In-painting using generative AI for the correctness evaluation of eXplainable AI (XAI) methods

Keywords: Explainable AI, Generative AI, Evaluation, Convolutional Neural Networks

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).

In the field of XAI, multiple metrics have been proposed to evaluate the correctness of an explanation, i.e. how well the explanation reflects the actual AI model behaviour. In the particular context of computer vision, most evaluation metrics from the state of the art propose to "de-activate" pixels (e.g. replacing them with a black pixels) to measure their impact on the model decision. However, recent work has shown that such metrics might not be informative, in the sense that they tend to evaluate the model behaviour on images that do not belong the training distribution and that can be considered as “out of distribution”: indeed, an image with entire regions painted in black can hardly be considered as a “normal” input that the model should expect during its lifecycle.

Internship

In this internship, we propose to use generative AI to de-activate pixels in a more subtle way - creating images that resemble the original one but with missing features while remaining “in distribution” - and to study the impact of such method on the evaluation of the correctness of XAI methods.

More precisely, the internship will be split in several subtasks as follows:

- Establish a baseline of existing metrics for evaluating the correctness of XAI methods, using the Quantus framework.
- Identify a body of existing works on the use of generative AI for in-painting and select a method based on a set of motivated criteria
- Implement the selected method and evaluate the advantages and drawbacks of the resulting evaluation metric, compared to the state of the art

Qualifications

As it is not realistic to be expert in machine-learning, computer vision and XAI, 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.

- **Minimal**
  - Master student or equivalent (2nd/3rd engineering school year) in computer science
  - knowledge of Python and the Pytorch framework
  - ability to work in a team, some knowledge of version control
- **Preferred**
  - notions of AI and neural networks
  - notions of Computer Vision
  - notions of explainable AI

Characteristics

The candidate will be supervised by a research engineer.

- **Duration:** 5 to 6 months from early 2024

- **Location:** CEA Grenoble
- **Compensation:**
  - 1400€ monthly stipend
  - possible allowance for housing and travel expense (in case a relocation is needed)
  - 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;

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:

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