



GENERATING SYNTHETIC DATA FOR TRAINING ARTIFICIAL INTELLIGENCE

Xudoyorov Shaxriyor

*Termez University of Economics and Service, Faculty of Economics and Information Technologies,
Information Systems and Technologies Department, 3rd year student*

Scientific supervisor: Ernazarov Mirzohid

E-mail: xudoyorovshaxriyor57@gmail.com

Annotation: This scientific article provides a comprehensive analysis of the generation of synthetic (artificially created) data for training artificial intelligence (AI) systems and the prospects of this direction. Modern AI models require enormous amounts of data for training, but collecting real data is often expensive, time-consuming, or impossible due to privacy concerns. The research scientifically substantiates the methods of creating artificial data that resembles real data (generative models, simulations), its advantages, and its limitations. The article examines the use of synthetic data in solving privacy problems, eliminating data shortages, and balancing datasets. The scientific novelty of the article lies in demonstrating that synthetic data is becoming an important tool for the development of AI. As a result of the analyses, recommendations are developed regarding the application of these technologies and their reliability.

Keywords: synthetic data, data generation, artificial intelligence, generative models, GAN, data privacy, simulation, machine learning, data augmentation, neural networks, training data, data scarcity.

Introduction

Data is the "fuel" of artificial intelligence. Modern AI models, especially deep learning models, require enormous amounts of data for training. The more high-quality data a model is trained on, the more accurate and reliable it becomes. However, collecting real data faces a number of serious problems. First, cost and time: collecting and labeling a large amount of data is expensive and time-consuming. Second, privacy: data such as medical records and financial information is confidential, and its use is restricted. Third, data shortage: in some areas (rare diseases, unusual situations), there is simply not enough real data. Fourth, imbalance: a dataset may not adequately represent some groups or situations. Synthetic data generation emerged as a solution to these problems.

Synthetic data is artificially created data that resembles real data but does not relate to actual people or events. For example, instead of using the real medical records of actual patients, it is possible to create artificial records that have the same statistical properties but do not belong to any



real person. Such data preserves privacy and at the same time is suitable for training models. Synthetic data is created using generative AI models or computer simulations.

The relevance of this article is that the demand for data for training AI is constantly growing, while the problems of collecting real data (cost, privacy, shortage) are becoming more acute. Synthetic data offers an effective solution to these problems. In the context of Uzbekistan as well, where the collection of real data may be limited, the use of synthetic data is relevant for the development of AI. In this article, we analyze the methods of generating synthetic data, its advantages, and its limitations.

Discussion

The generation of synthetic data is one of the rapidly developing topics in the AI community. The main issue under discussion is how well synthetic data can replace real data.

First, the issue of the quality of synthetic data. For synthetic data to be useful, it must accurately reflect the statistical properties and complexity of real data. If the synthetic data is too simple or differs from reality, a model trained on it will not work well on real data. Therefore, assessing the quality of synthetic data is an important problem.

Second, the issue of preserving privacy. One of the main goals of synthetic data is to preserve privacy. However, the question arises: can the original real data be partially reconstructed from synthetic data? If the synthetic data is too closely tied to the real data, privacy may not be fully protected. Therefore, methods that combine synthetic data generation with privacy guarantees (such as differential privacy) are being developed.

Third, the risk of perpetuating bias. If the real data used to create synthetic data contains biases, the synthetic data may also reproduce these biases. Therefore, the issue of the fairness of synthetic data is also relevant.

Main Part

1. Methods of generating synthetic data.

Synthetic data is created using several methods. Generative models (such as generative adversarial networks, GANs) learn the statistical properties of real data and create new data with similar properties. Simulations create artificial data by modeling real-world processes — for example, generating images of various traffic situations to train autonomous vehicles. Statistical methods, in turn, create new data based on the statistical distribution of the data. Each method has its own advantages and is suitable for different tasks.

2. Solving privacy problems.



One of the most important applications of synthetic data is solving privacy problems. In fields such as medicine and finance, real data is confidential, and its use is restricted. Synthetic data resembles real data in its statistical properties but does not relate to actual people. This makes it possible to develop and test AI models without violating privacy. For example, instead of using the medical records of real patients, models can be trained on synthetic records.

3. Eliminating data shortages.

In some areas, there is simply not enough real data. For example, data on rare diseases or unusual situations (a rare type of accident) is scarce. Synthetic data helps to fill this gap. By artificially creating examples of rare situations, it makes it possible to train models more comprehensively. This is especially important for safety-critical systems (medicine, autonomous vehicles).

4. Balancing datasets and augmentation.

A dataset may not adequately represent some groups or situations, which leads to the model being biased toward the underrepresented groups. By artificially creating additional examples for the underrepresented groups, synthetic data balances the dataset. In addition, the method of "data augmentation" — creating new examples by slightly modifying existing data — increases the amount of training data and improves the model's robustness.

5. Reliability, limitations, and prospects.

Although synthetic data offers many advantages, it has limitations. Synthetic data cannot fully replace real data — a model must ultimately be tested on real data. The quality of the synthetic data, the preservation of privacy, and the prevention of bias must be carefully verified. For Uzbekistan, the use of synthetic data is relevant for the development of AI, especially in fields where the collection of real data is limited. To this end, it is necessary to train specialists in this field.

Conclusion

Generating synthetic data is becoming an important tool for the development of artificial intelligence. By solving the problems of collecting real data — cost, privacy, shortage, and imbalance — synthetic data accelerates and facilitates the training of AI models.

Our research shows that synthetic data offers especially valuable solutions in fields where privacy is important (medicine, finance) and where there is a shortage of real data. However, synthetic data cannot fully replace real data. The quality of the data, the preservation of privacy, and the prevention of bias remain relevant problems.



In conclusion, synthetic data is one of the important technologies for the development of AI. For Uzbekistan, the use of synthetic data, especially in fields where the collection of real data is limited, is relevant for the development of AI. By training specialists in this field, we can effectively make use of these technologies.

References

1. Goodfellow, I., et al. (2014). "Generative Adversarial Nets." Advances in Neural Information Processing Systems.
2. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.
3. Nikolenko, S. I. (2021). Synthetic Data for Deep Learning. Springer.
4. Jordon, J., et al. (2019). "PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees." ICLR.
5. Bowles, C., et al. (2018). "GAN Augmentation: Augmenting Training Data using Generative Adversarial Networks." arXiv preprint.
6. Choi, E., et al. (2017). "Generating Multi-label Discrete Patient Records using Generative Adversarial Networks." Machine Learning for Healthcare.
7. Shorten, C., & Khoshgoftaar, T. M. (2019). "A survey on Image Data Augmentation for Deep Learning." Journal of Big Data.
8. Vaswani, A., et al. (2017). "Attention is All You Need." Advances in Neural Information Processing Systems.
9. Dwork, C., & Roth, A. (2014). "The Algorithmic Foundations of Differential Privacy." Foundations and Trends in Theoretical Computer Science.
10. Tremblay, J., et al. (2018). "Training Deep Networks with Synthetic Data: Bridging the Reality Gap by Domain Randomization." CVPR Workshops.
11. Xu, L., et al. (2019). "Modeling Tabular Data using Conditional GAN." Advances in Neural Information Processing Systems.
12. Russell, S., & Norvig, P. (2020). Artificial Intelligence: A Modern Approach (4th ed.). Pearson.
13. Kingma, D. P., & Welling, M. (2014). "Auto-Encoding Variational Bayes." ICLR.
14. Decree of the President of the Republic of Uzbekistan "On the approval of the strategy 'Digital Uzbekistan – 2030'" (October 5, 2020).
15. Assefa, S. A., et al. (2020). "Generating Synthetic Data in Finance: Opportunities, Challenges and Pitfalls." ACM International Conference on AI in Finance.



16. UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence.
17. Law of the Republic of Uzbekistan "On Personal Data" (2019).
18. Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.