Offre n°2022-05266

PhD Position F/M Membership Inference Attack in Machine Learning

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

Within the framework of the Chaire "IA Défense SAIDA".

Mission confiée

Membership Inference

One of the wonders of machine learning is that it turns any kind of data into mathematical equations. Once you train a machine learning model on training examples—whether it's on images, audio, raw text, or tabular data—what you get is a set of numerical parameters. In most cases, the model no longer needs the training dataset and uses the tuned parameters to map new and unseen examples to categories or value predictions. You can then discard the training data and publish the model on GitHub or run it on your own servers without worrying about storing or distributing sensitive information contained in the training dataset.

Nevertheless, a type of privacy-leak oriented attack against ML systems, namely membership inference, makes it possible to detect whether a given data instance was used to train a machine learning model. In many cases, the attackers can stage membership inference attacks without having access to the machine learning model's parameters. They just query the model and observe its output (soft decision scores or hard predicted labels). Membership inference can cause severe security and privacy concerns in cases where the target model has been trained with sensitive information. For example, identifying that a certain patient's clinical record was used to train a automatic diagnosis model reveals that the patient's identity and relevant personal information. Moreover, such privacy risk might lead commercial companies who wish to leverage machine learning-as-a-service to violate privacy regula- tions. [VBE18] argues that membership inference attacks on machine learning models increase greatly the vulnerability of machine learning service providers on privacy leaks. They may face further legal issues related to privacy information breaching in their business practices due to GDPR (General Data Protection Regulation).

Thesis

In this thesis, our plan is to first implement and benchmark typical membership inference attacks proposed in the literature [LZ21, SDS+19, SSSS17, CCTCP21, CCN+22]. We need to carefully outline the impact of crucial parameters such as the hardness of the classification task (dimension of the inputs, number of classes), the size (depth, number of parameters), the training procedure (data augmentation), and the potential overfitting of the target model. This also includes the working assumptions about the attacker's knowledge on the training data and his computation power. Indeed, some attacks rely on unrealistic assumptions. Designing more tractable attacks is key in order to clearly define when membership attacks are a real threat in practice.

In differential privacy [ACG+16, NST+21], a common defense is to randomize the procedure by adding noise either on the inputs (the training data set), the training procedure of the model, or the outputs (the trained model's parameters). This idea witnesses several implementations in modern machine learning like randomness in label smoothing, data augmentation, or penalization. The study focuses on evaluating the multiple trade-off between the loss of classification performance, the preven- tion of overfitting, and the gain of robustness against membership inference attacks but also against adversarial attacks [SSM19].

Beyond inferring the membership of a given instance, we will also study the feasibility of attribute inference attack targeting to reversely estimate the attributes of training data, which is an extension to

membership inference.

Expectations

The candidate for this thesis is expected to have accomplished courses on Machine Learning and/or have experience of implementing Machine Learning algorithms using Python for practical data mining problems. Especially, expertise in using Pytorch will be required in the project. Theoretical developments are also expected based on statistics and theory of machine learning and approximation.

References

[ACG+ 16] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, pages 308– 318, 2016.

[CCN+ 22] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In 2022 IEEE Symposium on Security and Privacy (SP), pages 1519–1519. IEEE Computer Society, 2022.

[CCTCP21] Christopher A Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Paper- not. Label-only membership inference attacks. In International conference on machine learning, pages 1964–1974. PMLR, 2021.

[LZ21] Zheng Li and Yang Zhang. Membership leakage in label-only exposures, 2021.

[NST+ 21] Milad Nasr, Shuang Songi, Abhradeep Thakurta, Nicolas Papemoti, and Nicholas Carlin. Adversary instantiation: Lower bounds for differentially private machine learning. In 2021 IEEE Symposium on Security and Privacy (SP), pages 866–882. IEEE, 2021.

[SDS+ 19] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. In International Conference on Machine Learning, pages 5558–5567. PMLR, 2019.

[SSM19] Liwei Song, Reza Shokri, and Prateek Mittal. Privacy risks of securing machine learning models against adversarial examples. In Proceedings of the 2019 ACM SIGSAC Confer- ence on Computer and Communications Security, pages 241–257, 2019.

[SSSS17] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership in- ference attacks against machine learning models. In 2017 IEEE symposium on security and privacy (SP), pages 3–18. IEEE, 2017.

[VBE18] Michael Veale, Reuben Binns, and Lilian Edwards. Algorithms that remember: Model inversion attacks and data protection law. Social Science Research Network, 2018.

Compétences

Technical skills and level required : Machine Learning, Statistics, Information theory, Pytorch

Languages : English

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 2051 euros for the first and second years and 2158 euros for the third year

Informations générales

- **Thème/Domaine :** Optimisation, apprentissage et méthodes statistiques
  Statistiques (Big data) (BAP E)
- **Ville :** Rennes
- **Centre Inria :** Centre Inria de l'Université de Rennes
- **Date de prise de fonction souhaitée :** 2022-10-01
- **Durée de contrat :** 3 ans
- **Date limite pour postuler :** 2022-08-31

Contacts

- **Équipe Inria :** LINKMEDIA
- **Directeur de thèse :**
  Furon Teddy / teddy.furon@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.

L'essentiel pour réussir

The candidate for this thesis is expected to have accomplished courses on Machine Learning and/or have experience of implementing Machine Learning algorithms using Python for practical data mining problems. Especially, expertise in using Pytorch will be required in the project. Theoretical developments are also expected based on statistics and theory of machine learning and approximation.

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

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.