DP-CGAN : Differentially Private Synthetic Data and Label Generation

Reihaneh Torkzadehmahani
University of California, Santa Cruz
rtorkzad@ucsc.edu

Peter Kairouz
Google AI
kairouz@google.com

Benedict Paten
University of California, Santa Cruz
bpaten@ucsc.edu

Abstract

Generative Adversarial Networks (GANs) are one of the well-known models to generate synthetic data including images, especially for research communities that cannot use original sensitive datasets because they are not publicly accessible. One of the main challenges in this area is to preserve the privacy of individuals who participate in the training of the GAN models. To address this challenge, we introduce a Differentially Private Conditional GAN (DP-CGAN) training framework based on a new clipping and perturbation strategy, which improves the performance of the model while preserving privacy of the training dataset. DP-CGAN generates both synthetic data and corresponding labels and leverages the recently introduced Rényi differential privacy accountant to track the spent privacy budget. The experimental results show that DP-CGAN can generate visually and empirically promising results on the MNIST dataset with a single-digit epsilon parameter in differential privacy.

1. Introduction

Recent studies have shown that deep neural networks (DNNs) can achieve state-of-the-art performance in various applications such as image recognition [25, 30], natural language processing [9], speech recognition [26, 20] and complex video games [41, 34]. It has not only achieved exceptional accuracy in different tasks but also surpassed human-level performance in some of them [41, 24]. DNNs have also been leveraged in health-related studies ranging from medical images [22, 39, 38, 21, 4] to human genome analyses [3, 29, 45].

Generative Adversarial Networks (GANs) [19] form a well-researched class of generative models [27, 5, 31, 6]. They can learn the distribution of the training data and generate synthetic data with a distribution very similar to the distribution of the training data. GAN models are particularly used by research communities to generate the synthetic datasets in cases where they cannot directly access sensitive datasets. However, using sensitive data to train GAN models raises privacy concerns for participating individuals. Indeed, recent works show that most machine learning models, including GAN models, are vulnerable to a slew of attacks (from model inversion attacks to membership inference attacks) that can expose significant information about training data [40, 23, 17, 46].

Differential Privacy (DP) [10, 11] is a common technique to protect the privacy of ML models trained on sensitive data. However, in spite of its popularity, there have been very few recent studies on training GANs in a differentially private way [43, 18, 44, 7, 8]. The standard procedure leveraged by these recent studies to enforce DP is to first clip the l2 norm of the gradients of the sum of the discriminator's loss on real and fake data and then add Gaussian noise to the clipped gradients. To keep track of the privacy budget, they typically use the Moment Accountant (MA) technique [1]. One of the limitations of these recent works is that they focus exclusively on generating synthetic data (e.g., images) without corresponding labels – an aspect that renders the synthetically generated data useless for supervised learning applications. More importantly, training high quality GANs with a single digit epsilon parameter (for differential privacy) has been absent so far even for the simplest of all tasks: generating MNIST-like digits.

In this work, we propose a Differentially Private Conditional GAN (DP-CGAN) training framework, which can preserve the privacy of conditional GAN models using DP [10, 11]. The main idea in DP-CGAN is that it clips the gradients of discriminator loss on real and fake data separately, which allows the designer to better control the sensitivity of the model to real (sensitive) data. Moreover, DP-CGAN can generate not only synthetic data but also corresponding labels. Further, DP-CGAN employs the newly introduced Rényi Differential Privacy (RDP) Accountant [32] to track the privacy budget. In comparison to the classical MA technique, RDP accounting provides a tighter bound on the privacy budget, allowing for the addition of less noise without compromising the privacy guarantees.

DP-CGAN framework has three main components: conditional generator network, differentially private discriminator network, and privacy accountant. At each step of the training process, the discriminator network is trained in a differentially private manner in which the gradients of loss on real and fake data are clipped separately. Afterwards, the sum of these two set of clipped gradients are computed and noised by adding Gaussian noise to them. Then, the privacy accountant, which is based on the RDP accountant [32], is updated by accumulating the spent privacy budget at each step. Next, the generator network is trained with a non-private optimizer. At any given point in time, if the privacy budget exceeds the target one, the training process is halted and the conditional generator network is ready for the creation of synthetic data and labels.

We make the following contributions in this work:

- We propose DP-CGAN based on a new gradient clipping and noising procedure, which improves the performance compared to the standard procedure to preserve privacy. To the best of our knowledge, DP-CGAN is the first differentially private GAN framework than can generate both the synthetic data and corresponding labels with promising results. It leverages the recently introduced RDP accountant and TensorFlow Privacy[1] package (by Google) to keep track of the privacy budget.

- We provide preliminary experimental results showing that DP-CGAN can generate good visual and empirical results on MNIST dataset with single-digit epsilon parameter. This suggests that our work can be viewed as the first stepping stone towards training high quality GANs with strong DP guarantees.

- We use the differnetially private conditional generative model to create synthetic data and labels which are used (together) in the training of machine learning models. We test the accuracy of the learned models on real data and show that they perform well. We get an area under the ROC (AUROC) of 87.57% using DP-CGANs compared to 92.17% if we were to train the classifier directly on real data.

The remainder of the paper is organized as follows: Section 2 provides a background on GAN, CGAN, and differential privacy. Section 3 overviews the previous related work in the area of preserving the privacy of deep learning models. Section 4 describes the DP-CGAN framework in detail. Section 5 provides the experimental results and Section 6 concludes the paper with a brief conclusion.

2. Preliminaries

In this section, we review Generative Adversarial Networks(GAN), Conditional Generative Adversarial Networks(CGAN) and differential privacy concepts used in DP-CGAN.

2.1. GAN and CGAN

Nowadays, there is a great interest in using generative models to create synthetic data that looks like the original one. Generative Adversarial Network(GAN) proposed by Goodfellow et. al [19] is one the primary methods to learn generative models for images. GANs consist of two main components: a generator and a discriminator. The generator takes noise as input and generates synthetic data by capturing the original data distribution while the discriminator takes the synthetic data (generator's output) as well as original data (training set) and learns to discriminate between the real (training) and fake (synthetic) data distribution. The discriminator returns two possible values as output which is the assigned score to a test sample representing whether it is real or fake data. The generator and discriminator always try hard to be as accurate as possible and the more the generator improves the quality of the fake data, it gets harder for discriminator to distinguish the difference between the original and fake data. These two components always play a game and are trained simultaneously.

Suppose $p_z(z)$ is the probability distribution that random noise z is taken from, $G(z)$ is the generator network that takes the random noise z as input and $D(x)$ is the discriminator network that takes the generator's output as well as the input data x taken form the distribution $p_{data}(x)$. The game that the generator and discriminator play to achieve a trade-off, encapsulates in the following objective function, $V(D,G)$, of a minimax game:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[log(D(x))] + \mathbb{E}_{z\sim p_z(z)}[log(1-D(G(z)))] \tag{1}$$

Conditional GAN [33] is an extension of GAN in which both generator and discriminator are conditioned on some side information, "y" that can be any kind of extra information like class labels or data from other modalities. The objective function of a minimax game for CGAN is as the following:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[log(D(x|y))] + \mathbb{E}_{z\sim p_z(z)}[log(1 - D(G(z|y))))] \tag{2}$$

[1] https://github.com/tensorflow/privacy

2.2. Differential Privacy

Differential privacy [10, 11] is a mathematical framework to express the level of privacy preservation of individuals in a statistical databases. It provides strong privacy guarantees for algorithms on aggregate databases. Intuitively, in differential privacy, the user should learn about population as a whole but not about particular individual. In other words, if we replace individual I with another random member of the population, the user should learn the same thing about the dataset in presence or absence of individual I. Differential privacy has become an actual standard in data protection in both academia and industry [35] (Apple [2], Google [42] and US Census [37]).

Definition 1. (*differential privacy*) A randomized mechanism M over a set of databases D, satisfies (ϵ, δ)-differential privacy if for any two adjacent databases $d, d^{'} \in D$, with only one different sample, and for any subset of output $S \in R$, the following inequality holds:

$$Pr[M(d) \in S] \leq e^{\epsilon} Pr[M(d') \in S] + \delta \qquad (3)$$

In pure differential privacy, $\delta = 0$ and the additive term δ does not exist while in approximate differential privacy [10], δ is used for approximation in the cases that pure differential privacy is broken. δ is the probability that privacy loss is not bounded by ϵ and its optimal value is smaller than $\frac{1}{|d|}$ (inverse of the database size).

Differential privacy is resistant to post-processing. That is, any arbitrary randomized mapping of an (ϵ, δ)-differentially private algorithm, is differentially private as well.

Theorem 1. (*post-processing*) Given a randomized algorithm $M : D \to R$ that is (ϵ, δ)-differentially private and an arbitrary randomized mapping $f : R \to R^{'}$, $f \circ M : D \to R^{'}$ is (ϵ, δ)-differentially private.

A routine approach to privatizing the output of a real-valued function $f : D \to \mathbb{R}$ is to add noise with variance in the scale of f's *sensitivity*, S_f, to the output. The sensitivity of a function f is defined as the maximum absolute distance $|f(d) - f(d^{'})|$ (d and $d^{'}$ are adjacent databases). In formal notion:

$$S_f \equiv \max_{d \sim d'} |f(d) - f(d^{'})|, \qquad (4)$$

Gaussian noise is one of the popular kinds of noise employed in differential privacy, in which $f(d)$ is perturbed by Gaussian noise $N(0, {S_f}^2.\sigma^2)$. That is:

$$M(d) \equiv f(d) + N(0, {S_f}^2.\sigma^2) \qquad (5)$$

Composability is one of the interesting properties of diferetnial privacy that makes it possible to combine multiple differentially private mechanisms into one. A standard analysis implies the composition of k mechanisms that each of them are (ϵ, δ)-differentially private, is at least $(k\epsilon, k\delta)$-differentially private [10, 11, 12]. One of the possible ways of accounting differential privacy in composition of additive-noise mechanisms is to use Moment Account technique introduced by Abadi et. al [1], which provides strong estimates of privacy loss compared to various versions of composition theorem [10, 12, 28, 15, 16] including strong composition theorem [16]. RDP accountant [32] is a new approach based on a new definition of privacy, Rényi differential privacy, which provides a tighter bound for privacy loss in comparison with Moment Accountant.

3. Related Work

Some previous studies have proposed approaches to addressing the problem of preserving privacy in Deep Learning. Shokri et al. [40] developed a distributed approach in which multiple parties train a model on their local training set independently. Then, each party selects a set of key parameters, and shares them with the other parties. Although this method has high training accuracy without sharing the input parameters, Abadi et al. [1] showed that the overall privacy loss for each party exceeds several thousands on MNIST dataset using Moment Accountant technique they introduced.

Moment accountant mechanism [1] can be used to track the overall spent privacy budget, (ϵ, δ), for composing Gaussian Mechanisms with random sampling (e.g. training process in Stochastic Gradient Descent). This method provides a much tighter estimation for privacy loss compared to standard composition theorem [13]. It computes the log moments of the random variable indicating privacy loss and then calculates the tail bound using moments bound and standard Markov inequality. The result is privacy loss estimation in terms of differential privacy. In addition to Moment Accountant technique, Abadi et al. [1] proposed a method to make the Stochastic Gradient Descent(SGD) process differentially private.

Private Aggregation of Teacher Ensembles (PATE) [36] is a framework that leverages the moment accountant mechanism to trace the privacy leakage of knowledge transfer task using differential privacy. It presents a differentially private semi-supervised learning method in which the training data is split into multiple disjoint sets and the teacher models are trained independently. The teacher ensemble predicts the labels after perturbing counts of teachers' votes by Laplace noise while the student model is trained on public data as well as labeled data from the teacher model and can be published publicly. Although this method outperforms Shokri et al. [40] work in terms of both accuracy and privacy, it assumes the model has access to public data which may not be the case in practice. Moreover, the teacher ensemble just responds to the queries for which the consensus among teachers is sufficiently high.

Some other previous researches focused on preserving privacy of GANs in particular. DPGAN method [43], enforces differential privacy during the training process of the discriminator by adding Gaussian noise to the gradient of Wasserstein distance in WGAN algorithm and uses post-processing theorem to guarantee differential privacy for the generator. However, it is unclear how the overall privacy budget is accounted, the results do not look promising even on MNIST dataset and there is no methodology for creating labels for synthetic images.

Similar to DPGAN method, PATE-GAN approach [44] enforces privacy by making the discriminator differentially private. In PATE-GAN, the discriminator is replaced with modified version of PATE [36] in which the student model allows back-propagation to the generator and there is no need to have access to public training data. It employs the generated data to train different classifiers and evaluate the quality of generated data by testing these classifiers on real test data. The limitation of PATE-GAN is that it assigns binary labels for synthetic data, and therefore, it is not applicable for multi-label datasets. Moreover, the datasets used to evaluate the model are small. The other work is a DP-GAN framework for time series, continuous, and discrete data [18]. This framework is alike the previous DPGAN work [43] except it employs moments accountant approach to account the privacy budget and clips the discriminator gradients while reducing the clipping parameter over time (adaptive clipping).

Unlike DPGAN method [43], our proposed method leverages RDP accountant technique to follow the consumed privacy budget, (ϵ, δ) and generates not only synthetic data but also the labels using a Conditional GAN model. In contrast to PATE-GAN [44] which generates only binary labels, our model generates multi-class labels. Finally, in DPGAN frameworks [43, 18] the discriminator gradients are clipped and perturbed by adding Gaussian noise to gradients of the discriminator loss, while in our framework, Gaussian noise is added to the accumulation of clipped gradients of discriminator loss on real data and clipped gradients of discriminator loss on fake data.

4. Our Approach

As mentioned before, DP-CGAN can generate the synthetic data as well as the corresponding labels while preserve privacy of training samples. To this end, the DP-CGAN makes the training process private by injecting random Gaussian noise into the optimization process of the discriminator network. Based on post-processing theorem[14] making the generative network differentially private results in having a differentially private generator too. DP-CGAN tracks the spent privacy loss using RDP accounting technique[ényi], which provides tighter estimation on privacy loss in comparison with moment accountant technique. The training procedure stops if the spent privacy budget (ϵ, δ) goes beyond the target ones.

DP-CGAN makes the optimization process of discriminator loss (discriminator training) differentially private by computing the per-example gradients of the discriminator loss on both real and fake data, clipping the per-example gradients on real data and fake data separately, summing up two sets of the clipped gradients, perturbing the clipped gradients by adding Gaussian noise $N(0, \sigma^2 C^2)$, σ is noise multiplier and C is clipping value, to them, and finally applying the perturbed gradients.

Algorithm 1 outlines the training process of DP-CGAN. According to the algorithm, the model updates the discriminator network and the generator network as long as the number of iterations is less than maximum iteration count and the spent privacy budget is less than the target ϵ. At each step, it minimizes the discriminator loss function by computing the discriminator gradients of loss on real data and clipping them by L_2-norm (lines 9-12), computing the discriminator gradients of loss on fake data(lines 13-15) and clipping them by L_2-norm, compute the overall clipped gradients of discriminator by adding these two sets of clipped gradients, adding Gaussian noise to them and taking average over all the perturbed clipped per-example gradients in the batch(line 16-17), and finally applying the gradients (line 18). The model tracks the spent privacy budget by accumulating the spent privacy budget and updating the RDP accountant every time that noise is injected into the model(line 20). Then, the generator the gradients of generator loss are computed and applied so that the generator network gets trained(line 21-25). The last step is to check the overall spent privacy budget so far. If the spent ϵ or the spent δ has exceeded the target values, training is stopped, otherwise it continues (line 26-27).

5. Experimental Results

We compare the performance of DP-CGAN to CGAN with no privacy and CGAN trained with standard differentially private approach.The CGAN architecture used in all models is a vanilla CGAN in which both generator and discriminator consist of two fully connected layers.The generator takes random noise sample z and the corresponding label y as inputs while the discriminator inputs are real training sample x and its label y. Figure 1 depicts the generator and discriminator architecture of the vanilla CGAN.

Differentially private CGAN models use the new privacy package of TensorFlow Privacy (by Google), a python library that includes the implementation of few differentially private optimizers as well as the privacy accountants to keep track of the privacy loss. They leverage differentially private Gradient Descent as optimizer and RDP accountant as privacy accountant from this package.

The dataset used used in the evaluation is MNIST hand-

Algorithm 1: DP-CGAN

```
Examples $\{x_1, x_2, ..., x_N\}$, labels $\{y_1, y_2, ..., y_N\}$, target epsilon $\epsilon$, target delta $\delta$, noise scale $\sigma$, clip norm bound $C$, learning rate $lr$, batch size $bs$
Differentially private Generator that generates synthetic data and labels
$should\_terminate = False$
while $(step \leq max\_step \ \ \& \ ! \ should\_terminate)$ do
    - Sample random batch $(X^t, Y^t)$ of size $bs$ with probability $bs/N$ from data distribution $p_{data}(X)$

    - Sample noise batch $Z_t$ of size $bs$ from noise prior $p_z(z)$
    /* Update the Discriminator Network */
    $d\_loss\_real \leftarrow log(D(X^t))$
    $d\_loss\_fake \leftarrow log(1 - D(G(Z^t)))$
    Compute per-example gradients of discriminator loss on real data $X_t$ and clip them
    for $i \in X_t$ do
        Compute ${grad_{d\_real}}^t \leftarrow \nabla_{\theta_d} d\_loss\_real({\theta_d}^t, X_i)$
    ${grad_{d\_real}}^t = {grad_{d\_real}}^t / max(1, \frac{||grad_{d\_real}||_2}{C})$
    Compute per-example gradients of discriminator loss on fake data $Z_t$ and clip them
    for $i \in Z_t$ do
        Compute ${grad_{d\_fake}}^t \leftarrow \nabla_{\theta_d} d\_loss\_fake({\theta_d}^t, Z_i)$
    ${grad_{d\_fake}}^t = {grad_{d\_fake}}^t / max(1, \frac{||grad_{d\_fake}||_2}{C})$
    Compute the overall gradients of discriminator and add Gaussian Noise to them
    ${grad_d}^t \leftarrow \frac{1}{bs} \sum {grad_{d\_real}}^t + {grad_{d\_fake}}^t + N(0, \sigma^2 C^2 I)$
    Take the gradient Descent step for discriminator
    $\theta_{dt+1} \leftarrow SGD(grads\_d^t, \theta_{dt}, lr))$
    /* Update RDP Accountant */
    Accumulate the spent privacy budget using RDP Accountant
    /* Update the Generator Network */
    $g\_loss \leftarrow log(1 - D(G(Z^t)))$
    Compute gradients of generator loss
    Compute $grad\_g^t \leftarrow \nabla_{\theta_g} g\_loss({\theta_g}^t, Z_i)$
    Take the gradient Descent step for generator
    ${\theta_g}^{t+1} \leftarrow ADAM(grad\_g^t, {\theta_g}^t)$
    if $spent\_epsilon > \epsilon \ OR \ spent\_delta > \delta$ then
        /* Running out of privacy budget */
        $should\_terminate = True$
```

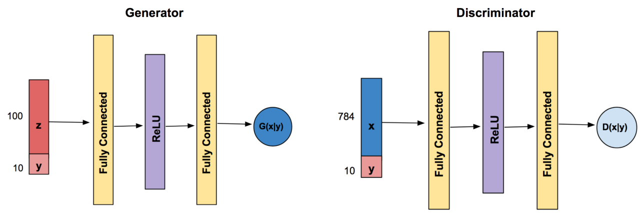


Figure 1: Vanilla CGAN Generator and Discriminator Architecture

written dataset containing 60k training samples and 10k test samples. In the experiments, batch size is set to 600, $\delta = 10^{-5}$ and learning rate is set by an adapative approach in which the initial learning rate is 0.15, it is decreased to 0.052 in iteration 10K and is fixed on 0.052 for the rest iterations.

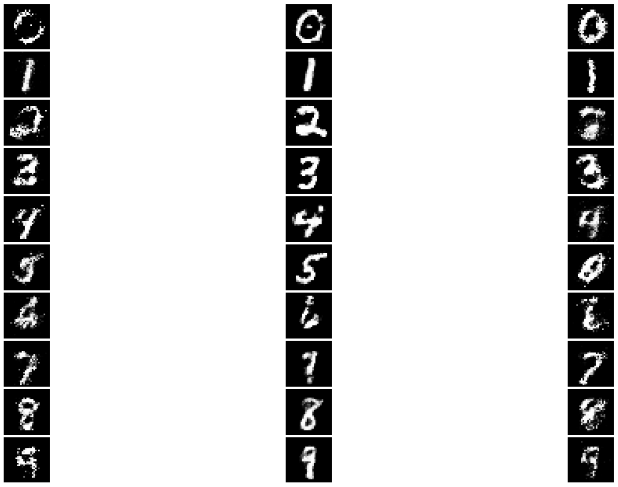
We trained Logistic Regression and Multi-Layer Perceptron classifiers using the synthetic data and labels generated by the models and tested the classifier on real test data. Closer performance of the classifier trained on synthetic data generated by differntially private models and on real data indicates that the model has captured the real data distribution better. We measured the performance of the classifier using the Area under ROC curve metric (AuROC). In the evaluation process, the generative model takes the 60k MNIST training data and the labels as input and generates 60k synthetic labeled data.Then, the classifier is trained on the generated data. Finally, performance of the trained classifier is evaluated on the 10k test data using AuROC metric.

Table 1. lists the results of AuROC for the three models as well as the case in which classifiers are trained on real data. According to the table, the AuROC of DP-CGAN is higher than CGAN trained with basic differentially private method, indicating that new clipping and perturbing technique used in DP-CGAN improves the performance. On the other hand, the AuROC of DP-CGAN is about 5% lower than that for real data and this is the price we pay to have privacy.

	Real	CGAN	**DP-CGAN**	CGAN with basic DP
LR	92.17%	91.10%	**87.57%**	83.42%
MLP	97.60%	91.06%	**88.16%**	83.29%

Table 1: Comparing AuROC for Logistic Regression(LR) and Multi-Layer Perceptron(MLP), which are trained on real data, data generated by CGAN (non-private), DP-CGAN and CGAN with basic differentially private approach using $\epsilon = 9.6$, and $\delta = 10^{-5}$

We also visualized the images generated by the models (Figure 2) . In the figure, the most left column shows the results for DP-CGAN, the left column represents the results for CGAN with no privacy, and the right column depicts the synthetic images generated by CGAN with basic differentially private approach. According to the figure, the quality of the images generated by DP-CGAN is better than CGAN with basic differentially private approach but worse than CGAN with no privacy.



((a)) ((b)) ((c))

Figure 2: (a) DP-CGAN, (b) CGAN with no privacy, (c) CGAN with basic differentially private approach

6. Conclusion

In this research, we proposed DP-CGAN framework that is a differentially private GAN model capable of generating both synthetic data and corresponding labels. The main idea behind DP-CGAN is that it clips the gradients of discriminator loss on real and fake data separately, sums up two sets of gradients, and adds Gaussian noise to the sum. DP-CGAN employs RDP account technique to track the spent privacy budget. The experimental results showed that DP-CGAN improves the performance compared to basic DP-CGAN and generates promising results on MNIST dataset.

The architectures we used for the generator and discriminator are rather simple. We are going to consider deep CGAN architectures with multiple convolutional layers to improve the quality of the synthetic data while spending the same privacy budget as we did for vanilla CGAN. Moreover, our results are still preliminary and we are going to show high quality differentially private CGANs on more challenging datasets such as CIFAR100 and CelebA/B. Finally, our preliminary results are very promising and we can extend our methodology to tackle the mentioned challenges.

References

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

[2] U. S. V. G. K. J. F. V. R. S. Abhradeep Thakurta, Andrew Vyrros and D. Davidson. Learning new words. *US Patent no. 9,594,741 B1*, 2017.

[3] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, 33(8):831, 2015.

[4] C. Angermueller, T. Pärnamaa, L. Parts, and O. Stegle. Deep learning for computational biology. *Molecular systems biology*, 12(7):878, 2016.

[5] G. Antipov, M. Baccouche, and J. Dugelay. Face aging with conditional generative adversarial networks. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 2089–2093, Sep. 2017.

[6] G. Antipov, M. Baccouche, and J.-L. Dugelay. Face aging with conditional generative adversarial networks. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 2089–2093. IEEE, 2017.

[7] B. K. Beaulieu-Jones, Z. S. Wu, C. Williams, R. Lee, S. P. Bhavnani, J. B. Byrd, and C. S. Greene. Privacy-preserving generative deep neural networks support clinical data sharing. *BioRxiv*, page 159756, 2018.

[8] Q. Chen, C. Xiang, M. Xue, B. Li, N. Borisov, D. Kaarfar, and H. Zhu. Differentially private data generative models. *arXiv preprint arXiv:1812.02274*, 2018.

[9] R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

[10] F. M. I. M. Cynthia Dwork, Krishnaram Kenthapadi and M. Naor. Our data,ourselves: Privacy via distributed noise generation. *EUROCRYPT*, 2006.

[11] K. N. Cynthia Dwork, Frank McSherry and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, page 265–284. Springer, 2006.

[12] C. Dwork and J. Lei. Differential privacy and robust statistics. In *STOC*, volume 9, pages 371–380, 2009.

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

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

[15] C. Dwork and G. N. Rothblum. Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*, 2016.

[16] C. Dwork, G. N. Rothblum, and S. Vadhan. Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 51–60. IEEE, 2010.

[17] M. Fredrikson, S. Jha, and T. Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1322–1333. ACM, 2015.

[18] L. Frigerio, A. S. de Oliveira, L. Gomez, and P. Duverger. Differentially private generative adversarial networks for time series, continuous, and discrete open data. *arXiv preprint arXiv:1901.02477*, 2019.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[20] A. Graves, A.-r. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE, 2013.

[21] H. Greenspan, B. Van Ginneken, and R. M. Summers. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging*, 35(5):1153–1159, 2016.

[22] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016.

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

[24] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[25] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[26] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, Jan. 2012.

[27] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[28] P. Kairouz, S. Oh, and P. Viswanath. The composition theorem for differential privacy. *IEEE Transactions on Information Theory*, 63(6):4037–4049, 2017.

[29] L. J. L. A. D. D. D. Killoran, Nathan and B. J. Frey. Generating and designing dna with deep generative models. *arXiv preprint arXiv:1712.06148*, 2017.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[31] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[32] I. Mironov. Renyi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275. IEEE, 2017.

[33] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[34] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. 518(7540):529, 2015.

[35] A. Myers and G. Nelson. Differential privacy: Raising the bar. *1 Geo. L. Tech. Rev.*, 1(1):135–142, November 2016.

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

[37] A. D. L. Phyllis E. Singer Daniel Kifer Jerome P. Reiter Ashwin Machanavajjhala Simson L. Garfinkel Scot A.Dahl Matthew Graham Vishesh Karwa Hang Kim Philip Leclerc Ian M. Schmutte William N. Sexton Lars Vilhuber Aref N. Dajani and J. M. Abowd. The modernization of statistical disclosure limitation at the u.s. census bureau. 2018.

[38] D. Ravı, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang. Deep learning for health informatics. *IEEE journal of biomedical and health informatics*, 21(1):4–21, 2017.

[39] D. Shen, G. Wu, and H.-I. Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221–248, 2017.

[40] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership inference attacks against machine learning models. In *Security and Privacy (SP), 2017 IEEE Symposium on*, pages 3–18. IEEE, 2017.

[41] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. 529(7587):484, Apr. 2016.

[42] V. P. U lfar Erlingsson and A. Korolova. Rappor: Randomized aggregatable privacypreserving ordinal response. *CCS*, 2014.

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

[44] J. Yoon, J. Jordon, and M. van der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. 2018.

[45] H. Zeng, M. D. Edwards, G. Liu, and D. K. Gifford. Convolutional neural network architectures for predicting dna–protein binding. *Bioinformatics*, 32(12):i121–i127, 2016.

[46] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.