minBERT and Downstream Tasks

Stanford CS224N Default Project

Jason Alexander Chan
Department of Computer Science
Stanford University
name@stanford.edu

1 Key Information to include

- External collaborators (if you have any): NA
- Mentor (custom project only): NA
- Sharing project: NA

2 Research paper summary (max 2 pages)

Title	Parameter-Efficient Transfer Learning For NLP
Venue	ICML
Year	2019
URL	http://proceedings.mlr.press/v97/houlsby19a/houlsby19a.pdf

Table 1: Bibliographic information [1].

Background. In 2018, BERT demonstrated that fine-tuning pre-trained language models works 'unreasonably well' for downstream Natural Language Processing (NLP) tasks. However, a major disadvantage is that BERT instantiated a new model for each downstream NLP task and fine-tuned each model's parameters. This is inefficient if we consider the three following questions:

1. What is the smallest subset of parameters required to be updated for the downstream NLP task to be considered 'successful'?
2. What techniques can be applied to reduce the inefficiencies of fine-tuning all the pre-trained parameters?
3. To what extent can a single pre-trained model be adjusted so that it can perform satisfactorily on many downstream tasks, also known as transfer learning?

Summary of contributions. This paper proposed a transfer learning strategy by adding adapter modules in between layers of a pre-trained model, where 'the model trained incrementally to solve new tasks without forgetting previous ones'. This paper applied the adapter strategy on the GLUE benchmark, a further 17 text classification tasks plus SQuAD v1.1 Question Answering. The authors obtained 'near state-of-the-art performance while only adding a few parameters per task'. As an example, on the GLUE benchmark the authors were 'within 0.4% of the performance of full fine-tuning' but only added 3.6% more parameters per task.

The adapters have two features: the first is a small addition of tuneable parameters and the second is a 'near identity initialization' which was found to be necessary for stable training. Each adapter does not interact and its parameters are frozen.

Limitations and discussion. The authors used the pre-trained BERT transformer as the base model and followed the approach in the original BERT paper's training procedure. This methodology is sound because the authors were intending to benchmark the adapter strategy for transfer learning vs. new model instantiations and fine-tuning. Other noteworthy discussion items in the paper include:

1. The authors tested robustness of the adapter strategy by removing adapters layers (also known as ablation) and checked performance degradation on the validation dataset without retraining the model. They found that removing layers at the top of the model had the most detrimental impact on downstream NLP tasks.
2. The authors acknowledged that the adapter model failed to train if the adapter initialization deviates too far from the identity function.
3. The authors tested their adapter strategy on 17 additional downstream NLP tasks vs. a fine-tuned BERT vs. a variable fine-tuned BERT vs. a non-BERT baseline. The non-BERT baseline was derived via Auto-ML to find the best the architecture and hyper-parameters because the authors claim that there is 'no state of the art numbers for these tasks'. The variable fine-tuned BERT only tunes the top n layers.

The paper does not have an explicit section on limitations. Some limitations identified in this paper include:

1. The authors experimented with extending the adapter architecture but discovered that it did not materially improve performance. In fact, it performed worse than their simplest proposal of the adapter architecture. This suggests that potential of the adapter strategy may be already at its limit. The authors document their failed experiments for readers, which is useful.
2. The authors don't convincingly report the effects of ablation vs. all the downstream NLP tasks. Since the objective is transfer learning, a more rigorous paper would discuss the impacts of ablating each adapter for all downstream NLP tasks.
3. It's unclear what is maximum number of new NLP tasks that can be added to the pre-trained model before performance across these previously trained tasks start to degrade.
4. It took AutoML one continuous week using 30 machines for each downstream task, which is infeasible to replicate for the majority of research scientists and engineers.

Why this paper? This paper was chosen because it made a notable contribution to the idea of extending foundational NLP models like BERT for transfer learning - adding only roughly 3% more parameters per downstream NLP task is quite remarkable. This paper was cited by 967 people, which gives it some gravitas. The authors work at Google Brain, who specialize in deep learning.

Wider research context. Since 2018, fine-tuning and its variants are yielding remarkable results on notable NLP benchmarks in a very short span of time. There is some concern that this progress should be met with some caution since they dont' reflect the wider space of practical NLP problems. And the hype over the public release of Chat-GPT also risks inflaming expectations about the potential of NLP in general. There is still much work to be done on machine reasoning particularly in the areas related to bias, fact-checking, physics informed reasoning, and higher-level problem solving so it is unlikely that fine-tuning is the final stop in NLP research. This paper demonstrated results that are comparable to fine-tuning as another vector of research whose objective is to improve generalisation of models for NLP tasks while minimising computation and memory requirements. This has real world applications in cloud computing and edge devices like mobile phones or robotics. This paper also identified related work in multi-task learning, transfer learning in vision and continual learning as adjacent fields.

3 Project description (1-2 pages)

Goal. The objective of this project is to assess the performance of taking a foundational pre-trained NLP model and extending it for multiple NLP downstream tasks. This project will demonstrate the generalisation potential of pre-trained NLP models and reveal its limitations.

This objective is motivated by the desire to better understand the transformer model and compare fine-tuning vs. the adapter strategy for transfer learning on downstream NLP tasks. The adapter strategy is promising because rather than retuning all the pre-trained parameters, it only adds a thin layer of tuneable parameters on top of the pre-trained parameters, which are all frozen. A new layer is added for each subsequent task. The benefits of transfer learning should be reduced computation and memory, which will be interesting to compare against the pre-trained BERT model and the pre-trained BERT model with fine-tuning.

Task. There are three specific tasks in this project:

1. Implement minBERT, which has the key features of the original BERT model that includes the multi-head self-attention and Transformer layer. The output of this task is a pre-trained minBERT model.
2. Evaluate minBERT on two downstream NLP tasks: sentiment analysis on the Stanford Sentiment Treebank (SST) and the CFIMDB movie reviews database. The output of this task is the performance result of the pre-trained minBERT model on the SST and IMDB movie review database
3. Adjust the minBERT parameters for three tasks: sentiment analysis, paraphrase detection and semantic textual similarity across the SST dataset, the Quora dataset, and the SemEval dataset.The output of this task is the performance result across these three tasks and datasets for pre-trained minBERT, pre-trained minBERT with fine-tuning, pre-trained minBERT with adapter strategy. The adapter strategy will involve implementing the paper presented in the paper summary.

Data. This project will employ the following datasets:

1. The Stanford Sentiment Treebank (SST) contains 11,885 single sentences from movie reviews. There are 215,154 unique phrases annotated by 3 human judges with the labels: negative, somewhat negative, neutral, somewhat positive and positive. Train/Dev/Test split is 72%, 9%, 19%.
2. The CFIMDB movie reviews database has 2434 movie reviews whose labels are positive or negative. There are reviews that are longer than one sentence. Train/Dev/Test split is 70%, 10%, 20%.
3. The Quora dataset has 400,000 question answer pairs with labels informing whether such instances are paraphrases of each other. Train/Dev/Test split is 70%, 10%, 20%.
4. SemEval dataset has 8628 different sentence pairs of varying similarity on a scale from 0 (unrelated) to 5 (equivalent in meaning). Train/Dev/Test split is 70%, 10%, 20%.

Methods. Applying the adapter strategy to extend the pre-trained minBERT model involves adding two adapter layers in each transformer layer's two sub-layers. According to the authors: 'The adapter is always applied directly to the output of the sub-layer after the projection back to the input size, but before adding the skip connection back'. After each skip connection a new normalization layer is also added and trained per task. Each adapter layer comprises a feed-forward layer, non-linearity layer, another feed-forward layer, plus a skip-connection.

Baselines. The baselines in the project will be presented by me from the results of the pre-trained minBERT model, pre-trained minBERT model with fine-tuning, and pre-trained minBERT model with the adapter strategy. These results will be compared with other students in CS224N on the leaderboard.

Evaluation. There are a number of evaluation metrics that will be employed in this project

1. The paper summarised in this research proposal employed accuracy, which would be used as the first evaluation metric.
2. To evaluate the effectiveness of the adapter strategy, the following evaluation metrics are considered:

(a) The number of additional parameters added via the adapter strategy vs. the number of parameters updated during fine-tuning for each down stream NLP task.

(b) The empirical training duration of the adapter strategy vs. fine-tuning, noting the compute platform employed.

(c) Accuracy performance of the adapter strategy vs. fine-tuning on the test data splits for each of the four data sets.

3. If the classes in a data set are unbalanced then using the F1 score could be considered because it combines precision and recall.

References

[1] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebsk, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Gelly Sylvain. Parameter-efficient transfer learning for nlp. In *Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97*, 2019.