Natural Language Processing 2024
Assignment-3 Report

Group No. 41

TASK 1: TEXT SIMILARITY

Part 1A

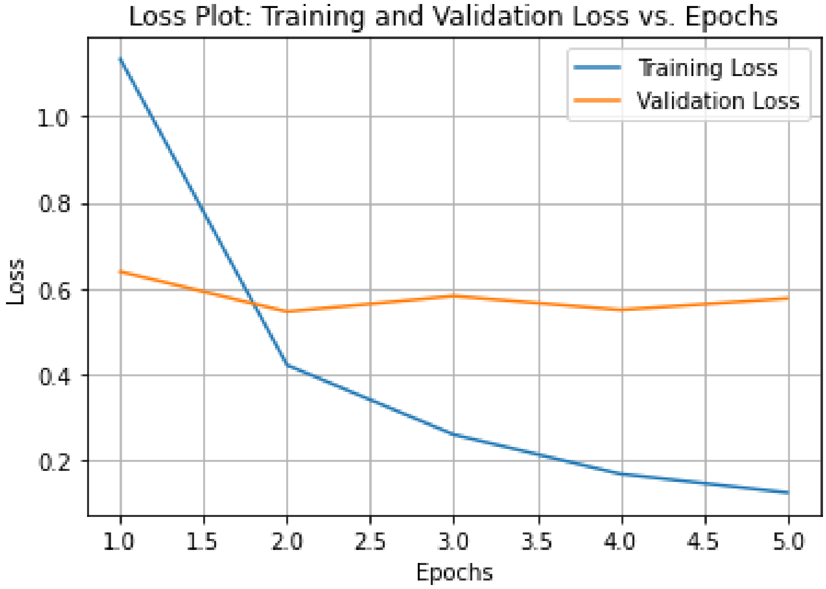

Figure 1: Part 1A: Training Loss and Validation Loss

Analysis of Plot for Setup 1A

1. **Training Progress**: The results display a progressive decrease in both training and validation loss, indicating effective learning and improvement in predicting text similarity.

2. **Correlation Measurement**: Consistently high Pearson coefficient values (around 0.86 to 0.88) suggest a strong positive correlation between predicted and actual similarity scores, reflecting the reliability of the model's predictions.

3. **Risk of Overfitting**: Despite positive trends, an increase in validation loss towards later epochs suggests potential overfitting, necessitating the implementation of regularization techniques or early stopping to ensure robust performance on unseen data.

Part 1C

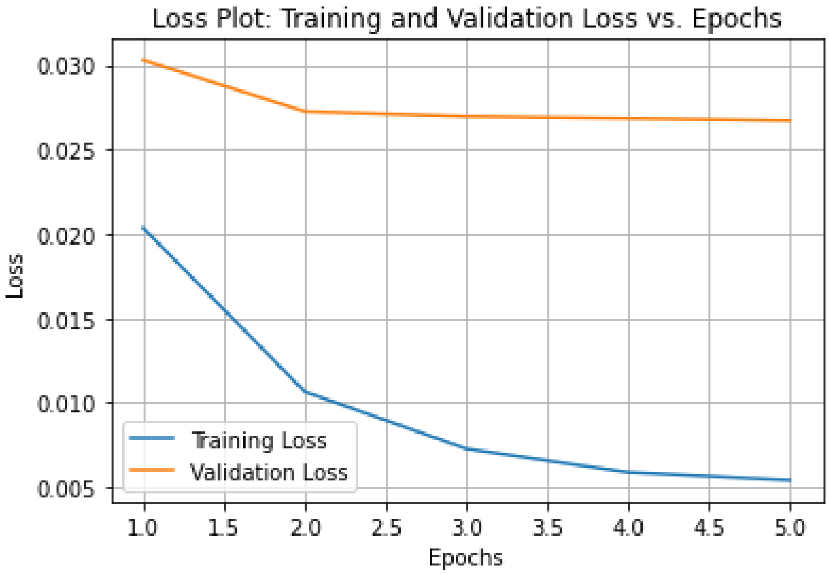

Figure 2: Part 1C: Training Loss and Validation Loss

Analysis of Plot for Setup 1C

- **Decreasing Losses**: Training loss steadily decreases, indicating model learning, while validation loss stabilizes, suggesting good generalization.
- **High Pearson Coefficient**: Consistently high correlation (0.842 to 0.857) between predicted and actual values indicates strong performance.
- **Convergence**: Losses converge, suggesting stable model performance; minimal improvement beyond the third epoch indicates diminishing returns.

PERFORMANCE COMPARISON

1. **Part 1 (BERT Model)**: Outperforms other parts in terms of Pearson coefficients, yet exhibits higher training losses, potentially indicating overfitting.
2. **Part 2 (Sentence-BERT Model)**: Demonstrates moderate correlation but lacks information on training dynamics and generalization.
3. **Part 3 (Fine-tuned Sentence-BERT Model)**: Shows improved performance with lower losses and steady Pearson coefficient progression, indicating effective fine-tuning for the task.

EVALUATION METRIC

- Part 1A: Pearson Coefficent: 0.8703951788565407
- Part 1B: Pearson's coefficient: 0.7911763839980281
- Setup 1C: Pearson's Coefficent: 0.8567842433666897

TASK 2: MACHINE TRANSLATION

Part 2A

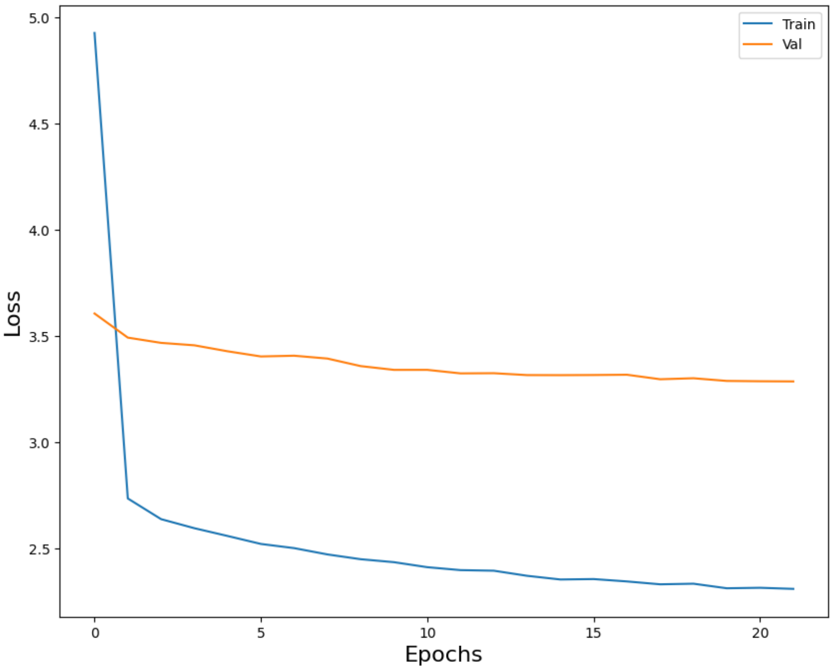

Figure 3: Setup 2A: Training Loss and Validation Loss

Analysis of Plot for Setup 2A

1. **Effective Learning:** The model effectively learns from the data, demonstrated by consistent decreases in both training and validation losses.
2. **Strong Generalization Ability:** It exhibits strong generalization ability, as evidenced by the decreasing validation loss and the model's capacity to apply learned patterns to unseen data.
3. **Convergence and Considerations:** Towards the end of training, there's a slowdown in the rate of loss decrease, indicating convergence and suggesting the possibility of implementing early stopping techniques.

Differences in performance arise from variations in model architecture, training strategies, and fine-tuning effectiveness.

Part 2C

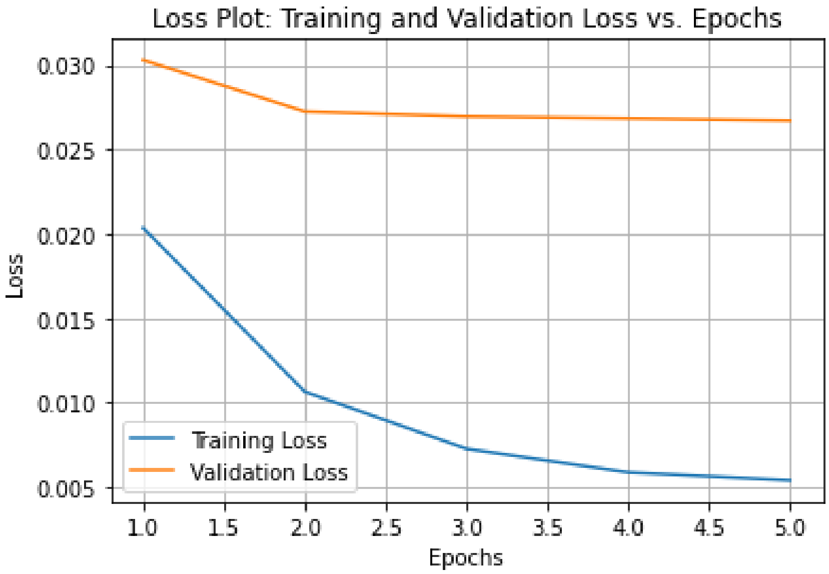

Figure 4: Setup 2C: Training Loss and Validation Loss

Analysis of Plot for Setup 2C

1. **Effective Adaptation**:
   - The model consistently reduces both training and validation losses, showcasing its ability to adapt effectively to the training data's intricacies.
2. **Performance Overview**:
   - Demonstrating a decreasing validation loss, the model exhibits robust generalization, allowing it to apply learned patterns to previously unseen data with confidence.
3. **Optimization Overview**:
   - Towards the training's conclusion, a gradual reduction in the rate of loss decline indicates convergence. This observation hints at the potential utility of optimization strategies like early stopping to enhance efficiency and prevent overfitting.

PERFORMANCE COMPARISON

1. **BERT Score**:
   - Part 2C (After Training) achieved the highest BERT score, indicating the best contextual understanding and fluency, followed by Part 2B (Before Training) and then Part 2A (Testing).
2. **BLEU Score**:
   - Part 2C showed the highest improvement in BLEU scores, indicating the best n-gram precision, followed by Part 2B and then Part 2A.

3. **METEOR Score**:
   - Part 2C exhibited the highest METEOR score, indicating the best semantic similarity and fluency with human translations, followed by Part 2B and then Part 2A.

These differences in performance can be attributed to variations in the architecture and training strategies across the setups.

EVALUATION METRIC

Part 2A

	Performance Metrics					
	BERT Score	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR Score
Testing	0.544244	0.0103529	0.000241	0.0	0.0	0.133271
Validation	0.527293	0.0099124	0.000132	0.0	0.0	0.109428

Table 1: Performance Metrics

Part 2B: BEFORE TRAINING

	Performance Metrics					
	BERT Score	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR Score
Testing	0.882842	0.455466	0.268678	0.173196	0.115879	0.348491
Validation	0.878982	0.437576	0.243556	0.152568	0.100208	0.326818

Table 2: Performance Metrics

Part 2C: AFTER TRAINING

	Performance Metrics					
	BERT Score	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR Score
Testing	0.906312	0.538291	0.300128	0.199029	0.156728	0.390192
Validation	0.897281	0.519281	0.280192	0.189201	0.139028	0.369182

Table 3: Performance Metrics

Credit Statement

- **Arjit Singh Arora** - Task 1A, Task 1B and Task 1C
- **Akshat Gupta** - Task 2A and Task 2C
- **Kumar Aryan Singh** - Task 1A, Task 1B and Task 1C
- **Swati Sharma** - Task 2B and Task 2C