# GPT 简介

Wuhan University

March 14, 2025

设输入为 $X \in \{0,1\}^{T\times d}$，其中 $d$ 为词元维度，$T$ 为序列长度，定义以下模块：

- **嵌入层**：

$$\mathrm{Enc}(X) = XE_1 \quad (E_1 \in \mathbb{R}^{d\times D}) \tag{1}$$

- **位置编码增强层**：

$$\mathrm{Enc}^{+}(X) = XE_1 + PE_2 \quad \begin{pmatrix} P \in \{0,1\}^{T\times T} \\ P_{i,j} = \delta_{ij} \\ E_2 \in \mathbb{R}^{T\times D} \end{pmatrix} \tag{2}$$

其中 $P$ 为单位矩阵，$\delta_{ij}$ 为 Kronecker delta 函数

- **规范化层**（中间层输入 $X \in \mathbb{R}^{T\times D}$）：

$$\mathrm{Norm}(X) := \frac{X - \mathbb{E}[X]}{\sqrt{\mathrm{Var}(X) + \epsilon}} \tag{3}$$

其中 $\mathbb{E}[X], \mathrm{Var}(X) \in \mathbb{R}^{T\times D}$ 为逐位置计算的均值与方差

**维度映射关系**：

- $\mathrm{Enc} : \{0,1\}^{T\times d} \to \mathbb{R}^{T\times D}$
- $\mathrm{Norm} : \mathbb{R}^{T\times D} \to \mathbb{R}^{T\times D}$

设 $H$ 为注意力头数，第 $h$ 个头的参数矩阵为 $W_{K,h} \in \mathbb{R}^{D\times D_K}$、$W_{Q,h} \in \mathbb{R}^{D\times D_K}$、$W_{V,h} \in \mathbb{R}^{D\times D_V}$，输出投影矩阵 $W_O \in \mathbb{R}^{HD_V\times D}$，定义**多头注意力机制**如下：

- **单头注意力计算**：

$$S_h = \text{Softmax}\left(\frac{XW_{Q,h}(XW_{K,h})^\top}{\sqrt{D_K}}\right)XW_{V,h} \tag{4}$$

- **掩码单头注意力**：

$$S_h = \text{Softmax}\left(\frac{XW_{Q,h}(XW_{K,h})^\top + M}{\sqrt{D_K}}\right)XW_{V,h} \tag{5}$$

其中掩码矩阵 $M \in \{-\infty, 0\}^{T\times T}$ 满足：

$$M_{ij} = \begin{cases} -\infty & \text{若} i < j \\ 0 & \text{其他情况} \end{cases}$$

**维度说明**：$\mathrm{S_h} : \mathbb{R}^{T\times D} \to \mathbb{R}^{T\times D_V}$

- **多头拼接与残差连接**：

$$\mathrm{ATT}(X) := X + [S_1 \| \cdots \| S_h \| \cdots \| S_H]\, W_O \tag{6}$$

**全局说明**：

- Softmax 为逐行 (row-wise) 归一化操作
- 最终映射 $\mathrm{ATT} : \mathbb{R}^{T\times D} \to \mathbb{R}^{T\times D}$ 保持维度不变性
- $\|$ 表示沿特征维度拼接操作

**MLP 层**定义为：

$$\mathrm{MLP}(X) := X + f_3 \circ \sigma \circ \sigma \circ f_2(X) \circ \sigma \circ f_1(X) \tag{7}$$

其中线性变换 $f_i(X) := XW_i + b_i \quad (1 \le i \le 3)$ 满足：

- 偏置项 $b_i \in \mathbb{R}^{D_i}$
- 权重矩阵维度：

$$W_1 \in \mathbb{R}^{D \times D_1}$$
$$W_2 \in \mathbb{R}^{D_1 \times D_2}$$
$$W_3 \in \mathbb{R}^{D_2 \times D}$$

**维度说明**：$\mathrm{MLP} : \mathbb{R}^{T \times D} \to \mathbb{R}^{T \times D}$。

定义逐元素 **GeLU 激活函数**：

$$\begin{aligned} \mathrm{GeLU}(x) &:= x\Phi(x) = x \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \\ &\approx 0.5x \left(1 + \tanh\left(\sqrt{\frac{2}{\pi}}\left(x + 0.044715x^3\right)\right)\right) \end{aligned} \tag{8}$$

**残差块**定义为多层操作的复合：

$$\text{Block}(X) := \text{MLP} \circ \text{Norm} \circ \text{ATT} \circ \text{Norm}(X) \tag{9}$$

**嵌入表示**通过多层块堆叠生成：

$$X^e := \text{Norm} \circ \text{Block}_M \circ \cdots \circ \text{Block}_1 \circ \text{Enc}(X) \tag{10}$$

**GPT 架构**的输出定义为：

$$\text{GPT}(X) := \arg\max_{\text{index}} \left( X^l = X^e W_{\text{head}} \right) \tag{11}$$

其中输出投影矩阵 $W_{\text{head}} \in \mathbb{R}^{D \times d}$

**维度说明**：

- $\text{ATT} : \{0,1\}^{T \times d} \to \mathbb{R}^{T \times d}$
- 所有残差操作保持维度不变性

Table: 典型 GPT 的参数设置

| Parameter | GPT-2(125M) | GPT-3/3.5(175B) | GPT-4(1800B) |
|---|---|---|---|
| d (vocab_size) | 50304 | * | * |
| T (block_size) | 1024 | 2048 | 8000(p.t.)->32000(f.t.) |
| D (n_embd) | 768 | 12288 | * |
| $D_V$ (n_embd) | 768/12=64 | 12288/96 =128 | * |
| $D_K$ (n_embd) | 768/12=64 | 12288/96 =128 | * |
| H (n_head) | 12 | 96 | * |
| L (MLP layer) | 2 | 2 | * |
| $W_1$ (first layer) | $\mathbb{R}^{4D\times D}$ | $\mathbb{R}^{4D\times D}$ | * |
| $W_2$ (second layer) | $\mathbb{R}^{D\times 4D}$ | $\mathbb{R}^{D\times 4D}$ | * |
| M (n_layer) | 12 | 96 | 120 |
| N(data_number) | 40G | 570G | * |

可以从以下两个角度理解损失函数形式：

- **交叉熵损失**（分类任务）：

$$L_{\text{CE}} = -\sum_{i=1}^{T}\sum_{j=1}^{d} P_{ij} \log \text{Softmax}(X^l)_{ij} \tag{12}$$

其中 $P_{ij} \in [0,1]$ 表示第 $i$ 个位置第 $j$ 个词元的真实概率分布。

- **负对数似然损失**（概率分布匹配）：

$$L_{\text{NLL}} = -\sum_{i=1}^{T} \log \text{Softmax}(X^l)_{i,j_{\text{true}}} \tag{13}$$

其中 $j_{\text{true}}$ 表示第 $i$ 个位置单词的索引号。