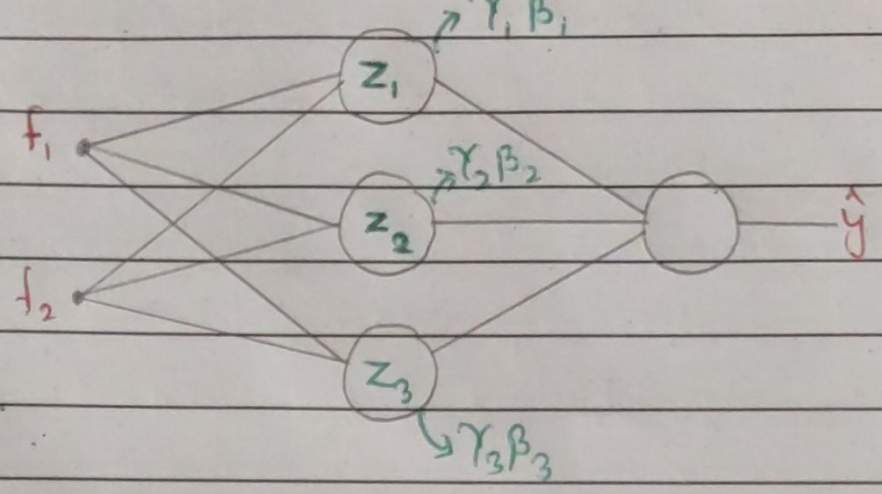
RAISONI GROUP
— a vision beyond —

- **How Batch normalization works**



		$\gamma_1\beta_1$	$\gamma_2\beta_2$	$\gamma_3\beta_3$
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
		μ_1 σ_1	μ_2 σ_2	μ_3 σ_3

Consider following setup with 2 features passing value in feed forward NN with 3 nodes where each node has it's pre-activation value z_1, z_2, z_3 & parameters $\gamma_1\beta_1, \gamma_2\beta_2, \gamma_3\beta_3$ respectively.

In batch normalization, mean (μ) & standard deviation (σ) of each column (z_1, z_2 & z_3) is ~~calle~~ calculated across the batch (columnwise) for normalization & readjusted with their respective γ & β value ($\gamma_1\beta_1, \gamma_2\beta_2, \gamma_3\beta_3$).

Batch Normalization in Sequential Data

Consider 4 sentences -

"Hi Probal"

"How are you today"

"I am good"

"How are you"

Assume each word to have a 3 dimensional embedding. We are passing these sentences in a batch of 2 to the self attention mechanism.

⇒

0.2	0.45	0.71		0.21	0.3	0.8
Hi				Probal		

* Values are hypothetical

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		

Since text lengths are unequal we will zero padding embedding vectors in the first sentence.

⇒

0.2	0.45	0.71		0.21	0.3	0.8		0	0	0		0	0	0
Hi				Probal				<Pad>				<Pad>		

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		

Now these are fed into the self attention mechanism to get contextual embeddings.

Hi	6.5	2.41	3.21		7.5	9.2	1.5	How
Probal	2.21	0.4	3.6		2.2	1.1	6.7	are
<Pad>	0	0	0		2.9	6	9	you
<Pad>	0	0	0		9.9	2.3	6.5	today

Self Attention

Hi	0.2	0.45	0.71		0.1	0.5	0.34	How
Probal	0.21	0.3	0.8		0.1	0.0	0.25	are
<Pad>	0	0	0		0.33	0.56	0.9	you
<Pad>	0	0	0		0.11	0.4	0.54	today

These contextual embeddings have varying value so now they are

normalized using batch normalization.

The two matrices are stacked up vertically -

$\gamma_1\beta_1$ $\gamma_2\beta_2$ $\gamma_3\beta_3$

⇒			
Hi	6.5	2.41	3.21
Probal	2.21	0.4	3.6
<Pad>	0	0	0
<Pad>	0	0	0
How	7.5	9.2	1.5
are	2.2	1.1	6.7
you	2.9	6	9
today	9.9	2.3	6.5
	μ_1	μ_2	μ_3
	σ_1	σ_2	σ_3

For each column/dimension ~~column~~, μ & σ are calculated ($\mu_1\sigma_1, \mu_2\sigma_2, \mu_3\sigma_3$).

The zero padding embedding vectors were added just to equal the length of texts. This unnecessary adding of padding vectors is resulting in not getting the true statistical representation of the embeddings (μ & σ)

The zero padding embedding vectors are affecting the true statistical representations of the text embeddings & that's why we don't use batch normalization.

Eg. for 1st dimension of embedding vector of word "How" -

$$\frac{7.5 - \mu_1}{\sigma_1} = 0.3 \longrightarrow \text{Hypothetical}$$

Column wise μ_1 & σ_1
Column wise γ_1 & β_1

Readjustment: $0.3\gamma_1 + \beta_1$