Differentially Private Active Learning with Latent Space Optimization

Sen-Ching Samson Cheung [1] **Xiaoqing Zhu** [2] **Herb Wildfeuer** [2] **Chongrou Wu** [3] **Wai-tian Tan** [2]

Abstract

Existing Active Learning (AL) schemes typically address privacy in the narrow sense of furnishing a differentially private classifier. Private data are exposed to both the labeling and learning functions, a limitation that necessarily restricts their applicability to a single entity. In this paper, we propose an AL framework that allows the use of untrusted parties for both labeling and learning, thereby allowing joint use of data from multiple entities without trust relationships. Our method is based on differentially private generative models and an associated novel latent space optimization scheme that is more flexible than the traditional ranking method. Our experiments on three datasets (MNIST, CIFAR10, CelebA) show that our proposed scheme produces better or comparable results than state-of-the-art techniques on two different acquisition functions (VAR and BALD).

1. Introduction

Supervised training of deep neural networks requires a large amount of labeled data, which are costly and time consuming to obtain. Reducing or even eliminating manual labeling effort has become an important problem for machine learning research. Much of the recent efforts have focused on new techniques for unsupervised learning, semi-supervised learning, and self-training methods. While unsupervised learning techniques do not require any labeled data, they might not be able to fully discover the hidden structure that is of interest to the application – for example, classification of generic face images based on specific facial attributes such as facial hair or eye glasses. Semi-supervised learning and self-training techniques greatly alleviate the burden of labeling as they require only a fraction of the labeled data compared to fully supervised learning. However, these techniques provide no guidance on *which* samples to label.

Active Learning (AL) methods, on the other hand, are designed to learn a model by judiciously selecting, from a pool of unlabeled samples, those that are most informative if labeled (Settles, 2011). The selection process is usually based on the optimization of a specific acquisition function, which is a function of both the unlabeled samples and the current state of the classifier. It has been demonstrated that AL can achieve the same level of training performance using fewer labeled samples (Gal et al., 2017).

Even though AL techniques have been studied for a long time, its impact on data privacy has not been fully investigated and only a handful of works can be found in the literature (Ghassemi et al., 2016; Bittner et al., 2020; Rane & Brito, 2019). The authors of (Ghassemi et al., 2016) proposed differentially-private (DP) mechanisms to screen out uninformative samples and perform a modified mini-batch update to add robustness against DP noise. These techniques were later extended to quantify the privacy-utility trade-off for online SVM and logistic regression learning (Bittner et al., 2020). The authors of (Rane & Brito, 2019) used the concept of version space to demonstrate the privacy guarantee provided by the inclusion of both informative and non-informative samples in AL.

A common thread of these works is their singular focus on protecting the privacy exposure of samples during the training update of the classifier. However, many applications require privacy protection throughout the entire AL workflow. For example, sensitive information collected by a private institution like a hospital or a university may need external experts or crowd-sourcing sites to provide labeling services. For multiple-site collaboration, external aggregation is often needed to support centralized/distributed training on a untrusted commercial cloud platform. The extensive sharing of data would certainly raise privacy concerns, even in the case when the original data are anonymized (Veiga & Eickhoff, 2016). As such, a more comprehensive solution would be needed to provide an end-to-end solution to protect the privacy of both the labeling process and the training of the model that can be scalable to arbitrary number of untrusted parties.

Our proposed solution to address this challenge is to provide privacy for both labeling and training services by employ-

[1]University of Kentucky, Department of ECE, Lexington, Kentucky [2]Innovation Labs, Intent Based Networking Group, Cisco Systems, San Jose, California [3]niversity of California, Department of CS, Davis, California. Correspondence to: Sen-Ching Samson Cheung <sccheung@ieee.org>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

ing local generative models trained on private data, and releasing only synthetic samples for labeling and training. Generative Adversarial Networks or GANs have become an important tool in protecting privacy (Fan, 2020). Formal differential privacy guarantee can be incorporated into the GAN-training process to mitigate the impact of any individual record (Xie et al., 2018). In addition, (Cheung et al., 2018) has showed that a well-trained GAN could obfuscate sensitive visual features even without differential privacy, making GANs suitable to protect privacy in the labeling process of an image-based active-learning task. While GAN-based AL has been previously investigated (Mayer & Timofte, 2020; Kim et al., 2020; Mottaghi & Yeung, 2019; Sinha et al., 2019; Tran et al., 2019), they focused on developing adversarial training processes to approximate the distributions between labeled data and the current pool of unlabeled data. These works offer no privacy protection as they assume that both the labeling process and the training of the central classifier have direct access of the original data. To the best of our knowledge, this paper is the first to apply GAN for protecting privacy of the AL process and propose a separate latent space optimization framework to identify informative samples to label. Our main contributions are as follows:

1. We proposed an end-to-end framework to protect both labeling and training of an AL pipeline using synthetic images generated by a local DP-GAN trained on private data.

2. We proposed an optimization framework that exploits the flexibility of GAN to avoid inadvertent selection of unrealistic generated samples by applying constrained stochastic gradient descent (SGD) on the latent space of the generator.

3. We have broadly tested our proposed scheme on multiple datasets (MNIST, CIFAR10, CelebA) and with different acquisition functions (BALD and VAR).

The rest of the paper is organized as follows: Section 2 reviews relevant concepts from GAN, differential privacy, and active learning. Section 3 describes the proposed AL system and the use of latent space optimization of acquisition functions in identifying informative samples to label. Experimentation results are presented in Section 4, followed by the conclusion in Section 5.

2. Background

In this section, we review concepts from GAN, differential privacy, and AL that are relevant to our proposed system.

2.1. Generative Adversarial Network

A Generative Adversarial Network or GAN consists of two separate networks: a generator $G(z) \in X$ that maps a randomly sampled d-dimensional latent vector $z \sim \mathcal{N}^d(0, I)$ to the target image space X, and a discriminator $C(x) \in \{0, 1\}$ that determines if an image input $x \in X$ looks real (1) or fake (0). Assuming that the real data come from a distribution P_X, the training goal of a GAN is to find G and C that solves the following optimization problem:

$$\min_G \max_C \mathbb{E}_{x\sim P_X}[\log(C(x))] + \mathbb{E}_{z\sim P_Z}[\log(1 - C(G(z)))] \quad (1)$$

2.2. Differential privacy

Differential privacy is a widely used framework in measuring and protecting data privacy. A randomized mechanism M applied onto a database D is called differentially private if the output of M will not change significantly when replacing D with a neighboring database D' that differs from D by at most one data record. Specifically, a differentially private mechanism can be defined as follows:

Definition 1 *A randomized mechanism M is (ϵ, δ)-differential privacy if any output set S and any neighboring databases D and D' satisfy the followings:*

$$\mathrm{P}(\mathcal{M}(D) \in S) \leq e^{\epsilon} \cdot \mathrm{P}\left(\mathcal{M}\left(D'\right) \in S\right) + \delta \quad (2)$$

A nice property of DP mechanism, which we state here without proof, is that any post-processing applied to a DP mechanism is automatically DP (Dwork et al., 2014). To provide a DP mechanism for training of a deep network on private data, the most straightforward approach is to use DPSGD (Xie et al., 2018), where zero-mean Gaussian noise with variance determined by ϵ and δ is added to the aggregate gradients during the training of the discriminator.

2.3. AL with Acquisition Functions

Identifying the unlabeled sample that contributes the most to the training of a classifier is an ill-posed problem because the contribution cannot be accurately evaluated without the actual label. Typical AL techniques follow the surrogate optimization framework by using a real-value *acquisition function* $f(x_c)$ to estimate the value of unlabeled sample x_c (Archetti & Candelieri, 2019). The surrogate optimization framework relates the cost function to the acquisition function, which guides the search to potentially low cost function values either because the prediction of $f(x_c)$ is high or the uncertainty is high. For AL in image classification, the surrogate model includes, in an iteration loop, the training of the classifier over the current set of labeled samples and the ranking of all unlabeled

samples based on their acquisition function values. The choice of acquisition function depends on the application. The detailed study in (Gal et al., 2017) has identified two high-performing acquisition functions. Given a classifier $P(y|x_c, \theta, D)$ with parameters θ, unlabeled image candidate x_c, output label $y \in \{0, 1, \ldots, M-1\}$, and training data $D = \{(x_i, y_i), i = 1, \ldots, N\}$, their definitions are as follows:

Variation Ratio (VAR) considers the probability that the pseudo output label is wrong:

$$\text{VAR}(x_c, \theta, D) := 1 - \max_y P(y|x_c, \theta, D) \tag{3}$$

Bayesian AL by Disagreement (BALD) is defined by the mutual information between the parameters and the output labels:

$$\text{BALD}(x_c, \theta, D) := H(y|x_c, D) - E_{P(\theta|D)}[H(y|x_c, \theta, D)], \tag{4}$$

where

$$H(y|x_c, \theta, D) := -\frac{1}{M} \sum_{i=0}^{M-1} P(i|x_c, \theta, D) \log P(i|x_c, \theta, D) \tag{5}$$

is the average entropy of $P(y|x_c, \theta, D)$. The posterior probability $P(\theta|D)$ in the second term depends on the Bayesian formulation of the classifier. An ensemble of Monte Carlo dropout samples of a deep neural network were used in (Gal et al., 2017), but even an ensemble of randomly initialized networks work relatively well in practice (Lakshminarayanan et al., 2017). Given an ensemble $\theta := \{\theta_i, i = 0, 1, \ldots, T-1\}$, the second term in (4) can be approximated by the sample mean:

$$E_{P(\theta|D)}[H(y|x_c, \theta, D)] \approx \frac{1}{T} \sum_{i=0}^{T-1} H(y|x_c, \theta_i, D) \tag{6}$$

while the first term in (4) is the average entropy of the Bayes estimate of the predictive probability:

$$P(y|x_c, D) \approx \frac{1}{T} \sum_{i=0}^{T-1} P(y|x_c, \theta_i, D). \tag{7}$$

Thus, the first term measures the total uncertainty of the prediction while the second terms measures the intrinsic uncertainty of the data, or aleatoric uncertainty. BALD considers their differences and measures the model uncertainty, or epistemic uncertainty, about a candidate sample.

3. Proposed Systems

3.1. The end-to-end AL framework

Figure 1 shows the proposed AL system based on synthetic images. At each of the local site with private unlabeled images, a DP-GAN is trained on these private images. As the DP-GAN is the only component that has access to these private data, its privacy guarantee (ϵ, δ) is supported by a properly implemented DP-GAN such as the one in (Abadi et al., 2016). The subsequent processing of expert labeling and classifier training will not have any access to the private data and therefore will not alter the privacy guarantee as explained in Section 2. In fact, with a well-regulated GAN, it has been argued in (Cheung et al., 2018) that DP-guarantee may not even be necessary to provide image privacy for practical image classification tasks.

Using the generator from the DP-GAN, the local site generates a set of synthetic images and sends them to the public central classifier. This approach is preferred over releasing the generator to the public as such a model release can lead to membership attacks (Hayes et al., 2017). The classifier evaluates on these synthetic images and sends the output class predictive probabilities back to the local site. These probabilities are used to compute the acquisition function (AF) value for each of the corresponding synthetic image. The AF value estimates the usefulness of each synthetic image to improve the central classifier if they were to be labeled. Unlabeled synthetic images with AF values higher than a threshold set by the central classifier are pooled together from different local sites and labeled by the external expert. The labels and the images are then added to the labeled training set of the central classifier, which is subsequently retrained. The updated classifier can evaluate on the remaining pool of unlabeled images and the active leaning process repeats.

3.2. Latent space optimization

In a traditional AL scheme based on real images (Gal et al., 2017), the AF values are ranked and those with the highest values are sent for expert labeling. There are two drawbacks for this ranking based approach. First, if a collection of $N > 1$ samples are sent out to be labeled, the ranking approach ignores the relationship among these samples and does not take into account the benefit of labeling the entire selection of N samples together. This could result, for example, in a highly unbalanced labeled dataset to train the central classifier. Second, as the local site has full access to the generator, it opens an interesting opportunity to select "optimal" latent vectors in synthesizing new fake images that have high acquisition function values. Our proposed latent space optimization is designed to provide more flexibility in the design of AF by integrating the generator into its optimization.

Given a trained generator $x = G(z)$ that maps a latent vector z to a synthetic image x, our goal is to identify informative x's to train a Bayesian classifier $P(y, \theta^t|x)$ at the t-th iteration. The surrogate framework aims to optimize

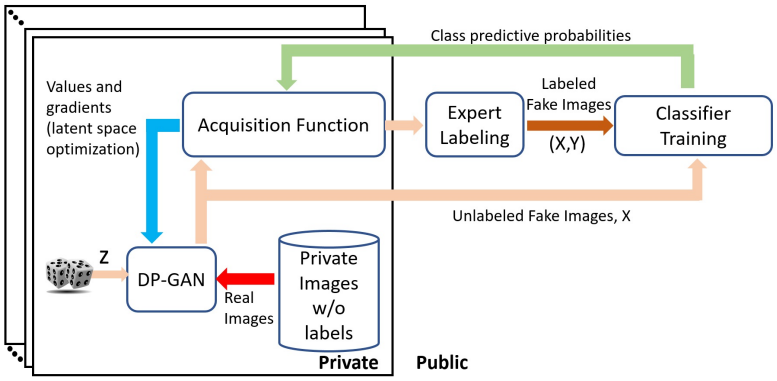

Figure 1. Proposed Privacy-Preserving AL system. The local site trains a DP-GAN from its collection of private unlabeled images. It leverages the current version of the publicly trained classifier for evaluating the acquisition functions (AF) of any given unlabeled synthetic image. Either ranking or latent space optimization is used to select the more informative unlabeled synthetic images for expert labeling in the public domain, followed by further training of the public classifier based on these labeled samples.

the acquisition function $f(\cdot)$:

$$\arg\max_{z\in\Gamma} f(P(y,\theta^t|G(z))) \tag{8}$$

The ranking approach relies on a pre-selected finite set Γ of latent vectors as the pool of candidates in evaluating $f(\cdot)$. If we want to identify $N > 1$ samples to label, selecting the top N samples based on (8) can result in choosing very similar samples of high AF values, thereby wasting valuable labeling effort. A better approach needs to consider the entire batch of N samples together:

$$\arg\max_{z_1,\ldots,z_N\in\Gamma} f\left(P(y_1,\ldots,y_N,\theta^t|G(z_1),\ldots,G(z_N))\right) \tag{9}$$

The joint distribution $P(y_1,\ldots,y_N,\theta^t|G(z_1),\ldots,G(z_N))$ allows the AF to optimally select, among all possible N samples, the best set to be labeled that results in the most improvement in training performance. This optimal set is therefore more diverse than those derived from ranking of individual samples. To solve this combinatorial maximization problem, the complexity of the ranking technique will grow exponentially with N. As such, this problem can only approximated, for example, by using the greedy heuristic on the BALD function as proposed in (Kirsch et al., 2019).

Since we have direct access to G and if f is a differentiable function, it is possible to solve the continuous version of this optimization by setting $\Gamma = \mathbb{R}^d$ and using stochastic gradient ascent methods. An unconstrained optimization, however, does not work as it ignores the quality of the resulting synthetic image $G(z)$. Since the GAN was originally trained with the latent vectors $z \sim P_Z = \mathcal{N}^d(0, I)$, a natural modification would be to use this Gaussian distribution as a constraint:

$$\arg\max_{\substack{z_1,\ldots,z_N\in\mathbb{R}^d,\\ P_Z(z_i)\geq P_Z(z_i^0)}} f\left(P(y_1,\ldots,y_N,\theta^t)|G(z_1),\ldots,G(z_N))\right) \tag{10}$$

where $z_i^0 \sim \mathcal{N}^d(0, I)$ for $i = 1,\ldots,N$ are the initial starting points of the SGD process. The constraint probabilistically guarantees that the optimization process will produce the optimal latent vector z_i^* such that the synthetic image $G(z_i^*)$ has the same or better quality than $G(z_i^0)$. All the constraints are convex as $P_Z(z_i) \geq P_Z(z_i^0)$ is the same as $\|z_i\| \leq \|z_i^0\|$. Convex constraint can be easily realized in SGD by projecting the search trajectory back onto the convex constraint (Shalev-Shwartz & Ben-David, 2014). This paper focuses on $N = 1$ and the extension to larger N will be forthcoming.

3.3. Adaptation of BALD and VAR to SGD

SGD requires the acquisition function to be differentiable, which is certainly the case for BALD in (4). The use of Bayesian classifier ensemble slightly complicates the calculations and requires separate handling of the gradients of the two terms in (4). The second term is straightforward as it is the average of the average entropy of each classifier. The gradient can be computed for individual classifiers and then averaged. The first term requires averaging of the output probabilities of the classifier before taking the average entropy. As such, its gradient needs to be explicitly computed based on the gradient and the output from each classifier as follows:

$$\nabla_x H(y|x,D) = -\frac{1}{M}\sum_{i=0}^{M-1}\left(\frac{1}{T}\sum_{j=0}^{T-1}\nabla_x P(y|x,\theta_j,D)\right)\cdot \left[\log\frac{1}{T}\sum_{j=0}^{T-1}P(y|x,\theta_j,D)+1\right] \tag{11}$$

VAR in (3) uses a maximum function over a finite set of non-negative numbers less than 1. As the maximum function

is not differentiable, we replace it with a differentiable log-exponential function as follows (Nielsen & Sun, 2016):

$$\max(x_1, \ldots, x_m) \approx \frac{1}{\alpha} \log \left[\sum_{i=1}^{m} e^{\alpha x_i} - (m-1) \right] \quad (12)$$

The computation of $e^{\alpha x_i}$ can cause overflow, which can be prevented by setting the scale parameter α appropriately. In our experiments, we set it to 86 to prevent overflow under single-precision floating point computations.

4. Experimental Results

In this section, we present preliminary results on the impact of different AFs, latent space optimization (LSO) vs ranking vs random, and differential privacy on using synthetic images for AL. To ensure fair comparisons, we use the same set of latent vectors, the same trained generator G, and the same initial set of synthetic images to start the feedback loop across all the schemes for the same dataset. For random and LSO schemes, the orders of latent vectors z's presented to the AL training are the same – the random scheme will simply use $G(z)$ while the LSO scheme will use z as the starting point for SGD. To emulate the expert labeling process, a labeling classifier, trained using the real training images, is used to provide soft labels for all synthetic images.

4.1. Datasets and Architectures

We have tested our system using three standard datasets: MNIST, CIFAR10, and Celeb-A. MNIST (LeCun et al., 1998) contains 60,000 training and 10,000 testing grayscale handwritten digit images of size 28×28. Its GAN is based on a simplified WGAN (Arjovsky et al., 2017) with only two convolutional layers. The generator has a dense layer followed by two transposed convolutional layers. The classifier follows the same structure as the discriminator, and 8x MC-dropout is to used to compute the BALD function (Gal et al., 2017). The classifier is trained three times to average the predictive probabilities to mitigate the variation due to random initialization.

CIFAR10 (Krizhevsky & Hinton, 2009) contains 50,000 training and 10,000 testing 32×32×3 natural images from 10 classes. The improved WGAN from (Gulrajani et al., 2017) is used while the classifier is based on Resnet-18. Deep ensembles with 3 different random initialization are used to estimate BALD (Lakshminarayanan et al., 2017).

Celeb-A (Liu et al., 2015) contains over 200,000 218×178×3 facial images in color with a rich set of ground-truth labels on binary attributes (e.g., male vs. female, with glasses vs. without glasses). All images are resized to $64 \times 64 \times 3$ to support the training of DC-GAN following its original architecture (Radford et al., 2016). We further combine two binary labels (male vs. female and smile vs. no smile) into 4 classification categories and randomly select 30,000 images for each combined category to form a balanced dataset. It is randomly partitioned into a training set with 96,000 images and a test set with 24,000 images. The classifier uses the same architecture (except for the output layer) as the Discriminator in the DC-GAN. BALD is also computed using deep ensembles.

4.2. Non-DP results

Figures 2, 3, and 4 show the improvements on accuracy when progressively adding labeled samples to the AL systems. We have tested the following schemes:

Rand (S)	Randomly selected synthetic samples
BALD (SR)	Ranked synthetic samples using BALD
VAR (SR)	Ranked synthetic samples using VAR
BALD (SO)	LSO synthetic samples using BALD
VAR (SO)	LSO synthetic samples using VAR
Rand (R)	Randomly selected real samples (ideal)

To facilitate visual comparisons of different plots, different point types are used to represent different AFs: triangles for BALD, squares for VAR, cross for random, and diamond for random selection of real images as a control reference. Different line types indicate manipulations on the images: solid means LSO and broken means that the images are used as is.

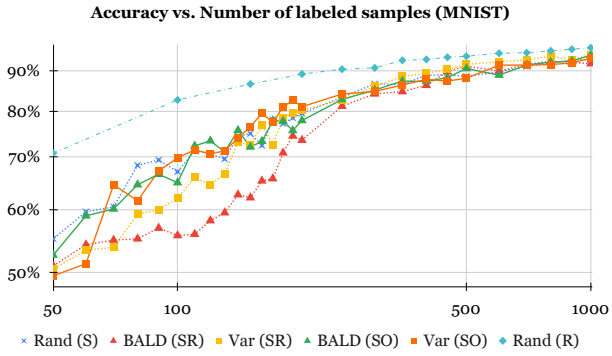

Figure 2. AL performance (MNIST)

For both MNIST and CIFAR-10, there is a sizeable gap in accuracy between using real versus synthetic images, especially when the number of labeled samples is small. However, there is no noticeable difference between real and synthetic for CelebA. For MNIST, the two LSO curves perform comparably with random but the two ranking schemes perform much worse. Upon closer inspection, we noticed that poor-quality synthetic images that do not contribute much to the learning actually get high AF values. For the more challenging CIFAR-10, the two LSO schemes outperform random and the ranking schemes. The trend, however,

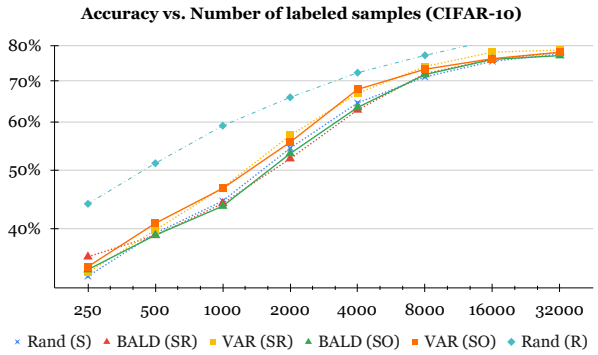

Figure 3. AL performance (CIFAR-10)

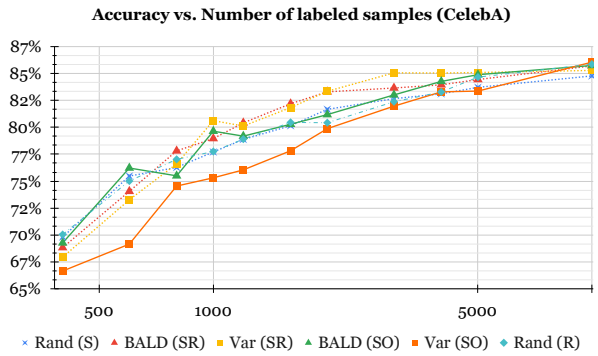

Figure 4. AL performance (CelebA)

is reversed for CelebA with both ranking schemes come out on top. One possible explanation is that the GAN for CelebA can produce images of higher quality than those from CIFAR-10, and the LSO scheme inadvertently affects the reconstruction quality of those high quality images. These two observations imply that LSO is likely to perform better than ranking for datasets that are more challenging to synthesize.

4.3. DP results

Figures 5 and 6 show the scenarios when a differentially private GAN is used for MNIST and CIFAR10 respectively. Under the DP framework, the amount of noise added to the training process is controlled by the parameter ϵ. The smaller the ϵ, the larger amount of noise is injected, which results in better privacy. However, the presence of the noise in training GAN can lead to instability or mode collapse. In our experiemnts, we have chosen the most stringent privacy protection (the smallest ϵ: 10 for MNIST and 50 for CIFAR010) before the training of the DP-GAN fails to converge. δ is set at 10^{-5} and a hyper-parameter search is performed on the clip-norm magnitude. For MNIST, similar to the non-DP case, the BALD-ranking scheme performs much worse than the others while both ranking and LSO schemes using VAR perform mostly better than random selection. A similar trend is also observed for CIFAR10, though the LSO scheme using VAR performs quite a bit better than the others. Again, this illustrates the advantage of the LSO schemes in handling lower-quality GANs, such as in the case for DP when the training of GAN is hindered by the added DP noises.

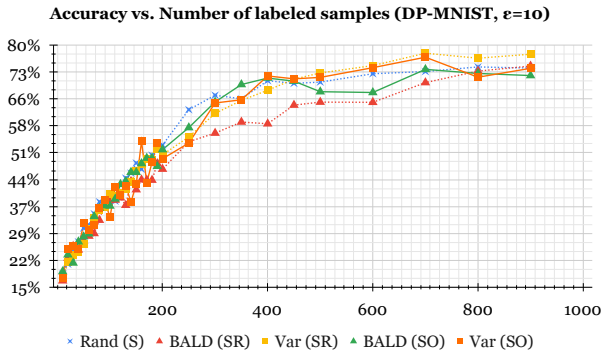

Figure 5. AL performance (DP-MNIST at $\epsilon = 10$ and $\delta = 10^{-5}$)

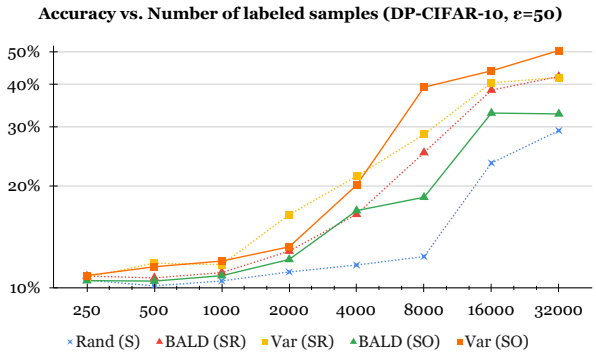

Figure 6. AL performance (DP-CIFAR-10 at $\epsilon = 50$ and $\delta = 10^{-5}$)

5. Conclusions

In this paper, we have proposed a novel differentially private AL system that protects the entire AL pipeline from labeling to training. Using a DP-GAN as a data release mechanism, we have considered both ranking and a novel latent-space optimization scheme to produce informative samples for the labeling process. Future investigation will focus on new acquisition functions on the entire labeling set and the use of semi-supervised learning at the central learner.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM*

SIGSAC Conference on Computer and Communications Security, pp. 308–318. ACM, 2016.

Archetti, F. and Candelieri, A. *Bayesian optimization and data science*. Springer, Cham, 2019.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

Bittner, D. M., Brito, A. E., Ghassemi, M., Rane, S., Sarwate, A. D., and Wright, R. N. Understanding Privacy-Utility tradeoffs in differentially private online active learning. *Journal of Privacy and Confidentiality*, 10(2), June 2020.

Cheung, S.-C., Wildfeuer, H., Nikkhah, M., Zhu, X., and Tan, W. Learning sensitive images using generative models. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 4128–4132. ieeexplore.ieee.org, October 2018.

Dwork, C., Roth, A., et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

Fan, L. A survey of differentially private generative adversarial networks. In *The AAAI Workshop on Privacy-Preserving Artificial Intelligence*. webpages.uncc.edu, 2020.

Gal, Y., Islam, R., and Ghahramani, Z. Deep bayesian active learning with image data. *arXiv:1703.02910*, March 2017.

Ghassemi, M., Sarwate, A. D., and Wright, R. N. Differentially private online active learning with applications to anomaly detection. In *Proceedings of the 2016 ACM Workshop on Artificial Intelligence and Security*, AISec '16, pp. 117–128, New York, NY, USA, October 2016. Association for Computing Machinery.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pp. 5767–5777, 2017.

Hayes, J., Melis, L., Danezis, G., and De Cristofaro, E. Logan: evaluating privacy leakage of generative models using generative adversarial networks. *arXiv preprint arXiv:1705.07663*, 2017.

Kim, K., Park, D., Kim, K. I., and Chun, S. Y. Task-aware variational adversarial active learning. *arXiv preprint arXiv:2002.04709*, 2020.

Kirsch, A., Van Amersfoort, J., and Gal, Y. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *arXiv preprint arXiv:1906.08158*, 2019.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. 2009.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 6402–6413. Curran Associates, Inc., 2017.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Mayer, C. and Timofte, R. Adversarial sampling for active learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3071–3079, 2020.

Mottaghi, A. and Yeung, S. Adversarial representation active learning. *arXiv preprint arXiv:1912.09720*, 2019.

Nielsen, F. and Sun, K. Guaranteed bounds on the kullback–leibler divergence of univariate mixtures. *IEEE Signal Processing Letters*, 23(11):1543–1546, 2016.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv:1511.06434*, January 2016.

Rane, S. and Brito, A. E. A version space perspective on differentially private Pool-Based active learning. In *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–6. ieeexplore.ieee.org, December 2019.

Settles, B. From theories to queries: Active learning in practice. In *Active Learning and Experimental Design workshop In conjunction with AISTATS 2010*, pp. 1–18. jmlr.org, 2011.

Shalev-Shwartz, S. and Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, May 2014.

Sinha, S., Ebrahimi, S., and Darrell, T. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5972–5981, 2019.

Tran, T., Do, T.-T., Reid, I., and Carneiro, G. Bayesian generative active deep learning. In *International Conference on Machine Learning*, pp. 6295–6304. PMLR, 2019.

Veiga, M. H. and Eickhoff, C. Privacy leakage through innocent content sharing in online social networks. *arXiv:1607.02714*, July 2016.

Xie, L., Lin, K., Wang, S., Wang, F., and Zhou, J. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*, 2018.