Youtube: Programming Cradle

✱ Recurrent Neural Networks (RNN)

✱ Why RNN?

MLP $x \rightarrow$ [MLP] $\rightarrow y$
(vector)

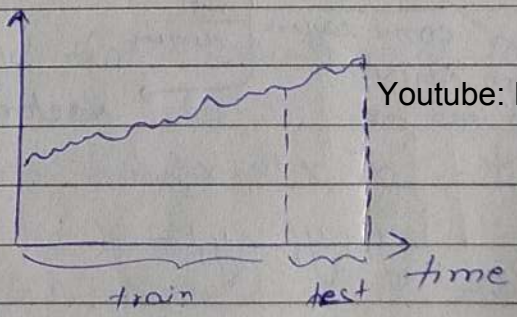
CNN $x \rightarrow$ [CNN] $\rightarrow y$
(img)
conv layer, fully connected layer, other layer like maxpool.

✱ x is a sequence we use RNN.
Sequence like amazon reviews (sequence of words)

NOTE:- When we use BoW, TfIdf, w2v we don't care/ completely discard the sequence of the words in a sentence. But sequence/ orientation of the words are important.

✱ Time-Series data

→ windowed data
→ fourier transform to decompose in constituent frequencies.



Youtube: Programming Cradle

✱ Language translation

English to Italian

✱ Speech recognition

↳ audio input → text

✱ Input image → output caption

Youtube: Programming Cradle

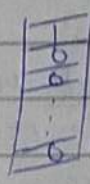
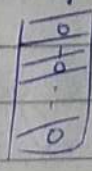
* Core Idea :- If Output depends on the Seq. or arrangment of the inputs we need new types of NN which retains and leverages the seq. info.

* If we try to train MLP for amazon review polarity.

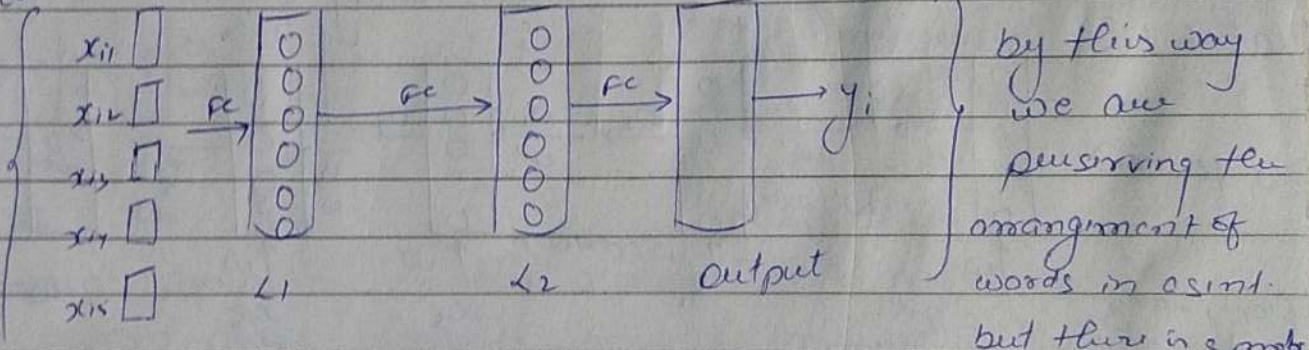
$x_1 \rightarrow x_i \rightarrow$ the phone is very fast

(if 10K unique words (BOW)) x_{i1} x_{i2} x_{i3} x_{i4} x_{i5}

$x_{ij} \in \mathbb{R}^{10K}$ → 5 vectors of 10K length

One-hot x_{i1} =  x_{i2} = 

FC → fully connected



by this way we are preserving the arrangment of words in a sent. but there is a prob

⇒ Let there are 2 reviews

5 words x_1 → This phone is very fast → 50K

4 words x_2 → This phone is good. → 40K

Sentences can be of diff lengths hence input ~~vector~~ size changes

10 words x_i → 200K

→ one soln to above prob can be we can fix the input size to ~~max~~ length of max length string let say 50 so for ~~word~~ a sent. with 5 word will have input filled till 50K and rest 450K will be zero vectors (zero padding)

→ Now what if in test data we have sent. of length 1000 words

→ Also we are getting huge vectors.

So these are the problems of using MLP for text data

Youtube: Programming Cradle

Youtube: Programming Cradle

⊛ Recurrent Neural Network (RNN)

↳ Repeating.

eg> Amazon food review → binary classf^n {0,1} {size of vocab is d'}

$D = \{x_i, y_i\}$ ↳ binary

Sentences (Seq. of words)

words: one hot encoded $x_{ij} \in R^d$ ↳ binary d' vector

$y_i \leftarrow x_i \rightarrow (x_{i1}), x_{i2}, x_{i3}, x_{i4}, x_{i5}$

$x_1 \rightarrow x_{11}, x_{12}, x_{13}$

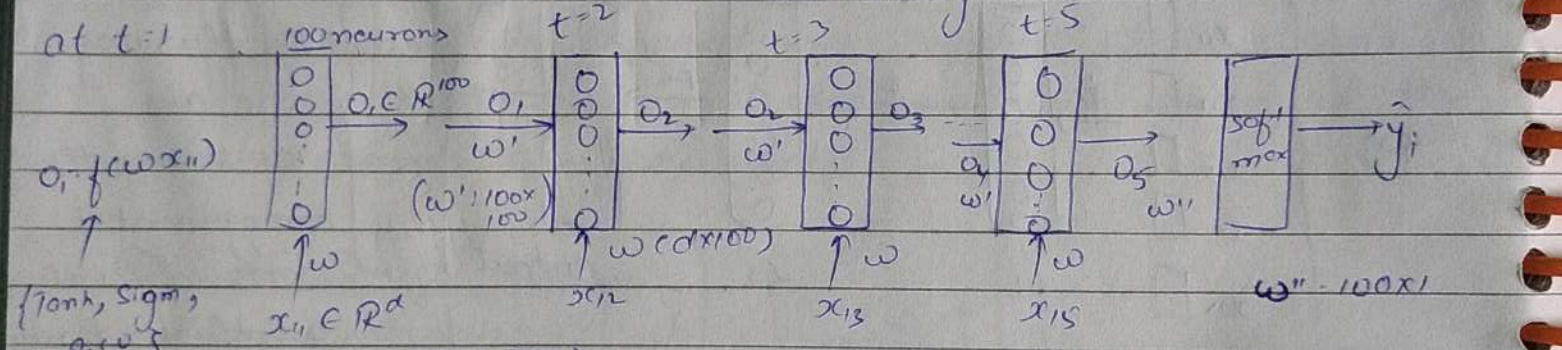
$x_i \rightarrow x_{i1}, x_{i2}, x_{i3}, \ldots x_{im}$

m_i : length of x_i

Youtube: Programming Cradle

Task: Seq. of words given; predict the polarity.

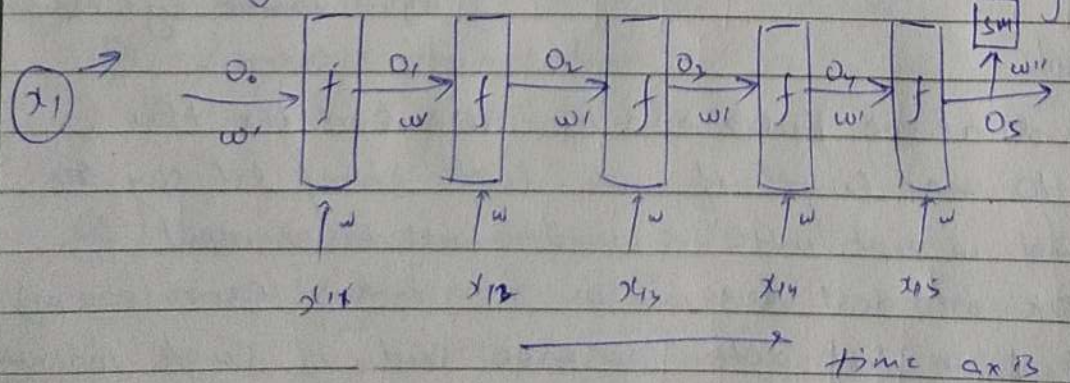
$x_i \rightarrow \{x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}\}$ y_i {0,1} ↳ binary



$O_2 = f(\omega x_{12} + \omega' O_1)$

$O_5 = f(\omega x_{15} + O_4 \times \omega')$

⊛ Above diagram Simplified below



⇒ forward prop (direct^n of arrow)

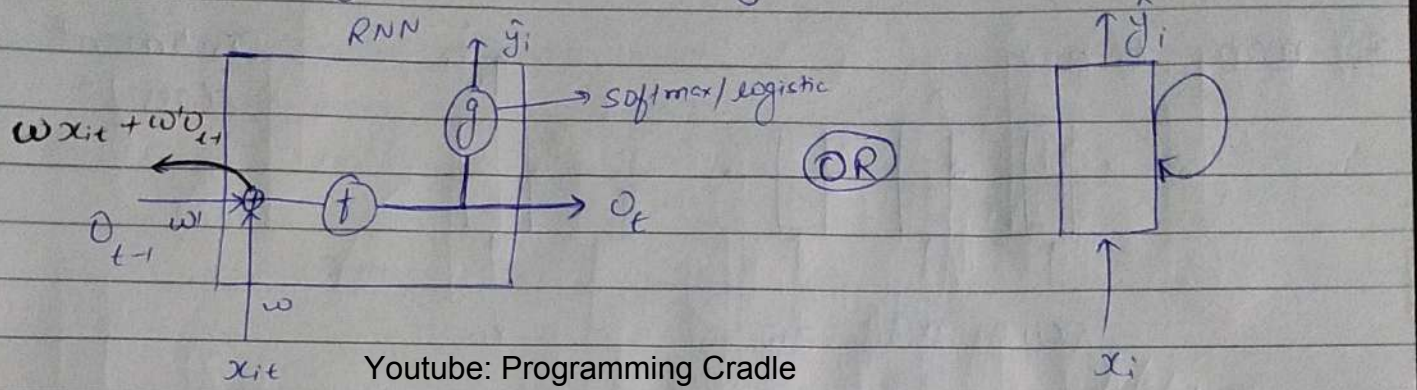
⇒ back prop (opposite dir. of the arrow)

O_0 : Zeros / random numbers

{ $\omega, \omega', \omega''$: init → Xavier / Xav }

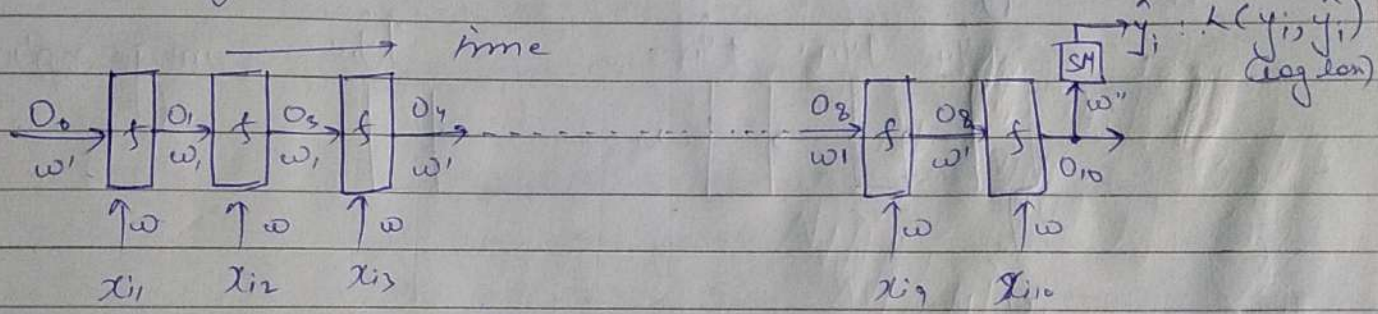
Youtube: Programming Cradle

✱ Other way of representing RNN (Box representn)



Youtube: Programming Cradle

✱ Training RNN : Backprop



- forward prop :- In the direct^n of arrows.
- backword prop :- Opposite to the arrows.

$$\frac{\partial L}{\partial o_{10}} = \frac{\partial L}{\partial \hat{y}_i} \cdot \frac{\partial \hat{y}_i}{\partial o_{10}} \quad ; \quad \frac{\partial L}{\partial w''} = \frac{\partial L}{\partial \hat{y}_i} \cdot \frac{\partial \hat{y}_i}{\partial w''} \quad ; \quad \frac{\partial L}{\partial w} = \frac{\partial L}{\partial \hat{y}_i} \cdot \frac{\partial \hat{y}_i}{\partial o_{10}} \cdot \frac{\partial o_{10}}{\partial w}$$

✱ Problm with this backprop

1st word in a sent : $\frac{\partial L}{\partial w}$: lots of multiplicat^n of partial derivatives

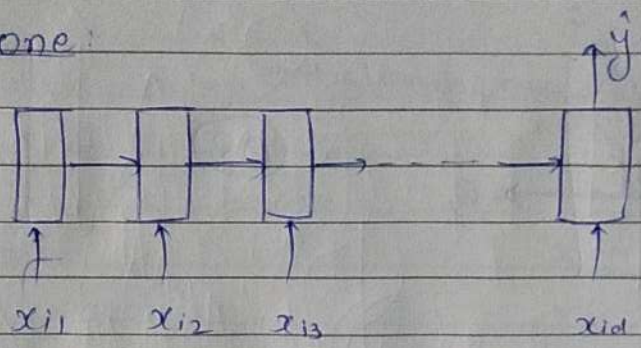
NOTE we are getting this error ↳ all are less than 1 hence not beoz of many layers but vanishing grad. problem.
beoz of repetition of the task ↳ possibility of exploding grad (performing forward and back prop over time) when derivat are more than 1

- Also SimpleRNN have short term memory.

Youtube: Programming Cradle

⊛ Types of RNN :-

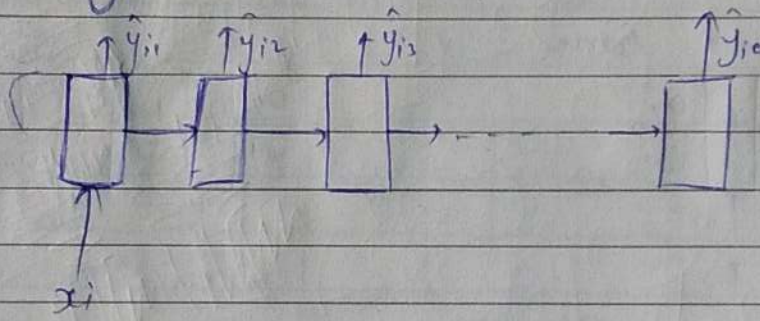
⊛ many to one :



∴ Sentiment classifⁿ

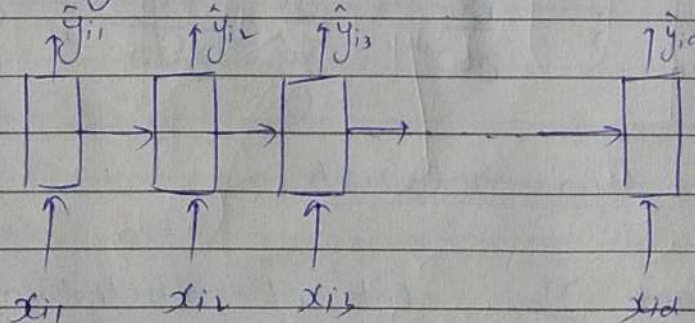
⊛ One to many :

Youtube: Programming Cradle



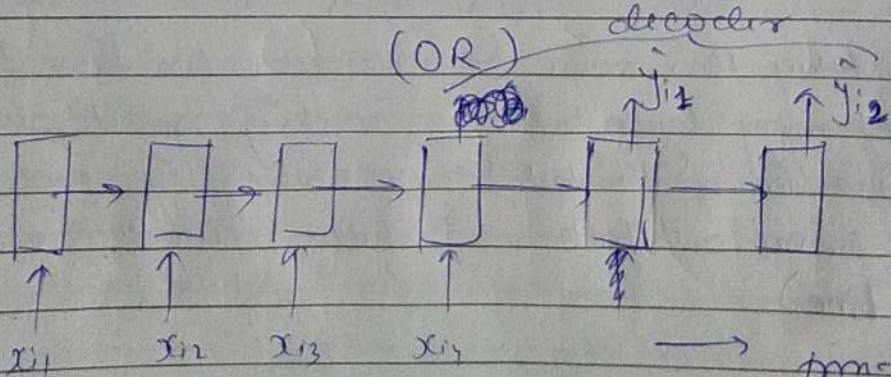
∴ image captioning

⊛ Many to Many :-



∴ part of speech

(OR)



this is also known as encoder decoder

encoder

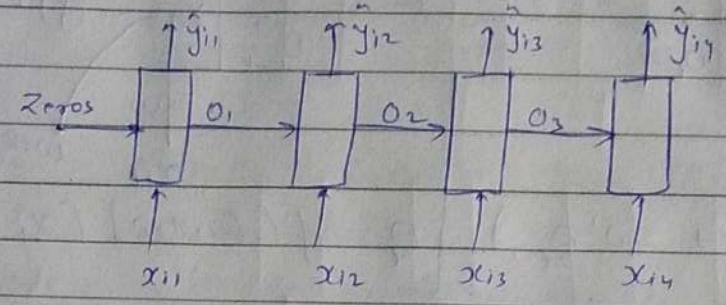
∴ language translation

Youtube: Programming Cradle

⊛ Need for LSTM / GRU

⊛ Problem with Simple RNN

→ many-many (same length)



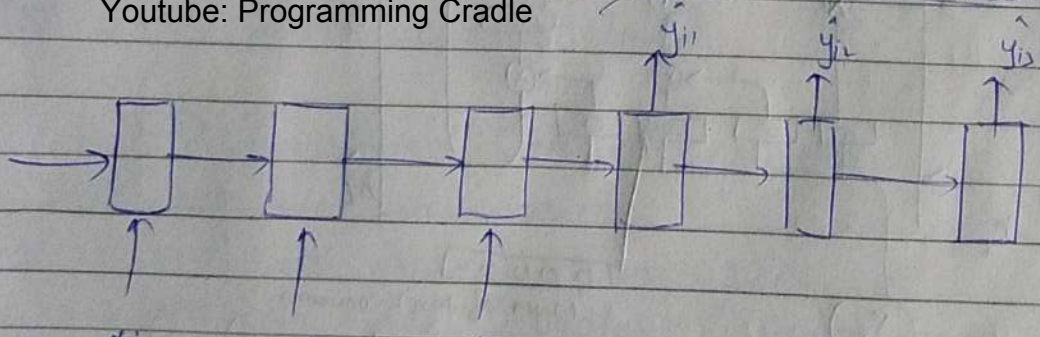
$\hat{y}_{i4}$ → depends a lot on x_{i4} and O_3 } we can think of it as short term memory.

$\hat{y}_{i4}$ → depends less on x_{i1} and O_1 }

∴ Because of back prop/~~forward prop~~, contribution of x_{i1} and O_1 will reduce, they are far from $\hat{y}_{i4}$

→ Simple RNN can not take care of long term ~~of dependency~~

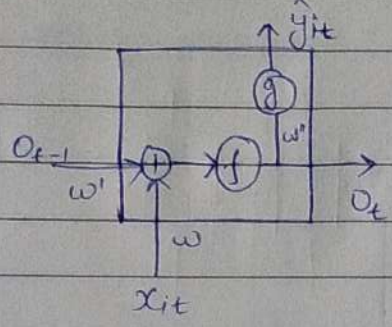
Youtube: Programming Cradle



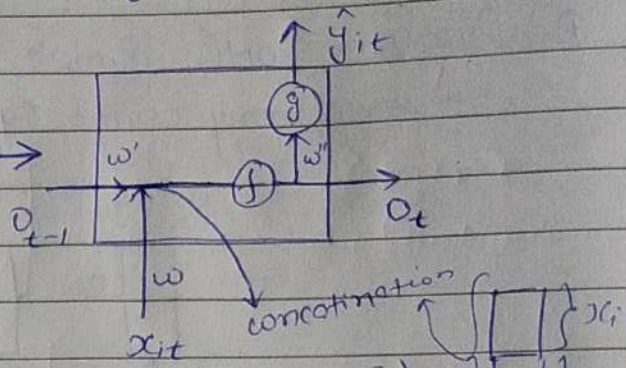
encoder (English)

→ in encoder-decoder problem can be even more as the distance b/w $\hat{y}_{i3}$ and x_{i1} ↑ even more

⊛ LSTM Long Short term Memory RNN



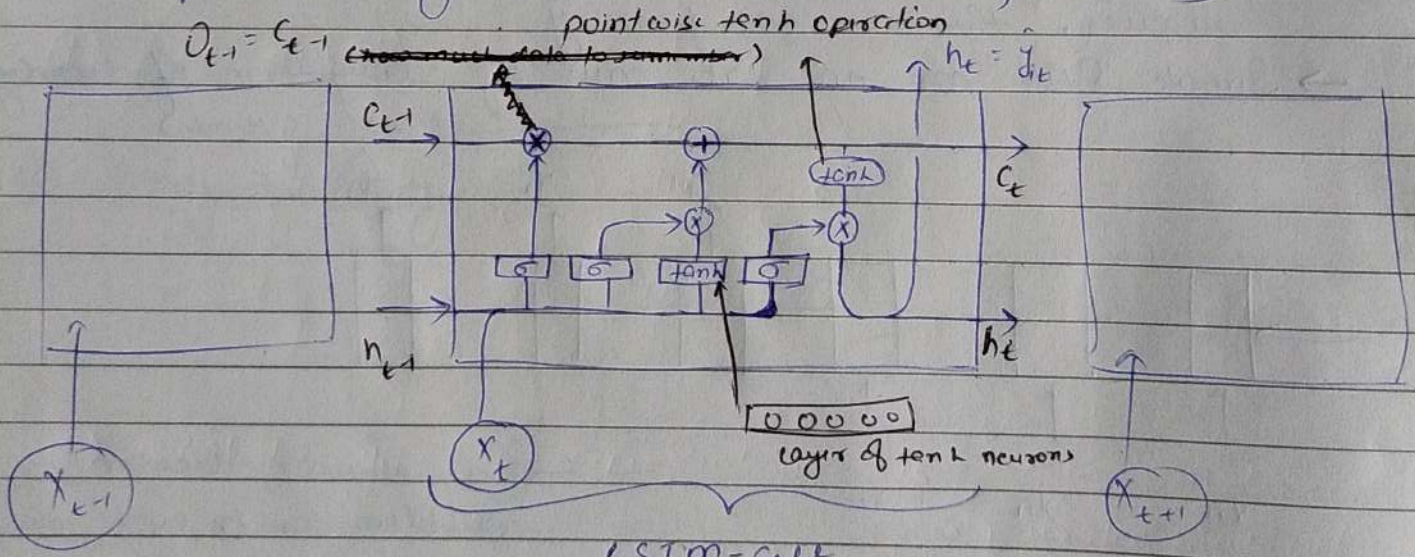
modified →

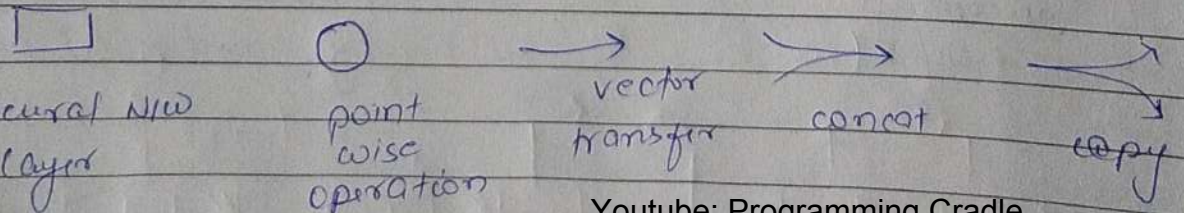


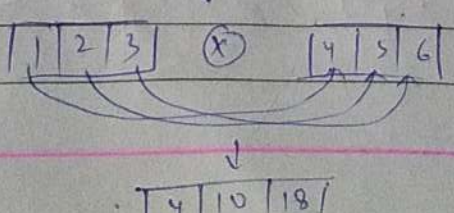
$O_t = f(wx_{it} + w'O_{t-1})$ | $O_t = f([w', w][x_t, O_t])$

$\hat{y}_{it} = g(w''O_t)$ | $\hat{y}_{it} = g(w''O_t)$

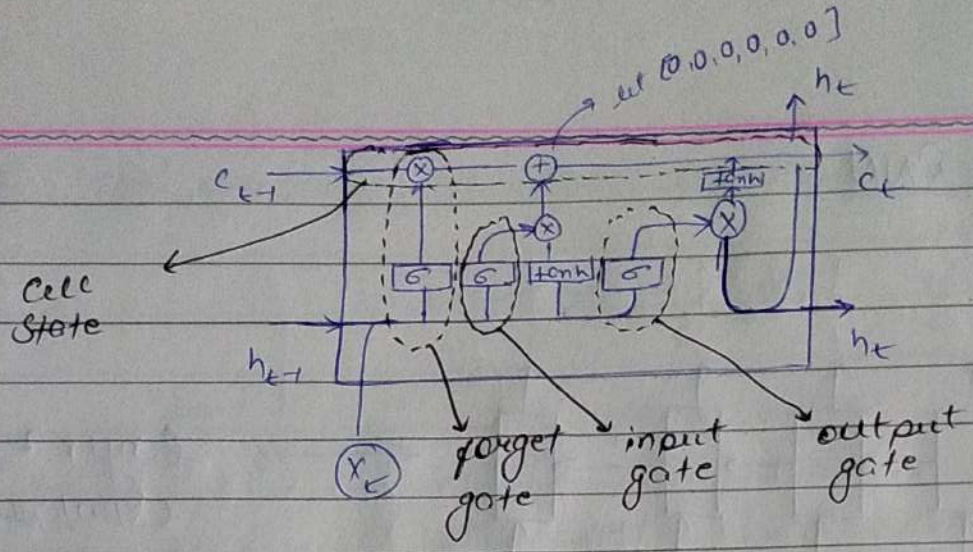
(refer colah.github.io understanding - LSTMs)







Youtube: Programming Cradle



-> Cell State : Works similar to skip connection in ResNet.

-> forget gate : Decides how much data to be retained or how much data to be forget

example :- let C_{t-1} : [2, 8, 9, 4, 0, 6]

forget gate : (1/2, 1/2, 1/2, 1/2, 1/2, 1/2)

$C_t = [[1, 4, 4.5, 2, 0, 3] \odot [0,0,0,0,0,0]]$

[1, 4, 4.5, 2, 0, 3]

forgetting 50% of input.

Youtube: Programming Cradle

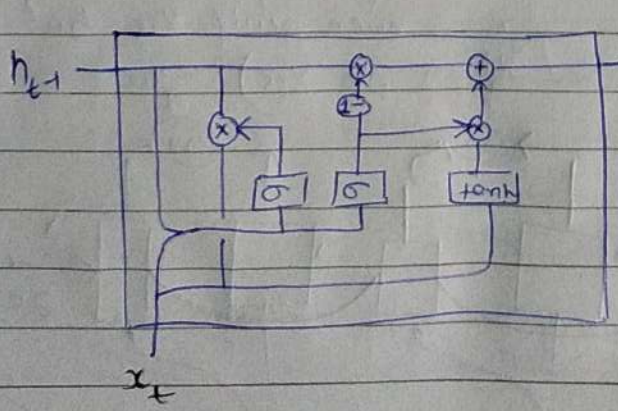
GRU (Gated Recurrent Unit)

↳ Simplified version of LSTM

↳ faster to train

↳ As powerful as LSTM

↳ has only 2 gates (reset and update gate)



$z_t = \sigma(W_z \cdot [h_{t-1}, x_t])$

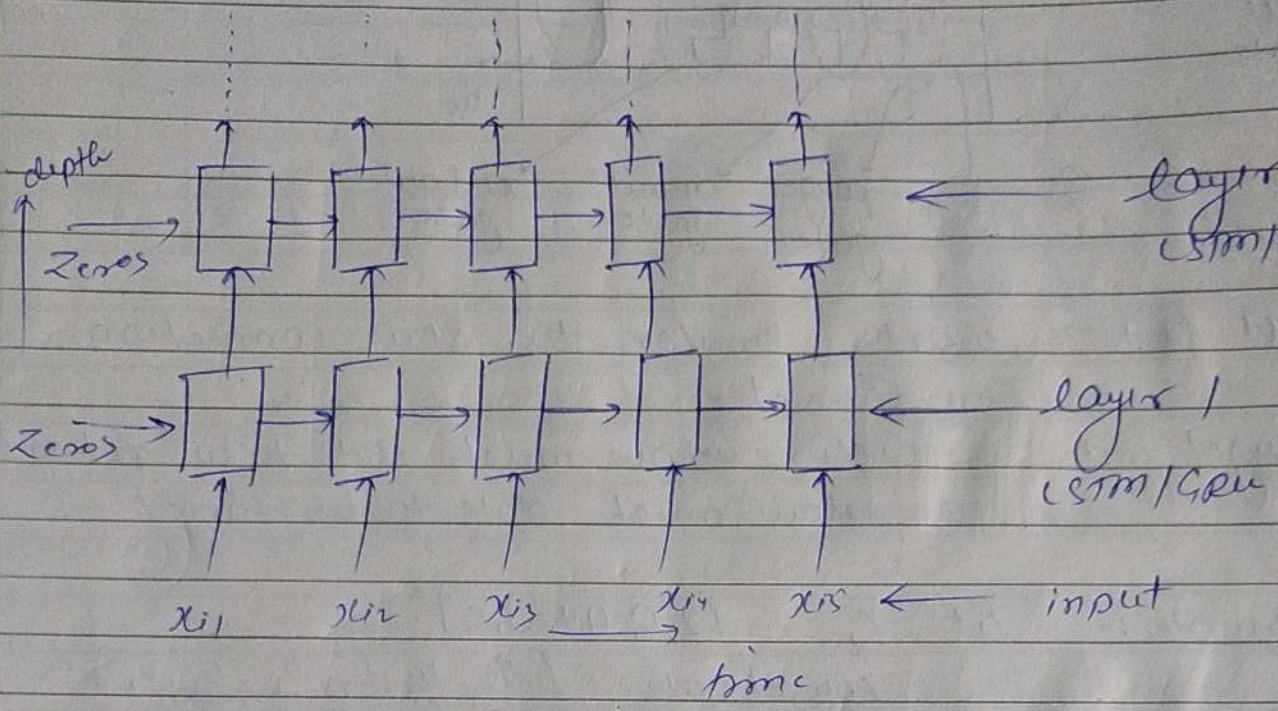
$r_t = \sigma(W_r \cdot [h_{t-1}, x_t])$

$\hat{h}_t = \tanh(W \cdot [r_t * h_{t-1}, x_t])$

$h_t = (1 - z_t) * h_{t-1} + z_t * \hat{h}_t$

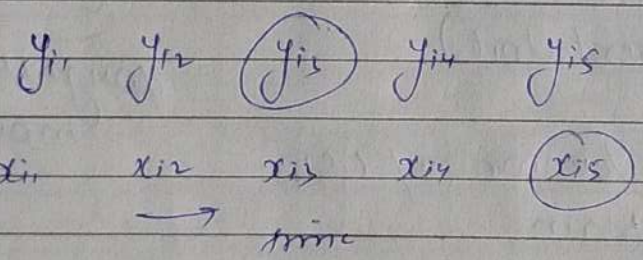
Youtube: Programming Cradle

✱ Deep RNN

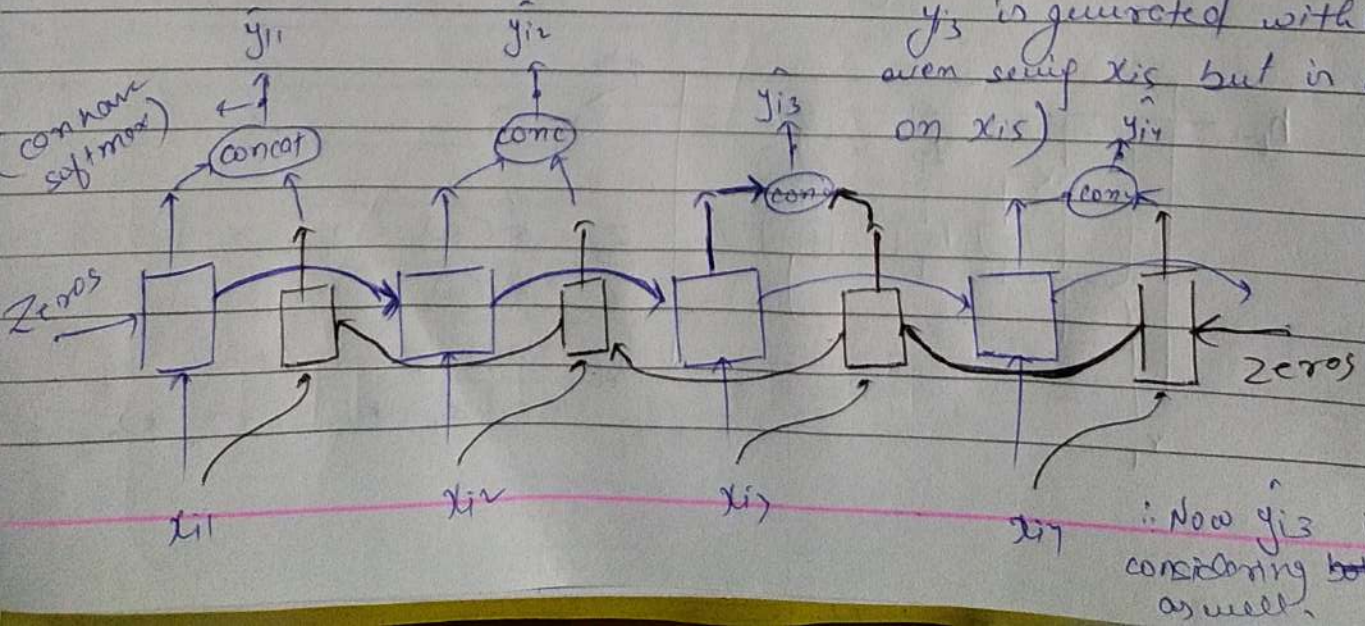


✱ 'Bi'-directional RNN

Youtube: Programming Cradle



Q) what if $\hat{y}_{i3}$ is dependent on x_{i5}? (in unidirectⁿ $\hat{y}_{i3}$ is generated without even seeing x_{i5} but in dependent on x_{i5})



Youtube: Programming Cradle

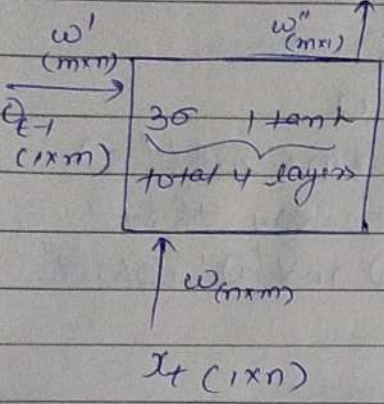
⊛ Total number of trainable params:-

$4(m^2+n\times m+m)$ or $4(m^2+n\times m)$

when bias(b) is considered ↑ when bias(b) is not considered

m: no. of artificial neurons in each layer

n: dimensions of input 'x_t'



⊛ Generative Adversarial Networks :-

Task :- Given MNIST images → D ⇒ Create/generate new MNIST imgs. very similar to images in D but not same.

Youtube: Programming Cradle

⊛ At first solve simpler problem:-

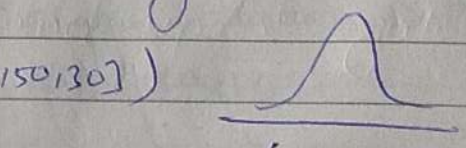
given D = Set of heights of students

$= \{h_1\ h_2\ h_3 \dots h_m\}$

generate $D' = \{h'_1\ h'_2 \dots h'_m\}$

Step ① find the distribution of the D; lets say it is normally distributed as $N(150,30)$

PDF : $P_\theta(H) = N(\theta=\{\mu,\sigma\}_\theta = [150,30])$



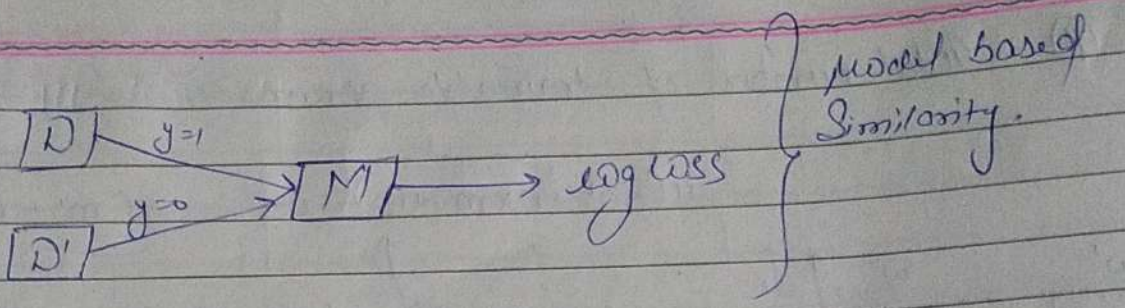
↳ it can be
- normal
- uniform
etc.

Step ② Generate random samples using parameters (μ,σ) (of) found in previous step.

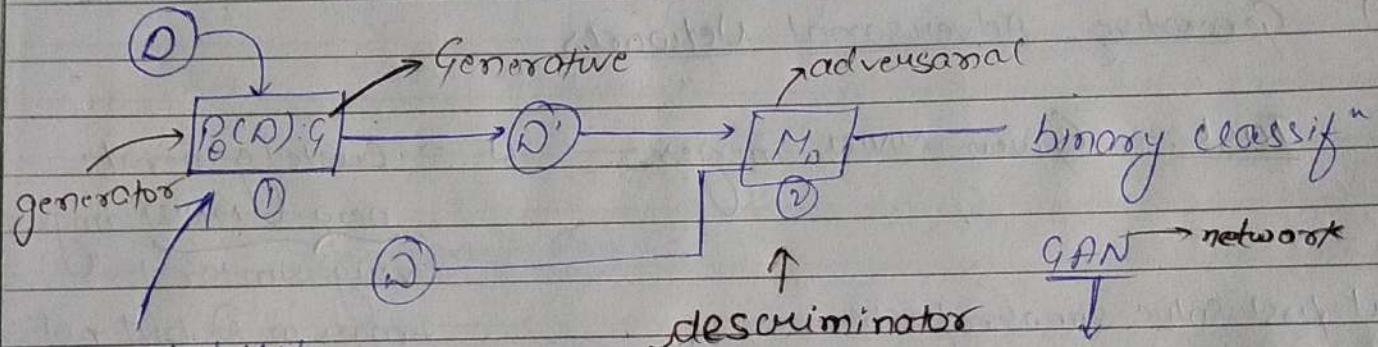
Step ③ measure / Test how similar $D \approx D'$ (how similar dist. of D and D' are)

→ We can use QQ plot, KS test, KL divergence but there is another way..... (model based Invariant)

Youtube: Programming Cradle



- label D as class 1 and D' as class 0
- if model M (can be any model) performs well => log loss will be less and hence we can say that model is able to distinguish b/w D and D' which means $D \neq D'$



probability dist. f" used to create new data

(JS-dist) } won't
(KL div) } work very well in case of high dim.

∴ PDF will not work well the data is high dimentional since we use DNN or CNN in ① box, ② box is always some model (can be NN also)

for training GAN

- fix M_D model's param and train G
- find G and train M_D using gradient ascent

Youtube: Programming Cradle

Youtube: Programming Cradle

✱ Encoder - Decoder Model

✱ Machine - Translation :-

$\langle x^1, x^2, x^3 \ldots x^T \rangle \xrightarrow{M} \langle y^1, y^2, y^3 \ldots y^K \rangle$

input (English) — output (Hindi)

NOTE :- t need not equal to K

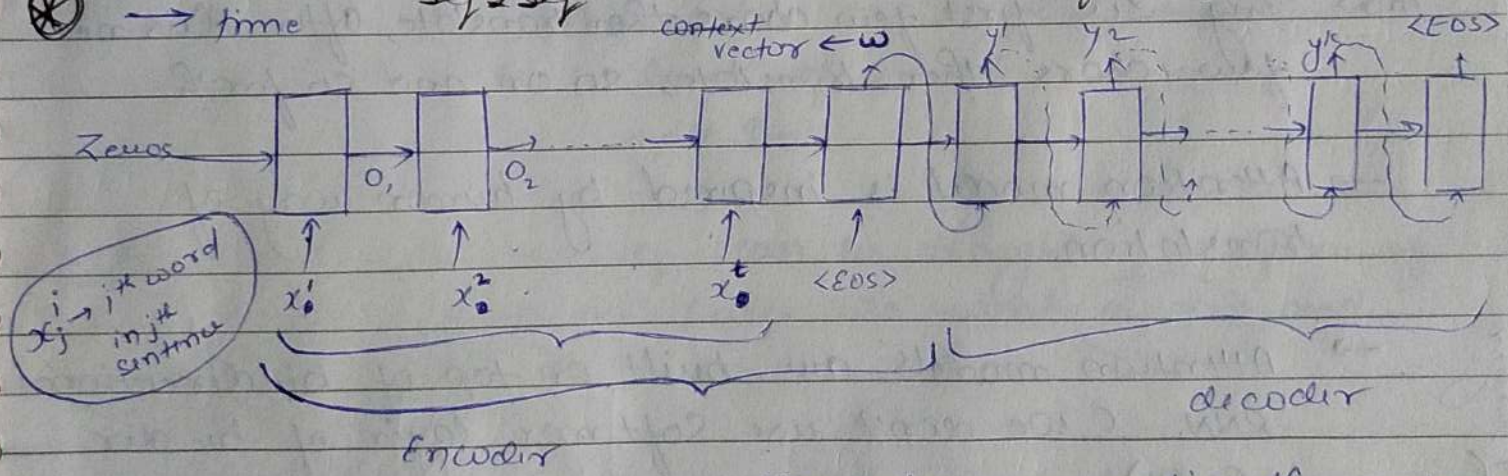
✱ Image - captioning / descriptions :

Youtube: Programming Cradle

Image (input) $\xrightarrow{M}$ $y^1 y^2 y^3 \ldots y^K$ (output)

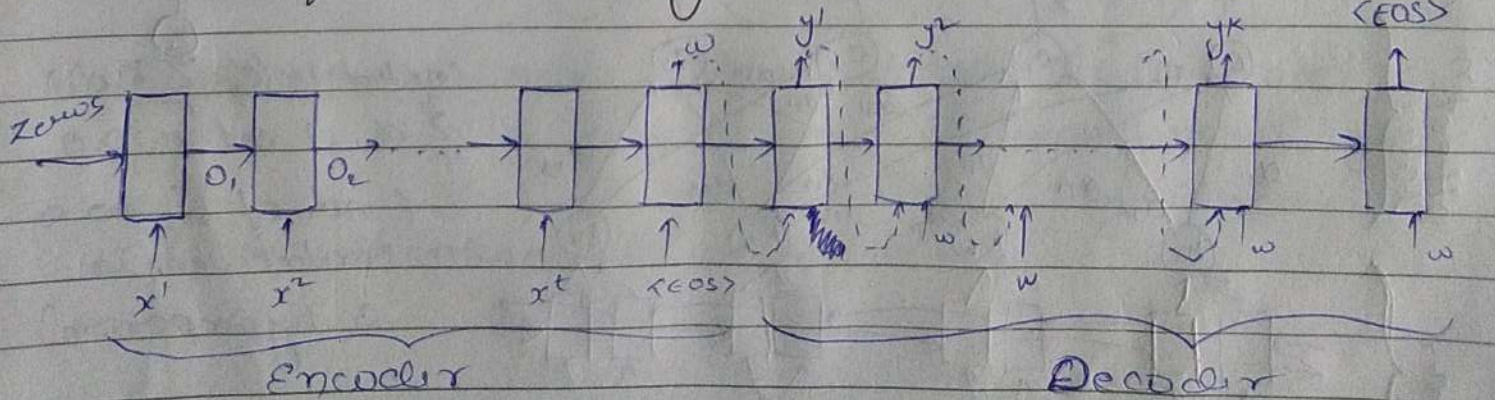
✱ Both the below models are seq 2 seq → time

EOS → End of string / sequence



ω → vector representing the sentence

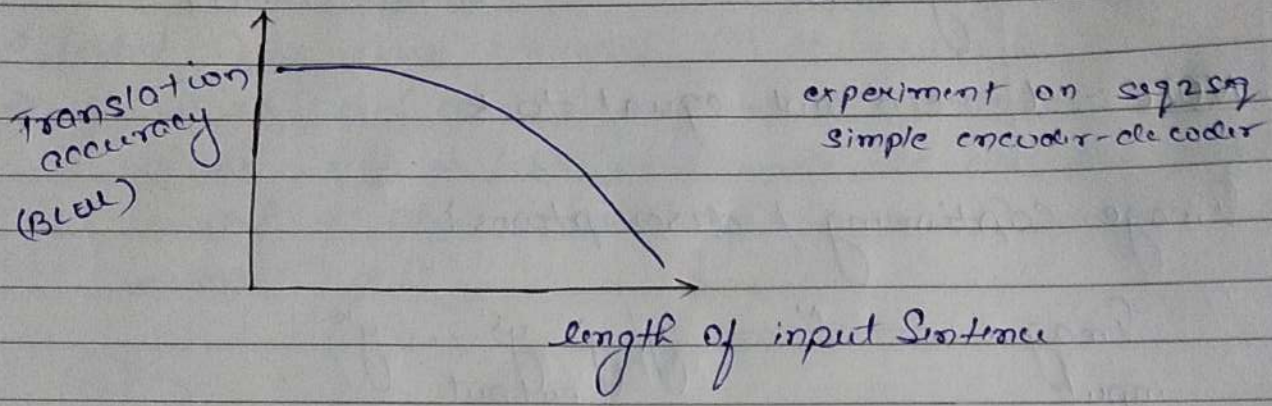
✱ Above model works really well for short sentences. But what if sent. are long.? we can use below model.



Youtube: Programming Cradle

Attention Models

* Biggest problem with Seq2Seq models is they don't perform very well with long sentences.

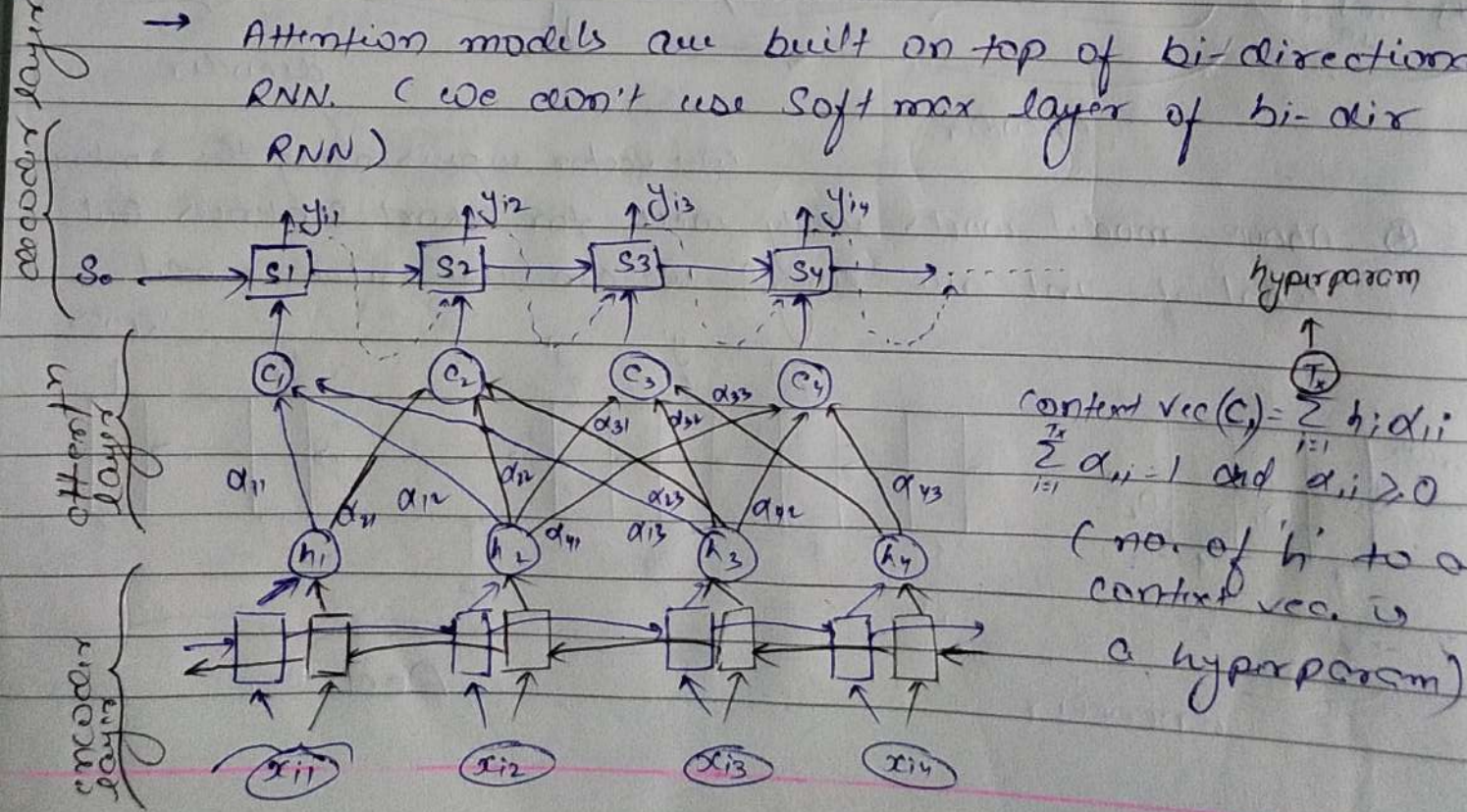


•> How humans translate? Youtube: Programming Cradle

Ans> They see first few chars then translate, after this next few chars then translate, so on and so forth.

→ Attention model is inspired by human way of translation.

→ Attention models are built on top of bi-directional RNN. (we don't use softmax layer of bi-dir RNN)



context vec $(c_i) = \sum_{i=1}^{T_x} h_i \alpha_{ij}$

$\sum_{i=1}^{T_x} \alpha_{ij} = 1$ and $\alpha_{ij} \geq 0$

(no. of 'h' to a context vec. is a hyperparam)

Youtube: Programming Cradle

Neural Machine ~~Transla~~ translation by jointly learning to align and translate.

Q) How to compute α_{ij}'s?

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j \quad \text{(eqn 5 in the research paper)}$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$

$$e_{ij} = a(s_{i-1}, h_j)$$

↑ attention-model

↓ feed forward NN (small)

example :-

$$e_{12} = a(s_0, h_2)$$

⊛ Drawback Youtube: Programming Cradle

Time Complexity : $O(T_1 \cdot T_2)$

T_1 → length of input sent.

T_2 → length of output sent.

NOTE :) constraint on α_{ij} can be thought of as a form of regularization. Since we want $\alpha_{ij} \geq 0$ and $\sum_{i=1}^{T_x} \alpha_{ij} = 1$

) we are using exp. ($e^{\wedge}$) in calculat^n of α_{ij} to make sure $\alpha_{ij} \geq 0$ and when we sum all α_{1j} or α_{2j} or α_{3j} ..... we will get it as 1

⊛ Transformers and BERT Blog: 'The illustrated transformer' by Jay Alammar

Blog: "Attention is all you need."

Youtube: Programming Cradle

✱ Image Segmentation

↳ input is an image and output is also an img.
↳ Diff. subject in the image will be colored in diff. colors.
→ coco dataset
→ image Segmentation traffic dataset (cityscapes)
→ kaggle medical segmentation (gg-mri-segment")

→ Segmentation vs Object detection :-

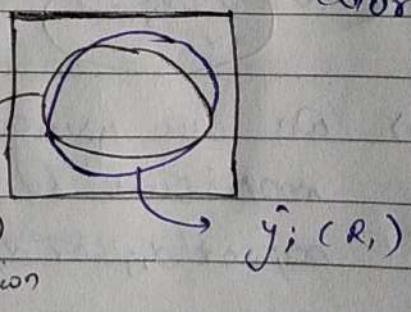
pixel-level Segment"	bounding box
(color whole object, diff obj in diff color)	(draw boxes around diff. objects) (yolo obj. det" algo)

1→ Performance Metric :- Youtube: Programming Cradle

i) Average Precision and Recall for each class/color (coco ds)

ii) Jaccard Sim = $\frac{R_i \cap \hat{R}_i}{R_i \cup \hat{R}_i}$ (intersect" over union)

ideal value = 1



→ classical approaches :-
① clustering. { K-means, ... }
② Edge-detection
③ Graph-Theoritic (spectral clustering)

Youtube: Programming Cradle

→ Paper: lmb.informatik.uni-freiburg.de/people/ronneber/u-net/

→ Deep-learning based method.
- Unets
- Fully convolutional Networks
- Mask R-CNN
- SegNet
- ⋮

⊛ Object Detection Youtube: Programming Cradle

- input : img/vid
- output : bounding box-1, obj. class1
  bounding box-2, obj. class2
  ⋮
- Dataset: coco (80-class-data)
- Trade off: Speed vs mAP (avg. precision)
  ↓ ↳ medical
  self driving cars
- Diff" algorithms :-

R-CNN	SSD (Single Shot detect")	Yolo v1
Fast R-CNN	Retina Net	Yolo v2 / Yolo 9000
Faster R-CNN		Yolo v3

Yolo v3 : feature extractor
→ fully convolutional Network
→ Conv-BatchNorm-LeakyRelu
→ Conv. with stride - down sample
→ no padding
→ pre-trained model.

paper → pjreddie.com/media/files/papers/YOLOv3.pdf.

Youtube: Programming Cradle

Q) Why Darknet-53 was developed?

A) To match the accuracy of ResNet-152 and at the same time to match the no. of operat^n of Darknet -19. and also for higher fps.

Youtube: Programming Cradle

✱ Per class Sigmoids

→ for multiclass classification Softmax can be used but it will give only one class which has the max. probability. But in case of situations like we have a class "person" and "women" like a women can be a person so 2 classes can be returned. BUT this can't be done using Softmax. Hence logistic regression / sigmoids are build for each classes.

✱ Loss:

$$\sum \text{coordinate} * \text{Squared loss} + \sum_{\text{no. obj in boxes}} \text{log-loss for } p_i$$

on $\{\sqrt{tw}, \sqrt{th}, tx \text{ and } ty\} \rightarrow (0 \text{ to } \infty)$ $\quad (0 \text{ to } 1)$

* bounding boxes containing an obj in ground truth

+ log loss for p_o

+ log loss for each p_i

YoloV3 Network architecture bcog.dmprof.com