ISSN: 2634 - 8853

Journal of Engineering and Applied Sciences Technology

SCIENTIFIC Research and Community

Review Article

Open Access

Synthetic Data Generation for Realtime Data Pipelines

Girish Ganachari

USA

ABSTRACT

This study discusses synthetic data synthesis for real-time data pipeline enhancements. Many companies can scale, cost-effectively, and privately train and test machine learning models using synthetic data. Key applications include advanced simulations, model effectiveness, and privacy. Despite data realism, computational complexity, and domain-specific requirements, generative models and integration approaches are promising. Legal and ethical issues must be resolved for acceptance. This study proves synthetic data's effectiveness, dependability, and regulatory compliance, revolutionising data-driven systems.

***Corresponding author**

Girish Ganachari, USA.

Received: February 01, 2022; **Accepted:**February 10, 2022; **Published:** February 26, 2022

Keywords: Synthetic Data, Real-Time Data Pipelines, Machine Learning, Privacy Preservation, Generative Models, Data Simulation

Introduction

Big data has altered many organisations, needing complicated data-driven systems to assess enormous data sets. Data pipelines that are able to function in real time are required in order to adequately meet the requirements of this demand. Companies can swiftly examine data via continuous data flow pipes. This facilitates quick, informed decisions. Avoid delays in financial trading, healthcare monitoring, and autonomous driving using real-time data pipelines. Creating realistic datasets for testing and training real-time data pipelines is tough despite its importance. Data privacy issues. Medical information and financial data are difficult to use for training due to privacy regulations. Specialised fields with costly or complex data collecting may have little data. Training unsuitable models without real-world data may cause biases and system performance issues. These issues may be resolved using synthetic data. Synthetic data resembles actual data. Advanced algorithms and models may simulate several data situations to create adaptable, scalable synthetic data. It anonymises sensitive data and provides huge, diverse datasets for testing and training. Synthetic data builds and optimises real-time pipelines. Synthetic data may help companies meet privacy rules, generate data-driven apps, and bypass data constraints. In data-driven systems, synthetic data production for real-time data pipelines has many uses, benefits, problems, and future prospects.

Applications of Synthetic Data in Real-Time Data Pipelines Enhancing Machine Learning Models

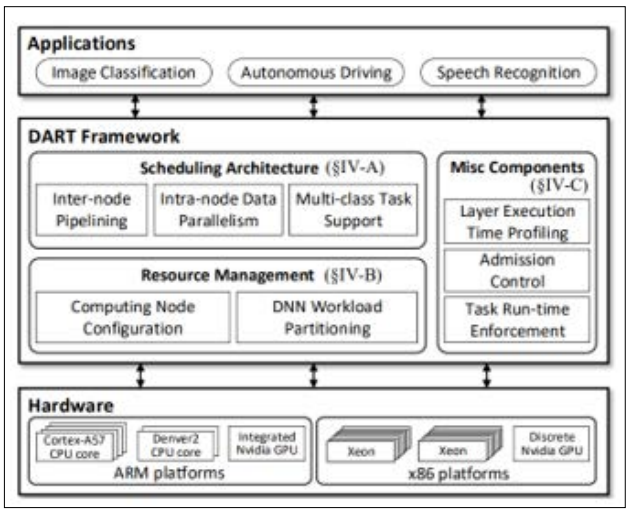

Figure: DART framework overview

Training datasets for machine learning algorithms may be improved by the creation of false data. Synthetic data may generate enormous volumes of training data cheaply. Planche et al. enhanced 2.5D recognition algorithms using CAD model synthetic data. This strategy provides a wide range of training samples for model performance while reducing data collection and annotation costs [1].

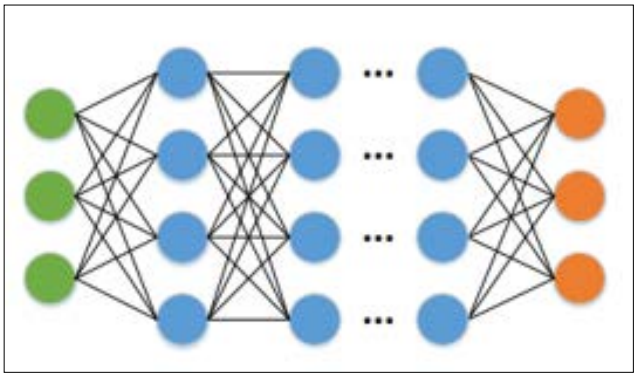

Figure 2: An example DNN

model with one input layer, several hidden layers, and one output layer

Compositing synthetic data by Tripathi et al. improved computer vision model dependability and effectiveness. Synthetic data provides diverse, well-annotated training situations to increase model generalisation. This improves real-world projections [2]. Researchers may use synthetic data tailored to their purposes to conduct experiments in difficult-to-replicate situations. Improves machine learning model efficiency.

Privacy Preservation

In sensitive sectors like banking and healthcare, real-time data privacy is essential. Synthetic data can tackle these problems by balancing privacy and utility. Synthetic data may preserve statistical traits without identifying information, according to El Emam, Mosquera, and Hoptroff [3]. This preserves user privacy when training and assessing bogus datasets. Synthetic data may also fulfil GDPR privacy requirements. Creating data without identifying individuals lets companies send and analyse information without infringing privacy rules. Cross-institutional studies and collaborative research benefit from this when data interchange is important but privacy considerations prohibit it.

Real-Time Simulation and Testing

Simulating and testing real-time data pipelines needs synthetic data. Studying monocular depth estimation. In controlled experiments, Atapour-Abarghouei and Breckon tested depth estimate algorithms using fake data. Researchers may test algorithms in synthetic settings to simulate dangerous or challenging conditions that are hard to replicate [4]. Before deployment, this function calibrates systems to reduce errors and increase reliability. To evaluate algorithm reliability and flexibility, autonomous automobiles may travel in simulated fog or clear sky. Medical items may be tested under different physiological settings using synthetic patient data.

Addressing Domain Shift

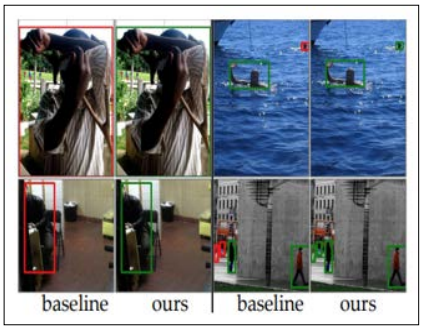

Figure 3: Comparison of object detection results using SSD

Domain shift issue solving is a key application of synthetic data. Domain shift separates a model's training and real-world data. It generally reduces model performance. Customising synthetic data may narrow this gap and make models more lifelike. Sankaranarayanan et al. addressed semantic segmentation domain shift using synthetic data. Researchers created phoney datasets that resembled the desired domain to increase model performance. It's beneficial when acquiring tagged data from the target location is hard or impossible. Synthetic data may simulate illumination, meteorological, and other environmental factors to assess model reliability and performance.

Benefits of Synthetic Data Generation

Scalability, Flexibility, and Cost-Effectiveness

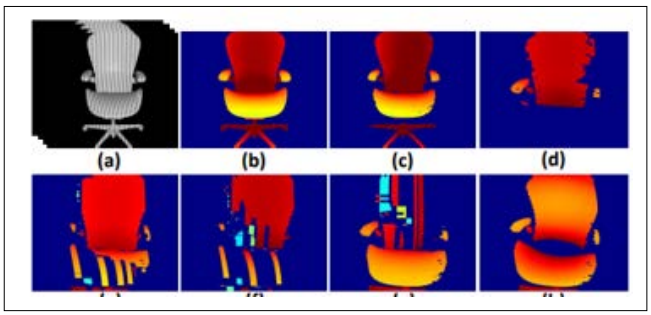

Figure 4: Examples of data generated, simulating a multi-shot depth sensor

Synthetic data outperforms real-world data in scalability and variety. Nikolenko advocated generating huge datasets for individual requirements to ensure systems can handle different data inputs and scenarios. [5]. Real-time data pipeline generation and assessment across industries need flexibility. Synthetic data synthesis is cheaper than real-world data collection and analysis. According to Tsirikoglou et al. [6], procedural modelling and physically based rendering save money in automotive applications. Synthetic data production cuts data preparation expenses by eliminating fieldwork and annotating. Synthetic data's scalability, adaptability, and affordability make it appealing for automotive and healthcare applications.

Improved Model Performance and Complex Simulations

Synthetic data may assist machine learning models by providing well-labeled training datasets. Wood et al. proved that high-quality synthetic data trains robust models like real-world data. They acquired cutting-edge facial analysis results from produced data [7]. Real-time data pipelines need accuracy and reliability, therefore model performance must improve. Synthetic data may also imitate complicated situations that are hard to replicate with real data. Wang et al. showed that synthetic data can train blind super-resolution models for real-world conditions, making training challenging [8]. Several real-time applications need unexpected event prediction. Synthetic instances may assist models handle unusual classes and conditions, enhancing generalisation, Beery et al. found [25]. Synthetic data enhances real-time data pipelines and application performance.

Challenges in Synthetic Data Generation
Data Realism, Validity, and Computational Complexity

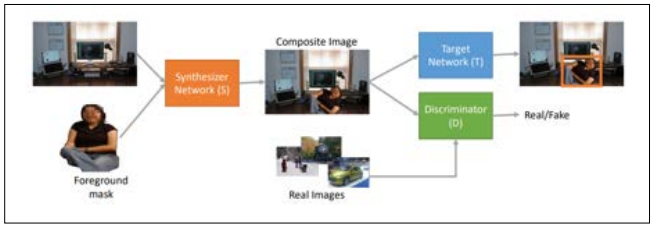

Figure 5: Our pipeline consists of three components: a synthesizer S, the target network T, and a natural image discriminator

Similarity to genuine data is prioritised in synthetic data development. High-order degradation algorithms may replicate realistic degradations, although ultimate realism is challenging because synthetic data may vary from actual data [8]. Synthetic data lacks realism, which may impair real-time data pipeline training and assessment. Quality synthetic data demands plenty of computing power. Domain adaptation may limit large-scale systems due to computationally demanding end-to-end synthetic data production [9]. To overcome computational challenges and make synthetic data creation viable and relevant, efficient methods and scalable infrastructure are required.

Domain-Specific Requirements and Integration with Real Data
The demands of each area differ, hence synthetic data must be customised. Developers need domain-specific synthetic datasets to verify appropriateness. This method requires specific instruments and subject area knowledge, making it challenging and time-consuming [10]. Healthcare synthetic data must accurately reflect medical issues, whereas autonomous automobile data must simulate various driving conditions. Integrating synthetic and real data is challenging for bias adjustment and consistency. The semantic dense fog scene interpretation model adaptation employs synthetic and real data, showing the difficulty of data integration [23]. In real-time data pipelines, synthetic data must match actual data without interrupting it for reliable results. These issues must be addressed to use synthetic data across applications without unintended repercussions.

Future Directions
In order to produce synthetic data, VAEs and GANs are absolutely necessary [3]. High-quality synthetic data showed promise from these models. Further study will enhance their real-time data stream. Real-time systems require artificial data generation to generate data quickly and increase data pipeline flexibility [5,7]. Transparency, justice, and accountability in synthetic data usage establish trust. Ethics and regulation are needed [3,12]. Enhancing synthetic data with additional scenarios enhances model generalisation and durability [16]. Wrap 20 in square brackets. Real-time synthetic data adjustment techniques are needed to accurately represent real-world occurrences in dynamic data adaption in real-time systems [21]. Setting clear norms and criteria for synthetic data generation and use would help address challenges and assure its ethical and successful use across several fields.

Case Studies and Practical Implementations

Figure 6: Using a discriminator improves the realism of generated images.

Many sectors employ synthetic data, demonstrating its adaptability. To create synthetic data, Tsirikoglou et al. researched procedural modelling and physically based rendering in cars. The study projected autonomous vehicle training and testing driving and environmental conditions [6]. Synthetic patient data lets healthcare companies securely share and analyse medical data. The technique improves data privacy and availability [3]. Agriculture benefits from semantic segmentation data synthesis. These technologies improved precision farming by creating new agricultural settings. This helps crop monitoring model training and disease detection [24]. Urban planning misrepresents population and transit demand to predict urban dynamics and justify infrastructure [22]. Natural language processing adapts question-answering systems to new areas using synthetic data. Large training datasets help models [9]. Environmental monitoring, especially severe weather modelling, uses synthetic data. Replicating densely crowded and foggy situations improves monitoring and navigation [15,23]. These case studies demonstrate synthetic data's benefits across industries.

Conclusion
Synthetic data is private, scalable, and cost-effective for real-time pipelines. Using this technique, enormous databases with diverse data may be created and changed. This trains and analyses industry-wide machine learning models. GANs, VAEs, and enhanced integration methods offer a bright future for synthetic data production despite data realism, computing complexity, and domain-specific needs. By creating these technologies, synthetic data can match real-world data and decrease disparities. Synthetic data confidence, accountability, equality, and transparency need ethical and legal resolution. Synthetic data production and usage must be managed for ethical and legal reasons. In current data-driven systems, synthetic data may increase real-time data pipeline efficiency, reliability, and flexibility. Companies may utilise data-driven insights legally and privately. Innovation and advancement in many fields will ensue.

References

1. Planche B, Wu Z, Ma K, Sun S, Kluckner S, et al. (2017) Depthsynth: Real-time realistic synthetic data generation from cad models for 2.5 d recognition. In 2017 International conference on 3d vision (3DV) IEEE 1-10.
2. Tripathi S, Chandra S, Agrawal A, Tyagi A, Rehg JM, et al. (2019) Learning to generate synthetic data via compositing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 461-470.
3. El Emam K, Mosquera L, Hoptroff R (2020) Practical synthetic data generation: balancing privacy and the broad availability of data. O'Reilly Media.
4. Atapour-Abarghouei A, Breckon TP (2018) Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In Proceedings of the IEEE conference on computer vision and pattern recognition 2800-2810.
5. Nikolenko SI (2021) Synthetic data for deep learning Springer Nature 174.
6. Tsirikoglou A, Kronander J, Wrenninge M, Unger J (2017) Procedural modeling and physically based rendering for synthetic data generation in automotive applications. arXiv preprint arXiv:1710.06270.
7. Wood E, Baltrušaitis T, Hewitt C, Dziadzio S, Cashman TJ, et al. (2021) Fake it till you make it: face analysis in the wild using synthetic data alone. In Proceedings of the IEEE/CVF international conference on computer vision 3681-3691.
8. Wang X, Xie L, Dong C, Shan Y (2021) Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In Proceedings of the IEEE/CVF international conference on computer vision 1905-1914.
9. Shakeri S, dos Santos C, Zhu H, Ng P, Nan F, et al. (2020) November. End-to-end synthetic data generation for domain adaptation of question answering systems. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) 5445-5460.
10. Bellovin SM, Dutta PK, Reitinger N (2019) Privacy and synthetic datasets. Stan. Tech. L. Rev 22: 1.
11. Ward D, Moghadam P, Hudson N (2018) Deep leaf segmentation using synthetic data. arXiv preprint arXiv:1807.10931.
12. Sankaranarayanan S, Balaji Y, Jain A, Lim SN, Chellappa R (2018) Learning from synthetic data: Addressing domain shift for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition 3752-3761.
13. Saleh FS, Aliakbarian MS, Salzmann M, Petersson L, Alvarez JM (2018) Effective use of synthetic data for urban scene semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV) 84-100.
14. Gupta A, Vedaldi A, Zisserman A (2016) Synthetic data for text localisation in natural images. In Proceedings of the IEEE conference on computer vision and pattern recognition 2315-2324.
15. Sakaridis C, Dai D, Van Gool L (2018) Semantic foggy scene understanding with synthetic data. International Journal of Computer Vision 126: 973-992.
16. Richardson E, Sela M, Kimmel R (2016) 3D face reconstruction by learning from synthetic data. In 2016 fourth international conference on 3D vision (3DV) IEEE 460-469.
17. Zamir SW, Arora A, Khan S, Hayat M, Khan FS, et al. (2020) Cycleisp: Real image restoration via improved data synthesis. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2696-2705.
18. Danielczuk M, Matl M, Gupta S, Li A, Lee A, et al. (2019) Segmenting unknown 3d objects from real depth images using mask r-cnn trained on synthetic data. In 2019 International Conference on Robotics and Automation (ICRA) IEEE 7283-7290.
19. Hinterstoisser S, Pauly O, Heibel H, Martina M, Bokeloh M (2019) An annotation saved is an annotation earned: Using fully synthetic training for object detection. In Proceedings of the IEEE/CVF international conference on computer vision workshops pp 0-0.
20. Marion P, Florence PR, Manuelli L, Tedrake R (2018) Label fusion: A pipeline for generating ground truth labels for real rgbd data of cluttered scenes. In 2018 IEEE International Conference on Robotics and Automation (ICRA) IEEE 3235-3242.
21. Xiang Y, Kim H (2019) December. Pipelined data-parallel CPU/GPU scheduling for multi-DNN real-time inference. In 2019 IEEE Real-Time Systems Symposium (RTSS) IEEE 392-405.
22. Hörl S, Balac M (2021) Synthetic population and travel demand for Paris and Île-de-France based on open and publicly available data. Transportation Research Part C: Emerging Technologies 130: 103291.
23. Sakaridis C, Dai D, Hecker S, Van Gool L (2018) Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In Proceedings of the european conference on computer vision (ECCV) 687-704.
24. Barth R, IJsselmuiden J, Hemming J, Van Henten EJ (2018) Data synthesis methods for semantic segmentation in agriculture: A Capsicum annuum dataset. Computers and electronics in agriculture, 144: 284-296.
25. Beery S, Liu Y, Morris D, Piavis J, Kapoor A, et al. (2020) Synthetic examples improve generalization for rare classes. In Proceedings of the ieee/cvf winter conference on applications of computer vision 863-873.