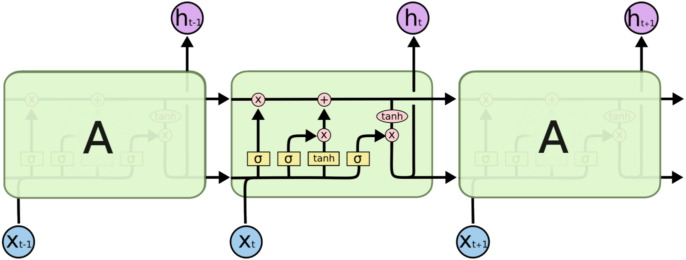
# Long Short Term Memory Networks

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 | |
|---|---|
| 1.0 | 0.0 |

| B | |
|---|---|
| 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 |
|---|
| 1.0  0.0 |

| B |
|---|
| 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:

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

## Model parameters for $x_t$

### Input's input gate

$$W_{ii} = \begin{bmatrix} 4 & 4 \\ 2 & 2 \end{bmatrix} \tag{1}$$

### forget gate

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

### cell params

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

### output gate

$$W_{io} = \begin{bmatrix} 5 & 5 \\ 3 & 5 \end{bmatrix} \tag{4}$$

Set all $b = 0$ for simplicity

## Model parameters for $h_t$

### input gate

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

### cell params

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

### forget gate

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

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

## Forwards at time step 1: $i_1$

**Input's input gate**

$$W_{ii} = \begin{bmatrix} 4 & 4 \\ 2 & 2 \end{bmatrix} \tag{9}$$

**input gate**

$$W_{hi} = \begin{bmatrix} 1 & 0 \\ 4 & -2 \end{bmatrix} \tag{10}$$

Compute

$$i_1 = \sigma(W_{ii} x_1 + W_{hi} h_0) \tag{11}$$

$$\tag{12}$$

## Forwards at time step 1: $i_1$

### Input's input gate

$$W_{ii} = \begin{bmatrix} 4 & 4 \\ 2 & 2 \end{bmatrix} \tag{9}$$

### input gate

$$W_{hi} = \begin{bmatrix} 1 & 0 \\ 4 & -2 \end{bmatrix} \tag{10}$$

Compute

$$\begin{aligned} i_1 =& \sigma\left(W_{ii}x_1 + W_{hi}h_0\right) && (11) \\ =& \sigma\left(\begin{bmatrix} 4 & 4 \\ 2 & 2 \end{bmatrix} \times \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}\right) && (12) \\ & && (13) \end{aligned}$$

## Forwards at time step 1: $i_1$

### Input's input gate

$$W_{ii} = \begin{bmatrix} 4 & 4 \\ 2 & 2 \end{bmatrix} \tag{9}$$

### input gate

$$W_{hi} = \begin{bmatrix} 1 & 0 \\ 4 & -2 \end{bmatrix} \tag{10}$$

Compute

$$i_1 = \sigma\left(W_{ii}x_1 + W_{hi}h_0\right) \tag{11}$$

$$= \sigma\left(\begin{bmatrix} 4 & 4 \\ 2 & 2 \end{bmatrix} \times \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}\right) \tag{12}$$

$$= \sigma\left([4.0, 2.0]^\top\right) \tag{13}$$

$$\tag{14}$$

## Forwards at time step 1: $i_1$

### Input's input gate

$$W_{ii} = \begin{bmatrix} 4 & 4 \\ 2 & 2 \end{bmatrix} \tag{9}$$

### input gate

$$W_{hi} = \begin{bmatrix} 1 & 0 \\ 4 & -2 \end{bmatrix} \tag{10}$$

Compute

$$i_1 = \sigma(W_{ii}x_1 + W_{hi}h_0) \tag{11}$$

$$= \sigma\left(\begin{bmatrix} 4 & 4 \\ 2 & 2 \end{bmatrix} \times \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}\right) \tag{12}$$

$$= \sigma([4.0, 2.0]^\top) \tag{13}$$

$$= [1.0, 0.9]^\top \tag{14}$$

## Forwards at time step 1: $f_1$

**forget gate**

$$W_{if} = \begin{bmatrix} -2 & 3 \\ 2 & 3 \end{bmatrix} \tag{15}$$

**forget gate**

$$W_{hf} = \begin{bmatrix} -1 & -2 \\ 0 & 0 \end{bmatrix} \tag{16}$$

Compute

$$f_1 = \sigma\left(W_{if} x_1 + W_{hf} h_0\right) \tag{17}$$

$$\tag{18}$$

## **Forwards at time step** 1: $f_1$

**forget gate**

$$W_{if} = \begin{bmatrix} -2 & 3 \\ 2 & 3 \end{bmatrix} \tag{15}$$

**forget gate**

$$W_{hf} = \begin{bmatrix} -1 & -2 \\ 0 & 0 \end{bmatrix} \tag{16}$$

Compute

$$f_1 = \sigma\left(W_{if}x_1 + W_{hf}h_0\right) \tag{17}$$

$$= \sigma\left(\begin{bmatrix} -2 & 3 \\ 2 & 3 \end{bmatrix} \times \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}\right) \tag{18}$$

$$\tag{19}$$

## Forwards at time step 1: $f_1$

**forget gate**

$$W_{if} = \begin{bmatrix} -2 & 3 \\ 2 & 3 \end{bmatrix} \tag{15}$$

**forget gate**

$$W_{hf} = \begin{bmatrix} -1 & -2 \\ 0 & 0 \end{bmatrix} \tag{16}$$

Compute

$$f_1 = \sigma\left(W_{if} x_1 + W_{hf} h_0\right) \tag{17}$$

$$= \sigma\left(\begin{bmatrix} -2 & 3 \\ 2 & 3 \end{bmatrix} \times \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}\right) \tag{18}$$

$$= \sigma\left([-2.0, 2.0]^\top\right) \tag{19}$$

$$\tag{20}$$

## Forwards at time step 1: $f_1$

**forget gate**

$$W_{if} = \begin{bmatrix} -2 & 3 \\ 2 & 3 \end{bmatrix} \tag{15}$$

**forget gate**

$$W_{hf} = \begin{bmatrix} -1 & -2 \\ 0 & 0 \end{bmatrix} \tag{16}$$

Compute

$$
\begin{aligned}
f_1 =& \sigma\left(W_{if} x_1 + W_{hf} h_0\right) && (17) \\
=& \sigma\left(\begin{bmatrix} -2 & 3 \\ 2 & 3 \end{bmatrix} \times \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}\right) && (18) \\
=& \sigma\left([-2.0, 2.0]^\top\right) && (19) \\
=& [0.1, 0.9]^\top && (20)
\end{aligned}
$$

## Forwards at time step 1

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

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

$$= \sigma([5.0, 3.0]^\top)$$

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

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

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

$$= \tanh([1.0, 0.0]^\top)$$

$$= [0.8, 0.0]^\top$$

## Forwards at time step 1

- $c_1 = f_1 * c_0 + i_1 * \tilde{c}_1$ $\qquad (c_0 = [0.0, 0.0]^\top)$

  $= [1.0, 0.9]^\top * [0.8, 0.0]^\top$

  $= [0.8, 0.0]^\top$

- $h_1 = o_1 * \tanh(c_1)$

  $= [1.0, 1.0]^\top * \tanh([0.8, 0.0]^\top)$

  $= [0.7, 0.0]^\top$

- $y_1 = \mathrm{softmax}(h_1)$

- successfully classify $\text{target}_1 = [1.0, 0.0]^\top$

## Forwards at time step 2

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

## Forwards at time step 2

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

## Forwards at time step 2

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

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

## Forwards at time step 2

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

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

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

## Forwards at time step 2

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

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

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

$$= \sigma([4.7, 4.8]^\top)$$

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

## Forwards at time step 2

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

## Forwards at time step 2

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

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

## Forwards at time step 2

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

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

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

## Forwards at time step 2

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

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

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

$$= \sigma([-2.7, 2.0]^\top)$$

$$= [0.1, 0.9]^\top$$

## Forwards at time step 2

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

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

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

$$= \sigma([5.7, 4.4]^\top)$$

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

## Forwards at time step 2

- $\tilde{c}_2 = \tanh(W_{ic}x_2 + W_{hc}h_1)$

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

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

$$= \tanh([-1.8, 2.8]^\top)$$

$$= [-0.9, 1.0]^\top$$

## Forwards at time step 2

- $c_2 = f_2 * c_1 + i_2 * \tilde{c}_2$
- $h_2 = o_2 * \tanh(c_2)$

## Forwards at time step 2

- $c_2 = f_2 * c_1 + i_2 * \tilde{c}_2$

  $= [0.1, 0.9]^\top * [0.8, 0.0]^\top + [1.0, 1.0]^\top * [-0.9, 1.0]^\top$

  $= [-0.8, 1.0]^\top$

- $h_2 = o_2 * \tanh(c_2)$

## Forwards at time step 2

- $c_2 = f_2 * c_1 + i_2 * \tilde{c}_2$

  $= [0.1, 0.9]^\top * [0.8, 0.0]^\top + [1.0, 1.0]^\top * [-0.9, 1.0]^\top$

  $= [-0.8, 1.0]^\top$

- $h_2 = o_2 * \tanh(c_2)$

  $= [1.0, 1.0]^\top * \tanh([-0.8, 1.0]^\top)$

  $= [-0.7, 0.8]^\top$

- successfully classify $\text{target}_2 = [0.0, 1.0]^\top$

## Keep forwarding in time...

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

- successfully classify $\text{target}_3 = [0.0, 1.0]^\top$

## Caveats

- 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:

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