Seminar Talk: "Enterprise Federated Learning: Challenges and Solutions" (Speaker: Dr. Dinesh Verma)

Matthew Whitesides

Abstract

Federated learning is a decentralized computational learning technique that utilizes multiple devices without needing to exchange data. In today's presentation, Dr. Dinesh Verma discusses the various challenges and benefits when implementing federated learning approaches in large-scale enterprise deployments.

I. Introduction and Background

FEDERATED machine learning is a collaborative training approach where training data is not exchanged to overcome constraints on training data sharing (policy, security, coalition constraints). Also, this approach sidesteps insufficient network capacity prevalent at the network's edge nodes as opposed to standard machine learning approaches that require centralizing the training data on one machine or network datacenter. Traditional machine learning approaches consolidate the training data from various nodes to one central node, usually a datacenter. On the other hand, Federated Learning enables nodes to collaboratively learn from a shared learning model while keeping all the training data on the individual nodes, decoupling the ability to do machine learning from the need to store the data in a centralized node.

At a high level, the federated learning process works as follows, each node downloads the latest model, improves upon it with the latest local data, and sends the updated model back to the central node. The central node then utilizes each node's individually improved models to create a new generalized model that is sent back to the individual nodes creating an iterative improvement feedback loop.

Typical federated learning research has focused on consumer edge devices such as smartphones or internet of things (IoT) devices. This use case has various challenges, such as data privacy and limited computing capabilities. However, the use case Dr. Verma has focused on here is federated learning among enterprise site computation. In the typical enterprise scenario, you have many individual sites with their own data centers and large datasets that need to combine information across those sites. These large sites lead to massive amounts of computing and data that would not be feasible to aggregate at a single location. Instead, each site computes its models and can fuse those models intelligently at the central data center.

II. Research Contributions and Results

In enterprise-scale approaches to federated learning, many challenges are presented across the various organizations and technologies used. For example, across datasets, there will exist many different methods of naming classes and features. A solution to this is auto-generation of feature mapping by detecting similar feature maps across datasets.

Both consumer and enterprise federated learning comes with various shared and unique challenges. Consumers need data privacy on a relatively small dataset across thousands of devices, leading to some of the users having malicious intent. While enterprise, on the other hand, has fewer sites with larger datasets, each site having its policies and systems, making it harder to centralize and synchronize, and it is very costly to migrate or even maintain regulatory concerns across sites. However, the benefits of federated learning apply to both scenarios. Due to data not needing to be sent to a centralized site, privacy and security are improved, and all nodes in the chain can benefit from merging models across the sites or nodes.

Another challenge is the quality of data in various sets. In this, you'll need to be able to drop data from the datasets that do not meet a quality threshold. Different sites and organizations may use other learning functions as well. For example, averaging different functions could lead to different results, and combining those does not lead to final quality models.

With the number of sites you are collecting data from, there are different ways you can split up the data to build the model. Do you randomly take a sample of data between the sites? As mentioned, other sites may have different features or functions utilized to calculated that data. You may need to take a more informed approach and combine models and datasets that continue the feature properties you require. This leads to a much better quality of model but requires more domain knowledge and time.

Synchronization between sites leads to the situation to take into account. Say different sites have various models with other functions used to build those models. They must share models across the enterprise, which often leads to out-of-date models being used by different sites and becoming out of sync with their desired model. A proposed solution is to use a "sketch" of the models with the desired architecture that all sites use as a template. These sites can then send their trained models to a generator that distributes the models across sites.

Take a example retail chain organization comprised of a large number of individual stores and corporate offices across the world. Each store is concerned with data confidentiality but wants to analyze buying patterns across the organization. A centralized approach would have each store send data to a data center and train the model there. A federated approach would train smaller models at each store, send those models to the data center, and there the models would be fused into the final model. This federated approach, while sometimes generates a different model than the centralized one, has shown to be comparably affected in model accuracy as the centralized approach. This approach also benefits from added security. Each site does not need to send sensitive data outside their local network, only their trained models, which do not contain concrete consumer data.

Depending on the business problem attempting to solve this final model, you may desire each site to use the same fused model. The combined model will want to apply to the site requesting it; this leads to the model generator building multiple fused models for different business problem categories and policies. You can utilize different feature mapping to generate these other models and prioritize various site models or features to benefit from incorporating all the models but still have the unique sites results. Or your business process may benefit more from each site having a very generalized model across sites, regardless of whether your federated model build should approach this based upon business needs.

III. Lessons Learned

Federated learning is an appealing application of exciting technology. I appreciate Dr. Verma's focus on a large enterprise-scale, not the typical consumer small edge devices but large datacenters combining their data to create even more valuable models. Working in a large organization, I see daily the challenges in sharing and incorporating improvements made by teams, even sitting across the room, let alone at different sites across the organization. Utilizing an informed architected approach like those laid out in this presentation would allow an automated or process-driven implementation of this idea. Each site could benefit from the data and improvements made by other sites without needing to share data or sanitize the data sent if it is sensitive. We could upload our newer models, and at the centralized data center, intelligent fusing could be done to improve our model with data across the organization.

While this approach is an excellent idea and would give massive benefits to many companies, it would also be a considerable challenge to incorporate. I would foresee a company would almost have to build from the ground up incorporating this approach. Each site would need to agree on the template proposed by the centralized data center and have relatively similar dataset features. As mentioned in the presentation incorporating this into existing sites with their unique datasets is a big challenge. Some solutions to intelligently selecting and building models of the individual datasets were proposed. To truly gain the benefits of an idea like this, a large amount of corporation and rework would need to be done in most companies. Regardless of the challenge in architecting a system like this across an enterprise, the upside is massive. Fully utilizing data and improvements across the enterprise is something I'm sure most companies would spare no expense to achieve. In particular, this makes sense for companies such as IBM, a massive company that has a primary focus on data science and enterprise solutions. The benefits to this kind of implementation could be in the billions of dollars.

IV. Conclusion

In the end, Dr. Verma has presented an excellent overview of the unique qualities and challenges of implementing a federated learning approach to machine learning model generation at an enterprise scale. While the challenge is considerable to implementing federated learning at this scale, I feel some of the proposed solutions presented in this talk should excite anyone looking to incorporate the quality and security benefits that a federated learning approach could bring.

Acknowledgment

The author would like to thank Professor Sajal Das with the Department of Computer Science, Missouri University of Science and Technology and Dr. Dinesh Verma with the IBM Research.

Matthew Whitesides Master's Student at Missouri University of Science and Technology.