Named Tensors: Transformer

FFN

$$X \in \mathbb{R}^{\mathsf{emb}:d_{\text{model}}}$$

$$W^1 \in \mathbb{R}^{\mathsf{emb}:d_{\text{model}},\mathsf{hid}:d_{\text{ff}}} \qquad b^1 \in \mathbb{R}^{\mathsf{hid}:d_{\text{ff}}}$$

$$W^2 \in \mathbb{R}^{\mathsf{hid}:d_{\text{ff}},\mathsf{emb}:d_{\text{model}}} \qquad b^2 \in \mathbb{R}^{\mathsf{hid}:d_{\text{ff}}}$$

$$\text{FFN}(X; W, b) = W^2 \underset{\mathsf{hid}}{\cdot} \text{ReLU}(W^1 \underset{\mathsf{emb}}{\cdot} X + b^1) + b^2$$

Masked Attention

$$Q \in \mathbb{R}^{\mathsf{key}:d_k,\mathsf{seq}':n}, K \in \mathbb{R}^{\mathsf{key}:d_k,\mathsf{seq}:n}$$

$$V \in \mathbb{R}^{\mathsf{seq}:n,\mathsf{val}:d_v fg}, M \in \mathbb{R}^{\mathsf{seq},\mathsf{seq}'}$$

$$\text{att}(Q, K, V, M) = V \underset{\mathsf{seq}}{\cdot} \underset{\mathsf{seq}}{\text{softmax}} \left(\frac{Q \underset{\mathsf{key}}{\cdot} K}{\sqrt{d_k}} + M \right)$$

Multiheaded Self Attention

$$W^Q \in \mathbb{R}^{\mathsf{head}:h,\mathsf{emb}:d_{\text{model}},\mathsf{key}:d_k}$$

$$W^K \in \mathbb{R}^{\mathsf{head}:h,\mathsf{emb}:d_{\text{model}},\mathsf{key}:d_k}$$

$$W^V \in \mathbb{R}^{\mathsf{head}:h,\mathsf{emb}:d_{\text{model}},\mathsf{val}:d_k}$$

$$W^O \in \mathbb{R}^{\mathsf{head}:h,\mathsf{val}:d_k,\mathsf{emb}:d_{\text{model}}}$$

$$X \in \mathbb{R}^{\mathsf{seq}:n,\mathsf{emb}:d_{\text{model}}}$$

$$\text{MHA}(X; W) = \left[W^O \underset{\mathsf{head},\mathsf{val}}{\cdot} \text{att}(Q, K, V, M) \right]_{\mathsf{seq}' \to \mathsf{seq}}$$

$$Q = W^Q \underset{\mathsf{emb}}{\cdot} [X]_{\mathsf{seq} \to \mathsf{seq}'}$$

$$K = W^K \underset{\mathsf{emb}}{\cdot} X$$

$$V = W^V \underset{\mathsf{emb}}{\cdot} X$$

$$M_{\mathsf{seq}':i,\mathsf{seq}:j} = \begin{cases} 0 & j \le i \\ -\infty & \text{otherwise} \end{cases}$$

Layer Norm

$$X \in \mathbb{R}^{\mathsf{emb}:d_{\text{model}}} \qquad \gamma, \beta \in \mathbb{R}^{\mathsf{emb}:d_{\text{model}}}$$

$$\text{lnorm}(X; \gamma, \beta) = \frac{X - \underset{\mathsf{emb}}{\text{mean}}(X)}{\sqrt{\underset{\mathsf{emb}}{\text{var}}(X) + \epsilon}} \odot \gamma + \beta$$

Position Encoding

$$X \in \{0,1\}^{\mathsf{seq}:n,\mathsf{vocab}:b} \qquad \sum_{\mathsf{vocab}} X = 1$$

$$E \in \mathbb{R}^{\mathsf{vocab}*:v,\mathsf{emb}:d_{\text{model}}}$$

$$\text{embed}(X;E) = (E \underset{\mathsf{vocab}}{\cdot} X)\sqrt{d_{\text{model}}} + P$$

$$P \in \mathbb{R}^{\mathsf{seq}:n,\mathsf{hidden}:d_{\text{model}}}$$

$$P_{\mathsf{hidden}:i,\mathsf{seq}:p} = \begin{cases} \sin((p-1)/10000^{(i-1)/d_{\text{model}}}) & i \text{ odd} \\ \cos((p-1)/10000^{(i-2)/d_{\text{model}}}) & i \text{ even} \end{cases}$$

Transformer

$$I \in \{0,1\}^{\mathsf{seq}:n,\mathsf{vocab}:b} \qquad \sum_{\mathsf{vocab}} X = 1$$

$$\begin{aligned} X^0 &= \text{embed}(I) \\ T^1 &= \text{lnorm}(\text{MHA}(X^0)) + X^0 \\ X^1 &= \text{lnorm}(\text{FFN}(T^1)) + T^1 \\ &\vdots \\ T^L &= \text{lnorm}(\text{MHA}(X^{L-1})) + X^{L-1} \\ X^L &= \text{lnorm}(\text{FFN}(T^L)) + T^L \\ O &= \underset{\mathsf{vocab}}{\text{softmax}}(W \underset{\mathsf{emb}}{\cdot} X^L) \end{aligned}$$