Offre n°2022-05165

PhD Position F/M Auditing the mutations of online AI-models

Le descriptif de l'offre ci-dessous est en Anglais

Type de contrat : CDD

Niveau de diplôme exigé : Bac + 5 ou équivalent

Fonction : Doctorant

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 institute, 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 AIbased 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 a 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 unexplored 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 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 datamining 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 it is 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 ?

Principales activités

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 indentifed.
· 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 tradeoff in the requests to the black-box)
· 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
[1] Le Chen, Alan Mislove, and Christo Wilson. Peeking beneath the hood of uber. In Proceedings of the 2015 internet measurement conference, pages 495–508, 2015.
[2] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In International Conference on Machine Learning, pages 1310–1320. PMLR, 2019.
[3] Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In Proceedings of the 10th ACM conference on recommender systems, pages 191–198, 2016.
[4] Zecheng He, Tianwei Zhang, and Ruby Lee. Sensitive-sample fingerprinting of deep neural networks. In
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4729–4737, 2019.
[5] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P Gummadi, and Karrie Karahalios. Quantifying search bias: Investigating sources of bias for political searches in social media. In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, pages 417–432, 2017.
[6] Erwan Le Merrer and Tre□dan Gilles. Tampernn: efficient tampering detection of deployed neural nets. In
2019 IEEE 30th International Symposium on Software Reliability Engineering (ISSRE), pages 424–434. IEEE, 2019.
[7] Erwan Le Merrer, Ronan Pons, and Gilles Tre□dan. Algorithmic audits of algorithms, and the law. working paper or preprint, February 2022.
[8] Erwan Le Merrer and Gilles Tre□dan. The topological face of recommendation. In International Conference on Complex Networks and their Applications, pages 897–908. Springer, 2017.
[9] Edward F Moore et al. Gedanken-experiments on sequential machines. Automata studies, 34:129–153, 1956.
[10] Maxim Naumov, Dheevatsa Mudigere, Hao-Jun Michael Shi, Jianyu Huang, Narayanan Sundaraman, Jongsoo Park, Xiaodong Wang, Udit Gupta, Carole-Jean Wu, Alisson G. Azzolini, Dmytro Dzhulgakov, Andrey Mallevich, Ilia Cherniavskii, Yinghai Lu, Raghuraman Krishnamoorthi, Ansha Yu, Volodymyr Kondratenko, Stephanie Pereira, Xianjie Chen, Wenlin Chen, Vijay Rao, Bill Jia, Liang Xiong, and Misha Smelyanskiy. Deep learning recommendation model for personalization and recommendation systems. CoRR, abs/1906.00091, 2019.
[11] Bashir Rastegarpanah, Krishna Gummadi, and Mark Crovella. Auditing black-box prediction models for
data minimization compliance. Advances in Neural Information Processing Systems, 34:20621–20632, 2021.
[12] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. Aequitas: A bias and fairness audit toolkit. arXiv preprint arXiv:1811.05577, 2018.

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

Informations générales

- **Thème/Domaine :** Algorithmique, calcul formel et cryptologie
  Systèmes d'information (BAP E)
- **Ville :** Rennes, Paris
- **Centre Inria :** Centre Inria de l'Université de Rennes
- **Date de prise de fonction souhaitée :** 2022-10-01
- **Durée de contrat :** 3 ans, 1 mois
- **Date limite pour postuler :** 2022-08-25

Contacts

- **Équipe Inria :** WIDE
- **Directeur de thèse :**
  Le Merrer Erwan / erwan.le-merrer@inria.fr

A propos d'Inria

Inria est l'institut national de recherche dédié aux sciences et technologies du numérique. Il emploie 2600 personnes. Ses 215 équipes-projets agiles, en général communes avec des partenaires académiques, impliquent plus de 3900 scientifiques pour relever les défis du numérique, souvent à l'interface d'autres disciplines. L'institut fait appel à de nombreux talents dans plus d'une quarantaine de métiers différents. 900 personnels d'appui à la recherche et à l'innovation contribuent à faire émerger et grandir des projets scientifiques ou entrepreneuriaux qui impactent le monde. Inria travaille avec de nombreuses entreprises et a accompagné la création de plus de 200 start-up. L'institut s'efforce ainsi de répondre aux enjeux de la transformation numérique de la science, de la société et de l'économie.

Attention: Les candidatures doivent être déposées en ligne sur le site Inria. Le traitement des candidatures adressées par d'autres canaux n'est pas garanti.

Consignes pour postuler

Please submit online : your resume, cover letter and letters of recommendation eventually

For more information, please contact erwan.le-merrer@inria.fr

Sécurité défense :
Ce poste est susceptible d'être affecté dans une zone à régime restrictif (ZRR), telle que définie dans le décret n°2011-1425 relatif à la protection du potentiel scientifique et technique de la nation (PPST). L'autorisation d'accès à une zone est délivrée par le chef d'établissement, après avis ministériel favorable, tel que défini dans l'arrêté du 03 juillet 2012, relatif à la PPST. Un avis ministériel défavorable pour un poste affecté dans une ZRR aurait pour conséquence l'annulation du recrutement.

Politique de recrutement :
Dans le cadre de sa politique diversité, tous les postes Inria sont accessibles aux personnes en situation de handicap.