Article:

Biagiola, M., Cardozo, N., Shin, D. orcid.org/0000-0002-0840-6449 et al. (3 more authors) (2023) Summary of the fourth international workshop on deep learning for testing and testing for deep learning (DeepTest 2023). ACM SIGSOFT Software Engineering Notes, 48 (4). pp. 39-40. ISSN 0163-5948

https://doi.org/10.1145/3617946.3617953

Summary of the Fourth International Workshop on Deep Learning for Testing and Testing for Deep Learning (DeepTest 2023)

Matteo Biagiola
Università della Svizzera italiana
Lugano, Switzerland
matteo.biagiola@usi.ch

Nicolás Cardozo
Universidad de los Andes
Bogotá, Colombia
n.cardozo@uniandes.edu.co

Donghwan Shin
The University of Sheffield
Sheffield, United Kingdom
d.shin@sheffield.ac.uk

Foutse Khomh
Polytechnique Montréal
Montréal, Canada
foutse.khomh@polymtl.ca

Andrea Stocco
Technical University of Munich
Munich, Germany
andrea.stocco@tum.de

Vincenzo Riccio
University of Udine
Udine, Italy
vincenzo.riccio@uniud.it

ABSTRACT

Deep Learning (DL) techniques help software developers thanks to their ability to learn from historical information which is useful in several program analysis and testing tasks (e.g., malware detection, fuzz testing, bug-finding, and type-checking). DL-based software systems are also increasingly adopted in safety-critical domains, such as autonomous driving, medical diagnosis, and aircraft collision avoidance systems. In particular, testing the correctness and reliability of DL-based systems is paramount, since a failure of such systems would cause a significant safety risk for the involved people and/or environment. The 4$^{\text{th}}$ International Workshop on Deep Learning for Testing and Testing for Deep Learning (DeepTest 2023) was co-located with the 45$^{\text{th}}$ International Conference on Software Engineering (ICSE), with the goal of targeting research at the intersection of software engineering and deep learning and devise novel approaches and tools to ensure the interpretability and dependability of software systems that depends on DL components.

1. INTRODUCTION

The 4$^{\text{th}}$ International Workshop on Deep Learning for Testing and Testing for Deep Learning (DeepTest 2023) was co-located in conjunction with the 45$^{\text{th}}$ International Conference on Software Engineering (ICSE). The workshop was held in a hydrid format, with nearly 20 registered participants in presence and a number of virtual attendees.

DeepTest is a high-quality workshop for research at the intersection of Deep Learning (DL) and Software Engineering (SE). It is a forum for researchers and industry practitioners to share and discuss new DL-related technology as well as applications of DL to practical SE problems. The attendees are expected to have expertise in the SE domain, as well as experience or interest in DL, and their inter-connection.

DL is widely adopted in modern software systems, including safety-critical domains, such as autonomous cars, medical diagnosis, and aircraft collision avoidance systems. It is crucial to rigorously test such applications to ensure high reliability and safety. However, standard notions of software reliability become irrelevant when considering DL systems, due to their non-deterministic nature and the lack of a transparent understanding of the models' semantics. On the other hand, SE researchers and practitioners increasingly resort to DL approaches to devise novel solutions to address existing problems in the software development life-cycle. In particular, we welcome submissions that investigate:

- How to ensure the quality of DL-based applications, both at model level and system level.
- The use of DL to support software engineering tasks, particularly software testing.

2. KEYNOTES

This year, the DeepTest workshop included two keynotes and a panel of experts.

Baishakhi Ray from Columbia University delivered the opening keynote on "Testing Autonomous Driving Systems". It covers their recent efforts in automatically generating critical test-driving scenarios using three techniques: (1) AutoFuzz [6], a grammar-based black-box fuzzing technique to generate failure-inducing scenarios for Autonomous Driving Systems (ADSs), (2) FusED [5], an evolutionary and causality-based grey-box fuzzing technique to generate critical scenarios for fusion component of ADSs, and (3) CTG [7], a conditional diffusion model that generates realistic and user-controllable scenarios for ADSs.

Zhenchang Xing from Australian National University delivered the second keynote on "Testing Generative Large Language Model: Mission Impossible or Where Lies the Path?". With the widespread adoption of pre-trained generative language models, such as ChatGPT, some people believe that the emergent capabilities of large language models are turning AI from engineering into natural science, as it is hard to think of these models as being designed for a specific purpose in the traditional sense. The keynote explores intuitive questions for software engineering researchers and practitioners in the era of large language models, such as: Will differential testing, metamorphic testing, and adversarial testing, which are effective for testing discriminative models in specific tasks, no longer be the saviors of open-ended task testing for large language models? How can we test and correct ethical issues and hallucinations in generative AI?

3. PAPER PRESENTATION

Four papers were presented, which we shortly summarize below.

Gao et al. [2] address the problem of testing for DL-based machine translation models using metamorphic testing. They presented a method using back-translation as a reference for machine translation testing, minimizing the use of external natural language processing tools in the target language, so that the same workflow can be applied to test systems translating English to multiple languages.

Sakuma et al. [4] focus on the problem of automatically identifying the causes of prediction errors in DL models. They presented AIEDF, a flowchart-structured analysis method that integrates the results of a comprehensive analysis of data and models in an interpretable form. AIEDF is also model-agnostic and applicable to various models, such as gradient-boosting trees and neural networks.

Lin and Yu [3] discuss the interpretability of DL models using decision logic. They showed that there is a different distribution of DL model explanation signatures (e.g., SHAP) between normal and adversarial examples, as well as diverse decision logic of networks to distinguish them. They further extended the idea to improve the effectiveness of adversarial examples for DL models.

Bu and Sun [1] address the problem of repairing DL models, especially for Deep Neural Networks (DNNs), when the original training data is not available. They presented DeepPatch, a patching-based DNN repair method that improves the DNN's accuracy on corrupted images while avoiding too much impact on clean images.

4. EXPERT PANEL

The panel discussed future challenges with respect to testing deep learning systems and featured three experts from academia, some with industrial collaboration experience. Topics covered include software in the loop testing, simulators fidelity, testing challenges for Large Language Models (LLMs), and the mitigation of reality gaps for cyber-physical systems.

List of Panelists:

- Lionel C. Briand · University of Ottawa, Canada
- Baishakhi Ray · Columbia University, USA
- Paolo Arcaini · National Institute of Informatics, Japan

5. CONCLUSIONS

The workshop concluded with a reflection on the challenges and opportunities that lay ahead for research on deep learning testing. In addition, the 5th DeepTest workshop is planned for 2024 and is expected to be co-located again with the 46th International Conference on Software Engineering (ICSE).

6. ACKNOWLEDGMENTS

We thank the program committee, the authors, the invited speakers, the panelists, and the participants who contribute to the success and quality of this workshop. We also thank the student volunteers, the ICSE 2023 Workshop Chairs and the entire ICSE 2023 organization team for providing us with the necessary support to organize and conduct the workshop, including an amazing social event.

7. REFERENCES

[1] Hao Bu and Meng Sun. DeepPatch: A Patching-Based Method for Repairing Deep Neural Networks. In *2023 IEEE/ACM 45th International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)*. IEEE Computer Society, 2023.

[2] Wentao Gao, Jiayuan He, and Van-Thuan Pham. Metamorphic Testing of Machine Translation Models using Back Translation. In *2023 IEEE/ACM 45th International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)*. IEEE Computer Society, 2023.

[3] Yi-Ching Lin and Fang Yu. DeepSHAP Summary for Adversarial Example Detection. In *2023 IEEE/ACM 45th International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)*. IEEE Computer Society, 2023.

[4] Keita Sakuma, Ryuta Matsuno, and Yoshio Kameda. A Method of Identifying Causes of Prediction Errors to Accelerate MLOps. In *2023 IEEE/ACM 45th International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)*. IEEE Computer Society, 2023.

[5] Ziyuan Zhong, Zhisheng Hu, Shengjian Guo, Xinyang Zhang, Zhenyu Zhong, and Baishakhi Ray. Detecting multi-sensor fusion errors in advanced driver-assistance systems. In *Proceedings of the 31st ACM SIGSOFT International Symposium on Software Testing and Analysis*, pages 493–505, 2022.

[6] Ziyuan Zhong, Gail Kaiser, and Baishakhi Ray. Neural network guided evolutionary fuzzing for finding traffic violations of autonomous vehicles. *IEEE Transactions on Software Engineering*, 2022.

[7] Ziyuan Zhong, Davis Rempe, Danfei Xu, Yuxiao Chen, Sushant Veer, Tong Che, Baishakhi Ray, and Marco Pavone. Guided conditional diffusion for controllable traffic simulation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3560–3566. IEEE, 2023.