14th International Defense and Homeland Security Simulation Workshop,006
21st International Multidisciplinary Modeling & Simulation Multiconference

2724-0363 © 2024 The Authors.
doi: 10.46354/i3m.2024.dhss.006

Leveraging Synthetic Data Generation for Military Decision Support Enhancement

Agostino Bruzzone[1,*], Umberto Battista[2], Carolina Badano[2] & Federico Taddei Santoni[2]

[1]Simulation Team & Genoa University
[2]Stam S.r.l., Via Pareto 8 AR, 16129, Genoa, Italy

[*]Corresponding author. Email address: agostino.bruzone@simulationteam.com

Abstract

Synthetic data generation has become an important approach to tackle challenges related to data scarcity, privacy concerns, and resource optimization in artificial intelligence applications. This paper examines recent advances in synthetic data generation methods, with a focus on generative learning, transfer learning, and modelling techniques. Generative learning uses machine learning models to replicate statistical patterns found in real-world data. Meanwhile, transfer learning allows for knowledge transfer across related tasks, reducing the impact of data scarcity. Modeling techniques, such as statistical and machine learning-based approaches, creates synthetic data that closely mirrors real data distributions. This paper examines various methodologies and case studies and their significance in different application domains, with an emphasis on the military. Additionally, benchmarking analyses demonstrate the effectiveness of Generative Adversarial Networks and Variational Autoencoders in synthetic data generation tasks. Transfer learning strategies are evaluated considering their advantages/disadvantages and the field of application. Modelling techniques are evaluated to generate synthetic scenarios. The paper concludes by discussing the importance of synthetic data generation to enhance decision support in military domain.

Keywords: Synthetic data generation, Generative learning, Transfer learning, Modeling techniques, Artificial intelligence, Military domain

1. Introduction

One of the most important topics in today's world is artificial intelligence. Artificial intelligence (AI) refers to the development of computer systems capable of performing tasks that typically require human intelligence. These tasks include understanding natural language, recognizing patterns, learning from experience, and making decisions. With AI, machines analyze vast amounts of data, identify trends, and make predictions, leading to improved decision-making, increased productivity, and enhanced user experiences. Developing AI is expensive and requires technical expertise, leading to a shortage of skilled professionals. AI systems could also perpetuate biases present in training data, limiting their fairness. One of the key requirements for AI is real-world data sets. Despite the large and growing number of datasets related to technological advances, one of the main challenges is the low quality and scarcity of data, especially in the military field. To address this challenge, two important issues need to be addressed: frugal AI and synthetic data.

Frugal AI is a technique that aims to achieve robustness in AI models while using less data and computational resources. It involves training AI systems with limited

© 2024 The Authors. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-NC-ND) license (https://creativecommons.org/licenses/by-nc-nd/4.0/).

resources, focusing on input frugality and learning frugality. The goal is to achieve prediction quality while using less data and optimizing the learning process. In certain domains, such as the military, complete databases may not be readily available for security reasons. In such cases, frugal AI becomes essential. To overcome the lack of data, researchers and engineers are exploring different methods. One approach is the so-called "transfer learning", which involves using an existing AI system that has already learned from a sufficient data set. Another technique is the data generation, where a virtual environment is used to generate data that closely resembles the conditions of a real environment. Data augmentation is another method that involves generating new data by applying transformations or modifications to existing data. The combination of data generation and augmentation aims to a more comprehensive and diverse data set.

Data has significant value, but quality is paramount. The need for high quality data and privacy has become increasingly important as both businesses and researchers rely more heavily on data. Synthetic data, which consists of artificially generated information, is emerging as a powerful solution to these challenges. Synthetic data is often of higher quality than real data. In addition, privacy safeguards should be enforced to prevent the disclosure of critical information. In the military domain, data collection is challenging due to the dynamic and high-stakes nature of operations. Yet, privacy concerns are prevalent in the military due to the sensitive nature of the data involved. Synthetic data generation is proving to be a valuable and innovative solution. While synthetic data is a compelling concept, its generation demands precision. It must be plausible and adhere to the underlying distribution of the original data. Consequently, the algorithms responsible for generating synthetic data must exhibit robustness and effectively capture the patterns inherent in real data.

2. State of the art

Synthetic data generation is mostly based on the following methodologies that are able to operate individually or in combination. These methodologies are generative learning, transfer learning, and modeling and simulation.

2.1. Generative Learning

Generative learning, in the context of synthetic data generation, is a cutting-edge technique that uses machine learning models to create artificial data that mimics the statistical patterns and characteristics of real-world data. This approach is particularly valuable in scenarios where obtaining sufficient and diverse real-world data may be difficult, expensive or privacy sensitive. The generative models learn the underlying data distribution by analyzing the patterns, relationships and structures present in authentic data. Once properly trained, they produce synthetic data samples that closely resemble the original data in terms of statistical properties and features. However, it's important to note that the quality and effectiveness of synthetic data depends on the accuracy and representativeness of the generative model. Careful evaluation and validation are essential to ensure that the synthetic data meet the requirements of the intended applications.

2.1.1. *Generative Adversarial Networks*

Generative Adversarial Networks (GANs) were introduced in 2014 and typically consist of a generator and a discriminator that learn simultaneously. The generator aims to capture the potential distribution of real samples and generate new data samples, while the discriminator acts as a binary classifier, accurately distinguishing between real and generated samples. GANs learn by implicitly computing the similarity between the distribution of a candidate model and the distribution corresponding to real data. The main issue related to GANs is that the generator could be trained even after and when the discriminator is optimal, aiming to act on the accuracy of the discriminator. If the generator distribution perfectly matches the distribution of real data, the discriminator is maximally confused, predicting 0.5 for all inputs. However, the discriminator practically cannot be optimally trained.

2.1.2. *Autoregressive models*

Autoregressive statistical models are widely used for predicting future values based on past data. They assume that past values significantly influence current values and are applied in various fields to analyze time-varying phenomena. In autoregressive models, white noise is used to represent random fluctuations in the data. However, these models have limitations when the underlying dynamics change over time, potentially leading to inaccurate predictions. To overcome this, advanced modeling techniques incorporate additional components like moving averages, seasonal patterns, and trend analysis to capture evolving dynamics and improve accuracy.

2.1.3. *Generative Moment Matching Networks*

Generative Moment Matching Networks (GMMNs) utilize a simple prior distribution for easy sampling. This prior is then deterministically propagated through the hidden layers of the neural network, and the output represents a sample from the model. This enables GMMNs to quickly generate independent random samples. The fundamental concept of GMMNs revolves around employing a neural network to acquire a deterministic mapping from samples of a straightforward and readily accessible distribution to samples derived from the data distribution. The generative network consists of a stochastic hidden layer with independent prior uniform distributions for each hidden unit. GMM are useful in generating

synthetic data that closely matches the statistical properties of real-world military data. However, they face challenges in capturing complex data dependencies due to the deterministic propagation of this prior through hidden layers.

2.1.4. *Variational Autoencoders*

A Variational Autoencoder (VAE) is a neural network architecture commonly used in unsupervised machine learning. Its primary objective is twofold: first, to acquire a concise and continuous representation of data, and second, to generate new data samples that closely resemble the input. An Autoencoder represents a straightforward neural network architecture comprising two fundamental components: the encoder and the decoder. The encoder's role is to condense the original input into a compact representation within a significantly smaller vector space. Conversely, the decoder aims to reconstruct the compressed data back into its original form, albeit with some degree of loss. In contrast, a VAE extends this concept by not only compressing the data but also learning its underlying distribution. By leveraging this distribution, the VAE could be able to decode and generate entirely new data points.

2.1.5. *Normalizing Flows*

Normalizing flows (NFs) are a sequence of simple functions that can be inverted or have an analytical inverse. These flows transform complex data points, like Modified National Institute of Standards and Technology (MNIST) Images, into simple Gaussian Distributions, and vice versa. Unlike GANs, where the generator is trained to produce images from random vectors, flow-based models transform data points into simple distributions during training. Flow-based models do not require noise on the output, allowing for powerful local variance models. The training process of flow-based models is more stable compared to GANs, which necessitate careful tuning of hyperparameters for both generators and discriminators. Normalizing flows also converge more easily. However, the quality of samples generated by flow-based models is not as good as those produced by GANs and VAEs.

2.2. Transfer Learning

Transfer learning is a technique often used when data sets are small and computational resources are limited. The concept revolves around reusing the parameters of pre-trained models, typically embodied by deep neural networks, in the creation of new models designed for related, albeit different, tasks. Transfer learning in deep learning applications offers two distinct approaches, each tailored to optimize model performance. The first approach involves "relearning the output layer". This involves replacing the final layer of a pre-trained model with a new output layer designed to match the expected output of the new task. The second approach is "fine tuning the whole model". This is similar to the first approach, but with one key difference: it allows the weights of the entire Deep Neural Network to be updated. Fine-tuning the entire model could be beneficial when the source and target tasks have a moderate to high degree of similarity, although it typically requires a larger amount of training data to achieve optimal results.

The main advantage of transfer learning is its potential to improve the generalization capabilities of a model, allowing it to perform effectively on a wide range of tasks. By transferring knowledge from pre-trained models, they could achieve better performance even with limited training data. However, it's important to recognize that the effectiveness of transfer learning diminishes as the dissimilarity between the source and target tasks increases.

2.2.1. *Style transfer learning*

Style transfer learning is a machine learning and computer vision technique that transfers the style of one image to another. The algorithm transforms the image style while preserving the content image's structure. Thus, the resulting composite image represents a seamless blend of the original image content and the desired style. An essential aspect is to analyze an image with a specific style and create a mathematical or statistical model that accurately represents that style. By comprehending the underlying patterns and features of the style image, it is possible to formulated a model for transformation process. This model is a reference for adjusting the target image to align better with the established style, resulting in visually pleasing outcomes. However, traditional style transfer approaches have limited versatility. They excel at replicating a specific style or scene but struggle to adapt to diverse styles or accommodate multiple stylistic elements within a single image.

2.2.2. *Domain adaptation*

Domain adaptation is a form of transfer learning that leverages labeled data from a related area to enhance the performance of a target model that lacks annotated data. Unlike transfer learning, where tasks may differ between source and destination, domain adaptation maintains the same tasks while varying only the domains. The aim of domain adaptation is to reduce the differences between the source and target domains, allowing for successful transfer of the model trained on the source domain to the target domain. These differences are expected to be due to variations in data collection, such as the use of different sensors, perspectives, or illumination conditions. Recent research recommends two methods of domain adaptation: open set domain adaptation and partial domain adaptation. Open set domain adaptation introduces 'unknown' classes in both domains and assumes knowledge of common classes during training. Partial domain adaptation requires that the

source labels include the target labels. Partial domain adaptation is appropriate in situations where the source label set adequately encompasses the target labels. In contrast, if the source label set shares classes or is a subset of the target label set, open set domain adaptation is the preferred method. The main challenge in the general scenario is the lack of prior knowledge about the label set of the target domain, making it impossible to select the appropriate domain adaptation method. Unsupervised Domain Adaptation provides a promising solution to this problem by using labelled source data and unlabeled target data to train a classification model.

2.3. Modeling and Simulation

Modelling and simulation techniques play a crucial role in generating synthetic data that closely mirrors real-world data across various fields. At its core, this process involves creating artificial data that captures the statistical and structural characteristics of real data, particularly when genuine data is scarce, sensitive, or costly to obtain. In complex systems, the use of Simulation to recreate te Scenario Dynamic by including AI elements, often encapsulated in Intelligent Agents to reproduce the dynamics of entities and units as well as Commander decision Processes, Actions and Reactions based on Situation Awareness (Bruzzone et al.,2015). These Simulators for the application sectors of this research are usually discrete event stochastic agent driven and result in constructive simulation supporting multiple purposes (Bruzzone & Massei, 2017). In facts, the synthetic data construction usually is based on defining a Data Generating Process (DGP), which outlines essential elements and relationships within the data, including distributions, correlations, and other attributes. Selecting an appropriate modelling approach, be it statistical or machine learning-based, depends on the complexity of the real data and the intricacies of the DGP. In statistical modeling, techniques such as parametric and non-parametric models are utilized to simulate data that closely resembles the original dataset in terms of distributional properties and relationships between variables. In contrast, machine learning-based approaches learn the underlying data distribution from the original dataset and generate new samples that mimic it, capturing complex patterns and dependencies in the data. Parameter estimation is pivotal in model-based approaches, as accurate estimates directly influence how closely synthetic data resembles real data. Once the DGP and model parameters are determined, the process moves to simulate data generation, leveraging the chosen modelling approach to generate synthetic data points that adhere to specified patterns and relationships. Validation and adjustment steps are crucial to ensure synthetic data quality and accuracy, involving comparison with real data to identify and address any disparities. The iterative nature of synthetic data generation enhances its reliability over time, offering systematic and adaptable solutions to various data challenges, from privacy concerns to research and development needs.

3. Comparative Analysis

A benchmarking analysis has been used to evaluate the GL algorithm's effectiveness in generating synthetic data when faced with data insufficiency. It compares various generative learning approaches using a comprehensive scale ranging from 1 to 5. Key characteristics considered during the evaluation of algorithm choice include: output quality, realism of the synthetic data produced; ease of use, user-friendliness of implementing the algorithm; stability, robustness across different datasets; adaptability, ability to self-adjust to different datasets; completeness, generating sufficient synthetic data; applicability, suitability for various real-world applications; feasibility, how practical is to use the algorithm. By systematically evaluating these characteristics, the benchmarking analysis provides valuable insights into the performance and suitability of the GL algorithm for generating synthetic data in scenarios of data insufficiency. To evaluate transfer learning methods, their advantages and disadvantages are considered. Style transfer is an image processing technique that blends the artistic style of one image with the content of another, resulting in visually appealing and artistic images. While it allows for customization and ease of use, it may lead to a loss of original content and could be computationally intensive, potentially introducing imperfections. Despite these limitations, it remains a powerful tool for achieving desired artistic effects. Domain transfer, on the other hand, is a machine learning technique that transfers knowledge from one domain to another, enabling models trained on one domain to perform well in a related domain. However, challenges such as domain gaps and differences in data distribution or labelling schemes could hinder its effectiveness. The quality and relevance of the source domain data heavily influence the success of domain transfer techniques, but leveraging pre-existing knowledge could be useful to mitigate and reduce training time and required resources in the target domain. The two main modelling approaches are compared in terms of parameter estimation and DGP. Statistical modelling involves explicit estimation of parameters based on assumed distributions. Instead, machine learning modelling learns parameters implicitly during training. Another important aspect of the comparison is the DGP. Statistical modelling explicitly defines the DGP by specifying relationships between variables. Machine learning modelling learns the DGP implicitly from the data during training.

4. Results and Discussion

Based on the benchmarking analysis presented in Table 1, the most efficient results are observed with Generative Adversarial Networks (GANs) and

Variational Autoencoders (VAEs). When the primary objective is to achieve high-quality data, GANs emerge as the preferred choice due to their ability to generate realistic samples with visually appealing characteristics. Conversely, if stability is prioritized, VAEs offer a more suitable solution, as they tend to provide consistent performance and robustness in generating synthetic data. Therefore, depending on the specific requirements of the task, GANs are recommended for optimizing data quality, while VAEs are preferred for ensuring stability.

Table 1. Benchmarking analysis of GL algorithms

	Output quality	Ease of use	Stability	Adaptability	Completeness	Applicability	Feasibility	TOT
GAN	5	4	3	4	4	5	4	29
AM	4	4	3	3	4	4	3	25
GMMN	3	4	4	3	3	3	3	23
VAE	4	4	4	5	4	5	4	30
NF	5	3	3	5	4	4	3	27

The most promising transfer learning method is style transfer because it does not just generate images, but has the desired style in the synthetic data. Given the context of lack of data, and thus the important aspect of being able to implement desired features in synthetic images, it is best to integrate these methods into generative learning frameworks and specific contexts.

Statistical modelling offers interpretability but it could struggle with complex patterns. Machine learning modelling excels at capturing complex patterns but may lack interpretability. Therefore, statistical modelling is preferred for well-understood data distributions and machine learning modelling is suitable for complex, high-dimensional data.

5. Conclusions

The power of generating synthetic images has opened new frontiers, revolutionizing various domains across industries and sectors. The ability to create realistic and immersive simulations has transformed the way professionals are prepared for their tasks. In the military domain, where the stakes are high and the challenges are constantly evolving, preparing commanders for novel tasks presents a formidable challenge given their well-honed skills and extensive experience. Conventional training approaches often struggle to keep pace with the dynamic nature of geopolitical landscapes and the intricacies of modern warfare. However, Simulation-Based Military Training offers a solution by providing immersive environments where commanders refine their abilities in response to evolving challenges. By leveraging synthetic data, training protocols incorporate a diverse range of scenarios, enriching the available datasets and enhancing the realism of the training experience. Recent advancements in style transfer techniques further amplify this capability, allowing for the creation of highly realistic and dynamic training scenarios that closely mimic real-world situations. This not only expands the repertoire of training scenarios available to commanders but also augments decision support capabilities, enabling more effective and efficient responses to complex and unpredictable situations. Thus, the integration of synthetic images into military training methodologies represents a significant advancement, empowering commanders to better prepare for the challenges of tomorrow's battlefield.

Acknowledgements

The authors received support from the FaRADAI project (ref. 101103386) funded by the European Commission under the European Defence Fund.

References

McCarthy, J. (2007). What is artificial intelligence.

Bruzzone, A., Remondino, M., Battista, U., Tardito, G., & Santoni, F. T. (2021). Modelling & Data Fusion to support Acquisition in Defence.

Bruzzone, A.G., Massei, M., Longo, F., Nicoletti, L., Di Matteo, R., Maglione, G., Agresta, M. (2015) Intelligent agents & interoperable simulation for strategic decision making on multicoalition joint operations, 5th International Defense and Homeland Security Simulation Workshop, DHSS 2015, DOI

Larsson, S. (2022). Frugal AI: Value at Scale Without Breaking the Bank. Retrieved from https://blog.dataiku.com/frugal-ai-value-at-scale-without-breaking-the-bank#:~:text=Simply%20put%2C%20Frugal%20AI%20is,(and%20cost%20efficiency)%20objectives.

Bruzzone, A. G., & Massei, M. (2017). Simulation-based military training. Guide to Simulation-Based Disciplines: Advancing Our Computational Future, 315-361.

Ioualalen, A. Limousin, L. (2021). How can frugal AI overcome the lack of data? Retrieved from https://www.linkedin.com/pulse/how-can-frugal-ai-overcome-lack-data-arnault-ioualalen

Lawrence, A. R., Kaiser, M., Sampaio, R., & Sipos, M. (2021). Data generating process to evaluate causal discovery techniques for time series data. arXiv preprint arXiv:2104.08043.

Figueira, A., Vaz, B. (2022). Survey on Synthetic Data Generation, Evaluation Methods and GANs. Mathematics, 10(15), 2733. https://doi.org/10.3390/math10152733

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2020). Generative adversarial networks. Communications of the ACM, 63(11), 139-144.

Wang, K., Gou, C., Duan, Y., Lin, Y., Zheng, X., & Wang, F. Y. (2017). Generative adversarial networks:

introduction and outlook. IEEE/CAA Journal of Automatica Sinica, 4(4), 588-598.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). Generative adversarial networks: An overview. IEEE signal processing magazine, 35(1), 53-65.

Fernando, J. Catalano, J. Munichiello, K. (2022). What Are Autoregressive Models? How They Work and Example. Retrieved from https://www.investopedia.com/terms/a/autoregressive.asp#:~:text=A%20statistical%20model%20is%20autoregressive,based%20on%20its%20past%20performance.

Li, Y., Swersky, K., & Zemel, R. (2015, June). Generative moment matching networks. In International conference on machine learning (pp. 1718-1727). PMLR.

Zu, D. (2020). Generate Images Using VariationalAutoencoder (VAE). Retrieved from https://medium.com/@judyyes10/generate-images-using-variational-autoencoder-vae-4d429d9bdb5

Omray, A. (2021). Introduction to Normalizing Flows. Retrieved from https://towardsdatascience.com/introduction-to-normalizing-flows-d002af262a4b

Svenmarck, P., Luotsinen, L., Nilsson, M., & Schubert, J. (2018, May). Possibilities and challenges for artificial intelligence in military applications. In Proceedings of the NATO Big Data and Artificial Intelligence for Military Decision Making Specialists' Meeting (pp. 1-16).

Liu, L., Xi, Z., Ji, R., & Ma, W. (2019). Advanced deep learning techniques for image style transfer: A survey. Signal Processing: Image Communication, 78, 465-470.

Farahani, A., Voghoei, S., Rasheed, K., & Arabnia, H. R. (2021). A brief review of domain adaptation. Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020, 877-894.

You, K., Long, M., Cao, Z., Wang, J., & Jordan, M. I. (2019). Universal domain adaptation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2720-2729).