TASK 01
Title: Report on Encoder Decoder Based Transformers

The following are the main results and key findings:

Part I (Preparation): I preprocess a toy dataset that consists of input arithmetic expression and an output result of the expression.

```
Dictionary created successfully!

preprocess input token error 1:  0.0
preprocess input token error 2:  0.0
preprocess input token error 3:  0.0
preprocess input token error 4:  0.0


preprocess output token error 1:  0.0
preprocess output token error 2:  0.0
preprocess output token error 3:  0.0
preprocess output token error 4:  0.0
```

Part II (Implement Transformer blocks): I implement the building blocks of a Transformer. It will consist of the following blocks:

1. MultiHeadAttention
2. FeedForward
3. LayerNorm
4. Encoder Block
5. Decoder Block

```
sacled_dot_product_two_loop_single error:  5.204997336388729e-06
```

```
scaled_dot_product_two_loop_batch error:  4.020571992067902e-06
```

```
scaled_dot_product_no_loop_batch error:  4.020571992067902e-06
```

Time Complexity:

```
%timeit -n 5 -r 2  y = scaled_dot_product_no_loop_batch(query, key, value)
```

```
543 ms ± 90 ms per loop (mean ± std. dev. of 2 runs, 5 loops each)
```

```
%timeit -n 5 -r 2  y = scaled_dot_product_no_loop_batch(query, key, value)
```

```
1.67 s ± 81.7 ms per loop (mean ± std. dev. of 2 runs, 5 loops each)
```

```
SelfAttention error:  5.567700453666357e-07
SelfAttention error:  0.4621903063309185
```

```
MultiHeadAttention error:  6.577880512731245e-07
MultiHeadAttention error:  0.844447827769596
```

```
LayerNormalization error:  0.07179646242633864
LayerNormalization grad error:  0.07179690055792502
```

```
FeedForwardBlock error:  2.1976866936034156e-07
FeedForwardBlock error:  1.0
```

```
EncoderBlock error 1:  0.5058199798801631
EncoderBlock error 2:  6.26799449492745e-07
```

```
get_subsequent_mask error:  0.0
```

```
scaled_dot_product_no_loop_batch error:  2.8390648478191238e-06
```

```
DecoderBlock error:  0.42327517346581917
DecoderBlock error:  0.4058814012411442
```

Part III (Data Loading): I use the preprocessing functions in part I and the positional encoding module to construct the Dataloader.

```
position_encoding_simple error:  0.0
position_encoding_simple error:  0.0
```

```
position_encoding error:  0.9947700500488281
position_encoding error:  0.4524955749511719
```

Part IV (Train a model): In the last part I fit the implemented Transformer model to the toy dataset.

Overfitted model on small data:

```
[epoch: 175] [loss:  0.0177 ] val_loss: [val_loss  0.0088 ]
[epoch: 176] [loss:  0.0197 ] val_loss: [val_loss  0.0087 ]
[epoch: 177] [loss:  0.0279 ] val_loss: [val_loss  0.0086 ]
[epoch: 178] [loss:  0.0167 ] val_loss: [val_loss  0.0085 ]
[epoch: 179] [loss:  0.0185 ] val_loss: [val_loss  0.0084 ]
[epoch: 180] [loss:  0.0353 ] val_loss: [val_loss  0.0083 ]
[epoch: 181] [loss:  0.0184 ] val_loss: [val_loss  0.0083 ]
[epoch: 182] [loss:  0.0272 ] val_loss: [val_loss  0.0082 ]
[epoch: 183] [loss:  0.0194 ] val_loss: [val_loss  0.0081 ]
[epoch: 184] [loss:  0.0195 ] val_loss: [val_loss  0.0081 ]
[epoch: 185] [loss:  0.0221 ] val_loss: [val_loss  0.0080 ]
[epoch: 186] [loss:  0.0146 ] val_loss: [val_loss  0.0079 ]
[epoch: 187] [loss:  0.0220 ] val_loss: [val_loss  0.0079 ]
[epoch: 188] [loss:  0.0163 ] val_loss: [val_loss  0.0078 ]
[epoch: 189] [loss:  0.0229 ] val_loss: [val_loss  0.0078 ]
[epoch: 190] [loss:  0.0179 ] val_loss: [val_loss  0.0077 ]
[epoch: 191] [loss:  0.0221 ] val_loss: [val_loss  0.0076 ]
[epoch: 192] [loss:  0.0150 ] val_loss: [val_loss  0.0075 ]
[epoch: 193] [loss:  0.0180 ] val_loss: [val_loss  0.0075 ]
[epoch: 194] [loss:  0.0166 ] val_loss: [val_loss  0.0074 ]
[epoch: 195] [loss:  0.0237 ] val_loss: [val_loss  0.0073 ]
[epoch: 196] [loss:  0.0145 ] val_loss: [val_loss  0.0073 ]
[epoch: 197] [loss:  0.0162 ] val_loss: [val_loss  0.0072 ]
[epoch: 198] [loss:  0.0136 ] val_loss: [val_loss  0.0071 ]
[epoch: 199] [loss:  0.0251 ] val_loss: [val_loss  0.0071 ]
[epoch: 200] [loss:  0.0166 ] val_loss: [val_loss  0.0070 ]
```

```
Overfitted accuracy:  1.0000
```

Experiments on whole data:

Experiment: 01

Hyperparameters:

- inp_seq_len = 9
- out_seq_len = 5
- learning_rate = 1e-2
- Batch size: 256

- Number of epochs: 10
- Loss: CrossEntropy
- Drop out = 0.2

Results: Training loss is 2.3461, and Validation loss is 2.3397

```
Training started...
[epoch: 1] [loss:  2.5651 ] val_loss: [val_loss  2.3527 ]
[epoch: 2] [loss:  2.3765 ] val_loss: [val_loss  2.3449 ]
[epoch: 3] [loss:  2.3644 ] val_loss: [val_loss  2.3451 ]
[epoch: 4] [loss:  2.3569 ] val_loss: [val_loss  2.3455 ]
[epoch: 5] [loss:  2.3550 ] val_loss: [val_loss  2.3470 ]
[epoch: 6] [loss:  2.3519 ] val_loss: [val_loss  2.3461 ]
[epoch: 7] [loss:  2.3505 ] val_loss: [val_loss  2.3425 ]
[epoch: 8] [loss:  2.3483 ] val_loss: [val_loss  2.3440 ]
[epoch: 9] [loss:  2.3470 ] val_loss: [val_loss  2.3422 ]
[epoch: 10] [loss:  2.3461 ] val_loss: [val_loss  2.3397 ]


Final Model accuracy:  0.2500
```

Experiment: 02

Hyperparameters:

- inp_seq_len = 9
- out_seq_len = 5
- learning_rate = 1e-3
- Batch size: 256
- Number of epochs: 10
- Loss: CrossEntropy
- Drop out = 0.2

Results: Training loss is 0.0228, and Validation loss is 0.0088

```
Training started...
[epoch: 1] [loss:  2.2302 ] val_loss: [val_loss  1.5272 ]
[epoch: 2] [loss:  1.3362 ] val_loss: [val_loss  0.8738 ]
[epoch: 3] [loss:  0.8233 ] val_loss: [val_loss  0.5102 ]
[epoch: 4] [loss:  0.4649 ] val_loss: [val_loss  0.1985 ]
[epoch: 5] [loss:  0.2152 ] val_loss: [val_loss  0.0689 ]
[epoch: 6] [loss:  0.1128 ] val_loss: [val_loss  0.0372 ]
[epoch: 7] [loss:  0.0691 ] val_loss: [val_loss  0.0218 ]
[epoch: 8] [loss:  0.0429 ] val_loss: [val_loss  0.0147 ]
[epoch: 9] [loss:  0.0297 ] val_loss: [val_loss  0.0111 ]
[epoch: 10] [loss:  0.0228 ] val_loss: [val_loss  0.0088 ]

Final Model accuracy:  1.0000
```

Experiment: 03

Hyperparameters:

- inp_seq_len = 9
- out_seq_len = 5
- learning_rate = 1e-4
- Batch size: 256
- Number of epochs: 10
- Loss: CrossEntropy
- Drop out = 0.2

Results: Training loss is 0.9837, and Validation loss is 0.7957

```
Training started...
[epoch: 1] [loss:  2.7242 ] val_loss: [val_loss  2.3020 ]
[epoch: 2] [loss:  2.3144 ] val_loss: [val_loss  2.0976 ]
[epoch: 3] [loss:  2.0983 ] val_loss: [val_loss  1.8285 ]
[epoch: 4] [loss:  1.8713 ] val_loss: [val_loss  1.5797 ]
[epoch: 5] [loss:  1.6726 ] val_loss: [val_loss  1.3873 ]
[epoch: 6] [loss:  1.4941 ] val_loss: [val_loss  1.2258 ]
[epoch: 7] [loss:  1.3288 ] val_loss: [val_loss  1.0867 ]
[epoch: 8] [loss:  1.1948 ] val_loss: [val_loss  0.9746 ]
[epoch: 9] [loss:  1.0851 ] val_loss: [val_loss  0.8754 ]
[epoch: 10] [loss:  0.9837 ] val_loss: [val_loss  0.7957 ]


Final Model accuracy:  0.7725
```

Experiment: 04

Hyperparameters:

- inp_seq_len = 9
- out_seq_len = 5
- learning_rate = 1e-5
- Batch size: 256
- Number of epochs: 10
- Loss: CrossEntropy
- Drop out = 0.2

Results: Training loss is 2.4652, and Validation loss is 2.3359

```
Training started...
[epoch: 1] [loss:  3.0951 ] val_loss: [val_loss  2.9265 ]
[epoch: 2] [loss:  2.9767 ] val_loss: [val_loss  2.7755 ]
[epoch: 3] [loss:  2.8715 ] val_loss: [val_loss  2.6636 ]
[epoch: 4] [loss:  2.7837 ] val_loss: [val_loss  2.5810 ]
[epoch: 5] [loss:  2.7135 ] val_loss: [val_loss  2.5189 ]
[epoch: 6] [loss:  2.6565 ] val_loss: [val_loss  2.4714 ]
[epoch: 7] [loss:  2.5983 ] val_loss: [val_loss  2.4326 ]
[epoch: 8] [loss:  2.5504 ] val_loss: [val_loss  2.3991 ]
[epoch: 9] [loss:  2.5099 ] val_loss: [val_loss  2.3670 ]
[epoch: 10] [loss:  2.4652 ] val_loss: [val_loss  2.3359 ]
```

```
Final Model accuracy:  0.3535
```

Experiment: 05

Hyperparameters:

- inp_seq_len = 9
- out_seq_len = 5
- learning_rate = 1e-6
- Batch size: 256
- Number of epochs: 10
- Loss: CrossEntropy
- Drop out = 0.2

Results: Training loss is 3.0363, and Validation loss is 2.8027

```
Training started...
[epoch: 1] [loss:  3.1635 ] val_loss: [val_loss  2.9445 ]
[epoch: 2] [loss:  3.1444 ] val_loss: [val_loss  2.9260 ]
[epoch: 3] [loss:  3.1269 ] val_loss: [val_loss  2.9082 ]
[epoch: 4] [loss:  3.1179 ] val_loss: [val_loss  2.8912 ]
[epoch: 5] [loss:  3.1008 ] val_loss: [val_loss  2.8749 ]
[epoch: 6] [loss:  3.1005 ] val_loss: [val_loss  2.8593 ]
[epoch: 7] [loss:  3.0756 ] val_loss: [val_loss  2.8443 ]
[epoch: 8] [loss:  3.0608 ] val_loss: [val_loss  2.8300 ]
[epoch: 9] [loss:  3.0422 ] val_loss: [val_loss  2.8161 ]
[epoch: 10] [loss:  3.0363 ] val_loss: [val_loss  2.8027 ]
```

```
Final Model accuracy:  0.1924
```

Experiment: 06

Hyperparameters:

- inp_seq_len = 9

- out_seq_len = 5
- learning_rate = 1e-7
- Batch size: 256
- Number of epochs: 10
- Loss: CrossEntropy
- Drop out = 0.2

Results: Training loss is 3.4504, and Validation loss is 3.4122

```
Training started...
[epoch: 1] [loss:  3.4804 ] val_loss: [val_loss  3.4464 ]
[epoch: 2] [loss:  3.4829 ] val_loss: [val_loss  3.4426 ]
[epoch: 3] [loss:  3.4548 ] val_loss: [val_loss  3.4388 ]
[epoch: 4] [loss:  3.4658 ] val_loss: [val_loss  3.4350 ]
[epoch: 5] [loss:  3.4717 ] val_loss: [val_loss  3.4312 ]
[epoch: 6] [loss:  3.4678 ] val_loss: [val_loss  3.4273 ]
[epoch: 7] [loss:  3.4592 ] val_loss: [val_loss  3.4236 ]
[epoch: 8] [loss:  3.4603 ] val_loss: [val_loss  3.4198 ]
[epoch: 9] [loss:  3.4558 ] val_loss: [val_loss  3.4160 ]
[epoch: 10] [loss:  3.4504 ] val_loss: [val_loss  3.4122 ]
```

```
Final Model accuracy:  0.0508
```

Experiment: 07

Hyperparameters:

- inp_seq_len = 9
- out_seq_len = 5
- learning_rate = 1e-8
- Batch size: 256
- Number of epochs: 10
- Loss: CrossEntropy
- Drop out = 0.2

Results: Training loss is 3.2717, and Validation loss is 3.0475

```
Training started...
[epoch: 1] [loss:  3.2667 ] val_loss: [val_loss  3.0494 ]
[epoch: 2] [loss:  3.2585 ] val_loss: [val_loss  3.0492 ]
[epoch: 3] [loss:  3.2770 ] val_loss: [val_loss  3.0490 ]
[epoch: 4] [loss:  3.2668 ] val_loss: [val_loss  3.0488 ]
[epoch: 5] [loss:  3.2624 ] val_loss: [val_loss  3.0486 ]
[epoch: 6] [loss:  3.2583 ] val_loss: [val_loss  3.0483 ]
[epoch: 7] [loss:  3.2659 ] val_loss: [val_loss  3.0481 ]
[epoch: 8] [loss:  3.2639 ] val_loss: [val_loss  3.0479 ]
[epoch: 9] [loss:  3.2583 ] val_loss: [val_loss  3.0477 ]
[epoch: 10] [loss:  3.2717 ] val_loss: [val_loss  3.0475 ]
```

```
Final Model accuracy:  0.1016
```

Experiment: 08

Hyperparameters:

- inp_seq_len = 9
- out_seq_len = 5
- learning_rate = 1e-9
- Batch size: 256
- Number of epochs: 10
- Loss: CrossEntropy
- Drop out = 0.2

Results: Training loss is 3.4178, and Validation loss is 3.2929

Accuracy: 0.0742

```
Training started...
[epoch: 1] [loss:  3.4269 ] val_loss: [val_loss  3.2930 ]
[epoch: 2] [loss:  3.4228 ] val_loss: [val_loss  3.2929 ]
[epoch: 3] [loss:  3.4111 ] val_loss: [val_loss  3.2929 ]
[epoch: 4] [loss:  3.4154 ] val_loss: [val_loss  3.2929 ]
[epoch: 5] [loss:  3.4194 ] val_loss: [val_loss  3.2929 ]
[epoch: 6] [loss:  3.4338 ] val_loss: [val_loss  3.2929 ]
[epoch: 7] [loss:  3.4167 ] val_loss: [val_loss  3.2929 ]
[epoch: 8] [loss:  3.4215 ] val_loss: [val_loss  3.2929 ]
[epoch: 9] [loss:  3.4296 ] val_loss: [val_loss  3.2929 ]
[epoch: 10] [loss:  3.4178 ] val_loss: [val_loss  3.2929 ]
```

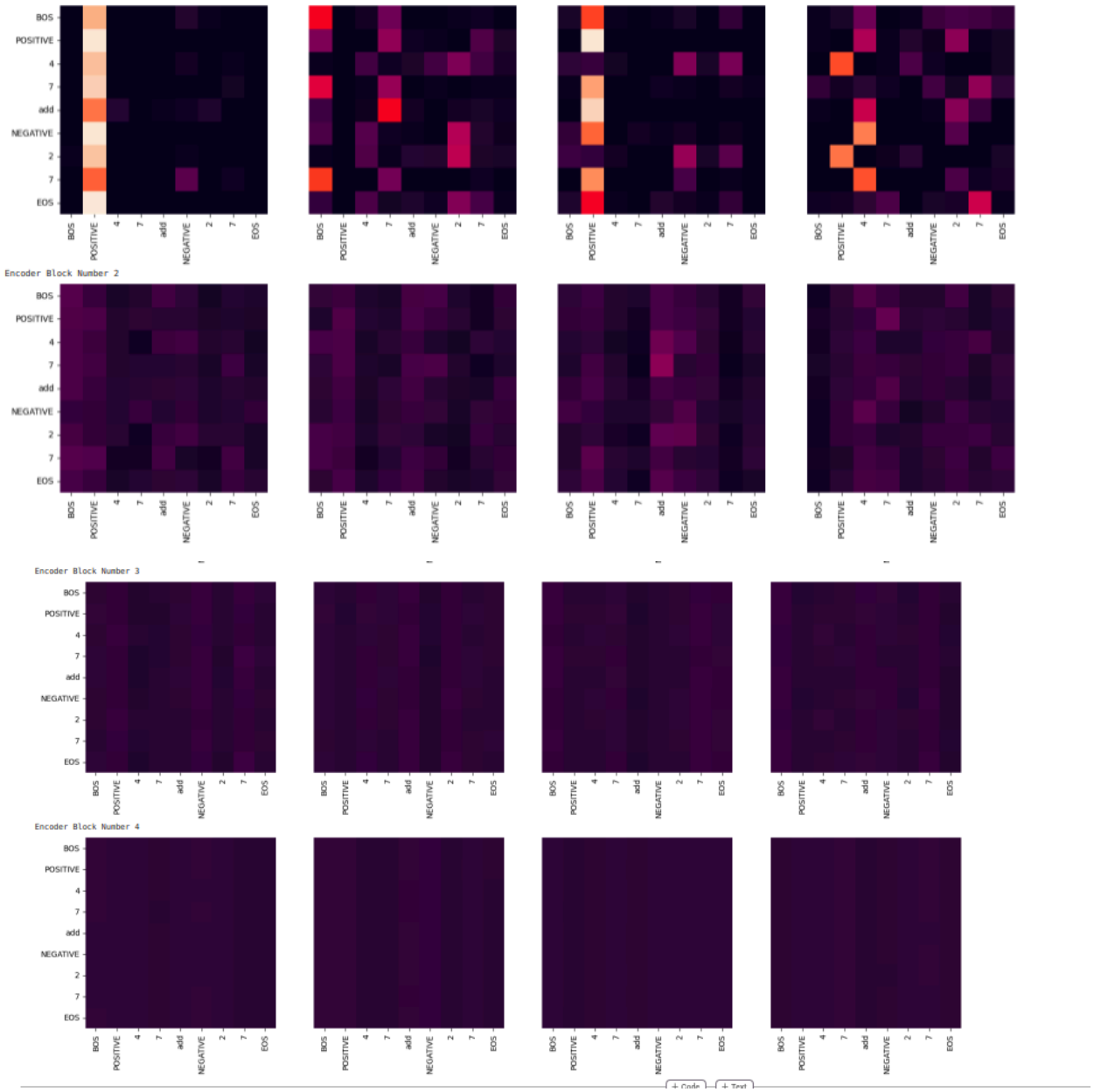
Learning rate	num_epochs	Training Loss	Validation Loss	Accuracy
1e-2	10	2.3461	2.3397	0.25
1e-3	10	0.0228	0.0088	1.00
1e-4	10	0.9837	0.7957	0.77
1e-5	10	2.4652	2.3359	0.35
1e-6	10	3.0363	2.8027	0.19
1e-7	10	3.4504	3.4122	0.050
1e-8	10	3.2717	3.0475	0.10
1e-9	10	3.4178	3.2929	0.074

```
Input sequence:
 BOS NEGATIVE 2 add NEGATIVE 4 EOS

Output Sequence:        BOS NEGATIVE 7 EOS
```

Own probing example:

```
custom_seq = "BOS POSITIVE 04 add NEGATIVE 11 EOS"

Output Sequence:        BOS NEGATIVE 7 EOS
```

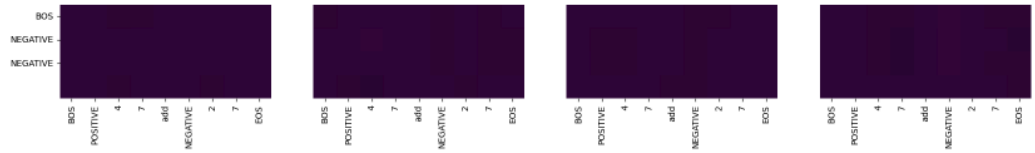
Visualizing attention weights:


Encoder Block Number 2
Encoder Block Number 3
Encoder Block Number 4
BOS
POSITIVE
4
7
add
NEGATIVE
2
7
EOS

Decoder:

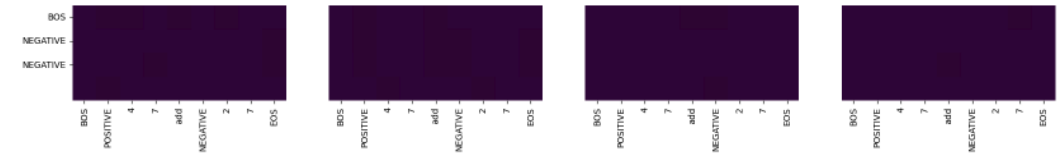
Decoder Block number 1
Decoder Self Attention 1

Decoder Cross attention 1

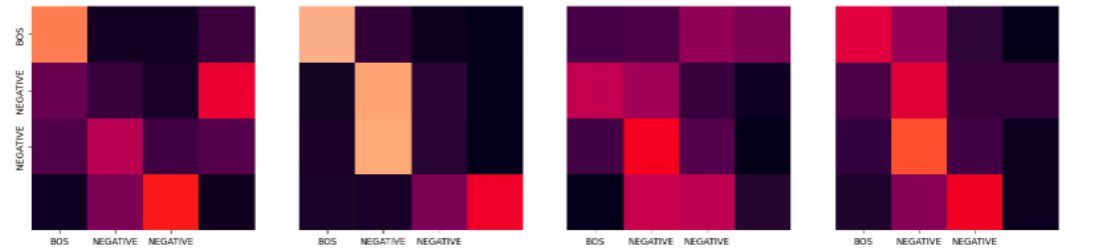


Decoder Block number 2
Decoder Self Attention 2

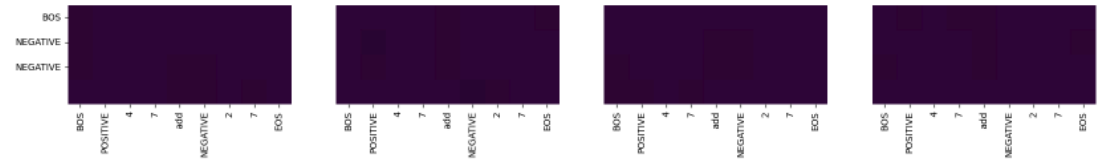
Decoder Cross attention 2

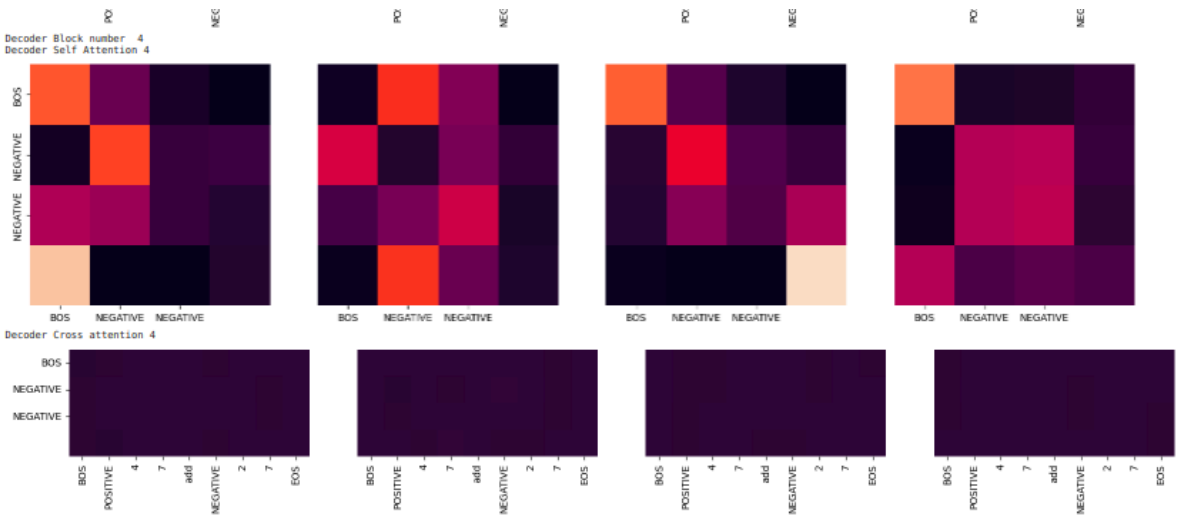


Decoder Block number 3
Decoder Self Attention 3



Decoder Cross attention 3




Decoder Block number 4
Decoder Self Attention 4
BOS
NEGATIVE
NEGATIVE
BOS
NEGATIVE
NEGATIVE
Decoder Cross attention 4
BOS
NEGATIVE
NEGATIVE
BOS
POSITIVE
4
7
add
NEGATIVE
2
7
EOS