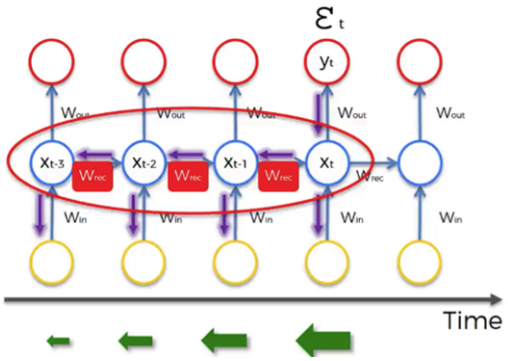
Long Short-Term Memory (LSTM)

August 1, 2024

Seoul National University

Young rae Cho

Problem of RNN



$$\frac{\partial \mathcal{E}}{\partial \theta} = \sum_{1 \le t \le T} \frac{\partial \mathcal{E}_t}{\partial \theta} \quad (3)$$

$$\frac{\partial \mathcal{E}_t}{\partial \theta} = \sum_{1 \le k \le t} \left(\frac{\partial \mathcal{E}_t}{\partial \mathbf{x}_t} \frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} \frac{\partial^+ \mathbf{x}_k}{\partial \theta} \right) \quad (4)$$

$$\frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} = \prod_{t \ge i > k} \frac{\partial \mathbf{x}_i}{\partial \mathbf{x}_{i-1}} = \prod_{t \ge i > k} \mathbf{W}_{rec}^T diag(\sigma'(\mathbf{x}_{i-1})) \quad (5)$$

W_{rec} ~ small ⇨ Vanishing

W_{rec} ~ large ⇨ Exploding

- Long-term dependency problem

Problem of RNN

$$|f'_{l_m}(net_{l_m}(t-m))w_{l_m l_{m-1}}|$$

- $net_{l_m}(t-m)$

  the aggregated input received by neuron l at time $t-m$.

- $f'_{l_m}(net_{l_m}(t-m))$

  the derivative of the activation function f at neuron l in layer m.

- $w_{l_m l_{m-1}}$

  the weight of the connection from neuron l in layer $m-1$ to neuron l in layer m

Problem of RNN

- $|f'_{l_m}(net_{l_m}(t-m))w_{l_m l_{m-1}}| > 1.0$

  : gradient exploding problem

- $|f'_{l_m}(net_{l_m}(t-m))w_{l_m l_{m-1}}| < 1.0$

  : gradient vanishing problem

Naïve approach of avoiding gradient vanishing

- **A single Unit**

$$\vartheta_j(t) = f_j'(net_j(t))\vartheta_j(t+1)w_{jj}.$$

$$f_j'(net_j(t))w_{jj} = 1.0.$$

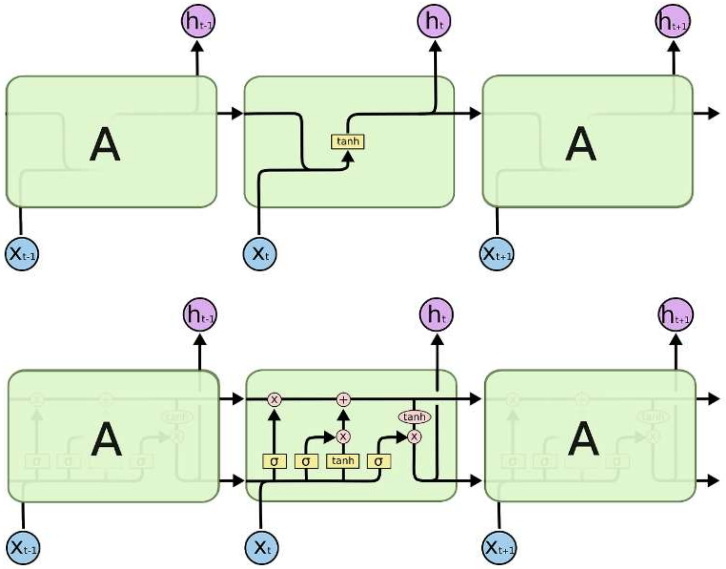
- **The Constant Error Carousel (CEC)**

$$f_j(net_j(t)) = \frac{net_j(t)}{w_{jj}}$$

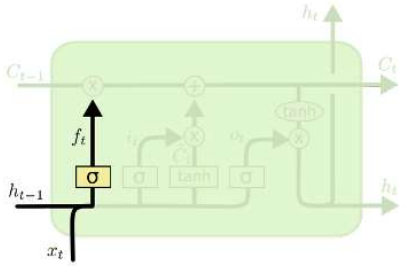
$$y_j(t+1) = f_j(net_j(t+1)) = f_j(w_{jj}y^j(t)) = y^j(t).$$

$f_j(x) = x, \forall x$, and by setting $w_{jj} = 1.0$.

The concept of LSTM



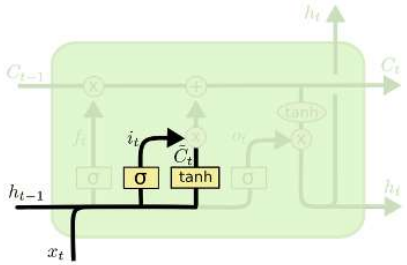
The concept of LSTM



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

LSTM의 forget gate layer

The concept of LSTM

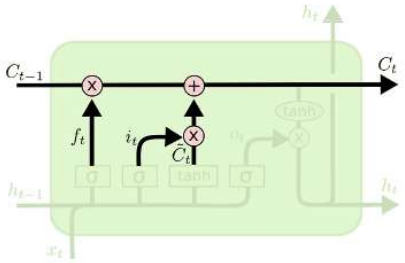


$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

LSTM의 input gate layer

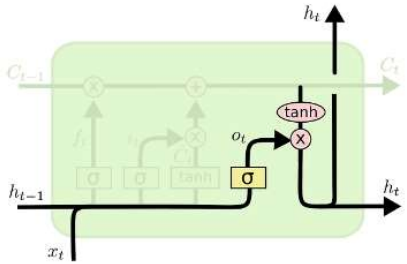
The concept of LSTM



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

LSTM의 cell state 업데이트

The concept of LSTM



$$o_t = \sigma\left(W_o\ [h_{t-1}, x_t]\ +\ b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$

LSTM의 output gate layer

Advantages of LSTM

- Long Time Lags
- Generalization
- Parameter Robustness
- Computational efficiency

Limitation of LSTM

- Delayed XOR Problem

  [0,1,0,0…….,1,0]

- Very long sequences

- Counting Discrete Time Steps

Reference

https://www.superdatascience.com/blogs/recurrent-neural-networks-rnn-the-vanishing-gradient-problem

https://dgkim5360.tistory.com/entry/understanding-long-short-term-memory-lstm-kr