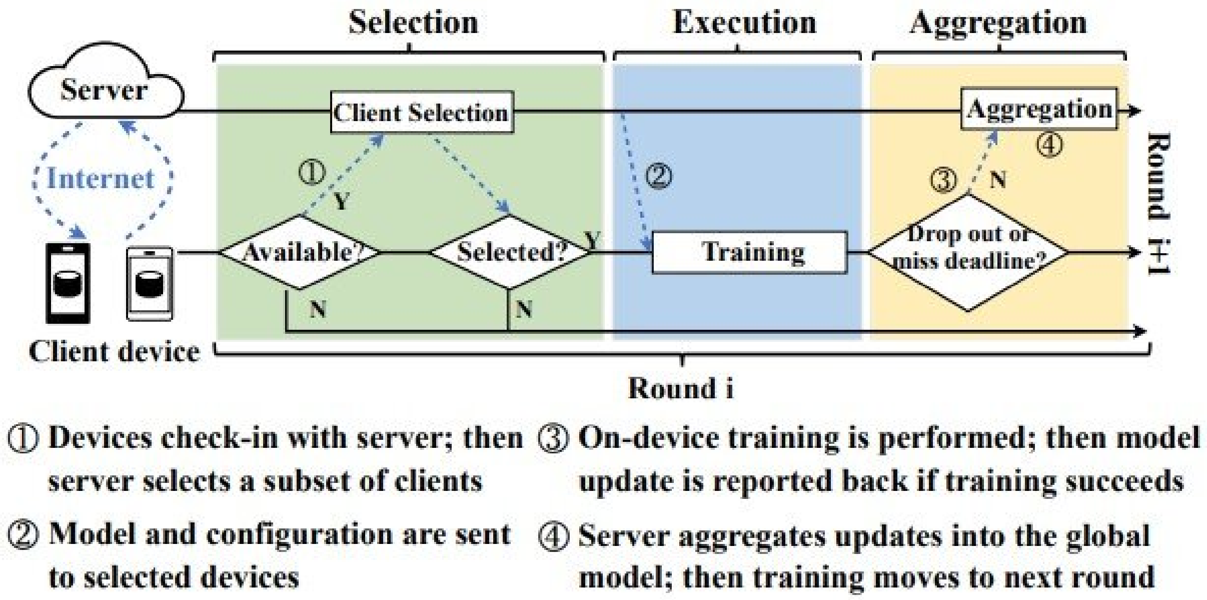
TechXplore

Open source platform enables research on privacy-preserving machine learning

July 20 2022, by Zach Champion



① Devices check-in with server; then server selects a subset of clients

② Model and configuration are sent to selected devices

③ On-device training is performed; then model update is reported back if training succeeds

④ Server aggregates updates into the global model; then training moves to next round

Standard FL protocol. Credit: https://arxiv.org/abs/2105.11367

The biggest benchmarking data set to date for a machine learning technique designed with data privacy in mind has been released open source by researchers at the University of Michigan.

Called federated learning, the approach trains learning models on end-user devices, like smartphones and laptops, rather than requiring the transfer of private data to central servers.

"By training in-situ on data where it is generated, we can train on larger real-world data," explained Fan Lai, U-M doctoral student in computer science and engineering, who presents the FedScale training environment at the International Conference on Machine Learning this week.

"This also allows us to mitigate privacy risks and high communication and storage costs associated with collecting the raw data from end-user devices into the cloud," Lai said.

Still a new technology, federated learning relies on an algorithm that serves as a centralized coordinator. It delivers the model to the devices, trains it locally on the relevant user data, and then brings each partially trained model back and uses them to generate a final global model.

For a number of applications, this workflow provides an added data privacy and security safeguard. Messaging apps, health care data, personal documents and other sensitive but useful training materials can improve models without fear of data center vulnerabilities.

In addition to protecting privacy, federated learning could make model training more resource-efficient by cutting down and sometimes eliminating big data transfers, but it faces several challenges before it can be widely used. Training across multiple devices means that there are no guarantees about the computing resources available, and uncertainties like user connection speeds and device specs lead to a pool of data options with varying quality.

"Federated learning is growing rapidly as a research area," said Mosharaf Chowdhury, U-M associate professor of computer science and engineering. "But most of the work makes use of a handful of data sets, which are very small and do not represent many aspects of federated learning."

And this is where FedScale comes in. The platform can simulate the behavior of millions of user devices on a few GPUs and CPUs, enabling developers of machine learning models to explore how their federated learning program will perform without the need for large-scale deployment. It serves a variety of popular learning tasks, including image classification, object detection, language modeling, speech recognition and machine translation.

"Anything that uses machine learning on end-user data could be federated," Chowdhury said. "Applications should be able to learn and improve how they provide their services without actually recording everything their users do."

The authors specify several conditions that must be accounted for to realistically mimic the federated learning experience: heterogeneity of data, heterogeneity of devices, heterogeneous connectivity and availability conditions, all with an ability to operate at multiple scales on a broad variety of machine learning tasks. FedScale's data sets are the largest released to date that cater specifically to these challenges in federated learning, according to Chowdhury.

"Over the course of the last couple years, we have collected dozens of data sets. The raw data are mostly publicly available, but hard to use because they are in various sources and formats," Lai said. "We are continuously working on supporting large-scale on-device deployment, as well."

The FedScale team has also launched a leaderboard to promote the most successful federated learning solutions trained on the U-M system.

More information: Fan Lai et al, FedScale: Benchmarking Model and System Performance of Federated Learning at Scale. arXiv:2105.11367v5 [cs.LG], arxiv.org/abs/2105.11367

Provided by University of Michigan

Citation: Open source platform enables research on privacy-preserving machine learning (2022, July 20) retrieved 17 April 2024 from https://techxplore.com/news/2022-07-source-platform-enables-privacy-preserving-machine.html