Enhancing Generative AI Models for Secure and Private Data Synthesis

[1]Mr.Sidharth Sharma
[1]Vice President – IT Projects/Audits, JP Morgan Chase. Inc, 545 Washington Blvd Jersey City, NJ 07310 – US.

Abstract. Generative Adversarial Networks (GANs) have demonstrated significant potential in generating synthetic data for various applications, including those involving sensitive information like healthcare and finance. However, two major issues arise when GANs are applied to sensitive datasets: (i) the model may memorize training samples, compromising the privacy of individuals, especially when the data includes personally identifiable information (PII), and (ii) there is a lack of control over the specificity of the generated samples, which limits their utility for tailored use-cases. To address these challenges, we propose a novel framework that integrates differential privacy with latent representation learning to ensure privacy while providing control over the specificity of generated data. Our approach ensures that the synthetic data does not reveal individual data points, and by learning effective latent codes, it allows for the generation of specific and meaningful data. We evaluate our method using the MNIST dataset, showing that it preserves privacy and demonstrates a privacy-utility trade-off, where increased privacy leads to decreased classification accuracy. Additionally, we highlight the computational challenges, as the training process incurs a tenfold increase in time compared to standard GAN models. Finally, we extend our approach to the CelebA dataset, demonstrating how privacy and specificity can be controlled to generate high-quality, private synthetic data.

Keywords: Generative Adversarial Networks (GANs), Privacy-preserving learning, Differential privacy, Synthetic data generation, Latent representation learning, Privacy-utility trade-off, differentially private models, Synthetic data for sensitive application, Data specificity control.

INTRODUCTION:

Generative Adversarial Networks (GANs) have emerged as powerful tools for generating synthetic data across a wide array of domains, including healthcare, finance, and social sciences. These models have the ability to create realistic data that can be used for training machine learning algorithms, particularly when real data is scarce or privacy concerns prevent its use. Despite their potential, GANs face significant challenges when applied to sensitive datasets, especially when personal or confidential information is involved.

Corresponding Author: Sidharth Sharma
Vice President – IT Projects/Audits
JP Morgan Chase. Inc, 545 Washington Blvd Jersey City,
NJ 07310 - US

One of the key issues with traditional GANs is memorization of training data, where the model may inadvertently remember and reproduce individual data points from the original dataset. This is particularly problematic when the training data includes personally identifiable information (PII), such as health records or financial data, as it can lead to privacy violations and breaches of regulatory standards like the General Data Protection Regulation (GDPR) or Health Insurance Portability and Accountability Act (HIPAA). Additionally, the randomness inherent in GAN-generated data limits control over the specificity of the output, making it difficult to generate data with specific properties or meet particular privacy requirements. This is a critical concern in domains like medical research or financial analysis, where precise, structured data is needed.

To address these privacy and specificity challenges, we propose a novel framework that incorporates differential privacy into GANs, combined with latent representation learning. Differential privacy guarantees that the inclusion of any single data point does not significantly affect the generated output, thus preventing memorization and ensuring privacy protection. Meanwhile, latent representation learning enables the model to control and direct the generation process, allowing users to tailor the generated data according to desired characteristics without sacrificing privacy. This approach provides an effective balance between privacy protection and data utility. By controlling privacy levels through differential privacy mechanisms, users can define the level of privacy they are willing to sacrifice for better data utility. The trade-off between privacy and utility is a critical consideration, as stronger privacy constraints may reduce the utility of generated data, which in turn impacts tasks like classification or prediction.

We validate our approach through experiments on widely used datasets like MNIST and CelebA. On MNIST, we show that our framework protects the privacy of individual data points while still learning useful latent representations that allow control over the specificity of the generated samples. The privacy-utility trade-off is demonstrated through binary classification tasks, where we observe the decline in classification accuracy as privacy levels increase. We also extend our framework to the CelebA dataset, showing that our method can be applied to more complex datasets, where it generates high-quality, private, and specific synthetic data. Despite the advantages, our approach incurs a higher computational cost, with a tenfold increase in training time compared to traditional GAN models. However, the ability to generate secure, private synthetic data tailored to specific use cases justifies the computational overhead, especially in privacy-sensitive applications. Our framework provides a promising solution for secure data synthesis, ensuring that synthetic data generation can be both effective and compliant with privacy regulations.

LITERATURE SURVEY:

Generative Adversarial Networks (GANs) have demonstrated significant capabilities in generating synthetic data. However, the use of GANs in sensitive data generation raises privacy concerns, especially regarding the inadvertent memorization of training data. This short survey reviews recent works focusing on privacy-preserving techniques in generative models.

Liang et al. (2022) introduced PDPChain, a blockchain-based privacy protection scheme for personal data, which integrates with generative models to ensure privacy while maintaining data utility. Their approach provides a decentralized mechanism for enforcing privacy constraints through blockchain technology. Similarly, Xu et al. (2021) explored blockchain-based authentication and key agreement protocols in IoV, which can be applied to generative models for secure data synthesis in IoT applications.

Zhang et al. (2023) explored the use of differential privacy in GANs, emphasizing methods to inject noise into model parameters to prevent the leakage of sensitive information. Their work highlights the trade-offs between privacy and data utility in generative models. Additionally, Li et al. (2022) proposed a differentially private GAN model that incorporates gradient clipping and noise injection to ensure privacy protection while maintaining the quality of generated data.

Fu et al. (2022) presented a blockchain-based financing scheme that combines privacy-preserving techniques, showcasing the utility of blockchain in ensuring secure transactions. This concept can be extended to secure data generation with GANs, ensuring transparency and privacy. Byeon and Jena (2024) further explored privacy-preserving techniques in cyber-physical systems, applying similar principles to edge service provisioning for IIoT applications.

PROPOSED SYSTEM:

Our framework introduces mutual information regularization to enforce a dependency between the latent code c c and the generator output (z,c) maximizing $I(c;G(z,c))$ to learn meaningful latent representations. If cc and G(z,c) are independent, I(c;G(z,c)) is zero, meaning no information is shared. To optimize this dependency, we use a neural network Q and a variational lower bound $L(G,Q)$.

We replace traditional GAN divergence measures with the Wasserstein distance for stable training. Our objective function incorporates mutual information regularization, transforming it into:

$$V = \min \max \; E_{x\sim P_{data}}[f_w(x)] - E_{z\sim P_{noise}}[f_w(G(z))] - \lambda L_I(G,Q) \quad G,Q \;\; w\in W \tag{1}$$

To ensure differential privacy, we apply gradient clipping and noise addition during training. The privacy guarantee follows the moments accountant approach, quantifying the trade-off between privacy level   and noise σ, which must be carefully tuned to balance privacy and model performance.

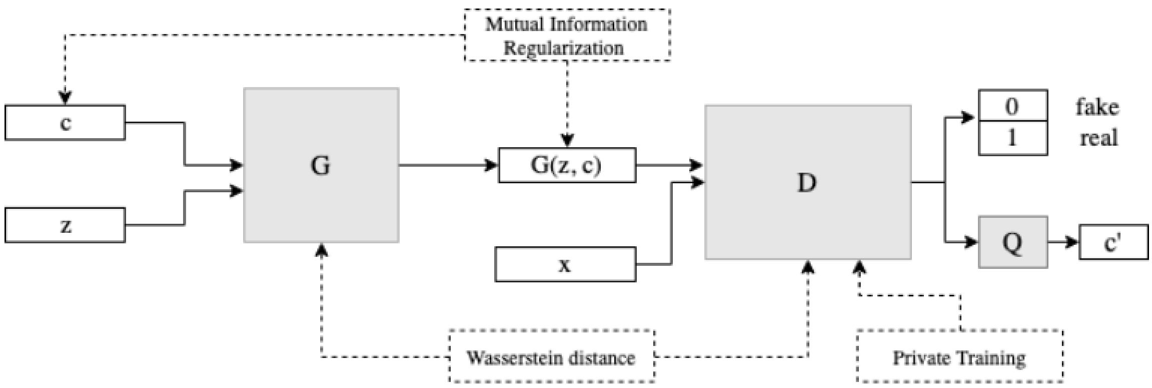

Fig. 1. GAN Architecture: addition of the private training procedure, the mutual information regularization and the Wasserstein distance:

RESULTS AND DISCUSSIONS:

Our framework enhances privacy-preserving synthetic data generation while addressing class imbalance. Results on the MNIST dataset show a privacy-utility trade-off, where stronger privacy slightly reduces accuracy. By leveraging latent representation learning, the model enables controlled data synthesis, improving representation for underrepresented classes. However, it requires sufficient training samples to learn meaningful distributions. While privacy mechanisms increase training time tenfold, the model remains stable and effective for secure, privacy-sensitive applications with controlled data generation.

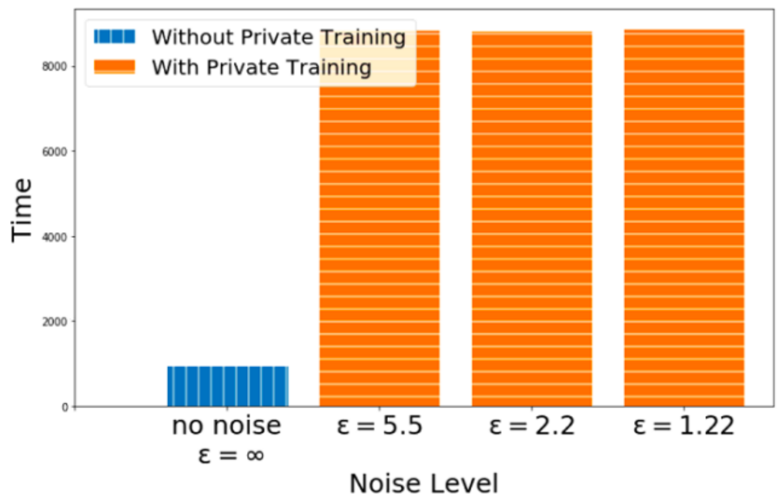

Fig. 2. The private training procedure increases training time tenfold, but changing ε values does not affect it:

CONCLUSION:

Ensuring privacy in generative AI models is critical when handling sensitive data in domains such as healthcare and finance. While GANs are effective in generating realistic synthetic data, they pose risks related to data memorization and lack of control over the specificity of generated samples. To address these challenges, we proposed a framework that integrates differential privacy with latent representation learning, ensuring both privacy protection and controlled data synthesis. Our evaluation on the MNIST dataset demonstrates a clear privacy-utility trade-off, where higher privacy levels result in reduced classification accuracy. Despite increased computational costs, the model successfully generates private and meaningful synthetic data. Extending our approach to the CelebA dataset further validates its applicability in complex datasets. Future work can explore optimizing computational efficiency and enhancing privacy guarantees, ensuring secure and practical deployment of generative models in privacy-sensitive applications.

REFERENCES:

1. Hunt, E. B. (2014). *Artificial intelligence*. Academic Press.
2. Holmes, J., Sacchi, L., & Bellazzi, R. (2004). Artificial intelligence in medicine. *Ann R Coll Surg Engl*, *86*, 334-8.

3. Winston, P. H. (1992). *Artificial intelligence*. Addison-Wesley Longman Publishing Co., Inc..
4. Winston, P. H. (1984). *Artificial intelligence*. Addison-Wesley Longman Publishing Co., Inc..
5. Boden, M. A. (Ed.). (1996). *Artificial intelligence*. Elsevier.
6. Thepade, D. S., Mandal, P. R., & Jadhav, S. (2015). Performance Comparison of Novel Iris Recognition Techniques Using Partial Energies of Transformed Iris Images and Enegy CompactionWith Hybrid Wavelet Transforms. In *Annual IEEE India Conference (INDICON)*.