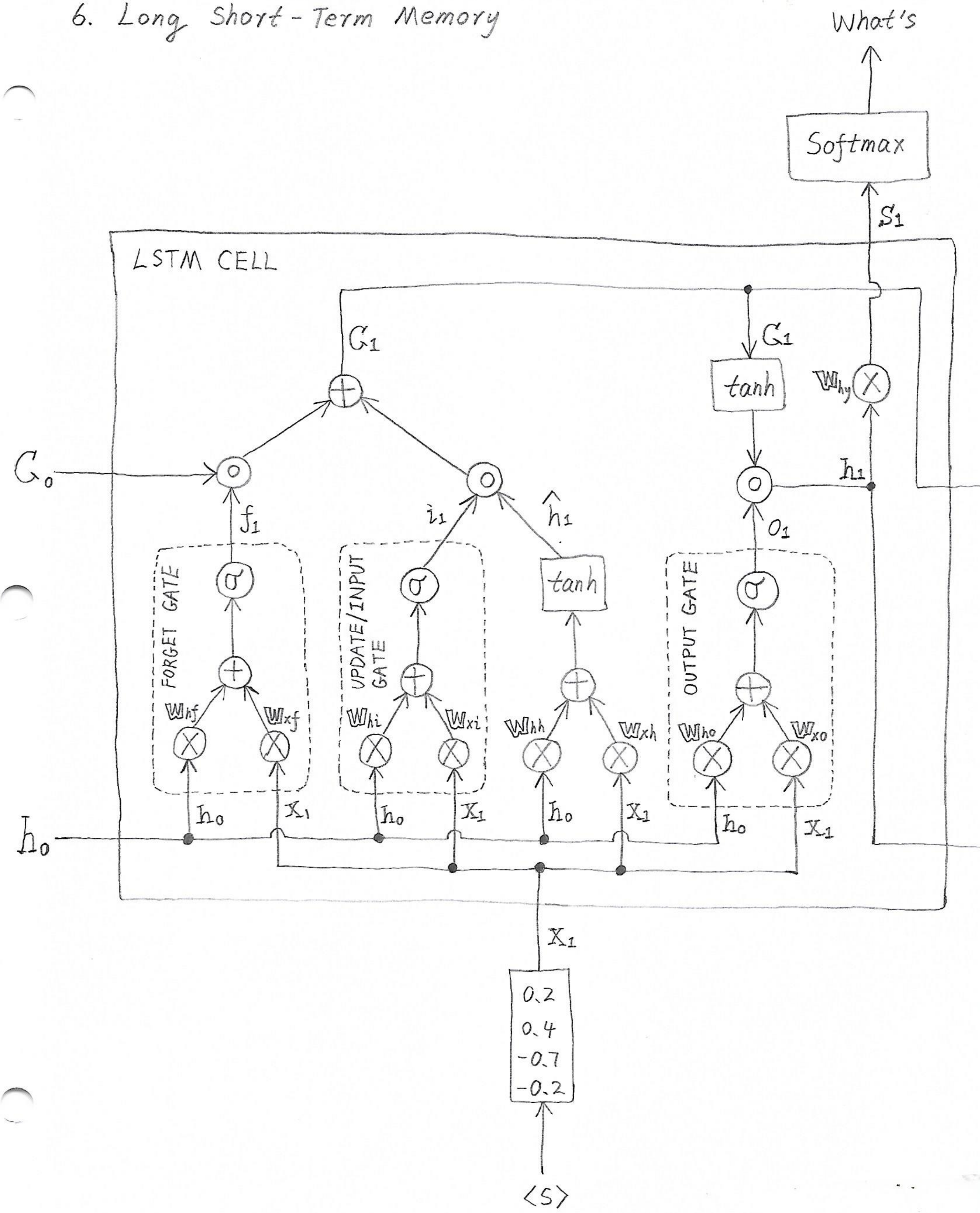
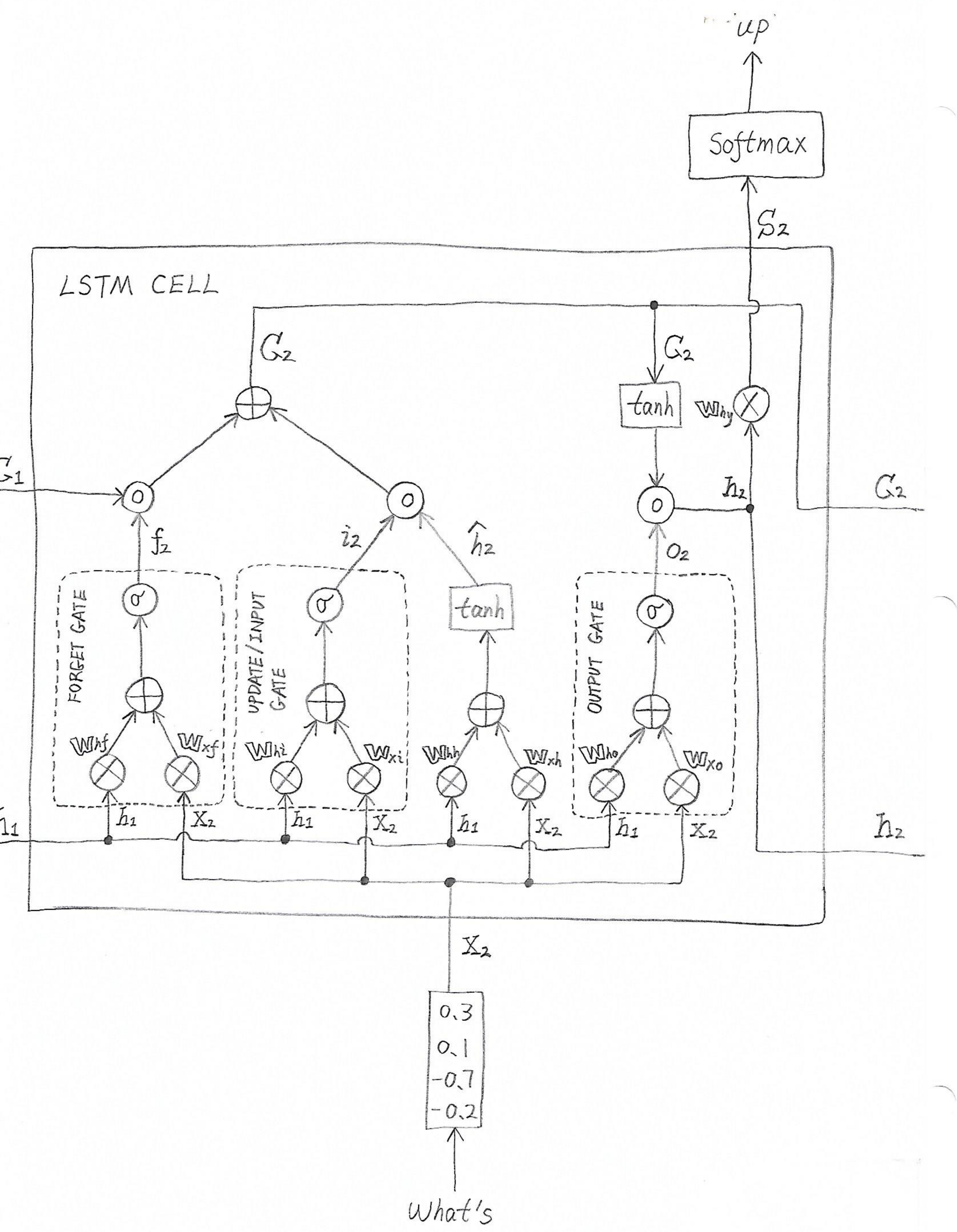
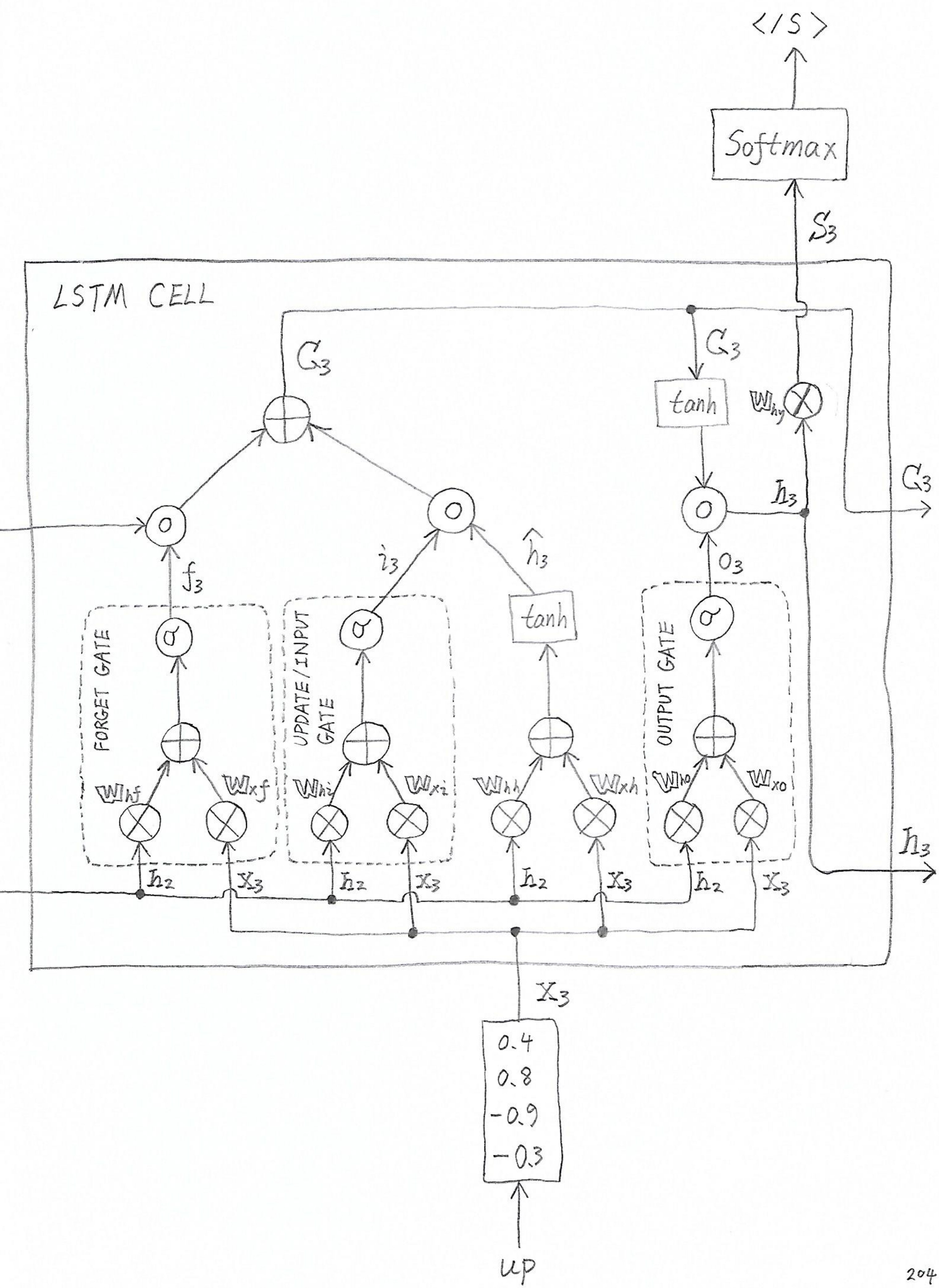
6. Long Short-Term Memory




up
Softmax
S2
LSTM CELL
C2
C2
tanh
Why
C1
C2
h2
f2
i2
ĥ2
O2
FORGET GATE
UPDATE/INPUT GATE
tanh
OUTPUT GATE
Whf
Wxf
Whi
Wxi
Whh
Wxh
Who
Wxo
h1
X2
h1
X2
h1
X2
h1
X2
h2
h1
X2
0.3
0.1
-0.7
-0.2
What's


</S>
Softmax
S_3
LSTM CELL
C_3
C_3
tanh
W_{hy}
C_3
h_3
f_3
i_3
$\hat{h}_3$
o_3
tanh
FORGET GATE
UPDATE/INPUT GATE
OUTPUT GATE
W_{hf}
W_{xf}
W_{hi}
W_{xi}
W_{hh}
W_{xh}
W_{ho}
W_{xo}
h_3
h_2
x_3
h_2
x_3
h_2
x_3
h_2
x_3
x_3
0.4
0.8
-0.9
-0.3
up

7. Forward Pass

$$\mathbb{W}_e = \{ \text{"hello"}: \mathbb{X}_1, \text{"what"}: \mathbb{X}_2, \cdots, \text{"zaltan"}: \mathbb{X}_{v\text{-}size} \}$$

$$\mathbb{X}_t = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{256} \end{bmatrix}, \quad \text{hidden\_size} = 256$$

$$i_t = \sigma(\mathbb{W}_{xi} \cdot \mathbb{X}_t + \mathbb{W}_{hi} \cdot h_{t-1}) \quad (\text{Update/Input Gate}) \quad (17)$$

$$f_t = \sigma(\mathbb{W}_{xf} \cdot \mathbb{X}_t + \mathbb{W}_{hf} \cdot h_{t-1}) \quad (\text{Forget Gate}) \quad (18)$$

$$O_t = \sigma(\mathbb{W}_{xo} \cdot \mathbb{X}_t + \mathbb{W}_{ho} \cdot h_{t-1}) \quad (\text{Output Gate}) \quad (19)$$

$$\hat{h}_t = \tanh(\mathbb{W}_{xh} \cdot \mathbb{X}_t + \mathbb{W}_{hh} \cdot h_{t-1}) \quad (20)$$

$$C_t = f_t \circ C_{t-1} + i_t \circ \hat{h}_t \quad (20.1)$$

$$h_t = O_t \circ \tanh(C_t) \quad (20.2)$$

(17) – (20) can be written as

$$\begin{bmatrix} i_t \\ f_t \\ O_t \\ \hat{h}_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{bmatrix} \left(\begin{bmatrix} \mathbb{W}_{xi} & \mathbb{W}_{hi} \\ \mathbb{W}_{xf} & \mathbb{W}_{hf} \\ \mathbb{W}_{xo} & \mathbb{W}_{ho} \\ \mathbb{W}_{xh} & \mathbb{W}_{hh} \end{bmatrix} \cdot \begin{bmatrix} \mathbb{X}_t \\ h_{t-1} \end{bmatrix} \right)$$

$$\mathcal{U}_t = g(T_{lstm} \cdot Z_t)$$

$$= g(T_x \cdot \mathbb{X}_t + T_h \cdot h_{t-1})$$

$$S_t = \mathbb{W}_{hy} \cdot h_t = \begin{bmatrix} S_t(y_1) \\ \vdots \\ S_t(y_m) \end{bmatrix} \quad (20.3)$$

$$P_t = \text{Softmax}(S_t) = \begin{bmatrix} \dfrac{e^{S_t(y_1)}}{\sum_{y'\in Y} e^{S_t(y')}} \\ \vdots \\ \dfrac{e^{S_t(y_m)}}{\sum_{y'\in Y} e^{S_t(y')}} \end{bmatrix} = \begin{bmatrix} P_t(y_1) \\ \vdots \\ P_t(y_m) \end{bmatrix} \quad (20.4)$$

$$loss_t = -\log P_t(y_t) = \log \sum_{y'\in Y} e^{S_t(y')} - S_t(y_t) \quad \text{(D.L. P19}$$

$$(20.5)$$

8. Backward Pass.

$$dS_t = \frac{\partial loss_t}{\partial S_t} = P_t - 1_{y=y_t} \quad (\text{from } ⓪)$$

$$dh_t = \mathbb{W}_{hy}^T \cdot dS_t \quad (\text{from } ⑨)$$

$$d\mathbb{W}_{hy} = dS_t \cdot h_t^T \quad (\text{from } ⑩)$$

$$dO_t = \tanh(C_t) \circ dh_t \quad (\text{by lemma } 2) \quad (21)$$

$$dC_t = \tanh'(C_t) \circ O_t \circ dh_t \quad (22)$$

Recall that

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

$$\tanh'(x) = \frac{(e^x + e^{-x})(e^x + e^{-x}) - (e^x - e^{-x})(e^x - e^{-x})}{(e^x + e^{-x})^2}$$

$$= 1 - \tanh^2(x)$$

$$df_t = C_{t-1} \circ dC_t \quad (\text{by corollary } 2) \quad (23)$$

$$dC_{t-1} = \frac{\partial l_t}{\partial C_{t-1}} = f_t \circ dC_t \quad (24)$$

$$di_t = \hat{h}_t \circ dC_t \quad (25)$$

$$d\hat{h}_t = i_t \circ dC_t \quad (26)$$

$$dU_t = \begin{bmatrix} di_t \\ df_t \\ do_t \\ d\hat{h}_t \end{bmatrix} = [\, di_t ;\ df_t ;\ do_t ;\ d\hat{h}_t \,]$$

Recall that

$$U_t = g(T_{lstm} \cdot Z_t)$$

$$= g(T_x \cdot X_t + T_h \cdot h_{t-1})$$

$$dT_x = (g'(T_{lstm} \cdot Z_t) \circ dU_t) \cdot X_t^T \quad (\text{by lemma } 1)$$

$$\begin{bmatrix} dW_{xi} \\ dW_{xf} \\ dW_{xo} \\ dW_{xh} \end{bmatrix} = \left(\begin{bmatrix} \sigma'(W_{xi} X_t + W_{hi} h_{t-1}) \\ \sigma'(W_{xf} X_t + W_{hf} h_{t-1}) \\ \sigma'(W_{xo} X_t + W_{ho} h_{t-1}) \\ \tanh'(W_{xh} X_t + W_{hh} h_{t-1}) \end{bmatrix} \circ \begin{bmatrix} di_t \\ df_t \\ do_t \\ d\hat{h}_t \end{bmatrix} \right) \cdot X^T$$

$$d\mathbb{W}_{xi} = (\sigma'(\mathbb{W}_{xi} X_t + \mathbb{W}_{hi} h_{t-1}) \circ d i_t) \cdot X^T \qquad (27)$$

$$d\mathbb{W}_{xf} = (\sigma'(\mathbb{W}_{xf} X_t + \mathbb{W}_{hf} h_{t-1}) \circ d f_t) \cdot X^T \qquad (28)$$

$$d\mathbb{W}_{xo} = (\sigma'(\mathbb{W}_{xo} X_t + \mathbb{W}_{ho} h_{t-1}) \circ d O_t) \cdot X^T \qquad (29)$$

$$d\mathbb{W}_{xh} = (\tanh'(\mathbb{W}_{xh} X_t + \mathbb{W}_{hh} h_{t-1}) \circ d \hat{h}_t) \cdot X^T \qquad (30)$$

Similarly,

$$dT_h = (g'(T_{lstm} \cdot Z_t) \circ d\mathcal{U}_t) \cdot h_{t-1}^T \quad (\text{by lemma } 1)$$

$$d\mathbb{W}_{hi} = (\sigma'(\mathbb{W}_{xi} X_t + \mathbb{W}_{hi} h_{t-1}) \circ d i_t) \cdot h_{t-1}^T \qquad (31)$$

$$d\mathbb{W}_{hf} = (\sigma'(\mathbb{W}_{xf} X_t + \mathbb{W}_{hf} h_{t-1}) \circ d f_t) \cdot h_{t-1}^T \qquad (32)$$

$$d\mathbb{W}_{ho} = (\sigma'(\mathbb{W}_{xo} X_t + \mathbb{W}_{ho} h_{t-1}) \circ d O_t) \cdot h_{t-1}^T \qquad (33)$$

$$d\mathbb{W}_{hh} = (\tanh'(\mathbb{W}_{xh} X_t + \mathbb{W}_{hh} h_{t-1}) \circ d \hat{h}_t) \cdot h_{t-1}^T \qquad (34)$$

and

$$dX_t = T_x^T \cdot (g'(T_{lstm} \cdot Z_t) \circ d\mathcal{U}_t) \quad (\text{by lemma } 1)$$

$$= \begin{bmatrix} \mathbb{W}_{xi}^T & \mathbb{W}_{xf}^T & \mathbb{W}_{xo}^T & \mathbb{W}_{xh}^T \end{bmatrix} \cdot \begin{bmatrix} \sigma'(\mathbb{W}_{xi} X_t + \mathbb{W}_{hi} h_{t-1}) \circ d i_t \\ \sigma'(\mathbb{W}_{xf} X_t + \mathbb{W}_{hf} h_{t-1}) \circ d f_t \\ \sigma'(\mathbb{W}_{xo} X_t + \mathbb{W}_{ho} h_{t-1}) \circ d O_t \\ \tanh'(\mathbb{W}_{xh} X_t + \mathbb{W}_{hh} h_{t-1}) \circ d \hat{h}_t \end{bmatrix}$$

$$= \mathbb{W}_{xi}^{T} \cdot (\sigma'(\mathbb{W}_{xi} X_t + \mathbb{W}_{hi} h_{t-1}) \circ di_t) +$$

$$\mathbb{W}_{xf}^{T} \cdot (\sigma'(\mathbb{W}_{xf} X_t + \mathbb{W}_{hf} h_{t-1}) \circ df_t) +$$

$$\mathbb{W}_{xo}^{T} \cdot (\sigma'(\mathbb{W}_{xo} X_t + \mathbb{W}_{ho} h_{t-1}) \circ dO_t) +$$

$$\mathbb{W}_{xh}^{T} \cdot (\tanh'(\mathbb{W}_{xh} X_t + \mathbb{W}_{hh} h_{t-1}) \circ d\hat{h}_t) \qquad (35)$$

Similarly,

$$dh_{t-1} = \frac{\partial l_t}{\partial h_{t-1}} = T_h^{T} \cdot (g'(T_{lstm} \cdot Z_t) \circ dU_t) \quad (\text{by lemma } 1)$$

$$= \mathbb{W}_{hi}^{T} \cdot (\sigma'(\mathbb{W}_{xi} X_t + \mathbb{W}_{hi} h_{t-1}) \circ di_t) +$$

$$\mathbb{W}_{hf}^{T} \cdot (\sigma'(\mathbb{W}_{xf} X_t + \mathbb{W}_{hf} h_{t-1}) \circ df_t) +$$

$$\mathbb{W}_{ho}^{T} \cdot (\sigma'(\mathbb{W}_{xo} X_t + \mathbb{W}_{ho} h_{t-1}) \circ dO_t) +$$

$$\mathbb{W}_{hh}^{T} \cdot (\tanh'(\mathbb{W}_{xh} X_t + \mathbb{W}_{hh} h_{t-1}) \circ d\hat{h}_t) \qquad (36)$$

Backpropagation_Through_Time_LSTM ($\{y_t\}_{t=1}^{T}, \{P_t\}_{t=1}^{T}, \{h_t\}_{t=1}^{T}, \{X_t\}_{t=1}^{T},$
$\{\mathbb{W}_{hy}, \mathbb{W}_{xi}, \mathbb{W}_{xf}, \mathbb{W}_{xo}, \mathbb{W}_{xh}, \mathbb{W}_{hi}, \mathbb{W}_{hf}, \mathbb{W}_{ho}, \mathbb{W}_{hh}\}$) :

for t from T to 1 :

$$\frac{\partial l_t}{\partial S_t} = P_t - 1_{y=y_t} \quad ⓪$$

$$\frac{\partial l_t}{\partial \mathbb{W}_{hy}} = \frac{\partial l_{t+1}}{\partial \mathbb{W}_{hy}} + \frac{\partial l_t}{\partial S_t} \cdot h_t^T \quad ⑩$$

$$\frac{\partial l_t}{\partial h_t} = \frac{\partial l_{t+1}}{\partial h_t} + \mathbb{W}_{hy}^T \cdot \frac{\partial l_t}{\partial S_t} \quad ⑨$$

$$\frac{\partial l_t}{\partial o_t} = \tanh(C_t) \circ \frac{\partial l_t}{\partial h_t} \quad ㉑$$

$$\frac{\partial l_t}{\partial C_t} = \frac{\partial l_{t+1}}{\partial C_t} + \tanh'(C_t) \circ o_t \circ \frac{\partial l_t}{\partial h_t} \quad ㉒$$

$$\frac{\partial l_t}{\partial f_t} = C_{t-1} \circ \frac{\partial l_t}{\partial C_t} \quad ㉓$$

$$\frac{\partial l_t}{\partial i_t} = \hat{h}_t \circ \frac{\partial l_t}{\partial C_t} \quad ㉕$$

$$\partial \hat{h}_t \qquad \partial C_t$$

$$\frac{\partial l_t}{\partial C_{t-1}} = f_t \circ \frac{\partial l_t}{\partial C_t} \quad (24)$$

$$\frac{\partial l_t}{\partial W_{xi}} = \frac{\partial l_{t+1}}{\partial W_{xi}} + (\sigma'(W_{xi} X_t + W_{hi} h_{t-1}) \circ \frac{\partial l_t}{\partial i_t}) \cdot X_t^T \quad (27)$$

$$\frac{\partial l_t}{\partial W_{xf}} = \frac{\partial l_{t+1}}{\partial W_{xf}} + (\sigma'(W_{xf} X_t + W_{hf} h_{t-1}) \circ \frac{\partial l_t}{\partial f_t}) \cdot X_t^T \quad (28)$$

$$\frac{\partial l_t}{\partial W_{xo}} = \frac{\partial l_{t+1}}{\partial W_{xo}} + (\sigma'(W_{xo} X_t + W_{ho} h_{t-1}) \circ \frac{\partial l_t}{\partial o_t}) \cdot X_t^T \quad (29)$$

$$\frac{\partial l_t}{\partial W_{xh}} = \frac{\partial l_{t+1}}{\partial W_{xh}} + (\tanh'(W_{xh} X_t + W_{hh} h_{t-1}) \circ \frac{\partial l_t}{\partial \hat{h}_t}) \cdot X_t^T \quad (30)$$

$$\frac{\partial l_t}{\partial W_{hi}} = \frac{\partial l_{t+1}}{\partial W_{hi}} + (\sigma'(W_{xi} X_t + W_{hi} h_{t-1}) \circ \frac{\partial l_t}{\partial i_t}) \cdot h_{t-1}^T \quad (31)$$

$$\frac{\partial l_t}{\partial W_{hf}} = \frac{\partial l_{t+1}}{\partial W_{hf}} + (\sigma'(W_{xf} X_t + W_{hf} h_{t-1}) \circ \frac{\partial l_t}{\partial f_t}) \cdot h_{t-1}^T \quad (32)$$

$$\frac{\partial l_t}{\partial W_{ho}} = \frac{\partial l_{t+1}}{\partial W_{ho}} + (\sigma'(W_{xo} X_t + W_{ho} h_{t-1}) \circ \frac{\partial l_t}{\partial o_t}) \cdot h_{t-1}^T \quad (33)$$

GPU

$$\frac{\partial l_t}{\partial \mathbb{W}_{hh}} = \frac{\partial l_{t+1}}{\partial \mathbb{W}_{hh}} + \left(\tanh'(\mathbb{W}_{xh} X_t + \mathbb{W}_{hh} h_{t-1}) \circ \frac{\partial l_t}{\partial \hat{h}_t}\right) \cdot h_{t-1}^T \quad (34)$$

$$\begin{aligned} \frac{\partial l_t}{\partial X_t} = \; & \mathbb{W}_{xi}^T \cdot \left(\sigma'(\mathbb{W}_{xi} X_t + \mathbb{W}_{hi} h_{t-1}) \circ \frac{\partial l_t}{\partial i_t}\right) + \\ & \mathbb{W}_{xf}^T \cdot \left(\sigma'(\mathbb{W}_{xf} X_t + \mathbb{W}_{hf} h_{t-1}) \circ \frac{\partial l_t}{\partial f_t}\right) + \\ & \mathbb{W}_{xo}^T \cdot \left(\sigma'(\mathbb{W}_{xo} X_t + \mathbb{W}_{ho} h_{t-1}) \circ \frac{\partial l_t}{\partial o_t}\right) + \\ & \mathbb{W}_{xh}^T \cdot \left(\tanh'(\mathbb{W}_{xh} X_t + \mathbb{W}_{hh} h_{t-1}) \circ \frac{\partial l_t}{\partial \hat{h}_t}\right) \quad (35) \end{aligned}$$

$$\begin{aligned} \boxed{\frac{\partial l_t}{\partial h_{t-1}}} = \; & \mathbb{W}_{hi}^T \cdot \left(\sigma'(\mathbb{W}_{xi} X_t + \mathbb{W}_{hi} h_{t-1}) \circ \frac{\partial l_t}{\partial i_t}\right) + \\ & \mathbb{W}_{hf}^T \cdot \left(\sigma'(\mathbb{W}_{xf} X_t + \mathbb{W}_{hf} h_{t-1}) \circ \frac{\partial l_t}{\partial f_t}\right) + \\ & \mathbb{W}_{ho}^T \cdot \left(\sigma'(\mathbb{W}_{xo} X_t + \mathbb{W}_{ho} h_{t-1}) \circ \frac{\partial l_t}{\partial o_t}\right) + \\ & \mathbb{W}_{hh}^T \cdot \left(\tanh'(\mathbb{W}_{xh} X_t + \mathbb{W}_{hh} h_{t-1}) \circ \frac{\partial l_t}{\partial \hat{h}_t}\right) \quad (36) \end{aligned}$$

$$\text{return} \left\{ \frac{\partial l_{total}}{\partial \mathbb{W}_{hy}}, \frac{\partial l_{total}}{\partial \mathbb{W}_{xi}}, \frac{\partial l_{total}}{\partial \mathbb{W}_{xf}}, \frac{\partial l_{total}}{\partial \mathbb{W}_{xo}}, \frac{\partial l_{total}}{\partial \mathbb{W}_{xh}}, \frac{\partial l_{total}}{\partial \mathbb{W}_{hi}}, \frac{\partial l_{total}}{\partial \mathbb{W}_{hf}}, \frac{\partial l_{total}}{\partial \mathbb{W}_{ho}}, \frac{\partial l_{total}}{\partial \mathbb{W}_{hh}} \right\}, \left\{ \frac{\partial l_t}{\partial X_t} \right\}_{t=}^{T}$$