USC University of Southern California

Automating Assessment of Computational Reproducibility in Machine Learning Research

Victoria Stodden
Department of Industrial and Systems Engineering
University of Southern California

Joint Statistical Meetings Invited Paper
Session: Statistics and the Reproducibility Crisis
Toronto, Ontario, Canada
August 10, 2023

USC

Agenda

1. The Model Checking Value Proposition for Machine Learning

2. Leveraging Steps in the Community

3. Reproscreener: A New Tool for ML Model Checking (Preliminary Empirical Results)

4. Next steps

1. Model Checking Value Proposition: Potential Gains

Scalability to large-scale complex models

- e.g. integration of multiple large data sources, deployment on large scale / high throughput computing systems.

Verification of model performance

- Consistency of results / predictions across time, systems, data / Reproducibility

Transparency / interpretability

Efficiency in resource use / Discovery speedup

- Computing systems: compute time, appropriate benchmarking;
- Engineers: code re-use, reduction in effort duplication

USC

Model Checking Value Proposition: Potential Drawbacks

Increased overhead

- Additional computational step(s) in model building, deployment

Culture change

- Increased emphasis on reproducibility, verification, correctness in ML models

Hewing to the wrong goals

2. Community Efforts

ML model publication standards

- Gunderson (AAAI 2018)
- Pineau (JMLR 2020)

Formal Verification for ML models

- Abate (MEMOCODE 2017)
- Urban and Miné (arxiv 2021)

Research Publication Standards

- Willis and Stodden (HDSR 2020)
- ML Commons (Github)
- National Academies Reproducibility Report (NASEM 2019)

Many many more…

USC

ML Model Checking: A Novel Approach

- Previous work in AI involves applying formal techniques using SMT (satisfiability modulo theory) solvers, constraint solving, or abstract numerical interpretation.
- We exploit specialized features of ML pipelines and propose a *reproducibility* approach (NASEM 2019):
  - Exposure of methods
  - Well-defined guarantees in correctness of results

3. Automating ML Model Checking: Reproscreener

- Automate ML model checking *at the point of publication*, to provide guarantees on correctness, scalability, and transparency.

- Reproscreener software tool verifies criteria and provides feedback[1].

Available at reproscreener.org and https://github.com/Machine-Learning-Pipelines/reproscreener/

USC

Criteria used by Reproscreener

1. ML model criteria for publication based on Gunderson 2018.
2. Code/repo criteria (when found by Reproscreener) based on Krafczyk et al 2020.

Implemented in a labelled testbed of arXiv publications

- 50 most recent arXiv preprint submissions in stat.ML and CS.GL from October 25 2022.

Reproscreener Performance on Testbed Preprints (based on Gunderson 2018)

Metric	Proportion Correct (n=50)
Code available	0.82
Hypothesis stated	0.60
Experimental setup	0.54
Dataset available	0.48
Problem stated	0.36
Predicted result	0.30
Research method	0.28
Objective/Goal	0.28
Research question	0.16

USC

Reproscreener Performance on Code (based on Krafczyk et al 2020)

Metric	Proportion Correct (n=22)
Readme has dependencies info	0.45
Readme has setup instructions	0.45
Readme has requirements info	0.41
Readme has install instructions	0.41
Wrapper scripts	0.36
Dependency tracking files	0.32

ChatGPT4 Performance on Abstracts

Metric	Proportion Correct (n=50)
Code available	1.00
Research question	0.96
Hypothesis stated	0.88
Dataset available	0.88
Objective/Goal	0.88
Problem stated	0.82
Predicted result	0.52
Research method	0.46
Experimental setup	0.46

USC

ChatGPT mistake examples

“Problem stated”
GPT’s found phrase: Upcoming large astronomical surveys are expected to capture an unprecedented number of strong gravitational lensing systems.
Manually found phrase: The absence of large quantities of representative data from current astronomical surveys motivates the development
GPT’s conclusion: The problem is stated in the abstract. FALSE
Notes: The Problem here is the fact that large amounts of data is *missing* from surveys and not that the surveys are expected to capture a large number of systems.

“Dataset available”
GPT’s found phrase: Our investigation on 59 different USB flash drives---belonging to 17 brands, including the top brands purchased on Amazon in mid-2019---reveals a minimum classification accuracy of 98.2% in the identification of both brand and model, accompanied by a negligible time and computational overhead.
GPT’s conclusion: Dataset is available. FALSE

Extending "Gunderson 2018" Criteria

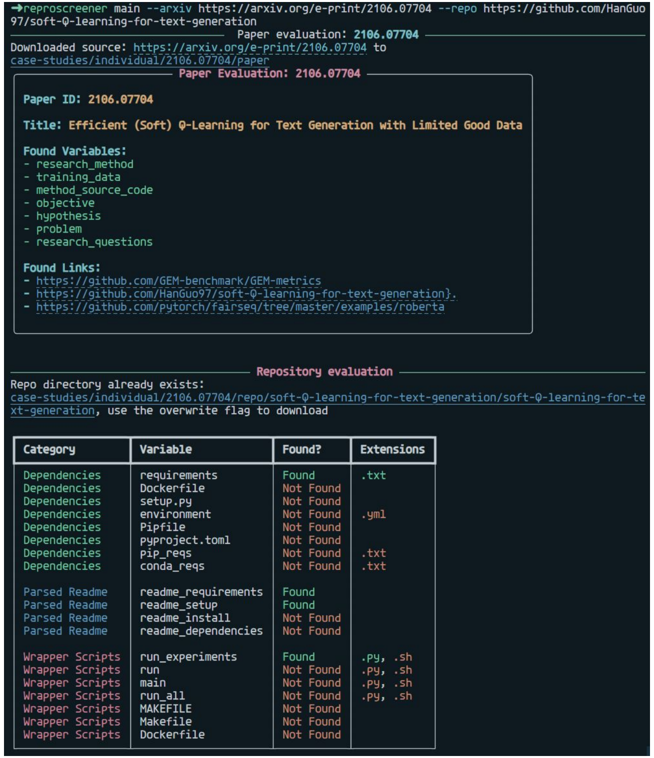
ReproScreener Goals:

- Automatically check specific guidances to improve correctness of ML models to predict error bounds, capture and identifies difference in model output at scale (due to architecture, non-determinism, etc.)
- Enable comparison of model code through:
  - Checking for modularity, file structure, dependencies.
  - Checking for steps/scripts to create figures & visualizations.
  - Tracking model benchmarks and provenance.

- Real world case studies to demonstrate ReproScreener's functionality

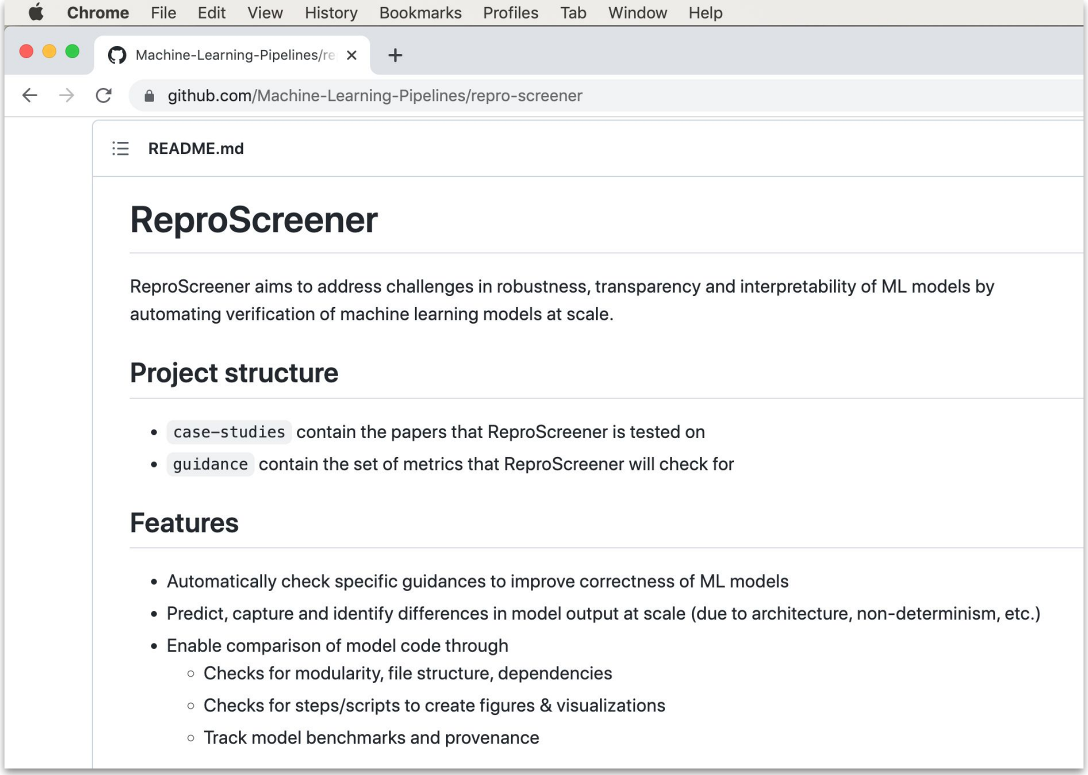
USC

Reproscreener Open Source Development (work in progress)



```
➜reproscreener main --arxiv https://arxiv.org/e-print/2106.07704 --repo https://github.com/HanGuo
97/soft-Q-learning-for-text-generation
                               Paper evaluation: 2106.07704
Downloaded source: https://arxiv.org/e-print/2106.07704 to
case-studies/individual/2106.07704/paper
                       Paper Evaluation: 2106.07704

 Paper ID: 2106.07704

 Title: Efficient (Soft) Q-Learning for Text Generation with Limited Good Data

 Found Variables:
 - research_method
 - training_data
 - method_source_code
 - objective
 - hypothesis
 - problem
 - research_questions

 Found Links:
 - https://github.com/GEM-benchmark/GEM-metrics
 - https://github.com/HanGuo97/soft-Q-learning-for-text-generation}.
 - https://github.com/pytorch/fairseq/tree/master/examples/roberta

                              Repository evaluation
Repo directory already exists:
case-studies/individual/2106.07704/repo/soft-Q-learning-for-text-generation/soft-Q-learning-for-te
xt-generation, use the overwrite flag to download
```

Category	Variable	Found?	Extensions
Dependencies	requirements	Found	.txt
Dependencies	Dockerfile	Not Found	
Dependencies	setup.py	Not Found	
Dependencies	environment	Not Found	.yml
Dependencies	Pipfile	Not Found	
Dependencies	pyproject.toml	Not Found	
Dependencies	pip_reqs	Not Found	.txt
Dependencies	conda_reqs	Not Found	.txt
Parsed Readme	readme_requirements	Found	
Parsed Readme	readme_setup	Found	
Parsed Readme	readme_install	Not Found	
Parsed Readme	readme_dependencies	Not Found	
Wrapper Scripts	run_experiments	Found	.py, .sh
Wrapper Scripts	run	Not Found	.py, .sh
Wrapper Scripts	main	Not Found	.py, .sh
Wrapper Scripts	run_all	Not Found	.py, .sh
Wrapper Scripts	MAKEFILE	Not Found	
Wrapper Scripts	Makefile	Not Found	
Wrapper Scripts	Dockerfile	Not Found	



ReproScreener

ReproScreener aims to address challenges in robustness, transparency and interpretability of ML models by automating verification of machine learning models at scale.

Project structure

- `case-studies` contain the papers that ReproScreener is tested on
- `guidance` contain the set of metrics that ReproScreener will check for

Features

- Automatically check specific guidances to improve correctness of ML models
- Predict, capture and identify differences in model output at scale (due to architecture, non-determinism, etc.)
- Enable comparison of model code through
  - Checks for modularity, file structure, dependencies
  - Checks for steps/scripts to create figures & visualizations
  - Track model benchmarks and provenance

USC

Conclusion: The Model Checking Value Proposition Revisited

Reproscreener:

1. can assist in automatically checking manuscripts and code for the satisfaction of relevant criteria,

2. is a research tool that enables us to study and refine criteria based on desired goals.

Goal: Boundedness guarantees regarding correctness of reproduced results compared to original ML pipeline.

Thank you!

Joint work with **Adhithya Bhaskar**, Ph.D. student
Department of Industrial and Systems Engineering
University of Southern California

This material is based upon work supported by the REAL@USC-META Center and National Science Foundation Grant No 2138776

USC