Date

M T W T F S

Castelli

Feature Selection:

f_1 f_2 f_3 ..... label

irrelevant

f_1 w_1

f_2 w_2

f_3 w_3

$$\chi^2 = \sum \frac{(O_i - e_i)^2}{e_i}$$

observe values / expected value

	Buy	Not buy	O_i values	e_i	
Red	10	20	30	13.63	16.36
Blue	15	10	25	$\frac{30\times25}{55}$	$\frac{30\times30}{55}$
				$\frac{25\times25}{55}$	$\frac{25\times30}{55}$
	25	30	55	11.36	13.63

$$\chi^2 = \frac{(10 - (30\times25/55)}{55}$$

30 → Red and 25 → Blue

$$\chi^2 = \left[\frac{(10-13.63)^2}{13.63}, \frac{(20-16.36)^2}{16.36}, \frac{(15-11.36)^2}{11.36}, \frac{(10-13.63)^2}{13.63}\right]$$

$$= [0.97, 0.80, 1.16, 0.97]$$

largest distance

Date (M|T|W|T|F|S)

Castelli

target distance

less relevant

selection criteria

top k features

Accept or reject a feature

f_1	f_2	f_3	f_4	Color	decision
				red	buy
				red	don't buy

buy not buy

red

Degree of freedom

[0.97, 0.80, 1.16, 0.97] = 3.9

0 ... 9 - 1st deg

10 .... 99 - 2nd

100 ... 99 - 3rd

Date _____ M T W T F S

dec 1 2 3

20% 12...

10% 14...

5% 16...

1% 21...

80% confidence

$3.9 < 12$ True

I accept these feature with 80% confidence.

$3.9 < 21$ Accept with 99% confidence

$[0.97, 0.80, 20, 0.97]$ $23 < 21$

False

I reject these feature, w/

Date

M T W T F S

Castelli

Individual feature check

$$\text{Z-test} = \frac{0.97 - \mu}{\sigma}$$

mean

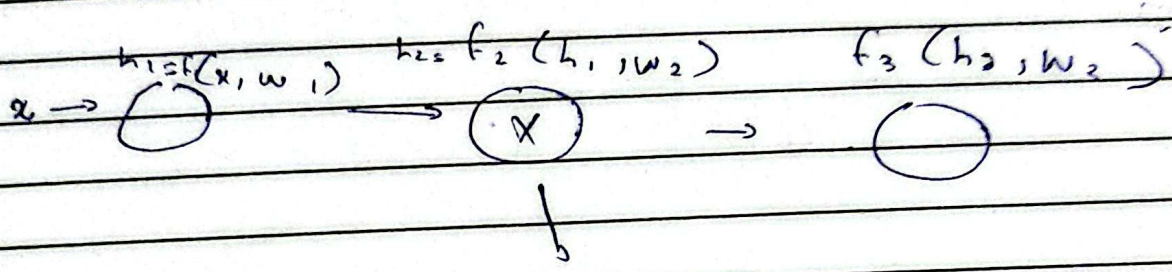
$$\frac{x - \mu}{\sigma}$$

Date (M T W T F S)

Residual Neural Network

CNN — Conv2D → MaxPooling → depth ↑

vanishing gradient issue

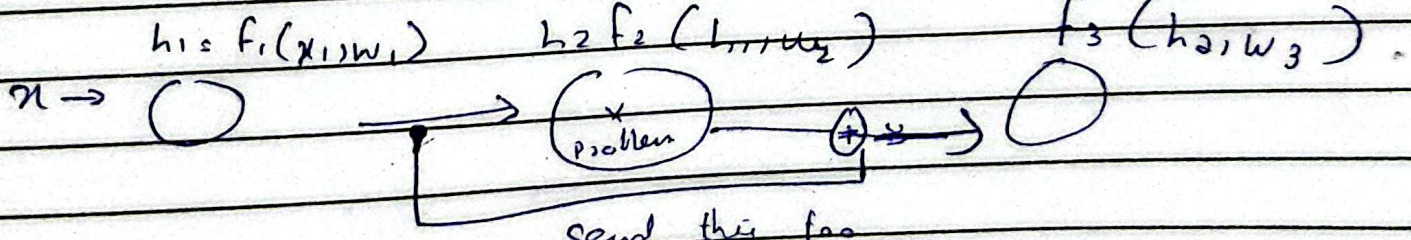


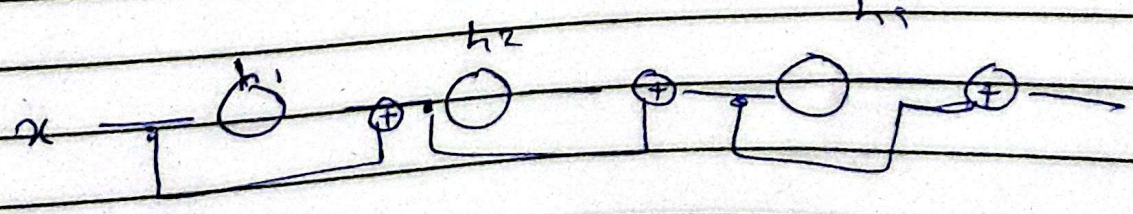
$y = f_3(h_{3,} w_3)$ * if problem is here it goes forward

$= f_3(f_2(h_1, w_2), w_3)$

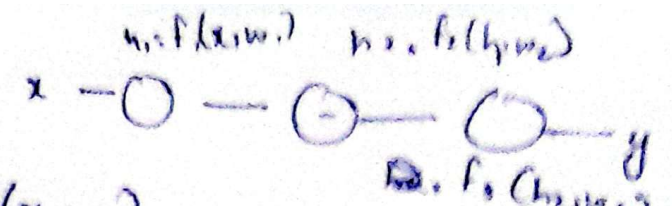
$= f_3(f_2(f_1(x_1, w_1), w_2), w_3)$ ↑ nested behavior

how residual solve it





Here →



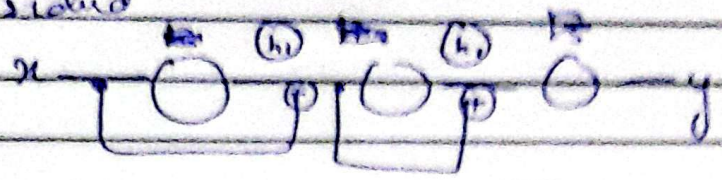
$h_1 = f_1(x, w_1)$

$h_2 = f_2(h_1, w_2)$

Date

M T W T F S

In residual



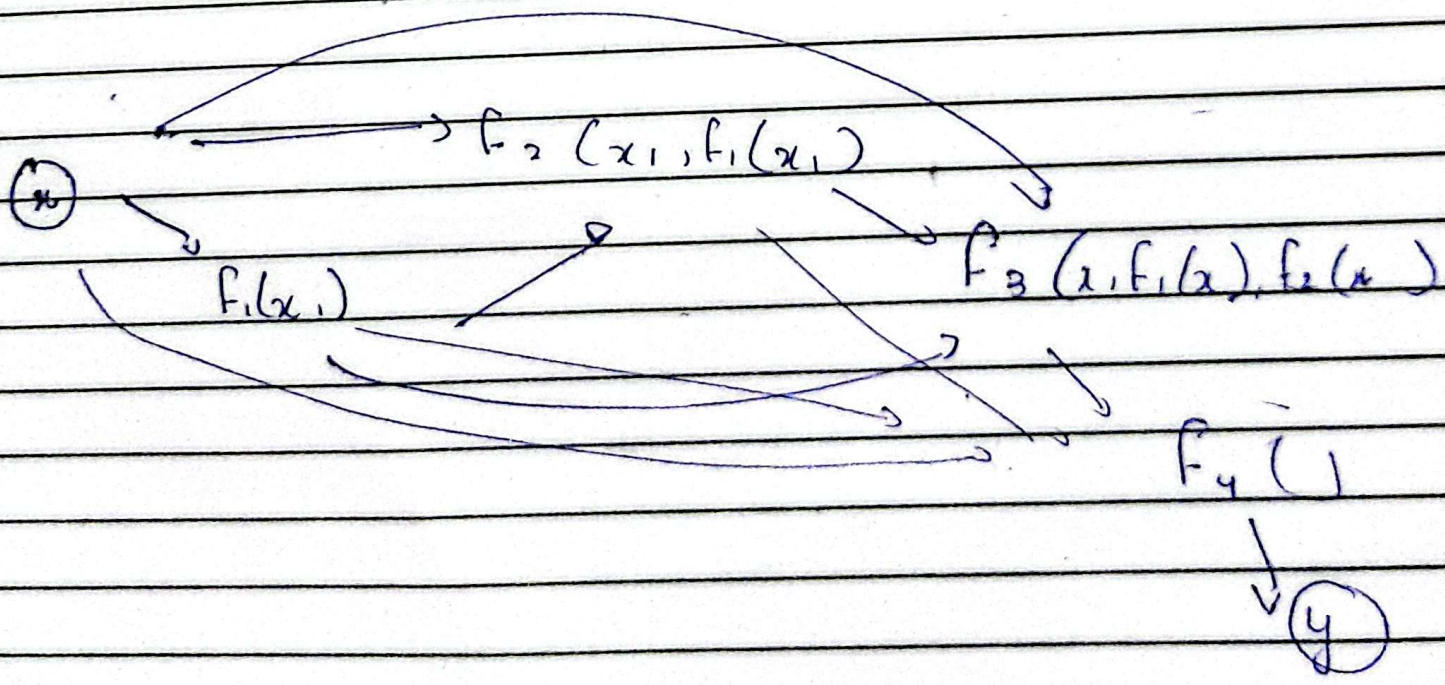
$h_1 = f_1(x, w_1) + x$

$h_2 = f_2(h_1 + w_2) + h_1$

$y = f_3(h_2, w_3) + h_2$

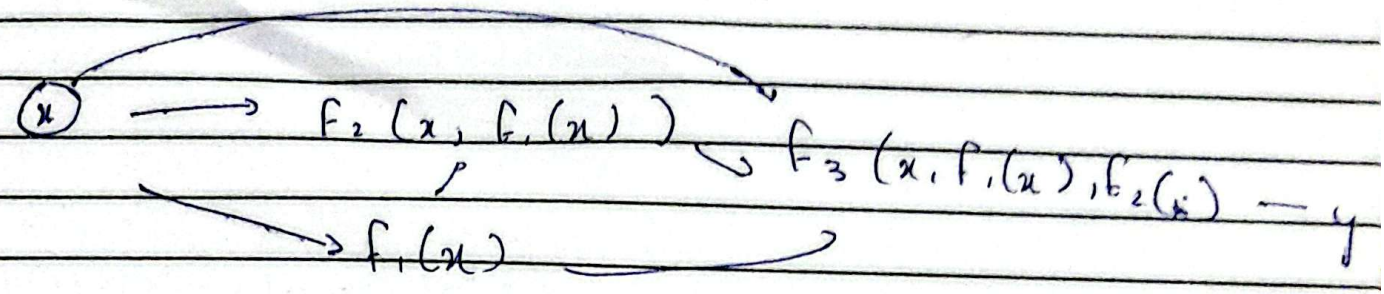
$y = x + f_1(x, w_1) + f_2(x + f_1(x, w_1)) +$

$f_3(x + f_1(x, w_1) + f_2(x + f_1(x, w_1)))$



$x \rightarrow y$

Date M T W T F S



Multiple paths

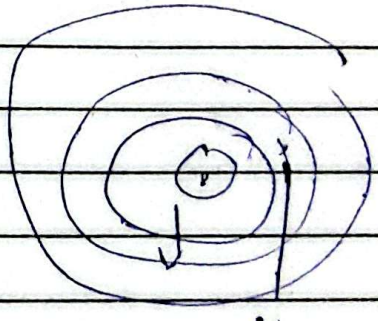
$x \to y$

$x \to f_2 \to f_3 \to y$

$x \to f_1 \to f_3 \to y$

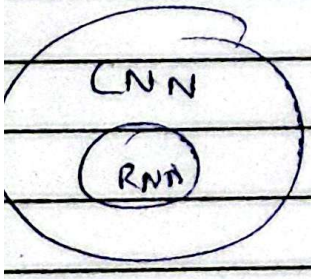
$x \to f_1 \to f_2 \to f_3 \to y$

Nested neural networks



layers in neural network

If some problematic path is going in upward layer good part is also going



You can bypass problematic part

Res NN + Relu

Relu $f(x) = \max(0, x_i)$

↓

solution to vanishing gradient

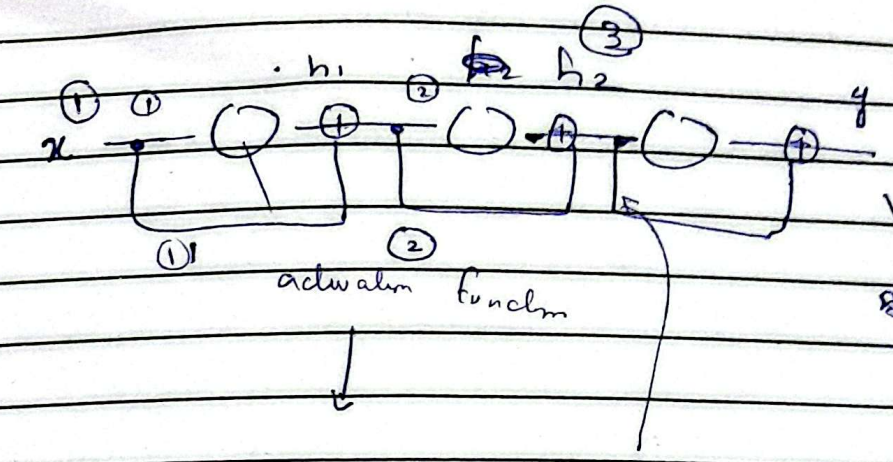
not a perfect solution

Whole solution

Res NN + Relu + Batch Normalization

Date

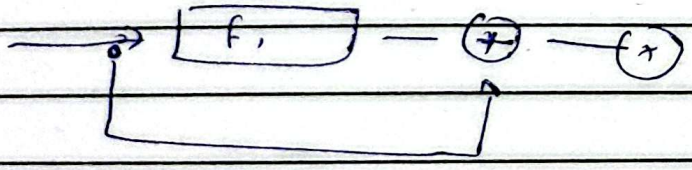
M T W T F S



$$h_1 = relu\left(f_1(x_1, w_1)\right) + x_1$$

but we do the solution only in the end

We want to do normalization on the every step



$$h_1 = \left(f_1(x) + x\right) \times \frac{1}{\sqrt{2}}$$

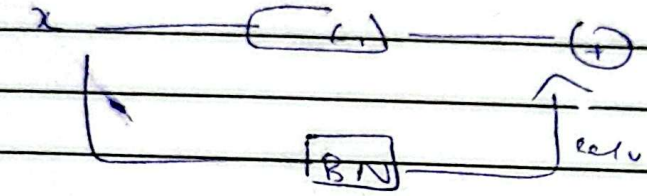
normalized

B

Date M T W T F S

Batch Normalization

$x = [5, 8, 10, 12]$



$\mu = 8.75$

$\sigma^2 = \frac{1}{4}[(5-8.75)^2 + (8-8.75)^2 + (10-8.75)^2 + (12-8.75)^2]$

$\sigma^2 = 6.68$

$\sigma = \sqrt{6.68}$

$\sigma = \frac{1}{N-1}$

zero problem $\rightarrow \sigma = \sqrt{6.68 + 0.00001}$

$x' = \left[\frac{5-\mu}{\sigma}\right.$

$x' = \left[\frac{5-8.75}{2.58}, \frac{8-8.75}{2.58}, \frac{10-8.75}{2.58}, \frac{12-8.75}{2.58}\right]$ 2.58

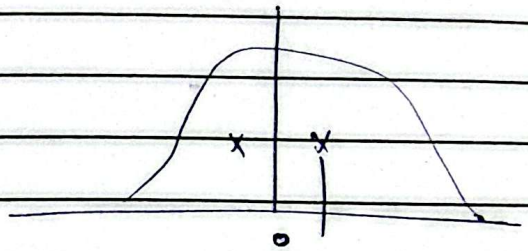
Date

M T W T F S

Castelli

[- - + +]

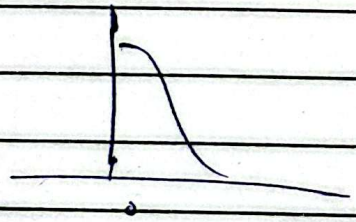
[-1.45 -0.29 +0.48 +1.25]



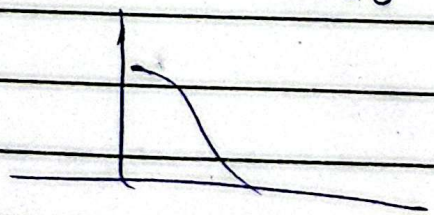
normalization

Symmetry

so we dont need other half



even with redu



positive values

Date M T W T F S

$$x' = \frac{x - \mu}{\sigma}$$

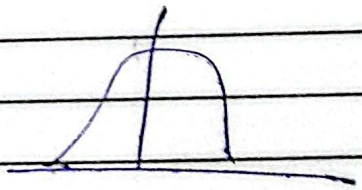
Batch normalization

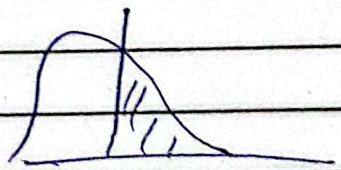
$$x'' = \gamma x' + \beta$$

Let $\gamma = 1$

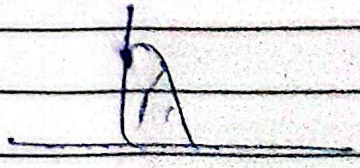
$\beta = 0$

[-1.45 -0.29 +0.48 1.25]





reduce



a small portion

Date

M T W T F S S

We update the value of γ and β

$$\frac{\partial \epsilon}{\partial \gamma} \qquad \frac{\partial \epsilon}{\partial \beta}$$

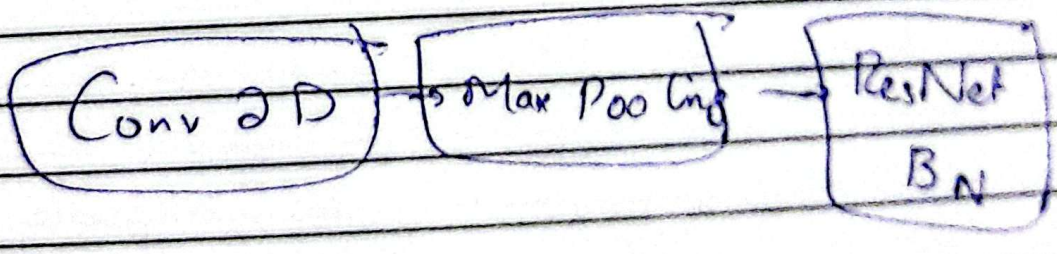
Sola

Vanishing Gradient

as near 0 values after γ, β

push to the side

CNN



Date (M T W T F S)

[Attention]

Regular → timeseries

text/ freeform

↳ cat ate a mouse } from computer pov it made
mouse ate a cat } sense

cat → mouse

ambiguity is always there

position → 10000
welcome

cat ate
ate 01005
a 00104
mouse 00011

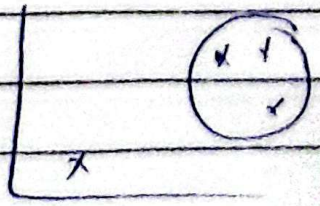
cat ate
ate a
a mouse

→ Vectorized Form → learning algo

1 gram 2 gram

position information is important

cosine → similarity



similar words ← databricks Corpus

Date

M T W T F S

Castelli

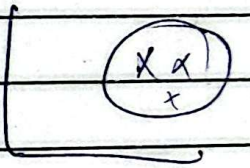
Attention

~~where~~ cat attention is toward whose - ate
~ mouse

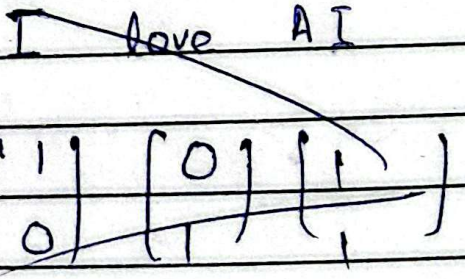
nlp techniques

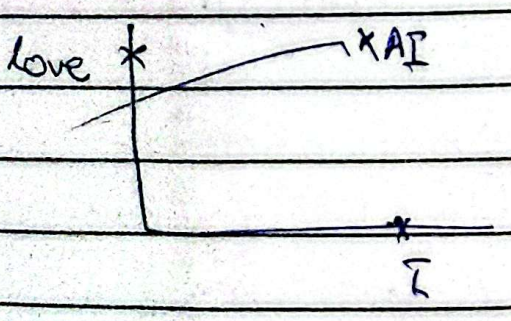
d will never learn

Attention



it learns





Date _____ M T W T F S

Castelli

vectors must be unique and less in size

I love AI

$[1\ 0]$ $[0\ 1]$ $[1\ 1]$

$\sigma(q_i k_i) + \rightarrow v_{ik}$

attention (Q, k, V

Query, key, value

love → AI

Weight

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

for I

$q_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

$k_1 = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

$v_1 = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

for love

$q_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

$k_2 = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

$v_2 = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

for AI

$q_3 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

$k_3 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$

$v_3 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

q	k1	k2	k3
q1	1	1	2
q2	0	1	1
q3	1	2	3

$q_i k_i$ dot product

softmax

$\frac{e^1}{e^1 + e^1 + e^2} = \frac{2.21}{10.82}$

e^1

$\begin{bmatrix} 0.21 & 0.21 & 0.58 \\ 0.16 & 0.42 & 0.42 \\ 0.09 & 0.24 & 0.66 \end{bmatrix}$

Output for I

$0.21 \begin{bmatrix} 1 \\ 1 \end{bmatrix} + 0.21 \begin{bmatrix} 0 \\ 1 \end{bmatrix} + 0.58 \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

$\rightarrow \begin{bmatrix} 0.21 + 0 + 0.58 \\ 0.21 + 0.21 + 1.16 \end{bmatrix}$

$\rightarrow \begin{bmatrix} 0.79 \\ 1.58 \end{bmatrix}$

Date M T W T F S

Castelli

Output for love

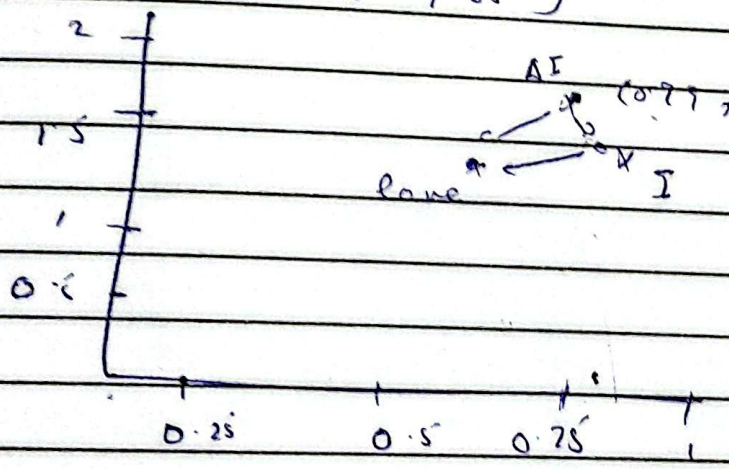
$$0.16\begin{bmatrix}1\\1\end{bmatrix} + 0.16\begin{bmatrix}0\\1\end{bmatrix} + 0.16\begin{bmatrix}1\\2\end{bmatrix}$$

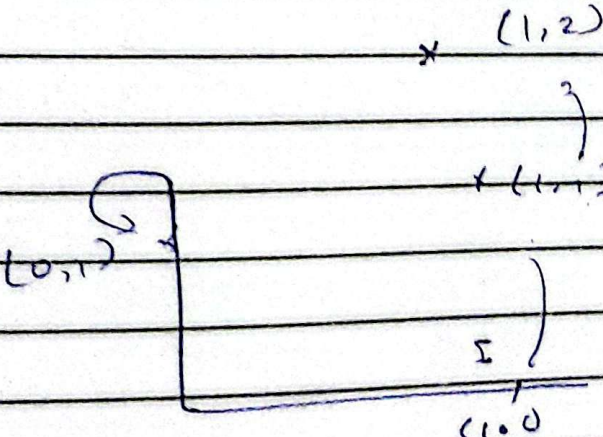
$$\begin{bmatrix}0.48\\1.42\end{bmatrix}$$

Output of AI

$$0.07\begin{bmatrix}1\\1\end{bmatrix} + 0.24\begin{bmatrix}0\\1\end{bmatrix} + 0.66\begin{bmatrix}1\\2\end{bmatrix}$$

$$= \begin{bmatrix}0.75\\1.65\end{bmatrix}$$





Weights

↓

Random values

Date M T W T F S

Transformers → Attention (Multi Head Attention)

x ← input $n \times 1 \in \mathbb{R}^{n \times 1}$

x ← ?

Matrix s.t. $\in \mathbb{R}^{n \times d}$

i	love	ai
1	0	0
0	1	0
0	0	0

↓ d

W_Q W_K W_V

$Q = x^T W_Q \in \mathbb{R}^{n \times d}$

$K = x^T W_K \in \mathbb{R}^{n \times d}$

$V = x^T W_V \in \mathbb{R}^{n \times D}$

$$\frac{\sigma(QK^T)}{\sqrt{d}} \in \mathbb{R}^{n \times n}$$

(A)

Date _____ M T W T F S

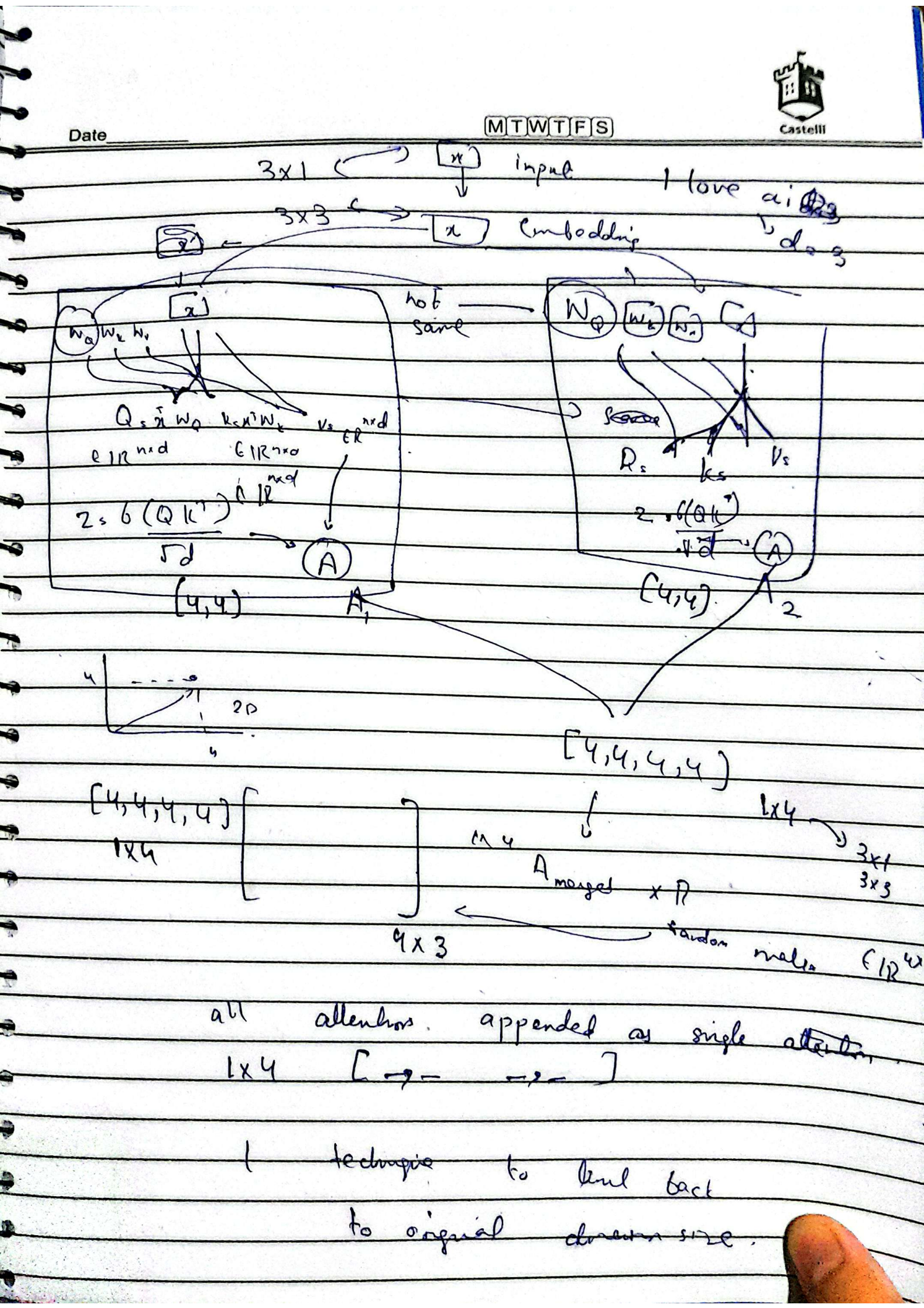
* Ambiguities

I saw the elephant with glasses.

Attention → single perspective drawback

multiple perspective

Date M T W T F S


3x1
x input
3x3
x Embedding
I love ai
dog
not same
Q = x W_Q K = x W_K V = x W_V
Q ∈ R^{n×d} K ∈ R^{n×d} V ∈ R^{n×d}
Z = softmax(QK^T / √d) V
A
[4,4]
A1
A2
[4,4,4,4]
1x4
4x3
A_merged x R
random matrix ∈ R^{4}
3x1
3x3
all attentions appended as single attention
1x4
technique to bring back to original dimension.

Positional Encoding

1) cat ate the mouse

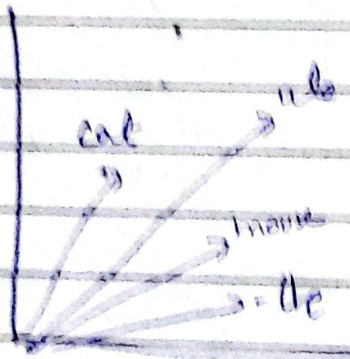
2) mouse ate the cat

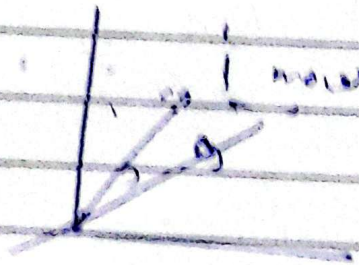
1 0 0 0
0 1 0 0
0 0 1 0
0 0 0 1

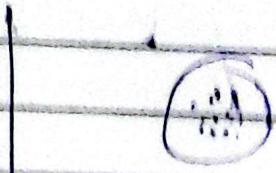
Same

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

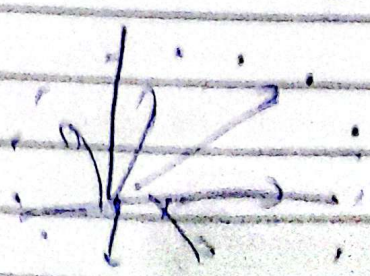
We can't relate our position







Position relation

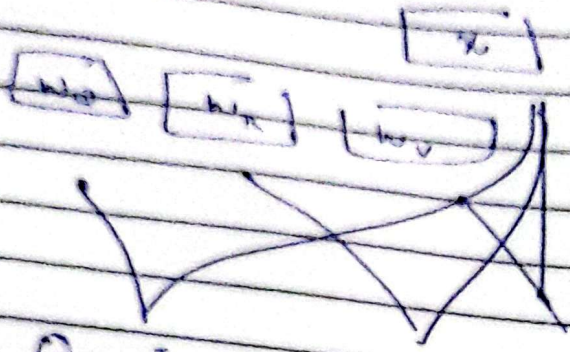


Find an angle

Date

M T W T F S

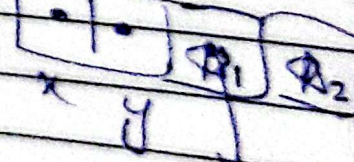
Castelli

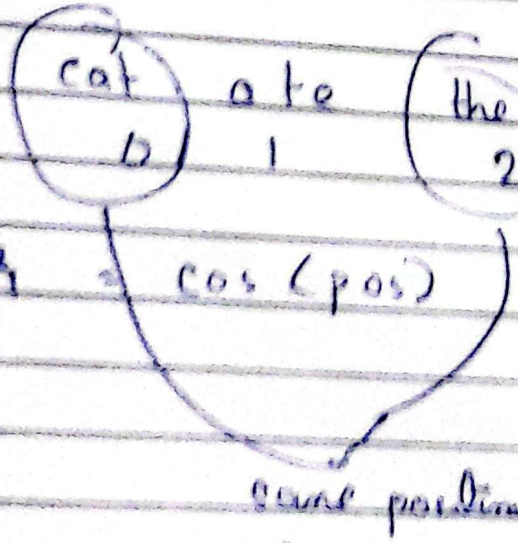


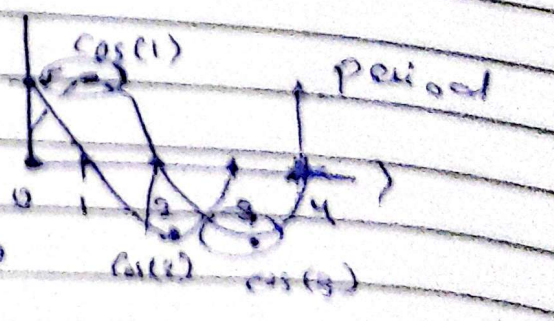
$Q = x^T W_Q$ $\in \mathbb{R}^{n \times d}$ $\quad K = x^T W_K$ $\in \mathbb{R}^{n \times d}$ $\quad V = \in \mathbb{R}^{n \times d}$

$Z = \sigma\left(\frac{QK^T}{\sqrt{d}}\right)$ $\in \mathbb{R}^{n \times n}$

$\Rightarrow$







$x_2 = \sin(pos)$

Date______ M T W T F S

	cat	ate	the	mouse
pos	0	1	2	3

cos(0) cos(1)
sin(0) sin(1)

cat
| x | y | cos(0) | sin(0) |

ate
| x | y | cos(1) | sin(1) |

the | x | y | cos(2) | sin(2) |

mouse

| x | y | cos(3) | sin(3) |

cos(pos) → cos(pos) → cos(pos)
cos(pos/2)
sin(pos) → sin(pos)
sin(pos 2)

(n) (n×1)

↓

sin(pos 1) cos(pos 1)
⋮ ⋮
sin(pos i) cos(pos(i))

(x_e) (n×d)

$i = d/2$

d = 4

$i = 4/2 = 2$

i = 0, 1

d = 8

$i = 8/2 = 4$

i = 0, 1, 2, 3

Subject:

Sr. No. | Date

d=4 $d/2 \rightarrow 2$ (0,1) Topics

		i=0		i=1	
x	pos	cos	sin	cos	sin
cat	0	PE(0,0)	P(0,1)	PE(0,2)	P(0,3)
ate	1	PE(1,0)	P(1,1)	PE(1,2)	P(1,3)
He	2	PE(2,0)	PE(2,1)	PE(2,2)	P(2,3)
mouse	3	P E (3,0)	P(3,1)	PE(3,2)	PE(3,3)

Unique position

cos → PE(Pos, 2i) $\cos\left(\frac{pos}{10{,}000^{2i/d}}\right)$

sin → PE(Pos, 2i+1) $\sin\left(\frac{pos}{10{,}000^{2i/d}}\right)$

d = 4

$\frac{2i}{d}$ $\frac{2}{4}i$ $\frac{1}{2}i$

$$\begin{bmatrix} \cos\left(\frac{0}{1}\right) & \sin\left(\frac{0}{1}\right) & \cos\left(\frac{0}{100}\right) & \sin\left(\frac{0}{100}\right) \\ \cos\left(\frac{1}{1}\right) & \sin\left(\frac{1}{1}\right) & \cos\left(\frac{1}{100}\right) & \sin\left(\frac{1}{100}\right) \\ \cos\left(\frac{2}{1}\right) & \sin\left(\frac{2}{1}\right) & \cos\left(\frac{2}{100}\right) & \sin\left(\frac{2}{100}\right) \\ \cos\left(\frac{3}{1}\right) & \sin\left(\frac{3}{1}\right) & \cos\left(\frac{3}{100}\right) & \sin\left(\frac{3}{100}\right) \end{bmatrix}$$

It solve ambiguities related to position in support of attention

$1000^{0/2} = 1$

$1000^{1/2} = 100$

for pseudo... depl. calm increase
cycles

safe figure 10 multiple (log scale)
decrease

$10{,}000^{\wedge} i$

Transformer Architecture
- Attention (M·H)
- Positional Encoding

→ RNN

↳ ANN / CNN

Natural Sequence (text)

↳ W_n W_{n-1} W_{n-2} ... W_0

Sequence 2 sequence architecture

1 Dare ai

W_2 W_1 W_0

W_0 W_1 W_2 (W_n)

Encoder ↔ Decoder

W_n → En → $W'_{n \cdot m}$ → Dec → W_n

(↑ Cmm → Dec)

Date M T W T F S

(2017) Encoder ⟷ Decoder

LSTM

① small text length ≤ threshold

↳ bad performance on large language

↓

(Generative text)

Transformers.

Encoder ⟷ Decoder

Attention

first

$x^i \in \mathbb{R}^{n \times 1} \rightarrow x \in \mathbb{R}^{512}$ → fixed

$n = 3$

$d = 512$

[1.1.1 0000 0000] embedding

$n = 5$

[1.1.1.1.1 0 0000]

Date M T W T F S

embeddings

I love AI

I [. . . . . 512]

love

AI

Skip gram . . . . . one hot

↳ 512 size vector

x' $\in \mathbb{R}^{n \times 1}$ → x $\in \mathbb{R}^{512}$ → PE $\in \mathbb{R}^{512}$ → (+)

! means $x + PE$

x' → x → PE → (+) → P, Q, V

x' $\in \mathbb{R}^{n \times 1}$ $\in \mathbb{R}^{512}$ $\in \mathbb{R}^{512}$

x'' { Embedding + PE

x'' — Multi Head Attention

$W_p \times x'' = P$

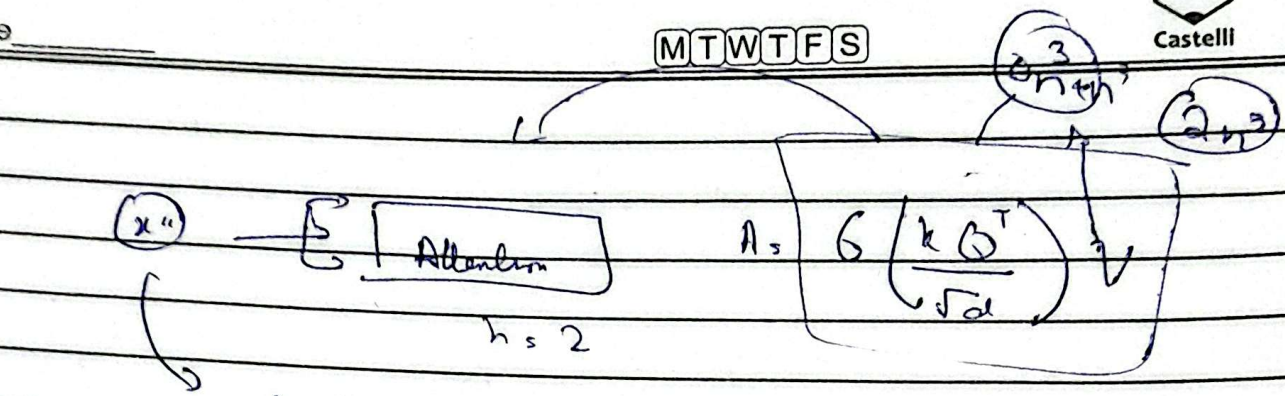
h : 8

$A = \left[\sigma \left(\frac{KQ^T}{\sqrt{d}} \right) V \right]$

Date

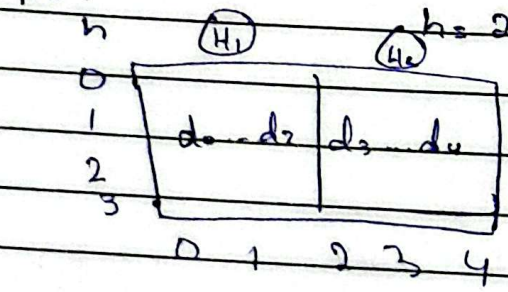
M T W T F S

Castelli

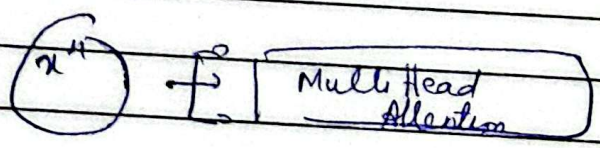


short sentences

ambiguities & sentence length



(1)



$O(p n^3)$

no of heads

$O(h(n^3 + n^3)) \Rightarrow O(n^3)$

$O(n^4)$

$A = \sigma\left(\frac{kQ^T}{\sqrt{d}}\right) V$

extreme bugs if $p > n$

$p \leq n$

$p = 2, 3, 4, 8$

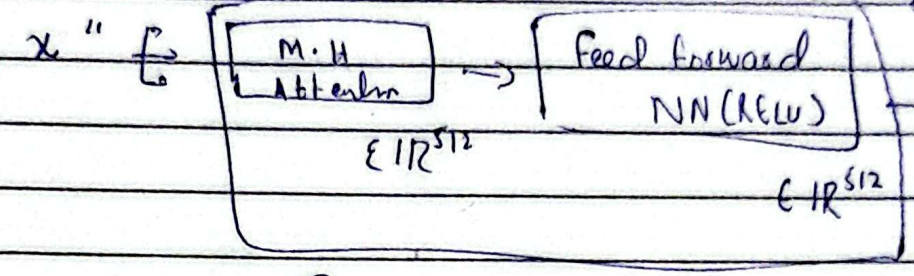
bias, gender, racial

↑

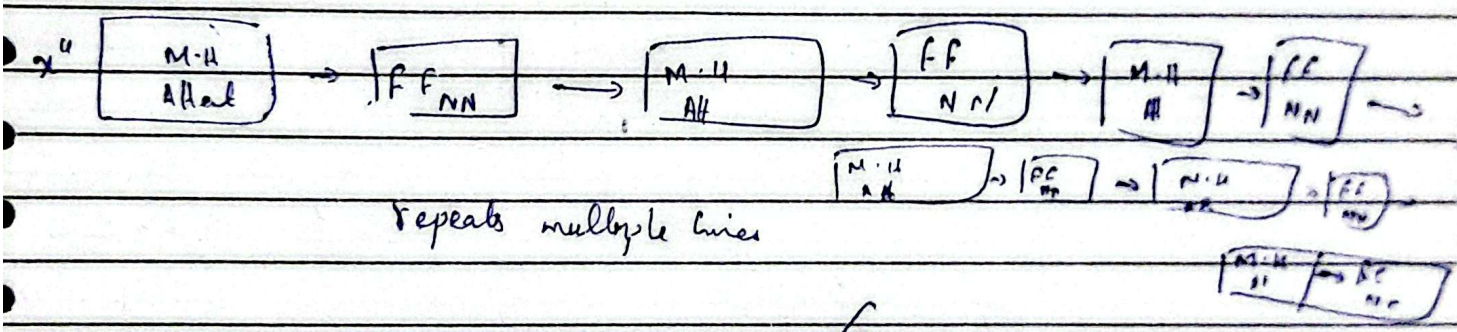
h should be increased

Date M T W T F S

$[W_Q, W_K, W_V]^T$ 512×512



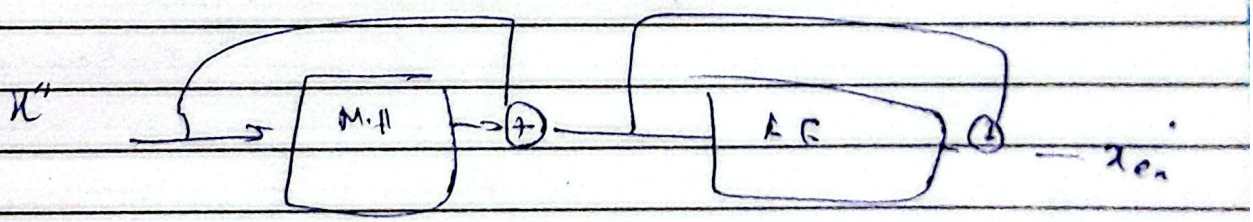
encoder

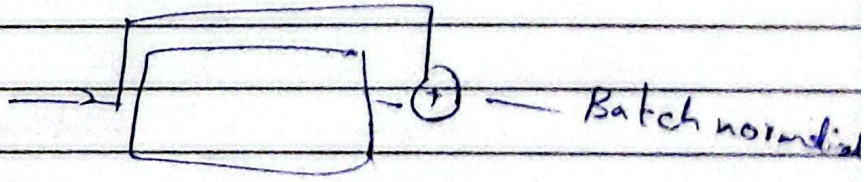


Repeats multiple times

6

Vanishing gadient problem if depth more





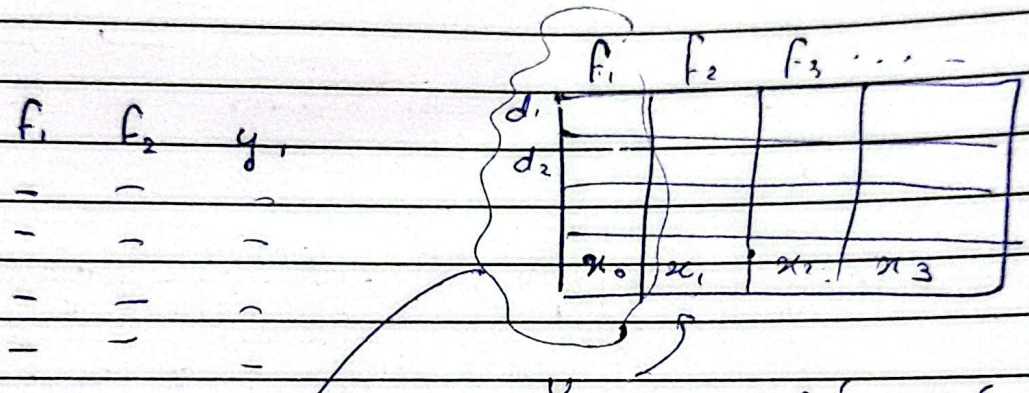
Date

M T W T F S

Castelli

Ex Here we have

layer Normalization



Batch Normalization for each row

layer normalization for each column.

1	2	3
4	5	6

Features me difference fir ek klin kr rahe

batch + layer both

Date [M][T][W][T][F][S]

Encoder

Learning

[MHA → FF] → ⊗

API

⊗ → P Encoding → Masked Multi Head Attention → Feed forward → output

↓

Decoder

[M·M H·A] → [FF] [M·M A] → FF

I love ai and fast → ?

→ → → → →

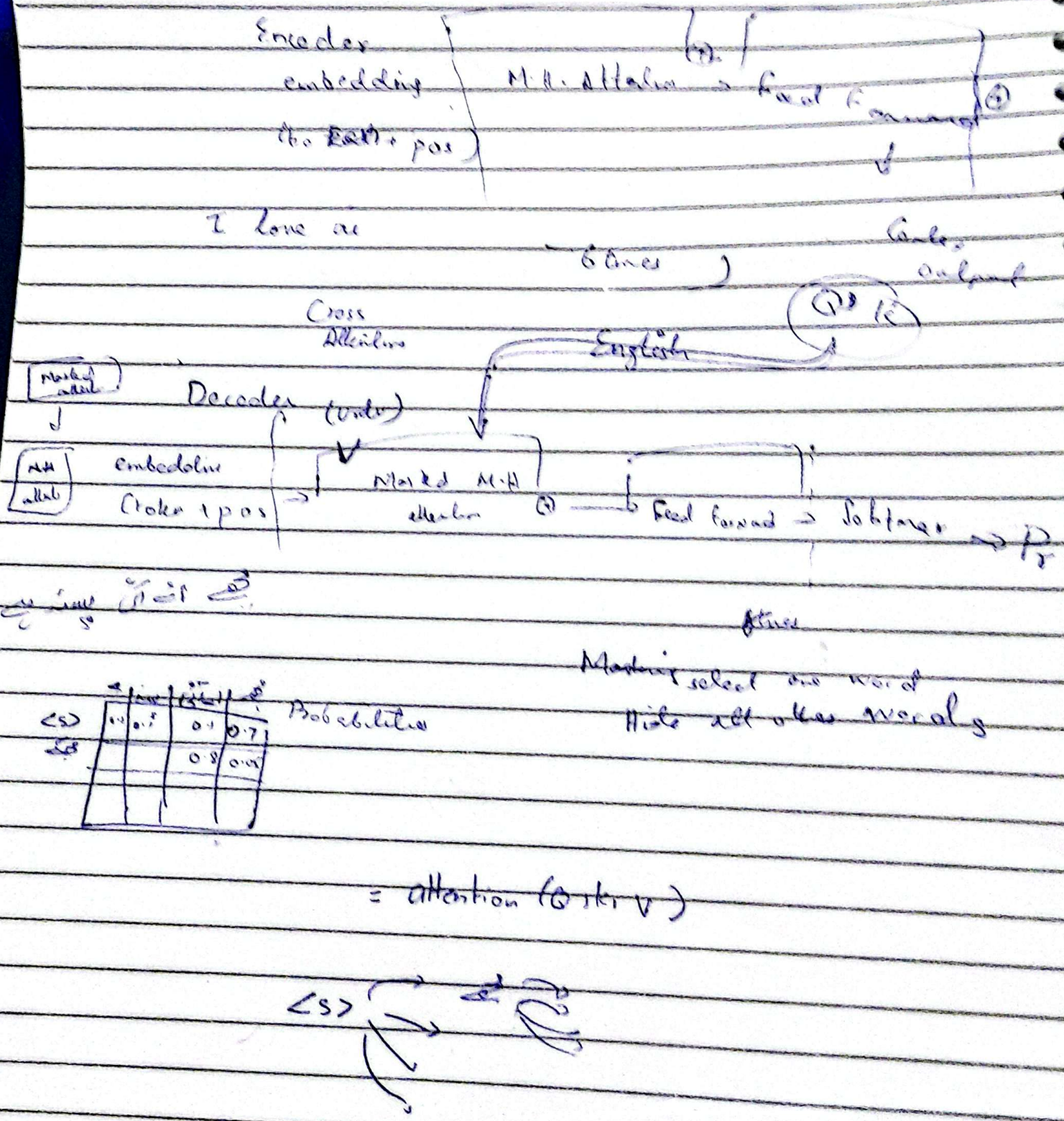
$x'' = [i \text{ love } \text{fast}] + [0 \; 0 \; -10^9]$

mask

[_ , _ , negative / large / value]

BERT / ROBERTA / GPT

Date [M][T][W][T][F][S]


Encoder
embedding
+ pos
M.H. Attention
Feed Forward
6 times
Cross Attention
Decoder (Outputs)
Marked M.H Attention
Feed forward
Softmax
Probabilities
Attention (Q K V)

M T W T F S

Generation

after trained model

just do encoding