ISSN: 0067-2904

Galaxy Morphological Image Classification using ResNet

Siddhartha Banerjee, Bibek Ranjan Ghosh, Ayan Gangapadhyay, Himadri Sankar Chatterjee
Ramakrishna Mission Residential College (Autonomous), Narendrapur, Kolkata, West Bengal, India

Received: 21/6/2020 Accepted: 21/12/2020

Abstract

Machine learning-based techniques are used widely for the classification of images into various categories. The advancement of Convolutional Neural Network (CNN) affects the field of computer vision on a large scale. It has been applied to classify and localize objects in images. Among the fields of applications of CNN, it has been applied to understand huge unstructured astronomical data being collected every second. Galaxies have diverse and complex shapes and their morphology carries fundamental information about the whole universe. Studying these galaxies has been a tremendous task for the researchers around the world. Researchers have already applied some basic CNN models to predict the morphological classes of the galaxies. In this paper, a residual network (ResNet) model is applied for this purpose. The proposed methodology classified the galaxies depending on their shape into 37 different classes. The performance of the methodology was evaluated using the data set provided by Kaggle. In this data set, 61,578 galaxy images are given, which are classified by human eye. The model achieved nearly 98% accuracy.

Keywords: Convolutional Neural Network, Residual Networks, ResNet-18, Galaxy Zoo Data set, Galaxy Morphological Classification.

1. INTRODUCTION

The cosmos is constructed by hundreds of millions of galaxies which define the structure of the universe. To understand the past, present, and future of the universe, studying the distribution of the physical properties of these galaxies is necessary. These properties are mostly measured by observation by astronomers and cosmologists to establish the theoretical models.

Physical information from the photometry of a galaxy is extracted by astronomers and cosmologists by looking at its morphology. For example, elliptical galaxies contain old and dying stars, whereas spiral galaxies contain young stars.

Advancement in technology in the past few decades allowed large scale surveys to categorize the galactic data at measureable rates. The Sloan Digital Sky Survey (SDSS) has categorized 1.2 billion objects across one third of the sky [1]. Thus, a robust automated system is needed to interpret and categorize such immense amounts of data.

The field of computer vision is enriched with the advancement of Convolutional Neural Network (CNN). It has been applied to classify and localize objects in images. Among many fields of application of CNN, it has been applied to understand this huge unstructured astronomical data. Researchers have already applied some basic CNN models to predict the morphological classes of the galaxies. In this paper, a residual network (ResNet) model is

*Email: bibekghosh2003@yahoo.co.in

applied for this purpose. Our result suggests that a simple ResNet model is able to achieve higher accuracy at the cost of training.

2. THE GALAXY ZOO DATA SET

Galaxies in the galaxy zoo data set provided by Kaggle [2] have already been classified through the help of hundreds of thousands of volunteers. They collectively classified the shapes of these images by eye in a successful citizen-science crowdsourcing project. But if data sets grow to contain millions or even billions of galaxies, this approach becomes less feasible.

The galaxy-zoo data set was classified using question tree which are given below.

Q1. Is the object a smooth galaxy, a galaxy with disk like shape or a star? *3 responses*
Q2. Is it edge-on? *2 responses*
Q3. Is there a bar? *2 responses*
Q4. Is there a spiral pattern? *2 responses*
Q5. How prominent is the central bulge? *4 responses*
Q6. Is there anything "odd" about the galaxy? *2 responses*
Q7. How round is the smooth galaxy? *3 responses*
Q8. What is the odd feature? *7 responses*
Q9. What shape is the prominence in the edge-on galaxy? *3 responses*
Q10. How tightly wounds are the spiral arms? *3 responses*
Q11. How many spiral arms are there? *6 responses*

2.1 PATHS AND THE DECISION TREE

Figure 1 showshow the questions lead to further questions to classify a particular image. For example, as shown in the Figure 1, if the response to the first question (*Is the object a smooth galaxy, a galaxy with features/disk or a star?)* is option one, i.e. the volunteer classifies the object as a smooth galaxy, then the next question would be question seven, i.e. *How rounded is the galaxy?*Thus, each galaxy is classified by recording the responses by following the decision tree.

Weighing the responses

Computation of the values for the morphological categories of the galaxies is performed using the following rule;the values for the first set of responses are computed by calculating the probability of the galaxy falling in each category. For each subsequent response, the probabilities for the response are first computed (these will sum to 1.0) and then multiplied by the value from which the new set of responses occur.

For example, if a galaxy is identified as smooth by 80%, as disk by 15%, and as star/ artifact by 5% users in first set of response, then the values for the morphological categories are the probability of the galaxies falling in each category, i.e. Class1.1 = 0.80, Class1.2 = 0.15, Class1.3 = 0.05.

Now, among these 80% users who identified the galaxy as “smooth”, 40% responded thatthe galaxy iscompletely round, 30% in between, and 30% cigar-shaped, while the values are:

Class 7.1 = 0.80 * 0.40 = 0.32
Class 7.2 = 0.80 * 0.30 = 0.24
Class 7.3 = 0.80 * 0.30 = 0.24

The concept of weighting is used to emphasize the fact that a good solution should get high rank for better classification.

3. PAST WORKS

Machine learning techniques are used extensively to classify galaxies and became an active area of research over the past two decades. Neural network, decision trees, and Naive-Bayes classifiers are applied on relatively small data sets and achieved 80% classification accuracy. Larger data sets, including Galaxy Zoo data set, are also categorized using more advanced techniques. 90% classification accuracy is achieved using neural network by

training the model using 900,000 objects from Galaxy Zoo data set with a novel set of features [4– 6].CNNswere also used to classify a subset of the 50,000 brightest objects in the Galaxy Zoo data set, with 99% accuracy [7]. Though these results are impressive, the main disadvantage of these studies is thatthat majority of these classifiers are classifying the objects into roughly three bins: elliptical galaxies, spiral galaxies, and other.

Several efforts have been applied on Galaxy Zoo data set for morphological classification of the galaxy images.Calleja *et al*. [8] used neural network and locally weighted regression method to classify morphological galaxy images. Principal component analysis was used to reduce the dimension of the data and extract relevant information from galaxy images. They achieve 91% accuracy considering three classes, whereas 95% accuracy was achieved considering two classes. Shamir [9] proposed an automatic galaxy image classifier using animage analysis-supervised learning algorithm.In this method, a large set of features isextracted from the images and then the most informative features are selected using Fisher scores. Images are classified using a simple Weighted Nearest Neighbor rule,where Fisher scores are used as the feature weights.Zhu *et al*. [10] proposed a residual networks (ResNets) model for galaxy morphology classification. The galaxy images are classified into five classes, i.e., completely round smooth, in-between smooth (between completely round and cigar-shaped), cigar-shaped smooth, edge-on, and spiral. 95.2% classification accuracywas achieved using their model.Jha *et al*. [11] applied various machine learning methods on the data set images to help understand the underlying theory. Chou [12] constructed a three-step system to classify the galaxy images. Feature extraction, machine learning regression, and probability normalization wereused for this purpose. Gauthier *et al*. [13] applied various supervised and unsupervised learning techniques to classify galaxy morphologies. Their direct classification technique, using random forest, achieved 67% accuracy. Dieleman [14] used a simple CNN for the task of classification.

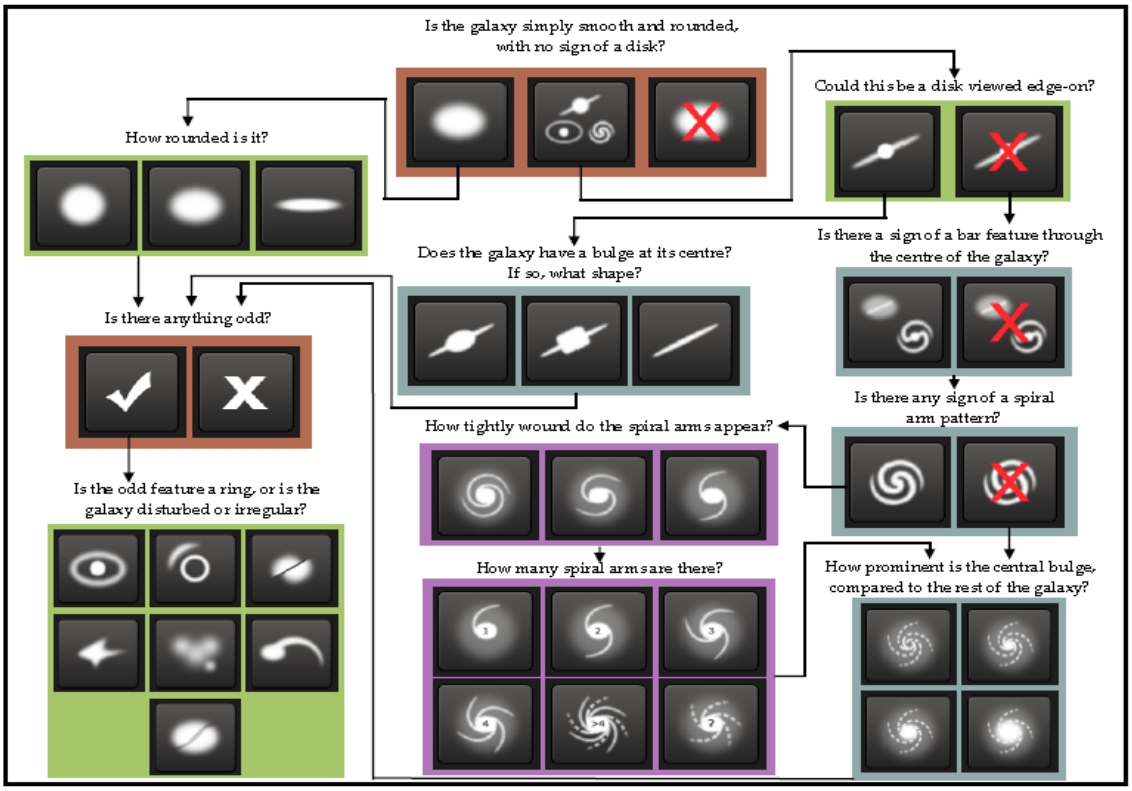

Figure 1-Flowchart of the classification task of Galaxy Zoo Data set, beginning at the top centre[3].

4. PROPOSED METHODOLOGY

4.1 Image Pre-processing

Each image is of 448X448 pixels. Most of the images contain the main data at the center and is surrounded by the black outer space. To increase efficiency, the best way is to retrieve the necessary part of the image. Thus the image was cropped from 424X424 to 312X312. This helped in eliminating most of the extra black region at the side of the image. Then the image was resized to 224X224 pixels.

4.2 Residual Model Architecture

Residual Model Architecture of CNN is one of the most successful architectures, with error rate of 3.57%. The problem with deep neural network models is that, during the training, the network starts converging. With more training, the accuracy gets saturated and then degrades rapidly, also known as the vanishing gradient problem. Most of the images in the data set are dominated by black surrounding and a bright object occupying the center of the image. With deeper models, the learning might saturate as the model learns less and less features. To avoid this, identity shortcut connections are created. It skips the training of one or more layers – creating the residual block. The reason why these skip of connections works is because they mitigate the problem of vanishing gradient, by allowing the alternate shortcut path. They also allow the model to learn an identity function which ensures that the higher layer will perform at least as good as the lower layers butnot worse.The Residual Neural Network with 18 layers (ResNet18) architecture was completely coded in Python using Keras with Tensorflow backend. Figure 2 describes the architecture behind the model.

Layer Name	Output Size	ResNet-18
conv1	$112 \times 112 \times 64$	7×7, 64, stride 2
conv2_x	$56 \times 56 \times 64$	3×3 max pool, stride 2
		$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$
conv3_x	$28 \times 28 \times 128$	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$
conv4_x	$14 \times 14 \times 256$	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$
conv5_x	$7 \times 7 \times 512$	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 2$
average pool	$1 \times 1 \times 512$	7×7 average pool
fully connected	1000	512×1000 fully connections
softmax	1000	

Figure 2-The ResNet 18 architecture.

The ResNet-18 model used in this project is shown in Figure 3.The model starts with a **Convolution Layer** that takes an image of galaxy as input. In this Layer,a simple two dimensional convolutional network containing 64 filters of size 7X7 is used. The output of the convolution network is normalized to avoid the problem of variable optimal target in deep convolution network. After that, the **rectified linear activation function** (RELU) is used to make the output nonnegative. Finally, some non-usefulparameter values are discarded.

The output of the Convolution Layer is used as the input of the **Intermediate Layer**. It consists of four repetitive sub-blocks; each consists of a combination of **Convolution block** and **Residual block**.

Convolution block is a simple block of Convolutional Neural Network that incorporates the skip-connection within itself. The block is divided into four components. The first two componentscontain a simple two dimensional Convolutional network, including theNormalization, followed by the rectified linearactivation, thatfunctions as **Convolution Layer**. The third component is the Convolutional Block for the shortcut. As mentioned earlier, the skip-connection provides a shortcut for outputs of some layers to jump to some layers ahead of it. The fourth component is the final part of the **Convolutional block**, where the output of the current block and the skip-connection are added. A simple addition operation on the two inputs is performed and the output is passed through a RELU activation function.

The **Residual block** is a simple block where the activation of a layer is fast-forwarded to a deeper layer of the network. This reduces the problem of vanishing gradient. Similar to the **Convolution block**, the **Residual block** also contains a sequence of Convolutional network, Normalization, and RELU activation function.

The output of the Intermediate Layer is down sampled by replacing each 7 X 7 block of the image by its average. The output is then flattened into a single dimensional tensor. The last component of the model is a Fully Connected Neural Network, containing 37 nodes, similar to the number of classes the input data is to be classified into.

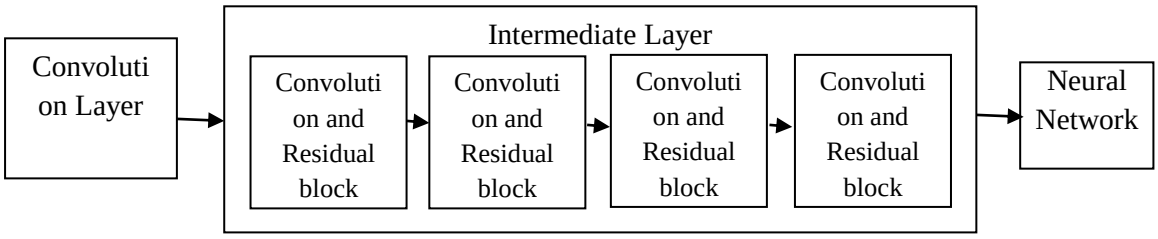

Figure 3- The ResNet Model.

4.3 Training

The model is now trained by using Galaxy Zoo Data set. It contains 61,578 galaxy images which are classified by human eye. 75% of this data set is used for training purpose. The training requires huge computational speed as the residual network essentially learns millions of parameters and the data to be passed is huge. The data is loaded and used for training in a controlled manner since the data set contains around 60 thousand images. Loading the total data set, preprocessing the images, converting the images into array, and passing it for training is not possible at once. Hence,the images are loaded in batches of 5000 images together. Each batch containsboth the training images (4000) and validation images (1000). Necessary pre-processing on the images, loading of the model, and training are done on these 5000 images. After that, the kernel is restarted. This allocates fresh memory to be used. Then the next batch of 5000 images is loaded and same tasks are performed.

Every result from the training batches are stored in *logs* folder and the model is saved to be loaded again. For each image, the model is trying to learn to output a tensor of length 37, corresponding to the morphology of the galaxy.The *categoricalhingeloss* function is used here because it is useful for multi-class classification when the expected value is within the range of (-1, +1). The Adam optimizer for the training is used to adjust the learning rate when training the same images repetitively.

5. EXPERIMENTAL RESULTS

The proposed methodology is tested using 25% of the Galaxy Zoo data set. The methodology

classifiedthe galaxy images into 37 categories by calculating the probability depending on morphological shape,as described in Section 2. Root Mean Squared Error (RMSE) was used as an evaluation metric, as follows:

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

N is the number of galaxies times the total number of responses, P_i is the predicted value, and A_i is the actual value. The model achieves nearly 98% accuracy on the validation data, but we are not sure of the loss value. For the *categorical_crossentropyloss* function, the values obtained are huge (greater than 80.46%), whereas for the *categorical_hinge*, the loss value decreases to be almost negligible. Figures 4 and 5 plot the training and the validation accuracy of the model through-out the training. In the training phase, the accuracy varies from 94% to 97% for each group of 5000 images, as shown in Figure 4, whereas in the validation phase, the accuracy varies from 97% to 99%, as shown in Figure 5. The x-axis shows the wall time relative to the first data point, i.e. the number of hours since the training run was started. This is useful to compare the performance of two or more different training runs that have notstarted simultaneously.

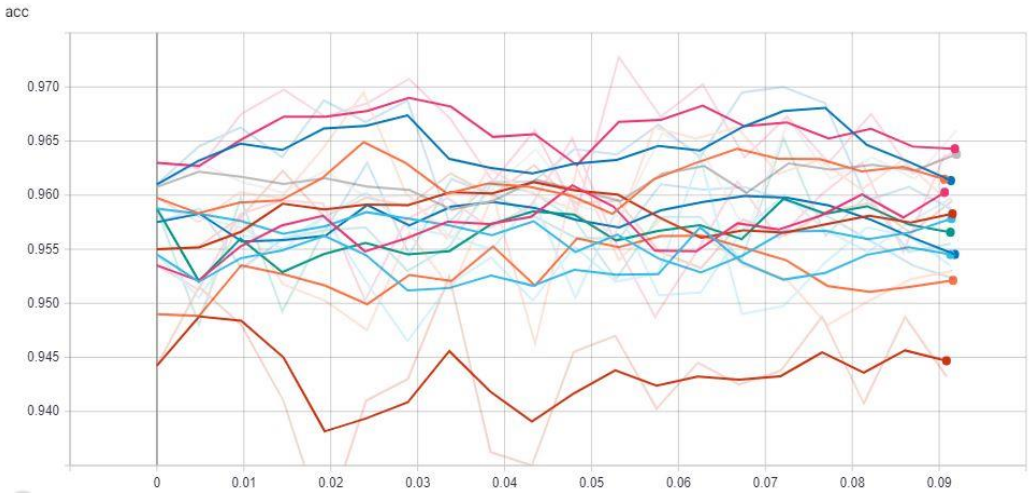

Figure 4-Accuracy in the training phase for every batch of 5000 images.

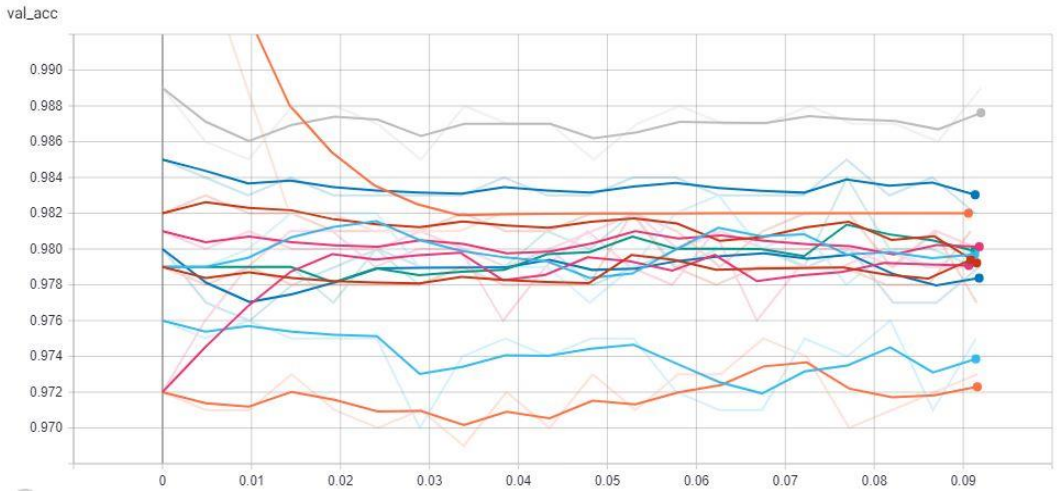

Figure 5-Accuracy in the validation phase for every batch of 5000 images.

6. CONCLUSIONS

In this paper, Residual Neural Network (ResNet) of convolutional neural network is used to classify the galaxy images depending on their shape. 37 different classes were considered depending on the morphology of the galaxy. The use of convolutional neural network eliminated the overhead of extraction of various features from the galaxy images. Our model does almost over-fit the data set with an accuracy of 98% at the cost of loss value. As for some future aspects, we expect to fix the problem of over-fitting by more data pre-processing, with some feature extractions that might render valuable information. Visualizing the activation functions of each layer might help in understanding what the model is trying to learn and in modifying the data for better classification.

References

[1] Albareti F. D., Prieto C. A., Almeida A., Anders F., Anderson S., AndrewsB. H., Aragón-SalamancaA., Argudo-FernándezM., ArmengaudE., AubourgE., "The Thirteenth Data Release of the Sloan Digital Sky Survey: First Spectroscopic Data from the SDSS-IV Survey Mapping Nearby Galaxies at Apache Point Observatory", *The Astrophysical Journal Supplement Series*, vol. 233, no. 2, 2017.

[2] Galaxy zoo data set, https://www.kaggle.com/c/galaxy-zoo-the-galaxy-challenge/overview.

[3] Willett K. W., Lintott C. J., Bamford S.P., Masters K. L., Simmons B. D., Casteels K.R., Edmondson M., Fortson L. F., Kaviraj S., Keel W. C.,Melvin T., Nichol R. C., Raddick M. J., Schawinski M., Simpson R. J., Skibba R. A., Smith A. M.,Thomas D., "Galaxy Zoo 2: detailed morphological classications for 304,122 galaxies from the Sloan Digital Sky Survey", *Monthly Notices of the Royal Astronomical Society*, vol, 435, no. 4, pp. 2835–2860, 2013.

[4] NaimA., LahavO., SodréL., Storrie –LombardiM. C., "Automated morphological classification of APM galaxies by supervised artificial neural networks", Monthly *Notices of the Royal Astronomical Society*, vol. 275, no. 3, pp. 567–590,1995.

[5] BazellD., AhaD. W., "Ensembles of Classifiers for Morphological Galaxy Classification", *The Astrophysical Journal*, vol. 548, no. 1, 2001.

[6] OwensE. A., GriffithsR. E., RatnatungaK. U., "Using oblique decision trees for the morphological classification of galaxies", *Monthly Notices of the Royal Astronomical Society*, vol. 281, no. 1, pp. 153–157, 1996.

[7] Huertas-CompanyM., Gravet R., Cabrera-Vives G., Pérez-González P.G., Kartaltepe J.S., Barro G., BernardiM., MeiS., ShankarF., DimauroP., BellE. F., Kocevski D., Koo D.C., FaberS.M., McintoshD.H., "A catalog of visual-like morphologies in the 5 CANDELS fields using deep-learning", *Astrophysical Journal Supplement Series*, vol. 221, no. 1, 2015.

[8] Calleja J. D. L., Fuentes O., "Machine learning and image analysis for morphological galaxy classification", Monthly *Notices of the Royal Astronomical Society*, vol. 349, no.1, pp. 87–93, 2004.

[9] Shamir L., "Automatic morphological classification of galaxy images". *Monthly notices of the Royal Astronomical Society*, vol. 399, no. 3, pp. 1367–1372, 2009.

[10] Zhu X., Dai J., Bian C., Chen S. , Chen S. , Chen H. . "Galaxy morphology classification with deep convolutional neural networks", *Astrophys Space Sci*ence, vol. 364, 2019.

[11] R. K. Jha and I. Huddal, Galaxy Zoo Challenge, *Project Report*, 2014.

[12] F. Chou, Galaxy Zoo Challenge: Classify Galaxy Morphologies from Images, 2014.

[13] A. Gauthier, A. Jain, E. S. Noordeh., Galaxy Morphology Classification, *Stanford University*, 2016.

[14] S. Dieleman, K. W. Willett, J. Dambre, "Rotation-invariant convolutional neural networks for galaxy morphology prediction", Monthly *Notices of the Royal Astronomical Society*, vol.450, no. 2, pp. 1441–1459, 2015.