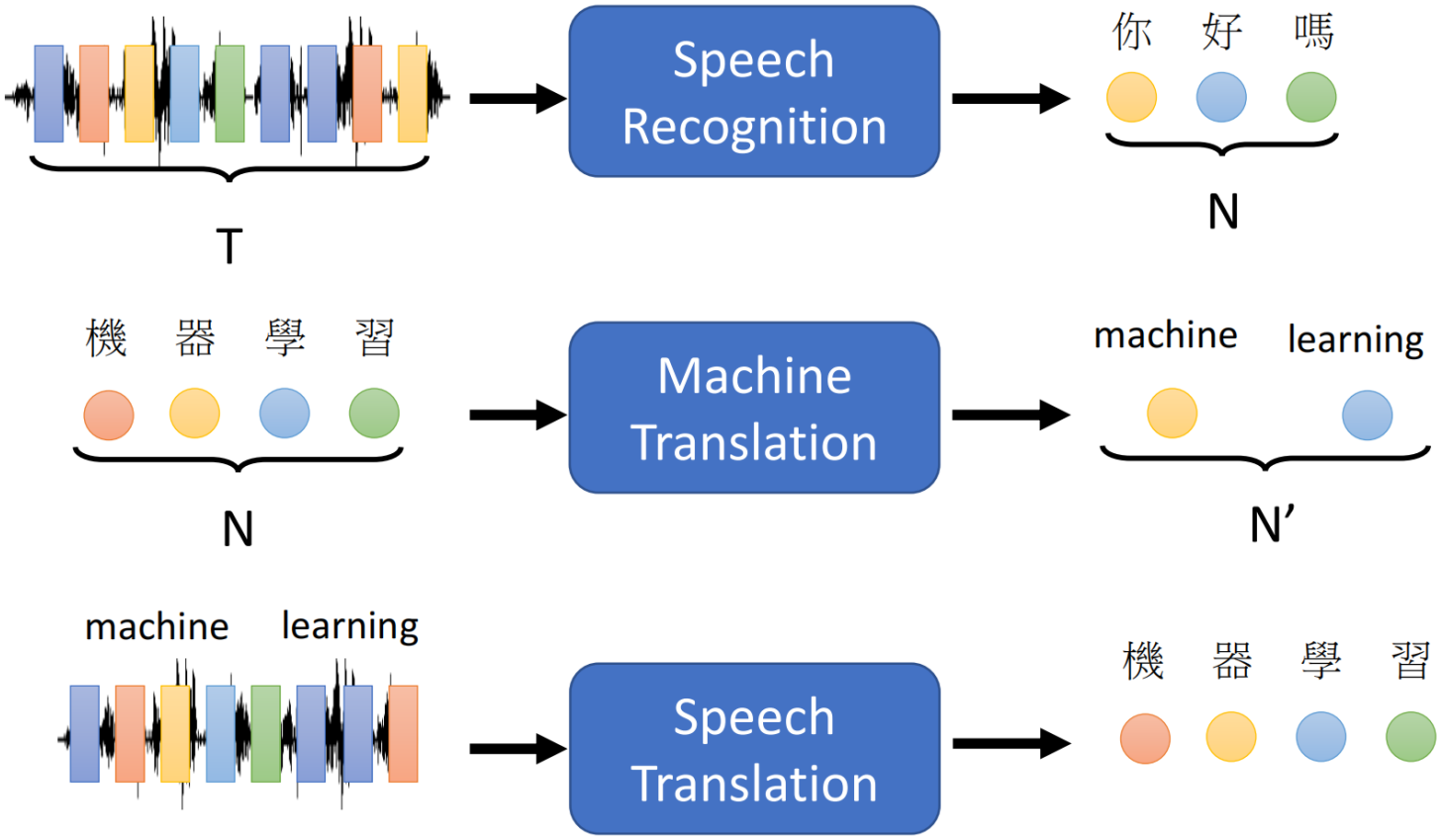
Sequence-to-sequence (Seq2seq)

東南大學 SOUTHEAST UNIVERSITY

Seq2Seq模型：输入一段序列，输出一段序列

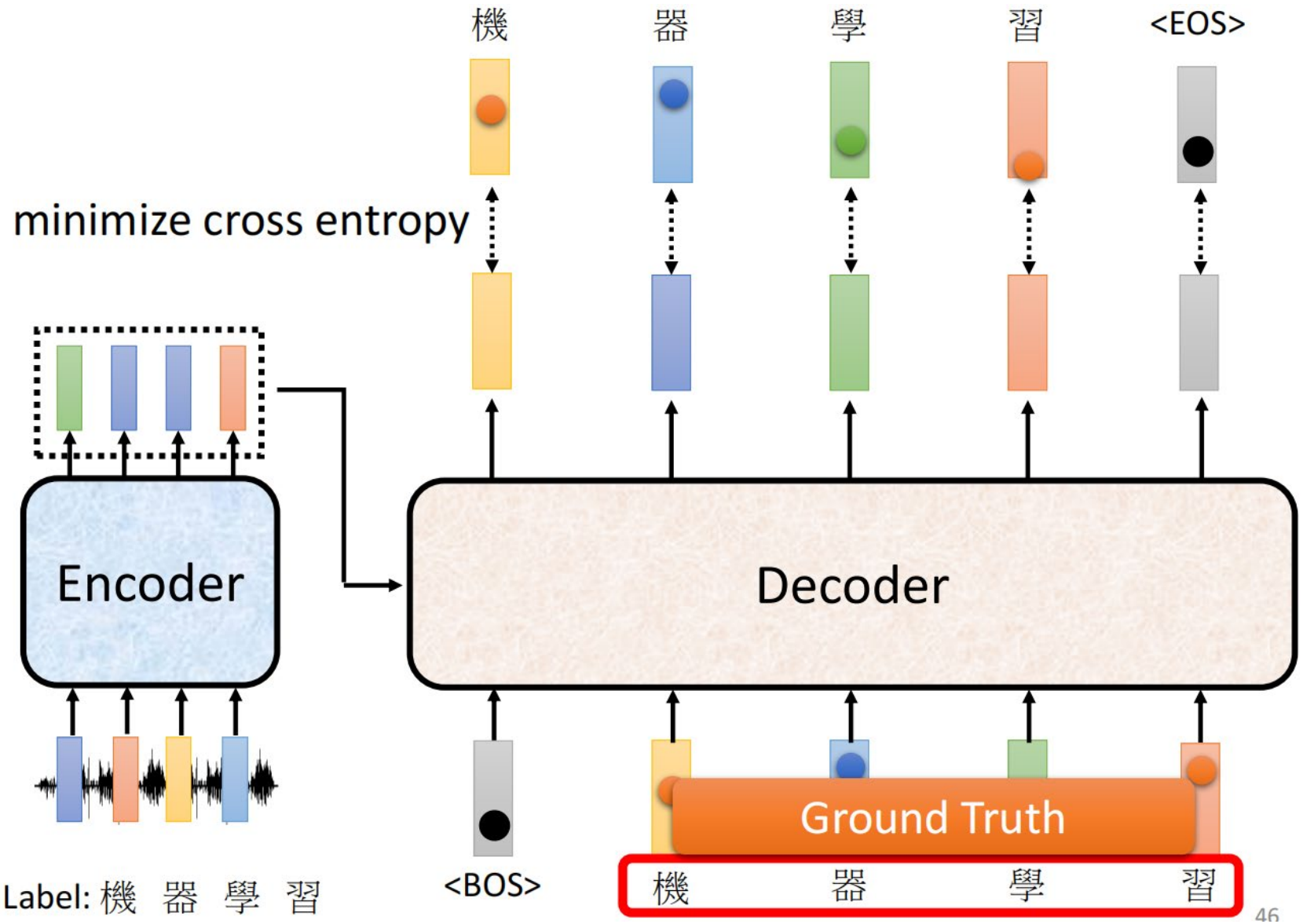
Input a sequence, output a sequence

The output length is determined by model.



Transformer本质也为Seq2Seq模型

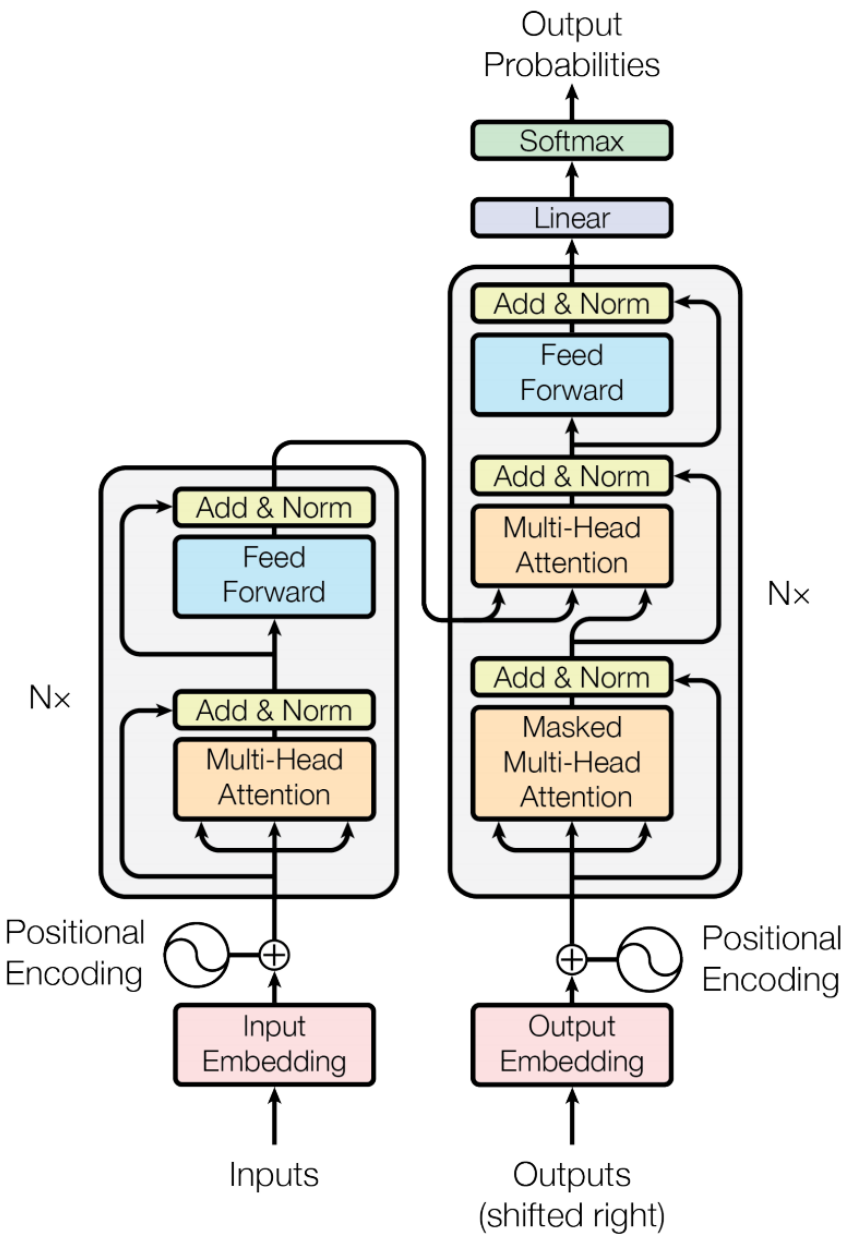
整体架构


機
器
學
習
<EOS>
minimize cross entropy
Encoder
Decoder
Ground Truth
<BOS>
機
器
學
習
Label: 機 器 學 習

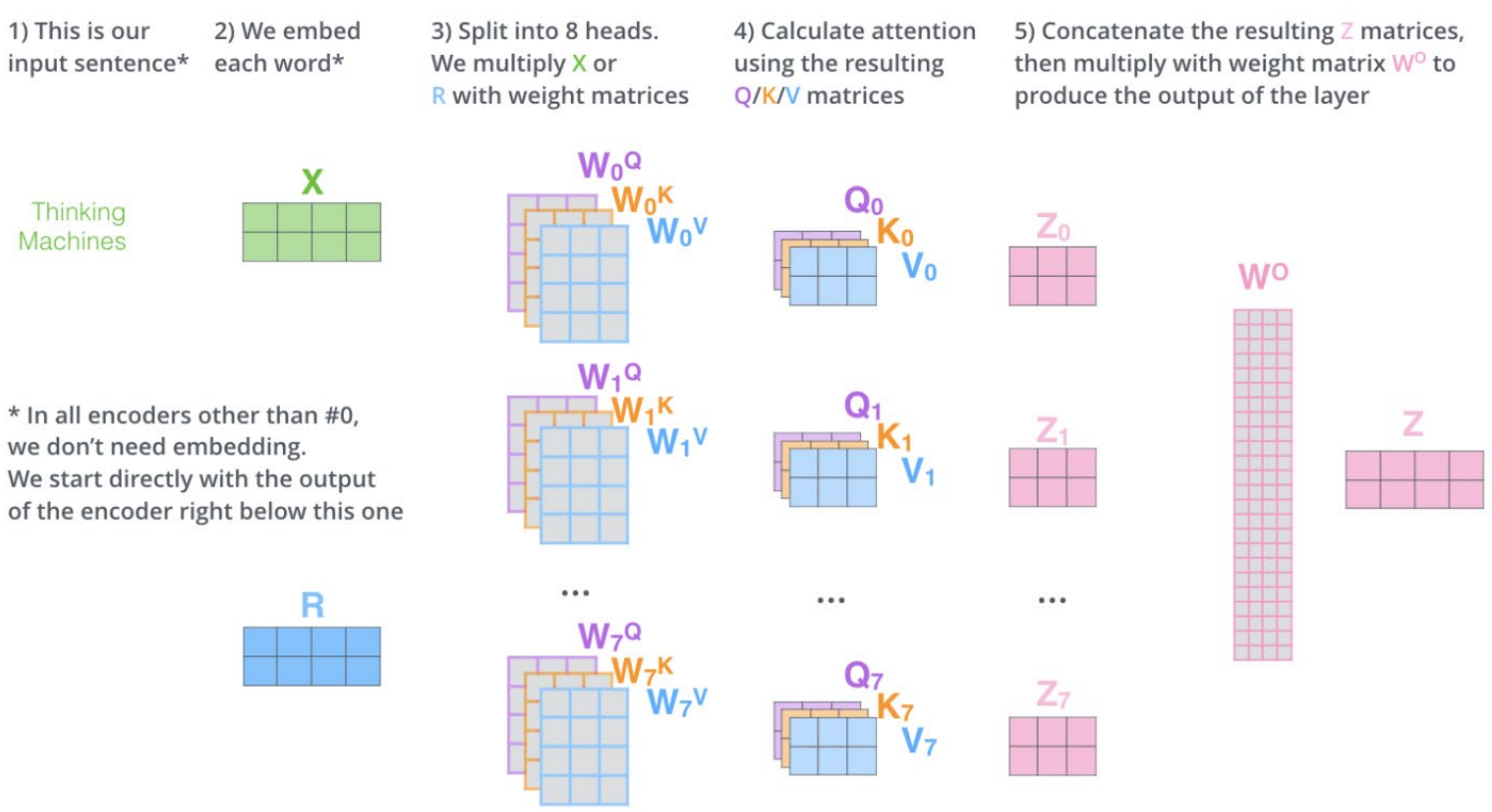
Transformer架构

Encoder：

- 以语音识别任务为例，一句话为一个**batch**，输入为一句话（n，512），n表示单词个数，每个单词经过embedding层得到512的向量，再加上位置编码（融入每个单词位置信息）。
- （n，512）复制三次，分为三个方向输入**多头注意力层**（其为**self-attention**的多头版本）；每个方向计算Q，K，V三个矩阵（得形状为（n，64）），每个头输出为（n，64），这里得的每个n是原输入样本的加权，因为多头，所以最终输出为（n，64*8=512）



Self-attention

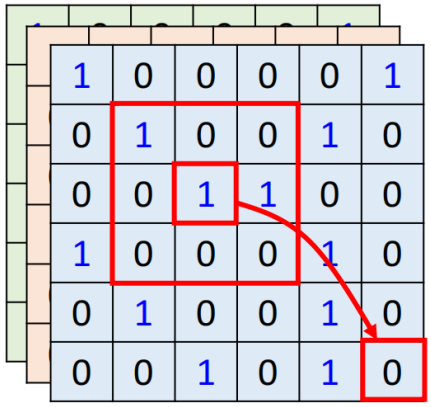


$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})W^O$$

$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

Self-attention vs CNN



CNN: self-attention that can only attends in a receptive field

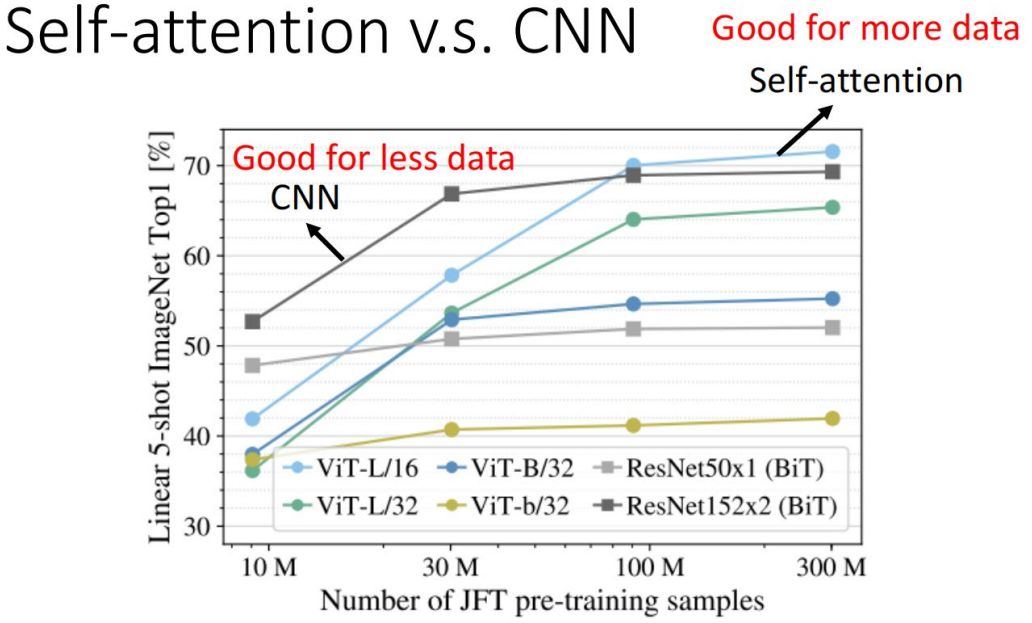
- CNN is simplified self-attention.

Self-attention: CNN with learnable receptive field

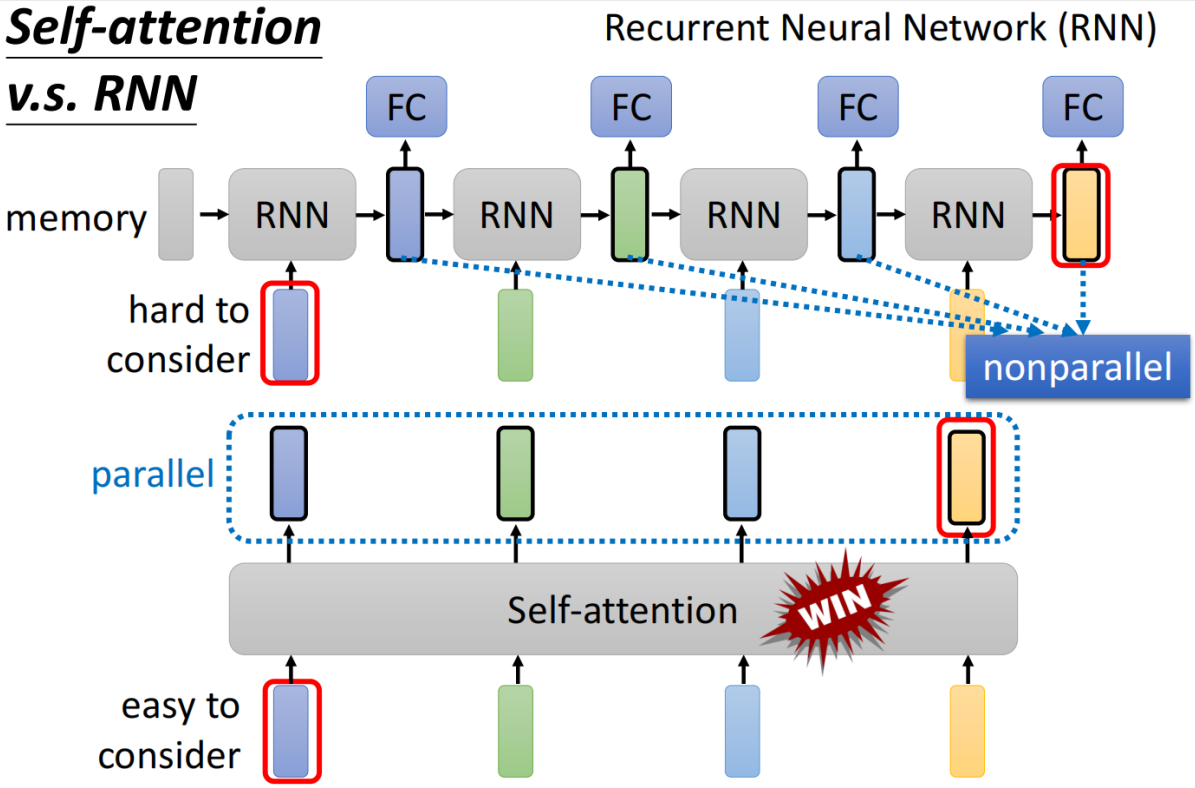
- Self-attention is the complex version of CNN.

结论：

- CNN是Self-attention特例（有严格数学证明）
- CNN适用于少数据量，self-attention适用于大数据量。因为transformer参数更多。



Self-attention vs RNN



口 结论：

- Self-attention解决了RNN的**长期依赖**问题。
- Self-attention可以**并行训练**，运算更快。

Encoder：

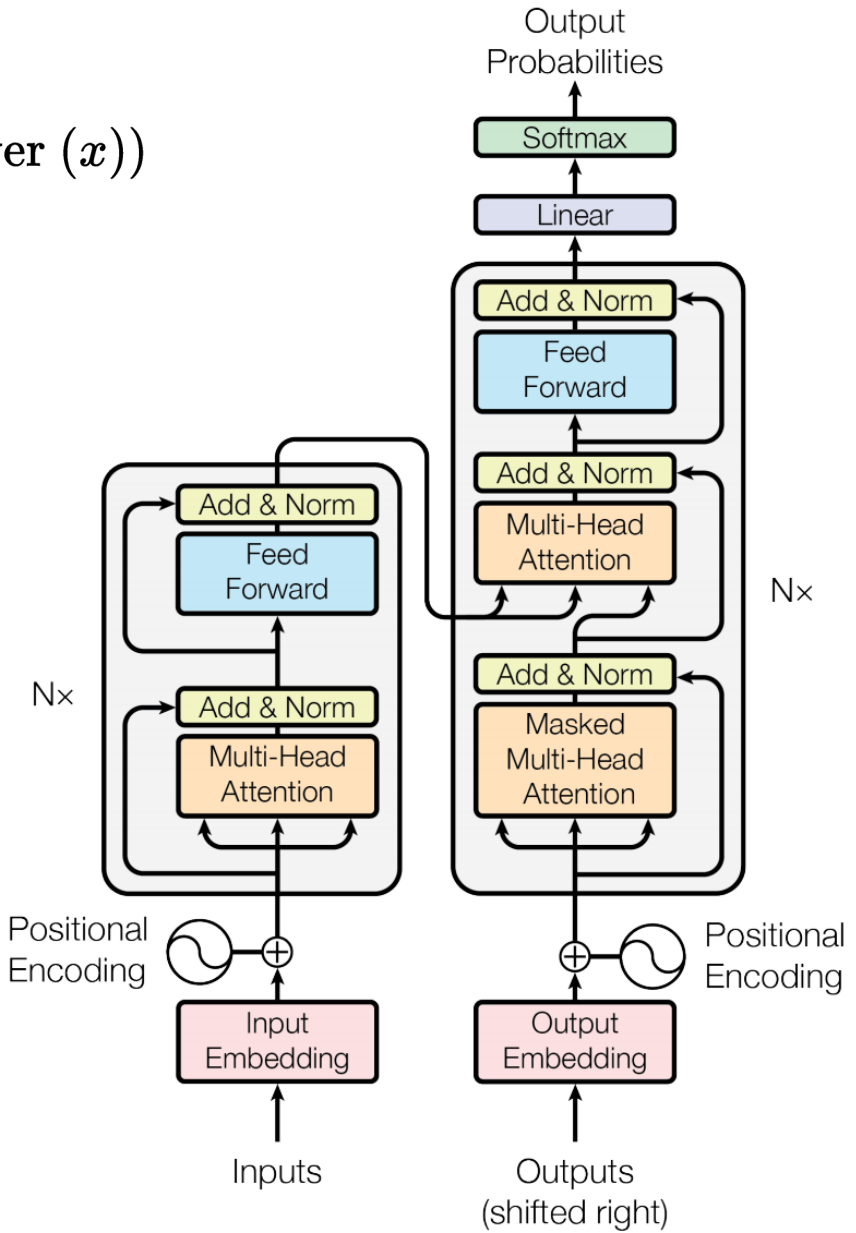
- **Add & Norm层：** $\mathrm{LayerNorm}\,(x+\mathrm{Sublayer}\,(x))$

  x为多头的输入，**LN**在每一个批量内正则化。

- **Feed Forward层：**

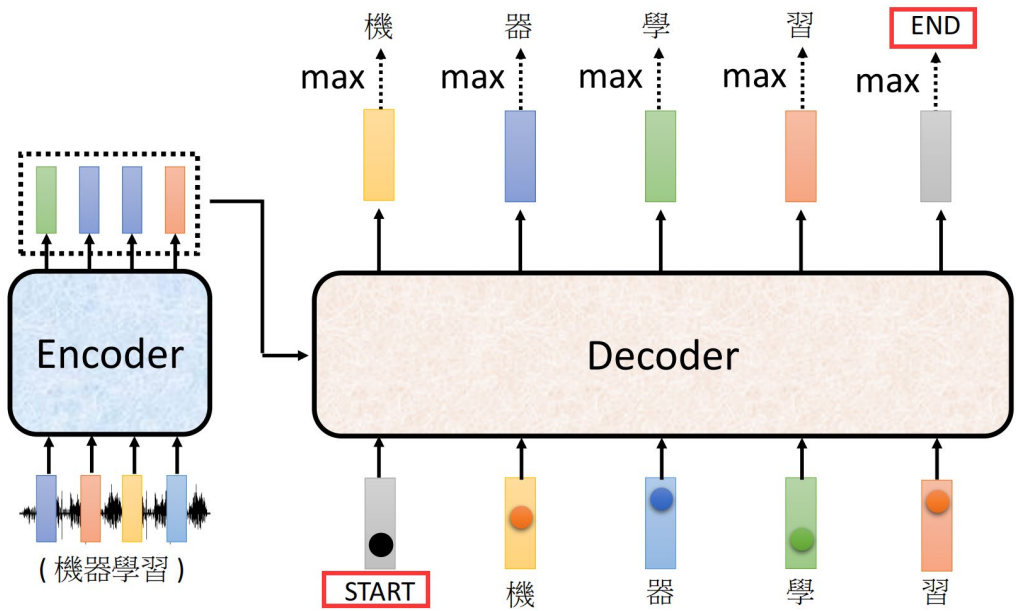
$$\mathrm{FFN}(x)=\max(0,xW_1+b_1)W_2+b_2$$

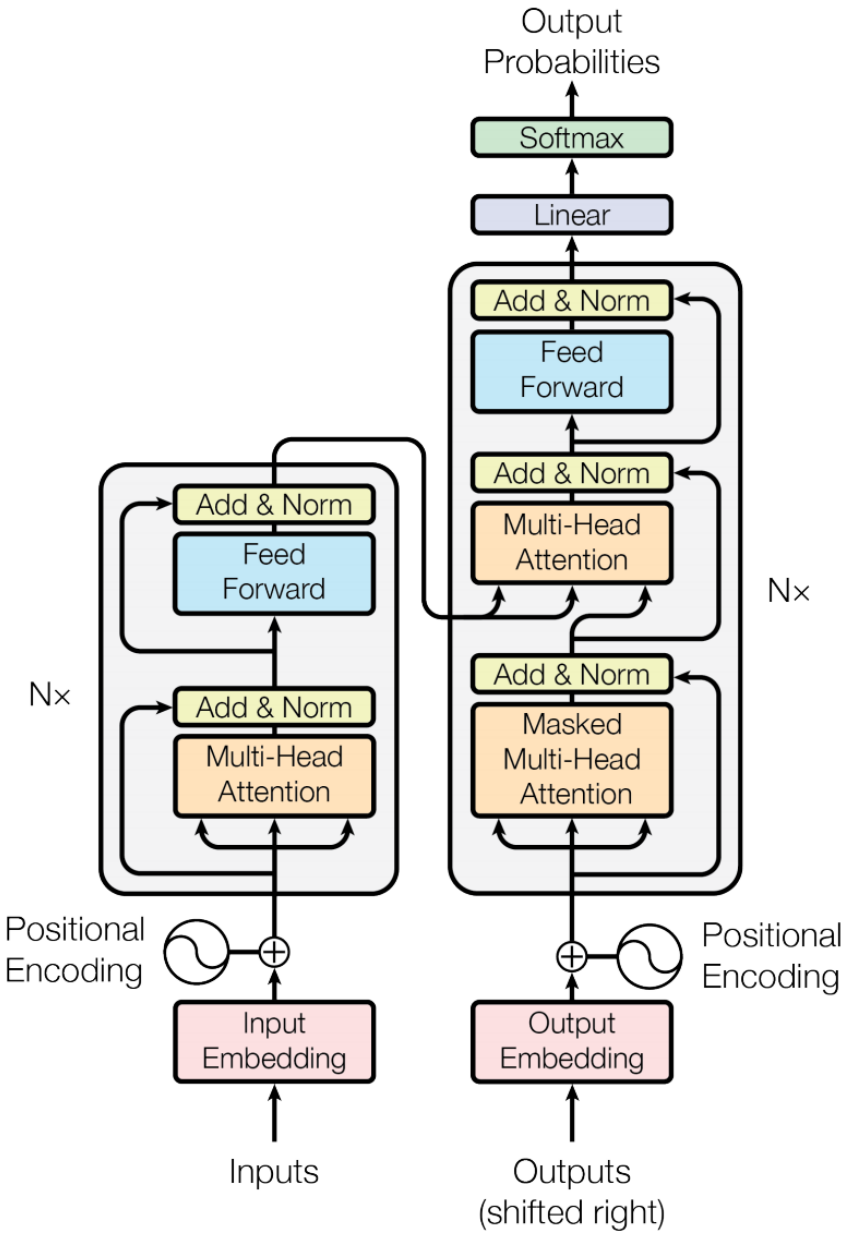
- 整个Encoder**重复N次**子结构块。



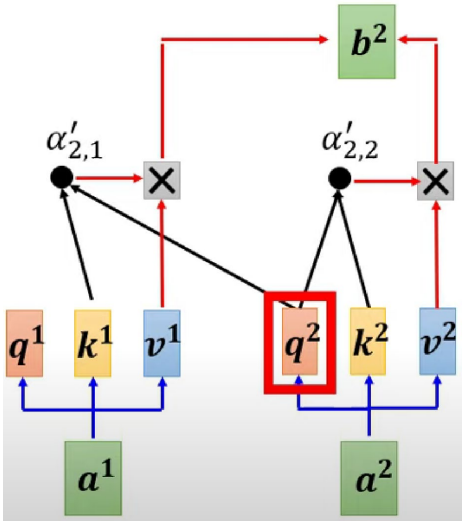
Decoder：

- 下端**输入**为：一个词（START），经过embedding为（1，512）；Decoder**输出**为（1，512）为概率，再将输出输入Decoder，这样一直做下去，直到输出END停止。

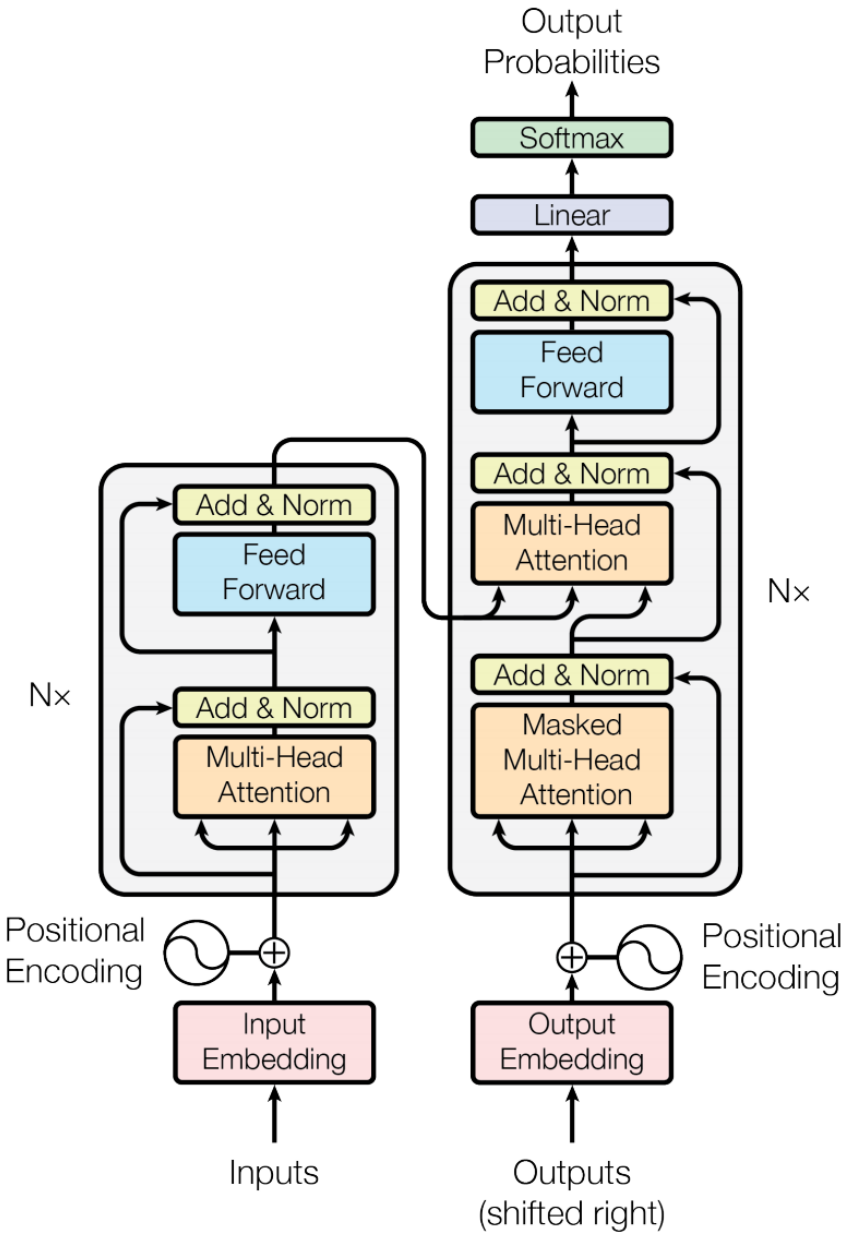




- 在**训练阶段**，每次Decoder输入为输出对应的标签的one-hot；**测试阶段**输入为输出。
- **Mask多头注意力机制**：

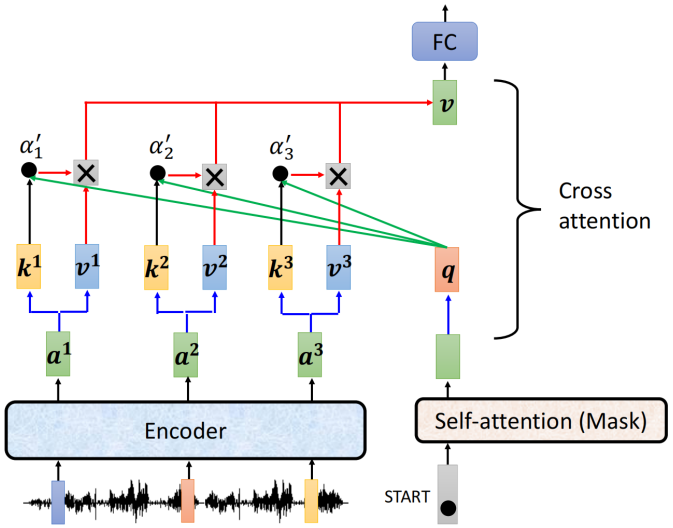


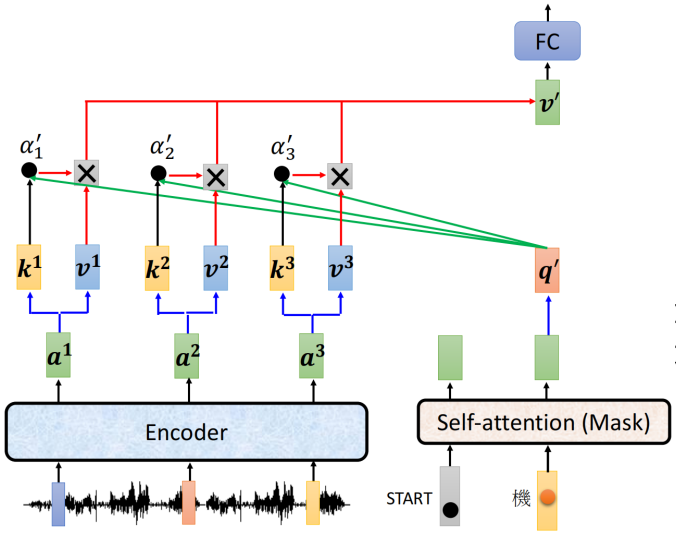
因为解码过程中会依次输入a^1，a^2等，在输入a^n时只使用前n项的q和k，不使用后面的输入，这与encoder在经过多头后得到每一项在每个输入上的加权不同。

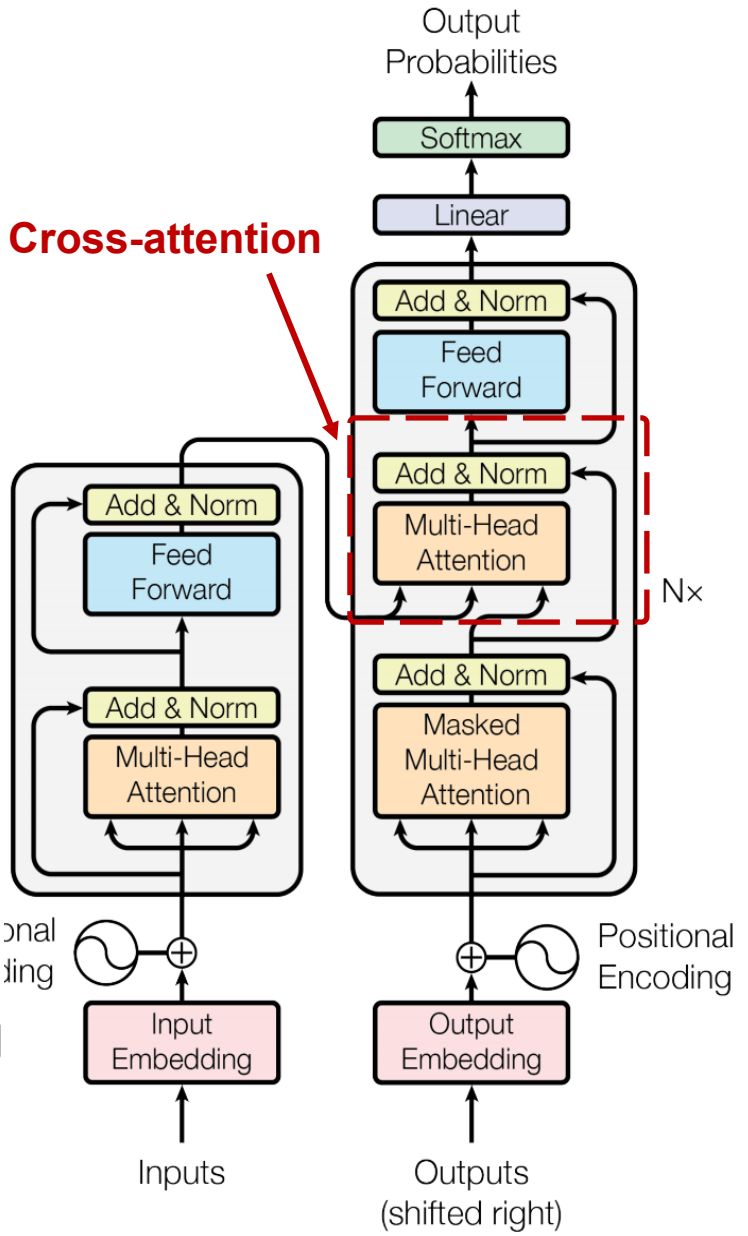


➢ Cross-attention：

每个q（1，64）和多个k（1，64）点乘得权重。







优缺点及作用

- **优：**
  - Self-attention解决了RNN的**长期依赖**问题。
  - Self-attention可以**并行训练**，运算更快。
  - Self-Attention模型更**可解释**，Attention结果的分布表明该模型学习到了一些语法和语义信息。
- **缺：**
  - 由于抛弃RNN和CNN使模型丧失了捕捉**局部特征**的能力，RNN + CNN + Transformer的结合可能会带来更好的效果。
  - Transformer失去的**位置信息**其实在NLP中非常重要，而论文中在特征向量中加入Position Embedding也只是一个权宜之计，并没有改变Transformer结构上的固有缺陷。
- **作用：**Transformer的核心结构块为**self-attention**，其可以取代**CNN中卷积核**和**RNN中循环核**，在CV和NLP领域取得许多sota结果。