AVIGNON UNIVERSITÉ

PROPOSITION SUJETS DE THESES

CONTRATS DOCTORAUX 2020-2023

Appel ciblé (merci de cocher la case correspondante):

X Contrat doctoral ministériel ED 536

□ **Contrat doctoral ministériel ED 537**

□ **Contrat doctoral EUR Implanteus**

Directeur de thèse : Rachid Elazouzi

Co-directeur éventuel : Mohamed Morchid

Co-encadrant éventuel :

Titre en français : Conception d'incitations pour un apprentissage fédéré efficace dans un système distribué

Titre en anglais : Incentive Design for Efficient Federated Learning in distributed system : a Game Theoretical Approach

Mots-clés : Federated learning, Game theory, Deep neural networks, Optimisation and stochastic games.

Co tutelle : Non **Pays :**

Opportunités de mobilité à l'international du doctorant dans le cadre de sa thèse : oui (avec l'université de Oulu, Finland)

Profil du candidat :

- Master in computer science or similar discipline
- Background in machine learning and optimization
- Ability to work in interdisciplinary teams and good communication skills in English

- Very good experience in python (Pytorch) or matlab
- NLP knowledge will be appreciated

Context of the PhD

There is an increasing interest in a new machine learning called Federated Learning (FL) [1]. This technique allows nodes to participate in the training model where each node contributes to the learning process by independently computing the gradient based on its local data. This process is repeated until reaching an accuracy level of the learning model. This paradigm is strongly supported by the machine learning community since it allows to protect the user data privacy because local training data are not shared. The growing demand for federated learning technology has produced a number of tools and frameworks such as Leafn [2], PySyft [3] and PaddleFL [4] while opening new challenges to tackle. One of the most important for the future of AI is to keep it sustainable. *The high-level goal of this thesis is to lead federated learning toward the road of sustainability: to evaluate and optimize the performance of this emerging new machine learning.*

Objectives of the PhD

Adapting techniques for handling **Non Informally, Identically Distributed** (Non-IID) datasets to federated learning remains an important open problem. For Non-IID data, performing local updates and communicating less frequently with the central server is the core challenge to reduce the traffic generated in the cloud in particular for non-IID data cases [5]. Given the heterogeneity of datasets, a natural theoretical analysis is to identify under what conditions the shared global model is better than independent models running at local devices. This of course depends on how the data is distributed among them. Thus a natural question of each node is how much improvement can be expected via federated learning? Or maybe it is more beneficial to join a subset of nodes to build a federated learning system. Here we will be addressing questions such as: Under which conditions would such a bargaining process result in any agreement, (i.e., can the nodes agree on anything, given their selfish nature?)? Would the outcome of such an agreement be beneficial from a system-wide/privacy perspective? What is the role of the central server (coordinator) at each coalition that guarantees the model accuracy? These give rise to novel mathematical approaches that we intend to tackle, including markov decision process, partially observable deep reinforcement learning, regret minimization and stackelberg equilibrium [6].

Many challenges will be handled in this thesis to put federated learning into practical use. The first challenge is to provide a general framework which subject to the following challenges :

- Non-IID data : The distribution of data at each node is not representative of the global data of distributions. This can reduce the accuracy of the federated learning, which is attributed to the weight divergence [ ]. As a solution, we investigate the process of bargaining and coalition formation among nodes by creating a small subset of data which is globally shared between nodes in a coalition. This procedure assists in

addressing the issue of trade off between the fairness and accuracy in federated learning.

- Fairness : Fairness receives increasing attention in machine learning and in particular for federated learning. The main goal here is to achieve fairness in federated learning when testing data distribution is different from training distribution or even unknown. This allows to build a learning that is robust against the possible unkwing testing distribution.
- Limited communication : One challenge of federated learning is to reduce the user-controller communication since the end end-users typically have very limited communication bandwidth. Based on this limitation, we aim to design an asynchronous model update strategy by introducing the previously trained local models.
- Sharing knowledge: During the learning process of federated nodes based on hidden spaces from the concentrator, the gradients are evaluated independently for each node. This allows the whole system to preserve a balance between nodes. Nonetheless, the node's hidden features in the case of connexionnists models have to share latent information to better represent their own knowledge (e.g. their datasets). Therefore, the thesis will study the latent representation of the nodes in regard to the global learning process to propose original learning algorithms of both the federated system as well as for each individual node.
- Effectiveness: Federated learning approaches based on connexionnist models employ feed-forward neural networks that hardly manage sequential data such as time-series. Nevertheless, most of the data processed in nowadays applications are based on sequential information such as videos, music, or even textual documents. Therefore, the client represented by the node is strongly inclined to share sequential information with the federated coalision. [7] has proposed a recurrent neural based model FedSL for nodes in a federated learning process to handle sequential data (time series). The proposed architecture considers that a sequence can be split and, therefore, two nodes (clients) have to share their knowledge (gradients) during the learning process. Sharing information is a weakness of this approach and the thesis will propose learning based algorithms to preserve sharing information between clients.
- Expressiveness: Federated learning is useful in different Natural Language Processing (NLP) related tasks such as Spoken Language Understanding (SLU)[8]. The thesis will also propose novel neural based architectures of nodes and data sharing strategies to learn efficient SLU systems in an end-to-end fashion. The proposed nodes and coalision strategy will allow the neural based system to extract robust spoken documents features for SLU.

Références bibliographiques :

[1] Vivienne Sze and Yu-Hsin Chen and Tien-Ju Yang and Joel Emer. *Efficient Processing of Deep Neural Networks: A Tutorial and Survey.* Proceedings of the IEEE, Vol 105, No 12, Dec 2017.
[2] The Leaf Authors. Leaf, 2019. URL https://leaf.cmu.edu/.
[3] Theo Ryffel, Andrew Trask, Morten Dahl, Bobby Wagner, Jason Mancuso, Daniel Rueckert, and Jonathan
Passerat-Palmbach. A generic framework for privacy preserving deep learning, 2018
[4] The PaddleFL Authors. PaddleFL, 2019. URL https://github.com/PaddlePaddle/PaddleFL.

[5] B. Mirzasoleiman, Big Data Summarization Using Submodular Functions, Doctoral Thesis, ETH Zurich, 2017.
[6] Mikael Touati, Rachid El-Azouzi, Marceau Coupechoux, Eitan Altman, Jean-Marc Kelif, “A controlled matching game for WLANs”, IEEE Journal on Selected Areas in Communications, 2017.
[7] FedSL: Federated Split Learning on Distributed Sequential Data in Recurrent Neural Networks, 2020, arxiv
[8]Federated Learning for Spoken Language Understanding, 2020, Coling

Les sujets devront être adressés à

gestion-ed@univ-avignon.fr

avant le 6 avril 2020