Mode Collapse Detection Strategies in Generative Adversarial Networks for Credit Card Fraud Detection

Farhat Lamia Barsha
Computer Science
Tennessee Tech University
Cookeville, TN
fbarsha42@tntech.edu

William Eberle
Computer Science
Tennessee Tech University
Cookeville, TN
weberle@tntech.edu

Abstract

A Generative Adversarial Network (GAN) is an artificial intelligence model developed specifically to produce synthetic data that resembles real data by training a generative model and a discriminative model simultaneously using adversarial training. A GAN can be extensively used for generating replicated data, however, it suffers from several issues, one of which is *mode collapse*. Mode collapse takes place when the generator is unable to capture the complete range of diversity in the target data distribution, resulting in the production of limited and repeating variations of samples. Multiple metrics exist to quantify mode collapse in GANs, although no individual metric is capable of consistently providing accurate results. This research focuses on the critical need for accurate mode collapse detection techniques in GANs, to strengthen the credit card fraud detection systems. In this work, we utilize a GAN to generate numerical data instead of image data. Our approach utilizes a wide range of measures, such as Generator and Discriminator Loss, Wasserstein Distance, precision, recall, and visualization tools, to provide a comprehensive framework for detecting mode collapse. In addition, we introduce an alert mechanism that identifies possible mode collapse at an early stage, allowing for earlier intervention and modifications to the training process. We have further proposed suggestions regarding monitoring and analyzing generator and discriminator loss values to identify potential instances of mode collapse to help the developer optimize GAN training and improve the quality of synthetic data.

1 Introduction

A generative adversarial network (GAN) is a machine learning (ML) model in which two neural networks, the generator and the discriminator, compete using deep learning approaches to improve the accuracy of their predictions (Goodfellow et al. 2014). The main purpose of the generator is to produce artificial data that closely resembles actual data samples. It acquires the ability to create a mapping between random noise or a hidden input and data points that are ideally indistinguishable from the actual dataset. The discriminator functions as a binary classifier, differentiating between real and synthetic data samples. It assesses the integrity of a provided input and assigns a probability indicating its likelihood of belonging to the real dataset. The training objective is to identify a Nash equilibrium in which the generator generates data that is indistinguishable from genuine data, and the discriminator is unable to consistently distinguish between real and generated samples.

The objective function is formulated as a minimax problem as follows (Dwivedi 2023):

$$\begin{aligned}\min_G \max_D V(D,G) = & E_{x\sim p_{\text{data}}(x)}[\log D(x)] \\ & + E_{z\sim p_z(z)}[\log(1-D(G(z)))]\end{aligned} \tag{1}$$

Here, $D(x)$ is the discriminator's assessment of the likelihood that the input x originates from the actual data distribution. The output of the generator, denoted as $G(z)$, is obtained by feeding a random noise input z. $p_{\text{data}}(x)$ represents the actual data distribution and $p_{\text{z}}(z)$ is the noise input distribution. The first part of this objective function encourages the discriminator to accurately categorize real data and the second part encourages the generator to generate samples that the discriminator classifies as real.

Since being developed in 2014 by Goodfellow et al., GANs have had significant success in producing duplicated data, but they also come across several challenges. The three primary constraints of GANs include the Vanishing gradient problem, Non-convergence, and Mode collapse (Saxena and Cao 2021). Mode collapse is a phenomenon that occurs in GANs when the generated outputs are limited to a small set of examples, therefore failing to represent the full range of diversity present in the training data distribution (Kossale, Airaj, and Darouichi 2022). Several innovative GAN techniques have been proposed to address the issue of mode collapse by showcasing the stability and resilience of their approach on specific architectures and datasets but there is a lack of research that systematically compares their performance. Mode collapse detection is important, and currently, no one particular metric can measure this effectively. However, by integrating many metrics, it is possible to identify occurrences of mode collapse.

The aim of this work is to reliably detect mode collapse by considering a range of diverse metrics. By detecting mode collapse at an early stage, it is possible to take appropriate steps during the training process of a GAN to enhance the quality of the generated data and eventually enhance its performance. The significance of this study lies in the application of a GAN to numerical data, which is distinct from the existing focus on image data, as conventional metrics designed for image data are not directly applicable due to the different characteristics of numerical data.

This work makes the following contributions:

1. **Diverse measures:** Our methodology integrates a range of criteria, such as Generator and Discriminator Loss, Wasserstein Distance, t-SNE visualization and precision, and recall to comprehensively assess the performance of GANs and identify mode collapse accurately.
2. **An alert system for earlier detection:** We propose an alert system that detects the occurrence of possible mode collapse at an early stage so that training adjustments can be made to stabilize the GAN.
3. **Monitoring and analysis:** We provide a methodology for examining the dynamics of generator and discriminator loss to actively monitor for indications of mode collapse.

We organize the rest of the paper as follows: In Section 2 we present a background of previous work. Then, we discuss mode collapse in detail in Section 3. Section 4 introduces experimental datasets. In Section 5, we discuss the experimental analysis, and in Section 6, we present our results. Finally, we conclude the paper with a final summary and discussion of future work in Section 7 and Acknowledgments in Section 8.

2 Background

Several mode collapse detection techniques have been developed in recent years. In this section, we will provide a thorough evaluation of these approaches.

Zhenyu et al. (Wu et al. 2021) investigates intra-mode collapse in state-of-the-art GANs within a novel black-box setting, without access to training data and trained model parameters. The researchers used faces and vehicles as subjects to measure and analyze the overall collapse inside a specific mode. They used statistical tools to quantify this collapse, examined potential causes, and proposed two novel black-box calibration methods to mitigate the mode collapse. Although the initial findings have been promising, the study has several limitations, such as discrepancies in predicting the description of the identity in generated images.

Acklyn et al. (Murray and Rawat 2021) examine the occurrence of mode collapse in the Intrusion Detection System (IDS) Control Flow GAN (ICF-GAN) model and suggests conditional strategies to mitigate instances of mode collapse. The paper presents a mini-batch approach that greatly improves the accuracy of the model. The ICF-GAN, which utilizes LSTM with mini-batch discrimination ingestion, exhibits enhanced precision in hazard control flow and botnet anomaly detection. The ICF-GAN outperforms current methods, as demonstrated by a comparative analysis and numerical findings.

Sayeri et al. (Lala et al. 2018) examine the latest GAN architectures, specifically AdaGAN, VEEGAN, Wasserstein GAN, and Unrolled GAN on both synthetic and real datasets. The comparison is based on several widely used metrics for measuring mode collapse. Their results indicate that AdaGAN consistently outperforms other GANs on almost all datasets, but Wasserstein GAN exhibits poor performance on these datasets. They also acknowledged that a single metric is inadequate for measuring mode collapse in GANs due to the lack of consistent results from these metrics.

Saad et al. (Saad, Rehmani, and O'Reilly 2022) compared mode collapse measures across datasets and found that AdaGAN outperforms other GANs whereas Wasserstein GAN performs poorly. Additionally, conflicting results imply that a single parameter for evaluating mode collapse in GANs is insufficient. AIIN (Adaptive Instance Normalization Initialization) is recommended for DCGAN to address this issue. DCGAN with AIIN generates more varied X-ray images than DCGAN without AIIN. The improved MS-SSIM and FID scores address mode collapse. In some image properties, AIIN preprocessing beats Gaussian and median filtering. Augmented images, which combine generated X-ray images with real photos, are used to train machine learning classifiers, validating the suggested approach.

Ding et al. (Ding, Jiang, and Zhao 2022) provide a comprehensive literature review on mode collapse. They analyze the resolution for this issue from an architectural and loss functions perspective. From an architectural point of view, the authors discuss multiple types of GANs, including multiple generator GANs, self-attention GANs, and big GANs. From a loss functions perspective, they explore Unrolled GAN and DRAGAN. They examine the future prospects of the GAN by providing a concise overview of the diverse applications of GANs.

Toi et al. (Tsuneda et al. 2021) introduce IRGANs as a solution to mitigate mode collapse, incorporating the concept of intrinsic rewards in reinforcement learning. The utilization of the RND approach in training the generator with simplified algorithms is aimed at stabilizing the learning process and is rooted in intrinsic incentives. In the comparison studies, they employed DCGAN as the default GANs and applied IRGAN with the intrinsic rewards of the proposed technique. Subsequently, they assess the performance of multiple datasets (MNIST and Fashion-MNIST) and find that their proposed approach consistently demonstrates superior performance in every experiment.

Durall et al. (Durall et al. 2020) identify a direct correlation between the eigenvalues of the Generator and the incidence of mode collapse through the presence of large eigenvalues, which may indicate the approach to a distinct minimum, is closely associated with the occurrence of mode collapse. Inspired by this result, the researchers propose a new optimization technique termed nudged-Adam (NuGAN) that utilizes second-order gradient information to deviate from sharp optima, thus avoiding mode collapse. Their findings indicate that examining the generalization aspects of GANs, such as analyzing the uniformity of the optimal solutions discovered during training, is a potential strategy for advancing toward more stable GAN training.

Adiban et al. (Adiban, Siniscalchi, and Salvi 2023) introduces STEP-GAN to identify and mitigate cyber risks. A multi-generator GAN-based model was employed to simulate potential attacks on the system. The model underwent step-by-step training involving the interaction between generators and a discriminator. They outperform state-of-the-art approaches on two severely unbalanced datasets, ICS and UNSW-NB15. The authors test their model on OCAN and MGOCAN, which have not been tested before, and get superior results. They also test their method on seven anomaly detection datasets to prove its universality. In general, STEP-GAN outperforms other semi-supervised methods and some supervised methods.

Jizheng et al. (Jia and Zhao 2019) develop a siamese network that can transform high-dimensional image data into a fixed real number embedding that accurately represents semantic information. Mode collapse is identified by emulating the Wasserstein distance with Euclidean distance. According to empirical evidence, they found that the Siamese Score is 59 times more efficient than the Inception Score which makes it an effective GAN evaluation method.

Adiga et al. (Adiga et al. 2018) introduce two principled measures, namely mode collapse divergence (MCD) and generative quality score (GQS), to accurately measure mode collapse and sample quality. The evaluation measures were used to compare numerous GAN architectures, including vanilla GAN, WGAN, LSGAN, PacGAN, and CGAN, across three datasets, namely MNIST, Fashion MNIST and CIFAR-10.

In contrast to these works, we propose a mode collapse detection strategy that includes diverse metrics, early detection warnings, and monitoring methodologies instead of individual metrics or architectural changes. The alert system allows rapid intervention and training changes to minimize extreme divergence and stabilize GAN performance. We also propose specific generator and discriminator loss dynamics monitoring recommendations to help developers actively track mode collapse warning indications and take preventative measures.

3 Mode Collapse

The primary objective of a GAN is to develop the ability to generate new data that closely resembles the pre-existing real data it was trained on. When replicating data, it is crucial to capture the diversity of the data to avoid the model producing redundant and identical data. Mode collapse is an issue faced by GANs in which the generator produces a limited set of similar or identical samples, instead of generating a diverse and realistic range of outputs (Kossale, Airaj, and Darouichi 2022). There are two possible indications of mode collapse: (1) The output data is missing a substantial fraction of the modes found in the input data. (2) The Generator exclusively acquires knowledge of a limited set of distinct patterns.

Multiple circumstances might lead to the phenomenon of mode collapse. For example: (1) When the discriminator becomes stronger than the generator because of imbalanced learning rates between the generator and the discriminator, the generator faces difficulties in properly using the gradients for learning. As a result, this leads to a limited variety of produced samples (Saad, Rehmani, and O'Reilly 2022). (2) In cases where the input data has a limited range of fluctuations, the generator may have challenges producing a wide range of outputs. (3) When the generator's convergence during training results in a local minimum instead of a global minimum, this may lead to the generation of a limited set of results, thereby inaccurately representing the complete range of data distribution.

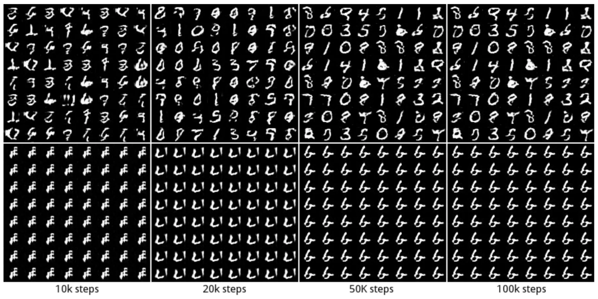

Figure 1: Example of Mode collapse (Metz et al. 2016)

The MNIST dataset contains 10 discrete categories corresponding to the numerical digits ranging from zero to nine. Figure 1 shows the samples generated by two separate GANs. The first row generates all ten modes, but the second row exclusively generates a single mode, precisely the digit "6". Mode collapse is a phenomenon that arises when the creation of data is restricted to a small number of modes.

4 Dataset

In this study, we use both synthetic (VU 2024) and real-world (ULB 2024) datasets from the Kaggle website. The real-world dataset (ULB 2024) consists of credit card transactions of two days performed by European cardholders in September 2013. It includes a total of 284,807 transactions, among which only 0.17% are fraudulent transaction data - which leads to the class imbalance issue. In this dataset, all confidential information is PCA-transformed and the only features that have not been transformed with PCA are 'Time' and 'Amount'. The synthetic dataset (VU 2024) presents Card Not Present Transaction Fraud which consists of 151,112 transactions that accurately replicate fraud patterns observed in the real world. The fraud prevalence rate is 9.36%. The features include sign-up time, purchase time, purchase value, device ID, user ID, browser, and IP address.

5 Experimental Setup

Our work introduces a new strategy where we utilize a combination of multiple metrics to detect mode collapse. For both synthetic and real-world datasets, we employ the same architecture and training technique of the GAN model, making slight adjustments as required for each dataset. This section will provide a comprehensive overview of our experimental setup, encompassing data preparation, the GAN architecture, and the training method.

5.1 Data Preprocessing

For both datasets, we primarily focused on transaction time, amount, and class features as shown in Table 1 as these are the only available features on the real-world dataset. The percentage of fraudulent data in both datasets is significantly lower than that of authentic data, resulting in an imbalanced dataset. To address the problem of class imbalance, we used a downsampling technique. Next, we split the dataset into training and test sets for applying GAN.

Table 1: Feature Categories

Feature Name	**Feature Type**	**Description**
Transaction Time	Numerical	Purchase time (synthetic dataset), number of seconds since first transaction (real dataset)
Transaction amount	Numerical	Transaction amount
Class	Numerical	Indicates whether a transaction is fraudulent or real

5.2 GAN Architecture

The GAN architecture includes a generator and a discriminator, which engage in a competitive training process. The purpose of the generator is to produce synthetic data samples that closely resemble the real data distribution, whereas the discriminator's objective is to distinguish between real and generated samples. The generator consists of three dense layers that utilize Rectified Linear Unit (ReLU) activation functions, batch normalization layers, and a sigmoid activation function in the output layer. The discriminator, however, consists of two dense layers utilizing ReLU activation functions, dropout layers to mitigate overfitting, and a sigmoid activation function in the output layer. The generator uses a latent dimension of 100 as its input, enabling it to acquire a significant and meaningful representation of the data. The GAN model is compiled using the Adam optimizer and binary *crossentropy* loss function. The Adam optimizer effectively updates model parameters and learning rates for optimal convergence and the binary *crossentropy* loss function quantifies the difference between predicted probability and true labels. The GAN trains both the generator and discriminator networks simultaneously to enhance the generator's capacity to generate authentic synthetic data that can subsequently increase the effectiveness of fraud detection models.

5.3 Training Process

The GAN undergoes training for a certain number of epochs, each consisting of a sequence of steps. During each iteration, the generator produces synthetic data by utilizing random noise, intending to replicate the distribution of authentic data. Simultaneously, the discriminator is trained to differentiate between genuine and generated samples. The generator and discriminator are compiled using the Adam optimizer and binary *crossentropy* loss function, which is optimal for the binary classification task performed by the GAN. During the training process, the losses of both the generator and discriminator are measured, and their convergence is observed. In addition, the training method includes an approach to identify mode collapse, which occurs when the generator generates a restricted range of samples. Careful monitoring ensures the generation of a broad and diverse set of synthetic data, which is essential for training fraud detection models and other related activities.

5.4 Alert System

We have implemented a mode collapse detection strategy that is inserted into the training loop. After a specific number of epochs, if there is an *increase* in the generator loss (suggesting a decline in its performance), and at the same time, the discriminator loss *decreases* (indicating that the discriminator is becoming excessively skilled at discriminating samples), it is assumed that a mode collapse has occurred. To resolve this problem, the training process is stopped, and a warning is displayed indicating the possible presence of mode collapse. This cautious interference prevents the GAN from reaching a suboptimal state where the generated data lacks the required variety to accurately represent the data distribution of the original dataset.

5.5 Mode Collapse Evaluation Metrics

For mode collapse detection, we propose a combination of four metrics: Wasserstein distance, t-SNE visualization, Generator and Discriminator loss, and precision and recall.

1. The Wasserstein distance is a metric that measures the dissimilarity or distance between two probability distributions. The function accepts two sets of synthetic data as input, converts them into one-dimensional arrays, and subsequently computes the Wasserstein distance. If mode collapse is present, the Wasserstein distance between distinct batches of synthetic data is expected to be minimal, indicating a lack of diversity in the generated samples.
2. The distribution of combined real and synthetic data is explored in a two-dimensional space using a t-SNE visualization. The real and synthetic data are combined into a unified data frame, and t-SNE is employed to reduce the dimensionality for the purpose of visualization. This representation facilitates the evaluation of the diversity and distribution of synthetic data produced by the GAN.
3. Monitoring the losses of the Generator and Discriminator is essential for detecting mode collapse during the training of GANs. The Generator loss measures the efficacy of the generator in producing synthetic data, whereas the Discriminator loss indicates the discriminator's proficiency in differentiating between real and synthetic samples. Mode collapse, a phenomenon in which the generator narrows its output to a restricted range of samples, can be identified by a simultaneous increase in the generator loss and a drop in the discriminator loss. This indicates that the generator faces difficulty producing diverse data, while the discriminator gets excessively skilled in differentiating the small range of generated samples.
4. Precision and recall play a vital role in detecting mode collapse. Precision evaluates the correctness of positive predictions, reflecting the discriminator's reliability in accurately detecting authentic incidents. On the other hand,

recall measures the model's capacity to correctly identify all positive cases, indicating the extent to which the generator's output is comprehensive. A high precision score means the discriminator can differentiate between real and fake data. A high recall score means the generator generates convincing fake data.

6 Result

We analyzed multiple scenarios on both datasets to illustrate the importance of each metric in detecting mode collapse.

- Scenario 1: Entire real and synthetic dataset
- Scenario 2: Real and synthetic dataset's subset

6.1 Generator and Discriminator loss

In Scenario 1, we applied GANs on both the real and synthetic datasets and graphed the loss of the generator and discriminator. Our goal is that both the generator and discriminator losses reach a point of stabilization. The decline in generator loss with time signifies the improvement of the generator in generating authentic and convincing data that misleads the discriminator. As the generator enhances, the discriminator's responsibility becomes more challenging, resulting in its loss stabilizing by increasing. Rapid fluctuations or exceptionally high values of any losses may suggest issues such as mode collapse. Figure 2 and Figure 3 show continuous and significant fluctuations in the discriminator loss, whereas the generator loss exhibits minimal fluctuation. However, neither of them reaches a stable state, indicating the occurrence of mode collapse. In Figure 2, we can see the fluctuations of discriminator loss, beginning at 935.14 in epoch 0, declining to 874.64 in epoch 30, further decreasing to 489.58 in epoch 60, and then increasing to 791.35 in epoch 100. Conversely, the loss of the generator began at 0.71 in epoch 0 and increased to 0.72 in epoch 20, then reduced to 0.69 in epoch 60 and further to 0.68 in epoch 100. These fluctuations indicate the possible occurrence of mode collapse. In Figure 3, we can observe a similar pattern for both the generator and discriminator loss, indicating the existence of mode collapse.

In Scenario 2, we applied GANs on subsets of both the real and synthetic datasets and observed the loss values. The results are a bit different here than the previous one because of the small amount of data. Figure 4 displays the fluctuation in discriminator loss. It started at a value of 3403.61 during epoch 0, subsequently increased to 4757.286 by epoch 40, and further increased to 5024.69 by epoch 90. Although it did not reach a stable peak, the positive aspect is the indication of a consistent upward trend, which is what we desire. The generator loss exhibits a consistent decline, decreasing from 0.72 in epoch 0 to 0.71 in epoch 40, and further to 0.70 by epoch 90. This constant decrease is a positive signal of improvement in generating high-quality data. In Figure 5, we can see a significant decrease in generator loss and a rise in discriminator loss, indicating a positive trend with no evidence of potential mode collapse.

6.2 t-SNE visualization

In Scenario 1, we worked with t-SNE visualization on both real and synthetic datasets. In both Figure 6 and Figure 7, the blue portion represents generated data, and the pink portion represents real data. The generated data displayed in Figure 6 and Figure 7 exhibit a limited range, lacking the diversity observed in real data. These visualizations demonstrate that the generated data fails to portray the true diversity of the real data, resulting in the occurrence of mode collapse, where similar types of data are repeatedly generated.

In Scenario 2, we utilized t-SNE visualization on a subset of both real and synthetic datasets. In Figure 8 and Figure 9, the generated data is distributed throughout the whole range combined with real data, indicating that it effectively captures the diverse characteristics of real data.

6.3 Wasserstein distance

The Wasserstein distance was also used to detect mode collapse. **In Scenario 1**, the Wasserstein distance between two synthetic data points in the real dataset is 0.00064. The presence of smaller Wasserstein distances indicates a potential shortage in the variety of the generated samples, which gives rise to concerns regarding mode collapse. And the Wasserstein distance between two synthetic data points in the synthetic dataset is 0.00104, indicating a similar situation.

In Scenario 2, the Wasserstein distance between two synthetic data points generated from a subset of the real dataset is 0.00917 - a higher Wasserstein distance indicates an improved variety in the generated samples. The Wasserstein distance between two synthetic data points inside a subset of the synthetic dataset is 0.01393, indicating a similar situation. So, we hypothesize that for Scenario 1, diversity is weaker, whereas the larger distances in Scenario 2 indicate improvements in capturing a wider variety of data patterns.

6.4 Precision-Recall

The evaluation of precision-recall is crucial for detecting mode collapse. **In Scenario 1**, we achieved a precision of 0.69 and a recall of around 0.67. **In Scenario 2**, we obtained a precision of 0.62 and a recall of 0.71. Therefore, while dealing with a small quantity of data, the recall rate is higher, while the precision rate is lower. A drop in precision indicates that the discriminator is encountering difficulties in distinguishing between generated and real data, perhaps due to the improved quality of generated data. On the other hand, a rise in recall signifies that the generator's capability to imitate real data has improved, which is a positive indication. Thus, the likelihood of mode collapse occurring is lower in scenario 2 compared to scenario 1.

It should be noted that we did not compare our work to existing approaches because most approaches use image data, while we use numerical data and different data formats require different evaluation metrics.

6.5 Alert system

In the next phase of our work, we developed an alert system with the purpose of promptly detecting mode collapse. At periods of 10 epochs, the function evaluates the progress

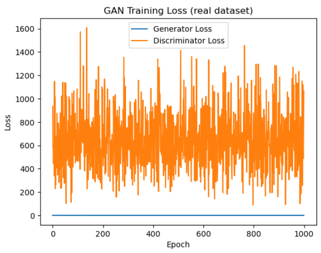

Figure 2: (1)

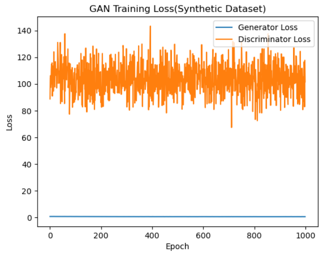

Figure 3: (2)

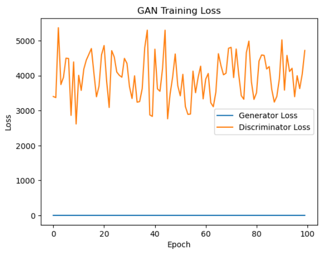

Figure 4: (3)

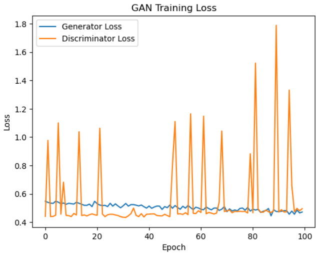

Figure 5: (4)

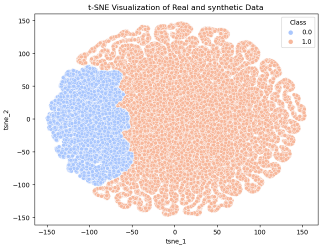

Figure 6: (5)

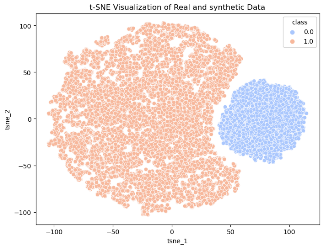

Figure 7: (6)

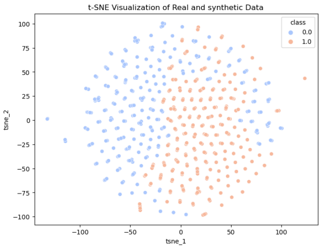

Figure 8: (7)

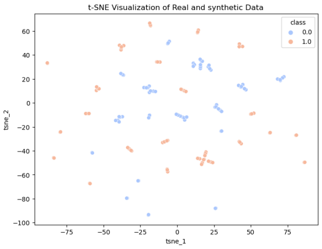

Figure 9: (8)

(1) GAN training loss on real dataset (2) GAN training loss on synthetic dataset (3) GAN training loss on real dataset's subset (4) GAN training loss on synthetic dataset's subset (5) t-SNE visualization of real data (6) t-SNE visualization of synthetic dataset (7) t-SNE visualization of real dataset's subset (8) t-SNE visualization of synthetic dataset's subset

of the training by comparing the losses of the generator and discriminator to determine if it is deviating from the desired path. If the generator's ability to mislead the discriminator declines while the discriminator becomes exceedingly proficient, an alarm is triggered, and the training process is stopped to prevent further loss of time and resources, thereby preventing the GAN from becoming stuck in a cycle of repetitive patterns.

Algorithm 1 Alert System for possible mode collapse

```
Initialize min_g_loss and min_d_loss to large values
for each epoch do
    Train the GAN model
    Calculate generator loss ($g\_loss$) and discriminator loss ($d\_loss$)
    if $epoch > 10$ and $epoch \mod 10 == 0$ then
        if $g\_loss > min\_g\_loss$ and $d\_loss < min\_d\_loss$ then
            Print("Mode collapse detected")
            Break
        else
            $min\_g\_loss \leftarrow g\_loss$
            $min\_d\_loss \leftarrow d\_loss$
        end if
    end if
end for=0
```

7 Conclusion and Future Work

In this work, we propose a comprehensive methodology for evaluating and addressing mode collapse in GANs. A combination of multiple metrics, such as Generator and Discriminator Loss, Wasserstein Distance, t-SNE visualization, and precision-recall metrics, provides a comprehensive evaluation of GAN's performance. The proposed alert system functions as an early warning for possible mode collapse, allowing for immediate actions to stabilize the GAN.

Based on an examination of the generator and discriminator loss across the epochs, we can conclude that there is no fixed threshold for the generator and discriminator loss. This is because the threshold depends on specific characteristics of the GAN architecture and dataset being used. A substantial rise in generator loss and also any extremely low or extremely high discriminator loss could suggest the occurrence of mode collapse. We recommend monitoring the fluctuations in both generator and discriminator losses to identify mode collapse as early as possible. Our future plans involve expanding our evaluation metrics and improving the alarm system through the incorporation of adaptive thresholding approaches and an automated intervention mechanism. In this work, we solely focused on improving credit card fraud detection with GANs. In the future, we plan to apply the proposed approach to other domains and problems. We also plan to apply explainability and interpretability to identify instances of mode collapse.

8 Acknowledgments

The authors wish to thank the Machine Intelligence and Data Science (MInDS) Center and the Computer Science Department at Tennessee Tech for providing resources and funding to work on this project.

References

Adiban, M.; Siniscalchi, S. M.; and Salvi, G. 2023. A step-by-step training method for multi generator gans with application to anomaly detection and cybersecurity. *Neurocomputing* 537:296–308.

Adiga, S.; Attia, M. A.; Chang, W.-T.; and Tandon, R. 2018. On the tradeoff between mode collapse and sample quality in generative adversarial networks. In *2018 IEEE global conference on signal and information processing (GlobalSIP)*, 1184–1188. IEEE.

Ding, Z.; Jiang, S.; and Zhao, J. 2022. Take a close look at mode collapse and vanishing gradient in gan. In *2022 IEEE 2nd International Conference on Electronic Technology, Communication and Information (ICETCI)*, 597–602. IEEE.

Durall, R.; Chatzimichailidis, A.; Labus, P.; and Keuper, J. 2020. Combating mode collapse in gan training: An empirical analysis using hessian eigenvalues. *arXiv preprint arXiv:2012.09673*.

Dwivedi, H. 2023. Understanding gan loss functions. `https://neptune.ai/blog/gan-loss-functions`. (Accessed on 01/18/2024).

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27.

Jia, J., and Zhao, Q. 2019. Siamese score: Detecting mode collapse for gans. In *2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 1–6. IEEE.

Kossale, Y.; Airaj, M.; and Darouichi, A. 2022. Mode collapse in generative adversarial networks: An overview. In *2022 8th International Conference on Optimization and Applications (ICOA)*, 1–6. IEEE.

Lala, S.; Shady, M.; Belyaeva, A.; and Liu, M. 2018. Evaluation of mode collapse in generative adversarial networks. *High Performance Extreme Computing*.

Metz, L.; Poole, B.; Pfau, D.; and Sohl-Dickstein, J. 2016. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163*.

Murray, A., and Rawat, D. B. 2021. On the performance of generative adversarial network by limiting mode collapse for malware detection systems. *Sensors* 22(1):264.

Saad, M. M.; Rehmani, M. H.; and O'Reilly, R. 2022. Addressing the intra-class mode collapse problem using adaptive input image normalization in gan-based x-ray images. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2049–2052. IEEE.

Saxena, D., and Cao, J. 2021. Generative adversarial networks (gans) challenges, solutions, and future directions. *ACM Computing Surveys (CSUR)* 54(3):1–42.

Tsuneda, T.; Kiriyama, T.; Shintani, K.; and Yamane, S. 2021. Gans with suppressed mode collapse using intrinsic rewards. In *2021 Ninth International Symposium on Computing and Networking Workshops (CANDARW)*, 187–192. IEEE.

ULB, M. L. G. 2024. Credit card fraud detection. `https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud`. (Accessed on 01/22/2024).

VU, B. 2024. Fraud ecommerce. `https://www.kaggle.com/datasets/vbinh002/fraud-ecommerce`. (Accessed on 01/22/2024).

Wu, Z.; Wang, Z.; Yuan, Y.; Zhang, J.; Wang, Z.; and Jin, H. 2021. Black-box diagnosis and calibration on gan intra-mode collapse: a pilot study. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 17(3s):1–18.