Assignment-3

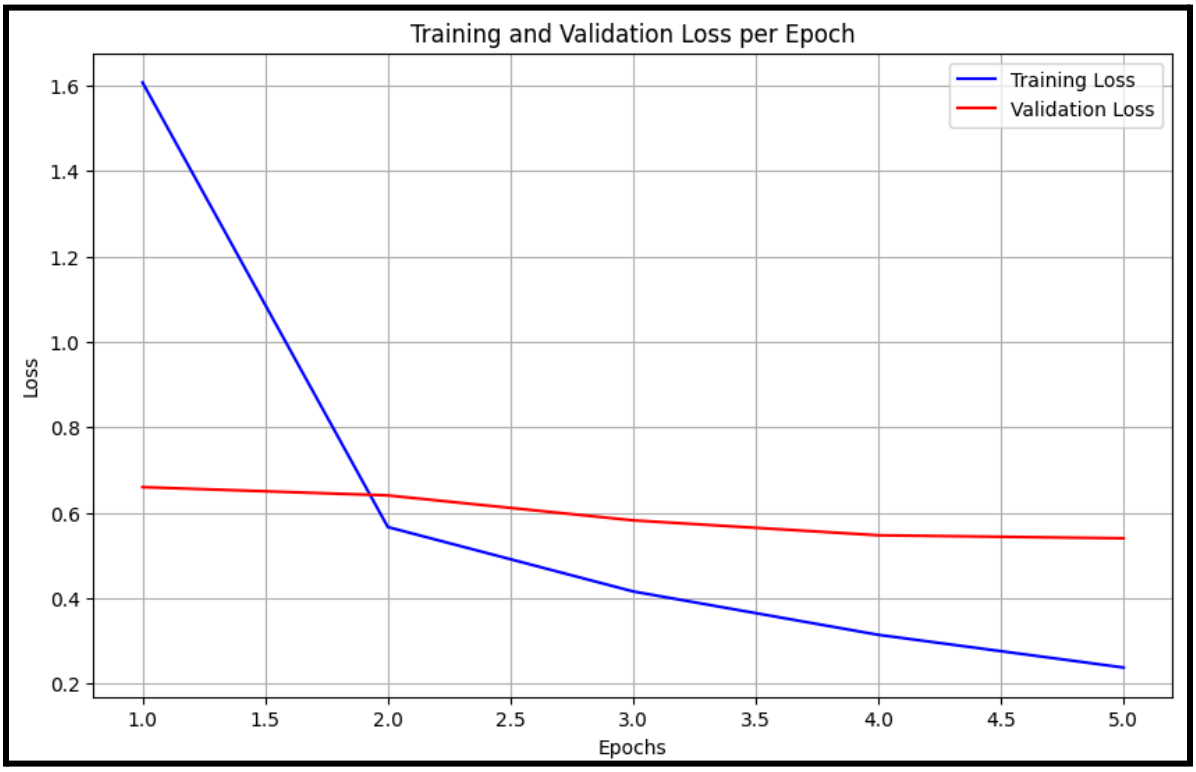
1. **Semantic Textual Similarity Task**

   Data Cleaning: We first cleaned the given datasets, train data had 38 rows and validation data had 30 rows which needed to be dropped as their formatting was different from the rest of the dataset and had some garbage data(wikipedia links, etc.)

   A. Using BERT Sequence Classification

   a. Used <SEP> special token to format inputs for the BERT model.
   b. Ignored all the null rows.
   c. Used BERTSequentialClassification model ‘bert-base-uncased’, returning one float output. Then applied the sigmoid function on it and scaled it from 0 to 5.
   d. Used Mean Square Error loss function to evaluate training as well as validation loss.
   e. Pearson Correlation for the Validation Set - 0.88

   Training and Validation Losses per Epoch Graph -



We can see that the training loss decreases with increasing epochs, since the model fits on this data. However, validation loss first decreases then begins increasing showing signs of the beginning of overfitting.

B. Sentence BERT to get Cosine Similarities

a. Import Sentence BERT model using sentence transformers.

b. Obtain cosine similarity scores for all the sentences in the validation set and scale them to the range 0 to 5. First, negative values were removed by taking the absolute value of the lowest cosine score and adding it to all scores, i.e. , the lowest score was scaled to 0 and the others subsequently increased by the same amount. These scores were then multiplied by the appropriate factor to bring them in the 0-5 range

c. Pearson Correlation for the Validation set - 0.860.

C. Fine-tuning sentence-BERT

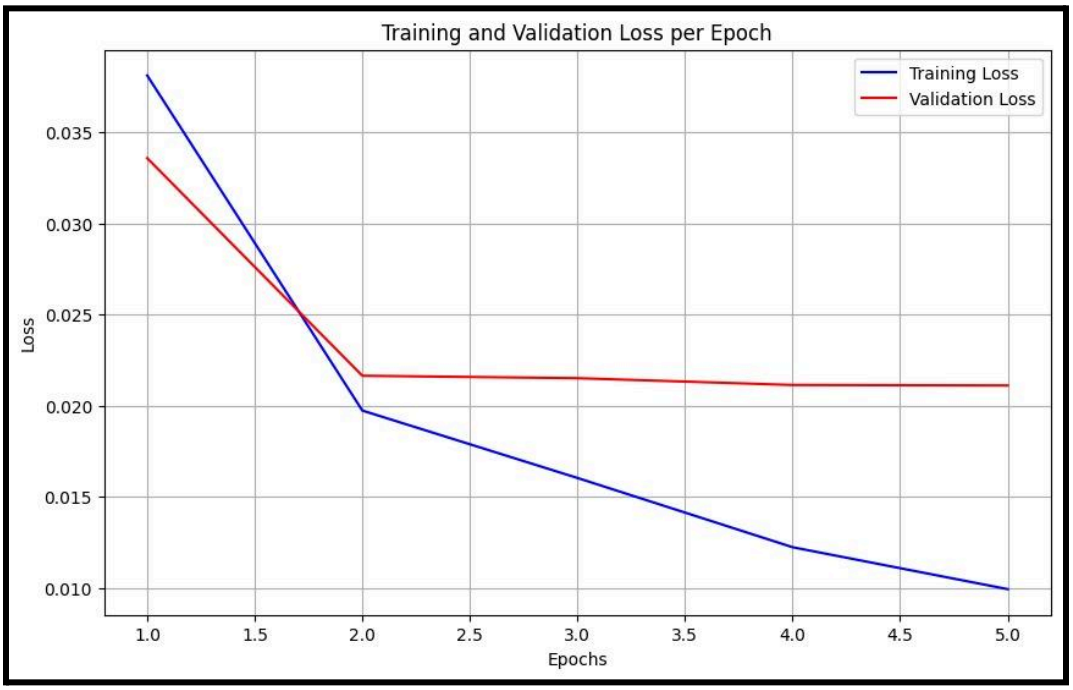
a. Import Sentence BERT model using sentence transformers.

b. Used the cosine similarity loss defined in pytorch to train the model.

c. Calculated training loss and validation loss for each epoch using cosine similarity and mean square error.

d. Pearson Correlation for the Validation Set - 0.8936678884907561. We can see that the fine-tuning has increased the accuracy of the sentence-BERT model.

Training and Validation Losses per Epoch Graph -

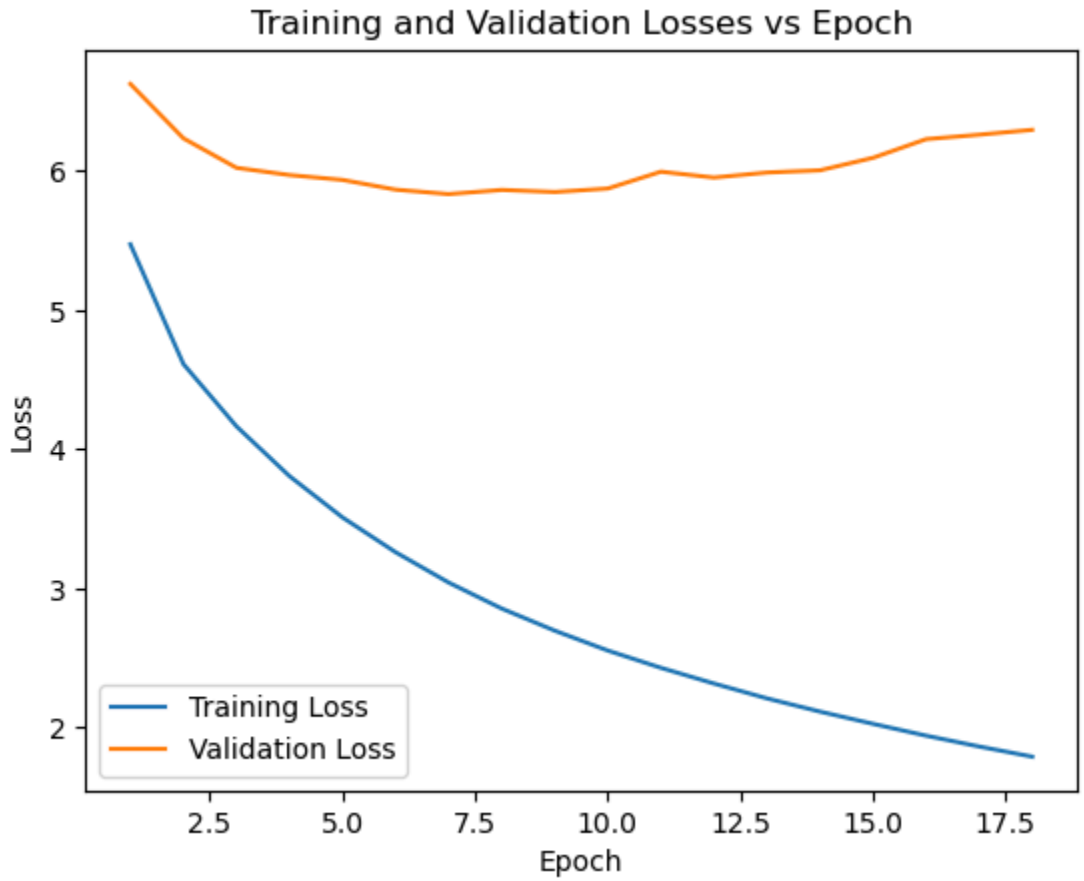


Again, we can see that both training and validation losses decrease with increase in the number of epochs. However, the magnitude of decrease in the validation loss is low and later becomes almost constant, showing that the model is moving towards overfitting.

We can see the best model was 1C. We can see that it is better than 1B as it has been fine-tuned on the data. Also, 1C is better than 1A since it has been specifically pre-trained to perform similarity matching on embeddings using the Siamese network architecture. Both 1A and 1C work better than 1B since they are fine-tuned on the current dataset.

2. Translation

A. Loss vs Epochs



Initially, both the training and validation losses start relatively high, which is expected when training a model from scratch.

As the training progresses, the training loss decreases steadily, indicating that the model is learning and improving its performance on the training data over the epochs. This is a desirable trend.

However, the validation loss does not decrease consistently. Instead, it fluctuates and even increases towards the later epochs, while the training loss continues to decrease.

Evaluation Metrics

Evaluation metrics for test set

BLEU Score 1	0.3209426733174293
BLEU Score 2	0.16163278911713322
BLEU Score 3	0.08968195242082655
BLEU Score 4	0.05231745082641526
METEOR Score	0.3558839072684144
BERTScore (Precision Score)	0.8229262863210377

Evaluation metrics for the validation set

BLEU Score 1	0.31581628494579833
BLEU Score 2	0.15169960866671464
BLEU Score 3	0.08174647886948301
BLEU Score 4	0.046390844376141084
METEOR Score	0.3516015046228352
BERTScore (Precision Score)	0.8219276331382284

B. Evaluation metrics:

Evaluation metrics for test set

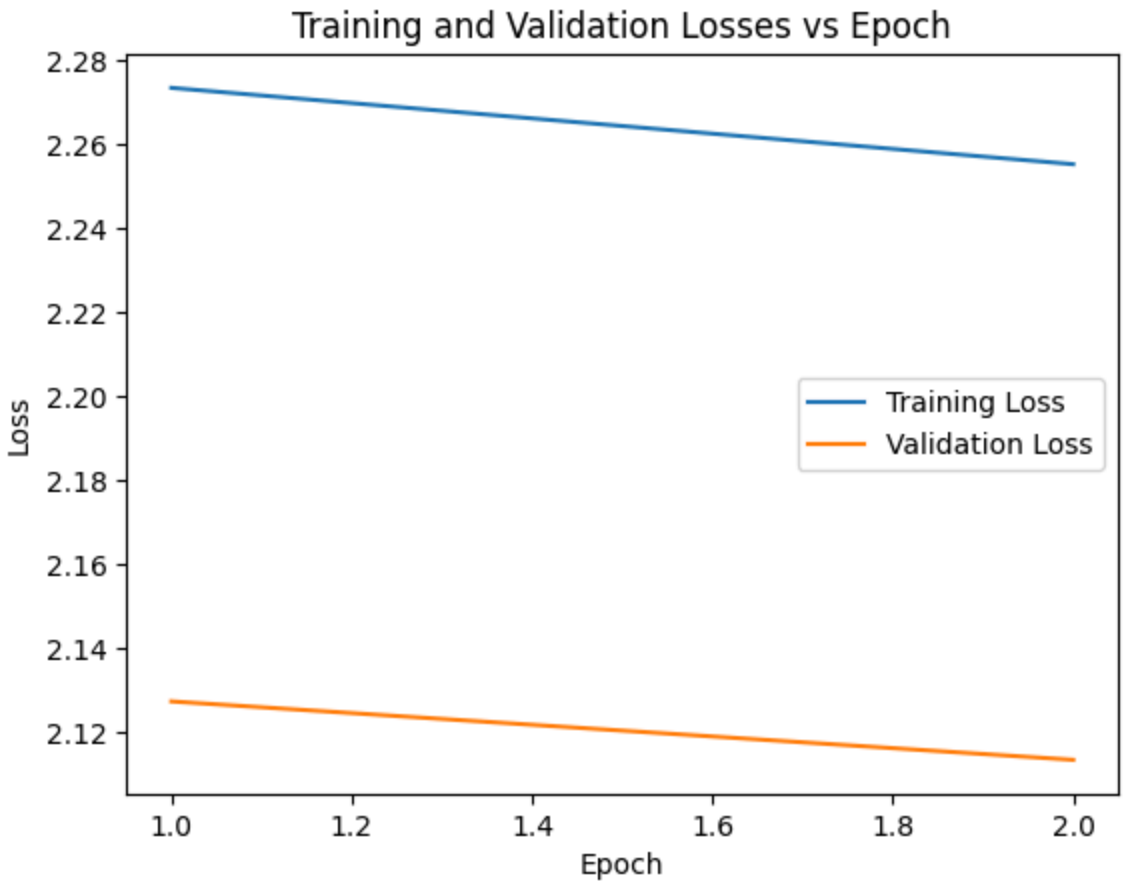
BLEU Score 1	0.27167469339738115
BLEU Score 2	0.2112397095608373

BLEU Score 3	0.16942337277959213
BLEU Score 4	0.13841273518987154
METEOR Score	0.3889479564070394
BERTScore (Precision Score)	0.8345420370900102

Evaluation metrics for the validation set

BLEU Score 1	0.2661676664462372
BLEU Score 2	0.20111319933833893
BLEU Score 3	0.15829949742184973
BLEU Score 4	0.1275631721885134
METEOR Score	0.3697782590377392
BERTScore (Precision Score)	0.8268642122704843

C. Loss vs Epochs



Both the training and validation losses appear to be decreasing over the course of the 2 epochs, which is a desirable trend and indicates that the model is learning and improving its performance on the translation task during the fine-tuning process.

However, it's important to note that the validation loss is consistently higher than the training loss. This is a common phenomenon known as overfitting, where the model performs better on the training data it has seen during fine-tuning but struggles to generalize well to unseen data (validation set).

Evaluation metrics:

Evaluation metrics for validation set

BLEU Score 1	0.2631952220379738
BLEU Score 2	0.16871324073053776

BLEU Score 3	0.1155930481126913
BLEU Score 4	0.08172596886950223
METEOR Score	0.3086951835762893
BERTScore (Precision Score)	0.8918597267871081

Evaluation metrics for test set

BLEU Score 1	0.27167469339738115
BLEU Score 2	0.2112397095608373
BLEU Score 3	0.16942337277959213
BLEU Score 4	0.13841273518987154
METEOR Score	0.3889479564070394
BERTScore (Precision Score)	0.8345420370900102

Contributors :

Ayush Srivastava : Task 1
Manas Narang : Task 1
Abhishek Sushil : Task 2
Kartik Gupta : Task 2