Transformer Architecture

$\vec{p_1}$ $\vec{p_2}$ $\vec{p_3}$ $\vec{p_4}$ logits

$\sigma(\vec{z_i}) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$ | softmax |

$LN(\vec{x}) = \frac{\vec{x} - ExpectedValue(\vec{x})}{\sqrt{Var(\vec{x})}}$ | Linear/LayerNorm |

$\vec{f_1}$ $\vec{f_2}$ $\vec{f_3}$ $\vec{f_4}$

contextualized word embedding $\vec{m_1}$ $\vec{m_2}$ $\vec{m_3}$ $\vec{m_4}$

Transformer Block

MLP

$\vec{m_i} = W_2 \cdot GeLu(W_1 \cdot \vec{h_i})$

$\vec{h_1}$ $\vec{h_2}$ $\vec{h_3}$ $\vec{h_4}$

repeat

| LayerNorm |

relevance matrix w/ causal masking

$$
\begin{bmatrix}
\vec{q_1}\cdot\vec{k_1} & \vec{q_1}\cdot\vec{k_2} & \vec{q_1}\cdot\vec{k_3} & \vec{q_1}\cdot\vec{k_4} \\
\vec{q_2}\cdot\vec{k_1} & \vec{q_2}\cdot\vec{k_2} & \vec{q_2}\cdot\vec{k_3} & \vec{q_2}\cdot\vec{k_4} \\
\vec{q_3}\cdot\vec{k_1} & \vec{q_3}\cdot\vec{k_2} & \vec{q_3}\cdot\vec{k_3} & \vec{q_3}\cdot\vec{k_4} \\
\vec{q_4}\cdot\vec{k_1} & \vec{q_4}\cdot\vec{k_2} & \vec{q_4}\cdot\vec{k_3} & \vec{q_4}\cdot\vec{k_4}
\end{bmatrix}
$$

$\div\sqrt{d_k}$ softmax $\times$ V $\times$ W^O $\rightarrow$

attention matrix

$$
\begin{bmatrix}
a_{1,1} & a_{1,2} & a_{1,3} & a_{1,4} \\
a_{2,1} & a_{2,2} & a_{2,3} & a_{2,4} \\
a_{3,1} & a_{3,2} & a_{3,3} & a_{3,4} \\
a_{4,1} & a_{4,2} & a_{4,3} & a_{4,4}
\end{bmatrix}
$$

| LayerNorm |

$\vec{z_1}$ $\vec{z_2}$ $\vec{z_3}$ $\vec{z_4}$ updated word embedding

$\vec{q_1}$ $\vec{z_1}$ W^Q

$\Box\Box\Box\Box = \Box\Box\Box\Box \times$

$\vec{k_1}$ $\vec{z_1}$ W^K

$\Box\Box\Box\Box = \Box\Box\Box\Box \times$

$\vec{v_1}$ $=$ $\vec{z_1}$ $\times$ W^V

$$
\begin{bmatrix} \vec{p_1} \\ + \\ \vec{w_1} \end{bmatrix}
\begin{bmatrix} \vec{p_2} \\ + \\ \vec{w_2} \end{bmatrix}
\begin{bmatrix} \vec{p_3} \\ + \\ \vec{w_3} \end{bmatrix}
\begin{bmatrix} \vec{p_4} \\ + \\ \vec{w_4} \end{bmatrix}
$$

position embedding

word embedding

The   ring   was   loud

$\times$

(input)