Practical Machine Learning

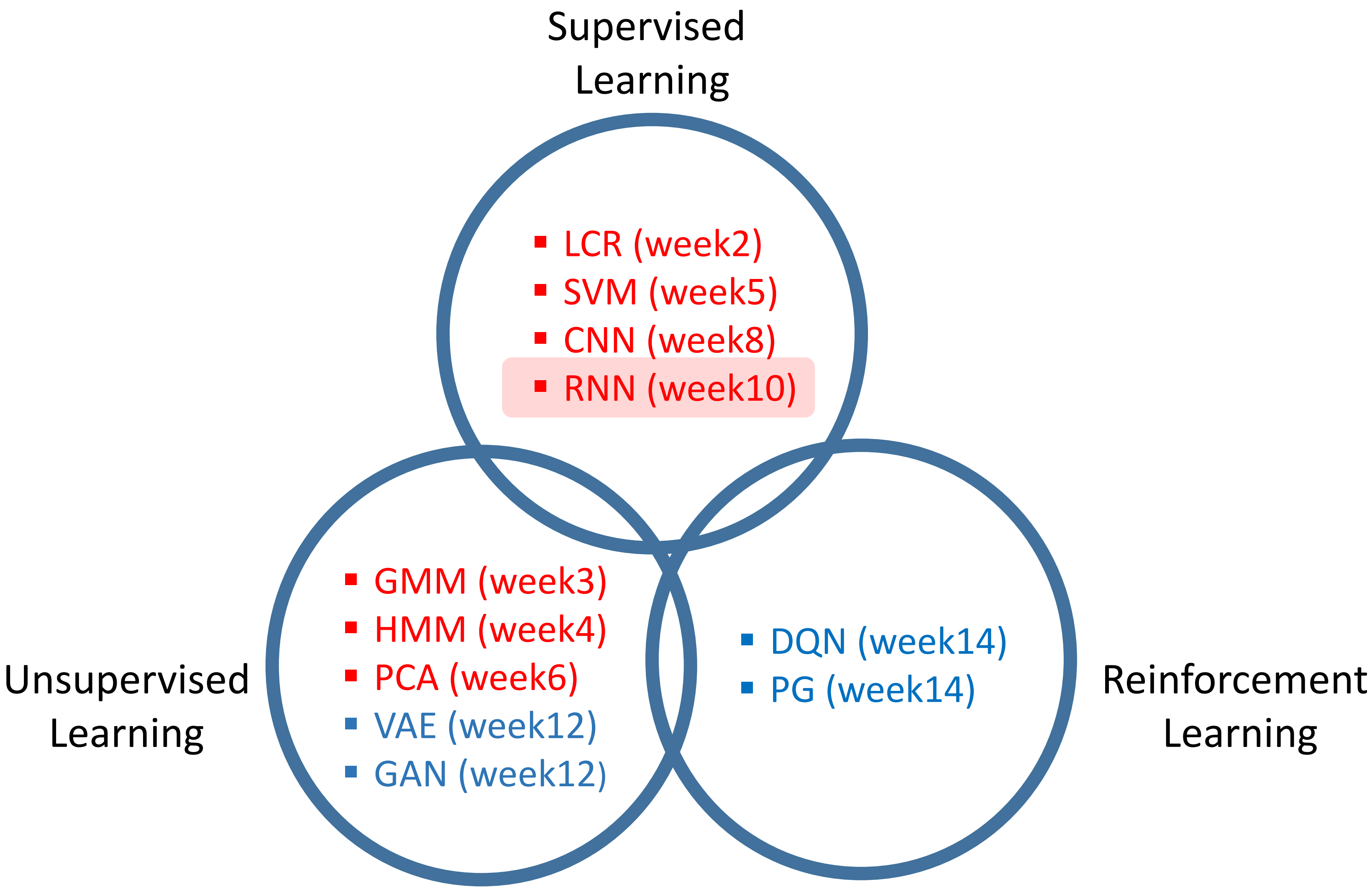
Lecture 10

Recurrent Neural Networks (RNN) & Long Short Term Memory (LSTM)

Dr. Suyong Eum

OSAKA UNIVERSITY

Where we are


Supervised Learning
LCR (week2)
SVM (week5)
CNN (week8)
RNN (week10)
Unsupervised Learning
GMM (week3)
HMM (week4)
PCA (week6)
VAE (week12)
GAN (week12)
Reinforcement Learning
DQN (week14)
PG (week14)

You are going to learn

- ❑ Recurrent Neural Networks (RNN)
- ❑ Long Short Term Memory (LSTM)
- ❑ Backpropagation in RNN and LSTM

Recurrent Neural Networks (RNN)

Recurrent Neural Network (RNN)

- ❑ Feed forward neural networks
  - Independent
  - Fixed length

- ❑ Recurrent Neural Networks
  - Temporal dependencies
  - Variable sequence length

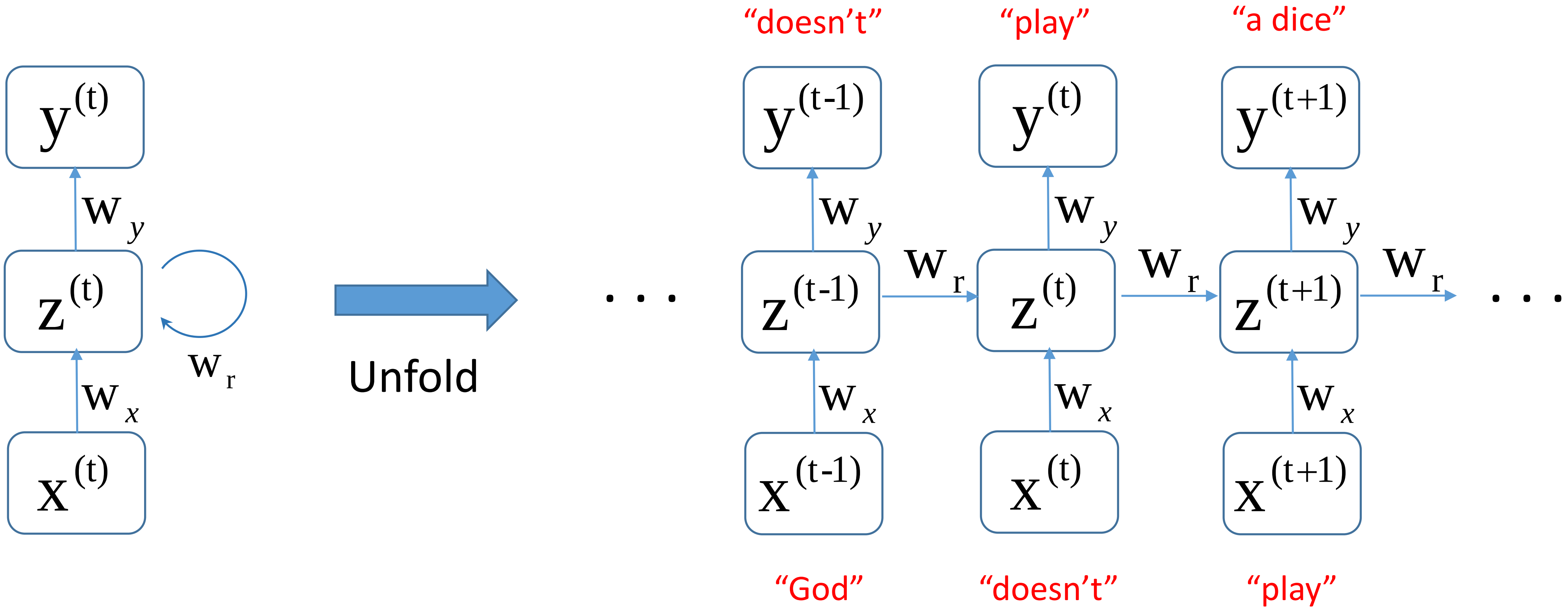
Some interesting applications

1. Music composition
   - http://people.idsia.ch/~juergen/blues/

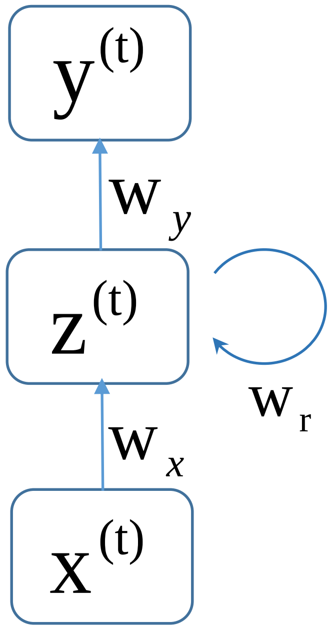
2. Writing a poem
   - https://github.com/dvictor/lstm-poetry

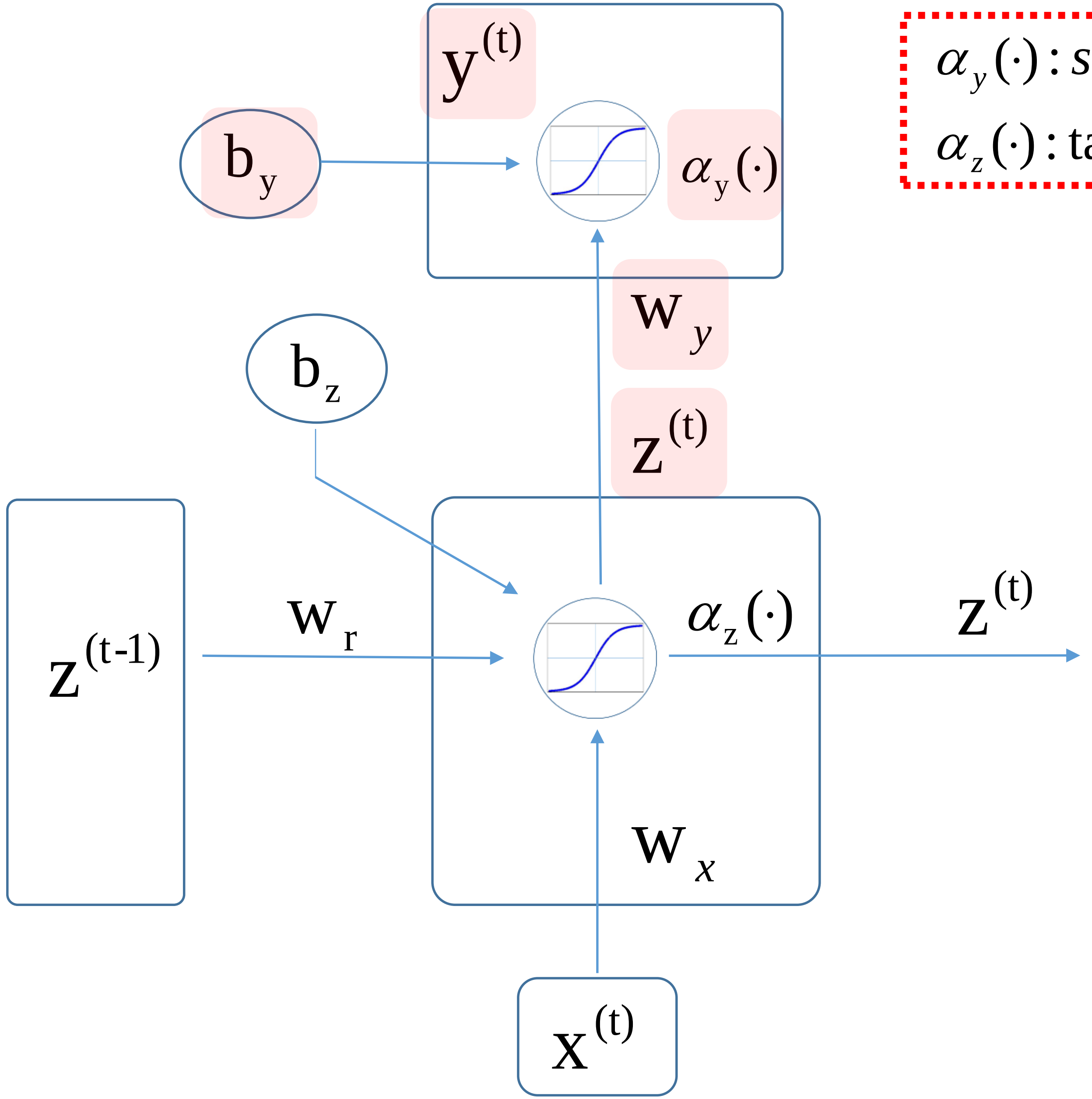
A butterfly in the sun
Just because I know that I should leave this heart for you
You said I was falling apart
I wish I were you
I wanted you to know how I feel
I could have settled it all
It's time to go and do it big and you can be my side
I can't believe it when I see you
I'm lost in the world and I can't see you cry
I'm asking you to love me then let me go
I can't stop this way

Recurrent Neural Network (RNN)



Recurrent Neural Network (RNN): inner structure





$\alpha_y(\cdot): soft\max$

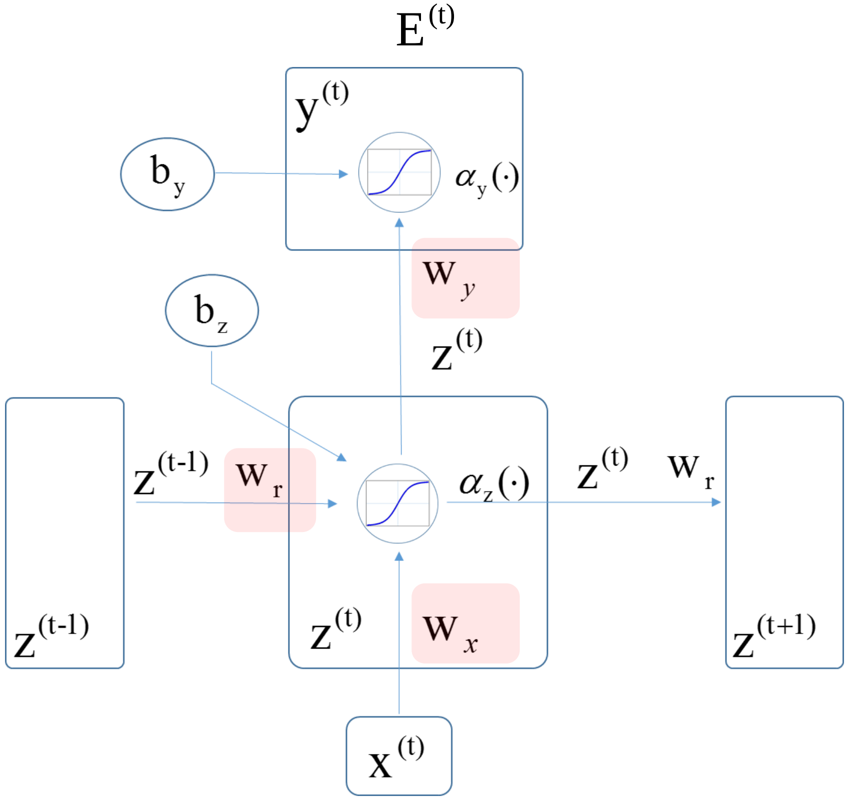
$\alpha_z(\cdot): \tanh$

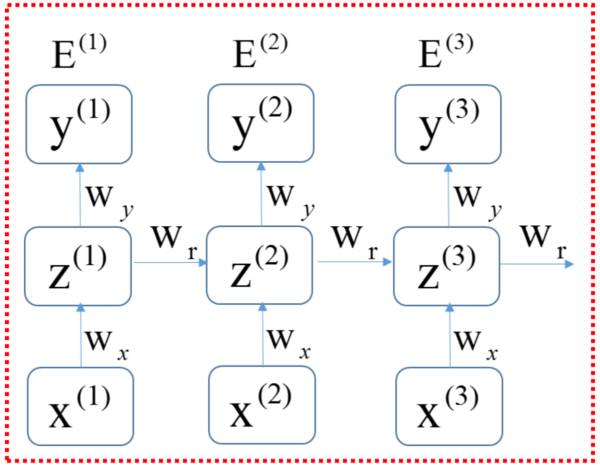
$$z^{(t)} = \alpha_z\left(w_x x^{(t)} + w_r z^{(t-1)} + b_z\right)$$

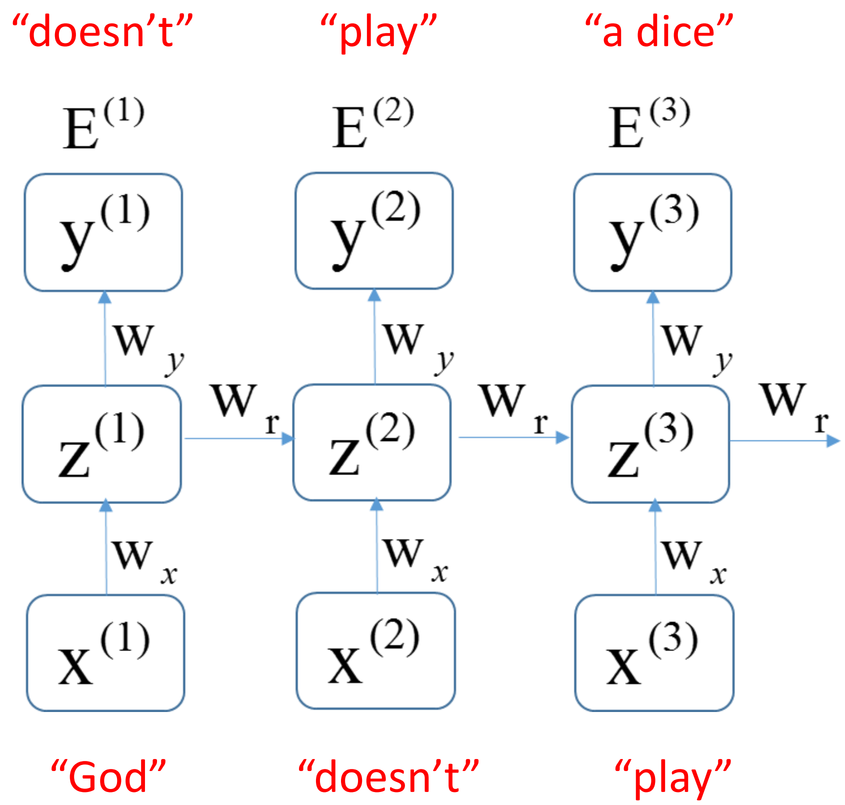
$$y^{(t)} = \alpha_y\left(w_y z^{(t)} + b_y\right)$$

Backpropagation through time (BPTT) in RNN

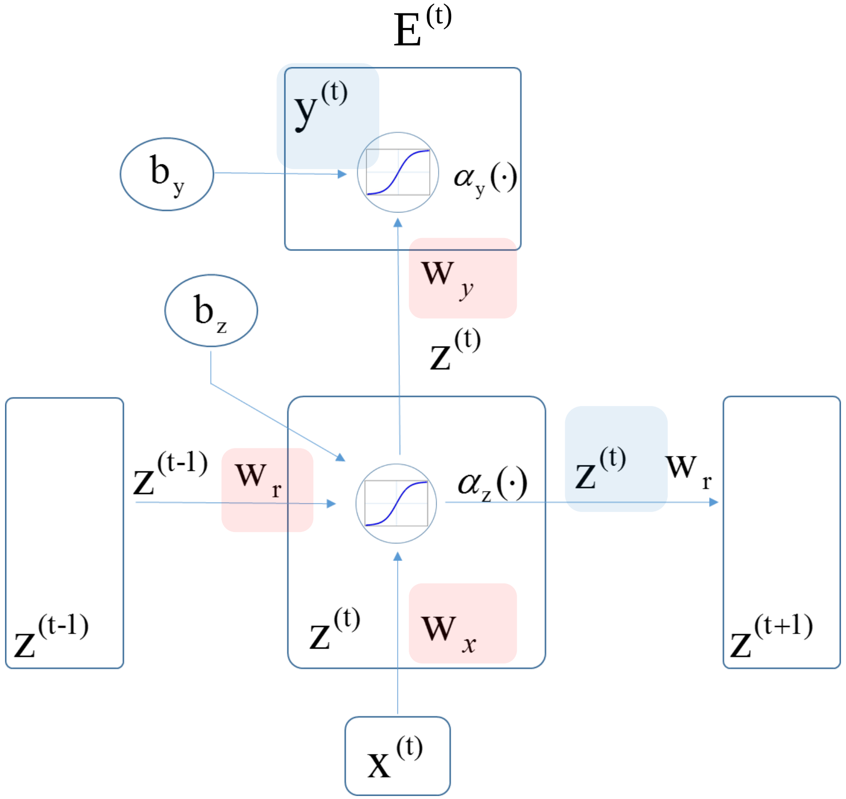
Recurrent Neural Network (RNN): Backpropagation through time (BPTT)

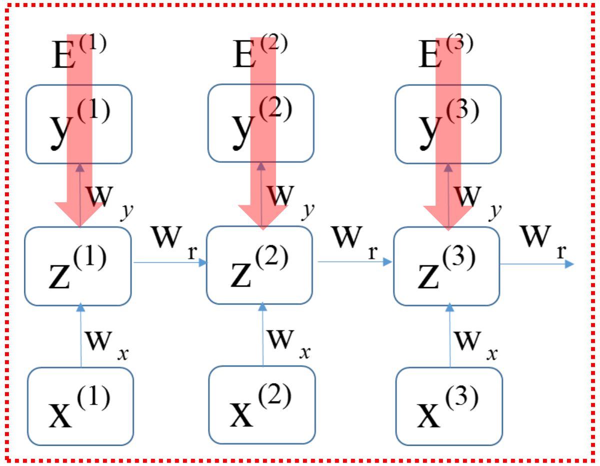






Recurrent Neural Network (RNN): Backpropagation through time (BPTT)





$$z^{(t)} = \alpha_z\left(w_x x^{(t)} + w_r z^{(t-1)} + b_z\right)$$
$$y^{(t)} = \alpha_y\left(w_y z^{(t)} + b_y\right)$$

$$s_z^{(t)} = w_x x^{(t)} + w_r z^{(t-1)} + b_z$$
$$s_y^{(t)} = w_y z^{(t)} + b_y$$

$$\frac{\partial E}{\partial w_y} = \sum_{t=1}^{3} \frac{\partial E^{(t)}}{\partial w_y} = \frac{\partial E^{(1)}}{\partial w_y} + \frac{\partial E^{(2)}}{\partial w_y} + \frac{\partial E^{(3)}}{\partial w_y}$$

$$\frac{\partial E}{\partial w_r} = \sum_{t=1}^{3} \frac{\partial E^{(t)}}{\partial w_r} = \frac{\partial E^{(1)}}{\partial w_r} + \frac{\partial E^{(2)}}{\partial w_r} + \frac{\partial E^{(3)}}{\partial w_r}$$

$$\frac{\partial E}{\partial w_x} = \sum_{t=1}^{3} \frac{\partial E^{(t)}}{\partial w_x} = \frac{\partial E^{(1)}}{\partial w_x} + \frac{\partial E^{(2)}}{\partial w_x} + \frac{\partial E^{(3)}}{\partial w_x}$$

$$\frac{\partial E}{\partial w_y} = \sum_{t=1}^{3} \frac{\partial E^{(t)}}{\partial w_y} = \frac{\partial E^{(1)}}{\partial w_y} + \frac{\partial E^{(2)}}{\partial w_y} + \frac{\partial E^{(3)}}{\partial w_y}$$

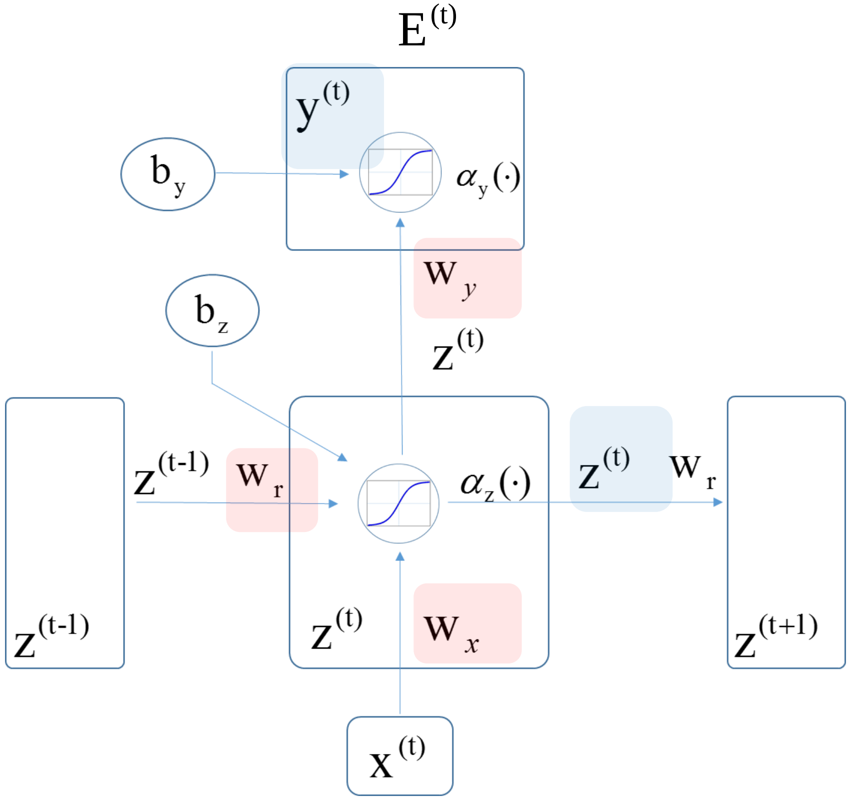
$$\frac{\partial E^{(1)}}{\partial w_y} = \boxed{\frac{\partial E^{(1)}}{\partial y^{(1)}}}\boxed{\frac{\partial y^{(1)}}{\partial s_y^{(1)}}}\boxed{\frac{\partial s_y^{(1)}}{\partial w_y}} \qquad \frac{\partial E^{(2)}}{\partial w_y} = \frac{\partial E^{(2)}}{\partial y^{(2)}}\frac{\partial y^{(2)}}{\partial s_y^{(2)}}\frac{\partial s_y^{(2)}}{\partial w_y} \qquad \frac{\partial E^{(3)}}{\partial w_y} = \frac{\partial E^{(3)}}{\partial y^{(3)}}\frac{\partial y^{(3)}}{\partial s_y^{(3)}}\frac{\partial s_y^{(3)}}{\partial w_y}$$

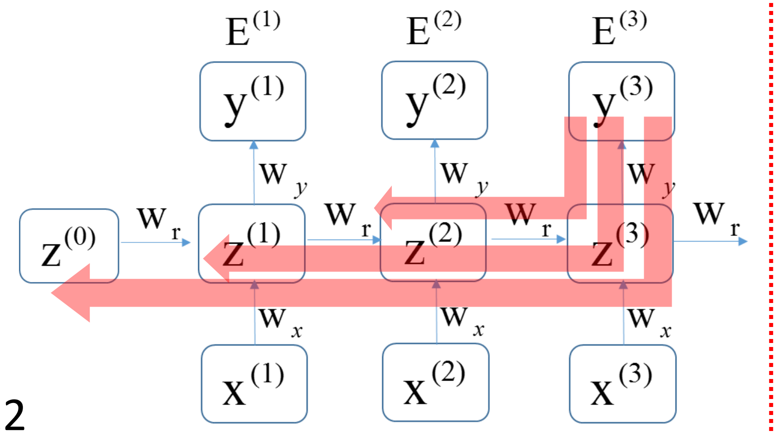
Linear input function

Activation function

Error function

Recurrent Neural Network (RNN): Backpropagation through time (BPTT)





$$z^{(t)} = \alpha_z\left(w_x x^{(t)} + w_r z^{(t-1)} + b_z\right)$$
$$y^{(t)} = \alpha_y\left(w_y z^{(t)} + b_y\right)$$

$$s_z^{(t)} = w_x x^{(t)} + w_r z^{(t-1)} + b_z$$
$$s_y^{(t)} = w_y z^{(t)} + b_y$$

$$\frac{\partial E}{\partial w_y} = \sum_{t=1}^{3} \frac{\partial E^{(t)}}{\partial w_y} = \frac{\partial E^{(1)}}{\partial w_y} + \frac{\partial E^{(2)}}{\partial w_y} + \frac{\partial E^{(3)}}{\partial w_y}$$

$$\frac{\partial E}{\partial w_r} = \sum_{t=1}^{3} \frac{\partial E^{(t)}}{\partial w_r} = \frac{\partial E^{(1)}}{\partial w_r} + \frac{\partial E^{(2)}}{\partial w_r} + \frac{\partial E^{(3)}}{\partial w_r}$$

$$\frac{\partial E}{\partial w_x} = \sum_{t=1}^{3} \frac{\partial E^{(t)}}{\partial w_x} = \frac{\partial E^{(1)}}{\partial w_x} + \frac{\partial E^{(2)}}{\partial w_x} + \frac{\partial E^{(3)}}{\partial w_x}$$

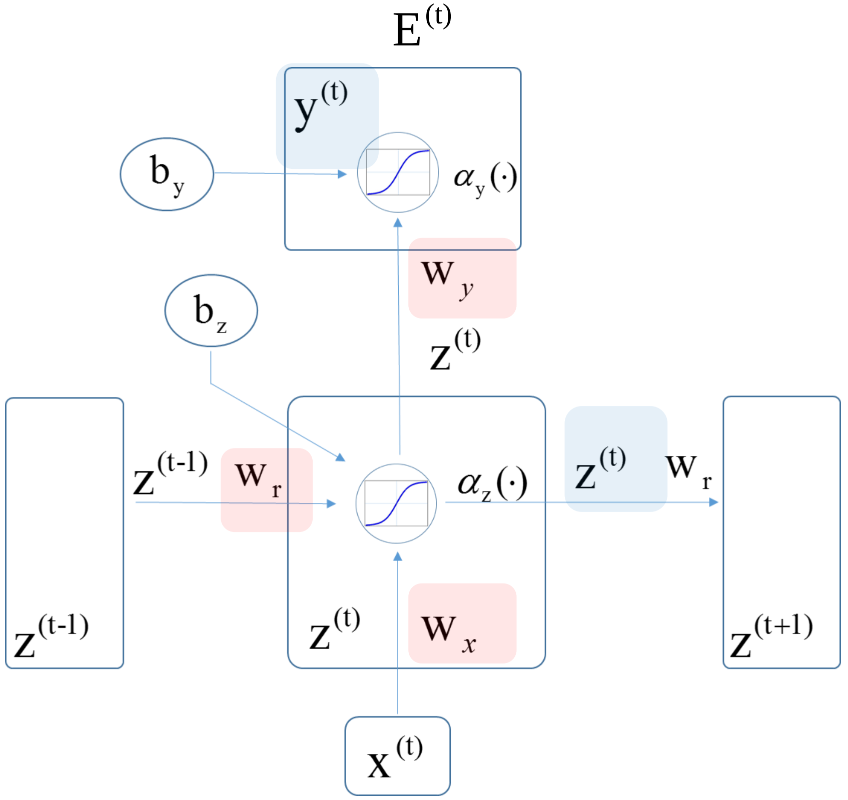
$$\frac{\partial E}{\partial w_r} = \sum_{t=1}^{3} \frac{\partial E^{(t)}}{\partial w_r} = \frac{\partial E^{(1)}}{\partial w_r} + \frac{\partial E^{(2)}}{\partial w_r} + \frac{\partial E^{(3)}}{\partial w_r}$$

$$\sum_{t=1}^{3} \frac{\partial E^{(t)}}{\partial w_r} = \boxed{\frac{\partial E^{(3)}}{\partial y^{(3)}} \frac{\partial y^{(3)}}{\partial s_y^{(3)}} \frac{\partial s_y^{(3)}}{\partial z^{(3)}} \frac{\partial z^{(3)}}{\partial s_z^{(3)}} \frac{\partial s_z^{(3)}}{\partial w_r}} \quad t = 3$$

$$+ \frac{\partial E^{(3)}}{\partial y^{(3)}} \frac{\partial y^{(3)}}{\partial s_y^{(3)}} \frac{\partial s_y^{(3)}}{\partial z^{(3)}} \frac{\partial z^{(3)}}{\partial s_z^{(3)}} \frac{\partial s_z^{(3)}}{\partial z^{(2)}} \boxed{\frac{\partial z^{(2)}}{\partial s_z^{(2)}} \frac{\partial s_z^{(2)}}{\partial w_r}} \quad t = 2$$

$$+ \frac{\partial E^{(3)}}{\partial y^{(3)}} \frac{\partial y^{(3)}}{\partial s_y^{(3)}} \frac{\partial s_y^{(3)}}{\partial z^{(3)}} \frac{\partial z^{(3)}}{\partial s_z^{(3)}} \frac{\partial s_z^{(3)}}{\partial z^{(2)}} \frac{\partial z^{(2)}}{\partial s_z^{(2)}} \frac{\partial s_z^{(2)}}{\partial z^{(1)}} \boxed{\frac{\partial z^{(1)}}{\partial s_z^{(1)}} \frac{\partial s_z^{(1)}}{\partial w_r}} \quad t = 1$$

Recurrent Neural Network (RNN): Backpropagation through time (BPTT)



$$z^{(t)} = \alpha_z\left(w_x x^{(t)} + w_r z^{(t-1)} + b_z\right)$$
$$y^{(t)} = \alpha_y\left(w_y z^{(t)} + b_y\right)$$

$$s_z^{(t)} = w_x x^{(t)} + w_r z^{(t-1)} + b_z$$
$$s_y^{(t)} = w_y z^{(t)} + b_y$$

$$\frac{\partial E}{\partial w_y} = \sum_{t=1}^{3} \frac{\partial E^{(t)}}{\partial w_y} = \frac{\partial E^{(1)}}{\partial w_y} + \frac{\partial E^{(2)}}{\partial w_y} + \frac{\partial E^{(3)}}{\partial w_y}$$

$$\frac{\partial E}{\partial w_r} = \sum_{t=1}^{3} \frac{\partial E^{(t)}}{\partial w_r} = \frac{\partial E^{(1)}}{\partial w_r} + \frac{\partial E^{(2)}}{\partial w_r} + \frac{\partial E^{(3)}}{\partial w_r}$$

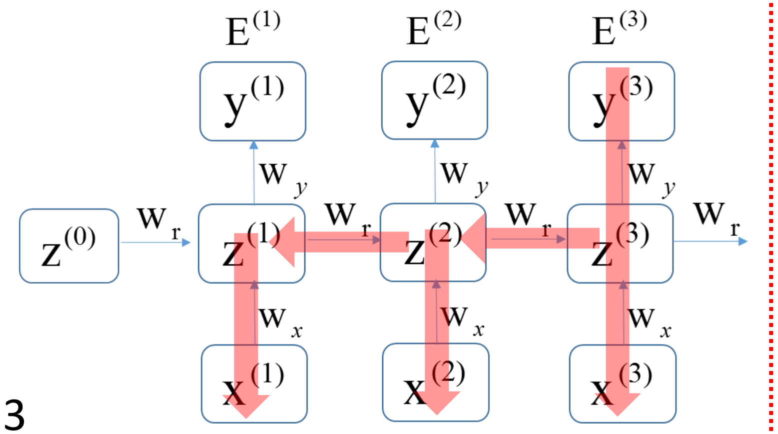
$$\frac{\partial E}{\partial w_x} = \sum_{t=1}^{3} \frac{\partial E^{(t)}}{\partial w_x} = \frac{\partial E^{(1)}}{\partial w_x} + \frac{\partial E^{(2)}}{\partial w_x} + \frac{\partial E^{(3)}}{\partial w_x}$$

$$\frac{\partial E}{\partial w_x} = \sum_{t=1}^{3} \frac{\partial E^{(t)}}{\partial w_x} = \frac{\partial E^{(1)}}{\partial w_x} + \frac{\partial E^{(2)}}{\partial w_x} + \frac{\partial E^{(3)}}{\partial w_x}$$

$$\sum_{t=1}^{3} \frac{\partial E^{(t)}}{\partial w_r} = \boxed{\frac{\partial E^{(3)}}{\partial y^{(3)}} \frac{\partial y^{(3)}}{\partial s_y^{(3)}} \frac{\partial s_y^{(3)}}{\partial z^{(3)}} \frac{\partial z^{(3)}}{\partial s_z^{(3)}} \frac{\partial s_z^{(3)}}{\partial w_x}} \quad t = 3$$

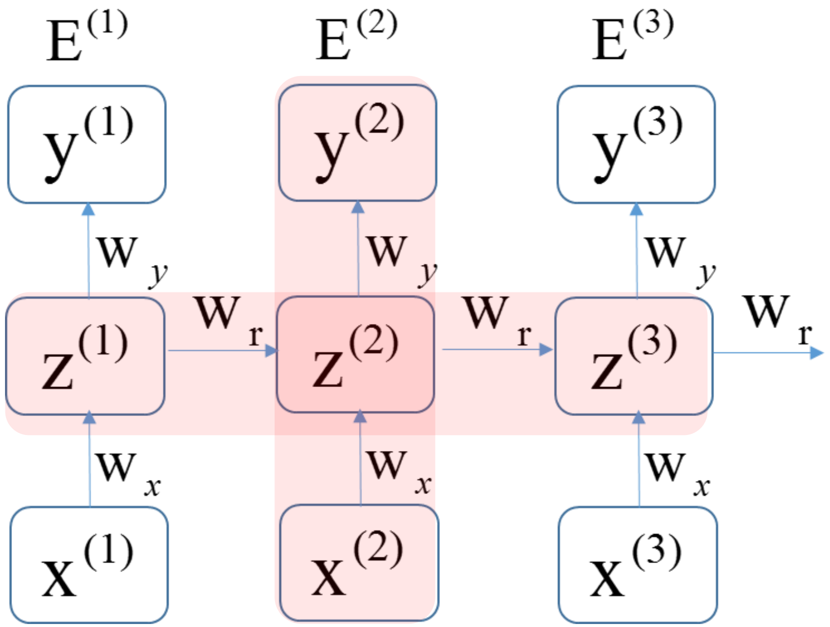
$$+ \frac{\partial E^{(3)}}{\partial y^{(3)}} \frac{\partial y^{(3)}}{\partial s_y^{(3)}} \frac{\partial s_y^{(3)}}{\partial z^{(3)}} \frac{\partial z^{(3)}}{\partial s_z^{(3)}} \frac{\partial s_z^{(3)}}{\partial z^{(2)}} \boxed{\frac{\partial z^{(2)}}{\partial s_z^{(2)}} \frac{\partial s_z^{(2)}}{\partial w_x}} \quad t = 2$$

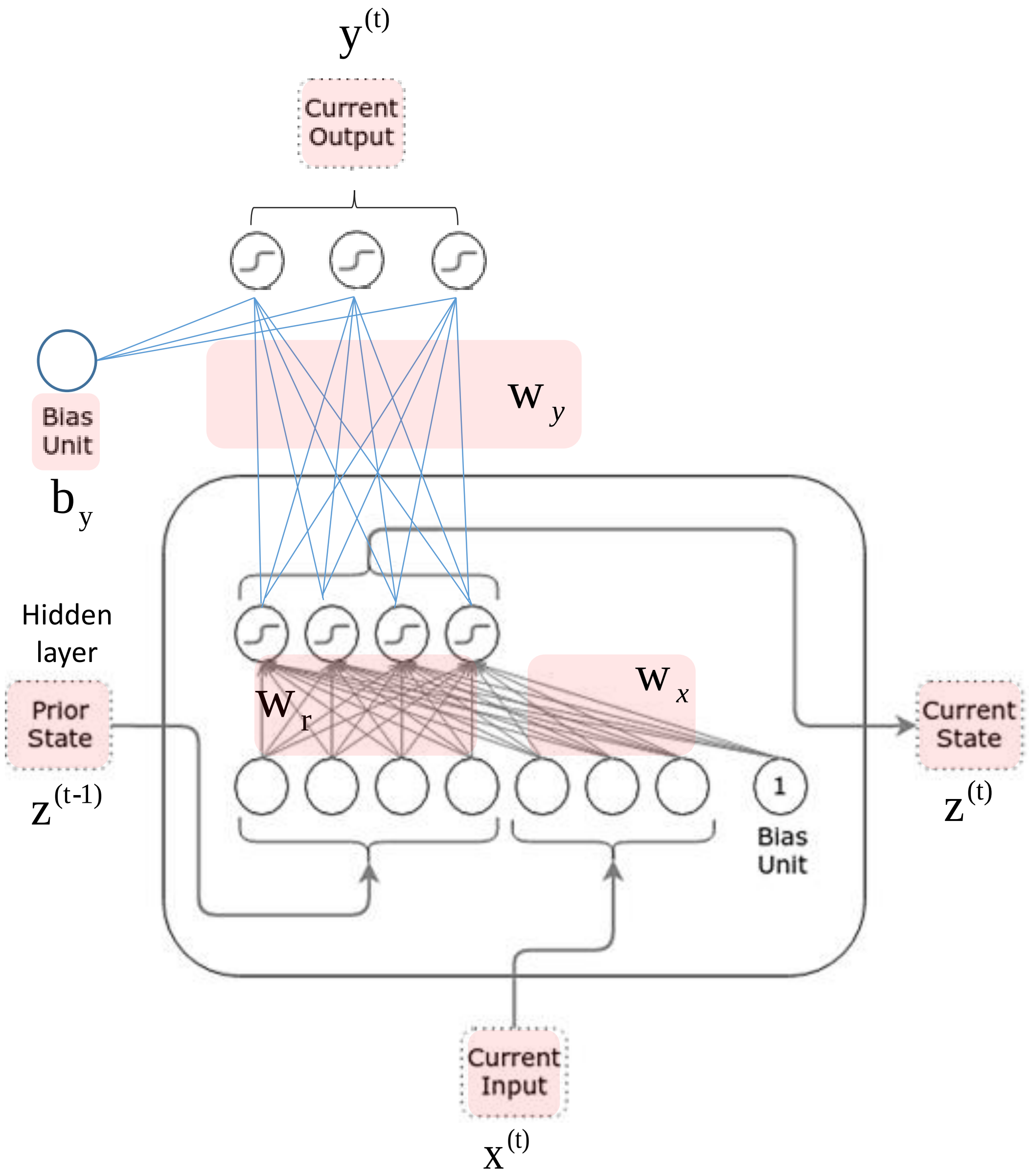
$$+ \frac{\partial E^{(3)}}{\partial y^{(3)}} \frac{\partial y^{(3)}}{\partial s_y^{(3)}} \frac{\partial s_y^{(3)}}{\partial z^{(3)}} \frac{\partial z^{(3)}}{\partial s_z^{(3)}} \frac{\partial s_z^{(3)}}{\partial z^{(2)}} \frac{\partial z^{(2)}}{\partial s_z^{(2)}} \frac{\partial s_z^{(2)}}{\partial z^{(1)}} \boxed{\frac{\partial z^{(1)}}{\partial s_z^{(1)}} \frac{\partial s_z^{(1)}}{\partial w_x}} \quad t = 1$$



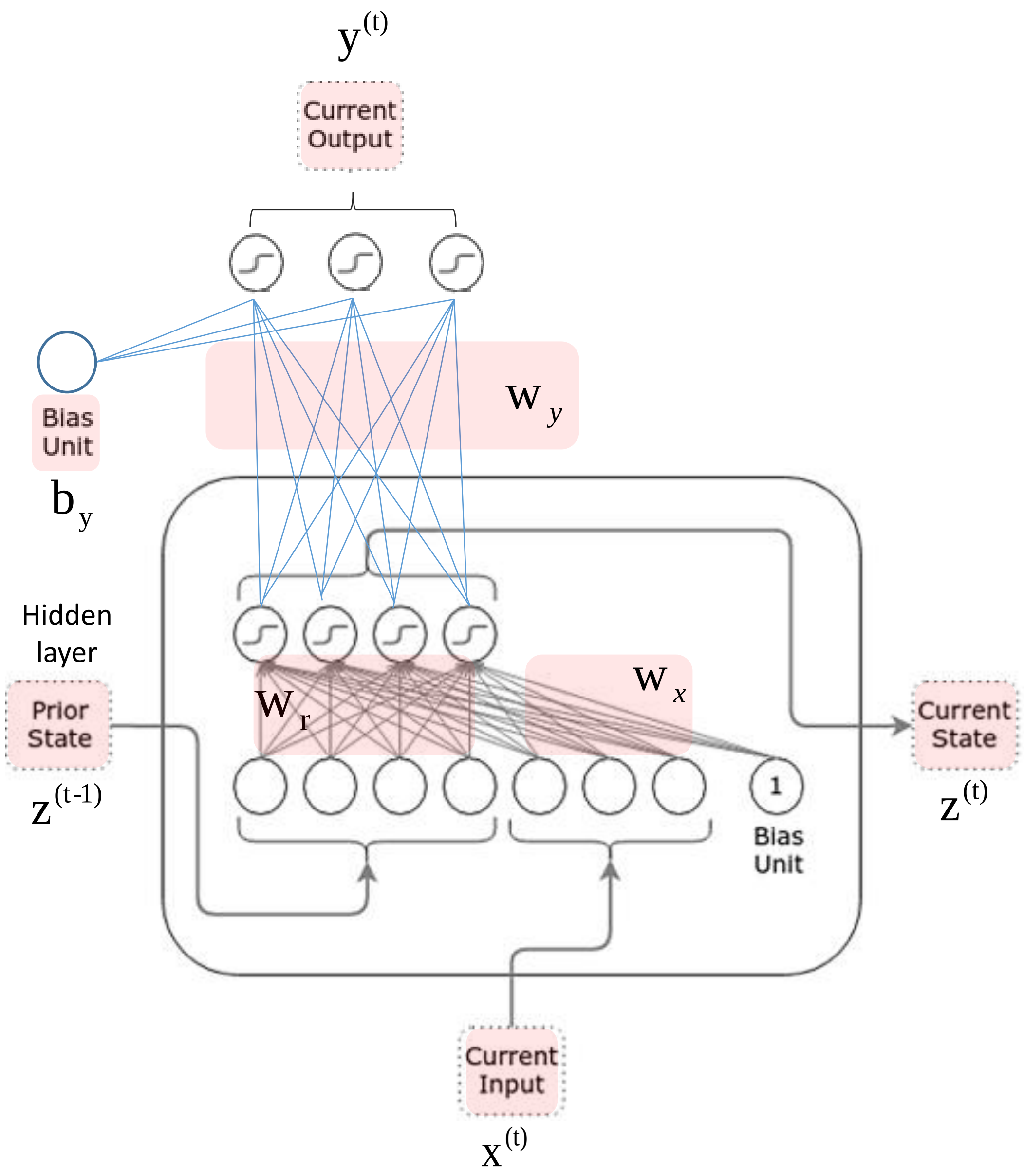
Connectivity of neurons in vanilla RNN component

Operation of RNN: Connectivity of neurons in vanilla RNN component





Operation of RNN: Connectivity of neurons in vanilla RNN component



where:

$z^{(t-1)}, z^{(t)} \in R^n$ $\quad W_r \in R^{n\times n}$ $\quad b_z \in R^n$

$x^{(t)} \in R^m$ $\quad W_x \in R^{n\times m}$ $\quad b_y \in R^k$

$y^{(t)} \in R^k$ $\quad W_y \in R^{k\times n}$

$$z^{(t)} = \alpha_z\left(w_x x^{(t)} + w_r z^{(t-1)} + b_z\right)$$

(n x 1) = (n x m) (m x 1) + (n x n) (n x 1) + (n x 1)

$$y^{(t)} = \alpha_y\left(w_y z^{(t)} + b_y\right)$$

(k x 1) = (k x n) (n x 1) + (k x 1)

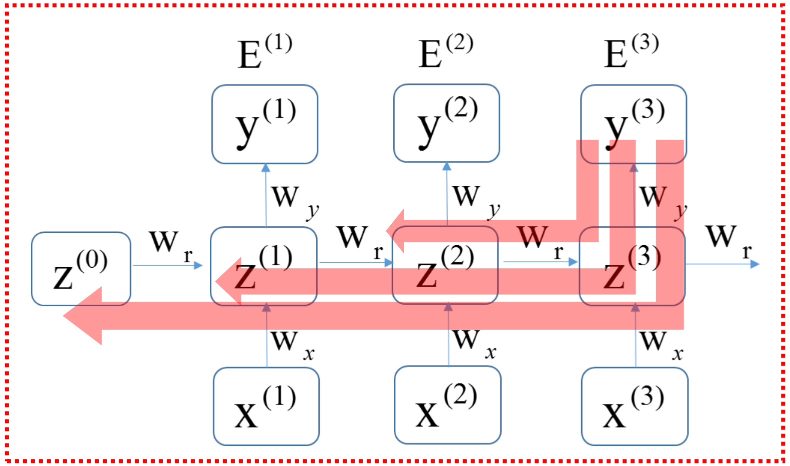
Long Short Term Memory (LSTM)

Long Short Term Memory (LSTM) network

- *Long Short Term Memory (LSTM) architecture was motivated to overcome the problem: error is not back-propagated properly to the end of RNN architecture.*

- *Hochreiter and Schmidhuber (1997) proposed the Long Short-Term Memory (LSTM) cell which includes "a memory unit":*
  1) *A cell with a number of components that together act similar to a memory cell.*
  2) *Inside one cell, multiple layers called "gates" are used.*
     1) *Forget gate*
     2) *Input gate*
     3) *Output gate*

RNN using BPTT: Vanishing and exploding problems



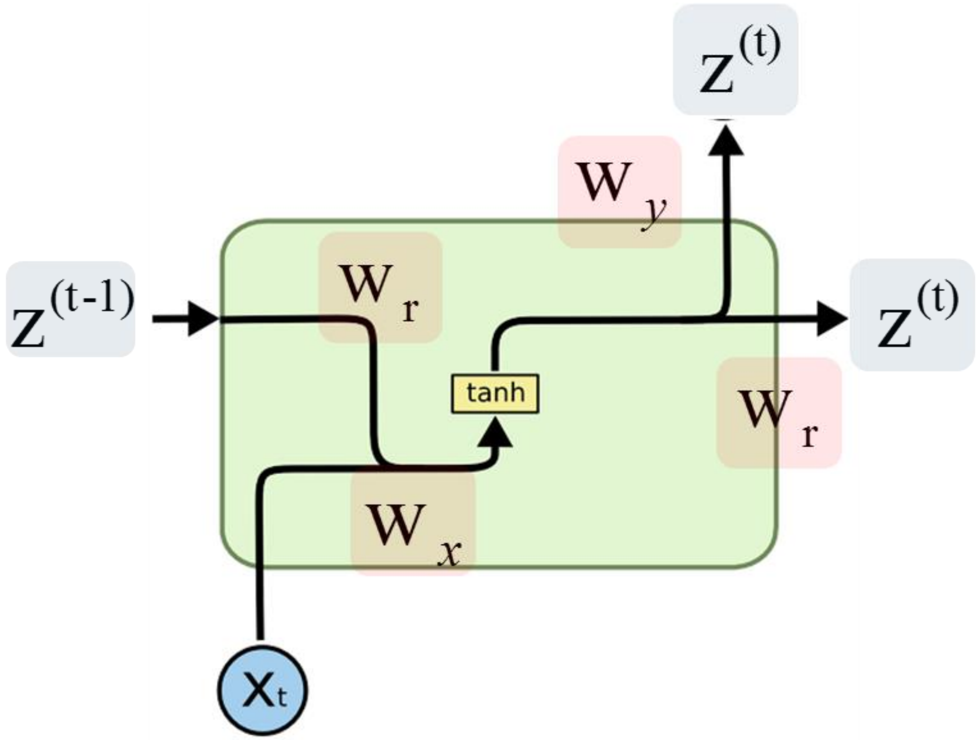
$$\sum_{t=1}^{3}\frac{\partial E^{(t)}}{\partial w_r}=\frac{\partial E^{(3)}}{\partial y^{(3)}}\frac{\partial y^{(3)}}{\partial s_y^{(3)}}\frac{\partial s_y^{(3)}}{\partial z^{(3)}}\frac{\partial z^{(3)}}{\partial s_z^{(3)}}\frac{\partial s_z^{(3)}}{\partial w_r}$$

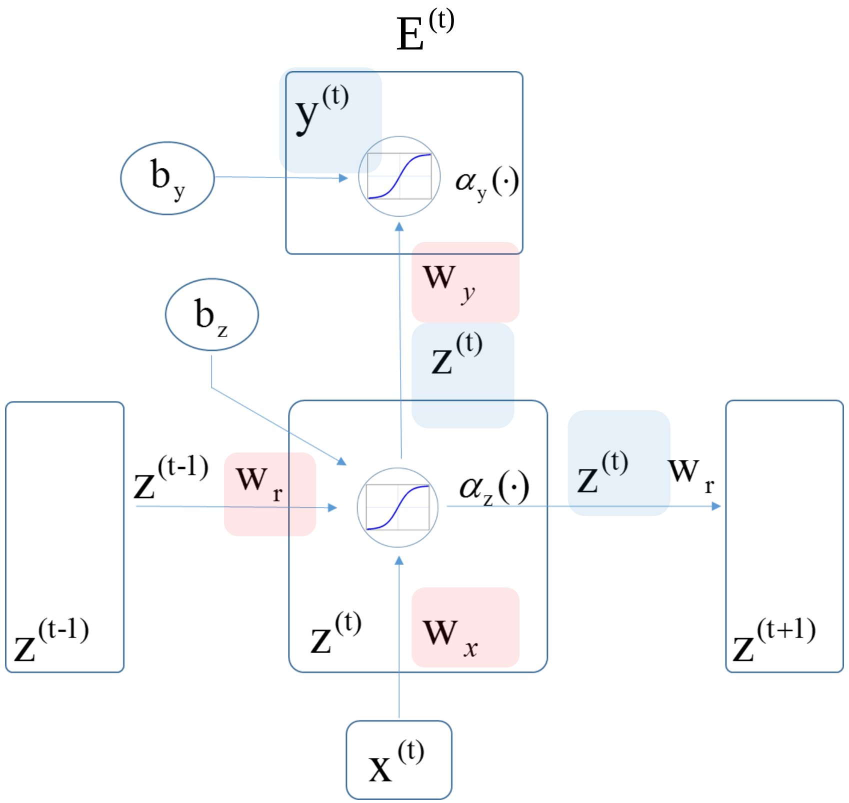
$$+\frac{\partial E^{(3)}}{\partial y^{(3)}}\frac{\partial y^{(3)}}{\partial s_y^{(3)}}\frac{\partial s_y^{(3)}}{\partial z^{(3)}}\frac{\partial z^{(3)}}{\partial s_z^{(3)}}\frac{\partial s_z^{(3)}}{\partial z^{(2)}}\frac{\partial z^{(2)}}{\partial s_z^{(2)}}\frac{\partial s_z^{(2)}}{\partial w_r}$$

$$+\frac{\partial E^{(3)}}{\partial y^{(3)}}\frac{\partial y^{(3)}}{\partial s_y^{(3)}}\overbrace{\frac{\partial s_y^{(3)}}{\partial z^{(3)}}}^{\mathrm{W}_r}\boxed{\frac{\partial z^{(3)}}{\partial s_z^{(3)}}}\overbrace{\frac{\partial s_z^{(3)}}{\partial z^{(2)}}}^{\mathrm{W}_r}\boxed{\frac{\partial z^{(2)}}{\partial s_z^{(2)}}}\overbrace{\frac{\partial s_z^{(2)}}{\partial z^{(1)}}}^{\mathrm{W}_r}\boxed{\frac{\partial z^{(1)}}{\partial s_z^{(1)}}}\overbrace{\frac{\partial s_z^{(1)}}{\partial w_r}}^{\mathrm{z}^{(0)}}$$

Activation function (tanh) $\quad \frac{d\sigma(a)}{da}=1-\sigma(a)^2$

Long Short Term Memory (LSTM) network



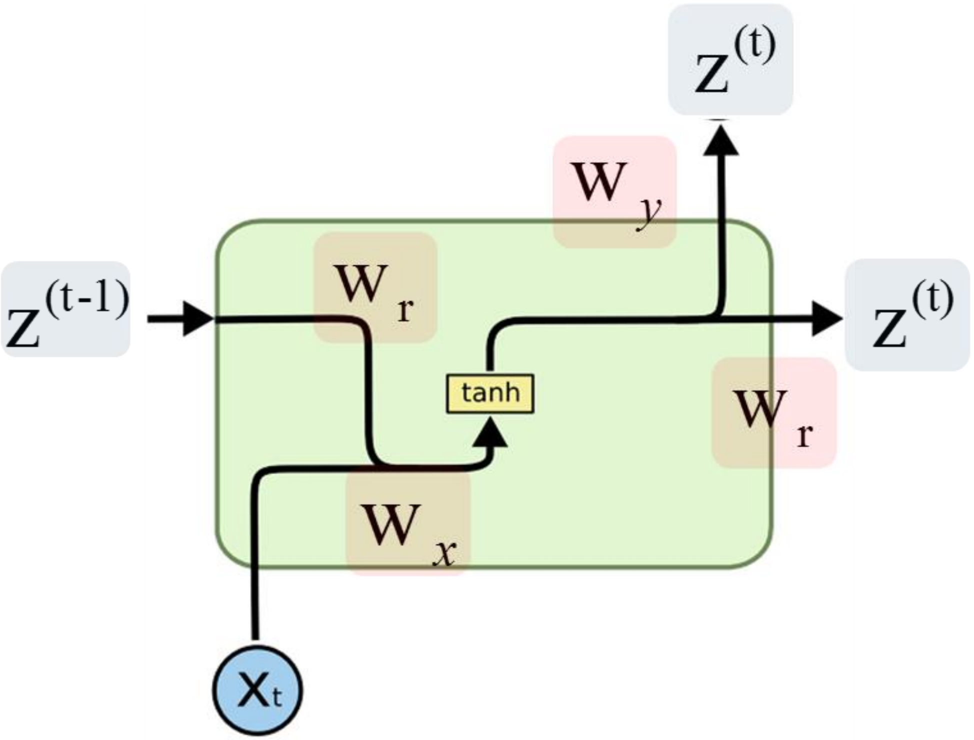


$$z^{(t)} = \alpha_z\left(w_x x^{(t)} + w_r z^{(t-1)} + b_z\right)$$

$$y^{(t)} = \alpha_y\left(w_y z^{(t)} + b_y\right)$$

RNN

Long Short Term Memory (LSTM) network

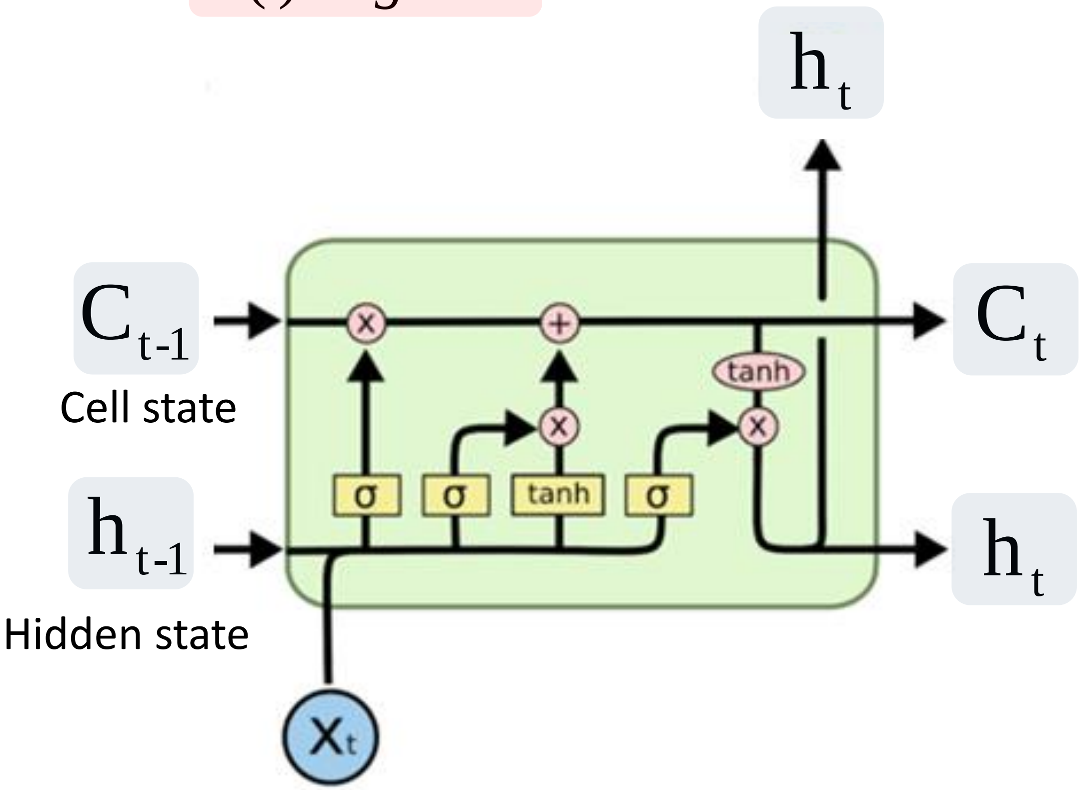


$$z^{(t)} = \alpha_z\left(w_x x^{(t)} + w_r z^{(t-1)} + b_z\right)$$

$$y^{(t)} = \alpha_y\left(w_y z^{(t)} + b_y\right)$$

RNN

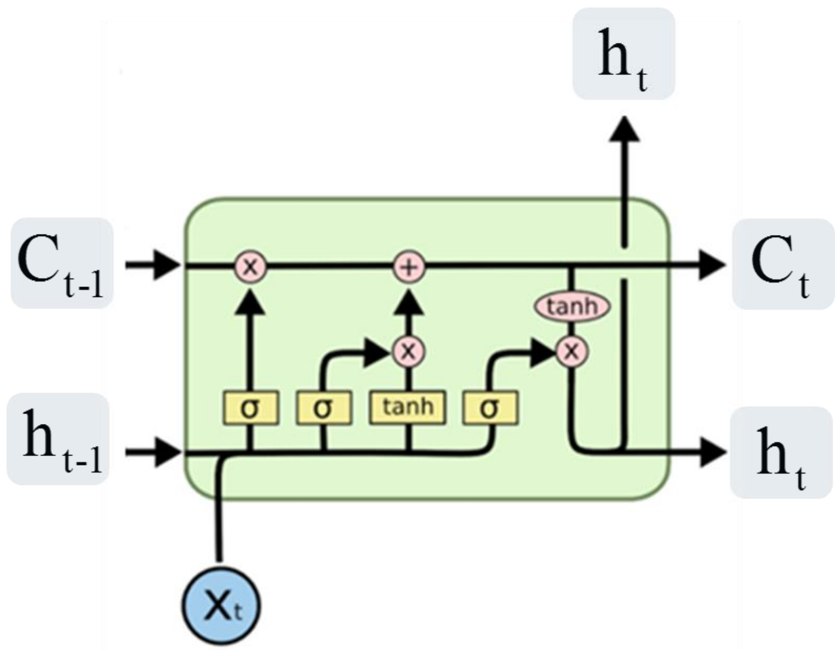
$\sigma(\cdot): sigmoid$



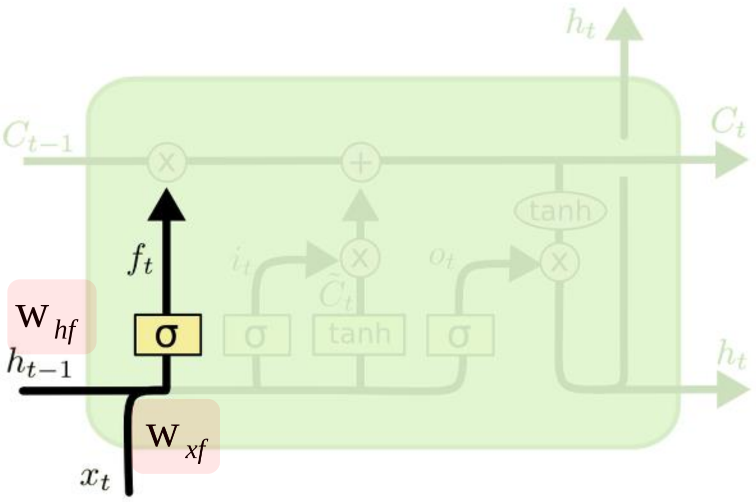
- ❑ *Cell state: a memory of LSTM cell*
- ❑ *Hidden state: an output of this cell*

LSTM

Long Short Term Memory (LSTM) network: forget gate



LSTM

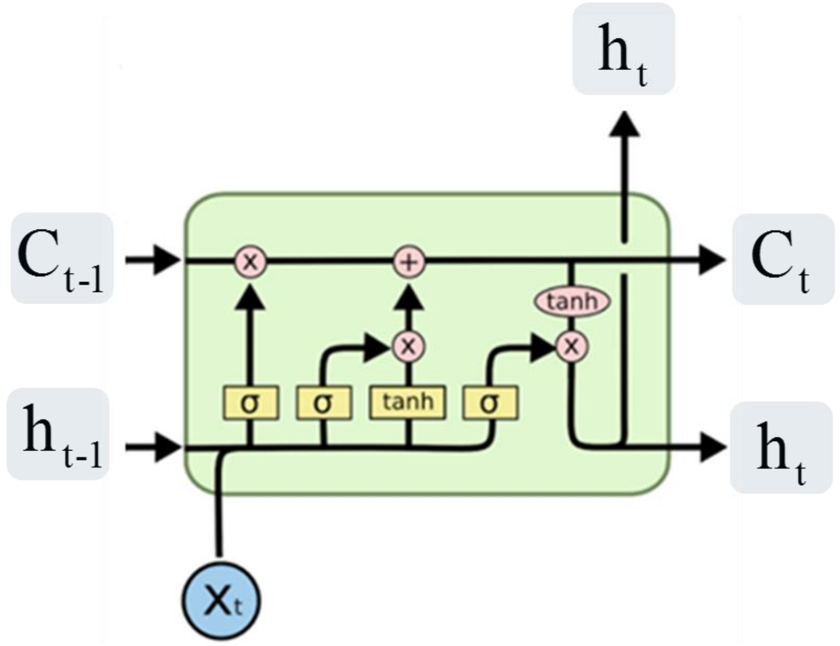


$\sigma(\cdot): sigmoid$

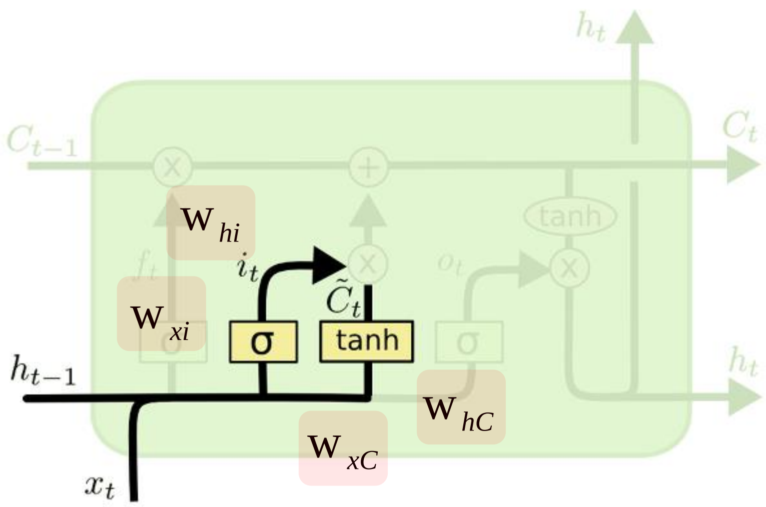
$$f_t = sigm(\mathrm{w}_{\mathrm{xf}}\mathrm{x}_{\mathrm{t}} + \mathrm{w}_{\mathrm{hf}}\mathrm{h}_{\mathrm{t\text{-}1}} + b_f)$$

- *Forget gate layer*
  - Decide how much "C_{t-1}" is forgotten?
  - If "f_t" is zero, forget "C_{t-1}" completely.
  - If "f_t" is one, do not forget "C_{t-1}" at all.

Long Short Term Memory (LSTM) network: input gate



LSTM



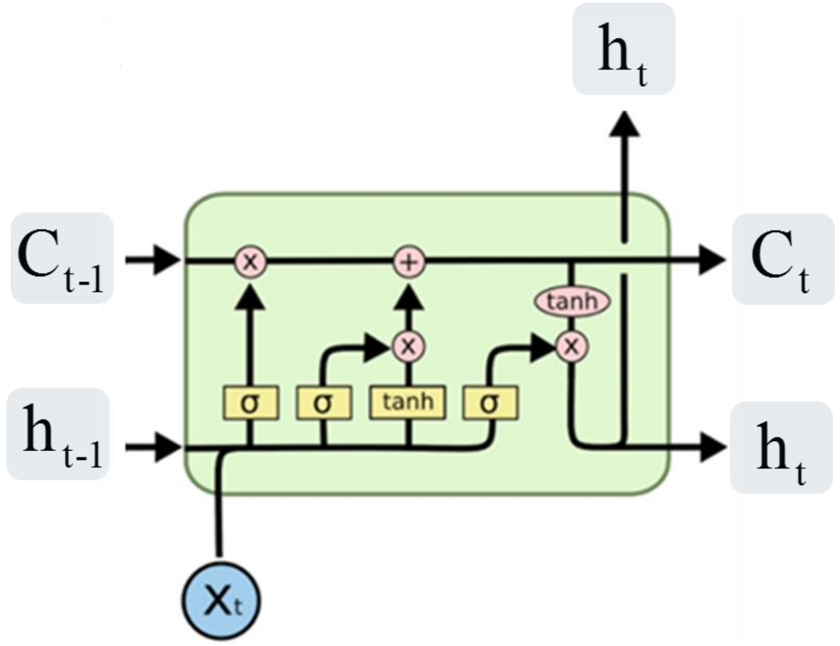
$\sigma(\cdot): sigmoid$

$$i_t = sigm(\mathrm{w}_{\mathrm{xi}}\mathrm{x}_{\mathrm{t}} + \mathrm{w}_{\mathrm{hi}}\mathrm{h}_{\mathrm{t\text{-}1}} + b_i)$$

$$\tilde{C}_t = \tanh(\mathrm{w}_{\mathrm{xC}}\mathrm{x}_{\mathrm{t}} + \mathrm{w}_{\mathrm{hC}}\mathrm{h}_{\mathrm{t\text{-}1}} + b_C)$$

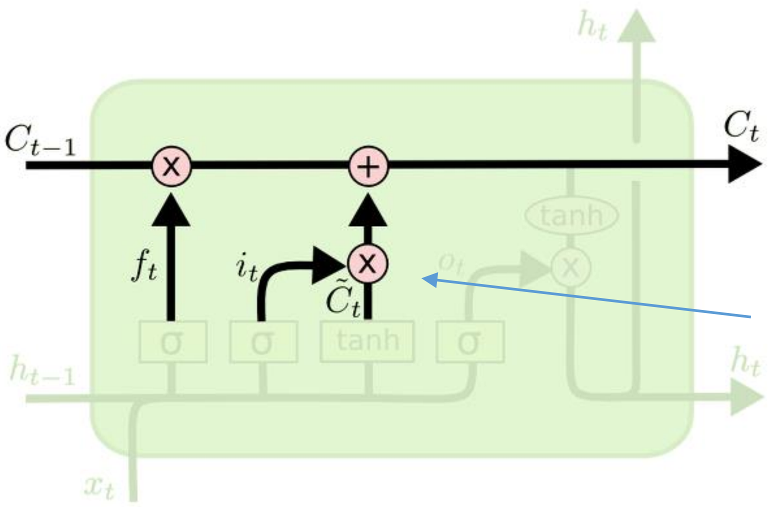
- *Input gate layer*
  - decide how much "$\tilde{C}_t$" is forgotten?
- *tanh layer:*
  - *decide which value is updated.*

Long Short Term Memory (LSTM) network: update cell



LSTM

$\sigma(\cdot): sigmoid$



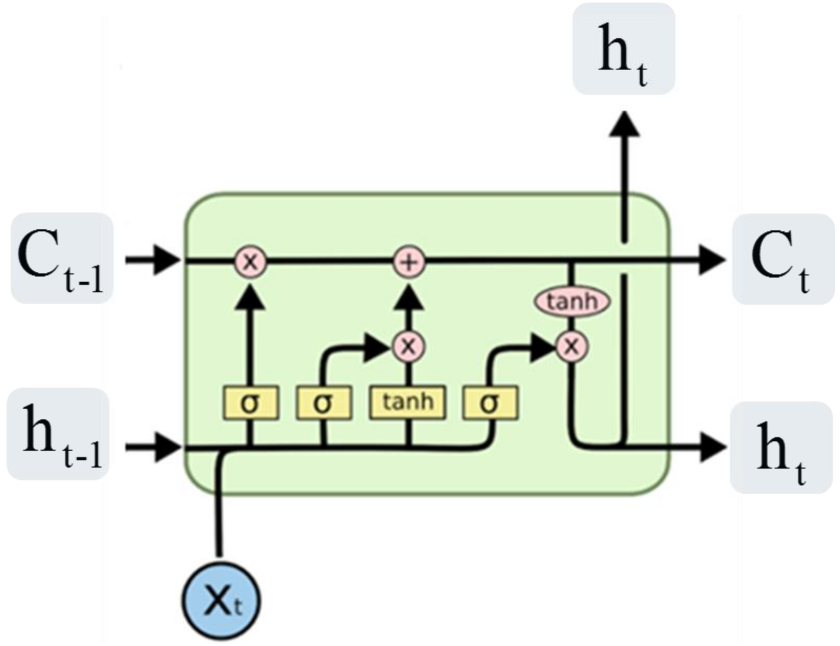
Hadamard product

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} * \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} x_1 y_1 \\ x_2 y_2 \end{bmatrix}$$

$$C_t = \underbrace{f_t * C_{t-1}}_{\text{past memory}} + \underbrace{i_t * \tilde{C}_t}_{\text{present memory}}$$

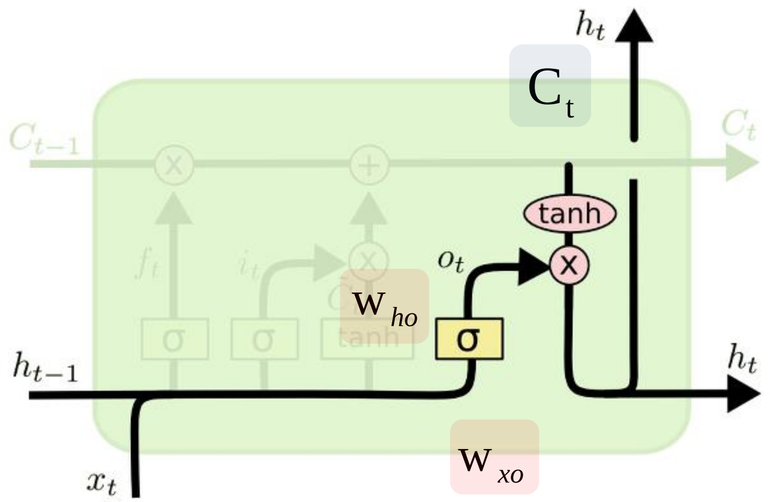
- *Update the output cell state of "C_t" by adding the past cell state of "C_{t-1}" to the present cell state of "$\tilde{C}_t$"*

Long Short Term Memory (LSTM) network: output gate



LSTM

$\sigma(\cdot): sigmoid$

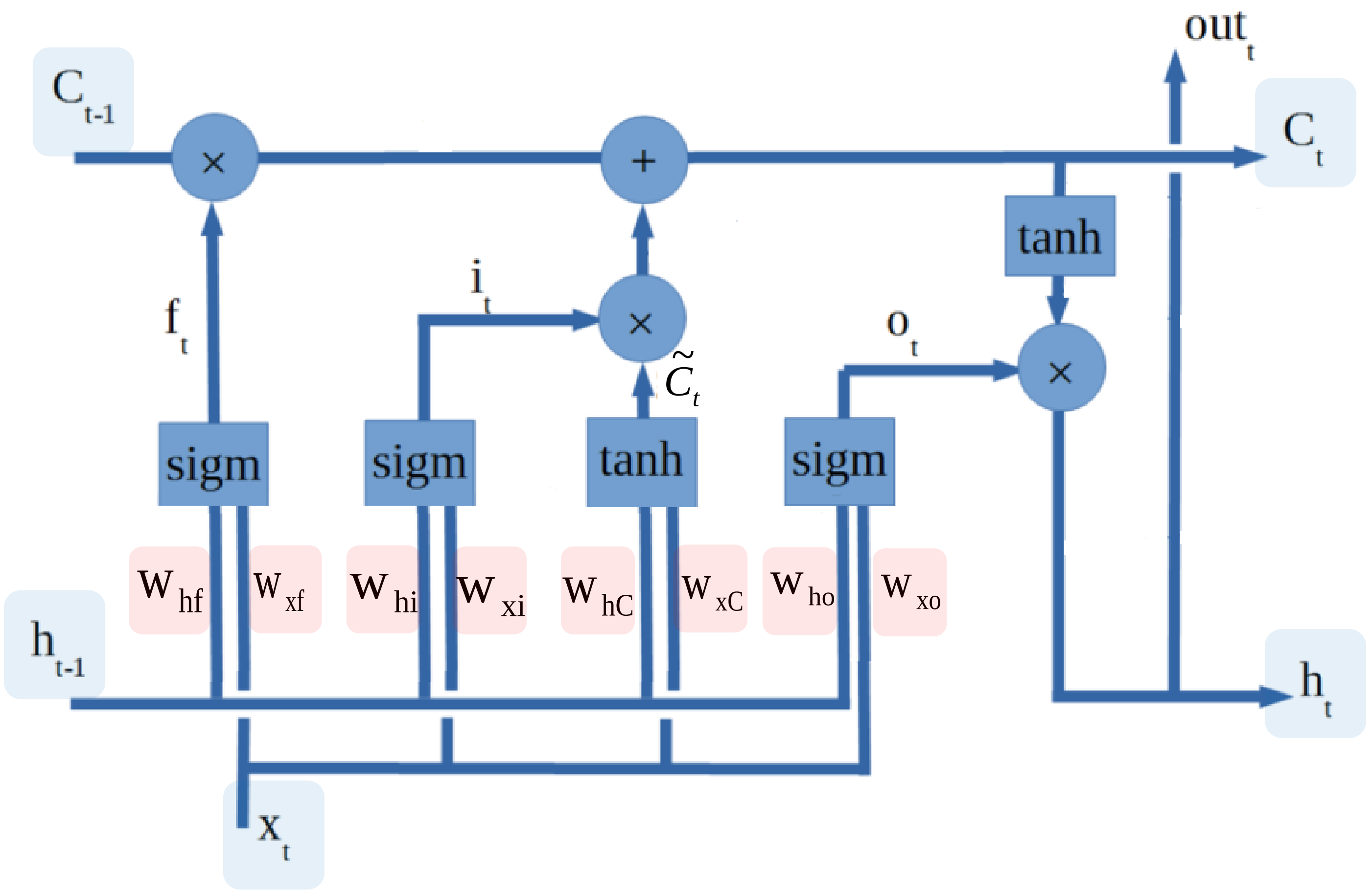


$$o_t = sigm(\mathrm{w}_{xo}\mathrm{x}_t + \mathrm{w}_{ho}\mathrm{h}_{t\text{-}1} + b_o)$$

$$h_t = o_t * \tanh(C_t)$$

- ***Output gate layer***
  - *The output cell state is put through tanh() and rescaled by the output of the sigmoid function.*

Long Short Term Memory (LSTM) network: summary



$$f_t = sigm(\mathrm{w}_{\mathrm{xf}}\mathrm{x}_\mathrm{t} + \mathrm{w}_{\mathrm{hf}}\mathrm{h}_{\mathrm{t-1}} + b_f)$$

$$i_t = sigm(\mathrm{w}_{\mathrm{xi}}\mathrm{x}_\mathrm{t} + \mathrm{w}_{\mathrm{hi}}\mathrm{h}_{\mathrm{t-1}} + b_i)$$

$$o_t = sigm(\mathrm{w}_{\mathrm{xo}}\mathrm{x}_\mathrm{t} + \mathrm{w}_{\mathrm{ho}}\mathrm{h}_{\mathrm{t-1}} + b_o)$$

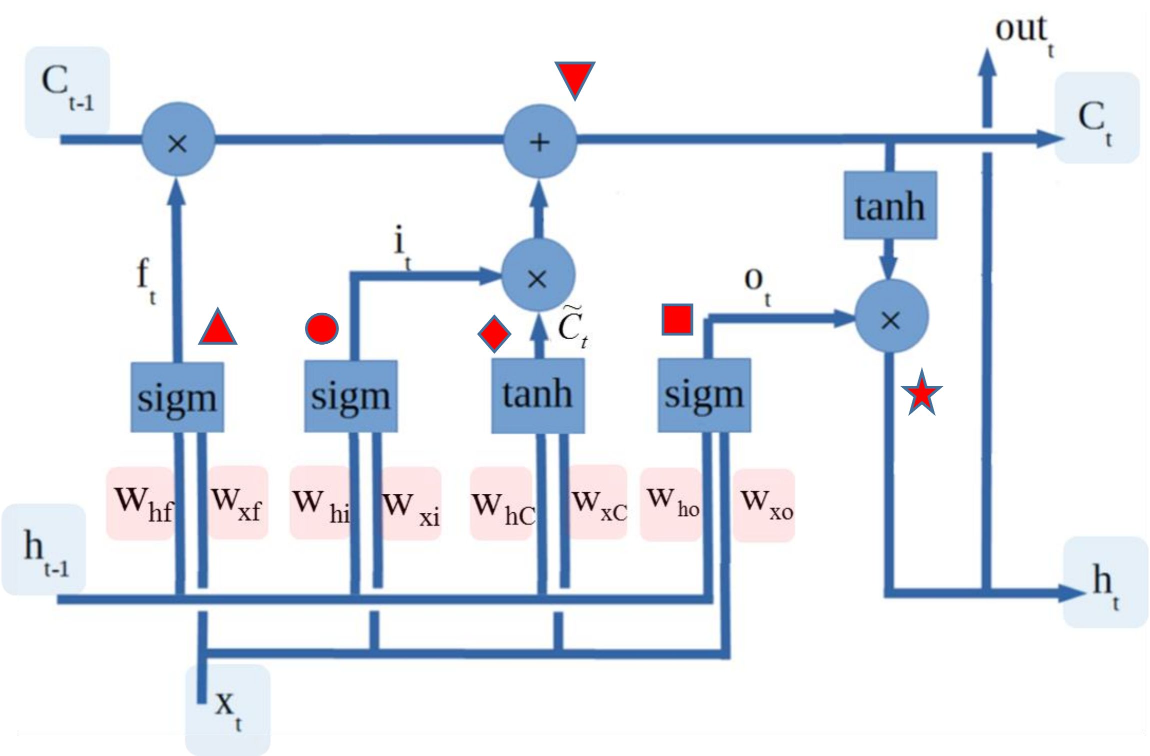
$$\tilde{C}_t = \tanh(\mathrm{w}_{\mathrm{xC}}\mathrm{x}_\mathrm{t} + \mathrm{w}_{\mathrm{hC}}\mathrm{h}_{\mathrm{t-1}} + b_C)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

$$h_t = o_t * \tanh(C_t)$$

LSTM Backpropagation

Long Short Term Memory (LSTM) network: Backpropagation



▲
$$a_f^{(t)} = w_{xf} x^{(t)} + w_{hf} h^{(t-1)}$$
$$b_f^{(t)} = sigm(a_f^{(t)})$$

●
$$a_i^{(t)} = w_{xi} x^{(t)} + w_{hi} h^{(t-1)}$$
$$b_i^{(t)} = sigm(a_i^{(t)})$$

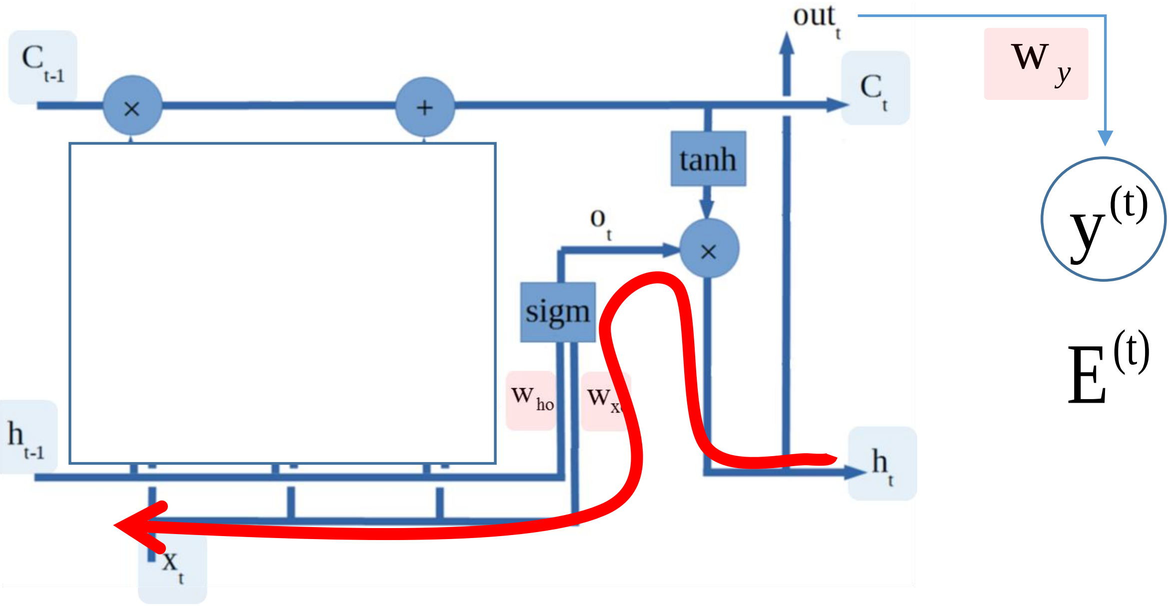
■
$$a_o^{(t)} = w_{xo} x^{(t)} + w_{ho} h^{(t-1)}$$
$$b_o^{(t)} = sigm(a_o^{(t)})$$

◆
$$a_c^{(t)} = w_{xc} x^{(t)} + w_{hc} h^{(t-1)}$$
$$b_c^{(t)} = \tanh(a_c^{(t)})$$

▼
$$C^{(t)} = b_f^{(t)} * C^{(t-1)} + b_i^{(t)} * b_c^{(t)}$$

★
$$h^{(t)} = b_o^{(t)} * \tanh(C^{(t)})$$

LSTM network: Backpropagation – output gate (t=1)



$$y^{(t)} = \sigma\left(w_y h^{(t)} + b_y\right)$$
$$s_y^{(t)} = w_y h^{(t)} + b_y$$
$$\sigma(\cdot): soft\max$$

$$a_f^{(t)} = w_{xf} x^{(t)} + w_{hf} h^{(t-1)}$$
$$b_f^{(t)} = sigm(a_f^{(t)})$$
$$a_i^{(t)} = w_{xi} x^{(t)} + w_{hi} h^{(t-1)}$$
$$b_i^{(t)} = sigm(a_i^{(t)})$$
$$a_o^{(t)} = w_{xo} x^{(t)} + w_{ho} h^{(t-1)}$$
$$b_o^{(t)} = sigm(a_o^{(t)})$$
$$a_c^{(t)} = w_{xc} x^{(t)} + w_{hc} h^{(t-1)}$$
$$b_c^{(t)} = \tanh(a_c^{(t)})$$
$$C^{(t)} = b_f^{(t)} * C^{(t-1)} + b_i^{(t)} * b_c^{(t)}$$
$$h^{(t)} = b_o^{(t)} * \tanh(C^{(t)})$$

$$\frac{\partial E^{(1)}}{\partial w_{ho}} = \boxed{\frac{\partial E^{(1)}}{\partial y^{(1)}}}\,\boxed{\frac{\partial y^{(1)}}{\partial s_y^{(1)}}}\,\boxed{\frac{\partial s_y^{(1)}}{\partial h^{(1)}}}\,\boxed{\frac{\partial h^{(1)}}{\partial b_o^{(1)}}}\,\boxed{\frac{\partial b_o^{(1)}}{\partial a_o^{(1)}}}\,\boxed{\frac{\partial a_o^{(1)}}{\partial w_{ho}}}$$

$\tanh(C^{(t)})$

Linear input function

Activation function (sigm)

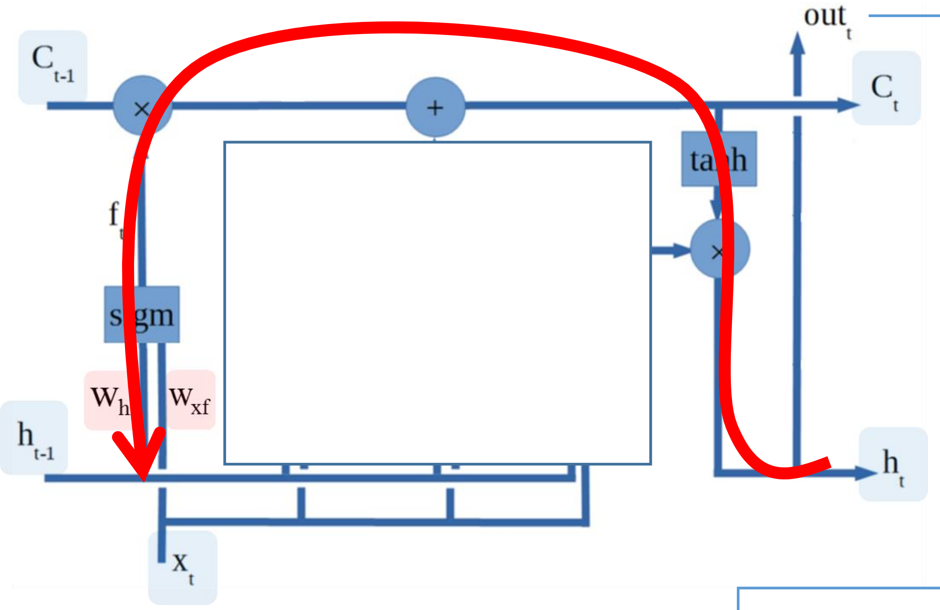
Linear input function

Activation function (softmax)

Error function

$$\frac{\partial E^{(1)}}{\partial w_{xo}} = ?$$

LSTM network: Backpropagation – forget gate (t=1)



$$y^{(t)}$$

$$E^{(t)}$$

$$y^{(t)} = \sigma\left(w_y h^{(t)} + b_y\right)$$
$$s_y^{(t)} = w_y h^{(t)} + b_y$$
$$\sigma(\cdot): soft\max$$

$$a_f^{(t)} = w_{xf} x^{(t)} + w_{hf} h^{(t-1)}$$
$$b_f^{(t)} = sigm(a_f^{(t)})$$
$$a_i^{(t)} = w_{xi} x^{(t)} + w_{hi} h^{(t-1)}$$
$$b_i^{(t)} = sigm(a_i^{(t)})$$
$$a_o^{(t)} = w_{xo} x^{(t)} + w_{ho} h^{(t-1)}$$
$$b_o^{(t)} = sigm(a_o^{(t)})$$
$$a_c^{(t)} = w_{xc} x^{(t)} + w_{hc} h^{(t-1)}$$
$$b_c^{(t)} = \tanh(a_c^{(t)})$$
$$C^{(t)} = b_f^{(t)} * C^{(t-1)} + b_i^{(t)} * b_c^{(t)}$$
$$h^{(t)} = b_o^{(t)} * \tanh(C^{(t)})$$

$$\frac{\partial E^{(1)}}{\partial w_{hf}} = \boxed{\frac{\partial E^{(1)}}{\partial y^{(1)}}}\boxed{\frac{\partial y^{(1)}}{\partial s_y^{(1)}}}\boxed{\frac{\partial s_y^{(1)}}{\partial h^{(1)}}}\boxed{\frac{\partial h^{(1)}}{\partial C^{(1)}}}\boxed{\frac{\partial C^{(1)}}{\partial b_f^{(1)}}}\boxed{\frac{\partial b_f^{(1)}}{\partial a_f^{(1)}}}\boxed{\frac{\partial a_f^{(1)}}{\partial w_{hf}}}$$

$C^{(0)}$

Activation function (sigm)

Linear input function

Activation function (tanh)

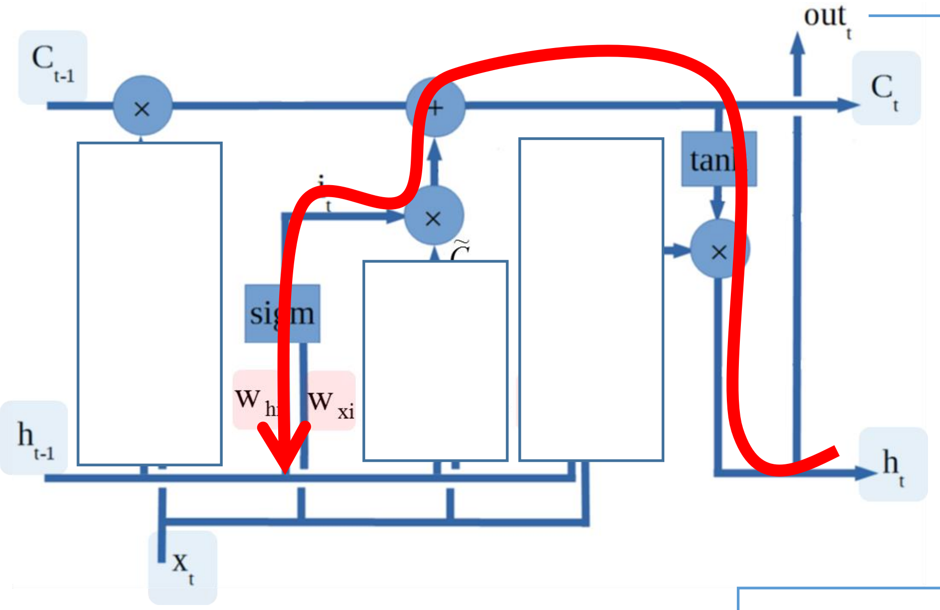
Linear input function

Activation function (softmax)

Error function

$$\frac{\partial E^{(1)}}{\partial w_{xf}} = ?$$

LSTM network: Backpropagation – input gate (t=1)



w_y

$y^{(t)}$

$E^{(t)}$

$$y^{(t)} = \sigma\left(w_y h^{(t)} + b_y\right)$$
$$s_y^{(t)} = w_y h^{(t)} + b_y$$

$$a_f^{(t)} = w_{xf} x^{(t)} + w_{hf} h^{(t-1)}$$
$$b_f^{(t)} = sigm(a_f^{(t)})$$
$$a_i^{(t)} = w_{xi} x^{(t)} + w_{hi} h^{(t-1)}$$
$$b_i^{(t)} = sigm(a_i^{(t)})$$
$$a_o^{(t)} = w_{xo} x^{(t)} + w_{ho} h^{(t-1)}$$
$$b_o^{(t)} = sigm(a_o^{(t)})$$
$$a_c^{(t)} = w_{xc} x^{(t)} + w_{hc} h^{(t-1)}$$
$$b_c^{(t)} = \tanh(a_c^{(t)})$$
$$C^{(t)} = b_f^{(t)} * C^{(t-1)} + b_i^{(t)} * b_c^{(t)}$$
$$h^{(t)} = b_o^{(t)} * \tanh(C^{(t)})$$

$b_c^{(t)}$

Activation function (sigm)

$$\frac{\partial E^{(1)}}{\partial w_{hi}} = \boxed{\frac{\partial E^{(1)}}{\partial y^{(1)}}}\,\boxed{\frac{\partial y^{(1)}}{\partial s_y^{(1)}}}\,\boxed{\frac{\partial s_y^{(1)}}{\partial h^{(1)}}}\,\boxed{\frac{\partial h^{(1)}}{\partial C^{(1)}}}\,\boxed{\frac{\partial C^{(1)}}{\partial b_i^{(1)}}}\,\boxed{\frac{\partial b_i^{(1)}}{\partial a_i^{(1)}}}\,\boxed{\frac{\partial a_i^{(1)}}{\partial w_{hi}}}$$

Linear input function

Activation function (tanh)

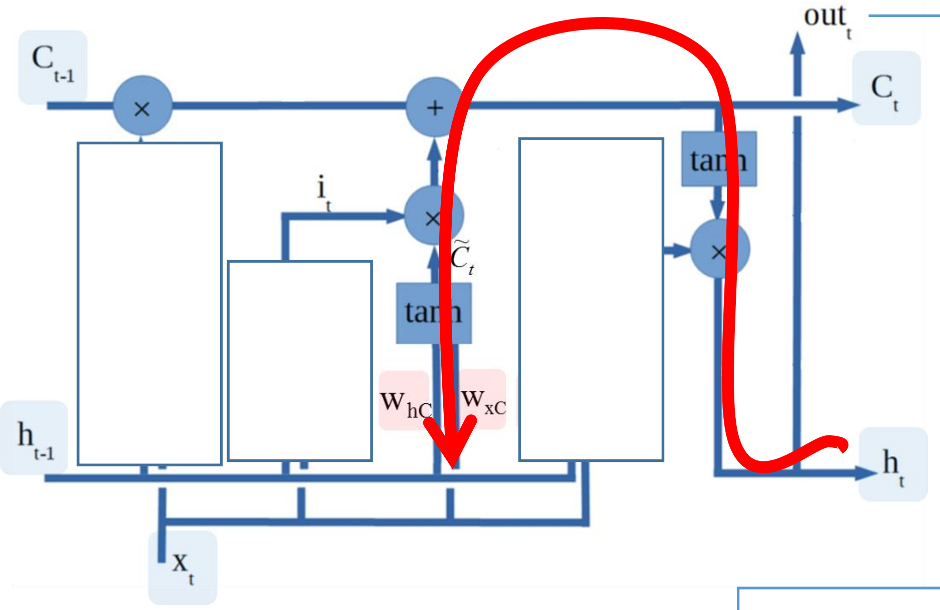
Linear input function

Activation function (softmax)

Error function

$$\frac{\partial E^{(1)}}{\partial w_{xi}} = ?$$

LSTM network: Backpropagation – block input (t=1)



$y^{(t)}$

$E^{(t)}$

$$y^{(t)} = \sigma\left(w_y h^{(t)} + b_y\right)$$
$$s_y^{(t)} = w_y h^{(t)} + b_y$$
$$\sigma(\cdot): soft\max$$

$$a_f^{(t)} = w_{xf} x^{(t)} + w_{hf} h^{(t-1)}$$
$$b_f^{(t)} = sigm(a_f^{(t)})$$
$$a_i^{(t)} = w_{xi} x^{(t)} + w_{hi} h^{(t-1)}$$
$$b_i^{(t)} = sigm(a_i^{(t)})$$
$$a_o^{(t)} = w_{xo} x^{(t)} + w_{ho} h^{(t-1)}$$
$$b_o^{(t)} = sigm(a_o^{(t)})$$
$$a_c^{(t)} = w_{xc} x^{(t)} + w_{hc} h^{(t-1)}$$
$$b_c^{(t)} = \tanh(a_c^{(t)})$$
$$C^{(t)} = b_f^{(t)} * C^{(t-1)} + b_i^{(t)} * b_c^{(t)}$$
$$h^{(t)} = b_o^{(t)} * \tanh(C^{(t)})$$

$b_i^{(t)}$

Activation function (tanh)

$$\frac{\partial E^{(1)}}{\partial w_{hc}} = \boxed{\frac{\partial E^{(1)}}{\partial y^{(1)}}}\boxed{\frac{\partial y^{(1)}}{\partial s_y^{(1)}}}\boxed{\frac{\partial s_y^{(1)}}{\partial h^{(1)}}}\boxed{\frac{\partial h^{(1)}}{\partial C^{(1)}}}\boxed{\frac{\partial C^{(1)}}{\partial b_c^{(1)}}}\boxed{\frac{\partial b_c^{(1)}}{\partial a_c^{(1)}}}\boxed{\frac{\partial a_c^{(1)}}{\partial w_{hc}}}$$

Linear input function

Activation function (tanh)

Linear input function

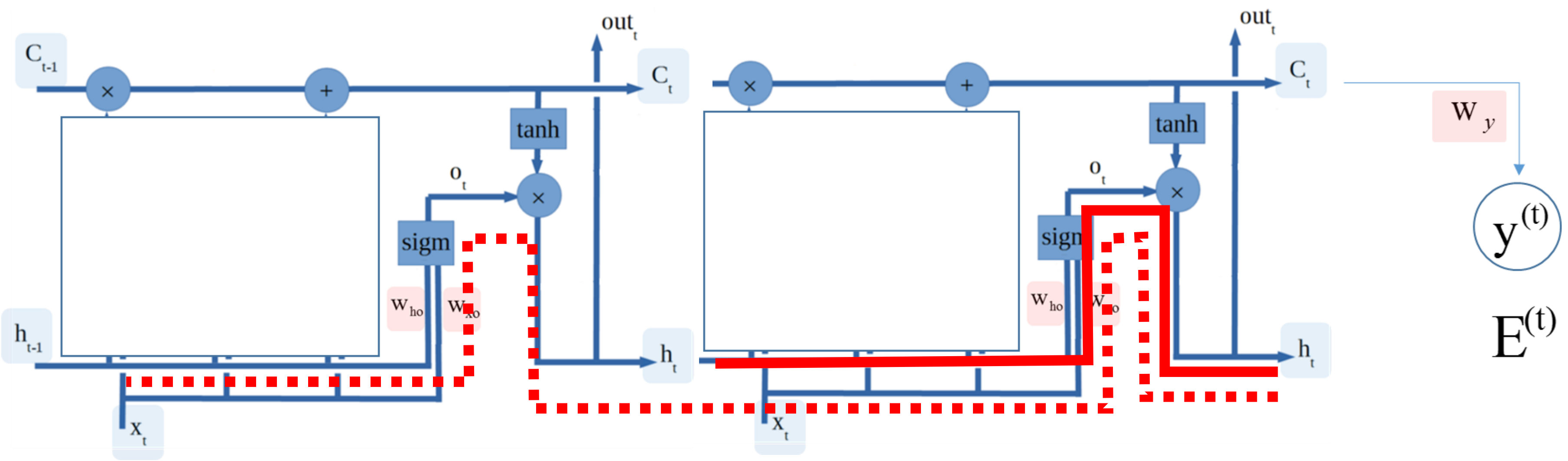
Activation function (softmax)

Error function

$$\frac{\partial E^{(1)}}{\partial w_{xc}} = ?$$

LSTM Backpropagation (t=2)

LSTM network: Backpropagation – output gate (t=2)



$$y^{(t)} = \sigma\left(w_y h^{(t)} + b_y\right)$$
$$s_y^{(t)} = w_y h^{(t)} + b_y$$
$$\sigma(\cdot): soft\max$$

$$\sum_{t=1}^{2}\frac{\partial E^{(t)}}{\partial w_{ho}} = \frac{\partial E^{(2)}}{\partial y^{(2)}}\frac{\partial y^{(2)}}{\partial s_y^{(2)}}\frac{\partial s_y^{(2)}}{\partial h^{(2)}}\frac{\partial h^{(2)}}{\partial b_o^{(2)}}\frac{\partial b_o^{(2)}}{\partial a_o^{(2)}}\frac{\partial a_o^{(2)}}{\partial w_{ho}}$$

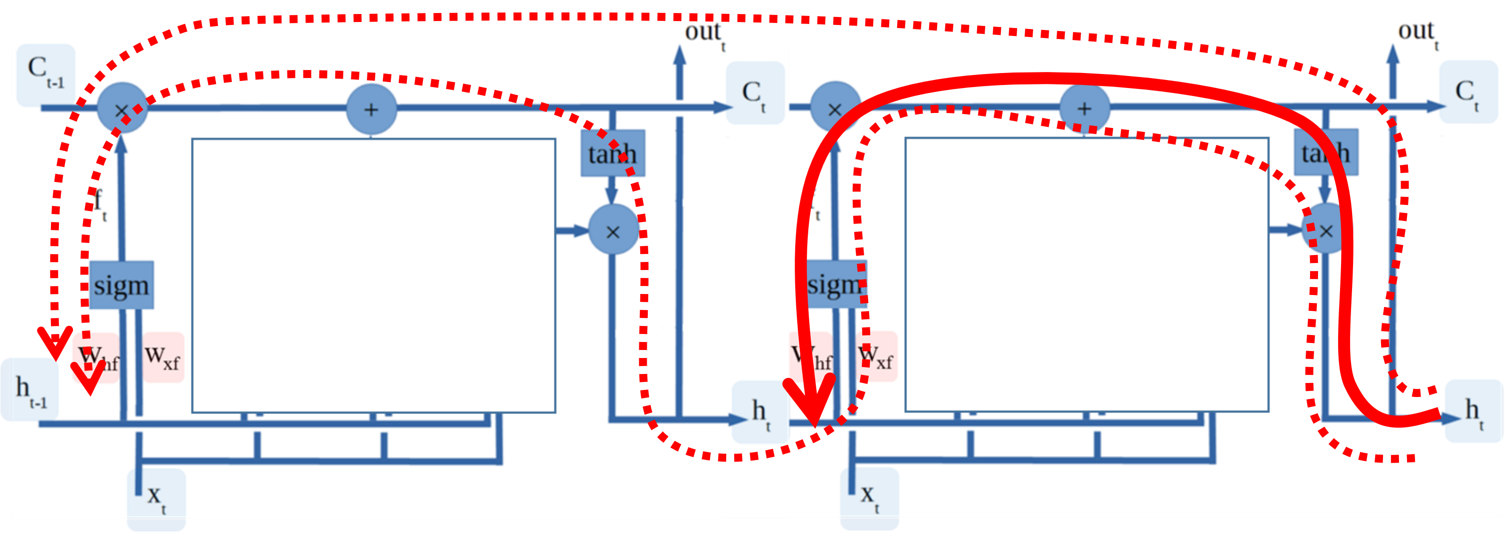
$$+\frac{\partial E^{(2)}}{\partial y^{(2)}}\frac{\partial y^{(2)}}{\partial s_y^{(2)}}\frac{\partial s_y^{(2)}}{\partial h^{(2)}}\underbrace{\boxed{\frac{\partial h^{(2)}}{\partial b_o^{(2)}}}}_{\tanh(C^{(t)})}\boxed{\frac{\partial b_o^{(2)}}{\partial a_o^{(2)}}}\frac{\partial a_o^{(2)}}{\partial h^{(1)}}\underbrace{\boxed{\frac{\partial h^{(1)}}{\partial b_o^{(1)}}}}_{\tanh(C^{(t)})}\boxed{\frac{\partial b_o^{(1)}}{\partial a_o^{(1)}}}\frac{\partial a_o^{(1)}}{\partial w_{ho}}$$

Activation function (sigm)

Activation function (sigm)

$$a_f^{(t)} = w_{xf}x^{(t)} + w_{hf}h^{(t-1)}$$
$$b_f^{(t)} = sigm(a_f^{(t)})$$
$$a_i^{(t)} = w_{xi}x^{(t)} + w_{hi}h^{(t-1)}$$
$$b_i^{(t)} = sigm(a_i^{(t)})$$
$$a_o^{(t)} = w_{xo}x^{(t)} + w_{ho}h^{(t-1)}$$
$$b_o^{(t)} = sigm(a_o^{(t)})$$
$$a_c^{(t)} = w_{xc}x^{(t)} + w_{hc}h^{(t-1)}$$
$$b_c^{(t)} = \tanh(a_c^{(t)})$$
$$C^{(t)} = b_f^{(t)} * C^{(t-1)} + b_i^{(t)} * b_c^{(t)}$$
$$h^{(t)} = b_o^{(t)} * \tanh(C^{(t)})$$

LSTM network: Backpropagation – forget gate (t=2)



$$a_f^{(t)} = w_{xf} x^{(t)} + w_{hf} h^{(t-1)}$$
$$b_f^{(t)} = sigm(a_f^{(t)})$$
$$a_i^{(t)} = w_{xi} x^{(t)} + w_{hi} h^{(t-1)}$$
$$b_i^{(t)} = sigm(a_i^{(t)})$$
$$a_o^{(t)} = w_{xo} x^{(t)} + w_{ho} h^{(t-1)}$$
$$b_o^{(t)} = sigm(a_o^{(t)})$$
$$a_c^{(t)} = w_{xc} x^{(t)} + w_{hc} h^{(t-1)}$$
$$b_c^{(t)} = \tanh(a_c^{(t)})$$
$$C^{(t)} = b_f^{(t)} * C^{(t-1)} + b_i^{(t)} * b_c^{(t)}$$
$$h^{(t)} = b_o^{(t)} * \tanh(C^{(t)})$$

$$\sum_{t=1}^{2} \frac{\partial E^{(t)}}{\partial w_{hf}} = \frac{\partial E^{(2)}}{\partial y^{(2)}} \frac{\partial y^{(2)}}{\partial s_y^{(2)}} \frac{\partial s_y^{(2)}}{\partial h^{(2)}} \frac{\partial h^{(2)}}{\partial C^{(2)}} \frac{\partial C^{(2)}}{\partial b_f^{(2)}} \frac{\partial b_f^{(2)}}{\partial a_f^{(2)}} \frac{\partial a_f^{(2)}}{\partial w_{hf}}$$

tanh’ C(t-1) sigm’ h(t-1)

$$+ \frac{\partial E^{(2)}}{\partial y^{(2)}} \frac{\partial y^{(2)}}{\partial s_y^{(2)}} \frac{\partial s_y^{(2)}}{\partial h^{(2)}} \frac{\partial h^{(2)}}{\partial C^{(2)}} \frac{\partial C^{(2)}}{\partial b_f^{(2)}} \frac{\partial b_f^{(2)}}{\partial a_f^{(2)}} \frac{\partial a_f^{(2)}}{\partial h^{(1)}} \boxed{\frac{\partial h^{(1)}}{\partial C^{(1)}}} \boxed{\frac{\partial C^{(1)}}{\partial b_f^{(1)}}} \boxed{\frac{\partial b_f^{(1)}}{\partial a_f^{(1)}}} \boxed{\frac{\partial a_f^{(1)}}{\partial w_{hf}}}$$

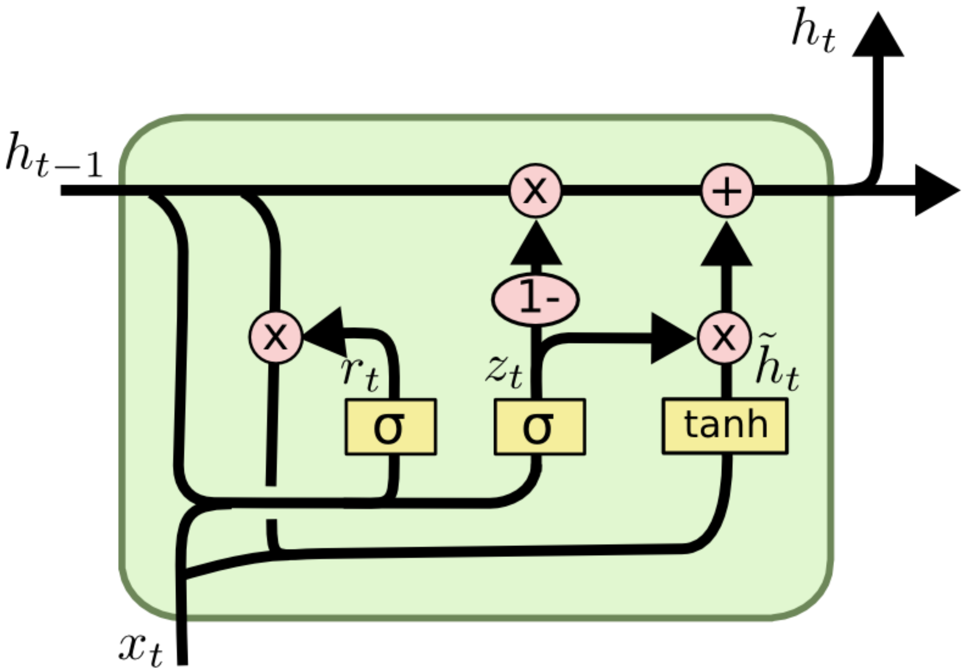
$$+ \frac{\partial E^{(2)}}{\partial y^{(2)}} \frac{\partial y^{(2)}}{\partial s_y^{(2)}} \frac{\partial s_y^{(2)}}{\partial h^{(2)}} \frac{\partial h^{(2)}}{\partial C^{(2)}} \boxed{\frac{\partial C^{(2)}}{\partial C^{(1)}}} \boxed{\frac{\partial C^{(1)}}{\partial b_f^{(1)}}} \boxed{\frac{\partial b_f^{(1)}}{\partial a_f^{(1)}}} \boxed{\frac{\partial a_f^{(1)}}{\partial w_{hf}}}$$

LSTM network: Backpropagation – input gate and block gate (t=2)

- Backpropagations of input gate and block gate are similar to that of forget gate in the previous slide.
- Please, refer to the backup slides.

Gated Recurrent Unit (GRU)

- Simpler than LSTM and so training is faster,
- Cell state (C) is replaced by hidden state (h),
- GRU has two gates: update gate (z), reset gate (r).



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

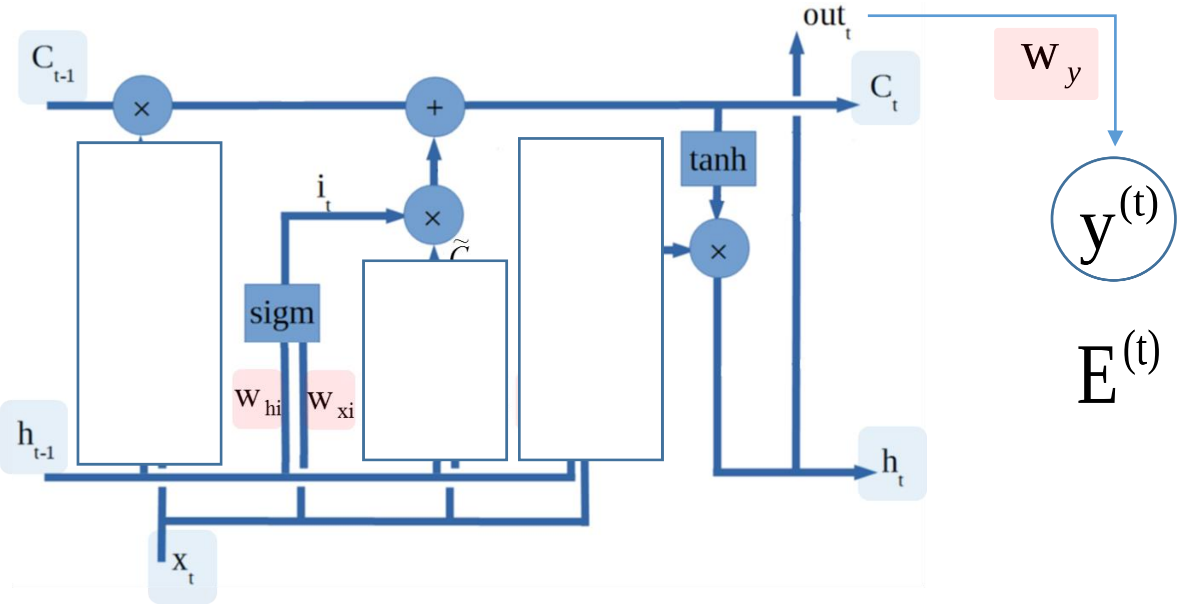
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

Backup Slides

LSTM network: Backpropagation – input gate (t=2)



$$y^{(t)} = \sigma\left(w_y h^{(t)} + b_y\right)$$
$$s_y^{(t)} = w_y h^{(t)} + b_y$$
$$\sigma(\cdot): soft\max$$

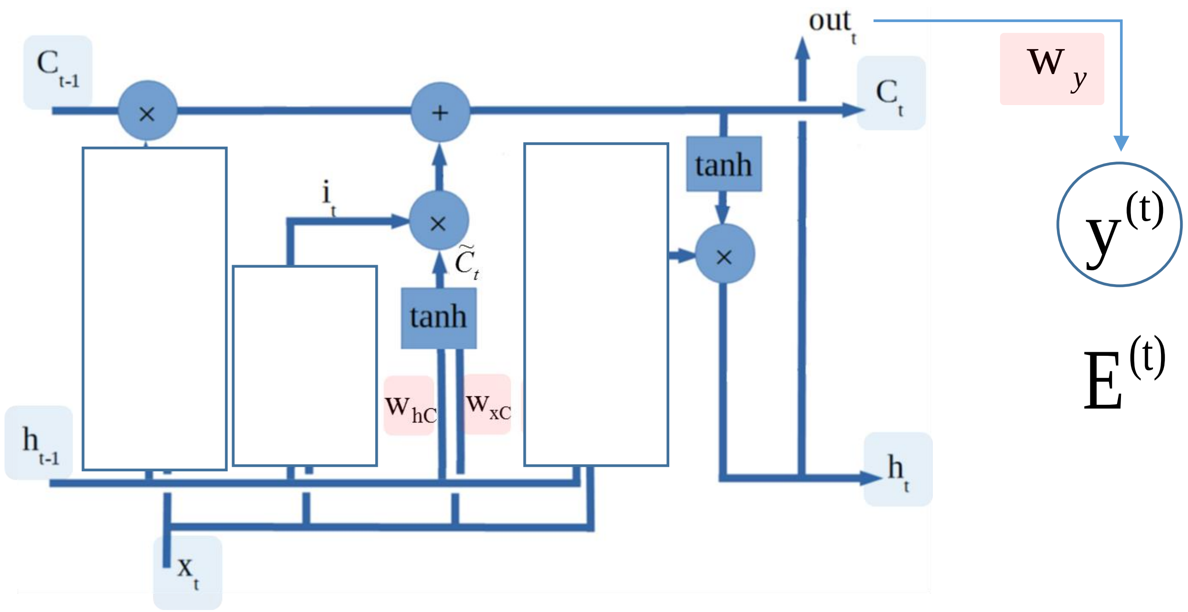
$$a_f^{(t)} = w_{xf} x^{(t)} + w_{hf} h^{(t-1)}$$
$$b_f^{(t)} = sigm(a_f^{(t)})$$
$$a_i^{(t)} = w_{xi} x^{(t)} + w_{hi} h^{(t-1)}$$
$$b_i^{(t)} = sigm(a_i^{(t)})$$
$$a_o^{(t)} = w_{xo} x^{(t)} + w_{ho} h^{(t-1)}$$
$$b_o^{(t)} = sigm(a_o^{(t)})$$
$$a_c^{(t)} = w_{xc} x^{(t)} + w_{hc} h^{(t-1)}$$
$$b_c^{(t)} = \tanh(a_c^{(t)})$$
$$C^{(t)} = b_f^{(t)} * C^{(t-1)} + b_i^{(t)} * b_c^{(t)}$$
$$h^{(t)} = b_o^{(t)} * \tanh(C^{(t)})$$

$$\frac{\partial E^{(2)}}{\partial w_{hi}} = \frac{\partial E^{(2)}}{\partial y^{(2)}} \frac{\partial y^{(2)}}{\partial s_y^{(2)}} \frac{\partial s_y^{(2)}}{\partial h^{(2)}} \frac{\partial h^{(2)}}{\partial C^{(2)}} \frac{\partial C^{(2)}}{\partial b_i^{(2)}} \frac{\partial b_i^{(2)}}{\partial a_i^{(2)}} \frac{\partial a_i^{(2)}}{\partial w_{hi}}$$

$$+ \frac{\partial E^{(2)}}{\partial y^{(2)}} \frac{\partial y^{(2)}}{\partial s_y^{(2)}} \frac{\partial s_y^{(2)}}{\partial h^{(2)}} \frac{\partial h^{(2)}}{\partial C^{(2)}} \frac{\partial C^{(2)}}{\partial b_i^{(2)}} \frac{\partial b_i^{(2)}}{\partial a_i^{(2)}} \frac{\partial a_i^{(2)}}{\partial h^{(1)}} \boxed{\frac{\partial h^{(1)}}{\partial C^{(1)}}} \boxed{\frac{\partial C^{(1)}}{\partial b_i^{(1)}}} \boxed{\frac{\partial b_i^{(1)}}{\partial a_i^{(1)}}} \boxed{\frac{\partial a_i^{(1)}}{\partial w_{hi}}}$$

$$+ \frac{\partial E^{(2)}}{\partial y^{(2)}} \frac{\partial y^{(2)}}{\partial s_y^{(2)}} \frac{\partial s_y^{(2)}}{\partial h^{(2)}} \frac{\partial h^{(2)}}{\partial C^{(2)}} \boxed{\frac{\partial C^{(2)}}{\partial C^{(1)}}} \boxed{\frac{\partial C^{(1)}}{\partial b_i^{(1)}}} \boxed{\frac{\partial b_i^{(1)}}{\partial a_i^{(1)}}} \boxed{\frac{\partial a_i^{(1)}}{\partial w_{hi}}}$$

LSTM network: Backpropagation – block input (t=2)



$$y^{(t)} = \sigma(w_y h^{(t)} + b_y)$$
$$s_y^{(t)} = w_y h^{(t)} + b_y$$
$$\sigma(\cdot): soft\max$$

$$a_f^{(t)} = w_{xf} x^{(t)} + w_{hf} h^{(t-1)}$$
$$b_f^{(t)} = sigm(a_f^{(t)})$$
$$a_i^{(t)} = w_{xi} x^{(t)} + w_{hi} h^{(t-1)}$$
$$b_i^{(t)} = sigm(a_i^{(t)})$$
$$a_o^{(t)} = w_{xo} x^{(t)} + w_{ho} h^{(t-1)}$$
$$b_o^{(t)} = sigm(a_o^{(t)})$$
$$a_c^{(t)} = w_{xc} x^{(t)} + w_{hc} h^{(t-1)}$$
$$b_c^{(t)} = \tanh(a_c^{(t)})$$
$$C^{(t)} = b_f^{(t)} * C^{(t-1)} + b_i^{(t)} * b_c^{(t)}$$
$$h^{(t)} = b_o^{(t)} * \tanh(C^{(t)})$$

$$\frac{\partial E^{(2)}}{\partial w_{hc}} = \frac{\partial E^{(2)}}{\partial y^{(2)}} \frac{\partial y^{(2)}}{\partial s_y^{(2)}} \frac{\partial s_y^{(2)}}{\partial h^{(2)}} \frac{\partial h^{(2)}}{\partial C^{(2)}} \frac{\partial C^{(2)}}{\partial b_c^{(2)}} \frac{\partial b_c^{(2)}}{\partial a_c^{(2)}} \frac{\partial a_c^{(2)}}{\partial w_{hc}}$$

$$+ \frac{\partial E^{(2)}}{\partial y^{(2)}} \frac{\partial y^{(2)}}{\partial s_y^{(2)}} \frac{\partial s_y^{(2)}}{\partial h^{(2)}} \frac{\partial h^{(2)}}{\partial C^{(2)}} \frac{\partial C^{(2)}}{\partial b_c^{(2)}} \frac{\partial b_c^{(2)}}{\partial a_c^{(2)}} \frac{\partial a_c^{(2)}}{\partial h^{(1)}} \boxed{\frac{\partial h^{(1)}}{\partial C^{(1)}}} \boxed{\frac{\partial C^{(1)}}{\partial b_c^{(1)}}} \boxed{\frac{\partial b_c^{(1)}}{\partial a_c^{(1)}}} \boxed{\frac{\partial a_c^{(1)}}{\partial w_{hc}}}$$

$$+ \frac{\partial E^{(2)}}{\partial y^{(2)}} \frac{\partial y^{(2)}}{\partial s_y^{(2)}} \frac{\partial s_y^{(2)}}{\partial h^{(2)}} \frac{\partial h^{(2)}}{\partial C^{(2)}} \boxed{\frac{\partial C^{(2)}}{\partial C^{(1)}}} \boxed{\frac{\partial C^{(1)}}{\partial b_c^{(1)}}} \boxed{\frac{\partial b_c^{(1)}}{\partial a_c^{(1)}}} \boxed{\frac{\partial a_c^{(1)}}{\partial w_{hc}}}$$

Operation of LSTM: weight update

$$\begin{bmatrix} \frac{\partial E}{\partial w_{xo}} \\ \frac{\partial E}{\partial w_{xf}} \\ \frac{\partial E}{\partial w_{xi}} \\ \frac{\partial E}{\partial w_{xc}} \end{bmatrix} = \begin{bmatrix} \frac{\partial E^{(1)}}{\partial w_{xo}} + \frac{\partial E^{(2)}}{\partial w_{xo}} + \cdots + \frac{\partial E^{(t)}}{\partial w_{xo}} \\ \frac{\partial E^{(1)}}{\partial w_{xf}} + \frac{\partial E^{(2)}}{\partial w_{xf}} + \cdots + \frac{\partial E^{(t)}}{\partial w_{xf}} \\ \frac{\partial E^{(1)}}{\partial w_{xi}} + \frac{\partial E^{(2)}}{\partial w_{xi}} + \cdots + \frac{\partial E^{(t)}}{\partial w_{xi}} \\ \frac{\partial E^{(1)}}{\partial w_{xc}} + \frac{\partial E^{(2)}}{\partial w_{xc}} + \cdots + \frac{\partial E^{(t)}}{\partial w_{xc}} \end{bmatrix}$$

$$\begin{bmatrix} \frac{\partial E}{\partial w_{ho}} \\ \frac{\partial E}{\partial w_{hf}} \\ \frac{\partial E}{\partial w_{hi}} \\ \frac{\partial E}{\partial w_{hc}} \end{bmatrix} = \begin{bmatrix} \frac{\partial E^{(1)}}{\partial w_{ho}} + \frac{\partial E^{(2)}}{\partial w_{ho}} + \cdots + \frac{\partial E^{(t)}}{\partial w_{ho}} \\ \frac{\partial E^{(1)}}{\partial w_{hf}} + \frac{\partial E^{(2)}}{\partial w_{hf}} + \cdots + \frac{\partial E^{(t)}}{\partial w_{hf}} \\ \frac{\partial E^{(1)}}{\partial w_{hi}} + \frac{\partial E^{(2)}}{\partial w_{hi}} + \cdots + \frac{\partial E^{(t)}}{\partial w_{hi}} \\ \frac{\partial E^{(1)}}{\partial w_{hc}} + \frac{\partial E^{(2)}}{\partial w_{hc}} + \cdots + \frac{\partial E^{(t)}}{\partial w_{hc}} \end{bmatrix}$$

where:

$$h^{(t-1)}, h^{(t)} \in R^{n} \qquad w_{xo} \in R^{n\times m} \qquad w_{ho} \in R^{n\times n}$$
$$C^{(t)} \in R^{n} \qquad w_{xf} \in R^{n\times m} \qquad w_{hf} \in R^{n\times n}$$
$$x^{(t)} \in R^{m} \qquad w_{xi} \in R^{n\times m} \qquad w_{hi} \in R^{n\times n}$$
$$w_{xc} \in R^{n\times m} \qquad w_{hc} \in R^{n\times n}$$

$$\begin{bmatrix} w_{xo}^{new} \\ w_{xf}^{new} \\ w_{xi}^{new} \\ w_{xc}^{new} \end{bmatrix} = \begin{bmatrix} w_{xo}^{old} - \lambda \frac{\partial E}{\partial w_{xo}} \\ w_{xf}^{new} - \lambda \frac{\partial E}{\partial w_{xf}} \\ w_{xi}^{new} - \lambda \frac{\partial E}{\partial w_{xi}} \\ w_{xc}^{new} - \lambda \frac{\partial E}{\partial w_{xc}} \end{bmatrix} \qquad \begin{bmatrix} w_{ho}^{new} \\ w_{hf}^{new} \\ w_{hi}^{new} \\ w_{hc}^{new} \end{bmatrix} = \begin{bmatrix} w_{ho}^{old} - \lambda \frac{\partial E}{\partial w_{ho}} \\ w_{hf}^{new} - \lambda \frac{\partial E}{\partial w_{hf}} \\ w_{hi}^{new} - \lambda \frac{\partial E}{\partial w_{hi}} \\ w_{hc}^{new} - \lambda \frac{\partial E}{\partial w_{hc}} \end{bmatrix}$$

λ: Learning rate