DistilBERT

Xinlin Wu, Bozhou Jin, Chengyi Jiang

http://github.com/TimesECS/5782final_DistilBERT

1 Introduction

Title: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter [1]
Authors: Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf
BERT as a transformer encoder-based [2] language model has 110 million parameters. This size makes it expensive to train or use, and hard to deploy on memory-limited devices. The key motivation for DistilBERT is to shrink the model while largely retaining its performance. In [1], a student (DistilBERT) model is trained by bringing its output close to both the teacher (BERT-base) model and the ground truth, thus "distilling" the knowledge from the teacher. The resulting student model is only about half the size (66 million parameters) of the teacher model, while retaining 97% of its performance on benchmark tests (GLUE [3]).

2 Chosen Result

The specific result we chose to reproduce is the performance of the DistilBERT model on the benchmark tests in the GLUE [3] dataset (relevant table: Table 1 [1]). These results are the fundamental measurements illustrating that DistilBERT, while smaller, can perform as good as the teacher model, thus key to their main contributions. Given the scope of this project, we focused on the 5 binary classification tasks (SST2, MRPC, QQP, CoLA, RTE). Both the teacher and the original DistilBERT have various performance (from 51% accuracy on CoLA to 91% on SST2). Therefore, these tests provide a comprehensive analysis from our model performance, despite their simplicity.

3 Methodology

Model Architecture: We first use the pretrained bert-base-uncased model (12 layers) as a teacher to guide the DistilBERTForMaskedLM student model [A] (6 layers) in performing knowledge distillation on the Wikipedia dataset, combining KL loss, Cross Entropy loss, and Cosine loss. The distilled DistilBERTForMaskedLM model is then used to initialize a DistilBERTForSequenceClassification model [B]. For each of the 5 tasks, this model was used to generate a fine-tuned model by adding a classification head and optimizing with Cross-Entropy loss.
We also include an alternative approach that uses the DistilBERTForSequenceClassification model for both teacher and student, and directly distilling using each of the aforementioned binary classification datasets without pretraining. This is to test whether we could achieve high performance on specific task classes given limited amount of time and resources.
Datasets: Wikipedia (20220301.en); CoLA, MRPC, QQP, RTE, and SST-2 from the GLUE benchmark. Due to computational power and data availability we only used the "train" split of Wikipedia and "train + validation" of the other datasets, and mannually carved out the train, validation and test splits.

Evaluation Metrics: For the full re-implementation we note its loss in test dataset after pretraining, then evaluate the fine-tuned model with the accuracy on CoLA, MRPC, QQP, RTE, and SST-2.
Modifications: (1) The original methodology uses a full concatenation of English Wikipedia and the Toronto Book Corpus for pretraining, we used Wikipedia only (more explained below). (2) The original methodology initializes the student model with the teacher's weights, whereas our student model is initialized with random weights. (3) The original model had much more training than us. We pretrain the model for 1 epoch. All trainings in fine tuning or the alternative approach was done for 3 epochs. (4) The alternative approach is one that was not explored in the original model.

4 Results & Analysis

The re-implementation achieves comparable performance on SST-2, QQP, and RTE, better performance on CoLA, but worse performance on MRPC compared to the original DistilBERT [C]. The worse performance on MRPC might be due to not initializing the model with the teacher model's weights, which, according to the original paper, might lead to a decrease in score. The better performance on CoLA may be explained by the fact that we pre-trained our DistilBERTForMaskedLM for only one epoch, making it less dependent on the teacher model (BERT-base). The BERT-base model has an accuracy of 56.3 on CoLA, and the original DistilBERT was pre-trained for more epochs. Thus, while the original DistilBERT achieves similar accuracy to BERT-base on CoLA, our model achieves better performance.
Surprisingly, the alternative approach that directly distilled without pretraining also generated greate results. The test on SST-2 was comparable to the original result; MRPC, QQP and RTE generated less but reasonably good outcome compared with original or our DistilBERT, which was likely due to pretraining; the results on CoLA was similar to our re-implementation. Although these results do not generalize to other datasets, they do indicate a potentially very light-weighted alternative solution for DistilBERT. Compared to the 19 hour pretraining in our re-implementation, this approach was completed in about 1 hour. The distillation architecture likely absorbs major part of the teacher's knowledge in very short training time, with little to no compromise in performance.
The main challenges we encountered were limitations in computational resources and data availability. Our training was mainly executed on an RTX 4070 GPU with 12GB VRAM. This is very limited compared with the 8 16GB V100 GPUs used in [1]. Our solutions were: (1) To reduce the dataset and training time. (2) To try the alternative approach as described above. The two solutions were expected to complement each other in resources required, expected test accuracy, and generalization. In terms of data availability, the original work was trained on a concatenation of English Wikipedia and Toronto Book Corpus datasets, but the latter was deprecated at the time of our work. The datasets in [3] hid their labels for the test split (online test was available for only twice a day), so we had to manually carve out train, validation and test sets from their available data.

5 Reflections

Key takeaways: We have successfully re-implemented DistilBERT and achieved comparable, reasonably close or better results compared to the original DistilBERT on various tasks. In addition, a very light-weighted, non-generalized options do exist for specific tasks, with a small amount of training using the corresponding dataset.

Lessons learned: before committing to a long training process, one should make sure that the code can run and the checkpoints are regularly saved. For CPU-intensive work such as tokenizing the data, parallelization and saving the tokenized dataset will be very helpful.
Future directions: Task-specific distillation for task types other than binary classification; possibility of further distillation of the model to compress it with fewer encoder layers.

6 References

[1] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[3] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *In the Proceedings of ICLR.*, 2019.

7 Appendix

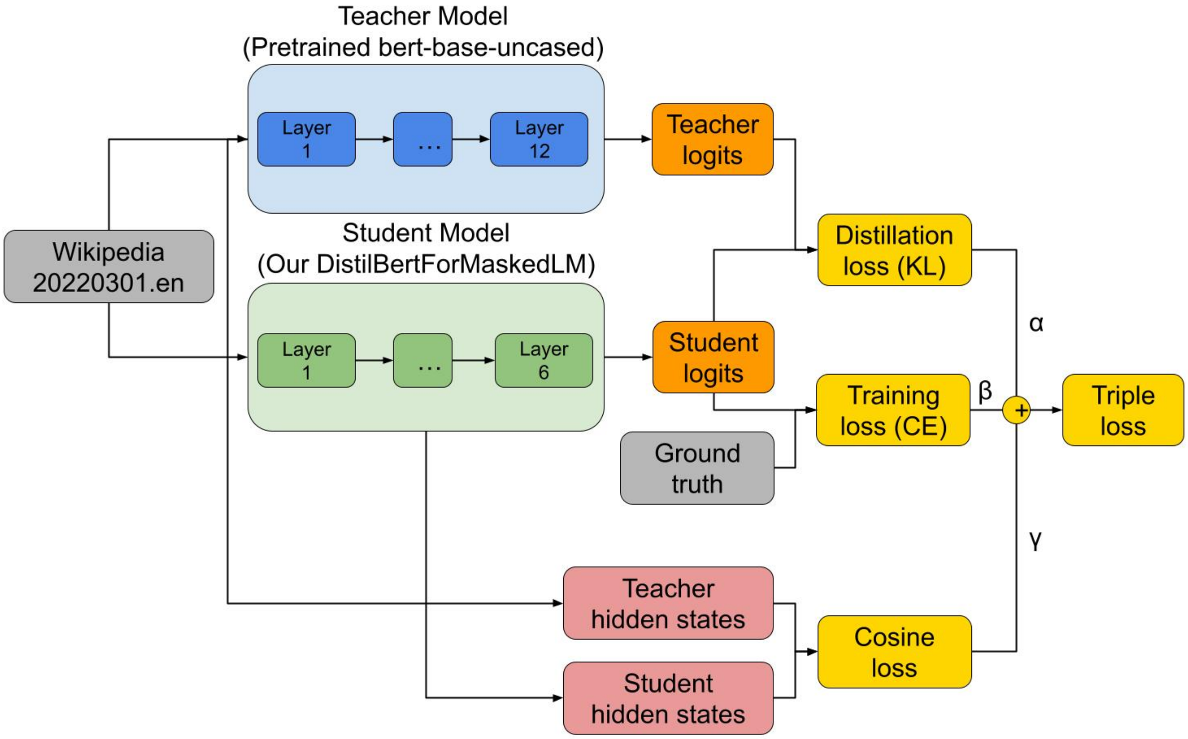

Figure 1: DistilBERT for Masked Language Modeling

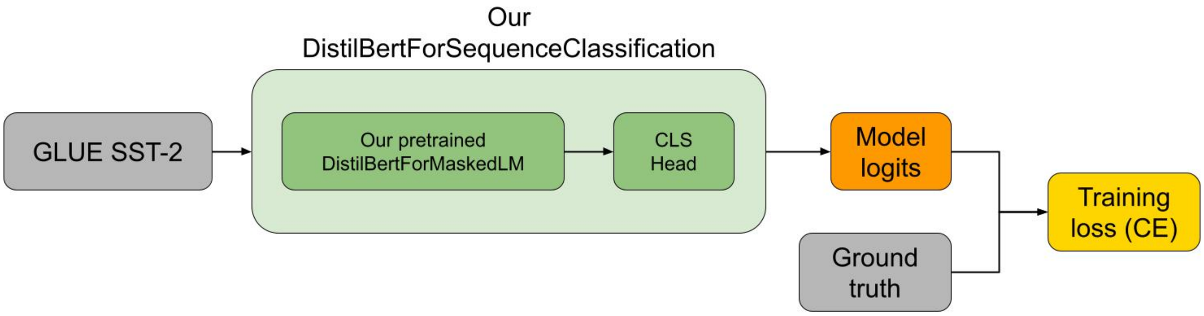

Figure 2: DistilBERT for Sequence Classification

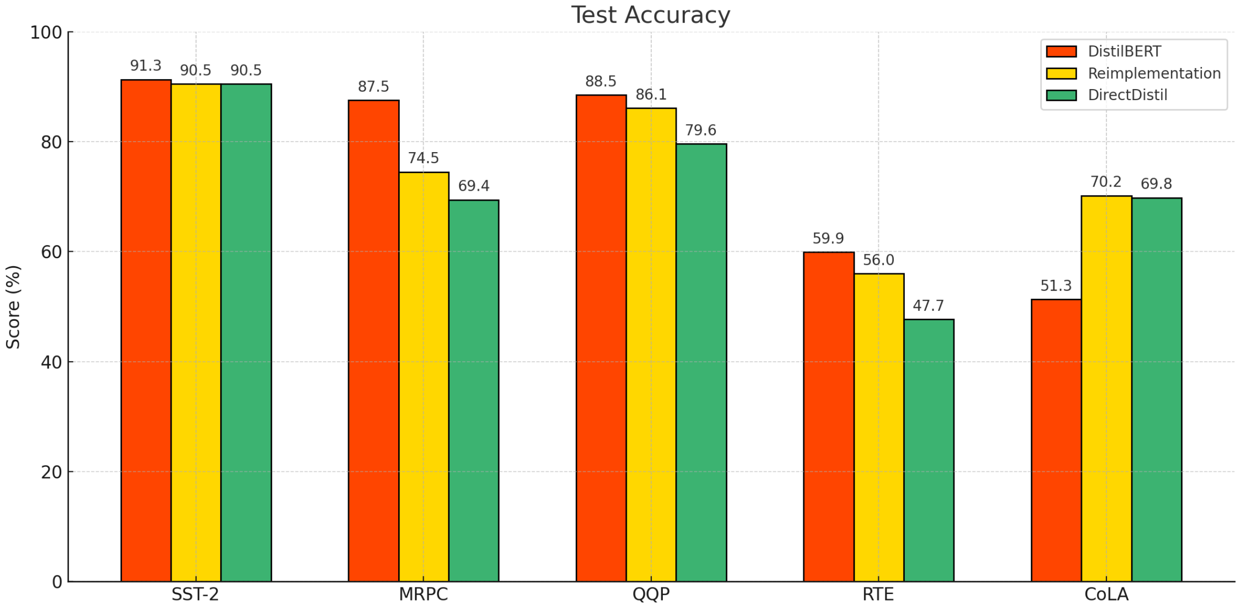

Figure 3: Accuracy on five GLUE binary classification tasks (CoLA, MRPC, QQP, RTE, SST-2)