ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2.
Available from http://www.i6doc.com/en/.

Federated Learning in a Semi-Supervised Environment for Earth Observation Data

Bruno Casella[1,†], Alessio Barbaro Chisari[2,†], Marco Aldinucci[1],
Sebastiano Battiato[2], and Mario Valerio Giuffrida[3] *

1 - Alpha Research Group, Computer Science Department
University of Turin, C.so Svizzera 185, Turin - Italy

2 - Image Processing Laboratory, Department of Mathematics and Computer Science
University of Catania - Viale Andrea Doria 6, Catania - Italy

3- School of Computer Science - University of Nottingham
Wollaton Road, Nottingham NG8 1BB - UK

† Both authors contributed equally to this work.

Abstract. We propose FedRec, a federated learning workflow taking advantage of unlabelled data in a semi-supervised environment to assist in the training of a supervised aggregated model. In our proposed method, an encoder architecture extracting features from unlabelled data is aggregated with the feature extractor of a classification model via weight averaging. The fully connected layers of the supervised models are also averaged in a federated fashion. We show the effectiveness of our approach by comparing it with the state-of-the-art federated algorithm, an isolated and a centralised baseline, on novel cloud detection datasets. Our code is available at `https://github.com/CasellaJr/FedRec`.

1 Introduction

In recent years, the massive adoption of data-driven technologies has required effective artificial intelligence methods addressing the increasing privacy requirements, such as the European GDPR regulation. Federated Learning (FL)[1] has emerged as a promising approach for dealing with private and sensitive data to train machine learning models. In a typical federated scenario, there are two entities: a server and many different clients. By aggregating locally trained deep learning (DL) models sent by the clients, the server produces a global model without sharing locally-stored data in each of the clients. The current state-of-the-art FL algorithm is FedAvg [1], which aggregates the locally trained models by averaging their parameters. Privacy preservation is achieved by keeping data locally. Built on the assumption that clients hold labelled data, most of the FL literature focuses on supervised learning problems. However, in real-world scenarios, as the parties involved may not have sufficient domain expertise or resources, the data may also be unlabelled. Therefore, the absence of annotated labels currently represents a challenge in FL.

[*] This work has been partly supported by the Spoke "FutureHPC & BigData" of the ICSC - Centro Nazionale di Ricerca in "High Performance Computing, Big Data and Quantum Computing", funded by European Union - NextGenerationEU, partially by the Horizon2020 RIA EPI project (G.A. 826647) and partially by Italy PON project E69J21011470006. All the data were collected thanks to EHT S.C.p.A. and ALICEnet cooperative network.

In this paper, we propose FedRec, an FL pipeline for image classification tasks, in which clients without ground-truth labels assist client training on annotated data. In our scenario, some clients are trained on labelled data, while others are trained on unlabelled data in an unsupervised fashion. We leverage an encoder-decoder model performing image reconstruction in those clients where labels are missing or lacking, in which the encoder architecture matches the feature extractor of those clients trained in a supervised fashion. This results in a mechanism of data augmentation for the labelled data, as the aggregation involves more parties contributing to the extraction of image features. At this purpose, we conducted experiments on five datasets collected using appropriate instruments, the ceilometers, across Italy. All the datasets contain images reconstructed from measurements taken by ceilometers, which, by counting the number of photons reflected by particles in the atmosphere, are able to estimate the height and presence of clouds in the sky.

Our contribution can be summarised as follows: (1) We propose FedRec, a federated semi-supervised learning (FSSL) approach in which unlabelled data are used in conjunction with labelled data to capture features serving as data augmentation for the latent space encoded from fully labelled data. We aim to improve the performance of the supervised model by using additional data from different locations in Italy. (2) We compare with the simplest approach based on FedAvg and labelled data and show its limitations. (3) We demonstrate the efficacy of FedRec through extensive experiments on novel datasets specially collected to support researchers in the field of deep learning applied to environmental phenomena, showing that we outperform the traditional method.

2 Related Work and Methodology

The primary goal of FL is to train a global inference model while keeping data scattered across different silos, thus preserving privacy. While most of the FL literature focuses on supervised tasks, some recent works adopt SSL techniques for exploiting the increasing volume of unlabelled data. In SemiFL [2], clients have completely unlabelled data, while the server holds a small amount of annotated samples. SemiFL proposes alternate training to fine-tune the global model with labelled data and generate pseudo-labels with the global model.

FedMatch [3] introduces an inter-client consistency loss that aims to maximise the agreement between the models trained at different clients. In particular, in FedMatch, each client samples the top-k nearest clients and ensures consistency by regularising the local model output with the top-k client models. Additionally, FedMatch adopts the decomposition of model parameters for disjoint learning on labelled and unlabeled data.

Another work addressing the increasing communication and computational cost due to the inter-client knowledge sharing based on model weights is ProtoFSSL [4], an approach based on prototype learning that exchanges lightweight prototypes between clients. Each federation client creates pseudo-labels based on shared prototypes to compute the loss. A prototype-based inter-client knowledge sharing significantly reduces both communication and computation costs.

In contrast to previous studies based on disjoint learning or prototype learning, we propose a method for FSSL based on parameter exchange leveraging both labelled and unlabelled data. In our approach, some clients hold only unlabelled data, while others hold ground-truth labels. Depending on the availability of annotated data, each client will solve a different task. Some clients will undertake a classification task, while others will tackle an unsupervised learning problem. Specifically, in FedRec, unlabelled data are utilised to solve an image reconstruction task with an encoder-decoder architecture. Our method involves aggregating the weights of model architectural components that perfectly match between supervised and unsupervised clients. A recent work [5] on vertical FL shows that aggregating only identical architectural parts of different models is a promising approach. Since our goal is to augment the supervised task with unlabelled data from other clients, we enforced the encoder architecture to align with the feature extractor of the image classification network. Although the different tasks, the encoder should extract image features serving as a data augmentation technique for the supervised clients, thus resulting in an increased FL generalization property. As a result, the weights of the encoder and the feature extractors of the classification models are averaged. Finally, we perform parameter aggregation of the fully connected layers of the classification models.

3 Experiments

Our experiments aim to investigate the classification performance of our proposed federated model trained by aggregating features extracted from both labelled and unlabelled data. We report the results of a typical isolated scenario in which a model is trained on data from a single institution, a centralised experiment in which the data are gathered in a single data lake, a naive federated approach based on FedAvg and only labelled data, and our proposed method.
Testbed setup: Our experiments were conducted in a simulated federation utilising an Intel® Xeon® processor (Skylake, IBRS, eight sockets of one core) and one Tesla T4 GPU. The federation comprised one server, two clients holding labelled data, and three clients holding unlabelled data. The baseline experiments, including the isolate, centralised, and FedAvg approach on the two clients with labelled data, were run on the same dedicated machine. For reproducibility purposes, the code and the data used for our experiments are available at the following link: `https://github.com/CasellaJr/FedRec`.
Datasets: All the datasets were obtained through measurement reconstructions of appropriate instruments, such as the ceilometers. A ceilometer is a measuring device mostly used in meteorology that can detect the height of a cloud base by emitting a modulated light beam directed to the sky. This makes it possible not only to recognise the clouds in the sky but also to determine their height. For this purpose [1] we used ALICEnet[2], a cooperative network of lidar-ceilometers coor-

[1] thanks to the company EHT S.C.p.A.

[2] `https://www.alice-net.eu/`

Location	Latitude	Longitude	Samples	Positive (%)	Period
S.G.L.P.	37d 34’ 44” N	15d 06’ 11” E	1568	1050 (66.96%)	01/01/23 - 14/03/23
C.G.	37d 34’ 16” N	12d 39’ 35” E	2193	890 (40.58%)	01/06/23 - 31/08/23
Roma	41d 50’ 32” N	12d 38’ 50” E	2208	N.A.	01/07/23 - 30/09/23
Taranto	40d 29’ 37” N	7d 13’ 01” E	2160	N.A.	01/01/21 - 31/03/21
Aosta	45d 44’ 32” N	7d 21’ 24” E	2180	N.A.	01/07/21 - 30/09/23

Table 1: Statistics of the datasets.

dinated by CNR-ISAC and operated in collaboration with other Italian research institutions, universities, and environmental agencies. We used five locations around Italy: San Giovanni La Punta (S.G.L.P.), Capo Granitola (C.G), Roma, Taranto, and Aosta. Table 1 summarizes the statistics of the datasets. S.G.L.P. and C.G. datasets contain annotated data, while the latter are unlabeled. The first dataset, S.G.L.P., has previously been publicly released [6] and it was labelled using the output of a *Weather Research and Forecasting* (WRF) model specially set up to produce weather simulations at the coordinates of the corresponding ceilometer. The authors labelled the C.G. dataset manually. Both S.G.L.P and C.G. are used to solve the cloud detection binary classification task. In cloud detection, we have a positive label if a cloud is detected and a negative otherwise. S.G.L.P. and C.G. clients split their data into training (70%) and testing (30%) data. As we were interested only in improving the classification performance, the clients holding unlabeled data used the entire set as training data to increase the generalisability of the extracted features.

Models: We employed a ResNet-18 as a feature extractor on S.G.L.P. and C.G., trained by minimising the cross-entropy loss with mini-batch gradient descent using the SGD optimizer with a learning rate of 10^{-4}, momentum 0.8 and weight decay 10^{-5}. The local batch size was 8. We employed an encoder-decoder architecture to solve the image reconstruction task in Roma, Taranto, and Aosta. A ResNet-18 serves as the backbone of the encoder part for capturing image features and encoding them into the latent space. The decoder, mapping the encoded representation back to the original feature space, is made of four convolutional and upsampling layers. This architecture was trained by minimising the MSE with mini-batch gradient descent using the SGD optimizer with a learning rate of 10^{-4}, momentum 0.8 and weight decay 10^{-5}. The local batch size was 8. As evaluation metrics, we focus on the accuracy and F1-score.

Discussion: Table 2 shows the results of our experiments in terms of accuracy and F1-score. Looking at the accuracies, FedRec performs slightly better than FedAvg with only labelled data on the S.G.L.P dataset, while it outperforms the naive approach on the C.G. data. Federated F1-scores are comparable between the two datasets and methods. FedRec achieves better results than the baseline on the C.G. dataset, while it is beaten on the S.G.L.P. data, even if scores are really closer to each other. Isolated accuracies show that the classifier correctly discriminates the majority of samples. However, accuracy alone may not adequately capture the classifier’s performance, especially in the presence

Dataset		Accuracy			F1-score	
	Isolated	Federated		Isolated	Federated	
		Naive	FedRec		Naive	FedRec
S.G.L.P.	.742 ± .012	.689 ± .005	.691 ± .038	.668 ± .0	.808 ± .002	.805 ± .001
C.G.	.988 ± .001	.427 ± .052	.587 ± .093	.405 ± .0	.576 ± .0	.581 ± .001

Table 2: Comparison between our proposed method, centralised baselines and state-of-the-art methods. Results (mean) are obtained with five averaged runs.

of class imbalance, as it will tend to favour the majority class. F1-score, being the harmonic mean of precision and recall, is a better metric for evaluating performance on unbalanced datasets. Isolated F1-scores show a moderate balance between precision and recall. This means that while the classifier identifies some true positives, it may also produce a notable number of false positives, or false negatives, or both. Moreover, the similarity between the class imbalance percentage and the isolated F1-scores on both datasets suggests that the classifier's performance is comparable to randomly guessing the positive class. Both the naive federated and FedRec approaches outperform the isolated F1-scores. We hypothesise that this is because the federated model has a higher ability to generalise, due to a greater number of image features it was trained on. Results that were obtained in a centralised setting, in which both the annotated datasets were aggregated in a single data lake, show that both accuracy and F1-score are a weighted mean (with a bigger impact of C.G. due to the greater amount of samples) of the isolated performance. In particular, in the centralized scenario, we obtained an accuracy of 0.933 ± .001 and an F1-score of 0.517 ± .00. We hypothesize that, while counter-intuitive, FL benefits from its intrinsic alternate training nature, leading to a better feature extraction process.

We did not report the results of the image reconstruction task on the unlabelled clients, as we were only interested in the supervised performance. We used the unlabelled data as a data augmentation technique to extract a broader and more general set of image features and assist in the supervised training.

Finally, Table 3 reports the global model results after fine-tuning on the local dataset for one epoch. Although the model's accuracy benefits from a fine-tuning iteration on the local dataset, there is a drop in F1-scores. In particular, the performance goes down to the isolated case. We hypothesise that this occurs due to the small size of the datasets. Even though the aggregated model benefits from a more accurate feature extraction process, an epoch of fine-tuning on the local datasets leads to overfitting, probably due to the small dataset size.

4 Conclusion and Future Work

This paper proposes FedRec, a method for FSSL leveraging unlabeled data to help supervised training on annotated data. In particular, clients with solely unlabeled data use an encoder-decoder architecture for doing image reconstruction, in which the encoder part matches the feature extractor of a classification

Dataset	Accuracy		F1-score	
	FedRec	FedRec fine-tuning	FedRec	FedRec fine-tuning
S.G.L.P.	.691 ± .038	.848 ± .009	.808 ± .002	.668 ± .0
C. G.	.587 ± .093	.984 ± .011	.576 ± .0	.405 ± .0

Table 3: Global model results after fine-tuning on the local dataset for one epoch. Results (mean) are obtained with five averaged runs.

model. The trained feature extractors are aggregated via weight averaging, as well as the fully connected layers of the classification models. We show the effectiveness of our method by comparing its accuracy and F1-score performance against the isolated, centralised and federated baseline based on FedAvg of just the supervised models.

For future work, we aim to study if there is room to improve our method's computation costs. Indeed, if the communication cost is reduced by sharing only a smaller subset of layers of the local models, the computation time required increases due to a bigger volume of data and clients participating in the federation. A possible strategy to address this issue may be an early stopping technique on the unlabeled data. Finally, we aim to deepen the possibility to improve the learning performance by determining the best model for both the supervised and unsupervised tasks.

References

[1] Brendan McMahan et al. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA*, 2017.

[2] Enmao Diao, Jie Ding, and Vahid Tarokh. Semifl: Semi-supervised federated learning for unlabeled clients with alternate training. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans*, 2022.

[3] Wonyong Jeong et al. Federated semi-supervised learning with inter-client consistency & disjoint learning. In *9th International Conference on Learning Representations, ICLR 2021, Austria, May 3-7, 2021*, 2021.

[4] Woojung Kim et al. Federated semi-supervised learning with prototypical networks. *CoRR*, abs/2205.13921, 2022.

[5] Bruno Casella and Samuele Fonio. Architecture-based fedavg for vertical federated learning. In *Proceedings of the IEEE/ACM 16th International Conference on Utility and Cloud Computing*, New York, NY, USA, 2024.

[6] Alessio Barbaro Chisari et al. On the cloud detection from backscattered images generated from a lidar-based ceilometer: Current state and opportunities. In *IEEE ICIP 2024, Abu Dhabi, UAE, October 27-30, 2024*, 2024.