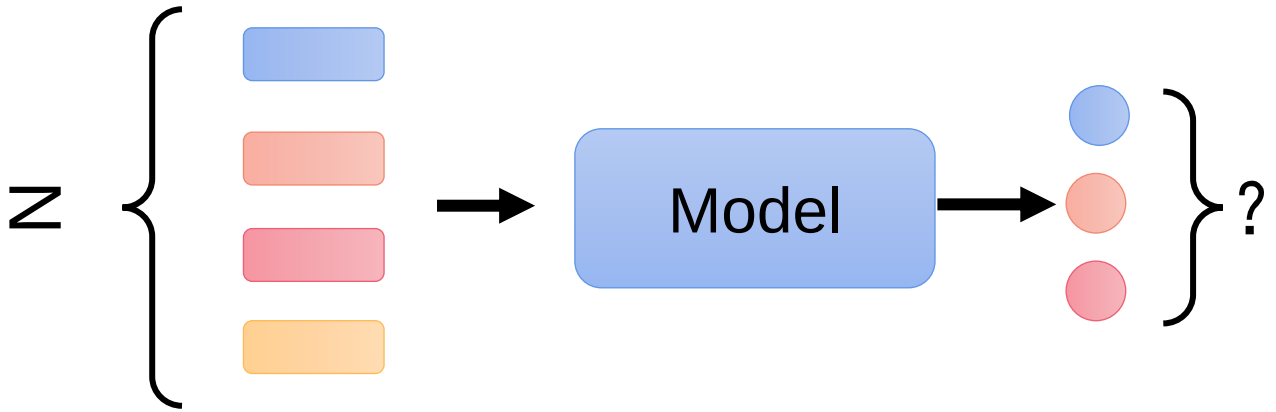
第四部分：

生成任务，大模型

常见输出。

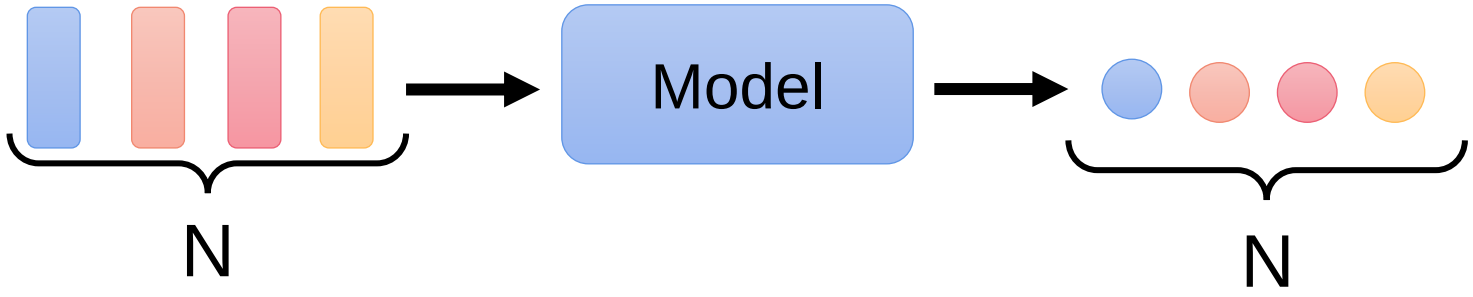
- 输入输出长度不对应。.



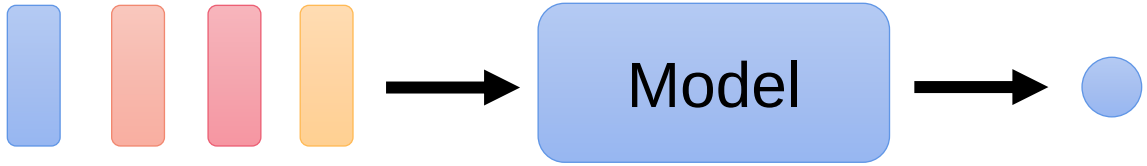
我爱中国 翻译 I love China

What is the output?

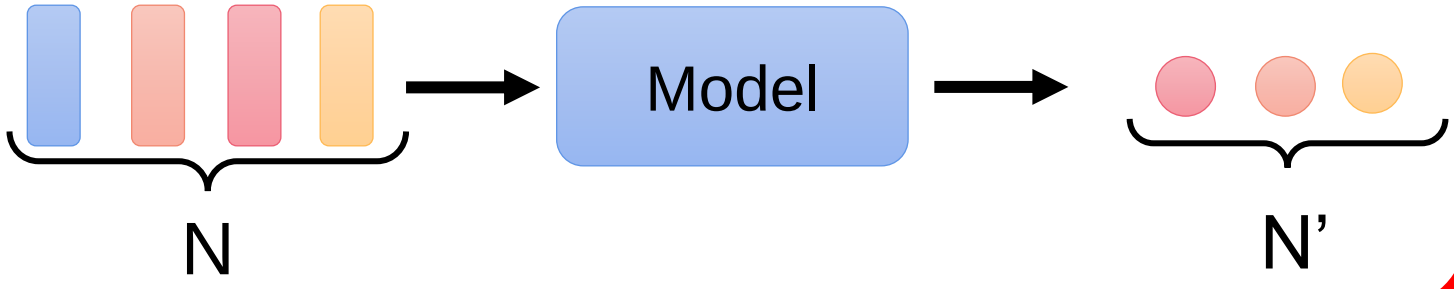
- Each vector has a label.



- The whole sequence has a label.

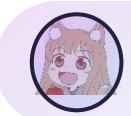


- Model decides the number of labels itself.



这节课

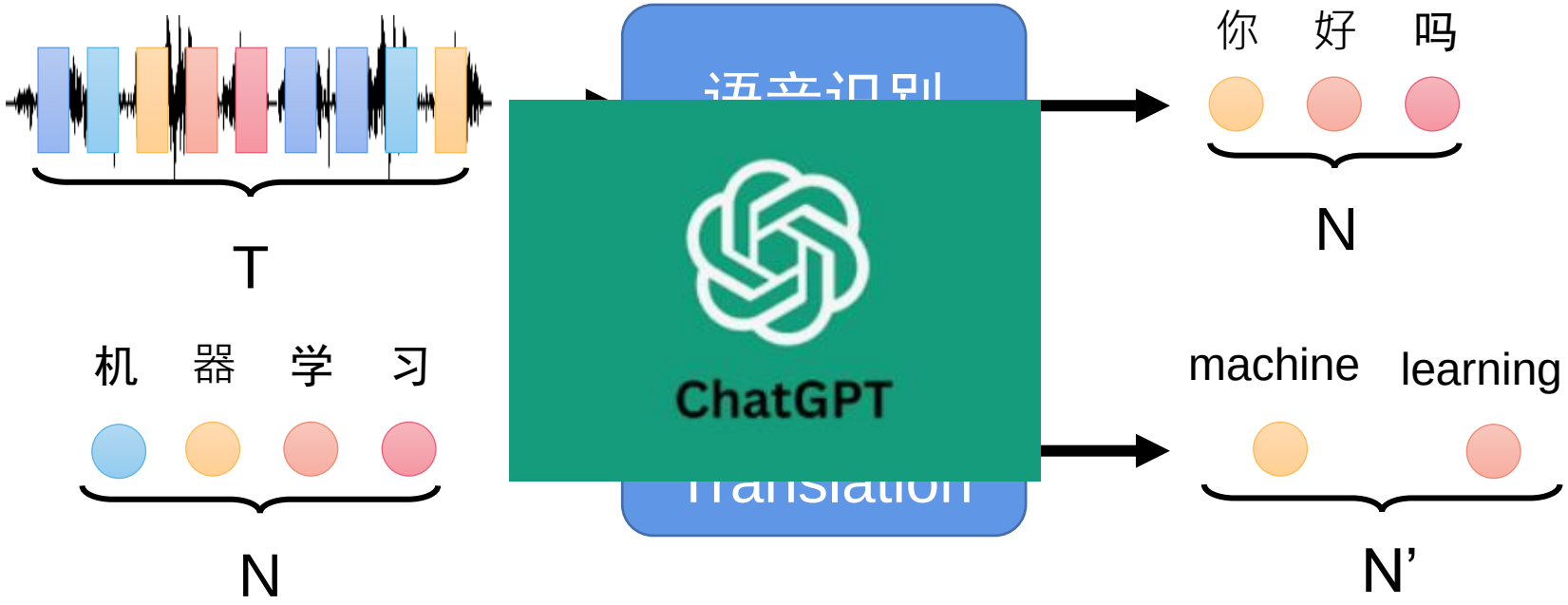
Sequence-to-sequence (Seq2seq)



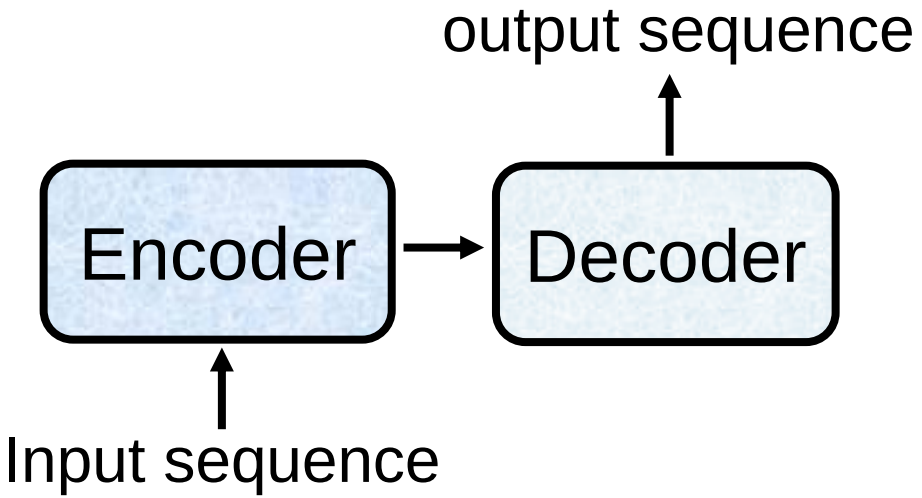
李哥考研

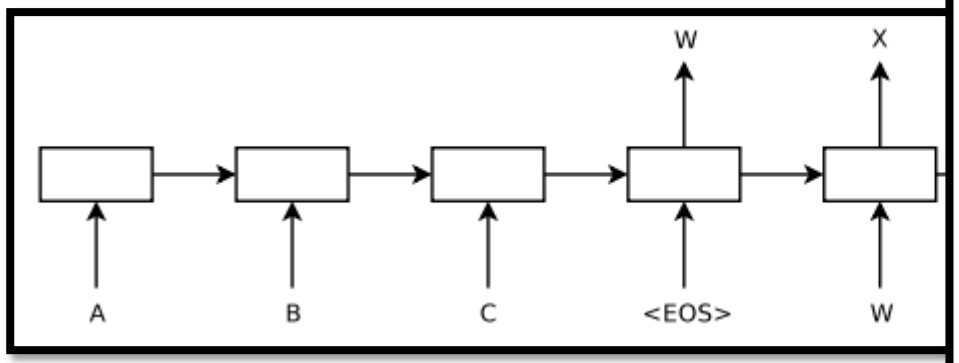
输入一段，输出一段

输出长度由模型自己来决定



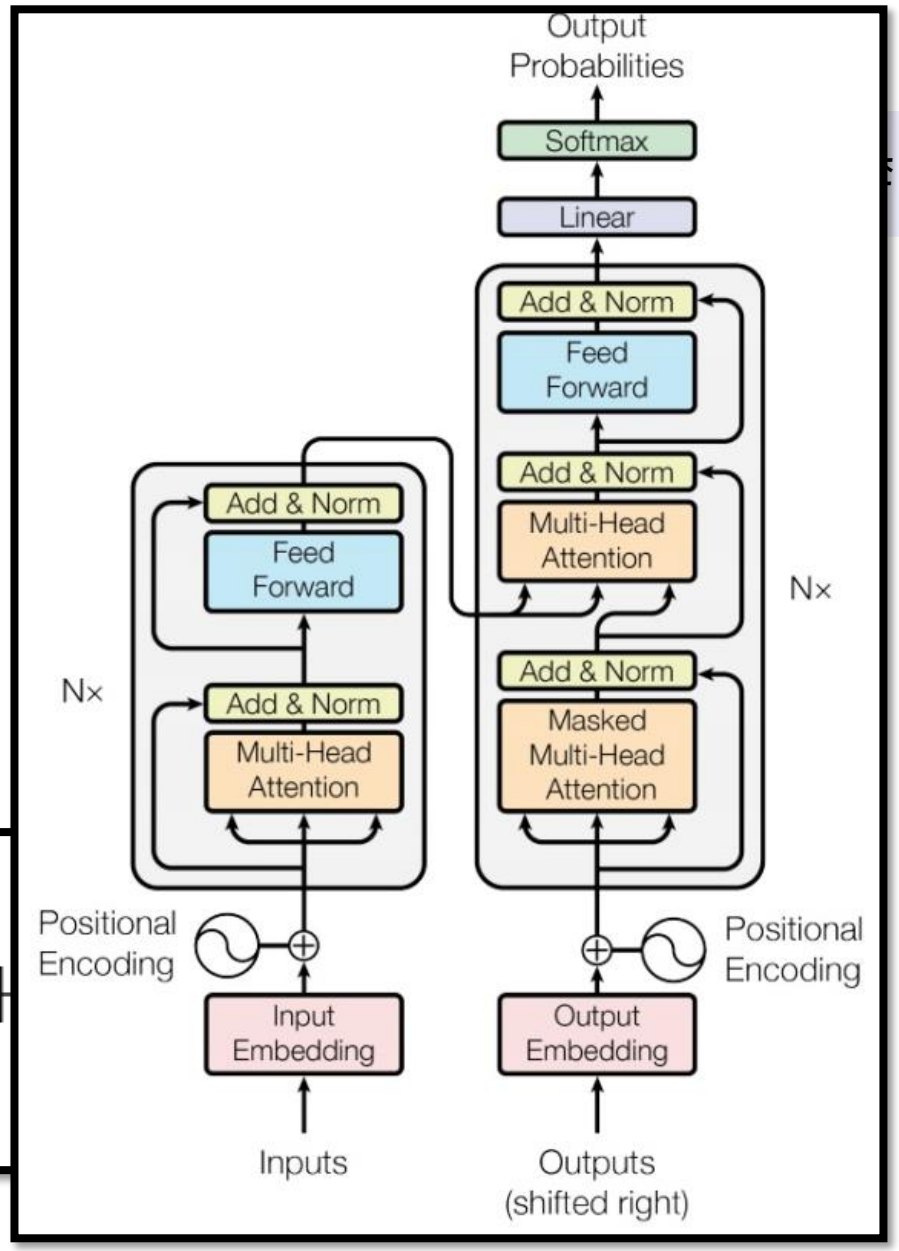
Seq2seq





Sequence to Sequence Learning with Neural Networks

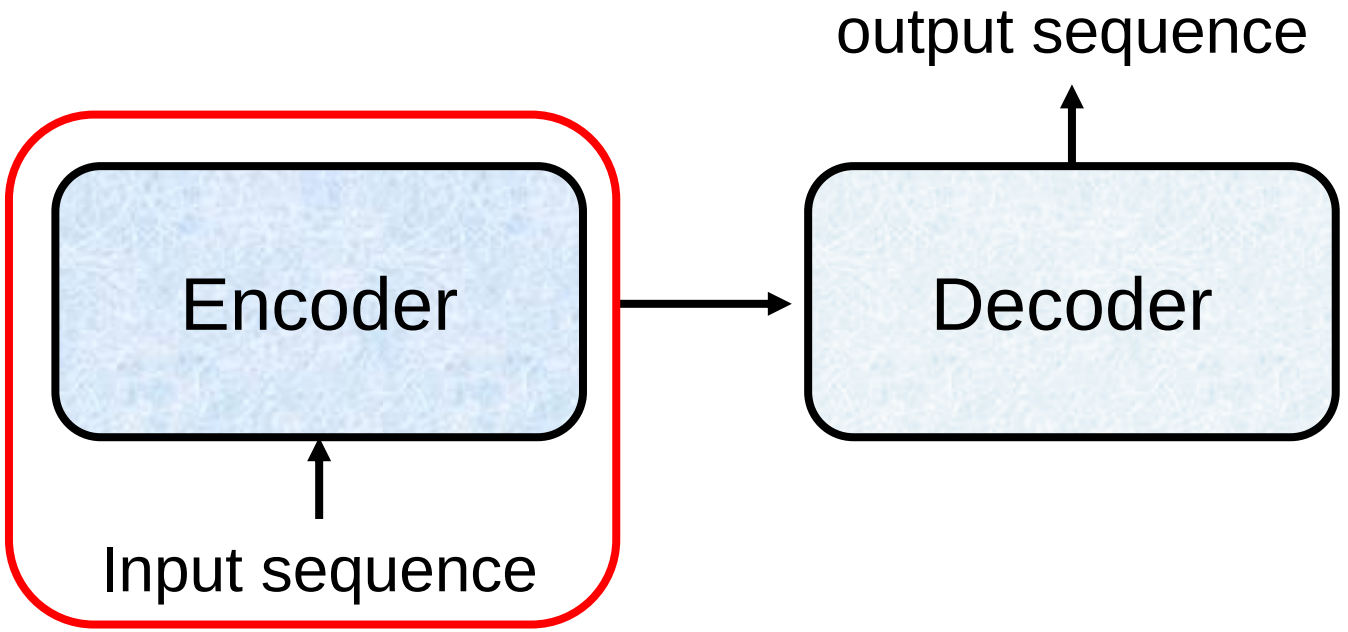
https://arxiv.org/abs/1409.3215



Transformer

https://arxiv.org/abs/1706.03762

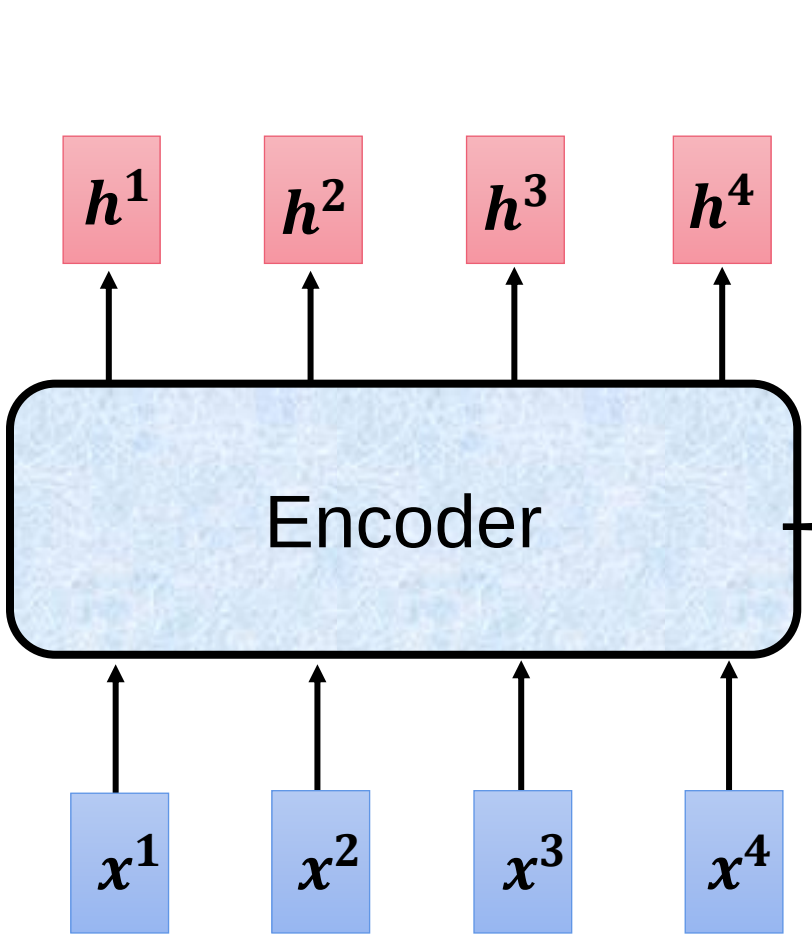
Encoder

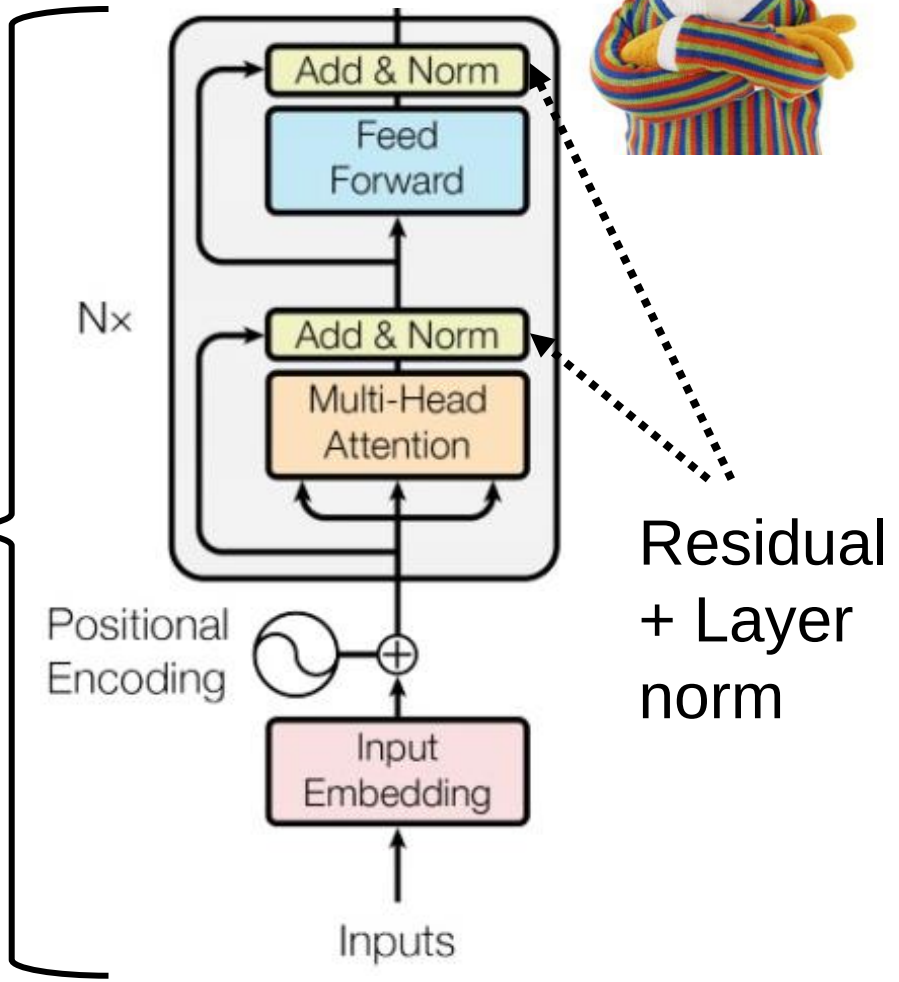


I use the **same** network architecture as **transformer encoder**.

BERT

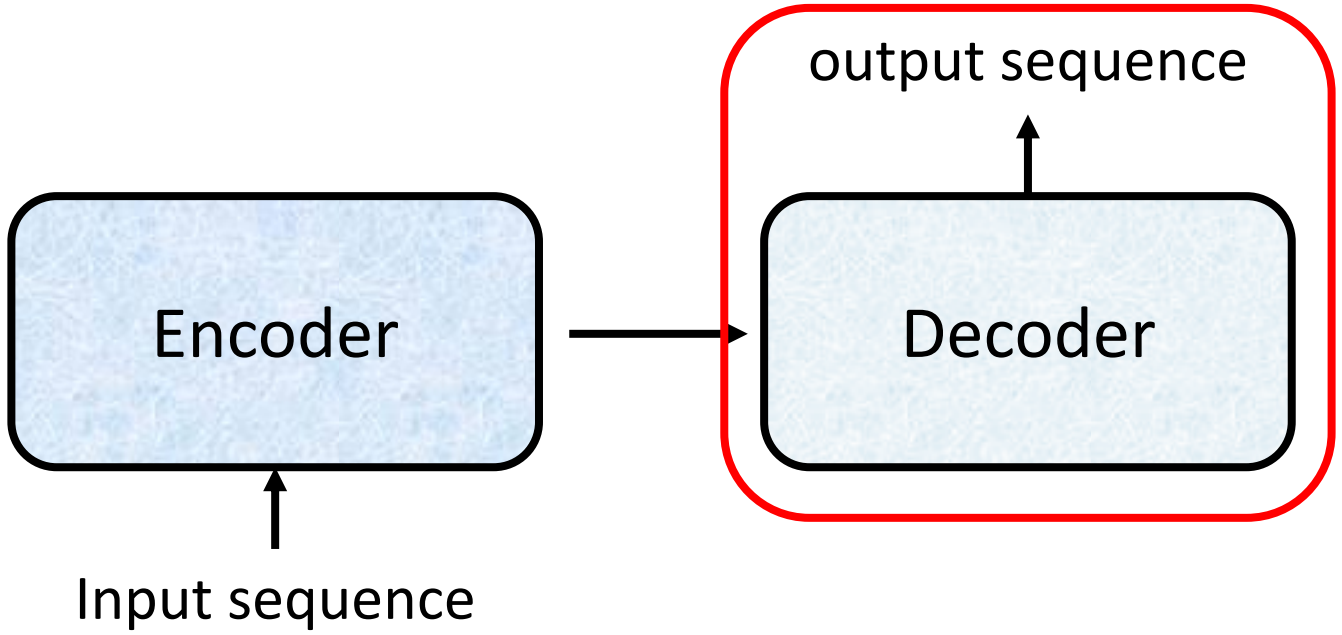
哥考研





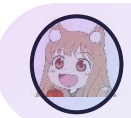
Residual + Layer norm

Decoder

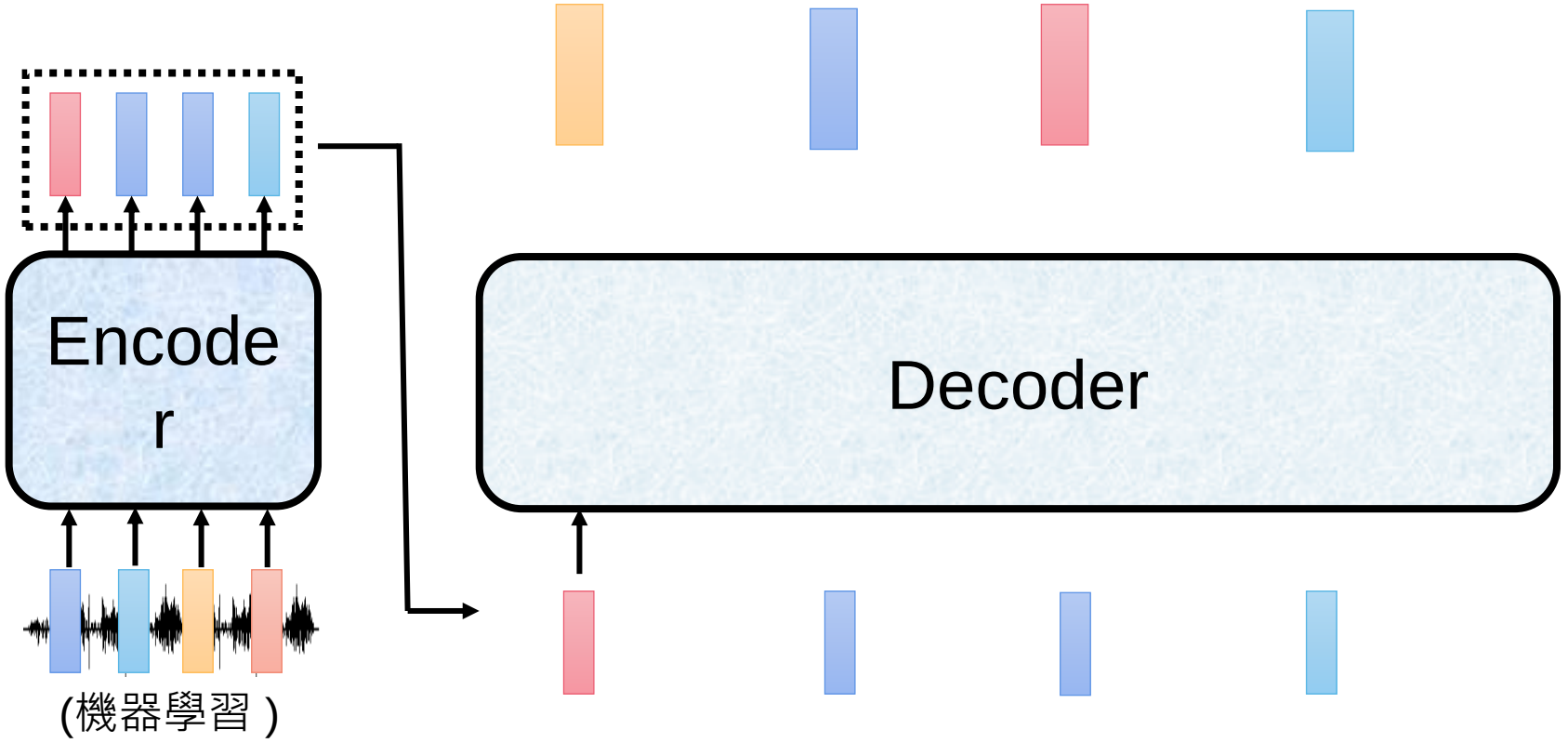


生成方式

distribution



李哥考研

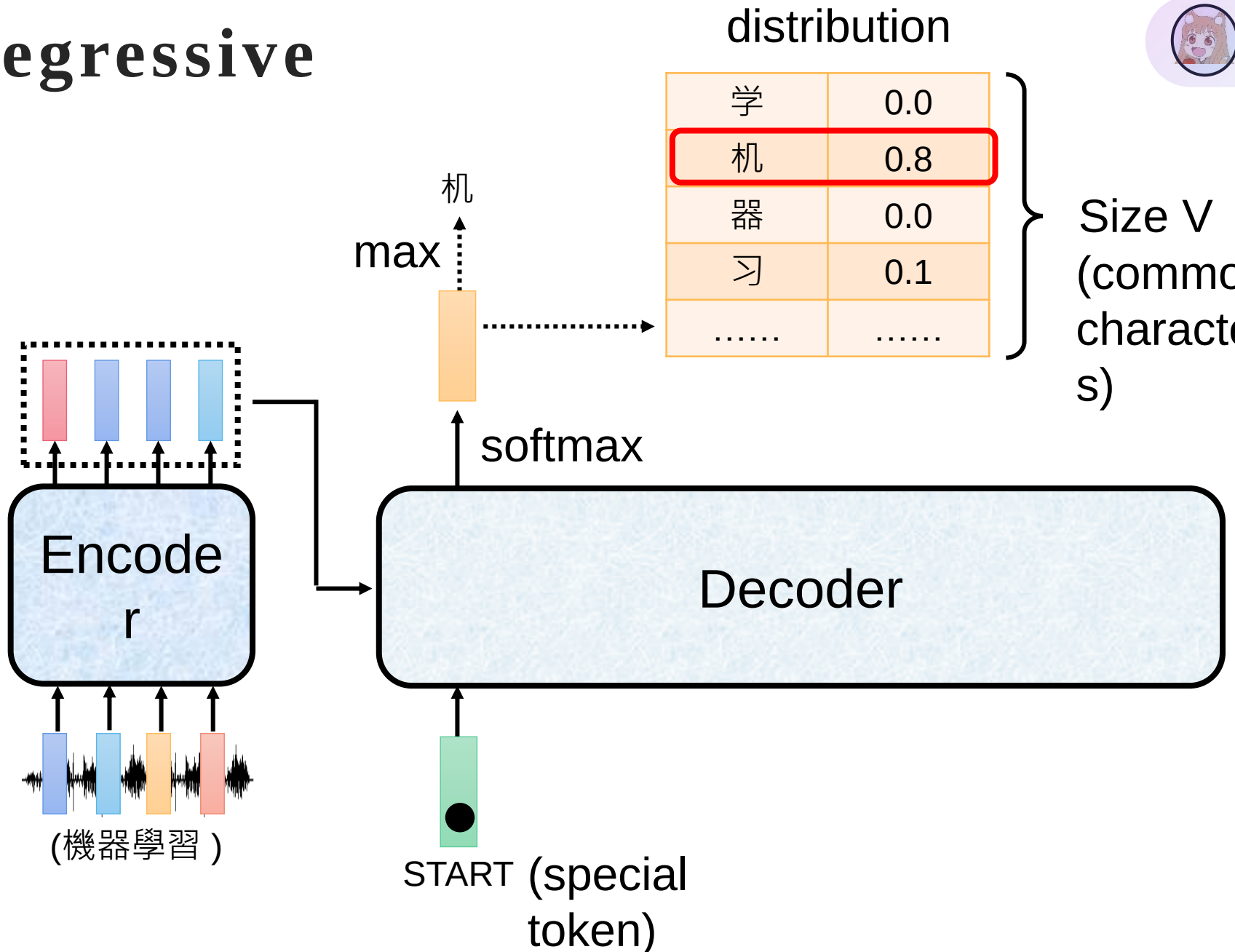


两个问题

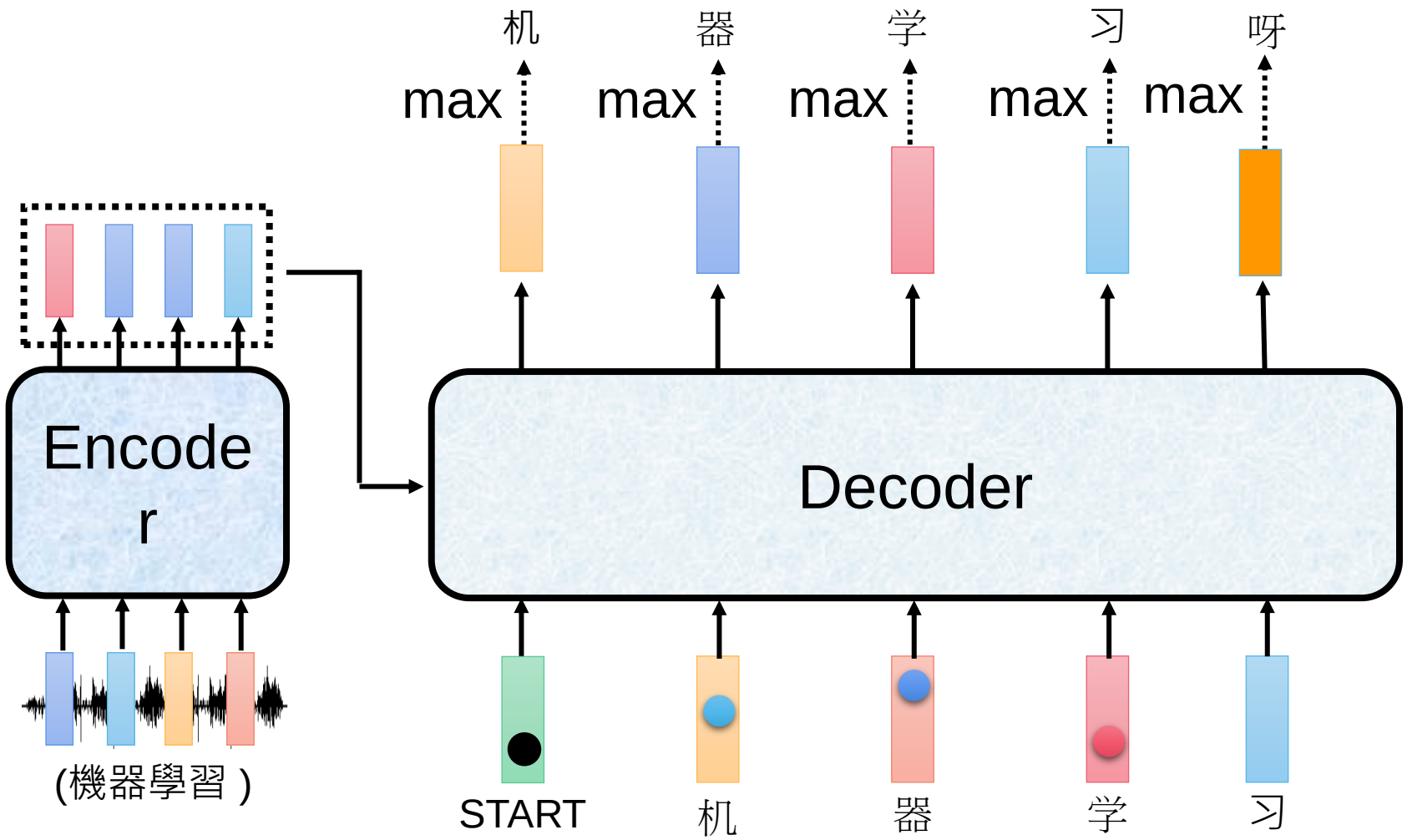
1， 效果差

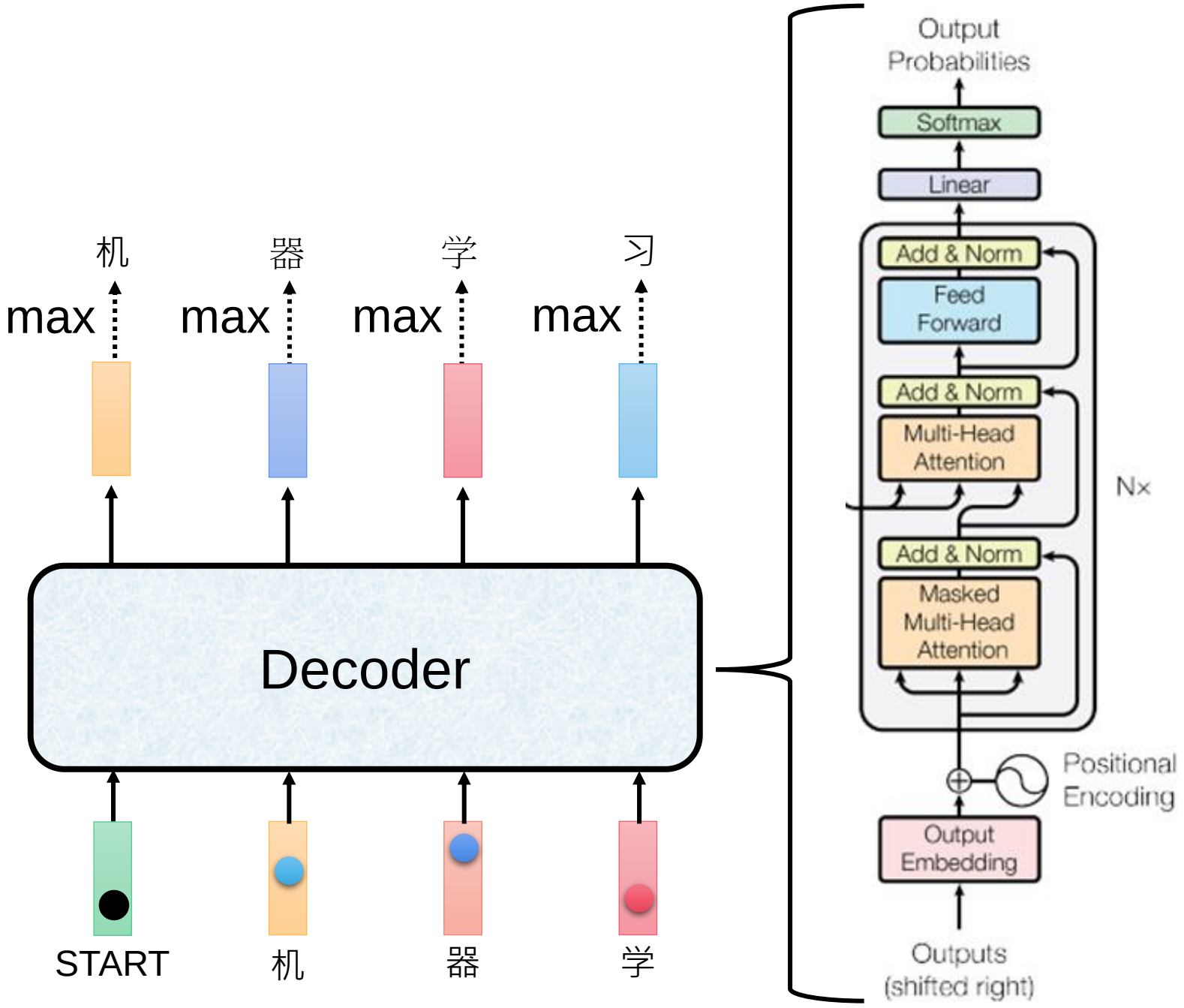
2， 长度

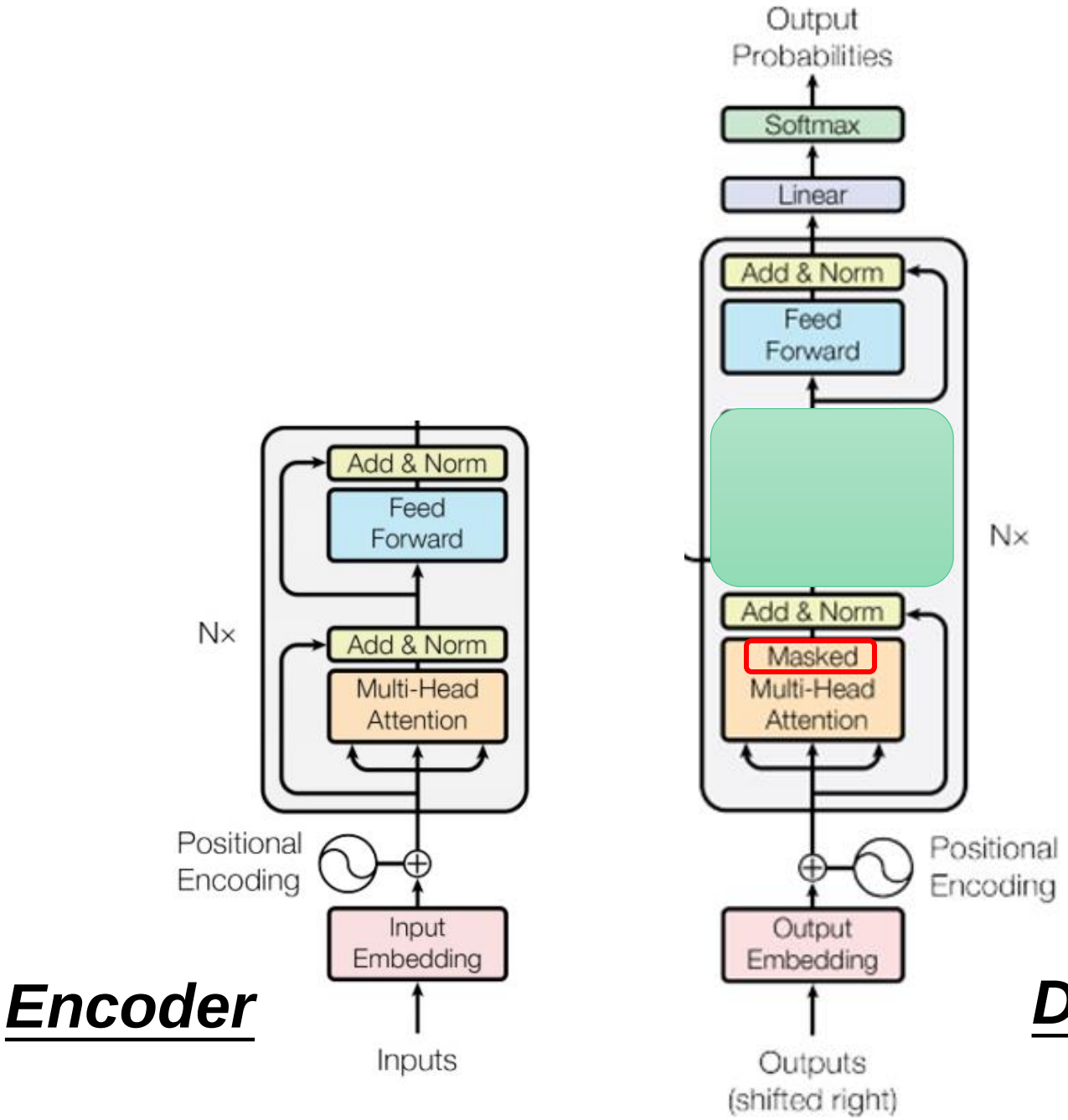
Autoregressive



Autoregressive



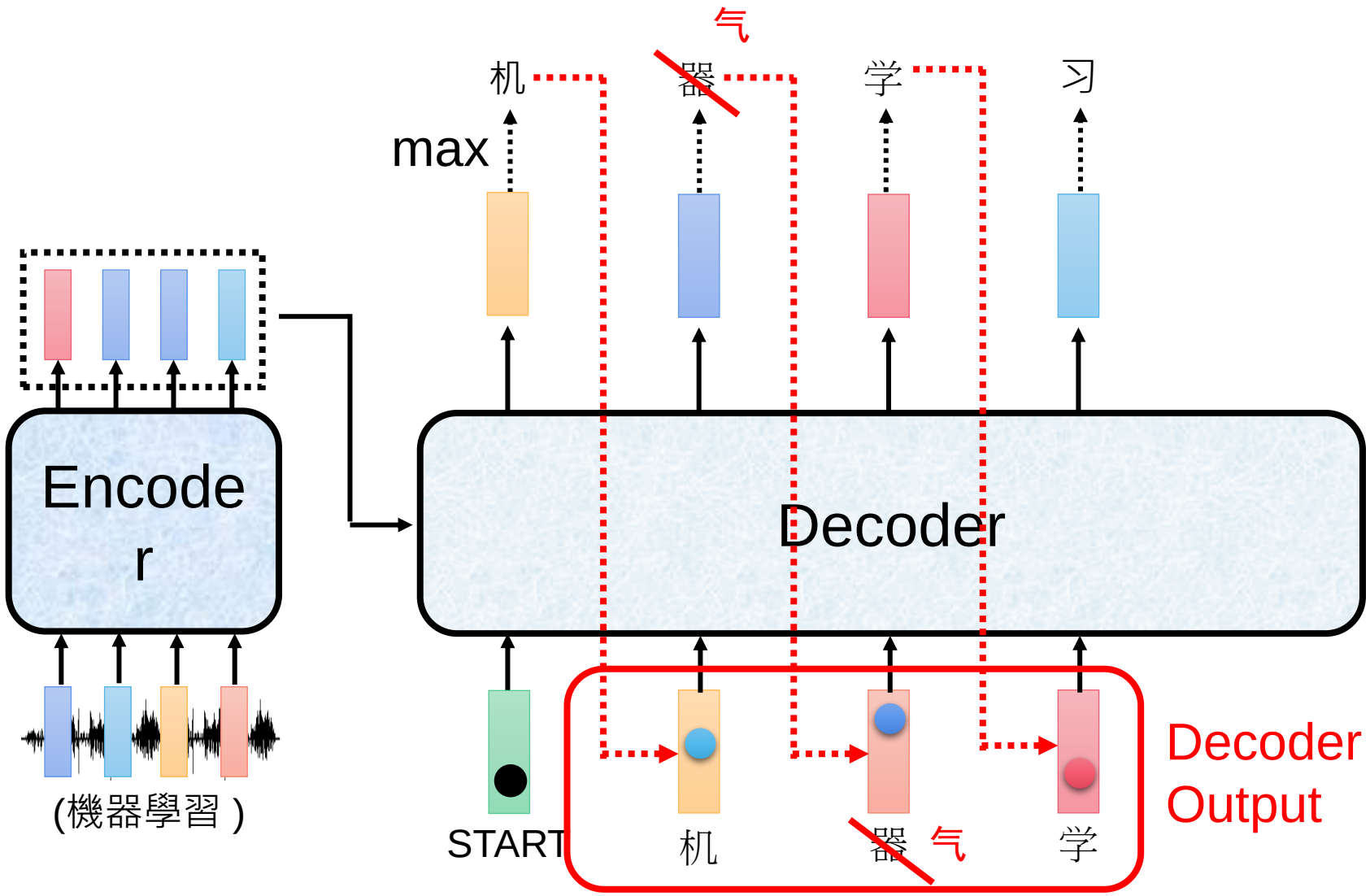

机
器
学
习
max
max
max
max
Decoder
START
机
器
学
Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Positional Encoding
Output Embedding
Outputs (shifted right)



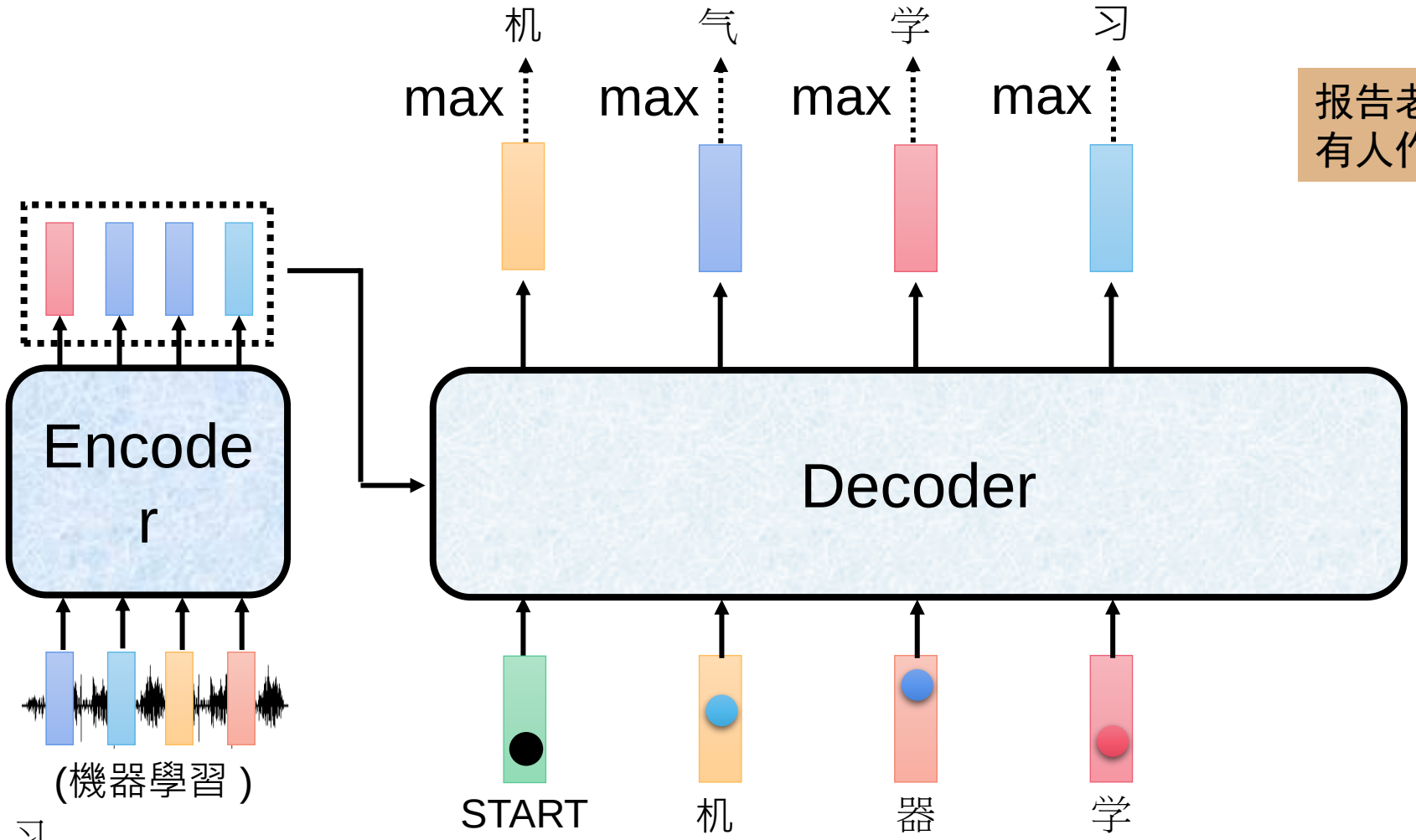
Encoder

Decoder

Autoregressive

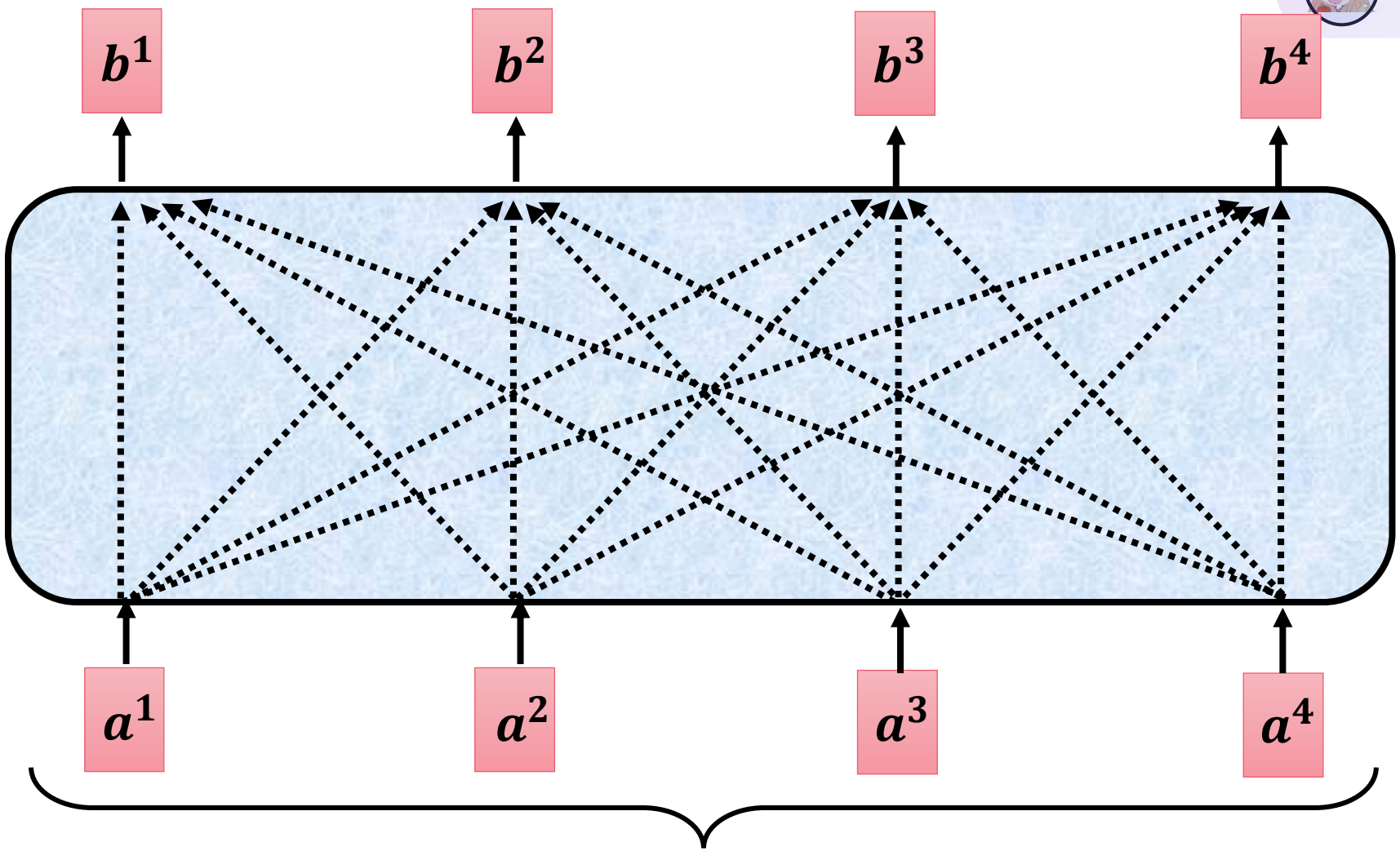


Autoregressive



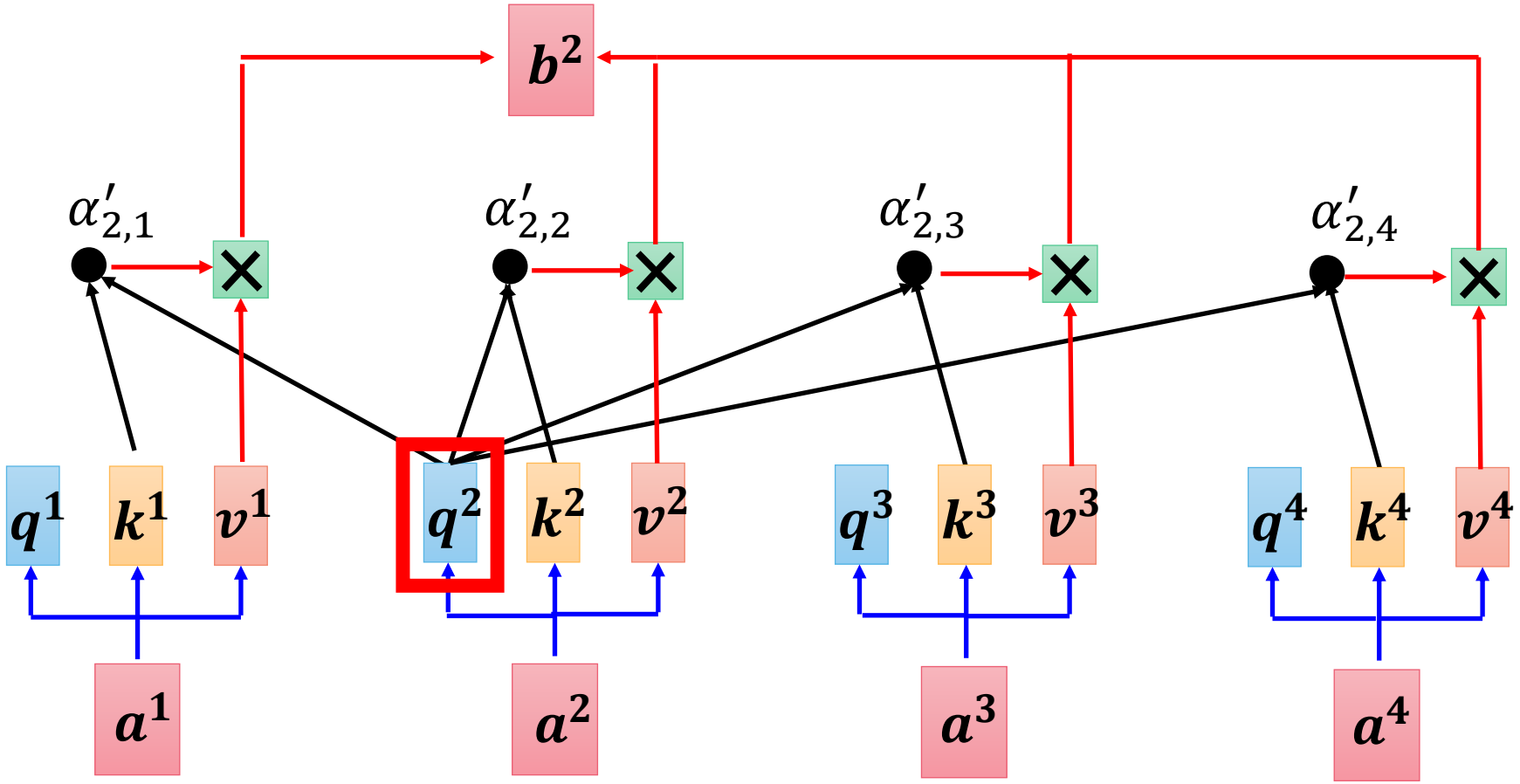
Label: 机 器 学 习

Self-attention ➤ Masked Self-attention



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

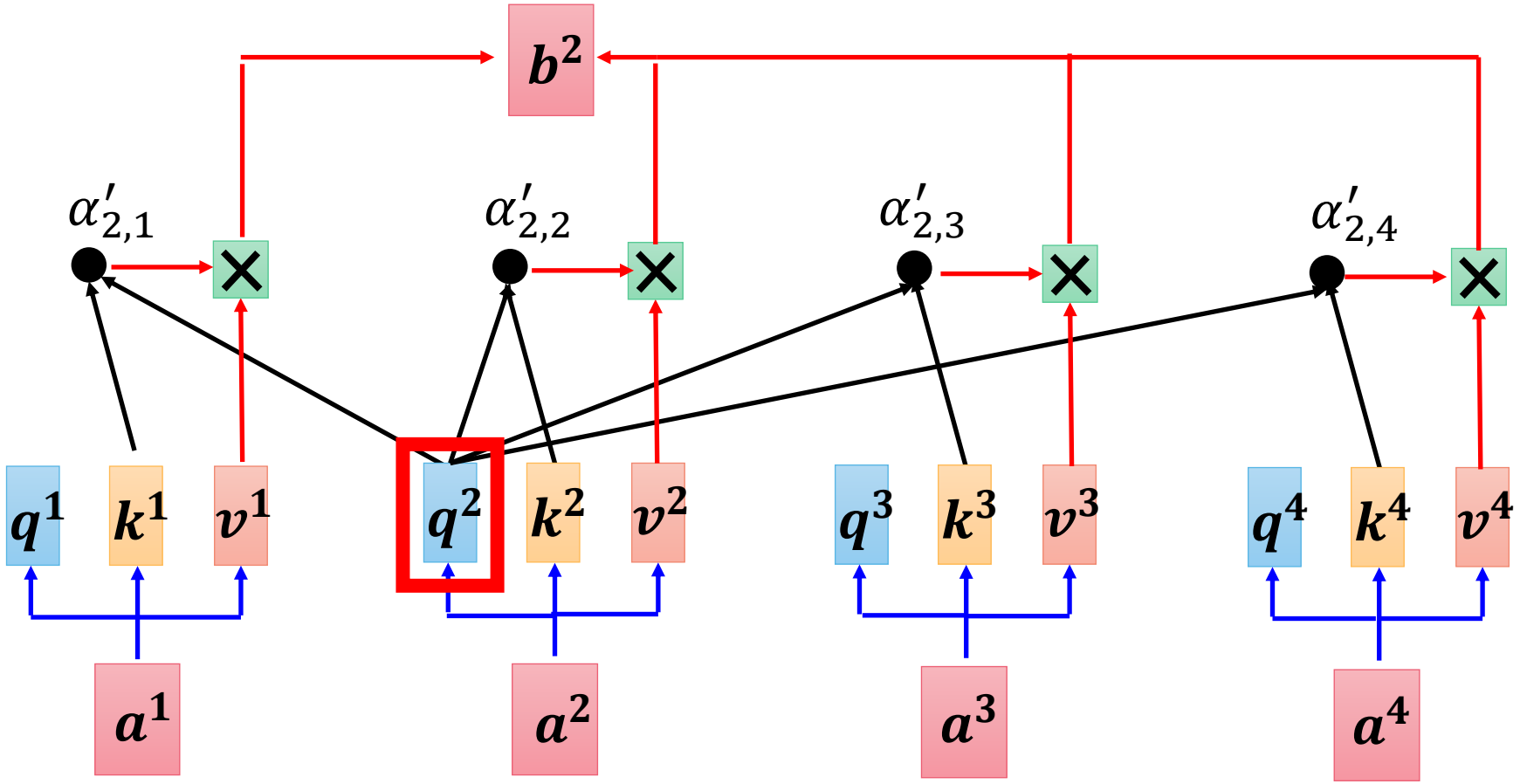
Self-attention → Masked Self-attention



思考为什么是mask的

Self-attention → Masked Self-attention

Mask
=
[
1, 0, 0, 0
1, 1, 0, 0
1, 1, 1, 0
1, 1, 1, 1
]

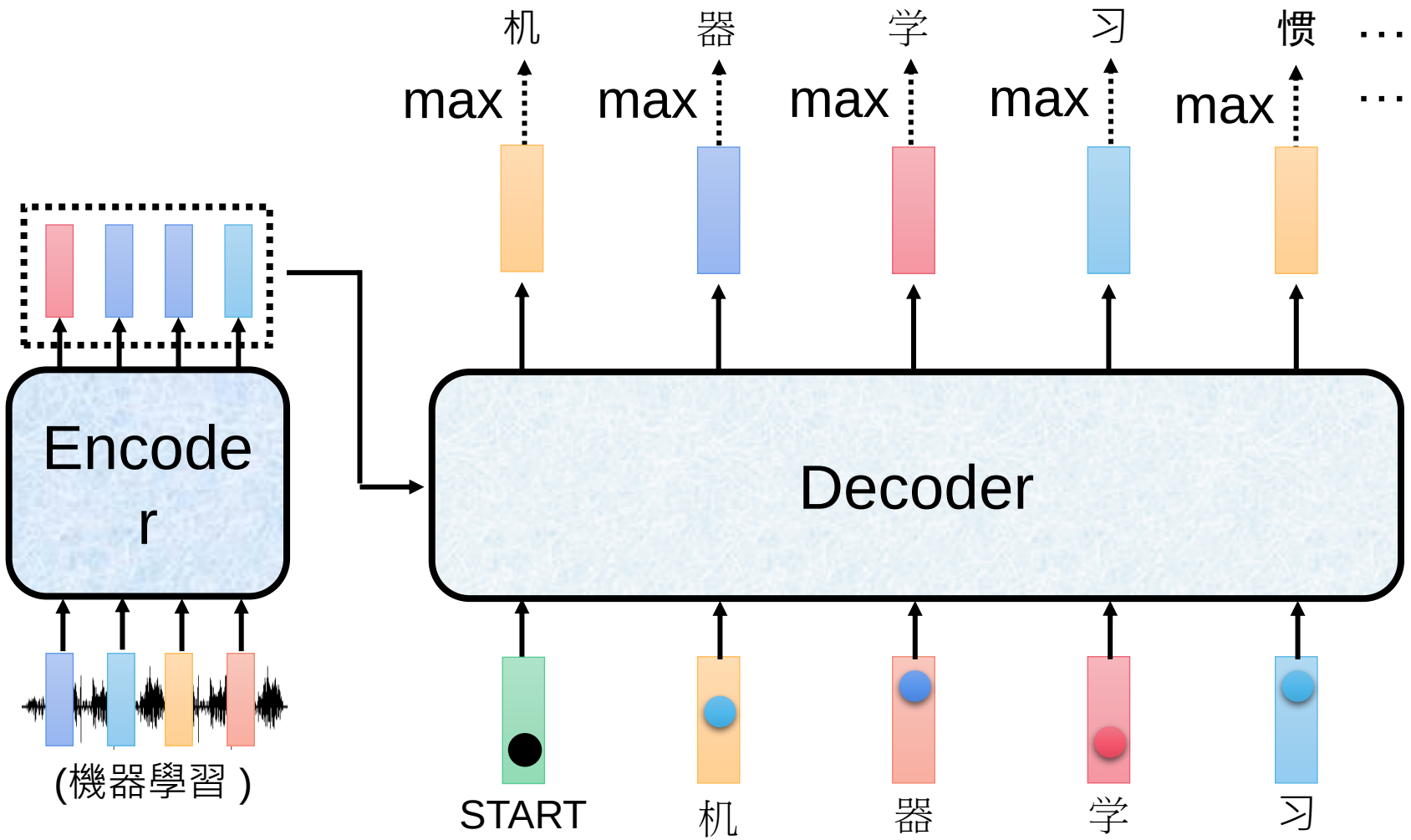


思考为什么是mask的

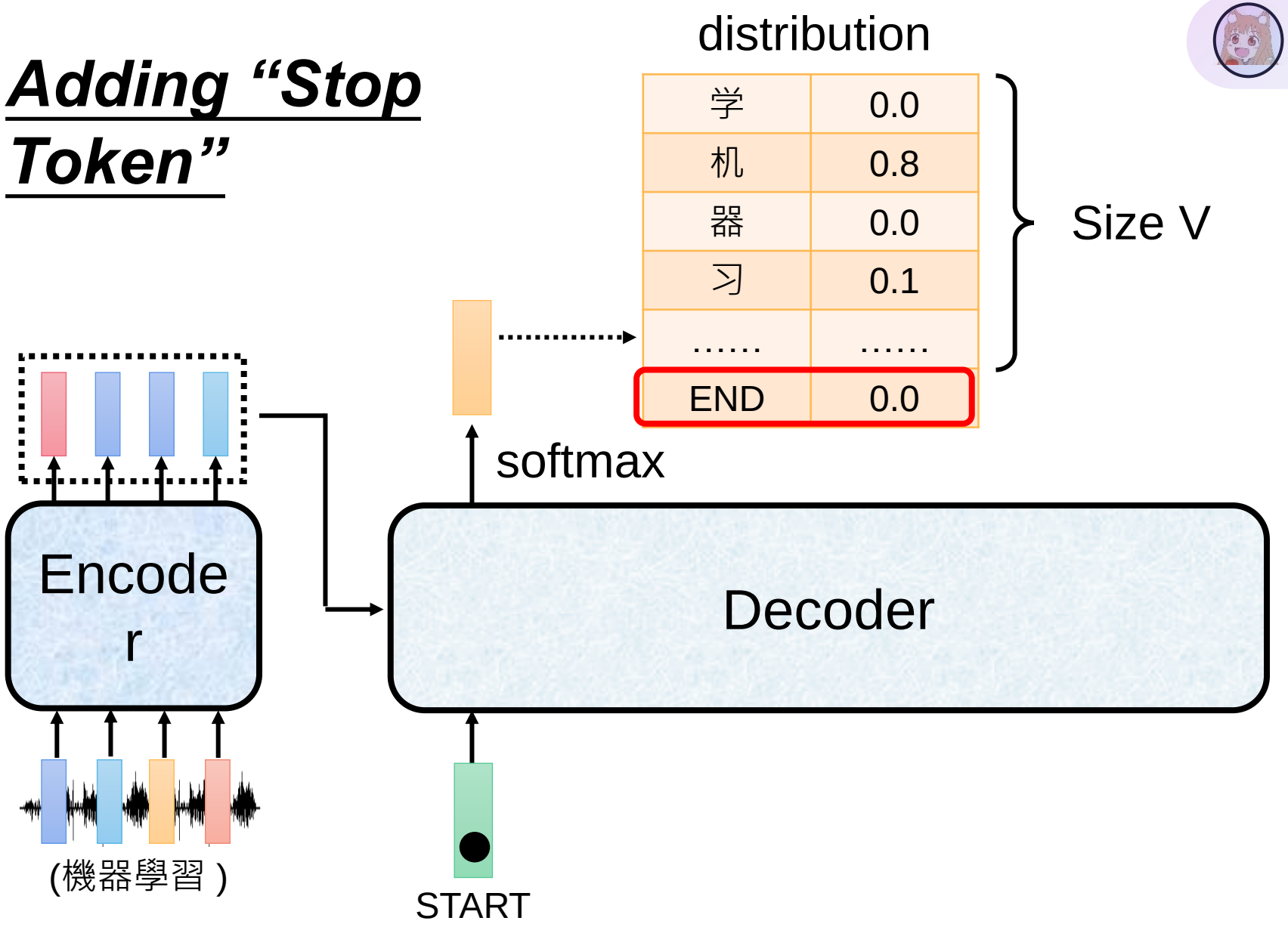
Autoregressive

如何知道要输出多长

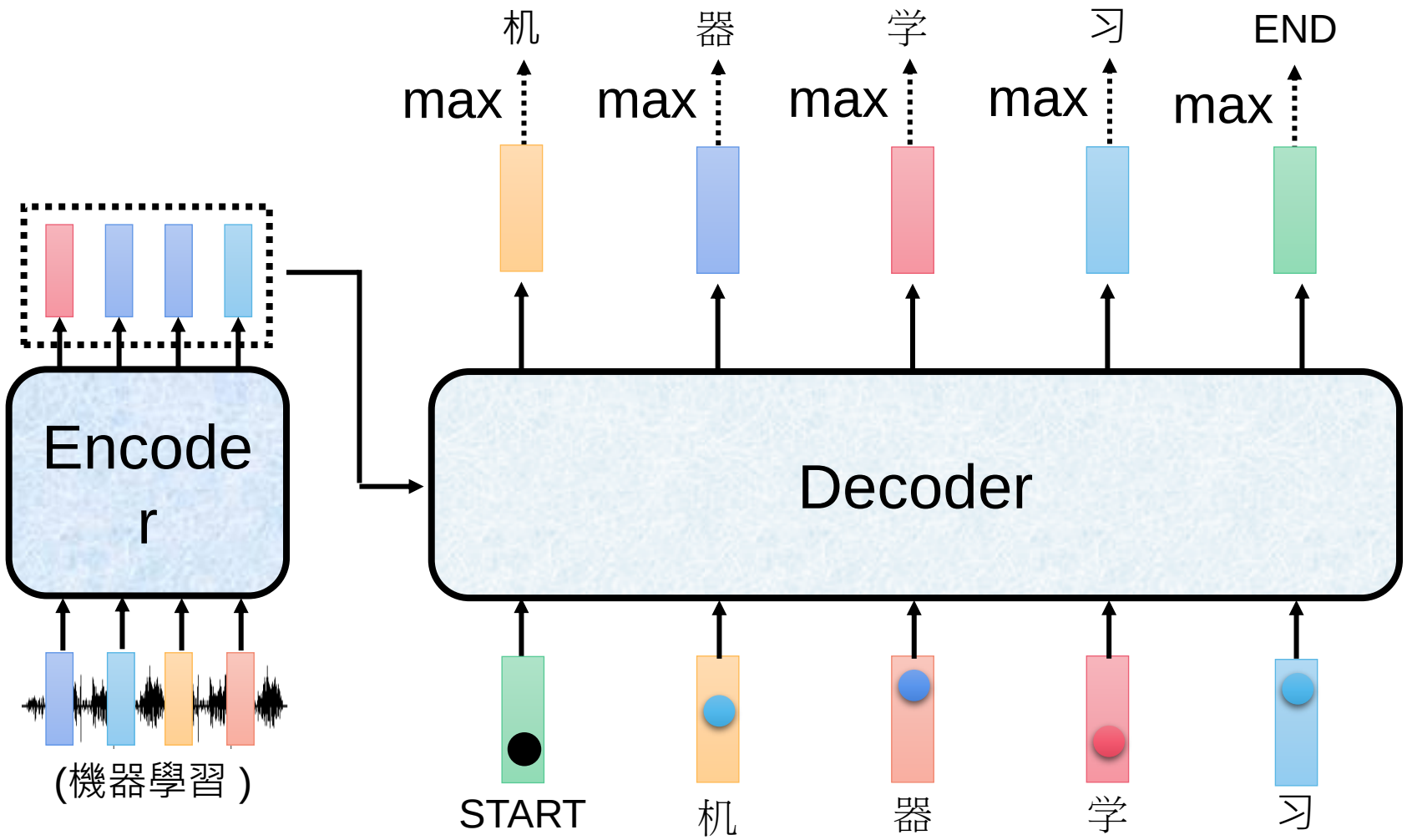
Never stop!



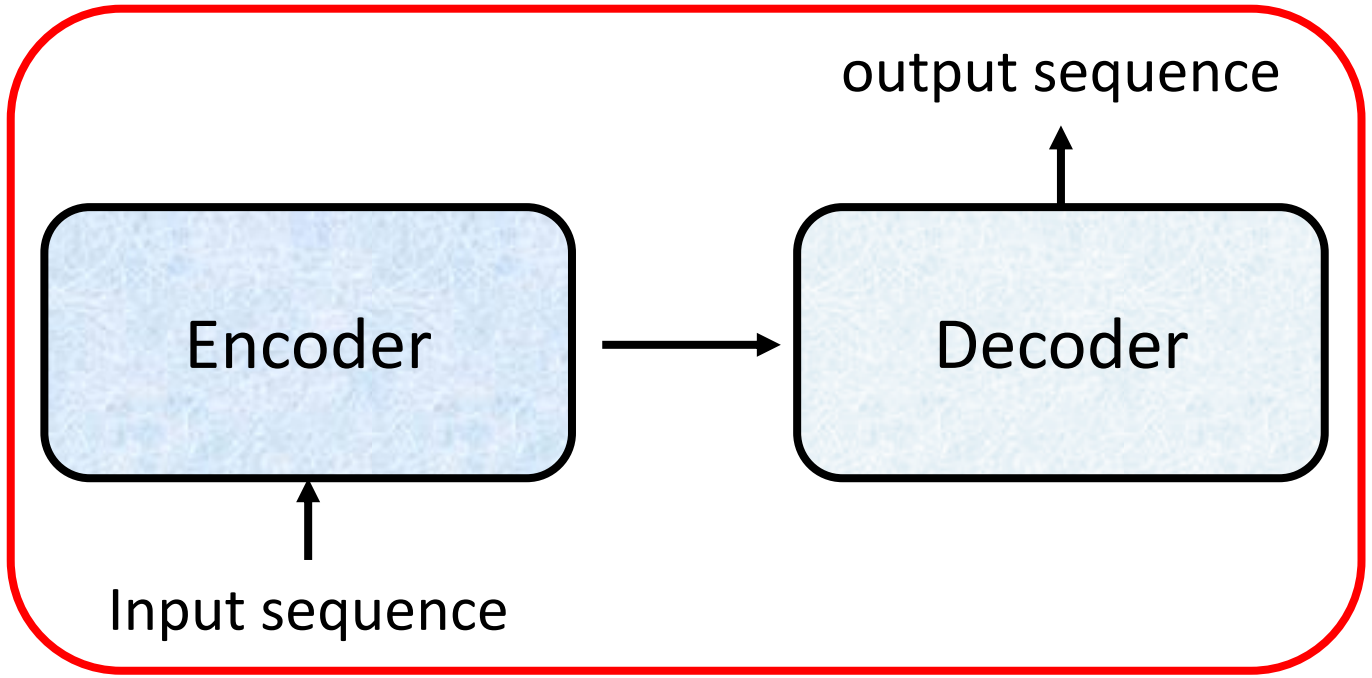
Adding "Stop Token"

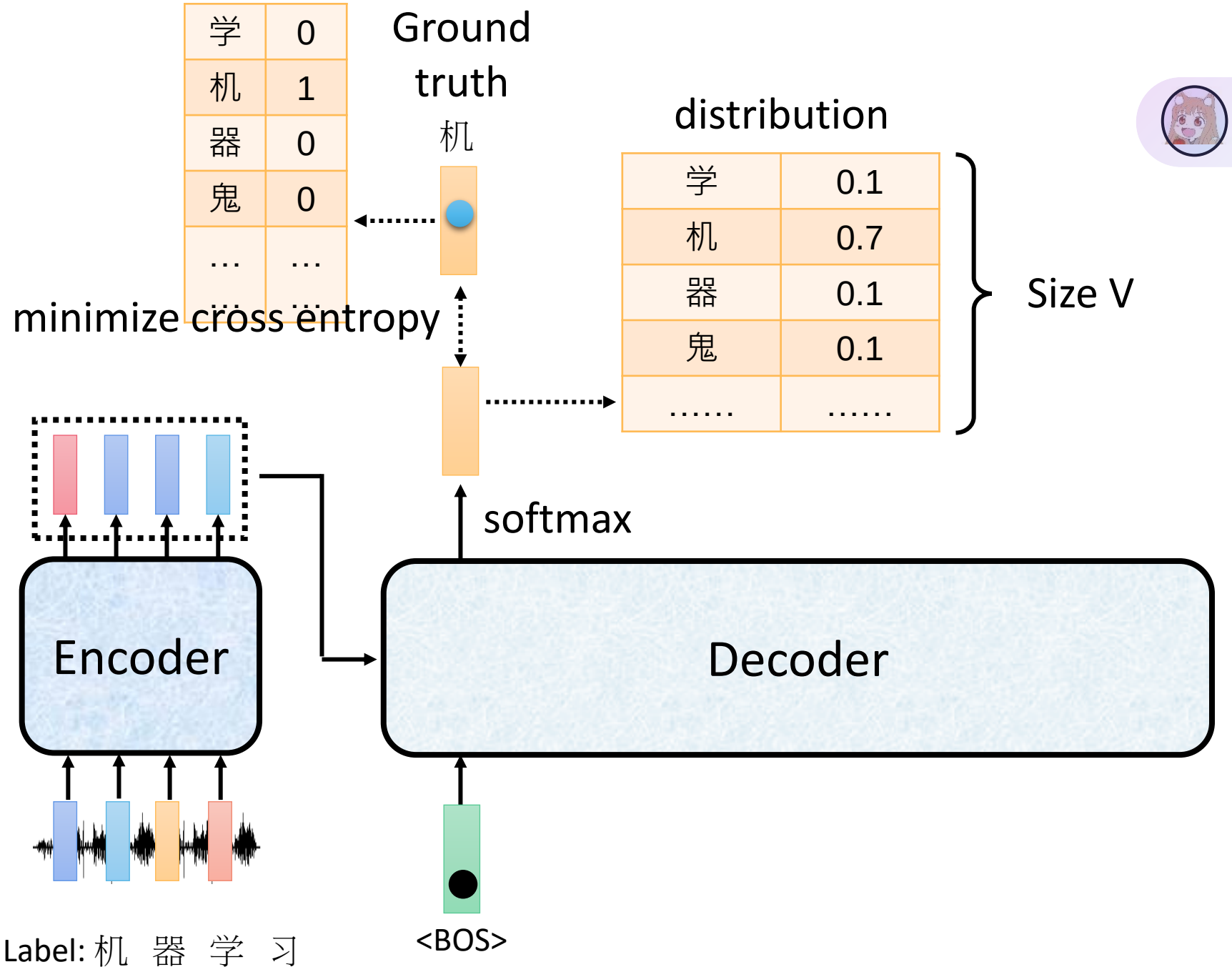


Autoregressive

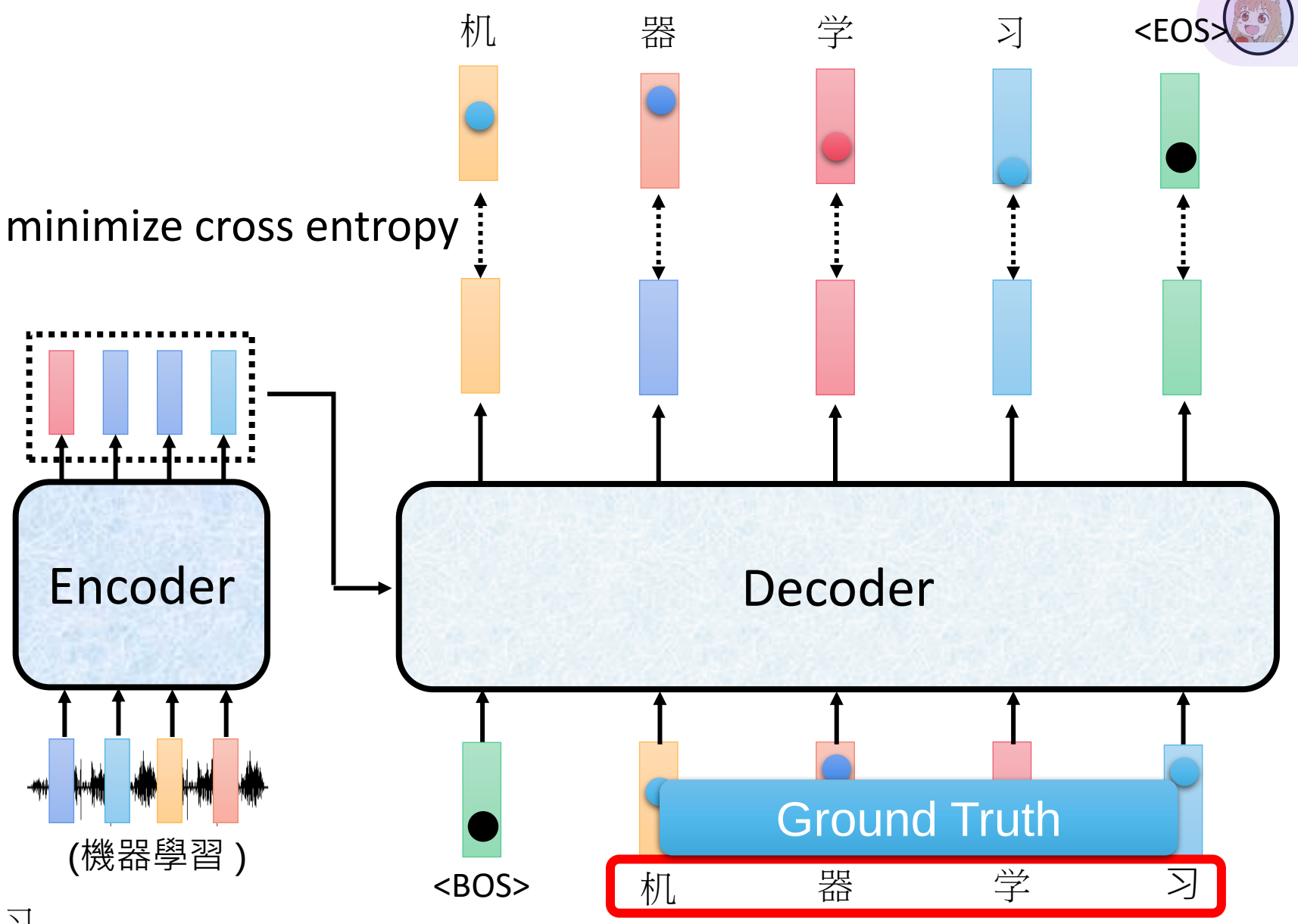


Training



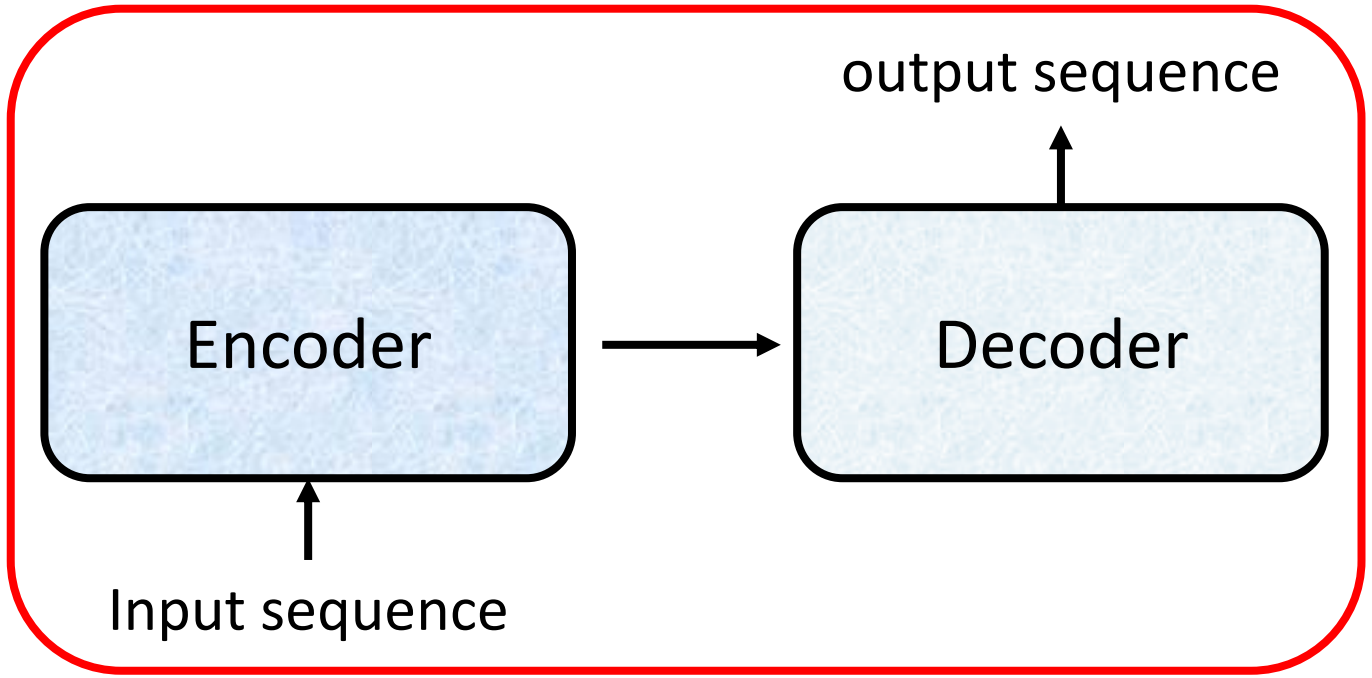

学 0
机 1
器 0
鬼 0
… …
Ground truth
机
distribution
学 0.1
机 0.7
器 0.1
鬼 0.1
…… ……
Size V
minimize cross entropy
softmax
Encoder
Decoder
<BOS>
Label: 机 器 学 习

训练: 要将标签当作输入



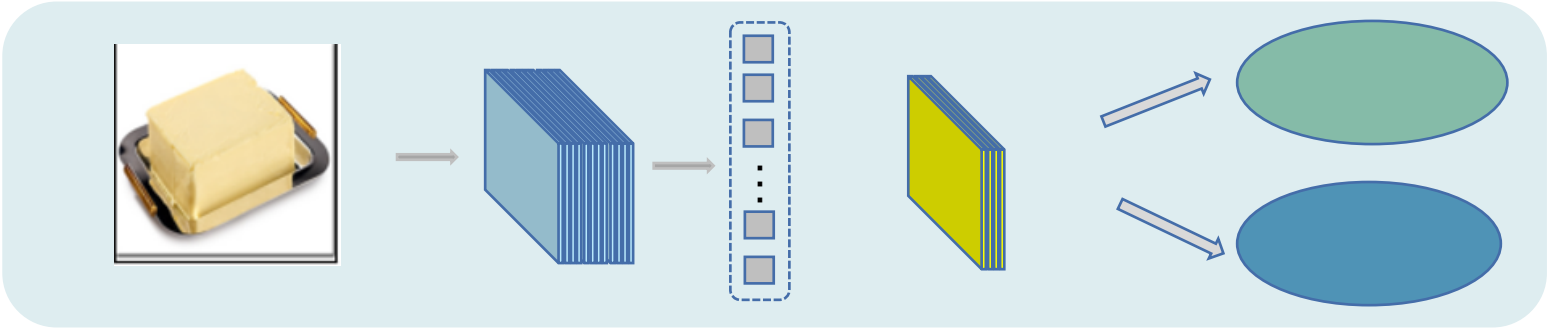
Label: 机 器 学 习

生成任务的训练，和推断



平常任务

训练时的方式，即为测试时的方式。

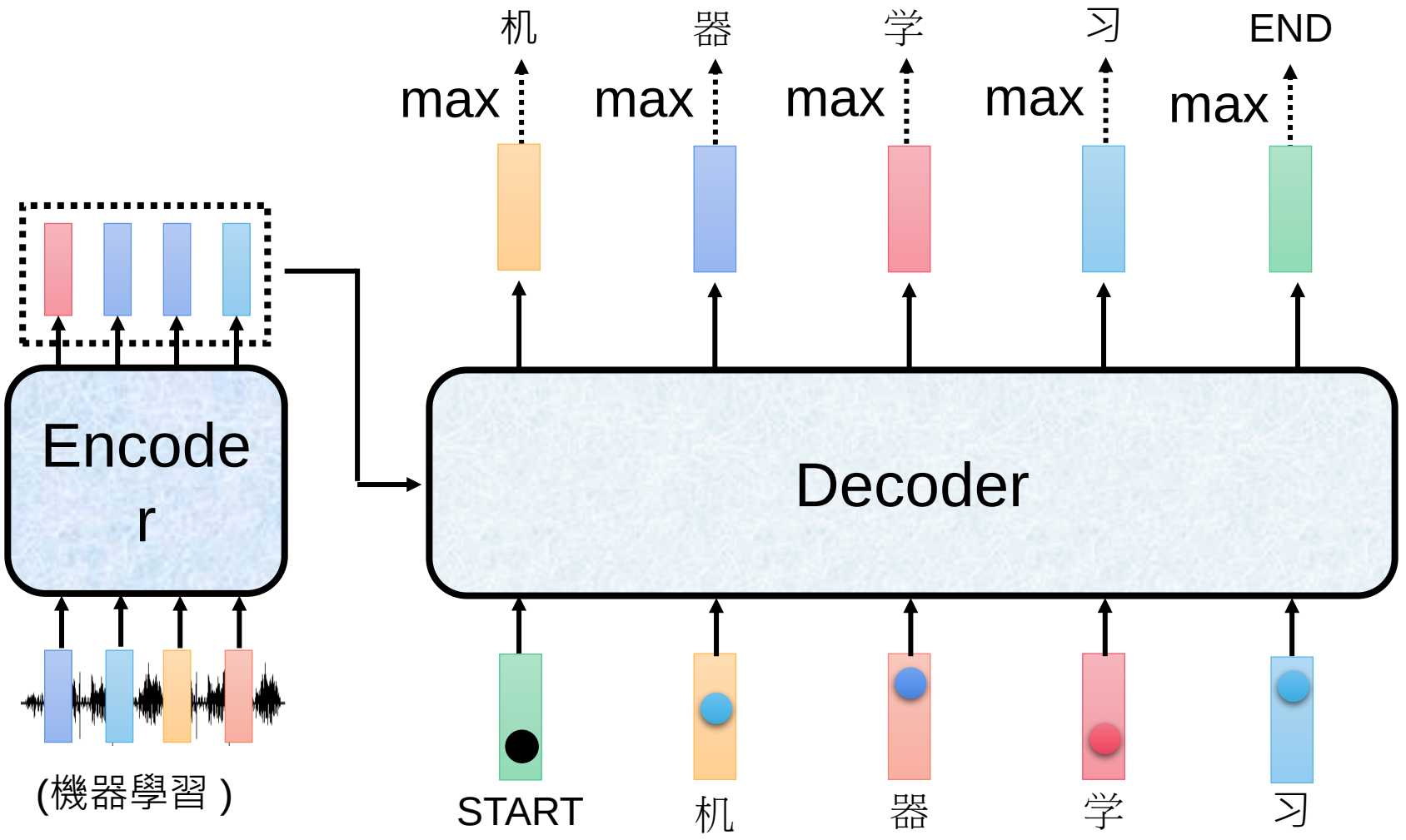


生成任务

训练和测试大为不同。

训练，虽有mask，但标签是直接作为输入的。

整体并行



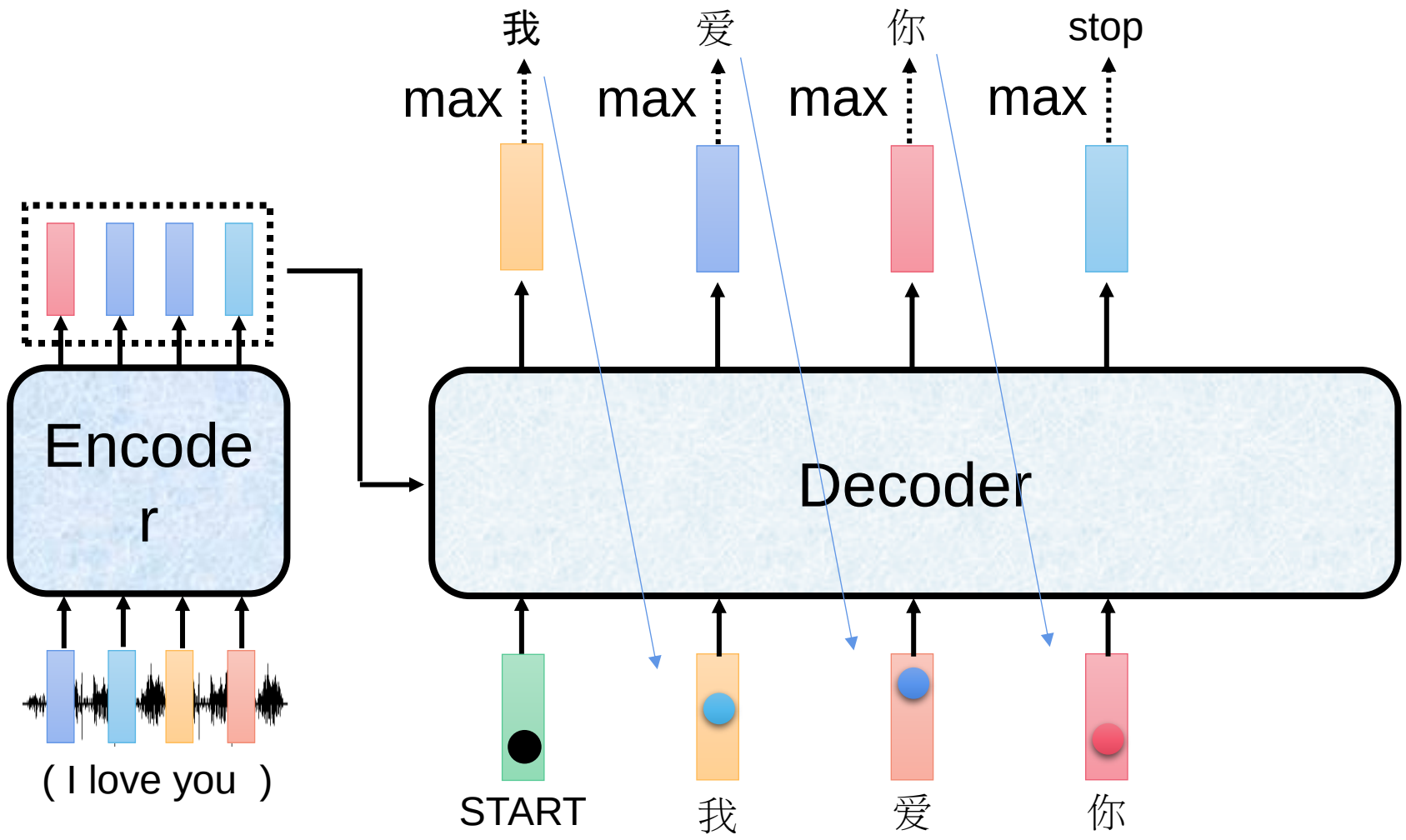
生成任务

训练和测试大为不同。

测试

只能一个字一个字来

整体串行

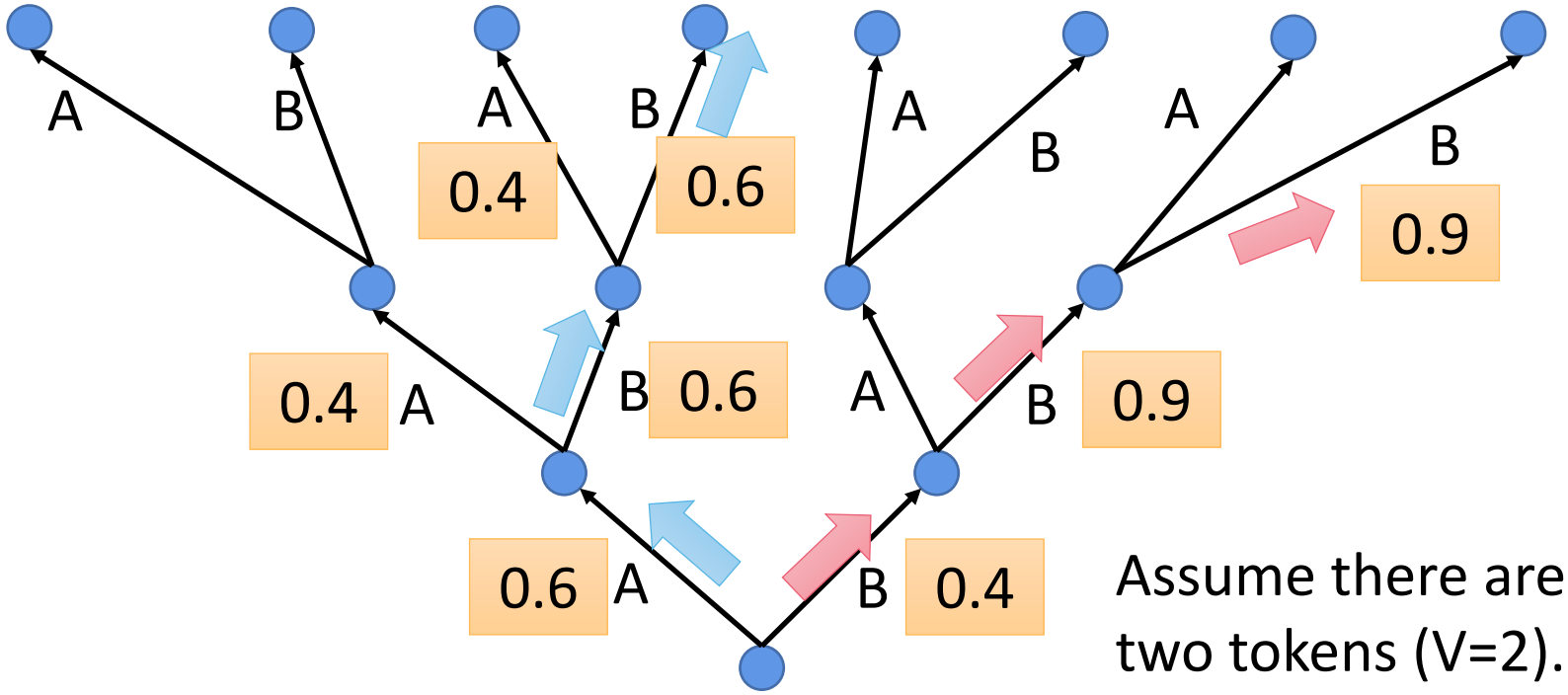


Beam Search

红色是最好策略

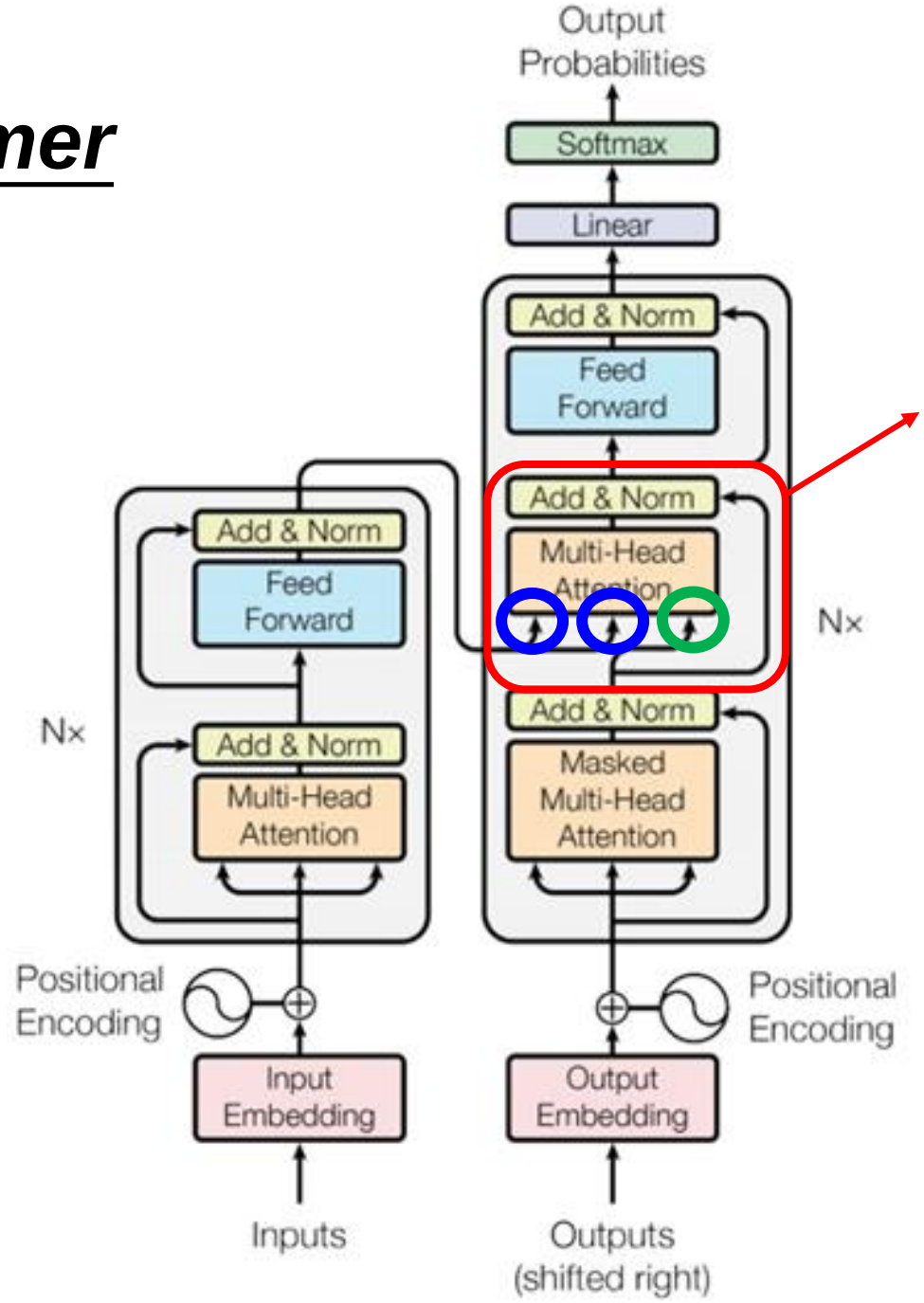
蓝色是贪婪策略

Not possible to check all the paths … → Beam Search

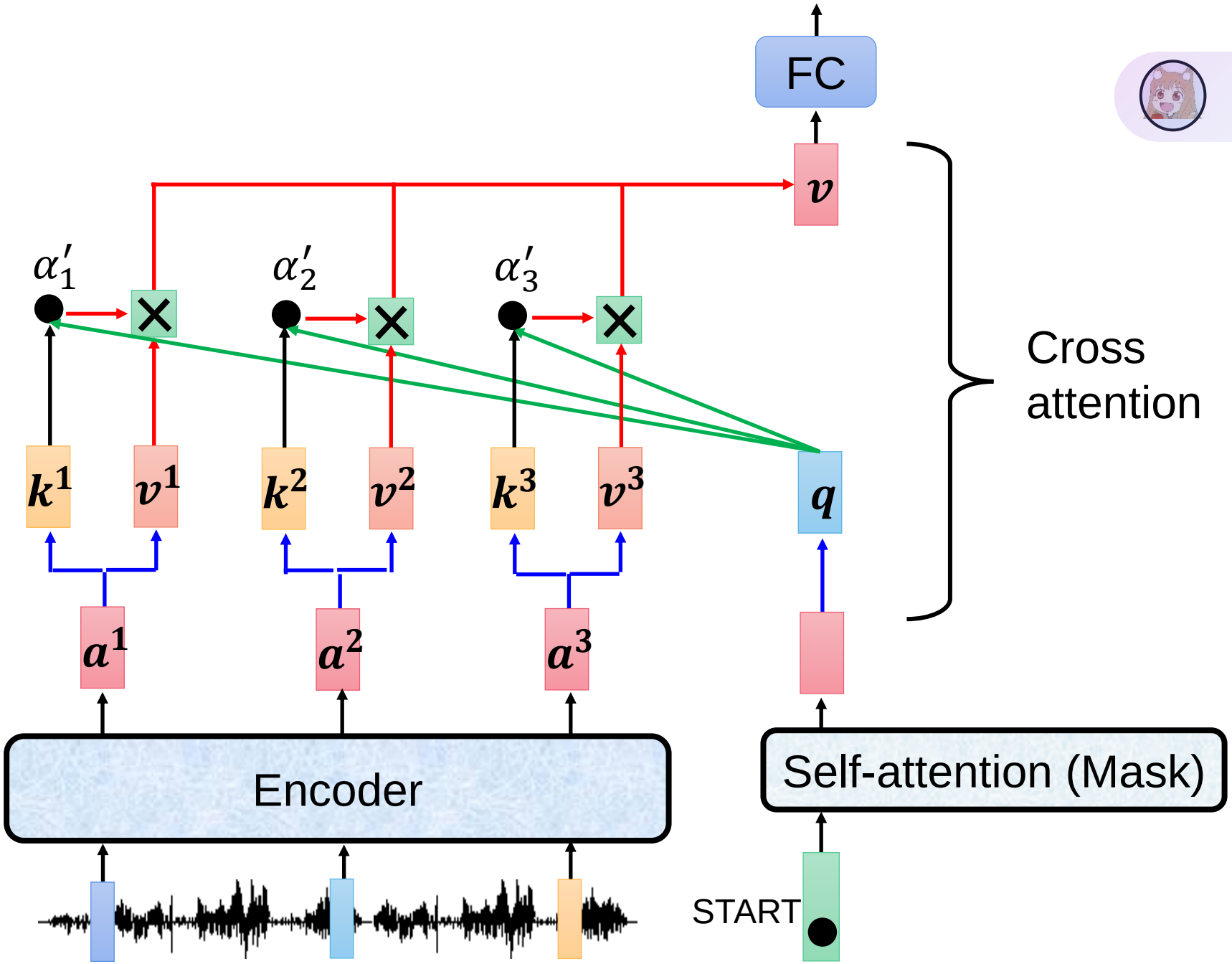


Assume there are only two tokens (V=2).

Transformer

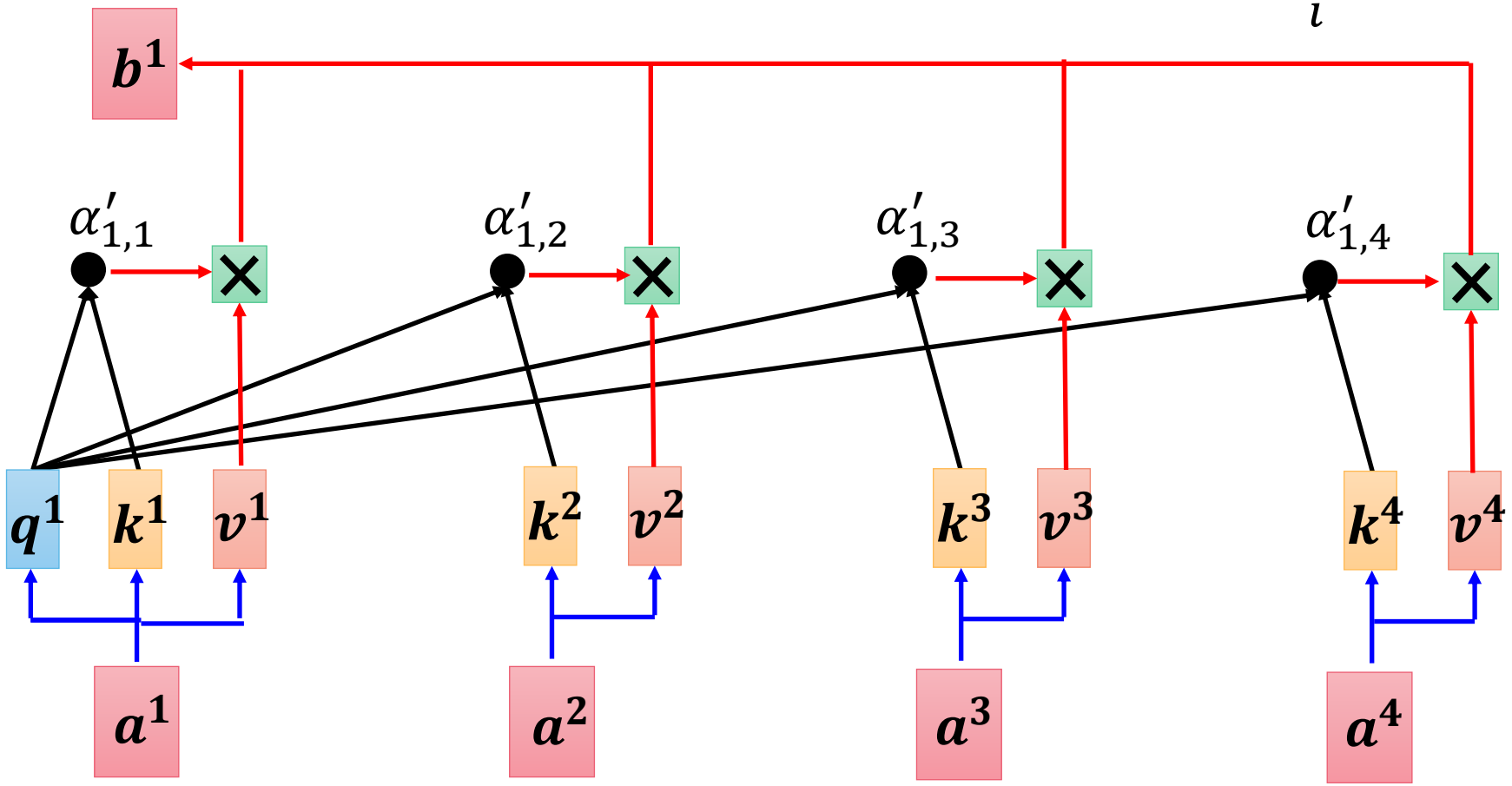


Cross attention

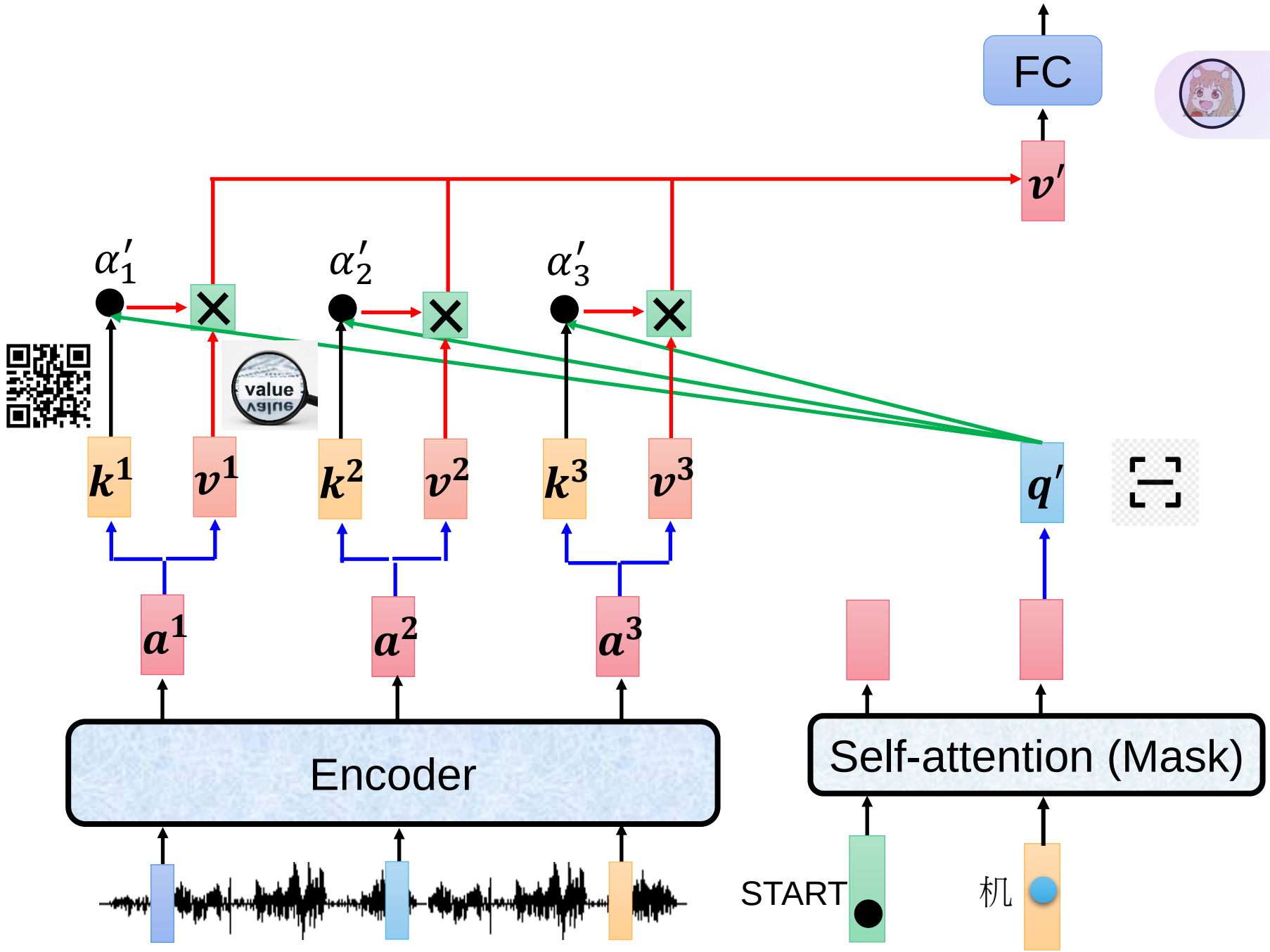

FC
v
α'_1
α'_2
α'_3
k^1
v^1
k^2
v^2
k^3
v^3
q
Cross attention
a^1
a^2
a^3
Encoder
Self-attention (Mask)
START

encoder

$$b^1 = \sum_i \alpha'_{1,i} v^i$$

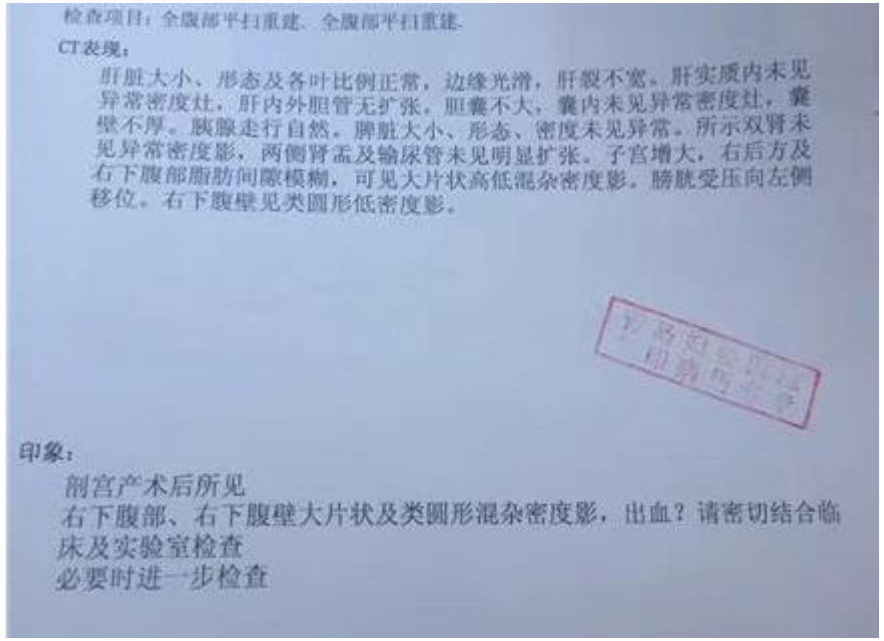


$$v^1 = W^v a^1 \qquad v^2 = W^v a^2 \qquad v^3 = W^v a^3 \qquad v^4 = W^v a^4$$


FC
v'
α'_1
α'_2
α'_3
value
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

带看一个生成项目

- 0,14 108 28 30 15 13 294 29 20 18 23 21 25 32 16 14 39 27 14 47 46 69 70 11 24 42 26 37 61 24 10 79 46 62 19 13 31 95 19 28 20 18 10 22 12 38 41 17 23 21 36 53 25 10,22 12 38 41 17 81 10


检查项目：全腹部平扫重建. 全腹部平扫重建.

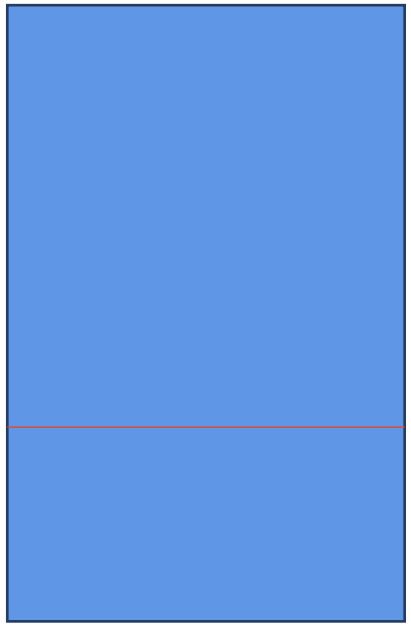
CT表现：

肝脏大小、形态及各叶比例正常，边缘光滑，肝裂不宽。肝实质内未见异常密度灶，肝内外胆管无扩张。胆囊不大，囊内未见异常密度灶，囊壁不厚。胰腺走行自然。脾脏大小、形态、密度未见异常。所示双肾未见异常密度影，两侧肾盂及输尿管未见明显扩张。子宫增大，右后方及右下腹部脂肪间隙模糊，可见大片状高低混杂密度影。膀胱受压向左侧移位。右下腹壁见类圆形低密度影。

印象：

剖宫产术后所见
右下腹部、右下腹壁大片状及类圆形混杂密度影，出血？请密切结合临床及实验室检查
必要时进一步检查

process_data.py



process_vocab.py

李哥考研

- 0,14 108 28 30 15 13 294 29 20 18 23 21 25 32 16 14 39 27 14 47 46 69 70 11 24 42 26 37 61 24 10 79 46 62 19 13 31 95 19 28 20 18 10 22 12 38 41 17 23 21 36 53 25 10,22 12 38 41 17 81 10

把这些数字转为id时

有些数字在词表里没有啊！！！

三种处理

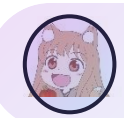
词表未见过词的方法

1， 直接数字当id

2， 直接加字

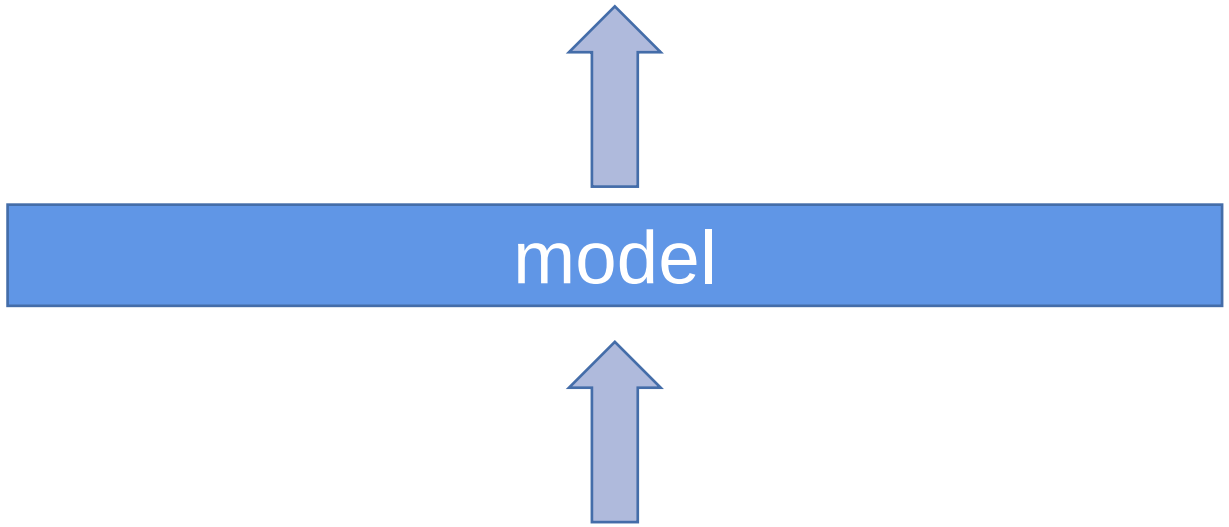
3， 重新制作词表

自监督预训练



李哥考研

去的尽管去了，来的尽管来着；去来的中间，又怎样地匆匆呢？早上我起来的时候，小屋里射进两三方斜斜的太阳。太阳他有脚啊，轻轻悄悄地挪移了；我也茫茫然跟着旋转。



没错，就是我

去的尽M去了，来的尽管来M；去来的中间，又怎样M匆匆呢？早上我起M的时候，小屋里射进两三方M斜的太阳。太M他有脚啊，轻轻悄悄地挪移了；我也茫茫然跟着旋转。

Bert-Pre-Training

李哥考研

Masked Language Model

80% : my dog is hairy → my dog is [MASK]

10% : my dog is hairy → my dog is apple

10% : my dog is hairy → my dog is hairy.

预训练任务

Next Sentence Prediction

Input = [CLS] the man went to [MASK] store [SEP]
he bought a gallon [MASK] milk [SEP]

Label = IsNext

Input = [CLS] the man [MASK] to the store [SEP]
penguin [MASK] are flight ##less birds [SEP]

Label = NotNext

BART模型考虑了以下五种噪声引入方式：

（1）**单词掩码**。与BERT模型类似，在输入文本中随机采样一部分单词，并替换为掩码标记（如[MASK]）；

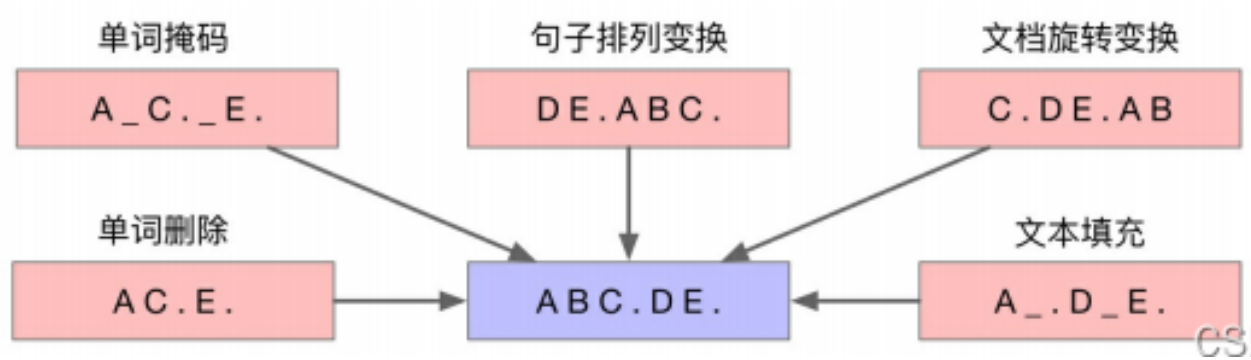
（2）**单词删除**。随机采样一部分单词并删除。要处理这类噪声，模型不仅需要预测缺失的单词，还需要确定缺失单词的位置；

（3）**句子排列变换**。根据句号将输入文本分为多个句子，并将句子的顺序随机打乱。为了恢复句子的顺序，模型需要对整段输入文本的语义具备一定的理解能力；

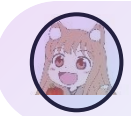
（4）**文档旋转变换**。随机选择输入文本中的一个单词，并旋转文档，使其以该单词作为开始。为了重构原始文本，模型需要从扰乱文本中找到原始文本的开头；

（5）**文本填充**。随机采样多个文本片段，片段长度根据泊松分布（λ=3）进行采样得到。用单个掩码标记替换每个文本片段。当片段长度为0时，意味着插入一个掩码标记。要去除这类噪声，要求模型具有预测缺失文本片段长度的能力。

下图对这五类噪声进行了概括：



预训练过程



李哥考研

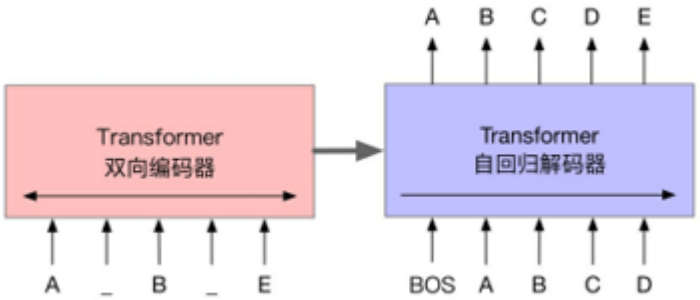
数据

14　　108　　28　　30　　15

x: 14　　108　　28　　103　　15

Y: -100，-100，-100，30，-100

模型



训练

微调过程

李哥考研

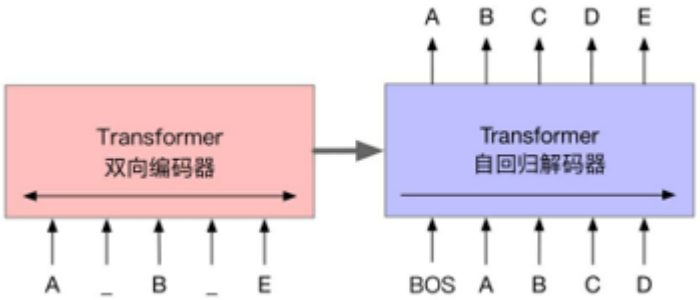
数据

14　　108　　28　　30　　15

14　　108　　28　　103　　15

-100，-100，-100，30，-100

模型



训练

CiderD_scorer

李哥考研

```
refs = {
    'image1': ['a man is surfing on a wave', 'surfing man on a big wave'],
    'image2': ['a dog is running in the park', 'dog running in grassy park’]
}


cand = {
    'image1': 'man surfing on ocean wave',
    'image2': 'dog runs in the park'
}


  scorer = CiderD(df='corpus')

  # 计算得分
  score, scores = scorer.compute_score(refs, cand）
```

下节预告。

李哥考研

Vit与多模态与大模型

作业：

完成外卖数据集用bert分类。

用调试跑一遍bert的维度变化。

答疑和结束

THANKS