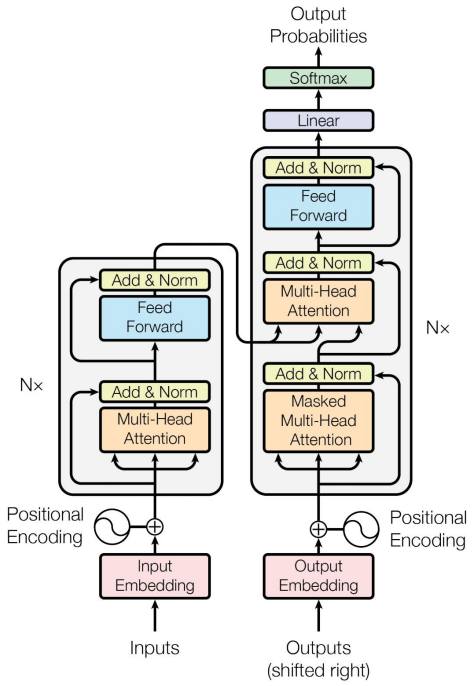
# 第09章 注意力机制

欧 新 宇

Thanks to LiMu-CS329 and Hung-yi Lee-MachineLearning


TRANSFORMERS
THE LAST KNIGHT

- Transformer 的架构概述
- Transformer 的模块
- Transformer 的训练和预测

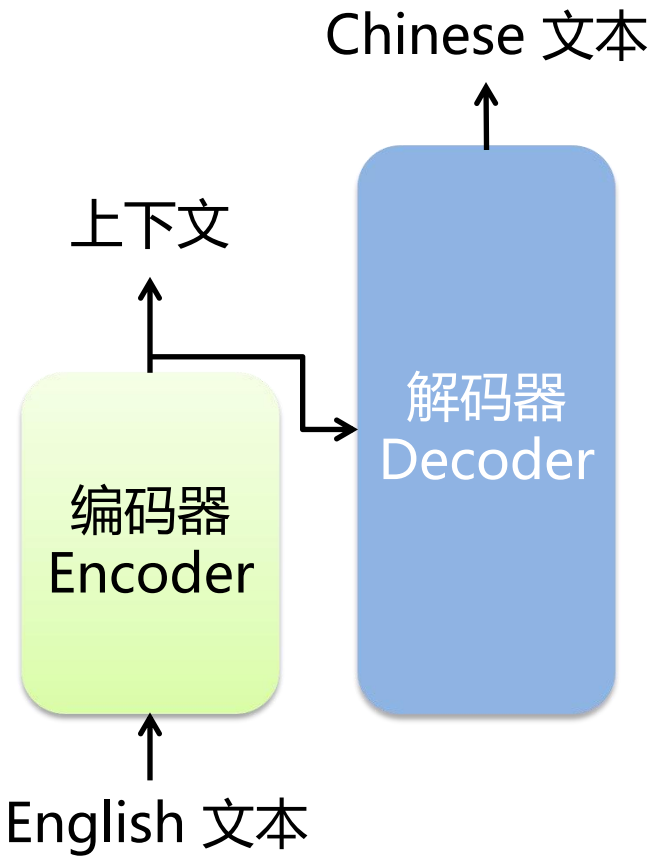



# Transformer的架构概述

Ahish Vaswani, Noam Shazeer, Niki Parmar, et.al. Attention is all you need. NIPS2017.

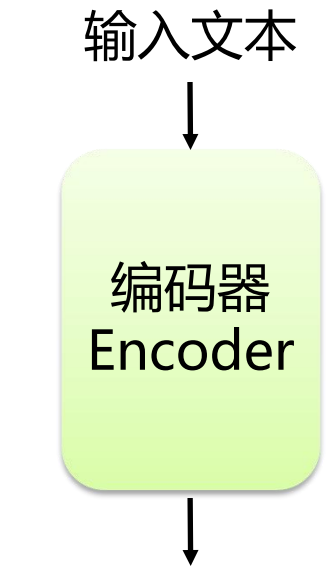
# Transformer的架构概述

## 原始 Seq2Seq




*将英文翻译成中文*

## 编码器模型




*Basis*：

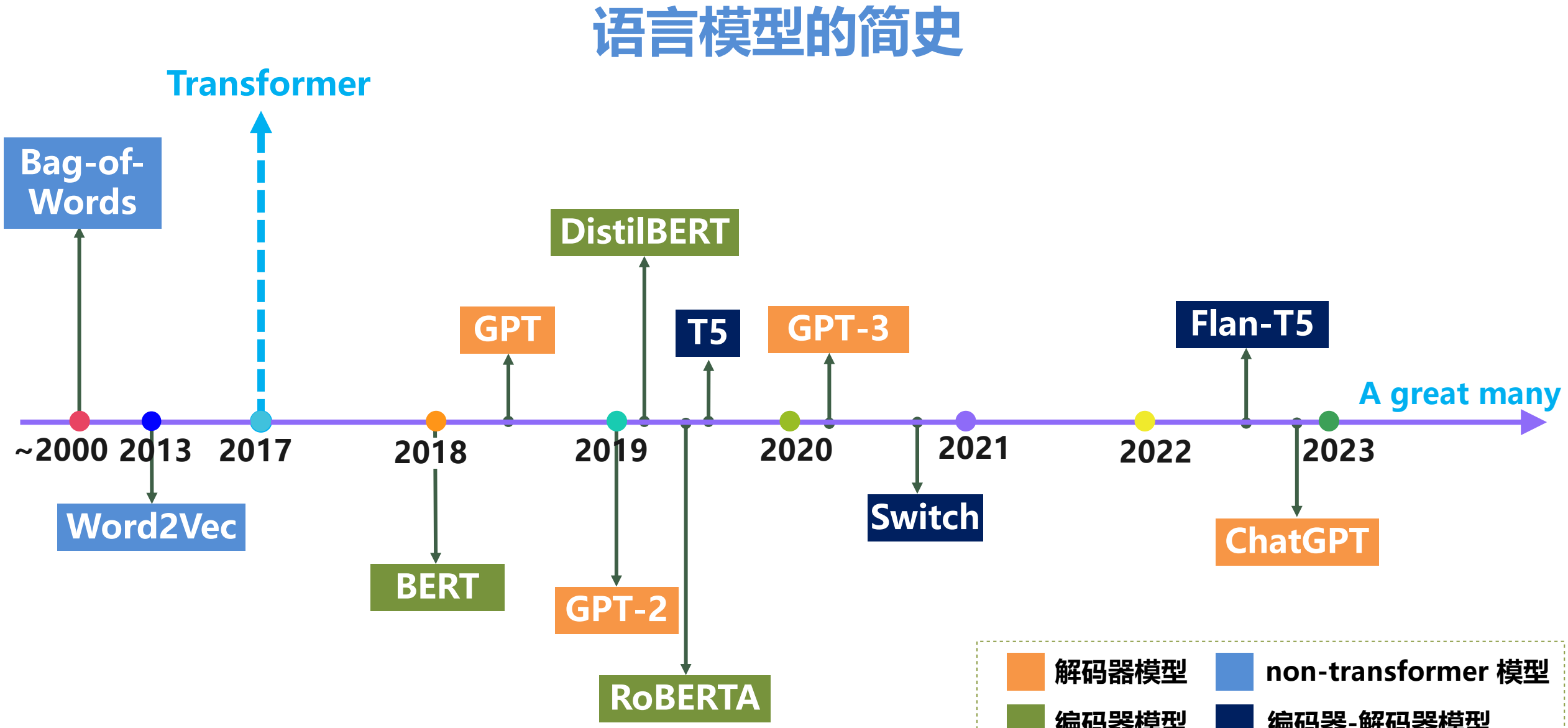
- BERT
- 大多数 embedding 模型

## 解码器模型


*Basis*：

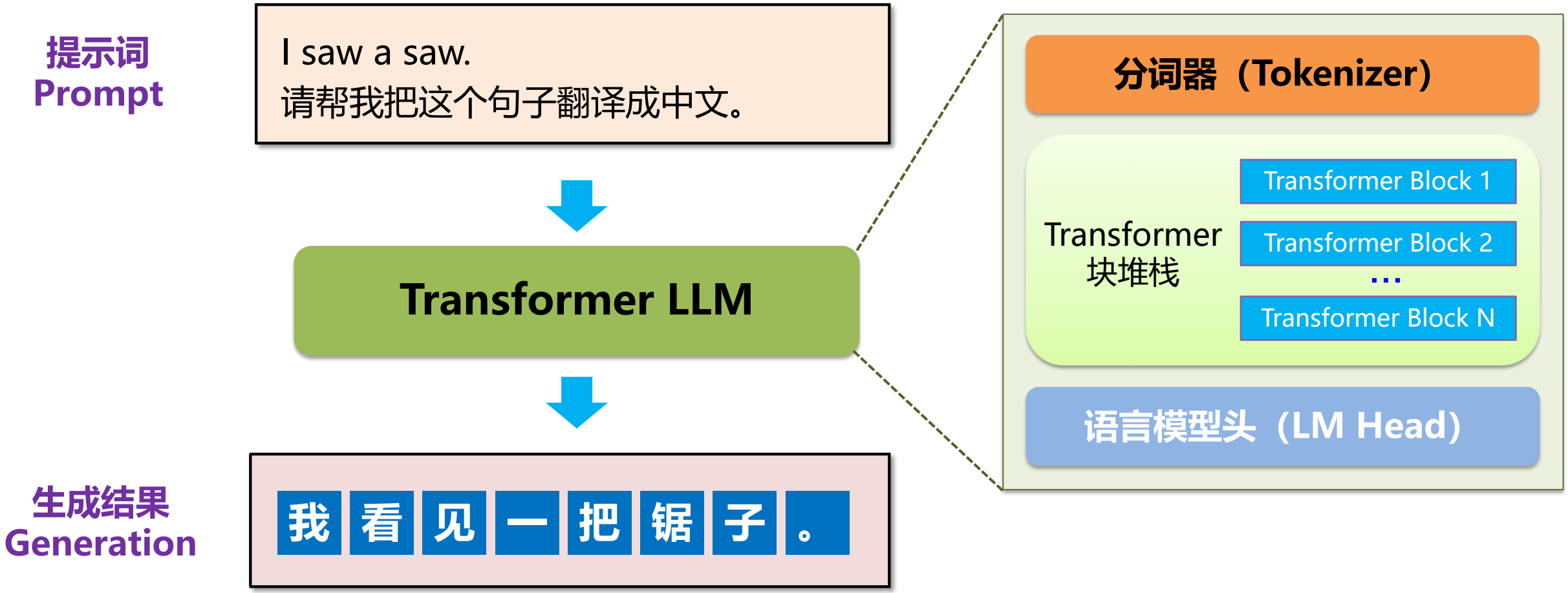
- 主流LLMs模型
- GPT、Claude、文心等

# Transformer的架构概述



# Transformer的架构概述

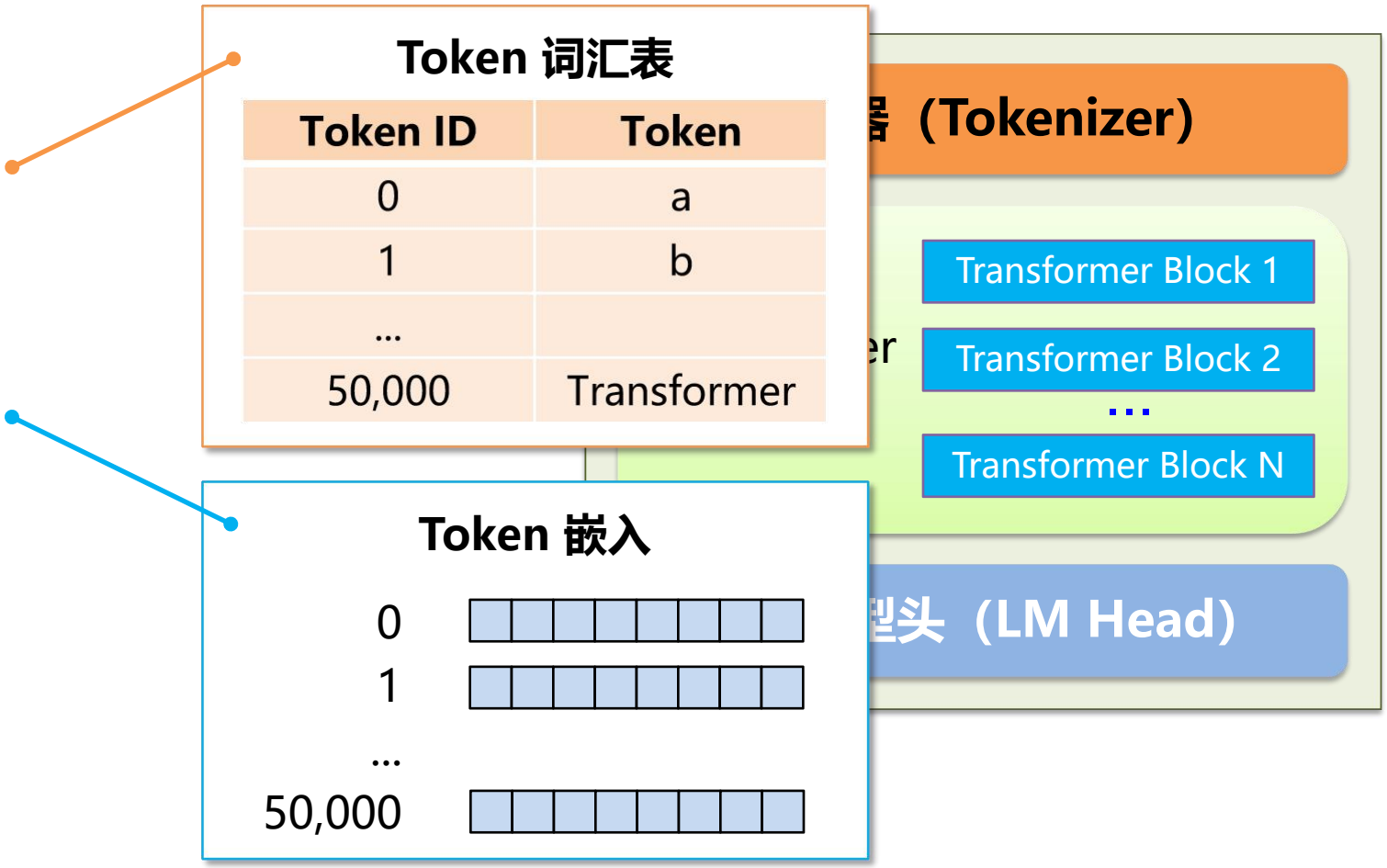
## Transformer 架构的主要组件



# Transformer的架构概述

## Transformer 架构的主要组件

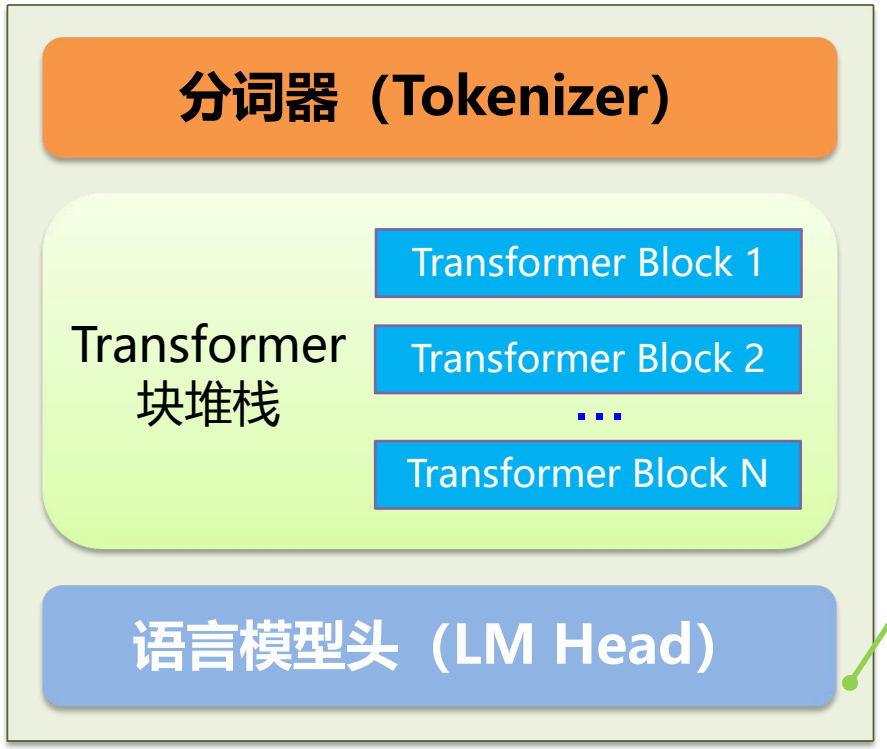
*Transformer LLM*



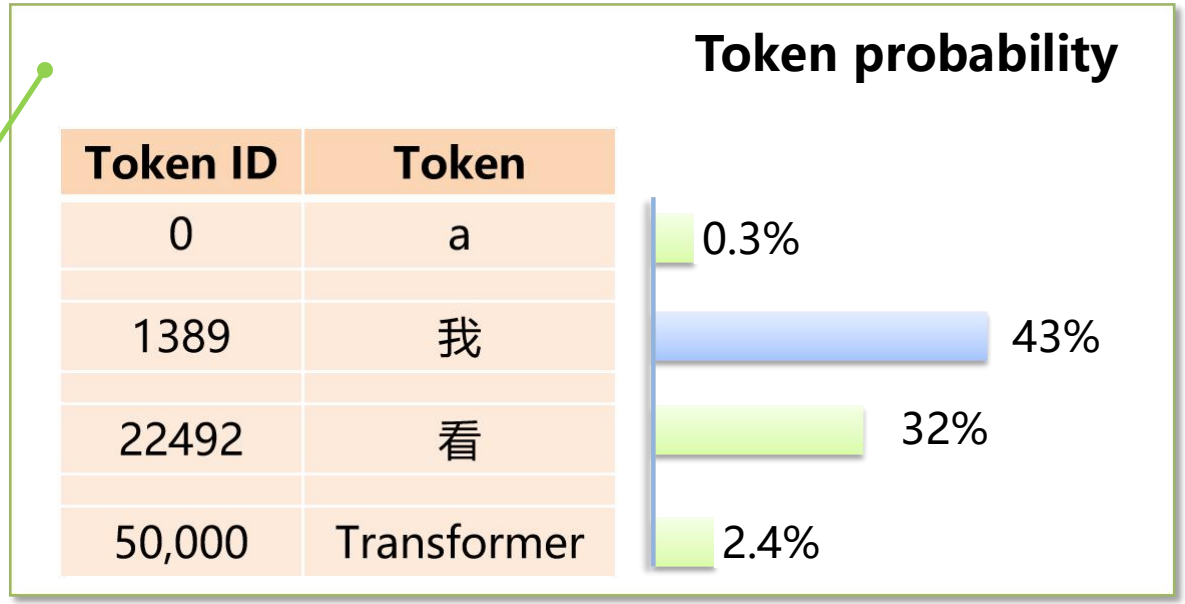
# Transformer的架构概述

## Transformer 架构的主要组件

*Transformer LLM*




$$P(y_i) = \frac{e^{z_i/T}}{\sum_j e^{z_j/T}}$$

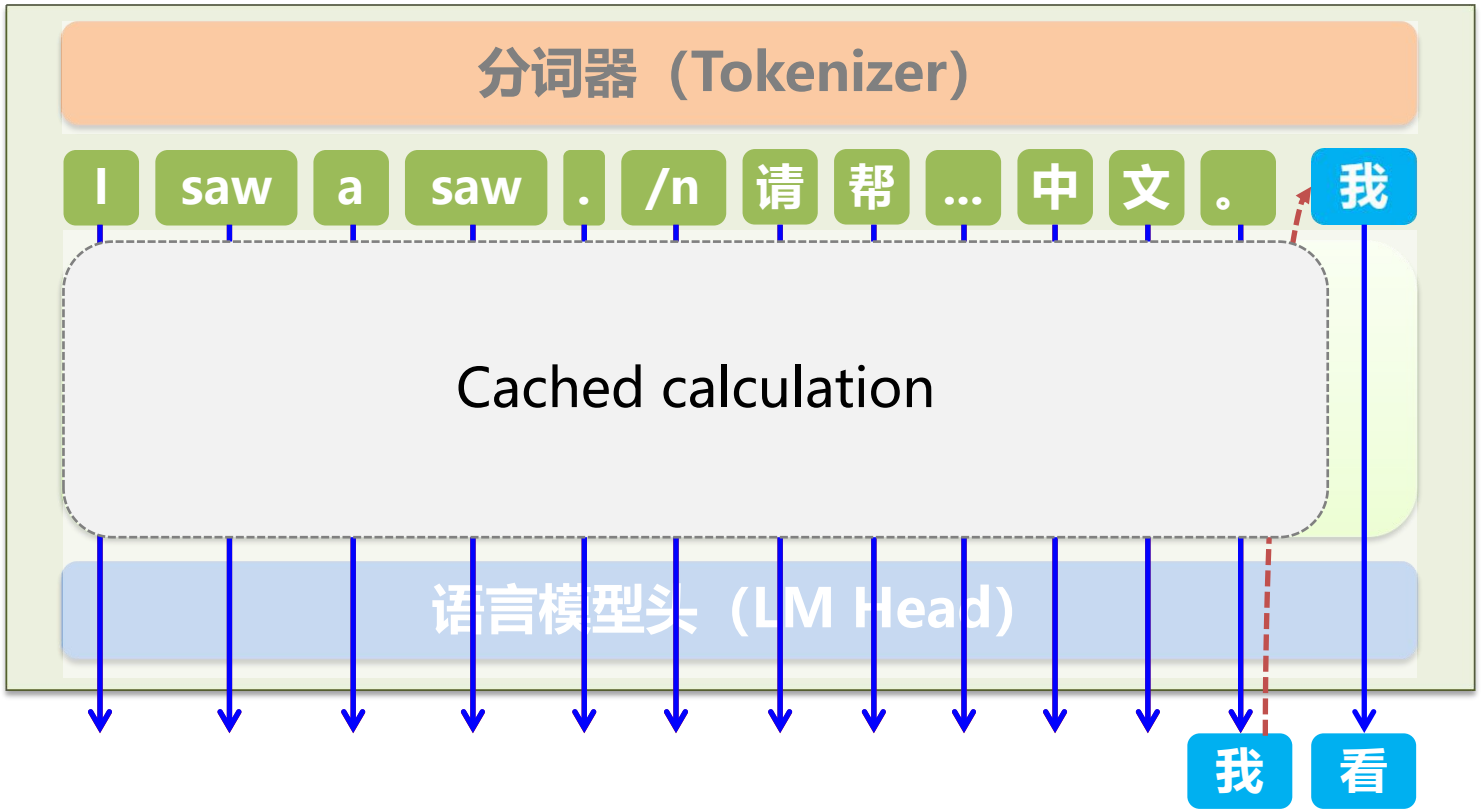


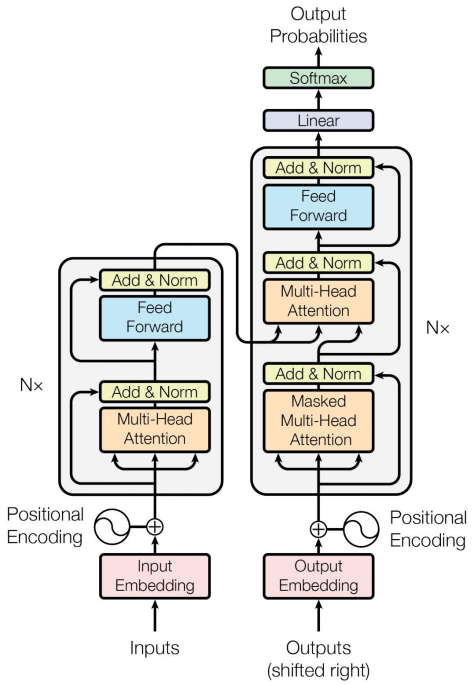
# Transformer的架构概述

## Transformer 架构的主要组件

I saw a saw.
请帮我把这个句子翻译成中文。

*Transformer LLM*



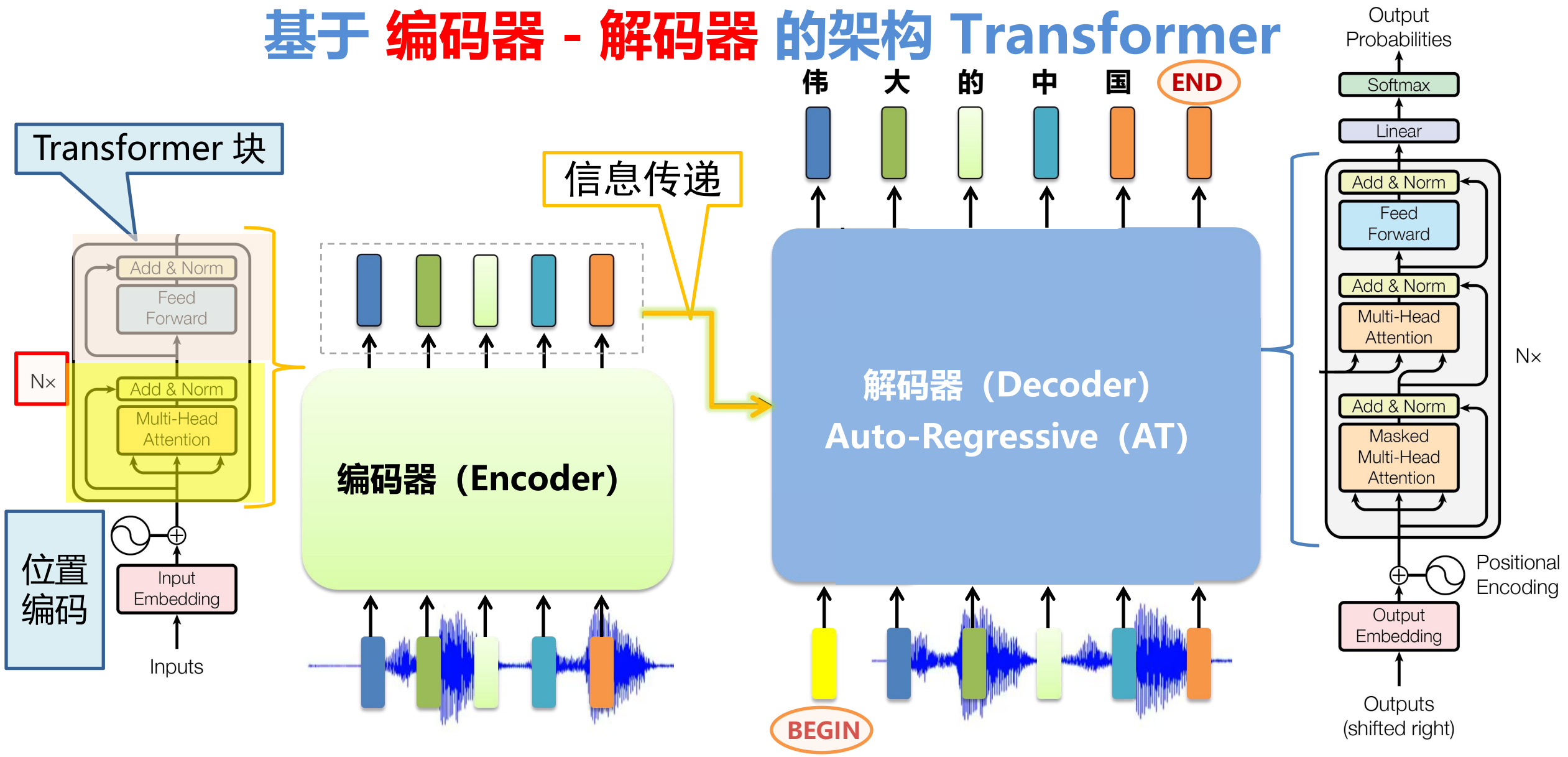



# Transformer 的模块

Ahish Vaswani, Noam Shazeer, Niki Parmar, et.al. Attention is all you need. NIPS2017.

# Transformer的架构概述

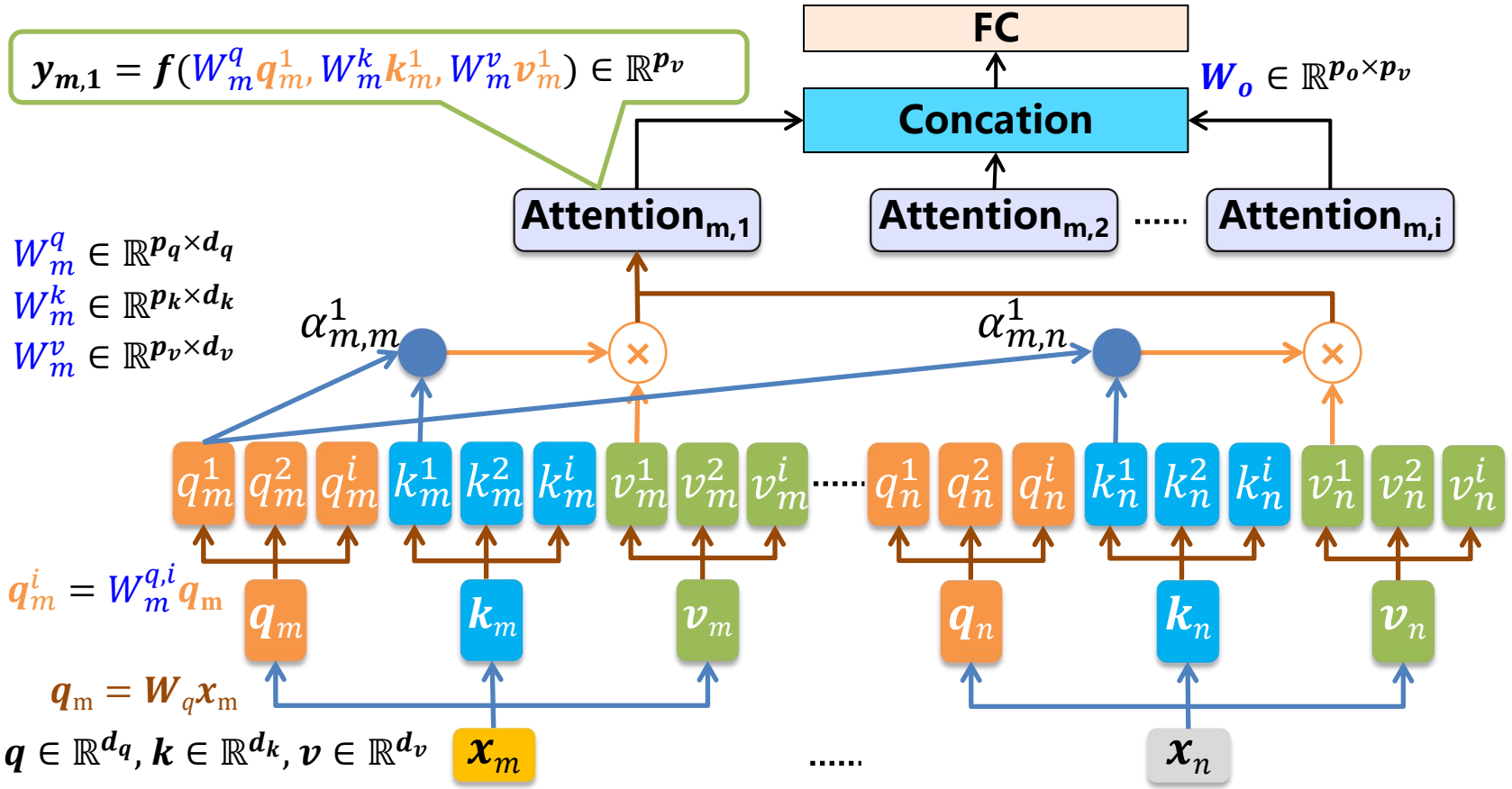
## 基于 编码器 - 解码器 的架构 Transformer

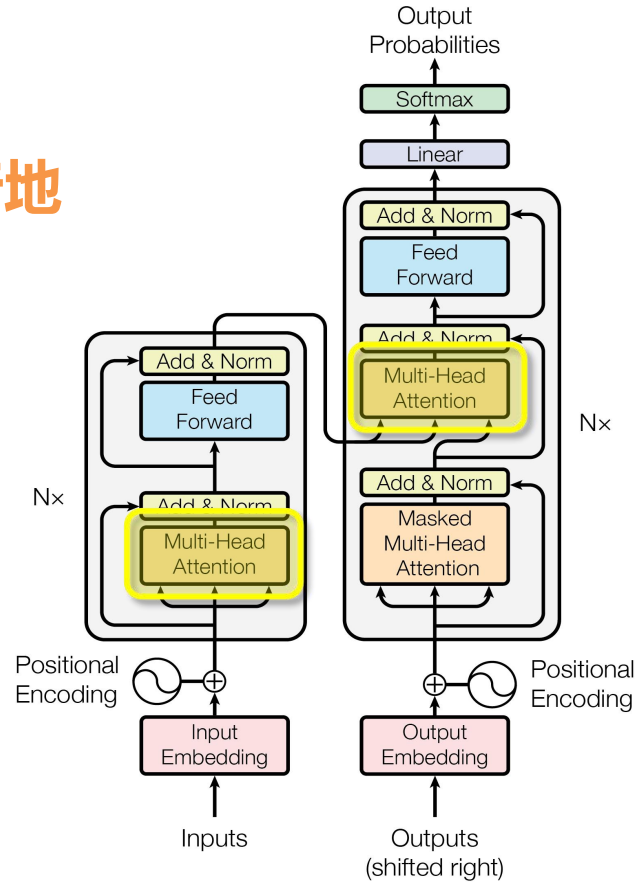


# Transformer的模块

## 多头注意力

- **多头注意力**：对于同一组 query, key, value，通过多组线性变换并行地捕捉不同的特征。短距离关系（句子）和长距离关系（段落）。



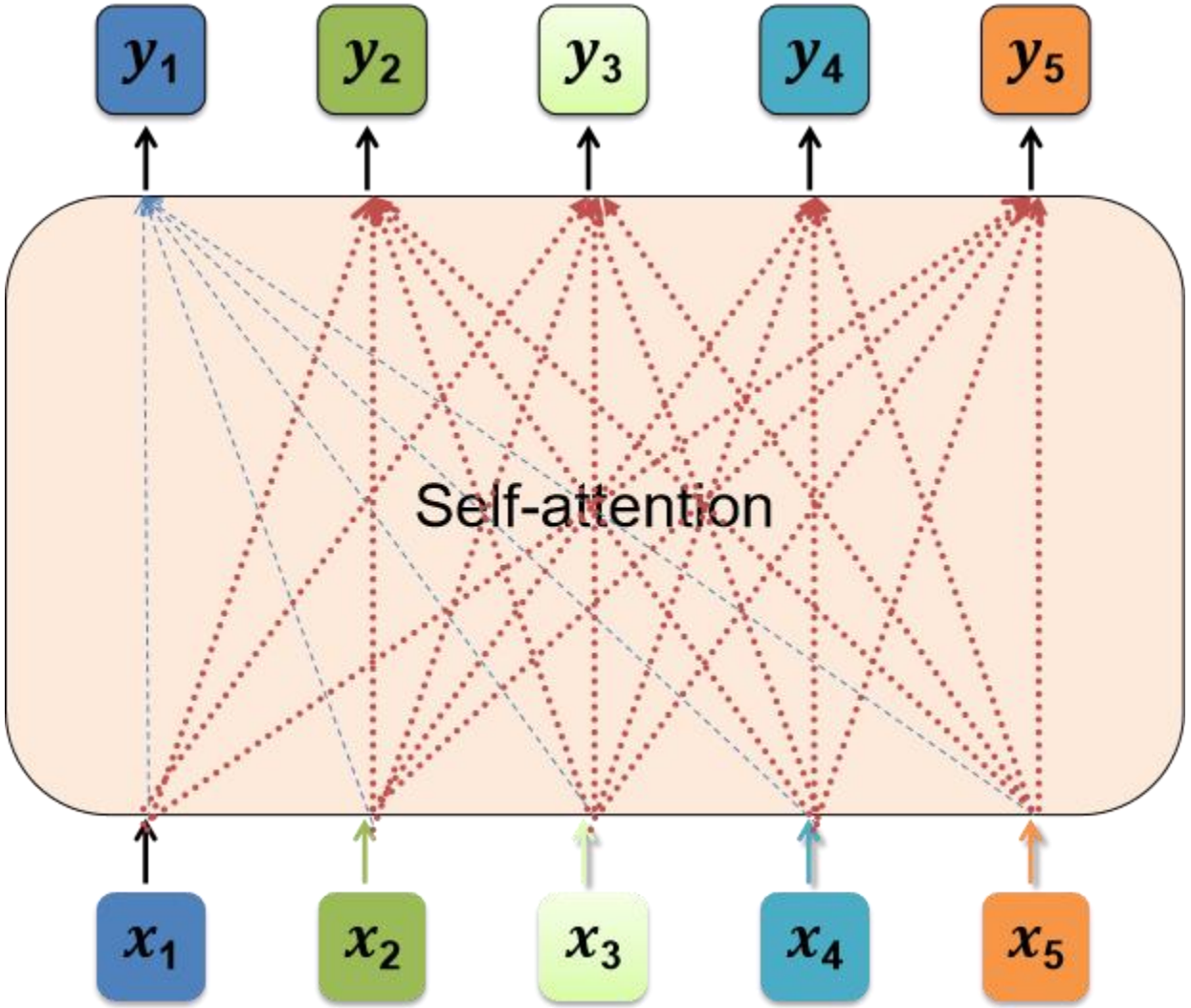



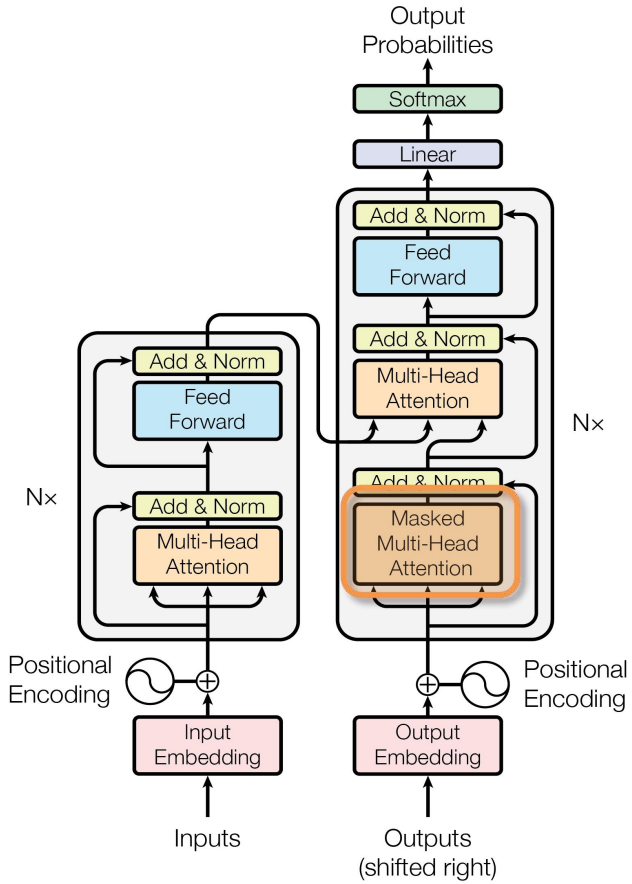

- ***多头注意力的输出***

$$y_m = W_o \begin{bmatrix} y_{m,1} \\ y_{m,2} \\ \cdots \\ y_{m,i} \end{bmatrix} \in \mathbb{R}^{p_o}$$

# Transformer的模块

## 有掩码的多头注意力

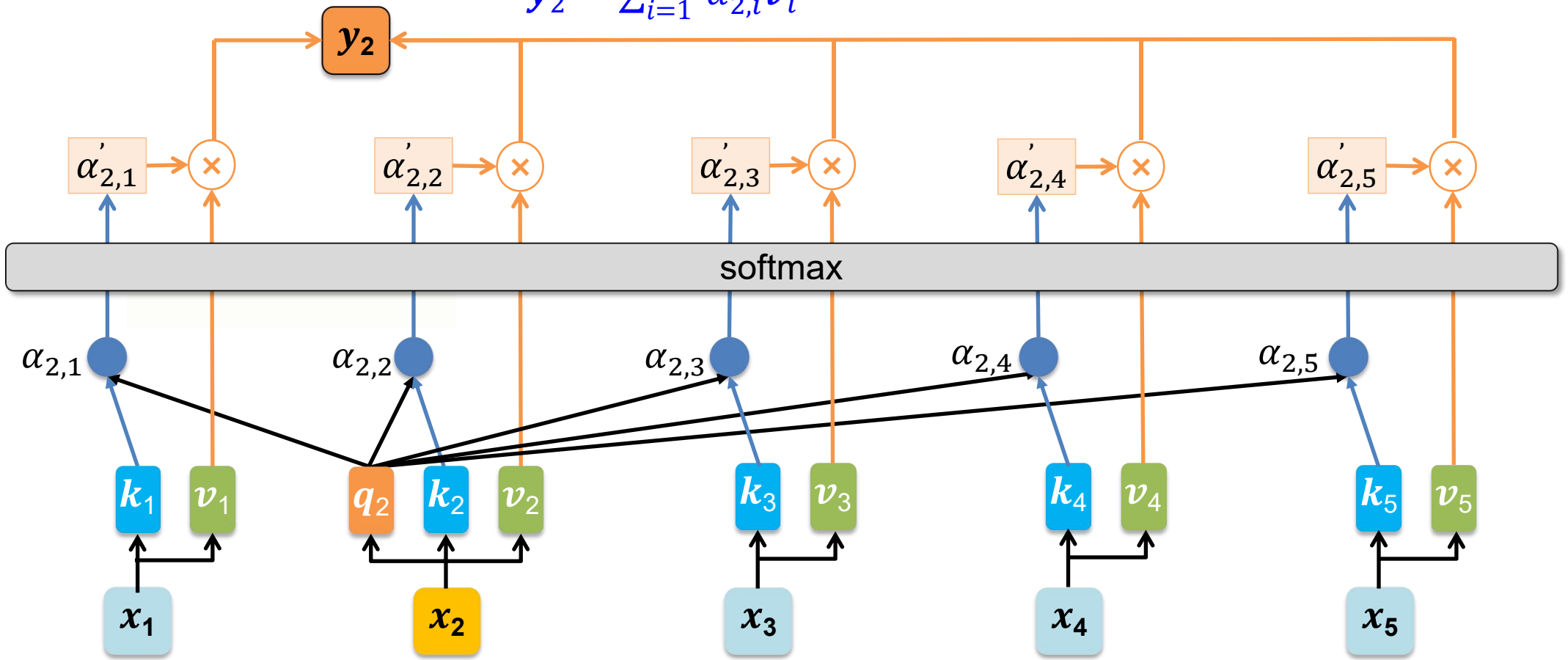


# Transformer的模块

## 有掩码的多头注意力

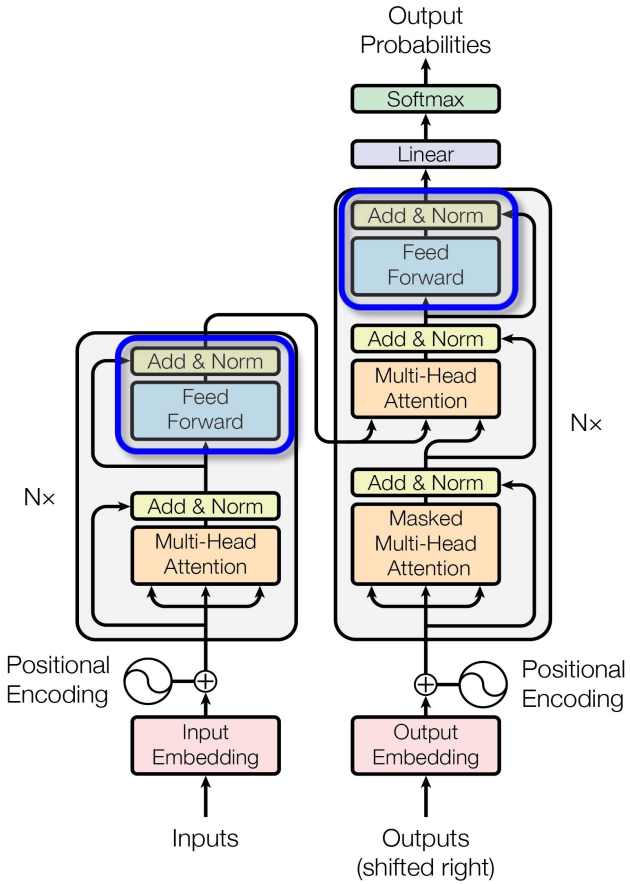
$$y_2 = \sum_{i=1}^{n=2} \alpha'_{2,i} v_i$$



# Transformer的模块

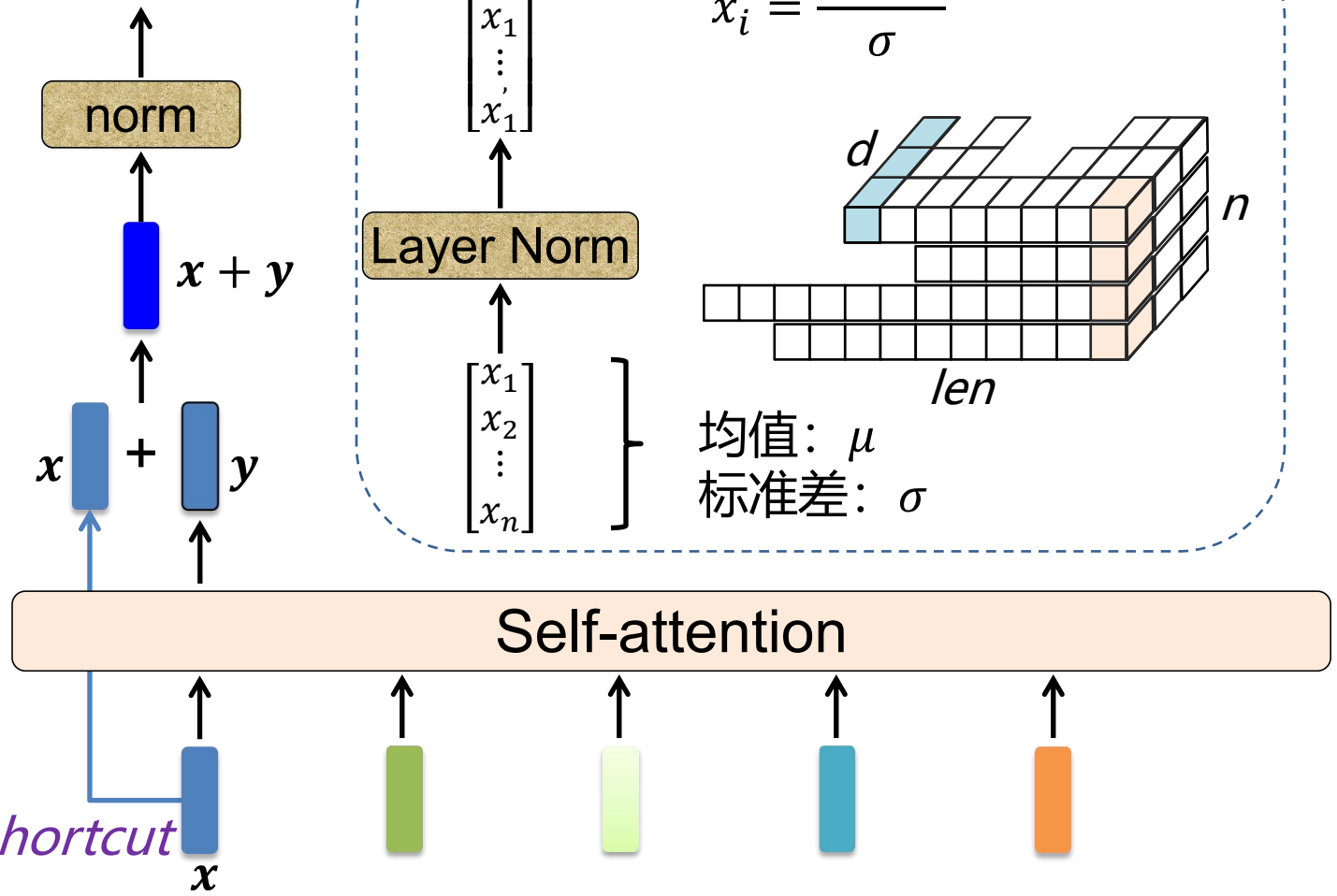
## 基于位置的前馈网络

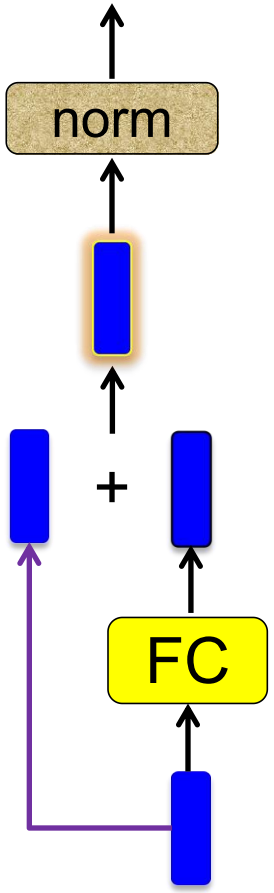
- ***卷积神经网络***
  - ✓ **输入（卷积层）形状：** （b，h，w，c）
  - ✓ **输出（全连接层）形状：** （b, h×w×c）
- ***Transformer***
  - ✓ **编码器：**将输入形状由（b，n，d）变换为（bn，d）
  - ✓ **解码器：**将输出形状由（bn，d）变回（b，n，d）
  - ✓ 等价于两个核窗口为1的一维卷积

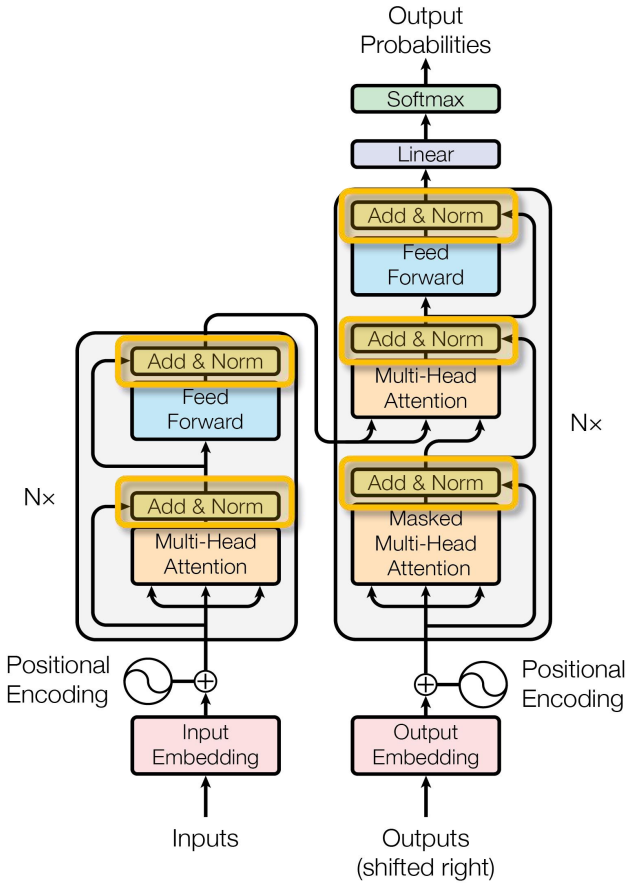


# Transformer的模块

## 层归一化

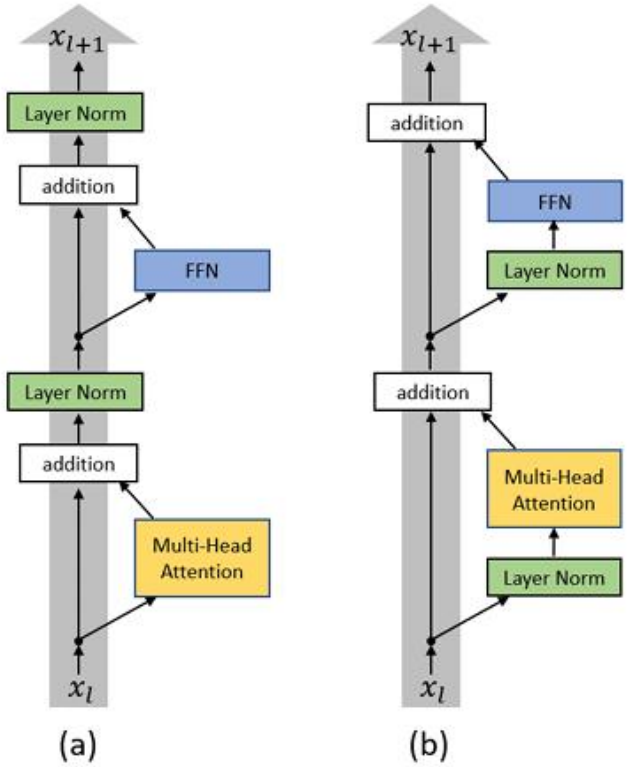


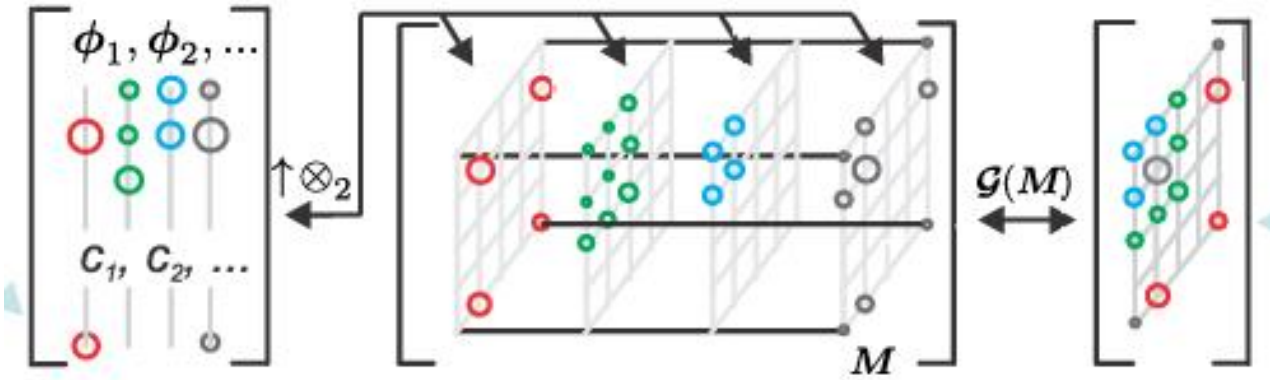



# Transformer的模块

## Tips1: Add & Norm

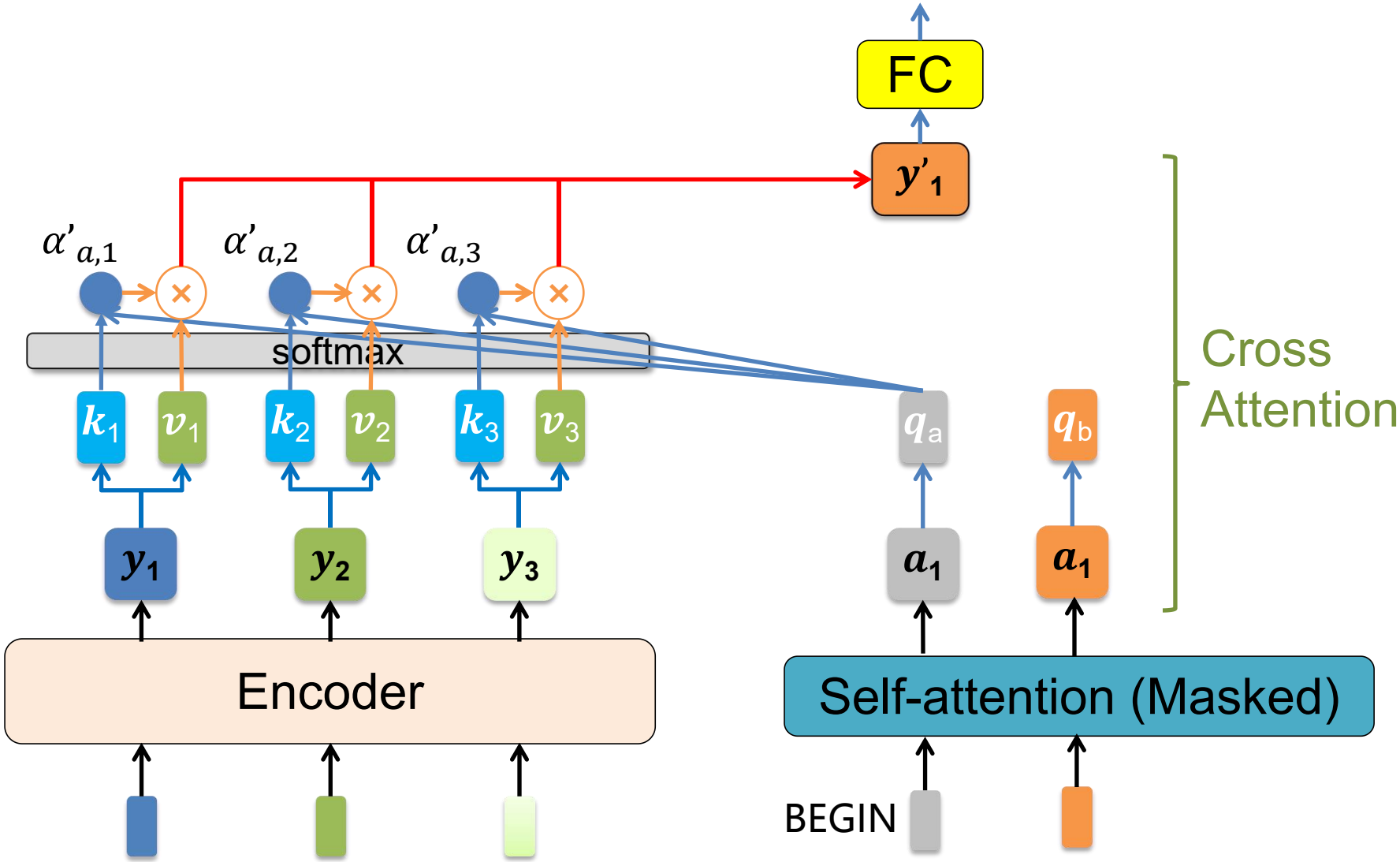


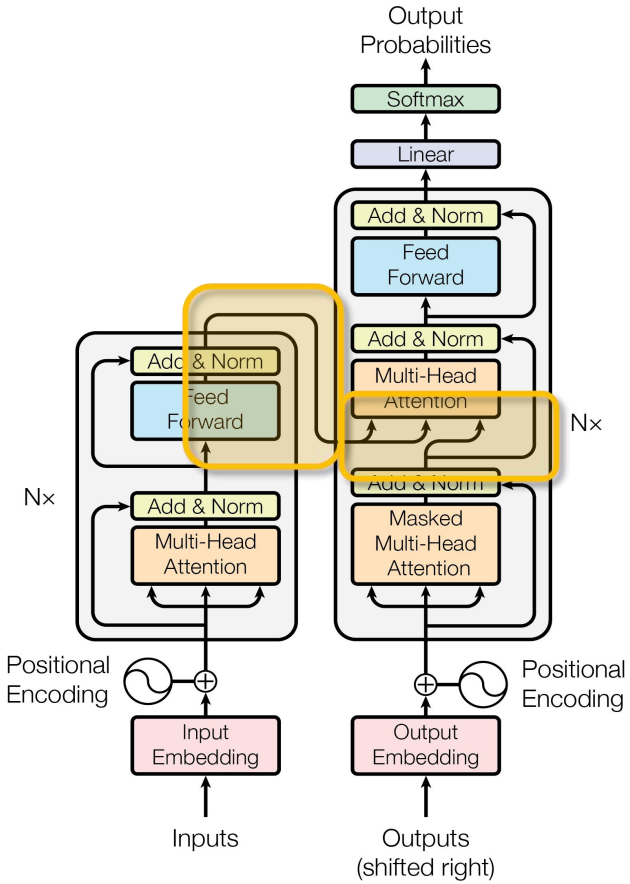




- On Layer Normalization in the Transformer Architecture
- PowerNorm: Rethinking Batch normalization in Transformers
- A Deeper Look at Power Normalizations

# Transformer的模块

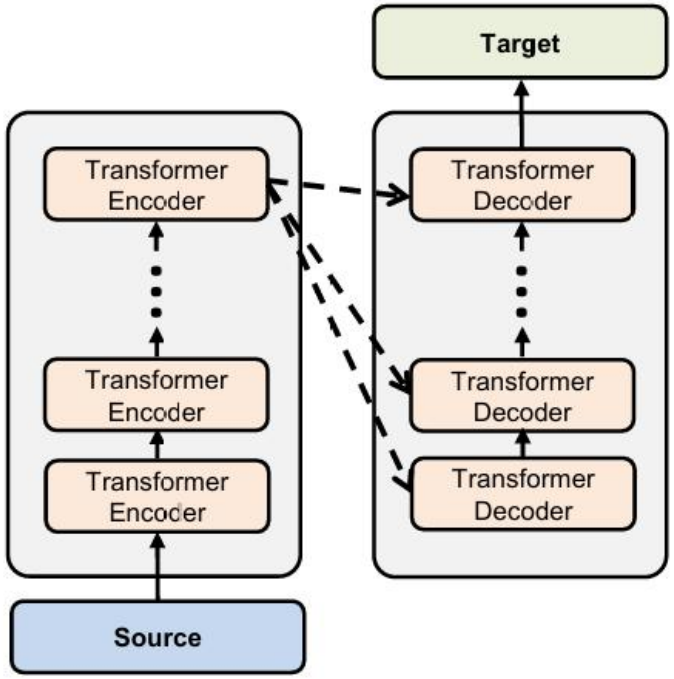
## 信息传递（Cross Attention）



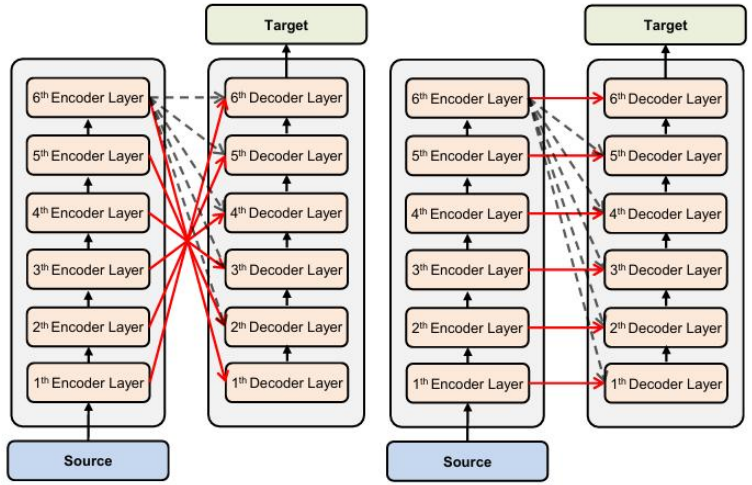



# Transformer的模块

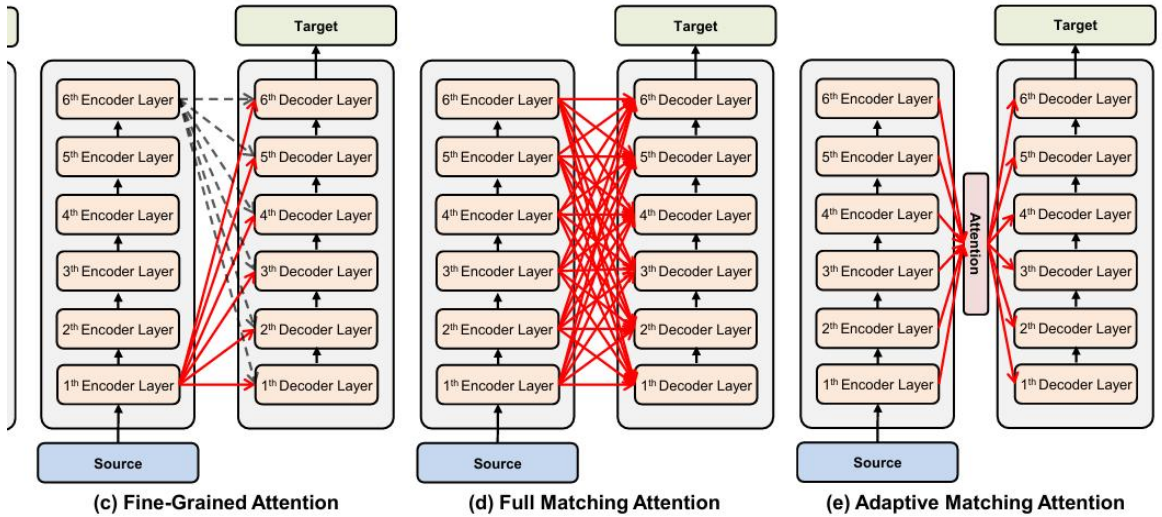
## Tips2: 交叉注意力




(a) Conventional Transformer




(a) Granularity Consistent Attention (b) Granularity Parallel Attention

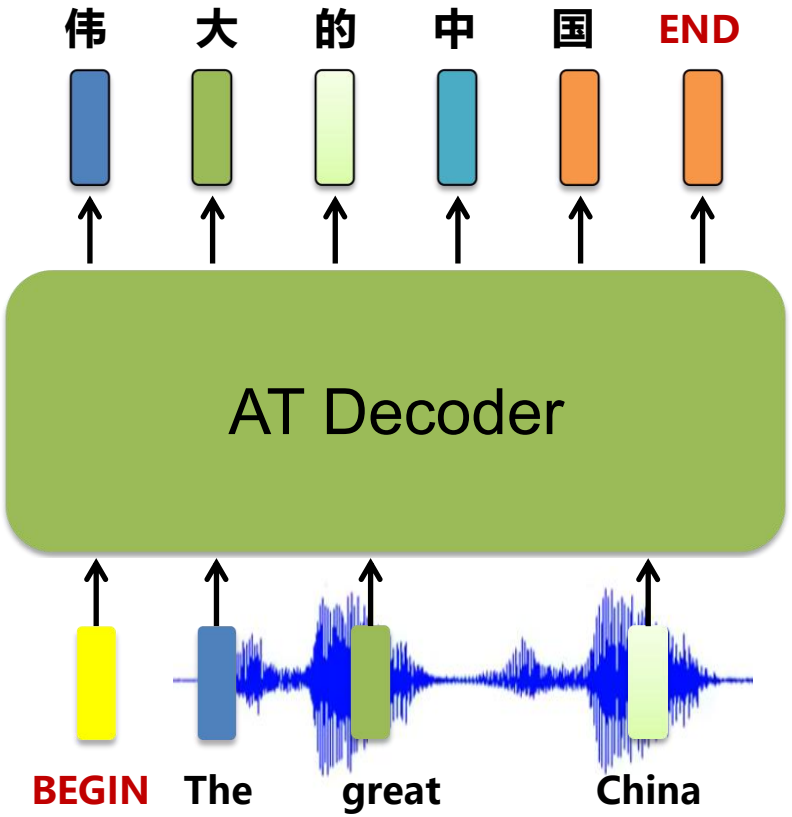


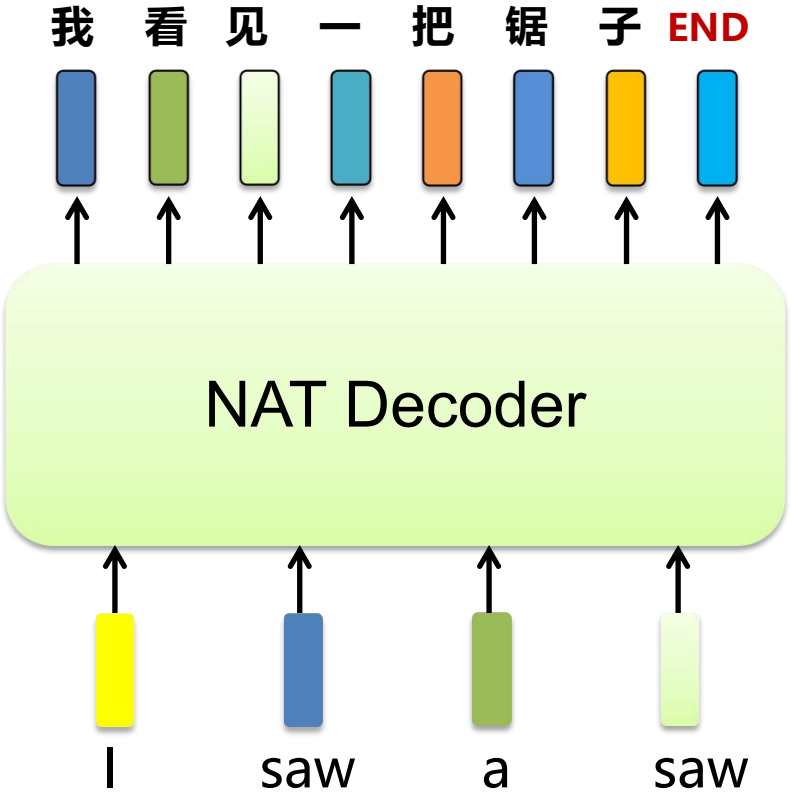

(c) Fine-Grained Attention (d) Full Matching Attention (e) Adaptive Matching Attention

- Rethinking and Improving Natural Language Generation with Layer-Wise Multi-View Decoding

# Transformer的模块

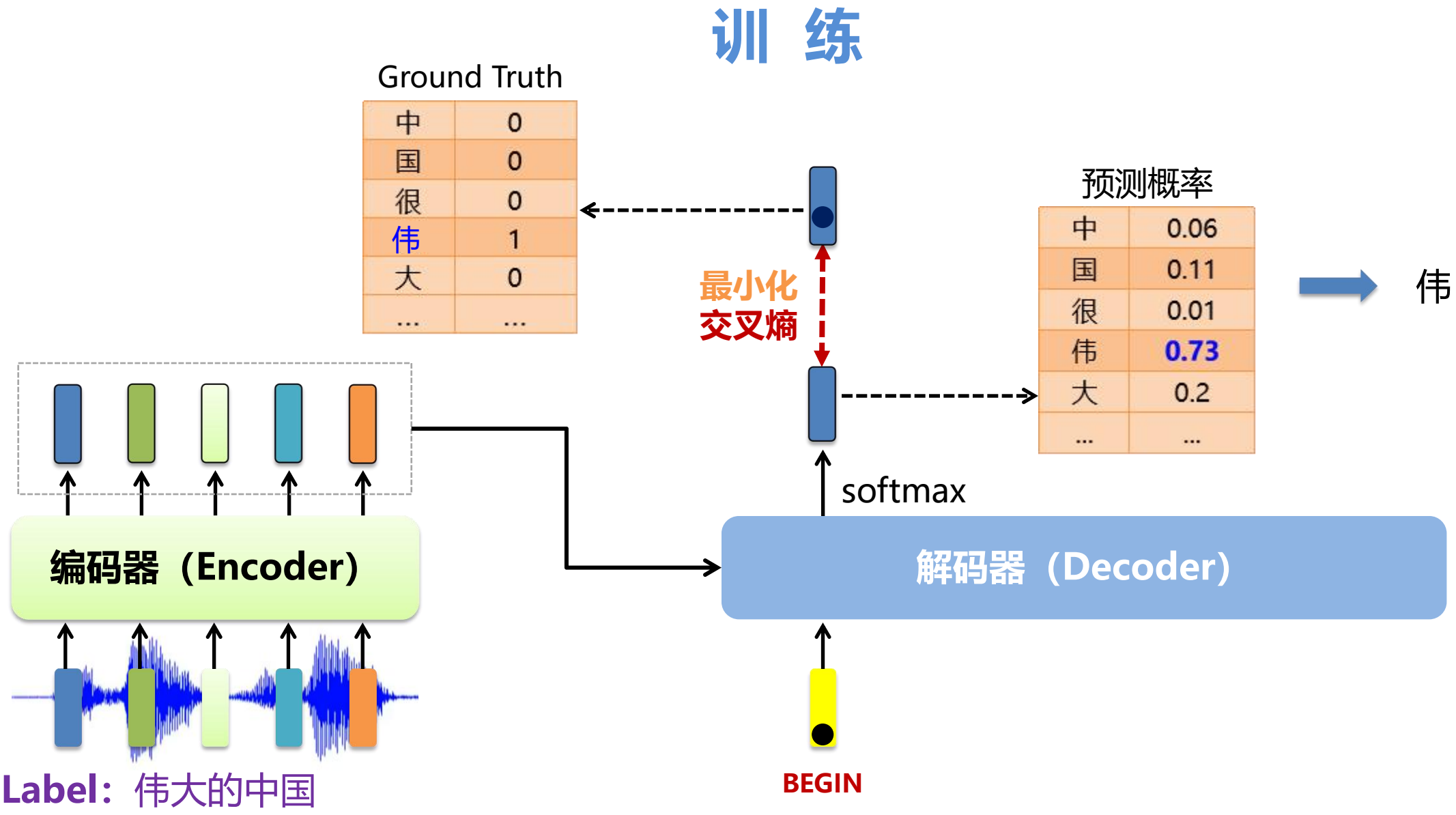
## Tips3: 自回归解码器和非自回归解码器



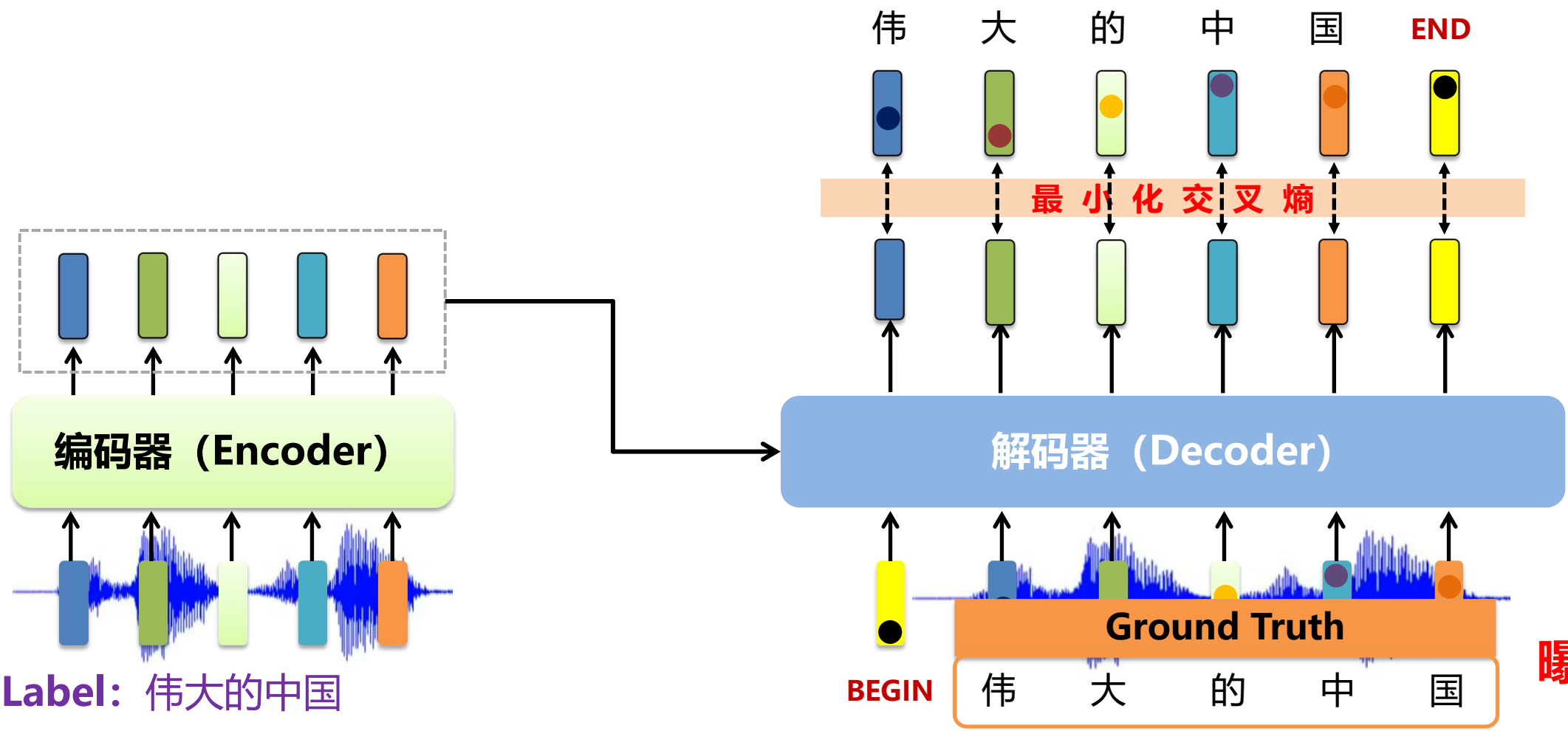
# Transformer 的训练和预测

# Transformer 的训练和预测



# Transformer 的训练和预测

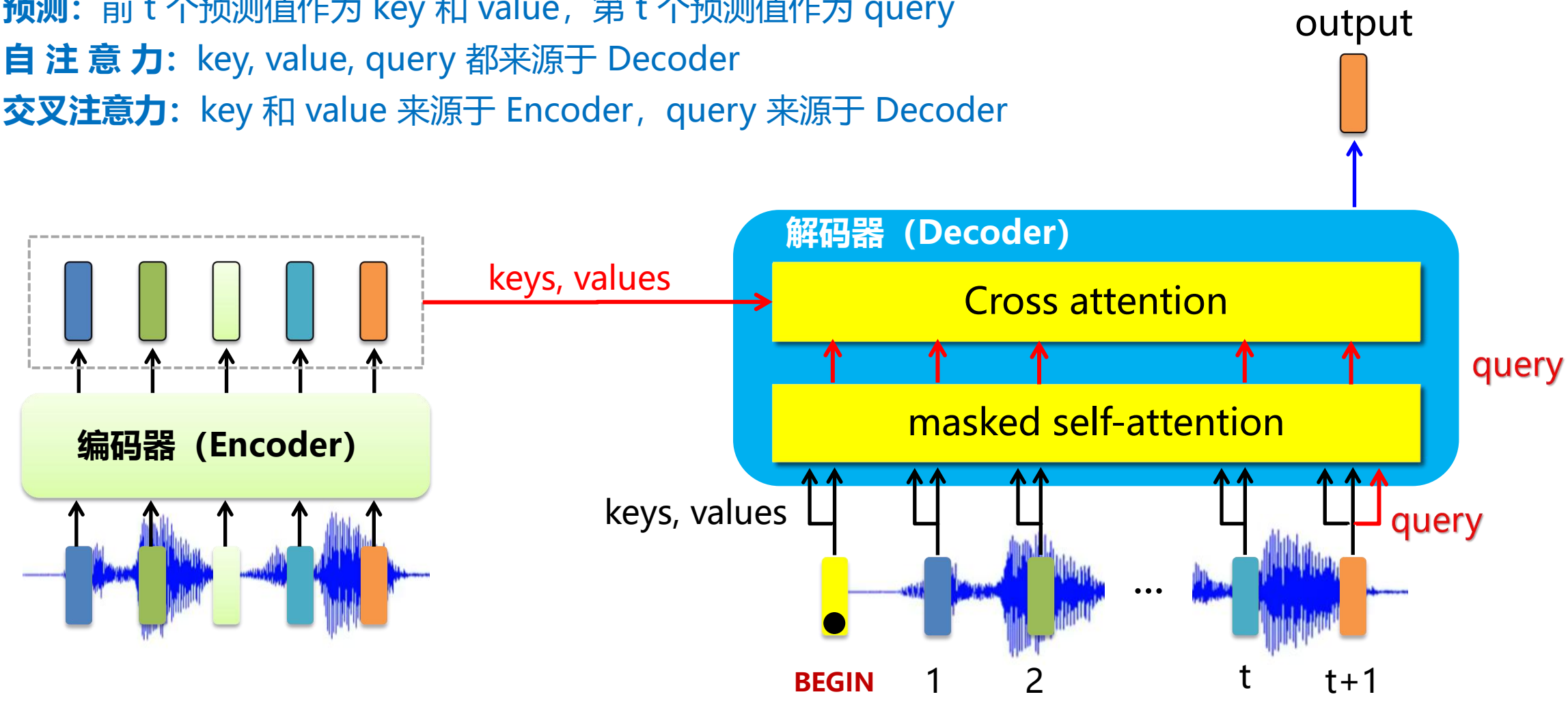
## 训 练



# Transformer 的训练和预测

## 预 测

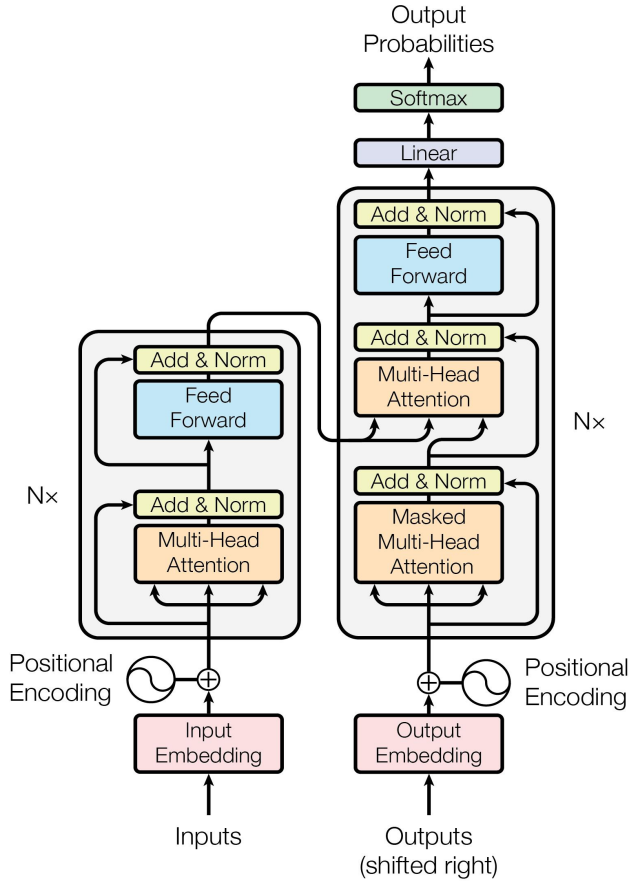
- **预测：**前 t 个预测值作为 key 和 value，第 t 个预测值作为 query
- **自 注 意 力：**key, value, query 都来源于 Decoder
- **交叉注意力：**key 和 value 来源于 Encoder，query 来源于 Decoder



# Transformer

## 小　结

- **Transformer** 是一个纯注意力的 **编码器-解码器** 模型
- **编码器** 和 **解码器** 都由多个 transformer 块构成，主要包括 *多头自注意力*、*基于位置的前馈网络*、*层归一化组成*
- **编码器使用** self-attention 捕捉全局依赖；**解码器**采用 masked self-attention 来屏蔽未来信息，并实现历史序列信息的获取
- 编码器和解码器之间使用 cross-attention 进行连接
- Positional Encoding 通过硬植入的方式将相对位置信息融合到每一个 token 嵌入向量中。



# 欧老师的联系方式

读万卷书 行万里路 只为最好的修炼


QQ：14777591 (宇宙骑士)

Email：ouxinyu@alumni.hust.edu.cn

Website: http://ouxinyu.cn

Tel：18687840023

地址：安宁校区 诚远楼201

南院 智能应用研究院A306-2