Plan

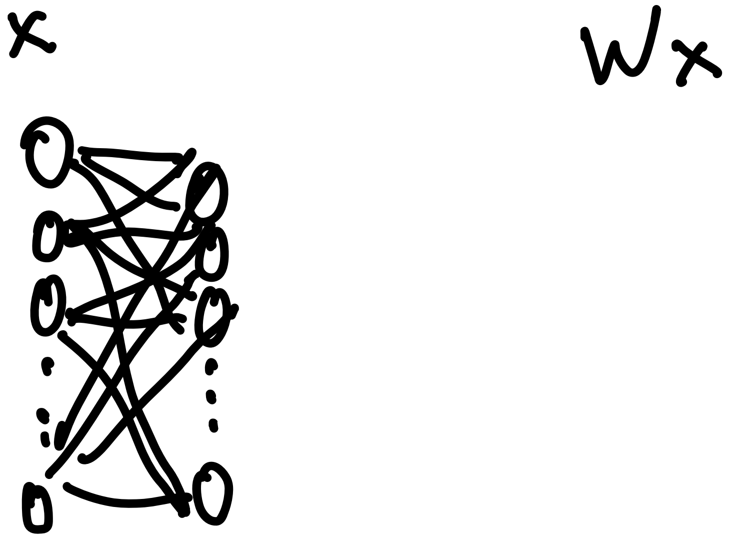
Recap

Reminders

Embedding

Motivation

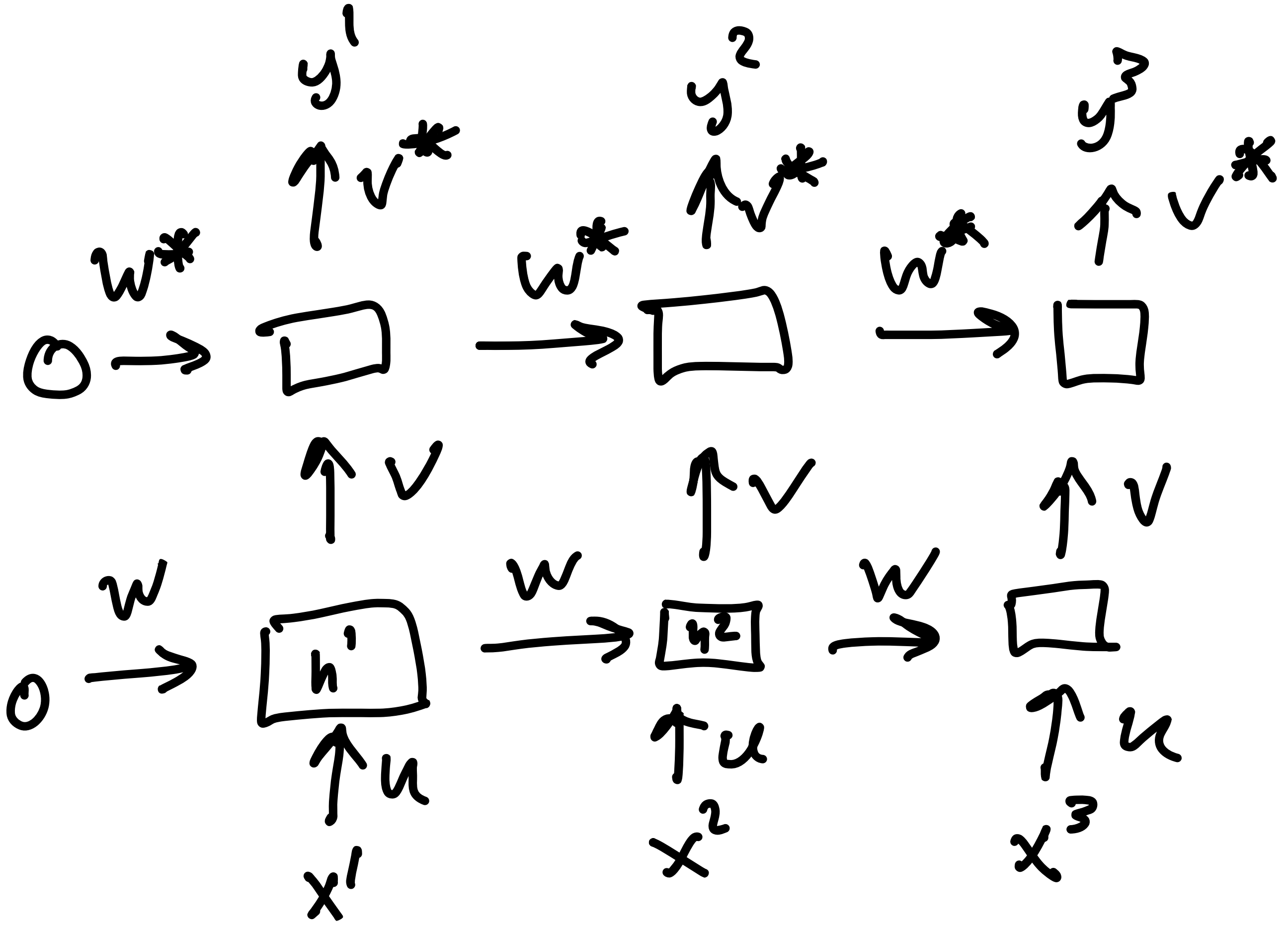
Self-attention



Recap

1) How to encode text?

2) How to include long range dependency?



$$h^2 = \sigma(Wh^1 + Ux^2)$$

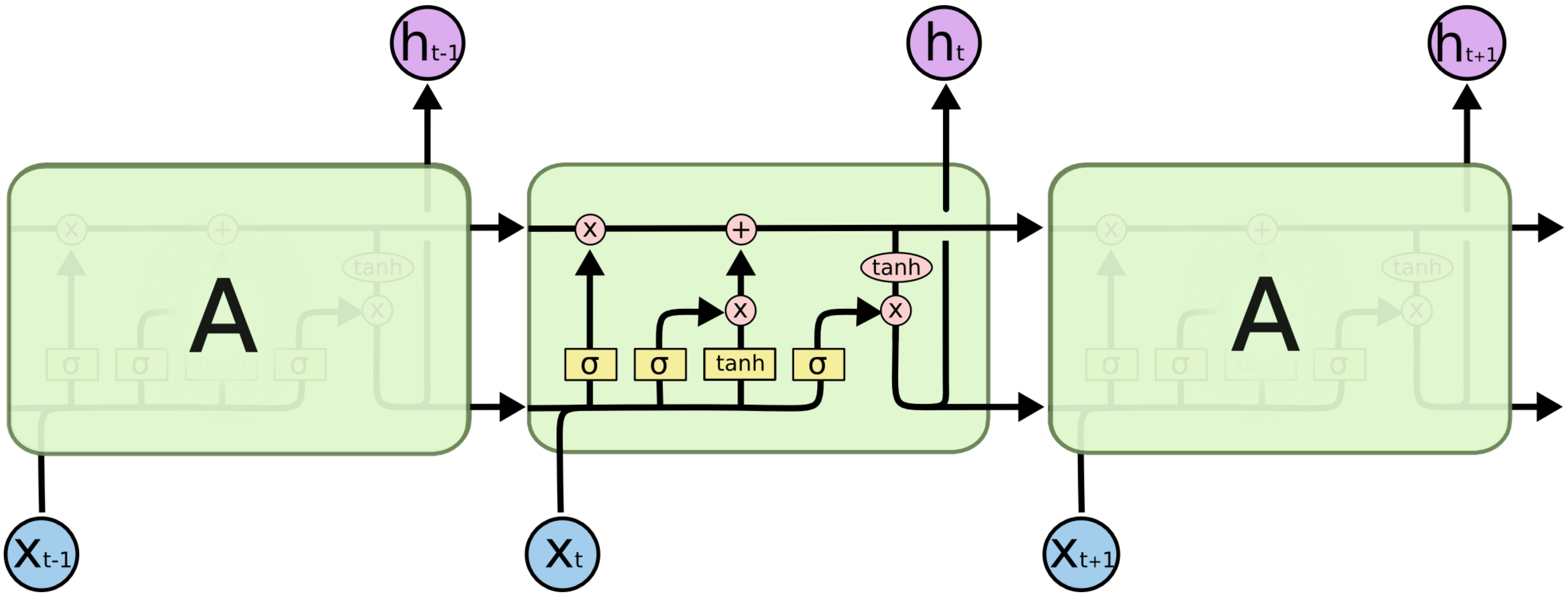
RNN

- variable length
- memory (but overwritten)

LSTM

cell state

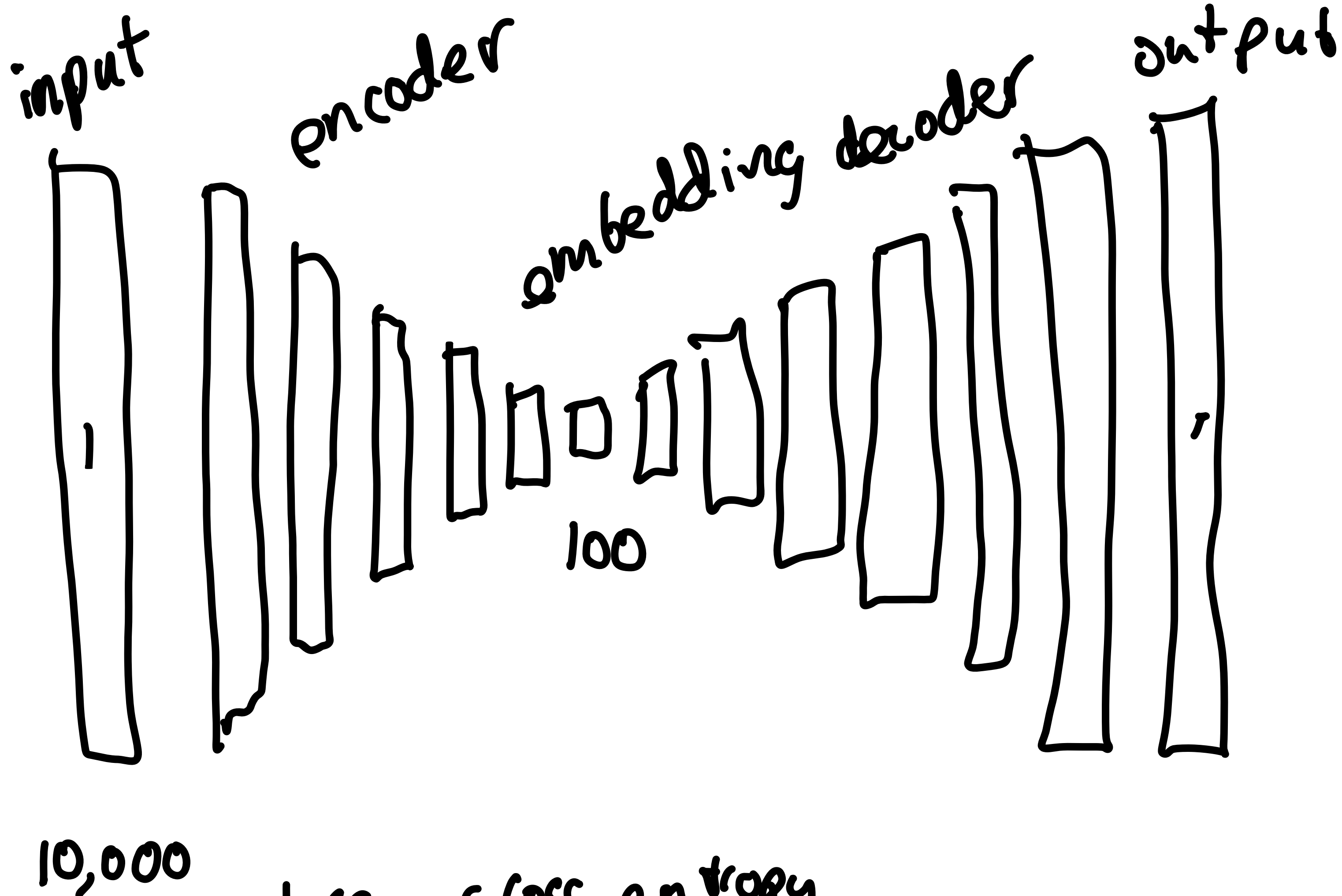
$$\begin{bmatrix} \\ \\ \\ \end{bmatrix} * \begin{bmatrix} 1 \\ 0 \\ .5 \end{bmatrix}$$



Reminders

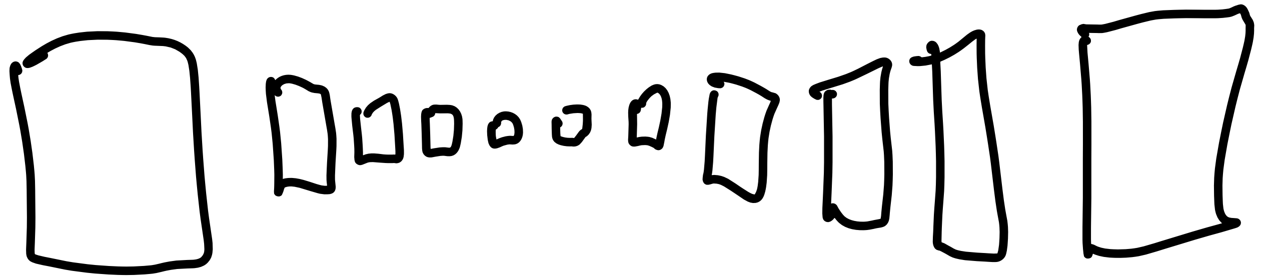
- Form 24/26
- 24 hours <u>total</u> lateness
- work in 202 during office hours
- Project proposal due Monday

Embedding



loss = cross entropy

- l_2 norm $= ||\text{input} - \text{output}||_2^2$



RNN work well on

- next word prediction
- sequence classification

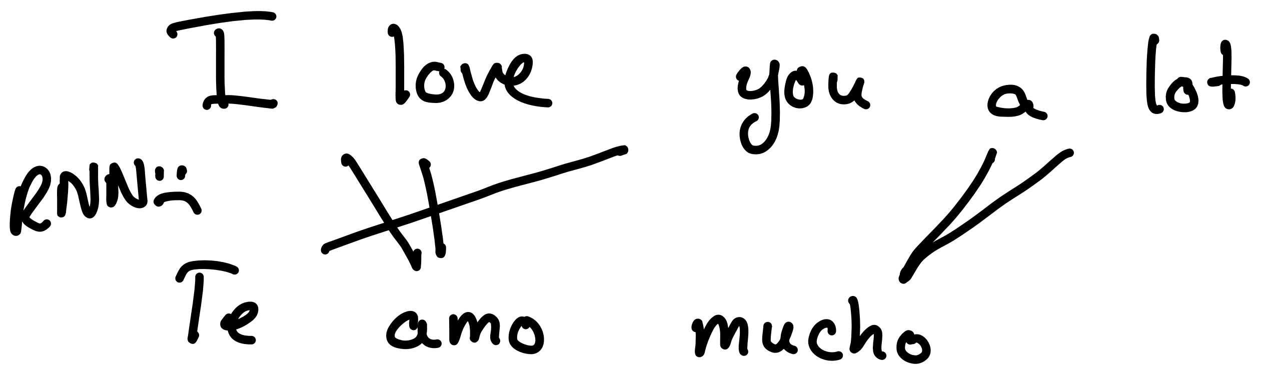
but less well on

- translation
- sentence generation

The dog ran really fast

RNN ☺

El perro corrio muy rapido



I love you a lot

RNN ☹

Te amo mucho

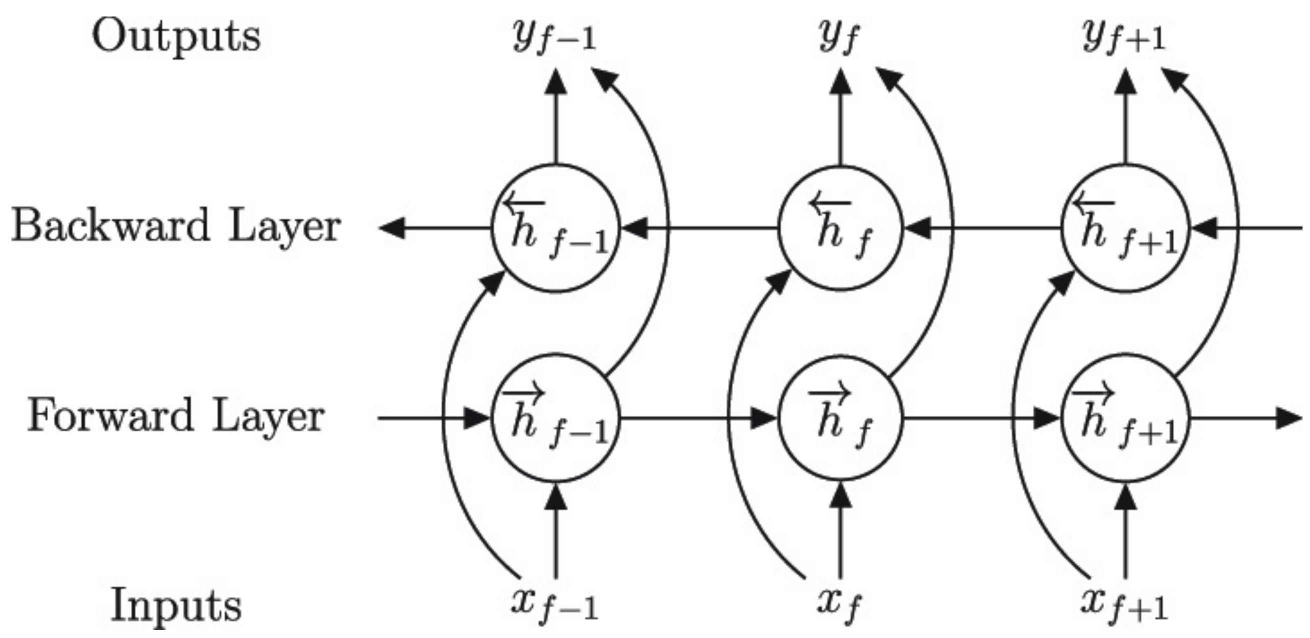
1. different # words 5 vs 3
2. mixed order
3. need whole context

Approach 1: sentences instead of words

sentences of length 5

$5^{\# \text{words}}$

Approach 2: bidirectional RNN



no long range dependency

Wishlist:

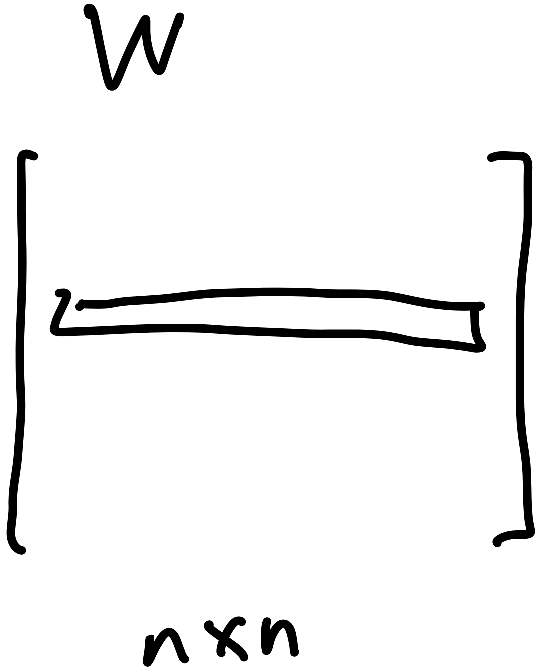
- long range dependency
- variable length
- association between output and input

input: $x_1, \dots, x_n \in \mathbb{R}^d$

output: $y_1, \dots, y_n \in \mathbb{R}^d$

$$y_i = \sum_{j=1}^{n} \underset{\uparrow \text{ row normalized}}{w_{ij}} \overset{\text{value}}{x_j}$$

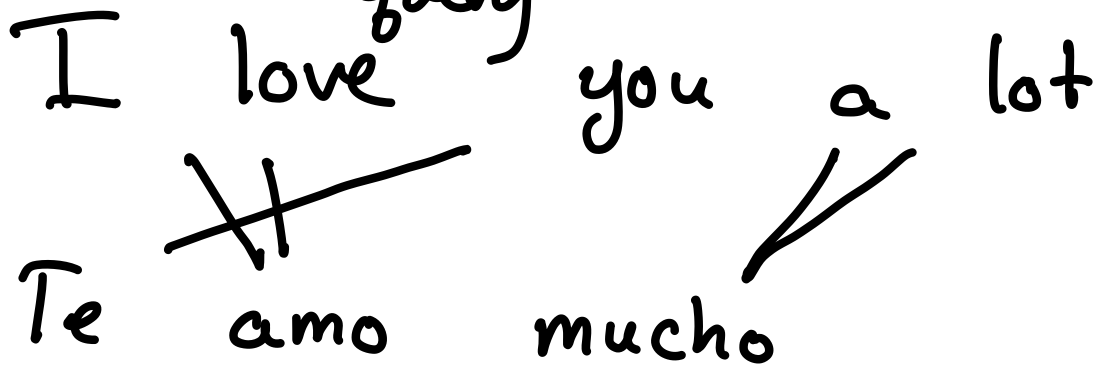
$$\sum_{j=1}^{n} w_{ij} = 1$$



$$w_{ij} = \frac{\overset{\text{softmax}}{\exp}(w_{ij})}{\sum_{j'} \exp(w_{ij'})}$$

w_{ij} = similarity between y_i, x_j

$$= \underset{\text{query}}{x_i^T} \underset{\text{key}}{x_j}$$



$$\text{amo} = \underset{1/4}{\text{I}} + \underset{3/4}{\text{love}} + \underset{0}{\text{you}} + \underset{0}{\text{a}} + \underset{0}{\text{lot}}$$

- set to set ☺
- see all input
  ↳ no distance
- parameters

x_i: appeared as

- part of output
- weight for own output
- weight for other output

$q_i = W_q x_i$ # query

$k_i = W_k x_i$ # key

$v_i = W_v x_i$ # value

$W_{ij} = q_i^T k_j \quad w_{ij} = \text{softmax}(w_{ij})$

$$y_i = \sum_{j=1}^{n} W_{ij} v_j$$

	I	love	you	a	lot
I	3/4	1/4			
love	1/2	1/2			
you			1		
a				3/4	1/4
lot				0	1

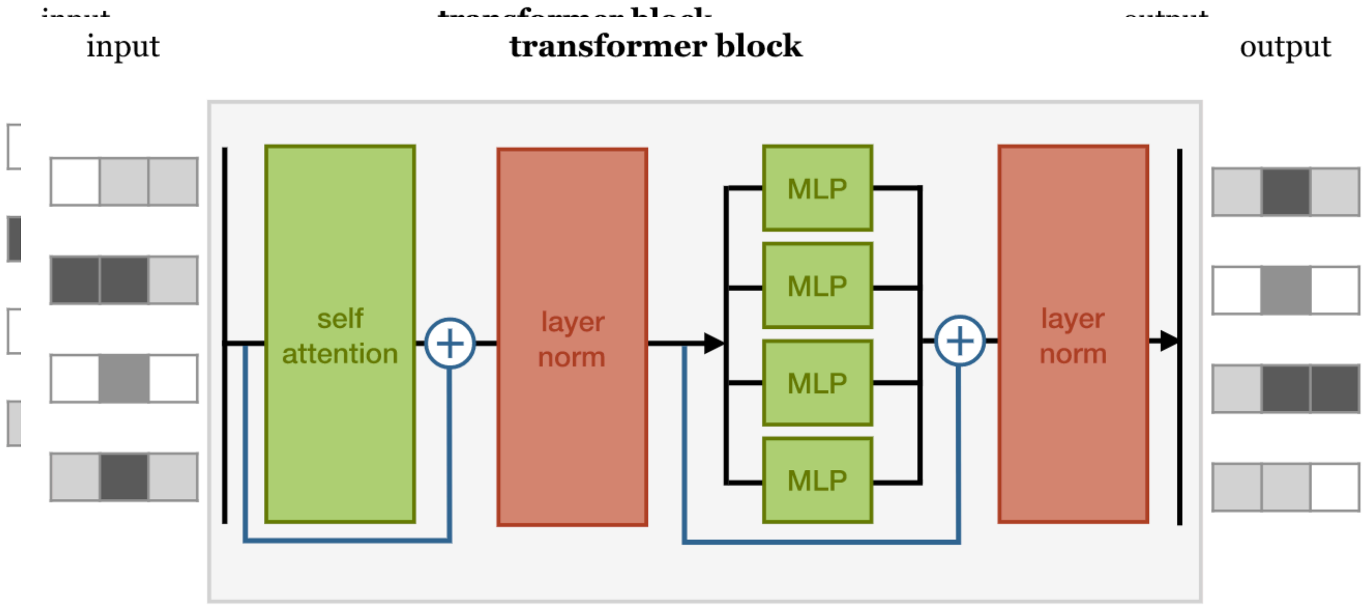
W_{ij} where i → love, j → you

$= q_i^T k_j$

W_{ji}

$= q_j^T k_i$

Transformer



Positional Encoding

t ↙

① Jack ② gave ③ water ④ to ⑤ Jill

Jill gave water to Jack

x_i

$\begin{bmatrix} \\ \\ \\ \end{bmatrix}$ ← embedding "Jack"

← encodes position

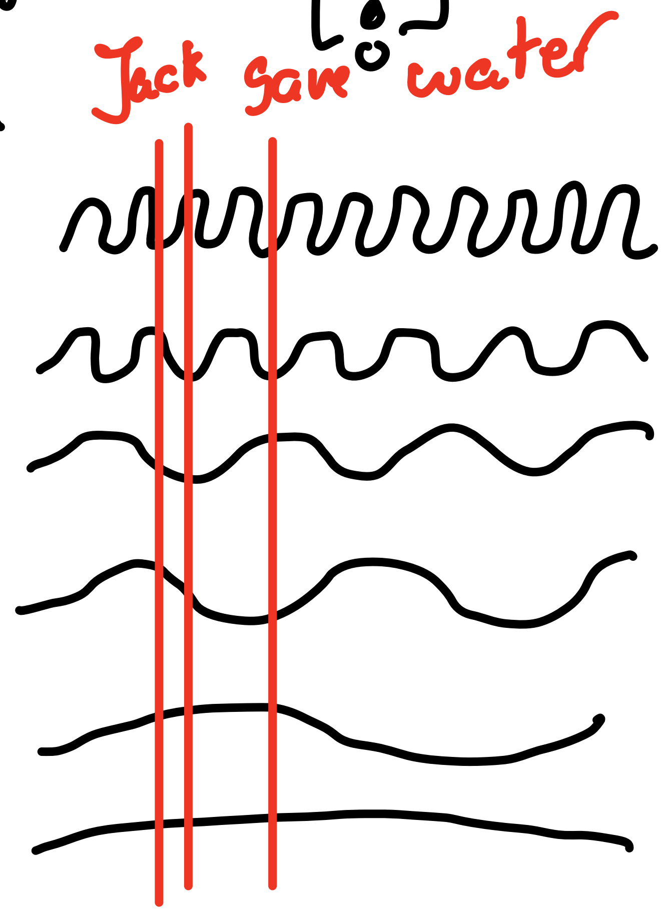
↳ put in index

↳ one hot encoding

Jack $\begin{bmatrix}1\\0\\0\\0\end{bmatrix}$ gave $\begin{bmatrix}0\\1\\0\\0\end{bmatrix}$ water $\begin{bmatrix}0\\0\\1\\0\\0\end{bmatrix}$

↙omega

$$\begin{bmatrix}\sin(w_1 t)\\ \sin(w_2 t)\\ \sin(w_3 t)\\ \vdots\end{bmatrix}$$


Jack
gave
water