

Exploring GAN Variants for Balancing Imbalanced Datasets

1. Problem Statement

Imbalanced datasets are a common challenge in many real-world machine learning applications, especially in classification problems. In such datasets, one class significantly outnumbers the other classes, which often leads to biased models that perform well on the majority class while failing to correctly classify minority class samples.

Traditional classifiers tend to maximize overall accuracy, which can be misleading in imbalanced scenarios, as the minority class is usually the most important one to detect.

This project focuses on addressing the problem of class imbalance using Generative Adversarial Networks (GANs). Instead of relying on traditional oversampling techniques, GANs are used to generate synthetic samples for the minority class. The goal is to study how different GAN architectures can help balance the dataset and improve classification performance, particularly in detecting minority class instances.

2. Dataset Description and Imbalance Analysis

The dataset used in this project is an SMS spam classification dataset, where each message is labeled as either spam or ham. This dataset is

naturally imbalanced, with legitimate (ham) messages forming the majority class and spam messages forming the minority class.

To prepare the data for model training, all SMS messages were converted into numerical representations using the TF-IDF (Term Frequency–Inverse Document Frequency) technique. This approach transforms text into fixed-length numerical vectors, allowing machine learning and deep learning models to process textual data effectively.

An analysis of the dataset revealed a clear imbalance between the two classes. The number of ham messages was significantly higher than spam messages, which confirms that the dataset is suitable for studying the effects of imbalance. The class distribution was visualized using summary tables and bar charts to clearly illustrate the imbalance and justify the need for data augmentation techniques.

3. GAN Architectures and Training

3.1 Vanilla GAN

The first model implemented was a Vanilla GAN, consisting of two neural networks: a generator and a discriminator. The generator takes random noise as input and attempts to generate synthetic TF-IDF feature vectors that resemble real spam messages. The discriminator tries to distinguish between real spam samples and generated ones.

The Vanilla GAN was trained only on the minority class (spam). This design choice ensures that the generator learns the distribution of spam

messages without interference from the majority class. Although Vanilla GANs are relatively simple, they provide a useful baseline for comparison with more advanced GAN variants.

3.2 Wasserstein GAN (WGAN)

As an advanced GAN variant, a Wasserstein GAN (WGAN) was implemented. Unlike the Vanilla GAN, WGAN replaces the discriminator with a critic and uses the Wasserstein distance as a training objective. This modification improves training stability and reduces common GAN issues such as mode collapse.

The WGAN was also trained exclusively on spam samples. The goal was to evaluate whether a more stable GAN architecture could generate higher-quality synthetic samples and lead to better classification performance after data augmentation.

3.3 Conditional GAN (CGAN)

In addition to WGAN, a Conditional GAN (CGAN) was implemented. CGAN introduces class information as an additional input to both the generator and discriminator, allowing the generation process to be guided by class labels. Since this project focuses on generating spam samples only, a fixed class condition was used during training.

Training the CGAN on the minority class allowed for controlled generation of spam-like samples while maintaining consistency with the class label. This approach helps evaluate whether conditional

information improves the usefulness of generated samples for balancing purposes.

4. Data Augmentation and Classification

After training the GAN models, synthetic spam samples were generated and used to balance the original dataset. Three different training scenarios were considered:

1. Training a classifier on the original imbalanced dataset
2. Training a classifier on a dataset balanced using Vanilla GAN generated samples
3. Training a classifier on a dataset balanced using samples generated by the GAN variant

A Multilayer Perceptron (MLP) classifier was selected for this task due to its suitability for handling TF-IDF feature vectors and its simplicity. The same classifier architecture and training configuration were used across all three scenarios to ensure a fair comparison.

By keeping the classifier unchanged and only modifying the training data, the impact of GAN-based data augmentation on classification performance could be directly observed.

5. Evaluation and Results

The performance of the classifier was evaluated using multiple metrics, including Accuracy, Precision, Recall, F1-Score, AUC-ROC, and Confusion Matrix. These metrics provide a comprehensive evaluation, particularly in imbalanced classification problems where accuracy alone is insufficient.

The results showed that the classifier trained on the original imbalanced dataset achieved high accuracy but relatively low recall for the spam class. This confirms the expected bias toward the majority class. After balancing the dataset using synthetic samples generated by the Vanilla GAN, a noticeable improvement in recall and F1-score was observed.

Further improvements were achieved when using samples generated by the GAN variant, particularly in terms of recall and overall balance between precision and recall. Confusion matrices clearly demonstrated a reduction in false negatives for the spam class after applying GAN-based data augmentation.

Visual comparisons using tables and bar charts were used to highlight the performance differences across the three scenarios. Additionally, examples of generated synthetic samples were presented to demonstrate the effectiveness of the GAN models.

6. Observations and Conclusions

This project demonstrates that GAN-based data augmentation is an effective approach for addressing class imbalance in text classification tasks. Training GANs exclusively on the minority class allowed the generation of meaningful synthetic samples that improved classifier performance.

Among the implemented models, the advanced GAN variants showed more stable training behavior and produced better results compared to the Vanilla GAN. The improvements were most noticeable in recall and F1-score, which are critical metrics in imbalanced classification problems.

Overall, this study highlights the importance of using appropriate evaluation metrics and advanced data augmentation techniques when dealing with imbalanced datasets. Future work could explore more complex text representations or transformer-based classifiers to further enhance performance.