· Self - Attention

input vectors X: $N_x \times D_x$ — N_x is the sequence length, not batch size

key matrix W_K: $D_x \times D_a$

value matrix W_V: $D_x \times D_v$

query matrix W_Q: $D_x \times D_a$

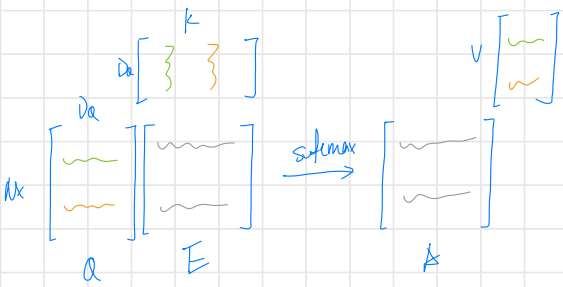
query: $Q = X \cdot W_Q$

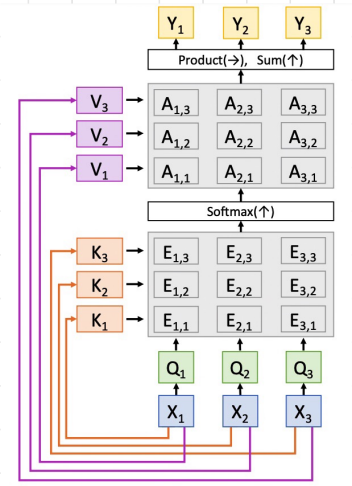
key: $K = X \cdot W_K$

value: $V = X \cdot W_V$

similarities $E = QK^T / \sqrt{D_a}$

$A = \text{softmax}(E, \text{dim}=1)$





Self-attention is permutation equivariant

· Positional Encoding

requirement: the encoding is deterministic

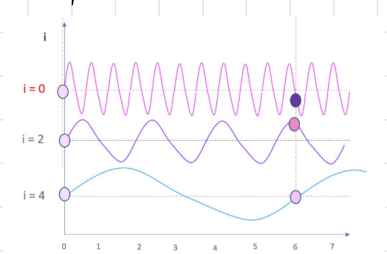
output a unique encoding for each time-step

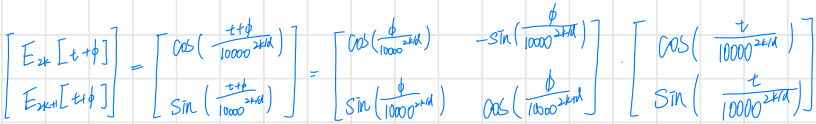
distance between any two timesteps should be consistent across sequences of different lengths

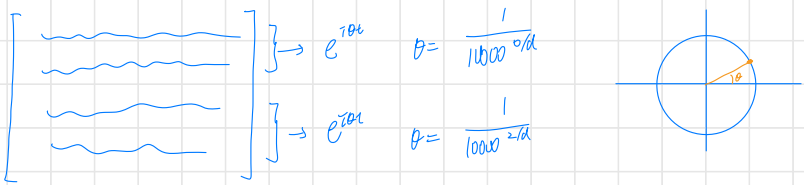
let $E_t \in \mathbb{R}^d$ be the positional encoding at time-step t

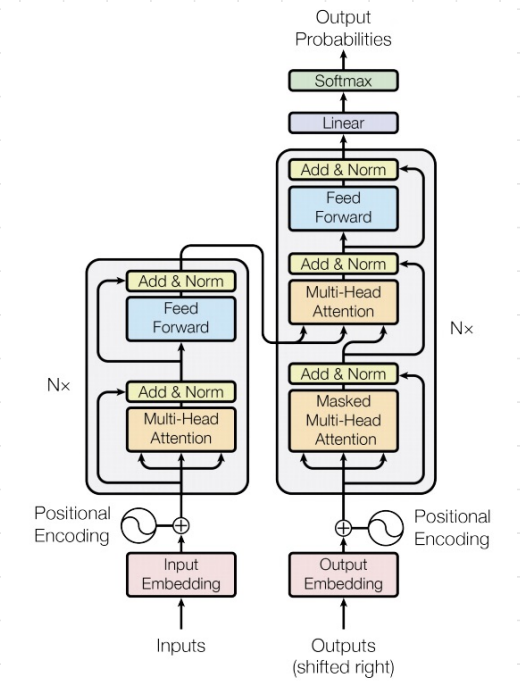
$$E_{2i}[t] = \sin\left(\frac{t}{10000^{2i/d}}\right)$$

$$E_{2i+1}[t] = \cos\left(\frac{t}{10000^{2i/d}}\right)$$








Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)