Multi-task minBERT using PALs with Gradient Episodic Memory

Stanford CS224N Default Project

Christopher Nguyen
Department of Computer Science
Stanford University
cnguye29@stanford.edu

1 Key information to include

- External collaborators: N/A
- Mentor: N/A
- Sharing project: N/A
- Short description: The project aims to extend an implemented minBERT model to improve its performance on sentiment analysis, paraphrase detection, and semantic textual similarity with Projected Attention Layers (PALs) for a more robust and diversely applicable model. Additionally, gradient episodic memory is applied for reducing task interference. The idea of behind a multi-task learning approach is that it provides an inductive bias meaning models have to learn features that are general enough to perform well on many tasks.

2 Research paper summary (max 2 pages)

Title	BERT and PALs: Projected Attention Layers for Efficient Adaptation in Multi-Task Learning
Venue	International Conference on Machine Learning
Year	2019
URL	https://proceedings.mlr.press/v97/stickland19a.html

Table 1: Bibliographical information (Stickland and Murray, 2019).

Background. The paper explores how to adapt a single large base model consisting of deep neural networks to work with multiple natural language understanding tasks. Specifically, they work with the BERT model as the base pre-trained model. The difference in their work in comparison to some other multi-task learning approaches is based mainly on parameter sharing. In their setting, they follow a hard parameter sharing approach where most parameters are shared across all tasks but have a small number of task-specific parameters which adapt the shared model (Stickland and Murray, 2019). In contrast, many other approaches shares all parameters across tasks within a general-purpose model. Furthermore, they explore questions such as where should the transformation of the base model be and what form it should take with the assumption that the task is always known. Key motivations behind PALs includes resource limitation in storage space, preserving life of batteries in mobile devices applying the model, and minimizing computational and energy overhead in mobile devices. Additionally, another factor in multi-task learning is to adapt the parameters of the already powerful and popular BERT architecture for other useful applications such as multilingual machine translation.

Summary of contributions. Stickland and Murray (2019) introduced the Projected Attention Layer (PALs). PAL is a low-dimensional multi-head attention layer that are task-specific functions added in parallel to the normal BERT layers. By incorporating PALs into BERT, the model can efficiently adapt to multiple tasks without significantly increasing computational complexity or memory requirements. This enables BERT to perform well on diverse tasks while maintaining computational efficiency. Another contribution introduced is a new method for scheduling training. The paper suggests sampling tasks proportional to their training set size and de-emphasize training set size as training proceeds.

Limitations and discussion. Although the paper provided valuable insight into the performance of the model using PALs, there were a few aspects that could have strengthened the findings of the paper. Firstly, their research was primarily based on English text and standard benchmark datasets like GLUE although they mentioned using multi-task learning for multilingual translation. Including experiments on multilingual datasets or diverse text corpora would provide a more comprehensive assessment of the model's generalization capability across different languages and domains and showcase practical usage for translation. Additionally, the paper did not provide any practical application for the model. The main motivation behind this paper was to provide a method that could reduce resource usage on mobile devices and web-applications but there was no further acknowledgement of how the PALs based model could be utilized given the results of the experiment. Another limitation of the paper that was directly addressed was "not considering different variations of training method and used no methods to reduce interference from training on separate tasks" (Stickland and Murray, 2019).
Even with these results, I still find PALs as a convincing method for multi-task learning. The paper offers valuable insights into the integration of PALs into BERT and highlights avenues for future research in multi-task learning and natural language processing.

Why this paper? I chose this paper as the main topic of it is multi-task learning and an adaptation of the BERT model similar to the one used for the project. The paper also has strong testing datasets and results that cover the same tasks as the project. Along with this, I found the idea of building a robust model with hard-parameter sharing an interesting topic. I also see potential to expand on the idea presented by Stickland and Murray (2019) as they discuss details in their experiment that can be improved or ideas that were not tested. The paper as a whole was very informative and the material presented provides a strong base for the context of the project.

Wider research context. Stickland and Murray (2019) contributes to several fundamental concepts in natural language processing, including language representation, structure, and the challenges of modeling language with computers. For language representation, BERT already involves masked language modeling and next sentence prediction which allows it to capture intricate linguistic patterns and semantics within a text. PALs further enhances the model's ability to adapt these representations to different tasks efficiently. As it is an adaptation of the BERT model, it captures both local and global dependencies within sentences. By leveraging self-attention mechanisms, BERT can effectively model syntactic and semantic relationships, thereby improving its understanding of language structure. And finally, the paper addresses challenges in modeling language with computers by pursuing a generalization approach with pre-trained models like BERT to diverse NLP tasks without extensive fine-tuning or task-specific architectures.
The premise of PALs naturally makes it broadly applicable to many NLP tasks. PALs offers a novel approach to enhancing representation learning and task adaptation in pre-trained models, which could potentially be extended to other architectures or learning frameworks. For example, the Transformer introduced in "Attention is All You Need" (**?**) might complement PALs in improving representation learning and task performance.

3 Project description (1-2 pages)

Goal. The main goal of the project is to investigate the performance of the minBERT model following a multi-task learning (MTL) approach. This is done by extending the model to use PALs and gradient episodic memory for continual learning(Lopez-Paz and Ranzato, 2017). Gradient episodic memory involves storing past gradients or training examples in memory and replaying them during training. The motivation for this goal is to see if MTL has the ability to match performance with a

model fine-tuned to a specific task. This adds an interesting challenge because fine-tuning separate models for each task often works better, but this may provide insight into how MTL can reduce computational time by reducing number of parameters required as described in the paper. As my goal aligns with the experiment of Stickland and Murray (2019) and is a continuation of the paper, I am looking to further extend this approach, if time permits, by implementing gradient surgery (Yu et al., 2020) and comparing the results to using gradient episodic memory. Gradient surgery involves altering conflicting gradients by projecting each onto the normal plane of the other, preventing the interfering components of the gradient from being applied to the network.

Task. Following the same premise proposed by Stickland and Murray (2019) that looked to have a BERT model with PALs match or out-perform separately fine-tuned BERT models, the project will be using the sentiment analysis task as an indicator for the success of the model. The project model will compare the extension of minBERT's performance on sentiment analysis with the original model's performance as the original is trained on this specific task. Additionally, NLP tasks such as paraphrase detection and semantic textual similarity (STS) will be addressed.

Data. The initial implementation of the minBERT model will utilize the Stanford Sentiment Treebank (SST) dataset that consists of 11,855 single sentences from movie reviews and the CFIMDB dataset consisting of 2,434 highly polar movie reviews. For training and testing the extension of the minBERT model with PALs, the Quora dataset consisting of 400,000 question pair with labels for paraphrase detection, SemEval dataset consisting of 8,628 different sentence pairs for semantic textual analysis, and again the SST dataset for sentiment analysis will be used. I expect to pre-process the data with tokenization, padding/truncating, and potential label encoding.

Methods. The method described in the paper and the one I will be using is based on hard-parameter sharing in multi-task learning. This is done by adding adapters to shared layers as well as the usual separate output layers (Stickland and Murray, 2019). Although the method follows a similar approach as the paper using PALs, I plan on slightly modifying it by attempting gradient episodic memory for continual learning (GEM). With GEM, it reduces interference from training on separate tasks by retaining information/gradients from past tasks. This method of replaying past gradients is done to reinforce previously learned knowledge and maintain performance on earlier tasks while learning new ones.
The main part of my method is integrating the open source codes of PALs and gradient episodic memory to the implemented minBERT model. Stickland and Murray (2019) already have methods defined for using PALs with a BERT model so the big part of the implementation is incorporating gradient episodic memory.

Baselines. The baseline to be utilized is comparing with results provided by Stickland and Murray (2019) based on the GLUE benchmark. The paper shows GLUE Test results for each task scored by the GLUE evaluation server. I will also use the original implementation of minBERT as a benchmark to know if my proposed method provided a better performance. Another baseline would be to compare accuracy/scores with previous projects using PALs or to strengthen my model more would be to compare with other methods. If time permits, I can implement a logistic regression model for sentiment analysis.

Evaluation. An evaluation metric to be used is the F1/accuracy score; and as stated in the baseline, I will be evaluating my extension of PALs against the score and accuracy obtained by Stickland and Murray (2019). Additionally, I can compare to the submitted class scores to determine the performance of my method.

References

David Lopez-Paz and Marc' Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Asa Cooper Stickland and Iain Murray. 2019. BERT and PALs: Projected attention layers for efficient adaptation in multi-task learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5986–5995. PMLR.

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