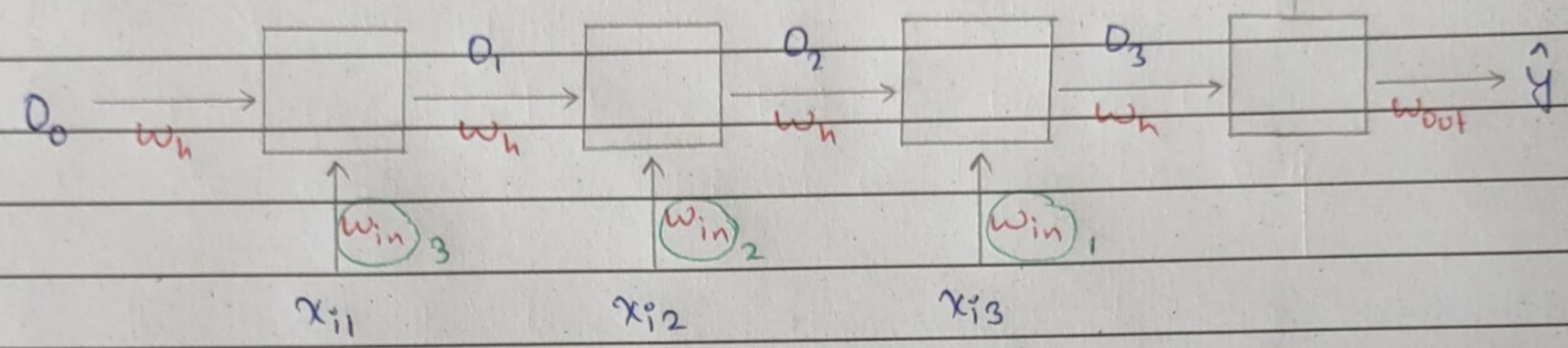
RAISONI GROUP
— a vision beyond —

PROBLEMS WITH RNN



Problem-1: Long Term Dependency Problem

$$w_{in} = w_{in} - \eta \frac{\partial L}{\partial w_{in}}$$

$$w_{out} = w_{out} - \eta \frac{\partial L}{\partial w_{out}}$$

$$w_h = w_h - \eta \frac{\partial L}{\partial w_h}$$

$$\frac{\partial L}{\partial w_{in}} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial O_3} \times \frac{\partial O_3}{\partial (w_{in})_1}$$ → Short term dependency

$$+ \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial O_3} \times \frac{\partial O_3}{\partial O_2} \times \frac{\partial O_2}{\partial (w_{in})_2}$$

$$+ \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial O_3} \times \frac{\partial O_3}{\partial O_2} \times \frac{\partial O_2}{\partial O_1} \times \frac{\partial O_1}{\partial (w_{in})_3}$$ → Long term dependency

This was for only 3 time steps. Now assume there are 100 time steps.

$$\Rightarrow \frac{\partial L}{\partial w_{in}} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial O_{100}} \times \cdots \times \frac{\partial O_1}{\partial w_{in}}$$ → Long term dependency for 100 time steps

$$\Rightarrow \frac{\partial L}{\partial w_{in}} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial O_{100}} \times \prod_{t=100}^{2} \left(\frac{\partial O_t}{\partial O_{t-1}} \right) \times \frac{\partial O_1}{\partial w_{in}}$$

$$O_1 = \tanh(x_{i1} w_{in} + O_0 w_h)$$

$$\Rightarrow O_t = \tanh(x_{it} w_{in} + O_{t-1} w_h)$$

$$\Rightarrow \frac{\partial O_t}{\partial O_{t-1}} = \tanh'(x_{it} w_{in} + O_{t-1} w_h) \cdot w_h$$

Derivative of tanh is always b/w 0 - 1

$$\Rightarrow \frac{\partial L}{\partial w_{in}} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial O_{100}} \times \prod_{t=100}^{2} \left(\tanh'(x_t w_{in} + O_{t-1} w_h) \cdot w_h \right) \times \frac{\partial O_1}{\partial w_{in}}$$

Derivative of tanh, i.e. tanh' is b/w 0-1
Now also assume value of w_h to be b/w 0-1

In this case the long term dependency $\frac{\partial L}{\partial w_{in}}$ will become a vanishing gradient & will not contribute & short term dependencies will have more responsibility in updating the parameters during backpropagation.

Problem-2 : Unstable Training Problem
This usually occurs because of exploding gradient problem. Suppose you use ReLU instead of tanh in your RNN & initialized your ~~weight~~ recurrent weight w_h with 1, in that case your long term dependency with explode & dominate other dependencies during backpropagation due to which unstable ~~gradient~~ training may happen. Similar thing can happen if you have a large value learning rate η.