BLOG ›

Federated Learning: Collaborative Machine Learning without Centralized Training Data

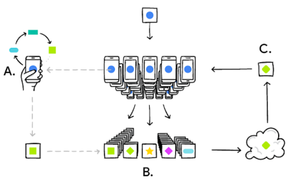
THURSDAY, APRIL 06, 2017

Posted by Brendan McMahan and Daniel Ramage, Research Scientists

Standard machine learning approaches require centralizing the training data on one machine or in a datacenter. And Google has built one of the most secure and robust cloud infrastructures for processing this data to make our services better. Now for models trained from user interaction with mobile devices, we're introducing an additional approach: *Federated Learning*.

Federated Learning enables mobile phones to collaboratively learn a shared prediction model while keeping all the training data on device, decoupling the ability to do machine learning from the need to store the data in the cloud. This goes beyond the use of local models that make predictions on mobile devices (like the Mobile Vision API and On-Device Smart Reply) by bringing model *training* to the device as well.

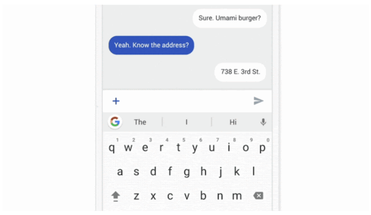
It works like this: your device downloads the current model, improves it by learning from data on your phone, and then summarizes the changes as a small focused update. Only this update to the model is sent to the cloud, using encrypted communication, where it is immediately averaged with other user updates to improve the shared model. All the training data remains on your device, and no individual updates are stored in the cloud.



Your phone personalizes the model locally, based on your usage (A). Many users' updates are aggregated (B) to form a consensus change (C) to the shared model, after which the procedure is repeated.

Federated Learning allows for smarter models, lower latency, and less power consumption, all while ensuring privacy. And this approach has another immediate benefit: in addition to providing an update to the shared model, the improved model on your phone can also be used immediately, powering experiences personalized by the way you use your phone.

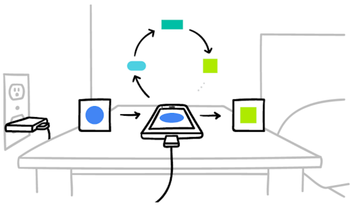
We're currently testing Federated Learning in Gboard on Android, the Google Keyboard. When Gboard shows a suggested query, your phone locally stores information about the current context and whether you clicked the suggestion. Federated Learning processes that history on-device to suggest improvements to the next iteration of Gboard's query suggestion model.



To make Federated Learning possible, we had to overcome many algorithmic and technical challenges. In a typical machine learning system, an optimization algorithm like Stochastic Gradient Descent (SGD) runs on a large dataset partitioned homogeneously across servers in the cloud. Such highly iterative algorithms require low-latency, high-throughput connections to the training data. But in the Federated Learning setting, the data is distributed across millions of devices in a highly uneven fashion. In addition, these devices have significantly higher-latency, lower-throughput connections and are only intermittently available for training.

These bandwidth and latency limitations motivate our Federated Averaging algorithm, which can train deep networks using 10-100x less communication compared to a naively federated version of SGD. The key idea is to use the powerful processors in modern mobile devices to compute higher quality updates than simple gradient steps. Since it takes fewer iterations of high-quality updates to produce a good model, training can use much less communication. As upload speeds are typically much slower than download speeds, we also developed a novel way to reduce upload communication costs up to another 100x by compressing updates using random rotations and quantization. While these approaches are focused on training deep networks, we've also designed algorithms for high-dimensional sparse convex models which excel on problems like click-through-rate prediction.

Deploying this technology to millions of heterogenous phones running Gboard requires a sophisticated technology stack. On device training uses a miniature version of TensorFlow. Careful scheduling ensures training happens only when the device is idle, plugged in, and on a free wireless connection, so there is no impact on the phone's performance.



Your phone participates in Federated Learning only when it won't negatively impact your experience.

The system then needs to communicate and aggregate the model updates in a secure, efficient, scalable, and fault-tolerant way. It's only the combination of research with this infrastructure that makes the benefits of Federated Learning possible.

Federated learning works without the need to store user data in the cloud, but we're not stopping there. We've developed a Secure Aggregation protocol that uses cryptographic techniques so a coordinating server can only decrypt the average update if 100s or 1000s of users have participated — no individual phone's update can be inspected before averaging. It's the first protocol of its kind that is practical for deep-network-sized problems and real-world connectivity constraints. We designed Federated Averaging so the coordinating server only needs the average update, which allows Secure Aggregation to be used; however the protocol is general and can be applied to other problems as well. We're working hard on a production implementation of this protocol and expect to deploy it for Federated Learning applications in the near future.

Our work has only scratched the surface of what is possible. Federated Learning can't solve all machine learning problems (for example, learning to recognize different dog breeds by training on carefully labeled examples), and for many other models the necessary training data is already stored in the cloud (like training spam filters for Gmail). So Google will continue to advance the state-of-the-art for cloud-based ML, but we are also committed to ongoing research to expand the range of problems we can solve with Federated Learning. Beyond Gboard query suggestions, for example, we hope to improve the language models that power your keyboard based on what you actually type on your phone (which can have a style all its own) and photo rankings based on what kinds of photos people look at, share, or delete.

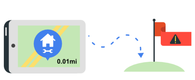
Applying Federated Learning requires machine learning practitioners to adopt new tools and a new way of thinking: model development, training, and evaluation with no direct access to or labeling of raw data, with communication cost as a limiting factor. We believe the user benefits of Federated Learning make tackling the technical challenges worthwhile, and are publishing our work with hopes of a widespread conversation within the machine learning community.

Acknowledgements

This post reflects the work of many people in Google Research, including Blaise Agüera y Arcas, Galen Andrew, Dave Bacon, Keith Bonawitz, Chris Brumme, Arlie Davis, Jac de Haan, Hubert Eichner, Wolfgang Grieskamp, Wei Huang, Vladimir Ivanov, Chloé Kiddon, Jakub Konečný, Nicholas Kong, Ben Kreuter, Alison Lentz, Stefano Mazzocchi, Sarvar Patel, Martin Pelikan, Aaron Segal, Karn Seth, Ananda Theertha Suresh, Iulia Turc, Felix Yu, Antonio Marcedone and our partners in the Gboard team.

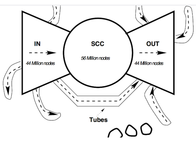
Labels: Collaboration Machine Learning On-device Learning optimization

Previous posts



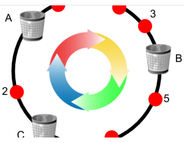
APR 7, 2017

Keeping fake listings off Google Maps



APR 5, 2017

And the award goes to...



APR 4, 2017

Consistent Hashing with Bounded Loads