Transformer

$$\text{output} = \text{Attention}\left(\overset{\text{„query"}}{q}, \overset{\text{„keys"}}{k}, \overset{\text{„values"}}{v}\right)$$

$$\alpha_t = \frac{\exp\left(q^T k_t / \sqrt{\dim(q)}\right)}{\sum_{j=1}^{T} \exp\left(q^T k_j / \sqrt{\dim(q)}\right)}$$

$$\text{output} = \sum_{j=1}^{T} \alpha_j \cdot v_j$$

$$q = \tilde{h}_{k-1}$$

$$k = \{h_1, h_2, \dots, h_T\}$$

$$v = \{h_1, h_2, \dots, h_T\}$$

$$Q \in \mathbb{R}^{\tilde{T} \times H}, \quad K \in \mathbb{R}^{T \times H}, \quad V \in \mathbb{R}^{T \times H}$$

$$O = \text{Attention}(Q, K, V)$$

$$\Leftrightarrow \quad o_k = \text{Att}(q_k, K, V) \qquad k = 1, \dots, \tilde{T}$$

$$\underset{\underset{\mathbb{R}^{\tilde{T} \times H}}{\uparrow}}{O} = \underset{\underset{\text{rowwise}}{\uparrow}}{\text{Softmax}} \Big(\underset{\underset{\mathbb{R}^{\tilde{T} \times T}}{\uparrow}}{Q \cdot K^T} / \sqrt{H} \Big) \cdot \underset{\underset{\mathbb{R}^{T \times H}}{\uparrow}}{V}$$

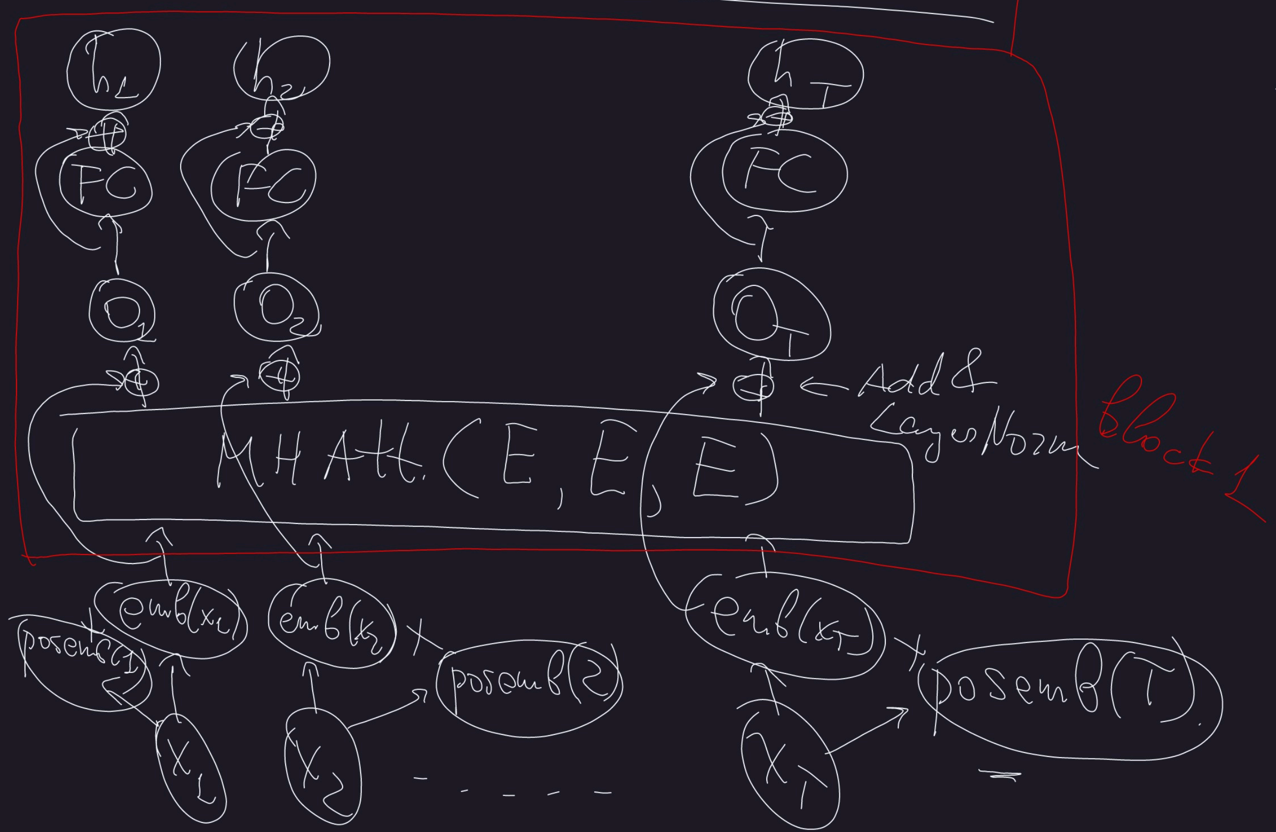
Multi Head Attention

$$\underset{\in \mathbb{R}^{\tilde{T} \times H_{head}}}{head_i} = Attention(\underset{\tilde{T}\times H}{Q} \cdot \underset{H \times H_{head}}{W_i^Q}, \underset{T \times H}{K} \cdot \underset{H \times H_{head}}{W_i^K}, \underset{T \times H}{V} \cdot \underset{H \times H_{head}}{W_i^V})$$

$\{W_i^Q, W_i^K, W_i^V\}_{i=1}^{n}$ – trainable params

$$\underset{\in \mathbb{R}^{\tilde{T} \times H}}{O} = MultiHeadAttention = Concat(\underset{\tilde{T} \times H_{head}}{head_1}, \underset{\tilde{T} \times H_{head}}{head_2}, \ldots, head_n) \cdot \underset{n \cdot H_{head} \times H}{W^O}$$

Transformer encoder



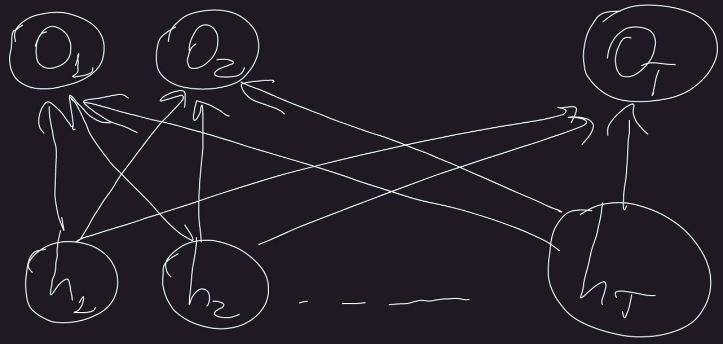
$x_t,\ t=1,\dots,T$

$h_t^0 = emb(x_t)$

for $\ell = 1, 2, \dots, L$:

$$O^\ell = LayerNorm\left(H^{\ell-1} + MHAtt.(H^{\ell-1}, H^{\ell-1}, H^{\ell-1})\right)$$

$$H^\ell = LayerNorm\left(O^\ell + g(\underset{T\times H}{O^\ell} \cdot \underset{H\times H}{W_1} + \underset{H\times 1}{b_1})\underset{H\times H}{W_2} + \underset{H\times 1}{b_2}\right)$$



$$h_{t+1} = g\left(\underset{H\times H}{W^h} \cdot \underset{H\times 1}{h_t} + W^x \cdot x_{t+1} + b\right)$$

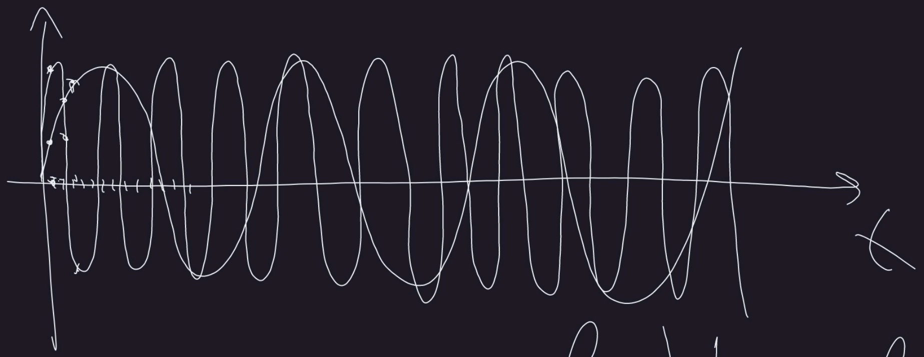
	RNN	MH Self Att.
Comp complexity:	$O(T\cdot H\cdot H)$	$O(n(T\cdot H\cdot H_{head} + T\cdot T\cdot H_{head}))$
		$= O(T\cdot H\cdot H + T\cdot T\cdot H)$
		$nH_{head} = H$

$T \quad H \quad H_{head} \quad n$

$$posemb_{t,2i} = \sin(f_i \cdot t)$$

$$posemb_{t,2i+1} = \cos(f_i \cdot t)$$

$$f_i = 10000^{\frac{emb\_dim}{2i}}$$



$$posemb_{t+k} = W_k \cdot posemb_t \Longleftarrow$$

$$posemb_{t+k,2i} = \sin(f_i(t+k)) =$$

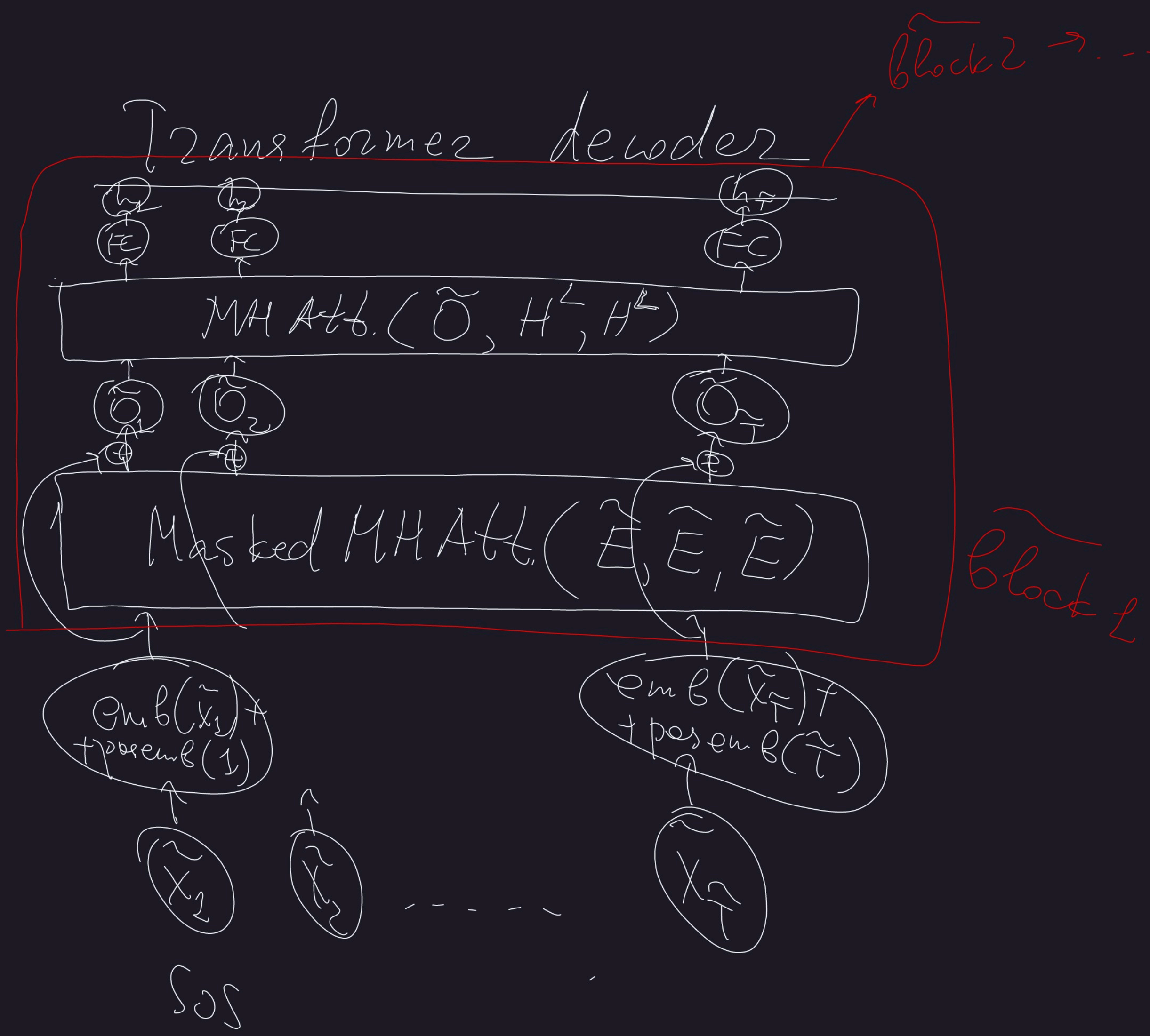
$$= \sin(f_i \cdot t)\cos(f_i \cdot k) +$$

$$+ \cos(f_i \cdot t) \cdot \sin(f_i \cdot k) =$$

$$= posemb_{t,2i} \cdot \cos(f_i k) +$$

$$+ posemb_{t,2i+1} \cdot \sin(f_i k)$$

Transformer decoder


block 2 → - - - → block L
h_1 h_2 h_T
FC FC FC
MH Att. $(\tilde{O}, H^L, H^L)$
$\tilde{O}_1$ $\tilde{O}_2$ $\tilde{O}_T$
⊕ ⊕ ⊕
Masked MHAtt. $(\tilde{E}, \tilde{E}, \tilde{E})$
block l
$emb(\tilde{x}_1)+$ $+posemb(1)$
$emb(\tilde{x}_T)+$ $+posemb(T)$
$\tilde{x}_1$ $\tilde{x}_2$ - - - - - $\tilde{x}_T$
SOS

$\tilde{x}_t \;, \quad t=1, \dots, T$

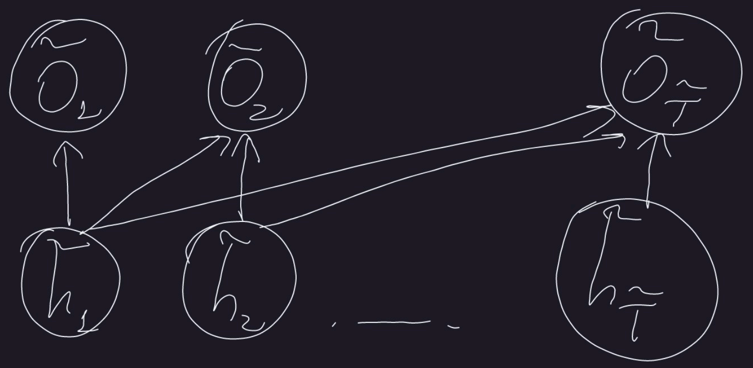
$\tilde{h}_t^0 = emb(\tilde{x}_t) + posemb(t)$

for $l = 1, 2, \dots, L$:

$$\tilde{O}^l = LayerNorm\big(\tilde{H}^{l-1} + + MaskedMHAtt.(\tilde{H}^{l-1}, \tilde{H}^{l-1}, \tilde{H}^{l-1})\big)$$

$$\tilde{O}^l = LayerNorm\big(\tilde{O}^l + + MHAtt.(\tilde{O}^l, H^L, H^L)\big)$$

$$\tilde{H}^l = LayerNorm\big(\tilde{O}^l + + g(\tilde{O}^l \cdot W_1 + b_1) \cdot W_2 + b_2\big)$$



$$m:\ \tilde{T}\ \begin{array}{c}\tilde{T}\\ \left[\begin{array}{ccccc} 0 & -\infty & -\infty & \cdots & -\infty \\ & 0 & -\infty & -\infty & -\infty \\ & & \ddots & & \\ 0 & & & \ddots & -\infty \\ & & & & 0 \end{array}\right]\end{array}$$

$$\tilde{O} = \text{Masked Attention}\left(\tilde{H}, \tilde{H}, \tilde{H}\right) =$$

$$= \text{Softmax}\left(\tilde{H}\cdot\tilde{H}^{T} / \sqrt{dim} + m\right)\cdot\tilde{H}$$