AMRITA VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY

Introduction to Deep Learning

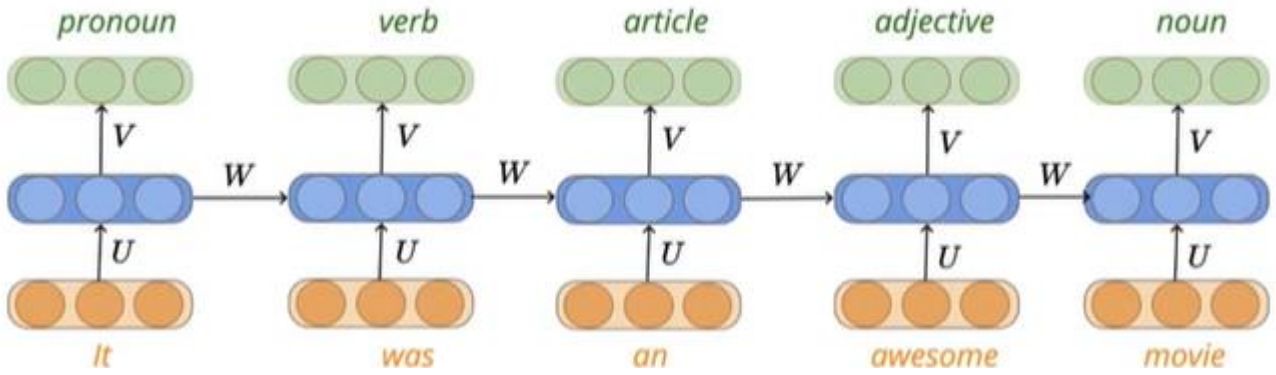
Sequential Models

Long Short-Term Memory Cells(LSTM)

RNN not dealing with longer sequence

(How the state record information when the sequence is very long)

RNN- Recap



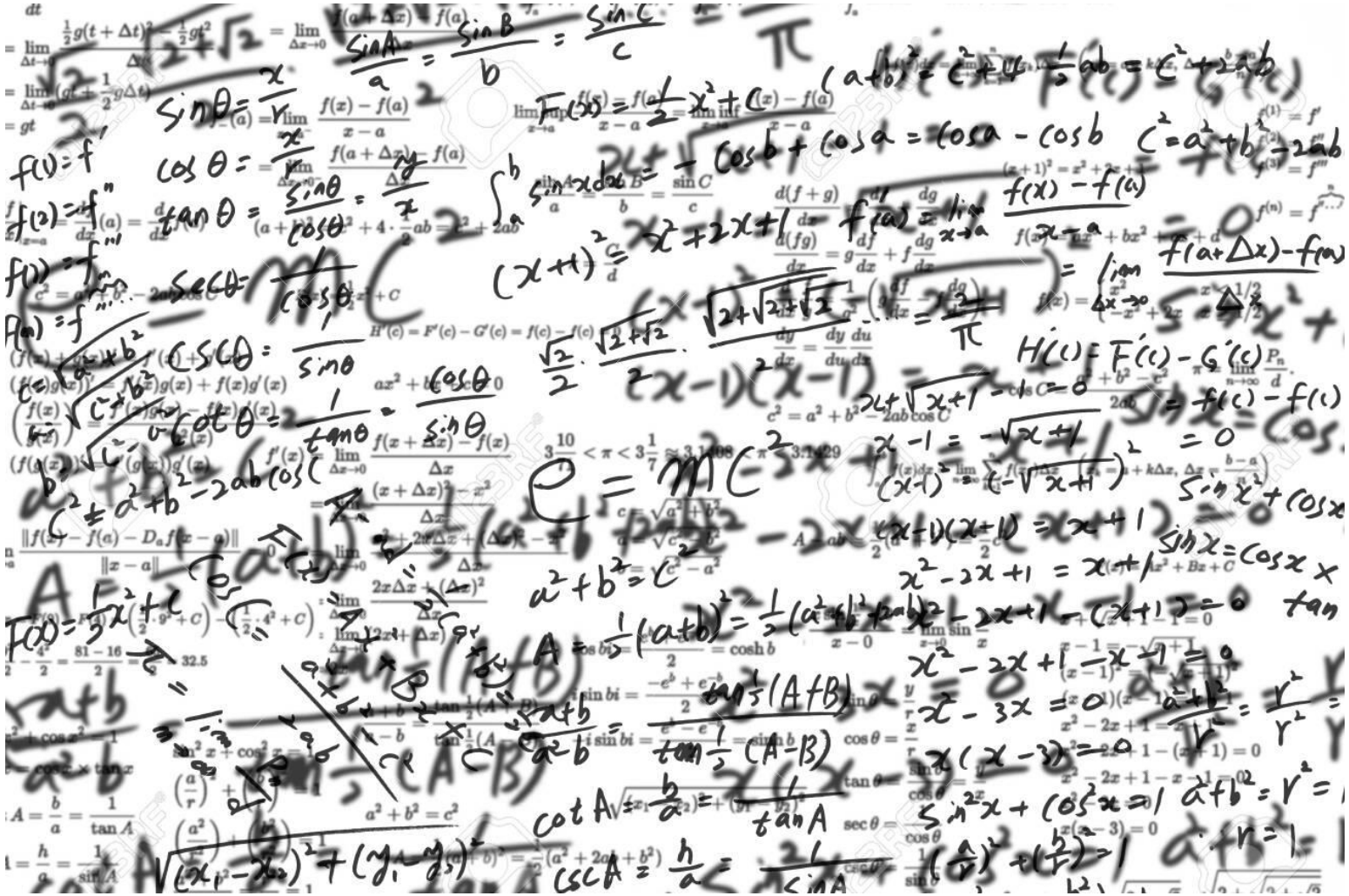
RNN: Exploding and vanishing gradient problem occurs!!
Not suitable for longer sequence

- At each new timestep the old information gets morphed by the current input
- One could imagine that after t steps the information stored at time step t − k (for some k < t) gets completely morphed
- Even during backpropagation the information does not flow well

$$s_i = \sigma(Ux_i + Ws_{i-1} + b)$$

$$y_i = O(Vs_i + c)$$

White board Analogy (over time white board become so messy and u cant make out anything)



White board Analogy

$a = 1 \quad b = 3 \quad c = 5 \quad d = 11$

Compute $ac(bd + a) + ad$

1. ac
2. bd
3. $bd + a$
4. $ac(bd + a)$
5. ad
6. $ac(bd + a) + ad$

$$ac = 5$$

$$bd = 33$$

$$bd + a = 34$$

Strategy

- Selectively write on the board
- Selectively read the already written content
- Selectively forget (erase) some content

White board Analogy

$a = 1 \quad b = 3 \quad c = 5 \quad d = 11$

Compute $ac(bd + a) + ad$

1. ac
2. bd
3. $bd + a$
4. $ac(bd + a)$
5. ad
6. $ac(bd + a) + ad$

$$ac = 5$$

$$bd + a = 34$$

Strategy

- Selectively write on the board
- Selectively read the already written content
- Selectively forget (erase) some content

White board Analogy

$a = 1 \quad b = 3 \quad c = 5 \quad d = 11$

Compute $ac(bd + a) + ad$

1. ac
2. bd
3. $bd + a$
4. $ac(bd + a)$
5. ad
6. $ac(bd + a) + ad$

$$ac = 5$$

$$ac(bd + a) = 170$$

$$bd + a = 34$$

Strategy

- Selectively write on the board
- Selectively read the already written content
- Selectively forget (erase) some content

White board Analogy

$a = 1 \quad b = 3 \quad c = 5 \quad d = 11$

Compute $ac(bd + a) + ad$

1. ac
2. bd
3. $bd + a$
4. $ac(bd + a)$
5. ad
6. $ac(bd + a) + ad$

$$ac = 5$$

$$ac(bd + a) = 170$$

$$ad = 11$$

Strategy

- Selectively write on the board
- Selectively read the already written content
- Selectively forget (erase) some content

White board Analogy

$a = 1 \quad b = 3 \quad c = 5 \quad d = 11$

Compute $ac(bd + a) + ad$

1. ac
2. bd
3. $bd + a$
4. $ac(bd + a)$
5. ad
6. $ac(bd + a) + ad$

$$ac(bd + a) = 170$$

$$ad = 11$$

Strategy

- Selectively write on the board
- Selectively read the already written content
- Selectively forget (erase) some content

White board Analogy

$a = 1 \quad b = 3 \quad c = 5 \quad d = 11$

Compute $ac(bd + a) + ad$

1. ac
2. bd
3. $bd + a$
4. $ac(bd + a)$
5. ad
6. $ac(bd + a) + ad$

$$ad + ac(bd + a) = 181$$
$$ac(bd + a) = 170$$
$$ad = 11$$

Strategy

- Selectively write on the board
- Selectively read the already written content
- Selectively forget (erase) some content

White board Analogy

$a = 1 \quad b = 3 \quad c = 5 \quad d = 11$

Compute $ac(bd + a) + ad$

1. ac
2. bd
3. $bd + a$
4. $ac(bd + a)$
5. ad
6. $ac(bd + a) + ad$

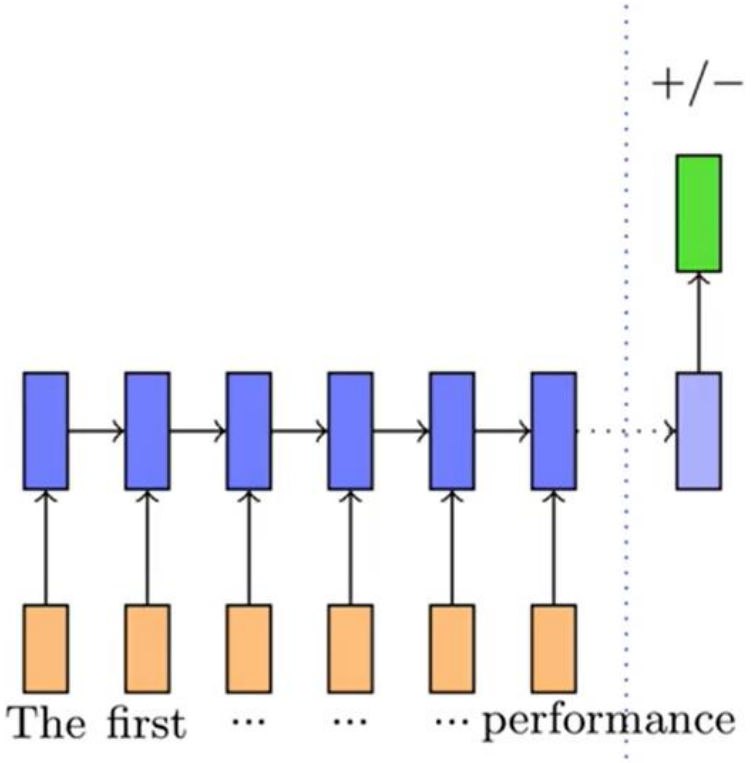
$$ad + ac(bd + a) = 181$$

Strategy

- Selectively write on the board
- Selectively read the already written content
- Selectively forget (erase) some content

Dealing with longer sequence

(An example where RNN need to selectively read write and forget)

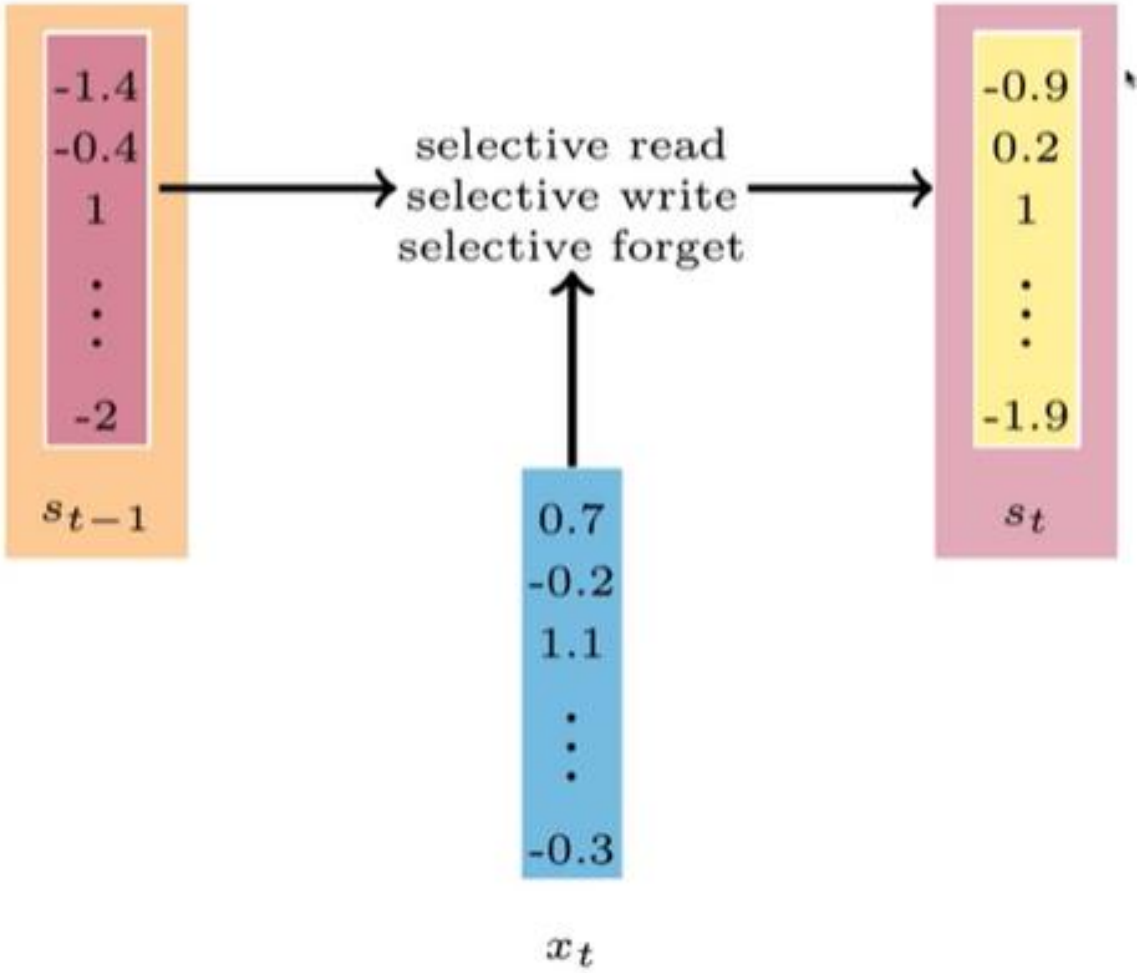


Ideally, we want to

- forget the information added by stop words (a, the, etc.)
- selectively read the information added by previous sentiment bearing words (awesome, amazing, etc.)
- selectively write new information from the current word to the state

Review: The first half of the movie was dry but the second half really picked up pace. The lead actor delivered an amazing performance

Long Short-Term Memory Cells – deals with longer sequences
(How to implement selective read write and Forget)



LSTM

While computing s_t from s_{t-1} we want to make sure that we use selective write, selective read and selective forget so that only important information is retained in s_t

Recap- RNN

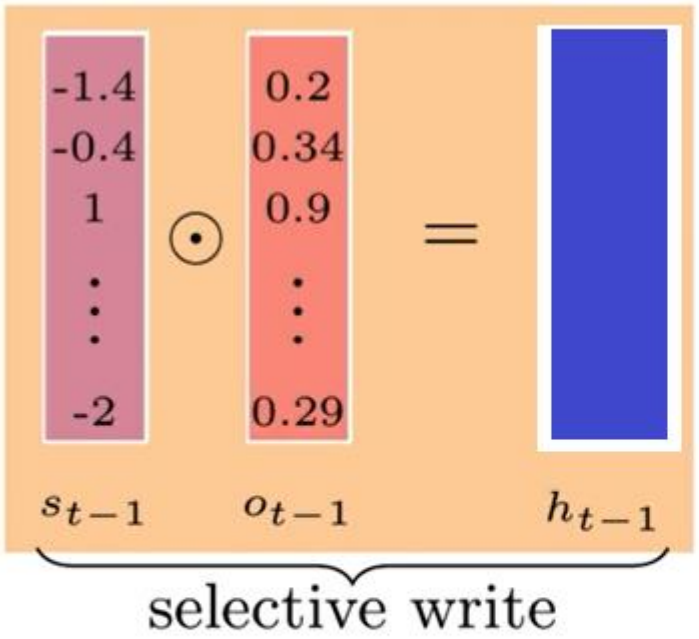
$$s_t = \sigma(Ux_t + Ws_{t-1} + b)$$

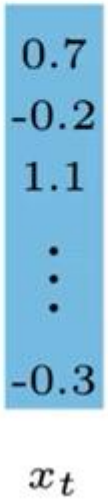
Selective write

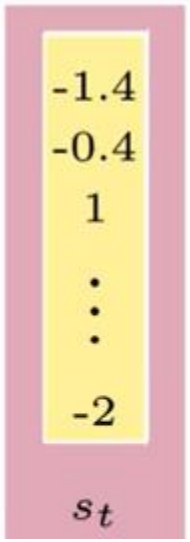
o_t is called the **output** gate

$$o_{t-1} = \sigma(U_o x_{t-1} + W_o h_{t-2} + b_o)$$

$$h_{t-1} = s_{t-1} \odot o_{t-1}$$







But how do we compute o_{t-1} ? How does the RNN know what fraction of the state to pass on?

- learn o_{t-1} from data
- the only thing that we learn from data is parameters
- **Solution:** express o_{t-1} using parameters

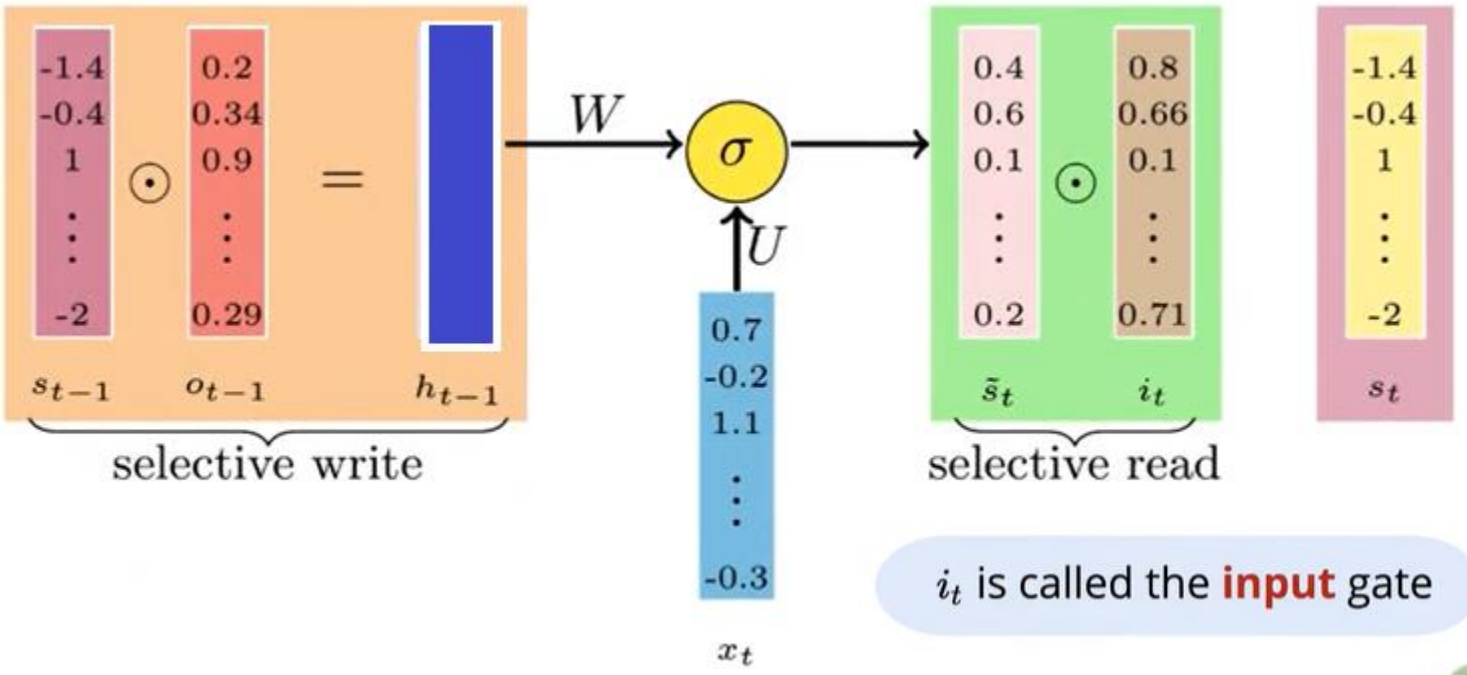
- instead of passing s_{t-1} as it is to s_t we want to pass (write) only some portions of it to the next state
- A reasonable way of doing this would be to assign a value between 0 and 1 which determines what fraction of the current state to pass on to the next state

Selective Read

$$\tilde{s}_t = \sigma(Ux_t + Wh_{t-1} + b)$$

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

$$= \tilde{s}_t \odot i_t$$



i_t is called the **input** gate

Previous state:

s_{t-1}

Output gate:

$o_{t-1} = \sigma(W_o h_{t-2} + U_o x_{t-1} + b_o)$

Selectively Write:

$h_{t-1} = o_{t-1} \odot \sigma(s_{t-1})$

Current (temporary) state:

$\tilde{s}_t = \sigma(Wh_{t-1} + Ux_t + b)$

Input gate:

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

Selectively Read:

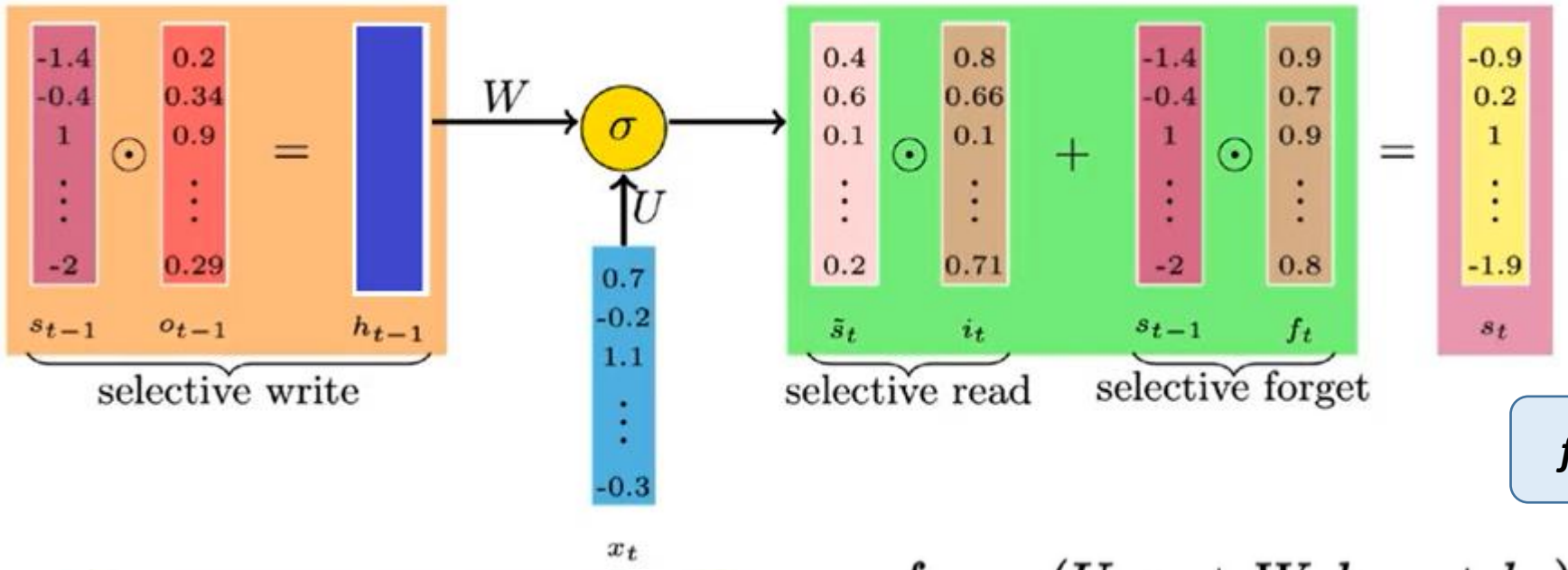
$i_t \odot \tilde{s}_t$

- $\tilde{s}_t$ thus captures all the information from the previous state h_{t-1} and the current input x_t
- However, we may not want to use all this new information and only selectively read from it before constructing the new cell

Selective Forget

Some LSTMs stop after selective read but the actual version LSTM has forget gate also

Want to make s_t depended on s_{t-1} but only the relevant portion of s_{t-1} (so forgetting some info of s_{t-1} and adding that to $\tilde{s}_t \odot i_t$)



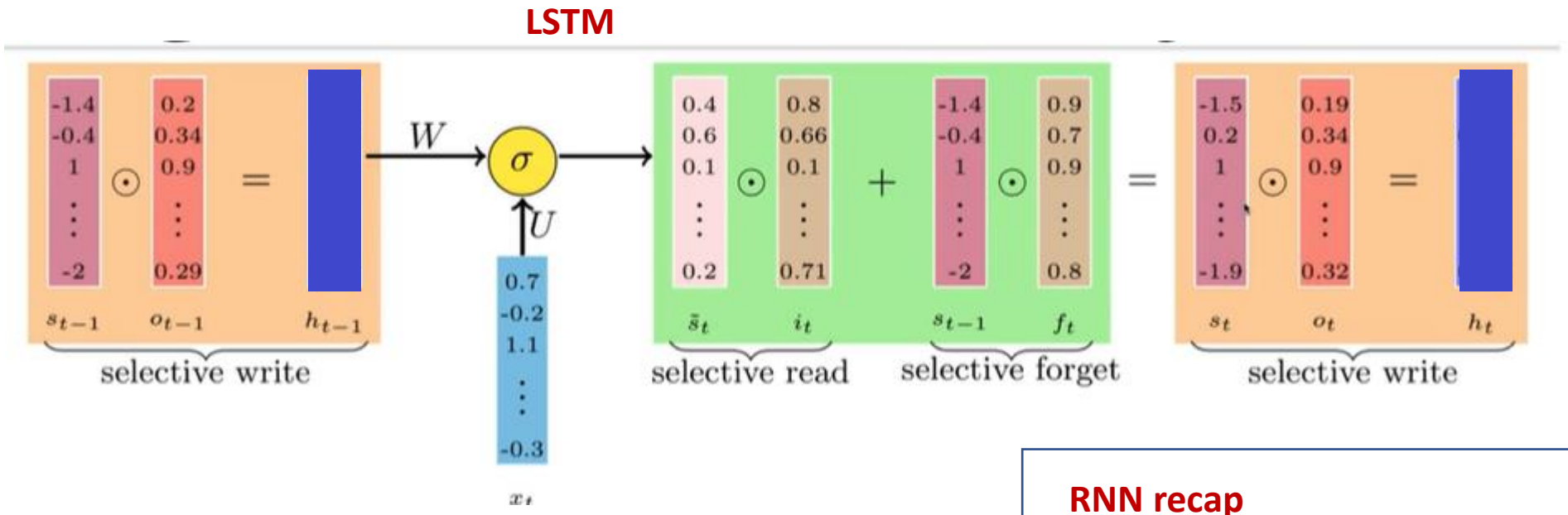
***ft* is called forget gate**

How do we combine $\tilde{s}_t$ and s_{t-1} to get the new state s_t

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

$$s_t = \tilde{s}_t \odot i_t + s_{t-1} \odot f_t$$

Summary-LSTM



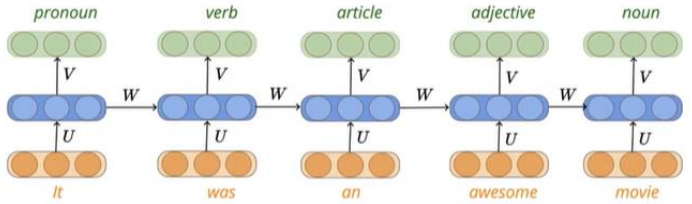
Gates:

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

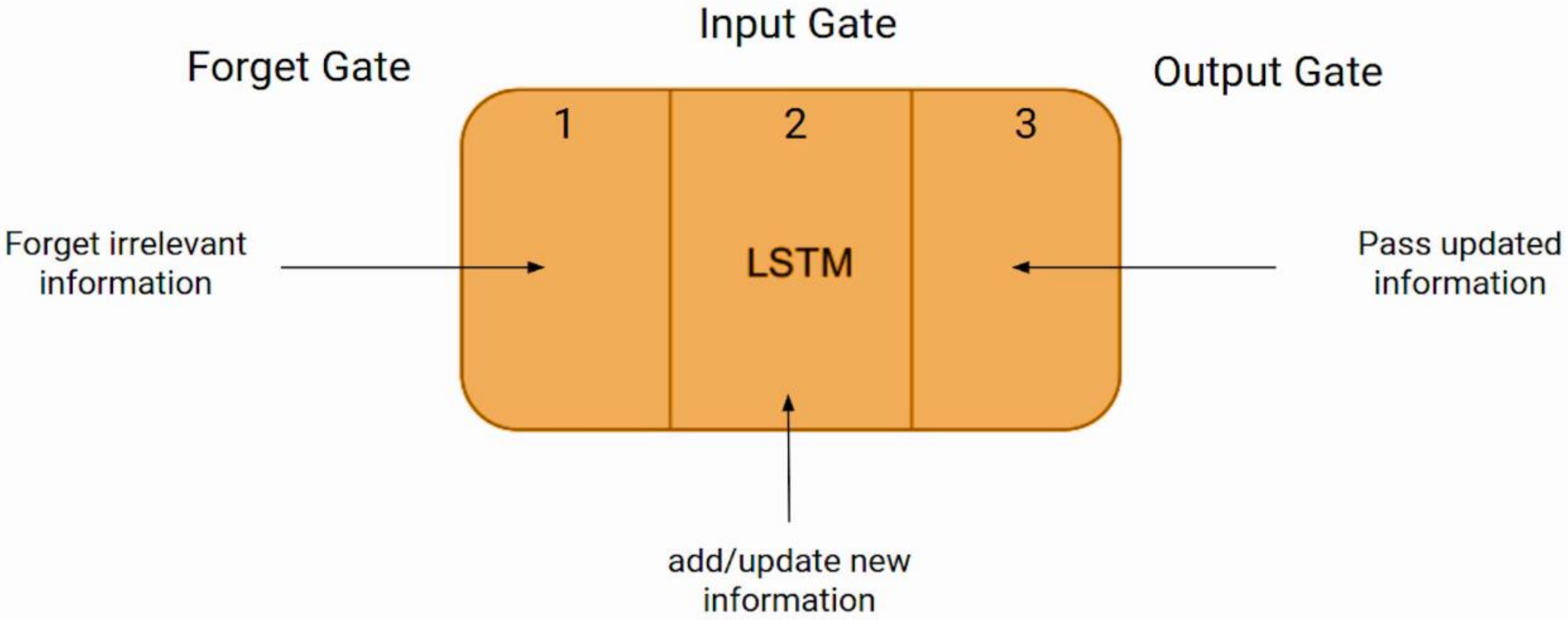
States:

$$\tilde{s}_t = \sigma(W h_{t-1} + U x_t + b)$$
$$s_t = f_t \odot s_{t-1} + i_t \odot \tilde{s}_t$$
$$h_t = o_t \odot \sigma(s_t)$$

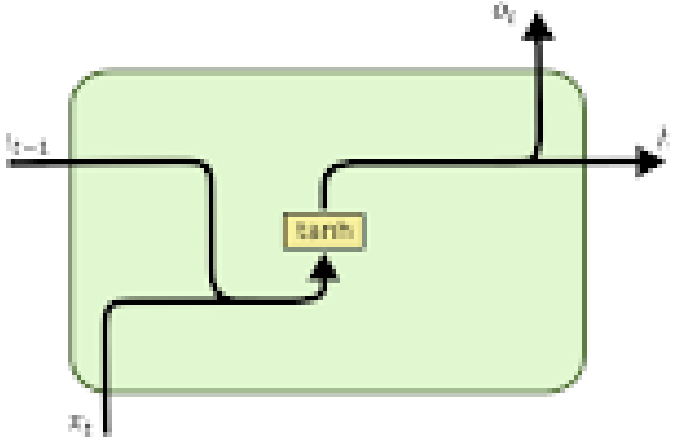
RNN recap



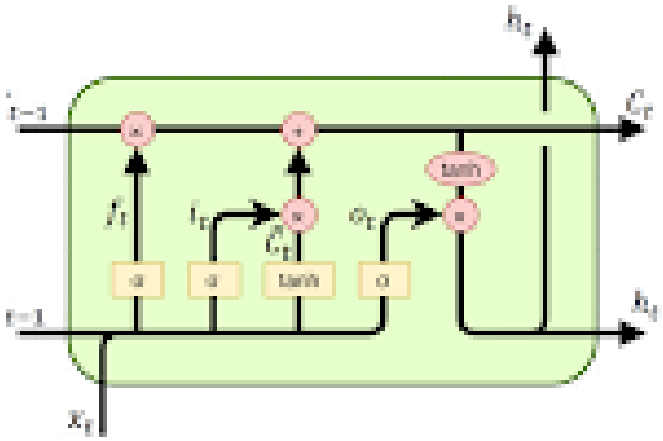
$$s_t = \sigma(U x_t + W s_{t-1} + b)$$


Input Gate
Forget Gate
Output Gate
1
2
3
Forget irrelevant information
LSTM
Pass updated information
add/update new information

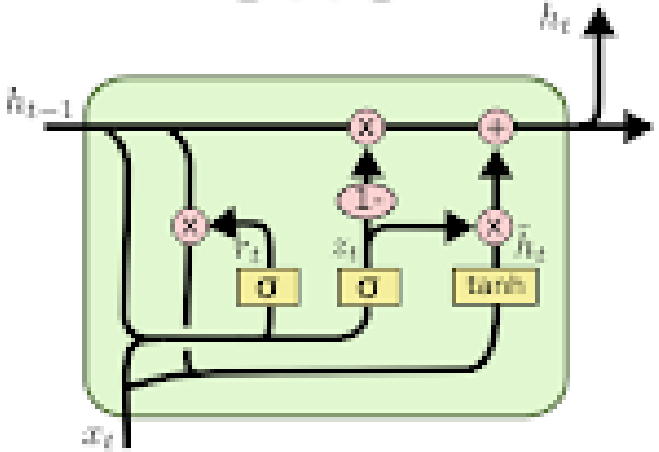
RNN



LSTM



GRU



Namah Shivaya

Courtesy : Video lectures of Dr.Mitesh Kapra