Titre

Stage M2 : développement d'algorithmes pour sécuriser l'apprentissage fédéré face aux attaques antagonistes

Mots clés

Machine Learning ; Apprentissage fédéré ; IA de confiance.

Présentation du service : CEA LIST / SID

Situé à Saclay, en Ile-de-France sud, le CEA LIST (http://www-list.cea.fr/) est un centre de recherche scientifique et technologique dédié au développement de logiciels, de systèmes embarqués et de capteurs pour des applications destinées à la défense, la sécurité, l'énergie, le nucléaire, l'environnement et la santé. Le CEA LIST fait partie de l'écosystème dynamique et stimulant de l'Université Paris Saclay - le plus grand pôle scientifique français comptant 60 000 étudiants. Il compte plus de 700 chercheurs se focalisant sur les systèmes numériques intelligents, centrés autour de l'intelligence artificielle, l'usine du futur, l'instrumentation innovante, les systèmes cyberphysiques et la santé numérique. Au sein de cet institut, le SID (Service d'Intelligence des Données) travaille sur les algorithmes et méthodologies de l'intelligence artificielle et du traitement du signal. Les recherches et avancées technologiques du service sont guidées par des applications variées, pour lesquelles les spécificités et contraintes sur les données ou l'environnement d'exécution nécessitent une conception fine des IA et de leur intégration comme briques unitaires de systèmes complexes.

Description du projet

En 2016, Google publie les principes fondateurs de l'apprentissage fédéré [1] avec la promesse de créer des IA sans compromettre les données des utilisateurs. Cette méthode est en train de changer le paradigme actuel de l'IA centralisée, où construire de meilleurs modèles se résume souvent à collecter toujours plus de données personnelles et les centraliser sur un serveur. L'apprentissage fédéré est une approche collaborative où tous les utilisateurs d'un service participent à l'apprentissage du modèle sans transmettre leurs données personnelles mais uniquement les paramètres du modèle mis à jour localement (voir Fig.1). Au lieu de centraliser les données, seuls les paramètres du modèle sont agrégés sur le serveur central ce qui permet d'améliorer la confidentialité des données et de limiter les coûts de communication.

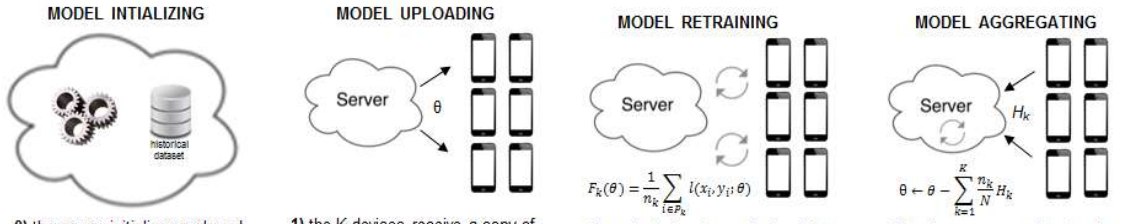

Fig-1 : illustration de l'apprentissage fédéré

Par construction, l'apprentissage fédéré assure la confidentialité des données client. Par contre, le processus d'apprentissage s'appuie sur des terminaux utilisateurs dont la fiabilité n'est pas éprouvée. Il est donc impératif de mettre en place des approches de défense face aux attaques antagonistes qui auraient pour objectif de détériorer les performances du modèle entrainé. Ce type d'attaques réalisées pour un ou plusieurs participants au processus d'apprentissage collaboratif a pour but de détériorer les performances globales

du modèle (untargeted attacks) ou plus spécifiquement la reconnaissance d'une classe particulière (targeted attacks).

L'objectif du stage va consister à identifier les différentes attaques qui pourraient intervenir lors de la phase d'apprentissage (data or model update poisoning) du modèle fédéré [2,3,4] puis proposer, développer et évaluer des solutions de défense pour contrer ces attaques.

Les cas d'applications envisagées concernent des données issues d'applications nucléaires, de vision par ordinateur ou de consommations électriques.

[1] Google AI blog: https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

[2] A.N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo. Analyzing Federated Learning through an Adversarial Lens, ArXiv, Nov. 2019.

[3] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov. How to Backdoor Federated Learning, ArXiv, Aug. 2019.

[4] H. Wang, K. Sreenivasan, S. Rajput, H. Vishwakarma, S. Agarwal, J.-Y. Sohn, K. Lee, and D. Papailiopoulos. Attack of the Tails: Yes, You Really Can Backdoor Federated Learning, ArXiv, Jul. 2020.

Profil recherché

Le stage s'adresse à un(e) étudiant(e) du cycle ingénieur/universitaire cherchant un stage M2 et manifestant l'envie de travailler dans le milieu de la recherche.

Idéalement, le candidat suit actuellement une formation en lien avec le domaine de l'Intelligence Artificielle/Machine Learning. La maîtrise de Python est indispensable.

Le/la candidat(e) devra être capable d'apporter ses idées novatrices, son enthousiasme, sa rigueur et devra faire preuve d'un esprit d'équipe prononcé.

La durée du stage est de 6 mois minimum. Le stage est rémunéré.

Informations administratives et contacts

Nature du contrat de travail : stage

Type du contrat de travail : Droit privé

Délai administratif pour début de contrat : environ 3 mois

Envoyer CV et lettre de motivation à :

Aurélien Mayoue (aurelien.mayoue@cea.fr)

Cédric Gouy-Pailler (cedric.gouy-pailler@cea.fr)