SMART-SBERT with SimCSE: A Robust Defense Against Catastrophic Forgetting in BERT

Submitted by:

Rooshan Khan 2021-EE-067
Hussnain Amjad 2021-EE-063
Abdul Samad 2021-EE-191
Areesha Noor 2021-EE-103

Supervised by:

Dr. Irfan Ullah Chaudhry

Abstract

We aim to build a robust model based on the BERT-base architecture that performs well on three NLP tasks—sentiment classification (SST-5), paraphrase detection (QQP), and semantic textual similarity (STS-B)—by training sequentially across tasks. A core challenge in this setup is *catastrophic forgetting*, where learning a new task can degrade performance on previously learned tasks due to destructive gradient updates.

To mitigate this, we implemented the **SMART algorithm**, which introduces adversarial regularization to preserve previously acquired knowledge and avoid aggressive parameter updates. This component was implemented by Rooshan Khan. Abdul Samad and Areesha contributed contrastive learning using supervised and unsupervised loss respectively. However, we used the best among these two that is supervised SimCSE loss to maximize cosine similarity between semantically similar sentence pairs and minimize it for dissimilar ones. Hussnain Amjad implemented the SBERT architecture, for task-specific classification and regression heads for QQP and STS-B, respectively.

We got two best models:

- **SMART SBERT (SS)**
- **SMART SBERT with SimCSE (SSS)**

SS performed better on the SST-5 task, while **SSS** showed superior results on the STS-B task due to the SimCSE loss function being optimized for similarity learning. Both models performed equally well on paraphrase detection. Results are shown below:

Task	**Metric**	**Score (SS)**
Sentiment Classification	Accuracy	0.537
	F1 Score	0.528
Paraphrase Detection	Accuracy	0.864
	F1 Score	0.864
Semantic Textual Similarity	Pearson Correlation	0.819
Overall Performance		**0.770**

$$\text{Overall performance (SS)} = \frac{\left(\frac{0.819+1}{2} + 0.537 + 0.864\right)}{3} = 0.770$$

Task	**Metric**	**Score (SSS)**
Sentiment Classification	Accuracy	0.507
	F1 Score	0.494
Paraphrase Detection	Accuracy	0.864
	F1 Score	0.863
Semantic Textual Similarity	Pearson Correlation	0.843
Overall Performance		**0.764**

$$\text{Overall performance (SSS)} = \frac{\left(\frac{0.843+1}{2} + 0.507 + 0.864\right)}{3} = 0.764$$

Depending on task priority, either SS or SSS may be selected. Our results show that a combination of SMART regularization and contrastive learning can effectively reduce catastrophic forgetting and achieve strong performance across all three tasks in a sequential training setup.