MACHINE INTELLIGENCE

UNIT - 5

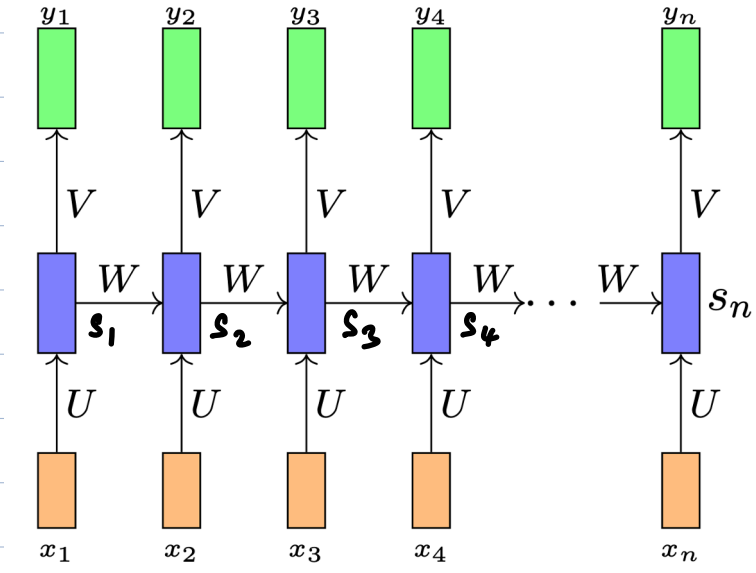
LSTM and GRU

feedback/corrections: vibha@pesu.pes.edu

VIBHA MASTI

Long Short Term Memory (LSTM) Cells & Gated Recurrent Units (GRU)

1. Selective Read, Selective Write, Selective Forget



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

- State at timestep i : s_i
- s_i contains information from all its previous timesteps
- contribution of earlier timesteps lost (and meaningless)

- state s_i is a fixed size memory (think of whiteboard)
- too much info makes it meaningless (impossible to see how s_{i-k} contributed to s_i)
- solution: selectively read from, write to, erase from (forget) board

Q: Assume a board that can only hold 3 statements at a time is being used to compute

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

where $a=1$ $b=3$ $c=5$ $d=11$

How would you compute with selective read, write and forget?

sel read

$ac = 5$
$bd = 33$
$bd+a = 34$

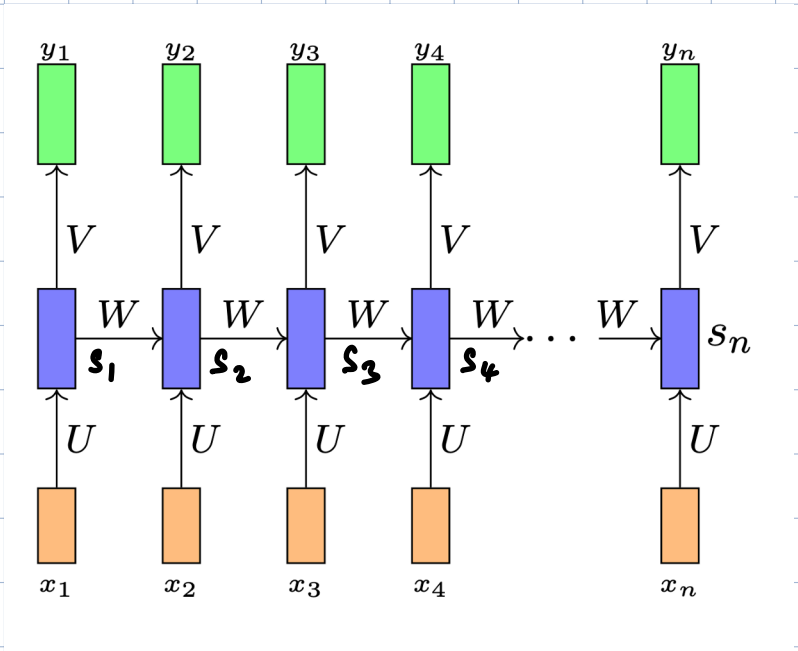
$\longrightarrow$

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

space exhausted

2. LSTM

- In an RNN, we want to overload state s_{t-1} with x_t to compute s_t
- Recall: dimensions



$x_i \in R^n$ (n-d input)
$s_i \in R^d$ (d-d state)
$y_i \in R^k$ (k classes)

$U \in R^{n \times d}$

$W \in R^{d \times d}$

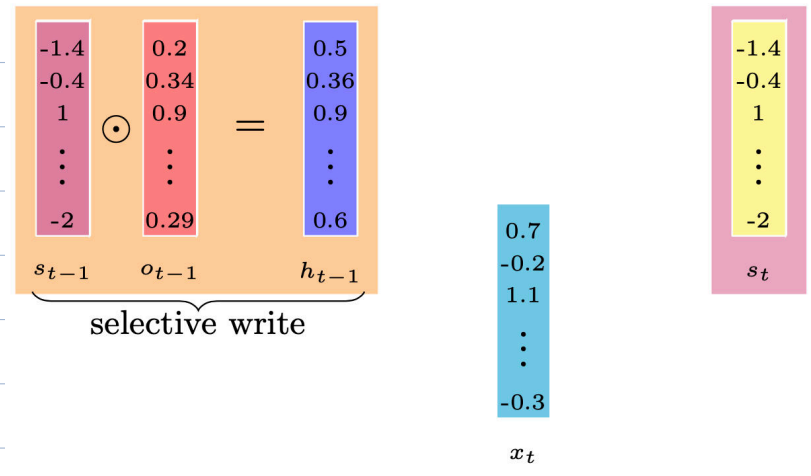
$V \in R^{d \times k}$

- Regular RNN,

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

2.1 Selective Write

- We want to selectively include only some elements of s_{t-1}
  (i) Binary – either include or not
  (ii) Probability – fraction/weight of each entry
- Vector o_{t-1} (gate) decides weights of s_{t-1} to be passed onto next state



Why sigmoid? weights b/w 0 and 1

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

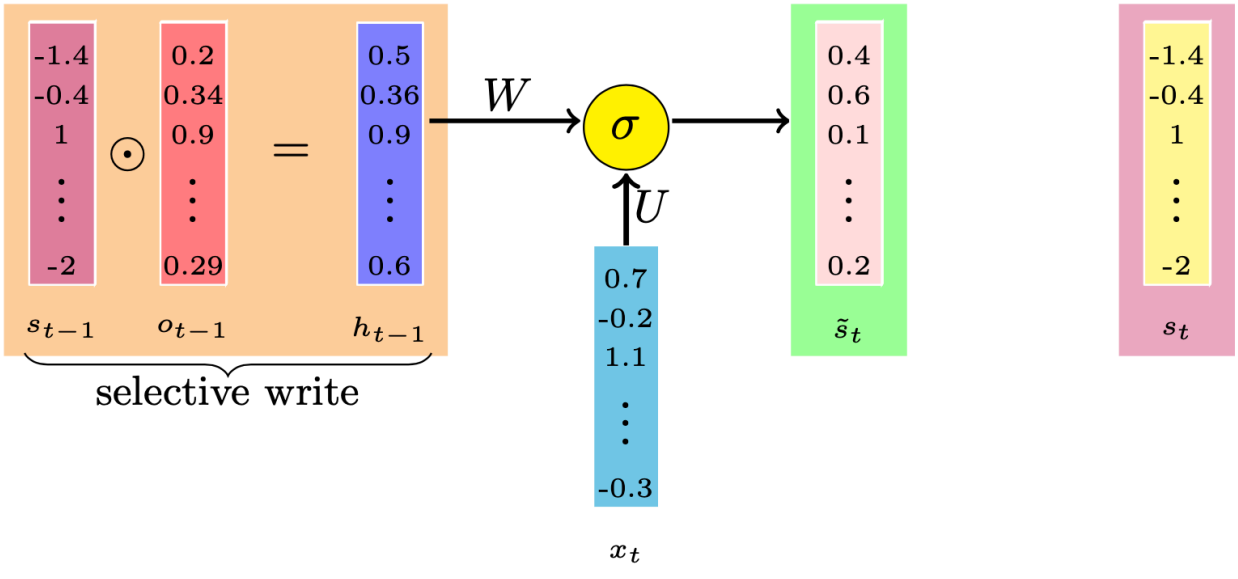
$$h_{t-1} = o_{t-1} \odot s_{t-1}$$

output gate

element-wise multiply

2.2 Selective Read

- Use h_{t-1} to compute $\tilde{s}_t$

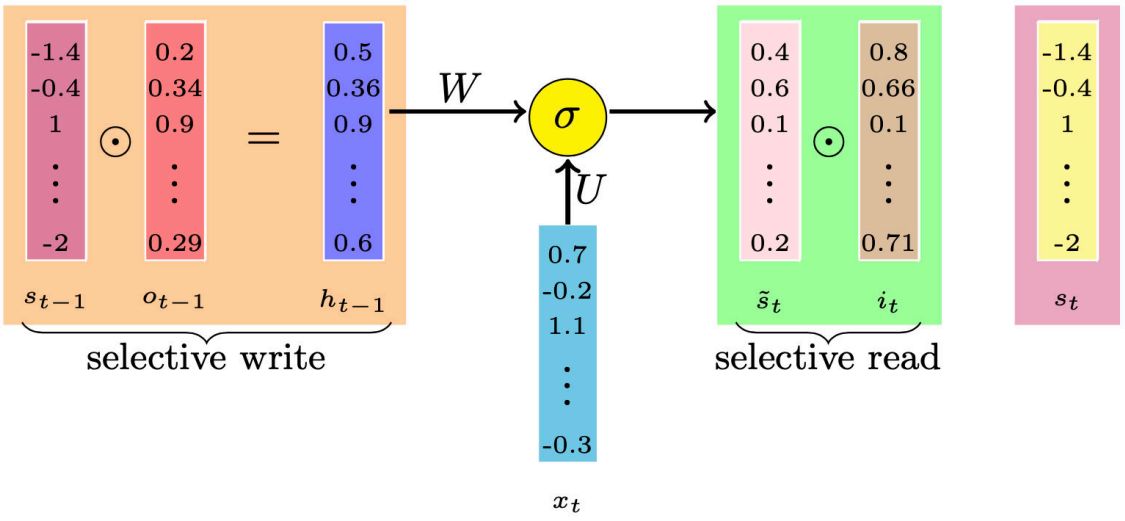


$$\tilde{s}_t = \sigma(Ux_t + Wh_{t-1} + b)$$

- selectively read from $\tilde{s}_t$ to get s_t
- Vector i_t decides input weights of $\tilde{s}_t$

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

$i_t \odot \tilde{s}_t$ is input



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:

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

Current state (temp):

$$\tilde{s}_t = \sigma(W h_{t-1} + U x_t + b)$$

Input gate:

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

Selectivey read:

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

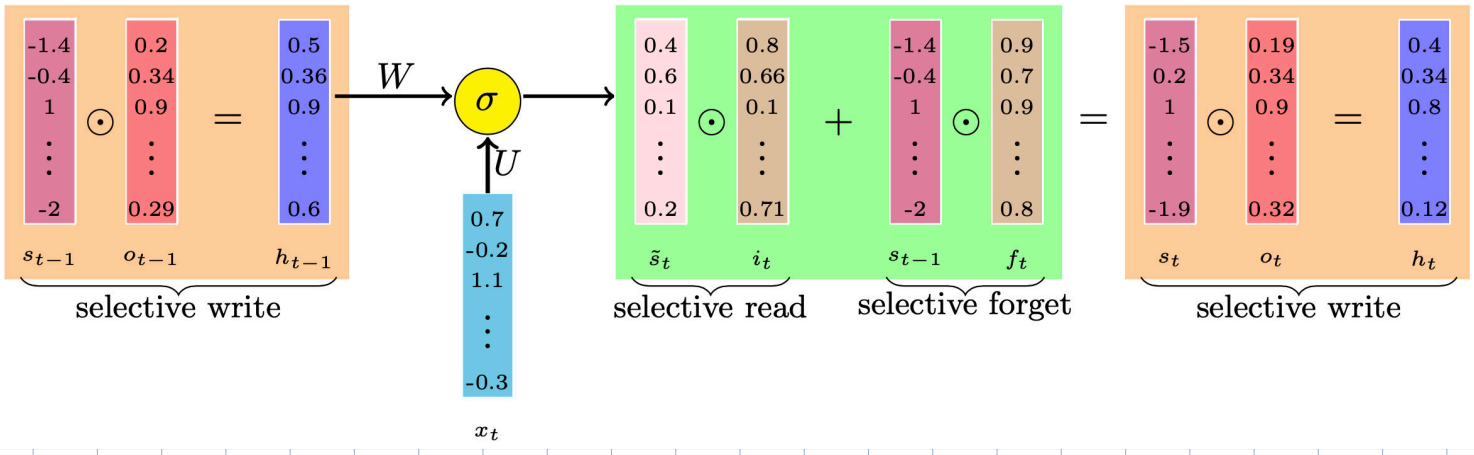
2.4 Selective Forget

- Combine $\tilde{s}_t$ and s_{t-1} to get new state s_t
- Simple: $s_t = s_{t-1} + i_t \odot \tilde{s}_t$
- To forget parts of s_{t-1}, use forget gate f_t

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

- Compute state as

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



- Note: we forget as well as selectively write to propagate information about s_{t-1} (merged in GRU)

- When LSTM run (eg: on TensorFlow), output is

h_t , s_t ← also called c_t (cell state)

LSTM Equations

Gates

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

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

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

States

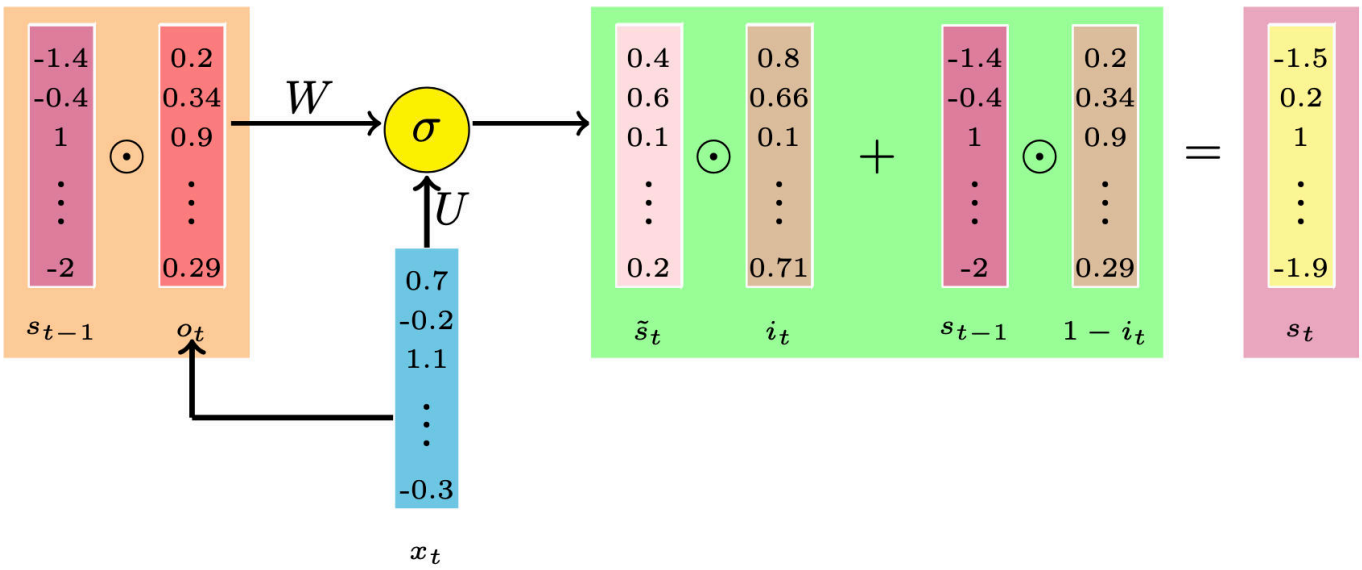
$$\tilde{s}_t = \sigma(W h_{t-1} + U x_t + b)$$

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

$$h_t = o_t \odot s_t$$

$$rnn_{out} = h_t$$

3. GRU



LSTM Equations

Gates

$$o_t = \sigma(W_o s_{t-1} + U_o x_{t-1} + b_o)$$

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

States

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

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

- Forget & input gates tied
- Update and reset gates