ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 12, December 2024

Enhancing Incomplete Image Datasets Using GANS: Generation and Integration

[1] C.H. Priyanka, [2] G. Ramya Sree, [3] Lasya Nama, [4] Dr. Sabitha. P

[1] [2] [3] Department of CSE-BDA, SRM Institute of Science and Technology, Ramapuram, Chennai, India
[4] Assistant Professor, Department of CSE, SRM Institute of Science, and Technology, Ramapuram, Chennai, India
Corresponding Author Email: [1] cc0560@srmist.edu.in, [2] gg5402@srmist.edu.in, [3] jt6466@srmist.edu.in, [4] sabithap@srmsit.edu.in,

***Abstract*— The dependability of AI and machine learning models in fast-evolving fields relies on having access to thorough and high-caliber data. Nevertheless, the problem of missing or incomplete data is widespread and can greatly reduce the accuracy and effectiveness of trained models. This study is centered on tackling this difficulty by using MISGAN, a new method created to produce missing images in datasets. The goal is to create lifelike and top-notch images that smoothly fill in the areas left empty due to missing data. This approach enhances the value of existing datasets and contributes to the development of more accurate and reliable machine learning models by leveraging the power of MISGANs.**

Index Terms*— Generative Adversarial Networks (GANs), Missing Data, Incomplete Data, MisGAN, Data Imputation, Machine Learning, MNIST, Missing Completely at Random (MCAR), Mask Generator.

I. INTRODUCTION

Generative Adversarial Networks (GANs) are transforming the computer science field, providing some of the most exciting breakthroughs in artificial intelligence and data science at present. The core of GANs is an interesting idea: the adversarial training process. In this procedure, two neural networks, referred to as the generator and the discriminator, are set in opposition to each other. The generator is responsible for producing images that appear as realistic as they can, while the discriminator's function is to differentiate between authentic images and those generated by the generator. Both models are driven by this adversarial relationship to constantly enhance their performance. The generator is pushed to produce more realistic images, while the discriminator becomes more adept at distinguishing real from generated images.

Data augmentation, specifically in dealing with the prevalent problem of missing data points in image datasets, is one of the major uses of GANs. Conventional approaches to managing missing data usually include interpolation or utilizing available data to predict the missing values. However, these methods may not be effective when dealing with intricate image data. However, GANs provide a sophisticated alternative. By creating artificial images to fill in the missing areas of a dataset, GANs not only supplement the absent data points but also improve the overall excellence of the dataset. The generator is designed to produce images that are indistinguishable from real images in order to preserve the integrity of the dataset.

Incorporating these artificial images into current datasets is an essential procedure to maintain uniformity and cohesion in the data. Adding realistic and high-quality images to fill in the gaps in datasets enhances the overall robustness and reliability of the dataset for training machine learning models. As a result, the performance and accuracy of these models are enhanced because they are trained with more comprehensive and reliable data. Utilizing GANs for filling in and increasing data is a potent resource for the progress of AI, enabling researchers and developers to conquer a long-standing obstacle in the field: the problem of inadequate or absent data.

II. LITERATURE REVIEW

[1] Proposes an advanced web-phishing detection and protection scheme integrating features from multiple images. It aims to enhance accuracy by comprehensively analyzing visual and textual elements. Machine Learning techniques are used to provide security.[2] Introduces Phishing detection method with the usage of image retrieval based on an improved Texton Correlation Descriptor. It aims to provide a more reliable and efficient defense against phishing attacks through advanced image processing and pattern recognition.[3] Explores the use of deep learning techniques for image detection, leveraging neural networks to analyze and identify phishing websites. Enhances the efficiency and effectiveness of phishing detection compared to traditional methods.[4] Provides a comprehensive review of various phishing detection methods, categorizing and evaluating techniques based on their approaches and effectiveness. Covers both traditional and modern methods, including machine learning and heuristic based strategies.[5] Presents a high accuracy phishing detection system using Convolutional Neural Networks. This method outperforms traditional machine learning techniques, offering enhanced security against phishing attacks.

[6] Explores the use of various machine learning algorithms to identify and prevent phishing attacks.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 12, December 2024

Highlights the effectiveness of techniques such as decision trees, random forests, support vector machines in detecting phishing websites by analysing key features in URLs and web content.[7] Focuses on developing an intelligent system to detect phishing attacks in online environments using advanced machine learning techniques. Emphasizes the integration of feature extraction, classification algorithms, and real-time analysis to accurately identify phishing threats.[8] Presents a novel approach to phishing detection by leveraging search engine results to identify and block phishing websites. Enhances detection accuracy by cross-referencing suspicious URLs with search engine indices and analyzing discrepancies in search results.[9] Introduces a method for detecting and preventing phishing attacks by monitoring web folders and employing customer image verification. Combines regular scanning of web directories for suspicious content with the verification of user-uploaded images to authenticate legitimate users [10] Explores the application of deep learning techniques to identify phishing URLs. Employs Convolutional Neural Networks (CNNs) to analyze URL structures and detect patterns indicative of phishing attempts.

[11] Presents a deep learning framework specifically designed to detect and prevent image-based spam. The framework uses Convolutional Neural Networks (CNNs) to analyze visual and textual features in images to accurately identify spam content.[12] Proposes a method that integrates both textual and visual features to detect cloned phishing websites. Enhances early phishing detection by leveraging a comprehensive analysis of both the textual and visual similarities between legitimate and cloned webpages.[13] Reviews various deep learning approaches applied to phishing detection. Highlights the strengths and limitations of different models, such as Convolutional Neural Networks (CNNs) and recurrent neural networks (RNNs), in identifying phishing threats.[14] Utilizes neural networks, particularly Convolutional Neural Networks (CNNs), to analyze and classify website features indicative of phishing. Demonstrates that deep learning techniques significantly enhance the accuracy and reliability of phishing website detection compared to traditional methods.[15] Explores the use of Convolutional Neural Networks (CNNs) to identify and filter out image-based spam. Shows that CNNs provide a high accuracy rate in detecting image spam, outperforming traditional image analysis methods.

III. PROPOSED WORK

A. Modified MisGAN Architecture:

The proposed work involves modifying the MisGAN architecture to improve the image quality, further shaping it to fit the model description and problem statement. That is, it includes introducing additional layers, activation functions, and/or connections to enhance the performance of the generator and the discriminator. This method depends on exploring different architectures like adding attention mechanisms to focus on specific image regions or using dilated convolutions to increase receptive field or incorporating conditional information to control image generation. Therefore, using this model assures high-quality work and provides better images.

B. Hyperparameter Tuning:

The proposed work also involves experimenting with multiple loss functions and hyperparameters, thus optimizing the model's performance. Effects of factors like learning rate, batch size, and number of training iterations are investigated to understand their impact on image quality. To be precise, learning rate and batch size are looked into for their effects on training stability and image quality, while the number of training iterations on convergence and overfitting. Weight initialization methods are observed for model performance and regularization techniques on overfitting.

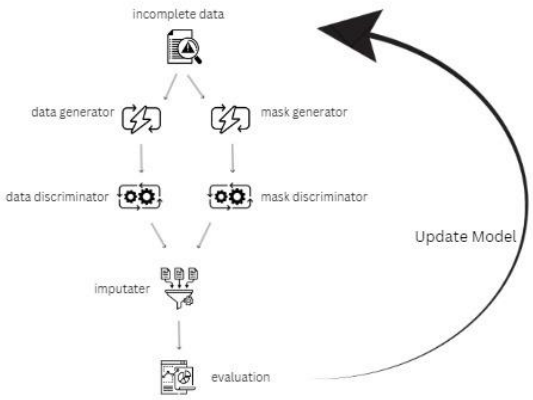

Figure 1. System Architecture

IV. METHODOLOGY

The proposed work can be implemented by implementing the following steps.

A. Data Preparation:

Data preparation includes collecting a dataset of images for training and testing. While doing so, it is important to ensure that the dataset is diversity and quality. Once dataset is collected and trained, images in the dataset are resized to uniform dimensions. This includes normalizing the pixel values to lie between 0 and 1. The images are then converted to suitable image formats, whether it be RGB or Grayscale.

B. Model Architecture:

Once data is prepared, a model must be created. The generator and the discriminator are designed here based on the problem statement. Appropriate layers are considered and chosen. These layers may include convolutional, transposed convolutional, and batch normalized. Activation functions are then selected. Later, output dimensions for the generator and the discriminator are defined based on the previous steps.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 12, December 2024

C. Training:

To train the model, the following functions should be defined. Each of these functions serves different purposes in the model. Firstly, the adversarial loss function must be defined, followed by the reconstruction loss function, and finally the total loss function. After the functions are defined, select the optimizer and hypermeters. Train the generator and discriminator, alternately.

D. Evaluation:

The images generated can be evaluated for their quality using metrics like Peak Signal-to-Noise Ratio (PSNR) and ***Structural Similarity Index Measure (SSIM). The proposed*** work is then compared to existing methods, to make sure that there is an improvement in the model. Moreover, Human evaluation studies are conducted to evaluate metric effectiveness.

E. Input/Output:

The input and the output refers to the data that flows into and out of the system.

a. Input:

Two types of possible inputs can be given to the model, namely Noise vector and Real image. The Noise vector is a random vector used as input to the generator to produce a synthetic image. While a real image is given as an input to the discriminator to evaluate its authenticity.

b. Output:

The model generates three possible outputs. They are Synthetic Image, probability of Real Image, probability of synthetic image. The output of the Generator when given the noise vector as input is called Synthetic image. This is the generated image that aims to mimic the real image. The output of the Discriminator when given the real image as input is said to be the probability of the generated image being the real image. This is the probability that the real image is real. The output of the Discriminator when given the synthetic image as input is said to be the probability of Synthetic image. This is the probability that the synthetic image is genuine.

Here the inputs and outputs are crucial in training the Generator and Discriminator networks. The Generator is amied to produce synthetic images that can fool the Discriminator, while the Discriminator aims to distinguish between real and synthetic images correctly.

F. System Flow:

The system flow consists of six steps. They are Data input, Generator forward pass, Discriminator forward pass, Loss calculation, Backward pass, and output. Every one of these steps are very crucial for the model to run smoothly.

V. RESULTS

A. MisGAN for the Imputation of Missing Data

This research assessed how well MisGAN performed in filling in missing data when classifying MNIST digits. The findings show that MisGAN is successful in producing realistic and precise replacements for incomplete data.

B. Main Discoveries:

Effective data generation: MisGAN successfully created MNIST digits that were visually indistinguishable from real examples.

Efficient imputation: The imputer part of MisGAN effective substitute missing values, leading to accurate and precise completions.

The model had shown adjusting to missing data by generating high-quality results regardless of the amount of missing information

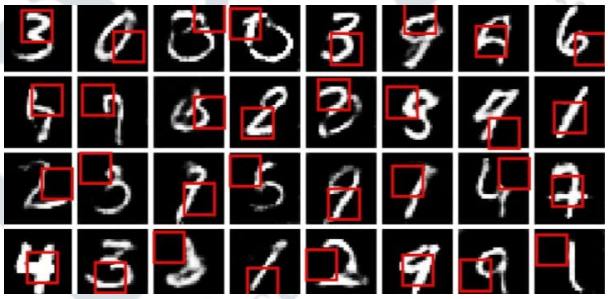

Figure 2. Result 1 (Epoch 99).

The MisGAN framework proved to be computationally efficient and able to be trained on a standard GPU, despite its complexity.

In general, MisGAN appears to be a hopeful method for filling in missing data in tasks involving images. It provides a versatile and efficient solution for managing incomplete data.

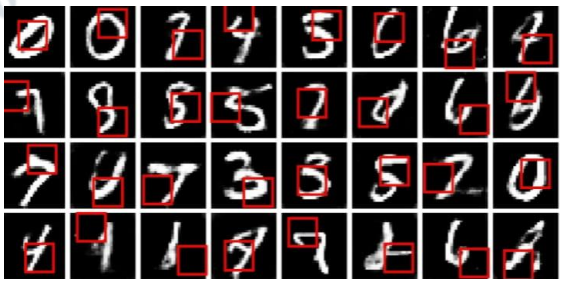

Figure 3. Result 2(Epoch 199)

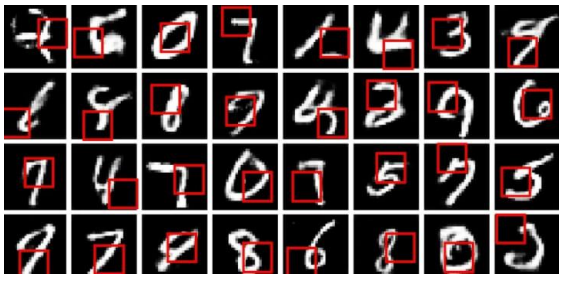

Figure 4. Result 3(Epoch 299)

VI. CONCLUSION

In our paper, we presented MisGAN, a visionary framework that utilizes Generative Adversarial Networks (GANs) for dealing with incomplete datasets. The MisGAN consists of two main elements: a data generator that fills in missing parts to form complete data, and a mask generator that represents the patterns of missing data. These togther allows MisGAN to perform well in both creating new data and filling in missing values. When evaluated using the MNIST dataset, MisGAN showed its efficacy and suitability for handling more intricate data situations. This development makes a hefty impact on machine learning and data analysis, offering a strong tool for handling incomplete datasets in a variety of applications.

REFERENCES

[1] Pattanayak, D., & Patel, K. (2022, January). Generative adversarial networks: solution for handling imbalanced datasets in computer vision. In 2022 International Conference for Advancement in Technology (ICONAT) (pp. 1-6). IEEE.

[2] Wan, Q., Guo, W., & Wang, Y. (2024). SGBGAN: minority class image generation for class-imbalanced datasets. Machine Vision and Applications, 35(2), 22.

[3] Huang, Y., Jin, Y., Li, Y., & Lin, Z. (2020). Towards imbalanced image classification: a generative adversarial network ensemble learning method. IEEE Access, 8, 88399-88409.

[4] Shoohi, L. M., & Saud, J. H. (2020). DCGAN for Handling Imbalanced Malaria Dataset based on Over-Sampling Technique and using CNN. Medico-Legal Update, 20(1).

[5] Andresini, G., Appice, A., De Rose, L., & Malerba, D. (2021). GAN augmentation to deal with imbalance in imaging-based intrusion detection. Future Generation Computer Systems, 123, 108-127.

[6] Zhai, J., Qi, J., & Zhang, S. (2022). Imbalanced data classification based on diverse sample generation and classifier fusion. International Journal of Machine Learning and Cybernetics, 1-16.

[7] Ali-Gombe, A., & Elyan, E. (2019). MFC-GAN: Class-imbalanced dataset classification using multiple fake class generative adversarial network. Neurocomputing, 361, 212-221.

[8] Pavan Kumar, M. R., & Jayagopal, P. (2021). Multi-class imbalanced image classification using conditioned GANs. International Journal of Multimedia Information Retrieval, 10(3), 143-153.

[9] Schonfeld, E., Ebrahimi, S., Sinha, S., Darrell, T., & Akata, Z. (2019). Generalized zero-and few-shot learning via aligned variational autoencoders. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8247-8255). Yu, Z., Jiang, X., Zhou, F., Qin, J., Ni, D., Chen, S., ... & Wang, T. (2018).

[10] Melanoma recognition in dermoscopy images via aggregated deep convolutional features. IEEE Transactions on Biomedical Engineering, 66(4), 1006-1016.Tai, Y., Yang, J., Liu, X., & Xu, C. (2017).

[11] Memnet: A persistent memory network for image restoration. In Proceedings of the IEEE international conference on computer vision (pp. 4539-4547). Dinov, I. D., Heavner, B., Tang, M., Glusman, G., Chard, K., Darcy, M., ... & Toga, A. W. (2016).

[12] Predictive big data analytics: a study of Parkinson's disease using large, complex, heterogeneous, incongruent, multi-source and incomplete observations. PloS one, 11(8), e0157077.Shahbazian, R., & Greco, S. (2023).

[13] Generative Adversarial Networks Assist Missing Data Imputation: A Comprehensive Survey & Evaluation. IEEE Access.Anwar, S., & Li, C. (2020).

[14] Diving deeper into underwater image enhancement: A survey. Signal Processing: Image Communication, 89, 115978.Huang, Y., Liu, M., & Yuan, F. (2021).

[15] Color correction and restoration based on multi-scale recursive network for underwater optical image. Signal Processing: Image Communication, 93, 116174.Wang, Y., Huang, F., Zhang, Y., Feng, R., Zhang, T., & Fan, W. (2020). Deep cascaded cross-modal correlation learning for fine-grained sketch-based image retrieval. Pattern Recognition, 100, 107148.