

AI by Hand

Volume 1

Basic x 5
Advanced x 20

Prof. Tom Yeh

2024

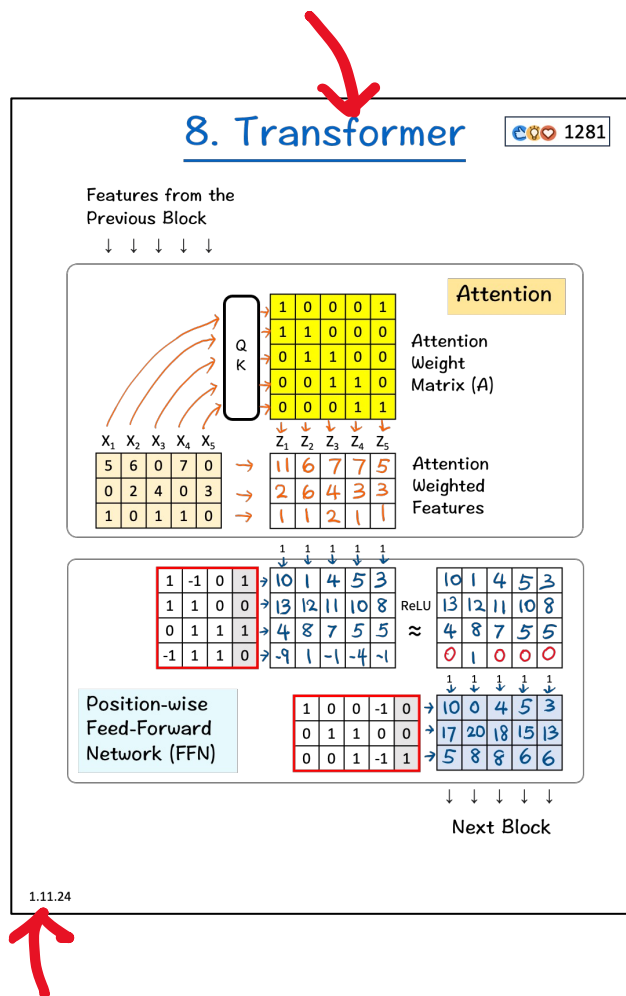
Basic

- I. One Node
- II. Four Nodes
- III. One Hidden Layer
- IV. Three Inputs
- V. Seven Layers

Advanced

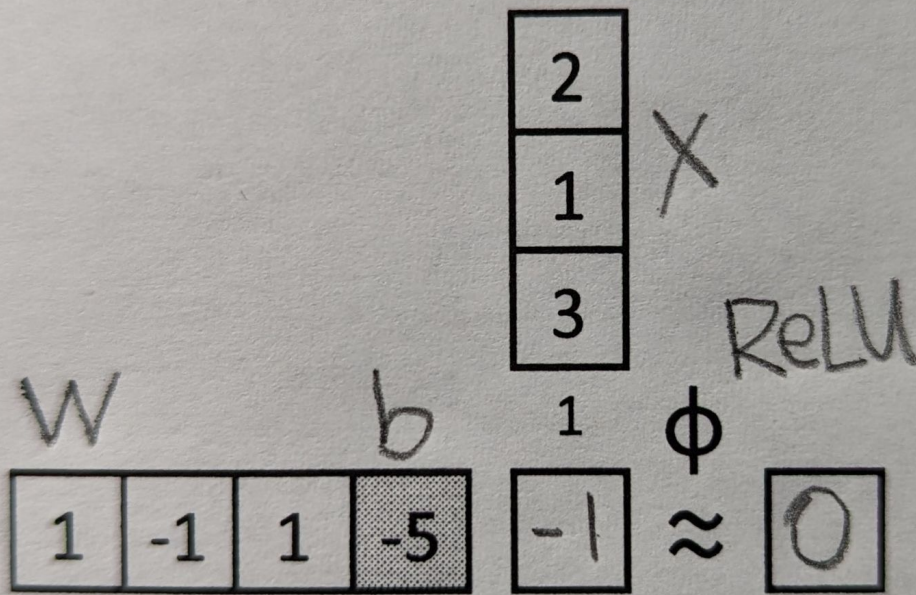
1. Mixture of Experts (MOEs)
2. Recurrent Neural Network (RNN)
3. Mamba
4. Matrix Multiplication
5. LLM Sampling
6. MLP in PyTorch
7. Backpropagation
8. Transformer
9. Batch Normalization
10. Generative Adversarial Network (GAN)
11. Self Attention
12. Dropout
13. Autoencoder
14. Vector Database
15. CLIP
16. Residual Network (ResNet)
17. Graph Convolution Network (GCN)
18. SORA's Diffusion Transformer (DiT)
19. Gemini 1.5's Switch Transformer
20. Reinforcement Learning with Human Feedback (RLHF)

Link to my original LinkedIn post
with animation and explanation



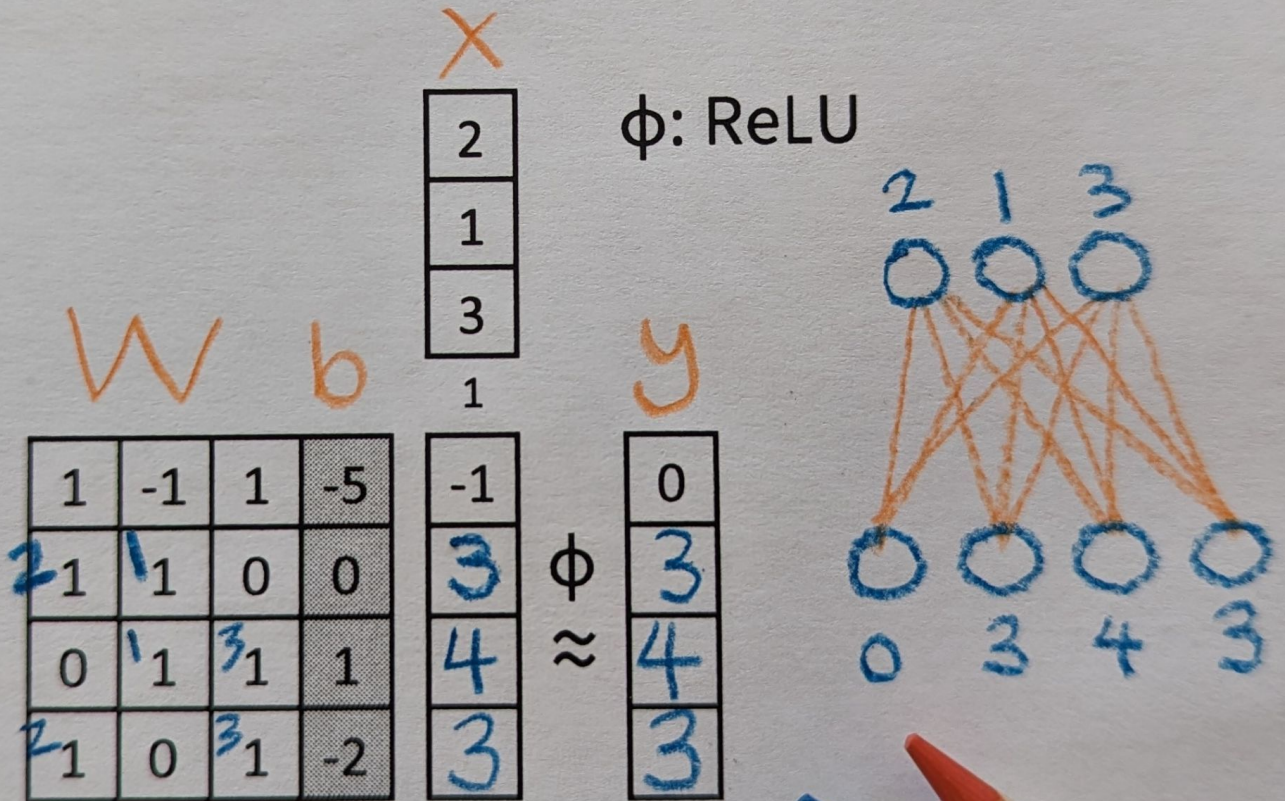
Date originally posted

I. One Node



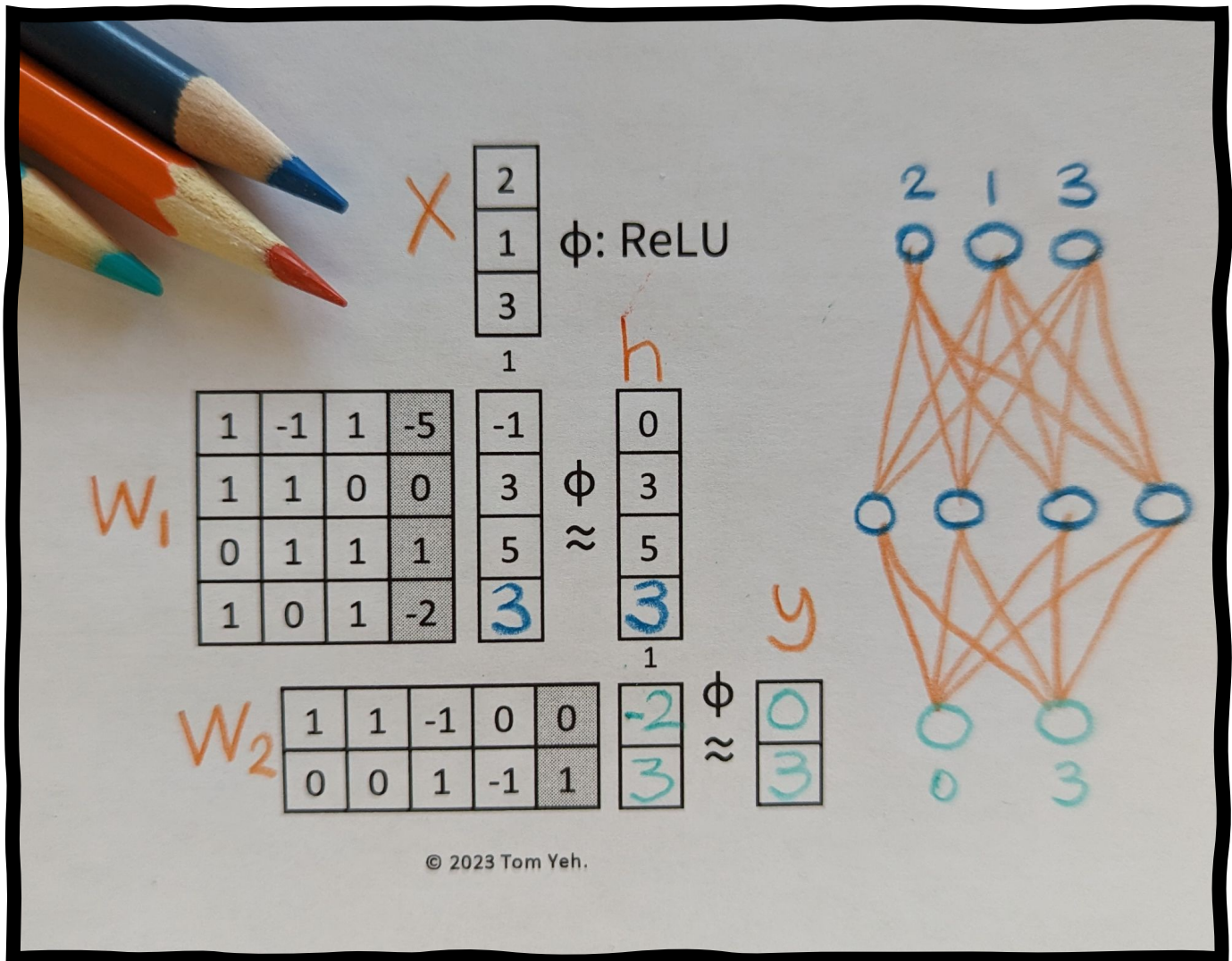
© 2023 Tom Yeh.

II. Four Nodes



© 2023 Tom Yeh.

III. Hidden Layer



IV. Three Inputs

2	1	0
1	1	1
3	0	1
1	1	1

$\phi: \text{ReLU}$

1	-1	1	-5
1	1	0	0
0	1	1	1
1	0	1	-2

-1	-5	-5
3	2	1
5	2	3
3	-1	-1

$\phi \approx$

0	0	0
3	2	1
5	2	3
3	0	0
1	1	1

1	1	-1	0	0
0	0	1	-1	1

-2	0	-2
3	3	4

ϕ

0	0	0
3	3	4

© 2023 Tom Yeh.

V. Seven Layers

W_1

0	0	1	5	3
0	1	0	4	4
1	0	0	3	5
1	1	0	7	9
0	1	1	9	7

W_2

1	1	-1	0	0	6	2
0	0	1	1	-1	1	7

W_3

1	1	7	9
1	-1	5	5 0
1	2	8	16

W_4

1	-1	0	2	9
0	-1	1	3	16

W_5

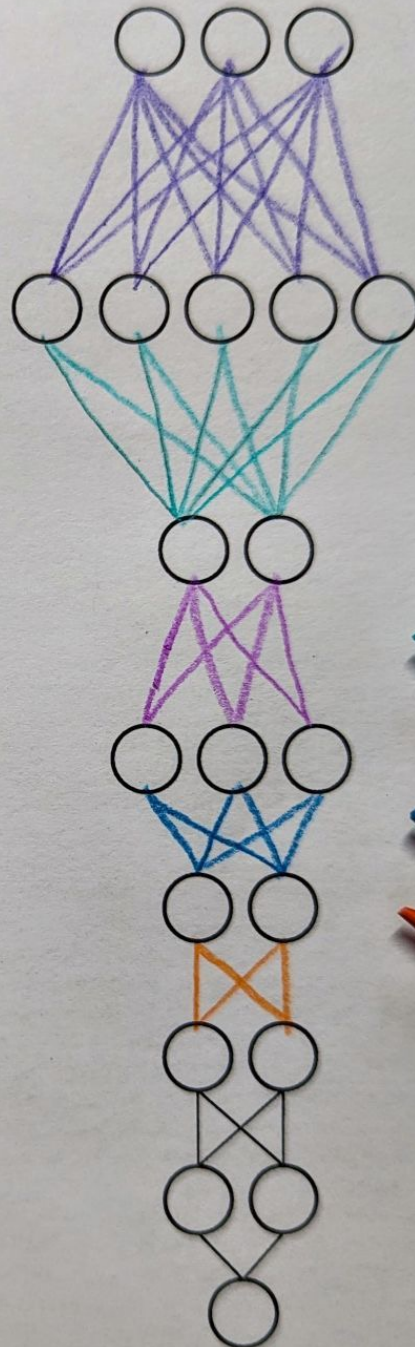
0	1	3	21
1	0	2	9

W_6

1	-1	1	12
1	1	5	30

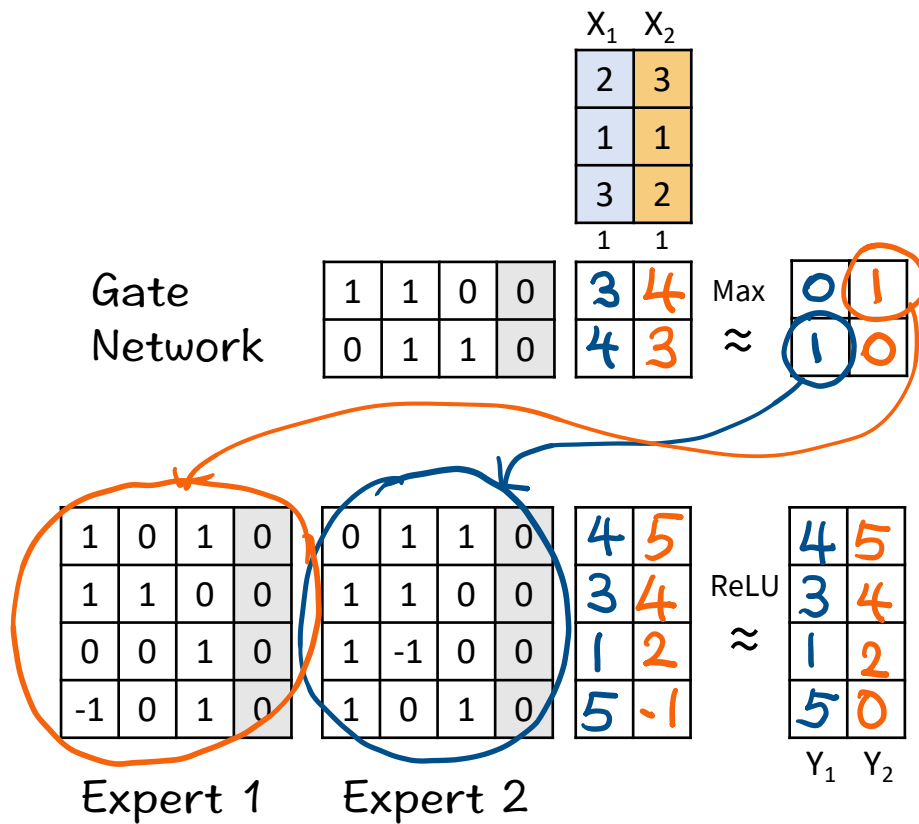
W_7

1	-1	-4	-18
---	----	----	-----



© 2023 Tom Yeh.

1. Mixture of Experts



2. Recurrent Neural Network (RNN)

🧠💡❤️ 406

Input Sequence

X

3	4	5	6
---	---	---	---

Parameters

A

1	-1
1	1

 B

1
2

 C

-1	1
----	---

Activation Function

ϕ : ReLU

Hidden States

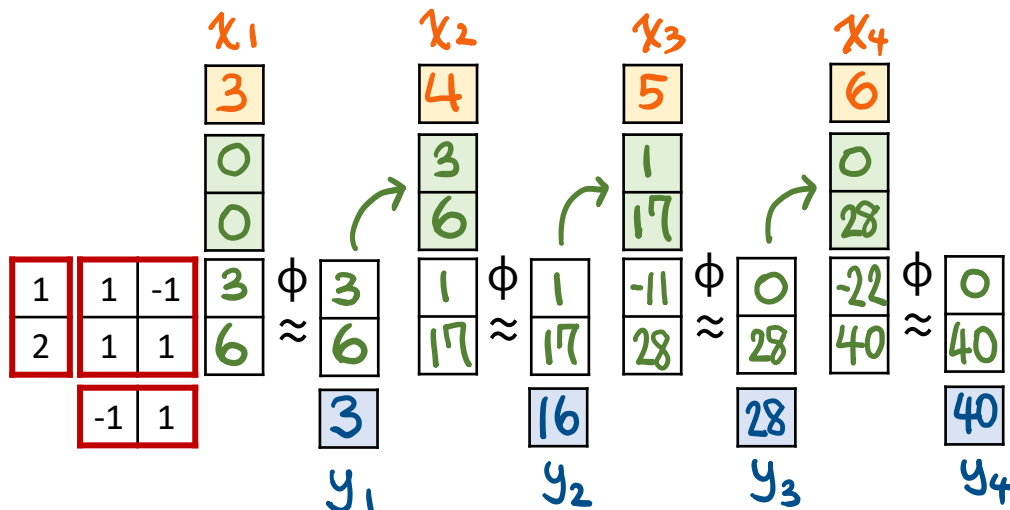
H_0

0
0

Output Sequence

Y

--	--	--	--

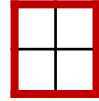


3. Mamba's S6 Model

Input Sequence

3 4 5 6

Parameters

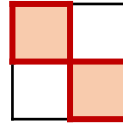


Output Sequence

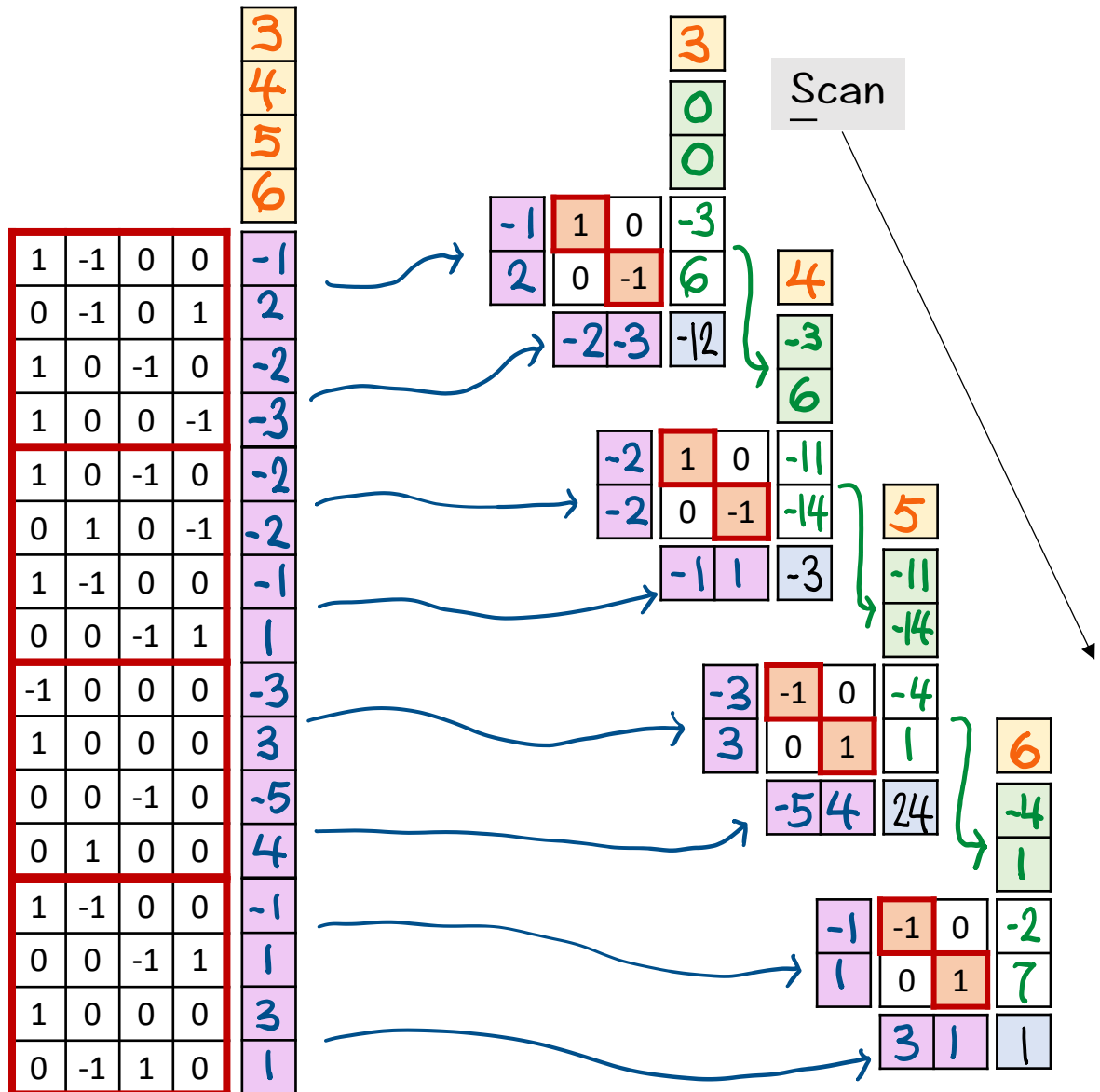
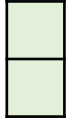
-12 -3 24 1

Selective

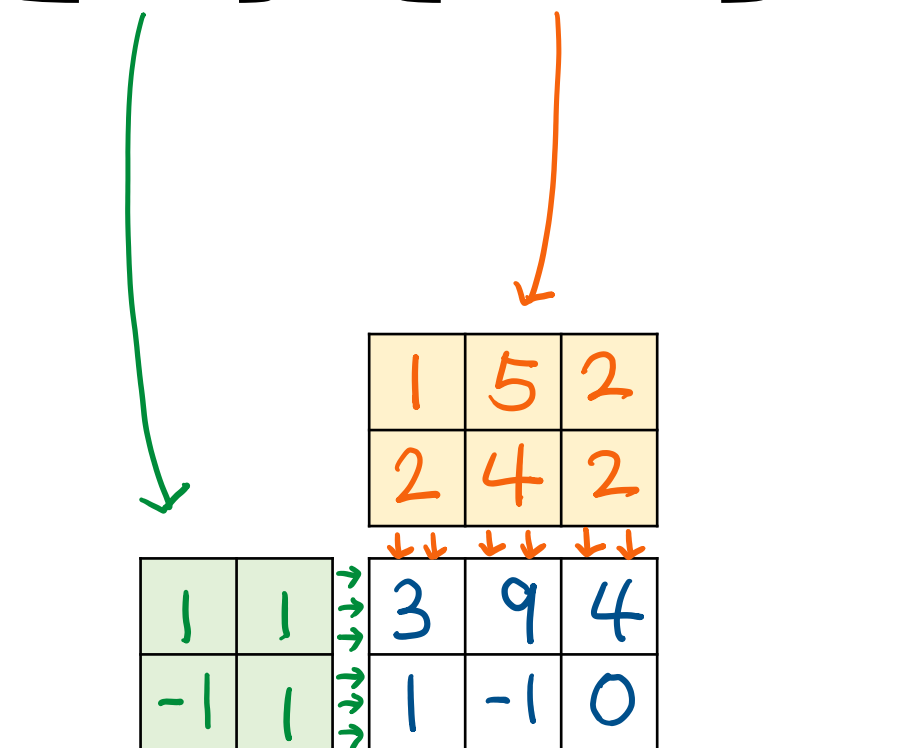
Structured



State-Space



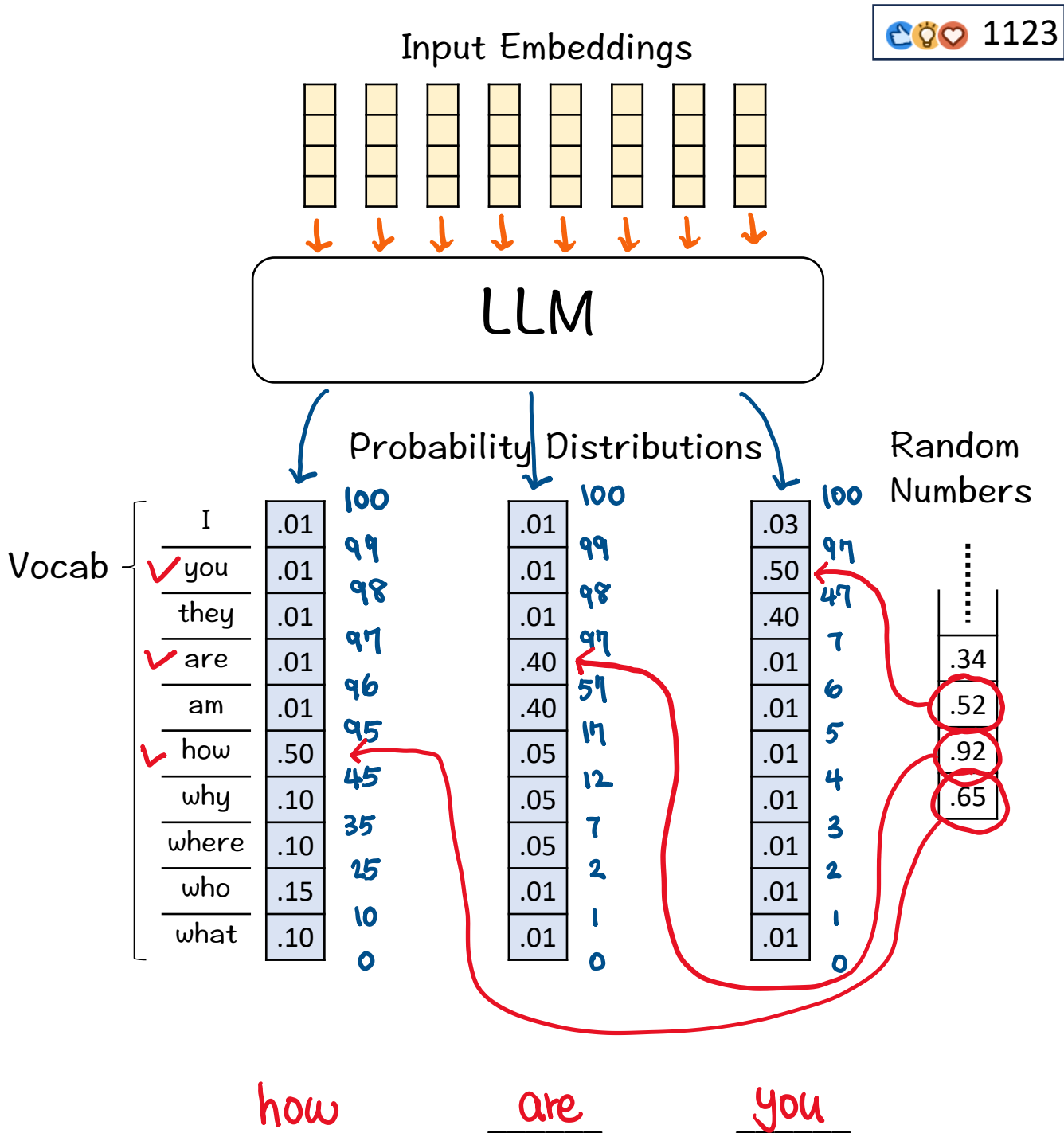
4. Matrix Multiplication

$$\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \times \begin{bmatrix} 1 & 5 & 2 \\ 2 & 4 & 2 \end{bmatrix} = ?$$


The diagram illustrates the dot product method for matrix multiplication. It shows the first matrix (green) and the second matrix (orange) being multiplied. The resulting matrix (blue) is shown below, with arrows indicating the dot products used to calculate each element.

1	1	→	3	9	4
-1	1	→	1	-1	0

5. How does an LLM sample a sentence?

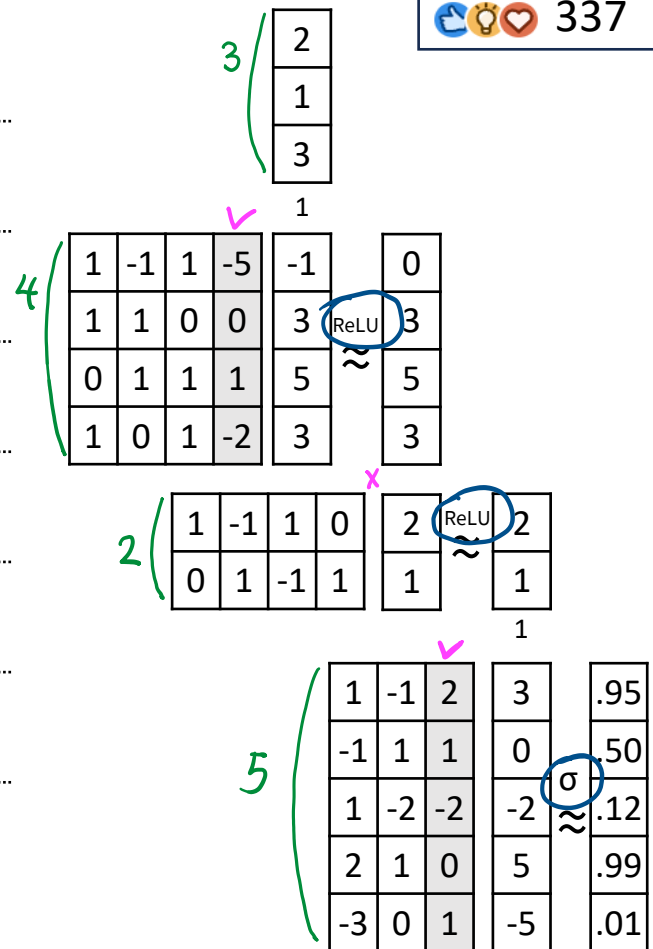


6. Multi Layer Perceptron in pytorch

🔗💡❤️ 337

```

1 mlp_model = nn.Sequential(
2     nn.Linear( 3, 4, bias = T ),
3     nn.ReLU(),
4     nn.Linear( 4, 2, bias = F ),
5     nn.ReLU(),
6     nn.Linear( 2, 5, bias = T ),
7     nn.Sigmoid()
8 )
    
```



Hints:

Linear Layer: { Identity | Linear | Bilinear }

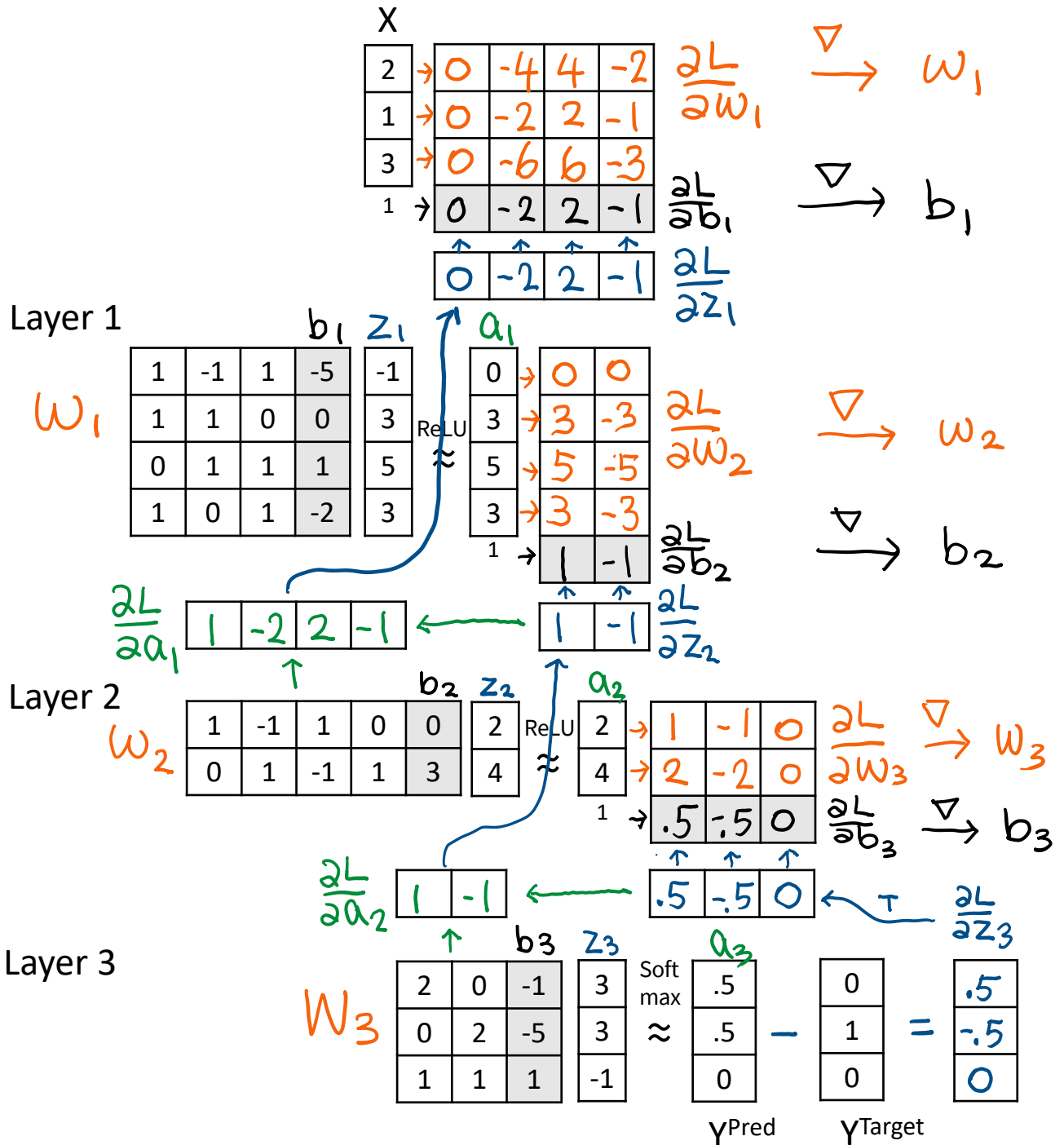
Activation Function: { ReLU | Tanh | Sigmoid }

in_features: { int }

out_features: { int }

bias: { T | F }

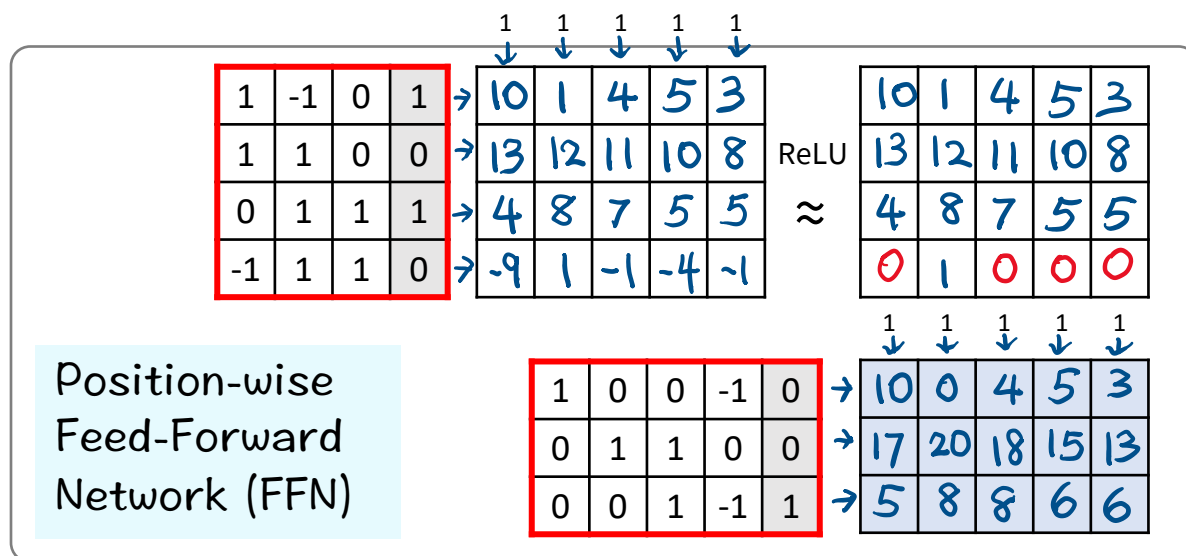
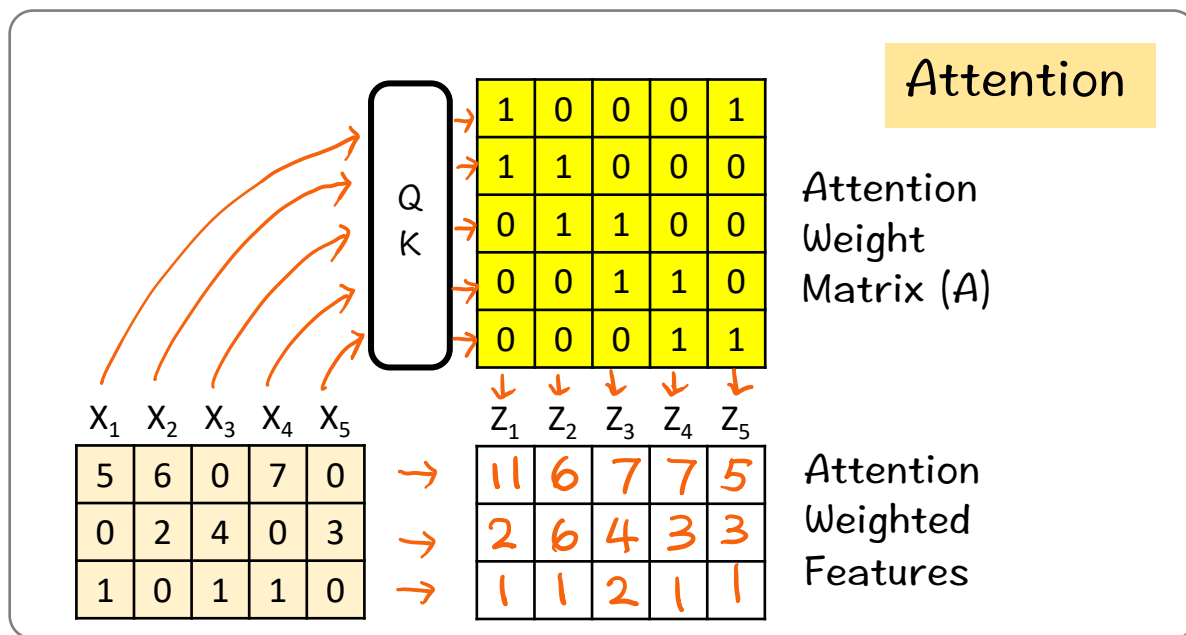
7. Backpropagation



L: Cross-Entropy Loss

8. Transformer

Features from the
Previous Block



Next Block

9. Batch Normalization

Mini-batch: X_1 X_2 X_3 X_4

1	0	3	0
0	3	1	1
2	1	0	2

Linear Layer

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

3	1	3	2
0	2	3	0
-2	5	2	0

ReLU

3	1	3	2
0	2	3	0
0	5	2	0

Batch Statistics

Σ	μ	σ^2	σ
9	2	1	1
5	1	1	1
7	2	4	2

Sum (Σ)
Mean (μ)
Variance (σ^2)
Std Dev (σ)

Normalize

μ	2
-	1
	2

1	-1	1	0
-1	1	2	-1
-2	3	0	-2

σ	1
$\div$	1
	2

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

Scale & Shift

2	0	0	0
0	3	0	0
0	0	-1	1

2	-2	2	0
-3	3	6	-3
2	0	1	2

Next Layer

Trainable
Parameters



10. Generative Adversarial Network (GAN)

1004

Noise: N_1 N_2 N_3 N_4

1	1	0	1
1	0	1	-1

Generator

1	1	0
0	1	2
-1	1	0

[$\approx$ ReLU]

2	1	1	0
3	2	3	1
0	2	1	2

Fake: F_1 F_2 F_3 F_4

-1	1	0	0
1	0	1	0
0	1	1	0
0	0	1	1

[$\approx$ ReLU]

1	1	2	1
2	1	2	0
3	2	4	1
1	1	2	1

Real:

X_1	X_2	X_3	X_4
2	3	3	4
1	1	1	1
2	3	4	3
1	1	1	1

Discriminator

1	0	0	-1	0
0	1	1	0	0
0	0	1	-1	1

[$\approx$ ReLU]

1	1	-1	-1
---	---	----	----

$\rightarrow Z$

0	0	0	0
5	3	6	1
3	2	3	1

[$\approx \sigma$]

1	2	2	3
3	4	5	4
2	3	4	3

[$\approx \sigma$]

Predictions:

.7	.5	.9	.3
----	----	----	----

.7	.9	.9	1
----	----	----	---

Training the Discriminator

Targets:

Y_D

0	0	0	0
---	---	---	---

1	1	1	1
---	---	---	---

Loss Gradients: $\frac{\partial L_D}{\partial Z}$

.7	.5	.9	.3
----	----	----	----

-.3	-.1	-.1	0
-----	-----	-----	---

Training the Generator

Targets:

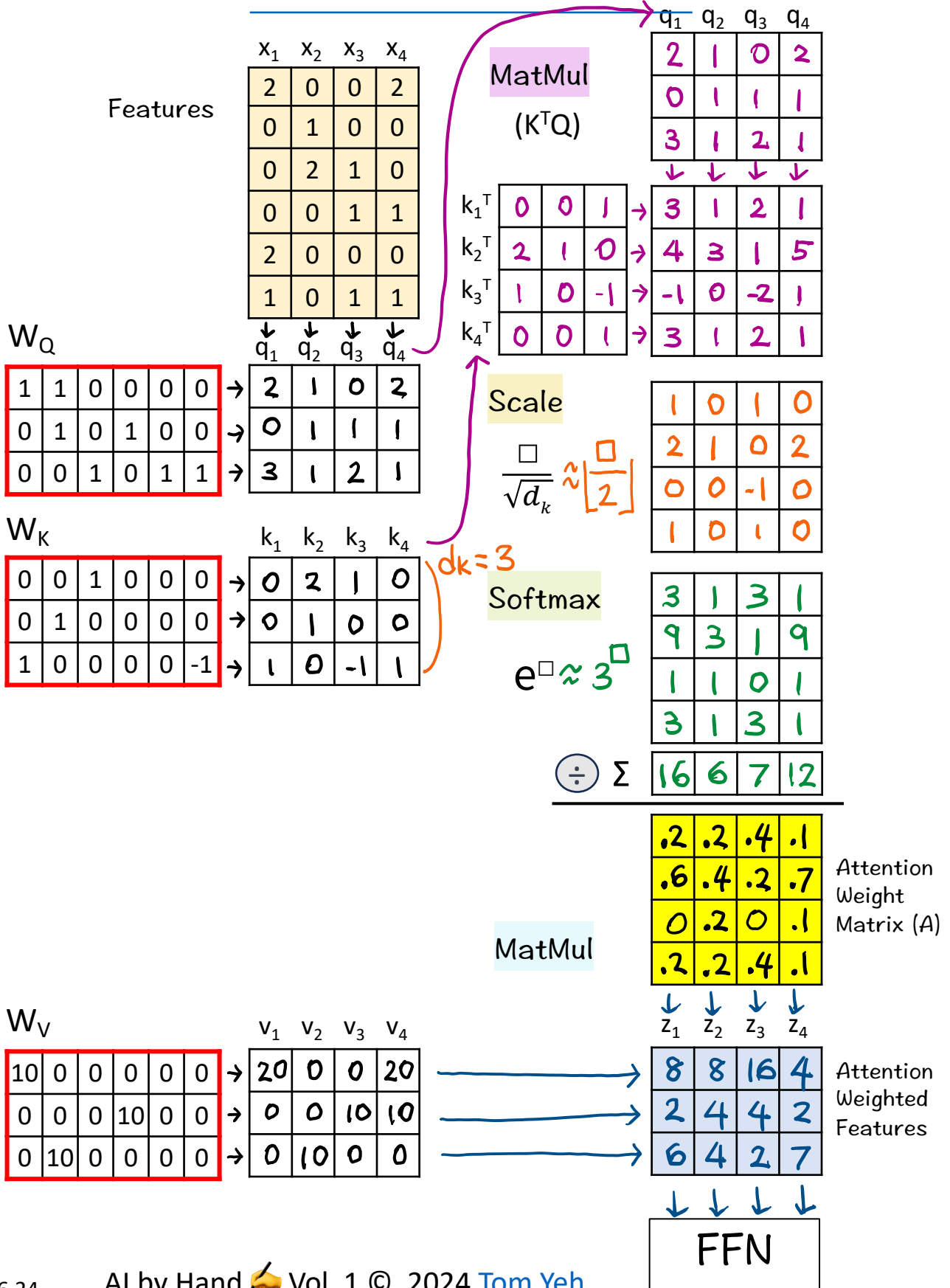
Y_G

1	1	1	1
---	---	---	---

Loss Gradients: $\frac{\partial L_G}{\partial Z}$

-.3	-.5	-.1	-.7
-----	-----	-----	-----

11. Self Attention



12. Dropout

Random Sequence

.61 >.5
 .39
 .75
 .40
 .65 >.33
 .42
 .23
 .19
 .93
 .42
 .87
 .53
 .27
 .69
 .50
 .11
 .42

Linear

Training Data:

X_1	X_2
3	5
4	1
1	1

1	0	0
1	1	0
0	1	1
1	-1	0
[$\approx$ ReLU]		
3	5	
7	6	
5	2	
4	4	
1	1	

Dropout
 (p=0.5)

$$\frac{1}{1-p}=2$$

2	0	0	0
0	1	0	0
0	0	2	0
0	0	0	1
6	10		
0	0		
10	4		
0	0		

Linear

1	0	0	1	0
0	1	1	0	0
1	0	-1	-1	1
[$\approx$ ReLU]				
6	10			
10	4			
4	6			
1	1			

Dropout
 (p=0.33)

$$\frac{1}{1-p}=1.5$$

1.5	0	0
0	1.5	0
0	0	1
9	15	
15	6	
0	0	
1	1	

Linear

1	-1	0	0
0	1	-1	-2
-6	9		
13	4		

Outputs
 Y

Inference

Unseen Data:

3	3
2	1
1	1

1	0	0
1	1	1
-1	1	1
1	-1	0
[$\approx$ ReLU]		
3	3	
6	5	
0	4	
1	2	
1	1	

1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1
3	3		
6	5		
0	0		
1	2		

1	0	1	1	0
1	1	1	0	0
1	0	-1	0	1
[$\approx$ ReLU]				
4	5			
9	8			
4	4			
1	1			

1	0	0
0	1	0
0	0	1
4	5	
9	8	
4	4	

1	1	0	0
0	1	-1	-1
13	13		
4	3		

Training

-	-4	7
	10	5

Targets
 Y'

-2	2
3	-1

x 2

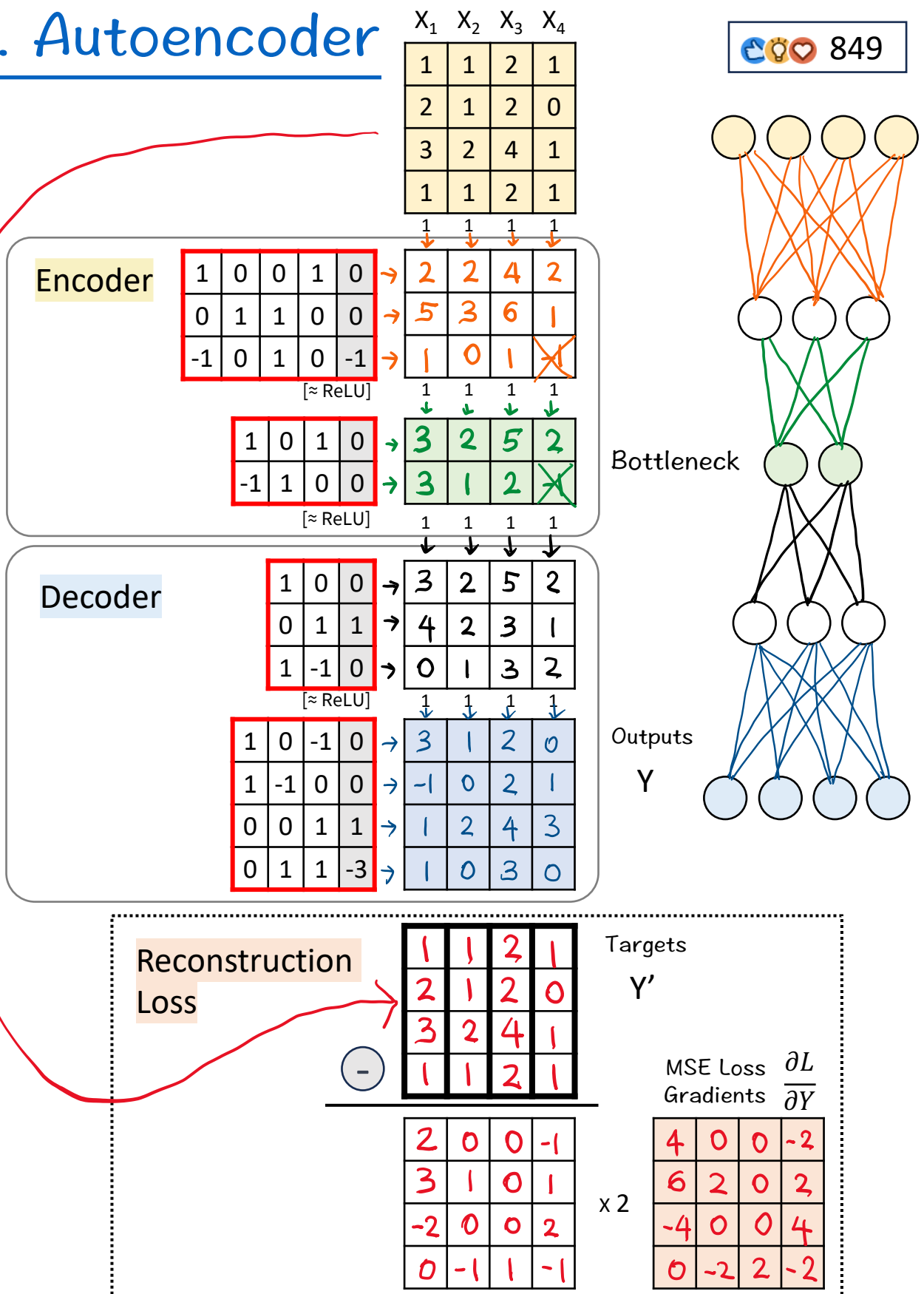
-4	4
6	-2

MSE Loss
 Gradients

$$\frac{\partial L}{\partial Y}$$

13. Autoencoder

849



14. Vector Database

Query

am I you

Data how are you who are you who am I

Word Embeddings

a	an	the	how	why	who	what	are	is	am	be	was	you	we	I	they	she	he	she	me	him	her
0	-1	0	1	0	1	0	0	-1	1	0	0	0	3	1	0	-1	0	0	0	-1	0
2	0	2	0	0	0	-1	1	0	0	0	2	1	0	2	0	2	0	0	2	0	0
-1	0	-1	1	2	0	0	1	0	1	-1	0	0	-1	0	3	0	0	-1	0	2	-1
0	1	0	0	1	0	1	0	1	0	1	-2	0	0	0	1	0	1	0	1	0	1

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

1

1

1

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

1

1

1

1	1	1
0	0	2
0	1	0
0	0	0

1

1

1

1	1	0
0	2	1
1	0	0
0	0	0

1

1

1

Text Embeddings

Encoder

1	1	0	0	0
0	1	0	1	0
1	0	1	0	-1
1	-1	0	0	0

Linear & ReLU

1	1	1
0	1	1
1	0	X
1	X	X

1	1	1
0	1	1
0	0	X
0	X	X

1	1	3
0	0	2
0	1	0
0	1	X

1	3	1
0	2	1
1	0	X
1	X	X

Mean Pooling

$$\Sigma / 3$$

3/3
2/3
1/3
1/3

3/3
2/3
0
0

5/3
2/3
1/3
1/3

5/3
3/3
1/3
1/3

Indexing

Projection

1	1	0	0
0	0	1	1

5/3
2/3

5/3
0

7/3
2/3

Vector Storage

8/3
2/3

Retrieval

Dot Products

$$44/9$$

$$40/9$$

$$60/9$$

$$8/3 \cdot 2/3$$

Nearest Neighbor (argmax)

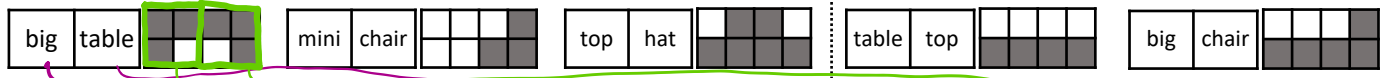


who am I

15. CLIP

mini batch of text-image pairs

400 millions more ...



word2vec

mini	big	top	hat	chair	table
0	1	1	0	1	0
1	0	0	0	1	1
0	1	0	1	0	1

Flatten Patches

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

Word Embeddings

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

Image Encoder

1 1 0 0 0	2 2	0 1	1 1
0 1 0 1 1	2 3	1 3	3 2
0 0 1 -1 1	2 0	1 1	1 1
0 1 1 1 -1	1 1	1 2	2 1

Text Encoder

1 0 1 0	2 1	0 1	1 0
0 3 0 -2	2 1	1 1	-2 -2
1 1 0 1	2 2	2 3	2 1

2	1	1
3	2	3
2	1	1
2	1	2

[Mean Pooling] (round)

[Mean Pooling] (round)

[Projection]

[Projection]

1	1	0	-2
0	1	1	-2

-1	1	1	0	-1
0	0	-1	1	0

2	1	2
0	0	1

Shared Embedding Space

T ₁	T ₂	T ₃
1	0	-1
1	2	0

[Softmax] $e^{\square} \approx 3^{\square}$ Σ

Similarity Image → Text	Target
.9 .1 0	1 0 0
.75 .25 0	0 1 0
.7 .3 0	0 0 1

Cross Entropy Loss Gradients

Image → Text
-.1 .1 0
.75 -.75 0
.7 .3 -1

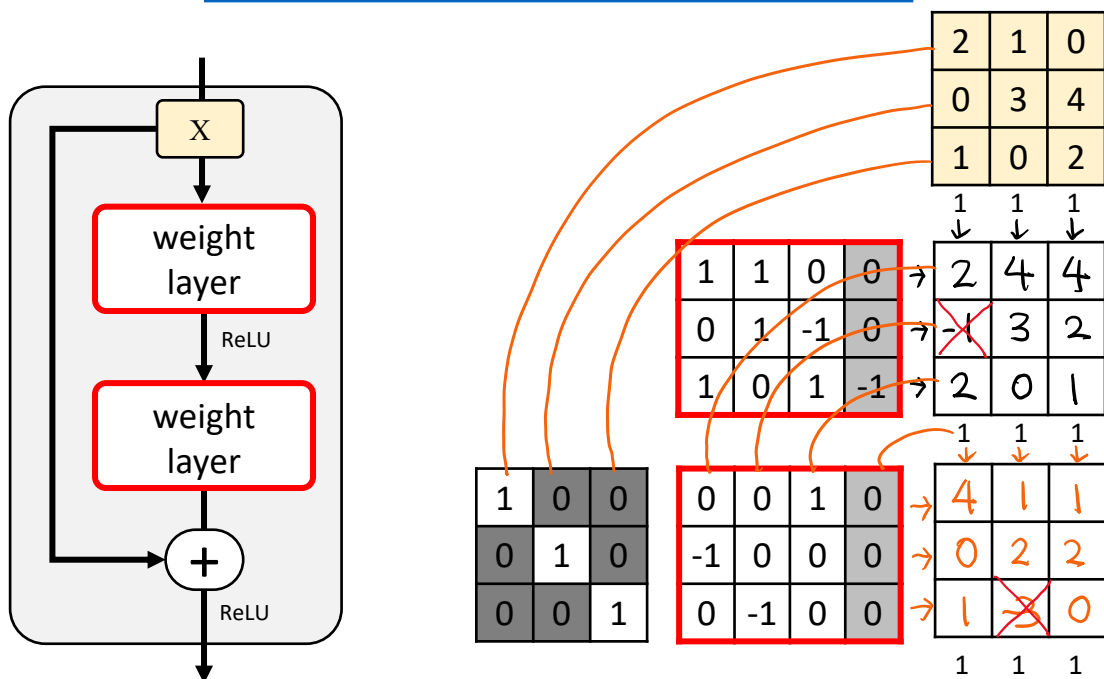
I ₁	I ₂	I ₃
2 0	2 0 -2	9 1 0 10
1 0	1 0 -1	3 1 0 4
2 1	3 2 -2	27 9 0 36

Similarity Text → Image
.2 .1 0
.1 .1 0
.7 .8 0

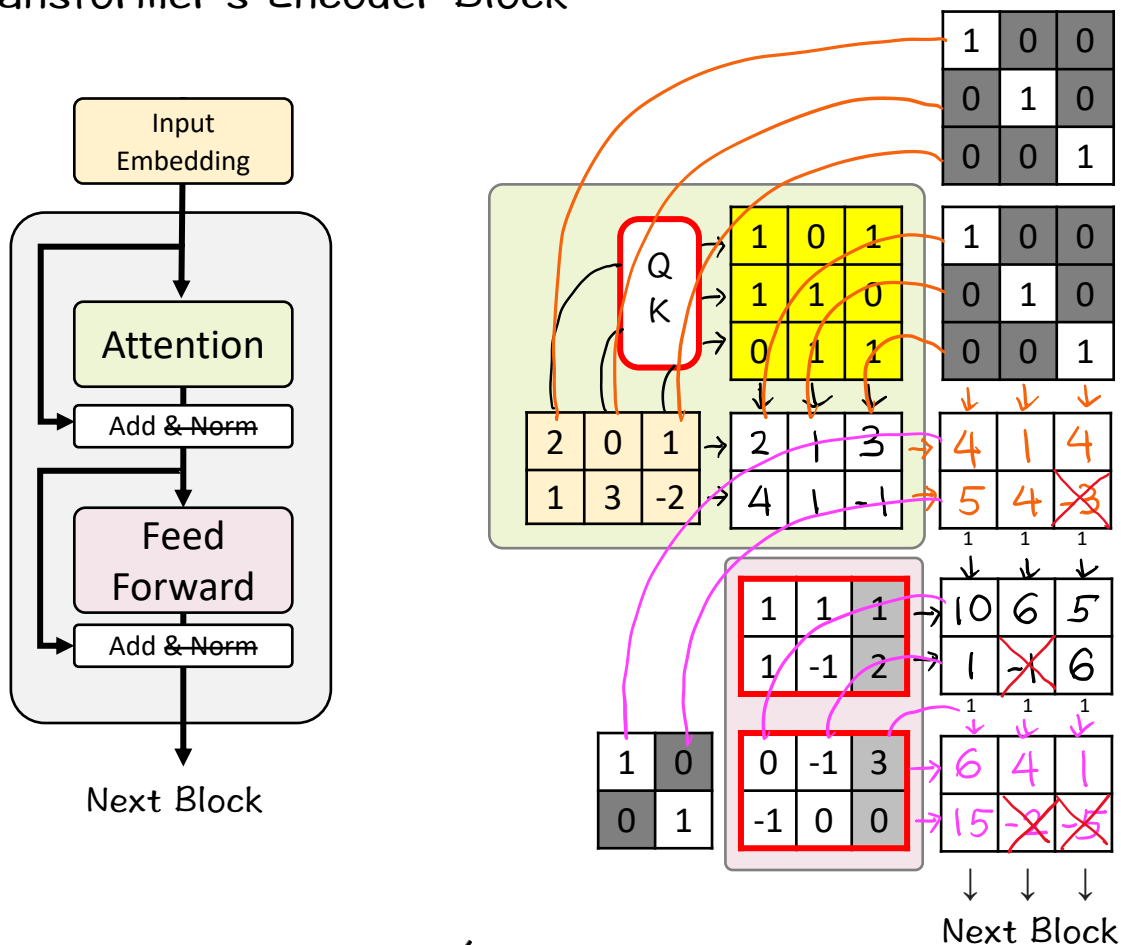
1 0 0
0 1 0
0 0 1

Text → Image
-.8 .1 0
.1 -.9 0
.7 .8 -1

16. Residual Network



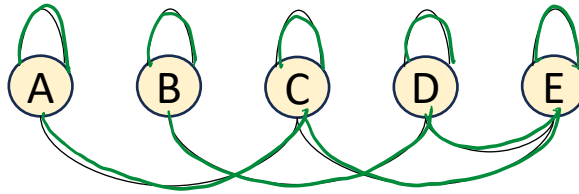
Transformer's Encoder Block



17. Graph Convolutional Network

🔗💡❤️ 573

Graph Data



Graph Convolutional Network

	A	B	C	D	E
A	1		1		
B		1		1	
C	1		1		1
D		1		1	1
E			1	1	1

Adjacency Matrix

A	B	C	D	E
1		1		
	1		1	
1		1		1
	1		1	1
		1	1	1

1	1	0	0	0
0	1	0	-1	0
1	0	0	1	-1

[ReLU]

1	1	1	1	1
3	1	1	0	1
0	1	2	1	0
1	2	0	0	0

Messages

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

[ReLU]

1	1	1	1	1
4	1	5	2	1
0	1	0	1	0
1	2	1	2	0

Messages

1	1	1	1	1
2	6	2	6	4
5	0	5	0	3
9	3	10	4	8

Fully Connected Network

1	0	0	-2
0	1	0	-2
0	0	1	-5
1	-1	0	0
1	0	-1	0

[ReLU]

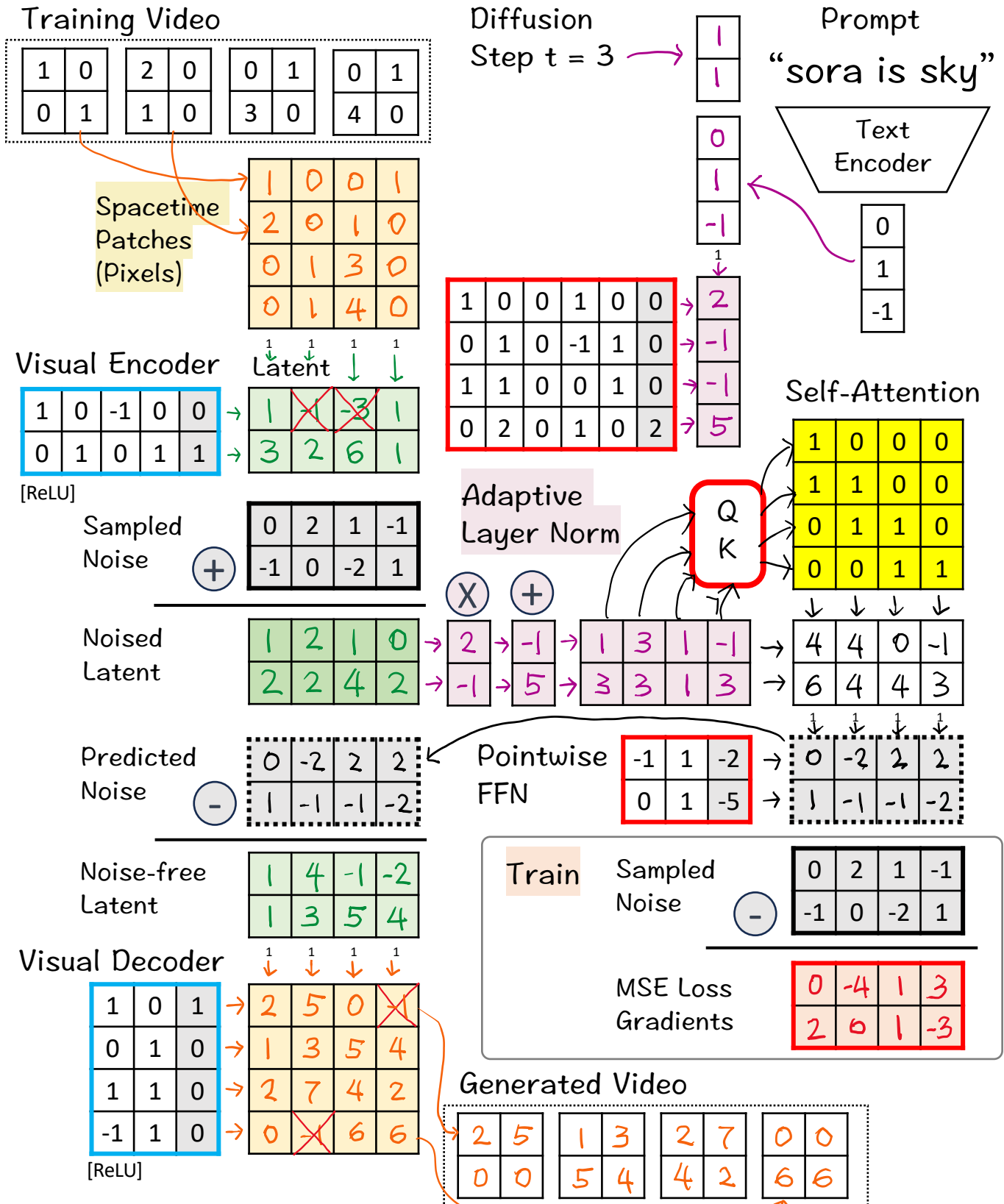
1	1	1	1	1
1	1	1	1	-9

1	1	1	1	1
0	4	0	4	2
3	2	3	2	1
2	4	3	1	1
3	6	3	6	1
7	3	8	2	4

σ

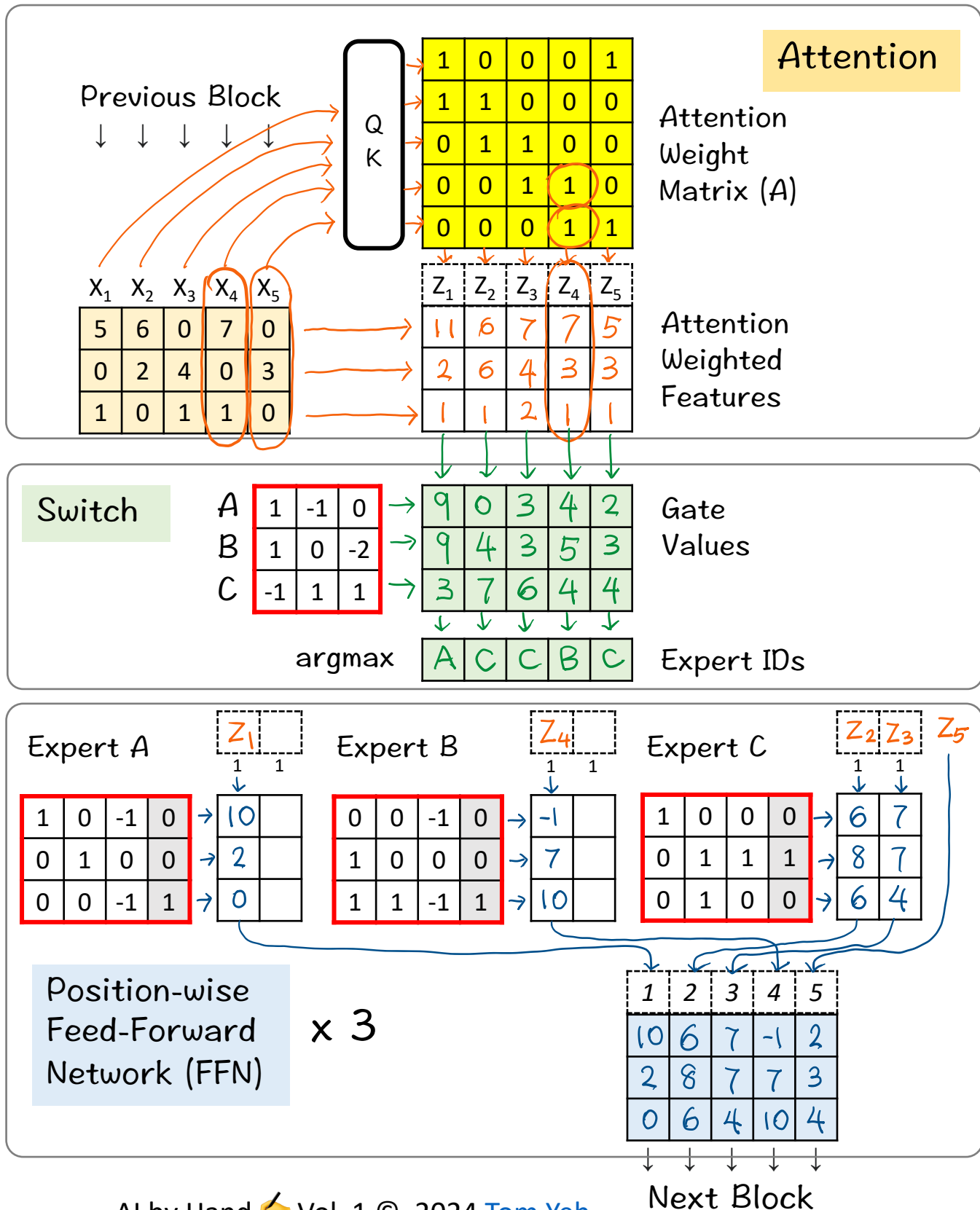
1	1	1	1	1
-4	4	5	3	0
0	1	1	1	0.5

18. SORA's Diffusion Transformer



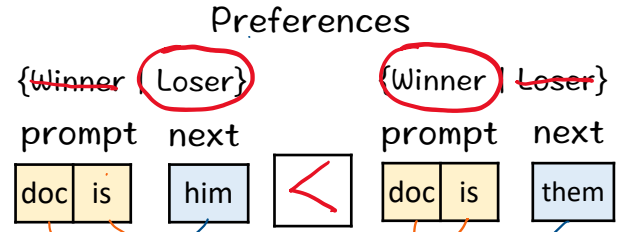
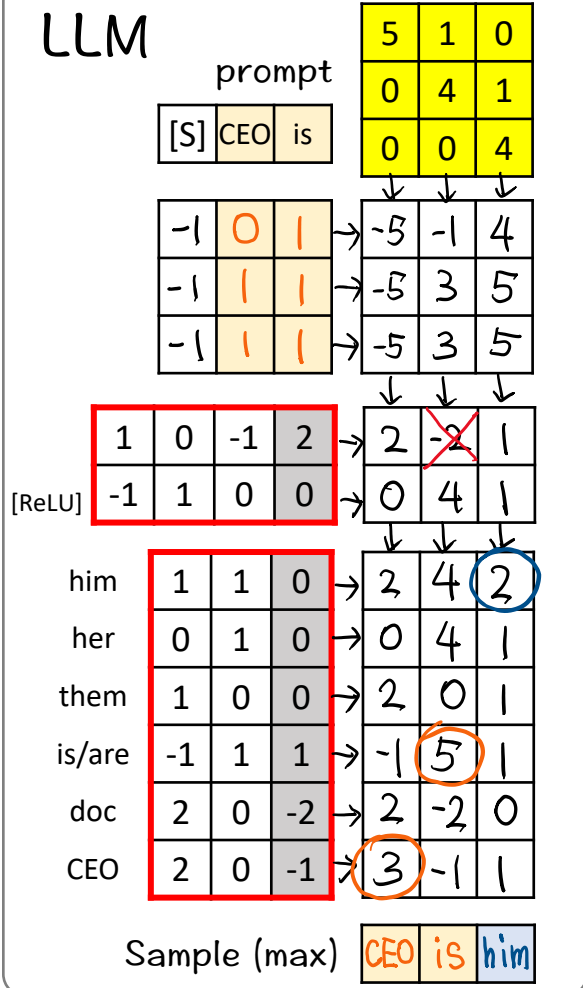
19. Switch Transformer

(Gemini 1.5's Sparse Mixture of Experts)



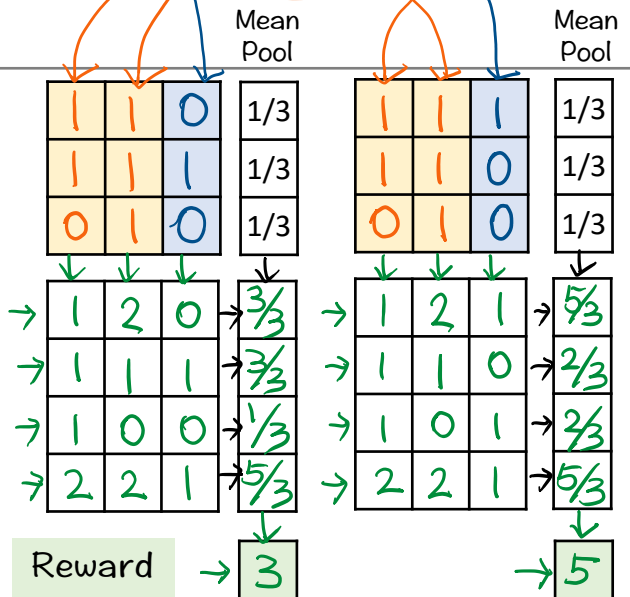
20. Reinforcement Learning with Human Feedback (RLHF)

🧠💡❤️ 550

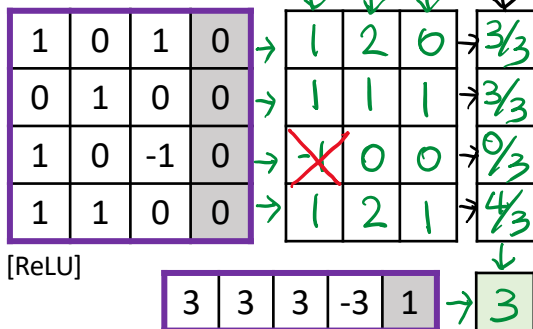


Word Embeddings

him	her	them	is	doc	CEO	[S]
0	1	1	1	1	0	-1
1	0	0	1	1	1	-1
0	1	0	1	0	1	-1



Reward Model (RM)



Align LLM

Loss -3

Loss Gradient -1

Train RM

5 - 3 Winner - Loser 2

Predicted σ 0.9

- Target -1

Loss Gradient -0.1