ESP-IJACT
ESP International Journal of Advancements in Computational Technology
ISSN: 2583-8628 / Volume 3 Issue 1 January 2025 / Page No: 1-4
Paper Id: IJACT-V3I1P1011 / Doi: 10.56472/25838628/IJACT-V3I1P101

Original Article

A Review of AI-Based Synthetic Data Generation Approaches

Anurag Bhagat
Senior Strategist, GenAI Innovation Center, Amazon Web Services (AWS), Sunnyvale, CA, USA.

Received Date: 06 November 2024 Revised Date: 15 December 2024 Accepted Date: 03 January 2025

Abstract: *Creating synthetic data, which closely resembles real data, using AI based techniques is becoming increasingly important in solving machine learning problems across the entire lifecycle of ML from training to tuning and testing. Synthetic data can solve multiple limitations like data being scarce or unavailable, data privacy concerns like in healthcare scenarios with PII and PHI data, or can just speed up the AI model development journey by providing fast access to data while the real data is being prepared. This review paper provides a view into various methodologies and key advancements in synthetic data creation with some examples, with a special focus on Generative AI based techniques which have really made this more accessible to a lot of people.*

Keywords: *Artificial Intelligence, Generative AI, Synthetic Data, AI/ML, GenAI.*

I. INTRODUCTION

We should first clarify two distinct terms which are often confused. It is important to distinguish between AI generated synthetic data and mock data.

- AI generated synthetic data is sample based. In order to generate it, you need a large enough sample of AI models to learn from. The resulting synthetic data has all the statistical properties of the original data. It looks real, but is completely artificial.
- Mock data does not need data samples. It could be generated based on simple rules (which could be as simple as random number generator), and may have very different statistical properties of the original dataset.

AI generated synthetic data generation approaches have gained significant attention recently, especially as Generative AI techniques have become more mainstream in the last few years. This has helped organizations which had been facing challenges around data privacy (e.g., PII/PHI data), data being scare or unavailable/imbalanced, or speed of data availability. This enables model building across all phases, from model training, hyperparameter tuning as well as testing without being constrained by hard-to-obtain datasets. This peer review work discusses recent advancements in this field, showcasing some practical applications and art-of-the-possible techniques.

II. APPLICATIONS OF SYNTHETIC DATA

A. Training Machine Learning Models

Having a large enough training dataset is sometimes a challenge to effective training. Synthetic data addresses this issue by expanding the data available for training models.

For example, Antoniou et al. (2019) proposed data augmentation generative adversarial networks (DAGANs), which takes data from a source domain and learns to take any data item and generalize it to generate other within-class data items. As this generative process does not depend on the classes themselves, it can be applied to novel unseen classes of data.

B. Testing and validation

Testing on diverse datasets ensures model robustness. Evaluating the performance of machine learning models on diverse and underrepresented subgroups is essential for ensuring fairness and reliability in real-world applications. However, accurately assessing model performance becomes challenging due to two main issues: (1) a scarcity of test data, especially for small subgroups, and (2) possible distributional shifts in the model's deployment setting, which may not align with the available test data. Breugel et al (2023) improved model evaluation by generating synthetic test sets for small subgroups and simulating distributional shifts. Hu et al (2023) used synthetic data as validation to improve AI robustness in both in-domain and out-domain test sets for early cancer detection in computed tomography (CT) volumes.

III. TECHNIQUES FOR SYNTHETIC DATA GENERATION

A. Generative Adversarial Networks (GANs)

GANs are composed of two opposing networks—the generator and the discriminator—trained in tandem in an adversarial process to produce realistic synthetic data.

This is an open access article under the CCBY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/2.0/)

Goodfellow et al. outlined the foundational architecture in 2014, which has since been extended and modified by many to enable use numerous situations.

a) Conditional Generative Adversarial Networks (CGAN):

Conditional Generative Adversarial Networks (CGANs) are an important extension of the original GAN. They contain two adversarial networks, a generator and a discriminator, both conditioned on some additional information. The extra information can be, for instance, class labels, feature correlations, or other auxiliary information.

b) CTGAN:

Tabular data usually contains a mix of discrete and continuous columns. Continuous columns may have multiple modes whereas discrete columns are sometimes imbalanced making the modelling difficult. To solve these challenges with structured datasets, GAN variations models like CTGAN (Xu et al., 2019) have been proposed to generate realistic tabular data.

c) Wasserstein Generative Adversarial Networks (WGAN):

Wasserstein Generative Adversarial Networks (WGANs) were proposed to improve the stability of learning in generative models. As opposed to the original GANs, WGANs introduce a critic network instead of a discriminator. A discriminator in the original GANs uses the probability estimation to distinguish between the real and generated data. The critic, on the other hand, estimates the distribution of both the real and generated data, and then minimizes the Wasserstein distance between them as a metric. This optimization increases the stability of the generative model and improves the quality of the generated data.

d) Other Variants:

GAN variants like CycleGAN (Zhu et al., 2017) and StyleGAN (Karras et al., 2019) have extended applications in image and video synth.

B. Variational Autoencoders (VAEs)

Kingma and Welling introduced VAEs, which model data distributions to generate realistic synthetic datasets. These are the second most common technique for synthetic data generation and have also seen multiple variants.

Khadka et al (2023) introduced a synthetic data generation approach called CT-VAE that uses Combinatorial Testing (CT) and Variational Autoencoder (VAE). We first use VAE to learn the distribution of the real-world data and encode it in a latent, lower-dimensional space. Next, we use CT to sample the latent space by generating a t-way set of latent vectors, each of which represents a data point in the latent space

Ally Salim Jr (2018) trained a VAE on available patient data to learn the latent distribution of the patient features given the diagnosis. This latent distribution would then be sampled to generate new accurate patient records given the patient diagnosis.

VAE architectures have also been proposed by Desai et al (2021) to generate time series data, with several distinct properties: interpretability, ability to encode domain knowledge, and reduced training times

IV. CASE STUDIES

A. Autonomous Vehicles

Synthetic data plays an extremely important role in training autonomous vehicles, as real-world data collection is not just expensive and hazardous, but would limit the kind of scenarios which the algorithm would be exposed to. NVIDIA's Omniverse Cloud Sensor RTX combines real-world data from various sensors with synthetic data, supposedly allowing developers to test sensor perception and associated AI software in realistic virtual environments before real-world deployment. The company claims this approach will enhance safety, reduce costs and save time in the development process. Tesla also is believed to be using synthetic data to provide more thorough training for its autonomous driving AI models.

B. Healthcare

Use of synthetic data is especially important in healthcare owing to data privacy concerns. It allows easy utilization and sharing of data for internal use while ensuring that identities cannot be linked back to specific records or used for re-identification purposes. This also makes sure that the organizations remain compliant with regulations such as GDPR and HIPAA throughout the process.

There are already some popular synthetic data sources that are easily available and widely used for medical research- DE-SynPUF files published by CMS, SyntheticMass and the US Synthetic Household Population database.

In 2021, a team from the Institute for Informatics at Washington University School of Medicine in St. Louis demonstrated synthetic data's potential to protect privacy while conducting clinical studies. Zhang et al in 2021 developed

SynTEG, a GAN based framework for generating synthetic EHR(Electronic Health Records) data. This synthetic data mimics the statistical properties of real patient records and is used for training clinical decision-support systems.

V. CHALLENGES AND LIMITATIONS

A. Data Quality

Poorly generated synthetic data can lead to inaccurate and unreliable AI models. Depending on the methodology employed, the integrity of synthetic data can vary. Frequently, the data generated by a generative adversarial network (GAN) are highly realistic, but it can be challenging to control their distribution. Statistical models can generate more evenly distributed data, but the resulting data may be less plausible.

B. Security

Synthetic data, if reverse-engineered, could potentially reveal information about the underlying real data or the process used to generate it, posing security risks. Re-identification is therefore a real risk for synthetic data, especially if the source data used is published with the synthetic data, or if the model used to create the synthetic data "overfits" the training data, meaning that it too closely resembles the original dataset.

C. Bias Amplification

Use of synthetic data carries concerns such as the risk of bias amplification, low interpretability, and an absence of robust methods for auditing data quality. Bias is defined as a systematic discrepancy or persistent deviation that originates during the data sampling or testing process. Consequently, if the primary dataset used to generate synthetic data carries inherent biases, the synthetic data could unintentionally magnify these biases. This can result in misinformed or discriminatory outcomes, which can have serious downstream implications when used in critical domains like healthcare and financial services.

VI. CONCLUSION

AI-based synthetic data approaches leveraging GAN and VAE have already revolutionized training, tuning, and testing of ML models by providing cheap access to synthetic data. While challenges persist, advancements in GANs, VAEs, and other techniques continue to push the boundaries of what synthetic data can achieve. There are applications across industries including autonomous vehicles, healthcare and financial services. By addressing limitations and exploring new frontiers, synthetic data will remain a critical asset in AI and ML research.

VII. REFERENCES

[1] Antoniou, A., Storkey, A., & Edwards, H. (2019). Data Augmentation Generative Adversarial Networks. arXiv preprint arXiv:1711.04340.

[2] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A Simple Framework for Contrastive Learning of Visual Representations. Proceedings of the 37th International Conference on Machine Learning.

[3] Qixin Hu, Alan Yuille, Zongwei Zhou(2023), Synthetic Data as Validation https://arxiv.org/abs/2310.16052

[4] Boris van Breugel, Nabeel Seedat, Fergus Imrie, Mihaela van der Schaar (2023), Can You Rely on Your Model Evaluation? Improving Model Evaluation with Synthetic Test Data. Neurips 2023

[5] Tshilidzi Marwala, Eleonore Fournier-Tombs, Serge Stinckwich(2023), The Use of Synthetic Data to Train AI Models: Opportunities and Risks for Sustainable Development, https://arxiv.org/pdf/2309.00652

[6] Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., & Sun, J. (2017). Generating Multi-label Discrete Patient Records using Generative Adversarial Networks. arXiv preprint arXiv:1703.06490.

[7] Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al. (2014). Generative Adversarial Networks. Advances in Neural Information Processing Systems.

[8] Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling Tabular Data using Conditional GAN. Advances in Neural Information Processing Systems.

[9] Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. Proceedings of the IEEE International Conference on Computer Vision.

[10] Abedi, Hempel, Sadeghi, Kirsten (2022). GAN-Based Approaches for Generating Structured Data in the Medical Domain. Appl. Sci., 12(14), 7075; https://doi.org/10.3390/app12147075

[11] Mirza, M.; Osindero, S. Conditional Generative Adversarial Nets. arXiv 2014, arXiv:1411.1784. [Google Scholar]

[12] Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein GAN. arXiv 2017, arXiv:1701.07875.

[13] M. Razghandi, H. Zhou, M. Erol-Kantarci and D. Turgut, "Variational Autoencoder Generative Adversarial Network for Synthetic Data Generation in Smart Home," ICC 2022 - IEEE International Conference on Communications, Seoul, Korea, Republic of, 2022, pp. 4781-4786, doi: 10.1109/ICC45855.2022.9839249.

[14] K. Khadka, J. Chandrasekaran, Y. Lei, R. N. Kacker and D. Richard Kuhn, "Synthetic Data Generation Using Combinatorial Testing and Variational Autoencoder," 2023 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW), Dublin, Ireland, 2023, pp. 228-236, doi: 10.1109/ICSTW58534.2023.00048

[15] Ally Salim Jr (2018); Synthetic Patient Generation: A Deep Learning Approach Using Variational Autoencoders. arXiv:1808.06444, https://doi.org/10.48550/arXiv.1808.06444

[16] Abhyuday Desai, Cynthia Freeman, Zuhui Wang, Ian Beaver, TimeVAE: A Variational Auto-Encoder for Multivariate Time Series Generation arXiv:2111.08095, https://doi.org/10.48550/arXiv.2111.08095
[17] https://www.pymnts.com/artificial-intelligence-2/2024/nvidias-new-ai-simulator-could-rev-up-robotics-self-driving-cars/
[18] Tesla's filed patent, Data Synthesis for Autonomous Control Systems https://ppubs.uspto.gov/pubwebapp/
[19] Ziqi Zhang, Chao Yan, Thomas A Lasko, Jimeng Sun, Bradley A Malin, SynTEG: a framework for temporal structured electronic health data simulation, Journal of the American Medical Informatics Association, Volume 28, Issue 3, March 2021, Pages 596–604, https://doi.org/10.1093/jamia/ocaa262