RNN:

1. 自回归模型：

一个变量 y_t 的历史信息来预测自己

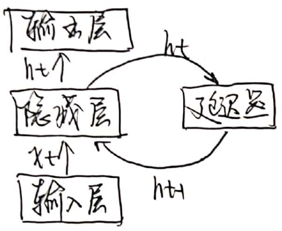
$$y_t = w_0 + \sum_{i=1}^{p} w_i y_{t-i} + \epsilon_t$$

有外部输入的非线性自回归.

$h_t =$

2. 循环 RNN

$$h_t = f(h_{t-1}, x_t) \quad h_0 = 0$$



可近似任意非线性动力系统

$$h_t = f(Uh_{t-1} + Wx_t + b)$$

3 BPTT

$$L = \sum_{t=1}^{T} L_t$$

$$\frac{\partial L}{\partial U} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial U}$$

$$\frac{\partial L_t}{\partial U_{ij}} = \sum_{k=1}^{t} \frac{\partial^+ z_k}{\partial u_{ij}} \frac{\partial L_t}{\partial z_k}$$

$$z_k = Uh_{k-1} + Wx_k + b$$

$$\frac{\partial^+ z_k}{\partial u_{ij}} = I_i([h_{k-1}]_j)$$

$$\delta_{t,k} = \frac{\partial L_t}{\partial z_k} = \frac{\partial h_k}{\partial z_k} \frac{\partial z_{k+1}}{\partial h_k} \frac{\partial L^t}{\partial z_{k+1}} = \mathrm{diag}(f'(z_k))U^T \delta_{t,k+1}$$

第t时刻的损失对第k时刻隐藏神经层的净输入 z_k 的导数

$$\therefore \frac{\partial L_t}{\partial U_{ij}} = \sum_{k=1}^{t} [\delta_{t,k}]_i [h_{k-1}]_j$$

$$\frac{\partial L_t}{\partial U} = \sum_{k=1}^{t} \delta_{t,k} h_{k-1}^T$$

同理

$$\frac{\partial L}{\partial U} = \sum_{t=1}^{T} \sum_{k=1}^{t} \delta_{t,k} h_{k-1}^T$$

$$\frac{\partial L}{\partial W} = \sum_{t=1}^{T} \sum_{k=1}^{t} \delta_{t,k} x_k^T$$

$$\frac{\partial L}{\partial b} = \sum_{t=1}^{T} \sum_{k=1}^{t} \delta_{t,k}$$

4 长程依赖问题

难以建模长时间间隔之间关系

$$\delta_{t,k} = \frac{\partial L_t}{\partial z_k} = \prod_{\tau=k}^{t-1} (\mathrm{diag}(f'(z_\tau))U^T) \delta_{t,t}$$

令 $\gamma \approx \|\mathrm{diag}(f'(z_k))U^T\|$，则

$$\delta_{t,k} \cong \gamma^{t-k} \delta_{t,t}$$

若 $t-k \to \infty$，$\begin{cases} \gamma > 1, & \delta_{t,k} \to \infty \\ \gamma < 1 & \delta_{t,k} \to 0. \end{cases}$ V

注意 RNN 中说的梯度消失并非指 $\frac{\partial L_t}{\partial U}$ 消失了，$\because \frac{\partial L_t}{\partial U} = \sum_{k=1}^{t} \delta_{t,k} h_{k-1}^T$

而 $\delta_{t,k} \cong \gamma^{t-k} \delta_{t,t}$

$\therefore$ 指的是那些离t很远的k，也就是更新主要依赖当前时刻t的几个相邻状态 h_k，长距离的状态对U没有影响

基础改进：$h_t = h_{t-1} + g(x_t, h_{t-1}, \theta)$

(1) 梯度爆炸

(2) 记忆容量

5 基于门控的RNN (LSTM)

三个门：遗忘门：$f_t = \sigma(W_f x_t + U h_{t-1} + b_f)$

输入门：$i_t = \sigma(W_i x_t + U h_{t-1} + b_i)$

输出门：$o_t = \sigma(W_o x_t + U h_{t-1} + b_o)$

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t$$

$$\tilde{c}_t = i_t \odot \tanh(W_c x_t + U_c h_{t-1} + b_c)$$

$$h_t = o_t \odot \tanh(c_t)$$

梯度消失?

注：一般DP网络参数初始化较小，但LSTM，较小会使f偏小，导致大量信息丢失，所以一般初始化较大（相邻时间间隔梯度非常小）

6. 门控循环单元网络 (GRU)

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot g(x_t, h_{t-1}; \theta)$$

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z)$$

$$\tilde{h}_t = g(x_t, h_{t-1}; \theta)$$

$$= \tanh(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h)$$

r_t 为重置门，是否清除 h_{t-1}

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r)$$