eGAN: Unsupervised approach to class imbalance using transfer learning

Ademola Okerinde, Lior Shamir, William Hsu, Tom Theis, and Nasik Nafi

Kansas State University, 2164 Engineering Hall, Manhattan, KS 43017-6221, USA
{okerinde,lshamir,bhsu,theis,nnafi}@ksu.edu

Abstract. Class imbalance is an inherent problem in many machine learning classification tasks. This often leads to trained models that are unusable for any practical purpose. In this study we explore an unsupervised approach to address these imbalances by leveraging transfer learning from pre-trained image classification models to encoder-based Generative Adversarial Network (eGAN). To the best of our knowledge, this is the first work to tackle this problem using GAN without needing to augment with synthesized fake images.

In the proposed approach we use the discriminator network to output a negative or positive score. We classify as minority, test samples with negative scores and as majority those with positive scores. Our approach eliminates epistemic uncertainty in model predictions, as the $P(minority) + P(majority)$ need not sum up to 1. The impact of transfer learning and combinations of different pre-trained image classification models at the generator and discriminator is also explored.

Best result of 0.69 F1-score was obtained on CIFAR-10 classification task with imbalance ratio of 1:2500.

Our approach also provides a mechanism of thresholding the specificity or sensitivity of our machine learning system.

Keywords: Class imbalance, Transfer Learning, GAN, nash equilibrium

1 INTRODUCTION

A dataset is said to be imbalanced when there is a significant, or in some cases extreme, disproportion between the number of examples of the different classes in the dataset. The class or classes with large number of samples are called the majority while the class with few examples are denoted as the minority. In many cases, the machine learning model is required to correctly classify the minority class while minimizing the misclassification of the majority class. However, the skewness in data often leads machine learning classification methods to favour the majority class.

Class imbalance problem in computer vision are normally approached in the following ways, either at the data or algorithm level. Data augmentation is a technique of transforming images via scaling, cropping, flipping, padding, rotation, brightness, contrast, saturation level etc. As a result, a humongous image dataset can be created from a single image of the minority class. By using data

augmentation, a class with a small number of samples can be expanded into a class with a much larger number of samples. The synthetic images can also be generated by using GAN to augment the minority class. Other approaches include deliberate undersampling of the majority class or oversampling of the minority class by mere copying. The earlier approach leads to loss of useful data information while the latter approach causes overfitting.

At the algorithm level, the objective function is tweaked to heavily penalize the network for mis-classifying the minority class. The most popular is cost-sensitive approach. Here, the classifier iis modified to incorporate varying penalty for each of considered groups of examples. This way by assigning a higher cost to less represented set of objects its importance is boost during training.

Transfer learning has been known to help improve the performance of machine learning models. By fine-tuning varying number of layers in the pre-trained image classification model, the pre-trained model can serve as a feature extractor, while adding a classifier head for more specific feature learning for the current task.

In this work we compared the performance of various pre-trained image classification models for the task of unsupervised image classification with varying imbalance ratios. Our architecture, named eGAN, is developed to serve as a basis for this comparison. Using GAN [2], we re-parameterise the job of the discriminator as a classifier giving a positive score to majority samples and negative to minority ones. While most GAN-based architectures focus on the output of the generator, in the proposed architecture, we intuitively adapted the vanilla GAN network by integrating an encoder module that takes minority samples as input and produces a latent code, from which the generator learns.

2 RELATED WORK

In [4] an ensemble method was proposed based on advanced generative adversarial network to generate new samples for the minority class to restore balance. Our opinion is that the computational demand of such approach is enormous, and not many low-income countries of the world have access to such compute power. We eliminate the need to generate a realistic image in our technique because our network is trained on imbalance dataset. Deep Cascading (DC) with a long sequence of decision trees could help to handle unbalanced data [1]. A DC is a sequence of n classifiers where each sample x passes to the next classifier only if the current one classifies it as positive according to a high-sensitivity decision threshold. However, this works well with foreground-background imbalance unlike classification task. Transfer learning with GAN was used to generate images from limited data in [8]. Their result showed that knowledge from pre-trained networks can shorten the convergence time and significantly improve the quality of generated images.

3 METHODOLOGY AND EXPERIMENTAL DESIGN

In this section, we discuss our proposed approach and the various testbeds that were used in our experiments.

3.1 ADDRESSING CLASS IMBALANCE WITH eGAN

The proposed architecture is based on adaptation of existing Deep Convolutional Generative Adversarial Network (DCGAN)[7] by incorporating an encoder module. This module encodes minority samples in latent space needed by the generator G to generate minority samples that are capable of fooling the discriminator D. On the other hand, the discriminator D is fed with data samples drawn from majority distribution and the generated output of the generator G. D and G are simultaneously optimized through the following two-player minimax game with value function V(G,D) in 1.

$$\overset{min}{G}\ \overset{max}{D}\ V(D,G) = E_{X_{ma}\sim P_{ma}}[\log D(X_{ma})] + E_{X_{mi}\sim P_{mi}}[\log(1 - D(G(X_{mi})))] \tag{1}$$

where X_{ma} and X_{mi} are majority and minority sample distributions respectively.

Over the course of iteration, the discriminator D is optimized to assign a negative score to the minority data distribution and a positive score to the majority data distribution. This enables the discriminator D to act as a classifier.

Encoder-Generator module Our latent space is composed of 128 units vector. Rather than feeding the generator with random noise as is typical of most GAN implementation, we added an encoder module that forces the generator to learn from known distribution (minority distribution). The encoder part consists of the pre-trained DenseNet121 followed by global average pooling layer and latent dimension space. The generator part has two transposed convolutional layers. We use leakyRelu activation function with alpha set to 0.2; batch normalization and Sigmoid function at the final layer.

Discriminator module The pre-trained discriminator has 7,038,529 parameters out of which only 39,937 are trainable. A layer of global average pooling follows the pre-trained DenseNet121. We use a dropout of 0.2 followed by the final one unit dense layer. The overall architecture of our encoder-based generator adversarial network is shown in Figure 1.

3.2 Selection of pre-trained image classification weights

We perform experiments on VGG16, VGG19, EfficientNetB2, ResNet101 and DenseNet121 pre-trained classification models on ImageNet dataset. Here, we fine-tuned only top five layers at each of the pre-trained models. Table 3 shows

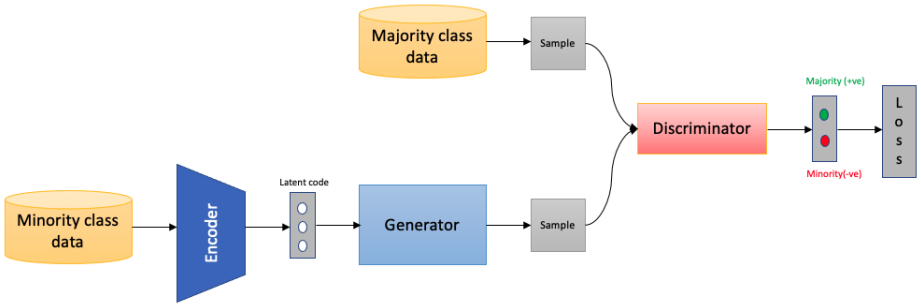

Fig. 1. Encoder-based Generative Adversarial Network (eGAN) architecture

the maximum precision, recall and F1-score obtained on CIFAR-100 with imabalance ratio 1:50 by using different combinations of pre-trained models.

DenseNet121 [3] was used for pretraining our eGAN. After experimenting different pre-trained architectures and different layers of fine-tuning, we obtained best result with fine-tuning only top 5-layer out 427 layers of DenseNet121.

3.3 Dataset

Several commonly used datasets were used in this study. In order to model the real-world scenario of heavy imbalance, we used only few samples of the minority class as input to the encoder module. Detail overview is shown in Table 4.

CIFAR-10: The CIFAR-10 dataset [6] consists of 60,000 32×32 colour images in 10 classes, with 6000 images per class. There are 50,000 training images and 10,000 test images. Here we use *airplane* as minority and *automobile* as majority.

CIFAR-100: CIFAR-100 [6] dataset is similar to CIFAR-10, except it has 100 classes containing 600 images each. Each class has 500 training images and 100 testing images. The 100 classes in the CIFAR-100 are grouped into 20 superclasses. Each image comes with a "fine" label (the class in which it belongs) and a "coarse" label (the superclass in which it belongs).

CheXpert: We use the pneumonia subset of Stanford CheXpert dataset [5] for experimenting on an inherent imbalanced dataset. The dataset contains 4576 and 167407 minority and majority samples, respectively.

4 RESULTS AND DISCUSSION

All discriminator scores that are less than zero are classified as minority, otherwise they are classified as majority class. Table 3 shows the result obtained by

Table 1. Confusion matrix of imbalance ratio 1:1000

	Predicted minority	Predicted majority
Actually minority	810	190
Actually majority	366	634

Table 2. Confusion matrix of imbalance ratio 1:2500

	Predicted minority	Predicted majority
Actually minority	821	179
Actually majority	767	233

combining various pre-trained models. Adam optimizer with a learning rate of 1e-4 was used to train all models for 100 epochs.

As can be seen in Figure 4, the model achieved a nash equilibrium on test data at around 10 epochs. Here, we perform inference on test data at every epoch and plot the number of samples correctly classified. Five layers of DenseNet121 were fine-tuned at each generator and discriminator module, while 422 layers' weight were kept fixed. At nash, the discriminator correctly classified roughly 700/1000 of each of the minority and majority test data. This result convinces us that transfer learning with GAN can be used to overcome the challenge of highly imbalanced dataset, owing to the fact that we train only with 10 samples of the minority class and 5000 samples of the majority class (ratio 1:500). Similar performance is observed in CIFAR-100 with imbalance ratio 1:50.

Without pre-training the discriminator, the effect of the high imbalance in the training set is revealed, as the discriminator is skewed towards the majority class in the training set, thereby missing all the minority samples in the test data. This can be seen in Figure 5 on CIFAR-100. This behaviour pattern is observed on CIFAR-10 as well. We experiment with no pre-training at all, neither in the discriminator nor generator, and observed exact same pattern. Therefore, we can safely conclude that the use of transfer learning helps unsupervised image classification in a highly imbalanced domain.

Training can be stopped as soon as nash equilibrium is reached, as this point gives the model best performance on the minority and majority class. An acceptable threshold can also be set for the absolute difference of the number of correctly classified samples of both classes. For instance, if the $|$ correctly_classified_minority - correctly_classified_majority $| \leq 20$. The precision, recall and F1-score curves on CIFAR-10 averaged over five folds at different imbalance ratios are shown in Figure 4.

We observed that at the early training epochs typically between 1 and 40 epochs, the generator tries to achieve its objective of fooling the discriminator by generating samples from the majority class fed into the discriminator, which results in more of the minority samples being mis-classified as the discriminator "knows" the distribution of the majority too well. A drastic change occurs when the generator start generating samples from latent vector which is a able to

Table 3. Comparative analysis of different pre-trained models configuration on Generator and Discriminator using CIFAR-100 dataset

Discriminator pre-trained	Generator pre-trained	Precision	Recall	F1
ResNet101	VGG19	0.72	1.0	0.78
VGG19	ResNet101	0.73	1.0	0.69
EfficientNetB2	VGG19	1.0	0.22	0.32
VGG19	EfficientNetB2	0.7	0.86	0.71
ResNet101	VGG16	1.0	0.17	0.27

conveniently fool the discriminator as can be seen in generator and discriminator loss shown in Figure 2.

4.1 Imbalance ratios

In this section we discuss the performance of the proposed approach on various class imbalance ratios. To eliminate bias in model performance, we conducted 5-fold-cross-validation on the minority samples and average the result.

Class ratio of 1:2500 We experiment on CIFAR-10 dataset by deliberately using an unbalanced subset of the training set. At the 4th epoch, our model correctly identifies 257 minority and 821 majority samples out of 1000 each. At epoch 5, a sharp change occurred that led to 821 minority samples being correctly classified, while only correctly classifying 233 majority samples as shown in Table 2. We also observed that at epoch 72 the performance of the network on the majority and minority classes reached a nash equilibrium with a threshold difference of less than or equal to 20.

Class ratio of 1:1000 At epoch 81 on CIFAR-10 dataset, nash equilibrium was reached. At this epoch, 532 and 525 minority and majority test data respectively were correctly classified. We observed that the classifier had another major shift between epoch 5 and 6. At epoch 5, best result was obtained. The network was able to classify 863 majority tests and 585 minority tests correctly out of the 1000 samples. At epoch 6, 634 majority and 810 minority tests were classified correctly as swown in Table 1. Maximum precision, F1-score and recall of 0.88, 0.74 and 0.99 were obtained at epochs 3, 6 and 37, respectively.

Class ratio of 1:500 Both CIFAR-10 and CIFAR-100 were used to experiment imbalance ratio 1:500. On CIFAR-10, maximum precision, recall and F1-score on averaging 5-fold-cross-validation are 0.75, 0.95 and 0.70 respectively as shown in Table 5. The maximum precision is slightly lower on CIFAR-100 with 0.60. However, the recall and F1-score which are 0.96 and 0.68 are roughly the same.

Class ratio of 1:50 We demonstrate our model performance on imbalance ratio 1:50 using CIFAR-100 and CIFAR-10. For CIFAR-100, a sudden change occurred between epoch 69 and 70 as follows majority: 57, minority: 59; and majority: 55, minority: 59. A nash equilibrium is attained at epoch 68, with 56 correctly classified minority as well as majority class. At epoch 97 maximum F1-score and recall of 0.66 and 0.77 were obtained respectively, while maximum precision of 0.6 was obtained at epoch 74.

Class ratio of 1:1 We use CIFAR-100 to demonstrate the performance of eGAN on a balanced dataset. We notice that the experimental performance follow the same pattern as imbalanced dataset. Training starts with mostly all the majority correctly classified and all the minority mis-classified. At epoch 48, a nash equilibrium (with threshold less than or equal to 5) is achieved, with 76 and 71 of minority and majority correctly classified respectively. The maximum F1-score of 0.78 is reached at epoch 53 as shown in Table 5. Instead of using 500 samples each of minority and majority class, by training on a single instance of minority and majority sample (*1:1) of CIFAR-100, we obtained an F1-score of 0.63. This demonstrates the impact of transfer learning on the training.

Class ratio of 1:36 For pneumonia subset of CheXpert dataset with imbalance ratio 1:36, the best performed model achieves 0.51, 0.97, and 0.67 precision, recall, and F1-score respectively. The results shown in Table 6 is evaluated on 1000 of each minority and majority test set. As can be seen in the table, our approach did not beat the baseline classification model because this task is more of an anomaly detection task rather than a classification problem. Also the pre-trained image classification model source dataset i.e. ImageNet is significantly different from medical domain. In our opinion, exploring more variants of complex GAN architectures like BigGAN, StyleGAN and ProGAN could possibly help.

Table 4. CIFAR-10 and CIFAR-100 dataset overview - CIFAR-100 in parenthesis

Class imbalance ratio	# training minor	# training major	# testing minor	# testing major
1:2500	2(-)	5000(-)	1000(-)	1000(-)
1:1000	5(-)	5000(-)	1000(-)	1000(-)
1:500	10(1)	5000(500)	1000(100)	1000(100)
1:50	100(10)	5000(500)	1000(100)	1000(100)
1:1	500(500)	500(500)	1000(100)	1000(100)
*1:1	-(1)	-(1)	-(100)	-(100)

Table 5. Maximum precision, recall and F1-score on CIFAR-10 and CIFAR-100 (avg. 5-fold) - CIFAR-100 in parenthesis

ratio	Precision	Recall	F1
1:2500	0.72(-)	0.94(-)	0.69(-)
1:1000	0.72(-)	0.96(-)	0.69(-)
1:500	0.75(0.60)	0.95(0.96)	0.70(0.68)
1:50	0.78(0.7)	0.97(0.86)	0.69(0.71)
1:1	**0.82**(0.72)	0.98(**1.0**)	0.72(**0.78**)
*1:1	-(0.53)	-(0.86)	-(0.63)

Table 6. Precision, recall and F1-score on pneumonia subset of CheXpert dataset

	#training		#testing					
Model	minor	major	minor	major	imbalance ratio	Precision	Recall	F1
eGAN	30	1080	1000	1000	1:36	0.51	0.97	0.67
baseline	30	1080	1000	1000	1:36	0.5	1.0	0.67

5 CONCLUSION

In this work, we explore the use of pre-trained image classification models on the task of image classification on varying levels of imbalance ratios in the training dataset. Our approach demonstrates unsupervised technique to addressing class imbalance by using pre-trained GAN. Rather than synthesizing fake images with the generator for data augmentation, we employ the discriminator as a classifier between the classes by scoring minorities and majorities negatives and positives respectively. The performance measure of interest plays a significant role in deciding what epoch of the trained model to deploy in production as various epochs favour different evaluation metrics for example sensitivity. Experimental results reveal that transfer learning plays a significant role in the model performance. In continuation of this work, we will explore the usage of this approach for anomaly detection task where distinguishing features between normal(majority) and ab-

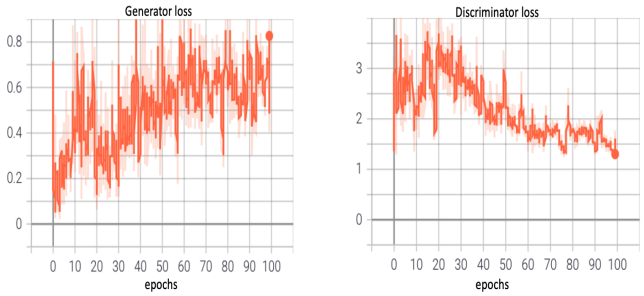

Fig. 2. Discriminator and Generator loss.

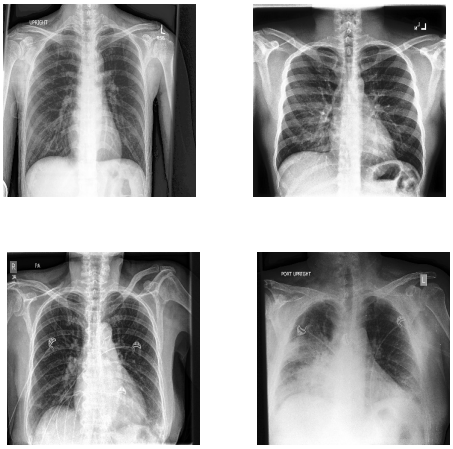

Fig. 3. CheXpert pneumonia Xray - majority class (top row), minority class (bottom row).

normal(minority) is less profound. Our work can be further explored in object detection tasks in cases of imbalance between foreground and background.

References

[1] Alessandro Bria, Claudio Marrocco, and Francesco Tortorella. "Addressing class imbalance in deep learning for small lesion detection on medical images". In: *Computers in Biology and Medicine* 120 (2020), p. 103735.

[2] Ian J Goodfellow et al. "Generative adversarial networks". In: *arXiv preprint arXiv:1406.2661* (2014).

[3] Gao Huang et al. "Densely connected convolutional networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 4700–4708.

[4] Yangru Huang et al. "Towards Imbalanced Image Classification: A Generative Adversarial Network Ensemble Learning Method". In: *IEEE Access* 8 (2020), pp. 88399–88409.

[5] Jeremy Irvin et al. "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01. 2019, pp. 590–597.

[6] Alex Krizhevsky, Geoffrey Hinton, et al. "Learning multiple layers of features from tiny images". In: (2009).

[7] Alec Radford, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks". In: *arXiv preprint arXiv:1511.06434* (2015).

[8] Yaxing Wang et al. "Transferring gans: generating images from limited data". In: *Proceedings of the European Conference on Computer Vision*. 2018, pp. 218–234.

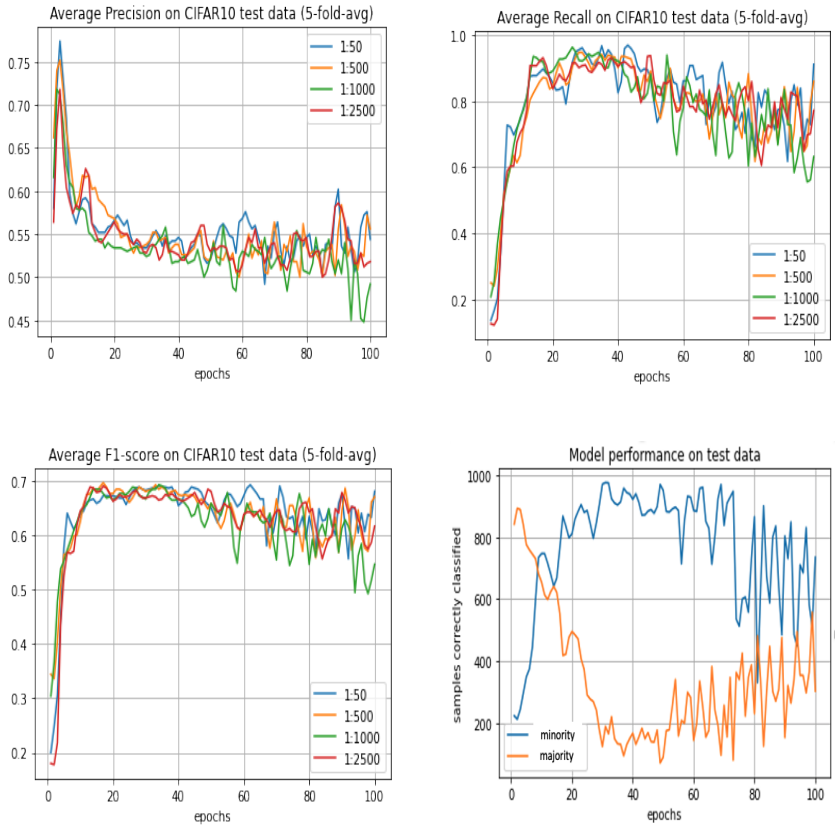

Fig. 4. eGAN's average precision (top left), average recall (top right), average F1-score (bottom left), and average performance (bottom right) on CIFAR-10 test dataset using DenseNet121.

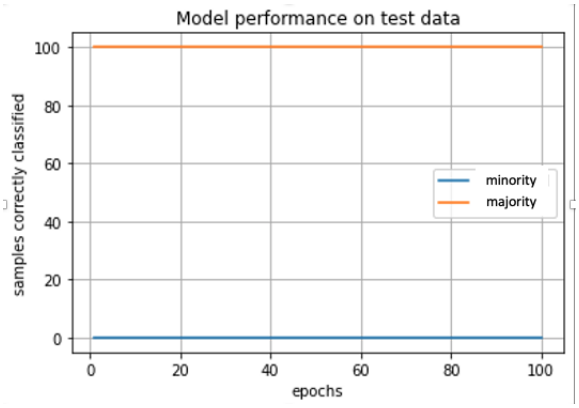

Fig. 5. Performance of eGAN on **CIFAR-100** test data using DenseNet121. Only the generator network is pre-trained