EECS 6322 Week 13, Paper 1
PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees

Shivani Sheth (Student #: 218011783)
shivs29@yorku.ca

I. Summary and contributions

The paper proposes a new framework, known as PATE-GAN, which ensures the privacy of the generator of the Generative Adversarial Nets (GAN) framework that can be used to generate synthetic data on which algorithms can be trained and validated, without compromising the privacy of the original dataset. The proposed framework modifies the Private Aggregation of Teacher Ensembles (PATE) framework and combines with Generative Adversarial Nets (GANs) to tightly bound the influence of individual samples on the model, thus resulting in tight differential privacy guarantees. The paper also proposes a new evaluation metric to measure the performance of the framework given by the relative performance of the trained and tested algorithms on the synthetic dataset with respect to the performance of the two algorithms on the original dataset. Finally, PATE-GAN is compared against the state of the art models in terms of synthetic data quality where it consistently outperforms its counterparts.

Differential privacy is required in machine learning since most large datasets include sensitive information that prevents data-holders from sharing the data. A naive way to resolve the issue would be to de-identify the records, but it is proved that it is easy to re-identify the data by linking them to other identifiable datasets. For developing and validating machine learning methods for a given task, real data is not necessary and it would suffice to have synthetic data that is sufficiently like the real data. The synthetic data can be used to train models that would be deployed directly on real data, or they can be used to identify the best methods to be applied on the real data. By modifying the training procedure of the discriminator in GANs, to be differentially private using a modified version of the PATE framework, the GAN generator (trained only using the differentially private discriminator), and thus the synthetic data it generates, can be proved to be differentially private.

The proposed model uses the GAN framework, where the discriminator is a differentiable module trained to classify samples as either real or generated. In addition, it uses the PATE framework as a differentially private mechanism for classification by training multiple teacher models on disjoint partitions of the data. To classify a new sample, each teacher model's output is evaluated on the sample and then all outputs are noisily aggregated. A student model is also added that trains on some public unlabelled data, which is labelled using the standard PATE mechanism. Differential privacy as used by the framework can be defined as $P(\mathcal{M}(\mathcal{D})\epsilon\mathcal{S}) \leq e^{\epsilon}\ P(\mathcal{M}(\mathcal{D}')\epsilon\mathcal{S}) + \delta$*, where* $\mathcal{M}$ *is a randomized algorithm which is* (ϵ, δ)*-differentially private if for all* $\mathcal{S} \subset \mathcal{O}$ *and for all neighboring datasets* $\mathcal{D}, \mathcal{D}'$ *the above inequality holds true. The intuition behind differential privacy is that a particular sample's inclusion or exclusion in the dataset should not change the probability of a particular outcome by a large margin.*

The PATE mechanism, which is a core part of the proposed framework, provides a differentially private method for classification, a core component of the GAN framework where the discriminator is a classifier trained to identify whether the samples are real/fake. To build a differentially private classifier, the dataset is divided into k *disjoint subsets and* k *different classifiers are trained separately on the* k *sub-datasets. When a feature vector* $\boldsymbol{x}$ *is input to the model for classification, a differentially private output is given by passing* $\boldsymbol{x}$ *to each of the* k *teachers, and then performing a noisy aggregation of the resulting outputs. This can be represented by* $PATE_\lambda(x) = \underset{j\epsilon[m]}{\text{argmax}}(n_j(x) + Y_j)$ *where* $n_j(x) = |T_i : T_i(x) = j|$ *for* $j = 1, \cdots, m$ *and where* $Y_1, \cdots, Y_m$ *are i.i.d.* $Lap(\lambda)$ *random variables. To make the discriminator differentiable, the PATE extension including a student model is used in the proposed framework. The student model (after training) is free to access both its outputs given inputs and its internal parameters, which makes the model differentially private. The student can also be made differentiable and can be modelled using any classifier, such as a neural net.*

Thus, PATE-GAN replaces the GAN discriminator with a PATE mechanism so that the discriminator is differentially private, but requires the (differentiable) student version to allow back-propagation to the generator. The framework trains the generator, G *to minimize its loss with respect to the student-discriminator. This can be given by:* $L_G(\theta_G; S) = \sum_{j=1}^{n} log(1 - S(G(z_j; \theta_G)))$*. The discriminator is replaced by the PATE mechanism in the framework. Thus, the adversarial training becomes unsymmetrical since the teachers are trained to improve their loss with respect to the generator, but the generator is trained to improve its loss with respect to the student*

S which in turn is being trained to improve its loss with respect to the teachers. The empirical loss of teacher i with weights θ_T^i for a fixed G is given by $L_T^i(\theta_T^i) = -[\sum_{u \epsilon D_i} logT_i(u;\theta_T^i) + \sum_{j=1}^{n} log(1 - T_i(G(z_j);\theta_T^i))]$ where each teacher is trained in the same way as the discriminator in a standard GAN framework, except that here the teacher only ever sees its partition of the real data. The student discriminator is trained on the teacher-labelled data to maximize the cross-entropy loss given by: $L_S(\theta_S) = \sum_{j=1}^{n} r_j logS(\hat{u}_j;\theta_S) + (1-r_j)log(1-S(\hat{u}_j;\theta_S))$.

The privacy of the framework is calculated using the moments accountant method which derives a data-dependent privacy guarantee at run-time. The moments accountant tightly bounds the total privacy cost of the framework and attributes a lower privacy cost to accessing the noisy aggregation of the teachers when the teachers have a stronger consensus. The PATE-GAN framework is also evaluated against the state-of-the-art benchmark DPGAN framework on six different datasets including a real-world Credit card fraud detection dataset from Kaggle. To empirically validate the quality of the generated dataset three different training-testing settings were used. The first setting trained the predictive models on the real training set and tested the model performance on the real testing set. The second setting trained on the synthetic training set and tested on the real testing set. Finally, the third setting trained on the synthetic training set, tested on the synthetic testing set. The second setting can be used to verify if the synthetic data has captured the relationship between features and labels in the real data, whereas the first and the third settings can be used to select the best method(s) to try on the real data.

Across all datasets in the second setting, PATE-GAN is seen to be capable of generating synthetic samples that better preserve the feature-label relationship (according to the AUROC and AUPRC datasets) than DPGAN. With respect to the trade-off between privacy constraint and utility, PATE-GAN is observed to be consistently better than DPGAN over the entire range of tested ϵ. The framework is also evaluated with the Synthetic Ranking Agreement (SRA) metric which compares the performance between the first and the third settings. When compared against DPGAN across various ϵ, PATE-GAN achieves the best SRA across all values of ϵ.

II. Strengths

The PATE-GAN framework, as proposed in the paper, produces high quality synthetic data while being able to give strict differential privacy guarantees. The high quality of the synthetic data produced by PATE-GAN is confirmed by the Synthetic Ranking Agreement (SRA) metric where PATE-GAN achieves the best SRA across all values of ϵ as compared to DPGAN. With respect to the trade-off between privacy constraint and utility, PATE-GAN is also observed to be consistently better than DPGAN over the entire range of tested ϵ. This is because the PATE mechanism tightly bounds the influence of a single sample on the discriminator, and hence provides tighter differential privacy guarantees. Thus, when the differential privacy guarantee is fixed, this results in higher quality synthetic data.

III. Weaknesses

The PATE-GAN framework shows significant decreases in performance than the original GAN in high-dimensional settings. This is because the student discriminator is trained only using data from the generator, which requires that some of the generated data look somewhat realistic from the start, which is a harder requirement to satisfy as the data has more dimensions.

IV. Correctness

The claims and empirical methodology as given by the paper are correct and are supported by the existing theoretical grounding, along with the competitive results against the state of the art methods. The framework also achieved high results in three different settings that measured the quality of the synthetic dataset produced, and the ability of the model to preserve the feature-label relationship from the real/ original dataset, thus proving the correctness of the model.

V. Clarity

The paper is well written in general with sufficient theoretical grounding, comparison with prior work, and evaluations against the state of the art models. The paper also covers different settings used to measure the performance of the model in different scenarios which supplements the clarity of the work proposed.

VI. Relation to prior work

Prior work in the field of differential privacy with respect to GANs includes frameworks like DPGAN the modifies the GAN framework to be differentially private, relying on the PostProcessing Theorem to change the problem of learning a differentially private generator to learning a differentially private discriminator. The main difference is that in DPGAN, noise is added to the gradient of the discriminator during training to create differential privacy guarantees, wheres in PATE-GAN, tighter bounds on the effect of a single sample are applied which gives tighter differential privacy guarantees and makes the framework more capable of producing higher quality synthetic data.

Other prior work in the domain is related to generating synthetic data using summary statistics of the original data or based on specific domain-knowledge, both of which are limited to low-dimensional feature spaces, specific fields and do not provide any differential privacy guarantees. A few other models also generate synthetic patient records using a GAN framework, but they focus only on generating discrete variables, whereas PATE-GAN is capable of generating mixed-type (continuous, discrete, and binary) variables. In addition, these methods also do not provide any differential privacy guarantees and instead use ad-hoc notions of privacy which are only validated empirically.

VII. Reproducibility

Sufficient details are provided in terms of code implementations, the hyperparameters used for the experiments, and the theoretical grounding which would aid in reproducing the major details of the work given in the paper.

VIII. Additional comments

A holistic visual representation of the framework architecture would have supplemented the explanation of the proposed model given in the paper.