Lecture 1 → Lecture 3

$x \in \mathbb{R}^D$

- binary Classification
- Multiclass classification
- Regression.

PS 3

→ Preprocessing ⟹ One hot encoding ⊗
Scaling

→ Feature engineering.

→ Model

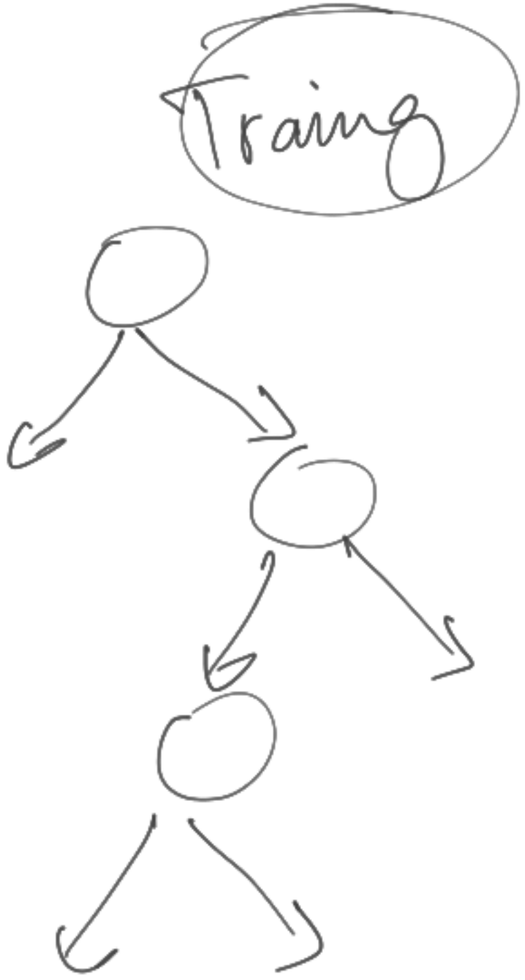
L2
Tree based Models

HP
- depth
- # trees
- Impurity
- etc.

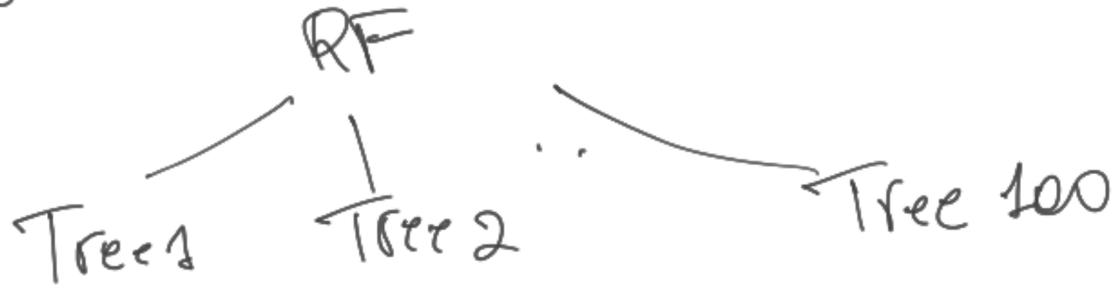
L4
NN

HP
- # neurons / layer
- # layers
- Act fcts ...

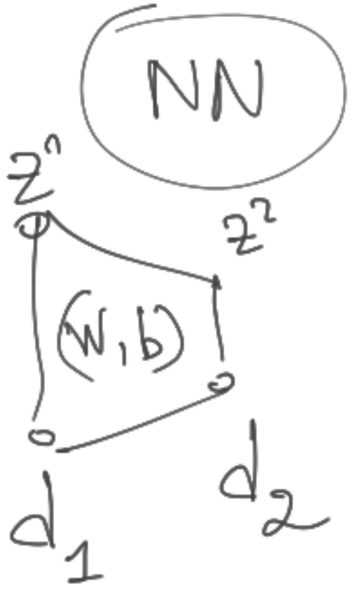
↘ Pick HP :



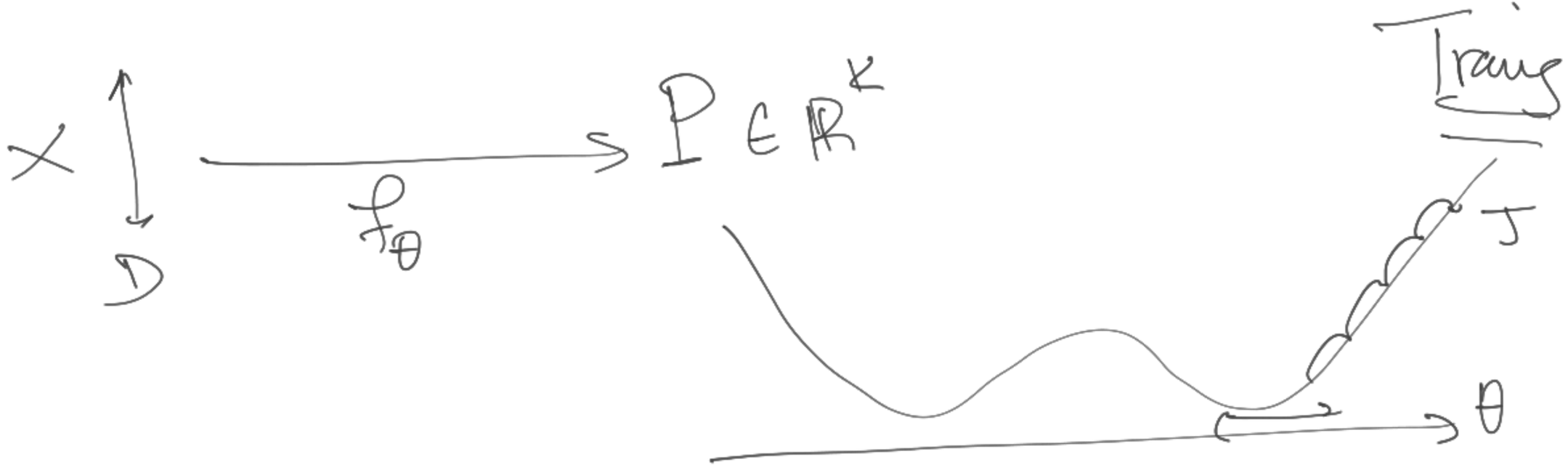
$h = (100 \text{ trees}, \text{Entropy}, 200)$



④ evaluation metrics (21)

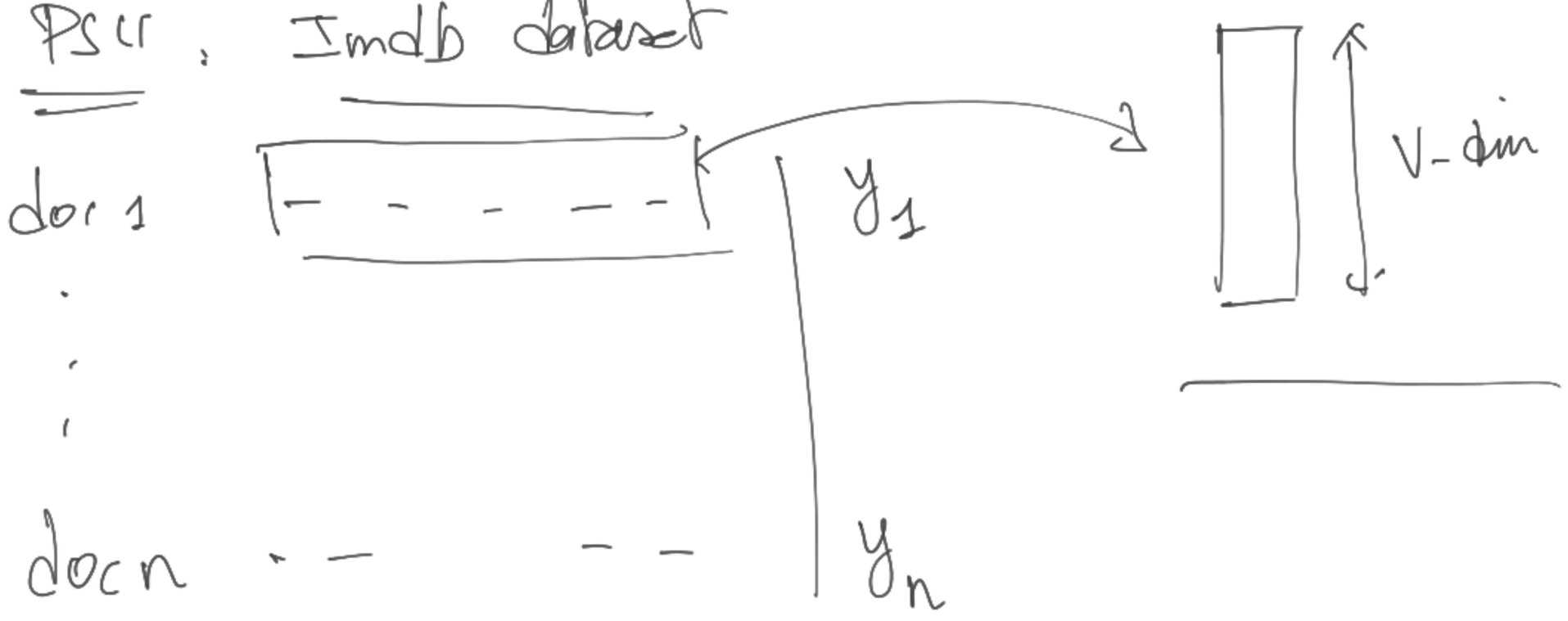


$$z^2 = \sigma(W^T z^1 + b)$$



Part 2 Introducing Sequential by

PSC : Imdb dataset



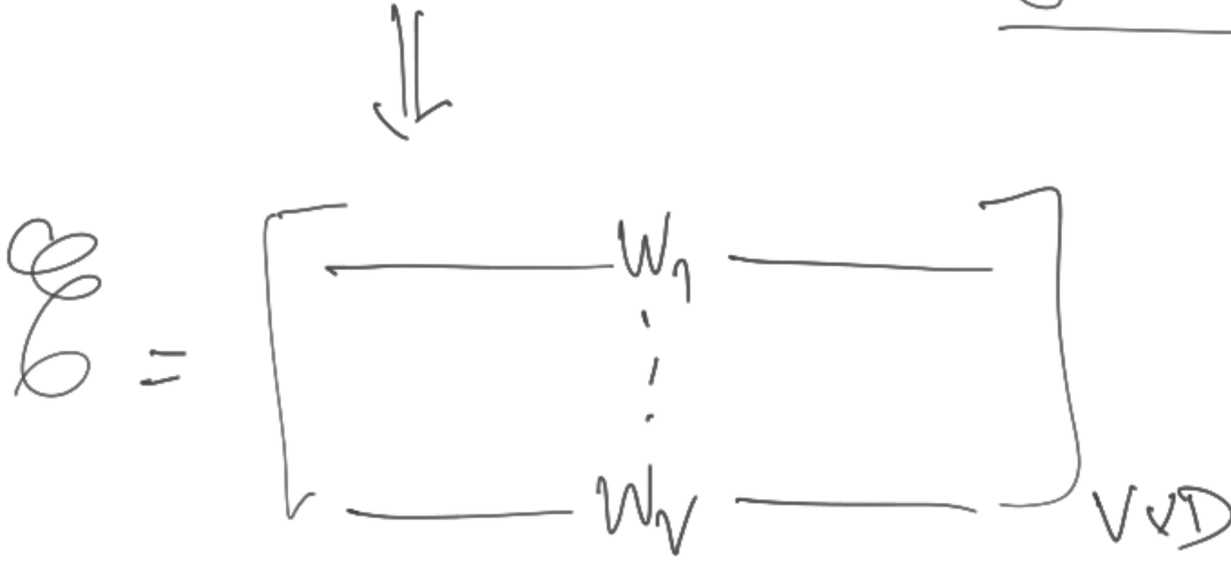
Limitation in RNN

→ Words represented as indices (L1)

→ Non sequential model. (L2)

Lecture 5 : Limitation 1

words ⟶ indices ⟶ D-dimensional representation of words



GloVe recap

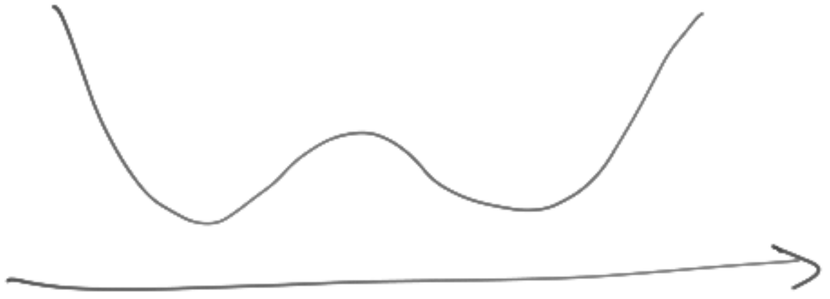
Dataset

doc$_1$
.
.
.
doc$_n$

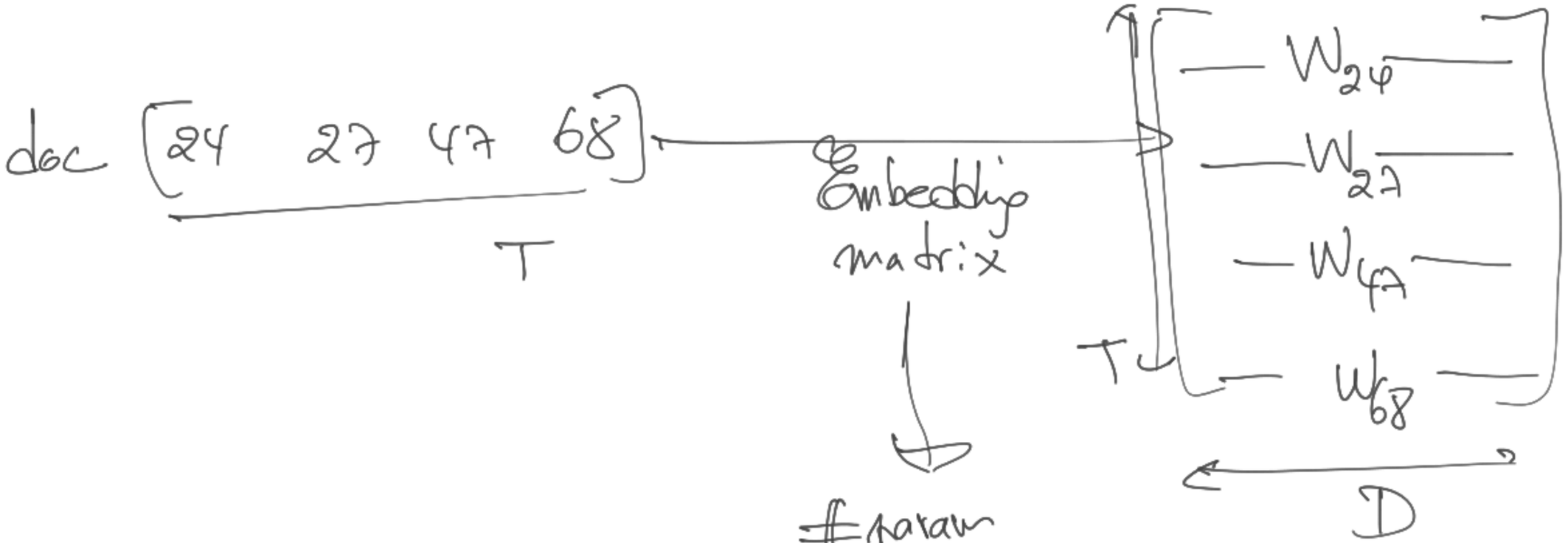
$\rightarrow$ X co-occurence matrix

$\rightarrow$ $\log X_{ij} \sim w_i^T \tilde{w}_j$

$$J = \sum_{i,j} \left(\log X_{ij} - w_i^T \tilde{w}_j\right)^2$$

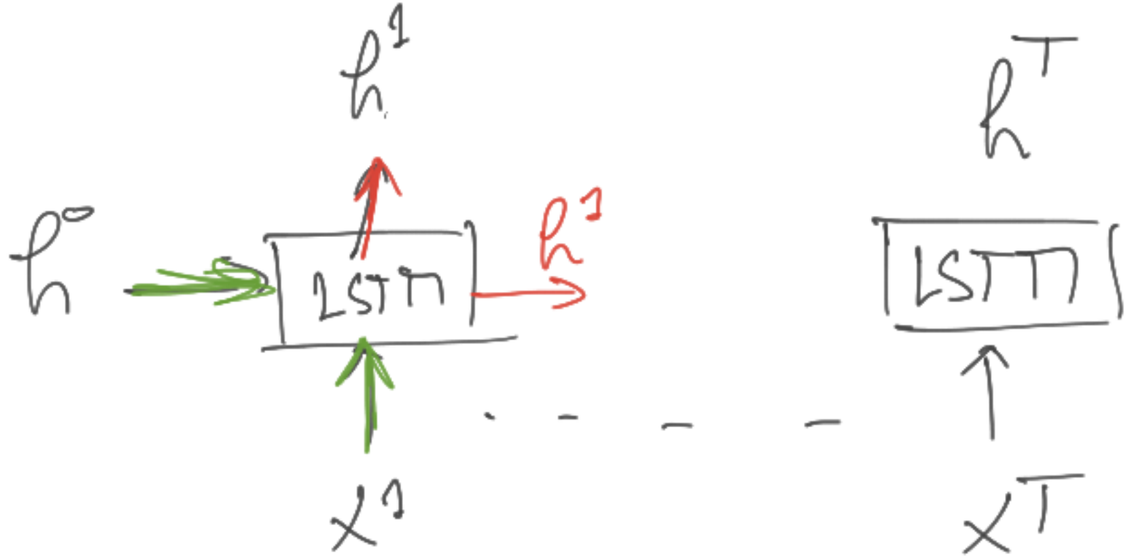


Embedding Layer (2/6)

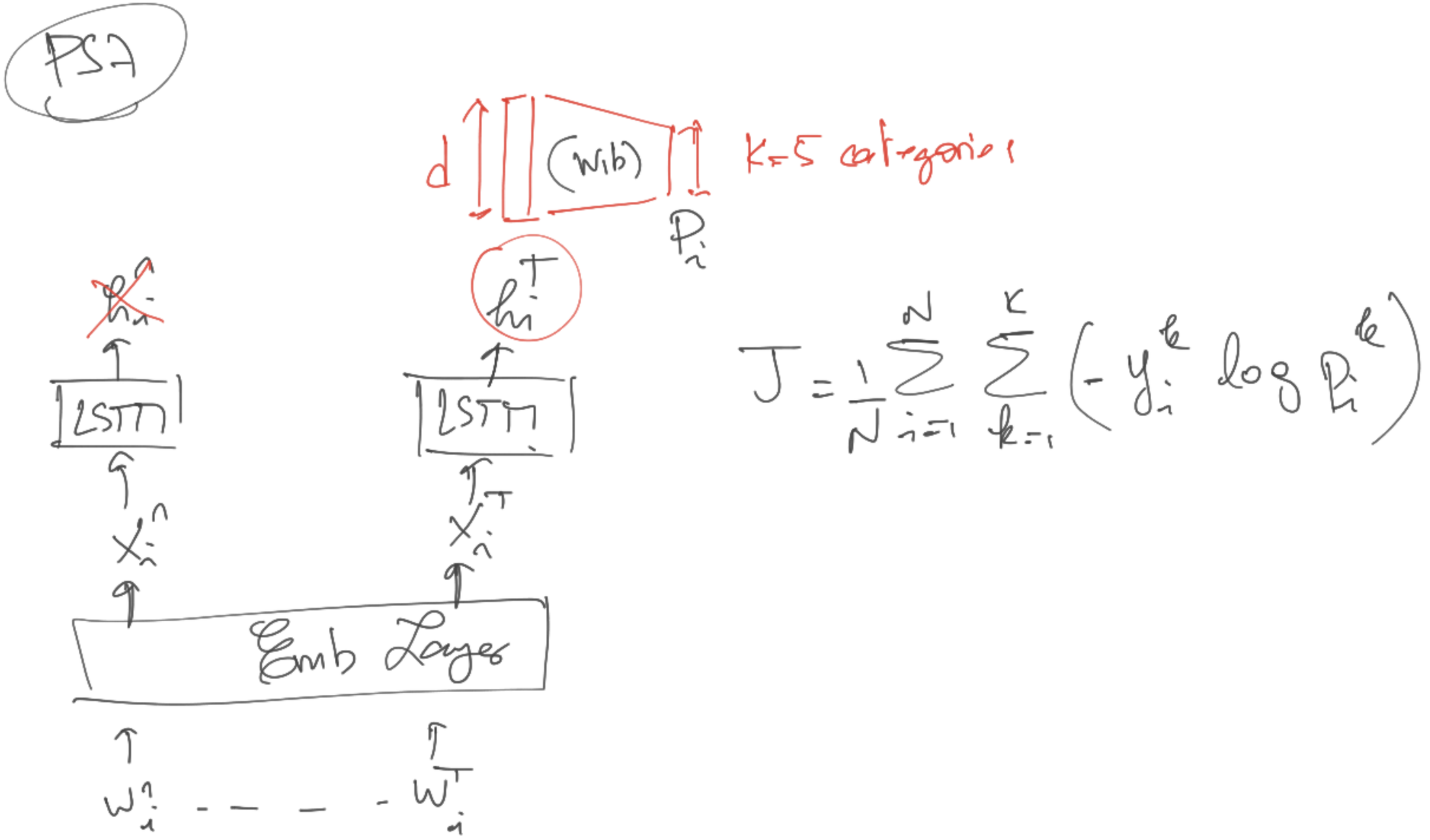


LSTM model

$x^1 \ldots x^T \Longrightarrow h^1 \ldots h^T$



PS7



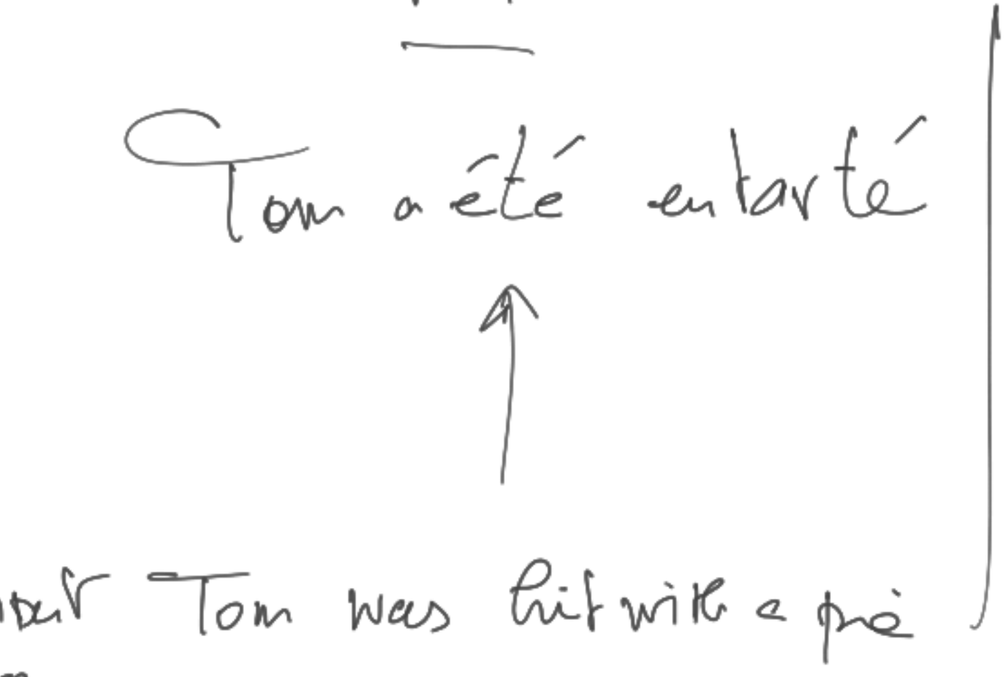
$$J = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\left(-y_i^k \log p_i^k\right)$$

Lecture 8

Many to one → Many to many

NLP

Tom a été entarté

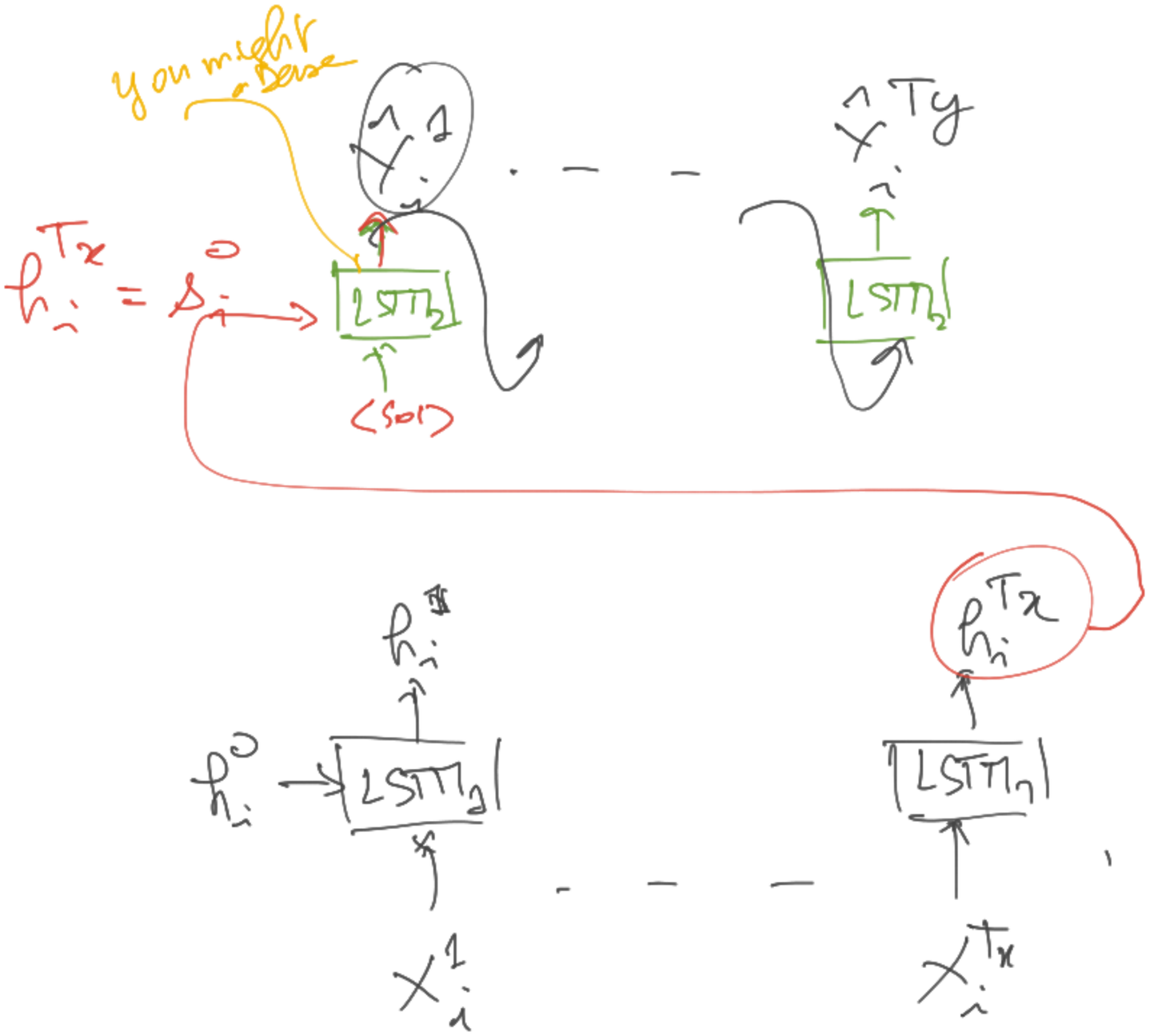


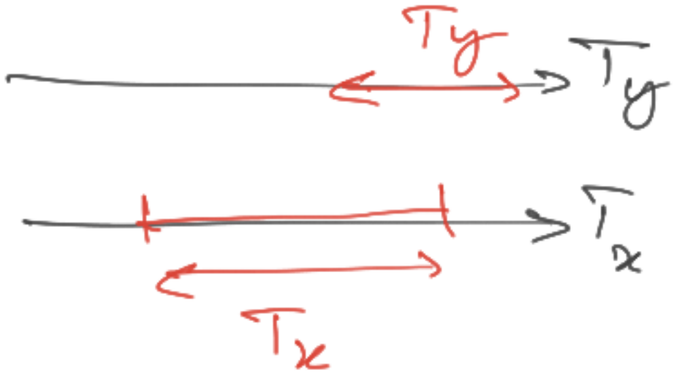
Input Tom was hit with a pie
Seq.

Finance

y^{T+1} y^{T+2} ... y^{T+H}

x^1 - - x^T



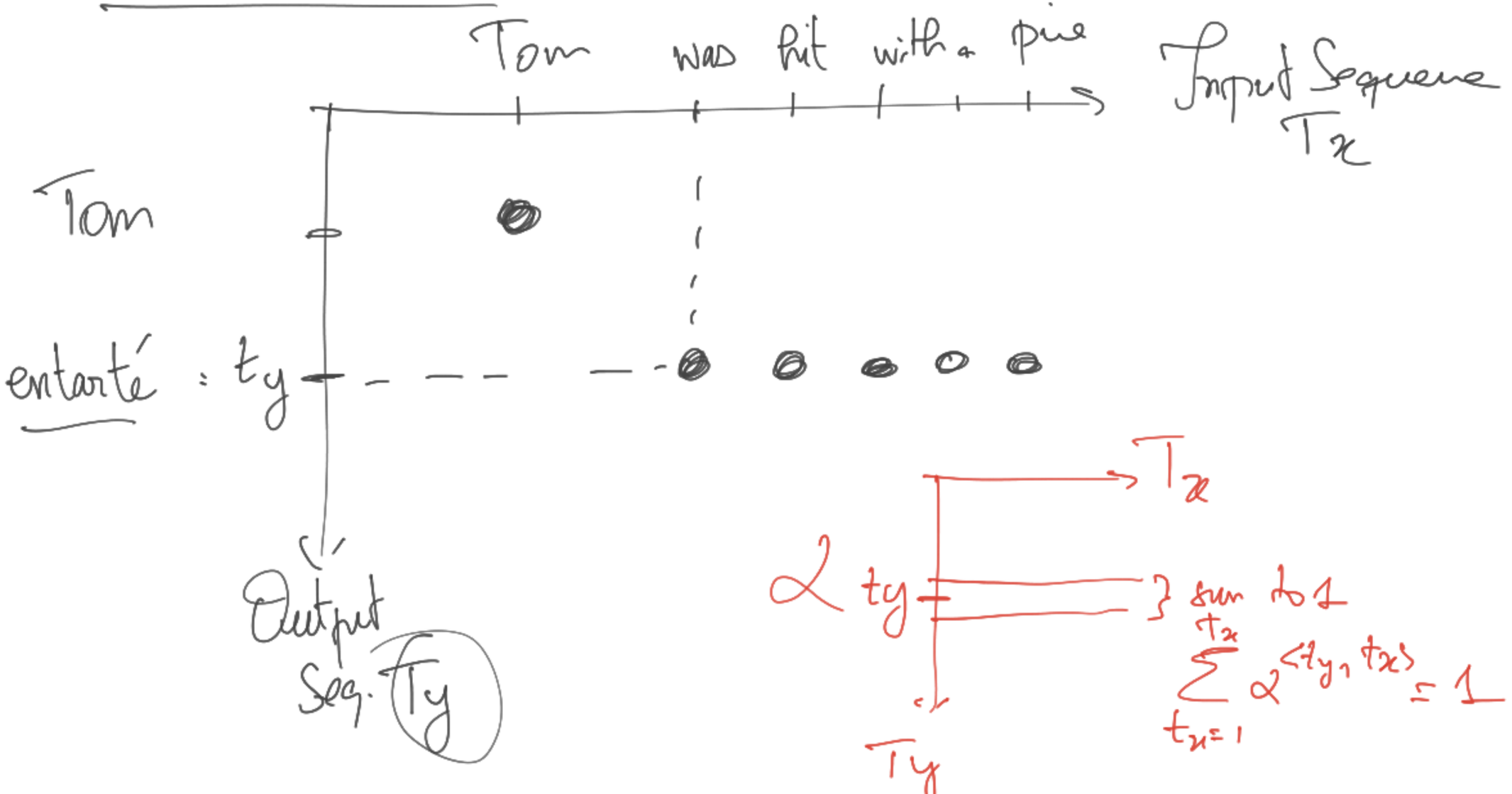


$\forall t_y \quad \hat{Y}_i^{t_y} \in \mathbb{R}^K$

$\begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} = Y_i^{t_y} \in \mathbb{R}^K$

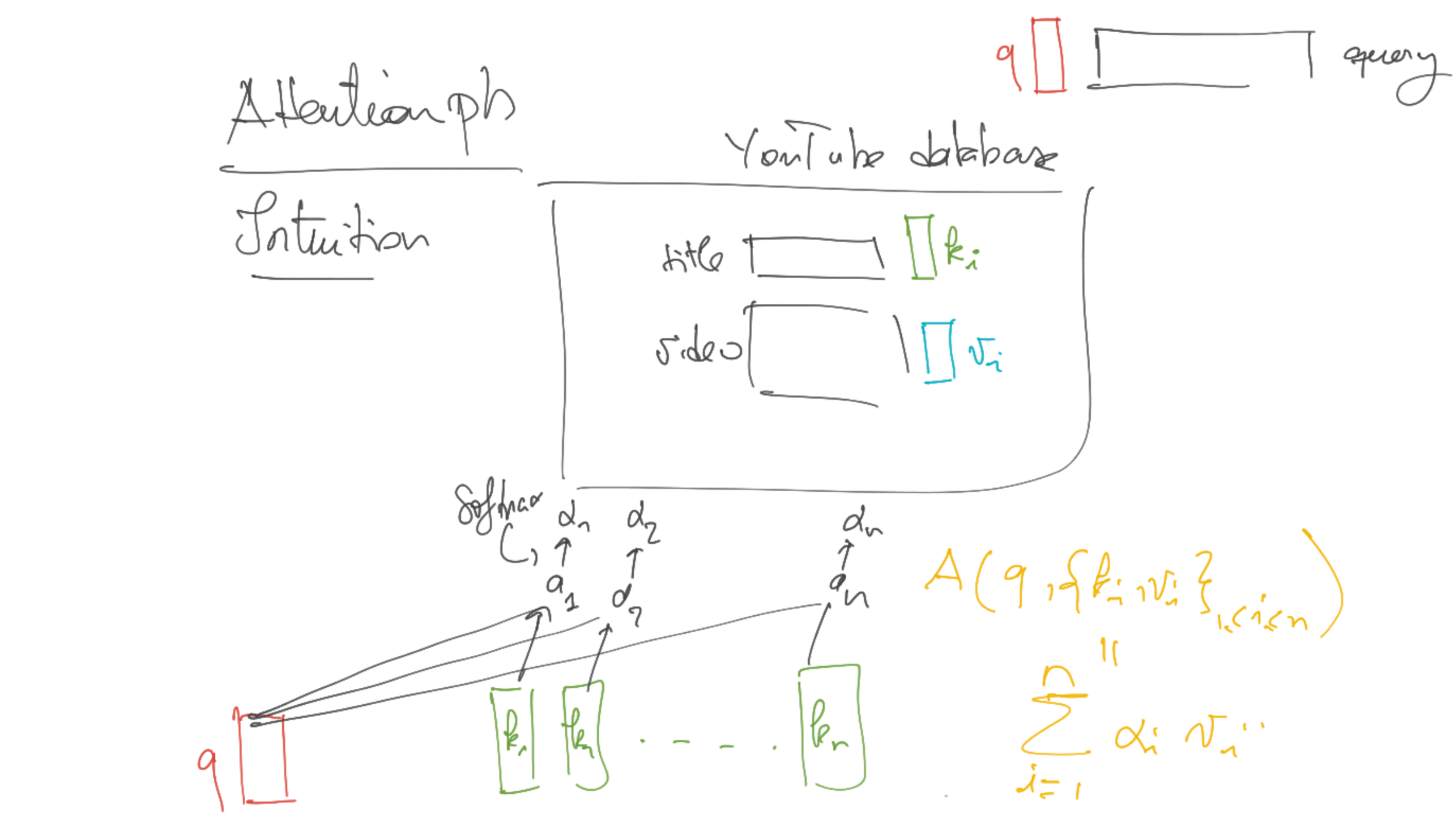
$$\frac{1}{T_y} \sum_{t_y=1}^{T_y} \sum_{k=1}^{K} \left(- Y_i^{t_y}(k) \log \hat{Y}_i^{t_y}(k) \right)$$

Alignment pb

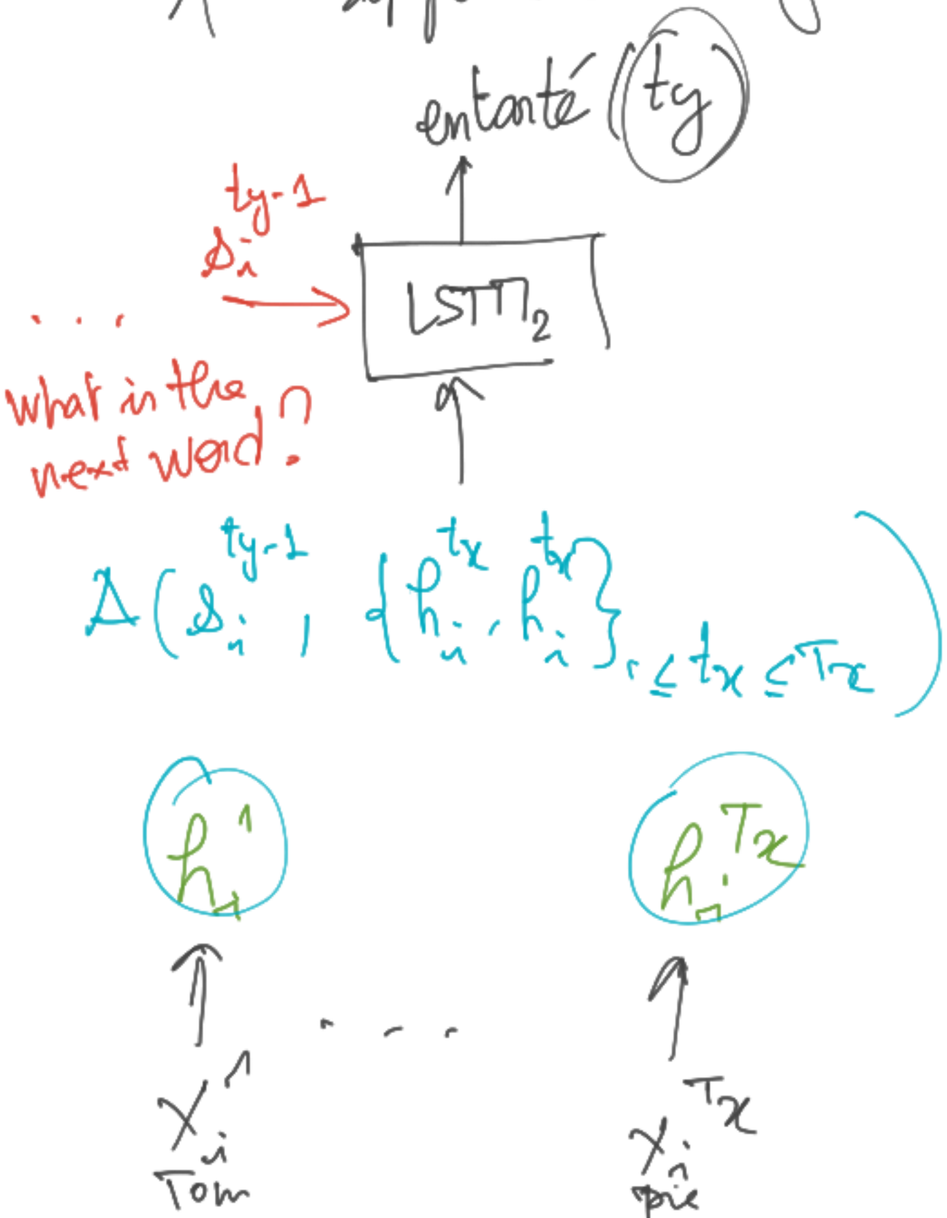


Attention

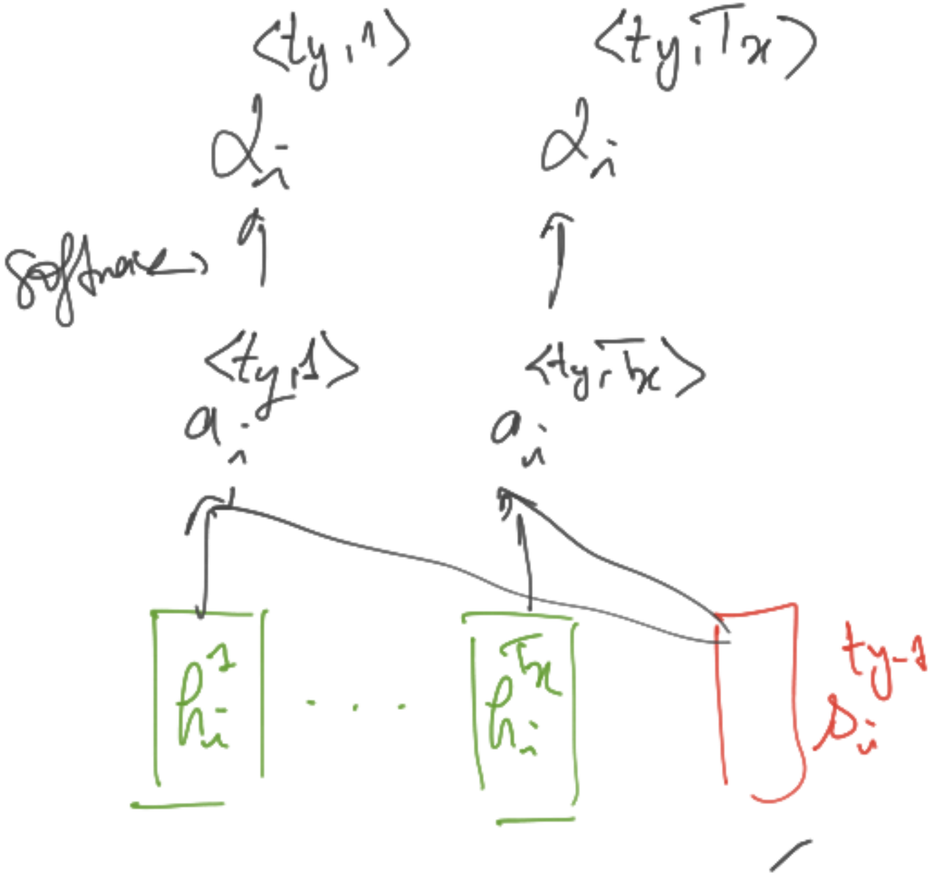
Intuition



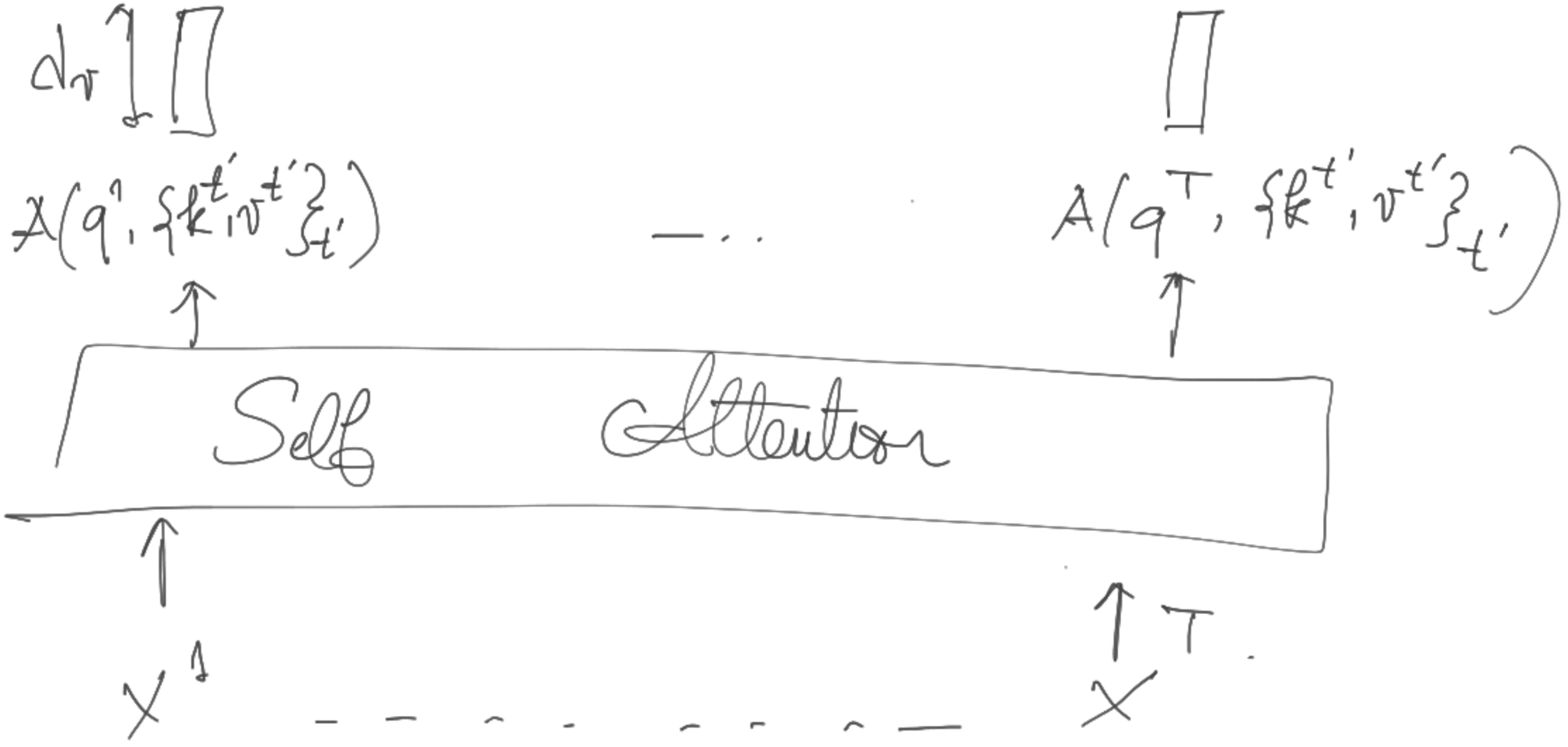
1st application of Attention (Not in real)

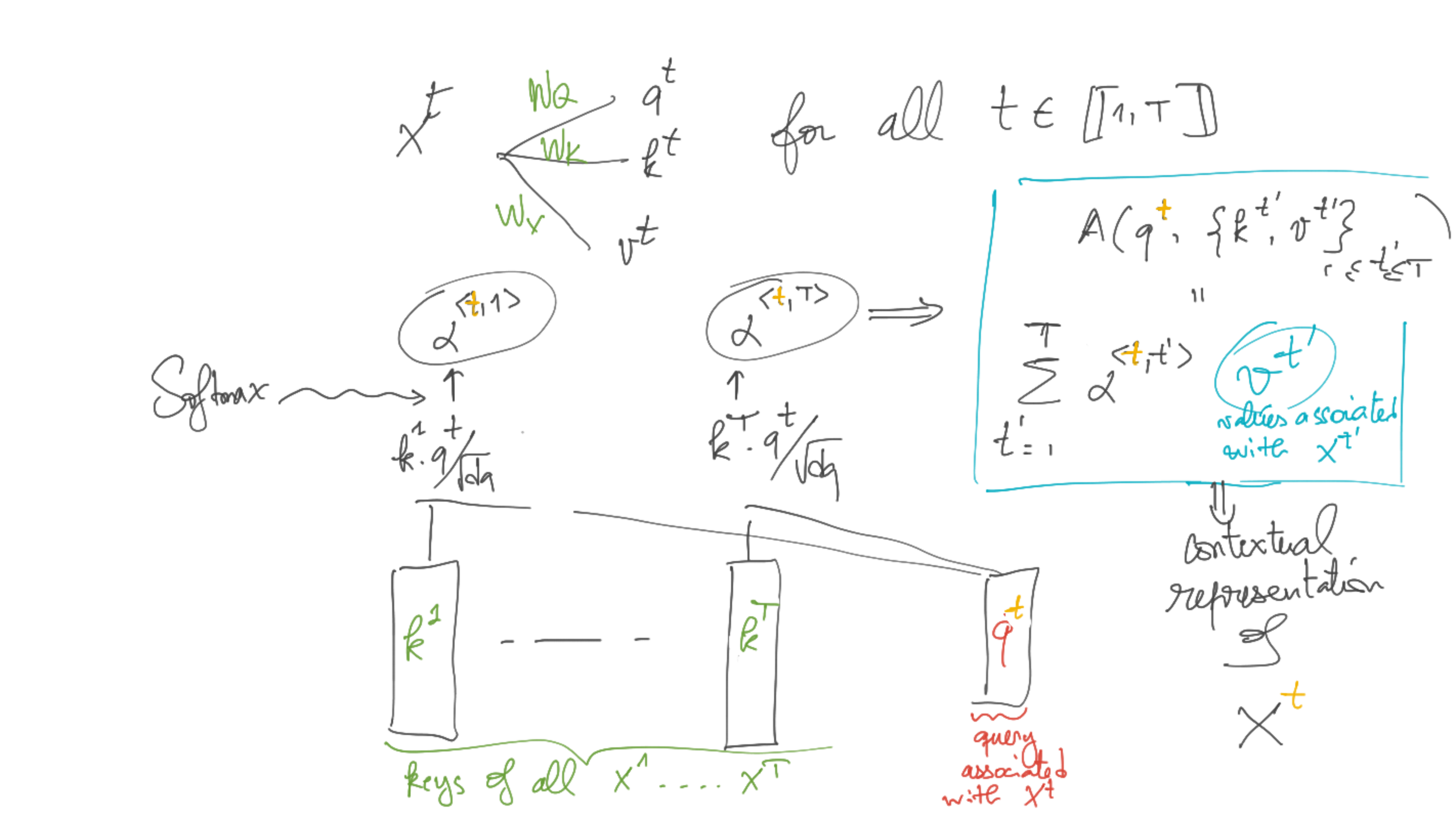


objective : put attention on "hit with a pie"



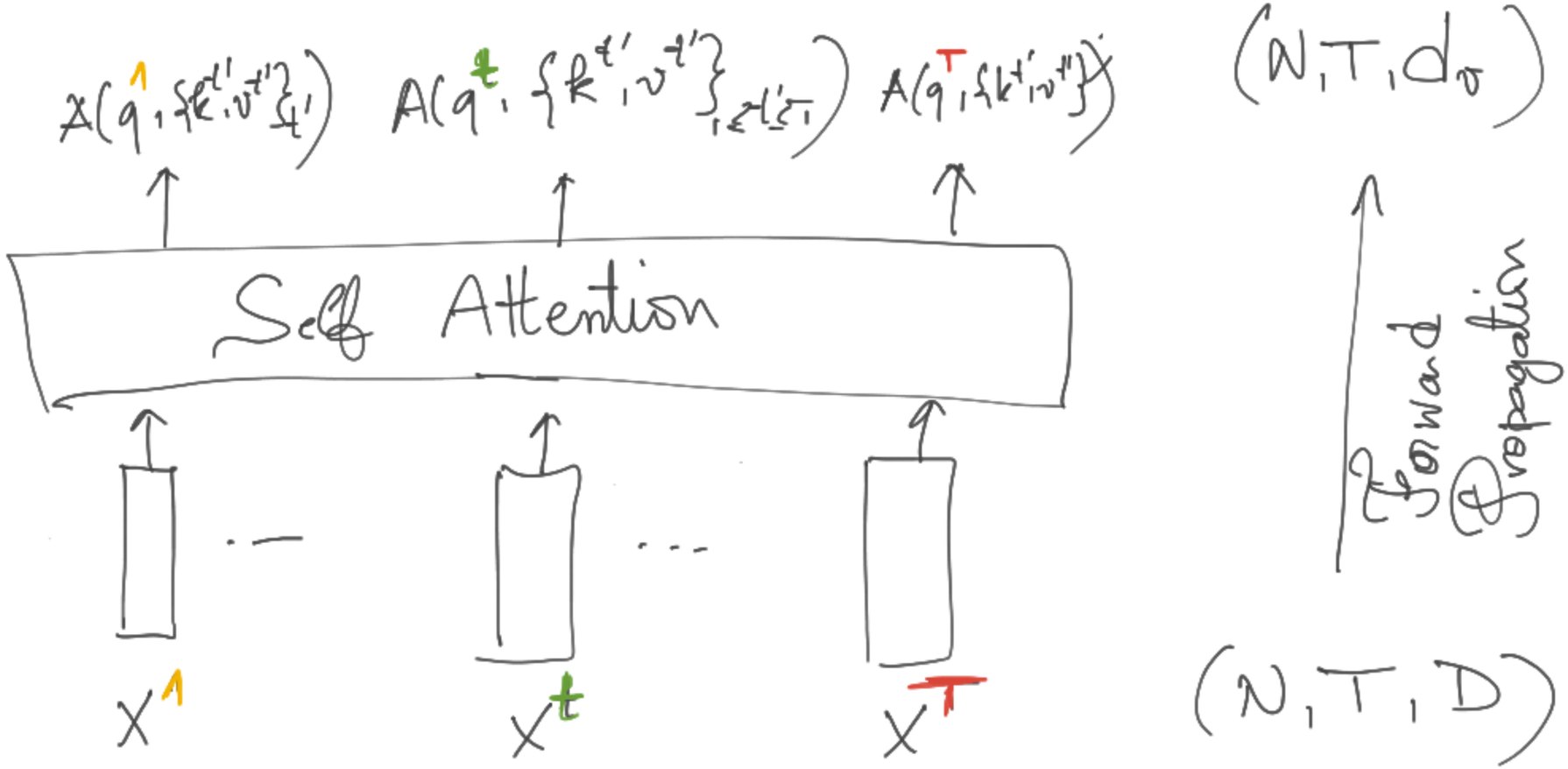
2nd Application of Attention: Contextual embedding vectors:




x^t
W_Q
W_K
W_V
q^t
k^t
v^t
for all $t \in [\![1, T]\!]$
$A(q^t, \{k^{t'}, v^{t'}\}_{r \le t' \le T})$
$= \sum_{t'=1}^{T} \alpha^{\langle t, t' \rangle} v^{t'}$
values associated with $x^{t'}$
$\alpha^{\langle t, 1 \rangle}$
$\alpha^{\langle t, T \rangle}$
Softmax
$k^1 \cdot q^t / \sqrt{d_q}$
$k^T \cdot q^t / \sqrt{d_q}$
k^1
k^T
q^t
keys of all $x^1 \ldots x^T$
query associated with x^t
contextual representation of x^t

The Self Attention Layer

Parameters : $W_Q \in \mathbb{R}^{D \times d_q}$; $W_k \in \mathbb{R}^{D \times d_k}$, $W_v \in \mathbb{R}^{D \times d_v}$



(N, T, d_v)

Forward Propagation

(N, T, D)

Optional Session on Monday, May 15th

→ Coding the Self Attention Layer

→ Introduce the Transformer architecture and Large Language Models

→ Correct part of exam 2022.