CHATGPT -> GPT

GPT -> Generative Pretrained transformer

↓

80%

(AI PART)

Tokenization

Dataset -> {NLP is good, machine learning is good ..... .}

-> {a, b, c ... A, B ... 7 .. 1, 2, $}

0 1 2 65

[65]

{39 [11 15 63 8 18
[63 6 14 14] 3 63}

Batch size -> (BS) = 2

Block size -> BL = 4

Device = cuda | cpu | mps

Embding size E = 4

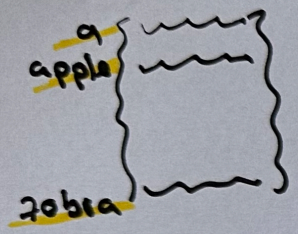
X -> [11 15 63 8
63 6 14 14]

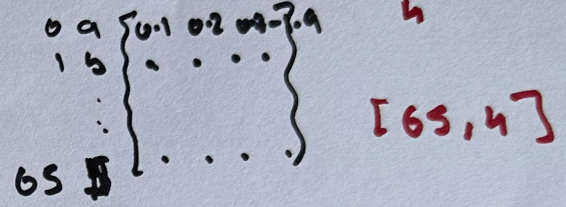
[BS, BL] = [2, 4]

Y -> [15 63 8 18
6 14 14 3]

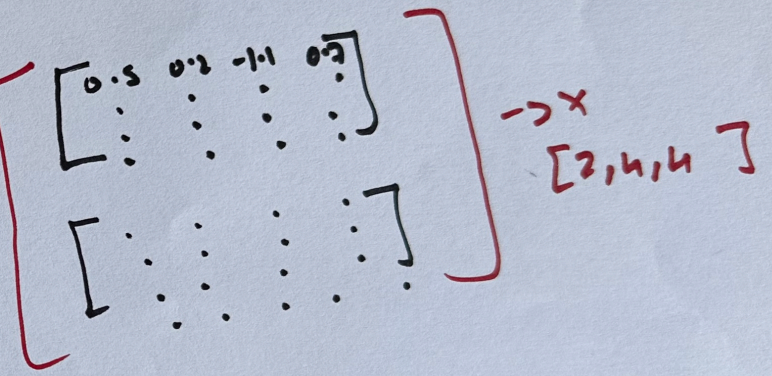
[2, 4]

Embedding and Positional Encoding





$X \rightarrow \begin{bmatrix} 11 & 65 & 63 & 8 \\ 63 & 6 & 14 & 14 \end{bmatrix}$ [2,4]



$X_n = X + p(x)$

[ ] [ ]

[2,4,4] [2,4]

-> [ ]

[2,4,4]

Masked Multihead Attention

$\frac{4}{2} = 2$

x_n · [ ]
[2,4,4]

head 1

$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$

Q $x_n W_1^T \rightarrow [2,4]$
[2,4,4][4,2] -> [ ]
[2,4,2]

K $x_n W_2^T$
[ ] [ ]
[2,4,4] [4,2] -> [ ]
[2,4,2]

V $x_n W_3^T$ [2,4,4][4,2] -> [ ]
[2,4,2]

QK^T
[2,4,2] [2,2,4] -> [
softmax [mask [[[ ] / $\sqrt{4}$]]]
[2,4,2]

[: : :]
[3,8]

[ ] [ ]
[2,4,2] [2,4,2]
[ ]
[2,4,2] -> 01

head 2

Q $x_n W_4^T$
[2,4,4][4,2] -> [2,4,2]

K $x_n W_5^T$ [ ] [2,4,2]

V $x_n W_6^T$
[ ] -> 02
[2,4,2]
[ ] -> x_0
[2,4,4]

$x_0 \cdot W_3^T$ [4,4] -> [ ] -> x_0
[2,4,4] [2,4,4]

x_f -> tanh weighted sum ($\alpha x_n + \alpha x_0$) -> [ ]
[2,4,4]

Feedforward -> Add -> Norm

x_s [ ]
[2, H/2]

fully connected 1

$x_s W_0^T$ -> [H/2 x H]
[2, H/2] [H/2, 16]
[ ]
[ ]
[2, H/2, 16] -> x_{s2}

relu

fc 2

$x_{s2} W_1^T$ [H/2 x H]
[2, H/2, 16] [16, H]
[ ]
[2, H/2, H] -> x_{s20}

$x_{s1} + x_{s20}$ -> x_0

LN

Attention(Q, K, V)

$$\text{Softmax}\left(\frac{QK^T}{\sqrt{h}}\right)V$$

$Q \rightarrow x_0 W_{10}^T$ [2,4] → [2,4]

$K \rightarrow x_0 W_{11}^T$ [2,4] → [2,4]

$V \rightarrow x_0 W_{12}^T$ [2,4] → [2,4]

[2,4,2] → [2,4,2] → x_{80}

$Q \rightarrow x_0 W_{13}^T$ [2,4] → [2,4]

$K \rightarrow x_0 W_{14}^T$ [2,4] → [2,4]

$Q \rightarrow x_0 W_{15}^T$ [2,4] → [2,4]

[2,4,4] → [2,4,4] → x_{81}

$x_{01} W_{10}^T$ [4,4] → [4,4,2]

$x_5 \rightarrow \text{LN}_1[x_0 + x_{01}]$

$x_5 W_{11}^T$ [4,4] → [4,4,2]

$x_{52} W_{12}^T$ [4,4] → [4,4,2]

[4,4,2] → [4,4,2] → x_{520}

$x_{52} W_{13}^T$ [4,4] → [4,4,4]

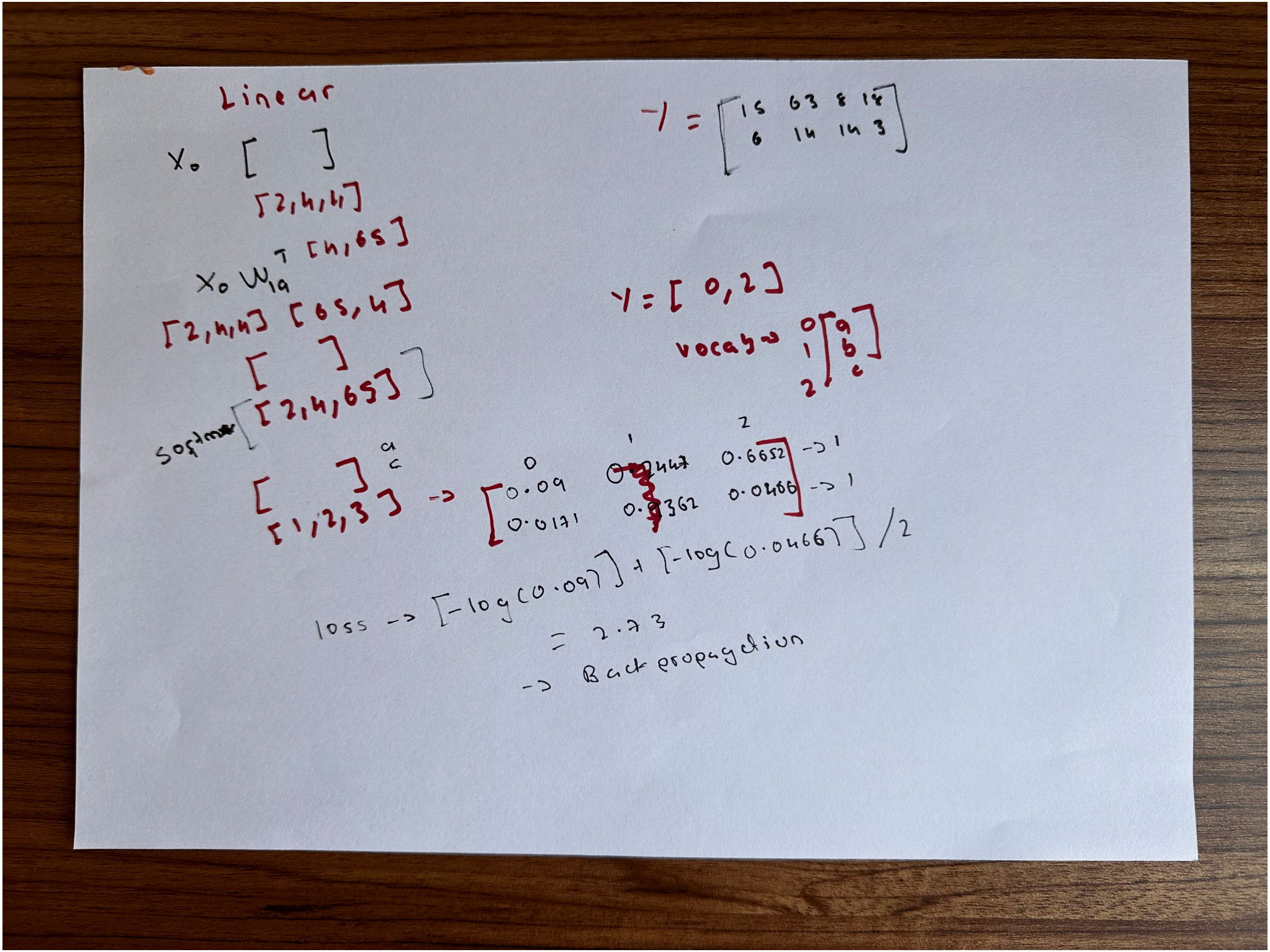
$x_{53} W_{14}^T$ [4,4] → [4,4,4]

$x_{54} W_{15}^T$ [4,4] → [4,4,4]

[4,4,4] → [4,4,4] → x_{53}

$x_0 = \text{LN}_1[x_5 + x_{530}]$

[4,4,2]


Linear
X_0 [ ]
[2,4,4]
W_{1a}^T [4,65]
$X_0 W_{1a}$
[2,4,4] [65,4]
[ ]
[2,4,65]
softmax
[2,4,65]
[ ] a c
[1,2,3]
-> 0 1 2
0.09 0.6652 -> 1
0.0171 0.0466 -> 1
Y = [0,2]
vocab -> 0 a 1 b 2 c
Y = [15 63 8 18
6 14 14 3]
loss -> [-log(0.09)] + [-log(0.0466)] /2
= 2.73
-> Back propagation