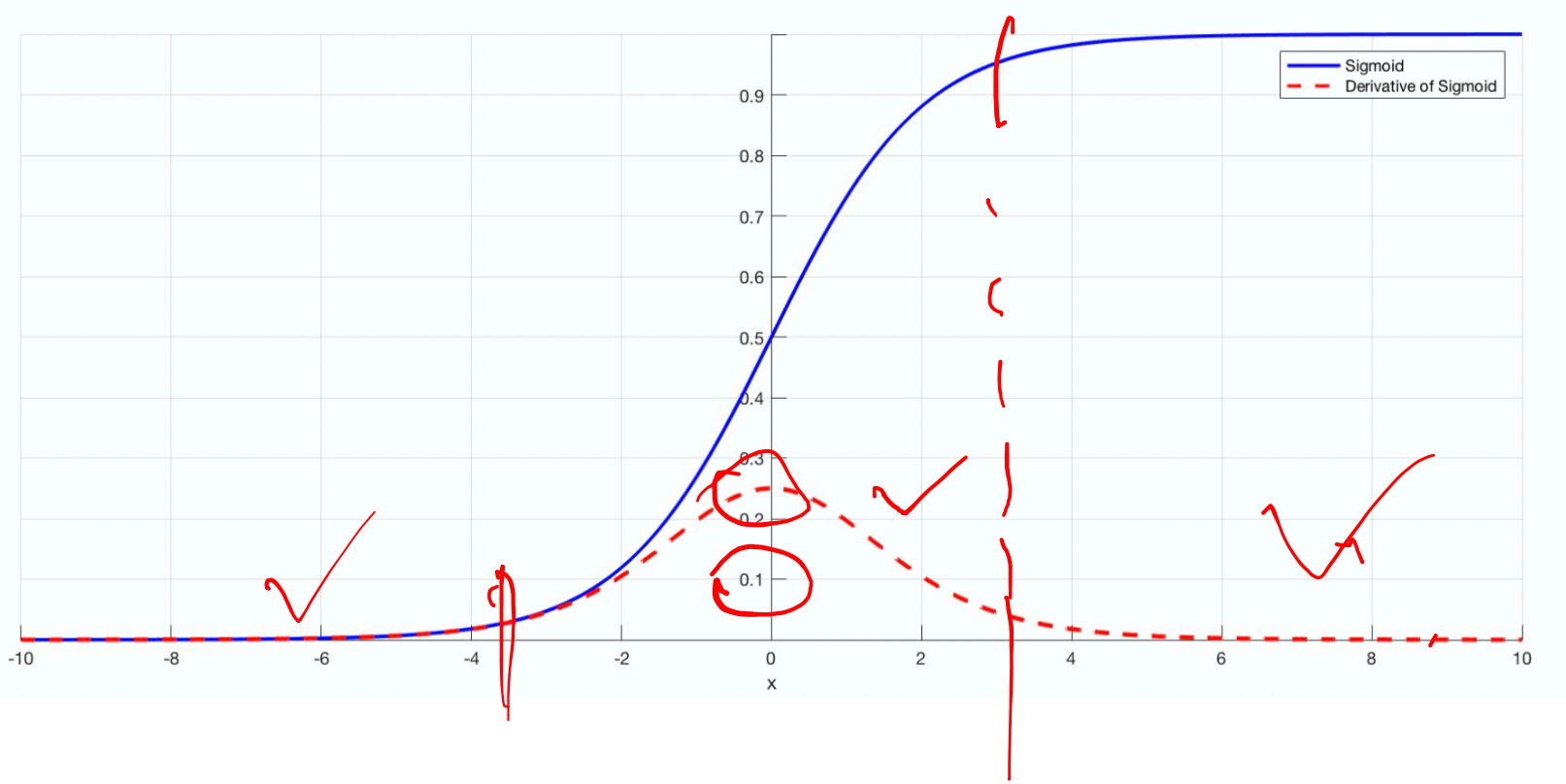
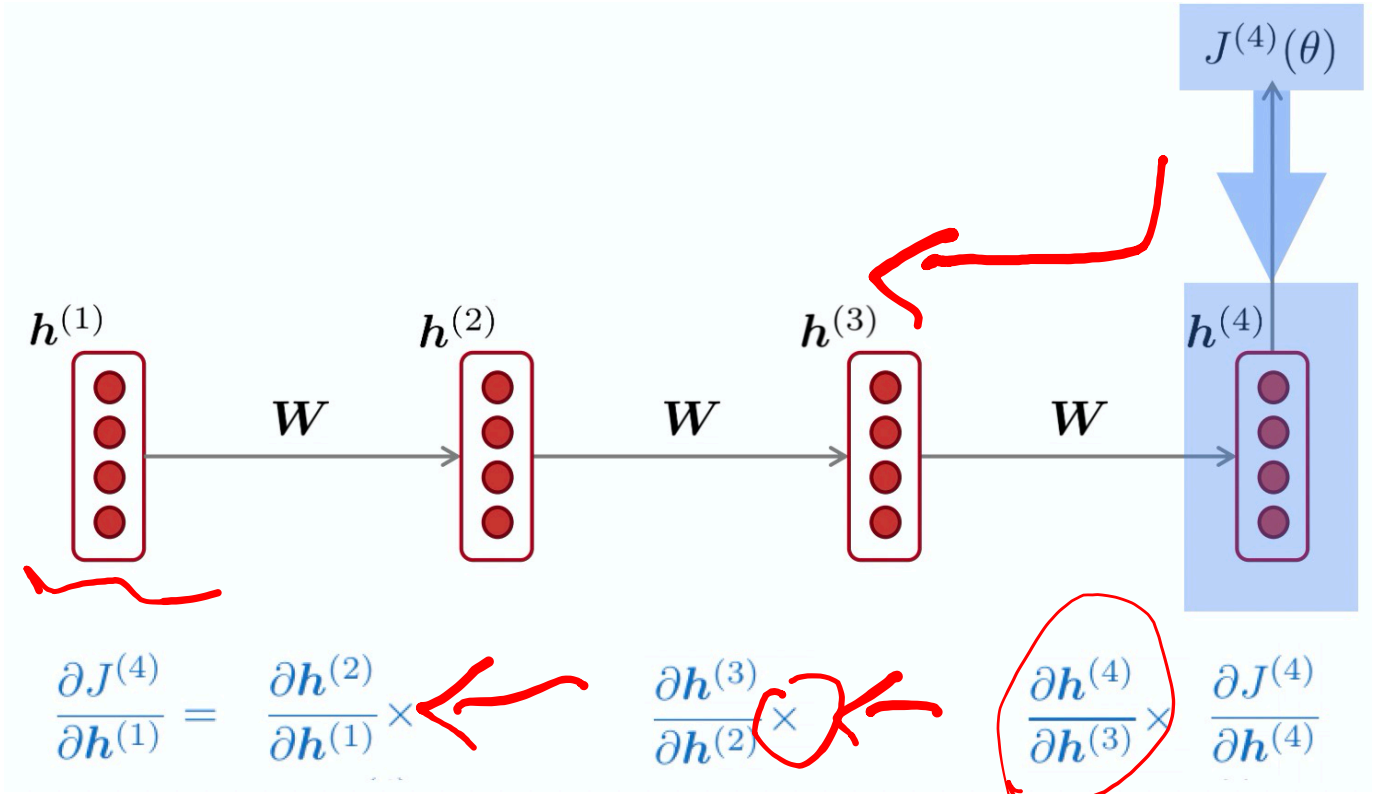
Exploding and Vanishing Gradient Problem



Exploding and Vanishing Gradient Problem



Chain rule

$$h^{(t)} = \sigma(W_h h^{(t-1)} + W_x x^{(t)} + b_1)$$

$$\frac{\partial h^{(t)}}{\partial h^{(t-1)}} = diag(\sigma'(W_h h^{(t-1)} + W_x x^{(t)} + b_1)) W_h$$

$\sigma' \cdot \sigma_1' \cdot \sigma_2'$

sigmoid

Exploding Gradient

gradient clipping

Vanishing Gradient

Identity Initialization

LSTM

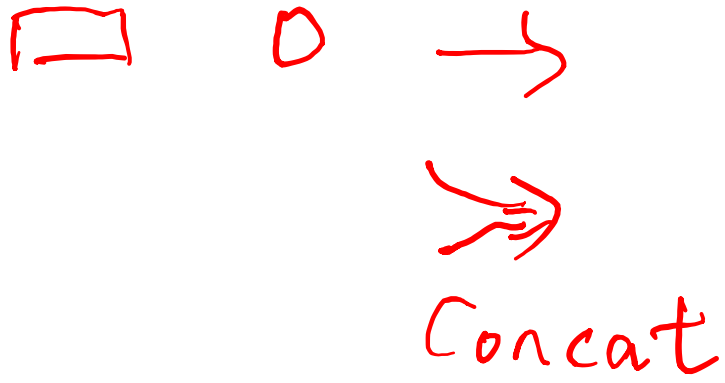
Residual Networks

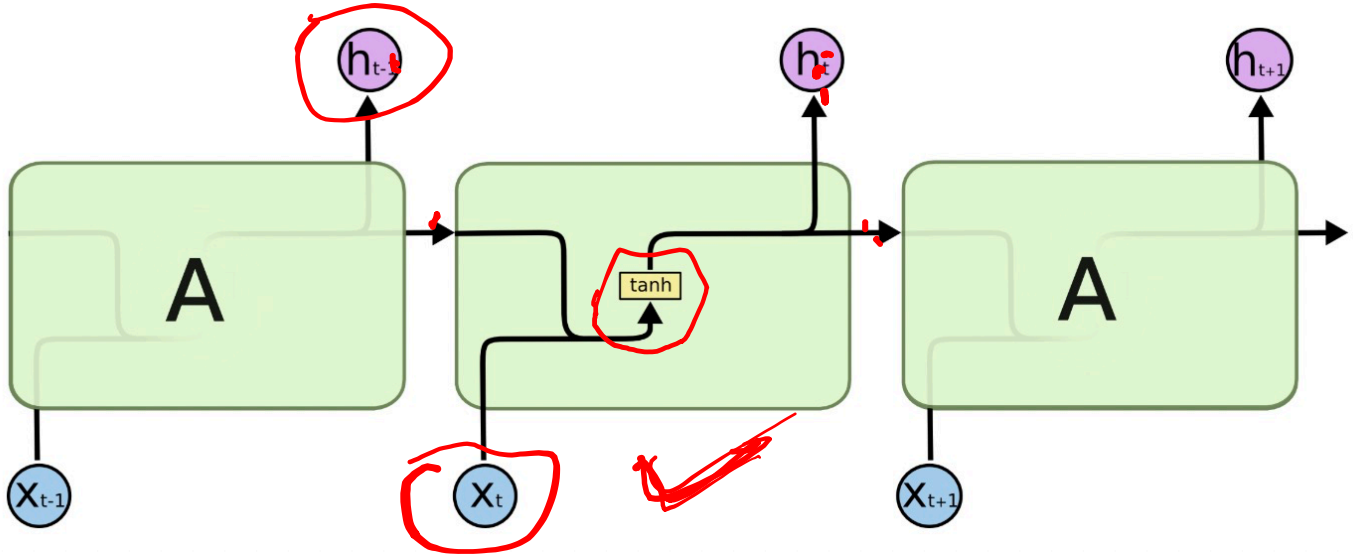
Batch Normalization

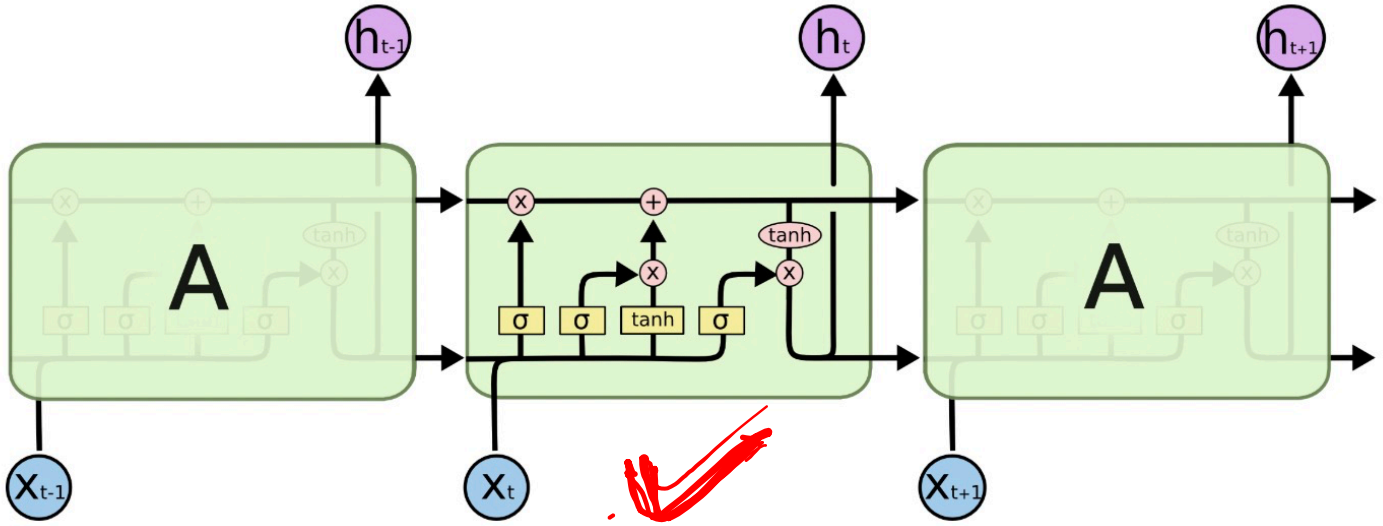
21:05亿

LSTM

Long Short Term Memory







LSTM

3 [i]

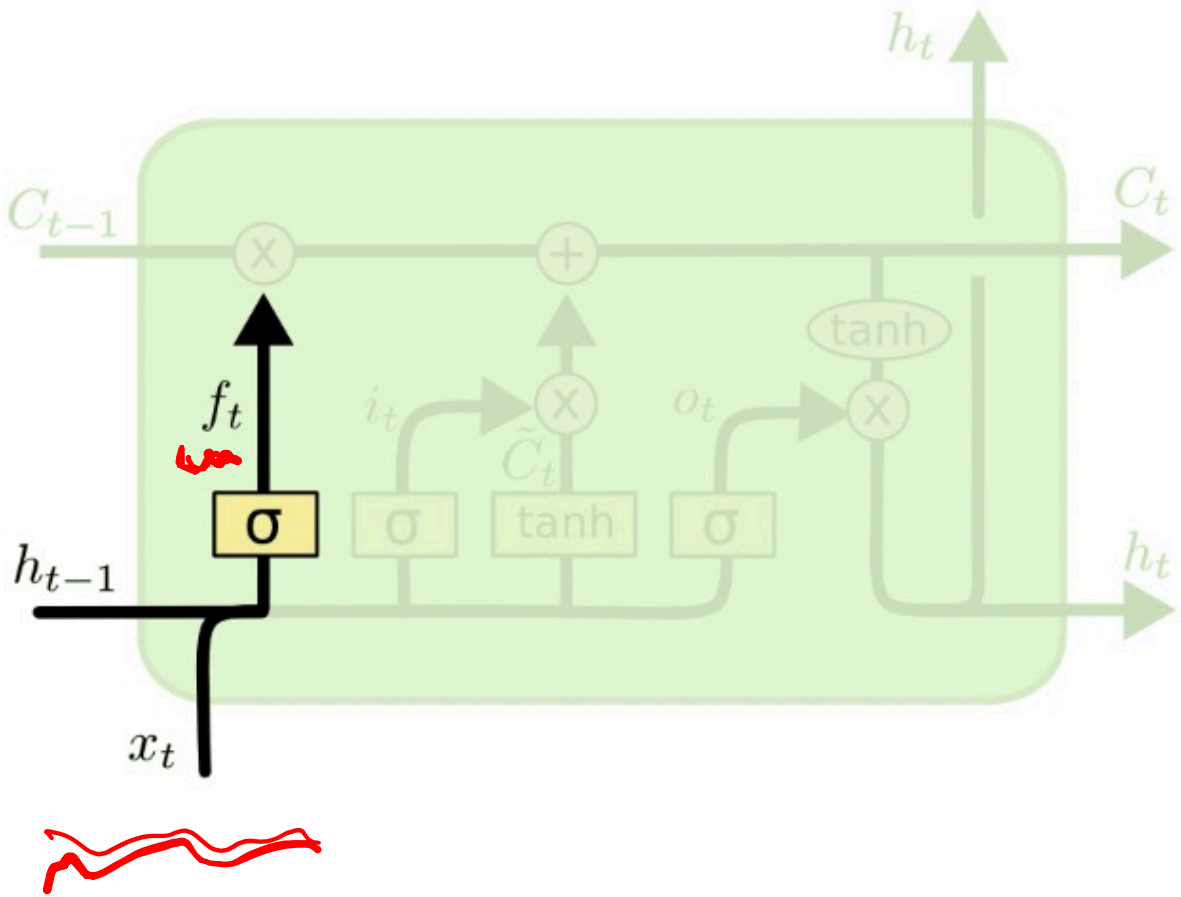
input

output

forget

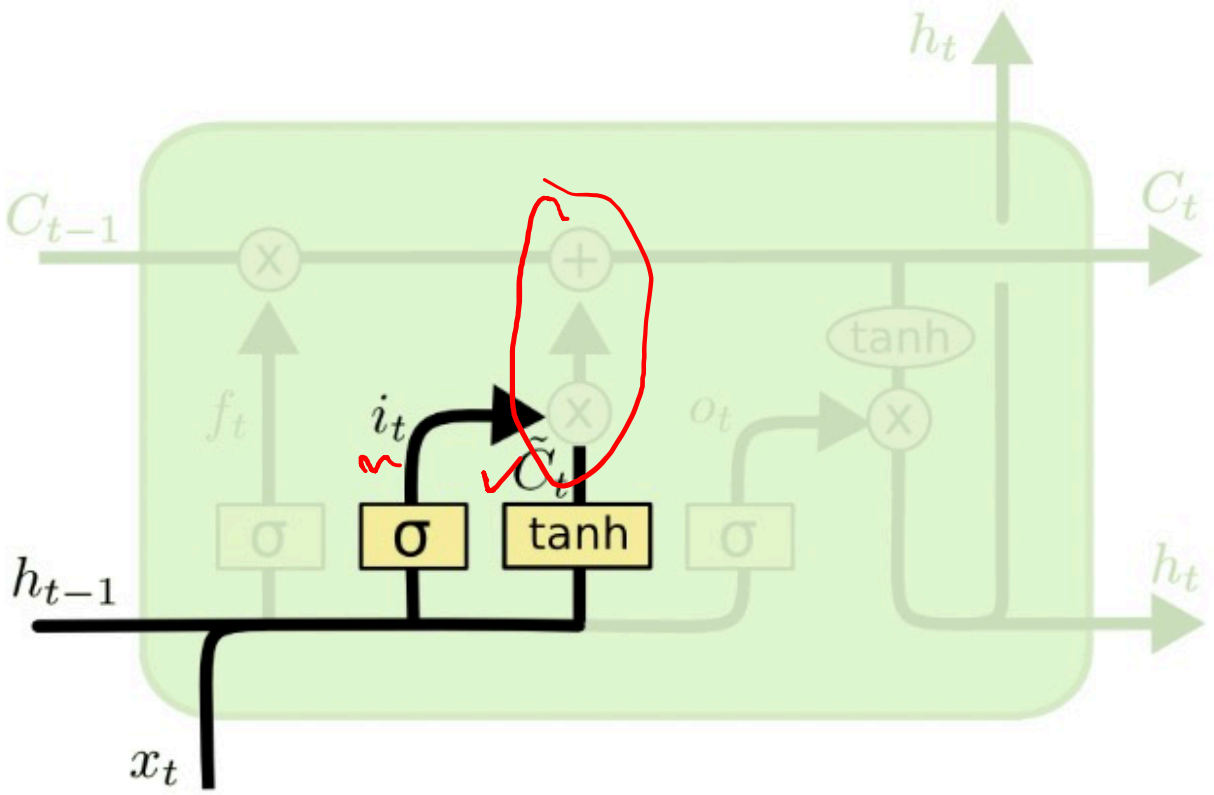
$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

sig



LSTM

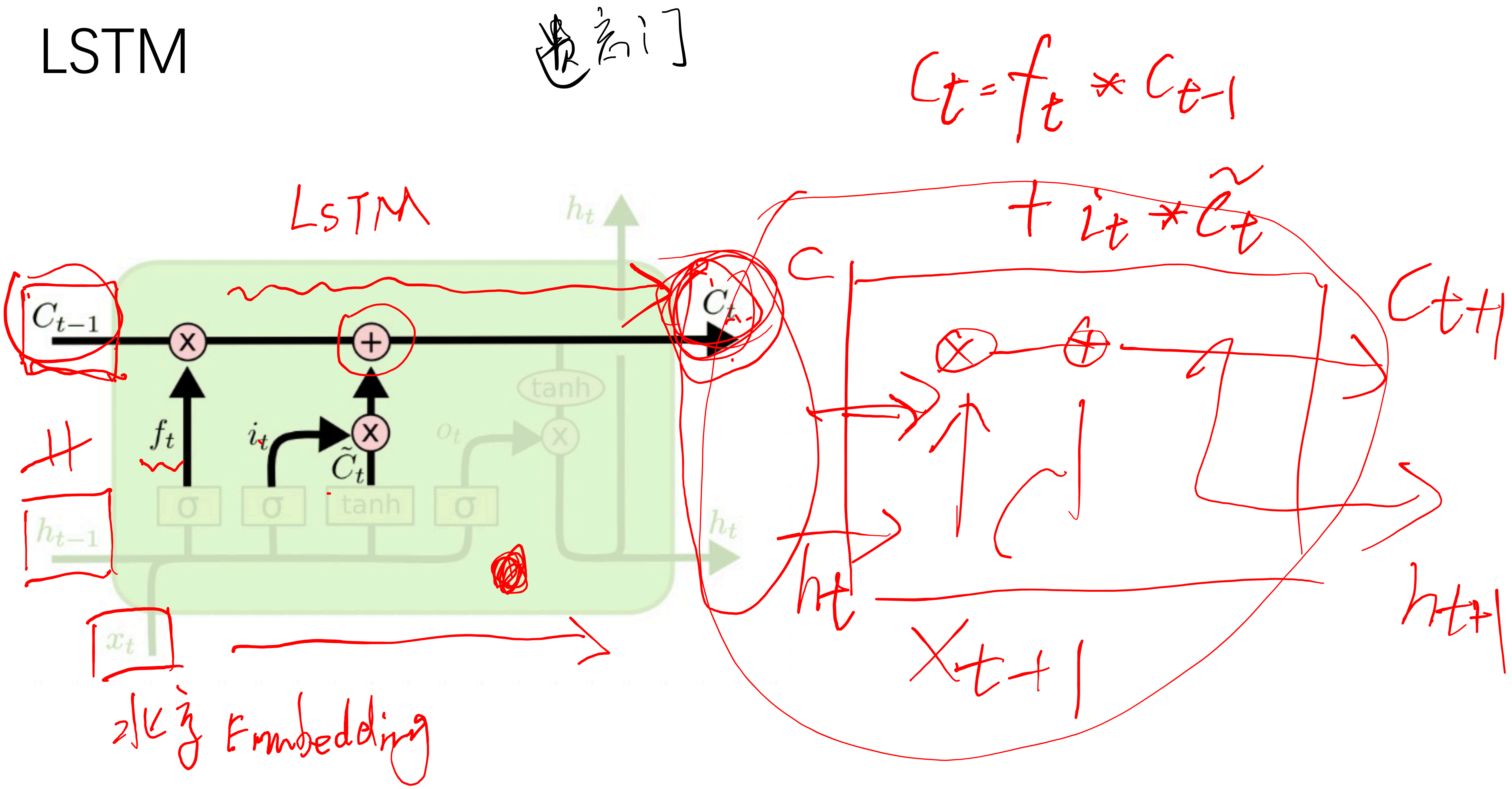
input gate



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

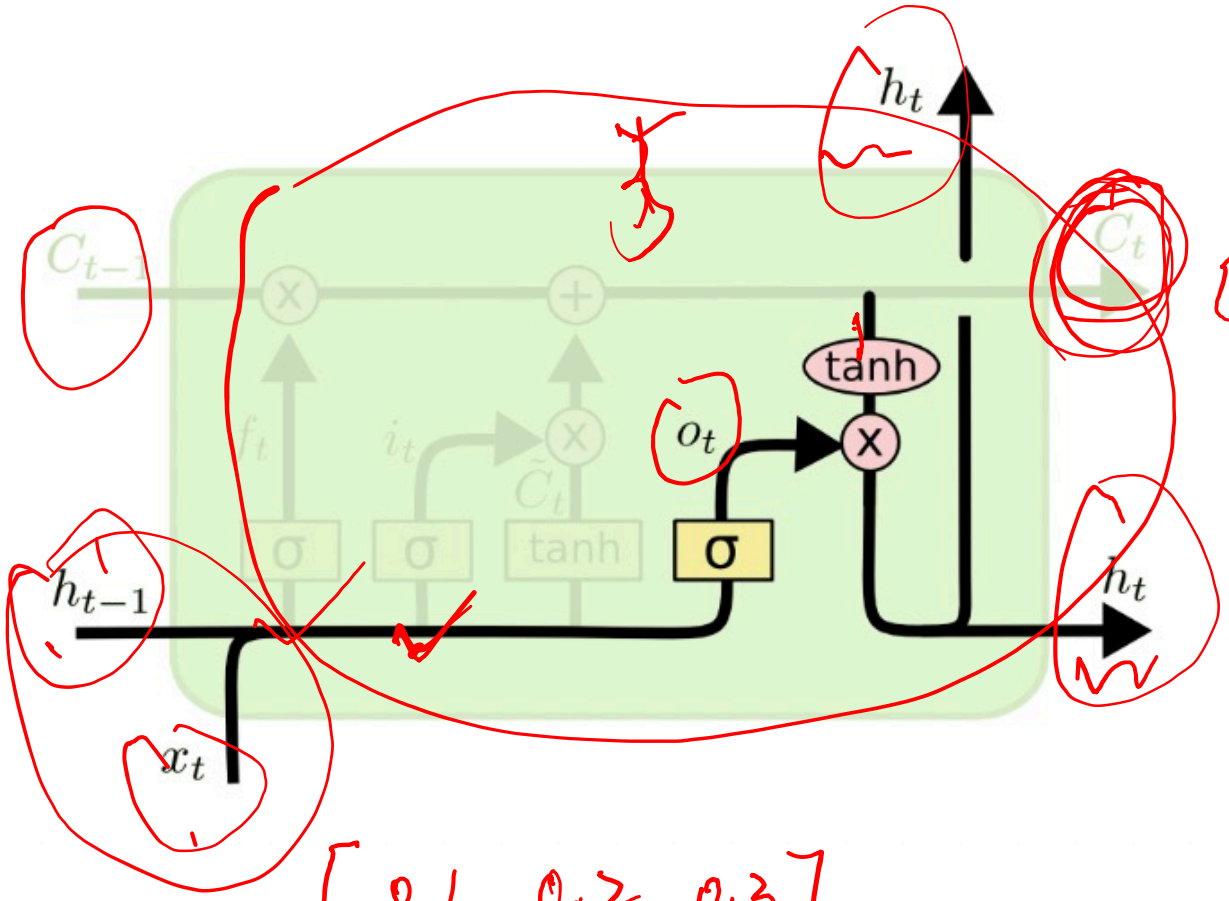
$$\tilde{C}_t = \tanh(W_C [h_{t-1}, x_t] + b_C)$$

LSTM


遗忘门
$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$
LSTM
C_{t-1}
f_t
i_t
$\tilde{C}_t$
o_t
tanh
σ
h_t
h_{t-1}
x_t
C_t
C
C_{t+1}
h_{t+1}
X_{t+1}
北京 Embedding

LSTM

Output gate



[0.1, 0.2, 0.3]

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$$

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

[1, 2, 3] ψ, t

f

i

o

σ

tanh

x

[0 0 0 0 0
0 0 0 0 0
. . . . .]