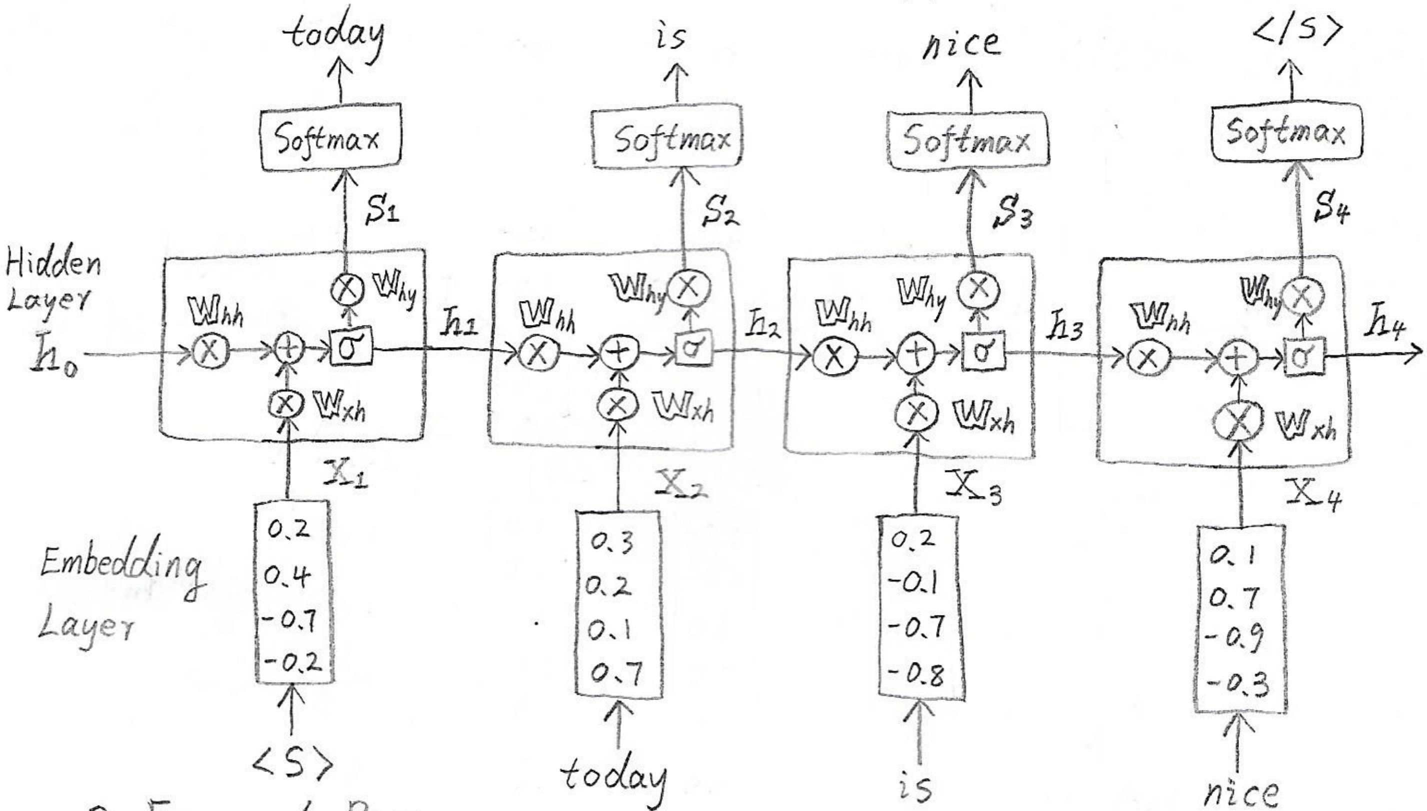
IX RECURRENT NEURAL NETWORK

1. Vanilla Recurrent Neural Network



2. Forward Pass

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

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

$$\underset{256\times 1}{h_t} = \sigma(\underset{256\times 256}{\mathbb{W}_{xh}} \cdot \underset{256\times 1}{\mathbb{X}_t} + \underset{256\times 256}{\mathbb{W}_{hh}} \cdot \underset{256\times 1}{h_{t-1}})$$

$$= \frac{1}{1 + e^{-(\mathbb{W}_{xh}\cdot \mathbb{X}_t + \mathbb{W}_{hh}\cdot h_{t-1})}}$$

$$\underset{|\mathbb{Y}|\times 1}{S_t} = \underset{|\mathbb{Y}|\times 256}{\mathbb{W}_{hy}} \cdot \underset{256\times 1}{h_t} = \begin{bmatrix} S_t(y_1) \\ S_t(y_2) \\ \vdots \\ S_t(y_m) \end{bmatrix}, \quad \mathbb{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

$$P_t = \text{softmax}(S_t)$$

$$= \text{Softmax}\left(\begin{bmatrix} S_t(y_1) \\ S_t(y_2) \\ \vdots \\ S_t(y_m) \end{bmatrix}\right)$$

$$= \begin{bmatrix} \dfrac{e^{S_t(y_1)}}{\Sigma_{y' \in Y} e^{S_t(y')}} \\ \vdots \\ \dfrac{e^{S_t(y_m)}}{\Sigma_{y' \in Y} e^{S_t(y')}} \end{bmatrix} = \begin{bmatrix} p_t(y_1) \\ \vdots \\ p_t(y_m) \end{bmatrix} = \begin{bmatrix} p_t(y_1) \\ \vdots \\ p_t(y_t) \\ \vdots \\ p_t(y_m) \end{bmatrix}$$

True output at index i — Estimated output at index i.

$$loss_t = -\sum_{i=1}^{m} y_i' \cdot \log(p_t(y_i)) = -y_t' \cdot \log(p_t(y_t))$$ (If $y_i' = y_t'$, $y_i' = 1$; if $y_i' \neq y_t'$, $y_i' = 0$. Cross Entropy Loss)

$$= -\log\left(\frac{e^{S_t(y_t)}}{\Sigma_{y' \in Y} e^{S_t(y')}}\right)$$

$$= \log \Sigma_{y' \in Y} e^{S_t(y')} - \log e^{S_t(y_t)}$$

$$= \log \sum_{y' \in Y} e^{S_t(y')} - S_t(y_t)$$

3. Backward Pass

$$dS_t(y) = \frac{\partial loss_t}{\partial S_t(y)} = \frac{\partial}{\partial S_t(y)}\left(\log \sum_{y' \in Y} e^{S_t(y')} - S_t(y_t)\right)$$

$$= \frac{\partial}{\partial S_t(y)}\left(\log \sum_{y' \in Y} e^{S_t(y')}\right) - \frac{\partial}{\partial S_t(y)}(S_t(y_t))$$

$$= \frac{1}{\sum_{y' \in Y} e^{S_t(y')}} \cdot \frac{\partial}{\partial S_t(y)}\left(\sum_{y' \in Y} e^{S_t(y')}\right) - \frac{\partial}{\partial S_t(y)}(S_t(y_t))$$

(chain rule)

when $y = y_t$

$$dS_t(y_t) = \frac{e^{S_t(y_t)}}{\sum_{y' \in Y} e^{S_t(y')}} - 1$$

$$= p_t(y_t) - 1$$

when $y \neq y_t$

$$dS_t(y) = \frac{e^{S_t(y)}}{\sum_{y' \in Y} e^{S_t(y')}} = p_t(y)$$

$$dS_t = \frac{\partial\, loss_t}{\partial S_t} = \begin{bmatrix} dS_t(y_1) \\ \vdots \\ dS_t(y_t) \\ \vdots \\ dS_t(y_m) \end{bmatrix} = \begin{bmatrix} p_t(y_1) \\ \vdots \\ p_t(y_t) - 1 \\ \vdots \\ p_t(y_m) \end{bmatrix}$$

$$= \begin{bmatrix} p_t(y_1) \\ \vdots \\ p_t(y_t) \\ \vdots \\ p_t(y_m) \end{bmatrix} - \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} = P_t - 1_{y = y_t} \quad 0$$

4. Mathematical Helpers

1). Definition 1

If $\mathcal{U}$ and $\mathcal{V}$ are vectors, then

$$diag(\mathcal{U}) \cdot \mathcal{V}$$

$$= diag\left(\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \right) \cdot \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = \begin{bmatrix} u_1 & 0 & \cdots & 0 \\ 0 & u_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & u_n \end{bmatrix} \cdot \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = \begin{bmatrix} u_1 v_1 \\ u_2 v_2 \\ \vdots \\ u_n v_n \end{bmatrix} = \mathcal{U} \circ \mathcal{V}$$

2). Lemma 1

$\mathcal{V}$ and h are vectors. $\mathbb{W}$ is a matrix. $d\mathcal{V} = \frac{\partial l}{\partial \mathcal{V}}$ is known.

$\mathcal{V} = f(\mathbb{W} \cdot h)$, where f is an element-wise function. Then

$$dh = \frac{\partial l}{\partial h} = \mathbb{W}^T \cdot (f'(\mathbb{W} \cdot h) \circ d\mathcal{V}), \quad ① \text{ and}$$

$$d\mathbb{W} = \frac{\partial l}{\partial \mathbb{W}} = (f'(\mathbb{W} \cdot h) \circ d\mathcal{V}) \cdot h^T \quad ②$$

Proof:

$$l = g(f(\mathbb{W} \cdot h)), \text{ where } \mathcal{V} = f(\mathbb{W} \cdot h), \ \mathbb{Z} = \mathbb{W} \cdot h$$

$$\frac{\partial l}{\partial h} = \frac{\partial f}{\partial h} \cdot \frac{\partial l}{\partial f} \quad (\text{Chain rule, D.L., P104, Denominator-Layout})$$

$$= \left(\frac{\partial \mathbb{Z}}{\partial h} \cdot \frac{\partial f}{\partial \mathbb{Z}} \right) \cdot \frac{\partial l}{\partial f} \quad (\text{Chain rule, again})$$

$$= \frac{\partial \mathbb{W} h}{\partial h} \cdot \frac{\partial f(\mathbb{Z})}{\partial \mathbb{Z}} \cdot \frac{\partial l}{\partial \mathcal{V}}$$

$$= \frac{\partial \mathbb{W} h}{\partial h} \cdot diag(f'(\mathbb{Z})) \cdot d\mathcal{V}$$

$$= \frac{\partial \mathbb{W} \cdot h}{\partial h} \cdot (f'(\mathbb{W} h) \circ dv) \quad (\text{Definition } 1)$$

$$\mathbb{W} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m1} & w_{m2} & \cdots & w_{mn} \end{bmatrix}, \quad h = \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_n \end{bmatrix}, \quad \text{then}$$

$$\mathbb{W} h = \begin{bmatrix} w_{11} h_1 + w_{12} h_2 + \cdots + w_{1n} h_n \\ w_{21} h_1 + w_{22} h_2 + \cdots + w_{2n} h_n \\ \vdots \\ w_{m1} h_1 + w_{m2} h_2 + \cdots + w_{mn} h_n \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_m \end{bmatrix}$$

$$\frac{\partial \mathbb{W} h}{\partial h} = \begin{bmatrix} \frac{\partial z_1}{\partial h_1} & \frac{\partial z_2}{\partial h_1} & \cdots & \frac{\partial z_m}{\partial h_1} \\ \frac{\partial z_1}{\partial h_2} & \frac{\partial z_2}{\partial h_2} & \cdots & \frac{\partial z_m}{\partial h_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial z_1}{\partial h_n} & \frac{\partial z_2}{\partial h_n} & \cdots & \frac{\partial z_m}{\partial h_n} \end{bmatrix} = \begin{bmatrix} w_{11} & w_{21} & \cdots & w_{m1} \\ w_{12} & w_{22} & \cdots & w_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ w_{1n} & w_{2n} & \cdots & w_{mn} \end{bmatrix} = \mathbb{W}^T$$

(Denominator - Layout)

$$dh = \frac{\partial l}{\partial h} = \mathbb{W}^T \cdot (f'(\mathbb{W} h) \circ dv)$$

$$\text{Let} \quad w_1 = \begin{bmatrix} w_{11} \\ w_{12} \\ \vdots \\ w_{1n} \end{bmatrix}, \quad w_2 = \begin{bmatrix} w_{21} \\ w_{22} \\ \vdots \\ w_{2n} \end{bmatrix}, \quad \cdots, \quad w_m = \begin{bmatrix} w_{m1} \\ w_{m2} \\ \vdots \\ w_{mn} \end{bmatrix}, \quad \text{then}$$

$$l = g(f(\mathbb{W} h))$$

$$= g(f(\begin{bmatrix} w_1^T \\ w_2^T \\ \vdots \\ w_m^T \end{bmatrix} \cdot h))$$

$$= g\left(f\left(\begin{bmatrix} w_1^T \cdot h \\ w_2^T \cdot h \\ \vdots \\ w_m^T \cdot h \end{bmatrix}\right)\right) = g\left(f\left(\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_m \end{bmatrix}\right)\right)$$

$$= g\left(\begin{bmatrix} f(z_1) \\ f(z_2) \\ \vdots \\ f(z_m) \end{bmatrix}\right) = g\left(\begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{bmatrix}\right)$$

$$\frac{\partial l}{\partial w_i} = \frac{\partial v_i}{\partial w_i} \cdot \frac{\partial l}{\partial v_i} \quad (\text{Chain Rule})$$

$$= \left(\frac{\partial z_i}{\partial w_i} \cdot \frac{\partial v_i}{\partial z_i}\right) \cdot \frac{\partial l}{\partial v_i} \quad (\text{Chain Rule})$$

$$= \frac{\partial w_i^T h}{\partial w_i} \cdot f'(z_i) \cdot dv_i$$

$$= \begin{bmatrix} \dfrac{\partial(w_{i1}h_1 + w_{i2}h_2 + \cdots + w_{in}h_n)}{\partial w_{i1}} \\ \vdots \\ \dfrac{\partial(w_{i1}h_1 + w_{i2}h_2 + \cdots + w_{in}h_n)}{\partial w_{in}} \end{bmatrix} \cdot f'(z_i) \cdot dv_i$$

$$= \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_n \end{bmatrix} \cdot f'(z_i) \cdot dv_i$$

$$= h \cdot f'(z_i) \cdot dv_i$$

$$dw_i^T = \frac{\partial l}{\partial w_i^T} = f'(z_i) \cdot dv_i \cdot h^T$$

$$d\mathbb{W} = \frac{\partial l}{\partial \mathbb{W}} = \begin{bmatrix} \frac{\partial l}{\partial w_1^T} \\ \frac{\partial l}{\partial w_2^T} \\ \vdots \\ \frac{\partial l}{\partial w_m^T} \end{bmatrix} = \begin{bmatrix} f'(z_1) \cdot dv_1 \\ f'(z_2) \cdot dv_2 \\ \vdots \\ f'(z_m) \cdot dv_m \end{bmatrix} \cdot h^T$$

$$= (f'(\mathbb{W} \cdot h) \circ dv) \cdot h^T \quad \square$$

3). Corollary 1

If $v = f(\mathbb{W}h) = \mathbb{W}h$, then

$$dh = \mathbb{W}^T \cdot dv \quad (3)$$

$$d\mathbb{W} = dv \cdot h^T \quad (4)$$

Proof :

$$f(\mathbb{W}h) = f\left(\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_m \end{bmatrix}\right) = \begin{bmatrix} f(z_1) \\ f(z_2) \\ \vdots \\ f(z_m) \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_m \end{bmatrix}$$

$$f'(\mathbb{W}h) = f'\left(\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_m \end{bmatrix}\right) = \begin{bmatrix} f'(z_1) \\ f'(z_2) \\ \vdots \\ f'(z_m) \end{bmatrix} = \begin{bmatrix} \frac{dz_1}{dz_1} \\ \frac{dz_2}{dz_2} \\ \vdots \\ \frac{dz_m}{dz_m} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

$$\therefore\ dh = \mathbb{W}^T \cdot dv, \quad d\mathbb{W} = dv \cdot h^T \quad (\text{From } (1) \text{ and } (2)) \quad \square$$

④ Lemma 2

u, v and s are vectors. $s = u \circ f(v)$.

$du = \frac{\partial l}{\partial u}$, $dv = \frac{\partial l}{\partial v}$, $ds = \frac{\partial l}{\partial s}$, then

$$du = f(v) \circ ds \quad ⑤$$

$$dv = f'(v) \circ u \circ ds \quad ⑥$$

Proof:

$$l = g(s) = g(u \circ f(v))$$

$$du = \frac{\partial s}{\partial u} \cdot \frac{\partial l}{\partial s} = f(v) \circ ds$$

$$dv = \frac{\partial s}{\partial v} \cdot \frac{\partial l}{\partial s} = \frac{\partial f}{\partial v} \cdot \frac{\partial s}{\partial f} \cdot \frac{\partial l}{\partial s} = f'(v) \circ u \circ ds \quad \square$$

⑤ Corollary 2

If $f(v) = v$, then

$$du = v \circ ds \quad ⑦$$

$$dv = u \circ ds \quad ⑧$$

5. Backward Pass - Continue

Since $s_t = \mathbb{W}_{hy} \cdot h_t$,

$$dh_t = \mathbb{W}_{hy}^T \cdot ds_t \quad (\text{by } ③). \quad ⑨$$

$$d\mathbb{W}_{hy} = ds_t \cdot h_t^T \quad (\text{by } ④) \quad ⑩$$

$$\mathbb{W}_{xh} \cdot X_t + \mathbb{W}_{hh} \cdot h_{t-1}$$

$$= \begin{bmatrix} w_{xh11} & w_{xh12} & \cdots & w_{xh1n} \\ w_{xh21} & w_{xh22} & \cdots & w_{xh2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{xhm1} & w_{xhm2} & \cdots & w_{xhmn} \end{bmatrix} \cdot \begin{bmatrix} x_{t1} \\ x_{t2} \\ \vdots \\ x_{tn} \end{bmatrix} + \begin{bmatrix} w_{hh11} & w_{hh12} & \cdots & w_{hh1n} \\ w_{hh21} & w_{hh22} & \cdots & w_{hh2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{hhm1} & w_{hhm2} & \cdots & w_{hhmn} \end{bmatrix} \cdot \begin{bmatrix} h_{t-11} \\ h_{t-12} \\ \vdots \\ h_{t-1n} \end{bmatrix}$$

$$= \begin{bmatrix} w_{xh11} & w_{xh12} & \cdots & w_{xh1n} & w_{hh11} & w_{hh12} & \cdots & w_{hh1n} \\ w_{xh21} & w_{xh22} & \cdots & w_{xh2n} & w_{hh21} & w_{hh22} & \cdots & w_{hh2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ w_{xhm1} & w_{xhm2} & \cdots & w_{xhmn} & w_{hhm1} & w_{hhm2} & \cdots & w_{hhmn} \end{bmatrix} \cdot \begin{bmatrix} x_{t1} \\ x_{t2} \\ \vdots \\ x_{tn} \\ h_{t-11} \\ h_{t-12} \\ \vdots \\ h_{t-1n} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbb{W}_{xh} & \mathbb{W}_{hh} \end{bmatrix} \cdot \begin{bmatrix} X_t \\ h_{t-1} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbb{W}_{xh} & \mathbb{W}_{hh} \end{bmatrix} \cdot \begin{bmatrix} X_t ; h_{t-1} \end{bmatrix}$$

$$= T_{rnn} \cdot Z_t$$

$$\therefore \quad h_t = \sigma(T_{rnn} \cdot Z_t)$$

$$dZ_t = T_{rnn}^T \cdot (\sigma'(T_{rnn} Z_t) \circ dh_t) \quad (\text{by } ①) \qquad ⑪$$

$$\begin{bmatrix} dX_t \\ dh_{t-1} \end{bmatrix} = \begin{bmatrix} \mathbb{W}_{xh}^T \\ \mathbb{W}_{hh}^T \end{bmatrix} \cdot (\sigma'(T_{rnn} Z_t) \circ dh_t)$$

$$dX_t = \mathbb{W}_{xh}^T \cdot (\sigma'(T_{rnn} Z_t) \circ dh_t) \qquad ⑫$$

$$dh_{t-1} = \mathbb{W}_{hh}^T \cdot (\sigma'(T_{rnn} Z_t) \circ dh_t) \qquad ⑬$$

$$dT_{rnn} = (\sigma'(T_{rnn} Z_t) \circ dh_t) \cdot Z_t^T \quad (\text{by } ②) \qquad ⑭$$

$$\begin{bmatrix} d\mathbb{W}_{xh} & d\mathbb{W}_{hh} \end{bmatrix} = (\sigma'(T_{rnn} Z_t) \circ dh_t) \cdot \begin{bmatrix} X_t^T & h_{t-1}^T \end{bmatrix}$$

$$d\mathbb{W}_{xh} = (\sigma'(T_{rnn} Z_t) \circ dh_t) \cdot X_t^T \qquad ⑮$$

$$d\mathbb{W}_{hh} = (\sigma'(T_{rnn} Z_t) \circ dh_t) \cdot h_{t-1}^T \qquad ⑯$$

Backpropagation_Through_Time_Vanilla_RNN ( ) :

for t from T to 1 :

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

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

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

$$\frac{\partial l_t}{\partial \mathbb{W}_{xh}} = \frac{\partial l_{t+1}}{\partial \mathbb{W}_{xh}} + \left(\sigma'(T_{rnn} Z_t) \circ \frac{\partial l_t}{\partial h_t} \right) \cdot X_t^T \quad ⑮$$

$$\frac{\partial l_t}{\partial \mathbb{W}_{hh}} = \frac{\partial l_{t+1}}{\partial \mathbb{W}_{hh}} + \left(\sigma'(T_{rnn} Z_t) \circ \frac{\partial l_t}{\partial h_t} \right) \cdot h_{t-1}^T \quad ⑯$$

$$\frac{\partial l_t}{\partial X_t} = \mathbb{W}_{xh}^T \cdot \left(\sigma'(T_{rnn} Z_t) \circ \frac{\partial l_t}{\partial h_t} \right) \quad ⑫$$

$$\frac{\partial l_t}{\partial h_{t-1}} = \mathbb{W}_{hh}^T \cdot \left(\sigma'(T_{rnn} Z_t) \circ \frac{\partial l_t}{\partial h_t} \right) \quad ⑬$$

Comments:

1) $\frac{\partial l_t}{\partial \mathbb{W}_{hy}}, \frac{\partial l_t}{\partial \mathbb{W}_{xh}}, \frac{\partial l_t}{\partial \mathbb{W}_{hh}}, \frac{\partial l_t}{\partial h_t}$ are aggregated through time since

$$\frac{\partial l_{total}}{\partial \mathbb{W}} = \frac{\partial\, l_{t=T} + l_{t=T-1} + \cdots + l_{t=1}}{\partial \mathbb{W}} = \frac{\partial l_{t=T}}{\partial \mathbb{W}} + \frac{\partial l_{t=T-1}}{\partial \mathbb{W}} + \cdots + \frac{\partial l_{t=1}}{\partial \mathbb{W}}$$

2) $\sigma'(T_{rnn} Z_t) = \sigma(T_{rnn} Z_t) - \sigma^2(T_{rnn} Z_t) = \sigma\left([\mathbb{W}_{xh}\ \mathbb{W}_{hh}] \cdot \begin{bmatrix} X_t \\ h_{t-1} \end{bmatrix} \right) - \sigma^2\left([\mathbb{W}_{xh}\ \mathbb{W}_{hh}] \cdot \begin{bmatrix} X_t \\ h_{t-1} \end{bmatrix} \right.$

(D.L., P48)

203