Synthetic Data Generation Using Generative Adversarial Networks for Machine Learning Applications

Dr. K. Anandharaj
Dept. of Computer Science
Sri Ramakrishna College of Arts and Science .
Coimbatore, India
anandharaj@srcas.ac.in

S.Adhi Shankar
Dept. of Computer Science
Sri Ramakrishna College of Arts and Science .
Coimbatore, India
anandharaj@srcas.ac.in

Abstract

Machine learning algorithms require vast amount of high quality data for learning. However, obtaining such practical sets for use can be a challenge due to privacy regulations, data a pittance, and utter costs for data collection and annotations. These issues pose significant barriers for researchers who rely heavily on vast data sets to engineer machine learning algorithms.

This paper presents a solution to this problem by proposing an innovative generative artificial intelligence framework that generate a large scale of synthetic data sets from Generative Adversarial Networks (GANs). The GAN framework learns the distribution of the underlying real data set, and then generate artificial samples that are statistically similar to the real data.

The synthetic data generated may be used to diversify the training data set to improve the quality and performance of machine learning algorithms. Based on experimental observation, the study reveal that the artificial data remain statistically similar to the original data set, and can be effectively apply in the field of machine learning and protect the privacy of the original data set.

Keywords: Generative Artificial Intelligence, Synthetic Data Generator, Generative Adversarial Networks (GANs), Deep Learning, Artificial Intelligence, Data Augmentation, Machine Learning.

1. Introduction

Machine learning and artificial intelligence (AI) are adopted to any domain in need to solve complex issues such as healthcare, finance, security, autonomous systems etc. On the other hand, the success of these intelligent systems relies on having a large set of training data with good quality. While data need to be collected and labeled to build AI systems is mainly hard to access for reasons of privacy, unavailability and expensiveness.

Lack of training data is one of the most fundamental issues of AI models in many applications, especially healthcare and financial systems, where a limited amount of sensitive data that is impossible to be published are available to build AI systems. Creating synthetic data appears to be an interesting technique to create large datasets without disclosing sensitive information of the original one. Thanks to last development on generative AI, especially Generative Adversarial Network (GAN), intense and realistic images or tables of synthetic data can be generated to develop ML systems.This paper suggests a GAN model to generate synthetic data in the purpose of machine learning applications.

2. Related Work

Synthetic data generation has been of growing interest in recent years, mainly when facing the data scarcity, privacy preservation, and expensive data collection bottleneck. Traditional machine learning algorithms require huge amount of data to reach promising performance, but in real world applications data collection is complicated or infeasible in the domains of healthcare, finance, and security, etc., because of privacy concern and data availability. Generative Adversarial Networks(GANs), pioneered by Ian Goodfellow et al., have become one of the most promising methods to generate synthetic data.

A GAN consists of a generator and a discriminator, which are adversaries during the training. Several research findings have shown that synthetically augmented data via GANs can reduce issues related to data sparseness, increase the accuracy of machine learning algorithms, and kept sensitive data private from a third party. Research on

controlling the quality of the generated data is still an open issue.

3. Problem Statement

Machine learning algorithms and systems require large and diverse quantities of data to enable the highly accurate and reliable system. However, other than the high cost involved in collecting such a large amount data, real-world datasets are not always accessible to use for training because of issues such as privacy and sensitive data restrictions and constraints, limited data availability or the unavailability of data sets at all. This lack of training data poses one of the biggest issues to researchers and developers working in the field with machine learning applications as the available data if limited, will generally mean the system can suffer from issues such as overfitting and lack of generalization, poor prediction accuracy due to the limited amount of training data can severely hamper the system. Another problem is the need for ever increasing amounts of data to train advanced deep learning architectures that require ideally thousands or even millions of data samples. Collecting such large quantities is often time-consuming and costly. It is therefore clear that there is a demand for intelligent methods to generate realistic artificial data sets that can sustain the training of such systems.

Synthetic data generation, using generative systems such as GANs provides a good candidate technology to generate such datasets which can be achieved to support machine learning training to gain more accurate systems

.4. Proposed methodology

The proposed methodology is based on creating artificial data by relying on generative models, specifically the Generative Adversarial Networks (GANs). The main goal of this methodology is overcoming the problem of limited training data in machine learning and specifically in deep learning applications. This methodology comprises data collection, data preprocessing, training the generator model, and generation of synthetic data models.

The first step is to gather real data and prepare it for training the generator model. For this research using a ready made dataset such as the MNIST handwritten digit dataset is a good example of a real dataset. Next, the data has to be preprocessed through normalization of the images and the scoping of data into appropriate trains and tests. Proper data would optimize the training process as well as help the model learn effective features from training dataset.

Finally, the core element of the framework is the GAN, which consists of the generator and the discriminator. Both are deep neural networks which play a competition game: while the generator is responsible for creating artificial data given some initial noise, the discriminator understands if the data is real or fake. During practice, the generator gets set against the discriminator and learns by competing of the real data distribution giving rise to subsequent artificial data that helps training in machine learning leading to better performance.

4.1 Architecture Overview

The architecture of the proposed synthetic data generation system is modular that allows the generating of artificial data set efficiently by using the Generative Adversarial Networks (GANs). The architecture of the proposed system is built with a modular approach such that each module in the architecture performs a specific task while needed to connect with the other modules of the system. Modular design improves the scalability of system, easily implement and smooth manage of overall synthetic data generation framework process.

The overall architecture of the proposed system is designed on the abstraction with the four modules, namely the Dataset Layer, Data Preprocessing Layer, GAN Model Layer and Synthetic Data Output Layer. Each layer of the system provides an important function in the overall system. System starts from collection of original data set to the transformation of input data for training purpose and generate a variety of sophisticated synthetic samples.

Layered architecture also supports individual management of component in the system and distinct separation of functions in the system. The Dataset Layer of the architecture provides the real data set which is used for training purpose of the generative model. The data preprocessing layer prepares the specific data for training in form of providing data cleaning, normalization and also formatting data for training procedure.

The core of architecture, the GAN model layer provides the learning for the generative networks where the generator and discriminator networks mastered the distribution of input data. The final layer of synthetic data output produces the artificial samples which pseudo qualified to original data set and useful for training of machine learning technique.

4.2 Architecture Diagram Description

The architect diagram shows the flow of proposed synthetic data generation system based on GAN. The input of system is a real dataset such as MNIST hand written digit dataset provided as input to the system. This dataset is used to train the generative model.It then passes the dataset to Data Preprocessing Layer where it is cleaned and normalized followed by formatting it to prepare it to train the model.

The dataset is then send to Generative Adversarial Network Model where the training of generator/ discriminator begins to learn the dataset patterns and generate the artificial training samples..

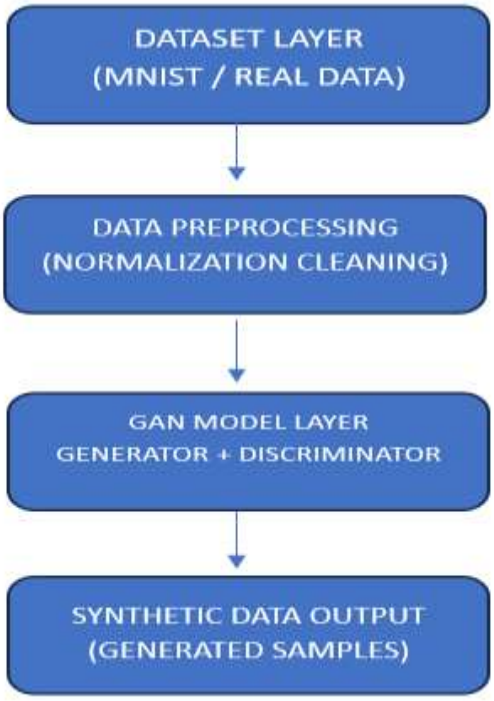

Figure 1: Architecture Diagram

finally send to synthetic data output layers where the generated samples are very similar to the dataset and used to build the large training dataset that enhances the learning of the model

4.3 Device Layer

The Device Layer constitutes the first step of the architecture of the synthetic data generation and it is the first block where the input data are fed into the system. In this layer, the real data set used to train the generative model is extracted from open-data repositories or existing ones. For the purposes of our work, a frequently used data set such as the MNIST data set of handwritten digits can be utilized as base data set.

In addition, data set collection will bring thousands of images of handwritten digits, which can be efficiently used for the learning process of the model, providing the necessary visual patterns and data features. The collected data set will then be used for training the generative model. These instances will help the system to identify the necessary statistical patterns and distribution of the initial data set.

The main function of the Device Layer is to extract and process usable input data for any kind of further processing. When the data is harvested, then it is transferred to the Data Preprocessing Layer for the training purposes.

4.4 Data Preprocessing Layer

Data Preprocessing Layer is responsible for preprocessing the gathered dataset so it can be used by the model for generative purposes. As raw dataset are unlikely to be perfect for machine learning, common dataset problems are addressed during data preprocessing. Results of the preprocessing are then used during training of the model.

Data preprocessing layer performs several operations on dataset, for example it normalizes images, resizes images and prepares dataset in proper format. Dividing dataset in train and test dataset is also part of data preprocessing. It is used to train the model by training data set and to validate the quality of synthetized data by testing dataset.

All in all, data preprocessing enhances the training process of the model..

4.5 GAN Model Layer

GAN Model LayerThis is the main layer of the framework from where, the actual process of synthetic data generation begins. It uses Generative Adversarial Networks that learns the pattern of the original data and generate artificial data samples. A GAN is made up of two neural networks, generator and discriminator.

Generator creates artificial data samples from various user-defined noise, whereas the discriminator classifier analyzes whether the generated samples are authentic or fake. These two neural networks pseudo-competitively train against one another over time. The generator produces more and more authentic data samples, on the other hand, the discriminator gets better and better at telling real samples apart from synthetic ones..

4.6 Synthetic Data Output Layer

The Synthetic Data Output Layer is the final component of the proposed architecture. Here the trained generator network generates some new artificial data samples similar to the original dataset. The new generated samples follow the same schemes and information distribution.

Therefore the generated samples have the same statistical properties as the real data, so that the generated data could be used to extend the original dataset. This would result in a more comprehensive and diverse training dataset for the machine learning model and leads to better generalization performance. Meanwhile, the performance of the generated data could be obtained using other performance metric calculation.

5. System Operation

The process of operation of the proposed system for synthetic data generation is initiated by collection of the original data set which is to be used for training generative model. Preprocessing operations such as cleaning, normalization are performed on the collected data set to prepare it to be fed as input to the model. Then the preprocessed dataset is fed as input to the targeted generative model.

The GAN model, then endures an iterative process of training in which generator produce fake samples while discriminator checks whether the sample is a genuine sample or a generated one. As the training continues, the generator improves its output to correctly predict the output thus the generated samples increasingly follow the original dataset. Once the training is over, the generator produces the artificial data samples which statistically follow the original dataset.

5.1 Data Acquisition

Data acquisition is responsible for obtaining the actual dataset for training the generative model. As mentioned previously in this thesis, a number of publicly available datasets e.g. the MNIST handwritten digits dataset are used for training the generative model. Such datasets offer a high number of samples from which the generative model can learn various useful characteristics and properties from real data.

The collection of real data is required to be used for training the synthetic data generation system.

5.2 Data Preprocessing

Once the dataset has been gathered, it gets preprocessed. This is where the data is prepared for training the model. This includes cleaning up data, normalizing it, and structuring the dataset in appropriate folders and formats.Preprocessing helps the training process by having the dataset be better standardized and more uniform. With this, the data the model trains on is more organized and can produce better training examples of synthetic data

5.3 GAN Model Training

At this stage the data is run through the Generative Adversarial Network to learn the features of the studied data set. This type of neural network contains two networks. The generator network creates fake data samples generated from random noise.The discriminator network detects the type of the data sample being studied as either real, fake or generated. After continually being trained in the features for the data, these two networks learn to; make the generator network generate more realistic data samples.

5.4 Synthetic Data Generation

Once can train the generator network, it will generate artificial data samples, which has similar properties as the original data set. The artificial samples will be similar to original data samples in terms of pattern or statistical distribution. The synthetic data can now be used to augment existing datasets and help train the machine learning models.

In order to preserve the privacy of original data it is always used in combination of it

.6. Proposed System vs Existing System

Explanation:

Compared to standard methods, that build AI systems only on real world data, the proposed system based on GANs for synthetic data generation presents several benefits. In standard systems, ML models are trained using only available real data, while the new system can generate new artificial data to train the model with, preventing in part this problem of unavailability. Moreover, the synthesized data serves to preserve important information from data release; finally, the

proposed method could be cheaper and contain less efforts.

Table

Feature	Existing system	Proposed GAN-Based System
Data Source	Depends only on real-world datasets	Uses both real datasets and GAN-generated synthetic data
Data Availability	Limited data availability	Synthetic data generation increases dataset size
Data Privacy	Risk of exposing sensitive information	Synthetic data protects privacy
Dataset Expansion	Difficult and time-consuming	GAN automatically generates additional data
Model Training	Limited training samples may affect performance	Larger datasets improve model training
Model Performance	Lower accuracy due to small datasets	Improved performance with augmented datasets

Table 1: Comparison of Proposed GAN Based system vs Existing System

7. Performance Evaluation

The system performance is then evaluated in terms of how well the artificial data produced from the proposed system improve the machine learning modeling process. This evaluation was made based on the advantage of the construction of the larger training samples, reduced training time, and the resemblance of real data sample and the generated data sample. The original dataset was expanded through the artificial data samples built from the GANs, which results in bigger and more diverse dataset.

The experiments shown in this research result in the models trained with both real data and artificial data can lead to better generalization and prediction.

8. Graph Analysis

Graph analysis the following graphical results clearly indicate some benefit of the proposed synthetic data generation framework, for example: increase in data size after synthetic data generation, comparison of accuracy of models while trained on real data only and real data along with synthesized data.

8.1 Latency Comparison Graph

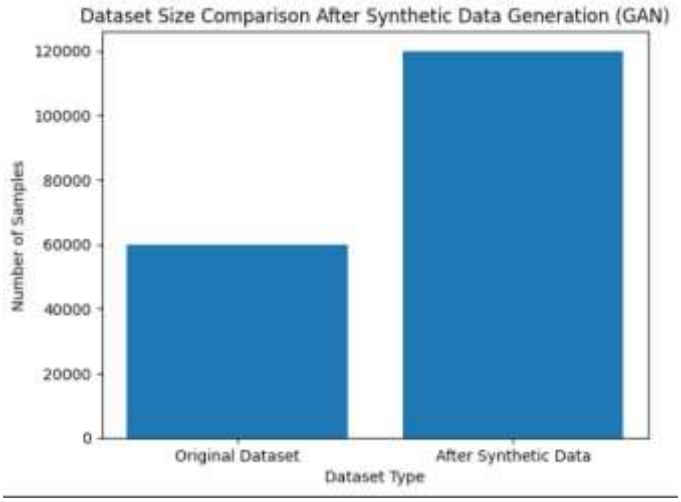

Figure 2: Latency Graph

Explanation:
The comparison between dataset sizes represents the expansion of the dataset size due to the new sample generator based on GAN method. There is just a small sample size in the original dataset that is used to train the machine learning model. The synthetic data is generated by the GAN model and added into the original dataset.

In this way, the size of the overall dataset size is improved greatly for better model training and prediction.

8.2 Model Accuracy Comparison

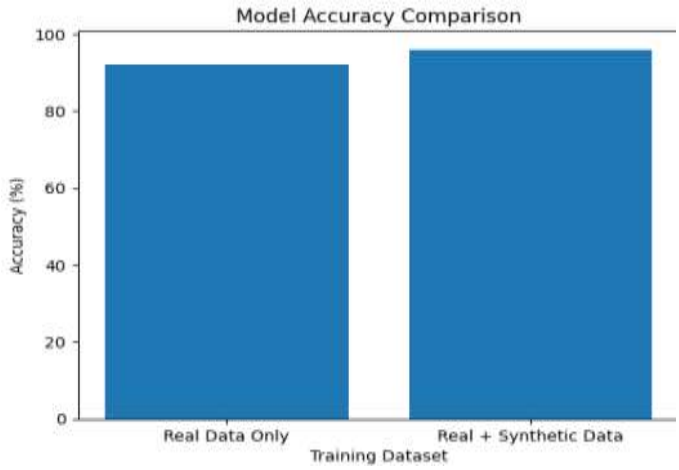

Figure 3: Performance Graph

Explanation:
This graph compares the accuracy of the model being trained with the original data only and with both real and synthetic data. The addition of synthetic data to the training set increased the accuracy of the model and should increase the machine learning ability of the model..

8.3 GAN Training Loss

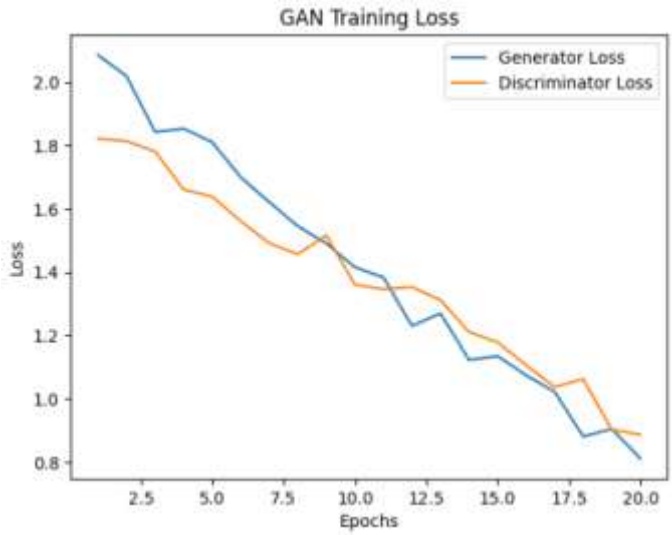

Figure 4: Security Comparison Graph

Explanation:
The GAN training loss graph represents the variation in the generator and discriminator loss throughout the training. It can be seen that as training continues both networks are slowly getting better. The decreased loss values imply the generator is creating more realistic synthetic data samples..

8.4 Summary

This project proposed a framework of synthetic data generation to overcome the problem of small datasets available for training in machine learning applications by means of a Generative Adversarial Network. The framework generate a set of synthetic data who can be used to emulate a real dataset without losing its statistical features. This approach contributes to a significant improvement in training time and model performance.

The advantages for synthetic data are the ability to optimize machine learning training time and protect data privacy.

9. Research Challenges

However, the GAN based synthetic data generation system proposed also has several issues when being implemented. The first issue is that training GAN model is hard. If the generator and discriminator are not maintained in balance during training, the model may collapse and unable to generate realistic data samples.

The second issue is that how to evaluate synthetic data quality. If the generated data samples do not faithfully reflect the properties of training data, this may hurt the performance of machine learning algorithm.

10. Conclusion

The author proposed a GAN-based framework of synthetically generate data to mitigate the limited dataset problem in machine learning applications. The framework produce realistic artificial data that resemble real dataset by retains the same statistical features. enlarging the original dataset by a dataset comprised of real and synthesized data thus enhances the size of dataset and improve machine learning application performance, in the meanwhile, protect data privacy.

11. Future Work

In order to further enhance the proposed framework for synthetically generating data, future studies can test the framework on more advanced generative models including Conditional GANs or diffusion based models. This way, better quality of individual synthetic data points may be achieved. Moreover, the framework can be also tested on larger and more complex datasets across domains including healthcare, finance and cyber security.

Improvements to the measurement system and reinforcement learning training may also allow better synthesis framework to be created..

References

- **I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Networks," Advances in Neural Information Processing Systems (NeurIPS), 2014.**
- **A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. Bharath, "Generative Adversarial Networks: An Overview," IEEE Signal Processing Magazine, vol. 35, no. 1, pp. 53–65, 2018.**
- **M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets," arXiv preprint arXiv:1411.1784, 2014.**
- **T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved Techniques for Training GANs," Advances in Neural Information Processing Systems (NeurIPS), 2016.**
- **L. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Synthetic Data Augmentation using GAN for Improved Liver Lesion Classification," IEEE Transactions on Medical Imaging, 2018.**
- **A. Radford, L. Metz, and S. Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," arXiv preprint arXiv:1511.06434, 2015.**