



Transformer

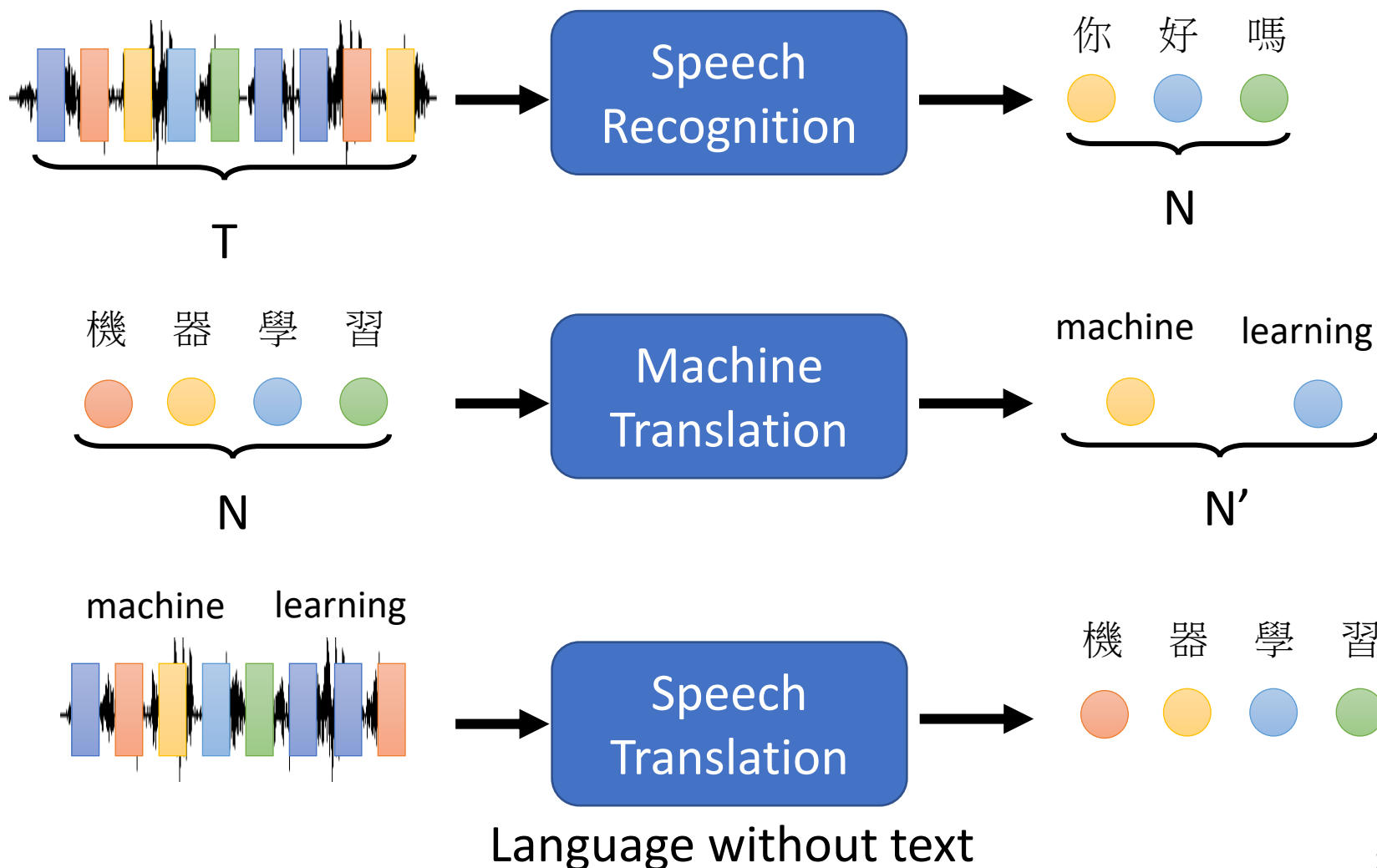
李宏毅

Hung-yi Lee

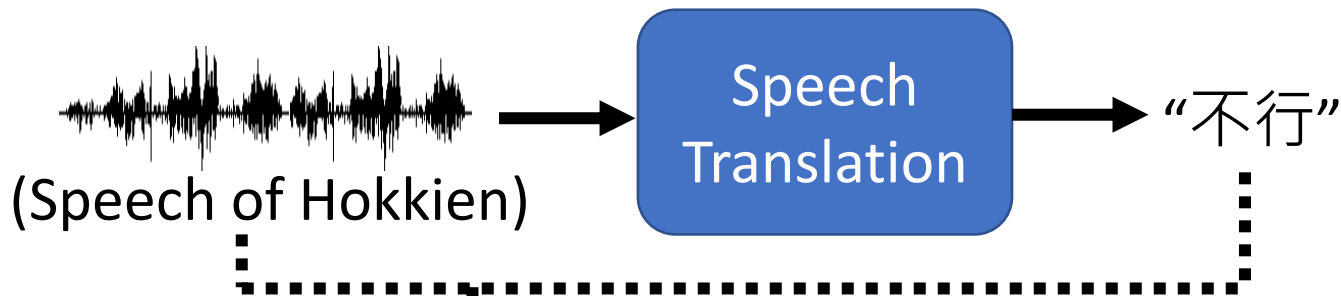
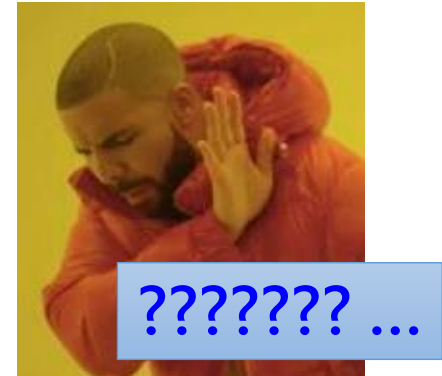
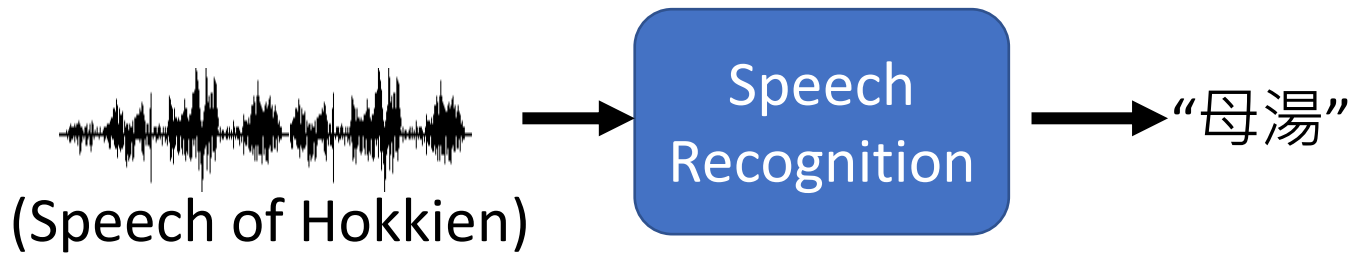
# Sequence-to-sequence (Seq2seq)

Input a sequence, output a sequence

The output length is determined by model.



# Hokkien (閩南語、台語)



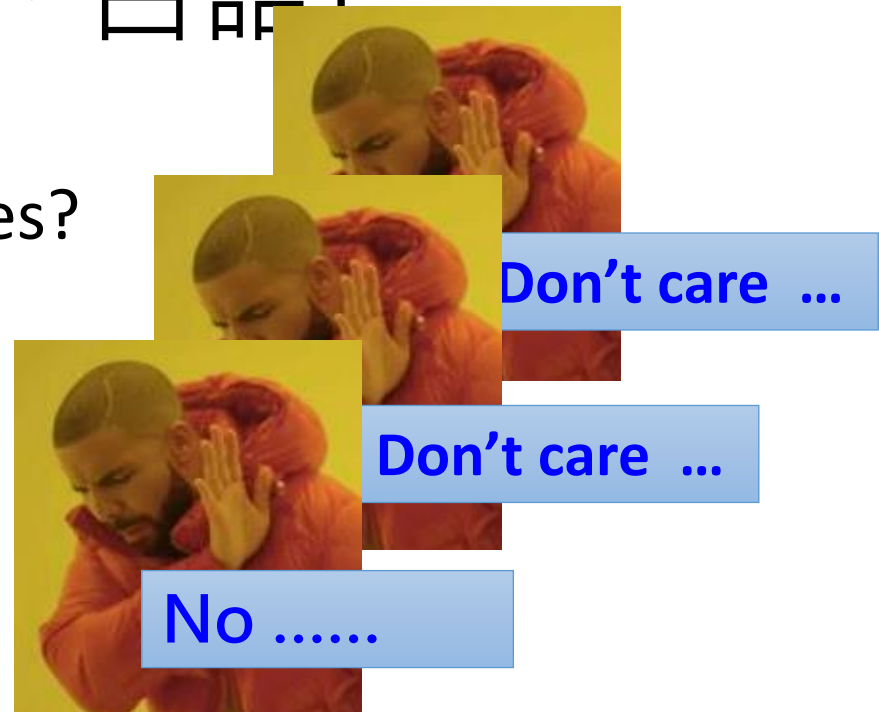
Local soap operas (鄉土劇) on YouTube  
(Speech of Hokkien, Chinese subtitle)

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

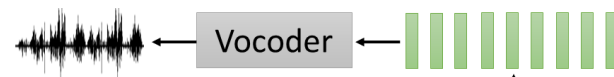


我有幫廠長拜託

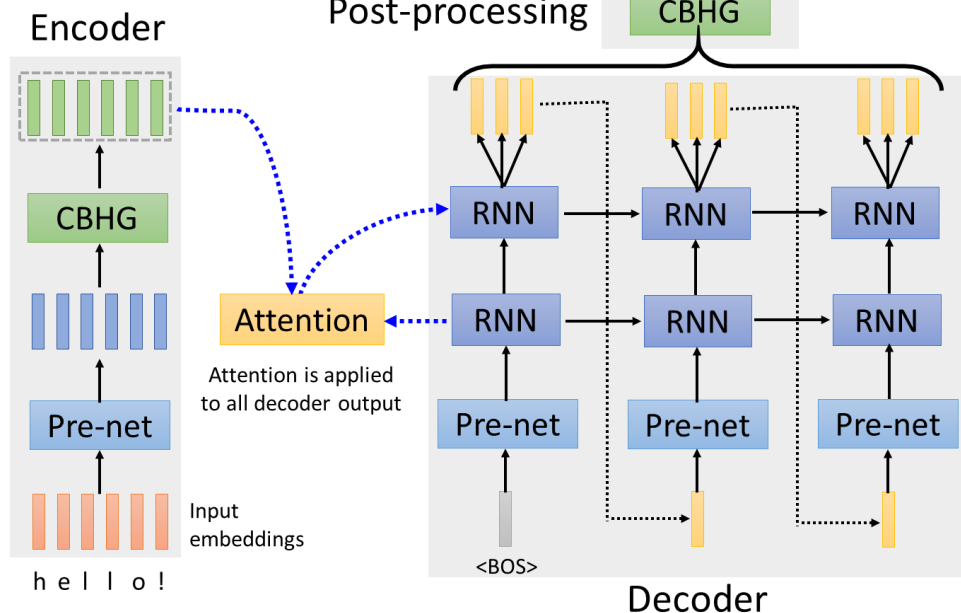
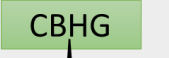
Answer:我拜託廠長了

# Text-to-Speech (TTS) Synthesis

感謝張凱為同學提供實驗結果



Post-processing



## 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

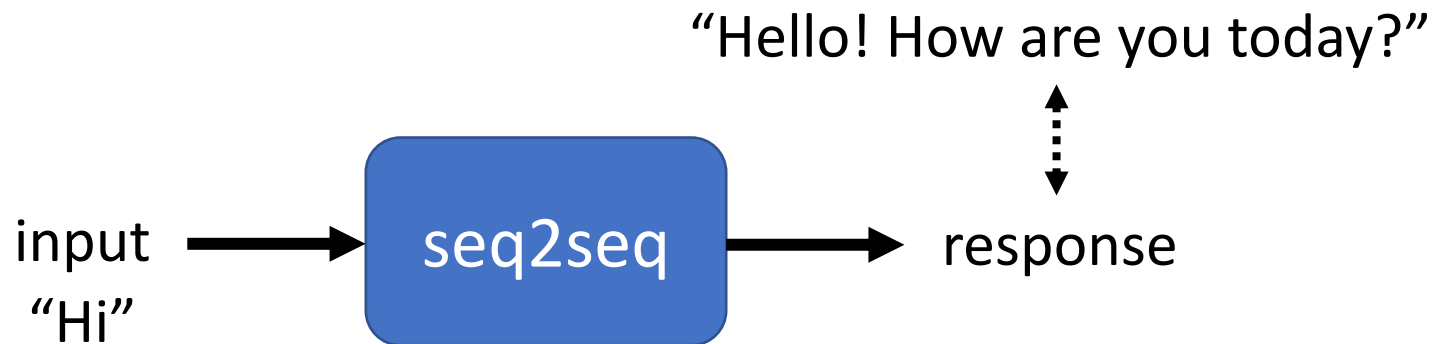


最近肺炎真嚴重，要記得戴口罩、  
勤洗手，有病就要看醫生





# 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.

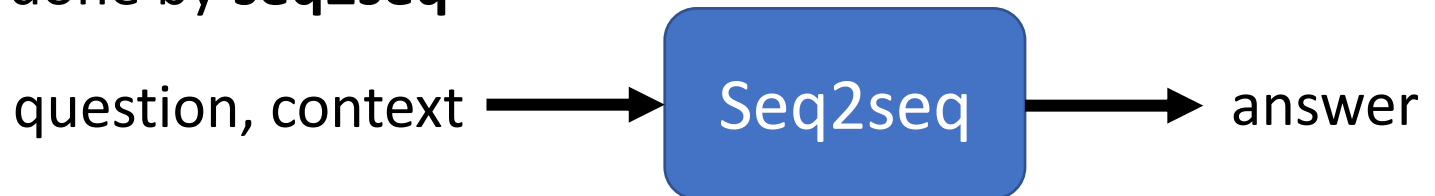
# Most Natural Language Processing applications ...

## 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 <b>major economic center</b> for the state of California and the US....	<b>major economic center</b>
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?	<b>Harry Potter star Daniel Radcliffe</b> gains access to a reported <b>£320 million fortune</b> ...	<b>Harry Potter star Daniel Radcliffe</b> gets <b>£320M fortune</b> ...
Hypothesis: Product and geography are what make cream skimming work. <b>Entailment</b> , neutral, or contradiction?	Premise: Conceptually cream skimming has two basic dimensions – product and geography.	<b>Entailment</b>
Is this sentence <b>positive</b> or negative? (sentiment analysis)	A stirring, funny and finally transporting re-imagining of Beauty and the Beast and 1930s horror film.	<b>positive</b>



QA can be done by seq2seq



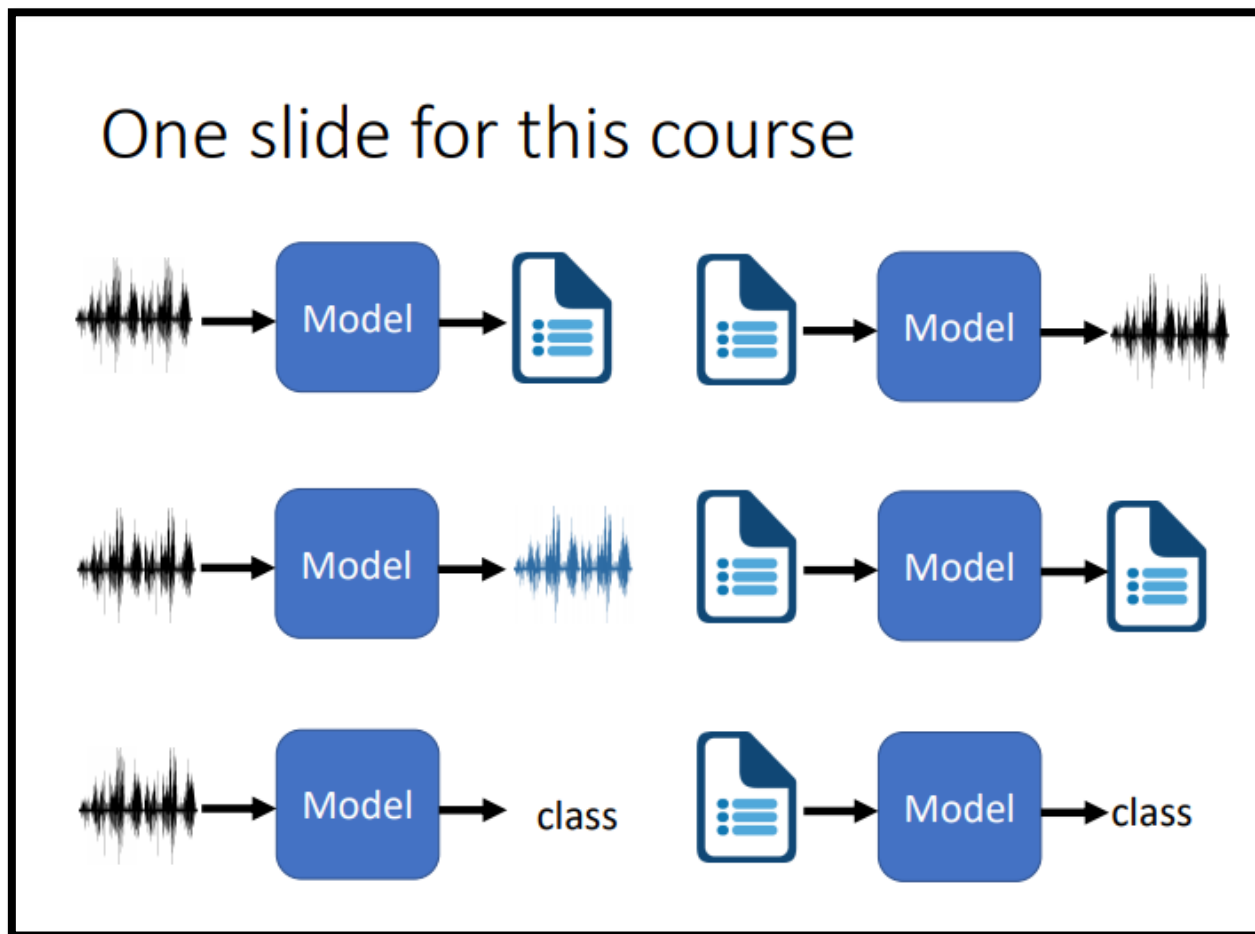
<https://arxiv.org/abs/1806.08730>

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



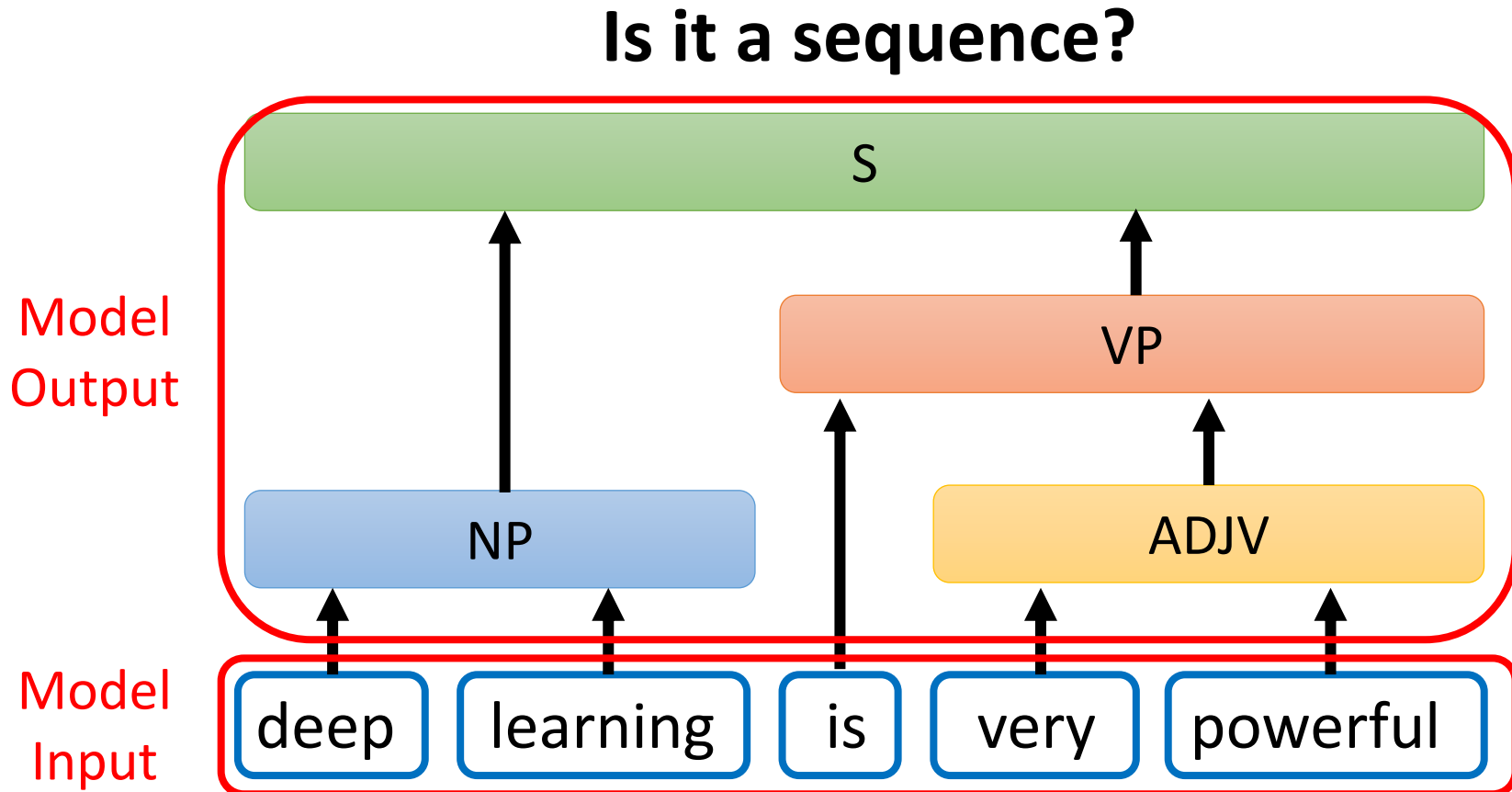
# Deep Learning for Human Language Processing

## 深度學習與人類語言處理



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

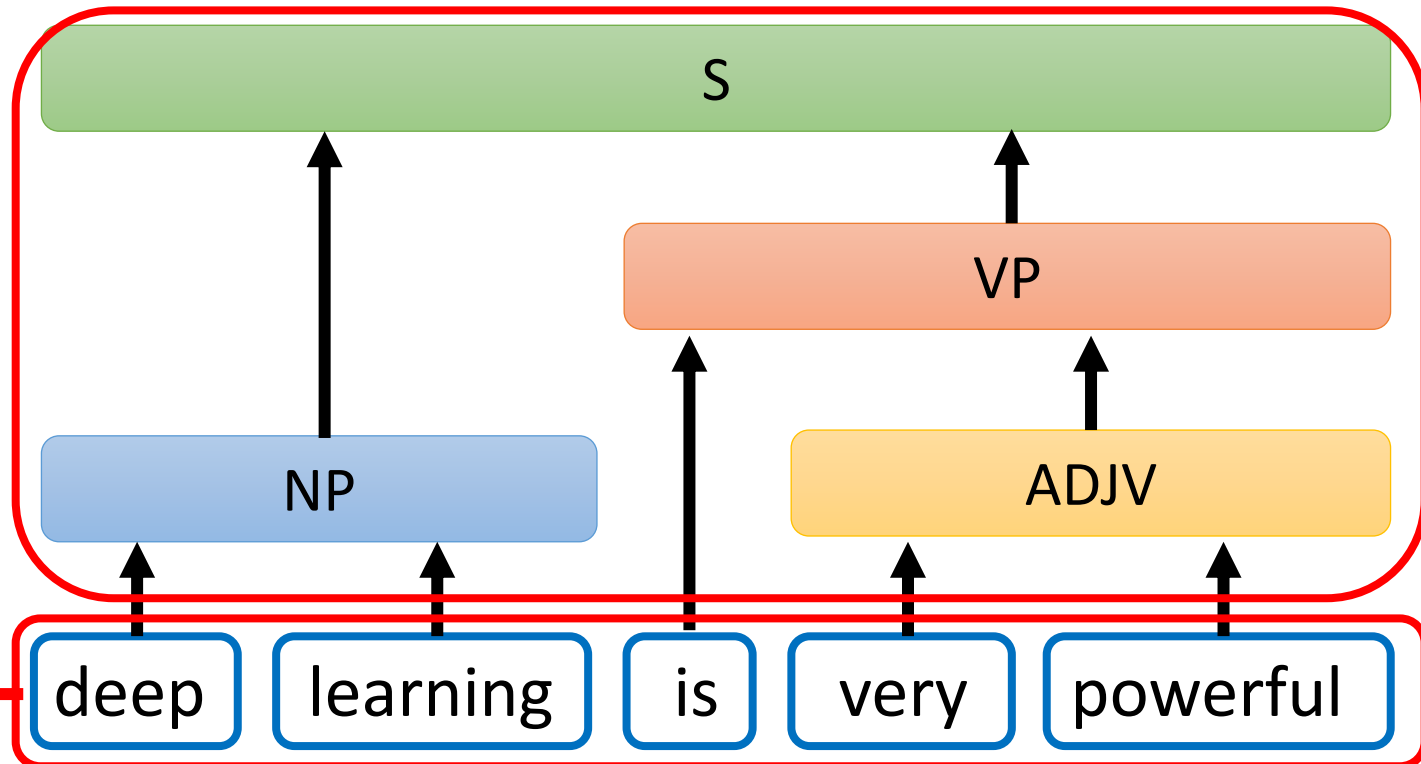
# Seq2seq for Syntactic Parsing



# Seq2seq for Syntactic Parsing

(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

# 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



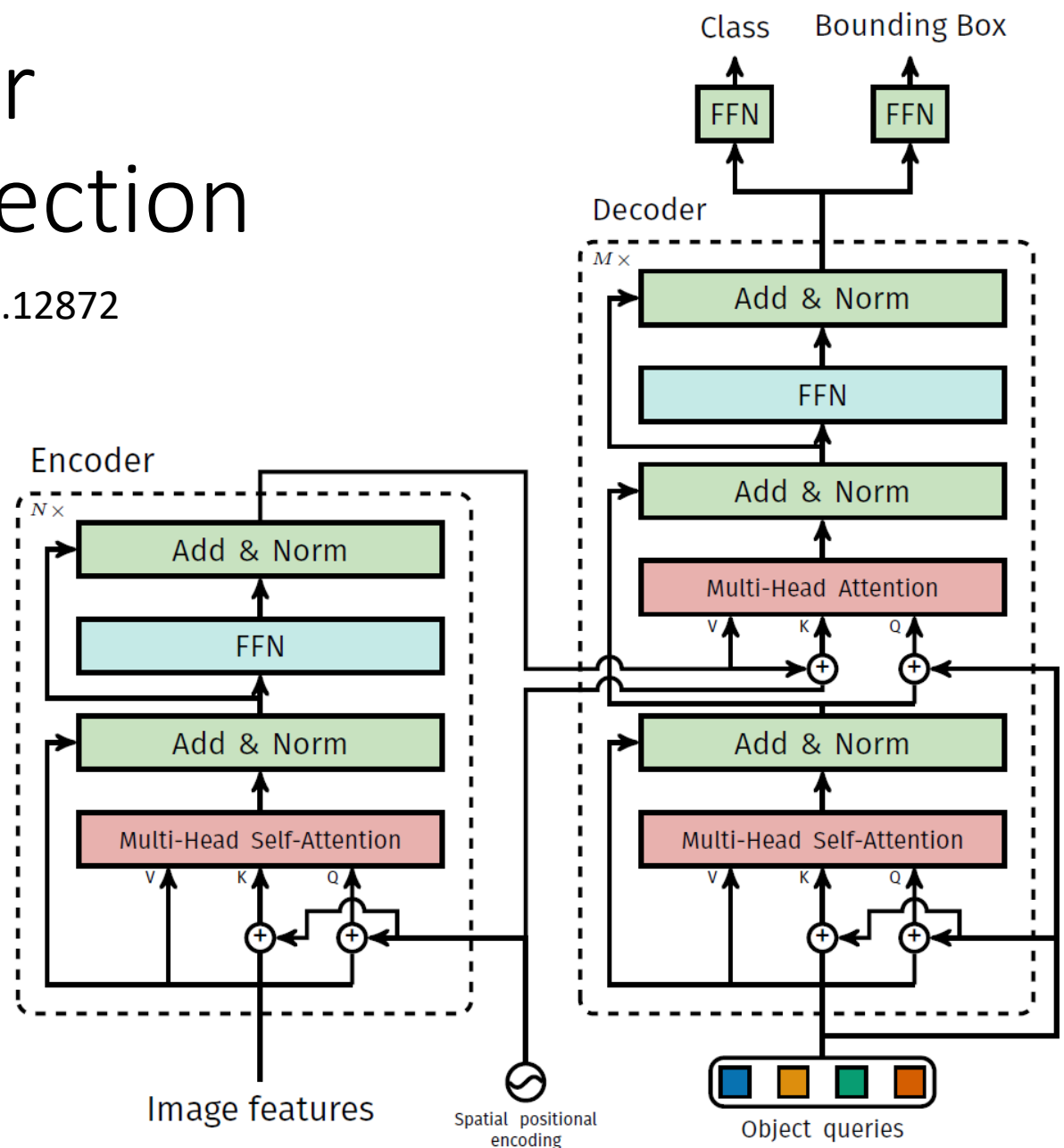
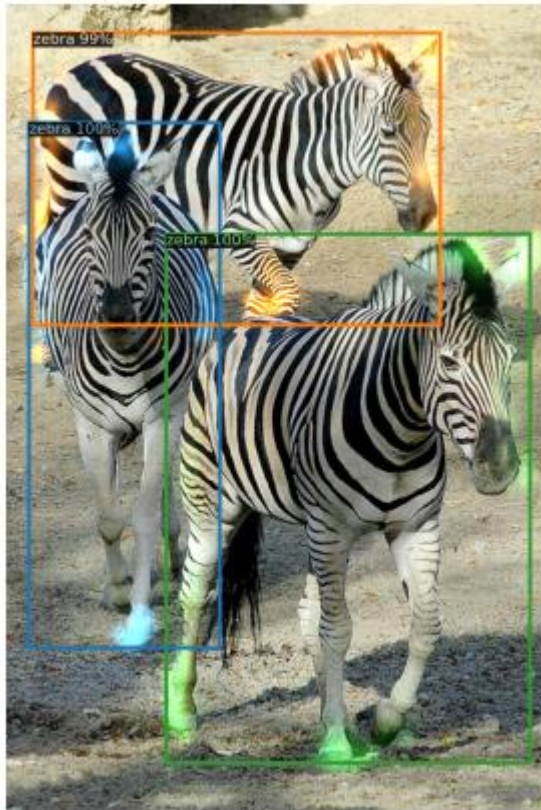
Class 13

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

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

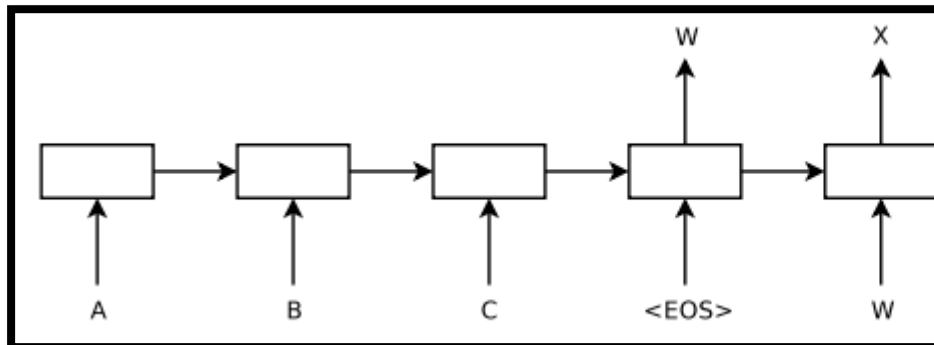
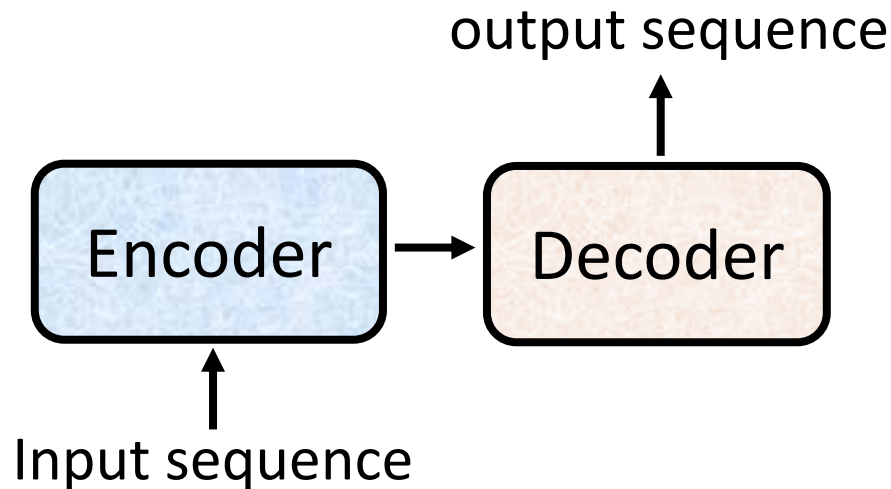
# Seq2seq for Object Detection

<https://arxiv.org/abs/2005.12872>



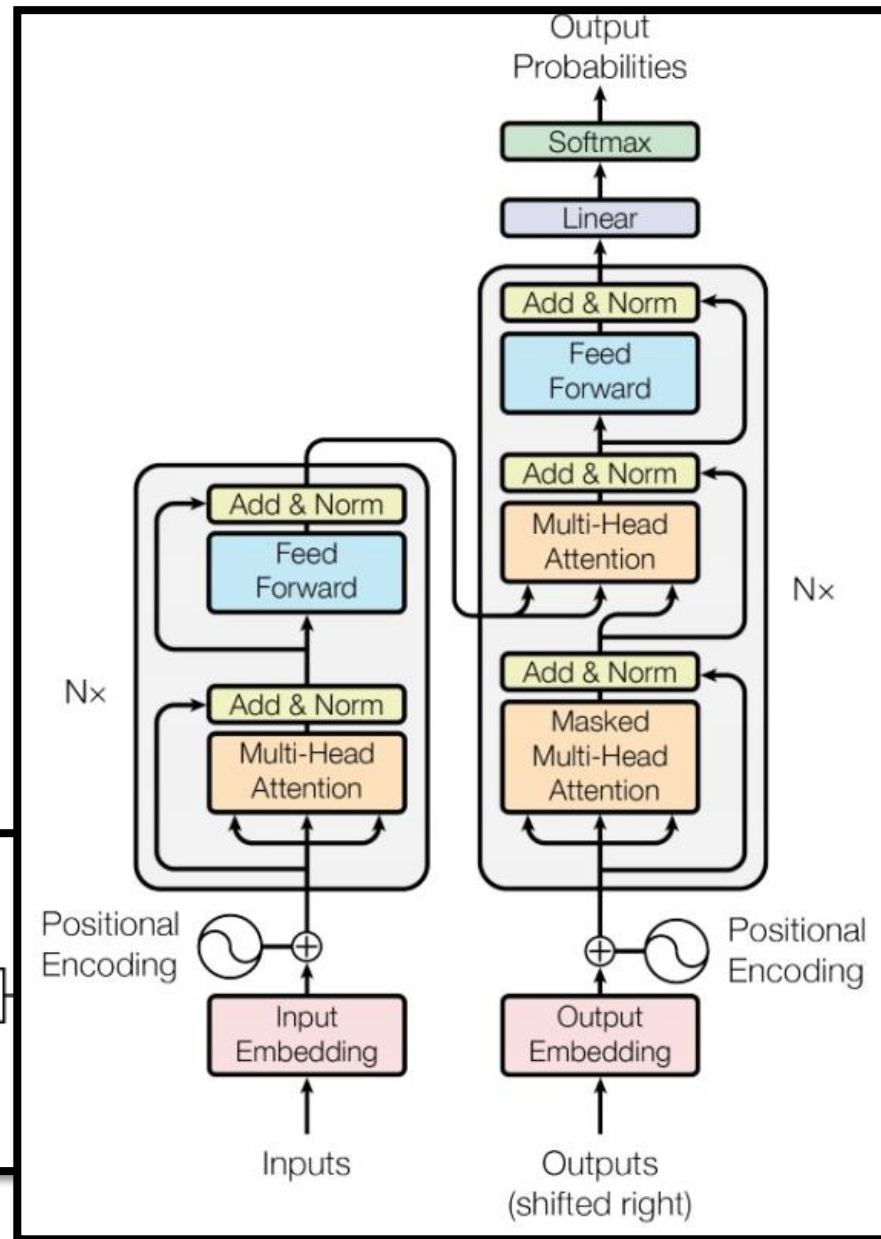


# Seq2seq



Sequence to Sequence Learning with Neural Networks

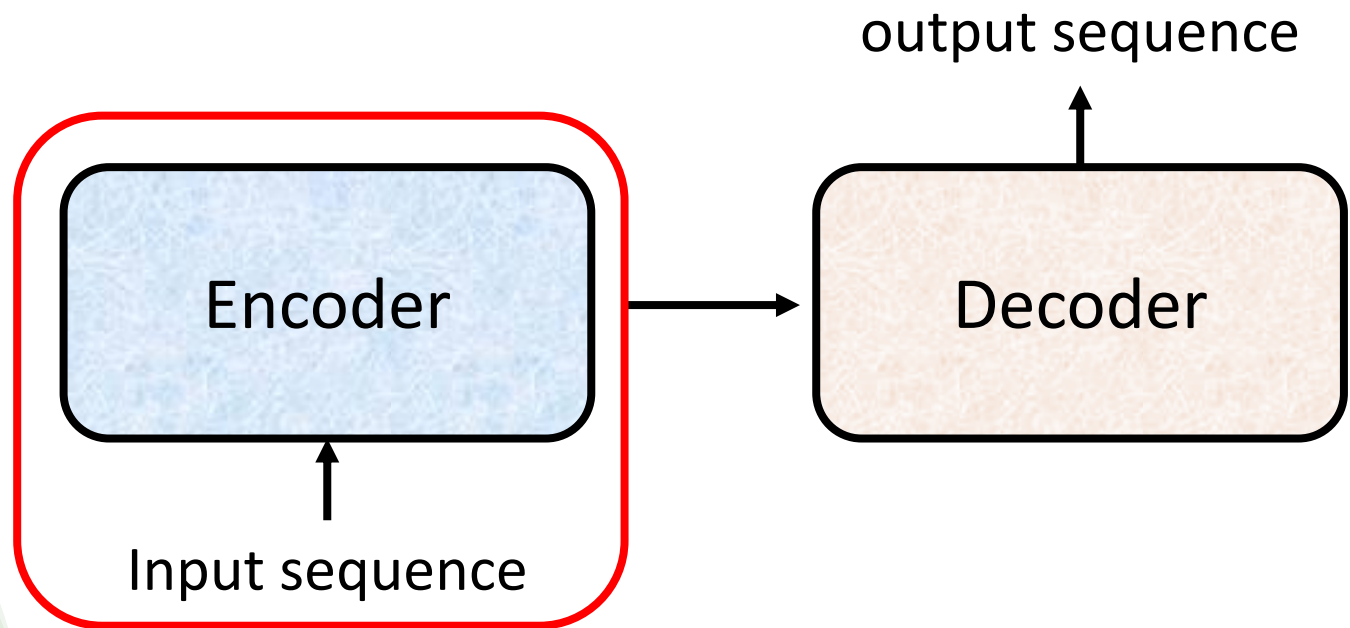
<https://arxiv.org/abs/1409.3215>



## Transformer

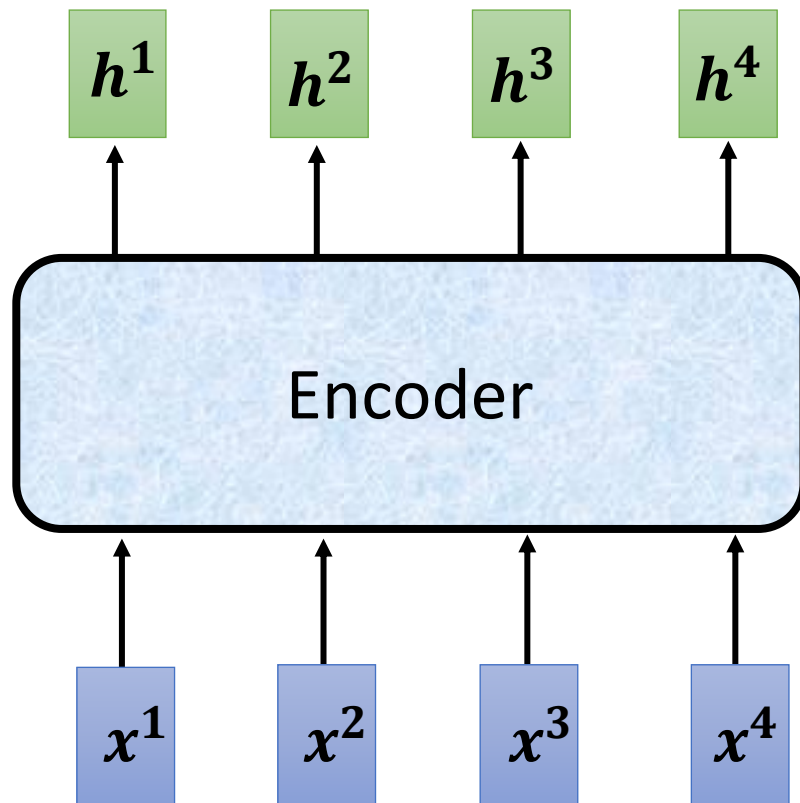
<https://arxiv.org/abs/1706.03762>

# Encoder

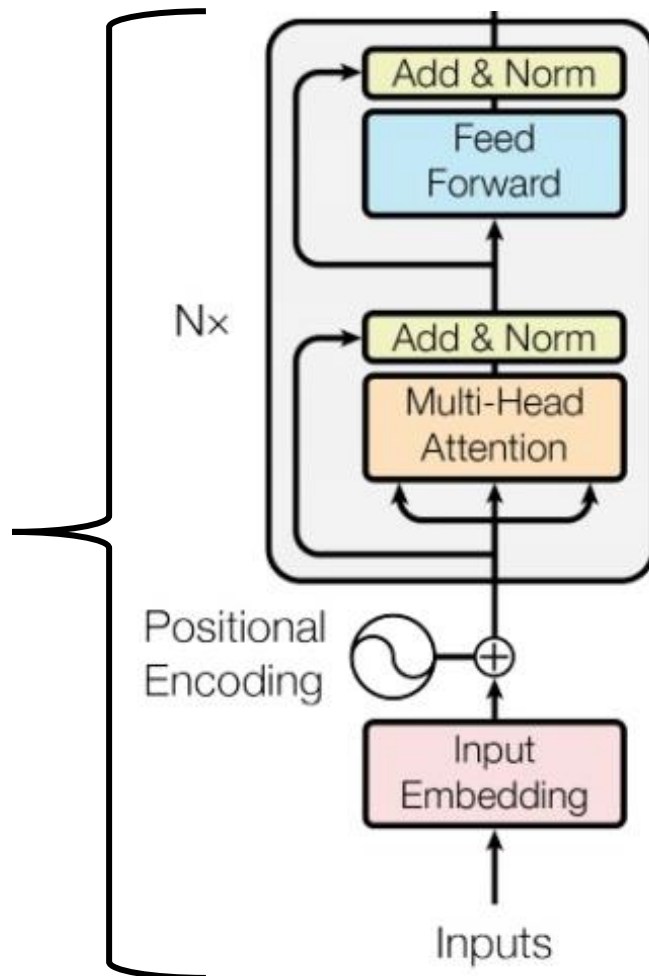


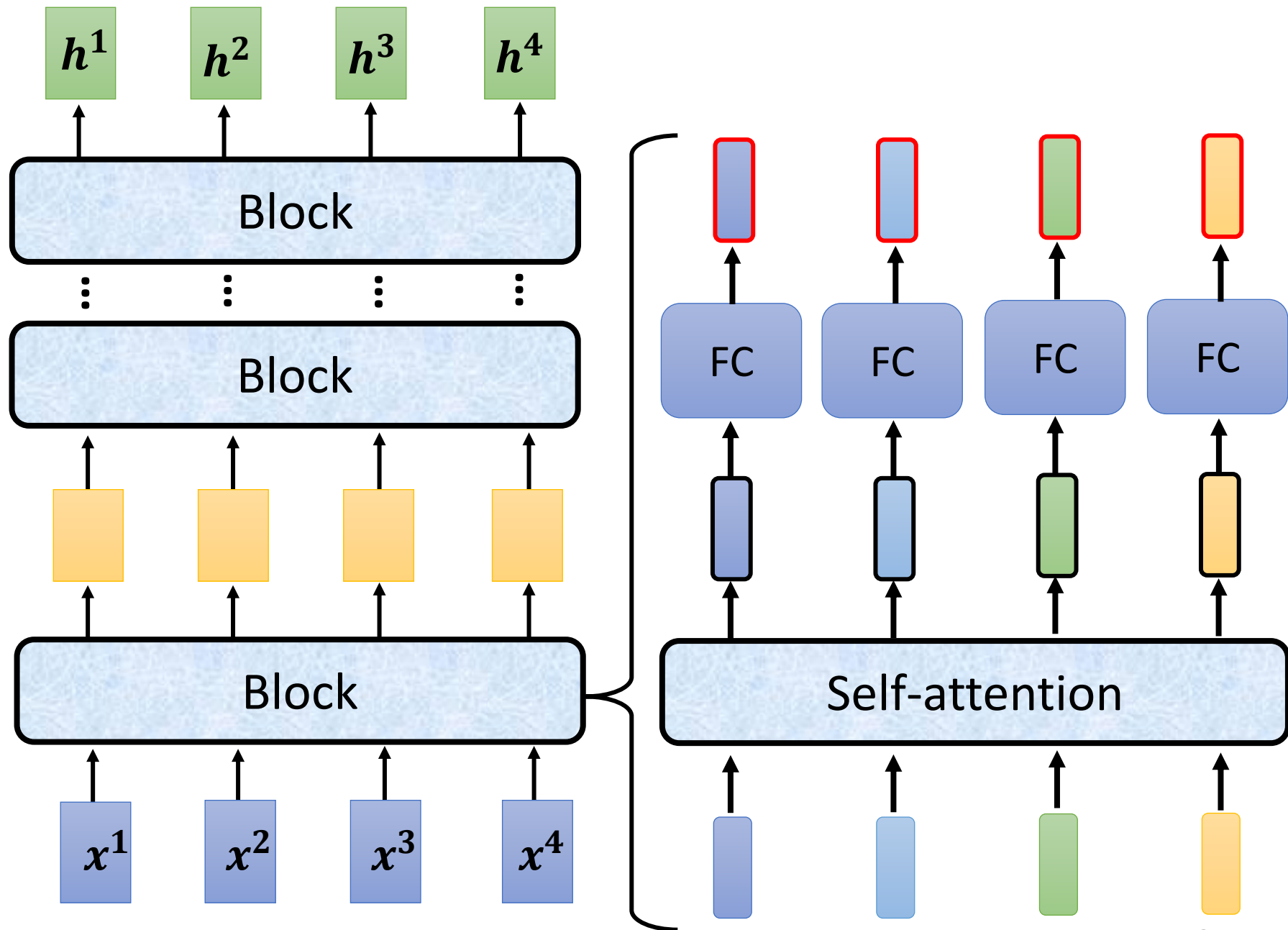
# Encoder

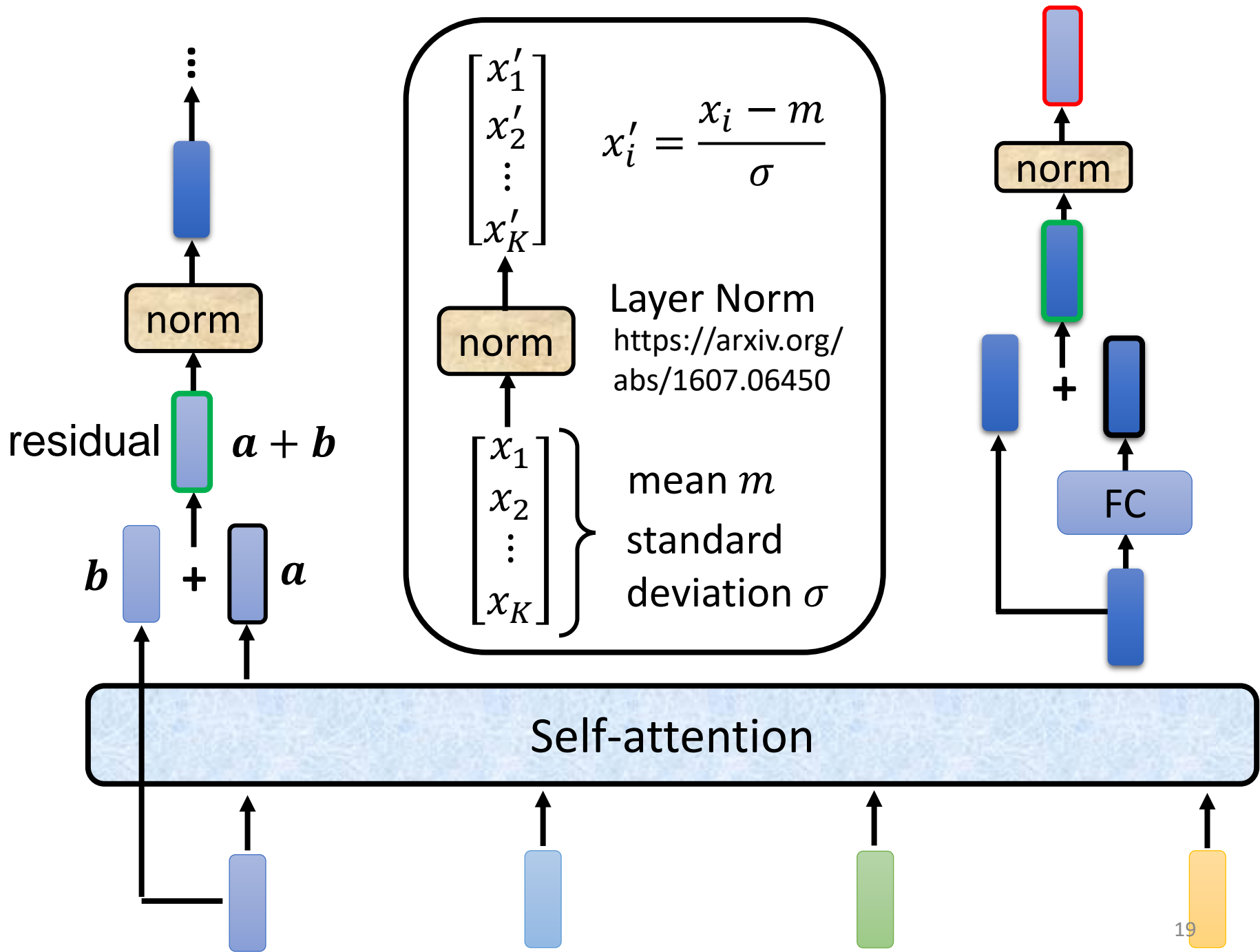
You can use **RNN** or **CNN**.



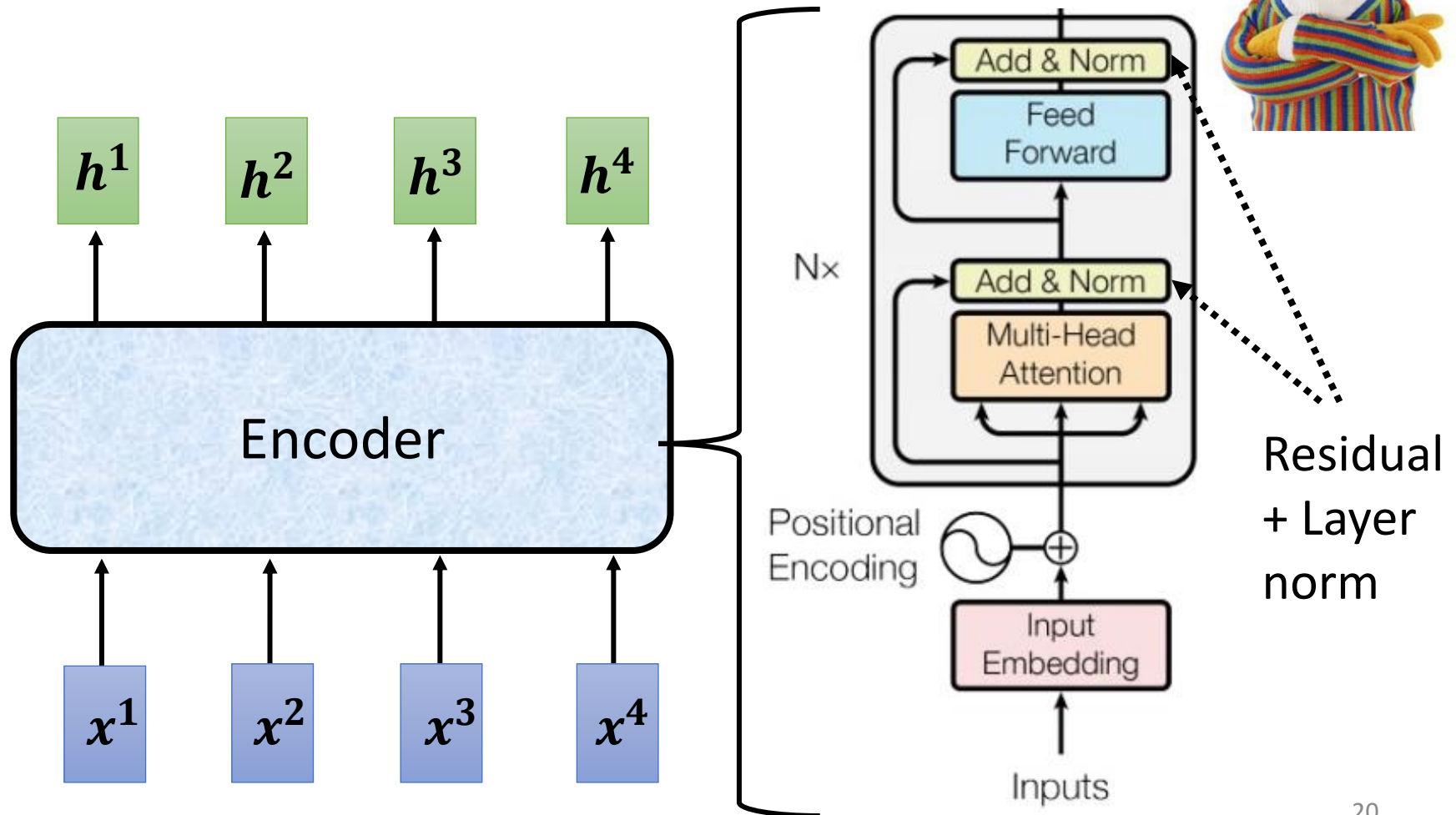
## Transformer's Encoder







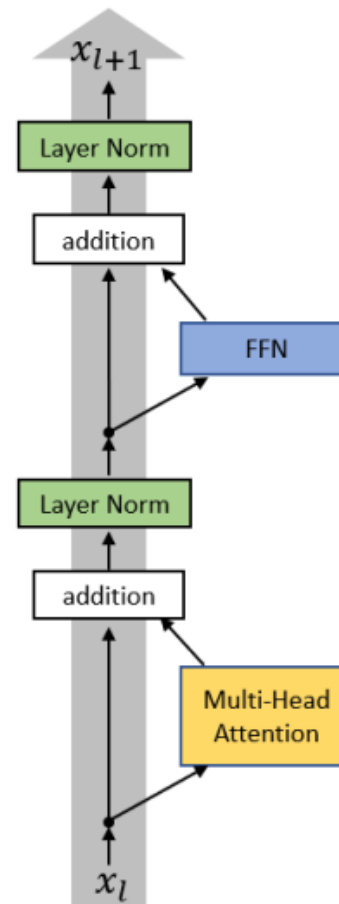
I use the **same** network architecture as **transformer encoder**.



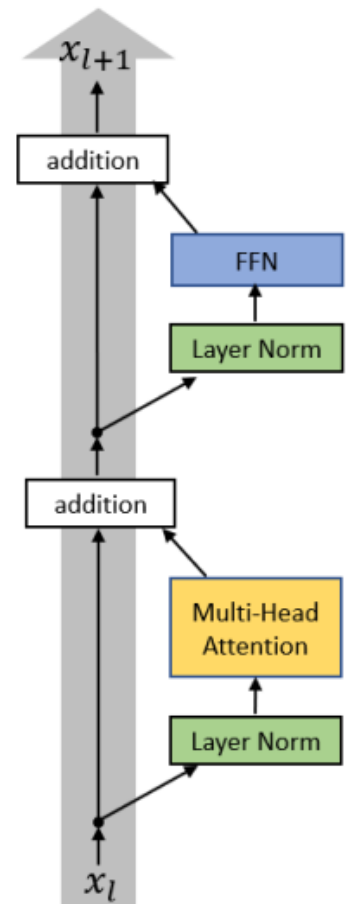


# To learn more .....

- 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>

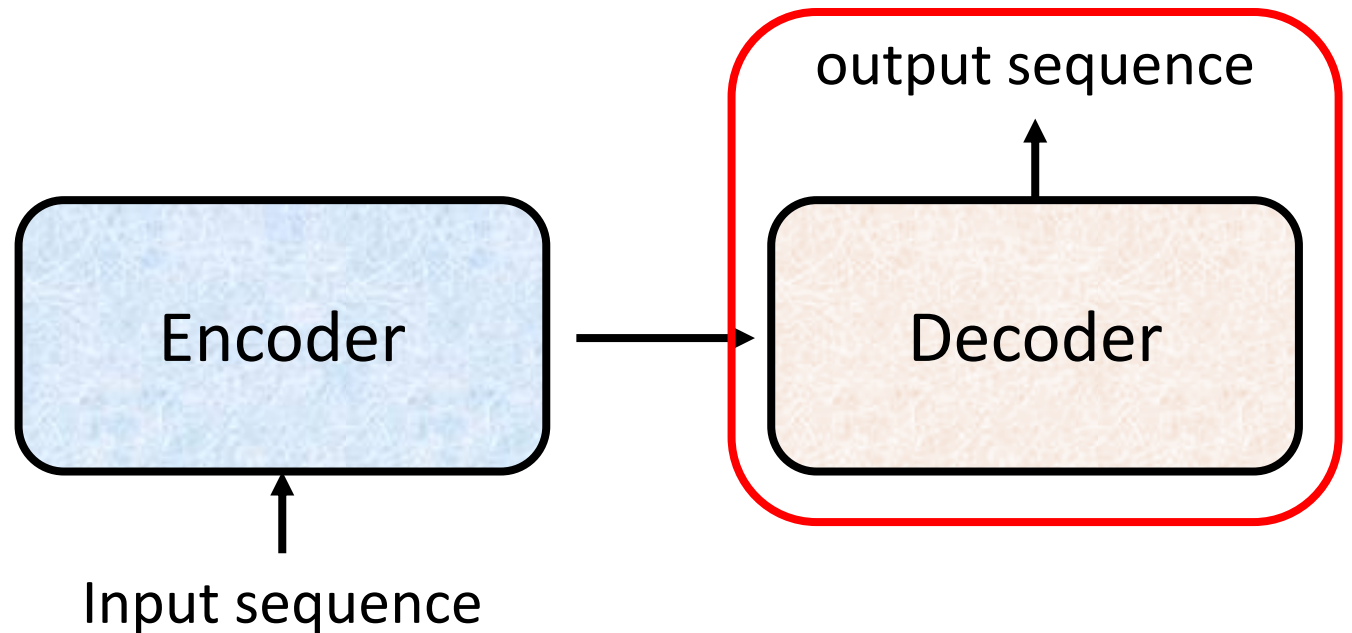


(a)



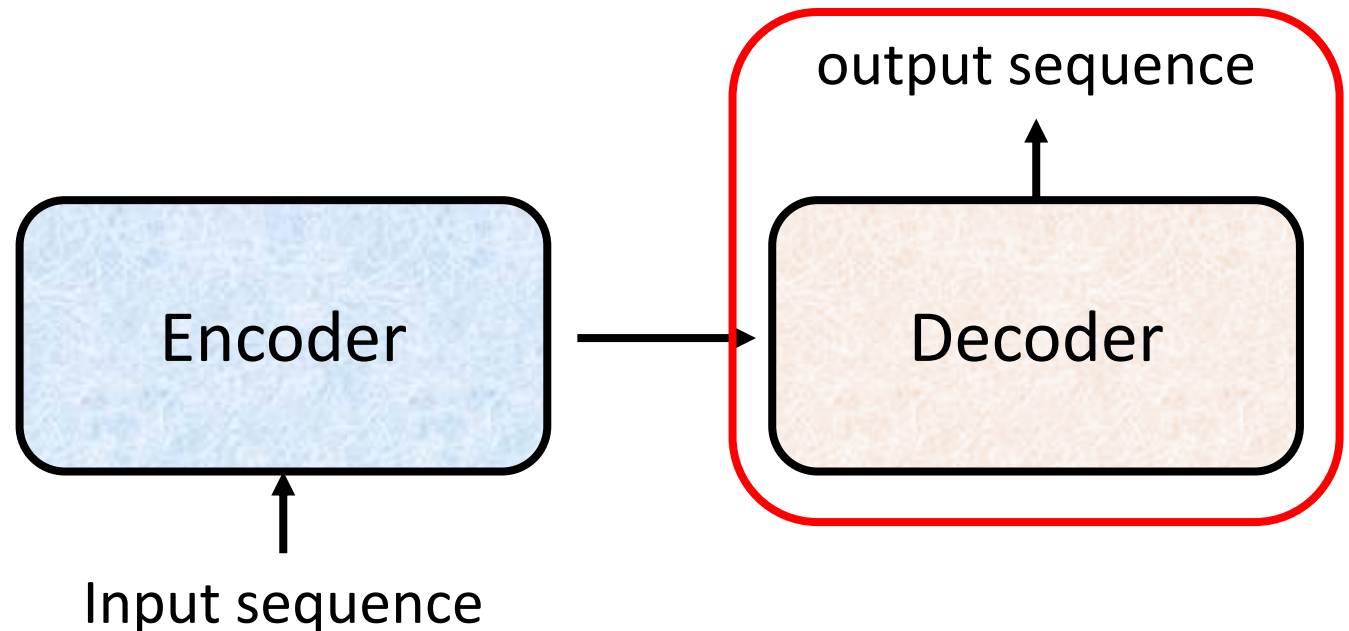
(b)

# Decoder



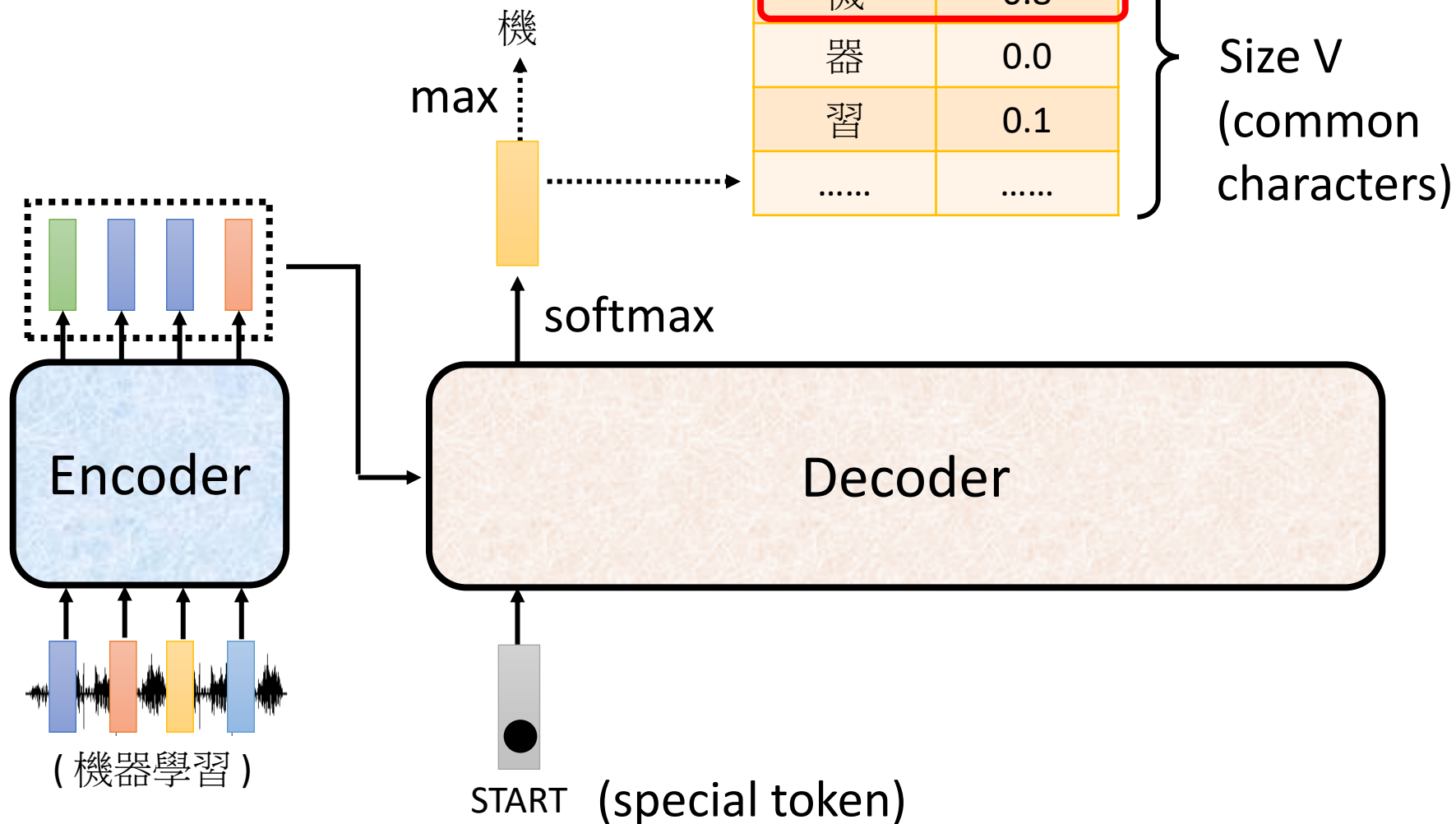
# Decoder

- Autoregressive (AT)

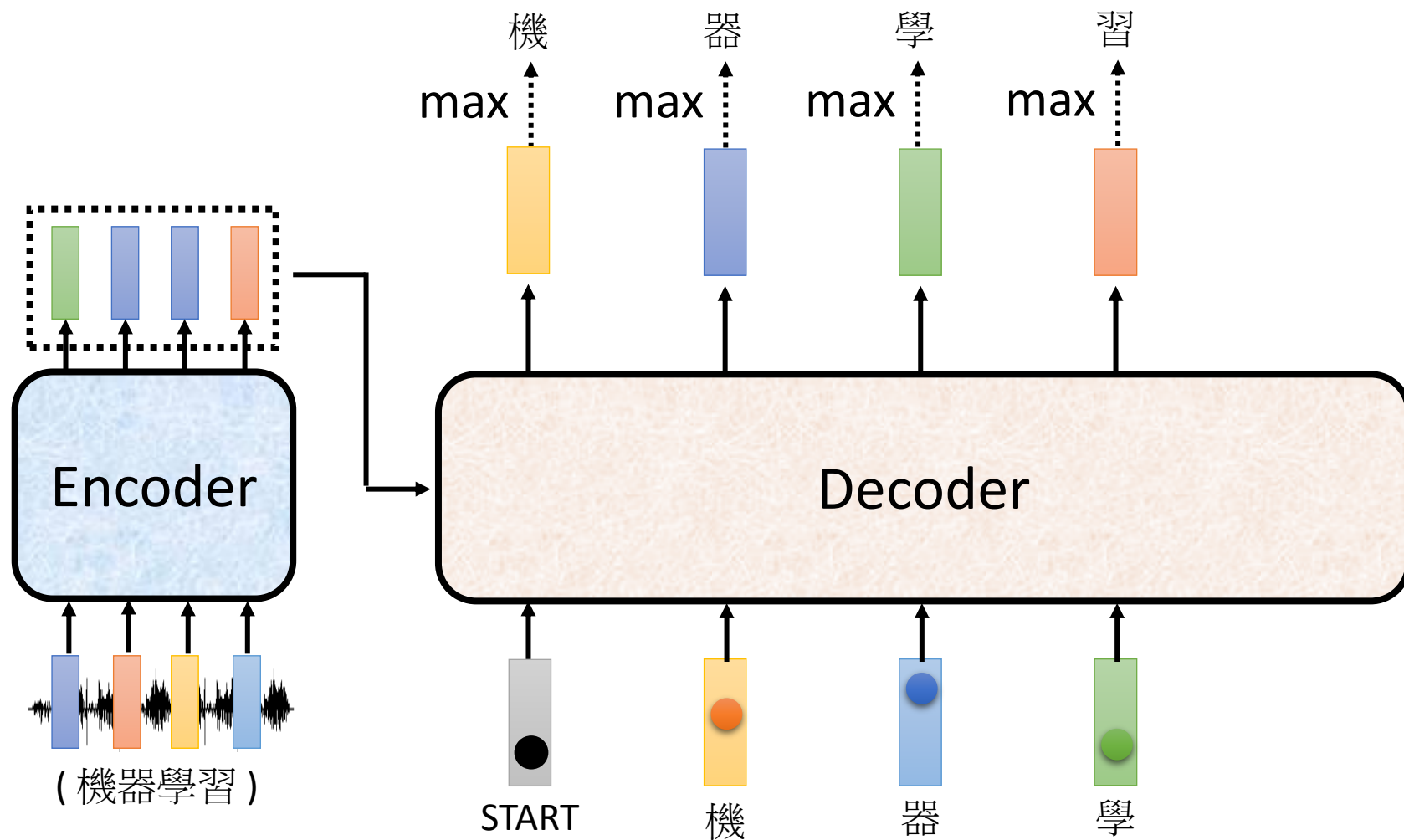


# Autoregressive

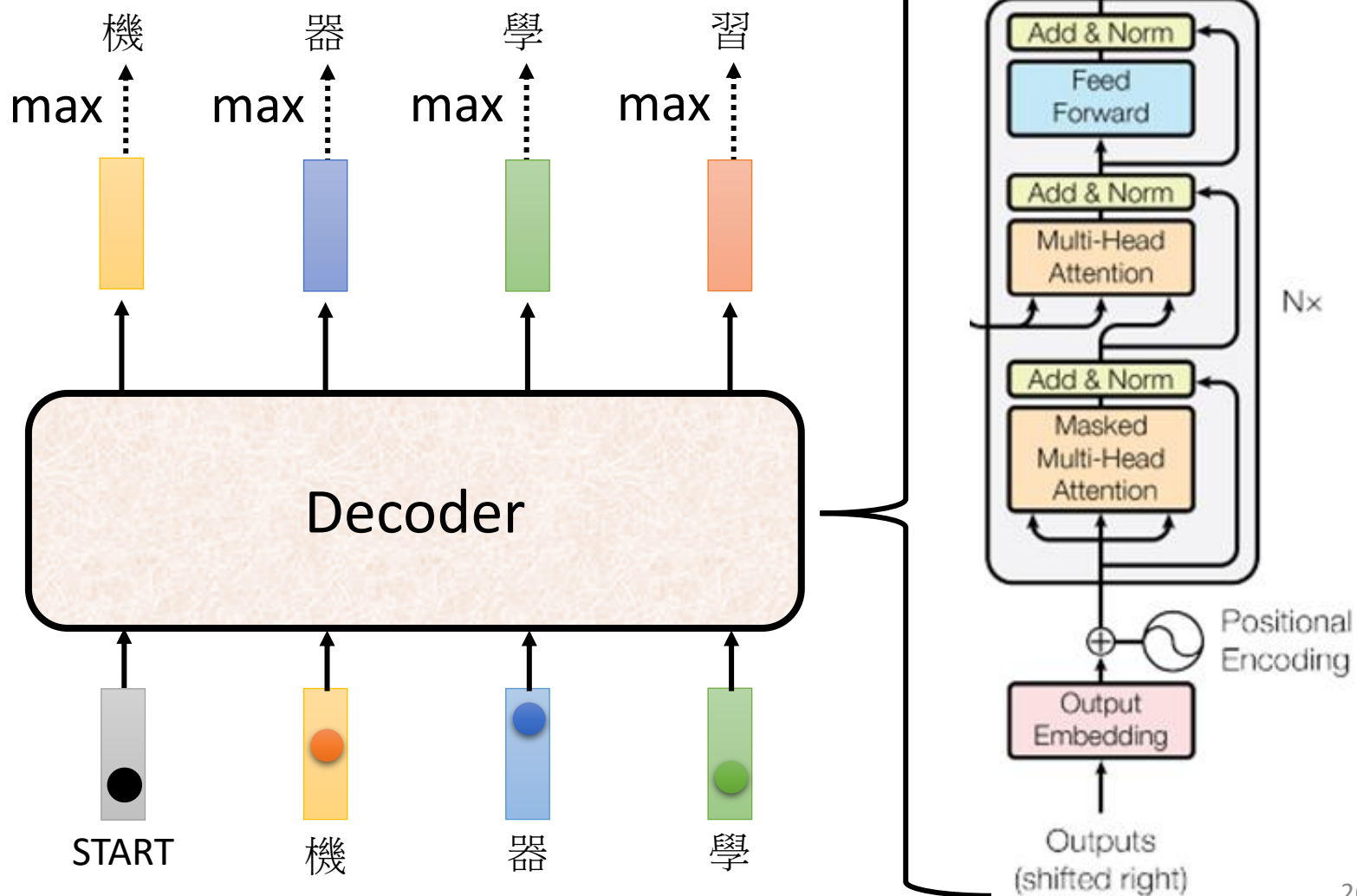
(Speech Recognition as example)



# Autoregressive

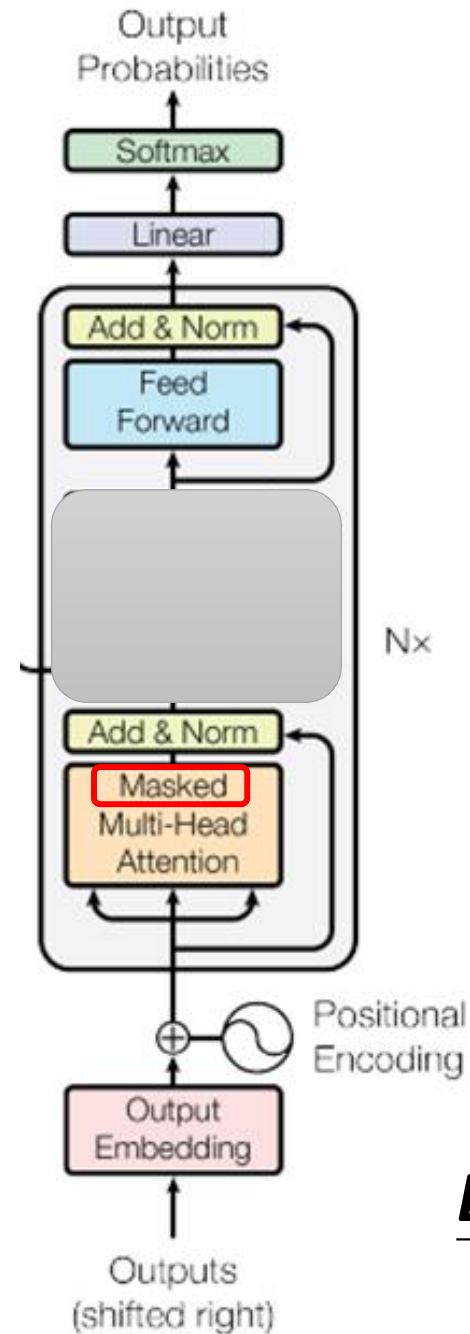
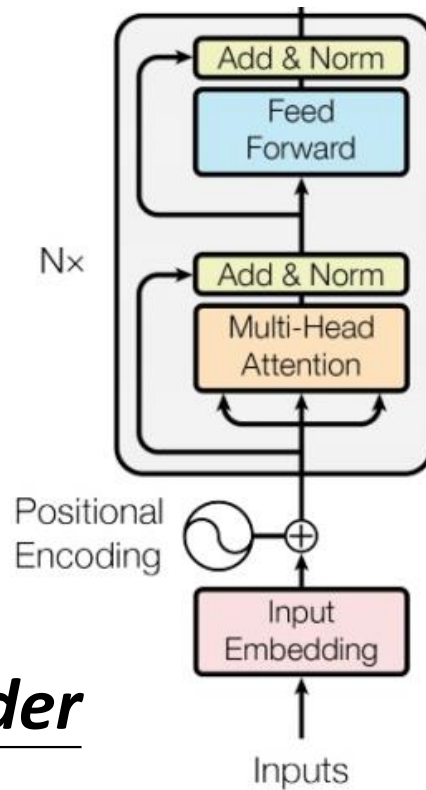


ignore the input from the encoder here ☺



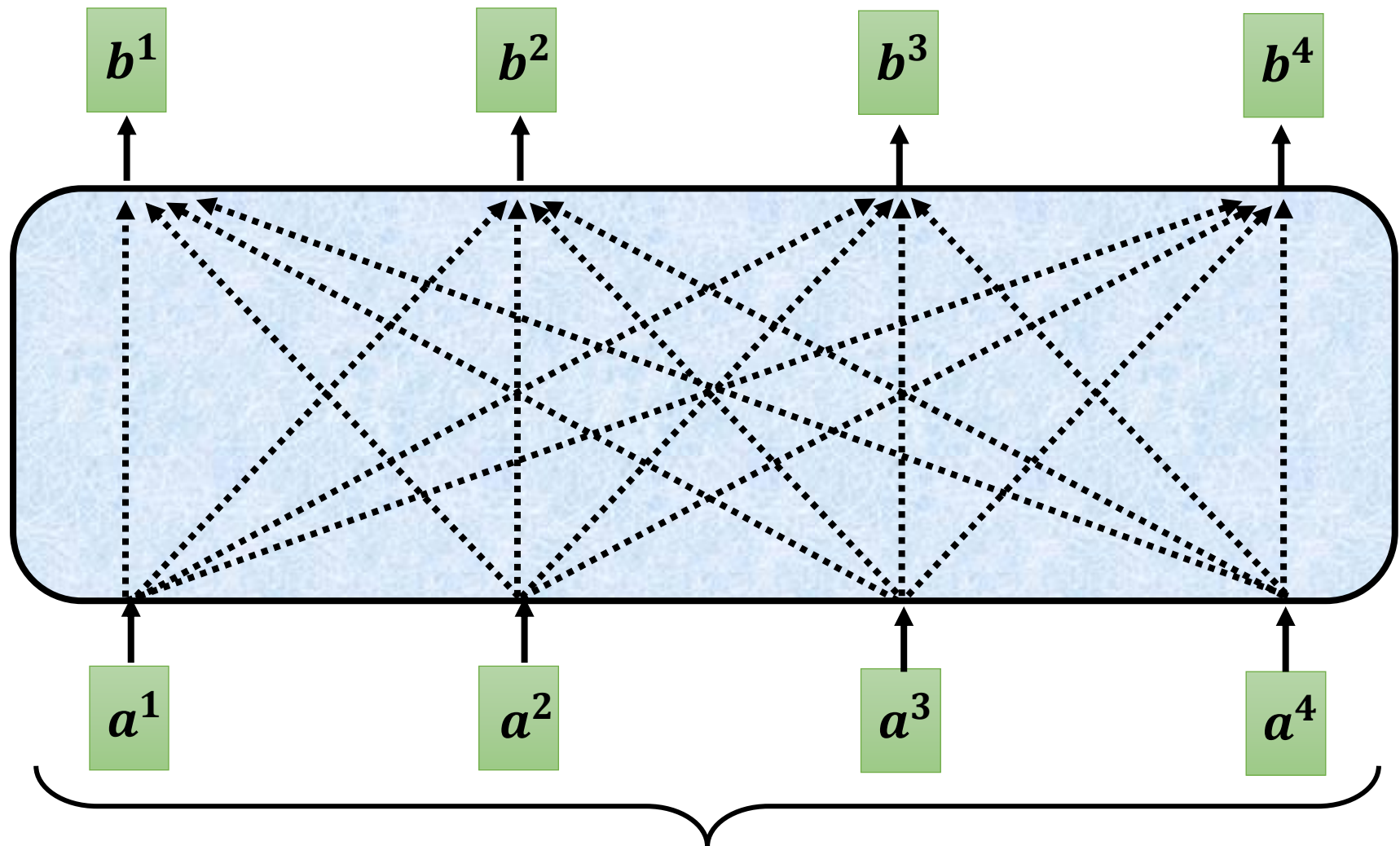


## Encoder



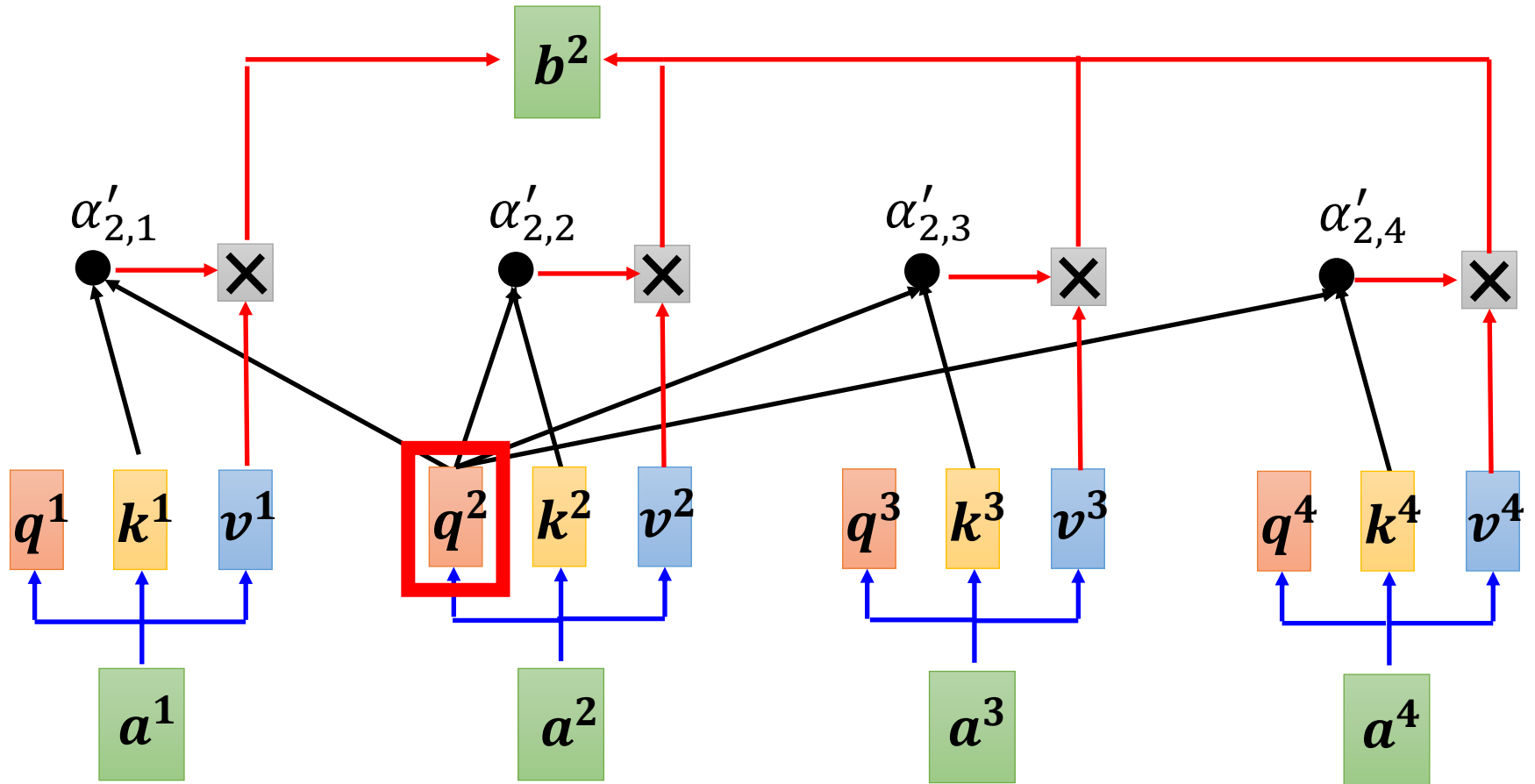
## Decoder

# Self-attention → Masked Self-attention



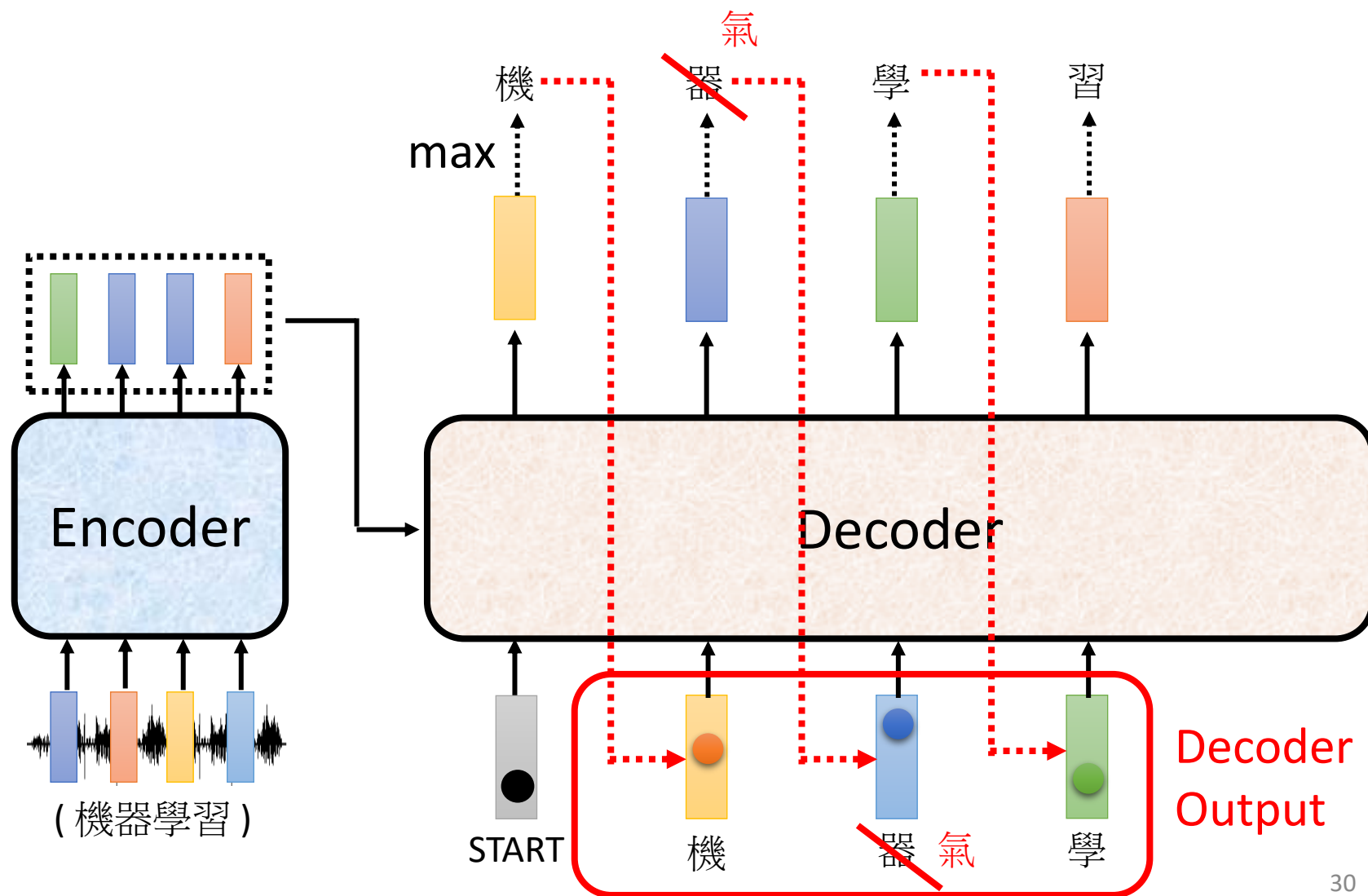
Can be either **input** or a **hidden layer**

# Self-attention $\rightarrow$ Masked Self-attention



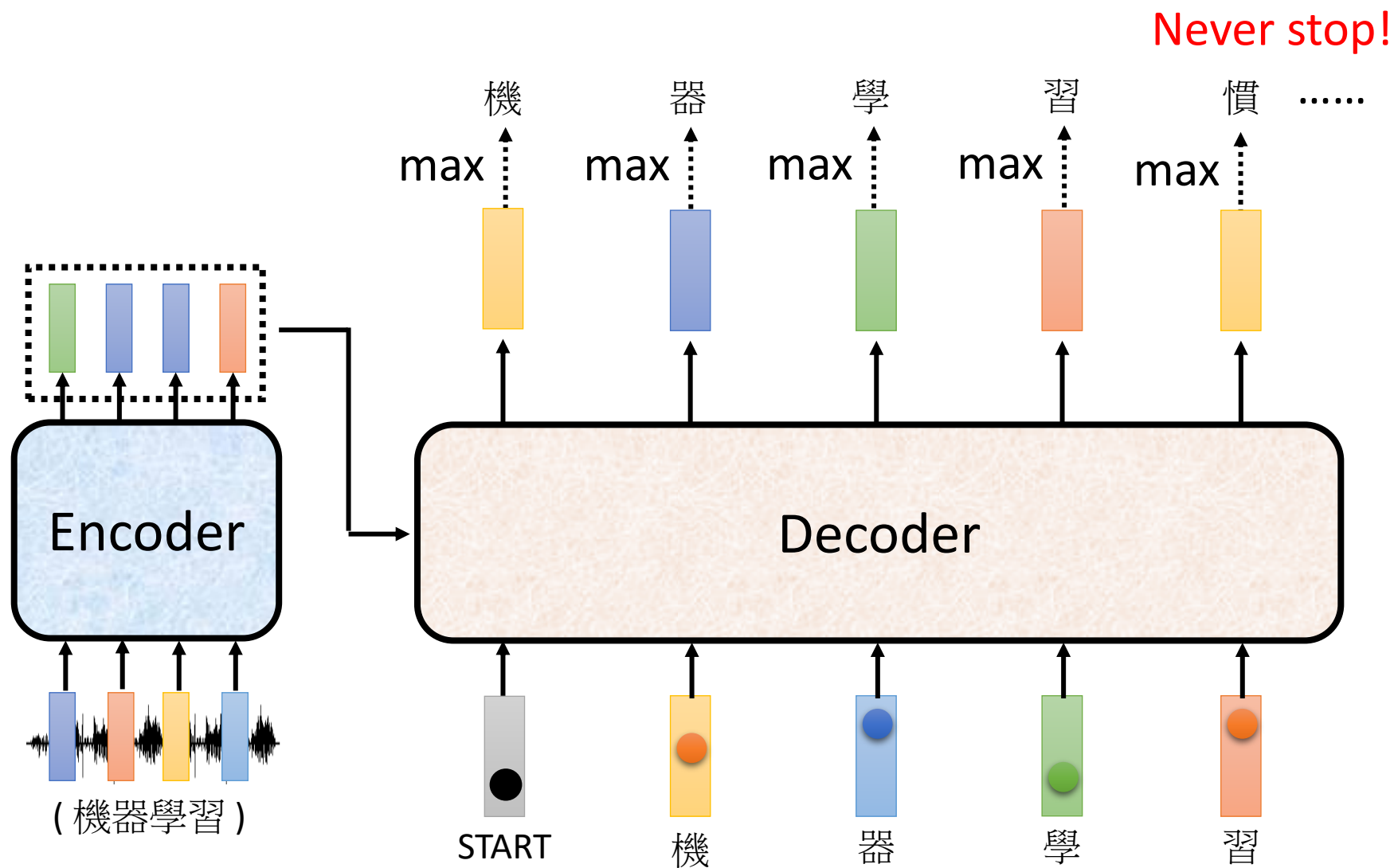
Why masked? Consider how does decoder work

# Autoregressive



# Autoregressive

We do not know the correct output length.

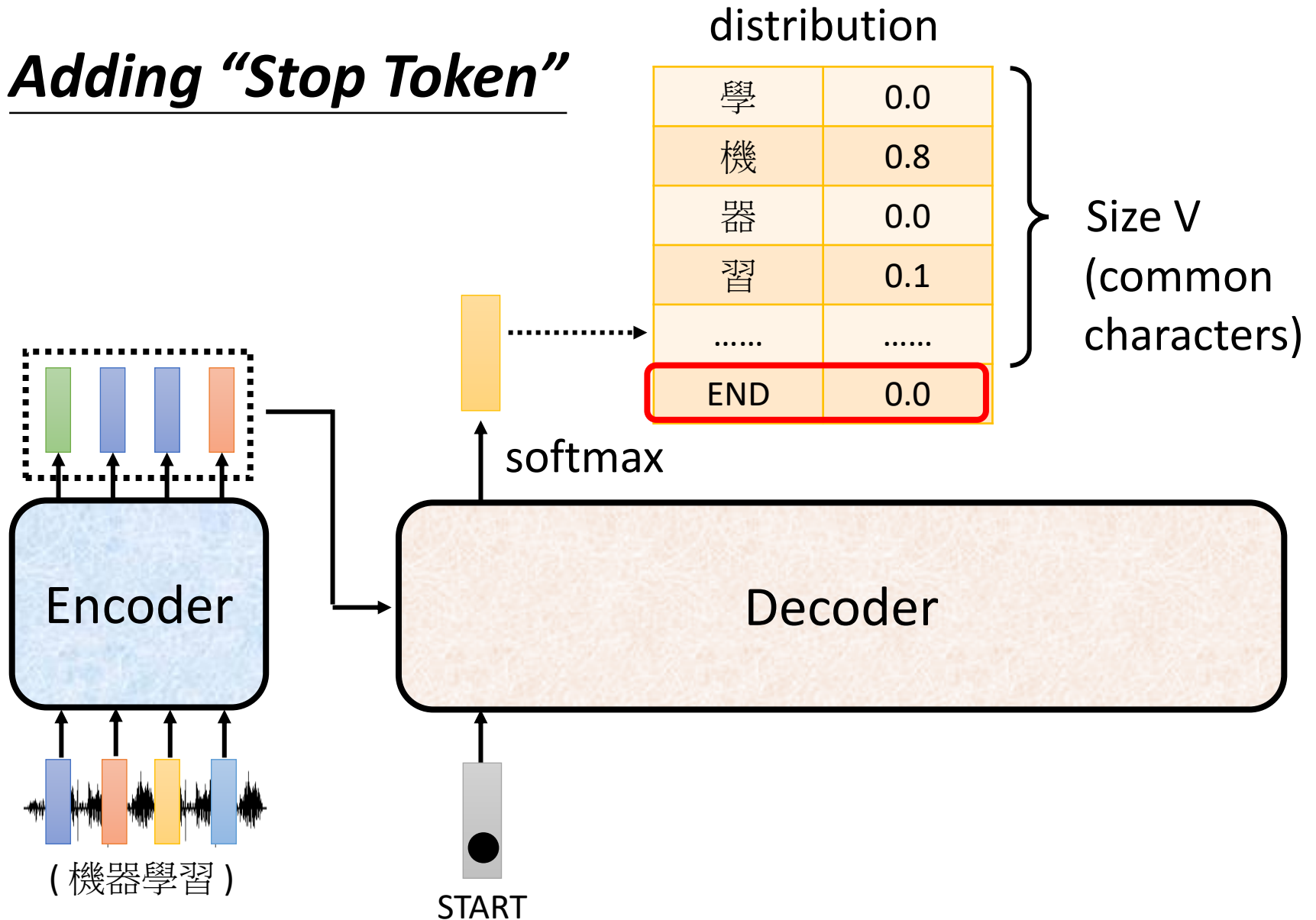


# 推文接龍 (Tweet Solitaire)

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	tlkagk:	=====斷=====	

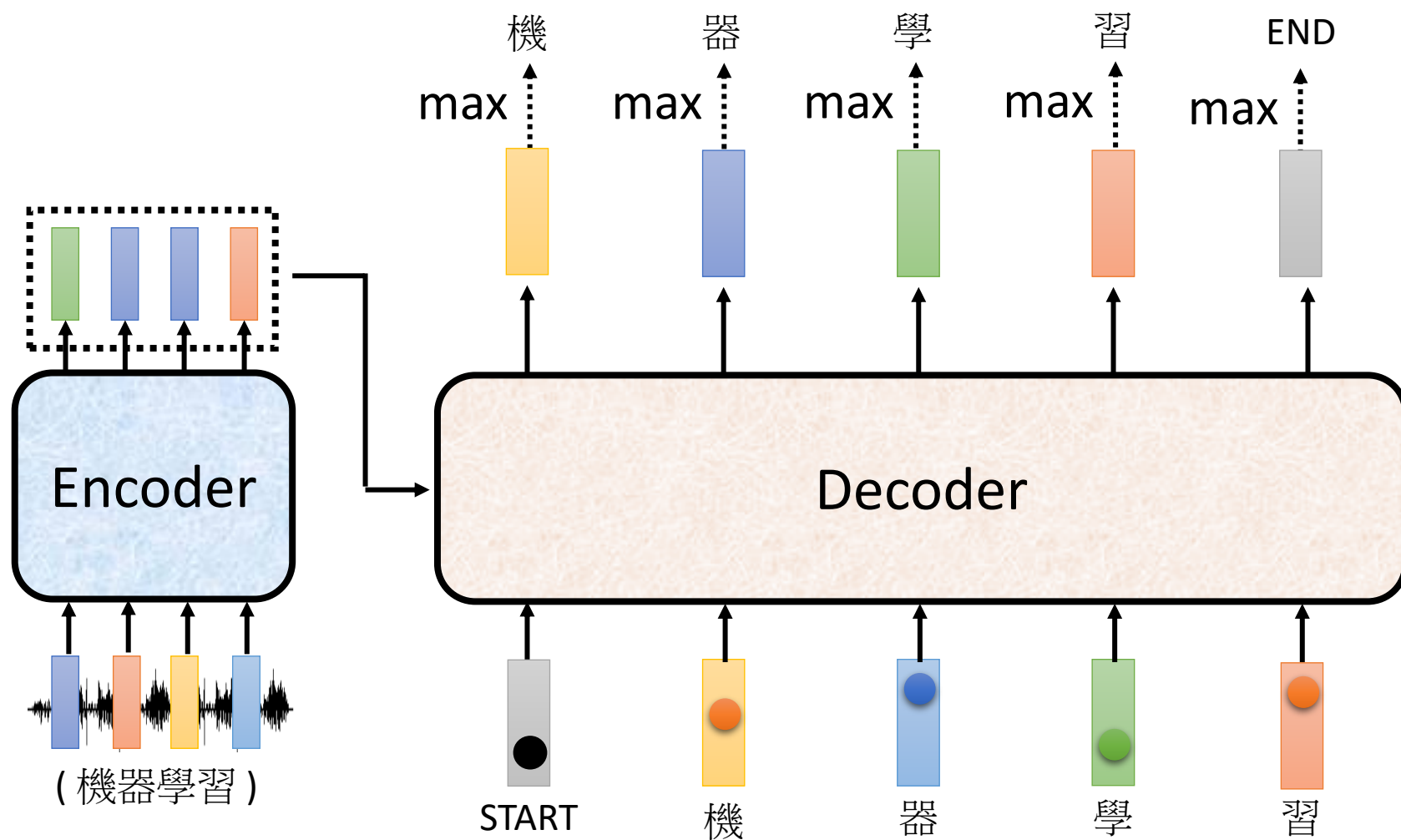


## Adding “Stop Token”



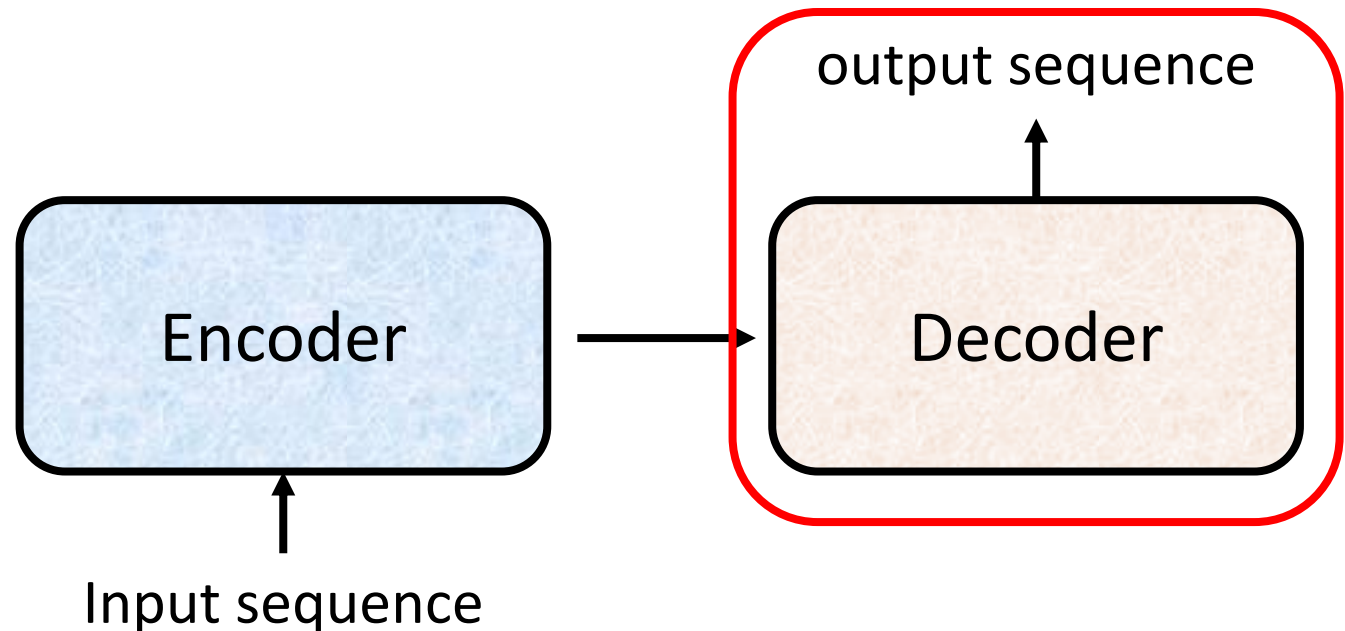
# Autoregressive

Stop at here!

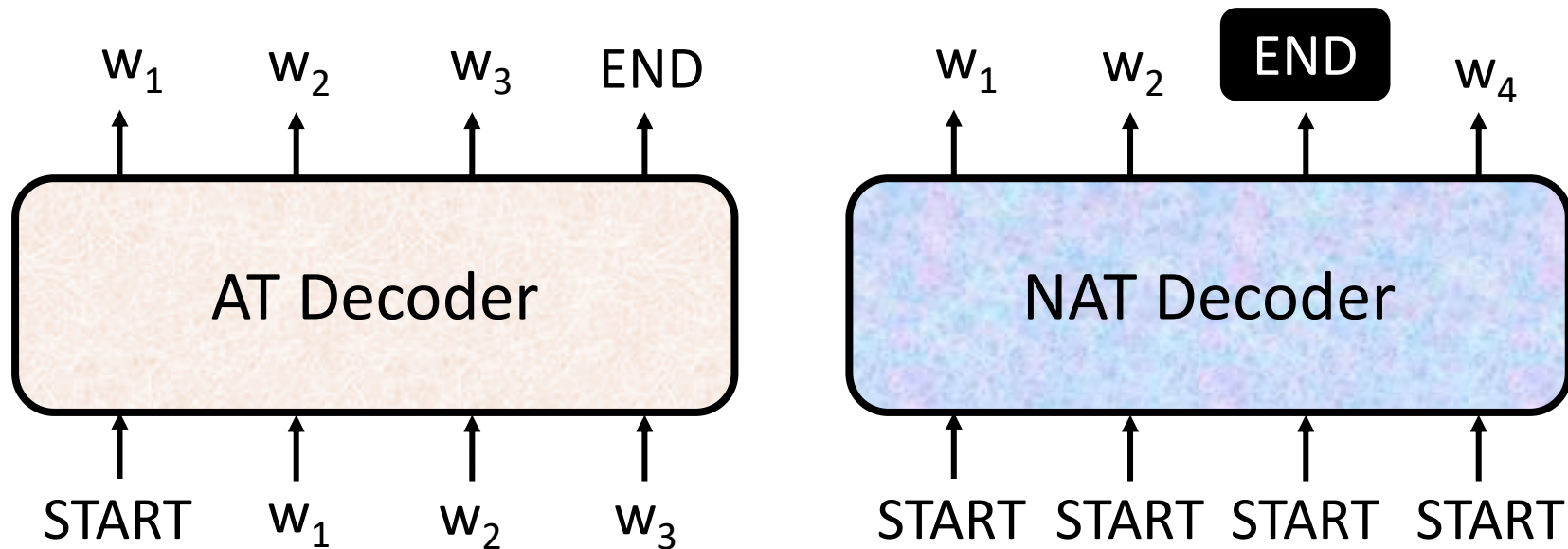


# Decoder

- Non-autoregressive (NAT)



# AT v.s. NAT



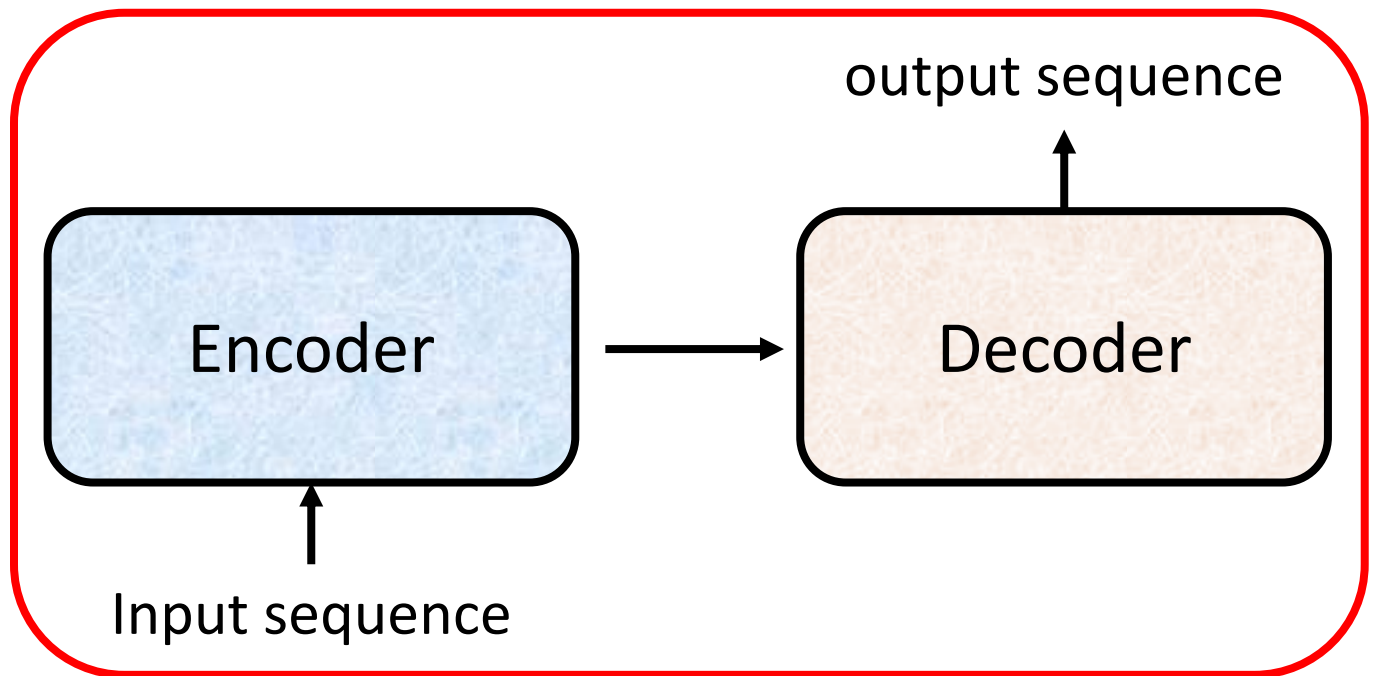
- How to decide the output length for NAT decoder?
  - Another predictor for output length
  - Output a very long sequence, ignore tokens after END
- Advantage: parallel, more stable generation (e.g., TTS)
- NAT is usually worse than AT (why? **Multi-modality**)

To learn more .....

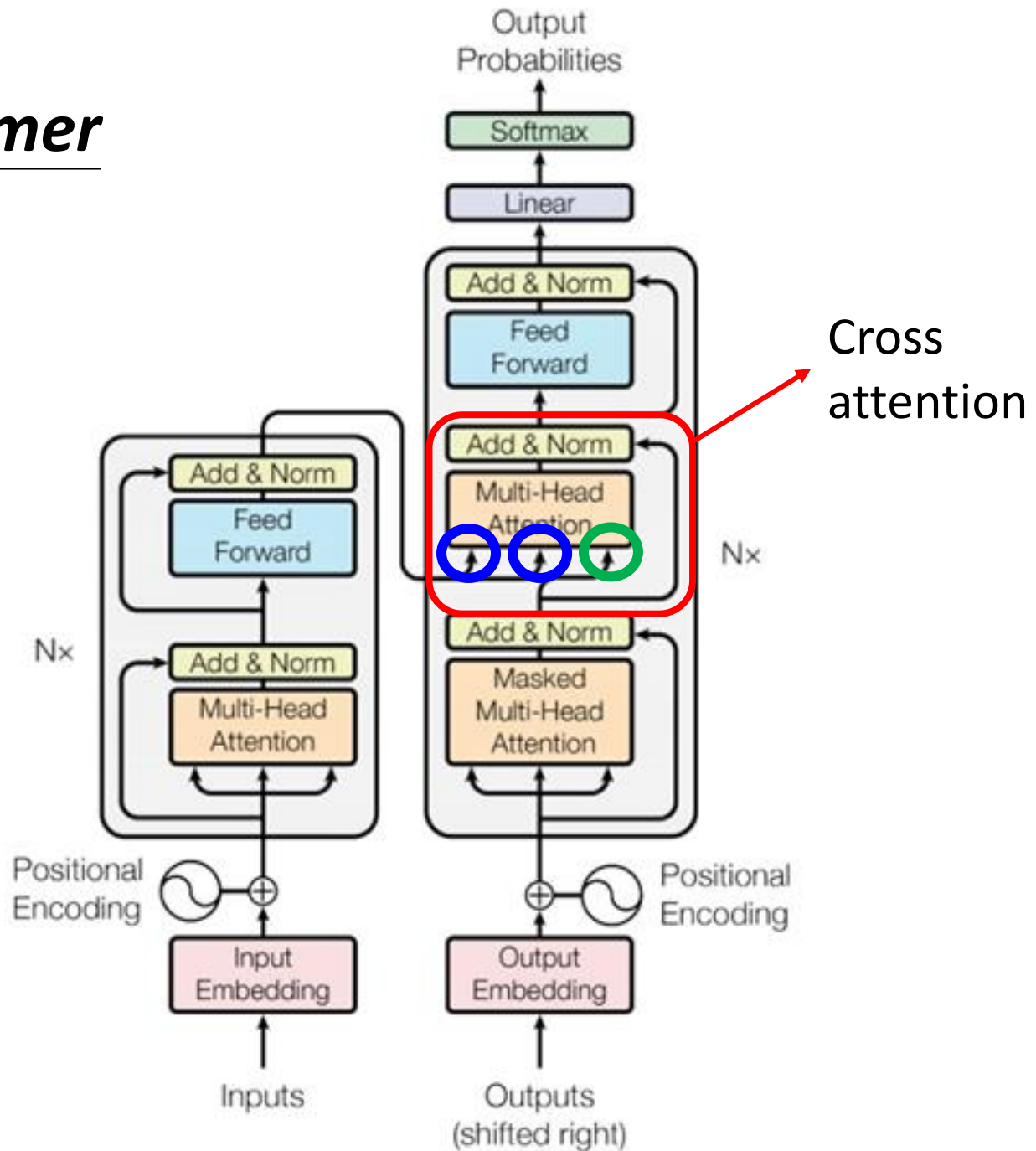


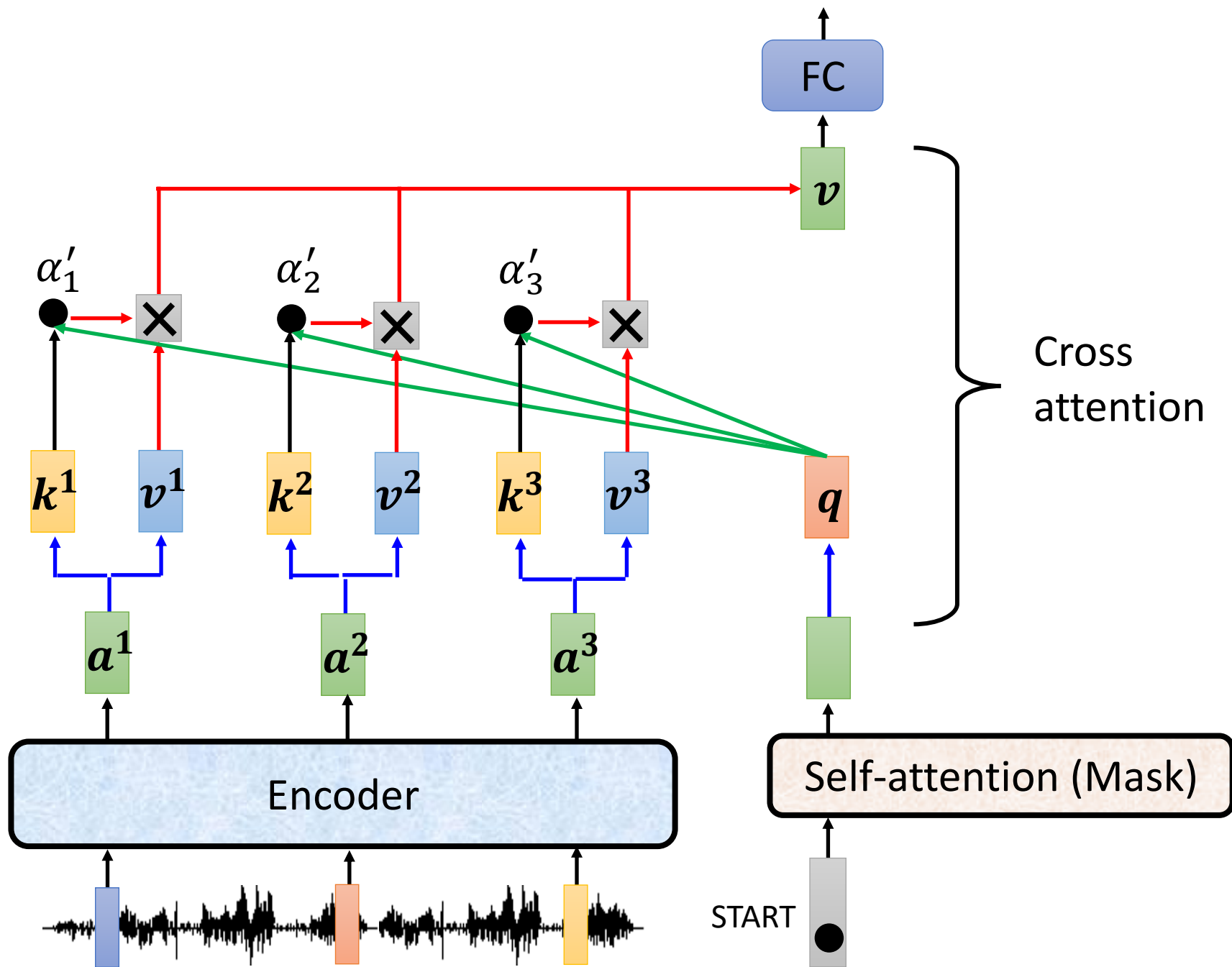
<https://youtu.be/jvyKmU4OM3c>  
(in Mandarin)

# Encoder-Decoder

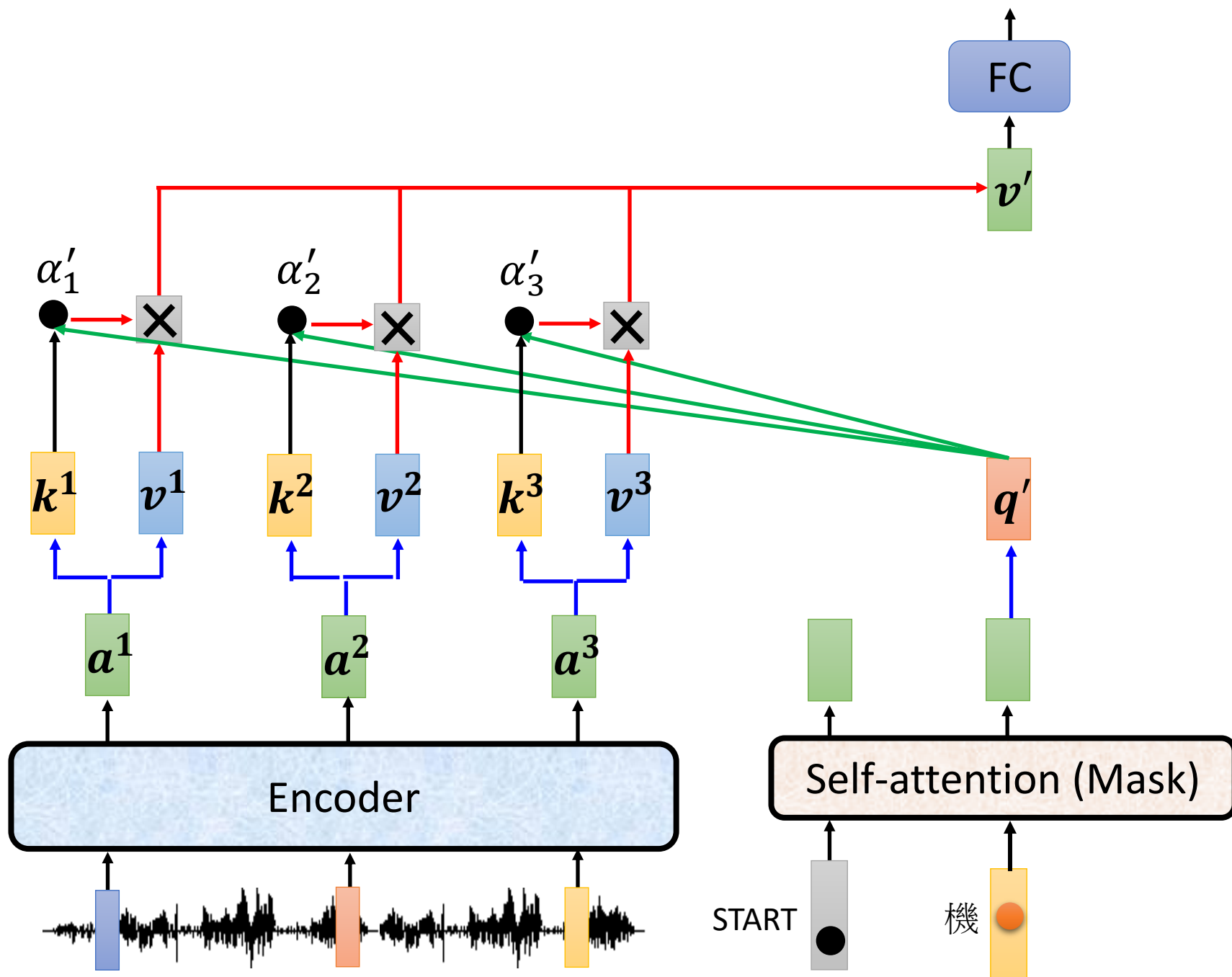


# Transformer





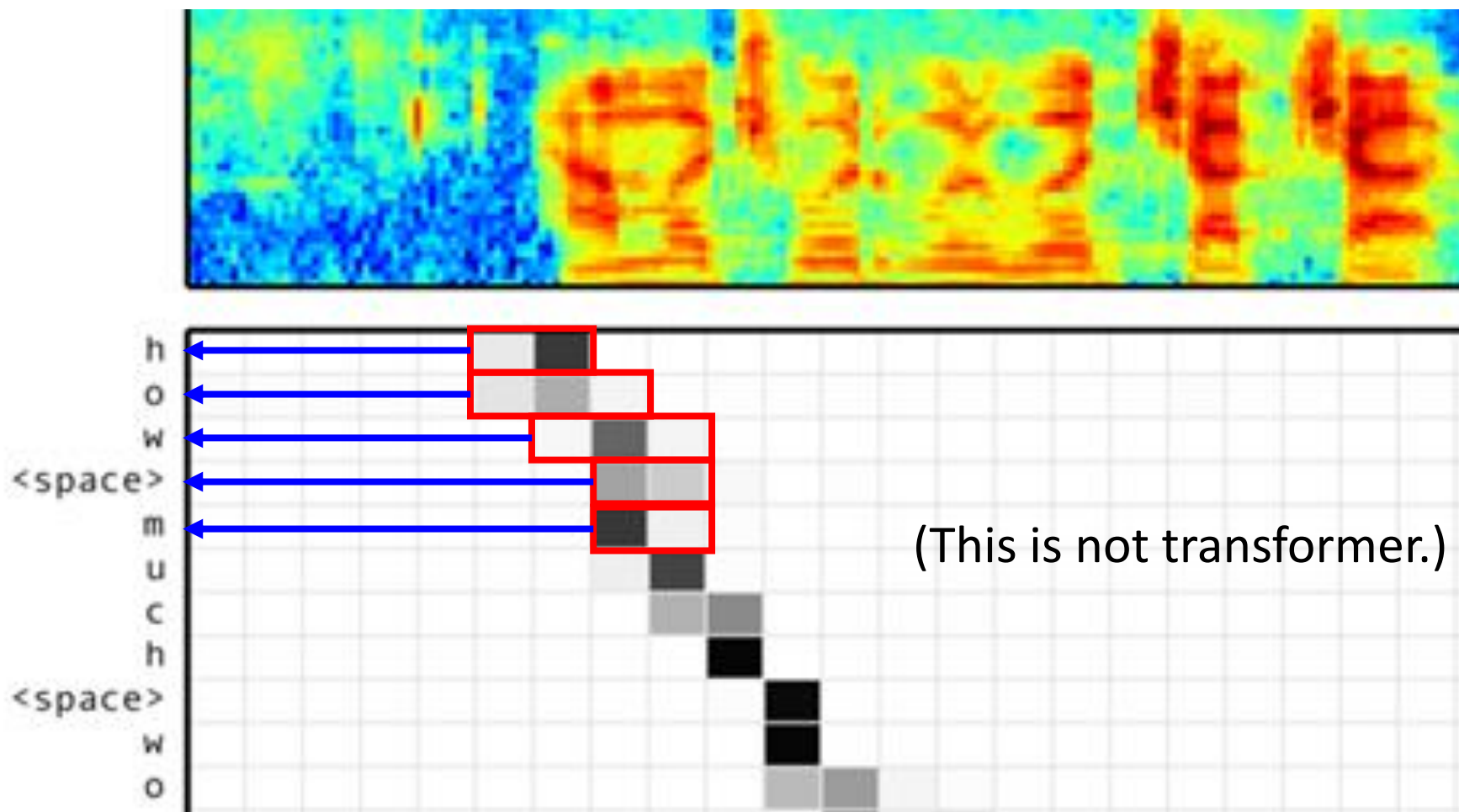




# Cross Attention

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

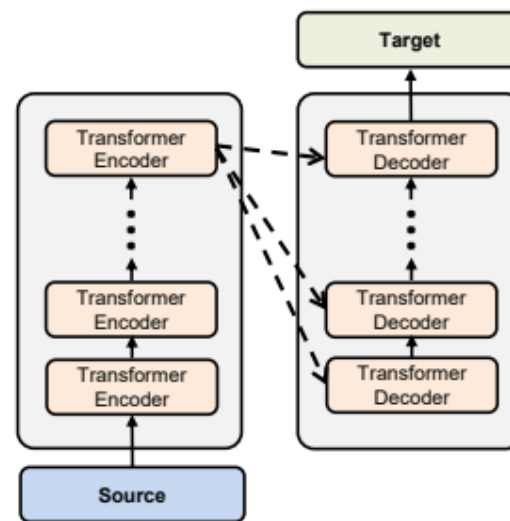
<https://ieeexplore.ieee.org/document/7472621>



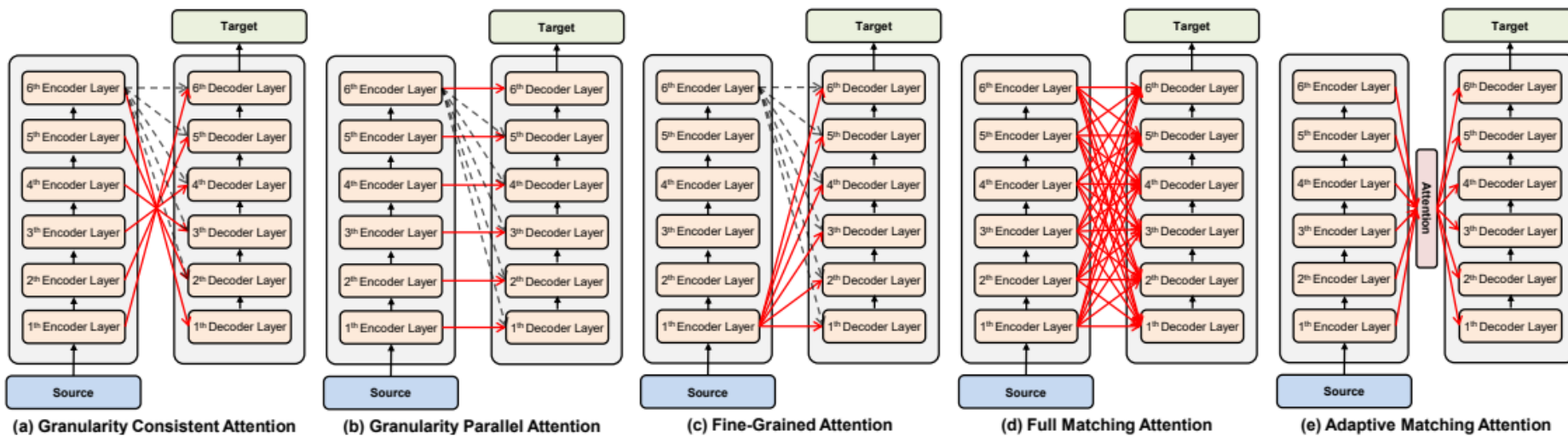
# Cross Attention

Source of image:

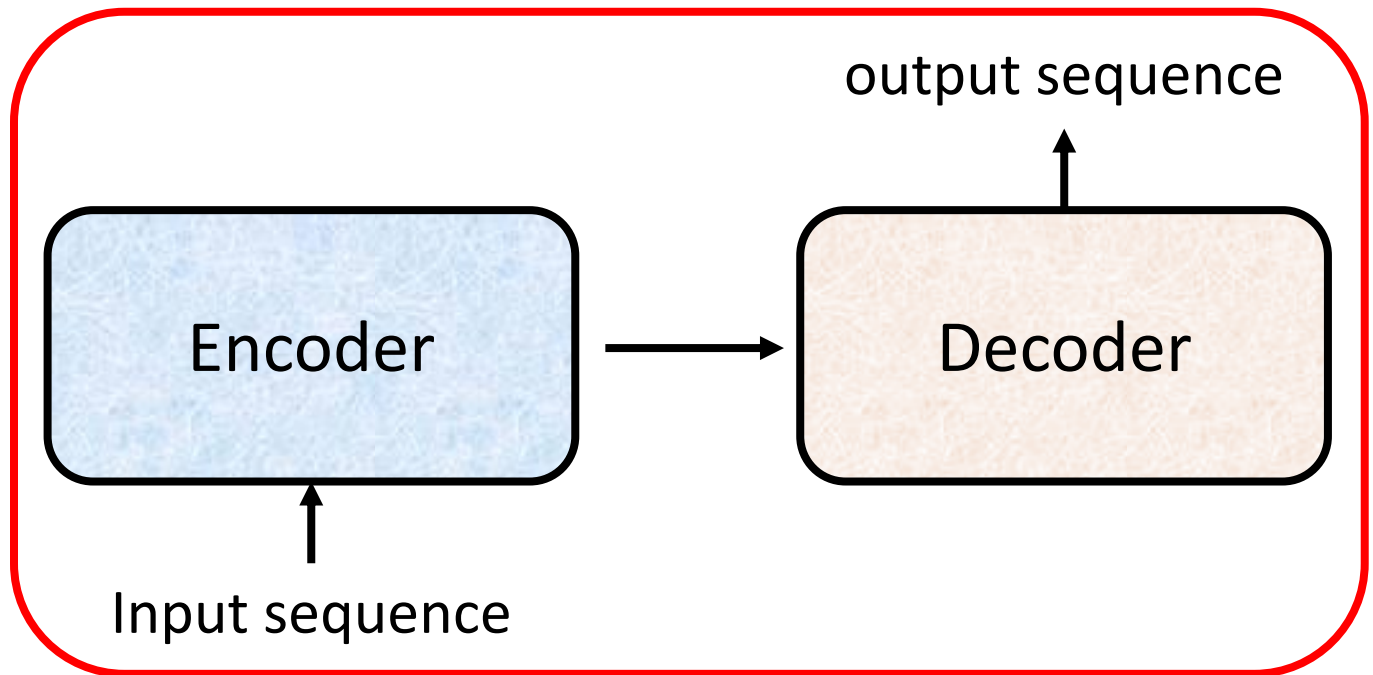
<https://arxiv.org/abs/2005.08081>



(a) Conventional Transformer



# 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

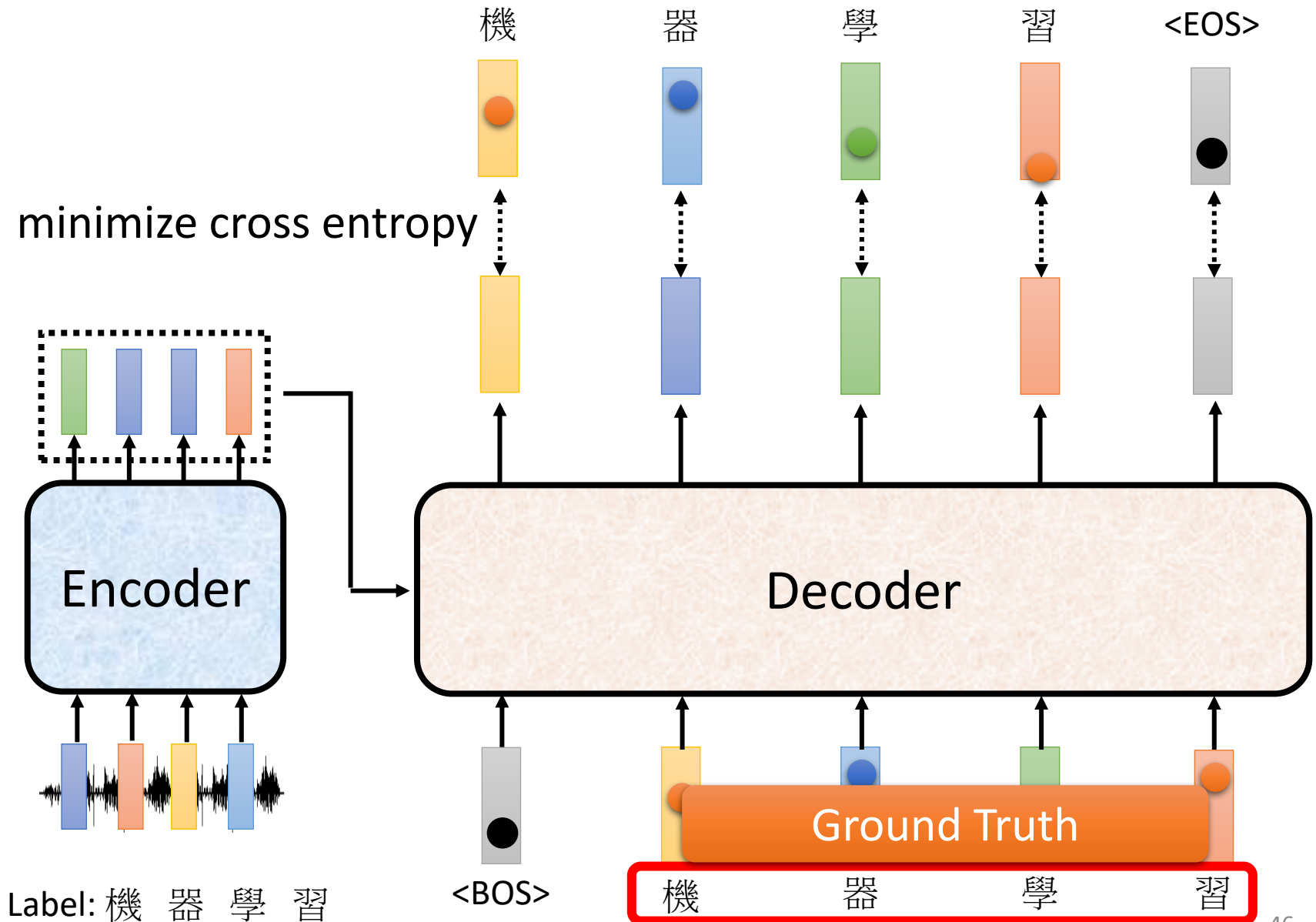
Decoder

Encoder

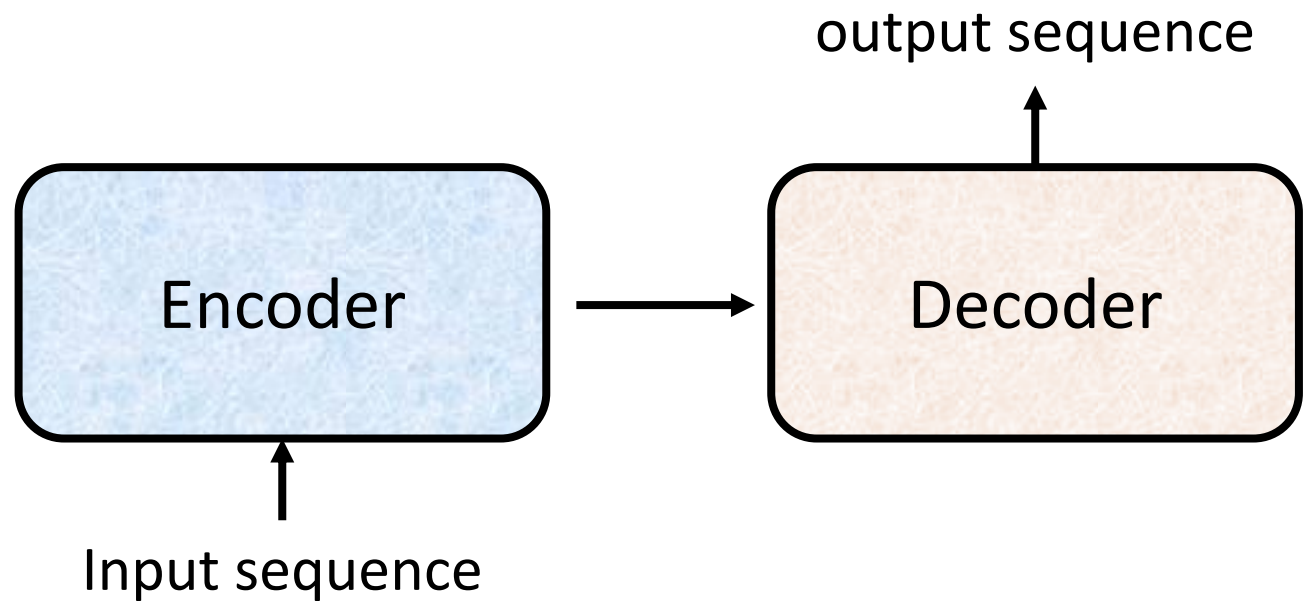
Label: 機 器 學 習

<BOS>

## Teacher Forcing: using the ground truth as input.

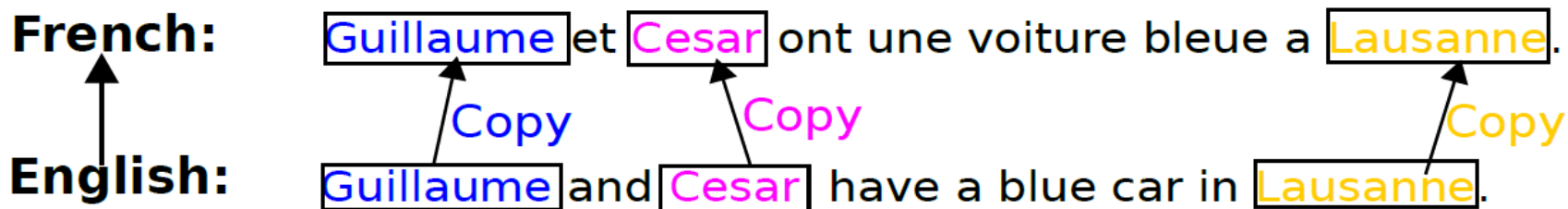


# Tips



# Copy Mechanism

## Machine Translation



## Chat-bot

User: X寶你好，我是庫洛洛

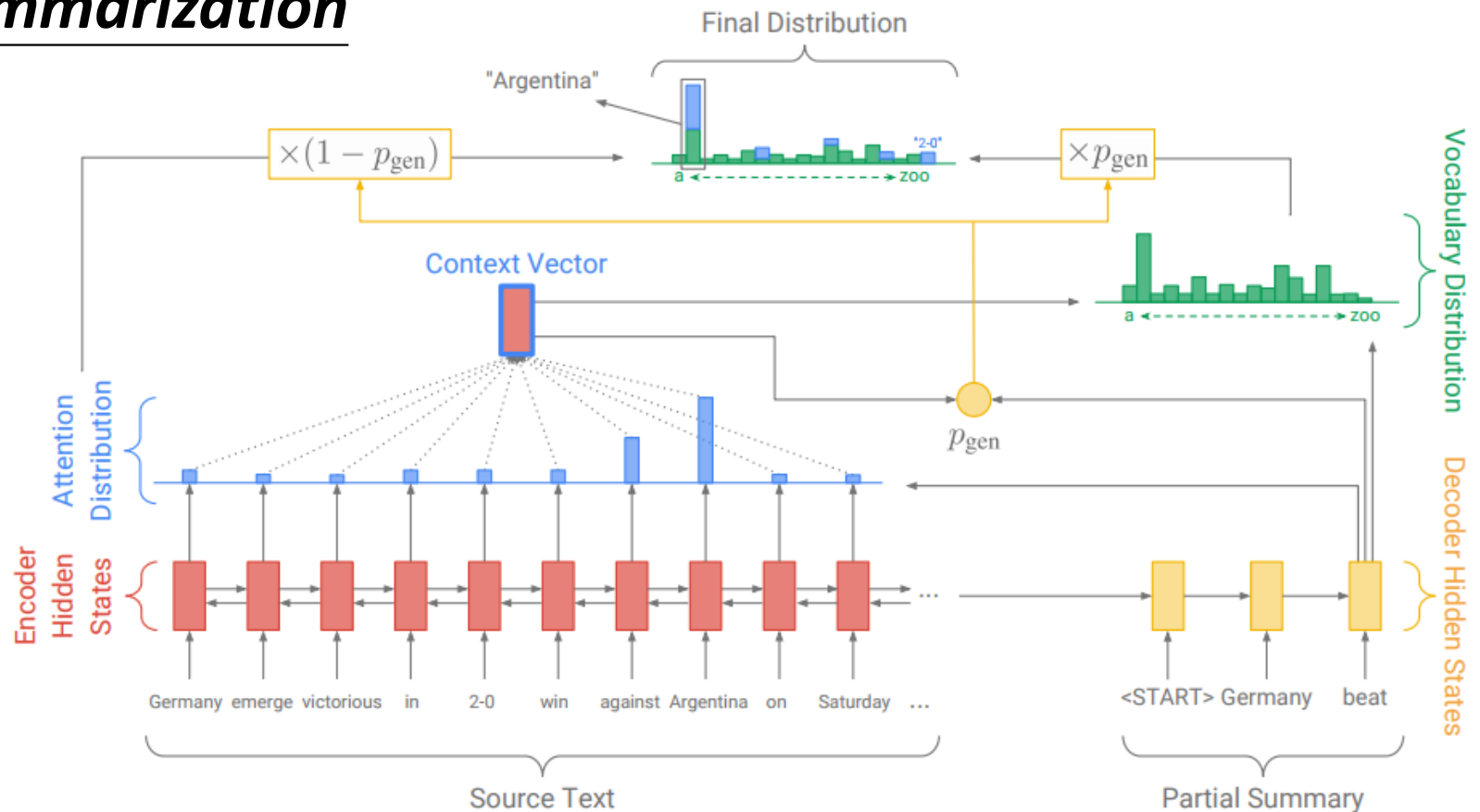
Machine: 庫洛洛你好，很高興認識你



# Copy Mechanism

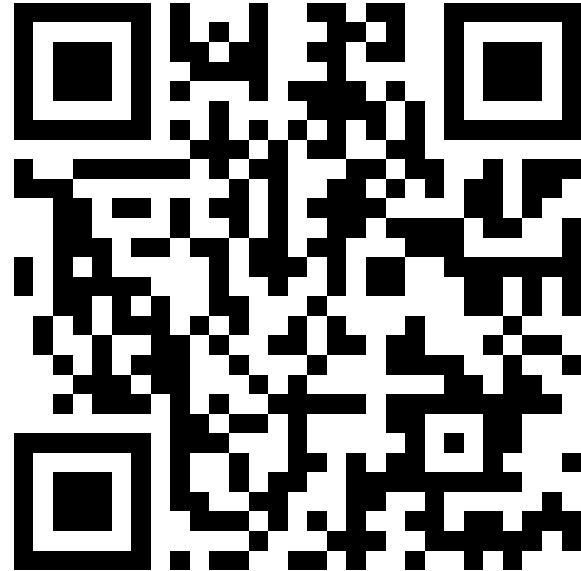
<https://arxiv.org/abs/1704.04368>

## 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

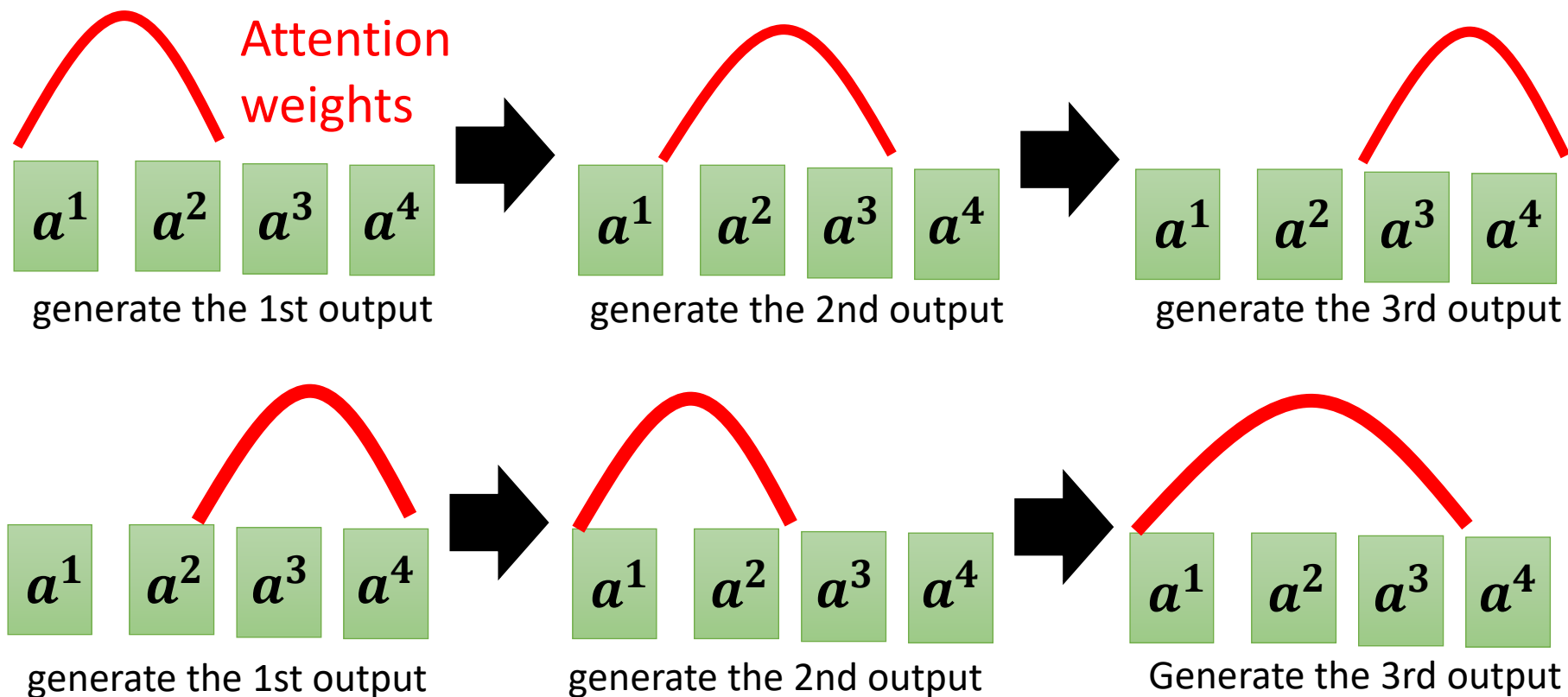


發財 (Missing an input character!)

# Guided Attention

Monotonic Attention  
Location-aware attention

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



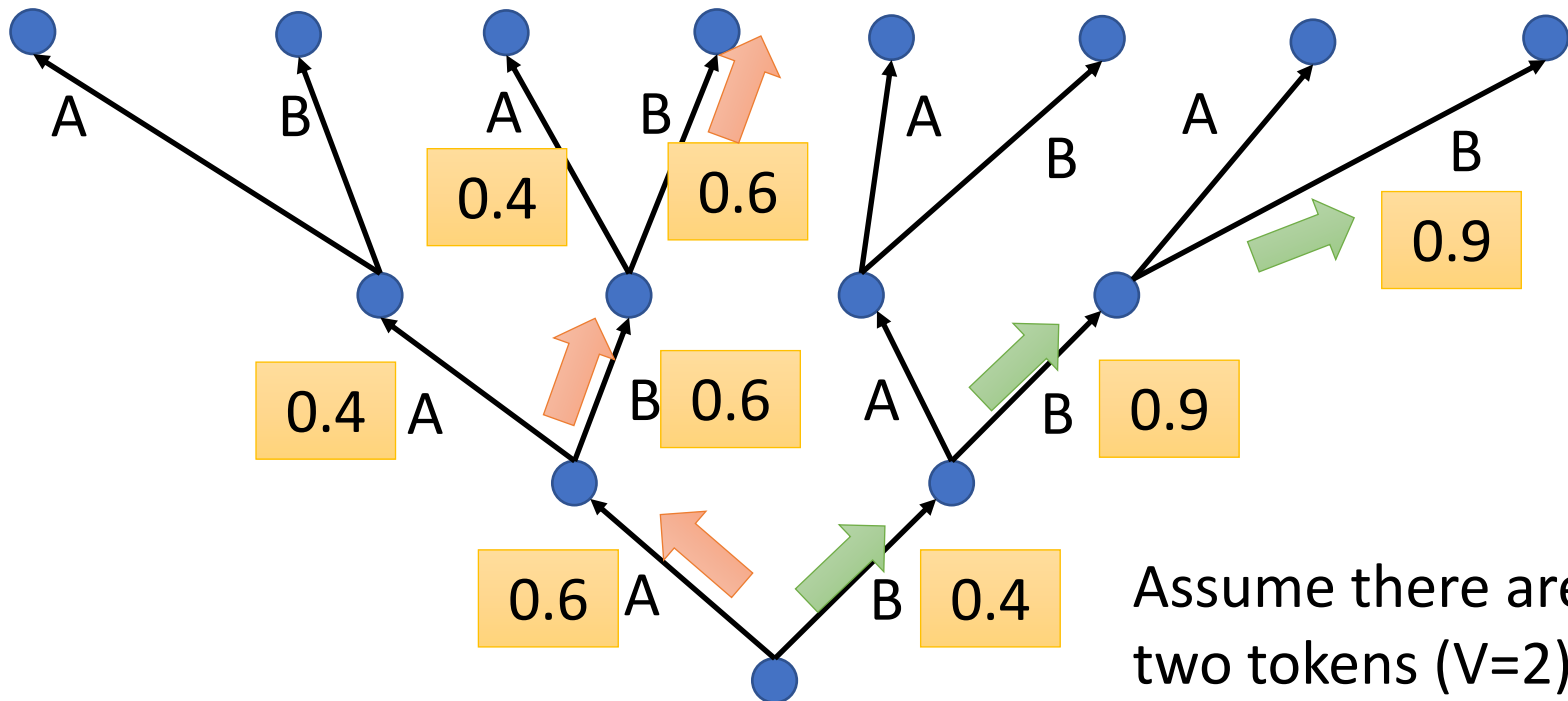
**Something wrong!**

# Beam Search

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

The **green** path is the best one.

Not possible to check all the paths ... → Beam Search



# 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 México/Universidad Nacional Autónoma de ..."

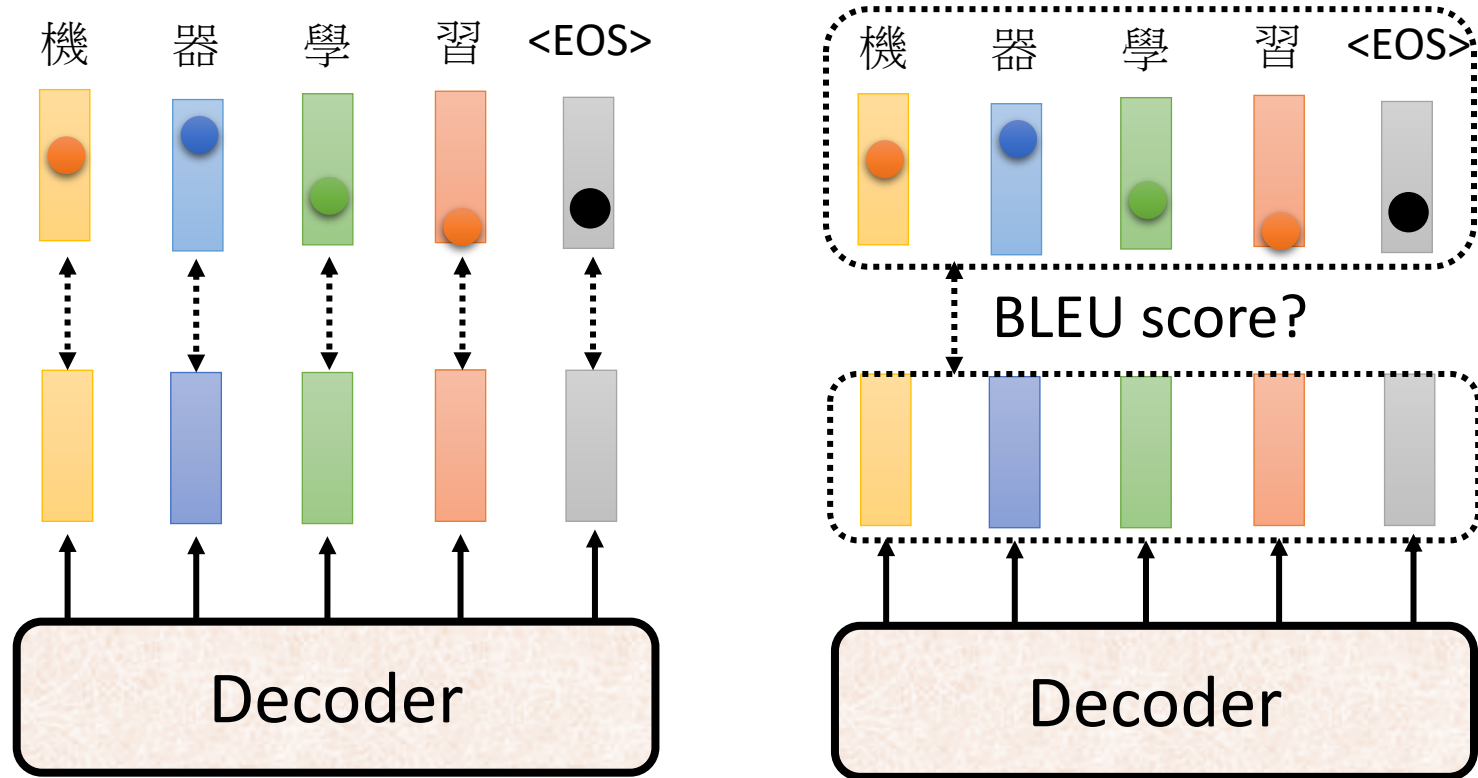
### Pure Sampling:

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. 😊

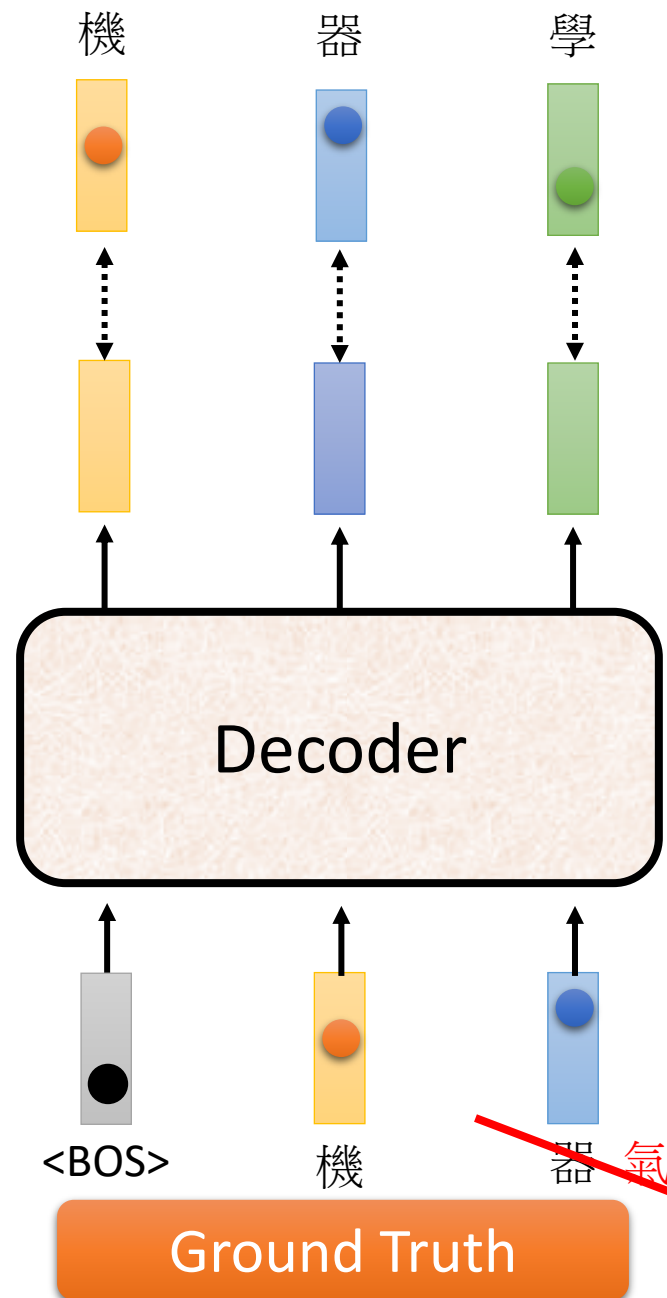
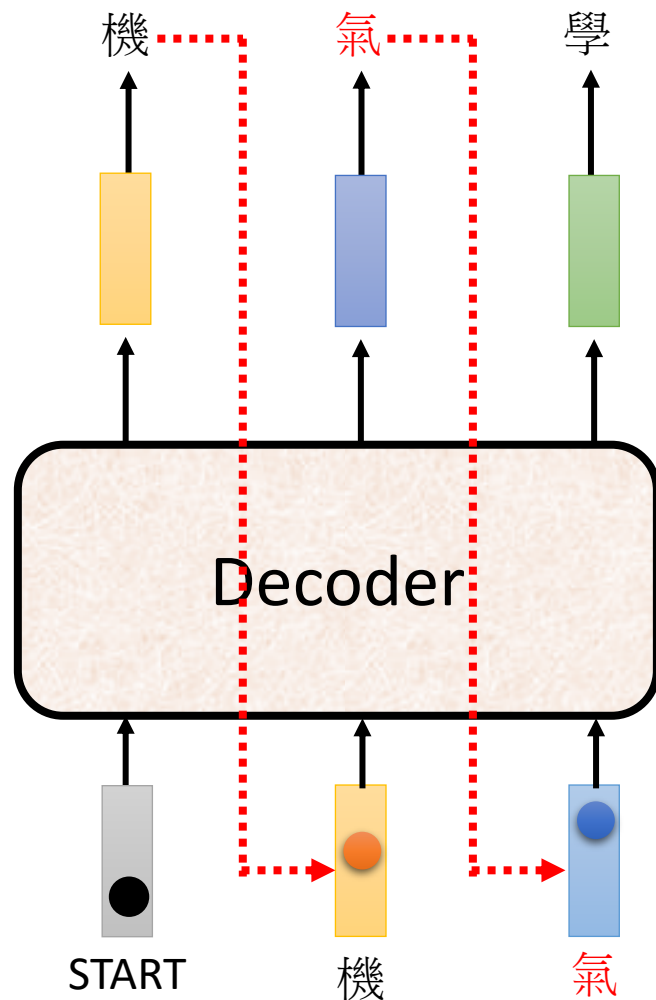
# Optimizing Evaluation Metrics?



How to do the optimization?

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

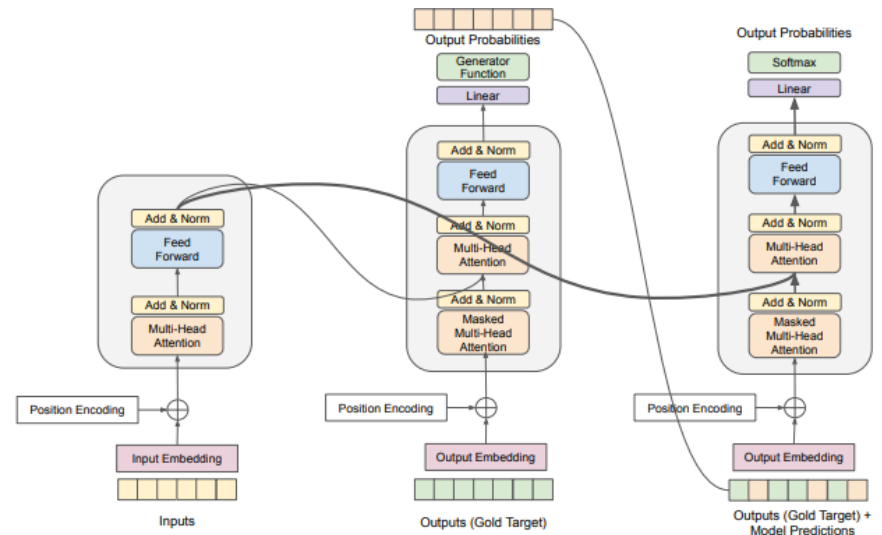
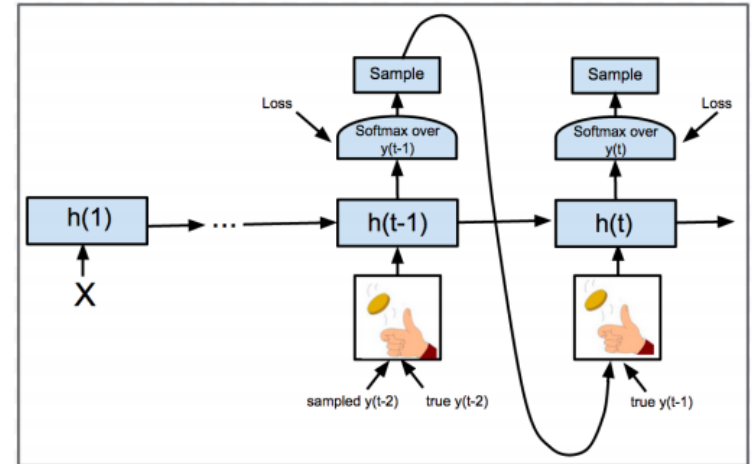
There is a mismatch! 😞  
**exposure bias**





# 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

