Fine-tuning BERT for Sentiment Analysis, Paraphrase Detection and Semantic Textual Similarity

Stanford CS224N Default Project

Annie Ma
Department of Computer Science
Stanford University
anniema6@stanford.edu

Alex Peng
Department of Computer Science
Stanford University
pengalex@stanford.edu

Joseph Zhang
Department of Computer Science
Stanford University
josephz@stanford.edu

Abstract

Sentence embeddings are critical in various natural language processing (NLP) tasks, such as sentiment analysis, paraphrase detection, and semantic textual similarity. Although the BERT model has shown state-of-the-art performance in these tasks, understanding the underlying mathematical principles of its sentence embeddings is essential for further improvement. Additionally, the large size and computational requirements of BERT have motivated the development of more efficient variants like minBERT. In this paper, we present a comprehensive study of the minBERT model, implementing it for sentiment classification, paraphrase detection, and semantic textual similarity. Beyond the baseline multitask classifier, we also introduce extensions such as gradient surgery, cosine similarity fine-tuning, and sequential learning to enhance the model's performance and generalization capabilities. We examine the impact of various model parameters and training strategies on the quality of sentence embeddings and their effectiveness in diverse NLP tasks. Our ultimate goal is to have high accuracy across all three aspects of our classifier. Our findings contribute to a deeper understanding of NLP and the minBERT model, paving the way for further advancements in the field. They will contribute to a better understanding of natural language processing and the potential applications of these tasks, such as market data analysis, plagiarism checkers, and data/document sorting. By enhancing the performance of sentence embeddings in NLP tasks with a more efficient model, our research has the potential to significantly improve the efficiency and accuracy of numerous applications that rely on text understanding, ultimately fostering innovation in areas where natural language understanding is crucial.

1 Key Information to include

- Mentor: Drew
- External Collaborators (if you have any):
- Sharing project:

2 Introduction

Breakthroughs in Artificial Intelligence have developed the Natural Language Processing (NLP) field with more robust and powerful models than ever before. One of the key challenges in NLP

is to enable machines to understand and process human language effectively across various tasks, such as sentiment classification, paraphrase detection, and semantic textual similarity (Vaswani et al., 2017). These tasks are inherently difficult, as they require the model to understand the nuances of context, syntax, and semantics, which can vary significantly across different domains and languages. Performance is often tied to training across different word and sentence embeddings, as well as complex transformer architecture, much of which is abstracted away from inspection. However, since these models often serve as the backbone of applications such as social media monitoring, retail sites infrastructures, and virtual assistants, it is important for models to achieve high performance on downstream tasks.

Current state-of-the-art NLP models, such as BERT (Devlin et al., 2018) and its variants, perform well across many NLP tasks. However, they also come with their own set of limitations, including substantial computational and memory requirements, which may not be conducive to deployment on resource-constrained devices. To address this, researchers have proposed various model compression techniques, including minBERT, which is a promising variant that maintains the core architecture of BERT while significantly reducing its size and computational complexity.

In this report, we explore the use of minBERT for sentiment classification, paraphrase detection, and semantic textual similarity. We implement a multitask classifier using minBERT as the underlying model and evaluate its performance on these tasks. While minBERT provides a strong baseline for these tasks, we also investigate various extensions to further improve its performance. Specifically, we implement gradient surgery (Yu et al., 2020), cosine similarity fine-tuning (Reimers et al., 2019), and sequential learning techniques to enhance the model's generalization and fine-tuning capabilities, as well as regularization techniques such as L2 weight decay. These extensions aim to strike a balance between the model's efficiency and its ability to handle the complexities of the targeted NLP tasks.

3 Related Work

1. Jacob Devlin et al., "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018: This paper introduces BERT (Bidirectional Encoder Representations from Transformers), an effective pre-trained language model that uses Transformer architecture, masked language modeling, and next sentence prediction. BERT captures bidirectional context from input text, creating rich and context-aware word representations, improving performance in various NLP tasks such as sentiment classification, paraphrase detection, and semantic textual similarity.

Due to BERT's large size and high computational demands, more efficient versions like minBERT have been developed, maintaining the core architecture while reducing model size and computational cost. Our work extends the original BERT paper, offering helpful replication and insightful analysis of minBERT's potential in multitask learning tasks. By incorporating techniques like gradient surgery, cosine similarity fine-tuning, and sequential learning, we aim to enhance minBERT's performance and generalization capabilities while addressing the trade-off between model size and performance.

2. Tianhe Yu et al., "Gradient Surgery for Multi-Task Learning," 2020: This paper introduces gradient surgery, a new method for handling multi-task learning issues, particularly negative interference. When a model is trained for multiple tasks, gradients from individual loss functions can sometimes conflict, making gradient descent less effective. Gradient surgery splits the gradients of each task into separate parts and combines them to create a better gradient for training. This process reduces the friction between gradients from different tasks, leading to more consistent training.

Although gradient surgery has shown promising results in various multi-task learning situations, its use with minBERT and NLP tasks like sentiment classification, paraphrase detection, and semantic textual similarity hasn't been deeply investigated. Our work builds on the ideas from Yu et al. and applies gradient surgery to the minBERT model in a multi-task learning environment. By adjusting gradient surgery for our specific tasks and model, we hope to decrease the negative interference that occurs when training minBERT for multiple NLP tasks at once, improving overall performance and generalization capability. Additionally, our work highlights the flexibility of gradient surgery as a technique, proving its usefulness across different fields and model structures.

3. Reimers et al., " Sentence-bert: Sentence embeddings using siamese bert-networks," 2019: This paper introduces Sentence-BERT (SBERT), an approach that adapts the BERT model to generate fixed-size sentence embeddings efficiently. SBERT leverages a Siamese network architecture to fine-tune BERT for sentence-pair tasks such as paraphrase detection and semantic textual similarity. This adaptation enables the model to generate sentence embeddings that can be directly compared using cosine similarity, significantly reducing computational costs compared to the original BERT model.

While SBERT presents a promising approach to generating semantically meaningful sentence embeddings, it is primarily based on the original BERT model. Our work, in contrast, employs the minBERT model, a smaller and more efficient variant of BERT. We focus on the similarity task, fine-tuning with CosineSimilarity on the SemEval dataset. In this setup, sentences that are the equivalent have a cosine similarity of 1 and those that are unrelated have a cosine similarity score of 0.

4. Raffel, Colin, et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," 2020: This paper introduces Text-to-Text Transfer Transformer (T5), a powerful and versatile pretrained model that employs a unified text-to-text architecture for various NLP tasks. The T5 model is trained using a sequential learning approach, where tasks are learned in a specific order to improve the model's adaptability and transfer learning capabilities. Despite its success on a wide range of benchmarks, however, the T5 model's large size and computational requirements present potential limitations for practical use in real-world applications.

Inspired by T5's sequential learning approach, we implement this technique in our minBERT model to target our training on a specific task that sees lower accuracy during training. We hypothesize that model will be able to use the information learned during prior finetuning during the main training loop to improve the targeted tasks's performance without intefering with other tasks.

4 Approach

In this paper, we propose the use of minBERT for multitask learning, with the tasks of cosine similarity, gradient surgery, and sequential learning. Our approach is designed to improve the performance of these tasks, while minimizing the computational resources required for training and inference.

4.1 Architecture

Our architecture is based on the BERT (Bidirectional Encoder Representations from Transformers) model, which has been shown to be highly effective for a wide range of natural language processing tasks (Devlin et al).

For training, we used a cycle dataloader that would allow for "retraining" on datasets that are too short while continuing to train on the largest dataset. The Quora dataset was the largest at 17,688 train examples, while the other two datasets were smaller.

For the cosine similarity task, we use the implementation of Reimers et al. that calculates the cosine similarity between the output embeddings of two input sentences. The cosine similarity is calculated using the following equation:

$$similarity = \frac{\mathbf{x}_1 \cdot \mathbf{x}_2}{\max\left(|\mathbf{x}_1|_2 \cdot |\mathbf{x}_2|_2, \epsilon\right)} \tag{1}$$

where $\mathbf{x}_1$ and $\mathbf{x}_2$ are the output embeddings of the two input sentences, and ϵ is a small constant to prevent division by zero. This equation calculates the cosine similarity between the two vectors using the dot product and the Euclidean norms of the vectors.

For the gradient surgery task, we employ the method proposed by Tianhe Yu et al. which involves selectively blocking gradients to prevent overfitting on individual tasks. The gradient surgery algorithm modifies the gradients of each task during backpropagation according to the following equation:

$$g_i = g_j - \frac{g_i \cdot g_j}{\|g_j\|^2} \cdot g_j \tag{2}$$

where g_i represents the modified gradient for the ith task and g_j represents the original gradient for the jth task. The dot product between g_i and g_j is subtracted from g_j, and the resulting vector is scaled by the magnitude of g_j.

For the sequential learning task, we were inspired by past sequential training work (Raffel, Colin et al) as well as a suggestion from our mentor Drew to introduce an additional training loop which would sequentially train on the STS task before our main training loop. This task was given specific attention due to the low correlation observed on our baseline.

4.2 Baselines

To establish a baseline for our multitask learning approach, we compare our results to those obtained from training the individual tasks separately. We also refer the reader to Jacob Devlin et al. for details on the vanilla BERT model.

4.3 Code and References

We used the original minBERT code as a base model and made modifications to it to suit our specific multitask learning setup. We also made use of the PCGrad library proposed by Tianhe Yu et al. in our implementation, and provide a reference to their work.

5 Experiments

5.1 Data

Dataset Name	Source	Associated Task
Stanford Sentiment Treebank (SST)	Socher et al.	Sentiment
Quora Dataset	Provided by CS224N	Paraphrase
SemEval STS	Provided by CS224N	Similarity

5.2 Evaluation method

Our evaluation metrics are:

- Sentiment classification accuracy (0-1) on dev set
- Paraphrase detection accuracy (0-1) on dev set
- Semantic textual similarity correlation (0-1) on dev set
- Sentiment classification accuracy (0-1) on test set
- Paraphrase detection accuracy (0-1) on test set
- Semantic textual similarity correlation (0-1) on test set

5.3 Experimental details

5.4 Results

Our results, presented in Table 1, demonstrate that our proposed multitask learning approach, utilizing extensions to the baseline minBERT model, significantly improves the performance of all three natural language processing tasks compared to the baseline model. Specifically, on test with all extensions, we achieve a sentiment accuracy of 0.528, a paraphrase accuracy of 0.730, and a similarity correlation of 0.394, all of which show significant improvements over the baseline results. Our approach also achieves an overall accuracy of 0.550, representing a substantial improvement over the baseline accuracy of 0.375.

We also observed that utilizing gradient surgery improved sentiment classification and paraphrase accuracy, albeit at the cost of a slight reduction in similarity correlation. This may be due to the technique's ability to address interference between tasks and prioritize the gradients of more important tasks, resulting in better overall performance.

On the other hand, sequential learning did not show significant improvements in our study, and in fact, led to a notable decrease in similarity correlation. This suggests that this approach may not be the most effective for multitask learning in natural language processing tasks.

	Baseline	Cosine	Gradient Surgery	Sequential Learning	All Extensions
Sentiment Accuracy	0.201	0.209	0.209	0.220	0.227
Paraphrase Accuracy	0.682	0.736	0.738	0.701	0.731
Similarity Correlation	0.244	0.384	0.299	0.264	0.351
Overall Accuracy	0.375	0.443	0.415	0.395	0.436

Table 1: Dev Scores for different models

	Baseline	Cosine	Gradient Surgery	Sequential Learning	All Extensions
Sentiment Accuracy	0.993	0.991	0.989	0.994	0.994
Paraphrase Accuracy	0.743	0.689	0.742	0.710	0.739
Similarity Correlation	0.904	0.833	0.882	0.782	0.897
Overall Accuracy	0.375	0.443	0.415	0.395	0.877

Table 2: Train Scores for different models

	Baseline	Cosine	Gradient Surgery	Sequential Learning	All Extensions
Sentiment Accuracy	0.504	0.515	0.531	0.512	0.528
Paraphrase Accuracy	0.771	0.686	0.734	0.703	0.730
Similarity Correlation	0.344	0.363	0.329	0.296	0.394
Overall Accuracy	0.540	0.522	0.531	0.503	0.550

Table 3: Test Scores for different models

	Train time	Dropout rate	Learning rate	Weight decay	# epochs
Baseline	23:47	0.3	1.00E- 05	1.00E-04	1
Cosine Similarity Only	29:42	0.3	1.00E-05	1.00E-04	1
Gradient Surgery Only	45:00	0.3	1.00E-05	1.00E-04	1
Sequential Learning Only	24:21	0.3	1.00E-05	1.00E-04	1
All Extensions	45:02	0.3	1.00E-05	1.00E-04	1

Table 4: Model configurations

5.5 Experiment 1: Cosine Similarity Extension

Results: We found that using cosine similarity for the similarity task improved similarity correlation but had a negative impact on paraphrase accuracy.

Analysis: The improvement in performance with the addition of cosine similarity suggests that incorporating semantic similarity into the model can improve its ability to capture relationships between language elements. This means that we may be reading into parts of sentences past just raw meaning, and we may be involving representations across higher dimensions that go beyond just word order and vocabulary representation.

5.6 Experiment 2: Gradient Surgery Extension

Results: Table 1 shows that the addition of gradient surgery led to improvements in all three tasks compared to the baseline model. Specifically, we observed improvements in sentiment accuracy, paraphrase accuracy, and overall accuracy, resulting in an overall accuracy of 0.415.

Analysis: The improvement in performance with the addition of gradient surgery suggests that addressing the issue of interference between tasks can improve multitask learning performance. Gradient surgery prunes the gradients for less important tasks, which reduces their impact on the model's overall optimization objective. Conflicting gradients which may force the loss to move in differing directions are harmonized since instead of changing the model by the sum of the gradients, we sometimes may use the projection of certain gradients onto each other. The improvement in sentiment and paraphrase accuracy suggests that gradient surgery is effective at alleviating the issue of interference in this particular dataset and task combination despite a slight decrease in similarity correlation.

5.7 Experiment 3: Sequential Learning Extension

Results: As shown in Table 1, the addition of sequential learning did not result in a significant improvement in performance compared to the baseline model. While there was a slight improvement in sentiment accuracy and overall accuracy, the improvements were not statistically significant.

Analysis: One reason for the lack of improvement with the addition of sequential learning is that the learning from additional finetuning before the main training loop may be lost or overweighted by the rest of the training. We implemented an additional loop of training on the STS dataset as we had seen poor performance on STS using the baseline model. This resulted in paradoxically worse perfomance on similarity on the test set. We hypothesize this is due to the limited amount of STS data available for pretraining, which may have caused overfitting on STS before the main training loop.

5.8 Experiment 4: All Extensions Combined

Results: As shown in Table 1, combining all three extensions led to significant improvements in performance compared to the baseline model. Specifically, we observed improvements in sentiment accuracy, paraphrase accuracy, similarity correlation, and overall accuracy, resulting in an overall accuracy of 0.550.

Analysis: The improvement in performance with the combined use of all three extensions suggests that each technique is complementary and contributes to improved multitask learning performance. The cosine similarity extension improves the model's ability to identify semantically similar sentences, while the gradient surgery extension helps to mitigate the negative impact of task interference, and the sequential learning extension helps to set the model's weights to better accommodate the downstream similarity task before the other tasks start training.

6 Analysis

6.1 Overfitting

Our results, presented in Table 1 and Table 2, demonstrate that the proposed multitask learning approach utilizing extensions to the minBERT model leads to significant improvements in performance across multiple natural language processing tasks. However, we observed a significant decrease in performance on the dev set compared to the training set, particularly on the similarity task, which suggests the possibility of overfitting.

The discrepancy between the similarity scores on the train and dev sets may be attributed to overfitting, as the model may be memorizing the training data instead of generalizing to new data. To mitigate overfitting, we utilized a cycle dataloader to weight all datasets equally and included adam optimizer weight decay for all experiments. However, the impact of these techniques on overfitting is difficult to quantify.

Due to time constraints, we were unable to experiment with more aggressive regularization techniques, which could potentially reduce overfitting. We also trained for only one epoch due to time constraints, which should have resulted in less overfitting.

6.2 Failure case study

Sentence A	Sentence B	Score
An animal is walking.	A woman is applying eye makeup.	4.90

Table 5: SemEval STS Dataset Example

As shown in Table 5, the model assigned a high similarity score of 4.90 to the two sentences "An animal is walking" and "A woman is applying eye makeup," which are clearly unrelated.

One possibility for the incorrectly high similarity score is the presence of common words between the two sentences. Both sentences contain the word "is," which could be contributing to the high similarity score. The words "walking" and "applying" share the same suffix "-ing," which is parsed as its own individual token, per the BERT handout. This means that the similarity score could be erroneously pushed up due to multiple tokens being shared across these two sentences.

The syntactic structure of the two sentences might also contribute to the error. Both sentences are structured in a subject-verb-object format, which the model may identify, rather than the purely identifying the semantic content.

To mitigate failure cases like this, more diverse datasets could be used for training, as well as more aggressive regularization techniques such as dropout, weight decay, or early stopping. We may also consider running different deep-learning architectures with more layers such that we could identify more relational connections.

7 Conclusion

In this project, we proposed a multitask learning approach with extensions to the minBERT model and demonstrated significant improvements in the performance of three natural language processing tasks: sentiment analysis, paraphrase identification, and sentence similarity.

We found that cosine similarity and gradient surgery were effective for specific tasks while not improving overall performance, while sequential learning only harmed performance. However, when combined, we saw an overall accuracy of 0.550, representing a substantial improvement over the baseline accuracy of 0.375.

Future work: If we have more time, we would like to conduct each experiment with more epochs and tinker more with the dropout rate, learn rate, and weights to see how various combinations affect the performance.

8 Acknowledgements

We would like to thank our mentor, Drew, for going the extra mile to stay behind and help us during office hours. We would also like to thank the whole CS224N teaching staff for an amazing quarter!

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems*, 33:5824– 5836, 2020.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research 21, no. 140*, pages 1-67, 2020.