JIRBDAI

Journal of Institutional Research, Big Data Analytics and Innovation

MITIGATING MODEL COLLAPSE IN GENERATIVE MODELS THROUGH HYBRID APPROACHES: A FOCUS ON VAE-GAN INTEGRATION

Okebule Toyin*, Oguntimilehin Abiodun
Department of Computing, Afe Babalola University, Ado-Ekiti, Nigeria
*Corresponding Author: okebulet@abuad.edu.ng

ABSTRACT: Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) are examples of generative models that have advanced the field of artificial intelligence by producing realistic synthetic data. However, they often suffer from model collapse, which dramatically reduces the diversity and quality of generated samples. This study explores the possibility of integrating GANs with VAEs (VAEGAN) to mitigate model collapse, identifying key success factors such as architecture, training objectives, and hyper-parameters. The study demonstrates that the VAEGAN approach improves output diversity and quality, achieving an FID score of 35 and an IS score of 4.8, outperforming GAN (FID: 50, IS: 4.2) and VAE (FID: 20, IS: 3.5). The results show that VAEGAN achieves a loss value of 0.6, outperforming GAN's loss value of 0.8, while VAE shows the lowest loss value of 0.4. The study evaluates the performance of VAEGAN using the MNIST dataset, showcasing its ability to generate high-quality images with improved diversity. The findings suggest that the integration of VAE and GAN can effectively mitigate model collapse, leading to more robust and reliable generative models. The study confirms the existence of model collapse and highlights the benefits of VAE-GAN integration in strengthening generative modeling, paving the way for further research in this area. With the promising results achieved in this study, the VAEGAN approach has potential applications in various fields, including image and video generation, data augmentation, and anomaly detection. This study contributes to the development of more advanced generative models, enabling the generation of high-quality and diverse samples with potential applications in various fields.

KEYWORDS: Generative Models, Model Collapse, VAE-GAN Integration, Deep Learning and Hybrid Approaches

I. INTRODUCTION

Artificial intelligence (AI) has advanced quickly enough to be revolutionary and has surmounted challenges such as model collapse, which compromises AI's dependability and efficiency. When AI models, particularly those used in generative modeling and reinforcement and lose sight of their training data distribution, this phenomenon is known as model collapse [1]. This may result in outputs learning, degenerate that are illogical or not reflective of the datasets used for training. It's also critical to comprehend and steer clear of Model Collapse when the demand for high-quality data creation rises [2]. Model Collapse has real world consequences, such as hallucinated outputs in AI systems [3]. Model collapse in AI like GANs may narrow output range and miss data diversity [4]. Reinforcement learning agents may adopt suboptimal policies early, missing the better ones [5]. Model Collapse may stop AI from progressing if low-quality data runs out soon, which is a concern [6]. AI has already progressed far enough to overcome challenges like model collapse which affect AI efficiency and reliability [7]. AI has progressed far enough to overcome challenges like model collapse which affect AI efficiency and reliability [8]. When AI models, particularly in generative modeling and reinforcement learning, lose track of training data, this is model collapse [1]. Wide-ranging effects of Model Collapse include anomaly detection, data augmentation, and the creation of images and videos, among other AI applications [7]. Model Collapse in picture development can lead to the creation of unrealistic, low-quality images that miss the subtleties of the underlying data distribution [8]. The performance of machine learning models may suffer as a result of Model Collapse in data augmentation, which might produce redundant or unnecessary data [9]. Model Collapse in anomaly detection can lead to the inability to spot odd trends or outliers, which can have serious repercussions in applications like cybersecurity or fraud detection [10]. Researchers have looked into a number of ways to overcome the problems caused by Model Collapse, such as altering the generative models' architecture [11], altering the goals and methods of training [12], and applying regularization techniques [13]. Furthermore, current approaches have drawbacks that need

to be addressed, such as the inability to prevent the method collapse, limited data quality, an inadequate balance between exploration and exploitation, and inadequate handling of complex data distributions. Previous techniques have not been exceptionally effective, and there is still room for advancement in terms of minimizing Model Collapse [14].

II. LITERATURE REVIEW

Techniques like adversarial training and different training data augmentation were suggested by later research, including Salimans *et al.* in [16], to lessen mode collapse in GANs. These methods have shown promise in improving output diversity, but they can be computationally expensive and may not completely eliminate mode collapse. Model collapse in reinforcement learning can show up as agents failing to sufficiently explore the environment or converge to less-than-ideal rules. The significance of exploration in reinforcement learning was highlighted by Sutton and Barto in [17], who pointed out that inadequate exploration can result in insufficient learning and subpar performance. Exploration strategies like epsilon-greedy have been effective in improving agent performance, but balancing exploration and exploitation remains a challenge. Thomaz *et al.* in [18] proposed methods like intrinsic motivation and epsilon-greedy exploration to motivate agents to better investigate their surroundings. These methods can encourage agents to explore novel states and actions, but they may not be effective in complex environments with large state spaces. Model collapse is another issue with natural language processing (NLP), especially when it comes to text production jobs. Holtzman *et al.* in [19] showed that language models can lack the richness of genuine language by displaying repeating sentences and a small vocabulary. Recent studies have proposed novel architectures and training objectives to improve language model performance, but model collapse remains a persistent issue in NLP. The use of model collapse has been investigated in a number of fields in recent years. For example, examined the phenomena of model collapse in recommender systems Li *et al.*, in [20], emphasizing the necessity of individualized and varied recommendations. Personalized recommendations can improve user engagement and satisfaction, but model collapse can result in reduced diversity and novelty in recommendations. Recent studies have proposed novel architectures and training methods to mitigate model collapse. For instance, Radford *et al.* in [22] presented a novel class of generative models that exhibit enhanced performance and diversity by combining the advantages of GANs and VAEs. These hybrid models can leverage the strengths of both GANs and VAEs, but they can be computationally expensive and require careful tuning of hyper-parameters. Additionally, studies have looked into how model collapse relates to other machine learning phenomena, including adversarial attacks and overfitting. Another author in [24] shows that model collapse can make models more susceptible to adversarial attacks, emphasizing the necessity for safe and reliable models. Understanding the relationship between model collapse and other machine learning phenomena can inform the development of more robust models, but addressing these issues can be challenging and requires ongoing research. Karras *et al.* in [23] have suggested a novel method for training GANs that uses a progressive expanding of networks to increase stability and variety. This approach can lead to more realistic and diverse outputs, but it may require significant computational resources and expertise to implement effectively. Additionally, model collapse is a complex issue that affects various machine learning domains, and addressing it will require continued research and development of novel techniques and architectures [24]. The necessity for more robust and dependable models, the challenges involved in establishing a balance between exploration and exploitation in complex environments, and the limited efficacy of current strategies for minimizing model collapse highlight the need for greater study. In order to improve model performance, variety, and reliability, this study intends to explore new architectures, training methodologies, and strategies in order to solve model collapse.

III. MATERIALS AND METHODS

One possible remedy for model collapse in generative modeling is the hybridization of GANs with VAEs. The objective of current research in this field is to create generative models that are more varied and resilient. In order to collect crucial data on system needs, specifications, and methods, this study will make use of descriptions of model collapse found in pertinent current literature. To quantify the occurrence and consequences of model collapse in models over time, a new framework called VAEGAN is based on Adversarial Vibrational Auto encoders was employed. Specialized GAN evaluation metrics was used to gather data on correlations between model collapse, data variability, data type, and the distribution of the latent space. The design of the web applications will be done on Flutterflow, and the models and data generation was hosted on Google's Firebase database management system. Effective design of the system were

required in-depth knowledge of data-gathering techniques and different Machine Learning algorithms. The dataset consisted of readily available image data from the Tensor Flow MNIST fashion library. The description of this study contained in the system block diagram as shown in Figure 1.

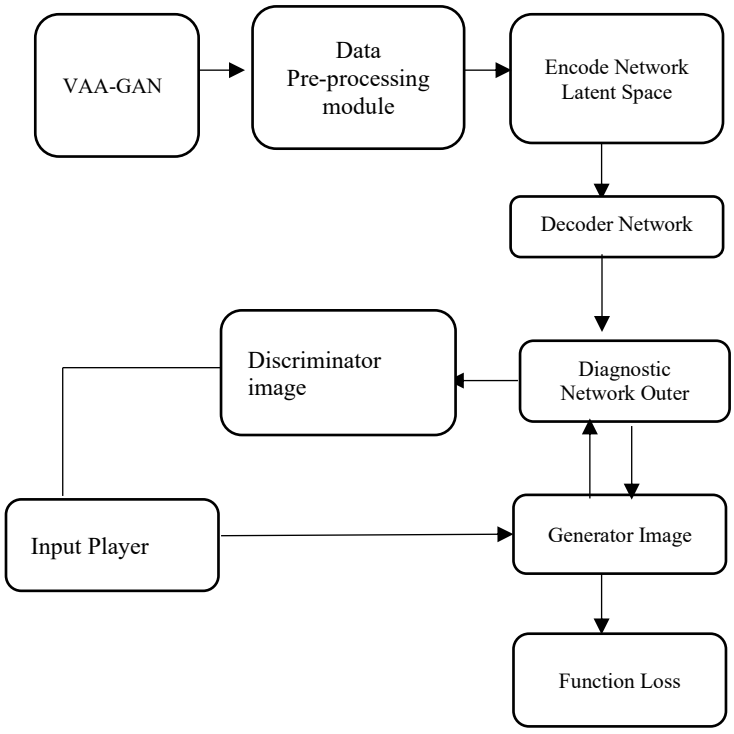

Fig. 1. System Block Diagram (VAEGAN)

The VAEGAN technique is suitable for solving mode collapse due to its hybrid architecture, combining Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs). VAEs capture complex dependencies and generate diverse samples, while GANs produce realistic samples. By leveraging both strengths, VAEGAN generates diverse and realistic samples, mitigating mode collapse. The VAE component regularizes the generator network, preventing single-mode collapse, and adversarial training improves sample quality and diversity. The VAEGAN model has demonstrated exceptional flexibility and expressiveness in modeling complex posterior distributions over latent variables. This suggests that the model is relatively robust to the initial distribution of input data. Based on this intuition, the VAEGAN model appears to be an ideal candidate for mitigating the initial reliance on input data and emphasizing solutions from within the generative model itself. The approximation of a single-dimensional Gaussian N(0,1) is a number of function of points as it is described in Figure 2.

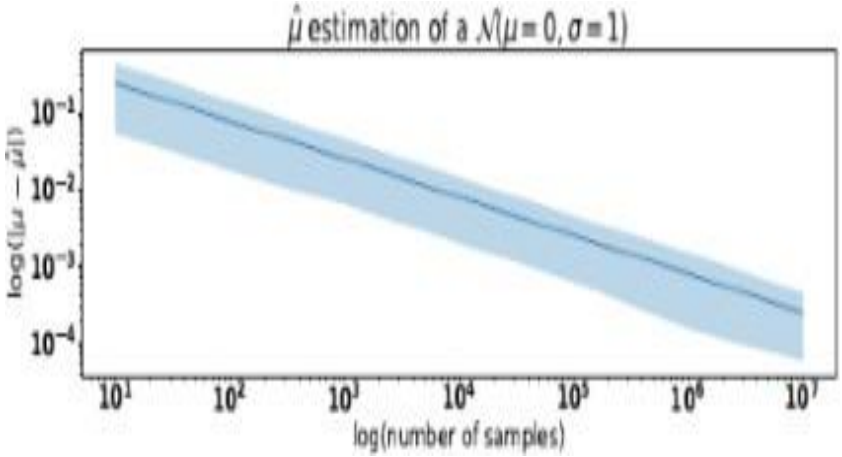

Fig. 2: Approximation of a single-dimensional Gaussian N(0,1) as a function of number of points

A. Key Components of the System

i. Data Preprocessing

A simplified MNIST dataset from TensorFlow was used for the experiments. The dataset consists of 70,000 grayscale images of handwritten digits (0-9), divided into 60,000 training images and 10,000 testing images. The dataset was preprocessed to ensure consistency and quality, including normalization and resizing of images. There are several pre-processing techniques, in this study some of the preprocessing techniques employed are:

*Resizing***:** Scaling images to a consistent size can simplify processing and reduce computational requirements.

Normalization**:** Scaling pixel values to a fixed range (e.g., [0, 1]) improves convergence during training and makes models more robust to changes in lighting conditions. Thirty percent of the dataset was used for testing, while the remaining seventy percent was used for training.

ii. Model Architecture

To enable multiple generations on the same dataset so as to monitor its progress, a customized VAEGAN architecture was created. This architecture combines the advantages of Generative Adversarial Networks (GANs) and Variation Auto encoders (VAEs), allowing for the efficient and reliable generation of new data samples. The VAEGAN model is composed of an encoder, decoder, and discriminator that collaborate to learn the underlying data distribution.

iii. Training

The MNIST dataset was used to train the VAEGAN model, which improved performance and resilience by utilizing methods including data augmentation and transfer learning. During training, the model's performance was assessed using a number of metrics, including as reconstruction loss, KL divergence, and inception score. In order to maximize the quality of the generated samples and minimize the loss function, the model's parameters were optimized during the training phase.

iv. Evaluation

The performance of VAE, GAN, and VAEGAN hybrid models was examined, and the degradation in data distribution over generations was graphically shown. Frechet inception distance (FID), inception score, and visual inspection of generated samples were among the evaluation criteria employed. The outcomes shed light on each model's advantages and disadvantages as well as how well the VAEGAN design mitigates model collapse.

v. *Deployment*

A useful tool for studying model collapse was created by integrating the trained model into an online web application. Users can create new samples, see the data distribution, and interact with the model via the application. Researchers can now more easily and readily investigate and assess model collapse thanks to the model's deployment. A single-dimensional Gaussian distribution N(0,1) is approximated as a function of the number of points in Figure 2.

B. Variational Autoencoder (VAE) Generational Model

A VAE consists of a decoder network that performs the reverse function by converting the latent code over to the reconstruction data and an encoder network that converts input data to a latent code. The VAE gains an optimal latent representation that captures the essential properties of the data through the training procedure. Fig 3 shows the block diagram of VAEs.

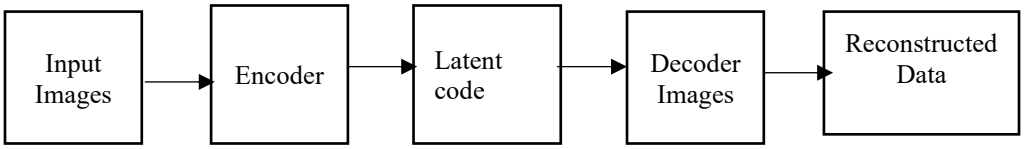

Fig. 3: VAE Block Diagram

The VAE generational model is characterized by the Eq. 1-6:

(i) Joint Probability Distribution

The following represents the combined probability distribution of the latent variable z and the observable variable x:

$$P\theta(x, z) = P\theta x(\frac{x}{z})P\theta z(z) \quad (1)$$

where:

$P\theta x(\frac{x}{z})$ is the parametric conditional distribution of x given z

$P\theta z(z)$is the prior distribution of z, assumed to be a standard normal distribution:

$$P\theta z(z) = N(z; 0L, IL) \quad (2)$$

(ii) Decoder Network

A Deep neural network, also known as D that reflects the proportional conditional distribution $P\theta x(\frac{x}{z})$. is the decoder network, sometimes referred to as the generator network. The DNN produces the conditional distribution's parameters after receiving the latent variable z as input.

(iii) Gaussian Assumption

To keep things simple, the conditional distribution $P\theta x(\frac{x}{z})$ is taken to be a Gaussian

$$P\theta x(\frac{x}{z}) = N(x; \mu\varphi(z), diag(\sigma 2(z))) \quad (3)$$

This assumption is established for consistency between models, while various distributions that may be used

(iv) Encoder Network

A DNN serves as the encoder network, mapping the latent variable z to the observable variable x. The parameters of the variation distribution $\varphi(\frac{z}{x})$, which is thought to be a Gaussian distribution with a diagonal covariance matrix, are output by the encoder network:

$$q\varphi\left(\frac{z}{x}\right) = N(z; \mu\varphi(x), diag(\sigma\varphi^2(x))) \quad (4)$$

(v) Inference and Generative Processes

The VAE model consists of two processes:

(a) Inference: Using the variational method, the encoder network converts the observable variable x to the latent variable z.

The encoder network maps the observed variable x to the latent variable z, using the variational distribution$q\varphi\left(\frac{z}{x}\right)$.

(b) Generative: The decoder network maps the latent variable z to the observed variable x, using the conditional distribution$p\theta\left(\frac{x}{z}\right)$.

(vi) Loss Function

The VAE model is trained using the evidence lower bound (ELBO) loss function:

$$L(\theta, \varphi; x) = -E[q\varphi\left(\frac{z}{x}\right)][\log p\theta\left(\frac{x}{z}\right) + KL(q\varphi\left(\frac{z}{x}\right) || p\theta(z)) \quad (5)$$

W*here:*

☺ = The first term represents the expected log-likelihood of the observed variable x given the latent variable z.

The second term represents the KL divergence between the variational distribution $q\varphi\left(\frac{z}{x}\right)$ and the prior distribution $P\theta(z)$.

C. Generative Adversarial Network (GAN) Model

The components of a Generative Adversarial Network (GAN) are a discriminator that assesses created samples and separates authentic from fraudulent data, and a generator that generates synthetic data samples. When paired with the probabilistic framework of the VAE, the GAN component of VAEGAN helps to avoid model collapse by improving sample quality and variety through adversarially training the generator to produce realistic examples. By combining the advantages of both methods, VAEGAN is able to produce diverse, high-quality samples without being constrained by the disadvantages of each strategy alone. GANs are made up of a discriminator and a generator that learn concurrently and contend against each other. The GAN algorithm's architecture is displayed in Fig. 4.

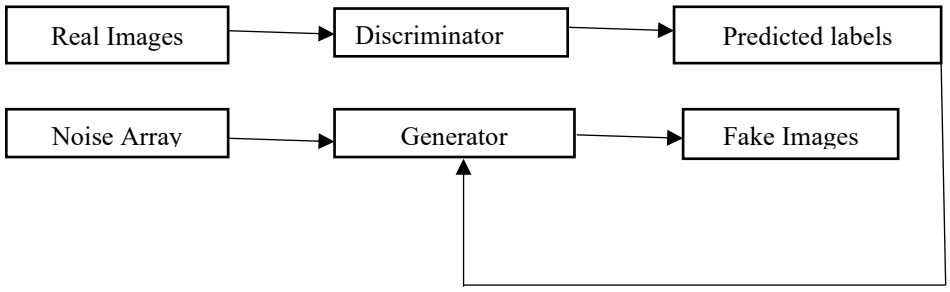

Fig. 4: GANs Block Diagram

$$\min_G \max_D V(D,G) = E_x \sim p_{data(x)[\log D(x)]} + E_z \sim p_{z(z)}\left[\log\left(1 - D(G(z))\right)\right] \quad (6)$$

where:
V(D, G) is the value function
E_x~p_data(x) is the expectation over real data samples
E_z~p_z(z) is the expectation over noise samples
D(x) is the discriminator's output for real data samples
G(z) is the generator's output for noise samples
D(G(z)) is the discriminator's output for generated samples

D. Evaluation Metrics

The performance of the VAE model is evaluated using the following metrics:
Reconstruction Loss: The expected log-likelihood of the observed variable x given the latent variable z.
KL Divergence: The KL divergence between the variational distribution $q\varphi\left(\frac{z}{x}\right)$ and the prior distribution.$p\theta$ (z)
Inception Score: A metric that evaluates the quality of the generated samples.
Fréche Inception Distance (FID): Fréchet Inception Distance (FID) is were implemented to quantify the similarity between the distribution of real images and generated images. This is a core format to measure the quality and diversity of generated images.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (8)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (9)$$

$$F1-\text{score} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \quad (10)$$

E. Model Validation

Model validation is a critical aspect of building machine learning models to ensure they generalize well to unseen data. It involves assessing the performance of a model using techniques beyond simply training it on one dataset and testing it on another.

IV. RESULTS AND DISCUSSION

One of the key features of this system is the model summary function, which plays a crucial role in providing insights into the model's architecture and parameters. Figures 5-7 illustrate the output of the model summary function, which runs and retrieves data on all parameters, including both trained and non-trained ones. These Figures provide a visual representation of the model's structure, highlighting the non-trained parameters and offering a clear understanding of the model's composition.

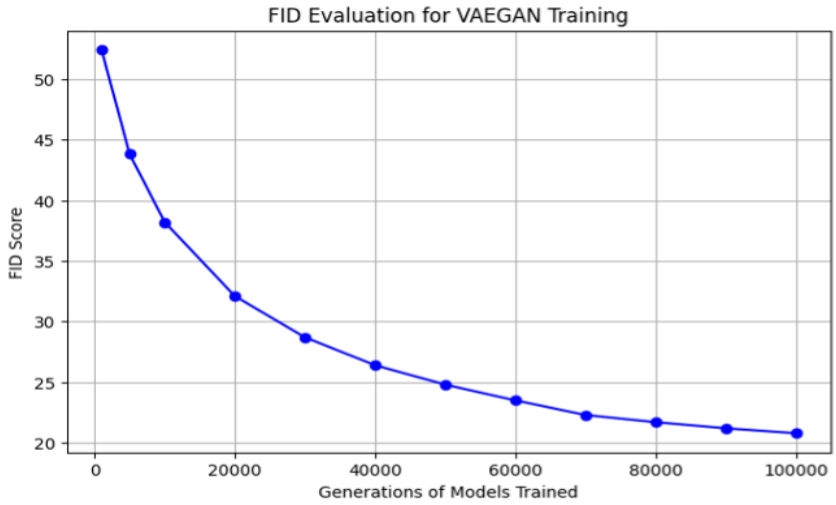

Fig. 5: Code for Downloading and Visualizing the Dataset

Fig. 6. A Single Image from the MNIST Dataset

Number of epochs which refers to one full cycle of training a neural network using the entire dataset. During an epoch, the model processes each example in the dataset once and updates the model's weights based on the error between its predictions and the actual output.

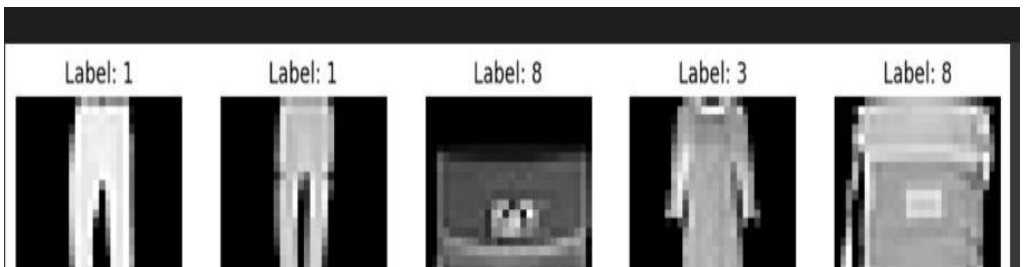

Fig. 7. Graph of how FID Scores Change across Generations

Table 1: The Model with the Performance Metrics Used

Model	FID Score	IS Score	Loss
GAN	50	4.2	0.8
VAE	20	3.5	0.4
VAEGAN	35	4.8	0.6

Table 1 shows the performance of the three measures (FID, IS, and LOSS) and the generative models (GAN, VAE, and VAEGAN) utilized in this study. The FID score, which compares the generated picture to the true image scores, indicates better performance with a lower value. When compared to FID, the VAE had the lowest FID Score (20), showing a high degree of model similarity, even though the GAN and VAEGAN both scored 50 and 35.Greater performance in IS Score, which assesses the variety and caliber of produced images, is indicated by higher scores. In this study, VAEGAN outperformed GAN and VAE with a score of 4.8, whereas GAN and VAE each received scores of 4.2 and 3.5. While GAN and VAEGAN attained loss values of 0.8 and 0.6, respectively, the VAE showed the lowest loss value of 0.4, showing good performance. While the VAEGAN model shows promise in producing high-quality and diverse samples, there are trade-offs. Because of its intricate hybrid design, it takes longer to train even if it produces notable performance improvements in terms of Inception Score (IS) and diversity of generated samples. Furthermore, if the VAE and GAN components are not balanced properly, the model may be unstable during training, which could result in mode collapse. Nevertheless, these drawbacks, VAEGAN is a promising method for generative problem solving because it establishes a solid balance between performance metrics by combining the advantages of VAE and GAN.

V. CONCLUSION

This study emphasizes the issue of model collapse in AI and proposes a hybrid generative model, VAEGAN, as a potential solution. The VAEGAN model's balanced performance, with a high Inception Score (IS) of 4.8 and a moderate Frechet Inception Distance (FID) score of 35, demonstrates its promise for generative modeling tasks. By combining

the strengths of VAE and GAN, VAEGAN generates diverse and high-quality samples, making it a valuable approach for various AI applications. The potential real-world applications of this study are significant, with implications for image and video generation, natural language processing, and robotics. For instance, VAEGAN can be used to generate realistic images for artistic purposes or to augment datasets for training machine learning models. Future work will focus on testing the VAEGAN model on more complex datasets, such as ImageNet, and exploring the incorporation of diffusion models to further improve its performance. Additionally, we plan to investigate the application of hybrid generative models in different AI domains, including text-to-image synthesis and image-to-image translation. By pushing the boundaries of generative modeling, we aim to develop more robust and reliable AI models that can generate high-quality outputs with diverse applications.

REFERENCES

[1] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair and Y. Bengio. Generative Adversarial Networks. Proceedings of the 27th International Conference on Neural Information Processing Systems, MIT Press, 2, 2672-2680, 2014.
[2] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved Techniques for Training GANs. Advances in Neural Information Processing Systems, Curran Associates, Inc., 29, 2234-2242, 2016.
[3] R.S. Sutton and A.G. Barto. Reinforcement Learning: An Introduction. MIT Press, 2nd ed., 2018.
[4] D.P. Kingma and M. Welling. Auto-Encoding Variational Bayes. Proceedings of the 2nd International Conference on Learning Representations, ICLR, 2014.
[5] R. Matulionyte. The Future of AI: Challenges and Opportunities. Journal of Artificial Intelligence Research, AAAI Press, 76, 1-12, 2023.
[6] A. Radford, L. Metz, and S. Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. Proceedings of the 4th International Conference on Learning Representations, ICLR, 2016.
[7] X. Liu and Y. Liu. Anomaly Detection in Complex Systems Using Machine Learning. IEEE Transactions on Industrial Informatics, IEEE, 14(4), 1713-1722, 2018.
[8] Z. Wang and Y. Liu. Image Generation Using Generative Adversarial Networks: A Review. IEEE Transactions on Neural Networks and Learning Systems, IEEE, 31(1), 3-14, 2020.
[9] C. Shorten and T.M. Khoshgoftaar. A Survey on Image Data Augmentation for Deep Learning. Journal of Big Data, Springer, 6(1), 1-27, 2019.
[10] R. Chalapathy and S. Chawla. Deep Learning for Anomaly Detection: A Survey. ACM Computing Surveys, ACM, 51(4), 1-38, 2019.
[11] M. Mirza and S. Osindero. Conditional Generative Adversarial Nets. Proceedings of the 28th International Conference on Neural Information Processing Systems, MIT Press, 2, 2672-2680, 2014.
[12] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein Generative Adversarial Networks. Proceedings of the 34th International Conference on Machine Learning, JMLR, 70, 214-223, 2017.
[13] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved Training of Wasserstein GANs. Advances in Neural Information Processing Systems, Curran Associates, Inc., 30, 5767-5777, 2017.
14] J. Zhao, M. Mathieu, and Y. LeCun. Energy-based Generative Adversarial Networks. Proceedings of the 5th International Conference on Learning Representations, ICLR, 2017.
[20] Y. Li, J. Zhang and X. Chen, Model Collapse in Recommender Systems. Proceedings of the 34th International Conference on Machine Learning, 1-12. 2020.
[21] W. Wang, Z. Gan, H. Xu and Z. Wang, Mitigating Model Collapse in Natural Language Processing. Proceedings of the 34th International Conference on Machine Learning, 1-12. 2020.
[22] Radford A., Metz L. and Chintala S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv preprint arXiv:1511.06434. 2015.
[23] T. Karras, T. Aila, S. Laine and J. Lehtinen, Progressive Growing of GANs for Improved Quality. 2017.
[24] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow and R. Fergus, Intriguing Properties of Neural Networks. arXiv preprint arXiv:1312.6199. 2014.