Generating Synthetic Data for Deep Neural Network Classification Using Global Differential Privacy Based Optimization

Shalini Agarwal

Amity School of Engg. and Technology Amity University, Uttar Pradesh Lucknow, India,
Email: drsagarwal.04@gmail.com

Received: 10.04.2024 Revised : 12.05.2024 Accepted: 22.05.2024

ABSTRACT
Anonymizing individual text samples before dissemination, is an open research problem in Natural Language Processing (NLP). Significant efforts have been devoted to constructing such mechanisms by employing Local Differential Privacy (LDP) in the model training phase. However, LDP requires substantial noise in the update rule and often comes at the expense of the output language's quality. In this study, we address this limitation by introducing Global Differential Privacy (GDP). Specifically, we first train a generative language model in a differentially private manner and subsequently sample data from it. To do so, a novel idea of Prompt Variance Loss (PVL) is introduced, that enables the model to generate correct samples for a given instruction, thereby giving remarkable results. Experiments demonstrate that the synthetic datasets maintain privacy without leaking sensitive information from the original data as well as exhibit high suitability for training models and doing further analysis on real-world data. Notably, we show that training classifiers on private synthetic data outperforms directly training classifiers on real data with DP-SGD.

Keywords: Classifier, Differential Privacy, Synthetic Dataset, Privacy Budget, Prompt Variance Loss

INTRODUCTION

In the realm of machine learning, much emphasis has been laid on fostering trust, transparency, and accountability while deploying machine learning models across various domains and applications. Many authors have critically discussed the contrast between three fundamental aspects, interpretability, accuracy and privacy. The quest for interpretability and accuracy aims to uncover the inner workings of machine learning models, giving details about the rationale behind specific decisions, which not only builds trust in the model's predictions but also enables domain experts to understand and potentially act upon its outputs. However, this pursuit of interpretability and accuracy has a significant side-effect. As such, it can inadvertently breach privacy, because itunveils sensitive information about individual data points leading to compromising confidentiality and anonymity. Addressing this issue is a formidable challenge, particularly in the context of vast and complex datasets commonly encountered in contemporary applications. Despite its paramount importance, this issue has largely remained unexplored within the research landscape.

In order to address this interplay among interpretability, privacy, and accuracy, several authors have given insights about Differential Privacy (DP), a seminal step towards measuring ‘privacy loss’ given by Dwork and Roth 2014. In simple words, the original model of DP measures privacy loss by considering adjacent datasets that differ by at most one record. The concept of privacy loss measures the degree to which an algorithm's output reveals the presence of an individual datapoint within the dataset. In a nutshell, it offers two main advantages. First, it ensures post-processing invariance that says that applying any data-independent mechanism to a DP quantity does not alter the privacy level of the resulting quantity. Second, it ensures composability which states that combining DP quantities degrades privacy in a quantifiable way.

In the present paper, we employ DP based optimization to craft highly accurate and fluent synthetic dataset that embody specific desired attributes such as sentiment or topic, while faithfully replicating the statistical properties of the personal or sensitive data. Experiments reveal that our synthetic datasets maintain data integrity, exhibiting high language quality and suitability for training models for subsequent analysis on real-world datasets. Thus, the paper offers a promising avenue for the development of ethically grounded and socially beneficial deep neural network classification model with DP-SGD as we combine

state-of-the-art deep learning methods with advanced privacy-preserving mechanisms, training neural networks with the concept of privacy budget.

BACKGROUND AND LITERATURE REVIEW

In this section we give formal introduction of Differential Privacy and explain the steps required to make a machine learning algorithm differentially private. To put in formal terms, consider an algorithm M and neighboring datasets D and D' in such a manner that dataset D'is obtained from D by removing one datapoint from D. By doing so, it is clear that datasets D and D' differ by a single entry only. Further, let M(D) and M(D') denote outputs obtained by applying algorithm M on D and D' respectively. Under this context, privacy loss (Lp) is defined as –

$$\mathcal{L}^{(o)} = \log \frac{\Pr(M(D) = 0)}{\Pr(M(D') = 0)} \ldots\ldots\ldots (1)$$

where Pr(M(D')=0) denotes the probability that M returns a specific output 'o'. When the two probabilities in Eq. (1) exhibit similarity, even a highly proficient adversary, having knowledge of all entries in dataset D except for one, would find it challenging to identify the excluded datapoint, solely based on the algorithm's output. Conversely, if the two probabilities significantly diverge, it is straightforward to detect the absence of the single datapoint in datasetD' thereby successfully quantifying how revealing an algorithm's output is about the presence or absence of an individual entry in D.

An algorithm M is termed as (ϵ, δ)-Differentially Private if and only if $|L\mathcal{L}^{(o)}| \leq \epsilon, \forall o$ with probability at least $1 - \delta$.

Making an algorithm differentially private requires, introducing noise or randomness to the algorithm's computations in a controlled manner and ensuring that the presence or absence of an individual entry does not significantly alter the output. In general, following roadmap must be followed to achieve this –

A. Define Sensitivity

Sensitivity parameter (s) of the algorithm measures how much the output of the algorithm can alter when a single data point is present or absent in the input dataset.

B. Introduce Noise or Randomness

Noise or randomness is added to the algorithm's computations to mask individual contributions, based on the sensitivity of the algorithm and the desired level of privacy. This may be done using one of the following approaches –

(i) Laplace Mechanism: Laplace-distributed noise is added to the output of the algorithm, proportional to the sensitivity of the function being computed.

(ii) Gaussian Mechanism: The Gaussian-distributed noise is added to the output of the algorithm.

(iii) Randomized Response Mechanism: The randomized response mechanism is commonly used for binary-valued data. It introduces randomization to the data before processing.

C. Compute Cumulative Privacy Loss

Given a sequence of computations of the algorithm, compute cumulative privacy loss introduced at each step, adjust the noise accordingly to maintain overall differential privacy guarantees.

D. Evaluate Privacy Guarantee

Assess the effectiveness of the differential privacy mechanism by measuring the privacy loss and the desired privacy guarantee.

E. Iterate and Refine

Adjust the privacy parameters and noise levels based on feedback and optimize the balance between privacy and utility.

Some of the empirical works related to application of DP in neural networks was proposed in [3-5] where the concept of Differential Privacy Stochastic Gradient Descent (DP-SGD) for training deep neural networks with privacy guarantees has been introduced. An interesting study [6] explored DP in specific applications like human behavior prediction and text transformation, respectively. In [7] authors proposed privacy-and utility-preserving textual analysis via calibrated multivariate perturbations, while in [8], an overview of protecting user private-attribute information on social networks has been given. The idea of scalable and differentially private distributed aggregation in the shuffled model, authors in [9-10] demonstrated scaling DP-SGD to large distributed settings and making it applicable for training deep neural networks on massive and distributed datasets. A comprehensive overview of privacy threats in machine learning and discusses various privacy-preserving techniques, including DP-SGD has been presented in [11].

Wei et al. [12] investigated federated learning with differential privacy and Ma et al. [13] proposed RDP-GAN, a Rényi-differential privacy based generative adversarial network. [14] demonstrated the use of large language models as strong differentially private learners and authors have also explored privacy-

preserving text representation learning using BERT[15]. Technical contributions as given in [16] introduced differentially private SGD with non-smooth losses, while in [17] authors have discussed stability and implicit bias of gradient methods. Fully adaptive composition in differential privacy addresses challenges in maintaining privacy guarantees in complex workflows [18-20].
Overall, the literature review reveals that there is a wide scope in integration of differential privacy into deep neural network training, spanning theoretical foundations, algorithmic developments, practical applications, and emerging challenges. These studies collectively contribute to advancing the state-of-the-art in privacy-preserving deep learning classification model and pave the way for future research in this critical area.

Privacy Guarantee Using Differentially Private Stochastic Gradient Descent (DP-SGD)
In machine learning, Stochastic Gradient Descent (SGD) is considered as the most popular algorithm for minimizing the loss function to train the models. Rather than considering the entire dataset, SGD proceeds by updating parameters of the model by taking into account, either a single randomly chosen sample or a small subset from the dataset, at each iteration. Basic SGD gives understandable explanations for model decisions; however, does not concern about privacy preservation of training dataset. Differentially Private Stochastic Gradient Descent (DP-SGD) is the most important extension of basic SGD algorithm that ensures privacy guarantees for the training data by adding noise to the gradients computed during each iteration. This noise aids in protecting the privacy of individual data points thereby obscuring individual data contributions. By adding carefully calibrated noise to the gradients, DP-SGD ensures that the updates to the parameters do not reveal sensitive or private information about datasets.
The update rule for DP-SGD is given by the following formula:

$$\theta_{t+1} = \theta_t - \left(\eta \nabla L(\theta_t) + N(0, \sigma^2 I)\right) \ldots\ldots\ldots (2)$$

where –
(i) θ_t is the parameter vector iteration t
(ii) η is the learning rate
(iii) $\nabla L(\theta_t)$ is the gradient of the loss function L with respect to θ computed on a mini-batch of data.
(iv)$N(0, \sigma^2 I)$ is the noise term sampled from a Gaussian distribution with mean 0 and variance σ^2
(v)I is the identity matrix
To transform SGD into DP-SGD, certain algorithmic modifications are required to be incorporated in the underlying optimizer function. Each of these are discussed below-

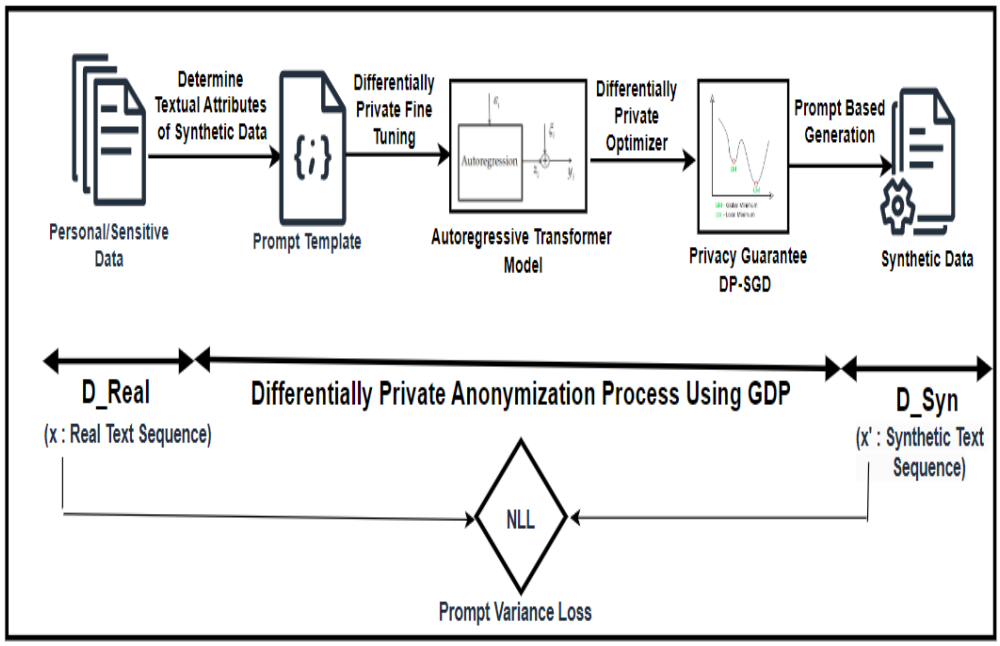

Figure 1. Model of Synthetic Data Generation using GDP optimizer

(A) Noise Addition
Incorporate noise into the gradient computation. This noise is drawn from a Gaussian distribution and scaled appropriately to achieve the desired level of privacy.

(B) Privacy Budget Management

Privacy budget is the mechanism to quantify maximum amount of privacy loss that is allowed during the model training process. This budget is directly proportional to the amount of noise added to the gradients and must be allocated judiciously to balance privacy and utility.

(C) Hyper parameter Tuning

DP-SGD introduces additional hyperparameters, such as the noise scale (σ) and privacy budget, thatdetermine the desired trade-off between privacy and model accuracy.

In summary, DP-SGD is preferred over SGD when privacy is a concern, as it provides strong guarantees against privacy breaches without sacrificing model performance significantly. By adding noise to the gradient computations and carefully managing the privacy budget, DP-SGD ensures that the trained models are robust against privacy attacks while still providing useful insights from the data.

Model of Synthetic Data Generation With Global Differential Privacy (GDP)

This section introduces a novel approach to anonymizing individual text samples using global differential privacy. The model proceeds by training a generative language model in a differentially private manner, which is then used for sampling data from it. By employing natural language prompts and introducing Prompt-Variance Loss (PVL), highly accurate and fluent textual datasets are created that possess desired attributes, while also resembling statistical properties of the training data.

This section gives deeper insights into the behavior and underlying mechanisms of the synthetic data generation model depicted in Fig. 1. Following are the components of the system pipeline –

(i) **Personal/Sensitive Data**- This comprises of large-scale data generated by individuals, often shared with third parties, for research purpose. However, this may involve numerous privacy risks which cannot be solely mitigated by pseudonymization. A variety of deanonymization attacks have been reported, that later allowed re-identification of individuals from tabulated data, such as movie ratings. An example is given in first two columns of table 1, that tabulates negative and positive sentiments for three product categories.

(ii) Prompt Template- Prompt template contains slots for verbal expressions of desired attribute values. A typical example of prompt template is shown in Table 1.

Table 1. Prompt Template

Generate **[Sentiment Type]** Review For **[Product Category]**	
Sentiment Type	Negative or Positive
Product Category	Home Appliance, Stationery, Apparel

(iii) Autoregressive Transformer- The Autoregressive Transformer model is a sophisticated architecture that can generate token sequences while remembering the context of previously generated tokens. Also, it allows parallelization during training leading to faster convergence, handles long-range dependencies and is scalable to process large datasets. The model operates by predicting the probability distribution of the next token given the preceding tokens using a parameterized function as given in equation 3.

$$P(y_t \text{ given } y < t) = \text{softmax}(f\theta(y < t)) \ldots\ldots\ldots (3)$$

where :

a) y_t is the token to be predicted
b) $y < t$ represents the token preceding y_t
c) $f\theta$ is a parametrized function for prediction
d) softmax is the function for output

By fine-tuning the pre-trained Autoregressive Transformer on our private dataset D_Real, the model can be tailored to perform optimally for synthetic text generation D_Syn. Fine-tuning involves adjusting the model's parameters$\theta$and refining them to better match the target dataset. Formally it is expressed as-

$$\theta_{\text{fine-tuned}} = \theta_{\text{pre-trained}} - \eta . \nabla\theta_{\text{pre-trained}} \mathcal{L} \ldots\ldots\ldots (4)$$

Where:

Table 2. Snapshot of Private and Synthetic Data

S.No.	Category (C)	Private / Sensitive Reviews (D_Real)	Synthetic Reviews by Prompting LLM (D_Syn)	Sentiment (T)
1.	Home Appliance	Disappointing performance	fails to meet expectations	Negative
2.	Home Appliance	Poor quality, constant malfunctions	Subpar craftsmanship, frequent breakdowns	Negative
3.	Home Appliance	Sleek design	Elegant aesthetic, adds sophistication to the home	Positive
4.	Home Appliance	Not worth the price, breaks easily	Overpriced for the quality, prone to premature failure	Negative
5.	Home Appliance	Innovative technology, top-notch	Cutting-edge technology, sets a new standard	Positive
6.	Home Appliance	Efficient and reliable	Effective and dependable, delivers consistent results	Positive
7.	Stationery	Variety of colors, perfect for creative projects	Wide range of hues, ideal for unleashing creativity	Positive
8.	Stationery	Ink smudges easily, ruining written work.	Smudging issues persist, compromising writing quality	Negative
9.	Stationery	High-quality paper, smooth writing experience	Luxurious paper texture, enhances writing experience	Positive
10.	Apparel	Comfortable fabric, perfect fit	Luxuriously soft fabric, contours to the body perfectly	Positive
11.	Apparel	Color faded quickly, looks worn out	Disappointingly fast color deterioration, worns prematurely	Negative
12.	Apparel	Thin material, feels cheaply made	Flimsy material, lacks durability and structure	Negative
13.	Apparel	Stylish design, received many compliments	Received numerous compliments on the chic design	Positive

a) η is the learning rate
b) $\nabla\theta_{\text{pre}-\text{trained}}\ \mathcal{L}$ is the gradient of the loss function $\mathcal{L}$

(iii) GDP-SGD – As elaborated in section III, Differentially Private Stochastic Gradient Descent (DP-SGD) ensures privacy guarantees for the training data by adding noise to the gradients computed during each iteration. DP-SGD ensures that the updates to the parameters do not disclose sensitive or private information about datasets. However, in the proposed model, we assume that the database operator is trustworthy and therefore introduce global differential privacy that leads to more accurate results with the same level of privacy protection. Also GDP is better suited for dataset of any scale.

(iv) Synthetic Data – The synthetic data is a secure data twin ideal for data sharing with mathematical privacy guarantees as granted by DP-SGD. An example is given in the last two columns of table 1. Through experimentation, the authors demonstrate that the synthetic datasets they generate do not leak information from the original data. Furthermore, they show that these synthetic datasets exhibit high language quality and are highly suitable for training models for further analysis on real-world data. The authors also illustrate that training classifiers on private synthetic data yields better performance compared to directly training classifiers on real data with DP-SGD.

(v) Prompt Variance Loss- In literature, Negative Log Likelihood (NLL) is recommended training objective for autoregressive language modeling during the parameter estimation process. Mathematically if $\mathcal{L}(\theta|x)$represents the likelihood function, where 'θ' is the parameter vector and 'x' is the observed data, then NLL function is given by

$\ell(\theta|x) = -\log \mathcal{L}(\theta|x) \ldots\ldots\ldots (4)$

This objective function encourages the model to generate correct data samples for a given prompt template, so that the D_Real and D_Syn text sequences are as similar as possible.

EXPERIMENTS AND RESULTS

(i) Dataset- In this study, we leveraged open source datasets namely, IMDb Movie Reviews and Amazon multiclass sentiment dataset for implementing the idea. The former consists of movie reviews written by various authors while the latter consists of more than two thousand product reviews from several categories including home appliances, apparels and stationery. The positive and negative sentiments, after converting into binary values are fed to the model for training. During the training stage, differentially

private optimizer is used to fine-tune our model in order to generate each text sample within the original dataset and based on the prompt corresponding to the desired attributes. This leads to generation of a private twin dataset that is derived using the same distribution of textual attributes as in the original dataset.

(ii) Privacy Parameter Setting-The experiments were conducted considering three different global privacy levels: 1)ϵ = 2, 2) ϵ = 8, and 3) ϵ = 10. It is observed that a smaller privacy parameter ensures stronger privacy but greater noise addition. An $\epsilon = \infty$ implies no privacy guarantee but also implies no noise. This non-private setting offers a baseline against which we assessed the performance of our model. The selection of these ε thresholds plays a significant role in experimentation and preliminary testing with the study dataset. In particular, careful attention was given to the selection of ϵ values so as to offer strong privacy but at the same time, not too strict that the added noise leads to completely useless synthetic data generation.

(iii)Accuracy vs Privacy trade-off- Figure 2 show the trade-off between accuracy and privacy strength of the proposed model. Ten random subsets of increasing size, ranging from 100 to 1000 are taken for two sets of experiments each 1) training done on real/private data 2) training done on synthetic data. The model was trained over five epochs while using a differentially private optimizer, with a smaller learning rate each time. Educated conjectures over a couple of iteration was applied, for setting the value of hyperparameter so as to avoid large computational effort. Following observations immediately come from figure 2. The synthetic or twin datasets leads to a classification model that is as accurate as the one trained on real data. Even for small samples, the model can synthesize datasets that lead to high-performing classifier because we employed pretrained knowledge through our prompts. Therefore, apart from anonymization of sensitive datasets, our model can also be utilized for enlarging small datasets with a privacy guarantee.

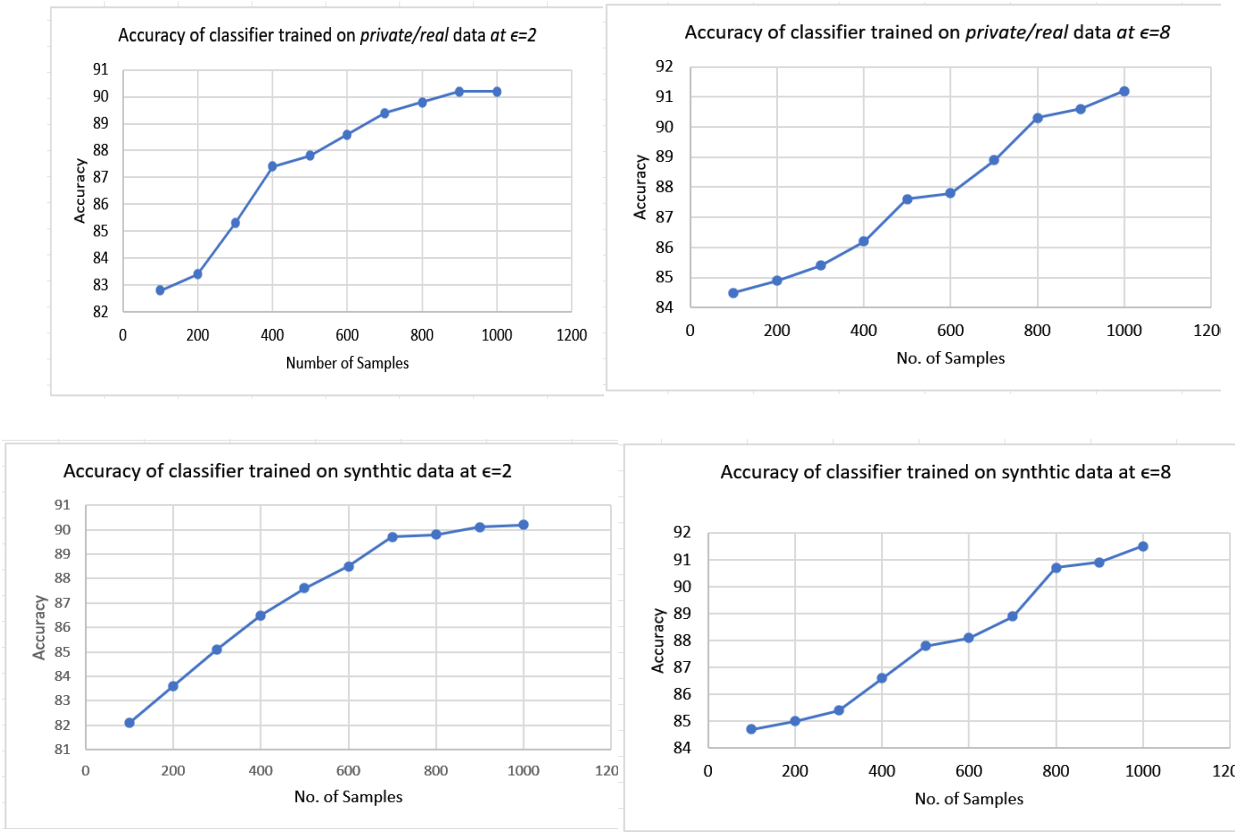

Figure 2. Accuracy of Classifier at different values of ϵ for different sample sizes

CONCLUSION AND FUTURE DIRECTION

This paper demonstrates creation of synthetic datasets using differentially private optimization method for publicly sharing textual data while safeguarding user's privacy. The experiments indicate that synthetic dataset generated has high quality and serves effectively as training data for subsequent classification tasks while diminishing the potential for original data leakage. Thus the model finds extensive applications in all domains where sensitive data is involved. We examine the effects of privacy levels on different

samples and show that the model is more tolerant of the higher levels of noise than the smaller values. Nevertheless, several open questions for future research still need exploration. Although DP is based on the ground that the inclusion or exclusion of a single individual should not significantly change the results of any analysis conducted on the dataset, however models trained with DP-SGD are noisier than the benchmark because of successive gradient clipping and addition of noise.

As a future work, the model shall be improvised for generating synthetic visual data using DP-SGD mechanism. For high-dimensional inputs such as images, it is expected that adding noise to the gradient shall lead to very low accuracies for private training. Hence, incorporate random projection matrices may be a good future idea to investigate.

REFERENCES

[1] Dwork, C. "Differential Privacy". In: van Tilborg, H.C.A., Jajodia, S. (eds) Encyclopedia of Cryptography and Security. Springer, Boston, MA. https://doi.org/10.1007/978-1-4419-5906-5_752 (2011)

[2] Dwork, C., and Roth, A. "The algorithmic foundations of differential privacy". Found. Trends Theor. Comput. Sci. 9:211–407 (2014)

[3] Chaudhuri, Kamalika, Monteleoni, Claire, and Sarwate, Anand D. "Differentially private empirical risk minimization", The Journal of Machine Learning Research, 12:1069–1109, (2011).

[4] Martín Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, Li Zhang, "Deep Learning with Differential Privacy", Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (ACM CCS), pp. 308-318, (2016)

[5] Phan, N.; Wang, Y.; Wu, X.; and Dou, D. "Differential privacy preservation for deep auto-encoders: An application of human behavior prediction", In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI'16, 1309–1316. AAAI Press. (2016)

[6] Satyapriya Krishna, Rahul Gupta, and Christophe Dupuy. "ADePT: Auto-encoder based differentially private text transformation" In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 2435–2439, Online. Association for Computational Linguistics, 2021

[7] Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. "Privacy-and utility-preserving textual analysis via calibrated multivariate perturbations", In Proceedings of the 13th International Conference on Web Search and Data Mining, pages 178–186 (2020)

[8] Alnasser, W., Beigi, G., Liu, H., "An overview on protecting user private-attribute information on social networks", In: Cruz-Cunha, M.M., Mateus-Coelho, N.R. (eds.) Handbook of Research on Cyber Crime and Information Privacy, Chap. 6 (2020)

[9] Ghazi, Badih, Rasmus Pagh, and Ameya Velingker. "Scalable and differentially private distributed aggregation in the shuffled model." arXiv preprint arXiv:1906.08320 (2019).

[10] Al-Rubaie, Mohammad & Chang, J.. "Privacy-Preserving Machine Learning: Threats and Solutions", IEEE Security & Privacy. 17. 49-58. 10.1109/MSEC.2018.2888775. (2019)

[11] Fredrikson, M.; Jha, S.; and Ristenpart, T. "Model inversion attacks that exploit confidence information and basic countermeasures", In Proceedings of the 22Nd ACM SIGSAC Conference on Computer and Communications Security, CCS '15, 1322–1333. New York, NY, USA: ACM. (2015)

[12] Wei, Kang & Li, Jun & Ding, Ming & Ma, Chuan & Yang, Howard & Farokhi, Farhad & Jin, Shi & Quek, Tony Q.S. & Poor, H. Vincent. "Federated Learning With Differential Privacy: Algorithms and Performance Analysis", IEEE Transactions on Information Forensics and Security. PP. 1-1. 10.1109/TIFS.2020.2988575. (2020)

[13] Ma, Chuan & Li, Jun & Ding, Ming & Liu, Bo & Wei, Kang & Weng, Jian & Poor, H. Vincent. "RDP-GAN: A Rényi-Differential Privacy Based Generative Adversarial Network", IEEE Transactions on Dependable and Secure Computing. PP. 1-15. 10.1109/TDSC.2022.3233580.(2023)

[14] Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori Hashimoto. 2021. Large language models can be strong differentially private learners

[15] Walaa Alnasser, Ghazaleh Beigi, and Huan Liu. 2021. Privacy preserving text representation learning using bert. In Social, Cultural, and Behavioral Modeling, pages 91–100, Cham. Springer International Publishing.

[16] Puyu Wang, Yunwen Lei, Yiming Ying, and Hai Zhang. Differentially private sgd with non-smooth losses. Applied and Computational Harmonic Analysis, 56:306–336, 2022.

[17] Matan Schliserman and Tomer Koren. Stability vs implicit bias of gradient methods on separable data and beyond. In Conference on Learning Theory, pages 3380–3394. PMLR, 2022.

[18] Justin Whitehouse, Aaditya Ramdas, Ryan Rogers, and Zhiwei Steven Wu. Fully adaptive composition in differential privacy. arXiv preprint arXiv:2203.05481, 2022.

[19] Da Yu, Gautam Kamath, Janardhan Kulkarni, Tie-Yan Liu, Jian Yin, Huishuai Zhang, "Individual Privacy Accounting for Differentially Private Stochastic Gradient Descent", https://doi.org/10.48550/arXiv.2206.02617 (2023)
[20] Puyu Wang, Yunwen Lei, Yiming Ying, Ding-Xuan Zhou, "Differentially Private Stochastic Gradient Descent with Low-Noise", https://doi.org/10.48550/arXiv.2209.04188 (2023)