Guided PATE for Scalable Learning

Himanshu Arora
Flixstock,Inc
himanshu@flixstock.com

Abstract

Deep learning models are often trained on data which is sensitive to the people whose data has been used. In order to handle this , *Private Aggregation of teacher ensembles* or PATE is an effective framework using which one can publish deep learning models with privacy guarantees about the data used. PATE framework involves training of teacher models on disjoint subsets of data. The predictions of student dataset given by teacher models are then used to train student model with privacy guarantees. The consensus among teachers while making predictions about the student set impacts the overall privacy guarantees and the efficiency of the student model. In order to improve this consensus, we perform analysis around data partitioning among teacher models. We suggest a more effective yet simple strategy to divide datapoints among teacher models. We show our strategy improves privacy guarantees as well as efficiency of the student model

Introduction

Differential privacy introduced by (Dwork et al. 2006) is a mathematical framework which is used widely to benchmark privacy for algorithms on statistical databases. Through differential privacy, we can train machine learning models with privacy guarantees about our dataset. Intuitively, through differential privacy, a model should learn about the population as a whole and not individual datapoints. Differential privacy has become a standard for privacy and is being used for in both academia and industry.(Myers and Nelson Novemeber 2016)

Private Aggregation of teachers(PATE) (Hamm, Cao, and Belkin 2016; Papernot et al. 2016) is a framework which has been proven to achieve private learning by carefully aggregating training of several machine learning models. PATE framework uses differential privacy techniques to prove its privacy guarantees. In PATE framework, teacher models are trained on disjoint datasets and are used to supervise learning of a student model in a privacy-preserving manner. Only the student model is published for the outside world and the teacher models are kept private. Notably, PATE framework is agnostic to learning algorithm and thus can be used in lot of applications.

In our work, we study the effect of data partitioning in the PATE framework using multiple strategies. We suggest a Guided PATE framework , wherein we suggest a k-Medoids based data partitioning technique and show its effectiveness with respect to privacy baselines and accuracy of student model.

Prelimnary

Differential Privacy

Differential privacy (Dwork et al. 2006) is a mathematical framework which is used to preserve privacy of individuals in algorithms trained on statistical databases. Intuitively, differential privacy ensures that the change in output distribution of algorithm is minimal when few data points are replaced or removed. This disallows a user to find out if a particular datapoint is used for training the algorithm or not. It has become a golden standard for privacy for both industry and academia.

ε-differential privacy A randomized mechanism M over a set of databases D, satisfies (ε, δ) - differential privacy if for any two adjacent databases d, d‘ϵ D, with only one different sample, and for any subset of output S R, the following inequality holds:

$$Pr[M(d) \epsilon S] \leq e^{\varepsilon} Pr[M(d) \epsilon S] \ + \ \delta \qquad (1)$$

δ denotes the error in privacy and $\delta = 0$ in pure differential privacy. It is the probability that privacy loss is not bounded by ε and its optimal value is less than $\frac{1}{|d|}$.

Differential privacy is robust to post-processing i.e any randomized mapping of differentially private algorithms is differentially private. Composability is another important aspect of differential privacy which allows combining multiple differentially private mechanisms into one. In other words, the composition of k differentially private mechanisms where each of them is (ε, δ) - differential private would make the overall mechanism $(k\varepsilon, k\delta)$ - differential private.

As compared to classical strong composition techniques, Moment Accountant(Abadi et al. 2016) technique provides stronger bound on privacy loss. Its basic idea is to accumulate the privacy expenditure by framing the privacy loss as a

random variable and using its moment-generating functions to understand it's distribution

PATE

Private Aggregation of Teacher Ensembles(PATE)(Hamm, Cao, and Belkin 2016; Papernot et al. 2016) takes use of moment accountant mechanism to track the privacy cost in a knowledge transfer task. In this framework, teacher models are trained on disjoint subsets and each of these models is made to predict labels for student queries. These predictions are aggregated into a single prediction after carefully adding random noise sampled from Laplacian or Gaussian distributions to the predictions. If most of the teachers agreed on the same class, adding noise to the vote counts will not change the class with maximum votes. However, if two classes have a similar number of votes the final class selection among those two depends on the random noise. The intuition behind this if the consensus among teacher models is high on a particular label, it is not revealing anything about a particular teacher model or data used to train that model. The student model is then trained on the aggregated student labels and is then published for the outside world to use. PATE algorithm stays independent of the learning algorithm used to train the teacher models or the student model.

Methodology

In our method, we build upon the intuition that a better consensus among teacher models on student queries would lead to better training of student model and less privacy loss. We add another layer to PATE algorithm for splitting data in a more goal-driven manner.

Usually, the combined dataset for all teacher models is uniformly divided among teachers for their models for training. This, however, is not an optimal approach for getting the best student model. Intuitively, if the disjoint teacher subsets have diverse and representative datapoints, individual teacher models would be trained more efficiently, and that will lead to a better consensus among teachers on student queries. Our method is based on this observation.

In our method, we train a model on the combined dataset and generate features using that model for every datapoint. The features are then clustered using k-Medoids clustering approach. This gives us clusters of datapoints where each cluster contains datapoints which are similar to other datapoints in the cluster. The datapoints are now added to teacher subsets from every cluster sequentially. The datapoints added this way to subset would be from every cluster and would be different from each other. Thus, teacher subsets formed this way would have diverse and representative datapoints for every class. In 1 we illustrate on your algorithm for data partition.

Algorithm 1 Data partitioning using Guided PATE

```
Generate features for all datapoints using model trained on complete dataset S
Cluster these features into C clusters
Initialize T lists for each storing datapoints of teacher model
for cluster in C
    teacher = 0
    for datapoint in cluster
        Add datapoint to T[teacher]
        if teacher = NumofTeachers then
            teacher = 0
```

Experimental Results

We show our results on MNIST and CIFAR-100 dataset. Till now, there has been no work on evaluating data partitioning in PATE framework. Therefore, we compare our results with following data partitioning techniques.

1.) Random sampling Randomly dividing datapoints among teacher models

2.) Sampling based on classes. In this baseline, we divide datapoints corresponding to every class in equal proportions among all teacher models

We use a generic 4-layer classification network in our experiments though PATE algorithm is agnostic of training algorithm used. To evaluate our privacy we use implementation of moment accountant's method in PySyft framework.

The MNIST and CIFAR-100 dataset contains 60k training images in train set and 10k images in test set. We've used 128 as batch size, δ =10^{-5} ,noise epsilon $\epsilon = 0.1$, n_teachers=100, and used laplacian noise in aggregation of label votes. The test set is divided in 9:1 ratio for training student model and reporting test accuracy respectively. Note, the labels used for reporting test accuracy are actual labels and not predicted through teacher models. In our approach we use number of clusters as 100 for both datasets. We observe that there is not much change in performance when number of clusters are in the range of 50-200 but the performance drops if the number of clusters is increased beyond that.

In , we report (ε, δ) - differential privacy guarantees provided for MNIST dataset as well as their corresponding test accuracies. It can be seen that our model showing higher consensus gives better accuracy for student model in both cases of the number of queries. The privacy loss achieved is also 0.2-0.3 lower than the usual case. This shows the efficiency of using a more sophisticated method for splitting data among teacher models.

Dataset		Queries	Non-Private Baseline	PATE (Random)	PATE(Equal class split)	**Guided PATE**
MNIST	ε	100	-	1.84	1.71	**1.58**
MNIST	ε	250	-	2.14	1.99	**1.82**
MNIST	Test Acc.		93.2	81.2	85.2	**89.2**
CIFAR-100	ε	100	-	5.64	5.2	**4.96**
CIFAR-100	ε	250	-	7.38	7.1	**6.94**
CIFAR-100	Test Acc.		88.4	76.3	78.8	**82.3**

Table 1: Comparing ε (privacy loss) and test accuracy in original PATE framework(Papernot et al. 2016) and with our modifications when n_teachers=100 and $\delta = 10^{-5}$ and noise epsilon $\epsilon = 0.1$

In order to validate our strategy, we also calculate the variance in the predictions of student model. We observe that the variance is directly proportional to the ε value. In the case of MNIST dataset, with a random data partitioning variance is 1.45 in the predictions of student model but it falls to 1.07 in our Guided PATE strategy.

Conclusion

Through this paper, the main objective was to show the effect of data partitioning in PATE framework. Given the sensitive data of users shared on social media platforms, using such privacy algorithms becomes essential. PATE framework is agnostic to training algorithm for both student and teacher model and thus can be used in deep learning-based tasks.

We also demonstrated our Guided PATE framework through a machine learning perspective. Using clustering, we performed data splitting in a more efficient manner leading to better student model and less privacy loss. However, the technique can further be improved by using active learning techniques to improve the splitting of data among student models. We look forward to further improving the efficiency of PATE framework from both privacy and machine learning perspective.

References

Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H. B.; Mironov, I.; Talwar, K.; and Zhang, L. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 308–318. ACM.

Dwork, C.; McSherry, F.; Nissim, K.; and Smith, A. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, 265–284. Springer.

Hamm, J.; Cao, Y.; and Belkin, M. 2016. Learning privately from multiparty data. In *International Conference on Machine Learning*, 555–563.

Myers, A., and Nelson, G. Novemeber 2016. Differential privacy: Raising the bar. In *1 Geo. L. Tech. Rev*. 1(1):135-142.

Papernot, N.; Abadi, M.; Erlingsson, U.; Goodfellow, I.; and Talwar, K. 2016. Semi-supervised knowledge transfer for deep learning from private training data. *arXiv preprint arXiv:1610.05755*.