Part_1:

1) 由BPTT相关性质，RNN中参数的更新要以求出损失函数 L 关于这些参数的梯度为前提。假设当前时刻为 τ ($\tau > 1$)。

$$\Rightarrow \nabla_V L = \sum_{t=1}^{\tau} (\nabla_{o^{(t)}} L)\, h^{(t)T}$$

其中 $\nabla_{o^{(t)}} L = \frac{\partial L}{\partial \hat{y}^{(t)}} \otimes (\hat{y}^{(t)} - y)$，$\otimes$ 是逐元素相乘符号。

$$\Rightarrow \nabla_W L = \sum_{t=1}^{\tau} (\nabla_{h^{(t)}} L) \frac{\partial h^{(t)}}{\partial h^{(t-1)}}$$

若 $t = \tau$，则 $\nabla_{h^{(t)}} L = \nabla_{h^{(\tau)}} L = V^T \nabla_{o^{(\tau)}} L$

若 $t \neq \tau$，则 $\nabla_{h^{(t)}} L = \left(\frac{\partial h^{(t+1)}}{\partial h^{(t)}}\right)^T (\nabla_{h^{(t+1)}} L) + \left(\frac{\partial o^{(t)}}{\partial h^{(t)}}\right)^T (\nabla_{o^{(t)}} L)$

$$= W^T (\nabla_{h^{(t+1)}} L)\, \mathrm{diag}(1 - (h^{(t+1)})^2) + V^T (\nabla_{o^{(t)}} L)$$

其中 $\frac{\partial h^{(t)}}{\partial h^{(t-1)}} = W^T \mathrm{diag}(1 - h^{(t)2})$

$$\therefore \nabla_W L = \sum_{t=1}^{\tau} \mathrm{diag}(1 - (h^{(t)})^2)(\nabla_{h^{(t)}} L)\, h^{(t-1)T}$$

$$\Rightarrow \nabla_U L = \sum_{t=1}^{\tau} \frac{\partial L}{\partial h^{(t)}} \nabla_U h^{(t)}$$

$$= \sum_{t=1}^{\tau} \mathrm{diag}(1 - (h^{(t)})^2)(\nabla_{h^{(t)}} L) x^{(t)T}$$

2) 当使用 sigmoid 函数和 tanh 函数时可能会产生梯度消失问题。

因为这两个激活函数的导数均小于 1，

在沿时间反向传播误差的过程中，反馈值将会趋近于 0。

这会使前面的信息无法有效被利用。