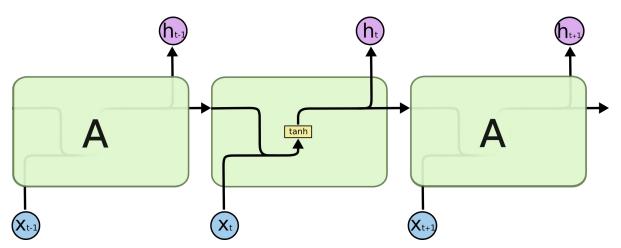
Long Short-Term Memory

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.



Input: xt

Output: yt,ht

Activation: tanh

Vanilla RNN Equations

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

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

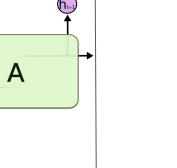
$$E_t = L(\hat{y}_t, y_t)$$

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

Α

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

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



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

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

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

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

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

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

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

 ∂E_t

 $\partial E_t \hspace{0.1cm} \hat{\partial y_t} \hspace{0.1cm} \partial h_t \hspace{0.1cm} \underline{\partial h_{t-k}}$ $= \overline{\partial \hat{y_t}} \; \overline{\partial h_t} \; \overline{\partial h_{t-k}} \; \overline{\partial W_h}$ $rac{\partial E_t}{\partial W_h} = \sum_{k=0}^{t-1} rac{\partial E_t}{\partial \hat{y_t}} rac{\partial \hat{y_t}}{\partial h_t} rac{\partial h_t}{\partial h_{t-k}} rac{\partial h_{t-k}}{\partial W_h}$

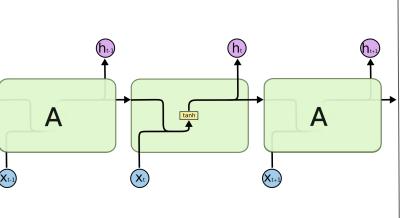
Vanilla RNN Equations

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

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

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

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



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

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

Vanilla RNN Equations

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

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

$$h_t = \zeta_h(W_x * x_t + W_h * h_{t-1} + b_h) \qquad \frac{\partial h_t}{\partial h_{t-k}} = \frac{\partial h_t}{\partial h_{t-1}} \frac{\partial h_{t-1}}{\partial h_{t-2}} \dots \frac{\partial h_{t-k+1}}{\partial h_{t-k}} = \prod_{i=1}^n \frac{\partial h_{t-i}}{\partial h_{t-i}}$$

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

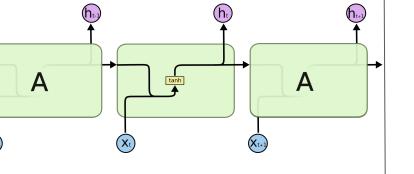
Vanilla RNN Equations $h_t = \zeta_h(W_x * x_t + W_h * h_{t-1} + b_h)$

$$\frac{\partial}{\partial t}$$

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

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

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



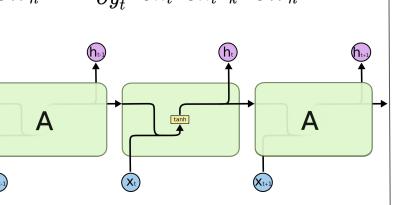
$\frac{\partial h_t}{\partial h_{t-k}} = \frac{\partial h_t}{\partial h_{t-1}} \frac{\partial h_{t-1}}{\partial h_{t-2}} \dots \frac{\partial h_{t-k+1}}{\partial h_{t-k}} = \prod_{i=1}^k \frac{\partial h_{t-i+1}}{\partial h_{t-i}}$ $rac{\partial h_k}{\partial h_{k-1}} = diag(\zeta_h^{'}(W_x st x_k + W_h st h_{k-1} + b_h)) st W_h$

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

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

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

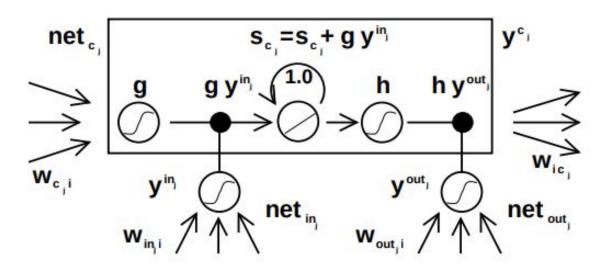
$$rac{E_t}{W_h} = rac{\partial E_t}{\partial \hat{y_t}} rac{\partial \hat{y_t}}{\partial h_t} rac{\partial h_t}{\partial h_{t-k}} rac{\partial h_{t-k}}{\partial W_h}$$

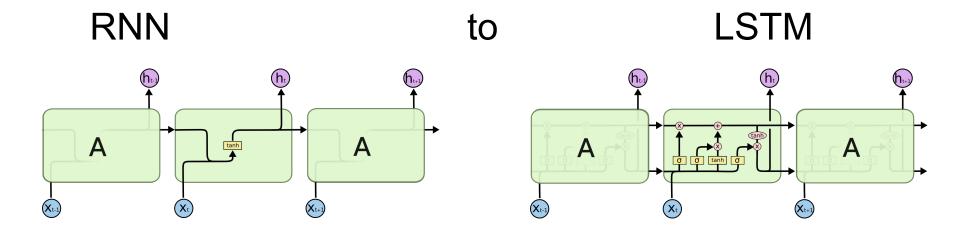


 $egin{aligned} rac{\partial h_t}{\partial h_{t-k}} &= rac{\partial h_t}{\partial h_{t-1}} rac{\partial h_{t-1}}{\partial h_{t-2}} \dots rac{\partial h_{t-k+1}}{\partial h_{t-k}} = \prod_{i=1}^k rac{\partial h_{t-i+1}}{\partial h_{t-i}} \ rac{\partial h_k}{\partial h_{k-1}} &= diag(\zeta_h'(W_x * x_k + W_h * h_{k-1} + b_h)) * W_h \end{aligned}$

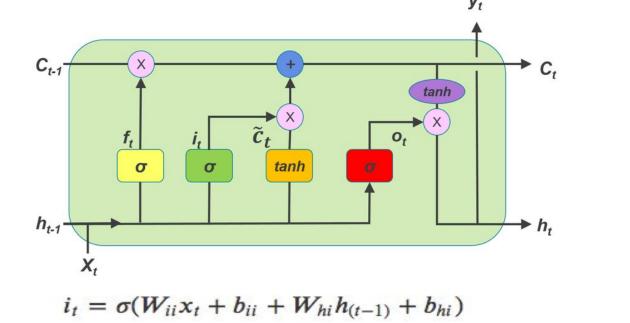
$$oxed{rac{\partial h_t}{\partial h_{t-k}}} = W_h^k \prod_{i=1}^k diag(\zeta_h^{'}(W_x * x_{t-i+1} + W_h * h_{t-i} + b_h))$$

Solution: Just Do This





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



$$h_{t-1}$$

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

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

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

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

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

 $h_t = o_t * \tanh(c_t)$

Gates

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

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

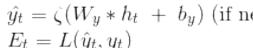
Update Rules

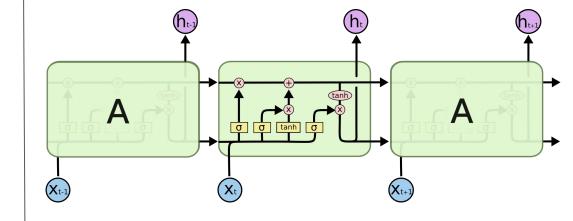
$$C_t = f_t \circ C_{t-1} + i_t \circ \tilde{C}_t$$

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

$$n_t = o_t \circ tann(C)$$

$$\hat{y_t} = \zeta(W_y * h_t + b_y)$$
 (if needed)





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

$$\tilde{C}_t = tanh(W^x *$$

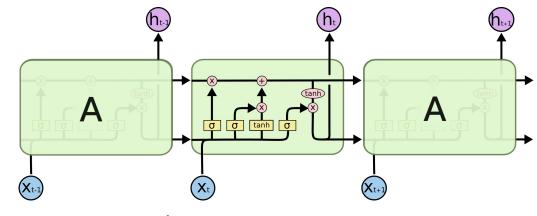
 $\tilde{C}_t = tanh(W_c^x * x_t + W_c^h * h_{t-1} + b_c)$

Update Rules

$$C_t = f_t \circ C_{t-1} + i_t \circ C_t$$

 $h_t = o_t \circ tanh(C_t)$

$$\hat{y_t} = \zeta(W_y * h_t + b_y)$$
 (if needed)
 $E_t = L(\hat{y_t}, y_t)$



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

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

$$C_t = f_t \circ C_{t-1} + i_t \circ C_t$$

 $h_t = o_t \circ tanh(C_t)$

$$h_t = o_t \circ tanh(C_t$$

 $\hat{u}_t = \mathcal{E}(W * h_t)^{\perp}$

$$\hat{y_t} = \zeta(W_y * h_t + b_y)$$
 (if needed)
 $E_t = L(\hat{y_t}, y_t)$

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

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

$$rac{\partial C_{t-i+1}}{\partial C_{t-i}}$$

Gates

$$f_t = \sigma(W_f^x * x_t + W_f^h * h_{t-1} + b_f)$$

$$i_t = \sigma(W_i^x * x_t + W_i^h * h_{t-1} + b_i)$$

Canidate Values

$$\tilde{C}_t = tanh(W_*^x *$$

Update Rules

$$C_t = f_t \circ C_{t-1} + h_t = o_t \circ tanh(C_t)$$

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

 $\hat{u}_t = \mathcal{E}(W_u * h_t) +$

 $h_t = o_t \circ tanh(C_t)$

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

$$o_t = \sigma(W_o^x * x_t + W_o^h * h_{t-1} + b_o)$$

$$Canidate\ Values$$

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

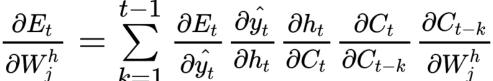
$$\lambda$$

$$\frac{\partial E_t}{\partial \hat{u}_t} \frac{\partial}{\partial t}$$





$$\frac{\partial C}{\partial C_{t}}$$



$$C_t = f_t \circ C_{t-1} + i_t \circ \tilde{C}_t \ h_t = o_t \circ tanh(C_t) \ \hat{y}_t = \zeta(W_y * h_t + b_y) \text{ (if needed)} \ E_t = L(\hat{y}_t, y_t) \ } \qquad \qquad \frac{\partial C_t}{\partial C_{t-k}} = \prod_{i=0}^k \frac{\partial C_{t-i+1}}{\partial C_{t-i}}$$

$$rac{\partial C_t}{\partial C_{t-k}} =$$

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

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

 $C_t = f_t \circ C_{t-1} + i_t \circ C_t$ $h_t = o_t \circ tanh(C_t)$

 $\hat{y_t} = \zeta(W_u * h_t + b_u)$ (if needed) $E_t = L(\hat{y}_t, y_t)$

 $rac{\partial C_k}{\partial C_{k-1}} = rac{\partial}{\partial C_{k-1}} (f_k \circ C_{k-1} + i_t \circ ilde{C_k})$

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

Update Rules

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

 $h_t = o_t \circ tanh(C_t)$

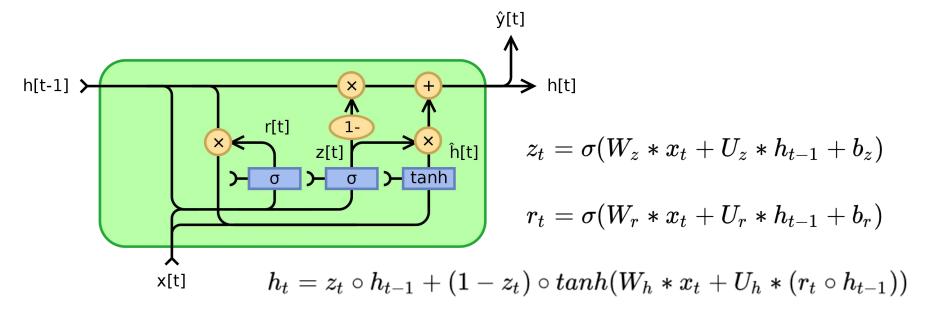
 $\hat{y_t} = \zeta(W_u * h_t + b_u)$ (if needed)

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

 $rac{\partial C_k}{\partial C_{k-1}} = rac{\partial}{\partial C_{k-1}} (f_k \circ C_{k-1} + i_t \circ ilde{C_k})$

 $rac{\partial C_k}{\partial C_{k-1}} = diag(C_{k-1})rac{\partial f_k}{\partial C_{k-1}} + diag(f_k) + diag(ilde{C_k})rac{\partial i_k}{\partial C_{k-1}} + diag(i_k)rac{\partial ilde{C_k}}{\partial C_{k-1}}$

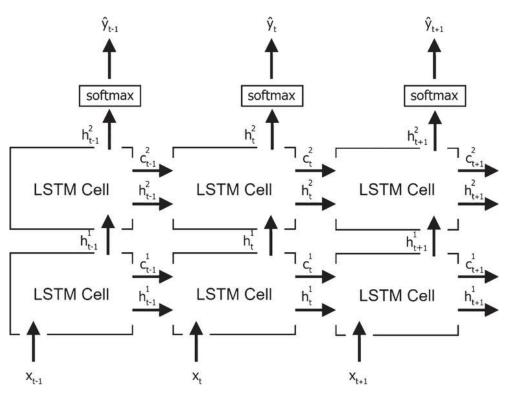
Gated Recurrent Unit(GRU)



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

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

- 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).