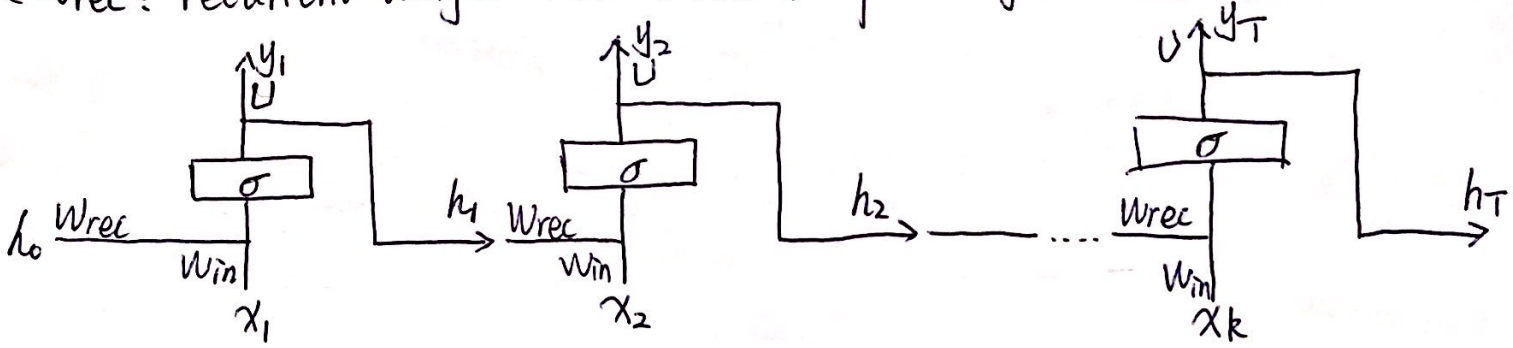
Recurrent Neural Network (RNN)

RNN基本结构

(Wrec: recurrent weight matrix, Win = input weight matrix)



$$h_t = \sigma(W_{rec} h_{t-1} + W_{in} x_t + b_t)$$

Back-prop ~~in~~ through time (BPTT). let $W = [W_{rec}, W_{in}]$

$$\frac{\partial E}{\partial W} = \sum_{t=1}^{T} \frac{\partial E_t}{\partial W}$$ (t: 第t个 timestep)

$$\frac{\partial E_k}{\partial W} = \frac{\partial E_k}{\partial y_k} \cdot \frac{\partial y_k}{\partial h_k} \left(\prod_{t=2}^{k} \frac{\partial h_t}{\partial h_{t-1}} \right) \cdot \frac{\partial h_1}{\partial W}$$

其中: $\frac{\partial h_t}{\partial h_{t-1}} = diag(\sigma'(W_{rec} h_{t-1} + W_{in} x_t + b_t)) W_{rec}$

so: $\prod_{t=2}^{k} \frac{\partial h_t}{\partial h_{t-1}} = \prod_{t=2}^{k} \left(diag(\sigma'(W_{rec} h_{t-1} + W_{in} x_t + b_t)) \cdot W_{rec} \right)$ ①

σ为sigmoid时，σ'最大为$\frac{1}{4}$；σ为tanh或ReLU，最大为1。当网络较深时，即K较大时，①式在连乘下会变得极小导致梯度消失；而当Wrec过大时，①又会导致梯度爆炸.