Exploring class imbalance using Generative Adversarial Networks in medical imaging

CSC2516: Neural Networks and Deep Learning

Vasudev Sharma
University of Toronto
vasu@cs.toronto.edu

Jose Gabriel Islas Montero
University of Toronto
gabriel@cs.toronto.edu

Mohammad Haddadnia
University of Toronto
m.nia@mail.utoronto.ca

Abstract

Imbalanced datasets pose a common and frequent challenge to classification tasks for neural networks, particularly in medical imaging. One of the most promising approaches to solve this problem is using Generative Adversarial Networks (GAN) to generate synthetic data capable of diminishing the imbalance of datasets. Although in some cases GANs suffer from overfitting on the discriminator side, approaches such as Balancing GAN (BAGAN) deal with this and other related issues. Unlike natural images, training GANs with imbalanced data has not been extensively explored in medical imaging, questioning the importance of prior works. To that extent, we extend Improved BAGAN in the medical domain using two public chest X-ray datasets. **GitHub Code:** `https://github.com/vasudev-sharma/CSC2516-GAN-class-imbalance`

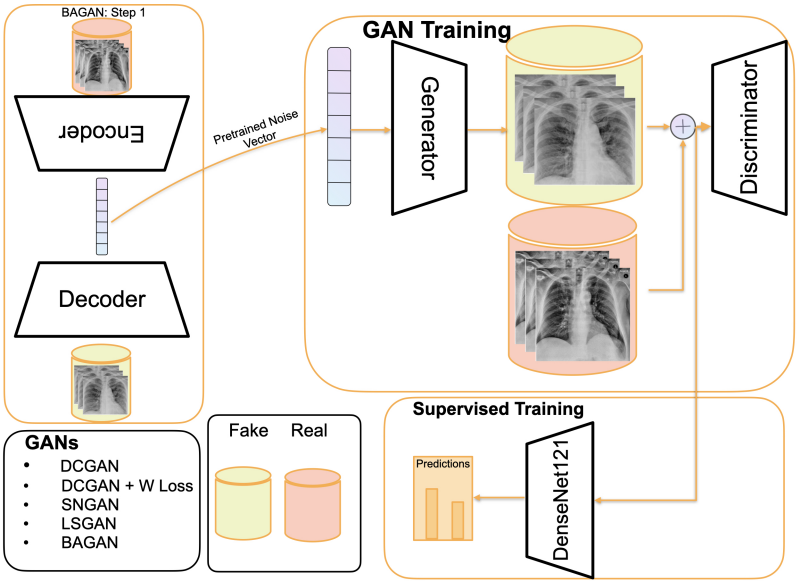

Figure 1: Overview of our approach. Different GAN architectures were used to synthesize fake X-ray data. The synthesized data were fed through a classifier for disease class prediction.

1 Introduction

Despite the success of deep neural networks in most use-cases, they fail to generalize when the dataset is imbalanced. With Generative Adversarial Networks (GANs), synthetic data can be generated so that the datasets have more samples as well as a more balanced distribution across different classes. The addition of these synthetic data can improve the performance and generalization of deep learning models. In this project, we wish to investigate recent advances and attempt to address their failure modes in GANs on medical imaging (Chest X-ray) to improve the generalization of X-ray image classifiers. Classification using X-ray images is a fast, cheap, and efficient approach to help in disease diagnosis. Nonetheless, traditional deep learning models fail in most scenarios, as medical data is often accompanied by class imbalance. One such example of class imbalance is COVID-19 being a minority class among other classes in the Chest X-ray dataset [1]. Hence, it is crucial to study class imbalance in the healthcare environment.

We aim to answer following research questions in the context of this project:

- Does incorporating synthetic images synthesized by GANs improve medical imaging classification in an imbalanced data setting?
- Are generative networks truly learning the distributions?

2 Related Works

The generation of new data based on the same training set using Generative Adversarial Nets [2] in an adversarial fashion is one of the avenues to tackle the challenge of class imbalance, where the acquisition of additional data is impractical or inconvenient, a very common problem in medical imaging.

Previous works [3, 4] based on data augmentations show that the synthesis of images using GANs improves the classification rate of convolutional neural networks models where the task of collecting a large dataset is unfeasible. One challenge that GAN-based data augmentation methods face is that the discriminator typically suffers from over-fitting when trained on limited data shown in [5, 6, 7], degrading the fidelity of generated images.

Imbalance in image classification is one of the fundamental problems affecting the accuracy of deep learning classifiers. Data augmentation methods based on GANs, such as BAGAN [8] and Improved BAGAN[9], have proven to improve the quality of images in the presence of imbalanced datasets.

In [10], the authors demonstrate how GANs can successfully augment training data in the presence of limited data in the medical imaging field. In particular, concerning X-ray images, empirical studies [11] show how Deep Convolutional GANs in conjunction with adaptive input-image normalization effectively address the intra-class mode collapse problem by successfully augmenting and balancing datasets. Another study [12] on X-ray classification supports how GAN techniques for data augmentation offer high performance for datasets with underrepresenting classes.

Our goal is to investigate how GANs behave with imbalanced medical datasets. We pay close attention to two factors: (1) the effect of dataset size in training and eventually in the quality of the generated images based on the Fréchet inception distance score, and (2) the extent to which the models we examine are learning the true embedding of the latent vectors. In this regard, we build on the work of Improved BAGAN [9].

3 Methodology

We trained a number of GAN architectures from scratch to produce synthetic data to mitigate the issue of class imbalance. The synthetic data along with the original dataset was then fed into a state-of-the-art supervised classifier (DenseNet121) for class prediction (Figure 1).

3.1 Experimentation Setup

The GAN architectures selected for this work were as follows:

- DCGAN: a Deep Convolutional Generative Adversarial Network, which, as opposed to fully connected layers in a simple GAN, uses convolutional layers and the transposed convolution technique for upsampling in its generator [13].
- Wasserstein GAN (WGAN): an extension of simple GAN, which is more stable in training and addresses the issue of mode collapse. This model uses Wasserstein distance and weight clipping, which leads to better theoretical properties of the cost function, such as smoother gradients [14].
- LSGAN: an extension of simple GAN that uses a Least Squares loss function in order to solve the problem of vanishing gradients [15].
- SNGAN: a type of GAN that uses spectral normalization as a means of normalizing weights to make the discriminator training more stable [16].
- Improved BAGAN: a variation of BAGAN [17] aimed to perform better than its predecessor when images in different classes look similar, e.g. flowers, x-ray scans, or cells. It uses a supervised encoder and an intermediate embedding model to disperse the labeled latent vectors.

3.2 Problem Formulation

The loss function of the original GAN [2] is described as follows:

$$\min_G \max_D \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[1 - \log D(G(z))] \tag{1}$$

where G denotes the generator function and D the discriminator, x is sampled from the real distribution and z from the generated distribution.

Our main objective is to overcome class imabalance issues when training GANs. In an scenario where the dataset is imbalanced, a vanilla GAN will train more on the majority class. To lessen this issue, Improved BAGAN adds the loss proposed in conditional DRAGAN [18] leveraging the concept of gradient penalty. Improved BAGAN's discriminator loss becomes:

$$L^D(X, Z, Y_r, Y_f, Y_s) = -\mathbb{E}[\log(D(x_r, y_r)] - \mathbb{E}[\log(1 - D(G(z, y_f)))] - \mathbb{E}[\log(1 - D(x_r, y_s))] \tag{2}$$

where z is a random noise vector $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, $y_f \sim \mathcal{U}(0, 1, 2...)$ and $y_s \sim \mathcal{U}(0, 1, 2...)$.

The loss function of the generator is:

$$L^G(Z, Y_f) = -\mathbb{E}[\log D(G(z, y_f))] \tag{3}$$

3.3 Autoencoder and Embeddings

In contrast to the other GAN baselines, Improved BAGAN employs an autoencoder as initialization. The autoencoder contains an embedding section with latent vectors computed by means and covariances of the encoded training data. Based on this idea, the generator can be seen as an extension of the pretrained embedding model and decoder model: a random latent vector and a random label are fed into the generator to obtain a synthetic image of a specific class. During training the embedding model inside the generator is updated.

3.4 Datasets

In an imbalanced medical imaging setting, we explored the following Chest X-ray datasets:

- **COVID-chest X-ray dataset [19]:** The class distribution of the dataset is shown in Table 3.
- **RSNA Pneumonia dataset:** This data comes from the NIH CXR14 dataset [20], which consists of 30,000 images. We only used 10% and 50% of the dataset, called RSNA_10 and RNSA_50 respectively. Table 3 summarizes the class distribution of the two datasets.

Table 1: Classification performances of different models on different fractions of the data as measured by their AUCROC score, as well as the FID scores of each of the GAN models.

Fraction of the training data	**Model**						
	Supervised baseline (DenseNet121)	**Supervised baseline with data augmentation**	**WGAN**	**DCGAN**	**SNGAN**	**LSGAN**	**Improved BAGAN**
5 %	0.4873	0.49509	0.6036	0.61309	0.61147	0.54912	**0.6330**
10 %	0.5417	0.5878	0.754219	0.85032	0.6367	0.83967	0.77249
50 %	0.7695	0.7439	0.82700	0.8464	0.860513	0.8715	0.832594
100 %	0.7757	0.79055	0.85376	0.75286	0.8513	0.83208	0.8646
FID Score	-	-	298.434	243	247.518	236.818	**223**

3.5 Baseline

We used a DenseNet121 as our classifier for predicting the pathology of the synthesized data. The classifier as a baseline for comparison was trained with and without augmentation with a split of 75:5:20 into Training:Validation:Testing sets. We employed horizontal and vertical flip as standard augmentation methods. The classification results can be seen in Table 2.

3.6 Experimental Results

To investigate our research questions, we performed extensive analysis. We trained various GANs: DCGAN, WGAN, LSGAN, and Improved BAGAN from scratch and evaluated on FID (Frechet Inception Distance) score for convergence. Further, we synthesized samples from minority classes to be in the same proportion as majority classes. Lastly, the new dataset synthesized by GANs was evaluated on AUC-ROC and it was compared against supervised baselines with and without augmentation (horizontal and vertical flipping). The hyperparameters tested for Improved BAGAN can be seen on the Table 5.

From Table 1, we can categorically observe that Improved BAGAN gives the best convergence (lowest FID score) on COVID dataset. In comparison to supervised baseline with and without augmentation, we noticed a performance gain in AUC-ROC score, suggesting that generative networks indeed improve the classification accuracy in an imbalanced setting. Upon evaluating the performance under different data settings, we noticed that in a highly imbalanced setting (5 % of the training data), Improved BAGAN outperforms other standard GAN architectures. However, this is not necessarily true in high data settings (10%, 50% or 100%) of the training data as there is no competitive advantage observed over others generative networks. We believe this is partly because the COVID dataset is a very small and noisy dataset (images obtained from different sources), resulting in a decrease in the performance of Improved BAGAN due to discriminator overfitting, like other GANs trained.

To examine this, we visualized the embedding of the encoded latent vectors using a two-dimensional t-SNE plot (Figure 2). Both labeled latent vectors of DCGAN and Improved BAGAN reflect a high concentration of samples in specific sectors. In particular, DCGAN has one unique cluster of blue points (synthetic or fake data) that does not mimic the original distribution of real data. Improved BAGAN plot displays similar results: two blue clusters containing synthetic data scarcely follow the real data points distribution. An additional insight from the plots is that both models show a mode collapse problem: the generators produce a consistent set of images with no variability as shown by the blue points around the same region. This indicates that Improved BAGAN suffers from discriminator overfitting in a highly imabalanced dataset.

In addition to that, we investigated the effect of varying the dataset size to see the impact on the performance. As illustrated in Table 2, we observed that more samples and longer training time leads to a lower FID score(4), generating high fidelity and diverse images. Also, it further improves the AUC-ROC score from 0.83 to 0.89 on 10% and 50% of the RSNA dataset.

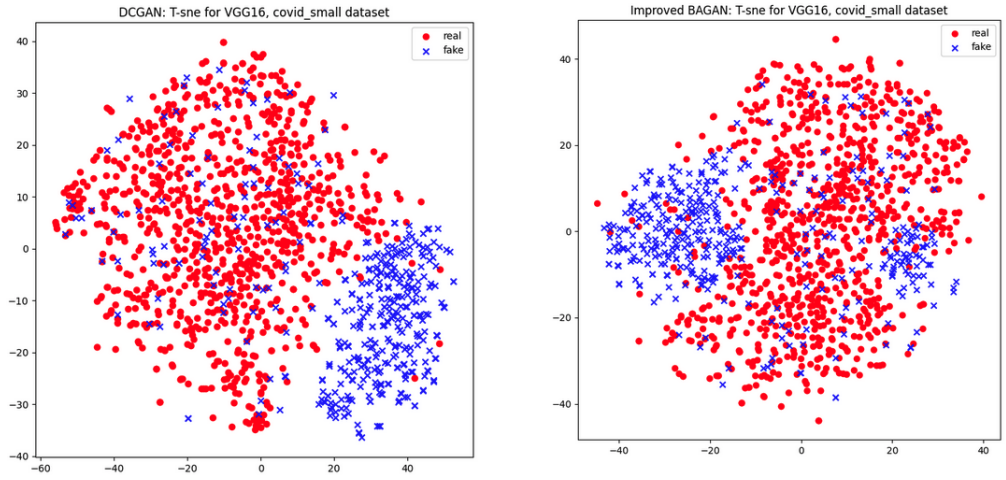

Figure 2: Embedding of the encoded latent vectors using a two-dimensional t-SNE plot. Left is the t-SNE plot for the DCGAN model and Right is the t-SNE plot for the improved BAGAN model.

Table 2: Classification results on the synthetic data generated from each dataset. The FID score was calculated with 100 epochs.

Dataset	FID	Samples	Classification (AUC-ROC)
Covid	223	1130	0.8646
RSNA_10	194.189	2668	0.83765
RSNA_50	98.485	12007	0.88706

4 Conclusion

In conclusion, in a medical imaging setting, where class imbalance is a fundamental problem, we comprehensively explored generative networks to evaluate their effectiveness. We observed that although synthesized images aid in better performance in downstream tasks in contrast to supervised baselines, it still suffers from discriminator overfitting due to class imbalance, especially in a highly imbalanced dataset. We showcased that not only standard GAN networks, but also the state-of-the-art Improved BAGAN, suffer from the issue if the imbalance is large enough as shown by our empirical results and t-SNE visualizations where the fake and real distributions do not converge successfully.

Future Work / Limitations In this section, we list down the limitations of our project. To assess the quality of synthesized images, it needs to be further evaluated by radiologists as FID score is not a good proxy as natural images are different from chest X-ray medical images. Given the project scope and duration, we didn't compare Improved BAGAN against state-of-the-art networks, such as Style-GAN and ProGAN, which would be vital to evaluate the effectiveness of Improved BAGAN in an imbalanced data setting. Although Improved BAGAN does not offer promising results on COVID dataset, we wish to further test the validity of the approach by comprehensively testing on a large dataset, such as RSNA. Contrary to what has been claimed by the authors of Improved BAGAN, we didn't observe better performance. This might be because Improved BAGAN has an exhaustive list of hyperparameters which weren't explored completely for the COVID dataset and as a next step we wish to investigate that too.

References

[1] Seung Hoon Yoo et al. "Deep Learning-Based Decision-Tree Classifier for COVID-19 Diagnosis From Chest X-ray Imaging". In: *Frontiers in Medicine* 7 (2020). ISSN: 2296-858X.

DOI: 10.3389/fmed.2020.00427. URL: https://www.frontiersin.org/article/10.3389/fmed.2020.00427.

[2] Ian J. Goodfellow et al. *Generative Adversarial Networks*. 2014. arXiv: 1406.2661 [stat.ML].

[3] Georgios Douzas and Fernando Bacao. "Effective data generation for imbalanced learning using conditional generative adversarial networks". In: *Expert Systems with Applications* 91 (2018), pp. 464–471. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2017.09.030. URL: https://www.sciencedirect.com/science/article/pii/S0957417417306346.

[4] Shobhita Sundaram and Neha Hulkund. "GAN-based data augmentation for chest X-ray classification". In: (July 2021). arXiv: 2107.02970 [eess.IV].

[5] Tero Karras et al. "Training Generative Adversarial Networks with Limited Data". In: *CoRR* abs/2006.06676 (2020). arXiv: 2006.06676. URL: https://arxiv.org/abs/2006.06676.

[6] Shengyu Zhao et al. "Differentiable Augmentation for Data-Efficient GAN Training". In: *CoRR* abs/2006.10738 (2020). arXiv: 2006.10738. URL: https://arxiv.org/abs/2006.10738.

[7] Ryan Webster et al. "Detecting Overfitting of Deep Generative Networks via Latent Recovery". In: *CoRR* abs/1901.03396 (2019). arXiv: 1901.03396. URL: http://arxiv.org/abs/1901.03396.

[8] Giovanni Mariani et al. "BAGAN: Data Augmentation with Balancing GAN". In: *CoRR* abs/1803.09655 (2018). arXiv: 1803.09655. URL: http://arxiv.org/abs/1803.09655.

[9] Gaofeng Huang and Amir H. Jafari. "Enhanced Balancing GAN: Minority-class Image Generation". In: *CoRR* abs/2011.00189 (2020). arXiv: 2011.00189. URL: https://arxiv.org/abs/2011.00189.

[10] Christopher Bowles et al. "GAN Augmentation: Augmenting Training Data using Generative Adversarial Networks". In: *CoRR* abs/1810.10863 (2018). arXiv: 1810.10863. URL: http://arxiv.org/abs/1810.10863.

[11] Muhammad Muneeb Saad, Mubashir Husain Rehmani, and Ruairi O'Reilly. "Addressing the Intra-class Mode Collapse Problem using Adaptive Input Image". In: *CoRR* (2022). URL: https://arxiv.org/abs/2201.10324.

[12] Shobbita Sundaram and Neha Hulkund. "GAN-based Data Augmentation for Chest X-ray Classification". In: *CoRR* (2021). URL: https://arxiv.org/pdf/2107.02970.pdf.

[13] Süleyman Aslan et al. "Deep Convolutional Generative Adversarial Networks Based Flame Detection in Video". In: *CoRR* abs/1902.01824 (2019). arXiv: 1902.01824. URL: http://arxiv.org/abs/1902.01824.

[14] Martin Arjovsky, Soumith Chintala, and Léon Bottou. "Wasserstein GAN". In: (2017). DOI: 10.48550/ARXIV.1701.07875. URL: https://arxiv.org/abs/1701.07875.

[15] Xudong Mao et al. "Least Squares Generative Adversarial Networks". In: (2017), pp. 2813–2821. DOI: 10.1109/ICCV.2017.304.

[16] Takeru Miyato et al. "Spectral Normalization for Generative Adversarial Networks". In: (2018). DOI: 10.48550/ARXIV.1802.05957. URL: https://arxiv.org/abs/1802.05957.

[17] Giovanni Mariani et al. "BAGAN: Data Augmentation with Balancing GAN". In: *CoRR* abs/1803.09655 (2018). arXiv: 1803.09655. URL: http://arxiv.org/abs/1803.09655.

[18] Naveen Kodali et al. *On Convergence and Stability of GANs*. 2017. DOI: 10.48550/ARXIV.1705.07215. URL: https://arxiv.org/abs/1705.07215.

[19] Tulin Ozturk et al. "Automated detection of COVID-19 cases using deep neural networks with X-ray images". In: *Computers in Biology and Medicine* 121 (2020), p. 103792. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2020.103792. URL: https://www.sciencedirect.com/science/article/pii/S0010482520301621.

[20] Xiaosong Wang et al. "ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases". In: (July 2017). DOI: 10.1109/cvpr.2017.369. URL: https://doi.org/10.1109%2Fcvpr.2017.369.

A Appendix

A.1 Experimental Results

Table 3: Class distribution of each of the datasets.

Dataset	**Class Names**	**# Samples**	
		Training	Validation
COVID	COVID	116	13
	Normal	450	50
	Pneumonia	451	50
RSNA_10	Normal	1542	171
	Disease	859	96
RSNA_50	Normal	9358	1040
	Disease	2649	295

Table 4: FID scores of the improved BAGAN model on different datasets with different number of epochs.

Model	**Dataset**	**# Samples**	**# Epochs**	**FID Score**
Improved BAGAN	Fashion MNIST	10,000	15	235.53
	COVID	1,130	15	298.774
			30	277.286
			50	258.323
			100	243.418
			200	243.366
	RSNA_10	2,668	15	243.689
			30	238.63
			50	207.84
			100	194.189
			200	181.425
	RSNA_50	12,007	15	164.323
			30	140.046
			50	123.377
			100	98.485

A.2 Contributions

- Vasudev:
  - Setting up, training supervised baseline (with and without augmentation) on COVID and RSNA dataset on PyTorch
  - Report: writing conclusion,limitations and experimental results.
  - Hyper parameter tuning
  - Training SNGAN, DCGAN, WGAN, LSGAN from scratch and synthesizing minority class on COVID and RSNA dataset
  - Evaluating AUC-ROC performance on the synthesized + original images
- Gabriel
  - Report: related works, methodology and contributed to other sections.
  - Hyperparameter tuning for Improved BAGAN.
  - Training BAGAN, Improved BAGAN models on COVID and RSNA datasets using Tensorflow.
  - Generating t-SNE embedding plots.
- Mohammad
  - Setting up data augmentations

- Setting up the augmentation pipeline for adaptive pseudo augmentation (dropped from the project)
- PyTorch implementation of BAGAN
- Report: methodology section, generated figures and table results

Table 5: Values tested for the hyperparameters of the improved BAGAN model on RNSA 10. Number of epochs used during the tuning process was set to 50.

Hyperparameters	**Tested values**
Learning Rate	0.0002
	0.0001
	0.00002
	0.00001
Gradient Penalty	0
	0.1
	1.0
Spectral Norm Generator	Yes
	No
Spectral Norm Discriminator	Yes
	No
Activation Function for Generator	ReLU
	Elu
	LeakyRelu
Activation Function for Discriminator	ReLU
	Elu
	LeakyRelu