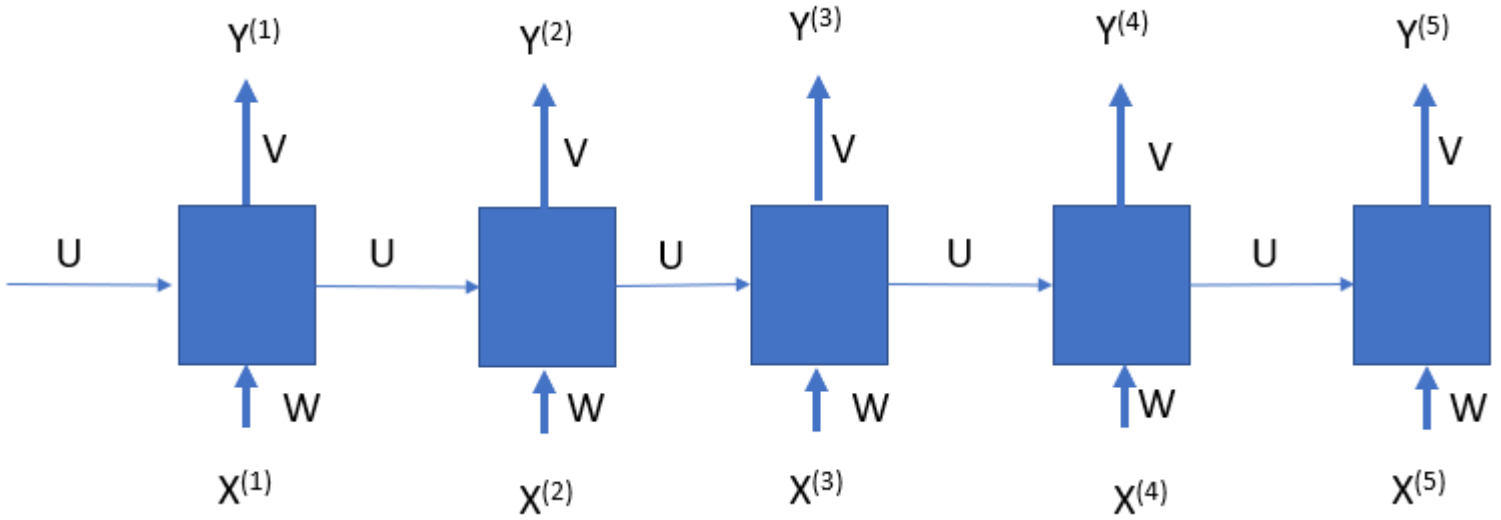
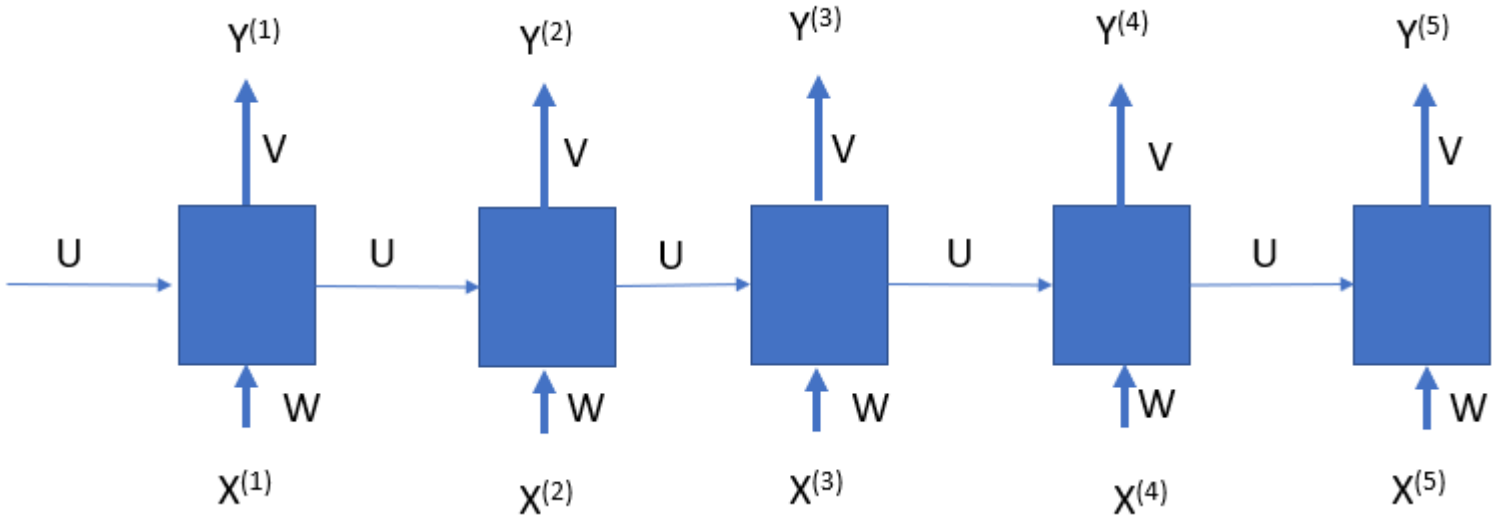
Handling Vanishing Gradient in RNN



$$\overline{h^t}^T = tanh\left(\overline{x^t}^T W + \overline{h^{t-1}}^T U\right) \qquad \overline{y^t}^T = softmax\left(\overline{h^t}^T V\right) \qquad E = \sum_t E^t = \sum_t -\acute{y}^t log(y^t)$$

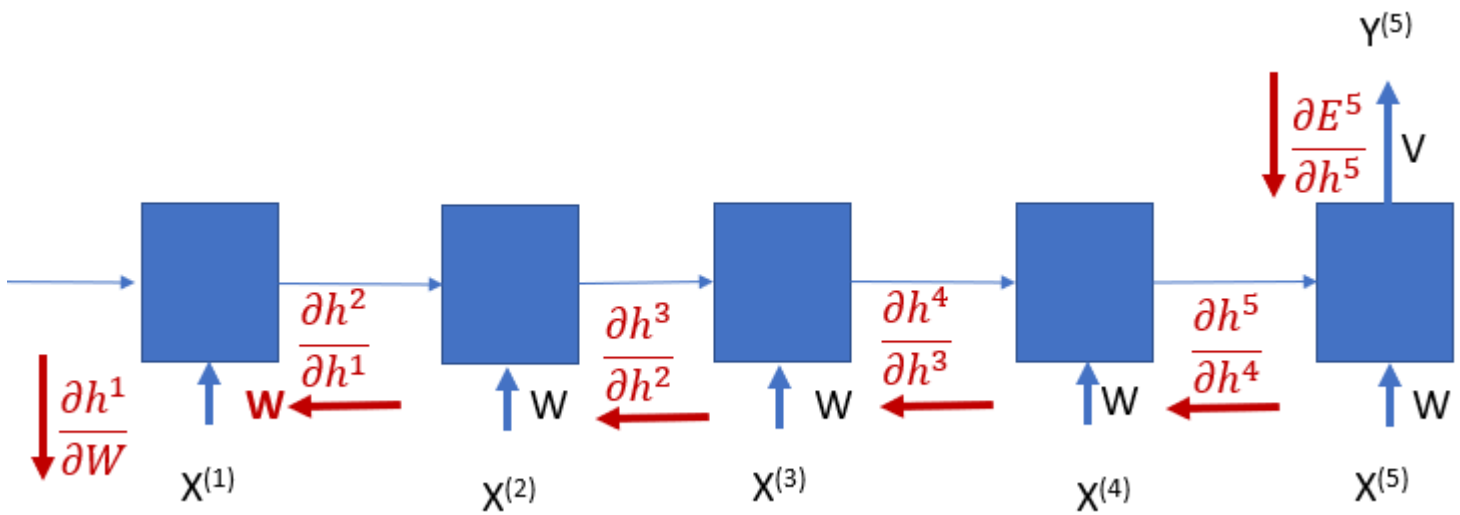
Handling Vanishing Gradient in RNN



$$\frac{\partial E^t}{\partial U} = \frac{\partial E^t}{\partial y^t} \sum_{i=1}^{t} \frac{\partial y^t}{\partial h^i} \frac{\partial h^i}{\partial U}$$

$$\frac{\partial E^t}{\partial W} = \frac{\partial E^t}{\partial y^t} \sum_{i=1}^{t} \frac{\partial y^t}{\partial h^i} \frac{\partial h^i}{\partial W}$$

Handling Vanishing Gradient in RNN



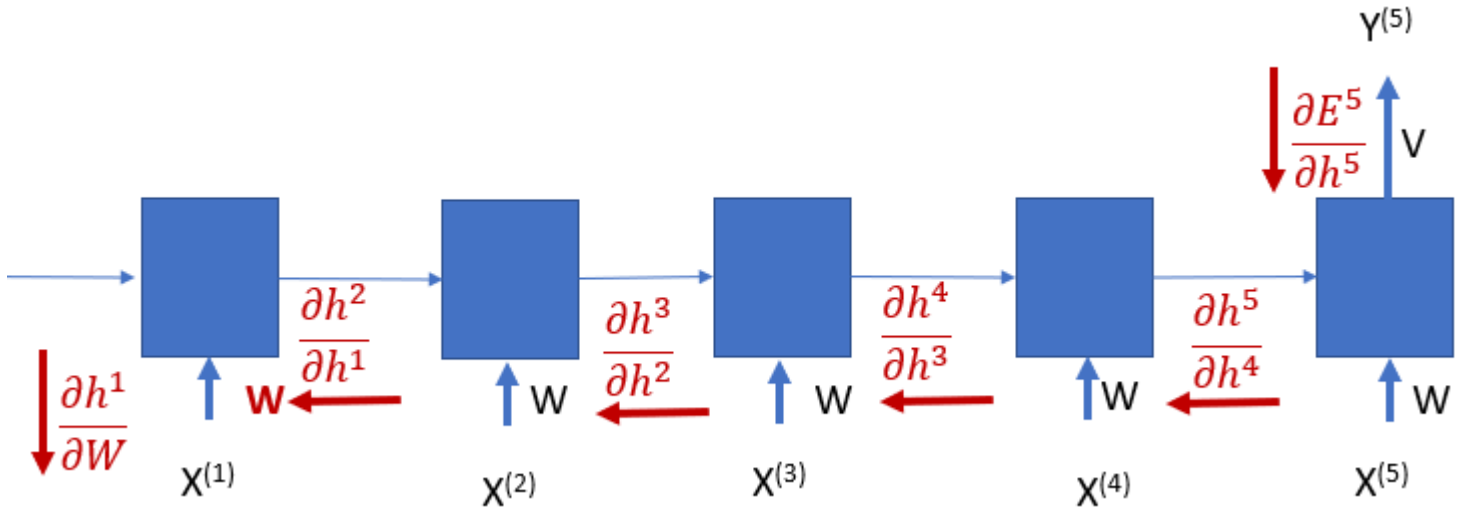
$$\frac{\partial E^t}{\partial W} = \frac{\partial E^t}{\partial y^t}\frac{\partial y^t}{\partial h^1}\frac{\partial h^1}{\partial W}$$

$$\frac{\partial E^5}{\partial y^5}\frac{\partial y^5}{\partial h^5}\frac{\partial h^5}{\partial h^4}\frac{\partial h^4}{\partial h^3}\frac{\partial h^3}{\partial h^2}\frac{\partial h^2}{\partial h^1}\frac{\partial h^1}{\partial W}$$

Handling Vanishing Gradient in RNN

RNN has issues when it experiences **Long Term Dependencies** because of the chain rule in its **hidden state**.

It means, while performing **Backpropagation,** outputs of ***every previous steps*** are considered.



$$\frac{\partial E^5}{\partial y^5}\frac{\partial y^5}{\partial h^5}\frac{\partial h^5}{\partial h^4}\frac{\partial h^4}{\partial h^3}\frac{\partial h^3}{\partial h^2}\frac{\partial h^2}{\partial h^1}\frac{\partial h^1}{\partial W}$$

Two popular methods for handling Long Term Dependency:

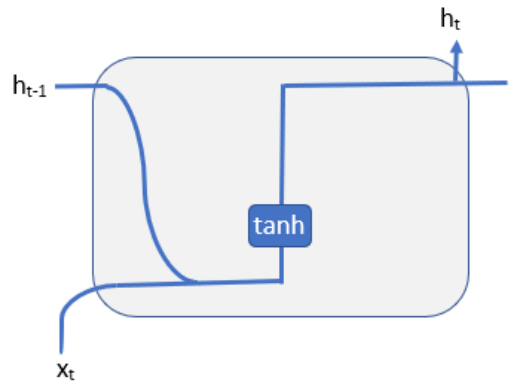
- Long Short Term Memory (LSTM)
- Gated Recurrent Unit (GRU)

Long Short Term Memory (LSTM)

Long Short Term Memory (LSTM)

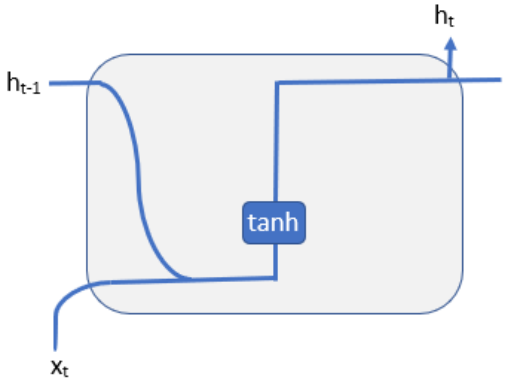
LSTM addresses the problem of long term dependency by introducing **cell state**, in addition to the existing hidden state of RNN.

Long Short Term Memory (LSTM)

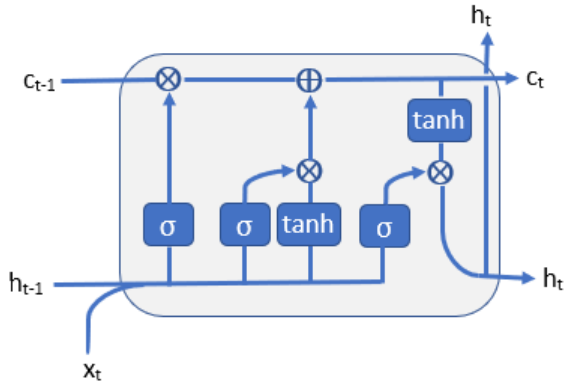


RNN Unit

Long Short Term Memory (LSTM)

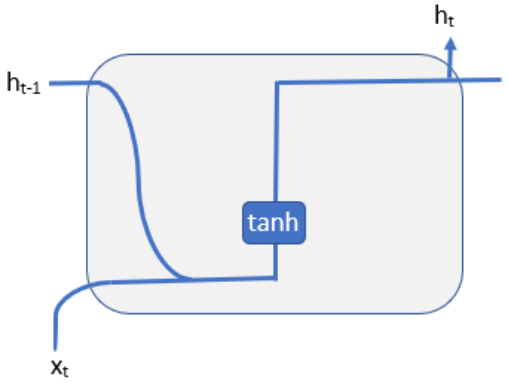


RNN Unit

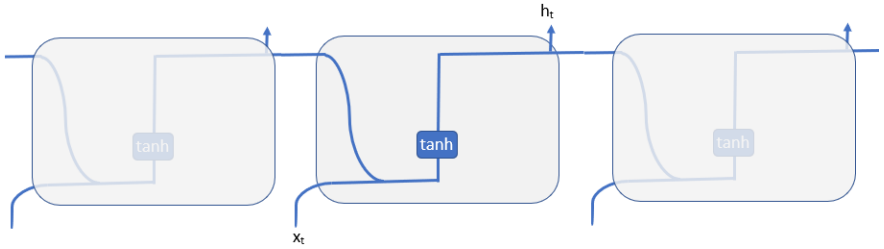


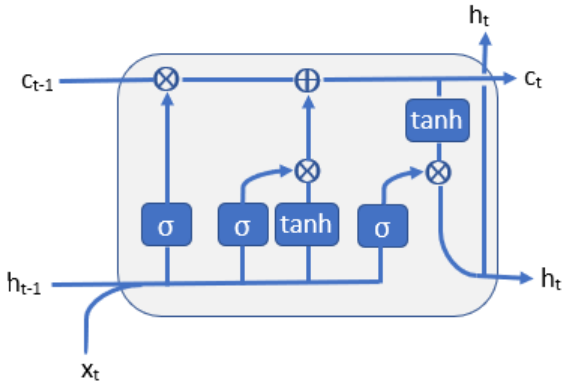
LSTM Unit

Long Short Term Memory (LSTM)

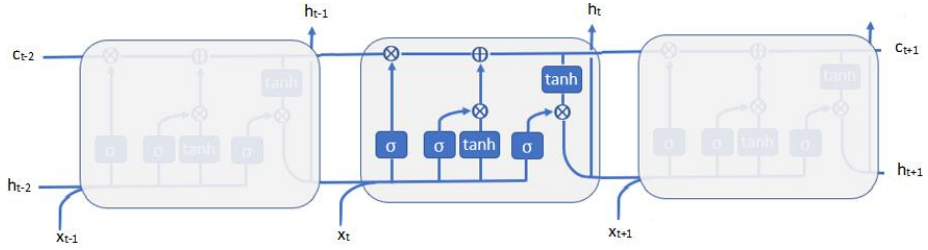


RNN Unit

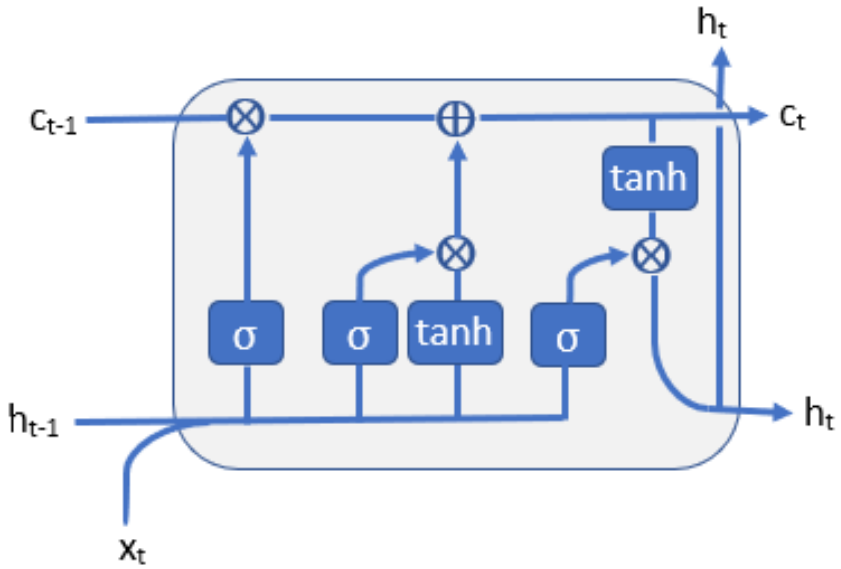




LSTM Unit



Long Short Term Memory (LSTM)

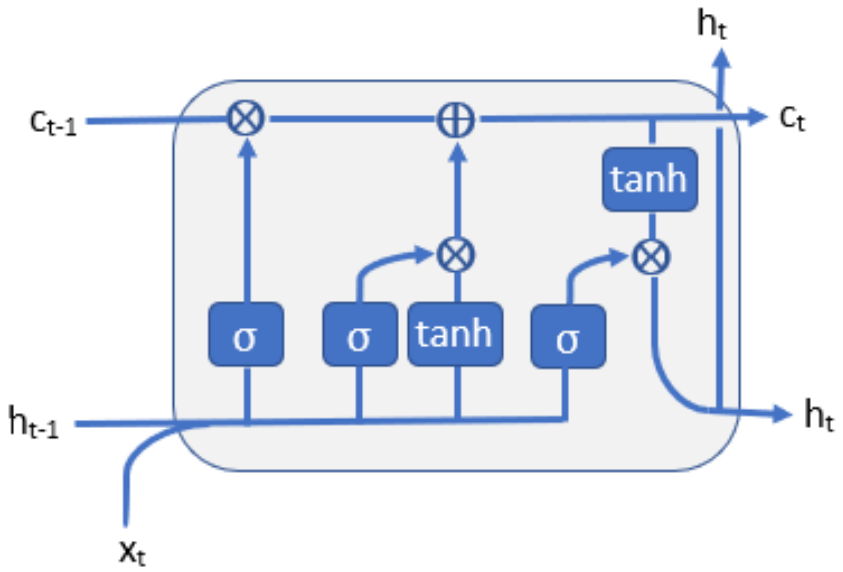


Unlike RNN, LSTM has Two States

Hidden State & Cell State

Cell State - Long Term Memory

Cell State



Input at time t

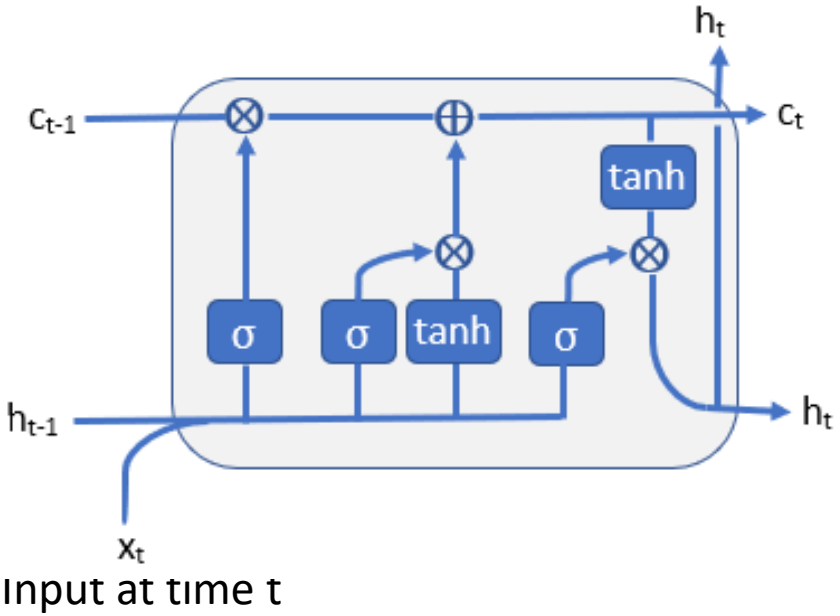
The cell state is a *global* or *aggregate* memory over all time-steps.

What is useful in the present input will be stored.

What is useful from the cell state for the present input will be used.

It is referred to as **Long Term Memory**

Hidden State: Short Term Memory

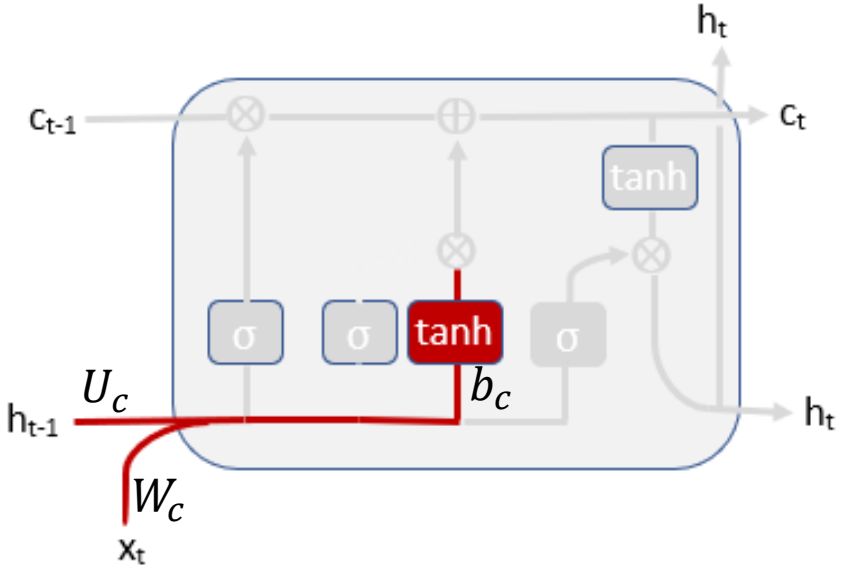


It merely encodes the previous inputs.

The hidden state concerns most recent time-steps.

Internal Information (output from hidden layer)

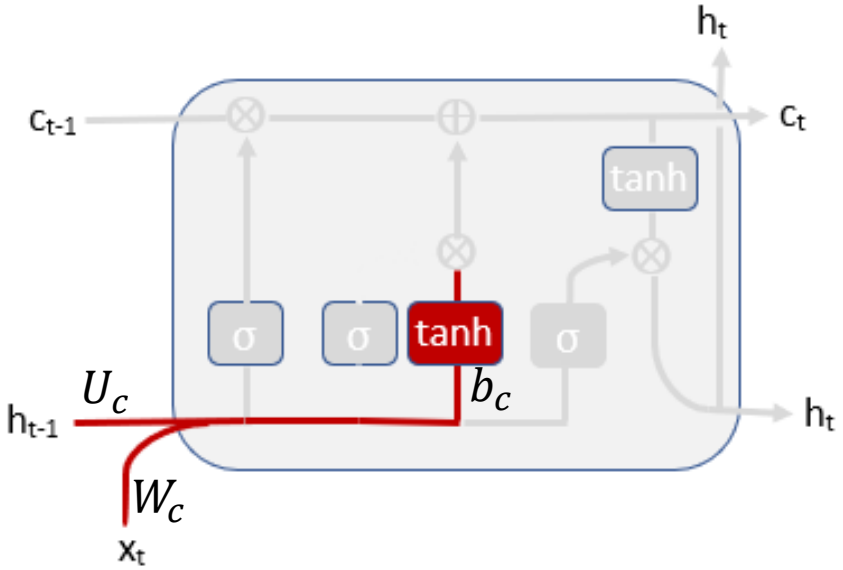
It takes the present input and hidden state of the previous time stamp, and then apply perceptron and activation function



$$\widetilde{c_t} = tanh(W_c \cdot x_t + U_c \cdot h_{t-1} + b_c)$$

Internal Information (output from hidden layer)

It takes the present input and hidden state of the previous time stamp, and then apply perceptron and activation function



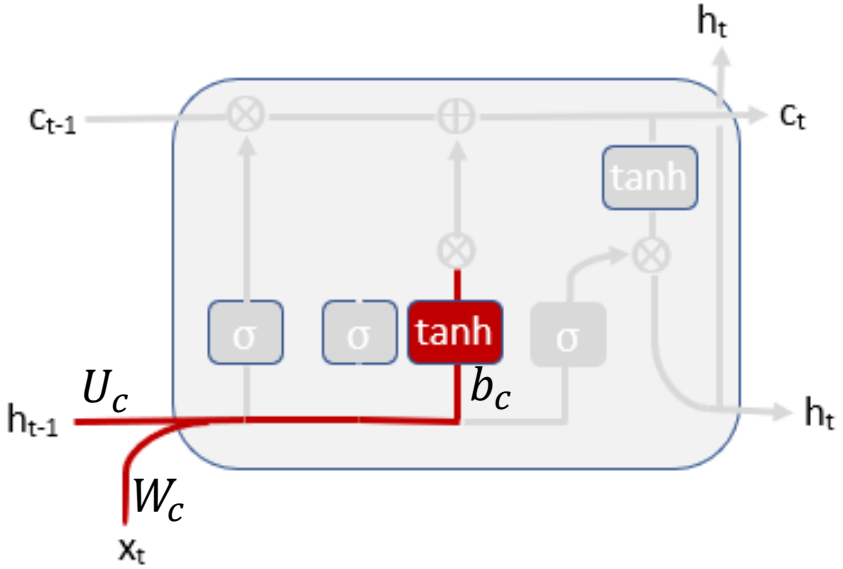
$$\widetilde{c_t} = tanh(W_c \cdot x_t + U_c \cdot h_{t-1} + b_c)$$

Weight matrices

bias

Internal Information (output from hidden layer)

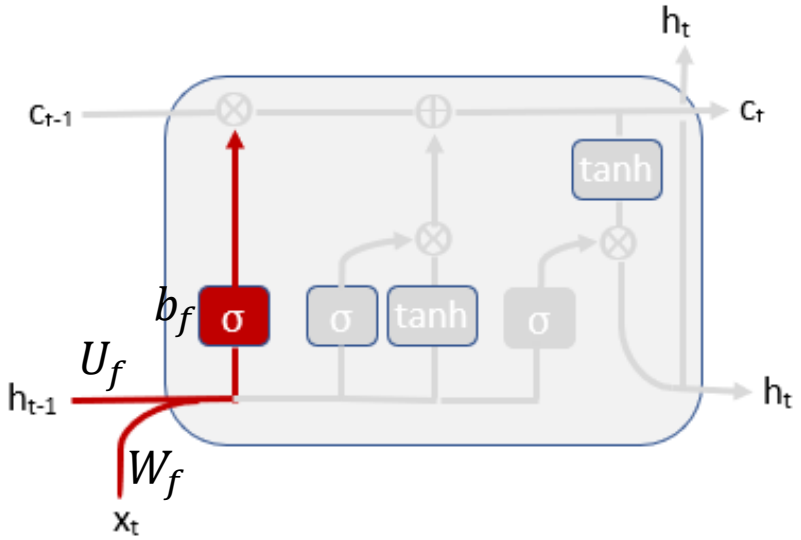
It takes the present input and hidden state of the previous time stamp, and then apply perceptron and activation function



$$\widetilde{c_t} = tanh(W_c \cdot x_t + U_c \cdot h_{t-1} + b_c)$$

Forget Gate

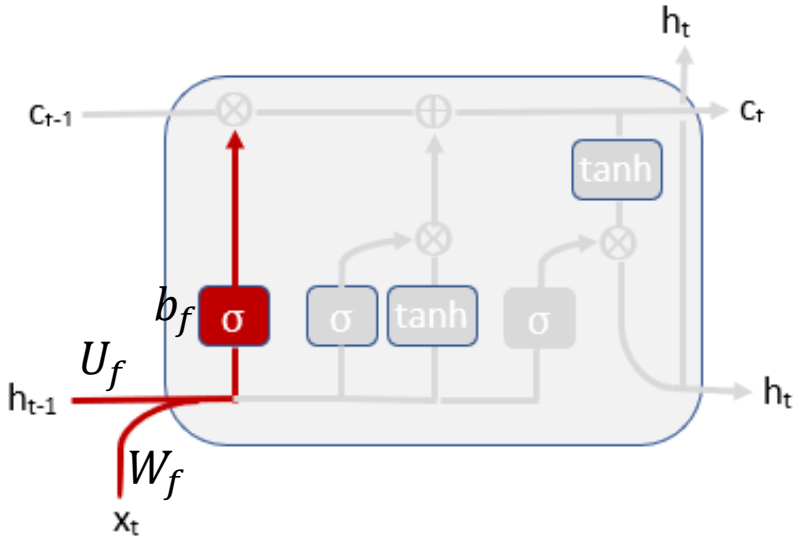
What information from the **Cell State (c_{t-1})**, we are going to throw away based on the present input and previous hidden state.



$$f_t = \sigma\left(W_f \cdot x_t + U_f \cdot h_{t-1} + b_f\right)$$

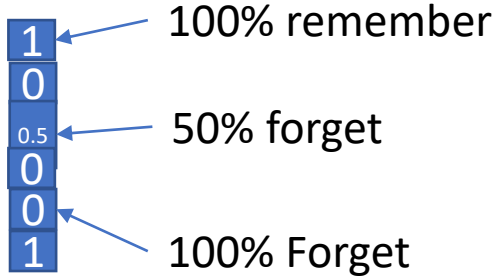
Forget Gate

What information from the **Cell State**, we are going to throw away based on the present input and previous hidden state.



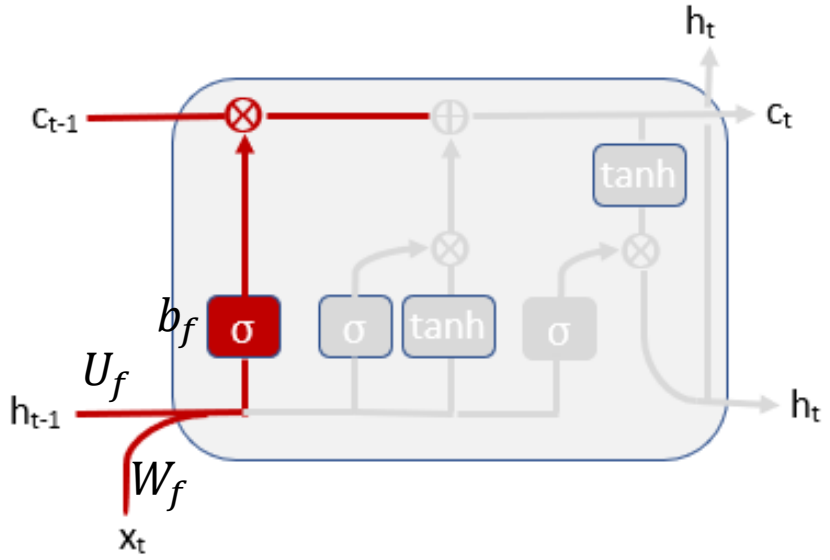
$$f_t = \sigma\left(W_f \cdot x_t + U_f \cdot h_{t-1} + b_f\right)$$

If f_t

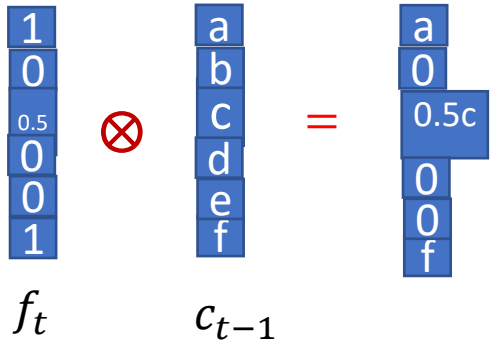


Forget Gate

What information from the **Cell State**, we are going to throw away based on the present input and previous hidden state.

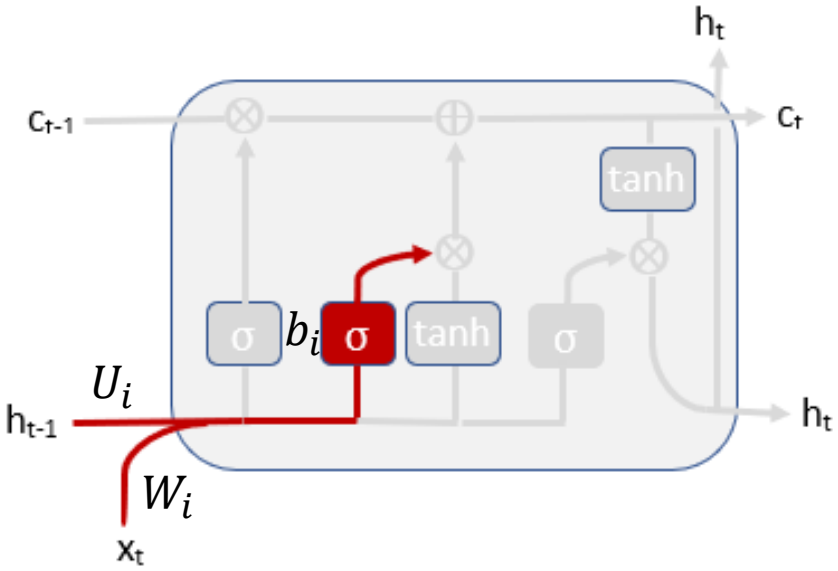


$$f_t = \sigma\left(W_f \cdot x_t + U_f \cdot h_{t-1} + b_f\right)$$



Input Gate

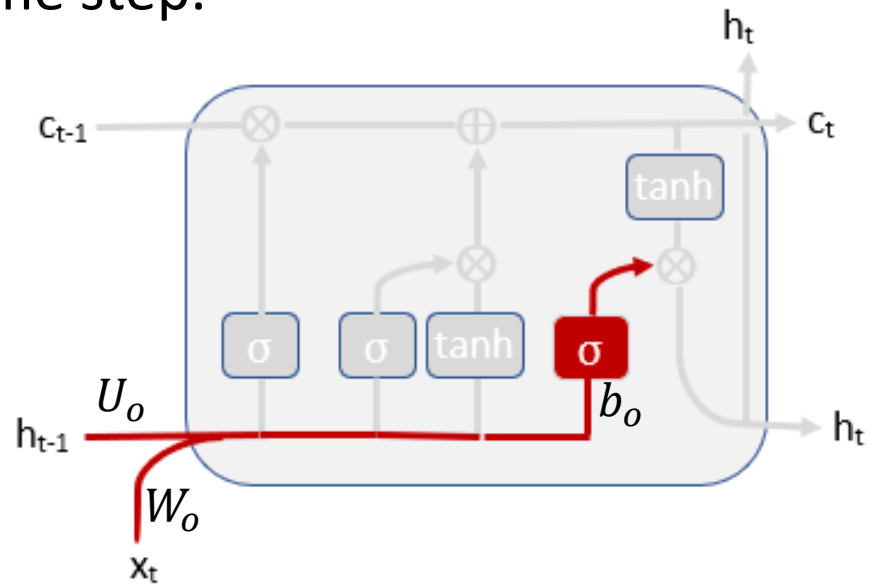
What information from **internal state** ($\widetilde{c_t}$), we are going to keep.



$$i_t = \sigma(W_i \cdot x_t + U_i \cdot h_{t-1} + b_i)$$

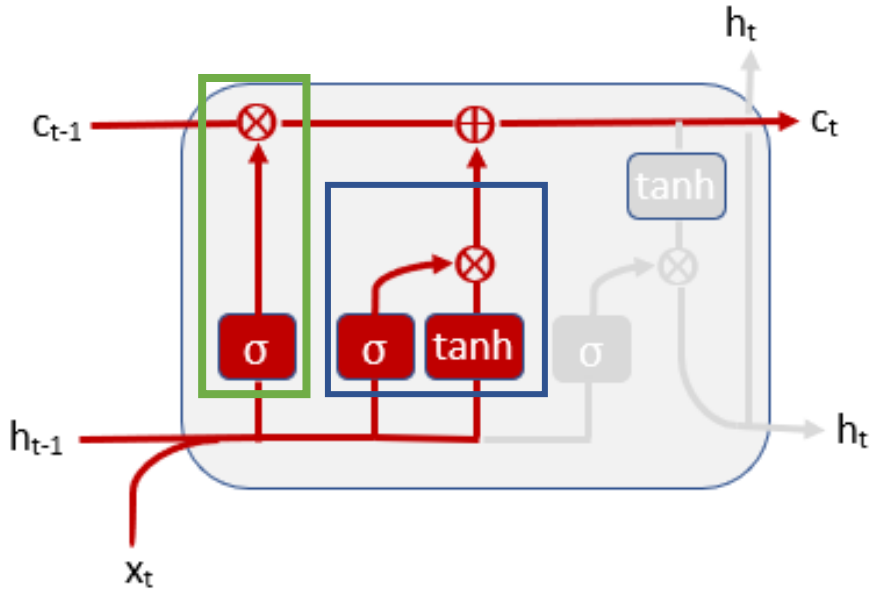
Output Gate

What information from **new cell state (c_t)** should carry forward to the next time step.



$$o_t = \sigma(W_o \cdot x_t + U_o \cdot h_{t-1} + b_o)$$

New Cell State c_t



$$f_t = \sigma(W_f \cdot x_t + U_f \cdot h_{t-1} + b_i)$$

$$i_t = \sigma(W_i \cdot x_t + U_i \cdot h_{t-1} + b_i)$$

$$\tilde{c_t} = tanh(W_c \cdot x_t + U_c \cdot h_{t-1} + b_c)$$

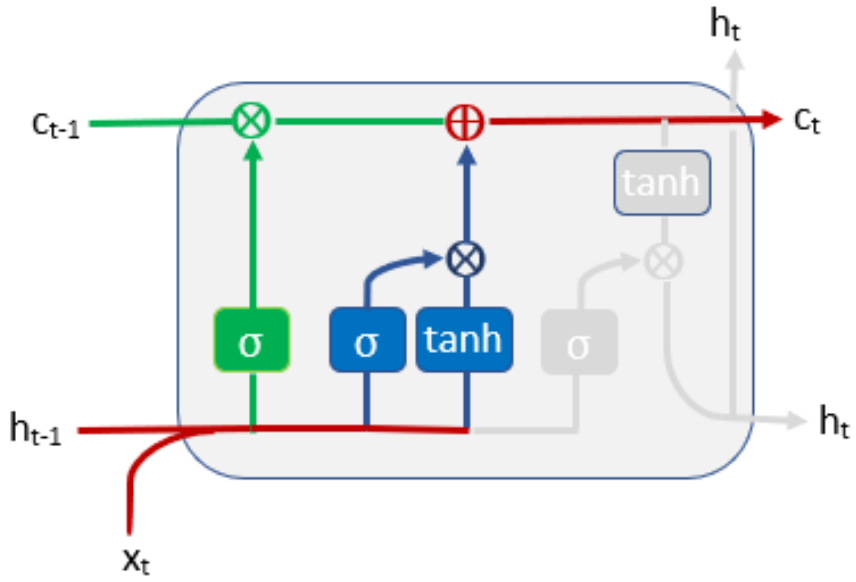
New Cell State

$$c_t = c_{t-1} \otimes f_t \oplus i_t \otimes \tilde{c_t}$$

Consider the useful information from the cell state

Consider the useful information from input

New Cell State c_t



$$f_t = \sigma(W_f \cdot x_t + U_f \cdot h_{t-1} + b_i)$$

$$i_t = \sigma(W_i \cdot x_t + U_i \cdot h_{t-1} + b_i)$$

$$\tilde{c_t} = tanh(W_c \cdot x_t + U_c \cdot h_{t-1} + b_c)$$

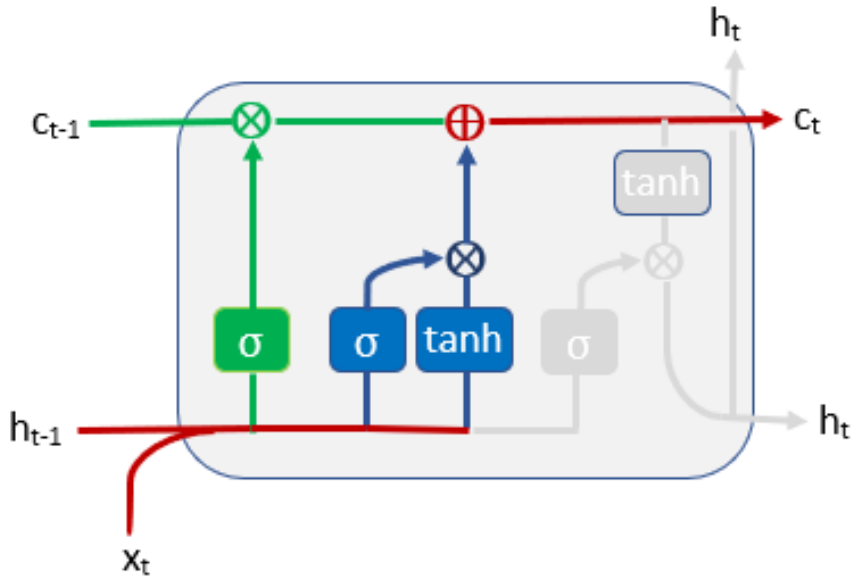
New Cell State

$$c_t = c_{t-1} \otimes f_t \oplus i_t \otimes \tilde{c_t}$$

Consider the useful information from the cell state

Consider the useful information from input

New Cell State c_t



$$f_t = \sigma(W_f \cdot x_t + U_f \cdot h_{t-1} + b_i)$$

$$i_t = \sigma(W_i \cdot x_t + U_i \cdot h_{t-1} + b_i)$$

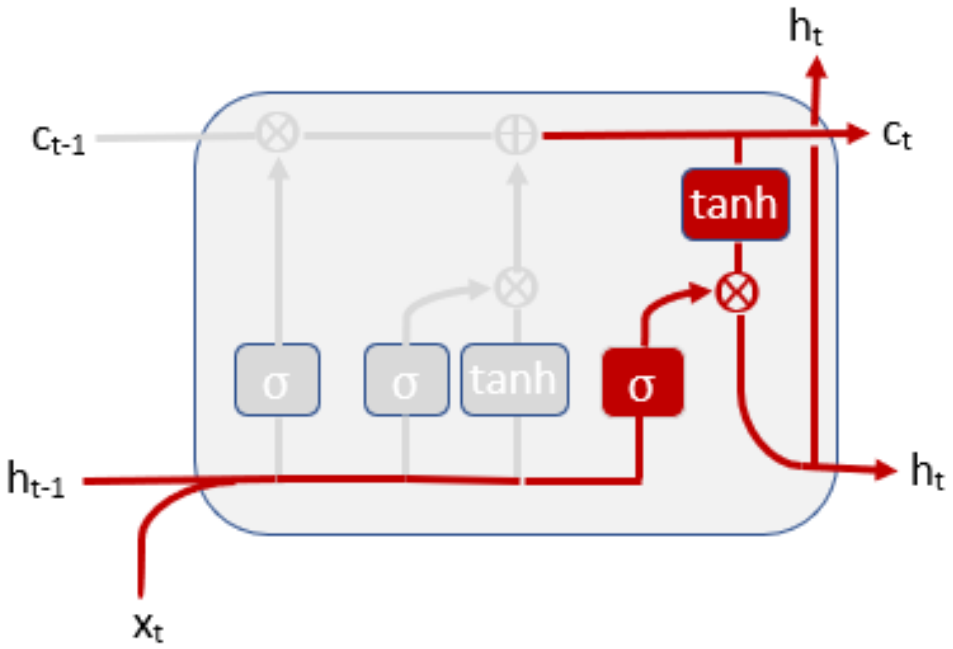
$$\tilde{c_t} = tanh(W_c \cdot x_t + U_c \cdot h_{t-1} + b_c)$$

New Cell State

$$c_t = c_{t-1} \otimes f_t \oplus i_t \otimes \tilde{c_t}$$

Useful information from cell state (c_{t-1}) and inputs (x_t, h_{t-1}) become the new cell state

New Hidden State ($\boldsymbol{h_t}$)



$$o_t = \sigma(W_o \cdot x_t + U_o \cdot h_{t-1} + b_o)$$

$$c_t = c_{t-1} \otimes f_t \oplus i_t \otimes \tilde{c}_t$$

New Hidden State

$$h_t = tanh(c_t) \otimes o_t$$

Useful information from new cell state (c_t) becomes the new hidden state h_t

Summary

Two States

- Cell State $c_t = c_{t-1} \otimes f_t \oplus i_t \otimes \tilde{c_t}$
- Hidden State $h_t = tanh(c_t) \otimes o_t$

Three Gates

- Forget Gate $f_t = \sigma(W_f \cdot x_t + U_f \cdot h_{t-1} + b_i)$
- Input Gate $i_t = \sigma(W_i \cdot x_t + U_i \cdot h_{t-1} + b_i)$
- Output Gate $o_t = \sigma(W_o \cdot x_t + U_o \cdot h_{t-1} + b_o)$