Evaluating differential privacy on language processing federated learning

Q.M.F. Van Opstal[1]

[1] Q.M.F.vanopstal-1@student.tudelft.nl

Final presentation, 2024

1 Introduction

- **Federated learning** allows a central server to train a machine learning model on different machines (clients) while keeping the training data used by the clients private [3]. A model is sent out to different clients, they train this model and send their updates back, these are the aggravated. In this project, it is used to train a natural language classifier.
- There is a threat against this kind of learning: a **backdoor attack**. Here, an adversary tries to make the model output a chosen label when a specific input is presented, without disrupting the general task when the input is not present [2]. A specific case which is used in this project is edge case attacks here, the input that triggers the malicious behavior is sparely present in the genuine training data [5].
- An existing defense against this is **differential privacy** [5], it works by clipping the updates received when they exceed a given threshold and then adding Gaussian noise to the aggravated model. The parameters that influence the performance of differential privacy are the threshold used, and the standard deviation used to add the Gaussian noise. When the standard deviation is relatively small, this defense is called weakDP [4].
- The data provided to the clients can be i.i.d. (independent identically distributed) here every client gets about the same amount of samples, and the classes of their targets are about evenly distributed. With **non-i.i.d.** both the amount of samples as in what class these samples lay is more unevenly distributed. Real life data is often non-i.i.d. [1]. This non-i.i.d. can be achieved with a heterogeneous Dirichlet distribution [6], a parameter β decides how uneven the data is distributed.

Research Question

How can Weak Differential Privacy provide a defense against backdoor attacks on a language processing federated learning model that is trained with non-i.i.d.?

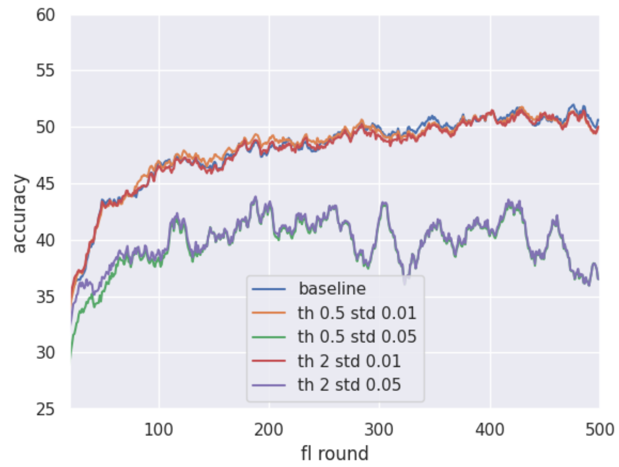
2 method

Execute a backdoor attack on a federated learning model training a natural language processing classifier with weakDP as its defense, with different parameters for the distribution of data and weakDP. All permutations of these parameters were tried:

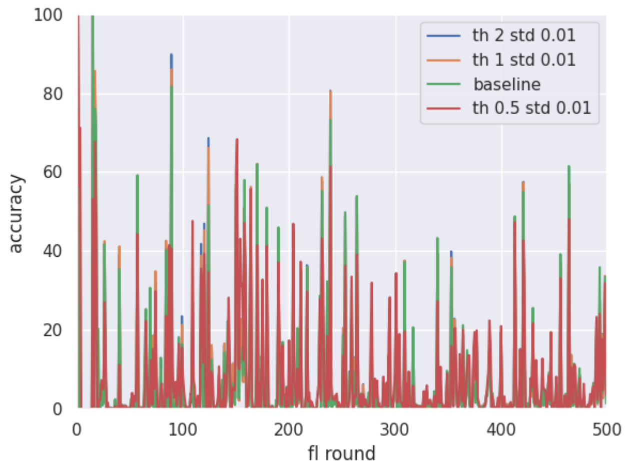
- β for a heterogeneous Dirichlet distribution: 0.5, 1, and 2
- A threshold of 0.5, 1 and 2
- A standard deviation of 0.01 and 0.05

The results were evaluated based on the main task performance and backdoor performance compared to each other and a model trained with no defense.

3 Results



Main task accuracy of running weakdp with threshold (th) of 0.5 and 2 in combination with standard deviation (std) of 0.01 and 0.05, $\beta = 2$ was used in the heterogeneous Dirichlet distribution. A rolling mean of 20 was used to improve the visibility of the results.



Backdoor accuracy comparison between no defense, weakDP with a threshold of 2, weakDP with a threshold of 1, and weakDP with a threshold of 0.5. A standard deviation of 0.01 was used for both weakDP tries, $\beta = 0.5$ was used in the heterogeneous Dirichlet distribution.

4 Conclusion

In conclusion, on this specific dataset [1] with the non-i.i.d. training of the specific model we used, a threshold of 0.5 and a standard deviation of 0.01 worked the best. **Future work** Further research is need to find out if these values are also the best when a different dataset or model is used. A different distribution might also affect the results.

Gitlab

The code used can be found in this github: `https://github.com/QuintenVanOpstal/OOD_Federated_Learning.git`.

References

[1] M. Adnan, S. Kalra, J. C. Cresswell, G. W. Taylor, and H. R. Tizhoosh. "Federated learning and differential privacy for medical image analysis". In: *Scientific reports* 12.1 (2022), p. 1953.

[2] L. Lyu, H. Yu, and Q. Yang. "Threats to Federated Learning: A Survey". In: *CoRR* abs/2003.02133 (2020). arXiv: 2003.02133.

[3] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. "Communication-efficient learning of deep networks from decentralized data". In: *Artificial intelligence and statistics*. PMLR. 2017, pp. 1273–1282.

[4] Z. Sun, P. Kairouz, A. T. Suresh, and H. B. McMahan. *Can You Really Backdoor Federated Learning?* 2019. arXiv: 1911.07963 [cs.LG].

[5] H. Wang, K. Sreenivasan, S. Rajput, H. Vishwakarma, S. Agarwal, J.-y. Sohn, K. Lee, and D. Papailiopoulos. "Attack of the Tails: Yes, You Really Can Backdoor Federated Learning". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 16070–16084.

[6] M. Yurochkin, M. Agarwal, S. Ghosh, K. Greenewald, T. N. Hoang, and Y. Khazaeni. *Bayesian Nonparametric Federated Learning of Neural Networks*. 2019. arXiv: 1905.12022 [stat.ML].

TU Delft Delft University of Technology