**Paper Review:** McMahan et al. “Communication-Efﬁcient Learning of Deep Networks." Accessed December 6, 2019. https://arxiv.org/pdf/1602.05629.pdf.

DATA590: University of Washington

Joel Stremmel

12-06-19­­

**Thesis:**

Federated Learning aims to train machine learning models in a distributed fashion without centralizing data but instead updating and passing model parameters from a central server to distributed entities and back to perform stochastic gradient descent. McMahan et al. propose the Federated Averaging algorithm in “Communication-Efficient Learning of Deep Networks from Decentralized Data” and report on a series of experiments suggesting that despite perceived limitations to training across non-independent-identically-distributed samples, Federated Averaging provides a surprisingly simple and effective method for secure and distributed on-device training with communication costs being the main constraint.

**Properties of Federated Optimization:**

Federated optimization can be distinguished from general, distributed optimization in that the training data is non-IID, with the data from one client device unlikely to be a representative sample of the population. Additionally, the data are unbalanced, with some clients having far more data than others. Finally, with many devices and limited communication options, federated optimization must handle device dropout and message passing across potentially hundreds of clients.

**The Federated Averaging Algorithm:**

In Federated Stochastic Gradient Descent “each client locally takes one step of gradient descent on the current model using its local data, and the server takes a weighted average of the resulting models.” It is also possible to have each device take multiple gradient steps before passing updates to the central server for averaging. This approach, termed Federated Averaging by the authors, can reduce the number of communication rounds and is the principle algorithm explored in the paper.

**Experimental Results:**

Experimental results indicate that by federated training of:

1. A two-layer CNN on MNIST with non-IID samples, where records and labels are *sorted* then distributed, and
2. A two-layer, character-level LSTM on Shakespeare plays, where lines in plays are distributed by speaking roles

Test accuracy compared to IID training with each dataset yields similar test set accuracy. Additionally, the authors report that test accuracy for these datasets remains stable up to a certain number of local training iterations before central averaging of model parameters, with local training sometimes demonstrating higher test accuracy than averaging on the central server after each round of local training. The authors conjecture that model averaging yields regularization benefit similar to dropout, and the reduced number of communication rounds, demonstrates the potential value of Federated Averaging in practice. Larger scale experiments on the CIFAR-10 dataset and social network posts further validate that Federated Averaging can reduce required communication rounds for the same level of accuracy as Federated SGD.

**Practical Takeaways:**

The authors detail the value of Federated Learning for training models on distributed devices without centrally aggregating data and the value of the Federated Averaging algorithm for reducing the number of training rounds and potentially providing regularization. With the increased use of personal computing devices and the desire for personal data privacy, Federated Learning represents an important advancement in machine learning technology. As a data scientist working in healthcare, I am particularly interested in potential healthcare applications of Federated Learning to model patient health, disease, and mortality risk without centrally aggregating sensitive clinical data. I am excited to work on Federated Learning experiments in my research project with Google and glad that the Tensorflow Federated API is open-source and provides the ability to simulate federated model training.