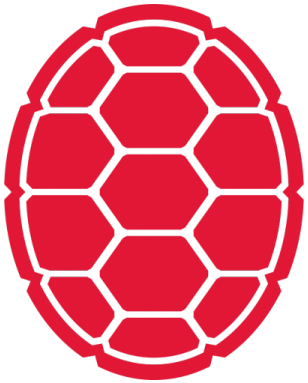
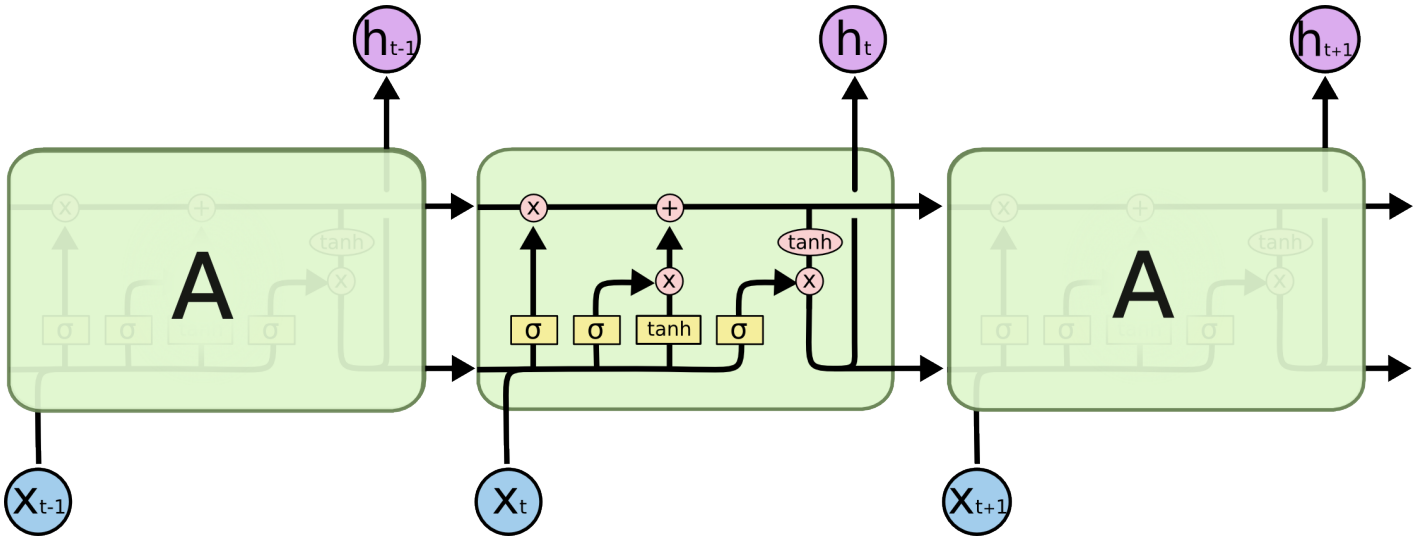


Long Short Term Memory Networks

Fenfei Guo and Jordan Boyd-Graber
University of Maryland

LSTM EXAMPLE

Recap of LSTM



Three gates: input (i_t), forget (f_t), out (o_t)

$$i_t = \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi})$$
$$f_t = \sigma(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf})$$
$$o_t = \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho})$$

New memory input: $\tilde{c}_t$

$$\tilde{c}_t = \tanh(W_{ic}x_t + b_{ic} + W_{hc}h_{t-1} + b_{hc})$$

Memorize and forget:

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t$$
$$h_t = o_t * \tanh(c_t)$$

Figuring out this LSTM

A	B
1.0 0.0	0.0 1.0

- input sequence: A, A, B

$$x_1 = [1.0, 0.0] \quad x_2 = [1.0, 0.0] \quad x_3 = [0.0, 1.0]$$

Figuring out this LSTM

A	B
1.0 0.0	0.0 1.0

- input: A, A, B

$$x_1 = [1.0, 0.0] \quad x_2 = [1.0, 0.0] \quad x_3 = [0.0, 1.0]$$

- prediction output:

$$y_t = \text{softmax}(h_t) \quad [\text{number of hidden nodes} = 2]$$

Model parameters for x_t

Input's input gate

$$W_{ii} = \begin{bmatrix} 4 & 4 \\ 2 & 2 \end{bmatrix} \quad (1)$$

forget gate

$$W_{if} = \begin{bmatrix} -2 & 3 \\ 2 & 3 \end{bmatrix} \quad (2)$$

cell params

$$W_{ic} = \begin{bmatrix} 1 & 3 \\ 0 & -3 \end{bmatrix} \quad (3)$$

output gate

$$W_{io} = \begin{bmatrix} 5 & 5 \\ 3 & 5 \end{bmatrix} \quad (4)$$

Set all $b = 0$ for simplicity

Model parameters for h_t

input gate

$$W_{hi} = \begin{bmatrix} 1 & 0 \\ 4 & -2 \end{bmatrix} \tag{5}$$

forget gate

$$W_{hf} = \begin{bmatrix} -1 & -2 \\ 0 & 0 \end{bmatrix} \tag{6}$$

cell params

$$W_{hc} = \begin{bmatrix} -4 & -8 \\ 4 & 3 \end{bmatrix} \tag{7}$$

output gate

$$W_{ho} = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} \tag{8}$$

Set all $b = 0$ for simplicity

Inputs

- Initial hidden states:

$$h_0 = [0.0, 0.0]^\top$$

- Initial memory input:

$$c_0 = [0.0, 0.0]^\top$$

- Input sequences in time:

$$x_1 = \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix} \quad x_2 = \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix} \quad x_3 = \begin{bmatrix} 0.0 \\ 1.0 \end{bmatrix}$$

Input Gate at $t = 1$: i_1

$$W_{ii} = \begin{bmatrix} 4.00 & 4.00 \\ 2.00 & 2.00 \end{bmatrix} \quad b_{ii} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix} \quad W_{hi} = \begin{bmatrix} 1.00 & 0.00 \\ 4.00 & -2.00 \end{bmatrix} \quad b_{hi} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix}$$

$$x_1 = [1.00, 0.00]^\top \qquad h_0 = [0.00, 0.00]^\top$$

Input Gate at $t = 1$: i_1

$$W_{ii} = \begin{bmatrix} 4.00 & 4.00 \\ 2.00 & 2.00 \end{bmatrix} \quad b_{ii} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix} \quad W_{hi} = \begin{bmatrix} 1.00 & 0.00 \\ 4.00 & -2.00 \end{bmatrix} \quad b_{hi} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix}$$

$$x_1 = [1.00, 0.00]^\top \qquad h_0 = [0.00, 0.00]^\top$$

$$i_1 = \sigma(W_{ii} x_1 + b_{ii} + W_{hi} h_0 + b_{hi}) \tag{9}$$

$$= \sigma([4.00, 2.00]^\top) \tag{10}$$

$$= [0.98, 0.88]^\top \tag{11}$$

Forget Gate at $t = 1$: f_1

$$W_{if} = \begin{bmatrix} -2.00 & 3.00 \\ 2.00 & 3.00 \end{bmatrix} \quad b_{if} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix} \qquad W_{hf} = \begin{bmatrix} -1.00 & -2.00 \\ 0.00 & 0.00 \end{bmatrix} \quad b_{hf} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix}$$

$$x_1 = [1.00, 0.00]^\top \qquad h_0 = [0.00, 0.00]^\top$$

Forget Gate at $t = 1$: f_1

$$W_{if} = \begin{bmatrix} -2.00 & 3.00 \\ 2.00 & 3.00 \end{bmatrix} \quad b_{if} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix} \qquad W_{hf} = \begin{bmatrix} -1.00 & -2.00 \\ 0.00 & 0.00 \end{bmatrix} \quad b_{hf} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix}$$

$$x_1 = [1.00, 0.00]^\top \qquad h_0 = [0.00, 0.00]^\top$$

$$f_1 = \sigma\left(W_{if} x_1 + b_{if} + W_{hf} h_0 + b_{hf}\right) \tag{12}$$

$$= \sigma\left([-2.00, 2.00]^\top\right) \tag{13}$$

$$= [0.12, 0.88]^\top \tag{14}$$

Output Gate at $t = 1$: o_1

$$W_{io} = \begin{bmatrix} 5.00 & 5.00 \\ 3.00 & 5.00 \end{bmatrix} \quad b_{io} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix} \quad W_{ho} = \begin{bmatrix} 1.00 & 0.00 \\ 2.00 & 1.00 \end{bmatrix} \quad b_{ho} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix}$$

$$x_1 = [1.00, 0.00]^\top \qquad h_0 = [0.00, 0.00]^\top$$

Output Gate at $t = 1$: o_1

$$W_{io} = \begin{bmatrix} 5.00 & 5.00 \\ 3.00 & 5.00 \end{bmatrix} \quad b_{io} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix} \quad W_{ho} = \begin{bmatrix} 1.00 & 0.00 \\ 2.00 & 1.00 \end{bmatrix} \quad b_{ho} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix}$$

$$x_1 = [1.00, 0.00]^\top \qquad h_0 = [0.00, 0.00]^\top$$

$$o_1 = \sigma(W_{io}x_1 + b_{io} + W_{ho}h_0 + b_{ho}) \tag{15}$$

$$= \sigma([5.00, 3.00]^\top) \tag{16}$$

$$= [0.99, 0.95]^\top \tag{17}$$

Memory Contribution at $t = 1$: $\tilde{c}_1$

$$W_{i\tilde{c}} = \begin{bmatrix} 1.00 & 3.00 \\ 0.00 & -3.00 \end{bmatrix} \quad b_{i\tilde{c}} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix} \qquad W_{h\tilde{c}} = \begin{bmatrix} -4.00 & -8.00 \\ 4.00 & 3.00 \end{bmatrix} b_{h\tilde{c}} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix}$$

$$x_1 = [1.00, 0.00]^\top \qquad h_0 = [0.00, 0.00]^\top$$

Memory Contribution at $t = 1$: $\tilde{c}_1$

$$W_{i\tilde{c}} = \begin{bmatrix} 1.00 & 3.00 \\ 0.00 & -3.00 \end{bmatrix} \quad b_{i\tilde{c}} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix} \qquad W_{h\tilde{c}} = \begin{bmatrix} -4.00 & -8.00 \\ 4.00 & 3.00 \end{bmatrix} b_{h\tilde{c}} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix}$$

$$x_1 = [1.00, 0.00]^\top \qquad h_0 = [0.00, 0.00]^\top$$

$$\begin{aligned} \tilde{c}_1 &= \tanh(W_{i\tilde{c}} x_1 + b_{i\tilde{c}} + W_{h\tilde{c}} h_0 + b_{h\tilde{c}}) && (18) \\ &= \tanh([1.00, 0.00]^\top) && (19) \\ &= [0.76, 0.00]^\top && (20) \end{aligned}$$

Forward message at time step 1

f_1	c_0	i_1	$\tilde{c}_1$
$[0.12, 0.88]^\top$	$[0.00, 0.00]^\top$	$[0.98, 0.88]^\top$	$[0.76, 0.00]^\top$

- Message forward (c_1)

$$c_1 = f_1 \circ c_0 + i_1 \circ \tilde{c}_1 \tag{21}$$

(22)

Forward message at time step 1

f_1	c_0	i_1	$\tilde{c}_1$
$[0.12, 0.88]^\top$	$[0.00, 0.00]^\top$	$[0.98, 0.88]^\top$	$[0.76, 0.00]^\top$

- Message forward (c_1)

$$c_1 = f_1 \circ c_0 + i_1 \circ \tilde{c}_1 \tag{21}$$

$$= [0.12, 0.88]^\top \circ [0.00, 0.00]^\top + [0.98, 0.88]^\top \circ [0.76, 0.00]^\top \tag{22}$$

$$\tag{23}$$

Forward message at time step 1

f_1	c_0	i_1	$\tilde{c}_1$
$[0.12, 0.88]^\top$	$[0.00, 0.00]^\top$	$[0.98, 0.88]^\top$	$[0.76, 0.00]^\top$

- Message forward (c_1)

$$
\begin{aligned}
c_1 &= f_1 \circ c_0 + i_1 \circ \tilde{c}_1 && (21)\\
&= [0.12, 0.88]^\top \circ [0.00, 0.00]^\top + [0.98, 0.88]^\top \circ [0.76, 0.00]^\top && (22)\\
&= [0.75, 0.00]^\top && (23)
\end{aligned}
$$

Forward message at time step 1

f_1	c_0	i_1	$\tilde{c}_1$
$[0.12, 0.88]^\top$	$[0.00, 0.00]^\top$	$[0.98, 0.88]^\top$	$[0.76, 0.00]^\top$

- Message forward (c_1)

$$c_1 = [0.75, 0.00]^\top \tag{21}$$

- New hidden (h_1)

$$h_1 \tag{22}$$

Forward message at time step 1

f_1	c_0	i_1	$\tilde{c}_1$
$[0.12, 0.88]^\top$	$[0.00, 0.00]^\top$	$[0.98, 0.88]^\top$	$[0.76, 0.00]^\top$

- Message forward (c_1)

$$c_1 = [0.75, 0.00]^\top \tag{21}$$

- New hidden (h_1)

$$h_1 = o_1 \circ \tanh(c_1) \tag{22}$$

$$\tag{23}$$

Forward message at time step 1

f_1	c_0	i_1	$\tilde{c}_1$
$[0.12, 0.88]^\top$	$[0.00, 0.00]^\top$	$[0.98, 0.88]^\top$	$[0.76, 0.00]^\top$

- Message forward (c_1)

$$c_1 = [0.75, 0.00]^\top \tag{21}$$

- New hidden (h_1)

$$h_1 = o_1 \circ \tanh(c_1) \tag{22}$$

$$= [0.99, 0.95]^\top \circ \tanh([0.75, 0.00]^\top) \tag{23}$$

$$\tag{24}$$

Forward message at time step 1

f_1	c_0	i_1	$\tilde{c}_1$
$[0.12, 0.88]^\top$	$[0.00, 0.00]^\top$	$[0.98, 0.88]^\top$	$[0.76, 0.00]^\top$

- Message forward (c_1)

$$c_1 = [0.75, 0.00]^\top \tag{21}$$

- New hidden (h_1)

$$h_1 = o_1 \circ \tanh(c_1) \tag{22}$$

$$= [0.99, 0.95]^\top \circ \tanh([0.75, 0.00]^\top) \tag{23}$$

$$= [0.63, 0.00]^\top \tag{24}$$

Forward message at time step 1

f_1	c_0	i_1	$\tilde{c}_1$
$[0.12, 0.88]^\top$	$[0.00, 0.00]^\top$	$[0.98, 0.88]^\top$	$[0.76, 0.00]^\top$

- Message forward (c_1)

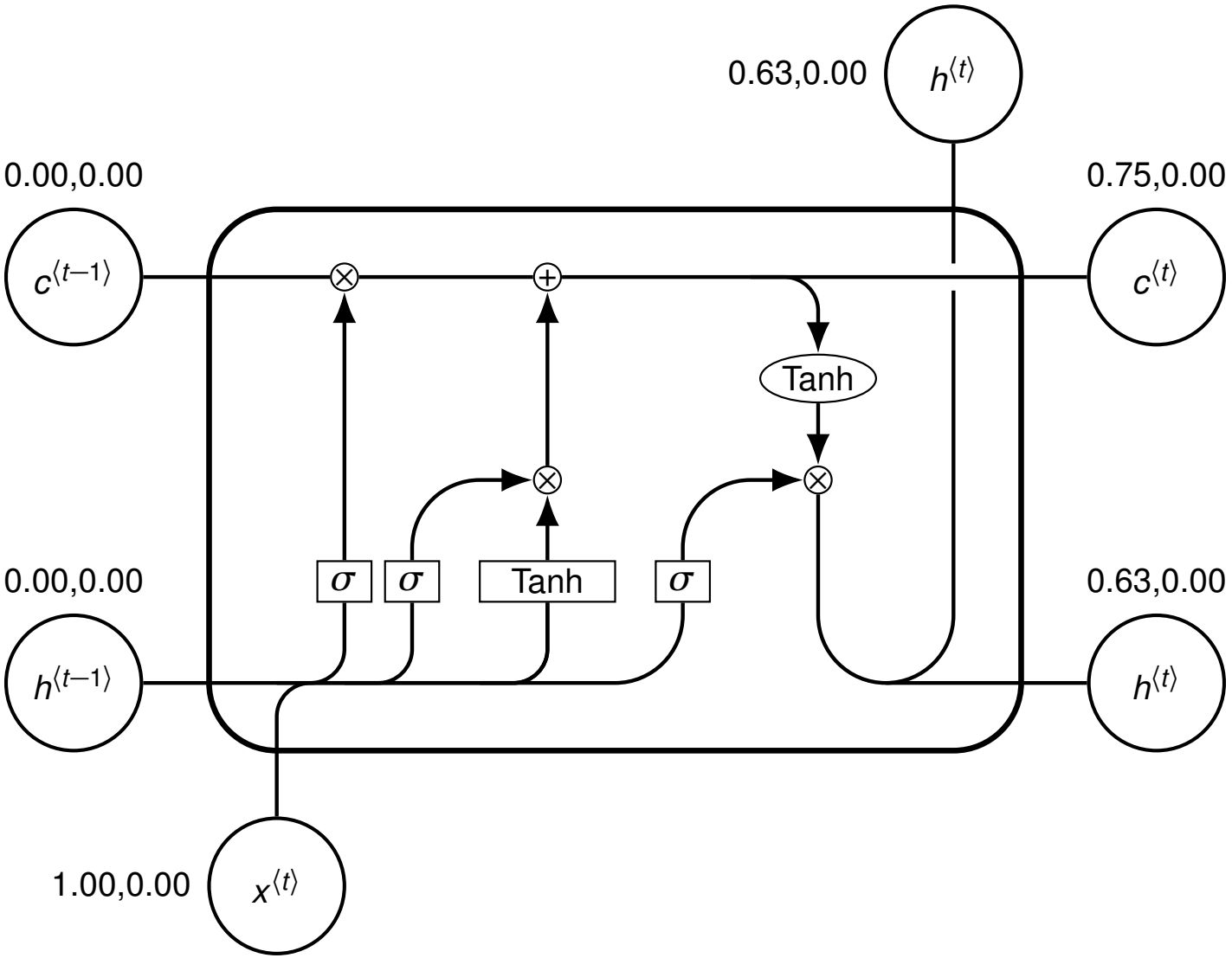
$$c_1 = [0.75, 0.00]^\top \tag{21}$$

- New hidden (h_1)

$$h_1 = [0.63, 0.00]^\top \tag{22}$$

- Prediction $y_1 = \text{softmax}(h_1) = 0$

Summary at $t = 1$



Input Gate at $t = 2$: i_1

$$W_{ii} = \begin{bmatrix} 4.00 & 4.00 \\ 2.00 & 2.00 \end{bmatrix} \qquad b_{ii} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix} \qquad W_{hi} = \begin{bmatrix} 1.00 & 0.00 \\ 4.00 & -2.00 \end{bmatrix} \qquad b_{hi} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix}$$

$$x_2 = [1.00, 0.00]^\top \qquad h_1 = [0.63, 0.00]^\top$$

Input Gate at $t = 2$: i_1

$$W_{ii} = \begin{bmatrix} 4.00 & 4.00 \\ 2.00 & 2.00 \end{bmatrix} \quad b_{ii} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix} \quad W_{hi} = \begin{bmatrix} 1.00 & 0.00 \\ 4.00 & -2.00 \end{bmatrix} \quad b_{hi} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix}$$

$$x_2 = [1.00, 0.00]^\top \qquad h_1 = [0.63, 0.00]^\top$$

$$i_2 = \sigma(W_{ii} x_2 + b_{ii} + W_{hi} h_1 + b_{hi}) \tag{23}$$

$$= \sigma([4.63, 4.52]^\top) \tag{24}$$

$$= [0.99, 0.99]^\top \tag{25}$$

Forget Gate at $t = 1$: f_1

$$W_{if} = \begin{bmatrix} -2.00 & 3.00 \\ 2.00 & 3.00 \end{bmatrix} \quad b_{if} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix} \qquad W_{hf} = \begin{bmatrix} -1.00 & -2.00 \\ 0.00 & 0.00 \end{bmatrix} \quad b_{hf} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix}$$

$$x_2 = [1.00, 0.00]^\top \qquad h_1 = [0.63, 0.00]^\top$$

Forget Gate at $t = 1$: f_1

$$W_{if} = \begin{bmatrix} -2.00 & 3.00 \\ 2.00 & 3.00 \end{bmatrix} \quad b_{if} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix} \qquad W_{hf} = \begin{bmatrix} -1.00 & -2.00 \\ 0.00 & 0.00 \end{bmatrix} \quad b_{hf} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix}$$

$$x_2 = [1.00, 0.00]^\top \qquad h_1 = [0.63, 0.00]^\top$$

$$f_2 = \sigma\left(W_{if} x_2 + b_{if} + W_{hf} h_1 + b_{hf}\right) \tag{26}$$

$$= \sigma\left([-2.63, 2.00]^\top\right) \tag{27}$$

$$= [0.07, 0.88]^\top \tag{28}$$

Output Gate at $t = 1$: o_1

$$W_{io} = \begin{bmatrix} 5.00 & 5.00 \\ 3.00 & 5.00 \end{bmatrix} \quad b_{io} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix} \quad W_{ho} = \begin{bmatrix} 1.00 & 0.00 \\ 2.00 & 1.00 \end{bmatrix} \quad b_{ho} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix}$$

$$x_2 = [1.00, 0.00]^\top \qquad h_1 = [0.63, 0.00]^\top$$

Output Gate at $t = 1$: o_1

$$W_{io} = \begin{bmatrix} 5.00 & 5.00 \\ 3.00 & 5.00 \end{bmatrix} \quad b_{io} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix} \quad W_{ho} = \begin{bmatrix} 1.00 & 0.00 \\ 2.00 & 1.00 \end{bmatrix} \quad b_{ho} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix}$$

$$x_2 = [1.00, 0.00]^\top \qquad h_1 = [0.63, 0.00]^\top$$

$$o_2 = \sigma(W_{io}x_2 + b_{io} + W_{ho}h_1 + b_{ho}) \tag{29}$$

$$= \sigma([5.63, 4.26]^\top) \tag{30}$$

$$= [1.00, 0.99]^\top \tag{31}$$

Memory Contribution at $t = 1$: $\tilde{c}_1$

$$W_{i\tilde{c}} = \begin{bmatrix} 1.00 & 3.00 \\ 0.00 & -3.00 \end{bmatrix} \quad b_{i\tilde{c}} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix} \qquad W_{h\tilde{c}} = \begin{bmatrix} -4.00 & -8.00 \\ 4.00 & 3.00 \end{bmatrix} b_{h\tilde{c}} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix}$$

$$x_2 = [1.00, 0.00]^\top \qquad h_1 = [0.63, 0.00]^\top$$

Memory Contribution at $t = 1$: $\tilde{c}_1$

$$W_{i\tilde{c}} = \begin{bmatrix} 1.00 & 3.00 \\ 0.00 & -3.00 \end{bmatrix} \quad b_{i\tilde{c}} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix} \qquad W_{h\tilde{c}} = \begin{bmatrix} -4.00 & -8.00 \\ 4.00 & 3.00 \end{bmatrix} b_{h\tilde{c}} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix}$$

$$x_2 = [1.00, 0.00]^\top \qquad h_1 = [0.63, 0.00]^\top$$

$$\tilde{c}_2 = \tanh(W_{i\tilde{c}} x_2 + b_{i\tilde{c}} + W_{h\tilde{c}} h_1 + b_{h\tilde{c}}) \tag{32}$$

$$= \tanh([-1.52, 2.52]^\top) \tag{33}$$

$$= [-0.91, 0.99]^\top \tag{34}$$

Forward message at time step 2

f_2	c_1	i_2	$\tilde{c}_2$
$[0.07, 0.88]^\top$	$[0.75, 0.00]^\top$	$[0.99, 0.99]^\top$	$[-0.91, 0.99]^\top$

- Message forward (c_2)

$$c_2 = f_2 \circ c_1 + i_2 \circ \tilde{c}_2 \tag{35}$$

(36)

Forward message at time step 2

f_2	c_1	i_2	$\tilde{c}_2$
$[0.07, 0.88]^\top$	$[0.75, 0.00]^\top$	$[0.99, 0.99]^\top$	$[-0.91, 0.99]^\top$

- Message forward (c_2)

$$c_2 = f_2 \circ c_1 + i_2 \circ \tilde{c}_2 \tag{35}$$

$$= [0.07, 0.88]^\top \circ [0.75, 0.00]^\top + [0.99, 0.99]^\top \circ [-0.91, 0.99]^\top \tag{36}$$

$$\tag{37}$$

Forward message at time step 2

f_2	c_1	i_2	$\tilde{c}_2$
$[0.07, 0.88]^\top$	$[0.75, 0.00]^\top$	$[0.99, 0.99]^\top$	$[-0.91, 0.99]^\top$

- Message forward (c_2)

$$c_2 = f_2 \circ c_1 + i_2 \circ \tilde{c}_2 \tag{35}$$

$$= [0.07, 0.88]^\top \circ [0.75, 0.00]^\top + [0.99, 0.99]^\top \circ [-0.91, 0.99]^\top \tag{36}$$

$$= [-0.85, 0.98]^\top \tag{37}$$

Forward message at time step 2

f_2	c_1	i_2	$\tilde{c}_2$
$[0.07, 0.88]^\top$	$[0.75, 0.00]^\top$	$[0.99, 0.99]^\top$	$[-0.91, 0.99]^\top$

- Message forward (c_2)

$$c_2 = [-0.85, 0.98]^\top \tag{35}$$

- New hidden (h_2)

$$h_2 \tag{36}$$

Forward message at time step 2

f_2	c_1	i_2	$\tilde{c}_2$
$[0.07, 0.88]^\top$	$[0.75, 0.00]^\top$	$[0.99, 0.99]^\top$	$[-0.91, 0.99]^\top$

- Message forward (c_2)

$$c_2 = [-0.85, 0.98]^\top \tag{35}$$

- New hidden (h_2)

$$h_2 = o_2 \circ \tanh(c_2) \tag{36}$$

$$\tag{37}$$

Forward message at time step 2

f_2	c_1	i_2	$\tilde{c}_2$
$[0.07, 0.88]^\top$	$[0.75, 0.00]^\top$	$[0.99, 0.99]^\top$	$[-0.91, 0.99]^\top$

- Message forward (c_2)

$$c_2 = [-0.85, 0.98]^\top \tag{35}$$

- New hidden (h_2)

$$h_2 = o_2 \circ \tanh(c_2) \tag{36}$$

$$= [1.00, 0.99]^\top \circ \tanh([-0.85, 0.98]^\top) \tag{37}$$

$$\tag{38}$$

Forward message at time step 2

f_2	c_1	i_2	$\tilde{c}_2$
$[0.07, 0.88]^\top$	$[0.75, 0.00]^\top$	$[0.99, 0.99]^\top$	$[-0.91, 0.99]^\top$

- Message forward (c_2)

$$c_2 = [-0.85, 0.98]^\top \tag{35}$$

- New hidden (h_2)

$$h_2 = o_2 \circ \tanh(c_2) \tag{36}$$

$$= [1.00, 0.99]^\top \circ \tanh([-0.85, 0.98]^\top) \tag{37}$$

$$= [-0.69, 0.74]^\top \tag{38}$$

Forward message at time step 2

f_2	c_1	i_2	$\tilde{c}_2$
$[0.07, 0.88]^\top$	$[0.75, 0.00]^\top$	$[0.99, 0.99]^\top$	$[-0.91, 0.99]^\top$

- Message forward (c_2)

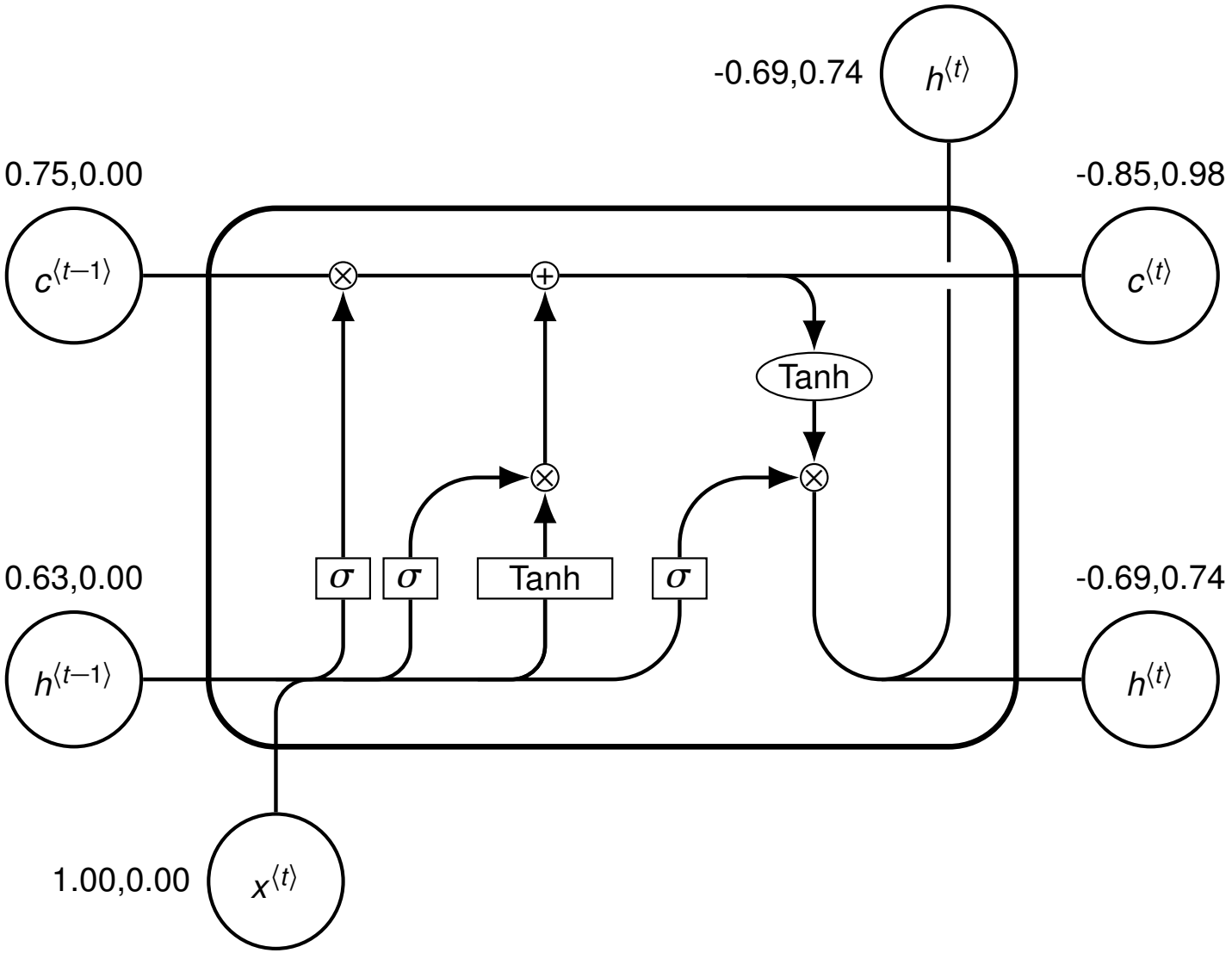
$$c_2 = [-0.85, 0.98]^\top \tag{35}$$

- New hidden (h_2)

$$h_2 = [-0.69, 0.74]^\top \tag{36}$$

- Prediction $y_2 = \mathrm{softmax}(h_2) = 1$

Summary at $t = 2$



Input Gate at $t = 3$: i_1

$$W_{ii} = \begin{bmatrix} 4.00 & 4.00 \\ 2.00 & 2.00 \end{bmatrix} \quad b_{ii} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix} \quad W_{hi} = \begin{bmatrix} 1.00 & 0.00 \\ 4.00 & -2.00 \end{bmatrix} \quad b_{hi} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix}$$

$$x_3 = [0.00, 1.00]^\top \qquad h_2 = [-0.69, 0.74]^\top$$

Input Gate at $t = 3$: i_1

$$W_{ii} = \begin{bmatrix} 4.00 & 4.00 \\ 2.00 & 2.00 \end{bmatrix} \quad b_{ii} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix} \quad W_{hi} = \begin{bmatrix} 1.00 & 0.00 \\ 4.00 & -2.00 \end{bmatrix} \quad b_{hi} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix}$$

$$x_3 = [0.00, 1.00]^\top \qquad h_2 = [-0.69, 0.74]^\top$$

$$i_3 = \sigma(W_{ii} x_3 + b_{ii} + W_{hi} h_2 + b_{hi}) \tag{37}$$

$$= \sigma([3.31, -2.24]^\top) \tag{38}$$

$$= [0.96, 0.10]^\top \tag{39}$$

Forget Gate at $t = 1$: f_1

$$W_{if} = \begin{bmatrix} -2.00 & 3.00 \\ 2.00 & 3.00 \end{bmatrix} \quad b_{if} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix} \qquad W_{hf} = \begin{bmatrix} -1.00 & -2.00 \\ 0.00 & 0.00 \end{bmatrix} \quad b_{hf} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix}$$

$$x_3 = [0.00, 1.00]^\top \qquad h_2 = [-0.69, 0.74]^\top$$

Forget Gate at $t = 1$: f_1

$$W_{if} = \begin{bmatrix} -2.00 & 3.00 \\ 2.00 & 3.00 \end{bmatrix} \quad b_{if} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix} \qquad W_{hf} = \begin{bmatrix} -1.00 & -2.00 \\ 0.00 & 0.00 \end{bmatrix} \quad b_{hf} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix}$$

$$x_3 = [0.00, 1.00]^\top \qquad h_2 = [-0.69, 0.74]^\top$$

$$f_3 = \sigma\left(W_{if} x_3 + b_{if} + W_{hf} h_2 + b_{hf}\right) \tag{40}$$

$$= \sigma\left([2.21, 3.00]^\top\right) \tag{41}$$

$$= [0.90, 0.95]^\top \tag{42}$$

Output Gate at $t = 1$: o_1

$$W_{io} = \begin{bmatrix} 5.00 & 5.00 \\ 3.00 & 5.00 \end{bmatrix} \quad b_{io} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix} \quad W_{ho} = \begin{bmatrix} 1.00 & 0.00 \\ 2.00 & 1.00 \end{bmatrix} \quad b_{ho} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix}$$

$$x_3 = [0.00, 1.00]^\top \qquad h_2 = [-0.69, 0.74]^\top$$

Output Gate at $t = 1$: o_1

$$W_{io} = \begin{bmatrix} 5.00 & 5.00 \\ 3.00 & 5.00 \end{bmatrix} \quad b_{io} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix} \quad W_{ho} = \begin{bmatrix} 1.00 & 0.00 \\ 2.00 & 1.00 \end{bmatrix} \quad b_{ho} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix}$$

$$x_3 = [0.00, 1.00]^\top \qquad h_2 = [-0.69, 0.74]^\top$$

$$o_3 = \sigma\left(W_{io} x_3 + b_{io} + W_{ho} h_2 + b_{ho}\right) \tag{43}$$

$$= \sigma\left([4.31, 4.36]^\top\right) \tag{44}$$

$$= [0.99, 0.99]^\top \tag{45}$$

Memory Contribution at $t = 1$: $\tilde{c}_1$

$$W_{i\tilde{c}} = \begin{bmatrix} 1.00 & 3.00 \\ 0.00 & -3.00 \end{bmatrix} \quad b_{i\tilde{c}} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix} \qquad W_{h\tilde{c}} = \begin{bmatrix} -4.00 & -8.00 \\ 4.00 & 3.00 \end{bmatrix} b_{h\tilde{c}} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix}$$

$$x_3 = [0.00, 1.00]^\top \qquad h_2 = [-0.69, 0.74]^\top$$

Memory Contribution at $t = 1$: $\tilde{c}_1$

$$W_{i\tilde{c}} = \begin{bmatrix} 1.00 & 3.00 \\ 0.00 & -3.00 \end{bmatrix} \quad b_{i\tilde{c}} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix} \qquad W_{h\tilde{c}} = \begin{bmatrix} -4.00 & -8.00 \\ 4.00 & 3.00 \end{bmatrix} b_{h\tilde{c}} = \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix}$$

$$x_3 = [0.00, 1.00]^\top \qquad h_2 = [-0.69, 0.74]^\top$$

$$\tilde{c}_3 = \tanh(W_{i\tilde{c}} x_3 + b_{i\tilde{c}} + W_{h\tilde{c}} h_2 + b_{h\tilde{c}}) \tag{46}$$

$$= \tanh([-0.18, -3.53]^\top) \tag{47}$$

$$= [-0.17, -1.00]^\top \tag{48}$$

Forward message at time step 3

f_3	c_2	i_3	$\tilde{c}_3$
$[0.90, 0.95]^\top$	$[-0.85, 0.98]^\top$	$[0.96, 0.10]^\top$	$[-0.17, -1.00]^\top$

- Message forward (c_3)

$$c_3 = f_3 \circ c_2 + i_3 \circ \tilde{c}_3 \tag{49}$$

(50)

Forward message at time step 3

f_3	c_2	i_3	$\tilde{c}_3$
$[0.90, 0.95]^\top$	$[-0.85, 0.98]^\top$	$[0.96, 0.10]^\top$	$[-0.17, -1.00]^\top$

- Message forward (c_3)

$$c_3 = f_3 \circ c_2 + i_3 \circ \tilde{c}_3 \tag{49}$$

$$= [0.90, 0.95]^\top \circ [-0.85, 0.98]^\top + [0.96, 0.10]^\top \circ [-0.17, -1.00]^\top \tag{50}$$

$$\tag{51}$$

Forward message at time step 3

f_3	c_2	i_3	$\tilde{c}_3$
$[0.90, 0.95]^\top$	$[-0.85, 0.98]^\top$	$[0.96, 0.10]^\top$	$[-0.17, -1.00]^\top$

- Message forward (c_3)

$$c_3 = f_3 \circ c_2 + i_3 \circ \tilde{c}_3 \tag{49}$$

$$= [0.90, 0.95]^\top \circ [-0.85, 0.98]^\top + [0.96, 0.10]^\top \circ [-0.17, -1.00]^\top \tag{50}$$

$$= [-0.93, 0.83]^\top \tag{51}$$

Forward message at time step 3

f_3	c_2	i_3	$\tilde{c}_3$
$[0.90, 0.95]^\top$	$[-0.85, 0.98]^\top$	$[0.96, 0.10]^\top$	$[-0.17, -1.00]^\top$

- Message forward (c_3)

$$c_3 = [-0.93, 0.83]^\top \tag{49}$$

- New hidden (h_3)

$$h_3 \tag{50}$$

Forward message at time step 3

f_3	c_2	i_3	$\tilde{c}_3$
$[0.90, 0.95]^\top$	$[-0.85, 0.98]^\top$	$[0.96, 0.10]^\top$	$[-0.17, -1.00]^\top$

- Message forward (c_3)

$$c_3 = [-0.93, 0.83]^\top \tag{49}$$

- New hidden (h_3)

$$h_3 = o_3 \circ \tanh(c_3) \tag{50}$$

(51)

Forward message at time step 3

f_3	c_2	i_3	$\tilde{c}_3$
$[0.90, 0.95]^\top$	$[-0.85, 0.98]^\top$	$[0.96, 0.10]^\top$	$[-0.17, -1.00]^\top$

- Message forward (c_3)

$$c_3 = [-0.93, 0.83]^\top \tag{49}$$

- New hidden (h_3)

$$h_3 = o_3 \circ \tanh(c_3) \tag{50}$$

$$= [0.99, 0.99]^\top \circ \tanh([-0.93, 0.83]^\top) \tag{51}$$

$$\tag{52}$$

Forward message at time step 3

f_3	c_2	i_3	$\tilde{c}_3$
$[0.90, 0.95]^\top$	$[-0.85, 0.98]^\top$	$[0.96, 0.10]^\top$	$[-0.17, -1.00]^\top$

- Message forward (c_3)

$$c_3 = [-0.93, 0.83]^\top \tag{49}$$

- New hidden (h_3)

$$h_3 = o_3 \circ \tanh(c_3) \tag{50}$$

$$= [0.99, 0.99]^\top \circ \tanh([-0.93, 0.83]^\top) \tag{51}$$

$$= [-0.72, 0.67]^\top \tag{52}$$

Forward message at time step 3

f_3	c_2	i_3	$\tilde{c}_3$
$[0.90, 0.95]^\top$	$[-0.85, 0.98]^\top$	$[0.96, 0.10]^\top$	$[-0.17, -1.00]^\top$

- Message forward (c_3)

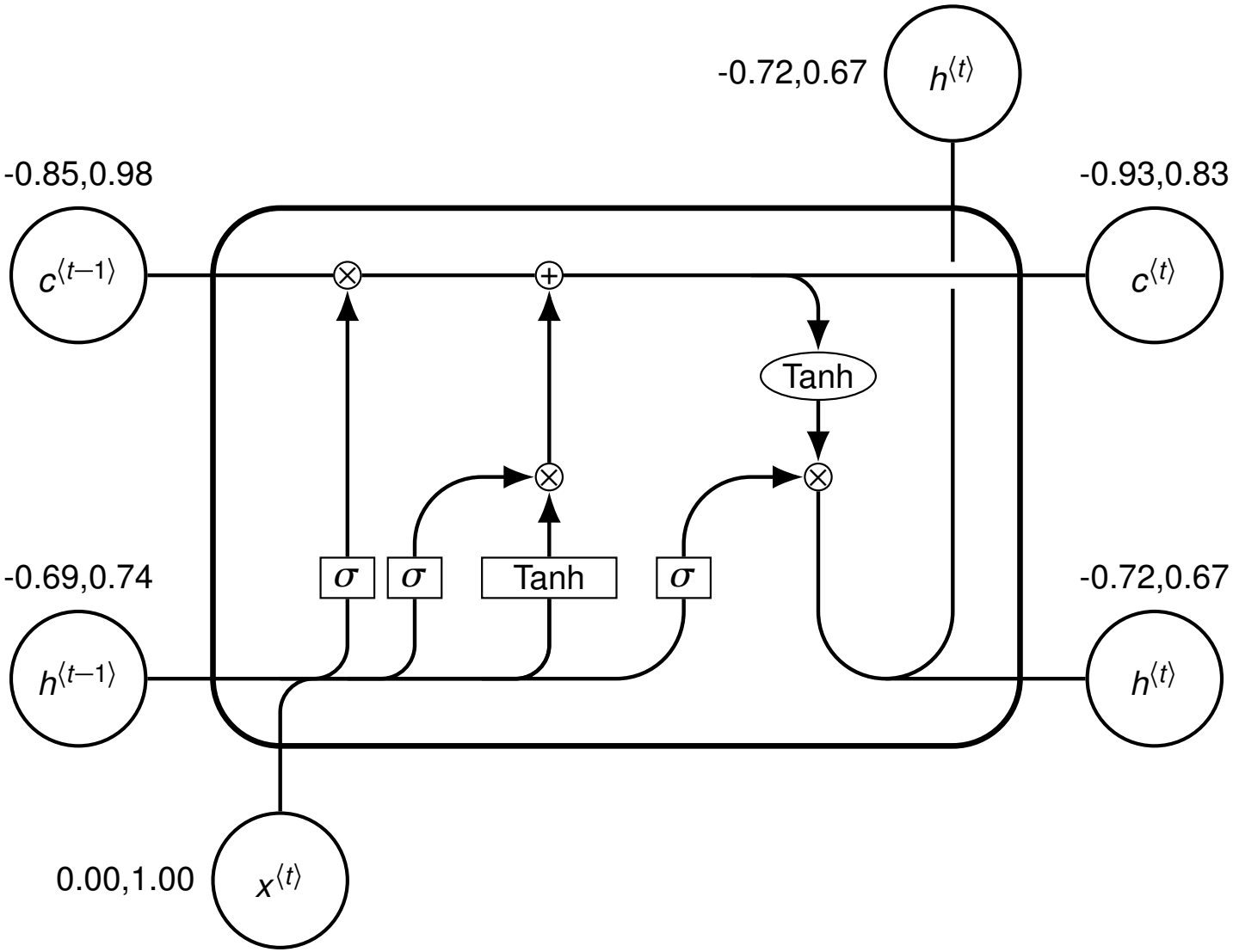
$$c_3 = [-0.93, 0.83]^\top \tag{49}$$

- New hidden (h_3)

$$h_3 = [-0.72, 0.67]^\top \tag{50}$$

- Prediction $y_3 = \text{softmax}(h_3) = 1$

Summary at $t = 3$



What's going on?

- What's the classification?
- What inputs are important?
- When can things be forgotten?
- How would other sequences be classified?

Training

- The parameters of LSTM showed in this example are obtained by training with cross-entropy loss function: (T=3)

$$\sum_{i=1}^{N}\sum_{t=1}^{T} H(y_{it},\ \text{target}_{it})$$

  - 0: accumulated number of A at time t is no larger than 1
  - 1: accumulated number of A at time t is larger than 1
  - Converted to binary classification problem:

$$\text{target}_1 = [1.0, 0.0] \quad \text{target}_2 = [0.0, 1.0] \quad \text{target}_3 = [0.0, 1.0]$$

Forwards at time step 1: i_1

Input's input gate

$$W_{ii} = \begin{bmatrix} 4 & 4 \\ 2 & 2 \end{bmatrix} \tag{51}$$

input gate

$$W_{hi} = \begin{bmatrix} 1 & 0 \\ 4 & -2 \end{bmatrix} \tag{52}$$

Compute

$$i_1 = \sigma\left(W_{ii} x_1 + W_{hi} h_0\right) \tag{53}$$

(54)

Forwards at time step 1: i_1

Input's input gate

$$W_{ii} = \begin{bmatrix} 4 & 4 \\ 2 & 2 \end{bmatrix} \tag{51}$$

input gate

$$W_{hi} = \begin{bmatrix} 1 & 0 \\ 4 & -2 \end{bmatrix} \tag{52}$$

Compute

$$i_1 = \sigma(W_{ii} x_1 + W_{hi} h_0) \tag{53}$$

$$= \sigma\left(\begin{bmatrix} 4 & 4 \\ 2 & 2 \end{bmatrix} \times \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}\right) \tag{54}$$

$$\tag{55}$$

Forwards at time step 1: i_1

Input's input gate

$$W_{ii} = \begin{bmatrix} 4 & 4 \\ 2 & 2 \end{bmatrix} \tag{51}$$

input gate

$$W_{hi} = \begin{bmatrix} 1 & 0 \\ 4 & -2 \end{bmatrix} \tag{52}$$

Compute

$$\begin{aligned} i_1 &= \sigma(W_{ii} x_1 + W_{hi} h_0) && (53) \\ &= \sigma\left(\begin{bmatrix} 4 & 4 \\ 2 & 2 \end{bmatrix} \times \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}\right) && (54) \\ &= \sigma([4.0, 2.0]^\top) && (55) \\ & && (56) \end{aligned}$$

Forwards at time step 1: i_1

Input's input gate

$$W_{ii} = \begin{bmatrix} 4 & 4 \\ 2 & 2 \end{bmatrix} \tag{51}$$

input gate

$$W_{hi} = \begin{bmatrix} 1 & 0 \\ 4 & -2 \end{bmatrix} \tag{52}$$

Compute

$$\begin{aligned} i_1 &= \sigma(W_{ii} x_1 + W_{hi} h_0) && (53) \\ &= \sigma\left(\begin{bmatrix} 4 & 4 \\ 2 & 2 \end{bmatrix} \times \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}\right) && (54) \\ &= \sigma([4.0, 2.0]^\top) && (55) \\ &= [1.0, 0.9]^\top && (56) \end{aligned}$$

Forwards at time step 1: f_1

forget gate

$$W_{if} = \begin{bmatrix} -2 & 3 \\ 2 & 3 \end{bmatrix} \tag{57}$$

forget gate

$$W_{hf} = \begin{bmatrix} -1 & -2 \\ 0 & 0 \end{bmatrix} \tag{58}$$

Compute

$$f_1 = \sigma\left(W_{if} x_1 + W_{hf} h_0\right) \tag{59}$$

(60)

Forwards at time step 1: f_1

forget gate

$$W_{if} = \begin{bmatrix} -2 & 3 \\ 2 & 3 \end{bmatrix} \tag{57}$$

forget gate

$$W_{hf} = \begin{bmatrix} -1 & -2 \\ 0 & 0 \end{bmatrix} \tag{58}$$

Compute

$$f_1 = \sigma\left(W_{if} x_1 + W_{hf} h_0\right) \tag{59}$$

$$= \sigma\left(\begin{bmatrix} -2 & 3 \\ 2 & 3 \end{bmatrix} \times \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}\right) \tag{60}$$

$$\tag{61}$$

Forwards at time step 1: f_1

forget gate

$$W_{if} = \begin{bmatrix} -2 & 3 \\ 2 & 3 \end{bmatrix} \tag{57}$$

forget gate

$$W_{hf} = \begin{bmatrix} -1 & -2 \\ 0 & 0 \end{bmatrix} \tag{58}$$

Compute

$$f_1 = \sigma(W_{if} x_1 + W_{hf} h_0) \tag{59}$$

$$= \sigma\left(\begin{bmatrix} -2 & 3 \\ 2 & 3 \end{bmatrix} \times \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}\right) \tag{60}$$

$$= \sigma([-2.0, 2.0]^\top) \tag{61}$$

$$\tag{62}$$

Forwards at time step 1: f_1

forget gate

$$W_{if} = \begin{bmatrix} -2 & 3 \\ 2 & 3 \end{bmatrix} \tag{57}$$

forget gate

$$W_{hf} = \begin{bmatrix} -1 & -2 \\ 0 & 0 \end{bmatrix} \tag{58}$$

Compute

$$f_1 = \sigma(W_{if} x_1 + W_{hf} h_0) \tag{59}$$

$$= \sigma\left(\begin{bmatrix} -2 & 3 \\ 2 & 3 \end{bmatrix} \times \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}\right) \tag{60}$$

$$= \sigma([-2.0, 2.0]^\top) \tag{61}$$

$$= [0.1, 0.9]^\top \tag{62}$$

Forwards at time step 1: o_1

output gate

$$W_{io} = \begin{bmatrix} 5 & 5 \\ 3 & 5 \end{bmatrix} \tag{63}$$

output gate

$$W_{ho} = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} \tag{64}$$

- $o_1 = \sigma(W_{io}x_1 + W_{ho}h_0)$

Forwards at time step 1: o_1

output gate

$$W_{io} = \begin{bmatrix} 5 & 5 \\ 3 & 5 \end{bmatrix} \tag{63}$$

output gate

$$W_{ho} = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} \tag{64}$$

- $o_1 = \sigma(W_{io}x_1 + W_{ho}h_0)$

$$= \sigma\left(\begin{bmatrix} 5 & 5 \\ 3 & 5 \end{bmatrix} \times \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}\right)$$

Forwards at time step 1: o_1

output gate

$$W_{io} = \begin{bmatrix} 5 & 5 \\ 3 & 5 \end{bmatrix} \tag{63}$$

output gate

$$W_{ho} = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} \tag{64}$$

- $o_1 = \sigma(W_{io}x_1 + W_{ho}h_0)$

$$= \sigma\left(\begin{bmatrix} 5 & 5 \\ 3 & 5 \end{bmatrix} \times \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}\right) \quad = \sigma([5.0, 3.0]^\top)$$

Forwards at time step 1: o_1

output gate

$$W_{io} = \begin{bmatrix} 5 & 5 \\ 3 & 5 \end{bmatrix} \tag{63}$$

output gate

$$W_{ho} = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} \tag{64}$$

- $o_1 = \sigma(W_{io}x_1 + W_{ho}h_0)$

$$= \sigma\left(\begin{bmatrix} 5 & 5 \\ 3 & 5 \end{bmatrix} \times \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}\right) \quad = \sigma([5.0, 3.0]^\top)$$

$$= [1.0, 1.0]^\top$$

Forwards at time step 1: $\tilde{c}_1$

cell params

$$W_{ic} = \begin{bmatrix} 1 & 3 \\ 0 & -3 \end{bmatrix} \tag{65}$$

cell params

$$W_{hc} = \begin{bmatrix} -4 & -8 \\ 4 & 3 \end{bmatrix} \tag{66}$$

- $\tilde{c}_1 = \tanh(W_{ic}x_1 + W_{hc}h_0)$

Forwards at time step 1: $\tilde{c}_1$

cell params

$$W_{ic} = \begin{bmatrix} 1 & 3 \\ 0 & -3 \end{bmatrix} \tag{65}$$

cell params

$$W_{hc} = \begin{bmatrix} -4 & -8 \\ 4 & 3 \end{bmatrix} \tag{66}$$

- $\tilde{c}_1 = \tanh(W_{ic}x_1 + W_{hc}h_0)$

$$= \tanh\left(\begin{bmatrix} 1 & 3 \\ 0 & -3 \end{bmatrix} \times \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}\right)$$

Forwards at time step 1: $\tilde{c}_1$

cell params

$$W_{ic} = \begin{bmatrix} 1 & 3 \\ 0 & -3 \end{bmatrix} \tag{65}$$

cell params

$$W_{hc} = \begin{bmatrix} -4 & -8 \\ 4 & 3 \end{bmatrix} \tag{66}$$

- $\tilde{c}_1 = \tanh(W_{ic}x_1 + W_{hc}h_0)$

$$= \tanh\left(\begin{bmatrix} 1 & 3 \\ 0 & -3 \end{bmatrix} \times \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}\right) \quad = \tanh([1.0, 0.0]^\top)$$

Forwards at time step 1: $\tilde{c}_1$

cell params

$$W_{ic} = \begin{bmatrix} 1 & 3 \\ 0 & -3 \end{bmatrix} \tag{65}$$

cell params

$$W_{hc} = \begin{bmatrix} -4 & -8 \\ 4 & 3 \end{bmatrix} \tag{66}$$

- $\tilde{c}_1 = \tanh(W_{ic}x_1 + W_{hc}h_0)$

$$= \tanh\left(\begin{bmatrix} 1 & 3 \\ 0 & -3 \end{bmatrix} \times \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}\right) \quad = \tanh([1.0, 0.0]^\top) \quad = [0.8, 0.0]^\top$$

Forwards at time step 1

f_1	c_0	i_1	$\tilde{c}_1$
$[0.1, 0.9]^\top$	$[0.0, 0.0]^\top$	$[1.0, 0.9]^\top$	$[0.8, 0.0]^\top$

- Message forward (c_1)

$$c_1 = f_1 \circ c_0 + i_1 \circ \tilde{c}_1 \tag{67}$$

$$\tag{68}$$

Forwards at time step 1

f_1	c_0	i_1	$\tilde{c}_1$
$[0.1, 0.9]^\top$	$[0.0, 0.0]^\top$	$[1.0, 0.9]^\top$	$[0.8, 0.0]^\top$

- Message forward (c_1)

$$c_1 = f_1 \circ c_0 + i_1 \circ \tilde{c}_1 \tag{67}$$

$$= [1.0, 0.9]^\top \circ [0.8, 0.0]^\top \tag{68}$$

$$\tag{69}$$

Forwards at time step 1

f_1	c_0	i_1	$\tilde{c}_1$
$[0.1, 0.9]^\top$	$[0.0, 0.0]^\top$	$[1.0, 0.9]^\top$	$[0.8, 0.0]^\top$

- Message forward (c_1)

$$c_1 = f_1 \circ c_0 + i_1 \circ \tilde{c}_1 \tag{67}$$

$$= [1.0, 0.9]^\top \circ [0.8, 0.0]^\top \tag{68}$$

$$\tag{69}$$

Forwards at time step 1

f_1	c_0	i_1	$\tilde{c}_1$
$[0.1, 0.9]^\top$	$[0.0, 0.0]^\top$	$[1.0, 0.9]^\top$	$[0.8, 0.0]^\top$

- Message forward (c_1)

$$c_1 = [0.8, 0.0]^\top \tag{67}$$

- New hidden (h_1)

$$h_1 \tag{68}$$

Forwards at time step 1

f_1	c_0	i_1	$\tilde{c}_1$
$[0.1, 0.9]^\top$	$[0.0, 0.0]^\top$	$[1.0, 0.9]^\top$	$[0.8, 0.0]^\top$

- Message forward (c_1)

$$c_1 = [0.8, 0.0]^\top \tag{67}$$

- New hidden (h_1)

$$h_1 = o_1 \circ \tanh(c_1) \tag{68}$$

$$\tag{69}$$

Forwards at time step 1

f_1	c_0	i_1	$\tilde{c}_1$
$[0.1, 0.9]^\top$	$[0.0, 0.0]^\top$	$[1.0, 0.9]^\top$	$[0.8, 0.0]^\top$

- Message forward (c_1)

$$c_1 = [0.8, 0.0]^\top \tag{67}$$

- New hidden (h_1)

$$h_1 = o_1 \circ \tanh(c_1) \tag{68}$$
$$= [1.0, 1.0]^\top \circ \tanh([0.8, 0.0]^\top) \tag{69}$$
$$\tag{70}$$

Forwards at time step 1

f_1	c_0	i_1	$\tilde{c}_1$
$[0.1, 0.9]^\top$	$[0.0, 0.0]^\top$	$[1.0, 0.9]^\top$	$[0.8, 0.0]^\top$

- Message forward (c_1)

$$c_1 = [0.8, 0.0]^\top \tag{67}$$

- New hidden (h_1)

$$\begin{aligned} h_1 &= o_1 \circ \tanh(c_1) && (68)\\ &= [1.0, 1.0]^\top \circ \tanh([0.8, 0.0]^\top) && (69)\\ &= [0.7, 0.0]^\top && (70) \end{aligned}$$

Forwards at time step 1

f_1	c_0	i_1	$\tilde{c}_1$
$[0.1, 0.9]^\top$	$[0.0, 0.0]^\top$	$[1.0, 0.9]^\top$	$[0.8, 0.0]^\top$

- Message forward (c_1)

$$c_1 = [0.8, 0.0]^\top \tag{67}$$

- New hidden (h_1)

$$h_1 = [0.7, 0.0]^\top \tag{68}$$

- Prediction $y_1 = \text{softmax}(h_1)$

Forwards at time step 2

$t = 2$ State

$x_2 = [1.0, 0.0]^\top;\ c_1 = [0.8, 0.0]^\top;\ h_1 = [0.7, 0.0]^\top$

Input's input gate

$$W_{ii} = \begin{bmatrix} 4 & 4 \\ 2 & 2 \end{bmatrix} \tag{69}$$

input gate

$$W_{hi} = \begin{bmatrix} 1 & 0 \\ 4 & -2 \end{bmatrix} \tag{70}$$

$$i_2 = \sigma\left(W_{ii} x_2 + W_{hi} h_1\right) \tag{71}$$

(72)

Forwards at time step 2

$t = 2$ State

$x_2 = [1.0, 0.0]^\top;\ c_1 = [0.8, 0.0]^\top;\ h_1 = [0.7, 0.0]^\top$

Input's input gate

$$W_{ii} = \begin{bmatrix} 4 & 4 \\ 2 & 2 \end{bmatrix} \tag{69}$$

input gate

$$W_{hi} = \begin{bmatrix} 1 & 0 \\ 4 & -2 \end{bmatrix} \tag{70}$$

$$i_2 = \sigma\left(W_{ii} x_2 + W_{hi} h_1\right) \tag{71}$$

$$= \sigma\left(\begin{bmatrix} 4 & 4 \\ 2 & 2 \end{bmatrix} \times \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 4 & -2 \end{bmatrix} \times \begin{bmatrix} 0.7 \\ 0.0 \end{bmatrix}\right) \tag{72}$$

(73)

Forwards at time step 2

$t = 2$ State

$x_2 = [1.0, 0.0]^\top;\ c_1 = [0.8, 0.0]^\top;\ h_1 = [0.7, 0.0]^\top$

Input's input gate

$$W_{ii} = \begin{bmatrix} 4 & 4 \\ 2 & 2 \end{bmatrix} \tag{69}$$

input gate

$$W_{hi} = \begin{bmatrix} 1 & 0 \\ 4 & -2 \end{bmatrix} \tag{70}$$

$$i_2 = \sigma(W_{ii} x_2 + W_{hi} h_1) \tag{71}$$

$$= \sigma\left(\begin{bmatrix} 4 & 4 \\ 2 & 2 \end{bmatrix} \times \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 4 & -2 \end{bmatrix} \times \begin{bmatrix} 0.7 \\ 0.0 \end{bmatrix}\right) \tag{72}$$

$$= \sigma([4.0, 2.0]^\top + [0.7, 2.8]^\top) \tag{73}$$

$$\tag{74}$$

Forwards at time step 2

$t = 2$ State

$x_2 = [1.0, 0.0]^\top;\ c_1 = [0.8, 0.0]^\top;\ h_1 = [0.7, 0.0]^\top$

Input's input gate

$$W_{ii} = \begin{bmatrix} 4 & 4 \\ 2 & 2 \end{bmatrix} \tag{69}$$

input gate

$$W_{hi} = \begin{bmatrix} 1 & 0 \\ 4 & -2 \end{bmatrix} \tag{70}$$

$$i_2 = \sigma(W_{ii} x_2 + W_{hi} h_1) \tag{71}$$

$$= \sigma\left(\begin{bmatrix} 4 & 4 \\ 2 & 2 \end{bmatrix} \times \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 4 & -2 \end{bmatrix} \times \begin{bmatrix} 0.7 \\ 0.0 \end{bmatrix}\right) \tag{72}$$

$$= \sigma([4.0, 2.0]^\top + [0.7, 2.8]^\top) = \sigma([4.7, 4.8]^\top) \tag{73}$$

$$= [1.0, 1.0]^\top \tag{74}$$

Forwards at time step 2

$t = 2$ State

$x_2 = [1.0, 0.0]^\top;\ c_1 = [0.8, 0.0]^\top;\ h_1 = [0.7, 0.0]^\top$

Input's input gate

$$W_{ii} = \begin{bmatrix} 4 & 4 \\ 2 & 2 \end{bmatrix} \tag{75}$$

input gate

$$W_{hi} = \begin{bmatrix} 1 & 0 \\ 4 & -2 \end{bmatrix} \tag{76}$$

$$f_2 = \sigma\left(W_{if} x_2 + W_{hf} h_1\right) \tag{77}$$

(78)

Forwards at time step 2

$t = 2$ State

$x_2 = [1.0, 0.0]^\top;\ c_1 = [0.8, 0.0]^\top;\ h_1 = [0.7, 0.0]^\top$

Input's input gate

$$W_{ii} = \begin{bmatrix} 4 & 4 \\ 2 & 2 \end{bmatrix} \tag{75}$$

input gate

$$W_{hi} = \begin{bmatrix} 1 & 0 \\ 4 & -2 \end{bmatrix} \tag{76}$$

$$f_2 = \sigma\left(W_{if} x_2 + W_{hf} h_1\right) \tag{77}$$

$$= \sigma\left(\begin{bmatrix} -2 & 3 \\ 2 & 3 \end{bmatrix} \times \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix} + \begin{bmatrix} -1 & -2 \\ 0 & 0 \end{bmatrix} \times \begin{bmatrix} 0.7 \\ 0.0 \end{bmatrix}\right) \tag{78}$$

(79)

Forwards at time step 2

$t = 2$ State

$x_2 = [1.0, 0.0]^\top;\ c_1 = [0.8, 0.0]^\top;\ h_1 = [0.7, 0.0]^\top$

Input's input gate

$$W_{ii} = \begin{bmatrix} 4 & 4 \\ 2 & 2 \end{bmatrix} \tag{75}$$

input gate

$$W_{hi} = \begin{bmatrix} 1 & 0 \\ 4 & -2 \end{bmatrix} \tag{76}$$

$$f_2 = \sigma(W_{if} x_2 + W_{hf} h_1) \tag{77}$$

$$= \sigma\left(\begin{bmatrix} -2 & 3 \\ 2 & 3 \end{bmatrix} \times \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix} + \begin{bmatrix} -1 & -2 \\ 0 & 0 \end{bmatrix} \times \begin{bmatrix} 0.7 \\ 0.0 \end{bmatrix}\right) \tag{78}$$

$$= \sigma([-2.0, 2.0]^\top + [-0.7, 0.0]^\top) \tag{79}$$

$$\tag{80}$$

Forwards at time step 2

$t = 2$ State

$x_2 = [1.0, 0.0]^\top;\ c_1 = [0.8, 0.0]^\top;\ h_1 = [0.7, 0.0]^\top$

Input's input gate

$$W_{ii} = \begin{bmatrix} 4 & 4 \\ 2 & 2 \end{bmatrix} \tag{75}$$

input gate

$$W_{hi} = \begin{bmatrix} 1 & 0 \\ 4 & -2 \end{bmatrix} \tag{76}$$

$$f_2 = \sigma(W_{if} x_2 + W_{hf} h_1) \tag{77}$$

$$= \sigma\left(\begin{bmatrix} -2 & 3 \\ 2 & 3 \end{bmatrix} \times \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix} + \begin{bmatrix} -1 & -2 \\ 0 & 0 \end{bmatrix} \times \begin{bmatrix} 0.7 \\ 0.0 \end{bmatrix}\right) \tag{78}$$

$$= \sigma([-2.0, 2.0]^\top + [-0.7, 0.0]^\top) \tag{79}$$

$$= \sigma([-2.7, 2.0]^\top) = [0.1, 0.9]^\top \tag{80}$$

Forwards at time step 2

$t = 2$ State

$x_2 = [1.0, 0.0]^\top;\ c_1 = [0.8, 0.0]^\top;\ h_1 = [0.7, 0.0]^\top$

output gate

$$W_{io} = \begin{bmatrix} 5 & 5 \\ 3 & 5 \end{bmatrix} \tag{81}$$

output gate

$$W_{ho} = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} \tag{82}$$

$$o_2 = \sigma\left(W_{io} x_2 + W_{ho} h_1\right) \tag{83}$$

(84)

Forwards at time step 2

$t = 2$ State

$x_2 = [1.0, 0.0]^\top;\ c_1 = [0.8, 0.0]^\top;\ h_1 = [0.7, 0.0]^\top$

output gate

$$W_{io} = \begin{bmatrix} 5 & 5 \\ 3 & 5 \end{bmatrix} \tag{81}$$

output gate

$$W_{ho} = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} \tag{82}$$

$$o_2 = \sigma\left(W_{io} x_2 + W_{ho} h_1\right) \tag{83}$$

$$= \sigma\left(\begin{bmatrix} 5 & 5 \\ 3 & 5 \end{bmatrix} \times \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} \times \begin{bmatrix} 0.7 \\ 0.0 \end{bmatrix}\right) \tag{84}$$

$$\tag{85}$$

Forwards at time step 2

$t = 2$ State

$x_2 = [1.0, 0.0]^\top;\ c_1 = [0.8, 0.0]^\top;\ h_1 = [0.7, 0.0]^\top$

output gate

$$W_{io} = \begin{bmatrix} 5 & 5 \\ 3 & 5 \end{bmatrix} \tag{81}$$

output gate

$$W_{ho} = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} \tag{82}$$

$$o_2 = \sigma(W_{io}x_2 + W_{ho}h_1) \tag{83}$$

$$= \sigma\left(\begin{bmatrix} 5 & 5 \\ 3 & 5 \end{bmatrix} \times \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} \times \begin{bmatrix} 0.7 \\ 0.0 \end{bmatrix}\right) \tag{84}$$

$$= \sigma([5.0, 3.0]^\top + [0.7, 1.4]^\top) \tag{85}$$

$$\tag{86}$$

Forwards at time step 2

$t = 2$ State

$x_2 = [1.0, 0.0]^\top;\ c_1 = [0.8, 0.0]^\top;\ h_1 = [0.7, 0.0]^\top$

output gate

$$W_{io} = \begin{bmatrix} 5 & 5 \\ 3 & 5 \end{bmatrix} \tag{81}$$

output gate

$$W_{ho} = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} \tag{82}$$

$$o_2 = \sigma(W_{io}x_2 + W_{ho}h_1) \tag{83}$$

$$= \sigma\left(\begin{bmatrix} 5 & 5 \\ 3 & 5 \end{bmatrix} \times \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} \times \begin{bmatrix} 0.7 \\ 0.0 \end{bmatrix}\right) \tag{84}$$

$$= \sigma([5.0, 3.0]^\top + [0.7, 1.4]^\top) \tag{85}$$

$$= \sigma([5.7, 4.4]^\top) = [1.0, 1.0]^\top \tag{86}$$

Forwards at time step 2

$t = 2$ State

$x_2 = [1.0, 0.0]^\top;\ c_1 = [0.8, 0.0]^\top;\ h_1 = [0.7, 0.0]^\top$

cell params

$$W_{ic} = \begin{bmatrix} 1 & 3 \\ 0 & -3 \end{bmatrix} \tag{87}$$

cell params

$$W_{hc} = \begin{bmatrix} -4 & -8 \\ 4 & 3 \end{bmatrix} \tag{88}$$

$$\tilde{c_2} = \tanh\left(W_{ic}x_2 + W_{hc}h_1\right) \tag{89}$$

(90)

Forwards at time step 2

$t = 2$ State

$x_2 = [1.0, 0.0]^\top;\ c_1 = [0.8, 0.0]^\top;\ h_1 = [0.7, 0.0]^\top$

cell params

$$W_{ic} = \begin{bmatrix} 1 & 3 \\ 0 & -3 \end{bmatrix} \tag{87}$$

cell params

$$W_{hc} = \begin{bmatrix} -4 & -8 \\ 4 & 3 \end{bmatrix} \tag{88}$$

$$\tilde{c_2} = \tanh(W_{ic}x_2 + W_{hc}h_1) \tag{89}$$

$$= \tanh\left(\begin{bmatrix} 1 & 3 \\ 0 & -3 \end{bmatrix} \times \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix} + \begin{bmatrix} -4 & -8 \\ 4 & 3 \end{bmatrix} \times \begin{bmatrix} 0.7 \\ 0.0 \end{bmatrix}\right) \tag{90}$$

(91)

Forwards at time step 2

$t = 2$ State

$x_2 = [1.0, 0.0]^\top;\ c_1 = [0.8, 0.0]^\top;\ h_1 = [0.7, 0.0]^\top$

cell params

$$W_{ic} = \begin{bmatrix} 1 & 3 \\ 0 & -3 \end{bmatrix} \tag{87}$$

cell params

$$W_{hc} = \begin{bmatrix} -4 & -8 \\ 4 & 3 \end{bmatrix} \tag{88}$$

$$\tilde{c_2} = \tanh(W_{ic}x_2 + W_{hc}h_1) \tag{89}$$

$$= \tanh\left(\begin{bmatrix} 1 & 3 \\ 0 & -3 \end{bmatrix} \times \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix} + \begin{bmatrix} -4 & -8 \\ 4 & 3 \end{bmatrix} \times \begin{bmatrix} 0.7 \\ 0.0 \end{bmatrix}\right) \tag{90}$$

$$= \tanh([1.0, 0.0]^\top + [-2.8, 2.8]^\top) \tag{91}$$

$$\tag{92}$$

Forwards at time step 2

$t = 2$ State

$x_2 = [1.0, 0.0]^\top;\ c_1 = [0.8, 0.0]^\top;\ h_1 = [0.7, 0.0]^\top$

cell params

$$W_{ic} = \begin{bmatrix} 1 & 3 \\ 0 & -3 \end{bmatrix} \tag{87}$$

cell params

$$W_{hc} = \begin{bmatrix} -4 & -8 \\ 4 & 3 \end{bmatrix} \tag{88}$$

$$\tilde{c_2} = \tanh(W_{ic}x_2 + W_{hc}h_1) \tag{89}$$

$$= \tanh\left(\begin{bmatrix} 1 & 3 \\ 0 & -3 \end{bmatrix} \times \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix} + \begin{bmatrix} -4 & -8 \\ 4 & 3 \end{bmatrix} \times \begin{bmatrix} 0.7 \\ 0.0 \end{bmatrix}\right) \tag{90}$$

$$= \tanh([1.0, 0.0]^\top + [-2.8, 2.8]^\top) \tag{91}$$

$$= \tanh([-1.8, 2.8]^\top) = [-0.9, 1.0]^\top \tag{92}$$

Forwards at time step 2

$\tilde{c}_2$	i_2	f_2	c_1
$[-0.9, 1.0]^\top$	$[1.0, 1.0]^\top$	$[0.1, 0.9]^\top$	$[0.8, 0.0]^\top$

- Message

$$c_2 = f_2 \circ c_1 + i_2 \circ \tilde{c}_2 \tag{93}$$

(94)

- Hidden

Forwards at time step 2

$\tilde{c}_2$	i_2	f_2	c_1
$[-0.9, 1.0]^\top$	$[1.0, 1.0]^\top$	$[0.1, 0.9]^\top$	$[0.8, 0.0]^\top$

- Message

$$c_2 = f_2 \circ c_1 + i_2 \circ \tilde{c}_2 \tag{93}$$

$$= [0.1, 0.9]^\top \circ [0.8, 0.0]^\top + [1.0, 1.0]^\top \circ [-0.9, 1.0]^\top \tag{94}$$

$$\tag{95}$$

- Hidden

Forwards at time step 2

$\tilde{c}_2$	i_2	f_2	c_1
$[-0.9, 1.0]^\top$	$[1.0, 1.0]^\top$	$[0.1, 0.9]^\top$	$[0.8, 0.0]^\top$

- Message

$$
\begin{aligned}
c_2 &= f_2 \circ c_1 + i_2 \circ \tilde{c}_2 && (93)\\
&= [0.1, 0.9]^\top \circ [0.8, 0.0]^\top + [1.0, 1.0]^\top \circ [-0.9, 1.0]^\top && (94)\\
&= [-0.8, 1.0]^\top && (95)\\
& && (96)
\end{aligned}
$$

- Hidden

Forwards at time step 2

$\tilde{c}_2$	i_2	f_2	c_1
$[-0.9, 1.0]^\top$	$[1.0, 1.0]^\top$	$[0.1, 0.9]^\top$	$[0.8, 0.0]^\top$

- Message

$$c_2 = f_2 \circ c_1 + i_2 \circ \tilde{c}_2 \tag{93}$$
$$= [0.1, 0.9]^\top \circ [0.8, 0.0]^\top + [1.0, 1.0]^\top \circ [-0.9, 1.0]^\top \tag{94}$$
$$= [-0.8, 1.0]^\top \tag{95}$$
$$\tag{96}$$

- Hidden

$$h_2 = o_2 \circ \tanh(c_2) \tag{97}$$
$$\tag{98}$$

Forwards at time step 2

$\tilde{c}_2$	i_2	f_2	c_1
$[-0.9, 1.0]^\top$	$[1.0, 1.0]^\top$	$[0.1, 0.9]^\top$	$[0.8, 0.0]^\top$

- Message

$$c_2 = f_2 \circ c_1 + i_2 \circ \tilde{c}_2 \tag{93}$$
$$= [-0.8, 1.0]^\top \tag{94}$$
$$\tag{95}$$

- Hidden

$$h_2 = o_2 \circ \tanh(c_2) \tag{96}$$
$$= [1.0, 1.0]^\top \circ \tanh([-0.8, 1.0]^\top) \tag{97}$$
$$\tag{98}$$

Forwards at time step 2

$\tilde{c}_2$	i_2	f_2	c_1
$[-0.9, 1.0]^\top$	$[1.0, 1.0]^\top$	$[0.1, 0.9]^\top$	$[0.8, 0.0]^\top$

- Message

$$c_2 = f_2 \circ c_1 + i_2 \circ \tilde{c}_2 \tag{93}$$
$$= [-0.8, 1.0]^\top \tag{94}$$
$$\tag{95}$$

- Hidden

$$h_2 = o_2 \circ \tanh(c_2) \tag{96}$$
$$= [1.0, 1.0]^\top \circ \tanh([-0.8, 1.0]^\top) \tag{97}$$
$$= [-0.7, 0.8]^\top \tag{98}$$

Forwards at time step 2

$\tilde{c}_2$	i_2	f_2	c_1
$[-0.9, 1.0]^\top$	$[1.0, 1.0]^\top$	$[0.1, 0.9]^\top$	$[0.8, 0.0]^\top$

- Message

$$c_2 = f_2 \circ c_1 + i_2 \circ \tilde{c}_2 \tag{93}$$
$$= [-0.8, 1.0]^\top \tag{94}$$
$$\tag{95}$$

- Hidden

$$h_2 = o_2 \circ \tanh(c_2) \tag{96}$$
$$= [-0.7, 0.8]^\top \tag{97}$$

- Output target$_2 = [0.0, 1.0]^\top$

Next time step . . .

- $i_3 = [0.4, 0.0]^\top$
- $f_3 = [0.4, 0.6]^\top$
- $o_3 = [0.5, 0.5]^\top$
- $\tilde{c}_3 = [-1.0, -0.6]^\top$
- $c_3 = [-0.7, 0.6]^\top$
- $h_3 = [-0.3, 0.3]^\top$

- Classify target$_3 = [0.0, 1.0]^\top$