Kaggle Galaxy Zoo Image Data Clustering using Deep Embedding Clustering

Anonymous Authors[1]

Abstract

Galaxies are one of the most fundamental entity of the universe. They come in all shapes, sizes and colors and in order to understand how the different shapes (or morphologies) of these galaxies relate to the physics that create them it is important to group similar galaxies based on their structure. Considering the number of galaxy images collected through numerous telescopes this project tries to evaluate the performance of a clustering algorithm: **"Deep Embedding Clustering"** on the Kaggle Galaxy Zoo data. The performance of the algorithm is evaluated based on the separation of *"Elliptical"* and *"Spiral"* galaxy images into distinct clusters.

1. Introduction

Understanding how and why we are here is one of the fundamental questions for the human race. In the quest of understand our origin we as human race also realized how important it is to understand the origin of our universe and all of its entities. According to many scientists galaxies, the most fundamental entity of the universe, might just hold the answer about the origin and evolution of the universe. Galaxies are of very different forms, size and shapes which has led scienties to believe that the inherent structure of galaxies might hold the clue to understanding the laws of physics that govern our universe. This quest for answers is what led to space exploration and development of thousands of telescopes to capture snapshots of galaxies in an effort to understand the structure of the galaxies and how it is critical to the origin and evolution of the universe itself. However in each passing moment, telescopes are capturing more and more images of galaxies far far away. Few decades ago the scientific community depended on massive citizen science crowdsourcing projects ((INSERT CITATION)) to help identofy the type of galaxies in each of the images captured. However because of the recent explosion in the number of images around million or even billions of images, it is virtually impossible for the community to depend on the citizen science projects to label each image individually.

To combat this problem the community pivoted to well establised Supervised Deep Learning techniques using Convolutional Neural Networks (CNN) for classifying each galaxy image and help in labelling the huge dataset. However, supervised learning deep learning techniqes are very much like a black box that takes in an input and returns an output, which in this case is just an image and a class label respectivesl. Using these models thus does not help understand what features define these galaxies and how do the inherent structure help in identifying the type of galaxy the image belongs to. Also the perfomance of some of these models decrease when tested on new images being captured. Because of this problem the community has to contnuously train these models which require volunteer labels for new images and thus leads to the same problem.

To address this problem, **Zooniverse** ((INSERT CITATION)) another citizen science crowdsourcing project has developed techniques which rely on clusters of images to efficiently collect labels for many images in much less time. Also using a clustering based approach helps in understanding what features of the galaxy are most influential and how do the structure relate to the galaxy type. Apart from that using clustering technique helps in extracting information about the most representative structure of the galaxies along with information about the similarity among galaxies of the same type. This experiment is part of their project to develop a more sophisticated technique which is able to cluster similar type of galaxy images together. In this experiment we evaluate the performance of the proposed architecture based on using a Deep Learning based clustering technique: **"Deep Embedding Clustering"** over the publicly available Kaggle Galaxy-Zoo dataset. For this experiment we are just evaluating the performance of the architecture on 2 classes of the galaxies which are: "Elliptical" and "Spiral". The evaluation metric for this experiment is the accuracy of the cluster assignments of each galaxy image where the cluster assignments are calculated based upon the dominant class

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. **AUTHORERR: Missing \icmlcorrespondingauthor.**

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

label in the various clusters formed.

2. Related Work

2.1. Deep Embedding Clustering

2.2. Galaxy Zoo Data Classification - Kaggle

2.3. SDSS and DEC Galaxy Zoo Data Classification

Khan Asad ‘et al.’

3. Approach

4. Experiment

5. Results

6. Analysis

7. Conclusion

A. Appendix