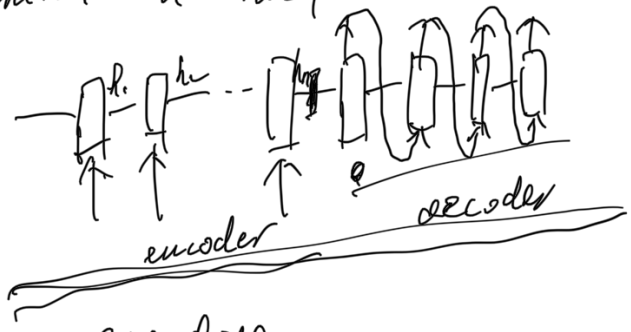
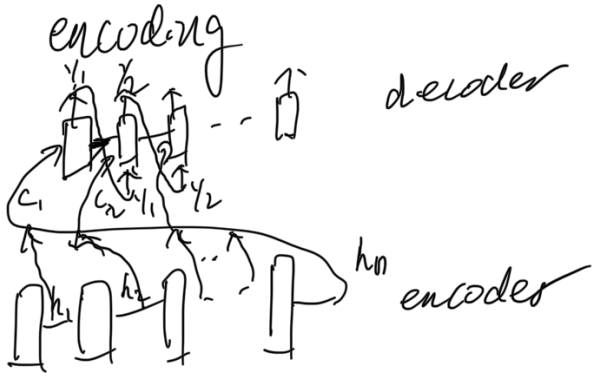
Lecture 15

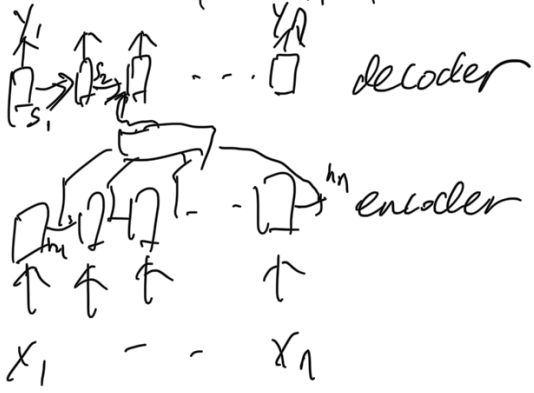
Attention and Transformers





Input sequence $x = [x_1, x_2, \dots, x_n]$, length $= n$

output sequence $y = [y_1, y_2, \dots, y_m]$, length $= m$



$$y_i = f(s_i), \quad s_i = g(s_{i-1}, c_i, y_{i-1})$$

$$\Rightarrow c_i = \sum_{k}^{n} a_{i,k} h_k$$

$k=1$ ↑ weights

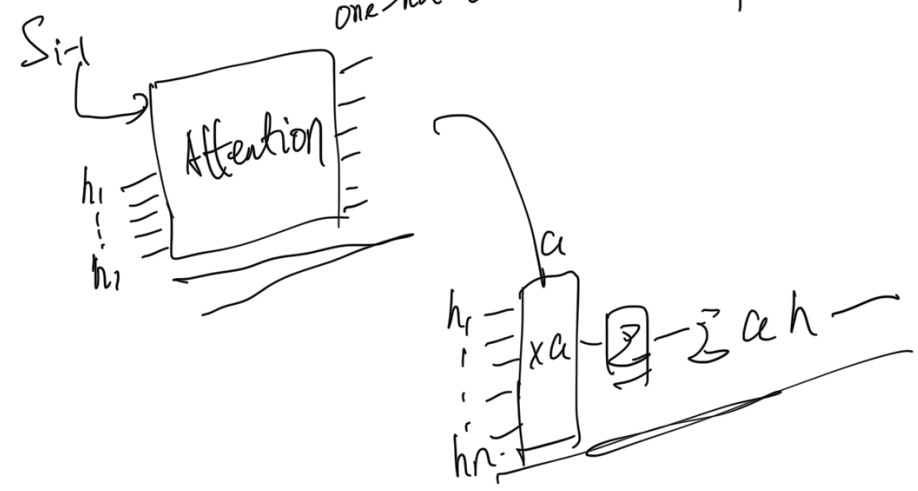
O_1 — learn $a_{i,k}$ directly from data, then a is fixed. think A is a matrix of a_{ik}.

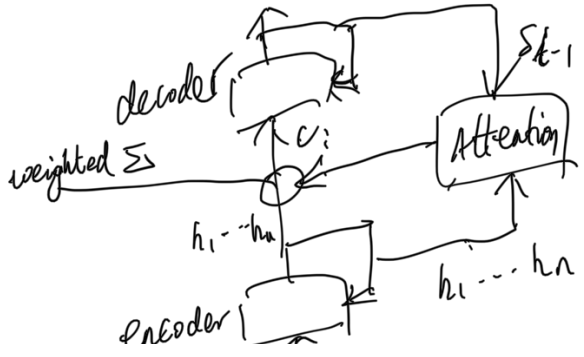
✓ O_2 — relate $a_{i,k}$ to some variables. A is a function

$$a_{i,k} = \text{align}(y_i, x_k)$$

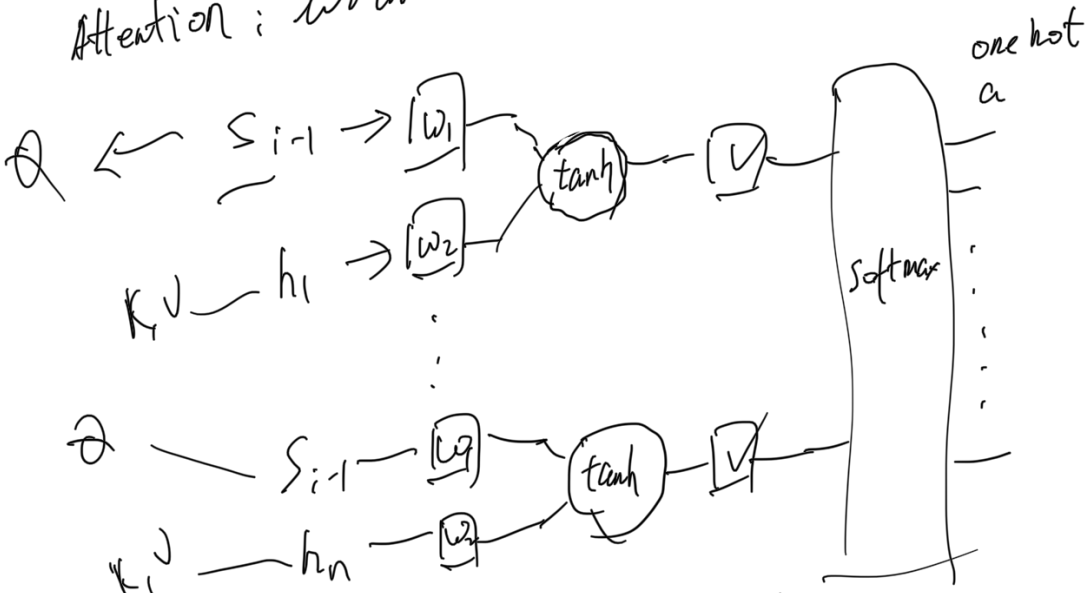
$$= \frac{\exp(\text{score}(s_{i-1}, h_k))}{\sum \cdots} \leftarrow \text{softmax}$$

$$\text{score}(s_{i-1}, h_k) = v \tanh(w_1 s_{i-1} + w_2 h_k) \leftarrow$$





Attention : with additive



learn W_1, W_2, V

several attention mechanisms, (score functions)

content-based attention $score(s_{i-1}, h_k) = \text{cosine}(s_{i-1}, h_k)$

General $score(s_{i-1}, h_k) = s_{i-1}^T W h_k$

W: to be learned

Dot-Product $score(s_{i-1}, h_k) = s_{i-1}^T h_i$

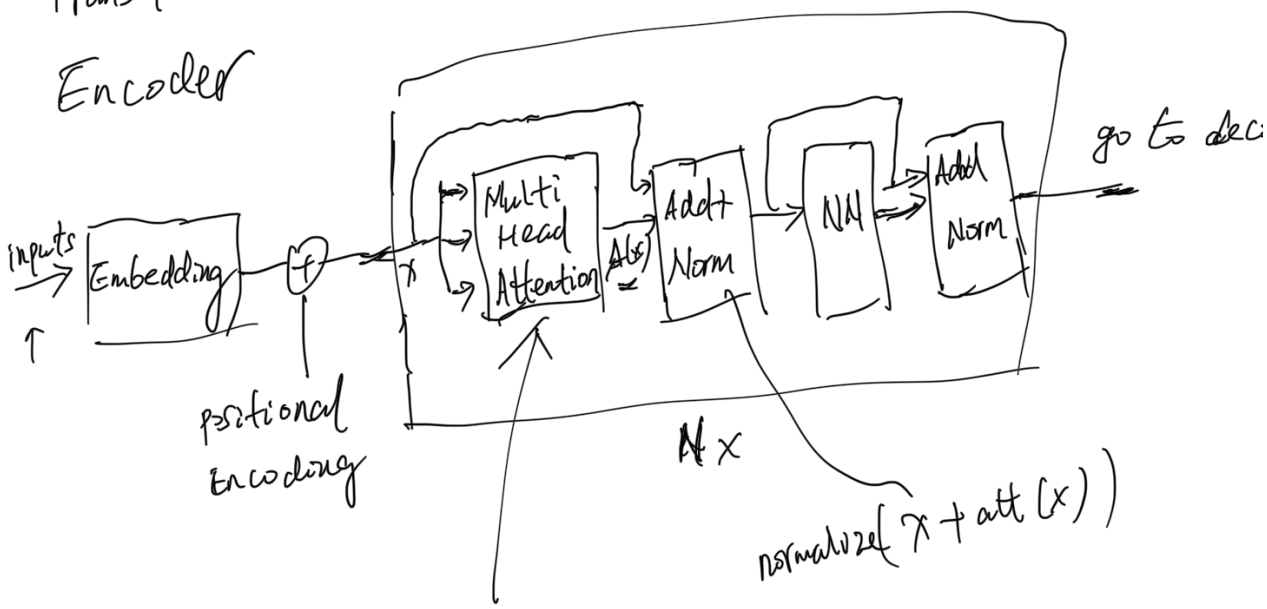
Scaled Dot product : $score(s_{i-1}, h_k) = \frac{s_{i-1}^T h_i}{\sqrt{n}}$

Global/Soft $\rightarrow$ $h_1 \cdots h_n$

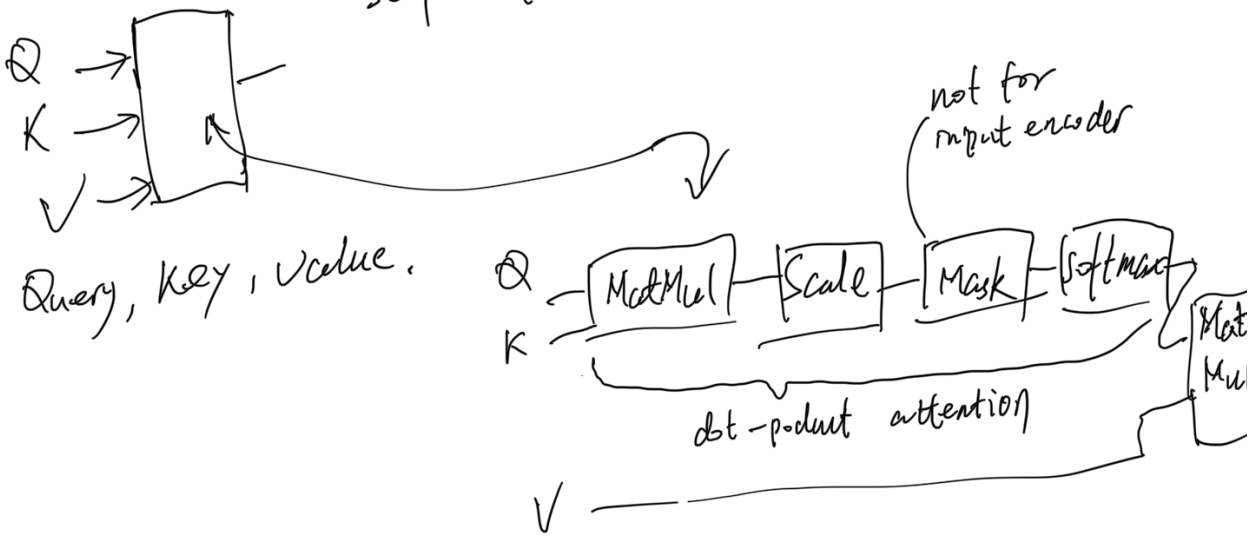
Local/Hard $\rightarrow$ $h_i \cdots h_j$ subset of $1, \cdots, n$

Transformer

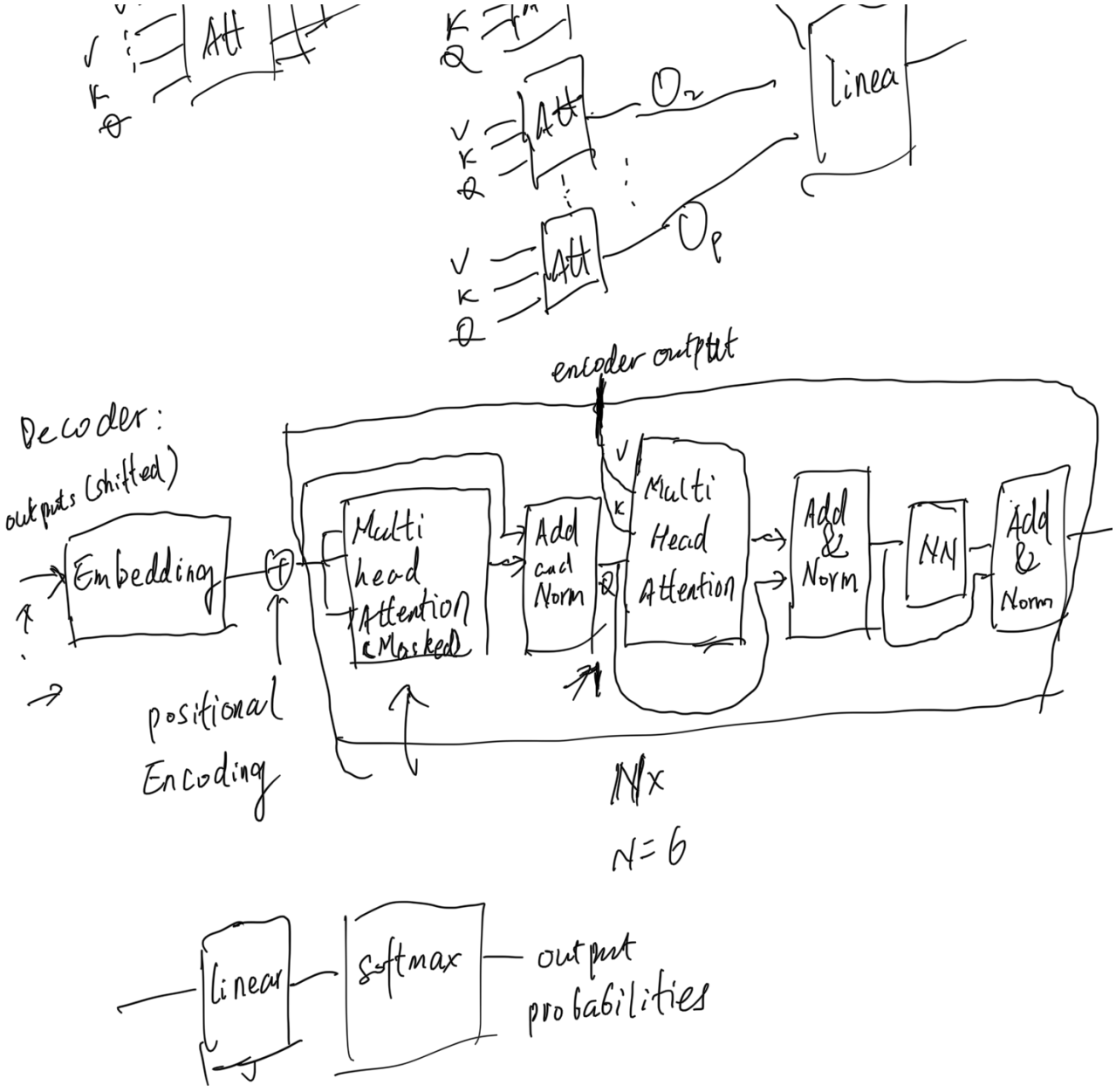
Encoder



self-Attention.



$$\text{Attention}(Q, K, V) = \underbrace{\text{softmax}\left(\frac{QK^T}{\sqrt{n}}\right)}_{a} V$$



Suppose the output is "I live in Tampa"

	I	live	in	Tamp
	1	2	3	4	5
Mask:					
For 1st input	0	0	0	0	
2nd	1	0	0	0	
3rd	1	1	0	0	

positional encoding

$$f(t) = \begin{cases} \sin(\omega_k t) & \text{if } i = 2k \\ \cos(\omega_k t) & \text{if } i = 2k+1 \end{cases}$$

$$\begin{bmatrix} \sin(\omega_1 t) \\ \cos(\omega_1 t) \\ \sin(\omega_2 t) \\ \cos(\omega_2 t) \\ \vdots \\ \sin(\omega_{d/2} t) \\ \cos(\omega_{d/2} t) \end{bmatrix} \quad \omega_k = \frac{1}{10000^{2k/d}}$$