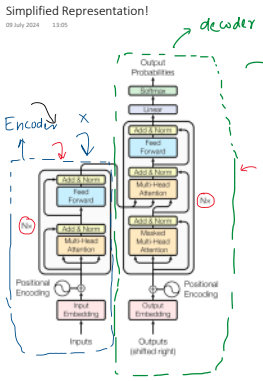
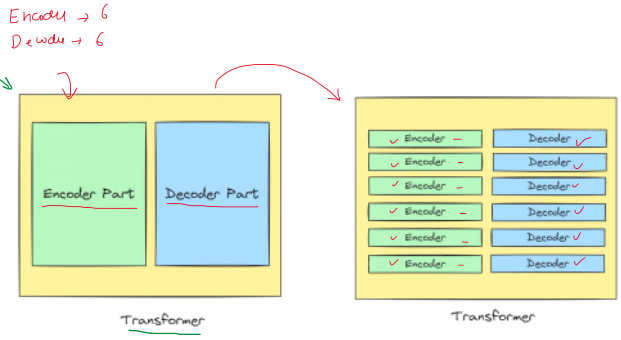
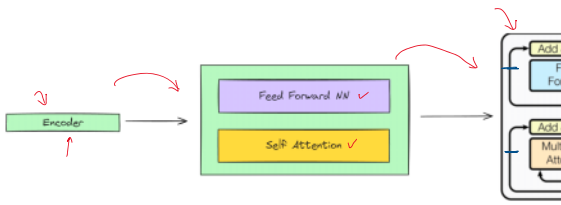
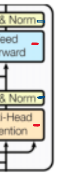
Simplified Representation!

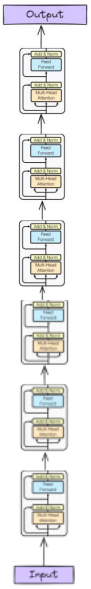
09 July 2024 13:05





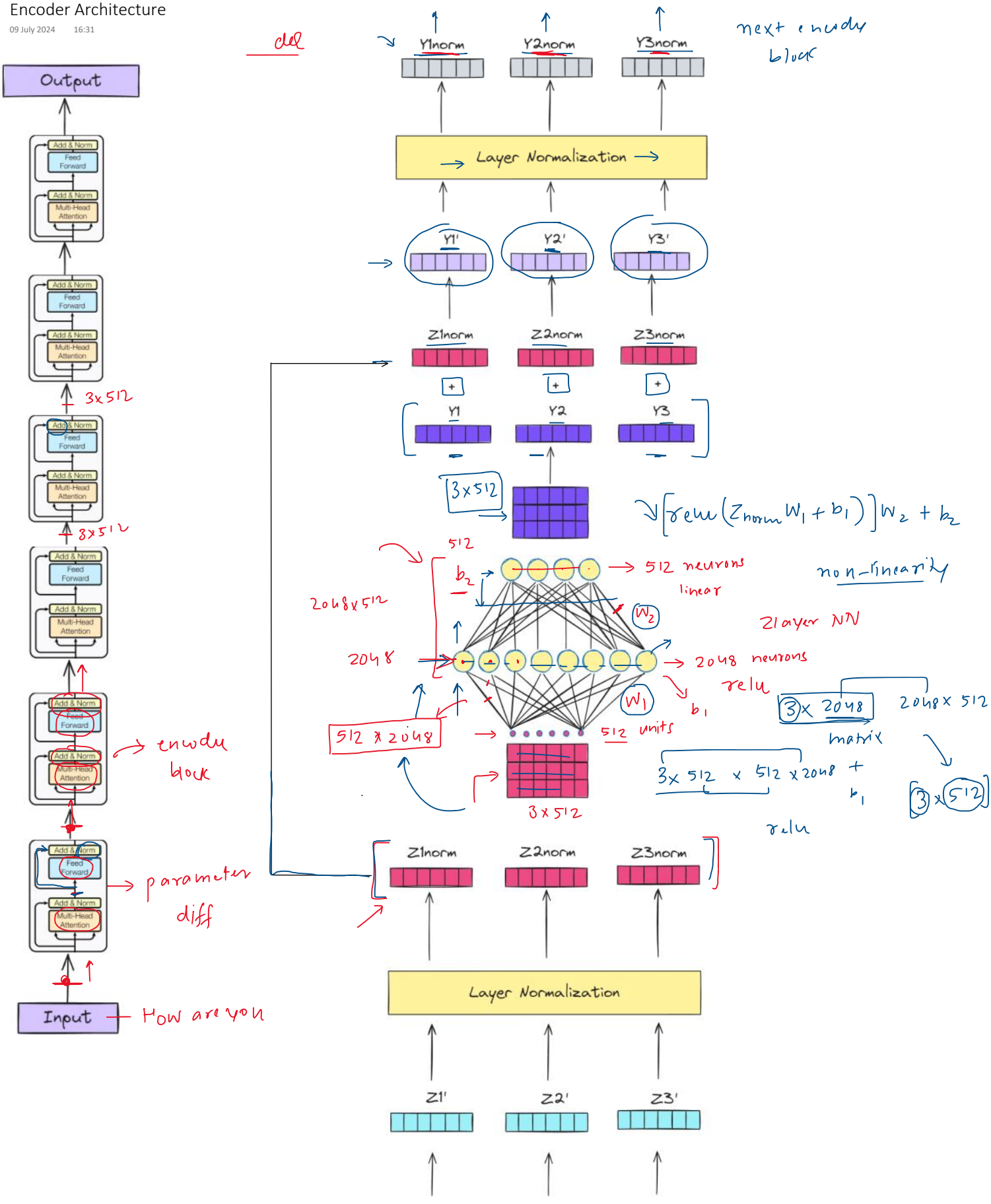


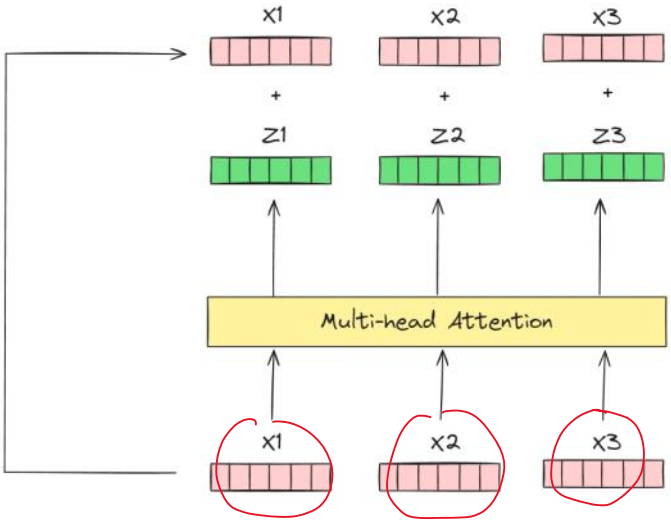

& Norm
eed
ward
& Norm
i-Head
ention

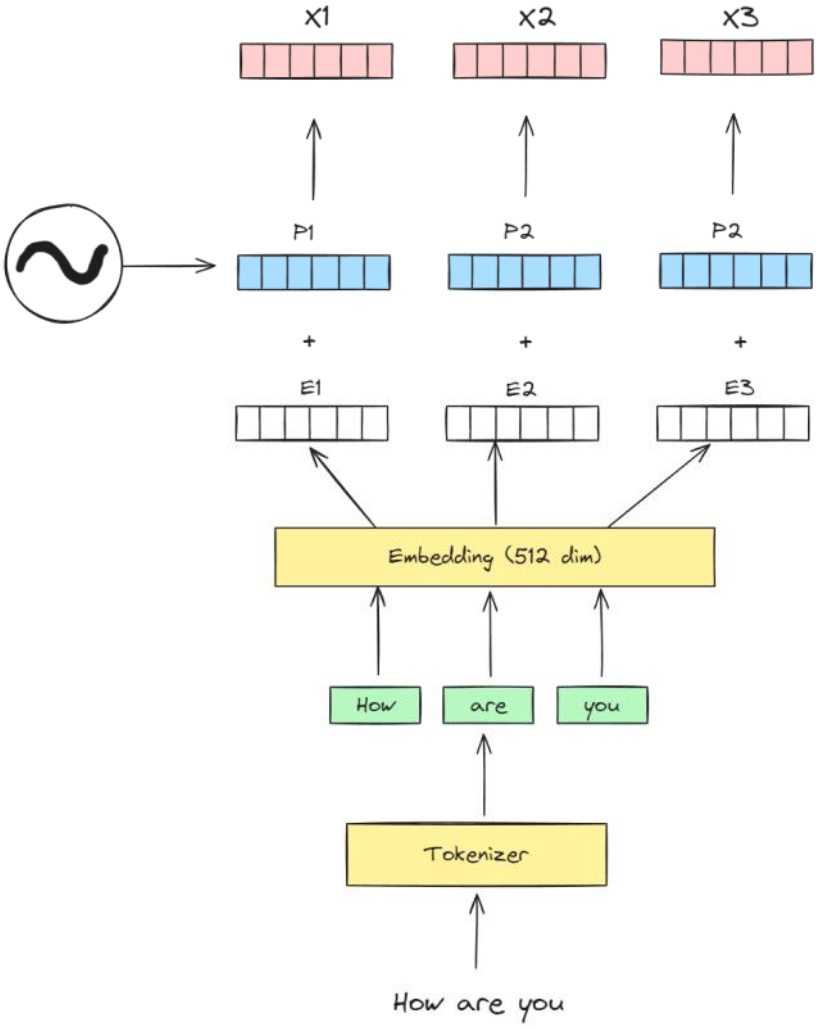

Output
Input

Encoder Architecture

09 July 2024 16:31




X1
X2
X3
+
+
+
Z1
Z2
Z3
Multi-head Attention
X1
X2
X3


X1
X2
X3
P1
P2
P2
+
+
+
E1
E2
E3
Embedding (512 dim)
How
are
you
Tokenizer
How are you

Some questions

09 July 2024 20:48

1. Why use residual connections? → Kaggle → comment
2. Why use a FFNN?
3. Why use 6 encoder blocks?

alternate

gradient flow

relu

non-linearity

scratch pytorch

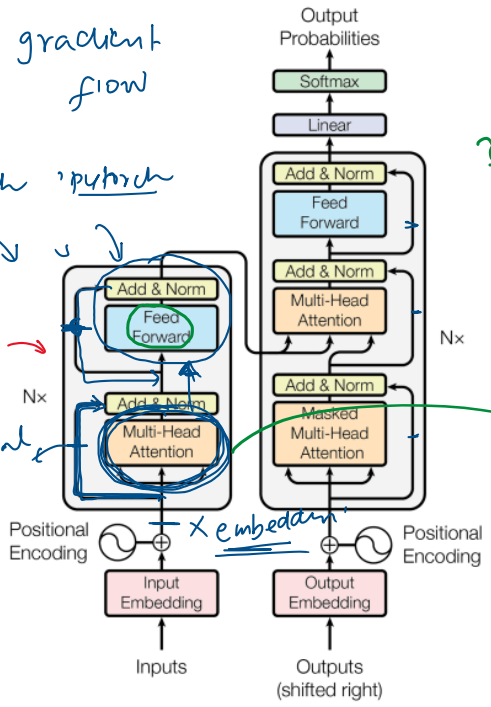
language → complex

1) stable training — deep NN

ResNet

vanishing Grad

residual



linear

× embedding

2)