Attention

$$\frac{\sum_j e^{\langle q, k_j \rangle}}{Z} v_j = \sum_j \text{softmax}(\langle q, k_\cdot \rangle)_j \, v_j \qquad (AT)$$

s_j $\quad \sum_j s_j = \sum_j s_{\tau(j)}$

Self-Attention

$X \in \mathbb{R}^{T \times d_x}$

$W^Q \in \mathbb{R}^{d_x \times d_k}$

$W^K \in \mathbb{R}^{d_x \times d_k}$

$W^V \in \mathbb{R}^{d_x \times d_v}$

$$\underset{\text{axis}=-1}{\text{softmax}}\left(XW^Q \cdot (XW^K)^T \right) \cdot XW^V \qquad (SAT)$$

$T \times d_k \quad d_k \times T \quad (T \times d_k)^T \quad T \times d_v$

cost $O(d_k \cdot T^2)$

Multihead

$$\left(SAT_1(X), \dots, SAT_H(X)\right), \quad M \sim H \cdot d_v \times d_x$$

$T \times H \cdot d_v \qquad (\dots) \cdot M$

$+ O(T^2 d_v)$

→ quadratic cost severely constrains the feasible context length .... more on this soon

Positional encoding

Note that (AT) is invariant to permuting the key/value pairs.

(SAT) is equivariant to permuting X:

$$Y_{ti} := X_{\sigma(t)i} \qquad \text{some } \sigma \in S_T$$

$$\Rightarrow \quad SAT(Y)_{ti} = SAT(X)_{\sigma(t)i}.$$

For many applications one wants to break this invariance/equivariance.

One wants to encode the positions of things.

↝ instead of above

$$q_m = f_q(x_m) \qquad x_m \cdot W^Q$$
$$\vdots$$
$$v_m = f_v(x_m) \qquad x_m \cdot W^V$$

take

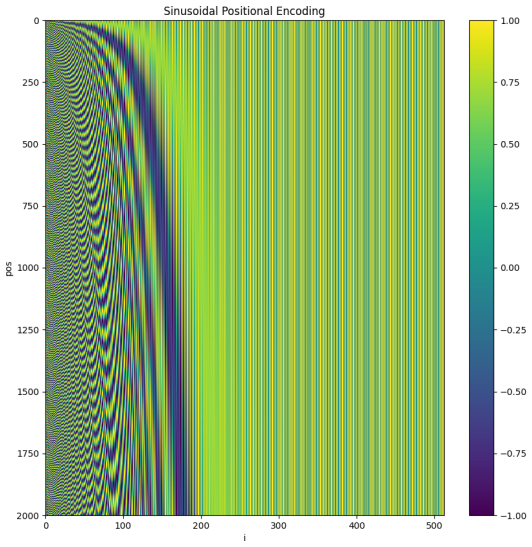
$$q_m = f_q(x_m, m)$$
$$\vdots$$
$$v_m = f_v(x_m, m)$$

sin/cos (additive)

Already in [vaswani2017attention] the following positional encoding is suggested.

$$\mathsf{pe}_{2i}^{\sin/\cos} = \left(\sin\left(\frac{\mathsf{pos}}{10000^{2i/d_{model}}}\right)\right)$$
$$\mathsf{pe}_{2i+1} = \left(\cos\left(\frac{\mathsf{pos}}{10000^{2i/d_{model}}}\right)\right), \qquad i = 0, \ldots, d_{model}/2 - 1, \mathsf{pos} = 1, \ldots, d_{model}.$$

$$f_\eta(x, m) = \cdots \quad (x + \mathsf{pe}_m) \cdot W^\eta$$



sin/cos (multiplicative)

$$f_\eta(x, m) = R_m W_\eta x_m$$

where

$$R_m = \begin{pmatrix} \cos(m\theta_1) & \sin(m\theta_1) & 0 & \ldots & 0 & 0 \\ -\sin(m\theta_1) & \cos(m\theta_1) & 0 & \ldots & 0 & 0 \\ 0 & 0 & \cos(m\theta_2) & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \cos(m\theta_{d/2}) & \sin(m\theta_{d/2}) \\ \ldots & \ldots & \ldots & \ldots & -\sin(m\theta_{d/2}) & \cos(m\theta_{d/2}) \end{pmatrix},$$

$\theta_i := 10000^{-2(i-1)/d}, i = 1, \ldots, d/2.$

Transformer architecture

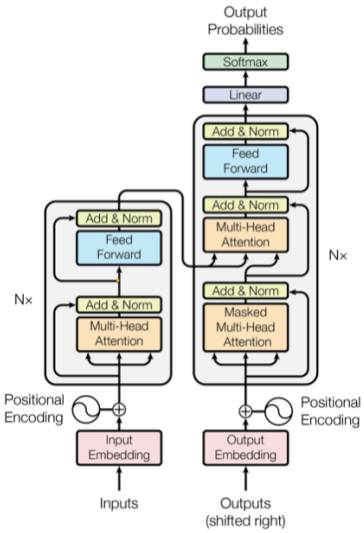


Figure 1: The Transformer - model architecture.

Vaswani et al '17

GPT, Llama .. modify this original Transformer architecture in that they are "decoder-only", they only have the right track (and discard the second Multi-Head-Attention, which is also called Cross-Attention

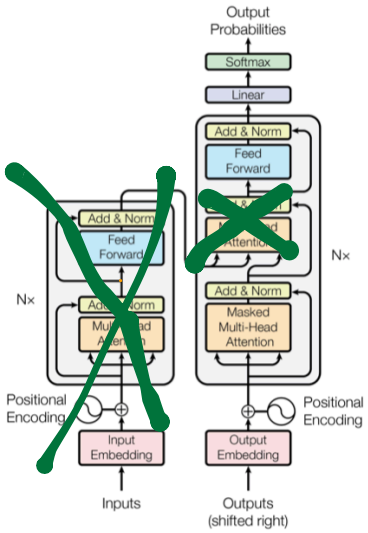


Figure 1: The Transformer - model architecture.

~> bck at makemore