原文链接：https://arxiv.org/pdf/1706.03762.pdf

# Transformer VS LSTM

- LSTM：串行训练（当前字处理完，才可以处理下一个字）
- Transformer：并行训练（所有字同时训练）

# Transformer特点

- 位置嵌入（positional encoding）理解语言顺序
- 自注意力机制（self Attention mechanism） & 全连接层

# Transformer的组成

- Encoder：输入→隐含层
- Decoder：隐含层→输出
- 输入 输出 都是语言序列

> **decoder的输出**
>
> - 通过N层Decoder Layer输出一个token
> - 而不是一层Decoder输出一个token

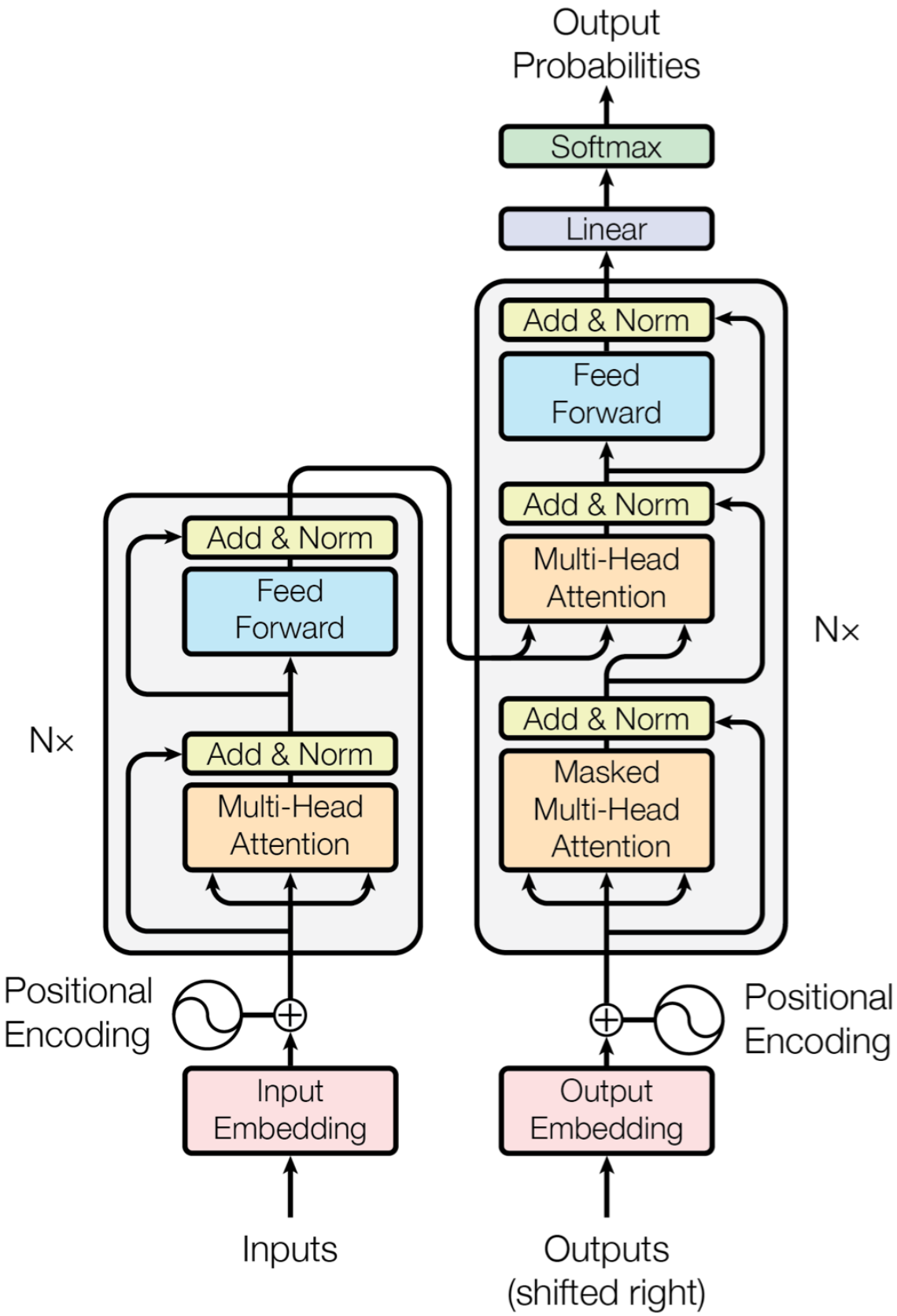
# Transformer 的作用？

机器翻译

输入：中文序列

输出：英文序列

# 架构图


Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
N×
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
N×
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)

# Encoder部分

- 白话：自然语言序列怎么变成的数学表示

## *part 1 位置编码*

- 位置编码的作用?

  > 让Transformer知道这个字在序列中的位置
  >
  > 因为Transformer所有字同时训练

- ？位置编码是什么?

  > 位置嵌入的维度为 `[max_sequence_length, embedding_dimension]`,
  >
  > `max_sequence_length`：限定每个句子最长由多少个词构成
  >
  > `embedding_dimension`：位置嵌入的维度与词向量的维度是相同的

- 初始化

  > 初始化字编码的大小为 `[vocab_size, embedding_dimension]`
  >
  > `vocab_size` 为字库中所有字的数量`embedding_dimension` 为字向量的维度,

## *part 2*