NCRA-TIFR Project Proposal

Archit Sakhadeo
Rathin Desai
Shadab Shaikh
Shubhankar Deshpande

Mentor
Dr. Yogesh Wadadekar
Dr. C. H. Ishwar Chandra

1 Introduction

1.1 Morphological Classes of Radio Galaxies

Radio galaxies with active nuclei can be distinguished based on their radio luminosity or brightness of their radio emissions in relation to their hosting environment. Some of the basic morphological classifications include point sources, extended sources i.e. sources with extended contours, double radio sources, jets, and lobes.

1.2 Problems faced with current classification

Currently Radio astronomers manually classify galaxies based on visual inspection of the images which is a slow procedure, and increases the time to production of scientific results. Further, it introduces uncertainities in the classification procedure, both of which are problems which can potentially be mitigated by using an automated approach.

Contemporary algorithms classify radio sources into at most three different classes. Our aim is to build a robust model capable of handling more than 2 classes.

2 Objective

- Potentially discovering rare forms of radio sources by classification in different classes.
- Reduction in time to generate scientific results by radio astronomers.
- Deeper insight into topological representation of radio data during classification.

3 Approach

3.1 Source Modelling

The first step would be source extraction using the standard technique of gaussian modelling. We propose to do this using the robust PyBDSM pipeline used for fitting gaussian distributions to radio sources. The software contains a plethora of features, from which we would be using a small subset. This would mainly include:

1. Source extraction using gaussian modelling of radio data.
2. Generation of a catalog file containing details of radio sources (RA, DEC, Size of Gaussian (min, max), etc.)

3.2 Cutout Generation

The second step would be to convert the RA(Right Ascension) and DEC (Declination) values generated from the catalog, to their corresponding pixel values in the original image. Based on these pixel values we generate 10*10 px cutouts using as reference the co-ordinates of the center of the radio source. This involves a multistep procedure briefly including:

1. Reading the FITS image in the form of a matrix
2. Parsing through the generated catalog file and extracting data for each radio source such as RA, DEC, etc.
3. Converting the RA and DEC values from WCS (World Coordinate System) to pixel values.
4. Processing pixel values to account for difference in addressing between FORTRAN and C family of languages.
5. Slicing the image matrix assuming the reference pixel co-ordinates as the center of the source.

Prototype code for section *3.1* and section *3.2* has been written mainly for testing purposes. We used a sample image from the TGSS survey which was then processed using the first two steps of our pipeline to generate 470 cutout images. More details can be found at: https://github.com/NCRA-TIFR/radiogen.

3.3 Data Preprocessing

Real world data are incomplete, inconsistent and noisy. Techniques like Data cleaning, integration, transformation, reduction, discretization are required to structure the data uniformly throughout the dataset in the required format.

Some regular techniques are:

- Mean subtraction

  It involves subtracting the mean across every individual feature in the data, and has the geometric interpretation of centering the cloud of data around the origin along every dimension.

- Normalization

  Normalization refers to normalizing the data dimensions so that they are of approximately the same scale. There are two common ways of achieving this normalization. One is to divide each dimension by its standard deviation, once it has been zero-centered. Another form of this preprocessing normalizes each dimension so that the min and max along the dimension is -1 and 1 respectively.

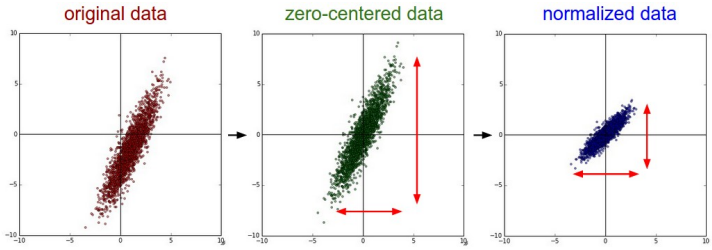

Figure 1: Mean Subtraction and Normalization

- Dimensionality reduction using Principal Component Analysis

  The main linear technique for dimensionality reduction, principal component analysis, performs a linear mapping of the data to a lower-dimensional space in such a way that the variance of the data in the low-dimensional representation is maximized. By finding the eigen vector with the highest eigen value we select the Principal component axis with the highest variance and the minimum reconstruction error. Thus we reduce the dimensions by eliminating Principal component axes with the eigen vectors with the least variance thereby not losing much information.

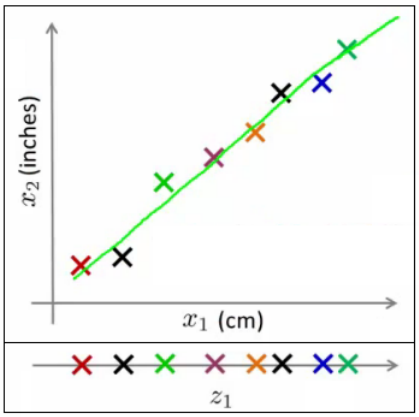

Figure 2: Reduction from 2D to 1D using PCA

3.4 Analytical Approach

Some approaches that we plan to use:

- Scale Invariant Feature Transform (SIFT

  Statistical modelling of data to manually extract features. For any object in an image, interesting points on the object can be extracted to provide a "feature description" of the object. This description, extracted from a training image, can then be used to identify the object when attempting to locate the object in a test image containing many other objects. To perform reliable recognition, it is important that the features extracted from the training image be detectable even under changes in image scale, noise and illumination. SIFT is invariant of scaling, transformation and rotation.

- Edge Detection

  1) Reduction of noise using a Gaussian filter 2) Finding gradients along the X and Y directions in the image 3) Suppression of local minima 4) Double Thresholing to determine probable edges 5) Suppression of all weak edges which lack connection

Figure 3: Edge detection

- Image Segmentation

  Image Segmentation is the process of partitioning a digital image into multiple segments. The goal of segmentation is to simplify and change the representation of an image into something that is more meaningful and easier to analyze. It is typically used to locate objects and boundaries (lines, curves, etc.) in images. More precisely, image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain characteristics.

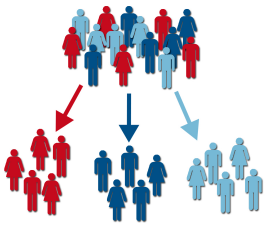

Figure 4: Segmentation of data into distinct characteristics

3.5 Empirical Approach

Empirical approach is a way of gaining knowledge by means of direct and indirect observation or experience. In the context of this project, we plan to employ Artificial Neural Networks(ANNs), which are a computational model based on large collection of simple artifical neurons, that learns hierarchical representations of the observational data. Each neural unit is connected with many other units which computes using

summation function. Neural networks typically consist of multiple layers in which a signal traverses from the first(input) to the last (output) layer of neural units.

Convolutional neural networks(CNNs), a particular type of ANN, currently provide the best solutions to many problems in computer vision such as image segmentation, recognition and classification[1].

Convolutional neural networks were designed to use minimal amounts of preprocessing. A typical architecture of a CNN consists of convolutional layers, activation layers and pooling layers. Convolutional layer performs the convolution operation on previous layers, activation layer defines the output of that node given an input or set of inputs, and the pooling layer serves to progressively reduce the spatial size of the representation, to reduce the number of parameters and amount of computation in the network.

The initial layers of the CNNs learn simple features in the input data such as straight edges, simple colors, and curves. The deeper layers learn the higher level representation of image and search for complex shapes and structures. Mathematically deeper layers can be thought of as compositions of previous layers. [1][2][3].

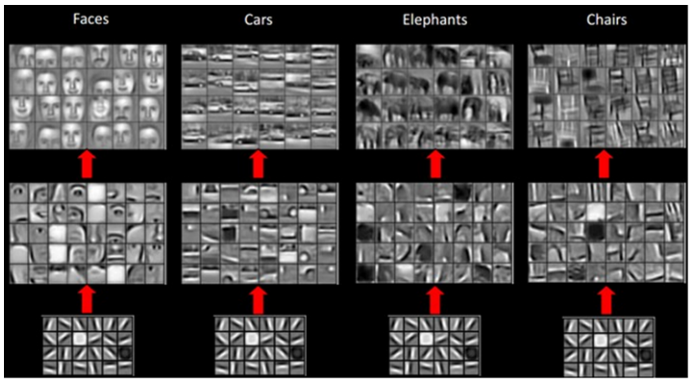

Figure 5: CNN learning simple features on initial layers for different classes

Variants of CNNs have achieved great successes for morphological galaxy classification and prediction on SDSS[4]. Sander Dieleman et al. won the Galaxy challenge, an international competition to build the best model for morphology classication based on annotated images from the Galaxy Zoo project.

4 Timeline

26th April to 11th May	literature survey
Mid-August to October	Basic prototype model
October to November	choosing the approaches which work, and implementing them on all data validation of the results
November to December	Refining the system, cleaning and commenting the code

5 Conclusion

We would like to thank Dr.Yogesh Wadadekar [1], Dr.C. H. Ishwara Chandra [1] for their supportive presence during the process of brainstorming potential research ideas. Their constant guidance has been an invaluable source of inspiration for us, and we are eager to continue working with them.

References

[1] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton. *ImageNet Classification with Deep Convolutional Neural Networks*.

[2] Matthew D. Zeiler, Rob Fergus. *Visualizing and Understanding Convolutional Networks*.
arXiv print : 1311.2901v3, 2013.

[3] Andrej Karparthy, Fei-Fei Li, Justin Johnson, Serena Yeung.
`http://cs231n.github.io/convolutional-networks/`

[4] Sander Dieleman, Kyle W. Willett and Joni Dambre
Rotation-invariant convolutional neural networks for galaxy morphology prediction.
arXiv print : 1503.0707v1, 2015.

[1] National Center for Radio Astrophysics - Tata Institute of Fundamental Research