RNN

HSK

June 2024

1 Introduce

Formula 1 has to go to Appendix for RNN BPPT.

$$x_t = x_t^o \ E \tag{1}$$
$$h_t = \Theta_h(W \ h_{t-1} \ + \ U \ x_t \ + \ b) \tag{2}$$
$$\Omega_t = V \ h_t \ + \ c \tag{3}$$
$$\hat{y}_t = \Theta_y(\Omega_t) \tag{4}$$
$$L_t = - \ y_t \ ln(\hat{y}_t) \tag{5}$$

in (1), we use embedding layer. Θ's represent all activation functions. In our example for analytics, Θ_h is tanh in (2), while Θ_y is softmax in (4), which is usually used.

2 Through Time for Recurrent Neural Network

2.1 Softmax

Softmax $(x_t) = S_t = \frac{e^{x_t}}{\sum e^{x_k}}$ for $t = 1, \ldots, k$

Since softmax is a $\mathbb{R}^k \to \mathbb{R}^k$ mapping function, most general Jacobian matrix for it:

$$\frac{\partial S}{\partial x} = \begin{bmatrix} \frac{\partial S_1}{\partial x_1} & \cdots & \frac{\partial S_1}{\partial x_k} \\ \vdots & & \\ \frac{\partial S_k}{\partial x_1} & \cdots & \frac{\partial S_k}{\partial x_k} \end{bmatrix}$$

Let's compute $\frac{\partial S_i}{\partial x_j}$ for some arbitrary i and j :

$$\frac{\partial S_i}{\partial x_j} = \frac{\partial}{\partial x_j} \frac{e^{x_i}}{\sum_k e^{x_k}}$$

Let's examine the formula for division

$$f(x) = \frac{g(x)}{h(x)},$$
$$f'(x) = \frac{g'(x)h(x) - g(x)h'(x)}{h(x)^2}$$

In our case $g_i = e^{x_i}$ and $h_i = \sum e^{x_k}$. No matter which x_j, when we com pute the derivative of h_i with respect to x_j, the answer will always be e^{x_j}.

$$\frac{\partial}{\partial x_j} h_i = \frac{\partial}{\partial x_j} \sum e^{x_k} = \sum \frac{\partial e^{x_k}}{\partial x_j} = e^{x_j}$$

because $\frac{\partial e^{x_k}}{\partial x_j} = 0$ for $k \neq j$. There are on the mean-fully derivatives for $i = j$ in $\frac{\partial S}{\partial x}$ matrices for our problem.

$$\begin{aligned}
\frac{\partial \frac{e^{x_i}}{\sum e^{x_k}}}{\partial x_j} &= \frac{e^{x_i} \sum e^{x_k} - e^{x_j} e^{x_i}}{\left(\sum e^{x_k}\right)^2} \\
&= \frac{e^{x_i} \left(\sum e^{x_k} - e^{x_j}\right)}{\left(\sum e^{x_k}\right)^2} \\
&= \frac{e^{x_i}}{\sum e^{x_k}} \cdot \left(\frac{\sum e^{x_k}}{\sum e^{x_k}} - \frac{e^{x_j}}{\sum e^{x_k}}\right)
\end{aligned}$$

$$S_i \left(1 - S_j\right) \tag{6}$$

Now we found what derivative of softmax. Let's go back to the loss function.

2.2 Derivative of Loss Function and Ω_t

Let's examine the derivative formula for logarithm

$$f(x) = \log_y x$$
$$f'(x) = \frac{x'}{x} \cdot \log_e y$$

$$
\begin{aligned}
L(\hat{y}, y) &= -\sum y_t \log\left(\operatorname{softmax}\left(\Omega_t\right)\right) \\
\frac{\partial L}{\partial \Omega_t} &= -\frac{\partial}{\partial \Omega_t} \sum y_t \log\left(\operatorname{softmax}\left(\Omega_t\right)\right) \\
&= -\sum y_t \frac{\partial \log\left(\operatorname{softmax}\left(\Omega_t\right)\right)}{\partial \Omega_t} \\
&= -\sum \frac{\partial \hat{y}_t}{\partial \Omega_t} \cdot \frac{y_t}{\hat{y}_t} \\
&= -\sum S_{t,i}\left(1 - S_{t,j}\right) \cdot \frac{y_t}{\hat{y}_t} \\
&= -\sum \left(1 - S_{t,j}\right) y_t \\
&= -\sum \left(y_t - S_{t,j} \hat{y}_t\right) \\
&= \sum S_{t,i} \hat{y}_t - \sum y_t \\
&= S_{t,j} \sum \hat{y}_t - \sum y_t \\
&= \sum (\hat{y}_t - y_t)
\end{aligned}
$$

$$
\frac{\partial L_t}{\partial \hat{\Omega}_t} = (\hat{y}_t - y_t) \tag{7}
$$

2.3 Derivative of V

The weight V is consistent across the entire time sequence, allowing us to perform differentiation at each time step and then aggregate the results.

$$
\begin{aligned}
\frac{\partial L}{\partial V} &= \sum_{t=1}^{S} \frac{\partial L_t}{\partial V} \\
&= \sum_{t=1}^{S} \frac{\partial L_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial \Omega_t} \frac{\partial \Omega_t}{\partial V} \\
&= \sum_{t=1}^{S} \frac{\partial L}{\partial \Omega_t} \frac{\partial \Omega_t}{\partial V}
\end{aligned}
$$

We know that this formula $\frac{\partial \hat{y}_t}{\partial \Omega_t}$ from (7) and no other function exists between Omega and V, so simply taking the derivative coefficient of V yields h, thus the answer is h.

$$
= \sum_{t=1}^{S} (\hat{y}_t - y_t) \cdot h_t^{\top} \tag{8}
$$

2.4 Derivative of c

Similar to V, but its derivative is easier to calculate since it stands alone in the function.

$$\begin{aligned}\frac{\partial L}{\partial c} &= \sum_{t=1}^{T} \frac{\partial L_t}{\partial c} \\ &= \sum_{t=1}^{T} \frac{\partial L_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial \Omega_t} \frac{\partial \Omega_t}{\partial c} \\ &= \sum_{t=1}^{T} \frac{\partial L}{\partial \Omega_t} \frac{\partial \Omega_t}{\partial c}\end{aligned}$$

In this case, The Analytical Derivatives of c becomes:

$$= \sum_{t=1}^{T} (\hat{y}_t - y_t) \tag{9}$$

2.5 Derivative of W

This function employs recursion, therefore, computing its derivative may take some time.

$$\begin{aligned}&h_t = tanh(W\ h_{t-1}\ +\ U\ x_t\ +\ b) \\ &h_1 = tanh(W\ h_0\ +\ U\ x_1\ +\ b) \quad h_2 = tanh(W\ h_1\ +\ U\ x_2\ +\ b) \\ &h_3 = tanh(W\ h_2\ +\ U\ x_3\ +\ b) \quad h_4 = tanh(W\ h_3\ +\ U\ x_4\ +\ b)\end{aligned}$$

By placing h_t into the last term, we get:

$$h_4 = tanh(W\ tanh(W\ tanh(W\ tanh(Wh_0 + Ux_1 + b) + Ux_2 + b) + Ux_3 + b) + Ux_4 + b)$$

We start from the first step go to the last step:

$$\begin{aligned}\frac{\partial L_1}{\partial W} &= \frac{\partial L_1}{\partial \hat{y}_1} \frac{\partial \hat{y}_1}{\partial h_1} \frac{\partial h_1}{\partial h_1} \frac{\partial h_1}{\partial W} \\ \frac{\partial L_2}{\partial W} &= \frac{\partial L_2}{\partial \hat{y}_2} \frac{\partial \hat{y}_2}{\partial h_2} \frac{\partial h_2}{\partial h_1} \frac{\partial h_1}{\partial W} + \frac{\partial L_2}{\partial \hat{y}_2} \frac{\partial \hat{y}_2}{\partial h_2} \frac{\partial h_2}{\partial h_2} \frac{\partial h_2}{\partial W} \\ \frac{\partial L_3}{\partial W} &= \frac{\partial L_3}{\partial \hat{y}_3} \frac{\partial \hat{y}_3}{\partial h_3} \frac{\partial h_3}{\partial h_1} \frac{\partial h_1}{\partial W} + \frac{\partial L_3}{\partial \hat{y}_3} \frac{\partial \hat{y}_3}{\partial h_3} \frac{\partial h_3}{\partial h_2} \frac{\partial h_2}{\partial W} + \frac{\partial L_3}{\partial \hat{y}_3} \frac{\partial \hat{y}_3}{\partial h_3} \frac{\partial h_3}{\partial h_3} \frac{\partial h_3}{\partial W} \\ \frac{\partial L_4}{\partial W} &= \frac{\partial L_4}{\partial \hat{y}_4} \frac{\partial \hat{y}_4}{\partial h_4} \frac{\partial h_4}{\partial h_1} \frac{\partial h_1}{\partial W} + \frac{\partial L_4}{\partial \hat{y}_4} \frac{\partial \hat{y}_4}{\partial h_4} \frac{\partial h_4}{\partial h_2} \frac{\partial h_2}{\partial W} + \frac{\partial L_4}{\partial \hat{y}_4} \frac{\partial \hat{y}_4}{\partial h_4} \frac{\partial h_4}{\partial h_3} \frac{\partial h_3}{\partial W} + \frac{\partial L_4}{\partial \hat{y}_4} \frac{\partial \hat{y}_4}{\partial h_4} \frac{\partial h_4}{\partial h_4} \frac{\partial h_4}{\partial W}\end{aligned}$$

Simplifying:

$$\frac{\partial L_1}{\partial W} = \frac{\partial L_1}{\partial \hat{y}_1}\frac{\partial \hat{y}_1}{\partial h_1}\frac{\partial h_1}{\partial W}$$
$$\frac{\partial L_2}{\partial W} = \frac{\partial L_2}{\partial \hat{y}_2}\frac{\partial \hat{y}_2}{\partial h_2}\frac{\partial h_2}{\partial h_1}\frac{\partial h_1}{\partial W} + \frac{\partial L_2}{\partial \hat{y}_2}\frac{\partial \hat{y}_2}{\partial h_2}\frac{\partial h_2}{\partial W}$$
$$\frac{\partial L_3}{\partial W} = \frac{\partial L_3}{\partial \hat{y}_3}\frac{\partial \hat{y}_3}{\partial h_3}\frac{\partial h_3}{\partial h_1}\frac{\partial h_1}{\partial W} + \frac{\partial L_3}{\partial \hat{y}_3}\frac{\partial \hat{y}_3}{\partial h_3}\frac{\partial h_3}{\partial h_2}\frac{\partial h_2}{\partial W} + \frac{\partial L_3}{\partial \hat{y}_3}\frac{\partial \hat{y}_3}{\partial h_3}\frac{\partial h_3}{\partial W}$$
$$\frac{\partial L_4}{\partial W} = \frac{\partial L_4}{\partial \hat{y}_4}\frac{\partial \hat{y}_4}{\partial h_4}\frac{\partial h_4}{\partial h_1}\frac{\partial h_1}{\partial W} + \frac{\partial L_4}{\partial \hat{y}_4}\frac{\partial \hat{y}_4}{\partial h_4}\frac{\partial h_4}{\partial h_2}\frac{\partial h_2}{\partial W} + \frac{\partial L_4}{\partial \hat{y}_4}\frac{\partial \hat{y}_4}{\partial h_4}\frac{\partial h_4}{\partial h_3}\frac{\partial h_3}{\partial W} + \frac{\partial L_4}{\partial \hat{y}_4}\frac{\partial \hat{y}_4}{\partial h_4}\frac{\partial h_4}{\partial W}$$

Since the calculation process needs to be simplified, let's expand $\frac{h_t}{h_k}$

$$\frac{\partial L_1}{\partial W} = \frac{\partial L_1}{\partial \hat{y}_1}\frac{\partial \hat{y}_1}{\partial h_1}\frac{\partial h_1}{\partial W}$$
$$\frac{\partial L_2}{\partial W} = \frac{\partial L_2}{\partial \hat{y}_2}\frac{\partial \hat{y}_2}{\partial h_2}\frac{\partial h_2}{\partial h_1}\frac{\partial h_1}{\partial W} + \frac{\partial L_2}{\partial \hat{y}_2}\frac{\partial \hat{y}_2}{\partial h_2}\frac{\partial h_2}{\partial W}$$
$$\frac{\partial L_3}{\partial W} = \frac{\partial L_3}{\partial \hat{y}_3}\frac{\partial \hat{y}_3}{\partial h_3}\frac{\partial h_3}{\partial h_2}\frac{\partial h_2}{\partial h_1}\frac{\partial h_1}{\partial W} + \frac{\partial L_3}{\partial \hat{y}_3}\frac{\partial \hat{y}_3}{\partial h_3}\frac{\partial h_3}{\partial h_2}\frac{\partial h_2}{\partial W} + \frac{\partial L_3}{\partial \hat{y}_3}\frac{\partial \hat{y}_3}{\partial h_3}\frac{\partial h_3}{\partial W}$$
$$\frac{\partial L_4}{\partial W} = \frac{\partial L_4}{\partial \hat{y}_4}\frac{\partial \hat{y}_4}{\partial h_4}\frac{\partial h_4}{\partial h_3}\frac{\partial h_3}{\partial h_2}\frac{\partial h_2}{\partial h_1}\frac{\partial h_1}{\partial W} + \frac{\partial L_4}{\partial \hat{y}_4}\frac{\partial \hat{y}_4}{\partial h_4}\frac{\partial h_4}{\partial h_3}\frac{\partial h_3}{\partial h_2}\frac{\partial h_2}{\partial W} + \frac{\partial L_4}{\partial \hat{y}_4}\frac{\partial \hat{y}_4}{\partial h_4}\frac{\partial h_4}{\partial h_3}\frac{\partial h_3}{\partial W} + \frac{\partial L_4}{\partial \hat{y}_4}\frac{\partial \hat{y}_4}{\partial h_4}\frac{\partial h_4}{\partial W}$$

Let us group them under a $\sum\prod$ for step t:

$$\frac{\partial L_t}{\partial W} = \sum_{k=1}^{t} \frac{\partial L_t}{\partial \hat{y}_t}\frac{\partial \hat{y}_t}{\partial h_t}\prod_{j=k}^{t-1}\left(\frac{\partial h_{j+1}}{\partial h_j}\right)\frac{\partial h_k}{\partial W}$$

Let us now present the formula for $\mathcal{S}$ steps:

$$\frac{\partial L}{\partial W} = \sum_{t=1}^{\mathcal{S}}\left(\sum_{k=1}^{t} \frac{\partial L_t}{\partial \hat{y}_t}\frac{\partial \hat{y}_t}{\partial h_t}\prod_{j=k}^{t-1}\left(\frac{\partial h_{j+1}}{\partial h_j}\right)\frac{\partial h_k}{\partial W}\right) \tag{10}$$

Finally, we insert the individual partial derivatives to calculate our final gradients of L with respect to W, where:

$$\frac{\partial L_t}{\partial \hat{y}_t} = (y_t - \hat{y}_t)$$
$$\frac{\partial \hat{y}_t}{\partial h_t} = V^\top$$
$$\frac{\partial h_{j+1}}{\partial h_j} = W^\top\ (1 - h_{j+1}^2)$$
$$\frac{\partial h_k}{\partial W} = (1 - h_k^2)\ h_{k-1}$$

In this case, The Analytical Derivatives of Eq. (10) becomes:

$$\frac{\partial L}{\partial W} = \sum_{t=1}^{\mathcal{S}}\left(\sum_{k=1}^{t} (y_t - \hat{y}_t)\ V^\top \prod_{j=k}^{t-1}\left(W^\top\ (1 - h_{j+1}^2)\right)(1 - h_k^2)\ h_{k-1}\right) \tag{11}$$

2.6 Derivative of U

Now, let's derive the gradient with respect to U. Similarly, we calculate the gradient with respect to U like Eq. (10) as follows:

$$\frac{\partial L}{\partial U} = \sum_{t=1}^{\mathcal{S}} \left(\sum_{k=1}^{t} \frac{\partial L_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial h_t} \prod_{j=k}^{t-1} \left(\frac{\partial h_{j+1}}{\partial h_j} \right) \frac{\partial h_k}{\partial U} \right) \tag{12}$$

we insert the individual partial derivatives to calculate our final gradients of L with respect to W, where:

$$\begin{aligned}
\frac{\partial L_t}{\partial \hat{y}_t} &= (y_t - \hat{y}_t) \\
\frac{\partial \hat{y}_t}{\partial h_t} &= V^\top \\
\frac{\partial h_{j+1}}{\partial h_j} &= W^\top \ (1 - h_{j+1}^2) \\
\frac{\partial h_k}{\partial U} &= (1 - h_k^2) \ x_k
\end{aligned}$$

In this case, The Analytical Derivatives of Eq. (12) becomes:

$$\frac{\partial L}{\partial W} = \sum_{t=1}^{\mathcal{S}} \left(\sum_{k=1}^{t} \ (y_t - \hat{y}_t) \ V^\top \prod_{j=k}^{t-1} \left(W^\top \ (1 - h_{j+1}^2) \right) (1 - h_k^2) \ h_{k-1} \right) \tag{13}$$

2.7 Derivative of b

Now, let's derive the gradient with respect to b. Similarly, we calculate the gradient with respect to b like Eq. (12) as follows:

$$\frac{\partial L}{\partial b} = \sum_{t=1}^{\mathcal{S}} \left(\sum_{k=1}^{t} \frac{\partial L_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial h_t} \prod_{j=k}^{t-1} \left(\frac{\partial h_{j+1}}{\partial h_j} \right) \frac{\partial h_k}{\partial b} \right) \tag{14}$$

we insert the individual partial derivatives to calculate our final gradients of L with respect to W, where:

$$\begin{aligned}
\frac{\partial L_t}{\partial \hat{y}_t} &= (y_t - \hat{y}_t) \\
\frac{\partial \hat{y}_t}{\partial h_t} &= V^\top \\
\frac{\partial h_{j+1}}{\partial h_j} &= W^\top \ (1 - h_{j+1}^2) \\
\frac{\partial h_k}{\partial b} &= (1 - h_k^2)
\end{aligned}$$

In this case, The Analytical Derivatives of Eq. (14) becomes:

$$\frac{\partial L}{\partial W} = \sum_{t=1}^{\mathcal{S}} \left(\sum_{k=1}^{t} \ (y_t - \hat{y}_t)\ V^\top \prod_{j=k}^{t-1} \left(\ W^\top \ (1 - h_{j+1}^2) \right) (1 - h_k^2) \ \right) \tag{15}$$

2.8 Derivative of E

ADD THIS ONE TO THE APPENDIX AS TH DIFFERENCE WHEN EMBEDDING LAYERS ARE USED

This function employs recursion, therefore, computing its derivative may take some time.

$$\begin{aligned}
&x_t = x_t^o \ E \\
&h_t = tanh(W \ h_{t-1} \ + \ U \ x_t \ + \ b) \\
&h_1 = tanh(W \ h_0 \ + \ U \ x_1 \ + \ b) \quad h_2 = tanh(W \ h_1 \ + \ U \ x_2 \ + \ b) \\
&h_3 = tanh(W \ h_2 \ + \ U \ x_3 \ + \ b) \quad h_4 = tanh(W \ h_3 \ + \ U \ x_4 \ + \ b)
\end{aligned}$$

By placing h_t into the last term, we get:

$$\begin{aligned}
&h_4 = \\
&\quad tanh(W \ tanh(W \ tanh(W \ tanh(W h_0 + U x_1 + b) + U x_2 + b) + U x_3 + b) + U x_4 + b)
\end{aligned}$$

We start from the first step go to the last step:

$$\begin{aligned}
\frac{\partial L_1}{\partial E} &= \frac{\partial L_1}{\partial \hat{y}_1} \frac{\partial \hat{y}_1}{\partial h_1} \frac{\partial h_1}{\partial h_1} \frac{\partial h_1}{\partial x_1} \frac{\partial x_1}{\partial E} \\
\frac{\partial L_2}{\partial E} &= \frac{\partial L_2}{\partial \hat{y}_2} \frac{\partial \hat{y}_2}{\partial h_2} \frac{\partial h_2}{\partial h_1} \frac{\partial h_1}{\partial x_1} \frac{\partial x_1}{\partial E} + \frac{\partial L_2}{\partial \hat{y}_2} \frac{\partial \hat{y}_2}{\partial h_2} \frac{\partial h_2}{\partial h_2} \frac{\partial h_2}{\partial x_2} \frac{\partial x_2}{\partial E} \\
\frac{\partial L_3}{\partial E} &= \frac{\partial L_3}{\partial \hat{y}_3} \frac{\partial \hat{y}_3}{\partial h_3} \frac{\partial h_3}{\partial h_1} \frac{\partial h_1}{\partial x_1} \frac{\partial x_1}{\partial E} + \frac{\partial L_3}{\partial \hat{y}_3} \frac{\partial \hat{y}_3}{\partial h_3} \frac{\partial h_3}{\partial h_2} \frac{\partial h_2}{\partial x_2} \frac{\partial x_2}{\partial E} + \frac{\partial L_3}{\partial \hat{y}_3} \frac{\partial \hat{y}_3}{\partial h_3} \frac{\partial h_3}{\partial h_3} \frac{\partial h_3}{\partial x_3} \frac{\partial x_3}{\partial E} \\
\frac{\partial L_4}{\partial W} &= \frac{\partial L_4}{\partial \hat{y}_4} \frac{\partial \hat{y}_4}{\partial h_4} \frac{\partial h_4}{\partial h_1} \frac{\partial h_1}{\partial x_1} \frac{\partial x_1}{\partial E} + \frac{\partial L_4}{\partial \hat{y}_4} \frac{\partial \hat{y}_4}{\partial h_4} \frac{\partial h_4}{\partial h_2} \frac{\partial h_2}{\partial x_2} \frac{\partial x_2}{\partial E} + \frac{\partial L_4}{\partial \hat{y}_4} \frac{\partial \hat{y}_4}{\partial h_4} \frac{\partial h_4}{\partial h_3} \frac{\partial h_3}{\partial x_3} \frac{\partial x_3}{\partial E} \\
&\quad + \frac{\partial L_4}{\partial \hat{y}_4} \frac{\partial \hat{y}_4}{\partial h_4} \frac{\partial h_4}{\partial h_4} \frac{\partial h_4}{\partial x_4} \frac{\partial x_4}{\partial E}
\end{aligned}$$

Simplifying:

$$
\begin{aligned}
\frac{\partial L_1}{\partial E} &= \frac{\partial L_1}{\partial \hat{y}_1}\frac{\partial \hat{y}_1}{\partial h_1}\frac{\partial h_1}{\partial x_1}\frac{\partial x_1}{\partial E}\\
\frac{\partial L_2}{\partial E} &= \frac{\partial L_2}{\partial \hat{y}_2}\frac{\partial \hat{y}_2}{\partial h_2}\frac{\partial h_2}{\partial h_1}\frac{\partial h_1}{\partial x_1}\frac{\partial x_1}{\partial E} + \frac{\partial L_2}{\partial \hat{y}_2}\frac{\partial \hat{y}_2}{\partial h_2}\frac{\partial h_2}{\partial x_2}\frac{\partial x_2}{\partial E}\\
\frac{\partial L_3}{\partial E} &= \frac{\partial L_3}{\partial \hat{y}_3}\frac{\partial \hat{y}_3}{\partial h_3}\frac{\partial h_3}{\partial h_1}\frac{\partial h_1}{\partial x_1}\frac{\partial x_1}{\partial E} + \frac{\partial L_3}{\partial \hat{y}_3}\frac{\partial \hat{y}_3}{\partial h_3}\frac{\partial h_3}{\partial h_2}\frac{\partial h_2}{\partial x_2}\frac{\partial x_2}{\partial E} + \frac{\partial L_3}{\partial \hat{y}_3}\frac{\partial \hat{y}_3}{\partial h_3}\frac{\partial h_3}{\partial x_3}\frac{\partial x_3}{\partial E}\\
\frac{\partial L_4}{\partial E} &= \frac{\partial L_4}{\partial \hat{y}_4}\frac{\partial \hat{y}_4}{\partial h_4}\frac{\partial h_4}{\partial h_1}\frac{\partial h_1}{\partial x_1}\frac{\partial x_1}{\partial E} + \frac{\partial L_4}{\partial \hat{y}_4}\frac{\partial \hat{y}_4}{\partial h_4}\frac{\partial h_4}{\partial h_2}\frac{\partial h_2}{\partial x_2}\frac{\partial x_2}{\partial E} + \frac{\partial L_4}{\partial \hat{y}_4}\frac{\partial \hat{y}_4}{\partial h_4}\frac{\partial h_4}{\partial h_3}\frac{\partial h_3}{\partial x_3}\frac{\partial x_3}{\partial E}\\
&\quad + \frac{\partial L_4}{\partial \hat{y}_4}\frac{\partial \hat{y}_4}{\partial h_4}\frac{\partial h_4}{\partial x_4}\frac{\partial x_4}{\partial E}
\end{aligned}
$$

Since the calculation process needs to be simplified, let's expand $\frac{h_t}{h_k}$

$$
\begin{aligned}
\frac{\partial L_1}{\partial E} &= \frac{\partial L_1}{\partial \hat{y}_1}\frac{\partial \hat{y}_1}{\partial h_1}\frac{\partial h_1}{\partial x_1}\frac{\partial x_1}{\partial E}\\
\frac{\partial L_2}{\partial E} &= \frac{\partial L_2}{\partial \hat{y}_2}\frac{\partial \hat{y}_2}{\partial h_2}\frac{\partial h_2}{\partial h_1}\frac{\partial h_1}{\partial x_1}\frac{\partial x_1}{\partial E} + \frac{\partial L_2}{\partial \hat{y}_2}\frac{\partial \hat{y}_2}{\partial h_2}\frac{\partial h_2}{\partial x_2}\frac{\partial x_2}{\partial E}\\
\frac{\partial L_3}{\partial E} &= \frac{\partial L_3}{\partial \hat{y}_3}\frac{\partial \hat{y}_3}{\partial h_3}\frac{\partial h_3}{\partial h_2}\frac{\partial h_2}{\partial h_1}\frac{\partial h_1}{\partial x_1}\frac{\partial x_1}{\partial E} + \frac{\partial L_3}{\partial \hat{y}_3}\frac{\partial \hat{y}_3}{\partial h_3}\frac{\partial h_3}{\partial h_2}\frac{\partial h_2}{\partial x_2}\frac{\partial x_2}{\partial E} + \frac{\partial L_3}{\partial \hat{y}_3}\frac{\partial \hat{y}_3}{\partial h_3}\frac{\partial h_3}{\partial x_3}\frac{\partial x_3}{\partial E}\\
\frac{\partial L_4}{\partial E} &= \frac{\partial L_4}{\partial \hat{y}_4}\frac{\partial \hat{y}_4}{\partial h_4}\frac{\partial h_4}{\partial h_3}\frac{\partial h_3}{\partial h_2}\frac{\partial h_2}{\partial h_1}\frac{\partial h_1}{\partial x_1}\frac{\partial x_1}{\partial E} + \frac{\partial L_4}{\partial \hat{y}_4}\frac{\partial \hat{y}_4}{\partial h_4}\frac{\partial h_4}{\partial h_3}\frac{\partial h_3}{\partial h_2}\frac{\partial h_2}{\partial x_2}\frac{\partial x_2}{\partial E} + \frac{\partial L_4}{\partial \hat{y}_4}\frac{\partial \hat{y}_4}{\partial h_4}\frac{\partial h_4}{\partial h_3}\frac{\partial h_3}{\partial x_3}\frac{\partial x_3}{\partial E}\\
&\quad + \frac{\partial L_4}{\partial \hat{y}_4}\frac{\partial \hat{y}_4}{\partial h_4}\frac{\partial h_4}{\partial x_4}\frac{\partial x_4}{\partial E}
\end{aligned}
$$

Let us group them under a $\sum\prod$ for step t:

$$\frac{\partial L_t}{\partial E} = \sum_{k=1}^{t} \frac{\partial L_t}{\partial \hat{y}_t}\frac{\partial \hat{y}_t}{\partial h_t}\prod_{j=k}^{t-1}\left(\frac{\partial h_{j+1}}{\partial h_j}\right)\frac{\partial h_k}{\partial x_k}\frac{\partial x_k}{\partial E}$$

Let us now present the formula for $\mathcal{S}$ steps:

$$\frac{\partial L}{\partial E} = \sum_{t=1}^{\mathcal{S}}\left(\sum_{k=1}^{t} \frac{\partial L_t}{\partial \hat{y}_t}\frac{\partial \hat{y}_t}{\partial h_t}\prod_{j=k}^{t-1}\left(\frac{\partial h_{j+1}}{\partial h_j}\right)\frac{\partial h_k}{\partial x_k}\frac{\partial x_k}{\partial E}\right) \tag{16}$$

Finally, we insert the individual partial derivatives to calculate our final gradients of L with respect to W, where:

$$\frac{\partial L_t}{\partial \hat{y}_t} = (y_t - \hat{y}_t)$$
$$\frac{\partial \hat{y}_t}{\partial h_t} = V^\top$$
$$\frac{\partial h_{j+1}}{\partial h_j} = W^\top \; (1 - h_{j+1}^2)$$
$$\frac{\partial h_k}{\partial x_k} = (1 - h_k^2) \; U$$
$$\frac{\partial x_k}{\partial E} = x_k^o$$

In this case, The Analytical Derivatives of Eq. (16) becomes:

$$\frac{\partial L}{\partial E} = \sum_{t=1}^{\mathcal{S}} \left(\sum_{k=1}^{t} \;\; (y_t - \hat{y}_t) \; V^\top \prod_{j=k}^{t-1} \left(\; W^\top \; (1 - h_{j+1}^2) \right) (1 - h_k^2) \; U \; x_k^o \right) \tag{17}$$