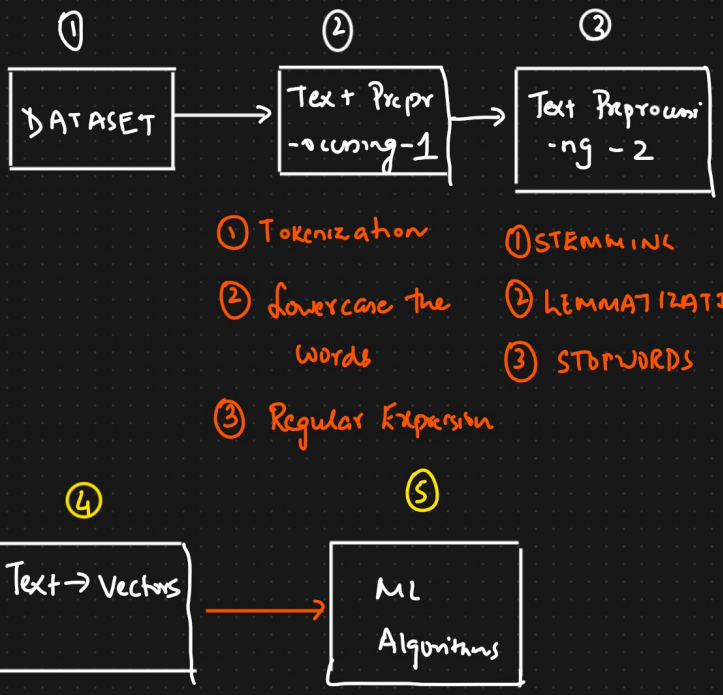
Text Preprocessing → What We have learnt?

Sentiment Analysis

	Text	O/P
D1	The food is good	1
D2	The food is bad	0
D3	Pizza is Amazing	1
D4	Burger is bad	0



D1 → | 1 | 0 | 1 | 1 | → 1

① One Hot Encoded
② Bag of Words (BOW)
③ TF-IDF
④ Word2Vec
⑤ Avg Word2Vec

Gensim

① One Hot Encoding

	Text	O/P
D1	The food is good	1
D2	The food is bad	0
D3	Pizza is Amazing	1

Test [Burger is bad]

Vocabulary size: 7

Vocabulary {unique words}

The	food	is	good	bad	Pizza	Amazing
1	0	0	0	0	0	0
0	1	0	0	0	0	0

D1 4x7
[[1 0 0 0 0 0 0],
[0 1 0 0 0 0 0],
[0 0 1 0 0 0 0],

D2 4x7.
[[1 0 0 0 0 0 0],
[0 1 0 0 0 0 0],
[0 0 1 0 0 0 0],

D3 3x7
[[0 0 0 0 0 1 0],
[0 0 1 0 0 0 0],
[0 0 0 0 0 0 1]]

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

{50K vocabulary size}

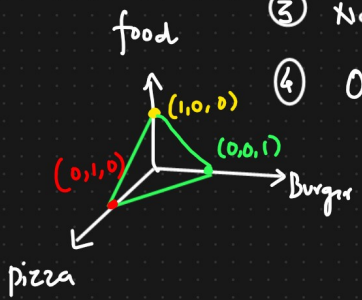
Advantages

① Easy to implement with python
[sklearn OneHotEncoder, pd.get_dummies()]

Disadvantages

① Sparse matrix → Overfitting
② ML Algorithm → Fixed size I/p
③ No semantic meaning is getting captured
④ Out of Vocabulary (OOV).

food	pizza	burger
1	0	0
0	1	0
0	0	1



② Bag of Words

Dataset-

Text	O/P
He is a good boy	1
She is a good girl	1
Boy and girl are good	1

Lower all the words case → Stopwords

S1 → good boy
S2 → good girl good
S3 → Boy girl good school

Test

Vocabulary	frequency
good	3
boy	2
girl	2

	[good	boy	girl]	O/p
S1	[1	1	0]	1
S2	[1	0	1]	1
S3	[1	1	1]	1

Binary BOW and BOW

{1 and 0} {Count will get updated based on frequency}

Advantages

① Simple and Intuitive

② Fixed Sized I/p → ML Algorithms

Disadvantages

① Sparse matrix or array → Overfitting

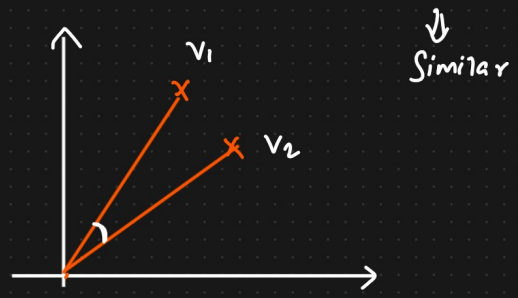
② Ordering of the word is getting changed

③ Out of Vocabulary (OOV).

④ Semantic meaning is still not captured.

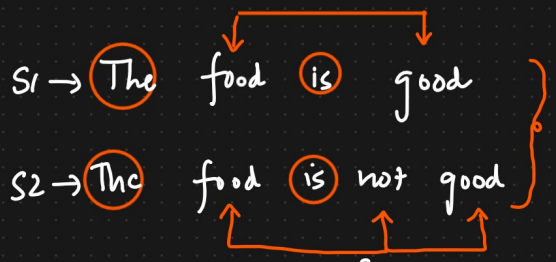
The food is good → [1 1 1 0 1] → v_1

The food is not good → [1 1 1 1 1] → v_2



④ N-grams

Eg: bigrams, trigrams



	food	not	good
→	1	0	1
→	1	1	1

	[food	not	good	food good	food not	not good]
S1	1	0	1	1	0	0
S2	1	1	1	0	1	1

Sklearn → n-grams = (1, 1) → unigrams

= (1, 2) → unigram, bigram

= (1, 3) → unigram, bigram, trigram

= (2, 3) → Bigram, trigram.

④ TF-IDF [Term Frequency – Inverse Document Frequency]

S1 → good boy

S2 → good girl

S3 → boy girl good

$$\text{Term Freq (TF)} = \frac{\text{No. of rep of words in sentence}}{\text{No. of words in sentence}}$$

$$\text{IDF} = \log_e\left(\frac{\text{No. of Sentences}}{\text{No. of sentences containing the word}}\right)$$

Term Frequency * IDF

	S1	S2	S3
good	1/2	1/2	1/3
boy	1/2	0	1/3
girl	0	1/2	1/3

Words	IDF
good	$\log_e(3/3) = 0$
boy	$\log_e(3/2)$
girl	$\log_e(3/2)$

Final TF-IDF

	[good	boy	girl]
Sent 1	0	$\frac{1}{2} \times \log_e(3/2)$	0
Sent 2	0	0	$\frac{1}{2}\log_e(3/2)$
Sent 3	0	$\frac{1}{3}\log_e(3/2)$	$\frac{1}{3}\log_e(3/2)$

Op

Bow

1	1	0
1	0	1
1	1	1

Advantages

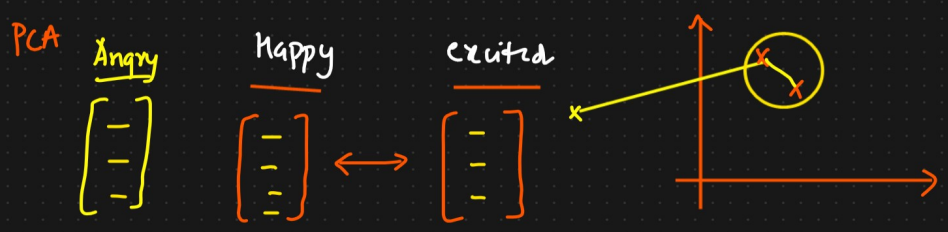
1. Intuitive
2. Fixed Size → Vocab size
3. Word Importance is getting captured

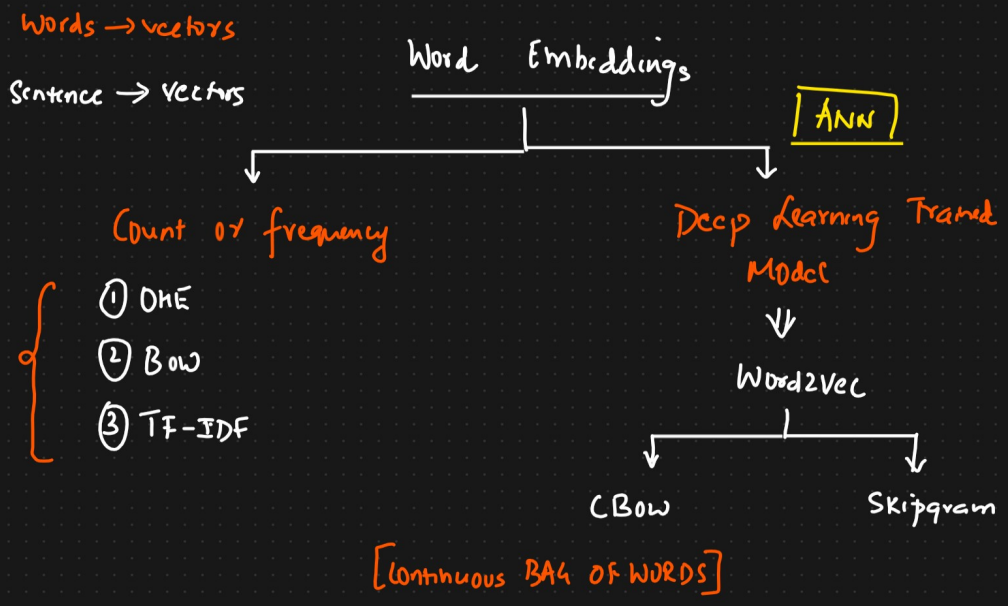
Disadvantages

1. Sparsity still exists
2. OOV

Word Embeddings [Wikipedia]

In natural language processing (NLP), word embedding is a term used for the representation of words for text analysis, typically in the form of a real-valued vector that encodes the meaning of the word such that the words that are closer in the vector space are expected to be similar in meaning.





Word2Vec → Feature Representation

Google

Word2vec is a technique for natural language processing published in 2013. The word2vec algorithm uses a neural network model to learn word associations from a large corpus of text. Once trained, such a model can detect synonymous words or suggest additional words for a partial sentence. As the name implies, word2vec represents each distinct word with a particular list of numbers called a vector.

Vocabulary → Unique words → Corpus.

Feature Representation	Boy	Girl	KING	QUEEN	Apple	Mango
Gender	-1	1	-0.92	+0.93	0.01	0.05
Royal	0.01	0.02	0.95	0.96	-0.02	0.02
Age	0.03	0.02	0.75	0.68	0.95	0.96
Food	-	-	-	-	0.91	0.92
					-	-

300 dimension !

$$[\text{KING} - \text{BOY} + \text{QUEEN} = \text{GIRL}]$$

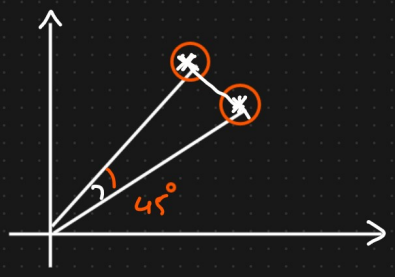
KING [0.95, 0.96] Man [0.95 0.98]

QUEEN [-0.96, 0.95] Women [-0.94, -0.96]

KING - MAN + QUEEN = WOMEN

Cosine Similarity

Distance = 1 − Cosine Similarity

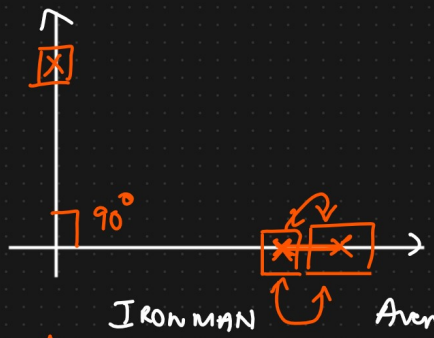


Cosine-Sim = $\cos 45 = \frac{1}{\sqrt{2}} = 0.7071$

Distance = 1 − 0.7071

↳ = 0.29

1 − 1 = 0



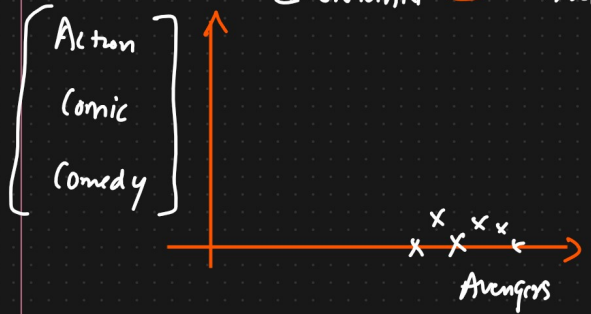
Distance = 1 − 0

= 1

Distance = 1 − cos0

= 1 − 1

= 0//.



ANN, Loss, Optimizers

(#) Word2Vec → CBOW [Continuous Bag of Words]

→ Skipgram.

Pretrained Model

Train A Model For Scratch

① CBOW [Continuous Bag of Words]

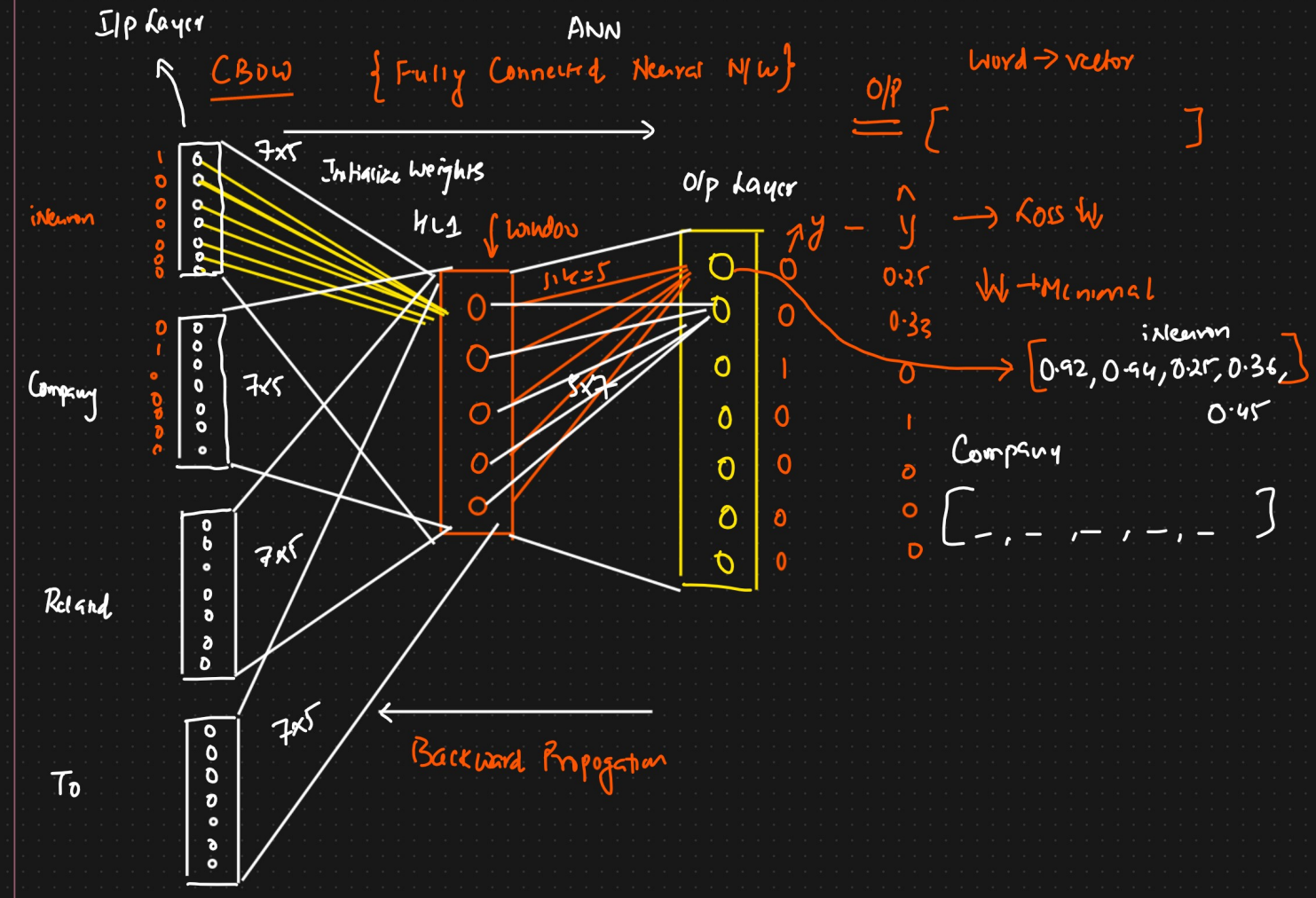
CORPUS ÷ DATASET

[iNeuron Company IS Related To DATA SCIENCE] → Vocabulary

300 dimension — 300 ← Window size = 5 ⇒ Feature Representation [_ _ ___] — OHE

I/P	O/P
→ [iNeuron, Company, Related, To]	IS
→ Company, IS, To, DATA	Related
→ Is, Related, DATA, SCIENCE	To

iNeuron	[1 0 0 0 0 0 0]
Company	[0 1 0 0 0 0 0]
Related	[0 0 0 1 0 0 0]
to	0 0 0 0 1 0 0



② Skipgram – Word2Vec

Window size=5

O/p

→ [Neuron, Company, Related, To]

→ Company, Is, To, DATA

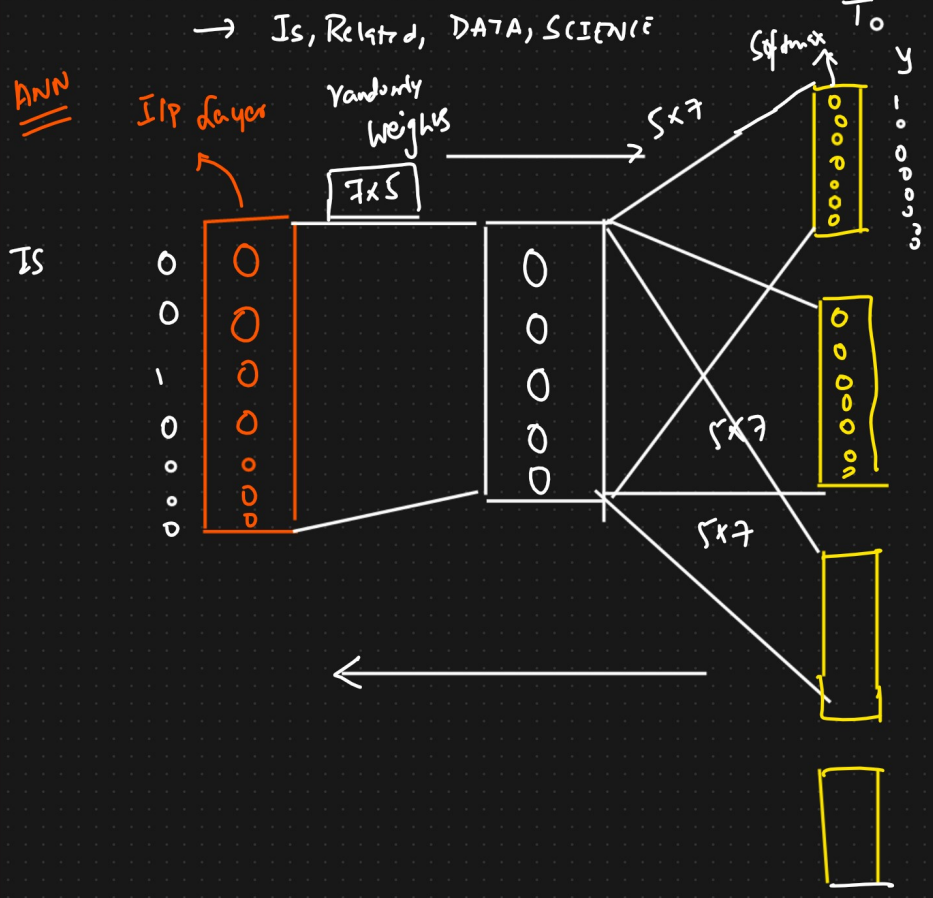
→ Is, Related, DATA, SCIENCE

I/p

Is

Related

To



Improve CBow or Skipgram

① Increasing the Training Data

② Increase the window size - vector dimension is also increasing.

When Should we apply CBow or SkipGram.

- Small Dataset → CBow
- Huge Dataset → SkipGram

Google Word2vec

3 billion words → Google News

feature representation of 300 dimension vectors]

Cricket → [- - - - - - - - - - - - - -].

Gensim

Advantages of Word2Vec

*) Sparse Matrix ⟶ Dense Matrix

*) Semantic Info is getting captured [Honest, good]

*) Vocabulary Size ⟶ Fixed set of dimension vectors
Google Word2Vec [300 dimension]

*) OOV is also solved

⑧ Avg Word2Vec

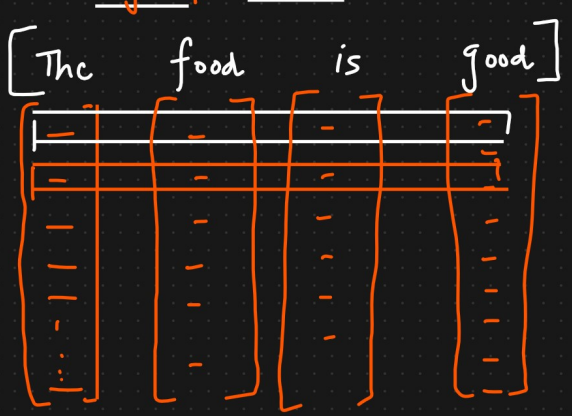
	Text	O/P
D1	The food is good	1
D2	The food is bad	0
D3	Pizza is Amazing	1

Avg word2vec

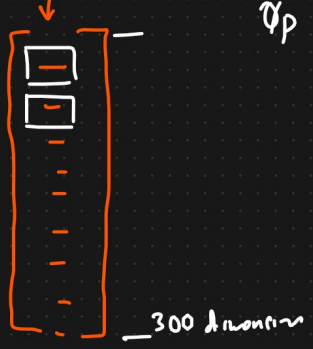
[. . . . . .]
[ ]
[ ]

O/P
1

Google pretrained Word2Vec Model



Avg Word2Vec

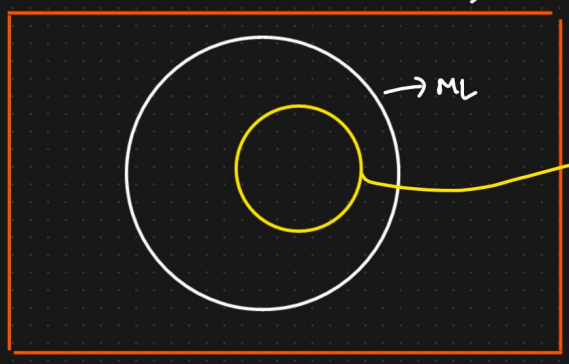


⑧ Gensim, Glove

↳ pretrained Google Word2Vec

↳ Train a word2vec from scratch

Deep Learning



Deep Learning

1. ANN → Artificial Neural N/W → Classification, Regression
2. CNN → Convolutional Neural N/W → I/P : Images, Video frames → RCNN, MASKED RCNN, Detectron, YOLO V5, V6, V7 (Computer Vision, Objection Detection)
3. RNN → Recurrent Neural N/W → NLP → NLP, Time Series

I/P: Text, Time Series

⇓

{ Word Embedding, LSTM RNN, GRU RNN, Bidirectional LSTM RNN, Encoder Decoder, Transformers, BERT }

FRAMEWORK

TENSORFLOW

End to End Project

② Why Deep Learning Is Becoming popular?

2005 ⟶ Facebook, Youtube, Whatsapp, Linked, Twitter } ⇒ Social Media platform

⋮

2011-2012

⇓

DATA WAS GENERATED

Exponentially

Image, Text, Documents, Videos

↳ Big DATA ⟶ HADOOP → unstructured DATA

⇓

2011 ⇒ Cloudera, Hortonworks

Netflix → movie streaming platform

⇓

Recommendation System

⇓

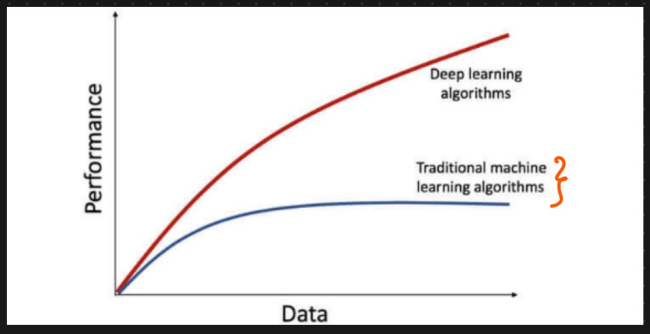
Increase Revenue

2011-2012 :

① Hardware Requirement ⟶ GPU's cost↓↓↓ → NVidia GPU Price↓↓↓

TRAIN MODELS

② Huge amount of data is getting generated → Deep Learning Models Perform well



③ Deep learning is been used in Many Domains

1. Medical
2. ECommerce
3. Retails
4. Marketing . . . . .

④ Frameworks Open source

Community size ↑↑

Tensorflow

⇓

Google

Pytorch

⇓

Facebook

⇓

More Research

③ Perceptron [Artificial Neuron or Neural Network Unit]

① Input layer ✓

② Hidden layer ✓

③ Weights ✓

④ Activation function ✓

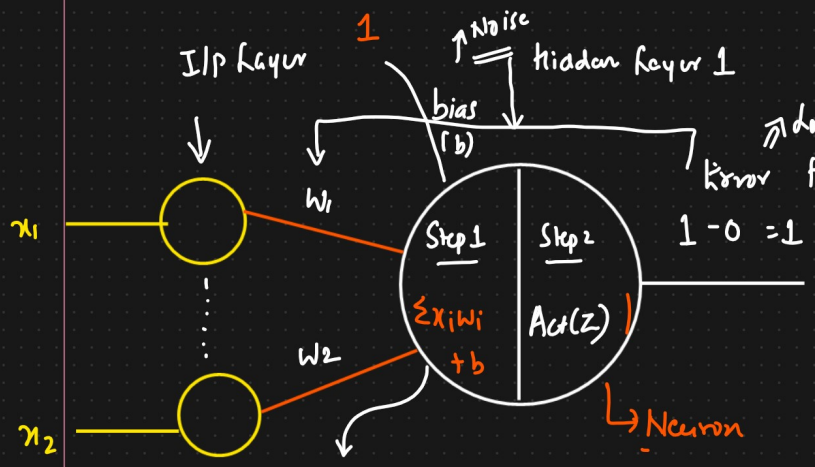
[Single layered NN]

⇓

Binary classifier

DATASET

	x_1 Id	x_2 No. of study hours	o/p Pass/Fail
→	95	3	0
→	110	4	1
→	100	5	1



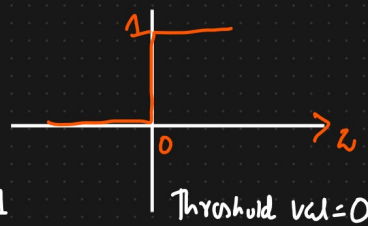
$$z = w_1 x_1 + w_2 x_2 + b$$

$$z = \sum_{i=1}^{n} x_i w_i + b$$

Activation function

Transform the output between 0 to 1, -1 to 1

Step Function

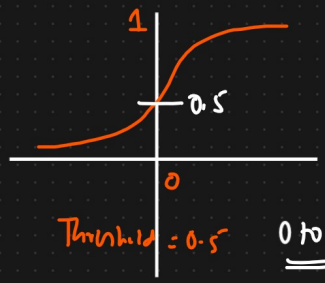


0 to 1

Threshold val = 0

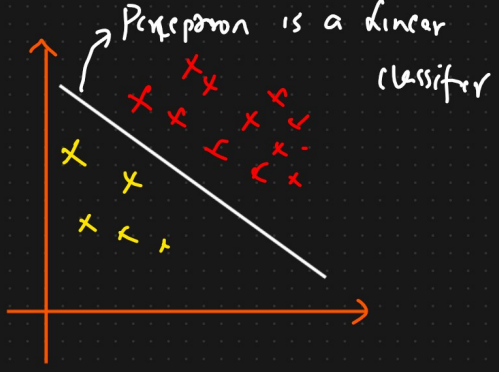
$$\begin{cases} 0 & z \le 0 \\ 1 & z > 0 \end{cases}$$

Sigmoid Function



Threshold = 0.5 0 to 1

$$\begin{cases} 1 & z > 0.5 \\ 0 & z \le 0.5 \end{cases}$$



Step 1

$$z = \sum_{i=1}^{n} w_i x_i + b$$

$$z = b + w_1 x_1 + w_2 x_2 + w_3 x_3 + \cdots + w_n x_n$$

$$y = mx + c$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \; - - - + \beta_n x_n$$

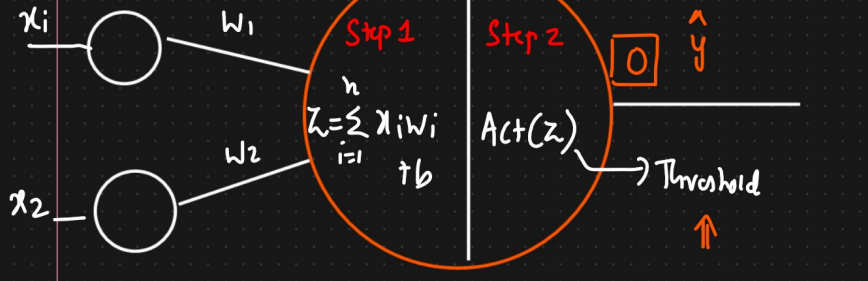
linear problem Statement

Perceptron Models

- Single Layer Perceptron Model
- Multi Layered Perceptron Model [ANN]

Single Layer Perceptron Model

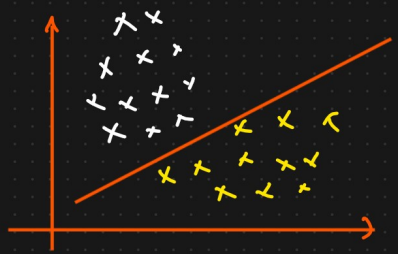
Feed Forward Neural NW



Multi Layered Perceptron Model

1. Forward Propogation
2. Backward "
3. Loss functions
4. Activation functions
5. Optimizers

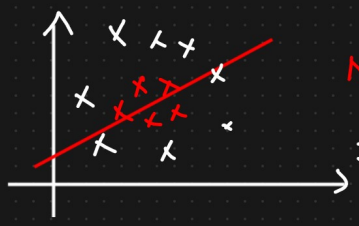
linear Separable problem



Binary classification

⇓

linearly separable



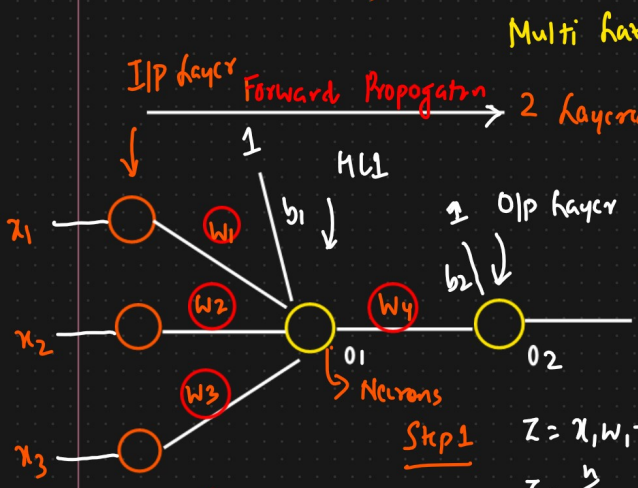
Non linear separable

⇒ Multi Layered NN

5 Multi Layered Perceptron Model [Artificial Neural N/W]

① Forward Propogation
② Backward Propogation → Geoffrey Hinton →

③ Loss function
④ Optimizers ✓
⑤ Activation function:



x_1	x_2	x_3	Pass/Fail
IQ	Study hours	Play hours	O/p
95	4	4	1
100	5	2	1 ←
95	2	7	0

Step 1 $z = x_1w_1 + x_2w_2 + x_3w + b_1$

$z = \sum_{i=1}^{n} w_i^T x_i + b$

① Forward Propogation 2: Activation(z)

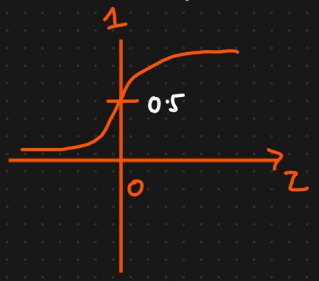
W_1 W_2 W_3 [0.01, 0.02, 0.03] [0.001]

Hidden Layer 1

Step 1: $z = 95 \times 0.01 + 4 \times 0.02 + 4 \times 0.03 + 1 \times 0.01$

$= 1.151$

Sigmoid $= \frac{1}{1+e^{-z}}$



Step 2: Activation (z)

$f(z) = \frac{1}{1+e^{-1.151}} = 0.759$

$O_1 = 0.759$

Hidden Layer 2

$W_4 = 0.02$

$b_2 = 0.03$

Step 1: $z = O_1 * W_4 + b_2$

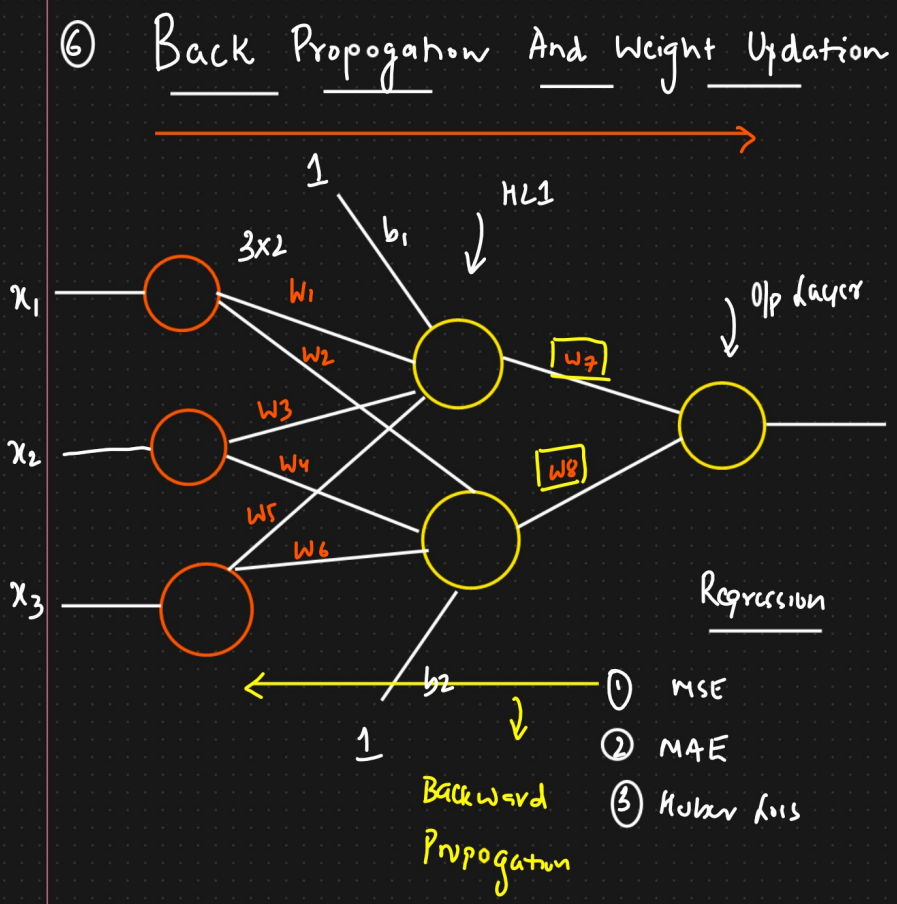
$= 0.759 * 0.02 + 0.03$

$= 0.045180$

Step 2 : Activation(z) $\frac{1}{1+e^{-(0.04518)}} = 0.511129$

$O_2 = 0.511129 \Rightarrow \hat{y}$

⑥ Back Propogation And Weight Updation Formula



I/p features →			↓
x_1	x_2	x_3	Pass/Fail
IQ	Study hours	Play hours	O/p
> 95	4	4	1
→ 100	5	2	1 ←
95	2	7	0

Loss function $(y-\hat{y})^2$ ↓↓↓↓↓↓

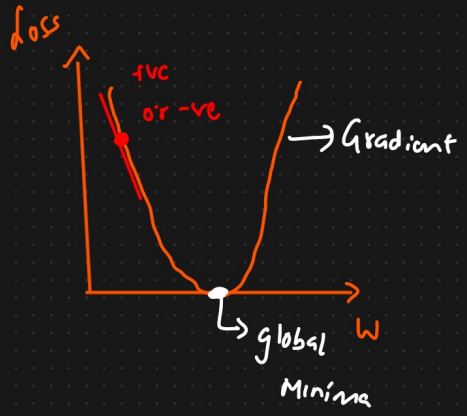
Regression
1. MSE
2. MAE
3. Huber Loss

Classification
1. Binary Cross Entropy
2. Categorical CROSS Entropy

Weight updation Formula

$$W_{7_{new}} = W_{7_{old}} - \eta \left[\frac{\partial L}{\partial W_{7old}} \right]$$ ← slope

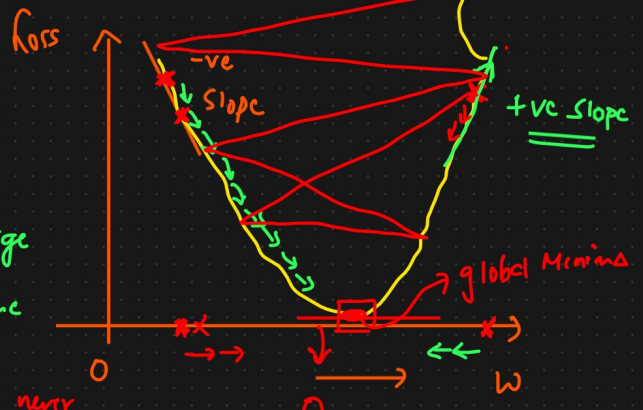
$$W_{8new} = W_{8old} - \eta \left[\frac{\partial L}{\partial W_{8old}} \right]$$ ← Learning Rate



$$W_{new} = W_{old} - \eta \frac{\partial L}{\partial W_{old}}$$ ⇒ Weight Updation Formula.

Gradient Descent ↗ Optimizers

Optimizers : To reduce the loss value



$W_{new} = W_{old} - \eta\,(-ve)$

$= W_{old} + \eta\,(+ve)$

$W_{new} >> W_{old}$

η = large value ⇓ It may never converge

↳ learning Rate

0.001 → Small value.

$W_{new} = W_{old} - \eta\,(+ve)$

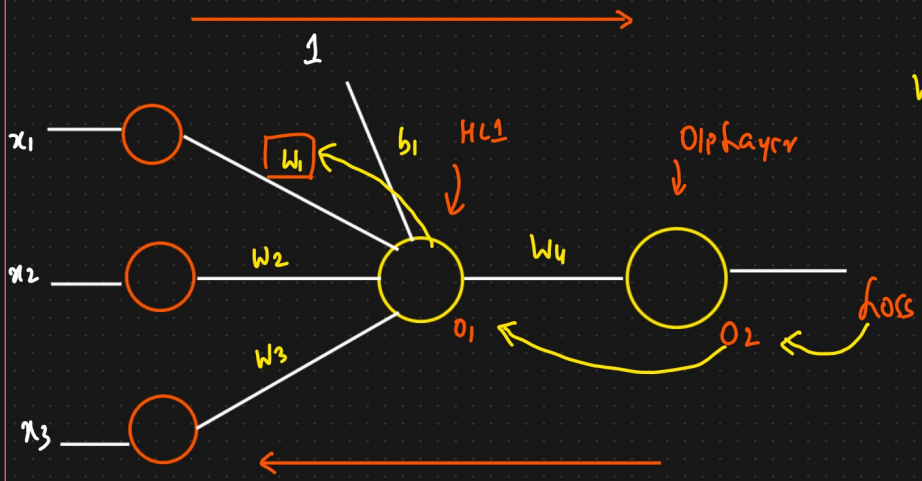
$= W_{old} - \eta\,(+ve)$

$W_{new} << W_{old}$

When W reaches global Minima

$W_{new} = W_{old}$

⑦ Chain Rule of Derivative



$$W_{new} = W_{old} - \eta \frac{\partial h}{\partial W_{old}}$$

$$W_{4_{new}} = W_{4_{old}} - \eta \frac{\partial h}{\partial W_{4_{old}}}$$

$$\frac{\partial h}{\partial w_{4old}} = \frac{\partial h}{\partial o_2} * \frac{\partial o_2}{\partial w_{4old}} \Rightarrow \text{Chain Rule of Derivation}$$

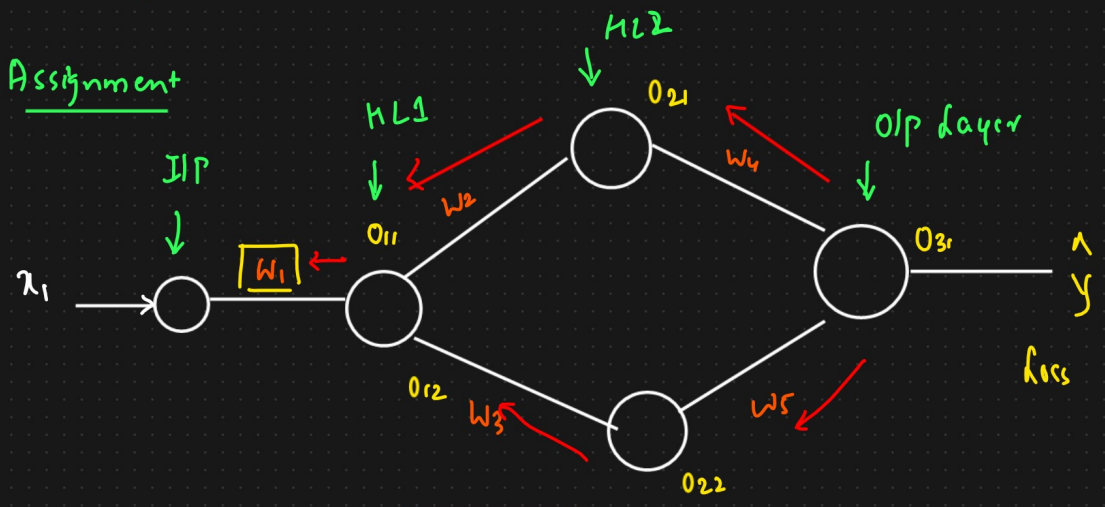
$$w_{1new} = w_{1old} - \eta \boxed{\frac{\partial h}{\partial w_{1old}}}$$

$$\boxed{\frac{\partial h}{\partial w_{1old}} = \frac{\partial h}{\partial o_2} * \frac{\partial o_2}{\partial o_1} * \frac{\partial o_1}{\partial w_{1old}}}$$

w_{2new}

w_{3new}.

Assignment



$$w_{1new} = w_{1old} - \eta \boxed{\frac{\partial h}{\partial w_{1old}}}$$

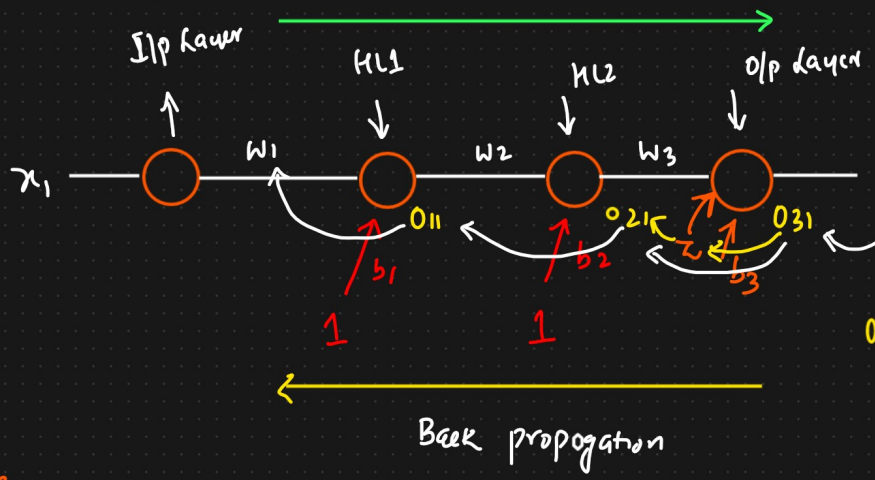
$$\Downarrow$$

$$\frac{\partial h}{\partial w_{1old}} = \left[\frac{\partial h}{\partial o_{31}} * \frac{\partial o_{31}}{\partial o_{21}} * \frac{\partial o_{21}}{\partial o_{11}} * \frac{\partial o_{11}}{\partial w_{1old}}\right]$$

$$+$$

$$\left[\frac{\partial h}{\partial o_{31}} * \frac{\partial o_{31}}{\partial o_{22}} * \frac{\partial o_{22}}{\partial o_{12}} * \frac{\partial o_{12}}{\partial w_{1old}}\right]$$

⑧ Vanishing Gradient Problem And Activation functions $\phi(z)$

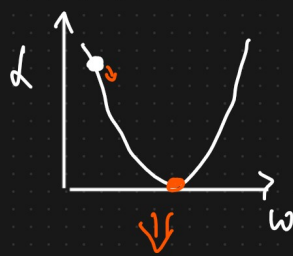


$$z = \sum_{i=1}^{n} w_i x_i + b$$

$\sigma(z) \Rightarrow 0$ to 1

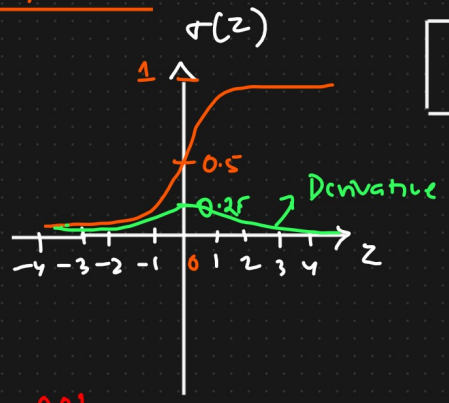
↳ Sigmoid

Optimizers



⑨ Sigmoid Activation function

$$\sigma(z) = \frac{1}{1+e^{-z}}$$



$$0 \le \sigma(z) \le 1$$

Derivative $(\sigma(z))$

$$0 \le \sigma(z) \le 0.25$$

Vanishing Gradient Problem.

$$W_{1\,new} = W_{1\,old} - \eta \left[\frac{\partial h}{\partial W_{old}} \right]$$

η = 0.01

$\frac{\partial h}{\partial W_{old}} \Rightarrow$ small value

$\Rightarrow$ small value

$$W_{1\,new} \approx W_{1\,old}$$ (Equal)

$$\frac{\partial h}{\partial W_{1old}} = \frac{\partial h}{\partial O_{31}} * \frac{\partial O_{31}}{\partial O_{21}} * \frac{\partial O_{21}}{\partial O_{11}} * \frac{\partial O_{11}}{\partial w_{11}}$$

$0-0.25 * 0-0.25 * 0-0.25 \Rightarrow 0.000009$

$\Downarrow$

$\Rightarrow$ Smaller value

$$O_{31} = \sigma(\underbrace{w_3 * O_{21} + b_3}_{z})$$

$z = w_3 * O_{21} + b_3$

$O_{31} = \sigma(z)$

$$\frac{\partial O_{31}}{\partial O_{21}} = \frac{\partial(\sigma(z))}{\partial(z)} * \frac{\partial z}{\partial O_{21}}$$ (Sigmoid)

{Chain Rule}.

$0 \leq \sigma(z) \leq 0.25$ * $\frac{\partial((w_3 * o_{21}) + b_3)}{\partial(o_{21})}$

⇓

Derivative of Sigmoid

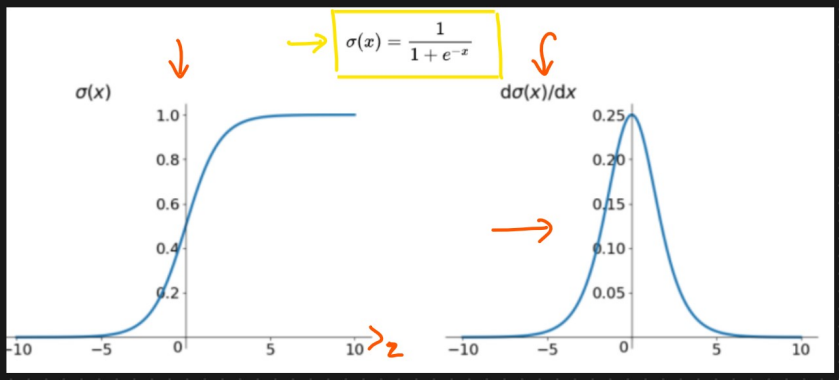
$$\frac{\partial o_{31}}{\partial o_{21}} = 0 \leq \sigma(z) \leq 0.25 * w_{3_{old}}$$

* To fix this problem Researchers started exploring other Activation function

① Tanh ② Relu ③ Prelu ④ Swiss

Activation Functions

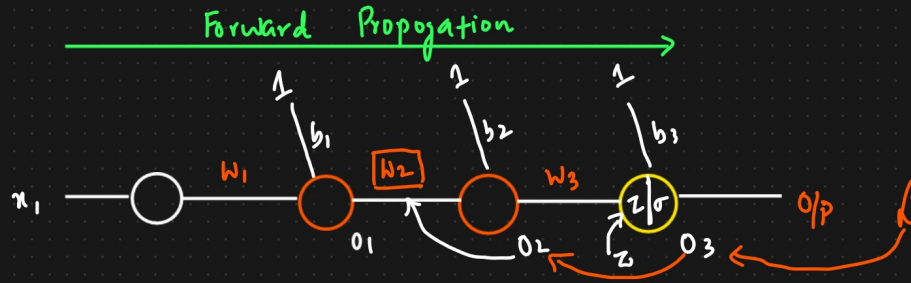
① Sigmoid Activation function [0 to 1] $z = \sum_{i=1}^{n} w^T x + b$



$\sigma(z) \Rightarrow$ 0 to 1

$\phi(z) \Rightarrow$

$\frac{\partial \sigma(z)}{\partial z}$ = 0 to 0.25



$$W_{2new} = W_{2old} - \eta \left[\frac{\partial h}{\partial W_{2old}} \right]$$

η = 0.001

$\frac{\partial h}{\partial W_{2old}}$ = 0.00001

$\Rightarrow W_{2new} \approx W_{2old} \Rightarrow$

$$\frac{\partial h}{\partial W_{2old}} = \frac{\partial h}{\partial O_3} * \frac{\partial O_3}{\partial O_2} * \frac{\partial O_2}{\partial W_2}$$

0.20 * 0.01 *

Let $z = (O_2 * W_3) + b_3$

$$\frac{\partial O_3}{\partial O_2} = \frac{\partial(\sigma(z))}{\partial z} * \frac{\partial z}{\partial O_2}$$

[0 to 1]

$$= [0 - 0.25] * \frac{\partial[(O_2 * W_3) + b_3]}{\partial O_2}$$

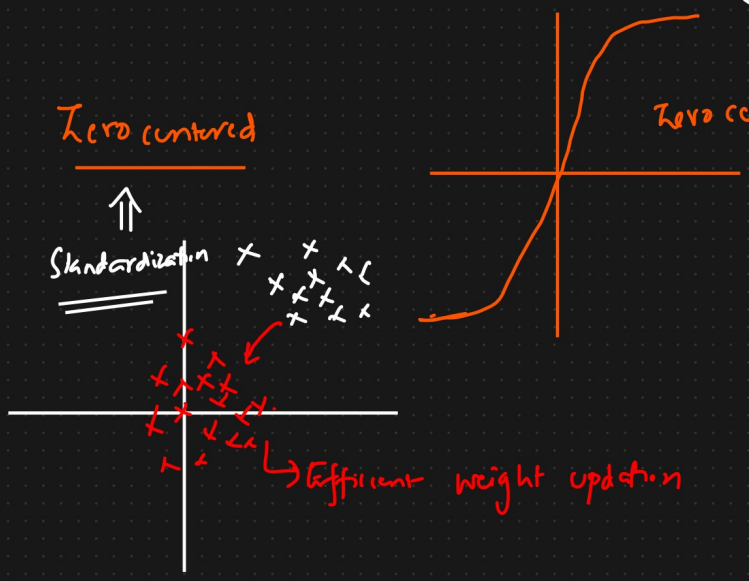
$$= [0 - 0.25] * W_3 \Rightarrow \text{Small value} \Rightarrow$$

Advantages

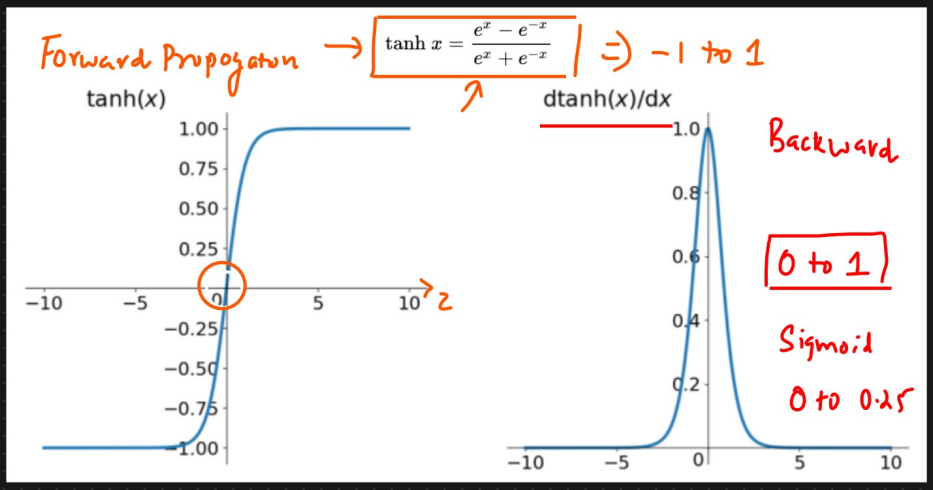
1. Binary classification Suitable.
2. Clear prediction i.e very close 1 or 0

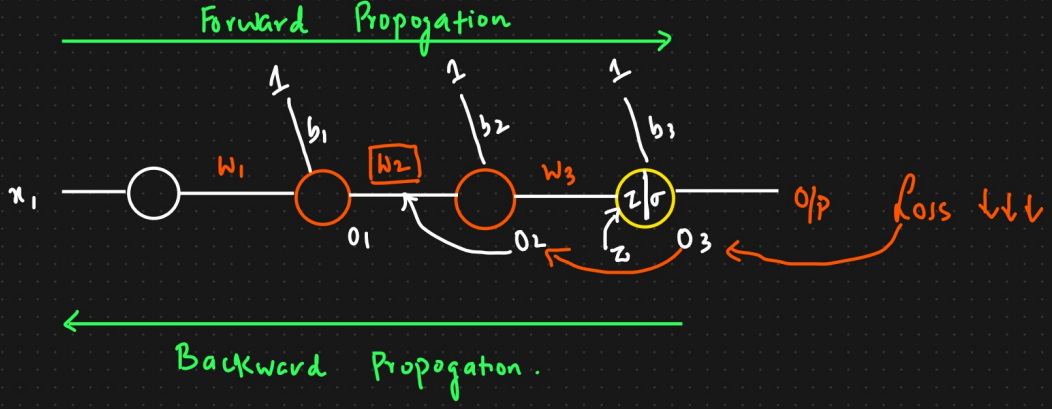
Disadvantages

1. Prone to vanishing gradient Problem.
2. Function output is not Zero centred ⇒ Efficient weight updation
3. Mathematical operation are relatively time Consuming



② Tanh Activation Function





$$\frac{\partial h}{\partial w_{2old}} = \frac{\partial h}{\partial o_3} * \boxed{\frac{\partial o_3}{\partial o_2}} * \frac{\partial o_2}{\partial w_2}$$

$0.20_{11} * 0.01 \times$

Let $z = (o_2 * w_3) + b_3$

$$\frac{\partial o_3}{\partial o_2} = \boxed{\frac{\partial(\tanh(z))}{\partial z}} * \frac{\partial z}{\partial o_2}$$

[0 to 1]

$$= [0-1] * \frac{\partial[(o_2 * w_3) + b_3]}{\partial o_2}$$

$$= [0-1] * w_3 \Rightarrow \text{Small value} \Rightarrow$$

Advantages

① Zero Centric ⇒ Weight updation is Efficient

Disadvantages

① Prone to Vanishing Gradient Problem

② Time complexity

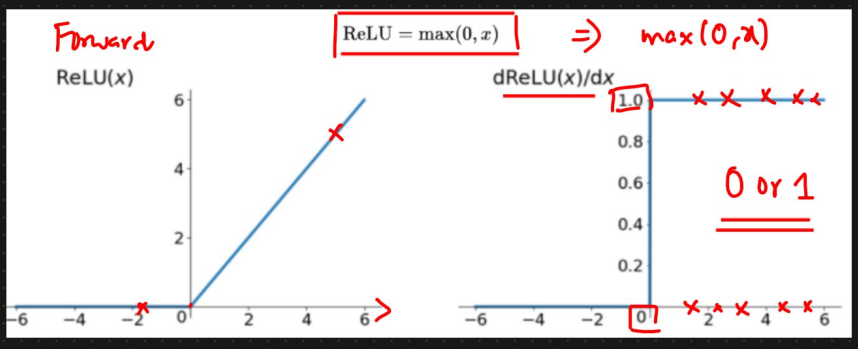
③ Relu Activation Function

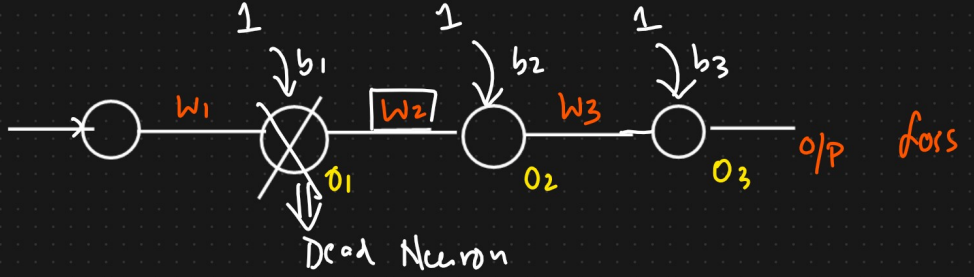
[Rectified Linear Unit].

Tanh ⇒ 0 to 1

Sigmoid ⇒ 0 to 0.25

Forward $\boxed{\text{ReLU} = \max(0, x)}$ ⇒ max(0,x)





$$\frac{\partial h}{\partial w_{2old}} = \frac{\partial h}{\partial o_3} * \boxed{\frac{\partial o_3}{\partial o_2}} * \frac{\partial o_2}{\partial w_2}$$

0.20_{11} * 0.01 ×

let $z = (o_2 * w_3) + b_3$

$$\frac{\partial o_3}{\partial o_2} = \boxed{\frac{\partial (\text{relu}(z))}{\partial z}} * \frac{\partial z}{\partial o_2}$$

[0 to 1]

$$= [0 \text{ or } 1] * \frac{\partial [(o_2 * w_3) + b_3]}{\partial o_2}$$

$$= [0 \text{ or } \underset{+ve}{1}] * w_3 \quad \Rightarrow \text{Small value} \Rightarrow$$

(-ve)

Derivative of

If Relu output is [1] ⇒ Weight updation will happen

If Relu output is [0] ⇒ Dead Neuron

$$W_{2new} = W_{2old} - \eta \boxed{\frac{\partial h}{\partial w_{2old}}} \Rightarrow 0$$

If Derivative of Relu(z) is 0

$$\boxed{W_{2new} \approx W_{2old}} \Rightarrow \text{Dead Neuron}$$

If $z = +ve$ $\quad \frac{\partial \text{Relu}(z)}{\partial z} = 1$

If $z = -ve$ $\quad \frac{\partial \text{Relu}(z)}{\partial z} = 0$

Advantages

① Solving Vanishing Gradient Problem

② Max(0, x) → Calculation is superfast. The Relu function has a linear Relationship

③ It is much faster than Sigmoid or Tanh.

Disadvantages

① Dead Neuron

② Relu function O/P

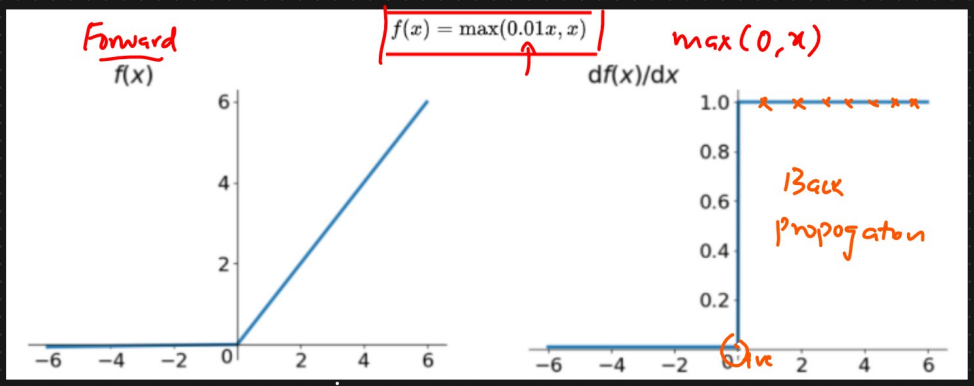
(0, x) ⇒ 0 or +ve number

⇓

It is not zero centric

④ Leaky Relu And Parametric Relu

max(αx, x) → hyperparameter α = alpha = 0.01, 0.02, 0.03



Relu ⟶ Dead Neuron ⟶ Dead Relu Problem

Advantages

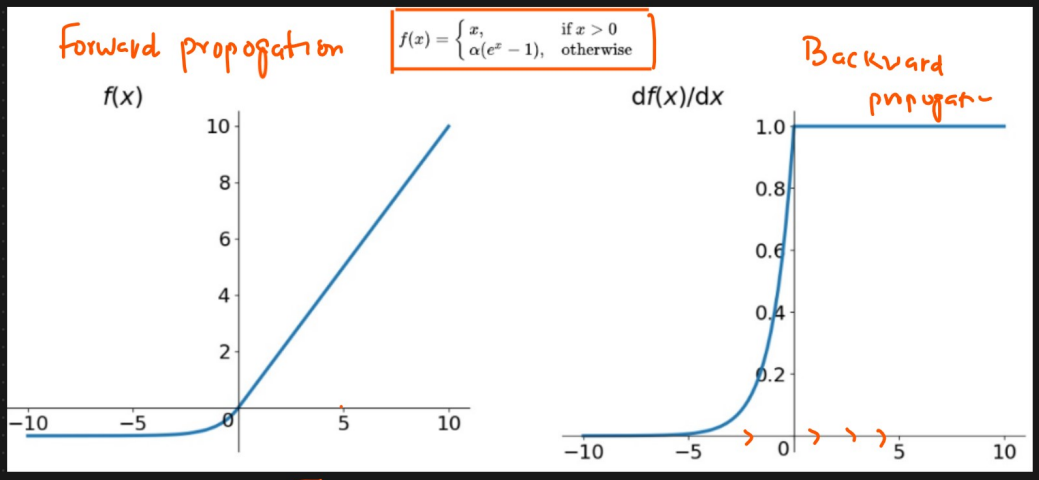
① Leaky Relu has all the advantage of Relu

② It removes the Dead Relu Problem

Disadvantage

① It is not zero centric

⑤ ELU (Exponential Linear Units)



[It is used to solve Relu problems]

Advantages

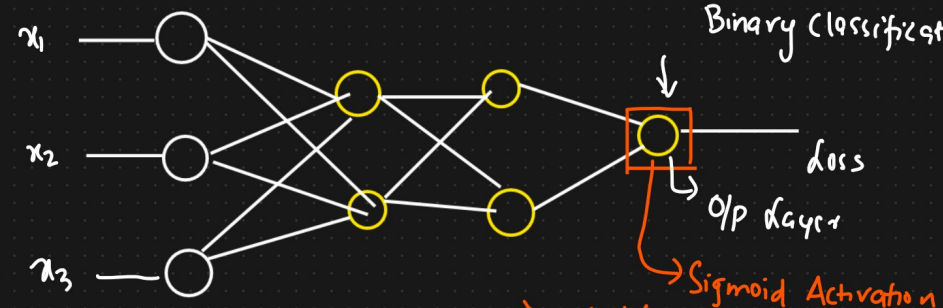
① No Dead Relu Issues

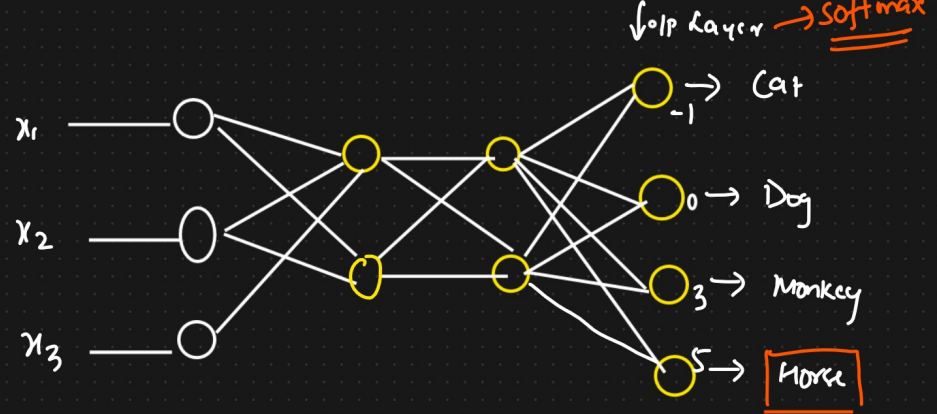
② Zero centered

Disadvantage

1) Slightly more computationally Intensive.

⑥ Softmax Activation function [Multiclass classification Problem]





Softmax Activation

$$\text{Softmax} = \frac{e^{y_i}}{\sum_{K=0}^{n} e^{y_K}}$$

$y_i = 0 * w + b$

Softmax $\Rightarrow$ Cat $= \frac{e^{-1}}{e^{-1+0+3+5}} = 0.00033$

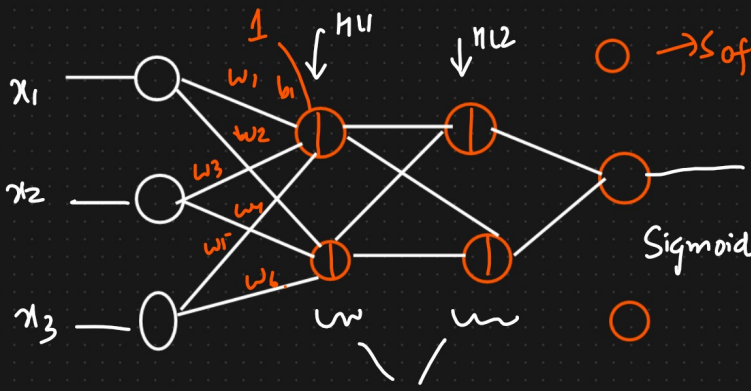
Dog $= \frac{e^{0}}{e^{-1+0+3+5}} = 0.0024$

Monkey $= \frac{e^{3}}{e^{-1+0+3+5}} = 0.0183$

Horse $= \frac{e^{5}}{e^{-1+0+3+5}} = 0.1353$

$Pr(\text{Horse}) = \frac{0.1353}{0.00033 + 0.0024 + 0.0183 + 0.1353} \approx 86\%$

⑦ Which Activation Function To Use When ?



↓ O/P layer

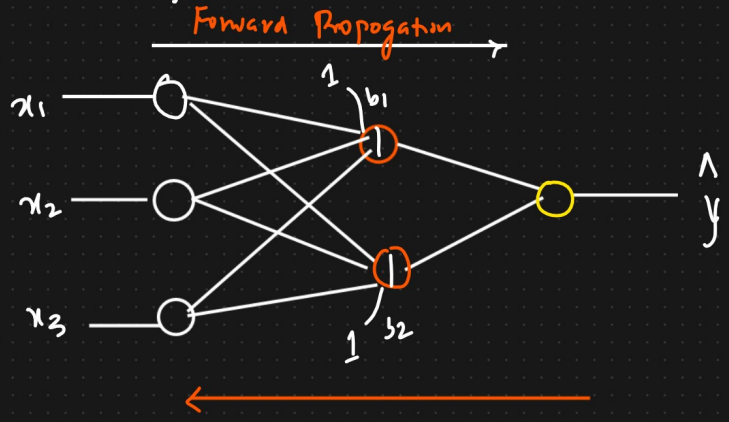
→ Softmax ⇒ ⇓ Multiclass Classification

Sigmoid → Binary classification

Hidden layers → Sigmoid or Tanh ⇓ Vanishing Gradient Problem

Relu or its variant Prelu, Leaky Relu, Elu

Loss function And Cost Function



$$\text{Cost fn} = \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \quad \text{(Error)}$$

$$\text{Loss} = (y-\hat{y})^2$$

Optimizers

Gradient Descent

x_1	x_2	x_3	$\hat{y}$ O/P
-	-	-	0
-	-	-	1
-	-	-	0
-	-	-	1

Loss function

$$MSE = (y-\hat{y})^2$$

Costfunction

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

ANN

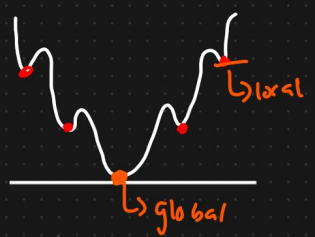
- Classification
- Regression
  1. Mean Squared Error (MSE)
  2. Mean Absolute Error (MAE)
  3. Huber Loss
  4. RMSE

1. Mean Squared Error (MSE)

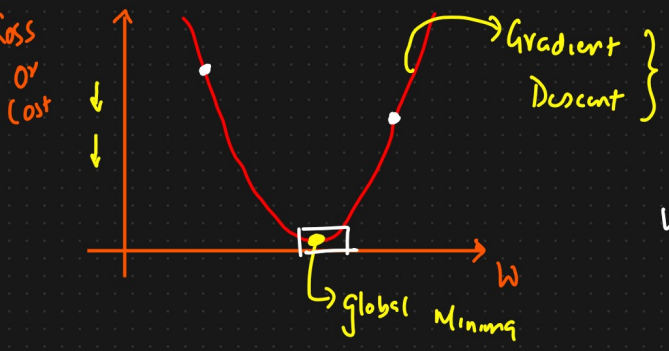
$$\text{loss function} = (y-\hat{y})^2$$

$$\text{Cost fn} = \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

↳ Quadratic Equations



Weight updation



$$W_{new} = W_{old} - \eta \left[\frac{\partial h}{\partial W_{old}} \right]$$

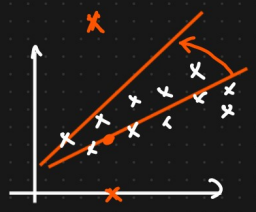
Slope

Advantages

1. MSE is Differentiable
2. It has 1 local or global Minima
3. It converges faster

Disadvantages

1. Not Robust to outliers



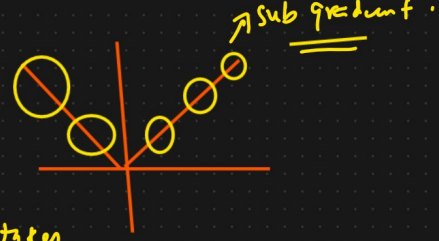
penalizing the Error

Cost function = $\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

2. Mean Absolute Error (MAE)

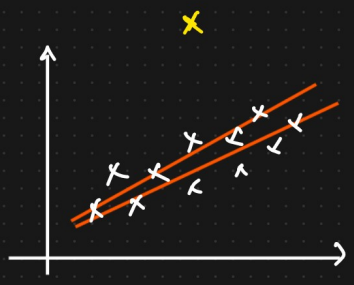
Loss fn : $|y - \hat{y}|$

Cost fn = $\frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$



Advantages

1. Robust to outliers



Disadvantages

1. Convergence usually takes time in MAE

3. Huber loss

1. MSE
2. MAE

Cost fn = $\begin{cases} \frac{1}{n} \sum_{i=1}^{n} (y - \hat{y})^2 & \text{if } |y - \hat{y}| \leq \delta \end{cases}$

MSE — No outlier — Hyperparameter

$$\begin{cases} \delta|y-\hat{y}| - \frac{1}{2}\delta^2, & \text{otherwise} \end{cases}$$

⇓

MAE

④ RMSE (Root Mean Squared Error)

$$\text{Cost function} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}}$$

Advantages

Disadvantages

② Loss or Cost function For Classification Problems

Classification ⟶ Cross Entropy
- ⟶ Binary Cross Entropy {Binary}
- ⟶ Categorical Cross Entropy {Multiclass}
- ⟶ Sparse Categorical Cross Entropy
  ↳ {Multiclass}

① Binary Cross Entropy

Log Loss

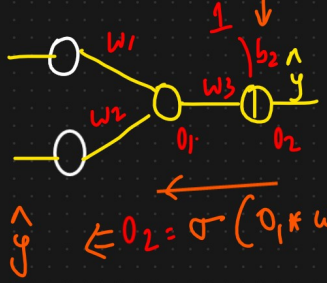
⇓

$$\text{loss} = -y * \log(\hat{y}) - (1-y) * \log(1-\hat{y})$$

y = Actual value

$\hat{y}$ = Predicted value

$$\text{loss} = \begin{cases} -\log(1-\hat{y}) & \text{if } y=0 \\ -\log(\hat{y}) & \text{if } y=1 \end{cases}$$



Binary classification

Sigmoid

$$\hat{y} = \frac{1}{1+e^{-z}}$$ ⇒ Sigmoid Activation function

② Categorical CROSS Entropy (Multiclass classification)

ONE → One Hot Encoding

	f_1	f_2	f_3	O/P	Good (j:1)	Bad (j=2)	Neutral (j=3)
→	2	3	4	Good	1	0	0
→	5	6	7	Bad	0	1	0
→	8	9	10	Neutral	0	0	1

C = No of categories

i : 1 to n

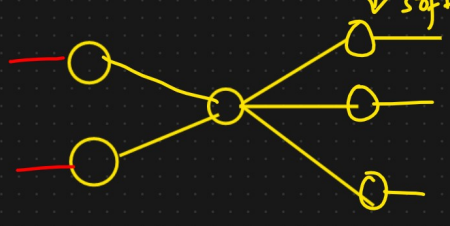
$$L(x_i, y_i) = -\sum_{j=1}^{C} y_{ij} * \ln(\hat{y}_{ij})$$

Actual value ⇐ $y_{ij} = [\, y_{11} \;\; y_{12} \;\; y_{13} \cdots y_{1C} \,]$

$[\, y_{21}, \; y_{22}, \; y_{23} \cdots y_{2C} \,]$

$$y_{ij} = \begin{cases} 1 & \text{if the element is in the class} \\ 0 & \text{Otherwise} \end{cases}$$

Prediction ⇐ $\hat{y}_{ij}$ ⇒ Softmax Activation $= Sof(z) = \dfrac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}}$



O/p $\hat{y}_{ij}$ = Probabilities

[0.1, 0.2, 0.3, 0.2, 0.2] ← 1

Categorical Cross Entropy ⇒ [0.2, 0.3, 0.5]

[This also gives the probability of other categories]

③ Sparse Categorical Cross Entropy

0 1st 2nd → Categories

[0.2, 0.3, 0.5]

2nd Index ⇒ O/p

0 1 2nd 3rd 4th

[0.2, 0.3, 0.1, 0.2, 0.2] ⇒ 1th Index ⇒ Category ⇒ O/p

Disadvantage

① Loosing info about the probability of other category.

④ Right Combination

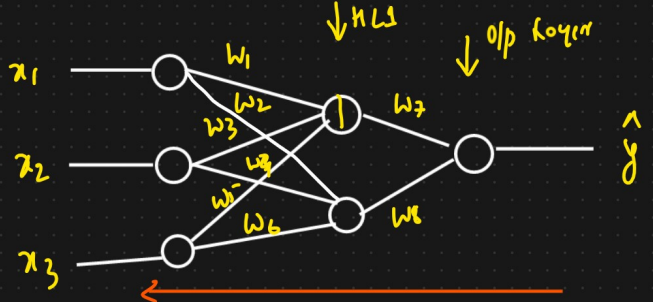
Activation applied ⇓

	Hidden Layers	O/p Layer	Problem Statement	Loss function
①	Relu or its Variants	Sigmoid	Binary classification	Binary Cross Entropy
②	Relu or its Variants	Softmax	Multi class	Categorical or Sparse CE
③	Relu or its Variants	linear	Regression	MSE, MAE, Huber loss, RMSE

Optimizers

1. Gradient Descent
2. Stochastic Gradient Descent (SGD)
3. Mini batch SGD
4. SGD With Momentum
5. Adagrad and RMSPROP
6. Adam Optimizers

1. Gradient Descent Optimizer



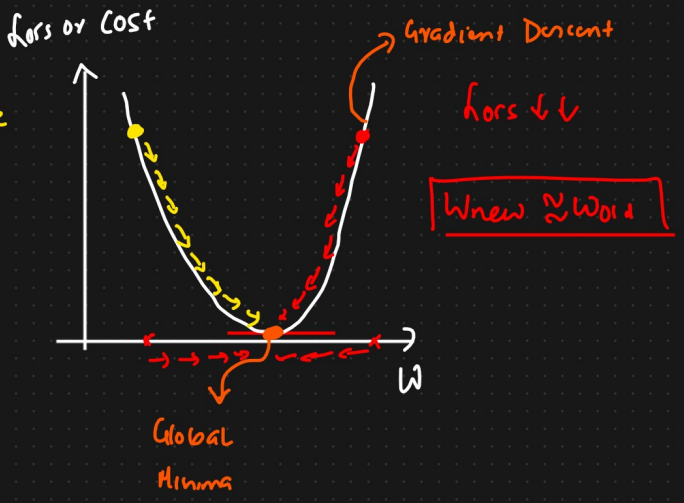
Backward Propogation

$$Loss = [\quad] \downarrow\downarrow\downarrow$$

Optimizers

Learning Rate

$$W_{new} = W_{old} - \eta \frac{\partial L}{\partial W_{old}}$$



MSE

$$\text{Loss fn} = (y-\hat{y})^2 \qquad \text{Cost fn} = \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

Dataset = 1000 Data points

Epochs, Iteration

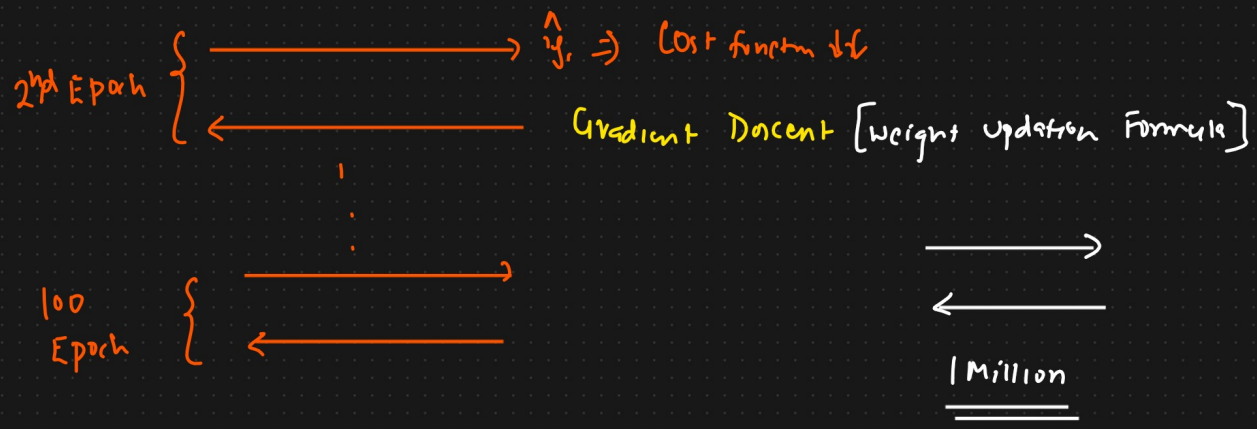
1 Epoch = 1 Iteration

10 Iteration

$100/10 = 100$

1 Epoch { 1000 datapoint → $\hat{y}_i$ ⇒ Cost function ↓↓ ; Weights will get updated ← } Gradient Descent optimizer [Weight updation]

100



Advantages

① Convergence will happen.

Disadvantage

① Huge Resource RAM, GPU

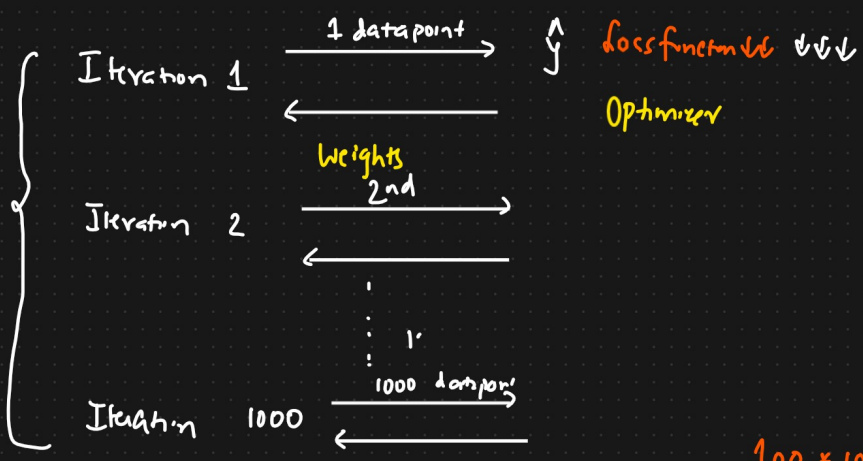
⇓

Resource Intensive.

② Stochastic Gradient Descent (SGD)

1000 datapoint

Epochs & Iteration



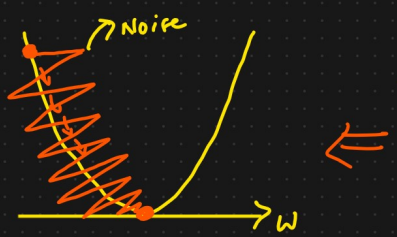
100 × 1000 = 100000

Epochs

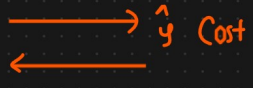
100000 datapoint

Advantage

Ⓐ Solve Resource Issue



Disadvantage



Ⓐ Time Complexity ↑↑

Ⓐ Convergence will also take more time.

Ⓐ Noise gets Introduced

③ Mini Batch SGD

No. of iterations = $\frac{100000}{1000}$ = 100 iterations

Epoch, Iteration, Batch_size

Datapoints = 100000

batch_size = 1000

MSE

Cost fn = $\sum_{i=1}^{1000} (y_i - \hat{y}_i)^2$ ↓↓

Epoch 1

Iteration 1 — 1000 →

← Optimizer ⇒ Mini Batch SGD

change the weights

Iteration 2 — 1000 →

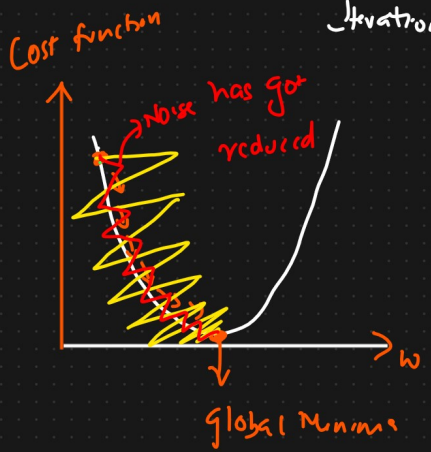
↓↓

8gb	1
16gb	5000

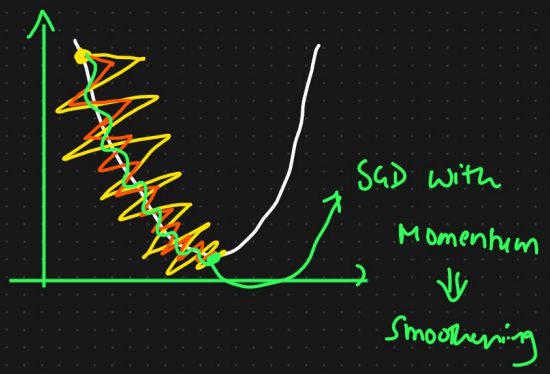
Iteration 3

!

Iteration 100



Batch Size



Advantages

① Convergence speed will increase =

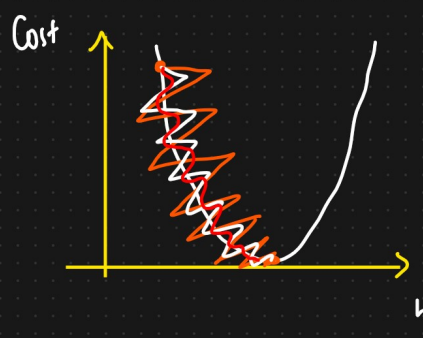
② Noise will be less when compared to SGD

③ Efficient Resource Usage (RAM)

Disadvantage

① Noise still exists

④ SGD With Momentum



Weight Updation formula

learning rate

$$W_{new} = W_{old} - \eta \frac{\partial h}{\partial W_{old}}$$

$$b_{new} = b_{old} - \eta \frac{\partial h}{\partial b_{old}}$$

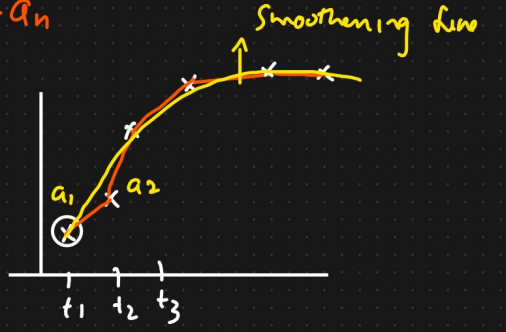
$$W_t = W_{t-1} - \eta \frac{\partial h}{\partial W_{t-1}}$$

Exponential Weight Average. {Smoothening} ⇒ ARIMA, SARIMAX

Time = t_1 t_2 t_3 t_4 - - - t_n

Values = a_1 a_2 a_3 a_4 - - - - a_n

Time Series



$$V_{t_1} = a_1$$

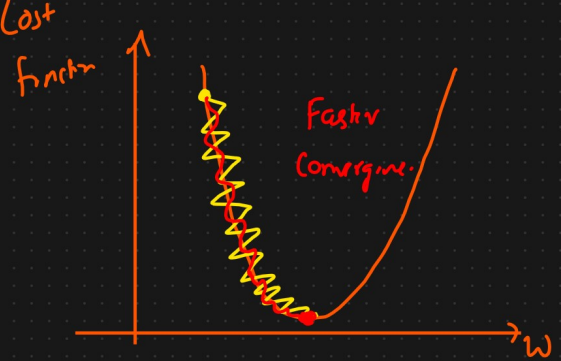
$$V_{t_2} = \boxed{\beta} * V_{t_1} + (1-\beta) * a_2$$

$\beta = 0$ to 1

$$= 0.95 * a_1 + (0.05)\, a_2$$

$$V_{t_3} = \beta * V_{t_2} + (1-\beta) * a_3$$

$$= 0.95 \left[0.95 * a_1 + (0.05)\, a_2 \right] + (0.05) * a_3$$



Advantage

① Reduces the noise

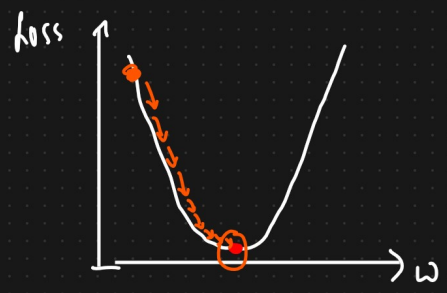
② Quick Convergence

⑤ Adagrad : Adaptive Gradient Descent

η = fixed ⇒ Dynamic learning Rate

$$W_t = W_{t-1} - \eta \frac{\partial h}{\partial w_{t-1}}$$

→ Learning Rate = 0.001



As the convergence happen the learning rate should change

$0.00001 \approx 0$

$$W_t = W_{t-1} - \eta' \frac{\partial h}{\partial w_{t-1}}$$

$$\eta' = \frac{\eta}{\sqrt{\alpha_t + \epsilon}}$$

$\eta' \downarrow\downarrow$ $\alpha_t \uparrow\uparrow$

ϵ → Epsilon ≈ small value

$\alpha_t \uparrow\uparrow$

$$\alpha_t = \sum_{i=1}^{t} \left(\frac{\partial h}{\partial w_t}\right)^2$$

$W_t \approx W_{t-1}$

$t=1 \rightarrow t=2 \rightarrow t=3$

$\eta = 0.01$ $\eta = 0.005$ $\eta = 0.003$

Disadvantage

① η' → Possibility to become a very small value ≈ 0

⑥ Adadelta And RMSPROP

Exponential Weight Average

$\beta = 0.95$ $\quad Sdw_t = 0$

$$Sdw_t = \beta * Sdw_{t-1} + (1-\beta)\left(\frac{\partial h}{\partial w_{t-1}}\right)^2$$

$$\eta' = \frac{\eta}{\sqrt{Sdw_t + \epsilon}}$$

[Dynamic LR + Smoothening EWA]

$$\boxed{W_t = W_{t-1} - \eta' \frac{\partial h}{\partial w_{t-1}}}$$

⑦ Adam Optimizer

SGD With Momentum + RMSPROP [Dynamic LR + Smoothening]

$$W_t = W_{t-1} - \eta' V dw$$

Weight Updation

$$b_t = b_{t-1} - \eta' V db$$

→ Bias Updation

$$\eta' = \frac{\eta}{\sqrt{Sdw_t + \epsilon}}$$

$Sdw_t = 0$

EWA

$$Sdw_t = \beta * Sdw_{t-1} + (1-\beta)\left(\frac{\partial h}{\partial w_{t-1}}\right)^2$$

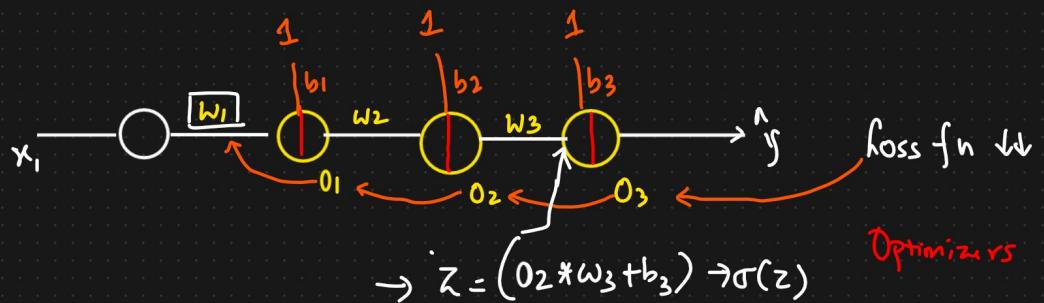
[EWA]

$$Vdw_t = \beta * Vdw_{t-1} + (1-\beta)\frac{\partial h}{\partial w_{t-1}}$$

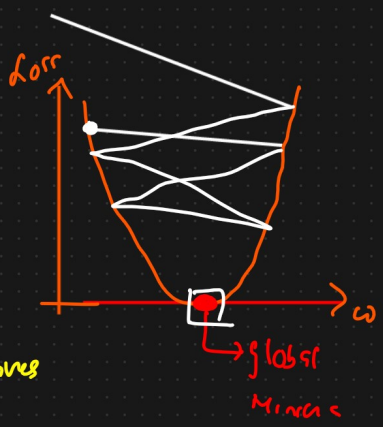
$$Vdb_t = \beta * Vdb_{t-1} + (1-\beta)\frac{\partial h}{\partial b_{t-1}}$$

⇒ Momentum
↳ Smoothening

Exploding Gradient Problem ⇒ Weight Initialization Technique



Optimizers



$$W_{new} = W_{old} - \left[\eta \frac{\partial h}{\partial W_{old}}\right]$$

$\{ W_{new} >>>> W_{old} \quad W_{new} <<<< W_{old} \}$

Chain Rule of Derivatives

$$\frac{\partial h}{\partial W_{old}} = \frac{\partial h}{\partial O_3} * \boxed{\frac{\partial O_3}{\partial O_2}} * \frac{\partial O_2}{\partial O_1} * \frac{\partial O_1}{\partial w_{old}}$$

big * big * big * big ⇒ big value

Weight initialization ⇓ Very high value

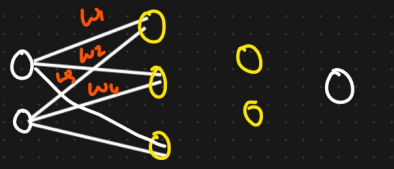
$$\frac{\partial O_3}{\partial O_2} = \boxed{\frac{\partial \sigma(z)}{\partial z}} * \frac{\partial z}{\partial O_2}$$

$$= [0 - 0.25] * \frac{\partial (O_2 * w_3 + b_3)}{\partial O_2}$$

$$= [0 - 0.25] * w_3 \Rightarrow 500 - 1000$$

Weight Initializing Techniques

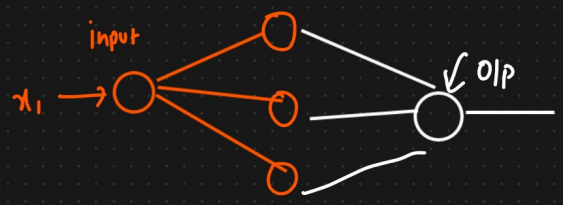
1. Uniform Distribution ✓
2. Xavier / Glorot Initialization ✓
3. Kaiming He Initialization ✓



Key Points

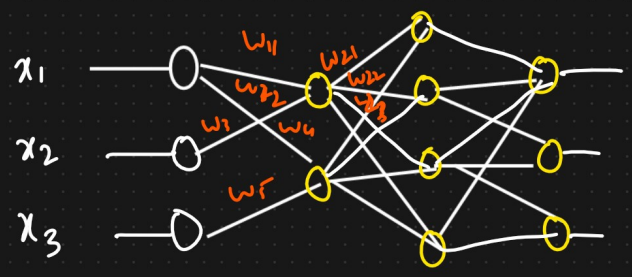
1. Weights should be small ✓
2. Weights should not be same ✓
3. Weights should have good variance

} ✓



input = 1

Output = 1



input = 3

O/p = 3

① Uniform Distribution

$W_{ij} \sim$ Uniform Distribution $\left[\frac{-1}{\sqrt{input}}, \frac{1}{\sqrt{input}}\right]$

$\left[\frac{-1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right]$

② Xavier/Glorot Initialization

Researcher $\longrightarrow$ Xavier Glorot

① Xavier Normal Init

$W_{ij} \sim N(0, \sigma)$

$$\sigma = \sqrt{\frac{2}{(input + output)}}$$

② Xavier Uniform

$W_{ij} \sim$ Uniform Distribution $\left[\frac{-\sqrt{6}}{\sqrt{input + output}}, \frac{\sqrt{6}}{\sqrt{i/p + o/p}}\right]$

③ Kaiming HC Initialization

① He Normal

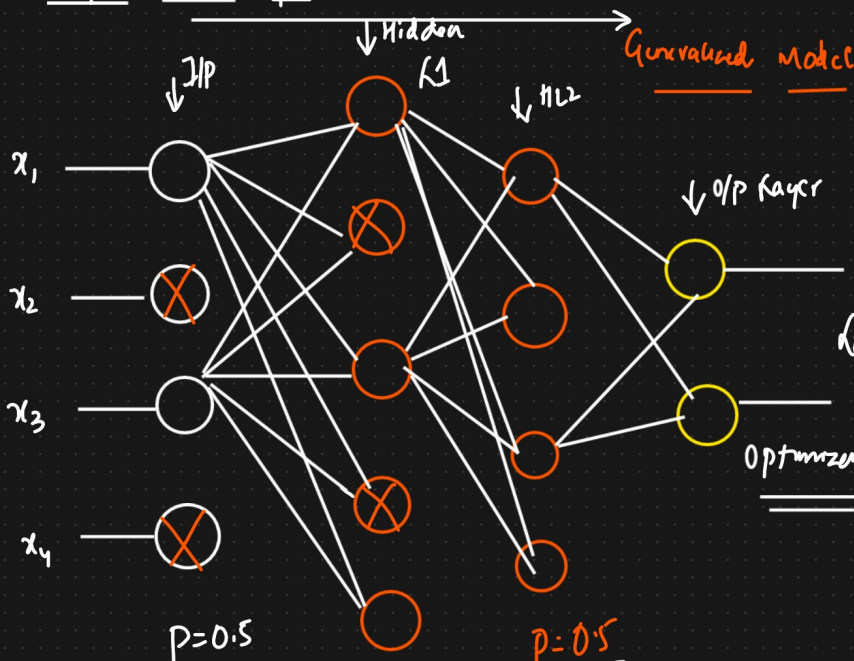
$$W_{ij} \sim N(0, \sigma)$$

$$\sigma = \sqrt{\frac{2}{input}}$$

② He uniform

$$W_{ij} \sim \text{Uniform Dimbution} \left[-\sqrt{\frac{6}{input}}, \sqrt{\frac{6}{input}} \right]$$

Drop Out Layer



Generalized Model ∈ ⇒ Overfitting

Model → Training Acc ⇒ 90%
Test Acc ⇒ 60%

Optimizers

Machine Learning

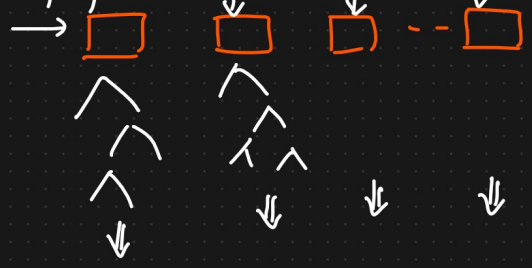
DT

⇒ Overfitting

Generalized Model

Random Forest

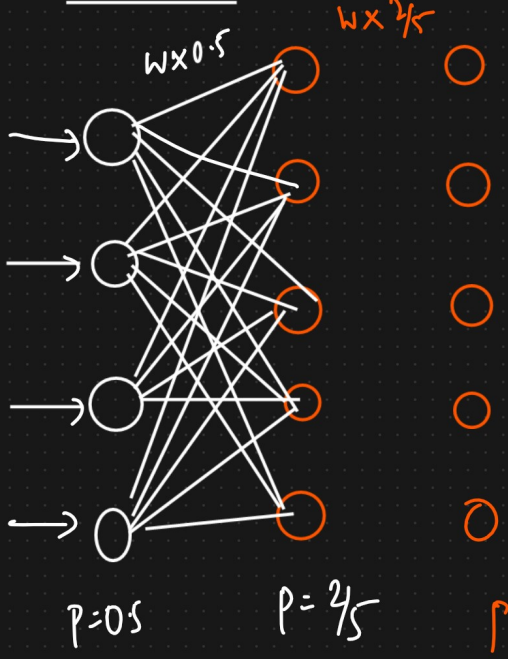
Feature Sampling
Row Sampling



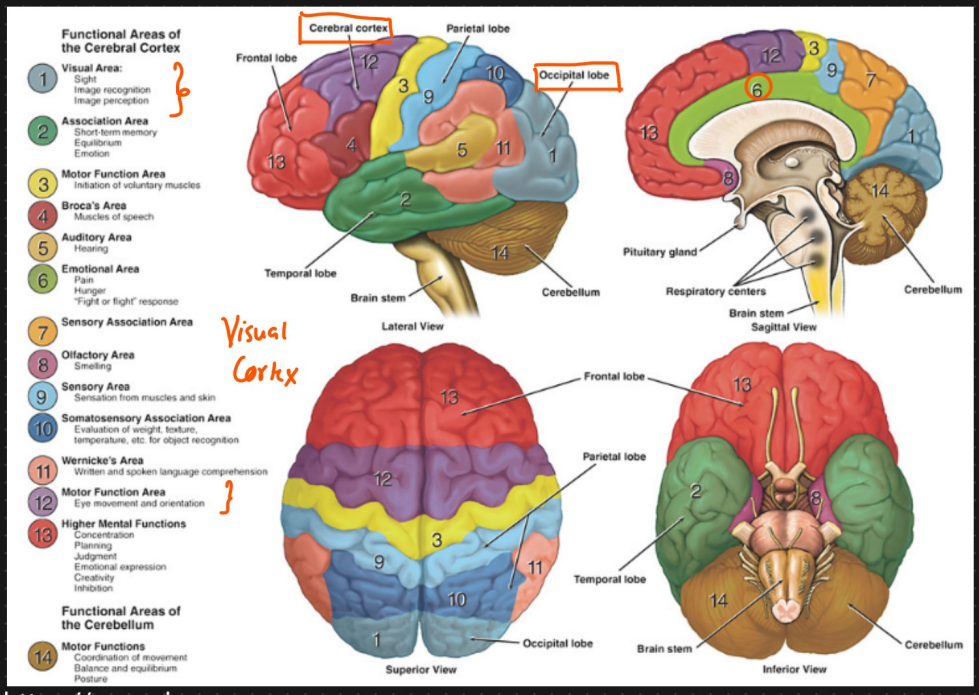
$0 \leq p \leq 1$

Hyperparameter

Test data



Convolutional Neural N/w



https://www.dana.org

① ANN → Supervised Learning → [Classification, Regression]

Dataset : I/p features O/p

② CNN : I/p => Images Eg: Image classification, Object Detection, Segmentation

② Cerebral Cortex And Visual Cortex

Visual Cortex (V1-V5) [Region of the brain that receives, integrates and processes visual information relayed from the retinas].

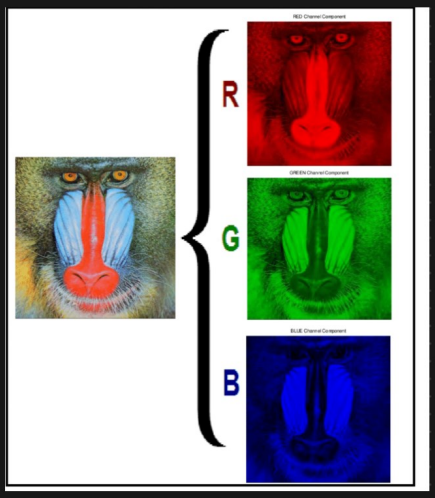
→ V1 → Primary Visual Cortex [Orientation of edges And Lines]

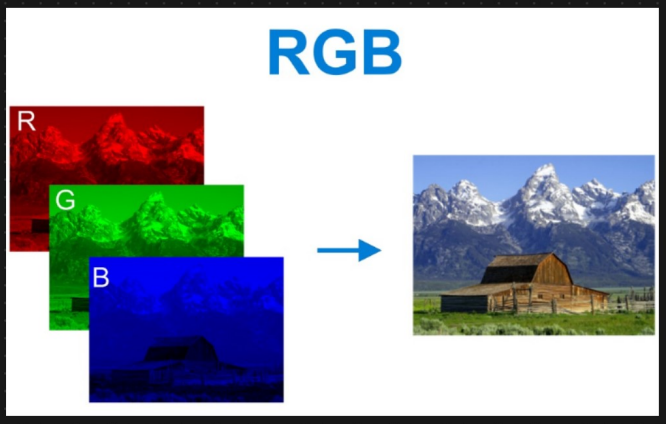
V2 [Differences in color, complex patterns, Object orientation]

V3 V4 V5 [Object Recognition]

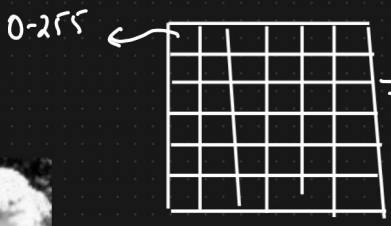
Visualize the Image

③ RGB Images And Gray Scale Images

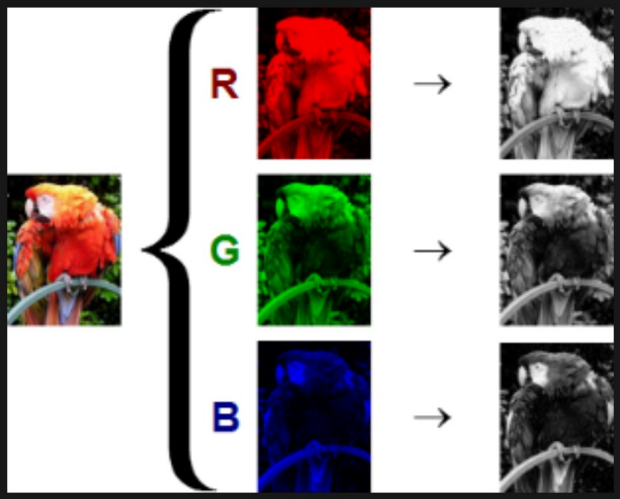




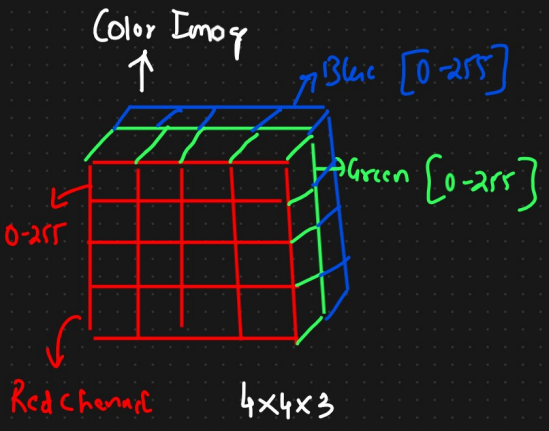
https://www.researchgate.net/



6x6x1



https://commons.wikimedia.org/



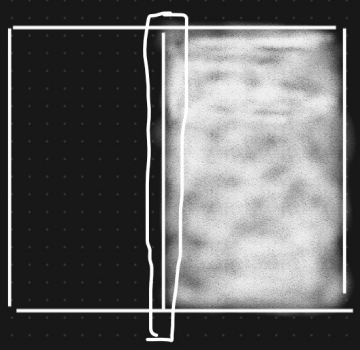
4x4x3

④ Convolution Operation In CNN

→ (0,1)

0	0	0	255	255	255
0	0	0	255	255	255
0	0	0	255	255	255
0	0	0	255	255	255
0	0	0	255	255	255
0	0	0	255	255	255

6x6x1



Step 1

(1) Normalize

Divide by 255

+1	+2	+1
0	0	0
-1	-2	-1

Convolution operation

Stride=1

0	0	0	1	1	1
0	0	0	1	1	1
0	0	0	1	1	1
0	0	0	1	1	1
0	0	0	1	1	1
0	0	0	1	1	1

6 x 6 x 1

n = 6

*

f = 3

+1	0	-1
+2	0	-2
+1	0	-1

3 x 3

filters

Vertical edge filters

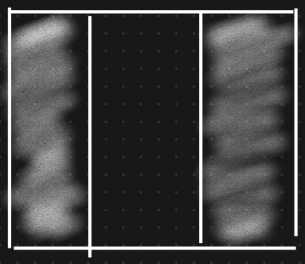
=

o/p = 4

0	-4	-4	0
0	-4	-4	0
0	-4	-4	0
0	-4	-4	0

4 x 4

$n - f + 1 = 6 - 3 + 1 = 4$



⇐

2rr	0	0	2rr
2rr	0	0	2rr
2rr	0	0	2rr
2rr	0	0	2rr

⑤ Padding In CNN

① Zero padding

② Neighbour padding

0	0	0	0	1	1	1	1
0	0	0	0	1	1	1	1
0	0	0	0	1	1	1	1
0	0	0	0	1	1	1	
0	0	0	0	1	1	1	
0	0	0	0	1	1	1	
0	0	0	0	1	1	1	
0							

6x6 n=6 8x8

*

f = 3

+1	0	-1
+2	0	-2
+1	0	-1

=

	0	-4	-4	0
	0	-4	-4	0
	0	-4	-4	0
	0	-4	-4	0

6x6

$n - f + 2p + 1$

$6 - 3 + 2p + 1 = 6$

$3 + 2p + 1 = 6$

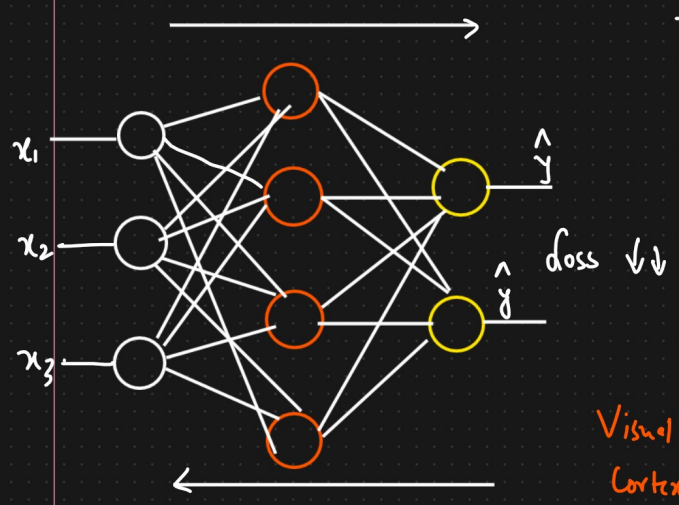
$2p = 6 - 4$

$p = 2/2 = 1$

7x7 3x3 7x7

How much padding you need to apply?

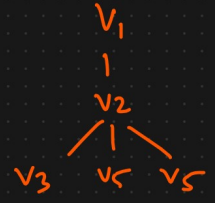
⑥ Operation of CNN Vs ANN



$$z = w^T x_i + b$$

$$relu(z)$$

Visual Cortex



Image

0	0	0	1	1	1
0	0	0	1	1	1
0	0	0	1	1	1
0	0	0	1	1	1
0	0	0	1	1	1
0	0	0	1	1	1

$*$

+1	0	-1
+2	0	-2
+1	0	-1

f_1

f_2

f_3

f_n

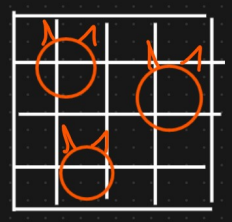
$=$

→ Relu operation $max(0, x)$

0	-4	-4	0
0	-4	-4	0
0	-4	-4	0
0	-4	-4	0

Convolution Layer

⑦ Max Pooling, Min Pooling, Mean Pooling



$\circledast$

2x2

P=1

S=1

filter

f_2

f_3

f_n

Convolution Layer

$=$

Relu

1	2	3
4	3	6
2	8	4

3x3

Stride = 2

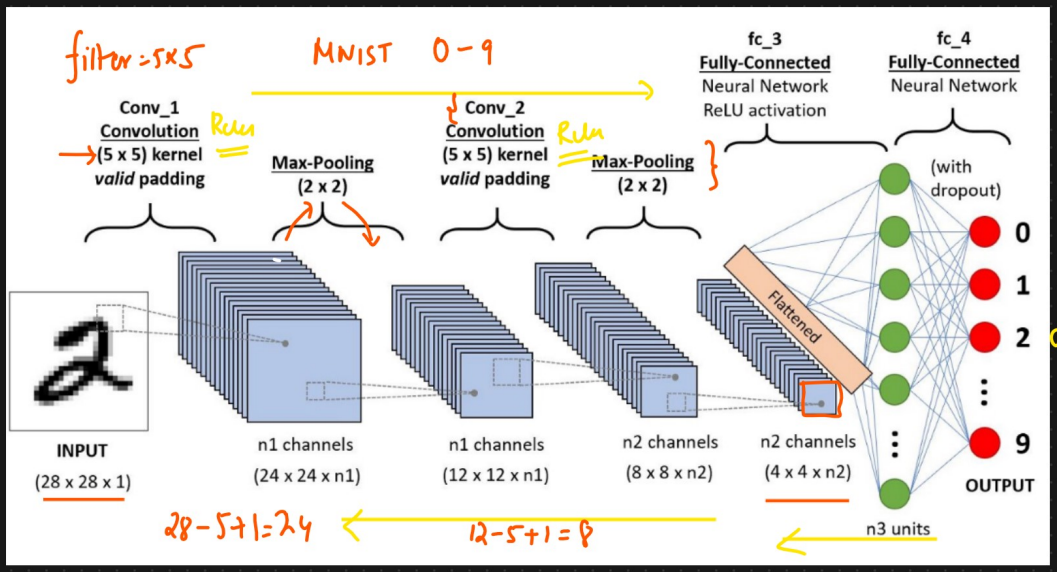
Max Pooling

2x2

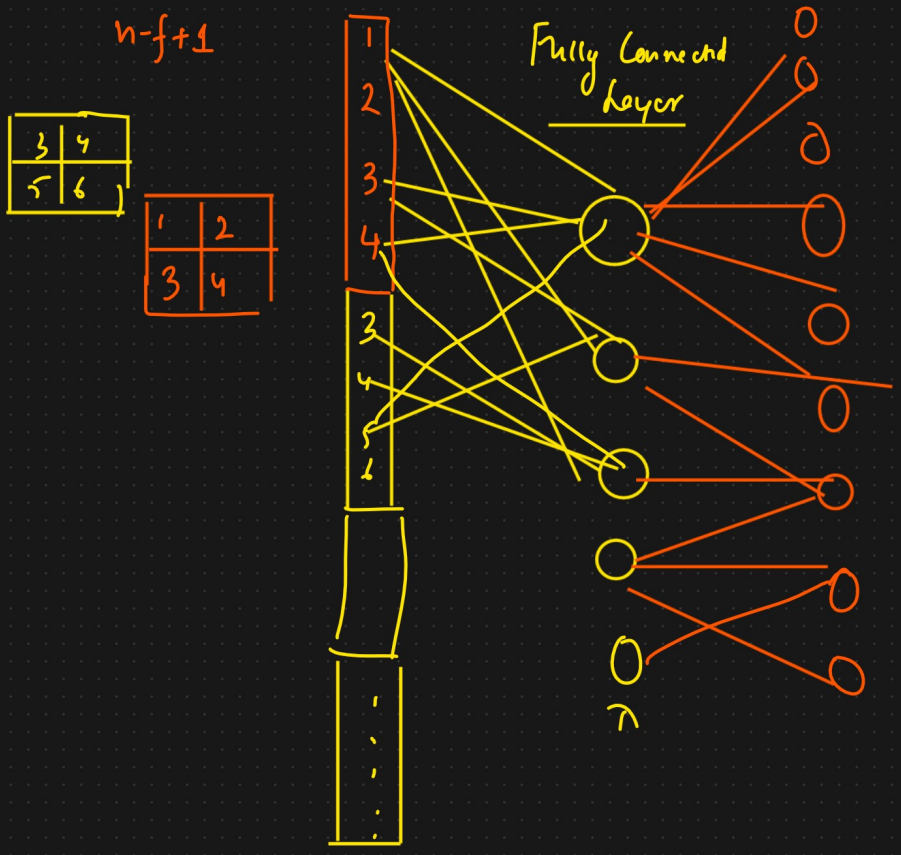
4	6
8	4

⑦ Location Invariant

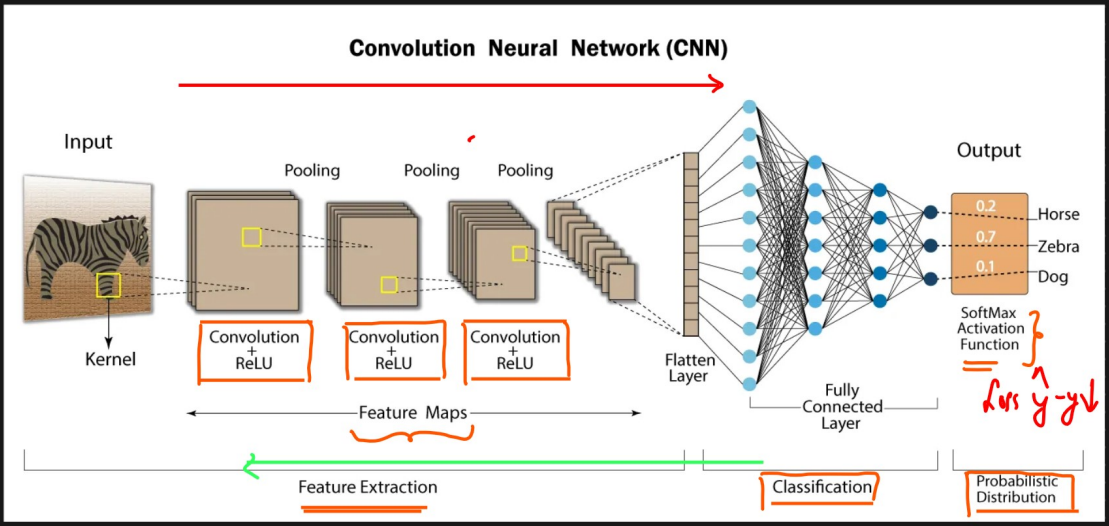
⑧ Fully Connected Layer In CNN [Flattened Layer]



Loss function ↓↓
Optimizer



⑨ CNN Complete Example



https://developersbreach.com/convolution-neural-network-deep-learning/

0	0	0			
0	0	0			
0	0	0			
0	0	0			
0	0	0			
0	0	0			

*

+1	0	-1
+2	0	-2
+1	0	-1

0	-4	-4	0
0	-4	-4	0
0	-4	-4	0
0	-4	-4	0

Day 6 - NLP {Recurrent Neural N/W}

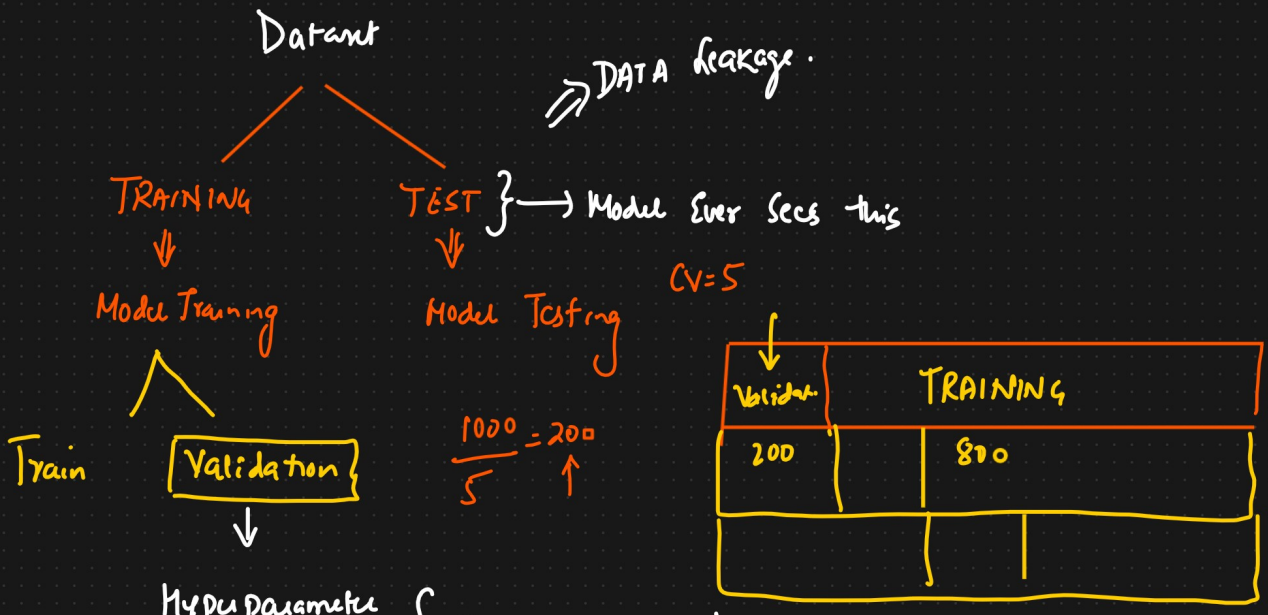
① BoW, TFIDF, Word2vec, Avg Word2Vec {Python Practical Implementation}

- CBOW
- Skip Gram

② Deep Learning → ① RNN ② LSTM RNN ③ GRU RNN ④ Bidirectional LSTM RNN ⑤ Encoders - Decoders ⑥ TRANSFORMERS ⑦ BERT

Interview Question

① Train Vs Test Vs Validation →



CROSS-VALIDATION

↓

GRIDSEARCHCV

RANDOMIZED SEARCH CV

② Why RandomForest instead of Decision Tree {Answer}.

DT → {Low Bias, High Variance}

↓

{ Low Bias Low Variance } → Random Forest }.

① Recurrent Neural N/W ⇒ Text → Vectors

Machine Learning

Word Embedding ← Word2vec, AvgWord2vec

Chatbot : Question and Answer → Sequence of words

Language Translation → Hindi → English

Text generation → A sentence → Suggestion → Completion of Sentences (Grammatically)

Gmail AutoSuggestion

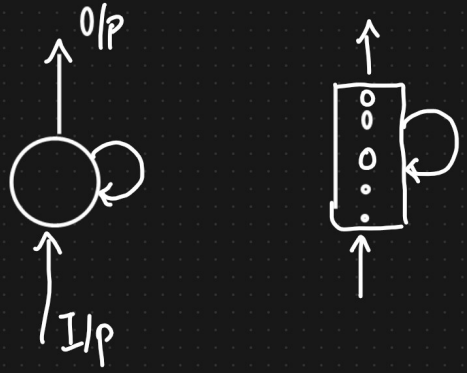
Deep Learning

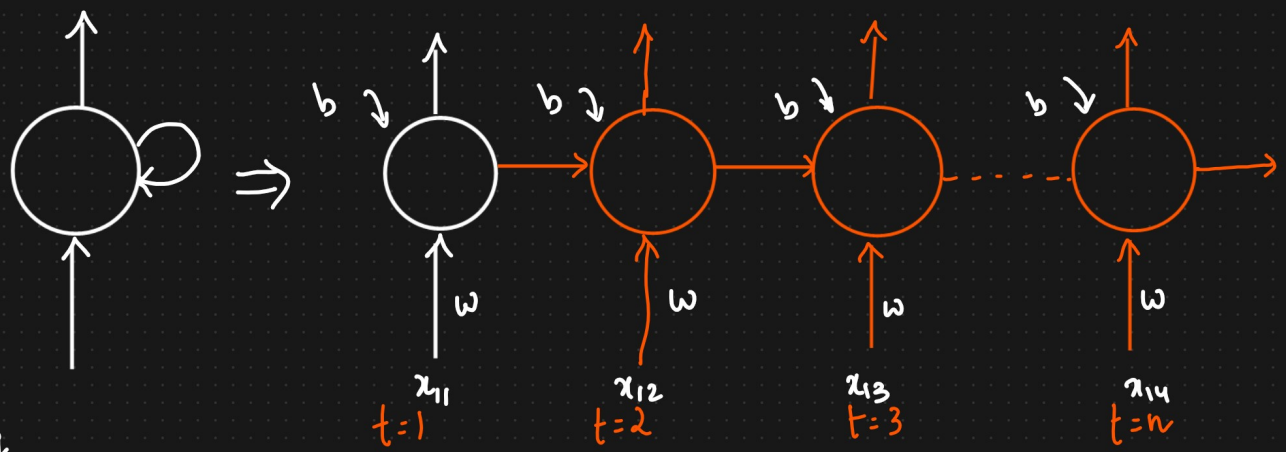
① RNN ② LSTM RNN ③ Transformers ④ BERT

Word2vec ⇒ Deep Learning

Words ⇒ Vec

① Recurrent Neural N/W





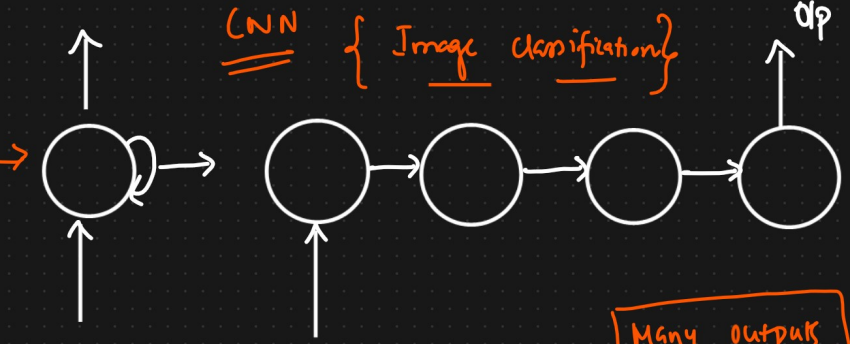
Sentiment Analysis → The food is good → Positive (O/p)

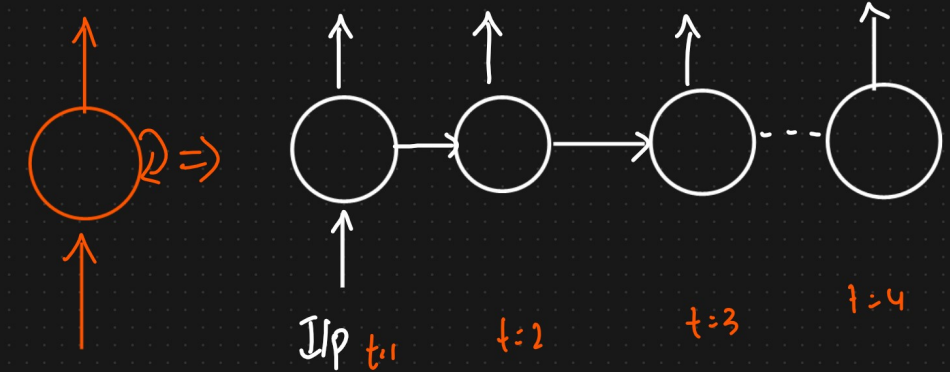
$\langle x_{11}, x_{12}, x_{13}, x_{14} \rangle$

x_{11} → Word2vec → Vectors → $d = 300$

Types of RNN

1. One to One RNN
2. One to Many RNN
3. Many to One RNN
4. Many to Many RNN



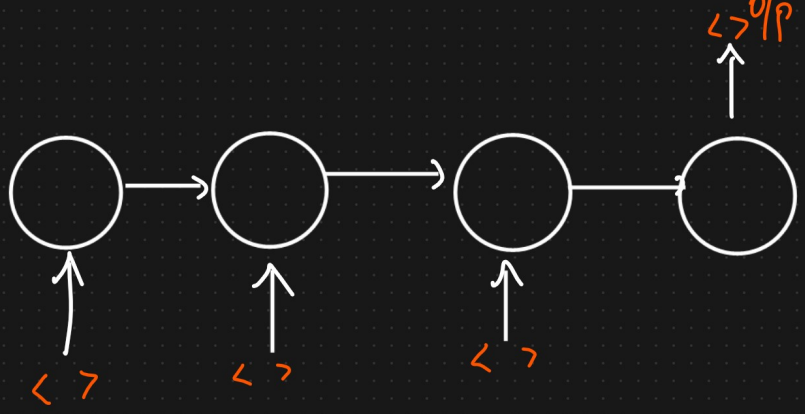


Eg: Music Generation, Text Generation

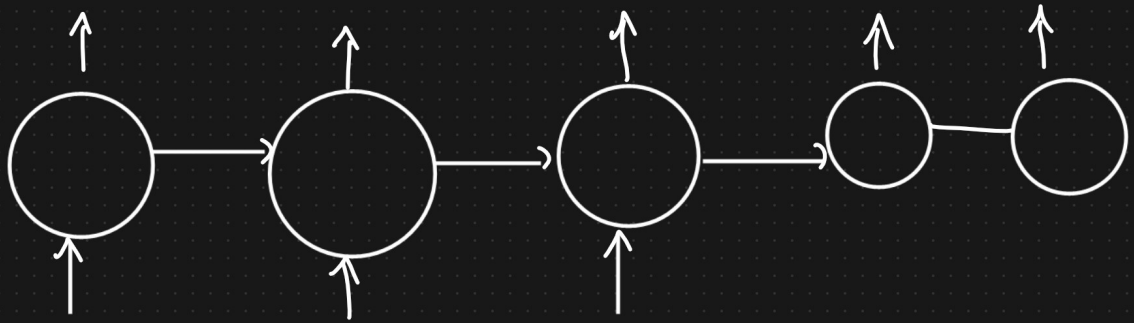
Google Search Suggestion, Movie Recommendation

③ Many to One

Sentiment Analysis



④ Many to Many



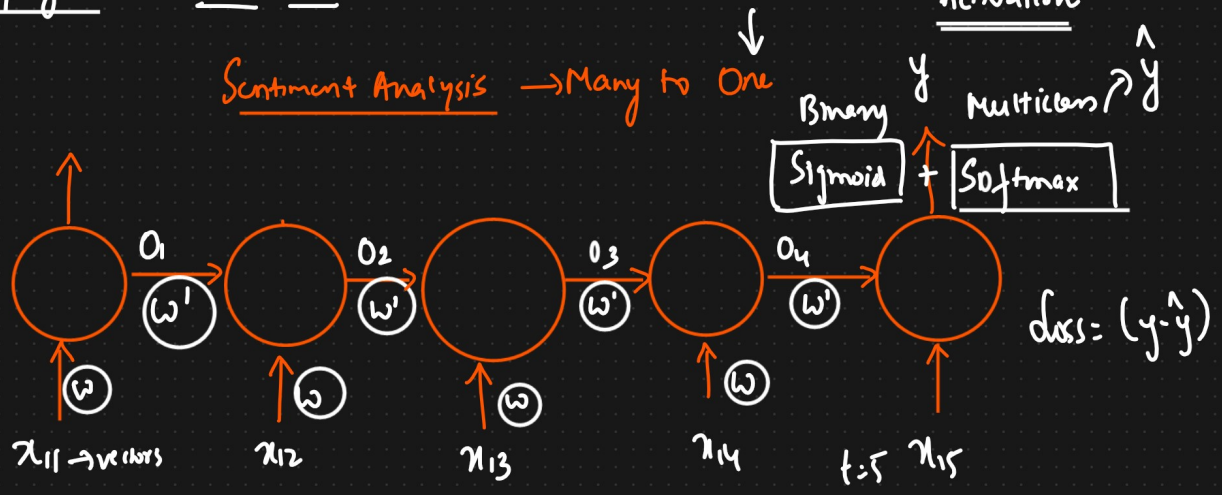
① Language Translation
② Question answers
③ Chatbots.

{ ① Forward Propogation
② Backward Propogation }

Time Series Data

① Forward Propogation In RNN

Sentiment Analysis → Many to One



t=1	t=2	t=3	t=4	O/P
The	food	is	very	good
x_{11}	x_{12}	x_{13}	x_{14}	x_{15}

Positive.

$O_1 = f(x_{11} * w)$

$O_2 = f(x_{12} * w + O_1 * w_1')$

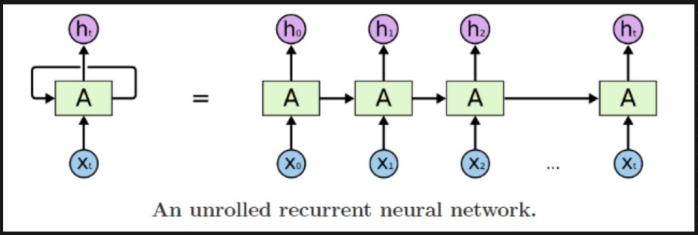
$O_3 = f(x_{13} * w + O_2 * w')$

LSTM RNN [Long Short Term Memory RNN]

RNN ⟶ Long Term Dependencies ⟶ Vanishing Gradient Problem

1. RNN ⟶ Problem ? ✓
2. Why LSTM RNN ? ✓ Basic Representation.
3. How LSTM RNN Works
   - Long Term Memory ✓
   - Short Term Memory ✓
4. LSTM Architecture
5. Working of LSTM RNN

Problems With RNN → Long Term Dependency



An unrolled recurrent neural network.

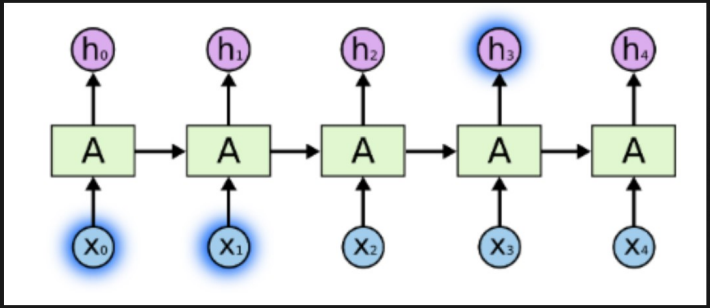
Vanishing Gradient Problem

Task:

Next Word In a Sentence

The color of the sky is blue

further context



Gap is less

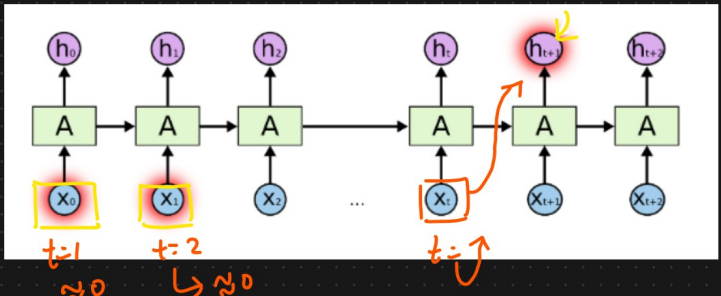
Huge gap O/p ← Context

I grew up in India .... I speak fluent

Context

Name of language

further Context



0-0.25

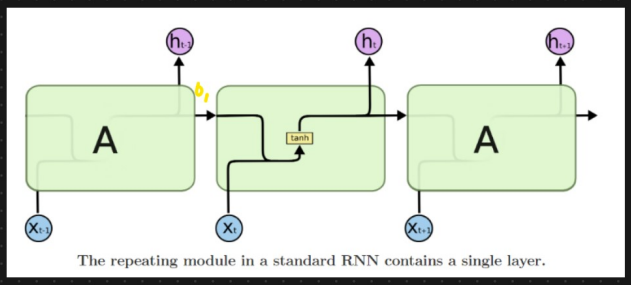
0-1

Huge Gap → Long Term Dependency

RNN → Long Term Dependency → Vanishing Gradient Problem

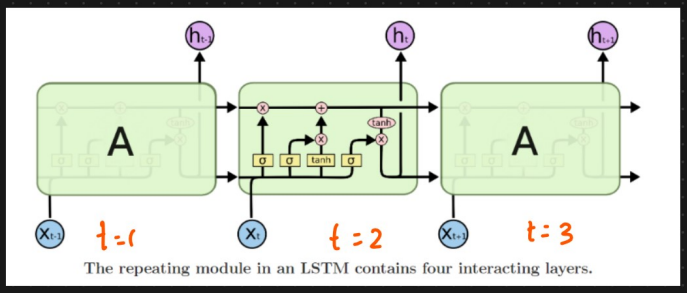
Chain Rule → ≈0.

Basic Representation of RNN And LSTM RNN

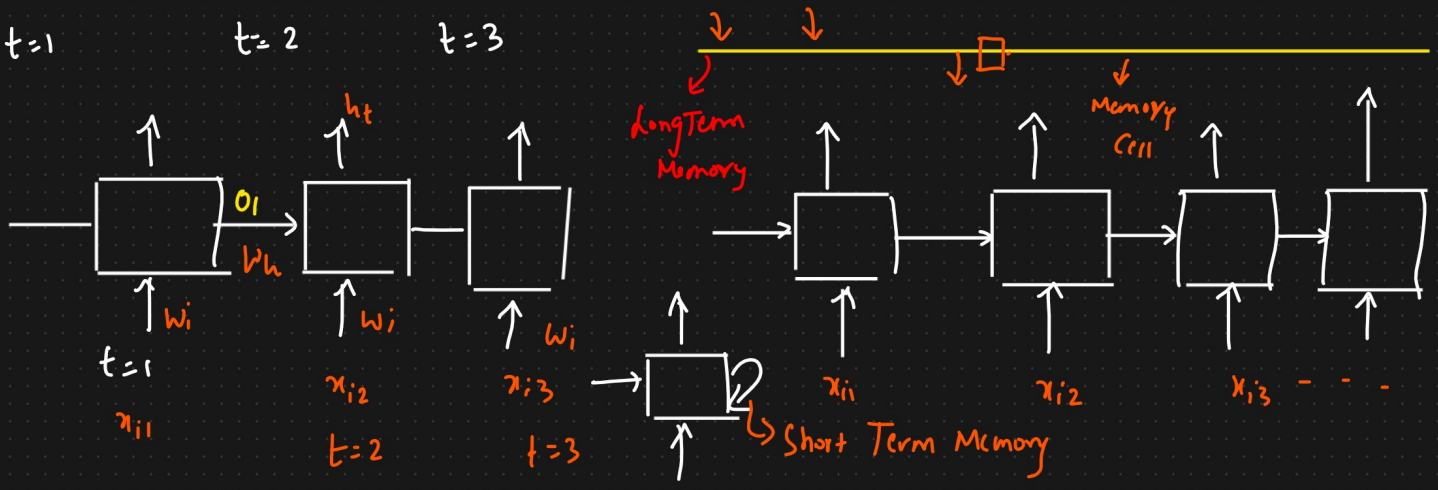


The repeating module in a standard RNN contains a single layer.

LSTM RNN



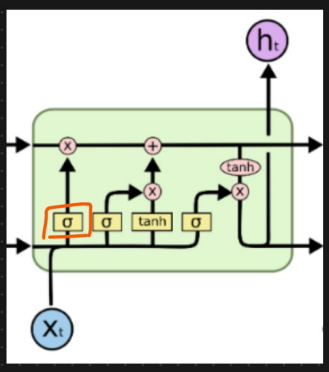
The repeating module in an LSTM contains four interacting layers.



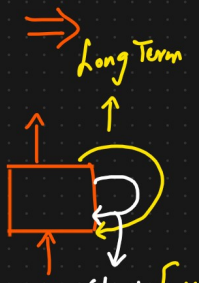
LSTM RNN → Long Term Memory
LSTM RNN → Short Term Memory

Conveyance But : Luggages

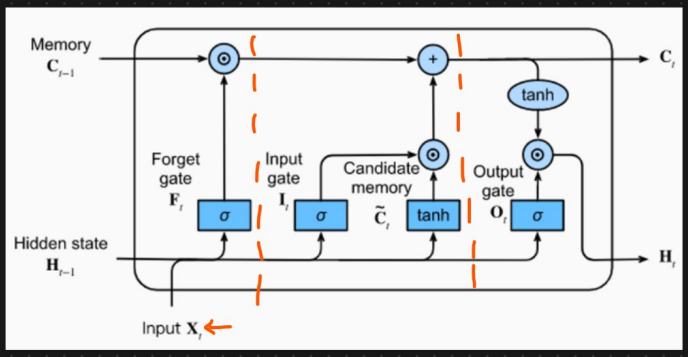
LSTM Architecture



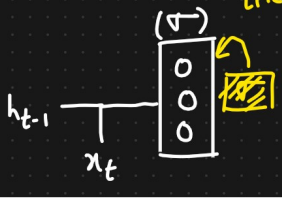
Basic Architecture

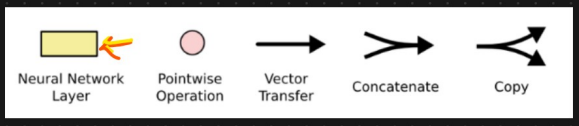


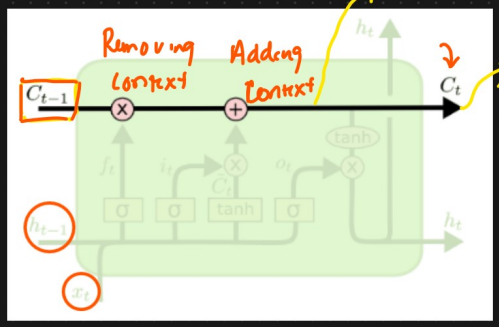
LSTM RNN



{Neural Nw Layer}





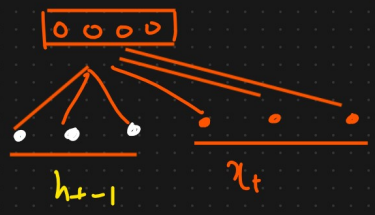


Long Term Memory ⊗ ⊕ tanh

Combining 2 vectors

$h_{t-1} = [1 \quad 2 \quad 3]$

$x_t = [2 \quad 3 \quad 4]$



Duplicate copy

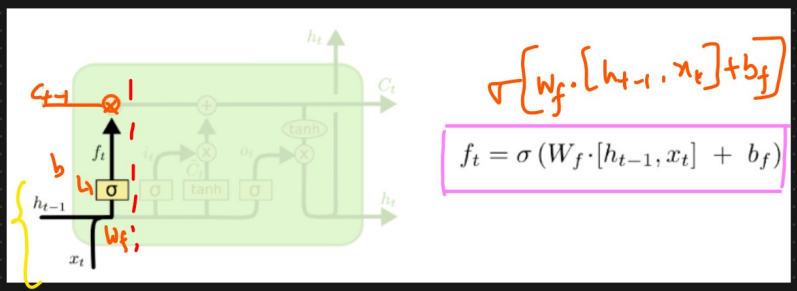
V1 $[1 \quad 2 \quad 3]$ ← tanh

V2 $[4 \quad 5 \quad 6]$

⊗ ⇒ [4 10 18]

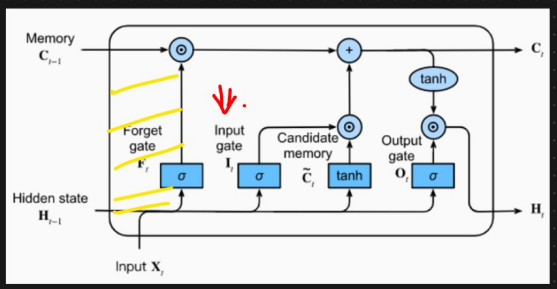
⊕ ⇒ [5 7 9]

tanh ⇒ [tanh(1) tanh(2) tanh(3)]



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right)$$

$\sigma\left[W_f \cdot [h_{t-1}, x_t] + b_f\right]$



Forget Gate

h_{t-1} = Hidden state of previous time stamp

x_t = Word passed as i/p in the current time stamp

Text — Next Word

$x_{11} \quad x_{12} \quad x_{13} \quad x_{14}$ — y_{15}

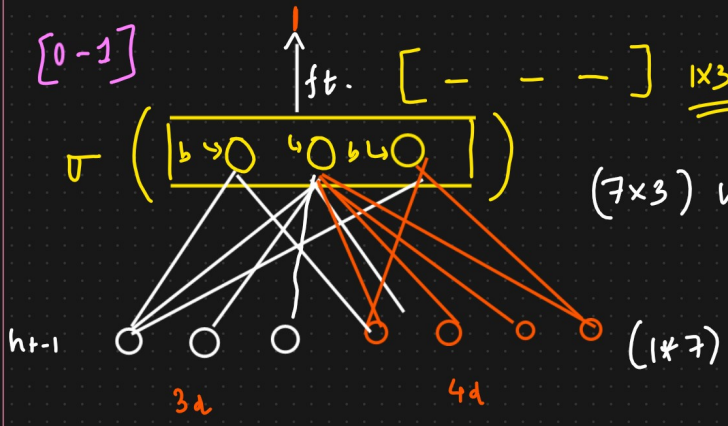
$[0 \quad 2 \quad 4 \quad 1]$ $[4 \quad 5 \quad 12]$ - - - - -

x_t x_{t+1}

Memory Cell

h_{t-1} $[1 \quad 2 \quad 4]$ 3d

C_{t-1} $[4 \quad 2 \quad 1]$ 3d ⇐



$$f_t = [0 \quad 0 \quad 0]$$ ← Complete Sentence

Context changed

① $C_{t-1} = [6 \quad 8 \quad 9]$ ⊗ $[0 \quad 0 \quad 0]$

$= [0 \quad 0 \quad 0]$ ← Removing all the previous Context

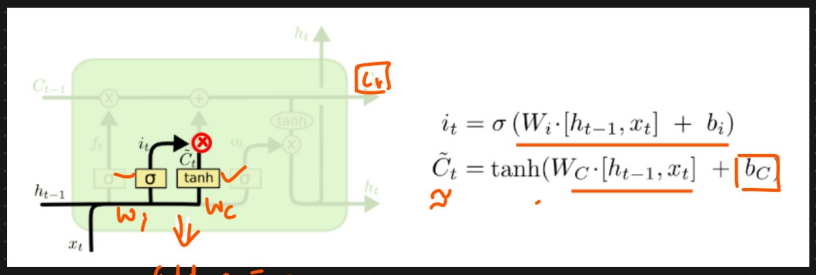
$f_t = [1 \quad 1 \quad 1]$

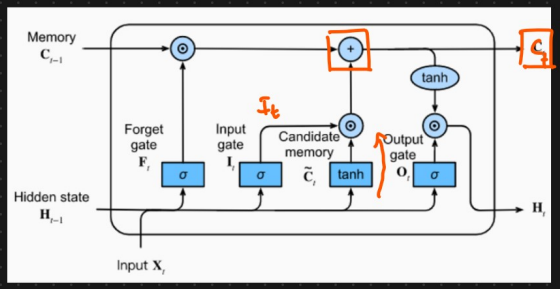
② $C_{t-1} = [6 \quad 8 \quad 9]$ ⊗ $[1 \quad 1 \quad 1] = [6 \quad 8 \quad 9]$

③ $C_{t-1} = [6 \quad 8 \quad 9]$ ⊗ $[0.5, 1, 0.5] = [3 \quad 8 \quad 4.5]$

Conclusion : Based on the context → Forget gate will let go some information or will not let go some Info {Forgetting}.

② Input Gate And Candidate Memory

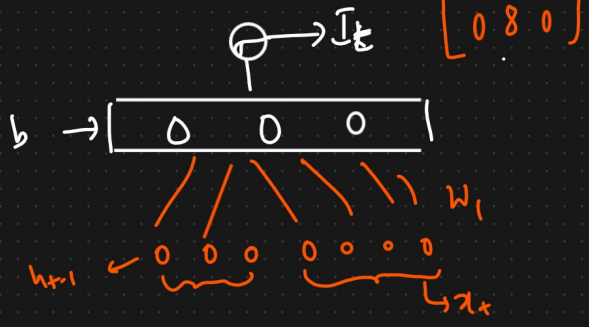


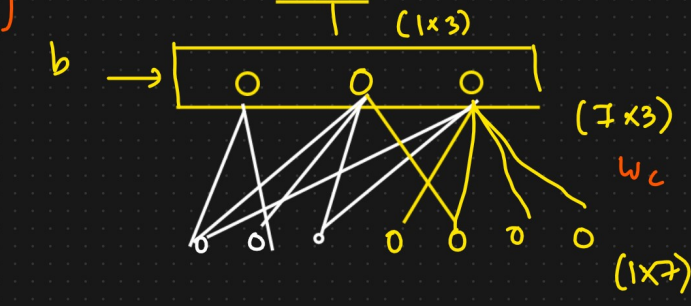


Adding Info

$I_t = [\,2\ \boxed{4}\ 6\,]$ —— ⊛ —— C_t $[\,0\ 2\ 0\,]$ ⇒ I/p Gate

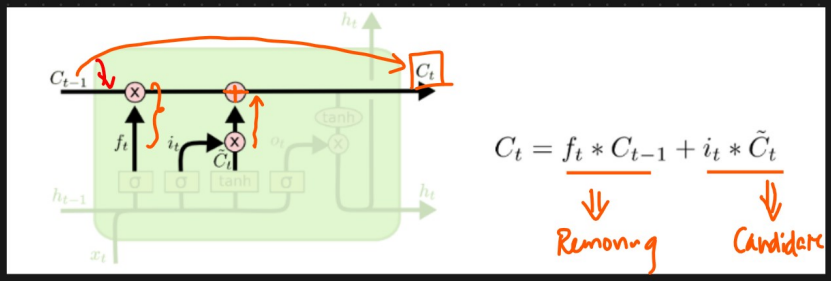
$[\,0\ 8\ 0\,]$





Context = If any information needed to be added it the memory

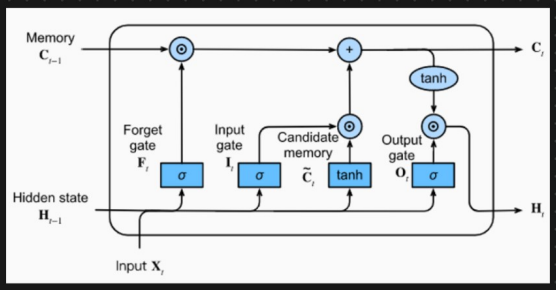
C_{t-1} ⟶ The information will be added



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

f_t ⇓ Removing

$i_t * \tilde{C}_t$ ⇓ Candidate



I stay in India . . . . . and I speak → English → Hindi

Or Forgetting Some info

Memory.

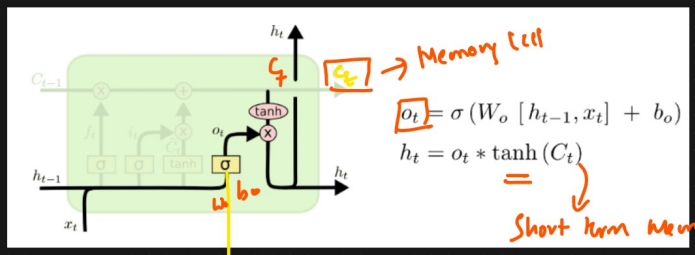
Forget Gate

I/p Gate (*) Candidate memory

+

Ct-1 => Cf.

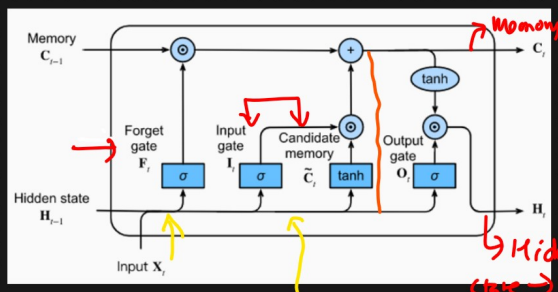
Output gate LSTM RNN



→ Memory Cell

→ Short term Memory.

→ Neural Network

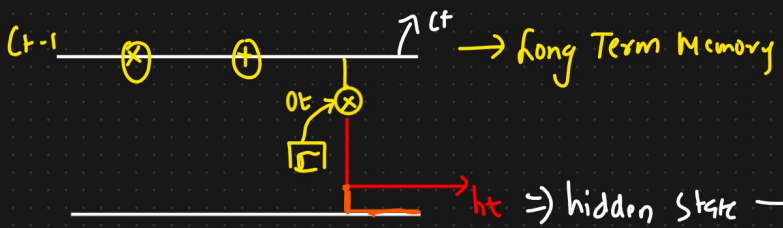


Memory

→ Hidden state → ST Memory.

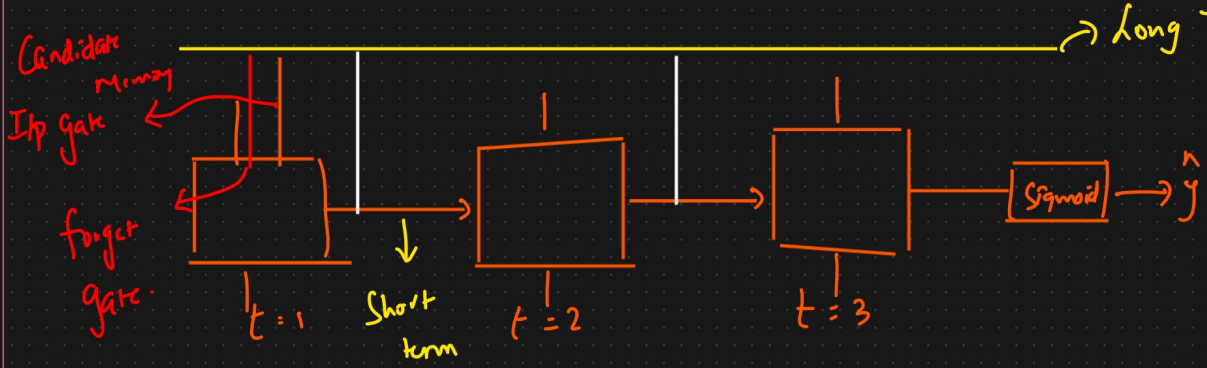
Forget → Some Info → context → Memory Cell

Add Info → Context



→ Long Term Memory

ht ⇒ hidden state → Short term Memory



→ Long Term

loss ↓↓.

[Wi, Wc, Wo] → Updating ← Back Propogation

GRU RNN ⇒ LSTM Variant

Training Data With LSTM RNN

{Training} Text Paragraph

I Went to Restaurant and order burger

The burger looked tasty and crispy

But burger is not good for health

It has lot of fats, cholestrol

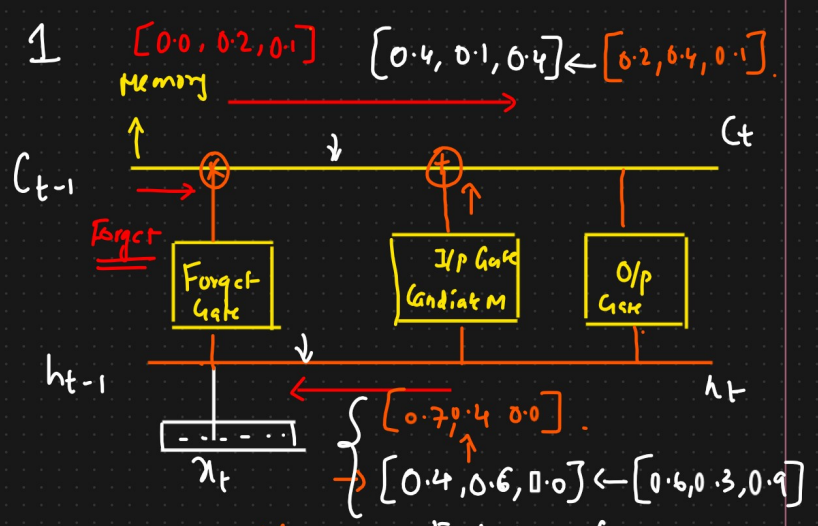
But this burger was made with whey protein and only vegetables were used, so it was good.

O/p [good/bad]

Forget gate

I/p gate

O/p gate



Word → Vectors → Embedding Layer

Word2Vec [3 dimension-vector]

→ [[Good] [Bad] [Healthy]] ← Black Box

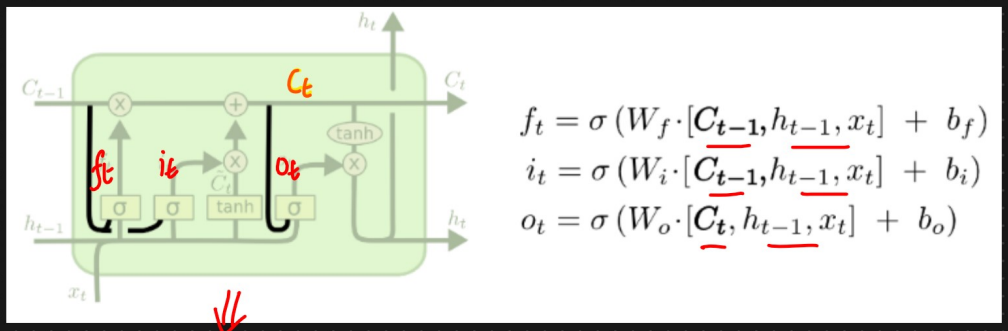
Tasty [0.9 0.0 0.1] ← 3 d.

Variants of LSTM RNN

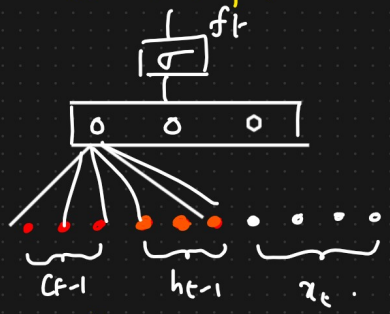
LSTM Variants Introduced By Gers & Schmidhuber [2000]

↳ Research paper

LSTM RNN [1970-80]



$$f_t = \sigma\left(W_f \cdot [C_{t-1}, h_{t-1}, x_t] + b_f\right)$$
$$i_t = \sigma\left(W_i \cdot [C_{t-1}, h_{t-1}, x_t] + b_i\right)$$
$$o_t = \sigma\left(W_o \cdot [C_t, h_{t-1}, x_t] + b_o\right)$$

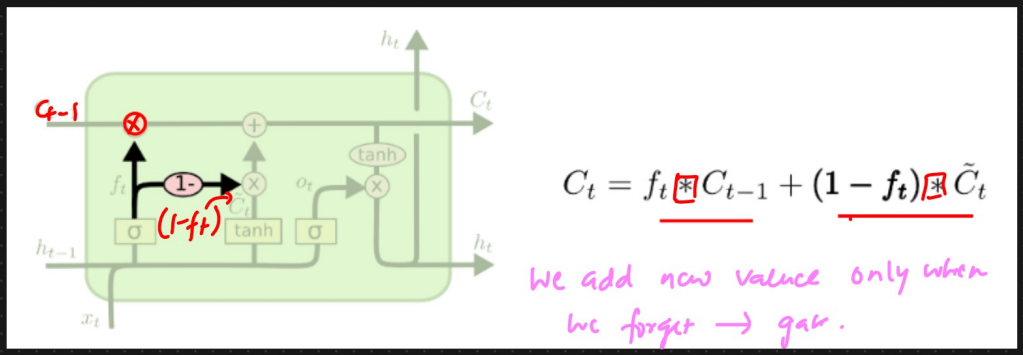


focus

Connections → From Memory cell to forget gate, i/p gate, O/p gate ⇒ Peephole Connections

Peephole Connections : We let the gate layers look at the cell state

Another variation → Coupling Forget And I/p Gates



$$C_t = f_t * C_{t-1} + (1 - f_t) * \tilde{C}_t$$

We add new value only when we forget → gate.

Goal : We only forget when we're going to i/p something in its place.

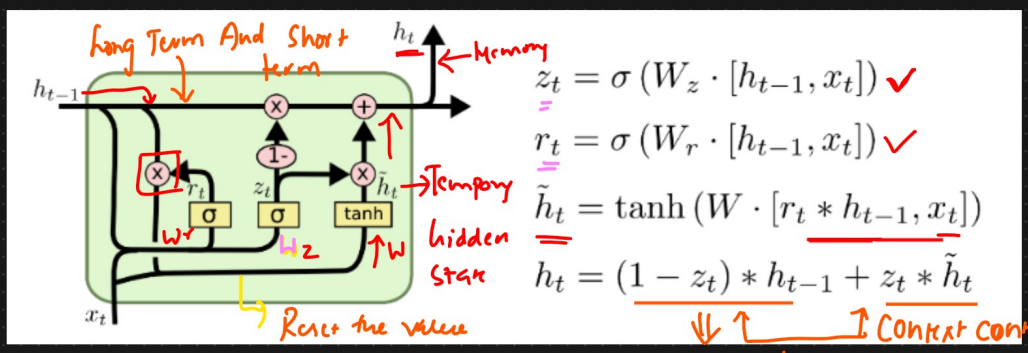
⇓

We only i/p new values to the state when we forget something older.

Instead of seperately deciding what to forget and what we should add new Info, we make this decision together.

GRU → Gated Recurrent Unit [Cho, et al [2014]]

1980 → LSTM
2000 - Variants
2014 → GRU



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

Context context

⇓ Coupled.

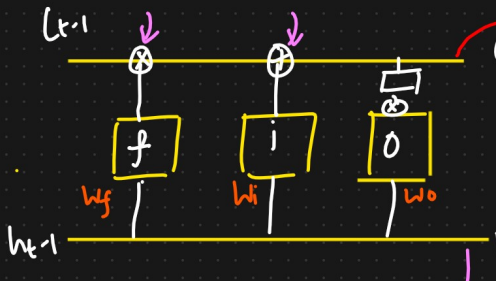
z_t ⇒ Update Gate ←

r_t ⇒ Reset Gate ←

$\tilde{h}_t$ ⇒ Temporary hidden state.

Trainable parameters

bias → $[W_f, W_i, W_o]$

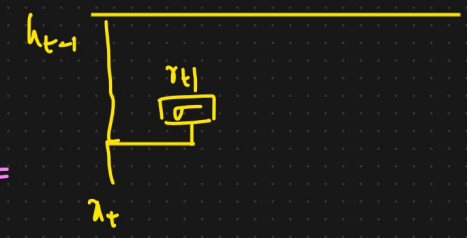


LSTM : Complex Architecture → forget, i/p + Candidate memory, O/p

Training Time ↑↑

z_t ,

Reset gate = r_t =

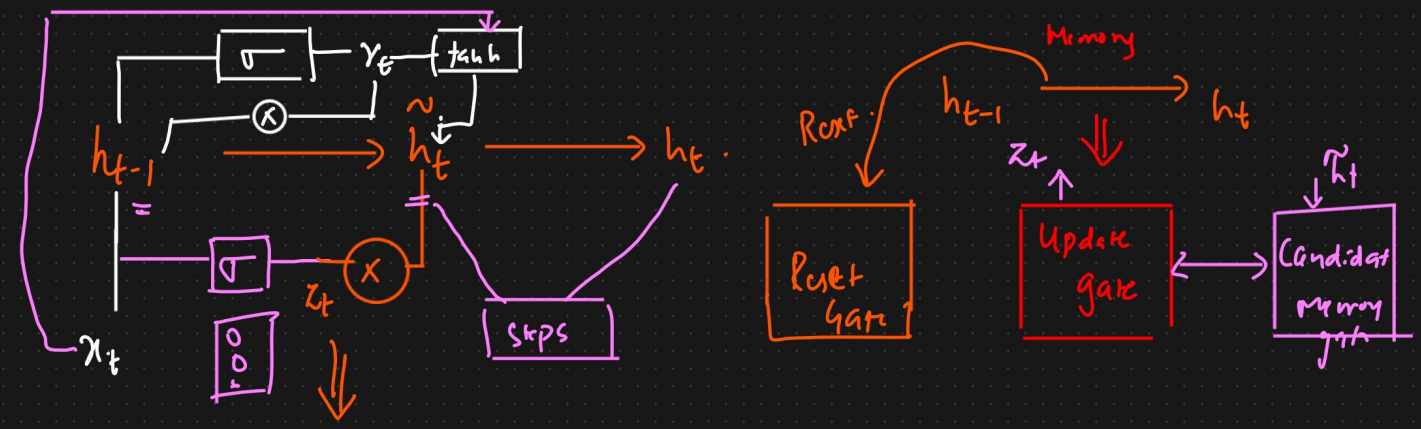


→ Reseting some info from h_{t-1} ⇒ Memory → LTM + STM

h_{t-1} [0.6 0.5 0.3 0.9]

$\otimes \leftarrow r_t$ [0.2 0.4 0.8 0.2]

$x_t \rightarrow$ [0.12 0.20 0.24 0.18] ← Resetting → Context



What context Info needs to be Added

⇓

Candidate hidden state · [current Context]

⇓

Imp → Add Info

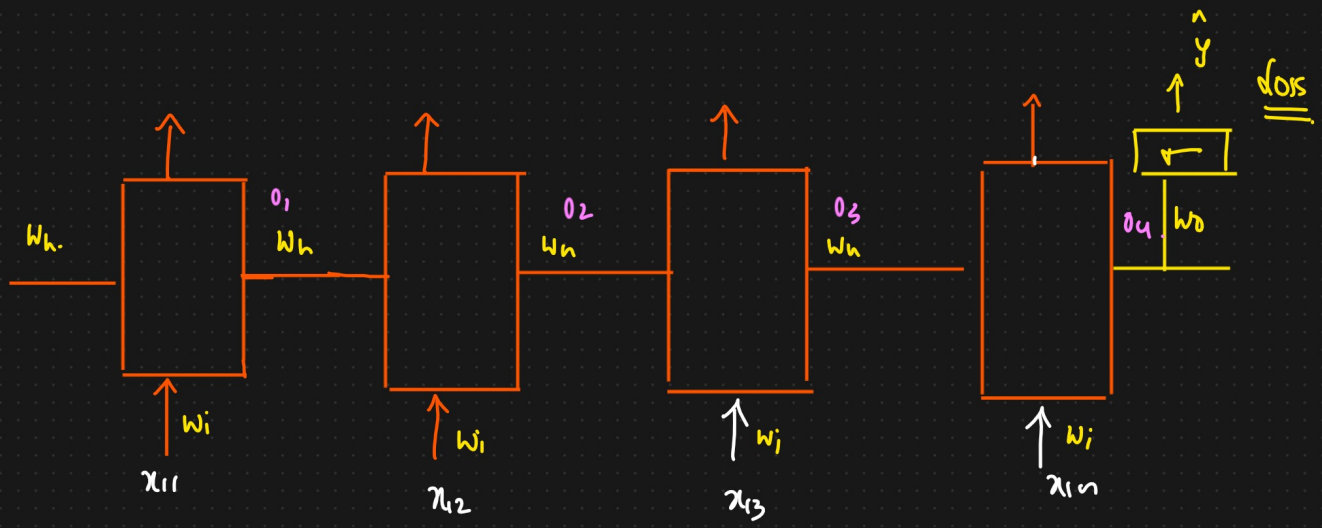
$z_t \Rightarrow$ Less Info

Bidirectional RNN

① Simple RNN ⟶ practical Implement ⟶ Embedding Layers

② LSTM, GRU Variants RNN ⟶ Practical Implement

③ Bidirectional RNN / LSTM RNN



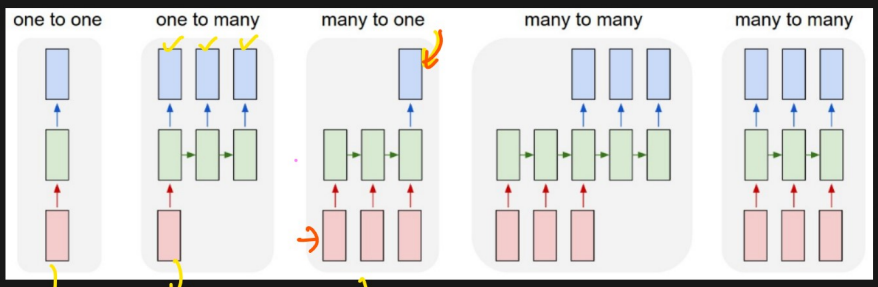
Types of RNN

① One to Many RNN

② Many to One RNN

③ Many to Many RNN.

④ One to One RNN

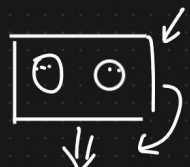


One i/p
One O/p

One i/p
O/p Many

Image scanch
I/p Many
O/p One.

Language Translation
Many I/p
Many O/p.

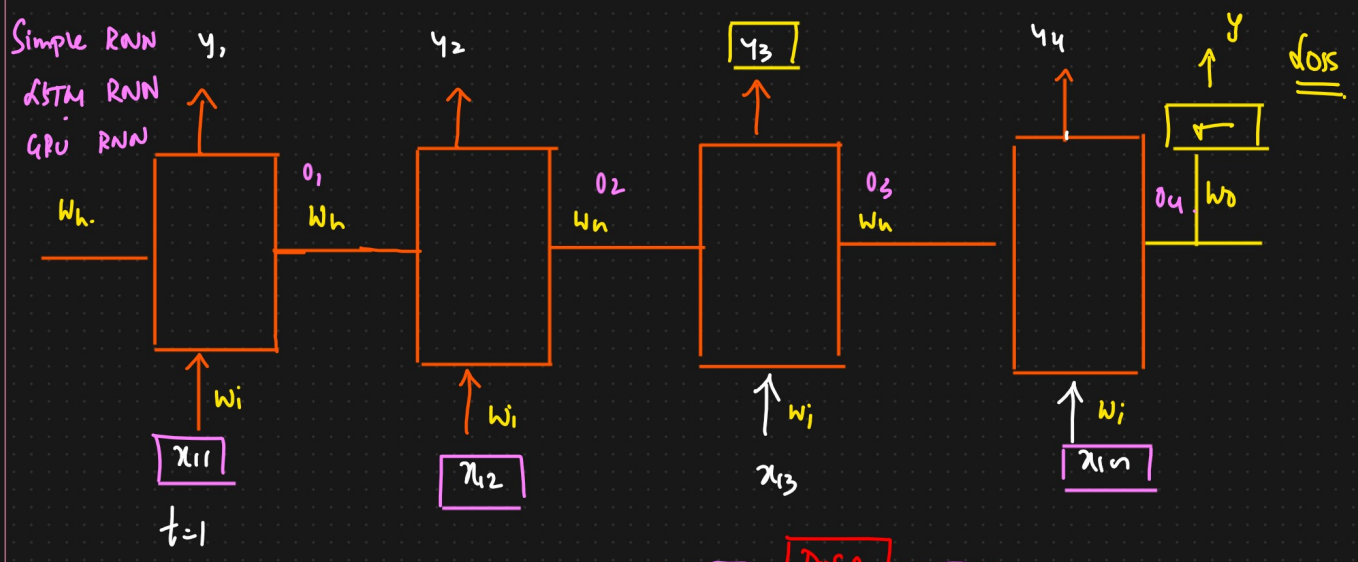


There is a dog and cat

Eg:

Eg: Image Captioning

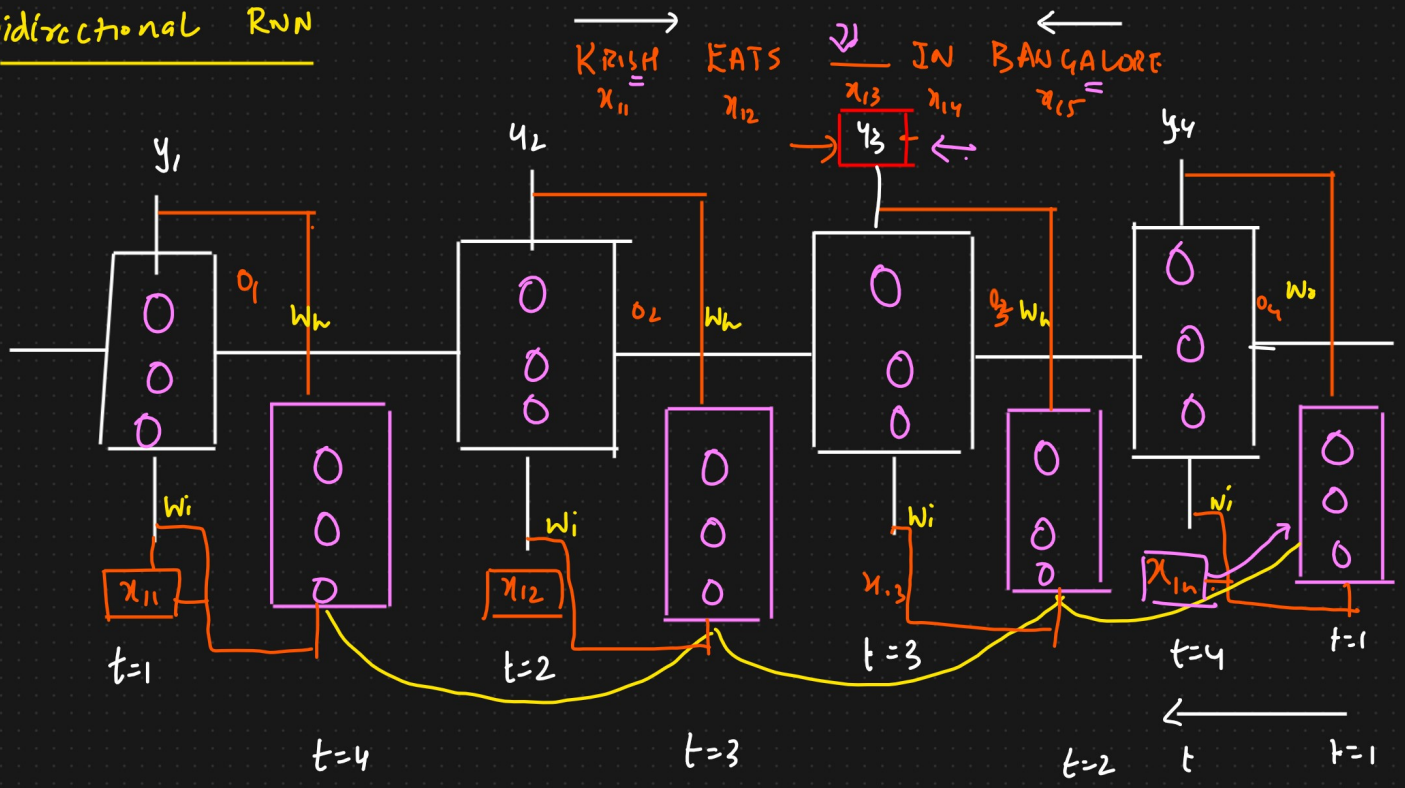
A cat eating food



Text Example = KRISH EATS DOSA IN BANGALORE

KRISH EATS PIZZA IN PARIS

Bidirectional RNN



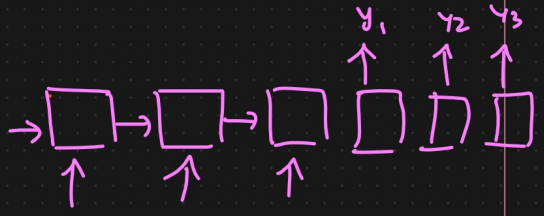
① Forward propogation → Equation ←

Encoder And Decoder

1. Simple RNN → Vanishing Gradient Problem
2. LSTM RNN →
3. GRU RNN → } Long Short Term Memory.
4. Bidirectional RNN ←

Type of RNN

1. Many to Many RNN



Encoder And Decoder

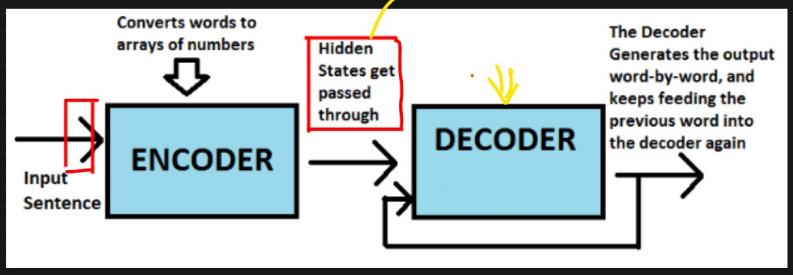
Sequences I/p O/p Sequence Of Words

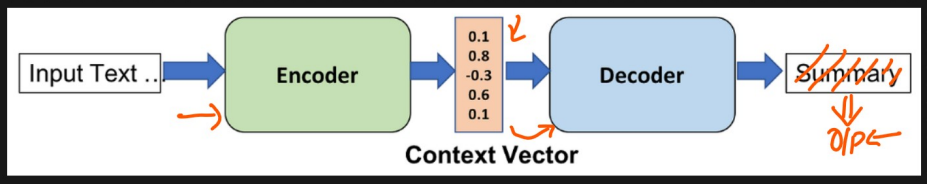
Eg: One Language To Other

English → French

Eg: Linked Chat → Hi, How are you?

Simple Working





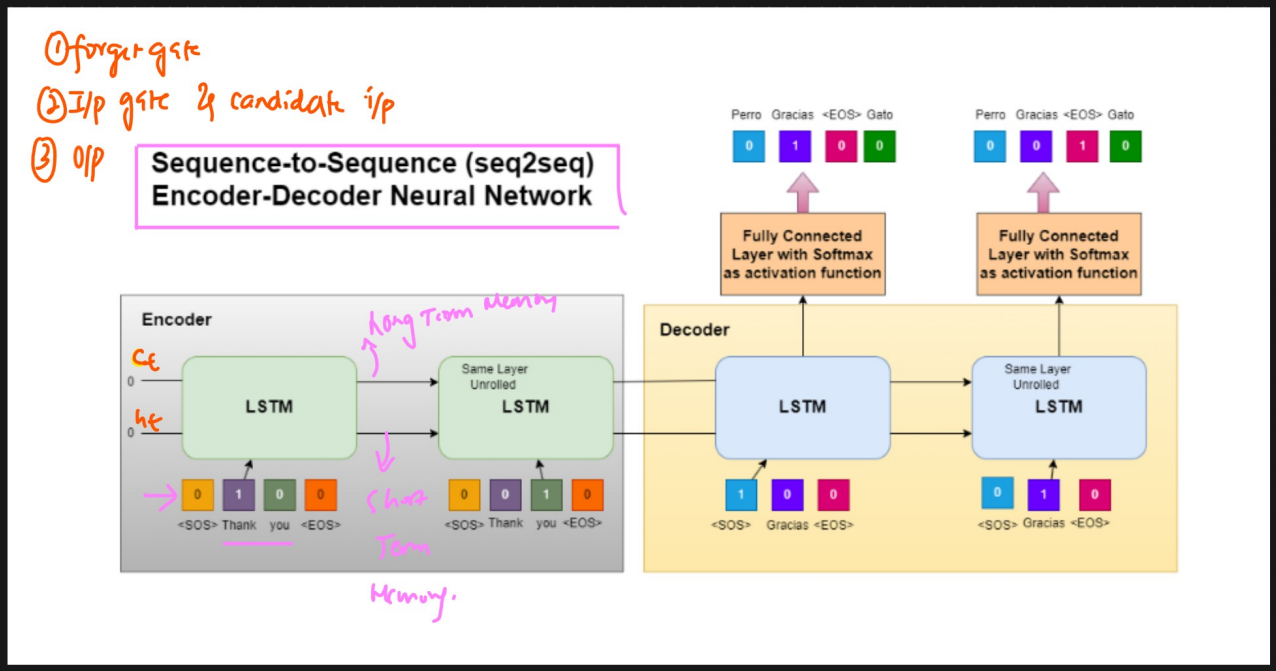
1. Encoder ⇒ I/p ⇒ Context Vector ⇐ Vectors
2. Decoder ⇒ ⇒ O/p

Usecase

1. Language Translation
2. Text Generation
3. Text Suggestion.

RNN → Vanishing Gradient Problem

① forget gate
② I/p gate & candidate i/p
③ O/p



Dataset

English: <SOS> Thank you <EOS>.

True O/p: French {Training}: Gracias

<SOS> GRACIAS → <EOS> . GRACIOUS <EOS>.

$\hat{y}$ [0.1, 0.6, 0.3] [0.1, 0.3, 0.6] $\hat{y}$ [- - -]

$[y\ \hat{y}]$ like

Encoder → Context Vectors

Long Term Memory

Short Term Memory

Softmax Softmax

Decoder

Ct, ht

ONE ⇒ ⇓ Context of this Sentence.

<SOS> Thank you <EOS>. <SOS> GRACIOUS

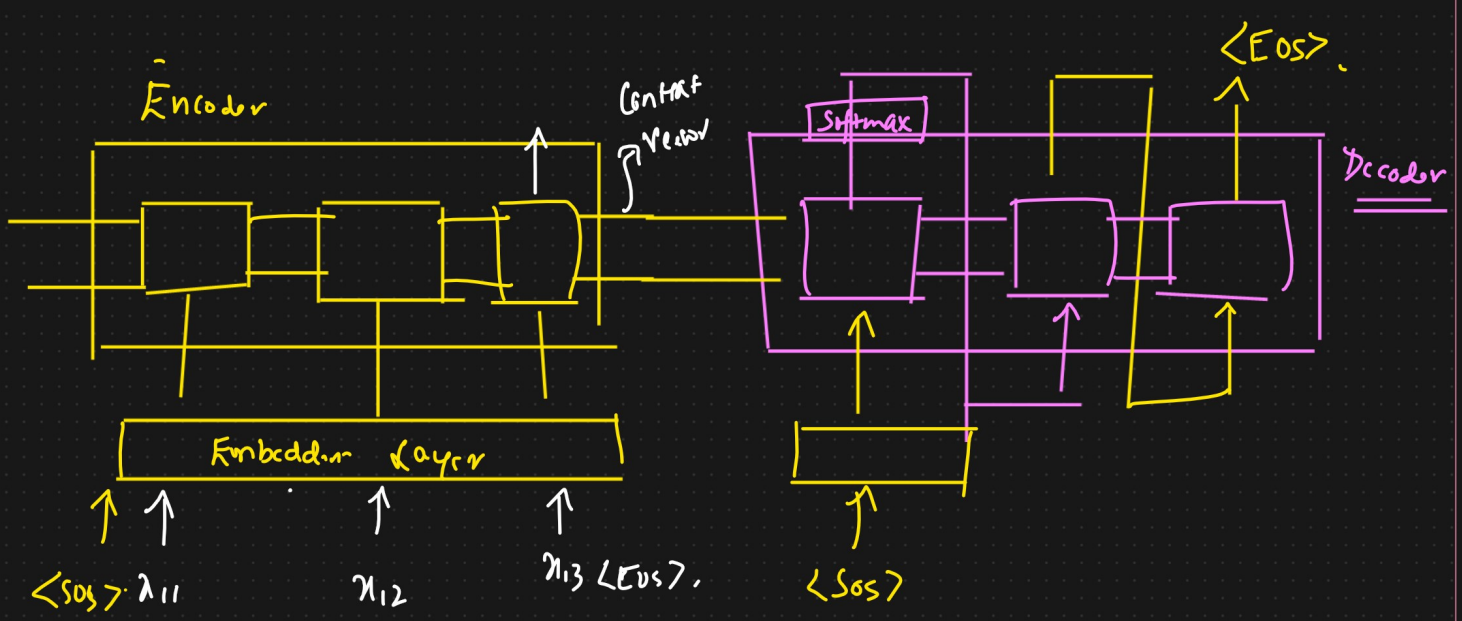
t=1 t=2 t=3 t=4 ← t=1 t=2 t=3

$\hat{y}$ [- - -] [- - -] [- - -]

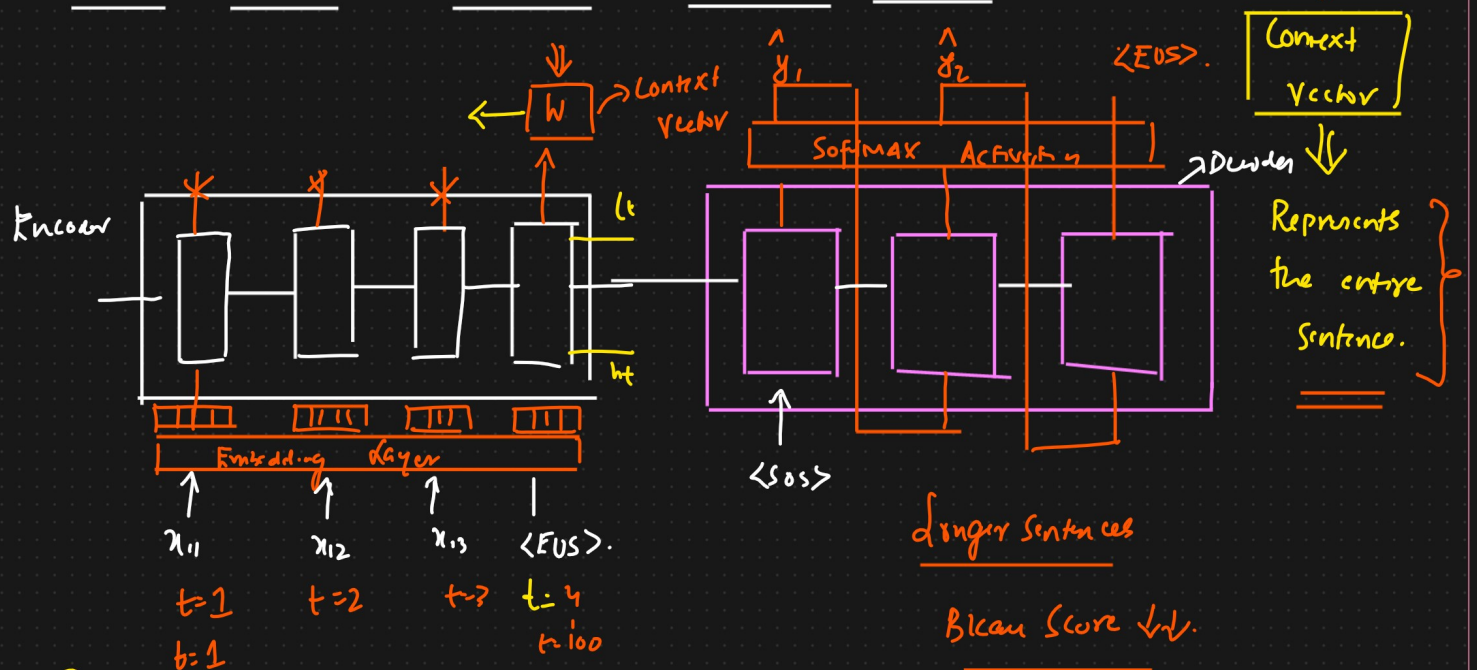
y [. . .] [. . .] [. . .]

Optimizer

Loss $(y-\hat{y})^2$



Problems With Encoder-Decoder Seq2seq Architecture

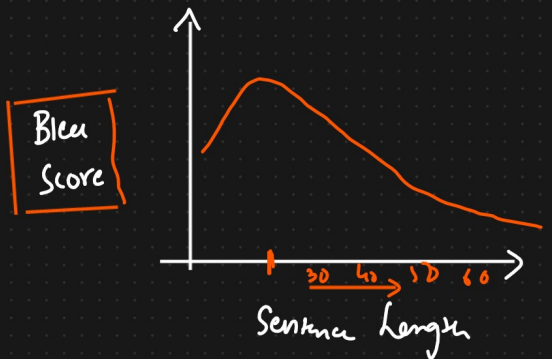


Context vector
Represents the entire sentence.

Longer Sentences

Bleau Score ↓↓.

Reasons : Sentences of varying length



⇒ Seq to Seq Data

(*) Attention Mechanism → Seq2Seq Network

Longer paragraph → {context vector}.

+

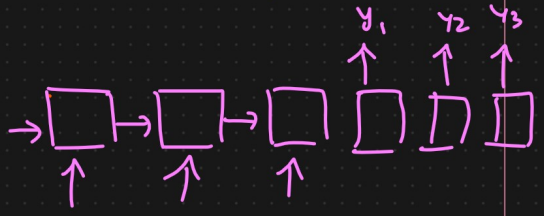
{context}

Encoder And Decoder

1. Simple RNN → Vanishing Gradient Problem
2. LSTM RNN →
3. GRU RNN → } Long Short Term Memory.
4. Bidirectional RNN ←

Type of RNN

1. Many to Many RNN



Encoder And Decoder

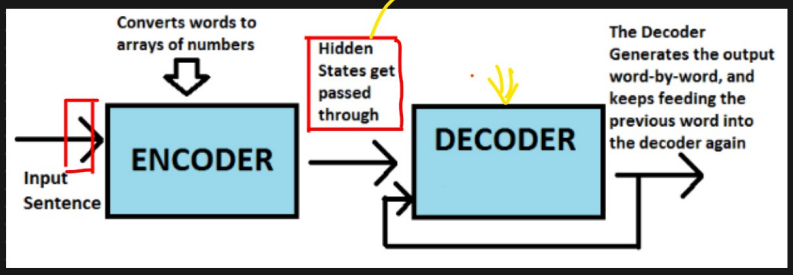
Sequences I/p O/p Sequence Of Words

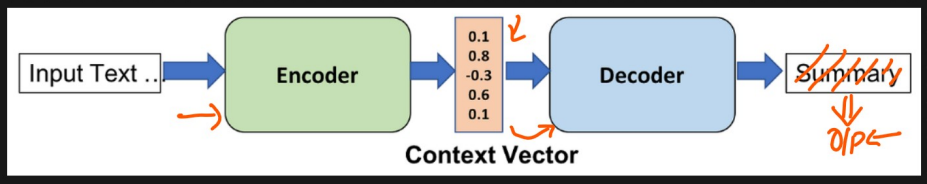
Eg: One Language To Other

English → French

Eg: Linked Chat → Hi, How are you?

Simple Working





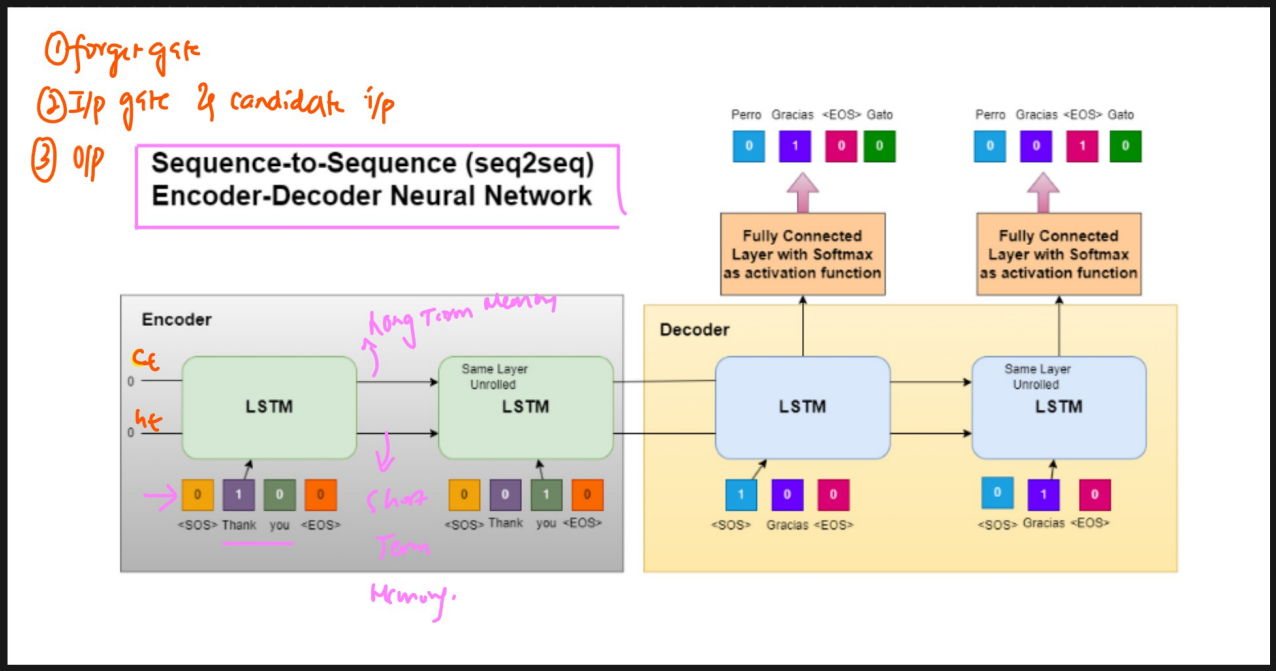
1. Encoder ⇒ I/p ⇒ Context Vector ⇐ Vectors
2. Decoder ⇒ ⇒ O/p

Usecase

1. Language Translation
2. Text Generation
3. Text Suggestion.

RNN → Vanishing Gradient Problem

① forget gate
② I/p gate & candidate i/p
③ O/p



Dataset

English: <SOS> Thank you <EOS>.

True O/p: French {Training}: Gracias

<SOS> GRACIAS → <EOS> . GRACIOUS <EOS>.

$\hat{y}$ [0.1, 0.6, 0.3] [0.1, 0.3, 0.6] $\hat{y}$ [- - -]

$[y\ \hat{y}]$ like

Encoder → Context Vectors

Long Term Memory

Short Term Memory

Softmax Softmax

Decoder

Ct, ht

ONE ⇒ ⇓ Context of this Sentence.

<SOS> Thank you <EOS>. <SOS> GRACIOUS

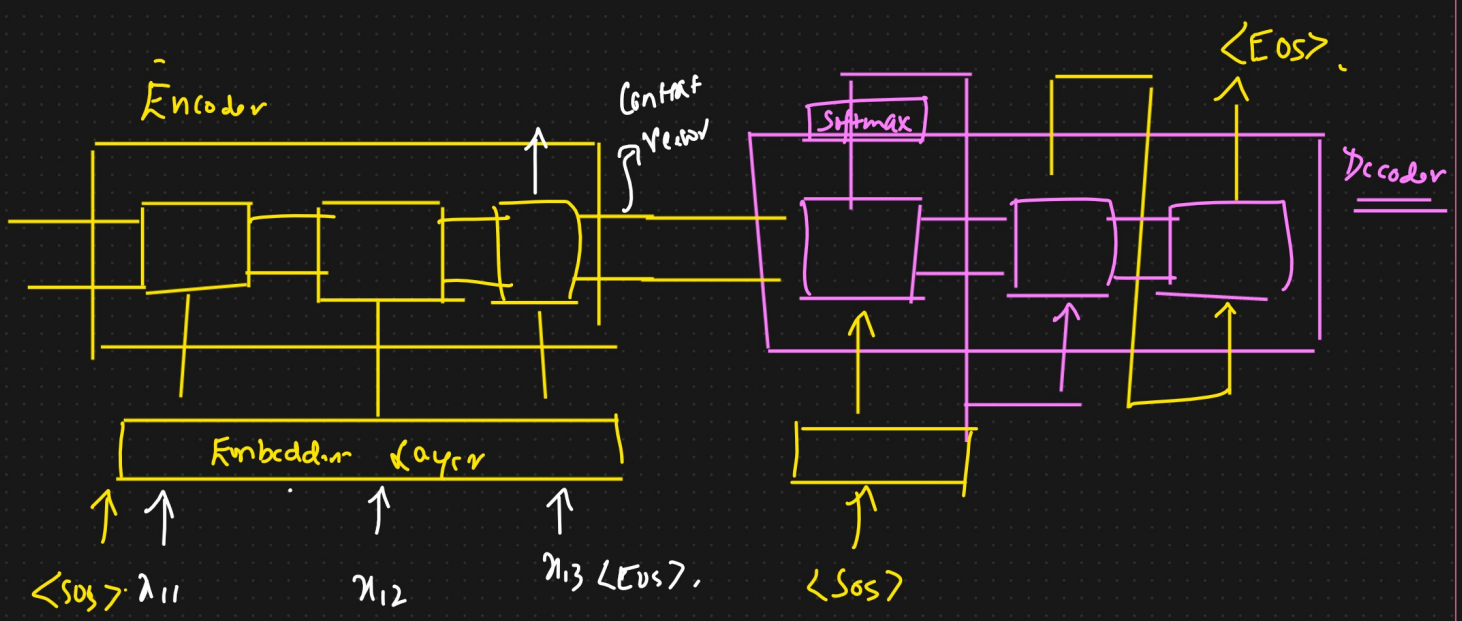
t=1 t=2 t=3 t=4 ← t=1 t=2 t=3

$\hat{y}$ [- - -] [- - -] [- - -]

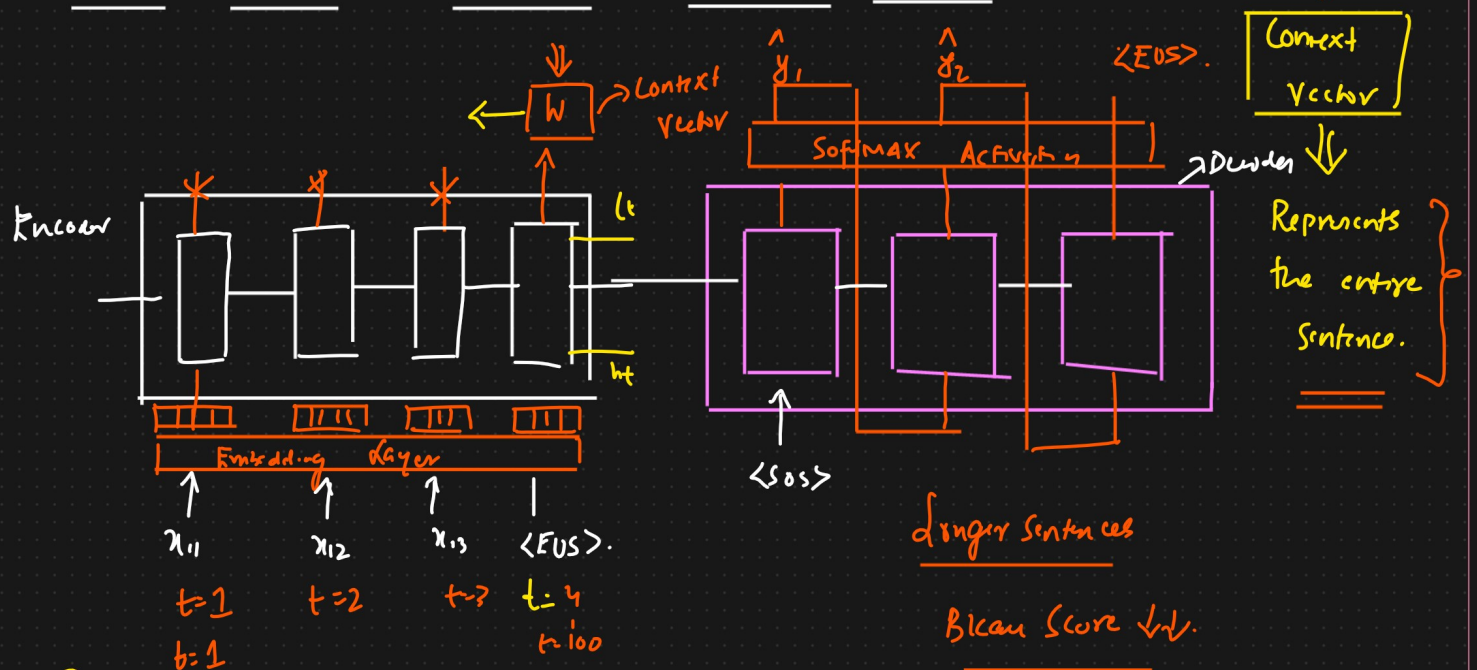
y [. . .] [. . .] [. . .]

Optimizer

Loss $(y-\hat{y})^2$



Problems With Encoder-Decoder Seq2seq Architecture

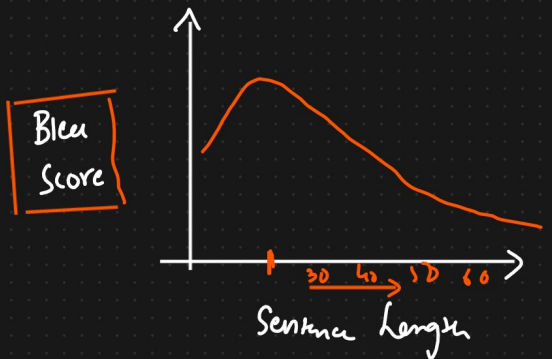


Context vector
Represents the entire sentence.

Longer Sentences

Bleau Score ↓↓.

Reasons : Sentences of varying length



⇒ Seq to Seq Data

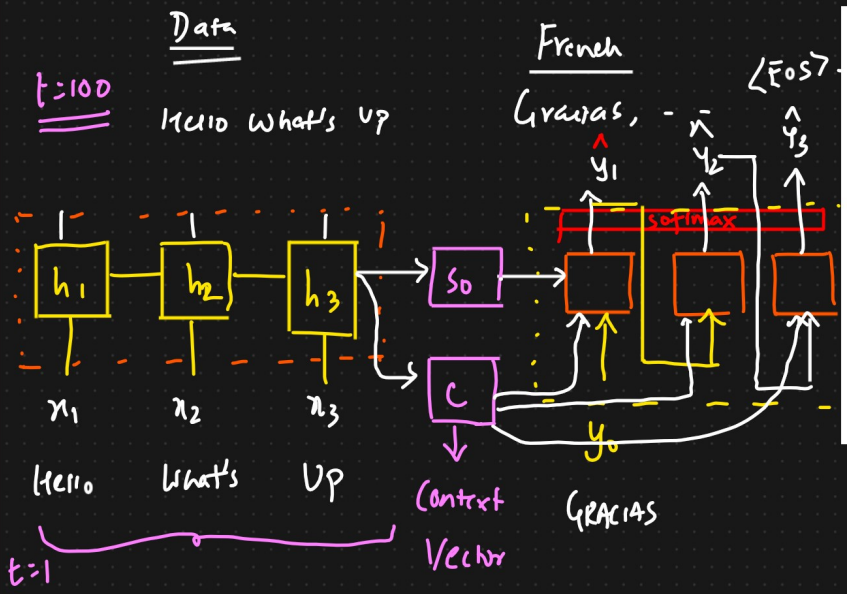
(*) Attention Mechanism → Seq2Seq Network

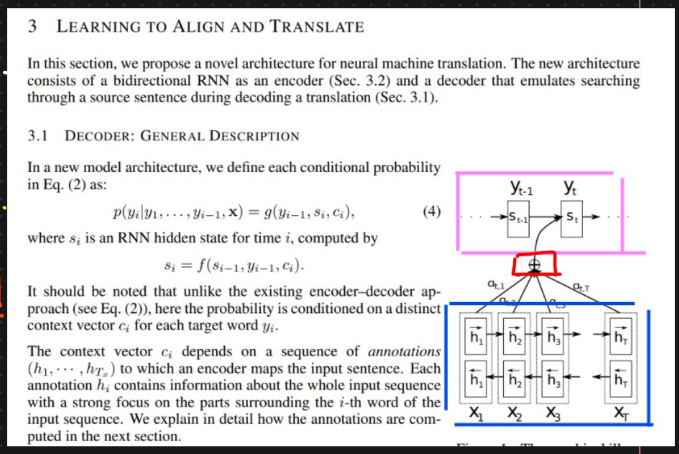
Longer paragraph → {context vector}.

+

{context}

Attention Mechanism | Seq2Seq Networks





3 LEARNING TO ALIGN AND TRANSLATE

In this section, we propose a novel architecture for neural machine translation. The new architecture consists of a bidirectional RNN as an encoder (Sec. 3.2) and a decoder that emulates searching through a source sentence during decoding a translation (Sec. 3.1).

3.1 DECODER: GENERAL DESCRIPTION

In a new model architecture, we define each conditional probability in Eq. (2) as:

$$p(y_i|y_1,\dots,y_{i-1},\mathbf{x}) = g(y_{i-1}, s_i, c_i), \quad (4)$$

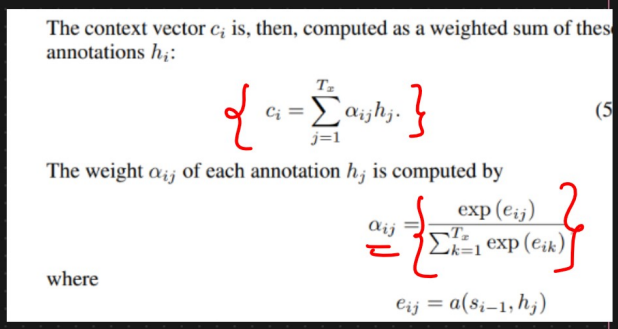
where s_i is an RNN hidden state for time i, computed by

$$s_i = f(s_{i-1}, y_{i-1}, c_i).$$

It should be noted that unlike the existing encoder–decoder approach (see Eq. (2)), here the probability is conditioned on a distinct context vector c_i for each target word y_i.

The context vector c_i depends on a sequence of *annotations* $(h_1,\cdots,h_{T_x})$ to which an encoder maps the input sentence. Each annotation h_i contains information about the whole input sequence with a strong focus on the parts surrounding the i-th word of the input sequence. We explain in detail how the annotations are computed in the next section.

Encoder Decoder Architecture



The context vector c_i is, then, computed as a weighted sum of these annotations h_i:

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j. \quad (5$$

The weight α_{ij} of each annotation h_j is computed by

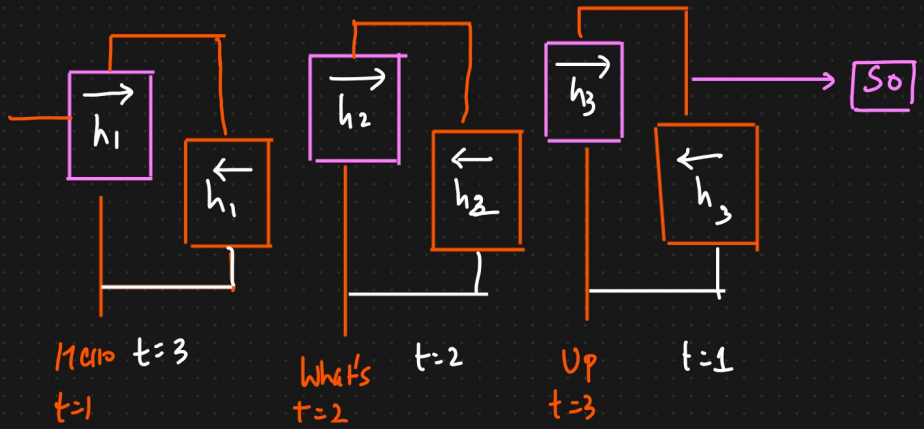
$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$

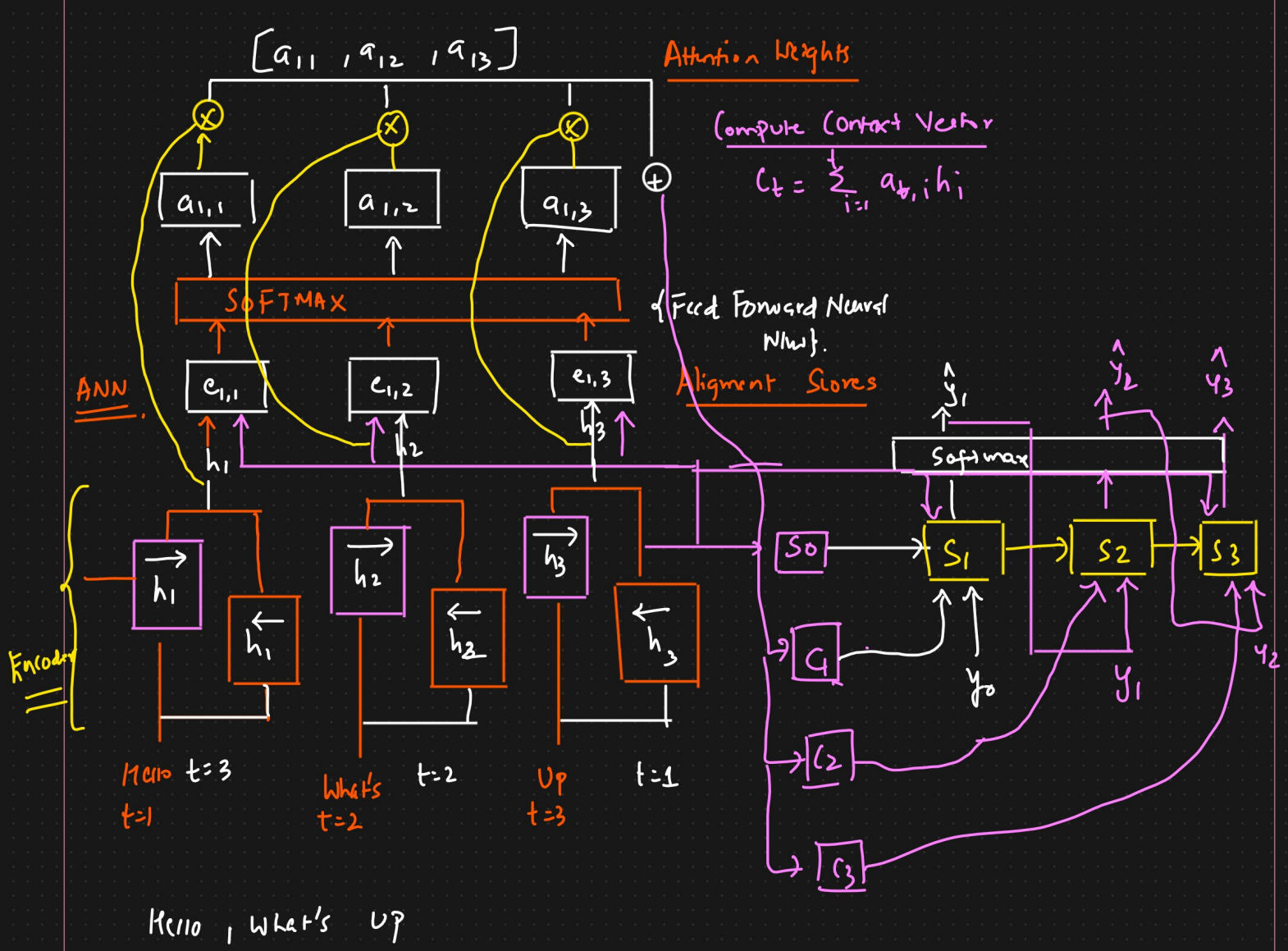
where

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

Attention Mechanism

https://erdem.pl/2021/05/introduction-to-attention-mechanism

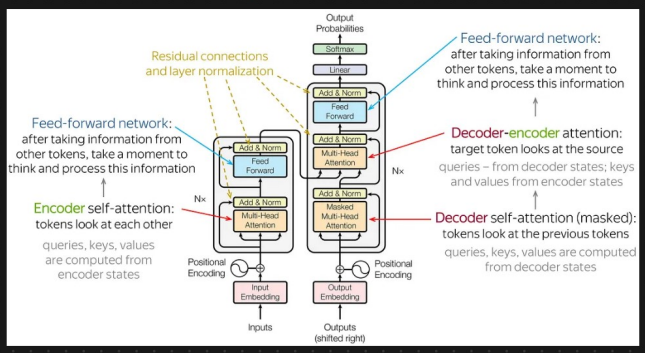



[a11 , a12 , a13]
Attention Weights
Compute Context Vector
$c_t = \sum_{i=1} a_{t,i} h_i$
a1,1
a1,2
a1,3
SOFTMAX
{Feed Forward Neural Nwk}.
Alignment Scores
e1,1
e1,2
e1,3
ANN
h1
h2
h3
Softmax
y1
y2
y3
Encoder
h1
h2
h3
S0
S1
S2
S3
C
C2
C3
y0
y1
y2
Hello t=3
t=1
What's t=2
t=2
Up t=3
t=1
Hello , What's Up

Introduction To Transformers

1) RNN / LSTM / GRU RNN
2) Encoder Decoder Architecture
3) ATTENTION MECHANISM
4) TRANSFORMERS
   - ① Why Transformers?
   - ② Architecture of Transformers?
   - ③ SELF ATTENTION → Q, K, V
   - ④ Positional Encoding
   - ⑤ Multi Head ATTENTION
   - ⑥ Combining the Working of Transformers

Architecture



Generative AI → LLM, Multimodel

BERT, GPT ←

↓

Open AI → Chat GPT

↓

GPT-4o

① What And Why → Transformers

Transformers in natural language processing (NLP) are a type of deep learning model that use self-attention mechanisms to analyze and process natural language data. They are encoder-decoder models that can be used for many applications, including machine translation. ⇒ Seq2Seq Task

Eg: Language Translation → Google Translation

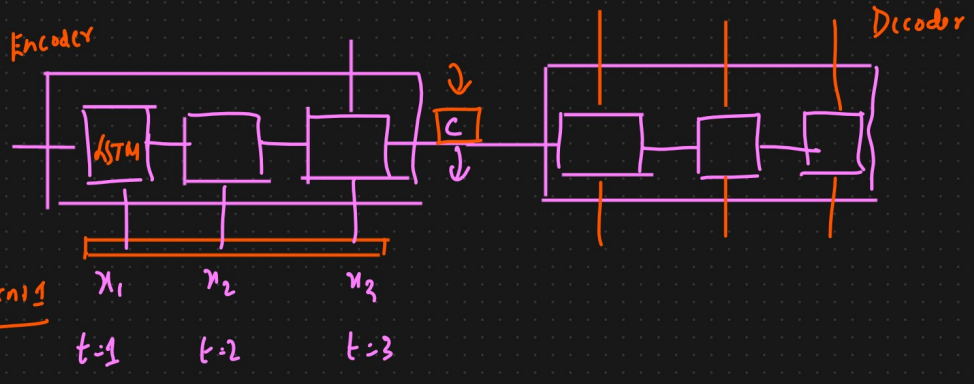
English → French

i/p ⇒ Many → o/p: Many. {Length of the sentence}.

Encoder - Decoder

Sentence Length ↑↑ → Bleau Score ↓↓

Length Sentence ↑↑.



3.1 Decoder: General Description

In a new model architecture, we define each conditional probability in Eq. (2) as:

$$p(y_i|y_1,\dots,y_{i-1},\mathbf{x}) = g(y_{i-1}, s_i, c_i), \quad (4)$$

where s_i is an RNN hidden state for time i, computed by

$$s_i = f(s_{i-1}, y_{i-1}, c_i).$$

It should be noted that unlike the existing encoder–decoder approach (see Eq. (2)), here the probability is conditioned on a distinct context vector c_i for each target word y_i.

The context vector c_i depends on a sequence of *annotations* $(h_1,\cdots,h_{T_x})$ to which an encoder maps the input sentence. Each annotation h_i contains information about the whole input sequence with a strong focus on the parts surrounding the i-th word of the input sequence. We explain in detail how the annotations are computed in the next section.

The context vector c_i is, then, computed as a weighted sum of these annotations h_i:

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j. \quad (5)$$

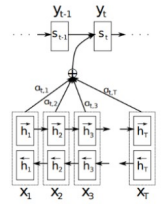
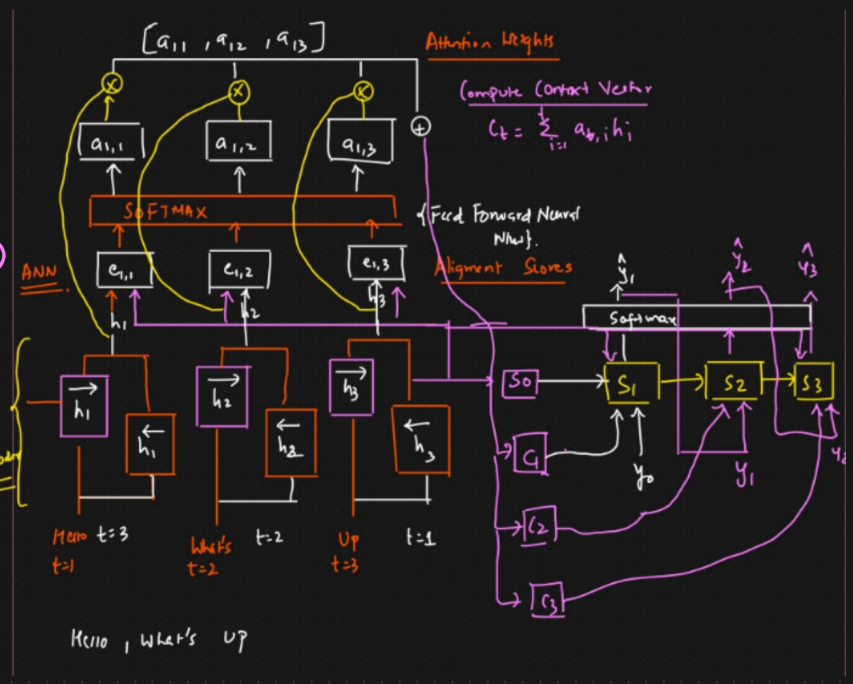


Figure 1: The graphical illustration of the proposed model trying to generate the t-th target word y_t given a source sentence $(x_1, x_2, \dots, x_T)$.



Additional Context ⟶ Decoder

long Sentence. Accuracy ↑↑.

Attention Mechanism ↓

① Parallely we cannot send all the words in a sentence ⟶ **Scalable**

DATASET ⟶ Huge ⟶ Scalable With Respect to Training.

TRANSFORMERS ≠ LSTM RNN

↳ Self Attention Module ← All the words will be parallely sent to encoder.

↓

Positional Encoding

Transformer ↑↑ DATASET ⟶ Amazing SOTA ← NLP

Transfer Learning ⟶ MultiModal Task ⟶ NLP + Image ←

Transformers ∴ AI Space ⟶ SOTA Model ⟶

↓ ↓

BERT GPT ⟶ Transfer learning ⟶ SOTA Models ⟶ DALLE } LLM's Generative AI

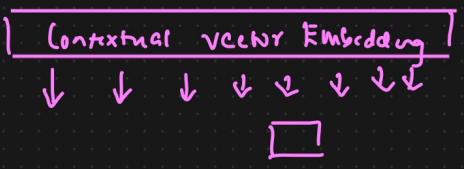
↑

Train Huge Data

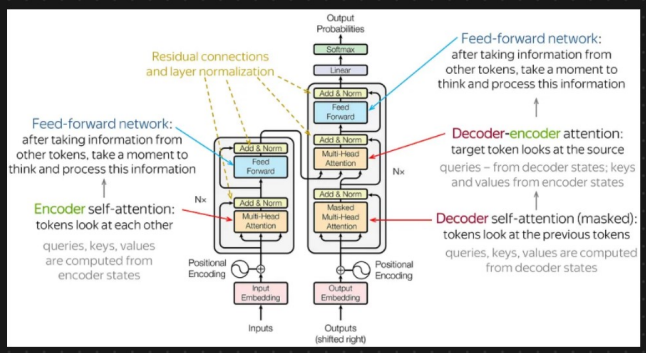
② Contextual Embedding → Self Attention — Contextual Vectors

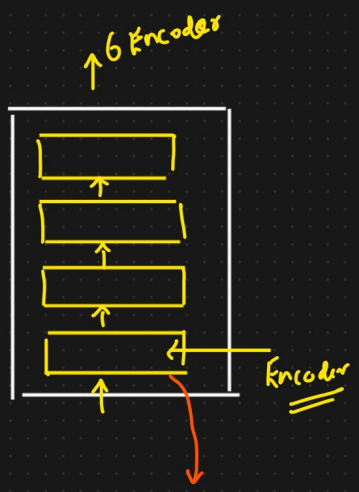
Eg: My name is KRISH And I play CRICKET

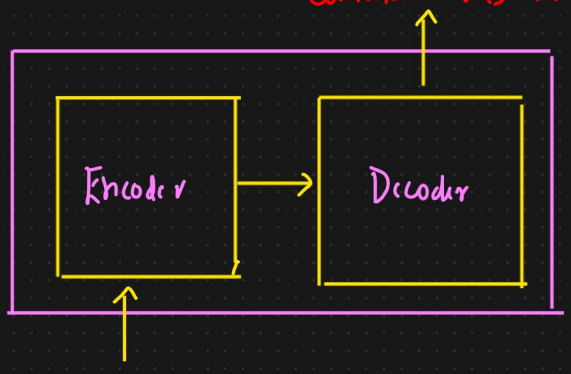
Word2Vec

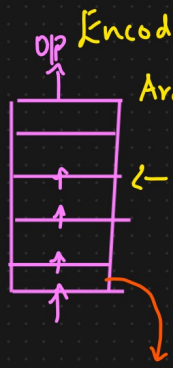


② Basic Transformer Architecture {Seq2Seq Task} → Language Translation {Eng → French}

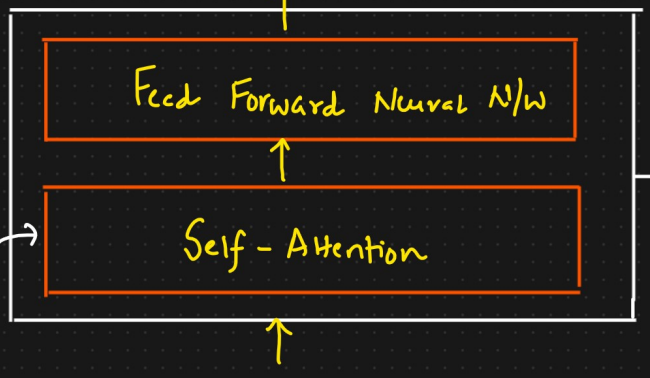






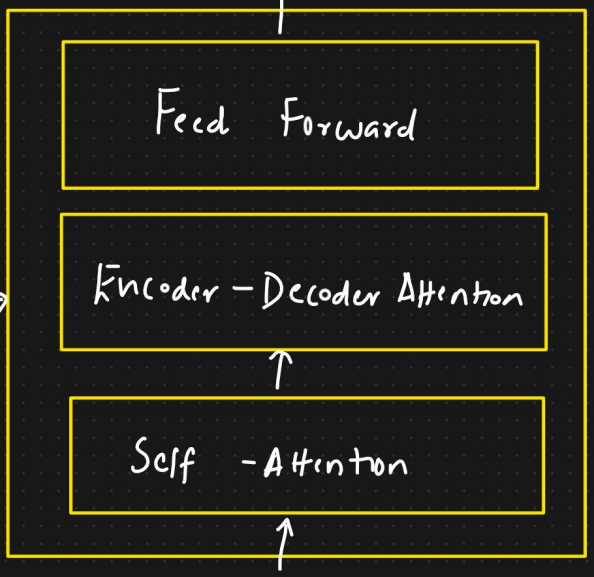


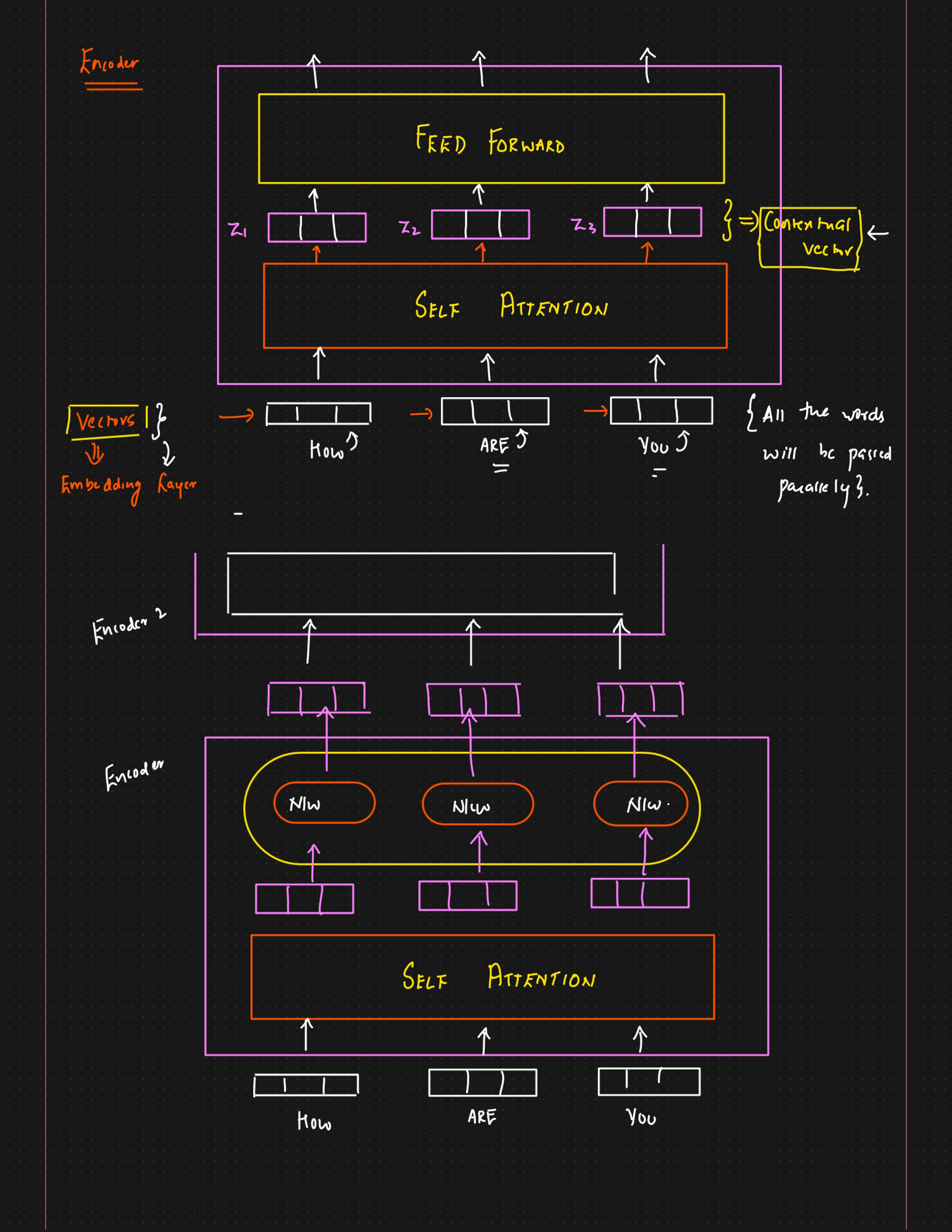
Encoder



{LSTM RNN

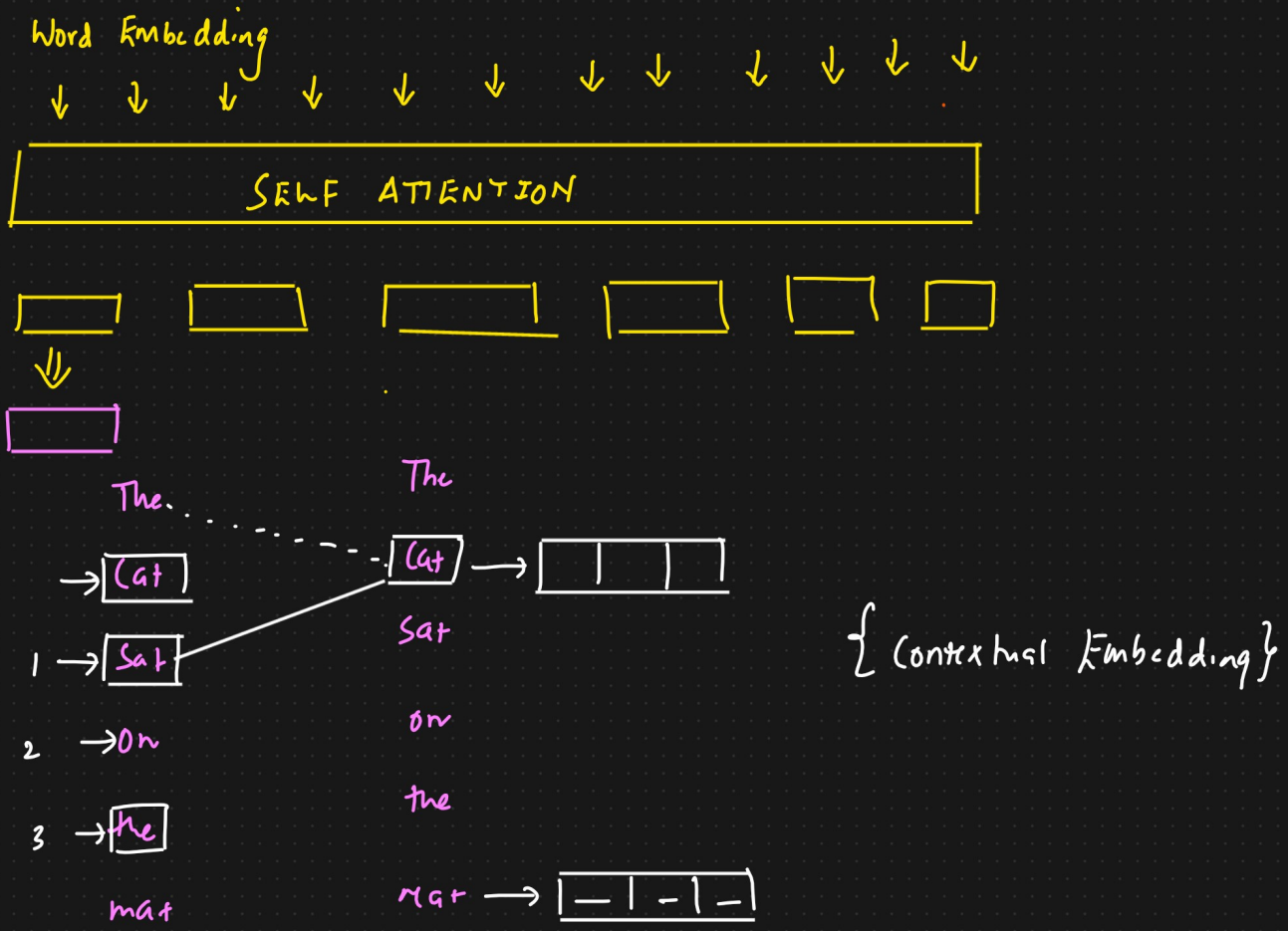
Decoder

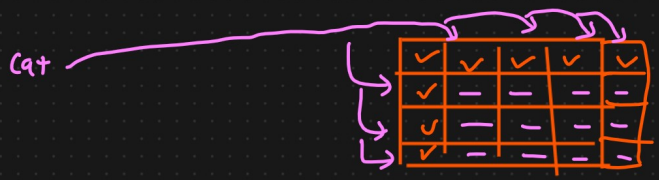



Encoder
FEED FORWARD
Z1
Z2
Z3
Contextual vector
SELF ATTENTION
Vectors
Embedding Layer
How
ARE
You
All the words will be passed parallely.
Encoder 2
Encoder
NN
NN
NN
SELF ATTENTION
How
ARE
You

Self Attention At a Higher Level

Eg: The cat sat on the mat, the cat lay on the rug.





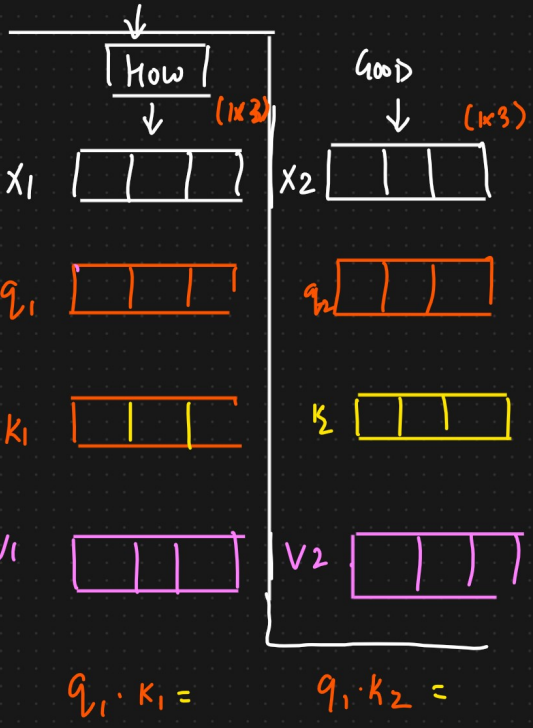
Self Attention In Detail

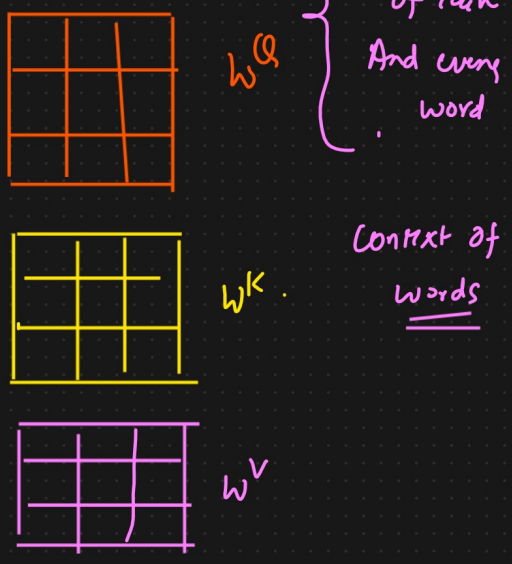
① To create 3 vectors from each of the encoder i/p. Query vector, Key vector, value vector. ⇒ Contextual Embedding

Eg: YT ⇒ Search keywords (Topic's) ⇒ {Query}.

Query ⟶ {Key} → Tags, Description, Title ⇒ Value ⇒ O/P Video

② Second step in calculating self attention is to calculate the score. The Score determines how much focus to place on the other part of the sentence.





(*) How much Focus to place on other part of the i/p sentence as we encode a word at certain position

Divide $\sqrt{dk}$ — —

Dimension=3 ↓ Variance / $\sqrt{dk}$ ≈ Variance ↑↑↑ / $\sqrt{dk}$ ⇒

Dimension ↑↑ T12

O/p ↑↑↑

{More stable gradients}

Softmax

[0-1] [0-1].

0.88 + 0.12 = 1

Softmax ⇒ The softmax score determines how much each word will be expressed at this position

X X X

Value Vectors V_1 [1 1 1] V_2 [1 0.6 1]

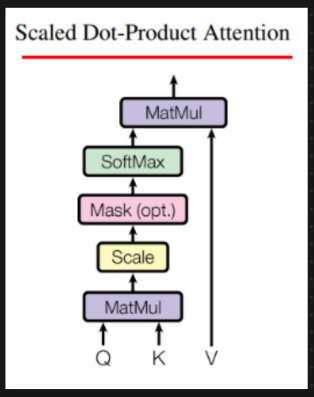
0.88 [1 1 1]

[0.88 0.88 0.88] + [0.12 0.6 0.12]

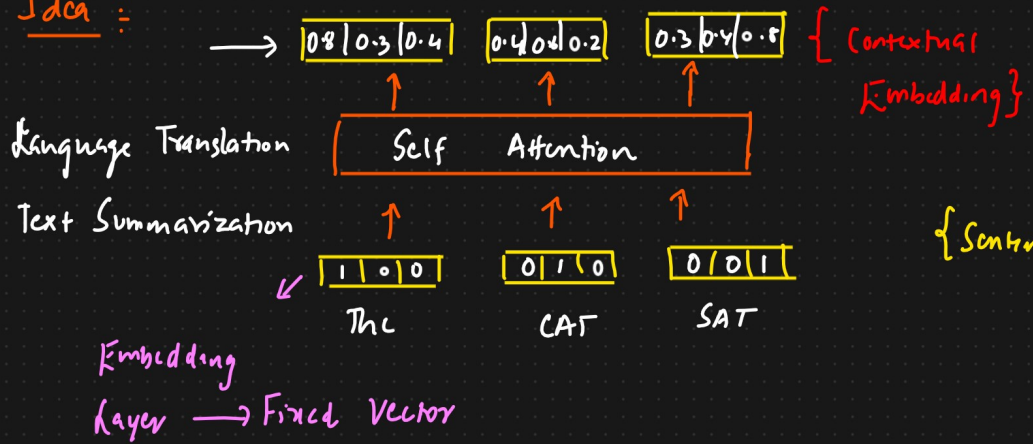
Contextual vector Z_1 → [ ] Z_2 → [ ]

Self Attention At Higher And Detailed Level

Self-attention, also known as scaled dot-product attention, is a crucial mechanism in the transformer architecture that allows the model to weigh the importance of different tokens in the input sequence relative to each other



Idea :



1) Inputs: Queries, Keys, And Values

Model → Queries, Keys And Values

1. Query Vectors (Q):
Role: Query vectors represent the token for which we are calculating the attention. They help determine the importance of other tokens in the context of the current token.

Importance:

Focus Determination: **Queries help the model decide which parts of the sequence to focus on for each specific token. By calculating the dot product between a query vector and all key vectors, the model assesses how much attention to give to each token relative to the current token.**

Contextual Understanding: **Queries contribute to understanding the relationship between the current token and the rest of the sequence, which is essential for capturing dependencies and context.**

2. Key Vectors (K):
Role: Key vectors represent all the tokens in the sequence and are used to compare with the query vectors to calculate attention scores.

Importance:

Relevance Measurement: **Keys are compared with queries to measure the relevance or compatibility of each token with the current token. This comparison helps in determining how much attention each token should receive.**
Information Retrieval: **Keys play a critical role in retrieving the most relevant information from the sequence by providing a basis for the attention mechanism to compute similarity scores.**

3. Value Vectors (V):
Role: Value vectors hold the actual information that will be aggregated to form the output of the attention mechanism.

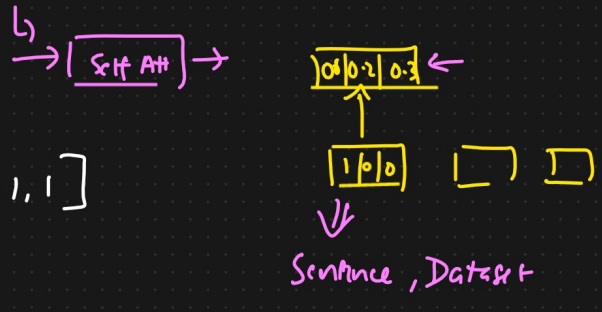
Importance:

Information Aggregation: Values contain the data that will be weighted by the attention scores. The weighted sum of values forms the output of the self-attention mechanism, which is then passed on to the next layers in the network.
Context Preservation: By weighting the values according to the attention scores, the model preserves and aggregates relevant context from the entire sequence, which is crucial for tasks like translation, summarization, and more.

Input Sequence = ["The", "CAT", "SAT"]

Embedding Size = 4

Q, K, V ⇒ 4



① Token Embedding

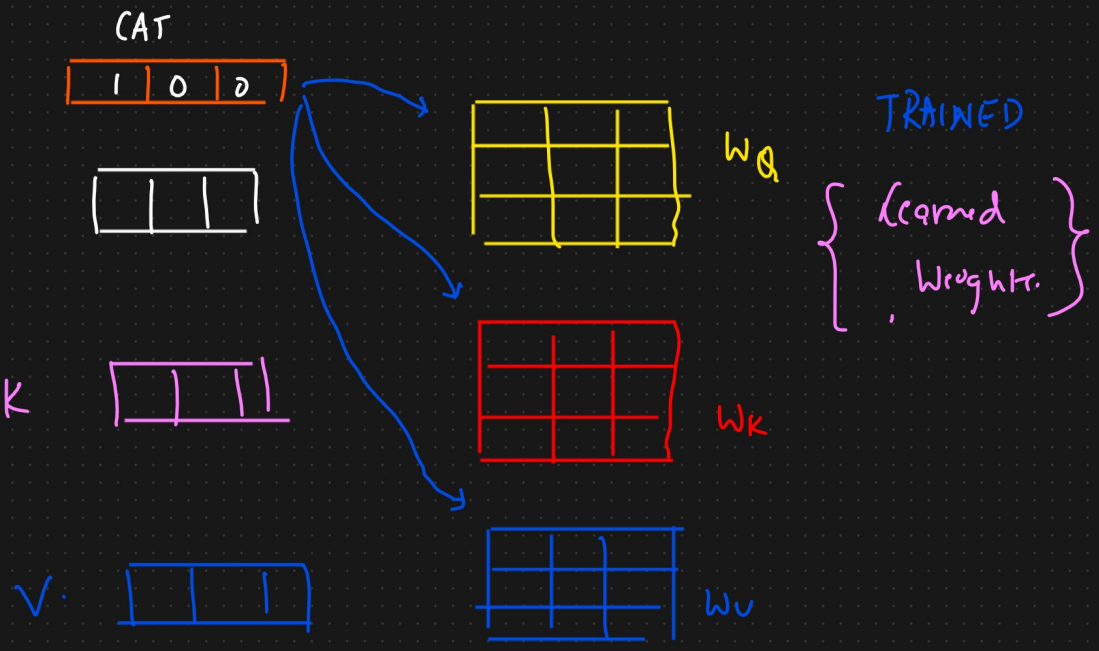
$E_{The} = [1\ 0\ 1\ 0]$

$E_{CAT} = [0\ 1\ 0\ 1]$

$E_{Sat} = [1, 1, 1, 1]$

② Linear Transformation

We create Q, K, V by multiplying the embeddings by learned weights matrices W_Q, W_K and W_V.



Lets consider

$$W_Q = W_K = W_V = I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \leftarrow$$

$$Q_{The} = [1\ 0\ 1\ 0] \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = [1\ 0\ 1\ 0]$$

$$K_{The} = [1\ 0\ 1\ 0]$$

$$V_{The} = [1\ 0\ 1\ 0]$$

① $Q_{The} = K_{The} = V_{The} = [1\ 0\ 1\ 0]$

② $Q_{CAT} = K_{CAT} = V_{CAT} = [0\ 1\ 0\ 1]$

③ $Q_{SAT} = K_{SAT} = V_{SAT} = [1,1,1,1]$

③ Compute Attention Scores

$$[1\,0\,1\,0]\begin{bmatrix}1\\0\\1\\0\end{bmatrix}$$

The

$$\text{Score}(Q_{The}, K_{The}) = [1010]\cdot[1010]^T = 2$$

$$\text{Score}(Q_{The}, K_{CAT}) = [1010]\cdot[0101]^T = 0$$

$$\text{Score}(Q_{The}, K_{SAT}) = [1010]\cdot[1,1,1,1]^T = 2$$

For the token CAT

$$\text{Score}(Q_{CAT}, K_{The}) = [0101]\cdot[1010]^T = 0$$

$$\text{Score}(Q_{CAT}, K_{CAT}) = [0101]\cdot[0101]^T = 2$$

$$\text{Score}(Q_{CAT}, K_{SAT}) = [0101]\cdot[1111]^T = 2.$$

For the Token SAT

$$\text{Score}(Q_{SAT}, K_{The}) = [1,1,1,1]\cdot[1010]^T = 2$$

$$\text{Score}(Q_{SAT}, K_{CAT}) = 2$$

$$\text{Score}(Q_{SAT}, K_{SAT}) = 4$$

④ Scaling ÷

We take up the scores and scale down by dividing the scores by the $\sqrt{d_k}$ $\Rightarrow$ $d_k = 4$ $\sqrt{d_k} = 2$.

Scaling in the attention mechanism is crucial to prevent the dot product from growing too large. $\Rightarrow$ Ensure stable gradients during Training.

d_k is large $\longrightarrow$

① Gradient Exploding

② Softmax Saturation $\{$ $\}$ $\rightarrow$ Vanishing Gradient Problem.

$Q = [2\ 3\ 4\ 1]$ $K_1 = [1\ 0\ 1\ 0]$ $K_2 = [0\ 1\ 0\ 1]$

Without Scaling

$Q \cdot K_1^T = 2\times1 + 3\times0 + 4\times1 + 1\times0 = 2 + 4 = 6.$

$Q \cdot K_2^T = 2\times0 + 3\times1 + 4\times0 + 1\times1 = 0 + 3 + 0 + 1 = 4$

*) Score $[6, 4]$ ⇒ Scaling Not Applied

$$\text{Softmax}([6,4]) = \left[\frac{e^6}{e^6+e^4}, \frac{e^4}{e^6+e^4}\right] = \left[\frac{e^6}{e^6(1+e^{-2})}, \frac{e^4}{e^4(e^2+1)}\right]$$

$$= \left[\frac{1}{(1+e^{-2})}, \frac{1}{(e^2+1)}\right]$$

$\approx [0.88, 0.12]$.

(*) Property of Softmax W_Q, W_K, W_V

$([10, 1]) = [0.99, 0.01]$

Dot product = Large values

Most of the attention weight is assigned to the first key vector, very little to the second vector,

With Scaling

① Compute Scaled Dot Product

$\sqrt{d_k} = \sqrt{4} = \sqrt{2}$ ⇒ Variance 2, 3

dimension ↑ Variance ↑2

$\sqrt{d_k}$ ↑ same

$[6, 4]$ ⇒ Scale ⇒ $\left[\frac{6}{2}, \frac{4}{2}\right] = [3, 2]$.

$$\text{Softmax}([3,2]) = \left[\frac{e^3}{e^3+e^2}, \frac{e^2}{e^3+e^2}\right] = \left[\frac{e^3}{e^3(1+e^{-1})}, \frac{e^2}{e^2(e^1+1)}\right]$$

$= [0.73, 0.27]$ ⇒ Attention Weights

(*) Here, the attention weights are more balanced compared to the unscaled case

Summary of Importance
Stabilizing Training: Scaling prevents extremely large dot products, which helps in stabilizing the gradients during backpropagation, making the training process more stable and efficient.

Preventing Saturation: By scaling the dot products, the softmax function produces more balanced attention weights, preventing the model from focusing too heavily on a single token and ignoring others.
Improved Learning: Balanced attention weights enable the model to learn better representations by considering multiple relevant tokens in the sequence, leading to better performance on tasks that require context understanding.

Scaling ensures that the dot products are kept within a range that allows the softmax function to operate effectively, providing a more balanced distribution of attention weights and improving the overall learning process of the model.

(4) Scaling $= \sqrt{d_k} = \sqrt{4} \Rightarrow 2$

Similarly Scaling will be done for all other Tokens.

$\text{Scaled\_Score}(Q_{The}, K_{The}) = 2/2 = 1$

$\text{Scaled\_Score}(Q_{The}, K_{CAT}) = 0/2 = 0$

$\text{Scaled\_Score}(Q_{The}, K_{SAT}) = 2/2 = 1$

(5) Apply Softmax

$\text{ATTENTION WEIGHTS}_{\text{"The"}} = \text{Softmax}([1,0,1]) = [0.4223, 0.1554, 0.4223]$

$\text{ATTENTION WEIGHTS}_{\text{"CAT"}} = \text{Softmax}([0,2,2]) = [0.1554, 0.4223, 0.4223]$

$\text{ATTENTION WEIGHTS}_{\text{"SAT"}} = \text{Softmax}([2,2,4]) = [0.2119, 0.2119, 0.5762]$.

(6) Weight Sum of Values

We multiply the attention weights by corresponding value vectors

For the Token The =

$$Output_{(The)} = 0.4223 * V_{The} + 0.1554 * V_{CAT} + 0.4223 \cdot V_{Sat}$$

$$= 0.4223 \begin{bmatrix} 1 & 0 & 1 & 0 \end{bmatrix} + 0.1554 * \begin{bmatrix} 0 & 1 & 0 & 1 \end{bmatrix} + 0.4223 * \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}$$

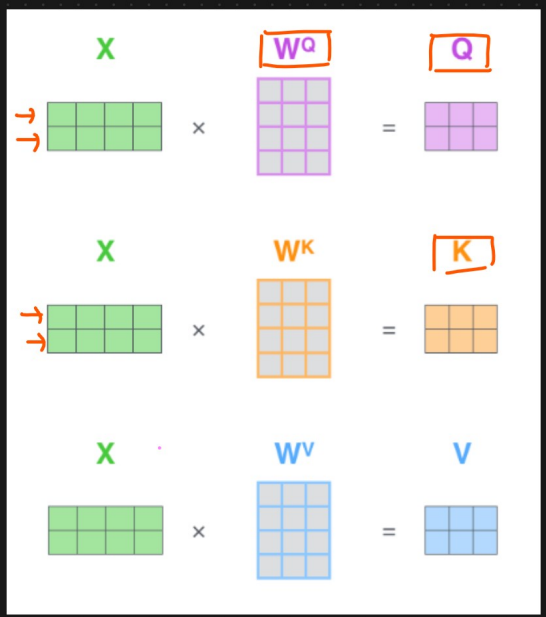
$$= [0.4223, 0, 0.4223, 0] + [0, 0.1554, 0, 0.1554] + [0.4223, 0.4223, 0.4223, 0.4223]$$

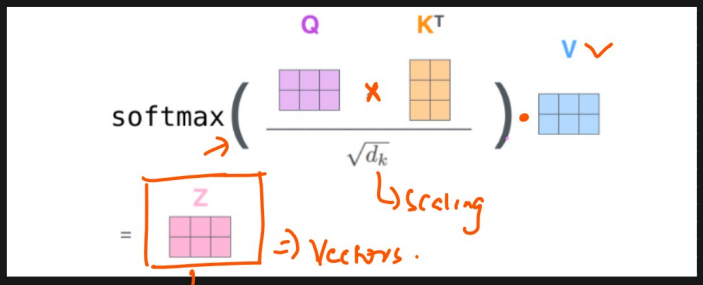
$$= [1.2669, 0.9999, 1.2669, 0.9999].$$

The $\boxed{1\,0\,1\,0}$ ⇒ Self Attention ⇒ [1.2669, 0.9999, 1.2669, 0.9999]. ⇓ Contextual vector

① ↳ Q, K, V [W^Q, W^K, W^V]
② ↳ Attention Score
③ ↳ Scaled
④ ↳ Softmax
⑤ ↳ Weighted Sum of Value (Softmax × V)

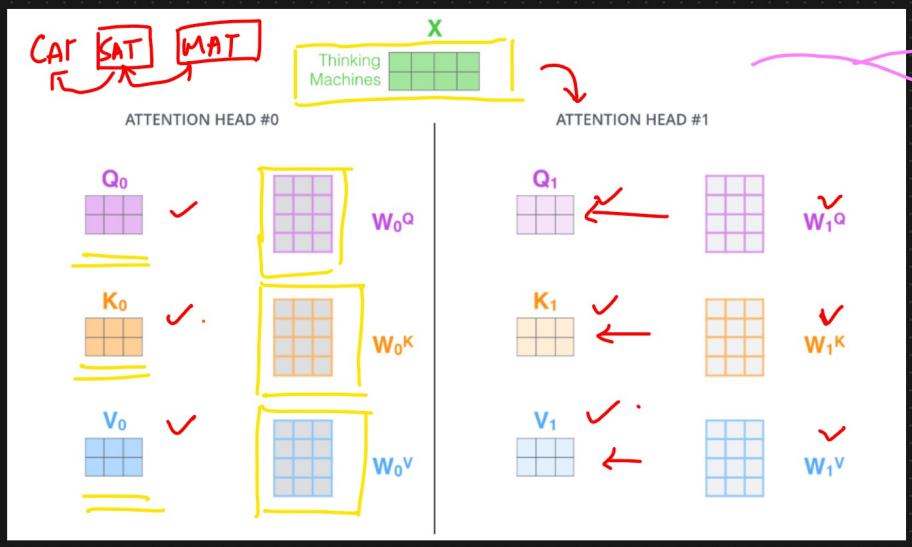
④ Multi Head Attention





Attention Head.

→ Self Attention with Multi Heads



It expands model's ability to focus on different position of tokens.

Score, Softmax X V

[_ _ _]

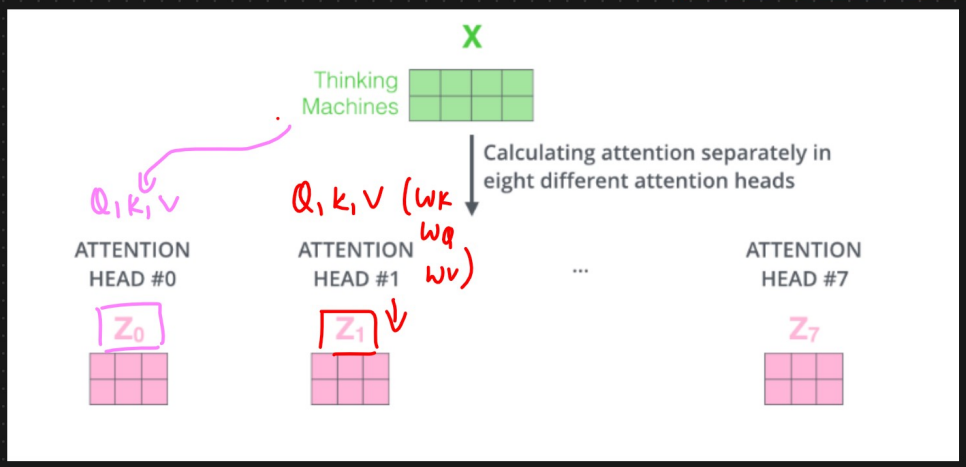
Z_0

Vectors

[_ _ _]

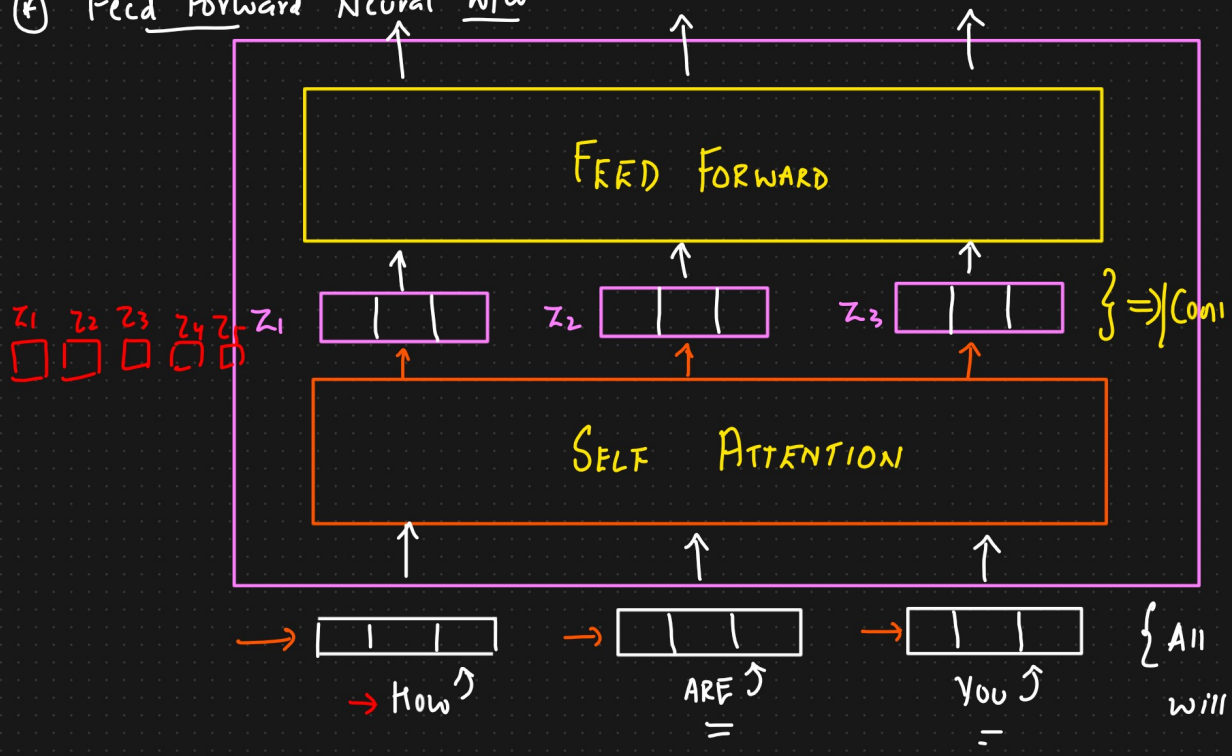
Z_1

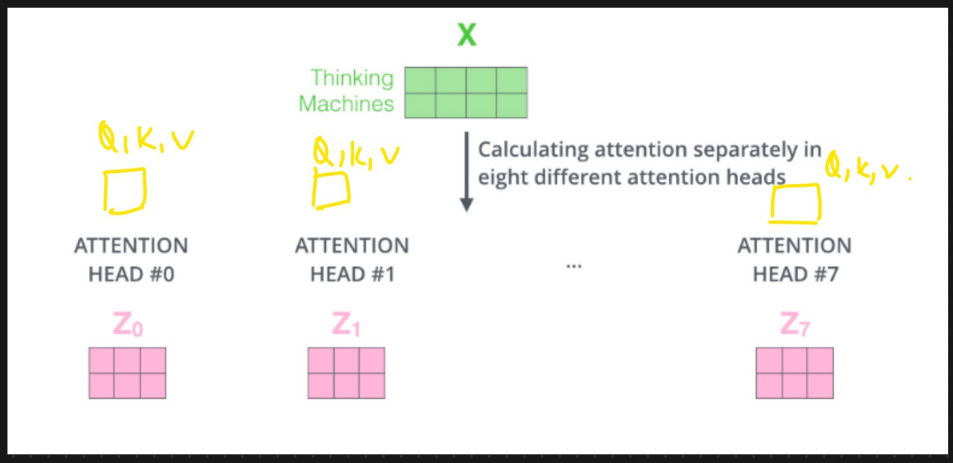
Multi Head Attention

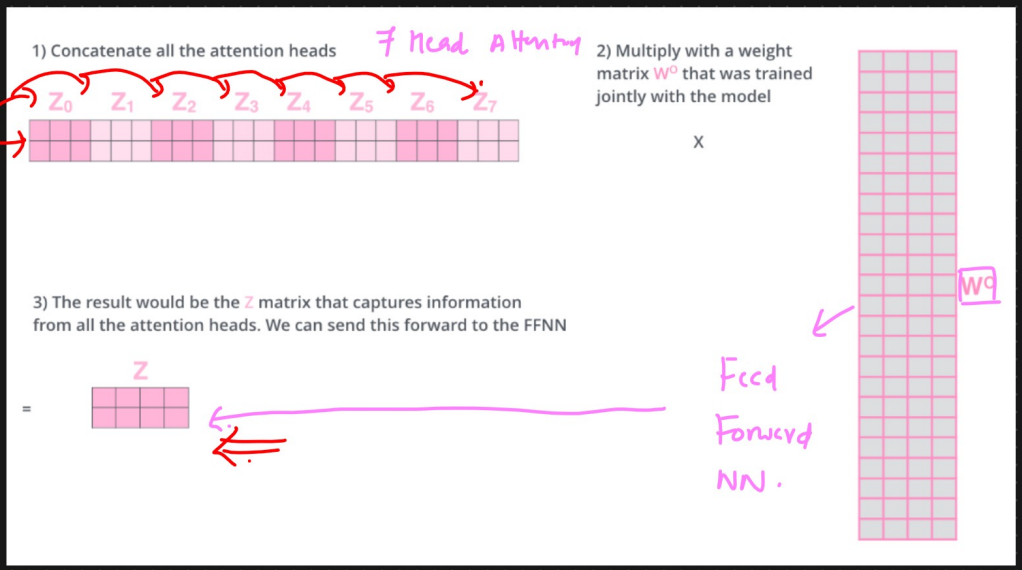


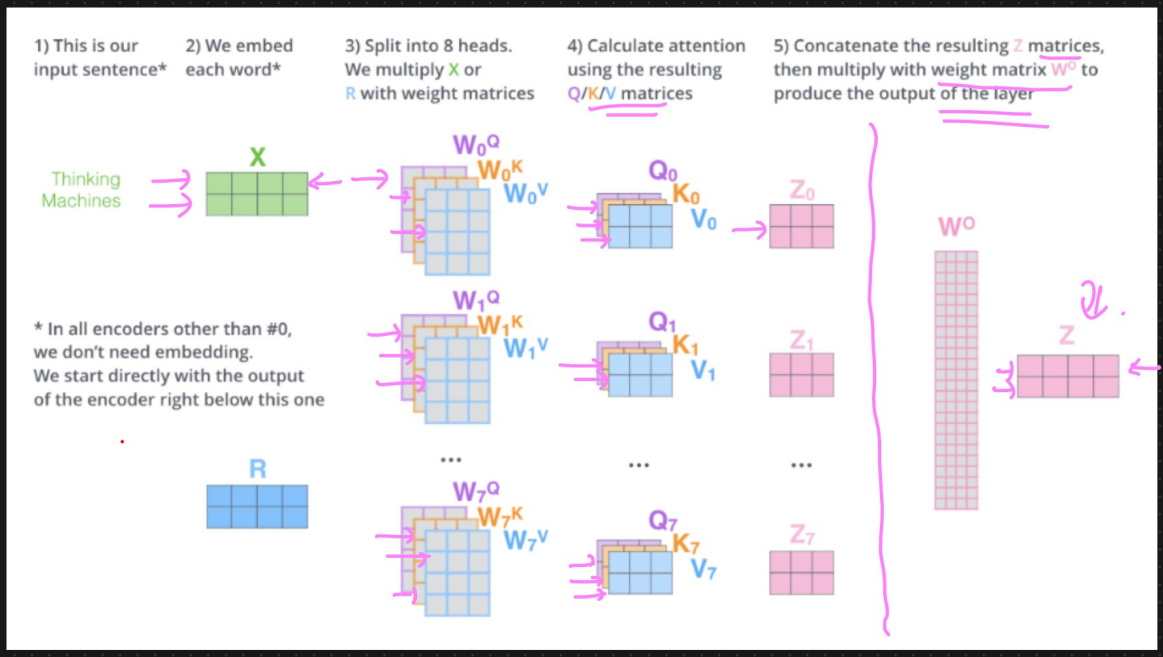
Z Z Z

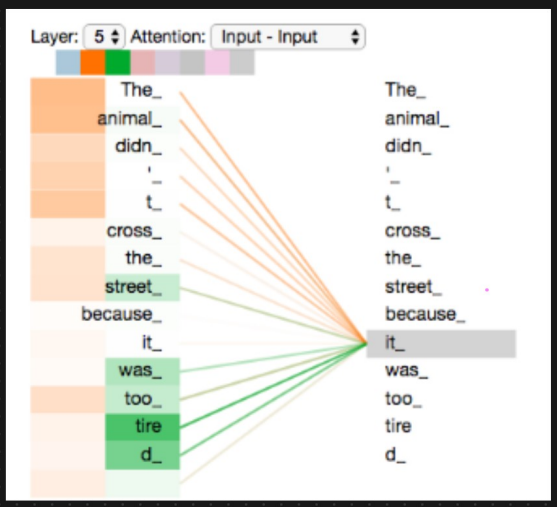
(4) Feed Forward Neural N/W

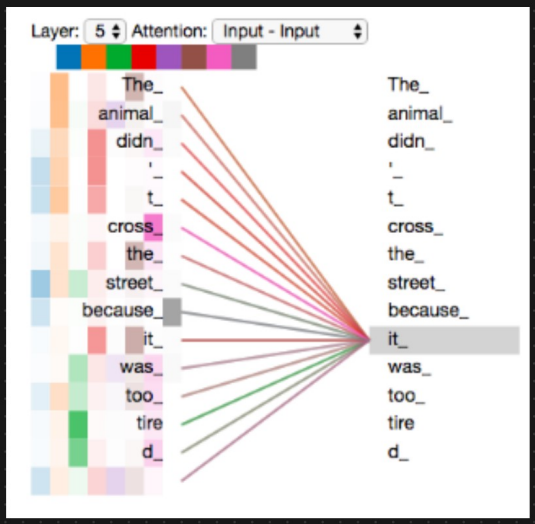










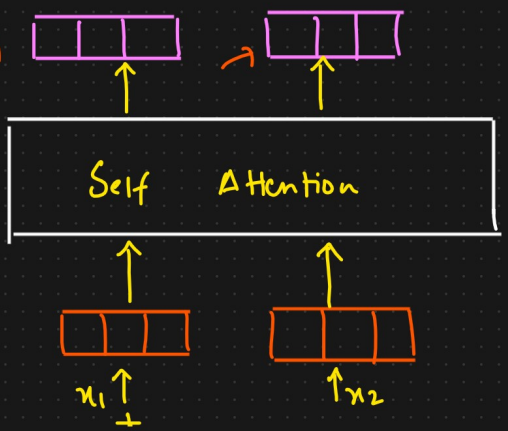


⑥ Positional Encoding – Representing Order of Sequence

① Lion KILLS TIGER
② TIGER KILLS LION

⇓

Missing



Advantage

① Word Tokens it can process parallelly

⇓

DRAW BACK

Lack the sequential structure of the words {order}

Attention Is All You Need

↓ Positional Encoded Vectors

Book, Journal, Novel
↳ 1 lakh words
↓ Backpropagation

Types of Position Encoding

1) Sinusoidal Position Encoding →
2) Learned positional Encoding ⇒ Positional Encoding are learned during Training ←

① Sinusoidal Positional Encoding : It uses sine and cosine functions of different frequencies to create positional encodings

Formula :

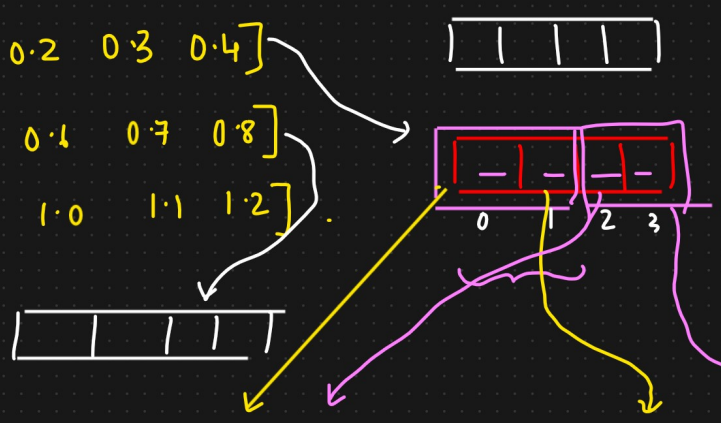
$$P.E(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$

$$P.E(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$

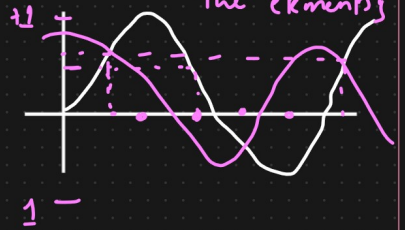
Where pos is the position
i is the dimension
d_{model} is the dimensionality of the embeddings.

Eg: The Cat Sat

The → [0.1 0.2 0.3 0.4]
CAT → [0.5 0.6 0.7 0.8]
SAT → [0.9 1.0 1.1 1.2]



{ Miss the order of the elements }



$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right) \qquad PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$

For our example $d_{model} = 4$

For position $pos = 0$

$$PE(0,0) = \sin\left(\frac{0}{10000^{0/4}}\right) = \sin(0) = 0$$

$$PE(0,1) = \cos\left(\frac{0}{1}\right) = \cos(0) = 1$$

$$PE(0,2) = \sin\left(\frac{0}{10000^{2/4}}\right) = \sin(0) = 0$$

$$PE(0,3) = \cos(0) = 1$$

$PE = [0, 1, 0, 1]$

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right) \qquad PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$

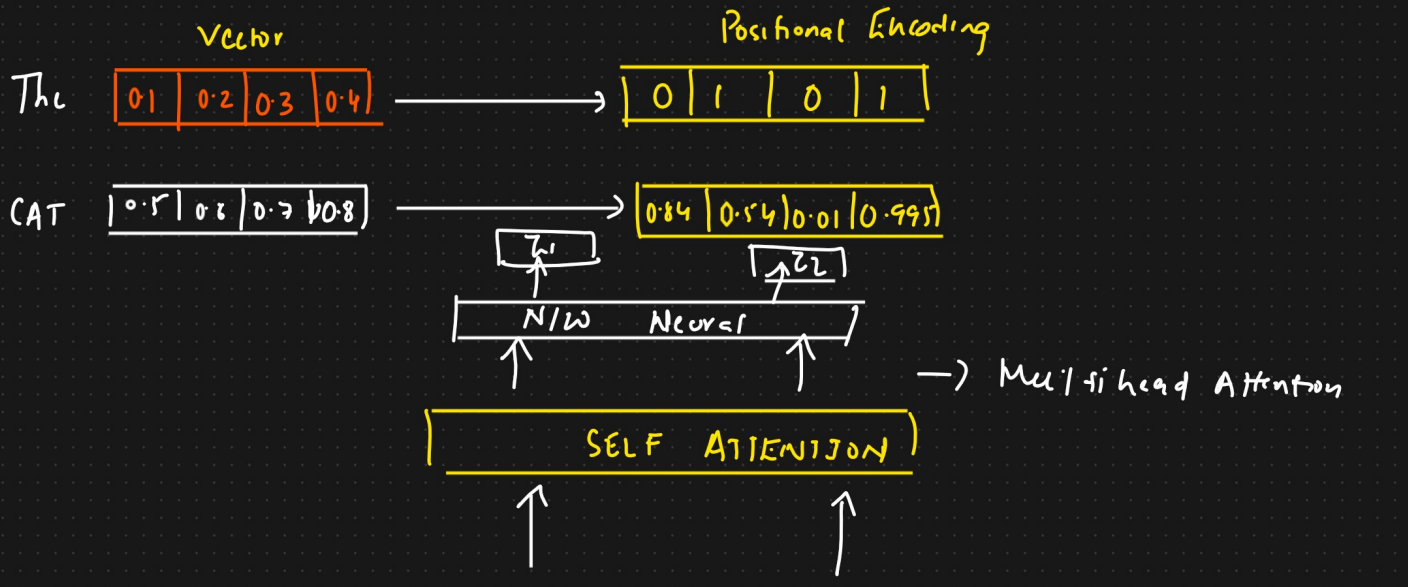
For $pos = 1$

$$PE(1,0) = \sin\left(\frac{1}{10000^{0/4}}\right) = \sin(1) = 0.8415$$

$$PE(1,1) = \cos\left(\frac{1}{10000^{2/4}}\right) \approx 0.5403$$

$$PE(1,2) = \sin\left(\frac{1}{10000^{2/4}}\right) \approx 0.01$$

$$PE(1,3) = \cos \approx 0.99995$$



Positional Encoding

0.84 | 0.54 | 0.01 | 0.99

+

0.5 | 0.6 | 0.7 | 0.8

↑ CAT

0 | 1 | 0 | 1 ← Positional Encoding

+

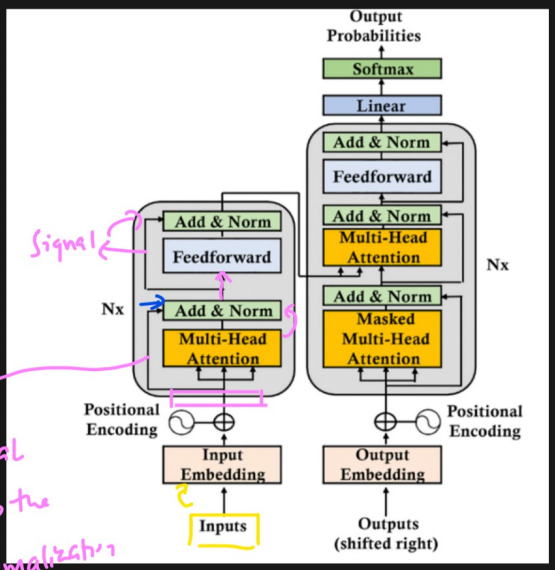
0.1 | 0.2 | 0.3 | 0.4

↑ The.

⑨ Layer Normalization In Transformers

Transformers

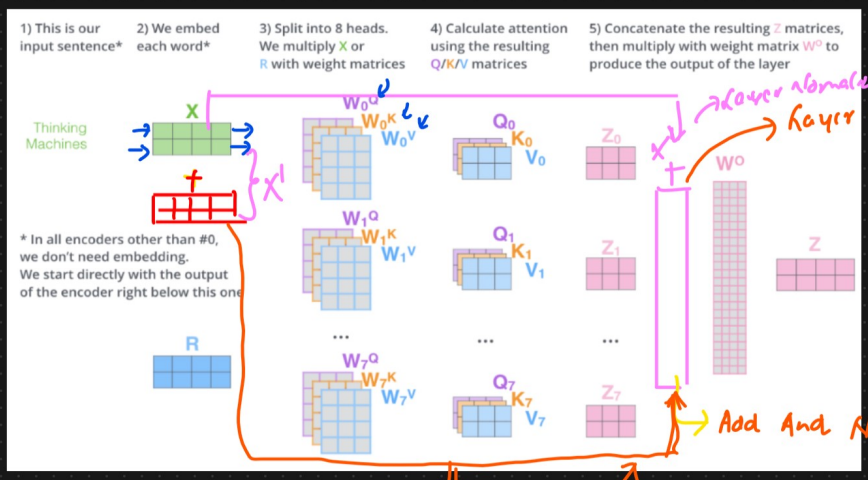
Residuals



Residuals
↳ Additional Signal to the Layer Normalization.

① Self Attention layer
② Multi head Attention
③ Positional Encoding
④ Layer Normalization

ADD AND Normalize



Normalization

→ Batch Normalization
→ Layer Normalization

f1 House Size	f2 No of Rooms	Price
1200	2	45
1500	3	70

Normalization : Standard Scaling

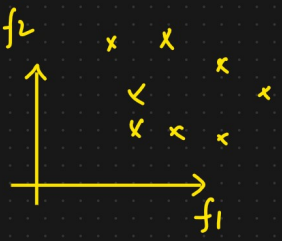
μ, σ ← | 2000 | 3.5 | 80

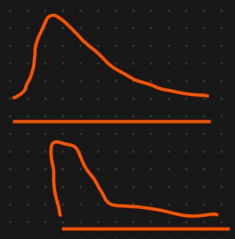
$$Z_{score} = \frac{x_i - \mu}{\sigma} \Rightarrow \boxed{\mu = 0, \sigma = 1}$$

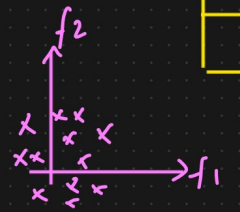
$f_1 \longrightarrow f_1' \Leftarrow \mu = 0, \sigma = 1$

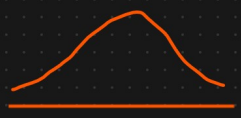
Deep learning : I/p Image ⇒ Min Max Scaler

⇒ Min Max Scaler ⇒





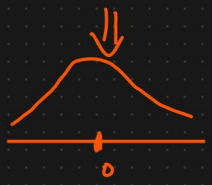




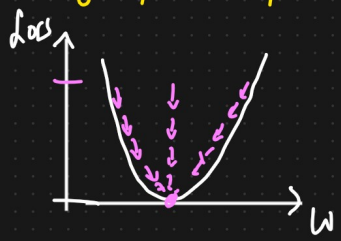
Advantages

① Improved Training Stability ⇓ Vanishing and exploding Gradient problem

$\{\mu = 0, \sigma = 1\}$ $\mu = 0, \sigma = 1$



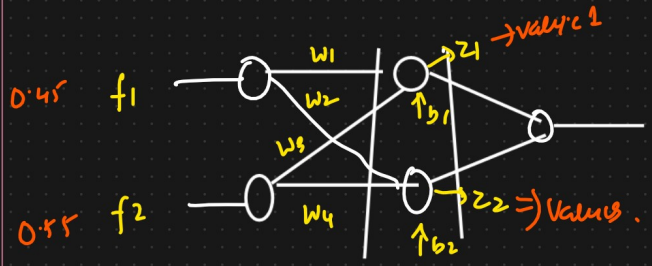
② Faster Convergence



⇒ Back propogation ⇒ Stable updates.

{Batch Normalization}

Normalize Normalized



	f_1 House size	f_2 Rooms	Price	Z_1 =	Z_2 =
→	0.45	0.55	45	—	—
	0.60	0.20	—	—	—
	—	—	—	—	—

$Z_1 = \sigma[(0.45 * W_1 + 0.55 * W_3) + b_1] = \text{Value 1}$ } Does this distribution

$Z_2 = \sigma(\quad + b_2) = \text{Value 2}$ }. $\boxed{\mu = 0, \sigma = 1}$ μ_1, σ_1 μ_2, σ_2

Batch Normalization Vs Layer Normalization

f_1	f_2	Z_1	Z_2	
—	—	—	—	→ μ_1, σ_1
—	—	—	—	→ $\mu_2 \sigma_2$
—	—	—	—	→ $\mu_3 \sigma_3$

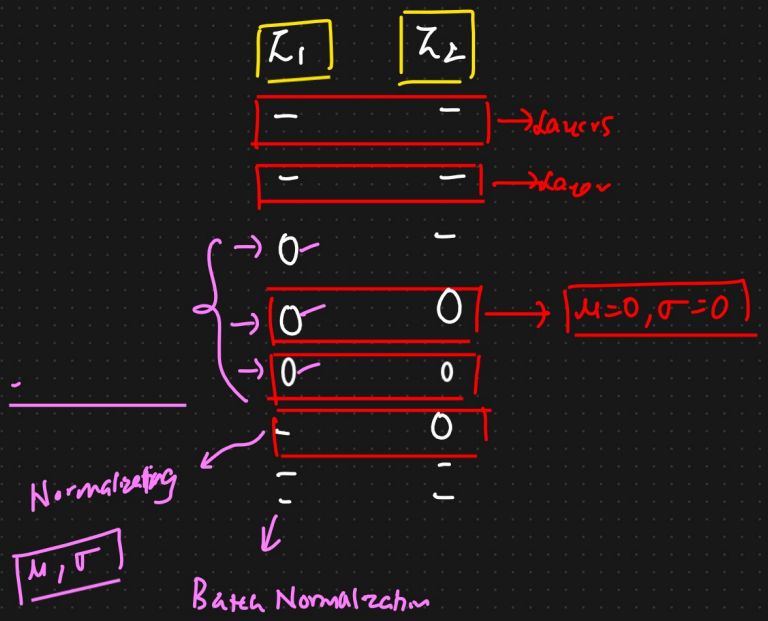
$Z_{score} = \frac{x_i - \mu_1}{\sigma_1}$

$Z_{score} = \frac{x_i - \mu_2}{\sigma_2}$

$Z_{score} = \frac{x_i - \mu_3}{\sigma_3}$ ..

Layer Normalization

$\gamma, \beta \rightarrow$ learnable parameters



Normalization

γ, β

$Z_1 = \sigma[w_1^T x + b_1]$

$Y = \gamma\left[\frac{Z_1 - \mu_1}{\sigma_1}\right] + \beta$

learnable parameters

Scale And Shift parameters

1) "CAT" = [2.0, 4.0, 6.0, 8.0] ← {Vectors} — Normalized.

2) Parameters = γ = [1.0, 1.0, 1.0, 1.0] → Learned Scale

β = [0.0, 0.0, 0.0, 0.0] → Shift

⇒ Scale And Shift param

$Z_{score} = \frac{x_i - \mu}{\sigma}$

i) Compute the mean

$\mu = \frac{1}{4}(2.0 + 4.0 + 6.0 + 8.0)$

$= \frac{20.0}{4} = 5.0$

ii) Compute the variance (σ^2)

$\sigma^2 = \frac{1}{4}\left[(2.0-5.0)^2 + (4.0-5.0)^2 + (6.0-5.0)^2 + (8.0-5.0)^2\right] = 5.0$

iii) Normalize the i/p

$\hat{x_i} = \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}}$ $\quad \epsilon = 1e^{-5}$ ⇒ Avoid division by 0

$\sqrt{\sigma^2 + \epsilon} = \sqrt{5.0 + 1e-5} \approx \sqrt{5.00001} = 2.236$

$$\hat{x}_1 = \frac{2.0 - 5.0}{2.236} \approx -1.34$$

$$\hat{x}_2 = \frac{4.0 - 5.0}{2.236} \approx -0.45$$

$$\hat{x}_3 = \frac{6.0 - 5.0}{2.236} \approx 0.45$$

$$\hat{x}_4 = \frac{8.0 - 5.0}{2.236} \approx 1.34.$$

Normalized vector

$$\hat{x} = [-1.34, -0.45, 0.45, 1.34]$$

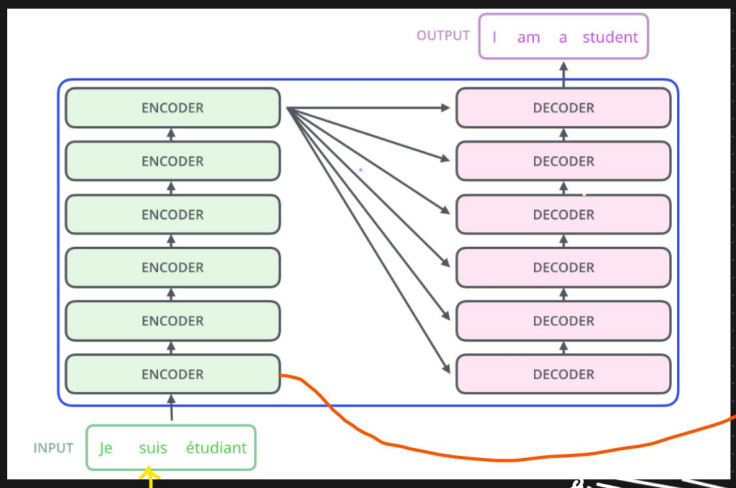
4) Scale And Shift

$$\gamma = [1.0, 1.0, 1.0, 1.0] \quad \beta = [0.0, 0.0, 0.0, 0.0]$$

$$y_i = \gamma_i \hat{x}_i + \beta_i$$

$$y = [-1.34, -0.45, 0.45, 1.34].$$

⑥ Encoder Architecture [Research Paper]



Sequence to Sequence ↓ (complex)

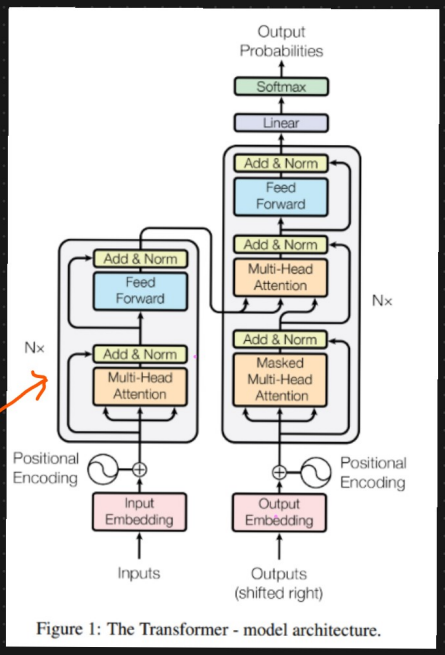


Figure 1: The Transformer - model architecture.

Feed Forward NN. →

(Add And Norm)

Multi Head Attention → $Z_1, Z_2, Z_3, Z_4, Z_5, Z_6, Z_7, Z_8$ {Research Paper}.

Residuals

Text Embeddings + Positional Encoding ⇒ 512. → {Research paper}.

I/P Sequence

Every word = 512.

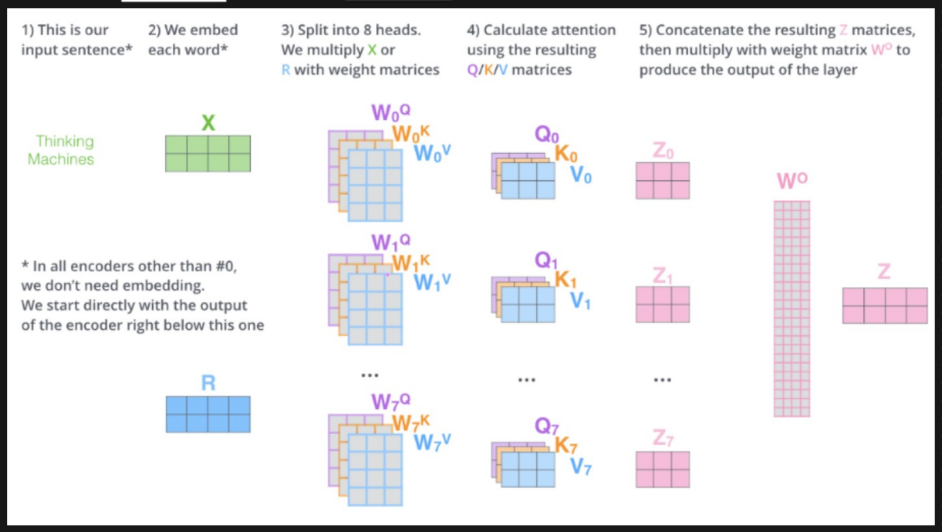
$Q_1 = 64$

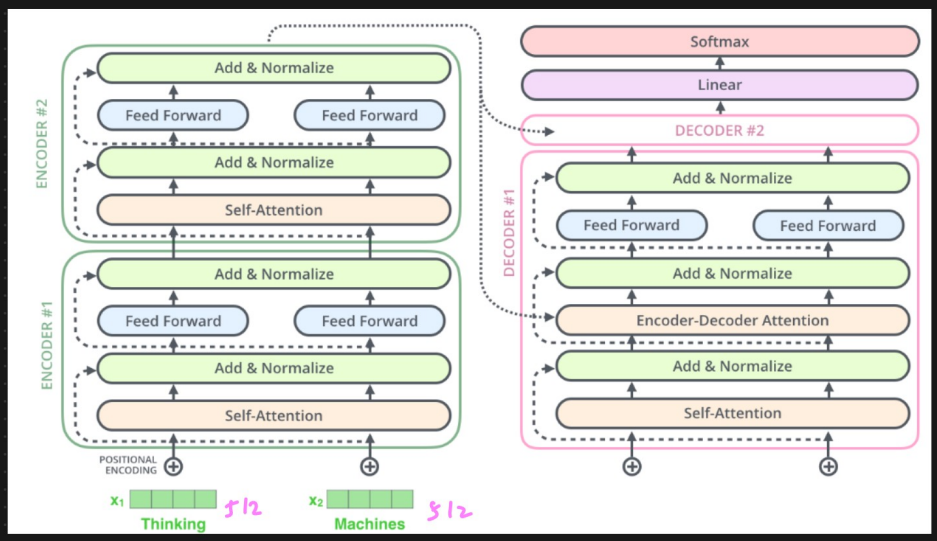
$K = 64$

$V = 64$

$\sqrt{64} = 8$ //.

Self Attention to Feed Forward Neural N/W





Embedding vector: 512

Q, K, V = 64

Head Attention: 8

⑧ Residual connection : Skip connection NN.

1) Addressing the Vanishing Gradient Problem

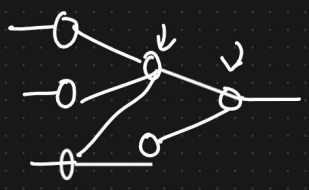
Residuals : Residual connection create a short paths for gradients to flow directly through the n/w. Gradient remains sufficiently large.

2) Improve Gradient flow

Convergence will be faster.

3) Enables Training of Deeper Networks.

④ Feed Forward NN



linear function problem

{Non linear functions}.

① Adding Non linearity

② Processing Each position Independently.

Self Attention → Capture relationships

FFN → Each token representation Independently.

↓

Transforming these representations further

and allows the model to learn

Richer Representation.

ANN ⇒

③ FFN → Deeper ⇒ Adds Depth to the Model.

Depth ↑↑ ⇒ More learnings → DATA

⑤ Decoders In Transformers

3 main Components

The transformer decoder is responsible for generating the output sequence one token at a time, using the encoder's output and the previously generated tokens.

① Masked Multi Head Self Attention ✓.

② Multi Head Attention (Encoder Decoder Attention)

③ Feed Forward Neural Network.

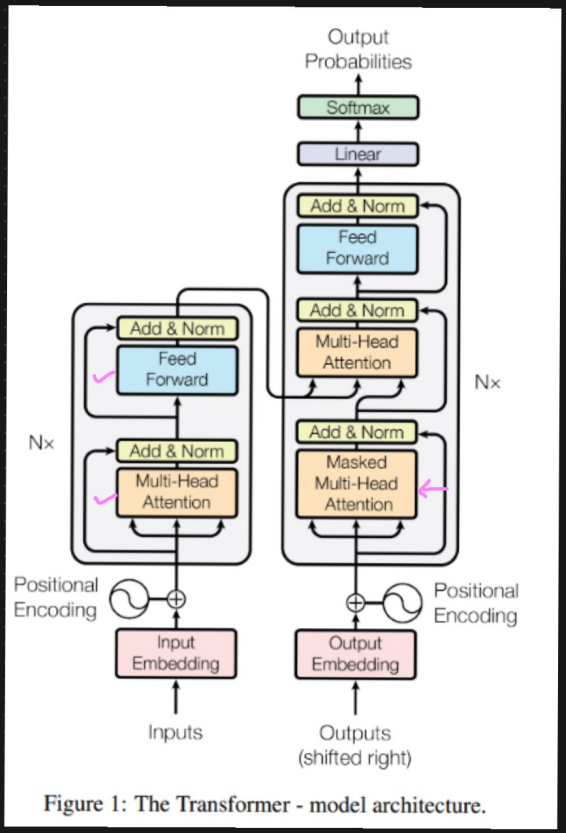
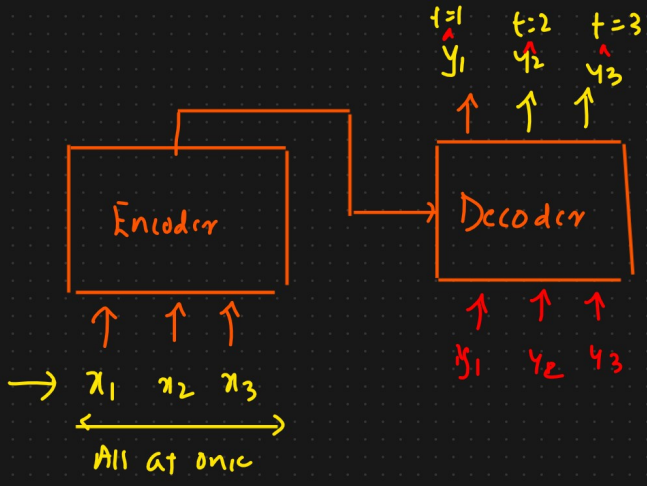


Figure 1: The Transformer - model architecture.



$\langle y_1, y_2, y_3 \rangle$

① Training Mechanism

② Inference Mechanism

⑧ Masked Multi Head Self Attention

① I/P Embedding And Positional Embedding ✓ → Zero padding → Sequence length Equal

② Linear Projection for Q, K, V

③ Scaled Dot Product Attention

✓④ Mask Application } ⇒ Try to understand the imp.

⑤ Multi Head Attention

⑥ Concatenation And Final Linear Projection

⑦ Residual connection And Layer Normalization

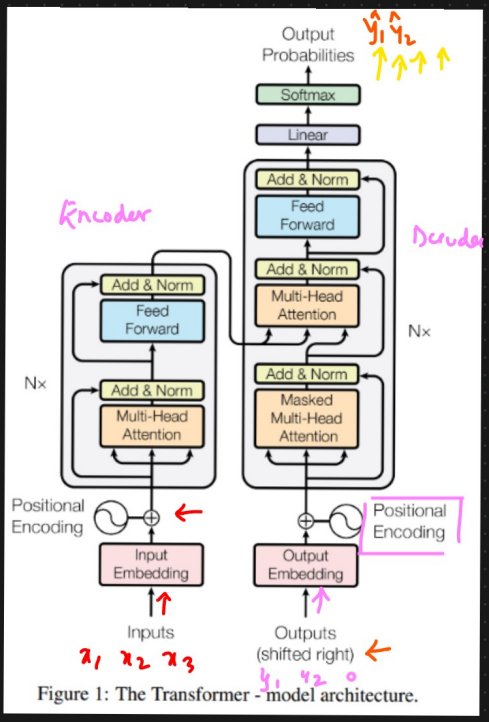


Figure 1: The Transformer - model architecture.

Dataset

Eng: $\langle x_1, x_2, x_3 \rangle$

Hindi: $\langle y_1, y_2 \rangle$ ← ⇓ $\langle y_1, y_2, 0 \rangle$ ↓ Zero padding

① Linear Projections Q, K, V

② Scaled Dot Product Attention

③ Mask Application → Look Ahead Mask, → Padding Mask

Masked Multi Head Attention.

I/P [4 5 6 7] O/P [1 2 3]

[1 2 3 0]

↓

4 dimension vector

i) Input Embedding and Positional Encoding

Output Embedding [STEP 1]

$$\begin{bmatrix} [0.1, 0.2, 0.3, 0.4], \\ [0.5, 0.6, 0.7, 0.8], \\ [0.9, 1.0, 1.1, 1.2], \\ [0.0 \quad 0.0 \quad 0.0 \quad 0.0] \end{bmatrix} + PE \Rightarrow 0$$

Step 2 : Linear Projection for Q, K, and V.

$W_Q = W_K = W_V = I$

Create query (Q), Key(K) and value(V) vectors

Q = Output Embedding * W_Q = Output Embedding

K = " * W_K = "

V = " * W_V = "

$$Q = K = V = \begin{bmatrix} [0.1, 0.2, 0.3, 0.4], \\ [0.5, 0.6, 0.7, 0.8], \\ [0.9, 1.0, 1.1, 1.2], \\ [0.0 \quad 0.0 \quad 0.0 \quad 0.0] \end{bmatrix} \begin{bmatrix} 0.1 \\ 0.2 \\ 0.3 \\ 0.4 \end{bmatrix}^T \begin{bmatrix} 0.5 \\ 0.6 \\ 0.7 \\ 0.8 \end{bmatrix} \cdot \begin{bmatrix} 0.9 \\ 1.0 \\ 1.1 \\ 1.2 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

③ Scaled Dot Product Attention Calculation

$$\text{Scores} = Q * K^T / \sqrt{dk}$$

$$= Q * K^T / 2.$$

$$\begin{bmatrix} [0.1*0.1 + 0.2*0.2 + 0.3*0.3 + 0.4*0.4, \\ 0.1*0.5 + 0.2*0.6 + 0.3*0.7 + 0.4*0.8, \\ 0.1*0.9 + 0.2*1.0 + 0.3*1.1 + 0.4*1.2 \\ 0.1*0 + 0.2*0 + 0.3*0 + 0.4*0] \end{bmatrix}$$

Scores : [[0.3, 0.7, 1.1, 0.0]
[0.7, 1.9, 3.1, 0.0]
[1.1 3.1 5.1 0.0]
[0.0 0.0 0.0 0.0]]

[ ]
[ ]
[ ]]

⑥ Masked Application

It helps manage the structure of the sequences being processed and ensures the models behaves correctly during training And Inferencing

Reasons

① Handling Variable length Sequences with Padding MASK

Purpose

① To handle sequences of different length in batch

② To ensure that padding tokens, which are added to make sequences of uniform length, do not affect the model prediction

Eg: i/p ← Sequence 1 [1, 2, 3]

o/p ← Sequence 2 [4, 5, 0]

→ 100.

0 is the padding token

→ Influence the Attention Mechanism

⇓

lead to Incorrect or biased predictions.

o/p

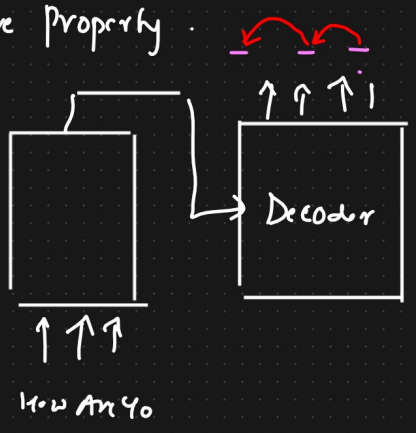
→ 4, 42 0 0 0 0 0

A padding mask ⇒ The tokens Are ignored.

Masking → Padding Mask
→ Look Ahead Mask.

Padding Mask [1 1 1]
[1 1 0]

② Look Ahead Mask → Maintain Auto Regressive Property.



① To ensure that each position in the decoder output sequence can only attend to the previous position, but no future position

② Sequence → Language Modelling, Translation

Eg: [4, 5, 0] → [1, 1, 0] → 1D Mask.

Convert 1D to 2D Mask

For each token in the sequence, the mask should indicate which tokens it can attend to.

$$\rightarrow \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Token 1 attends to token 1, 2

Attention Mechanism ⇐ 1, 2

③ Look Ahead Mask → Decoder Output

$$\begin{bmatrix} [1 & 0 & 0] \\ [1 & 1 & 0] \\ [1 & 1 & 1] \end{bmatrix}$$

④ Combine Padding And Looking Ahead Mask

Element Wise multiplication of 2 mask

$$\text{Combine Mask} = \begin{bmatrix} [1, 0, 0] \\ [1, 1, 0] \\ [0, 0, 0] \end{bmatrix}$$

④ Mask

Scores :
$$\begin{bmatrix} [0.3, 0.7, 1.1, 0.0] \\ [0.7, 1.9, 3.1, 0.0] \\ [1.1 \quad 3.1 \quad 5.1 \quad 0.0] \\ [0.0 \quad 0.0 \quad 0.0 \quad 0.0] \end{bmatrix}$$

Look Ahead Mask

$$\begin{bmatrix} [1 & 0 & 0 & 0] \\ [1 & 1 & 0 & 0] \\ [1 & 1 & 1 & 0] \\ [1 & 1 & 1 & 1] \end{bmatrix}$$

Padding Masking [extended to 2D Format]

$$\begin{bmatrix} [1,1,1,0] \\ [1,1,1,0] \\ [1,1,1,0] \\ [0,0,0,0] \end{bmatrix}$$

Combined Mask : Look Ahead Mask * Padding Mask

$$\begin{bmatrix} [1*1, 0*1, 0*1, 0*0] \\ [1*1, 1*1, 0*1, 0*0] \\ [1*1, 1*1, 1*1, 0*0] \\ [1*1, 1*1, 1*1, 1*0] \end{bmatrix} = \begin{bmatrix} [1,0,0,0], \\ [1,1,0,0], \\ [1,1,1,0], \\ [1,1,1,0] \end{bmatrix} \Rightarrow \begin{bmatrix} [1,-\infty,-\infty,-\infty], \\ [1,1,-\infty,-\infty], \\ [1,1,1,-\infty] \\ [1,1,1,-\infty] \end{bmatrix}$$

Masked Score

$$\begin{bmatrix} [0.3, -\infty, -\infty, -\infty] \\ [0.7, 1.9, -\infty, -\infty] \\ [1.1, 3.1, 5.1, -\infty] \\ [0.0, 0.0, 0.0, -\infty] \end{bmatrix}.$$

Zero out the influence when the Softmax is applied.

Attention weights.

⑧ Softmax

Softmax Score = Softmax (Masked Scores)

$$= [[1.0, 0.0, 0.0, 0.0],$$
$$[0.3, 0.7, 0.0, 0.0],$$
$$[0.1, 0.3, 0.6, 0.0],$$
$$[1.0, 0.0, 0.0, 0.0]]$$

⑧ Weigth Sum of Value

Attention O/p = Softmax Scores * V.

Masking

Masking in the transformer architecture is essential for several reasons. It helps manage the structure of the sequences being processed and ensures the model behaves correctly during training and inference. Here are the key reasons for using masking:

1. Handling Variable-Length Sequences with Padding Mask

Purpose
To handle sequences of different lengths in a batch.
To ensure that padding tokens, which are added to make sequences of uniform length, do not affect the model's predictions.

2. Maintaining Autoregressive Property with Look-Ahead Mask

Purpose
To ensure that each position in the decoder output sequence can only attend to previous positions and itself, but not future positions.
This is crucial for sequence generation tasks like language modeling and translation, where the model should not have access to future tokens when predicting the current token.

④ Encoder Decoder Multi Head Attention

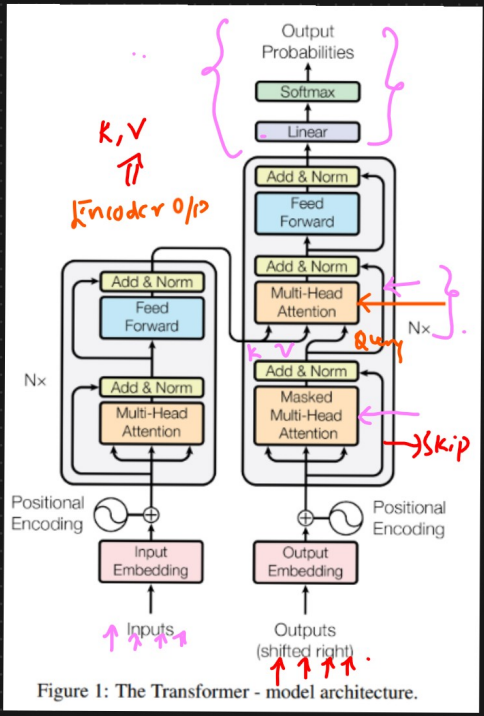


Figure 1: The Transformer - model architecture.

① Encoder O/p → Set of Attention vector K & V

② Masked Multihead → Attention vector Q {Query Vector}

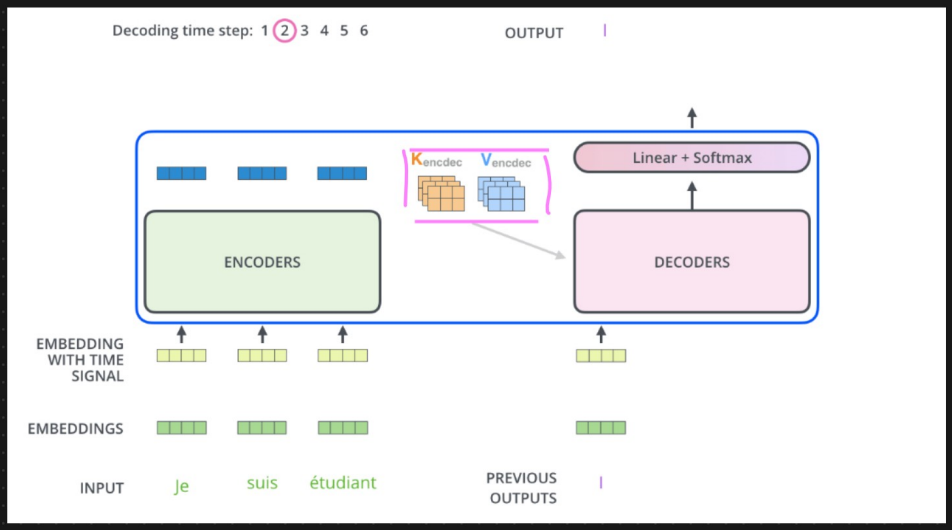
These are to be used by each decoder in its "encoder – decoder" attention layer

⇓

Helps the Decoder to focus on appropriate places in the i/p sequence

Dataset

i/p $\langle x_1\ x_2\ x_3 \rangle$ O/p $\langle y_1\ y_2\ y_3 \rangle$.



④ The Final Linear And Softmax Layer

{Vectors ⟶ o/p word}

{BLOG} ⇒ Transformers

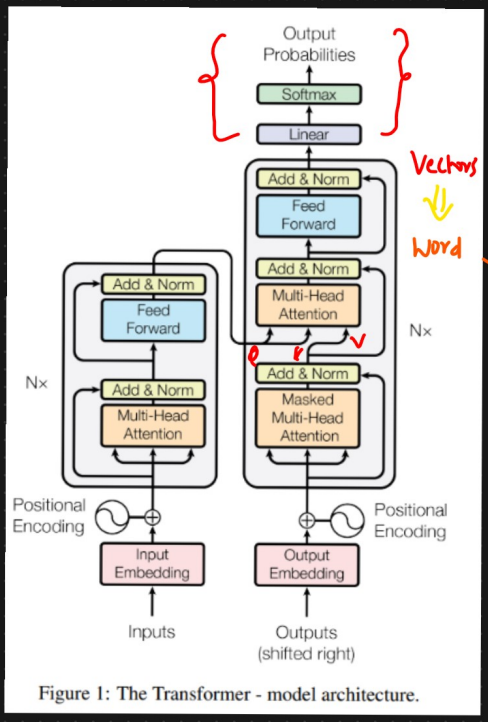
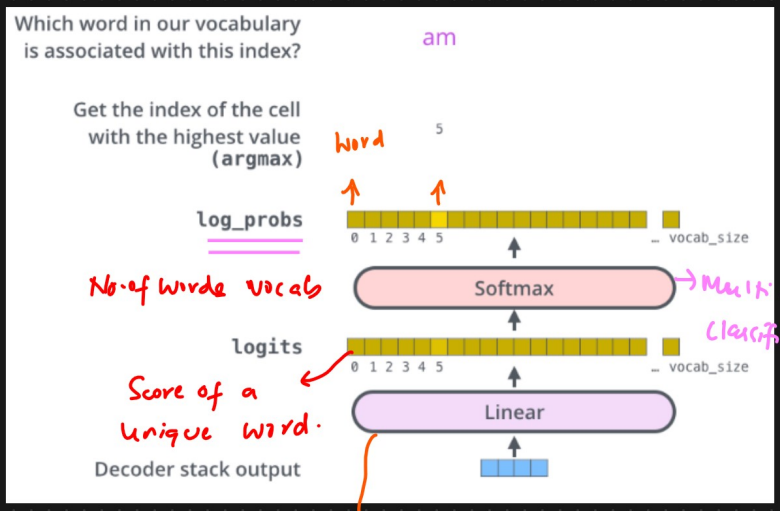


Figure 1: The Transformer - model architecture.



Linear ⇒ The linear layer is a simple fully connected neural n/w that projects the vector produced by the stack of Decoder ⇒ logits vector

⇒

Model ⇒ 10,000 ⇒ vocabulary ⇒ logits vector = 10000 cells wide

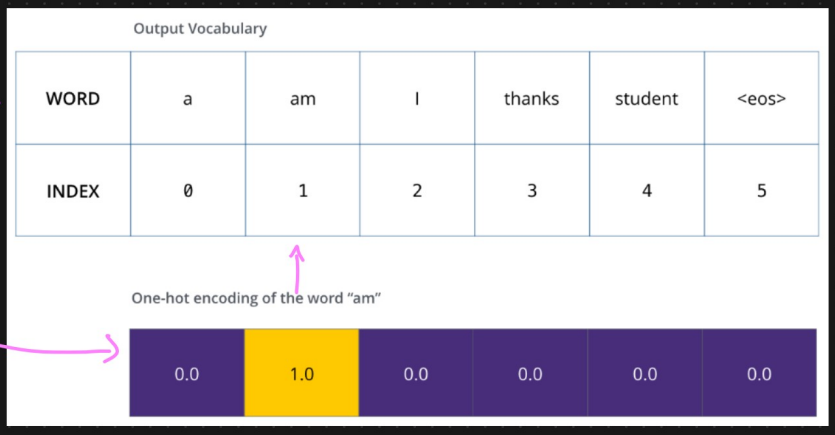
② Softmax layer turns those scores into probabilities (all add upto 1.0).

The cell with the highest probability is chosen, and the word associated with it is produced as the o/p. ⇒ time stamp.

Recap of Training

Output Vocabulary

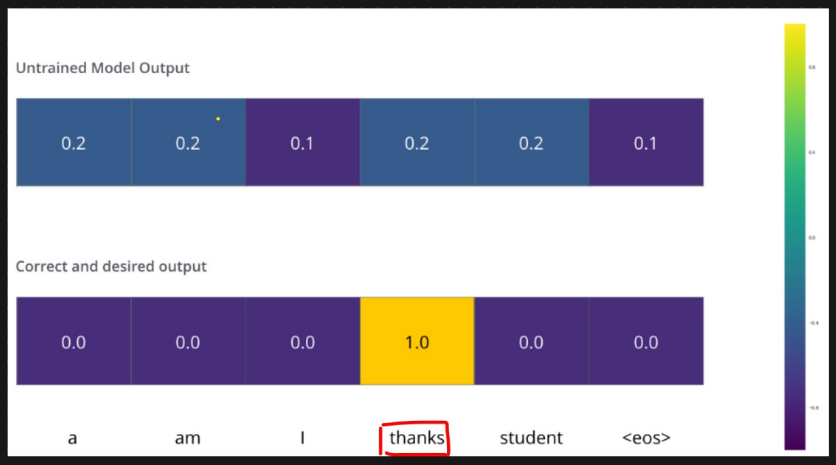
WORD	a	am	I	thanks	student	<eos>
INDEX	0	1	2	3	4	5



OHE ←

Merci → Thanks

I am a student <eos>.



⇒ Loss function ↓↓.

Back Propogation

i/p

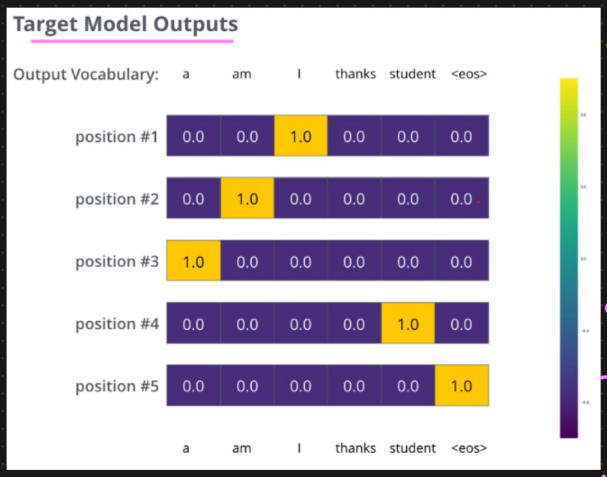
a, am, i, student

o/p

I am a student

↓ ↓ ↓ ↓

OHE OHE OHE OHE



Loss