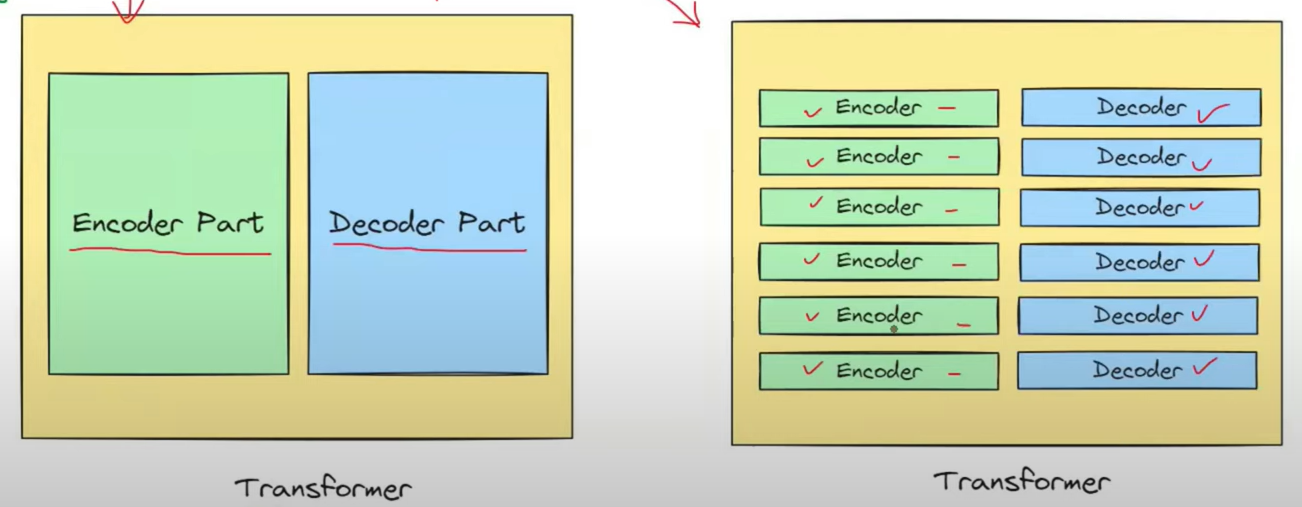
## horizontal line

Transformer (Encoder)

12.06.2025

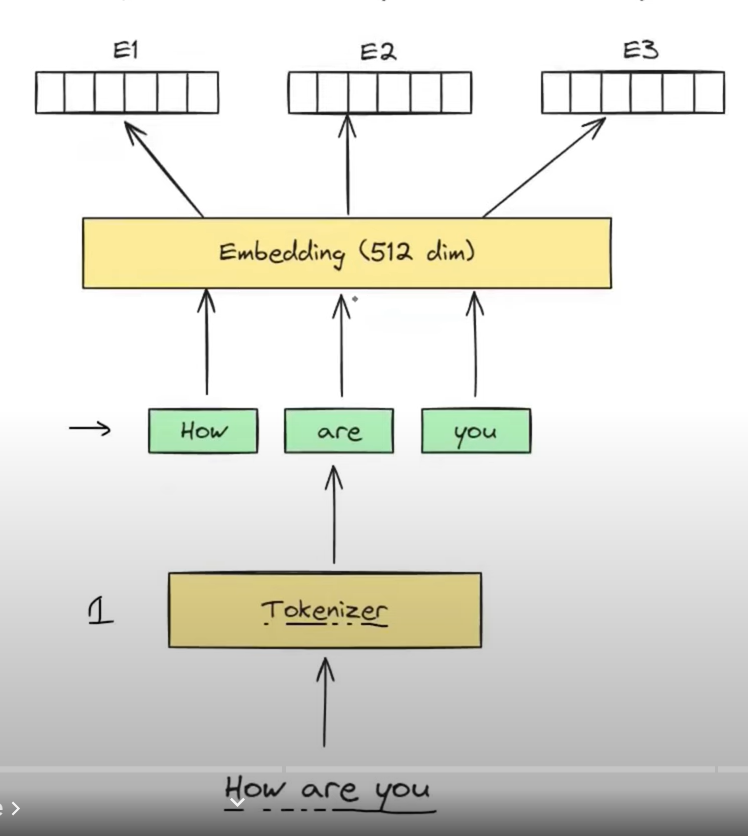
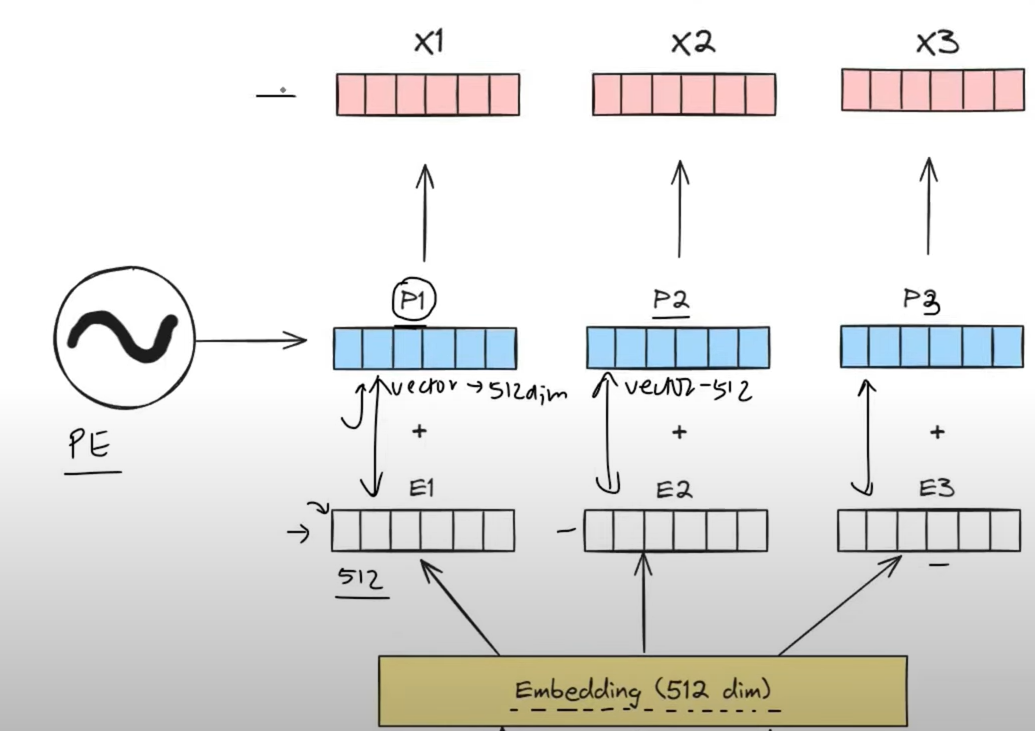
# Transformer Encoder

Transformer is composed of Encoder and Decoder blocks where we use 6 Encoder blocks . It is just experimentally proven that 6 blocks would give better results.

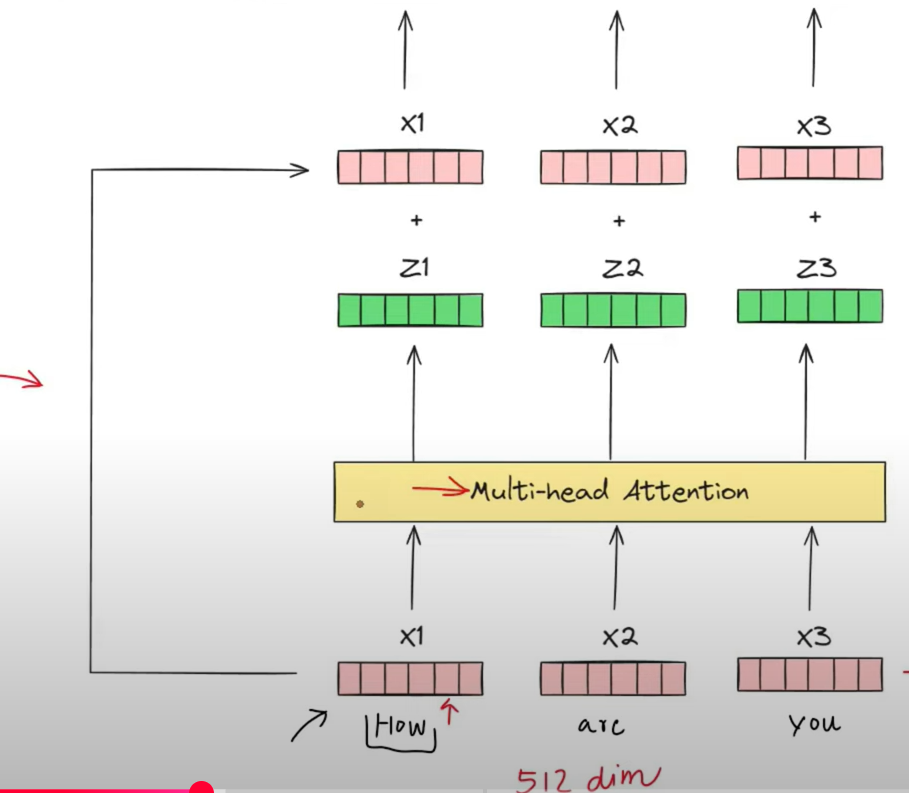
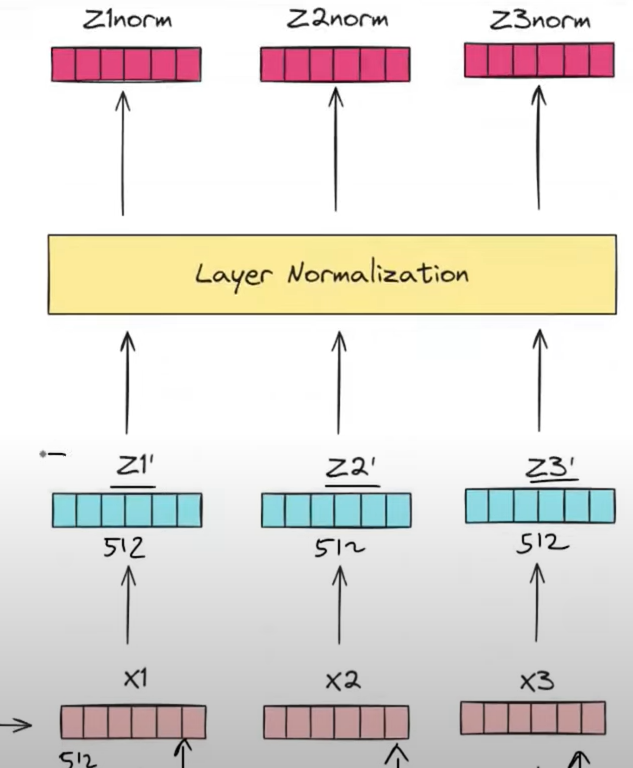
We will study one encoder block because all 6 are identical in architecture. Encoder block contains a Self Attention and a FFNN.

# Input Block

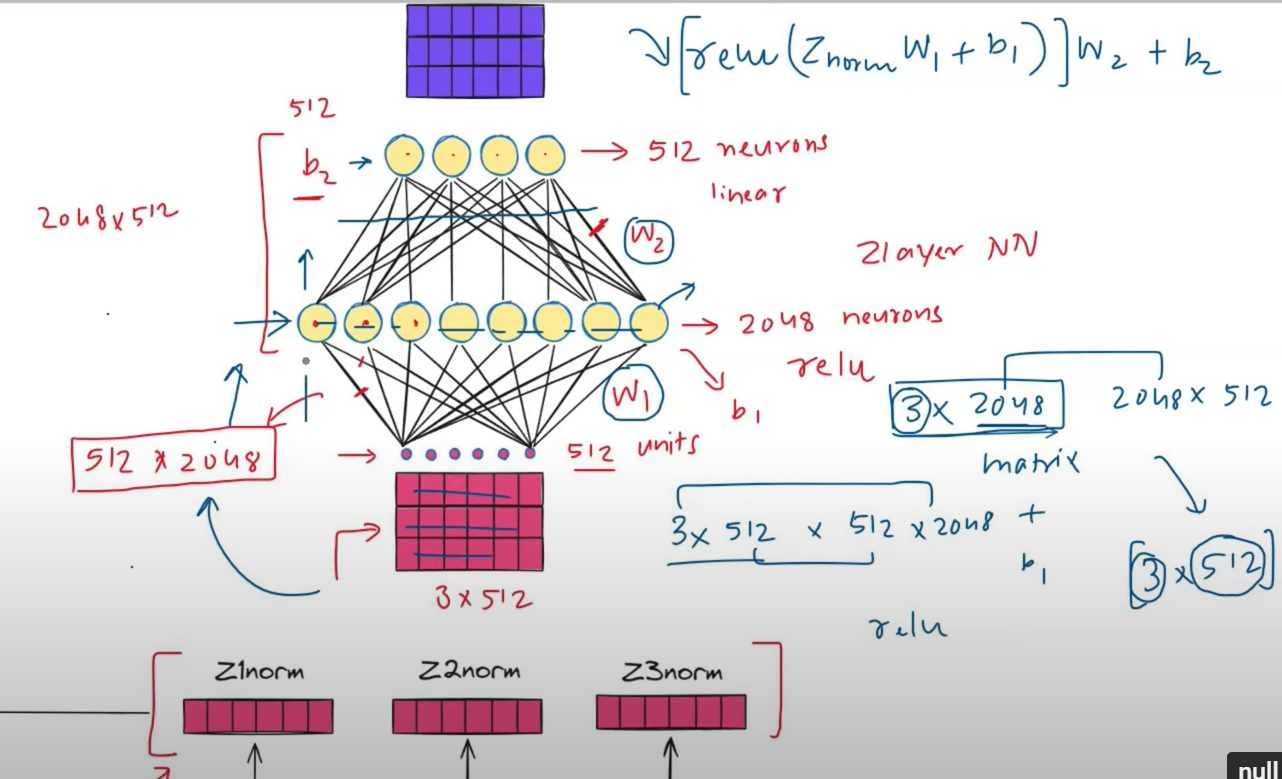
Input Blocks contain 3 processes i.e. tokenization , Embeddings and positional Encoding.



## Multi Head Attention in Encoder



## Feed Forward Neural Network in Encoder



We have 3 rows of 512 dim embeddings.

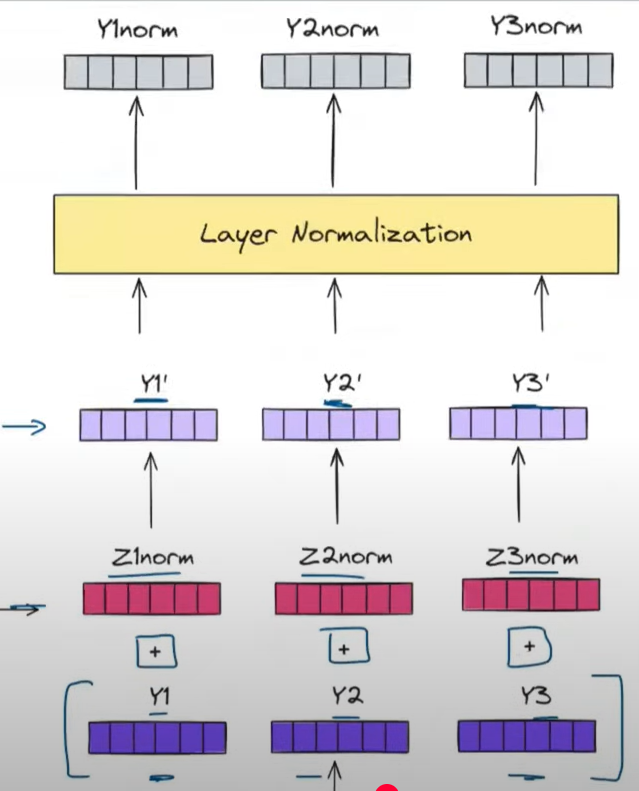
Combinedly we send them to the neural network i.e. 512 input units.

First layer has 2048 neurons , having W1 [512 X 2048] and inputs come as [3 X 512].

(3 X 512) (512 X 2048) = (3 X 2048) increasing dimensionality.

Second layer has 512 nodes , having W2 (2048 X 512) and inputs to it will be (3 X 2048).

(3 X 2048) (2048 X 512) = (3 X 512) again reducing dimensionality. This helps to capture non linearity. We get the same outputs as input (3 X 512).

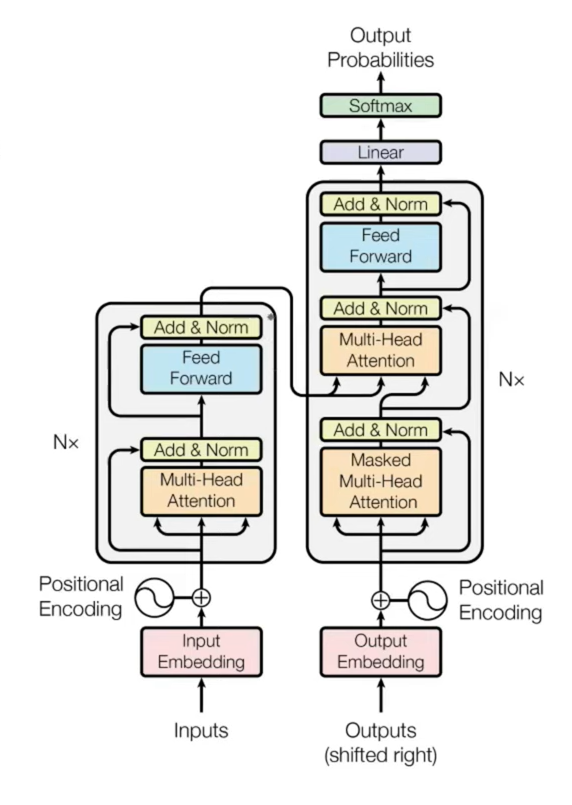


After Self Attention Add and norm applied.

Also same after FFNN.

Why do we use residual Connections ??

1. Stable training - In deep RNNs , like Resnet , due to various layers in Backpropagation it causes Vanishing Gradient Problem.
2. Every time an operation does not cause a positive effect , sometimes to maintain originality it is required.



Why use a FFNN ??

To capture non linearity.

Why use 6 encoder blocks ?

Human language is complex and it requires high representation power which can be attained by multiple blocks to understand language completely.