



Transformer

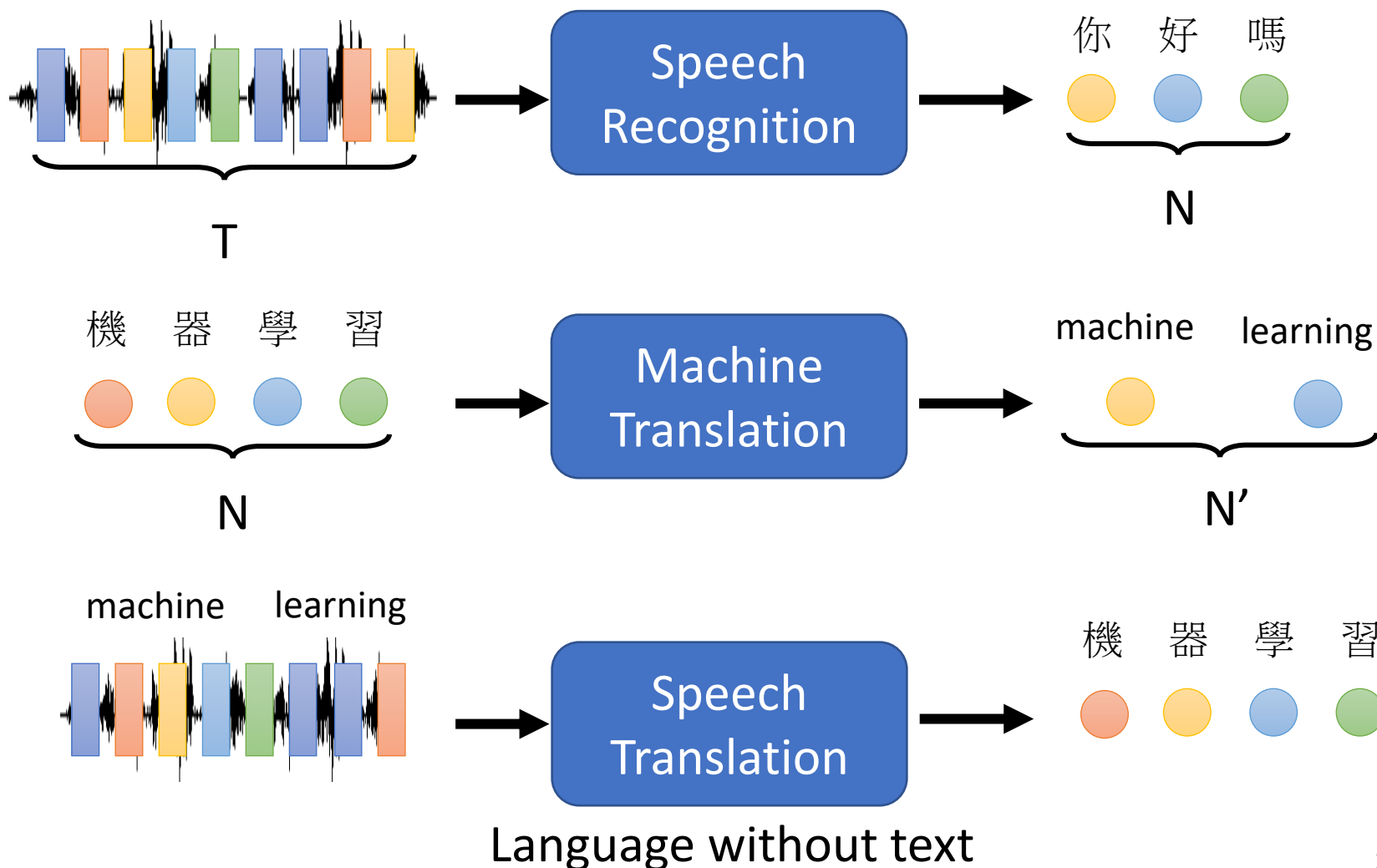
李宏毅

Hung-yi Lee

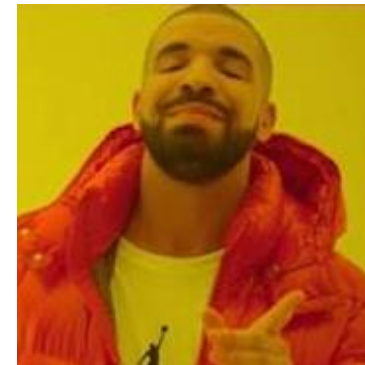
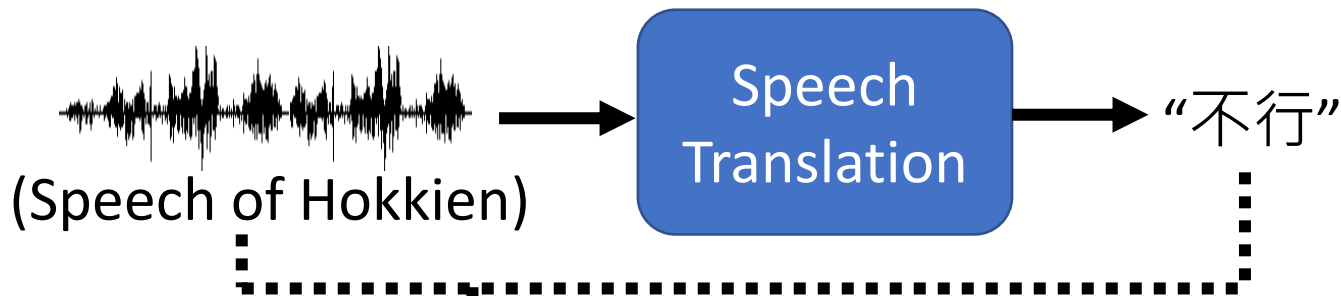
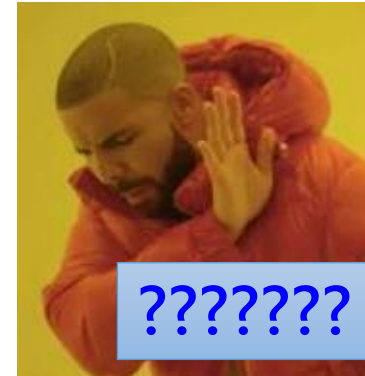
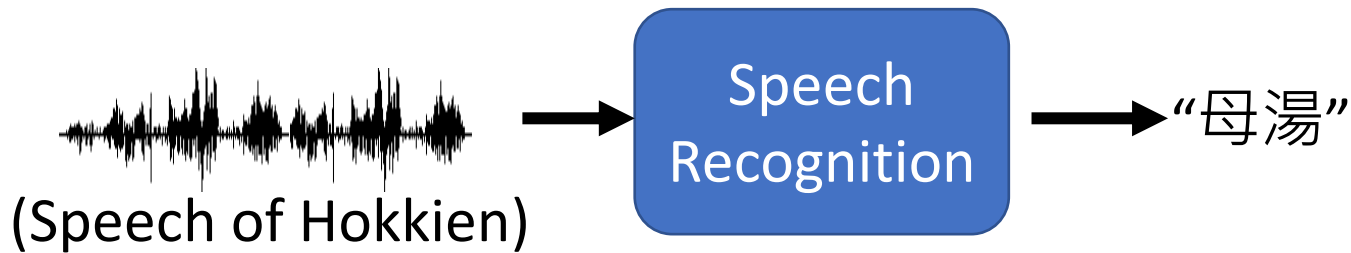
Sequence-to-sequence (Seq2seq)

Input a sequence, output a sequence

The output length is determined by model.



Hokkien (閩南語、台語)



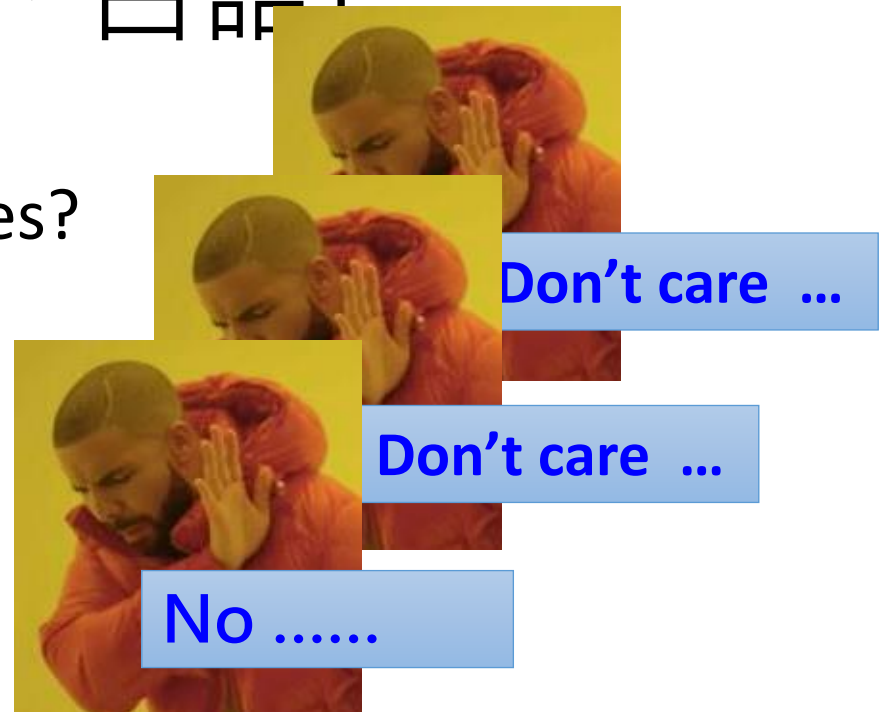
Local soap operas (鄉土劇) on YouTube
(Speech of Hokkien, Chinese subtitle)

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)

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

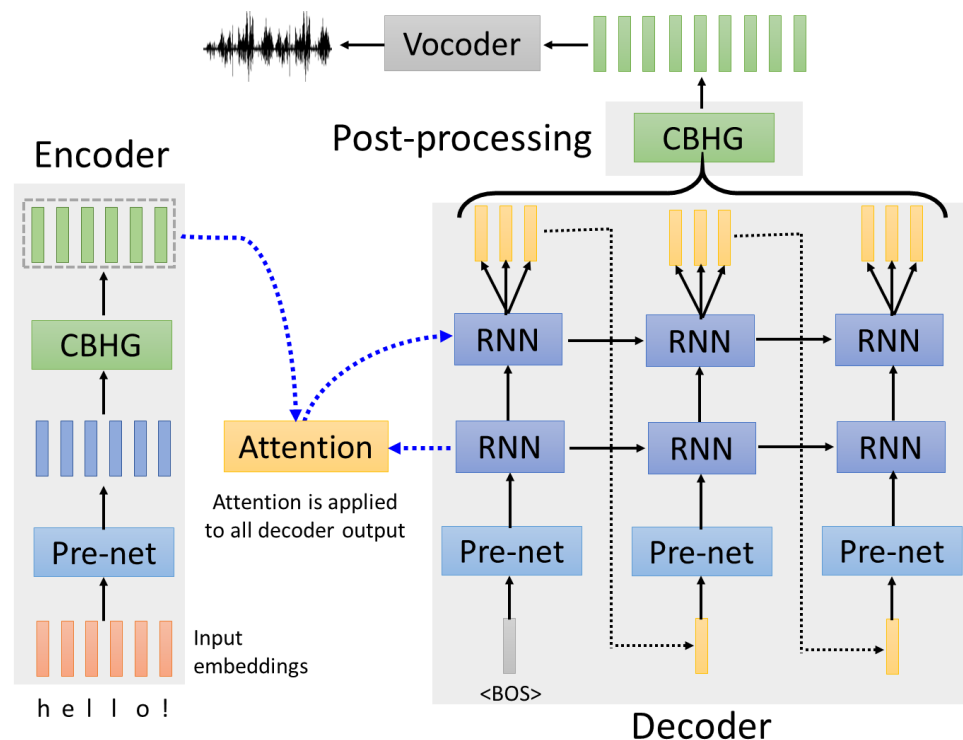
Answer:我拜託廠長了

Text-to-Speech (TTS) Synthesis

感謝張凱為同學提供實驗結果

Taiwanese Speech Synthesis

Source of data: 台灣嬌聲2.0



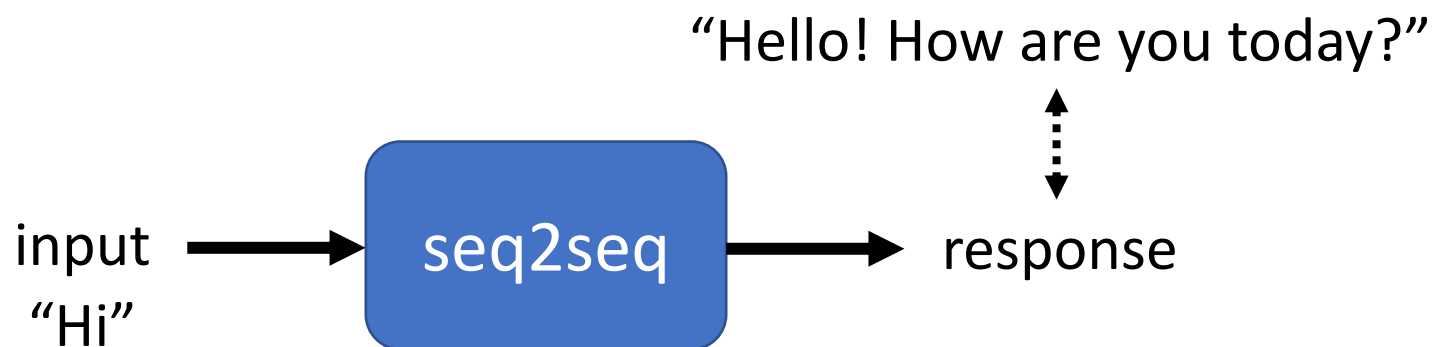
歡迎來到台大語音處理實驗室



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.

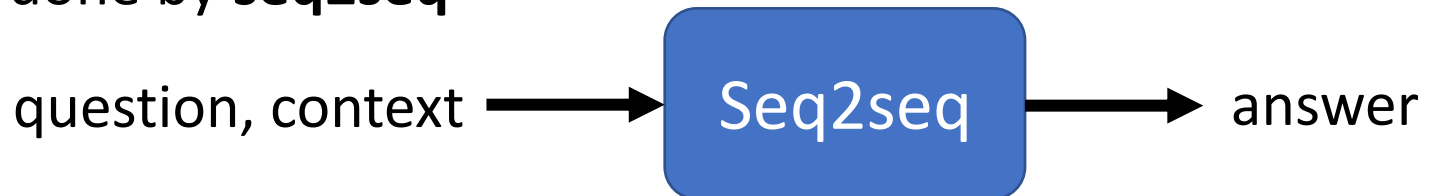
Most Natural Language Processing applications ...

Question Answering (QA)

<u>Question</u>	<u>Context</u>	<u>Answer</u>
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



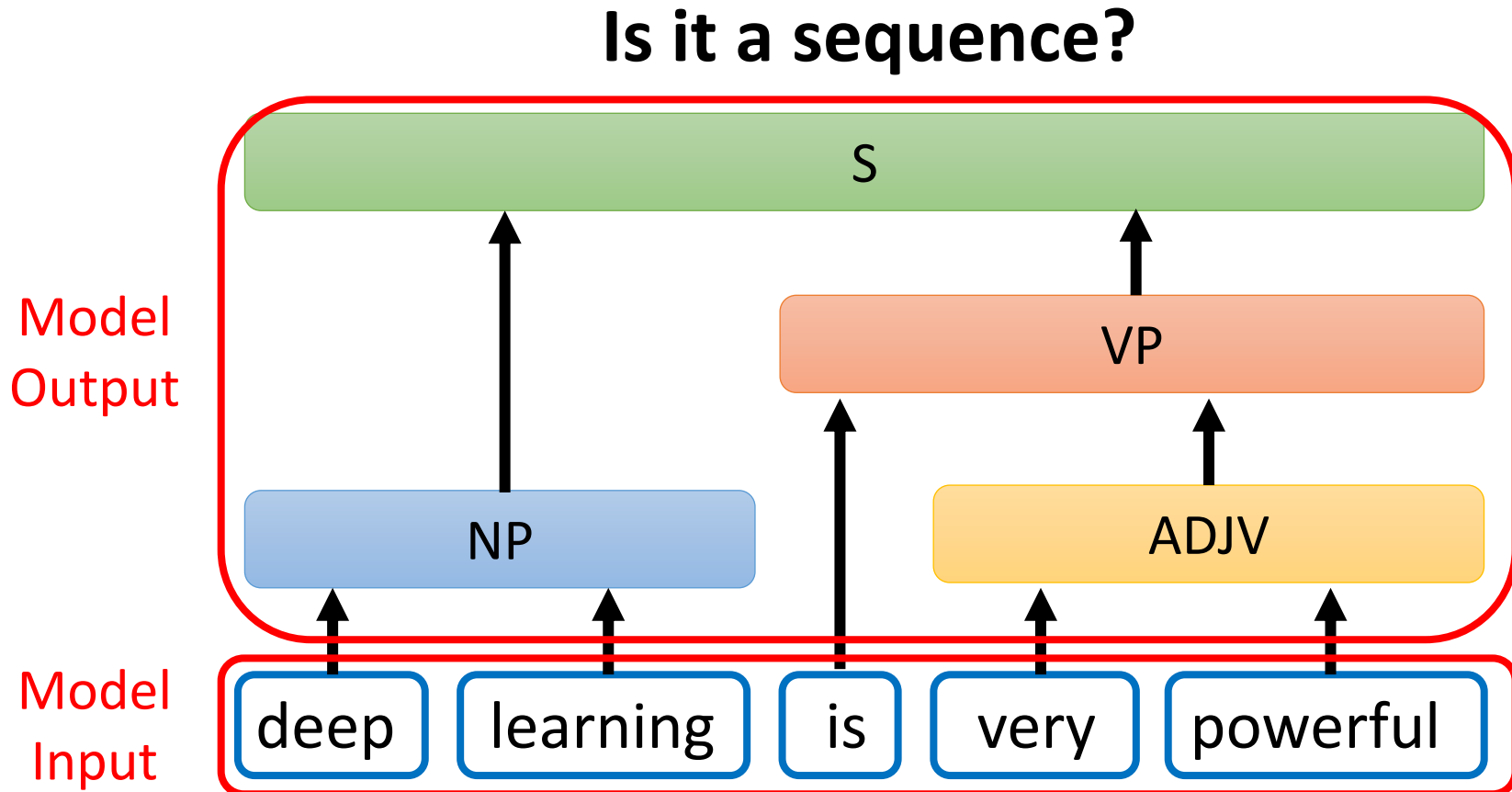
QA can be done by seq2seq



<https://arxiv.org/abs/1806.08730>

<https://arxiv.org/abs/1909.03329>

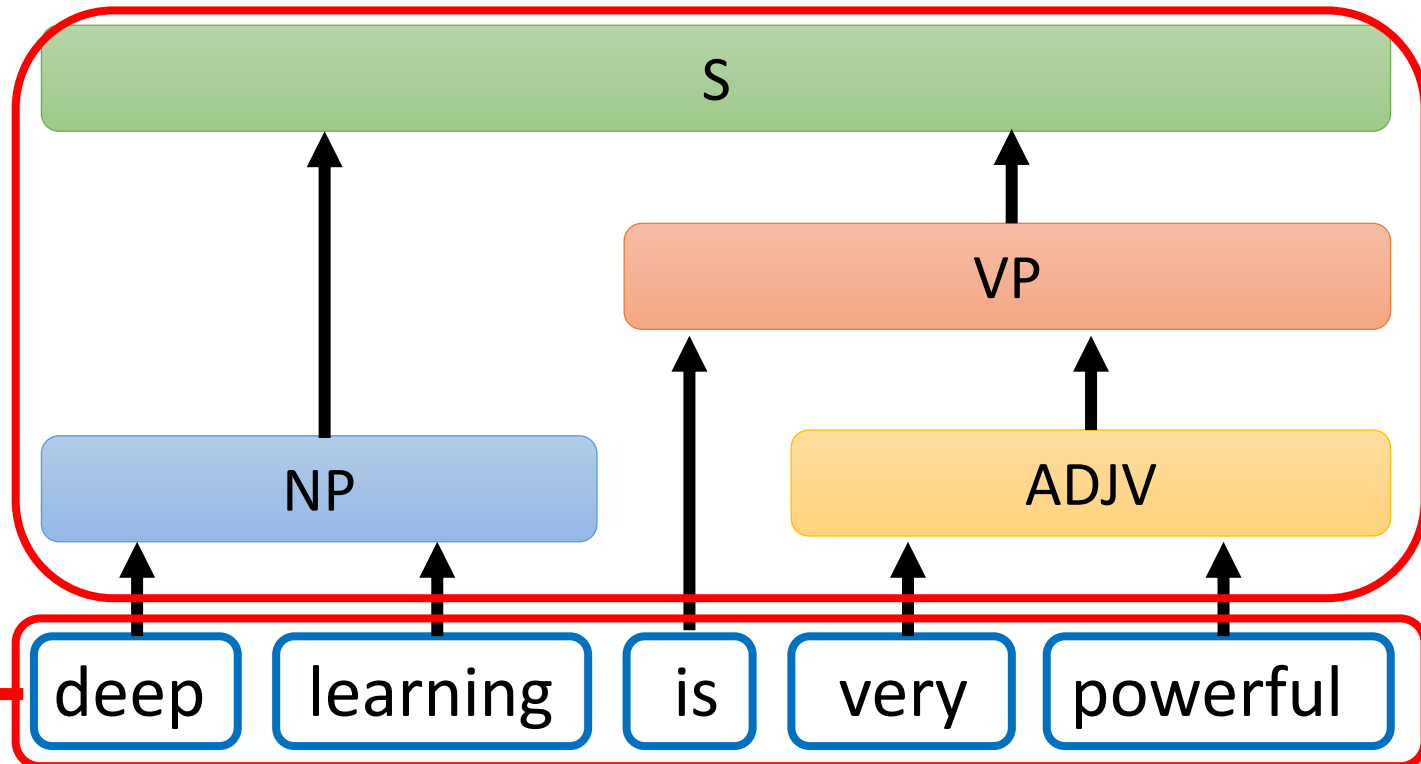
Seq2seq for Syntactic Parsing



Seq2seq for Syntactic Parsing

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

Seq2seq!



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
lukaszkaizer@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

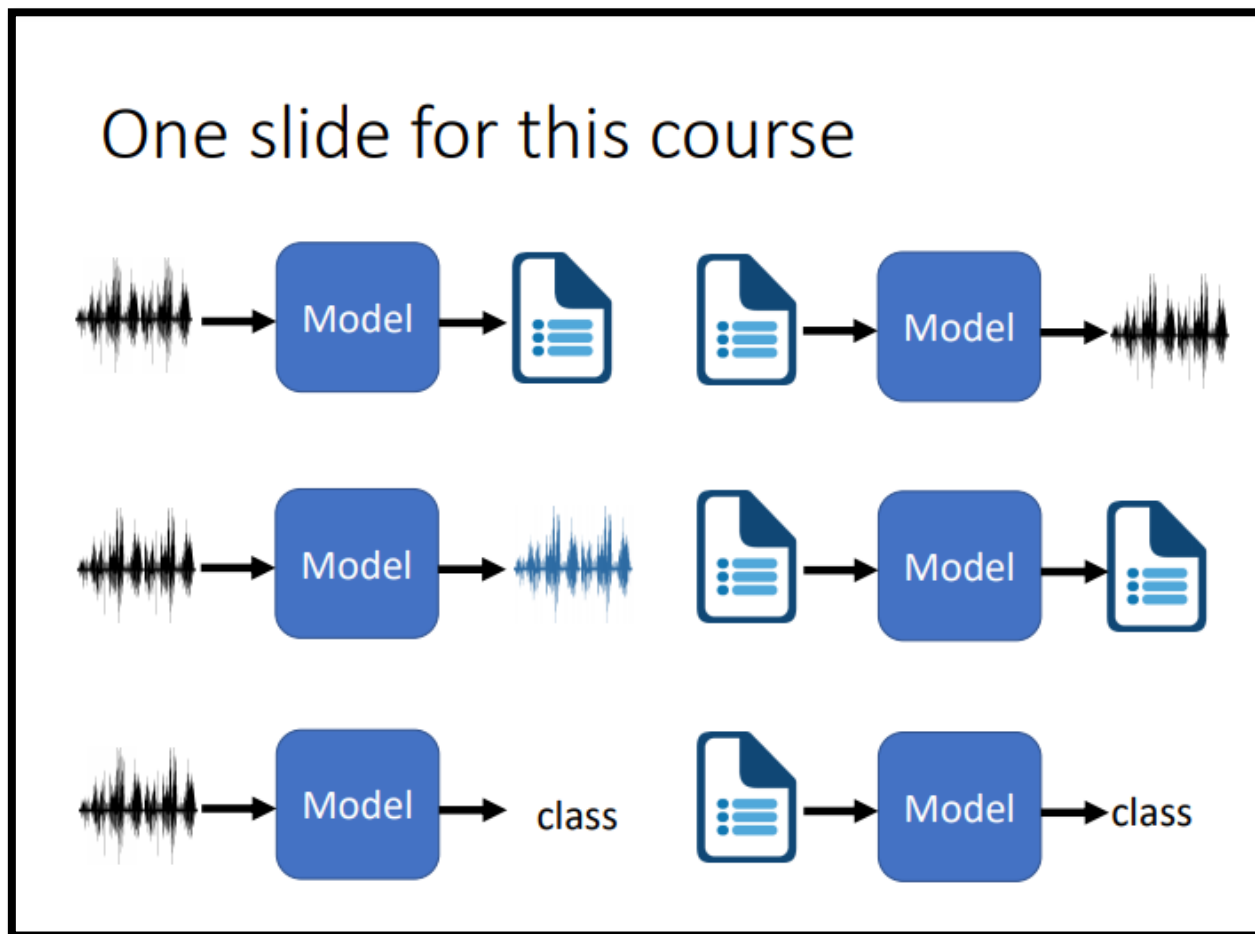
is

very

powerful

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 Multi-label Classification

An object can belong to multiple classes.



Class 1
Class 3



Class 1



Class 3
Class 9
Class 17



Class 10



Class 9



Class 7

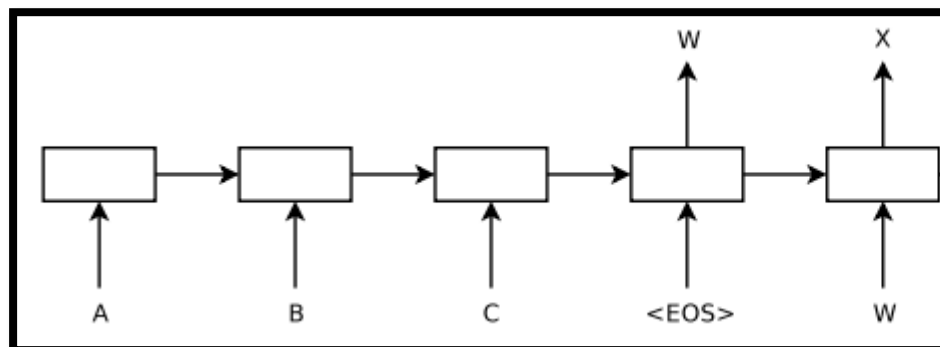
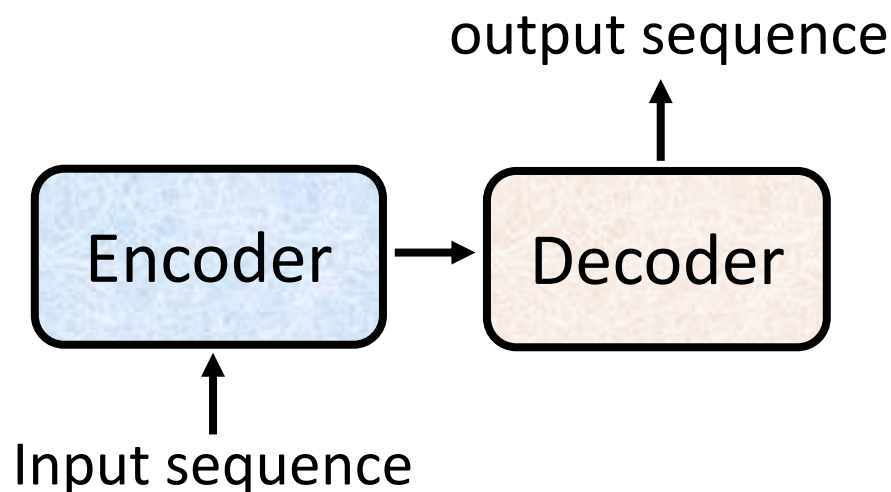


Class 13

<https://arxiv.org/abs/1909.03434>

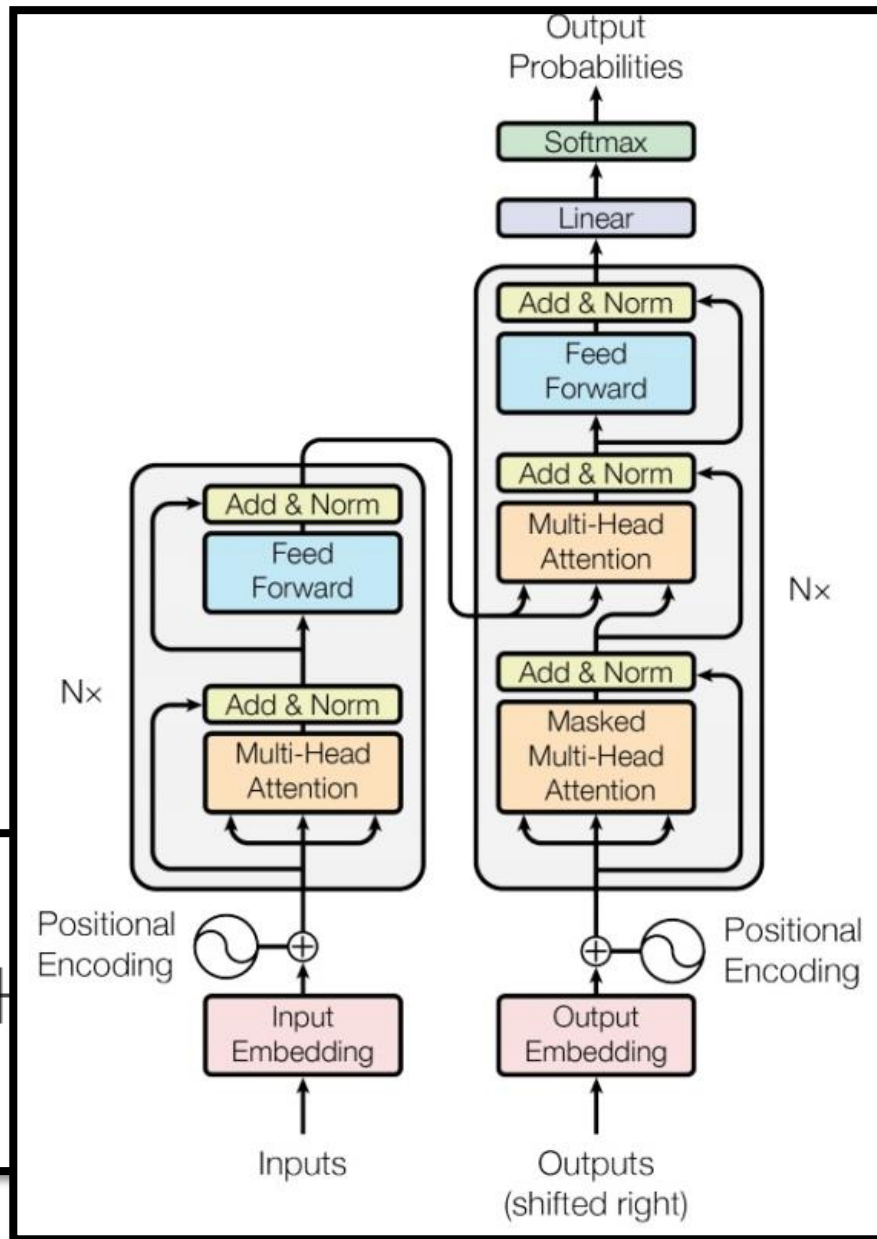
<https://arxiv.org/abs/1707.05495>

Seq2seq



Sequence to Sequence Learning with Neural Networks

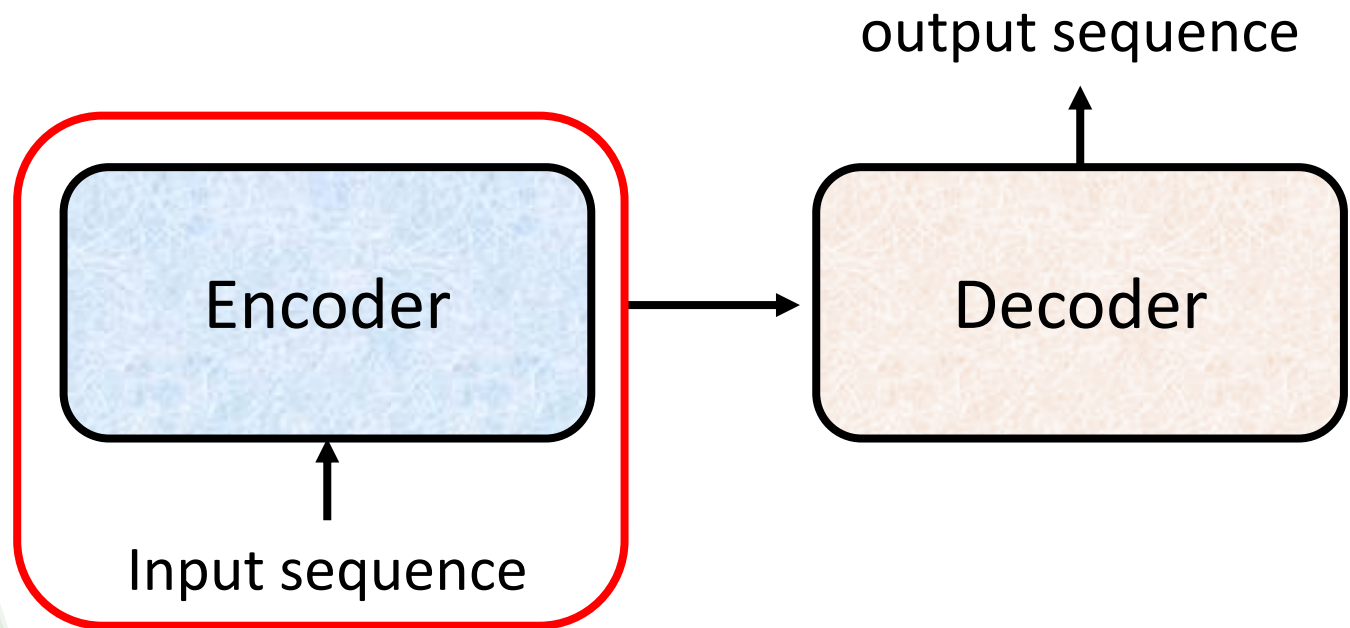
<https://arxiv.org/abs/1409.3215>



Transformer

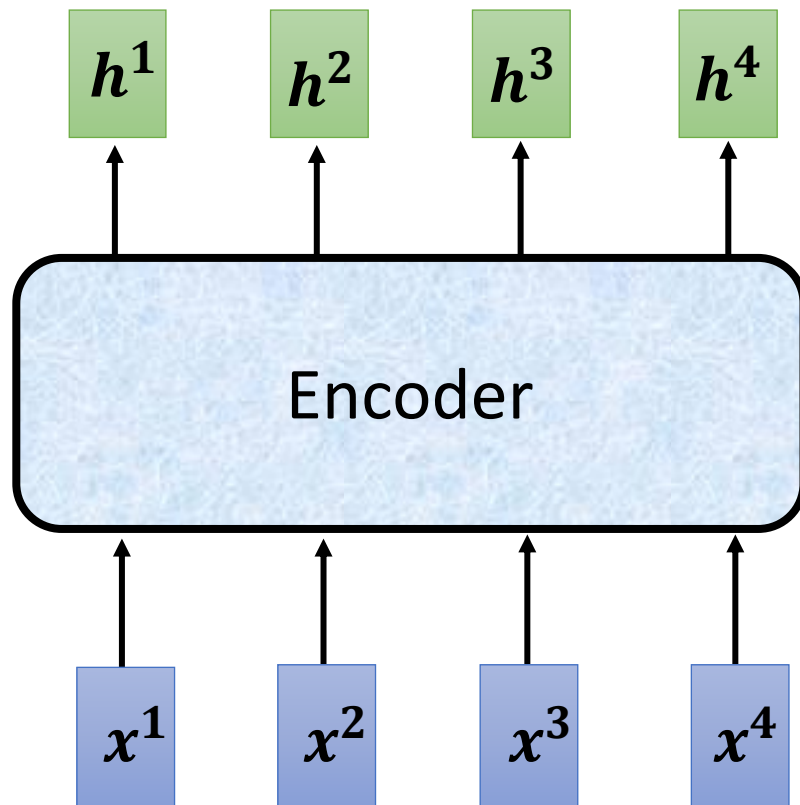
<https://arxiv.org/abs/1706.03762>

Encoder

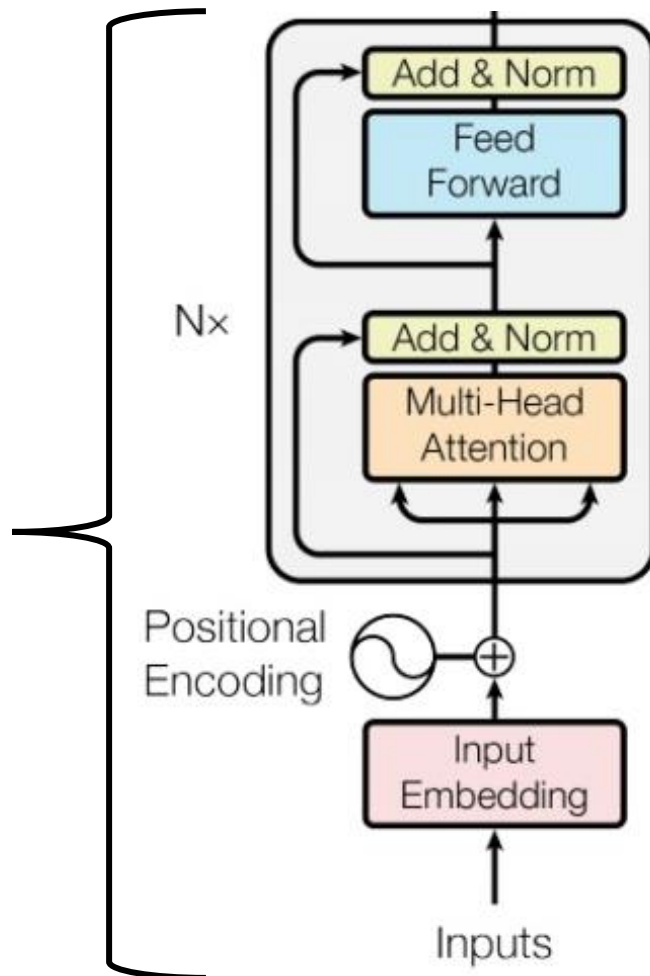


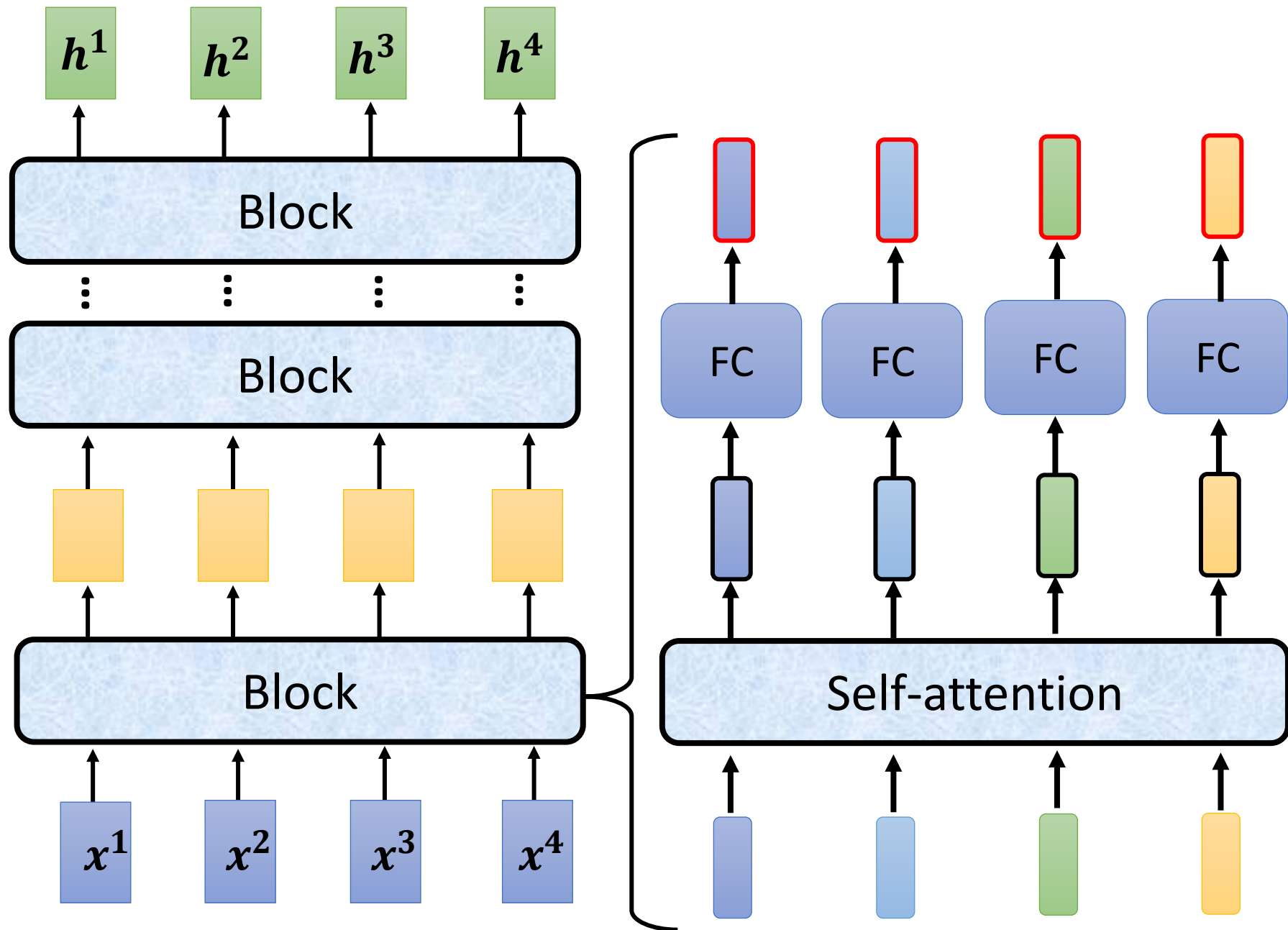
Encoder

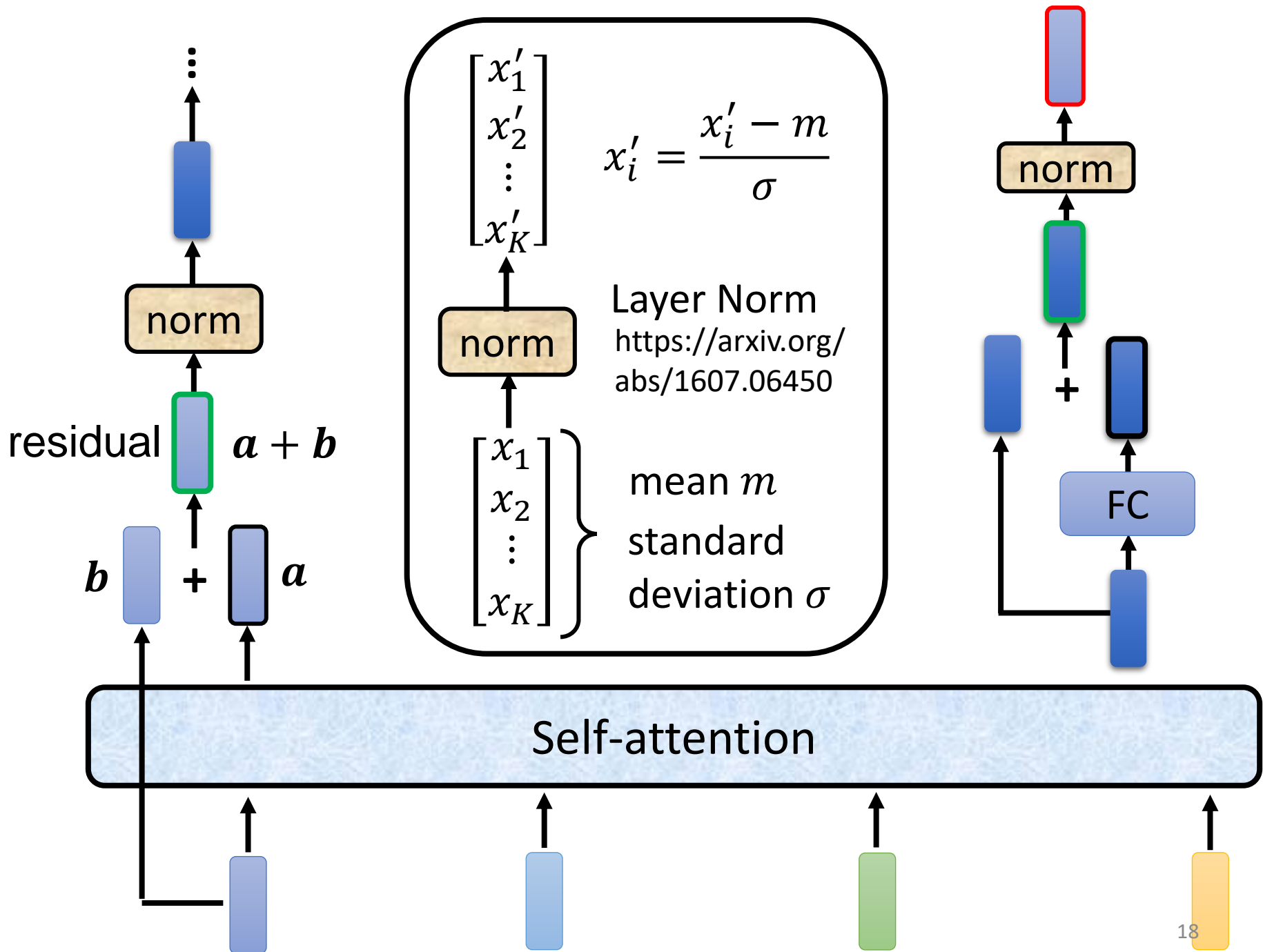
You can use **RNN** or **CNN**.



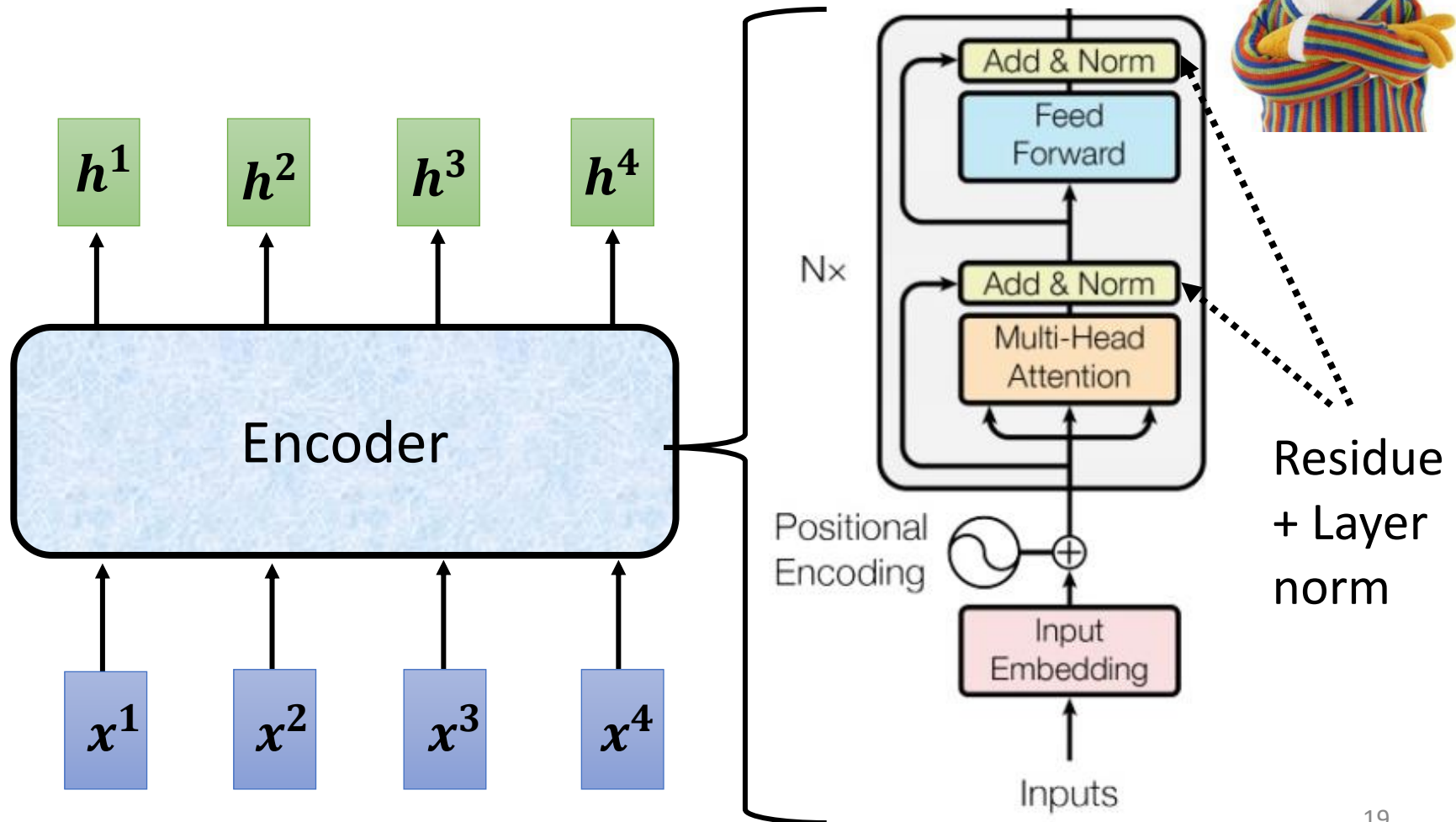
Transformer's Encoder





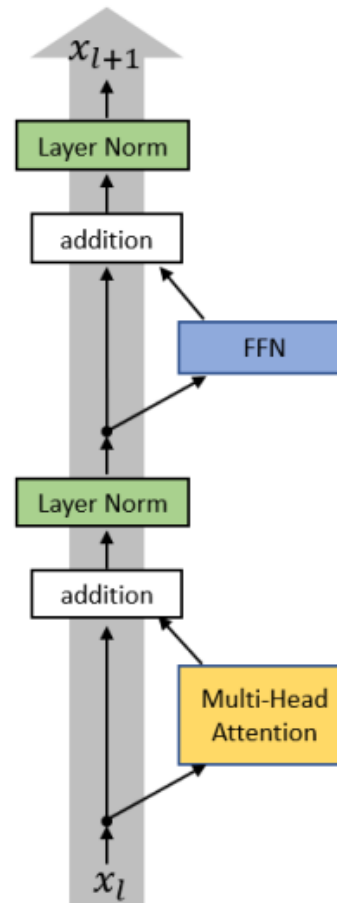


I use the **same** network architecture as **transformer encoder**.

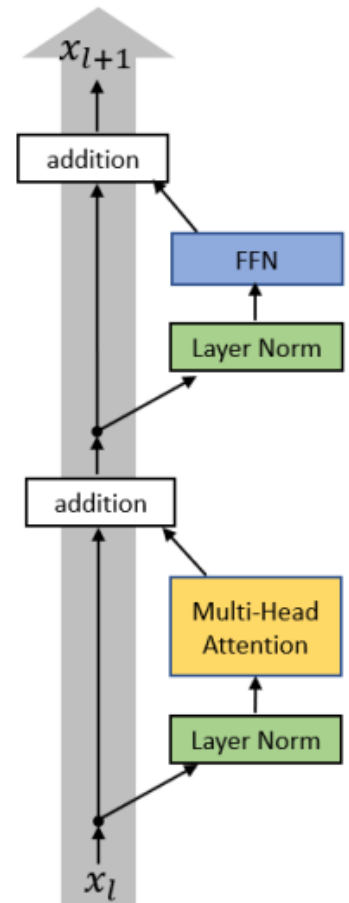


To learn more

- 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>

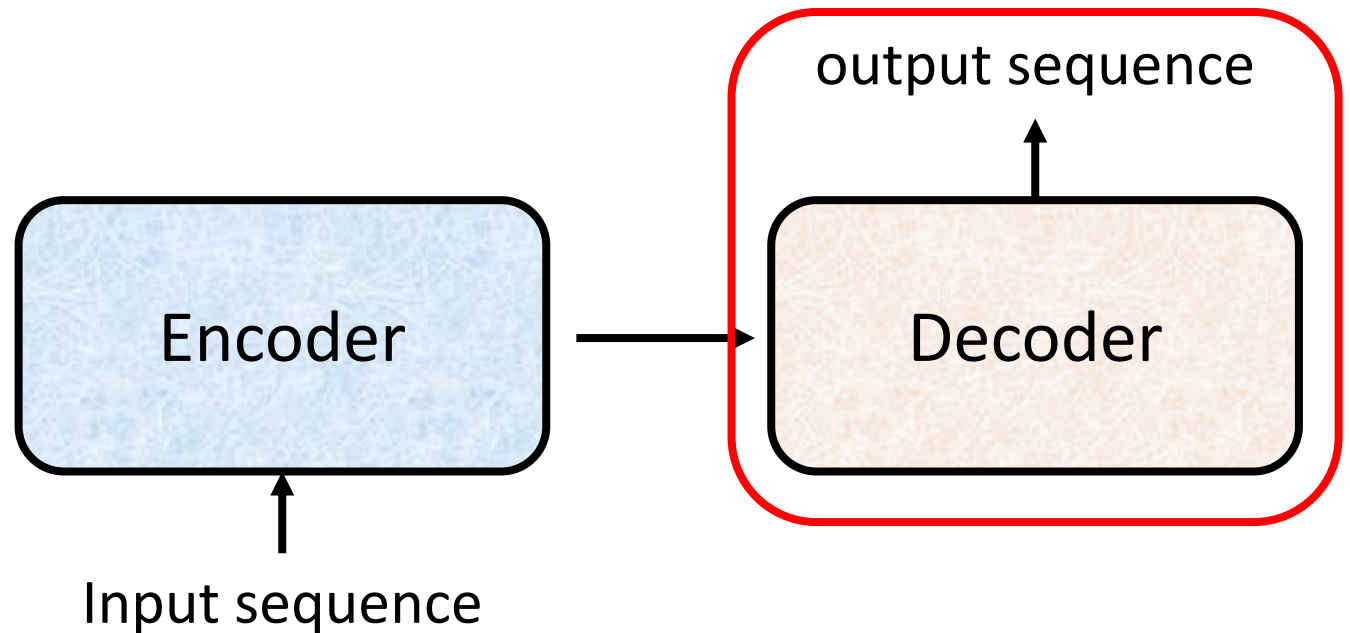


(a)



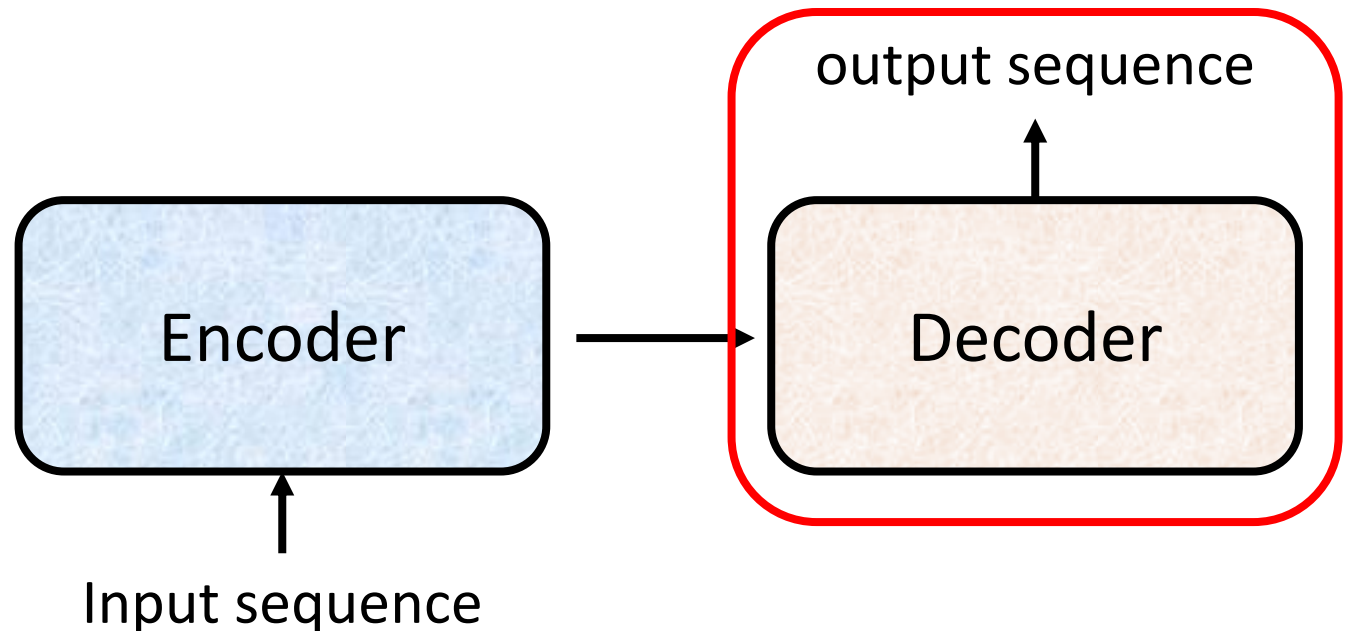
(b)

Decoder



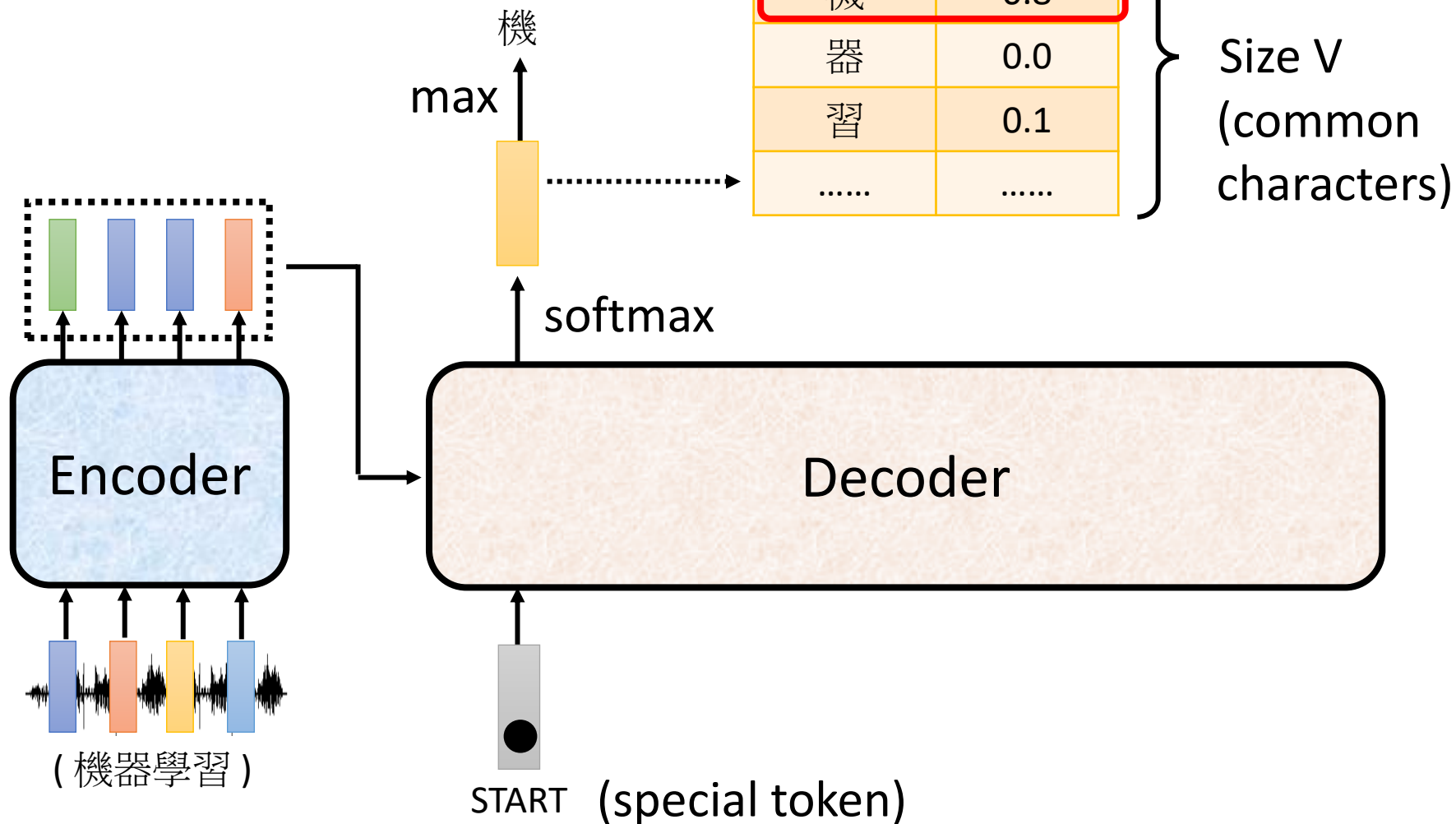
Decoder

- Autoregressive (AT)

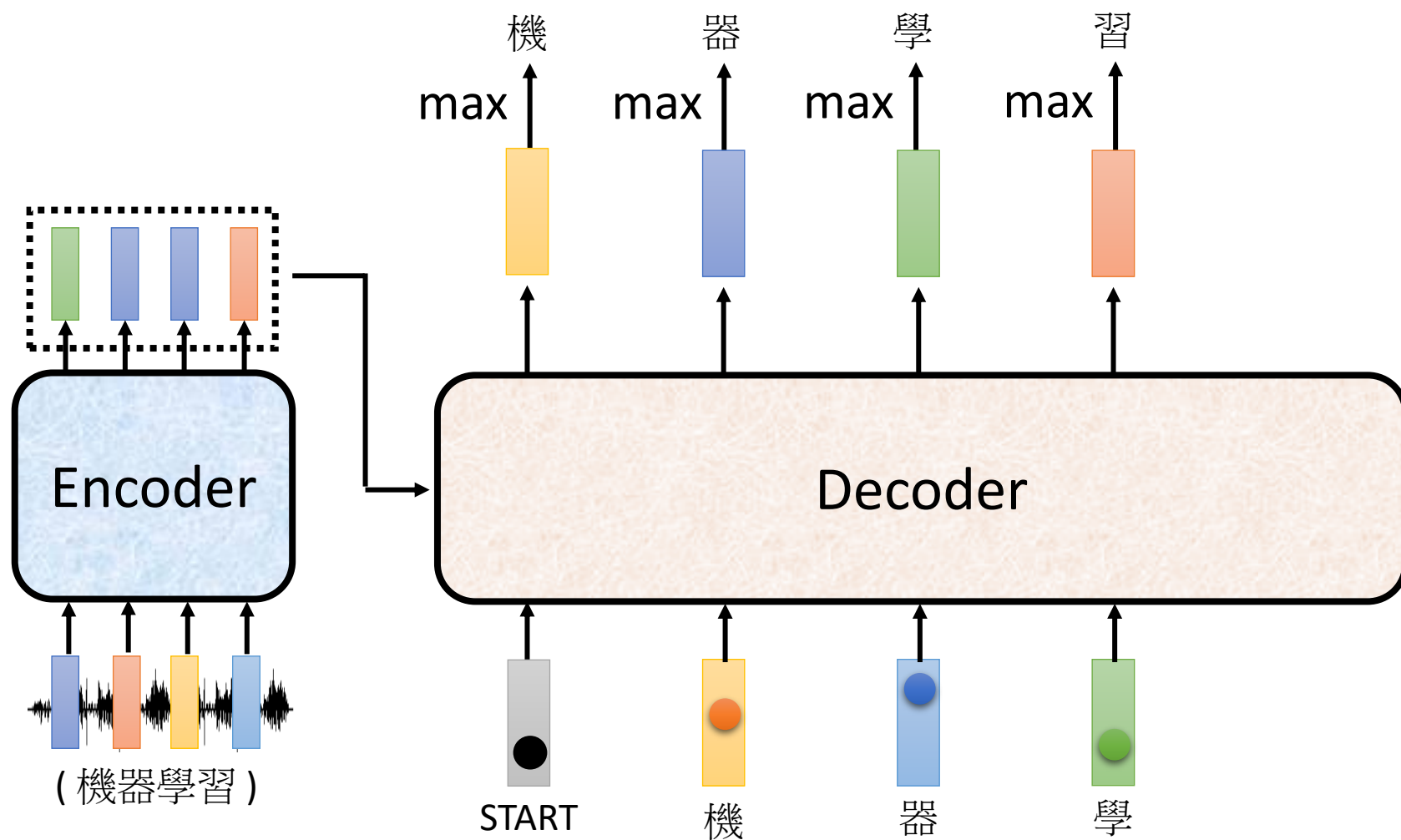


Autoregressive

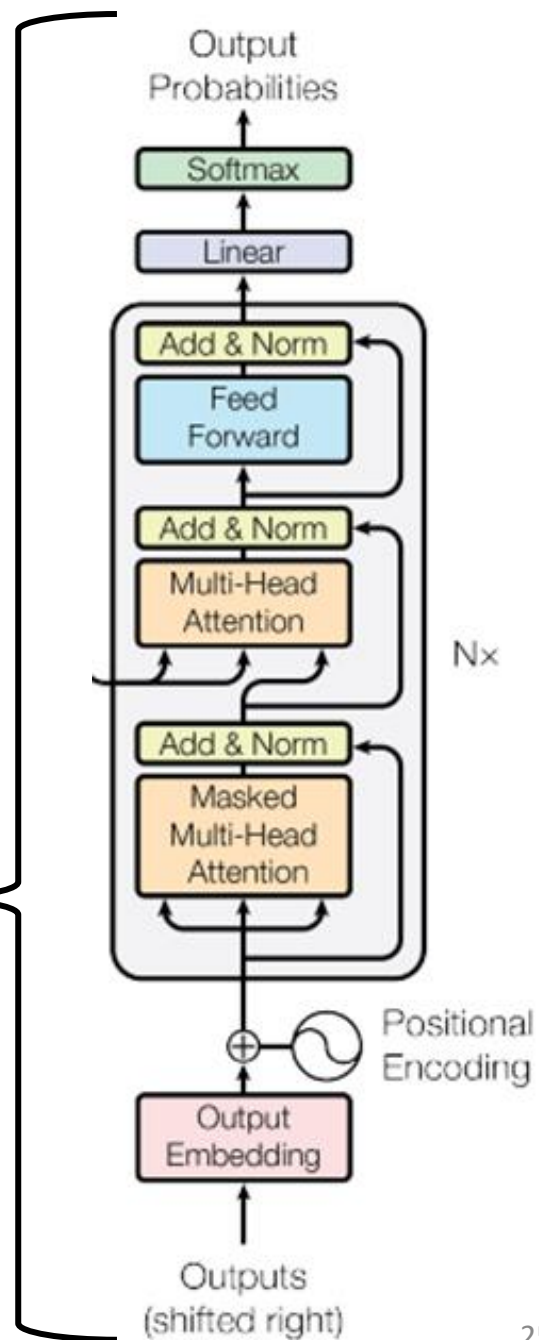
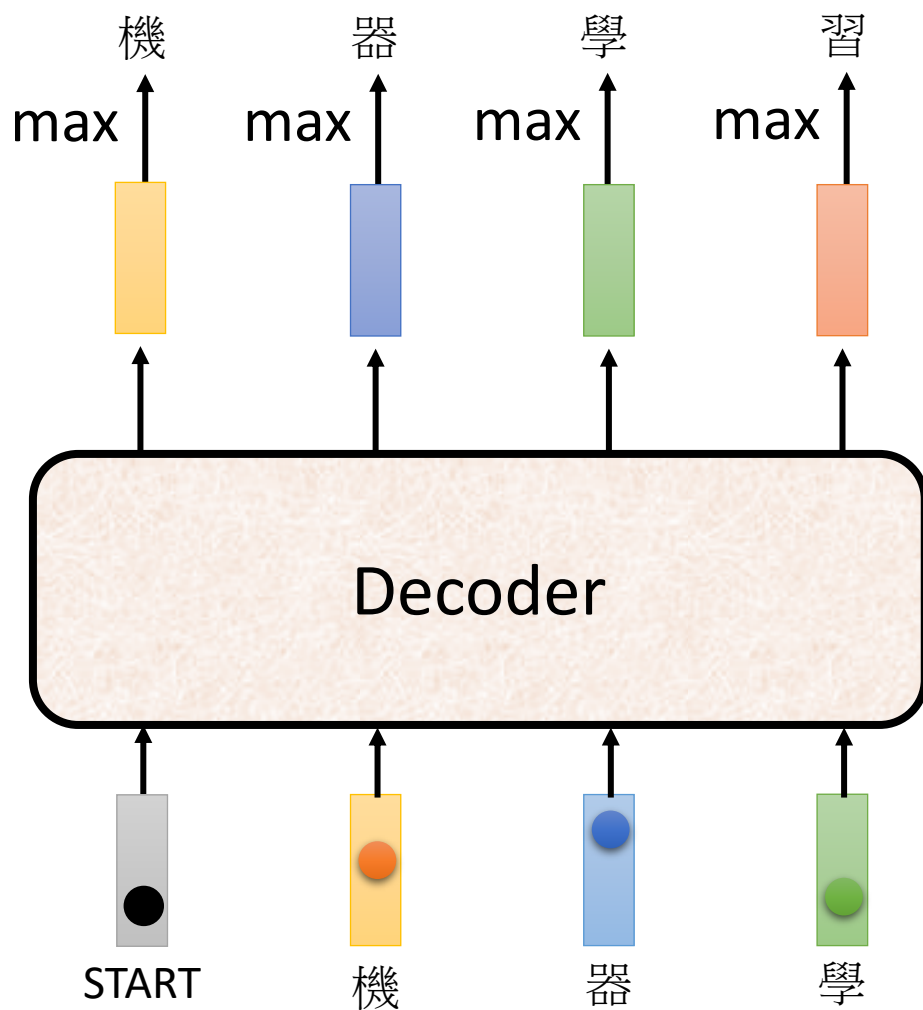
(Speech Recognition as example)



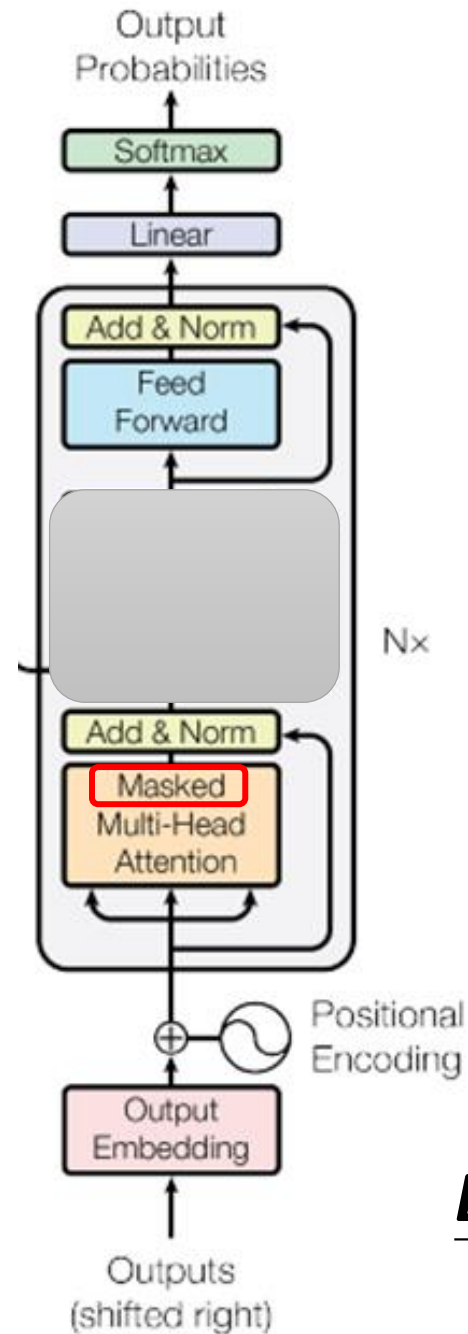
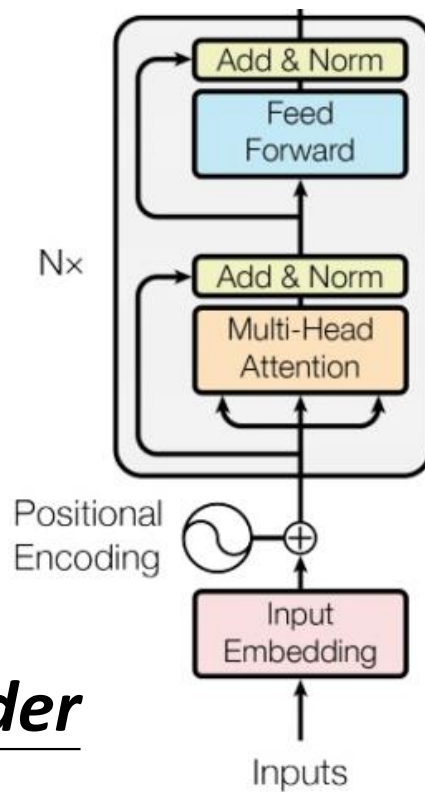
Autoregressive



ignore the input from the encoder here 😊

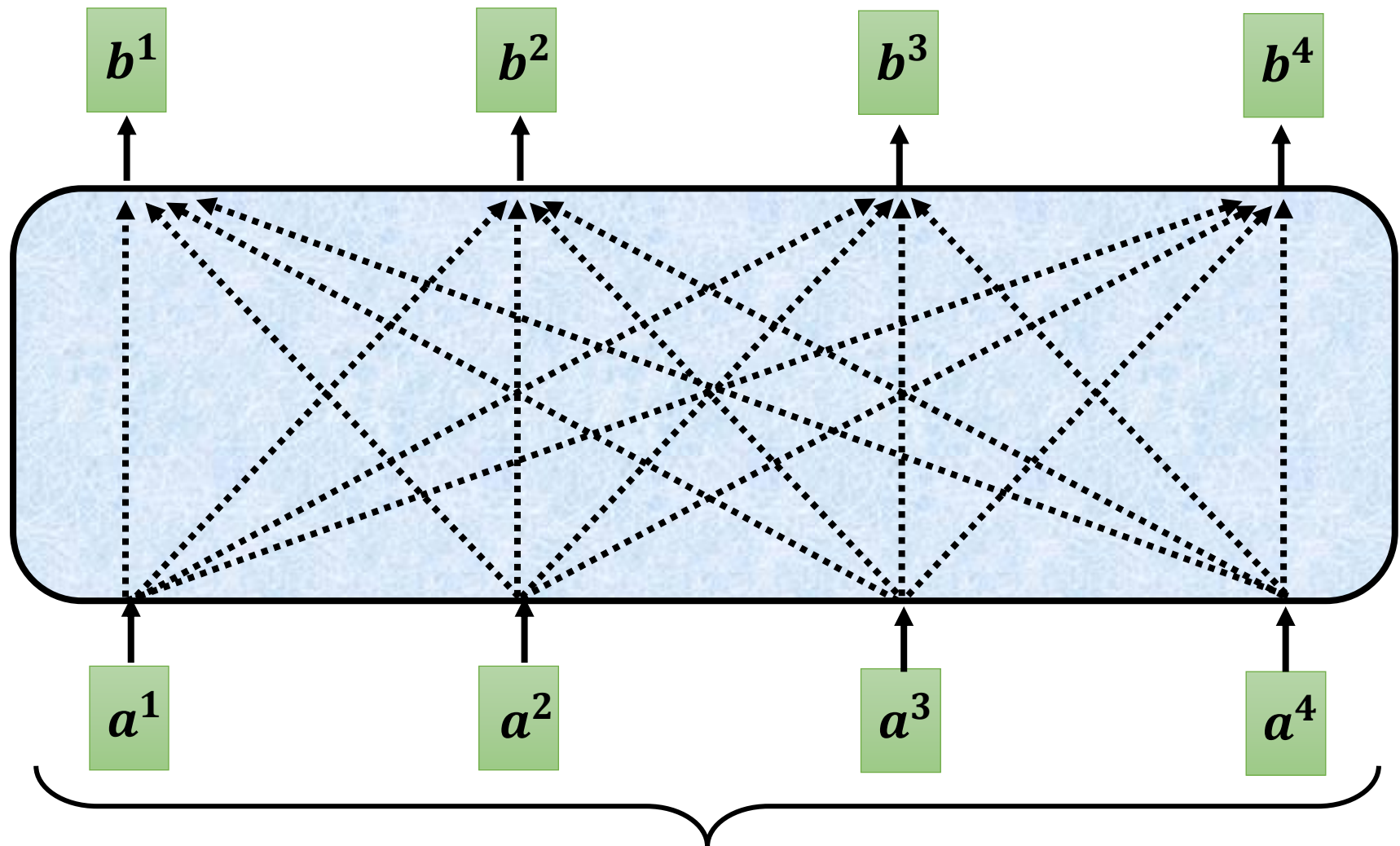


Encoder



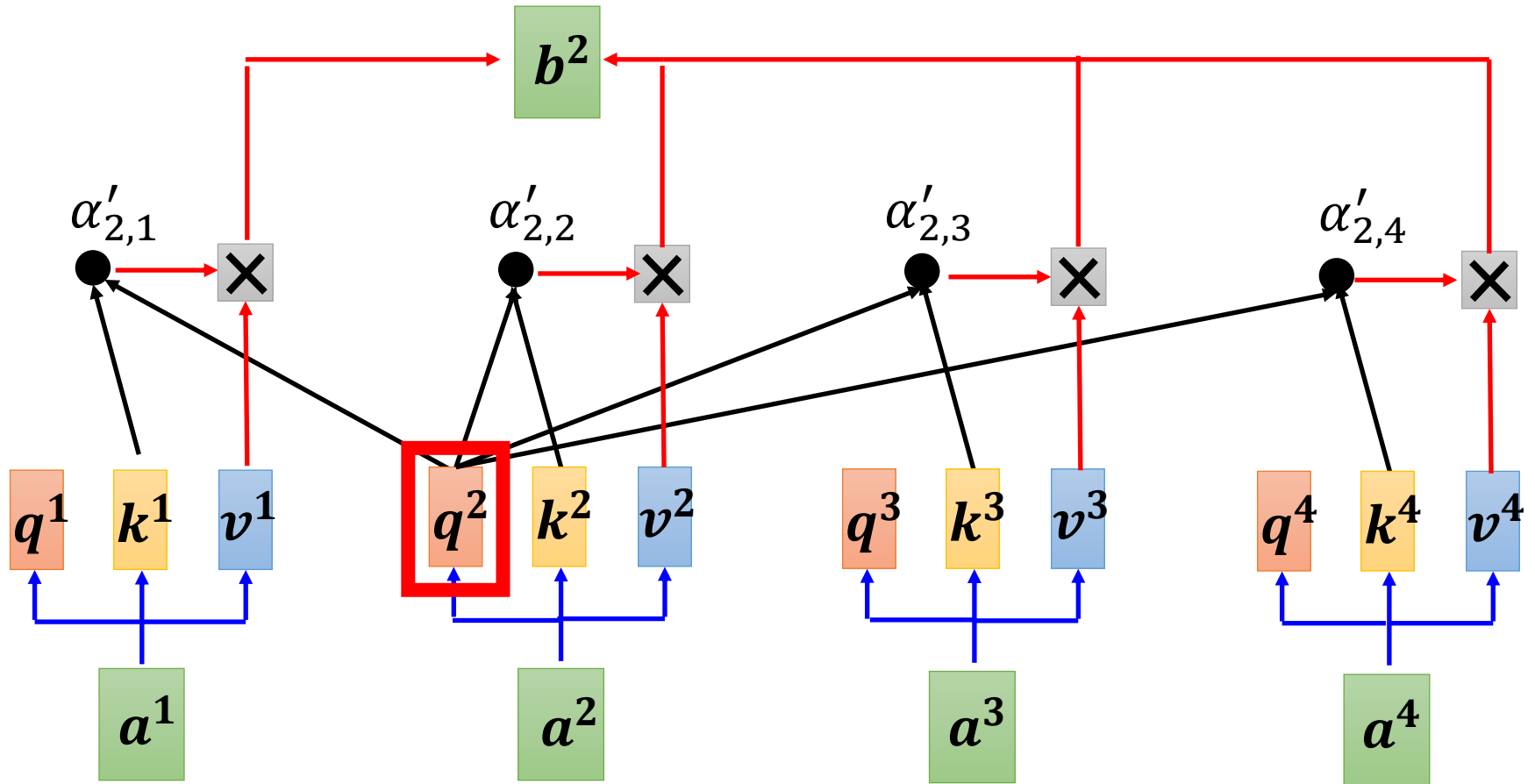
Decoder

Self-attention \longrightarrow Masked Self-attention



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

Self-attention \longrightarrow Masked Self-attention

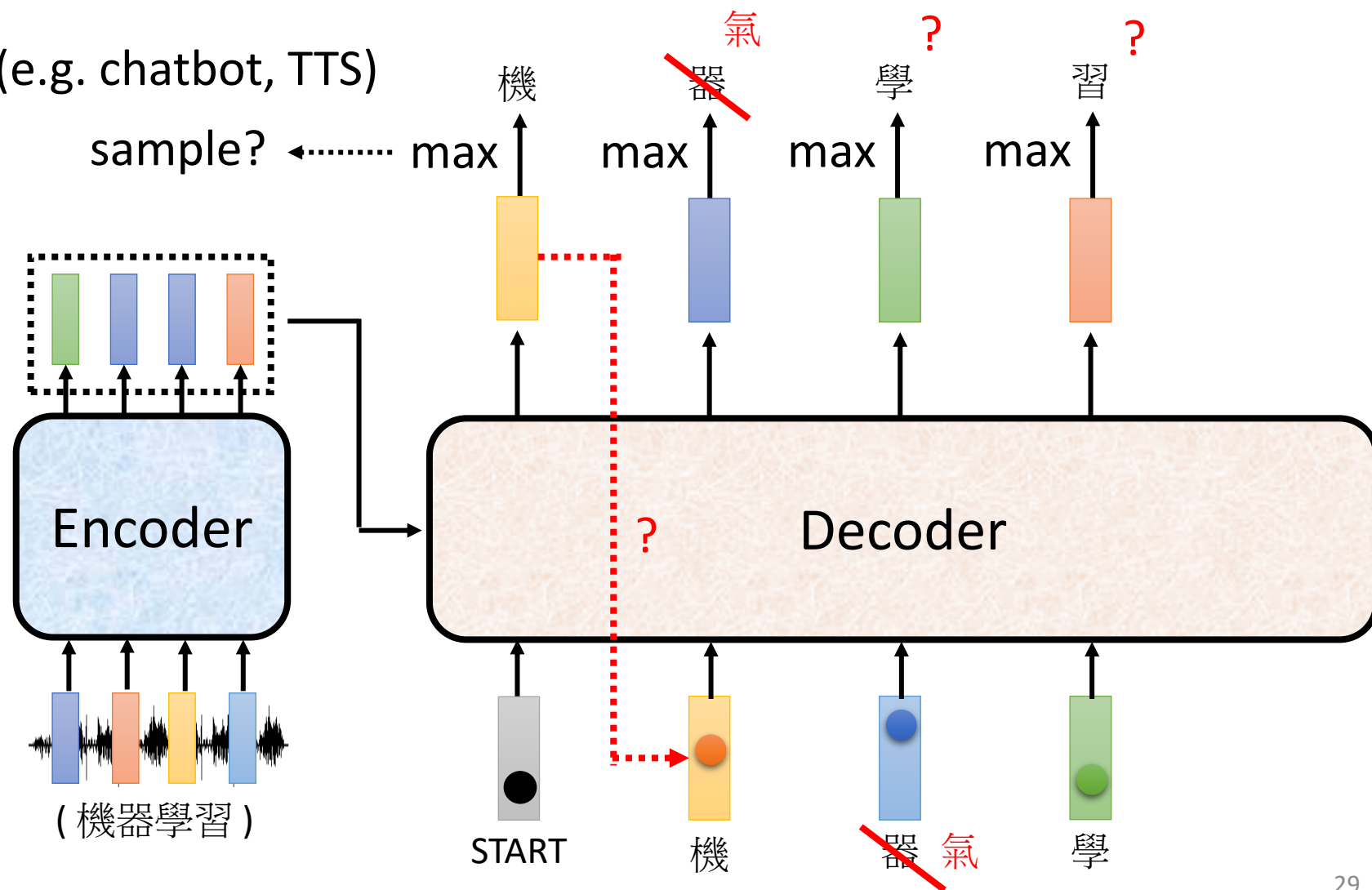


Why masked? Consider how does decoder work

Autoregressive

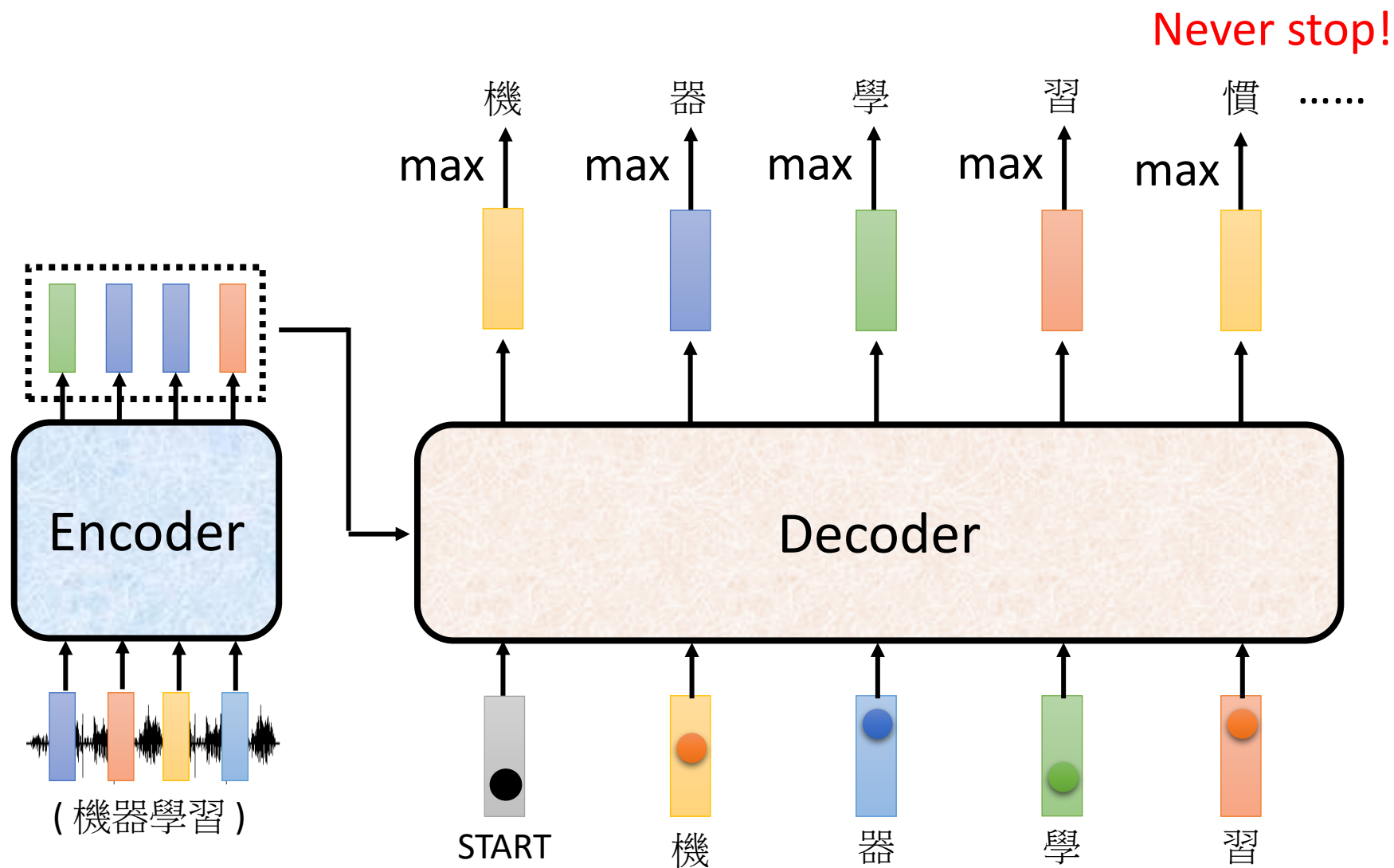
- Question?
- Error propagation
 - Distribution as input?

(e.g. chatbot, TTS)



Autoregressive

We do not know the correct output length.

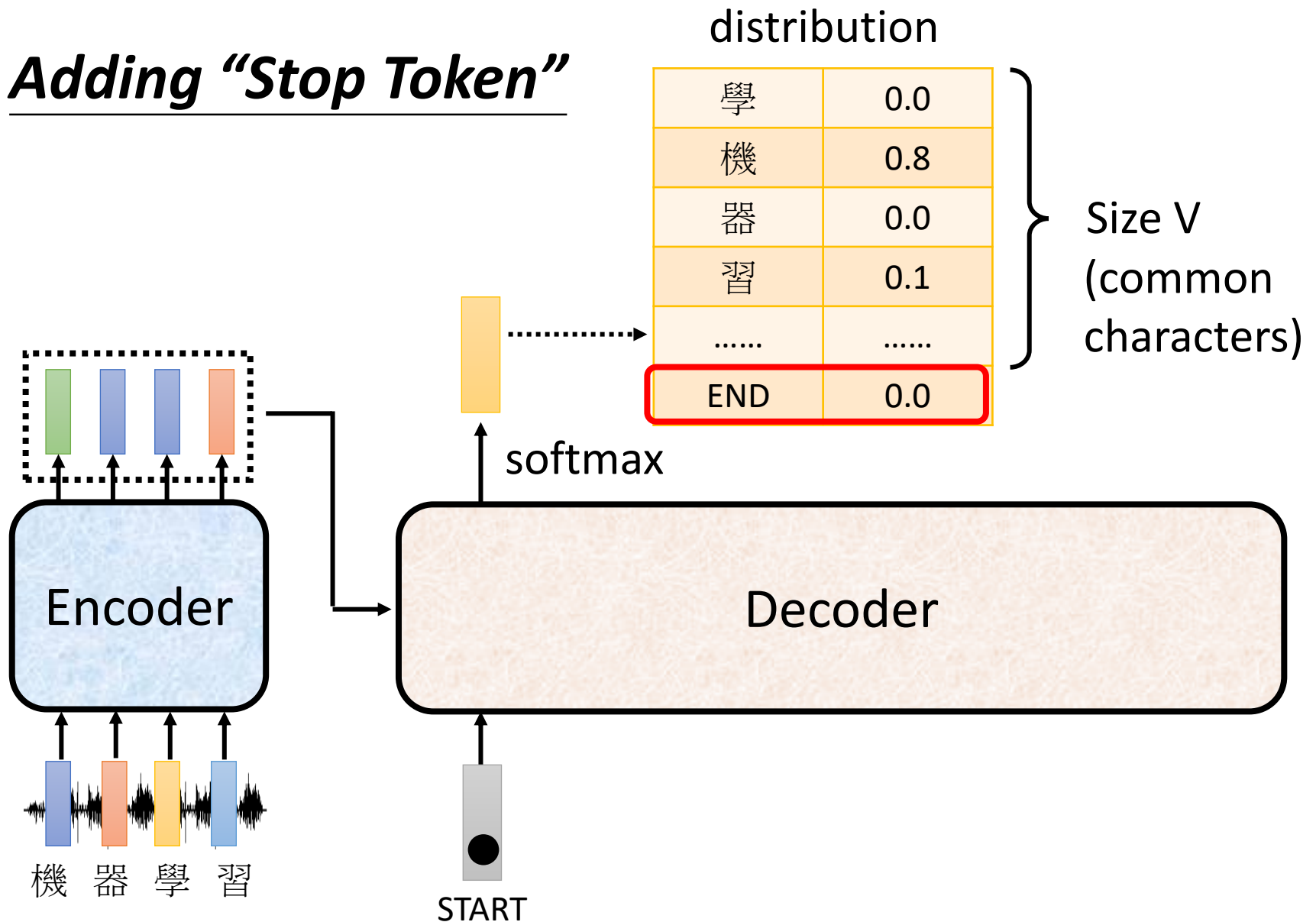


推文接龍 (Tweet Solitaire)

推	:	超	06/12 10:39
推	n:	人	06/12 10:40
推	tion:	正	06/12 10:41
→	host:	大	06/12 10:47
推	:	中	06/12 10:59
推	403:	天	06/12 11:11
推	:	外	06/12 11:13
推	527:	飛	06/12 11:17
→	990b:	仙	06/12 11:32
→	512:	草	06/12 12:15

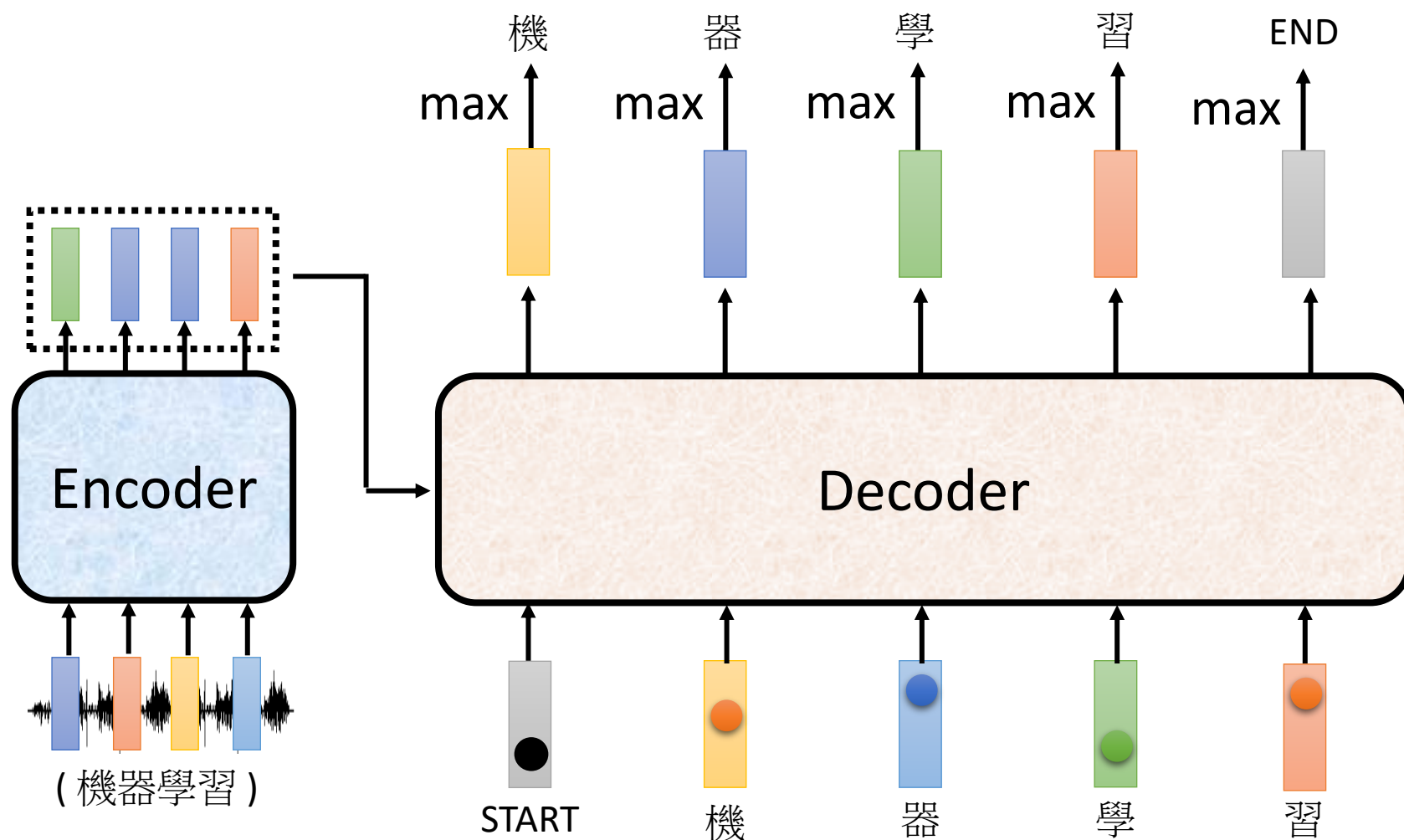
推 tkagk: =====斷=====

Adding “Stop Token”



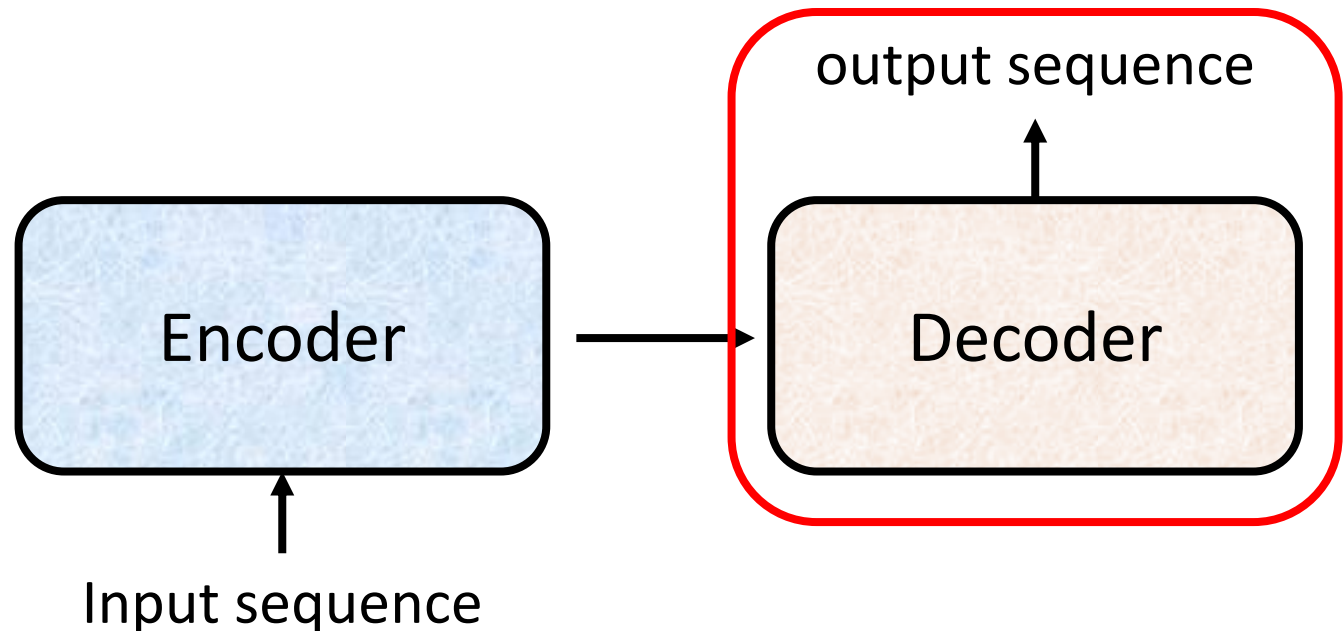
Autoregressive

Stop at here!

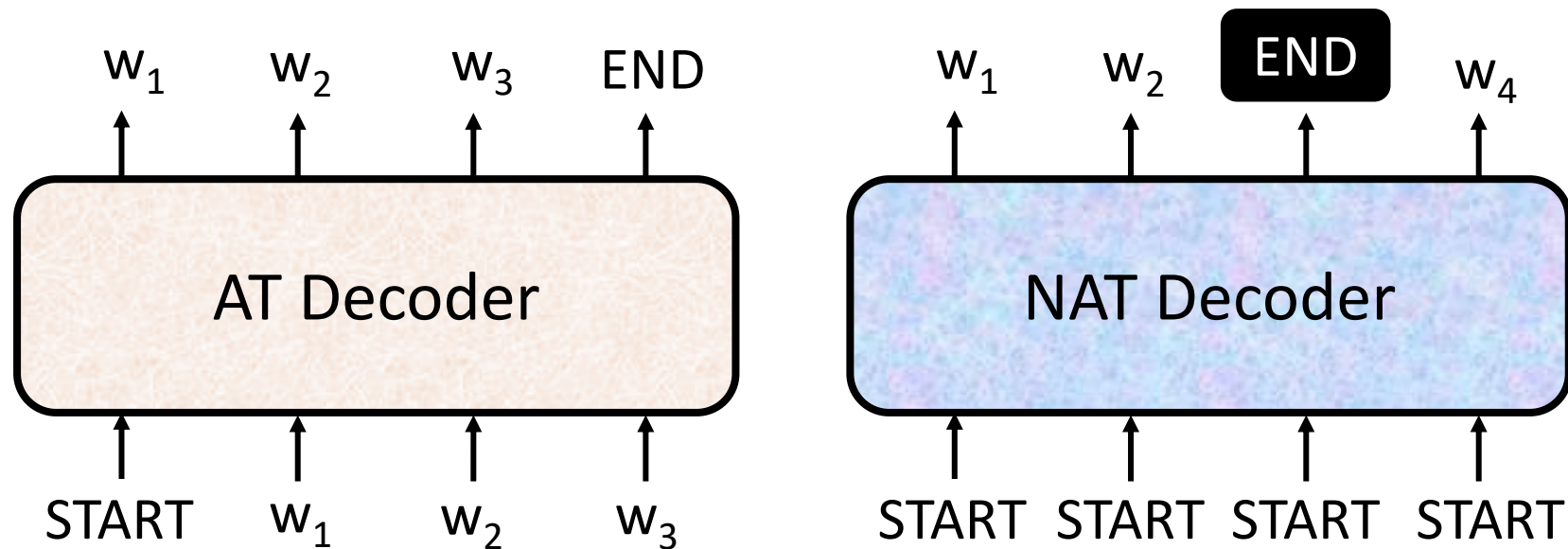


Decoder

- 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)

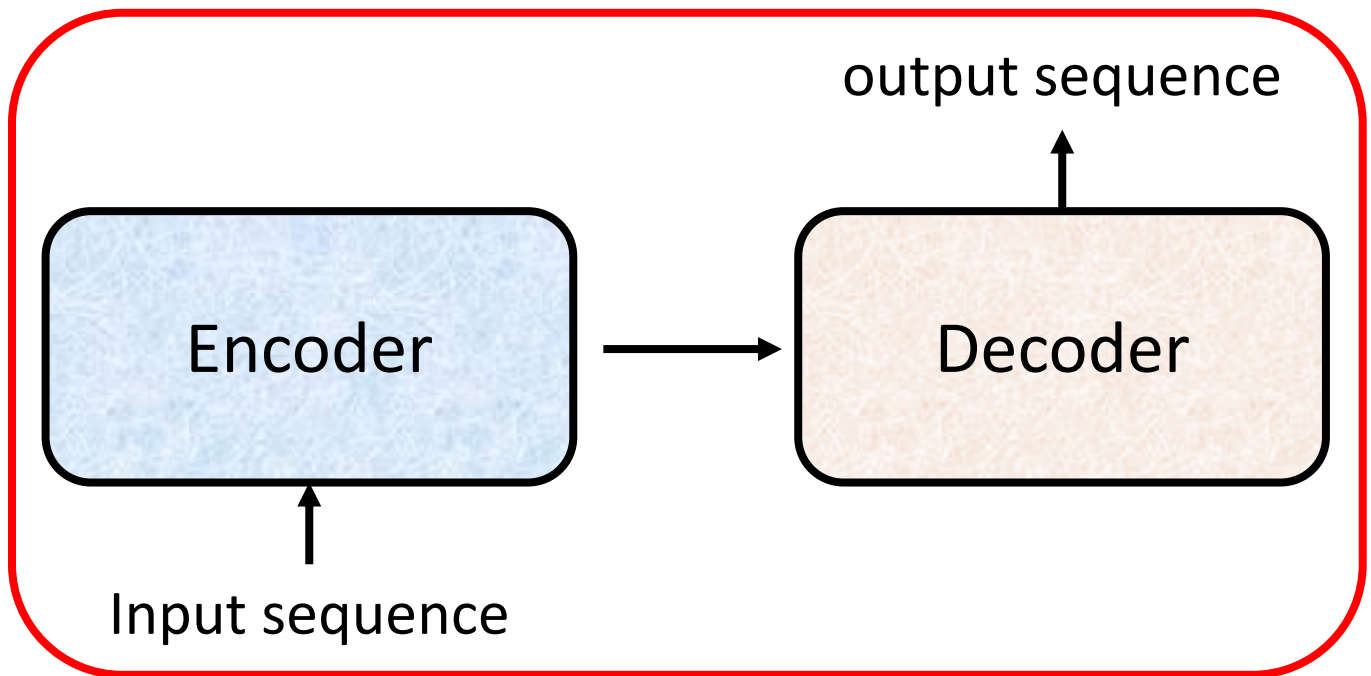
To learn more

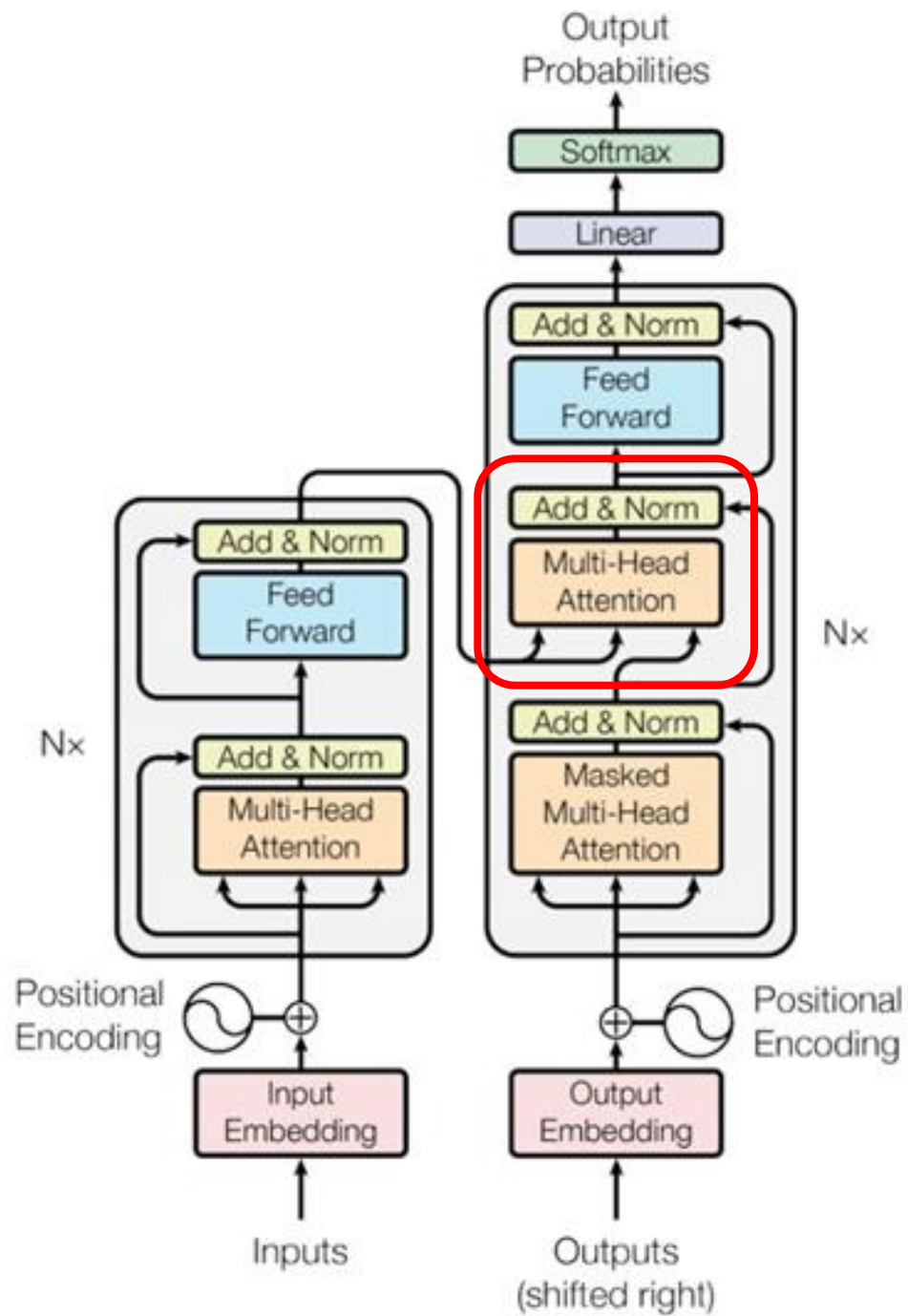


<https://youtu.be/jvyKmU4OM3c>
(in Mandarin)



Encoder-Decoder





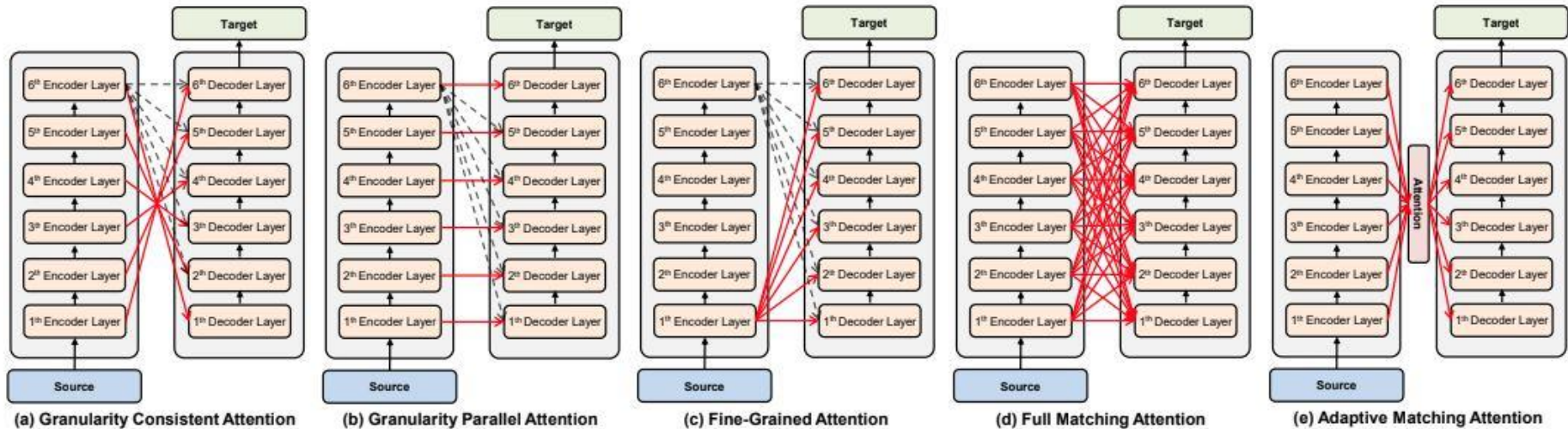
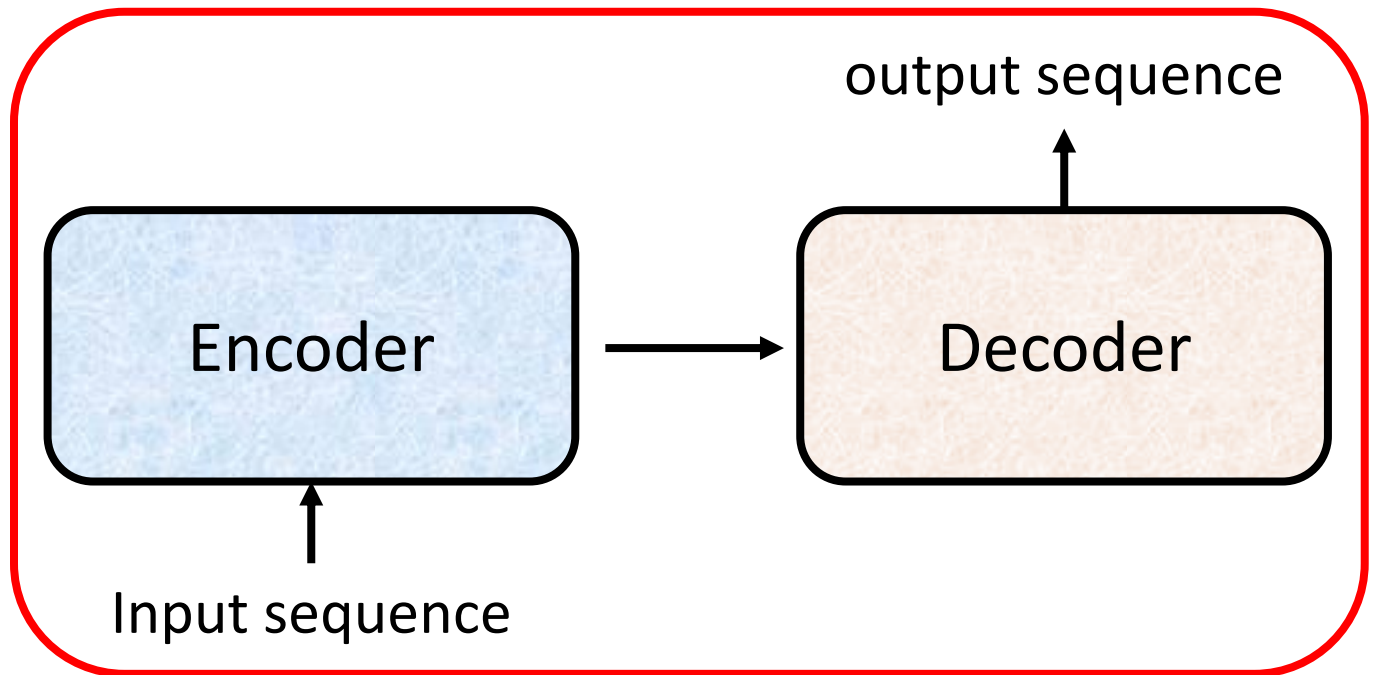


Figure 2: We present the proposal on Transformer with various strategies for routing the source representations: (a) Granularity Consistent Attention; (b) Granularity Parallel Attention; (c) Fine-Grained Attention; (d) Full Matching Attention; (e) Adaptive Matching Attention. The dashed lines represent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

Training



Training

