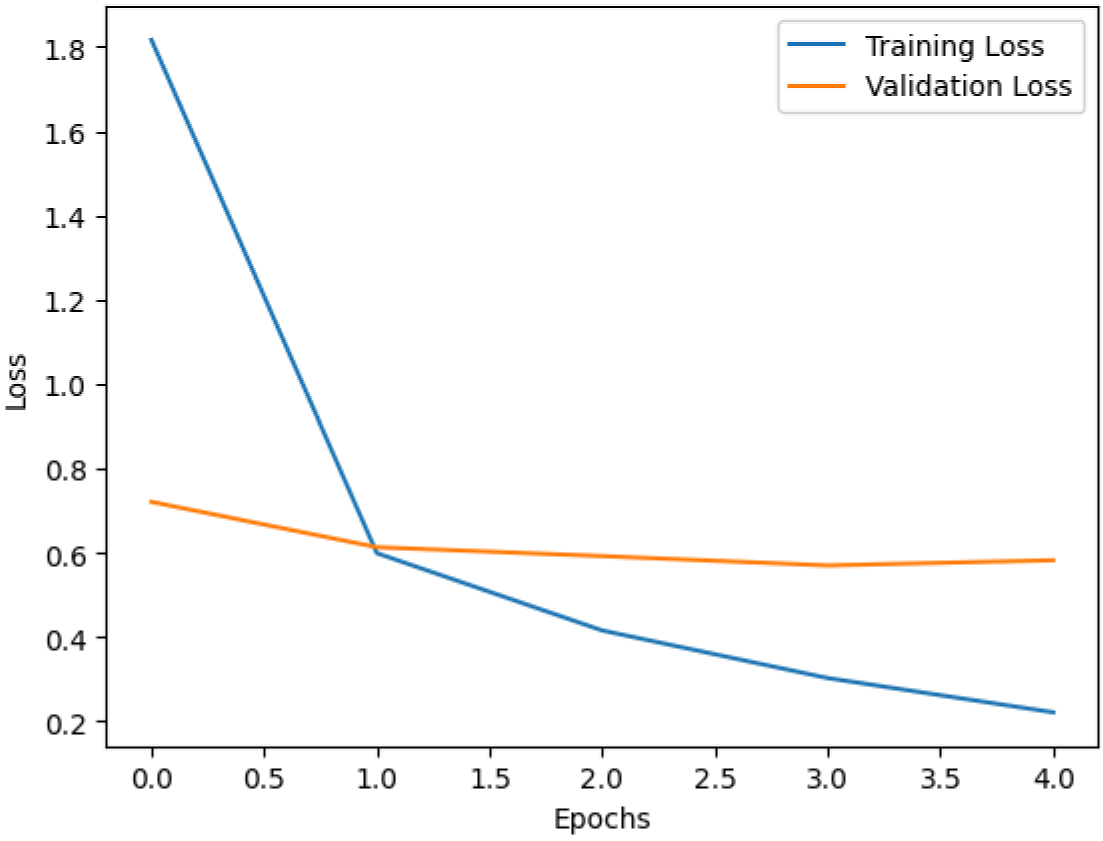
Group 36 – Assignment 3

CSE 556: Natural Language Processing

Task 1 - Text Similarity

Setup 1A

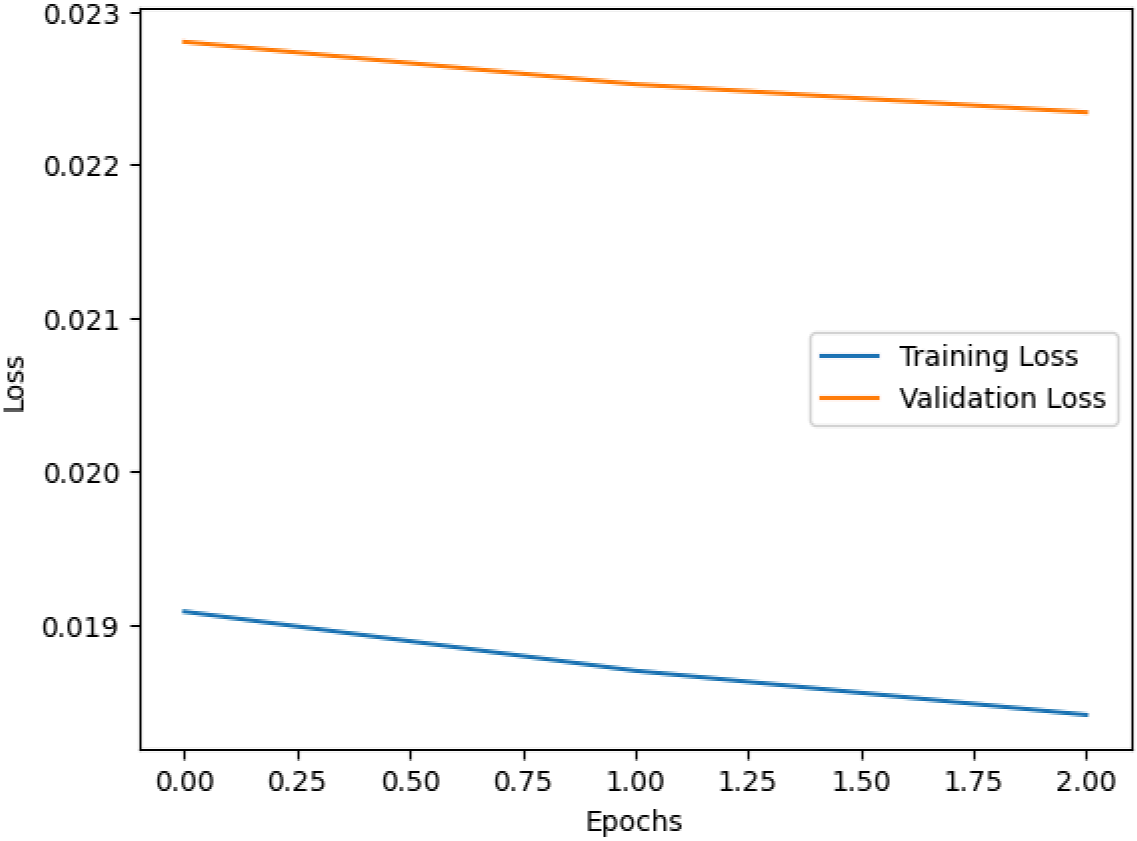


Validation loss doesn’t decrease after a few epochs, while training loss does.
Validation Pearson Correlation: 0.8676728865750375
Test Pearson Correlation: 0.9784852902947985

Setup 1B

Test-Set Pearson Correlation: 0.982463644410194
Validation-Set Pearson Correlation: 0.8631423846336786

Setup 1C



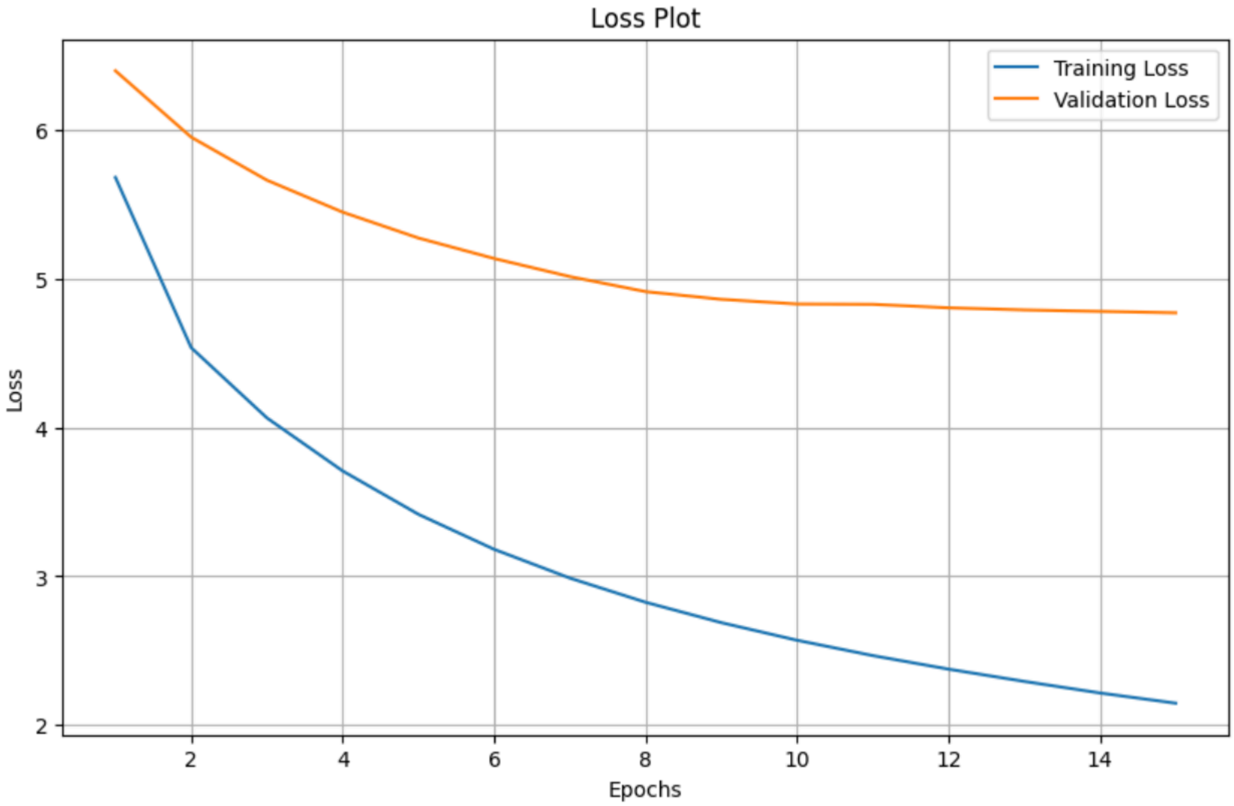
Both training and validation loss doesn’t decrease much and somewhat converge on fine-tuning.
Validation Pearson Score: 0.8904479023415723
Test Pearson Score: 0.986926101342789

All the models performed similarly to some extent. Task1B and Task1C use the same model for the STS task. However, Task1C model provides better results on Pearson correlation due to fine-tuning.

Task 2 - Machine Translation

Setup 2A

We trained the model on 100K training data for 15 epochs. We noticed that validation loss decreases to a point then plateaus and starts to rise again while the training loss continues to go down. This is due to overfitting and hence we increased the training data to 100K, but the results improved only slightly. Also, we added weight decay to minimize the overfit. We employed early stopping and let the model train further past the plateau while retaining the best model.



Smooth loss and the expected trend in both training and validation, then validation loss plateaus

Evaluation Score on Validation Set:

BLEU-1: 100.0
BLEU-2: 71.42857142857143
BLEU-3: 30.76923076923077
BLEU-4: 8.333333333333334
BLEU score: 36.78763249927777
Average METEOR Score for validation data: 0.3043
Bert-Score for Validation data:
Average Precision: 0.8538
Average Recall: 0.8632

Average F1: 0.8584

Evaluation Score on Test Set:

BLEU-1: 100.0
BLEU-2: 25.0
BLEU-3: 7.142857142857143
BLEU-4: 4.166666666666667
BLEU score: 16.515821590069034
Average METEOR Score for test data: 0.3153
Bert-Score for Test data:
Average Precision: 0.8536
Average Recall: 0.8607
Average F1: 0.8570

Setup 2B

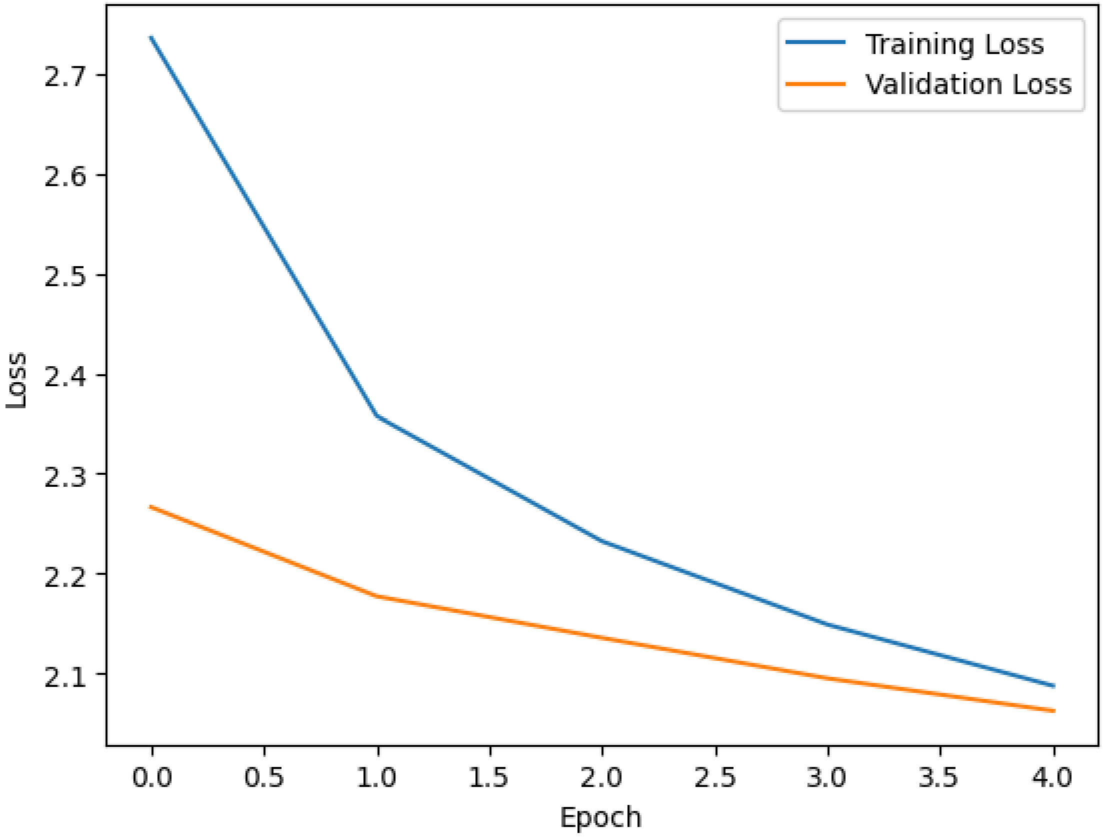
Evaluation Score on Validation Set:

BLEU Score: 39.281465090051306
BLEU-1: 100.0
BLEU-2: 50.0
BLEU-3: 28.571428571428573
BLEU-4: 16.666666666666668
METEOR Score: 0.3697782590377402
BERT Scores:
Precision in BERT Score: 0.8268641829490662
Recall in BERT Score: 0.7662411332130432
F1 Score in BERT Score: 0.7942118644714355

Evaluation Score on Test Set:

BLEU Score: 0.0
BLEU-1: 66.66666666666667
BLEU-2: 25.0
BLEU-3: 25.0
BLEU-4: 0.0
METEOR Score: 0.3889479564070389
BERT Scores:
Precision in BERT Score: 0.8345420956611633
Recall in BERT Score: 0.7703850269317627
F1 Score in BERT Score: 0.8000085949897766

Setup 2C



Both training and validation loss are converging very quickly.

<u>Evaluation Score on Validation Set:</u>

BLEU Score: 11.478744233307168
BLEU-1 Score: 70.0
BLEU-2 Score: 11.11111111111111
BLEU-3 Score: 6.25
BLEU-4 Score: 3.5714285714285716
METEOR Score: 0.3227082170618216
BERT Scores:
Precision in BERT Score: 0.79283207654953
Recall in BERT Score: 0.7551372647285461
F1 Score in BERT Score: 0.7728002071380615

Evaluation Score on Test Set:
BLEU Score: 0.0
BLEU-1 Score: 66.66666666666667
BLEU-2 Score: 25.0
BLEU-3 Score: 25.0
BLEU-4 Score: 0.0
METEOR Score: 0.35205323785503273
BERT Scores:
Precision in BERT Score: 0.8002291321754456
Recall in BERT Score: 0.7628726363182068
F1 Score in BERT Score: 0.7803449034690857

We were able to finetune ‘t5-small’ for German to English translation. The results were also good to a great extent. The translations visible in the Setup files are intuitive.
Results for Task 2A were also good but not as good as those of the fine-tuned model for Task 2C because t5-small is already pre-trained.
Task 2B model performed somewhat similar to the fine-tuned model as expected because it is the same model but on a different problem statement.
Results on the model for Task-2A are expected to increase provided that we use more training data, Transformers require a lot of data to perform well.

Contributions

1. Khushdev Pandit:
    a. helped in Task1A & Task1B
    b. attempted Task1C, Task2B and Task2C
2. Arjun Mehra: Task1A, Task1B
3. Pankaj: Helped in Task1C
4. Apurv Dube - Task 2A