Privacy in Machine Learning

Project Report

Generating Differential Private Synthetic Data

By:
Zahra Bashir, Mahtab Farrokh, Fatemeh Tavakoli

Course Instructor:
Prof. Nidhi Hegde

April 24, 2021

Contents

List of Figures

1 Introduction

Machine learning is a widely-used tool for many applications, but one of the essential concerns is having appropriate training data. When competitions want to share their datasets, some privacy issues may arise, such as revealing individual-related information. One approach is to make a synthetic dataset that is close to the real dataset. To generate this synthetic dataset, we can use a Generative Adversarial Network (GAN) [9] model; however, it is non-private due to information leakage, so it will reveal some information about the real dataset [5]. One way to avoid this problem is to use differential privacy [8]. Several approaches have been developed to preserve the training samples' privacy by applying differential privacy in GAN training. Hence, the synthetic generated data would be differentially private. Some of the recent and prominent approaches for differentially private GANs that we will work on are DPGAN, DPCGAN, SPRINT-gan, and PATE-GAN. This project compares the mentioned approaches using the Banknote Authentication Dataset [7], and Credit Card Fraud Detection [6] datasets which are from two different categories. Furthermore, we introduce another method that applies differential privacy with PATE [11] method on ACGAN [10].

2 Related Works

Generative Adversarial Network (GAN) is an excellent tool for generating synthetic data with respect to the real data [9]. However, traditional GAN provides no guarantee on what the synthetic data reveals about the real data. One way to make synthetic data private is to use differential privacy [1]. In this project, we investigated SPRINT-GAN, PATE-GAN, DPGAN, and DPCGAN.

The first introduced differentially private GAN was DPGAN [14] which uses the differentially private stochastic gradient descent method (DPSGD) [1] for training the discriminator in a Wasserstein GAN (WGAN) [2]. They introduced a gradient descent method that limits each training sample's influence by clipping and adding noise to guarantee differential privacy in networks.

Following the previous work, [13] used the same training method (DPSGD) to train a GAN for generating synthetic data, which is conditional on given labels. For this purpose, they used a Conditional GAN instead of a simple GAN.

Differential privacy and its application in Neural Networks and specifically Deep Learning has been reviewed in different papers such as ”Deep Learning with Differential Privacy” [1] which studied a gradient clipping approach to make our training procedure private. ”Deep Learning with Differential Privacy”[1] introduced a Private Aggregation of Teacher Ensembles (PATE) for data privacy which was applicable to any semi-supervised model. It aggregated the Laplacian mechanism with a machine teaching Framework. However, this approach suffers from intolerance of privacy loss because the volume of labeled data in the public dataset affects the privacy loss. DPGAN solves this problem by working on the generator part to make it differentially private, generating lots of data points while preserving the privacy of training data.

SPRINT-GAN [3] trains an Auxiliary classifier gan (AC-GAN) [10] under differential privacy and follows DPSGD for training the discriminator. SPRINT-GAN limits the effect of each data instance by clipping the norm of the discriminator's training gradient and adding proportionate Gaussian noise. SPRINT-GAN uses ACGAN on SPRINT clinical dataset [7], and provides summary statistics, and studies the correlation between variables in the real and simulated data.

In PATE-GAN [15], they have used the methods in ”Semi-Supervised Knowledge Transfer For Deep Learning From Private Training Data” [12] and ”Scalable Private Learning with PATE” [11], but modified the PATE framework to be used explicitly for a GAN. As opposed to the method in [12], PATE-GAN trains the student discriminator while it does not need to have access to the original dataset.

3 Methodology

In this project, we trained DPGAN, DPCGAN, SPRINT-GAN, and PATE-GAN on two datasets to evaluate and compare their results. We trained all of the models on the Banknote Authentication Dataset [7], and the Credit Card Fraud Detection Dataset[6]. We also proposed a new method PATE-ACGAN, which applies the PATE idea to an Auxiliary Conditional GAN (ACGAN) model. Up to our knowledge, the PATE method has not been applied to the Auxiliary Classifier GANs before. In the end, we compared our proposed method to the state-of-the-art methods. The PATE-ACGAN architecture is shown in Figures 1 and 2.

In this method, we use ACGAN for the Generator, Teachers, and Students models. In ACGAN, the generator gets a class label C along with a noise vector; the class label C will be produced given the class ratio of the real dataset. Then, given latent noise and a class label, the generator's output will be the synthetic data. We have N teacher discriminators that, given the generated data, will predict the data class label and whether the input data is fake or real. Given the teachers' prediction, we compute teachers' loss using binary cross-entropy. Each teacher's loss is the sum of the class label loss and the real/fake loss. Up to this step is shown in Figure 1. Teacher discriminators try to minimize the classification loss. It is to be noted that during this step only the parameters of the teachers are getting updated (not the generator).

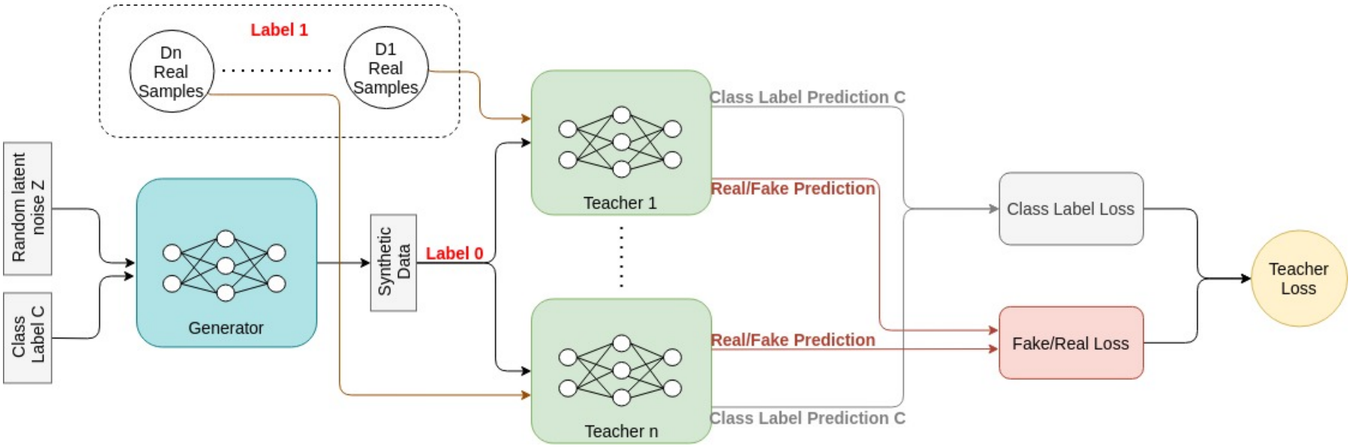

Figure 1: Block diagram of the training procedure for the teacher discriminator.

In the next part of our method, like baseline PATE-GAN, the student discriminator is trained using noisy teacher-labeled generated samples (the noise provides the DP guarantees). Moreover, the student will use an ACGAN structure. The student output is the class label C and the real/fake prediction. The student is trained to minimize classification loss on this noisily labeled dataset, while the generator is trained to maximize the student loss. Notably, the teachers are not updated during this step, only the student and the generator.[15]

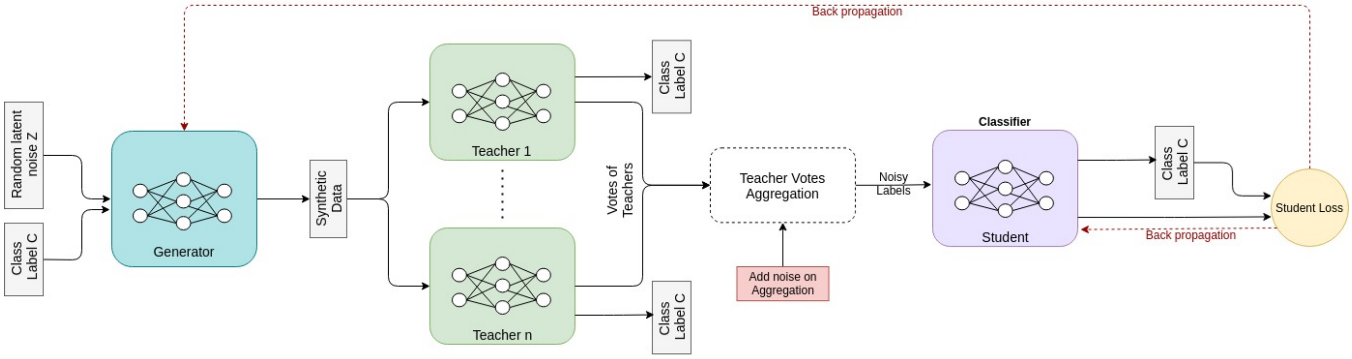

Figure 2: Block diagram of the training procedure for the student discriminator and the generator.

The proposed PATE-ACGAN generates a private dataset because it is using the same privacy approach proposed in the PATE-GAN paper.[15] Our model assures privacy by adding noise to the teacher votes and aggregating the noisily votes. The student discriminator does not have access to the real dataset and is only affected by the noisy teacher votes. So, this approach guarantees to generate a differentially private dataset.

4 Evaluation & Results

For evaluating the similarity of synthetic and real data, we applied the method used in [15] paper, which compares the performance of models trained on the synthetic dataset and tested on the real dataset on a prediction task (classification/regression). So, we can make sure that it can be used instead of the real sensitive datasets on machine learning tasks, and therefore, we can share it publicly.[4]

To evaluate our models, we used the Credit Card Fraud Detection dataset. This dataset is imbalanced and has 492 positive samples among 28k instances(0.173%). The low number of positive examples is due to the fact that fraud in bank transactions happens rarely. When we trained a simple Neural Network on this dataset, the accuracy result was about 99.9%.

When we trained our GAN models, the Neural Network results trained on the generated private data were not satisfying. The test accuracy on private data was about 97%. However, other metrics such as precision, recall, and f1-score were zero. The reason is that the true positive value was zero, and the model only learned to predict zero. So, we tried doing undersampling and made a new dataset consisting of 492 positive instances and 1500 negative instances. Though, we think, due to the low number of data instances, our GAN models could not learn the whole data distribution and failed to achieve satisfying results.

Therefore, we decided to work on a balanced dataset. Banknote Authentication Dataset has four features and has 1372 instances .Labels are 0 for authentic and 1 for forgery, So it is a binary prediction task. There is 55% negative and 45% positive instances. Therefore, this datset is balanced. For evaluating the models, we splitted the real dataset (”Banknote Authentication Dataset”) to the test (30% of the whole dataset) and train dataset. We trained the GANs with the training set. Each of the GANs generated a differentially private dataset. Then, we trained a simple Neural Network with four layers on the synthetic datasets and tested the Neural Network model on the real test set (30% of the Banknote Authentication Dataset).

To be noted, the GAN models are not trained on the real test dataset. The Neural Network accuracy was 96% using the real data for both training and testing. Among previous works, the best result was 0.79 and was achieved by DPCGAN. The PATE-ACGAN method had a better result of 83%, and increased the accuracy by 4%. A summary of accuracy results is shown in Figure 3. These results are averaged over multiple runs (10 runs).

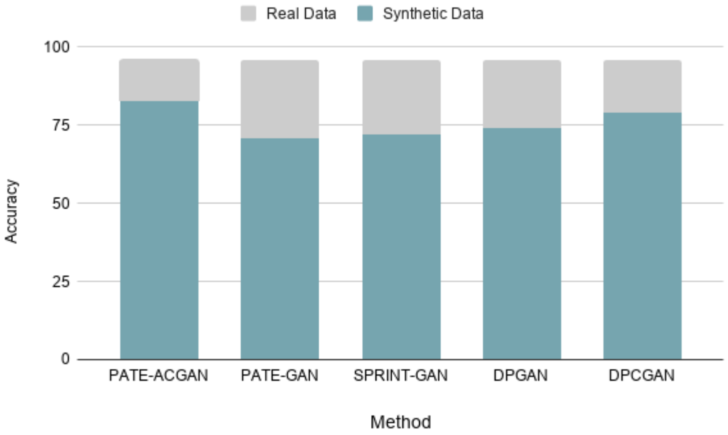

Figure 3: Accuracy of the Neural Network trained and tested on the real dataset vs. trained on the synthetic dataset(generated by different methods) and tested on the real dataset.

Furthermore, we evaluated our models using f1-score, precision, and recall metrics to measure the ability of each model in label prediction. Results are shown in the following table.

Evaluation on Banknote Authentication Dataset						
Metrics	DPGAN	DPCGAN	PATE-GAN	SPRINT-GAN	PATE-ACGAN	Real Data
Accuracy	0.74	0.79	0.71	0.72	0.83	0.96
Loss	0.83	0.62	1.17	4.65	1.97	0.12
F1-score	0.66	0.80	0.52	0.58	0.79	0.95
Precision	0.85	0.69	0.90	0.89	0.93	0.95
Recall	0.55	0.98	0.38	0.43	0.70	0.97

As shown in the table, our new proposed method could outperform other models in some metrics. In other words, we got the highest precision(0.93%) and accuracy(0.83%) for the PATE-ACGAN model. The f1-score(0.79%) for the PATE-ACGAN was very close to the highest existing f1-score, which belongs to DPCGAN(0.80%). We think our model outperforms other models(especially the PATE-GAN) because it is considering labels in the generation phase. In general, models like

CGAN and ACGAN are expected to perform better because they consider labels. That is why we observed result improvement in PATE-ACGAN and DPCGAN comparing to PATE-GAN and DPGAN respectively.

5 Conclusion

To conclude, in this project, we trained and compared five proposed differentially private GANs naming DPGAN, DPCGAN, SPRINT-GAN, PATE-GAN, and PATE-ACGAN on two datasets (Banknote Authentication Dataset and Credit Card Fraud Detection) and tested their performance according to several metrics. As our contribution, we applied the PATE method on ACGAN (PATE-ACGAN), which has not been suggested before up to our knowledge. In the evaluations, we observed improved precision and accuracy on the synthetic data generated by PATE-ACGAN.

6 Future Works

Due to time constraints, we could not execute our experiments with sufficient different settings. Approaches like PATE need more hyperparameter tuning because there are a lot of attributes like number of teachers, number of each students' and teacher's iterations, etc, that can be manipulated and tested. Although we modified the hyperparameters several times, we think there is more hyperparameter tuning needed and we may reach better results for all the models.

References

[1] Martin Abadi et al. "Deep Learning with Differential Privacy". In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (Oct. 2016). DOI: 10.1145/2976749.2978318. URL: http://dx.doi.org/10.1145/2976749.2978318.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. *Wasserstein GAN*. 2017. arXiv: 1701.07875 [stat.ML].

[3] Brett K. Beaulieu-Jones et al. *Privacy-preserving generative deep neural networks support clinical data sharing*. 2017. DOI: 10.1101/159756. eprint: https://www.biorxiv.org/content/early/2017/07/05/159756.full.pdf. URL: https://www.biorxiv.org/content/early/2017/07/05/159756.

[4] Andrew P. Bradley. *The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms*. USA, July 1997. DOI: 10.1016/S0031-3203(96)00142-2. URL: https://doi.org/10.1016/S0031-3203(96)00142-2.

[5] Dingfan Chen et al. "GAN-Leaks: A Taxonomy of Membership Inference Attacks against GANs". In: *CoRR* abs/1909.03935 (2019). arXiv: 1909.03935. URL: http://arxiv.org/abs/1909.03935.

[6] *Credit Card Fraud Detection*. 2018. URL: https://www.kaggle.com/mlg-ulb/creditcardfraud.

[7] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. 2017. URL: http://archive.ics.uci.edu/ml.

[8] Cynthia Dwork. "Differential Privacy". In: *33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006)*. Vol. 4052. Lecture Notes in Computer Science. July 2006, pp. 1–12. ISBN: 3-540-35907-9.

[9] Ian J. Goodfellow et al. *Generative Adversarial Networks*. 2014. arXiv: 1406.2661 [stat.ML].

[10] Augustus Odena, Christopher Olah, and Jonathon Shlens. *Conditional Image Synthesis with Auxiliary Classifier GANs*. Ed. by Doina Precup and Yee Whye Teh. International Convention Centre, Sydney, Australia, June 2017. URL: `http://proceedings.mlr.press/v70/odena17a.html`.

[11] Nicolas Papernot et al. *Scalable Private Learning with PATE*. 2018. arXiv: `1802.08908 [stat.ML]`.

[12] Nicolas Papernot et al. *Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data*. 2017. arXiv: `1610.05755 [stat.ML]`.

[13] Reihaneh Torkzadehmahani, Peter Kairouz, and Benedict Paten. *DP-CGAN: Differentially Private Synthetic Data and Label Generation*. 2020. arXiv: `2001.09700 [cs.LG]`.

[14] Liyang Xie et al. *Differentially Private Generative Adversarial Network*. 2018. arXiv: `1802.06739 [cs.LG]`.

[15] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. "PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees". In: *International Conference on Learning Representations*. 2019. URL: `https://openreview.net/forum?id=S1zk9iRqF7`.

Appendix

Work Distribution

- Literature Review: Everyone
- Data Preprocssing: Everyone
- Implementing DPGAN: Fatemeh
- Implementing DPCGAN: Fatemeh
- Implementing SPRINT-GAN: Mahtab
- Implementing PATE-GAN: Zahra
- Implementing PATE-ACGAN: Everyone
- Evaluation: Everyone
- Result analysis and Visualization: Everyone
- Debugging the codes: Zahra, Mahtab
- Writing Reports: Everyone

We all three spend the same time amount on the project. Depending on the challenges we faced, some tasks took a longer time.

Github Link:

`https://github.com/zahrabashir98/Generating-Differential-Private-Synthetic-Data`