Prithvi Kannan

405110096

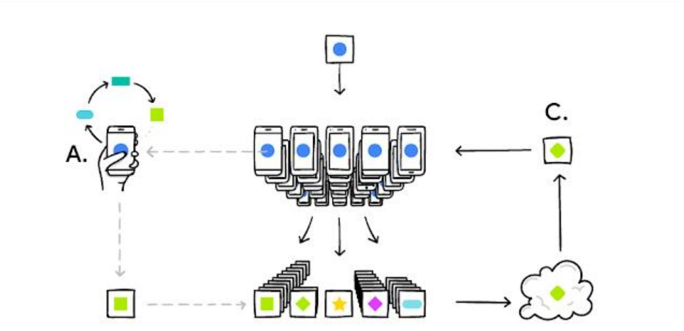
Assignment 10: Federated Learning

One of the hottest fields today is machine learning and big data. Both ideas are centered on the idea that by aggregating data millions of user's trends, we can best predict what to show to future users. Developments in the fields of distributed systems and cloud infrastructure have made it easier to centralize the training data on one machine or datacenter, but this has a few downsides. One of which is that a user's data is being harvested and sent to the cloud, where it is stored, and therefore vulnerable to data breaches. Another disadvantage is the latency of updates in the model, having to do with the time to send new training data to the cloud, recompute the model, and then redeploy the model back to the user's device. Lastly, consider the affect of a single drop in a large bucket of water, analogous to one's data being fed through a model that has already been trained on millions of datapoints. The model is not going to change much because of one user, so that user's version of the product is not personalized to them.

Federated Learning is a new approach that attempts to tackle some of these challenges by collaboratively learning a shared model, while keeping an individual's data on the device, meaning massive centralized training sets are no longer needed in the cloud.

Federated Learning works as follows:

1. Download the current master model to the device
2. As your device produces more data, train the locally stored model
3. Batch up the changes and send the update to the cloud (via encrypted channels)
4. In the cloud, process the update and make chances to the master model



Source: Google AI Blog

All the training data remains on your device, and no human readable individual updates are stored in the cloud.

This idea has been around for a while, but the recent news is that a team of engineers at Google have figured out how to deploy this at scale on the GBoard product (Google keyboard). The solution they have built uses federated learning to give users the best typing experience on their mobile devices. The phone downloads the model locally when the user is not using their phone, then stores predicted results versus the user's actual keystrokes as the user is typing. Once the user is no longer using their phone, GBoard locally runs stochastic gradient descent, which is a way of training the model using only a finite number of additional data points, and sends those encrypted weights back up to the cloud. This creates a seamless and secure user experience while improving the master GBoard model. In the approach that Google has taken, federated learning seems to be achievable without reducing performance even though they are making calculations locally instead of in the cloud. I anticipate that for more difficult calculations, such as video processing at high scale, it may not be efficient to use federated learning in this way. As with most research, this blog post shows that this technology is feasible, and now other researchers at Google and elsewhere will attempt to build off the results.

When I heard about this new technology, I was surprised that there was a way to train models without centralized datasets. In my (limited) experience with machine learning models, I've always figured that the data had to be in the same place. I think this technology can be disruptive, especially in light of recent privacy concerns. To me, the most important benefit is that my data is not being sent to Google or Facebook or whatever app I am using, but rather only an encrypted "changelog" of updated model weights.

The applications of federated learning are immense. The authors of the paper concede that certain machine learning problems, such as recognizing dog breeds, cannot be solved using Federated Learning, but some domains include improving the language models based on what you actually type on your phone (since people's phones can have different layouts) and photo rankings based on what kinds of photos people look at, share, or delete.

Another realm of potential impact is in advertising, where there is currently a lot of buzz regarding data harvesting and selling. A common critique is that "Facebook is selling my data to marketers", and ignoring the validity of that concern, it's unquestionable that people would be more comfortable if their personal data was not directly handled by the platform that they are using.

Healthcare is another segment ripe for disruption by Federated Learning. Drug research is more and more dependent on real world data, but that presents two problems: aggregating data from diverse medical institutions and securing sensitive patient data. Federated learning is starting to be applied to medicine to enable researchers to train their models without requiring hospitals to upload patient data to a centralized repository.

Beyond the concerns of privacy, there may also be performance benefits to federated learning, which can be instrumental to the self-driving car industry. Due to the sheer volume of self-driving cars and the amount of data collected by each one, the traditional central cloud approach may not work. Instead, applying federated learning can limit the volume of data transfer and accelerate machine learning models by making more calculations onboard and then discarding the data.

Works Cited:

Brendan McMahan and Daniel Ramage. Federated Learning: *Collaborative Machine Learning without Centralized Training Data*. 2017. https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

Fred Donovan. *Federated Learning Balances Machine Learning with Patient Privacy*. 2019. https://hitinfrastructure.com/news/federated-learning-balances-machine-learning-with-patient-privacy

Xinle Liang, Yang Liu, Tianjian Chen, Ming Liu, Qiang Yang. *Federated Transfer Reinforcement Learning for Autonomous Driving*. 2019. https://arxiv.org/pdf/1910.06001.pdf

Keith Bonawitz, Hubert Eichner. *TOWARDS FEDERATED LEARNING AT SCALE: SYSTEM DESIGN*. 2019.https://arxiv.org/pdf/1902.01046.pdf