# Transformer

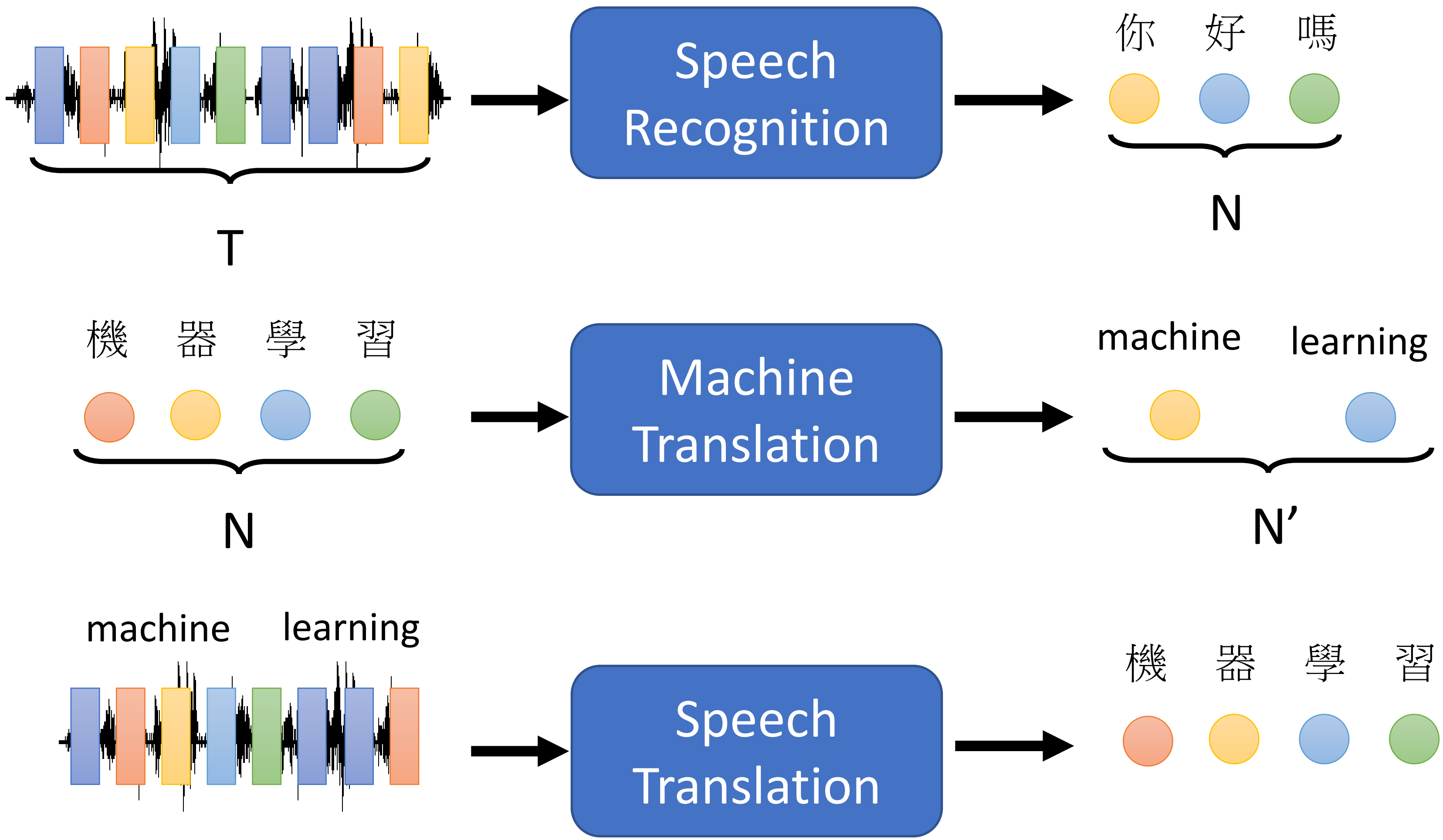

李宏毅

Hung-yi Lee


# ***Sequence-to-sequence (Seq2seq)***

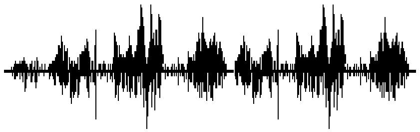
Input a sequence, output a sequence

The output length is determined by model.

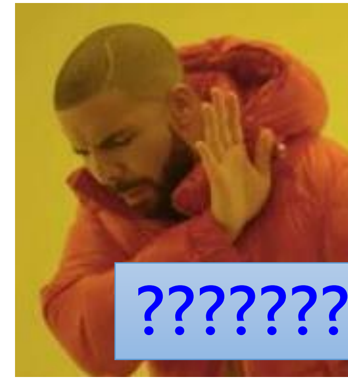



Language without text

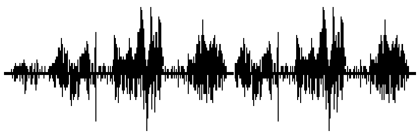
# Hokkien (閩南語、台語)



(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

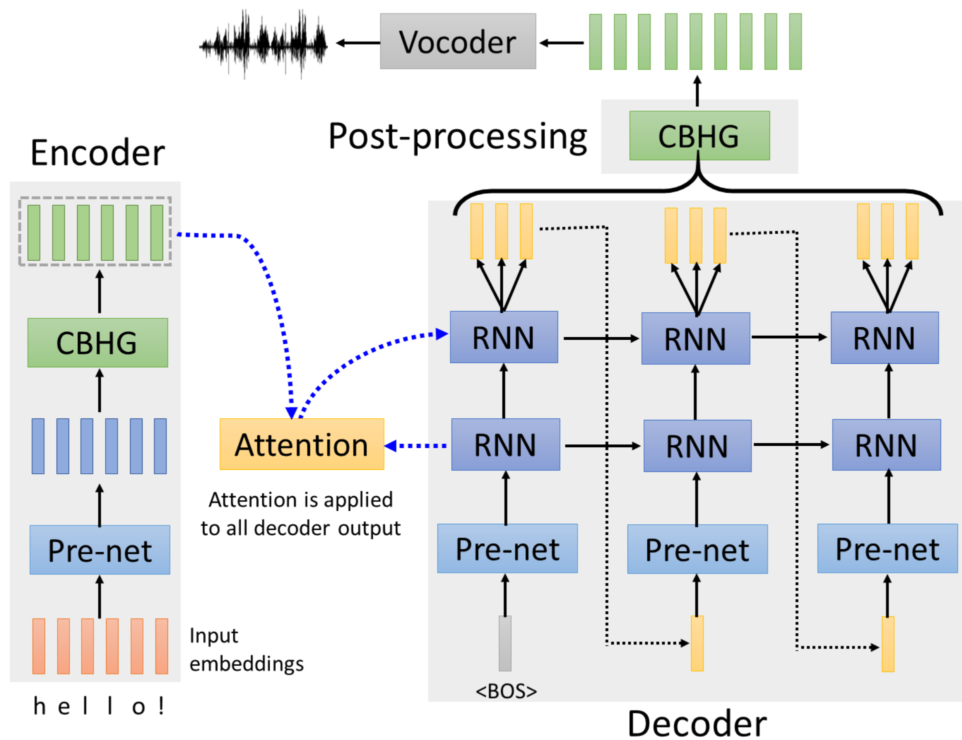
我 有 幫 廠 長 拜 託

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



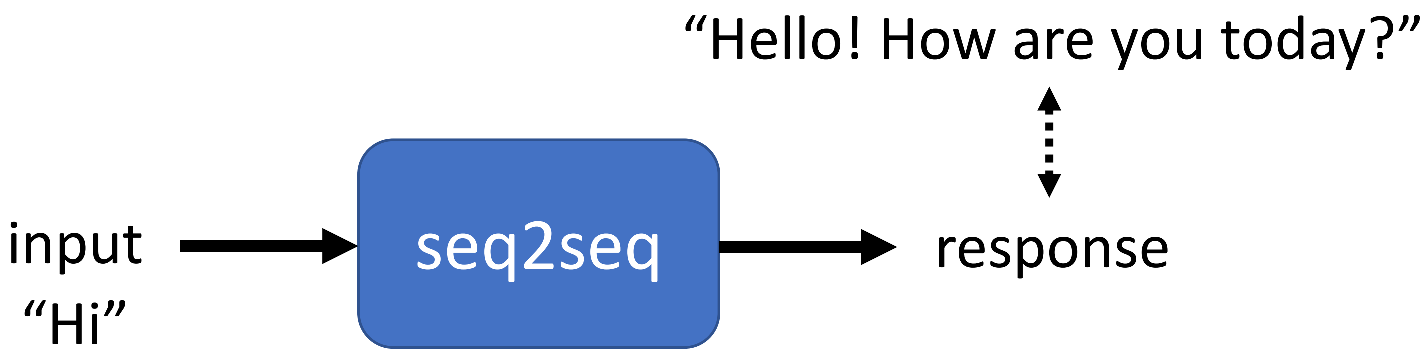

***Taiwanese Speech Synthesis***

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 |

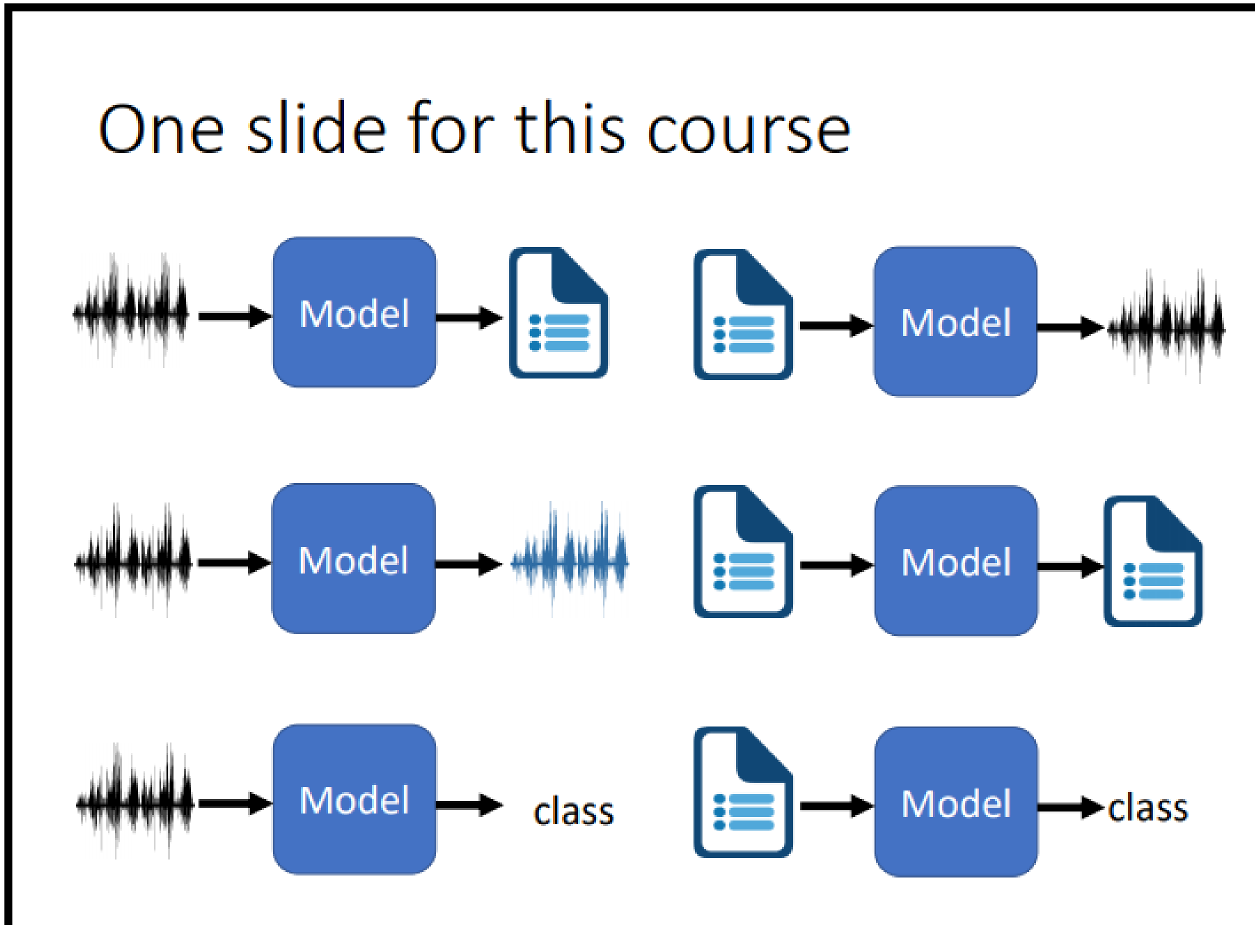
decaNLP

**QA** can be done by **seq2seq**

question, context → Seq2seq → answer

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

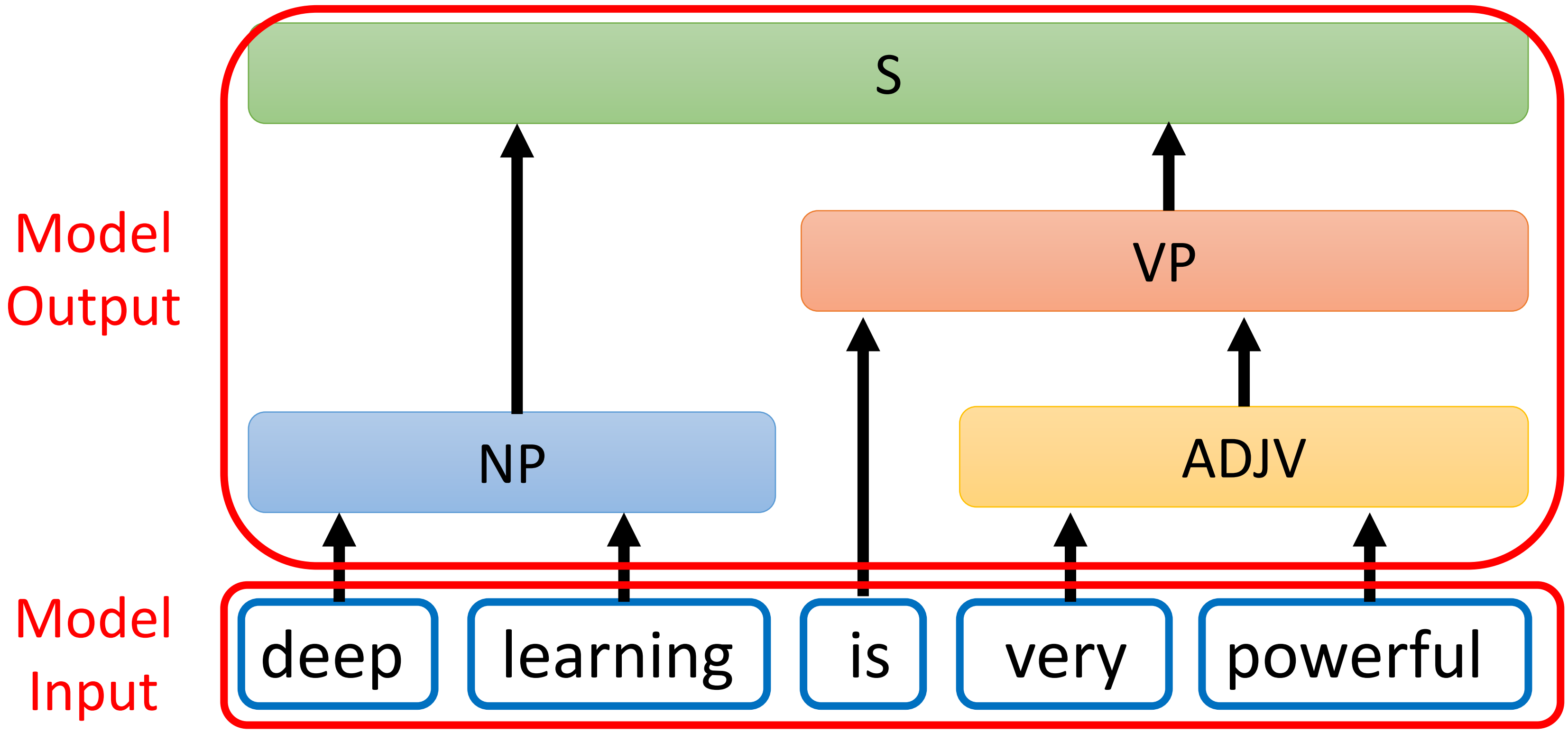
# Deep Learning for Human Language Processing
# 深度學習與人類語言處理



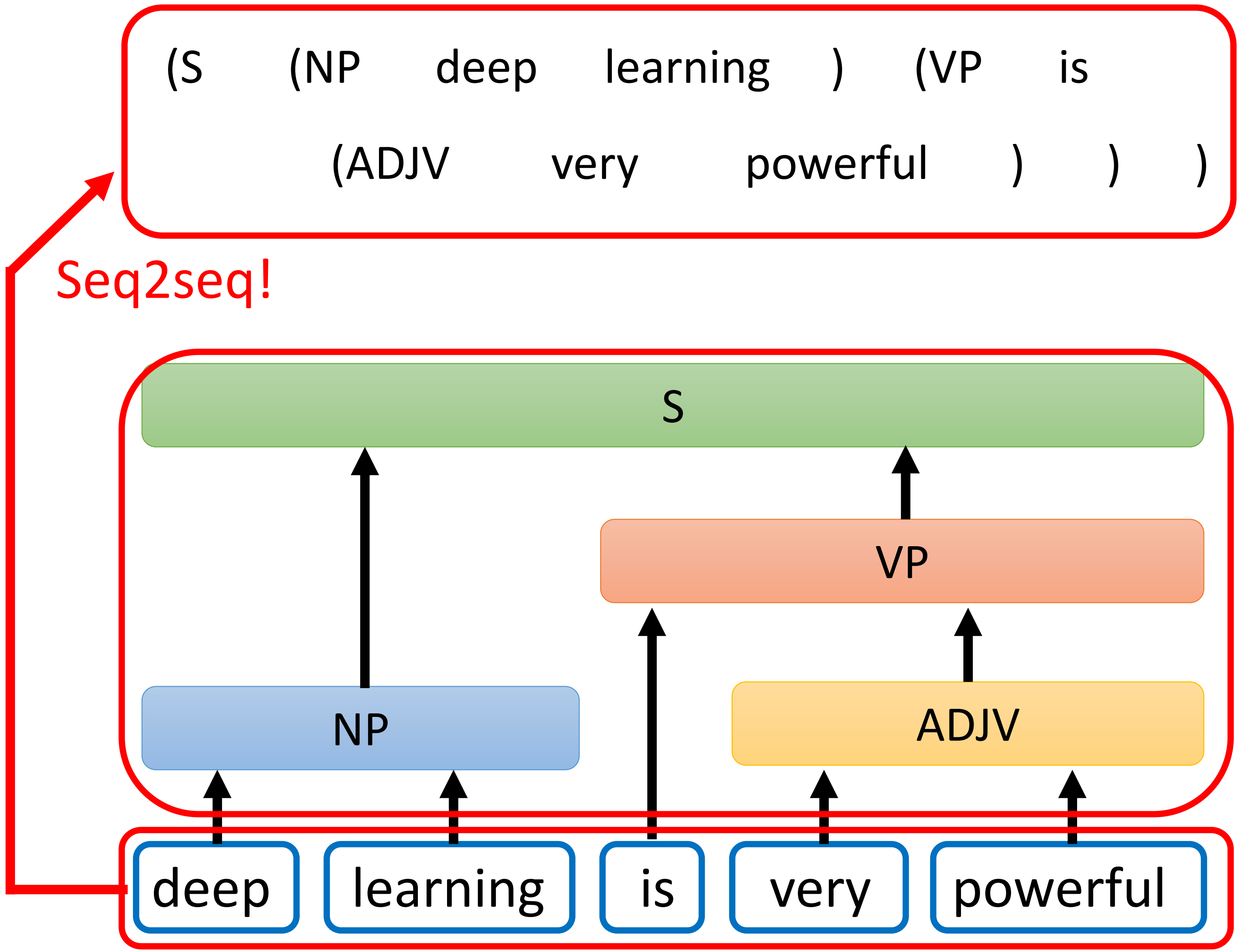

Source webpage: https://speech.ee.ntu.edu.tw/~hylee/dlhlp/2020-spring.html

# *Seq2seq for Syntactic Parsing*

**Is it a 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

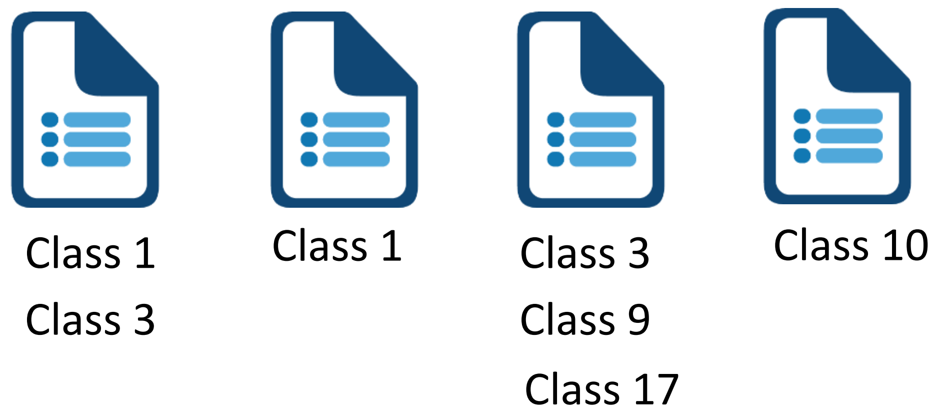
https://arxiv.org/abs/1412.7449

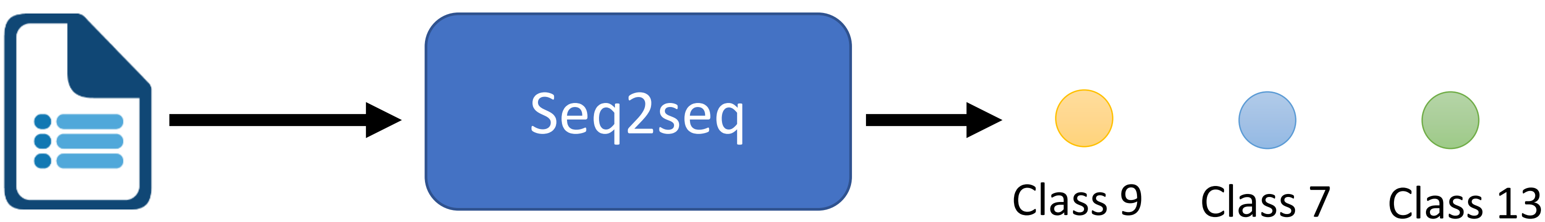
deep learning is very powerful

# Seq2seq for Multi-label Classification

c.f. Multi-class Classification

An object can belong to multiple classes.

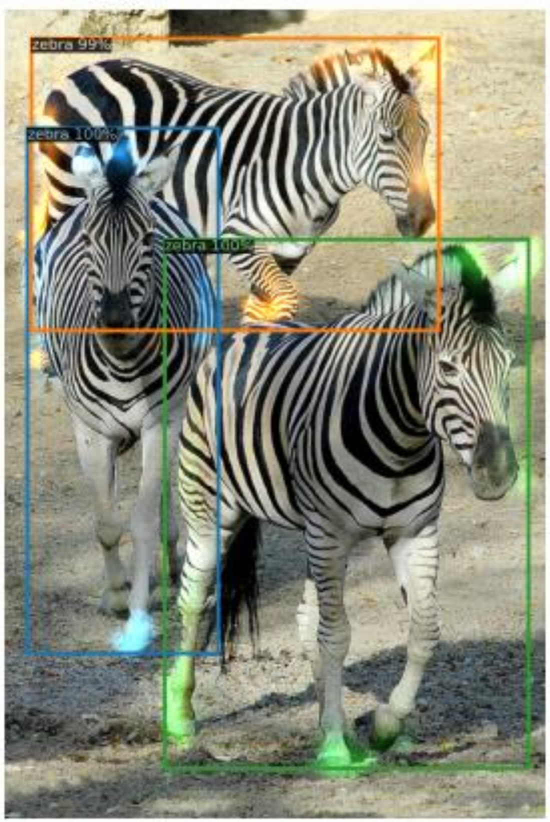


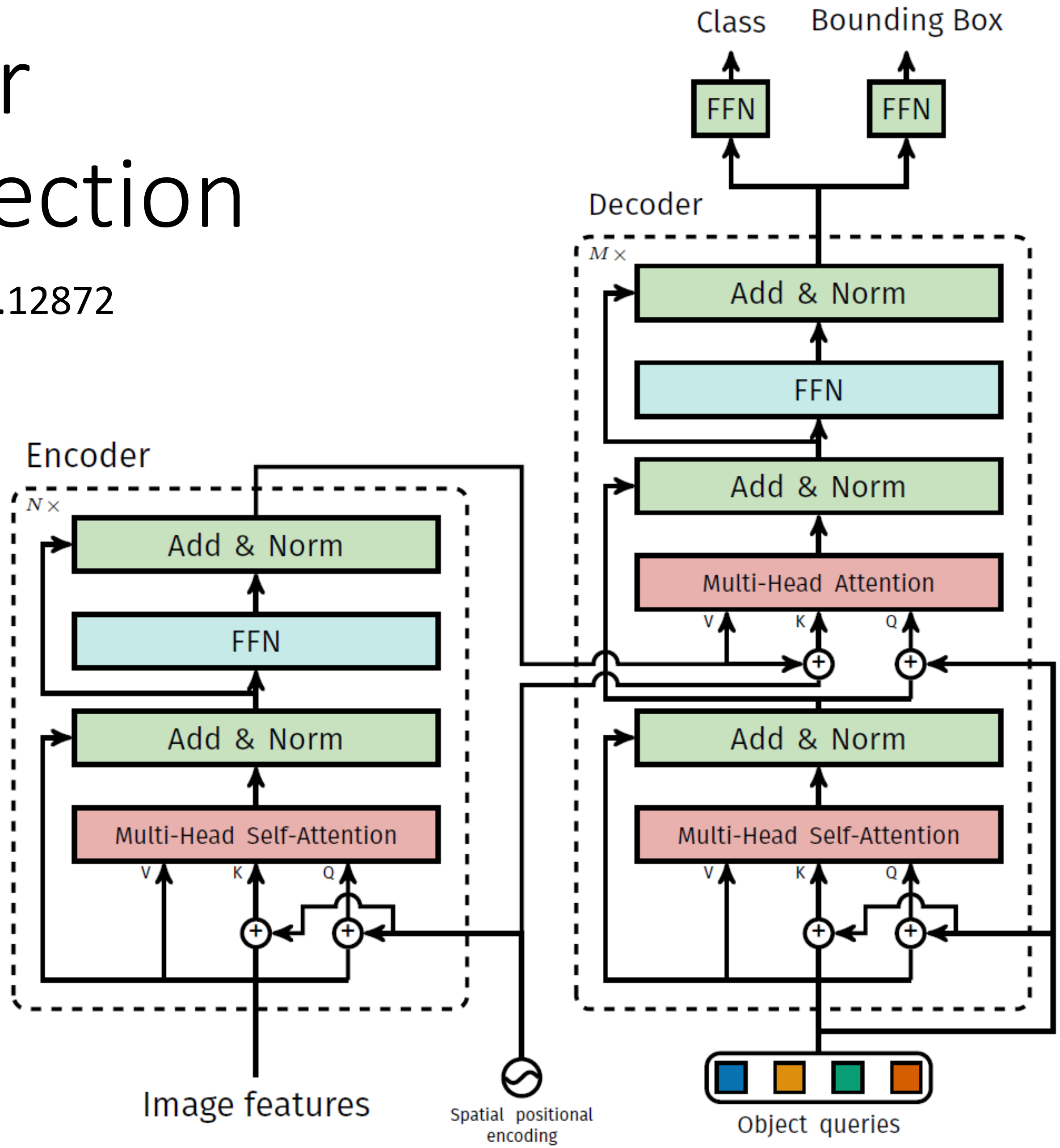




https://arxiv.org/abs/1909.03434
https://arxiv.org/abs/1707.05495

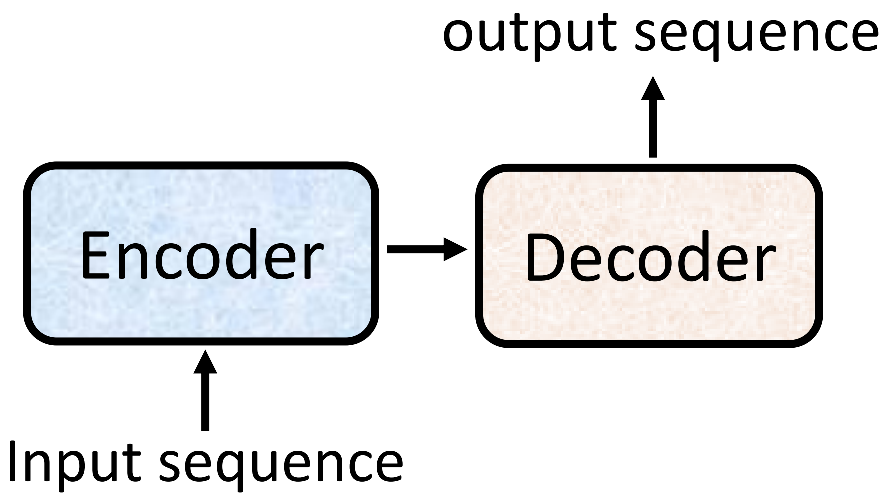
# Seq2seq for Object Detection

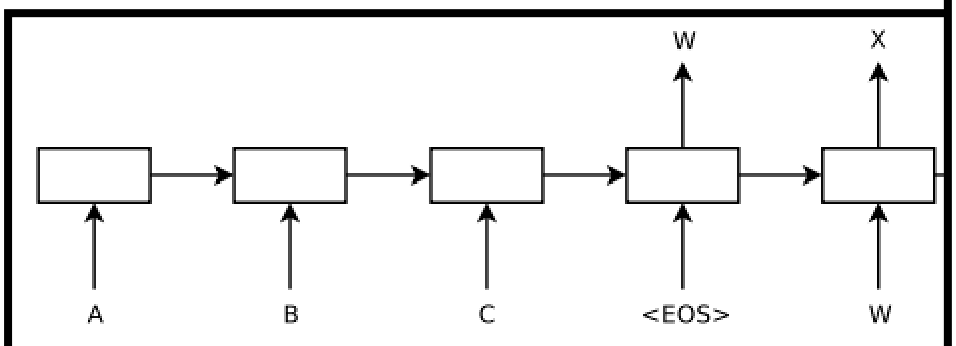
https://arxiv.org/abs/2005.12872



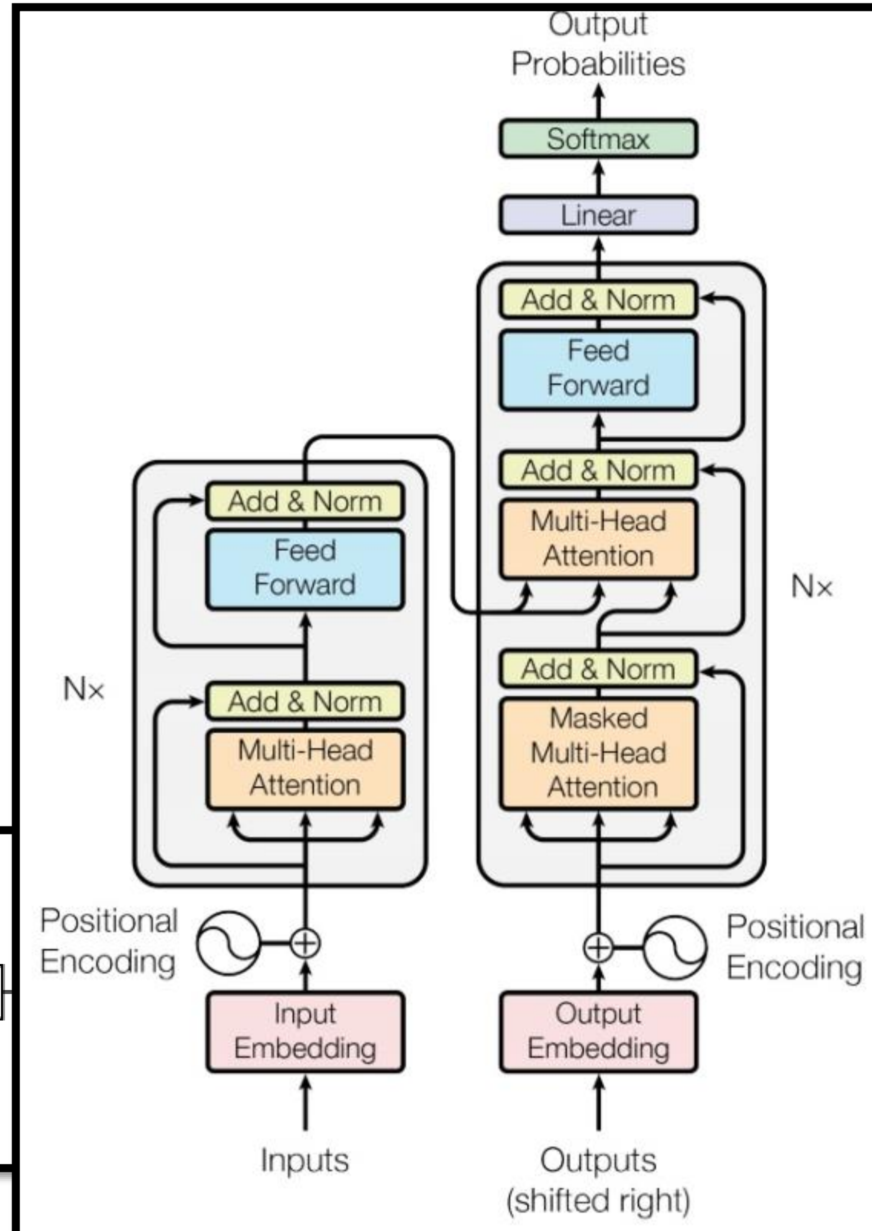



# Seq2seq



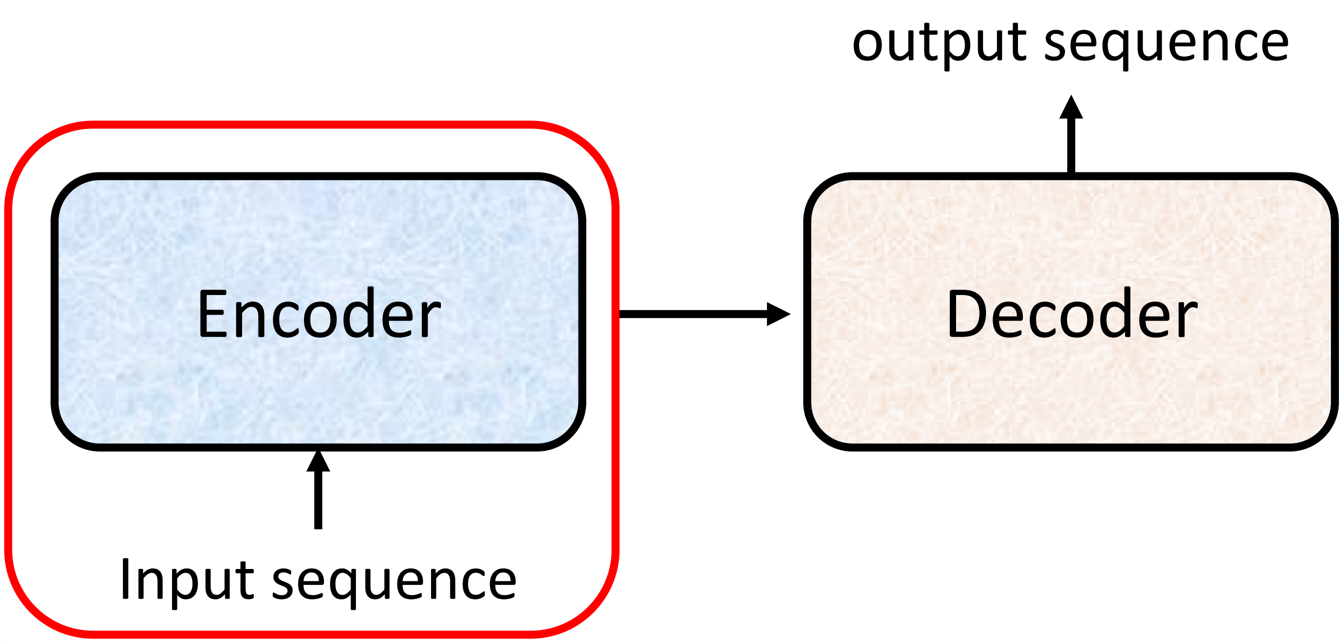




Sequence to Sequence Learning with Neural Networks
https://arxiv.org/abs/1409.3215



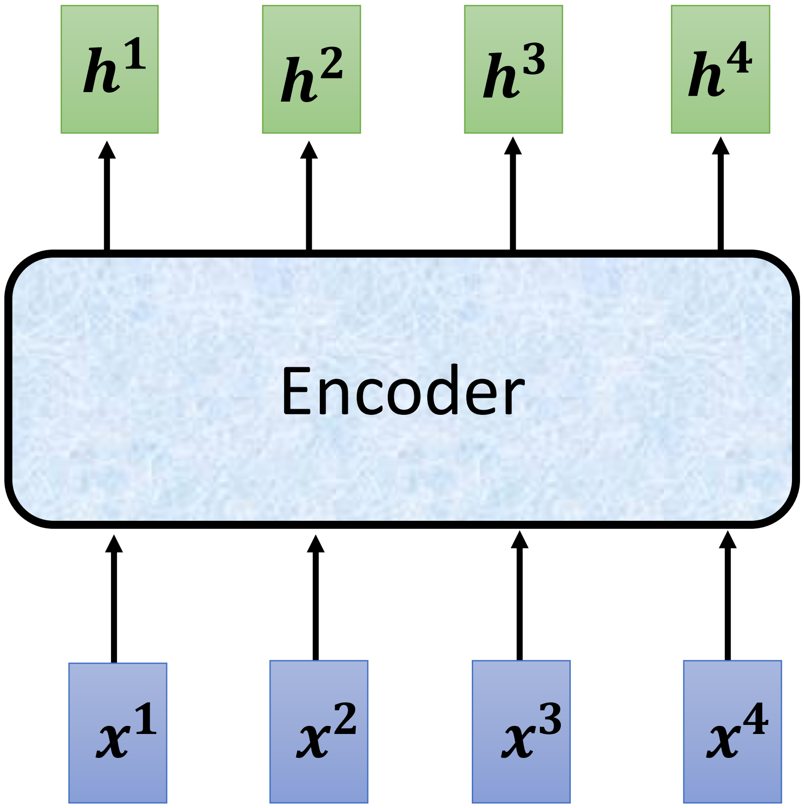

Transformer
https://arxiv.org/abs/1706.03762

# Encoder

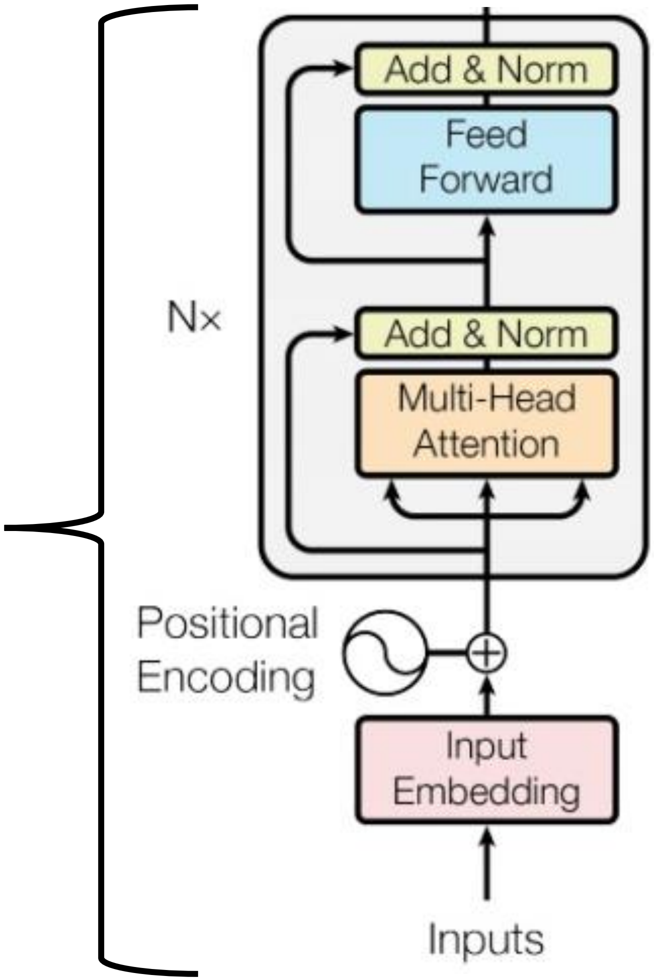


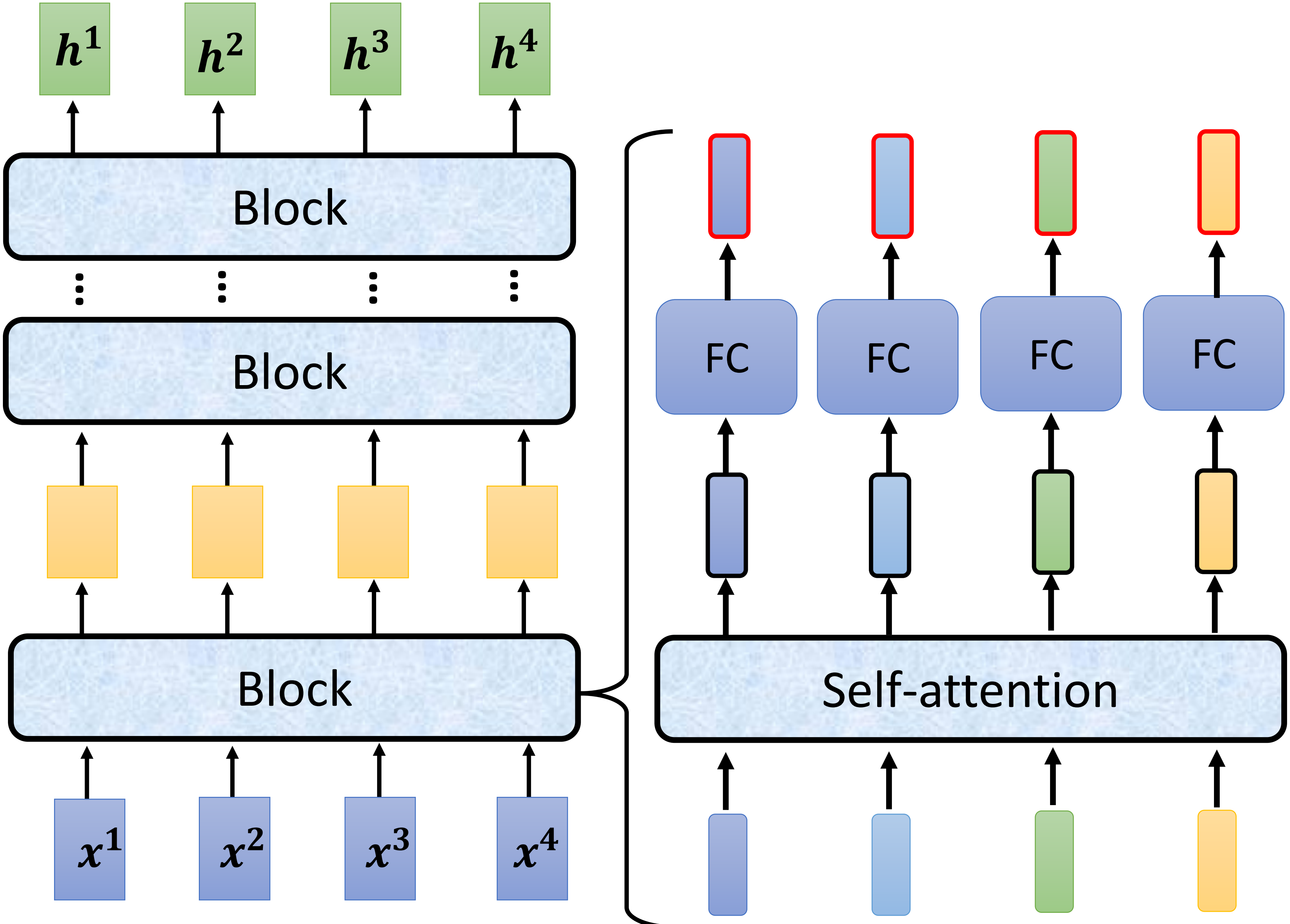
# Encoder

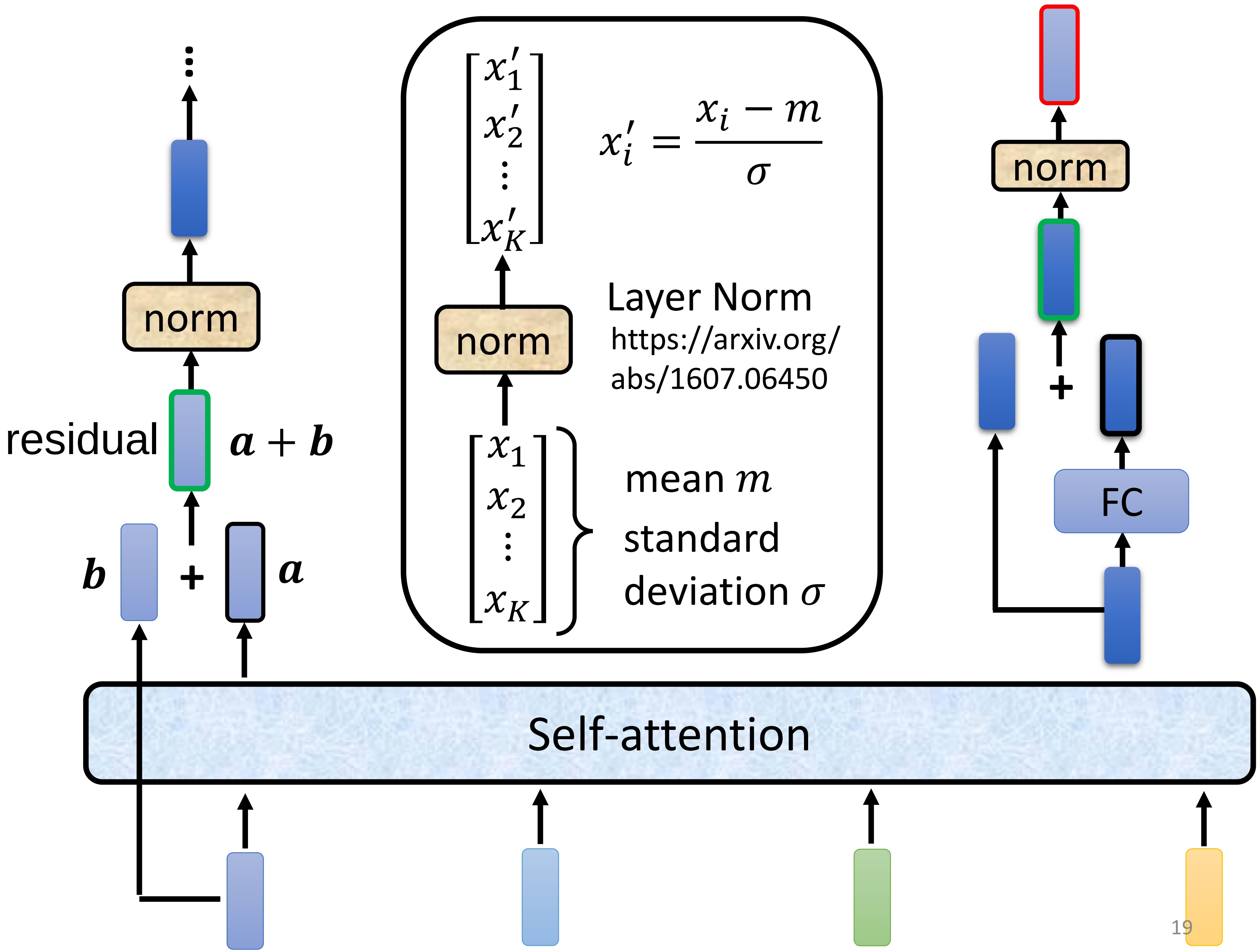
You can use **RNN** or **CNN**.




Transformer's Encoder



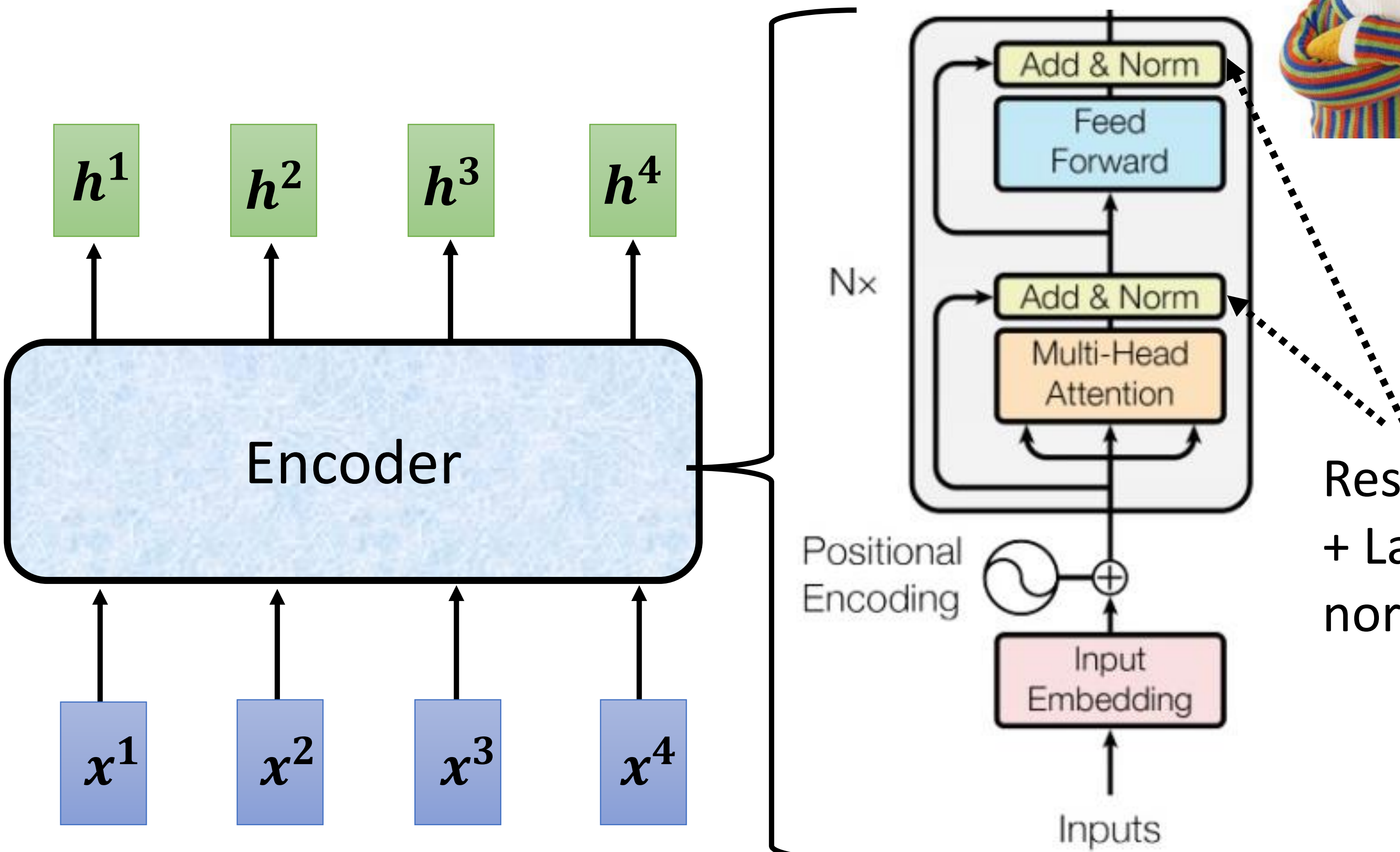

$h^1$
$h^2$
$h^3$
$h^4$
Block
Block
Block
FC
FC
FC
FC
Self-attention
$x^1$
$x^2$
$x^3$
$x^4$



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

BERT

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

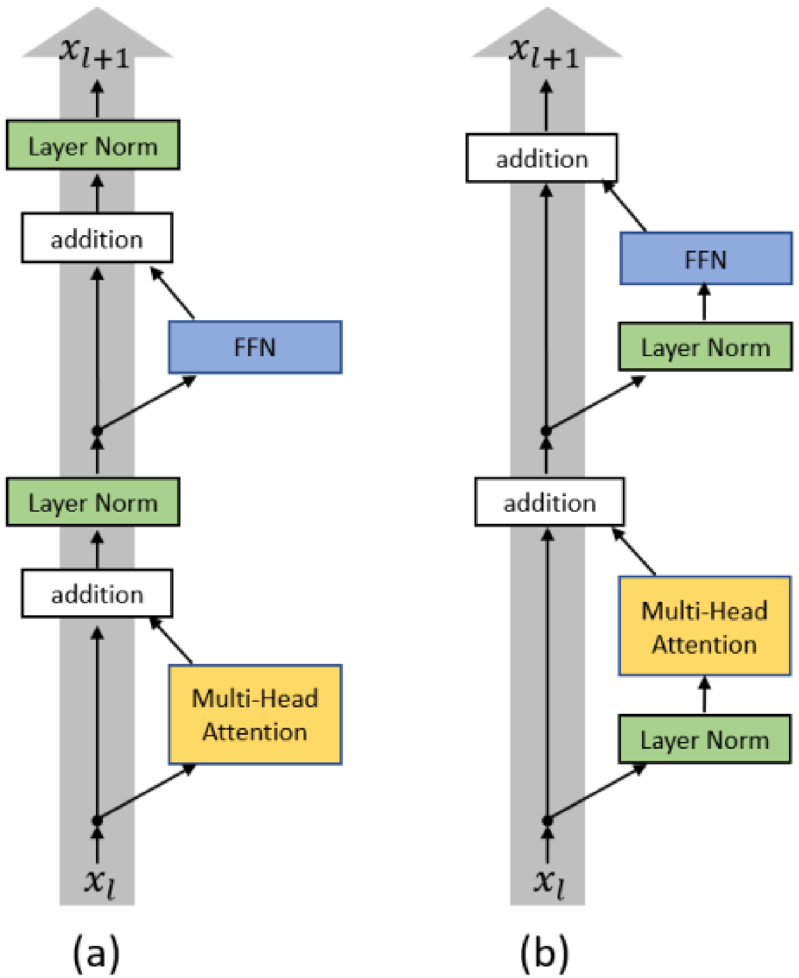



Residual + Layer norm

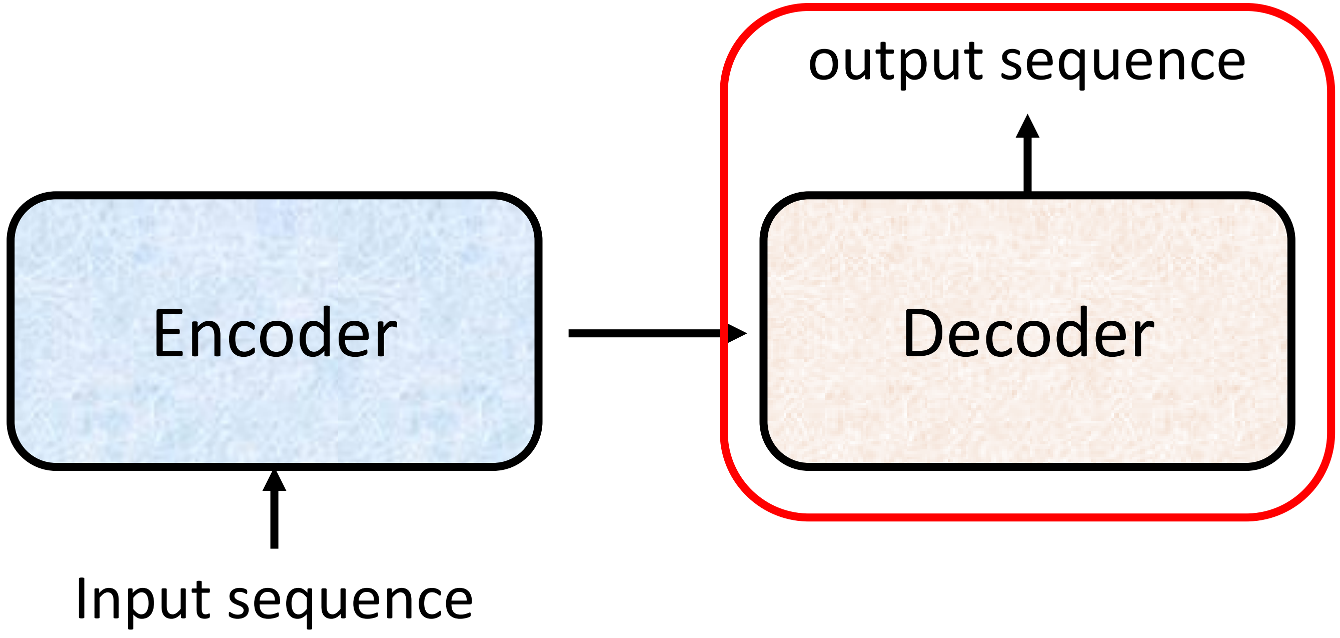
# To learn more ……

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

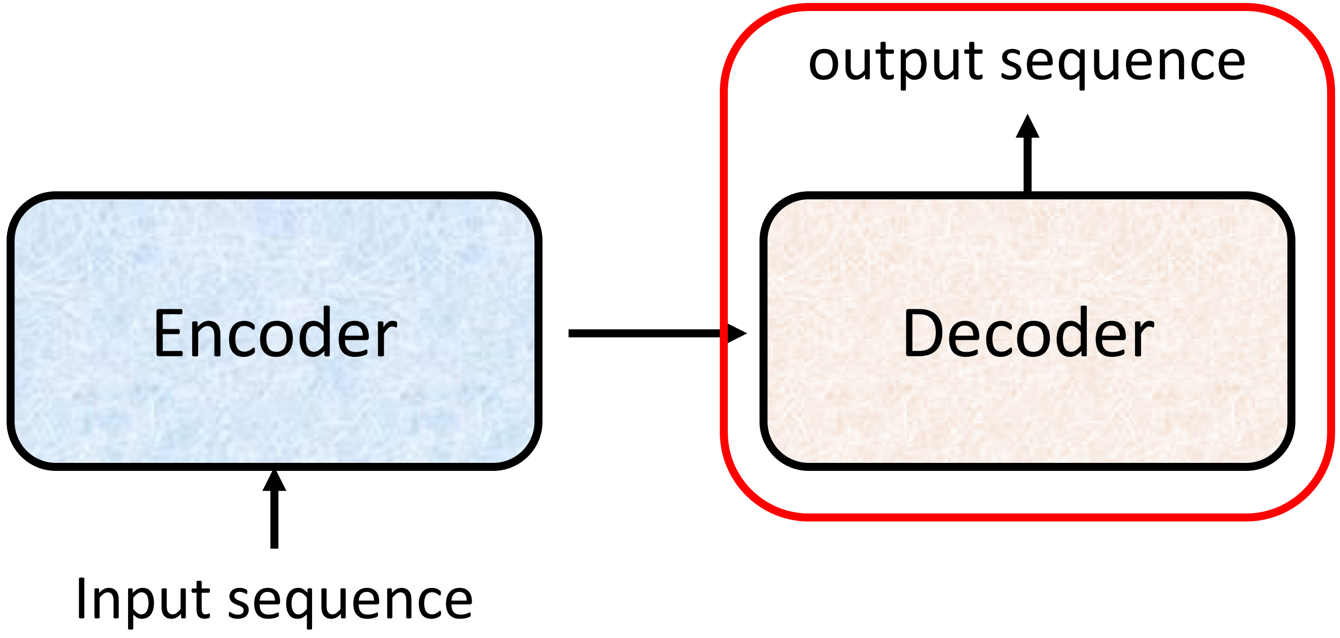
- PowerNorm: Rethinking Batch Normalization in Transformers
- https://arxiv.org/abs/2003.07845



# Decoder

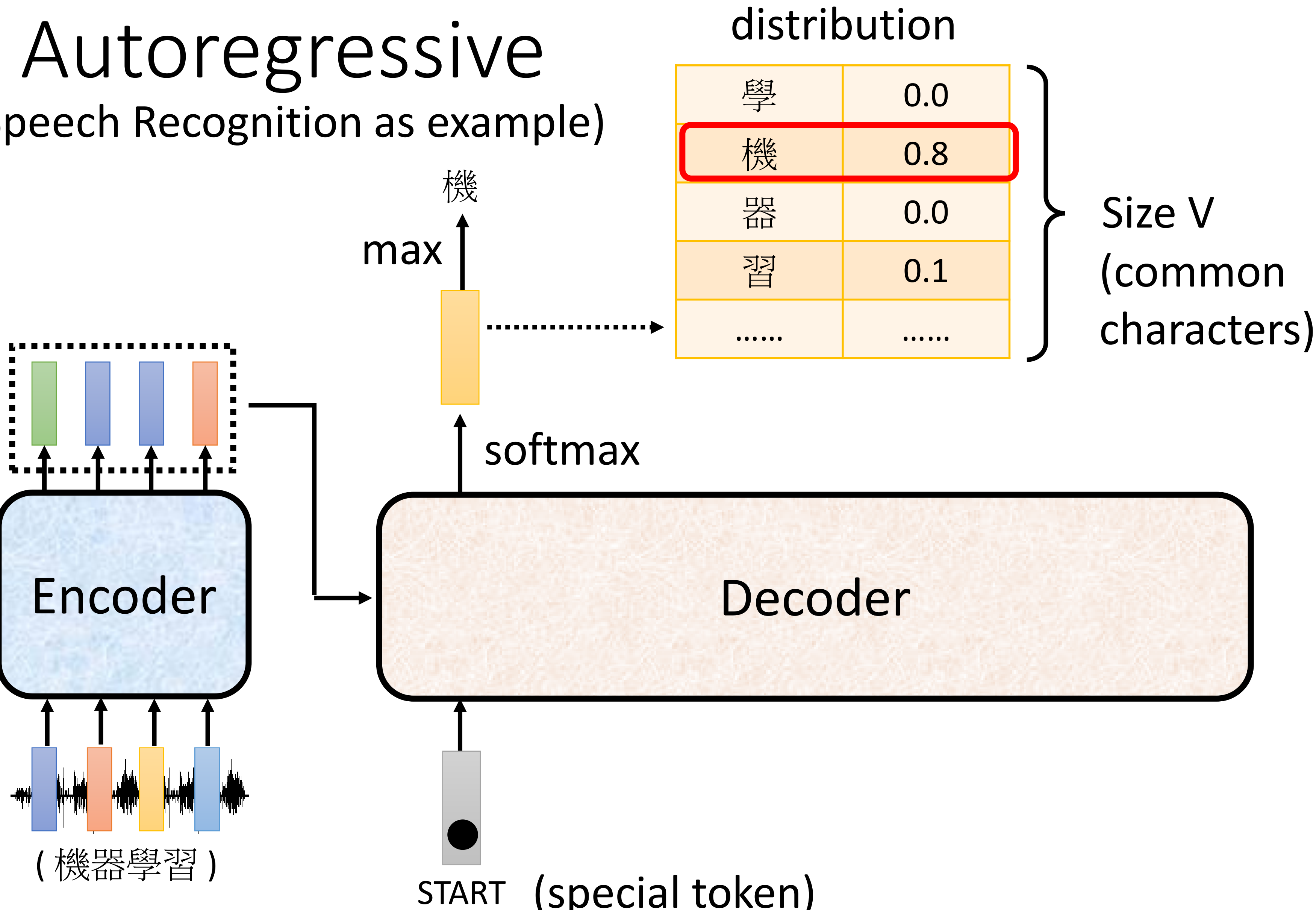


# Decoder
# – Autoregressive (AT)

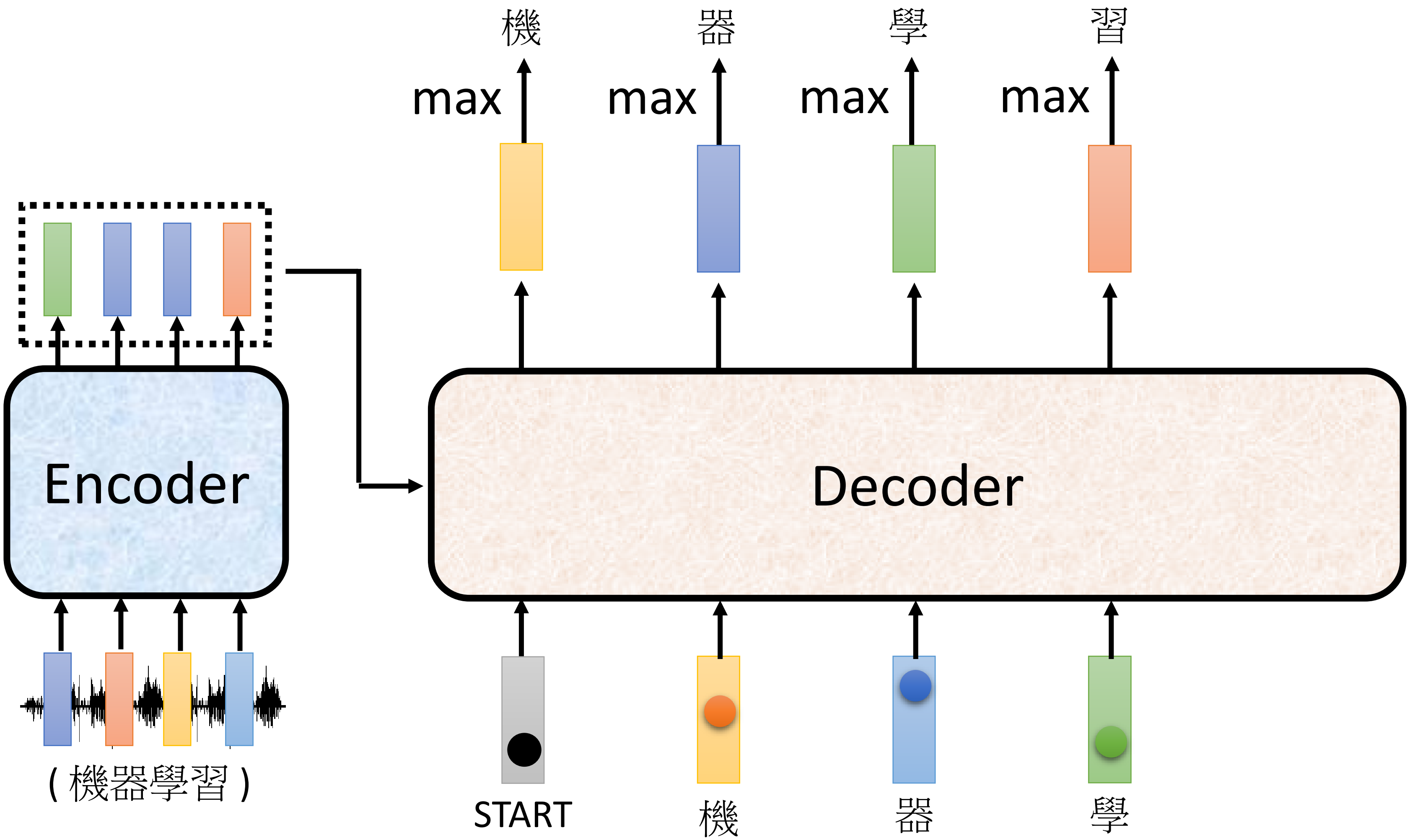


# Autoregressive

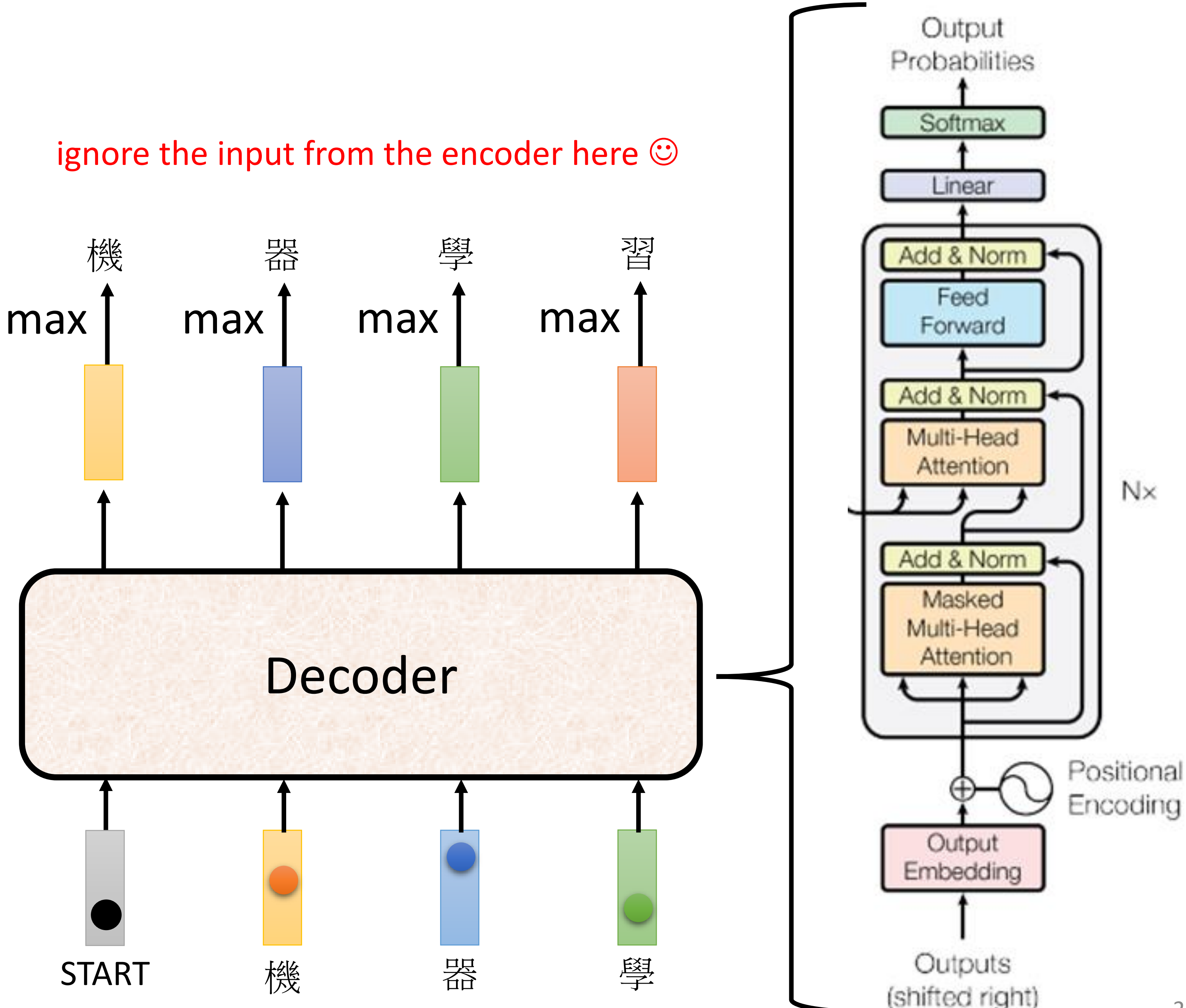
(Speech Recognition as example)

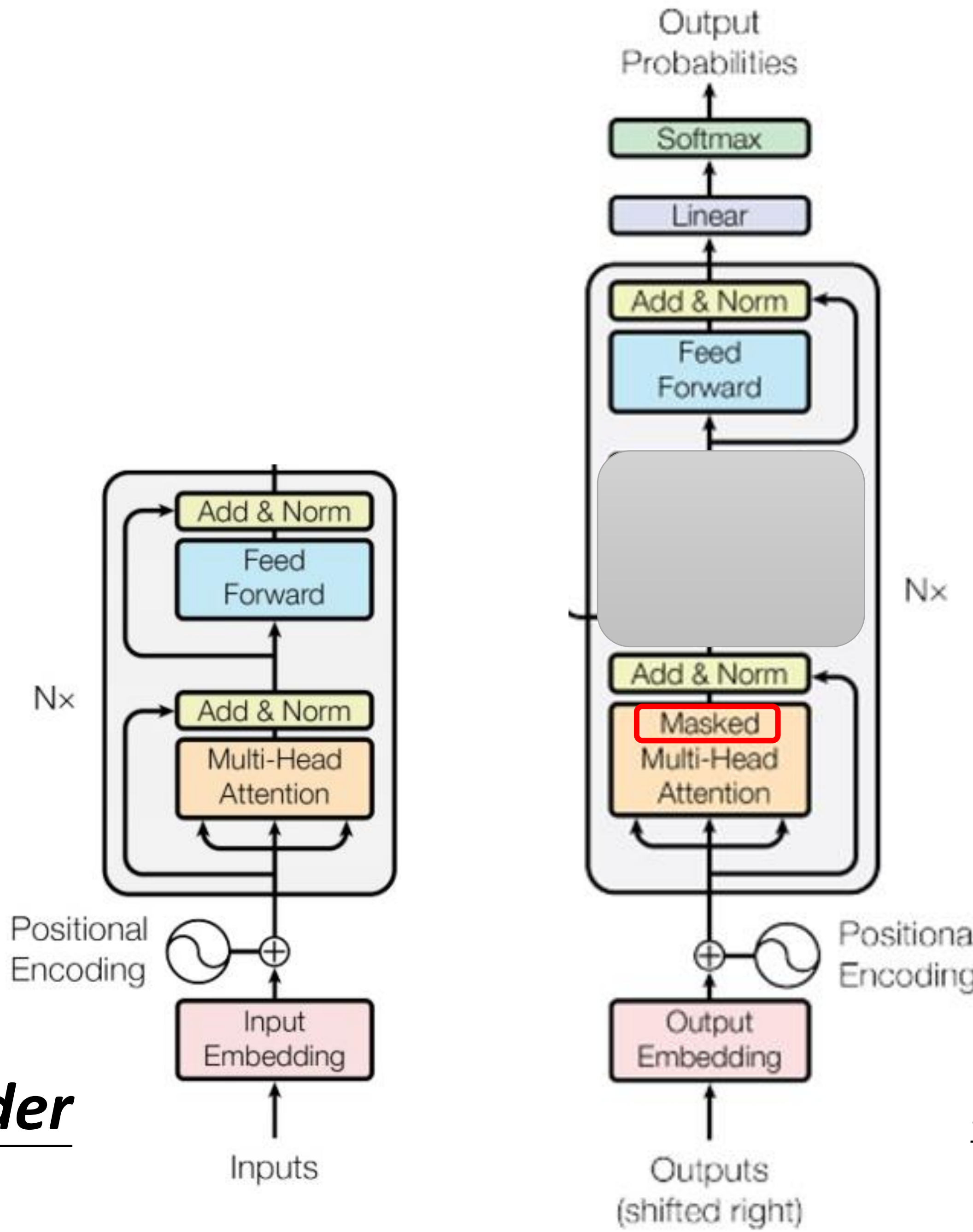


# Autoregressive



ignore the input from the encoder here ☺

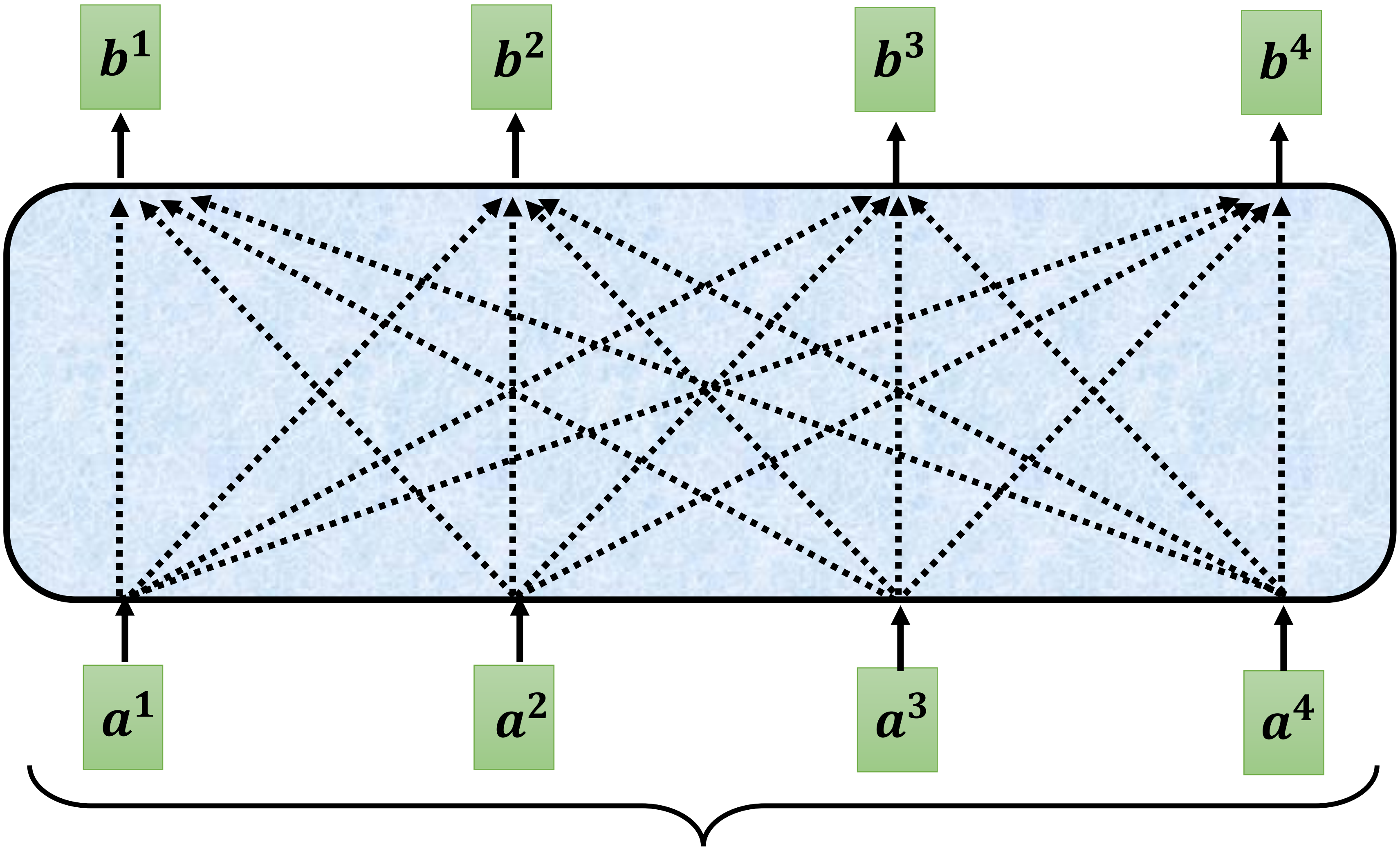





***Encoder***

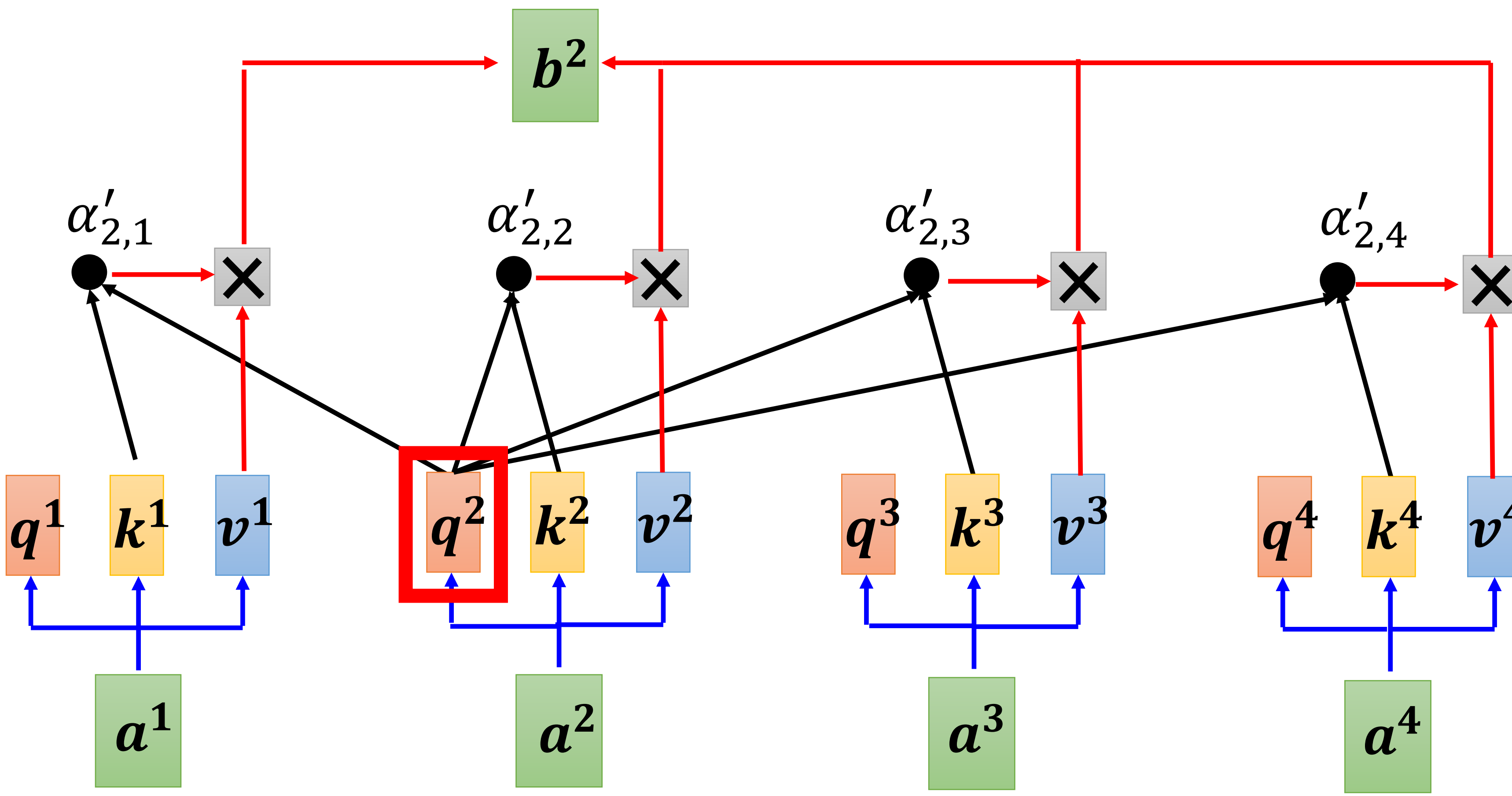
***Decoder***

# *Self-attention* → *Masked Self-attention*




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

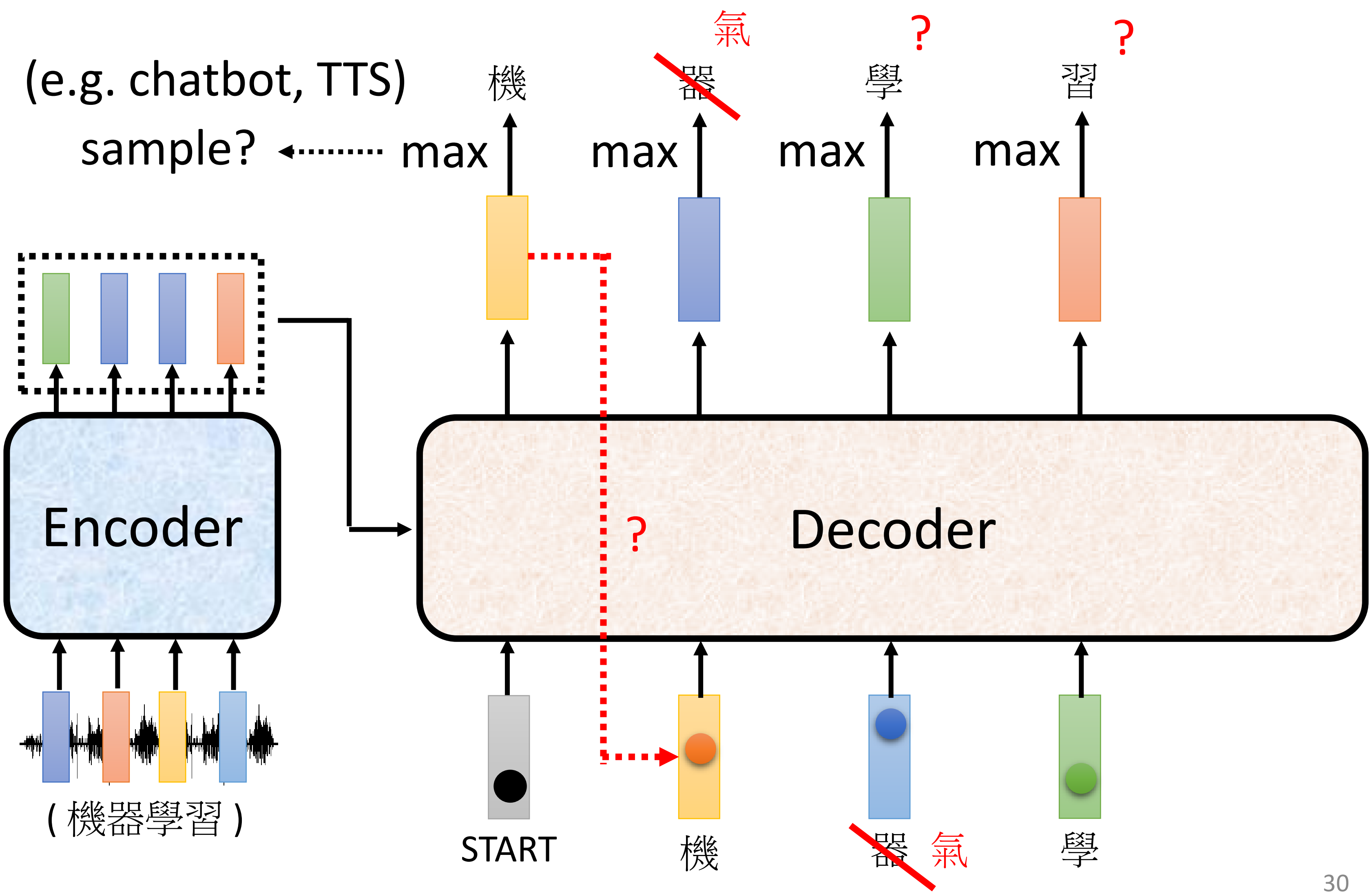
# *Self-attention* → *Masked Self-attention*




Why masked? Consider how does decoder work

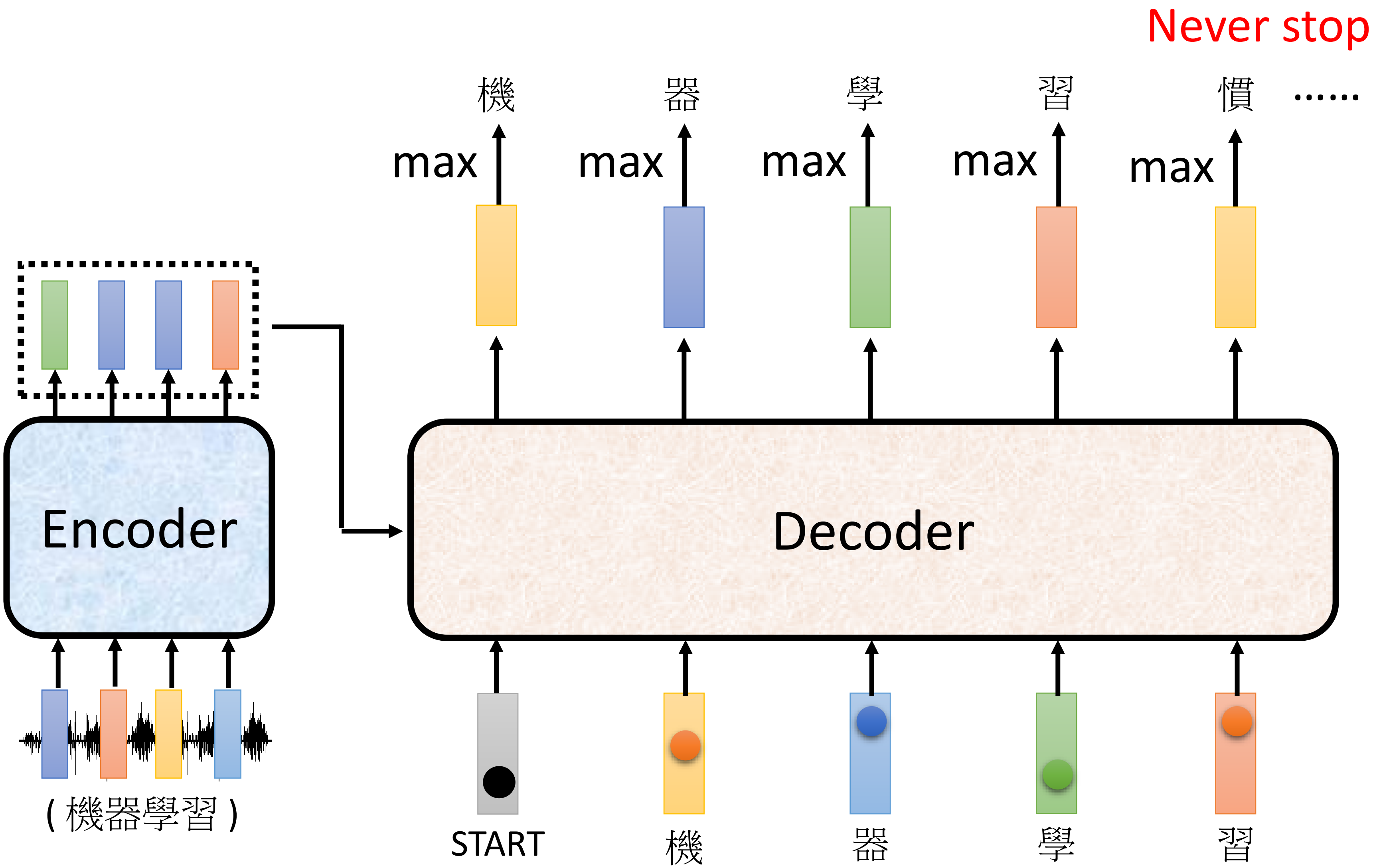
# Autoregressive

Question?
- Error propagation
- Distribution as input?

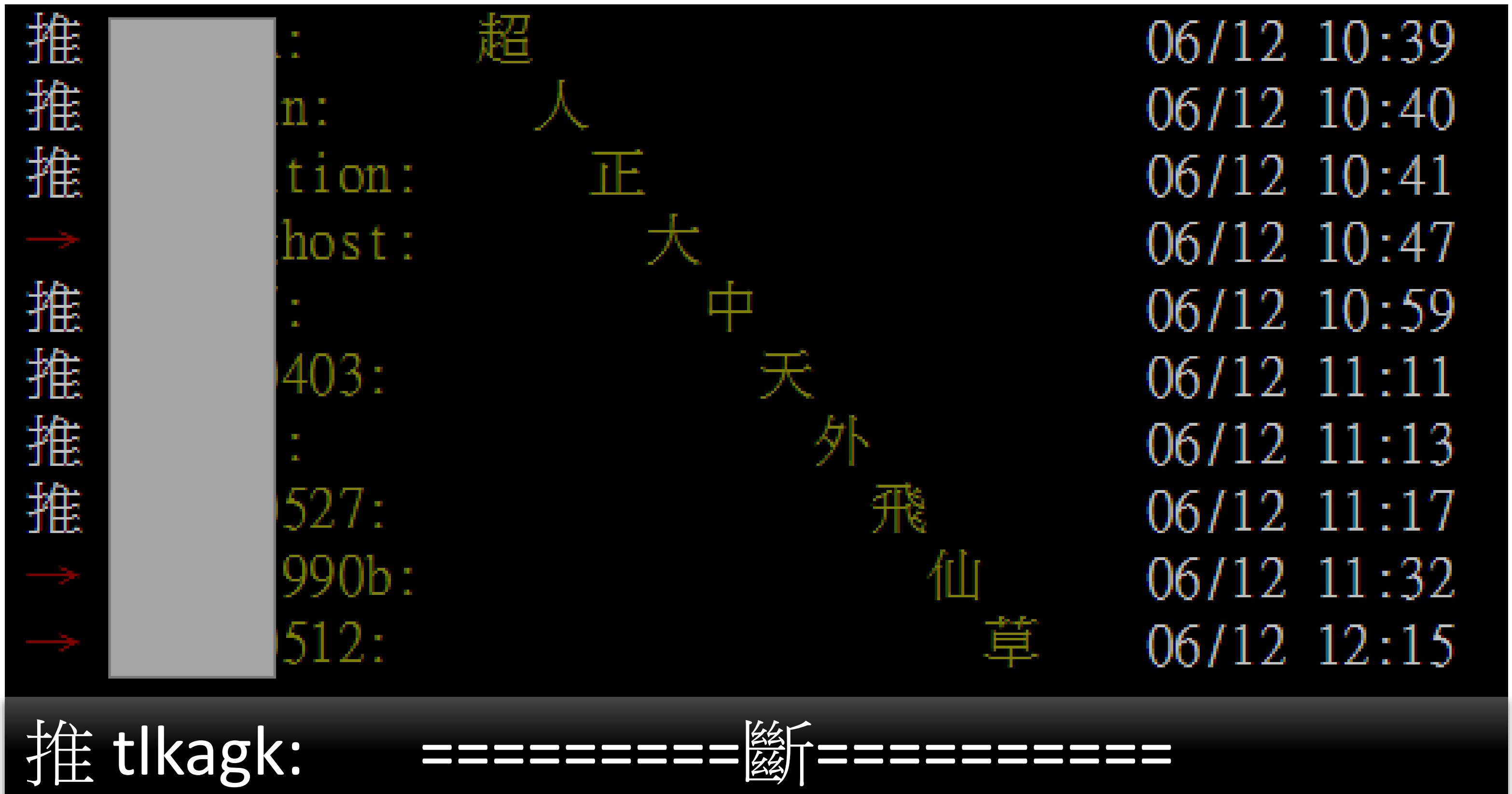


# Autoregressive

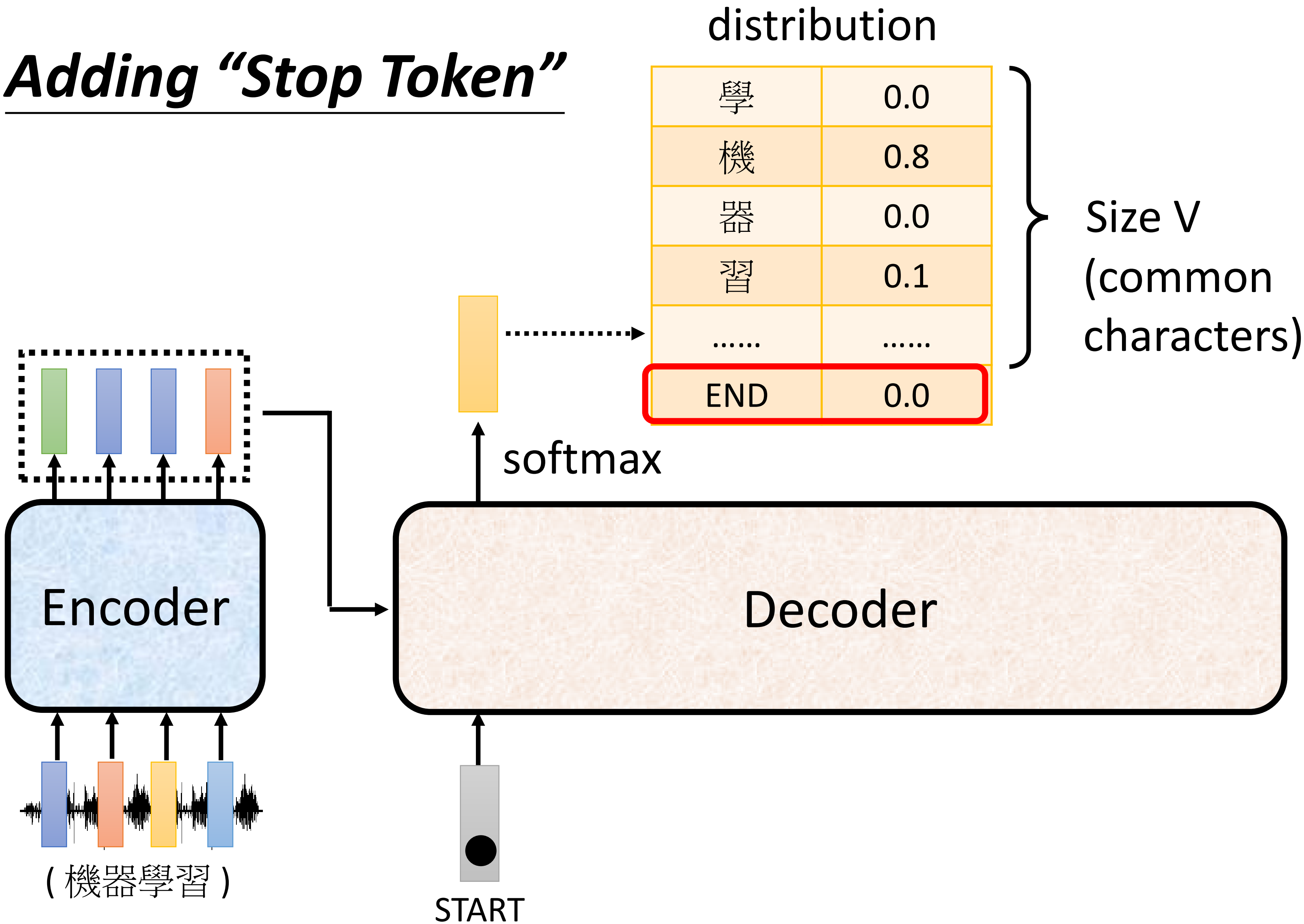
We do not know the correct output length.



# 推文接龍 (Tweet Solitaire)

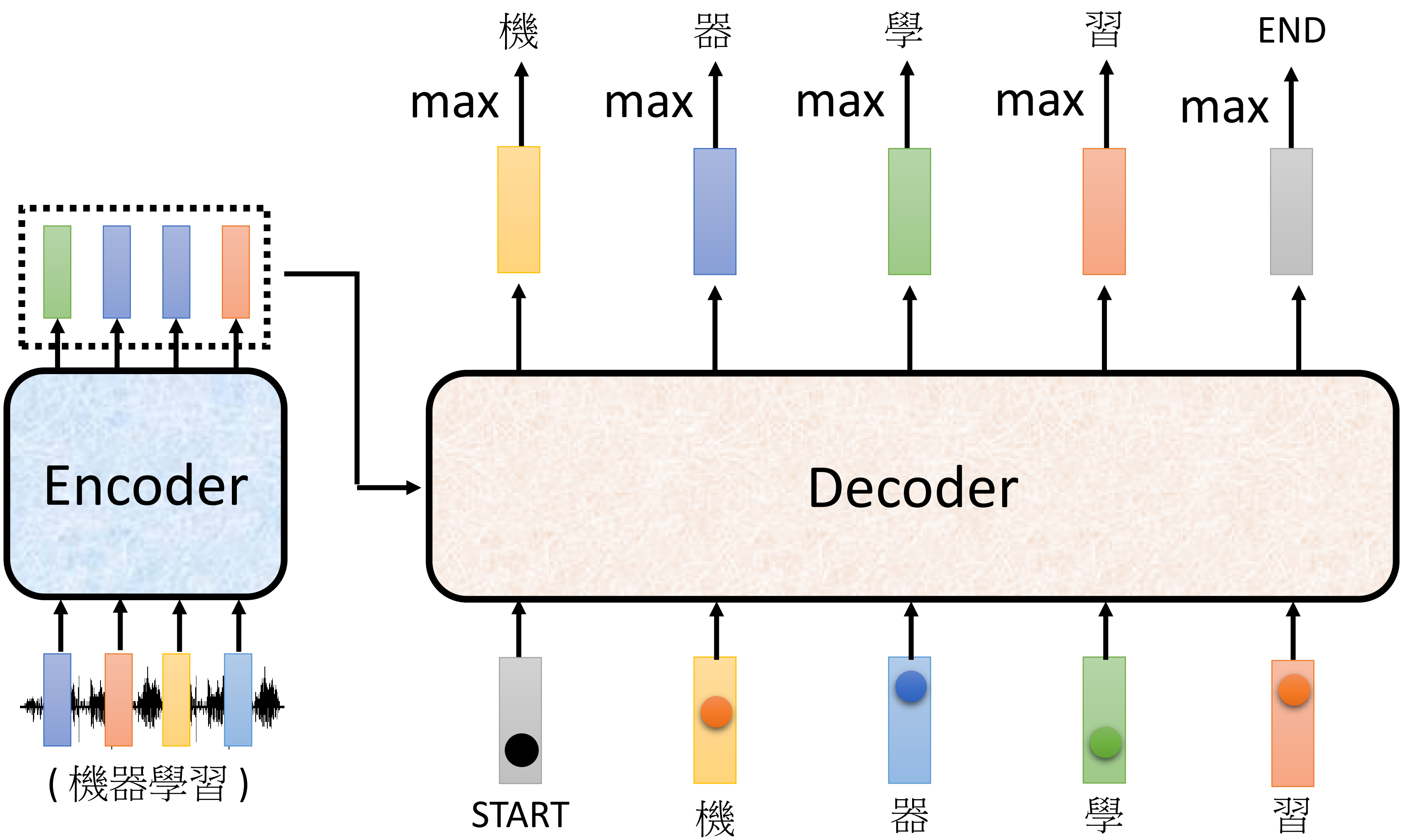


# ***Adding "Stop Token"***

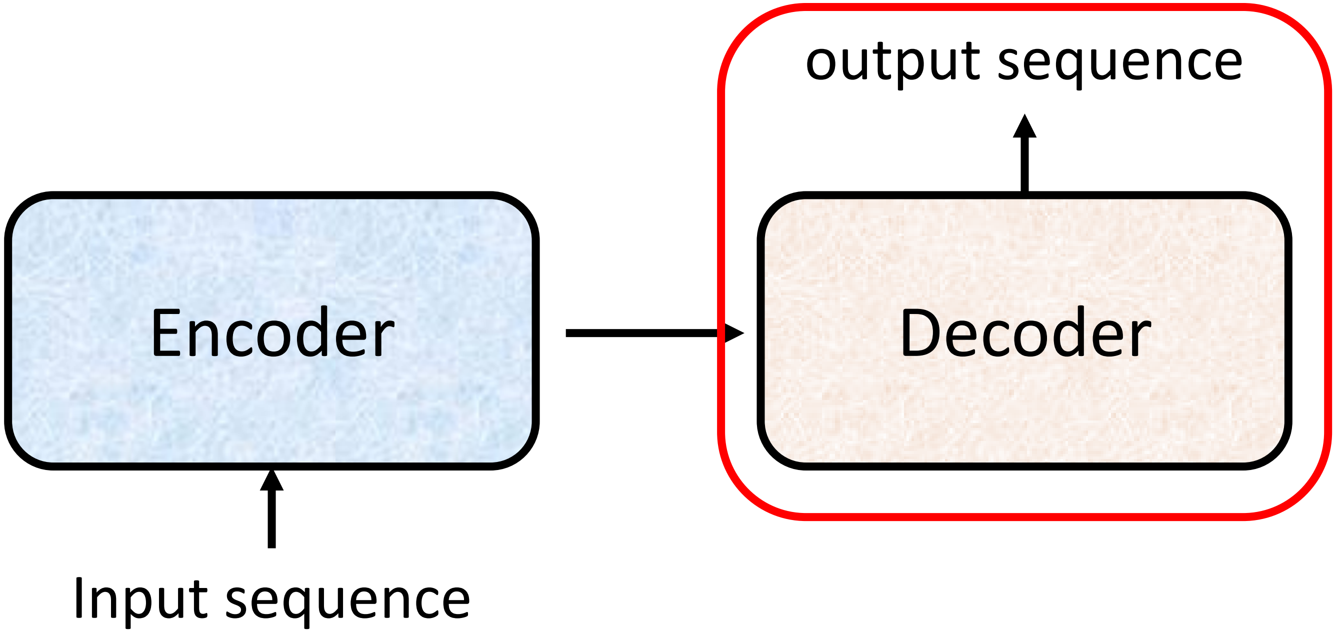


# Autoregressive

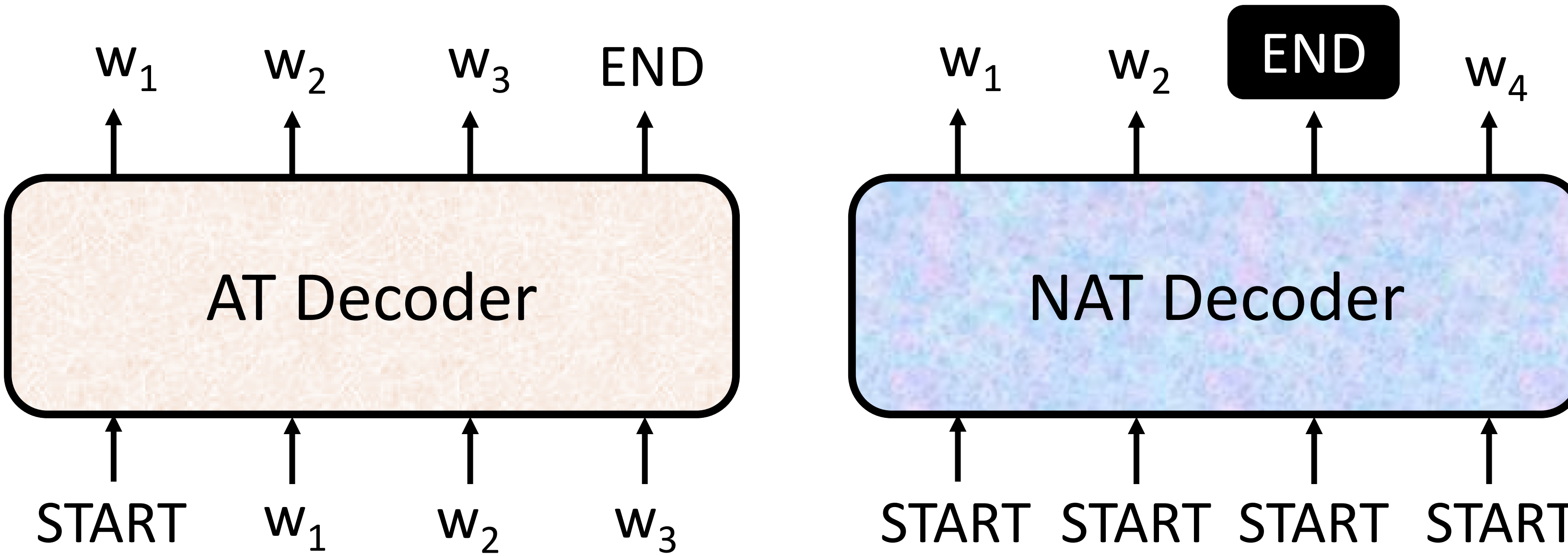
Stop at here!



# 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)

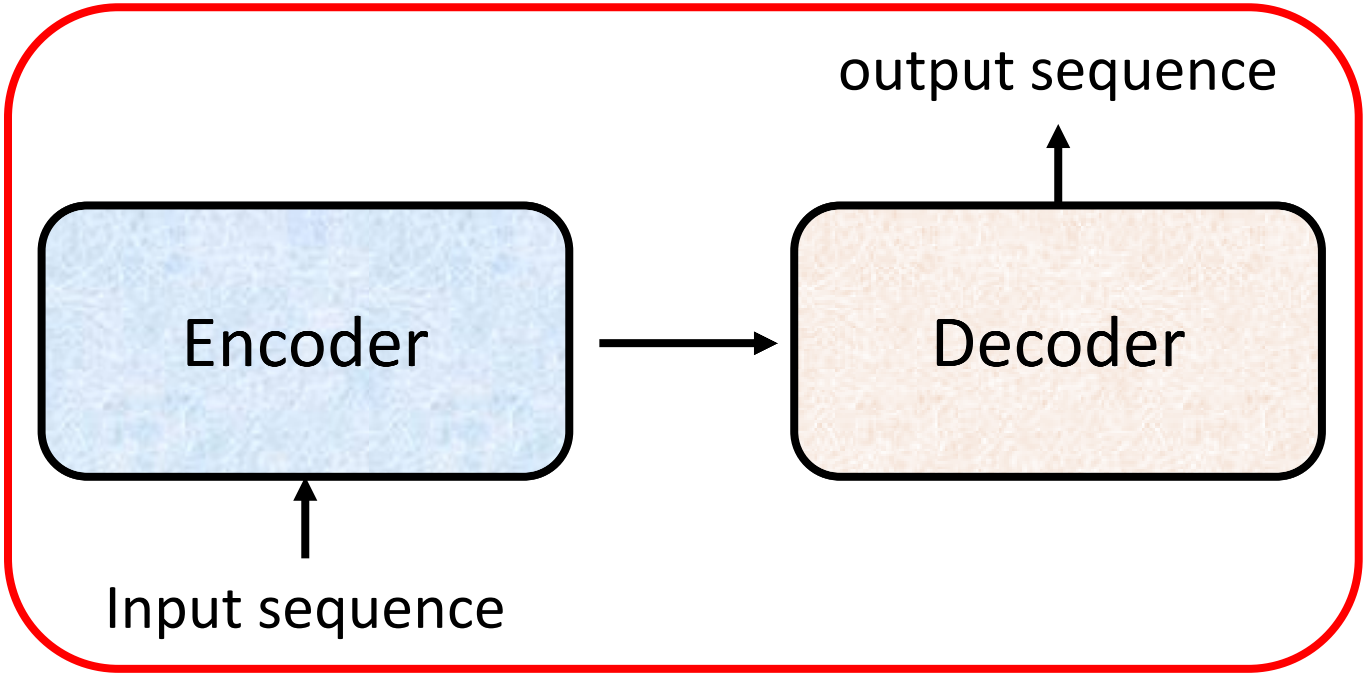
# To learn more ……

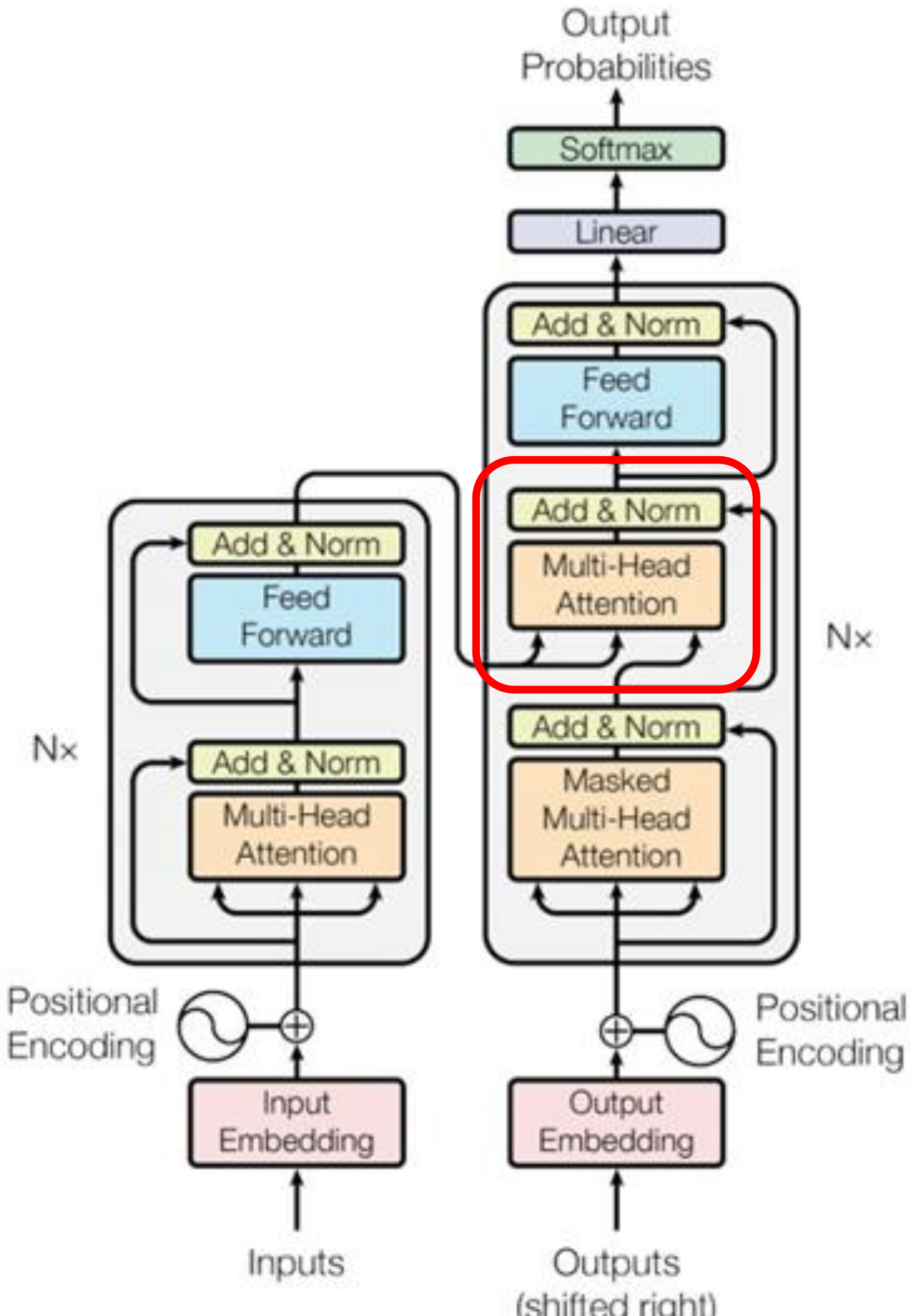
https://youtu.be/jvyKmU4OM3c

(in Mandarin)


Q&A

# Encoder-Decoder




Output
Probabilities
Softmax
Linear
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Nx
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Nx
Add & Norm
Masked
Multi-Head
Attention
Positional
Encoding
Positional
Encoding
Input
Embedding
Output
Embedding
Inputs
Outputs
(shifted right)

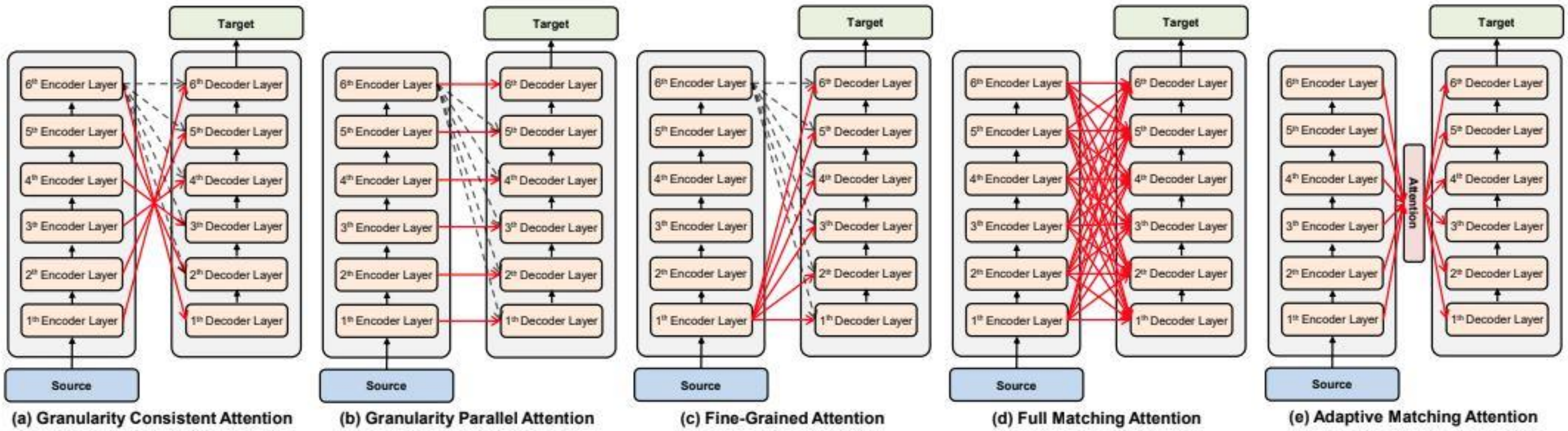
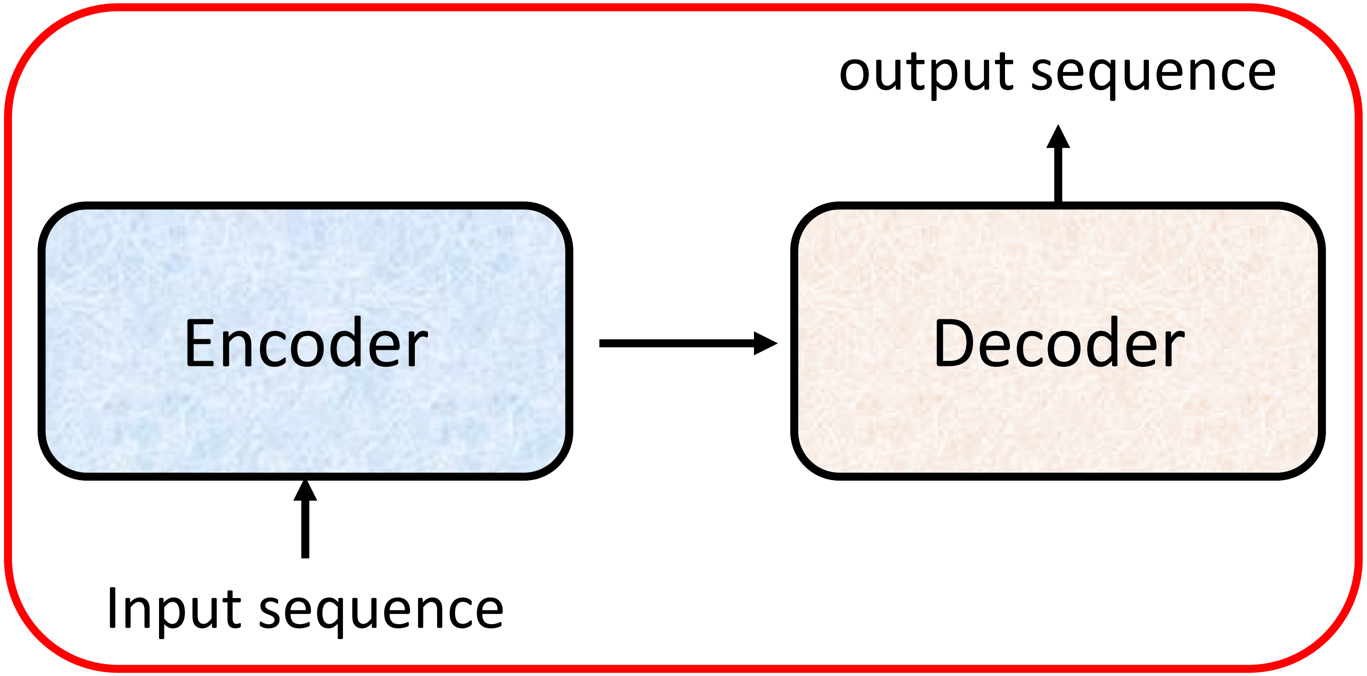
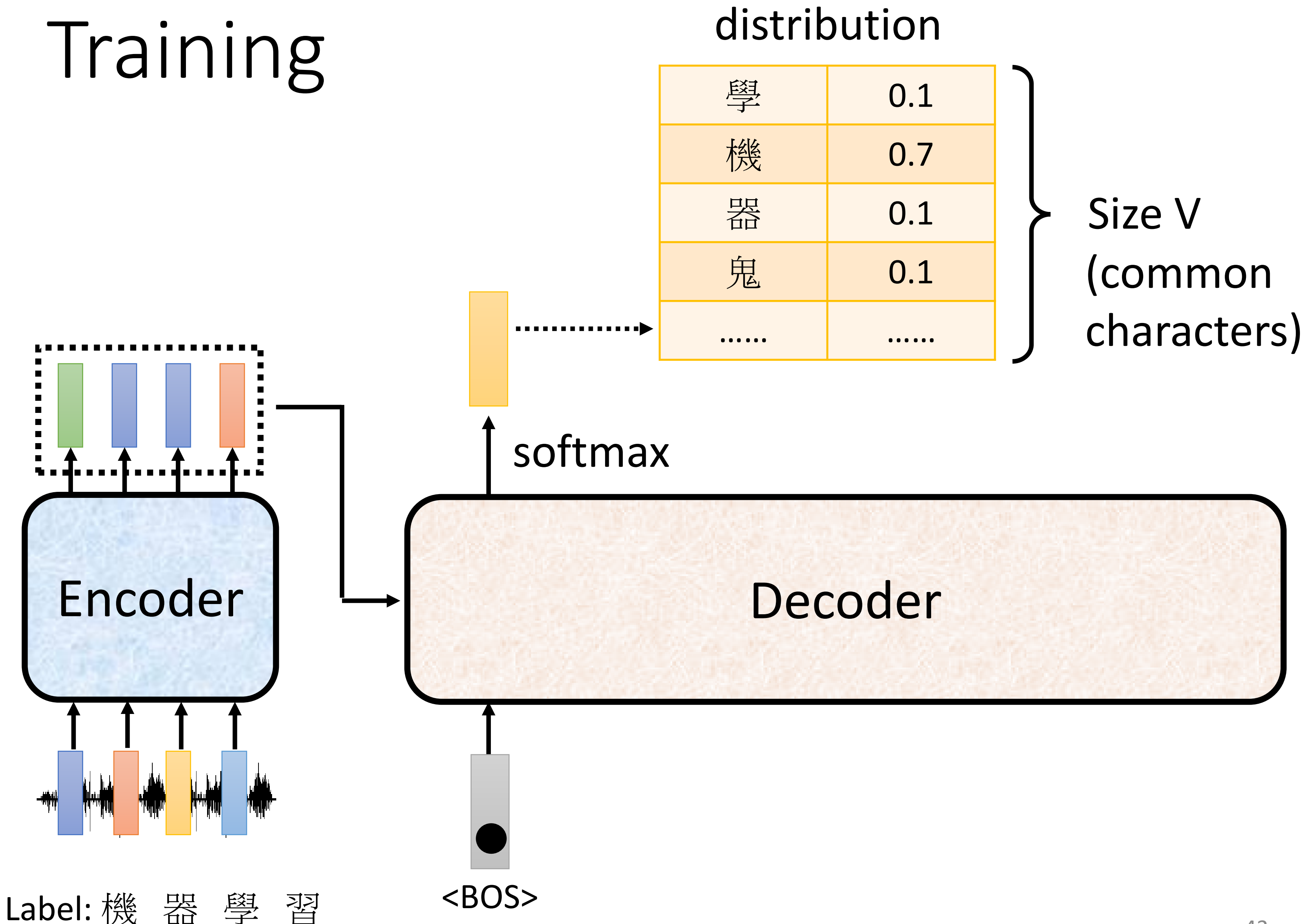


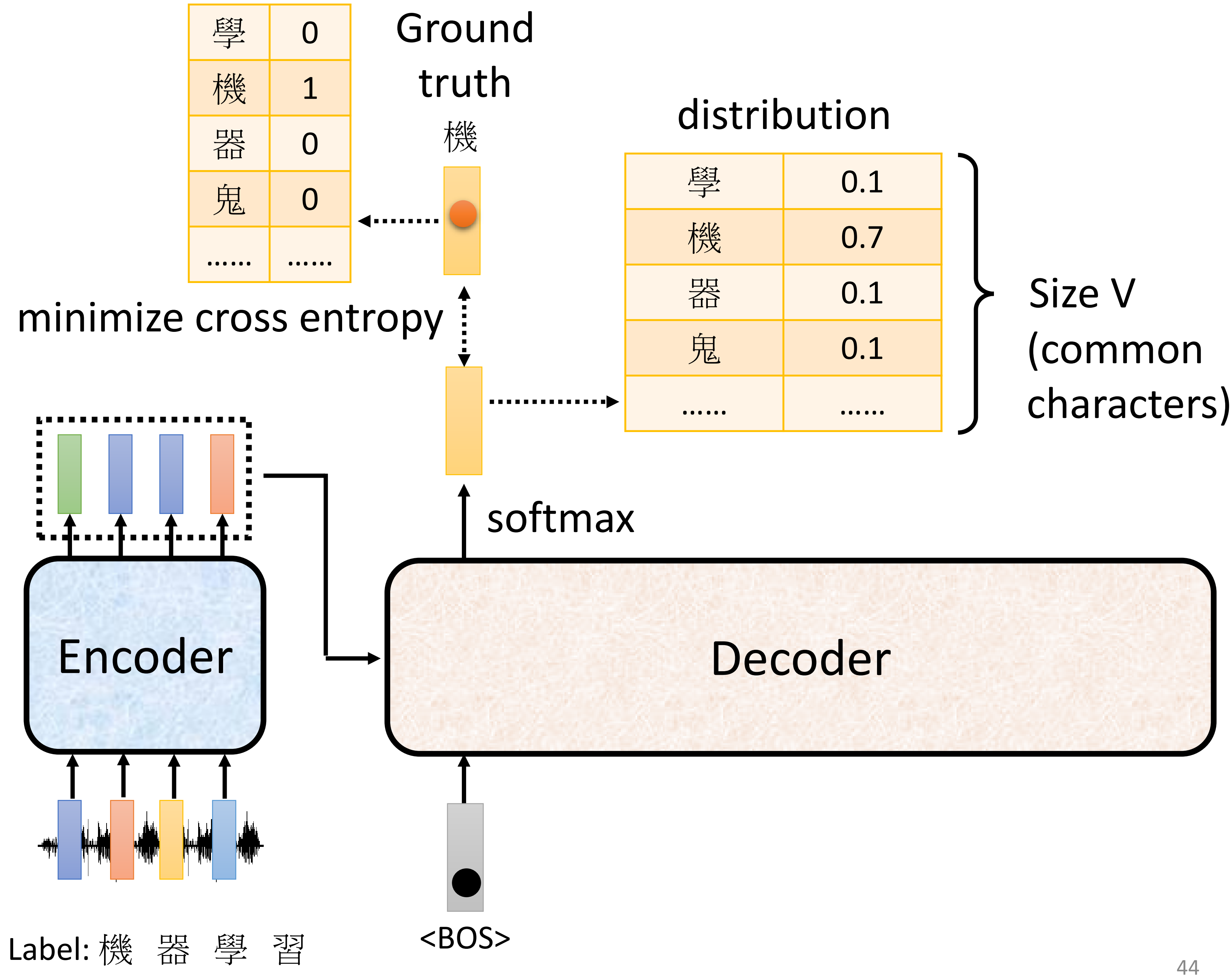

Figure 2: We present the proposal on Transformer with various strategies for routing the source representations: (a) Granularity Consistent Attention; (b) Granularity Parallel Attention; (c) Fine-Grained Attention; (d) Full Matching Attention; (e) Adaptive Matching Attention. The dashed lines represent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

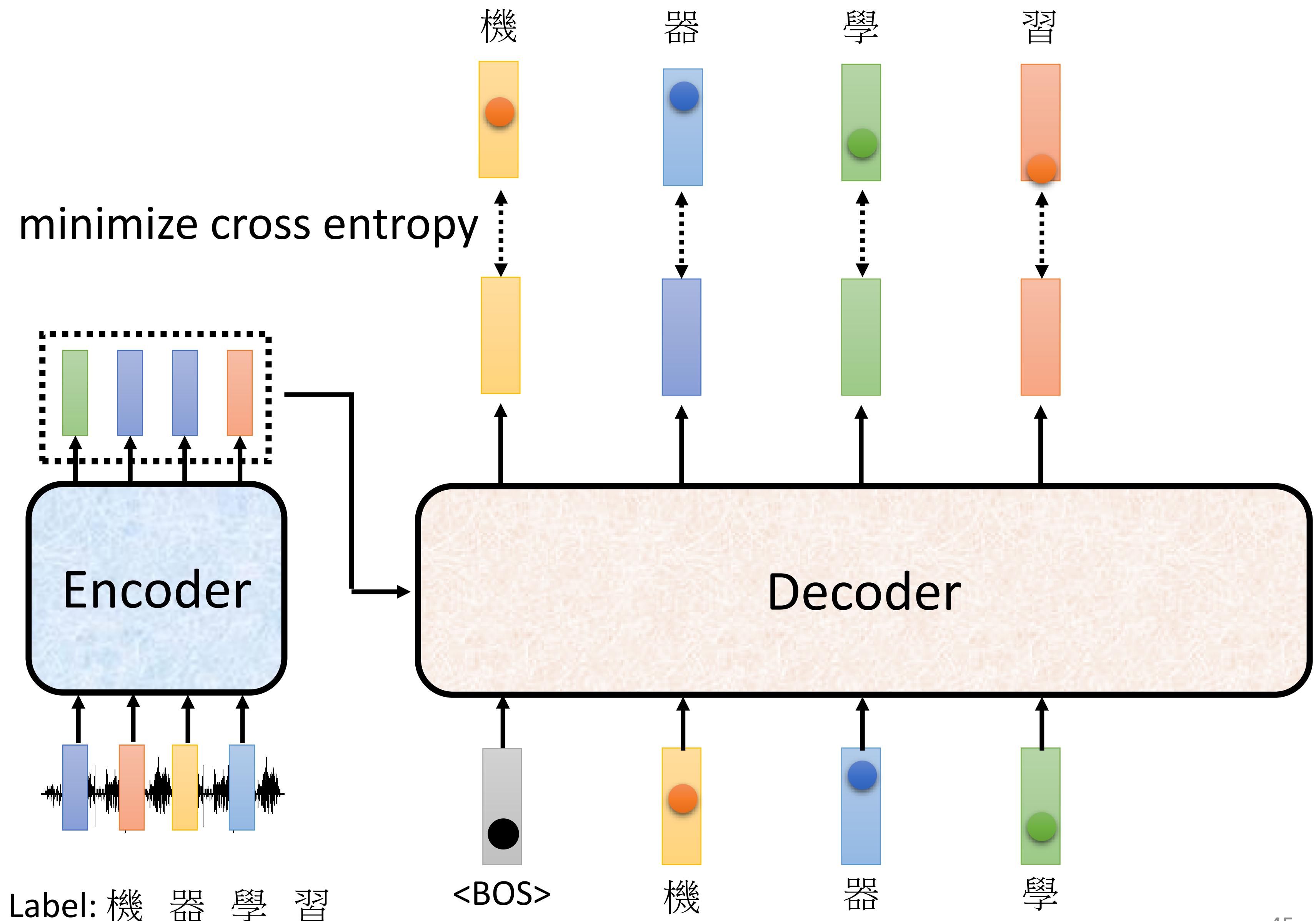
# Training



# 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
minimize cross entropy
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



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