Do Your Homework – Poster Presentation

Author: Ammar Amjad – 5992-1730
Supervisor: Dr. Patrick Traynor

UF UNIVERSITY of FLORIDA

Introduction

- Most Published Research Findings Are False!
- We need trust to build on the work of others without verifying their authenticity each time.
- Yet there is no way to differentiate the good papers from the bad and Almost all the time can be spent on verifying the results.
- In this study, three papers are peer-reviewed based on the availability of open-source data and the code base on GitHub among other metrics.
- We identify the most common pitfalls of reproducibility.
- We further introduce improvements to the traditional way of reviewing papers which improves the time to validate the authenticity of a paper by up to 80%.

Methodology

The procedure employed was to observe the following steps for each paper, also shown in figure 1:

1. Reading Paper 1st time
2. Reading Paper 2nd time
3. Collecting Data
4. Writing Code
5. Run Code
6. Verifying Results
7. Perform Analysis

Metrics like Time taken to :

- ❖ Find Dataset
- ❖ Read the Paper
- ❖ Recreate Code if unavailable
- ❖ Verify Results

were used as credible criteria for reproducibility shown in Figure 3.

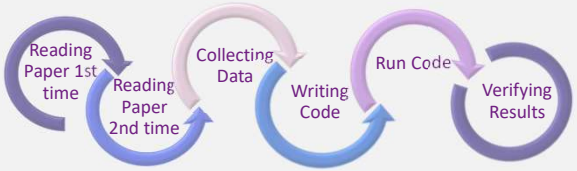

Figure 1: Steps in Chronological Order

Results

- Van Ede, Thijs, et al. DEEPCASE paper[1] is reproducible.
- Fu, Chuanpu, et al. Realtime robust malicious traffic detection paper[2] is partly reproducible due to the lack of clear instructions and incoherent directory structure.
- Yang, Yijun, et al. What You See is Not What the Network Infers paper[3] is not reproducible due to hardware limitations.

PREDICTION RESULTS. SYSTEMS TRAINED ON FIRST 20% OF DATA AND EVALUATED ON REMAINING 80% OF DATA. TIME SHOWS THE AVERAGE AMOUNT OF TIME FOR 1 EPOCH OF TRAINING. BEST PERFORMANCE IS HIGHLIGHTED IN **BOLD**.

	System	Precision	Recall	F1-score	Accuracy	Train time
HDFS	DeepLog	89.71%	89.34%	89.35%	89.34%	**1.0 s**
	Tiresias	89.70%	87.63%	87.96%	87.63%	15.0 s
	DEEPCASE	**90.41%**	**90.64%**	**90.40%**	**90.64%**	1.3 s

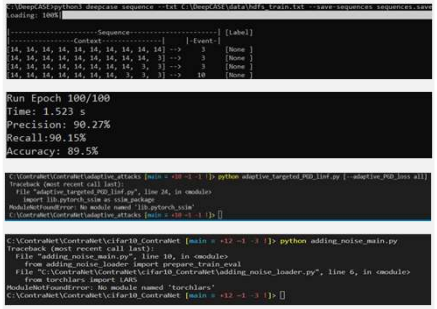

Figure 2: Execution of Codes and Results

Analysis

Most time is Spent on Recreating Code

Our analysis of the data collected has led us to believe that more than 80% of time is spent on verifying whether the paper is reproducible.

The first read of the paper takes only 20% of the total time, while the other steps in the methodology take the remaining time. So, we should aim to eliminate the remaining part.

Proper Code versioning and Docker-izing can alone solve this problem.

Papers with uncommon Hardware Components are Unlikely to Reproduce

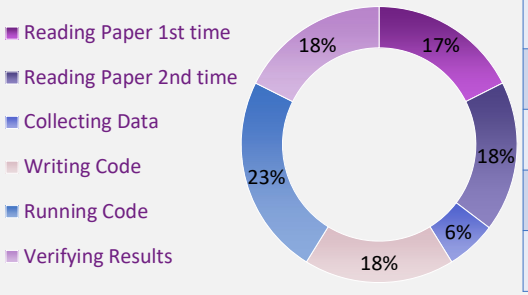

Figure 3: Time Taken per Activity

The most common errors are in the Data and Code

Our research has shown that most errors encountered while reproducing a paper are not in the paper itself but in addressing and dealing with data, code, and directory errors as shown in table 1.

More precisely, the bulk of the problems arise from not having clean usable data and code available online. Ranging from requirements and libraries not being clearly mentioned, runtime errors to the dataset requiring data preprocessing.

	Deep Case	Real Time Robust	What You See	Average
Data Errors	5	4	3	4
Code Errors	4	3	2	3
Directory Errors	1	2	1	1
Paper Errors	1	2	1	1

Table 1. Errors Encountered For Each Paper

Tools

pandas, Python, Excel, ANACONDA, Keras, TensorFlow, jupyter, NumPy, Windows

Future Direction

- Papers should have a YouTube video, a GitHub repo, and a Dockerized environment which can quicken the reproducibility process by up to 80%.
- A reproducible paper is more likely to more cited.
- Conclusively, Our research shows that rather than delve straight into paper reproduction, the reader should "Do Your Homework" and first check the availability of all assets needed to recreate the results.

References

1. van Ede, Thijs, et al. "DEEPCASE: Semi-Supervised Contextual Analysis of Security Events." IEEE Security and Privacy. 2022.
2. Fu, Chuanpu, et al. "Realtime robust malicious traffic detection via frequency domain analysis." Proceedings of 2021 ACM SIGSAC Conference on CSS. 2021.
3. Yang, Yijun, et al. "What You See is Not What the Network Infers: Detecting Adversarial Examples Based on Semantic Contradiction." 2022.