Comparing the Effectiveness of Deep Learning Models on the Categorisation of Galaxy Morphology
CMP3753M Project Proposal

Luke Roberts

University of Lincoln
25722923@students.lincoln.ac.uk

1 Introduction

Introduction

Understanding the universe has been a pursuit of humanity for thousands of years. Because of the increasing effectiveness of deep learning image classification models, categorising distant objects in space has become much easier. Therefore, this rapid development in deep learning will allow astrophysicists to further their research into the early universe.

Due to advances in observational technology such as the Hubble Space Telescope (HST) and more recently the James Webb Space Telescope (JWST) [], we can observe galaxies billions of lightyears from earth. The time taken for the light of distant galaxies to reach us means that our night sky is a window into the past, which allows astrophysicists to understand the evolution of our universe in more depth []. Photographs such as the Hubble Ultra-Deep field show us an ancient universe full of developing galaxies [], and the amount of observed galaxies in astronomy databases is only increasing.

By identifying how the structure of galaxies, or galaxy morphology, changes over time, we have a clearer picture of the changes of galactic structure over the last several billion years. However, it is extremely time-consuming for scientists to categorise galaxies in their research. A study in 2016 [] calculated that there are 2.0×10^{12} galaxies in the observable universe. While it would be unfeasible to categorise them all, it would be much more efficient to automate the process. One powerful method for automation of images is to train and deploy a deep learning model.

Deep Learning has recently revolutionised both scientific research and modern life in a profound way []. From the categorisation of X-ray images in the medical field; machine translation of natural language such as Google Translate and large large language models such as ChatGPT, deep learning has become the best way to categorise, analyse and generate unstructured data []. To perform many of these tasks, machine learning engineers create deep learning models which utilise a dataset that learns patterns about that data.

Image classification is a form of deep learning that uses images as input data and is used for a wide variety of applications in scientific research []. Like all forms of deep learning, image processing models are trained so that they more accurately categorise new input data []. At the start of training, the model performs poorly when tested. However, through analysing how the output fails, the model can tweak its own parameters in order to achieve greater accuracy []. The choice of which machine learning model or architecture to use and the specific hyperparameters are important for maximising the efficiency of the model []. This can be due to the content and quality of the training data being used.

Recent studies such as 'Galaxy classification: deep learning on the OTELO and COSMOS databases' [] concluded that using a DNN "outperforms" other adopted models for galaxy classification. However,

to extend this research I will evaluate which deep learning neural network is most effective at classifying the morphology of distant galaxies by implementing deep learning architectures and evaluating their computational efficiency and performance in categorisation.

2 Aims and Objectives

2.1 Aim

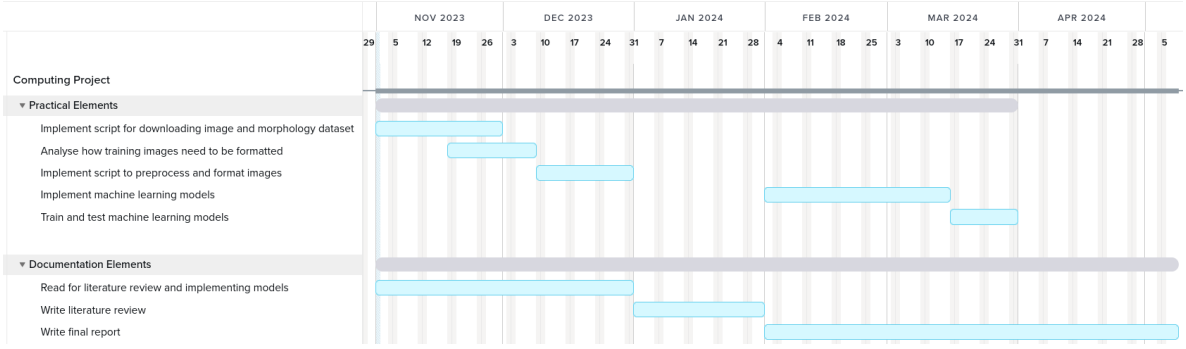
One of the most widely adopted deep learning image classification is the convolutional neural network (CNN) [] which has been used in many studies to categorise a wide range of image types. New deep learning image classification architectures have been developed which have either greater accuracy or are faster at categorising data. ConvXGB has been shown to improve on CNNs [Tables 5a-c] for different datasets. The aim of this project is to determine whether an implementation of traditional CNN model or ConvXGB is more suitable to classify deep space galaxy morphology from existing databases of galaxy images. This will be training the models and calculating ML performance heuristics.

2.2 Objectives

1. Gather a dataset of galaxy images, coupled with their morphological type. I will use an official astronomy database like the ESA astronomy archives [] using the astroquery library [].
2. Implement a method of preprocessing images for training in python throgh an image procseeing library such as PIL [].
3. Implement a traditional CNN model and ConvXGB based model for predicting the categorisation of images of galaxies. The real type of morphology associated with that image will be used in backpropergation the of the model to train it. This will be done using the python libraries Sci-kit Learn [], TensorFlow [] and XGBoost []. GPU parallelisation will be needed to train the dataset much more quickly.
4. Split the dataset into training and testing. Train the models with the training dataset gathered. Once the models have been successfully trained, the models will be tested using the testing dataset and ML performance heuristics will be calculated. The performance heuristics will determine which model is better at categorising galaxy morphology.

3 Project Plan and Risk Analysis

3.1 Project Plan



Gant Chart of project

3.2 Risk Analysis

References

1. Castelvecchi, D. (2016). Universe has ten times more galaxies than researchers thought. Nature. https://doi.org/https://doi.org/10.1038/nature.2016.20809.