



# 什么是Self-Attention和Transformer?

用开会和工厂的例子，解释大模型“怎么想事情”。一起探索AI语言模型的核心原理！



作者：Yueming Ni

# 我们今天讲啥？

1 什么是“注意力”？

解释如何确定重点信息。

2 什么是Transformer？

了解AI模型如何处理文本。

3 怎么一起工作？

揭示配合方式，理解句子。

4 总结提升

归纳关键知识，展望发展。



# 大模型组件分工对比

大模型就像一座智能工厂，将语言原料加工成高质量成品。每个组件都有特定职责，共同构建强大的语言理解系统。

组件	工厂比喻	在大模型中的作用
Token	原料小块	把语言拆成小片段，模型处理的最小单位
Embedding	给每块原料贴上隐形说明书	把每个词转换成可以被模型理解的数字向量
位置编码	给每块原料编号	告诉模型这些词的顺序（谁先谁后）
注意力机制	项目经理：调度谁更重要	决定每个词要参考哪些其他词，理解重点关系
Transformer	整个流水线系统	将所有输入加工成有用输出，主干架构
输出层	交付组装成品	将内部理解转成我们看得懂的文字或答案

这个精密的"语言工厂"让AI能够理解、生成和处理人类语言，每个零件缺一不可。

# 什么是“注意力”？



## 区分重点信息

像开会时，挑选对任务最关键的信息，确保资源集中。



## 词与词之间的关注

每个词都可以灵活关注句子中其他词的重要程度，形成动态联系。



## 动态调整关注强度

通过调整注意力权重（线的粗细）来体现词之间关联的强弱，帮助模型抓住重点。



# 注意力机制：重点关注！

## 加权信息

给重要词更高权重，减少噪音误导，提升模型的准确性和效率。  
相关性决定影响力，模型动态调整注意力权重，突出关键信息。  
通过计算词与词之间的关联程度，实现信息的加权过滤和强化。

## 例子讲解

在句子“我 爱 北京”中，“爱”会重点关注“北京”，强化其语义联系。

这种机制帮助模型精准理解上下文，实现更自然流畅的语言生成。

通过不断调整注意力，模型能够捕捉到长距离依赖，理解复杂语义结构。

# Transformer：自动化智能加工厂

## 输入原始信息

文本中的每个词汇作为“原材料”进入工厂，等待加工处理。

## 多步智能加工

通过多层注意力机制自动识别词间关系，再经过前馈神经网络进行深度特征提取。

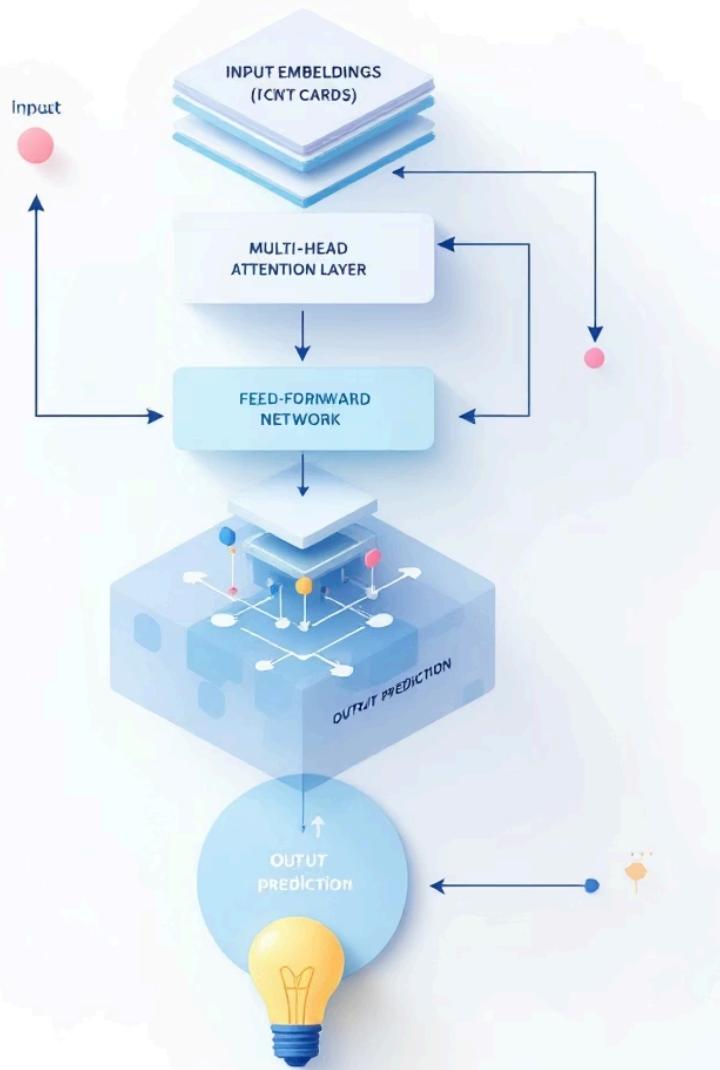
这一过程帮助模型动态调整关注点，实现信息加权和上下文理解。

## 输出理解结果

加工后的信息被转换成模型理解的语义表示，最终输出符合语义的文本答案或预测结果。



## Transformer



# Transformer架构

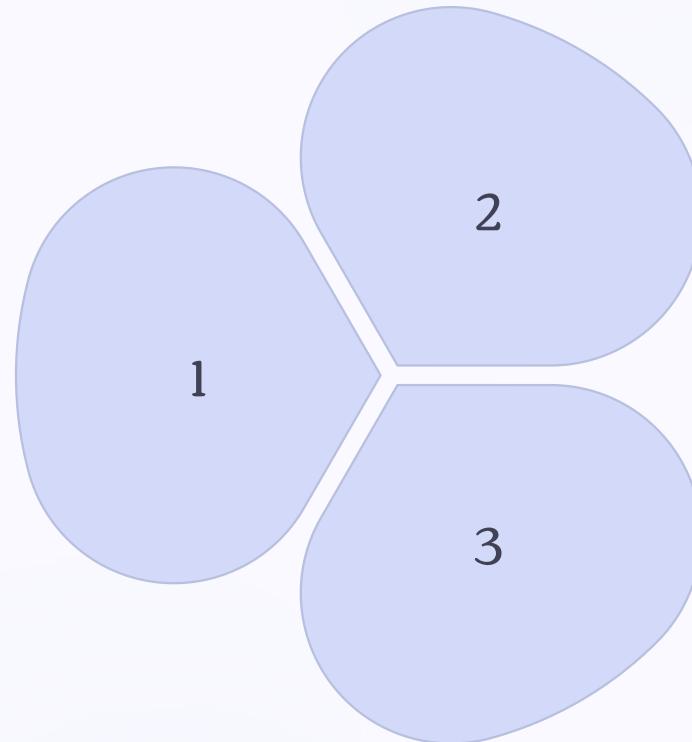
Transformer就像一个多层工厂流水线，每层都有特殊功能。

- 多头注意力：同时关注多个重点，类似多位项目经理各自负责不同方面
- 残差连接：保留原始信息，避免加工过程丢失重要细节
- 前馈网络：深度处理每个词的独特特征
- 层归一化：调整数据范围，确保稳定运行

这种架构让AI能理解语言复杂性，实现真正的"理解"而非简单记忆。

# 注意力 + 工厂 = 超强大脑

多轮关注  
每一步都相互作用，信息不断增强。



理解能力强  
能理解复杂长句和模糊表达。

多种用法  
适合翻译、写作、问答等任务。

# 大模型层数图解

AB

## 输入层

接收文本并转换为向量表示



## 中间层

多层Transformer模块堆叠，层数决定复杂度

## 深层理解

越深层次捕获越抽象的语义关系



## 输出层

将内部表示转换为有意义的文本结果

现代大模型通常有几十到上百层，每层都为最终输出增添智能。GPT-4拥有超过100层的复杂结构，让它能理解极其复杂的语义关系。

# 总结：注意力机制 + Transformer



抓重点

注意力机制让模型看清核心。



强运算

Transformer框架高效处理复杂结构。



智能语言

理解、生成和应用多种场景。



未来无限

AI模型还会更强大、更聪明！

