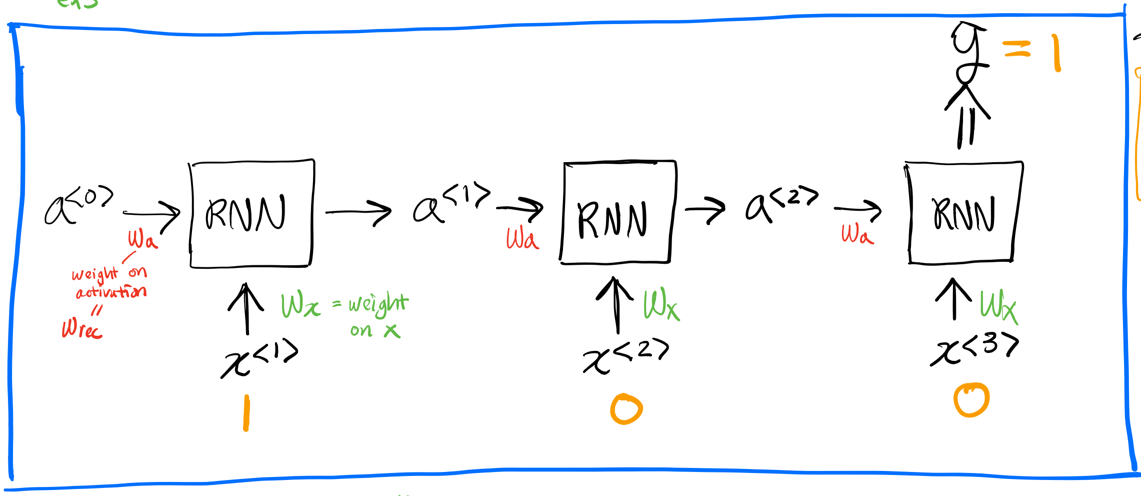
< RNN by Hand >

$$\text{Data} = \begin{array}{c|cc} & & y \\ \text{ex1} & [1\ 0\ 0] & \rightarrow [1] \\ \text{ex2} & [0\ 1\ 1] & \rightarrow [2] \\ \text{ex3} & [1\ 1\ 1] & \rightarrow [3] \end{array}$$

3 examples, length of 3

target (sum of input)

Goal: have RNN sum the input sequence as the output.



→ cost (loss) function

$$L = \frac{\sum^{i} \left(y^{[i]} - \hat{y}^{[i]}\right)^2}{n}$$

↳ Mean Squared Error.

$[i]$ denotes ith example.

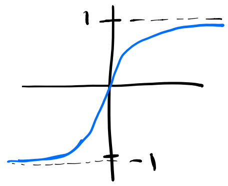
n = # of examples

others refer it as $s^{<t>}$ for state

$$\boxed{a^{<t>} = \tanh\left(W_a a^{<t-1>} + W_x x^{<t>} + b_a\right)}$$

we will ignore bias for

※ $\tanh(x) = \dfrac{\sinh(x)}{\cosh(x)}$



Step 1: initialize variables. + first forward prop.

$$\boxed{\begin{aligned} W_a &= 1 \\ W_x &= 1 \\ a^{<0>} &= 0 \end{aligned}}$$

ex1 →

$a^{<1>} = \tanh(1 \cdot 0 + 1 \cdot 1) = 0.76$

$a^{<2>} = \tanh(1 \cdot 0.76 + 1 \cdot 0) = 0.64$

$a^{<3>} = \tanh(1 \cdot 0.64 + 1 \cdot 0) = \underline{0.57} = \hat{y}^{[1]} \rightarrow y^{[1]} = 1$

ex2 →

$a^{<1>} = \tanh(1 \cdot 0 + 1 \cdot 0) = 0$

$a^{<2>} = \tanh(1 \cdot 0 + 1 \cdot 1) = 0.76$

$a^{<3>} = \tanh(1 \cdot 0.76 + 1 \cdot 1) = \underline{0.94} = \hat{y}^{[2]} \rightarrow y^{[2]} = 2$

ex3 →

$a^{<1>} = \tanh(1 \cdot 0 + 1 \cdot 1) = 0.76$

$a^{<2>} = \tanh(1 \cdot 0.76 + 1 \cdot 1) = 0.94$

$a^{<3>} = \tanh(1 \cdot 0.94 + 1 \cdot 1) = \underline{0.96} = \hat{y}^{[3]} \rightarrow y^{[3]} = 3$

1st Epoch

$$\text{Loss} = L = \frac{(1-0.57)^2 + (2-0.94)^2 + (3-0.96)^2}{3} = \boxed{1.82}$$

Step 2: Back propagation

MSE loss function for some parameters θ and function h_θ

$$L(\theta) = \frac{1}{n}\sum_i^n \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 = \frac{1}{n}\sum_i^n \left(a^{<3>(i)} - y^{(i)}\right)^2$$

To keep things simple, we simplify the activation function to be:

$$\boxed{a^{<t>} = W_a\, a^{<t-1>} + W_x\, x^{<t>}}$$

h_θ θ_0 θ_1

$$\frac{\partial}{\partial\theta_0} L(\theta_0,\theta_1) = \frac{\partial}{\partial\theta_0}\left(\frac{1}{n}\sum_i^n \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2\right)$$

$$= \frac{1}{n}\sum_i^n \frac{\partial}{\partial\theta_0}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

?

$$= \frac{1}{n}\sum_i^n 2\left(h_\theta(x^{(i)}) - y^{(i)}\right)\cdot\frac{\partial}{\partial\theta_0}\left(h_\theta(x^{(i)}) - y^{(i)}\right)$$

$$= \frac{a^{<t-1>}}{n}\sum_i^n \left(h_\theta(x^{(i)}) - y^{(i)}\right)$$

State 3 (Last State)

$$\frac{\partial L}{\partial W_x} = \frac{\partial L}{\partial a^{<3>}}\cdot\frac{\partial a^{<3>}}{\partial W_x} = \frac{2}{n}\sum_i^n \left(a^{<3>(i)} - y^{(i)}\right)\cdot x^{<3>(i)} \quad \text{— (a)}$$

$$\frac{\partial L}{\partial W_a} = \frac{\partial L}{\partial a^{<3>}}\cdot\frac{\partial a^{<3>}}{\partial W_a} = \frac{2}{n}\sum_i^n \left(a^{<3>(i)} - y^{(i)}\right)\cdot a^{<2>(i)} \quad \text{— (b)}$$

Helper

$$\frac{\partial a^{<2>}}{\partial W_x} = x^{<2>}, \quad \frac{\partial a^{<2>}}{\partial W_a} = a^{<1>}$$

$$\frac{\partial a^{<3>}}{\partial a^{<2>}} = \frac{\partial}{\partial a^{<2>}}\left(W_a a^{<2>} + W_x x^{<3>}\right) = W_a$$

$$\frac{\partial L}{\partial a^{<3>}} = \frac{2}{n}\sum_i^n \left(a^{<3>(i)} - y^{(i)}\right)$$

State 2 ($a^{<2>}$)

$$\frac{\partial L}{\partial W_x} = \frac{\partial L}{\partial a^{<3>}}\cdot\frac{\partial a^{<3>}}{\partial a^{<2>}}\cdot\frac{\partial a^{<2>}}{\partial W_x} = \frac{2}{n}\sum_i^n \left(a^{<3>(i)} - y^{(i)}\right)(W_a)\left(x^{<2>(i)}\right) \quad \text{— (c)}$$

$$\frac{\partial L}{\partial W_a} = \frac{\partial L}{\partial a^{<3>}}\cdot\frac{\partial a^{<3>}}{\partial a^{<2>}}\cdot\frac{\partial a^{<2>}}{\partial W_a} = \frac{2}{n}\sum \left(a^{<3>(i)} - y^{(i)}\right)(W_a)\left(a^{<1>(i)}\right) \quad \text{— (d)}$$

helper

$$\frac{\partial a^{<2>}}{\partial a^{<1>}} = W_a$$

$$\frac{\partial a^{<1>}}{\partial W_x} = x^{<1>}, \quad \frac{\partial a^{<1>}}{\partial W_a} = a^{<0>}$$

State 1 ($a^{<1>}$)

$$\frac{\partial L}{\partial W_x} = \frac{\partial L}{\partial a^{<3>}}\cdot\frac{\partial a^{<3>}}{\partial a^{<2>}}\cdot\frac{\partial a^{<2>}}{\partial a^{<1>}}\cdot\frac{\partial a^{<1>}}{\partial W_x} = \frac{2}{n}\sum_i^n \left(a^{<3>(i)} - y^{(i)}\right)\cdot(W_a)^2\cdot\left(x^{<1>(i)}\right) \quad \text{— (e)}$$

$$\frac{\partial L}{\partial W_a} = \quad '' \quad \cdot\frac{\partial a^{<1>}}{\partial W_a} = \frac{2}{n}\sum_i^n \left(a^{<3>(i)} - y^{(i)}\right)(W_a)^2\cdot\left(a^{<0>(i)}\right) \quad \text{— (f)}$$

* Remember that I did not use consistent activation function in forward prop & backward prop. But the flow of calculations should be same

(a) $\frac{2}{3}\left[\cancel{(0.57-1)(0)} + (0.94-2)(1) + (0.96-3)(1)\right] = -1.36$ ← probably need to multiply by learning rate since -1.36 is too big

(b) $\frac{2}{3}\left[(0.57-1)(0.64) + (0.94-2)(0.76) + (0.96-3)(0.94)\right] = \cdots$

(c) (d) (e) (f) → do the same for these equations

Step 3: Update Weights

$$W_x = W_x - \left((a) + (c) + (e)\right)$$

$$W_a = W_a - \left((b) + (d) + (f)\right)$$

Step 4: Repeat!

Repeat step 1 until the model converges to minimum loss.

⟨ Vectorization ⟩ vector implementation of above

$$x = \begin{array}{cl} \left[\begin{array}{c} [\,1\,,\,0\,,\,0\,] \\ [\,0\,,\,1\,,\,1\,] \\ [\,1\,,\,1\,,\,1\,] \end{array}\right] & \begin{array}{l} \leftarrow \vec{x}^{(1)} \\ \leftarrow \vec{x}^{(2)} \\ \leftarrow \vec{x}^{(3)} \end{array} \\ \quad \vec{x}_1 \quad \vec{x}_2 \quad \vec{x}_3 & \end{array} \qquad \vec{y} = [1,2,3]$$

$$\begin{tikzcd} \vec{a}_0 \arrow[r] & \vec{a}_1 \arrow[r] & \vec{a}_2 \arrow[r] & \vec{a}_3 \\ & \vec{x}_1 \arrow[u] & \vec{x}_2 \arrow[u] & \vec{x}_3 \arrow[u] \\ & \begin{bmatrix}1\\0\\1\end{bmatrix} & \begin{bmatrix}0\\1\\1\end{bmatrix} & \begin{bmatrix}0\\1\\1\end{bmatrix} \end{tikzcd}$$

where $\vec{a}_0 = \begin{bmatrix}0\\0\\0\end{bmatrix}$

$$\left[\begin{array}{l} \vec{a}_1 = \vec{a}_0 \cdot W_a + \vec{x}_1 \cdot W_x = \begin{bmatrix}0\\0\\0\end{bmatrix} W_a + \begin{bmatrix}1\\0\\1\end{bmatrix} W_x = \begin{bmatrix}W_x\\0\\W_x\end{bmatrix} \\ \vec{a}_2 = \vec{a}_1 \cdot W_a + \vec{x}_2 \cdot W_x = \begin{bmatrix}W_x\\0\\W_x\end{bmatrix} W_a + \begin{bmatrix}0\\1\\1\end{bmatrix} W_x = \begin{bmatrix}W_x W_a + 0\\0 + W_x\\W_x W_a + W_x\end{bmatrix} = \begin{bmatrix}W_x W_a\\W_x\\W_x W_a + W_x\end{bmatrix} \\ \vec{a}_3 = \vec{a}_2 W_a + \vec{x}_3 \cdot W_x = \cdots \end{array}\right.$$

$$L = \frac{1}{n}(\vec{a}_3 - \vec{y})^2 = \frac{1}{n}\left(\begin{bmatrix}a_{31}\\a_{32}\\a_{33}\end{bmatrix} - \begin{bmatrix}1\\2\\3\end{bmatrix}\right)^2 = \boxed{}$$