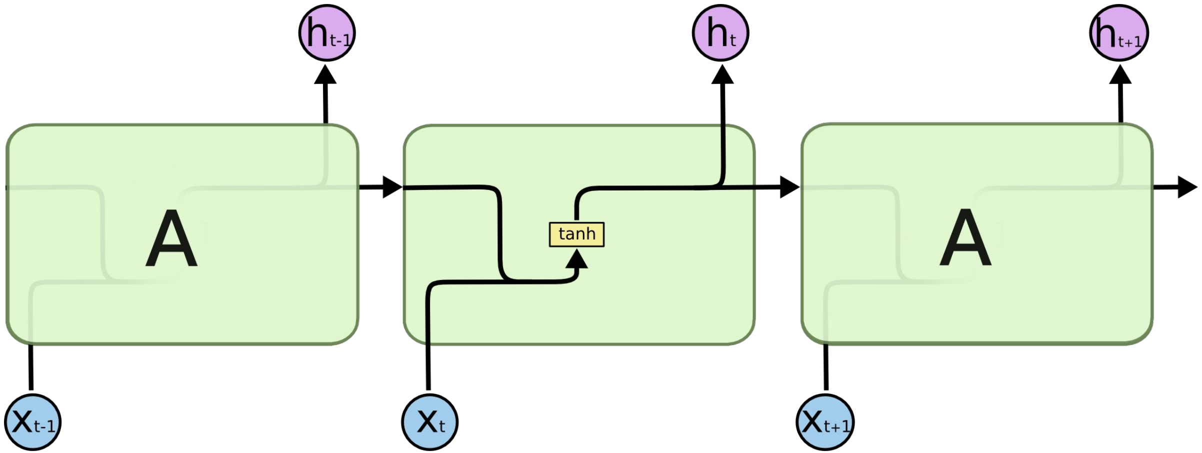
Long Short-Term Memory

Kellen Donahue
Anthony Mayer

Recurrent Neural Networks

- A strategy to deal with sequential data.
- Notoriously difficult to train until LSTM networks were invented in 1997.
- Vanilla RNN (Elman network) fails due to the vanishing or exploding gradient.
- LSTM was designed to solve this problem.



Vanilla RNN Equations

- Input: xt
- Output: yt,ht
- Activation: tanh

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

$$y_t = \zeta_y(W_y * h_t \ + \ b_y)$$

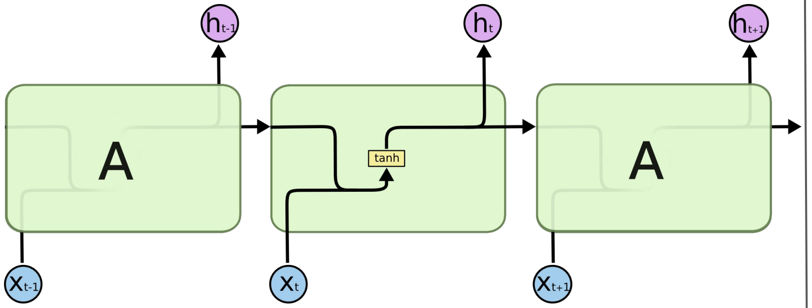
$$E_t = L(\hat{y}_t, y_t)$$

Vanilla RNN Equations

$$h_t = \zeta_h(W_x * x_t \; + \; W_h * h_{t-1} \; + \; b_h)$$

$$y_t = \zeta_y(W_y * h_t \; + \; b_y)$$

$$E_t = L(\hat{y}_t, y_t)$$

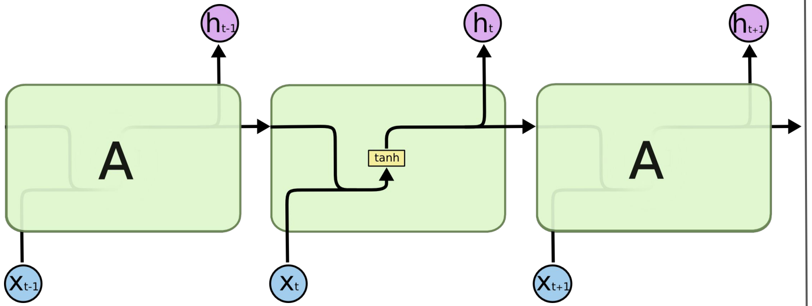


Vanilla RNN Equations

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

$$y_t = \zeta_y(W_y * h_t \ + \ b_y)$$

$$E_t = L(\hat{y}_t, y_t)$$



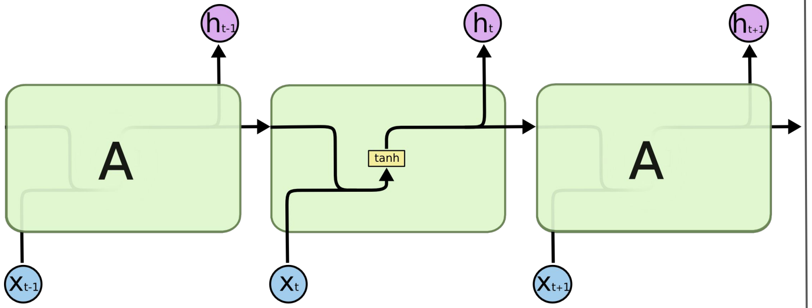
$$\frac{\partial E_t}{\partial W_h} = \frac{\partial E_t}{\partial \hat{y_t}} \frac{\partial \hat{y_t}}{\partial h_t} \frac{\partial h_t}{\partial h_{t-k}} \frac{\partial h_{t-k}}{\partial W_h}$$

Vanilla RNN Equations

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

$$y_t = \zeta_y(W_y * h_t \ + \ b_y)$$

$$E_t = L(\hat{y_t}, y_t)$$



$$\frac{\partial E_t}{\partial W_h} = \frac{\partial E_t}{\partial \hat{y_t}} \frac{\partial \hat{y_t}}{\partial h_t} \frac{\partial h_t}{\partial h_{t-k}} \frac{\partial h_{t-k}}{\partial W_h}$$

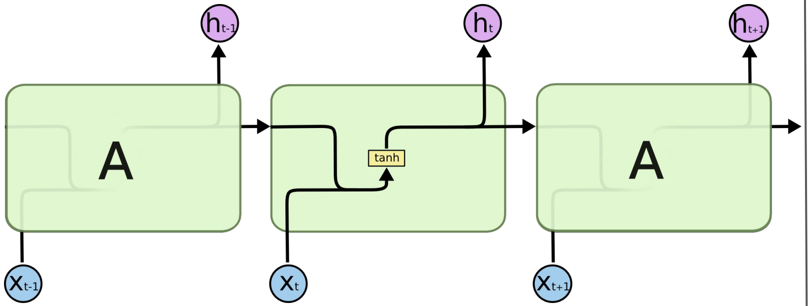
$$\frac{\partial E_t}{\partial W_h} = \sum_{k=0}^{t-1} \frac{\partial E_t}{\partial \hat{y_t}} \frac{\partial \hat{y_t}}{\partial h_t} \frac{\partial h_t}{\partial h_{t-k}} \frac{\partial h_{t-k}}{\partial W_h}$$

Vanilla RNN Equations

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

$$y_t = \zeta_y(W_y * h_t \ + \ b_y)$$

$$E_t = L(\hat{y_t}, y_t)$$



$$\frac{\partial E_t}{\partial W_h} = \frac{\partial E_t}{\partial \hat{y_t}} \frac{\partial \hat{y_t}}{\partial h_t} \frac{\partial h_t}{\partial h_{t-k}} \frac{\partial h_{t-k}}{\partial W_h}$$

$$\frac{\partial E_t}{\partial W_h} = \sum_{k=0}^{t-1} \frac{\partial E_t}{\partial \hat{y_t}} \frac{\partial \hat{y_t}}{\partial h_t} \frac{\partial h_t}{\partial h_{t-k}} \frac{\partial h_{t-k}}{\partial W_h}$$

$$\frac{\partial h_t}{\partial h_{t-k}} = \frac{\partial h_t}{\partial h_{t-1}} \frac{\partial h_{t-1}}{\partial h_{t-2}} \cdots \frac{\partial h_{t-k+1}}{\partial h_{t-k}} = \prod_{i=1}^{k} \frac{\partial h_{t-i+1}}{\partial h_{t-i}}$$

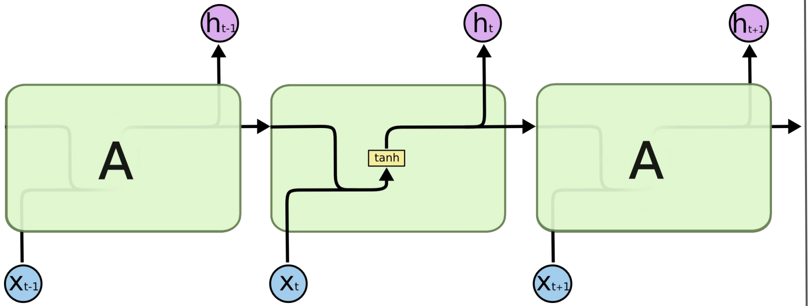
Vanilla RNN Equations

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

$$y_t = \zeta_y(W_y * h_t \ + \ b_y)$$

$$E_t = L(\hat{y_t}, y_t)$$

$$\frac{\partial E_t}{\partial W_h} = \frac{\partial E_t}{\partial \hat{y_t}} \frac{\partial \hat{y_t}}{\partial h_t} \frac{\partial h_t}{\partial h_{t-k}} \frac{\partial h_{t-k}}{\partial W_h}$$



$$\frac{\partial h_t}{\partial h_{t-k}} = \frac{\partial h_t}{\partial h_{t-1}} \frac{\partial h_{t-1}}{\partial h_{t-2}} \cdots \frac{\partial h_{t-k+1}}{\partial h_{t-k}} = \prod_{i=1}^{k} \frac{\partial h_{t-i+1}}{\partial h_{t-i}}$$

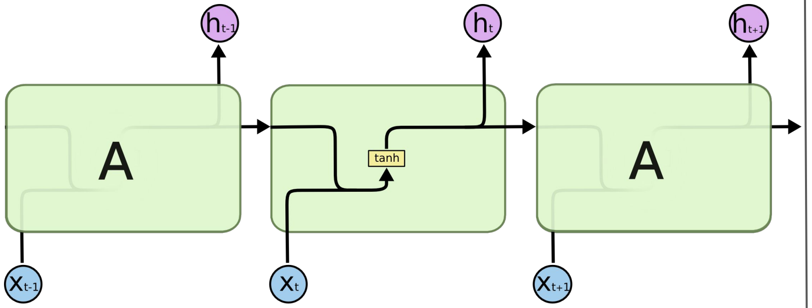
Vanilla RNN Equations

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

$$y_t = \zeta_y(W_y * h_t \ + \ b_y)$$

$$E_t = L(\hat{y_t}, y_t)$$

$$\frac{\partial E_t}{\partial W_h} = \frac{\partial E_t}{\partial \hat{y_t}} \frac{\partial \hat{y_t}}{\partial h_t} \frac{\partial h_t}{\partial h_{t-k}} \frac{\partial h_{t-k}}{\partial W_h}$$



$$\frac{\partial h_t}{\partial h_{t-k}} = \frac{\partial h_t}{\partial h_{t-1}} \frac{\partial h_{t-1}}{\partial h_{t-2}} \cdots \frac{\partial h_{t-k+1}}{\partial h_{t-k}} = \prod_{i=1}^{k} \frac{\partial h_{t-i+1}}{\partial h_{t-i}}$$

$$\frac{\partial h_k}{\partial h_{k-1}} = diag(\zeta_h'(W_x * x_k + W_h * h_{k-1} + b_h)) * W_h$$

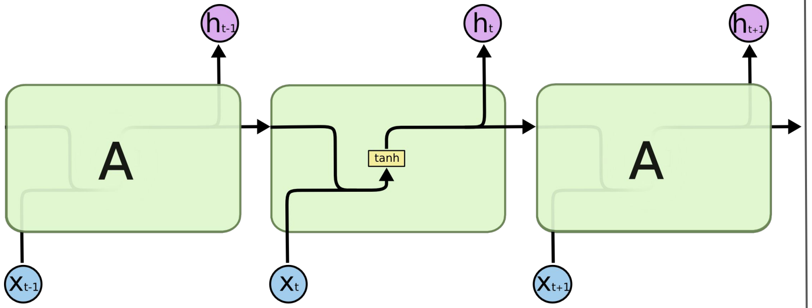
Vanilla RNN Equations

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

$$y_t = \zeta_y(W_y * h_t \ + \ b_y)$$

$$E_t = L(\hat{y}_t, y_t)$$

$$\frac{\partial E_t}{\partial W_h} = \frac{\partial E_t}{\partial \hat{y_t}} \frac{\partial \hat{y_t}}{\partial h_t} \frac{\partial h_t}{\partial h_{t-k}} \frac{\partial h_{t-k}}{\partial W_h}$$

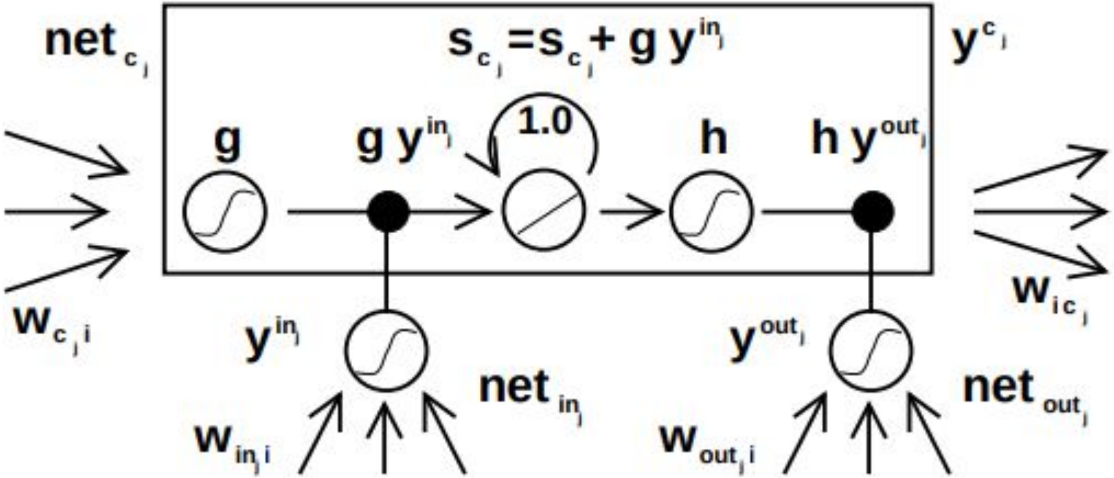


$$\frac{\partial h_t}{\partial h_{t-k}} = \frac{\partial h_t}{\partial h_{t-1}} \frac{\partial h_{t-1}}{\partial h_{t-2}} \cdots \frac{\partial h_{t-k+1}}{\partial h_{t-k}} = \prod_{i=1}^{k} \frac{\partial h_{t-i+1}}{\partial h_{t-i}}$$

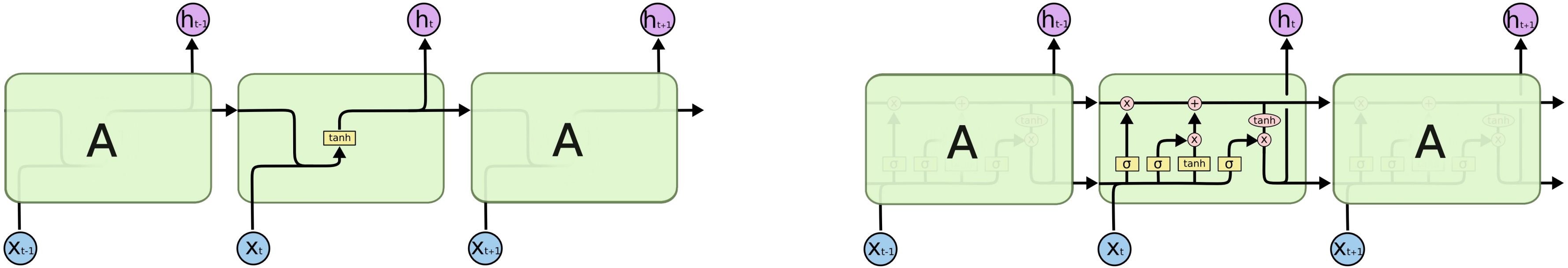
$$\frac{\partial h_k}{\partial h_{k-1}} = diag(\zeta_h'(W_x * x_k + W_h * h_{k-1} + b_h)) * W_h$$

$$\frac{\partial h_t}{\partial h_{t-k}} = W_h^k \prod_{i=1}^{k} diag(\zeta_h'(W_x * x_{t-i+1} + W_h * h_{t-i} + b_h))$$

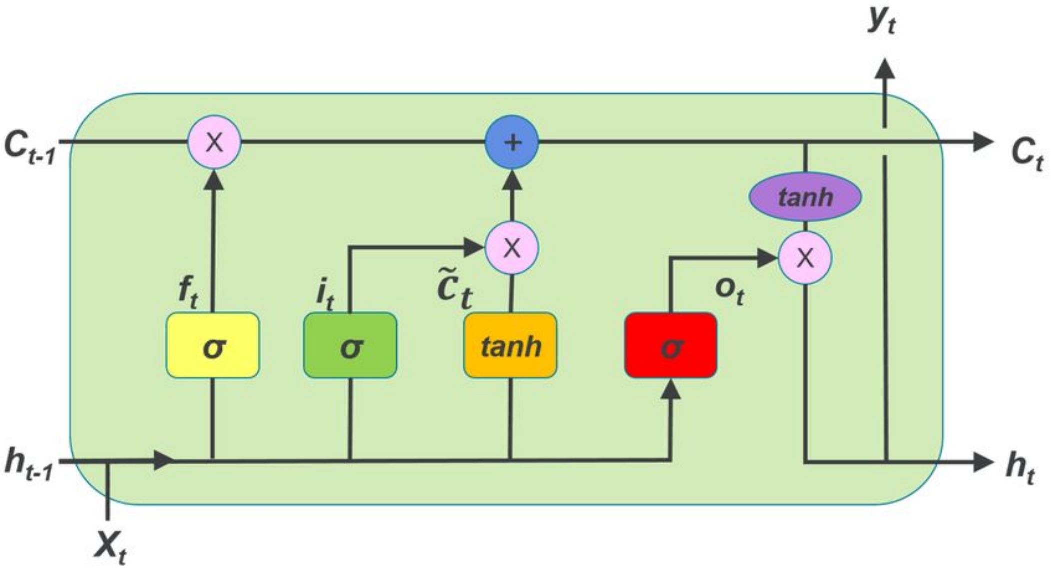
Solution: Just Do This



RNN to LSTM



- The cell, input gate, output gate, and forget gate
- The cell stores the information
- The gates control the information



$$i_t = \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{(t-1)} + b_{hi})$$

$$f_t = \sigma(W_{if}x_t + b_{if} + W_{hf}h_{(t-1)} + b_{hf})$$

$$\widetilde{c}_t = \tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{(t-1)} + b_{hg})$$

$$o_t = \sigma(W_{io}x_t + b_{io} + W_{ho}h_{(t-1)} + b_{ho})$$

$$c_t = f_t * c_{(t-1)} + i_t * \widetilde{c}_t$$

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

LSTM Equations

Gates

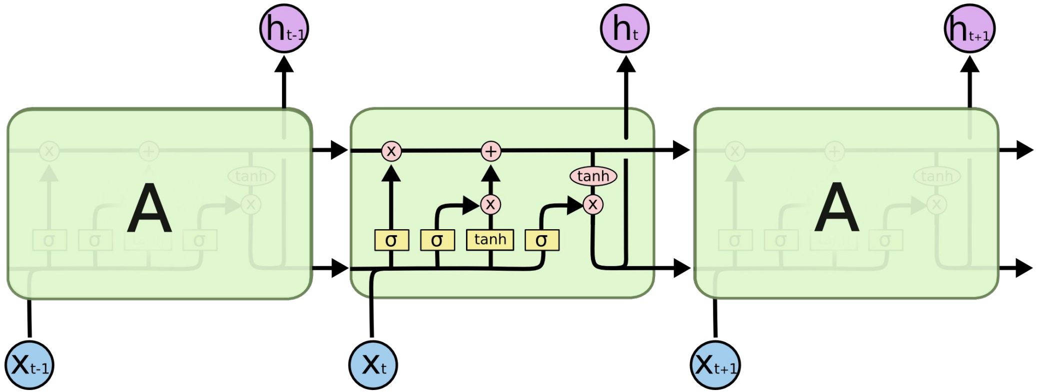
$$f_t = \sigma(W_f^x * x_t \ + \ W_f^h * h_{t-1} \ + \ b_f)$$
$$i_t = \sigma(W_i^x * x_t \ + \ W_i^h * h_{t-1} \ + \ b_i)$$
$$o_t = \sigma(W_o^x * x_t \ + \ W_o^h * h_{t-1} \ + \ b_o)$$

Canidate Values

$$\tilde{C}_t = tanh(W_c^x * x_t \ + \ W_c^h * h_{t-1} \ + \ b_c)$$

Update Rules

$$C_t = f_t \circ C_{t-1} \ + \ i_t \circ \tilde{C}_t$$
$$h_t = o_t \circ tanh(C_t)$$
$$\hat{y}_t = \zeta(W_y * h_t \ + \ b_y) \text{ (if needed)}$$
$$E_t = L(\hat{y}_t, y_t)$$



LSTM Equations

Gates

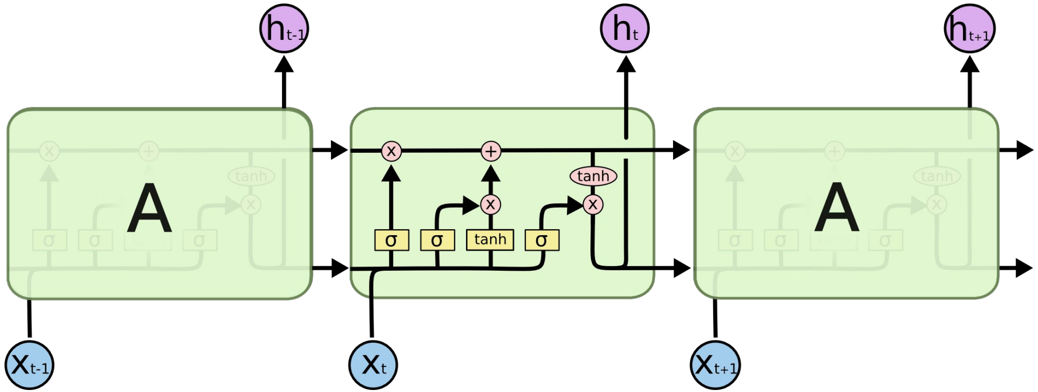
$$f_t = \sigma(W_f^x * x_t + W_f^h * h_{t-1} + b_f)$$
$$i_t = \sigma(W_i^x * x_t + W_i^h * h_{t-1} + b_i)$$
$$o_t = \sigma(W_o^x * x_t + W_o^h * h_{t-1} + b_o)$$

Canidate Values

$$\tilde{C}_t = tanh(W_c^x * x_t + W_c^h * h_{t-1} + b_c)$$

Update Rules

$$C_t = f_t \circ C_{t-1} + i_t \circ \tilde{C}_t$$
$$h_t = o_t \circ tanh(C_t)$$
$$\hat{y}_t = \zeta(W_y * h_t + b_y) \text{ (if needed)}$$
$$E_t = L(\hat{y}_t, y_t)$$



$$\frac{\partial E_t}{\partial W_j^h} = \sum_{k=1}^{t-1} \frac{\partial E_t}{\partial \hat{y_t}} \frac{\partial \hat{y_t}}{\partial h_t} \frac{\partial h_t}{\partial C_t} \frac{\partial C_t}{\partial C_{t-k}} \frac{\partial C_{t-k}}{\partial W_j^h}$$

LSTM Equations

Gates

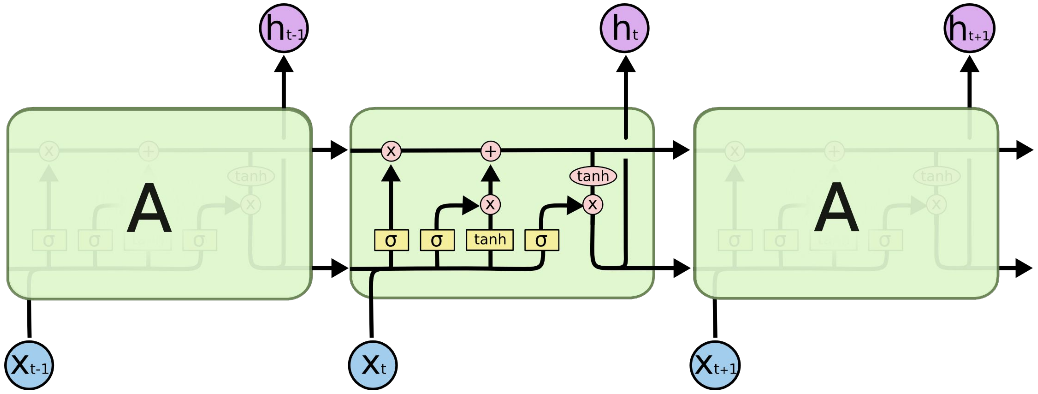
$$f_t = \sigma(W_f^x * x_t \ + \ W_f^h * h_{t-1} \ + \ b_f)$$
$$i_t = \sigma(W_i^x * x_t \ + \ W_i^h * h_{t-1} \ + \ b_i)$$
$$o_t = \sigma(W_o^x * x_t \ + \ W_o^h * h_{t-1} \ + \ b_o)$$

Canidate Values

$$\tilde{C}_t = tanh(W_c^x * x_t \ + \ W_c^h * h_{t-1} \ + \ b_c)$$

Update Rules

$$C_t = f_t \circ C_{t-1} \ + \ i_t \circ \tilde{C}_t$$
$$h_t = o_t \circ tanh(C_t)$$
$$\hat{y}_t = \zeta(W_y * h_t \ + \ b_y) \text{ (if needed)}$$
$$E_t = L(\hat{y}_t, y_t)$$



$$\frac{\partial E_t}{\partial W_j^h} = \sum_{k=1}^{t-1} \frac{\partial E_t}{\partial \hat{y_t}} \frac{\partial \hat{y_t}}{\partial h_t} \frac{\partial h_t}{\partial C_t} \frac{\partial C_t}{\partial C_{t-k}} \frac{\partial C_{t-k}}{\partial W_j^h}$$

$$\frac{\partial C_t}{\partial C_{t-k}} = \prod_{i=0}^{k} \frac{\partial C_{t-i+1}}{\partial C_{t-i}}$$

LSTM Equations

Gates

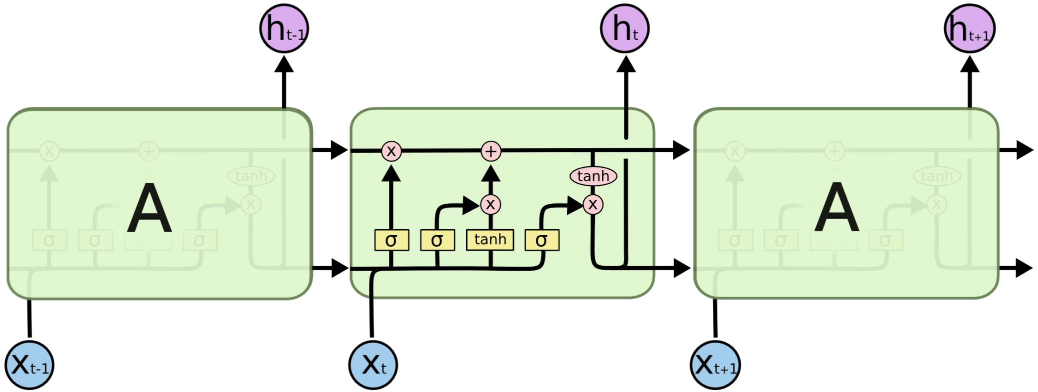
$$f_t = \sigma(W_f^x * x_t + W_f^h * h_{t-1} + b_f)$$
$$i_t = \sigma(W_i^x * x_t + W_i^h * h_{t-1} + b_i)$$
$$o_t = \sigma(W_o^x * x_t + W_o^h * h_{t-1} + b_o)$$

Canidate Values

$$\tilde{C}_t = tanh(W_c^x * x_t + W_c^h * h_{t-1} + b_c)$$

Update Rules

$$C_t = f_t \circ C_{t-1} + i_t \circ \tilde{C}_t$$
$$h_t = o_t \circ tanh(C_t)$$
$$\hat{y}_t = \zeta(W_y * h_t + b_y) \text{ (if needed)}$$
$$E_t = L(\hat{y}_t, y_t)$$



$$\frac{\partial E_t}{\partial W_j^h} = \sum_{k=1}^{t-1} \frac{\partial E_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial h_t} \frac{\partial h_t}{\partial C_t} \frac{\partial C_t}{\partial C_{t-k}} \frac{\partial C_{t-k}}{\partial W_j^h}$$

$$\frac{\partial C_t}{\partial C_{t-k}} = \prod_{i=0}^{k} \frac{\partial C_{t-i+1}}{\partial C_{t-i}}$$

$$\frac{\partial C_k}{\partial C_{k-1}} = \frac{\partial}{\partial C_k}(f_k \circ C_{k-1} + i_k \circ \tilde{C}_k)$$

LSTM Equations

Gates

$$f_t = \sigma(W_f^x * x_t \ + \ W_f^h * h_{t-1} \ + \ b_f)$$
$$i_t = \sigma(W_i^x * x_t \ + \ W_i^h * h_{t-1} \ + \ b_i)$$
$$o_t = \sigma(W_o^x * x_t \ + \ W_o^h * h_{t-1} \ + \ b_o)$$

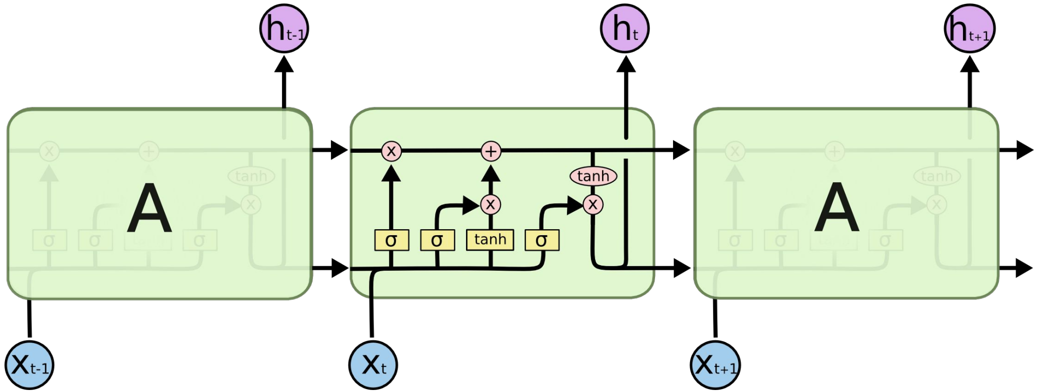
Canidate Values

$$\tilde{C}_t = tanh(W_c^x * x_t \ + \ W_c^h * h_{t-1} \ + \ b_c)$$

Update Rules

$$C_t = f_t \circ C_{t-1} \ + \ i_t \circ \tilde{C}_t$$
$$h_t = o_t \circ tanh(C_t)$$
$$\hat{y}_t = \zeta(W_y * h_t \ + \ b_y) \text{ (if needed)}$$
$$E_t = L(\hat{y}_t, y_t)$$

$$\frac{\partial C_t}{\partial C_{t-k}} = \prod_{i=0}^{k} \frac{\partial C_{t-i+1}}{\partial C_{t-i}}$$



$$\frac{\partial C_k}{\partial C_{k-1}} = \frac{\partial}{\partial C_{k-1}}(f_k \circ C_{k-1} + i_t \circ \tilde{C}_k)$$

LSTM Equations

Gates

$$f_t = \sigma(W_f^x * x_t + W_f^h * h_{t-1} + b_f)$$
$$i_t = \sigma(W_i^x * x_t + W_i^h * h_{t-1} + b_i)$$
$$o_t = \sigma(W_o^x * x_t + W_o^h * h_{t-1} + b_o)$$

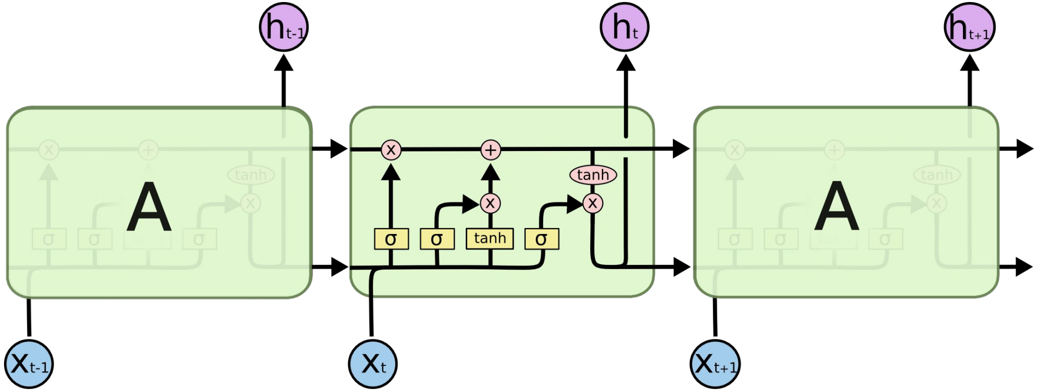
Canidate Values

$$\tilde{C}_t = tanh(W_c^x * x_t + W_c^h * h_{t-1} + b_c)$$

Update Rules

$$C_t = f_t \circ C_{t-1} + i_t \circ \tilde{C}_t$$
$$h_t = o_t \circ tanh(C_t)$$
$$\hat{y}_t = \zeta(W_y * h_t + b_y) \text{ (if needed)}$$
$$E_t = L(\hat{y}_t, y_t)$$

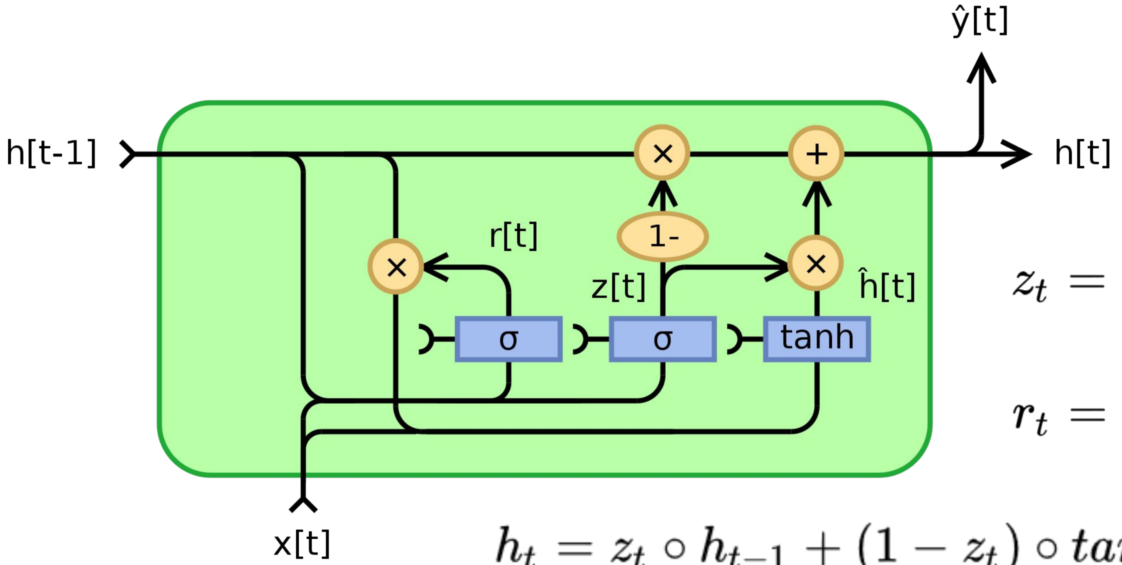
$$\frac{\partial C_t}{\partial C_{t-k}} = \prod_{i=0}^{k} \frac{\partial C_{t-i+1}}{\partial C_{t-i}}$$



$$\frac{\partial C_k}{\partial C_{k-1}} = \frac{\partial}{\partial C_{k-1}}(f_k \circ C_{k-1} + i_t \circ \tilde{C}_k)$$

$$\frac{\partial C_k}{\partial C_{k-1}} = diag(C_{k-1})\frac{\partial f_k}{\partial C_{k-1}} + diag(f_k) + diag(\tilde{C}_k)\frac{\partial i_k}{\partial C_{k-1}} + diag(i_k)\frac{\partial \tilde{C}_k}{\partial C_{k-1}}$$

Gated Recurrent Unit(GRU)



$$z_t = \sigma(W_z * x_t + U_z * h_{t-1} + b_z)$$

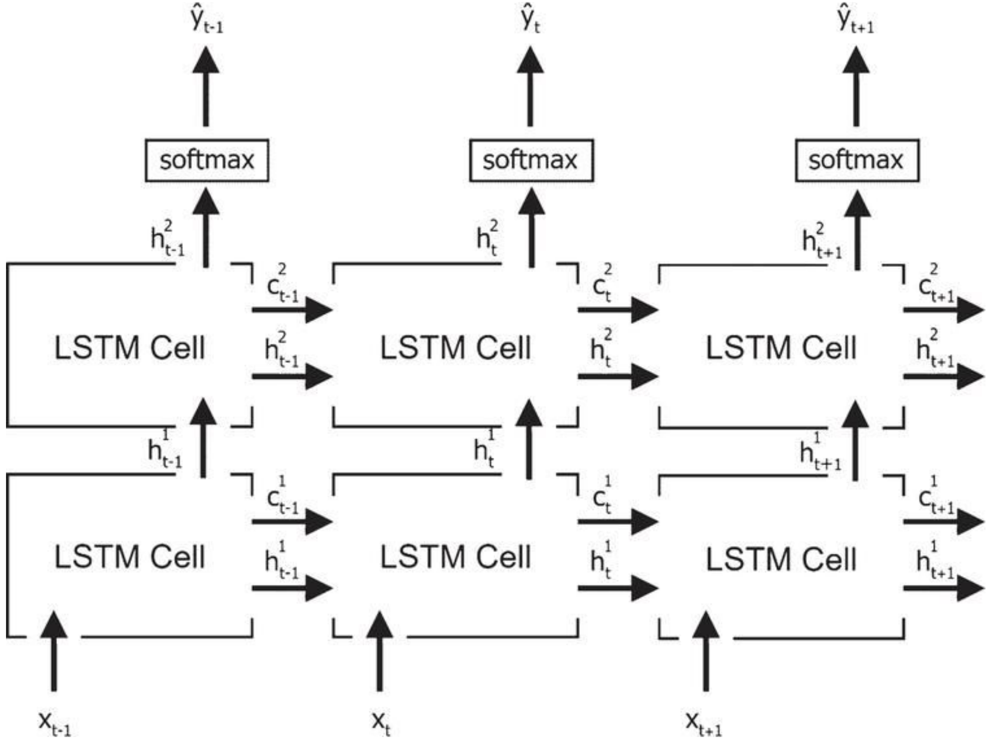
$$r_t = \sigma(W_r * x_t + U_r * h_{t-1} + b_r)$$

$$h_t = z_t \circ h_{t-1} + (1 - z_t) \circ tanh(W_h * x_t + U_h * (r_t \circ h_{t-1}))$$

$$\frac{\partial h_t}{\partial h_{t-1}}$$

Is additive and well behaved just like in LSTM.

Stacked LSTM



Sources

LSTM:

https://weberna.github.io/blog/2017/11/15/LSTM-Vanishing-Gradients.html (Highly recommended)
https://colah.github.io/posts/2015-08-Understanding-LSTMs/ (Highly Recommended)

Matrix Calculus.
https://explained.ai/matrix-calculus/

Discussion

1. Why does LSTM outperform GRU on complicated Data
2. What does stacking LSTM layers do?
3. Can you think of any variants of lstm?(there are a ton).