Ensemble-Based Decentralized Machine Learning

Tianbo Hao[1], Yi Zhou[2], and Justin Qu[3]

[1, 2, 3]Carnegie Mellon University
{*tianboh, yizhou2, jpqu*}*@andrew.cmu.edu*

ABSTRACT

Modern machine learning tasks can often require training on massive amounts of data that are difficult to save and work with on a single server. In many settings, training data are also generated from different contexts and geographical locations and are not easily sharable due to privacy or security concerns. Training on decentralized data partitions poses a ”fundamental and pervasive” challenge to training accuracy, since such data are often not independent and identically distributed (IID) and thus not amenable to traditional training techniques such as batch normalization [4]. [4] proposes SkewScout to enhance global training accuracy (accuracy across different data partitions) under the condition of data skew, and meanwhile tries to control the costs of parameter transfer by model traveling. Two challenges remain: 1. There is still room for improving accuracy (SkewScout gives about 79%); 2. Updates based on transferred parameters could de-emphasize distinctive patterns in the local data set, but SkewScout does not consider local accuracy.

In our work, we aim to tackle these two remaining challenges. We propose a decentralized ensemble learning algorithm based on ensemble learning [1]. The key idea is that the prediction of a data center will be based on an ensemble of each data center's local model with predictions from each local model weighted with an entropy-based confidence calculation. It is worth noting that our training and testing do not assume similar data distributions, a condition assumed by most traditional machine learning algorithms. We also give a novel definition of skewness and show that it is reasonable. We evaluate our model's performance in terms of global and local accuracy and compare it against Federated Learning, a popular decentralized learning architecture.

We do not consider model transfer costs in our experiments, but propose clustering, which could reduce parameter/model transfer costs, as the possible next step of research after ensemble.

Keywords: ensemble learning, non-IID distribution, decentralized learning

1. INTRODUCTION

Many machine learning applications require large-scaled data sets, and it may not be possible to train a model on a single server due to storage, memory, and compute bottlenecks.

However, this problem has been well studied from system's perspective and people can utilize cloud infrastructure to achieve the necessary memory, storage, and compute specifications by distributing the load across multiple machines.

Decentralized learning is a machine learning approach that enables multiple servers to train a single global model while only possessing a local subset of the global data set. Servers (data centers), periodically share their model or gradients with other machines in the group to update the global model[5]. These algorithms focus on maximizing the prediction accuracy over the global data set (combined data of all data centers) and often assume that all of the local data sets are similarly distributed (IID). However, since data may be collected from different geographical locations and posses unique labels or features due to the locality, this assumption of IID data sets across multiple