SMART Multitask MinBERT

Stanford CS224N Default Project

Weilun Chen
Department of Computer Science
Stanford University
chen1108@stanford.edu

Xiaowen Zhang
seanxwzhang@gmail.com

Abstract

In this study, we introduce a novel approach that combines SMART with multitask finetuning to address the challenges of model overfitting and performance enhancement in data-scarce environments. This methodology leverages regularization and joint task training to improve model robustness and generalizability. We validate our strategy using an adversarial dataset designed to test resilience against spurious correlations, demonstrating that our approach not only mitigates overfitting but also aligns model performance with specific tasks.

1 Key information to include

- (Optional) External collaborators: Xiaowen Zhang

2 Introduction

In the rapidly evolving landscape of machine learning, the paradigm of pretraining models on extensive datasets Conneau et al. (2020) followed by finetuning them on specific tasks has become a cornerstone strategy. This approach capitalizes on the abundance of large-scale data to bolster model performance in domains where data is scarce. However, when it comes to finetuning models with limited downstream data, practitioners often encounter significant challenges. The paucity of data heightens the risk of overfitting, making it difficult to align the model's capabilities with the nuances of the task at hand. Such challenges underscore the need for innovative solutions that can navigate the intricacies of model finetuning under data constraints.

Addressing these concerns, we introduce a novel methodology that integrates SMART Jiang et al. (2020) (Self-Regulated Model Attention for Regularization) regularization with a multitask finetuning framework. Our approach is designed to mitigate the risks of overfitting by introducing a regularization technique that enhances the model's generalizability. By incorporating multitask finetuning, we not only bolster the model's robustness but also improve its performance on specific tasks by leveraging synergies from ancillary downstream tasks. This multifaceted strategy enables our model to draw on a broader spectrum of data, enhancing its ability to generalize and adapt to new tasks.

Furthermore, to validate the effectiveness of our approach, we have constructed an adversarial dataset characterized by spurious correlations. This dataset serves as a rigorous testing ground to assess the resilience of our combined regularization and joint training strategy against misleading patterns in the data. Through this innovative validation method, we aim to demonstrate the efficacy of our approach in overcoming the prevalent challenges in finetuning models with limited data. Our findings suggest that this combined approach not only addresses the issues of overfitting and task alignment but also sets a new benchmark for model performance in data-constrained environments.

In summary, our contribution is twofold: we introduce a sophisticated regularization technique complemented by a multitask finetuning framework to enhance model performance in the face of scarce data, and we validate our approach against an adversarially designed dataset to demonstrate its effectiveness in mitigating overfitting and improving task-specific performance. This work not only advances our understanding of model finetuning strategies but also opens new avenues for research in the optimization of machine learning models under data limitations.

3 Related Work

The landscape of machine learning research is rich with efforts aimed at enhancing model performance in the face of limited data availability. Our approach builds upon and diverges from several key areas within this broad domain, particularly in the realms of regularization techniques, multitask learning, and the utilization of adversarial datasets for model validation.

3.1 Regularization

Regularization has long been a staple in machine learning to combat overfitting, with techniques such as dropout, L2 regularization, and more recently, attention-based methods like SMART (Self-Regulated Model Attention for Regularization) gaining prominence. Our work is mostly affiliated with the works by TODO(CITE) have laid the groundwork for understanding how different regularization techniques can be tailored to improve model generalization. Our approach extends this line of inquiry by integrating SMART regularization in a novel finetuning context, aiming to leverage the self-attention mechanisms of transformers to enhance model robustness.

3.2 Multitask Learning

Multitask learning (MTL) strategies, as discussed by Caruana (1997) and Ruder (2017), involve training a model on multiple related tasks simultaneously to improve its performance on each. The intuition behind MTL is that learning related tasks together allows the model to share representations and leverage commonalities among tasks, thus improving generalization. Our work contributes to this area by exploring how multitask finetuning, when combined with advanced regularization techniques, can further enhance model performance, especially in scenarios with limited task-specific data.

4 Approach

At the core of our methodology lies the implementation of multitask learning. This technique leverages the information in different datasets concurrently, thereby enriching its learning experience and generalizability. We employ a BERT-based model, renowned for its powerful representational capabilities, as the foundational backbone. This pretrained model is equipped with three distinct task-specific heads, each tailored to a different downstream task. These heads operate in parallel, sharing the BERT model's base layers while retaining the ability to specialize through their task-specific components. This design allows for the efficient propagation of knowledge across tasks, leveraging shared patterns and reducing the model's susceptibility to overfitting on any single task's limited data. By harnessing the strengths of multitask learning, we aim to cultivate a model that is not only versatile across various tasks but also exhibits enhanced performance on each individual task due to the cross-pollination of insights and learnings.

To further safeguard against the risk of overfitting and to ensure the model's generalization capabilities are maximized, we incorporate SMART (Self-Regulated Model Attention for Regularization) regularization into our training regime. SMART regularization acts as a sophisticated mechanism that fine-tunes the model's focus, encouraging it to distribute its attention more evenly across the input features and preventing it from relying too heavily on potentially misleading or noisy signals. This regularization technique is particularly effective in the context of multitask learning, where the model's ability to generalize across diverse datasets is paramount. By implementing SMART on top of our multitask learning framework, we imbue the model with an enhanced capacity for discernment, enabling it to make more informed decisions and improve its overall robustness and accuracy.

A critical aspect of our approach is the introduction of an adversarial dataset designed to challenge the model with spurious correlations and misleading patterns. This dataset is purposely constructed trick the model to cheat by assigning spurious correlation with some keyword such as "good" and "bad", while ignoring the semantic meaning. For this dataset, we leverage GPT-4 to rewrite the CFIMDB dataset, where we add a word "good" for positive review, and "bad" for negative review in training data. However, for dev and test dataset, we do the exactly the other way around, adding a "good" for negative review (e.g. "nothing is good about this moview") and adding "bad" for positive ones.

In summary, our approach combines multitask learning, SMART regularization and demonstrate its effectiveness using a spuriously constructed dataset.

5 Experiments

5.1 Data

Our experiment is maily conducted with the following 3 datasets

Quora Paraphrase Dataset: The Quora dataset, as previously described in Section 1, consists of 400,000 question pairs with labels indicating whether particular instances are paraphrases of one another.
SemEval STS Benchmark Dataset: The SemEval STS Benchmark dataset as described in Section 1 consists of 8,628 different sentence pairs of varying similarity on a scale from 0 (unrelated) to 5 (equivalent meaning).
SST Benchmark Dataset: The Stanford Sentiment Treebank consists of 11,855 single sentences from movie reviews extracted from movie reviews, each labeled with a 5 scale score.

Additionally, we use GPT-4 to rewrite the CFIMDB dataset to artificially inject a word "good" into positive training data, and "bad" into negative training data. While doing the converse thing for test/dev dataset. Which gave us the following dataset.

CFIMDB Rewrite Benchmark Dataset: This consists of 1706 training moview review with positive/negative labeling. We utilize GPT-4 to modify the CFIMDB dataset, incorporating the word "good" into positive reviews and "bad" into negative reviews within the training data. Conversely, in the development and test datasets, we invert this approach, appending "good" to negative reviews (for example, "nothing is good about this movie") and "bad" to positive reviews.

5.2 Evaluation method

We use Pearson correlation for regression task (STS), accuracy for classification (Quora, SST, CFIMDB).

5.3 Experimental details

We finetune Bert model with 10 epochs, with learning rate of 1e-5 on all datasets. To address the issue of dataset not having equal length, we loop smaller dataset while keeping the largest dataset (Quora) only processed once per epoch. For SMART settings, we use the default setting in the original SMART paper.

5.4 Results

We first report the performance on SST/STS and Quora result with different training setup.

	SST Dev Set	Quora Dev Set	STS Dev Set
Baseline	0.414	0.702	0.284
Smart	0.425	0.702	0.286
SST Quora STS MTL	0.441	0.699	0.289
SST Quora STS MTL Smart	0.448	0.703	0.294
SST Quora MT	0.442	0.702	N/A
SST Quora MT Smart	0.443	0.694	N/A
Quora STS MT	N/A	0.692	0.285
Quora STS MT Smart	N/A	0.694	0.292

Table 1: Result in pretrain mode

	SST Dev Set	Quora Dev Set	STS Dev Set
Baseline	0.531	0.790	0.369
Smart	0.504	0.797	0.380
SST Quora STS MTL	0.495	0.789	0.403
SST Quora STS MTL Smart	0.519	0.796	0.404
SST Quora MT	0.509	0.786	N/A
SST Quora MT Smart	0.526	0.794	N/A
Quora STS MT	N/A	0.785	0.403
Quora STS MT Smart	N/A	0.797	0.403

Table 2: Result in finetune mode

We then report the SMART finetuning result on the adversarial dataset below.

	Train Accuracy	Dev Accuracy
Baseline	1	0.946
Smart	0.998	0.967

Table 3: Result in CFIMDB rewritten dataset

6 Analysis

6.1 Quantitative Analysis

Our analysis yields three key insights regarding the training configurations employed.

1. Generally, the SMART training approach enhances model performance across various scenarios.
2. Implementing multi-task learning (MTL) benefits models trained on smaller datasets but appears to detrimentally affect those trained on larger datasets, such as Quora.
3. As demonstrated in table 3. SMART shows significant regularization against bias within the data. Handling the bias (false association of "good"/"bad") much better.

7 Conclusion

In conclusion, our research introduces a novel framework that synergistically combines multitask learning, SMART regularization, and an adversarial dataset to tackle the challenges of finetuning models with limited data. Through the innovative integration of these elements, we demonstrate significant advancements in model robustness, generalizability, and task-specific performance. Our approach not only mitigates the risks of overfitting but also paves the way for models to better navigate the complexities and nuances of diverse tasks. The utilization of an adversarially crafted dataset further underscores our model's capability to withstand and adapt to challenging data scenarios, reflecting its practical applicability in real-world settings.

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2020. SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2177–2190, Online. Association for Computational Linguistics.