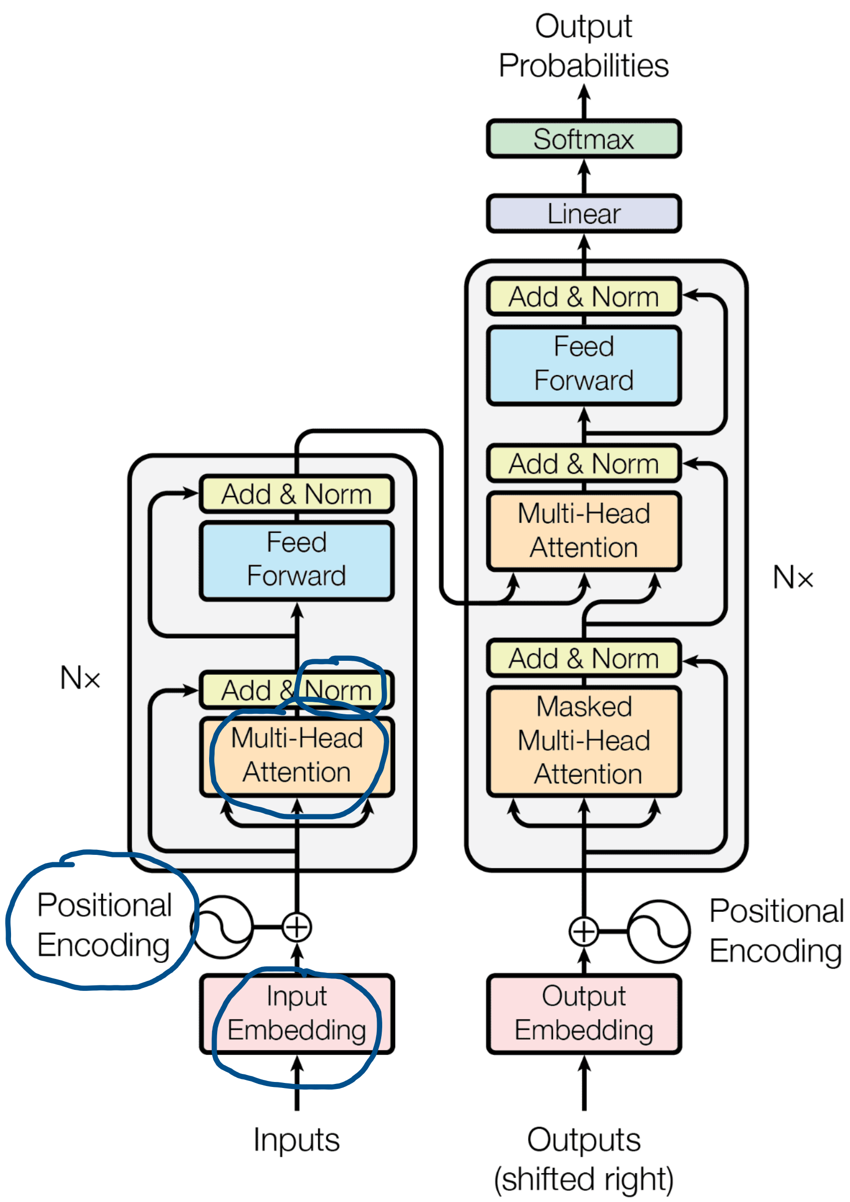
Agenda

07 June 2024 02:03

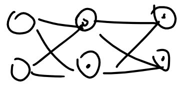


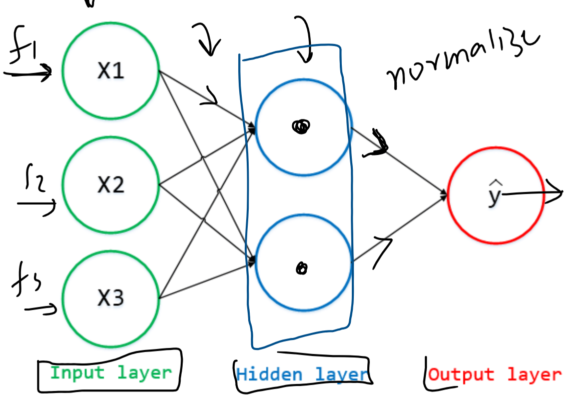
What is Normalization

05 June 2024 10:32

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.

What do we normalize?





f1
27
23
41

min-max

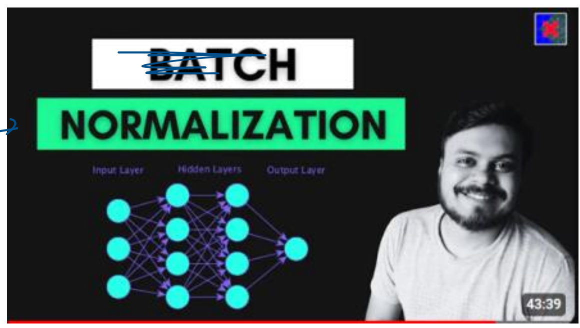
$$\frac{x_i - \mu}{\sigma}$$

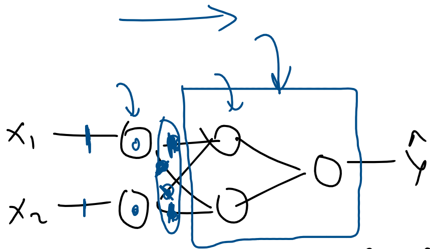
$\mu = 0$
$\sigma = 1$

f1	f2	f3
–	–	–
–	–	–
–	–	–

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 backpropagation.
- 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.
- Regularization Effect:
  - Some normalization techniques, like batch normalization, introduce a slight regularizing effect by adding noise to the mini-batches during training. This can help to reduce overfitting.





input data → dist

covariate shift

internal

red roses

pred
↓
yellow white

Activation

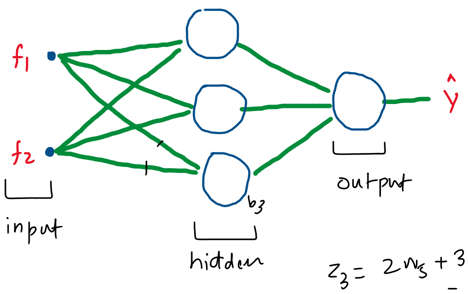
input

Batch Norm → Layer

+ Batch Norm $\rightarrow$ $\boxed{\text{Layer Norm}}$

Batch Norm(Revision)

05 June 2024 10:39



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

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

$$z_1 = \underbrace{2w_1 + 3w_2 + b_1} = 7$$

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

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

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

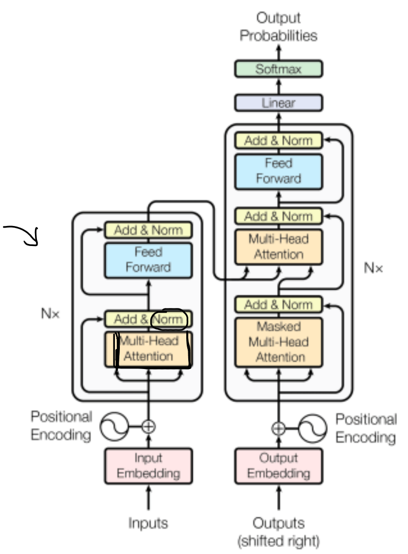
$$\frac{5-\mu_2}{\sigma_2} = -0.21\,\gamma_2 + \beta_2 = -0.21$$

$$\frac{4-\mu_3}{\sigma_3} = 0.12\,\gamma_3 + \beta_3 = 0.12$$

Why don't we use Batch Norm in Transformers?

05 June 2024 10:40

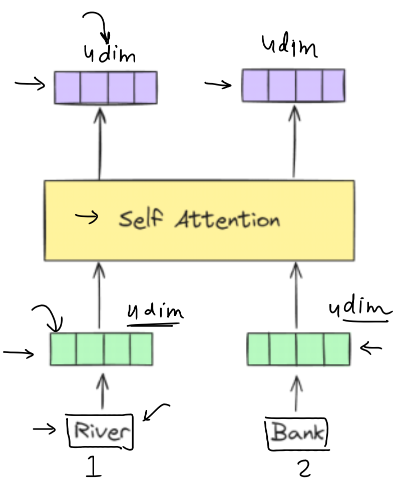
↳ self attention



→ Batch Norm

↑

Self - attention



Batches

dim=3

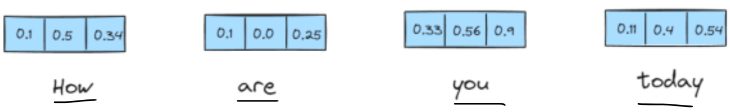
	Review	Sentiment
r1	Hi Nitish	1
r2	How are you today	0
r3	I am good	0
r4	You?	1

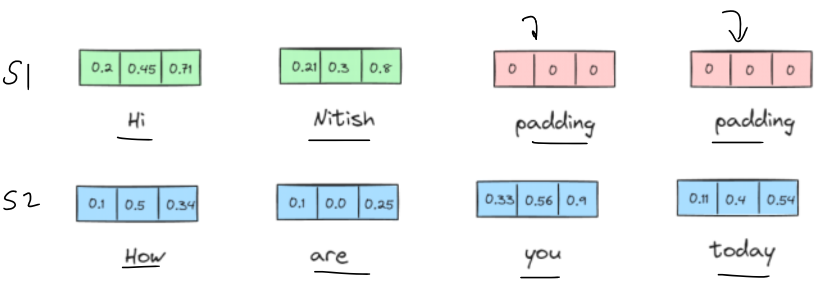
Embedding dimension - 3

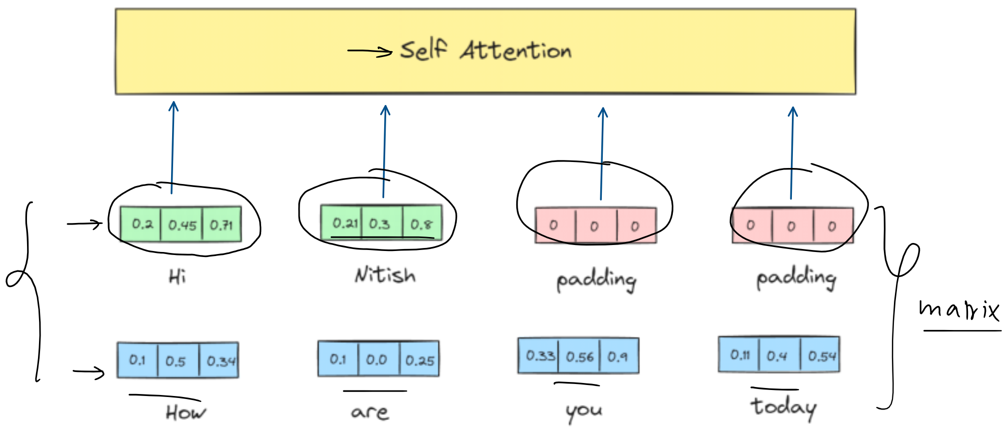
Batch Size - 2

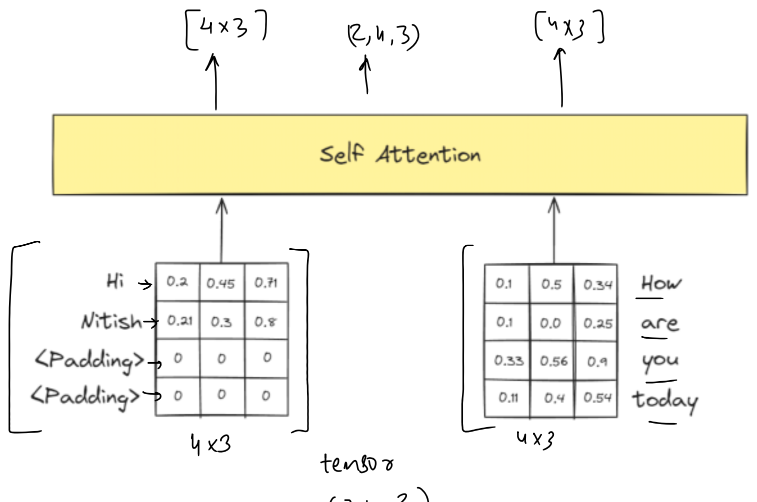
padding

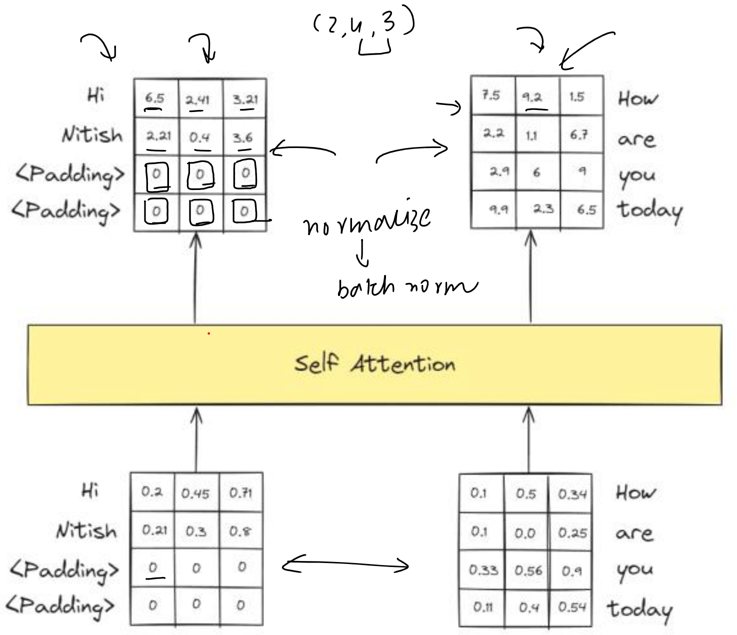
<pad> <pad>

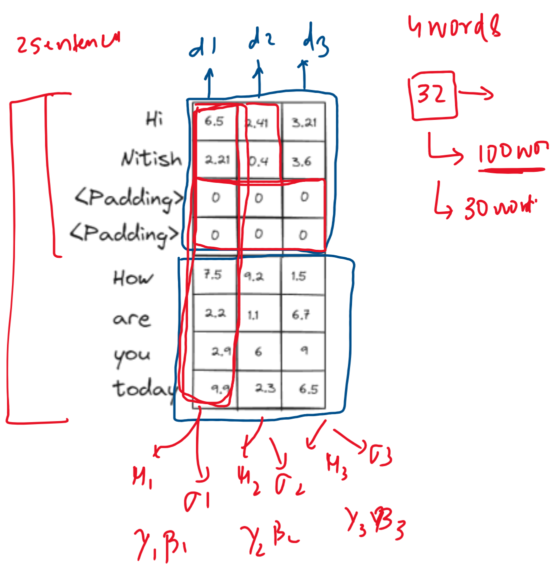





Self Attention
0.2 0.45 0.71
Hi
0.21 0.3 0.8
Nitish
0 0 0
padding
0 0 0
padding
0.1 0.5 0.34
How
0.1 0.0 0.25
are
0.33 0.56 0.9
you
0.11 0.4 0.54
today
matrix

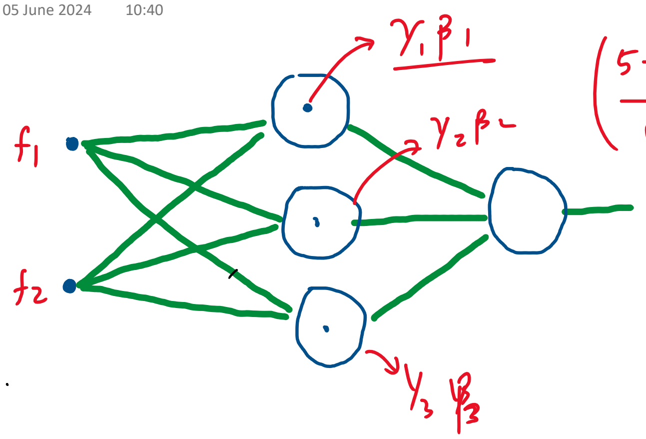

[4x3]
(2,4,3)
[4x3]
Self Attention
Hi
Nitish
<Padding>
<Padding>
4x3
How
are
you
today
4x3
tensor
(2,4,3)


Hi
Nitish
<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
normalize
batch norm
Self Attention
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
0.11 0.4 0.54


2 sentence
d1 d2 d3
4 words
32
100 words
30 words
Hi
Nitish
<Padding>
<Padding>
How
are
you
today
μ1 σ1
μ2 σ2
μ3 σ3
γ1 β1
γ2 β2
γ3 β3

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

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

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

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

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

Layer Norm in Transformers

05 June 2024 10:40