Google’s Federated Learning Architecture Can Enhance Privacy While Ending The Centralised Dataset

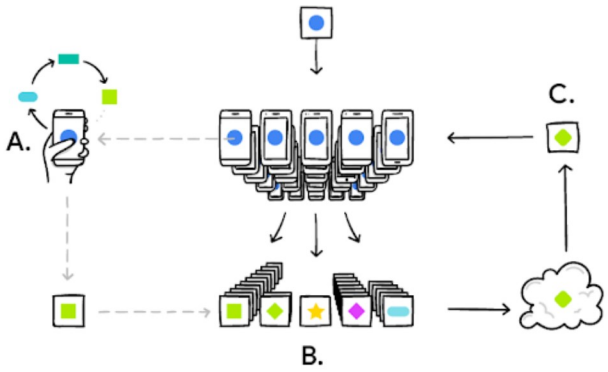
Traditionally, datasets are a huge part of the machine learning pipeline. With the advent of deep learning techniques, the amount of high quality, well-labelled data is paramount to the success of the machine learning project. The standard way of executing machine learning also needs the data to be on the datacentre or the machine where the model is being trained. But now, engineers at Google have come up with a new secure and robust cloud infrastructure architecture for processing data called Federated Learning. It has been created for models which are trained from user interactions on mobile devices.

A significant amount of research has been done on enabling the efficient distributed training of neural networks. We can take several approaches to distribute the training of deep learning networks.

- Model parallelism: Here, different machines in the distributed system are responsible for the computations in different parts of a single network
- Data parallelism: Here different machines have a complete copy of the model; each machine simply gets a different portion of the data, and results from each are somehow combined.

Federated Learning : The Architecture

Federated Learning takes advantage of mobile phones to collaborate and learn a shared prediction model. This is done while all the training data stays on the device and is not sent over to the cloud. This decoupling helps us to execute machine learning on the device without the the need to store the data in the cloud. This is different from the use cases such as the Mobile Vision API and On-Device Smart Reply where prediction is done on device; in federated learning the model training happens on the device as well.



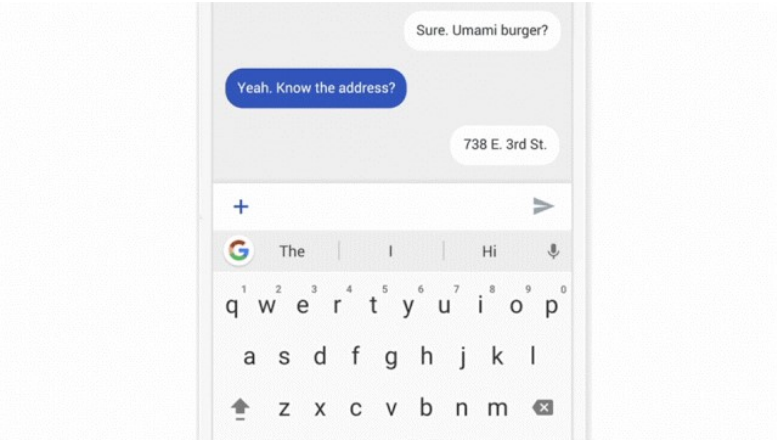
The phone personalises the model locally, based on your usage (A). Many users' updates are aggregated (B) to form a consensus change (C) to the shared model, after which the procedure is repeated.

Here how it works: The mobile device downloads the current model, improves it by learning from data on your phone, and then makes a summary of the knowledge it has learned from the data 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.

This kind of architecture results in smarter models, lower latency, and less power consumption; all this while ensuring complete privacy of user data. The architecture also sends an update to the shared model, but the improved model on the phone can now be used immediately, and this results in a very powerful user experience.

Making Federated Learning Possible

The system of Federated Learning is already being tested in Gboard on Android, the very popular Google Keyboard. Whenever Gboard shows a suggested query, the mobile device locally stores information about the current context and whether you used the suggestion. Federated Learning processes that history on-device to suggest improvements to the next iteration of Gboard's query suggestion model.



The research team at Google had to overcome many algorithmic and research challenges to make federated learning possible. Optimisation algorithm like Stochastic Gradient Descent (SGD) which are typically used in many machine learning systems runs on a large dataset. Most of the times these datasets are partitioned homogeneously across servers in the cloud. Many highly iterative algorithms require low-latency, high-throughput connections to the training data. But in this particular setting, the data is distributed across millions of mobile and cellular devices in a highly heterogeneous fashion. In addition, these devices have significantly higher-latency, lower-throughput connections and are only intermittently available for training.

To put such a system in deployment to millions heterogenous phones running Gboard requires a fairly advanced technology stack. On-device training uses a minimised version of TensorFlow. Upload speeds are typically much slower than download speeds, the researchers also developed a novel way to reduce upload communication costs up to another 100x by compressing updates using random rotations and quantisation.

The Future Of Federated Learning

The researchers still feel that the work is only the beginning. They said in their official blog, "Our work has only scratched the surface of what is possible. Federated Learning can't solve all machine learning problems (for example, learning to recognise 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)."

The application of Federated Learning requires that machine learning practitioners to use new tools and a new way of looking at the problem. Model development, training, and evaluation with no direct access to or labelling of raw data, with communication cost as a limiting factor. The researchers believe that the user benefits of Federated Learning make tackling the technical challenges worthwhile.

Provide your comments below

comments