README *for* On Finetuning Large Language Models

Yu Wang
Fudan Institute for Advanced Study for Social Sciences
Fudan University

This README file contains four sections. The first section, *General Description*, describes the contents in this replication folder. The second section, *Hardware Requirements*, describes what hardware is used for running the scripts. The third section, *Instructions*, reports on how to run the scripts. The last section, *Package Versions*, reports on the versions of the packages used in the paper.

1 General Description

These are the materials to replicate results displayed in the paper "On Finetuning Large Language Models." The replication package contains the following files:

- README.pdf
- 01_Baseline_Replication.ipynb
- 02_ConfliBERT_Unrestricted_V0.ipynb
- 03_ConfliBERT_Unrestricted_V1.ipynb
- 04_ConfliBERT_Max_Length.ipynb
- 05_ConfliBERT_Parameters.ipynb
- 06_Corrected_Baseline.ipynb
- 07_Sequence_Length.ipynb
- 08_ConfliBERT_Unrestricted_V1_Large_Learning_Rate.ipynb
- cw_texts_clean_bert.csv[1]

[1]Should the Google Drive link in the Google Colab scripts become inaccessible, readers can download and use this dataset.

2 Hardware Requirements

The replication package has been tested for running on A100 GPUs on Google Colab.[2] Exact performance metrics would vary to some extent depending on the type of GPUs used, but they would not affect the key results. Exact performance metrics should be replicable on the same type of GPUs.

3 Instructions

Main paper

The statistics in Figure 1 in the main paper come from ConfliBERT_Parameters.ipynb.

Table 1 in the main paper is created using the logs in (1) 01_Baseline_Replication.ipynb and (2) 03_ConfliBERT_Unrestricted_V1.ipynb.

Table 2 in the main paper is created using results from the following three files: (1) 01_Baseline_Replication.ipynb for Column 3, (2) 03_ConfliBERT_Unrestricted_V1.ipynb for Column 4, and (3) 04_ConfliBERT_Max_Length.ipynb for Column 5.

In addition, statistics on sequence length in Footnote 8 are calculated using the script 07_Sequence_Length.ipynb.

Appendix

Table 1 in the appendix builds on top on Table 2 in the main paper. It contains new results for ConfliBERT+Pooler Layer. This set of results is created by 06_Corrected_Baseline.ipynb.

Table 2 in the appendix records the running time of training and evaluation for the following scripts: 01_Baseline_Replication.ipynb for Column 1, 06_Corrected_Baseline.ipynb for Column 2, 02_ConfliBERT_Unrestricted_V0.ipynb for Column 3, 03_ConfliBERT_Unrestricted_V1.ipynb for Column 4, and 04_ConfliBERT_Max_Length.ipynb for Column 5.

The key results, mean squared error (MSE) and R^2, are directly read from the logs. R^2 is printed directly about the line *"Now whole corpus."* MSE is one line above R^2.

[2]Note that usage of A100 GPUs on Google Colab is not free. Usage of some other types GPUs (e.g. T4) is free.

Figure 1: How to identify MSE and R^2 from the logs. The R^2 in this example is 0.77 (the line above *"Now whole corpus"*) and the MSE is 0.99.

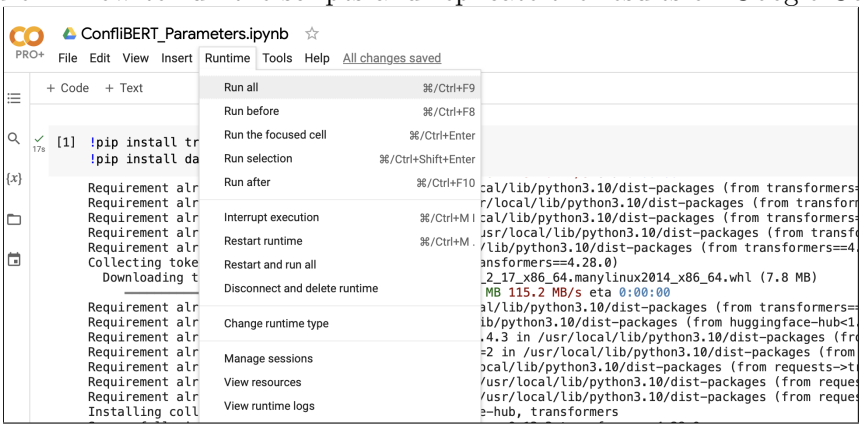
```
0.9905870703083635
0.7738313664183092
Now whole corpus
```

In addition, 08_ConfliBERT_Unrestricted_V1_Large_Learning_Rate.ipynb illustrates the importance of setting an appropriate learning rate and supports Footnote 1 in the appendix. When we set the learning rate to 2e-3, for example, the MSE that we get is 4.38 as compared to 0.99 when we set the learning rate to 2e-5.

Execution

Each of the scripts above can be run independently. In Google Colab, we only need to hit the **Run all** button under **Runtime** (Figure 2).[3] All the key results will be printed out during execution.

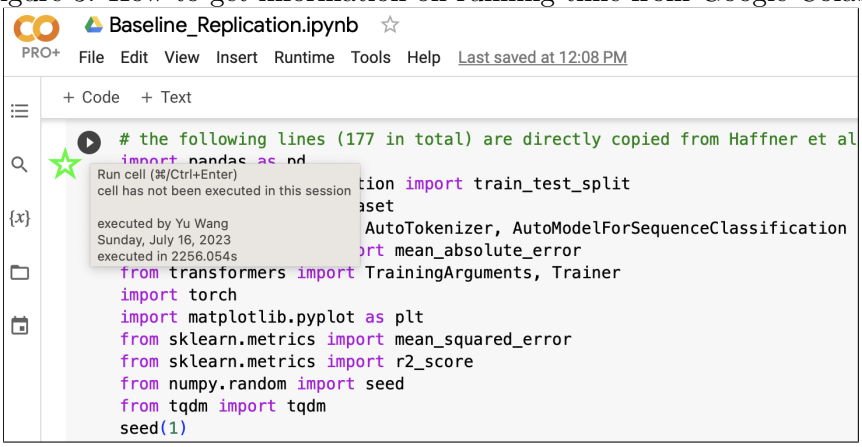
Figure 2: How to run the scripts and replicate the results on Google Colab.



To check the time that each step takes to complete, please hover over the dark run button, as illustrated in Figure 3. For example, running the training and evaluation steps in 01_Baseline_Replication.ipynb took about 2,256 seconds or 37 minutes.

[3] We can also run the scripts interactively if readers so prefer.

Figure 3: How to get information on running time from Google Colab.



4 Package Versions

We used Python 3.10.12 and pinned our transformer package to version 4.28.0. For all other packages, we used the default version provided on Colab. The versions of these other packages should not affect the replication results.