Transformer

Language Modeling (MLE)

$$p_\theta(\, y_{1:L} \mid x_{1:T}\,) = \prod_{i=1}^{L} p_\theta(\, y_l \mid y_{<l}; enc(x_{1:T})\,)$$

Attention

$$Attention(Q,K,V;a) = a\left(\frac{QK^\top}{\sqrt{d_q}}\right)V$$

$$Q = (q_1, q_2, ..., q_N)^\top \in \mathbb{R}^{N\times d_q} \quad W_Q \in \mathbb{R}^{d_X\times d_q} \quad X \in \mathbb{R}^{\mathrm{N}\times d_X}$$
$$K = (k_1, k_2, ... k_M)^\top \in \mathbb{R}^{M\times d_q} \quad W_K \in \mathbb{R}^{d_X\times d_q} \quad QK^\top \in \mathbb{R}^{N\times M}$$
$$V = (v_1, v_2, ..., v_N)^\top \in \mathbb{R}^{M\times d_v} \quad W_V \in \mathbb{R}^{d_X\times d_v} \quad Y \in \mathbb{R}^{N\times d_v}$$

Soft Attention: $a = \text{Softmax}$, Hard Attention: $a = \text{argmax}$

Complexity

Time: $(m\times n)\cdot(n\times p) : \mathcal{O}(mnp)$. Space: $(m\times n)\cdot(n\times p) : \mathcal{O}(mp)$

Time: $QK^\top$: $\mathcal{O}(NMd_q)$, $a(QK^\top)$: $\mathcal{O}(NM)$, AV: $\mathcal{O}(NMd_v)$ All: $\mathcal{O}(NMd_q + NMd_v)$

Space: $\mathcal{O}(MN + Nd_v)$

MHA

$$Multihead(Q,K,V;a) = concat(head_1, head_2, ...)W_O$$
$$head_i(Q,K,V;a) = Attention\big(QW_i^Q, KW_i^K, VW_i^V; a\big); h = |head|$$
$$W_i^Q \in \mathbb{R}^{d_q\times\tilde{d}_q} \quad W_i^K \in \mathbb{R}^{d_q\times\tilde{d}_q} \quad W_i^V \in \mathbb{R}^{d_v\times\tilde{d}_v} \quad W_O \in \mathbb{R}^{h\tilde{d}_v\times d_o}$$
$$Q^*: h\times N\times\tilde{d}_q \quad K^*: h\times M\times\tilde{d}_q \quad V^*: h\times M\times\tilde{d}_v \quad H: h\times N\times\tilde{d}_v$$

$h = 1$: Time: $\mathcal{O}(NM\tilde{d}_q + NM\tilde{d}_v)$. Space: $\mathcal{O}\big(MN + N\tilde{d}_v\big)$

Time: $\mathcal{O}\Big(h\Big(d_q\tilde{d}_q(M+N) + d_v\tilde{d}_vM\Big) + h\Big(NM\tilde{d}_q + NM\tilde{d}_v\Big) + Nh\tilde{d}_vd_o\Big)$

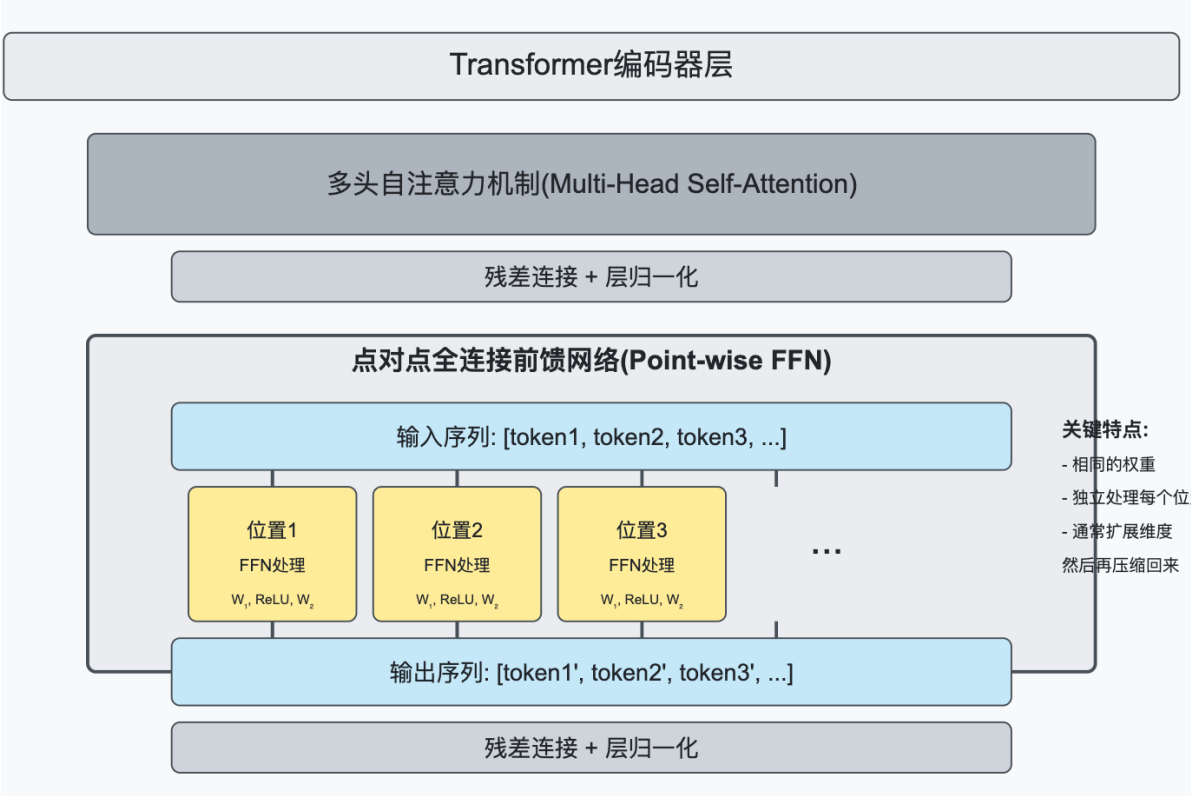
Projection + Attention + Output Projection

Space: $\mathcal{O}\Big(h\Big(\tilde{d}_q(M+N) + \tilde{d}_vM\Big) + h\Big(MN + N\tilde{d}_v\Big) + Nd_o\Big)$

Layer Norm

$$Y = \{y_1, y_2, ..., y_N\}, \qquad \text{where } y_i = XW + b$$
$$y_i \leftarrow \frac{y_i - \mu}{\sigma}, \qquad \text{where } \mu = \mathbb{E}Y, \sigma = \sqrt{\mathbb{E}_{y\in Y}[y-\mu]^2}$$
$$\text{Add \& Norm}(X) = LN(X + \text{Sublayer}(X))$$

Point-wise FFN (PFFN)



假设输入 X: $Batch \times Context \times Embedding : B \times C \times E$

FFN $f(X) = ReLU\big((XW_1 + b_1)W_2 + b_2\big) : E \to V$

$$(B \times C \times E) \xrightarrow{PFFN:E \to V} (B \times C \times V)$$