Deep Dive into Transformer Architecture

1 Architectural Foundations

1.1 The Architectural of GPT-2

我们以GPT-2的**模型架构**为切入，分析整个Transformer Block的**结构**及其**内在机制**。GPT-2的架构是在GPT-1的基础上改进的，而GPT-1的模型架构则是拿掉了Multi-Head Cross Attention （多头交叉注意力），只保留了Masked Multi-Head Self-Attention的**Transformer的解码器**。GPT-2的模型架构在GPT-1的基础上做了如下改进：

- Layer normalization被移动到每一个sub-block（两个子层：**解码器自注意力**与**基于位置的前馈神经网络**）的**输入**位置，类似于一个**预激活**的残差网络。同时在**最后的**自注意力块后添加一个额外的layer normalization。
- 采用一种改进的初始化方法，该方法考虑了残差路径与模型深度的累积。在初始化阶段使用缩放因子$\frac{1}{\sqrt{N}}$对residual layer的权重进行缩放操作，其中N为residual layer的数量（深度）。
- 字典大小设置为50257；无监督预训练可看到的上下文的 context 由512扩展为1024；Batch Size 大小调整为512。

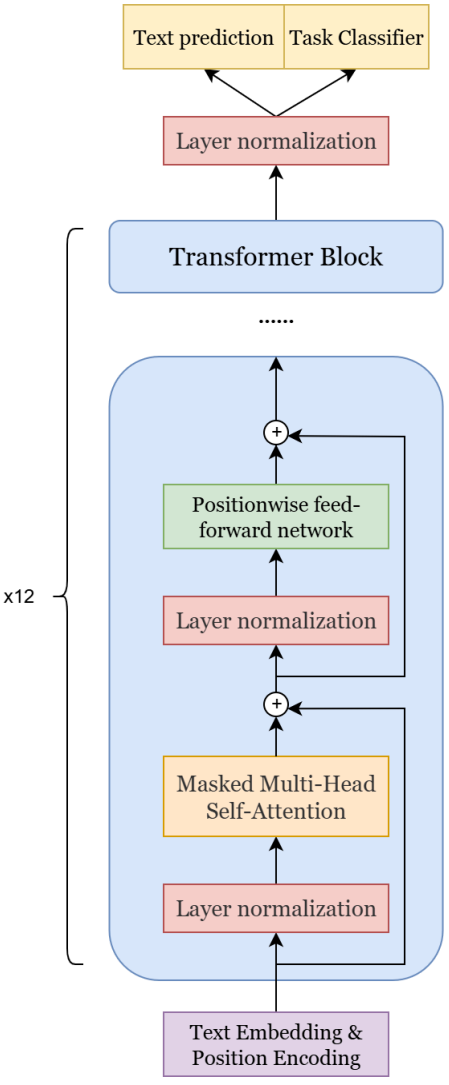

图1 GPT-2的Transformer Block