Compare Fine-tuned Bert Models for Sentiment Analysis

Haoran Hu Duc Minh Nguyen Quan Pham
hhran,nguymi01,phamquan@bu.edu

Abstract

This project explores the efficacy of fine-tuning three distinct transformer models—BERT-base, BERT-large, and DistilBERT—on sentiment analysis tasks using the IMDB dataset. The first objective is to investigate the impact of model size on sentiment classification accuracy. The models are fine-tuned on the IMDB dataset, a benchmark for sentiment analysis in the movie review domain. Subsequently, the fine-tuned models are evaluated on a different domain by testing them on the Amazon review dataset.

In addition to the fine-tuned models, the original, untuned versions of the transformer models are also assessed for sentiment analysis performance. This comparative analysis aims to quantify the improvement in accuracy achieved through fine-tuning. The relationship between model size and accuracy increment is examined to discern patterns and insights into the scalability of transformer models for sentiment analysis tasks.

By controlling various hyperparameters and variables, we strive to maintain consistency in all aspects except for the pretrained models. We observed that larger models yield better fine-tuned performance. While DistilBERT has undergone fine-tuning for sentiment analysis tasks, it exhibits the least improvement. Additionally, BERT-base demonstrates slightly worse performance compared to BERT-large.

Our findings provide valuable insights into the transferability of sentiment analysis models across domains and shed light on the trade-offs associated with model size in relation to task performance. This research contributes to the ongoing exploration of transformer models in natural language processing and their adaptability to diverse datasets.

1. Introduction

1.1. Motivation

Since the emergence of transformer models, the size of parameters has been increasing to achieve better performance. To compensate for this and avoid creating more models for different tasks, people have employed fine-tuning techniques. We are particularly interested in how model size will affect fine-tuning results, specifically whether a larger model will yield greater improvements during fine-tuning. Therefore, we want to test three different sizes of models on the same fine-tuning task to observe their performance.

This is interesting, as the findings could help us better understand the impact of model sizes on fine-tuning performance and, consequently, overall performance. This, in turn, will assist us in designing our model more efficiently and utilizing fewer resources.

2. Dataset

The dataset chosen for fine-tuning our pretrained models for sentiment analysis is the IMDB reviews dataset, comprising 25,000 movie reviews, each labeled as positive or negative based on sentiment. We aim to demonstrate that by fine-tuning on this relatively small dataset, even the base BERT model can achieve high performance for sentiment analysis, which can be generalized to a much larger dataset.

For evaluating the models in the sentiment analysis task, we selected the Amazon reviews dataset, consisting of 400,000 reviews for various products listed on the Amazon website, not limited to movies. The labels have been categorized as positive or negative as well.

The link for the datasets is: https://drive.google.com/dnrive/folders/1JSk8zg4HsnAYTGURyGMX9JHAwRCtVgH_?usp=drive_link

3. Approach

We initially selected four distinguished models, each with distinct features or sizes from the others. We chose BERT-base-uncased and BERT-base-cased for a case-wise performance comparison. Additionally, we included BERT-base-uncased, BERT-large-uncased, and DistilBERT for a model size comparison.

Next, we gathered IMDB movie reviews data from Kaggle for fine-tuning the sentiment analysis task. We converted the ”negative” and ”positive” labels into 0 and 1, respectively. Subsequently, we split the dataset into training and testing datasets. Following this, we utilized the BERT

	text	sentiment
0	Now, I won't deny that when I purchased this o...	neg
1	The saddest thing about this "tribute" is that...	neg
2	Last night I decided to watch the prequel or s...	neg
3	I have to admit that i liked the first half of...	neg
4	I was not impressed about this film especially...	neg
...	...	...
24995	This film is fun, if your a person who likes a...	pos
24996	After seeing this film I feel like I know just...	pos
24997	first this deserves about 5 stars due to actin...	neg
24998	If you like films that ramble with little plot...	neg
24999	As interesting as a sheet of cardboard, this d...	neg

Figure 1. IMDB Reviews Dataset

	Score	Summary	Text
0	2	One of the best game music soundtracks - for a...	Despite the fact that I have only played a sma...
1	1	Batteries died within a year ...	I bought this charger in Jul 2003 and it worke...
2	2	works fine, but Maha Energy is better	Check out Maha Energy's website. Their Powerex...
3	2	Great for the non-audiophile	Reviewed quite a bit of the combo players and ...
4	1	DVD Player crapped out after one year	I also began having the incorrect disc problem...
...	...	...	...
399994	1	Unbelievable- In a Bad Way	We bought this Thomas for our son who is a hug...
399995	1	Almost Great, Until it Broke...	My son recieved this as a birthday gift 2 mont...
399996	1	Disappointed !!!	I bought this toy for my son who loves the "Th...
399997	2	Classic Jessica Mitford	This is a compilation of a wide range of Mitfo...
399998	1	Comedy Scene, and Not Heard	This DVD will be a disappointment if you get i...

Figure 2. Amazon Reviews Dataset

tokenizer to convert the review texts into vectors, enabling us to input them into the model, as neural networks can only process numerical data.

For optimization, we opted for Adam, a powerful optimizer widely employed in contemporary machine learning tasks. Although we experimented with different learning schedulers, we observed little to no impact on the overall performance. To expedite the training process, we transferred all tensors and models to the GPU before initiating the training. Given that we were fine-tuning the model rather than pretraining it, we utilized only 3 to 5 epochs, with 3 epochs yielding the best results.

After fine-tuning, we repeated the same preprocessing steps for the Amazon review dataset. We set the model to evaluation mode for testing and also tested the original model on the Amazon dataset without fine-tuning on the IMDB movie review dataset. Finally, we compared the results of these two different approaches to identify patterns.

4. Model

We chose four different transformer models for this project, they are BERT-base-cased, BERT-base-uncased, BERT-large-uncased and DistilBERT. BERT-base, with a comparatively smaller number of parameters, strikes a balance between model size and performance, making it more suitable for resource-efficient applications. On the other hand, BERT-large, characterized by a significantly larger parameter count, excels in capturing intricate linguistic nuances, albeit at the cost of increased computational demands. DistilBERT, a distilled version of BERT designed for efficiency, maintains a competitive performance with reduced parameters, making it a compelling choice for scenarios with limited computational resources and general purpose tasks. From figure 3, we can see can that BERT-base has 110 million parameters and BERT-large has 340 million parameters, while DistilBERT only has 66 million parameters.

Comparison	BERT October 11, 2018	RoBERTa July 26, 2019	DistilBERT October 2, 2019	ALBERT September 26, 2019
Parameters	**Base:** 110M **Large:** 340M	**Base:** 125 **Large:** 355	**Base:** 66	**Base:** 12M **Large:** 18M
Layers / Hidden Dimensions / Self-Attention Heads	**Base:** 12 / 768 / 12 **Large:** 24 / 1024 / 16	**Base:** 12 / 768 / 12 **Large:** 24 / 1024 / 16	**Base:** 6 / 768 / 12	**Base:** 12 / 768 / 12 **Large:** 24 / 1024 / 16
Training Time	**Base:** 8 x V100 x 12d **Large:** 280 x V100 x 1d	1024 x V100 x 1 day (4-5x more than BERT)	**Base:** 8 x V100 x 3.5d (4 times less than BERT)	[not given] **Large:** 1.7x faster
Performance	Outperforming SOTA in Oct 2018	88.5 on GLUE	97% of BERT-base's performance on GLUE	89.4 on GLUE
Pre-Training Data	BooksCorpus + English Wikipedia = 16 GB	BERT + CCNews + OpenWebText + Stories = 160 GB	BooksCorpus + English Wikipedia = 16 GB	BooksCorpus + English Wikipedia = 16 GB
Method	Bidirectional Transformer, MLM & NSP	BERT without NSP, Using Dynamic Masking	BERT Distillation	BERT with reduced parameters & SOP (not NSP)

Figure 3. https://360digitmg.com/blog/bert-variants-and-their-differences

4.1. Pretrained Models

All of our pretrained model are first fine-tuned on HuggingFace IMDB dataset.

1. BERT-large-uncased: First, we created a new classifiers with the backbone as the pretrained model Bert-large-uncased from Huggingface, and append a new Linear layers to satisfy our text classification task. Since our BERT-large-uncased model is relatively big in terms of model size, we will finetune on 2 V100 GPUs using DataParralell. We finetune our model over 10 epochs, and record the training / validation loss over epochs.

2. BERT-base-cased: Similar to our BERT-large-uncased, we also append a new Linear layers for our classification task. This pretrained model only require 1 GPU V100 to train. We also finetune our model over 10 epochs, and record the training / validation loss over epochs.

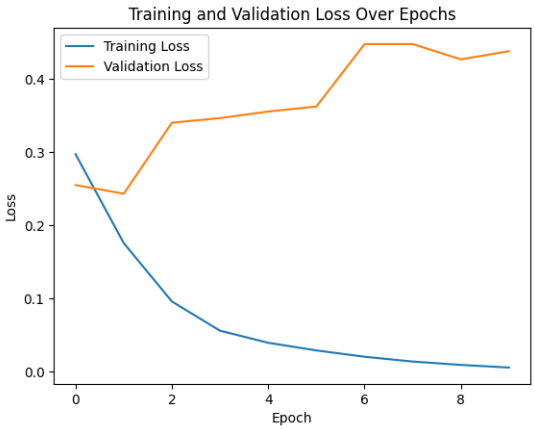

Figure 4. Finetuning BERT-large-uncased over 10 epochs

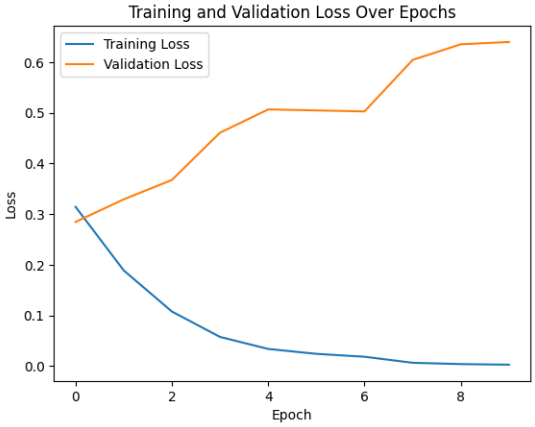

Figure 5. Finetuning BERT-base-cased over 10 epochs

3. BERT-base-uncased: Similar to BERT-base-cased, this pretrained model only require 1 GPU V100 to train. We want to check how is the performance of the model when all of the text has been preprocessed to be lower case.

4. DistilBERT: Similar to other Bert models, the original model was the pretrained model from hugging face. Unlike the previous models, DistilBERT was fine-tuned to be a more lightweight and efficient version of BERT, designed to compress the original model's size while retaining its performance. DistilBERT aims to provide a more resource-efficient alternative for various natural language processing tasks, making it suitable for deployment in environments with limited computational resources.

5. Final Result

After finetuning, we tested our model on Amazon review datasets. With reviews ratings 1-2, it will be labeled as 0 (Negative), 4-5 will be labeled as 1 (Positive). All of the neutral reviews with a rating of 3 will be ignored, since it is a neutral opinion. We then use our fine-tuned model predicts the Amazon dataset and record its accuracy. It is interesting to see that the pretrained BERT-large-uncased at the beginning when trying to predict if an Amazon review is negative or positive is worse compared to the BERT base model, having the accuracy of only 38.81% compared to 50% of BERT-base-uncased. For sentiment analysis task, the normal BERT model seems to do more than a sufficient job compared to much larger model.

BERT model	Before Finetune (%)	After Finetune (%)
DistilBERT	88.20	88.27
base-uncased	50.00	90.33
base-cased	50.01	89.01
large-uncased	38.81	90.62

Table 1. Performance of BERT models on Amazon reviews dataset

We also noticed that for DistilBERT the improvement is only 0.07%, which is lower than expected. After conducting the research, we found that during the distillation process, the DistilBERT has already been fine-tuned on sentiment analysis tasks. Even so, we can still see the benefit of fine-tuning it on IMDB dataset.

Overall, we have accomplished the goal we set at the beginning by evaluating the percentage increment on the test set. We realized even smaller model can have reasonably good performance comparing to the larger models. We also learned that Distillation process can inherit the tasks' features that larger models were trained for.

All of the code for training, testing is available in this Github link: https://github.com/QuanPham99/FinetuneBert-TextClassification

6. Future Work

If we proceed with the work, we would like to improve the project in the following ways. First, we aim to test our theory on more models, such as ALBERT and RoBERTa. This will allow us to identify a more distinguished trend or exceptions to our theory. Second, we intend to fine-tune our models on a larger dataset; the IMDB dataset only contains 25,000 reviews. It is necessary to investigate whether there is a relationship between dataset size and fine-tuned results. Third, we can explore different structural options. For example, we can attempt to add more layers to the current models and assess if this leads to improved performance.

7. Contribution List

Duc Minh Nguyen: Fine-tuned and tested BERT based uncased model. Wrote up approach, model, dataset and part

of pretrained models and final result sections.

Quan Pham: Fine-tuned and tested BERT-large-uncased, BERT-base-cased. Wrote up models, dataset, and final results.

Haoran Hu: Fine-tuned and tested DistilBERT. Wrote up abstract, motivation, approach, model, future work and part of pretrained models and final result sections.