Writing: BPTT

AIG 309505002 鄭詔文

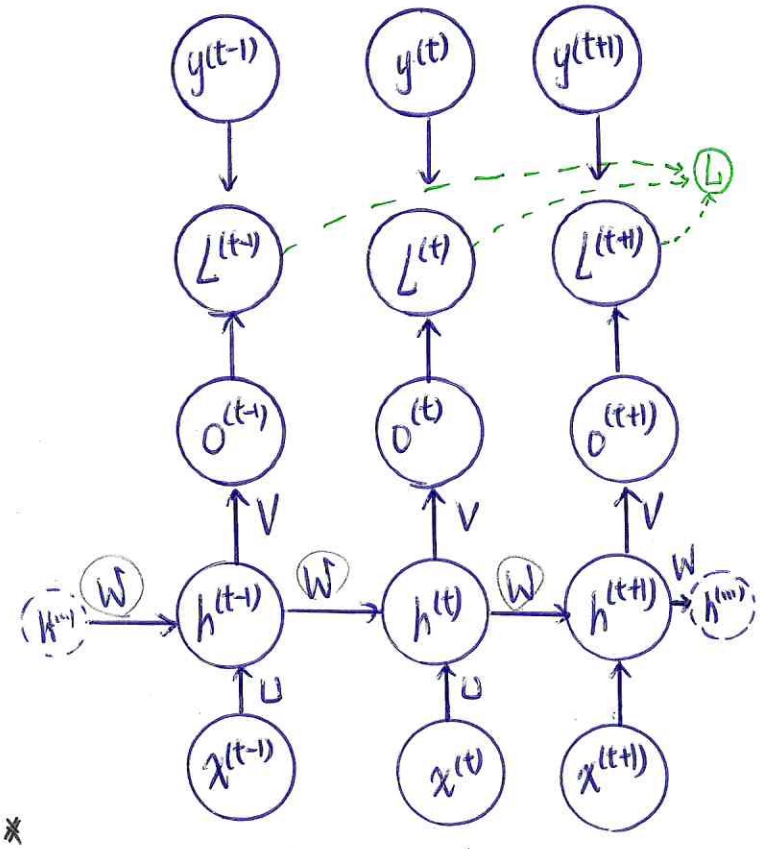
Ex: $\underline{a}^{(t)} = b + \underline{W}\underline{h}^{(t-1)} + \underline{U}\underline{x}^{(t)}$

$\underline{h}^{(t)} = \tanh(\underline{a}^{(t)})$

$\underline{o}^{(t)} = c + \underline{V}\underline{h}^{(t)}$

$L = \sum_t L^{(t)}$

$\nabla_{\underline{o}^{(t)}} L = \underline{\hat{y}}^{(t)} - \underline{y}^{(t)}$



* $J_{ij} = \partial f_i / \partial x_j$, $f: R^n \to R^m$

$f = Ax$, $\partial f / \partial x = A^T$

bias:

$$\nabla_c L = \sum_t \left(\frac{\partial o^{(t)}}{\partial c}\right)^T \left(\nabla_{\underline{o}^{(t)}} L\right)$$

$$= \sum_t \underbrace{\nabla_c o^{(t)}}_{=1} \cdot \nabla_{\underline{o}^{(t)}} L^{(t)}$$

$$= \sum_t \nabla_{\underline{o}^{(t)}} L^{(t)} = \sum_t \nabla_{\underline{o}^{(t)}} L \quad *$$

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

$$= W^T H^{(t+1)} \left(\nabla_{h^{(t+1)}} L\right) + V^T \left(\nabla_{\underline{o}^{(t)}} L\right) \quad *$$

* $h^{(t+1)}$ 依賴 $h^{(t)}$, $o^{(t)}$.

where

$$\underline{H}^{(t+1)} = \left(\frac{\partial h^{(t+1)}}{\partial \underline{a}^{(t+1)}}\right)^T$$

$$= \begin{bmatrix} 1-(\underline{h}_1^{(t+1)})^2 & 0 & \cdots & 0 \\ 0 & 1-(\underline{h}_2^{(t+1)})^2 & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & 1-(\underline{h}_n^{(t+1)})^2 \end{bmatrix}$$

$d\tanh(x)/dx = 1 - \tanh^2(x)$

$$\nabla_b L = \sum_t \underbrace{\left(\frac{\partial h^{(t)}}{\partial b}\right)^T \nabla_{h^{(t)}} L}_{\| \; \sum_t \sum_i \left(\frac{\partial L}{\partial \underline{h}_i^{(t)}}\right)\left(\nabla_b h_i^{(t)}\right)} = \sum_t \operatorname{diag}\left(1-(\underline{h}^{(t)})^2\right)\left(\nabla_{h^{(t)}} L\right) = \sum_t \underline{H}^{(t)} \cdot \nabla_{h^{(t)}} L \quad *$$

$$\nabla_{\underline{\underline{V}}} L = \sum_t \sum_i \left(\frac{\partial L}{\partial h_i^{(t)}}\right)\left(\nabla_{\underline{\underline{V}}} h_i^{(t)}\right) = \sum_t \left(\frac{\partial o^{(t)}}{\partial V}\right)\left(\nabla_{o^{(t)}} L\right) = \sum_t \left(\nabla_{o^{(t)}} L\right)\left(h^{(t)}\right)^T$$

$$\nabla_{\underline{\underline{W}}} L = \sum_t \sum_i \left(\frac{\partial L}{\partial h_i^{(t)}}\right)\left(\nabla_{\underline{\underline{W}}} h_i^{(t)}\right) = \sum_t \left(\frac{\partial L}{\partial h^{(t)}}\right)\left(\frac{\partial h^{(t)}}{\partial W}\right) = \sum \operatorname{diag}\left(1-\left(h^{(t)}\right)^2\right)\left(\nabla_{h^{(t)}} L\right)\left(h^{(t-1)}\right)^T$$

$$= \sum_t H^{(t)}\left(\nabla_{h^{(t)}} L\right)\left(h^{(t-1)}\right)^T$$

$$\nabla_{\underline{\underline{U}}} L = \sum_t \sum_i \left(\frac{\partial L}{\partial h_i^{(t)}}\right)\left(\nabla_{\underline{\underline{U}}} h_i^{(t)}\right) = \sum_t \left(\frac{\partial L}{\partial h^{(t)}}\right)\left(\frac{\partial h^{(t)}}{\partial U}\right)$$

$$= \sum_t \operatorname{diag}\left(1-\left(h^{(t)}\right)^2\right)\left(\nabla_{h^{(t)}} L\right)\left(x^{(t)}\right)^T = \sum_t H^{(t)}\left(\nabla_{h^{(t)}} L\right)\left(x^{(t)}\right)^T$$