Azra Avci
MVWWJQ

Evaluating individual Participant Contributions Using SecAgg

Abstract

In Federated Learning (FL), it can be challenging to meet the need for both efficient evaluation of participants and ensuring participant privacy simultaneously. This project implements Leave-One-Out (L1O) and Include-One-In(I1I) scores using the Flower framework in order to provide an evaluation of FL participants in a privacy-preserving way by using Secure Aggregation(SecAgg). This project provides L1O and I1I score computation as well as giving the option to choose between global and local evaluation modes. The achieved results has shown this approach to provide an effective way of calculating accurate evaluation metrics while maintaining the privacy of participants.

Azra Avci
MVWWJQ

1. Introduction

Motivation

Federated Learning (FL) is a powerful paradigm for training machine learning models across decentralized data sources. One notable challenge of it is evaluating the contributions of individual FL participants while preserving their privacy. Moreover, this challenge is particularly important in areas where data is sensitive and privacy is crucial such as healthcare. Being able to evaluate individual contributions without compromising privacy is important in such fields.

FL makes organizations training models together without sharing raw data possible. This allows for easier compliance with data protection regulations on well as preserving privacy. This decentralization makes individual participant contribution computation challenging especially when privacy is a concern.

Problem definition

The goal of this project is to solve the evaluation of individual FL participants' contributions without compromising their privacy. Other methods such as the Shapley value are computationally infeasible in a FL setting. This is not only due to its exponential complexity but also in order to perform the calculation, it needs gradient sharing. This does not correspond with privacy preserving protocols like SecAgg. Since the use of SecAgg prevents direct access to individual gradients, alternative evaluation metrics that correspond to the privacy level of SecAgg are needed.

Challenges

Main challenges are:

1. **The computational complexity of Shapley value:** Since the calculation of the Shapley value needs all possible subsets of participants, this gets computationally impossible especially for larger numbers of participants.
2. **Privacy concerns with gradient sharing:** Direct sharing of gradients can expose sensitive information about a participant's data.

3. **Integration within the Flower framework:** Implementing a privacy-preserving evaluation program with an existing and actively developed FL framework like Flower needs careful integration.

Solution

Using L1O and I1I scores as approximations to the Shapley Value while also implementing SecAgg. These custom metrics can be computed within the Flower Framework while taking care of participants' privacy needs through SecAgg. L1O measures the effect of excluding a participant while I1I measures the effect of including a participant's update. These scores can be used to evaluate individual contributions without revealing gradients.

Contributions

- **Implementation of I1I computation:** Developed algorithm for computing I1I and integrated into the Flower framework.
- **Implementation of L1O computation:** Developed algorithm for computing L1O and integrated into the Flower framework.
- **Development of global and local evaluation modes:** Implemented different evaluation modes that can be configured from the configuration file to evaluate contributions using both shared global test sets and individual local test sets.

2. Related Work

Evaluating FL participants can be done through using the Shapley Value [1], which provides strong insight into individual contributions. However, just like a line in a Pascal triangle, it increases very fast and therefore is exponentially hard to compute. These constraints have limited the practical application of the Shapley Value in FL.

There is an approach that focuses on gradient aggregation and privacy-preserving protocols [2]. This method has the similar goals to that of this project. It has become standard to utilize Federated Averaging (FedAvg) to aggregate updates from participants without sharing raw data. [3] Differential privacy is also an approach that is employed in machine learning in general [4], and also in FL and even in Flower Framework [5]. A

downside to this method is that it often introduces significant noise which affects the accuracy of evaluation metrics.

This project's approach differs from prior works by offering a feasible alternative to the Shapley value. The integration of L1O and I1I scores and maintenance of privacy through SecAgg gives a practical and efficient option for FL participant evaluation. By eliminating the need to share individual gradients, it provides custom evaluation metrics while preserving privacy. The balance between accuracy, computational feasibility and privacy of this approach makes it a good option for real world FL deployments.

3. Background

Federated Learning

FL provides a way for decentralized data sources to train a model together without sharing raw data. This approach provides more privacy and utilizes diverse datasets. Multiple participants each with their own local data collectively train a global model by sharing updates with a central server.

One use case of FL is scenarios where data is not and cannot be centralized due to various concerns such as privacy, security and concerning regulations. The local training and central aggregation structure allow for sensitive data to remain on the participant's device. The ability to use more diverse data sources has the potential to lead to more robust models.

Secure Aggregation (SecAgg)

SecAgg makes sure individual participants' updates are aggregated without revealing their gradients. By encrypting participants' updates before aggregation, it prevents inferring individual data impossible.

The SecAgg protocol's workflow allows for the server to compute the weighted average of model parameters across all participants while ensuring the individual contributions remain private [6]. The approach is such that clients send two versions of the locally updated parameters, and they are both masked for privacy. In order to compute the weighted average of model parameters, server aggregates these contributions.

Shapley value

The Shapley value comes from cooperative game theory: it is a method that fairly distributes total gains to participants based on specific contributions they made. Yet, it is exponentially hard to compute, thus it is not a practical option for larger FL scenarios. Not only that but also, as the computation needs to evaluate all subsets this goes against privacy concerns. L1O and I1I are alternative options.

L1O and I1I scores

L1O and I1I scores offer approximations to the Shapley value by evaluating the two subsets the calculation of which do not violate privacy concerns. L1O measures the global model without the contribution of the participant, while I1I measures the impact of including a participant's update. These custom metrics offer efficient computation and reflect individual contributions without compromising privacy.

4. Model

System model

This system assumes a Federated Learning (FL) setting in which participants can access local training data. The MNIST dataset is used to train participants. The MNIST dataset is a widely used dataset used for machine learning. It is structured into training and testing sets thus is appropriate for evaluating machine learning algorithms. [1] In this FL setting, the data is partitioned among participants. The participants then train their local models, after which they share their updates with the server in an encrypted manner.

In this federated learning environment, the server coordinates the training by sending the global model to participants and aggregating their updates. Local models are trained on respective datasets by the clients who then compute their contribution scores without sharing raw data.

In the absence of secure aggregation methods, there exists a threat to the privacy of the participant's data. If the updates were sent without encryption, the data of participants can be inferred from the shared gradients. SecAgg mitigates this threat through the aggregation of encrypted updates. This implementation ensures the individual contributions of participants stay private even if the aggregated updates were accessed by an adversary.

5. Solution

The approach this project uses computes L1O and I1I scores using the Flower framework for simulating a FL environment and utilizes the SecAgg mode of Flower framework for privacy reasons. The program trains local models, computes their evaluation scores and aggregates updates to create a global model. The important steps are:

1. **Initialization:** The global model is initialized and shared with the participants by the server.
2. **Local training and I1I computation:** The model is trained locally, the I1I score is computed, then encrypted updates are shared with the server by participants.
3. **Global model update:** The updates are aggregated by the server to form a new global model.
4. **L1O computation:** The L1O score is computed based on the updated global model by participants.
5. **Evaluation:** The participants evaluate the models using global or local test sets, these are two evaluation modes the program provides the user with.

The I1I and L1O computations are optional and can be controlled by their respective flags in the configuration file.

Local training and I1I computation

During this process, the participants receive the global model and train it locally using local datasets. The I1I score is computed by evaluating the model's performance on the local validation set before and after training. The Include One In score shows how much the model's performance improves as a result of the participant's contribution, providing insight on the impact. The trained model's parameters are encrypted and sent to the server.

Global model update

The server receives the encrypted updates from participants and aggregates them in order to form a new global mode. The individual gradients of the participants are kept private by utilizing SecAgg, only the aggregated result is provided. For the next training round, the updated global model is shared with the participants.

During aggregation, updates from all participants are averaged to form the updated global model. FedAvg strategy offered by the Flower Framework is used in this project. This way the contribution of each client can be proportionately reflected in the updated model.

L1O computation

The Leave-One-Out score is computed through the evaluation of the updated global model and comparison of it with a model that excludes the contribution of the participant. This way, the contribution that can be credited to each client is measured. The global model is evaluated on the local validation set and compared with a model that excludes the contribution of the participant.

Through using L1O score, he importance of the participant in isolation from other participants can be seen through the performance difference. It is an evaluation metric that can be utilized without compromising privacy.

Evaluation modes

There are two evaluation modes implemented in this project:

- **Global evaluation:** Participants have different local training sets but use the same global test set for evaluation. With this mode, the evaluation metrics that are used are consistent across participants.
- **Local evaluation:** Participants use both their different local training and different local test sets for evaluation. With this mode, more practical scenarios in which participants have diverse datasets can be simulated.

Global evaluation provides consistent evaluation of contributions while local evaluation provides insight into the performance of the model in more realistic, diverse environments and data distributions. The choice between two modes depends on the goals of the user.

6. Evaluation

Dataset description

This project uses the MNIST dataset, a widely used dataset that consists of 60,000 training and 10,000 test images. Each image shows a digit from 0 to 9 in handwriting. [1] In order to simulate a realistic federated learning scenario where each participant only has

access to a subset of the data, the MNIST dataset is partitioned according the configuration file.

In the default configuration of the program that is used in all trial runs, the data is partitioned among 10 participants and every participant gets a random subset of the training data.

Baselines

A traditional federated learning environment without the computation of L1O and I1I scores and no secure aggregation mode can be compared to the approach of this project. In the baseline the global model is trained without evaluating individual contributions and the updates are not encrypted. The comparison between two approaches shows the value of this project's evaluation metrics and privacy approach while using federated learning.

Parameter settings

In the experiments, the following parameters are used:

- Learning rate: 0.01
- Momentum: 0.9
- Batch size: 20
- Number of local epochs: 1
- Number of clients: 10
- Number of rounds: 5

Balancing training efficiency and model performance was considered when the parameters were set. The values of the learning rate and momentum were chosen based on standard practices in training convolutional neural networks (CNNs) on the MNIST dataset.

Evaluation metrics

These were the metrics used for evaluation:

- **Accuracy**: The proportion of correctly classified images.
- **Loss**: The cross-entropy loss between the predicted and true labels.
- **L1O and I1I scores**: Custom evaluation metrics for individual contribution calculations.

While accuracy and loss demonstrate overall model performance, the L1O and I1I scores demonstrate individual contributions in addition to that.

Results

The experiments and result indicate that the approach taken in this project is able to evaluate individual contributions of participants while maintaining their privacy. The global evaluation mode gives an option that ensures consistency across participants and the local evaluation mode gives an option that is more representative of realistic scenarios with diverse datasets.

Results from example runs:

- **Global evaluation results**
    - Initial accuracy: 0.0957
    - Final accuracy: 0.9803
    - L1O scores ranged from -0.0089 to 0.0086
    - I1I scores ranged from -0.0167 to 0.795
- **Local evaluation results:**
    - Initial accuracy: 0.1032
    - Final accuracy: 0.9795
    - L1O scores ranged from -0.0013 to 0.0137
    - I1I scores ranged from -0.0017 to 0.8283

Initial low accuracy is as expected from an untrained model at the start. The significant improvement in the final accuracy demonstrates that the global model showed high performance on the MNIST dataset after being trained with FL.

As for the L1O scores, the overall range is minimal. This suggest that the impact of any participant is relatively small, which is expected from a balanced dataset like MNIST. A negative score shows that a participant had a minimal negative impact on the performance of the overall model. Positive scores show the participant had a positive contribution and enhanced the performance of the model.

I1I scores show the improvement in accuracy when a participant's update is included. There is a wide range here which reflects that some participants had contributions that

significantly improved the performance of the model while others had negligible or even slightly negative effects.

For the local evaluation, L1O scores have a higher positive end which might be attributed to some participants' data being comparatively more important for improving the model than others.

7. Conclusion

This project implements the computation of L1O and I1I scores using Flower Framework and utilizes SecAgg mode in order to evaluate the contribution of participants in FL while also preserving their privacy. While not as precise as some other evaluation methods, this approach to evaluating individual contributions is a privacy preserving option with achievable computability. The results show that these are reliable evaluation metrics.

In conclusion, this project's approach gives a practical solution for evaluating federated learning participants while maintaining their privacy. It also offers global and local evaluation modes, each of which is appropriate for different scenarios. The L1O and I1I scores are shown to be reflecting reliable data about individual contributions.

Future work

The implementation should be thoroughly tested with diverse datasets and under different conditions.

The output of evaluation metrics should be structured further.

8. Sources

[1]: https://medium.com/the-modern-scientist/what-is-the-shapley-value-8ca624274d5a

[2]: https://arxiv.org/abs/1611.04482

[3]: https://arxiv.org/abs/1602.056294:

[4]: https://arxiv.org/abs/1607.00133

[5]: https://flower.ai/docs/framework/how-to-use-differential-privacy.html

[6]: https://flower.ai/docs/framework/ref-api/flwr.server.workflow.SecAggPlusWorkflow.html

[7]: https://medium.com/@binaya.puri/mnist-dataset-for-machine-learning-8987e0b20bb3