Review

1. No sequentiality
2. Introducing Sequentiality.
3. Introducing Attention

Part 1

Lec 1 → Lec 3

$X \in \mathbb{R}^D$ →

- binary classificati
- Multiclass classificat
- Regression

PS3

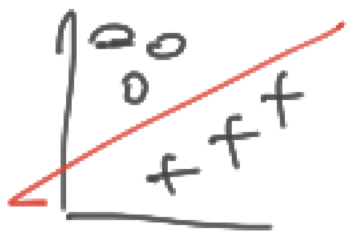
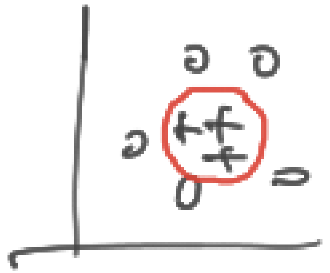
→ Preprocessing
- one hot encoding x
- scaling

→ Feature engineering

→ Feature Importance

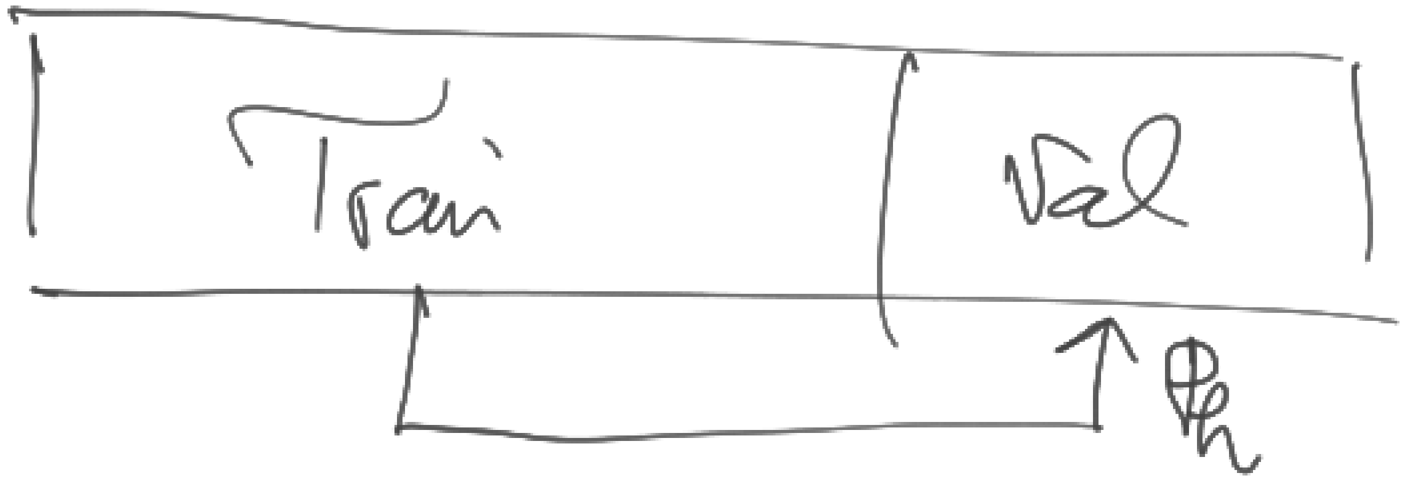
→ prediction

~> Models

 ⟹ 

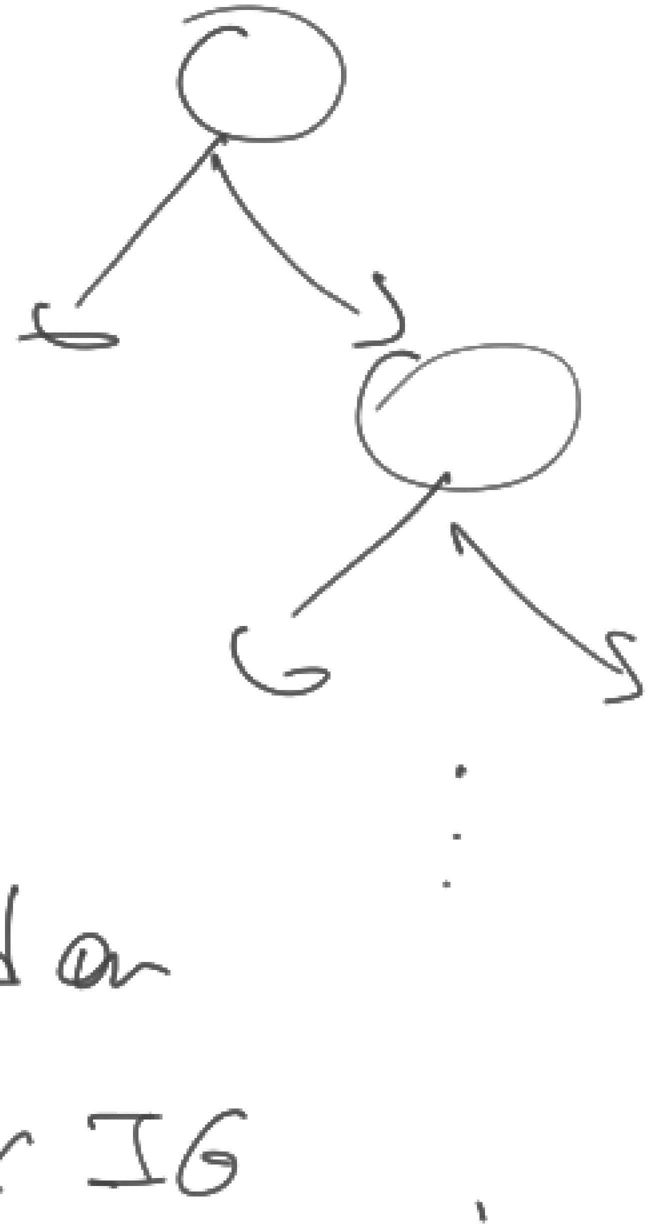
Tree based Models	NN (DL)
RF	
→ # trees	→ # Layers
→ Depth of trees	→ # neurons / Layer
→ # feature / split	→ Act. function
→ Impurity measure	→ Train → lr → batch size → # epochs
HP	HP

HP optimization for RF (PS3)



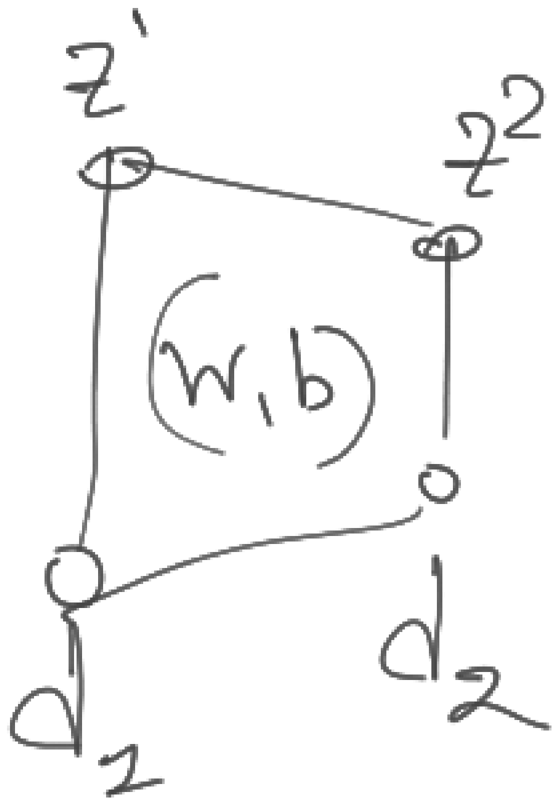
- h = (100 trees, Entropy, 200 depth)
- evaluation metrics (L1)

Trees (L2)



based on max IG

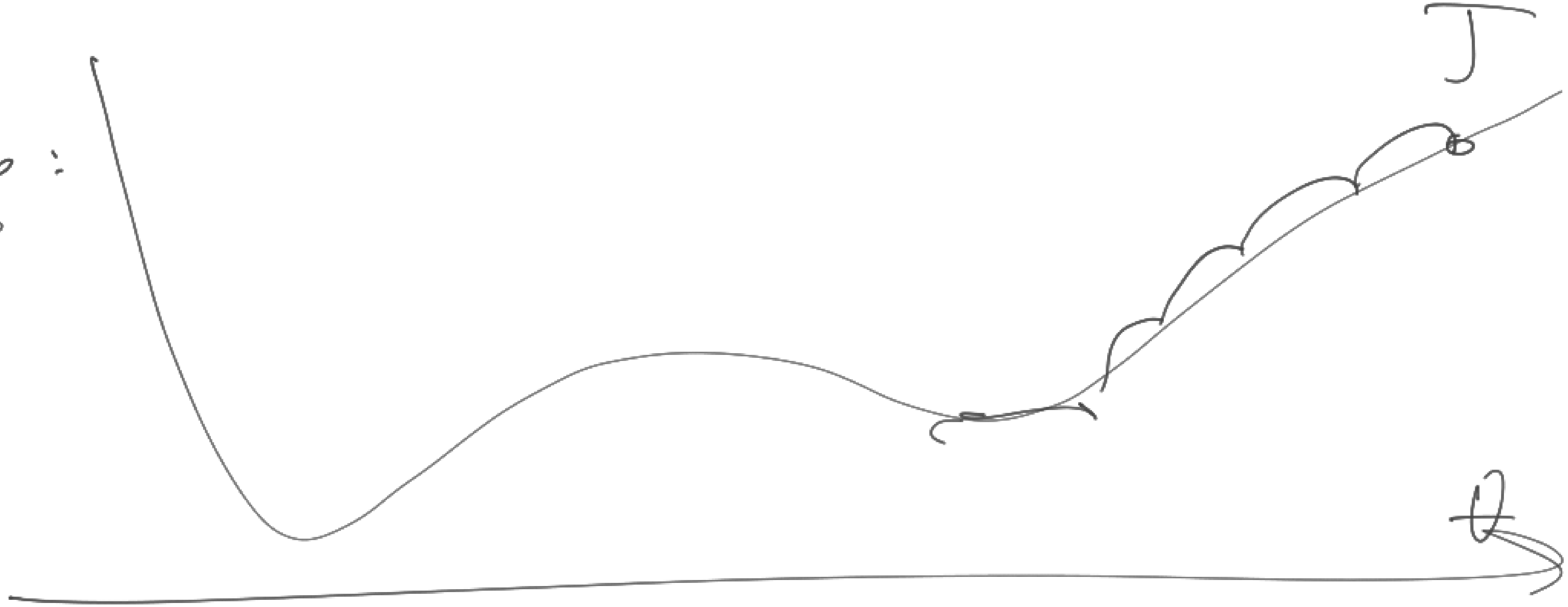
NN



$$z^2 = (W^T z^1 + b)$$

$W \in \mathbb{R}^{d_1 \times d_2}$, $b \in \mathbb{R}^{d_2}$

Train:

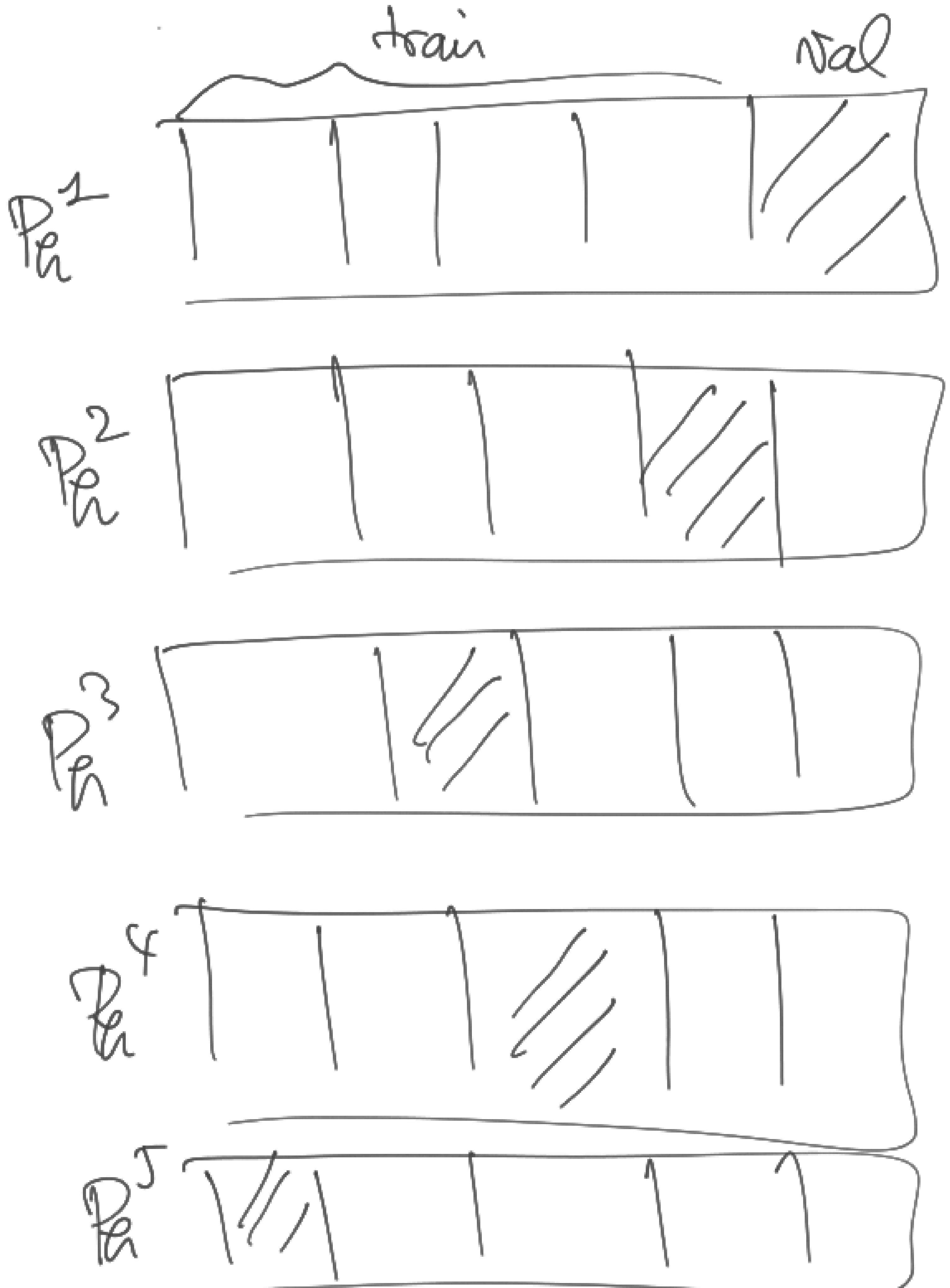


Cross validation

(small datasets)

$h = (100 \text{ trees, Entropy, } 200 \text{ depth})$

$$P_h = \underset{1 \le k \le 5}{\text{mean}} P_h^k$$



Part 2 : Introducing Sequentiality :

PSL Introducing sequential data : Imdb dataset.

Dataset

doc_1 :	w_1^1 ——— $w_1^{T_1}$	y_1
$\vdots$		
doc_N	w_N^1 - - - $w_N^{T_N}$	y_N

$\Downarrow$

2 limitations

→ non sequentiality

→ poor representation of words (mere indices)

Lecture 5

Dataset

doc_1 ————————

⋮

doc_N ————————

$\rightarrow$ word_index = { 'a' : 1

⋮

'zoo' : V }

objective :

$$\mathcal{E} = \begin{bmatrix} \text{——} & w_1 & \text{——} \\ & \vdots & \\ \text{——} & w_i & \text{——} \\ & \vdots & \\ \text{——} & w_V & \text{——} \end{bmatrix}_{V \times D}$$

: w_i : D dim $\vec{v}$ representing word of index i

GloVe:

Dataset $\Longrightarrow$ X: co-occurence matrix $\Longrightarrow$ $\overbrace{\log X_{ij}}^{\text{"target"}} \sim \overbrace{w_i^T \, \tilde{w}_j}^{\text{prediction}}$

X_{ij}: # times word of index j is in the context of word of index i

Loss: $$J_\theta = \sum_i \sum_j \left(\log X_{ij} - w_i^T \tilde{w}_j\right)^2$$

$$\theta = \begin{bmatrix} w_1 \\ \vdots \\ w_V \\ \tilde{w}_1 \\ \vdots \\ \tilde{w}_V \end{bmatrix}$$

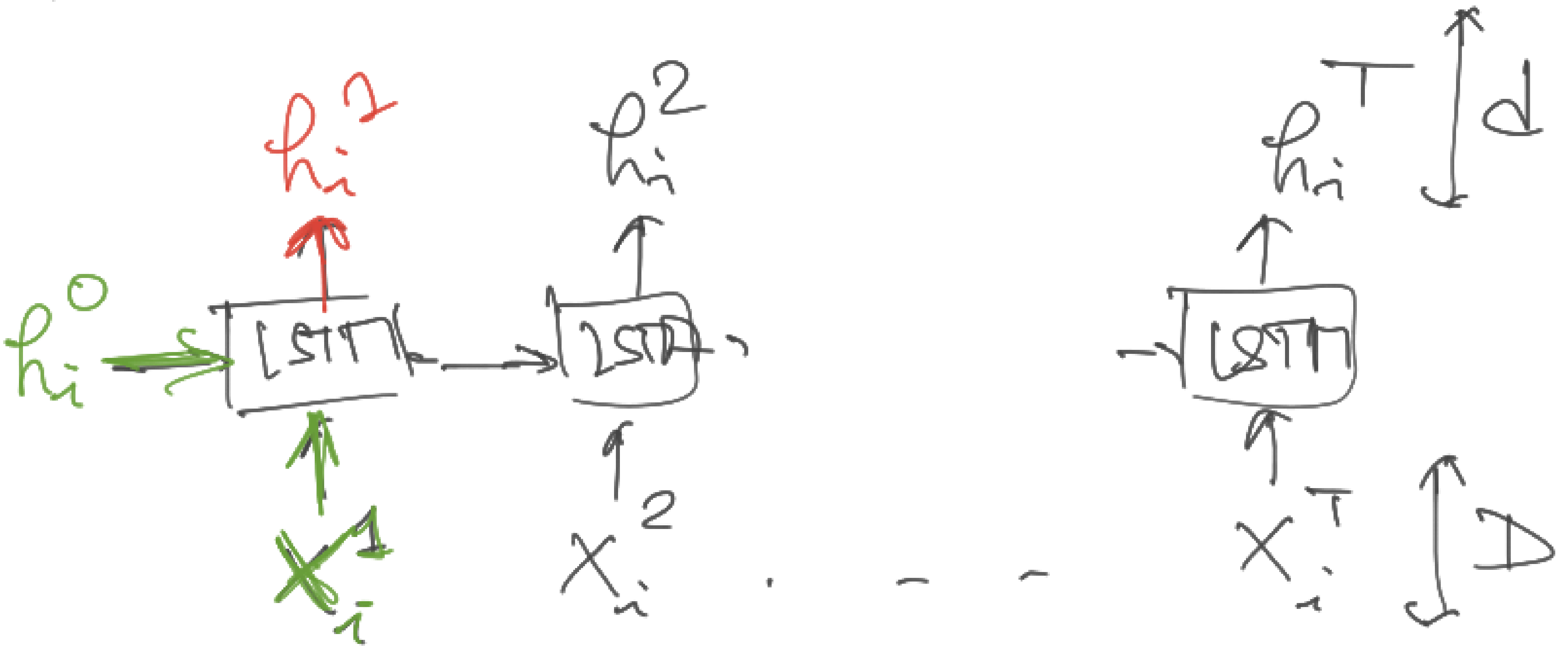
Embedding Layer

Parameters: embedding matrix: $E = \begin{bmatrix} — w_1 — \\ \vdots \\ — w_V — \end{bmatrix}_{V \times D}$

$$\text{doc}_i \; \underbrace{[\,17,\ 78,\ \ldots,\ 47\,]}_{T} \longrightarrow \boxed{\text{Embedd. Layer}} \longrightarrow \underbrace{\begin{bmatrix} — w_{17} — \\ — w_{28} — \\ \vdots \\ — w_{47} — \end{bmatrix}}_{D} T$$

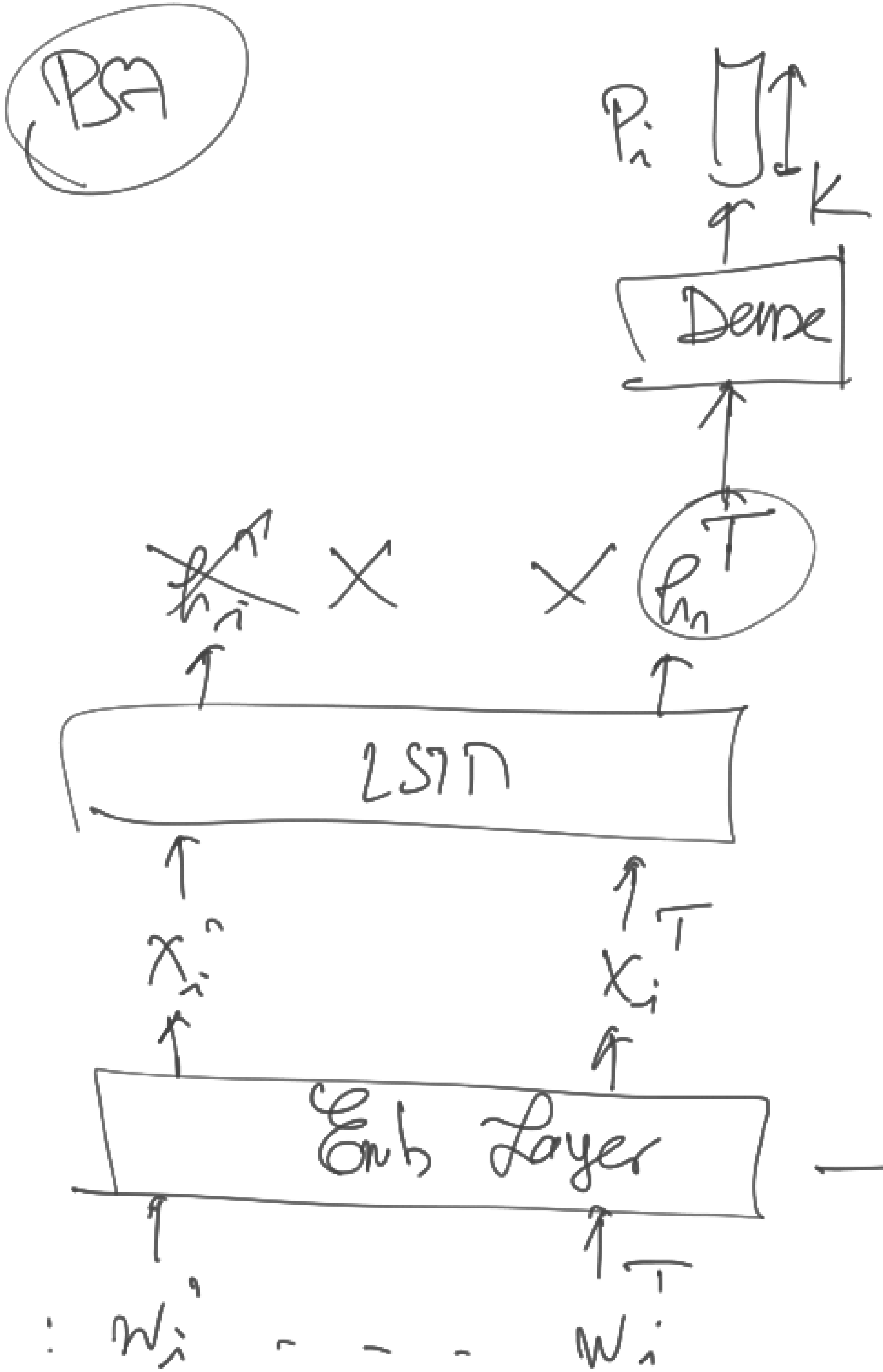
$$(N, T) \longrightarrow (N, T, D)$$

Lecture 6 Introducing Seq Models : LSTM.



$(N, T, D) \longrightarrow (N, T, d)$

if I only keep h_i^T $(N, T, D) \longrightarrow (N, d)$

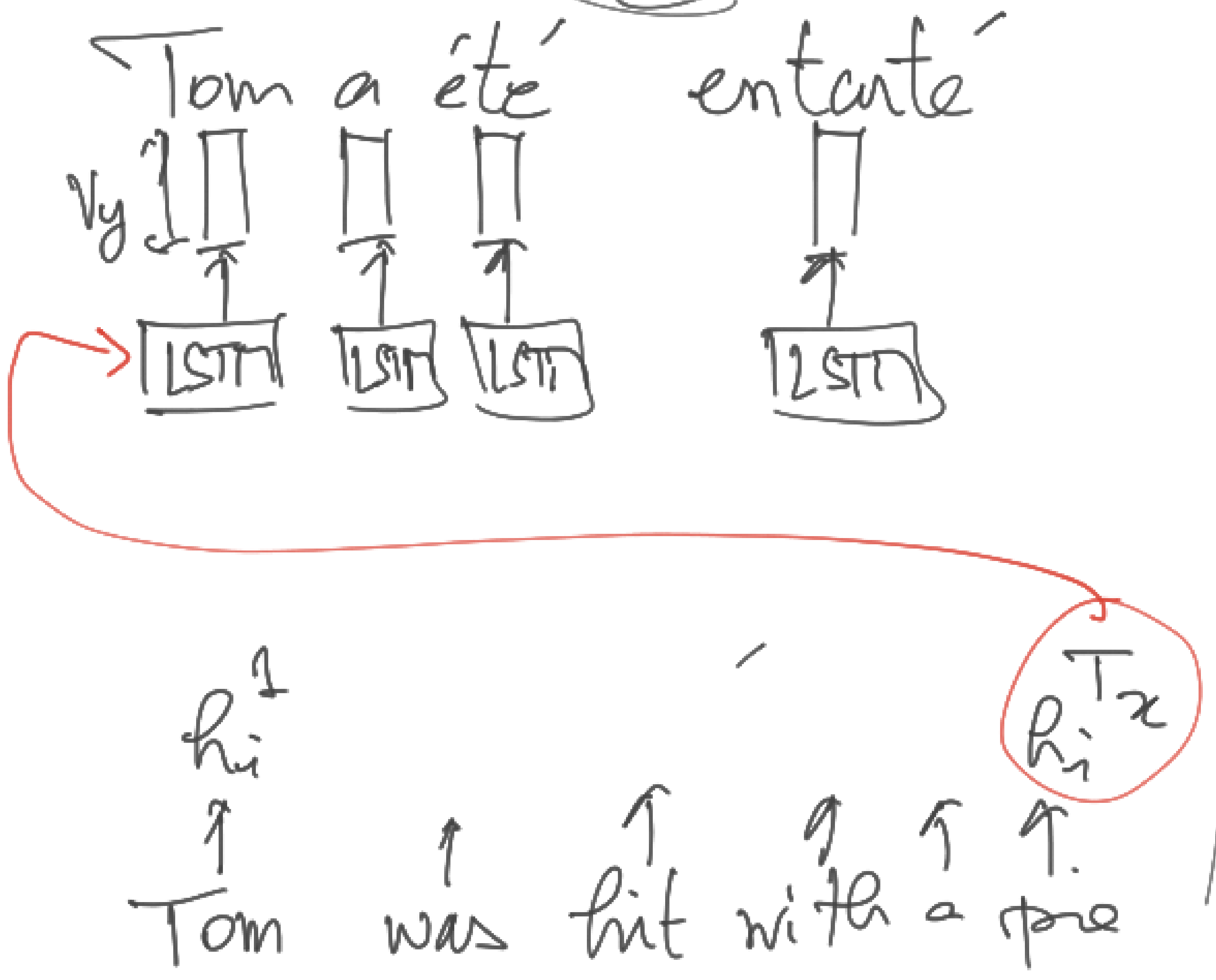


→ ~2 million param

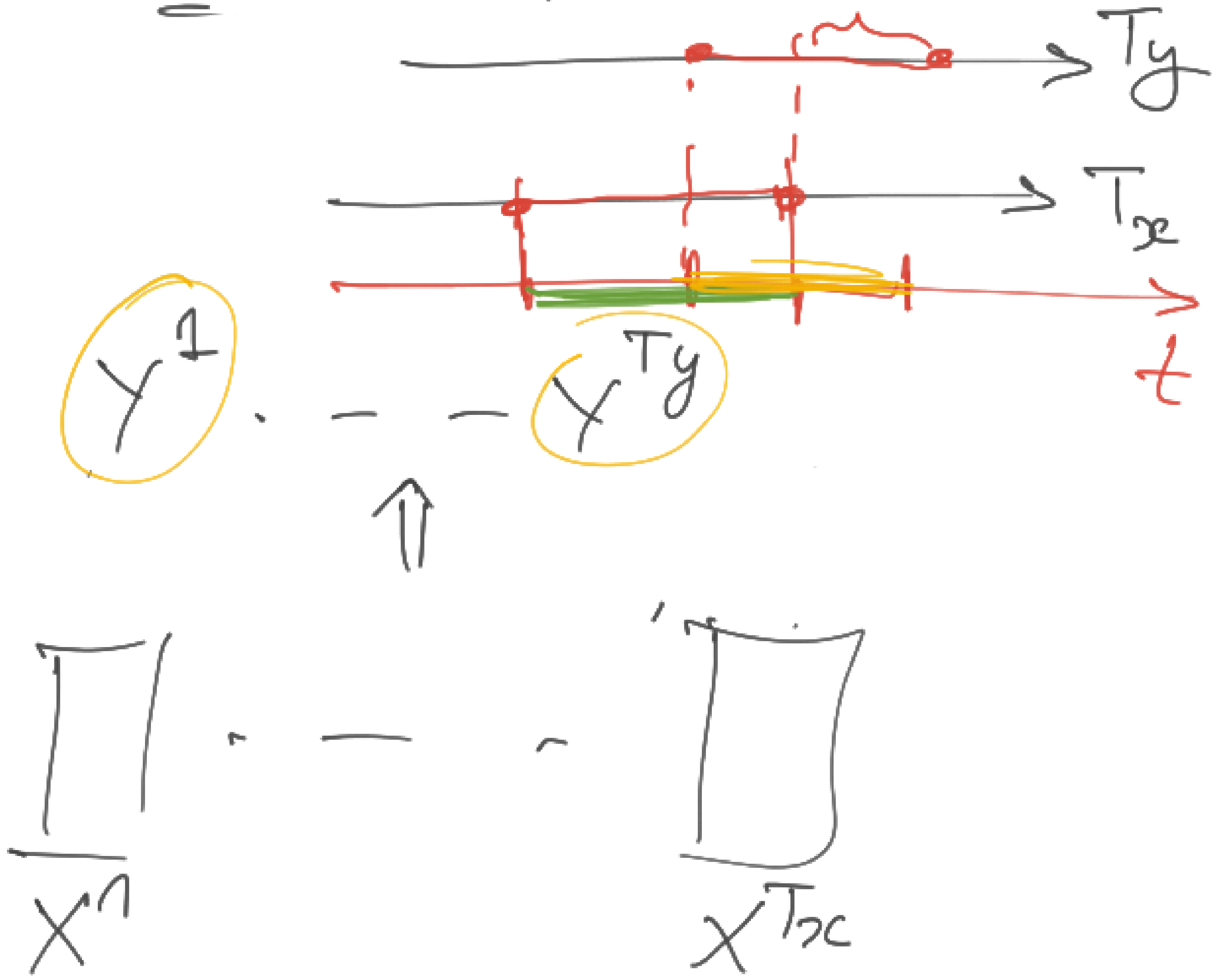
using pretrained word vectors reduced drastically # parameters to train

28 From may to one (PSA) → Many to many.

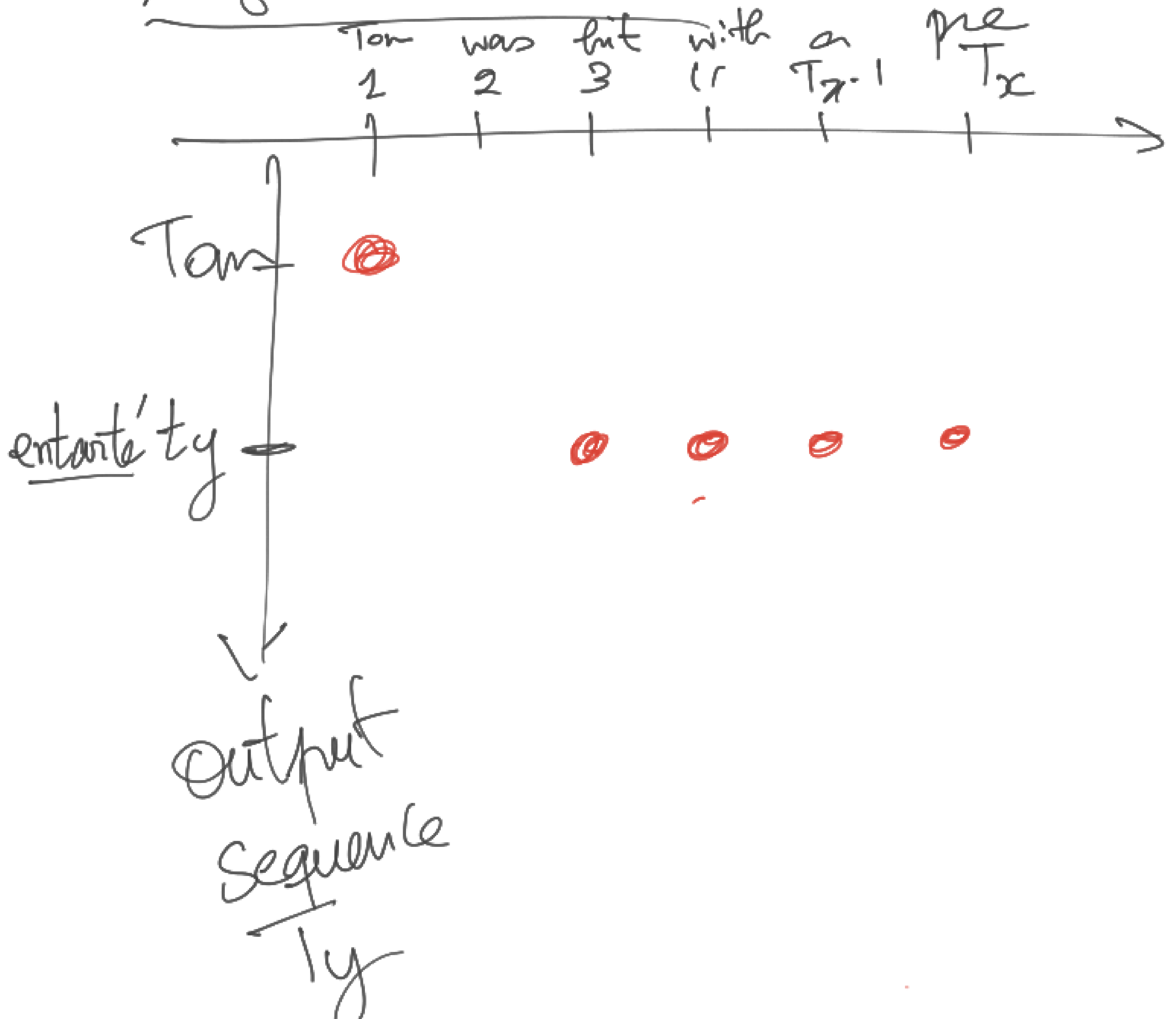
NLP : NMT

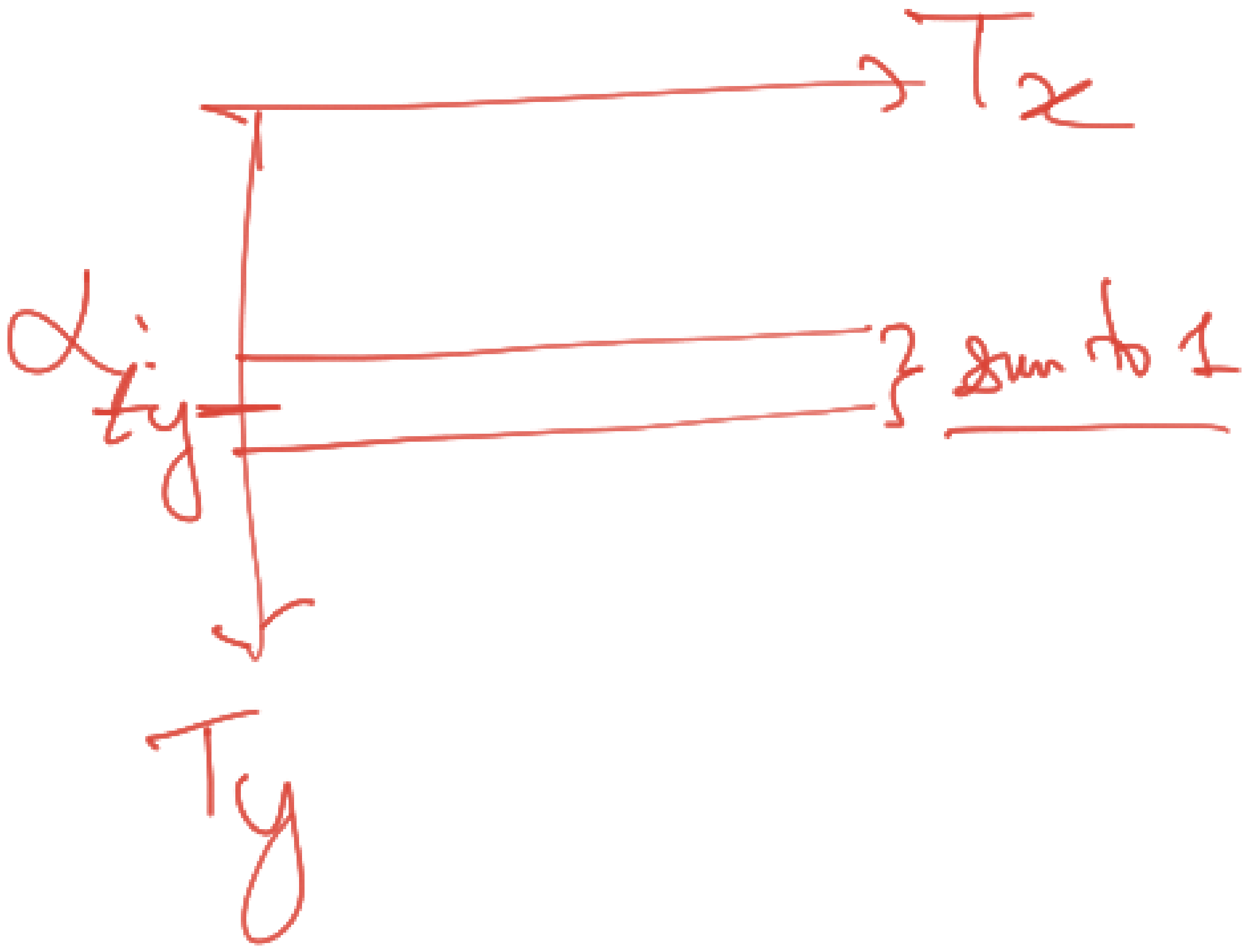


Finance



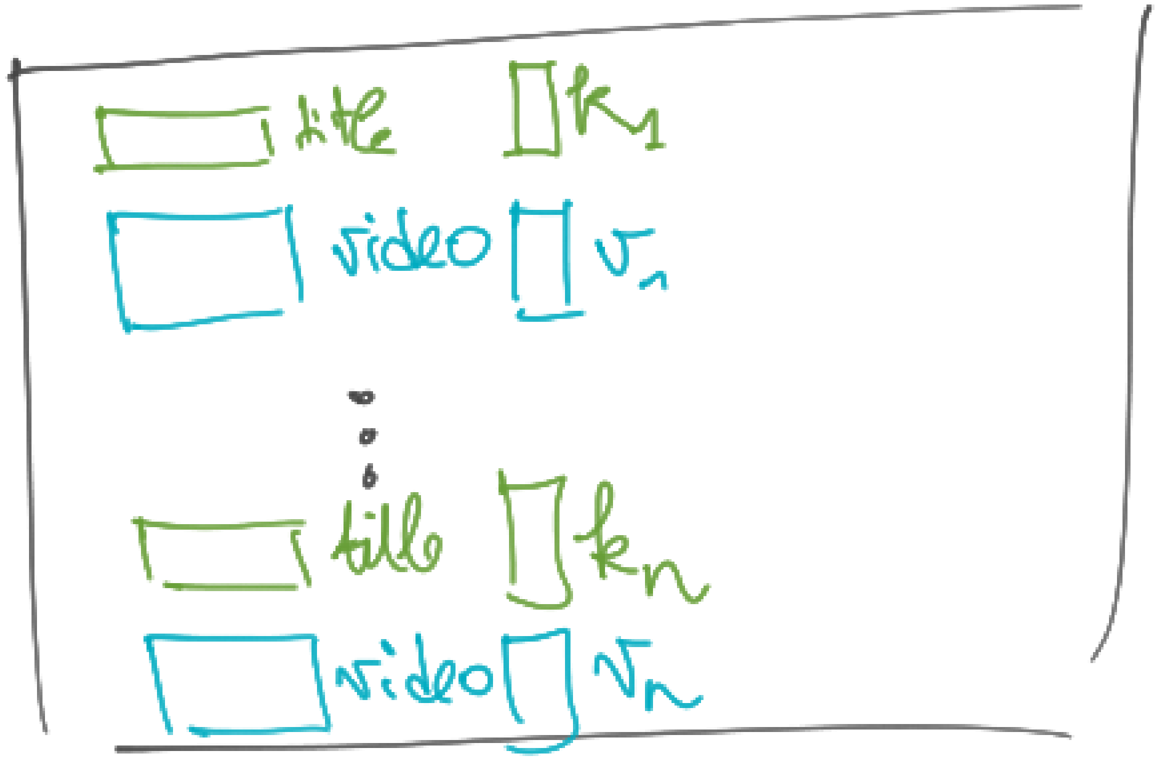
Alignment matrix:



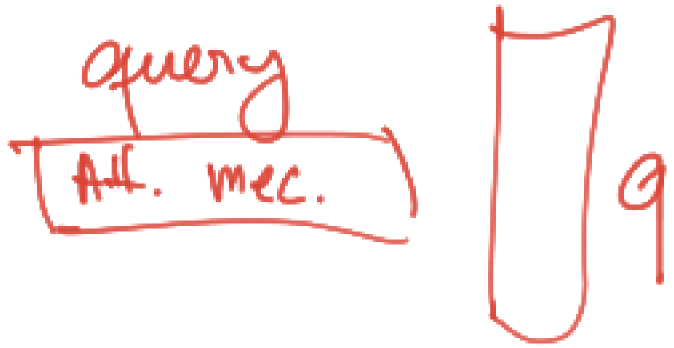


Part 3 : Introducing the Attention Mechanism :

Intuition :



Database

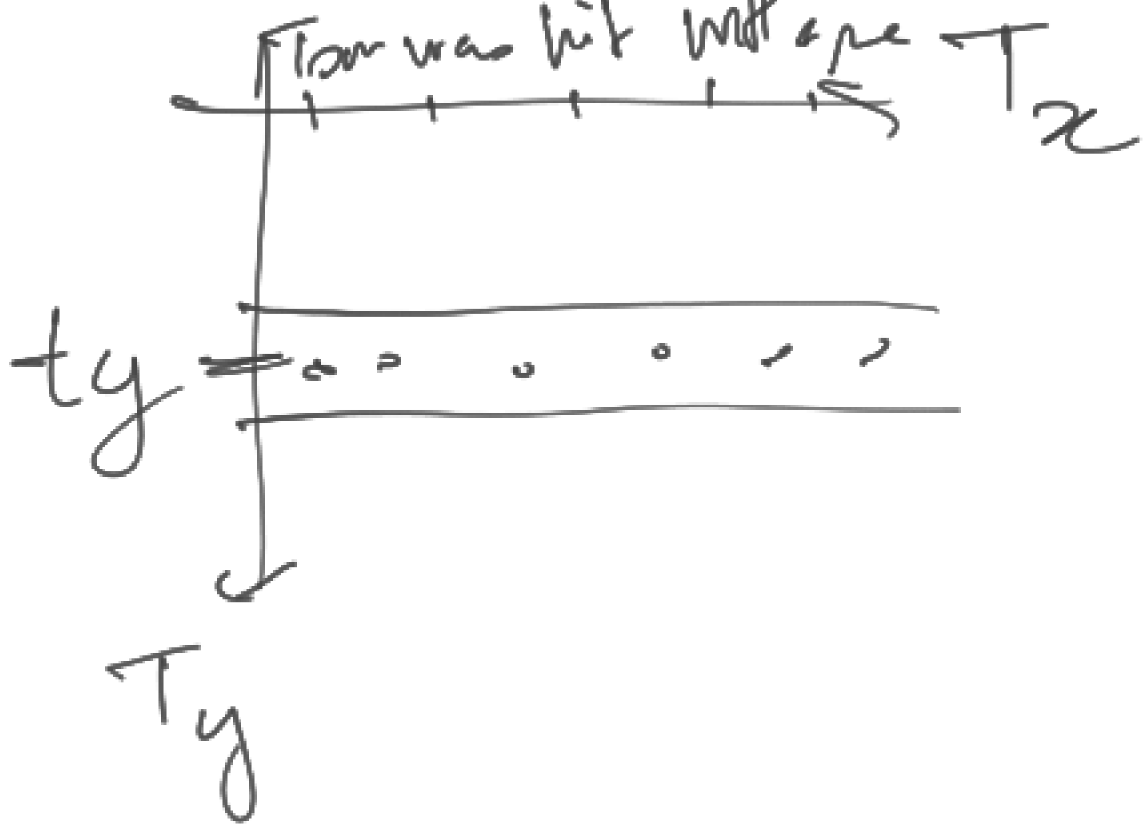


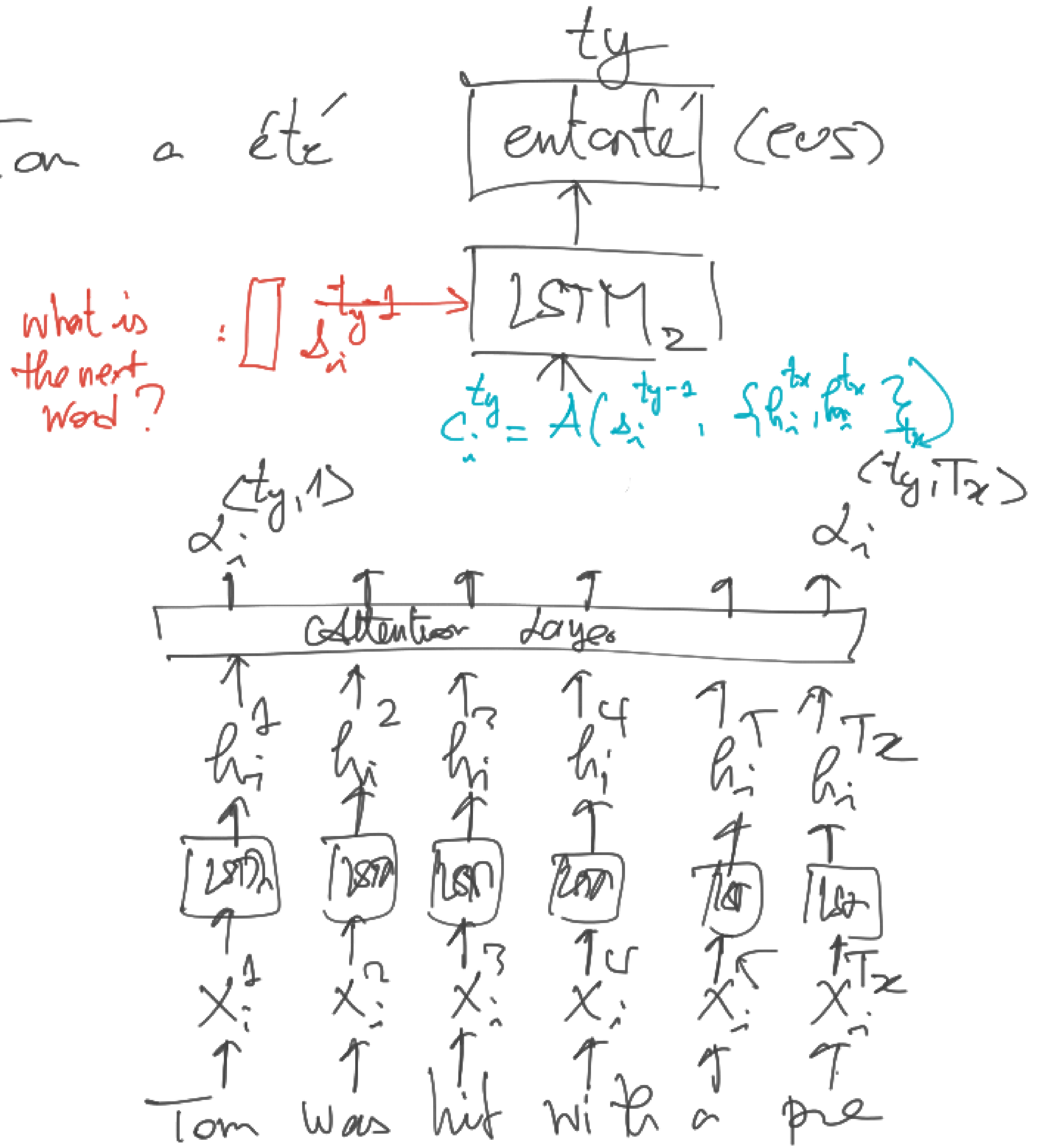
$$A(q, \{k_i, v_i\}_{1\leq i\leq n}) = \sum_{i=1}^{n} \alpha_i \, v_i$$

how relevant v_i is to q

α_i ? $\quad \left(\mathrm{sim}(q, k_i)\right)_{1\leq i\leq n} \xrightarrow{\text{Softmax}} \left(\alpha_i\right)_{1\leq i\leq n}$

1st application:





↳ The input to the Decoder LSTM in the sequence to sequence with Attention framework is:

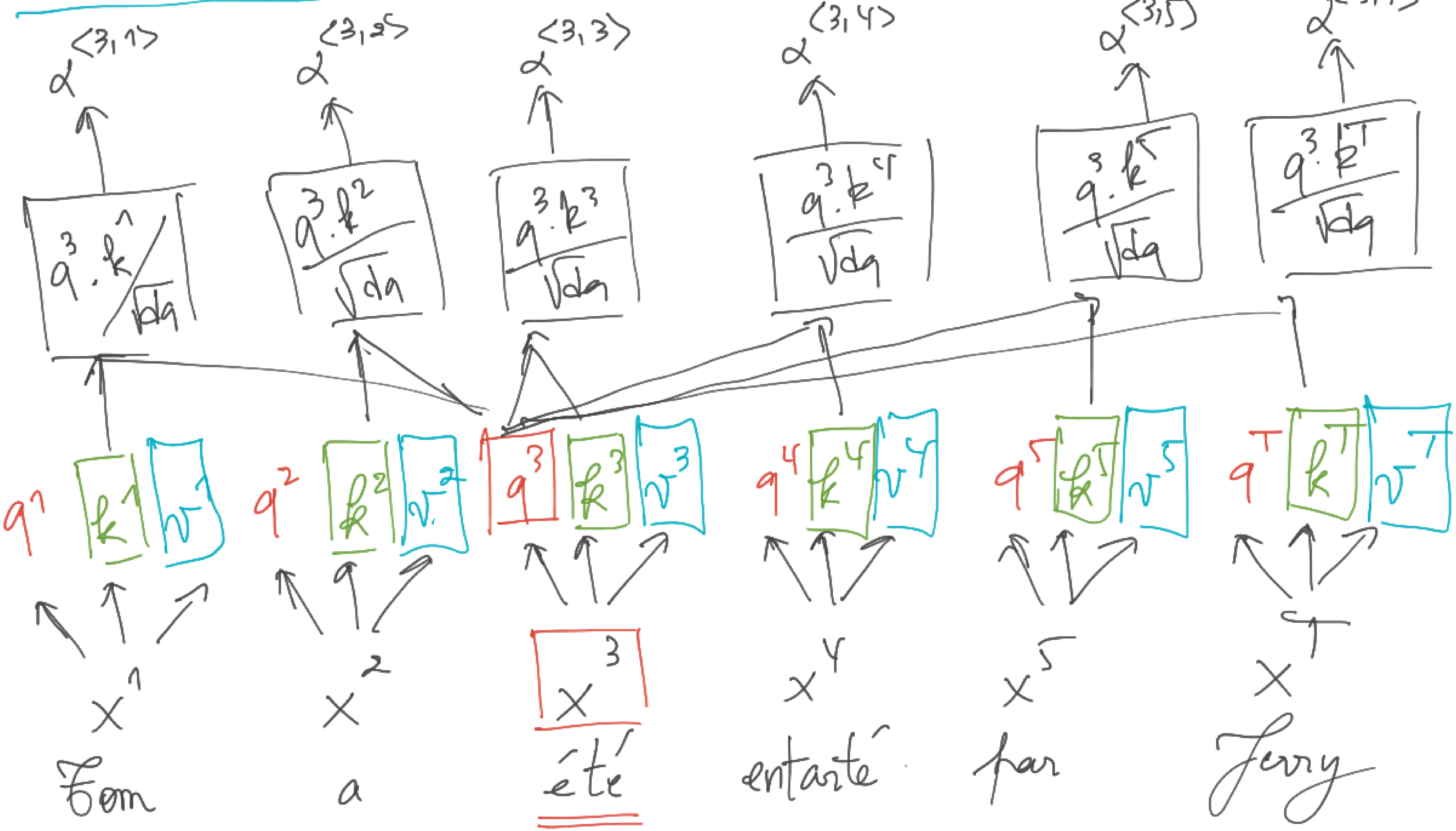
$$c_i^{t_y} = \sum_{t_x=1}^{T_x} \underbrace{\alpha_i^{\langle t_y, t_x \rangle}}_{=\,\frac{\exp\left(s_i^{t_y-1} \cdot h_i^{t_x}\right)}{\sum_{t_x'=1}^{T_x} \exp\left(s_i^{t_y-1} \cdot h_i^{t_x'}\right)}} h_i^{t_x}$$

2^{nd} Application : Self Attention layer.

- Introduced to handle the polysemy problem
  - → Tom a $\boxed{\text{été}}$ entarté par Jerry.
    (i.e, Tom WAS hit with a pie by Jerry)
  - → Cet $\boxed{\text{été}}$ il fera horriblement chaud
    (i.e, This summer will be unbearably hot)
- GloVe approach : Static embedding vectors

↳ Need to create contextual embedding vectors

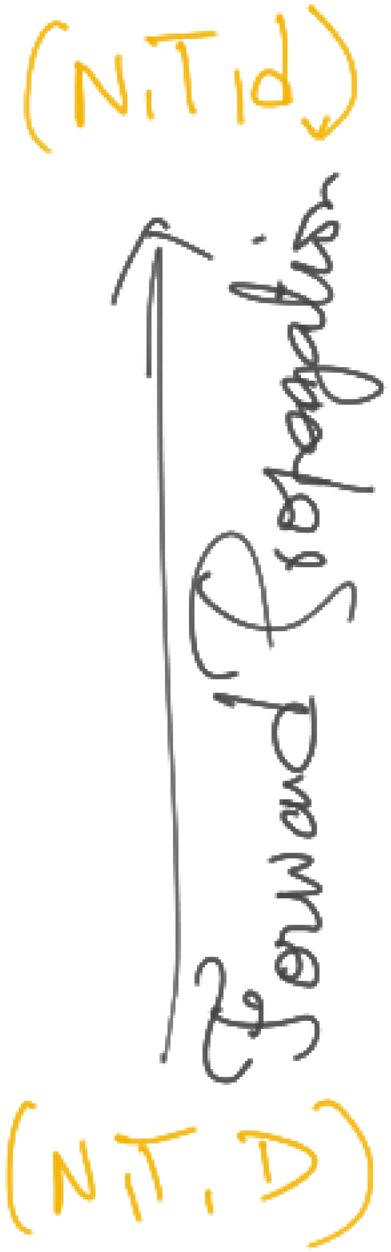
$$\boxed{A\left(q^3, \{k^t, v^t\}_{1 \le t \le T}\right) = \sum_{t'=1}^{T} \alpha^{\langle 3,t' \rangle} v^{t'}}$$



The Self Attention Layer

Parameters $W_Q \in \mathbb{R}^{D \times d_q}$; $W_K \in \mathbb{R}^{D \times d_k}$, $W_V \in \mathbb{R}^{D \times d_v}$

$\boxed{A(q^t, \{k^{t'}, v^{t'}\}_{1 \le t' \le T}) \text{ is the contextual representation of } X^t}$



$$A(q^t, \{k^{t'}, v^{t'}\}_{1 \le t' \le T}) = \sum_{t'=1}^{T} \frac{\exp\left(\frac{q^t \cdot k^{t'}}{\sqrt{d_q}}\right)}{\sum_{t''=1}^{T} \exp\left(\frac{q^t \cdot k^{t''}}{\sqrt{d_q}}\right)} v^{t'}$$

The end.