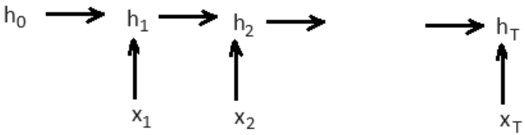
Lecture 28

Agenda

Recurrent Neural Networks (RNN)

- Simple RNN
- Vanishing Gradients Problem
- LSTM

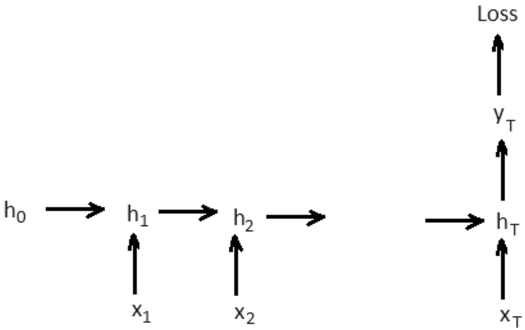
RNN



$$h_t = g(W_x x_t + b_x + W_h h_{t-1} + b_h)$$

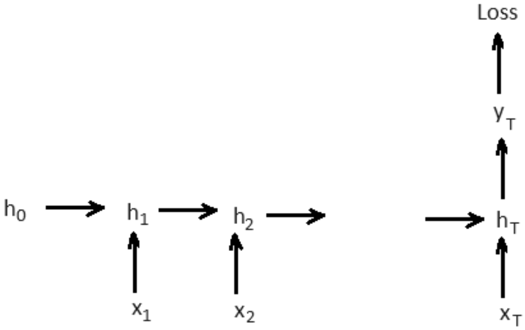
$$h_t = g(W_x x_t + W_h h_{t-1} + b_h)$$

RNN



$$h_t = g(W_x x_t + W_h h_{t-1} + b_h)$$

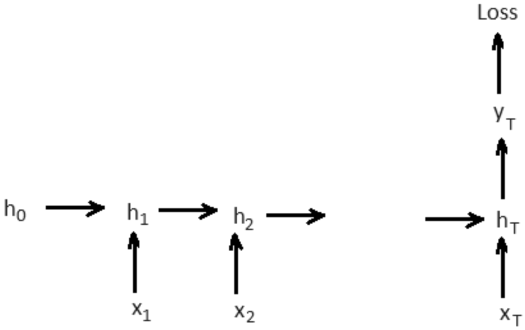
RNN



$$h_t = g(W_x x_t + W_h h_{t-1} + b_h)$$

$$h_t = g\left(W\begin{pmatrix} x_t \\ h_{t-1} \end{pmatrix} + b_h\right)$$

Vanishing Gradients Problem

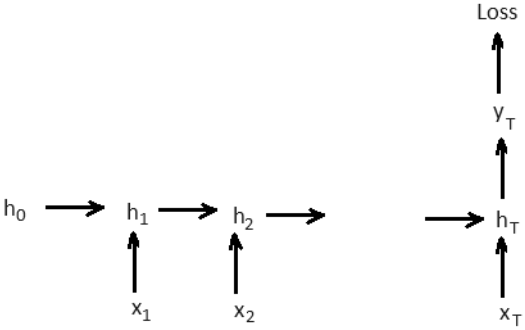


$$h_t = g(W_x x_t + W_h h_{t-1} + b_h)$$

$$\frac{\partial L}{\partial W_x} = \sum_{t=1}^{T} \frac{\partial L}{\partial h_t} \frac{\partial h_t}{\partial W_x}$$

$$\frac{\partial L}{\partial h_{T-1}} = \frac{\partial L}{\partial h_T} \frac{\partial h_T}{\partial h_{T-1}} = \frac{\partial L}{\partial y_T} \frac{\partial y_T}{\partial h_T} \frac{\partial h_T}{\partial h_{T-1}} \frac{\partial h_T}{\partial h_{T-1}}$$

Vanishing Gradients Problem



$$h_t = g(W_x x_t + W_h h_{t-1} + b_h)$$

$$\frac{\partial L}{\partial W_x} = \sum_{t=1}^{T} \frac{\partial L}{\partial y_T} \frac{\partial y_T}{\partial h_T} \frac{\partial h_t}{\partial W_x} \prod_{k=t}^{T-1} \frac{\partial h_{k+1}}{\partial h_k}$$

$$\left\| \frac{\partial h_{k+1}}{\partial h_k} \right\|_2 < (>) 1 \quad \text{vanishing (exploding) gradient}$$

Ways to fix vanishing gradients problem

- Residual connections:

$$y = x + f(x) \quad \Rightarrow \quad \frac{\partial y}{\partial x} = I + \frac{\partial f}{\partial x}.$$

- Batch normalization
- Xavier initialization
- Orthogonal initialization

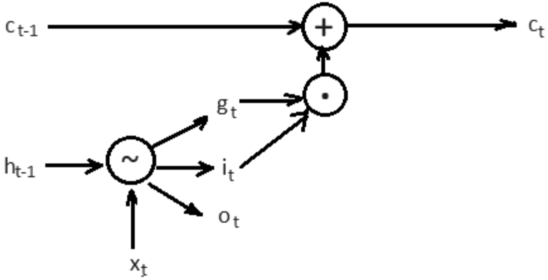
$$\frac{\partial h_{k+1}}{\partial h_k} = g' W_h \quad \Rightarrow \quad \left\| \frac{\partial h_{k+1}}{\partial h_k} \right\|_2 = \|g'\|_2 \text{ if } W_h W_h^T = I$$

- Orthogonal regularization

$$\ldots + \lambda \| W_h^T W_h - I \|_F^2$$

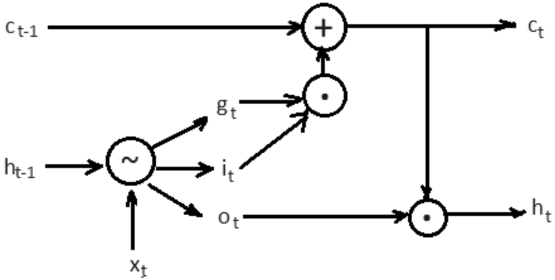
- Orthogonal optimization, e.g., Riemannian optimization

Long Short-Term Memory (LSTM)



$$g_t = g(W_x x_t + W_h h_{t-1} + b_h)$$
$$i_t = \sigma(W_x^i x_t + W_h^i h_{t-1} + b_h^i)$$
$$c_t = c_{t-1} + i_t g_t$$

Long Short-Term Memory (LSTM)

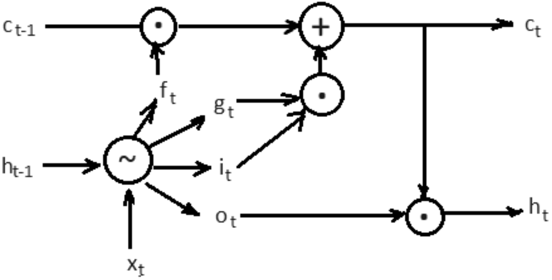


$$g_t = g(W_x x_t + W_h h_{t-1} + b_h)$$
$$i_t = \sigma(W_x^i x_t + W_h^i h_{t-1} + b_h^i)$$
$$c_t = c_{t-1} + i_t g_t$$
$$o_t = \sigma(W_x^o x_t + W_h^o h_{t-1} + b_h^o)$$
$$h_t = o_t g_t(c_t)$$

R · I · T

Computer Science @ RIT

Long Short-Term Memory (LSTM)



$$g_t = g(W_x x_t + W_h h_{t-1} + b_h)$$
$$i_t = \sigma(W_x^i x_t + W_h^i h_{t-1} + b_h^i)$$
$$f_t = \sigma(W_x^f x_t + W_h^f h_{t-1} + b_h^f)$$
$$c_t = f_t c_{t-1} + i_t g_t$$
$$o_t = \sigma(W_x^o x_t + W_h^o h_{t-1} + b_h^o)$$
$$h_t = o_t g_t(c_t)$$