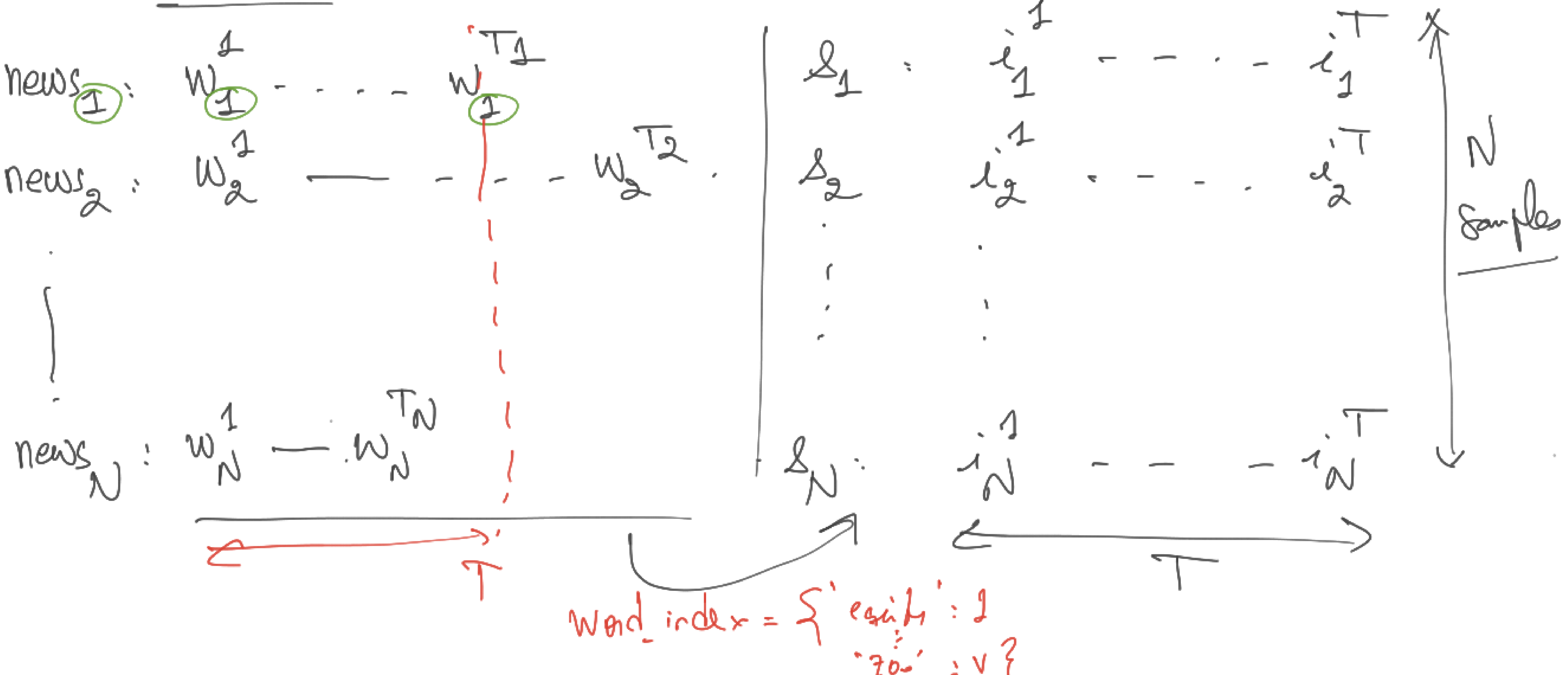
Outline : Programming Session 7

1. Preprocessing documents (texts) $\longrightarrow$ (N, T) tensors
2. Embedding Layer (N, T) $\longrightarrow$ (N, T, D) tensors
3. The Model.

1 Processing Dataset.

$\forall m \in [\![1,N]\!] \ \forall t \in [\![1,T]\!] \quad i_m^t \in \{1,\dots,V\}$



Processing the targets:

$$\text{news}_i \xrightarrow[NN]{f_\theta} p_i = \begin{bmatrix} p_i^1 \\ \vdots^K \\ p_i^K \end{bmatrix} \quad ; \quad y_i \in \{1, \dots, K\} \xrightarrow[\text{hot encoding}]{\text{one}} \tilde{y}_i = \begin{bmatrix} \tilde{y}_i^1 \\ \vdots^K \\ \tilde{y}_i^K \end{bmatrix}.$$

⇒ Loss associated with (news_i, y_i)

$$J_i = - \sum_{k=1}^{K} \tilde{y}_i^k \log p_i^k$$

⇒ Loss associated with the dataset

$$J = - \sum_{i=1}^{N} \sum_{k=1}^{K} \tilde{y}_i^k \log p_i^k .$$

② Embedding Layer:

↳ Glove approach

400K english words

$$X = \begin{bmatrix} & X_{ij} & \end{bmatrix}_{V \times V}$$

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

learning : 27
statistics : 54

$X_{27,54}$??

statistics 54
banana 98

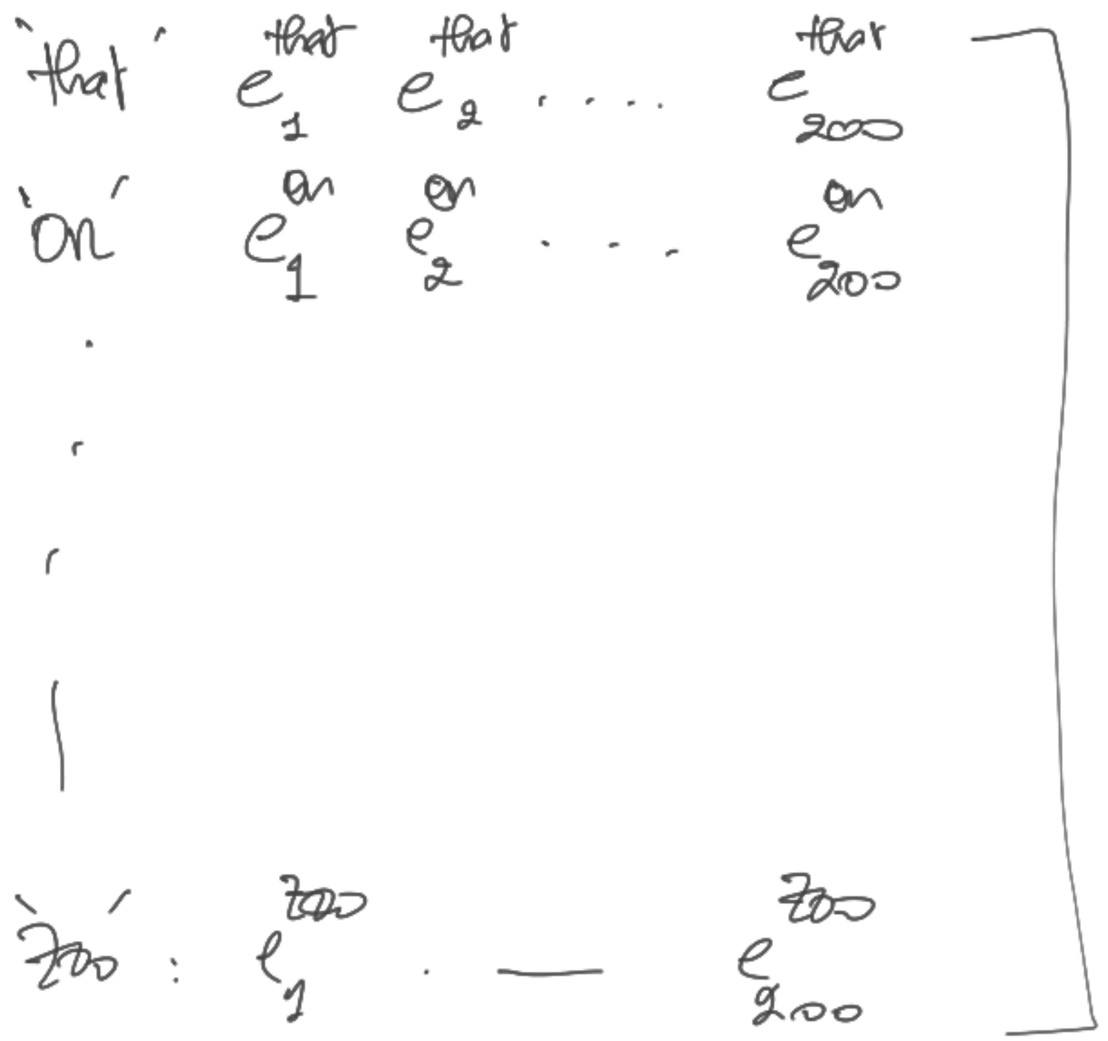
$X_{54,98}$??

objective : $W = \begin{bmatrix} — & w_1 & — \\ & \vdots & \\ — & w_V & — \end{bmatrix} \in \mathbb{R}^{V \times D}$

$\tilde{W} = \begin{bmatrix} — & \tilde{w}_1 & — \\ & \vdots & \\ — & \tilde{w}_V & — \end{bmatrix} \in \mathbb{R}^{V \times D}$

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

file: glove.6B.200d.txt



embedding =

$\{$ 'that' : $[e_1^{that}, \ldots, e_{200}^{that}]$,

.

.

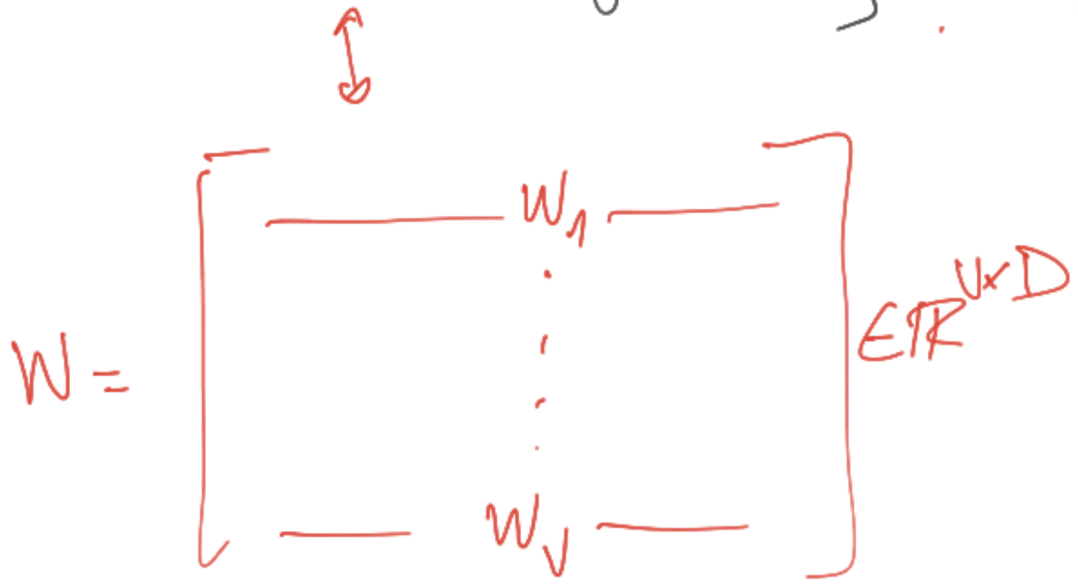
.

'zoo' : $[e_1^{zoo}, \ldots, e_{200}^{zoo}]\}$

400k words

word_index = { 'equity' : 1,
'market' : 2,
⋮
'finance' : V }

embedding = { 'that' : $[e_1^{that} \dots e_{200}^{that}]$,
⋮
'200' : $[e_1^{200} \dots e_{200}^{200}]$ }

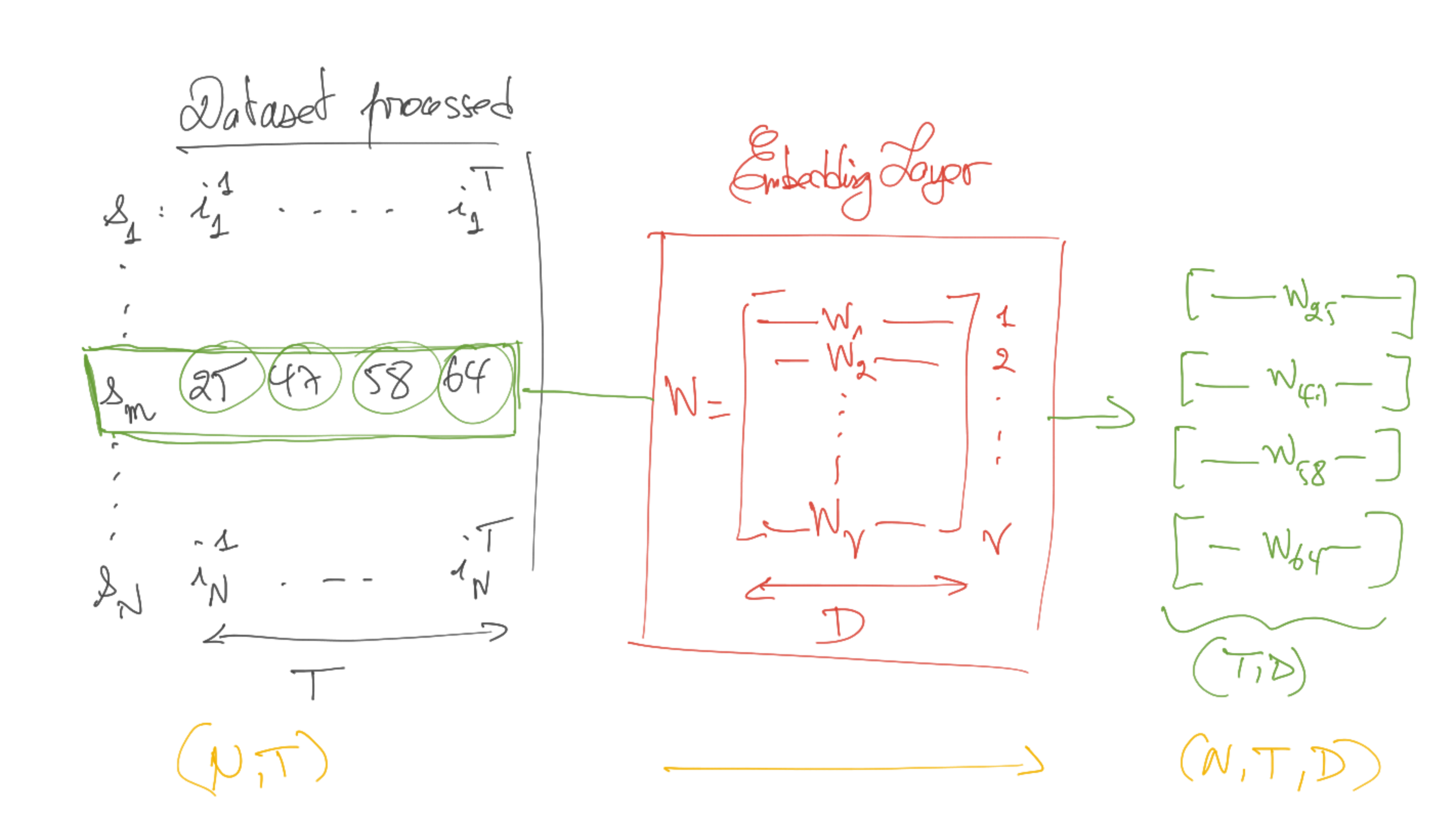


W_i : D-dim vector representing the word of index i in word_index

pseudo-code for W (V x D)

=> init W with zeros.

=> for word, index in word_index:
    embedding_vector = embedding[word]
    W[index] = embedding_vector.


Dataset processed
1
T
s_1 : i_1^1 . . . . i_1^T
s_m
25 47 58 64
s_N i_N^1 . . . i_N^T
T
(N, T)
Embedding Layer
$W =$
w_1
w_2
w_V
1
2
V
D
w_{25}
w_{47}
w_{58}
w_{64}
(T, D)
(N, T, D)

LSTM Layer:

Inputs: embedding vectors $x_1 \dots x_T \in \mathbb{R}^D \Longrightarrow h_1 \dots h_T \in \mathbb{R}^d$

Parameters: $\{(W_f, b_f), (W_i, b_i), (W_\theta, b_\theta), (W_c, b_c)\}$ · Each $W_* \in \mathbb{R}^{(D+d)\times d}$, $b_* \in \mathbb{R}^d$

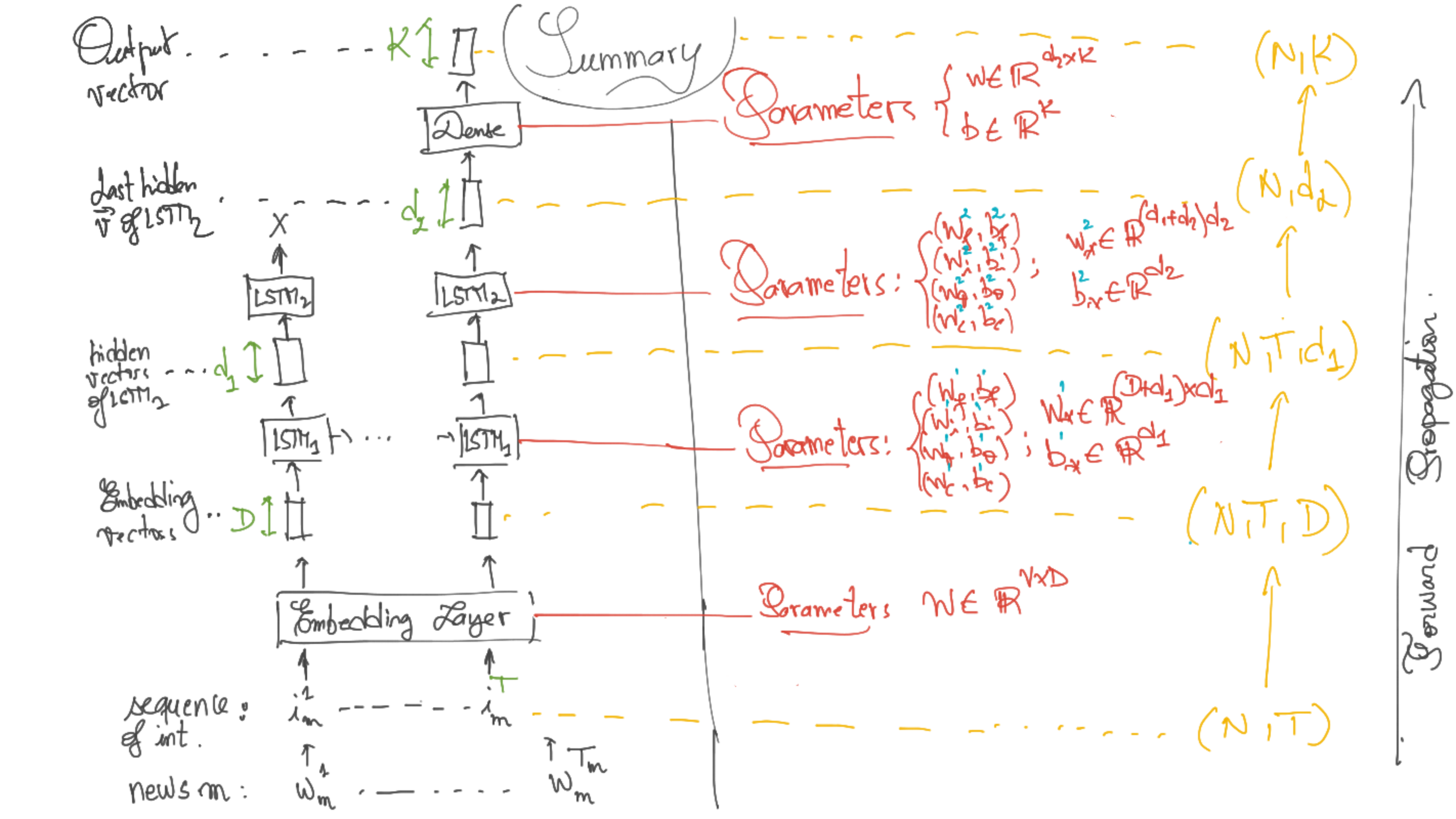
#parameters: $4((D+d)d + d)$

Transformation:

At $(h_{t-1}, C_{t-1}, x_t) \longrightarrow (h_t, C_t)$

$$f_t = \sigma\left(W_f^T [h_{t-1}, x_t] + b_f\right)$$
$$i_t = \sigma\left(W_i^T [h_{t-1}, x_t) + b_i\right)$$
$$\theta_t = \sigma\left(W_\theta^T [h_{t-1}, x_t) + b_\theta\right)$$

$$C_t = C_{t-1} \odot f_t + \tilde{C}_t \odot i_t \quad : \text{update gc}$$
$$\tilde{C}_t = \tanh\left(W_c^T [h_{t-1}, x_t) + b_c\right) : \text{candidate}$$
$$h_t = \theta_t \odot \tanh(C_t)$$


Output vector
K
Summary
Dense
Parameters $\begin{cases} W \in \mathbb{R}^{d_2 \times K} \\ b \in \mathbb{R}^K \end{cases}$
(N, K)
last hidden $\overrightarrow{v}$ of LSTM$_2$
d_2
X
LSTM$_2$
LSTM$_2$
Parameters: $\begin{cases} (W_f^2, b_f^2) \\ (W_i^2, b_i^2) \\ (W_o^2, b_o^2) \\ (W_c^2, b_c^2) \end{cases}$; $W_x^2 \in \mathbb{R}^{(d_1+d_2) d_2}$, $b_x^2 \in \mathbb{R}^{d_2}$
(N, d_2)
hidden vectors of LSTM$_2$
d_1
LSTM$_1$
LSTM$_1$
Parameters: $\begin{cases} (W_f, b_f) \\ (W_i, b_i) \\ (W_o, b_o) \\ (W_c, b_c) \end{cases}$; $W_x \in \mathbb{R}^{(D+d_1) \times d_1}$, $b_x \in \mathbb{R}^{d_1}$
(N, T, d_1)
Embedding vectors
D
Embedding Layer
Parameters $W \in \mathbb{R}^{V \times D}$
(N, T, D)
sequence of int.
news m:
(N, T)
Forward Propagation