Internship position: Out-of-distribution detection for adversarial attacks evasion

Keywords: Explainability, Neural Network, Out-of-distribution detection

Institution

The French Alternative Energies and Atomic Energy Commission (CEA) is a key player in research, development, and innovation. Drawing on the widely acknowledged expertise gained by its 16,000 staff spanned over 9 research centers with a budget of 4.1 billion Euros, CEA actively participates in more than 400 European collaborative projects with numerous academic (notably as a member of Paris-Saclay University) and industrial partners. Within the CEA Technological Research Division, the CEA List institute addresses the challenges coming from smart digital systems.

Among other activities, CEA List's Software Safety and Security Laboratory (LSL) research teams design and implement automated analysis in order to make software systems more trustworthy, to exhaustively detect their vulnerabilities, to guarantee conformity to their specifications, and to accelerate their certification. The lab recently extended its activities on the topic of AI trustworthiness and gave birth to a new research group: AISER (Artificial Intelligence Safety, Explainability and Robustness).

Scientific context

Through the recent developments of AI, the use of models produced by machine learning has become widespread, even in industrial settings. However, studies are flourishing showing the dangers that such models can bring, in terms of safety, privacy or even fairness. To mitigate these dangers and improve trust in AI, one possible avenue of research consists in designing methods for generating explanations of the model behaviour. Such methods, regrouped under the umbrella term "eXplainable AI" (XAI), empower the user by providing them with relevant information to make an informed choice to trust the model (or not).

Another important topic is to assess the correct domain of operation of a neural network. Indeed, inputs of a neural network are expected to be drawn from a distribution similar to the training set. To put it bluntly, a program trained to detect pedestrians on a road should not be expected to perform well when presented with pictures of planes. As

embedding such limitation in a neural network is unfeasible, there had been a lot of work on the field of "out-of-distribution detection" (OOD-detection).

Through several work Xu-Darme, Girard-Satabin, et al. (2023), the AISER team bridged XAI and OOD-detection together, using case-based reasoning techniques to detect distribution shift from an input. This ability can be used to other means, for instance monitoring the presence of maliciously modified samples (for instance, adversarial examples (Szegedy et al. 2014).

Internship

During this internship, you will study the use of AISER's OOD-detection method, PARTICUL, to identify whether new inputs were tampered with. You will work using the open-source library CaBRNet(Xu-Darme et al. 2024), developped at CEA LIST, which provides an implementation of PARTICUL.

The broad internship goals are:

- familiarization with the state-of-the-art on XAI (Molnar 2022), OOD-detection Tajwar et al. (2021) and adversarial examples Carlini and Wagner (2016);
- getting started with the PARTICUL implementation in CaBRNet;
- design and implementation of benchmarks involving the tampering of whole datasets with adversarial examples;
- evaluation against other OOD-detection methods using for instance the Open-OOD benchmark (Yang et al. 2022)

Qualifications

The candidate will work at the confluence of numerous topics: artificial intelligence, machine learning and cybersecurity. As it is not realistic to be expert in all fields, we encourage candidates that do not meet the full qualification requirements to apply nonetheless. We strive to provide an inclusive and enjoyable workplace. We are aware of discriminations based on gender (especially prevalent on our fields), race or disability, we are doing our best to fight them. One of our team member is formally trained against psychological harassment and sexual abuse.

Minimal

- master student or equivalent (2nd/3rd engineering school year) in computer science;
- ability to work in a team;
- some knowledge of version control;

Preferred

- formal training in machine learning and/or statistics

- experience of machine learning theory

Characteristics

The candidate will be monitored by two research engineers of the team, with at least one weekly meeting.

Duration

5 to 6 months from early 2024

Locations

Two locations are possible for this internship

- CEA Nano-INNOV, Paris-Saclay Campus, France
- CEA Grenoble, 17 Av. des Martyrs, 38000 Grenoble, France

Compensation and funding

- 1400€ monthly stipend
- possible allowance for housing and travel expense (in case a relocation is needed) with a limit of €229 per month
- 75% refund of transit pass
- subsidized lunches
- 3 days of remote work

Application

If you are interested in this internship, please send to the contact persons an application containing:

- your resume;
- a cover letter indicating how your curriculum and experience match the qualifications expected and how you would plan to contribute to the project;
- your bachelor and master 1 transcripts;
- the contact details of two persons (at least one academic) who can be contacted to provide references.

Applications are welcomed until the position is filled. Please note that the administrative processing may take up to 3 months.

Contact persons

For further information or details about the internship before applying, please contact:

- Julien Girard-Satabin (julien.girard2@cea.fr)

- Romain Xu-Darme (romain.xu-darme@cea.fr)

References

Carlini, Nicholas, and David Wagner. 2016. "Towards Evaluating the Robustness of Neural Networks." arXiv. https://doi.org/10.48550/ARXIV.1608.04644.

Ilyas, Andrew, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. 2019. "Adversarial Examples Are Not Bugs, They Are Features." arXiv. https://doi.org/10.48550/ARXIV.1905.02175.

Molnar, Christoph. 2022. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2nd ed. https://christophm.github.io/interpretable-ml-book.

Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. "Intriguing Properties of Neural Networks." In *2nd International Conference on Learning Representations, ICLR 2014*.

Tajwar, Fahim, Ananya Kumar, Sang Michael Xie, and Percy Liang. 2021. "No True State-of-the-Art? OOD Detection Methods Are Inconsistent Across Datasets." *ArXiv* abs/2109.05554.

Xu-Darme, Romain, Julien Girard-Satabin, Darryl Hond, Gabriele Incorvaia, and Zakaria Chihani. 2023. "Contextualised Out-of-Distribution Detection Using Pattern Identification." In *Computer Safety, Reliability, and Security. SAFECOMP 2023 Workshops*, edited by Jérémie Guiochet, Stefano Tonetta, Erwin Schoitsch, Matthieu Roy, and Friedemann Bitsch, 423–35. Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-40953-0_36.

Xu-Darme, Romain, Georges Quénot, Zakaria Chihani, and Marie-Christine Rousset. 2023. "PARTICUL: Part Identification with Confidence Measure Using Unsupervised Learning." In *Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges*, edited by Jean-Jacques Rousseau and Bill Kapralos, 173–87. Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-37731-0_14.

Xu-Darme, Romain, Aymeric Varasse, Alban Grastien, Julien Girard-Satabin, and Zakaria Chihani. 2024. "CaBRNet, an Open-Source Library for Developing and Evaluating Case-Based Reasoning Models." *The 2nd World Conference on eXplainable Artificial Intelligence*.

Yang, Jingkang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, WENXUAN PENG, Haoqi Wang, et al. 2022. "OpenOOD: Benchmarking Generalized Out-of-Distribution Detection." In *Thirty-Sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.