Transformer 模型简介

Transformer 是一种基于注意力机制（Attention Mechanism）的深度学习模型架构，最早由 Vaswani 等人在 2017 年论文《Attention Is All You Need》中提出。该模型最初用于机器翻译任务，随后成为自然语言处理（NLP）和大规模语言模型（LLM）的核心基础。

一、提出背景

在 Transformer 出现之前，序列建模任务主要依赖 RNN、LSTM 和 GRU 等循环神经网络。这些模型在长序列建模中存在梯度消失、训练难以并行化等问题。Transformer 通过完全抛弃循环结构，仅依赖注意力机制，实现了高效并行训练和更强的建模能力。

二、核心思想：注意力机制

注意力机制的核心思想是：在处理序列中的某个位置时，动态地关注序列中其他位置的信息。Transformer 中使用的是缩放点积注意力（Scaled Dot-Product Attention），其计算过程包括 Query、Key 和 Value 三个向量。

三、模型结构

Transformer 由编码器（Encoder）和解码器（Decoder）两部分组成。编码器通常由多层相同结构的模块堆叠而成，每一层包含多头自注意力（Multi-Head Self-Attention）和前馈全连接网络（Feed-Forward Network）。解码器在此基础上增加了对编码器输出的交叉注意力（Cross-Attention）。

四、关键组件

1. 多头注意力（Multi-Head Attention）：通过多个注意力头并行学习不同子空间的关系。
2. 位置编码（Positional Encoding）：为模型提供序列中位置信息，弥补无循环结构的不足。
3. 残差连接与层归一化（Residual Connection & LayerNorm）：提高训练稳定性和收敛速度。

五、优势与影响

Transformer 具备高度并行化、长距离依赖建模能力强等优势。在此基础上诞生了 BERT、GPT、T5 等一系列重要模型，推动了自然语言处理、多模态学习和生成式 AI 的快速发展。

六、总结

Transformer 已成为现代人工智能领域最重要的模型架构之一。理解 Transformer 的基本原理，对于深入学习大模型、智能体系统以及 AI 应用开发具有重要意义。