1. The introduction of RNN's propagation algorithm

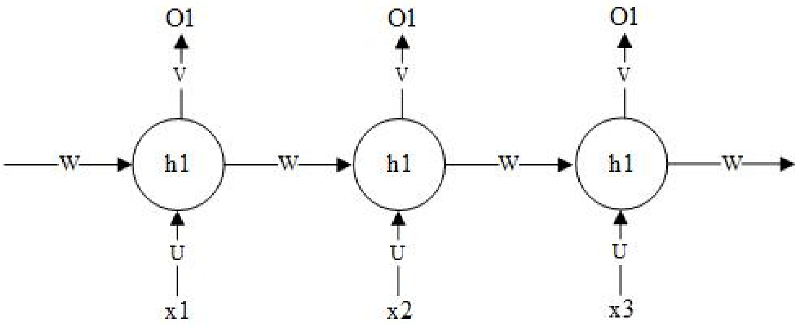

Figure 1 RNN

Symbol introduction:

$h_t = f(U \cdot x_t + W \cdot h_{t-1}), f\ is\ an\ activation\ function.$

$O_t = g(V \cdot h_t), g\ is\ an\ activation\ function.$

$\mathcal{L}_t = \frac{1}{2}(Y_t - O_t)^2$ and so on.

BP procedure:

Time step one:

$h_1 = f(U \cdot x_1 + W \cdot h_0), f\ is\ an\ activation\ function.$

$O_1 = g(V \cdot h_1), g\ is\ an\ activation\ function.$

$\mathcal{L}_1 = \frac{1}{2}(Y_1 - O_1)^2$

Compute the gradient of V:

$$\frac{\partial L_1}{\partial V} = \frac{\partial \mathcal{L}_1}{\partial V} = \frac{\partial \mathcal{L}_1}{\partial O_1} \cdot \frac{\partial O_1}{\partial V}$$

$$\Delta V =- \eta \frac{\partial \mathcal{L}_1}{\partial V}$$

$$V \leftarrow V + \Delta V$$

Compute the gradient of W:

$$\frac{\partial L_1}{\partial W} = \frac{\partial \mathcal{L}_1}{\partial W} = \frac{\partial \mathcal{L}_1}{\partial O_1} \cdot \frac{\partial O_1}{\partial h_1} \cdot \frac{\partial h_1}{\partial W}$$

$$\Delta W =- \eta \frac{\partial \mathcal{L}_1}{\partial W}$$

$$W \leftarrow W + \Delta W$$

Compute the gradient of U:

$$\frac{\partial L_1}{\partial U} = \frac{\partial \mathcal{L}_1}{\partial U} = \frac{\partial \mathcal{L}_1}{\partial O_1} \cdot \frac{\partial O_1}{\partial h_1} \cdot \frac{\partial h_1}{\partial U}$$

$$\Delta U =- \eta \frac{\partial \mathcal{L}_1}{\partial U}$$

$$U \leftarrow U + \Delta U$$

Time step two:

$h_2 = f(U \cdot x_2 + W \cdot h_1), f\ is\ an\ activation\ function.$

$O_2 = g(V \cdot h_2), g\ is\ an\ activation\ function.$

$\mathcal{L}_2 = \frac{1}{2}(Y_2 - O_2)^2$

Compute the gradient of V:

$$\frac{\partial L_2}{\partial V}=\frac{\partial(\mathcal{L}_1+\mathcal{L}_2)}{\partial V}=\frac{\partial\mathcal{L}_1}{\partial V}+\frac{\partial\mathcal{L}_2}{\partial V}=\frac{\partial\mathcal{L}_1}{\partial O_1}\cdot\frac{\partial O_1}{\partial V}+\frac{\partial\mathcal{L}_2}{\partial O_2}\cdot\frac{\partial O_2}{\partial V}$$

$$\Delta V=-\eta\frac{\partial(\mathcal{L}_1+\mathcal{L}_2)}{\partial V}$$

$$V\leftarrow V+\Delta V$$

Compute the gradient of W:

$$\frac{\partial L_2}{\partial W}=\frac{\partial(\mathcal{L}_1+\mathcal{L}_2)}{\partial W}=\frac{\partial\mathcal{L}_1}{\partial W}+\frac{\partial\mathcal{L}_2}{\partial W}=\frac{\partial\mathcal{L}_1}{\partial O_1}\cdot\frac{\partial O_1}{\partial h_1}\cdot\frac{\partial h_1}{\partial W}+\frac{\partial\mathcal{L}_2}{\partial O_2}\cdot\frac{\partial O_2}{\partial h_2}\cdot\frac{\partial h_2}{\partial h_1}\cdot\frac{\partial h_1}{\partial W}$$

$$\Delta W=-\eta\frac{\partial(\mathcal{L}_1+\mathcal{L}_2)}{\partial W}$$

$$W\leftarrow W+\Delta W$$

Compute the gradient of U:

$$\frac{\partial L_2}{\partial U}=\frac{\partial(\mathcal{L}_1+\mathcal{L}_2)}{\partial U}=\frac{\partial\mathcal{L}_1}{\partial U}+\frac{\partial\mathcal{L}_2}{\partial U}=\frac{\partial\mathcal{L}_1}{\partial O_1}\cdot\frac{\partial O_1}{\partial h_1}\cdot\frac{\partial h_1}{\partial U}+\frac{\partial\mathcal{L}_2}{\partial O_2}\cdot\frac{\partial O_2}{\partial h_2}\cdot\frac{\partial h_2}{\partial h_1}\cdot\frac{\partial h_1}{\partial U}$$

$$\Delta U=-\eta\frac{\partial(\mathcal{L}_1+\mathcal{L}_2)}{\partial U}$$

$$U\leftarrow U+\Delta U$$

<u>Time step three:</u>

$h_3=f(U\cdot x_3+W\cdot h_2), f\ is\ an\ activation\ function.$

$O_3=g(V\cdot h_3), g\ is\ an\ activation\ function.$

$\mathcal{L}_3=\frac{1}{2}(Y_3-O_3)^2$

Compute the gradient of V:

$$\frac{\partial L_3}{\partial V}=\frac{\partial(\mathcal{L}_1+\mathcal{L}_2+\mathcal{L}_3)}{\partial V}=\frac{\partial\mathcal{L}_1}{\partial V}+\frac{\partial\mathcal{L}_2}{\partial V}+\frac{\partial\mathcal{L}_4}{\partial V}=\frac{\partial\mathcal{L}_1}{\partial O_1}\cdot\frac{\partial O_1}{\partial V}+\frac{\partial\mathcal{L}_2}{\partial O_2}\cdot\frac{\partial O_2}{\partial V}+\frac{\partial\mathcal{L}_3}{\partial O_3}\cdot\frac{\partial O_3}{\partial V}$$

$$\Delta V=-\eta\frac{\partial(\mathcal{L}_1+\mathcal{L}_2+\mathcal{L}_3)}{\partial V}$$

$$V\leftarrow V+\Delta V$$

Compute the gradient of W:

$$\frac{\partial L_3}{\partial W}=\frac{\partial(\mathcal{L}_1+\mathcal{L}_2+\mathcal{L}_3)}{\partial W}=\frac{\partial\mathcal{L}_1}{\partial W}+\frac{\partial\mathcal{L}_2}{\partial W}+\frac{\partial\mathcal{L}_3}{\partial W}$$

$$=\frac{\partial\mathcal{L}_1}{\partial O_1}\cdot\frac{\partial O_1}{\partial h_1}\cdot\frac{\partial h_1}{\partial W}+\frac{\partial\mathcal{L}_2}{\partial O_2}\cdot\frac{\partial O_2}{\partial h_2}\cdot\frac{\partial h_2}{\partial h_1}\cdot\frac{\partial h_1}{\partial W}+\frac{\partial\mathcal{L}_3}{\partial O_3}\cdot\frac{\partial O_3}{\partial h_3}\cdot\frac{\partial h_3}{\partial h_2}\cdot\frac{\partial h_2}{\partial h_1}\cdot\frac{\partial h_1}{\partial W}$$

$$\Delta W=-\eta\frac{\partial(\mathcal{L}_1+\mathcal{L}_2+\mathcal{L}_3)}{\partial W}$$

$$W\leftarrow W+\Delta W$$

Compute the gradient of U:

$$\frac{\partial L_3}{\partial U}=\frac{\partial(\mathcal{L}_1+\mathcal{L}_2+\mathcal{L}_3)}{\partial U}=\frac{\partial\mathcal{L}_1}{\partial U}+\frac{\partial\mathcal{L}_2}{\partial U}+\frac{\partial\mathcal{L}_3}{\partial U}$$

$$=\frac{\partial\mathcal{L}_1}{\partial O_1}\cdot\frac{\partial O_1}{\partial h_1}\cdot\frac{\partial h_1}{\partial U}+\frac{\partial\mathcal{L}_2}{\partial O_2}\cdot\frac{\partial O_2}{\partial h_2}\cdot\frac{\partial h_2}{\partial h_1}\cdot\frac{\partial h_1}{\partial U}+\frac{\partial\mathcal{L}_3}{\partial O_3}\cdot\frac{\partial O_3}{\partial h_3}\cdot\frac{\partial h_3}{\partial h_2}\cdot\frac{\partial h_2}{\partial h_1}\cdot\frac{\partial h_1}{\partial U}$$

$$\Delta U=-\eta\frac{\partial(\mathcal{L}_1+\mathcal{L}_2+\mathcal{L}_3)}{\partial U}$$

$$U \leftarrow U + \Delta U$$

The gradient explosion and the gradient disappears：

The general formula in anytime step for the gradient of U and W：

$$\frac{\partial L_t}{\partial U} = \sum_{k=0}^{t} \frac{\partial L_t}{\partial O_t} \cdot \frac{\partial L_t}{\partial O_t} \cdot \left(\prod_{j=k+1}^{t} \frac{\partial h_j}{\partial h_{j-1}} \right) \cdot \frac{\partial h_k}{\partial U}$$

If the activation function is tanh, $h_j = \tanh(Ux + Wh + b)$,and $\prod_{j=k+1}^{t} \frac{\partial h_j}{\partial h_{j-1}} = \prod_{j=k+1}^{t} tanh' U$

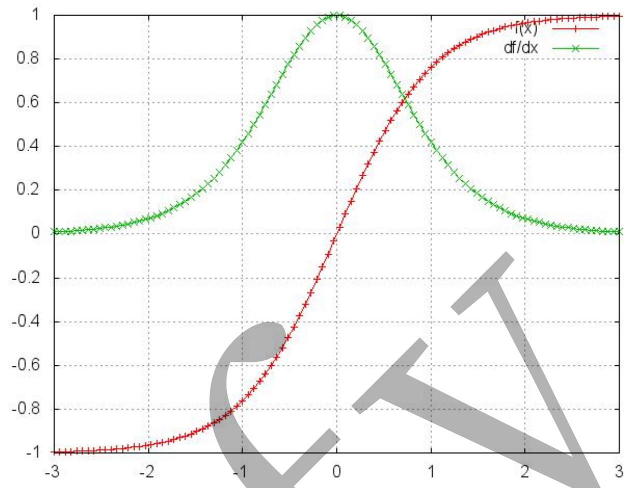

Figure2 tanh and $tanh'$

As shown in Figure2, in the case of $Ux + Wh + b = 0$, $tanh'$=1. In the most cases, $tanh' < 1$.

When 0<U<1, $\lim_{t\to\infty} \prod_{j=k+1}^{t} \frac{\partial h_j}{\partial h_{j-1}} = \lim_{t\to\infty} \prod_{j=k+1}^{t} tanh' U = 0$**(Gradient disappears)**.

When U is large, $tanh' U > 1$, and $\lim_{t\to\infty} \prod_{j=k+1}^{t} \frac{\partial h_j}{\partial h_{j-1}} = \lim_{t\to\infty} \prod_{j=k+1}^{t} tanh' U = \infty$ **(Gradient explosion)**

2.LSTM

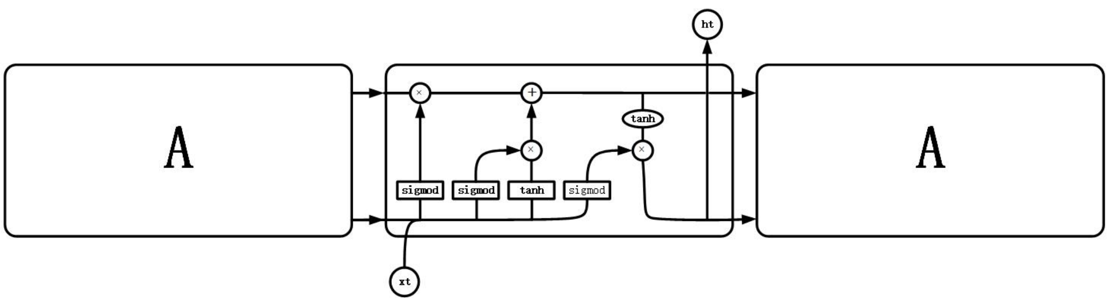

Figure3 LSTM

Three gates to control the data stream, f_t, i_t, o_t are respectively forget gate, input gate, and output gate. The value of sigmod function is between 0 and 1, 0 means enable to pass the gate, and 1 means able to pass the gate.

$$f_t = \sigma(W_f X_t + b_f)$$

$$i_t = \sigma(W_i X_t + b_i)$$
$$o_t = \sigma(W_o X_t + b_o)$$

In LSTM, $h_t = tanh[f_t h_{t-1} + i_t X_t] = tanh[\sigma(W_f X_t + b_f) h_{t-1} + \sigma(W_i X_t + b_i) X_t]$.

In the question of RNN, $\prod_{j=k+1}^{t} \frac{\partial h_j}{\partial h_{j-1}}$

In LSTM, it also has $\prod_{j=k+1}^{t} \frac{\partial h_j}{\partial h_{j-1}}$, but $\prod_{j=k+1}^{t} \frac{\partial h_j}{\partial h_{j-1}} = \prod_{j=k+1}^{t} tanh^{'} \sigma(W_f X_t + b_f)$, let's make $Z = tanh^{'}(x)\sigma(y)$, and the Figure of Z is shown as Figure 4.

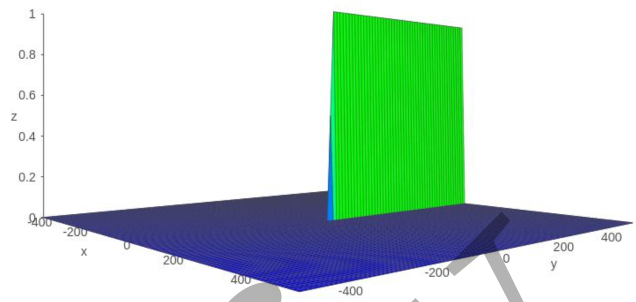

Figure 4 Function Z

From this figure, the value of this function is 0 and 1.

$$\frac{\partial L_3}{\partial U} = \frac{\partial(\mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3)}{\partial U} = \frac{\partial \mathcal{L}_1}{\partial U} + \frac{\partial \mathcal{L}_2}{\partial U} + \frac{\partial \mathcal{L}_3}{\partial U}$$
$$= \frac{\partial \mathcal{L}_1}{\partial O_1} \cdot \frac{\partial O_1}{\partial h_1} \cdot \frac{\partial h_1}{\partial U} + \frac{\partial \mathcal{L}_2}{\partial O_2} \cdot \frac{\partial O_2}{\partial h_2} \cdot \frac{\partial h_2}{\partial h_1} \cdot \frac{\partial h_1}{\partial U} + \frac{\partial \mathcal{L}_3}{\partial O_3} \cdot \frac{\partial O_3}{\partial h_3} \cdot \frac{\partial h_3}{\partial h_2} \cdot \frac{\partial h_2}{\partial h_1} \cdot \frac{\partial h_1}{\partial U}$$
$$\prod_{j=k+1}^{t} \frac{\partial h_j}{\partial h_{j-1}} = \prod_{j=k+1}^{t} tanh^{'} \sigma(W_f X_t + b_f) \approx 0 \ or \ 1$$

So that it could solve the problem of gradient explosion and disappears.