DOI: 10.21917/ijivp.2022.0396

DIMENSIONALITY REDUCTION BASED CLASSIFICATION USING GENERATIVE ADVERSARIAL NETWORKS DATASET GENERATION

G. Narendra and D. Sivakumar

Department of Electronics and Instrumentation Engineering, Annamalai University, India

Abstract

The term data augmentation refers to an approach that can be used to prevent overfitting in the training dataset, which is where the issue first manifests itself. This is based on the assumption that extra datasets can be improved by include new information that is of use. It is feasible to create an artificially larger training dataset by utilizing methods such as data warping and oversampling. This will allow for the creation of more accurate models. This idea is demonstrated through the application of a variety of different methods, some of which include neural style transfer, adversarial training, and erasure by random erasure, amongst others. By utilizing oversampling augmentations, it is feasible to create synthetic instances that can be incorporated into the training data. This is made possible by the generation of synthetic instances. There are numerous illustrations of this, including image merging, feature space enhancements, and generative adversarial networks, to name a few (GANs). In this paper, we aim to provide evidence that a Generative Adversarial Network can be used to convert regular images into Hyper Spectral Images (HSI). The purpose of the model is to generate data by including a certain amount of unpredictable noise.

Keywords:

Data Augmentation, GAN, Hyper Spectral Images, Classification

1. INTRODUCTION

When there is an imbalance in the number of samples coming from one class compared to another class, we say that a dataset has class imbalance. This is because we call this an imbalance in the number of samples. This might be a problem with binary classification, in which there is a clear division between majority and minority classes, or it might be a problem with multi-class classification, in which there is more than one majority class and more than one minority class. Either way, it possible that the issue lies with the way the data is organized. In either case, it is possible that the classification system is the root of the issue [1]-[4].

When applying machine learning models to datasets that have an unequal composition, there is a high potential for errors due to the fact that these datasets have a propensity to favor predictions for the majority of classes. As a direct consequence of this, accuracy is not a reliable performance indicator whenever there are inequities present in the dataset [5]. One of the methods that can be used to deal with class disparity and one of the ways that it can be employed is called data augmentation, which is a data-level technique that can leverage multiple implementation strategies. This is one of the ways that it can be used [6].

The following is an easy method for making use of data augmentation that does not lead to the introduction of any bias as a result of oversampling. You will need to carry out a random oversample by making some easy alterations to the sample geometry, such as rotating it by 30 degrees, in order to achieve this goal [7]. It is easy to generate the illusion of oversampling by applying very basic visual adjustments such as color augmentations, blending images, kernel filters, and random wipes. This will provide the impression that the image has been oversampled. As a result of this, putting together and executing tests with classes of varied sizes and make-ups is a simple and straightforward operation [8]. When a significant number of people are included in the sample, there is a greater possibility that the results will be skewed in favor of the underrepresented group that is being oversampled. This is one of the potential issues that might crop up in the future. When the sample has been completed, the prejudices held by the group that is underrepresented will be more apparent than they have ever been before in the history of the sample.

An oversampling strategy can be made more intelligent by using other learning techniques, such as adversarial training, neural style transfer (NST), generative adversarial networks (GANs), and meta-learning methods. One such method is referred to as Neural Style Transfer, and it makes it feasible to produce completely original pieces of visual art [9]. These one-of-a-kind images can be produced in one of two ways: either by mixing elements of multiple visual languages that are already present in the collection or by making use of an entirely separate visual language.

If we oversampled the data using GANs, we would be able to maintain the underlying distribution while simultaneously increasing the number of classes set aside for populations that are underrepresented. In order to train GANs, either a subset of the minority class or the minority class as a whole can be used as input. Both of these options are available [10]. It is highly conceivable that in the not-too-distant future, we will pick which categories should be utilized as inputs to a GAN by employing a technique that is known as evolutionary sampling.

In this research, we analysed GAN for label generation, then we applied dimensionality reduction techniques and then classifier in HSI image classification systems so that they are better able to tolerate unexpected inputs. These networks are used to classify images and are used in image classification systems.

2. BACKGROUND

The method that is known as data augmentation is a strategy that can be utilized to address the issue of insufficient data for training purposes. The creation of new instances is carried out in a manner that is compatible with the primary data set. This is the case. It possible that data augmentation could be misunderstood as implicit regularization due to the fact that it enables learners to improve their generalization after being presented with more comprehensive training sets. The implementation of easy affine and elastic changes to the image data is the common method utilized in computer vision for the purpose of augmenting data [11]. Even while these methods can be used on HSI data, they do not model what is actually important in a way that makes the most

of the data that is available to them. This prevents them from maximizing the potential of the data.

Because augmentation is used by only one of the deep-learning procedures that we have discussed up to this point, there is not a huge amount of research on HSI data augmentation. This is due to the fact that augmentation is used by only one of the deep-learning strategies.

In [12], the standard deviation of the training set is calculated on an individual basis for each and every spectral band. Where is a diagonal matrix that contains the standard deviations for all classes along its major diagonal and is a hyper-parameter, the augmented samples are taken from a multivariate normal distribution with zero mean.

The synthesis of samples in reference [13] made use of both spectral and spatial information, whereas in reference [14], Gaussian filtering was linked with label-based augmentation. Both of these references may be found in the same article.

The second method is based on the idea that HSI pixels that are adjacent to one another should belong to the same class [15]. This is the principle that underpins the second method. As the process moves forward, more examples are generated and added to the training set as the label of a pixel spreads to its neighboring pixels. The use of this approach may result in the samples being wrongly labelled as a direct consequence.

It is possible for the deployment of generative adversarial networks, more commonly referred to as GANs, to create invariance with regard to affine and appearance fluctuations. Both the generator and the discriminator work together to construct a model of an unidentified data distribution by basing it on the samples that are now at their disposal. It is the job of a generator, according to the results of a discriminator, to generate samples of data that are statistically indistinguishable from the data that was initially obtained.

The authors in [16] makes use of GAN conditioning in the process of generating HSI instances to ensure that these instances are an accurate representation of the category that is being sought after in the data. All of the most recent modifications that have been made to the HSI have been made with the intention of achieving the same objective, which is to increase the variety and scope of the training data.

3. PROPOSED DATA AUGMENTATION USING GAN

The GAN generator that has been presented includes several different potential architectures, one of which is depicted in Figure 2. In order for G to function properly, it needs to have a random Gaussian noise vector and a set of training images fed into it as inputs. It is feasible to reduce the dimensionality of a image by using discriminator and convolution layers before combining it with the projected noise vector. This can be done before merging the image with the noise vector (concatenation occurs after a dense layer and non-linearity). We are able to extend the range of the images that can be produced by the generator as a result of its various inputs by adding images from other classes to our own training data class. This allows us to broaden the scope of the images that the generator is capable of producing. This ensures that the generator will deliver results that are as precise as they possibly can be given the parameters that it has been given.

Generating Adversarial Network (GAN), which is a model for a network that has a high-quality generative effect. Goodfellow was the one who came up with the term Generating Adversarial Network.

In order to develop fresh data, it is required to make use of a generator, and the very first thing that has to be done is to feed it noise *Z* that is based on a random distribution. This is the very first step that needs to be taken. In Eq.(1), we can see the formula that will be used for the computations performed by Generator *G*:

$$f(\theta_G,\theta_D)=E_{x\sim pdata}[\log(D(x;\theta_D))]+E_{z\sim N(0,I)}[\log(1-D(G(z;\theta_G);\theta_D))] \quad (1)$$

In the event that the input sample *Z* was faked, the following expressions will describe the mathematical expectation *E*, the Gaussian noise distribution PZ(Z), and the discriminator output probability D(G(Z)): The objective of the training is to achieve G at a level that is as low as is humanly practicable. The discriminator D takes into consideration both the new data produced by the generator as well as the starting data Pdata(x). It then determines which of the two data sets is more likely to contain legitimate information based on the information in both sets. In Eq.(2), the formula for determining how to arrive to the discriminator D is presented as follows:

$$D(x)= p_{data}(x)/(p_{data}(x)+ p_G(x)) \quad (2)$$

We are able to describe the probability of the discriminator output as *D* if we begin with a genuine input sample *X*. It is recommended that the worth of the discriminator D be enhanced in direct proportion to the degree to which the data that are generated are precise. When a GAN is trained, the process of teaching it how to optimize the degree to which the data it creates may trick the discriminator is referred to as training. In the event that the discriminator is unable to distinguish the created data from the original, the authenticity of the generated data is validated. Concurrently, the discriminator adjusts its discriminant ability in order to change its interaction with the generator into that of an adverbial game. The Eq.(3), which may be found below, contains the fundamental theoretical formula, which is as follows:

$$f_K(\theta_G,\theta_D) = \mathrm{f}\ (\theta_G,\ \theta_K^D(\theta_G,\theta_D)) \quad (3)$$

According to Eq.(3), in order to make the discriminator $\theta(D,G)$ as large as possible, it is important to optimize $D(X)$ while concurrently decreasing G. This is required in order to maximize the discriminator $\theta(D,G)$. Therefore, the only element that matters for the purpose of reducing the value of $\theta(D,G)$ for a generator is the second one on the right, which involves the maximization of $D(G(Z))$. This is because this element involves the maximization of $D(G(Z))$.

4. OPERATION USING GAN

In GANs and their derivatives, the objective function is used to determine the degree to which the produced sample distribution deviates from the true distribution in question. This is done by comparing it to the true distribution. GANs have seen a lot of use; however, there have been several issues with their objective functions, such as the gradient disappearing and the model collapsing, which have hampered their efficacy. These issues have been caused by particular problems.

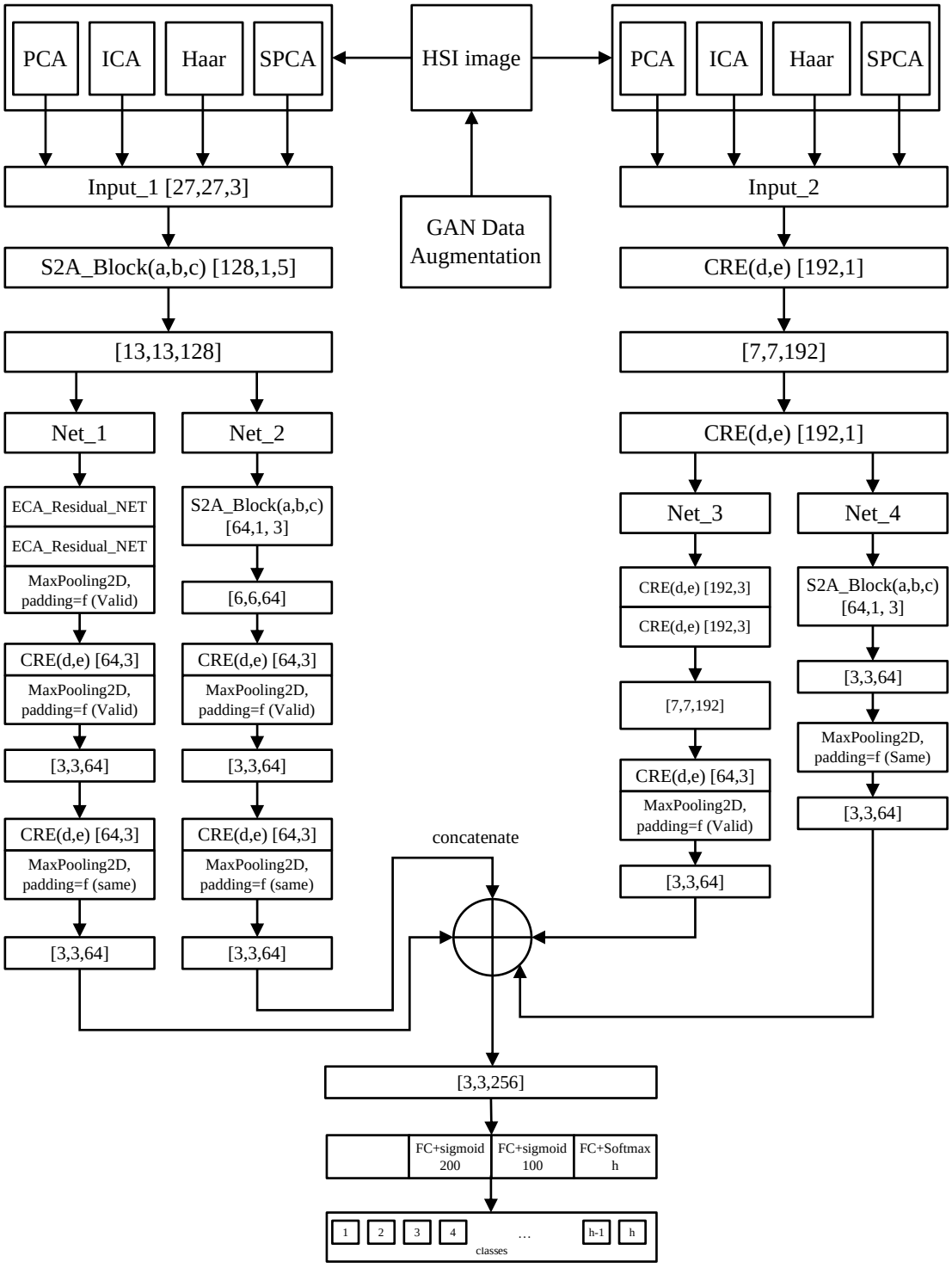

Fig.1. Proposed GAN Classification

Table.1. Classification Accuracy

Class	Dataset	SNR = 5 dB				SNR = 10 dB				SNR = 15 dB			
		ICA	PCA	Haar Wavelet	SuperPCA	ICA	PCA	Haar Wavelet	Super PCA	ICA	PCA	Haar Wavelet	Super PCA
OA (%)	Indian Pines	97.45	98.33	99	99.06	96.68	97.21	98.93	99.61	95.55	96.09	97.8	99.94
AA (%)		98.01	98.77	99.31	99.56	97.43	97.96	98.99	99.55	96.49	97.02	98.37	99.38
Kappax100		97.58	97.56	97.05	96.91	97.7	97.74	97.3	96.97	97.92	97.94	97.52	97.1
OA (%)	KSC	94.06	96.15	97.67	92.43	92.94	94.58	97.22	91.31	91.81	93.46	96.63	94.06

AA (%)	University	94.96	96.48	96.64	93.48	93.98	95.48	96.55	92.36	92.86	94.50	96.25	94.96
Kappax100		89.52	88.76	88.61	89.67	89.78	89.12	88.7	89.99	90.08	89.45	88.87	89.52
OA (%)	Pavia University	97.28	86.87	97.79	87.32	98.61	88.04	99.44	88.77	97.75	86.13	98.26	86.57
AA (%)		94.375	95.268	97.452	99.021	93.608	94.143	95.88	98.571	92.484	93.019	94.758	97.894
Kappax100		94.08	93.99	94.7	94.78	94.31	94.04	94.9	95	94.51	94.19	94.67	96.28

Changing the organizational structure of GANs in any way will not be of any assistance in resolving these issues. Reformulating the objective function as an effective approach of tackling these difficulties has been confirmed to be the case through extensive research and testing. In order to circumvent these deficiencies, this step has been taken. An explanation of numerous objective functions, followed by a discussion of those functions, is provided below.

5. RESULTS

When random noise vectors were launched, the result was the production of three brand-new images for each and every one of the original photos that was fed into the GAN. To create 3225 photos from 1075 standard images, the same process was performed for each image that was not utilized to evaluate the performance of the model. The standard images were used as a basis for the images. After that, these images were utilized in order to compare the outcomes of the model.Two trained generators were deployed in a manner that was comparable to that of a GAN in order to carry out the processing of the training images. The first generation of generators was instructed with the help of typical images, whereas the second generation was instructed with the assistance of images of pneumonia. Pneumonia and images of individuals who were not models were used to create three unique graphics, one for each of the three generators.

The Table.1 and Table.2 provide the results of a classification that was carried out by GAN at Indian Pines University and KSC University, respectively, making use of the dimensionality reduction technique. In each table, the rates of noise grow from their lowest to their greatest values. In addition, Table.1 displays the classification results attained by GAN at Pavia University through the application of the dimensionality reduction technique.

These findings were obtained after gradually increasing the noise level from 0 to 15 over the course of 5 dB intervals. It has been established that there is an inverse relationship between the degree of accuracy for each variable in the simulation results and the availability of samples in each class. This relationship has been shown to have a negative correlation.

According to the findings of the simulation, the proposed method, which is known as SuperPCA, is superior to the typical methods used in the industry, which are known as Haar Wavelet, PCA, and ICA, respectively.

When there is an increase in the quantity of noise that is present, the accuracy of an image categorization diminishes; this is indicated by a commensurate decline in the quality of following noise samples. When the noise level is increased from 10 dB to 15 dB, the classification accuracy rate of SuperPCA is higher than that of Haar Wavelet, PCA, and ICA techniques. SuperPCA also has the highest accuracy rate overall.

6. CONCLUSION

According to the findings, the GAN architecture performs better than existing methods of dimensionality reduction when it comes to demarcating and classifying the various instances or groups. This conclusion was reached after comparing the GAN architecture to other methods of dimensionality reduction. According to the findings, it would appear that other approaches of dimensionality reduction, such as superPCA, are superior to superPCA in terms of accuracy of classification.

REFERENCES

[1] Ian Goodfellow and Yoshua Bengio, “Generative Adversarial Networks”, *Communications of the ACM*, Vol. 63, No. 11, pp. 139-144, 2020.

[2] Z. Xu, Jiawei Luo and Zehao Xiong, “scSemiGAN: A Single-Cell Semi-Supervised Annotation and Dimensionality Reduction Framework based on Generative Adversarial Network”, *Bioinformatics*, Vol. 83, No. 2, pp. 1-14, 2022.

[3] Farajzadeh Zanjani, Maryam, Ehsan Hallaji, Roozbeh Razavi Far and Mehrdad Saif, “Generative Adversarial Dimensionality Reduction for Diagnosing Faults and Attacks in Cyber-Physical Systems”, *Neurocomputing*, Vol. 440, pp. 101-110, 2021.

[4] E. Lin, C. Lin and H.Y. Lane, “Relevant Applications of Generative Adversarial Networks in Drug Design and Discovery: Molecular De Novo Design, Dimensionality Reduction, and De Novo Peptide and Protein Design”, *Molecules*, Vol. 25, No. 14, pp. 3250-3265, 2020.

[5] H. Ding, L. Chen and X. Cui, “Imbalanced Data Classification: A KNN and Generative Adversarial Networks-Based Hybrid Approach for Intrusion Detection”, *Future Generation Computer Systems*, Vol. 131, pp. 240-254, 2022.

[6] S. Chan and A.H. Elsheikh, “Parametrization of Stochastic Inputs using Generative Adversarial Networks with Application in Geology”, *Frontiers in Water*, Vol. 2, pp. 1-5, 2020.

[7] D. Li, C. Du and H. He, “Semi-Supervised Cross-Modal Image Generation with Generative Adversarial Networks”, *Pattern Recognition*, Vol. 100, pp. 107085-107098, 2020.

[8] S. Latif, S., Jurdak and B.W. Schuller, “Augmenting Generative Adversarial Networks for Speech Emotion Recognition”, *Proceedings of International Conference on Recent Trends in Computer Science*, pp. 1-8, 2020.

[9] M. Marouf, P. Machart, V. Bansal, C. Kilian and S. Bonn, “Realistic in Silico Generation and Augmentation of Single-Cell RNA-Seq Data using Generative Adversarial Networks”, *Nature Communications*, Vol. 11, No. 1, pp. 1-12, 2020.

[10] F. Zhu, F. Ye, Y. Fu, Q. Liu and B. Shen, "Electrocardiogram Generation with a Bidirectional LSTM-CNN Generative Adversarial Network", *Scientific Reports*, Vol. 9, No. 1, pp. 1-11, 2019.

[11] J. Viola, Y. Chen and J. Wang, "FaultFace: Deep Convolutional Generative Adversarial Network (DCGAN) based Ball-Bearing Failure Detection Method", *Information Sciences*, Vol. 542, pp. 195-211, 2021.

[12] Z. Zhong, J. Li and A. Wong, "Generative Adversarial Networks and Conditional Random Fields for Hyperspectral Image Classification", *IEEE Transactions on Cybernetics*, Vol. 50, No. 7, pp. 3318-3329, 2019.

[13] S. Surana, P. Arora, D. Singh, D. Sahasrabuddhe and J. Valadi, "Pandoragan: Generating Antiviral Peptides using Generative Adversarial Network", *Proceedings of International Conference on Machine Learning*, pp. 1-7, 2021.

[14] J. Feng, J. Chen, X., Zhang and T. Yu, "Generative Adversarial Networks based on Collaborative Learning and Attention Mechanism for Hyperspectral Image Classification", *Remote Sensing*, Vol. 12, No. 7, pp. 1149-1162, 2020.

[15] H. Nguyen, D. Zhuang, P.Y. Wu and M. Chang, "Autogan-based Dimension Reduction for Privacy Preservation", *Neurocomputing*, Vol. 384, pp. 94-103, 2020.

[16] W. Zhao, J. Chen and Y. Qu, "Sample Generation with Self-Attention Generative Adversarial Adaptation Network (SaGAAN) for Hyperspectral Image Classification", *Remote Sensing*, Vol. 12, No. 5, pp. 843-852, 2020.