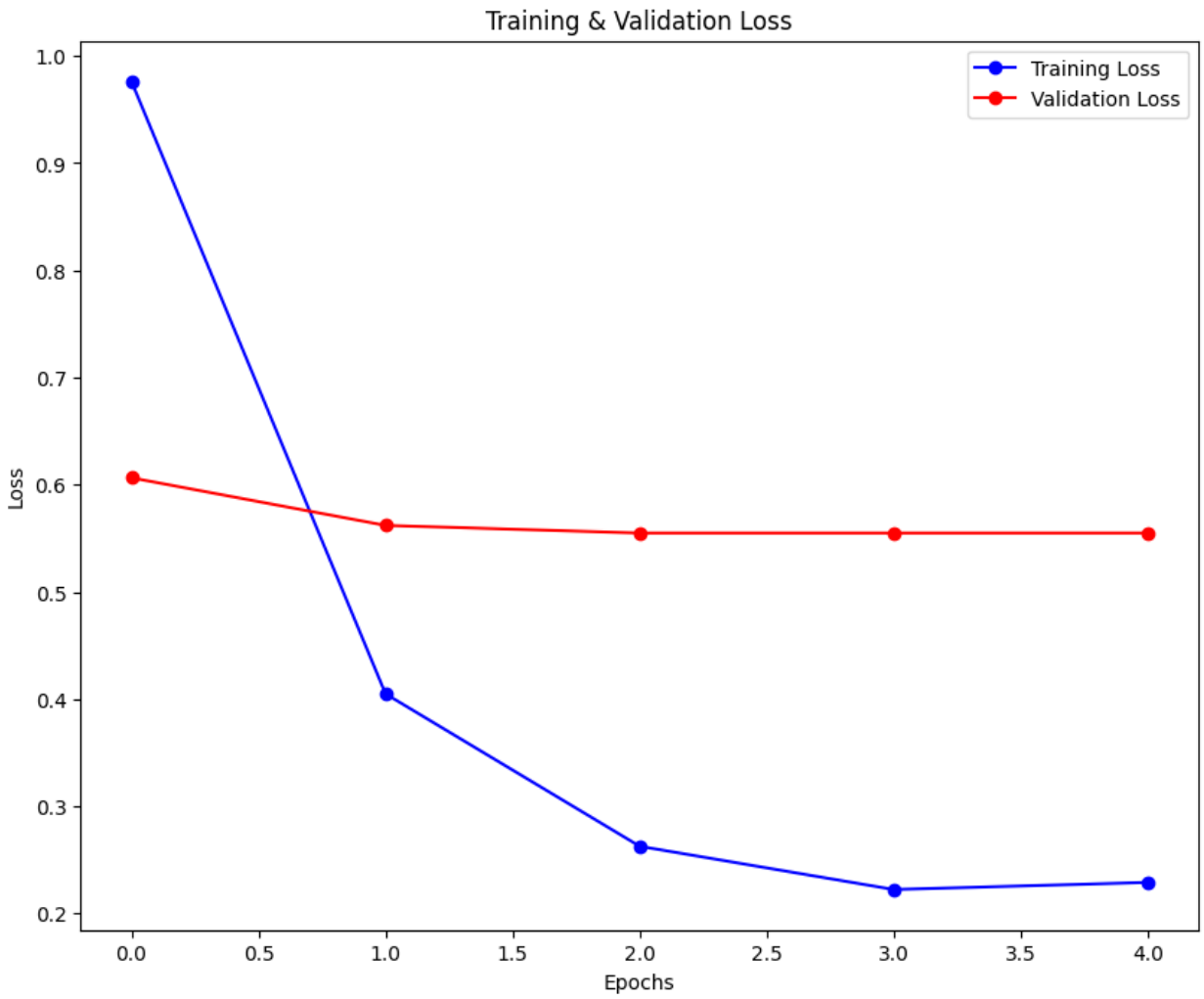
# Natural Language Processing (Assignment 3 REPORT)

## Task 1:

- **Setup 1A:**

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



Loss is consistently decreasing over each epoch, it can be observed that the decreasing rate of validation loss is comparatively lower than training loss.

Overall loss decreasing over each epoch → Model is learning

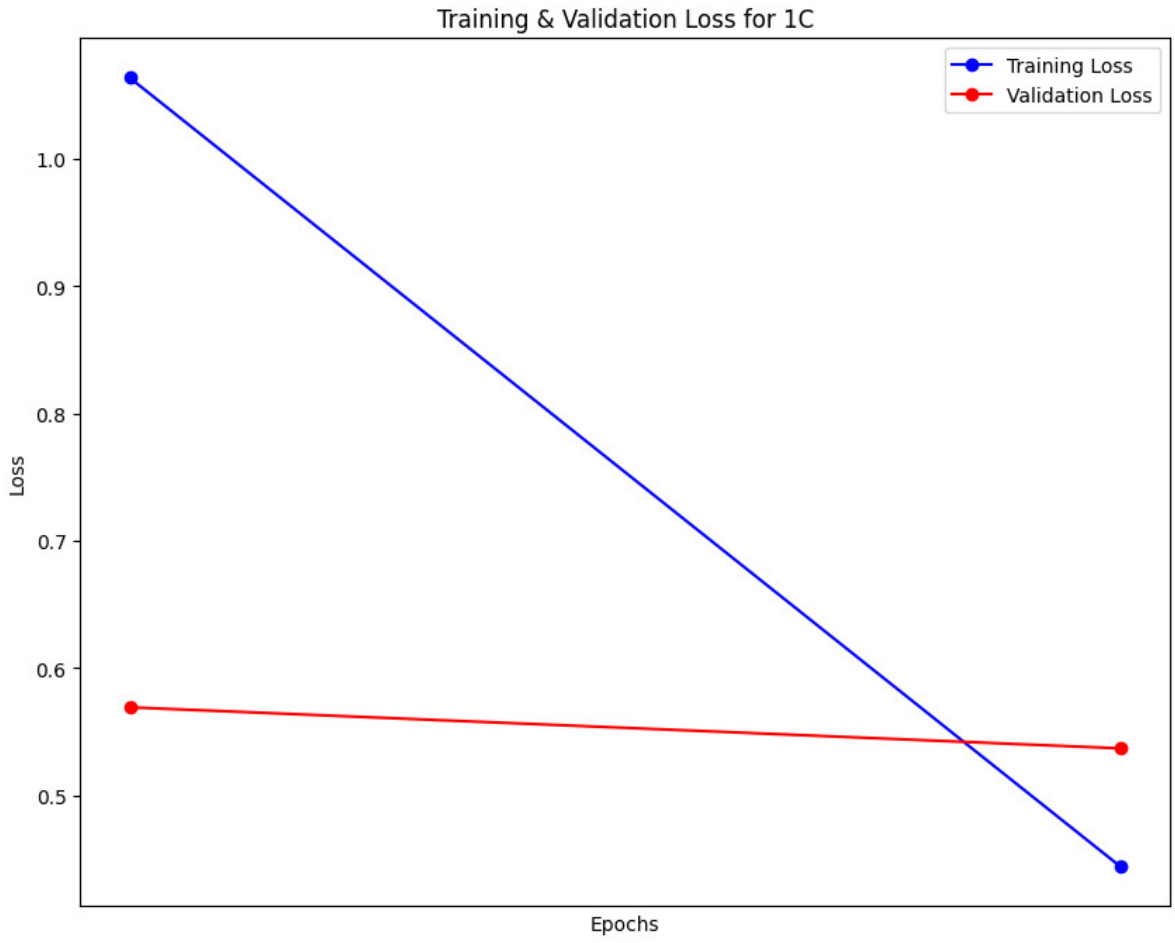
Pearson Correlation Coefficient: 0.8701540028942458

- **Setup 1B:**

  Used Cosine similarity between embeddings for validation.

  Pearson Correlation Coefficient On the Validation Set: 0.872128606029852

- **Setup 1C:**




  Used Cosine Similarity loss function.

  Pearson Correlation on Validation Set: 0.8944839734464778

- **Brief comparison and explanation of the performance differences between the three setups**

  The BERT model used in setup 1A is a base BERT model without any task-specific fine-tuning, so it captures semantic similarity to a reasonable extent without task-specific fine-tuning that's why it has not performed optimally compared to models fine-tuned specifically for the STS task.

  In setup 1B, the Sentence-BERT model from the SentenceTransformers framework, whereas in setup 1C, the Sentence-BERT model is specifically fine-tined for the STS task and yields the best performance.

# Task 2

## PART 2A

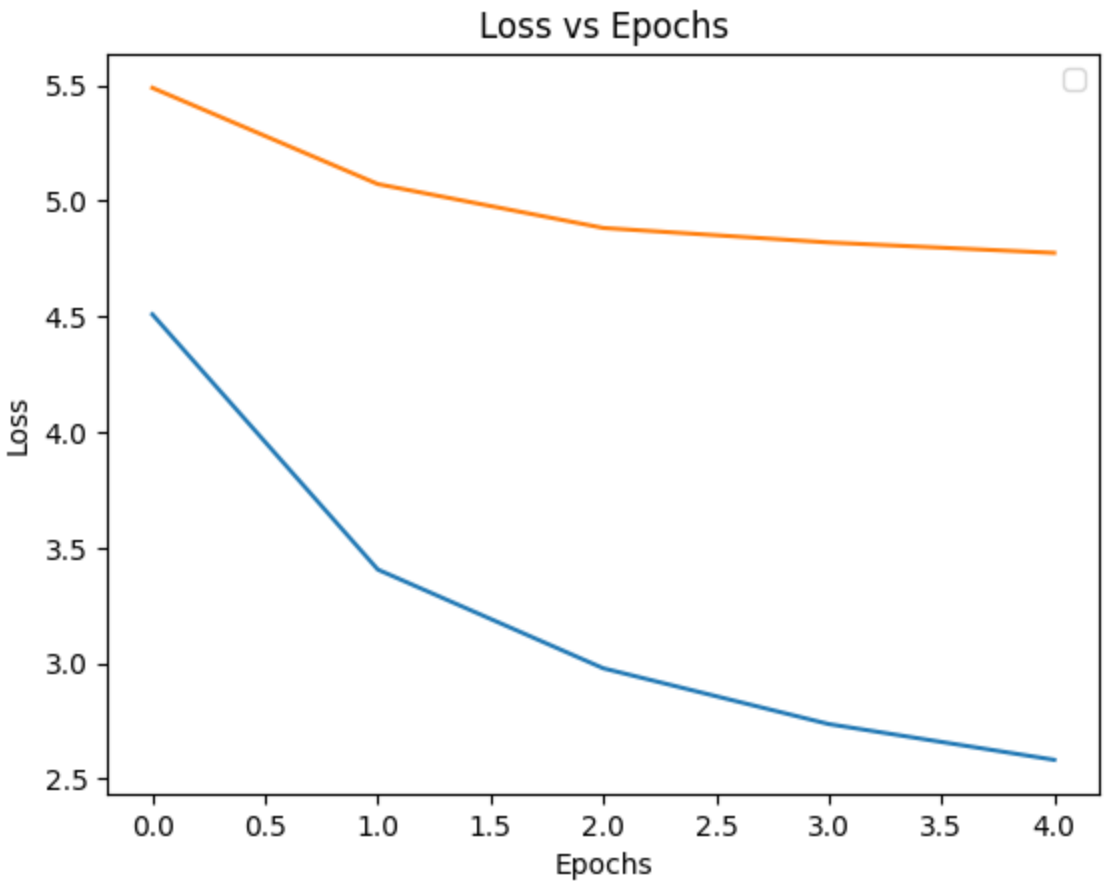
### Metrics

1. BLEU 1
   a. Validation - 0.36666801080688743
   b. Test - 0.37265239818746615
2. BLUE 2
   a. Validation - 0.2153545035148271
   b. Test - 0.223102289300297
3. BLUE 3
   a. Validation - 0.13552011516501977
   b. Test - 0.1423458596909431
4. BLUE 4
   a. Validation - 0.08876632014376198
   b. Test - 0.09420957621736477

5. Meteor
    a. Validation - 0.316223179323813
    b. Test - 0.33026008261526635
6. BERTScore
    a. Validation
        i. Precision - 0.8567657980461613
        ii. Recall - 0.864513774767841
        iii. F1 - 0.8605014695006618
    b. Test
        i. Precision - 0.855689146432053
        ii. Recall - 0.861997166129898
        iii. F1 - 0.8587042163832659

## Plot

a. ORANGE - Validation
   BLUE - Train




b. We can clearly see the training loss as well as validation loss decreasing upto 4 epochs but there is significant difference between the rate of decrement. It can be clearly seen that after 4 epochs model will start overfitting and its variance will increase so early stopping is done.

## Metrics

1. BLEU 1
    a. Validation -0.2600828236841529
    b. Test - 0.37265239818746615
2. BLEU 2
    a. Validation - 0.2153545035148271
    b. Test - 0.223102289300297

3. BLEU 3
    a. Validation - 0.13552011516501977
    b. Test - 0.1423458596909431
4. BLEU 4
    a. Validation - 0.08876632014376198
    b. Test - 0.09420957621736477
5. Meteor
    a. Validation - 0.316223179323813
    b. Test - 0.33026008261526635
6. BERTScore
    a. Validation
        i. Precision - 0.8567657980461613
        ii. Recall - 0.864513774767841
        iii. F1 - 0.8605014695006618
    b. Test
        i. Precision - 0.855689146432053
        ii. Recall - 0.861997166129898
        iii. F1 - 0.8587042163832659

PART 2A

| Metric | Val | Test |
|---|---|---|
| BLEU-1 | 0.2600828236841529 | 0.37265239818746615 |
| BLEU-2 | 0.2153545035148271 | 0.223102289300297 |
| BLEU-3 | 0.13552011516501977 | 0.1423458596909431 |
| BLEU-4 | 0.08876632014376198 | 0.09420957621736477 |
| Meteor | 0.316223179323813 | 0.33026008261526635 |
| BERT-SCORE | | |

| Metric | Val | Test |
|---|---|---|
| Precision | 0.8567657980461613 | 0.855689146432053 |
| Recall | 0.864513774767841 | 0.861997166129898 |
| F1 | 0.8605014695006618 | 0.8587042163832659 |

PART 2B

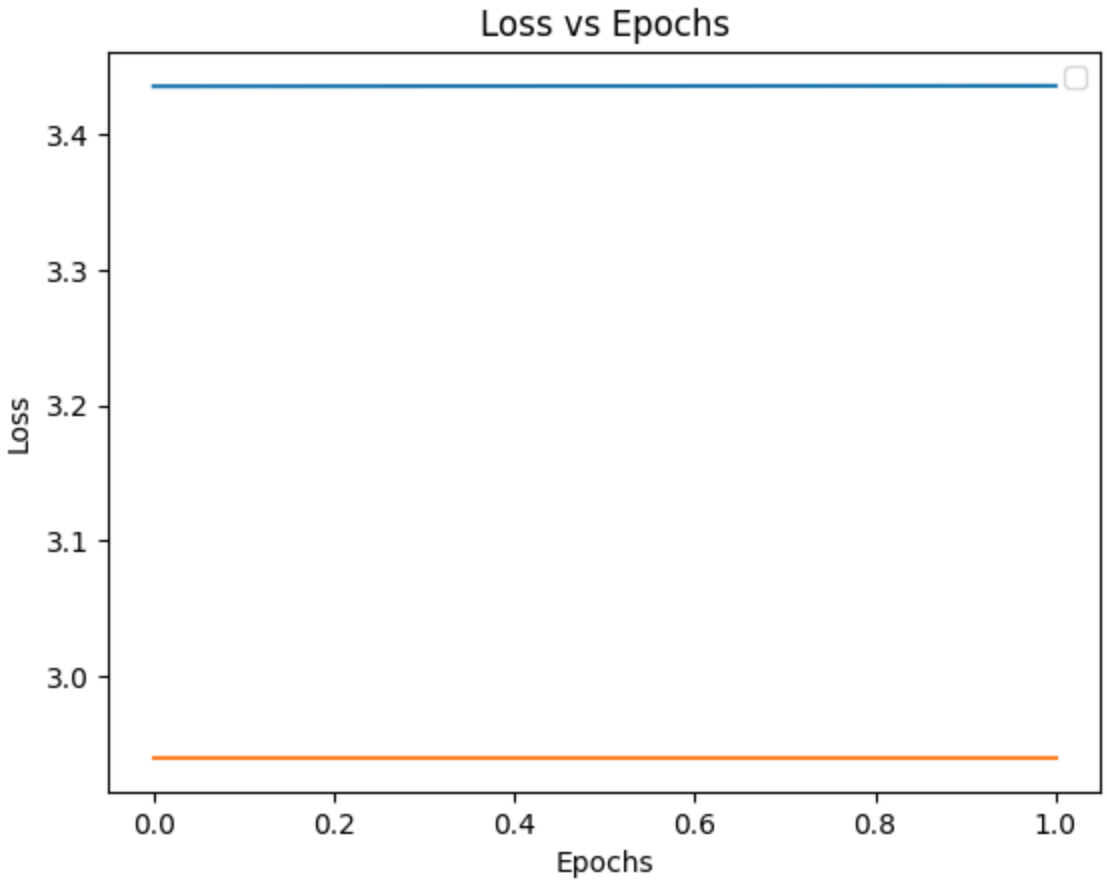
| Metric | Val | Test |
|---|---|---|
| BLEU-1 | 0.2600828236841529 | 0.26747473119807746 |
| BLEU-2 | 0.19594776651894452 | 0.20670570423053988 |
| BLEU-3 | 0.15392176677126235 | 0.16492522825450043 |
| BLEU-4 | 0.12388498267558495 | 0.13401190862752402 |
| Meteor | 0.35849881537769956 | 0.37840198623189225 |
| BERT-SCORE | | |
| Precision | 0.8202985864512213 | 0.8293289487939868 |
| Recall | 0.7605131295157116 | 0.7654230316486625 |
| F1 | 0.7880993610485846 | 0.7949184006196175 |

PART 2C

| Metric | Val | Test |
|---|---|---|
| BLEU-1 | 0.2600828236841529 | 0.26747473119807746 |
| BLEU-2 | 0.19594776651894452 | 0.20670570423053988 |
| BLEU-3 | 0.15392176677126235 | 0.16492522825450043 |
| BLEU-4 | 0.12388498267558495 | 0.13401190862752402 |
| Meteor | 0.35849881537769956 | 0.37840198623189225 |
| BERT-SCORE | | |
| Precision | `0.8202985920571907` | `0.8293289566842983` |
| Recall | `0.7605131319889334` | `0.7654230377104887` |
| F1 | `0.7880993663797515` | `0.7949184092850476` |

PART 2C

a. BLUE - Train
ORANGE - Validation




b. Analysis

As the model's only one layer is trained(loss back propogated) and only 2 epoch be able to run due to lack of computation power the loss is not decreacing to a larger extent in t5 fine-tuning.

## Analysis

> pre-trained google-t5/t5-small vs (2A) Transformer:

It is fairly lage model 60 million parameters.
Trained on very large dataset as a common knowledge model.

The BLEU 1, 2 are worse for T5 than the trained transformer in  Task 2A whereas Bleu 3, 4 is better for T5.
This shows t5 performs better for the  longer sentence context  whereas  it struggles with shorter sentences. As transformer is tranined on given data for the same particular task it is texted for.

T5 is generic model which understands the language more precisely but is not for translation task while our model is trained for only translation task only. That is why we can clearly see that BLEU 1, 2 scores of our Seq2Seq transformer are higher while BLEU 3, 4 scores of t5 is higher. Similar observations can be made with meteor and BERT score.

t5 perform a bit better than Transformer, wrt to Meteor score.

t5 performs worse to transformer(2A), wrt BERTScore

> pre-trained google-t5/t5-small vs fine-tuned google-t5/t5-small :
> {2B vs 2C}

As the t5 model is so  big and has many parameters; we are not able to train the model much.
Only one layer of t5 is finetuned (due to computation boundation)

Thus the specific task to translate german to english is not learned by the fine-tuned one.

## Conribution

1A - Utkarsh, Manav

1B - Utkarsh, Akash

1C - Akash, Lakshay

2A - Manav, Akash

2B - Lakshay Utkarsh

2C - Manav Lakshay

Report - Everyone