



人工智能技术及应用

Artificial Intelligence and Application

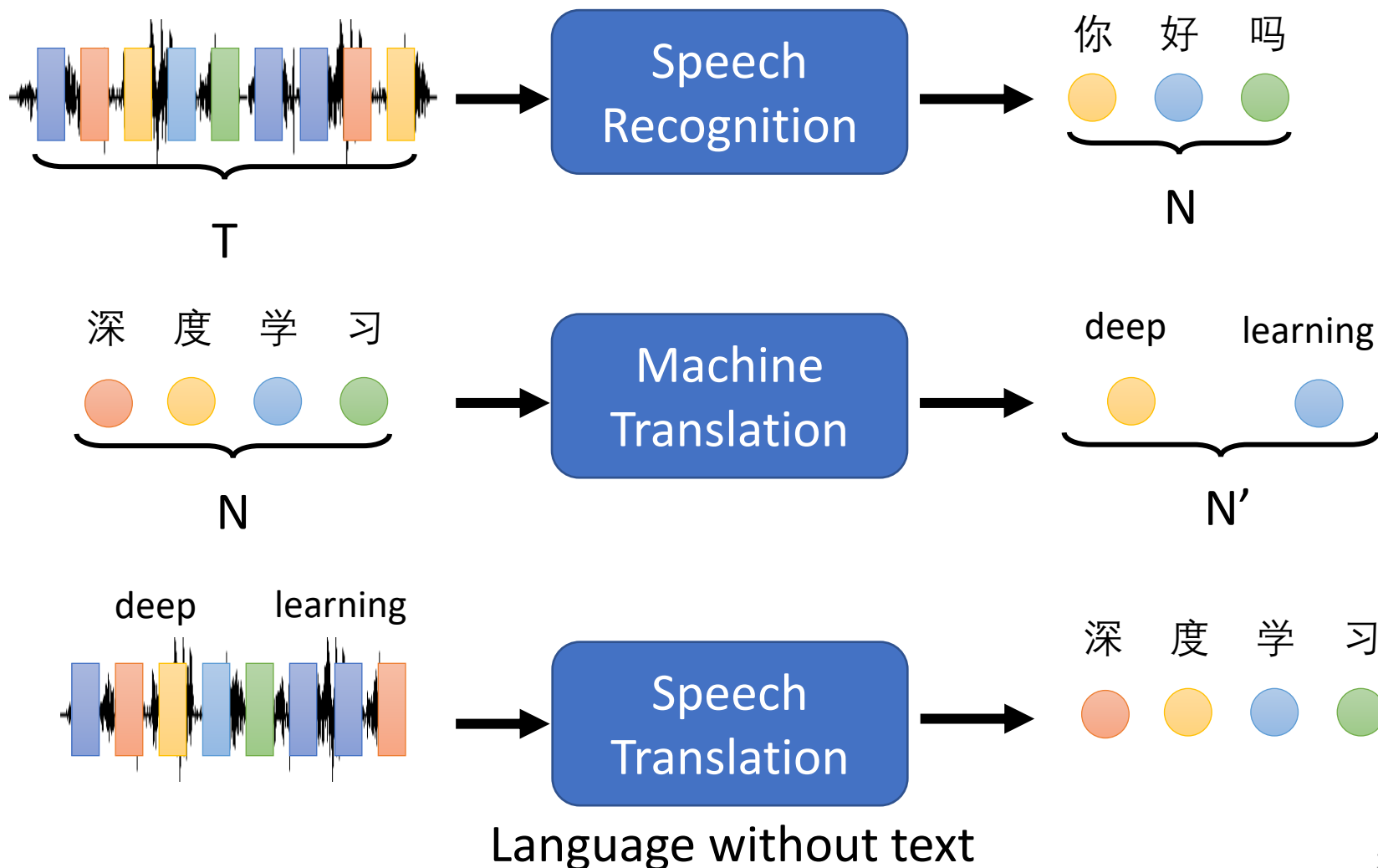
Transformer



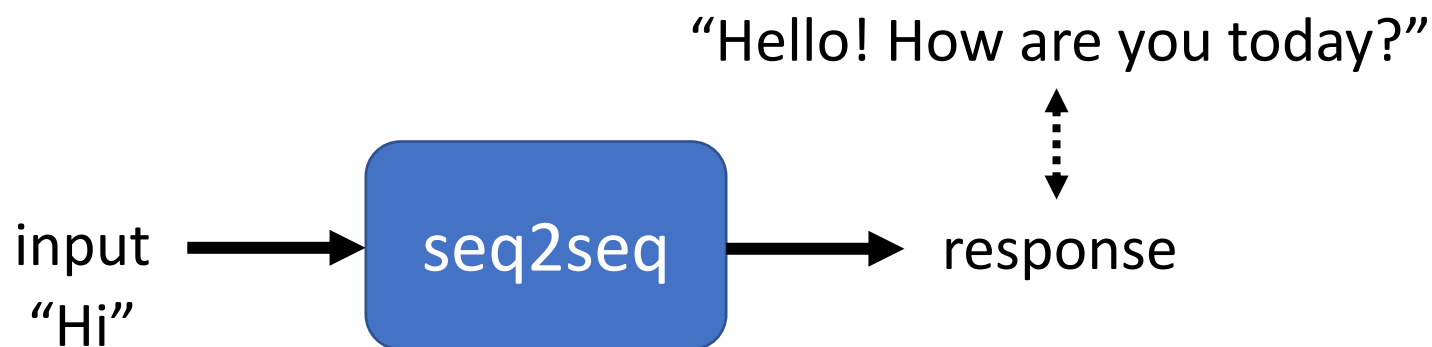
Sequence-to-sequence (Seq2seq)

Input a sequence, output a sequence

The output length is determined by model.



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.

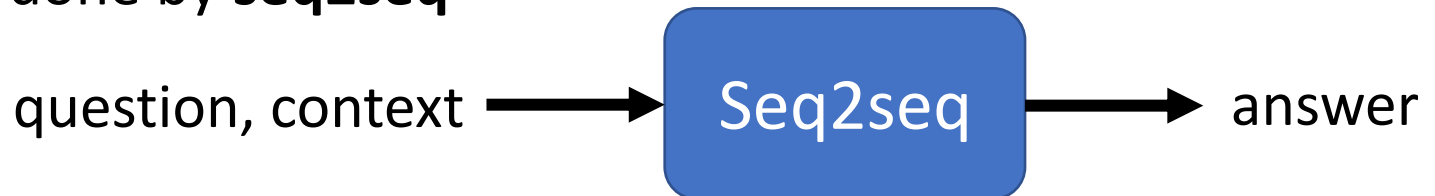
Most Natural Language Processing applications ...

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



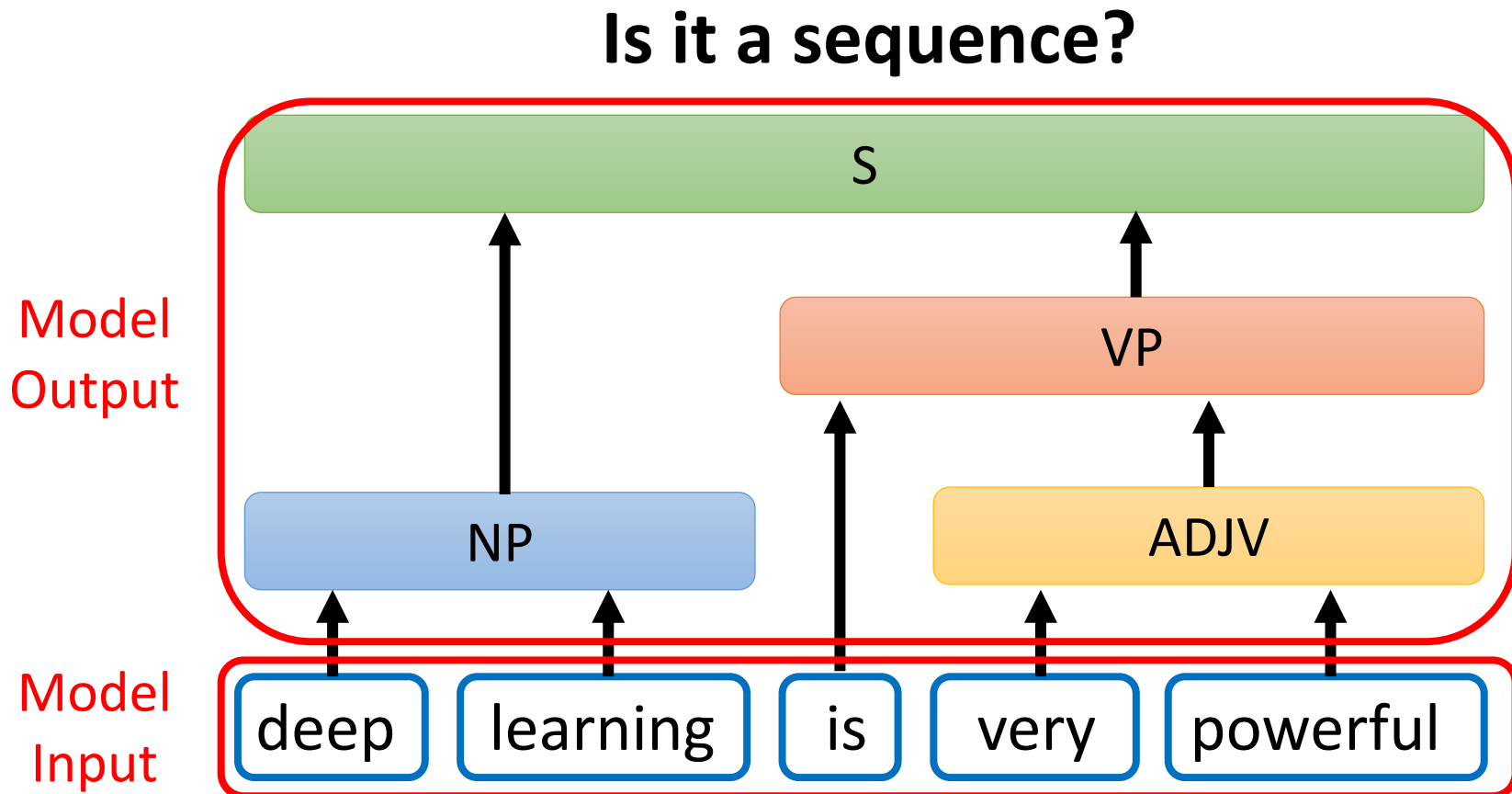
QA can be done by seq2seq



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

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

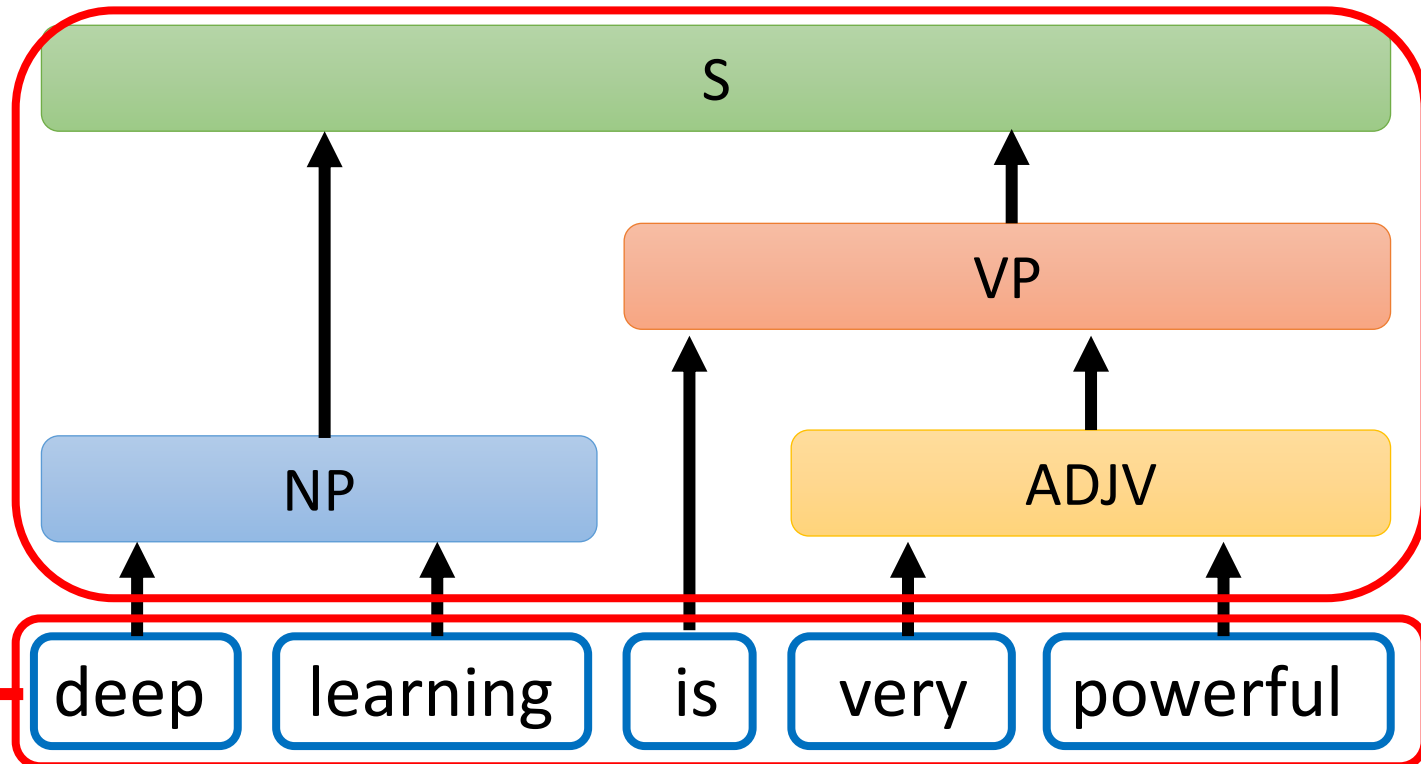
Seq2seq for Syntactic Parsing



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

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



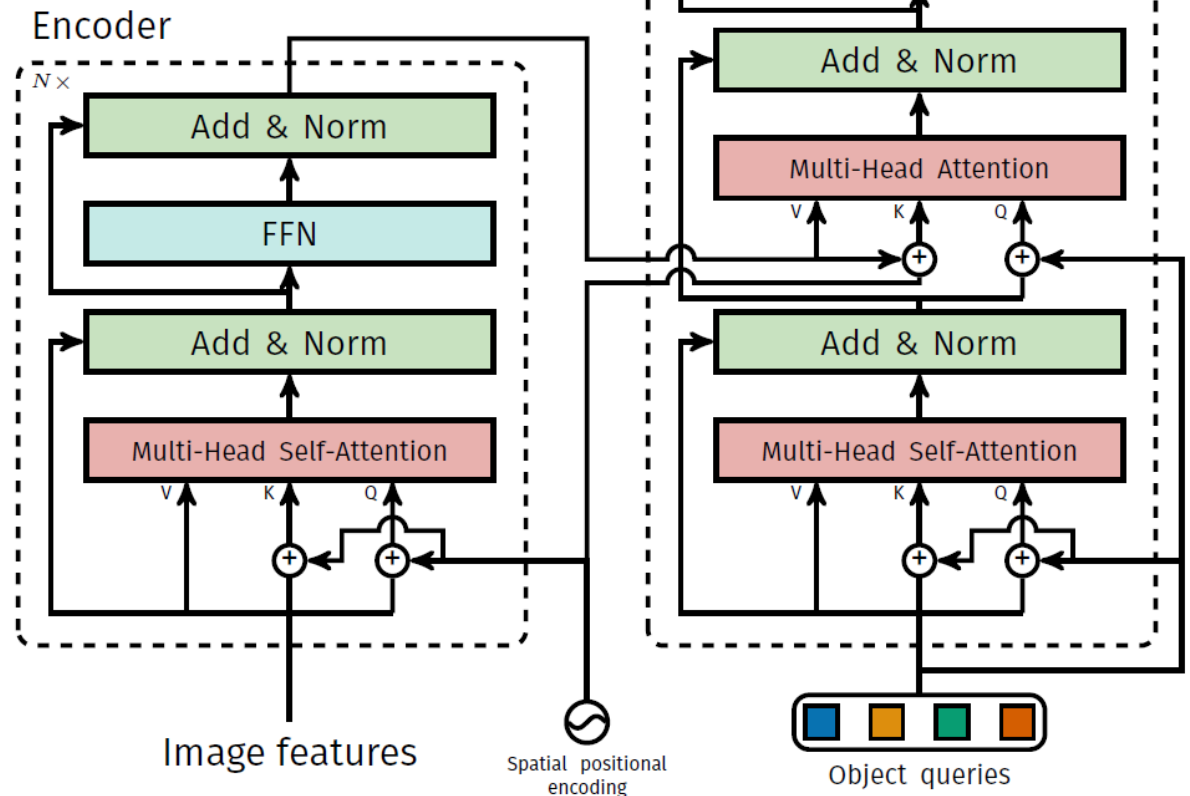
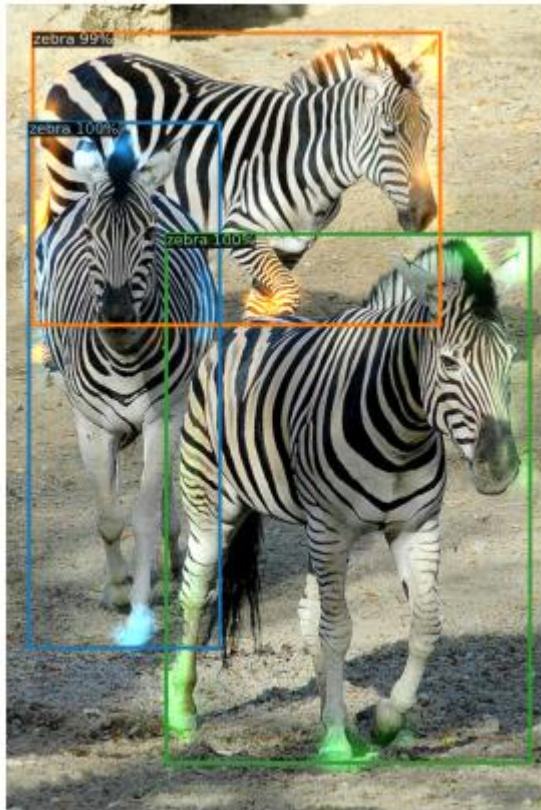
Class 13

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

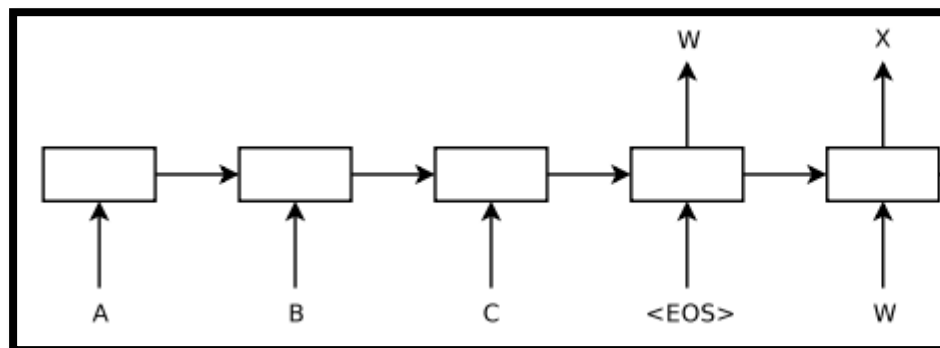
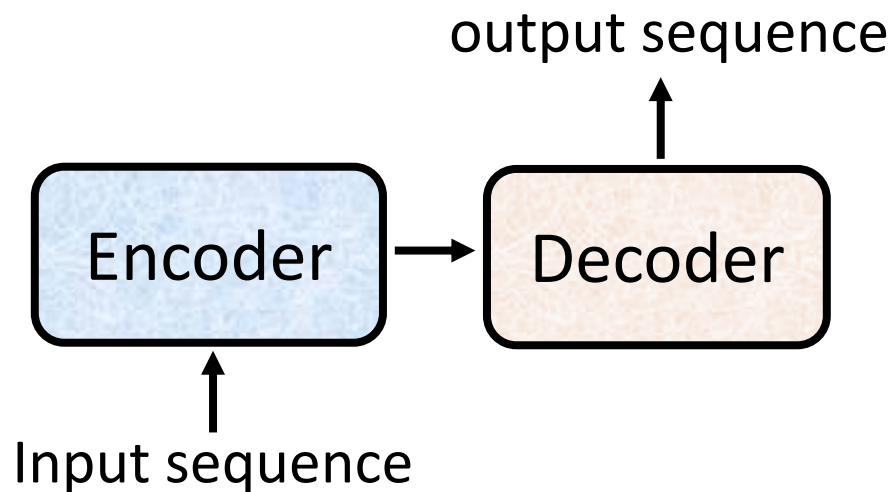
<https://arxiv.org/abs/1707.05495>

Seq2seq for Object Detection

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

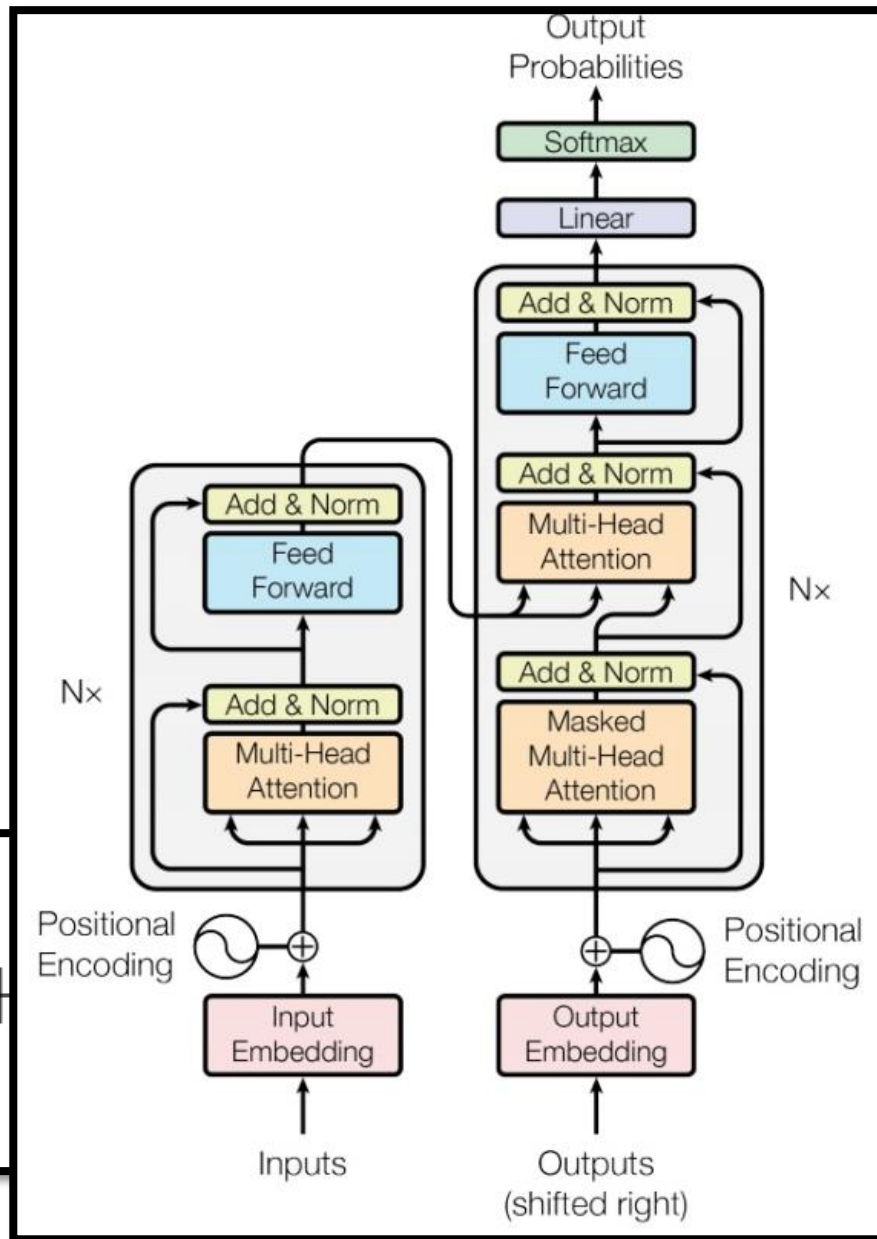


Seq2seq



Sequence to Sequence Learning with Neural Networks

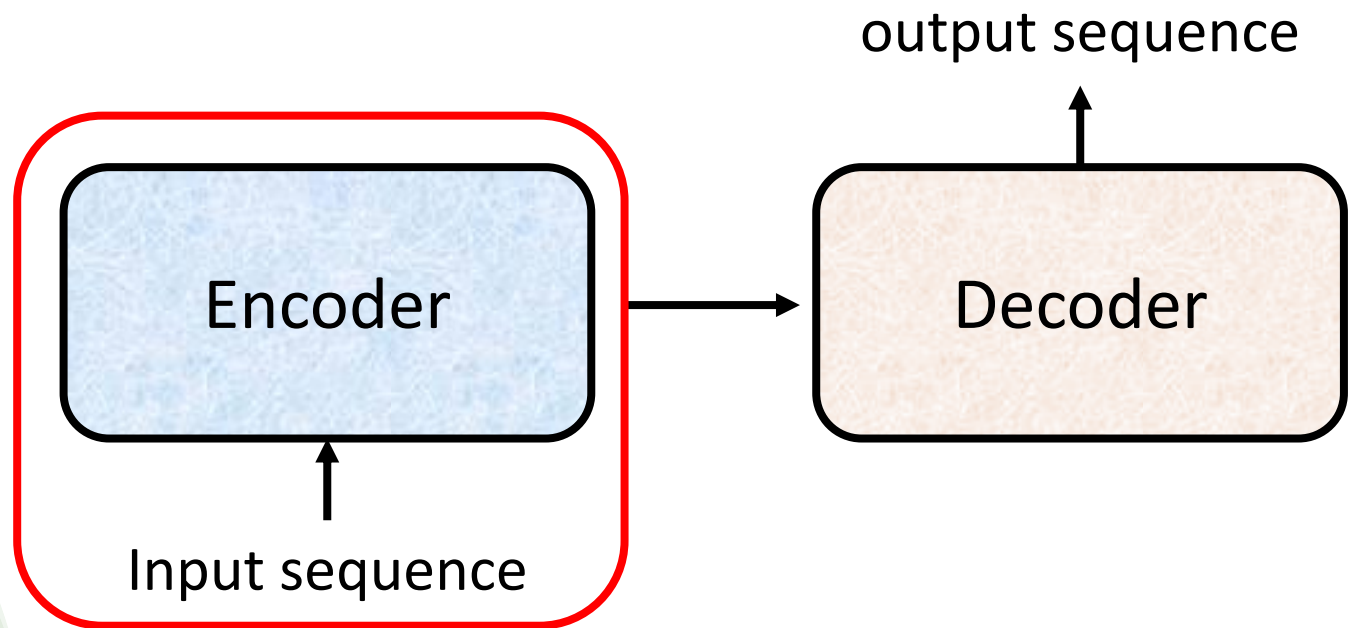
<https://arxiv.org/abs/1409.3215>



Transformer

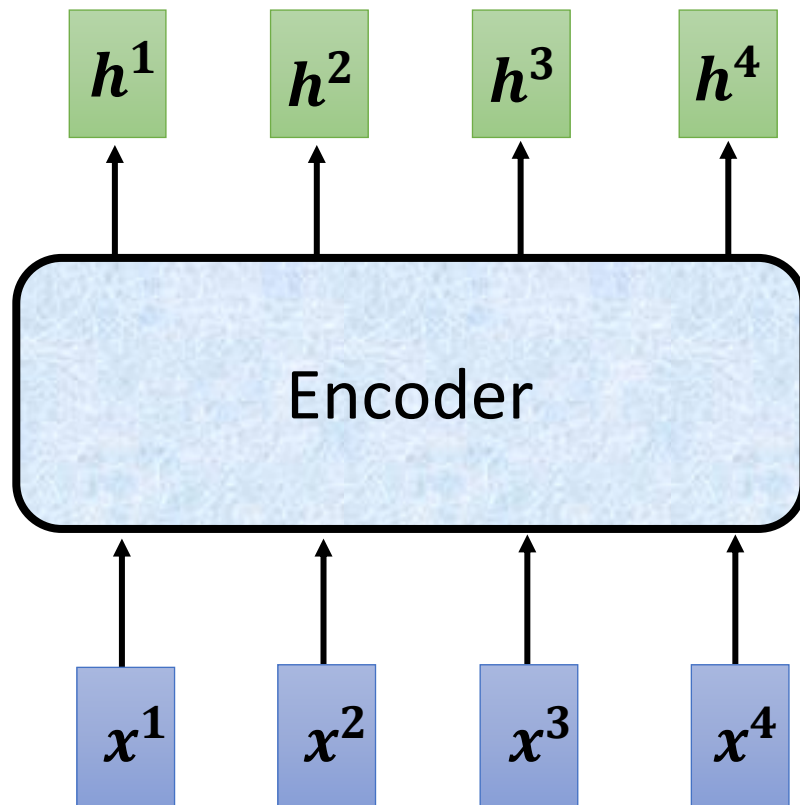
<https://arxiv.org/abs/1706.03762>

Encoder

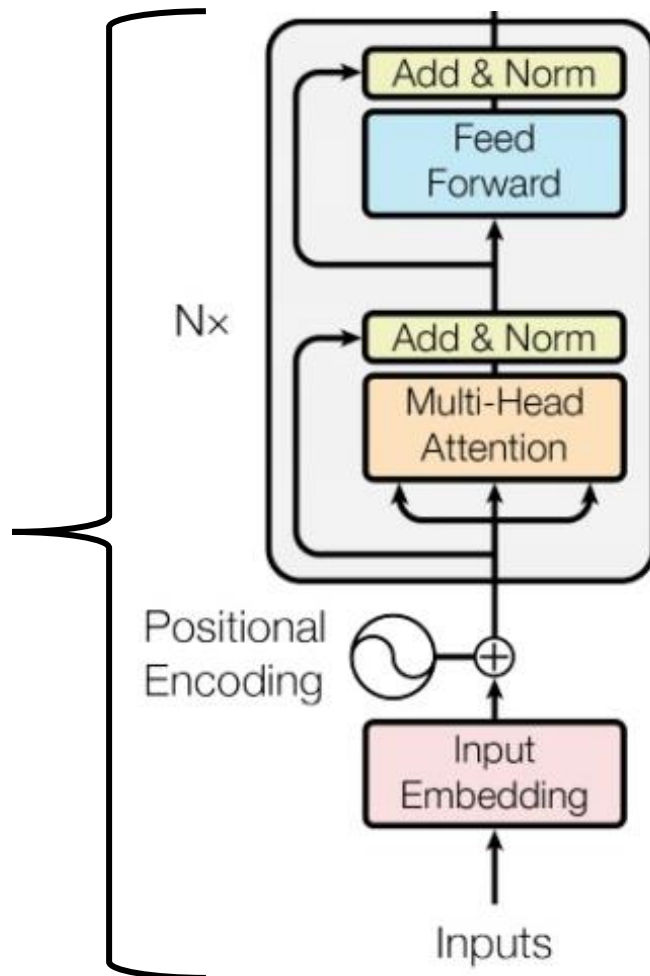


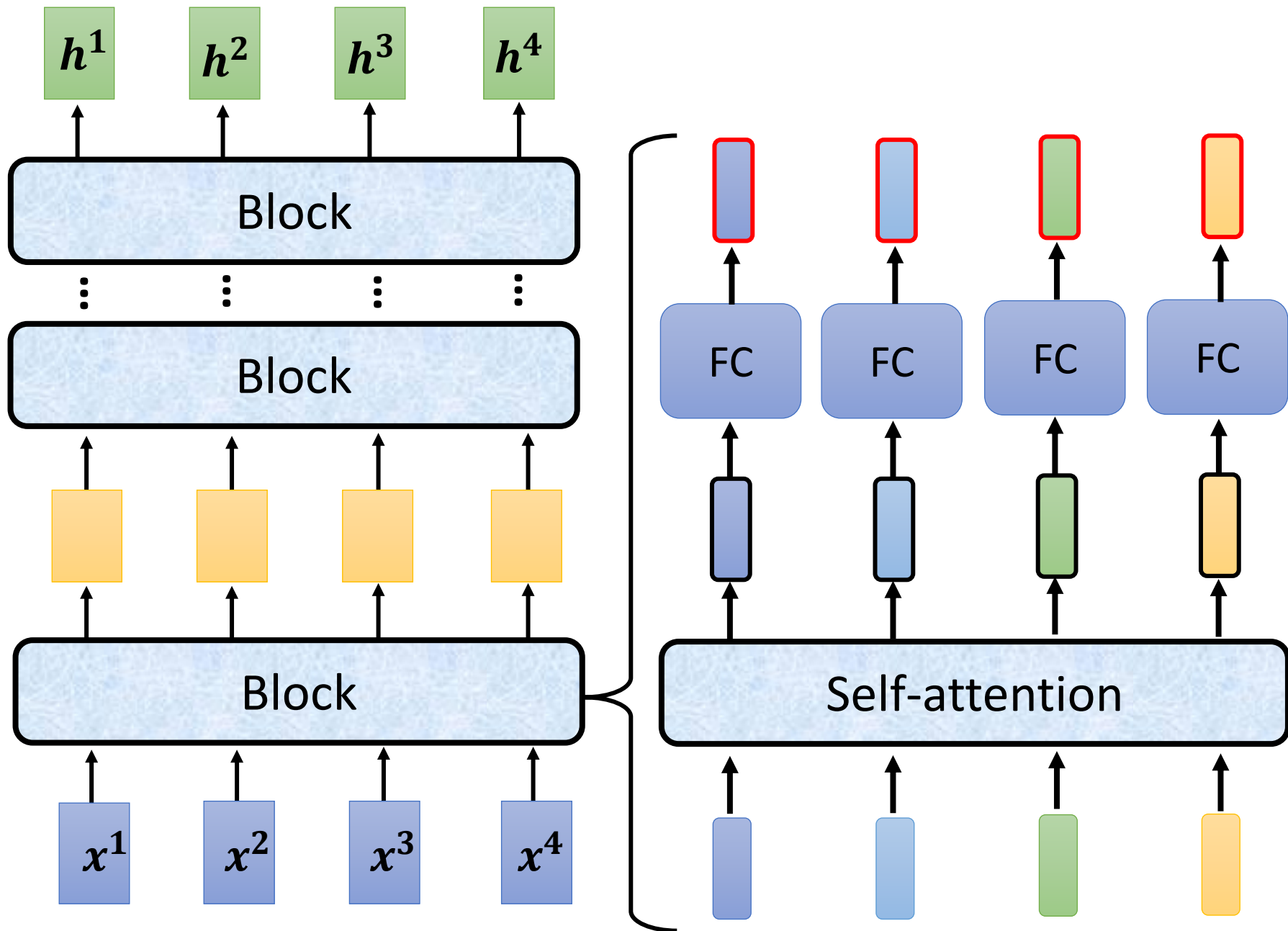
Encoder

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



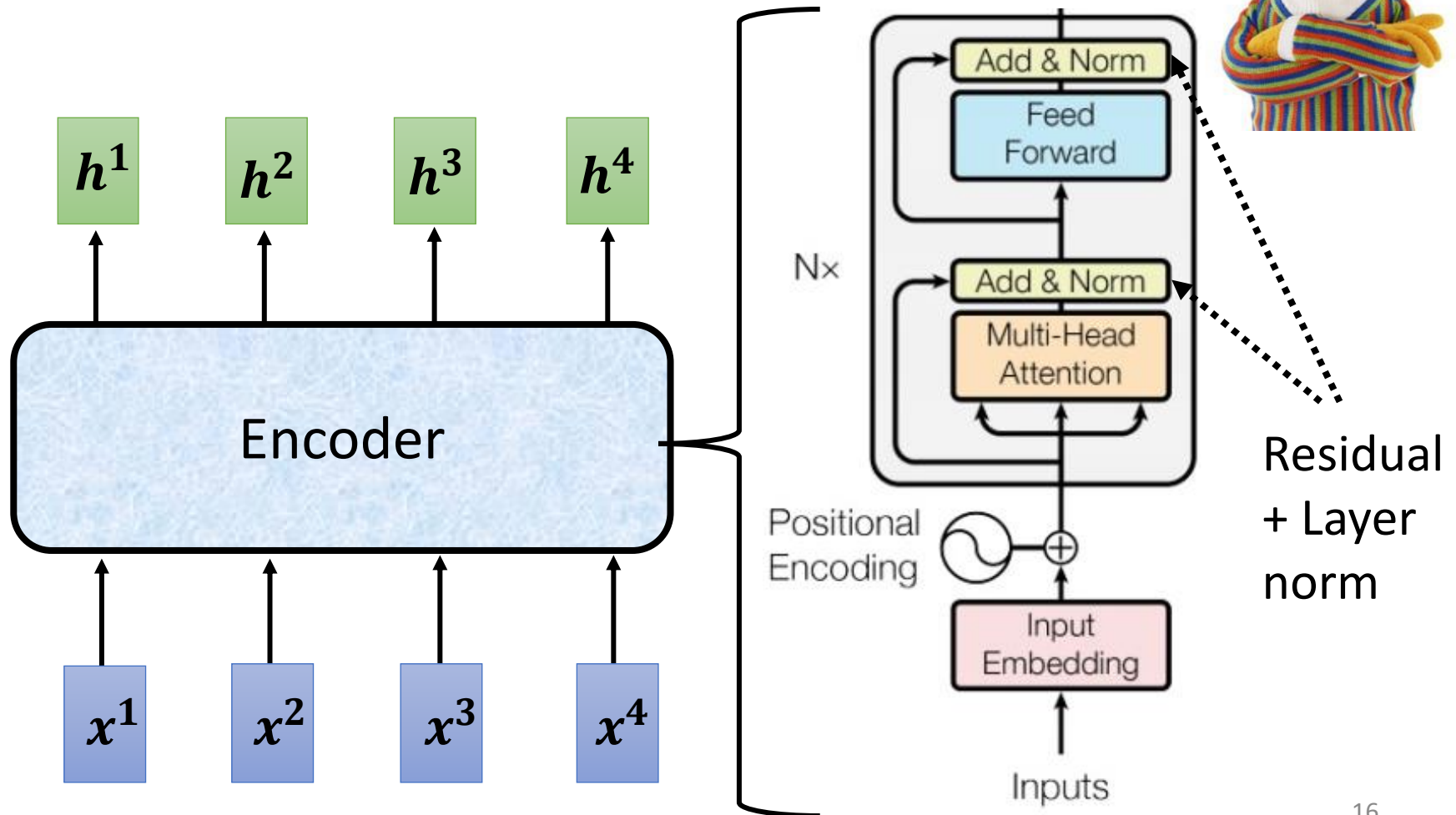
Transformer's Encoder





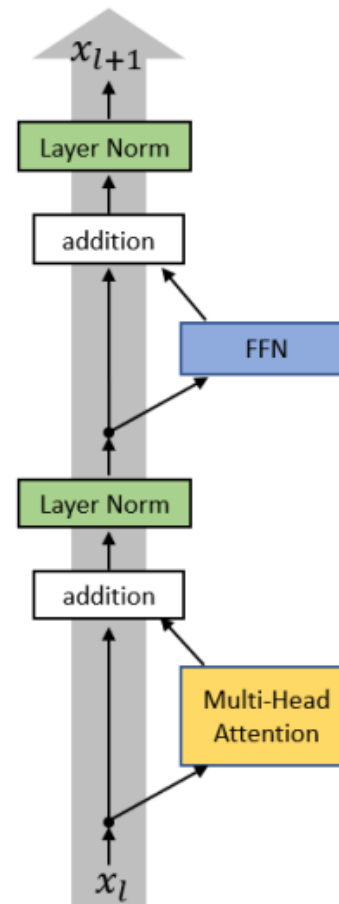


I use the **same** network architecture as **transformer encoder**.

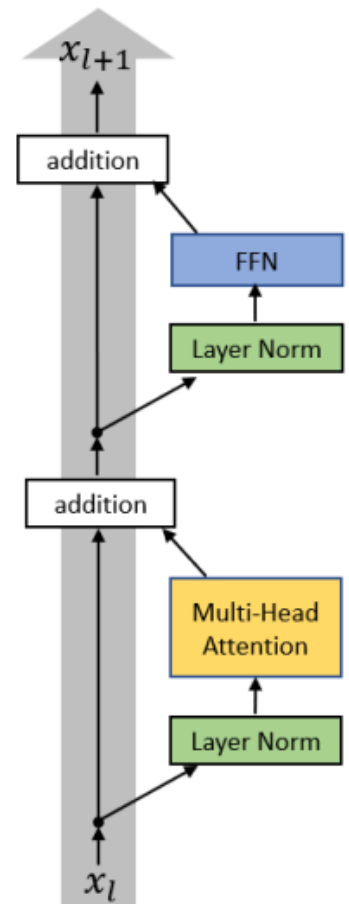


To learn more

- 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>

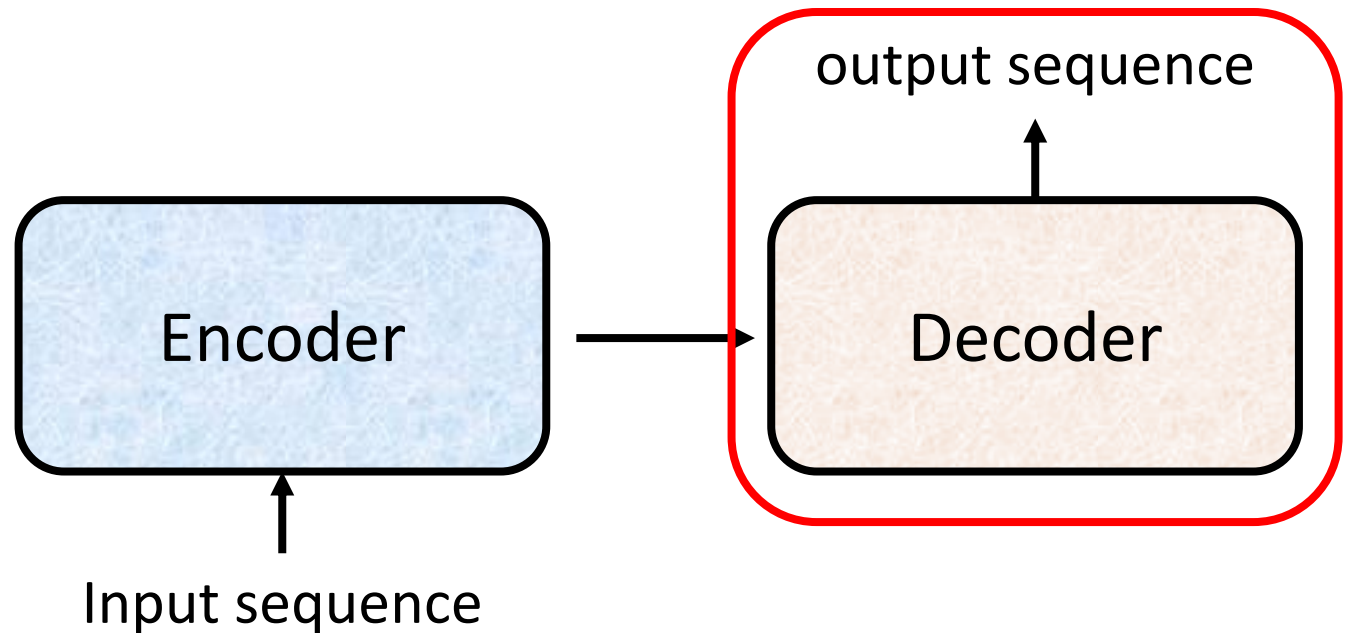


(a)



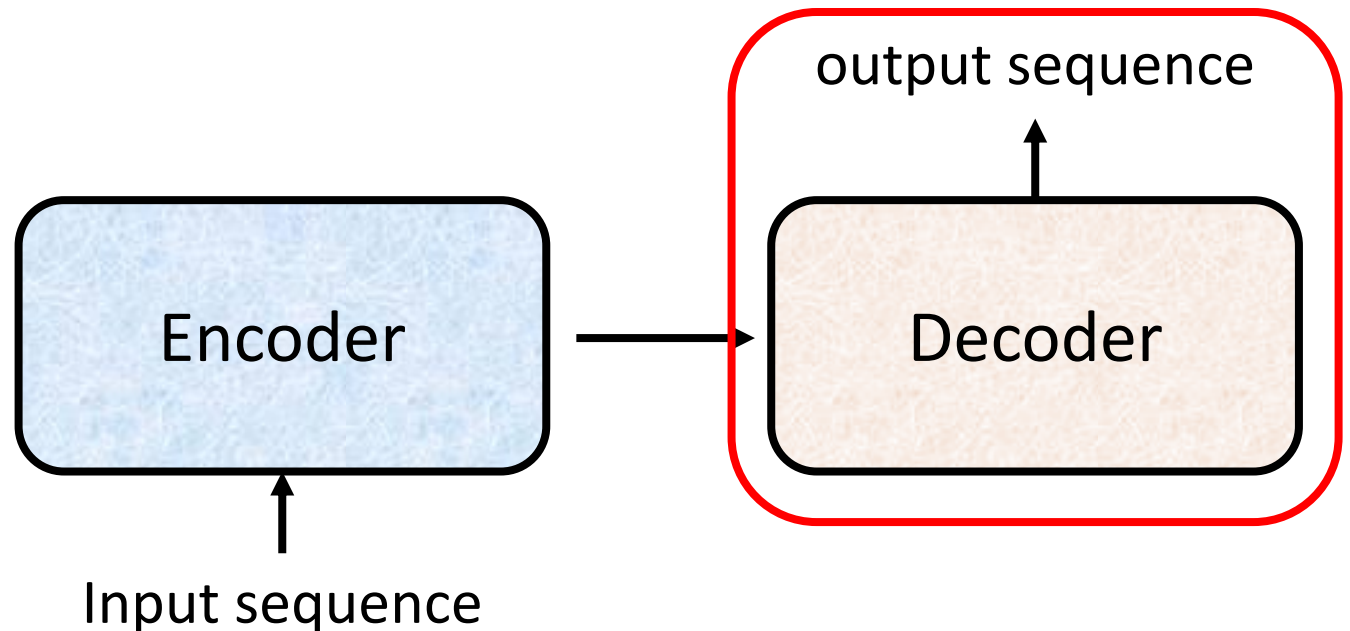
(b)

Decoder



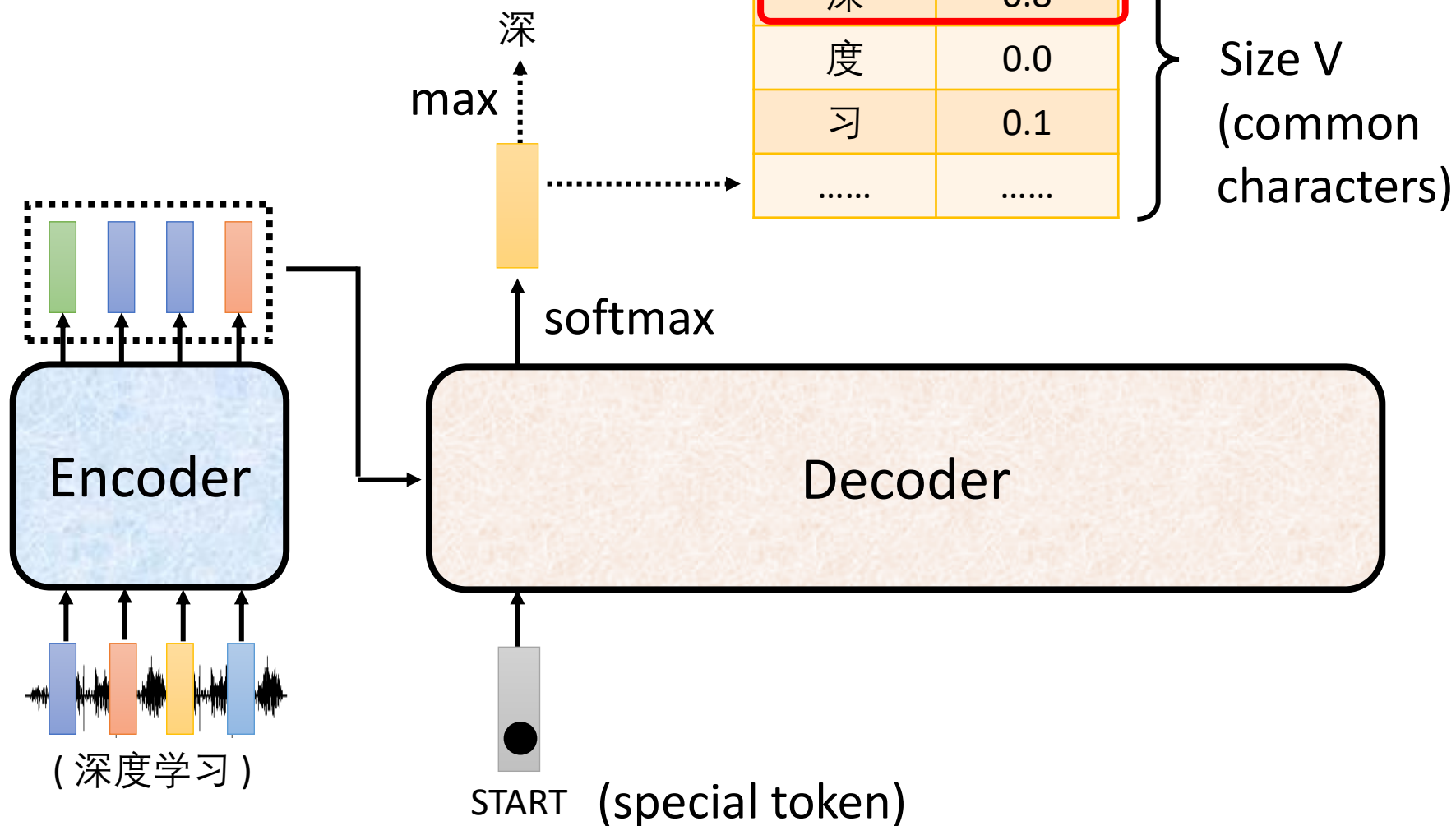
Decoder

- Autoregressive (AT)

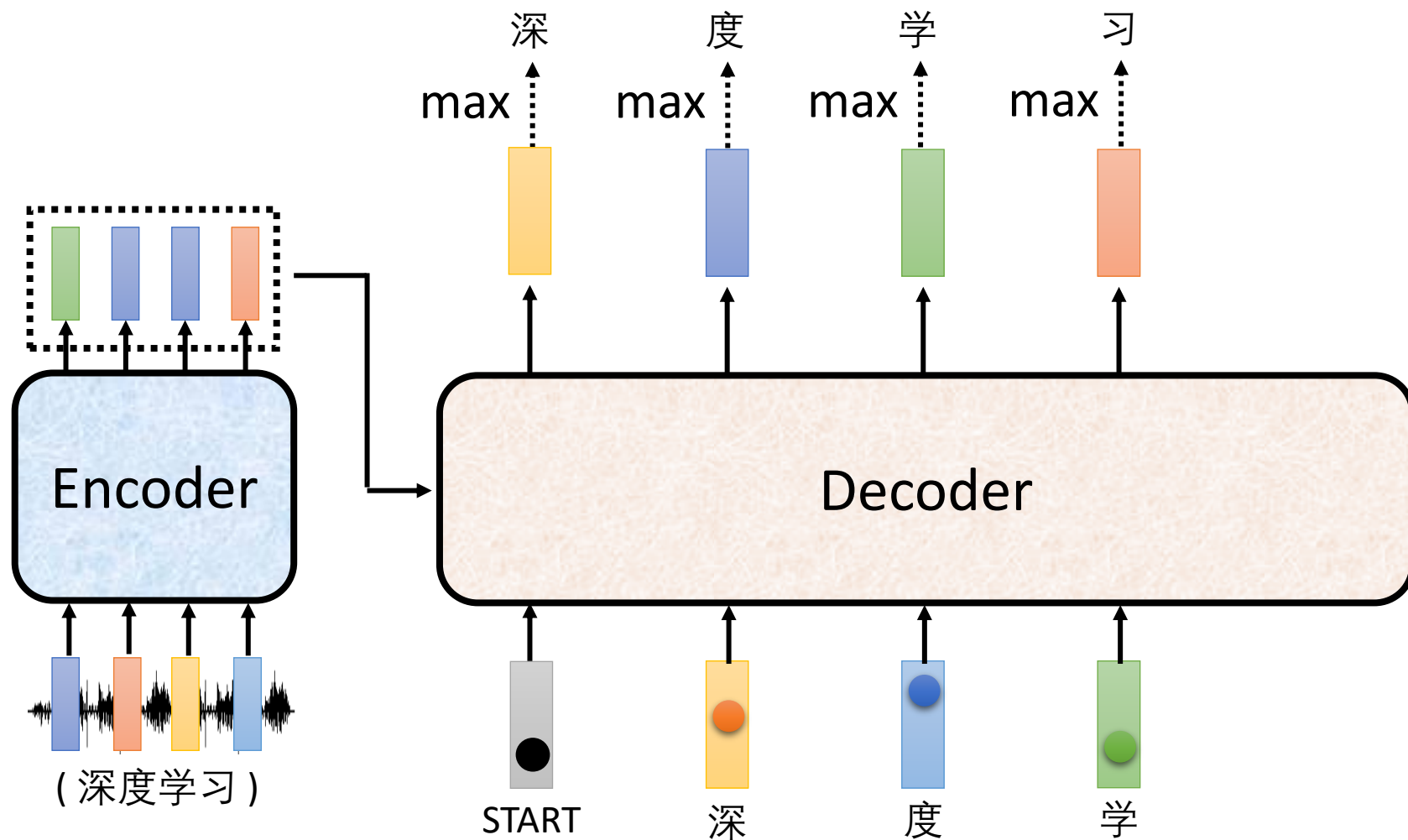


Autoregressive

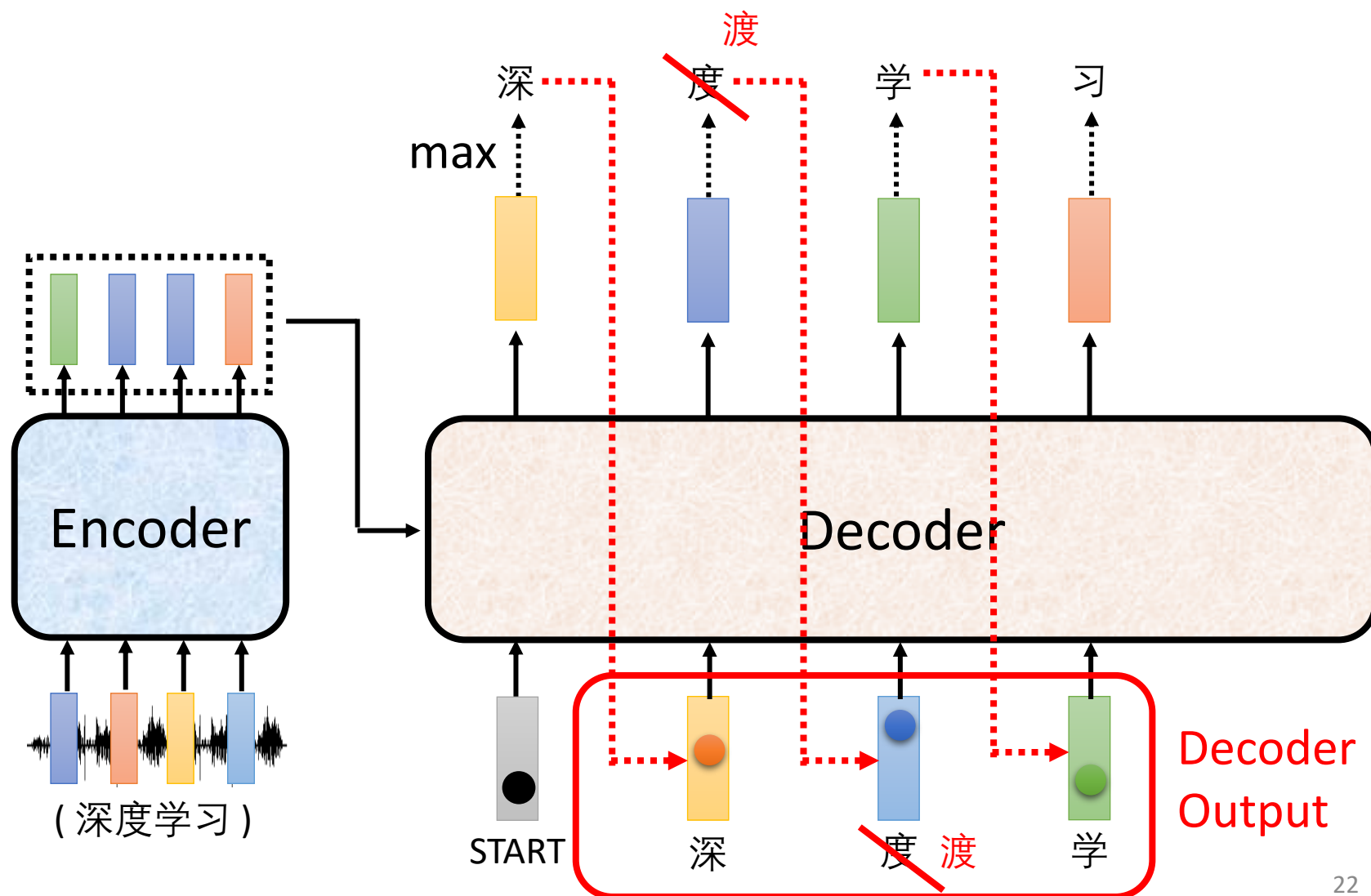
(Speech Recognition as example)



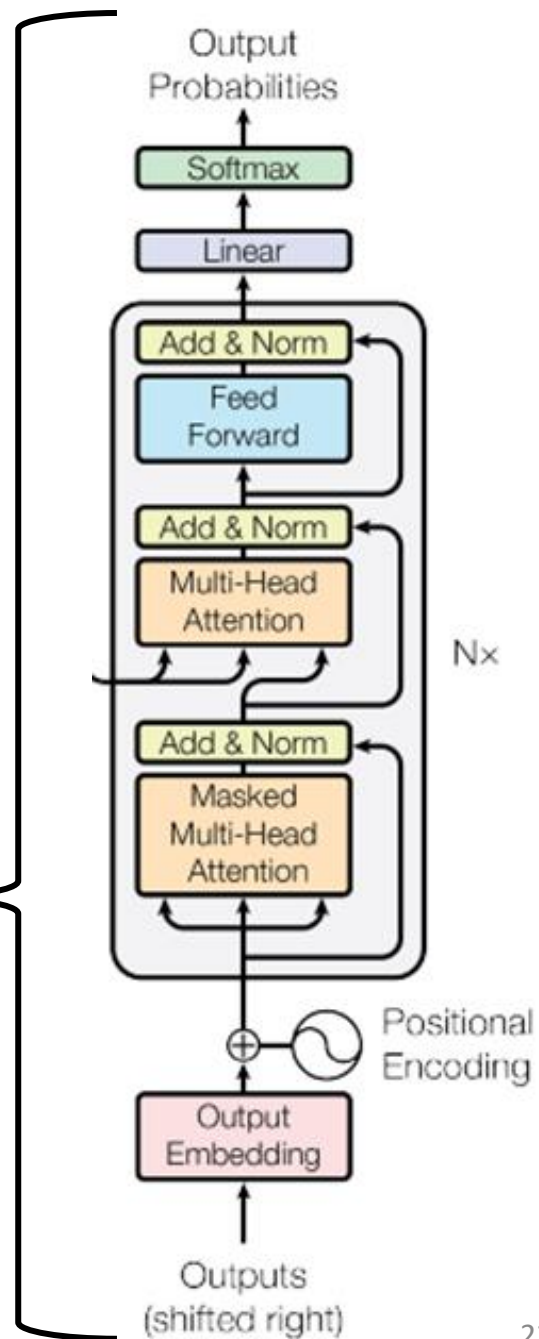
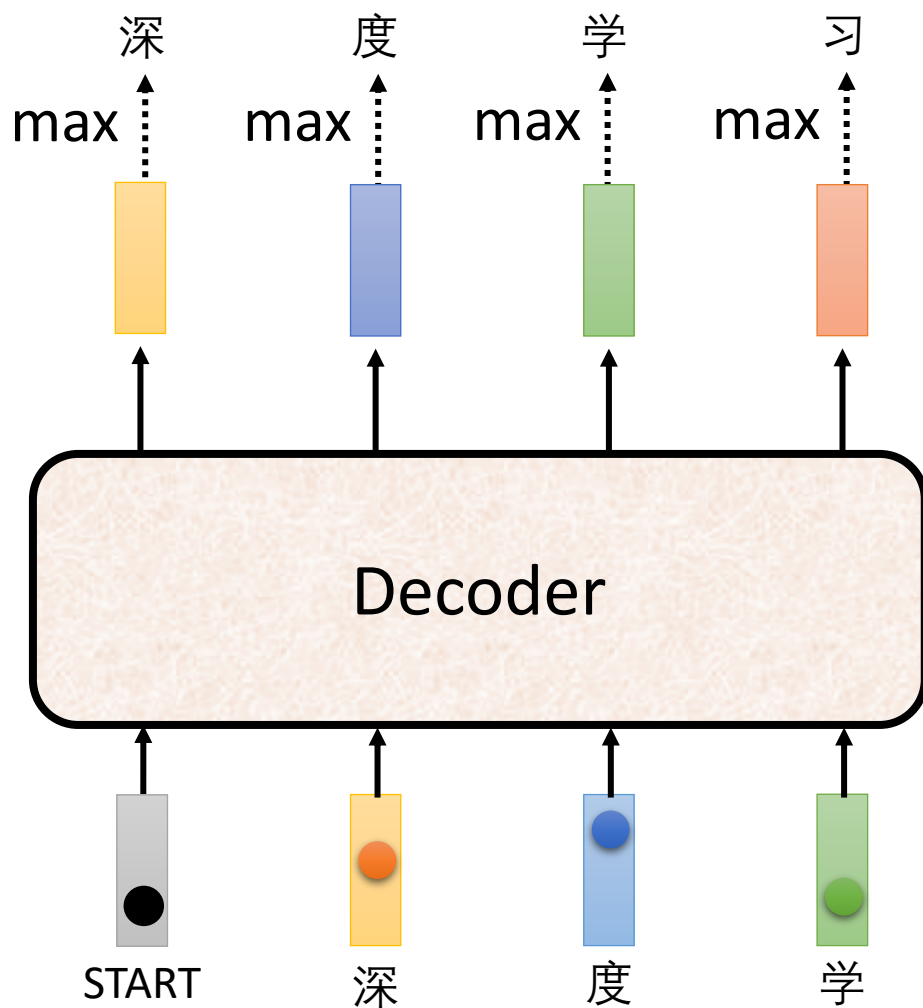
Autoregressive



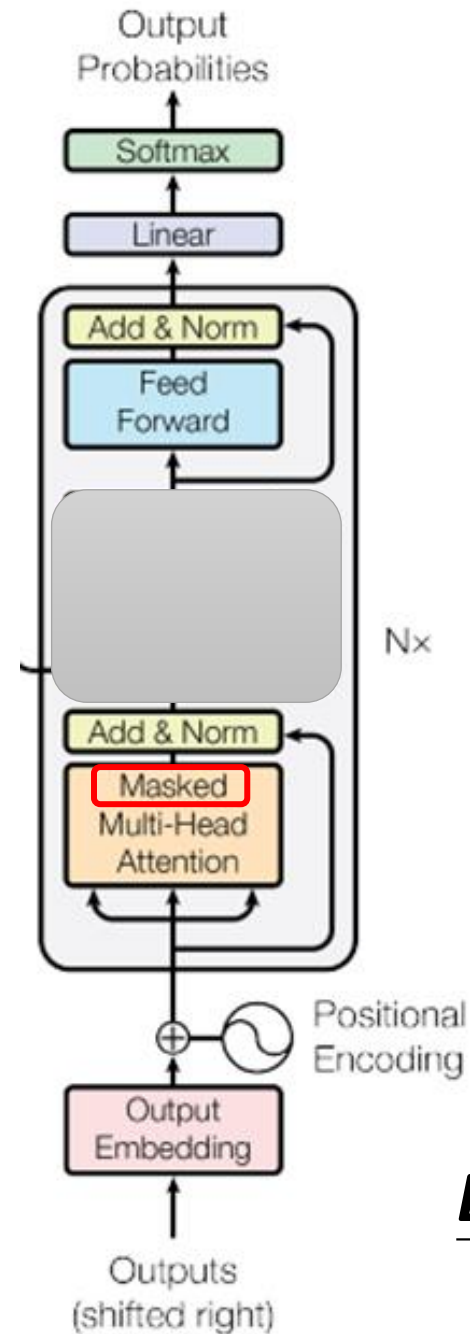
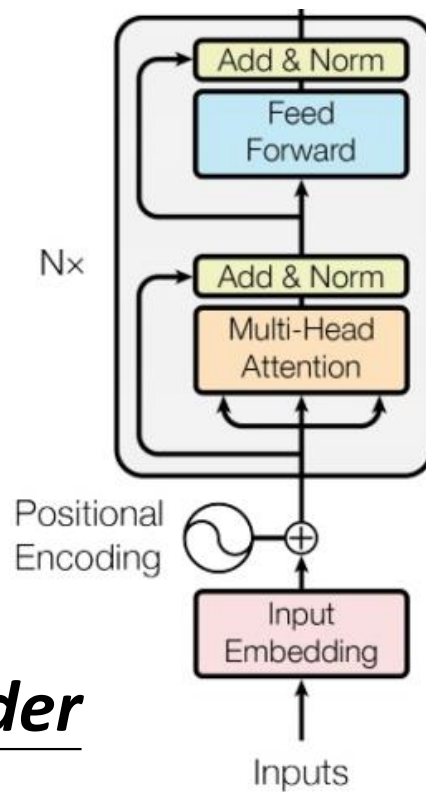
Autoregressive



ignore the input from the encoder here 😊

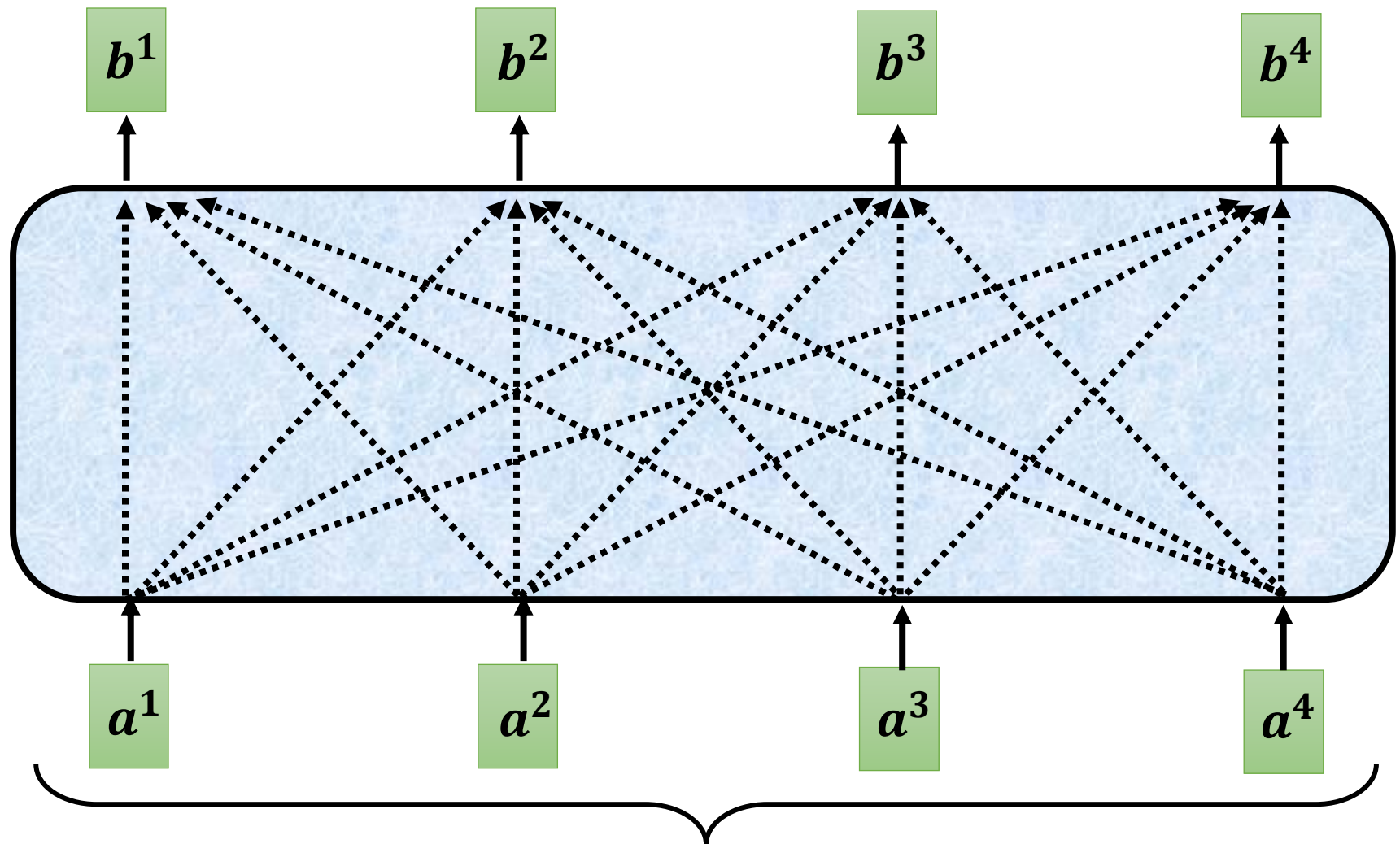


Encoder



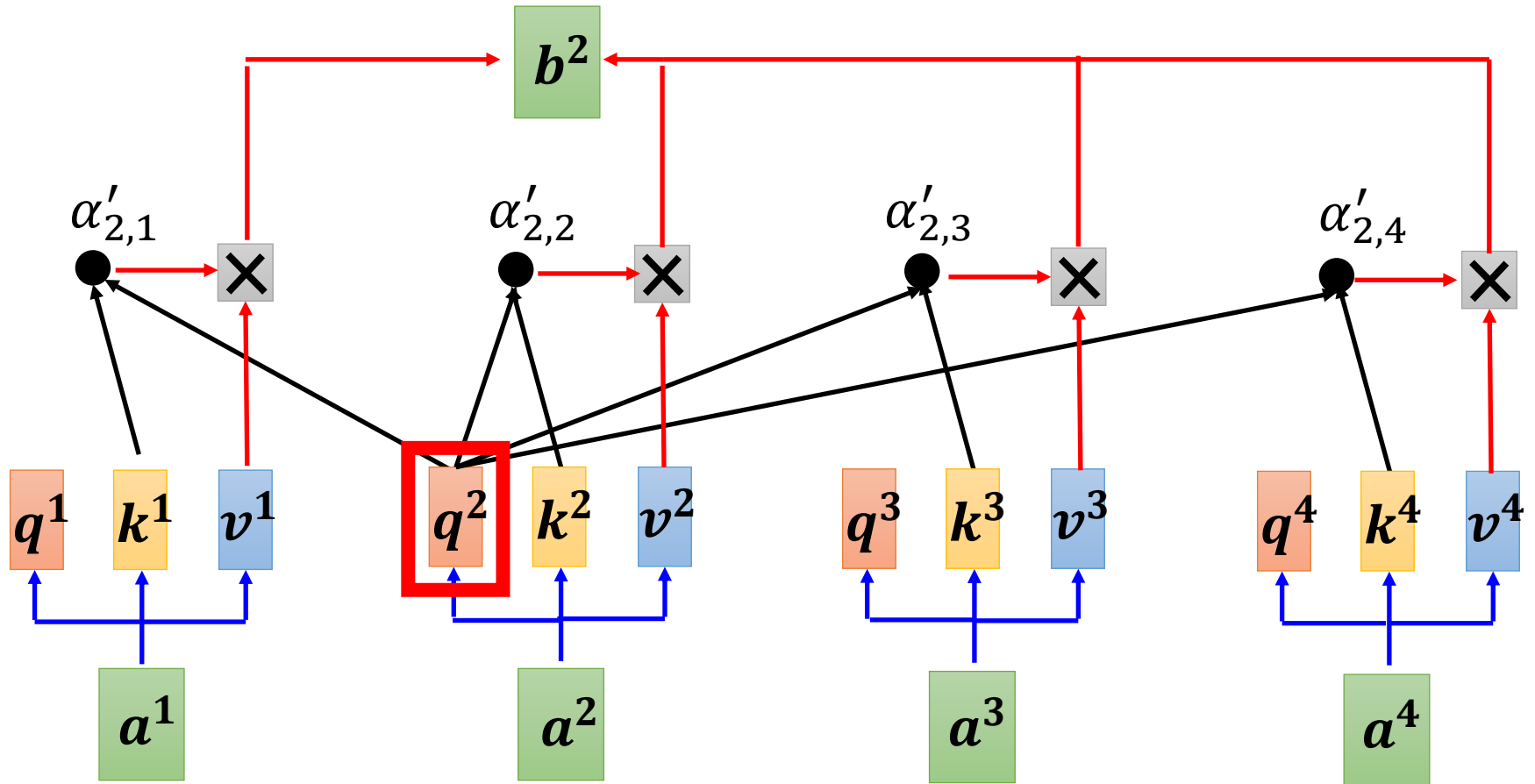
Decoder

Self-attention → Masked Self-attention



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

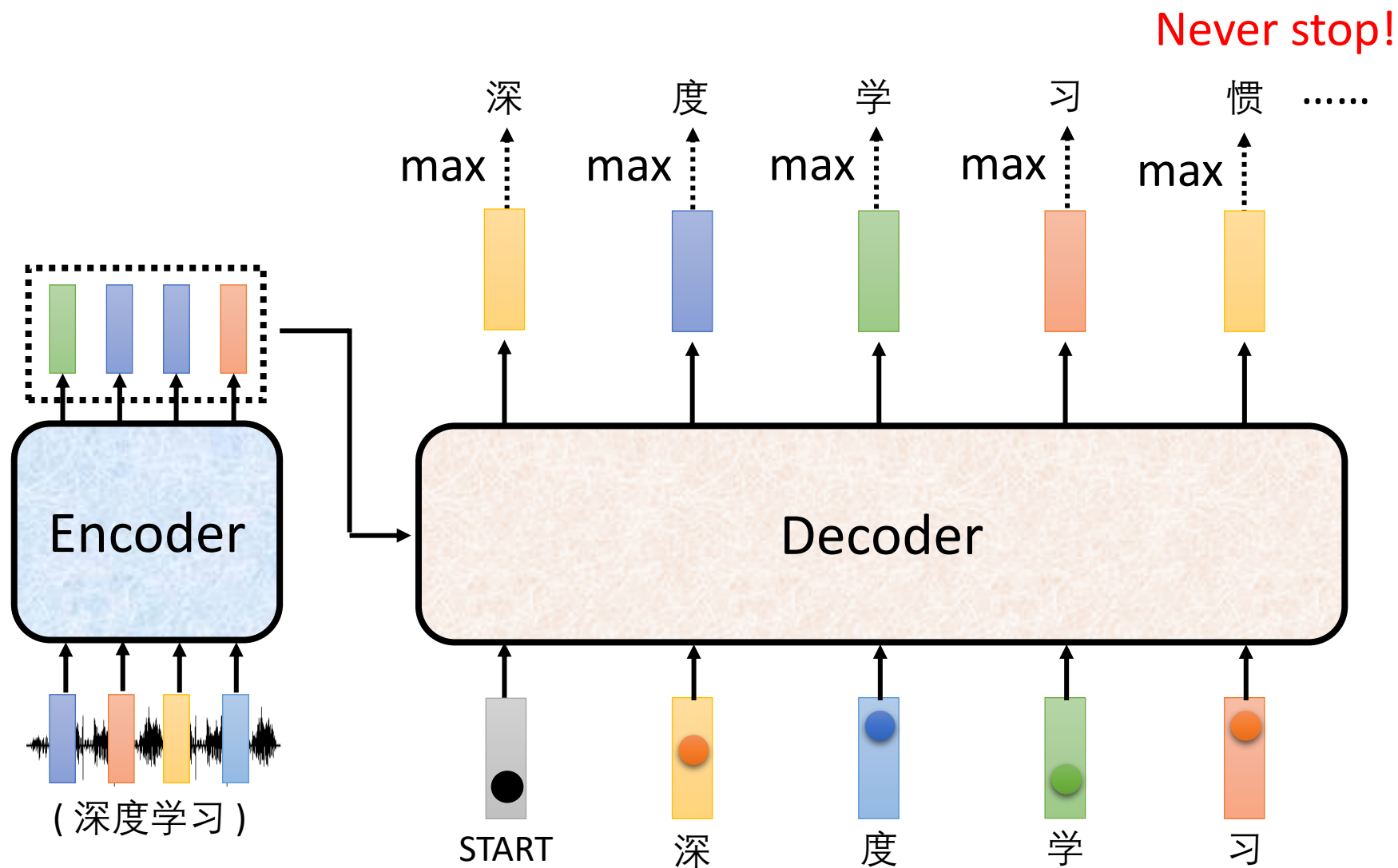
Self-attention \rightarrow Masked Self-attention



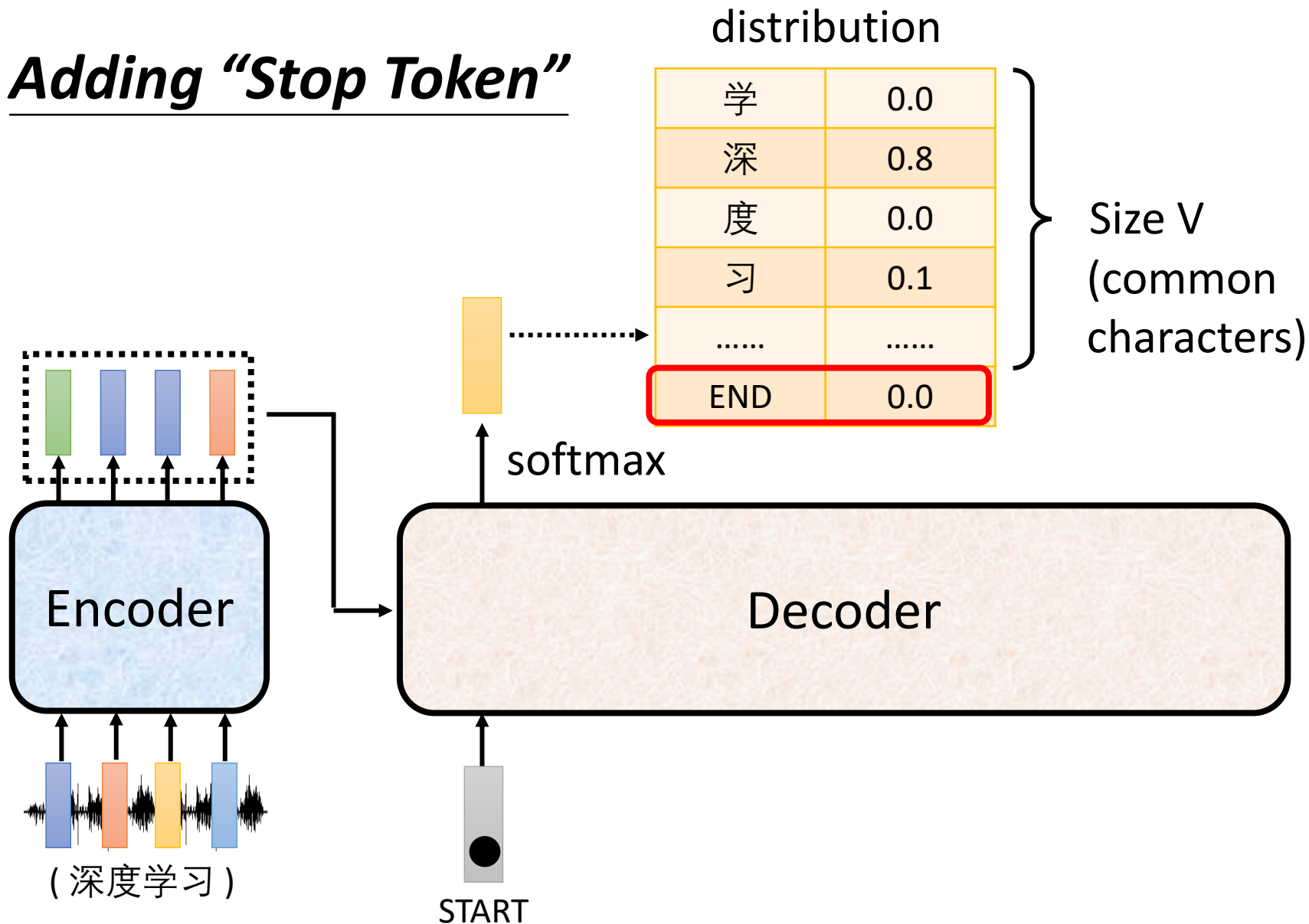
Why masked? Consider how does decoder work

Autoregressive

We do not know the correct output length.

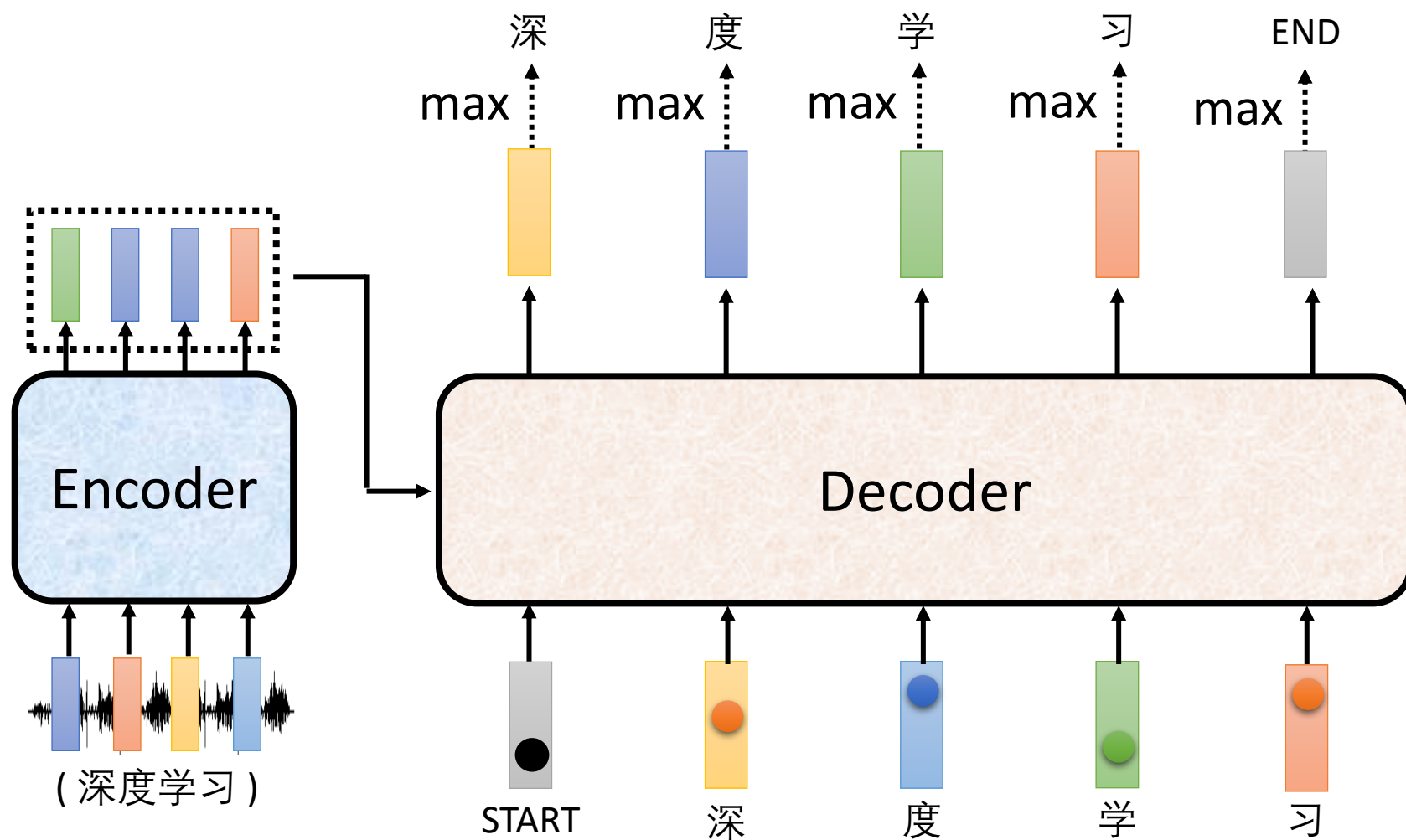


Adding “Stop Token”



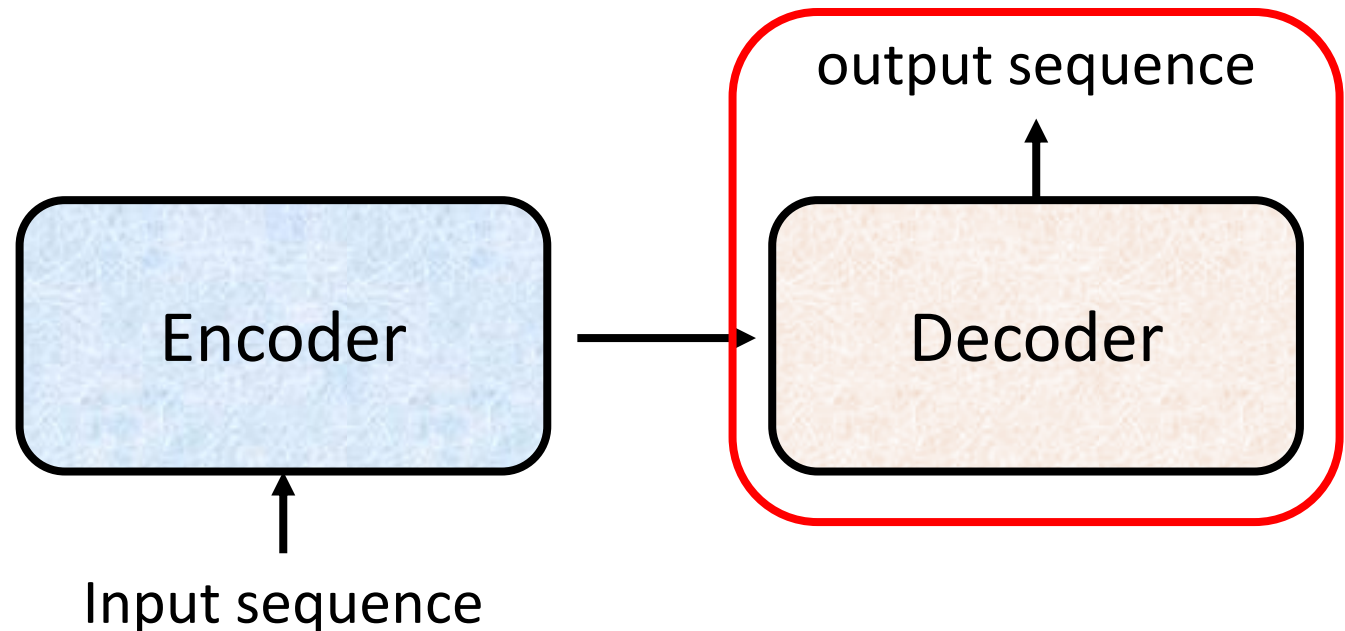
Autoregressive

Stop at here!

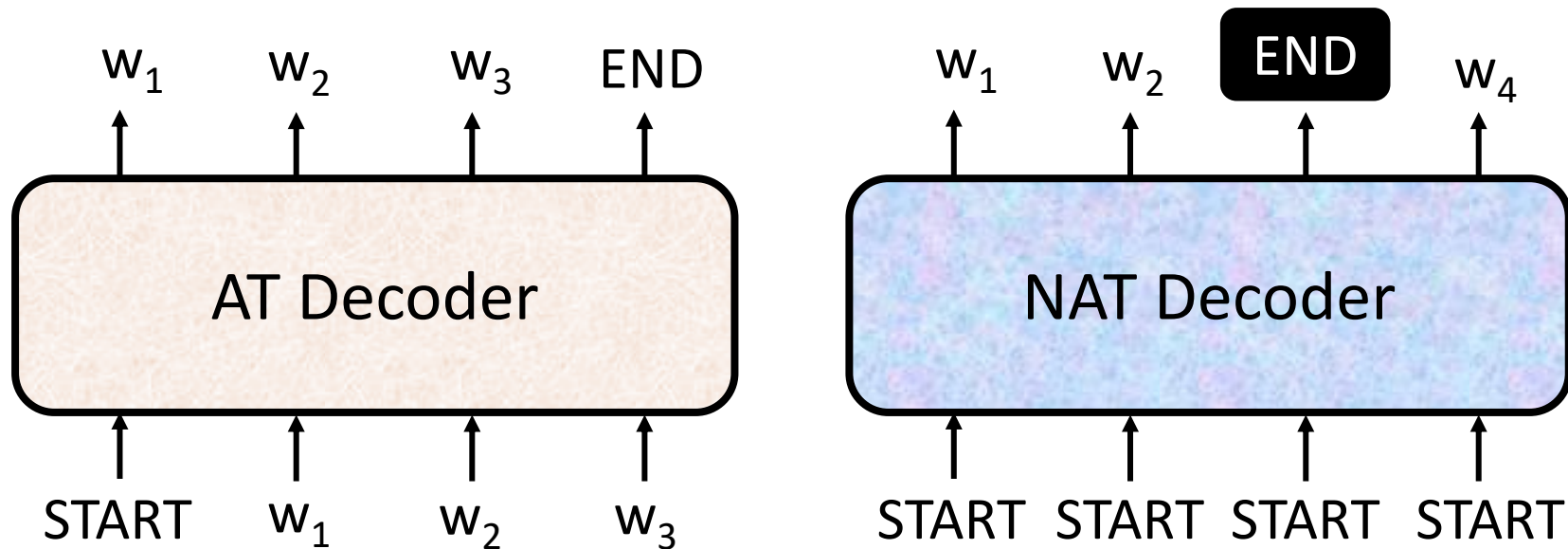


Decoder

- Non-autoregressive (NAT)

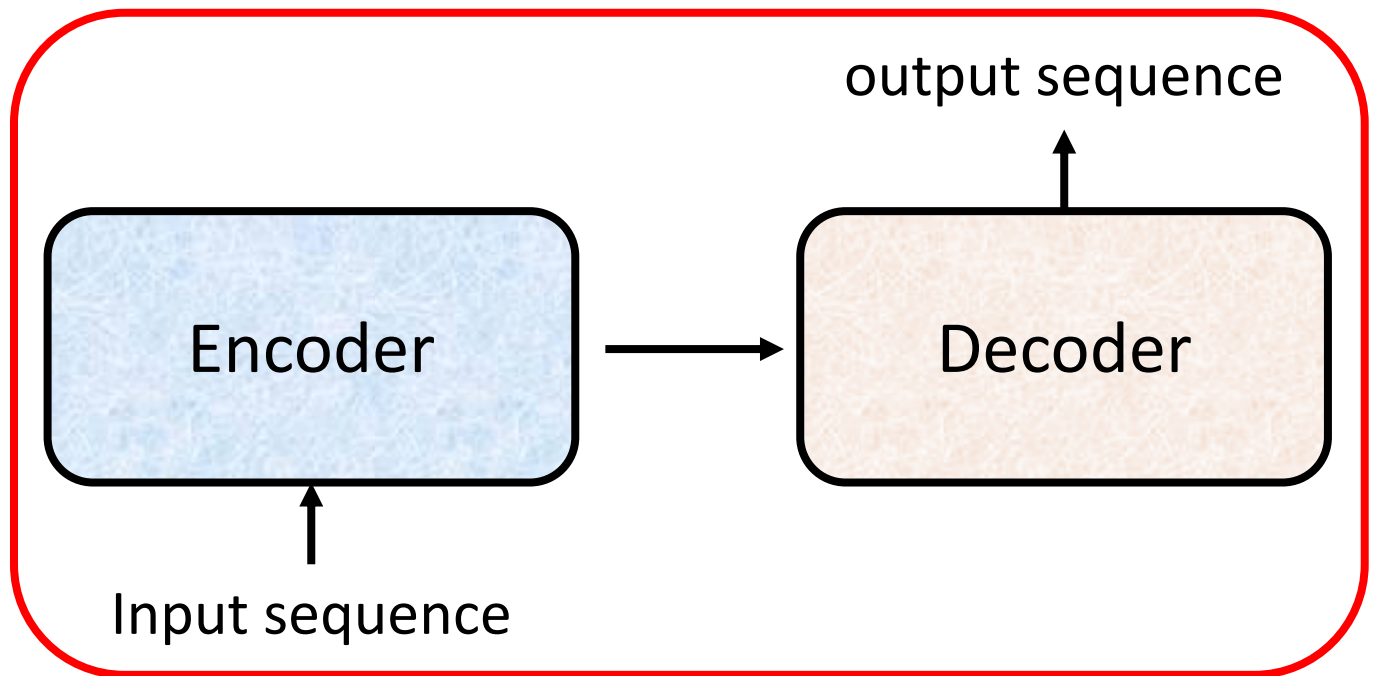


AT v.s. NAT

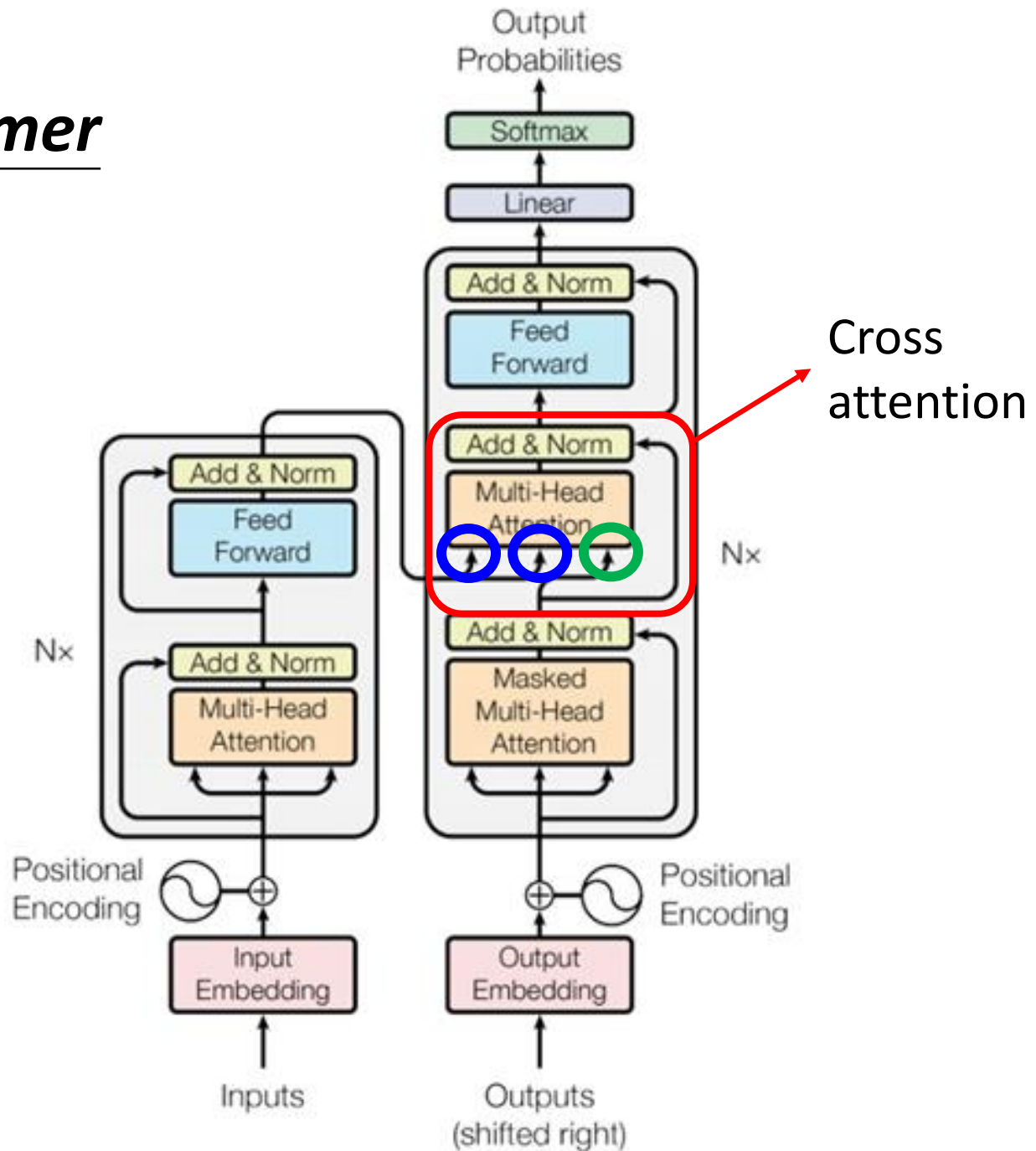


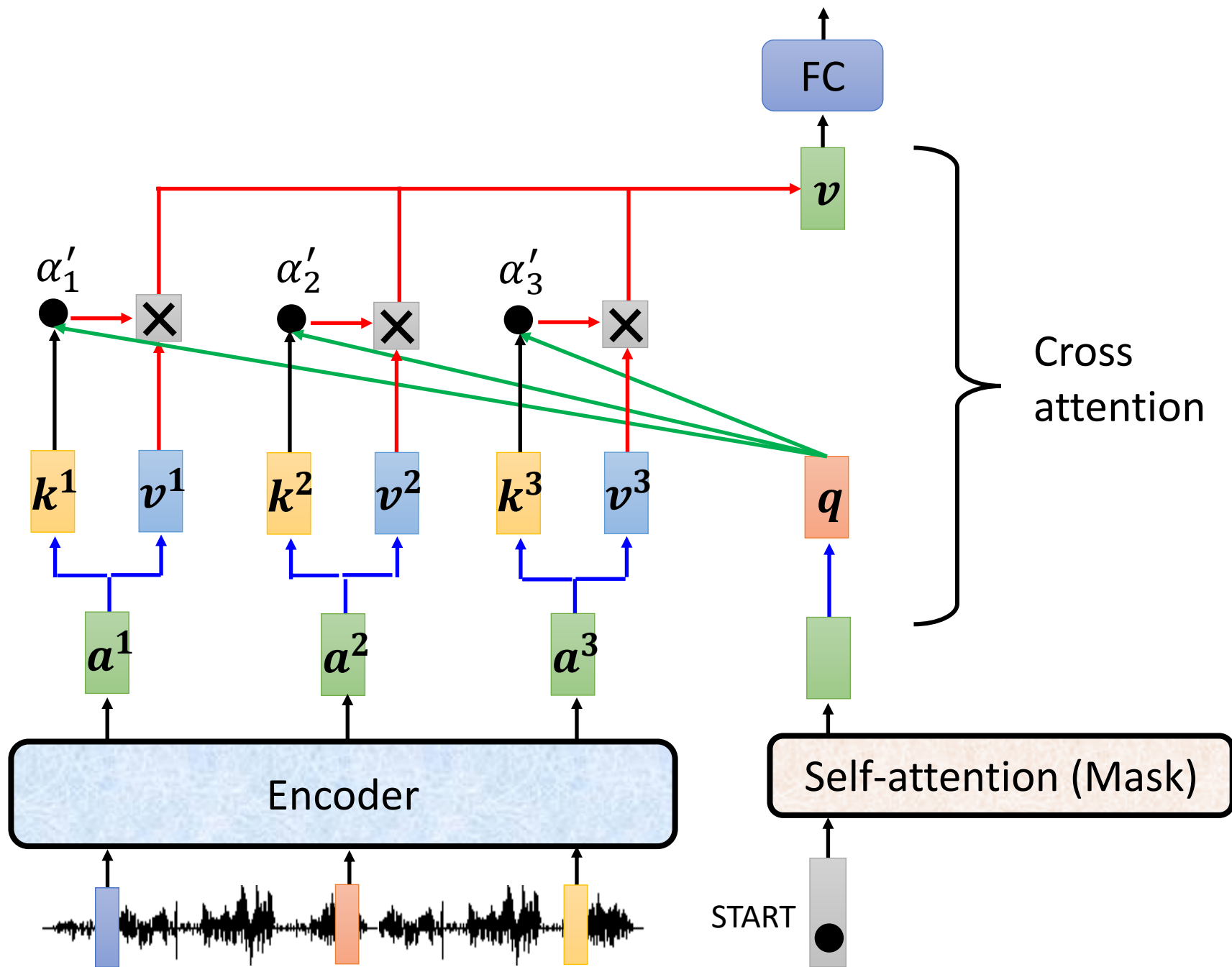
- How to decide the output length for NAT decoder?
 - Another predictor for output length
 - Output a very long sequence, ignore tokens after END
- Advantage: parallel, more stable generation (e.g., TTS)
- NAT is usually worse than AT (why? **Multi-modality**)

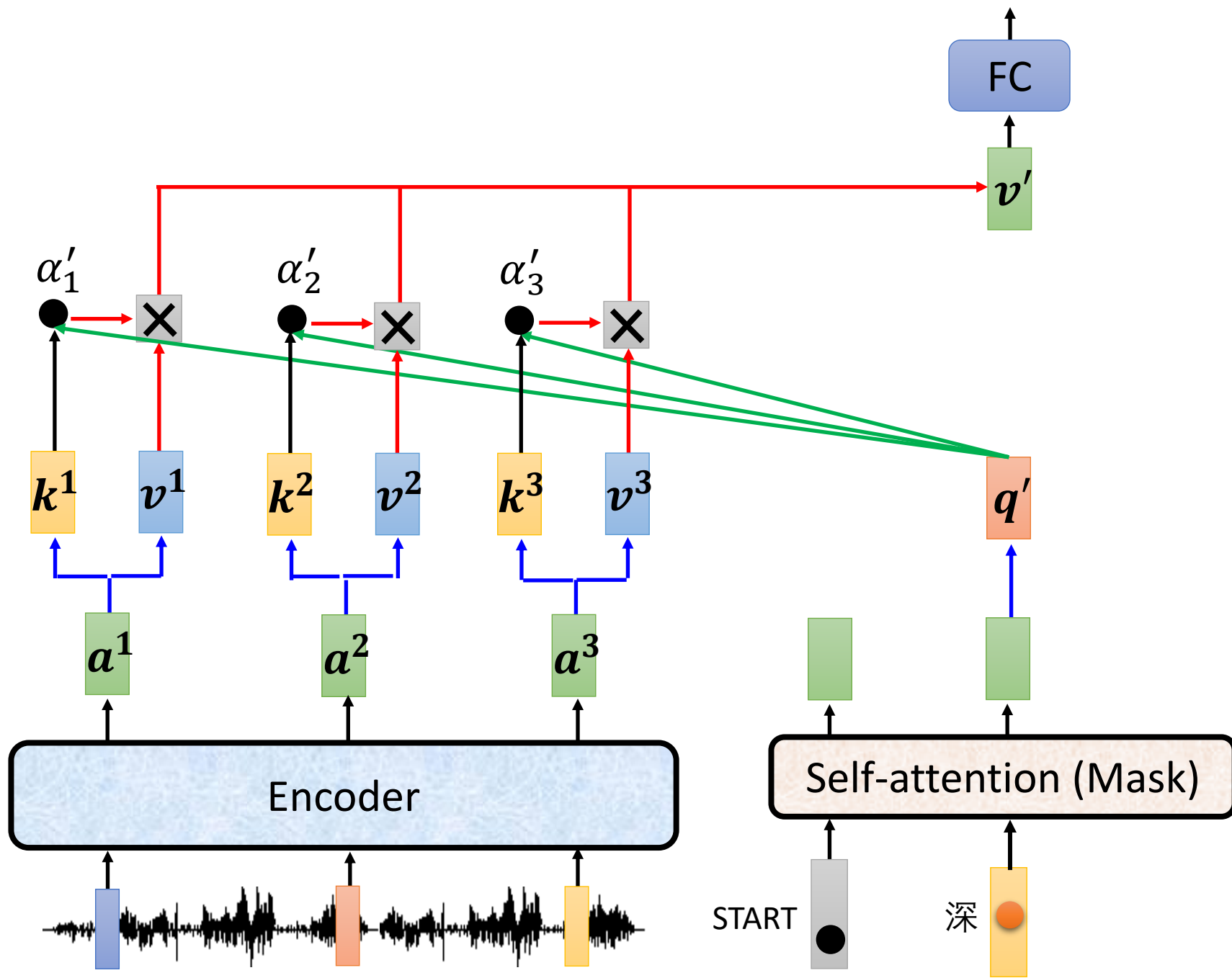
Encoder-Decoder



Transformer



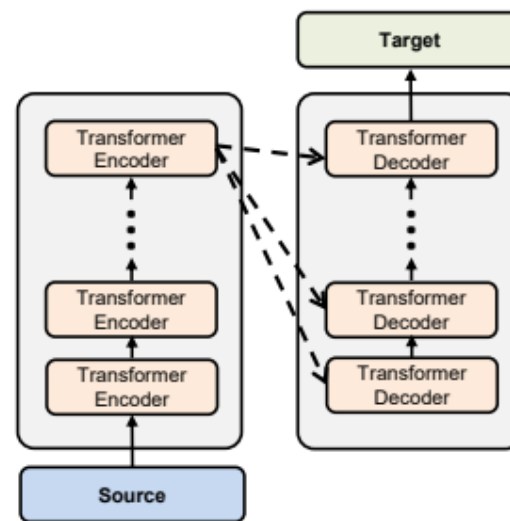




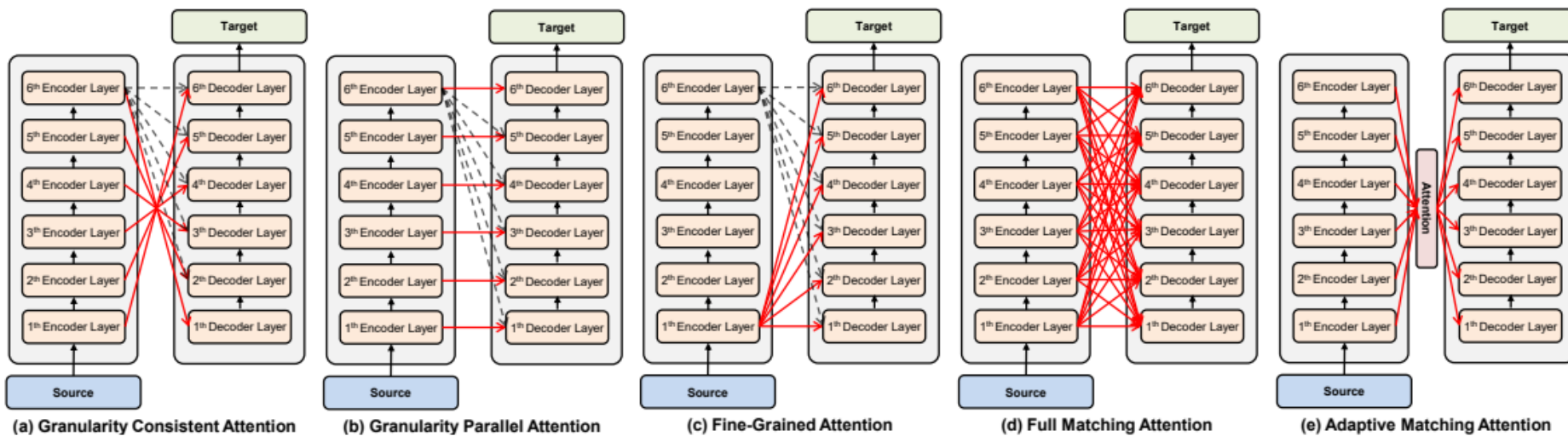
Cross Attention

Source of image:

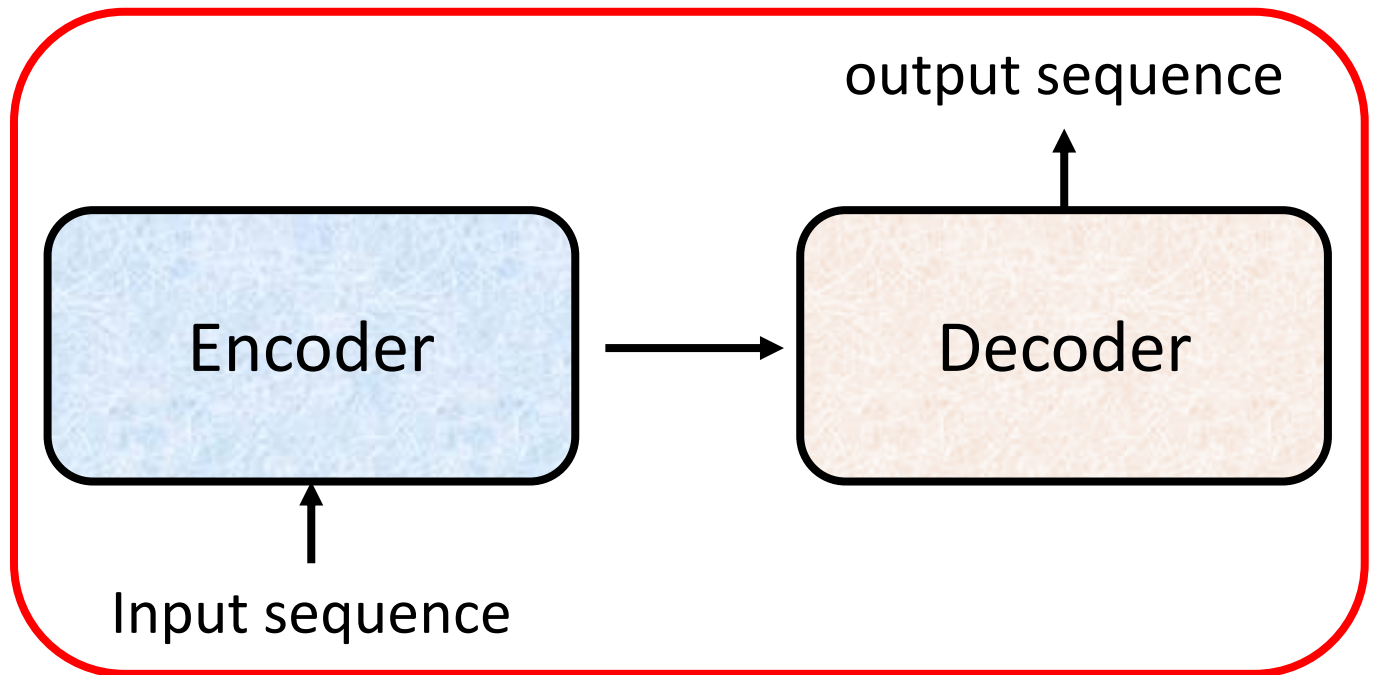
<https://arxiv.org/abs/2005.08081>

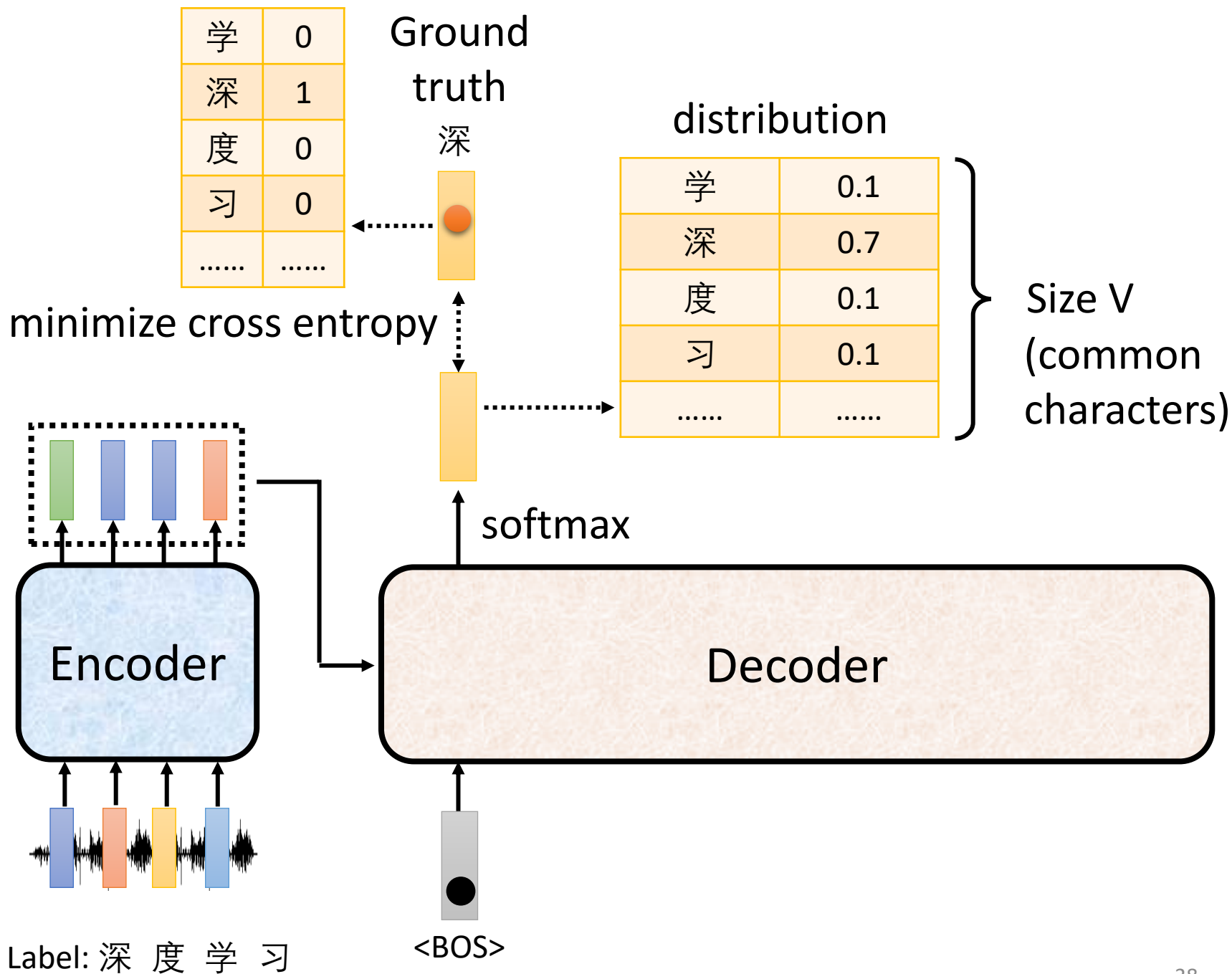


(a) Conventional Transformer

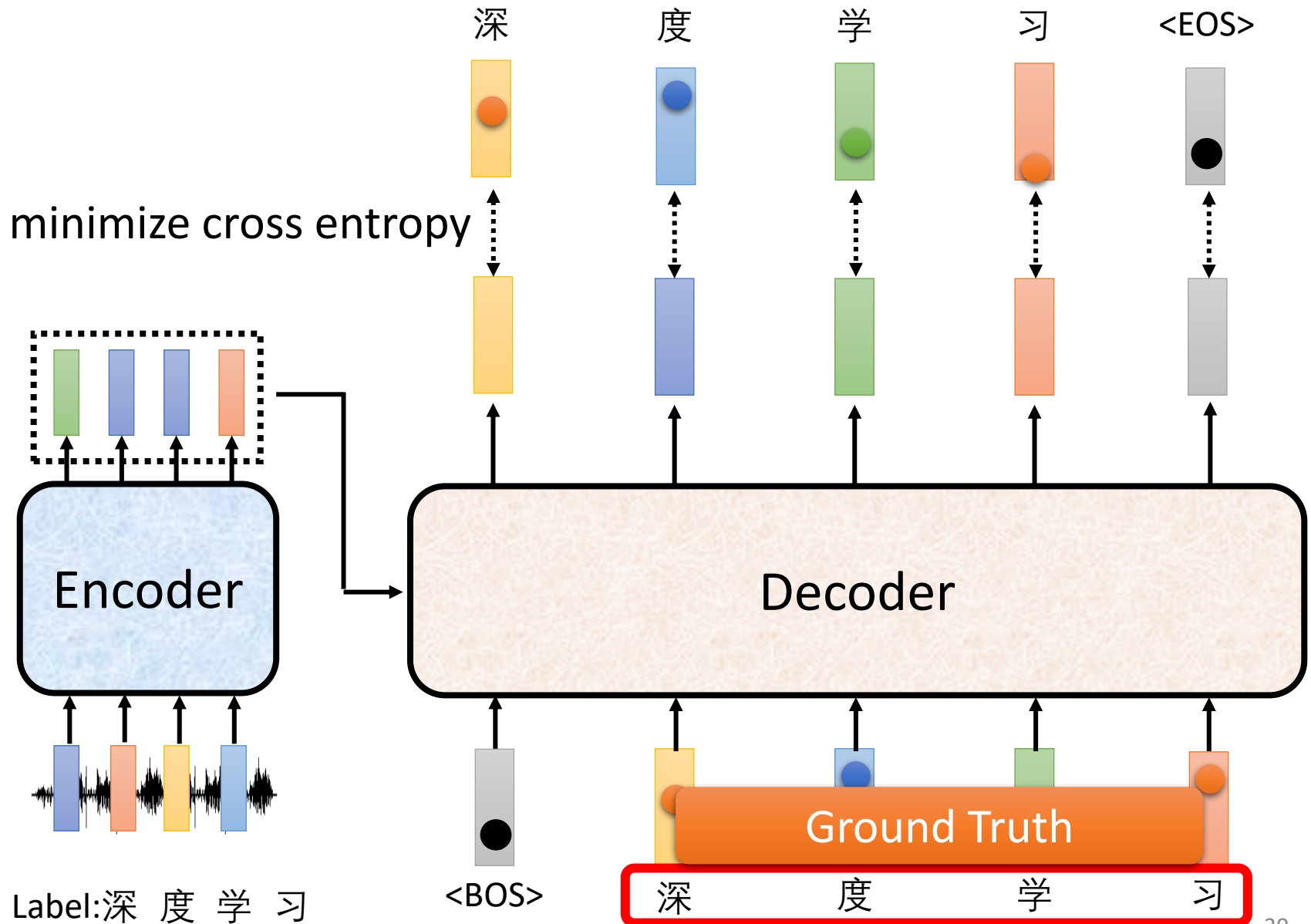


Training

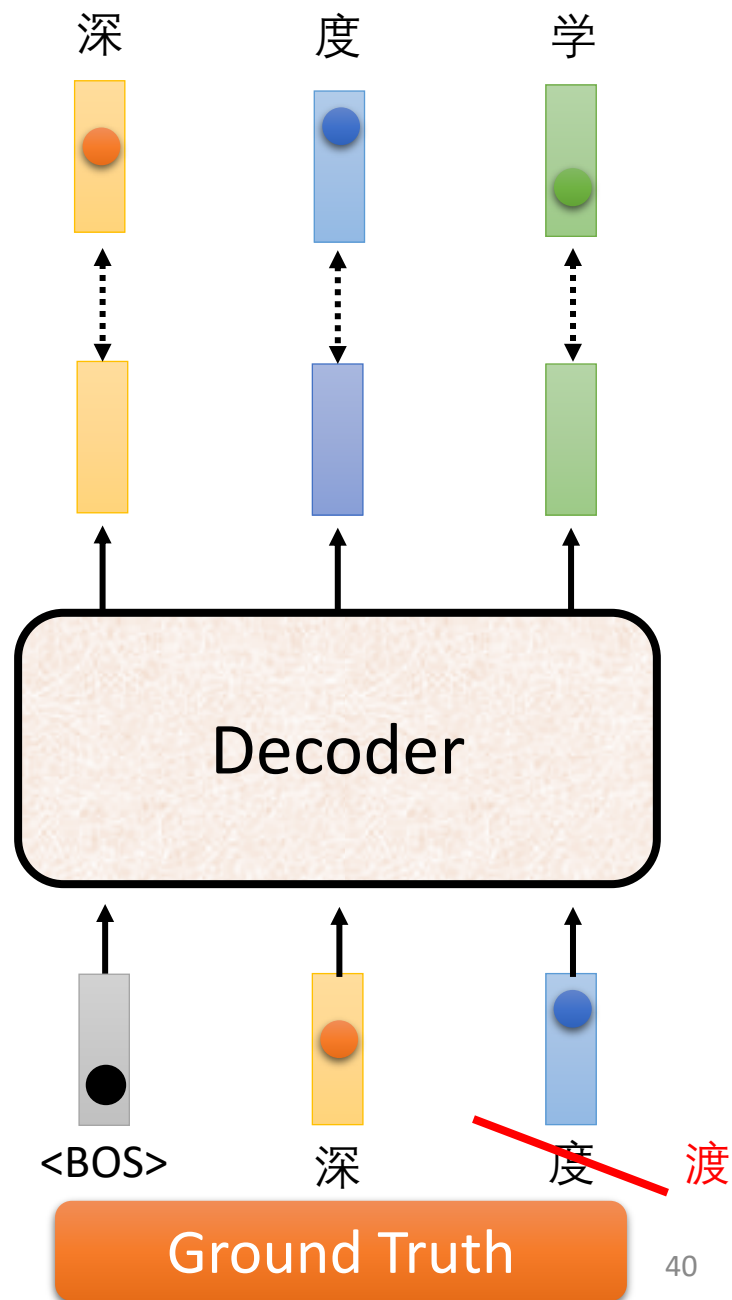
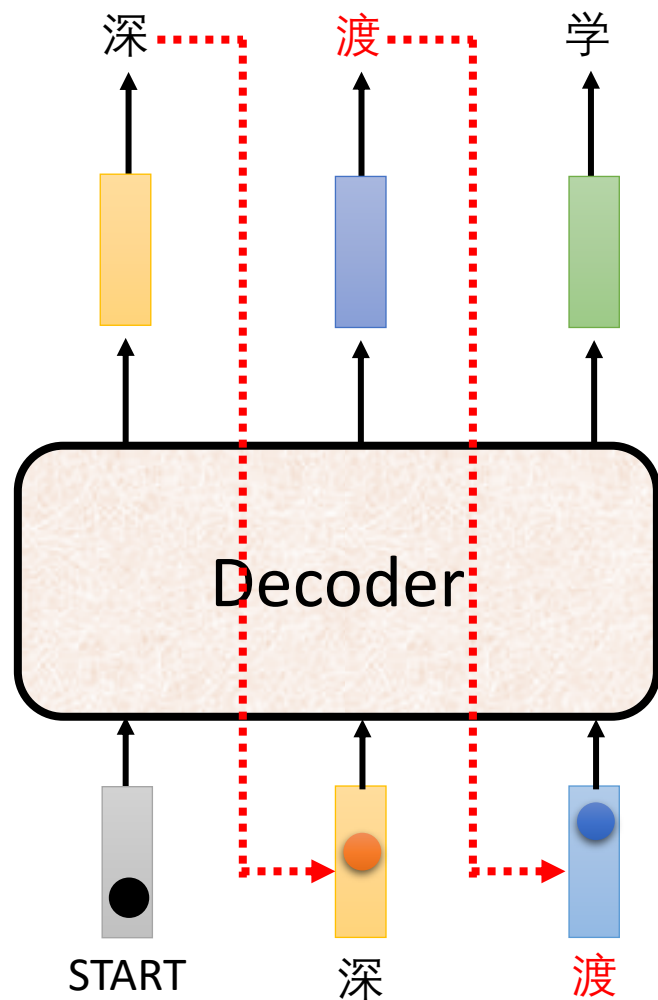




Teacher Forcing: using the ground truth as input.



There is a mismatch! 😞
exposure bias



Scheduled Sampling

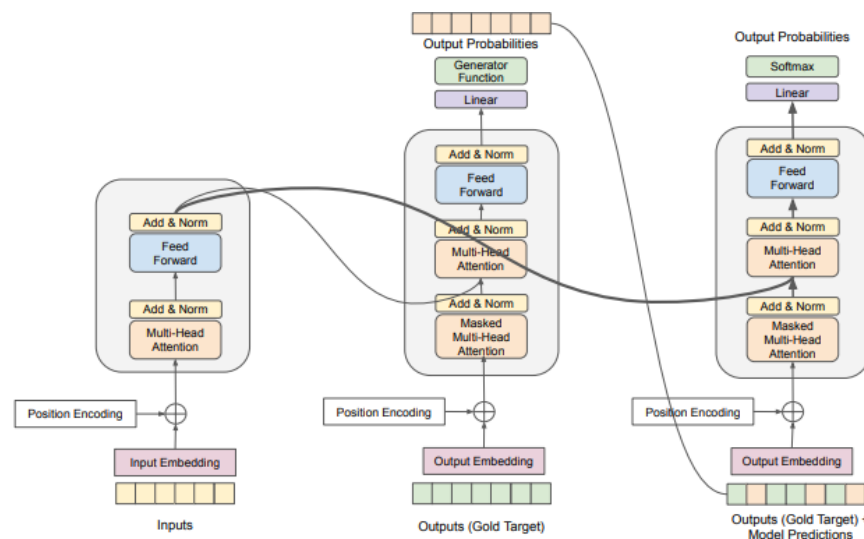
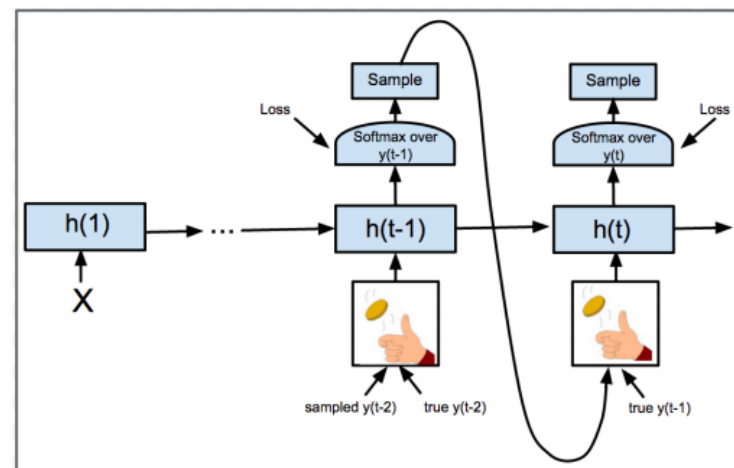
- Original Scheduled Sampling
- Scheduled Sampling for Transformer

<https://arxiv.org/abs/1506.03099>

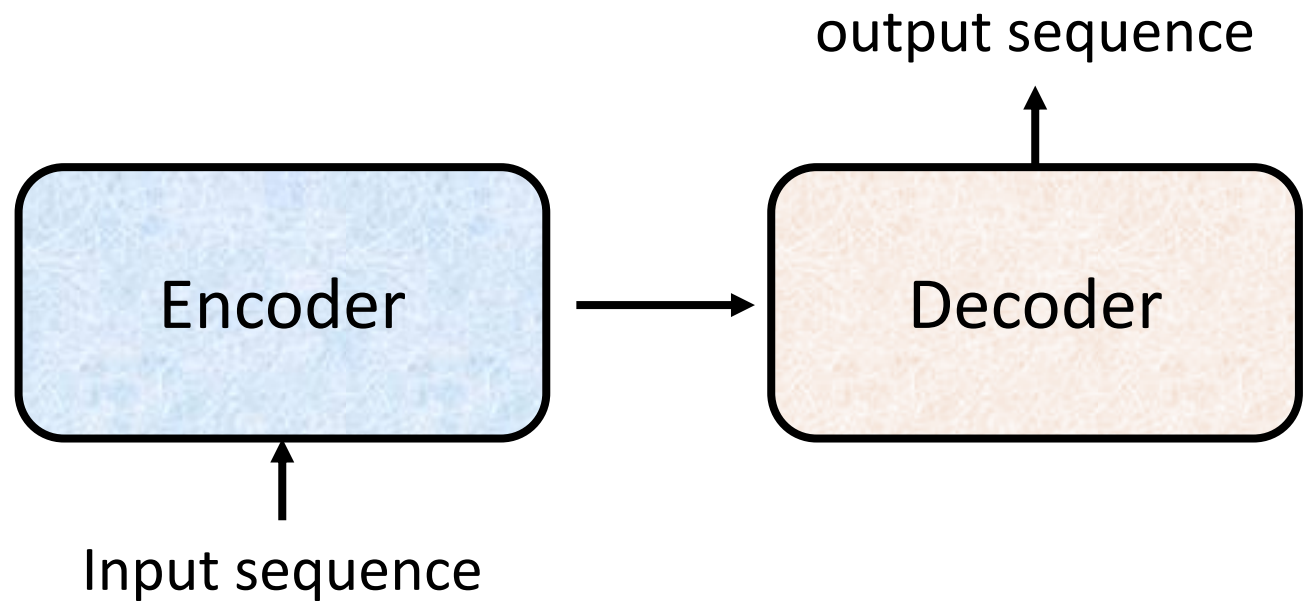
<https://arxiv.org/abs/1906.07651>

- Parallel Scheduled Sampling

<https://arxiv.org/abs/1906.04331>

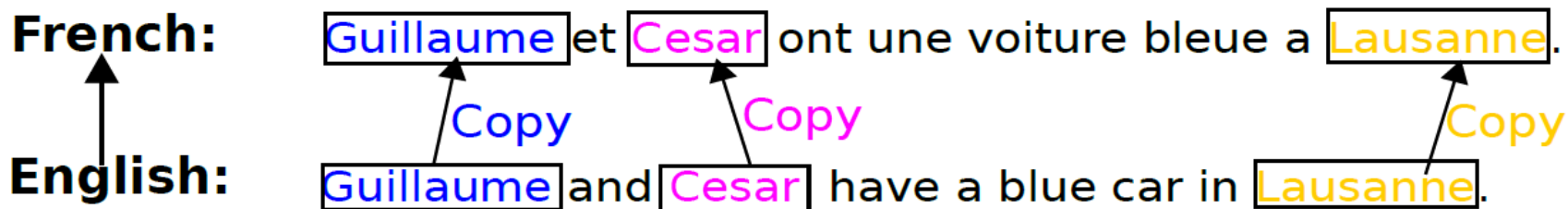


Tips



Copy Mechanism

Machine Translation



Chat-bot

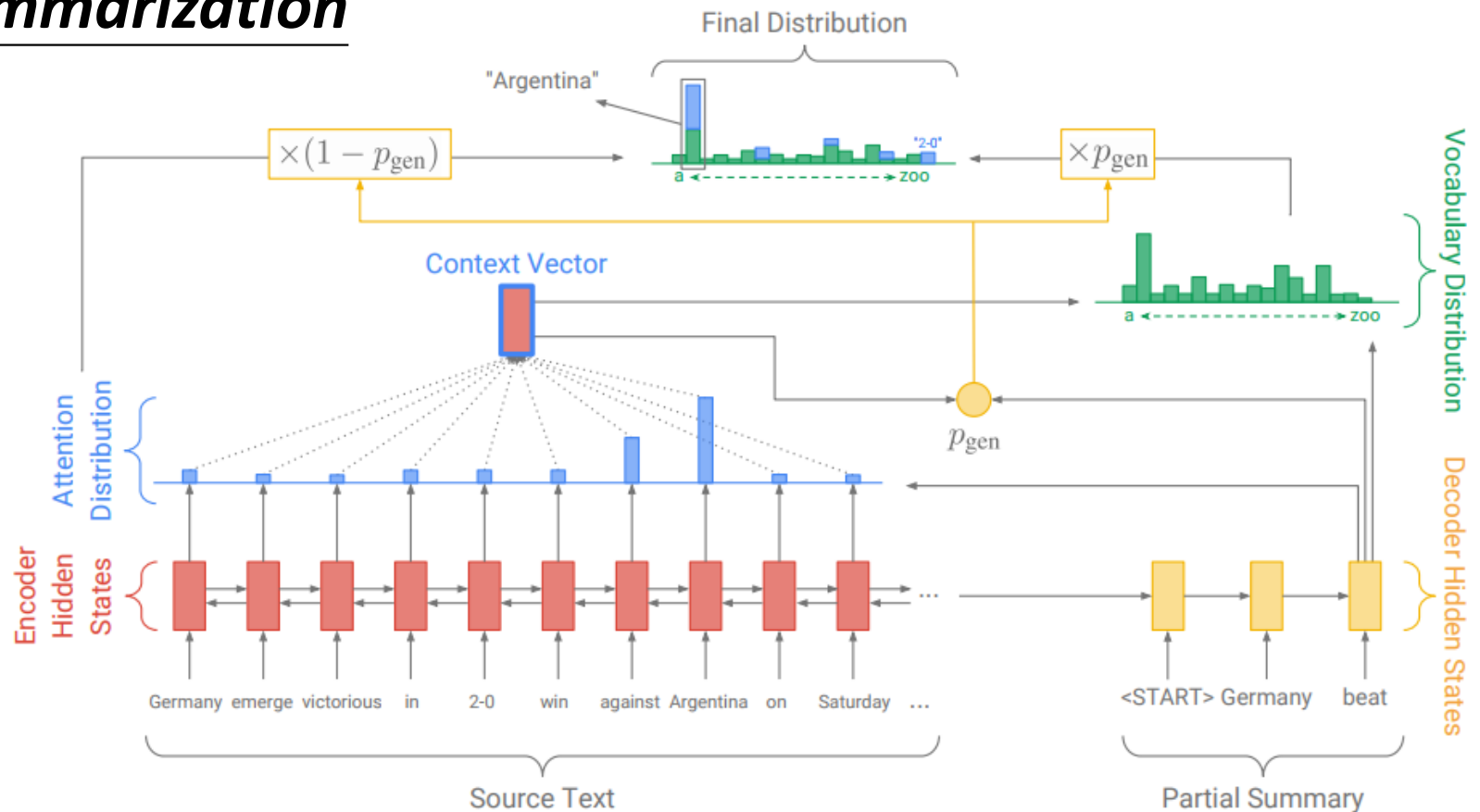
User: 你好，我是甘道夫

Machine: 甘道夫你好，很高兴认识你

Copy Mechanism

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

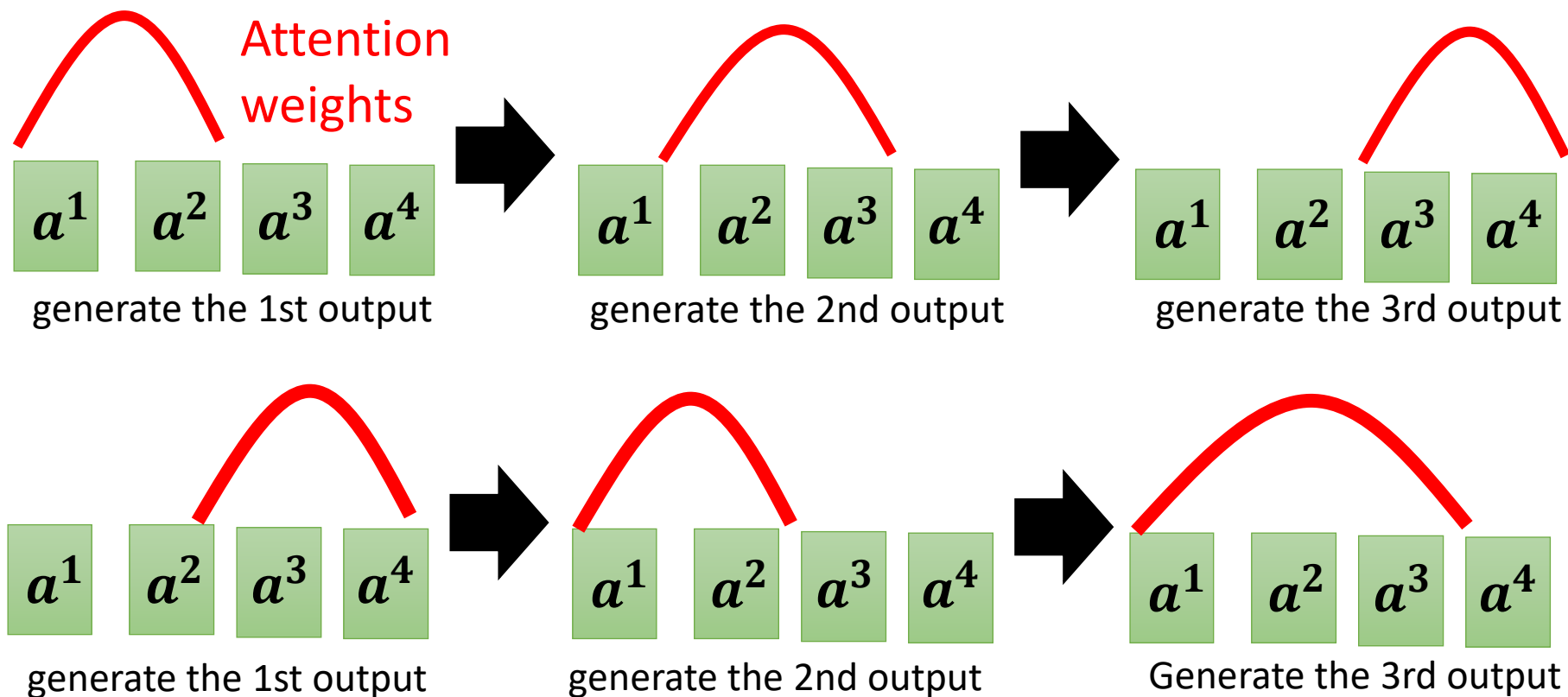
Summarization



Guided Attention

Monotonic Attention
Location-aware attention

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



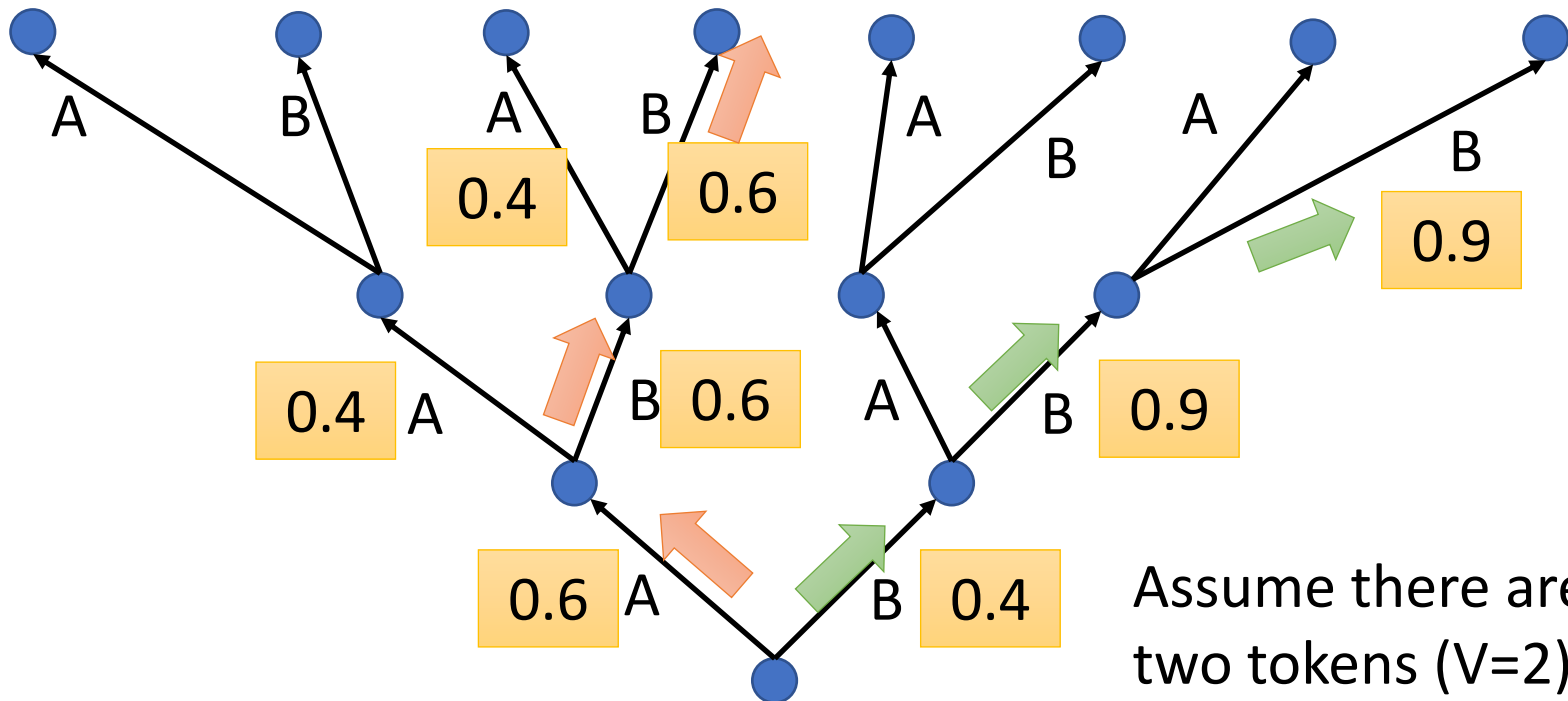
Something wrong!

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, $b=32$:

[illegible]

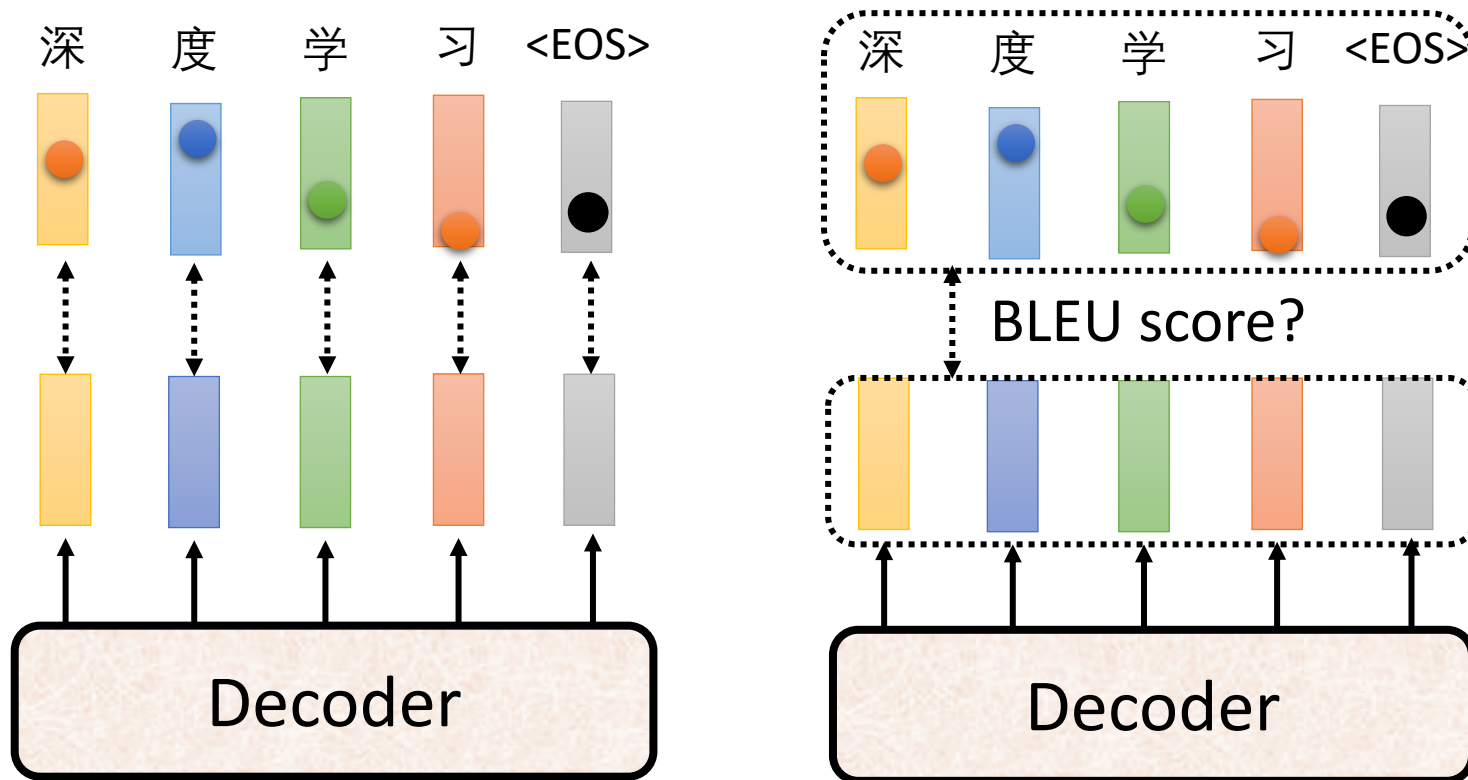
Pure Sampling:

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>

Concluding Remarks: Transformer

