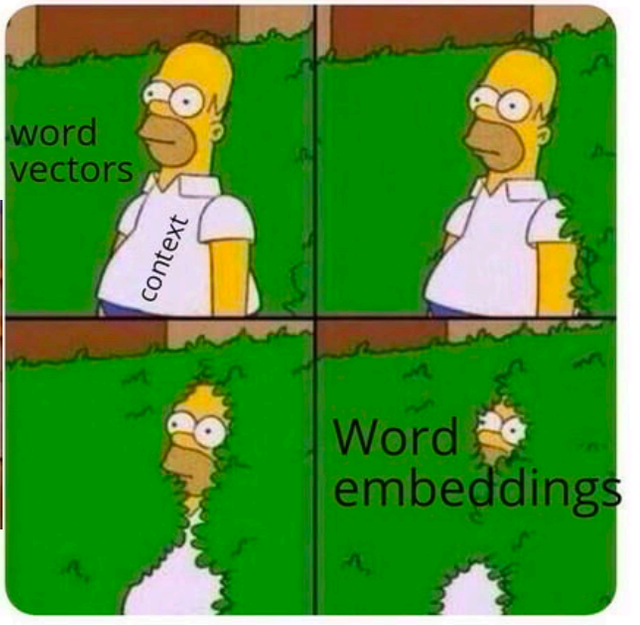
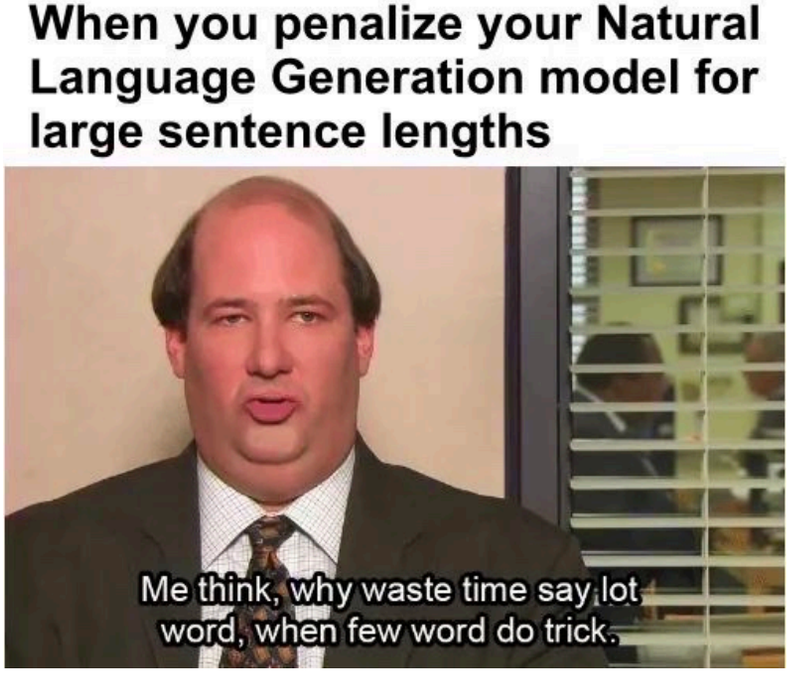
May 5, 2023.

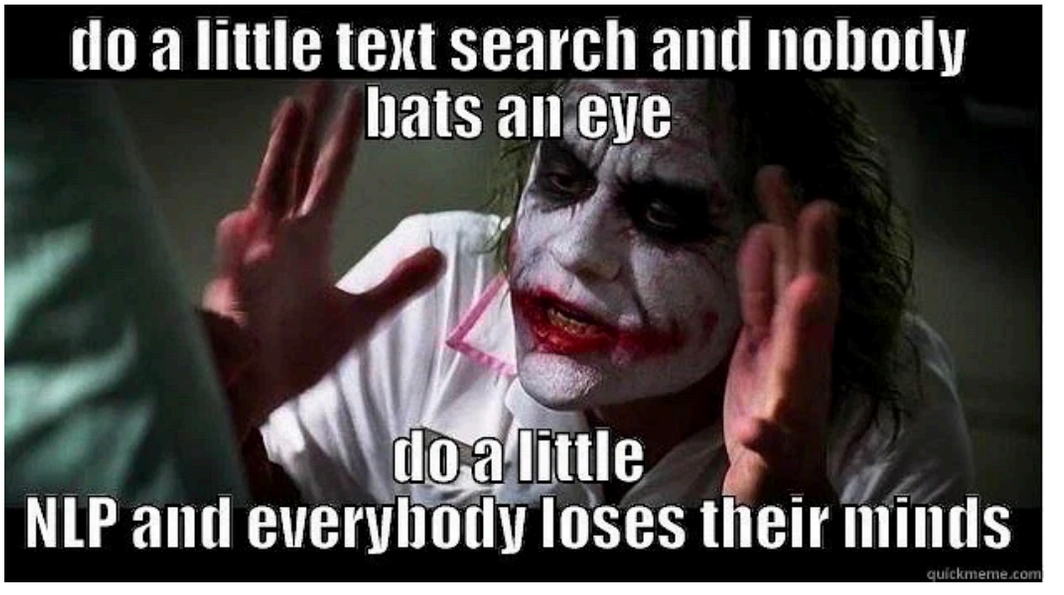
DSML : NLP module.

Long Short-Term Memory (LSTM)

Class starts @ 9:05

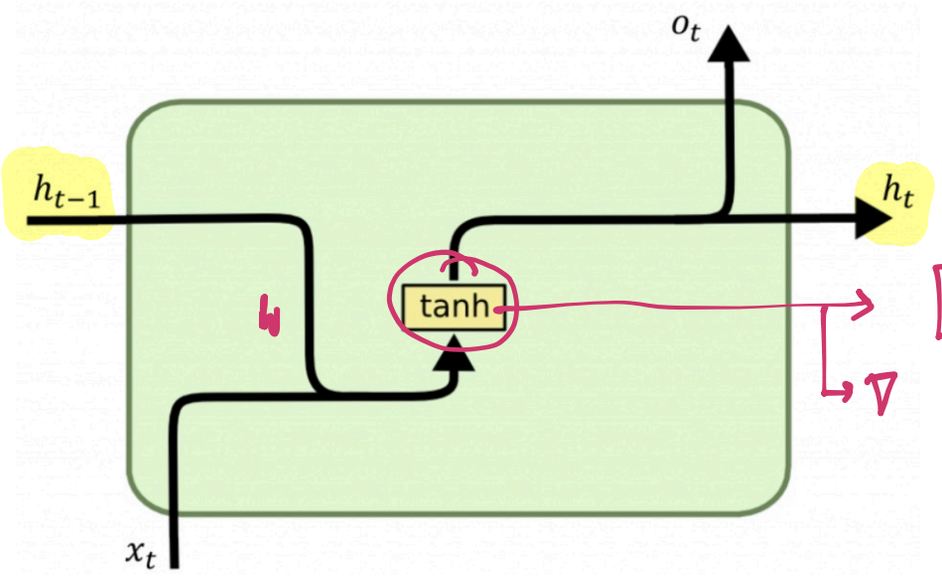






Recap:

ReLU → ∇ [-∞, ∞]



→ [-1 to 1]

→ ∇ [0 to 1.]

* Recurrent Neural Networks (RNNs).

→ What problems do they solve?

* Variable input/output lengths. * Explicit encoding of context - hidden state.

→ Types?

* one-to-one * one-to-many * Many-to-one * Many-to-many

→ Training?

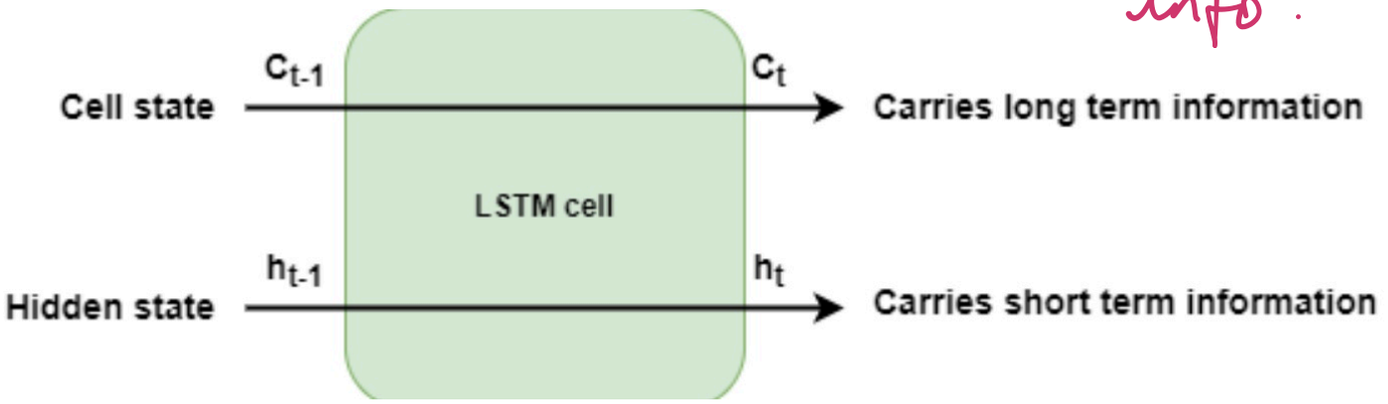
Backpropagation through time (BPTT).

Agenda:

* A new type of Architecture:
  Long Short-term Memory (LSTM).
  → Why? RNN training is difficult because of vanishing / exploding gradients.
* Application: Text Summarization.
  → Encoder - Decoder Architecture.
  → Input, Forget and Output Gates.
  → Gated Recurrent Units.
  → ROUGE score.

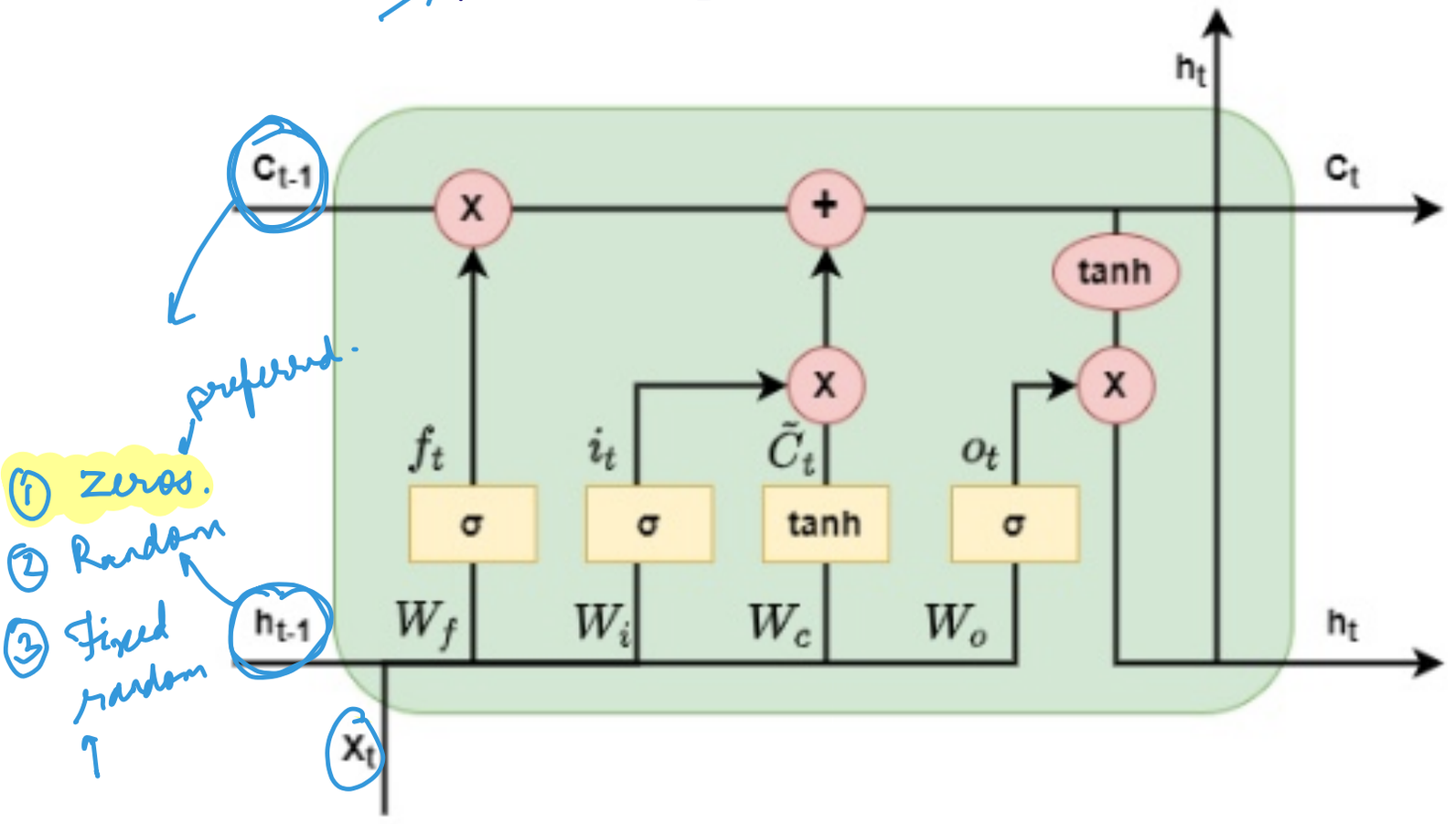
Let's understand the Industry Requirement !!

Conceptual change: Add memory to store long-term info.



Practical Implementation.

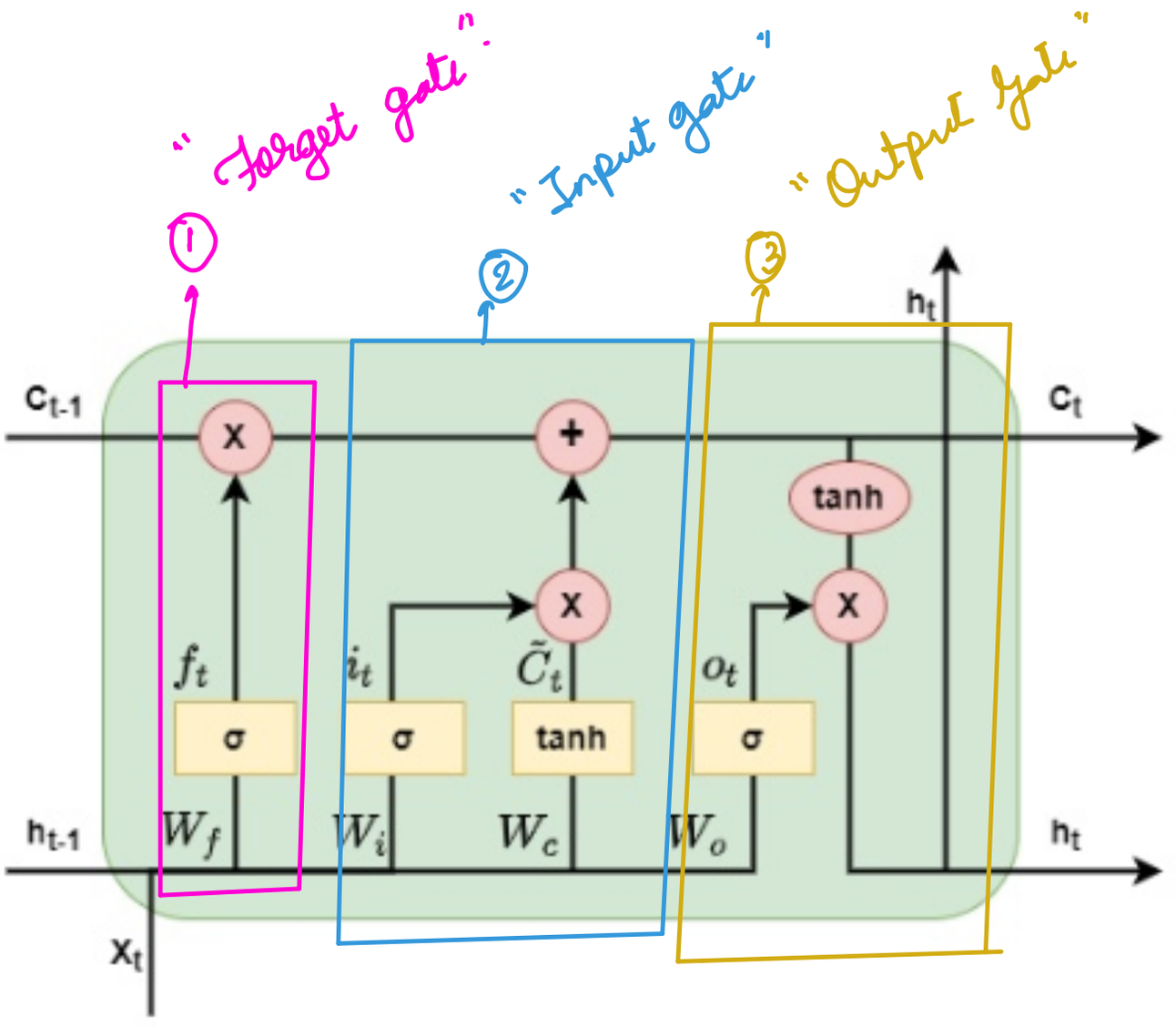
I have a fat cat.
1 2 3 4 5



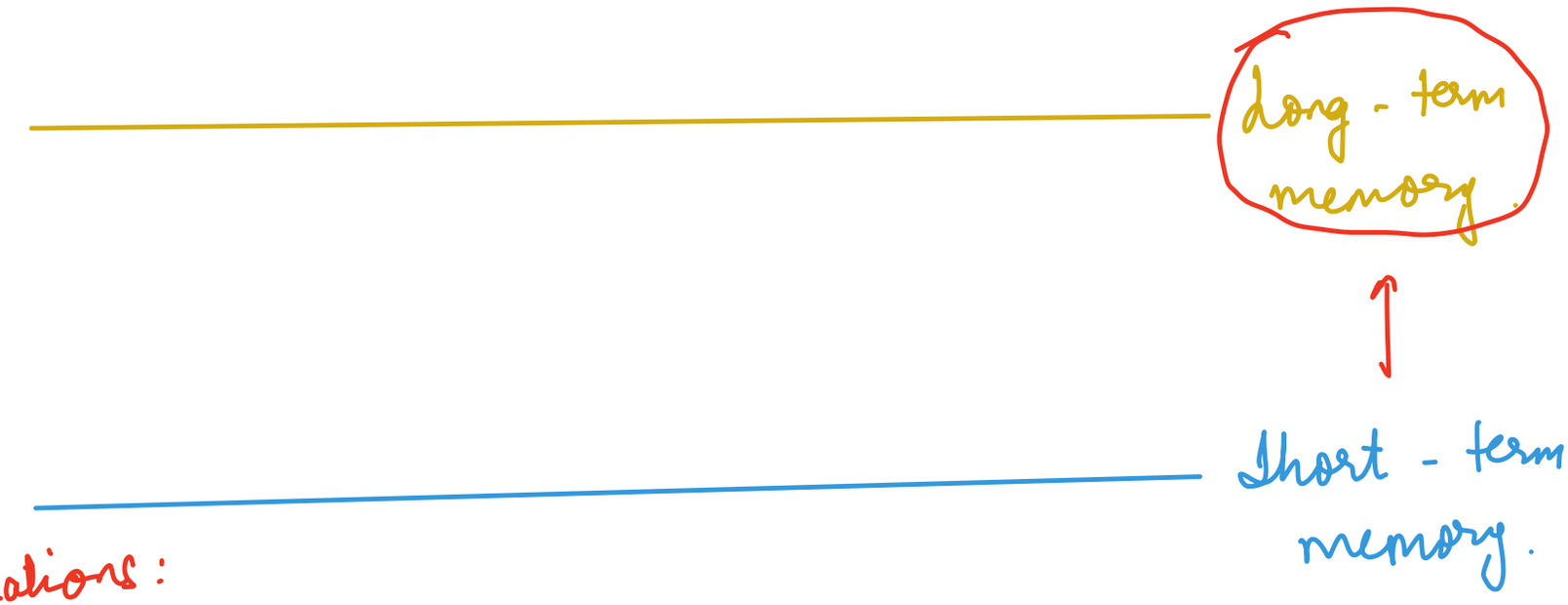
preferred.

① zeros.
② Random
③ Fixed random

Input word at time step t.


" Forget gate "
" Input gate "
" Output gate "
1
2
3
C_{t-1}
C_t
h_t
h_{t-1}
h_t
X_t
X
+
X
X
tanh
tanh
σ
σ
σ
f_t
i_t
$\tilde{C}_t$
o_t
W_f
W_i
W_c
W_o

Conceptual Explanation.



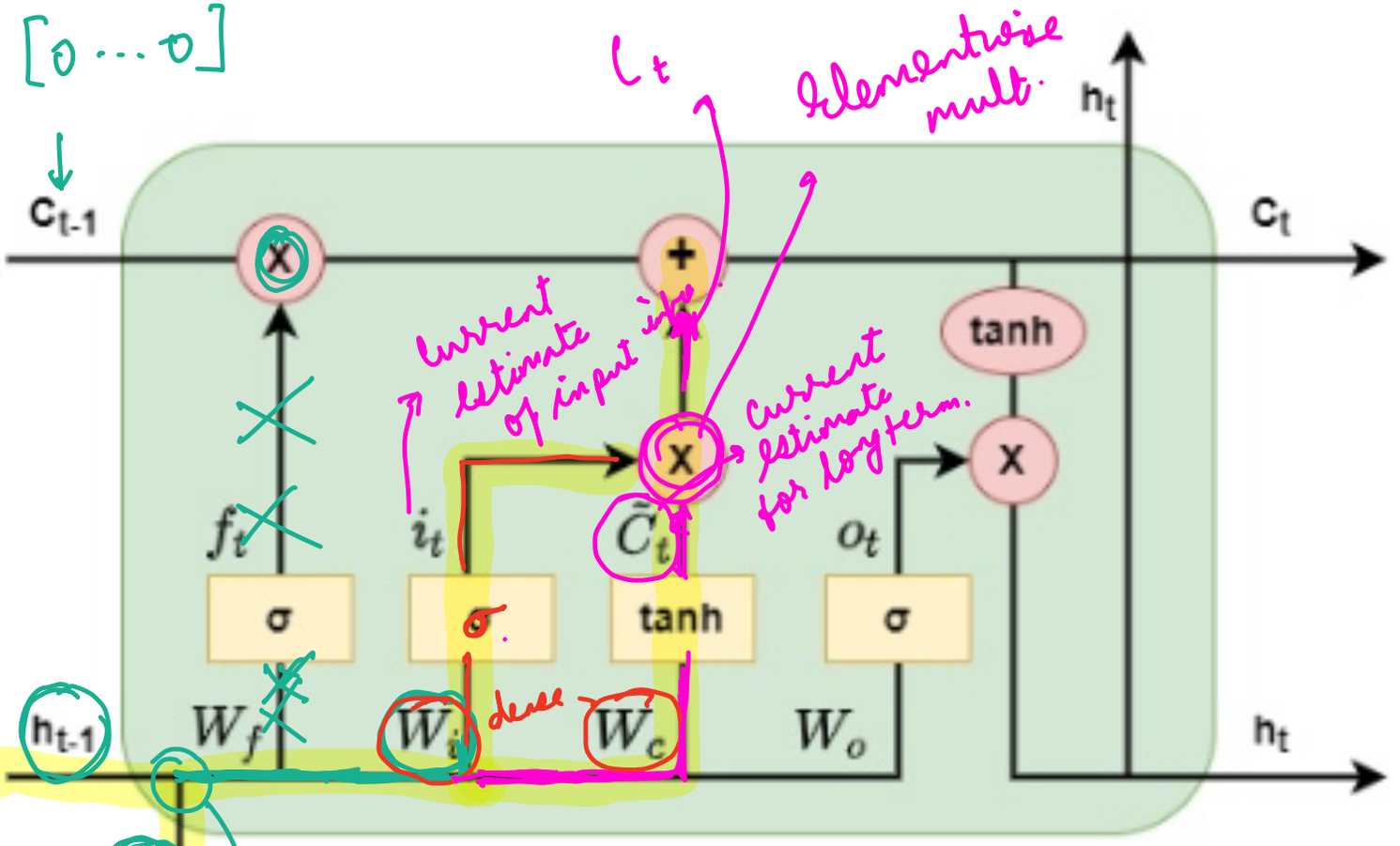
Operations:

(i) **Input operation**: Allow info to flow from short-term to long-term.

(ii) **Forget operation**: Allow unnecessary information to be removed from the long-term.

(iii) **Output operation**: Retrieve info from long term-

Input Gate: How information flows into long-term memory.



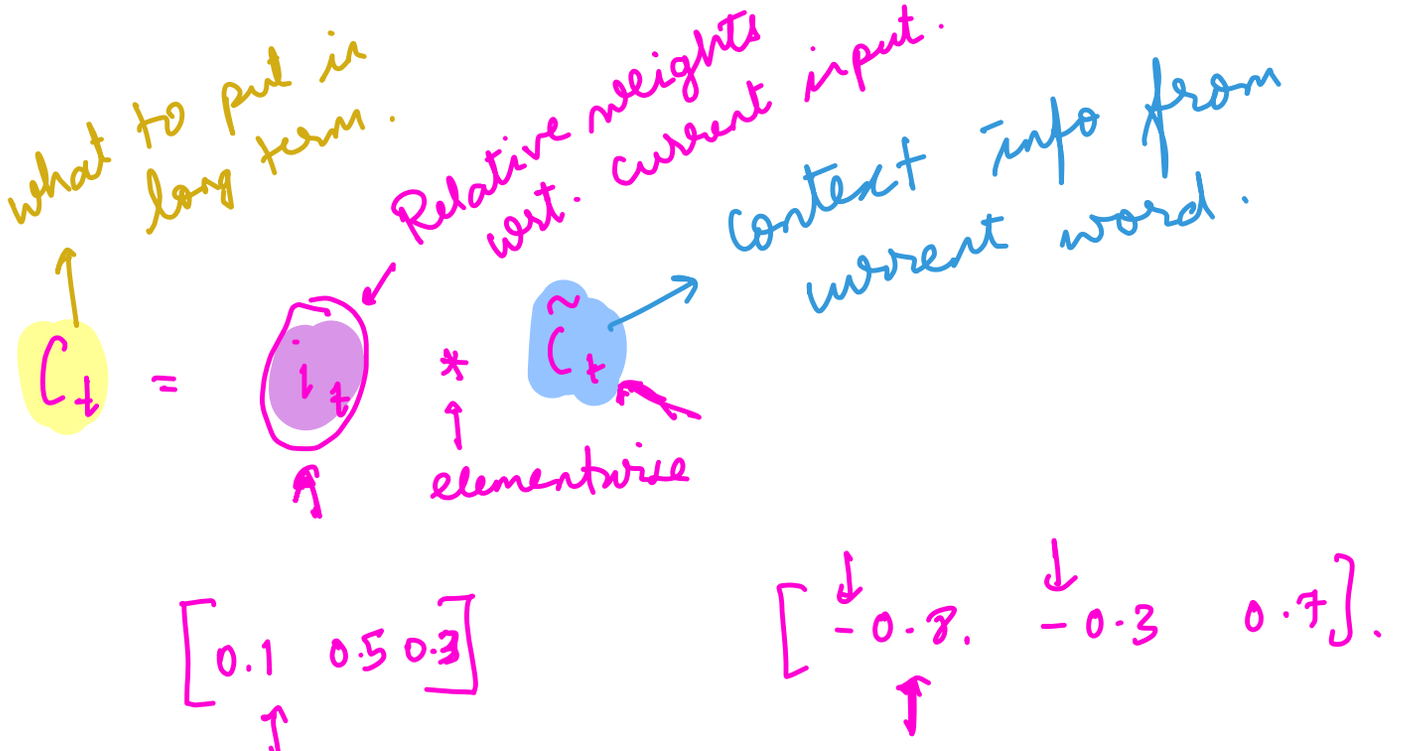
[0, 1]

concatenation. $[x_t, h_{t-1}] \rightarrow 500$.

200 300

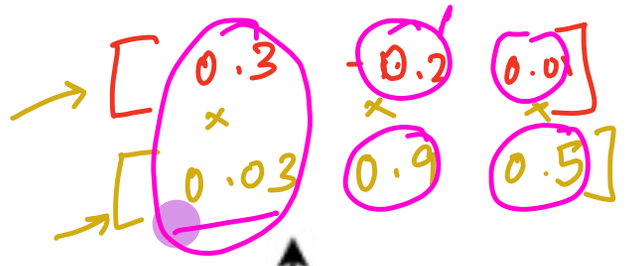
$i_t = \sigma(W_i [x_t, h_{t-1}] + b_i)$

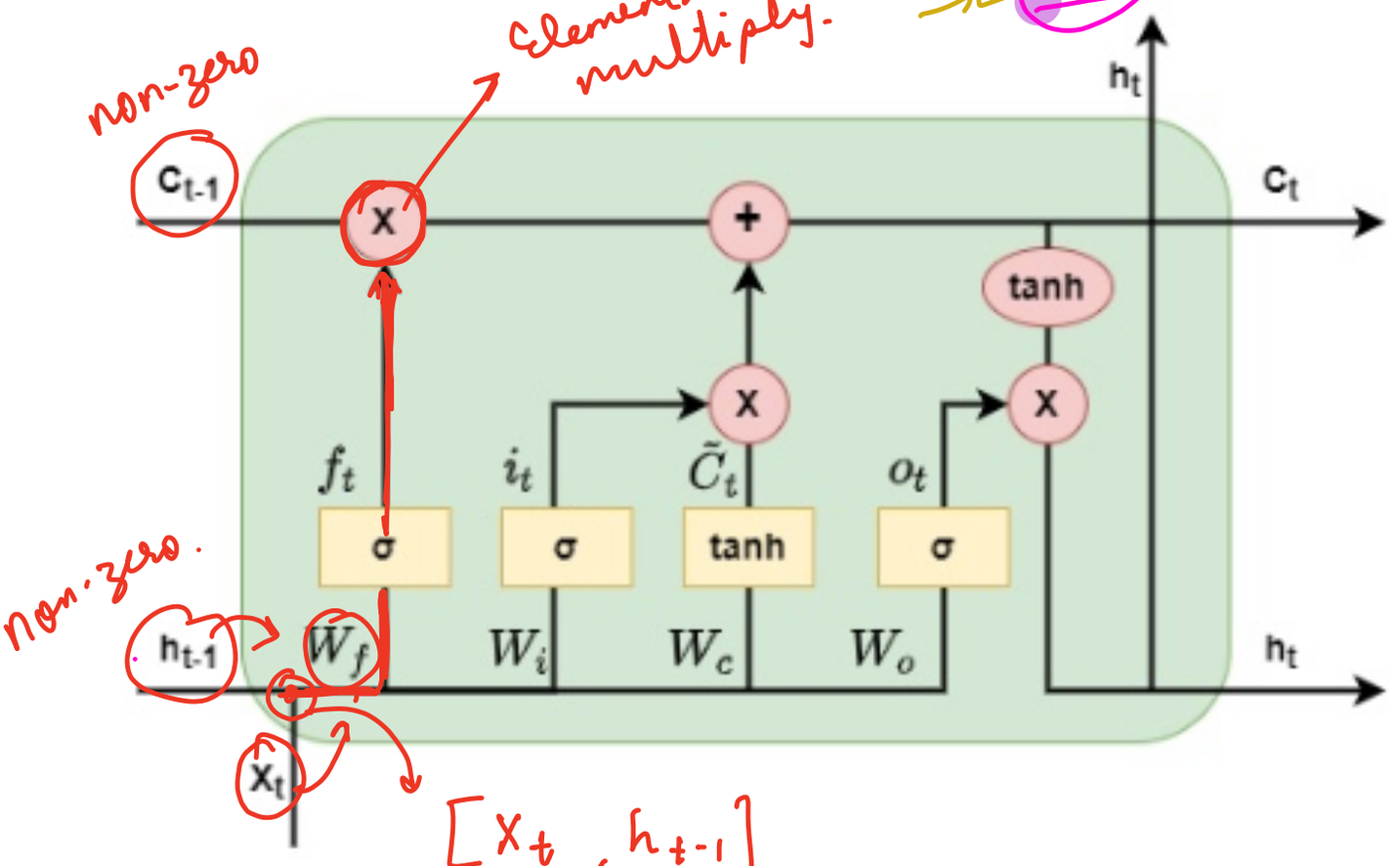
$\rightarrow \tilde{C}_t = \tanh(W_c [x_t, h_{t-1}] + b_c) \rightarrow (-1, 1)$



$$\begin{bmatrix} 0.1 & 0.5 & 0.3 \end{bmatrix} \qquad \begin{bmatrix} -0.8 & -0.3 & 0.7 \end{bmatrix}.$$

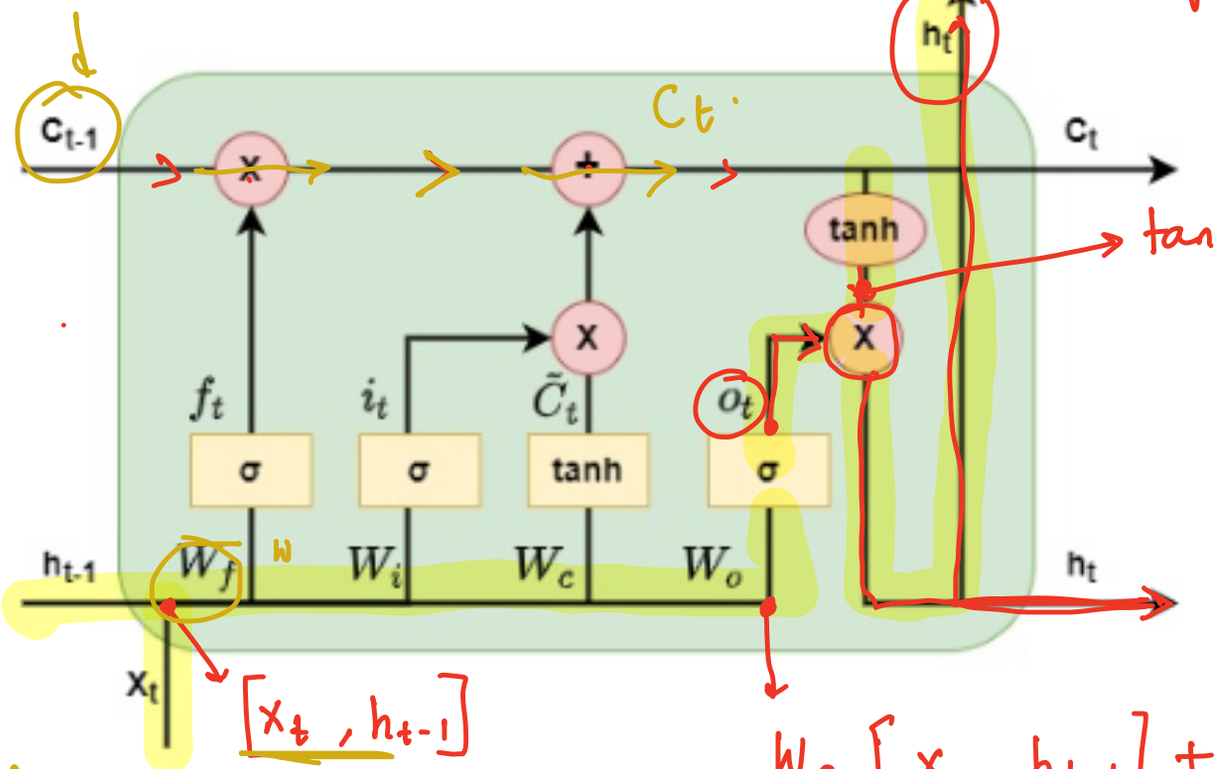
Forget Operation :





$$\sigma\left(W_f\,[x_t, h_{t-1}] + b_f\right)$$

Output Gate:



Output of the LSTM.

C_t

$\tanh(C_t)$

$[x_t, h_{t-1}]$

$W_o [x_t, h_{t-1}] + b_o$

contribution from long-term

$$o_t = \sigma\left(W_o [x_t, h_{t-1}] + b_o\right)$$

contribution from short term.

$$h_t = \tanh(C_t) * o_t$$

↑ elementwise

Hidden State

Which of the following statements is correct about Hidden State ?

HINTS | Complete Solution

You will get full points if and only if you give CORRECT ANSWER in first attempt. All later attempts will get you ZERO score.

- [ ] **It is present in both MLP and RNN**
- [ ] **When we are passing the first input the Hidden states are all generally initialized according to HE initialization.**
- [x] **When we are passing the first input the Hidden states are all typically initialized to zeros**
- [x] **Activation Function typically used in Vanilla RNN is tanh**

Feedback:
Include more detailed explanations.