Generative Adversarial Networks for Galaxy Classification Using Convolutional Neural Networks

Aryan Singh
University of Massachusetts, Amherst
arysingh@umass.edu

Polina Petrova
University of Massachusetts, Amherst
ppetrova@umass.edu

October 3, 2024

1 Proposal

Please follow the steps outlined below when submitting your manuscript to the IEEE Computer Society Press. This style guide now has several important modifications (for example, you are no longer warned against the use of sticky tape to attach your artwork to the paper), so all authors should read this new version.

1.1 Group Members

Our group consists of Aryan Singh and Polina Petrova. Aryan will be working on implementing both the GANs for the project. This includes training a Super Resolution Generative Adversarial Network (SRGAN) to improve the image quality in the dataset, and implementing a conditional Generative Adversarial Network (cGAN) to create more new images for each galaxy class. This includes finding appropriate values for layer size, learning rate, selecting regularization parameters, and choosing an appropriate loss function. Polina will be responsible for implementing the Convolutional Neural Network (CNN) for galaxy classification. This includes designing the architecture, selecting appropriate layers and configuring activation functions tailored to galaxy image data. Polina will also manage the training process, including correcting for overfitting. Each group member will complete hyperparameter tuning on their respective components of the model to achieve optimal accuracy. This entails adjusting parameters such as batch size, learning rates and the number of layers for both the GAN and CNN, ensuring balanced tradeoffs between model complexity and performance. Together, we will complete a high-level analysis of our GAN-CNN pipeline based on selected evaluation criteria and observe the effectiveness of classifying test instances on a classical cGAN compared to a SR-cGAN.

1.2 Motivation

We are addressing the problem of accurately classifying galaxies using CNNs when there is an imbalance of data with the classes. In such a situation a cGAN can help generate data. This problem is interesting because data imbalance is a real word problem which is encountered often, and cGANS can be used to mitigate this issue.

1.3 Literature Review

We will examine several key papers to provide context and background for our research. Works such as *Deep generative models for galaxy image simulations* [1] and *Forging new worlds: high-resolution synthetic galaxies with chained generative adversarial networks* [2] will provide a foundational understanding of the use of Generative Adversarial Networks as applied to galaxy morphology classification problems. Additionally, we will explore papers such as *Galaxy classification: a deep learning approach for classifying*

Sloan Digital Sky Survey images [3], *Star-galaxy classification using deep convolutional neural networks* [4] and *Neural computation as a tool for galaxy classification: methods and examples* [8], which will inform our approach to constructing a functional Convolutional Neural Network and integrate it with our implementation of GAN.

1.4 Data

We plan to use the Galaxy Zoo 2 dataset, which is publicly available on Kaggle. The dataset includes 250k images labeled by volunteers based on characteristics like shape, size, smoothness. If required, we plan to use Google Colab to accelerate the training process, particularly for GANs as they require significant computational power.

1.5 Approach

We propose a combined GAN-CNN model aimed at improving galaxy classification by augmenting training data with synthetic images. The goal is to address the challenge of limited labeled data, particularly for galaxy datasets, eliminating the need for large deep networks as a solution to this data limitation. Our approach involves a two-step process: first, generating synthetic galaxy images using a cGAN and then using these generated images along with existing data to train a CNN for galaxy classification. Firstly, we use a SRGAN to improve the image resolution, then we train the cGAN on these enhanced images. The cGAN will be used to produce additional images for each galaxy class. Afterwards, we will build the CNN using a feed-forward architecture, which has been proven effective in image classification tasks. We will modify the architecture as needed, depending on its performance on the galaxy data, ensuring that the CNN is robust and capable of learning from both types of images effectively. Initially, the CNN will be trained on real galaxy data alone to provide a baseline for the model performance, then on both real and synthetic images generated by the cGAN. The addition of these synthetic images will help balance the dataset and improve the network's ability to classify underrepresented galaxy classes. Lastly, we will train the network using real and synthetic images generated by a progressive GAN. Our model will be evaluated based on its performance in classifying test instances with real available data alone, then on cGAN-generated data and SRGAN images.

1.6 Evaluation Metrics

To evaluate the effectiveness of our model, we will use standard quantitative metrics such as accuracy, precision, recall and F-1 score, which we will use to fine-tune the architecture of our model. These metrics will be plotted as a visual representation of the GAN-CNN performance for each class label. This plot will represent the performance of the CNN trained on the original dataset versus the augmented dataset. We may also use statistical significance tests, such as t-tests, to ensure that performance improvements are not due to random chance. These evaluations will be replicated using cGAN and resolution progressive GAN, providing a statistical comparison between the two models.