Federated Learning with Student Models and Attention Transfer across Heterogeneous Devices

Hongrui Shi and Valentin Radu

University of Sheffield

Federated Learning (FL) is considered as the present best solution for training a machine learning model on the private data of many devices. The wide adoption of smartphones where a vast amount of data is generated and stored has encourage the emergence of this privacy preserving learning method. With an increasing diversity of devices, the system heterogeneity challenge is becoming more relevant.

Unlike traditional machine learning where the model is trained in a centralized fashion, FL trains a global model on decentralized data. Under the coordination of a server, in each FL round a global model is copied to many devices (clients) for local training. Then, the locally updated client models are aggregated on the server side as global update. These clients participating in a FL round are expected to complete their local updates and transfer these to the server by a deadline. Recent work has shown that FL converges on the assumption that clients perform a similar amount of local updates. But with a growing spectrum of devices, some slower than smartphones, such as IoT devices, and others faster, such as home data boxes, the standard FL method of distributing the same model is starting to break down. In current FL settings, the slower clients either perform less local updates or are completed dropped from the round, risking the convergence of the global FL model.

The best solution we see fit for this problem is to adapt the amount of work each client performs based on their system computing capabilities. Since distributing a single global model cannot fit devices with different hardware size, we propose to replace the single global model with a set of smaller-size models that match the computing performance of each system. As such, the important question we need to answer is how we can best transfer the knowledge between the client side and the server side. We leverage the Attention Transfer (AT) technique with a student-teacher learning framework to address this question. By this method, the model receiving the attention during training is guided to align its activation maps at different levels with the model providing the attention. We keep a global model and a generic dataset on the server side to facilitate the training with AT. Additionally, we extract the activation maps from the client models to create metadata, which helps to transfer client side knowledge to the global model in non-IID scenarios (some classes are never seen by the global model directly).

We evaluate our method on both IID (Independent and Identically Distributed) and non-IID data scenarios. From experiments, we make the following observations:

- On IID data, training the global model with AT from the updated client models improves the test accuracy of the global model. But, in non-IID data, using AT alone is not sufficient for the global model to recognise unseen classes present in the local data. This determined us to integrate metadata (aggregated fractions of attention maps) in our learning process.
- Learning with Attention Transfer and metadata shows significant improvement in non-IID data scenario.
- We identify the best level in the network from where to extract attention maps as metadata. Our results show that lower level attention maps benefit the training better, with higher test accuracy.
- Our ablation study to understand the role of Attention Transfer and metadata, as well as the impact of fractions of metadata shows that a combination of AT, followed by as little as 10% of metadata, followed by an additional step of AT on the generic data achieves the best test accuracy.

Our experiments demonstrate that Attention Transfer is effective in transferring the knowledge between the client side and the server side. We also show a realistic FL setup with multiple and varied-size clients, validating the effectiveness of our method.