From Collaborative Learning to Collaborative Inference

Akash Dhasade
EPFL
Lausanne, Switzerland
akash.dhasade@epfl.ch

Anne-Marie Kermarrec
EPFL
Lausanne, Switzerland
anne-marie.kermarrec@epfl.ch

1 INTRODUCTION

Training machine learning models at the edge has become increasingly popular given the decentralized nature of data and the issues with data privacy. Referred to as "Collaborative Learning" (CL), such training at the edge takes the form of Federated Learning (FL) [3] or Decentralized Learning (DL) [2] where edge devices train on their local data and learn by exchanging models. The outcome of this learning is a global model that is used to perform inferences in real deployments.

While most current research has focused on several aspects of collaborative learning, the ultimate goal has always been inference. The potential use of collaboration solely for the purpose of inference has largely been overlooked. In this work, we consider the problem of "Collaborative Inference" (CI) in its own light and demonstrate that CI can be as competitive as CL at extremely low-cost. We emphasize that while CL has become a go-to solution for most tasks, organisations today should (re)consider and benefit from low cost and performant paradigm of CI.

2 CI DEFINITION & METHODS

We begin by formally defining collaborative inference as the following task. Suppose that we are given a set of M models $\pi_1, \pi_2, \ldots, \pi_M$ that are respectively trained on local datasets $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_M$ sampled from a common space $\mathcal{X} \times \mathcal{Y}$. The goal of collaborative inference is to find a function $f : \mathcal{Y}^M \rightarrow Y$ that best aggregates individual inferences from the models $\{\pi_i\}_{i=1}^M$ to maximize performance on the dataset $\mathcal{D}^* \in \mathcal{X} \times \mathcal{Y}$. Typically the local data distributions $\mathcal{D}_i$ are heterogeneous across nodes. Next we identify some methods for performing CI aggregation. We classify the methods as *training-free* and *training-required* methods depending on whether the method needs to be trained. The first two methods below are training-free while last two are training-required.

(1) **Averaging** – aggregate inferences using simple mean.
(2) **Weighted Averaging** – incorporates a weighted scheme based on number of classwise training samples.
(3) **Polychotomous Voting** – optimal voting rule that maximizes benefit based on Bayesian expectation maximization.
(4) **Neural network (NN)** – our approach is to model f as an arbitrary function that can be learnt by a neural network.

3 COST COMPARISON TO CL

Since nodes in CL exchange large models with the server for several communication rounds, CL incurs significant communication costs. On the other hand, when training f for CI, nodes must send only the inferences to the server, incurring very low communication cost. Additionally, for training-free aggregators f, no communication cost is incurred. However, CI must store all M local models to perform inference during production as compared to one final model in CL, thus requiring M times more memory than CL during production.

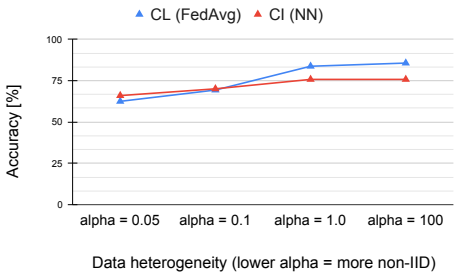

Figure 1: CI is very competitive with CL when the data is highly heterogeneous or non-IID ($\alpha = 0.05$ and $\alpha = 0.1$).

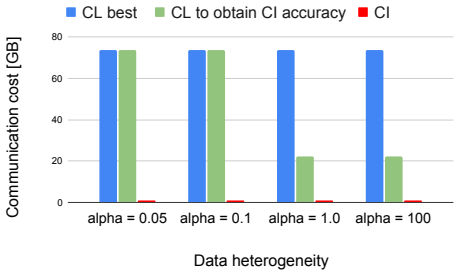

Figure 2: CL needs 67× more communication than CI.

4 PRELIMINARY RESULTS

We present results on the CIFAR-10 dataset with $M = 20$ nodes. Figure 1 charts the accuracy across different values of heterogeneity for CL using the FedAvg algorithm [3] and CI using the NN approach. Figure 2 charts the corresponding communication costs. The data is distributed heterogeneously using the Dirichlet approach [1]. Our results validate that CI can achieve competitive accuracies with CL while saving 67× in communication under realistic non-IID data distributions.

REFERENCES

[1] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. 2020. Federated visual classification with real-world data distribution. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*. Springer, 76–92.
[2] Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. 2017. Can Decentralized Algorithms Outperform Centralized Algorithms? A Case Study for Decentralized Parallel Stochastic Gradient Descent. In *NIPS'17*, Vol. 30. Long Beach, California, USA, 5336–5346. https://proceedings.neurips.cc/paper/2017/file/f75526659f31040afeb61cb7133e4e6d-Paper.pdf
[3] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*. PMLR, 1273–1282.