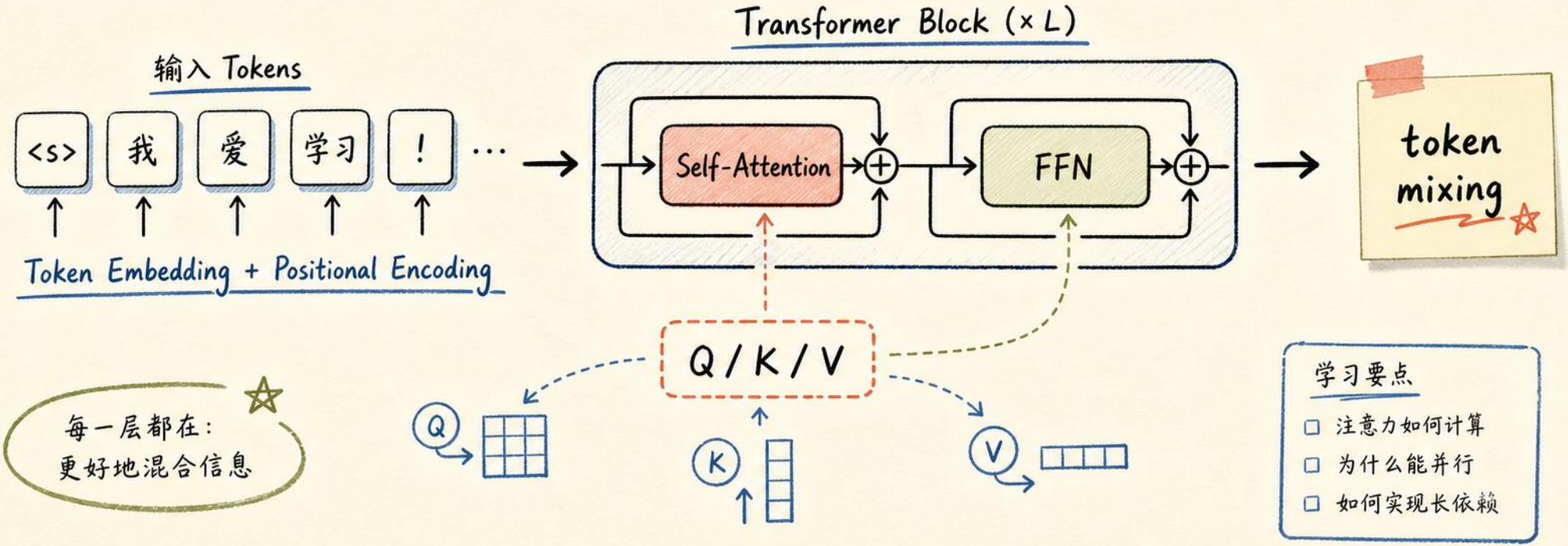
阅读笔记

- [x] 理解结构
- [x] 抓住核心
- [x] 融会贯通

Attention Is All You Need 解读

Transformer 从这里开始

核心思想：
注意力机制
替代循环与卷积



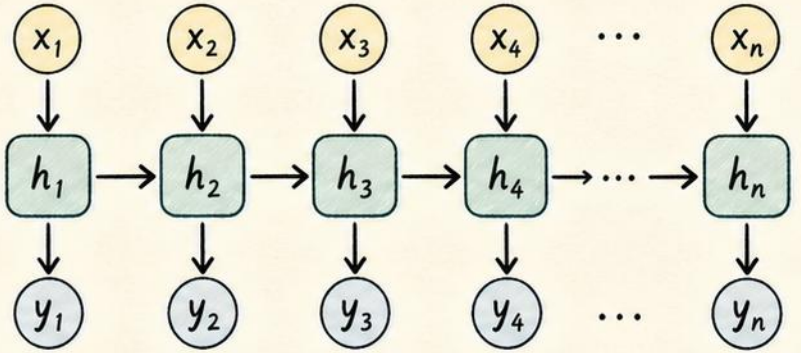
每一层都在：
更好地混合信息

学习要点

- [ ] 注意力如何计算
- [ ] 为什么能并行
- [ ] 如何实现长依赖

序列计算太慢

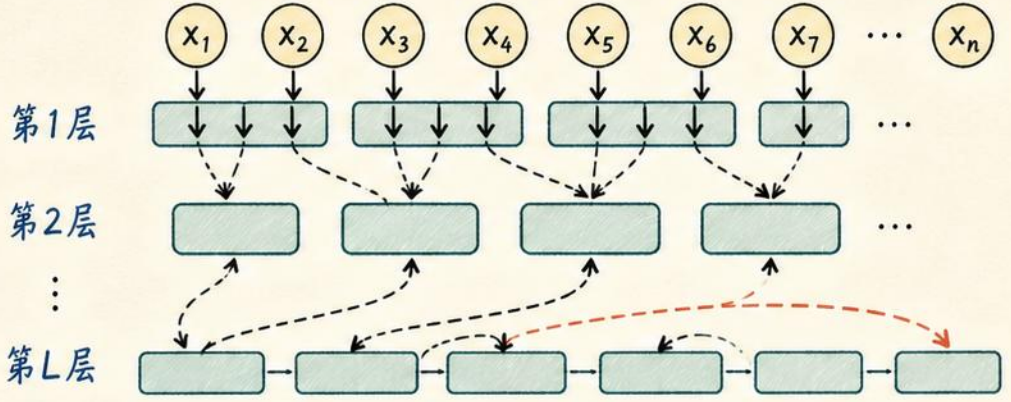
RNN



O(n) 顺序

必须一步步传递

CNN

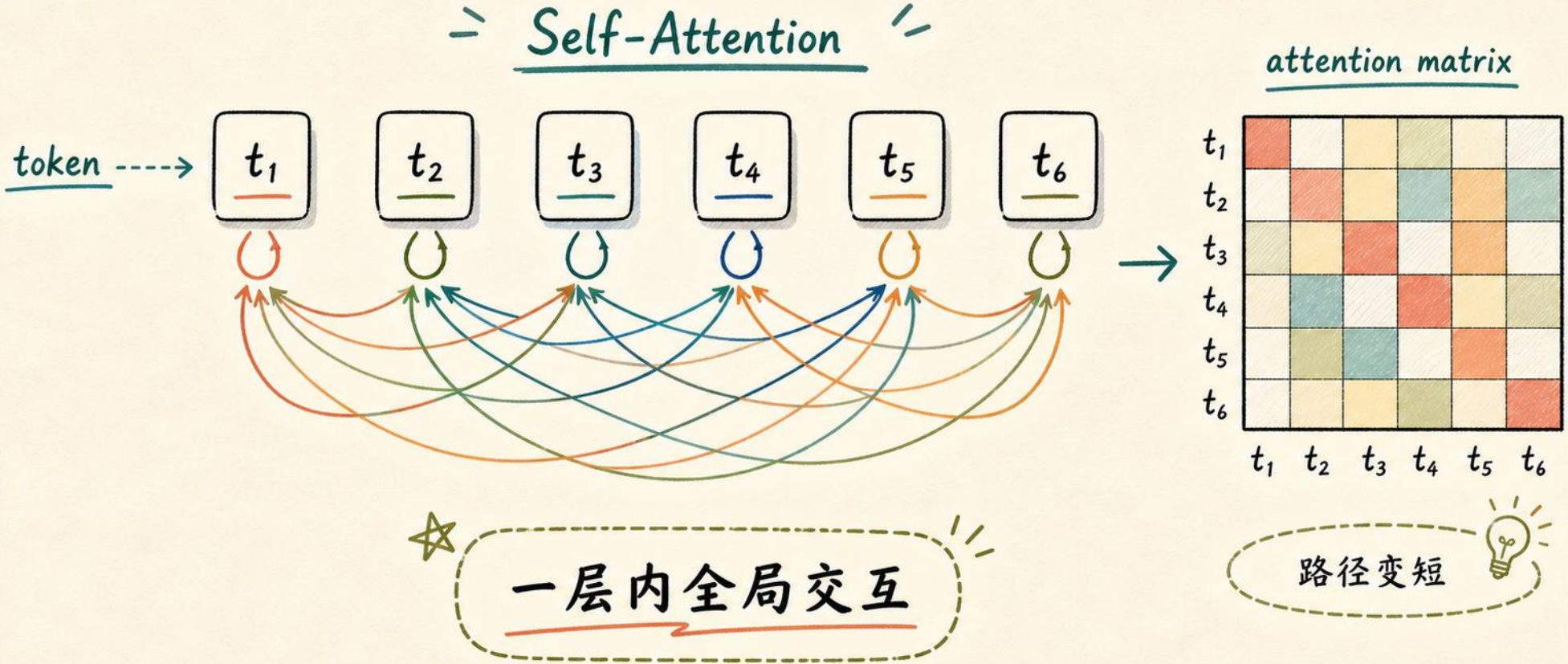


多层传递

要堆很多层才能连到远处

☆ 无论是 RNN 的 O(n) 顺序，还是 CNN 的多层传递，都会形成 **长依赖路径**

全局可见


Self-Attention
token
t_1
t_2
t_3
t_4
t_5
t_6
attention matrix
t_1
t_2
t_3
t_4
t_5
t_6
一层内全局交互
路径变短

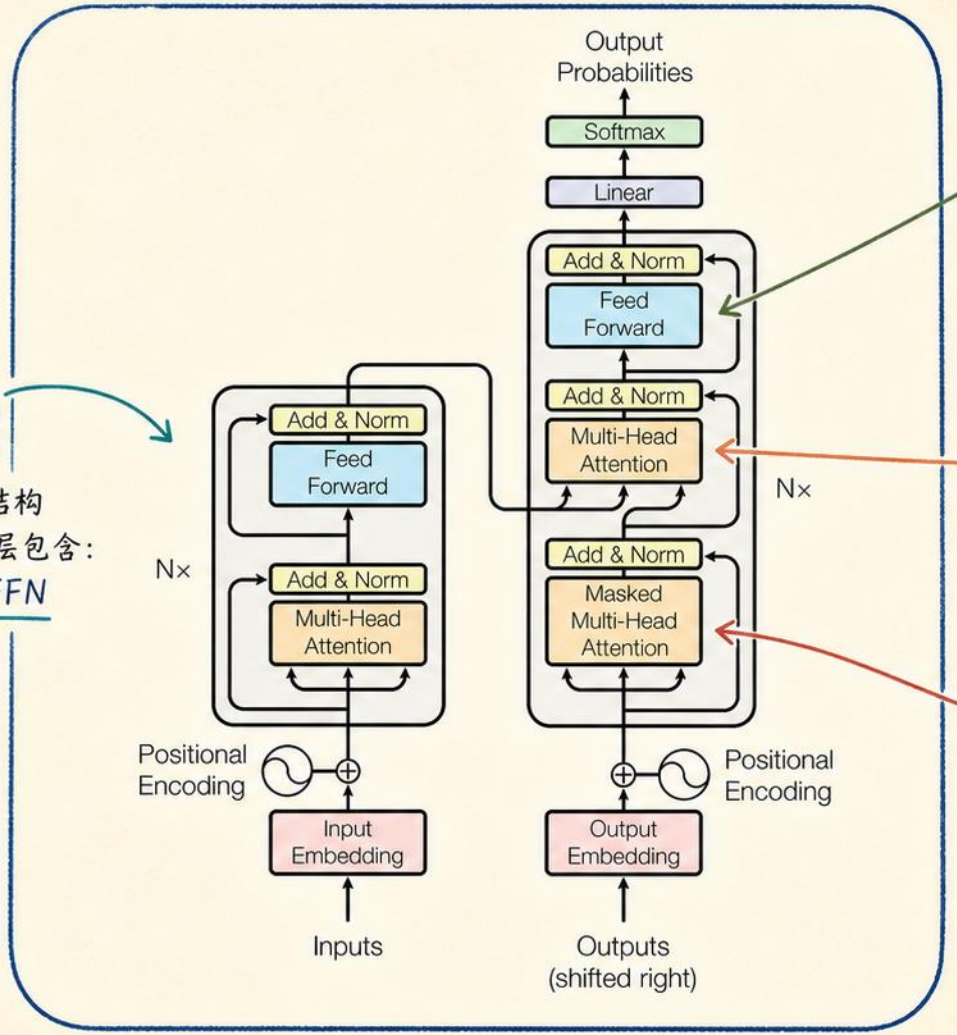
Transformer 总装

核心结论

Transformer 仍然是 Encoder-Decoder 架构，但它的“积木块”由注意力层和 FFN 层组成。

要点

- ☑ 注意力驱动
- ☑ 并行计算友好
- ☑ 长距离依赖建模强



Encoder

由 N 层相同结构堆叠而成，每层包含：自注意力 + FFN

FFN

位置逐点的前馈网络（两层线性变换 + 非线性），增强表达能力。

Cross Attention

编码器输出作为 K、V，解码器当前表示作为 Q，让解码器“看懂”输入信息。

Masked

带掩码的自注意力，防止看到未来位置，保证自回归生成。

QK 匹配，V 汇聚

Scaled Dot-Product Attention = 可微分检索

Q Query 查询
想要找什么？

K Key 键
有哪些内容？

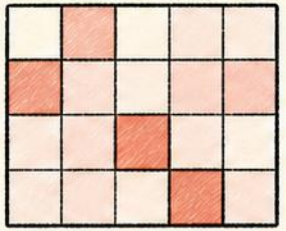
V Value 值
内容本身

Q 查 K

$$\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

scale
防止 softmax 饱和

权重

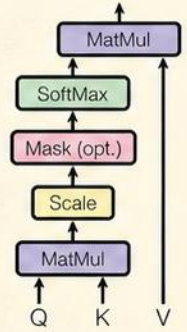


softmax 后的权重
（行和为 1）

汇聚 V

输出
加权后的信息

参考：论文 Figure 2
(Scaled Dot-Product Attention)



① Q 与所有 K 计算相关性 → ② softmax 得到权重 → ③ 用权重加权汇聚 V

全程可微，可端到端训练

多个视角并行看

8 heads

子空间

并行读取

输入 Token 流

Scaled Dot-Product Attention

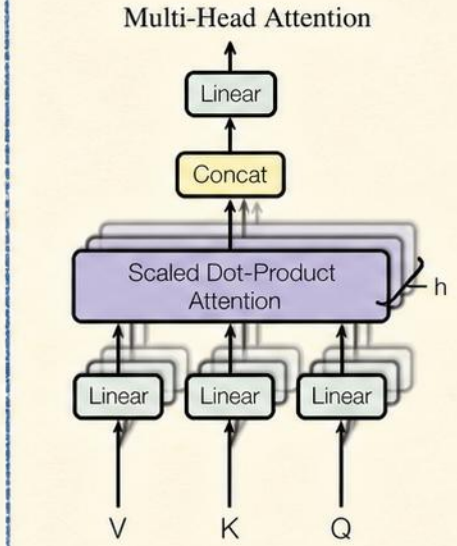
Concat

Linear

单头会平均，
多头分工

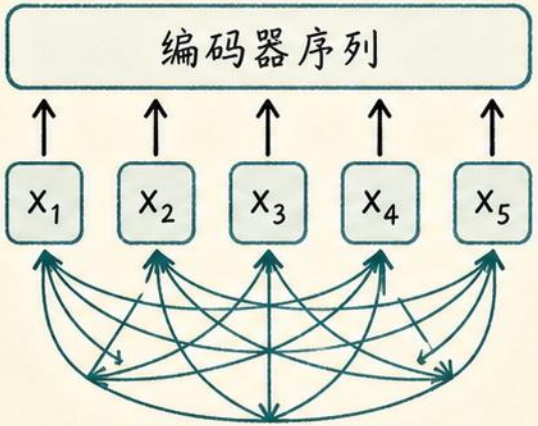
分工

来源：论文 Figure 2 (right)



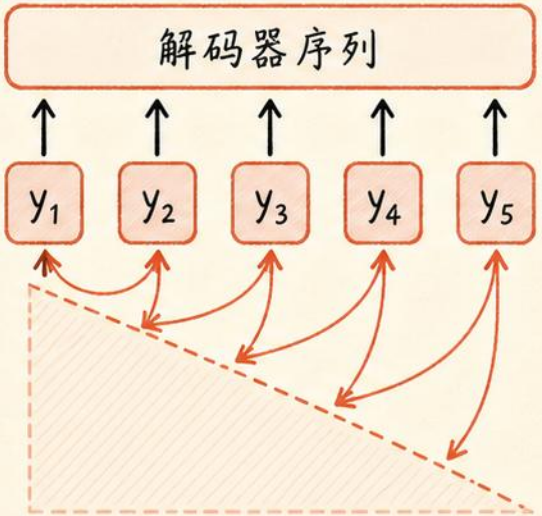
三种 Attention

1 Encoder self



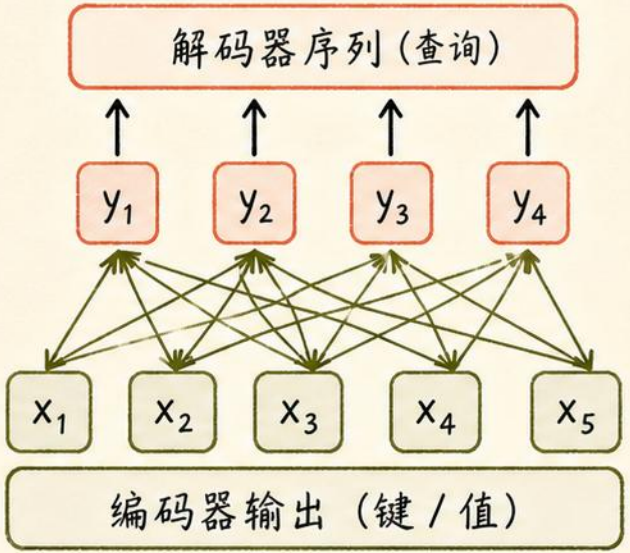
全部可见，彼此关注

2 Decoder masked



只看过去

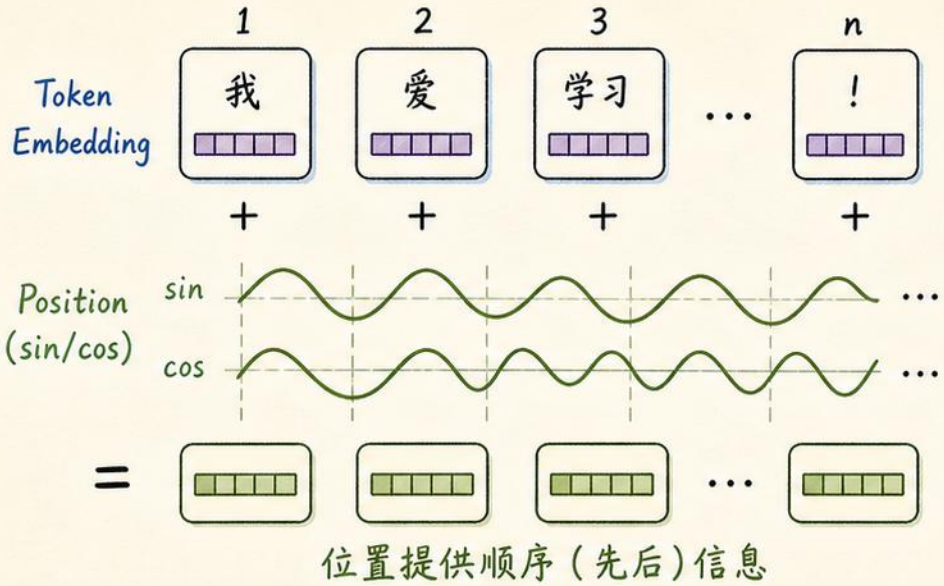
3 Cross attention



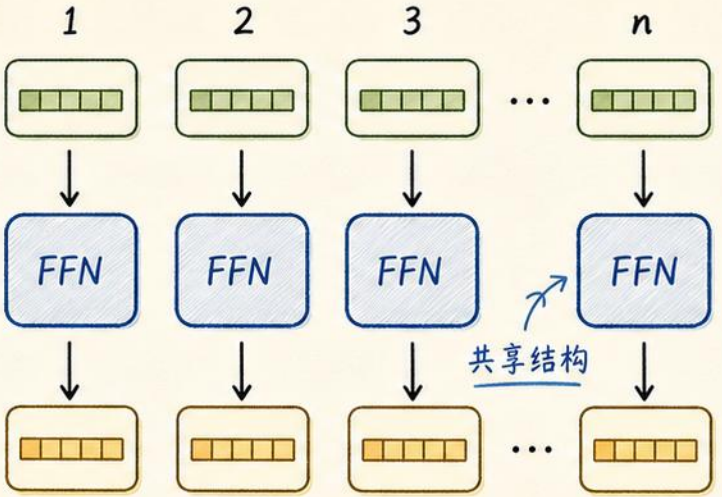
读取源句

位置 + FFN

1) 位置编码提供顺序信息



2) FFN 对每个 token 独立变换



逐 token

每个 token 独立处理，不看其他 token

位置编码（PE）

$$PE(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right), \quad PE(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$

pos：位置（从 0 开始）　　i：维度索引　　d_{model}：模型维度

FFN 公式

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

x：输入向量　　W_1, W_2：权重矩阵　　b_1, b_2：偏置

位置编码在加法后与词向量同维度相加

阅读笔记

- ☑ 理解结构
- ☑ 抓住核心
- ☑ 融会贯通

短路径，高 BLEU

架构优势 → 实验收益

核心思想：
注意力机制
替代循环与卷积

证据 1：结构优势（来自 Table 1）

Layer Type	Sequential Operations	Maximum Path Length
Self-Attention	O(1)	O(1)

O(1) sequential
常数次顺序操作

O(1) path
常数路径长度

结构优势
→ 实验收益

证据 2：实验收益（来自 Table 2）

Model	BLEU (WMT14)	
	EN-DE	EN-FR
Transformer (base model)	27.3	38.1
Transformer (big)	28.4	41.0

28.4 BLEU

41.0 BLEU

WMT14

常数步骤，短路径 → 更高质量的翻译（更高 BLEU）

核心遗产

同一个核心：token mixing

BERT

- mask：双向（随机）
- 目标：MLM + NSP
- token：文本序列
- 规模：中等

GPT

- mask：因果（左到右）
- 目标：Next Token 预测
- token：文本序列
- 规模：大

Transformer Block

token mixing

T5

- mask：编码器双向，解码器因果
- 目标：Text-to-Text 生成
- token：文本序列（编解码）
- 规模：中到大

ViT

- mask：无（全可见）
- 目标：图像分类 / 预训练
- token：图像 Patch 序列
- 规模：中到大

同一骨架，换用法

mask：注意力可见性　目标：预训练/学习目标　token：输入表示形式　规模：参数量级