Distribution bias, federated learning, and meta learning

Erik Louie [1]

Abstract

In this project, various federated learning methods will be compared for their effectiveness in handling non-IID data. Furthermore, meta learning techniques and their application to federated learning will be investigated. Inspired by the design of meta learning algorithms, each method will be evaluated against both the global performance and the individual task performance.

1. Federated Learning

In recent years, federated learning has received wide attention. In part, this is due to claims of mitigating privacy issues and conforming to recent privacy law as well as minimizing costs and data collection requirements of centralized systems (Kairouz et al., 2019, pg. 4). McMahan, in their seminal paper, defines federated learning as the approach when "the learning task is solved by a loose federation of participating devices ... which are coordinated by a central server (McMahan et al., 2017, pg. 1)." In practice, federated learning is a nebulous term including any approaches to the challenges in coordinating learning of local data on devices by a centralized server to collaboratively solve a machine learning problem (Kairouz et al., 2019, pg. 4-5). Thus, federated learning includes a wide set of disciplines, including data privacy, systems security, distributed systems, parallelism, fault tolerance analysis, and even bias and fairness in machine learning.

In this project, the primary focus will be on the specific challenge of the effectiveness of federated learning algorithms. More specifically, one of the underlying assumptions of several federated learning algorithms is the identically and independently distributed data (IID Data). However, in real systems this is not an effective assumption, since there are several sources of non-IID data:

- Skewed label partitions (Hsieh et al., 2020),
- violations of independence, eg. available data is time-dependent (Kairouz et al., 2019, pg. 19),
- Dataset shifts (differences between the training sets and the test set) (Kairouz et al., 2019, pg. 19), and
- self-selection bias induced by algorithmic assumptions.

For a more thorough analysis of non-IID data issues, see (Hsieh et al., 2020) and (Kairouz et al., 2019). For self-selection bias, consider the design proposed by (Bonawitz et al., 2019) where devices are oversampled and only the fastest and most reliable devices are included for aggregation.

In this project, several popular approaches to federated learning will be considered as well as their consequences on the effectiveness and accuracy under several metrics. The first popular algorithm for federated learning is federated averaging (FedAvg) first proposed by (McMahan et al., 2017). In this paper, local SGD minibatches are aggregated at the server level and returned to the clients for the following round (epoch). Another approach, varianced reduced local SGD (VRL-SGD), attempts to overcome the non-IID problem while providing a linear speed up in convergence in the federated learning setting by using an SVRG-like variance reduction technique to the FedAvg algorithm (Liang et al., 2019). Stochastic Controlled Averaging for Federated Learning (SCAFFOLD) was proposed concurrently with VRL-SGD, but instead draws inspiration from Distributed Approximate NEwton (DANE), a federated learning-like distributed learning algorithm for the IID setting that estimates Newton steps to improve convergence similar to ADMM (Shamir et al., 2014), and has been shown to converge faster and with lower error than FedAvg or global SGD (Karimireddy et al., 2019).

Finally, this project will explore a new approach with an algorithm that has similarities with federated learning, Model Agnostic MetaLearning (MAML). MAML was designed for heterogeneous few-shot metalearning by iterating over tasks during each epoch to generate updates to train a metalearner that can be fine-tuned to adapt to new tasks in few iterations (Finn et al., 2017). MAML has been proposed recently in two papers as a method for personalized federated learning

(Jiang et al., 2019; Khodak et al., 2019). Of particular interest is that concept drift has long been studied in the context of meta learning due to the periods over which task data is collected (Klinkenberg, 2005; Jiang et al., 2019; Khodak et al., 2019; Kairouz et al., 2019). However, the models proposed have not been evaluated in the context of non-IID performance, which this project will evaluate. Also, the metalearning approaches suggest an alternative metric by which performance should be measured: per-device data accuracy rather than a global held-out dataset. An approach to this in the federated learning context will be explored.

2. Simulation environment

The algorithms will be studied in the TensorFlow federated learning framework. The framework provides models and interfaces in order to simulate a federated learning experiment, including implemented versions of common federated learning algorithms and local and global aggregation isolation (ten, 2020).

3. Datasets

Three datasets will be explored for benchmarking: MNIST, Omniglot, and NABirds. As a baseline metric, the project will use the MNIST example from TensorFlow to implement each of FedAvg, VRL-SGD, and MAML federated learning methods. The Omniglot dataset is a common benchmark for meta learning algorithms and the one used in the Model Agnostic Meta Learner paper. Omniglot is a collection of scripts from various language, treated as separate tasks in meta learning. The NABirds dataset is freely available for the research community and includes ”a collection of 48,000 annotated photographs of the 400 species of birds that are commonly observed in North America.” Given the size of this dataset and the imbalanced nature of it, NABirds should be effective in demonstrating differences in data distribution.

4. Experimental design

The goal of the experiments is to demonstrate the effectiveness of various algorithms in response to non-IID and biased data samples in devices.

4.1. Hypothesis

In this project, it is hypothesized that FedAvg and MAML-based models will have a high effective performance due to bias to the localized data, but will perform worse against a global, unbiased benchmark of performance.

4.2. Proxy

In this experiment, the convergence time will be measured as

1. the total number of iterations of the data (statistical performance), and
2. the sum length of time of the longest client computation per round (hardware performance).

While the simulation cannot infer the effects of communication time, the length of the longest computation will serve as the substitutive metric. Given sufficient time, the effective performance will be measured in terms of

1. the accuracy against a global holdout test set,
2. the average accuracy against a holdout test set per device,
3. the F-measure treating each device as a separate task, and
4. a single unseen task / device measuring the generalizability of the model.

The last measure is inspired by the meta learning approach to benchmarking. In this case, the device will use the global model and train for a small number of iterations against a subset of the task and test with the remaining.

4.3. Protocol

For each of the algorithms of global SGD, FedAvg, VRL-SGD, and MAML-based algorithm will be run with a standard CNN against the MNIST, Omniglot, and NABirds. Each algorithm will be run until test performance peaks or reduces.

4.4. Expected Results

It is expected that

1. FedAvg will perform worse than global SGD, VRL-SGD, and SCAFFOLD in both convergence and accuracy on a global holdout set,
2. algorithms that are tailored to reduce effects of non-IID will perform worse against local holdout test sets than FedAvg or MAML-based methods, and
3. VRL-SGD and SCAFFOLD will converge fastest in iterations, but slower in wall clock time.

Algorithm 1 FedAvg

```
Input: data $x_i$, size $m$
repeat
  Initialize $noChange = true$.
  for $i = 1$ to $m - 1$ do
    if $x_i > x_{i+1}$ then
      Swap $x_i$ and $x_{i+1}$
      $noChange = false$
    end if
  end for
until $noChange$ is $true$
```

5. Prior Work

5.1. Federated Averaging

The FedAvg algorithm attempts to balance latency between clients and servers with the consequences of statistical heterogeneity. As in distributed SGD, global updates are sent to the local clients. Unlike distributed SGD, the client is permitted to run multiple SGD steps over batches of local data. By computing the local SGD multiple times mcmahan2017communicationefficient argue that FedAvg should converge in fewer rounds than FedSGD. Thus, the algorithm can be computed in two stages as in 1.

6. Experimental Setup

7. Acknowledgement

Data provided by the Cornell Lab of Ornithology, with thanks to photographers and contributors of crowd-sourced data at AllAboutBirds.org/Labs. This material is based upon work supported by the National Science Foundation under Grant No. 1010818.

References

Federated learning, 08 2020. URL https://www.tensorflow.org/federated/federated_learning.

Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konečnỳ, J., Mazzocchi, S., McMahan, H. B., et al. Towards federated learning at scale: System design. arXiv preprint arXiv:1902.01046, 2019.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks, 2017.

Hsieh, K., Phanishayee, A., Mutlu, O., and Gibbons, P. B. The non-iid data quagmire of decentralized machine learning, 2020.

Jiang, Y., Konečný, J., Rush, K., and Kannan, S. Improving federated learning personalization via model agnostic meta learning, 2019.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. Advances and open problems in federated learning. arXiv preprint arXiv:1912.04977, 2019.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S. J., Stich, S. U., and Suresh, A. T. Scaffold: Stochastic controlled averaging for on-device federated learning. arXiv preprint arXiv:1910.06378, 2019.

Khodak, M., Balcan, M.-F., and Talwalkar, A. Adaptive gradient-based meta-learning methods, 2019.

Klinkenberg, R. Meta-learning, model selection, and example selection in machine learning domains with concept drift. In LWA, volume 2005, pp. 164–171, 2005.

Liang, X., Shen, S., Liu, J., Pan, Z., Chen, E., and Cheng, Y. Variance reduced local sgd with lower communication complexity, 2019.

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data, 2017.

Shamir, O., Srebro, N., and Zhang, T. Communication efficient distributed optimization using an approximate newton-type method, 2014.