References:-

why RNN? - https://www.youtube.com/watch?v=4KpRP-YUw6c

RNN architecture and forward propogation: https://www.youtube.com/watch?v=BjWqCcbusMM&list=RDCMUCCWi3hpnq_Pe03nGxuS7isg&start_radio=1&rv=BjWqCcbusMM&t=958

RNN backpropogation: https://www.youtube.com/watch?v=OvCz1acvt-k&list=RDCMUCCWi3hpnq_Pe03nGxuS7isg&index=6

why do we need RNN?

RNN → Recurrent Neural Network

↳ is type of sequential model to work on sequential data, i.e, when sequence of data matters.

for example → text, speech

Let's say we have a text classification problem, so given a sentence we should classify it. And we know that length of sentences vary, but ANN (Artificial Neural Networks) takes a fixed length vector as input. You might solve this problem by 0 padding all sentences to make the length of all sentences same as the sentence with maximum length.

But this is not practicle because of

↳ unnecessary computation

↳ Prediction problems

→ Sequence of words is being ignored.

Data for RNN

↓

(timestamps, input-features)

example: movie review → sentiment

| Review X | Sentiment |
|---|---|
| $x_1$: $x_{11}$ movie $x_{12}$ was $x_{13}$ good | 1 |
| $x_2$: $x_{21}$ movie $x_{22}$ was $x_{23}$ bad | 0 |
| $x_3$: $x_{31}$ movie $x_{32}$ was $x_{33}$ not $x_{34}$ good | 0 |

vocab → 5 words

movie was

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

review 1 → movie was good

↓

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

↓

(3, 5) ← # of features

↑ no. of time stamps

# Architecture

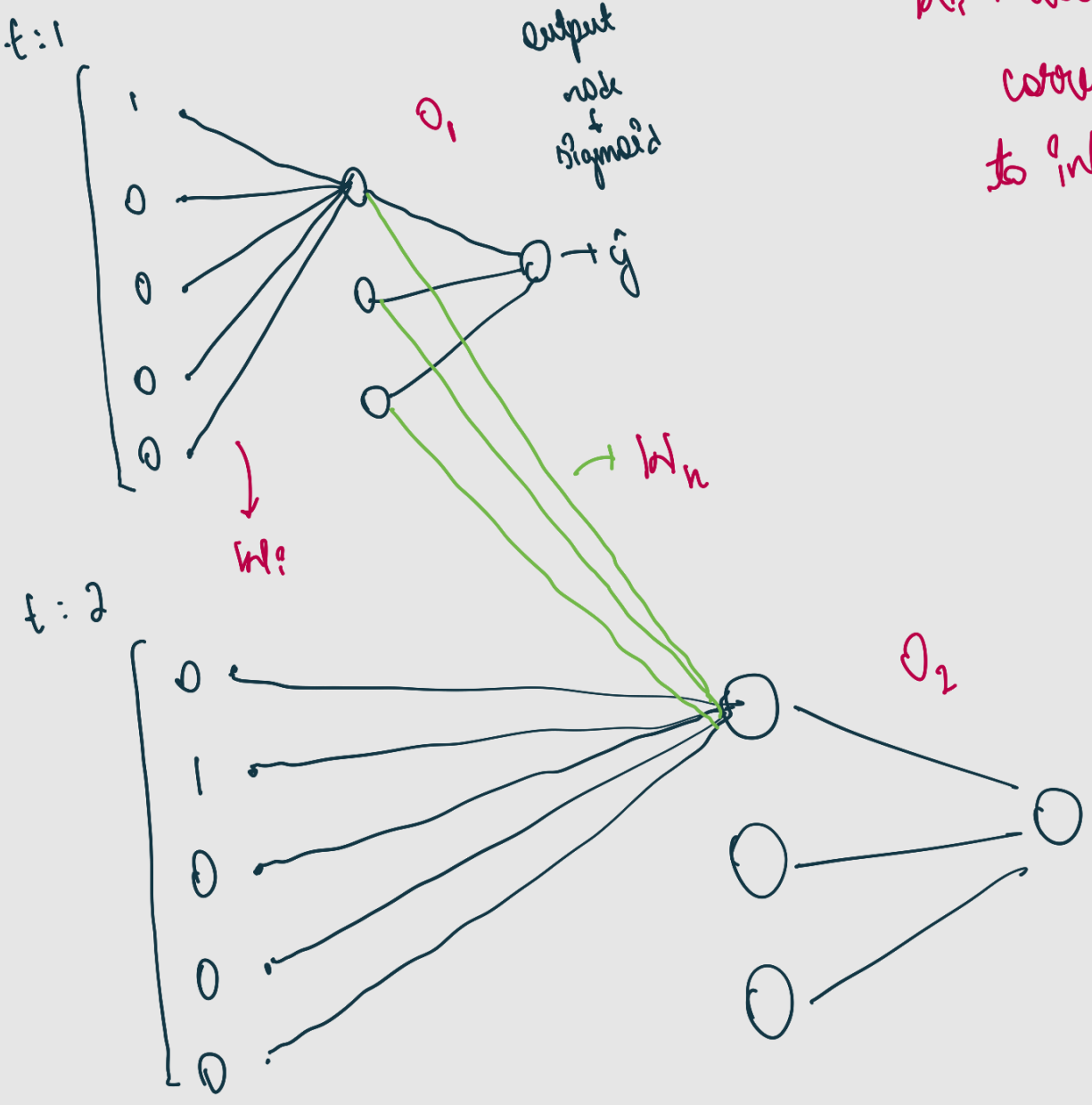
* Data is not given as input in single timestamp

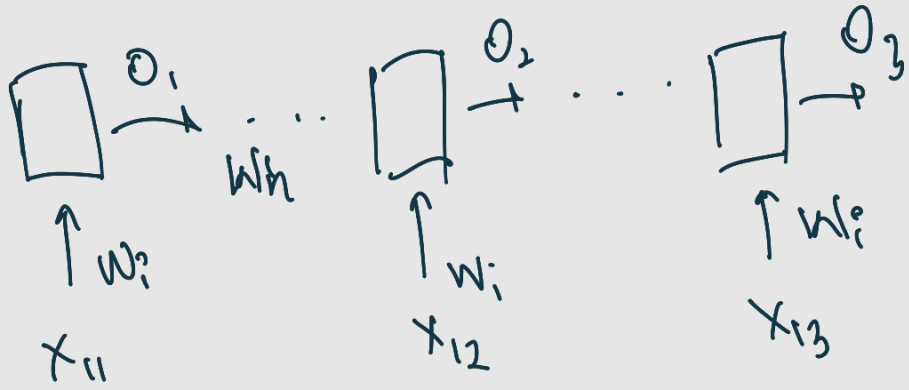

$x_{11}$ at $t:1$

$x_{12}$ at $t:2$

$x_{13}$ at $t:3$

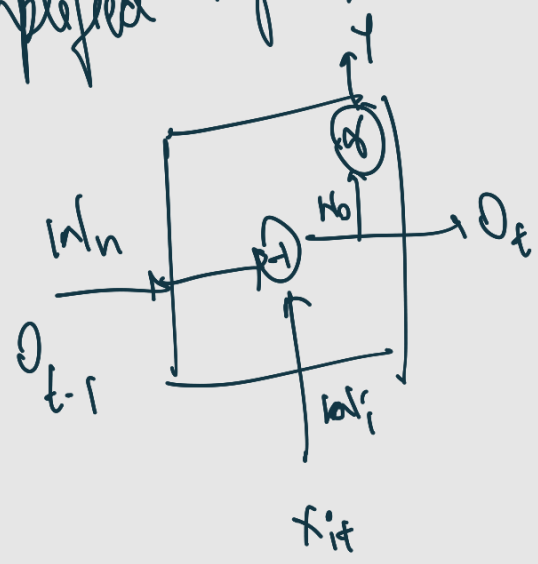
$W_i$ + weights corresponding to input feature






$O_1 : f(x_{11} W_i)$ $O_2 : f(x_{12} W_i + O_1 W_h)$

Simplified Representation:



# Backpropagation in RNN

(* Single Hidden Layer *)

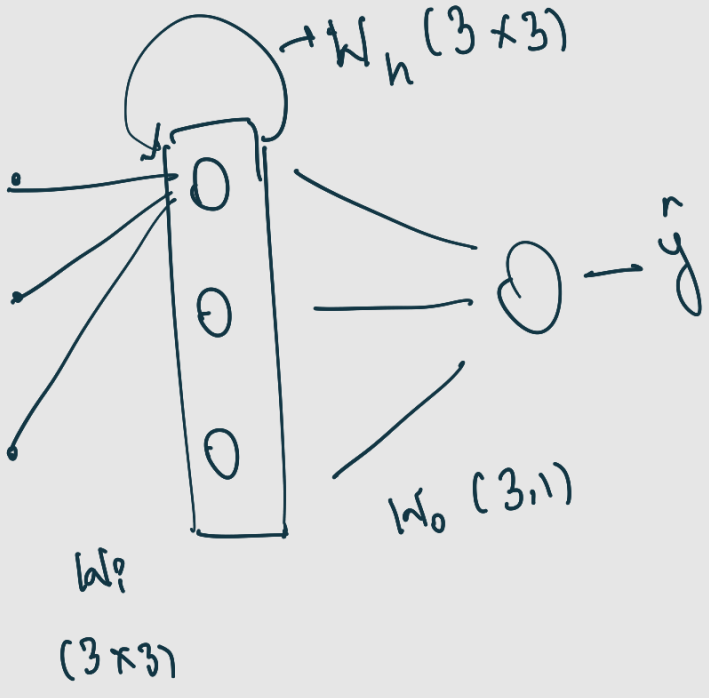
Many to one

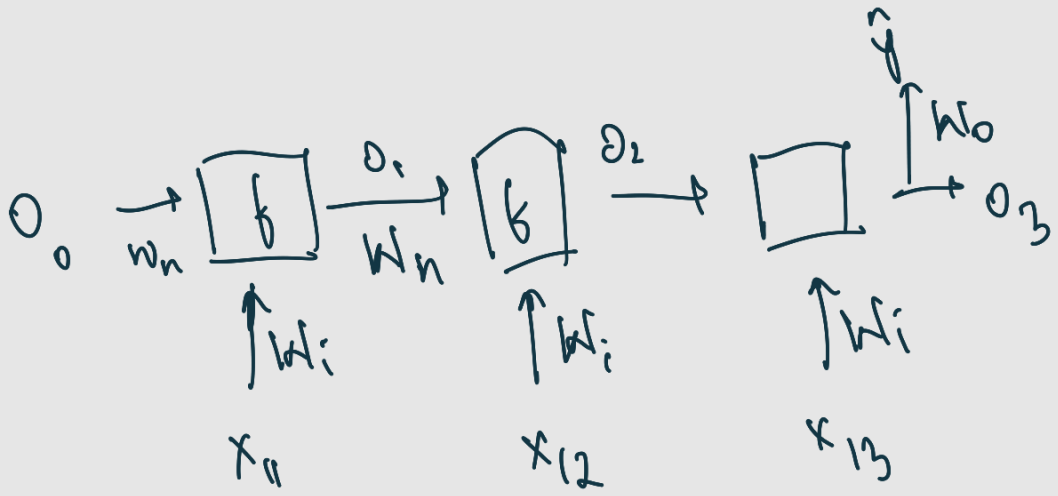
let's take a toy dataset

| text | sentiment |
|---|---|
| cat mat rat | 1 |
| rat rat mat | 1 |
| mat mat cat | 0 |

vocab → cat mat rat

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






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

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

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

$$y - \hat{y} : \sigma(O_3 W_o)$$

$$L : -y_i \log \hat{y}_i - (1-y_i) \log(1-y_i)$$

learnable parameters

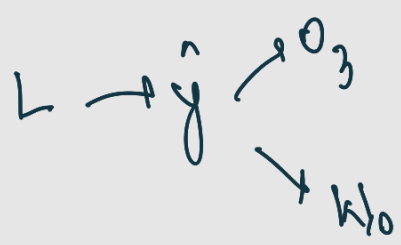
↳ $W_i$ $W_h$ $W_o$

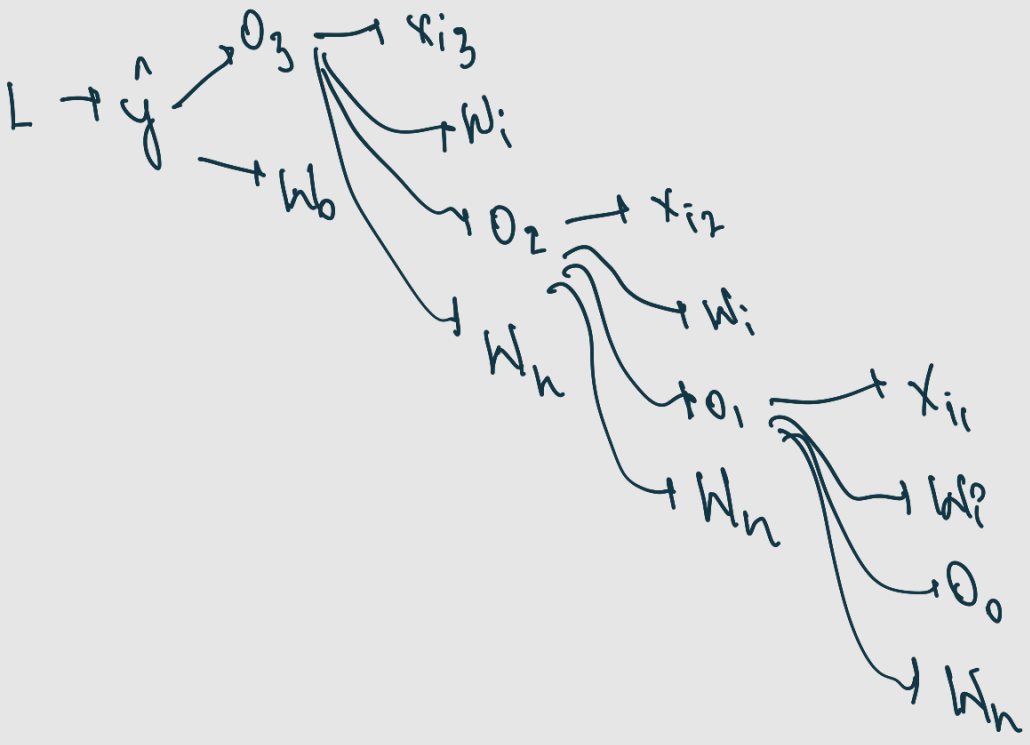
$$W_i : W_i - \eta \frac{\partial L}{\partial W_i}$$

$$W_h : W_h - \eta \frac{\partial L}{\partial W_h}$$

$$W_o : W_o - \eta \frac{\partial L}{\partial W_o}$$

$$\frac{\partial L}{\partial W_0} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial W_0}$$




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

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

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

$$\rightarrow \quad \frac{\partial L}{\partial W_i} = \sum \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial O_j} \frac{\partial O_j}{\partial W_i}$$

Now for

$$\frac{\partial L}{\partial W_h} = \sum \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial O_i} \frac{\partial O_i}{\partial W_h}$$