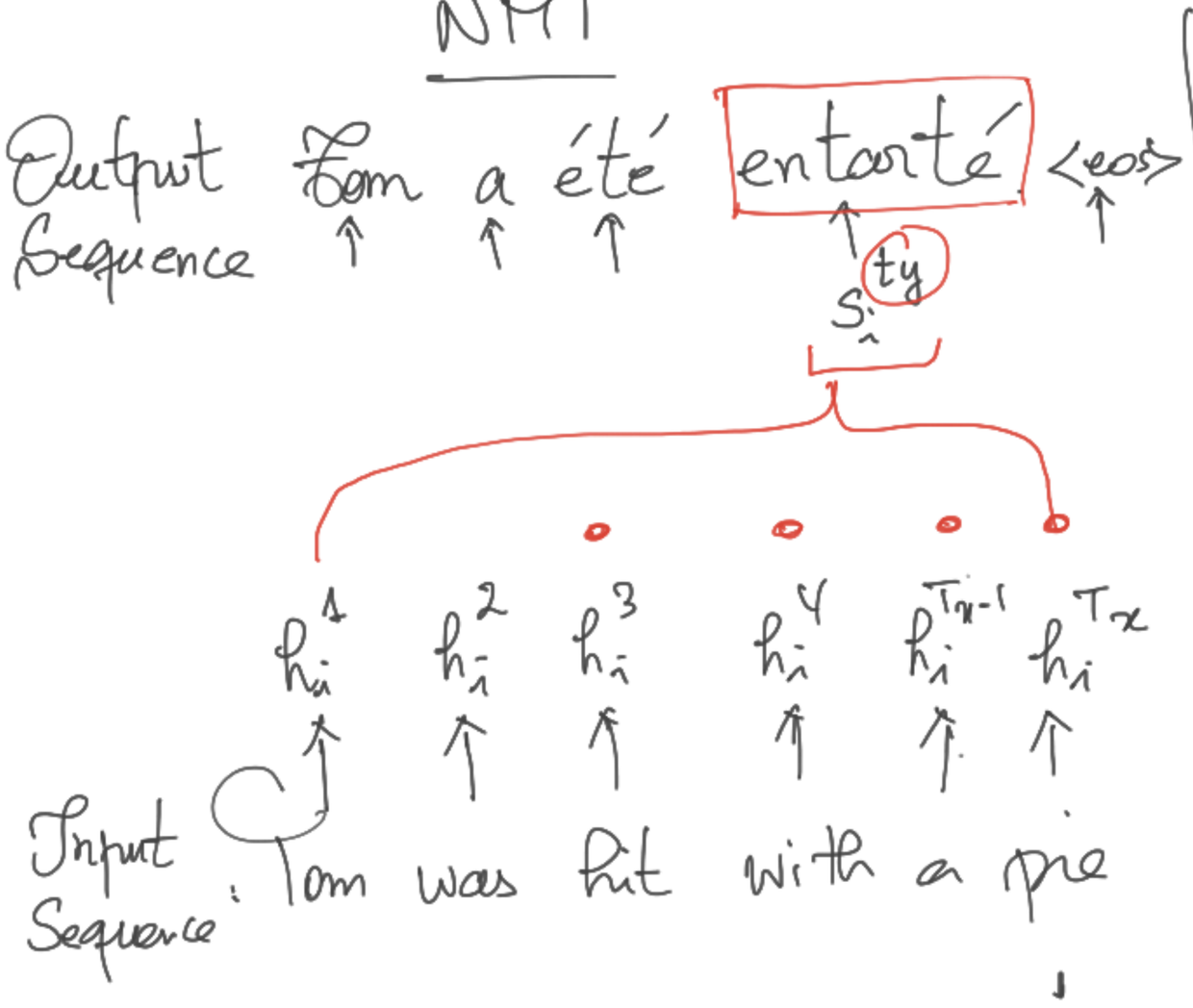
Outline:

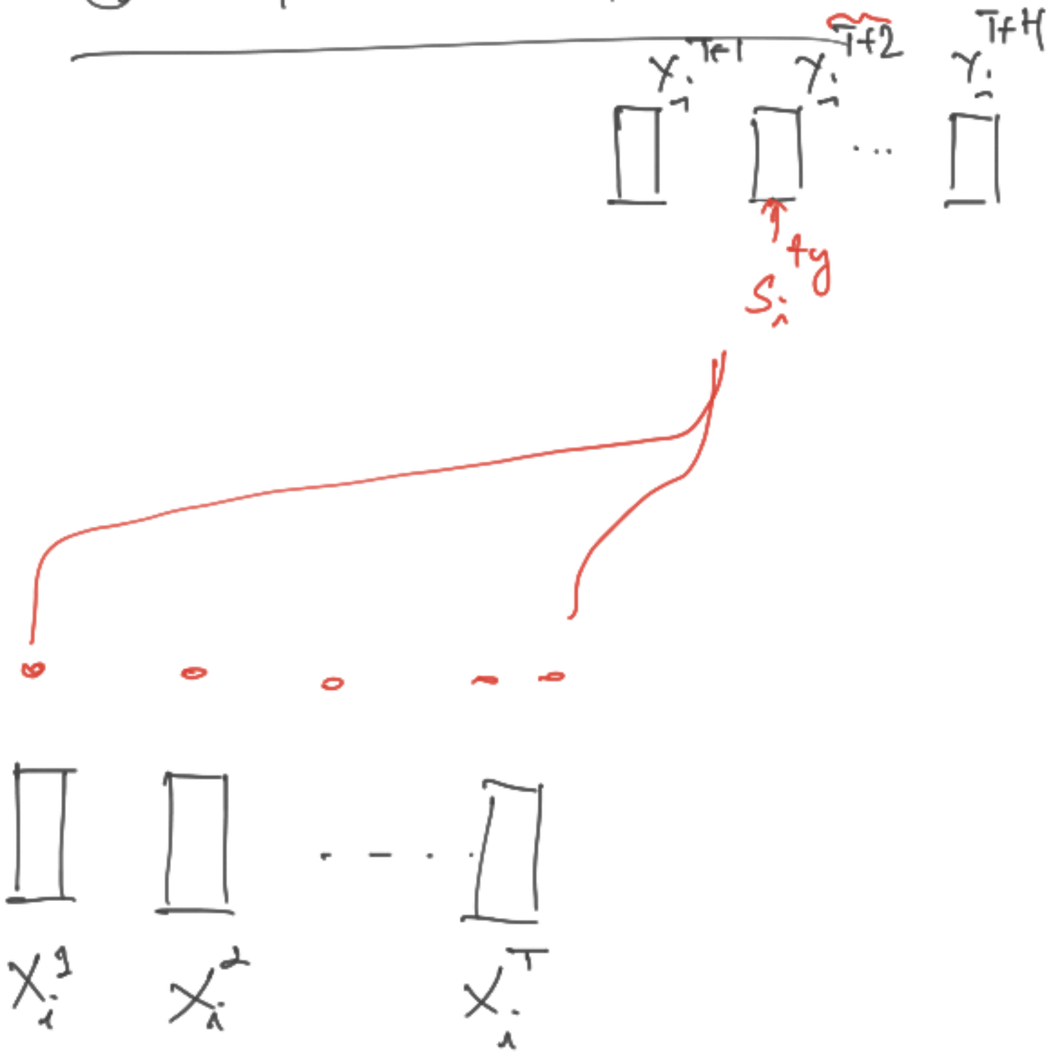
1. Applications of S2SWA (Many to many).
2. Introduce Self Attention in the S2SWA.
3. Introduce: Attention is all you need.
4. Correcting part of Exam 2022.

(1) Applications of S2S WA:

NMT

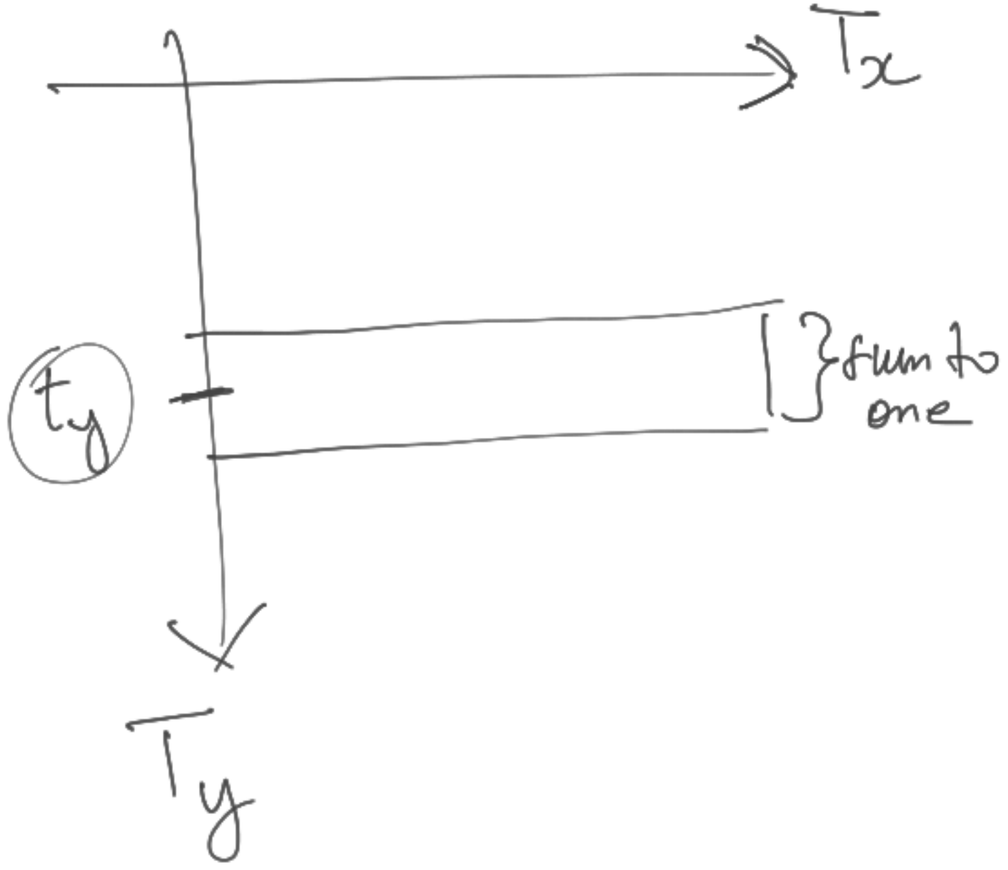


Time Series Predictions.

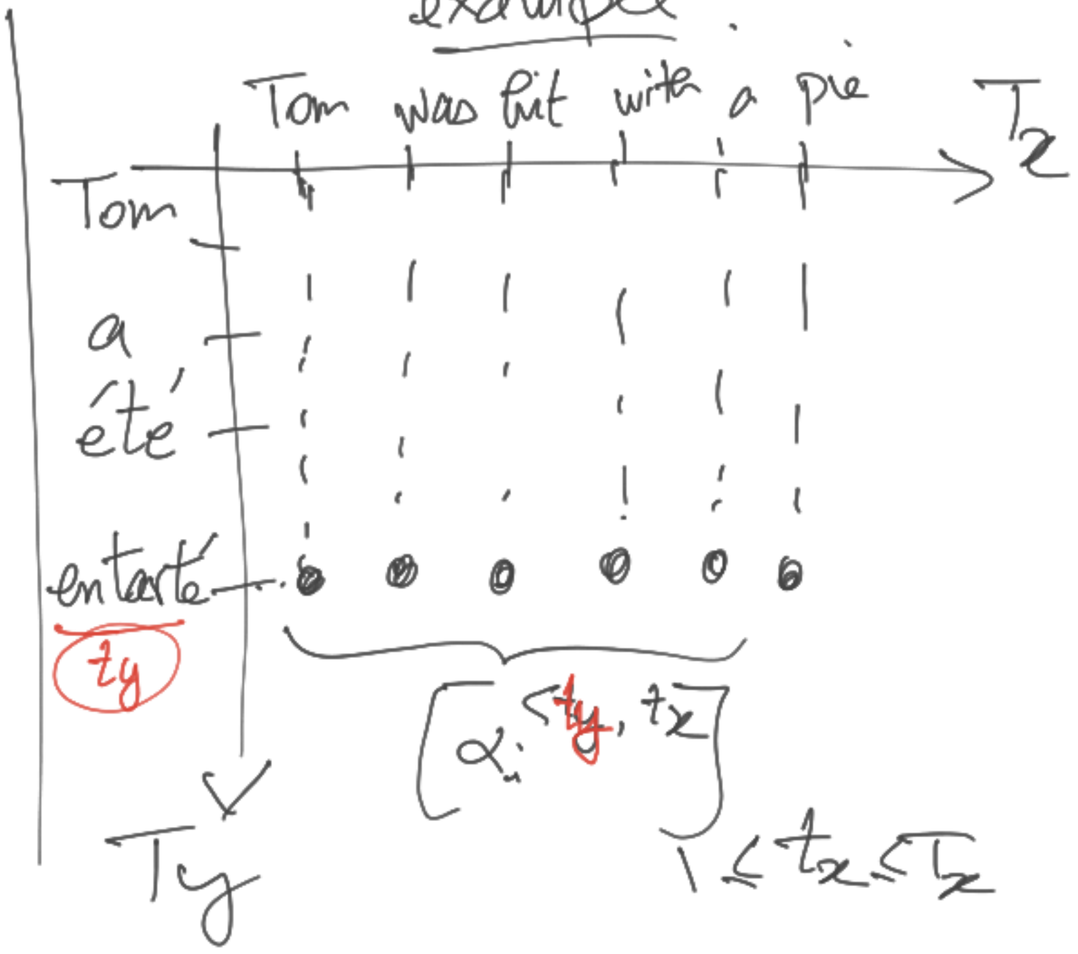


Alignment pb

Objective : $x_i^1 \dots x_i^{T_x} \longrightarrow \hat{y}_i^1 \dots \hat{y}_i^{T_y}$

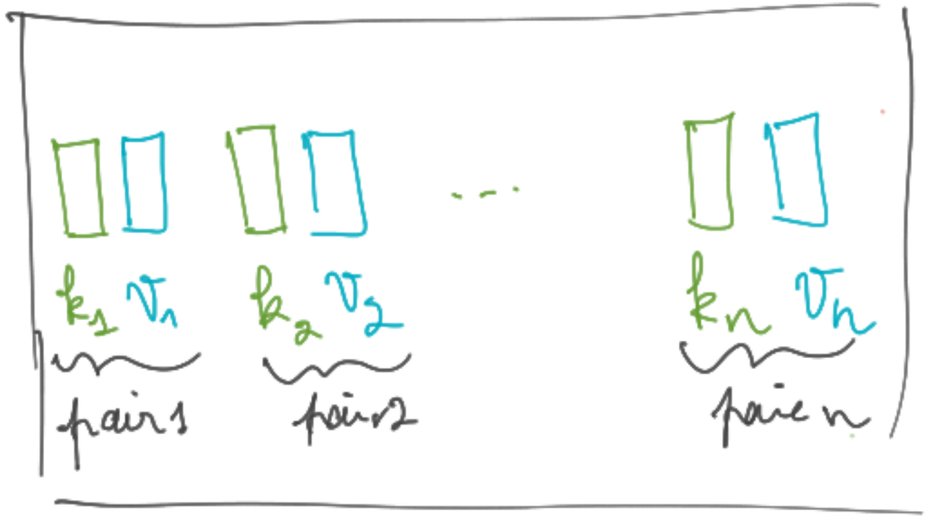


example



Attention :

$q \in \mathbb{R}^{d_q}$



Database

$$A(q, \{k_i, v_i\}_{1 \le i \le n}) = \sum_{i=1}^{n} \frac{\exp(q \cdot k_i)}{\sum_{j=1}^{n} \exp(q \cdot k_j)} v_i$$

α_i

we can use other similarity measures.

Self Attention:

objective : $X_i^1 \dots X_i^T$ (they represent embedding vectors in NLP)

↳ Create a sequence of contextual embedding vectors.

ex (28) : Tom a été entarté. (Tom was hit with a pie)

was

⇒ Cet été il fera horriblement chaud.

Summer

Self Attention Layer

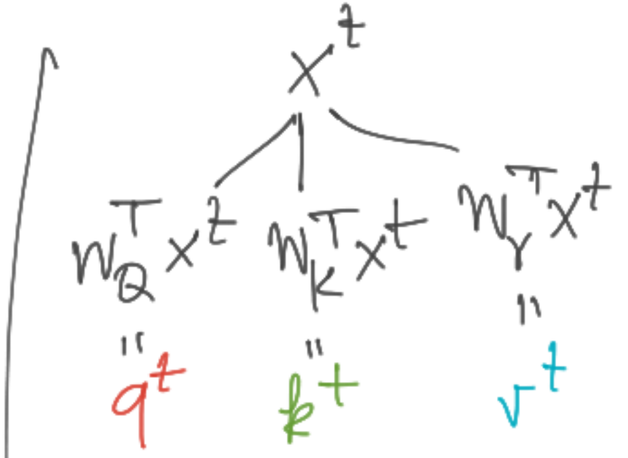
$(d_q = d_k)$

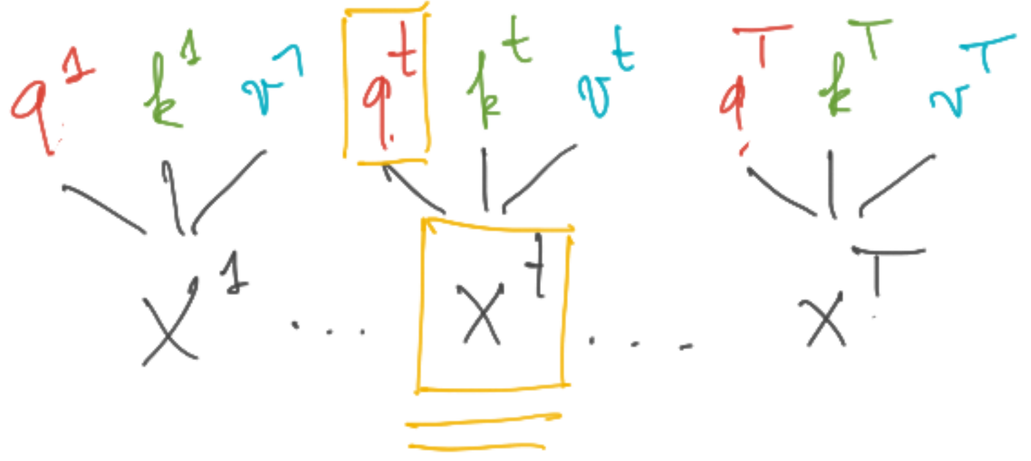
Parameters : $W_Q \in \mathbb{R}^{D \times d_q}$; $W_K \in \mathbb{R}^{D \times d_k}$, $W_V \in \mathbb{R}^{D \times d_v}$

Contextual representation of x^t :

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

$\alpha^{\langle t, t' \rangle}$: Impact of $x^{t'}$ in generation the contextual representation of x^t.

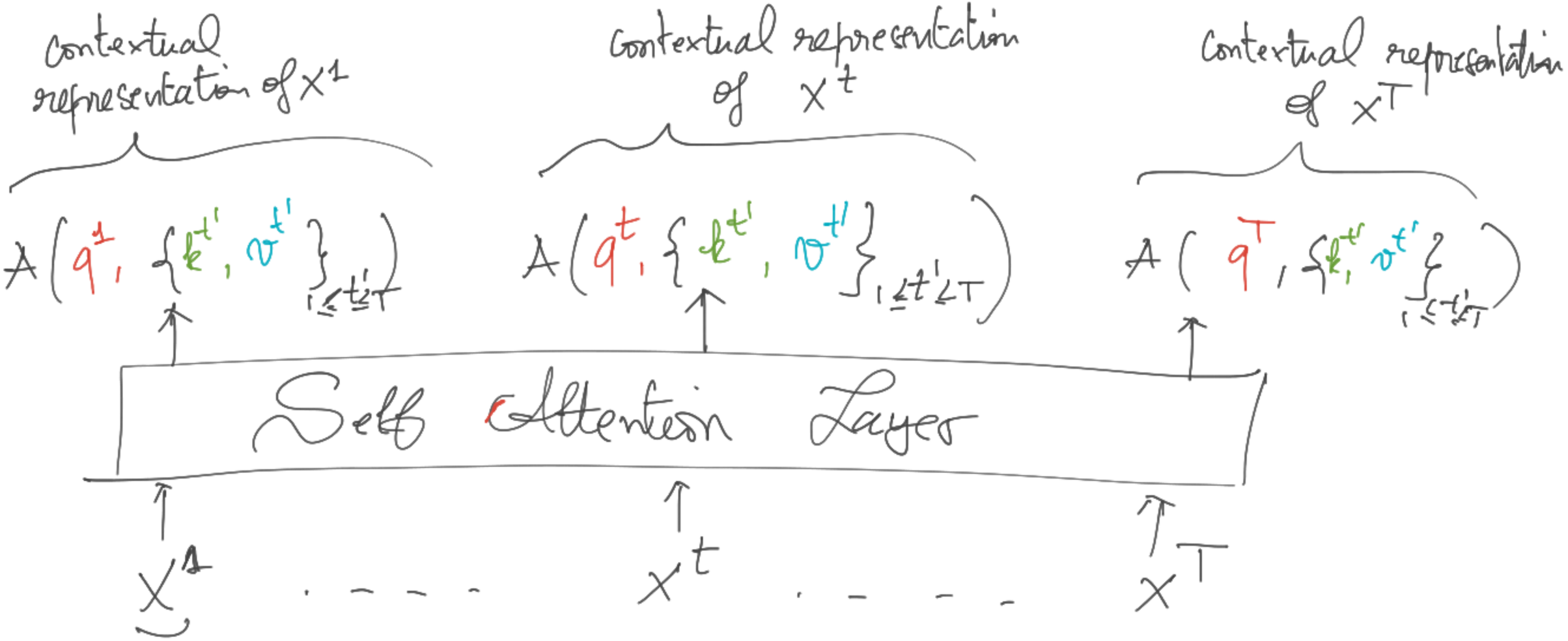




Self Attention Layer

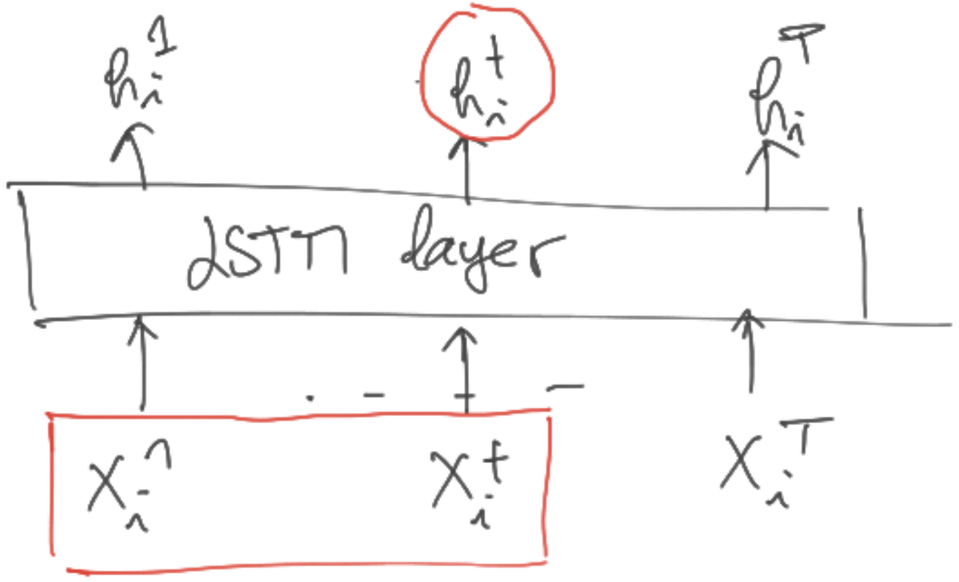
Forward Propagation

(N, T, D) — Self Attention. → (N, T, d_v)



LSTM Layer

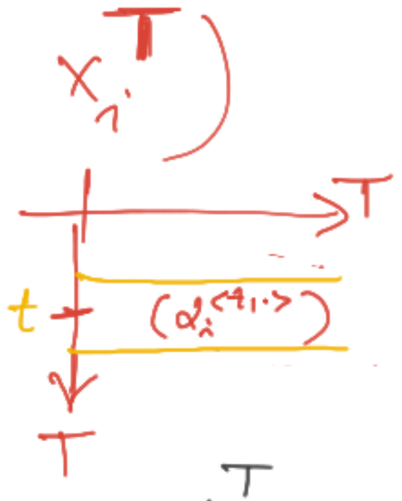
$$h_i^t = f(x_i^1, \ldots, x_i^t)$$

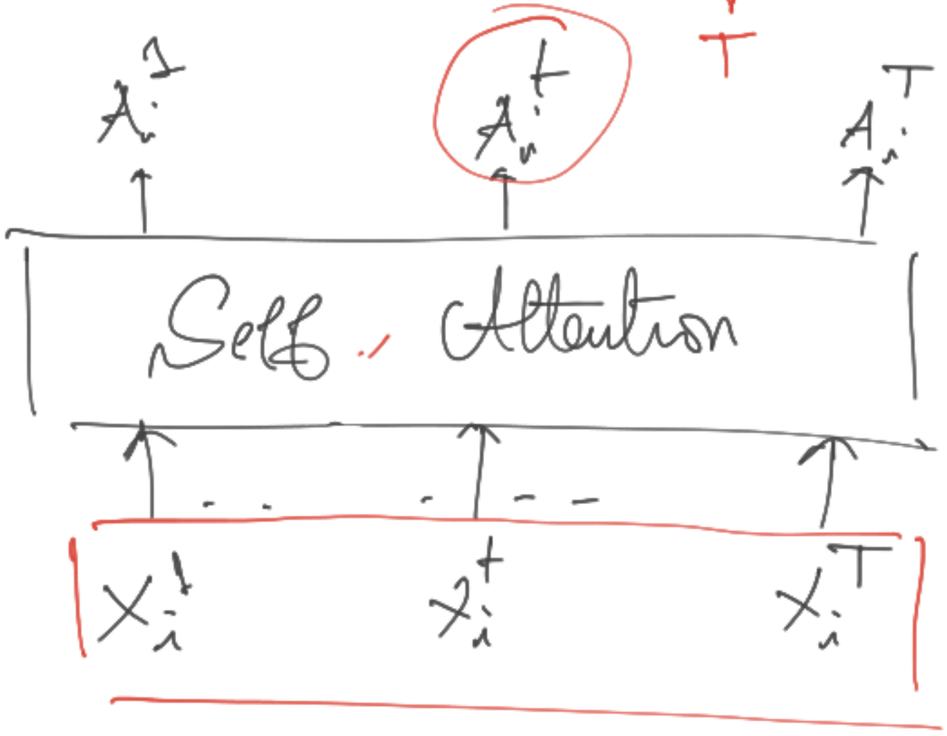


Self Attention Layer

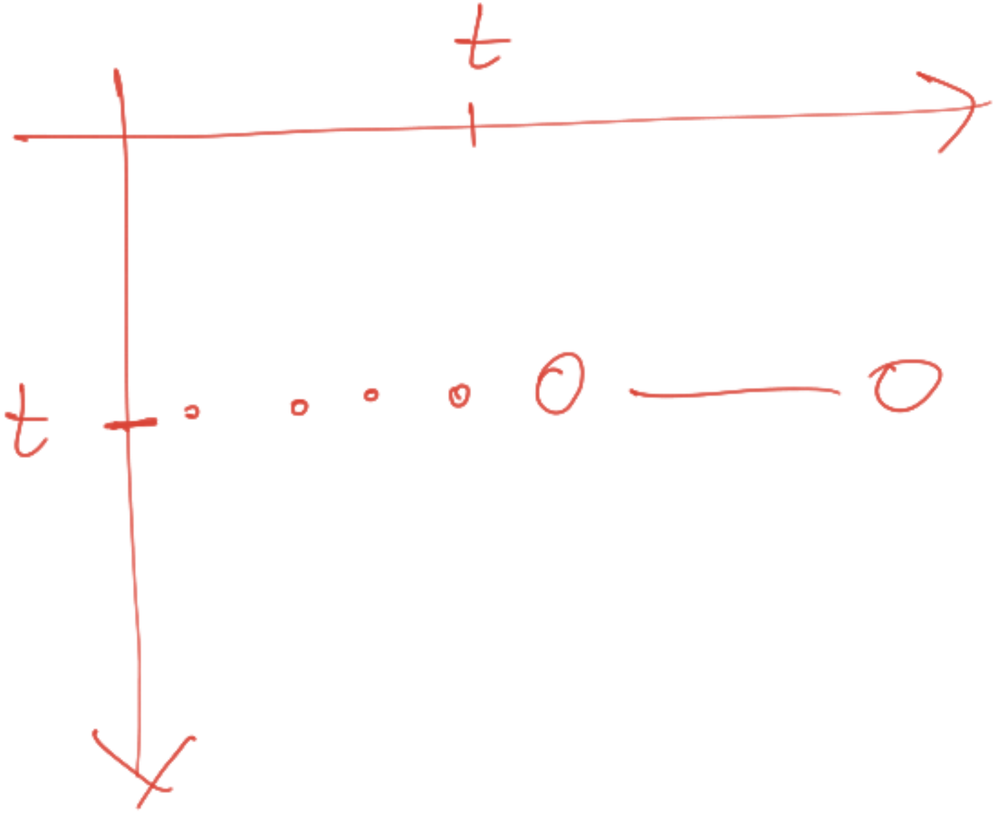
$$A_i^t = f(x_i^1 \quad \_\_\_\_\_ \quad x_i^T)$$

$$A_i^t = \sum_{t'=1}^{T} \alpha_i^{\langle t,t' \rangle} v_i^{t'}$$





↝ Causal Self Attention Layer (Masked Attention).



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

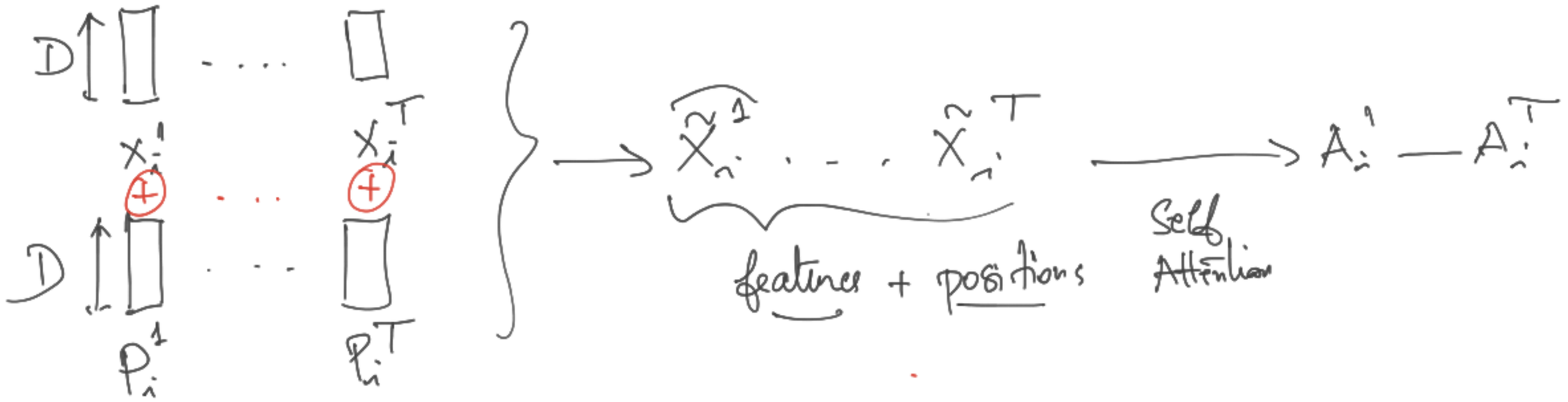
$$\forall\, t' > t \quad \alpha^{\langle t, t' \rangle} = 0$$

$$A(q^t, \{k^{t'}, v^{t'}\}_{t' \le t}) = f(x^1 \cdots x^t)$$

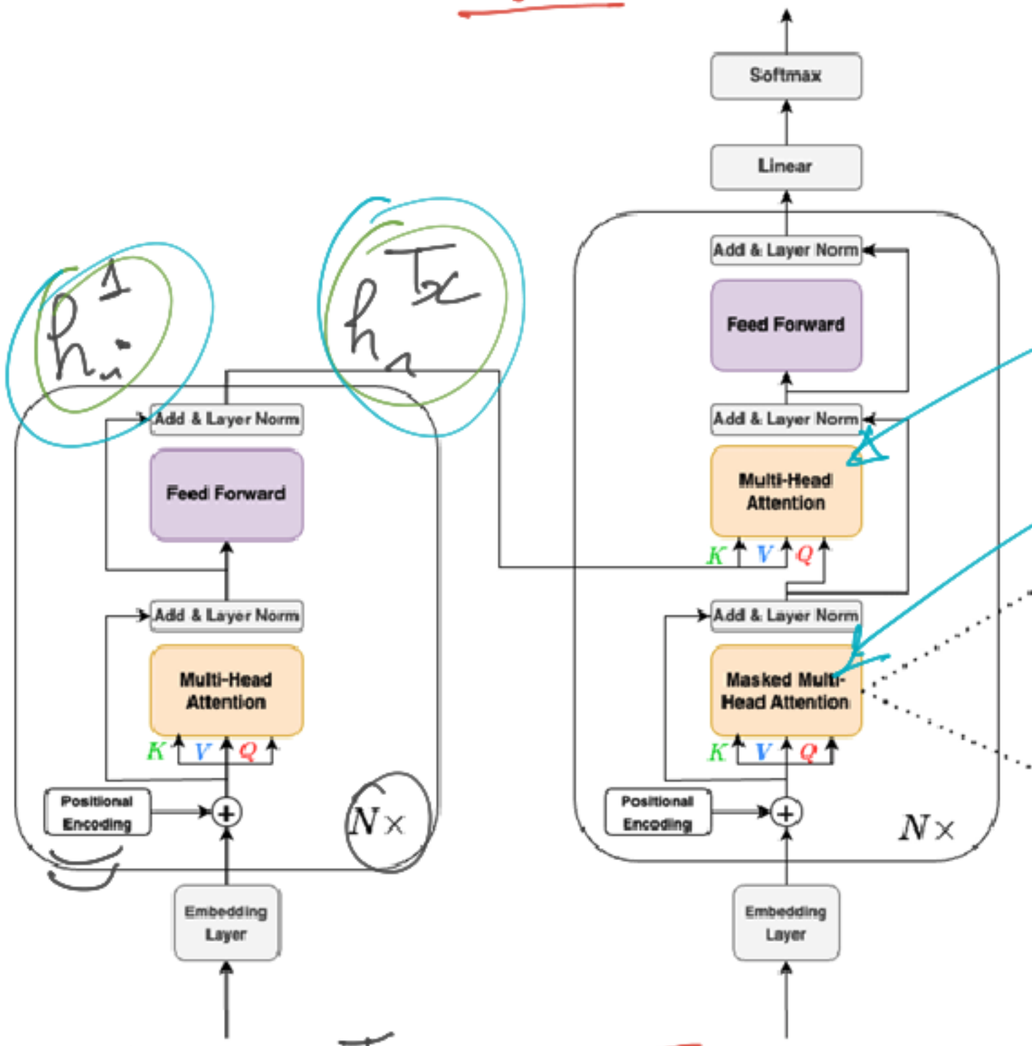
Positional Encoding:

- Pb : $A(q^t, \{k^{t'}, v^{t'}\}_{1 \le t' \le T})$ doesn't take into consideration the order of $x_i^1 \dots x_i^T$.

- Sol :



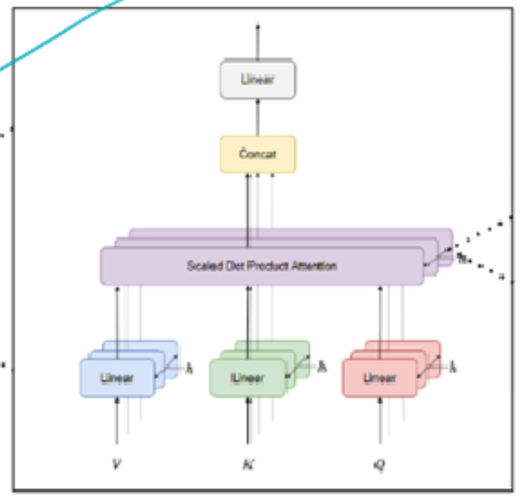
objective "Tom a été entarté <eos>" : Decoder Output

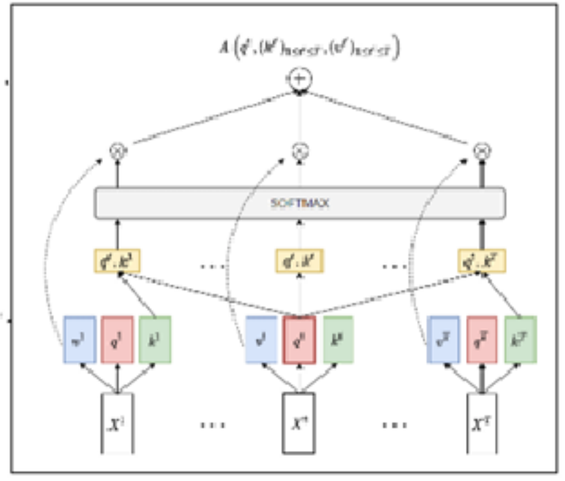


Cross Attention Layer (same as before)

Decoder → query
Encoder → key, value

Causal Attention Layer



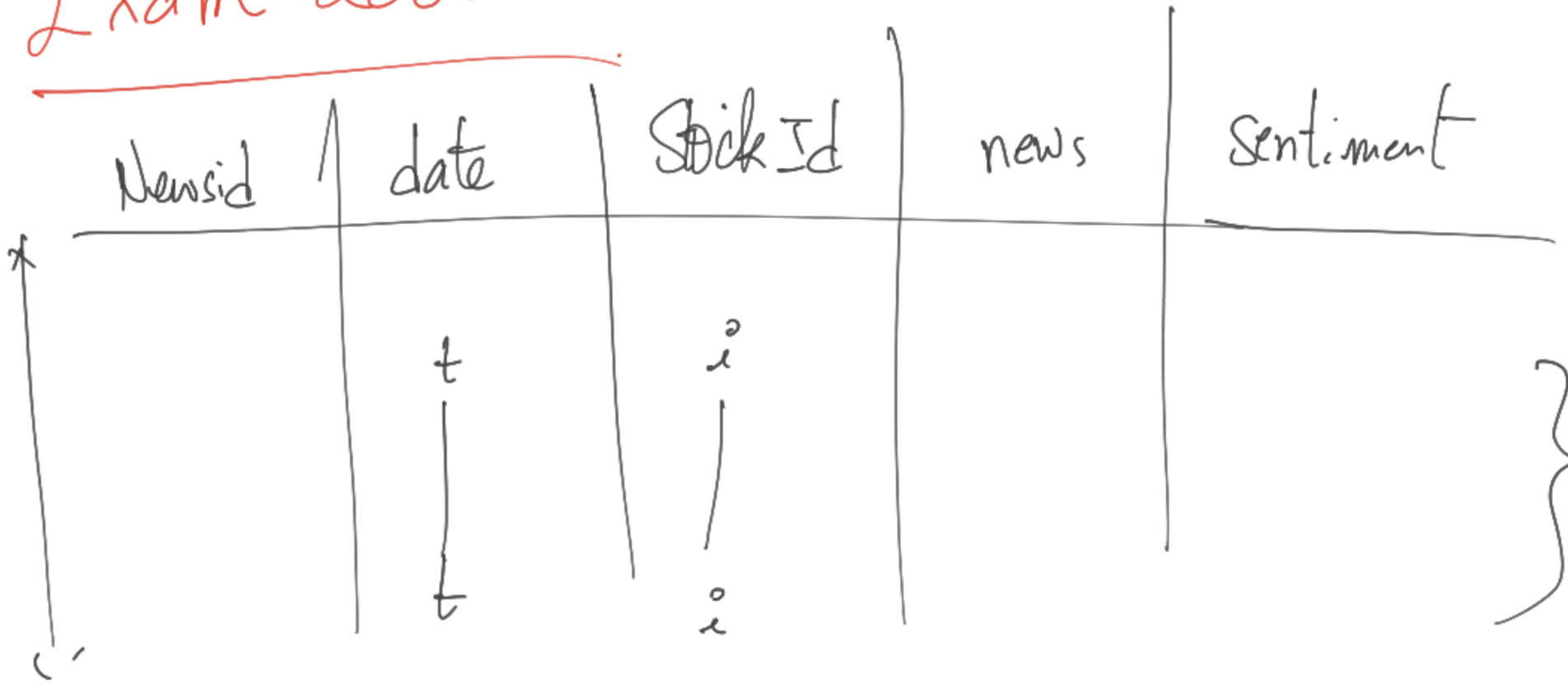


$X_{\hat{}}^{1} \dots X_{\hat{}}^{T_x}$

Tom was hit with a pie

[<sos> Tom a été entarté] : Decoder Input

Exam 2022



$n_{t,i}$ news related to stock of id i at time t.

Q10 $N = \sum_{t=1}^{T} \sum_{i=1}^{Nu} n_{t,i}$

Q11

newsId	News
1	$w_1 = (w_1^0 \dots w_1^L)$
⋮	⋮
N	$w_N = (w_N^0 \dots w_N^L)$

→ describe the word_index to map words ⟶ integers

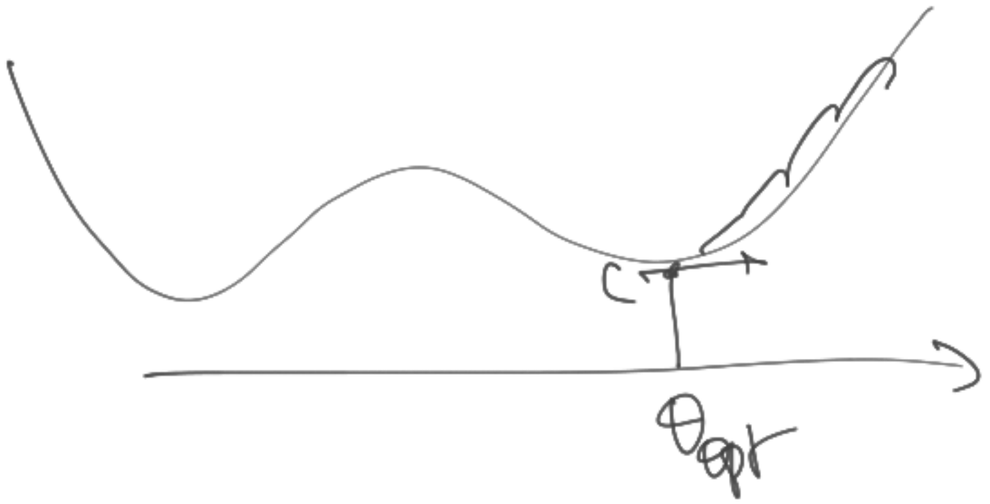
→ Padding to make all the sequences of the same size. (See Review2 Slide2)

Q12 (N, L) tensor.

Q14 ⚠ — How to gather the data to create X
— Explain the train of:

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

J ?



in θ_{opt} we have the embeddings

matrix $E = \begin{bmatrix} — w_1 — \\ \vdots \\ — w_V — \end{bmatrix}$

(Slide 9 Review2)

Q15 $W_e \in \mathbb{R}^{V \times D}$

Q16

$x_i^1 \; \cdots \; \boxed{x_i^l} \; \cdots \; x_i^2$] sequence of embeddings $\vec{v}$

Emb Layer ⇝

$w_i^1 \; \cdots \; \boxed{w_i^l} \; \cdots \; w_i^2$] sequence of integers

27

$$x_i^l = W_e[w_i^l]$$

↑ row # w_i^l in W_e.

Q17

$\begin{bmatrix} h_i^{l-1} \\ C_i^{l-1} \\ x_i^l \end{bmatrix}$ past memory $\vec{v}$ — new obs

LSTM

W_F, b_F
W_I, b_I
W_O, b_O
W_C, b_C

$\rightarrow (h_i^l, C_i^l)$

Gates

$$F_i^l = \sigma\left[W_F^T \left[h_i^{l-1}, x_i^l\right] + b_F\right]$$

$$I_i^l = \sigma\left[W_I^T \left[h_i^{l-1}, x_i^l\right] + b_I\right]$$

$$O_i^l = \sigma\left(W_O^T \left[h_i^{l-1}, x_i^l\right] + b_O\right)$$

Update equations

$$\tilde{C}_i^l = \tanh\left(W_C^T \left[h_i^{l-1}, x_i^l\right] + b_C\right)$$

$$C_i^l = F_i^l \odot C_i^{l-1} + I_i^l \odot \tilde{C}_i^l$$

$$h_i^l = O_i^l \odot \tanh\left(C_i^l\right)$$