Replication of DistilBERT evaluation results

Group M - Nadun Chandrabahu, Tai Ho, Muhammad Umer Bashir, Thanushreyas

November 1, 2024

1 Introduction

Reproducible research is vital in machine learning for several reasons. Firstly, it enables the validation of results, when research can be consistently reproduced, it enhances the credibility of the original findings. Additionally, reproducibility supports benchmarking for comparing algorithms and models, allowing the machine learning community to identify the best-performing models. It also plays an important role in identifying errors and biases, such as coding mistakes or data handling issues, improving the reliability of the research.

The paper, (Sanh et al, 2019)[1] "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter", we selected for this project presents a method for training a more compact general-purpose language representation model called DistilBERT, which is a streamlined version of the larger BERT model (Bidirectional Encoder Representations from Transformers). Although BERT is highly effective for Natural Language Processing (NLP) tasks, its size and complexity can lead to slower inference times, particularly in on-the-edge or resource-constrained environments, potentially impacting user experience. This paper aims to pre-train the DistilBERT model to enhance inference performance by reducing both its size and inference time while maintaining accuracy. According to the paper, DistilBERT achieves a 40% reduction in model size, preserves 97% of BERT's language understanding capabilities, and improves processing speed by 60%.

The source files and collaborative work done on this project is available for review on GitHub[2].

2 Project Justification

In this project, we will not replicate the transfer learning or model training processes of DistilBERT. Instead, we aim to reproduce the evaluation metrics shown in Tables 1 and 2 on page 3 of the referenced research paper, across various datasets and downstream

[1](Sanh et al, 2019): https://arxiv.org/pdf/1910.01108

[2]GitHub Repository for project: https://github.com/nadunchandrabahu/COMP8240-GroupM

tasks. This will allow us to assess DistilBERT's performance both on the datasets from the paper and on some of our own datasets.

The DistilBERT model is available for use through Hugging Face's transformers library (Wolf et al., 2019). In the original research, DistilBERT was evaluated using three datasets: the GLUE (General Language Understanding and Evaluation) benchmark [3], which includes nine sentence-pair language understanding tasks; IMDb [4], used for sentiment classification of user movie reviews; and SQuAD [5] (Stanford Question Answering Dataset), designed for question-answering based on a provided passage.

We will run some Python scripts to make predictions on the data with the DistilBERT model and calculate various metrics: the average score across the nine GLUE benchmark tasks, test accuracy on the IMDb dataset, and the Exact Match (EM) and F1 score on the SQuAD task. These metrics are reported in Tables 1 and 2 of the research paper and it is the aim of this project to replicate these results. Each group member will also show the metrics using new data sources to perform similar NLP tasks with DistilBERT.

Table 1 below shows a summary of GLUE average score as shown in the research paper.

Table 1: BERT and DistilBERT results on GLUE tasks

	Metrics									
Model.Name	Score	CoLA	MNLI	MRPC	QNLI	QQP	RTE	SST.2	STS.B	WNLI
BERT-base	79.5	56.3	86.7	88.6	91.8	89.6	69.3	92.7	89	53.5
DistilBERT	77	51.3	82.2	87.5	89.2	88.5	59.9	91.3	86.9	56.3

Table 2 below shows the test accuracy of IMDb sentiment analysis tasks, and EM/F1 scores of the SQuAD question answering task as shown on the research paper.

Table 2: IMDb and SQuAD metrics

	IMDb	SQuAD	
Model Name			
	acc.	EM	F1
BERT-base	93.46	81.2	88.5
DistilBERT	92.82	77.7	85.8

[3]GLUE Benchmark: https://gluebenchmark.com/

[4]IMDb dataset: https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

[5]SQuAD v1.1 dataset: https://rajpurkar.github.io/SQuAD-explorer/

3 DistilBERT

DistilBERT is a student model of BERT

The trained model is already available through a python library called transformers which is provided by hugging face.com (add reference). There are a few variations of the model available as well depending on what kind of data the model was trained on.

online link to SQuAD data set: https://github.com/rajpurkar/SQuAD-explorer/blob/master/dataset/dev-v1.1.jsonhttps://github.com/rajpurkar/SQuAD-explorer/blob/master/dataset/dev-v1.1.json

3.1 Evaluation Datasets

We'll have a look at the benchmark datasets GLUE, IMdb, SQuAD, we can also talk about how we can use

3.2 Replication of Evaluation Results

Show the coding and use the model on the datasets used in the paper such as GLUE benchmark, IMdb, SQuAD. Calculate the metrics and show how close they are to the ones in the paper

Include link to our github repo. And how we can replicate. Maybe we can include a notebook file running the evaluation on the datasets. We're using python version 3.12.7 to install transformers library so we can get the model. Also need to install pytorch

We can run python code as follows, calculate metrics and report on them. Let's have a discussion about the results on the datasets used in the paper in this section as well.

```
from transformers import pipeline
print("Hello GroupM!")
```

3.3 New data construction

Each member can talk about how they constructed their data.

3.3.1 Tai Ho's dataset

Talk about how you created your dataset, and what your goal /context of the dataset is.

3.3.2 Nadun Chandrahu's dataset

I'll be talking about a question and answer dataset, similar to SQuAD, I can manually anotate whether the answers are correct or not.

3.3.3 Evaluation using new datasets

3.3.4 Tai Ho's dataset

Talk about how your dataset evaluates with the model

3.3.5 Nadun Chandrabahu's dataset

Talk about how your dataset evaluates with the model

3.4 Reflections

Each of us can write a paragraph about our reflections on the project.

4 References

1. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. https://doi.org/10.48550/arXiv.1910.01108

2. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Platen, P. V., M, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., … Drame, M. (2021). HuggingFace's Transformers: State-of-the-art Natural Language Processing. Webology. https://doi.org/10.48550/arXiv.1910.03771