Transformer

李宏毅

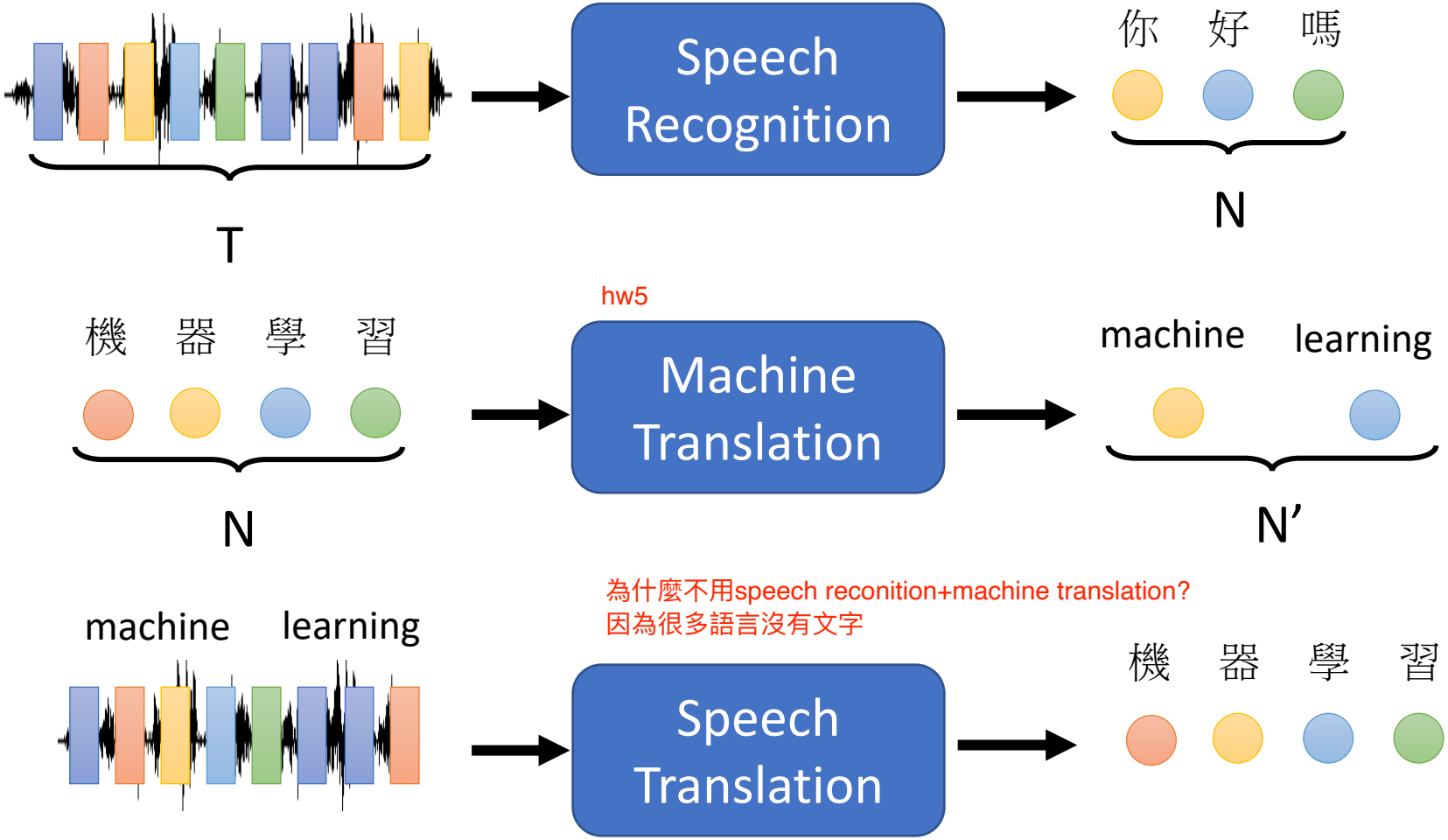
Hung-yi Lee

Sequence-to-sequence (Seq2seq)

hw2: input: N, output: N
hw4: input: N, output: 1
hw5: input: N, output: N'

Input a sequence, output a sequence

The output length is determined by model.



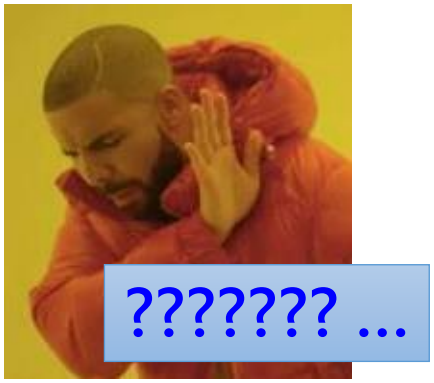
Language without text

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)

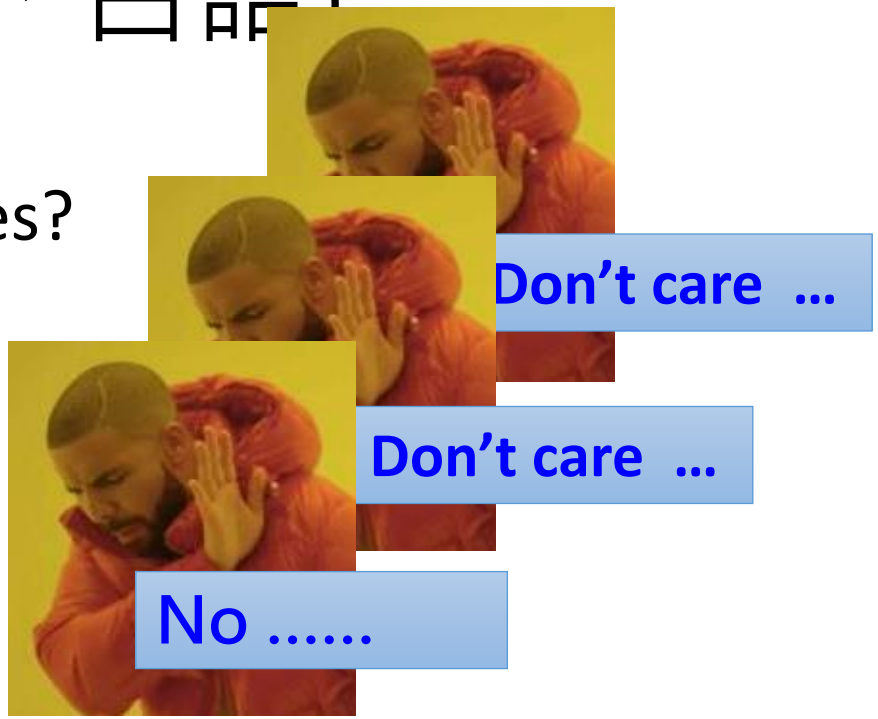
Using 1500 hours of data for training

李宏毅真的去抓了1500小時的data來做訓練

Hokkien (閩南語、台語)

但可能會遇到很多問題
例如背景音樂、音訊跟字幕沒對起來
但都先不要管它

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



或許可以用台羅拼音來當作中介
幫助model學習
但這也先不管他
總之就是直接硬train一發

"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

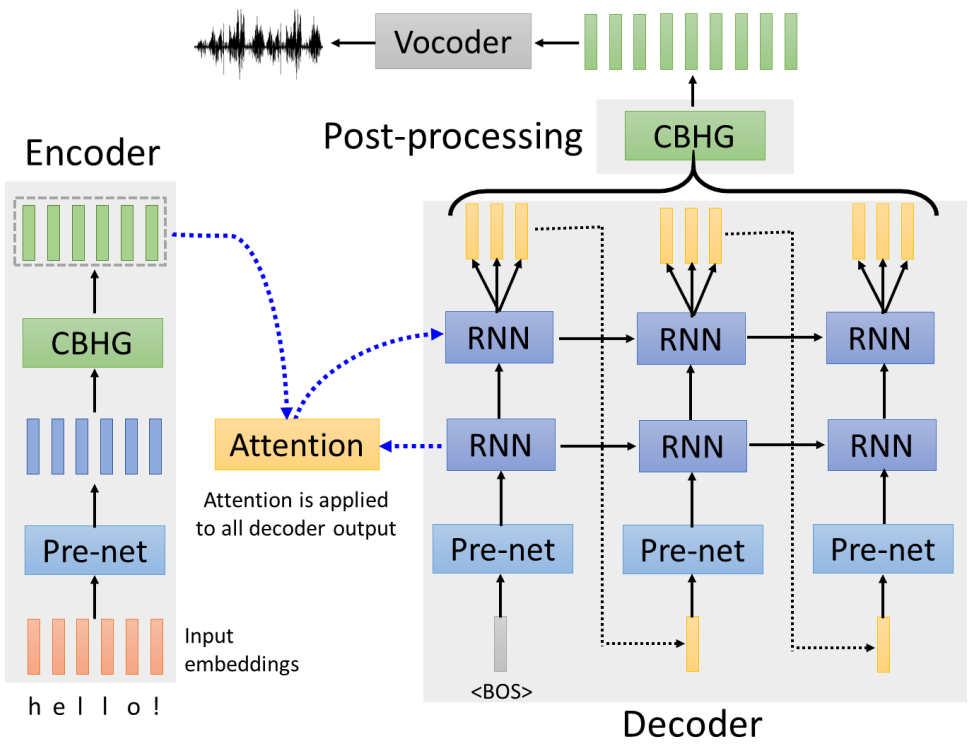
上面兩個是model有學到的
下面兩個是model學錯的
（第三個其實音很像）
（第四個是倒裝句）

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

語音合成：文字 -> 音訊
語音辨識：音訊 -> 文字



Taiwanese Speech Synthesis

Source of data: 台灣媠聲2.0

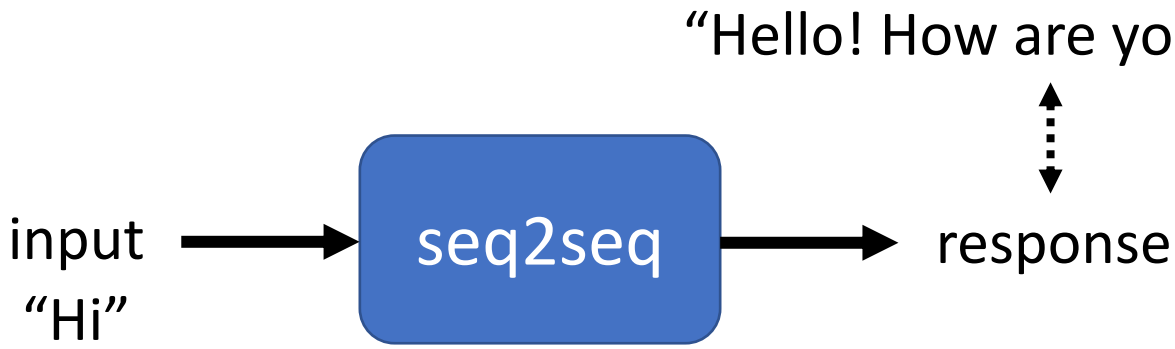
這個訓練是先將文字轉成台羅拼音（需要transformer），再轉乘台語

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

得到的結果表現還不錯

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**

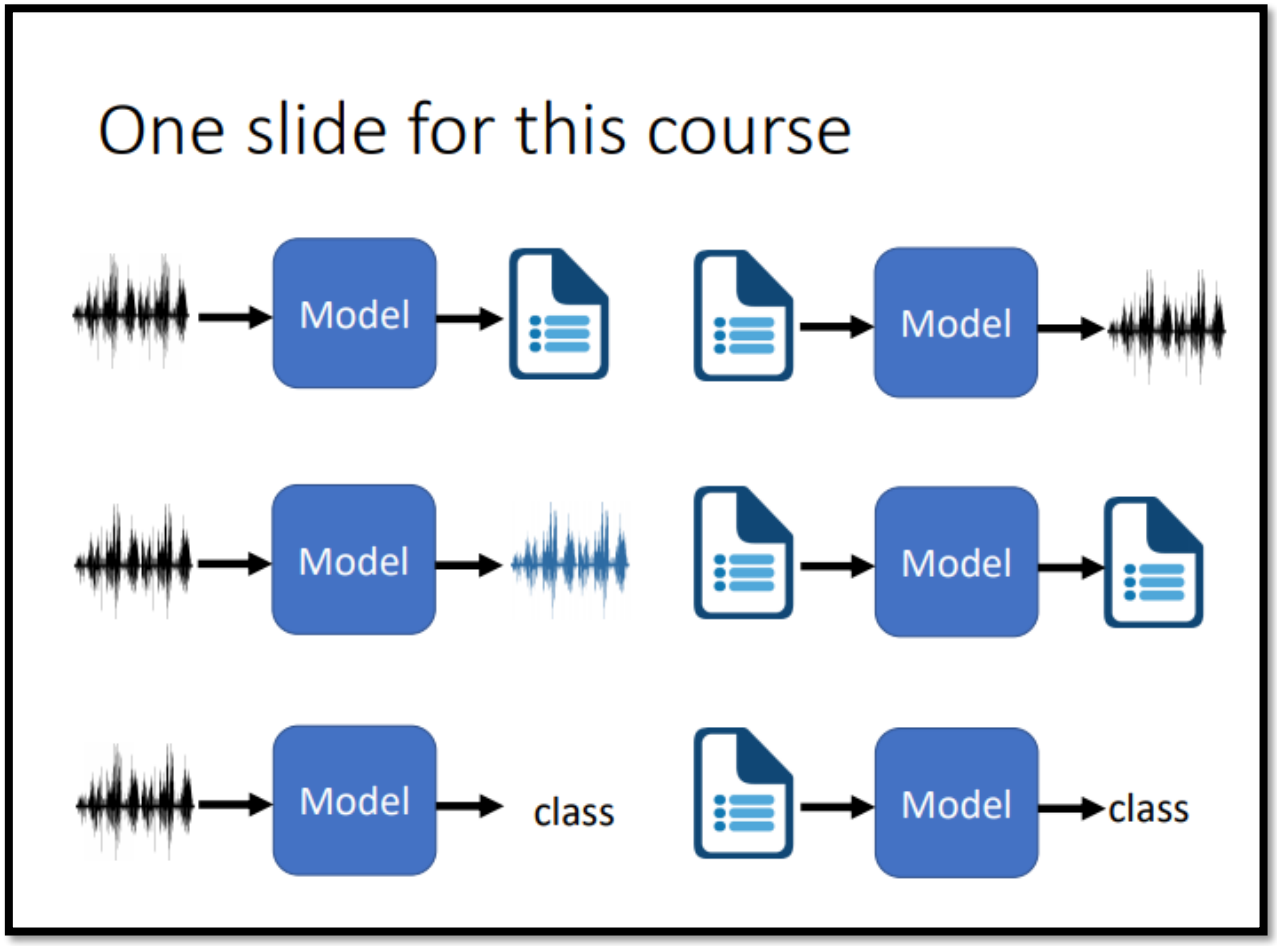
question, context → Seq2seq → answer

各式各樣nlp的問題
都可以化成Q&A
然後有人就用這個方式
去train了一個QA的model
可以去解決各式nlp問題

但往往客製化模型，
會得到更好的結果

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



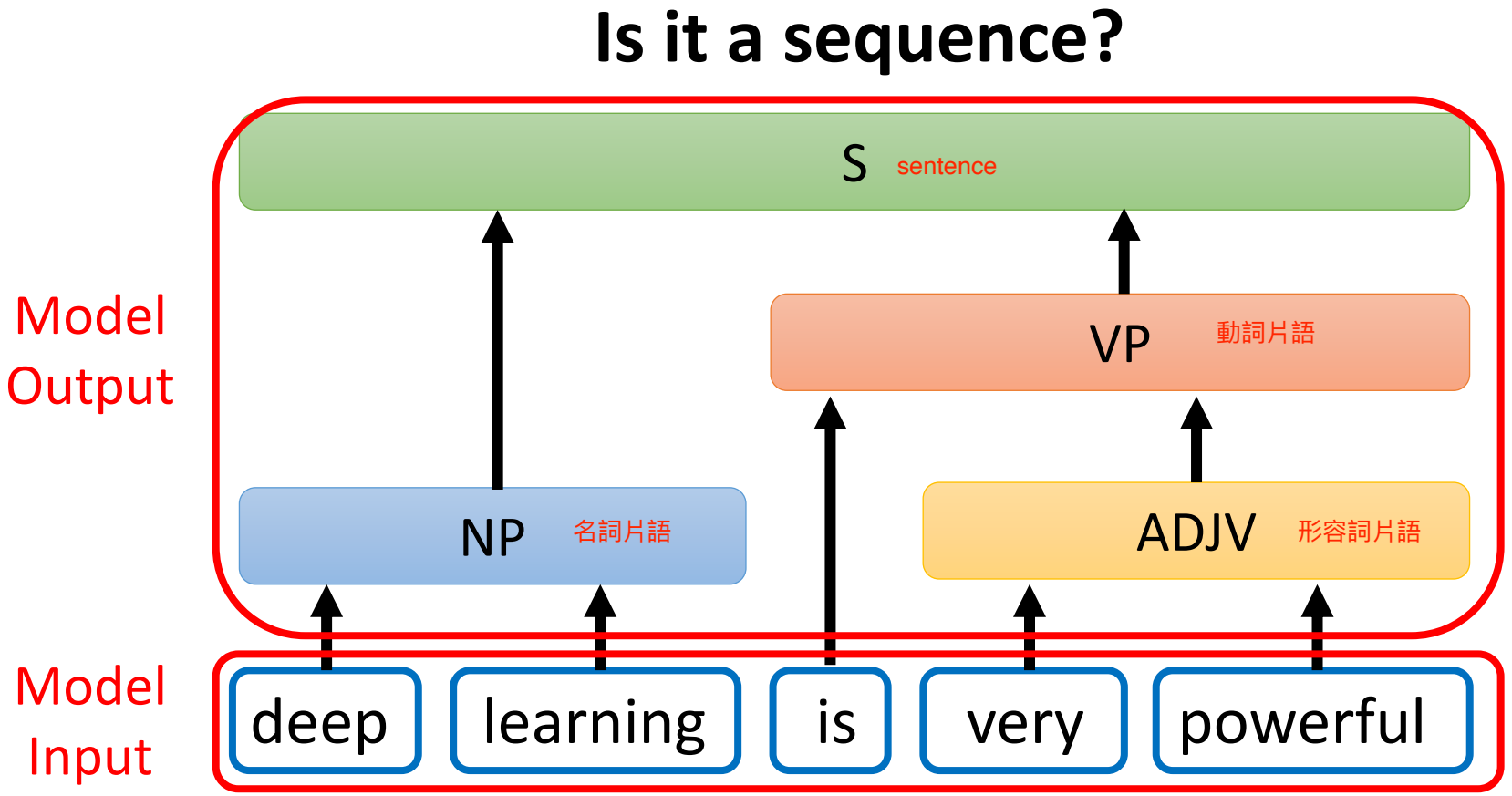
Source webpage: https://speech.ee.ntu.edu.tw/~hylee/dlhlp/2020-spring.html

nlp的相關課程

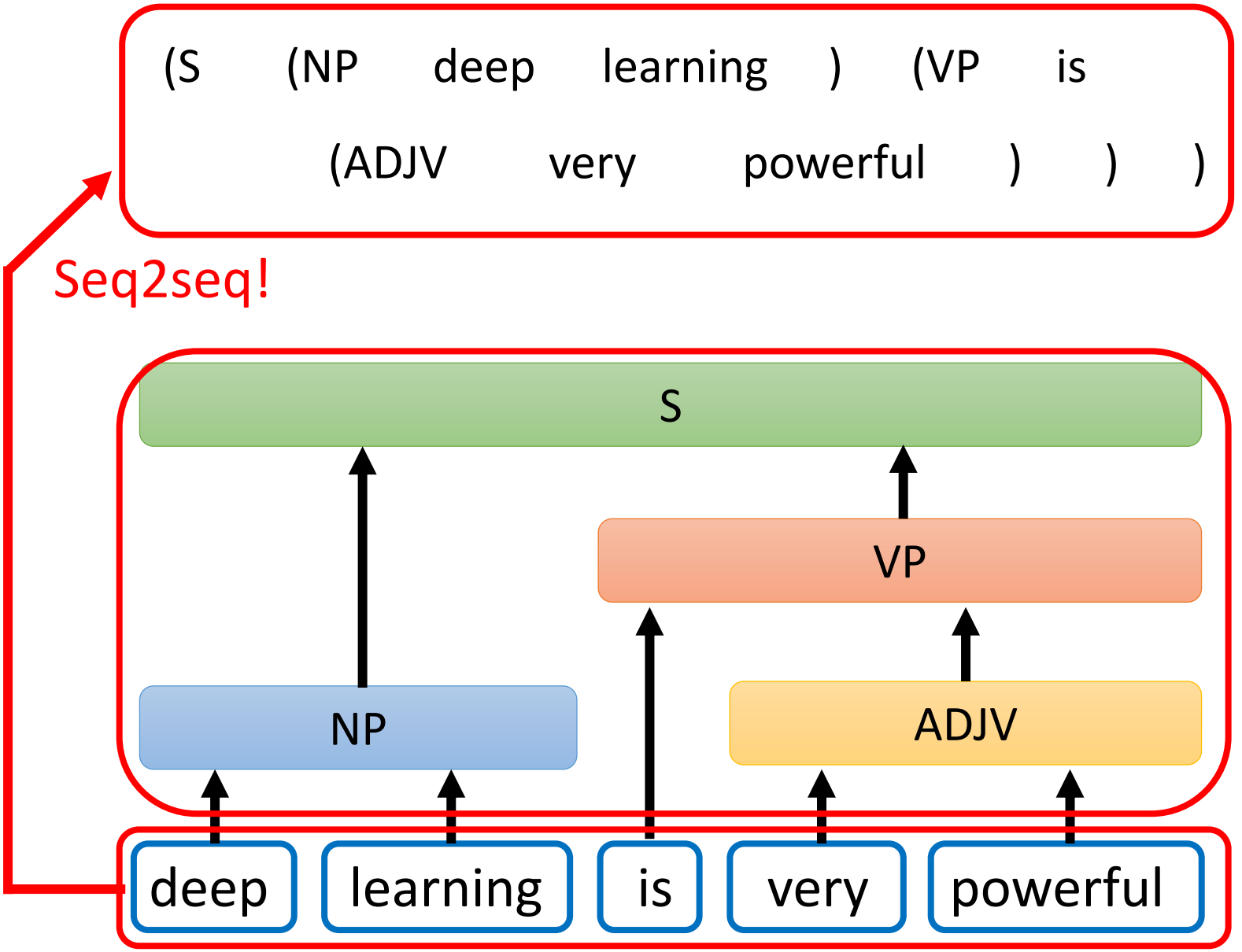
Seq2seq for Syntactic Parsing

文法剖析

若想要output出一個樹狀結構
其實也可以把它看成是一個sequence



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

在2014時
seq_to_seq大多只用在翻譯上
所以這篇文的名字才會叫做
grammar as s foreign language

https://arxiv.org/abs/1412.7449

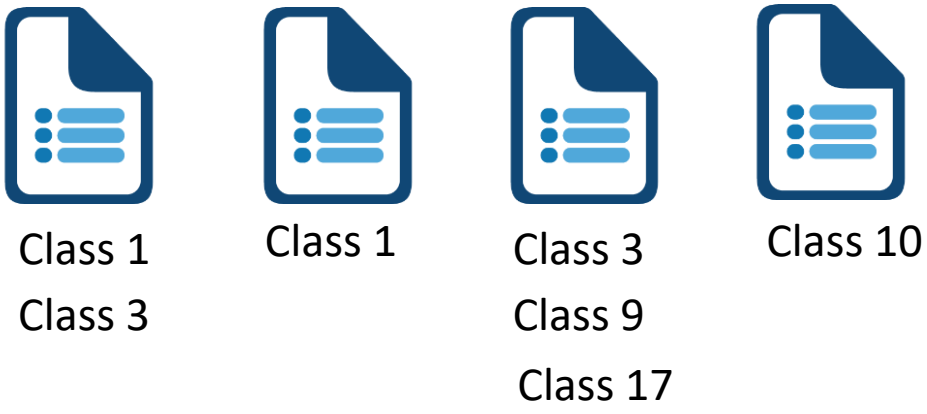
deep learning is very powerful

Seq2seq for Multi-label Classification

c.f. Multi-class Classification

multi-class classification: 機器要從數個class選擇一個class
multi-label classification: 同一個東西可以屬於多個class

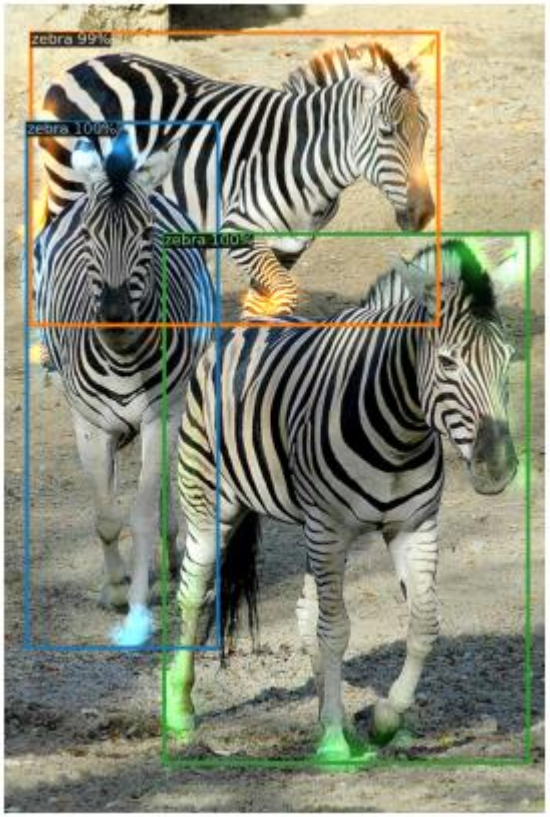
An object can belong to multiple classes.

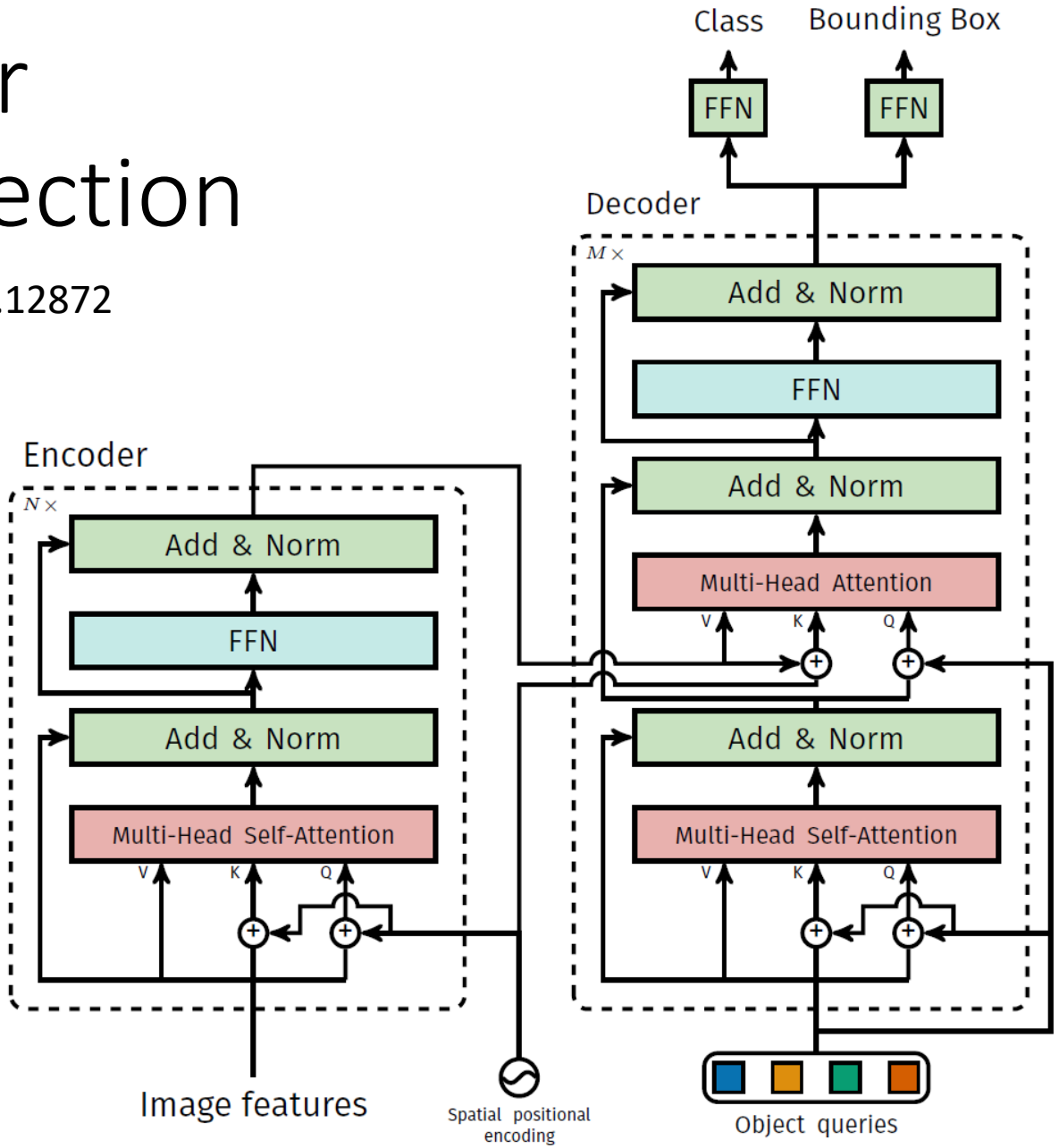


https://arxiv.org/abs/1909.03434
https://arxiv.org/abs/1707.05495

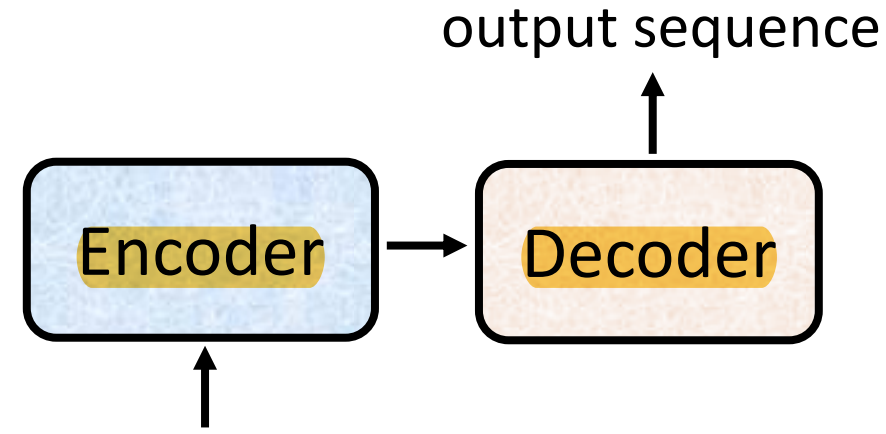
Seq2seq for Object Detection

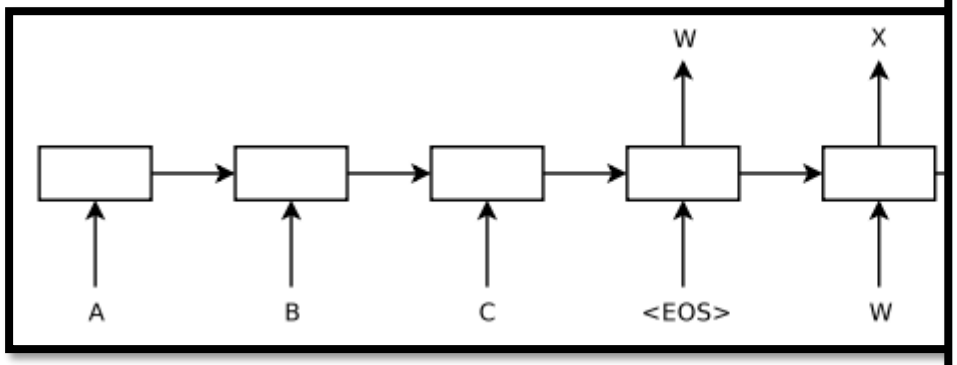
https://arxiv.org/abs/2005.12872





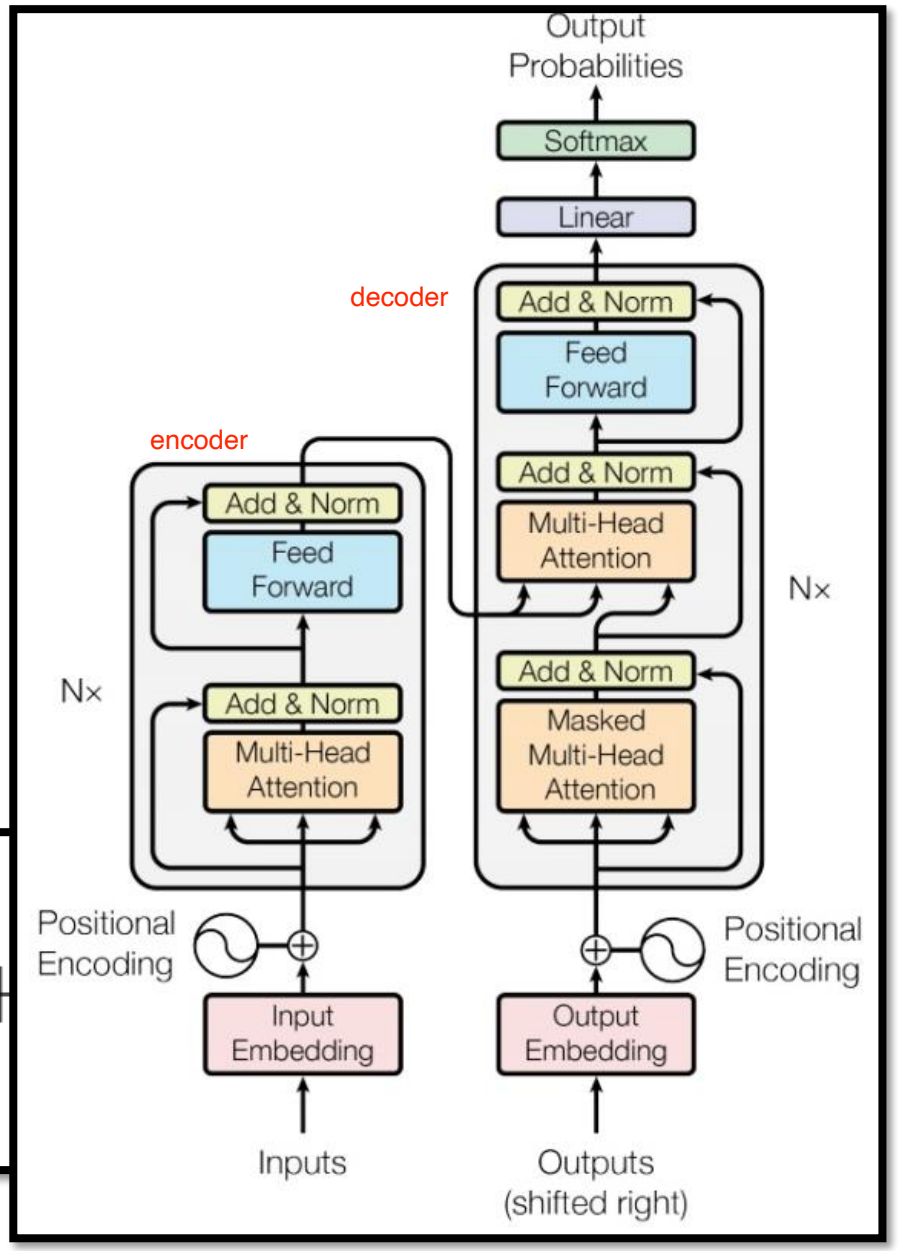
Seq2seq





Sequence to Sequence Learning with Neural Networks

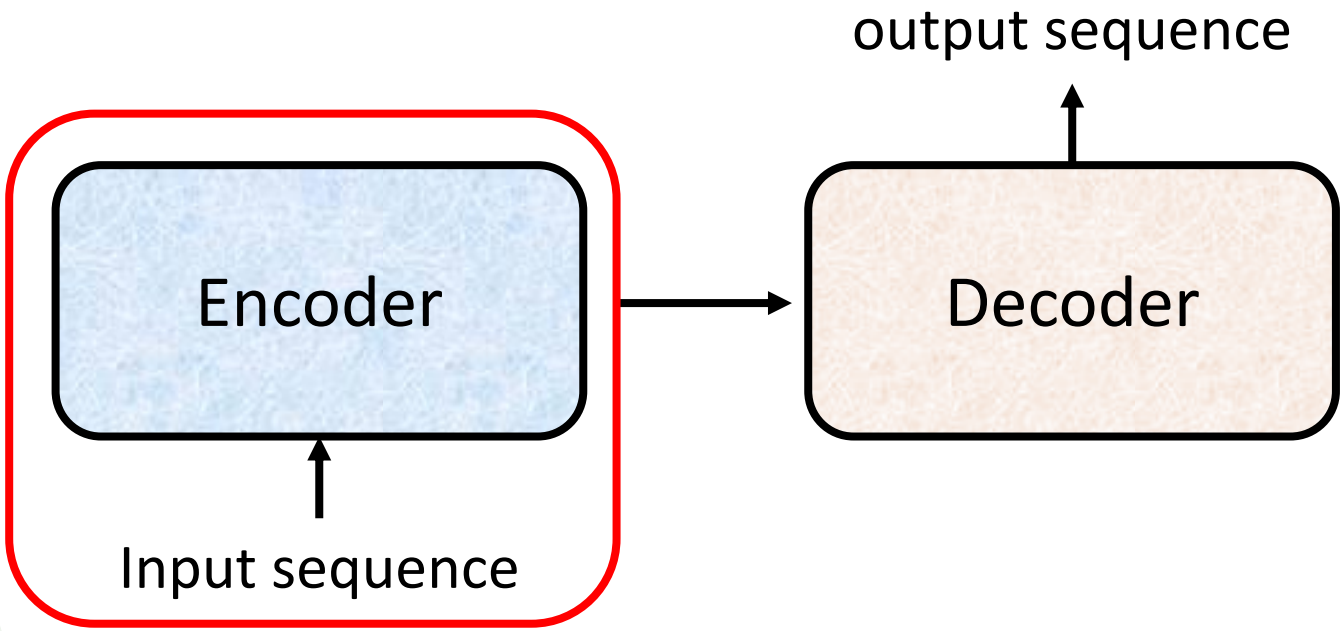
https://arxiv.org/abs/1409.3215



Transformer

https://arxiv.org/abs/1706.03762

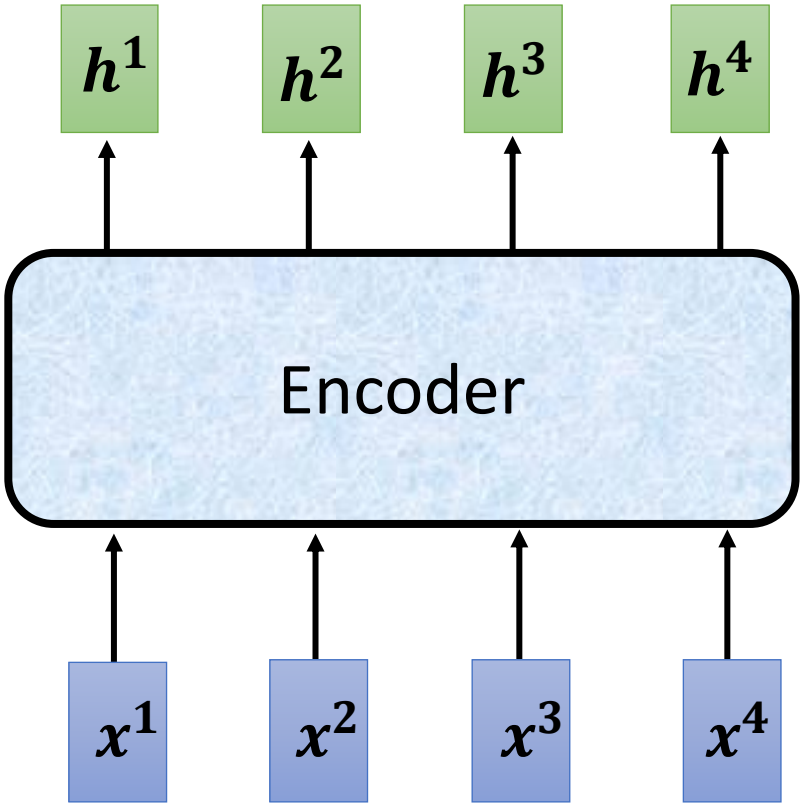
Encoder



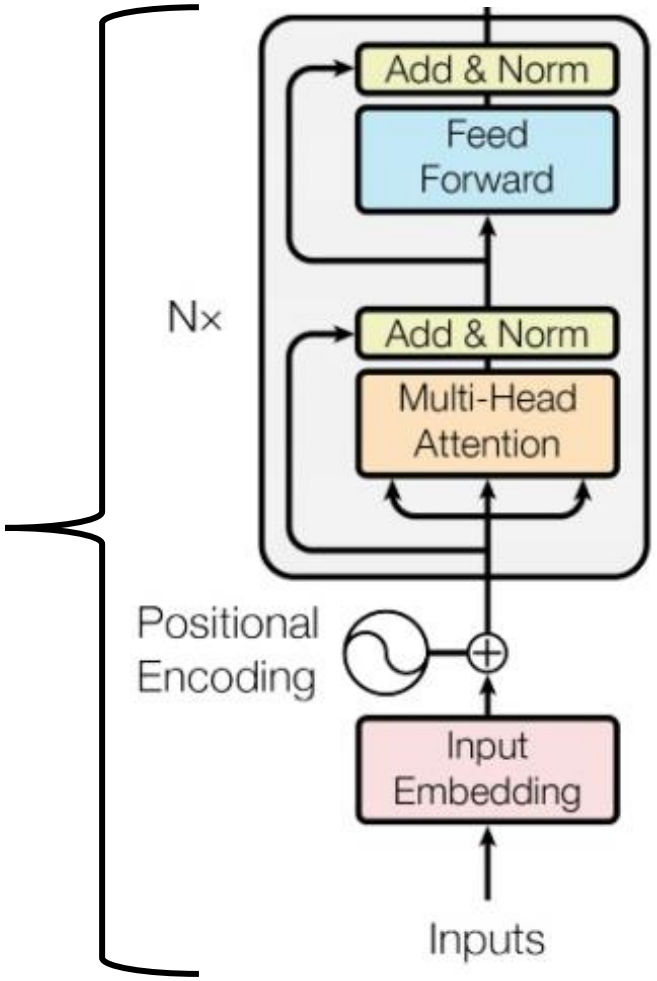
Encoder

input: 一排向量
output: 一排向量
有很多種模型可以做到這件事情
包括前一章節講述的self-attention還有cnn和rnn也都可以做到
而transformer的encoder就是self attention

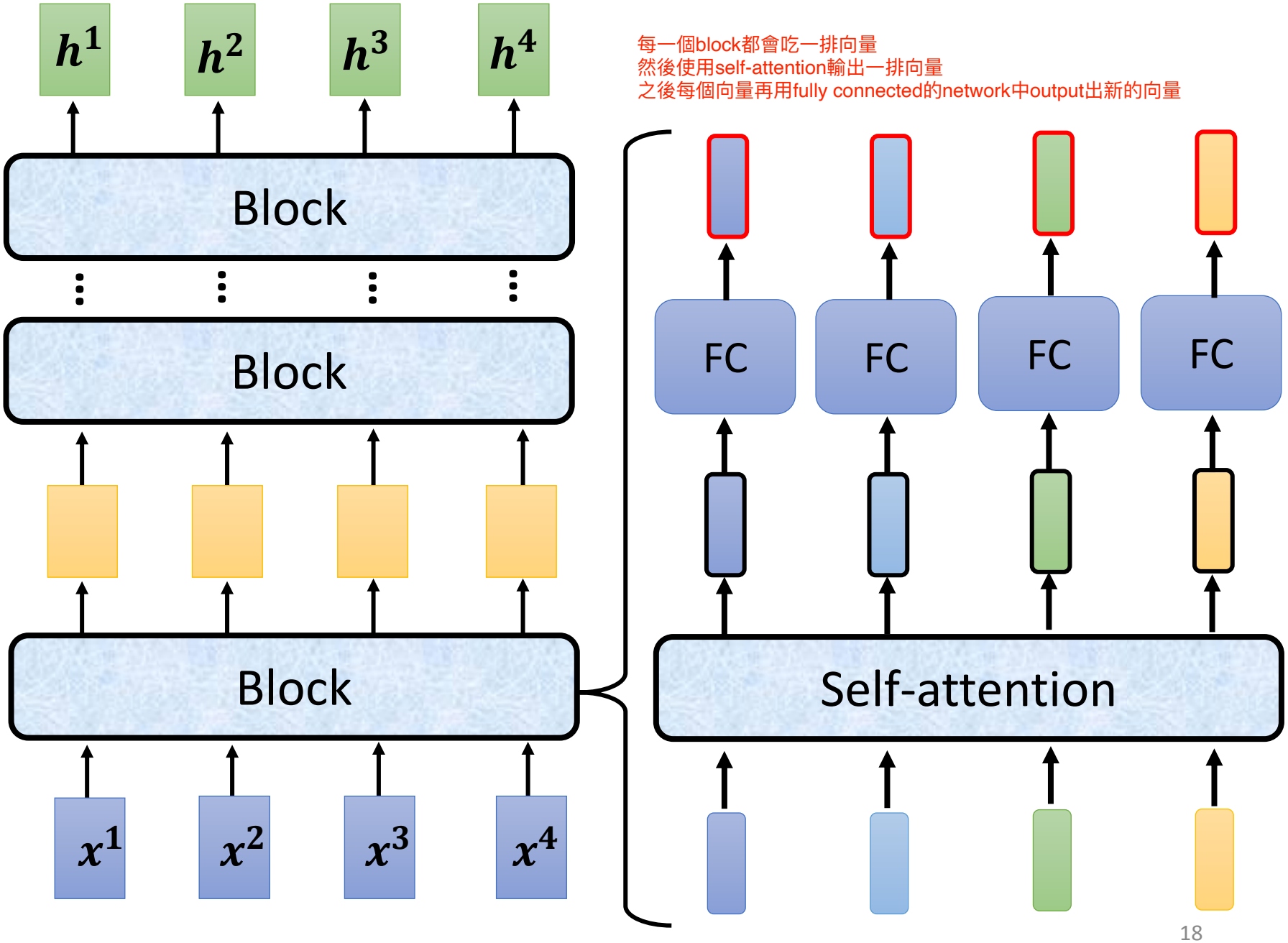
You can use **RNN** or **CNN**.



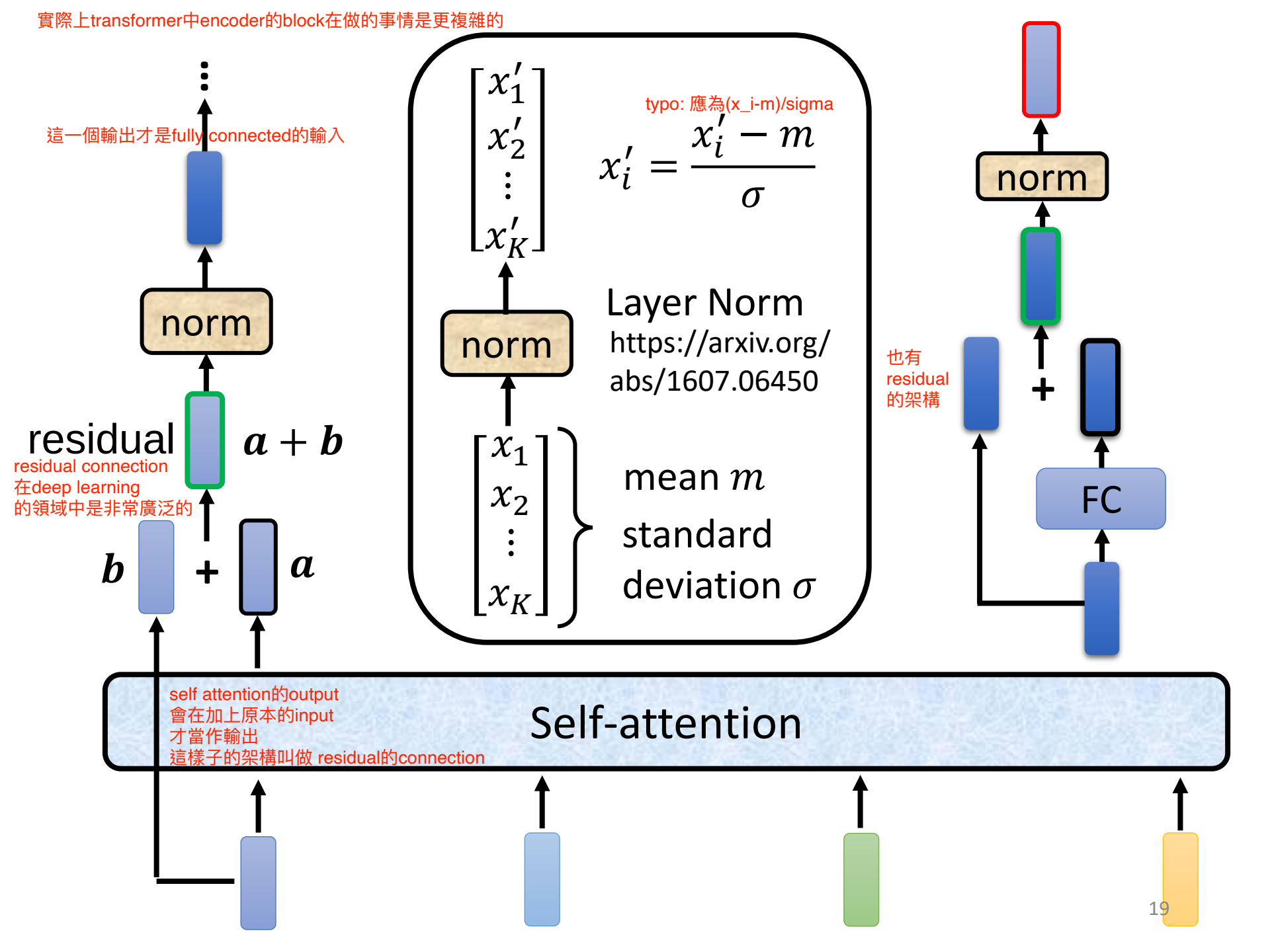
Transformer's Encoder



在encoder裡面會有很多個block，每一個block都是輸入一排向量，輸出一排向量



實際上transformer中encoder的block在做的事情是更複雜的



BERT

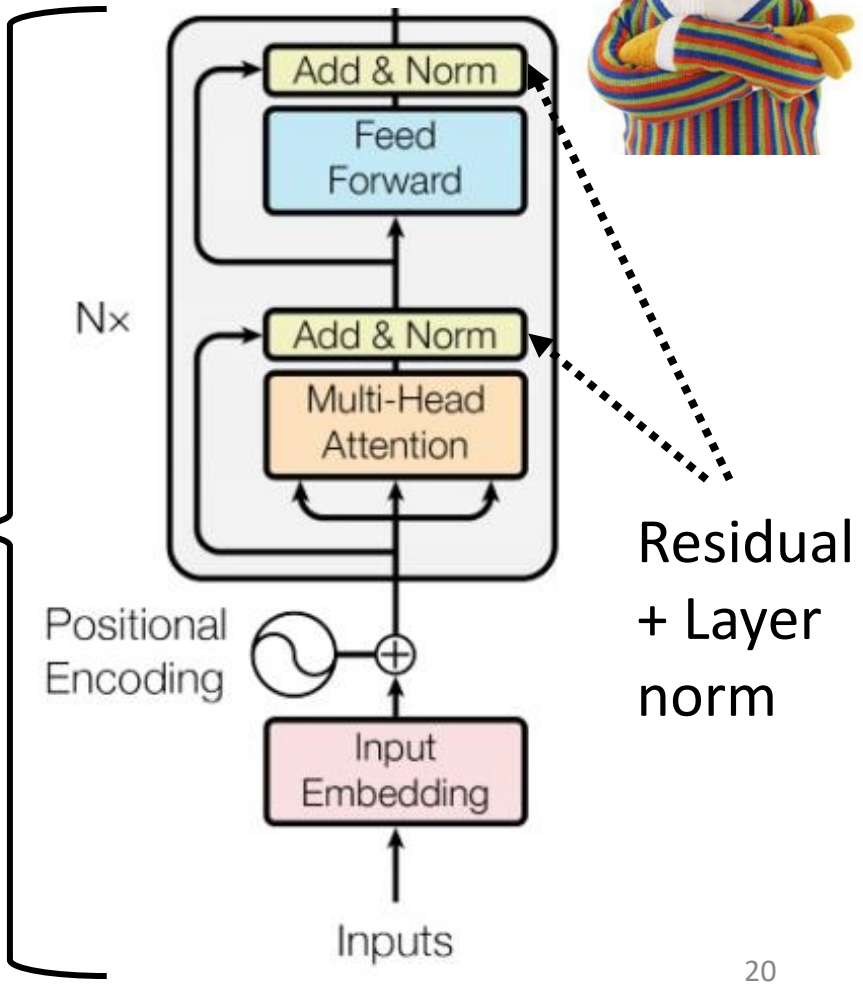
I use the **same** network architecture as **transformer encoder**.

這只是原始論文的設計
並不代表他是最好的
還有其他種設計encoder的方式

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

Encoder

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



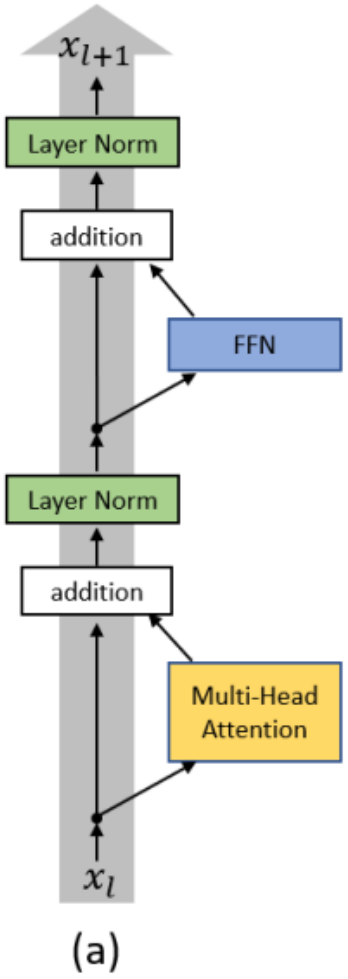
To learn more ……

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

這篇paper在討論為什麼使用layer norm而不是batch norm
並提出power norm

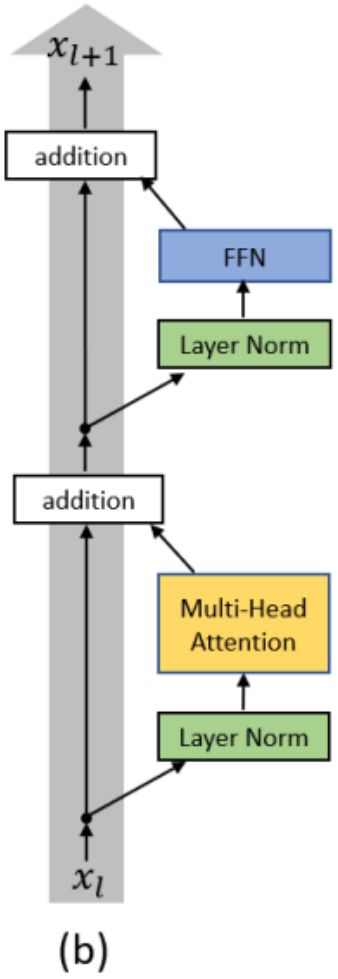
- PowerNorm: Rethinking Batch Normalization in Transformers
- https://arxiv.org/abs/2003.07845

原始的



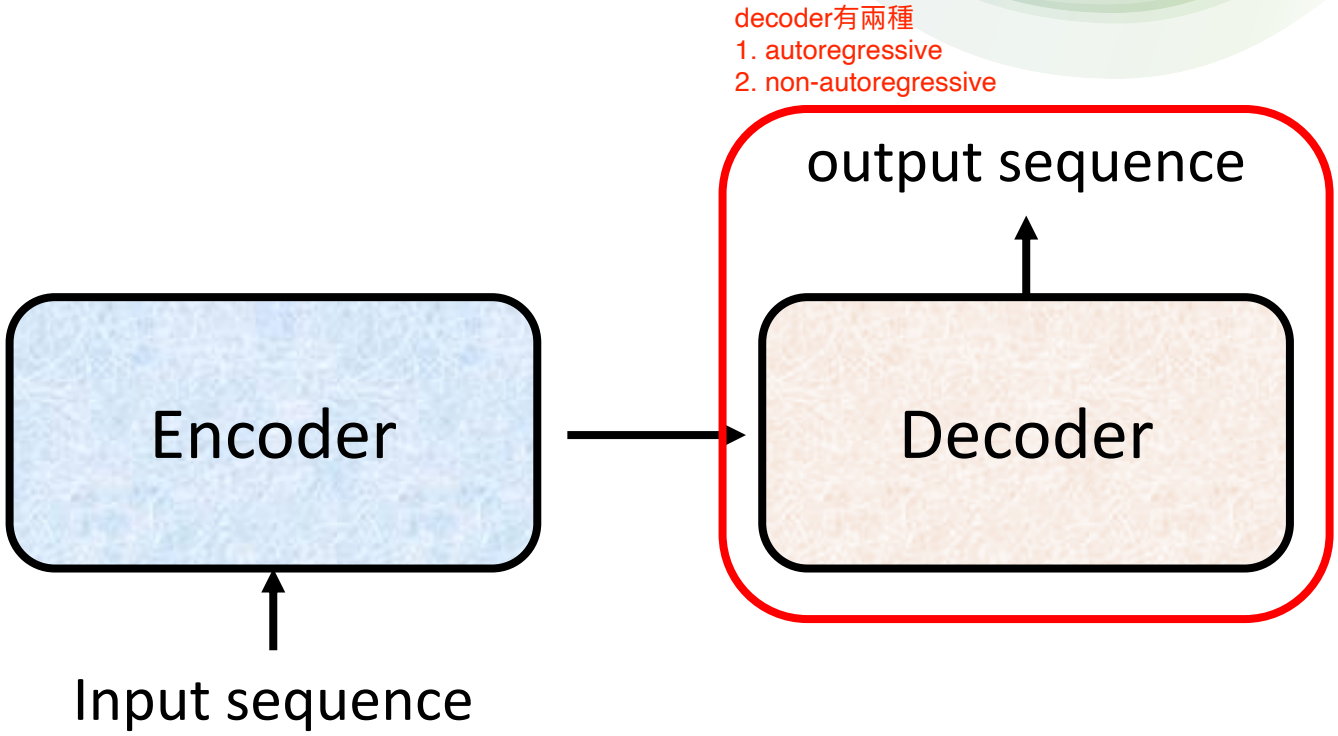
(a)

更換順序過後的
結果發現這樣子的比較好

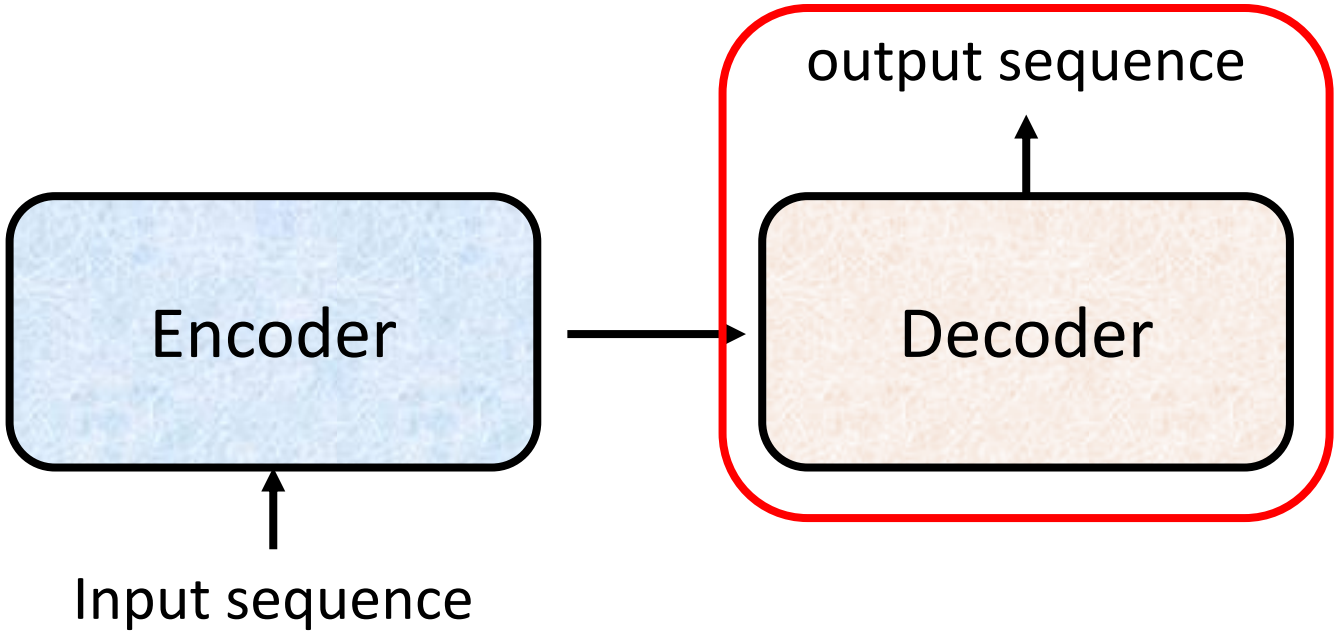


(b)

Decoder

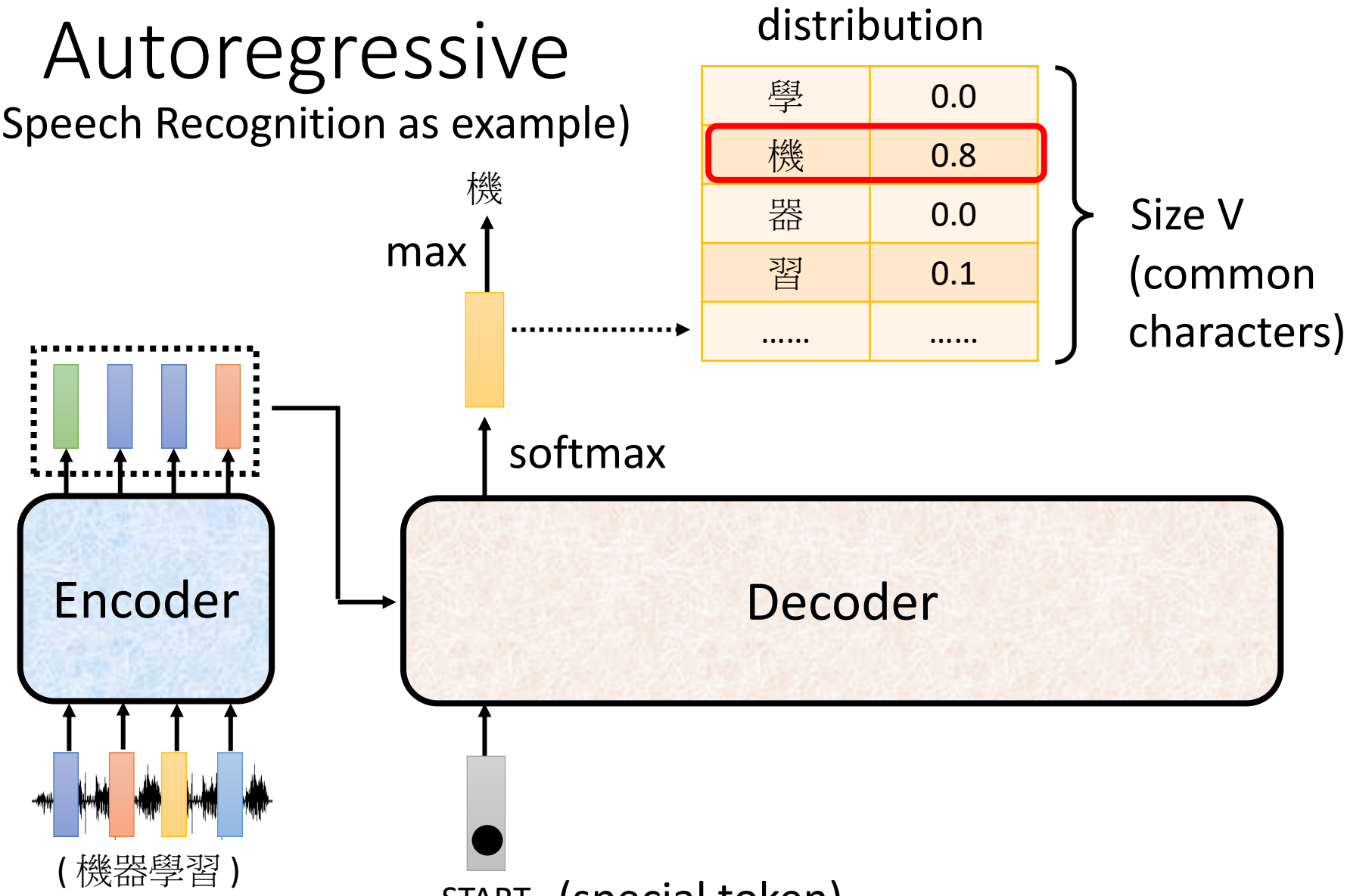

decoder有兩種
1. autoregressive
2. non-autoregressive
output sequence
Encoder
Decoder
Input sequence

Decoder – Autoregressive (AT)



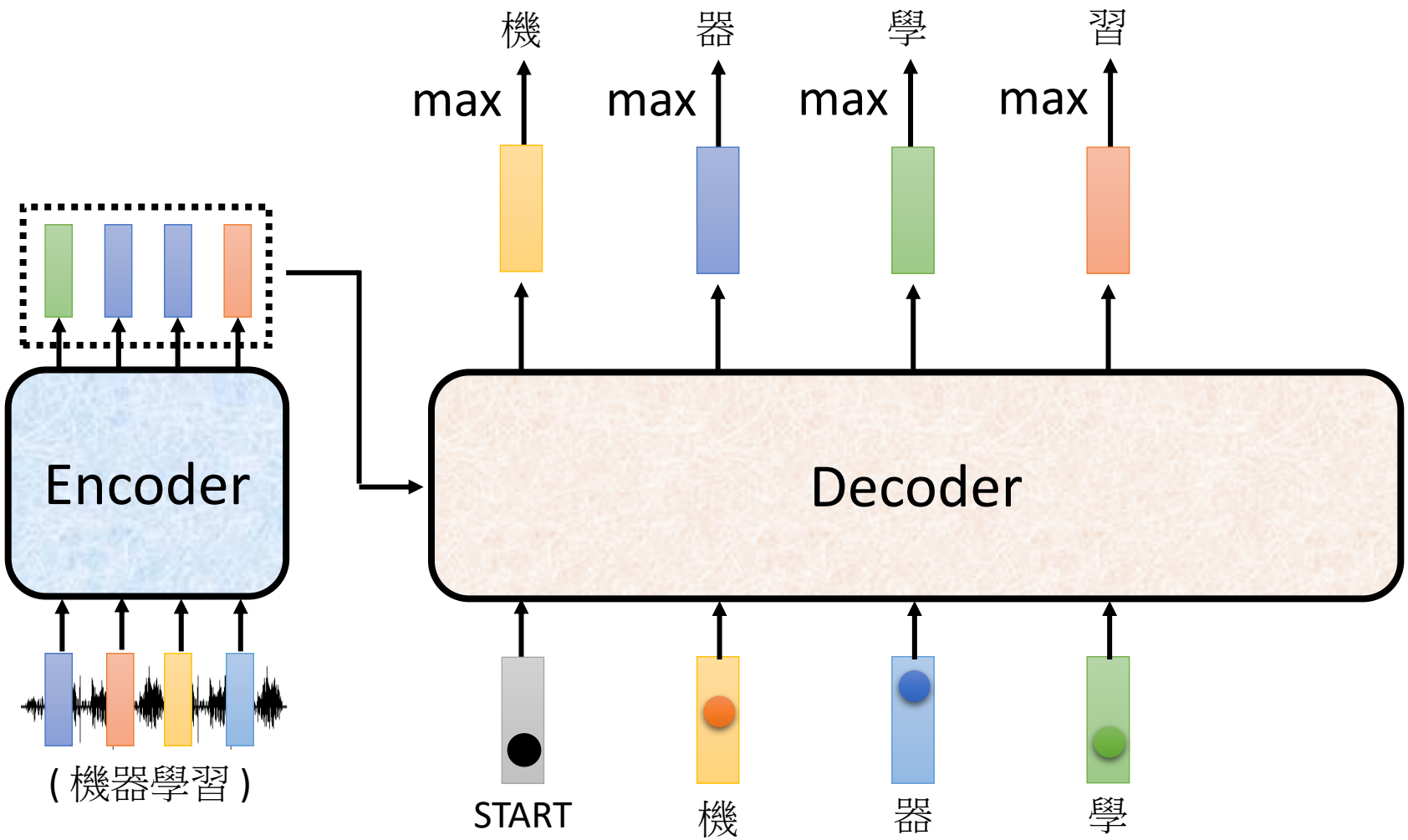
Autoregressive

(Speech Recognition as example)

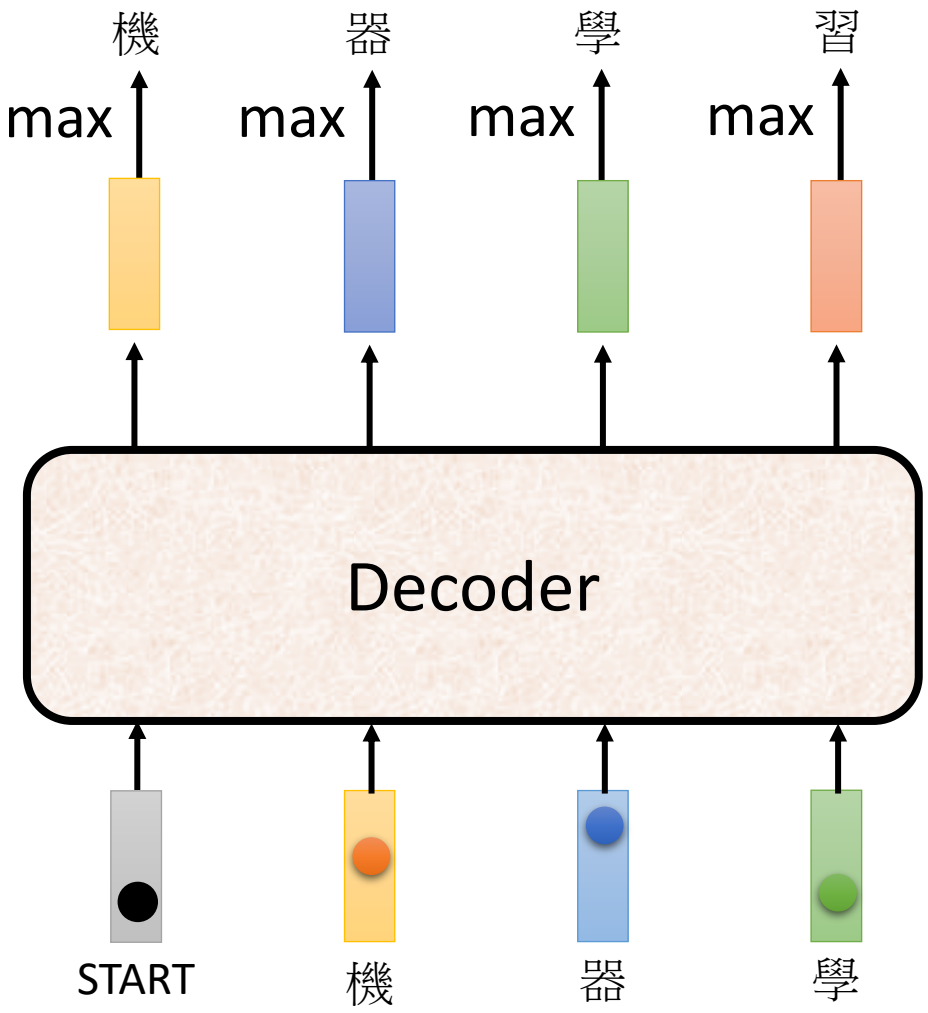


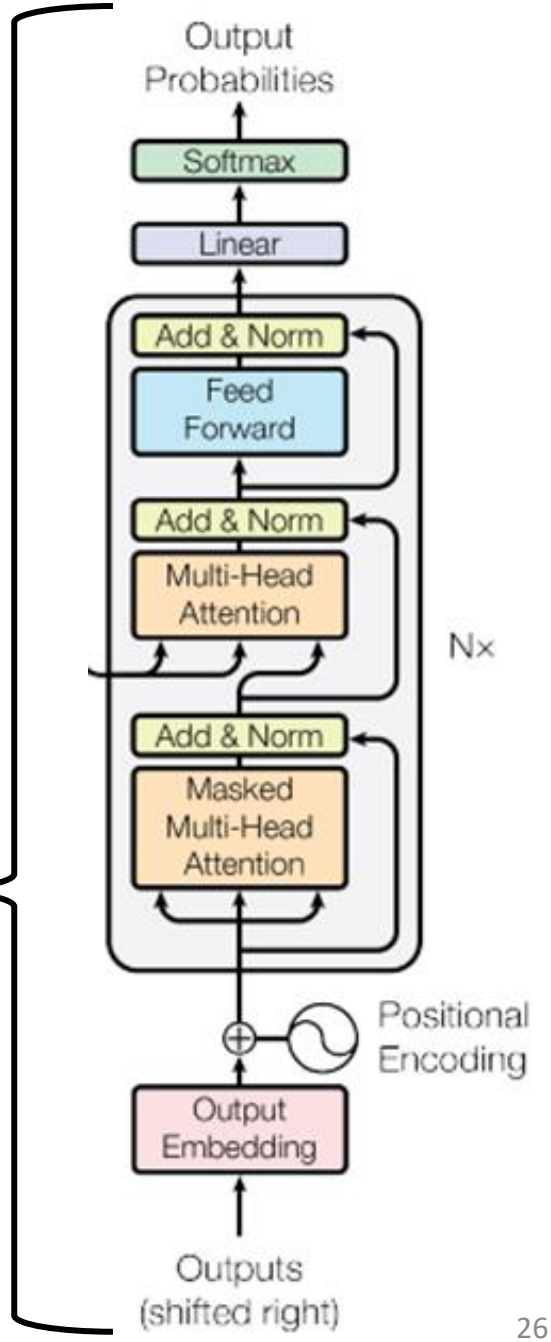
Autoregressive

decoder會把前一個output當成是下一個input
但這樣有可能會一步錯步步錯
（例如把「器」判斷成「氣」，接下來的output都會根據這個錯誤的input來產生結果）
因此後面會討論怎麼解決這個問題

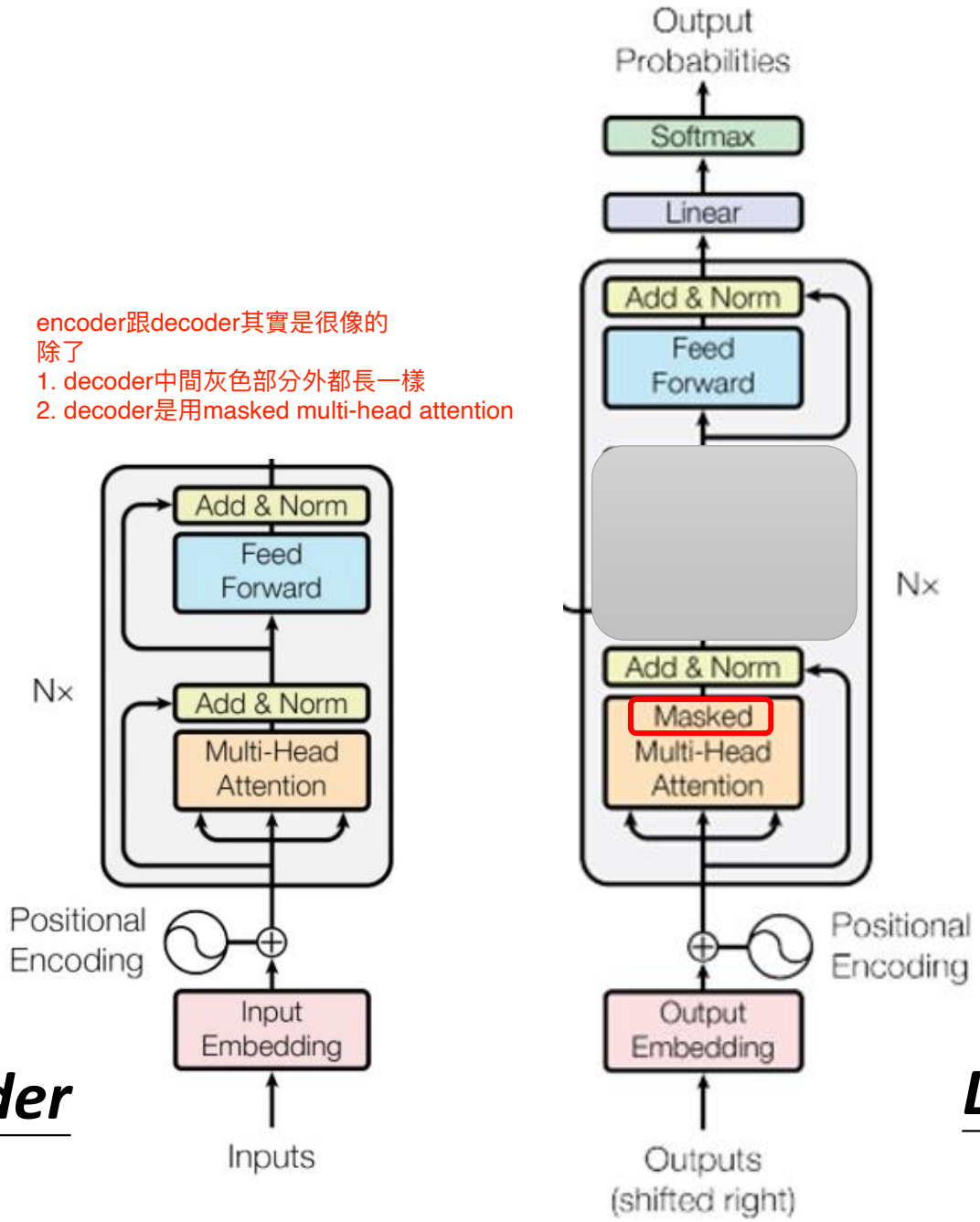


ignore the input from the encoder here ☺





encoder跟decoder其實是很像的
除了
1. decoder中間灰色部分外都長一樣
2. decoder是用masked multi-head attention



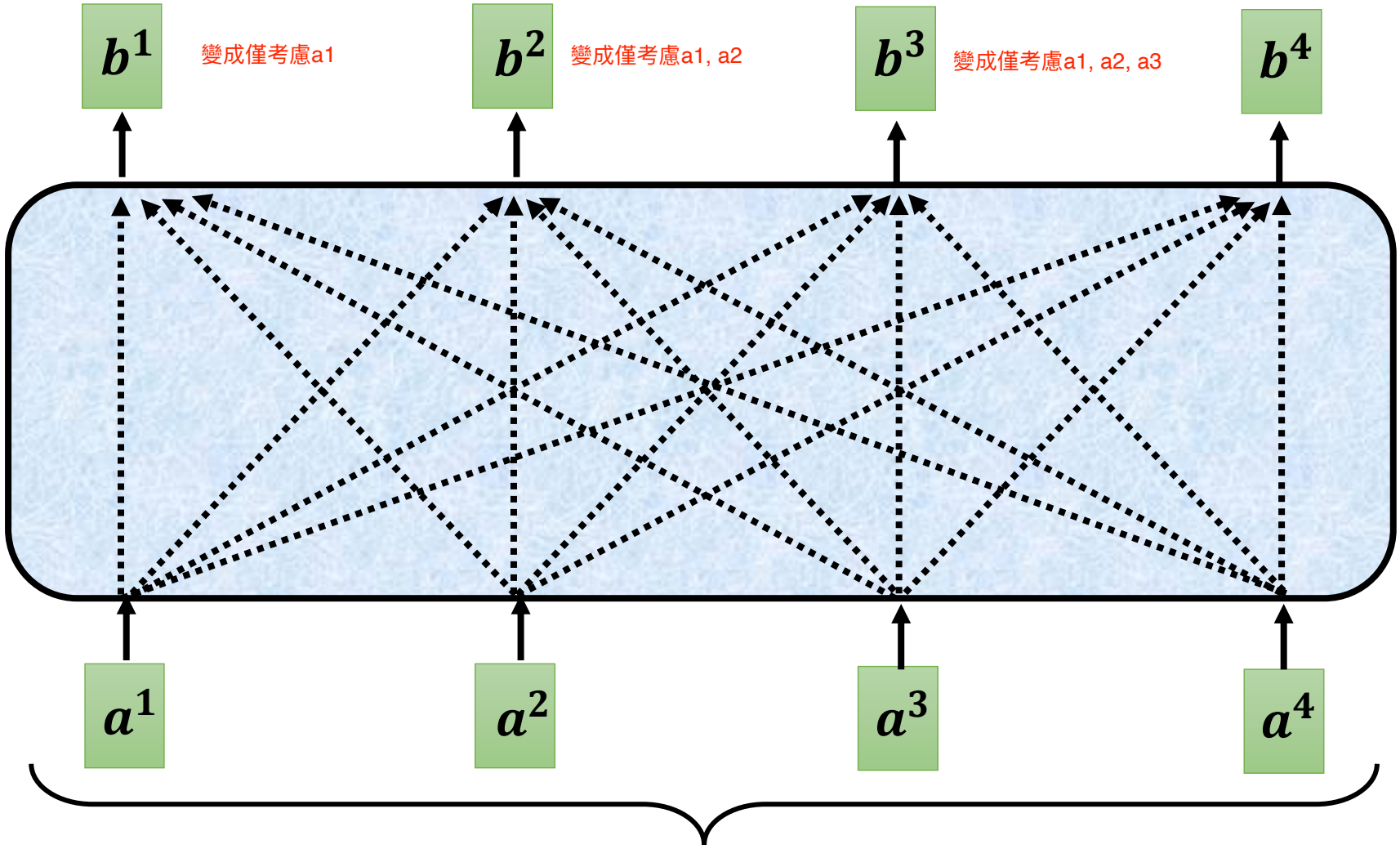
Encoder

Decoder

Self-attention → Masked Self-attention

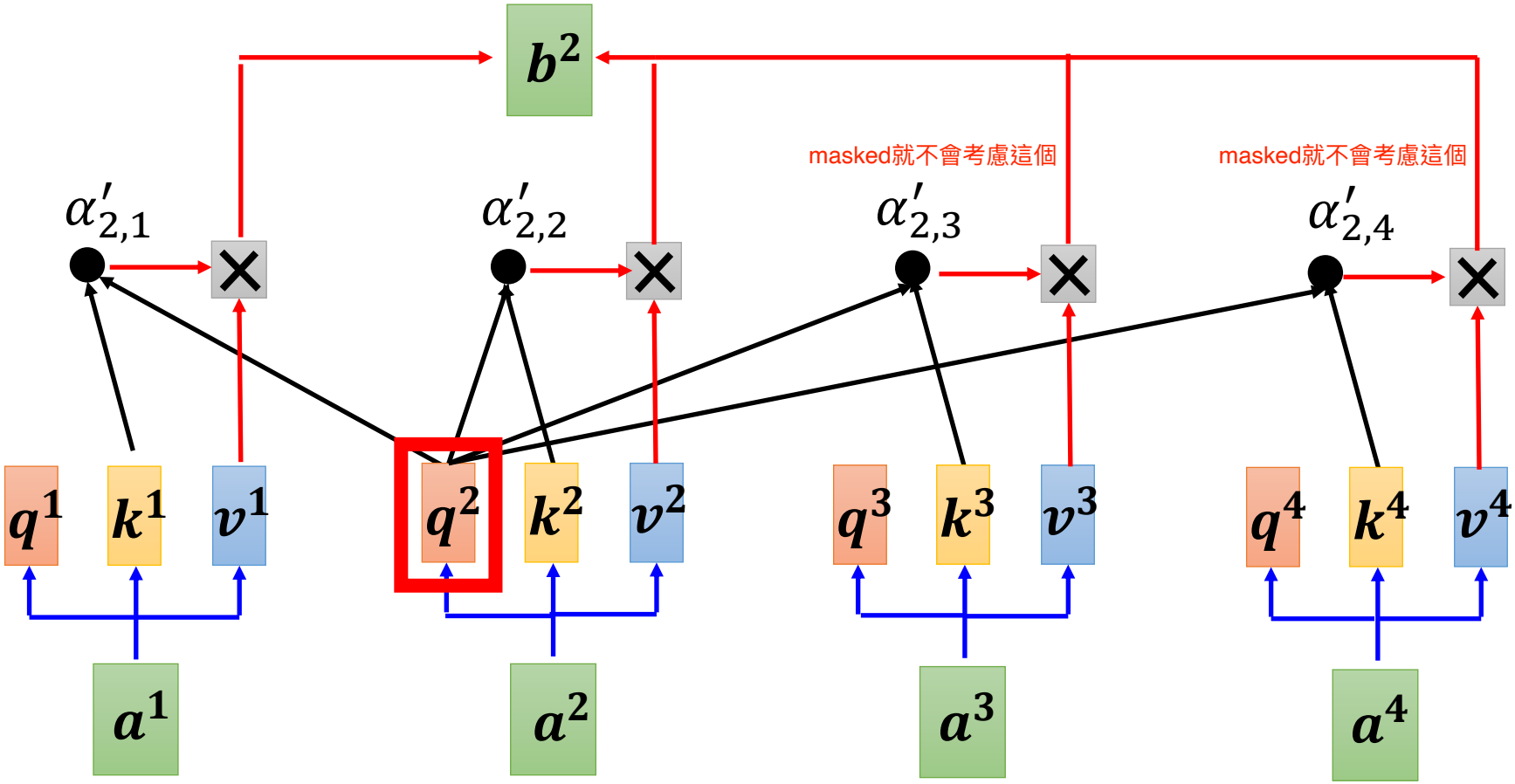
要改用masked的原因很直覺，因為在計算bi時還不知道bj for j > i

因為是最後一個
所以可以考慮全部



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

Self-attention → Masked Self-attention

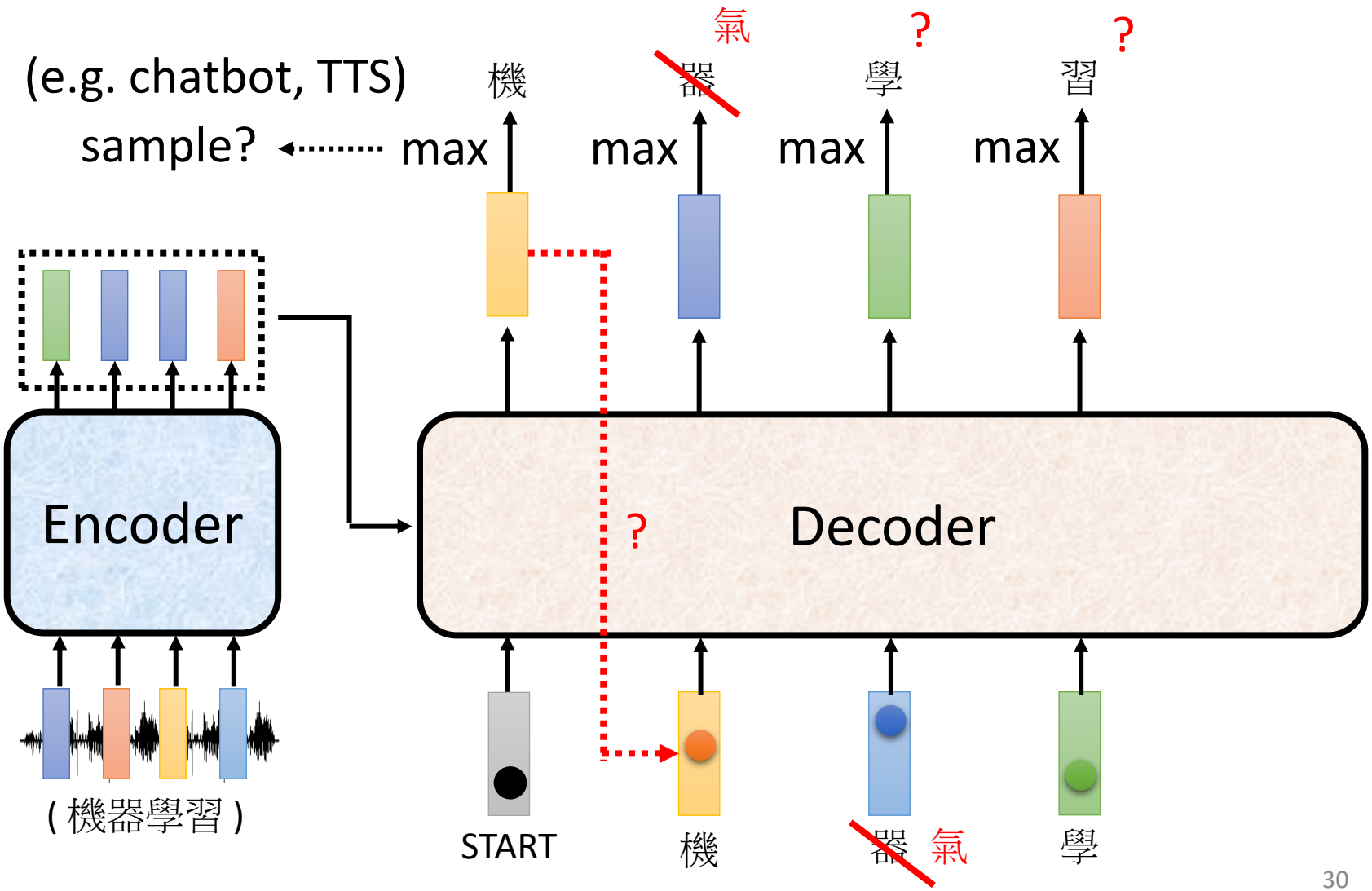


Why masked? Consider how does decoder work

Question?

- Error propagation
- Distribution as input?

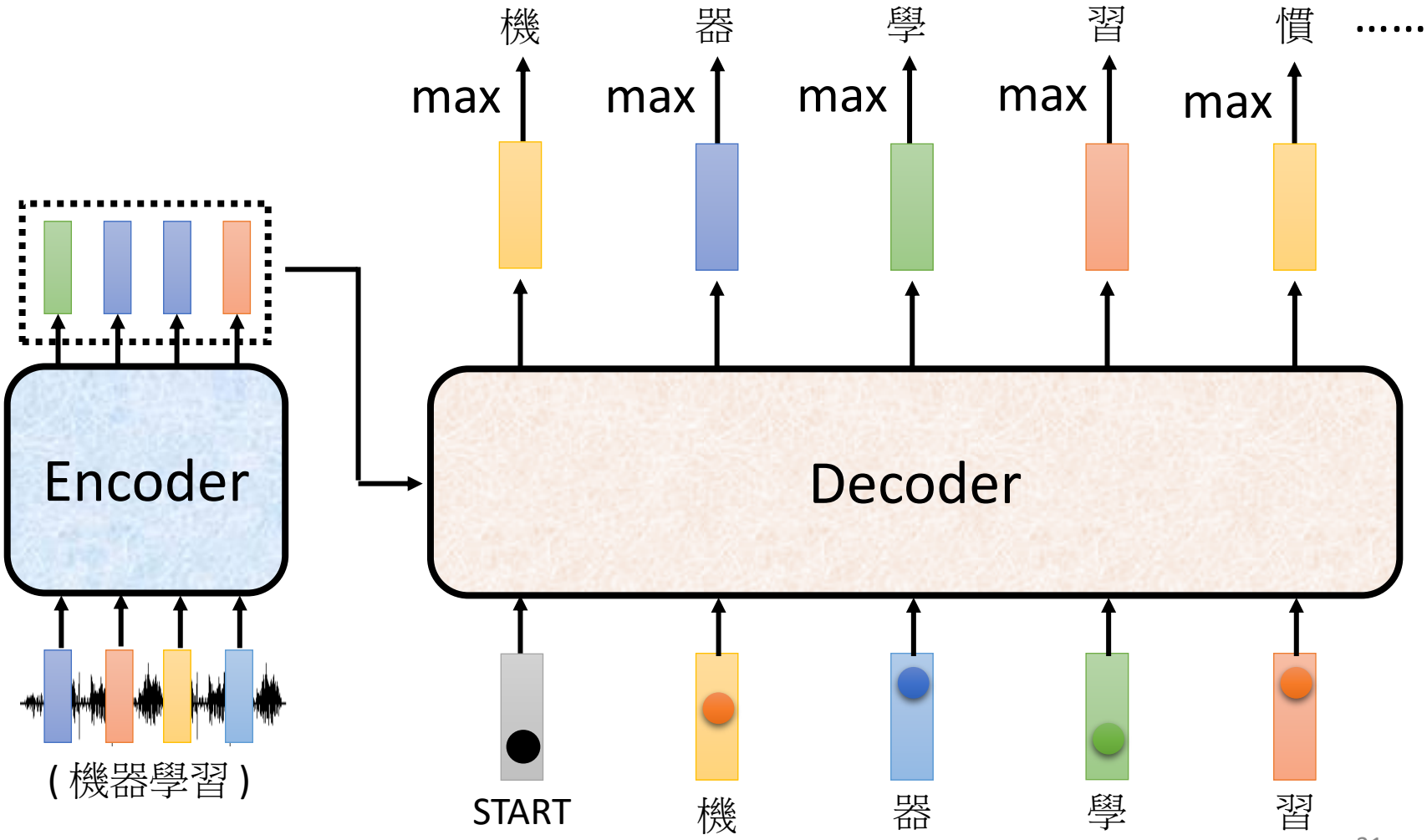
Autoregressive



Autoregressive

We do not know the correct output length.

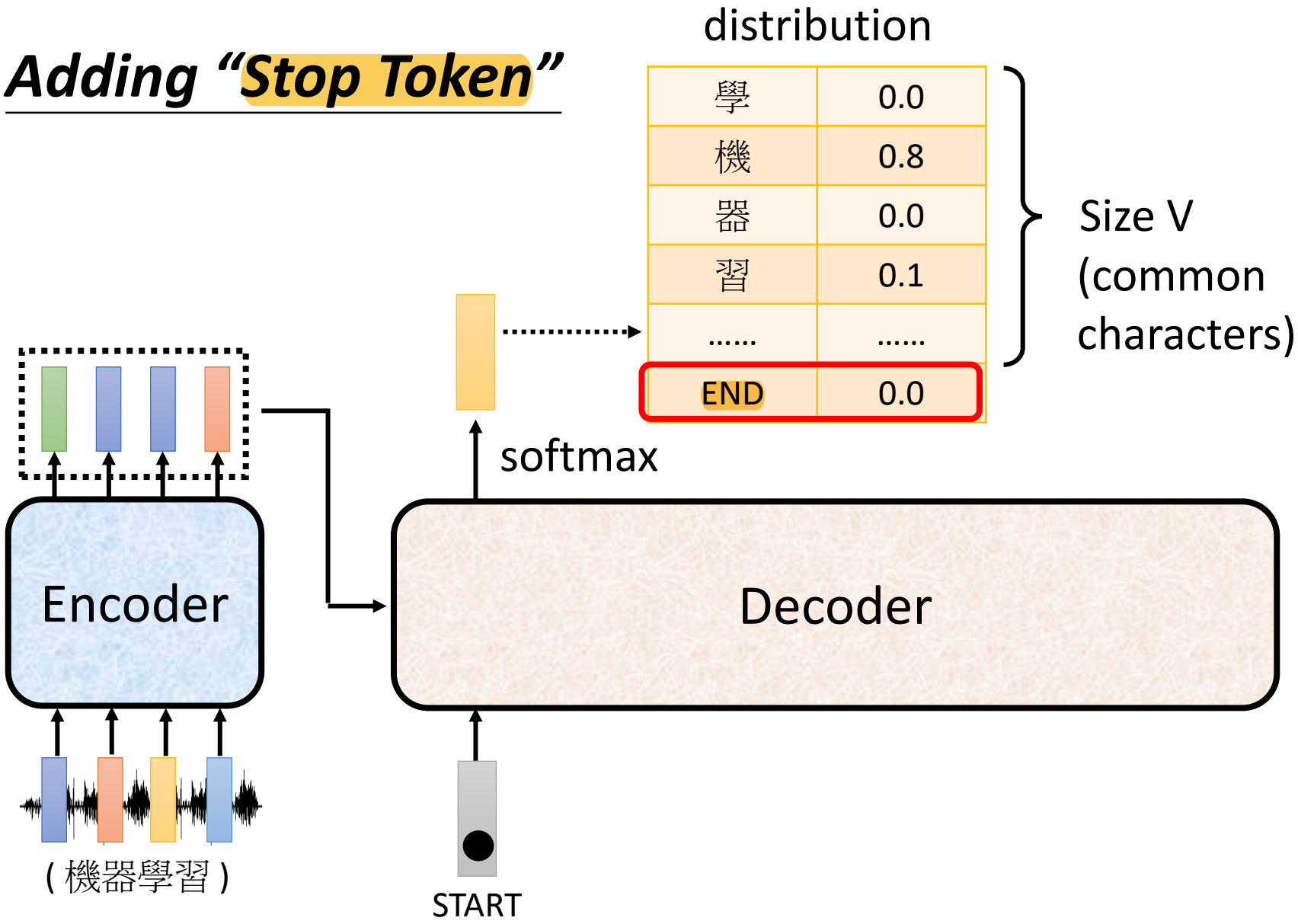
Never stop!



推文接龍 (Tweet Solitaire)

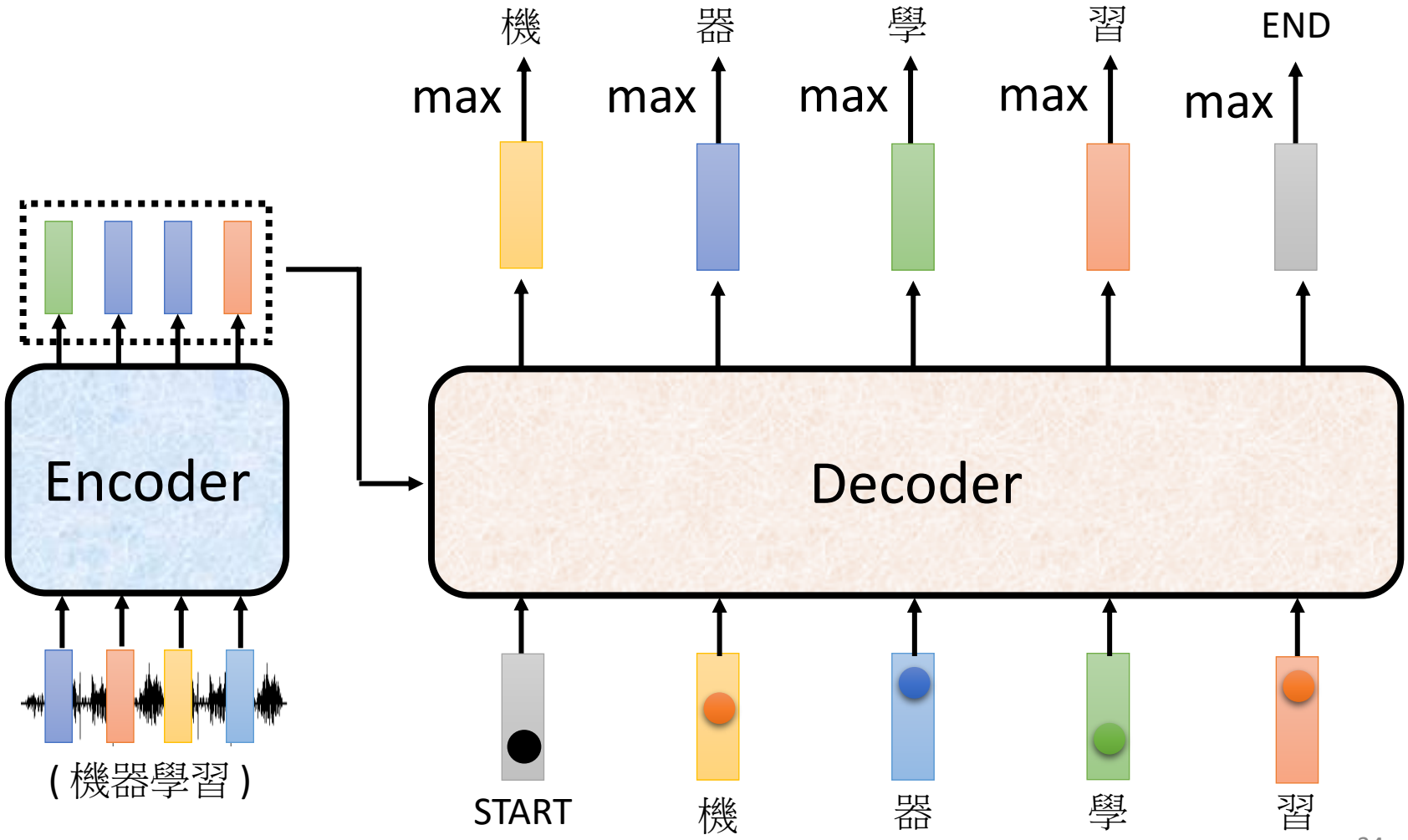
在助教的程式當中，START跟END是用同一個符號，其實也是可以的

Adding "Stop Token"

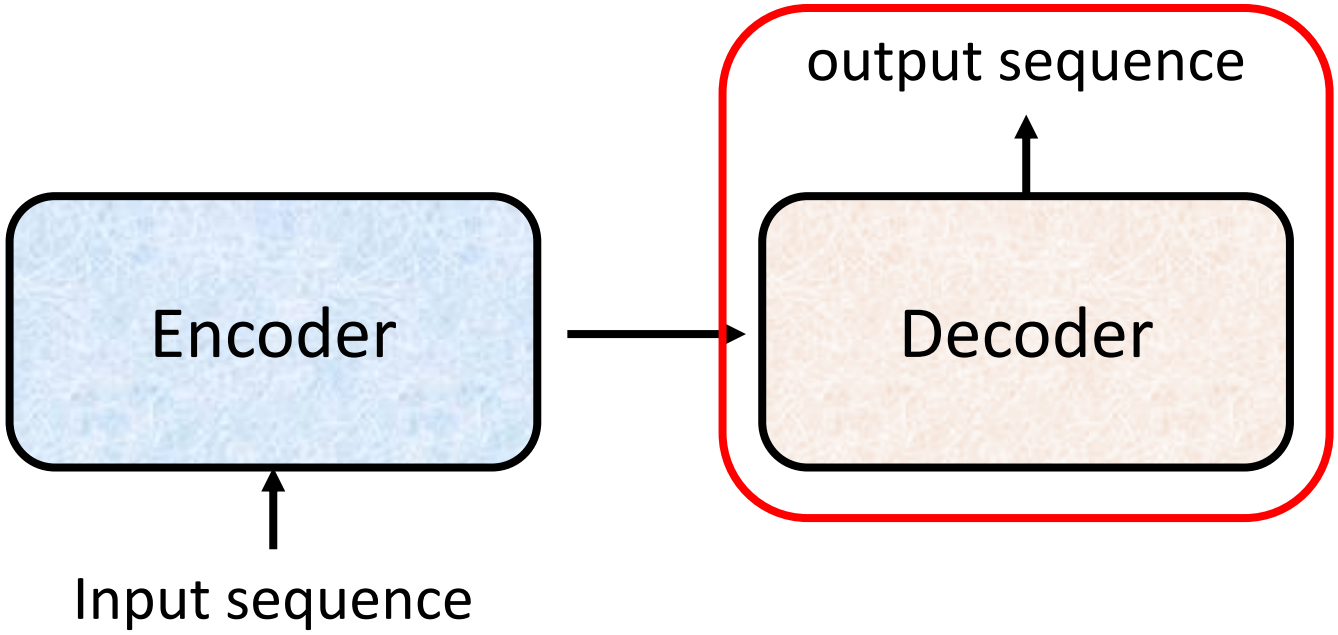


Autoregressive

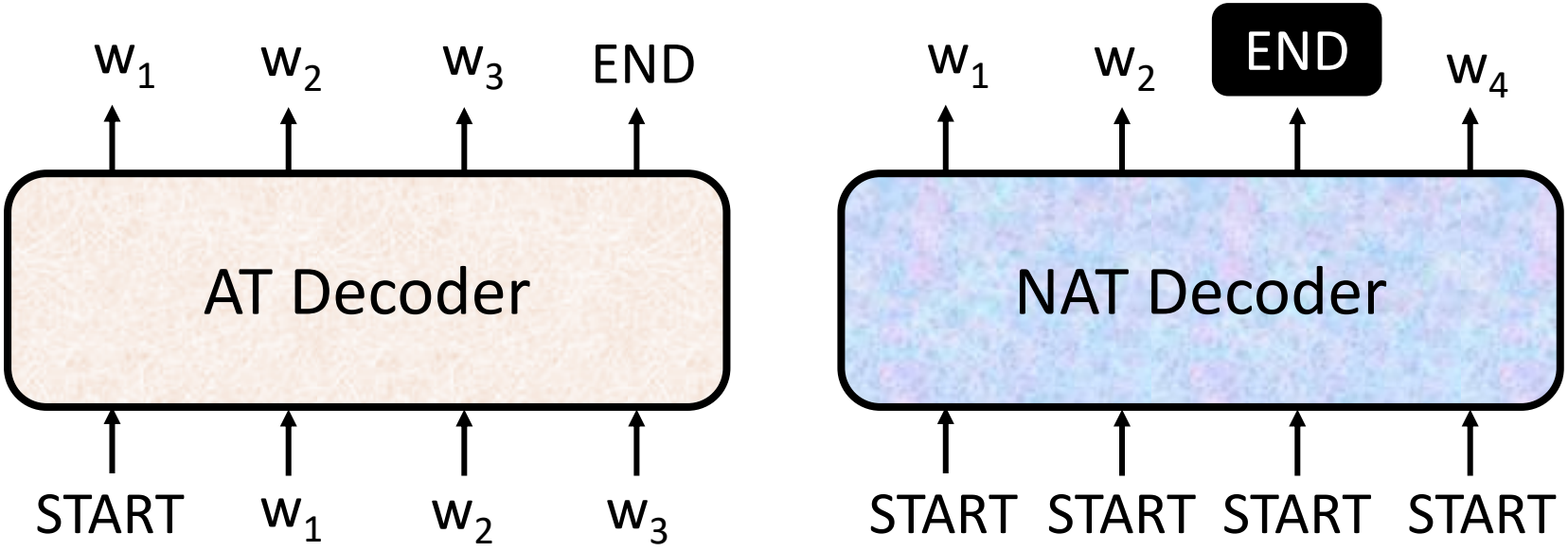
Stop at here!



Decoder
– Non-autoregressive (NAT)



AT v.s. NAT



- How to decide the output length for NAT decoder? 有兩種辦法
  - Another predictor for output length
  - Output a very long sequence, ignore tokens after END
- Advantage: parallel (顯然是在transformer之後才有的), more stable generation (e.g., TTS)
- NAT is usually worse than AT (why? Multi-modality)

To learn more ……

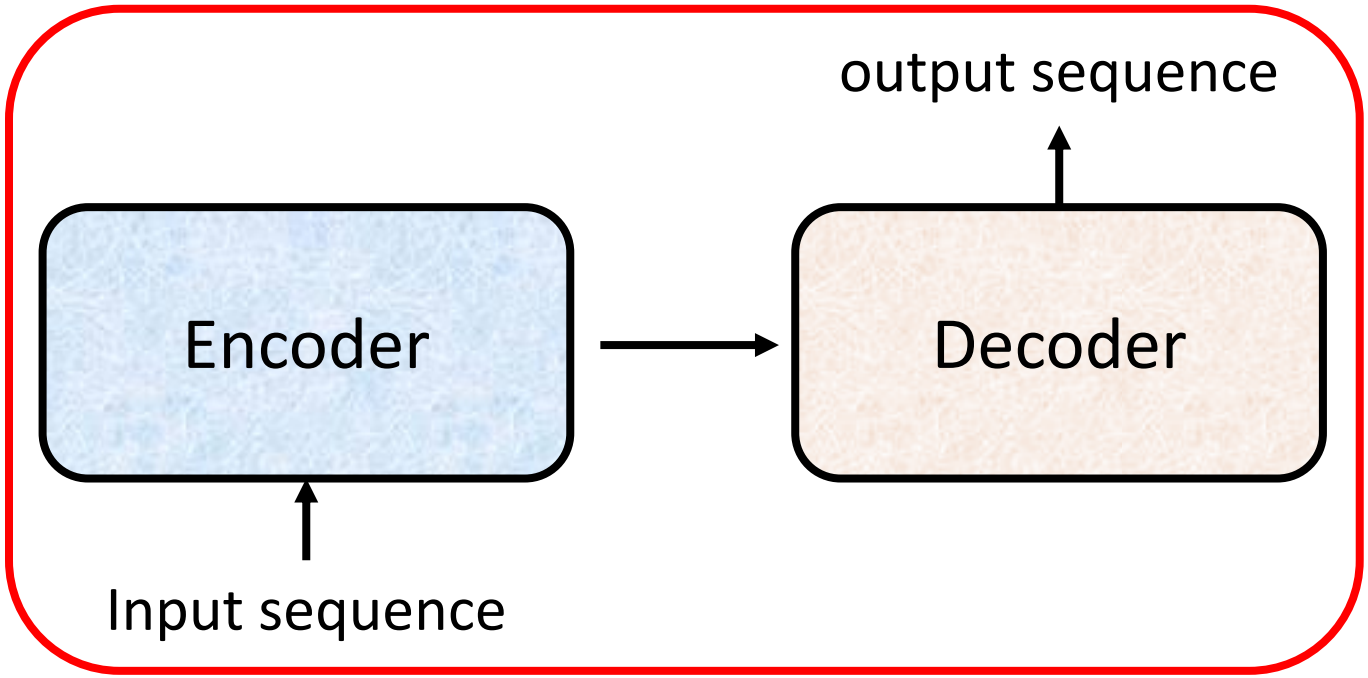
更多NAT有關課程

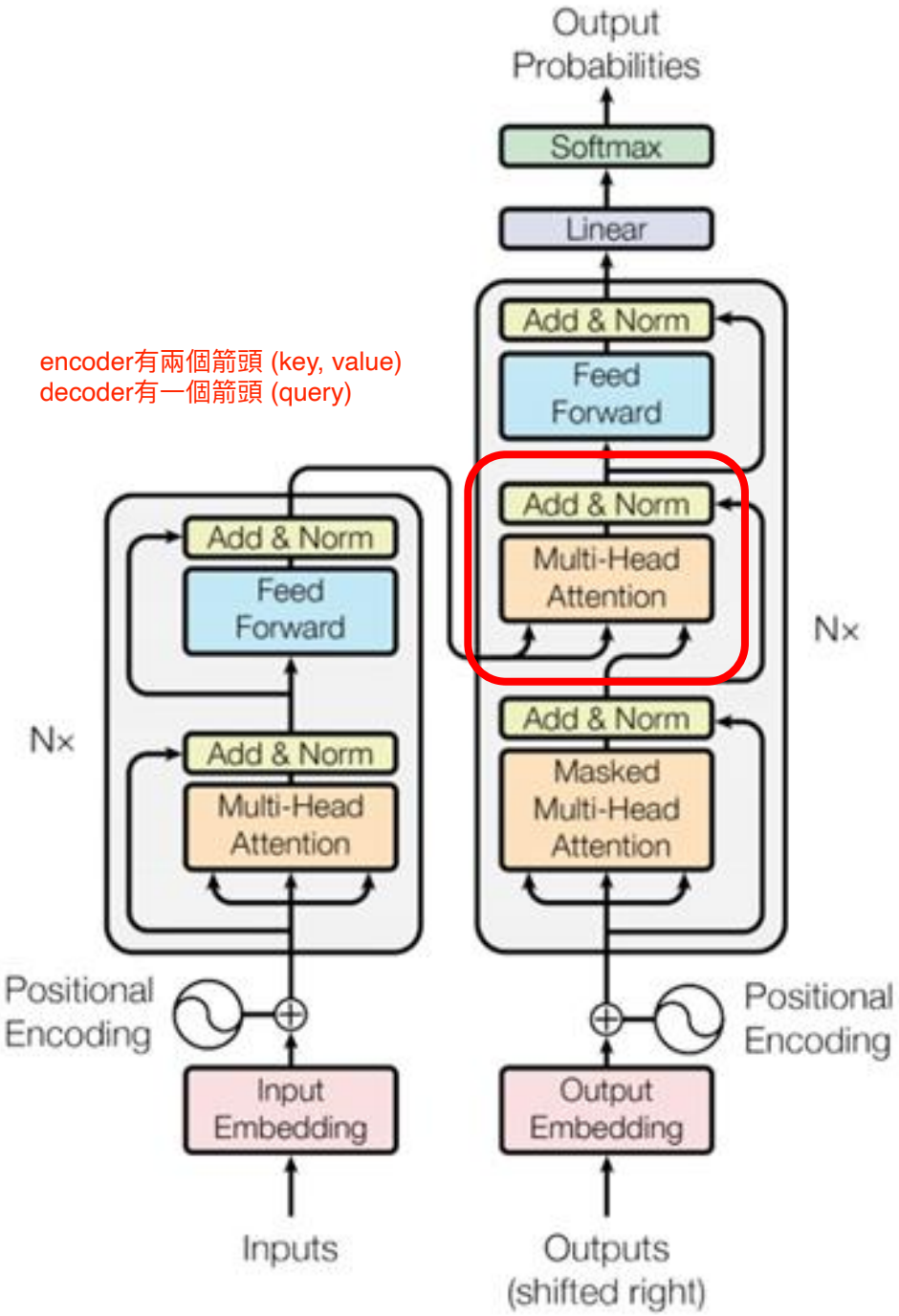
https://youtu.be/jvyKmU4OM3c

(in Mandarin)


Q&A

Encoder-Decoder





encoder有兩個箭頭 (key, value)
decoder有一個箭頭 (query)

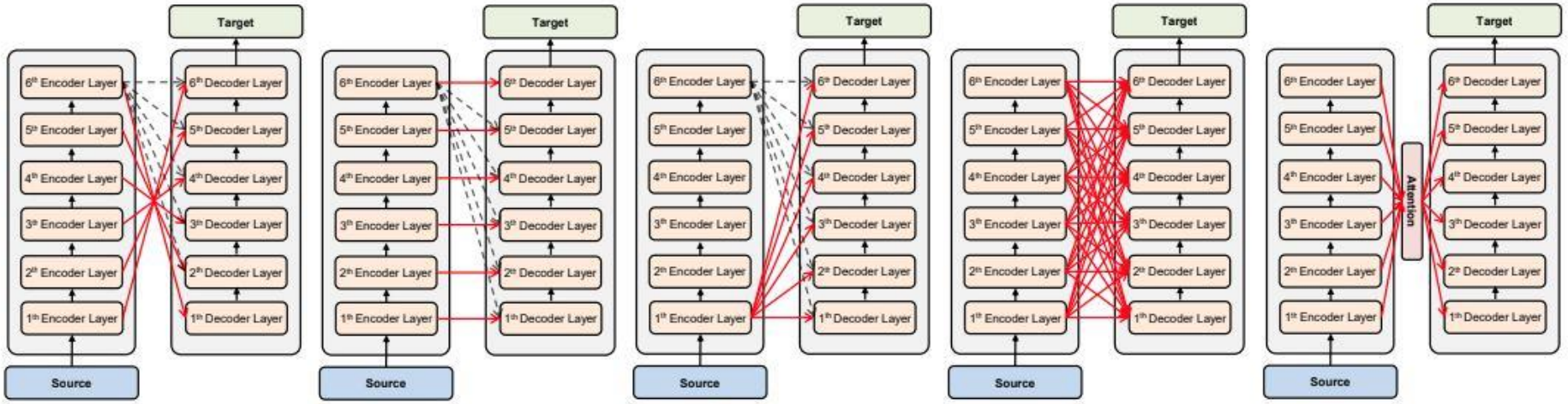
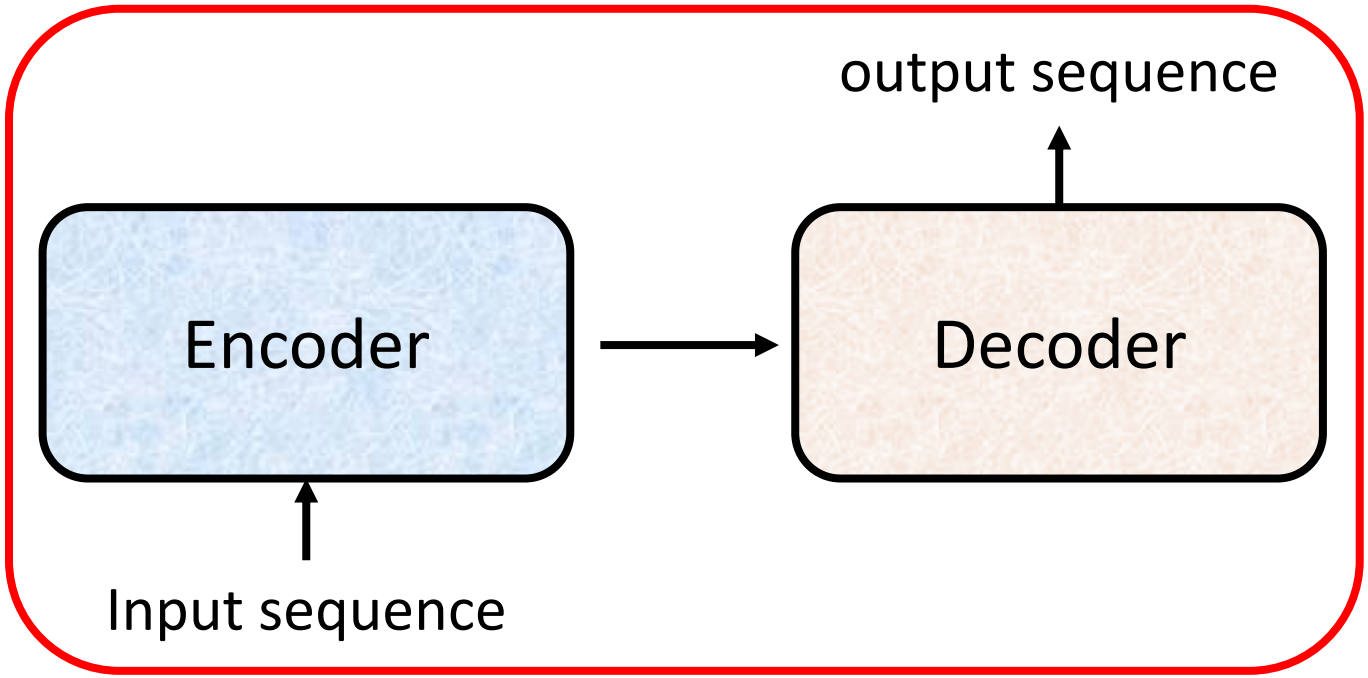
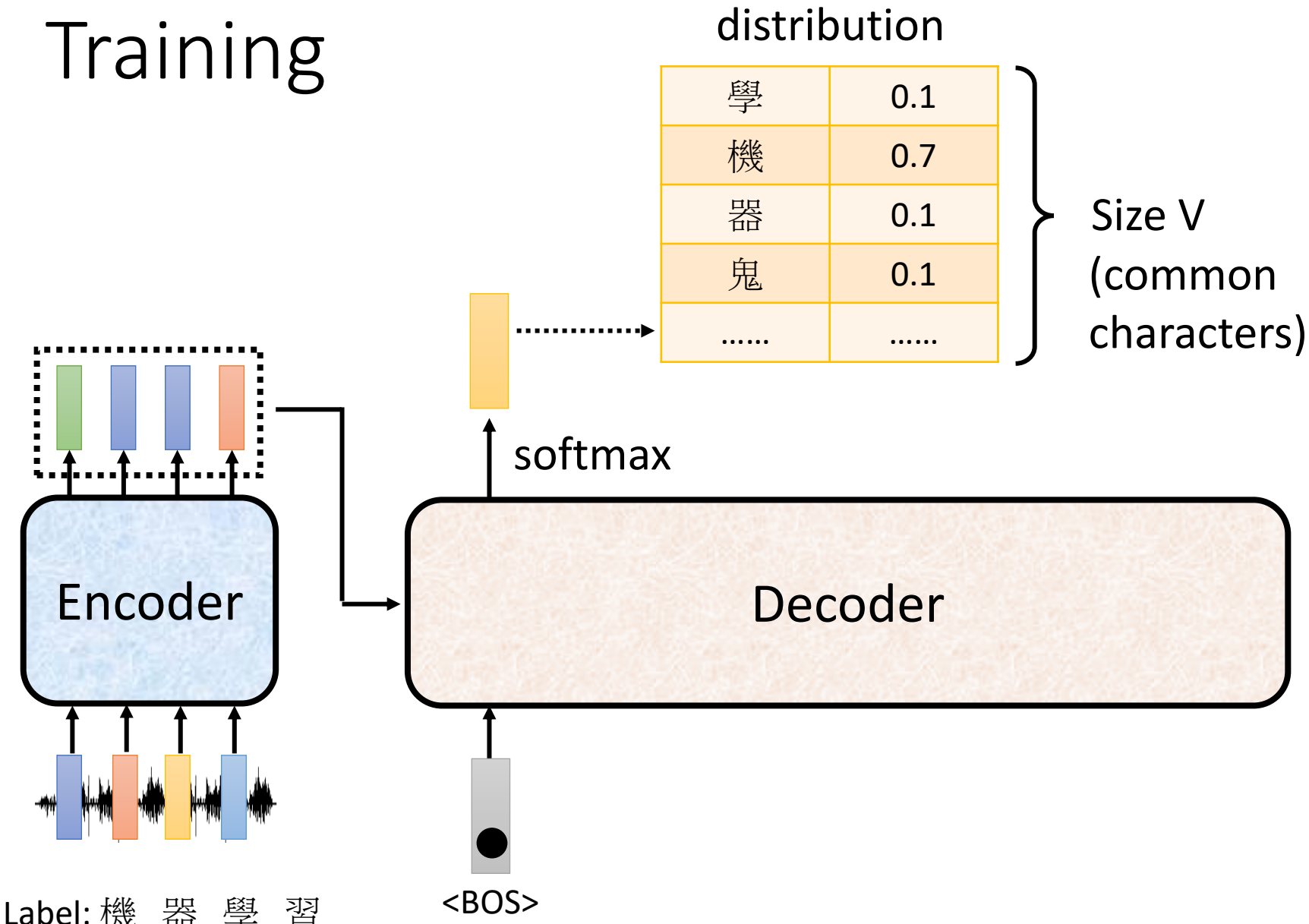


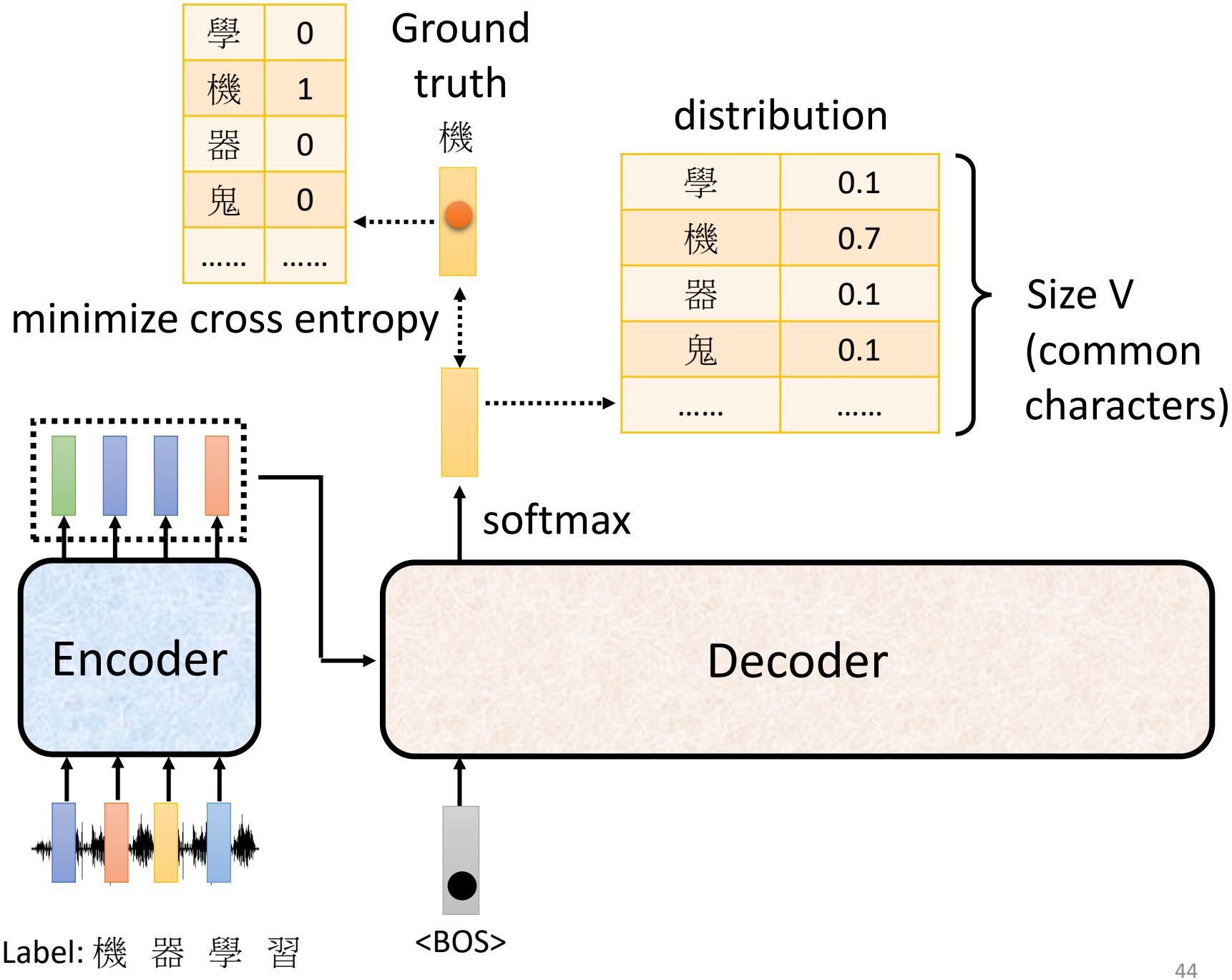
Figure 2: We present the proposal on Transformer with various strategies for routing the source representations: (a) Granularity Consistent Attention; (b) Granularity Parallel Attention; (c) Fine-Grained Attention; (d) Full Matching Attention; (e) Adaptive Matching Attention. The dashed lines represent the original attention to the last encoder layer and we omit them in (e) for clarity.

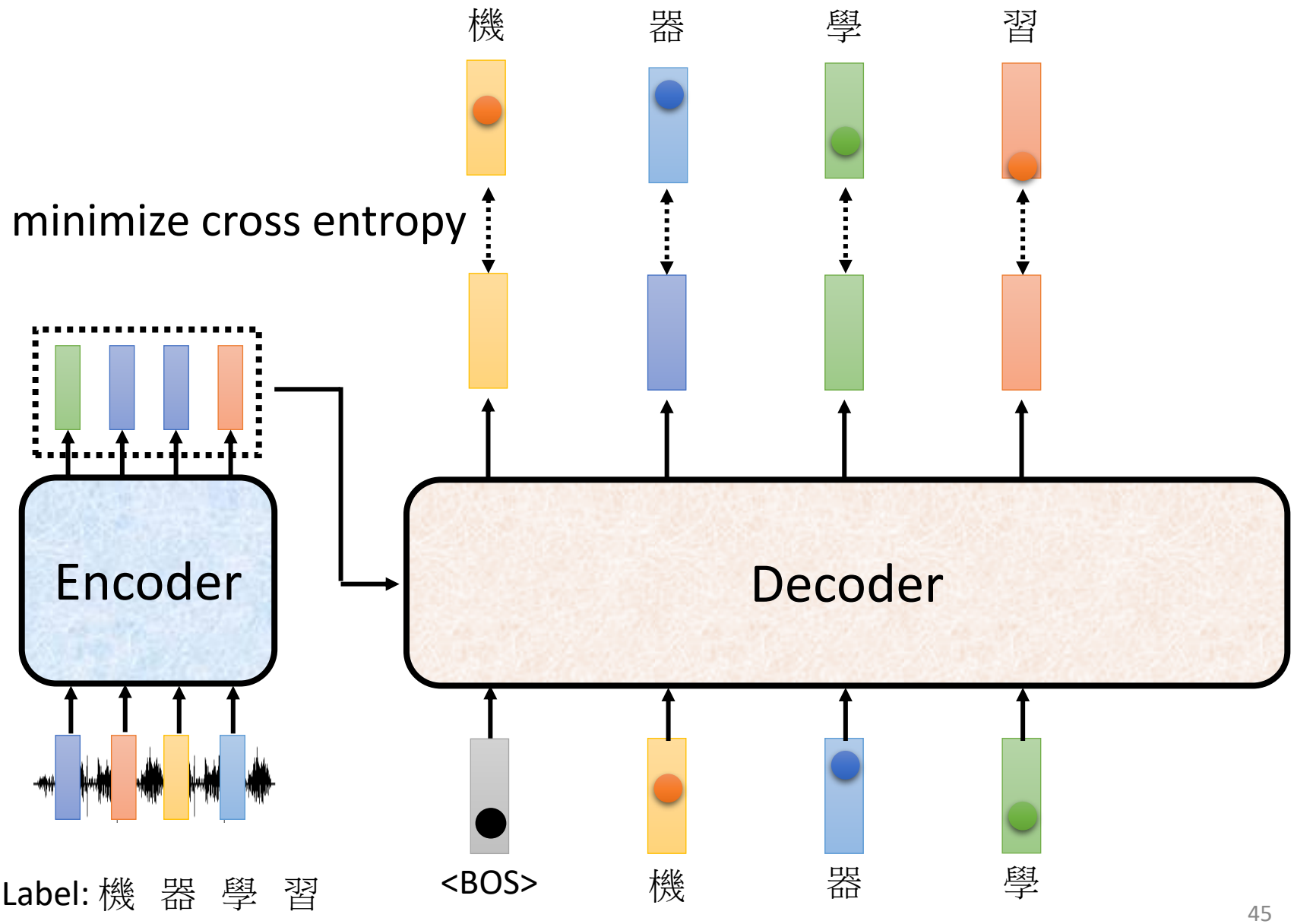
Training



Training


distribution
學 0.1
機 0.7
器 0.1
鬼 0.1
…… ……
Size V (common characters)
softmax
Encoder
Decoder
<BOS>
Label: 機 器 學 習


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


機
器
學
習
minimize cross entropy
Encoder
Decoder
<BOS>
機
器
學
Label: 機 器 學 習