Homework 3

Zhengyun Zhao

November 4, 2019

Part 1

1.

$$h^{(1)} = \varphi(UX^{(1)} + Wh^{(0)} + b)$$

$$h^{(2)} = \varphi(UX^{(2)} + Wh^{(1)} + b)$$

$$h^{(3)} = \varphi(UX^{(3)} + Wh^{(2)} + b)$$

$$O^{(3)} = Vh^{(3)} + c$$

$$\tilde{y}^{(3)} = softmax(O^{(3)})$$

$$L^{(3)} = CrossEntropy(\tilde{y}^{(3)}, y^{(3)})$$

$$\frac{\partial L^{(3)}}{\partial V} = \frac{\partial L^{(3)}}{\partial O^{(3)}} \frac{\partial O^{(3)}}{\partial V} = (\tilde{y}^{(3)} - y^{(3)})h^{(3)}$$

$$\frac{\partial L^{(3)}}{\partial W} = \frac{\partial L^{(3)}}{\partial O^{(3)}} \frac{\partial O^{(3)}}{\partial h^{(3)}} \frac{\partial h^{(3)}}{\partial W} = (\tilde{y}^{(3)} - y^{(3)})V(1 - (h^{(3)})^2)(h^{(2)} + W(1 - (h^{(2)})^2)(h^{(1)} + W(1 - (h^{(1)})^2)h^{(0)}))$$

$$\frac{\partial L^{(3)}}{\partial W} = \frac{\partial L^{(3)}}{\partial O^{(3)}} \frac{\partial O^{(3)}}{\partial h^{(3)}} \frac{\partial h^{(3)}}{\partial W} = (\tilde{y}^{(3)} - y^{(3)})V(1 - (h^{(3)})^2)(X^{(3)} + W(1 - (h^{(2)})^2)(X^{(2)} + W(1 - (h^{(1)})^2)X^{(1)}))$$

2.

When the sequence is long, the term $\frac{\partial h^{(t)}}{\partial W} = \sum_{i=1}^{t} \frac{\partial h^{(t)}}{\partial h(i)} \frac{\partial h^{(i)}}{\partial W} = \sum_{i=1}^{t} \frac{\partial h^{(t)}}{\partial s(t)} \frac{\partial s^{(t)}}{\partial h(i)} \frac{\partial h^{(i)}}{\partial W}$,where $s^{(i)} = UX^{(i)} + Wh^{(i-1)} + b$ is a multiple of many terms. For sigmoid and tanh activate function, $\frac{\partial h^{(t)}}{\partial s^{(t)}}$ can be so large or so small that the multiple of many terms makes the gradient so small or large. For Relu, it can be zero thus the gradient for previous cells are all zero.