Image Generation Analysis for Imbalanced Datasets

Ujjwal Asthana [1], Shruti Wagle [2], Shivani Singh [3]

New York University [1 2 3]
una207@nyu.edu [1], sgw6735@nyu.edu [2], ss16866@nyu.edu [3]

Abstract

Class imbalance is a common challenge in various domains, including image classification tasks. This project aims to analyze different image generation techniques as a means to address the class imbalance and improve the performance of image classifiers. We investigate the impact of different image generation techniques, such as Non-synthetic generation, General Adversarial Network (GAN), and Stable Diffusion, to mitigate the class imbalance issue in datasets. We use these techniques to balance the dataset and evaluate its performance on a ResNet-18 image classifier.

Upon evaluating our own customized version of the CIFAR-10 dataset, we observe that the test accuracies of non-synthetic generation, GAN, and Stable Diffusion are 85.40%, 85.55%, and 86.96%, respectively. Through this investigation, we shed light on the potential of generative models such as GAN and Stable Diffusion in mitigating class imbalance challenges within datasets. We gain insights into the effectiveness of each image generation method and assess their impact on improving classification accuracy.

Introduction

During the training of neural networks, input data may be equally or even more crucial than the network architecture. Imbalanced datasets, characterized by the unequal representation of classification categories, pose challenges in real-world machine learning applications. However, even with sufficient budget and resources, labeled data can be challenging to obtain. This has led to a rise in the occurrence of imbalanced datasets.

Non-synthetic methods, such as over-sampling or data augmentation, have been used to balance datasets but suffer from biases. Meanwhile, recent advancements in generative models, particularly Generative Adversarial Networks (GANs), have shown promise in generating high-quality synthetic images. However, GANs have limitations in capturing diversity and can be challenging to train and scale. GAN training is challenging due to its instability, which leads to difficulties in achieving convergence and generating consistent high-quality results.

Alternatively, Stable Diffusion, a class of likelihood-based models, has emerged as a viable option, producing high-quality images with beneficial properties such as distribution coverage and scalability. But unfortunately, generating images from stable diffusion models takes exponentially more time than GANs. The time required for image generation in stable diffusion models is primarily due to their sequential nature. Images are generated by refining a latent noise vector through diffusion. This iterative process results in more time-consuming computations than GANs, which generate images in a single forward pass through a generator network.

In this project, we comprehensively analyze various methods for balancing imbalanced datasets. We compare and evaluate the effectiveness of three distinct image generation approaches: Non-Synthetic Generation, Generative Adversarial Networks (GAN), and Stable Diffusion. To evaluate the performance of these methods, we train a ResNet-18 image classifier on customized CIFAR-10 datasets generated from the said methods and analyze their test accuracy.

In the course of our study, we generated images using non-synthetic techniques, trained our own GAN from scratch and used it to generate images, and lastly, utilized an open-source Stable Diffusion model to synthesize new images. The results from all these techniques were analyzed and compared to those received from training on an untouched (vanilla) CIFAR-10 dataset and an imbalanced CIFAR-10 dataset.

Literature Survey

In the context of addressing dataset imbalance, several traditional techniques have been employed. These techniques include:

1. **Dataset re-sampling:** This approach involves creating a balanced dataset by re-sampling the existing data. (Japkowicz and Stephen 2002) demonstrated the effectiveness of this method in mitigating class imbalance by either oversampling the minority class or undersampling the majority class. However, undersampling the majority class may discard useful instances, while oversampling the minority class can introduce duplicate or synthetic examples that may lead to overfitting. Additionally, resampling techniques may not capture the true underlying distribution of the data, potentially affecting the model's generalization.
2. **Data augmentation:** Another approach is to augment the existing data by artificially increasing the number of samples in the underrepresented class. (Simard,

Steinkraus, and Platt 2003) proposed this technique, which involves introducing variations to the existing data through methods such as rotation, translation, or flipping, thereby providing additional training examples for the minority class. However, the effectiveness of augmentation heavily depends on the specific domain and task. Data augmentation may sometimes introduce unrealistic variations or distortions that could negatively impact the model's ability to generalize to real-world scenarios.

3. **Weighted cost:** (Thai-Nghe, Gantner, and Schmidt-Thieme 2010) proposed using a weighted cost approach, assigning a higher weight to the underrepresented class during the learning process. This technique emphasizes the importance of learning from the minority class, ensuring it contributes significantly to the overall model training. Determining the appropriate weight can be subjective and requires careful tuning. Moreover, assigning higher weights to the underrepresented class may result in the model focusing excessively on that class, potentially leading to biased predictions or reduced performance of the majority class.

In 2014, the introduction of Generative Adversarial Networks (GANs) by (Goodfellow et al. 2020) revolutionized image synthesis by producing highly realistic synthetic images. GANs have also been applied for data augmentation in various domains, as demonstrated by (Sandfort et al. 2019). However, GAN training can be unstable, leading to suboptimal results. The instability of GAN training makes it difficult to achieve convergence and produce high-quality results consistently. GANs are notoriously sensitive to hyperparameters and can suffer from mode collapse, where the generator fails to capture the entire training data distribution. GAN training can be computationally expensive and time-consuming due to the adversarial nature of the learning process.

Recent advancements in diffusion models, such as the work of (Sohl-Dickstein et al. 2015), have shown superior image generation capabilities. Stable diffusion demonstrates advantages in terms of sample quality, noise control, integration of text embeddings, and availability as an open-source implementation, making it a promising approach for generating synthetic data in various domains.

Methodology

To ensure an objective evaluation of our synthetic method, we utilize an existing dataset and generate an artificially imbalanced dataset from it. For this purpose, we select CIFAR-10, a well-known dataset introduced by (Krizhevsky, Hinton et al. 2009). CIFAR-10 is preferred for our experimentation due to its relatively compact size. It enables us to efficiently iterate through various methods and generate synthetic data that adheres to the desired class imbalance for the classification task. By leveraging this dataset, we can effectively compare and assess the performance of different balancing techniques and evaluate the efficacy of our synthetic approach in addressing dataset imbalance.

The CIFAR-10 dataset comprises 60,000 labeled color images with a resolution of 32x32 pixels, divided into 10 distinct classes, each containing 6,000 images. The dataset consists of 50,000 training images and 10,000 test images. In our study, we construct three specific datasets: a full set, an imbalanced one, and a synthetic one. The original CIFAR-10 training set is the full set, containing images from all classes. To create the imbalanced set, we specifically target the "cat" class and randomly remove 99% of the images belonging to this class from the training set. Subsequently, we employ classical data augmentation techniques, a diffusion model, and a GAN to generate synthetic images representing the imbalanced class. The test set remains untouched, allowing us to evaluate the performance of the models on real-world, unbiased data.

Data Synthesis Techniques

We experiment with three techniques of image generation:

1. Non-synthetic data
2. Synthetic data: GAN
3. Synthetic data: Stable Diffusion

Non-synthetic data Since we have already removed 99% of the images from the "cat" category to create imbalance, we use the remaining 1% images to non-synthetically generate new images and add them back to the dataset to address the newly created imbalance. We randomly apply different data augmentation techniques to these images and keep appending them back to the dataset until it becomes balanced again.

These data augmentation techniques help the model become more impervious to variations in the viewpoint of the data and help the model improve its generalization capability. The transformations include:

- **Random Perspective Shift**: This involves creating a geometric alteration of an image by changing an image's viewpoint and applying a random change to its four corners. It was applied with a distortion scale 0.2 and constant padding of 0.
- **Random Rotation**: It is also a geometric transformation of an image that involves rotating the original image by an arbitrary angle. The scale was set between 0 to 180 degrees with constant padding of 0.
- **Color Jitter**: This technique randomly alters various aspects of the image's color, such as brightness, contrast, saturation, hue, etc. The brightness factor was set to 0.5 and the hue factor to 0.3.
- **Random Solarization**: It involves applying a random threshold to the pixel intensities of an image, causing an inversion or distortion effect. The technique was applied with a threshold of 0.75 and an apply probability of 0.5.

Synthetic data: GAN A Generative Adversarial Network (GAN) is a type of machine learning model used for generative modeling. It consists of two main components: a generator and a discriminator. These components are trained simultaneously in a competitive manner.

Here, the GAN is being employed to generate images of cats. The training process starts with a dataset of 5000 real cat images from the CIFAR-10 dataset. The generator, which

has 3,576,704 parameters, creates synthetic cat image samples that resemble the real data. Its objective is to generate images that can deceive the discriminator.

The generator comprised of total 5 layers. The size of the input noise vector fed into the generator (nz) was set to 100. The first four layers comprise a transposed convolution operation, a 2D batch normalization, and a ReLU activation function. The final layer has a transposed convolution operation with a Tanh activation function. The output is a 3 x 64 x 64 size image which is the number of channels (nc) x image width x image height. Batch Normalization stabilizes learning by normalizing the input to each unit to have zero mean and unit variance. This helps deal with training problems that arise due to poor initialization and helps the gradients flow in deeper models.

The discriminator, on the other hand, is a binary classifier with 2,765,568 parameters. It attempts to distinguish between real cat images from the dataset and the synthetic images produced by the generator. The discriminator aims to become increasingly skilled at correctly classifying the images.

The discriminator is also made of 5 layers. The input size was 3 x 64 x 64, which is the number of channels (nc) x image width x image height. The first layer comprises a convolutional operation followed by a LeakyReLU activation function. The next three layers comprise a convolutional operation followed by batch normalization and a LeakyReLU activation function. The final layer comprises a convolutional operation followed by a sigmoid activation function. The output is a binary value of size 1 indicating whether the input is real or fake. Leaky ReLUs is used because they help the gradients flow easier through the architecture.

The number of feature maps in the generator (ngf) and discriminator (ndf) was set to 64. The hyper-parameters for the training process included a batch size of 128 and 500 epochs. The Adam optimizer was selected with a learning rate of 0.00005 for the generator and 0.0002 for the discriminator, with beta1 as 0.5 for both. The entire training process is performed on a single GPU.

The weights of the generator and discriminator were initialized using a custom initialization function that initialized the weights of the convolutional layers with a normal distribution. The loss function used was Binary Cross Entropy. The labels used during training were smoothed for better training performance, with the real label set to 0.9 and the fake label set to 0.1.

During training, the generator and discriminator engage in a competitive mini-max game in which the generator aims to minimize the discriminator's ability to correctly classify the generated samples. In contrast, the discriminator aims to maximize its ability to correctly classify between real and generated samples. As training progresses, the generator learns to generate higher-quality and more realistic cat images, while the discriminator becomes more effective at differentiating between real and synthetic images. This minimax game is continued until a Nash equilibrium is reached.

The Fréchet Inception Distance (FID) is utilized to evaluate the performance of the GAN. FID is a metric that measures the similarity between two sets of images, typically the

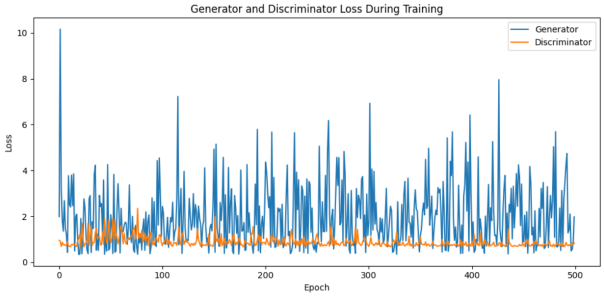

Figure 1: Generator and Discriminator Train Loss

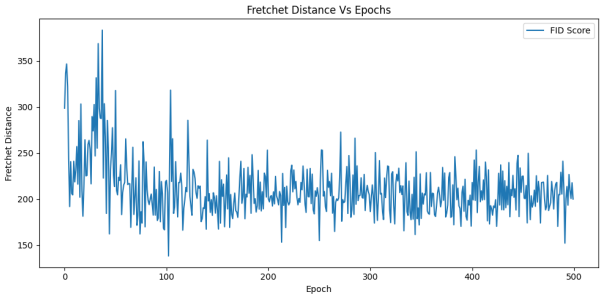

Figure 2: FID vs Epoch for GAN Training

real and generated images. It combines features extracted from a pre-trained Inception-v3 network with the Frechet distance, which is a statistical measure of similarity between probability distributions. The pre-trained Inception-v3 network is used as a feature extractor.

Both the real and generated images are passed through the network, and the activations from one of the intermediate layers are extracted. The mean and covariance of the extracted feature representations are computed separately for the real and generated images. The FID score is calculated as the Fréchet distance between the two multivariate Gaussian distributions defined by the mean and covariance computed in the previous step. The Fréchet distance measures the similarity between two distributions, considering their location and spread.

The quality and resemblance of the generated samples can be assessed by calculating the FID score between the real and generated images. Lower FID scores indicate higher similarity between the real and generated image distributions, suggesting the better performance of the GAN. The model with the lowest FID score was saved for generating images.

Synthetic data: Stable Diffusion Stable Diffusion is an open-source implementation of a latent diffusion model that combines various components to generate high-quality images. The architecture incorporates a variational autoencoder (VAE), a latent-space scheduler, a conditional U-Net based on the work of (Ronneberger, Fischer, and Brox 2015), and utilizes CLIP text embedding introduced by (Radford et al. 2021).

During inference, a random vector in the latent space un-

dergoes denoising through multiple iterations within the U-Net architecture. The noise is predicted and subtracted at each step from the vector, gradually refining the image representation. Simultaneously, text inputs are embedded and serve as pre-conditioning for the U-Net, enabling the denoising process to be influenced by the provided prompt. Finally, the image is obtained by passing the resulting latent vector through the decoder of the VAE.

By leveraging Stable Diffusion's architecture and techniques, we can generate synthetic images with improved quality and coherence. Combining denoising, conditioning, and the VAE framework generates visually appealing and contextually relevant images. This enables us to utilize Stable Diffusion as a powerful tool for data synthesis in our research project on image generation analysis for imbalanced datasets.

We incorporated the Stable Diffusion model into our implementation, utilizing pre-trained weights to ensure efficient and effective image generation. While it is often possible to engineer a specific prompt to generate a few impressive images for demonstration purposes, our goal is to generate a large amount of diverse and high-quality data for the CIFAR-10 dataset, which requires a more systematic approach.

To achieve this, we first initialize the diffusion model pipeline using a pre-trained Hugging Face Model Hub model. Each image is generated based on a randomly generated textual prompt, formed by selecting random words from predefined lists of cat breeds, furniture, prepositions, and actions. The stable diffusion model generates the images based on the textual prompts. If the generated image contains detected NSFW (Not Safe for Work) content, the prompt generation and image generation process is repeated until a suitable image is generated. This process is repeated till the "cat" category becomes balanced again.

This approach allowed us to efficiently sample diverse and realistic images, enabling us to explore the potential of Stable Diffusion for synthesizing large-scale datasets. By leveraging this strategy, we obtained a robust and extensive dataset that we can further analyze and evaluate in our study on image generation analysis for imbalanced datasets.

Classifier Network

To assess the effectiveness of the different balancing methods, we train a ResNet-18 classifier on five datasets, i.e., vanilla CIFAR - 10 dataset, imbalanced dataset (only 1% cat images present), datasets balanced by images produced from the non-synthetic method, GAN, and Stable Diffusion. We analyze the accuracy obtained on the test dataset of CIFAR-10 to examine the efficacy of the balancing techniques and understand their impact on the overall classification performance.

The ResNet-18 architecture consists of multiple levels, each containing a specific number of residual blocks. The model includes four residual block layers and a final output layer. The first level consists of a single convolutional layer with 64 output channels, followed by a max pooling layer. The second level comprises two residual blocks with 64 output channels and another max pooling layer. The third and fourth levels include two residual blocks each, with 128 or 256 output channels. The final level has two residual blocks with 512 output channels. After an average pooling operation with a kernel size of 4, the output is flattened, and the final predictions are obtained through a fully connected layer with ten categories corresponding to the classes in the CIFAR-10 dataset.

Results

We conducted a comprehensive analysis to evaluate the performance of image generation methods on imbalanced datasets. Specifically, we compared the test accuracies achieved by three ResNet-18 classifiers trained on datasets rebalanced using non-synthetic methods, GAN-generated images, and Stable Diffusion-generated images.

The ResNet-18 classifier trained on the imbalanced dataset showed a test accuracy of 0.1% on the 'cat' category and an overall test accuracy of 85.27%.

The ResNet-18 classifier trained on the dataset rebalanced using non-synthetic methods achieved a test accuracy of 2.61% on the "cat" data and an overall test accuracy of 85.4%. This method involved performing random data augmentations on existing samples. While this approach showed improvement compared to the original imbalanced dataset, its performance was limited.

In contrast, the ResNet-18 classifier trained on the rebalanced dataset using GAN-generated images demonstrated a significant boost in test accuracy, achieving 5.9% on the "cat" data and overall test accuracy of 85.55%. The Stable Diffusion model performed even better, achieving 21.1% test accuracy on the "cat" data and an overall test accuracy of 86.96%. The GAN and Stable Diffusion model generated synthetic samples representing the underrepresented class, addressing the class imbalance issue more effectively.

Data Generation Technique	"Cat" Class Test Accuracy	Overall Test Accuracy	Overall Test Loss
Vanilla CIFAR-10 Dataset	85.2%	92.49%	0.282
Imbalanced Dataset	0.1%	85.27%	0.665
Non-synthetic Generation	2.61%	85.4%	0.831
Synthetic Generation: GAN	5.9%	85.55%	0.747
Synthetic Generation: Stable Diffusion	**21.1%**	**86.96%**	**0.680**

These results indicate the superiority of synthetic data generation methods, particularly Stable Diffusion, in mitigating the challenges of imbalanced datasets. Training with the synthetic dataset yields a more notable improvement, as it offers a diverse and informative learning signal for the network to capture various distinctive features related to the cat class. By generating synthetic samples that closely resemble the original data distribution, both GAN and Stable Diffusion methods significantly enhance the classifier's ability to

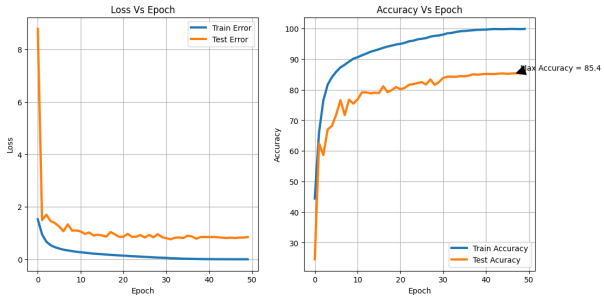

Figure 3: Test Loss and Test Accuracy for Non-Synthetic Dataset

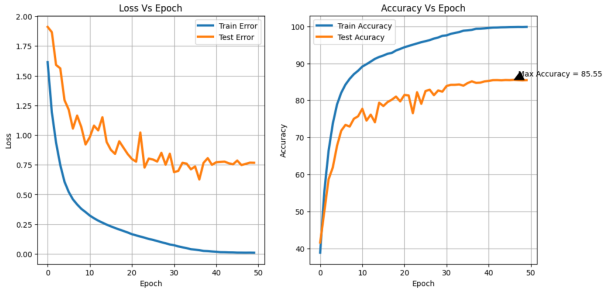

Figure 4: Test Loss and Test Accuracy for Synthetic GAN Dataset

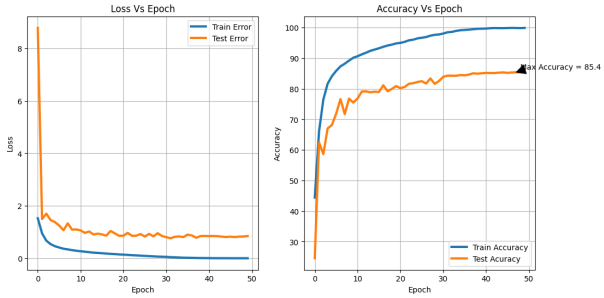

Figure 5: Test Loss and Test Accuracy for Synthetic Stable Diffusion Dataset

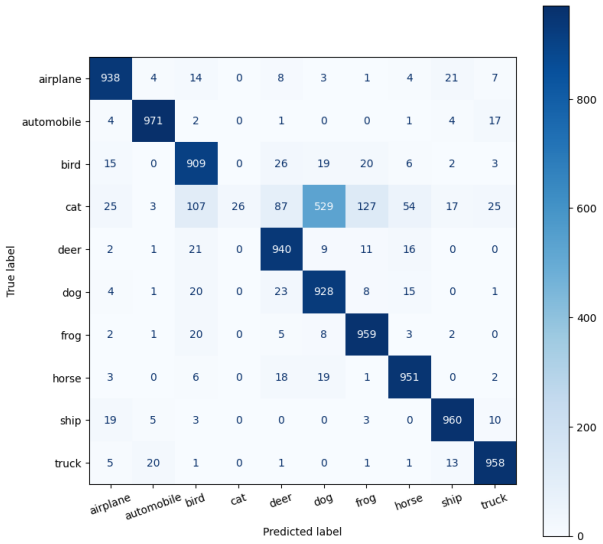

Figure 6: Confusion Matrix for Non-Synthetic Dataset

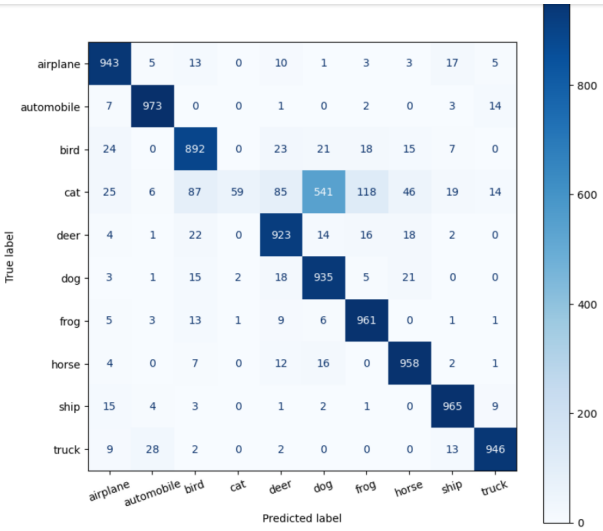

Figure 7: Confusion Matrix for Synthetic GAN Dataset

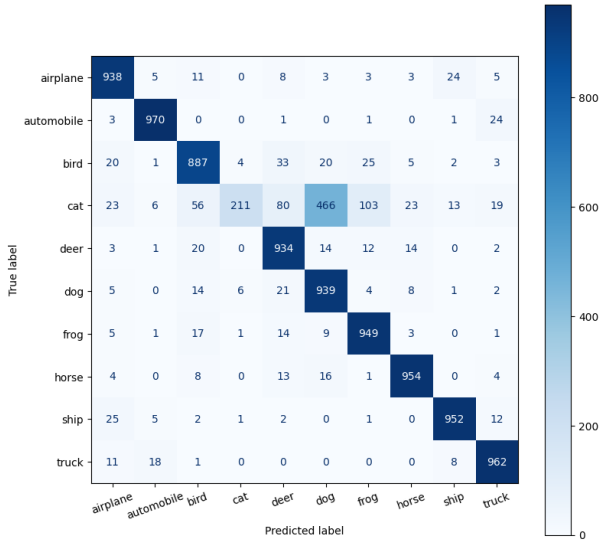

Figure 8: Confusion Matrix for Synthetic Stable Diffusion Dataset

generalize and achieve higher accuracy on the imbalanced dataset.

Overall, our findings highlight the effectiveness of generative models, specifically GAN and Stable Diffusion, in addressing the class imbalance problem and improving the performance of classifiers on imbalanced datasets.

But as evident from the results achieved on the vanilla CIFAR-10 dataset, there is still a long way to go when it comes to alleviating class imbalance in image classification.

Conclusion

This project delves into the exploration of generative models, including GAN and Stable Diffusion, to generate synthetic datasets for mitigating the challenges posed by class imbalance. Our investigation focuses on the CIFAR-10 dataset, wherein we intentionally create an imbalanced set by removing 99% of the cat class. By comparing the performance of synthetic methods with traditional augmentation techniques, we observe a significant performance advantage for the synthetic approaches compared to the non-synthetic process.

The ResNet-18 classifier, trained on datasets rebalanced using non-synthetic methods, GAN-generated images, and Stable Diffusion-generated images, yields test accuracies of 2.61%, 5.9%, and 21.1%, respectively, on the "cat" dataset. The test loss was also reduced for the synthetically generated dataset giving a loss of 0.81, 0.747, and 0.680 overall for non-synthetic methods, GAN-generated images, and Stable Diffusion-generated images, respectively. While the accuracy for the vanilla dataset is still considerably higher, with an accuracy of 85.2% on the "cat" dataset, our model proves that synthetic data generation techniques are greatly preferred when data generation is required, especially in imbalanced datasets.

These outcomes highlight the efficacy of synthetic data in addressing class imbalance challenges and improving the overall accuracy of the trained models. Furthermore, our results demonstrate that Stable Diffusion, a class of likelihood-based models, exhibits superior sample quality compared to GANs. This finding underscores the advantages of Stable Diffusion in generating high-quality synthetic images that closely resemble real data.

References

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Japkowicz, N.; and Stephen, S. 2002. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5): 429–449.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241. Springer.

Sandfort, V.; Yan, K.; Pickhardt, P. J.; and Summers, R. M. 2019. Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks. *Scientific reports*, 9(1): 16884.

Simard, P.; Steinkraus, D.; and Platt, J. 2003. Best practices for convolutional neural networks applied to visual document analysis. 958–962.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, 2256–2265. PMLR.

Thai-Nghe, N.; Gantner, Z.; and Schmidt-Thieme, L. 2010. Cost-sensitive learning methods for imbalanced data. In *The 2010 International joint conference on neural networks (IJCNN)*, 1–8. IEEE.

Github Repository Link

https://github.com/unasthana/ImageGenAnalysis