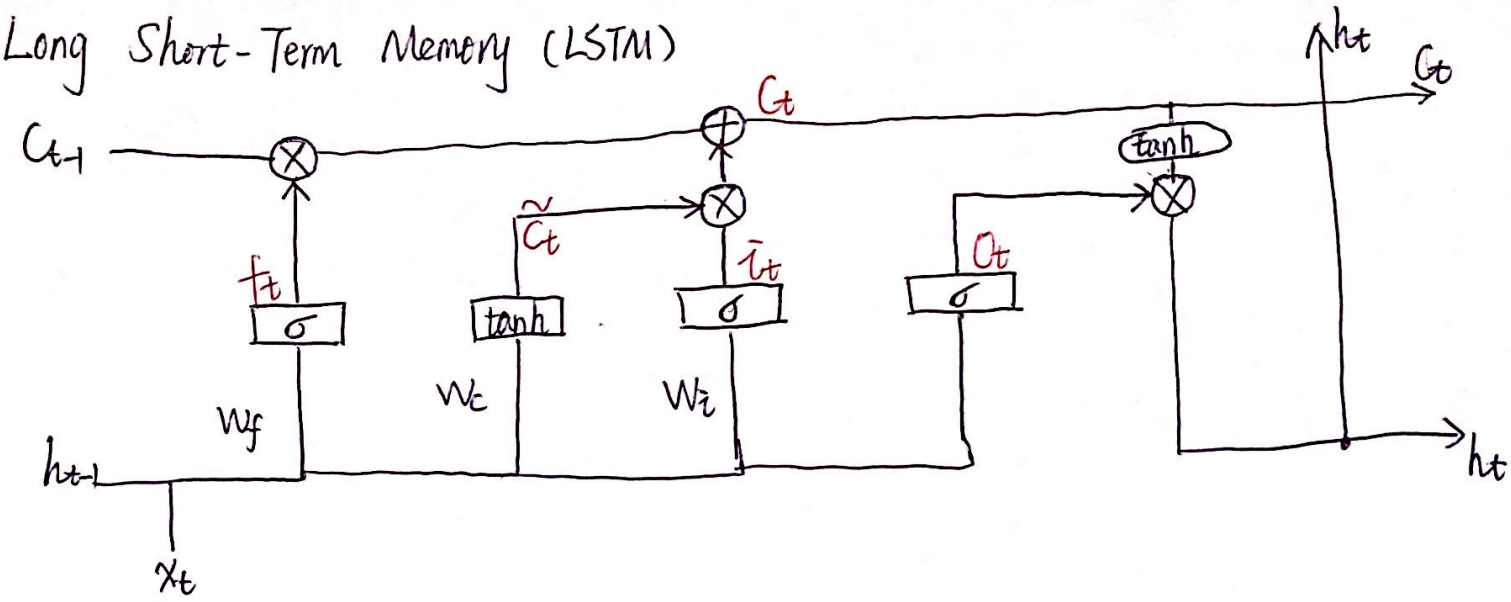
Long Short-Term Memory (LSTM)



C_t: knowledge encoded in C_t captures long-term dependencies and relations in the sequential ~~order~~ data

h_t: predictive vectors (hidden state)

3 gates: forget (f_t) / input (i_t) / output (O_t) gates

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t]) \qquad \tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t])$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t]) \qquad C_t = f_t \otimes C_{t-1} \oplus i_t \otimes \tilde{C}_t$$

$$O_t = \sigma(W_o \cdot [h_{t-1}, x_t]) \qquad h_t = O_t \otimes \tanh(C_t)$$

BPTT:

$$\frac{\partial E_k}{\partial w} = \frac{\partial E_k}{\partial h_k} \cdot \frac{\partial h_k}{\partial C_k} \cdot \left(\prod_{t=2}^{k} \frac{\partial C_t}{\partial C_{t-1}}\right) \cdot \frac{\partial C_1}{\partial w} \quad ②$$

$$\frac{\partial C_t}{\partial C_{t-1}} = \frac{\partial}{\partial C_{t-1}}(C_{t-1} \otimes f_t) + \frac{\partial}{\partial C_{t-1}}(\tilde{C}_t \otimes i_t)$$

$$= \underbrace{\frac{\partial f_t}{\partial C_{t-1}} C_{t-1}}_{①} + \underbrace{\frac{\partial C_{t-1}}{\partial C_{t-1}} \cdot f_t}_{②} + \underbrace{\frac{\partial i_t}{\partial C_{t-1}} \cdot \tilde{C}_t}_{③} + \underbrace{\frac{\partial \tilde{C}_t}{\partial C_{t-1}} \cdot i_t}_{④}$$

$= \sigma'(W_f \cdot [h_{t-1}, x_t]) \cdot W_f \cdot O_{t-1} \otimes \tanh'(C_{t-1}) \cdot C_{t-1}$ ① = A

$+ f_t$ ② = B

$+ \sigma'(W_i \cdot [h_{t-1}, x_t]) \cdot W_i \cdot O_{t-1} \otimes \tanh'(C_{t-1}) \cdot \tilde{C}_t$ ③ = C

$+ \tanh'(W_c \cdot [h_{t-1}, x_t]) \cdot W_c \cdot O_{t-1} \otimes \tanh'(C_{t-1}) \cdot i_t$ ④ = D

$$② = \frac{\partial E_k}{\partial h_k} \cdot \frac{\partial h_k}{\partial c_k} \left(\prod_{t=2}^{k} [A_t + B_t + C_t + D_t] \right) \cdot \frac{\partial c_1}{\partial w}$$

BPTT of LSTM 不易发生梯度消失/爆炸的原因：

(1) 连乘项中，$B_t = f_t \in (0,1)$，只要控制好 f_t 的取值使其不至于太小/太大梯度不消失/爆炸不会出现

(2) 连乘项是四项加和 (additive property) 不再是单个值，四项加和不至于趋近于0，可以通过控制使其接近1.