Galaxy Zoo Challenge

Rohit Kumar Jha, Irfan Hudda

Problem Statement

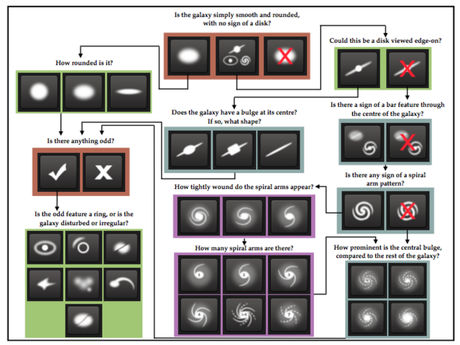

Figure 1. Flowchart of the classification tasks for GZ2, beginning at the top centre. Tasks are colour-coded by their relative depths in the decision tree. Tasks outlined in brown are asked of every galaxy. Tasks outlined in green, blue, and purple are (respectively) one, two or three steps below branching points in the decision tree. Table 2 describes the responses that correspond to the icons in this diagram.

Figure 1: Decision Tree of Classification[1]

We are given 61578 images of galaxies with probabilities corresponding to each of the 37 classes shown in above figure. The test set contains 79975 images for which we need to predict the probabilities corresponding to each class.

Motivation

Morphological information about the galaxies is very important in studying the dynamical history of galaxies and also it surroundings.

Central Pixel Approach

The simplest approach here was to crop the image to 10×10 at the center and take the average of RGB values.
And then we normalize this value and cluster these for each of the 37 classes. We perform a naive clustering using hash values for the clusters. The hash value is $Rf^2 + Gf + b$. Here f is constant.
Each cluster has an associated probability which is the average of the probabilities of all its members. The $RMSE$ for this approach was 0.16 when trained on entire training set.

Classifier-based Approach

In this approach we took the image as feature vector and to make classification with this feature vector feasible we cropped the image to 150×150 at center and resized it to 50×50 while retaining the color information.
We used RGB values of each pixel in this image to train a random forest regressor for each of the 37 classes. The $RMSE$ for this approach by training on one-fifth training and testing on other fifth was 0.165.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(p_i - a_i)^2}$$

These results can be further improved by stacking these regressors: Random Forest Regressor, Extra Trees Regressor, Ada Boost Regressor and Gradient Boosting using Meta Decision Trees. For each of the regressor, we can have 3-4 entries with varying number of parameters. It has been observed in practice that this performs really well in contrast to any single Regressor.

Preprocessing

The galaxy images are invariant under rotation. In other words there is no up or down in space. They are also scale and translation invariant to a limited extent. These invariance can be exploited by creating new data with same probabilities.

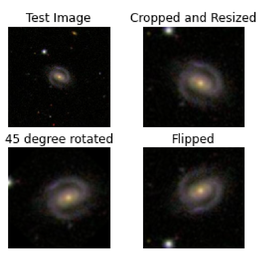

Figure 8: Example Preprocessing

Features-based Approach

This is a regression problem with probability values given for each of 37 classes. We converted it into a classification problem. For each of the 11 question, see image above, and each sub-class, we chose the images unambiguously belonging to that sub-class. Then we extracted the below mentioned features and ran Randomized PCA on it to obtain the top 50 eigenfaces. Some of the major features used:

- Gini, Kurtosis and Skewness Coefficients for Image Histogram Distribution Inequality

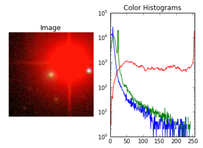

Figure 2: Color Histogram

- Entropy feature of an image
- Contour feature: Ellipse fitting

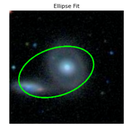

Figure 3: Ellipse fit for center galaxy

- Contour Feature: Bounding box
- Chirality: An object is chiral if it is not identical to its mirror image
- Central Moments (Contours) which are invariant to size, position and orientation
- Blob features: Blobs provide a complementary description of image structures in terms of regions, as opposed to corners that are more point-like.
- Light-Based features which are specific to galaxy images.

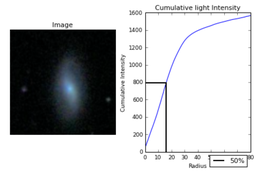

Figure 4: Radius of % intensity of light

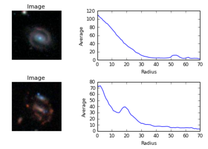

Figure 5: Radial Profile

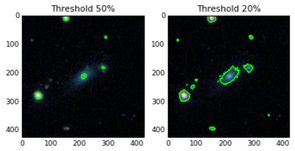

Figure 6: Contours for different Thresholds

Three set of different features were selected and for each set we obtained eigenfaces and used SVM to get predicted probabilities for each class. These results were used as input to Extra Trees Regressor to get the final values for each of the 37 classes.

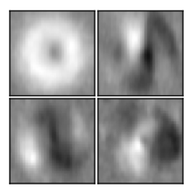

Figure 7: Top 4 Eigenfaces for Question 4 (Spiral)

Results and Conclusion

The central pixel approach resulted in RMSE of 0.16 and single classifier-based approach resulted in RMSE of 0.15 on $\frac{1}{5}$ training data. Other approaches were memory and computation extensive and could not be performed on out systems.
This challenge has shown that we can get good results with handcrafted features and it can get close to the results we obtain with CNN. But it is impossible to beat a CNN with good architecture using handcrafted features. Top three contestants used CNN with deep (7-9 layers) architecture.

References

[1] K. W. Willett et al. *Galaxy Zoo 2: detailed morphological classifications*. 2012.

[2] M. Banerji et al. *Galaxy Zoo: Reproducing Galaxy Morphologies Via Machine Learning*. 2009.

—-