Methods for Classification of Galaxy Types using Machine Learning

Christopher Lee
Seoul International School; Gyeonggido, South Korea
Email: chrisjlee2002@gmail.com

Abstract - Lately there has been a great sharing of astronomical data, especially the images from large telescopes such as the Hubble Space Telescope. Although all galaxies were formed by gravitational pull acting on stars, each resulted in a quite different shape. Galaxy Zoo contest in Kaggle.com offered an ideal dataset to apply machine learning to the galaxy classification problem. The dataset's labeling process was unique in that they were statistical, directly reflecting the people's judgement. Although the labels of the galaxy images went into subcategories as well, this research focused on the first level classification between the spiral and the elliptical galaxies. The training data they offered were numerous enough that over 10,000 images of each class could be split into training and testing data, to measure the accuracy of the classifier. Variations of LeNet were chosen, to squeeze more performance from it. The resulting accuracies were within about 95-97% in agreement to the labels, which were only 80% confident of the classification themselves.

Key Words - Galaxy Classification, Machine Learning, Galaxy Zoo, Kaggle

INTRODUCTION

The galaxy's morphology is an important piece of knowledge that would lead to a better understanding of the physics of galaxies, and the universe. Lately, there has been a great sharing of astronomical data, especially the images from large telescopes such as the Hubble Space Telescope [1] as shown in Figure 1. The images of the galaxies could be classified by an expert, or by volunteers on websites such as Galaxy Zoo [2]-[3]. Galaxy Zoo is an astronomy project that people contribute to by classifying a huge number of different galaxies. However, even with this crowdsourcing, galaxy classification is an overwhelming task due to its scale: for example, the Hubble Telescope captured more than 256,000 galaxies in a single image, according to MIT Technology Review May 2020.

Kaggle is the largest data science/machine learning community website. It includes projects and competitions with different data sets. "Machine learning is a field of computer science that uses algorithms and techniques to give the computer system the ability to learn, i.e., improving performance on a specific classification, with data, without being directly programmed to do certain steps" [4]. These algorithms work by building a model from a training data set of input and output pairs. Without being explicitly programmed, machine learning tries to produce an algorithm that matches the input output pairs in the training dataset as closely as possible. When the algorithm is presented with an input it has never seen, it will produce an output that is called "prediction." The goal is to have these "predictions" match the observed reality as closely as possible.

Different industries and agencies use machine learning for various purposes. For example, financial services such as credit card companies use machine learning for fraud prevention. Online merchants such as Amazon or Netflix use it to recommend books or movies that customers would actually enjoy.

Machine learning can also be used for image classification. The goal of this study is to present automated morphological classification of galaxies using machine learning algorithms and image analysis. This research uses machine learning to match categorizations of the galaxies done by humans, so that future images can be accurately classified with computers alone.

A pre-existing convolutional neural network (i.e. LeNet [5]) is used with a varying number of layers. Further tweaking made it possible to enhance the performance to the point of matching the human classification result.

Incidentally, the data that Kaggle provides does not definitely state the types of galaxies. Instead, it statistically scores the types of galaxies through the results of surveys.

FIGURE 1: Procedure of morphology classification of galaxies using the Hubble Telescope, Galaxy Zoo, and Kaggle [1]-[2]

CLASSIFICATION METHODOLOGY

I. Experimental Procedure

Mathematica version 11.3 and MacBook Pro (2.5 GHz Intel Core i7, 16 GB 2133 MHz LPDDR3) are used for machine learning programming. The hypothesis is that if machine learning is used to classify galaxy images into two categories, its accuracy should match human classifications with an error range of 5%.

Experimental procedure is shown in Figure 2. Galaxy Zoo data from Kaggle is used. This data contains 61,578 jpeg images. These images are labeled into several layers of galaxy classifications, and sub-classifications, and then sub-sub-classifications as shown in Figure 3. This research focuses on level 1 classification into the first (elliptical) or the second (spiral) types of galaxies. They are named Class 11 and Class 12 respectively throughout this report. There is also a third class, "Others," (Class 13) but this is not considered because the probabilities are negligible. The sum of the probabilities of all these three classes will always add up to 1.

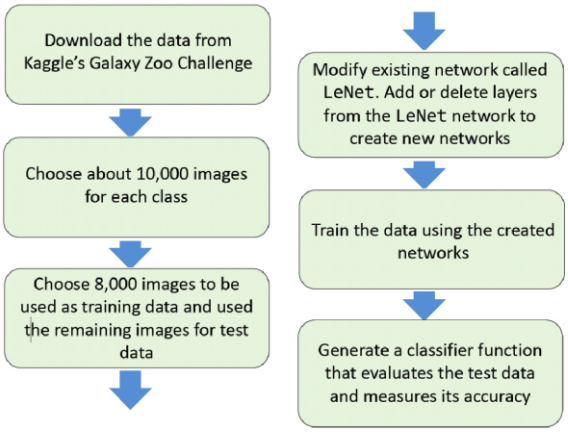

FIGURE 2: Experimental procedure of automated classification

These galaxy images are not classified nominally. They are only given the probability of belonging to that class according to many human classifiers who performed the task in the past. Only the top ranked (in terms of probability) 10,000 images of each class are chosen out of a total of 61,578 images. The chosen 10,000 images in each category have probabilities higher than 76.2% in Class 11, and higher than 87.2% in Class 12. These disparate probabilities stem from selecting a matching number of 10,000 images from each category. Class 12 images are labeled with higher confidence. Those images are stored in their respective folders. These two groups of high-probability images serve as labeled, nominal data. 8,000 (80%) images from each group are randomly selected for training data and the rest are used as testing data in the first phase.

A pre-existing network called LeNet is used as the basis to form the basic layers. The encoders and decoders, which are the inputs and outputs of the chain respectively, are altered based on the dimensions and the color space of the astronomical images. More layers and nodes are added or deleted to the LeNet in each cycle in search of higher accuracy. The training is done with the NetTrain command. Accuracy is measured using the ClassifierMeasurements command.

II. Galaxies Classification

As shown in Figure 3, Galaxy images are provided by the Hubble Space Telescope. The images were acquired during the "Galaxy Zoo: Hubble" project [2].

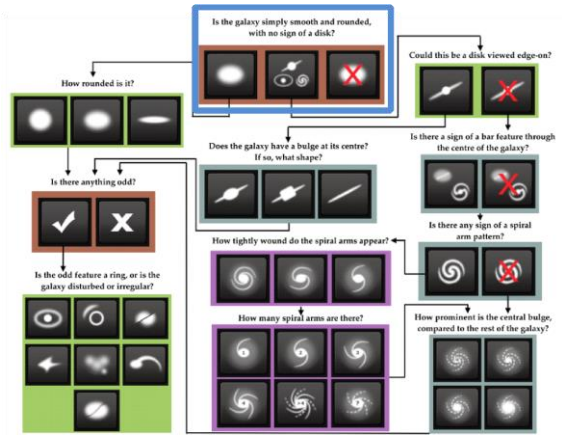

FIGURE 3: Classification of Galaxies [6]

Galaxy classification using Machine Learning was a success. A network called LeNet [5], which is initially used for handwritten character recognition, was utilized in this galaxy classification task. Several readily available networks were tried. Although some premade networks, such as VGG-16, produced very poor results while taking days to compute, LeNet in particular produced excellent results, achieving a 90% accuracy range.

III. Overview of the Whole Dataset

Figure 4 shows the distribution of the Class 11 probabilities out of 61,578 images. As shown in the histogram, the area of the data is more on the left side than on the right, which means that most of the classifiers thought that there were more images that belong to Class 12 than Class 11.

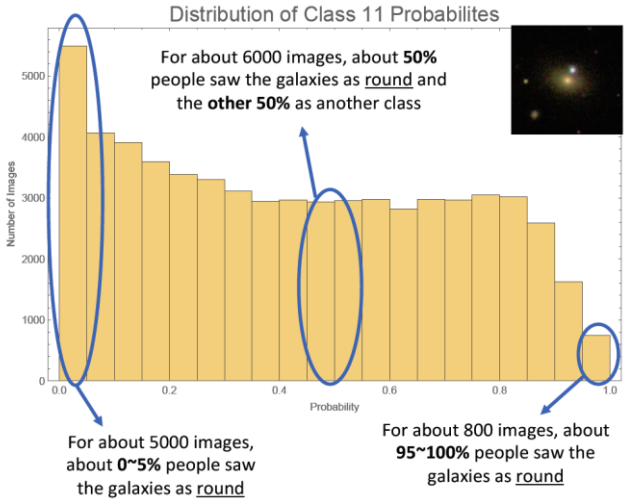

FIGURE 4: Distribution of Class11 Probabilities

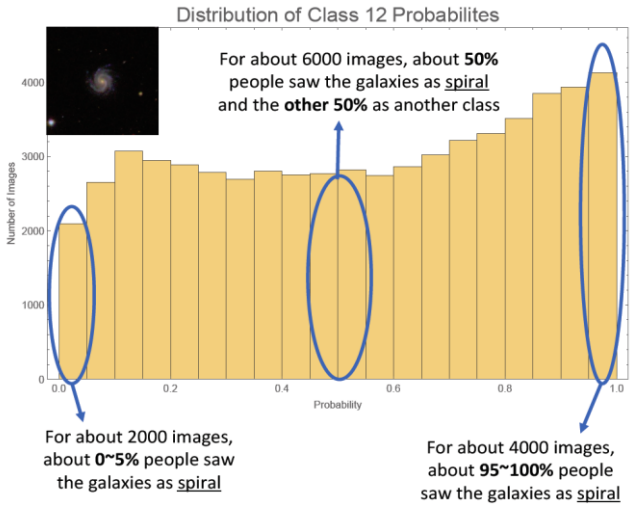

FIGURE 5. Distribution of Class12 Probabilities

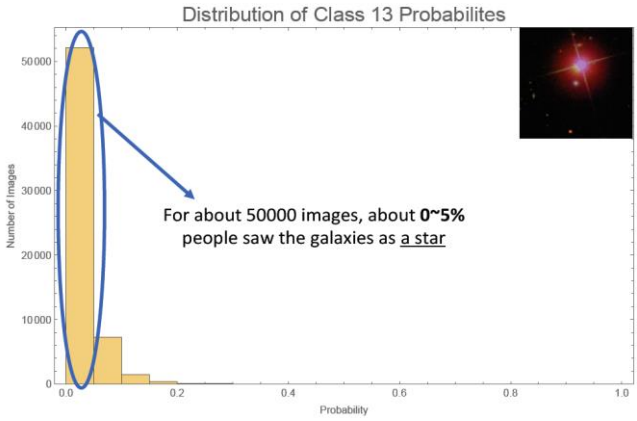

FIGURE 6: Distribution of Class13 Probabilities

In contrast, Figure 5 shows the distribution of the probabilities of Class 12 images in the whole dataset. As shown in the histogram, the area of the data is more tilted towards the right side than the left, which signifies that the classifiers thought that most of the images fit into Class 12 than others. The histograms in Figure 4 are not exact mirror images to Figure 5 because some of the images were classified as Class 13, which makes the count imprecise.

Figure 6 shows the distribution of the probabilities of Class 13 images in the whole dataset, which are non-galaxy images. However, as seen in the histogram, people who saw these pictures thought that they were mostly galaxies, and not stars. So since the probabilities of Class 13 images are negligible, this class can be ignored.

RESULTS AND DISCUSSION

I. Classification Results

Figure 7 indicates the structure of the chain with 20 layers. Each box in the diagram indicates a layer in the actual chain; they are connected together and going in one direction. This chain has the greatest number of layers out of all the chains that have been tested, and took the longest to compute. The classification also had an accuracy of about 97%, which is one of the highest results obtained in this research. Overall, counterintuitively, the more layers there are, the less time it takes to classify the images. After that, the computation time dramatically shoots up. This strange phenomenon needs to be further investigated.

In addition, the more layers there are, the more accurate the classification becomes. The optimal number of layers to have in a chain is around 14 layers, after which the accuracy plateaus while the computing time increases dramatically.

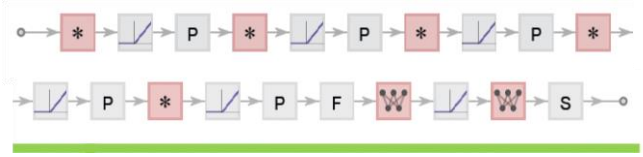

FIGURE 7: Structure of the chain with 20 layers

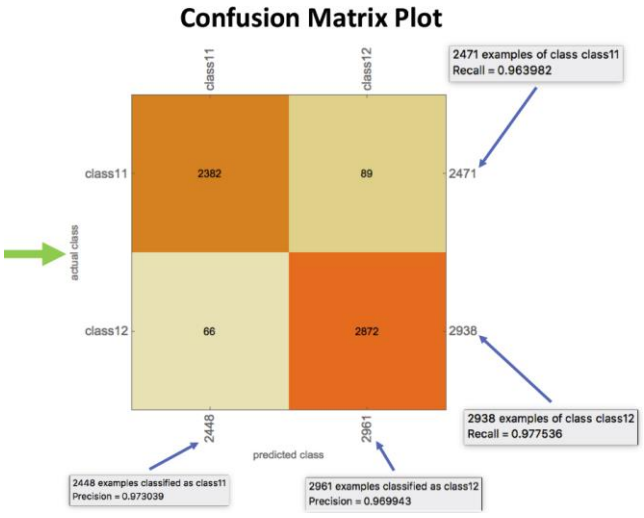

FIGURE 8: Confusion matrix showing the performance of machine learning

Figure 8 shows the result of the machine learning performance on the aforementioned data. 2,382 images of Class 11 were classified as Class 11, while 89 images were misclassified as Class 12. 2,872 images of Class 12 were classified as Class 12, while 66 images of Class 12 were misclassified as Class 11. This means that 96% of the images were classified correctly while 4% were classified incorrectly for Class 11. On the other hand, 98% of the images were classified correctly and 2% were classified incorrectly for Class 12. The overall classification resulted in a 97% accuracy for both classes, which is excellent.

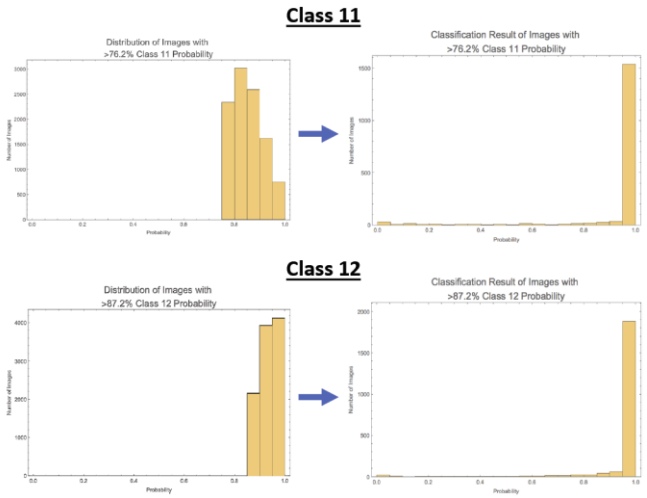

FIGURE 9: Distribution result of class 11 and 12 Probabilities

The confusion matrix plot shown in Figure 8 is binary, which means that it either classifies the images as Class 11 or Class 12 without any regards to the confidence of the prediction.

The middle 50% line is the threshold for the classification. For the top histogram in Figure 9, any bar that is on the right side of the middle line would be classified as Class 11 while the left side would be classified as Class 12. The classes are reversed for the pair of histograms at the bottom.

Histograms in Figure 9 show what is happening behind the scenes. The left top histogram is the input of the classifier, while the right top histogram is the result of the classification. Although the right histogram seems to have pushed all bars against the right wall, a low height bar can be observed near the floor, reading all the way to the left wall. This means that some images were classified as Class 12. The fact that some were misclassified was clear from the confusion matrix. This histogram shows just how many of them displayed a degree of Class 12-ness.

A similar story repeats for Class 12 histograms shown at the bottom of Figure 9. While appearing to have become more radicalized Class 12, there was an undercurrent of misclassification into Class 11 by varying degrees all the way to the left wall, which indicates near 100% Class 11-ness.

II. Network Optimization

Top in Figure 10 illustrates the effect of the number of layers of a chain on the accuracy of the classification of the galaxy images. As shown on the line graph, the line seems to increase then plateau. Although the accuracy continues to improve (with a slight dip at 17 layers, which is a mystery), it reaches a diminishing return around 14 layers.

At the bottom of Figure 10, the line graph shows the effect of the number of layers of a chain on the computation time. The graph seems to go on an exponential decay, as the amount of time it takes seems to decrease as the number of layers in a chain increases. This is a complete mystery. More layers should result in a longer computation time. But this is what was observed.

Then the line suddenly shoots up after 17 layers. The abruptness with which the computation time increases is also counterintuitive. This mystery in computation time must be investigated further. However, if this phenomenon is real, then this indicates that there might be a sweet spot where the highest accuracy can be attained without proportional expense in computation time.

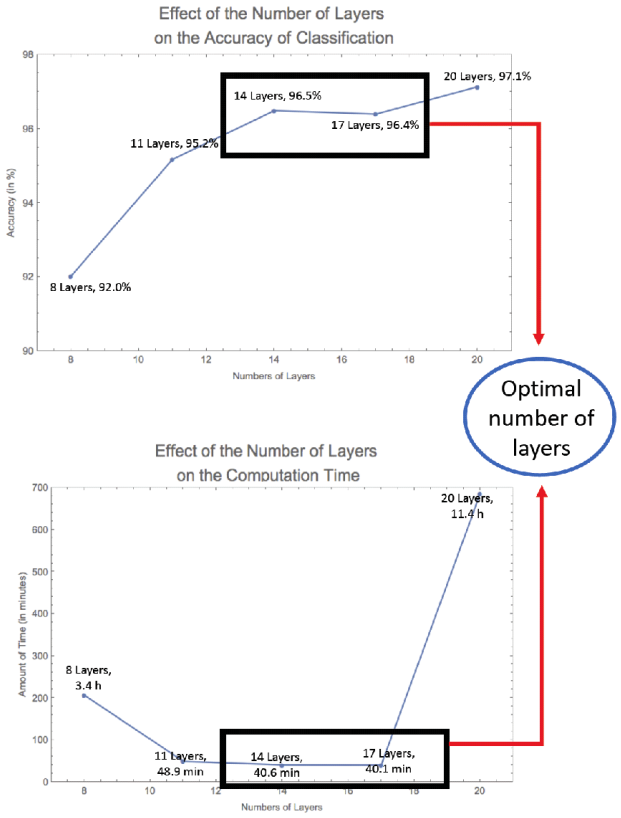

FIGURE 10: Effect of the number of layers on the performance (i.e., accuracy and computation time) of automated classification using machine learning method

CONCLUSION

Automated morphological classification of galaxies using machine learning algorithms and image analysis was examined. Machine learning was used to match the human categorization accuracy of the galaxy images. Results showed that the accuracy was high and the classification was successful, which was about 97%. The optimal number of LeNet layers for maximum performance per computation time was between 14 and 17. Future research is needed to resolve the strange phenomenon of “less computing time for more layers” observed during research. This machine learning based image classification is only the beginning: more applications shall come into effect as the performance continues to increase.

REFERENCES

"Hubble Space Telescope," Wikipedia, 9 October 2018. [Online]. Available: http://en.wikipedia.org/wiki/Hubble_Space_Telescope.

T. Zooniverse, "GALAXY ZOO: HUBBLE," Galaxy Zoo, 10 April 2010. [Online]. Available: https://blog.zooniverse.org/2010/04/23/galaxy-zoo-hubble/.

M. Banerji et.al., “Galaxy Zoo: reproducing galaxy morphologies via machine learning,” Mon. No. R. Astron. Soc., vol.406, pp.342-353, 2010.

"Machine Learning," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Machine_learning.

W. N. N. Repository, "LeNet Trained on MNIST Data," Wolfram, 30 January 2017. [Online]. Available: https://resources.wolframcloud.com/NeuralNetRepository/resources/LeNet-Trained-on-MNIST-Data.

K.W. Wilett et.al., “Galaxy Zoo 2: detailed morphological classifications for 304,122 galaxies from the Sloan Digital Sky Survey,” Mon. No. R. Astron. Soc., vol.435, pp.2835-2860 , 2013.