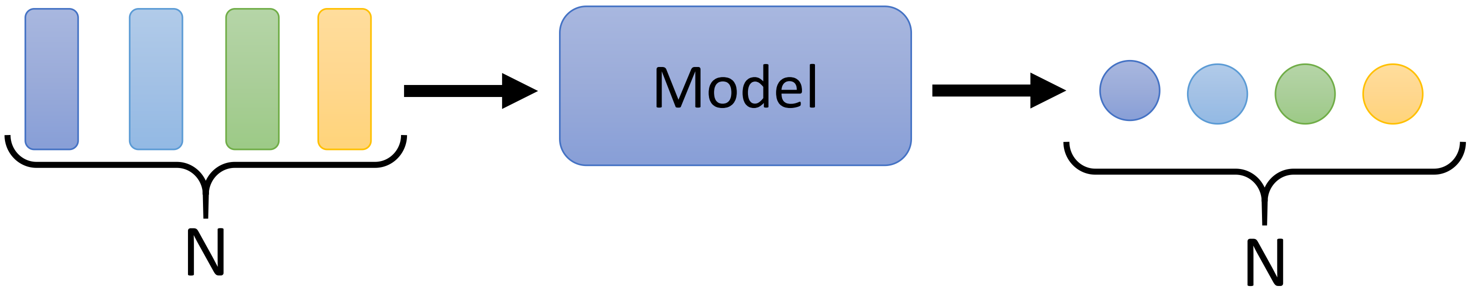
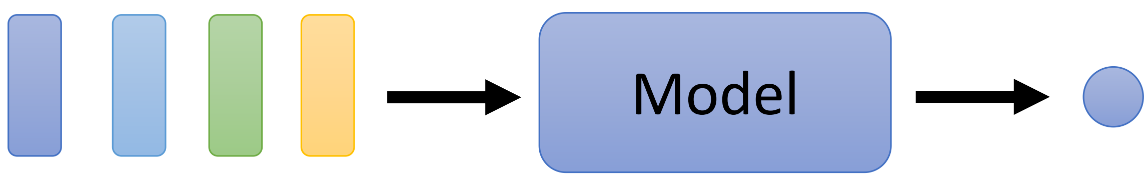
Transformer

Sequence-to-sequence (Seq2seq)

- Each vector has a label.



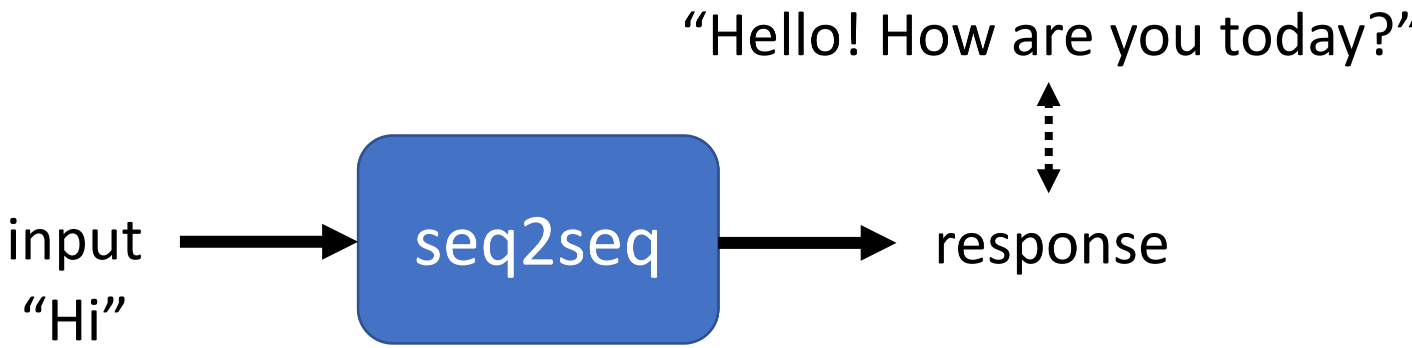
- The whole sequence has a label.



- Model decides the number of labels itself.

seq2seq

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

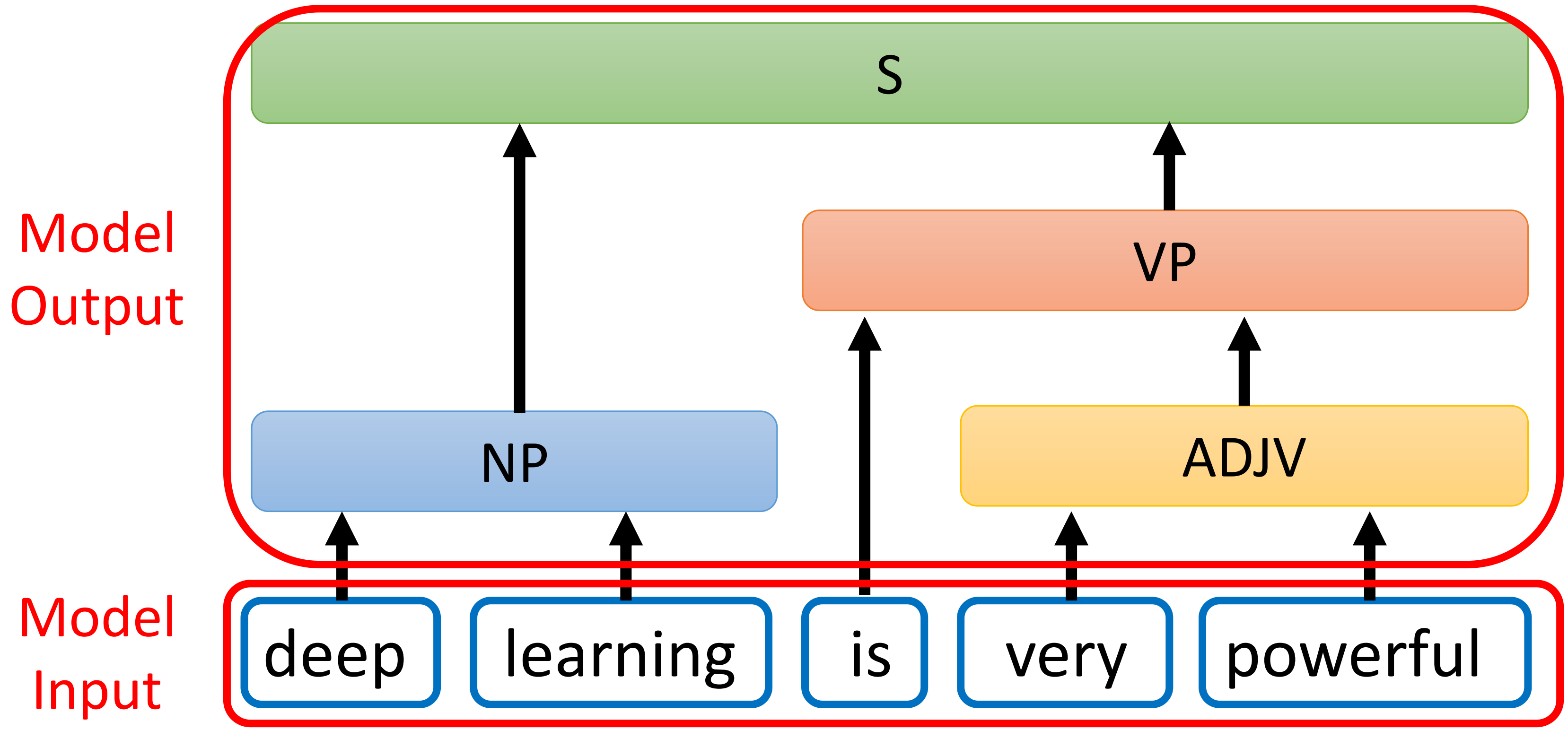
QA can be done by **seq2seq**

question, context → Seq2seq → answer

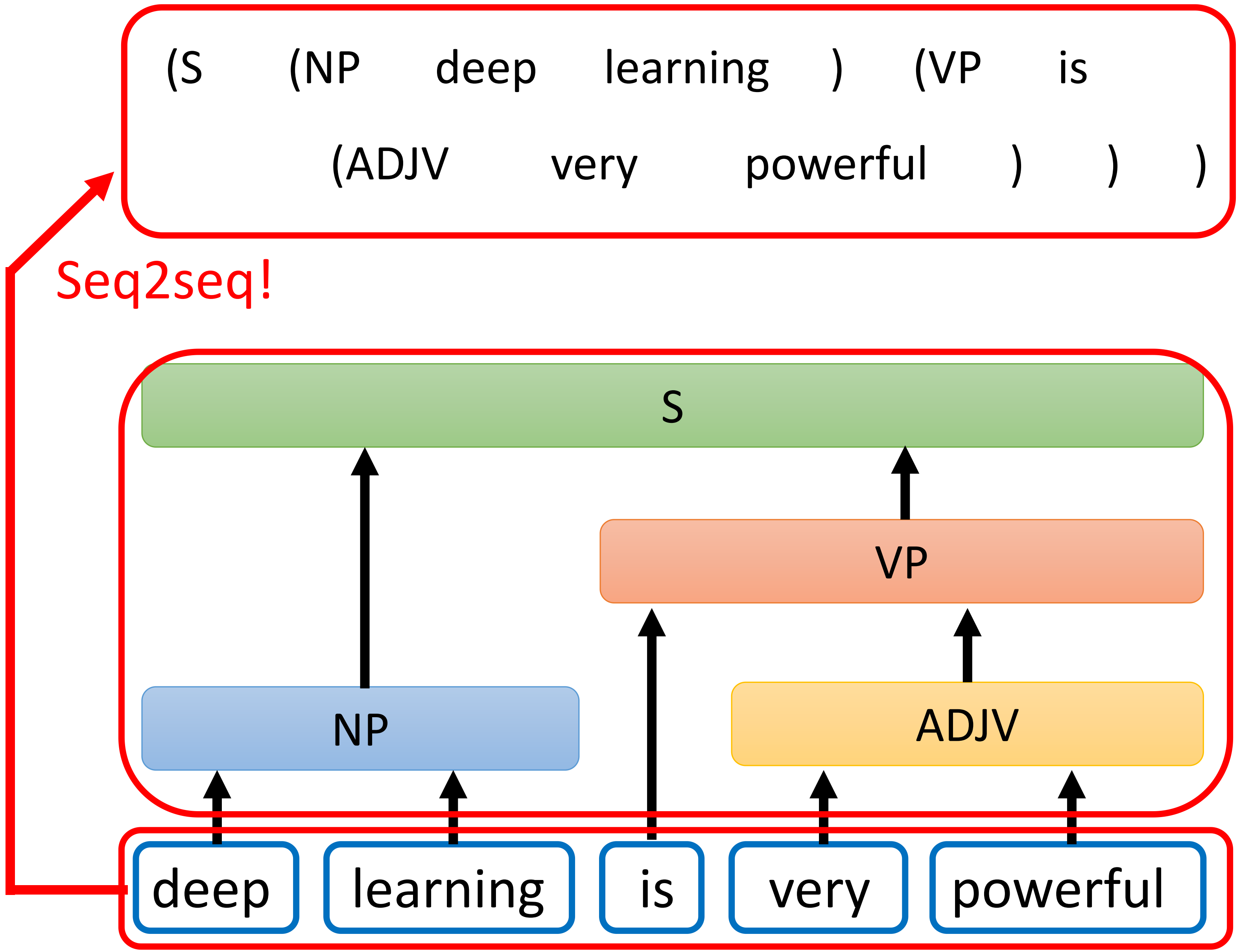
https://arxiv.org/abs/1806.08730
https://arxiv.org/abs/1909.03329

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

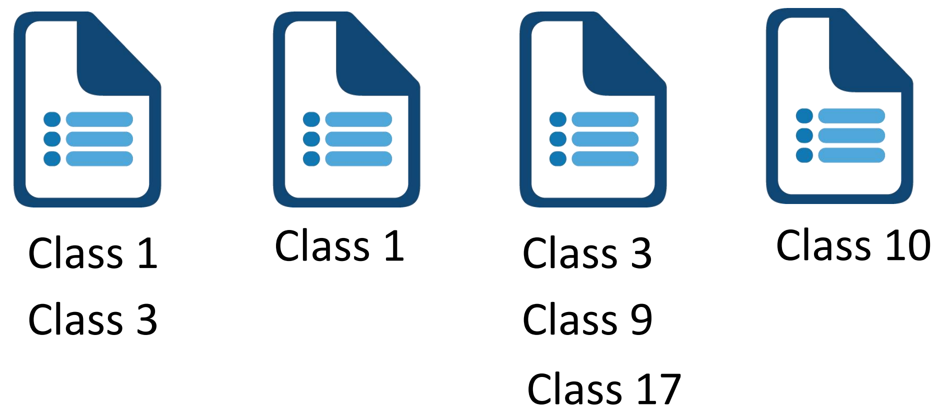
https://arxiv.org/abs/1412.7449

deep learning is very powerful

c.f. Multi-class Classification

Seq2seq for Multi-label Classification

An object can belong to multiple classes.

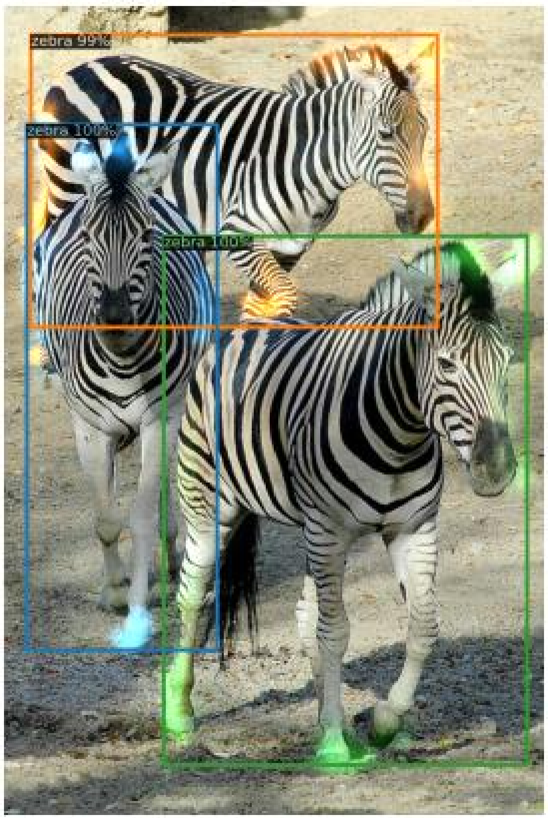


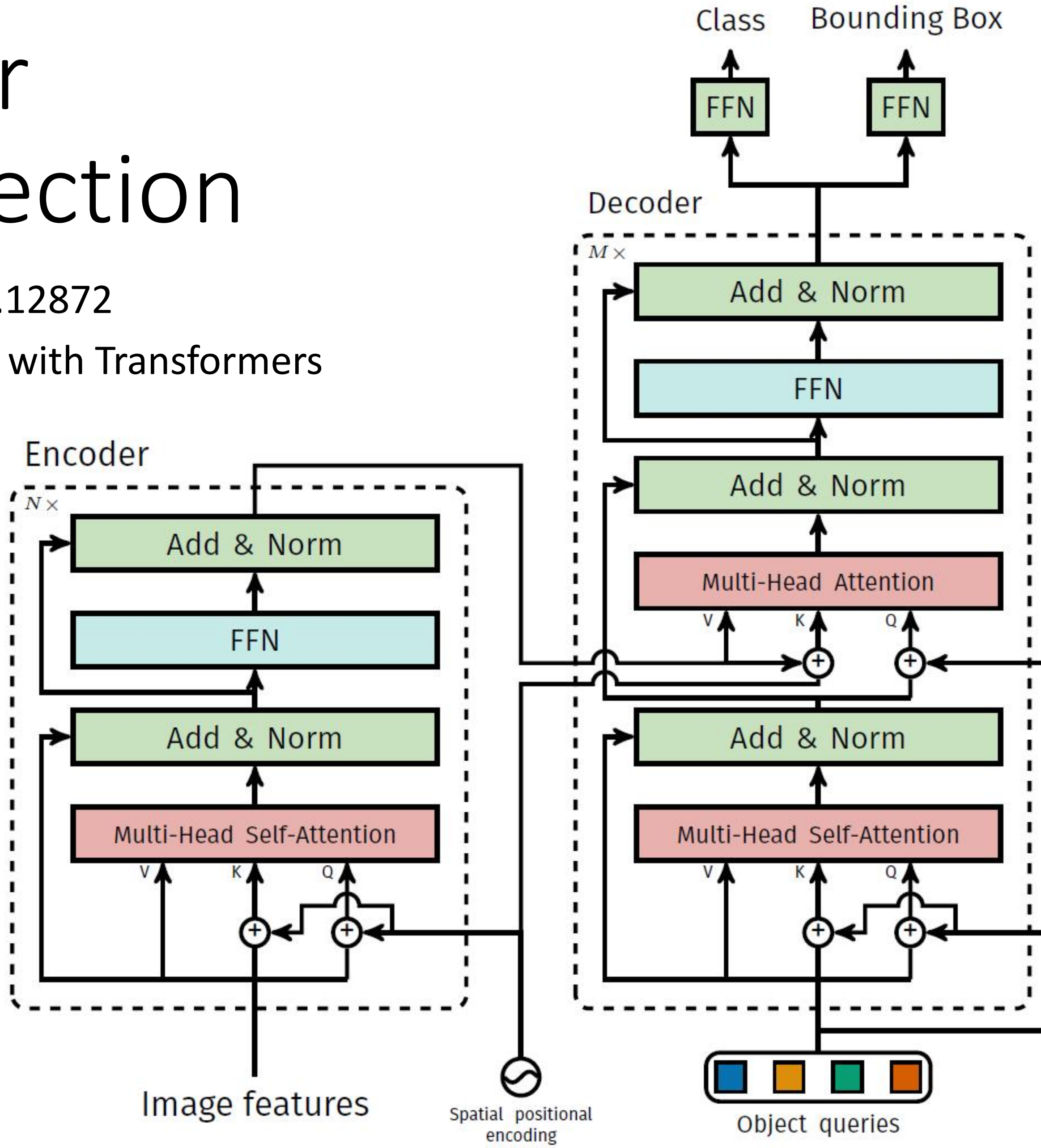
https://arxiv.org/abs/1909.03434
https://arxiv.org/abs/1707.05495

Seq2seq for Object Detection

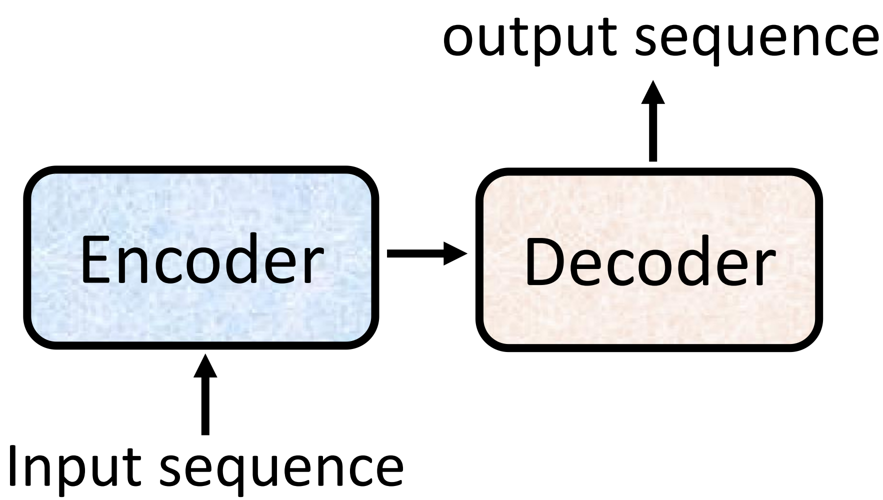
https://arxiv.org/abs/2005.12872

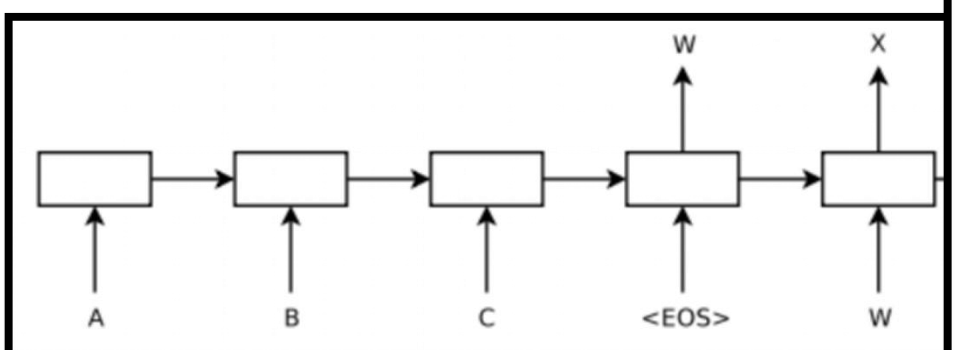
End-to-End Object Detection with Transformers

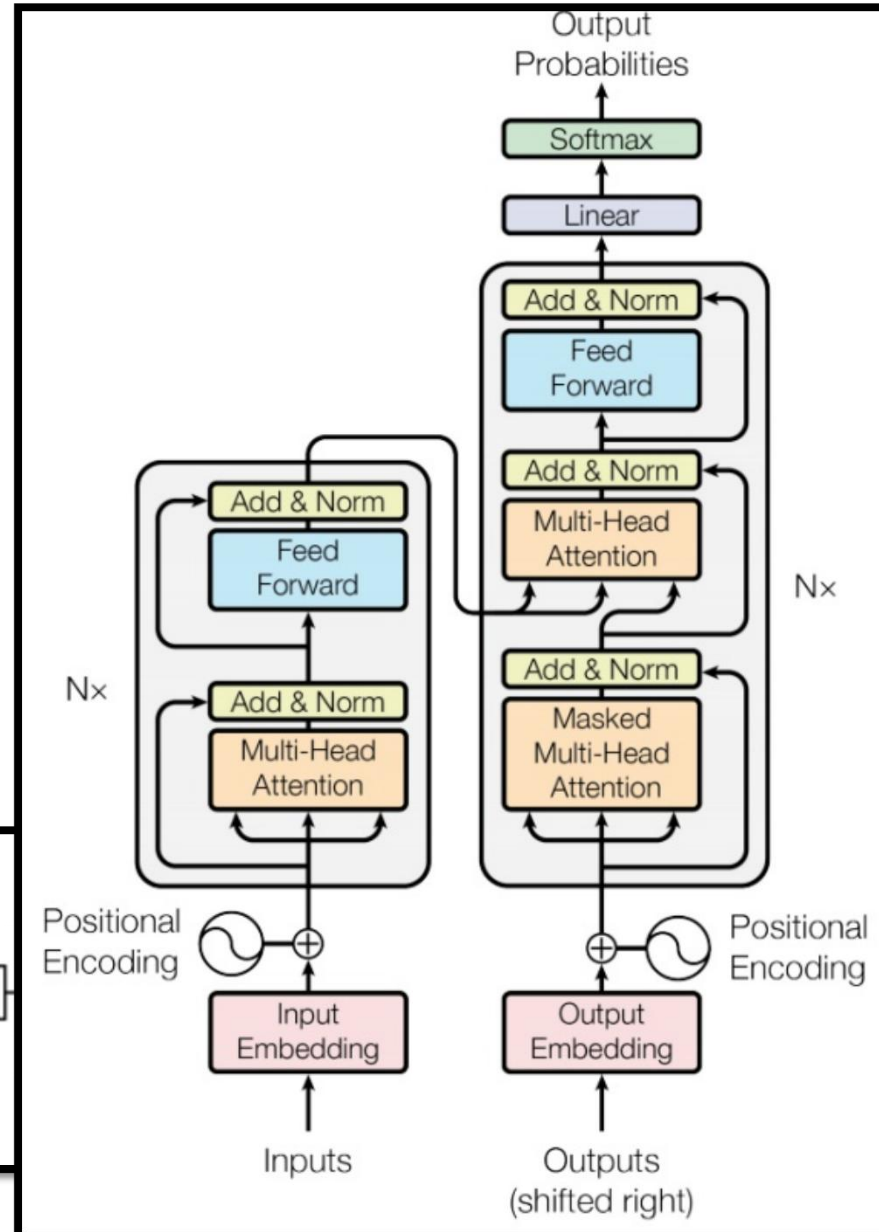




Seq2seq







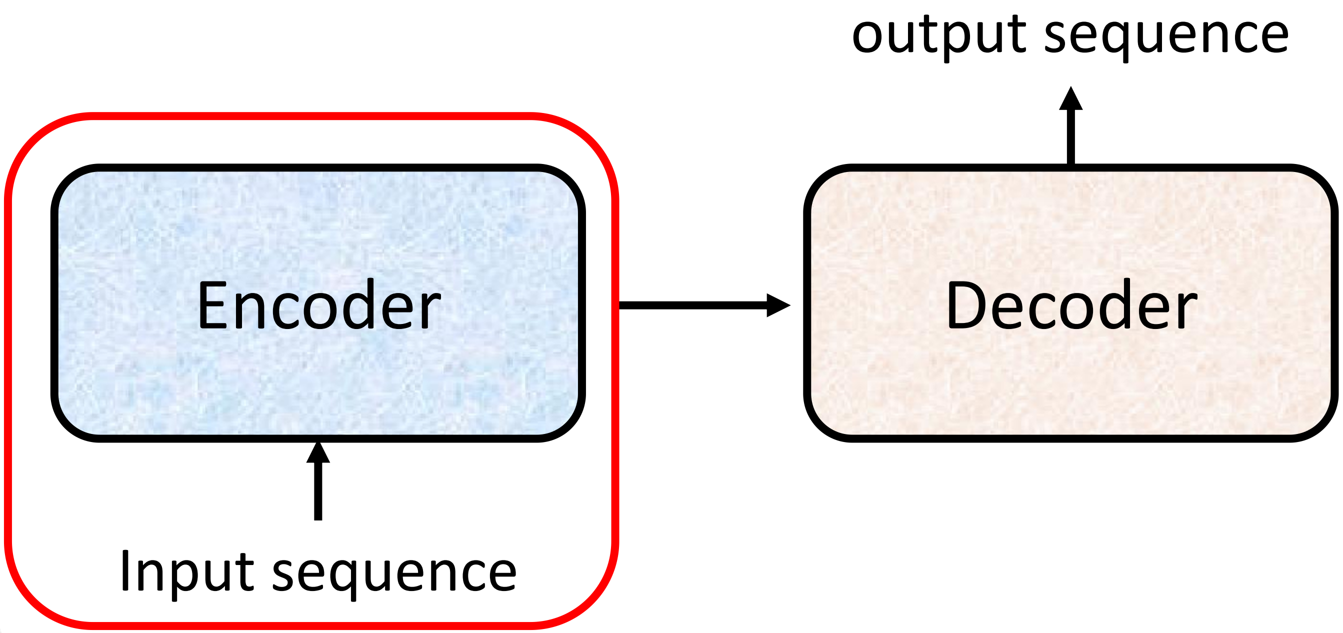
Sequence to Sequence Learning with Neural Networks

https://arxiv.org/abs/1409.3215

Transformer

https://arxiv.org/abs/1706.03762

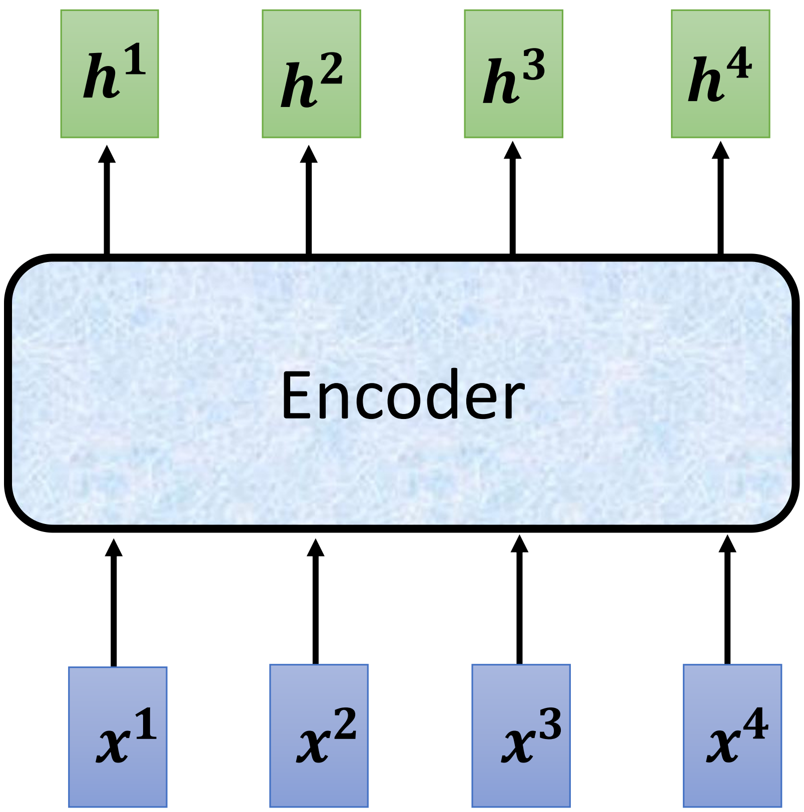
Encoder

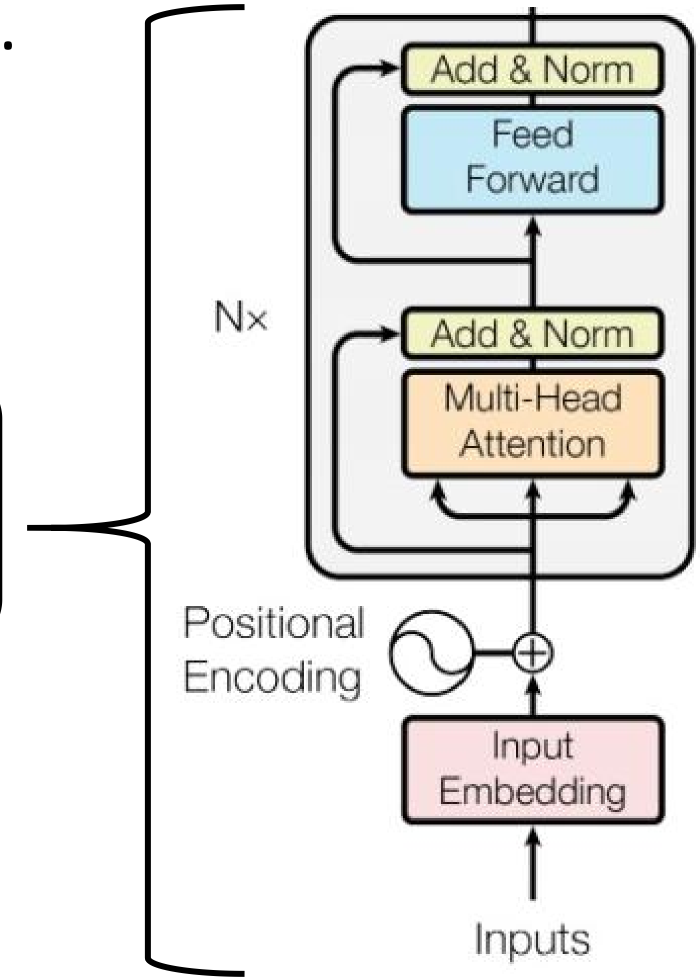


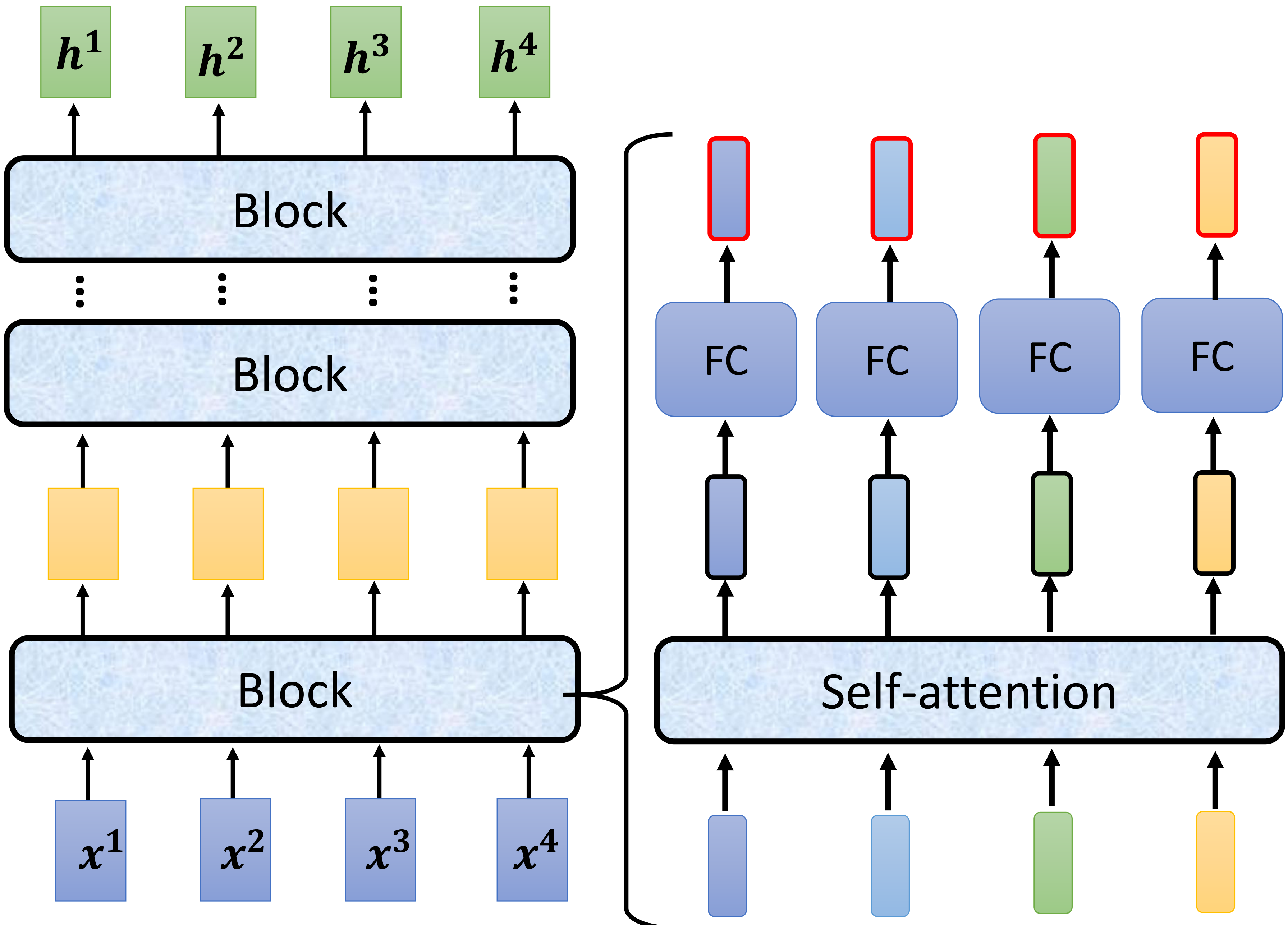
Encoder

Transformer's Encoder

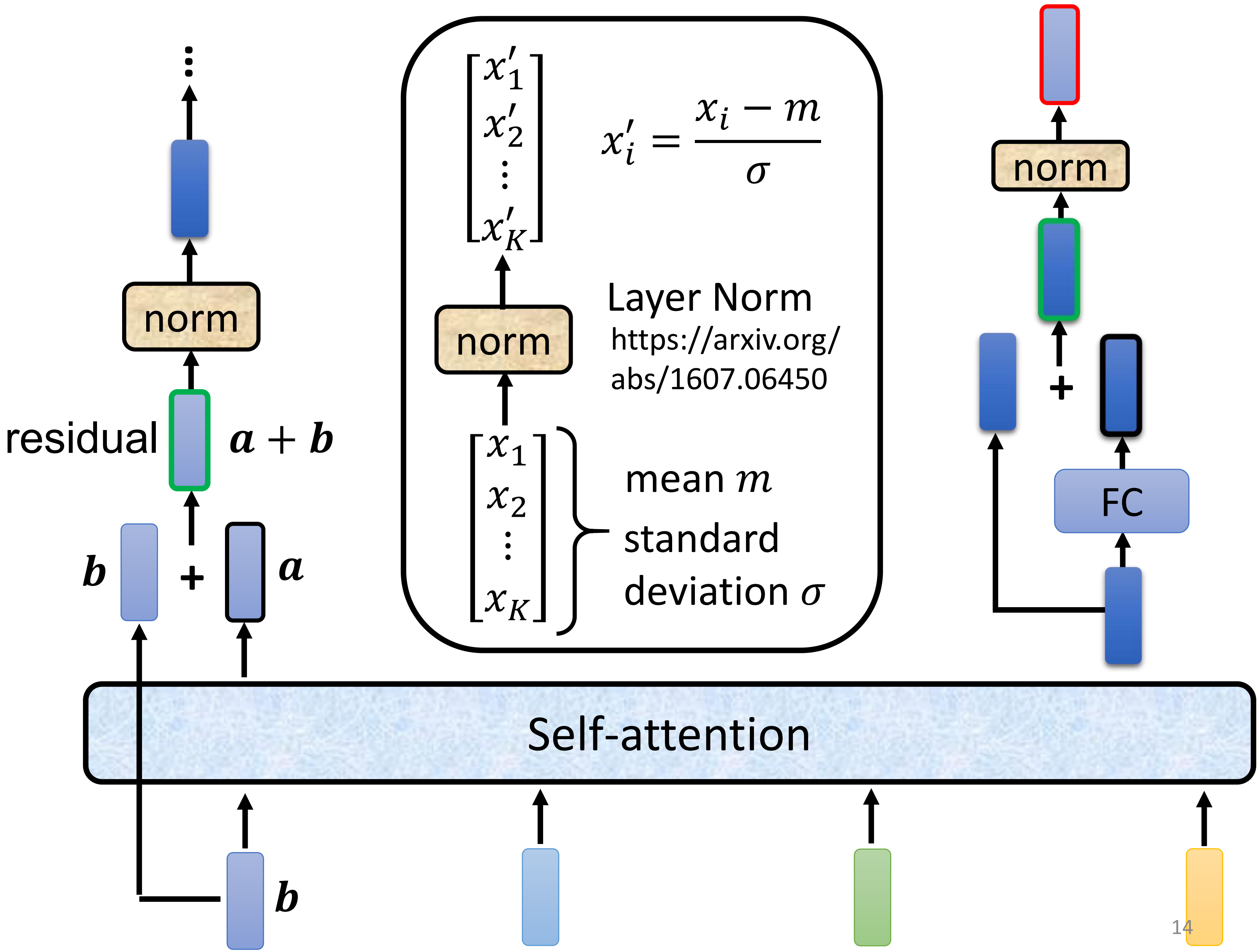
You can use **RNN** or **CNN**.





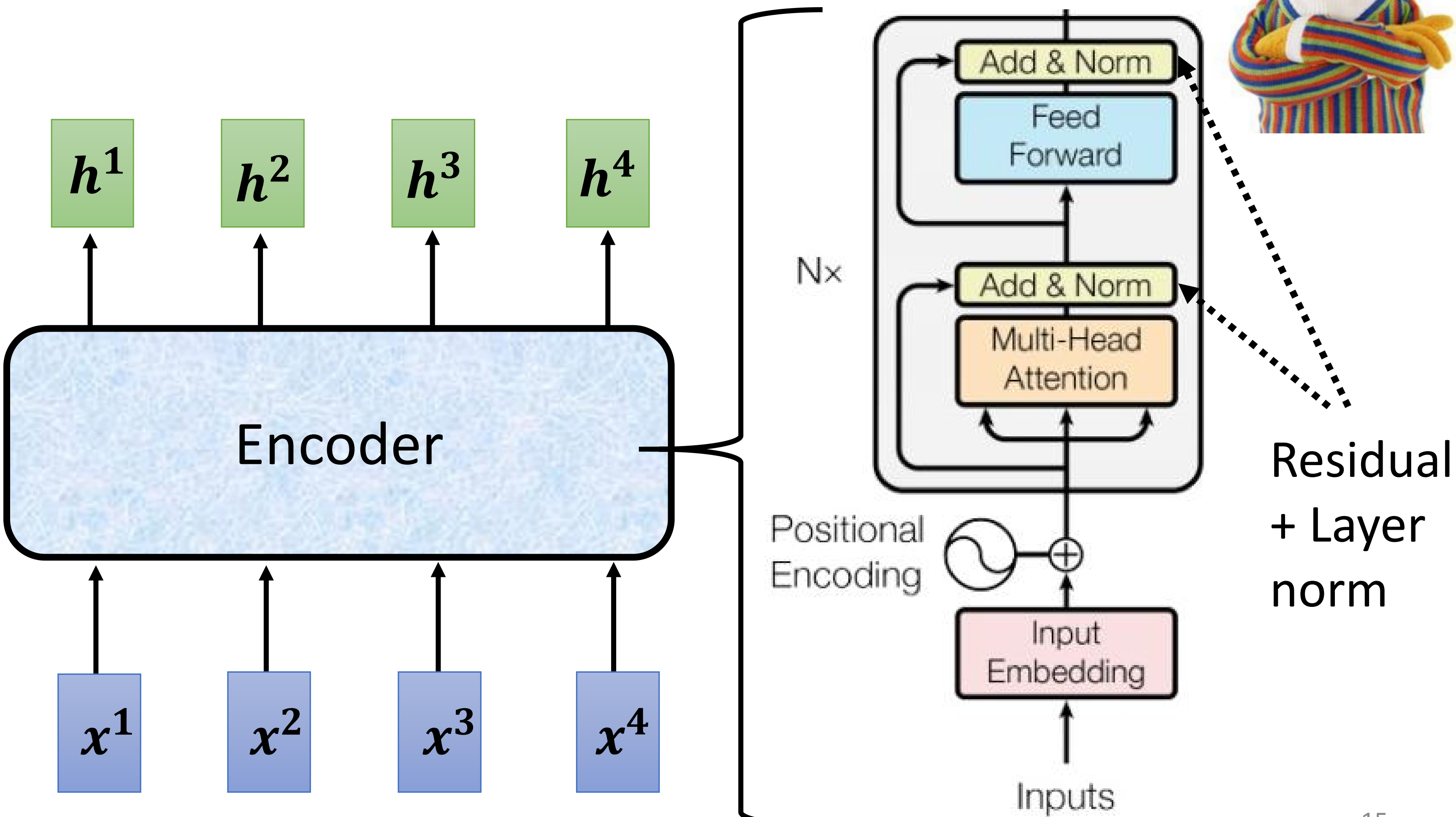


h^1
h^2
h^3
h^4
Block
Block
Block
x^1
x^2
x^3
x^4
FC
FC
FC
FC
Self-attention



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

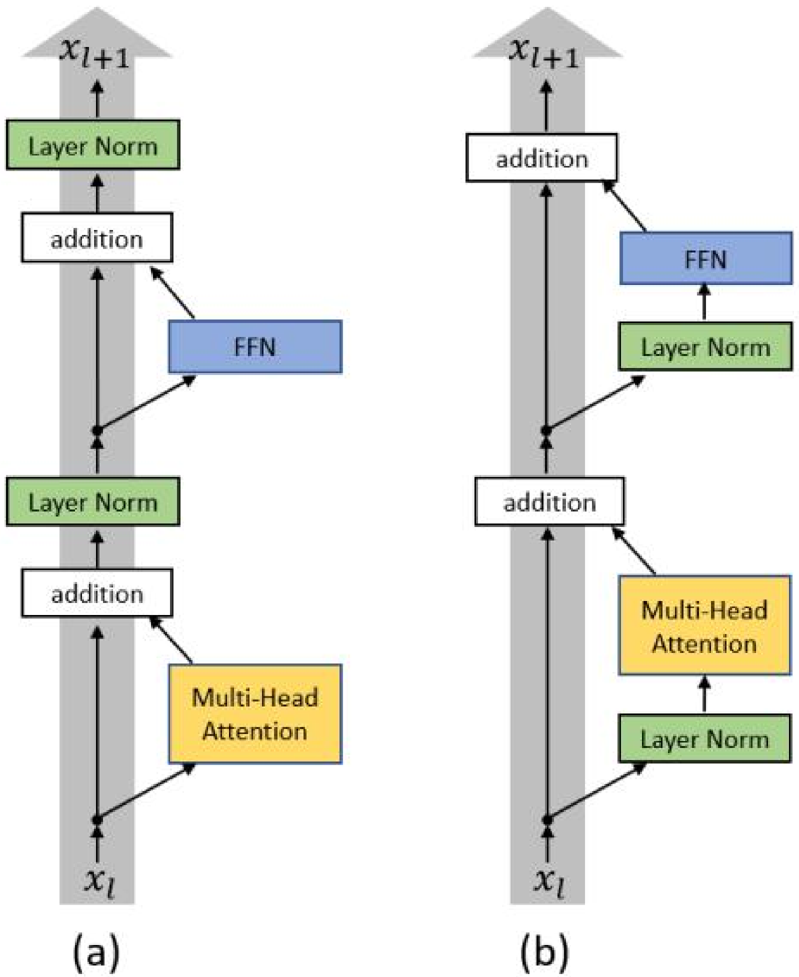
I use the **same** network architecture as **transformer encoder**.



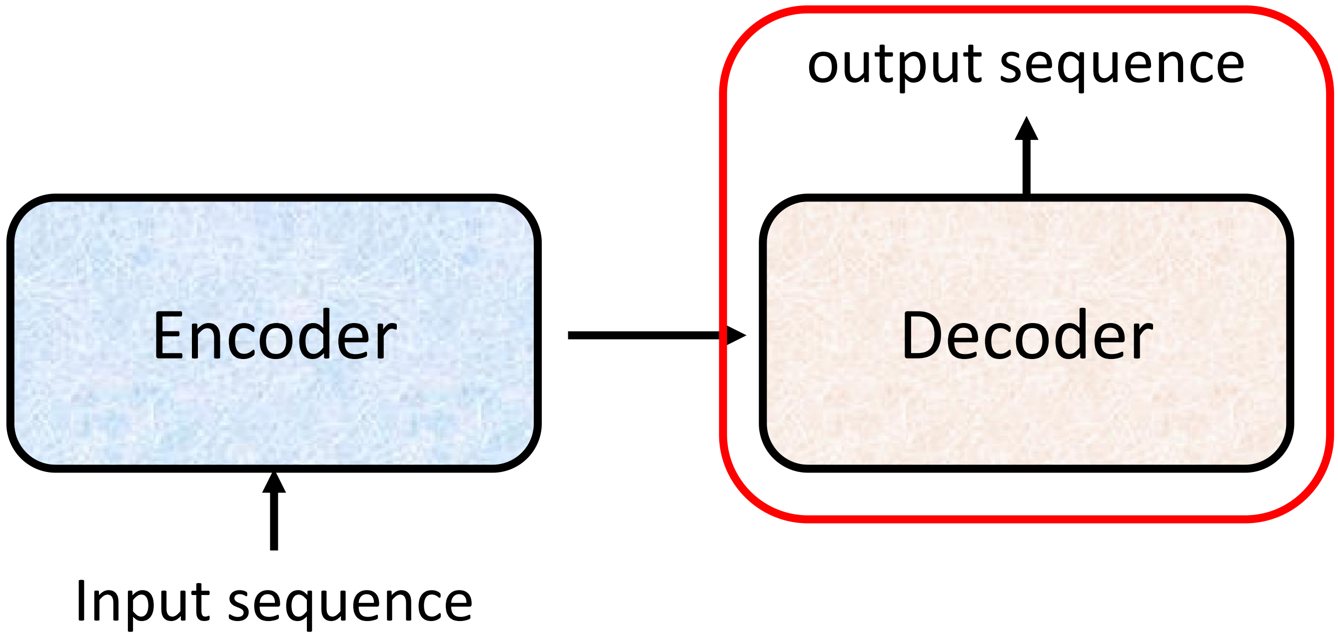
To learn more ……

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

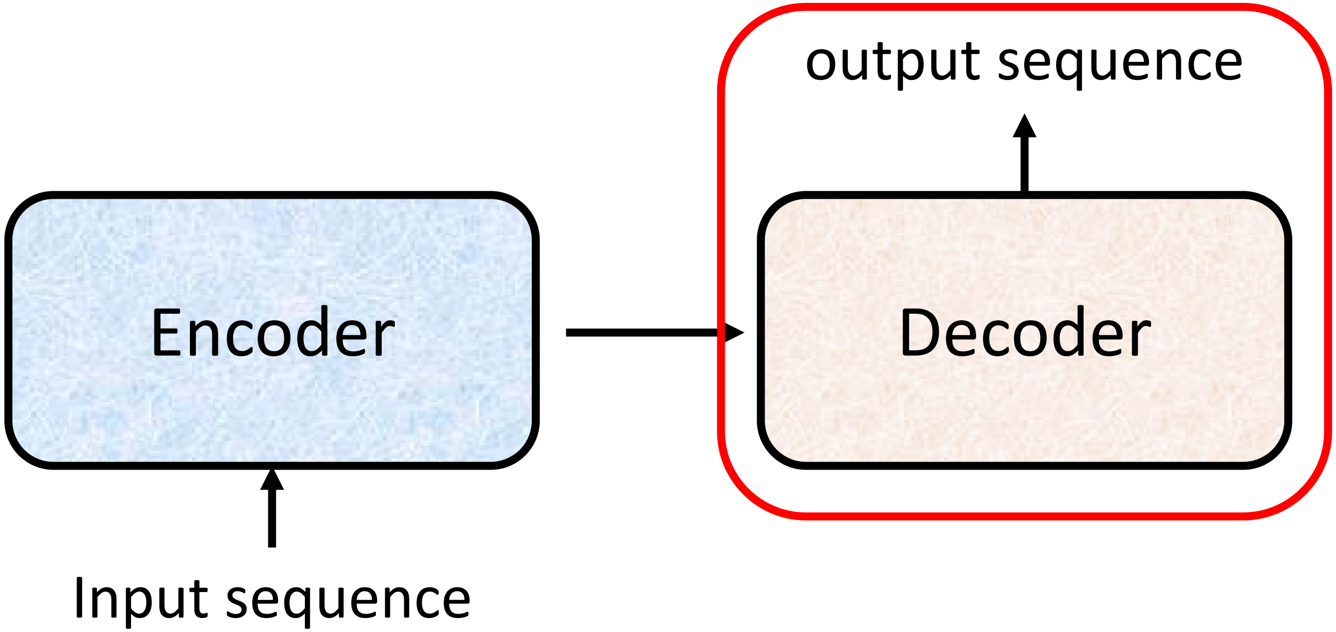
- PowerNorm: Rethinking Batch Normalization in Transformers
- https://arxiv.org/abs/2003.07845



Decoder

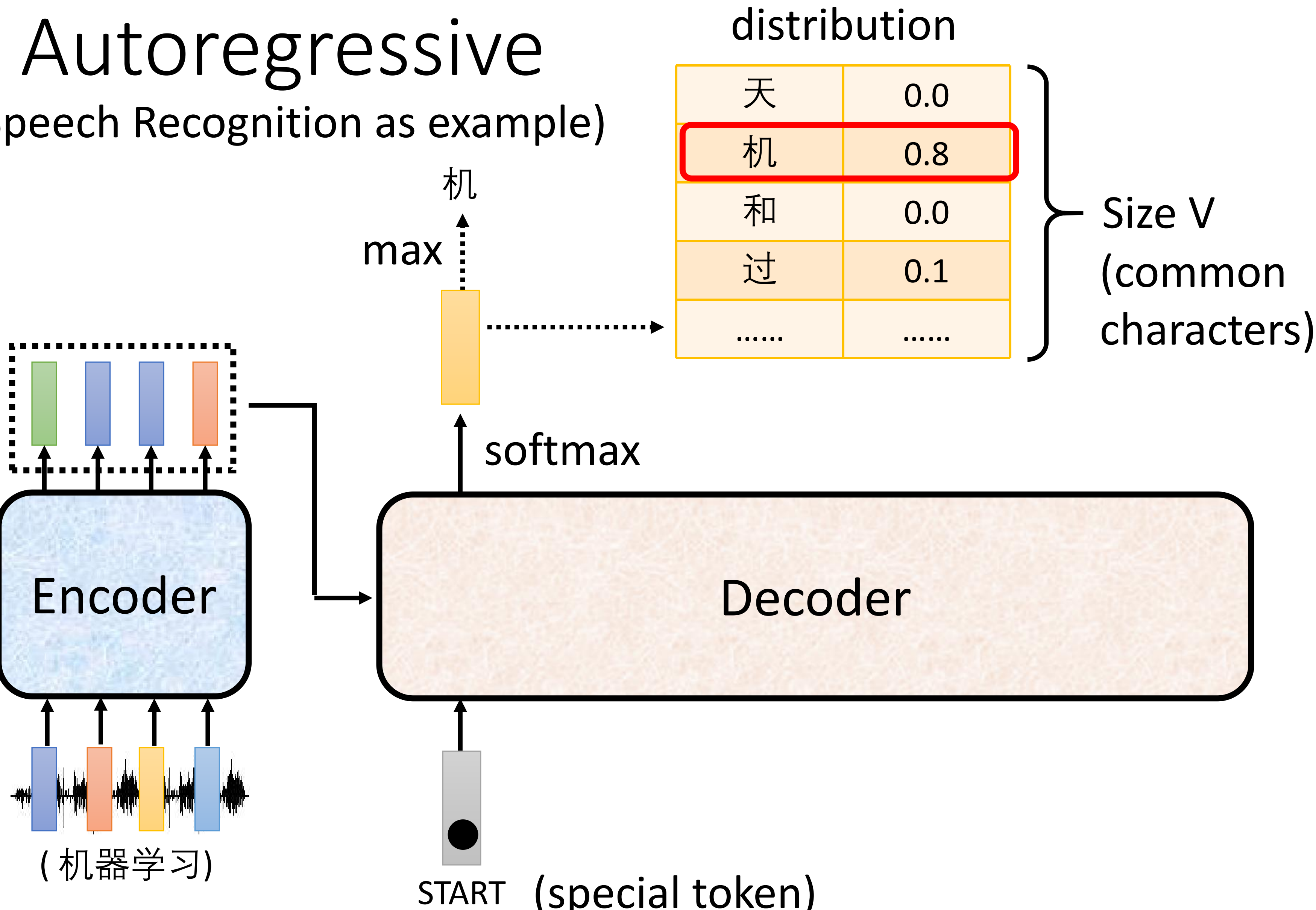


Decoder
– Autoregressive (AT)

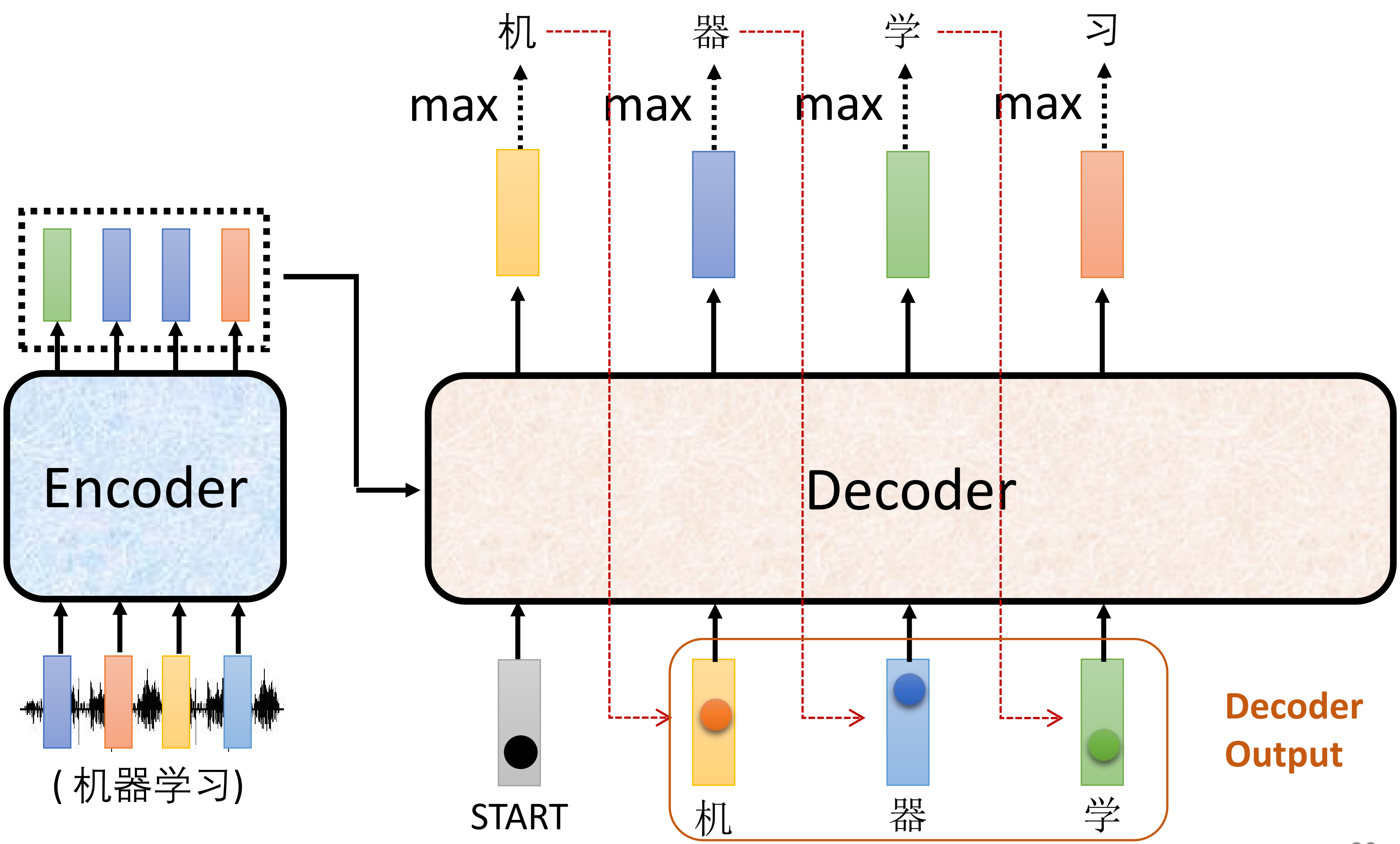


Autoregressive

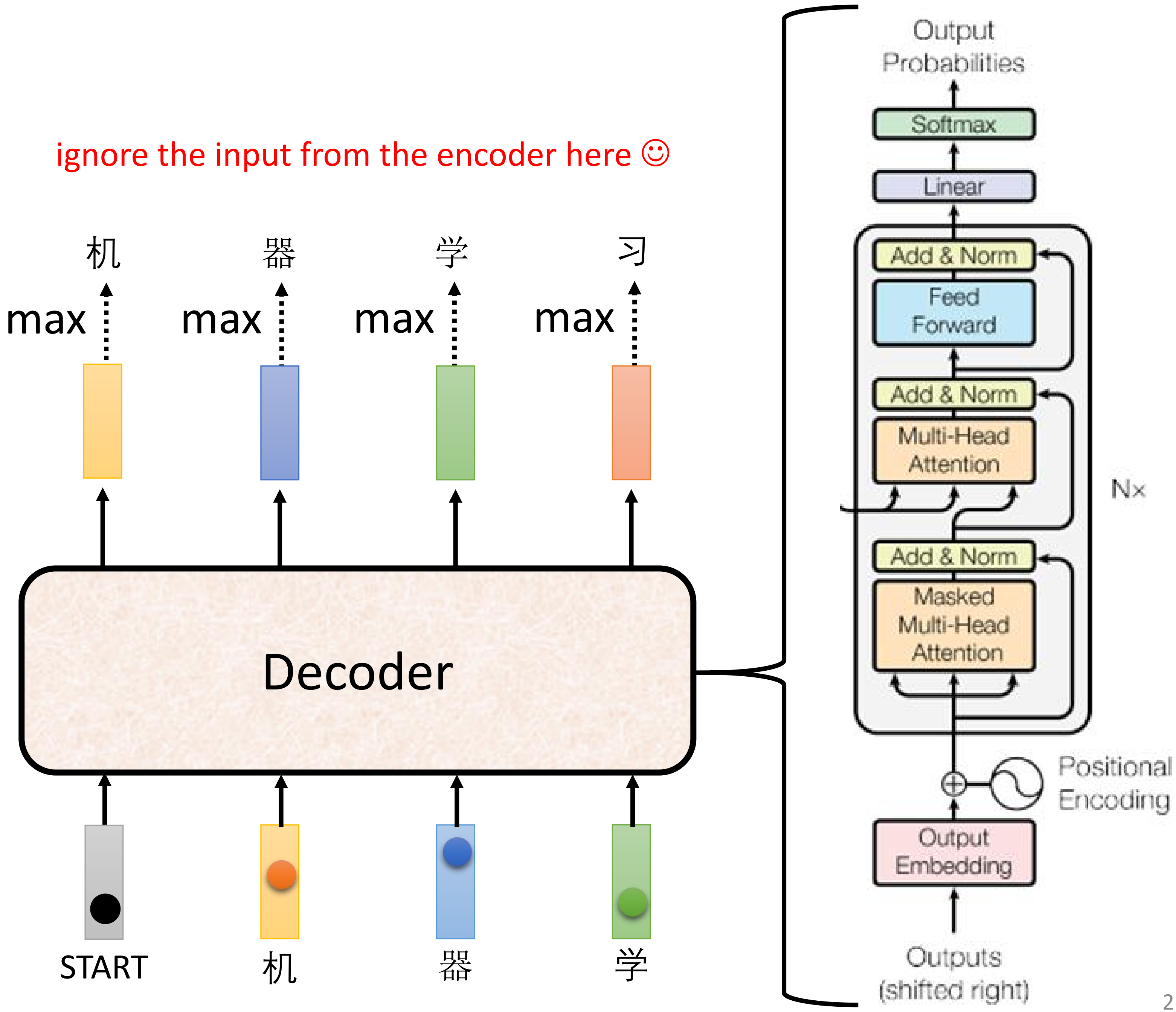
(Speech Recognition as example)

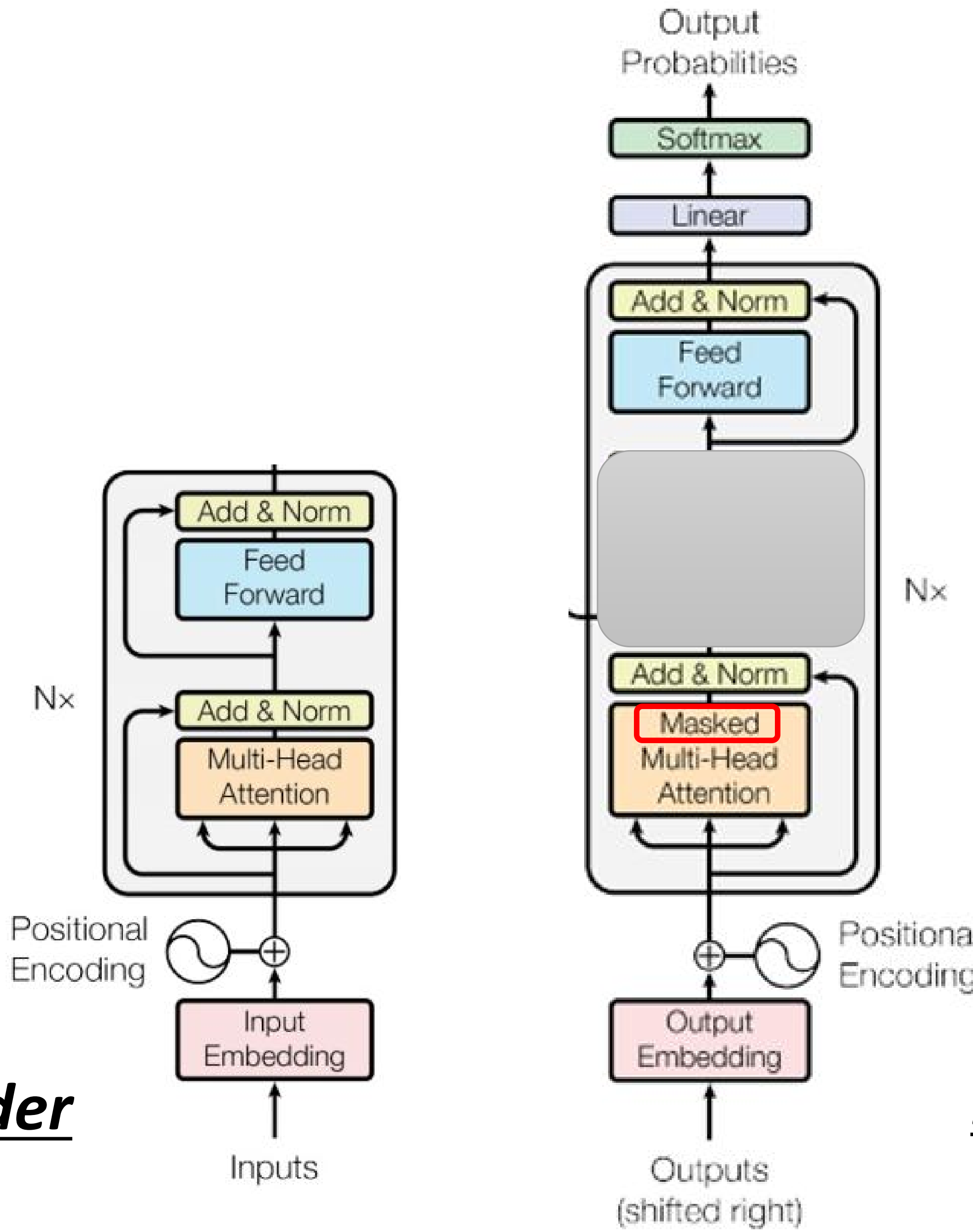


Autoregressive



ignore the input from the encoder here ☺

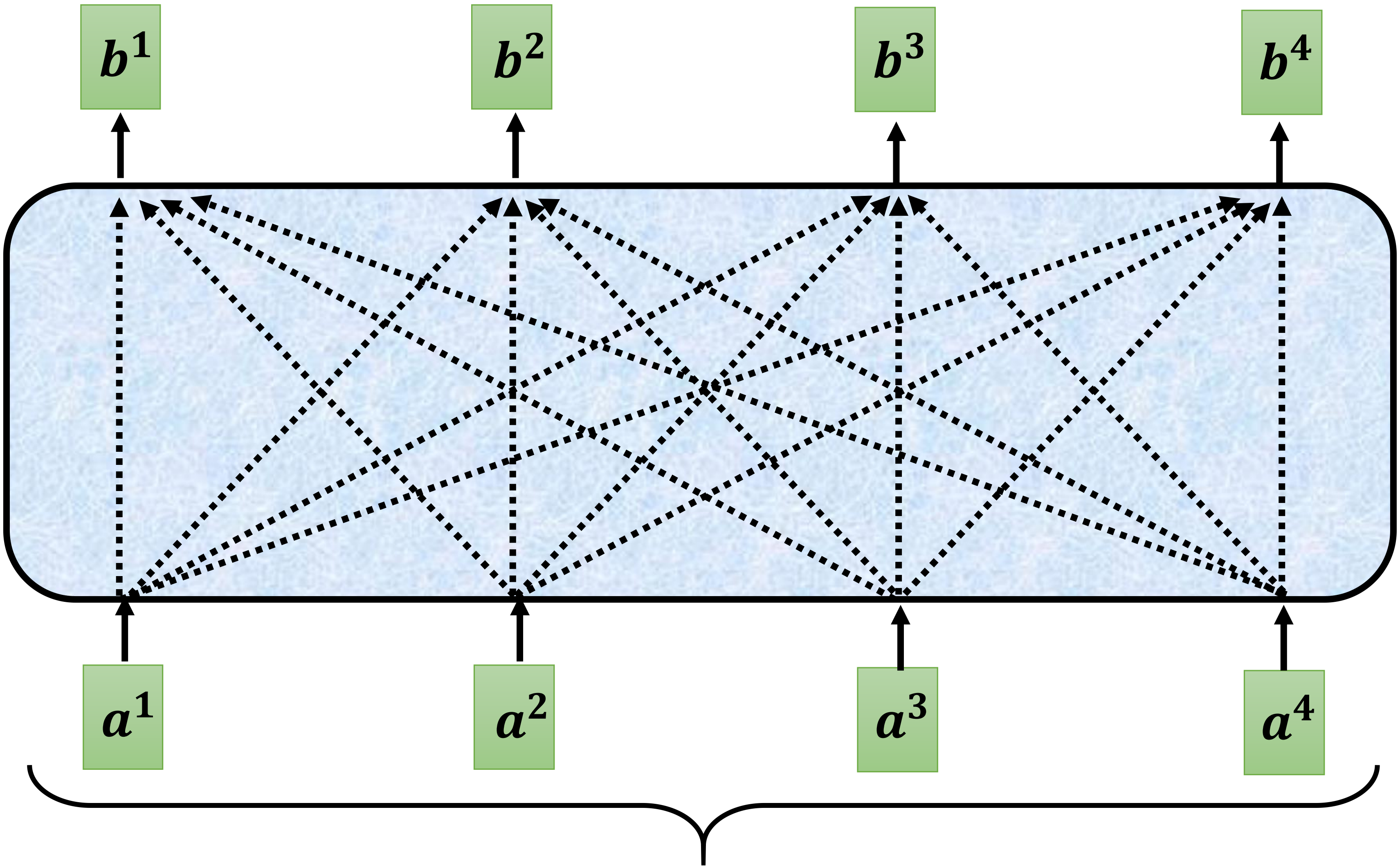




Encoder

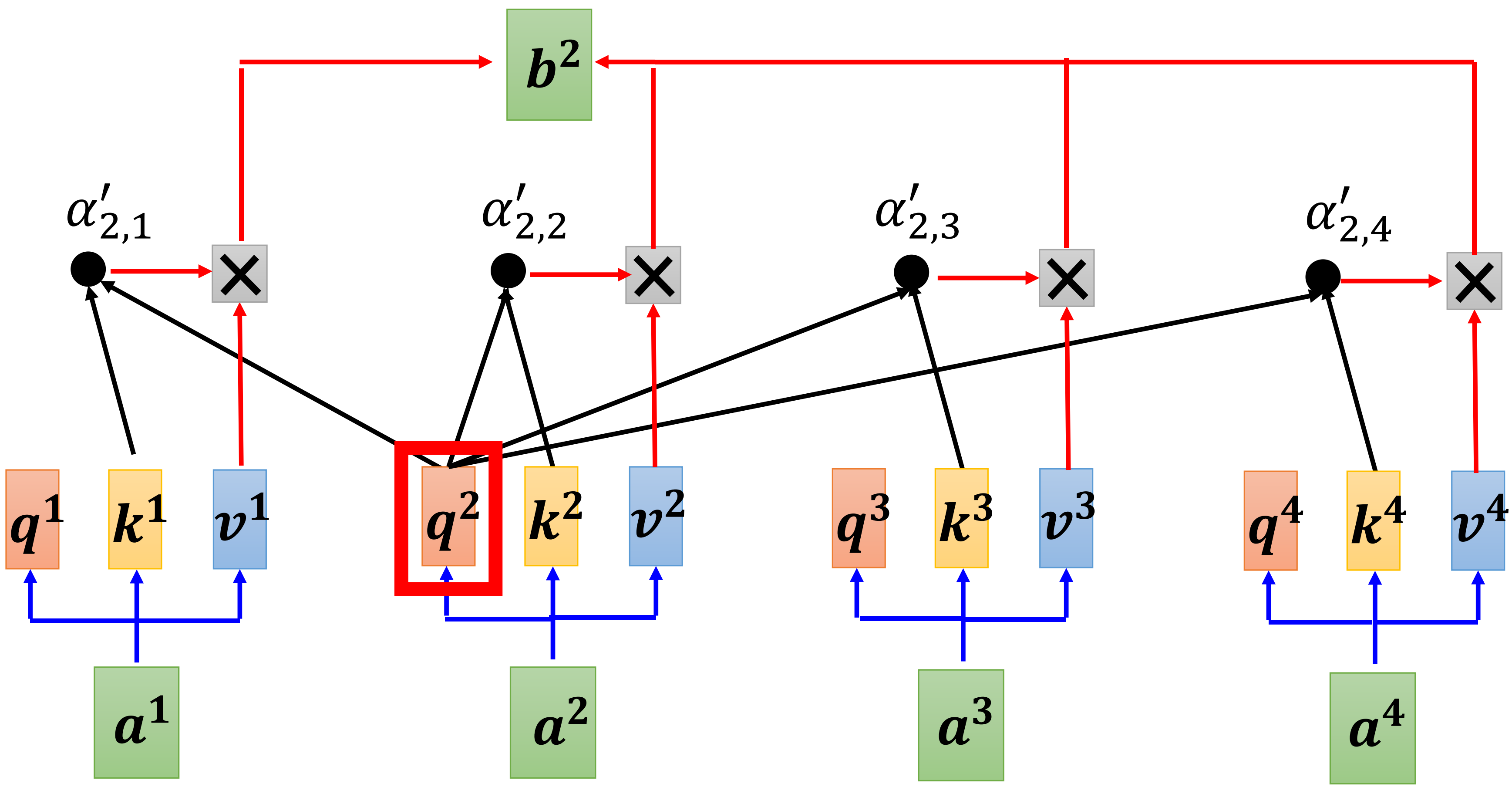
Decoder

Self-attention → *Masked Self-attention*



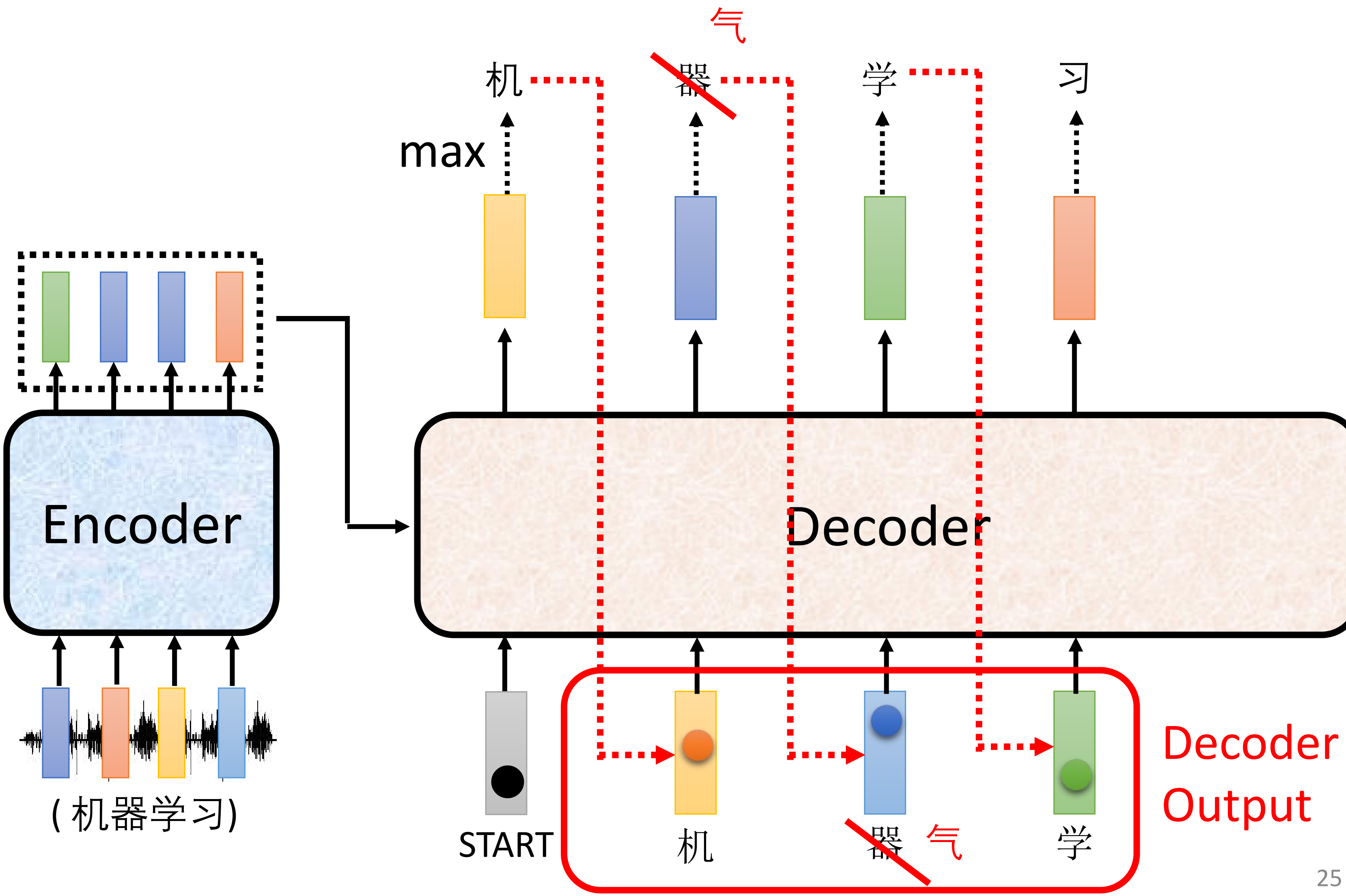
Can be either **input** or **a hidden layer**

Self-attention → Masked Self-attention



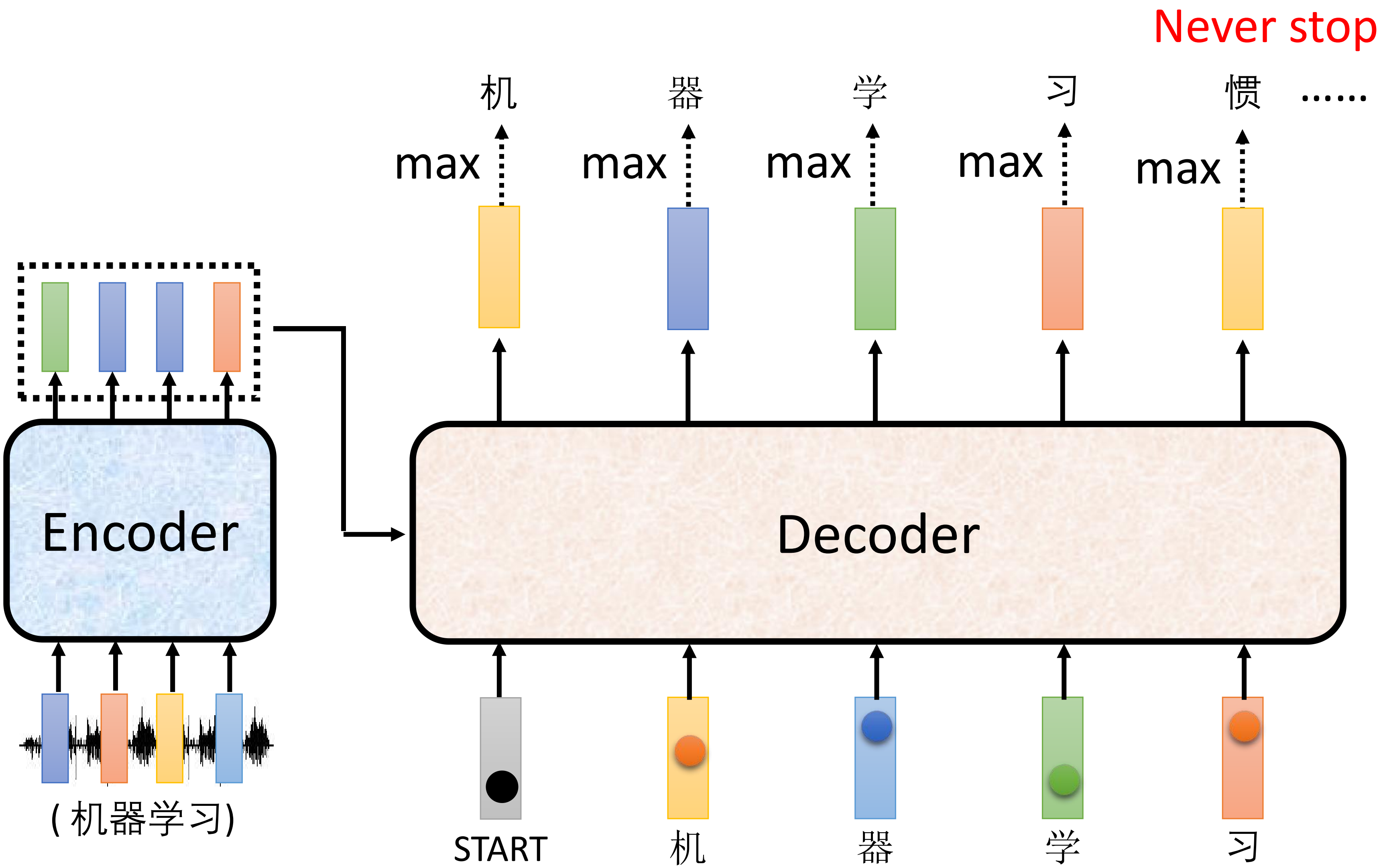
Why masked? Consider how does decoder work

Autoregressive

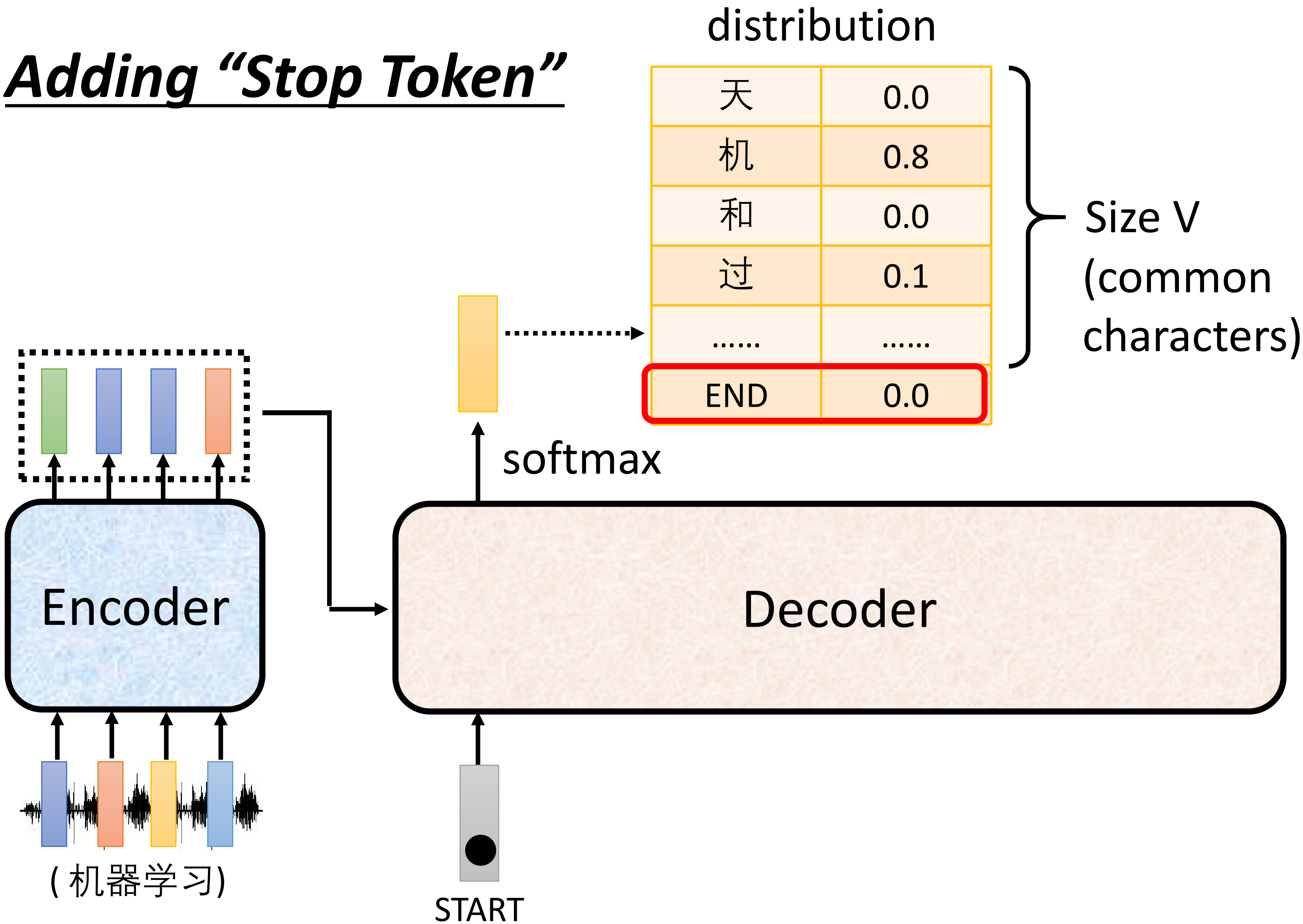


Autoregressive

We do not know the correct output length.

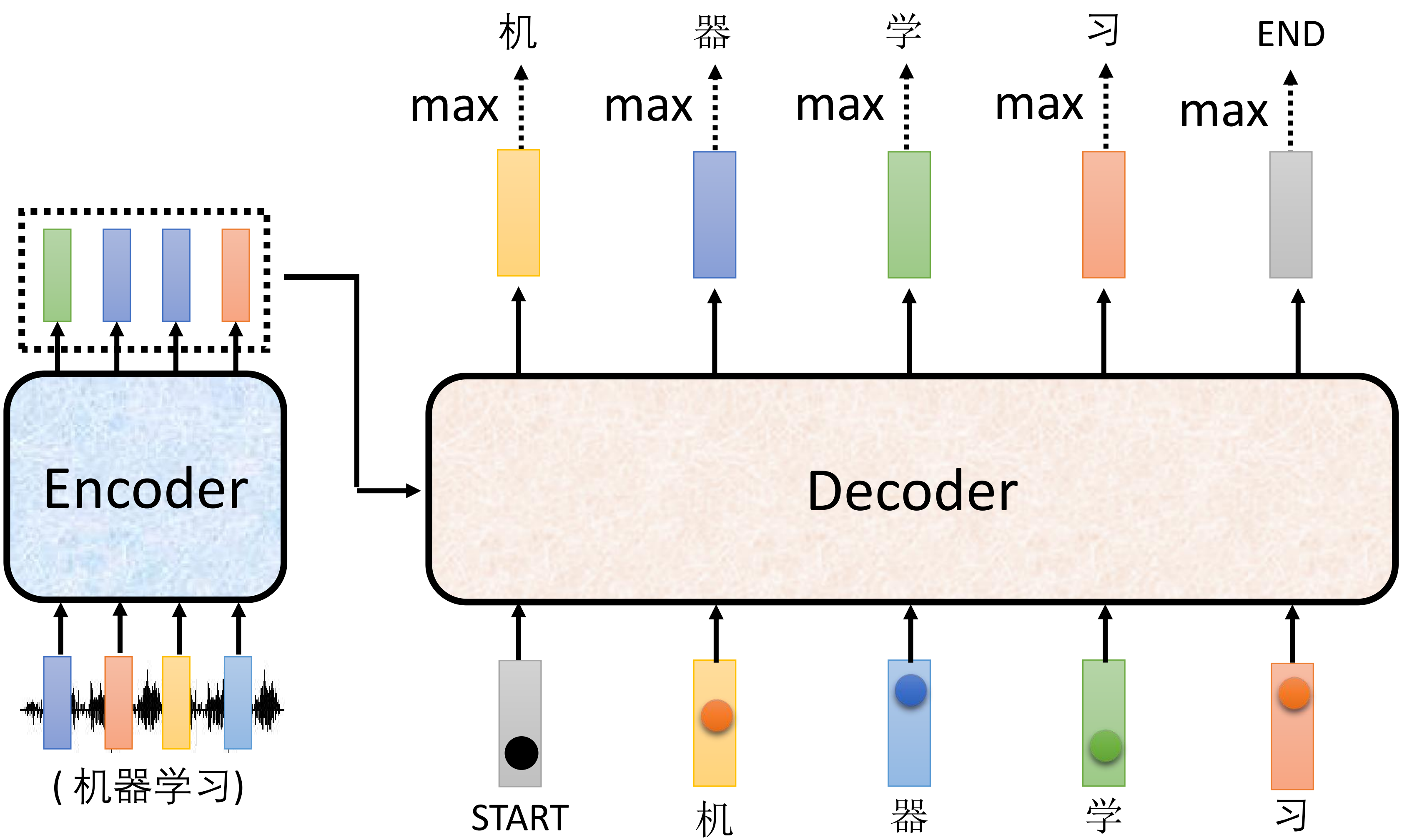


Adding "Stop Token"

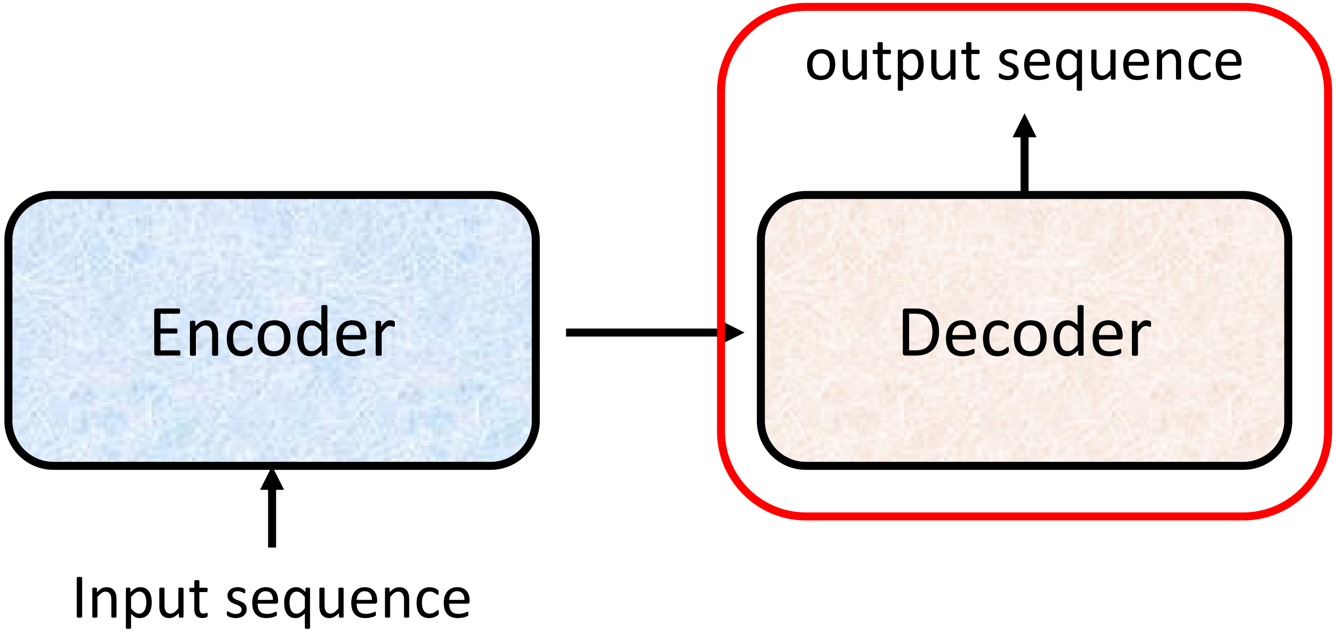


Autoregressive

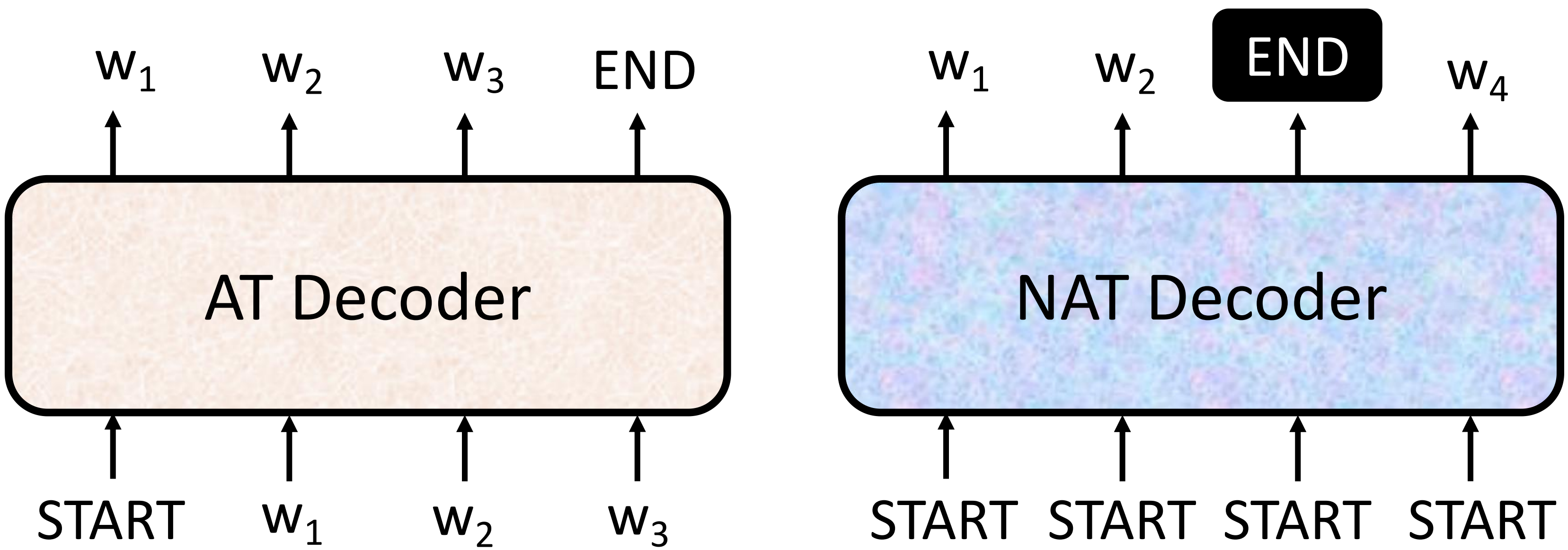
Stop at here!



Decoder
– Non-autoregressive (NAT)

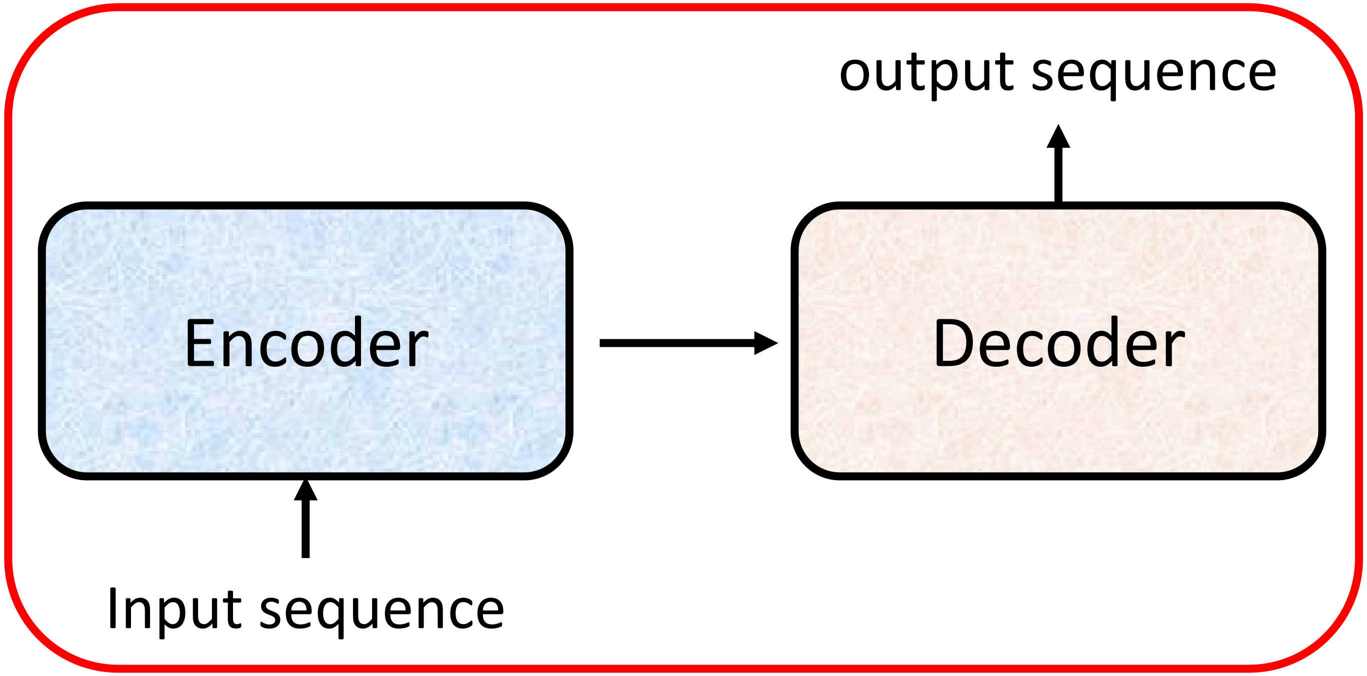


AT v.s. NAT

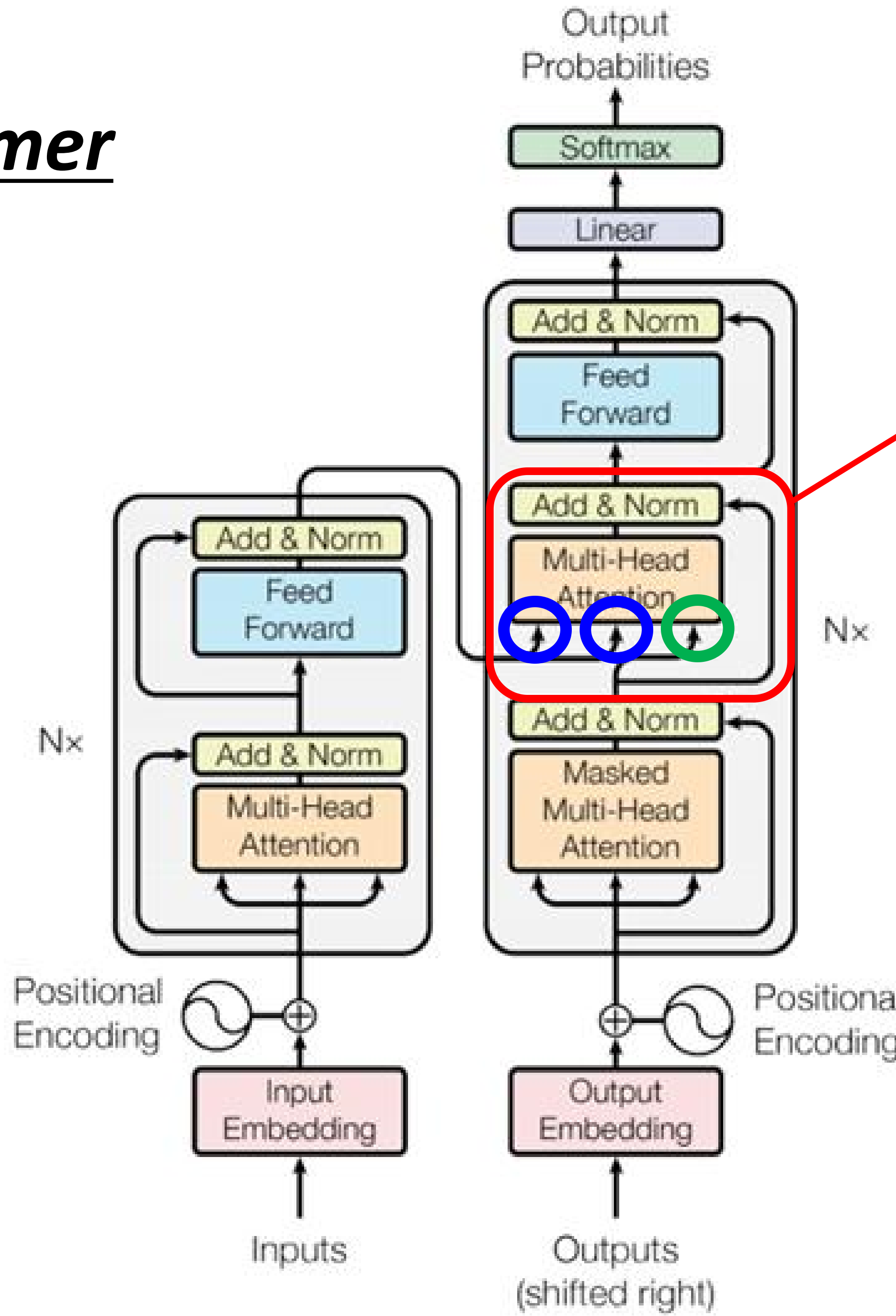


- How to decide the output length for NAT decoder?
  - Another predictor for output length
  - Output a very long sequence, ignore tokens after END
- Advantage: parallel, more stable generation (e.g., TTS)
- NAT is usually worse than AT (why? **Multi-modality**)

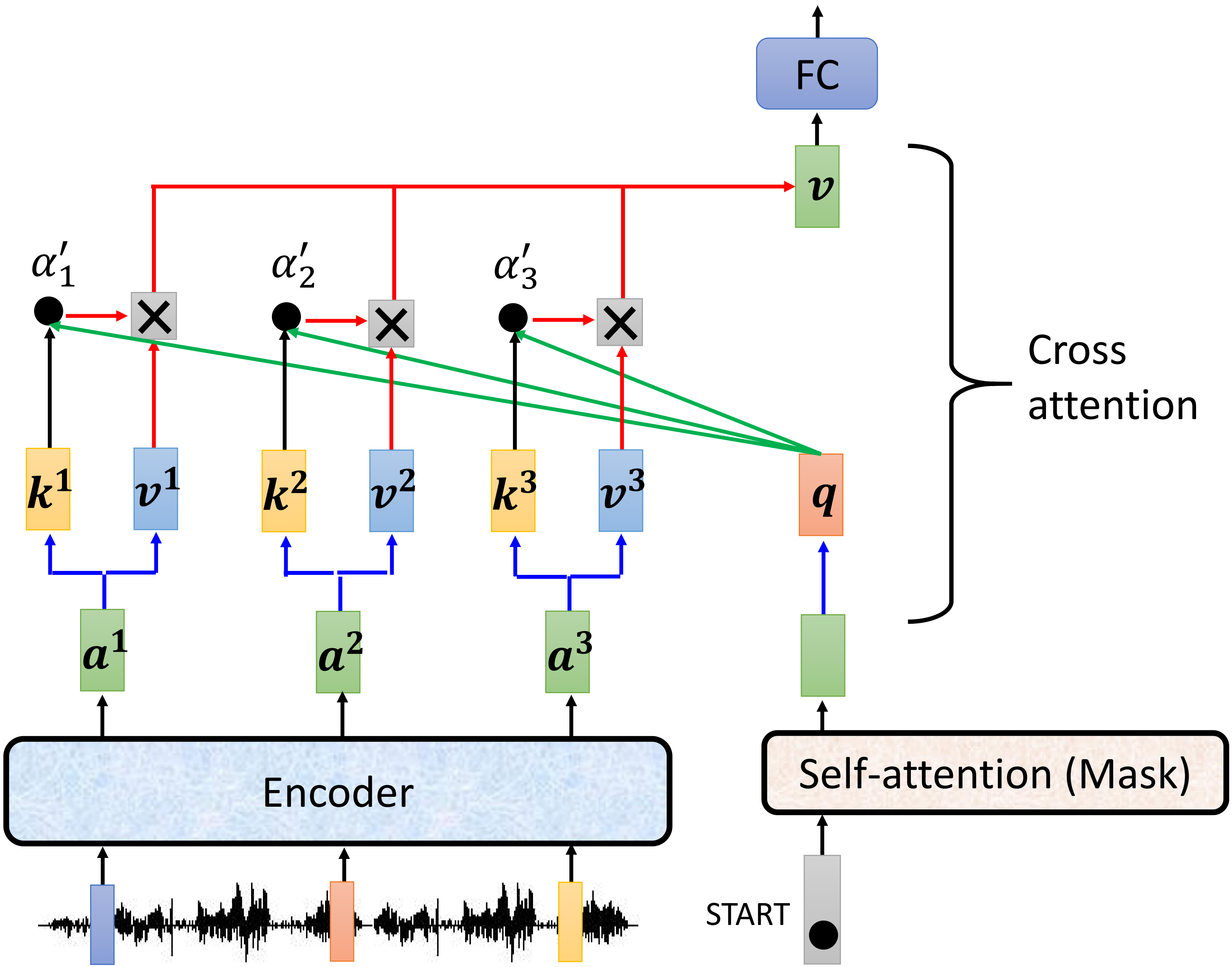
Encoder-Decoder



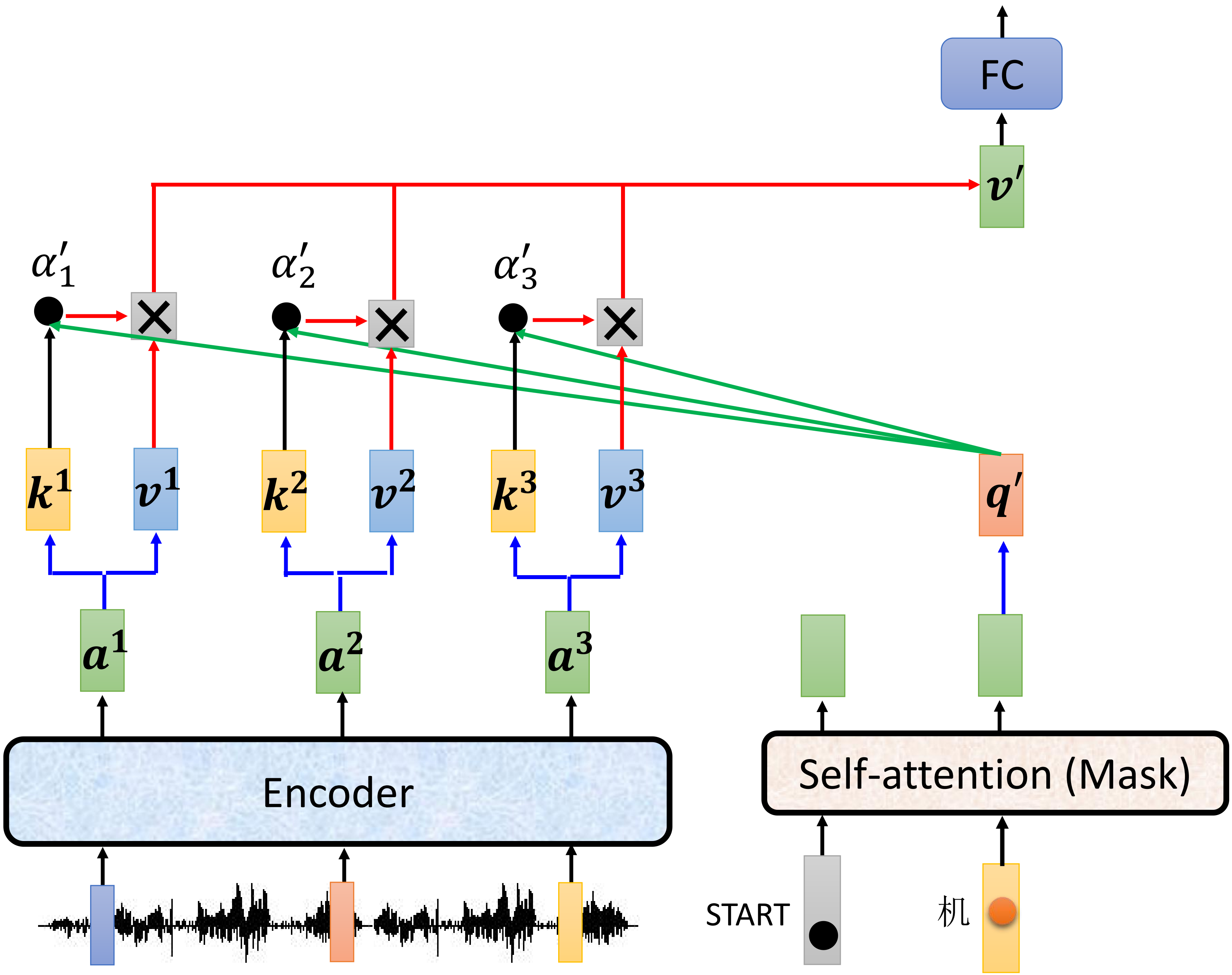
Transformer



Cross
attention



FC
v
α'_1
α'_2
α'_3
Cross attention
k^1
v^1
k^2
v^2
k^3
v^3
q
a^1
a^2
a^3
Encoder
Self-attention (Mask)
START

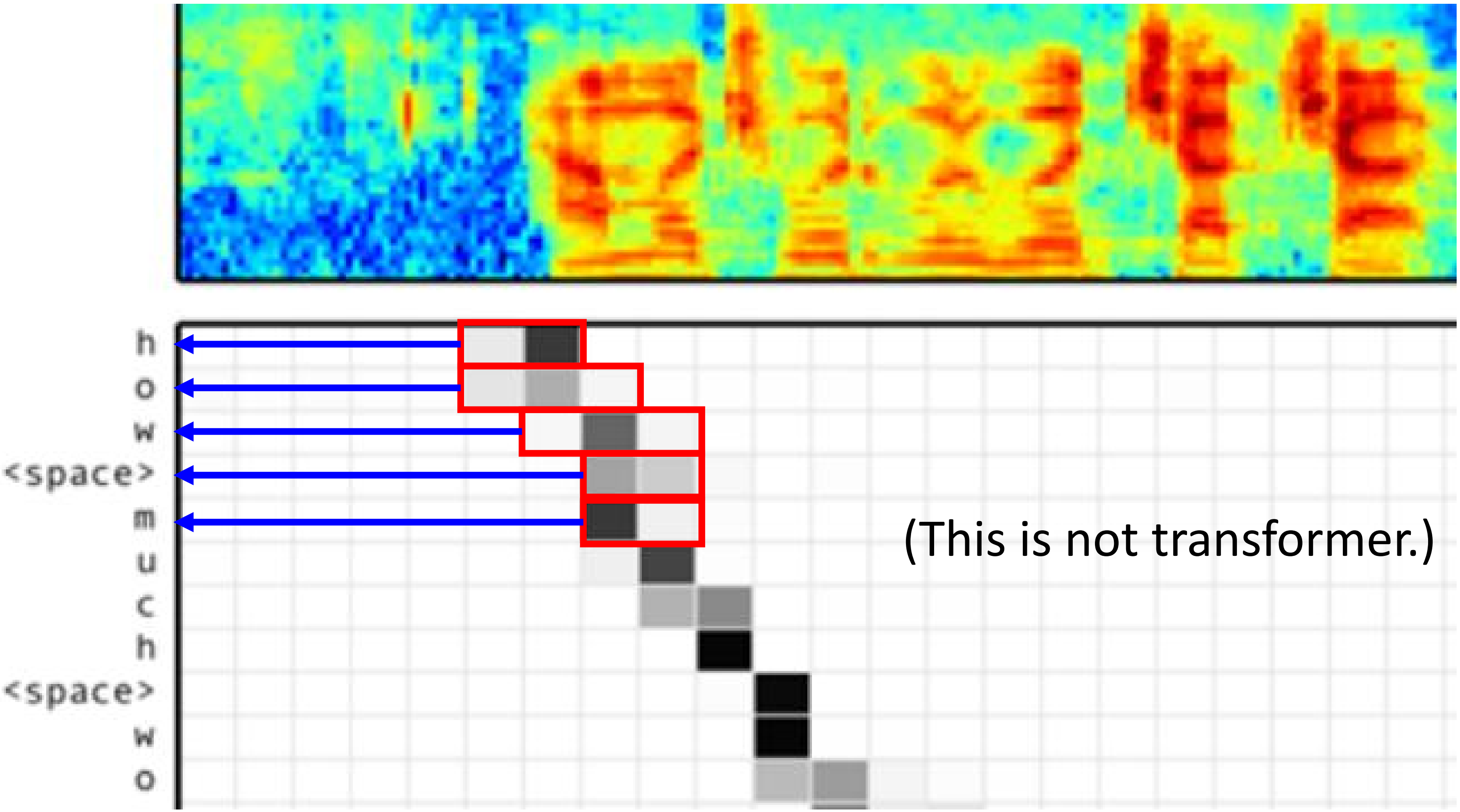


FC
v'
α'_1
α'_2
α'_3
k^1
v^1
k^2
v^2
k^3
v^3
q'
a^1
a^2
a^3
Encoder
Self-attention (Mask)
START
机

Cross Attention

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

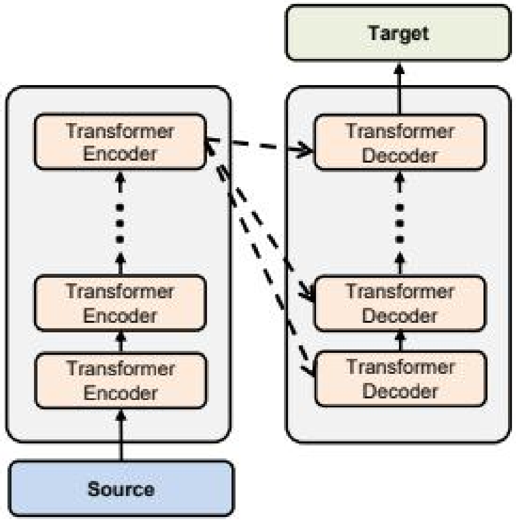
https://ieeexplore.ieee.org/document/7472621



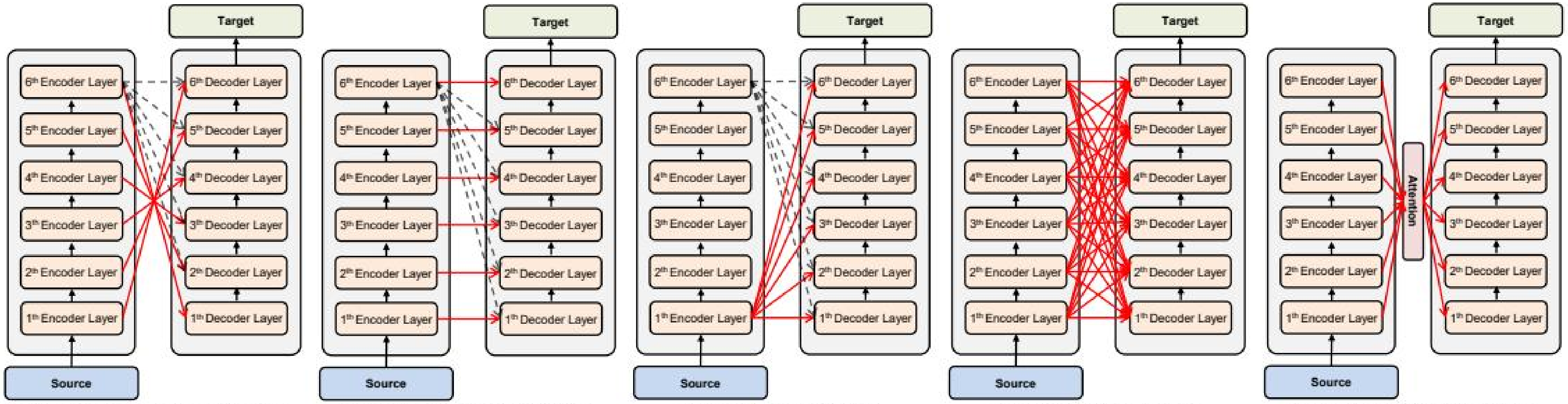
(This is not transformer.)

Cross Attention

Source of image:
https://arxiv.org/abs/2005.08081

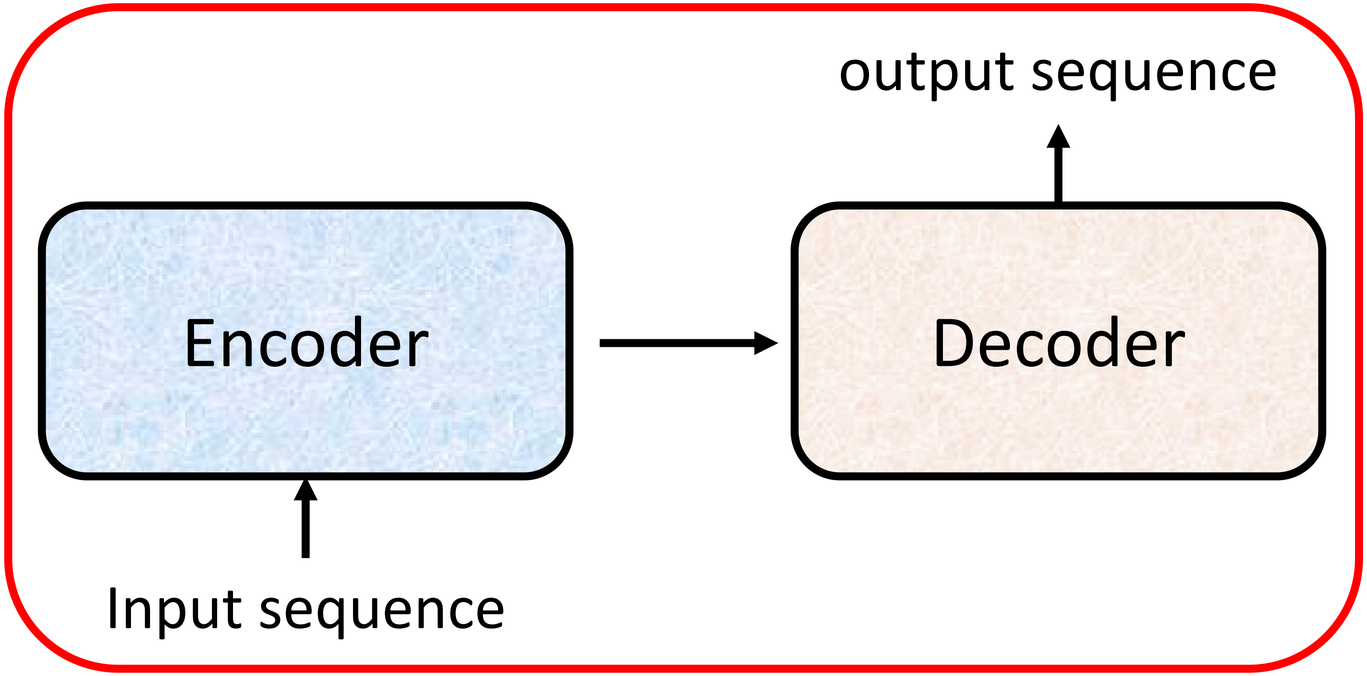


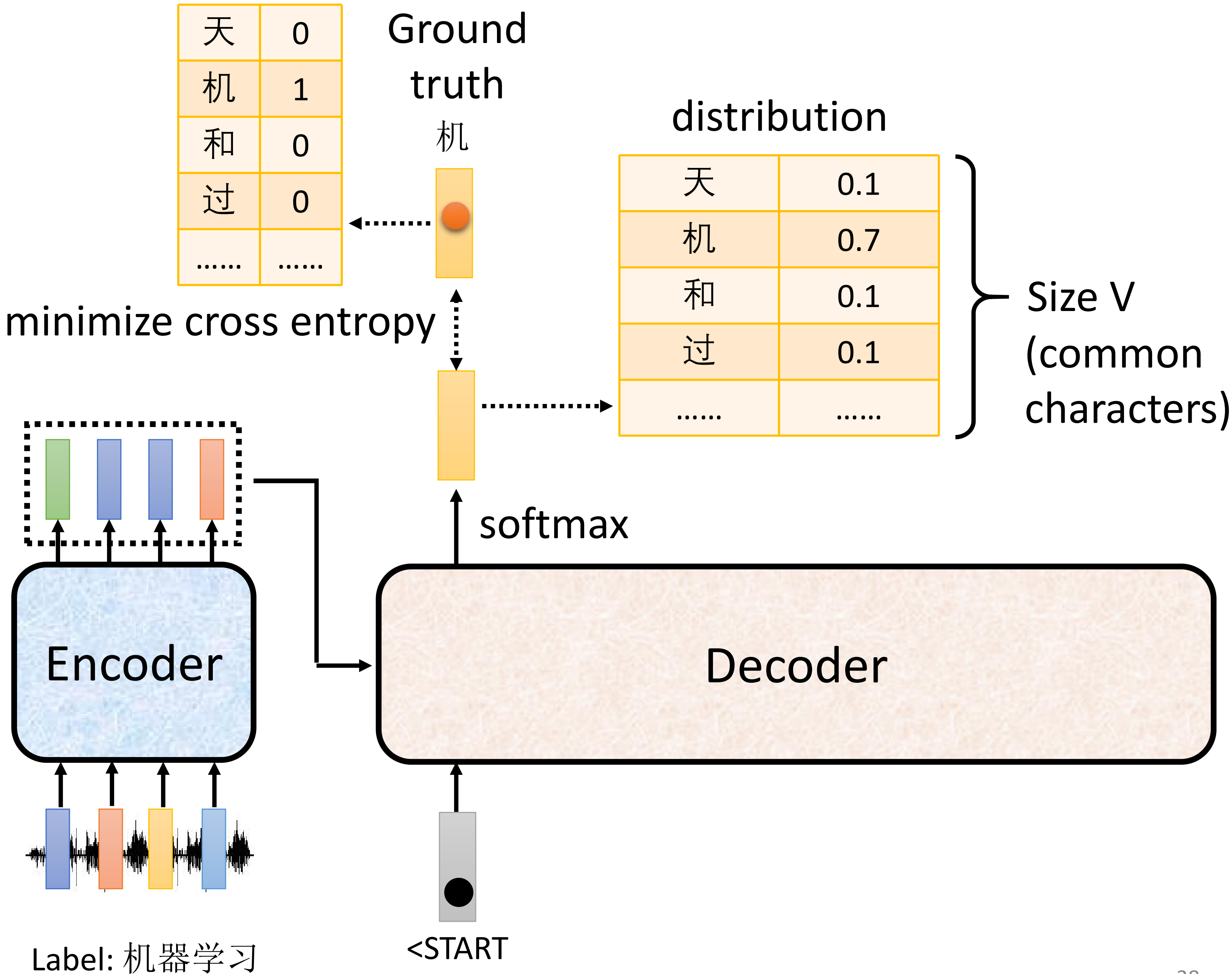
(a) Conventional Transformer



(a) Granularity Consistent Attention (b) Granularity Parallel Attention (c) Fine-Grained Attention (d) Full Matching Attention (e) Adaptive Matching Attention

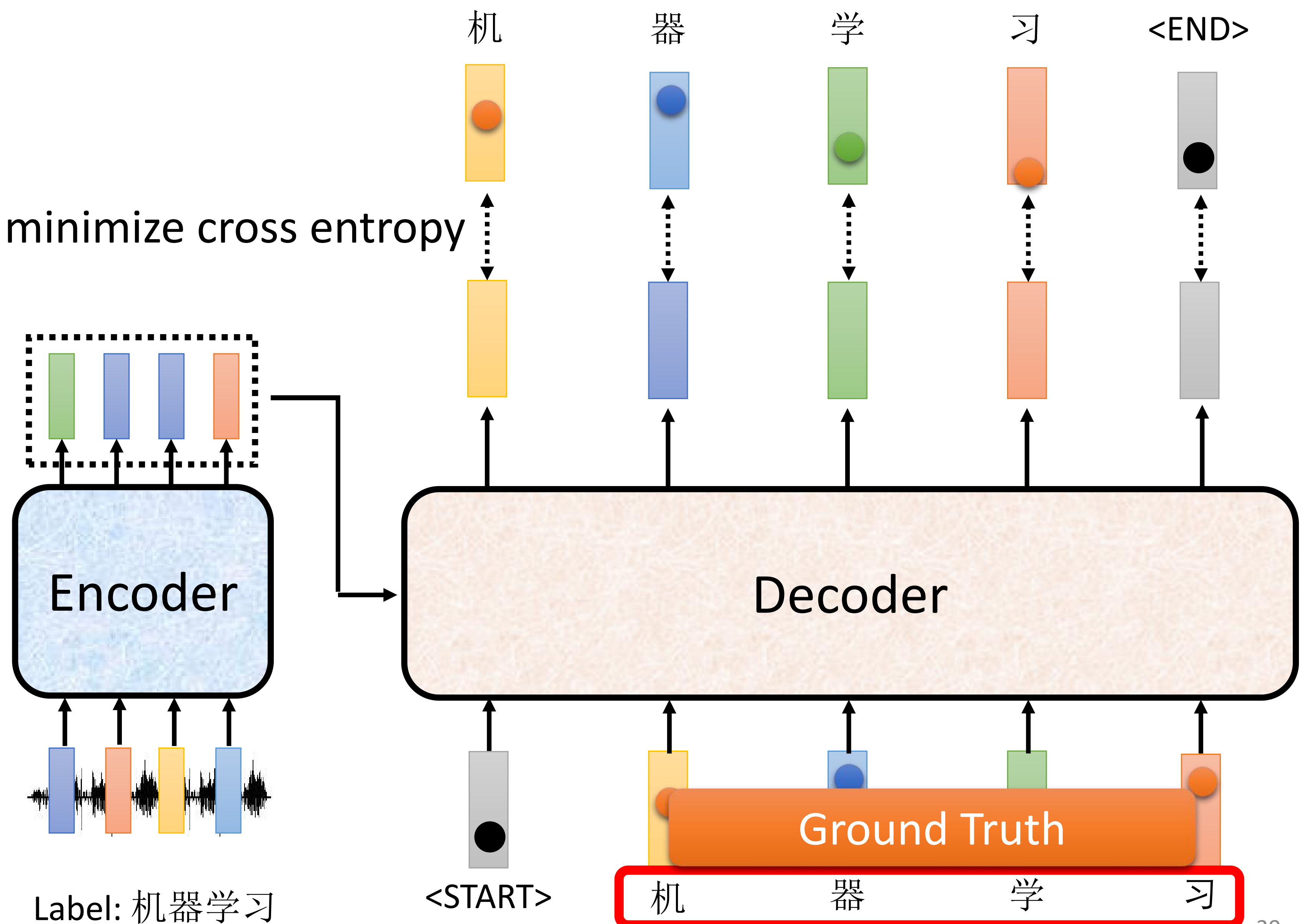
Training



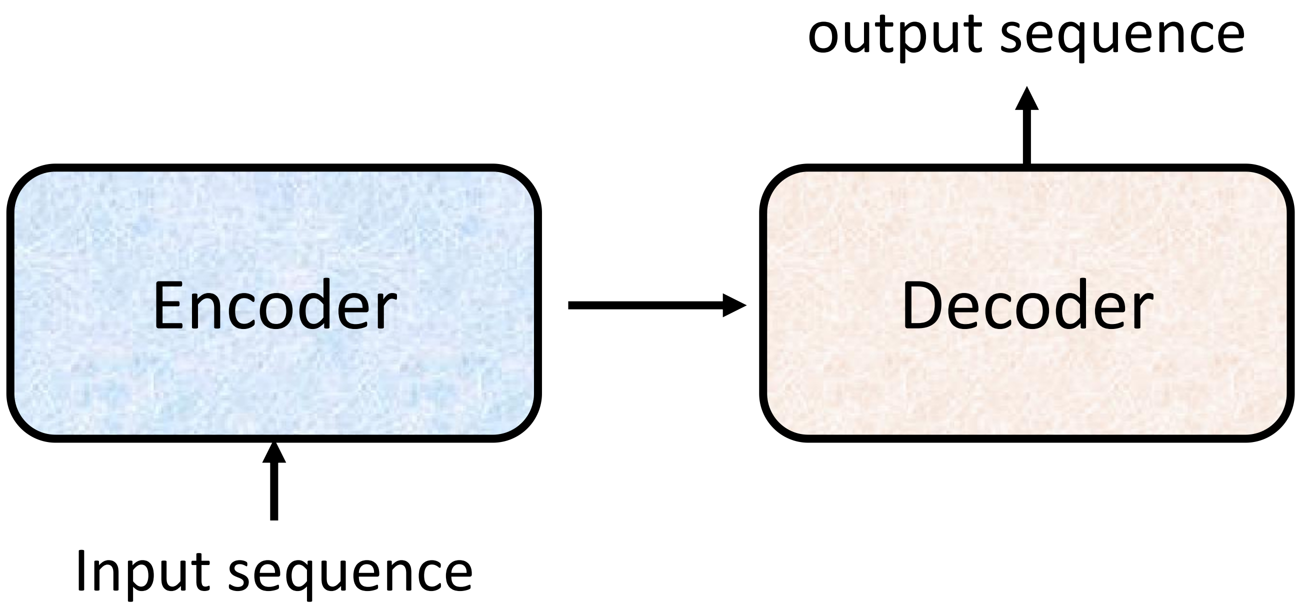


天 0
机 1
和 0
过 0
…… ……
Ground truth
机
minimize cross entropy
distribution
天 0.1
机 0.7
和 0.1
过 0.1
…… ……
Size V (common characters)
softmax
Encoder
Decoder
Label: 机器学习
<START

Teacher Forcing: using the ground truth as input.

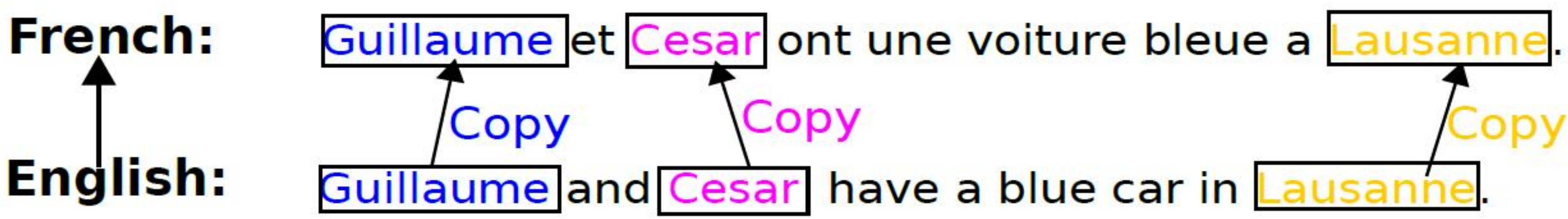


Tips



Copy Mechanism

Machine Translation



Chat-bot

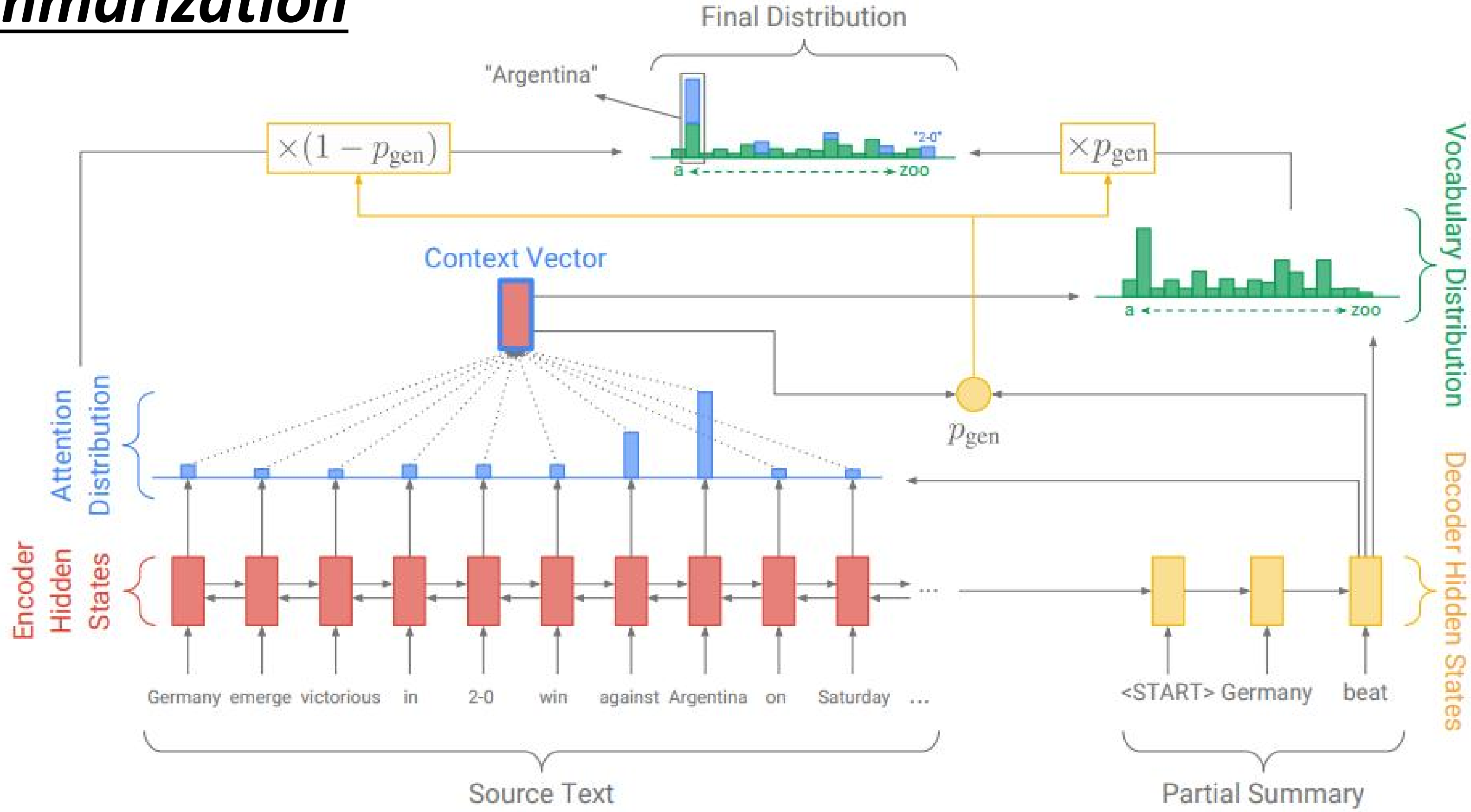
User: 李雷你好，我是韩梅梅

Machine:韩梅梅你好，很高兴认识你

Copy Mechanism

https://arxiv.org/abs/1704.04368

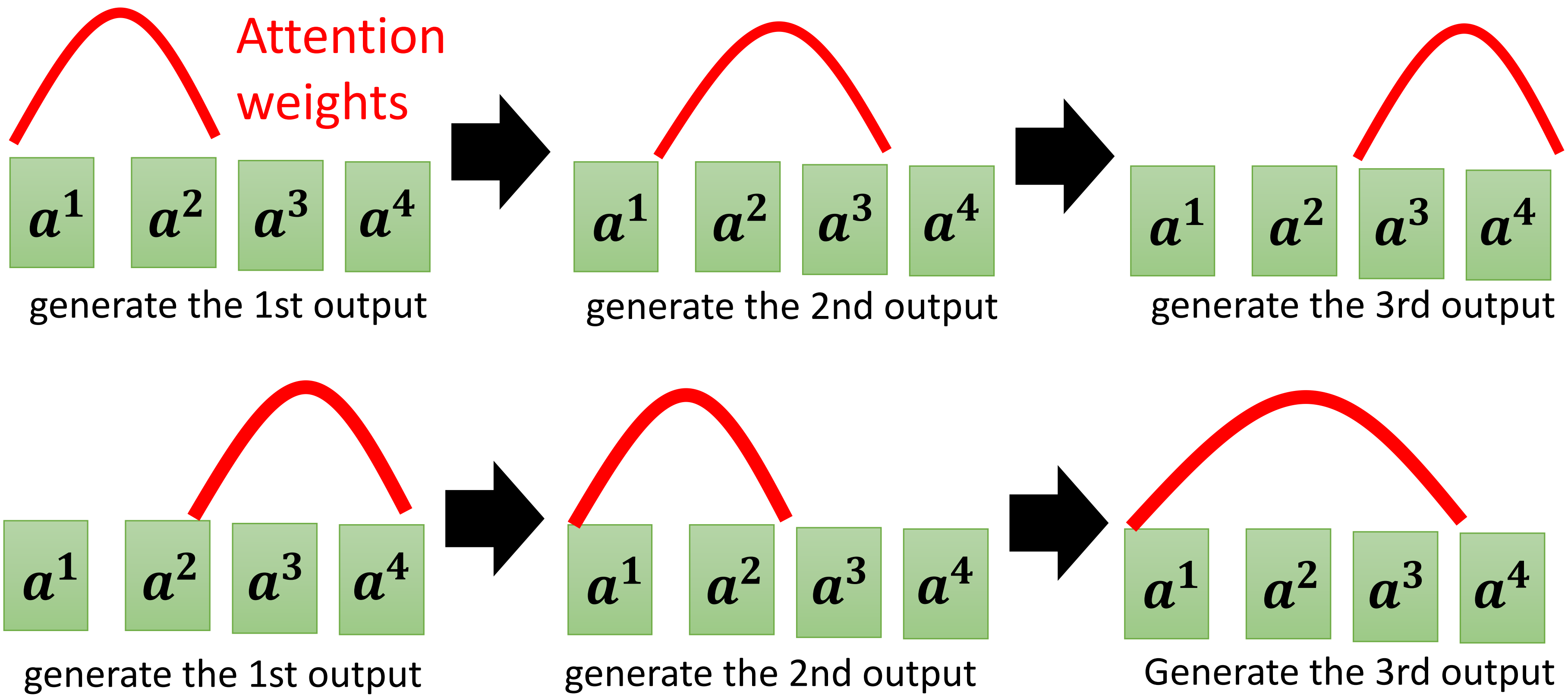
Summarization



Guided Attention

Monotonic Attention
Location-aware attention

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



Something wrong!

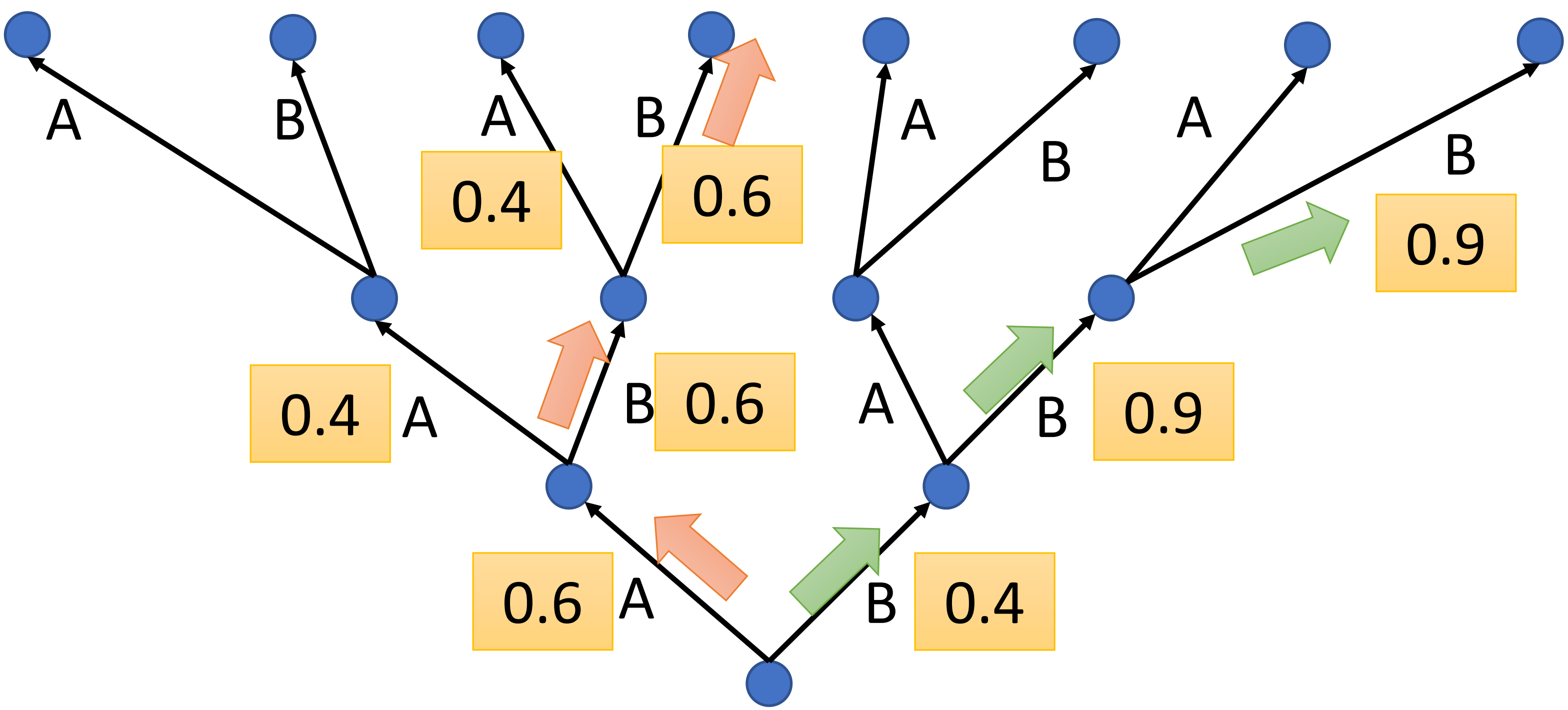
Beam Search

Assume there are only two tokens (V=2).

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.

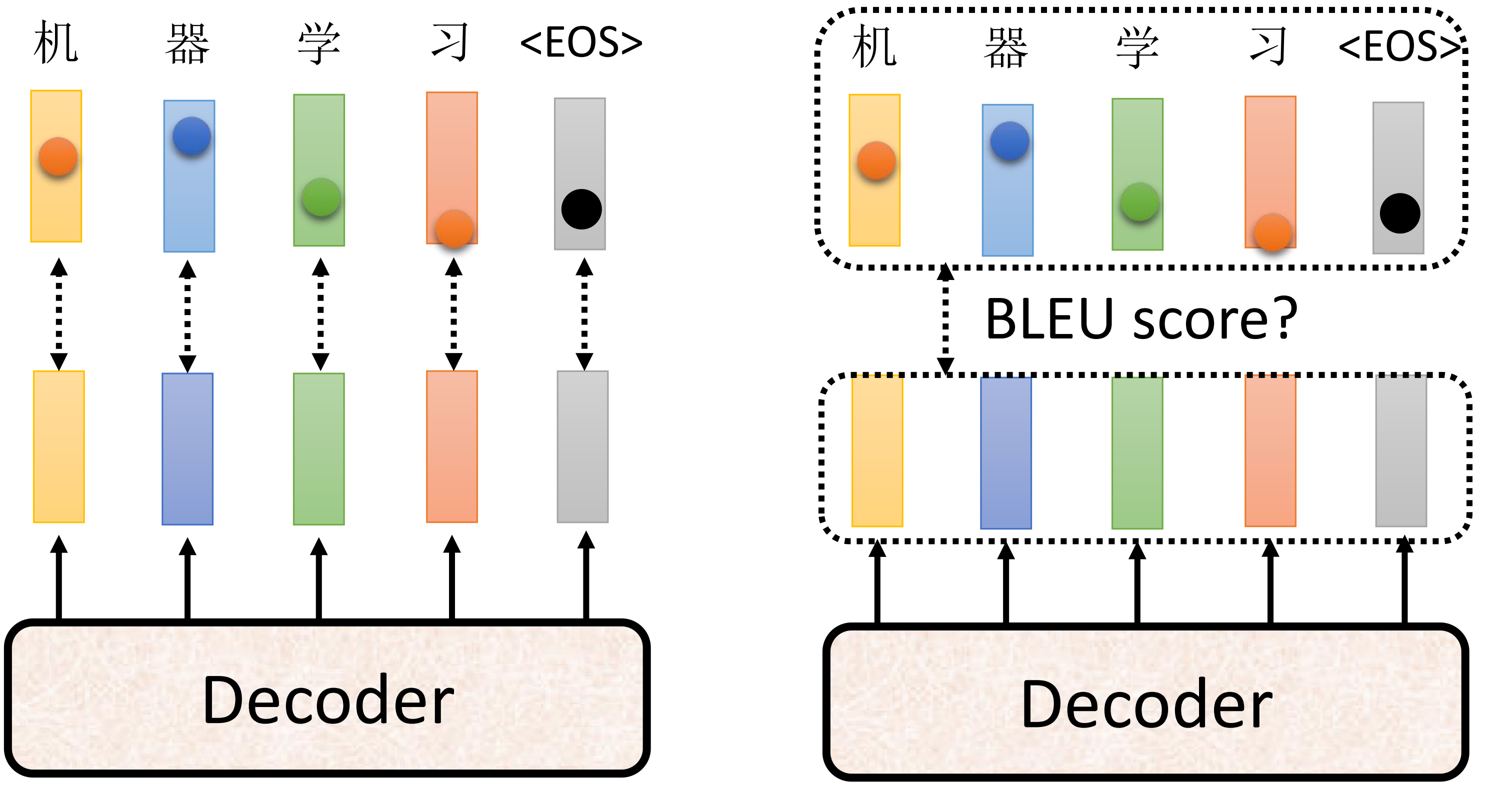
Beam Search, *b*=32:
"The study, published in the Proceedings of the National Academy of Sciences of the United States of America (PNAS), was conducted by researchers from the Universidad Nacional Autónoma de México (UNAM) and the Universidad Nacional Autónoma de México (UNAM/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de ..."

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?



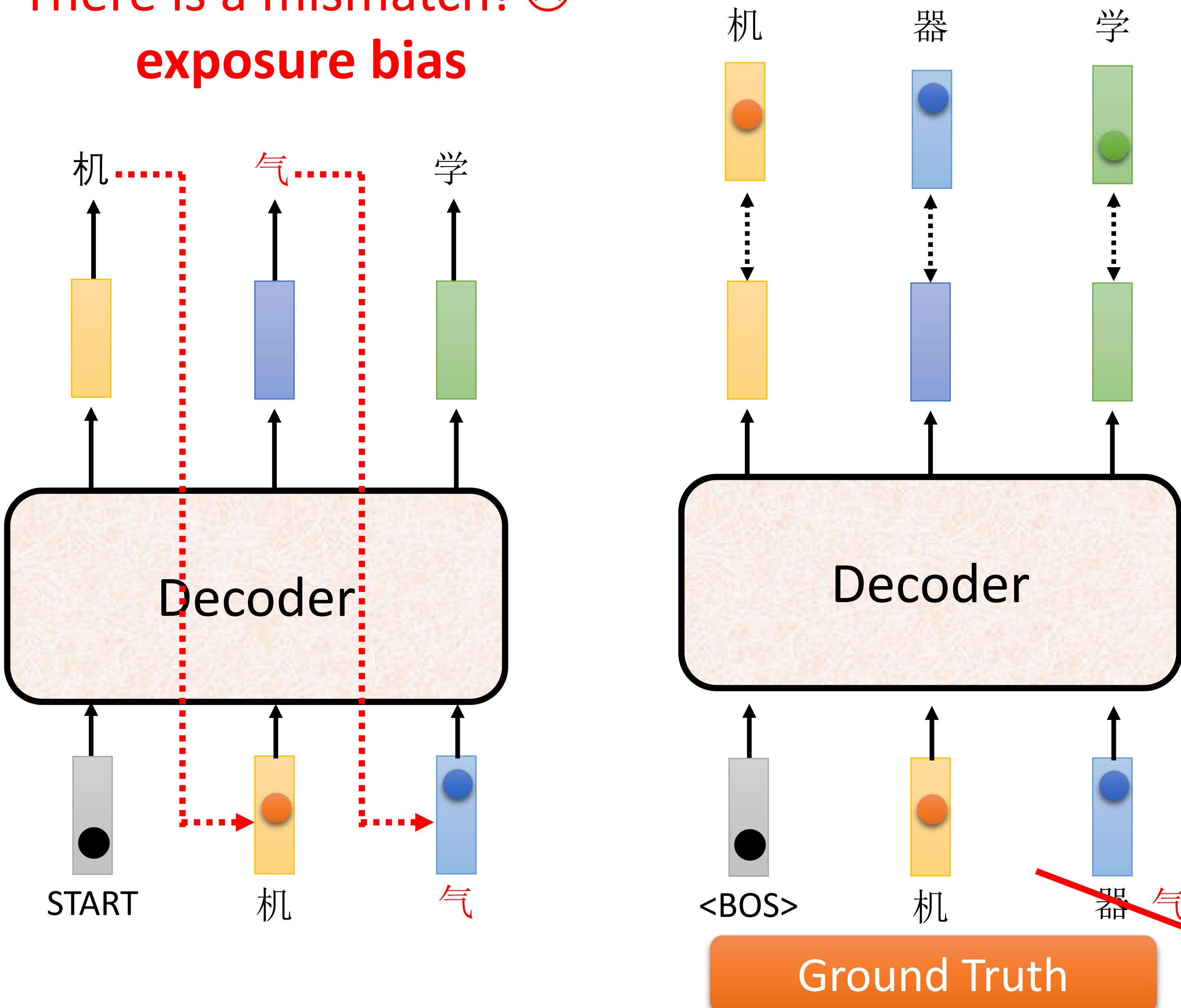
How to do the optimization?

When you don't know how to optimize, just use reinforcement learning (RL)!

https://arxiv.org/abs/1511.06732

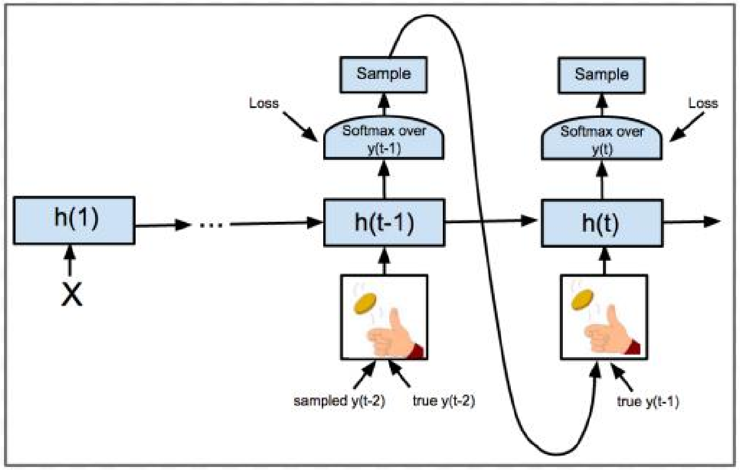
There is a mismatch! ☹

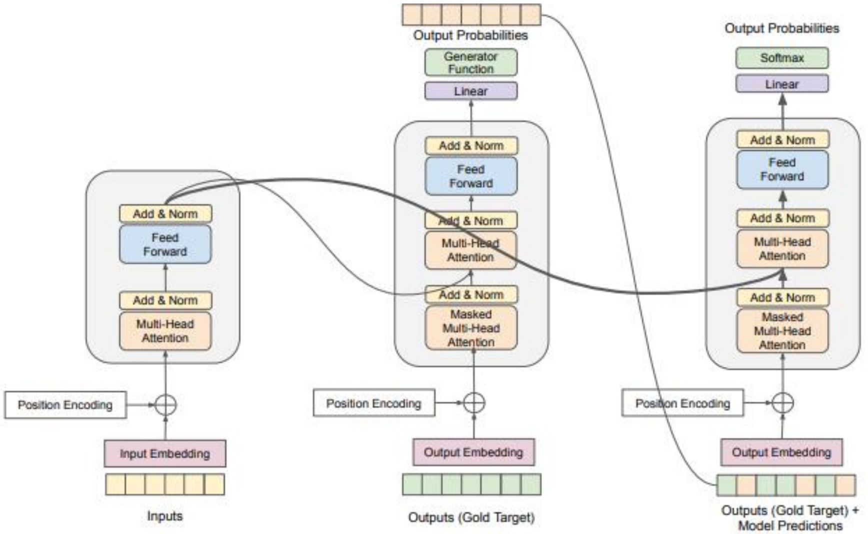
exposure bias



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





Concluding Remarks: Transformer