WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Generative AI in context of Data Augmentation

Aashna Arora[1] and Dr Thiruvenkadam[2]

[1] Jain (Deemed to be University), Bangalore KA 560069, India
[2] Professor, Jain (Deemed to be University), Bangalore KA 560069, India

ABSTRACT :

This paper presents Aashna Arora's system, involves using data generated by a Large Language Model (LLM) and how in many realistic settings we need to achieve goals with limited datasets; in those case deep neural networks seem to fall short, overfitting on the training set and producing poor generalization on the test set. The system uses machine learning and neural representations to train a model that predicts how useful code- comment pairs are. Techniques have been developed over the years to help combat overfitting such as dropout (Hinton et al., 2012b), batch normalization (Ioffe & Szegedy, 2015), batch renormalization (Ioffe, 2017) or layer normalization (Ba et al., 2016).

This research represents a pioneering step in exploring the intersection of Generative AI and data augmentation. As we delve deeper into this innovative realm, there is immense potential for advancements in machine learning capabilities, offering opportunities for real-world applications across various domains. This paper aims to set the stage for future studies, encouraging continued exploration and practical implementation of Generative AI in the context of data augmentation.

INTRODUCTION :

In the fast pace landscape of artificial intelligence (AI), the synergy between Generative AI and data augmentation has emerged as a game-changing influence, promising to elevate the efficiency and effectiveness of various machine learning tasks. Generative AI, consists of various high-tech technologies like Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), has shown and portrayed potential in its capabilities in generating data which is synthetic in nature and that resembles the patterns and structures present in real-world datasets.

Data augmentation is as one of the most crucial strategy to address challenges relating with limited of varied and the availability of abundant datasets is often a constraint in developing resilient models and generalizable machine learning models. training data. In multiple arenas such as computer vision and natural language processing (NLP), and healthcare data augmentation involves creating variations of existing data by applying transformation as rotations, translations, or changes in color. While traditional data augmentation methods have proven effective, Generative AI introduces a paradigm shift by allowing the creation of entirely new, realistic data samples.

Furthermore, incorporating Generative AI into workflows has the potential to streamline processes, resulting in increased productivity across various roles, including solution architects and software developers. This technology also contributes to a more detailed and precise requirement gathering process, empowering businesses to refine their data analytics strategies with greater effectiveness. Integrating Generative AI into data quality management processes goes beyond addressing immediate concerns about data accuracy; it can also yield long-term organizational benefits, including increased efficiency and more effective decision-making. This research marks a pioneering advancement in the convergence of Generative AI and data quality, laying the foundation for future studies and practical applications in the real world.

Data augmentation method also creates synthetic data, although that is done artificially This is done to elevate the real ones which is obtained by procuring observations and experiments. It is already in use for many years within the image and visual recognition, its use is mainly based on the simple idea of creating a lot of variations (on the basis of multiple factors like scaling etc) of the images that are put in use whilst training the Deep Learning algorithms, and therefore helps in bettering the durability and the performance of the machine.

The fusion of Generative AI and data augmentation not only addresses immediate challenges related to data accuracy but promises enduring organizational benefits. These encompass heightened efficiency in model training, improved generalization capabilities, and, fundamentally, more effective decision-making.

REVIEW OF LITERATURE :

In recent years, the intersection of Generative Artificial Intelligence (Generative AI) and data augmentation has garnered substantial attention within the machine learning community. This survey aims to provide an overview of key studies and developments in this evolving field.

1. A brief on Generative AI and Data Augmentation:

- The new emergence of Generative AI, which includes technologies like Generative Adversarial Networks also known as GANs and Variational Autoencoders known as VAEs, has created new opportunities for production of synthetic or artificial data. This capability is particularly impactful in the realm of data augmentation, where the scarcity and quality of training data pose challenges.

2. Early Applications and Methods –

Early studies by Goodfellow et al. (2014) [1] introduced GANs, revolutionizing data generation. Researchers explored GANs for generating realistic data samples, expanding their application beyond image data to diverse domains.

3. Improving Data Diversity and Robustness:

- Subsequent research delved into enhancing data diversity and model robustness through generative techniques. Methods like cycleGAN (Zhu et al., 2017) [2] extended GANs to unpaired image-to-image translation, demonstrating versatility in augmenting datasets.

4. Addressing Domain-Specific Challenges:

- Studies by Antun et al. (2019) [3] focused on addressing domain-specific challenges through generative techniques, particularly in medical imaging. The research demonstrated the unlocking of possibilities of generative AI in overcoming data limitations for specialized applications.

5. Beyond Image Data:

- Recent advancements extended Generative AI to diverse data types. For example, Sequential GANs (SeqGAN) proposed by Yu et al. (2017) [4] tackled the generation of realistic sequential data, showcasing applicability in time series and natural language data augmentation.

6. Challenges and Limitations:

- Despite the promising advancements, challenges such as mode collapse and biases in generated data persist. Ongoing research explores methods to mitigate these issues and ensure the reliability of augmented datasets.

7. Performance Evaluation and Metrics:

- Evaluating the performance of generative models and augmented datasets is a critical aspect. Metrics like Inception Score (Salimans, 2016) [5] and Frechet Inception Distance (Heusel et al., 2017) [6] provide quantitative measures, but ongoing work aims to establish comprehensive evaluation frameworks.

8. Real-world Applications

- Practical applications of Generative AI in data augmentation span industries, including healthcare, finance, and natural language processing. Case studies, such as the work by Brownlee (2020) [7], demonstrate the impact on real-world problems.

Components Of GAN'S :

In recent years, the advent of Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) has ushered in a new era of generative modeling in machine learning. These techniques enable the creation of synthetic data that closely resembles real-world examples, facilitating a wide range of applications such as image generation, data augmentation, and anomaly detection. This paper provides an in-depth exploration of GANs and VAEs, elucidating their core concepts, architectures, and applications. By understanding the intricacies of these two techniques, we can harness their full potential to drive innovation and progress in the field of machine learning.

Generative Adversarial Networks (GANs): Generative Adversarial Networks (GANs) are a class of deep learning models that consist of two neural networks: a generator and a discriminator. The generator learns to generate realistic data samples from random noise, while the discriminator learns to distinguish between real and generated samples. Through adversarial training, the generator and discriminator engage in a game of cat-and-mouse, with the generator continually improving its ability to generate realistic data while the discriminator becomes better at discerning real from fake. The result is a generator that produces high-fidelity data samples that closely resemble real-world examples. GANs have demonstrated remarkable success in

generating images, videos, and even text, with applications ranging from image synthesis to style transfer.

Variational Autoencoders (VAEs): Variational Autoencoders (VAEs) are another class of generative models that learn to generate data by compressing and decompressing input samples. VAEs consist of two components: an encoder and a decoder. The encoder maps input samples to a latent space where they are represented as latent variables, while the decoder reconstructs the input samples from these latent variables. Unlike GANs, VAEs are trained using a probabilistic framework, where the model learns to maximize the likelihood of generating realistic data samples. By learning a probabilistic distribution over the latent space, VAEs can generate diverse and realistic data samples by sampling from this distribution. VAEs have found applications in image generation, data augmentation, and unsupervised learning, demonstrating their versatility and efficacy across various domains.
While both GANs and VAEs are capable of generating realistic data samples, they differ in their underlying principles and training objectives. GANs focus on generating data that is indistinguishable from real examples by training a generator to fool a discriminator. In contrast, VAEs learn to generate data by modeling the underlying probability distribution of the input data and sampling from this distribution. This fundamental difference results in distinct characteristics for each technique. GANs tend to produce sharper and more realistic samples, but they may suffer from mode collapse, where the generator learns to generate only a subset of the possible data distribution. On the other hand, VAEs produce more diverse samples but may sacrifice some level of realism compared to GANs. Additionally, VAEs provide a principled framework for learning latent representations of data, enabling tasks such as data interpolation and manipulation..

Research Objective :

The good things about Artificial Intelligence (AI) only work well when the AI has enough data to learn from. But getting this data can be tricky because it might be expensive, hard to find, or take too much time. Also, sometimes there's not enough data, and that can make the AI decisions less accurate. Lately, it's become even more challenging because Deep Learning, a type of AI, has become more powerful thanks to better computer chips (like GPUs). This has led to making fancier AI models that need even more data to avoid mistakes by learning too much from just a little bit of information.
Data augmentation provides a feasible solution for tackling the substantial challenges that have been discussed priorly by going ahead with training the model and computing it to face such scenarios . Consequently, there is a heightened demand for dependable data augmentation methods today. These methods would allow the utilization of high-dimensional models while reducing the likelihood of overfitting. Data augmentation has it's roots deep in providing application in different fields, some of the examples for the same are clinical trials, scientific experiments, industrial testing, and financial risk management. Given the current computational capabilities, data augmentation is now becoming an essential technique technique for training Machine Learning algorithms more effectively.

Finding one single, data augmentation method which is optimal and can be applicable in all scenarios is not practically possible. The selection of a data augmentation method for a specific situation is expected to depend on at least four key factors:

- Type of Data: Different types of data, such as tabular, images, time series, or chemical structures, may necessitate distinct data augmentation methods tailored to the unique structure and intricacies of each data type.
- Downstream Task: The nature of the downstream processing of augmented data influences the choice of data augmentation method. Factors like the neural network architecture and whether the task involves classification play a role in this decision.
- Performance Evaluation Metrics: To determine the optimal data augmentation method, it is essential to compare its performance against competing methods using formally defined evaluation metrics. Further research studies are needed to delve deeper into comparing various data augmentation methods.
- Computation Time, Latency, and Determinism Constraints: Operational constraints related to the method's nature and execution, including computation time, latency, and determinism, impact the selection of a data augmentation method. Thorough testing is necessary to identify the most suitable data augmentation method for a given situation, and additional research is warranted to explore these critical operational constraints further.
  The comprehensive review encompasses a detailed evaluation of commonly employed data augmentation methods for time series data, with a specific focus on the first two categories: transformation-based methods and pattern-mixing methods. The methodology employed to identify an optimal data augmentation method involves training classification algorithms on both the original and augmented datasets. The evaluation entails assessing their performance on a test dataset composed of non-augmented data. While the comparison in the evaluation study incorporates 12 data augmentation methods for benchmarking, it overlooks a recently developed and increasingly important approach: Generative Adversarial Networks (GANs), acknowledged for their significance since 2014.

Our pilot study aims to extend the evaluation conducted by incorporating GANs for time series data augmentation. Additionally, we assess their generative performance in a classification task utilizing the synthetic dataset. It is crucial to highlight that the application of GANs for generating high-dimensional multivariate time series remains an active area of research, presenting numerous potential applications.

Methodology:

GANs are a class of generative networks that use adversarial training to jointly optimize two neural networks, a generator and a discriminator. Similar to encoder-decoder networks, in order to generate samples, the GAN is trained using the training dataset and then z-vector is sampled and used with the generator to create new time series. There have been numerous time series GANs proposed. However, most target generation only and not data aug-

mentation. Due to this, we will only focus on the works that are specifically used for data augmentation.
Introduced in 2014, GANs are neural networks jointly trained in order to learn a non-linear mapping which transforms normally-distributed variables into samples which mimic the real training data and inherit their statistical properties. The goal of our pilot study was to expand the evaluation study from by including GANs for time series data augmentation, along with testing their generative performance for a classification task relying on the synthetic dataset. It is important to note that the use of GANs to generate high-dimensional multivariate time series remains an active research area with many potential applications.

For the study, we opted to utilize the Recurrent Conditional GANs architecture. This choice is based on its distinction as one of the early GANs designed for generating continuous sequential data. Its straightforward structure facilitates swift training, without being exclusively tailored to a specific data type, and it allows for conditional generation. For an in-depth exploration of available generative algorithms, readers can refer to.

All training sessions consisted of 10,000 epochs, epochs in the context of machine learning refer to the number of times a learning algorithm processes the entire training dataset. It's a measure of how many times the algorithm goes through the entire dataset to update its weights and improve its understanding of the patterns within the data, where an epoch denotes one update of the generator's weights. Given that the discriminator might undergo multiple updates per generator update, this definition holds significance. Expanding upon the model introduced in, our implementation features two discriminators: one MultiLayer Perceptron (MLP)-based and the other Long Short Time Memory (LSTM)-based. The relative impact of their feedback on the generator loss is fine-tuned through a dynamic parameter α, evolving across epochs. In the initial training phases, the MLP-based discriminator exerts a more pronounced gradient signal for the generator weights, while the LSTM-based discriminator predominantly influences the later stages of training.

The hybrid approach adopted in study is motivated by the average classification performance of models detailed in Tables 1 [8]. These models were trained on non-augmented datasets, revealing that MLP-based models outperform LSTM-based models. However, LSTM-based models prove valuable in capturing time series patterns overlooked by non-recurrent models, such as time-to-time conditional probabilities and long-short memory patterns. Optimal generative performance is achieved by leveraging both discriminators in the specified order.

Table 1. Selected datasets from the 2018 UCR Time Series Classification Archive.

Dataset	Type	Train	Test	Class	Length
FordB (D1)	Sensor	3636	810	2	500
ECG5000 (D2)	ECG	500	4500	5	140
Strawberry (D3)	Spectro	613	370	2	235

Literature Review:

The generators trained on the three datasets listed in Table 1 are publicly accessible at https://storage.googleapis.com/ucrgen/generators.zip, accessed on 29 July 2022. [8]

The primary aim of the assessment outlined in was to appraise the overall effectiveness of each data augmentation method incorporated in the investigation. Various time series datasets were scrutinized, specifically those sourced from the publicly available 2018 UCR Time Series Classification Archive, accessible AT https://www.cs.ucr.edu/~eamonn/time_series_data_2018/,[9]

In this pilot exploration, we focused on the application of Generative Adversarial Networks (GANs) for augmenting time series data. It is crucial to acknowledge that not all datasets from the public repository are suitable for GANs training due to minimum sample size requirements to prevent overfitting. The exclusion of generative model-based augmentation methods, such as GANs, from the evaluation study was justified by the necessity for external training. Additionally, time series derived through image flattening were omitted since this procedure inherently disrupts the geometrical structure of the underlying image. Notably, GANs have been extensively researched and utilized for images, with well-established models for this data type. Considering these factors, three datasets representing distinct data types—Sensor, ECG, and Spectro—were selected. These data types are drawn from domains with pronounced requirements for either data augmentation or data anonymization [10,11,12] the latter being a notable advantage of synthetic data generation. Key attributes of the selected datasets are succinctly presented in Table 1.

Challenges:

In the initial investigation, creation of synthetic data by augmenting an existing set of real data. This involved training a classifier on a combined dataset comprising both synthetic and real instances, followed by an evaluation on a genuine test dataset. As suggested in, a more rigorous evaluation approach would involve two steps: training on synthetic data and testing on real data, along with training on real data and testing on synthetic data. Detecting potential issues, such as low variability within the synthetic dataset due to factors like mode-collapse, requires careful consideration. In our pilot study, the presence of real highly variable data helped conceal potential generative defects. It is imperative to formally assess the diversity of the generated synthetic time series based on these insights.

For the datasets incorporated in study, the neural networks employed for classification exhibited overall subpar performance, with accuracy in some instances nearing random chance levels demonstrated that simpler approaches to such tasks may outperform neural network-based methods. Given that a classifier reporting ∼50% accuracy lacks informativeness regarding the effectiveness of underlying data augmentation, it becomes valuable to complement this evaluation with more precise algorithms, such as Dynamic Time Warping and its variations.

Limitations:

- -The research acknowledges the inherent challenge of evaluating generative models due to issues like mode-collapse, and it emphasizes the need for more rigorous assessment metrics.
- -The study highlights the limited applicability of GANs for certain datasets due to minimum sample size requirements, preventing overfitting. This constraint underscores the importance of carefully selecting suitable datasets for effective training.
- -The discussion points out potential generative defects within synthetic datasets, particularly in scenarios with low variability, and highlights that the presence of highly variable real data in the pilot study may obscure these issues.
- -The limitations include the suboptimal performance of neural networks in classification tasks, raising concerns about the effectiveness of these models in certain contexts. The study suggests exploring alternative algorithms like Dynamic Time Warping for more accurate assessments.
- -The research underscores the preliminary nature of the pilot study, indicating that future evaluations should explore advanced GAN architectures, incorporate Wasserstein loss functions, employ a two-step evaluation process, and conduct a formal assessment of synthetic time series diversity.
- -A recognition of the underdeveloped state of time series data augmentation in comparison to image data is highlighted. The survey points out the limited exploration of combining multiple augmentation methods for improved results, emphasizing the untapped potential for growth in this area.
- -The study emphasizes the lack of utilization of various advanced data augmentation methods prevalent in the image domain, such as style transfer, meta-learning, and filters, 1indicating a notable gap in current time series data augmentation research.
- -The survey identifies the absence of exploration into synergizing multiple data augmentation methods sequentially, with only a few exceptions in existing literature. This limitation suggests a potential avenue for further research in enhancing the effectiveness of data augmentation.
- -More research in time series data augmentation is underscored, emphasizing the need for innovative methodologies and improved model performance in this less-explored domain compared to image data augmentation.

Conclusion:

Preliminary study on GAN-based data augmentation for time series has furnished vital initial insights, laying the groundwork for future evaluations. Building upon these initial findings, forthcoming studies in this domain should explore advanced GAN architectures, incorporate Wasserstein loss functions, adopt a two-step evaluation process, and undertake a formal assessment of the diversity inherent in synthetic time series generation.
Because the idea of improving time series data by creating new versions of it is not as well-established as it is for images, there's a lot of untapped potential for making it better. In this survey and many other studies, only one way of improving the data is used for each model. However, it could be beneficial to combine different ways of improving the data to get better results. Surprisingly, not many studies have tried this approach, except for Um et al. [30]. Additionally, there are several advanced methods used to improve images, like style transfer, meta-learning, and filters, which have not been explored for time series data. This suggests that there's a lot of room for new research in making time series data better by trying out different methods and coming up with innovative ways to improve model

REFERENCES :

1. A Pilot Study on the Use of Generative Adversarial Networks for Data Augmentation of Time Series by Nicolas Morizet 1, Matteo Rizzat David Grimbert, and George Luta
2. DATA AUGMENTATION GENERATIVE ADVERSARIAL NETWORKS by Antreas Antoniou, Amos Storkey, Harrison Edwards
3. An empirical survey of data augmentation for time series classification with neural networks by Brian Kenji Iwana , Seiichi Uchida

4. Goodfellow, I., et al. (2014). "Generative Adversarial Nets."
5. Zhu, J. Y., et al. (2017). "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks."
6. Antun, V., et al. (2019). "Generative adversarial networks for brain lesion detection."
7. Yu, L., et al. (2017). "SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient."
8. Salimans, T., et al. (2016). "Improved Techniques for Training GANs."
9. Heusel, M., et al. (2017). "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium."
10. Brownlee, J. (2020). "Generative Adversarial Networks for Data Augmentation in Image Classification."
11. https://storage.googleapis.com/ucrgen/generators.zip,
12. https://www.cs.ucr.edu/~eamonn/time_series_data_2018/
13. Esteban, C.; Hyland, S.L.; Rätsch, G. Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs. [**Google Scholar**]
14. Ghorbani, A.; Natarajan, V.; Coz, D.; Liu, Y. DermGAN: Synthetic Generation of Clinical Skin Images with Pathology.
15. Frid-Adar, M.; Diamant, I.; Klang, E.; Amitai, M.; Goldberger, J.; Greenspan, H. GAN-based Synthetic Medical Image Augmentation for increased CNN Performance in Liver Lesion Classification. [**CrossRef**]