Case Study

Artificial Intelligence | Federated Learning

Secure Federated Learning for a Better World

The medical industry's largest collaborative federated learning project to date between Intel Labs, the University of Pennsylvania and dozens of healthcare institutions improves accuracy of brain tumor detection by up to 33%[1]

What's New

The project has demonstrated FL's influence in revolutionizing healthcare as well as the value of running FL on Intel technology:

- More than double the number of participating healthcare and research sites across six continents (compared to 2020 numbers).[2]
- The largest and most diverse dataset of glioblastoma patients ever considered in the literature (5 TB of data from 6,314 glioblastoma patients).[3]
- Up to 33% more accurate brain tumor detection, compared to models trained on publicly available datasets.[3]
- Up to 4.48x lower latency and 2.29x lower memory utilization, compared to the first consensus model, resulting from model optimization using the Intel® Distribution of OpenVINO™ toolkit[3]—enabling the consensus model to run on edge devices in clinics.
- An open-source framework for deploying FL that is easy to use, secure and scalable.

Intel Labs sponsors science and technology centers at universities around the world to encourage collaboration and bring world-changing concepts from ideation to production. These partnerships with academic institutions have the power to transform research methodology into real-world applications that have the potential to revolutionize industries and even save lives.

Challenge

Early detection of brain tumors can reduce the impact of surgery and treatment, improving the prognosis for many patients. But as healthcare shifts from reactive to proactive scanning, the number of skilled technicians cannot keep up with the number of brain scans being generated. Machine learning (ML) models can help automate scan analysis, but model accuracy is a concern. Using larger training datasets increases accuracy, but healthcare institutions are historically reluctant to share data to create these larger datasets due to privacy concerns.

Solution

Intel Labs and the Perelman School of Medicine at the University of Pennsylvania spearheaded a three-year project that used data from over 70 geographically distinct sites to apply federated learning (FL) to ML model training. With FL powered by hardware and software from Intel, brain scan data can be used for collaborative model training without the need for collaborators to share their raw data.

> "Using Intel software and hardware, a team of highly motivated researchers from Intel Labs and the University of Pennsylvania worked with a federation of collaborating medical centers to advance the detection of brain tumors while protecting sensitive patient data."
>
> **— Jason Martin**
> Principal Engineer, Intel Labs

Technical Components of the Solution

- **Intel® Software Guard Extensions** (Intel® SGX) protects data and code in use.
- **OpenFL** is an open-source framework for federated learning.
- **Gramine Project** provides several tools and infrastructure components for running unmodified applications on confidential computing platforms based on Intel SGX.
- **Intel® Distribution for OpenVINO™** toolkit helps streamline models' performance.

Healthcare Needs AI, but AI Needs More—and More Secure—Data

Recent technological advancements in healthcare, coupled with patients' culture shifting from reactive to proactive, have resulted in a large increase in health screenings, such as brain scans. This places a tremendous burden on clinical experts, as such scans require expert assessment. To alleviate this situation, there have been numerous efforts to develop ML models that can simulate expert human analysis to detect brain tumors.

While there has been some success in developing such models, there are concerns about their generalizability on data from sources that were not used in model training (called out-of-sample data). Training robust and accurate models requires large amounts of data, the diversity of which affects model generalizability to "out-of-sample" cases. To address these concerns, models need to be trained on data originating from numerous sites that represent diverse population samples.

The current paradigm for such multi-site collaborations is centralized learning (CL), in which data from different sites are shared to a centralized location per inter-site agreements. However, such data centralization is nearly impossible to scale globally due to concerns relating to privacy, data ownership, intellectual property, technical challenges (e.g., network and storage limitations) and compliance with varying national and regional regulatory policies.

Helping Advance Federated Learning for Healthcare Use Cases

Working with Intel Labs, the Perelman School of Medicine at the University of Pennsylvania (Penn Medicine)[4] co-developed technology to enable a federation using data from more than 70 geographically distinct healthcare and research sites[5] from around the world to train ML models that detect brain tumors using FL. The three-year project applied confidential computing concepts to FL to encourage collaboration in the healthcare community (see the next section for technology details of the solution).

Unlike CL, FL is a distributed ML approach (see Figure 1) that enables healthcare organizations to collaborate without sharing sensitive patient data, thereby increasing patient privacy. With FL, models are trained by sharing only model parameter updates from decentralized data, meaning each site retains its data locally. FL models have similar performance compared to CL-trained models.[6] Thus, FL has the potential to increase access to geographically distinct collaborators, thereby increasing the size and diversity of data used to train ML models.

The project used FL for brain tumor segmentation to determine the boundary of a tumor for a rare form of cancer called glioblastoma. Institutions collaborated to train models across private data that, if centralized, would represent a greatly expanded version of the International Brain Tumor Segmentation (BraTS) challenge dataset. The consensus dataset was based on 25,256 MRI scans (over 5 TB) from 6,314 glioblastoma patients across six continents.[7] Notably, this describes the largest and most diverse dataset of glioblastoma patients ever considered in the literature.

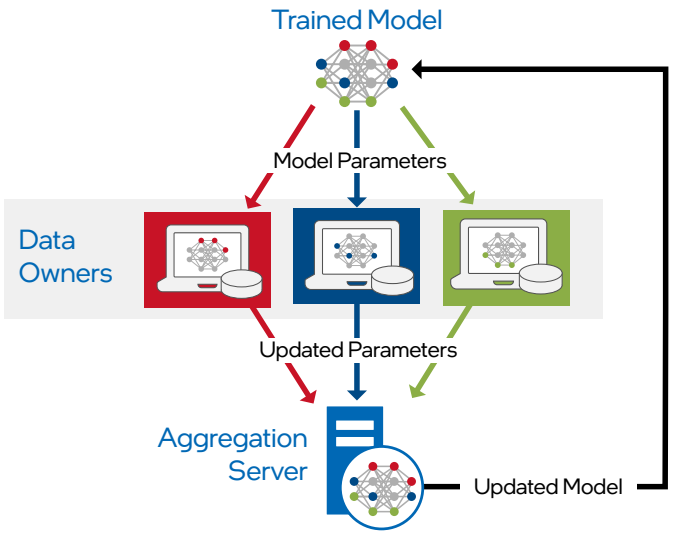

Figure 1. Overview of federated learning.

Software and Hardware Innovations from Intel Take Federated Learning to the Next Level

In 2018, Penn Medicine and Intel Labs presented results in a paper on federated learning in the medical imaging domain, demonstrating that FL could train a model with over 99% of the accuracy of a model trained in the traditional, non-private method. Since then, the project has taken advantage of Intel software and hardware to implement FL in a manner that:

- Provides additional privacy to the model and data.
- Makes the solution available on an open-source platform.
- Takes advantage of pre-packaged tools and libraries to deploy applications more easily with Intel SGX.
- Uses optimization techniques to enable the models to run on devices with limited compute and storage resources.

Enhanced Data and Code Protection with Intel SGX

Data has historically been protected in motion (over the network) and at rest (in storage). But what about protecting data while in use? Over a decade ago, Intel Labs began groundbreaking research into how data—and the code that uses the data—could be protected while an application is running. The result was Intel SGX, now a mature, enterprise-grade technology available with 3rd Generation Intel® Xeon® Scalable processors. Intel SGX allows organizations to isolate the software and data from the underlying infrastructure (hardware or OS) by using hardware-level encryption, which creates secure enclaves (see Figure 2). In addition to helping defend against the myriad of more common software-based attacks, Intel SGX's attestation mechanisms can verify that an application has not been compromised and that the processor it is running on has the latest security updates.

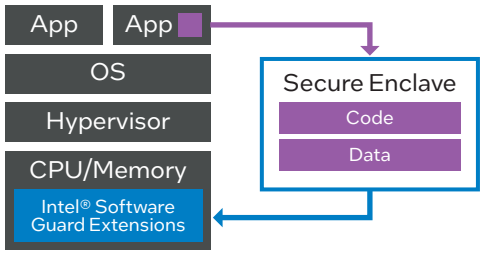

Figure 2. Intel® SGX is designed to prevent exposure of protected code and data even if a breach in the OS and hypervisor layers occurs, because they do not have access even at a privileged level.

Application code that performs within an Intel SGX enclave executes within the context of its parent application, thereby benefiting from the full power of the Intel processor. Within the enclave, code and data remain protected even when the BIOS, VM Manager, hypervisor, OS or drivers are compromised, implying that an attacker with full execution control over the platform can be kept at bay. Intel SGX also features memory protections that thwart memory bus snooping, memory tampering and "cold boot" attacks on images retained in DRAM.

The Penn Medicine FL project used Intel SGX to protect ML algorithms (essentially intellectual property), client data and the compute integrity of the FL components, which can include additional algorithmic privacy solutions. With confidence that both data and code are safe from unauthorized use, many more healthcare institutions may be willing to join similar FL projects that can enhance healthcare outcomes for many use cases beyond brain tumors.

> Intel Labs began groundbreaking research into how data—and the code that uses the data—could be protected while an application is running.

Easy-to-Use Framework for Getting Started with Secure Federated Learning

OpenFL is an open-source, Python 3-based framework for FL that is designed to be an easy-to-use, secure, scalable and extensible tool for data scientists. OpenFL is available on GitHub, along with tutorials and documentation to help organizations get started with their own FL projects. OpenFL is designed to be compatible with any ML or deep learning (DL) framework, and has tutorials and multiple examples using TensorFlow, PyTorch and MXNet.

OpenFL combines hardware and software to enable privacy-preserving AI using Intel SGX and Gramine (more on Gramine in the next section). OpenFL 1.3 was recently released, which can run OpenFL within an Intel SGX enclave. OpenFL helps institutions to collaborate and run their models in a federated manner while helping to improve the protection of sensitive information with the help of Intel SGX and Gramine.

Open-Source Library OS for Deploying Applications with Intel SGX

As with many powerful technologies, deploying Intel SGX can be challenging for those not steeped in IT lore, such as data scientists. The Gramine Project (formerly known as Graphene) is an open-source library OS (LibOS) for Linux multi-process applications, with Intel SGX support.

The Gramine Project is now a Confidential Computing Consortium project, of which Intel is a founding member. The Gramine Project provides tools and infrastructure components for running unmodified applications on confidential computing platforms based on Intel SGX. Gramine fast-tracks secure deployment of complex software stacks within Intel SGX by eliminating additional developer effort. It also provides tools for developing end-to-end secure solutions with Intel SGX enclaves that shield proprietary code and sensitive data from hackers, whether the data is in a state of use, in transit or at rest. Using Gramine, organizations seeking additional privacy protection for their FL projects can use Intel SGX more easily.

Intel SGX is the most researched, updated and deployed hardware-based trusted execution environment (TEE) for the data center, and Gramine is one of the few frameworks that supports multi-process applications by providing a complete and secure fork implementation.

Optimized Code with Intel Distribution of OpenVINO Toolkit

To further facilitate use in low-resource environments, we provide a post-training runtime optimized version of the final consensus model. The Intel Distribution of OpenVINO toolkit includes a Model Optimizer, which is a cross-platform command-line tool that facilitates the transition between training and deployment environments, performs static model analysis, and (DL) models for optimal execution on endpoint target devices. Optimizations include reducing the model's size (such as the number of parameters and layers) and speeding models up. Using the Model Optimizer results in an Intermediate Representation (IR), which is then passed to the Inference Engine, where the model undergoes further optimizations based on the target end-device.

> The number of participants in the Penn Medicine project has more than doubled and the brain tumor detection model is up to 33% more accurate than models run on publicly available datasets.

For the Penn Medicine project, using the Intel Distribution of OpenVINO toolkit resulted in up to 4.48x lower latency and 2.29x lower memory utilization, compared to the first consensus model created in 2020.[8] With lower latency and memory utilization, the model can run on edge systems at clinics, instead of requiring large data center resources.

Up To 4.48x Lower Latency

Up To 2.29x Lower Memory Utilization

Substantial Increase in Model Accuracy Has Potential to Improve Patient Prognosis and Quality of Life

More than 300,000 people around the world were diagnosed with a brain tumor in 2020.[9] Typically, the earlier a brain tumor is detected, the more options doctors have to treat it, and surgery can be less extensive—both of which can lead to a more positive medical outcome and a better quality of life for patients. Since the beginning of the Penn Medicine project, the number of participants more than doubled and the brain tumor detection model is now up to 33% more accurate than the original model based on the standard BraTS dataset. The Penn Medicine project has demonstrated how the power of collaboration can transform an idea—secure FL—to reality impactfully.

Learn More

You may find the following resources helpful:

- Intel® Software Guard Extensions
- OpenFL
- Gramine Project
- Intel® Distribution for OpenVINO™ toolkit
- Securing Sensitive Data Using Confidential Computing Powered by Intel® SGX video
- A Path Towards Secure Federated Learning article
- Confidential Computing with Gramine blog
- Confidential Computing Consortium Announces Gramine 1.0 press release
- Federated Learning through Revolutionary Technology white paper
- Federated Learning for AI Analytics eBook
- Optimized Edge Analytics using Intel® Distribution of OpenVINO™ toolkit white paper
- Confidential Computing Consortium open-source community

Learn more about **Open Framework for Federated Learning**.

intel®

[1] arXiv, "Federated Learning Enables Big Data for Rare Cancer Boundary Detection," https://arxiv.org/abs/2204.10836

[2] Ice Lake launch presentation by Lisa Spelman, https://www.intel.com/content/www/us/en/newsroom/news/3rd-gen-intel-xeon-scalable-video.html#gs.xkavmk, which shows 23 participating institutions, compared to a study from arXiv, https://arxiv.org/abs/2204.10836, which shows 55 participating institutions.

[3] See endnote 1.

[4] Penn Medicine's work is funded by the Informatics Technology for Cancer Research (ITCR) program of the National Cancer Institute (NCI) of the National Institutes of Health (NIH), through a three-year, $1.2 million grant awarded to principal investigator Dr. Spyridon Bakas at the Center for Biomedical Image Computing and Analytics (CBICA) of the University of Pennsylvania.

[5] See endnote 1.

[6] Mark R. Gilbert et al. "RTOG 0825: Phase III double-blind placebo-controlled trial evaluating bevacizumab (Bev) in patients (Pts) with newly diagnosed glioblastoma (GBM)." 2013. https://ascopubs.org/doi/abs/10.1200/jco.2013.31.18_suppl.1

[7] See endnote 1.

[8] The optimization of the consensus model inference workload was performed via OpenVINO99 (https://github.com/openvinotoolkit/openvino/tree/2021.4.1), which is an open-source toolkit enabling acceleration of neural network models through various optimization techniques. The optimizations were evaluated on an Intel® Core™ i7-1185G7E processor @ 2.80 GHz with 2x 8 GB DDR4 3200 MHz memory on Ubuntu 18.04.6 OS and Linux kernel version 5.9.0-050900-generic.

[9] ASCO, "Brain Tumor: Statistics," https://www.cancer.net/cancer-types/brain-tumor/statistics