Github

Homework_3

ShiQiang

1、 Assume the time t=1,…,T, the total loss of rnn model is

$$L=\sum_{t=1}^{T}L_t$$

So,the update formula of V is:

$$\begin{aligned}V&=V-lr*\frac{\partial L}{\partial V}\\&=V-lr*\sum_{t=1}^{T}\frac{\partial L^t}{\partial V}\\&=V-lr*\sum_{t=1}^{T}\frac{\partial L^t}{\partial\hat{y}^{(t)}}\frac{\partial\hat{y}^{(t)}}{\partial O^{(t)}}\frac{\partial O^{(t)}}{\partial V}\end{aligned}$$

So,the update formula of W is:

$$\begin{aligned}W&=W-lr*\frac{\partial L}{\partial W}\\&=W-lr*\sum_{t=1}^{T}\frac{\partial L^{(t)}}{\partial W}\end{aligned}$$

$$\frac{\partial L^{(1)}}{\partial W}=\frac{\partial L^{(1)}}{\partial\hat{y}^{(1)}}\frac{\partial\hat{y}^{(1)}}{\partial O^{(1)}}\frac{\partial O^{(1)}}{\partial h^{(1)}}\frac{\partial h^{(1)}}{\partial W}$$

$$\frac{\partial L^{(2)}}{\partial W}=\frac{\partial L^{(2)}}{\partial\hat{y}^{(2)}}\frac{\partial\hat{y}^{(2)}}{\partial O^{(2)}}\frac{\partial O^{(2)}}{\partial h^{(2)}}\left(\frac{\partial h^{(2)}}{\partial W}+\frac{\partial h^{(2)}}{\partial h^{(1)}}\frac{\partial h^{(1)}}{\partial W}\right)$$

$$\frac{\partial L^{(3)}}{\partial W}=\frac{\partial L^{3}}{\partial\hat{y}^{(3)}}\frac{\partial\hat{y}^{(3)}}{\partial O^{(3)}}\frac{\partial O^{(3)}}{\partial h^{(3)}}\left(\frac{\partial h^{(3)}}{\partial W}+\frac{\partial h^{(3)}}{\partial h^{(2)}}\frac{\partial h^{(2)}}{\partial W}+\frac{\partial h^{(3)}}{\partial h^{(2)}}\frac{\partial h^{(2)}}{\partial h^{(1)}}\frac{\partial h^{(1)}}{\partial W}\right)$$

$$when\ T=3,$$

$$W=W-lr*\sum_{t=1}^{T}\frac{\partial L^{(t)}}{\partial W}=W-lr*\left(\frac{\partial L^{(1)}}{\partial W}+\frac{\partial L^{(2)}}{\partial W}+\frac{\partial L^{(3)}}{\partial W}\right)$$

the update formula of V is:

$$
\begin{aligned}
U &= U - lr * \frac{\partial L}{\partial U} \\
&= W - lr * \sum_{t=1}^{T} \frac{\partial L^{(t)}}{\partial U}
\end{aligned}
$$

$$
\frac{\partial L^{(1)}}{\partial U} = \frac{\partial L^{(1)}}{\partial \hat{y}^{(1)}} \frac{\partial \hat{y}^{(1)}}{\partial O^{(1)}} \frac{\partial O^{(1)}}{\partial h^{(1)}} \frac{\partial h^{(1)}}{\partial U}
$$

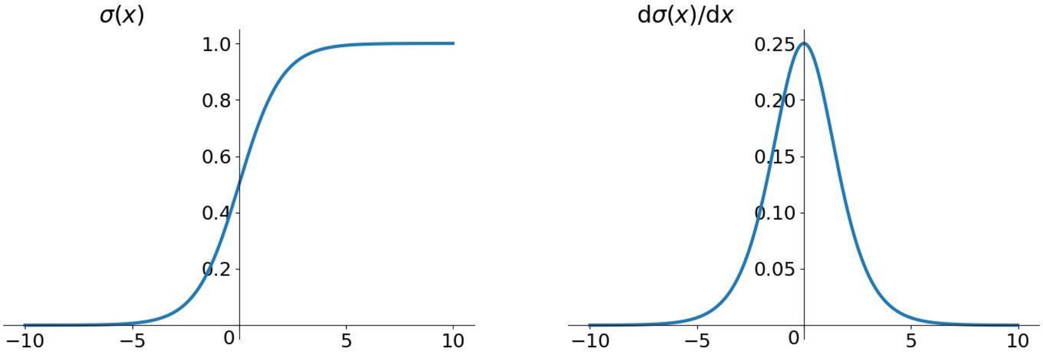
$$
\frac{\partial L^{(2)}}{\partial U} = \frac{\partial L^{(2)}}{\partial \hat{y}^{(2)}} \frac{\partial \hat{y}^{(2)}}{\partial O^{(2)}} \frac{\partial O^{(2)}}{\partial h^{(2)}} \left(\frac{\partial h^{(2)}}{\partial U} + \frac{\partial h^{(2)}}{\partial h^{(1)}} \frac{\partial h^{(1)}}{\partial U} \right)
$$

$$
\frac{\partial L^{(3)}}{\partial U} = \frac{\partial L^{3}}{\partial \hat{y}^{(3)}} \frac{\partial \hat{y}^{(3)}}{\partial O^{(3)}} \frac{\partial O^{(3)}}{\partial h^{(3)}} \left(\frac{\partial h^{(3)}}{\partial U} + \frac{\partial h^{(3)}}{\partial h^{(2)}} \frac{\partial h^{(2)}}{\partial U} + \frac{\partial h^{(3)}}{\partial h^{(2)}} \frac{\partial h^{(2)}}{\partial h^{(1)}} \frac{\partial h^{(1)}}{\partial U} \right)
$$

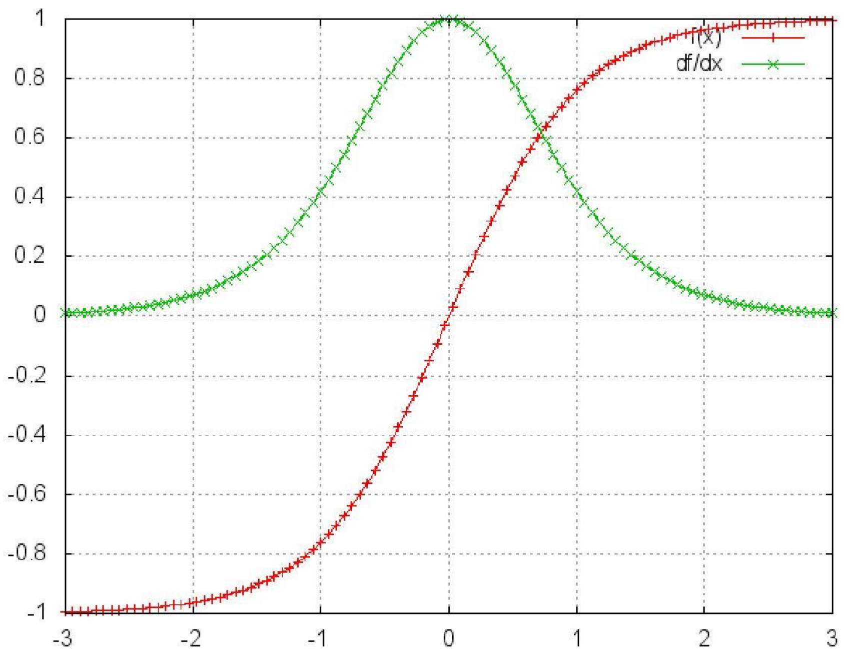
$$
when\ T = 3,
$$

$$
U = U - lr * \sum_{t=1}^{T} \frac{\partial L^{(t)}}{\partial U} = U - lr * \left(\frac{\partial L^{(1)}}{\partial U} + \frac{\partial L^{(2)}}{\partial U} + \frac{\partial L^{(3)}}{\partial U} \right)
$$

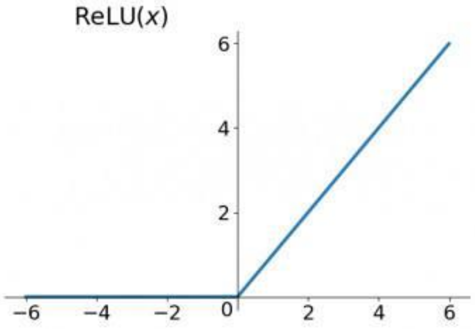
2）This is a function graph and a derivative graph of the sigmoid function.From the derivative graph, the derivative range of the sigmoid function is (0,0.25). As the time series continues to deepen, the multiplication of the decimals will cause the gradient to become smaller and smaller until it is close to zero. This is the phenomenon of “gradient disappearance”.

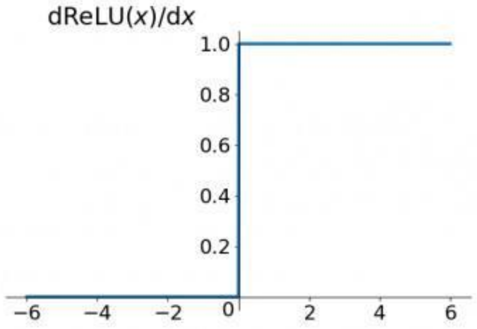


The function graph and derivative graph of the tanh function are as follows. From the derivative graph, the derivative range of the tanh function is (0,1]. Same as sigmoid,as the time series continues to deepen, the multiplication of the decimals will cause the gradient to become smaller and smaller until it is close to zero. This is the phenomenon of “gradient disappearance”, but the phenomenon more slowly than sigmoid.



The left derivative of the ReLU function is 0, and the right derivative is always 1, which avoids the occurrence of "gradient disappearance". But a derivative of constant 1 is easy to cause a "gradient explosion".


ReLU(x)


dReLU(x)/dx