Doctoral School of Information and Biomedical Technologies Polish Academy of Sciences - TIB-PAN

Research domain: Informatyka Techniczna i Telekomunikacja

Topic: 1.4 Uczenie maszynowe – zagadnienia specjalne

Robustness of machine learning models considering adversarial conditions

Supervisor; contact information

dr hab. Joanna Kołodziej (main supervisor);
tel. 601688140, joanna.kolodziej@nask.pl;
Naukowa i Akademicka Sieć Komputerowa – Państwowy Instytut Badawczy (NASK-PIB),
ul. Kolska 12, Warszawa
dr inż. Mateusz Krzysztoń (co-supervisor)
mateuszkr@nask.pl
Naukowa i Akademicka Sieć Komputerowa – Państwowy Instytut Badawczy (NASK-PIB),
ul. Kolska 12, Warszawa

Scope

The rapid development of machine learning (ML) has enabled the solution of many complex problems in various fields. However, ML models are subject to certain types of logical weaknesses due to the inherent limitations of the learning algorithms [1]. Therefore, specific testing techniques and the consideration of additional thread models are required when deciding to use ML. It is essential in areas where the cost of decision model failure is high. Thus, standards and methods for testing ML models are now a solid and pressing industry need [2].

During the standard ML model development process, several quality metrics are considered. These include decision quality, performance stability, learning speed or decision latency. However, the metrics obtained in the standard testing process may be overly optimistic - ML models are rarely tested:

- with data sets that reflect real-world data characteristics (e.g. class distributions) [3];
- for robustness to changes in data characteristics (i.e. concept drift);
- in an adversarial environment. A new branch of machine learning called Adversarial Machine Learning (AML) [4] studies attacks on machine learning algorithms and defences against such attacks. AML techniques make it possible to measure the model's resistance to adversarial user behaviour (so-called adversarial robustness);

As part of the research, existing methods, tools and measures will be reviewed and selected. If necessary, new ones will be developed to study the general resilience of machine learning models. The overall process of ML model testing will be designed.

It is often necessary to update models frequently and quickly, and therefore usually automatically, based on new data, making it necessary to automate the testing process. In

many cases, to reduce the refresh rate of the models and increase the system’s resilience, it is required to monitor the quality of the model on real data in real-time to react appropriately to possible quality degradation. Therefore, considering the automation of the proposed methods in a production environment or enabling real-time model monitoring using the proposed strategies could be an additional area of work.
The work will be carried out in cooperation with the Centre for Standardization and Certification of NASK-PIB.

Requested skills:

- MSc degree in computer sciences telecommunication or similar discipline,
- Backgrounds in machine learning,
- Advanced practical knowledge of Python/Java
- Experience in at least one ML tool (e.g. Tensorflow, PyTorch, scikit-learn)
- Advanced Level in English (speaking and writing).

References

1. Bitton, Ron, et al. "Adversarial machine learning threat analysis in open radio access networks." arXiv preprint arXiv:2201.06093 (2022).
2. ETSI 5G PoC Consortium Steering Committee and Contributors, Artificial Intelligence (AI) in Test Systems, Testing AI Models and ETSI GANA Model's Cognitive Decision Elements (DEs) via a Generic Test Framework for Testing GANA Multi-Layer Autonomics & their AI Algorithms for Closed-Loop Network Automation, 2020, online: https://intwiki.etsi.org/images/ETSI_5G_PoC_White_Paper_No_5.pdf
3. Borovicka, Tomas, et al. "Selecting representative data sets." Advances in data mining knowledge discovery and applications 12 (2012): 43-70.
4. Laskov, Pavel, and Richard Lippmann. "Machine learning in adversarial environments." Machine learning 81 (2010): 115-119.

Warsaw, May, 2024