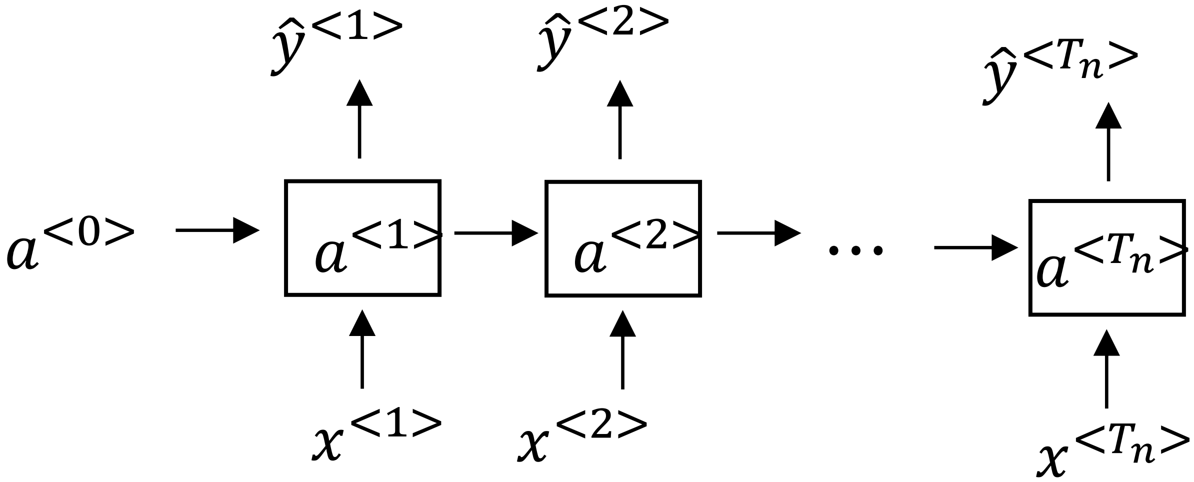

如上图所示 RNN，我们此时来推导从 n 时刻的损失回传的公式。
这里我们假设没有激活函数，则 rnn 的计算公式如下所示：

$$a^{<t>} = (w_x x + w_a a^{<t-1>} + b)$$

$$y^{<t>} = (w_y a^{<t>} + b)$$

求 Ln 对 w 参数的梯度：（Ln 为损失函数，即损失函数为预测值$\hat{y}_n$ 的函数）

$$\frac{\partial L_n}{\partial W_{ax}} = \frac{\partial L_n}{\partial \hat{y}_n}\frac{\partial \hat{y}_n}{\partial W_{ax}} = \frac{\partial L_n}{\partial \hat{y}_n}\frac{\partial \hat{y}_n}{\partial a_n}\frac{\partial a_n}{\partial W_{ax}} = \frac{\partial L_n}{\partial \hat{y}_n}\frac{\partial \hat{y}_n}{\partial a_n}\left(x^{<n>} + \frac{\partial a_n}{\partial a_{n-1}}\frac{\partial a_{n-1}}{\partial W_{ax}}\right)$$

上式课展开为两项 第一项为 $\frac{\partial L_n}{\partial \hat{y}_n}\frac{\partial \hat{y}_n}{\partial a_n} x^{<n>}$ 第二项为 $\frac{\partial L_n}{\partial \hat{y}_n}\frac{\partial \hat{y}_n}{\partial a_n}\frac{\partial a_n}{\partial a_{n-1}}\frac{\partial a_{n-1}}{\partial W_{ax}}$

第一项为 Ln 传到 Tn 的导数， 第二项为继续沿着时间点回传的梯度。

由第二项递推可得回传到 T=0 的梯度为

$$\frac{\partial L_n}{\partial \hat{y}_n}\frac{\partial \hat{y}_n}{\partial a_n}\frac{\partial a_n}{\partial a_{n-1}}\frac{\partial a_{n-1}}{\partial a_{n-2}}\frac{\partial a_{n-2}}{\partial a_{n-3}}\cdots\cdots\frac{\partial a_2}{\partial a_1}\frac{\partial a_1}{\partial a_0}\frac{\partial a_0}{\partial W_{ax}}$$

上述公式中有一串连乘，若都小于 1 则多次乘法后的数值接近零因此产生长距离的梯度消失问题。