Handling Class Imbalance of Radio Frequency Interference in Deep Learning-Based Fast Radio Burst Search Pipelines Using a Deep Convolutional Generative Adversarial Network Postprint

Authors: Wenlong Du, Yanling Liu, Maozheng Chen, Yanling Liu

Date: 2025-03-14T00:00:00+00:00

Abstract

This paper addresses the performance degradation issue in a fast radio burst search pipeline based on deep learning. This issue is caused by the class imbalance of the radio frequency interference samples in the training dataset, and one solution is applied to improve the distribution of the training data by augmenting minority class samples using a deep convolutional generative adversarial network. Experimental results demonstrate that retraining the deep learning model with the newly generated dataset leads to a new fast radio burst classifier, which effectively reduces false positives caused by periodic wide-band impulsive radio frequency interference, thereby enhancing the performance of the search pipeline.

Full Text

Preamble

Astronomical Techniques and Instruments, Vol. 2, January 2025, 10–15

Article

Open Access

Handling class imbalance of radio frequency interference in deep learning-based fast radio burst search pipelines using a deep convolutional generative adversarial network

Wenlong Du[12], Yanling Liu[12]*, Maozheng Chen[123]

[1]Xinjiang Astronomical Observatory, Chinese Academy of Sciences, Urumqi 830011, China

[2]Xinjiang Key Laboratory of Microwave Technology, Urumqi 830011, China
[3]Key Laboratory of Radio Astronomy and Technology, Chinese Academy of Sciences, Beijing 100101, China

*Correspondence: liuyanling@xao.ac.cn

Received: September 18, 2024; Accepted: September 30, 2024; Published Online: October 15, 2024
https://doi.org/10.61977/ati2024053; https://cstr.cn/32083.14.ati2024053

Citation: Du, W. L., Liu, Y. L., Chen, M. Z. 2025. Handling class imbalance of radio frequency interference in deep learning-based fast radio burst search pipelines using a deep convolutional generative adversarial network. *Astronomical Techniques and Instruments*, 2(1): 10−15. https://doi.org/10.61977/ati2024053.

Abstract: This paper addresses the performance degradation issue in a fast radio burst search pipeline based on deep learning. This issue is caused by the class imbalance of the radio frequency interference samples in the training dataset, and one solution is applied to improve the distribution of the training data by augmenting minority class samples using a deep convolutional generative adversarial network. Experimental results demonstrate that retraining the deep learning model with the newly generated dataset leads to a new fast radio burst classifier, which effectively reduces false positives caused by periodic wide-band impulsive radio frequency interference, thereby enhancing the performance of the search pipeline.

Keywords: Fast radio burst; Deep convolutional generative adversarial network; Class imbalance; Radio frequency interference; Deep learning

---

1. INTRODUCTION

Astrophysical millisecond-duration Fast Radio Bursts (FRBs) are extremely energetic, coherent phenomena detectable over cosmological distance scales. The first FRB was detected using Parkes archival data in 2006 by Lorimer et al. [1], but the phenomenon was not officially named until Thornton et al. discovered four more high dispersion measure (DM) FRBs in 2013 [2]. Since then, with increasing enthusiasm for FRB research, over 800 FRBs (including 67 repeating FRBs) have been reported by different radio telescopes all over the world, such as the Canadian Hydrogen Intensity Mapping Experiment (CHIME), the Australian Square Kilometre Array Pathfinder (ASKAP), the Parkes radio telescope, the Molonglo Observatory Synthesis Telescope (UTMOST), the Low-Frequency

Array (LOFAR), and the Meer Karoo Array Telescope (MeerKAT), at frequencies ranging from 110 MHz to 8 GHz, with most of these discoveries occurring in the last 2–3 years [1]. More observational samples are required for deeper analysis.

The search methods for FRBs include two approaches: one based on dedispersion search software and the other on deep learning classifiers. Thanks to the development of computer technology, especially graphics processing units, deep learning has achieved remarkable success in signal classification, pattern recognition, and other areas of data science. Compared with traditional machine learning methods, deep learning does not require experienced experts to spend extensive time on feature engineering, which avoids the incompleteness caused by manual design and selection of features. Moreover, deep learning has far surpassed the capabilities of traditional machine learning, especially when dealing with large datasets. Classifiers based on deep learning have shown great potential and advantages in the search for rare varieties of FRB within vast quantities of astronomical observation data.

Researchers such as Connor & van Leeuwen [3], Zhang et al. [4], and Agarwal et al. [5] have successfully used deep learning techniques to achieve fast, high-precision detection of FRB events, facilitating FRB voltage data dumps and multi-wavelength follow-up observations. The Xinjiang Astronomical Observatory, Chinese Academy of Science (XAO) has detected the bursts FRB 20201124A, FRB 20200428A, and FRB 20220912A using a self-developed FRB search terminal and a dispersed dynamic spectra search (DDSS) pipeline based on deep learning [6, 7]. Practical applications like these have demonstrated that the deep learning-based FRB search method offers significant advantages in both search accuracy and speed.

Whether using traditional dispersion search methods or deep learning-based FRB classifiers, radio frequency interference (RFI) remains a significant challenge. RFI can drown out FRB signals, reduce the signal-to-noise ratio, and cause algorithms to miss detections of astronomical events. At the same time, RFI can also trigger search algorithms to generate false-positive candidates, making post-processing and filtering more difficult. In using deep learning to classify FRBs, the issue of interference class imbalance can significantly affect the training performance of the model. To achieve optimal classifier performance, deep learning models require a large number of training samples to cover all possible distributions of future data. Manual preparation of such a dataset is therefore not feasible.

Because of the insufficient number of existing FRB samples to form a representative training set, we generate positive samples through simulation, while negative samples are primarily drawn from actual observational data without single pulses. However, the complexity of the negative samples is uncontrollable, especially with the presence of RFI. When there is a lack of training samples for one or several types, the deep learning classifier fails to effectively learn the characteristics needed to accurately classify them as negative or positive, lead-

ing to the class imbalance problem. This issue had indeed arisen in our FRB search data processing from the Nanshan 26 m radio telescope (NSRT) at XAO [6].

This paper presents experimental research on the interference class imbalance issue encountered during searches of existing observational data and the significant and beneficial results achieved. In the second section, we introduce the previously encountered interference class imbalance issue and the optional solutions for these problems. The third section focuses on the deep convolutional generative adversarial network (DCGAN) model we used to address the interference class imbalance issue in this study. The fourth section presents our experimental process and results, and the fifth section provides a summary of the research.

2. THE CLASS IMBALANCE PROBLEM AND SOLUTIONS

When searching for FRBs in observational data from March 17, 2022, despite using multi-level RFI mitigation processes [6, 8], some false positives still appeared in the FRB candidate data output by the deep learning-based FRB classifier, as shown in Fig. 1 [Figure 1: see original paper]. These false positives were filled with periodic wide-band impulsive radio frequency interference (pwi-RFI) in their dynamic spectrograms [6]. This type of RFI is difficult to compensate for using current methods, and directly cutting out affected sections would result in significant signal loss. In the FRB observation data, collected over several hours each day for a week, we only encountered this type of RFI in the data from March 17, 2022. The time distribution is sporadic, as shown in Fig. 2 [Figure 2: see original paper]. This kind of interference is also rarely present in the training dataset, making it difficult for the deep learning model to learn to recognize such interference. At present, the origin of this interference remains undetermined.

For addressing the issue of class imbalance in the deep learning model training process, common solutions can generally be divided into two categories [9-11]:

(1) **Model-level solutions**: These include feature selection, cost-sensitive learning, and ensemble learning. By improving feature extraction and training methods, these approaches make the generated model more inclined toward data from minority classes.

(2) **Data-level solutions**: These primarily involve either oversampling the minority class or undersampling the majority class to change the class distribution in the training set, thereby reducing the negative impact of data imbalance on model performance.

To elaborate on the second category, undersampling reduces the number of majority class samples to balance the ratio between majority and minority classes, which can improve the classification accuracy of the model for minority classes.

However, reducing majority class samples may degrade overall model performance. Oversampling, however, increases the number of minority class samples to balance the dataset. This allows the model to better learn the features of the minority class during training without risking feature loss for the majority class. Given the problem we are currently facing, it is evident that, without compromising model performance, the most straightforward and efficient solution is to oversample the minority class at the data level.

The basic version of oversampling is random oversampling, which simply increases the number of minority class samples by randomly duplicating existing ones. This method is simple and effective, but it has the drawback of potentially leading to model overfitting. An improved version, the Synthetic Minority Oversampling Technique (SMOTE), increases the number of minority class samples by generating linear interpolations between existing samples, which reduces the risk of overfitting to some extent but it may introduce noise, sample overlap, and boundary fuzziness. With the rise of deep learning, generative models like generative adversarial networks (GANs) have been used to generate new minority class samples by simulating the distribution of real data, thereby alleviating the problem of class imbalance. The DCGAN is an improved version of the GAN. By incorporating convolutional neural networks (CNNs), the DCGAN replaces fully connected layers with convolutional and transposed convolutional layers, making it more suitable for capturing the spatial features of high-dimensional data. Compared with traditional GANs, the DCGAN generates more realistic images with greater detail. Through the multi-level feature extraction of CNNs, the DCGAN can generate samples similar to the original minority class samples while reducing the risks of overfitting, sample overlap, and boundary fuzziness.

3. THE DCGAN

Generative adversarial networks were first proposed by Goodfellow et al. [12] and consist of two neural networks: a generator and a discriminator. The generator is responsible for creating realistic images from random noise inputs, while the discriminator attempts to distinguish these generated images from real samples in the training set. These two networks compete against each other, continuously adjusting their parameters based on feedback generated by the discriminator. In an ideal scenario, the discriminator becomes capable of easily differentiating real from fake data, while the generator produces data that is nearly indistinguishable from real samples. The DCGAN combines CNNs and GANs by replacing fully connected layers with convolutional and transposed convolutional layers, making them more suitable for processing high-dimensional data.

The basic framework of the DCGAN is shown in Fig. 3 [Figure 3: see original paper]. The generator takes a random vector sampled from the latent space as input, using transposed convolution layers to gradually increase the dimensionality of the data, transforming the small vector into progressively larger feature maps, and finally generating an image the same size as the target image.

The discriminator, however, uses convolutional layers to progressively reduce the input image size, extracting features from the image. It then outputs a probability indicating whether the image is a real sample or a generated one. Except for the generator' s output layer and the discriminator' s input layer, all convolutional layers are followed by a batch normalization layer, which scales the output of the previous layer to have zero mean and unit variance, helping with smoother gradient updates and making the training process more stable. LeakyReLU activation functions are used in all convolutional layers except the output layer, which help prevent the vanishing gradient problem, allowing for better training of deep networks. The generator' s output layer uses a Tanh activation function to map the pixel values of the generated image to the range of $(-1, 1)$, while the discriminator' s output layer employs a Sigmoid activation function to map the classification result to the range of $(0, 1)$, representing the probability that the input is real. This result is fed back to both networks to update their parameters accordingly.

4. EXPERIMENTS AND RESULTS

We use the DCGAN to generate additional pwi-RFI samples, which are added to the training dataset of the DDSS pipeline' s deep learning model. By retraining the Xception model, we enable the FRB classifier to better recognize pwi-RFI. The number of detected false positives is related to the deep learning model and the setting of the confidence threshold. To obtain more pwi-RFI samples for conducting experiments, we use the DDSS pipeline with the Xception model, trained on simulated FRBs in the DM range of 100–500 pc cm^{-3}, with the confidence threshold set to 0.5 to search for FRB observational data from March 17, 2022. Then all observation data packets containing pwi-RFI are divided into two groups based on the sample quantity and time distribution. We ensure that the number of pwi-RFI samples and their time distribution are as balanced as possible in each group, to avoid model bias due to data imbalance. Each group contains 20 data packets with varying numbers of pwi-RFI samples. The pwi-RFI samples of one group are used for sample augmentation, while the other group' s data are used to experimentally validate the optimized model classifier.

4.1 Generating Minority Class Samples to Extend the Dataset

To ensure the quality of generated samples, the real samples used for the DCGAN training are best sourced from original observational data. Here, we extract a total of 82 pwi-RFI samples from the observation data group designated for the DCGAN training. Given the limited sample size, we first apply data augmentation techniques, such as data concatenation and the addition of low-intensity Gaussian noise, to expand the dataset. Additionally, this helps prevent model overfitting and enhances feature learning capability. Data concatenation involves temporally cropping the original observation data in a 1:1 and 1:3 ratio, then combining the cropped data pairwise to recreate samples of the original size. This method produces 6,642 new samples in total. For the Gaussian noise

augmentation, we set 20 different parameter groups with a mean of 0 and standard deviations ranging from 0.05 to 0.25, generating 168 samples per group, yielding 3,306 in total.

Finally, all the samples of real pwi-RFI are used to train the DCGAN, and the generated pwi-RFI samples are shown in Fig. 4 [Figure 4: see original paper].

A total of 6,560 pwi-RFI samples are generated using the DCGAN, and after 16× downsampling, the newly generated samples are added to the original dataset. The composition of the expanded dataset is shown in Fig. 5 [Figure 5: see original paper]. The total number of samples in the dataset is 158,240, with a positive-to-negative ratio of 92:100, and the minority class of pwi-RFI accounts for 4.15%.

4.2 Training the New Model

The Xception model is retrained using the new expanded dataset, with a 9:1 split between the training set and the test set. Typically, after more than six training iterations, the network model achieves optimal performance, with both training and validation errors converging, showing only minor differences between the two. Fig. 6 [Figure 6: see original paper] illustrates the changes in accuracy and loss function during one of the model training sessions.

4.3 Comparison of Classification Performance Between the New and Old Models

The original and new Xception classifiers are retrained with the new dataset and used to search for FRBs in the validation data packet from March 17, 2022. The classification results of the two models are shown in Table 1 .

The limited number of pwi-RFI samples used to train the DCGAN come with some concerns that this could lead to overfitting in the newly generated data. However, the classification results reveal very few false positives caused by the pwi-RFI. Furthermore, the number of other types of false positives has also decreased. This could be due to certain features in those false positives resembling the pwi-RFI signals, especially considering that the selection of pwi-RFI samples is based on manual selection of downsampled images. The newly retrained model' s classification performance is fairly effective, considering that the minority class data used for training and validation in the experiment are all acquired in a single day. There is still room for improvement, however, and the generalization ability of this method still requires further verification in the future.

5. CONCLUSIONS

Class imbalance is a common issue in deep learning. This paper addresses the performance degradation problem in FRB deep learning classifiers caused by pwi-RFI. Using the DCGAN, we augment the minority class samples in

the training dataset and retrain the model. Experimental results show that the new FRB classifier has gained the ability to identify pwi-RFI, effectively resolving the performance degradation caused by class imbalance. However, because all the pwi-RFI training and validation data are acquired from a single day of observations, the generalization ability of this approach and its optimal configuration in practical applications will require further research in the future.

ACKNOWLEDGMENTS

We thank the referee for their valuable comments, which improved the presentation of the paper. This work is supported by the Chinese Academy of Science "Light of West China" Program (2022-XBQNXZ-015), the National Natural Science Foundation of China (11903071), and the Operation, Maintenance and Upgrading Fund for Astronomical Telescopes and Facility Instruments, budgeted from the Ministry of Finance of China and administered by the Chinese Academy of Sciences.

AUTHOR CONTRIBUTION

Wenlong Du and Yanling Liu wrote and revised the paper. Wenlong Du was responsible for programming, algorithm implementation, and data analysis. Yanling Liu contributed to the design and development of experimental methods. Maozheng Chen supervised the project and reviewed the paper. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

Lorimer, D. R., Bailes, M., McLaughlin, M. A., et al. 2007. A bright millisecond radio burst of extragalactic origin. *Science*, 318(5851): 777−780.

Thornton, D., Stappers, B., Bailes, M., et al. 2013. A population of fast radio bursts at cosmological distances. *Science*, 341(6141): 53−56.

Connor, L., van Leeuwen, J. 2018. Applying deep learning to fast radio burst classification. *The Astronomical Journal*, 156(6): 256.

Zhang Y G, Gajjar V, Foster G, et al. 2018. Fast radio burst 121102 pulse detection and periodicity: a machine learning approach. *The Astrophysical Journal*, 866(2): 149.

Agarwal, D., Lorimer, D. R., Surnis, M. P., et al. 2020. Initial results from a real-time FRB search with the GBT. *Monthly Notices of the Royal Astronomical Society*, 497(1): 352−360.

Liu, Y. L., Li, J., Liu, Z. Y., et al. 2022. A search technique based on deep learning for fast radio bursts and initial results for FRB 20201124A with the NSRT. *Research in Astronomy and Astrophysics*, 22(10): 105007.

Liu, Y. L., Chen, M. Z., Li, J., et al. 2024. Design and application of an S-band Fast Radio Burst search pipeline for the Nanshan 26 m radio telescope. *Research in Astronomy and Astrophysics*, 24(7): 075008.

Liu, Y. L., Chen, M. Z., Yuan, J. P., et al. 2024. Radio frequency interference mitigation methods for fast radio burst observation data. *Acta Astronomica Sinica*, 65(2): 123−133. (in Chinese)

Buda, M., Maki, A., Mazurowski, M. A. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106: 249−259.

Li, A. Z. 2024. Research on generative adversarial networks for class imbalasnce in image classification. Xi' an: Xi Dian University. (in Chinese)

Mahmoud, P. 2021. Learning from imbalanced pulsar data by combine DCGAN and PILAE algorithm. *New Astronomy*, 85: 101561.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al. 2014. Generative adversarial nets. In *Proceedings of Advances in Neural Information Processing Systems* 27.

Note: Figure translations are in progress. See original paper for figures.