Assignment 3

Fudi(Fred) Wang

November 2019

Part 1

(a) Since the loss of RNN accumulates with time, i.e. $L = \sum_{t=1}^{n} L^{(t)}$, we have:

$$\frac{\partial L}{\partial V} = \sum_{t=1}^{n} \frac{\partial L^{(t)}}{\partial o^{(t)}} \cdot \frac{\partial o^{(t)}}{\partial V}$$

Therefore the updating process on V is:

$$V =: V - \eta \frac{\partial L}{\partial V}$$

where η is the learning rate and $\frac{\partial L}{\partial V}$ is given as above.
Now the partial derivative with respect to U at time $t = 3$ is given below:

$$\begin{aligned}\frac{\partial L^{(3)}}{\partial U} &= \frac{\partial L^{(3)}}{\partial o^{(3)}} \cdot \frac{\partial o^{(3)}}{\partial h^{(3)}} \cdot \Big(\frac{\partial h^{(3)}}{\partial U} + \frac{\partial h^{(3)}}{\partial h^{(2)}} \cdot \frac{\partial h^{(2)}}{\partial U} + \frac{\partial h^{(3)}}{\partial h^{(2)}} \cdot \frac{\partial h^{(2)}}{\partial h^{(1)}} \cdot \frac{\partial h^{(1)}}{\partial U}\Big) \\ &= \frac{\partial L^{(3)}}{\partial o^{(3)}} \cdot \frac{\partial o^{(3)}}{\partial h^{(3)}} \cdot \frac{\partial h^{(3)}}{\partial U} + \frac{\partial L^{(3)}}{\partial o^{(3)}} \cdot \frac{\partial o^{(3)}}{\partial h^{(3)}} \cdot \frac{\partial h^{(3)}}{\partial h^{(2)}} \cdot \frac{\partial h^{(2)}}{\partial U} + \frac{\partial L^{(3)}}{\partial o^{(3)}} \cdot \frac{\partial o^{(3)}}{\partial h^{(3)}} \cdot \frac{\partial h^{(3)}}{\partial h^{(2)}} \cdot \frac{\partial h^{(2)}}{\partial h^{(1)}} \cdot \frac{\partial h^{(1)}}{\partial U}\end{aligned}$$

Similarly,

$$\begin{aligned}\frac{\partial L^{(3)}}{\partial W} &= \frac{\partial L^{(3)}}{\partial o^{(3)}} \cdot \frac{\partial o^{(3)}}{\partial h^{(3)}} \cdot \Big(\frac{\partial h^{(3)}}{\partial W} + \frac{\partial h^{(3)}}{\partial h^{(2)}} \cdot \frac{\partial h^{(2)}}{\partial W} + \frac{\partial h^{(3)}}{\partial h^{(2)}} \cdot \frac{\partial h^{(2)}}{\partial h^{(1)}} \cdot \frac{\partial h^{(1)}}{\partial W}\Big) \\ &= \frac{\partial L^{(3)}}{\partial o^{(3)}} \cdot \frac{\partial o^{(3)}}{\partial h^{(3)}} \cdot \frac{\partial h^{(3)}}{\partial W} + \frac{\partial L^{(3)}}{\partial o^{(3)}} \cdot \frac{\partial o^{(3)}}{\partial h^{(3)}} \cdot \frac{\partial h^{(3)}}{\partial h^{(2)}} \cdot \frac{\partial h^{(2)}}{\partial W} + \frac{\partial L^{(3)}}{\partial o^{(3)}} \cdot \frac{\partial o^{(3)}}{\partial h^{(3)}} \cdot \frac{\partial h^{(3)}}{\partial h^{(2)}} \cdot \frac{\partial h^{(2)}}{\partial h^{(1)}} \cdot \frac{\partial h^{(1)}}{\partial W}\end{aligned}$$

Hence the updating process follows similarly:

$$U_{i+1} =: U_i - \eta \frac{\partial L^{(i)}}{\partial U_i} \qquad \text{and} \qquad W_{i+1} =: W_i - \eta \frac{\partial L^{(i)}}{\partial W_i}$$

for $i = 1, 2, 3$.

(b) From the derivation above, we can summarize the following:

$$\frac{\partial L^{(t)}}{\partial U} = \sum_{k=0}^{t} \frac{\partial L^{(t)}}{\partial o^{(t)}} \cdot \frac{\partial o^{(t)}}{\partial h^{(t)}} \Big(\prod_{j=k+1}^{t} \frac{\partial h^{(j)}}{\partial h^{(j-1)}} \Big) \frac{\partial h^{(k)}}{\partial U}$$

Same for W. Note that, when we apply activation functions, we have the following:

$$\prod_{j=k+1}^{t} \frac{\partial h^{(j)}}{\partial h^{(j-1)}} = \prod_{j=k+1}^{t} tanh' \cdot U_s$$

or

$$\prod_{j=k+1}^{t} \frac{\partial h^{(j)}}{\partial h^{(j-1)}} = \prod_{j=k+1}^{t} sigmoid' \cdot U_s$$

or even

$$\prod_{j=k+1}^{t} \frac{\partial h^{(j)}}{\partial h^{(j-1)}} = \prod_{j=k+1}^{t} ReLU' \cdot U_s$$

We can see that, in fact, it is the product of different activation functions. Since the range of the derivative of tanh function is (0, 1] and the range of the derivative of the sigmoid function is (0, 0.25], we know that the gradient vanishes when we have a product of numbers that are less than one. On the other hand, the derivative of ReLU function is zero when $x < 0$ and one when $x \geq 0$, this will lead to gradient explosion.