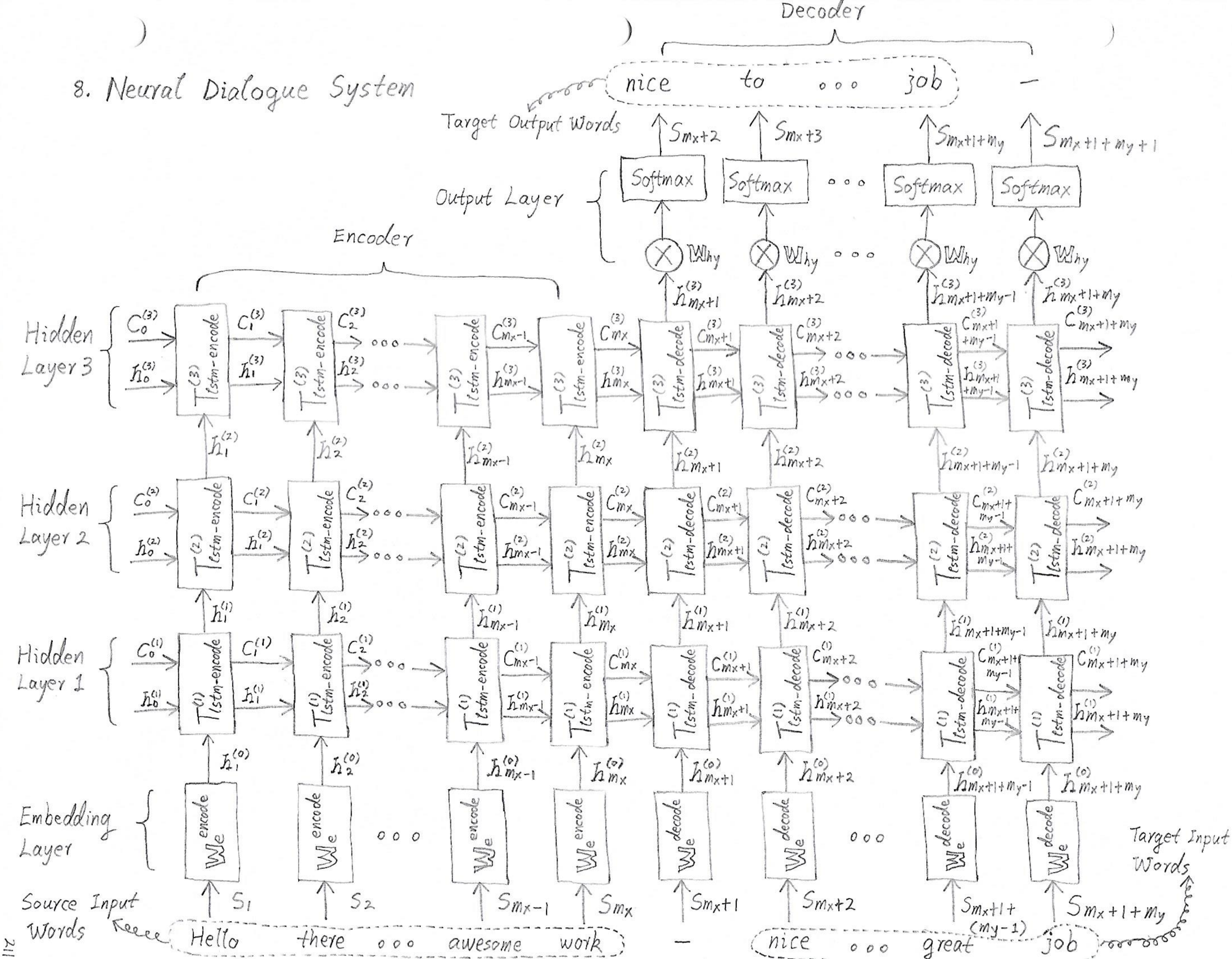
8. Neural Dialogue System



9.

$\text{Neural\_Dialogue\_System\_Forward\_Pass}(\{x_i\}_{i=1}^{m_x}, \{y_i\}_{i=1}^{m_y}, \mathbb{W}_e^{encode}, \mathbb{W}_e^{decode}, \{T_{lstm\text{-}encode}^{(l)}\}_{l=1}^{L}, \{T_{lstm\text{-}decode}^{(l)}\}_{l=1}^{L}, \mathbb{W}_{hy})$:

$S = [x_1, x_2, \cdots, x_{m_x}, \_, y_1, y_2, \cdots, y_{m_y}, \_] = \{S_t\}_{t=1}^{m_x+1+m_y+1}$

for l from 1 to L:

$\quad h_0^{(l)} = 0$ # Reset hidden state for the new training sequence

$\quad C_0^{(l)} = 0$ # so that the model doesn't depend on the previous

$\quad$ # training sequence. But who said it's not a good idea

$\quad T_{lstm}^{(l)} = T_{lstm\text{-}encode}^{(l)}$ # for an intelligent chat bot?

$\mathbb{W}_e = \mathbb{W}_e^{encode}$

for t from 1 to $m_x + 1 + m_y$:

$\quad$ if $t == m_x + 1$:

$\quad\quad$ for l from 1 to L:

$\quad\quad\quad T_{lstm}^{(l)} = T_{lstm\text{-}decode}^{(l)}$

$\quad\quad \mathbb{W}_e = \mathbb{W}_e^{decode}$

$\quad h_t^{(0)} = \text{Embedding\_Look\_Up}(S_t, \mathbb{W}_e)$

$\quad$ for l from 1 to L:

$\quad\quad h_t^{(l)}, C_t^{(l)} = \text{LSTM\_Forward\_Pass}(h_{t-1}^{(l)}, C_{t-1}^{(l)}, h_t^{(l-1)}, T_{lstm}^{(l)})$

$\quad\quad$ # By D.L. P205, (17) - (20.2), replace x_t by $h_t^{(l-1)}$

$\quad$ if $t >= m_x + 1$:

$\quad\quad l_t, P_t = \text{Predict}(S_{t+1}, h_t^{(L)}, \mathbb{W}_{hy})$

$\quad\quad$ # By D.L. P206, (20.3) - (20.5), replace y_t by S_{t+1}

return $\{l_t, P_t\}_{t=m_x+1}^{m_x+1+m_y+1}$, $\{h_t^{(l)}, C_t^{(l)}\}_{t=1,2,\cdots,m_x+1+m_y,\ l=0,1,\cdots,L}$

10. Neural_Dialogue_System_Backward_Pass $\left(\{x_i\}_{i=1}^{m_x}, \{y_i\}_{i=1}^{m_y}, \{T_{lstm\text{-}encode}^{(l)}\}_{l=1}^{L}, \{T_{lstm\text{-}decode}^{(l)}\}_{l=1}^{L}, \{h_t^{(l)}, C_t^{(l)}\}_{t=1,2,\cdots,m_x+1+m_y,\ l=0,1,\cdots,L}, \mathbb{W}_e^{encode}, \mathbb{W}_e^{decode}, \mathbb{W}_{hy} \right)$:

$S = [x_1, x_2, \cdots, x_{m_x}, -, y_1, y_2, \cdots, y_{m_y}, -] = \{S_t\}_{t=1}^{m_x+1+m_y+1}$

for t from m_x+1+m_y to 1:

for l from L to 1:

$$\frac{\partial L_t}{\partial h_t^{(l)}}, \frac{\partial L_t}{\partial C_t^{(l)}}, \frac{\partial L_t}{\partial T_{lstm}^{(l)}} = 0$$

$$\frac{\partial L_t}{\partial \mathbb{W}_{hy}} = 0$$

$$\frac{\partial L_t}{\partial \mathbb{W}_e} = 0$$

for t from m_x+1+m_y to 1:

if $t == m_x$:

for l from 1 to L:

$$\frac{\partial L}{\partial T_{lstm\text{-}decode}^{(l)}} = \frac{\partial L_{t+1}}{\partial T_{lstm}^{(l)}}$$

$$\frac{\partial L_{t+1}}{\partial T_{lstm}^{(l)}} = 0$$

$$\frac{\partial L}{\partial \mathbb{W}_e^{decode}} = \frac{\partial L_{t+1}}{\partial \mathbb{W}_e}$$

$$\frac{\partial L_{t+1}}{\partial \mathbb{W}_e} = 0$$

if $t >= m_x + 1$:

$$\frac{\partial L_t}{\partial h_t^{(L)}}, \frac{\partial L_t}{\partial \mathbb{W}_{hy}} = \text{Output\_Grad}\left(S_{t+1}, P_t, h_t^{(L)}, \frac{\partial L_{t+1}}{\partial h_t^{(L)}}, \frac{\partial L_{t+1}}{\partial \mathbb{W}_{hy}}, \mathbb{W}_{hy} \right)$$

\# Refactored from Backpropagation_Through_Time_LSTM() –

\# ⓪, ⑨ and ⑩, and replace y_t by S_{t+1}

for l from L to 1 :

if $t >= m_x + 1$:

$$T_{lstm}^{(l)} = T_{lstm\text{-}decode}^{(l)}$$

else :

$$T_{lstm}^{(l)} = T_{lstm\text{-}encode}^{(l)}$$

$$\frac{\partial L_t}{\partial h_{t-1}^{(l)}}, \frac{\partial L_t}{\partial C_{t-1}^{(l)}}, \frac{\partial L_t}{\partial X}, \frac{\partial L_t}{\partial T_{lstm}^{(l)}} = \text{LSTM\_Grad}$$

$$\left(C_{t-1}^{(l)}, h_{t-1}^{(l)}, h_t^{(l-1)}, \frac{\partial L_{t+1}}{\partial T_{lstm}^{(l)}}, T_{lstm}^{(l)}, \frac{\partial L_{t+1}}{\partial C_t^{(l)}}, C_t^{(l)}, \frac{\partial L_t}{\partial h_t^{(l)}} \right)$$

\# Refactored from Backpropagation_Through_Time_LSTM() –

\# ㉑ – ㊱, and replace x_t by $h_t^{(l-1)}$

$$* \quad \frac{\partial L_t}{\partial h_t^{(l-1)}} = \frac{\partial L_{t+1}}{\partial h_t^{(l-1)}} + \frac{\partial L_t}{\partial X} \quad *$$

if $t >= m_x + 1$:

$$\mathbb{W}_e = \mathbb{W}_e^{decode}$$

else :

$$\mathbb{W}_e = \mathbb{W}_e^{encode}$$

$$\frac{\partial L_t}{\partial \mathbb{W}_e} = \text{Embedding\_Grad}\left(S_t,\ \frac{\partial L_t}{\partial h_t^{(0)}},\ \mathbb{W}_e,\ \frac{\partial L_{t+1}}{\partial \mathbb{W}_e} \right)$$

for l from 1 to L :

$$\frac{\partial L}{\partial T^{(l)}_{lstm\text{-}encode}} = \frac{\partial L_{t=1}}{\partial T^{(l)}_{lstm}}$$

$$\frac{\partial L}{\partial \mathbb{W}_e^{encode}} = \frac{\partial L_{t=1}}{\partial \mathbb{W}_e}$$

return $\dfrac{\partial L_{t=m_x+1}}{\partial \mathbb{W}_{hy}},\ \dfrac{\partial L}{\partial \mathbb{W}_e^{encode}},\ \dfrac{\partial L}{\partial \mathbb{W}_e^{decode}},\ \left\{ \dfrac{\partial L}{\partial T^{(l)}_{lstm-encode}} \right\}_{l=1}^{L},\ \left\{ \dfrac{\partial L}{\partial T^{(l)}_{lstm-decode}} \right\}_{l=1}^{L}$

11. Stochastic_Gradient_Descent ($\mathbb{X} = \{ \mathbb{X}^{(1)}, \mathbb{X}^{(2)}, \cdots, \mathbb{X}^{(m)} \}$,

$\mathbb{Y} = \{ \mathbb{Y}^{(1)}, \mathbb{Y}^{(2)}, \cdots, \mathbb{Y}^{(m)} \}$, ε, δ):

$\theta' =$ Random ()

while ($\left\| \frac{1}{m} \cdot \sum_{i=1}^{m} L (\mathbb{X}^{(i)}, \mathbb{Y}^{(i)}, \theta) \Big|_{\theta=\theta'} \right\|_2 > \delta$):

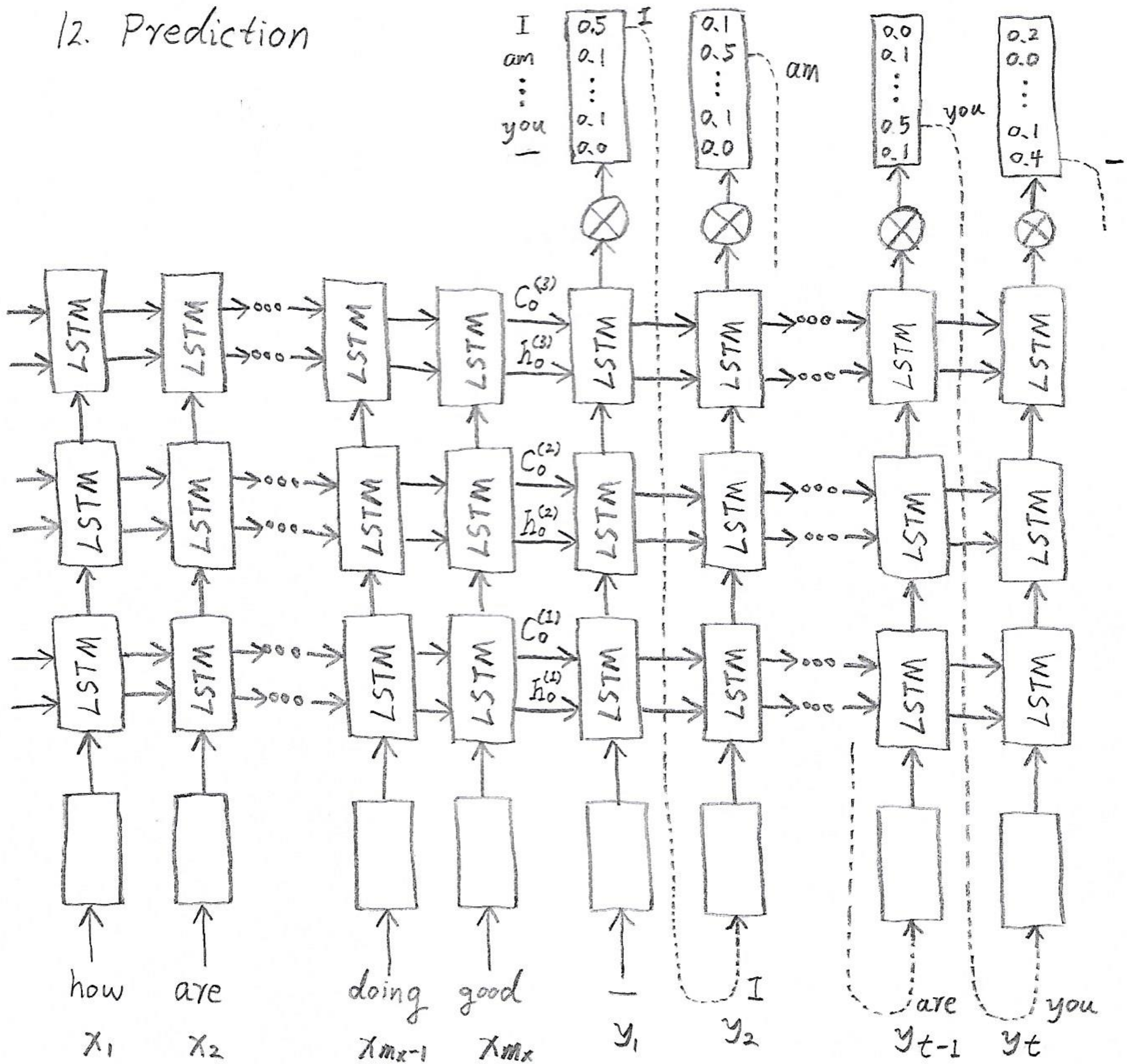
($\mathbb{X}' = \{ \mathbb{X}^{(i)} \}_{i=1}^{m'}$, $\mathbb{Y}' = \{ \mathbb{Y}^{(i)} \}_{i=1}^{m'}$)

= Sample_Mini_Batch ($\mathbb{X}$, $\mathbb{Y}$)

$$\theta' = \theta' - \varepsilon \cdot \left(\frac{1}{m'} \cdot \sum_{i=1}^{m'} \nabla_\theta L (\mathbb{X}^{(i)}, \mathbb{Y}^{(i)}, \theta) \Big|_{\theta=\theta'} \right)$$

return θ'

12. Prediction



Neural_Dialogue_System_Greedy_Predict ($\{x_i\}_{i=1}^{m_x}$, α,

$\{T_{lstm-encode}^{(l)}\}_{l=1}^{L}$, $\{T_{lstm-decode}^{(l)}\}_{l=1}^{L}$, $\mathbb{W}_e^{encode}$, $\mathbb{W}_e^{decode}$, $\mathbb{W}_{hy}$):

$\{h_0^{(l)}\}_{l=1}^{L}, \{C_0^{(l)}\}_{l=1}^{L} = Encode(\{x_i\}_{i=1}^{m_x}, \{T_{lstm-encode}^{(l)}\}_{l=1}^{L}, \mathbb{W}_e^{encode}$

$t = 1$

$y_1 =$ "_"

while $t <= \alpha \cdot m_x$:

$\{h_t^{(l)}, C_t^{(l)}\}_{l=1}^{L} = LSTM(\{h_{t-1}^{(l)}, C_{t-1}^{(l)}\}_{l=1}^{L}, \{T_{lstm-decode}^{(l)}\}_{l=1}^{L}, \mathbb{W}_e^{decode}, y_t)$

$\mathbb{P}_t = Softmax(h_t^{(L)}, \mathbb{W}_{hy}) = [p_t(y_1), \cdots, p_t(y_m)]^T$

$y_{t+1} = \arg\max_{y_i} p_t(y_i)$

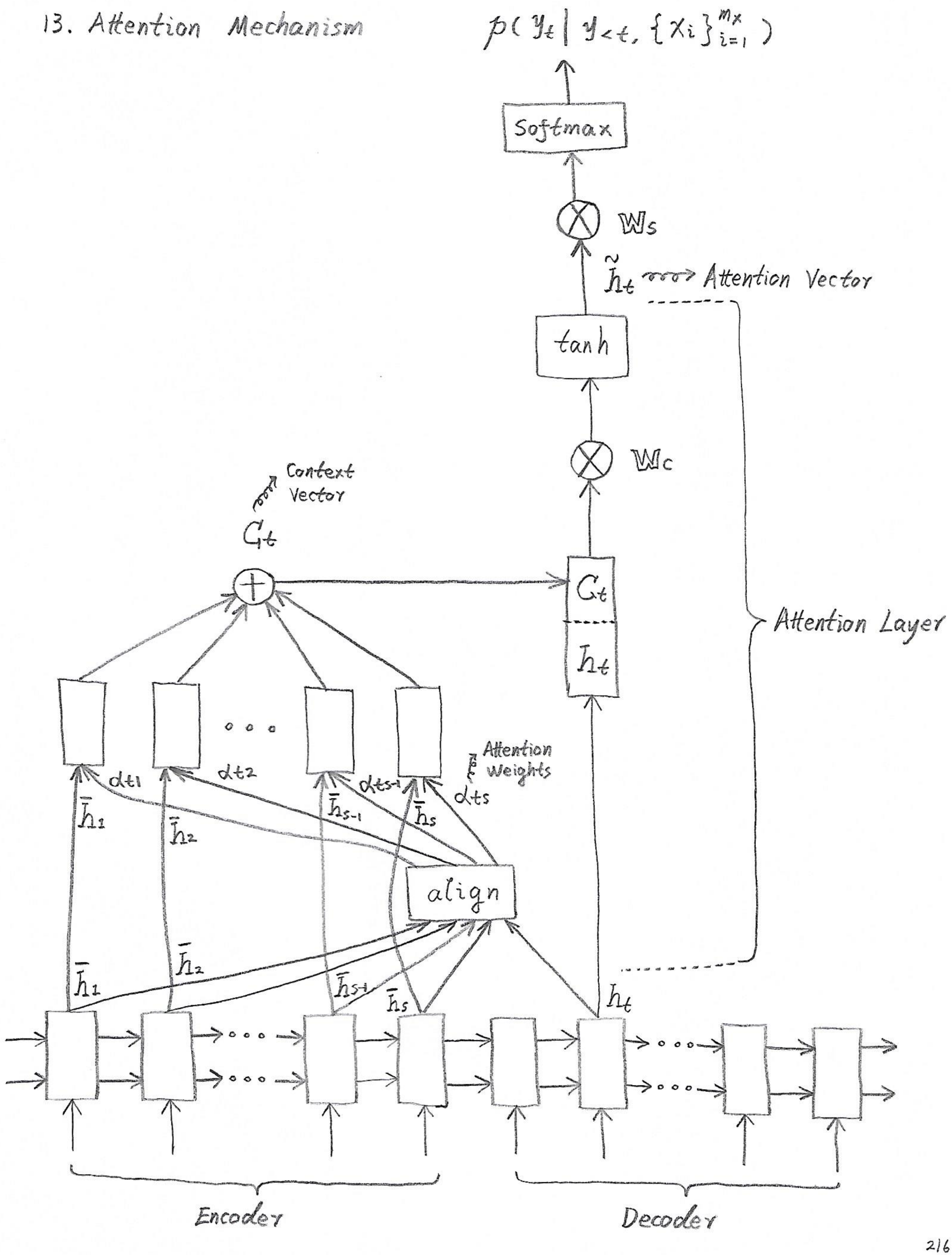
if y_{t+1} == "_" :

break

$t = t + 1$

return $\{y_2, y_3, \cdots, y_t\}$

13. Attention Mechanism



$$\alpha_{ts} = \text{align}(h_t, \bar{h}_s)$$

$$= \frac{e^{\text{score}(h_t, \bar{h}_s)}}{\sum_{s'} e^{\text{score}(h_t, \bar{h}_{s'})}}$$

$$\text{score}(h_t, \bar{h}_s) = \begin{cases} h_t^T \cdot \bar{h}_s \\ h_t^T \cdot \mathbb{W}_a \cdot \bar{h}_s \\ v_a^T \cdot \tanh(\mathbb{W}_a \cdot [h_t ; \bar{h}_s]) \end{cases}$$

$$c_t = \sum_s \alpha_{ts} \cdot \bar{h}_s$$

$$\tilde{h}_t = \tanh(\mathbb{W}_c \cdot [c_t ; h_t])$$

$$p(y_t \mid y_{<t}, \{x_i\}_{i=1}^{m_x}) = \text{Softmax}(\mathbb{W}_s \cdot \tilde{h}_t)$$