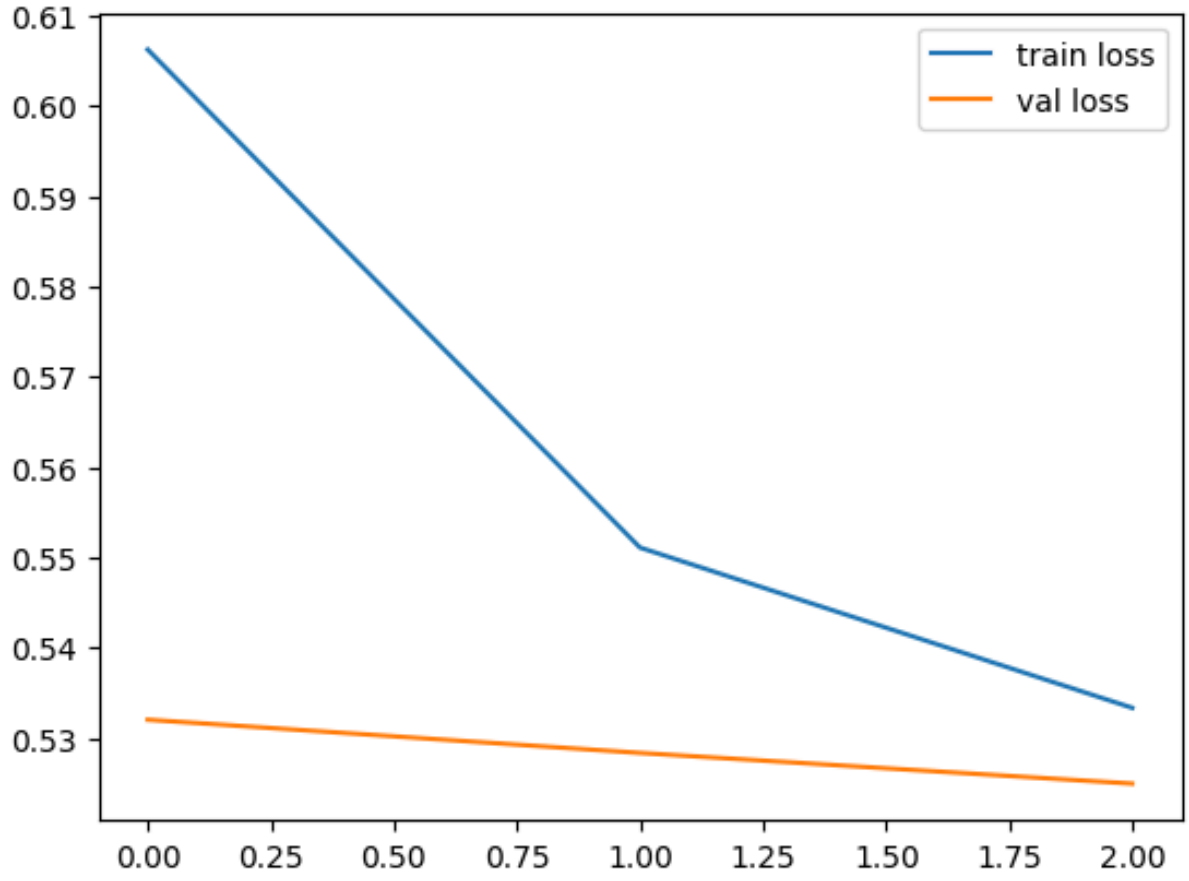
CSE 556: Natural Language Processing
Assignment 3 Report
Group 16

Task 1 - Text Similarity

Part 1A-

Loss Plot: Training Loss and Validation Loss V/s Epochs-



Analysis and Explanation-

- The training loss (blue line) falls sharply in the first few epochs, which is typical as the model initially has a lot to learn from the training data. The plot shows the training loss decreasing at a decelerating rate and starting to flatten as it progresses, indicating the model is beginning to saturate in terms of learning from the training data.
- The validation loss (orange line) decreases slightly throughout the plot, indicating that the model is not overfitting to the training data but the model is learning and improving over time, at a slower rate.

Pearson Correlation on validation dataset-0.7956897850304394

Part 1B

Pearson Correlation on validation dataset- 0.7923100957816998

Part 1C

Pearson Correlation on validation dataset- 0.8572838200887052

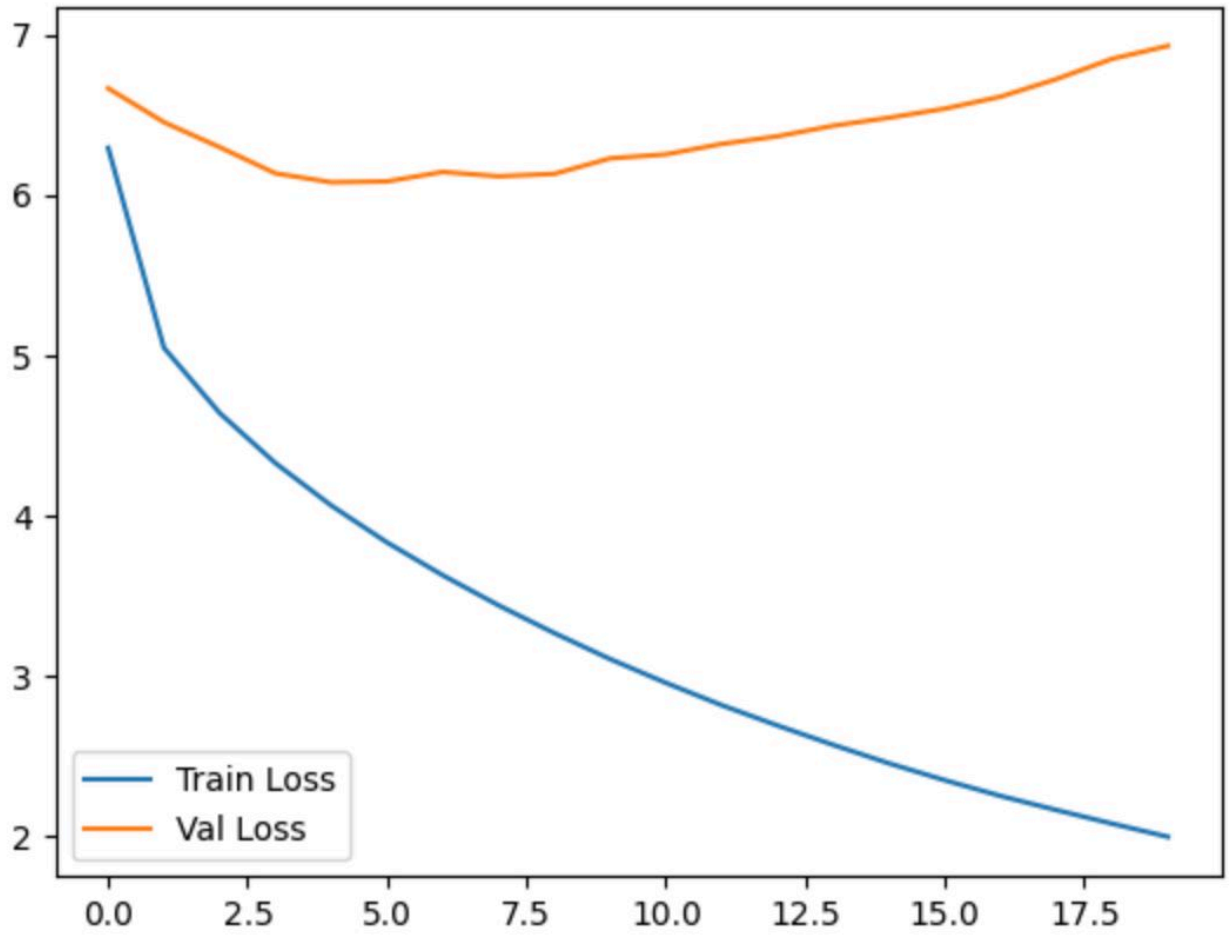
Comparison and explanation for the performance differences

- Part 1C outperforms both Part 1A and Part 1B in terms of the Pearson Correlation on the validation dataset.
- This improvement in performance in Part 1C could be attributed to fine-tuning the Sentence-BERT model specifically for the STS task, which allows the model to learn task-specific representations.
- In Part 1A, although using a pre-trained BERT model is beneficial, it might not capture the nuances of the specific task as effectively as fine-tuning
- Part 1B's performance falls between Part 1A and Part 1C, indicating that while Sentence-BERT offers better performance than vanilla BERT for text similarity tasks, fine-tuning for the specific task further enhances performance.

Task 2 - Machine Translation

Part 2A-

Loss Plot: Training Loss and Validation Loss V/s Epochs-



Analysis and Explanation-

- The plot shows the training loss(blue line) and the validation loss(orange line) across the epochs. It can be observed that train loss decreases rapidly from around 6.2 to approximately 2. This steep decline indicates that the model is learning quickly from the training data and is effectively minimizing the error between its predictions and the true values. At the beginning (near 0 epochs), the training loss is significantly higher than the validation loss.

- The validation loss starts out lower than the training loss, which is not common in many training scenarios. However, it shows a steady, although minor, upward trend. This gradual increase might indicate that the model is starting to overfit the training data. As epochs increase, the two loss curves diverge: training loss continues to decrease while validation loss slightly increases. This is a classic indication of overfitting.

Validation Metrics-

- **BLEU Score:** {'bleu': 0.03851418841522835, 'precisions': [0.2934494653986163, 0.06612326043737575, 0.01861547411285631, 0.006091458904407605], 'brevity_penalty': 1.0, 'length_ratio': 1.12975044678423, 'translation_length': 52469, 'reference_length': 46443}
  BLEU1:0.2934494653986163
  BLEU2: 0.06612326043737575
  BLEU3:0.01861547411285631
  BLEU4:0.006091458904407605

- **METEOR Score:**{'meteor': 0.2415273378585499}

- **BERTScore-precision:** 0.8400157972545105

- **BERTScore-recall:** 0.8507186676731523

- **BERTScore-f1:** 0.8452254099579878

Test Metrics-

- **BLEU Score:** {'bleu': 0.044099405884553494, 'precisions': [0.2912421432320758, 0.07173913043478261, 0.022474048696212355, 0.008054513436949698], 'brevity_penalty': 1.0, 'length_ratio': 1.1587672920663141, 'translation_length': 74299, 'reference_length': 64119}
  BLEU1:0.2912421432320758
  BLEU2: 0.07173913043478261
  BLEU3:0.022474048696212355
  BLEU4:0.008054513436949698

- **METEOR Score:** {'meteor': 0.24615395968645679}
- **BERTScore-precision:**0.8390632928828868
- **BERTScore-recall:** 0.8474407107601885
- **BERTScore-f1:** 0.8431200475722959

Part 2B

Validation Metrics-

- **BLEU Score:** {'bleu': 0.2823125666465826, 'precisions': [0.5945855614973262, 0.3374072253049566, 0.21694413614502406, 0.14595002996118073], 'brevity_penalty': 1.0, 'length_ratio':

1.0140081337550837, 'translation_length': 44880, 'reference_length': 44260}
BLEU1:0.5945855614973262
BLEU2: 0.3374072253049566
BLEU3: 0.21694413614502406
BLEU4:0.14595002996118073

- **METEOR Score:** {'meteor': 0.5515488896080301}
- **BERTScore-precision:** 0.8656165647363817
- **BERTScore-recall:** 0.8643085971412817
- **BERTScore-f1:** 0.8647710810118746

Test Metrics-

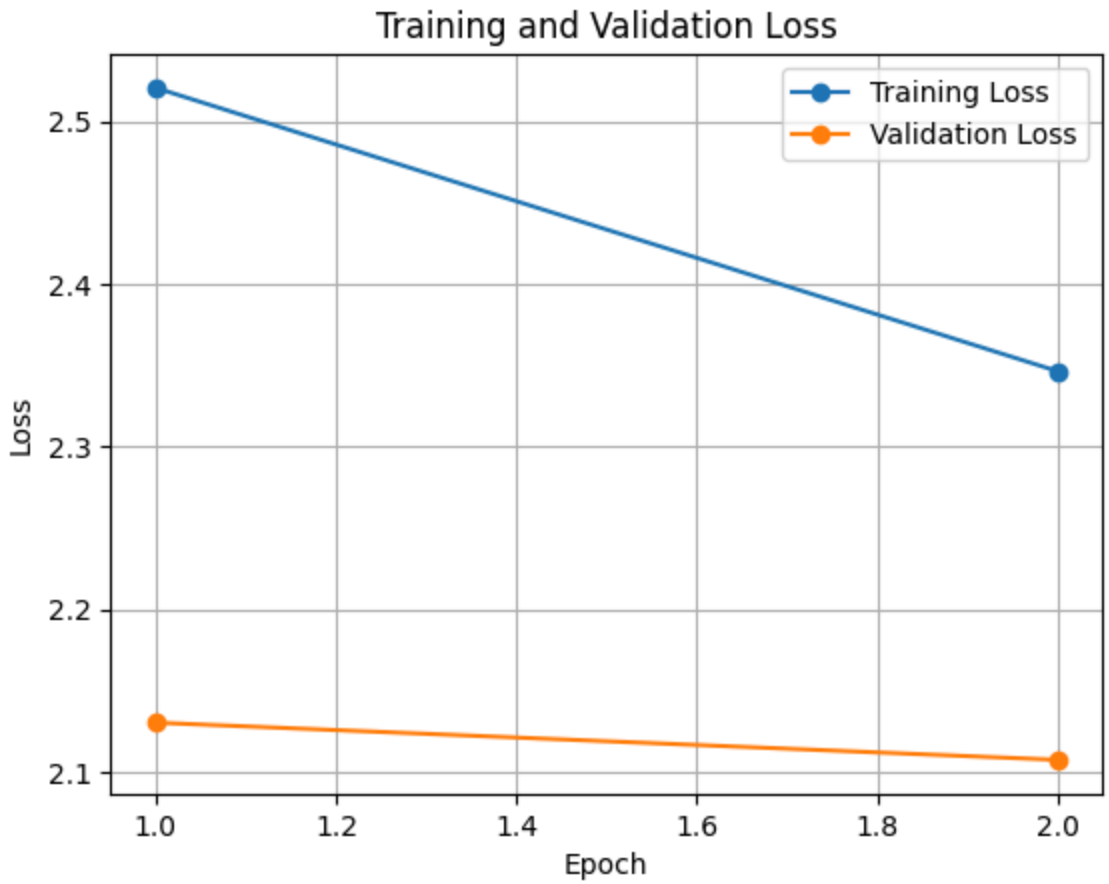
- **BLEU Score:** {'bleu': 0.3268254810286919,
 'precisions': [0.6340116232529098, 0.38745753590528403, 0.26043603775924057, 0.18072829969887969],
'brevity_penalty': 0.9966754605913349,
'length_ratio': 0.9966809746445611,
'translation_length': 62461, 'reference_length': 62669}
BLEU1:0.6340116232529098
BLEU2: 0.38745753590528403
BLEU3: 0.26043603775924057

BLEU4:0.18072829969887969

- **METEOR Score:** {'meteor': 0.5902469429945494}
- **BERTScore-precision:** 0.8778189009290888
- **BERTScore-recall:** 0.8741395660502468
- **BERTScore-f1:** 0.8757985283152665

Part 2C

Loss Plot: Training Loss and Valida
tion Loss V/s Epochs-




```
/usr/local/lib/python3.10/dist-packages/accelerate/accelerator.py:432: FutureWarning: Passing the fo
dataloader_config = DataLoaderConfiguration(dispatch_batches=None, split_batches=False, even_batches
  warnings.warn(
[6250/6250 13:16, Epoch 2/2]
```

Epoch	Training Loss	Validation Loss
1	2.520400	2.130549
2	2.346400	2.107708

```
TrainOutput(global_step=6250, training_loss=2.43339609375, metrics={'train_runtime': 798.2362, 'trai
125.276, 'train_steps_per_second': 7.83, 'total_flos': 2687376416047104.0, 'train_loss': 2.433396093
```

Analysis and Explanation-

The plot displays the training loss (blue line) and validation loss (orange line) over 2 epochs. The training loss starts around 2.52 and decreases gradually to 2.34 as the model learns from the training data. The validation loss starts around 2.13 and decreases to 2.10.

While the training loss is higher than the validation loss throughout the epochs, the gap between them narrows over time. This convergence between training and validation losses indicates that the model is not overfitting excessively to the training data and is generalizing well to unseen data. The slight decrease in validation loss over epochs suggests that the model is continuing to learn and improve its performance on the validation set.

Key observations from these plots are:

- The model is learning effectively from the training data, as evidenced by the decreasing losses.
- There is a gap between the training and validation metrics, indicating some degree of overfitting, where the model performs better on the training data than on the unseen validation data.

Validation Metrics-

- **BLEU Score:** {'bleu': 0.16651898208894583, 'precisions': [0.5077333215827973, 0.22288808163076423, 0.11548486103329841, 0.06456340635606295], 'brevity_penalty': 0.9770240322294348, 'length_ratio': 0.9772839825162026, 'translation_length': 45388, 'reference_length': 46443}

  BLEU1:0.5077333215827973
  BLEU2: 0.22288808163076423
  BLEU3: 0.11548486103329841
  BLEU4:0.06456340635606295

- **METEOR Score:** {'meteor': 0.44203751154922416}

- **BERTScore-precision:** 0.898599992621709
- **BERTScore-recall:** 0.9083440600314301
- **BERTScore-f1:** 0.9032908998749233

Test Metrics-

- **BLEU Score:** {'bleu': 0.19209094686911987, 'precisions': [0.5258708255848479, 0.24942615424973766, 0.13566920715936132, 0.0771258704387193], 'brevity_penalty': 0.9980017179398332, 'length_ratio': 0.9980037118482821, 'translation_length': 63991, 'reference_length': 64119}

  BLEU1:0.5258708255848479
  BLEU2:0.24942615424973766
  BLEU3: 0.13566920715936132
  BLEU4:0.0771258704387193

- **METEOR Score:** {'meteor': 0.48101835919583075}
- **BERTScore-precision:** 0.9000328742054631
- **BERTScore-recall:** 0.9114645639432594
- **BERTScore-f1:** 0.9055535279341085

Comparison and explanation for the performance differences -

- Results Comparison:
    - In terms of BLEU scores, Setup 2C tends to outperform Setup 2A on both validation and test datasets. This could be because the 't5-small' model, being pre-trained on a large corpus, already captures significant linguistic knowledge that benefits translation tasks.
    - METEOR and BERTScore also show improvements in Setup 2C, indicating better overall translation quality compared to Setup 2A.

- Overfitting:
    - Setup 2A might be prone to overfitting, as indicated by the divergence between training and validation losses. However, in Setup 2C, the convergence between training and validation losses suggests better generalization and less overfitting.

- Computational Efficiency:
    - Fine-tuning a pre-trained model like 't5-small' (Setup 2C) is usually more computationally efficient compared to training a model from scratch (Setup 2A), as it requires fewer epochs and less data for training.

In summary, the performance differences between Setup 2A and Setup 2C can be attributed to the differences in model architectures, training approaches, and the utilization of pre-trained weights. Setup 2C, leveraging the 't5-small' model, generally achieves better translation quality with less overfitting compared to the model trained from scratch in Setup 2A

Contribution

1. Part-1 Task 1A - Nalish Jain(2021543)
2. Part-1 Task 1B - Shobhit Pandey(2021287), Aniket Malik(2021231)
3. Part-1 Task 1C - Sanmay Sood(2021095), Shobhit Pandey(2021287)
4. Part-2 Task 2A -Nalish Jain(2021543)
5. Part-2 Task 2B - Aniket Malik(2021231)
6. Part-2 Task 2C - Sanmay Sood(2021095), Nalish Jain(2021543)
7. Inference - Nalish Jain(2021543) and Aniket Malik(2021231)
8. Report - Sanmay Sood(2021095), Shobhit Pandey(2021287)