Generative Adversarial Networks for Galaxy Classification Using Convolutional Neural Networks

Polina Petrova
ppetrova@umass.edu

Aryan Singh
arysingh@umass.edu

Abstract

The classification of galaxies poses significant challenges due to inherent data imbalances among various morphological classes. This research addresses the pressing issue of accurately classifying galaxies using Convolutional Neural Networks (CNNs), where traditional approaches often struggle with underrepresented categories. To mitigate this problem, we propose a novel combined model utilising a Conditional Generative Adversarial Network (cGAN) to generate synthetic images that augment the Galaxy Zoo 2 dataset, which contains approximately 250,000 labelled galaxy images. Our methodology begins with enhancing image resolution using Super Resolution GANs (SRGAN), followed by training the cGAN to produce additional samples for each class, thereby addressing class imbalance without resorting to deeper networks. We draw on foundational works, including studies on deep generative models for galaxy image simulations and prior CNN applications in galaxy classification, to inform our approach. Our model's performance will be rigorously evaluated using metrics such as accuracy, precision, recall, and F-1 score, comparing results from training on both real and synthetic data against traditional CNNs trained on the original dataset. Through this work, we aim to contribute to the ongoing efforts to improve automated galaxy classification and provide a scalable solution to the prevalent issue of data imbalance in astronomical research.

1. Introduction

With rapid development in technology comes a phenomenon that is referred to as the *data deluge*, – a consequence of an overwhelming influx of data and insufficient resources to process it. Astronomers, in particular, contend with this challenge; estimates suggest the observable universe holds between 100 billion and 200 billion galaxies [3]. A single photograph of a small sky section can capture up to 25,000 galaxies, and the resulting daily data volume overwhelms the limited pool of experts available to classify them [9]. To address this challenge, astrophysicists launched the *Galaxy Zoo* project in 2007, inviting citizen scientists to help classify over 900,000 galaxies, marking a transformative moment in data processing through public participation [8].

Following the success of the *Galaxy Zoo*, advancements in machine learning spurred efforts to automate galaxy classification. Modern approaches employ Convolutional Neural Networks (CNNs) to recognise patterns with minimal human input. However, a persistent challenge is the quality and distribution of available data. While the *Galaxy Zoo* project provided a substantial dataset, imbalances in class representation can skew training, causing models to favour more frequent classes. Our research specifically addresses this class imbalance in the Galaxy Zoo 2 dataset – a collection of categorised images taken from the Sloan Digital Sky Survey (SDSS) – where certain galaxy types are underrepresented. Building deeper networks is a common workaround to address this issue, but we argue that this approach only sidesteps the core problem.

Ideally, a large, balanced dataset would improve model accuracy, but limitations in space imaging and classification complexity make this difficult. To tackle this, we propose a novel solution using Generative Adversarial Networks (GANs) to generate synthetic images for underrepresented galaxy classes in the Galaxy Zoo 2 dataset, which we then use to train a CNN. We expect that by augmenting our data with synthetic images, we can improve classification accuracy across all classes. Our evaluation will compare the GAN-CNN model's performance against traditional CNNs trained on imbalanced data, particularly examining accuracy gains in classifying underrepresented classes. With this combined GAN-CNN model, we aim to address class imbalance directly, eliminating the need for deeper networks as a compensatory measure.

2. Related work

Lahav et al. [5] offers one of the first discussions of using neural networks in the galaxy classification problem. The study clarifies the role of Artificial Neural Networks (ANNs) in galaxy classification by demonstrating their ability to replicate human classification using ESO-LV galaxy data. ANNs achieve comparable accuracy to human experts, operating within 2 T-type units. The authors argue that ANNs provide a robust statistical framework, improving on linear methods through their capacity for non-linear modelling. While the paper does not cover all classification methods, it emphasises the potential of unsupervised algorithms to discover new features in galaxy data without external guidance. It also highlights the importance of integrating dynamic properties and multiwavelength data to enhance the classification process, as we now have in the Galaxy Zoo 2 dataset. This study lays the groundwork for our exploration into data-driven galaxy classification.

Fussell and Moews [2] demonstrate the effectiveness of using GANs to augment limited datasets of galaxy images. The authors find that the original DCGAN architecture can generate realistic galaxy images that align closely with real galaxy data when statistically evaluated. To achieve higher-resolution synthetic galaxies, they introduce a chained approach using StackGAN as a second stage, which overcomes DCGAN's limitations at higher resolutions. By evaluating physical property distributions of the generated galaxies and confirming their similarity to real data, the study suggests these synthetic images can effectively expand real galaxy datasets. This augmentation is beneficial for various tasks, including galaxy classification, segmentation, deblending, and calibration of shape measurement algorithms. Ultimately, this research highlights GAN architectures as valuable resources for astronomy, providing scalable data for deep learning models that require extensive training samples. We will adopt a similar approach, benchmarking our GAN-generated data against real data in our model pipeline before feeding them into our CNN.

Kim and Brunner [4] address the limitations and potential improvements for using CNNs in galaxy classification, emphasising concerns about overfitting due to limited training data. The authors note that while CNNs have shown promise, their model did not significantly outperform traditional machine learning models relying on summary catalog data, likely due to data constraints. Collecting additional spectroscopic training images could mitigate overfitting and enhance CNN performance. However, they argued that the process is costly and time-intensive. For future work, the authors suggest training multiple network architectures and combining them, a strategy that has proven effective in other galaxy classification challenges. Integrating CNN with other classifiers in a hybrid model could also yield improvements, as demonstrated in past studies where blending diverse classification approaches outperformed any single method. We will directly address these concerns in our research, providing a hybrid model that will even out and expand our training data to mitigate overfitting.

Walmsley et al. [1] focus on the limitations of deep learning for galaxy morphology, which often overlook uncertainty in labelling. They introduce a Bayesian CNN model to capture probabilistic predictions for galaxy morphology, leveraging sparse Galaxy Zoo labels. Using Monte Carlo Dropout and active learning, their model selects informative galaxies for labelling, enhancing classification accuracy with fewer labels. This approach, essential for large-scale surveys, offers insights into morphology and astrophysical connections. We aim to incorporate similar probabilistic and active learning strategies in future iterations of our model for effective scaling.

3. Method

Describe the methods you intend to apply to solve the given problem.

4. Results

State and evaluate your results upto the milestone.

5. Conclusion

State your conclusions upto the milestone. This section might be empty for now but your final report should contain the conclusions of your project.

References

[1] M. Walmsley et al. Galaxy zoo: probabilistic morphology through bayesian cnns and active learning. *Monthly Notices of the Royal Astronomical Society*, 491(2):1554–1574, 2019. 2

[2] L. Fussell and B. Moews. Forging new worlds: high-resolution synthetic galaxies with chained generative adversarial networks. *Monthly Notices of the Royal Astronomical Society*, 485(3):3203–3214, 2019. 2

[3] E. Howell. How many galaxies are there?, 2018. Accessed: Oct. 30, 2024. 1

[4] E. J. Kim and R. J. Brunner. Star–galaxy classification using deep convolutional neural networks. *Monthly Notices of the Royal Astronomical Society*, 464(4):4463–4475, 2016. 2

[5] O. Lahav, A. Nairn, L. Sodré, and M. C. Storrie-Lombardi. Neural computation as a tool for galaxy classification: methods and examples. *Monthly Notices of the Royal Astronomical Society*, 283(1):207–221, 1996. 1

[6] F. Lanusse, R. Mandelbaum, S. Ravanbakhsh, C.-L. Li, P. Freeman, and B. Póczos. Deep generative models for galaxy image simulations. *Monthly Notices of the Royal Astronomical Society*, 504(4):5543–5555, 2021.

[7] V. Lukic and M. Brüggen. Galaxy classifications with deep learning. *Proceedings of the International Astronomical Union*, 12(S325):217–220, 2016.

[8] C. McGourty. Scientists seek galaxy hunt help, 2007. Accessed: Oct. 30, 2024. 1

[9] E. Sauers. Webb telescope reveals more galaxies in a snapshot than hubble's deepest survey, 2023. Accessed: Oct. 30, 2024. 1

[10] K. Sharma, A. Kembhavi, T. Sivarani, S. Abraham, and K. Vaghmare. Application of convolutional neural networks for stellar spectral classification. *Monthly Notices of the Royal Astronomical Society*, 491(2):2280–2300, 2019.