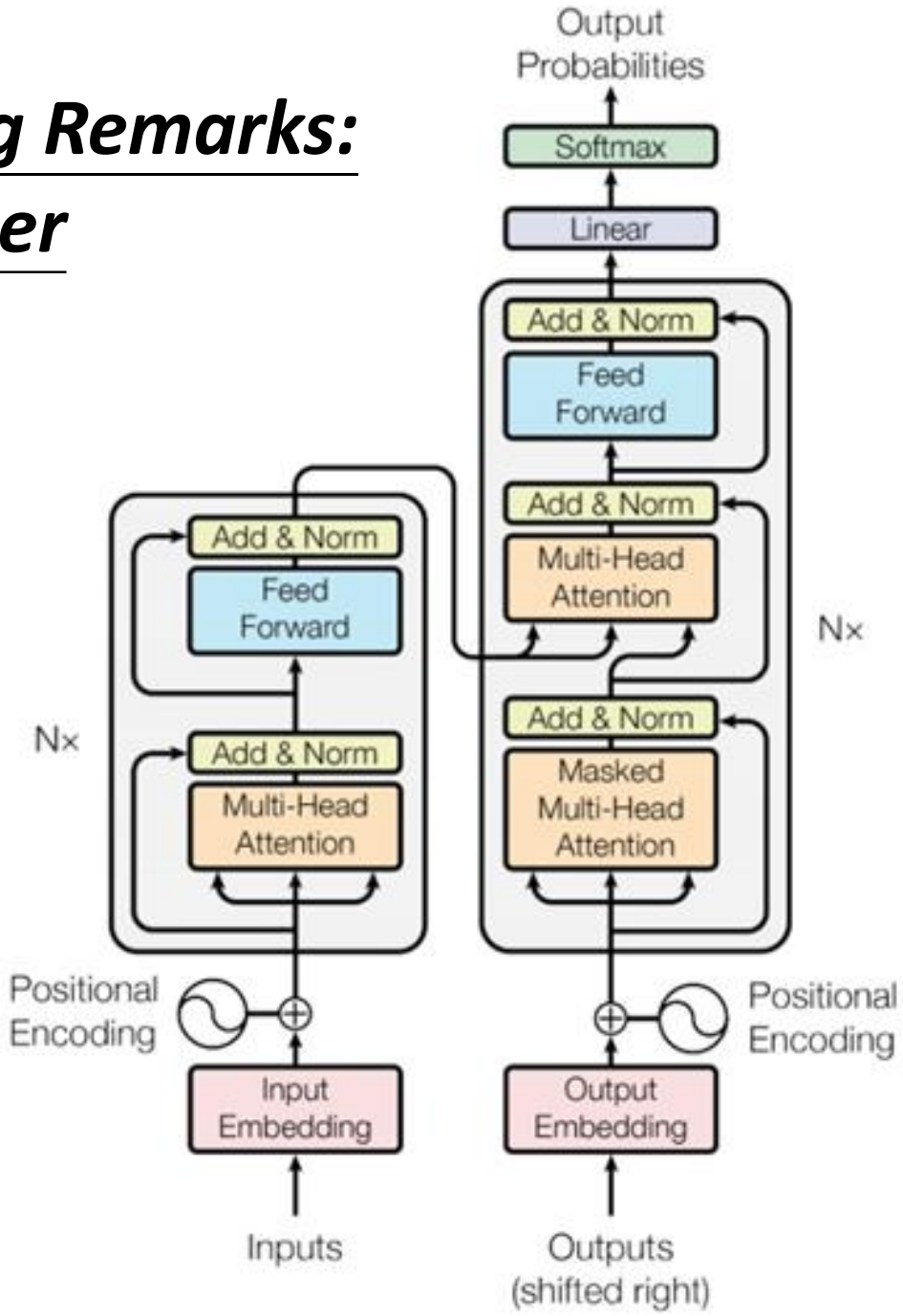
- Parallel Scheduled Sampling

  https://arxiv.org/abs/1906.04331



Schedule Sampling

Concluding Remarks: Transformer




Q&A