BERT-Base Transformer Forward Pass

Ana Marasović

Initialize:

$W_T \in \mathbb{R}^{\text{vocab size} \times d} = \mathbb{R}^{\text{vocab size} \times 768} \ldots$ token embeddings
$W_P \in \mathbb{R}^{\text{max input length} \times d} = \mathbb{R}^{512 \times 768} \ldots$ positional embeddings
$h \in \{1, \ldots, n_{\text{heads}}\}, n_{\text{heads}} = 12$
$l \in \{1, \ldots, n_{\text{layers}}\}, n_{\text{layers}} = 12$
$W_{h,l}^{Q} \in \mathbb{R}^{d \times d_q} = \mathbb{R}^{768 \times 64} \ldots$ query *weight* matrices
$W_{h,l}^{K} \in \mathbb{R}^{d \times d_k} = \mathbb{R}^{768 \times 64} \ldots$ key *weight* matrices
$W_{h,l}^{V} \in \mathbb{R}^{d \times d_q} = \mathbb{R}^{768 \times 64} \ldots$ value *weight* matrices
$W_{l}^{ffnn} \in \mathbb{R}^{d \times d_{ffnn}} = \mathbb{R}^{768 \times 3072} \ldots$ feedforward layer's weight matrix
$b_{l}^{ffnn} \in \mathbb{R}^{1 \times d_{ffnn}} = \mathbb{R}^{1 \times 3072} \ldots$ feedforward layer's bias vector
$W_{l}^{out} \in \mathbb{R}^{d_{ffnn} \times d} = \mathbb{R}^{3072 \times 768} \ldots$ output layer's weight matrix
$b_{l}^{out} \in \mathbb{R}^{1 \times d} = \mathbb{R}^{1 \times 768} \ldots$ output layer's bias vector
$W^{final} \in \mathbb{R}^{d \times d} = \mathbb{R}^{768 \times 768} \ldots$ final layer's weight matrix
$I = (i_1, \ldots, i_{512}) \in \mathbb{N}_0^{1 \times \text{max input length}} = \mathbb{N}_0^{1 \times 512} \ldots$ input vocab indices
$T = \texttt{lookup}(W_T, I) \in \mathbb{R}^{\text{max input length} \times d} = \mathbb{R}^{512 \times 768} \ldots$ input token embeddings
$X = T + W_P \in \mathbb{R}^{\text{max input length} \times d} = \mathbb{R}^{512 \times 768} \ldots$ input embeddings
$Z_0 = X$

Forward algorithmm:

for ($l = 1;\ l \le n_{layers} = 12;\ l++$) {

for ($h = 1;\ h \le n_{heads} = 12;\ h++$) {

$$Q_{h,l} = Z_{l-1} W_{h,l}^{Q} \in \mathbb{R}^{\text{max input len} \times d_q} = \mathbb{R}^{512 \times 64} \ldots \text{query matrix}$$

$$K_{h,l} = Z_{l-1} W_{h,l}^{K} \in \mathbb{R}^{\text{max input len} \times d_k} = \mathbb{R}^{512 \times 64} \ldots \text{key matrix}$$

$$V_{h,l} = Z_{l-1} W_{h,l}^{V} \in \mathbb{R}^{\text{max input len} \times d_v} = \mathbb{R}^{512 \times 64} \ldots \text{value matrix}$$

$$A_{h,l} = \mathtt{Softmax}\left(\frac{Q_{h,l} K_{h,l}^{T}}{\sqrt{d_k}}\right) \in \mathbb{R}^{\text{max input len} \times \text{max input len}} = \mathbb{R}^{512 \times 512}$$

$$Z_{h,l} = A_{h,l} V_{h,l} \in \mathbb{R}^{\text{max input len} \times d_v} = \mathbb{R}^{512 \times 64}$$

}

$$\tilde{Z}_l = \mathtt{concat}(Z_{1,l}, \ldots, Z_{n_{\text{heads}},l}) \in \mathbb{R}^{\text{max input len} \times (d_v \cdot n_{\text{heads}})} = \mathbb{R}^{512 \times (64 \cdot 12)} = \mathbb{R}^{512 \times 768}$$

$$\bar{Z}_l = \mathtt{LayerNorm}(Z_{l-1} + \tilde{Z}_l) \in \mathbb{R}^{512 \times 768}$$

$$Z_l^{ffnn} = \max(0, \bar{Z}_l W_l^{ffnn} + b_l^{ffnn}) \in \mathbb{R}^{\text{max input len} \times d_{ffnn}} = \mathbb{R}^{512 \times 3072}$$

$$Z_l^{out} = Z_l^{ffnn} W_l^{out} + b_l^{out} \in \mathbb{R}^{\text{max input len} \times d} = \mathbb{R}^{512 \times 768}$$

$$Z_l = \mathtt{LayerNorm}(\bar{Z}_l + Z_l^{out}) \in \mathbb{R}^{512 \times 768}$$

}

Pass $\tanh(W^{final} Z_{n_{\text{layers}}}[0,:])$ to the final `Softmax` that predicts the class, where $Z_{n_{\text{layers}}}[0,:]$ is the hidden state corresponding to the first token.