Utilizing minBERT for Multiple Sentence-level Tasks

Stanford CS224N Default Project

Qianhui Zheng
Department of Biomedical Data Science
Stanford University
zhengqh@stanford.edu

Abstract

In the rapidly evolving field of natural language processing, the development of efficient and versatile models remains a critical goal. This project addresses these demands by using a single, multi-task BERT model, which aims to maximize computational efficiency, enhance model generalization, enable effective transfer learning, and simplify deployment in production environments. We leverage a pre-trained BERT model for a spectrum of sentence-level tasks, including sentiment analysis, paraphrase detection, and semantic textual similarity. The first part of this project focuses on completing and finetuning a pre-trained BERT model to conduct sentiment analysis on the SST and the CFIMDB datasets. The subsequent part focuses on engineering innovative improvements and experimenting with various techniques to refine and adapt BERT for optimal performance across multiple tasks. These techniques include finetuning on additional datasets and projecting conflicting gradients (PCGrad). While fine-tuned exclusively with the SST dataset, the last-linear-layer BERT model and the full BERT model achieved development scores of 0.387 and 0.476 respectively on multi-task learning. However, with simultaneous fine-tuning on the SST, Quora, and SemEval datasets, combined with the integration of PCGrad, the full BERT model attained an improved overall development score of 0.652. Through systematic experimentation, this project seeks to explore effective strategies that could contribute to the broader applicability of transformer models in diverse NLP applications.

1 Introduction

In the rapidly advancing domain of natural language processing (NLP), the challenge of building models that are both efficient and versatile is more pertinent than ever. As the scope and scale of language-based applications grow, so does the demand for models that can adapt to a variety of tasks while maintaining computational and deployment efficiency. The complexity of NLP tasks such as sentiment analysis, paraphrase detection, and semantic textual similarity requires models not only to understand superficial textual features but also to capture deeper semantic meanings. Traditional approaches often involve training separate models for each task, leading to significant resource consumption and inefficiency in deploying these models in real-world applications. Furthermore, these models often fail to generalize across different tasks, a limitation that multi-task learning seeks to overcome.

This project attempts to resolve these demands through the innovative use of a single, multi-task BERT (Bidirectional Encoder Representations from Transformers) model. The versatility of BERT, initially introduced by Devlin et al. (2019), has been widely recognized for its deep contextual representations, but its adaptation to multiple NLP tasks without compromising performance remains an ongoing challenge in the field. Our approach utilizes a pre-trained BERT model, which we fine-tune to simultaneously address multiple sentence-level tasks, including sentiment analysis, paraphrase detection, and semantic textual similarity, thus maximizing computational efficiency, enhancing model generalization, and simplifying production deployment.

The first phase of this project focuses on implementing and fine-tuning a pre-trained BERT model on the SST and CFIMDB datasets for sentiment analysis. This step is crucial as it lays the foundation for understanding how well the pre-trained model can be adapted to specific tasks with minimal adjustments. The subsequent phase of the project involves engineering innovative improvements aimed at refining and adapting BERT for optimal performance across various downstream tasks. Key techniques utilized include additional fine-tuning phases and the use of Projecting Conflicting Gradients (PCGrad) to manage conflicting updates during training. Initial experiments that solely utilized the SST dataset reveal that while the BERT model fine-tuned on its last linear layer (the last-linear-layer BERT model) achieved a development score of 0.387 for multiple sentence-level tasks mentioned afore, the fully fine-tuned BERT model (the full BERT model) enhanced it to 0.476. Furthermore, with simultaneous fine-tuning on the SST, Quora, and SemEval datasets, coupled with the integration of PCGrad, the full BERT model achieved a significantly improved overall development score of 0.652.

This project not only highlights the potential of BERT in a multi-task framework but also provides insights into effective strategies for enhancing the versatility and performance of transformer models in diverse NLP applications. By systematically experimenting with different fine-tuning and optimization techniques, we aim to contribute to the broader applicability of transformer-based models in the field of natural language processing.

2 Related Work

Multi-task learning has been pivotal in enhancing the performance and generalization of models across various tasks. Previous work has shown that sharing representations among related tasks can lead to better performance compared to training separate models (Caruana (1997)). Building on this foundation, multi-task learning has been extended to transformer-based models, highlighting that joint training on multiple tasks can significantly improve efficiency and robustness (Liu et al. (2019)). These insights are fundamental to our research, where we employ a multi-task BERT model to streamline training and enhance generalization across diverse NLP tasks.

Concurrent fine-tuning on multiple datasets is a strategy that further strengthens the capabilities of multi-task learning. By fine-tuning models on diverse datasets simultaneously, researchers have been able to achieve more robust and versatile models. In this project, we adopt this approach by concurrently fine-tuning our BERT model on the SST, Quora, and SemEval datasets. This method not only improves the model's performance on individual tasks but also enhances its overall adaptability.

A critical challenge in multi-task learning is managing gradient conflicts that arise from optimizing multiple tasks simultaneously. Gradient conflicts can lead to suboptimal convergence and hinder model performance. To address this issue, Yu et al. (2020) proposed Projecting Conflicting Gradients (PCGrad) as a solution to mitigate the negative effects of conflicting gradients during the simultaneous optimization of multiple objectives. PCGrad projects conflicting gradients onto a plane where they can coexist without interfering, thus stabilizing the training process. By incorporating PCGrad into our multi-task learning framework, we aim to improve the stability and effectiveness of our model's optimization process. This integration is expected to reduce the negative impact of gradient conflicts, thereby enhancing the overall performance of the multi-task BERT model.

In summary, our research builds on the significant advancements in multi-task learning, concurrent fine-tuning on multiple datasets, and gradient optimization techniques like PCGrad. By integrating these approaches, we aim to address existing limitations and contribute to the development of more efficient and versatile transformer-based models in NLP.

3 Approach

The architecture of the BERT model is described in the default project handout of CS 224N. In the first phase of the project, the task of sentiment classifications was conducted on the SST dataset and the CFIMDB dataset separately. For the SST dataset, BERT takes one sentence as the input and outputs classification labels from 0 (corresponding to negative) to 4 (corresponding to positive). For the CFIMDB dataset, BERT takes one sentence as the input and outputs binary classification labels.

In the second phase of the project, the SST, Quora, and SemEval datasets were used for fine-tuning simultaneously. Since each dataset contains different numbers of training examples, random samples were taken from datasets with surplus to match the dataset with least number of training examples for each epoch. For sentiment analysis, the model takes one sentence as the input, passes through the BERT layer, and outputs classification labels from 0 to 4. For paraphrase detection, the model takes a pair of sentences as the input, passes each sentence through the BERT layer, concatenates the two embeddings, and outputs binary classification labels. For semantic textual similarity, the model takes a pair of sentences as the input, passes through the BERT layer, concatenates the two embeddings and the absolute difference between them, and outputs the similarity on a scale from 0 (unrelated) to 5 (equivalent meaning). PCGrad was applied to resolve the issue of conflicting gradients (Tseng (2020)).

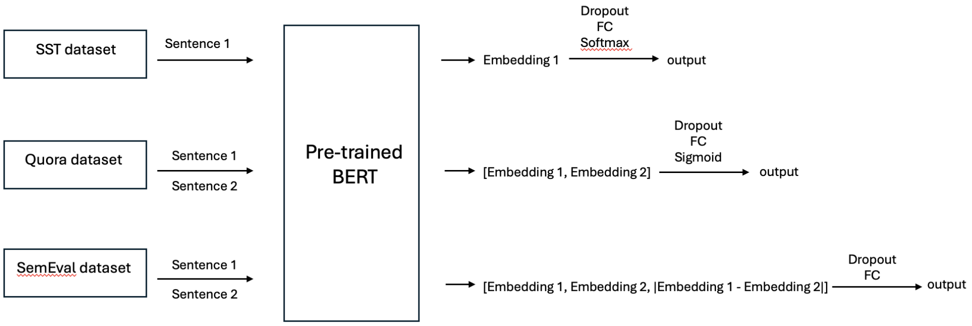

Figure 1: Model Architecture for Multi-task Learning

4 Experiments

4.1 Data

The SST dataset (Socher et al. (2013)) and the CFIMDB dataset are used for the task of predicting sentiment classifications in the first phase of the project. The Stanford Sentiment Treebank (SST) dataset comprises 11,855 single sentences from movie reviews, parsed into 215,154 unique phrases and annotated with sentiment labels (negative, somewhat negative, neutral, somewhat positive, positive). The dataset splits include 8,544 training examples, 1,101 development examples, and 2,210 test examples. The CFIMDB dataset contains 2,434 highly polar movie reviews with binary sentiment labels (negative, positive). The dataset splits include 1,701 training examples, 245 development examples, and 488 test examples.

The SST dataset, the Quora dataset, and the SemEval STS Benchmark dataset (Agirre et al. (2013)) are used for sentiment analysis, paraphrase detection, and semantic textual similarity respectively in the multi-task learning setting in the seconf phase of the project. The Quora dataset consists of 404,298 question pairs labeled to indicate whether they are paraphrases. The dataset splits include 283,010 training examples, 40,429 development examples, and 80,859 test examples. The SemEval STS Benchmark dataset includes 8,628 sentence pairs with similarity scores ranging from 0 (unrelated) to 5 (equivalent meaning). The dataset splits include 6,040 training examples, 863 development examples, and 1,725 test examples.

4.2 Evaluation method

Accuracy was used as the evaluation metric for sentiment classification and paraphrase detection tasks. Pearson correlation of the true similarity values against the predicted similarity values was used as the evaluation metric for semantic textual similarity.

4.3 Experimental details

All models were trained with the hyperparameters summarized in the table below. All trainings were conducted on Google Colab. For fine-tuning last-linear-layer BERT model, each epoch takes about 3 minutes and the entire training process takes about 30 minutes. For fine-tuning full BERT model, each epoch takes about 8 minutes and the entire training process takes about 90 minutes.

Hyperparameter	**Value**
Epochs	10
Batch Size	16
Dropout Probability	0.3
Learning Rate	1×10^{-5}

Table 1: Training Hyperparameters

4.4 Results

We evaluated the performance of our models on three sentence-level tasks: sentiment analysis, paraphrase detection, and semantic textual similarity. The results we obtained are summarized in Table 2. The full BERT model outperforms the last-linear-layer BERT model in terms of the overall development score. This suggests that fine-tuning the entire model, rather than just the last layer, allows the network to better adapt the pre-trained representations across all layers to the specifics of the task, leading to a more effective learning and better generalization on development data. Both models perform poorly on the STS (Semantic Textual Similarity) task when trained only on the SST dataset, indicating difficulty in modeling fine-grained semantic similarities using the given setup. The negative correlation suggests that the model predictions are inversely related to the actual similarity scores, which is problematic. This might be due to the model not being effectively trained on tasks requiring nuanced semantic understanding.

The full BERT model with additional fine-tuning and PCGrad achieved the highest overall development score of 0.652, demonstrating the substantial gains from our proposed enhancements. Notably, it achieved significant improvements for the tasks of paraphrase detection and semantic textual similarity. The Paraphrase Dev Accuracy increased from around 0.4 to 0.718 and the STS Development Correlation increased from nearly 0 to o a substantial 0.457.

Despite these improvements, the SST development accuracy showed a slight decrease in the full BERT model with additional fine-tuning and PCGrad (51.0%) compared to the full BERT model trained solely on SST (51.6%). This decrease suggests that while our multi-task approach and optimization techniques significantly benefit tasks like paraphrase detection and semantic textual similarity, they also reveal the presence of conflicting gradients in sentiment analysis. The integration of multiple datasets and tasks can sometimes lead to gradient conflicts, which in turn can slightly hinder performance in specific tasks like sentiment analysis.

Metric	**Last-linear-layer BERT trained on SST**	**Full BERT trained on SST**	**Full BERT with additional fine-tuning and PCGrad**
Overall Dev Score	0.387	0.476	0.652
SST Dev Accuracy	0.297	0.516	0.510
Paraphrase Dev Accuracy	0.369	0.414	0.718
STS Dev Correlation	-0.009	-0.006	0.457

Table 2: Performance Comparison of BERT Models in Multi-task Learning

5 Analysis

For sentiment analysis, the models struggled with sentences containing nuanced or mixed sentiments. Even with additional fine-tuning and PCGrad integration, the models did not significantly improve their handling of such cases, suggesting a need for more sophisticated sentiment analysis techniques. The models tend to rely heavily on words with strong emotions to make predictions. For instance, the sentence "You'll gasp appalled and laugh outraged and possibly, watching the spectacle of a promising young lad treading desperately in a nasty sea, shed an errant tear." has a true label of 3 (neutral) but was predicted to be 1 (negative) by the model. This misclassification highlights the model's inability to accurately interpret sentences with mixed emotional cues and complex structures.

In the paraphrase detection task, the full BERT model with additional fine-tuning and PCGrad showed remarkable improvement, achieving a development accuracy of 71.8%, compared to 36.9% for the last-linear-layer BERT and 41.4% for the full BERT trained solely on SST. However, error cases often involved sentences with more complex logical relationships. For instance, the sentence pair "Why do Quorans answer questions that are already answered?" and "Why do Quorans downvote questions they cannot answer?" was incorrectly classified as a paraphrase. Another example involves the pair "How do I upload photos to Quora with a pc?" and "How do you include a photo with your post on Quora?" where the model incorrectly identified the sentences as paraphrases due to misinterpreting the context and meaning of specific words such as "pc" and "photo." These examples indicate that while our model performs well on clear paraphrases, it could benefit from enhanced semantic understanding capabilities to handle nuanced differences and contextual meanings between sentences.

The semantic textual similarity (STS) task posed significant challenges. The full BERT model with additional fine-tuning and PCGrad improved the correlation to 0.457. However, the model still struggled with sentence pairs that had high similarity scores. For example, "Two black dogs are playing on the grass" and "Two black dogs are playing in a grassy plain" are predicted to have a similarity score of 2.4 but actually have a similarity score of 4.6.

6 Conclusion

In this report, we explored the effectiveness of multi-task learning and advanced optimization techniques in enhancing the performance of BERT models across a variety of NLP tasks. By leveraging a full BERT model with additional fine-tuning and integrating PCGrad, we aimed to address the challenges of gradient conflicts and improve the model's generalization capabilities.

Our results demonstrate that the full BERT model with additional fine-tuning and PCGrad achieved the highest overall development score of 0.652, significantly outperforming the last-linear-layer BERT model and the full BERT model trained solely on SST. Notably, our approach led to substantial improvements in paraphrase detection and semantic textual similarity tasks, indicating the efficacy of multi-task learning and gradient optimization in enhancing model performance.

However, the slight decrease in SST development accuracy for the full BERT model with additional fine-tuning and PCGrad highlights the presence of conflicting gradients, suggesting that while multi-task learning and PCGrad offer substantial benefits, they also introduce complexities that require careful management. This underscores the need for further refinement of these techniques to mitigate gradient conflicts and optimize performance across all tasks.

In conclusion, our research demonstrates the potential of multi-task learning and PCGrad to significantly improve the versatility and effectiveness of transformer-based models in NLP. Future work should focus on refining these techniques, exploring their application to a broader range of tasks and datasets, and investigating additional strategies to address gradient conflicts. By continuing to build on these foundations, we can contribute to the development of more robust and adaptable NLP models capable of excelling in diverse real-world applications.

7 Ethics Statement

Firstly, training with datasets like SST and CFIMDB may perpetuate or even amplify existing biases present in the training data. These biases can manifest in several forms, such as gender, racial, or ideological biases, which might affect the model's outputs and decisions, leading to unfair or prejudiced outcomes when deployed in real-world applications. The particular connection to our project lies in its reliance on data that may not have been adequately scrutinized for such biases or may reflect the inherent biases of the contexts in which the data was generated. To address this, we can implement a comprehensive bias audit of the datasets and the model outputs. This would involve analyzing the data for known biases, testing the model's performance across different demographic groups, and adjusting the training data or model architecture where necessary to minimize these biases. Additionally, continual monitoring post-deployment can help identify and correct unforeseen biases.

Secondly, the deployment of our enhanced BERT model across various NLP tasks can lead to an over-dependency on automated decisions, particularly in environments where users may not fully understand the underlying technology. This over-reliance can result in automation bias, where users trust the machine learning outputs unduly without critical scrutiny. This risk is especially pertinent to our project as it could potentially lead users to assume infallibility in its judgments. To combat this risk, it is essential to implement measures that promote transparency and user education. One effective approach could be the integration of "explanation" features that provide users with understandable reasons for the model's decisions. This can be achieved through techniques like SHAP values, which help elucidate how the model processes inputs to arrive at its conclusions. Additionally, training sessions for users, focusing on the model's capabilities, limitations, and the interpretation of its outputs, can help mitigate undue reliance and encourage more informed decision-making. This strategy not only reduces the potential for automation bias but also enhances user trust and engagement by demystifying the model's operations.

References

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, Georgia, USA. Association for Computational Linguistics.

Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28(1):41–75.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Wei-Cheng Tseng. 2020. Weichengtseng/pytorch-pcgrad.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. Gradient surgery for multi-task learning.