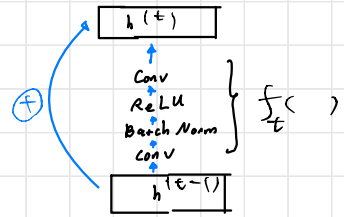
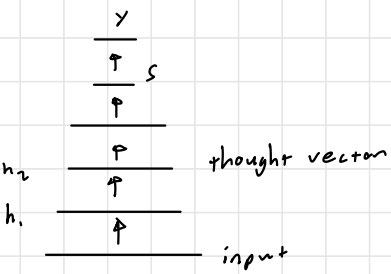
Lecture 18

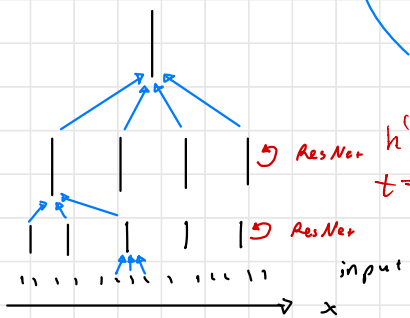
- Neural Network : Team of Vectors
  (agent or person)
  Formation, message passing



- MLP



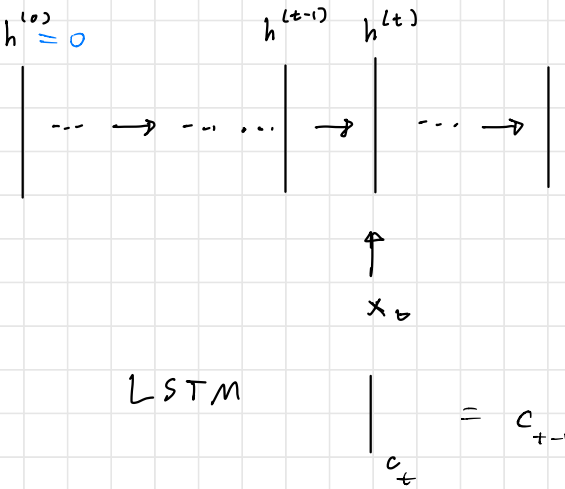
- CNN



$h^{(t)} = h^{(t-1)} + f_t(h^{(t-1)})$

$t = 1, \dots, 6$

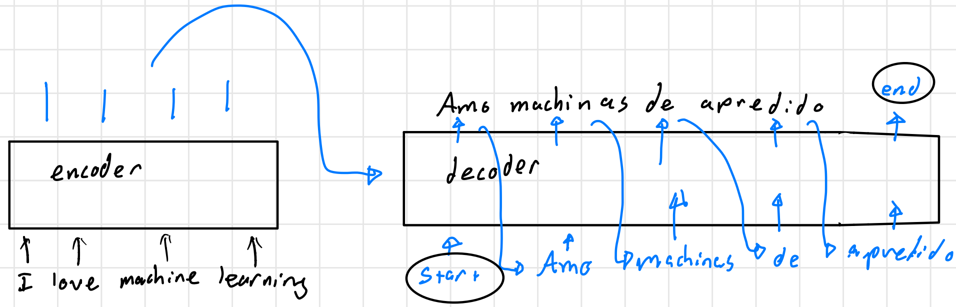
- RNN



LSTM

$= c_{t-1} + \Delta c_t$

- Transformer :

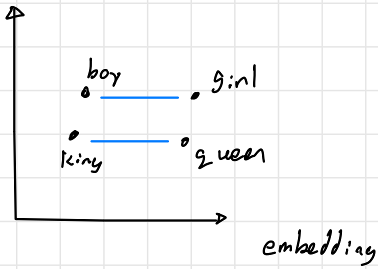


- Word 2 vec :

vector

↑ encoding / embedding

word

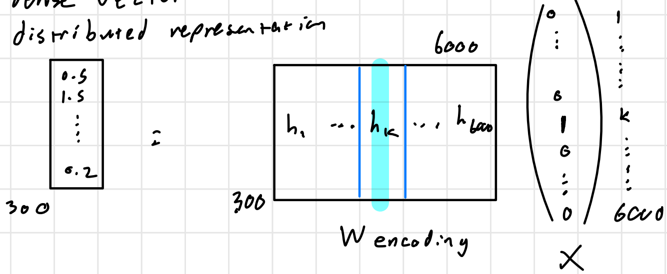


- Dictionary : 6000 words (50k tokens)

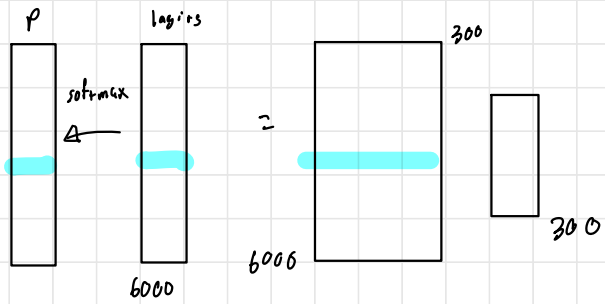
a word = one hot vector

$$k^{th} \text{ word} : \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \begin{matrix} 1 \\ 2 \\ \vdots \\ k \\ \vdots \\ c = 6000 \end{matrix}$$

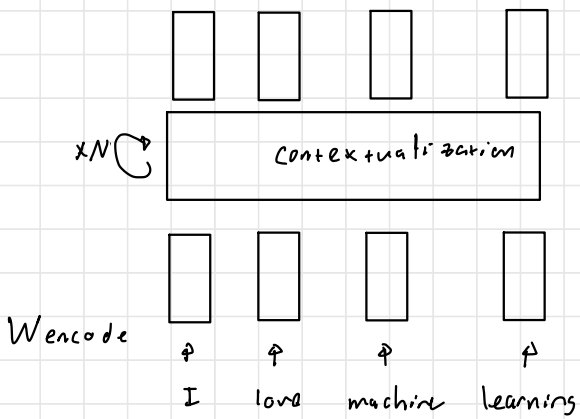
dense vector
distributed representation



- Decoding:
- vector ⟶ word

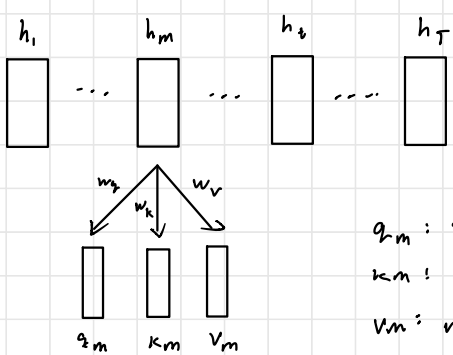


- Back to Encoder:



- One Round:

Party / Meeting



q_m: what you want to know? query
k_m: can you answer? key
v_m: what is your content? value

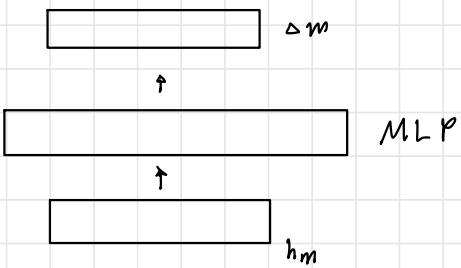
- Self - Attention : (During Party)

$$A_{m \to t} = \frac{\exp \langle q_m, k_t \rangle / \sqrt{d}}{\sum_{t=1}^{T} \exp(\quad)}, \quad d = \dim(q \text{ or } k)$$

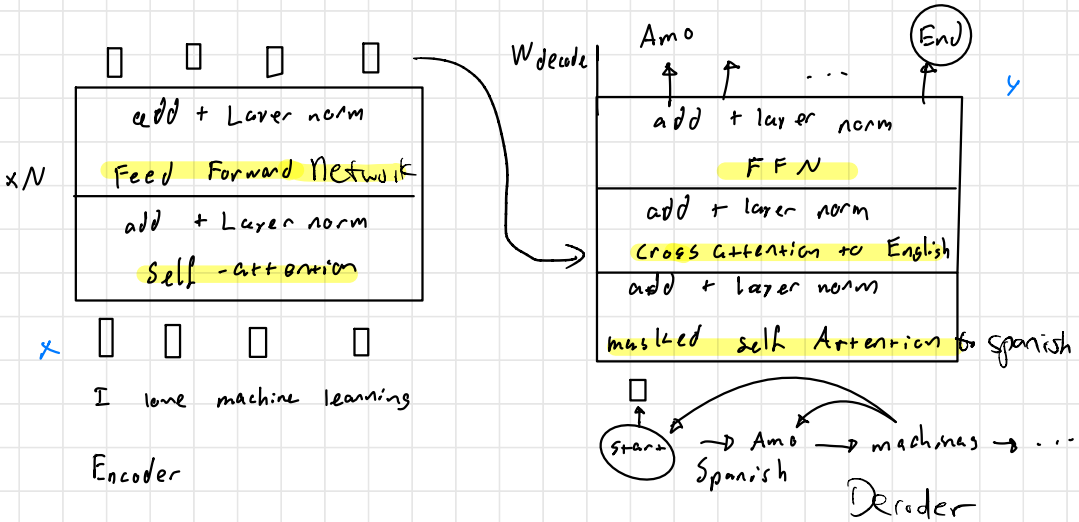
$$\Delta m = \sum_{t=1} A_{m \to t} V_t$$

$$h_m \leftarrow h_m + W \cdot \Delta m$$

- After the Party :



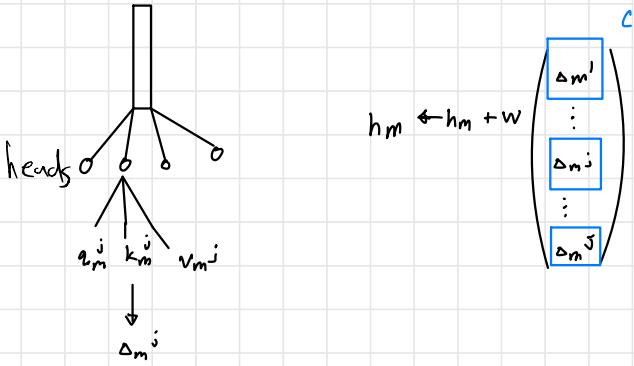
$$h_m = h_m + W \cdot \Delta m$$



- loss :

$$-\log p(y|x) = -\sum_t \log p(y_t \mid y_{<t}, x)$$

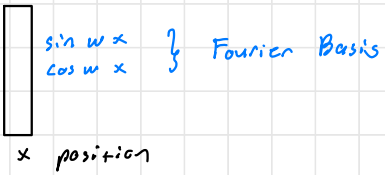
- Note : Multihead Self-Attention



- Position Embedding :

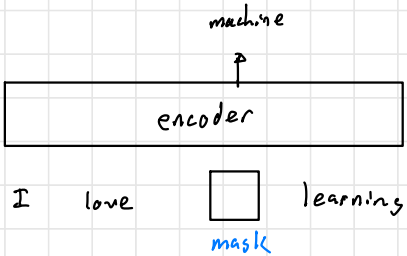
	I	love	machine	learning
$x =$	1	2	3	4

First Layer : $h_m \leftarrow h_m + P_m$



- Self - Supervised Learning :

  - Masked language modeling



  - BERT

- Generative pre-training: (GPT)

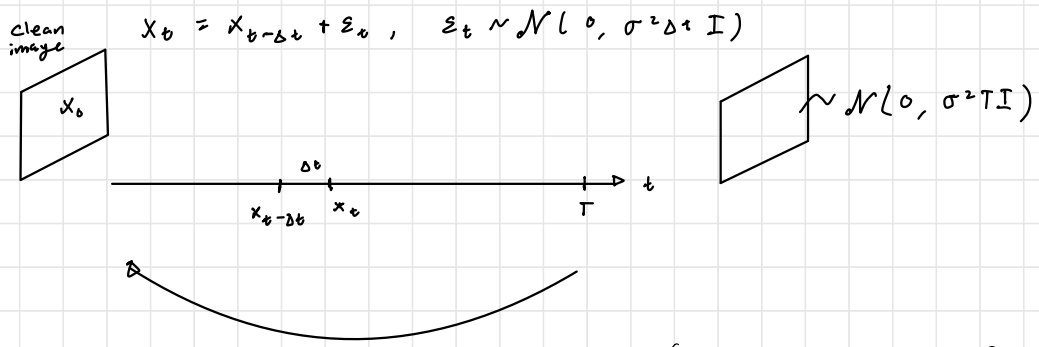
I love machine learning

$x_{<t}$ x_t

loss: $\log p(x_t \mid x_{<t})$

decoder

- Generative AI
- Diffusion Model:



$$p(x_{t-\Delta t} \mid x_t) \sim N\left(\mu_\theta(x_t, t), \sigma^2 \Delta t\, I\right)$$

denoising

$$x_t + \sigma^2 \Delta t\, \nabla_x \log p_t(x_t)$$