A propos du centre ou de la direction fonctionnelle

The Inria Rennes - Bretagne Atlantique Centre is one of Inria's eight centres and has more than thirty research teams. The Inria Center is a major and recognized player in the field of digital sciences. It is at the heart of a rich R&D and innovation ecosystem: highly innovative PMEs, large industrial groups, competitiveness clusters, research and higher education players, laboratories of excellence, technological research institutes, etc.

Contexte et atouts du poste

Context AI-based decision-making systems are now pervasive, serving populations in most parts of their online interactions (i.e., curation such as recommendation [3], pricing [1] or search algorithms [5]). These systems have demonstrated high level performances lately [10], so it comes with no surprise that having AI-based models to face users is now a common ground for the tech industry (called the platforms hereafter).

Yet, the massive use of AI-based model raises concerns, for instance regarding their potentially unfair and/or discriminatory decisions. It is then of societal interest to develop methods to audit the behavior of an online model, to verify its lack of bias [12], proper use of user data [11], or compliance to laws [7].

The growing list of known audit methods is slowly consolidating into the emerging field of algorithmic audit of AI-based decision making algorithms, and multiple directions are yet to be explored for expanding that nascent field.

Mission confiée

While audits are by essence punctual, the audited models often continuously evolve, for instance because of reinforcement learning, retraining on new user inputs, or simply because of code updates pushed by the platform operators. An unexpected direction of interest, that might be crucial for instance to regulators, is to provide means to observe the mutation of an online model.

Assume a platform model under scrutiny, and an auditor that has only access to that model solely by means of queries/responses. This is coined as a black-box access to a model in the literature. Through these basic actions, an open research question is the proper definition of what is a stable model, i.e., a model that is consistent in time with regards to its decisions, (and consequently does not mutate).

While there has been a couple of approaches to define techniques of tampering-detection of a model [6, 4], this definition is bound to classifiers and to the sole capability of checking if the model is the same or if it is different.

A more refined way would be to provide a quantification for mutation, that is a notion of a distance between two instances A and B of a model, possibly owned locally by an auditor, with a variant B of a model that has already mutated. How to define and design a practical and robust distance measure is the topic of this Ph.D. thesis.

This opens up multiple questions:
- How should such a setup be modeled (statistical modeling, use of information theory, similarities from the GATz Mining field, etc.), so that we are able to provide a well defined measure for that problem.
- Moreover, while standard approaches exist to evaluate the divergence between two models, those need to be adapted to the context. In particular, we seek practical approaches that estimate divergence using few requests.
- An example of a modeling can rely on graphs. One can indeed structure the data collected from the observed model under relations forming a graph (see e.g., [8] in the context of the YouTube recommender), and compare that graph to the structure of a desirable graph while considering the properties that are awaited from the platform.
- Such AI models are nowadays used in a large variety of tasks (such as classification, recommendation or search). How does the nature of the tasks influences the deviation estimation/detection ?
- Considering that the auditor tracks deviation tracking, with regards to a reference point, is it possible to identify the direction in the mutation? That is particularly interesting in order to assess if a model evolves towards compliance with law requirements.
- Taking the opposite (platform) side: are there ways to make this distance measurements impossible, or at least noisy, so that is it impossible for the auditor to issue valuable observations? (we will relate this to impossibility proofs). In other words, can we model adversarial platform behaviours that translate into increased auditing difficulty ?

Principal activitaires

Work Plan
- A state of the art will review past approaches to observe algorithms in a black-box. This relates to the fields of security (reverse engineering), machine learning (with e.g., adversarial examples), and computability [9].
- We plan to approach the problem by leveraging a large AI model made public (e.g., https://pytorch.org/torchrec/), and mutate it by fine-tuning for instance, so that we can get intuition about the problem, as well as testing the first distances we have indentified.
- Provide a first consistent benchmark from these various distances. In particular, an important aspect will be their precision depending on the query budget necessary to obtain them (precision/cost of exploration)
Once the optimum distance for our problem has been found, the followup work will be devoted to prevent its construction by designing countermeasures on the platform side. In short, design an adversary capable to create important noise in the measurement by the auditor. This can relate for instance to the notion of randomized smoothing in the domain of classifiers [2].

This cat-and-mouse game between the auditor and the platform will structure and help create the impossibility proofs we are seeking to propose, in order to provide algorithmic landmarks for scientists and regulators.

References

Compétences
- Advanced skills in machine learning (classification, regression, adversarial examples)
- A strong formal and theoretical background. Interest in the design of algorithms is a plus.
- Good scripting skills (e.g., Python) and/or familiar with statistical analysis tools (e.g., R)

Avantages
- Subsidized meals
- Partial reimbursement of public transport costs
- Possibility of teleworking (90 days per year) and flexible organization of working hours
- Partial payment of insurance costs

Rémunération
monthly gross salary amounting to 1982 euros for the first and second years and 2085 euros for the third year