Distributed and Federated Learning Optimization with Federated Clustering of IID-users

Lucas Pacheco, Eric Samikwa, Torsten Braun

Communications and Distributed System Group
Institute of Computer Science – University of Bern

Topics: DS algorithms with a view towards Machine Learning and Artifical Intelligence

{lucas.pacheco, eric.samikwa, torsten.braun}@inf.unibe.ch

Abstract:

Federated Learning (FL) is one of the leading learning paradigms for enabling a more significant presence of intelligent applications in networked and Internet of Things (IoT) systems. It consists of individual user devices performing machine learning (ML) models training locally, so that only trained models due to privacy concerns, but not raw data, is transferred through the network for aggregation at the edge or cloud data centers [Li et al. 2019]. Due to the pervasive presence of connected devices such as smart phones and IoT devices in peoples lives, there is a growing concern about how we can preserve and secure users' information. FL reduces the risk of exposing user information to attackers during transmission over networks or information leakages at the central data centers. Another advantage of FL is scalability and maintainability of intelligent applications in networked and IoT systems. Considering highly distributed environments in which such systems are deployed, collecting and transmitting raw user data for training of ML models at central data centers is a challenging task as it imposes huge workload on the networks and consumes high bandwidth. Training of ML models is distributed over locations and transmitting the trained models for aggregation alleviates these challenges.

Among others, distributed and federated learning have applications in smart healthcare systems, where very sensitive user data is involved, and industrial IoT applications, where the amount of data for training may be too large and cumbersome to transport to central data centers. However, FL has the significant shortcoming of requiring user data to be Independent Identically Distributed (IID) (*i.e.*, users which have similar data statistical distributions and are not mutually dependent) and make reliable predictions for a given group of users aggregated into a single model. IID users have similar statistical features, and thus can be aggregated into the same ML models. Since raw data is not available at the model aggregator, it is necessary to find IID users based solely on their trained machine learning models.

We present a Neural Network-based Federated Clustering mechanism capable of clustering IID with no access to their raw data called Neural-network SIMilarity estimator, NSIM. Such mechanism performs significantly better than competing techniques for neural-network clustering [Pacheco et al. 2021]. We also present an alternative to the FedAvg aggregation algorithm used in traditional FL, which significantly increases the aggregated models' reliability in terms of Mean Square Error by creating several training models over IID users in a real-world mobility prediction dataset. We observe improvements of up to 97.52% in terms of Pearson correlation between the similarity estimation by NSIM and ground truth based on the LCSS (Longest Common Sub-Sequence) similarity metric, in comparison with other state-of-the-art approaches. Federated Clustering of IID data in different geographical locations can improve performance of early warning applications such as flood prediction [Samikwa et al. 2020], where the data for some locations may have more statistical similarities. We further present a technique for accelerating ML inference in resource-constrained devices through distributed computation of ML models over IoT networks, while preserving privacy. This has the potential to improve the performance of time sensitive ML applications.

References

Li, X., Huang, K., Yang, W., Wang, S., and Zhang, Z. (2019). On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*.

Pacheco, L., Rosário, D., Cerqueira, E., and Braun, T. (2021). Federated user clustering for non-iid federated learning. International Conference on Networked Systems 2021 (NetSys 2021).

Samikwa, E., Voigt, T., and Eriksson, J. (2020). Flood prediction using iot and artificial neural networks with edge computing. In *2020 International Conferences on Internet of Things (iThings)*, pages 234–240. IEEE.