LABORATOIRE D'INFORMATIQUE & DES SYSTÈMES

Sujet de stage - M2
Année 2024—2025

# Analyse de la robustesse de l'apprentissage fédéré

**Lieu :** Laboratoire d'Informatique et des Systèmes, Marseille
**Encadrants :** François-Xavier Dupé (francois-xavier.dupe@lis-lab.fr),
Arnaud Labourel (arnaud.labourel@lis-lab.fr),
Ronan Sicre (ronan.sicre@lis-lab.fr).

**Durée du stage :** entre 4 et 6 mois

## Contexte

Ce stage s'intéresse à la robustesse des modèles de *machine learning* dans le cadre de l'apprentissage fédéré. L'apprentissage fédéré vise à répartir l'apprentissage sur un grand nombre d'agents en évitant ainsi la centralisation des données et modèles. Chaque agent reçoit une version du modèle qu'il pourra mettre à jour en fonction des données qu'il aura collectées. Régulièrement les agents envoient des données pour mettre à jour le modèle global, qui pourra ensuite être repartagé aux agents.

L'apprentissage fédéré a plusieurs avantages. Par exemple, il ne nécessite pas l'envoi de données, qui pourraient être sensibles. L'envoi des données peut aussi être très coûteux et on se contentera d'envoyer les poids du modèle.

Ce modèle d'apprentissage est assez générique et peut être utilisé pour diverses méthodes d'apprentissages. Toutefois, il est sensible à différents types d'attaques pouvant rendre le modèle inutilisable ou biaisé. Par exemple, il existe des attaques où un site malveillant transmet des mauvais poids tout en conservant les bons.

Ce stage s'inscrit dans le cadre d'un projet national, le PEPR TrustInCloud, qui regroupe plusieurs équipes de différents laboratoires à travers la France. Il se déroulera au sein du Laboratoire d'Informatique et des Systèmes (LIS) avec une collaboration entre deux équipes de recherche. L'équipe Qarma travaille sur l'apprentissage automatique et l'équipe DALGO pour la partie algorithmique distribuée. Le travail fourni pourra faire l'objet de présentation au sein du projet.

## Objectifs du stage

Le stage se découpe en deux objectifs autour de deux types de problèmes de sécurités rencontrés par ces méthodes : les attaques de type *adversarial* et les fuites d'information privées.

Le premier objectif du stage est d'identifier la surface d'attaque de l'apprentissage fédéré appliqué à des réseaux de neurones [1, 4] et de comment ces attaques peuvent être contrées ou amenuisées. Parmi les attaques, nous nous focaliserons ensuite sur les attaques de type *adversarial* (générales et ciblées) [2]. Une étude sur l'état de l'art des attaques et des mitigations est à faire, avec des preuves de concepts pour des réseaux de neurones classiques. Les preuves de concepts serviront ensuite d'études des différentes mitigations et techniques pour renforcer les modèles. Dans le cadre du stage, nous nous concentrerons sur un réseau convolutif de type ResNet pour une tâche de classification d'images.

Ce premier objectif demandera de prendre en main la plateforme *Flower.AI* proposée pour travailler avec PyTorch. Cette prise en main sera l'occasion d'implémenter une méthode classique d'apprentissage fédéré pour ensuite tester des attaques. Selon les résultats, nous irons vers des méthodes plus récentes avec des attaques et mitigations (ou autre protection) plus complexes.

Le deuxième objectif prend le problème dans l'autre sens en regardant les fuites de données possibles. Cette problématique entre dans le cadre de la *differential privacy* (DP) où on regarde à quel point une méthode d'apprentissage est sensible à un élément donné dans son ensemble d'entraînement. Nous allons ici nous focaliser sur cette problématique dans le cadre de l'apprentissage fédéré [5]. Enfin nous regarderons comment étendre le cadre aux réseaux de confiance [3]. Ces réseaux permettent de modéliser des acteurs qui peuvent échanger des données confidentielles de façon légitimes et légales.

Ce deuxième objectif repose sur les outils mis en place lors du premier objectif. L'analyse de l'efficacité de la *differential privacy* se fera aussi à travers les attaques et mitigations proposées auparavant. De même les solutions examinées pour renforcer la DP seront aussi testées dans le cadre des attaques adversariales.

## Informations supplémentaires

Ce stage se déroulera donc en plusieurs étapes :

1. Prise en main du cadre de développement ;
2. État de l'art sur les attaques *adversarial* dans un contexte d'apprentissage fédéré ;
3. Implémentation d'une preuve de concept avec un ResNet pour de la classification d'images ;
4. Étude de plusieurs techniques de mitigations ;
5. État de l'art sur la *differential privacy* ;
6. Implémentation des techniques de mesure et de robustesse (en conservant le cadre précédent).

Le langage de programmation utilisé sera Python avec le module Flower (`https://flower.ai/`) de PyTorch.

## Cadre du stage et pré-requis

Pour son bon déroulement, nous requérons les compétences suivantes,
— un bon niveau de programmation en Python ;
— des connaissances en apprentissage automatique (dont apprentissage profond) ;
— une bonne compréhension de l'apprentissage statistique ;
— un goût pour les mathématiques discrètes et les graphes.

Ce stage sera rémunéré. Il se déroulera à Marseille et le stagiaire sera accueilli dans l'équipe Qarma dont les locaux sont situés sur le technopôle de Château-Gombert.

## Références


[1] Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal, and Seraphin Calo. Analyzing federated learning through an adversarial lens. In *International conference on machine learning*, pages 634–643. PMLR, 2019.

[2] Shuhao Fu, Chulin Xie, Bo Li, and Qifeng Chen. Attack-resistant federated learning with residual-based reweighting. *arXiv preprint arXiv :1912.11464*, 2019.

[3] Badih Ghazi, Ravi Kumar, Pasin Manurangsi, and Serena Wang. Differential privacy on trust graphs. *arXiv preprint arXiv :2410.12045*, 2024.

[4] Youpeng Li, Xinda Wang, Fuxun Yu, Lichao Sun, Wenbin Zhang, and Xuyu Wang. Fedcap : Robust federated learning via customized aggregation and personalization. *arXiv preprint arXiv :2410.13083*, 2024.

[5] Lingjuan Lyu, Han Yu, Xingjun Ma, Chen Chen, Lichao Sun, Jun Zhao, Qiang Yang, and S Yu Philip. Privacy and robustness in federated learning : Attacks and defenses. *IEEE transactions on neural networks and learning systems*, 2022.