Balancing Imbalanced Image Datasets Using GAN Variants: A Case Study on Fashion-MNIST

UNIVERSITY OF JORDAN

King Abdullah II School of Information Technology

AUTHOR: Nedal Abushadouf / 0223734

Supervised By: Prof. Yousef Sanjalawi

Table Of Contents

TABLE OF FIGURES

1. Problem Statement

Class imbalance is a common challenge in real-world machine learning applications, where certain classes are significantly underrepresented compared to others. This imbalance often leads to biased classifiers that favor majority classes while performing poorly on minority classes. In image classification tasks, this issue can result in low recall and misclassification of rare but important categories.

Traditional approaches for handling class imbalance include data resampling, cost-sensitive learning, and algorithmic modifications. However, these methods may introduce overfitting or fail to capture the true data distribution of minority classes. Recently, Generative Adversarial Networks (GANs) have emerged as a promising approach for addressing class imbalance by generating synthetic data samples that augment minority classes.

This project aims to explore the effectiveness of GAN-based data augmentation for balancing imbalanced image datasets. Specifically, the performance of a Vanilla GAN is compared with an advanced GAN variant, Wasserstein GAN with Gradient Penalty (WGAN-GP), in improving classification performance on an imbalanced version of the Fashion-MNIST dataset.

2. Description of Dataset & Imbalance Analysis

2.1 Dataset Description

The Fashion-MNIST dataset is a widely used benchmark dataset consisting of 70,000 grayscale images of size 28×28 pixels. The dataset is divided into 60,000 training images and 10,000 test images, distributed evenly across 10 clothing categories, including T-shirts, trousers, dresses, and footwear.

Each image belongs to one of the following classes:

- T-shirt/top
- Trouser
- Pullover
- Dress
- Coat
- Sandal
- Shirt
- Sneaker
- Bag
- Ankle boot

Fashion-MNIST was selected because it provides a more complex and realistic image classification challenge compared to the original MNIST dataset while remaining computationally efficient for experimentation.

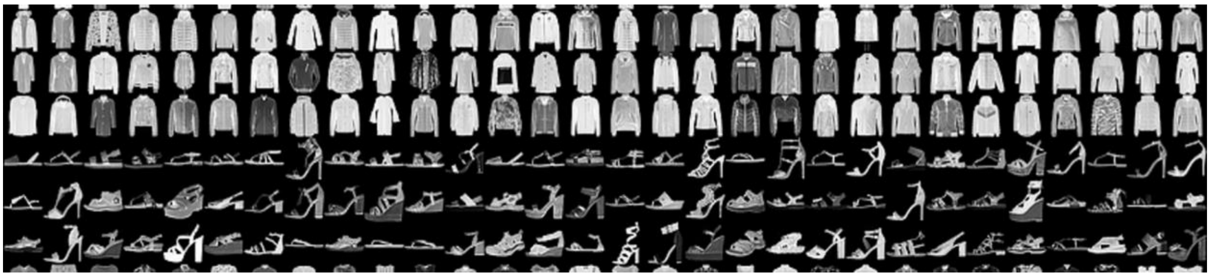

Figure 1: Sample images from the Fashion-MNIST dataset representing different clothing categories.

2.2 Imbalance Creation and Analysis

To simulate a real-world imbalanced classification scenario, the *Sneaker* class (label 7) was selected as the minority class. Approximately 90% of the Sneaker samples were removed from the training dataset, while all other classes were retained in their original quantities. The test dataset remained unchanged to ensure fair evaluation.

The resulting class distribution showed a severe imbalance, with the Sneaker class having significantly fewer samples compared to the remaining classes. This imbalance was visualized using class distribution plots and sample image inspection. The imbalance analysis confirmed the need for augmentation techniques to improve minority class representation.

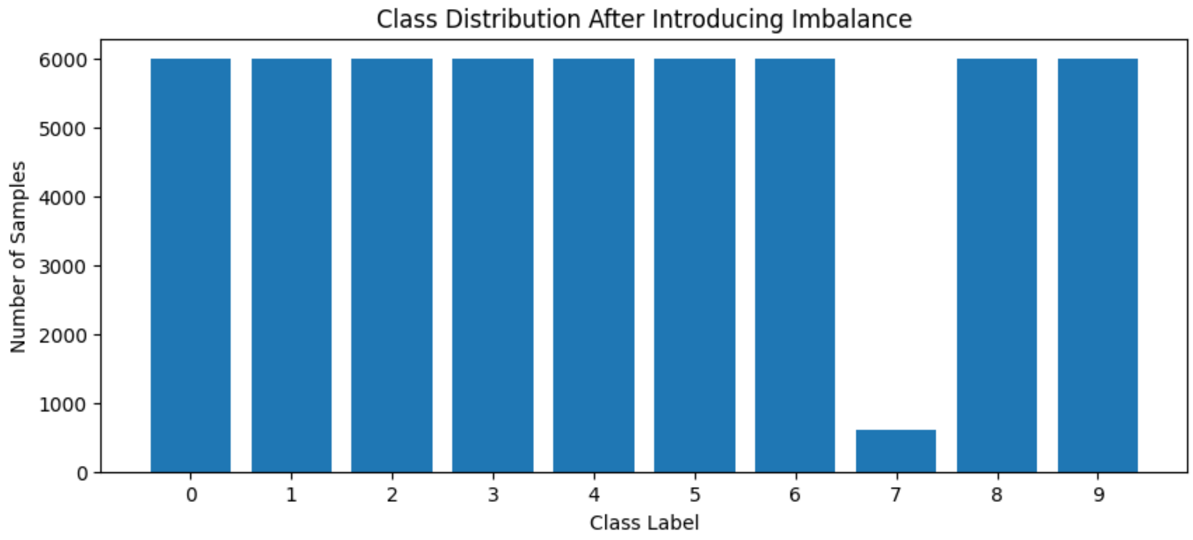

Figure 2: Class distribution of the training dataset after introducing class imbalance.

3. GAN Architectures & Training

3.1 Vanilla GAN

The Vanilla GAN consists of two neural networks trained adversarially:

- **Generator:** Maps random noise vectors from a latent space to synthetic image samples.
- **Discriminator:** Attempts to distinguish between real and generated images.

Both networks were implemented using fully connected layers. The discriminator used a sigmoid activation function and binary cross-entropy loss. The Vanilla GAN was trained exclusively on Sneaker images extracted from the imbalanced training dataset.

Despite successful convergence, the Vanilla GAN exhibited training instability, with the discriminator rapidly outperforming the generator. Generated samples were noisy and lacked clear semantic structure, reflecting known limitations of Vanilla GANs when applied to complex image distributions.

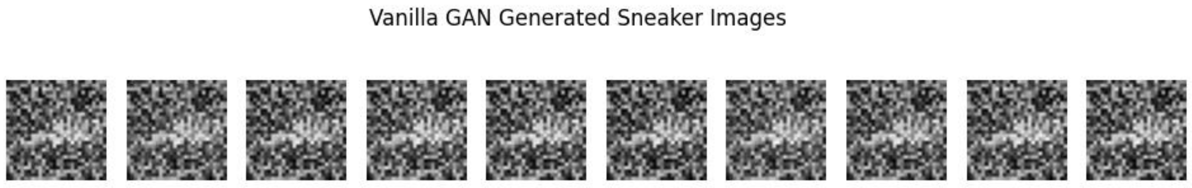

Figure 3: Synthetic Sneaker images generated using the Vanilla GAN after training.

3.2 WGAN-GP (Wasserstein GAN with Gradient Penalty)

To address the instability observed in Vanilla GAN training, a WGAN-GP model was implemented. Unlike Vanilla GANs, WGAN-GP replaces the discriminator with a critic that estimates the Wasserstein distance between real and generated data distributions. Additionally, a gradient penalty term is introduced to enforce the Lipschitz constraint, resulting in more stable training dynamics.

The WGAN-GP generator and critic were trained using the same minority-class Sneaker images. Compared to Vanilla GAN, WGAN-GP produced more realistic and structured sneaker images with significantly improved visual quality and training stability.

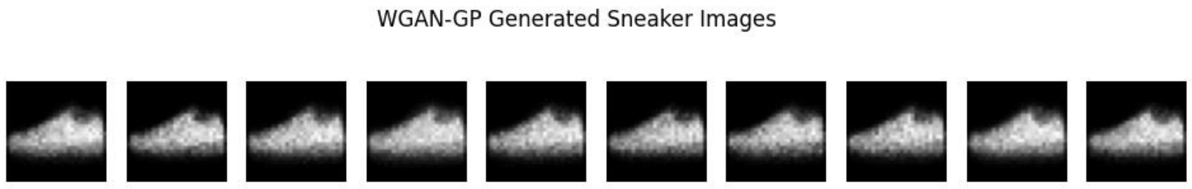

Figure 4: Synthetic Sneaker images generated using the WGAN-GP model, demonstrating improved visual quality.

4. Classifier Setup and Evaluation

4.1 Classification Model

A Convolutional Neural Network (CNN) was used as the classification model due to its effectiveness in image recognition tasks. The CNN architecture consisted of two convolutional layers with ReLU activations and max-pooling, followed by fully connected layers for classification into the 10 Fashion-MNIST classes.

The same CNN architecture, optimizer, learning rate, and training epochs were used across all experiments to ensure fair comparison.

4.2 Training Scenarios

The classifier was trained and evaluated under three different scenarios:

1. **Baseline:** Training on the original imbalanced dataset.
2. **Vanilla GAN Augmentation:** Training on a dataset balanced using synthetic Sneaker images generated by the Vanilla GAN.
3. **WGAN-GP Augmentation:** Training on a dataset balanced using synthetic Sneaker images generated by the WGAN-GP.

In both augmentation scenarios, the number of synthetic Sneaker samples was chosen to match the average class size across all classes.

4.3 Evaluation Metrics

The classifier performance was evaluated using the following metrics:

- Accuracy
- Macro-averaged Precision

- Macro-averaged Recall
- Macro-averaged F1-score
- Area Under the ROC Curve (AUC-ROC)
- Confusion Matrix

Macro-averaged metrics were emphasized to better reflect performance on the minority class.

5. Results & Comparisons

The evaluation results showed that overall accuracy remained relatively stable across all scenarios. This behavior is expected, as majority classes dominate accuracy in multi-class classification tasks. However, macro-averaged metrics and confusion matrices provided more meaningful insights.

Training on the Vanilla GAN-augmented dataset resulted in moderate improvements in minority class recognition. In contrast, the WGAN-GP-augmented dataset produced the most balanced classification performance, with improved recall and F1-score for the Sneaker class and fewer misclassifications observed in the confusion matrix.

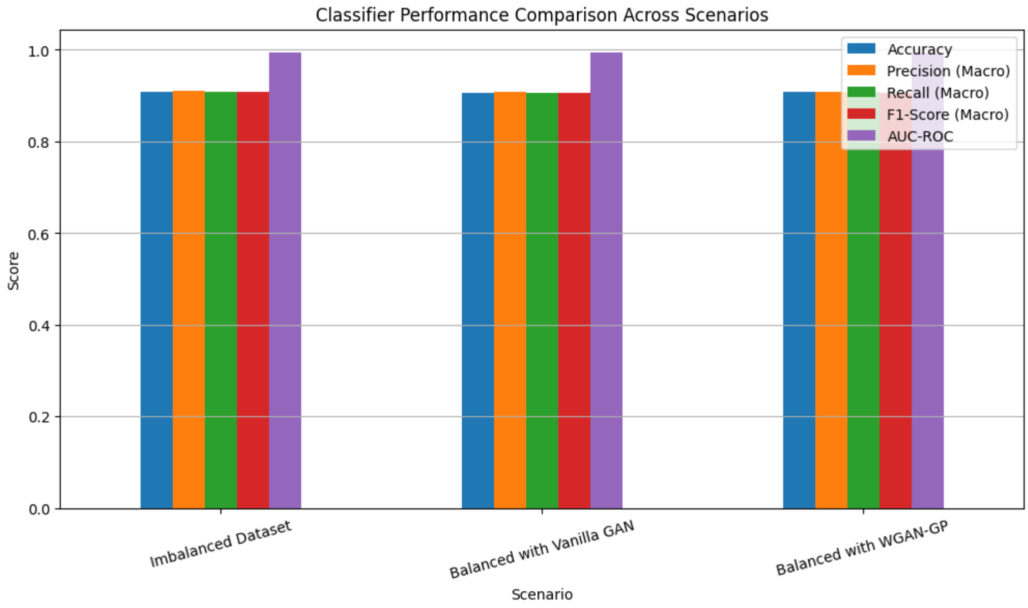

Figure 5: Comparison of classifier performance across different training scenarios using multiple evaluation metrics.

Visual comparison of confusion matrices demonstrated that WGAN-GP significantly reduced confusion between Sneaker and other footwear-related classes compared to both the baseline and Vanilla GAN scenarios.

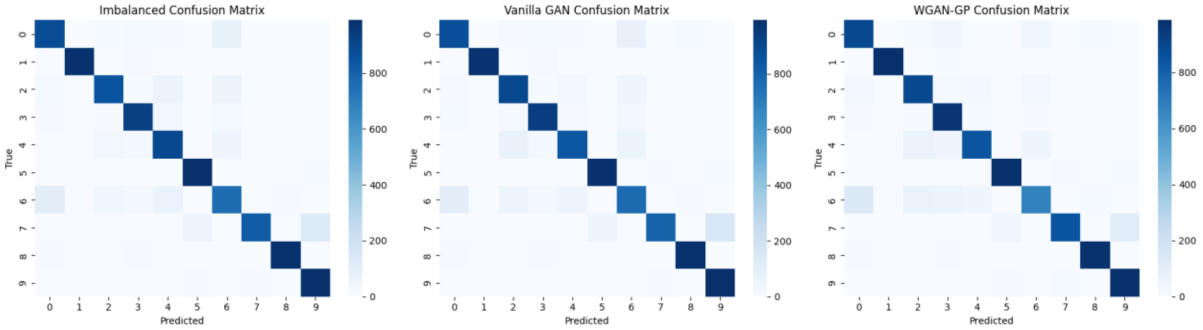

Figure 6: Confusion matrices for (a) imbalanced dataset, (b) Vanilla GAN augmented dataset, and (c) WGAN-GP augmented dataset.

6. Observations and Conclusions

Observations

- Vanilla GANs suffer from training instability and generate lower-quality synthetic samples.

- WGAN-GP provides more stable training and produces higher-quality images.

- GAN-based augmentation improves minority class recognition without modifying the classifier architecture.

- Macro-averaged metrics and confusion matrices are essential for evaluating imbalanced datasets.

Conclusions

This project demonstrated that GAN-based data augmentation is an effective approach for handling class imbalance in image classification tasks. While Vanilla GANs offer limited improvements, advanced GAN variants such as WGAN-GP significantly enhance minority class representation and classification performance.

The results highlight the importance of selecting stable GAN architectures when generating synthetic data for imbalanced learning problems. Future work could explore conditional GANs, larger image resolutions, or class-conditional generation to further improve performance.