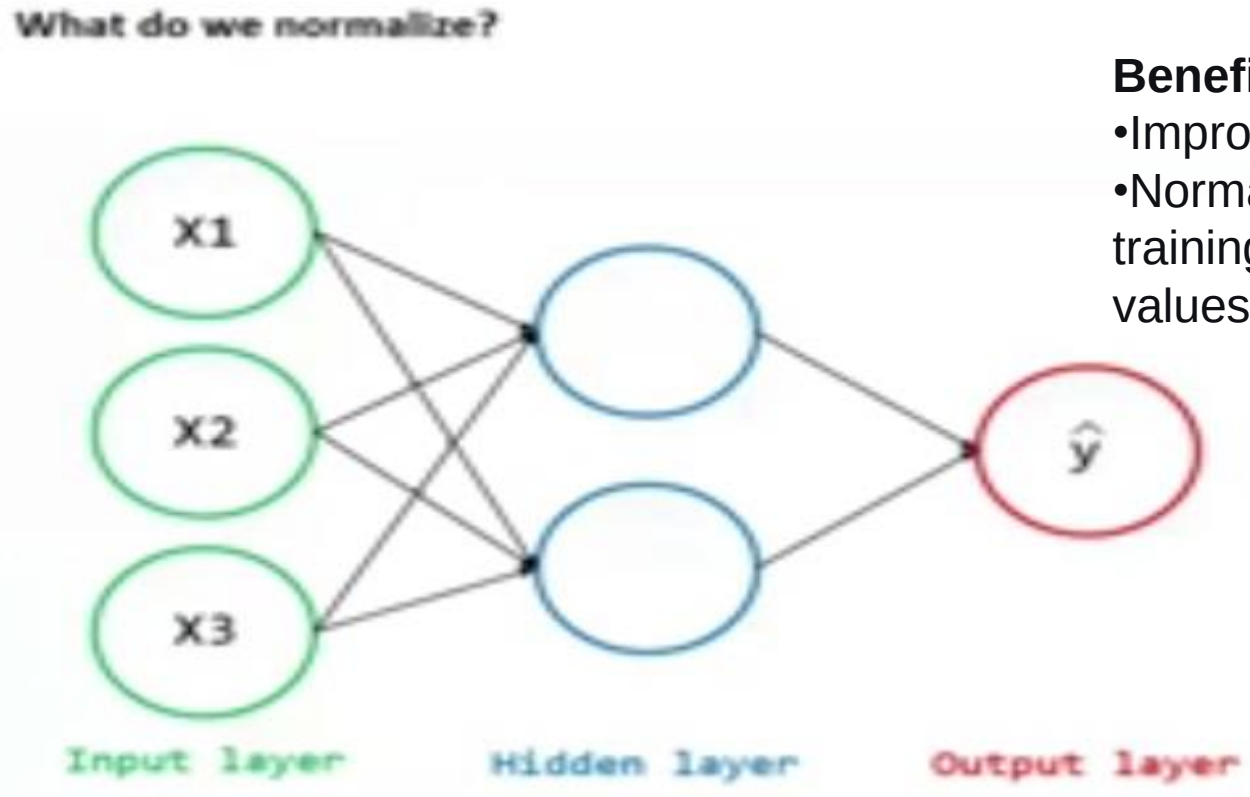
Normalization-Transformer

Dr. Muhammad Safyan

- Normalization in deep learning refers to the process of transforming data or model outputs to have specific statistical properties, typically a mean of zero and a variance of one.



Benefits of Normalization in Deep Learning

•Improved Training Stability:

•Normalization helps to stabilize and accelerate the training process by reducing the likelihood of extreme values that can cause gradients to explode or vanish.

Benefits of Normalization in Deep Learning

- Improved Training Stability:
  - Normalization helps to stabilize and accelerate the training process by reducing the likelihood of extreme values that can cause gradients to explode or vanish.
- Faster Convergence:
  - By normalizing inputs or activations, models can converge more quickly because the gradients have more consistent magnitudes. This allows for more stable updates during back propagation.
- Mitigating Internal Covariate Shift:
  - Internal covariate shift refers to the change in the distribution of layer inputs during training. Normalization techniques, like batch normalization, help to reduce this shift, making the training process more robust.

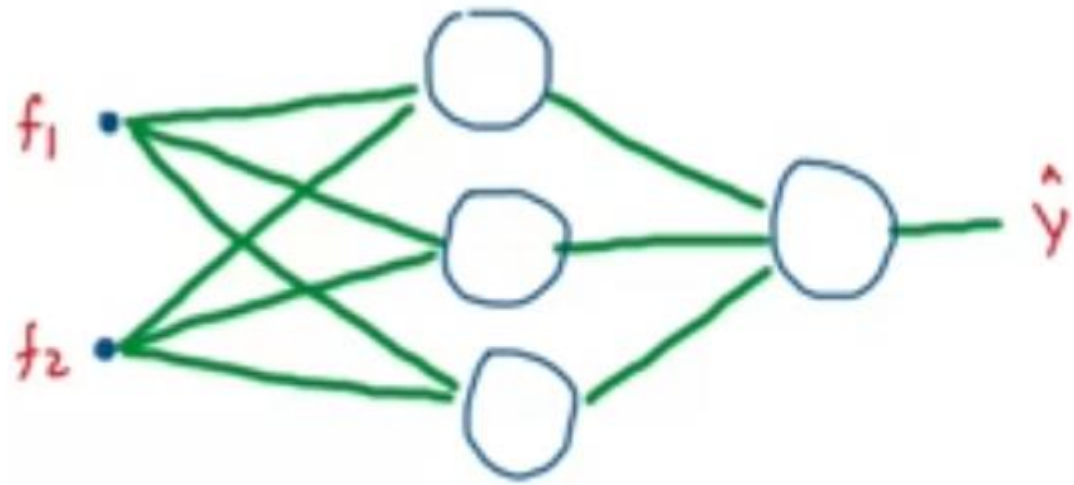
input data → dist

covariate shift

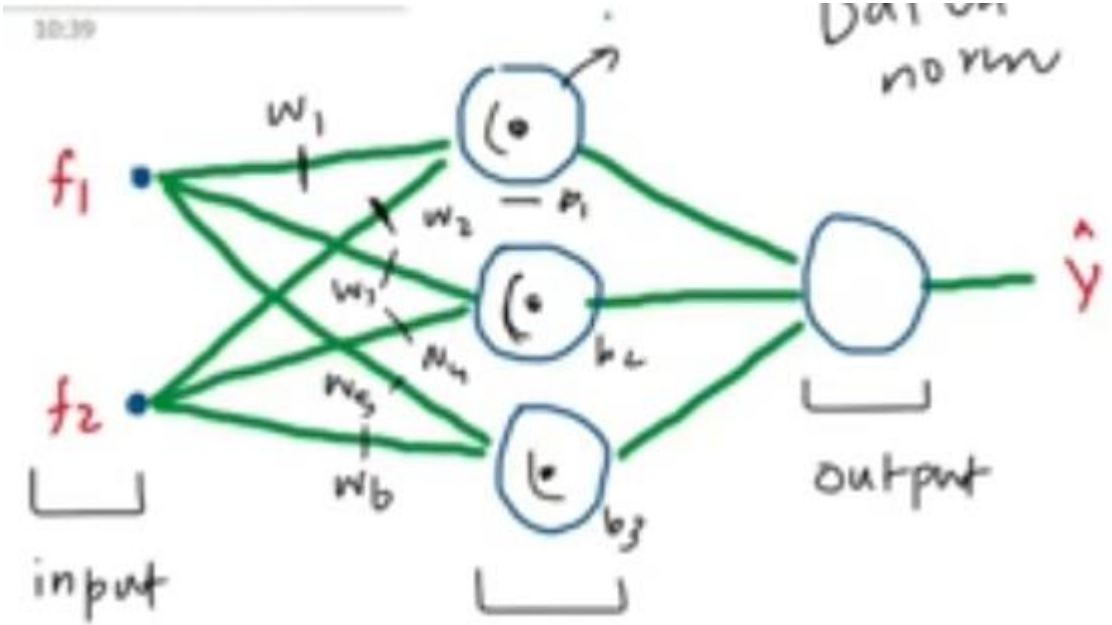
red roses

pred

yellow white



f_1	f_2
2	3
1	1
5	4
6	1
7	1



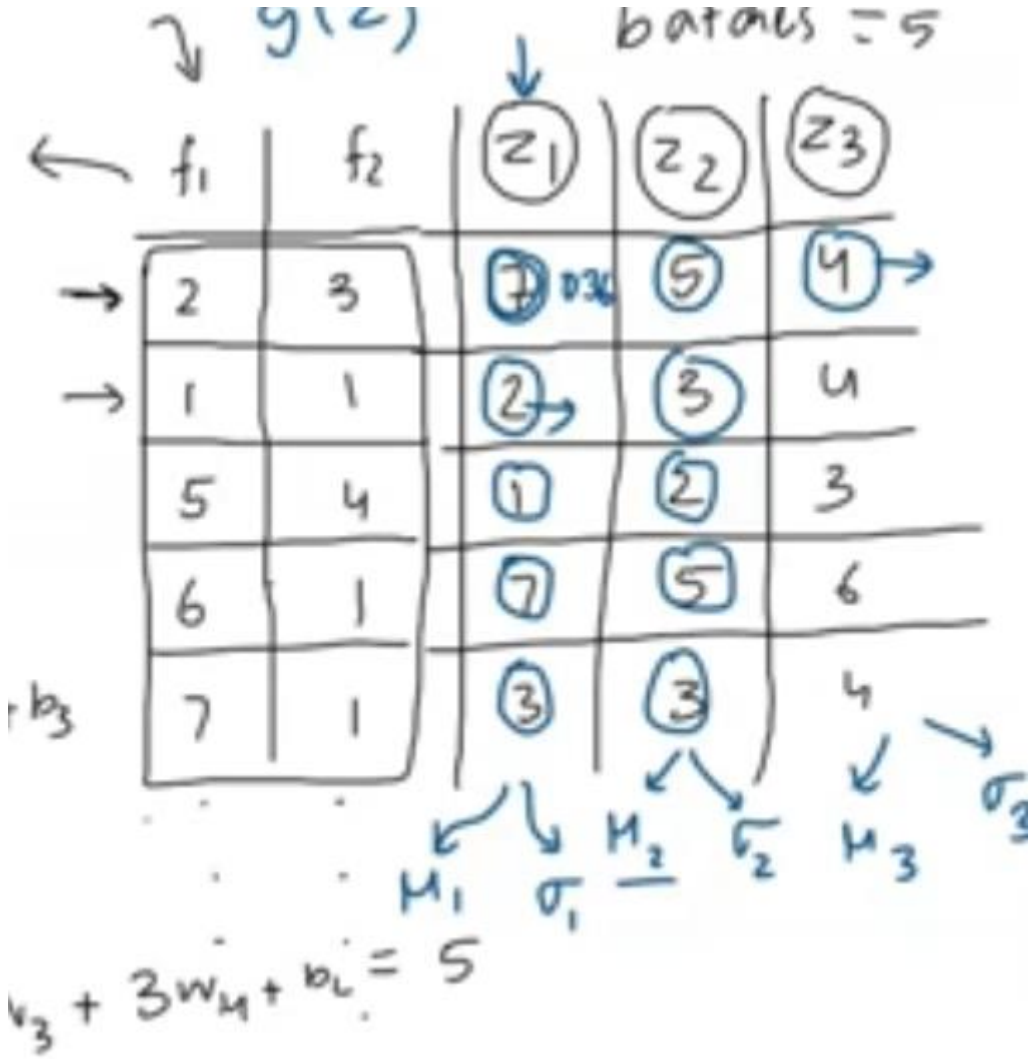
f_1	f_2	z_1	z_2	z_3
2	3	7	5	4
1	1	2	3	4
5	4	1	2	3
6	1	7	5	6
7	1	3	3	4

$z_3 = 2w_5 + 3w_6 + b_3 = 4$

$z_1 = 2w_1 + 3w_2 + b_1 = 7$

$z_2 = 2w_3 + 3w_4 + b_2 = 5$

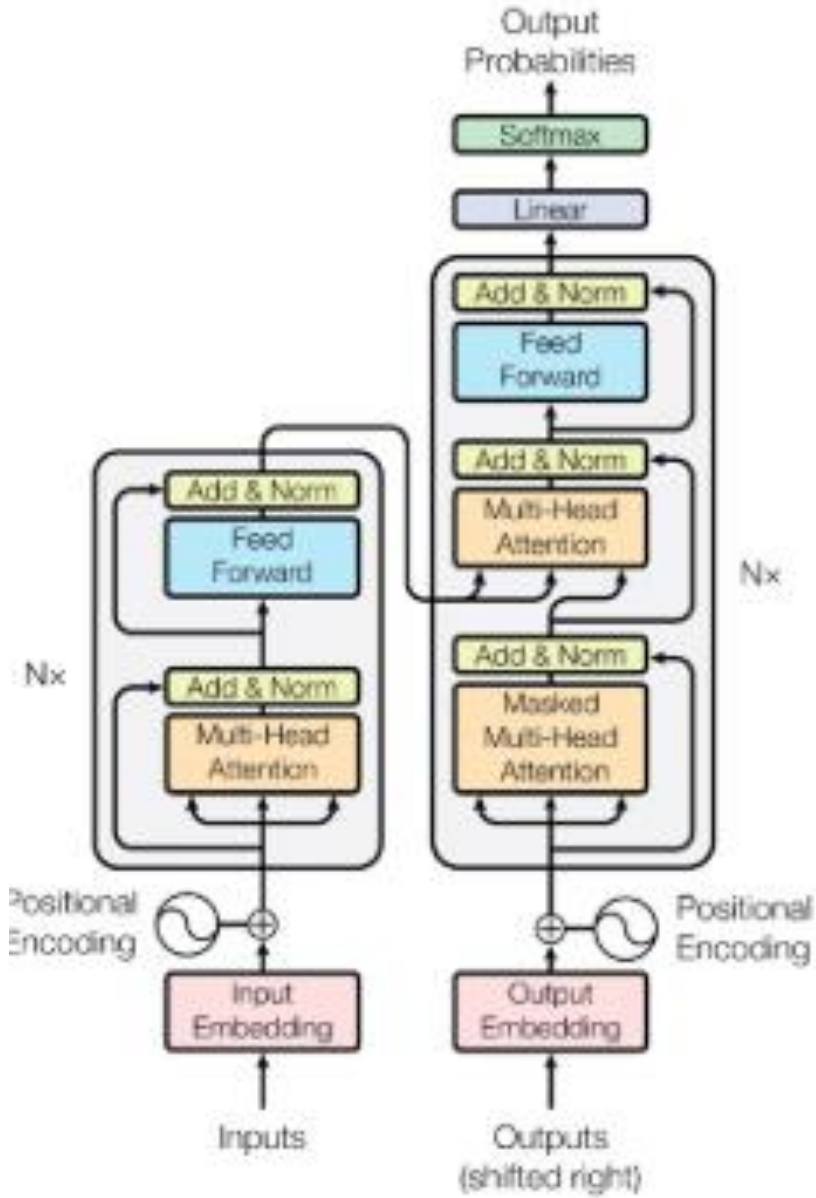
μ_1 σ_1 μ_2 σ_2 μ

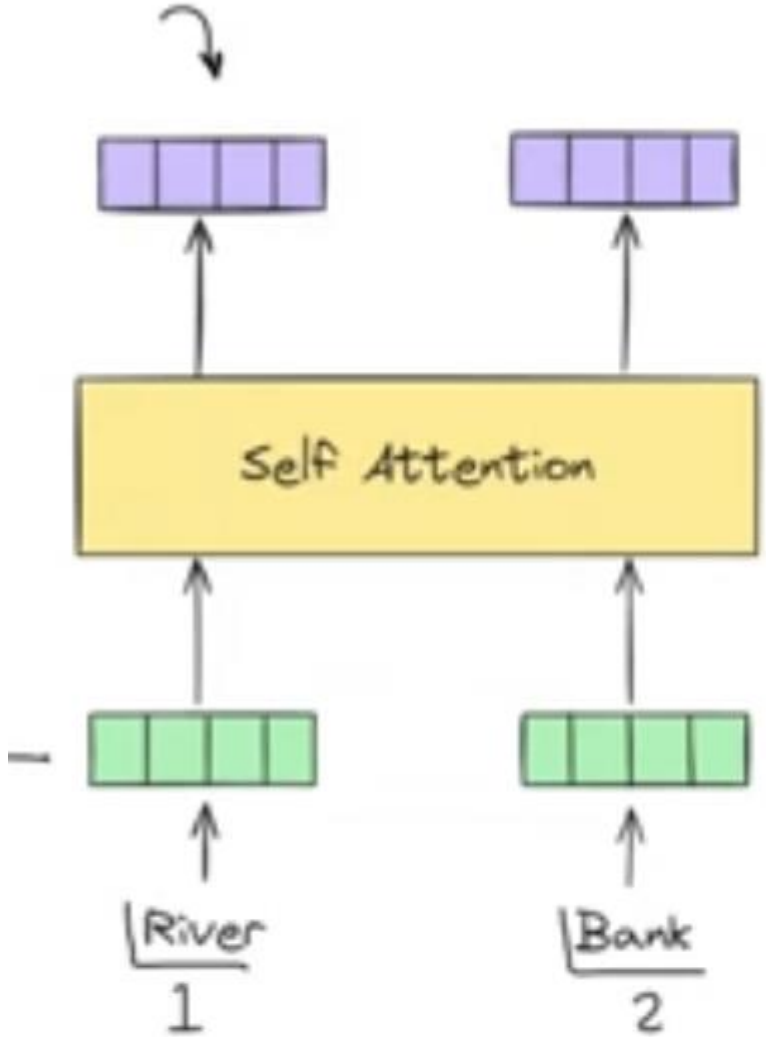


$$w_3 + 3w_4 + b_2 = 5$$

$$\frac{7-\mu_1}{\sigma_1} = 0.36 \quad \underset{(1)}{\gamma_1} + \underset{(0)}{\beta_1} = 0.36$$

$$\frac{2-\mu_1}{\sigma_1} = 0.71\,\gamma_1 + \beta_1 = 0.71$$


Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)



Review	Sentiment
Hi Babar	1
How are you today	0
I am good	0
You?	1

Embedding dimension - 3

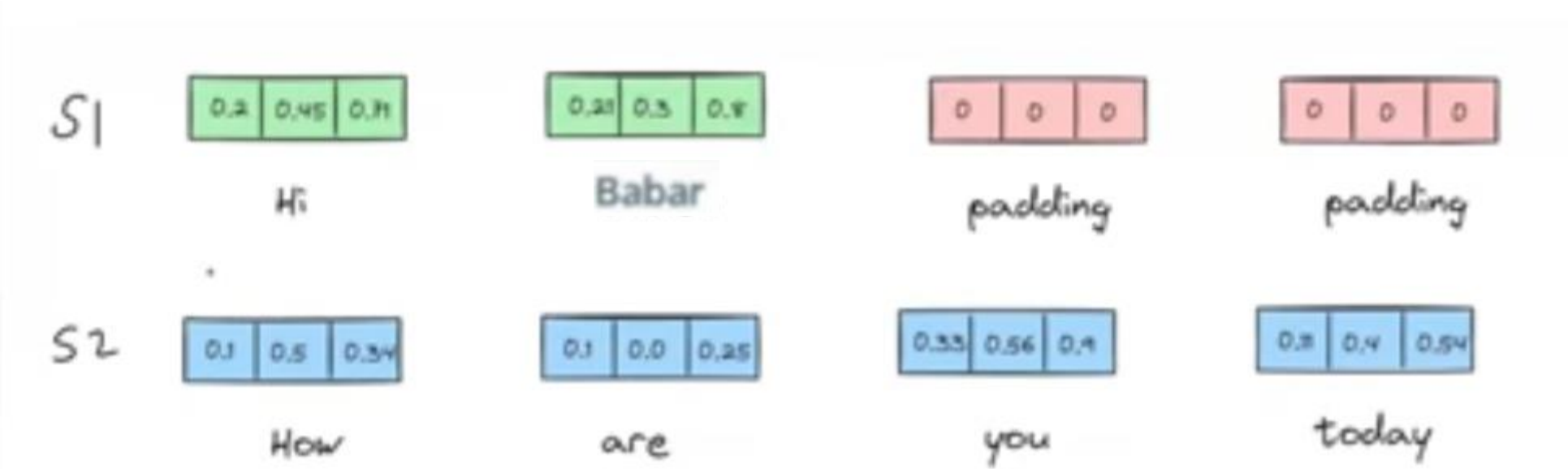
Batch Size - 2

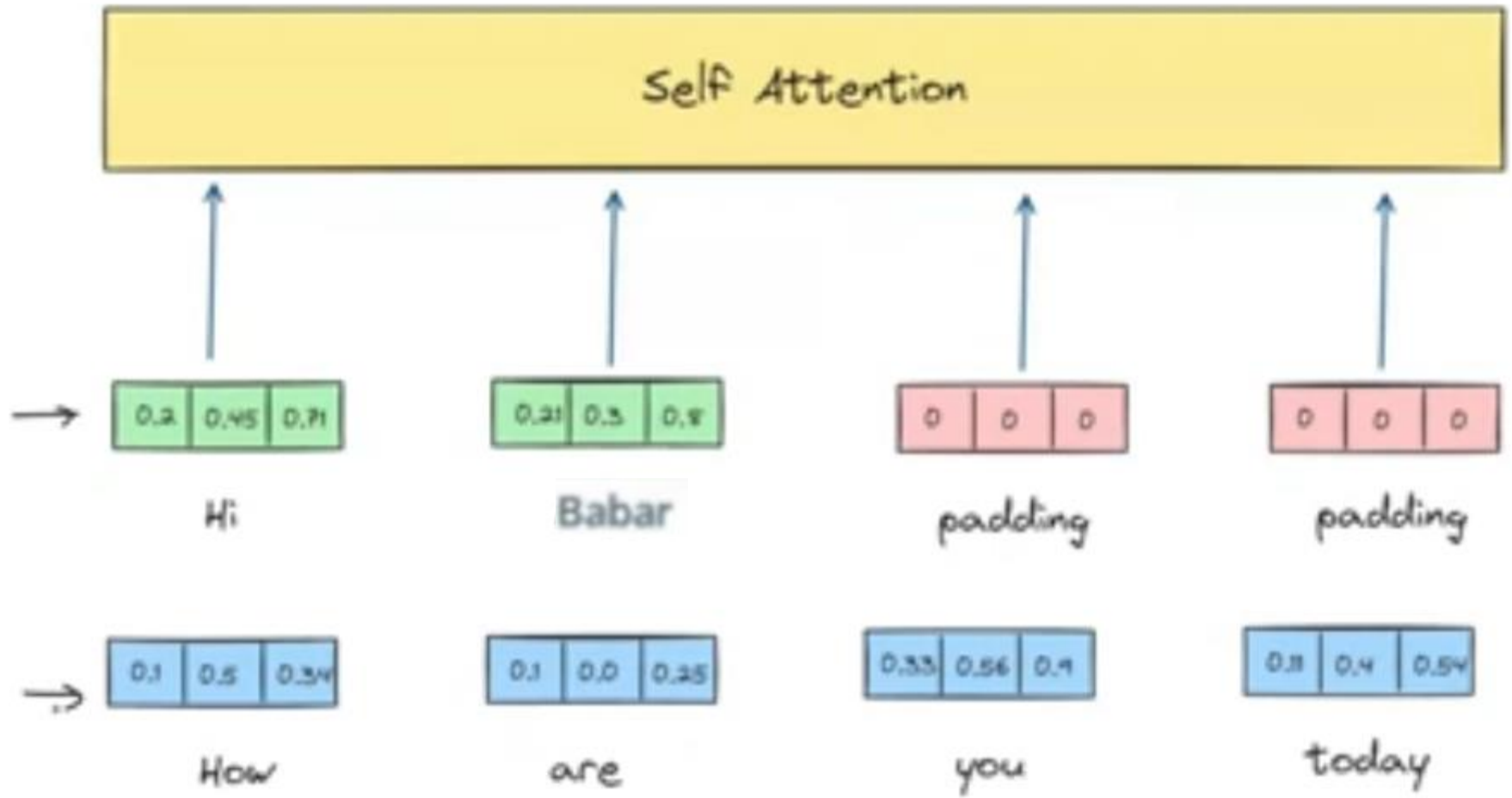
Review	Sentiment
Hi Babar	1
How are you today	0
I am good	0
You?	1

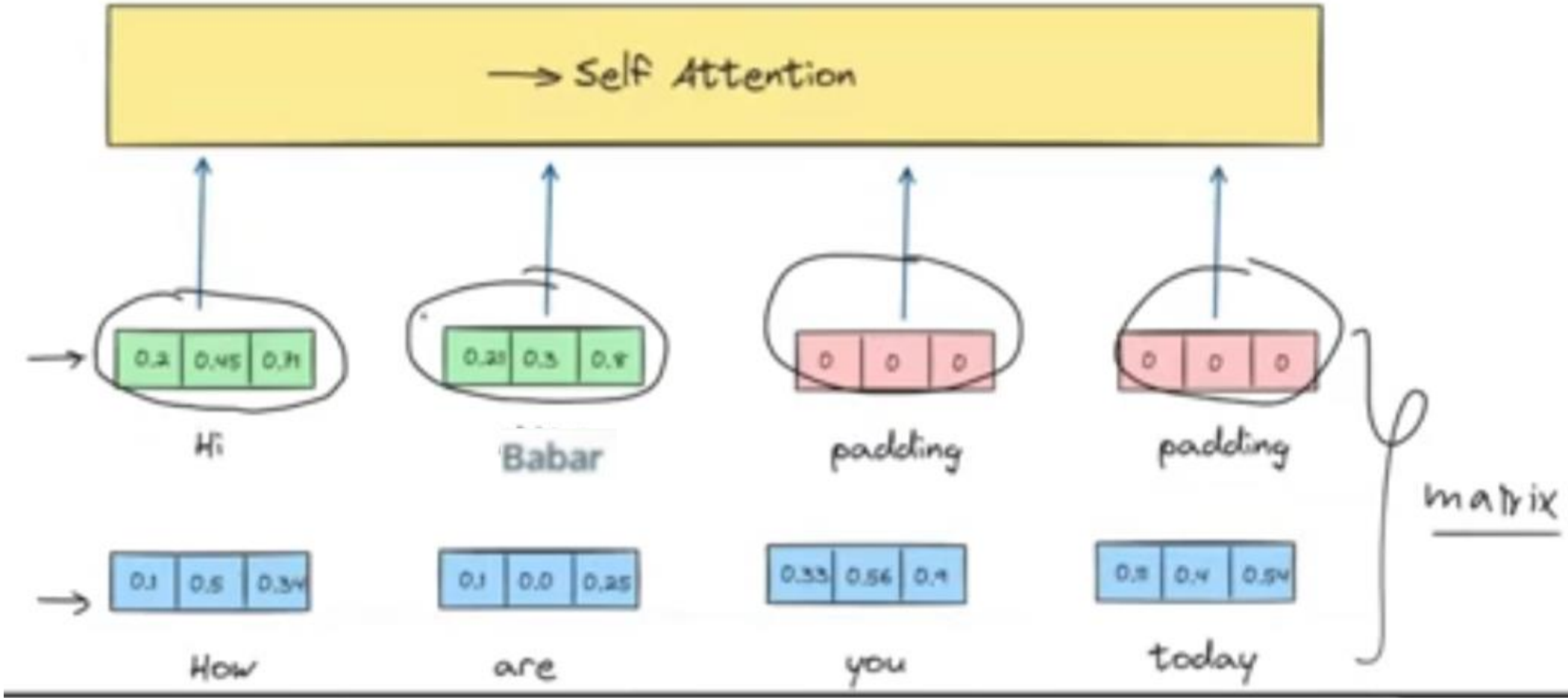
Embedding dimension - 3

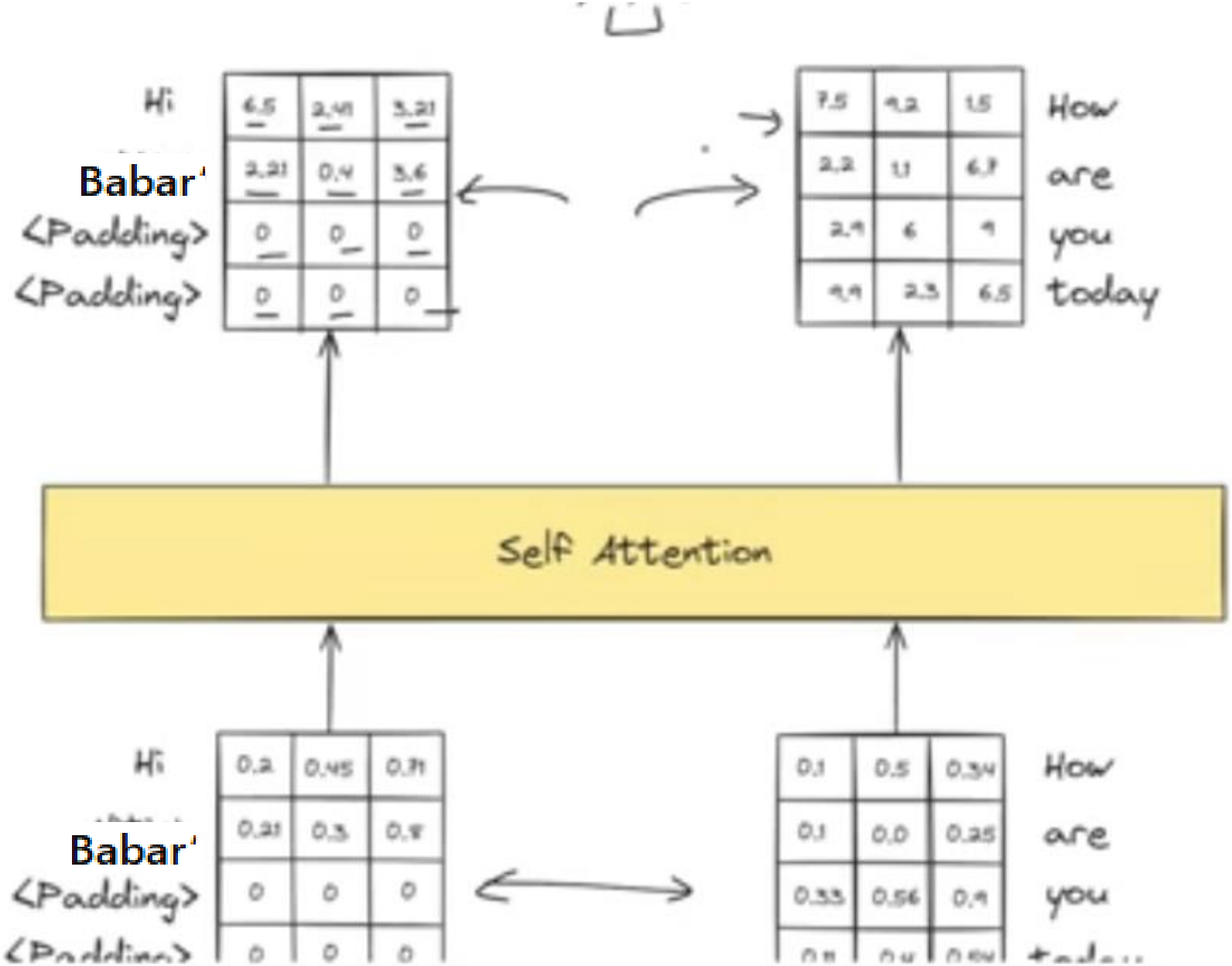
Batch Size - 2

padding

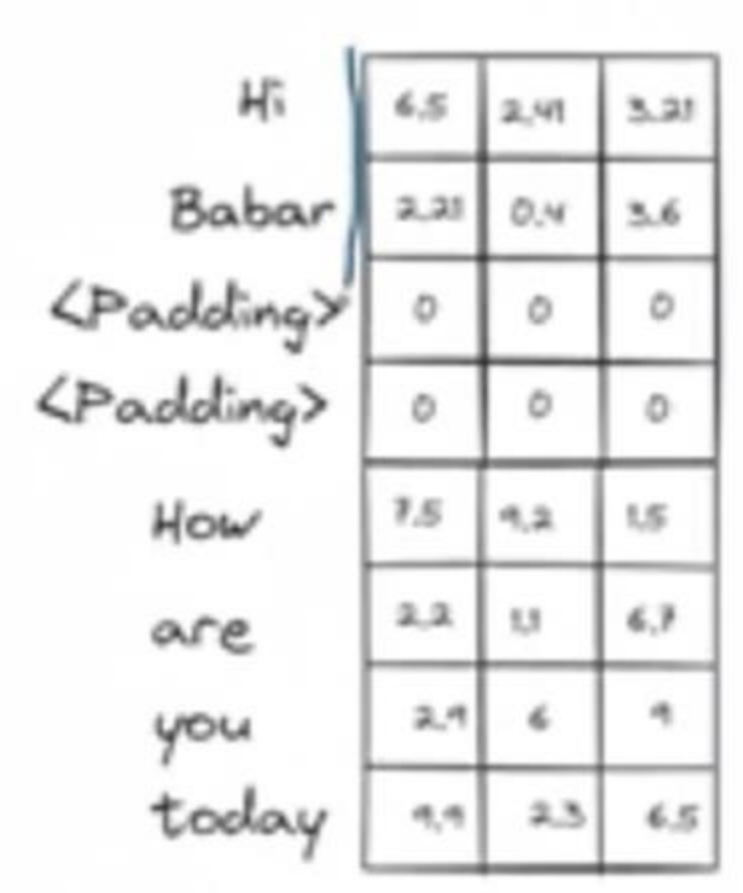

S1
0.2
0.45
0.71
Hi
0.21
0.3
0.8
Babar
0
0
0
padding
0
0
0
padding
S2
0.1
0.5
0.54
How
0.1
0.0
0.25
are
0.33
0.56
0.9
you
0.31
0.4
0.54
today


Self Attention
0.2
0.45
0.71
0.21
0.5
0.8
0
0
0
0
0
0
Hi
Babar
padding
padding
0.1
0.5
0.34
0.1
0.0
0.25
0.33
0.56
0.9
0.11
0.4
0.54
How
are
you
today

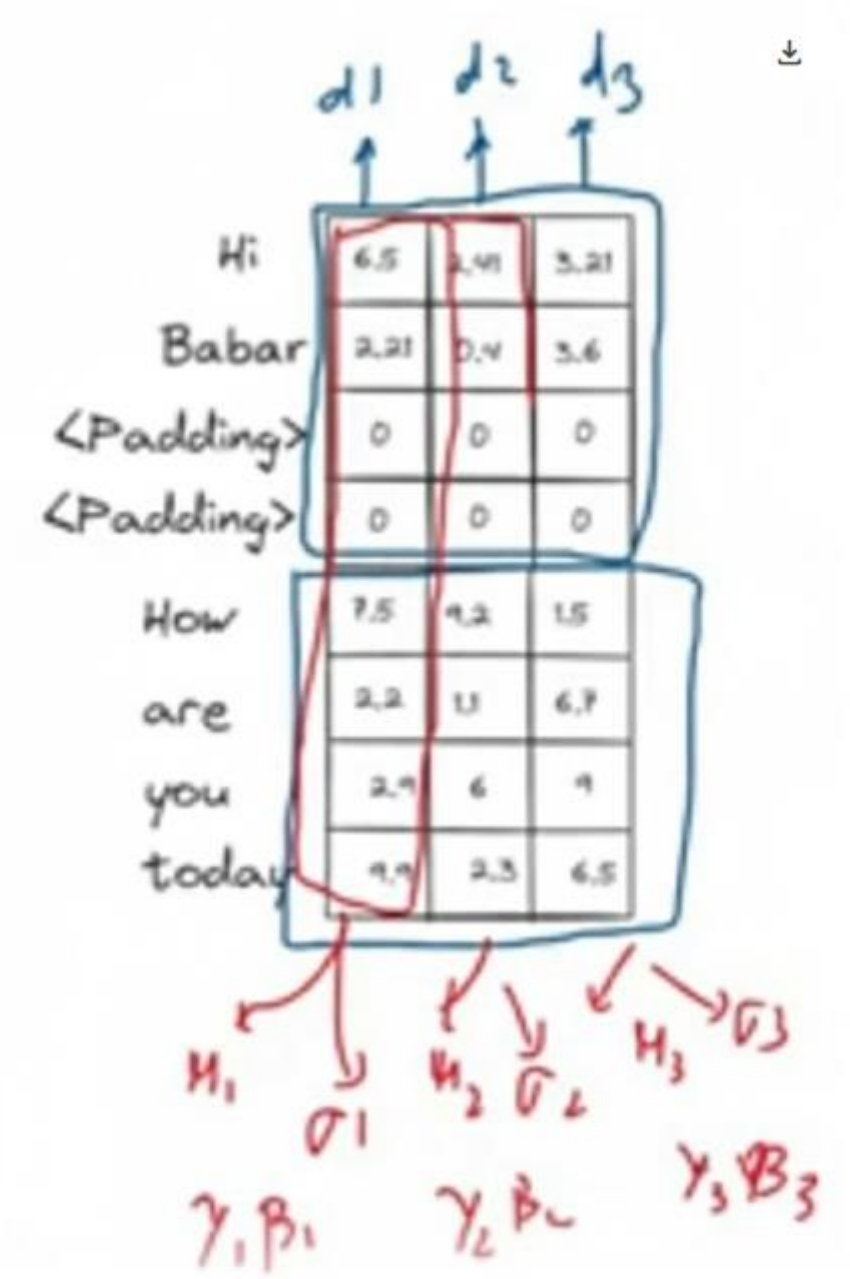

Self Attention
0.2
0.45
0.31
0.21
0.5
0.8
0
0
0
0
0
0
Hi
Babar
padding
padding
matrix
0.1
0.5
0.34
0.1
0.0
0.25
0.33
0.56
0.9
0.8
0.4
0.54
How
are
you
today


Hi
Babar'
<Padding>
<Padding>
6.5 2.41 3.21
2.21 0.4 3.6
0 0 0
0 0 0
7.5 9.2 1.5
2.2 1.1 6.7
2.9 6 9
9.9 2.3 6.5
How
are
you
today
Self Attention
Hi
Babar'
<Padding>
0.2 0.45 0.71
0.21 0.3 0.8
0 0 0
0 0 0
0.1 0.5 0.34
0.1 0.0 0.25
0.33 0.56 0.9
How
are
you

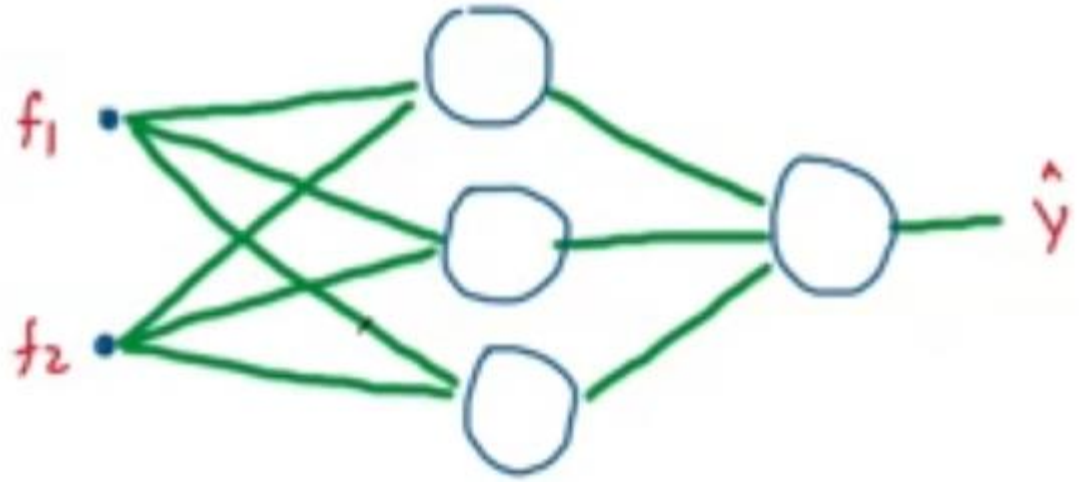
Vertically stackup



- Imagine
- 100 sentences batch
- Averge sentence lentg 20
- Longest sentence 60.



Layer Normalization

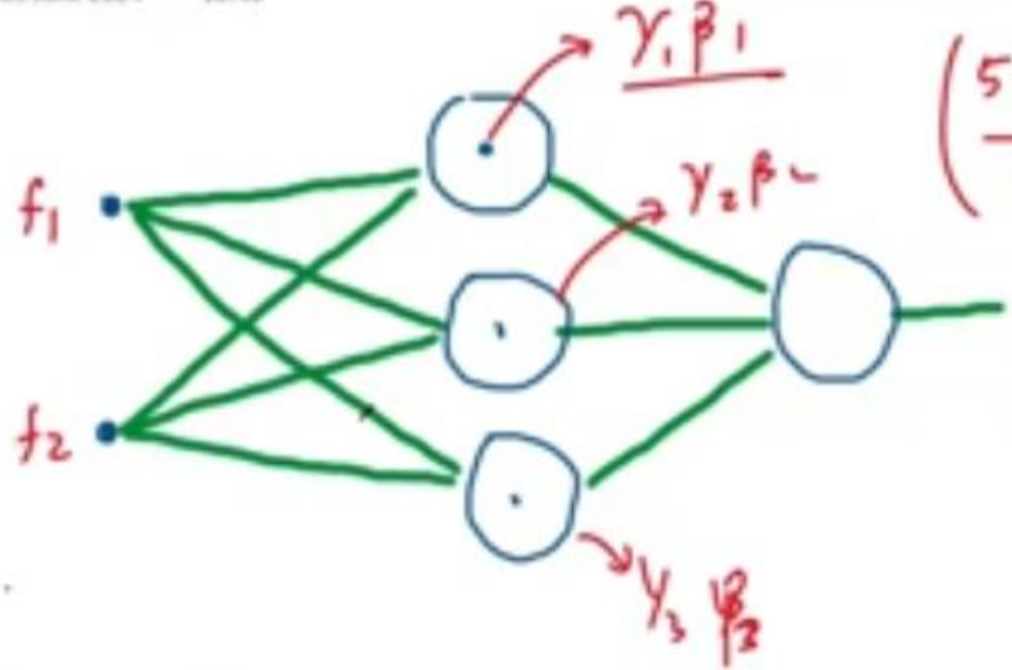


f_1	f_2	z_1	z_2	z_3
2	3	7	5	4
1	1	2	3	4
5	4	1	2	3
6	1	7	5	6
7	1	3	3	4

Layer Norm

05 June 2024 10:40

across batch across feature



$\left(\frac{5-\mu_4}{\sigma_4}\right)\gamma_2+\beta_2$

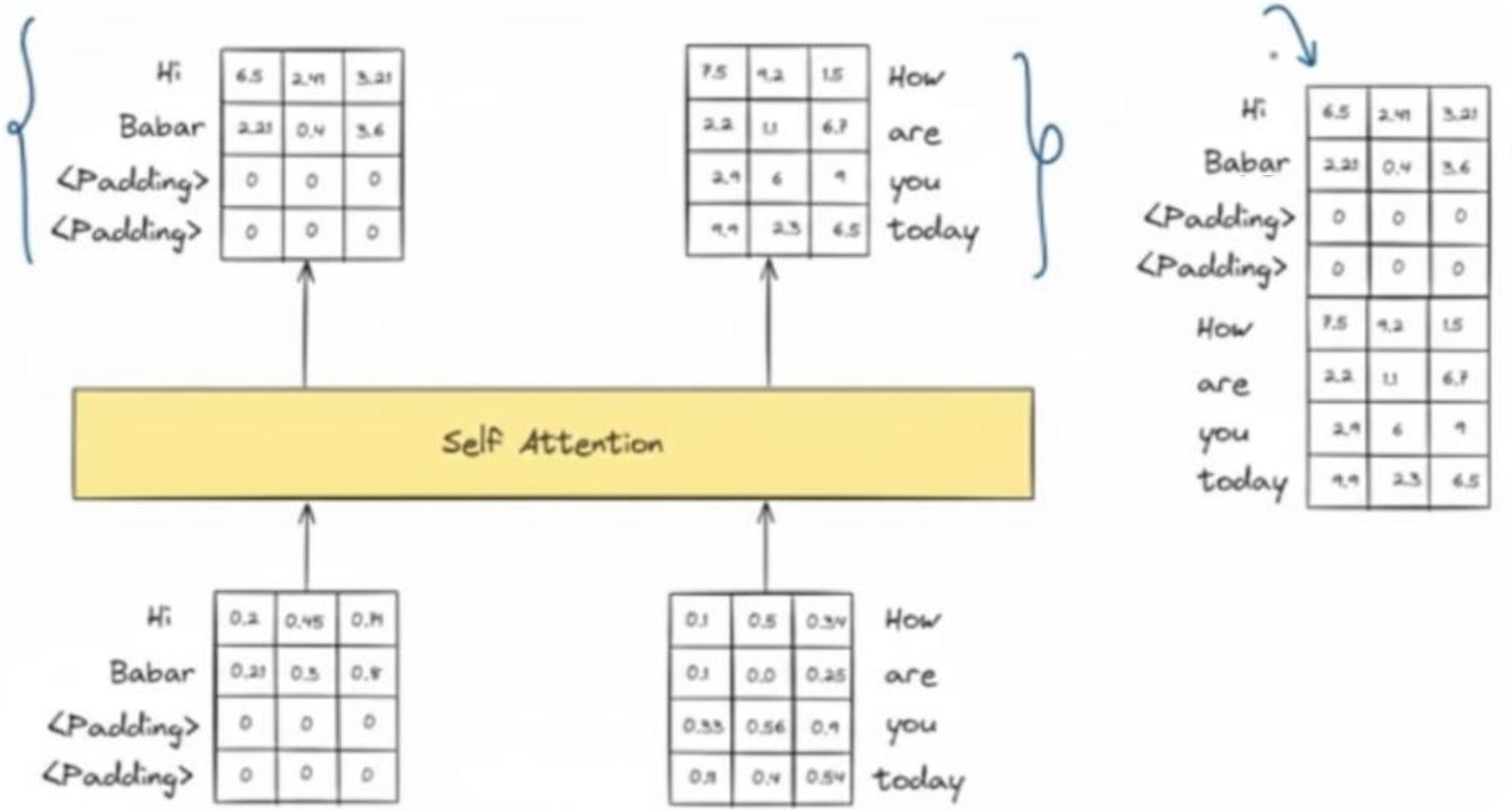
f_1	f_2	z_1	z_2	z_3	
2	3	7	5	4	μ_1, σ_1
1	1	2	3	4	μ_2, σ_2
5	4	1	2	3	
6	1	7	5	6	
7	1	3	3	4	μ_5, σ_5

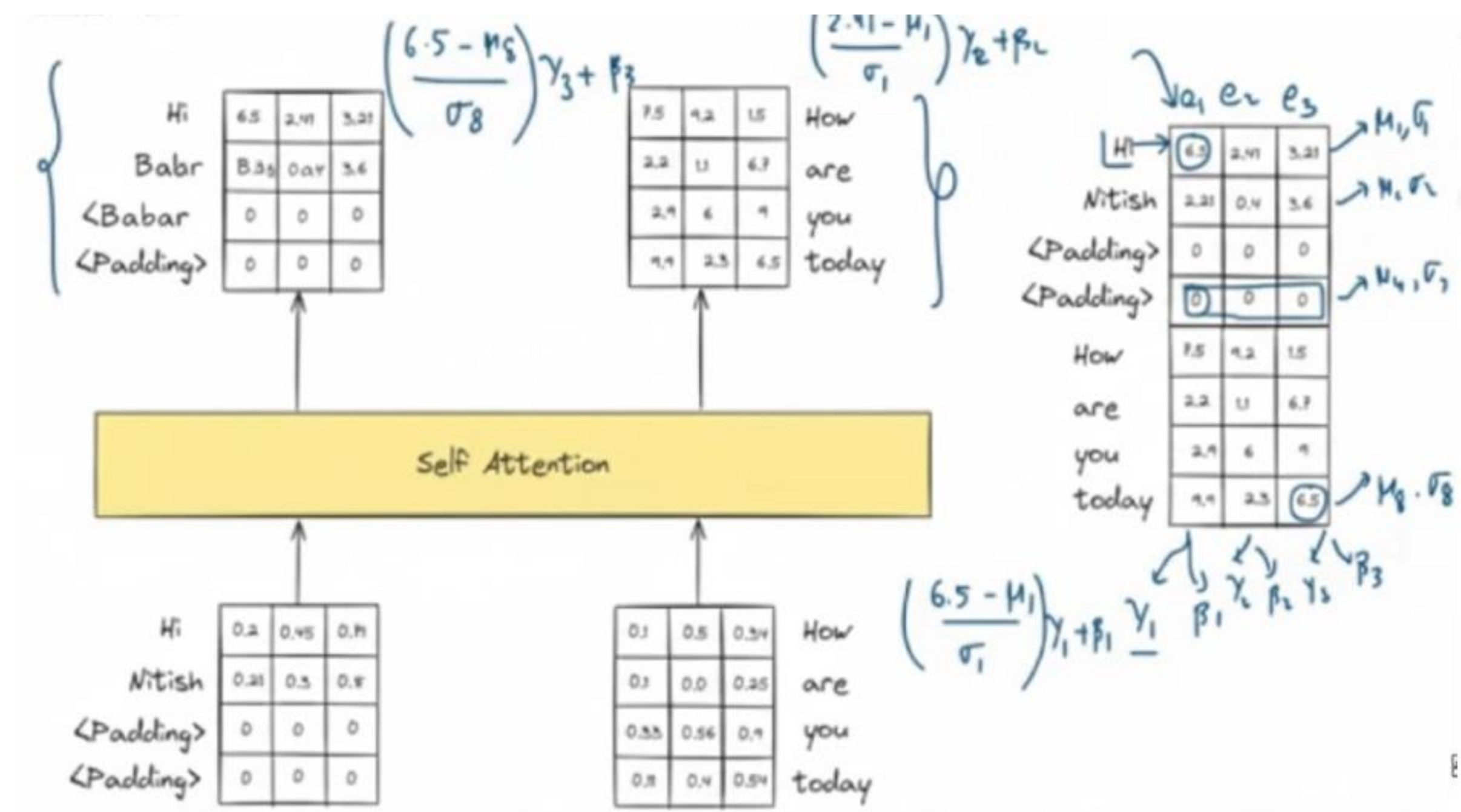
$\left(\frac{2-\mu_2}{\sigma_2}\right)\gamma_1+\beta_1$

$\left(\frac{4-\mu_1}{\sigma_1}\right)\gamma_3+\beta_3$

$\left(\frac{5-\mu_1}{\sigma_1}\right)\gamma_2+\beta_2$

$\frac{7-\mu_1}{\sigma_1} = \boxed{0.3\,\gamma_1+\beta_1}$


Hi
Babar
<Padding>
<Padding>
How
are
you
today
Self Attention
Hi
Babar
<Padding>
<Padding>
How
are
you
today
Hi
Babar
<Padding>
<Padding>
How
are
you
today


Hi
Babr
<Babar
<Padding>
How
are
you
today
Self Attention
Hi
Nitish
<Padding>
<Padding>
How
are
you
today
Nitish
<Padding>
<Padding>
How
are
you
today