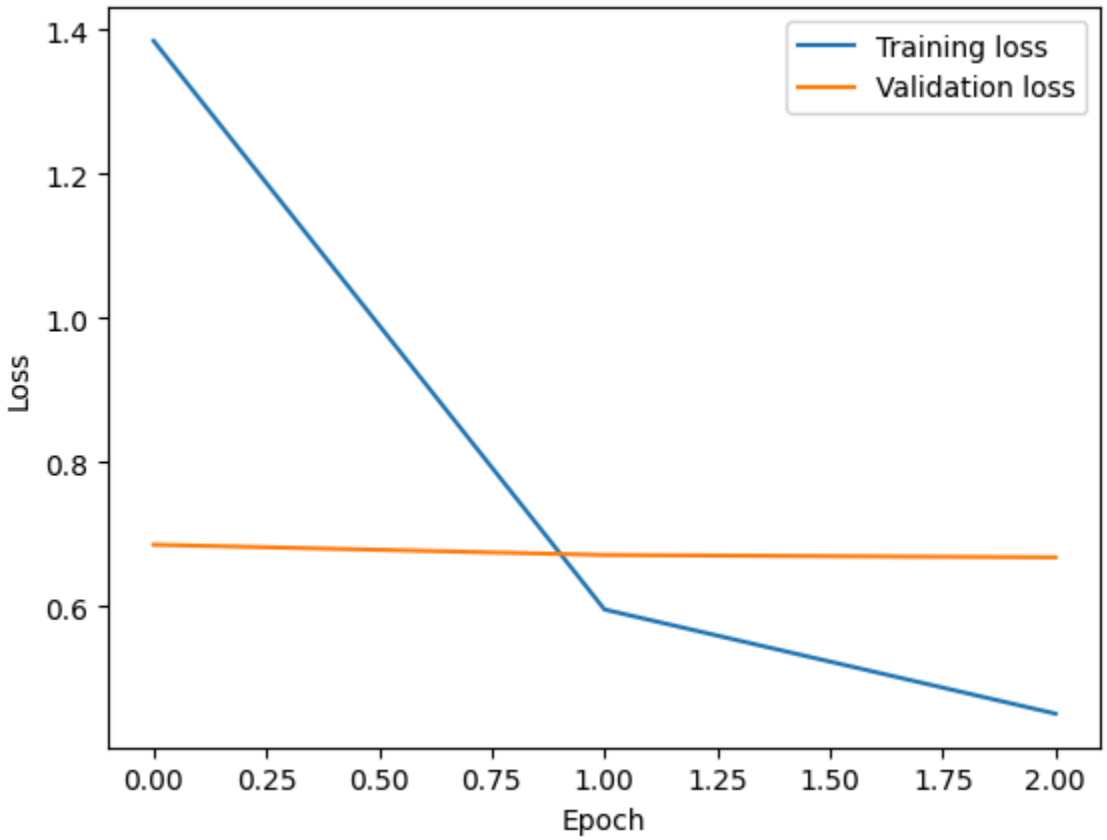
CSE556 NLP Assignment-3

Arnav Goel *2021519*
Medha Hira *2021265*
Siddharth Rajput *2021102*
Amil Bhagat *2021309*

1. TASK 1:

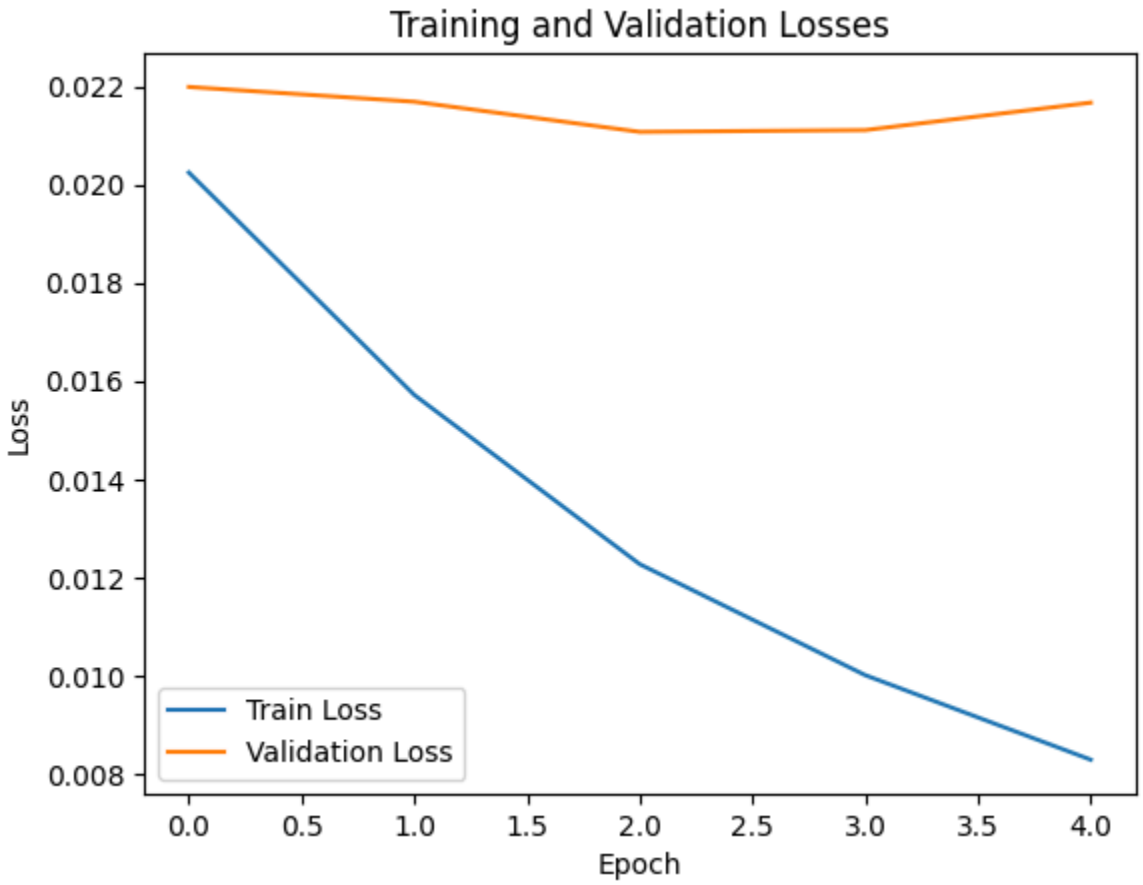
Analysis and Explanation:

Setup 1A:



- The training loss starts at a relatively high value of 1.3841 but decreases steadily with each epoch, indicating that the model is learning and improving its performance on the training data.
- The validation loss also decreases but at a slower rate and reaches a plateau after a few epochs, suggesting that the model might start overfitting to the training data.

Setup 1C:



- The training Loss starts at a much lower value of 0.0202 compared to Setup 1A, indicating that the model initialized with pre-trained weights is already well-performing.
- The validation loss starts higher than the training loss, around 0.0220, but then it drops more slowly and levels off after a few training rounds. This could mean that the model is starting to focus too much on the training data and might need to perform better on new, unseen data.

Comparison:-

- Both setups show a decrease in both training and validation losses, indicating successful learning.
- The loss curves for Setup 1C may exhibit smoother trends due to the nature of the loss function used (Cosine Similarity Loss) compared to Setup 1A, which uses Mean Squared Error (MSE) loss.
- As validation loss starts to increase or plateau while the training loss is still decreasing, it could indicate a bit of overfitting.
- The models in both setups are overfitting, likely due to the limited amount of data available for training. Transformers, with their complex architectures, typically require a large amount of data (in the order of hundreds of thousands of samples) to achieve optimal performance. The small dataset used in this scenario may not provide enough diverse examples for the model to learn robust representations, leading to overfitting.

Performance Comparison:

To provide a brief comparison and explanation for the performance differences between the three setups (1A, 1B, and 1C), we need to consider several factors:

Model Architecture:

Setup 1A: Uses a BERT-based model fine-tuned for regression tasks. It directly predicts the similarity score between two input sentences.
Setup 1B: Utilizes the Sentence-BERT architecture, which is designed specifically for semantic textual similarity tasks. Embeddings are extracted for the text, and cosine similarity between the two is reported.
Setup 1C: It also employs the Sentence-BERT architecture, but it fine-tunes the model using a custom cosine similarity loss function.

Loss Function:

Setup 1A: Employs Mean Squared Error (MSE) loss for regression.
Setup 1B: No loss function was used here as this was a zero-shot Evaluation Task.
Setup 1C: CosineSimilarityLoss was utilized here from the sentence transformers library.

Training / Experimentation Procedure:

Setup 1A: Trains the model using DataLoader with batched data and directly predicts similarity scores.
Setup 1B: Sentence-BERT Embeddings are extracted for the two given sentences, and cosine similarity between the two is reported.
Setup 1C: Fine-tunes the Sentence-BERT model using DataLoader and CosineSimilarityLoss with batched data.

Preprocessing and Feature Engineering:

Setup 1A: Applies basic text preprocessing (lowercasing, punctuation removal, lemmatization) and tokenization.
Setup 1B: Uses the SentenceTransformer library, which automatically handles tokenization and embedding extraction.
Setup 1C: Utilizes the SentenceTransformer library for fine-tuning the model using a CosineSimilarityLoss.

Evaluation Metrics:

The evaluation metrics for semantic textual similarity tasks given in the assignment was the Pearson correlation coefficient between predicted similarities and actual scores for the validation dataset:- Here's a summary of the evaluation metrics for each setup:

Setups	Pearson Correlation Coefficient
Setup 1A	0.8510
Setup 1B	0.8631
Setup 1C	0.8898

Performance Differences:

Setup 1A: Performs the worst among the three setups as a vanilla BERT model has been utilized. Objectively, the performance is optimal but can be improved.

Setup 1B: This shows improved performance compared to Setup 1A as it uses the pre-trained Sentence-BERT for our task of semantic similarity.

Setup 1C: Demonstrates the best performance among the three setups as it further fine-tunes the SBERT model on the STS task.

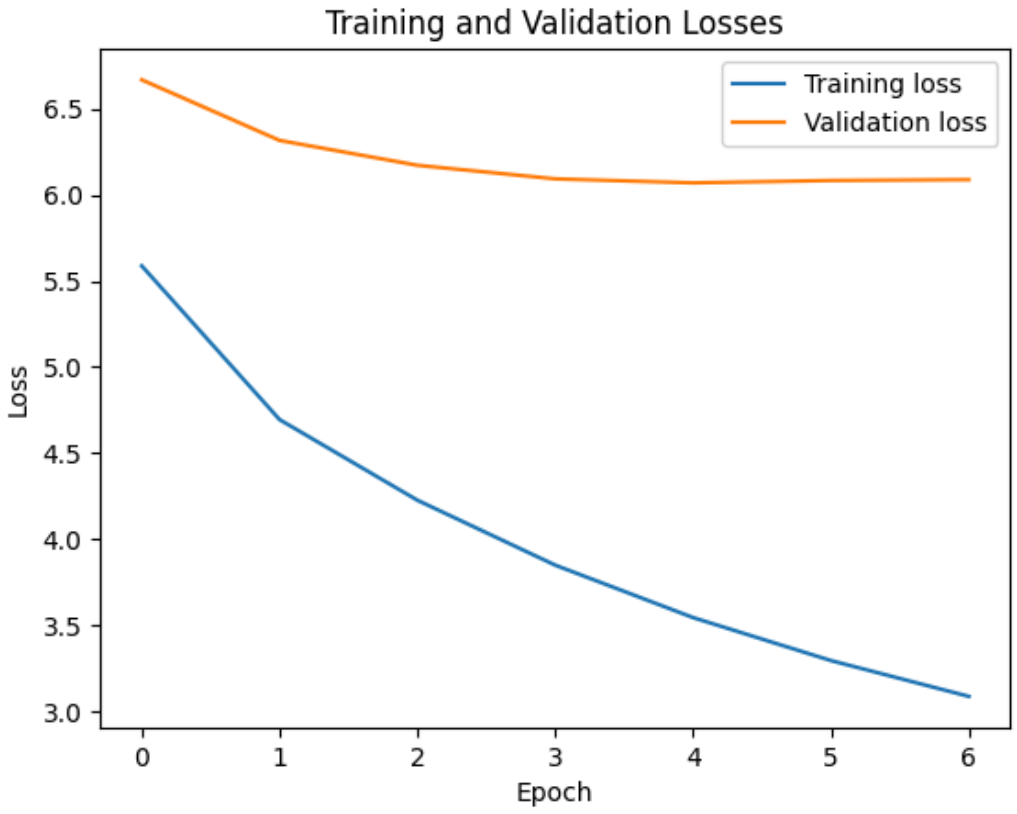
Explanation:

- Setup 1C outperforms the other setups because it customizes the loss function to directly optimize for cosine similarity, which is more aligned with the task requirements of semantic textual similarity. Additionally, fine-tuning the Sentence-BERT model with a cosine similarity loss allows it to capture semantic similarities more effectively.
- Setup 1B is a good contender as SBERT was pre-trained using a Siamese Network on a sentence similarity task and thus gives rich sentence-level embeddings.
- Setup 1A, while objectively performing well, is not able to perform as well as 1B and 1C as BERT gives word-level embeddings and lacks when it comes to sentence level representations.
- In conclusion, the performance differences between the setups can be attributed to variations in model architecture, loss function, training procedure, and preprocessing techniques. Setup 1C exhibits superior performance due to its tailored approach to the task.

2. TASK 2

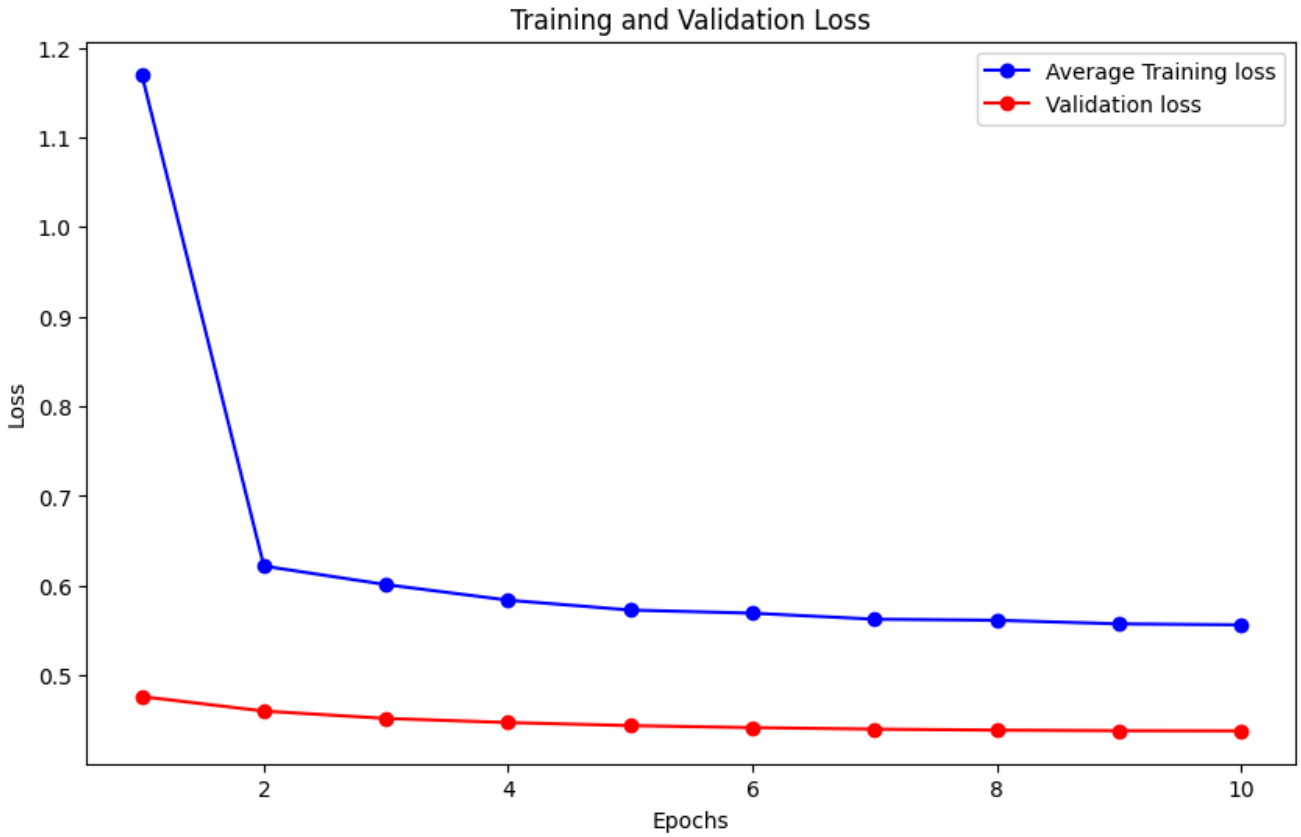
Analysis and Explanation:

Setup 2A(T5 Model without Tokenization):



- Initially, both training and validation losses decrease steadily, indicating effective learning and generalization of the model.
- Towards the end of the training, particularly in the last two epochs, the validation loss plateaus, suggesting that the model's performance stabilizes.
- This is a case of overfitting and could be attributed to the lack of data (only 50000 samples used) and the transformer being a very complex architecture causing overfitting on the training data here.

Setup 2C (T5 Model with Tokenization):



- Both training and validation losses decrease consistently over epochs, indicating effective learning and generalization of the model.
- The validation loss is consistently lower than the training loss, suggesting that the model generalizes well to unseen data. However, the validation loss plateaus after a few epochs, indicating that the model's performance stabilizes, and further training might not yield substantial improvements in validation loss. As this is a transformer-based pre-trained model, less data here, too, leads to overfitting and does not enforce a consistent drop in the validation loss.

Comparison:

- Both setups exhibit similar initial trends of decreasing loss, indicating effective learning.
- However, Setup 2A experiences a later onset of plateauing in the validation loss compared to Setup 2C, where it occurs earlier.
- This suggests that Setup 2C might reach its optimal performance level sooner compared to Setup 2A.
- The plateauing might be due to the limited amount of data, which could restrict the model's ability to learn further and improve performance as the transformer is a very complex architecture.

Performance Comparison:

Here we interpret the results:

1. Data Preprocessing

Setup 2a: No specific preprocessing steps are mentioned, implying that the data might not undergo any preprocessing before being fed into the T5 model.
Setup 2b & 2c: Both setups involve tokenization using the spaCy tokenizer. This preprocessing step ensures that the input text is broken down into individual tokens, which can improve the model's understanding of the text.

2. Model Architecture

All three setups utilize the same T5 model with its default configuration. Therefore, there are no differences in the architecture of the model across the setups.

3. Evaluation Metrics

Setup 2a, 2b, & 2c: All setups use the same evaluation metrics for assessing translation performance, including METEOR, BLEU, and BERTScore. This ensures consistency in the evaluation process across all setups. For Setup 2b, since the task at hand is to translate from English to German, BERTscore is computed using multilingual BERT by passing the language as an argument.

4. Performance Differences

Setup 2a: The METEOR scores for both validation and test data are notably lower compared to the other setups, indicating poorer translation quality. The BLEU and BERTScore metrics also suggest inferior performance.
Setup 2b: This setup demonstrates better performance than Setup 2a across all metrics, with significantly higher METEOR, BLEU, and BERTScore scores for both validation and test data. The use of tokenization and the fact that T5 is pre-trained on this task helps give better performance.
Setup 2c: Setup 2c outperforms Setup 2a in terms of BLEU, METEOR, and BERTScore metrics for both validation and test data. Incorporating a pre-trained model like T5 and fine-tuning helps us, but overfitting was also observed here due to only 50,000 samples being used for fine-tuning the model.

5. Conclusion

- Setup 2c achieves the highest translation quality among the three setups, followed by Setup 2b. Setup 2a lags significantly, underscoring that vanilla transformers need high quantities of data and memory power to be able to learn properly. A vanilla transformer being trained on only 50000 samples overfits easily due to its complex architecture.
- The consistent improvement in performance from Setup 2a to Setup 2c underscores the importance of pre-trained models, particularly in improving the model's understanding of complex language structures like German.
- While the choice of evaluation metrics remains consistent across all setups, the incorporation of preprocessing techniques significantly influences translation quality, as evidenced by the variations in performance observed across the three setups.

Evaluation Metrics:

Results on Validation Data:-

Setups	BLEU-1 Score	BLEU-2 Score	BLEU-3 Score	BLEU-4 Score	METEOR Score	BERT Score Precision:	BERT Score Recall	BERT Score-F1
Setup 2a	0.146	0.051	0.015	0.005	0.0789	0.7648	0.8221	0.7922
Setup 2b	0.266	0.201	0.157	0.126	0.2681	0.8268	0.7662	0.7942
Setup 2c	0.409	0.264	0.178	0.124	0.3576	0.9028	0.9080	0.9053

Results on Test Data:-

Setups	BLEU-1 Score	BLEU-2 Score	BLEU-3 Score	BLEU-4 Score	METEOR Score	BERT Score Precision:	BERT Score Recall	BERT Score-F1
Setup 2a	0.132	0.044	0.013	0.0045	0.0731	0.7632	0.8191	0.7899
Setup 2b	0.270	0.210	0.168	0.137	0.2879	0.8345	0.7703	0.800
Setup 2c	0.441	0.312	0.213	0.151	0.3913	0.9049	0.9115	0.9081

Credit Statement:

While each member contributed equally to all tasks, we are mentioning a division of tasks just for the sake of the question here.

- **Medha Hira (Roll No: 2021265)** - Setup 1A, Setup 1B
- **Arnav Goel (Roll No: 2021519)** - Setup 2A, Setup 2B
- **Siddharth Rajput (Roll No: 2021102)** - Setup 1C, Report and Analysis
- **Amil Bhagat (Roll No: 2021309)** - Setup 2C, Report and Analysis