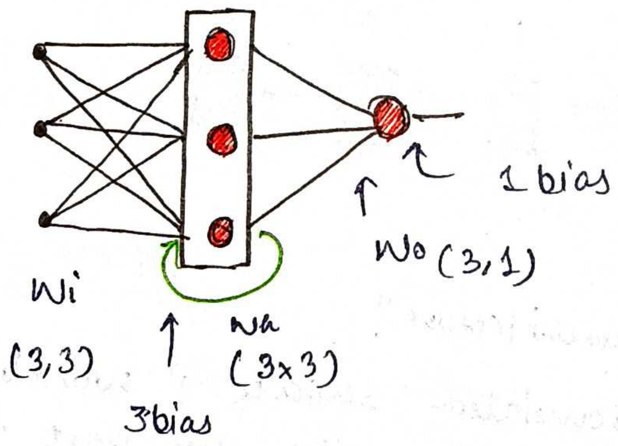
Backpropagation in RNN

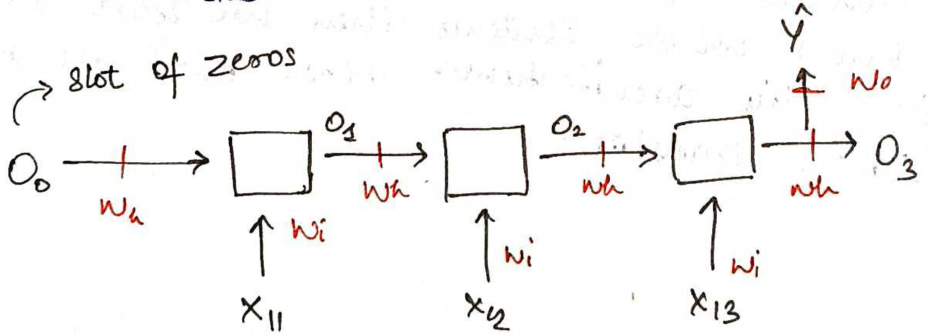
We gonna to take an example → Many to one RNN

text				X				Y
cat	mat	rat	1	x_1	[1 0 0]	[0 1 0]	[0 0 1]	1
rat	rat	mat	1 →	x_2	[0 0 1]	[0 0 1]	[0 1 0]	1
mat	mat	cat	0	x_3	[0 1 0]	[0 1 0]	[1 0 0]	0

Vocab → cat Mat Rat

[1 0 0] [0 1 0] [0 0 1]





$O_1 = f(x_{11} W_i + O_0 W_h)$

$O_2 = f(x_{12} W_i + O_1 W_h)$

$O_3 = f(x_{13} W_i + O_2 W_h)$

$\hat{Y} = \sigma(O_3 W_o)$

Loss $= -y_i \log \hat{y}_i - (1-y_i) \log(1-\hat{y}_i)$

Loss Calculate → minimize

using gradient descent ↵

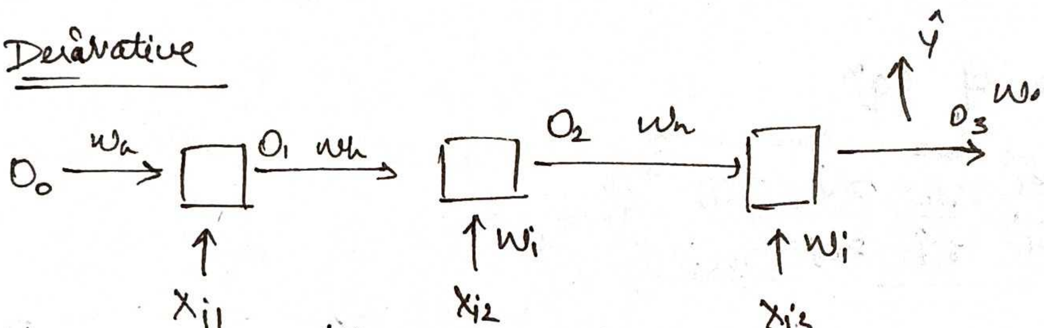
Find the value of Wi Wh Wo where Loss should be minimum.

$$W_i = W_i - \eta \boxed{\frac{\partial L}{\partial W_i}} \qquad W_o = W_o - \eta \boxed{\frac{\partial L}{\partial W_o}}$$

$$W_h = W_h - \eta \boxed{\frac{\partial L}{\partial W_h}}$$

Now, we have to find the derivative

Derivative



(i) $\boxed{\frac{\partial L}{\partial W_o}}$, (ii) $\boxed{\frac{\partial L}{\partial W_i}}$, (iii) $\boxed{\frac{\partial L}{\partial W_h}}$

$$\boxed{\begin{aligned} O_1 &= f(x_{i1} W_i + O_0 W_h) \\ O_2 &= f(x_{i2} W_i + O_0 W_h) \\ O_3 &= f(x_{i3} W_i + O_0 W_h) \\ \hat{y} &= \sigma(O_3 W_o) \end{aligned}}$$

$\frac{\partial L}{\partial W_o}$ $L \rightarrow \hat{y} \rightarrow O_3, W_o$

(i) $\boxed{\frac{\partial L}{\partial W_o} = \boxed{\frac{\partial L}{\partial \hat{y}}} \boxed{\frac{\partial \hat{y}}{\partial W_o}}}$ → $L = y_i \log \hat{y}_i - (1-y_i)\log(1-\hat{y})$

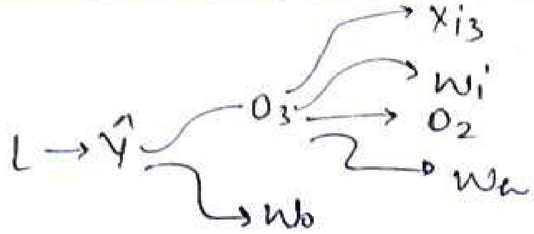
differentiation

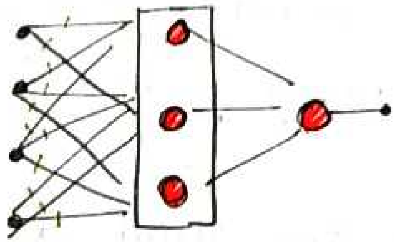
$\hat{y} = \sigma(O_3 W_o)$

(ii) $\frac{\partial L}{\partial W_i}$ → ∂W_i ⇒ change

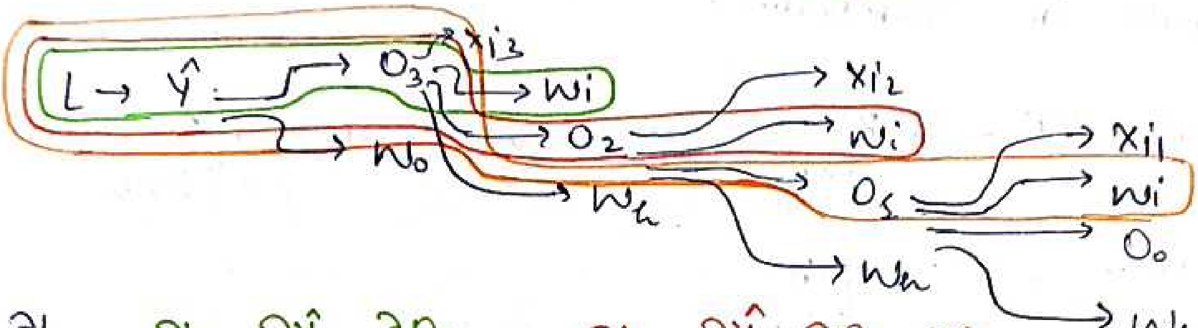
L ⇒ how much L change when Wi change

$\frac{\partial L}{\partial w_i} \rightarrow$





differention



$$\frac{\partial L}{\partial w_i} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial O_3} \frac{\partial O_3}{\partial w_i} + \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial O_3} \frac{\partial O_3}{\partial O_2} \frac{\partial O_2}{\partial w_i} + \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial O_3} \frac{\partial O_3}{\partial O_2} \frac{\partial O_2}{\partial O_1} \frac{\partial O_1}{\partial w_i}$$

Summarization of eq^n

bcz we have 3 words

$$\boxed{\frac{\partial L}{\partial w_i} = \sum_{j=1}^{t=3} \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial O_j} \frac{\partial O_j}{\partial w_i}}$$

Proof

expand j=1 $\Rightarrow \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial O_1} \frac{\partial O_1}{\partial w_i} \Rightarrow \boxed{\frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial O_3} \frac{\partial O_3}{\partial O_2} \frac{\partial O_2}{\partial O_1} \frac{\partial O_1}{\partial w_i}}$

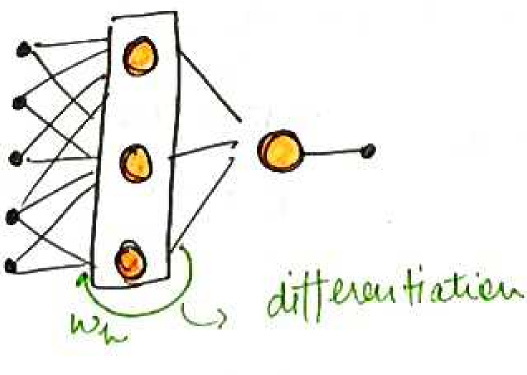
Same

Same

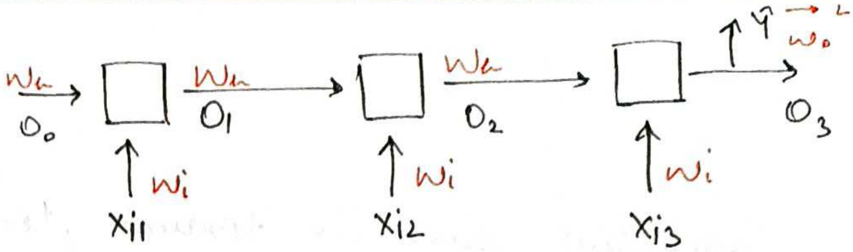
expand j=2 $\Rightarrow \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial O_2} \frac{\partial O_2}{\partial w_i} \Rightarrow \boxed{\frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial O_3} \frac{\partial O_3}{\partial O_2} \frac{\partial O_2}{\partial w_i}}$

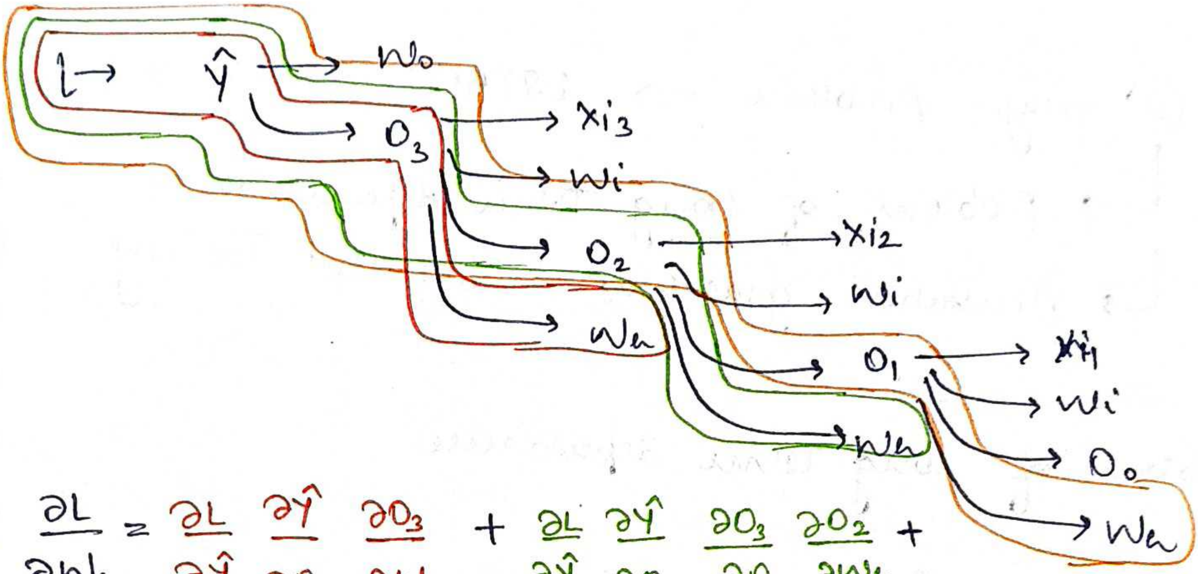
expand j=3 $\Rightarrow \boxed{\frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial O_3} \frac{\partial O_3}{\partial w_i}}$

(iii) $\frac{\partial L}{\partial w_h}$



$$\boxed{\frac{\partial L}{\partial W_h}}$$





$$\frac{\partial L}{\partial W_h} = \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial O_3}\frac{\partial O_3}{\partial W_h} + \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial O_3}\frac{\partial O_3}{\partial O_2}\frac{\partial O_2}{\partial W_h} +$$

$$\frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial O_3}\frac{\partial O_3}{\partial O_2}\frac{\partial O_2}{\partial O_1}\frac{\partial O_1}{\partial W_h}$$

Summarize

$$\boxed{\frac{\partial L}{\partial W_h} = \sum_{j=1}^{n} \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial O_j}\frac{\partial O_j}{\partial W_h}}$$

n = timesteps

Problem with RNN

RNN → Sequential data → textual, time, series

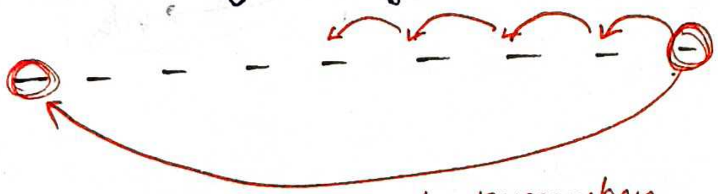
↓

Suffer ② major problem → LSTM

↳ problem of long term dependence

↳ unstable gradients / Stagnated Training

(1) Problem of long term dependence



* short term memory loss.

This data not remember starting word (just like gajani)

eg:- Marathi is spoken in Maharashtra.

remember

Maharashtra is a beautiful place. Went there last year but I could not enjoy properly because I don't understand marathi

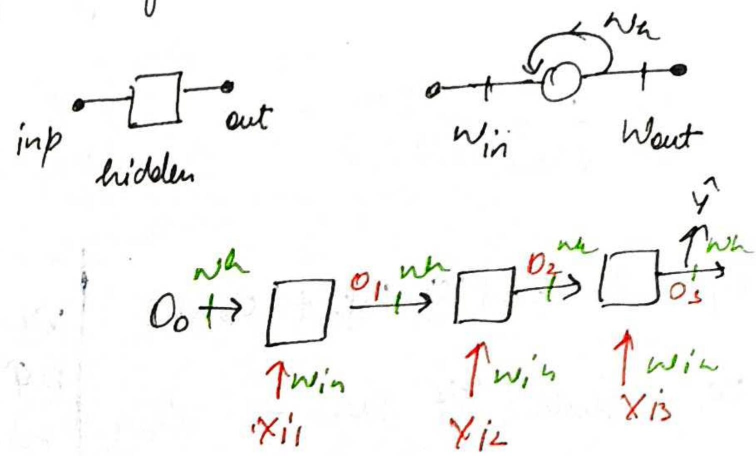
Not remember.

In Auto-suggestion (keypad) → Suggest short sentence but not long sentence

* 2nd sentence RNN not remember "Maharashtra" word. So, model not predict "Marathi" word at last. (Vanishing gradient problem)

Problem # 1 Problem of long term dependency → vanishing

Input			out
$\boxed{1}$	$\boxed{0}$	$\boxed{1}$	1
0	0	1	0
0	0	0	0
1	1	1	1



3 time stamps

Loss → minimum → gradient descent

short term dependency

$$\frac{\partial L}{\partial W_{in}} = \boxed{\frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial O_3} \frac{\partial O_3}{\partial W_{in}}} +$$

$$\frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial O_3} \frac{\partial O_3}{\partial O_2} \frac{\partial O_2}{\partial W_{in}} +$$

$$\boxed{\frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial O_3} \frac{\partial O_3}{\partial O_2} \frac{\partial O_2}{\partial O_1} \frac{\partial O_1}{\partial W_{in}}} \rightarrow \text{long term dependency}$$

$$W_{in} = W_{in} - \eta \boxed{\frac{\partial L}{\partial W_{in}}}$$

$$W_{out} = W_{out} - \eta \boxed{\frac{\partial L}{\partial W_{out}}}$$

$$W_h = W_h - \eta \boxed{\frac{\partial L}{\partial W_h}}$$

This is only for 3 timesteps

100 timesteps
long term dependency

$$\frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial O_{100}} \frac{\partial O_{100}}{\partial O_{99}} \text{-------} \frac{\partial O_2}{\partial O_1} \frac{\partial O_1}{\partial W_{in}}$$

During ~~the~~ calculating gradient descent, → value of long term dependency will be small

value of short term dependency will be large.

So, short term dependency contribute more as compare to long term memory to find gradient descent.

time stamps ↑↑ ⇔ long term dependency ↓↓

$$\frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial O_{100}} \boxed{\frac{\partial O_{100}}{\partial O_{99}} \cdots \frac{\partial O_2}{\partial O_1}} \frac{\partial O_1}{\partial w_{in}}$$

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

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

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

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

$$\frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial O_{100}} \prod_{t=2}^{100} (\underbrace{\tanh'()}_{\text{bet}^n\ 0-1}\ \underbrace{w_h}_{\text{bet}^n\ 0-1}) \frac{\partial O_1}{\partial w_{in}} \rightarrow \text{Very close to zero.}$$

Sol → Different activation function → relu / leaky (not bet^n 0-1)

→ better weight initialization

→ Skip RNNs (try) → self study topic

→ LSTM

#2 Problem → Unstable Training (Exploding problem)

Long term dependency → Very large
↳ ≃ infinite

* If Long term dep. is ≃ infinite then Gradient update will be infinite then weights also infinite And Model not train.

for example → (i) Using relu and Weight initialize is 1. So, weights multiply and get large number. Long Term dep → dominate and short term dep. not contribute in gradient update.

(ii) If Learning Rate is ↑↑ (very higher).

<u>Solution</u>. 1.) Gradient Clipping → Search (self study)

2) Controlled Learning Rate

3) LSTM use.

LSTM (Long Short Term Memory)

The what

<u>LSTM core idea</u>

Now we have to decide this story is good / Bad, decision depend on geo-Area. Hero of the story.

Story