Galaxy Morphology Classification

Shreyas Kalvankar

Guided by -

Prof. Seema Gondhalekar

Seminar Report

K. K. Institute of Engineering Education and Research

Department of Computer Engineering

A.Y. 2019-20 Sem II

Acknowledgements

First and foremost, I would like to thank my seminar guide, **Prof. S. K. Gondhalekar**, for her guidance and support. I will forever remain grateful for the constant support and guidance extended by my guide, in making this seminar report. Through our many discussions, she helped me to form and solidify ideas.

With a deep sense of gratitude, I wish to express my sincere thanks to, **Prof. Dr. S. S. Sane** for his immense help in planning and executing the works in time. My grateful thanks to the departmental staff members for their support.

I would also like to thank my wonderful colleagues and friends for listening my ideas, asking questions and providing feedback and suggestions for improving my ideas.

Mr. Shreyas Kalvankar

Abstract

Galaxy morphology classification is of immense importance to study the origin of the galaxy to analyze the space surrounding the galaxy. A number of features such as the age, existence of gravitational anomalies, position of the galaxy in the supercluster, etc can be extracted by the shape and size of a particular galaxy. The count of the shape of galaxies can be helpful in determining the true nature of the growth of galaxies. The belief that the Hubble Tuning fork diagram shows the evolution of galaxies from simple to complex (Hubble, 1926) can be tested in accordance with the research. The paper Dai and Tong (2018) presented aims to propose a variant of residual network (ResNets), along with other popular Convolutional Neural Networks, for Galaxy Morphology Classification. It has been applied over a large set of training images from the Galaxy Zoo 2 dataset to classify the galaxies into five shapes namely spiral, edge-on, cigar-shaped smooth, in-between smooth, completely round smooth. The paper has provided different metrics such as accuracy, precision, recall values, etc to prove the performance to be better than the state of the art networks such as AlexNet, VGG, Inception and ResNets. The classification accuracy of the model on the testing set is given to be 95.2803% with accuracy of each class separately defined. The model can be applied to large-scale galaxy classification in forthcoming surveys.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Studying galaxies and classifying them into different classes is a problem not of ephemeral nature. Physicists have been trying to identify and segregate galaxies into individual groups and study their discrete traits to understand the formation of these galaxies and relating the physics that creates them. Morphology is determined by the physical characteristics and the orbital structure of the galaxies. The shape of the galaxy potentials determines which orbital families are present. Stars moving in the specific parts of the space phase generate morphological features such as bars, rings, peanut bulges, pseudo-bulges, etc. Gases pile up close to orbital resonances producing regions of star formation. By looking at the morphology of a particular galaxy, we can learn about the history of star formation and secular evolution of galaxies. The stability of such features tells us about distribution of mass throughout the galaxy both dark and luminous.

1.1 History of Galaxy Morphology and classes

The first classification scheme of galaxy morphology was proposed by Edwin Hubble, 1926. The classification diagram was termed "The Hubble Sequence" that used visual inspection by classifying fewer than 400 galaxies. The Hubble Sequence (also called "Hubble Tuning Fork") classified the galaxies into three basic classes: spiral, elliptical, irregular (Hubble, 1926).

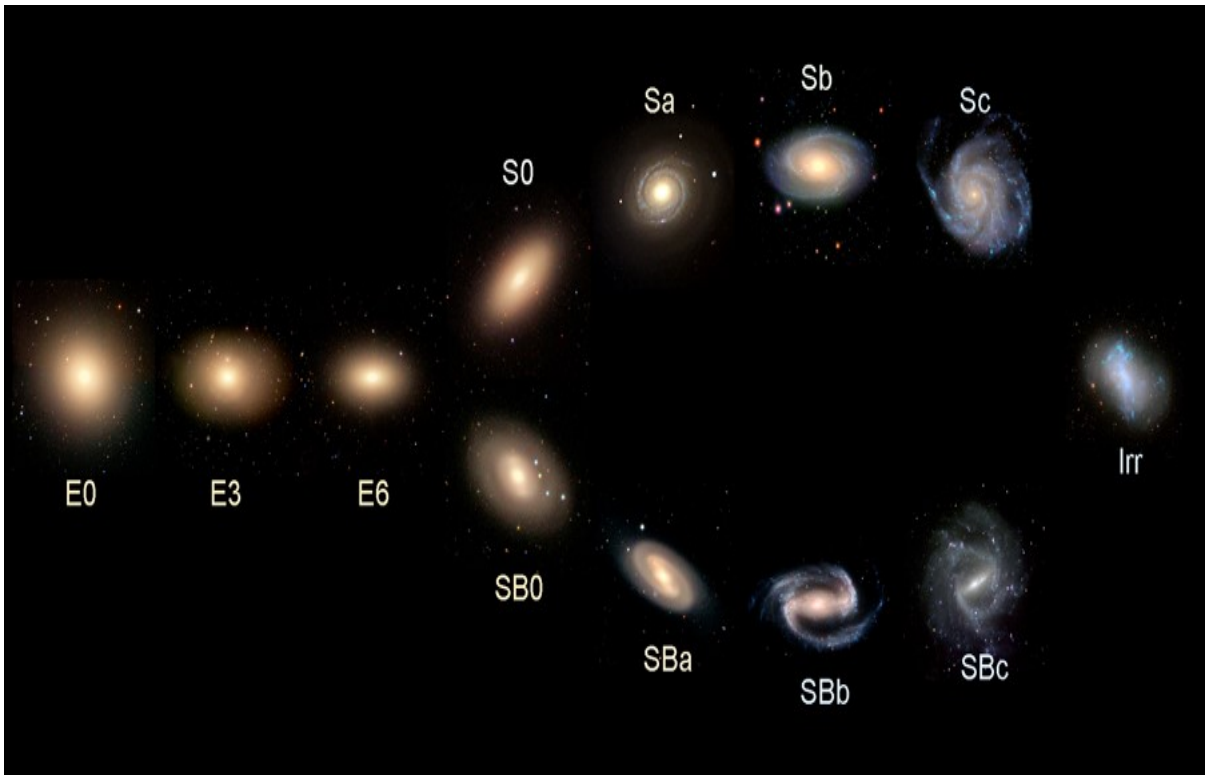

Figure 1.1: The Hubble Sequence

One of the major challenges in studying the morphologies is the techniques used for measurements. Before the computerized era of astrophysics, the technique of visual inspection and classification has been used by experts for several decades (Hubble, 1926; de Vaucouleurs, 1959; Edmondson, 1961; van den Bergh, 1976).

The computerized era of astrophysics has revolutionized galaxy morphology classification. Both parametrized (Sérsic and de Córdoba. Observatorio Astronómico, 1968; Cohen et al., 2003) and non-parametrized approaches have been used along with combined approaches (Conselice, 2003; Lotz et al., 2004) to reduce each galaxy to one number. This approach enables the processing of large scale images from different sky surveys (Djorgovski et al., 2013) and also helps provide a uniform quantitative set of parameters.

The previous methods of classification like visual inspection, however effective, were not able to cope with the sheer volume of data provided by the modern sky sur-

veys such as the SDSS (Sloan Digital Sky Survey). This called for a better classification methodology which could process a huge amount of data with much more efficiency.

1.2 Space surveys and galaxy classification

The previous methods of classification like visual inspection, however effective, were not able to cope with the sheer volume of data provided by the modern sky surveys such as the SDSS (Sloan Digital Sky Survey). This called for a better classification methodology which could process a huge amount of data with much more efficiency. Some experts have performed extensive work in the detailed classification of the subset of the SDSS images.

The The Galaxy Zoo 1 project obtained more than 40,000,000 classifications made by approx. 100,000 participants. In the next phase Lintott et al. (2010) included the data release of nearly 900,000 galaxies. The paper presented the measures of classification accuracy and bias. The data from Lintott et al. (2008) was substantially reduced by comparing with the professional catalogues. Data reduction was performed and no prominent changes were observed after cross-checking. The samples with 80% agreement among the users were labelled as 'clean' and with 95% agreement among the users were called 'superclean' samples. The accuracy of the samples that were created by collecting data clicks and forming them into a scientific catalogue was very close to the actual morophology feature that the galaxy possessed.

The Galaxy Zoo 2 project was launched later and Willett et al. (2013) produced the data release of nearly 16 million morphological classification with 304,122 galaxies, drawn from the SDSS. While the original Galaxy Zoo project identified galaxies as early-types, late-types or mergers, GZ2 measures finer morphological features Willett et al. (2013). This data release allowed a complete study of the finer morphological features and the co-relation of these features with properties of the galaxies i.e. mass, stellar and gas content, environment.As mentioned in the paper, although proxies such as spectral features, surface brightness profile, have been used extensively, they cannot be a replacement for for full morphological classification as pointed out by Lintott et al. (2010).This

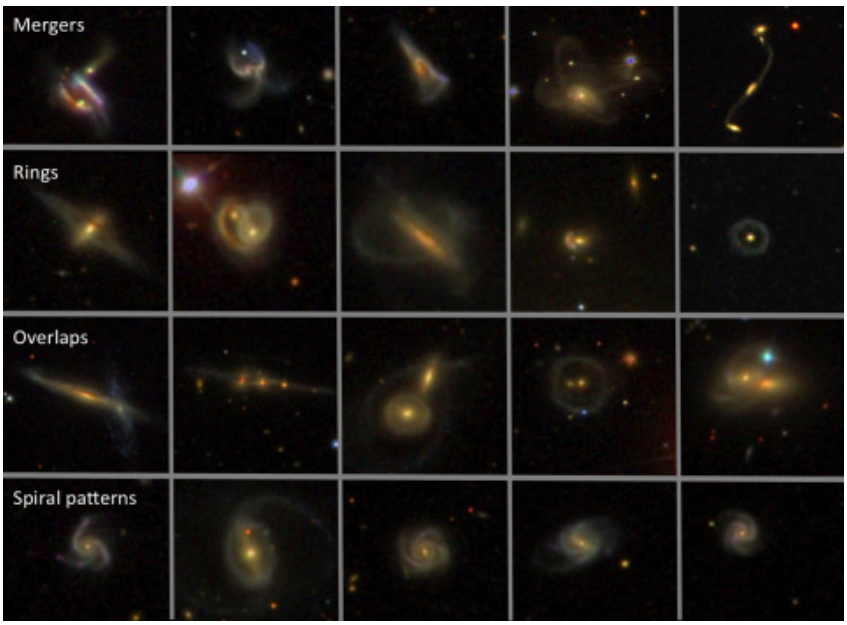

Figure 1.2: Galaxy features laid out in the Galaxy Zoo project

data has been used in studies of galaxy formation and evolution (Land et al., 2008; Schawinski et al., 2009; Willett et al., 2015).

The third phase in the crowd sourced visual classification was brought forward by Willett et al. (2016). The new Galaxy Zoo : HST legacy imaging presented the data release of the Galaxy Zoo: Hubble (GZH). The GZH mainly focused on drawing surveys conducted by the Hubble Space Telescope to view the earlier epochs of galaxy formation. It's data contained measurements of disc- and bulge-dominated galaxies, spiral disc structure details that relate to Hubble type, identification of bars and measurements of clump identification. Willett et al. (2016) also suggests new methods of calibrating galaxies at different luminosities and at different redshifts by artificially redshfiting the galaxies and using them as a baseline. The present Galaxy Zoo i.e. the 4th incarnation of the project combines new imaging from the Sloan Digital Sky Survey (SDSS) with the most distant images from the Hubble Space Telescope CANDLES survey.

Chapter 2

Literature Review

2.1 ANN and analysing the RMS dispersion

Naim et al. (1995) attempted the first ever automated classification using supervised artificial neural networks. They trained the network based on the morphology of the galaxy images as they appear on the blue survey plates. The network was used to classify images from the Automated Plate Measurement (APM) Equitorial Catalogue of Galaxies while reducing each image to the morphological features such as number of spiral arms and bulge size. By using this Supervised approach,the RMS dispersion between the ANN type and correct mean type was comparable to the overall RMS dispersion between the experts.

2.2 Implementaion of decision trees

Decision trees had proven themselves to be quite useful in automated classification techniques in a number of astronomical domains. Owens et al. (1996) employed oblique decision trees for the morphological classification using the data from Storrie-Lombardi et al. (1992).They also indicated that the data could be classified into lesser, but well defined categories. Galaxies could now be confidently classified to larger, overlapping regions, i.e. that multiple decision trees could, in fact, be generated to distinguish easily

between different regions along the continuum of classifications.This essentially meant that, while the non-neighbour classes could be easily separated, the neighbouring classes could not. While trees grown to distinguish E-type galaxies from Sa+Sb-types were accurate but those grown to distinguish E-types from S0-types would be very inaccurate (Refer Fig. 1.1).

2.3 Naive Bayes and Decision Tree induction algorithm

Bazell and Aha (2001) showed a comparison of three algorithms for automated galaxy classification. They used a Naive Bayes classifier and a decision-tree induction algorithm with pruning for automated classification of 800 galaxies, proving that an ensemble of classifiers decreases the classification error.The results show that (1) the neural network produced the best individual classifiers (lowest classification error) for the majority of cases, (2) the ensemble approach significantly reduced the classification error for the neural network and the decision-tree classifiers but not for the Naive Bayes classifier, (3) the ensemble approach worked better for decision trees (typical error reduction of 12%-23%) than for the neural network (typical error reduction of 7%-12%), and (4) the relative improvement when using ensembles decreases as the number of output classes increases. While more extensive comparisons are needed (e.g., a variety of data and classifiers), our work is the first demonstration that the ensemble approach can significantly increase the performance of certain automated classification methods when applied to the domain of morphological galaxy classification.

Feature Name	Description
Peak brightness	Maximum brightness level in image
m_{q2q3}	Ratio of fitted slope of $I(r)$ vs r for the second and third quartiles
Ellipticity	Ratio of the semimajor to semiminor axis length
Area	Number of pixels contained in the object
Max(rI)	Maximum value of the plot of $rI(r)$ vs. r
Asym	Comparision between original galaxy and galaxy rotated 180°
r_25/r_75	Ratio of radii at which 25% and 75% of light is enclosed in a plot of $I(r)$ vs r
R_{Bulge}	Radius where $I(r)$ falls to 90% of peak value
C_3, C_6	Concentration indices for the annuli 3 and 6
Isophotal displacement	Maximum displacement of the centres of five isophotal levels
Isophotal filling factor	Area of an isophotal level relative to the area of the enclosing ellipse
P_{max}	Maximum value of the normalized co-occurence matrix, c_{ij}
Entropy	$-\sum_{i,j} c_{ij} log(c_{ij})$

Table 2.1: Description of features used in morphological classification

2.4 Neural networks with Locally weighted regression

De La Calleja and Fuentes (2004) used a neural network along with a locally weighted regression method, and implemented a homogeneous ensembles of classifiers. It was found that accuracy dropped from 95.11% to 92.58% while classifying galaxies into two classes (E and S) to three classes (E, S, Irr). Further increase in classes caused the accuracy to drop further i.e. accuracy of 56.33% for a five-case classification which further reduced to 48.50% for a seven-case classification.

Banerji et al. (2010) employed Artifical Neural Network to classify galaxies into three classes ie.spirals,early types,point sources. A combination of the profile fitting and adaptive weighted fitting parameters resulted in better than 90% accuracy, it was observed that the input parameters are more decisive in achieving greater accuracy than the completeness of magnitude of the training set. Gauci et al. (2010) applied and compared Decision tree algorithms using CART and C4.5, Random Forests and fuzzy logic algorithms. While promising results were achieved in all the employed algorithms, Random Forests gave the highest accuracy.

2.5 Linear Discriminant Analysis Technique

Ferrari et al. (2015) presented an extented morphometric system which classified galaxies automatically based on the CASGM coefficients (Concentration, Asymmetry, Smoothness) along with some new parameters such as Entropy and spirality. Using the data from the Galaxy zoo project, spiral and elliptical galaxies were used for training, a Linear Discriminant Analysis (LDA) technique was employed to classify the galaxy samples. The cross validation showed that the accuracy achieved was about 90%. The approach was not simple visual inspection but rather based the scheme on physical characteristics of the galaxies (e.g. age weighted using luminosity or mass) and it was observed that the results matched closely with the visual classification.

2.6 Feature extraction and Convolutional Networks

LeCun et al. (2015) showed how the performance of classification depends on feature engineering. The feature engineering and feature extraction played an important role in classification models as further models exploited the elemental features of a galaxy image and used advanced algorithms to classify them. Deep learning models consists of multiple non-linear layers which learn data representations and automatically extract features from the raw data that is fed into it (Bengio et al., 2013; LeCun et al., 2015).

After a series of non-linear transformations, the higher levels have abstract representations of data and these can essentially be used for discrimination and classification purposes. Thus, deep convolutional neural networks (CNNs) have become a particularly dominant approach for image classification and feature extraction. The tremendous datasets such as the Galaxy Zoo which are well equipped with the predefined data representations allow faster and efficient implementation of these CNNs, with many works yielding excellent results.

Mairal et al. (2014) proposed to train convolutional neural networks to approximate kernel feature maps, which allow the desired invariance properties to be encoded in the choice of kernel, and subsequently be learnt.

The first ever attempt at classifying galaxies using a convolutional architecture

was made by Dieleman et al. (2015) by implementing a 7-layer model. He classified the galaxies based on their morphological features by translation of the images and exploiting the rotational invariance of galaxies.Then, Huertas-Company et al. (2011) used the Dieleman model to classify high redshift galaxies in the 5 Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey (CANDELS). Kim and Brunner (2016) presented a star–galaxy classification framework that uses deep ConvNets directly on the reduced, calibrated pixel values. Using data from the Sloan Digital Sky Survey and the Canada–France–Hawaii Telescope Lensing Survey, they demonstrated that ConvNets are able to produce accurate and well-calibrated probabilistic classifications that are competitive with conventional machine learning techniques.

Later, Dai and Tong (2018) implemented a combination of the Dieleman model along with the residual network (ResNets) proposed by He et al. (2016a) to classify the galaxies into 5 classes and achieved an accuracy of 95.2083% on the testing set.

Chapter 3

Problem Definition

Through numerous methods and approaches, galaxies were being classified into their respective classes rapidly. The problem of efficiency still remained as the importance of galaxy classes kept increasing. The methods that achieved an excellent accuracy for binary classifications mainly spiral and elliptical were particularly failing to achieve the same if the number of classes were increased. Furthermore, the classes that formed were refined over the years as more and more galaxies were discovered. It was observed that networks with over 90% testing accuracy with binary classifications had their accuracy reduced by almost half when introduced with a new class (e.g. irregular). This was making it harder to develop a method that was efficient. On the other hand, the data that was provided by the surveys kept on increasing. Thus, a new approach was necessary for galaxy classification with more refined classes and maximum efficiency.

Chapter 4

Solution

4.1 Dieleman Model

4.1.1 Methodology

Dieleman was the first person to propose a deep convolution model for galaxy morphology classification. He observed the drawback of the restricted connectivity patterns in convolutional networks. These result in reduced parameters required for modelling large images. But by exploiting the translational symmetry of the images, it could function to increase them. Furthermore, other invariances exist that could essentially improve the amount of parameters used for modelling a better network. Rotational invariance could be one of these as rotating a galaxy would have no effect on the morphology of the galaxy. Thus, exploiting the rotational invariance would provide better results in generalizing the classifications. Though, it was easier to use the translational symmetry for a convolutional network the same was not the case to implement a rotational invariance. Application of the same rotation filter would have to have different instances of translation applied. Rotating the image by any other angle which is not a multiple of 90° would result in missing pixels which would need interpolation to fill. These complications made exploiting the rotational symmetries much more difficult.

The dataset provided by the Galaxy Zoo project had been observed to have some images that were vertically and horizontally flipped as an experiment with the crowd-

sourced classification. Land et al. (2008) showed showed that the raw votes had an excess of 2.5% for S-wise (anticlockwise) spiral galaxies over Z-wise (clockwise) galaxies. Since this effect was seen in both the raw and mirrored images, it was interpreted as a bias due to preferences in the human brain, rather than as a true excess in the number of apparent S-wise spirals in the Universe. The Galaxy Zoo 2 probabilities do not contain any structures related to handedness or rotation-variant quantities, and no rotational or translational biases have yet been discovered in the data. If such biases do exist, however, this would presumably reduce the predictive power of the model since the assumption of rotational invariance to the output probabilities would no longer apply.

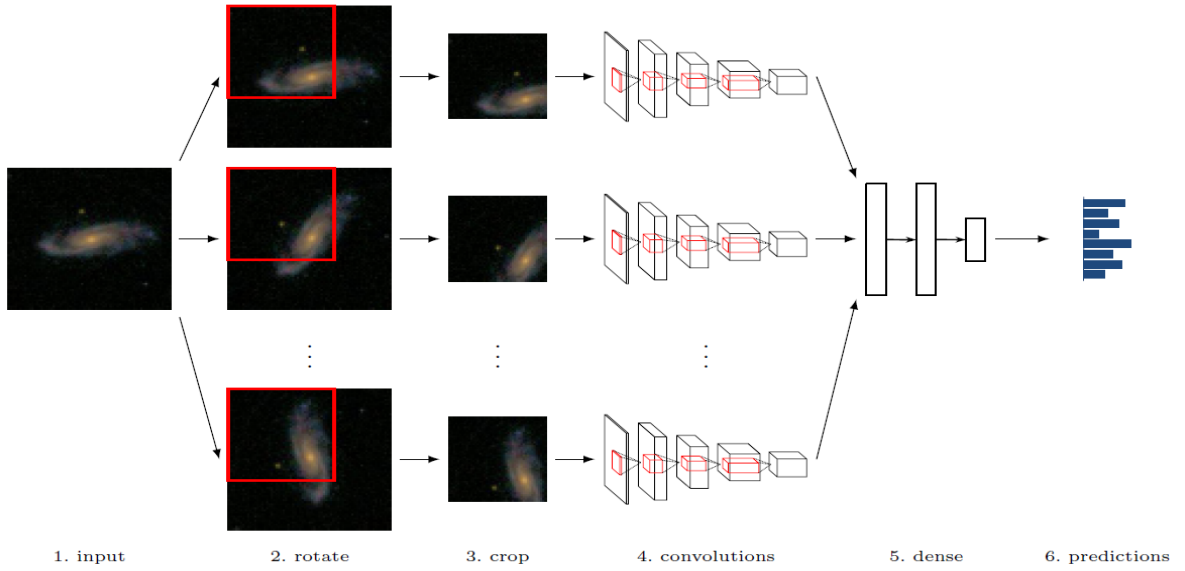

Figure 4.1: Schematic overview of Dieleman approach

Deep neural networks have a large number of lernable parameters as opposed to the limited training data. As a result, there is a high risk of overfitting, i.e. a network will memorize the training example because it has enough capacity. To tackle this, Dieleman et al. (2015) used many techniques:

- **data augmentation** : extending the training set by randomly perturbing images in a way that leaves their associated answer probabilities unchanged;
- **regularization** : penalizing model complexity through use of dropout;
- **parameter sharing** : reducing the number of model parameters by exploiting translational and rotational symmetry in the input images;
- **model averaging** : averaging the predictions of several models.

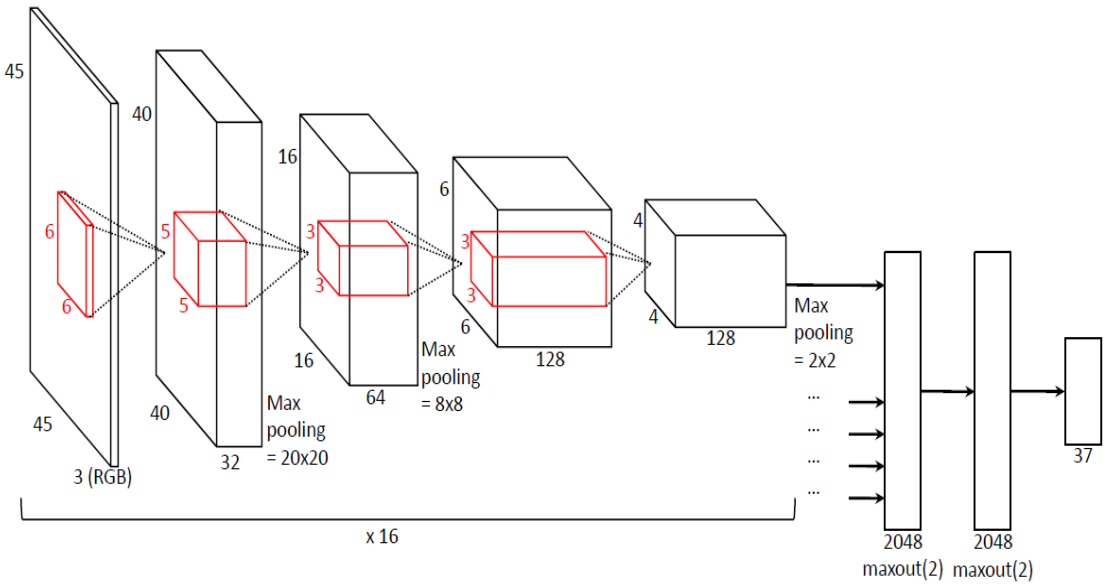

Figure 4.2: Schematic overview of the model architecture

4.1.2 Results

The dataset used was from the Galaxy Zoo 2 which was a collection of answers from the decision tree in table 4.1).

Dieleman achieved an RMSE score of 0.7671 for the best performing network and later, he averaged the results over 17 networks to get an RMSE score of 0.7467. The

results for precision and recall pf each question in the GZ2 tree (refer table 4.1) have been summarized in the table 4.2.

Task	Question	Responses	Next
01	Is the galaxy simply smooth and rounded, with no sign of a disk?	smooth	07
		features or disk	02
		star or artifact	**end**
02	Could this be a disk viewed edge-on?	yes	09
		no	03
03	Is there a sign of a bar feature through the centre of the galaxy?	yes	04
		no	04
04	Is there any sign of a spiral arm pattern?	yes	10
		no	05
05	How prominent is the central bulge, compared with the rest of the galaxy?	no bulge	06
		just noticible	06
		obvious	06
		dominant	06
06	Is there anything odd?	yes	08
		no	**end**
07	How rounded is it?	completely round	06
		in between	06
		cigar-shaped	06
08	Is the odd feature a ring, or is the galaxy distributed or irregular?	ring	**end**
		lens or arc	**end**
		distributed	**end**
		irregular	**end**
		other	**end**
		merger	**end**
		dust lane	**end**
09	Does the galaxy have a bulge at its centre? If so, what shape?	rounded	06
		boxy	06
		no bulge	06
10	How tightly wound do the spiral arms appear?	tight	11
		medium	11
		loose	11
11	How many spiral arms are there?	1	05
		2	05
		3	05
		4	05
		more than four	05
		can't tell	05

Table 4.1: The GZ2 decision tree, comprising of 11 tasks and 37 responses.

		precision	recall	#example
Q1: smoothness				6144
A1.1	smooth	0.8459	0.8841	2700
A1.2	features or disk	0.9051	0.8742	3435
A1.3	star or artifact	1.0000	0.4444	9
Q2: edge-on				3362
A2.1	yes	0.9065	0.8885	655
A2.2	no	0.9732	0.9778	2707
Q3: bar				2449
A3.1	yes	0.7725	0.7101	483
A3.2	no	0.9302	0.9486	1966
Q4: spiral				2449
A4.1	yes	0.8715	0.8270	1451
A4.2	no	0.7659	0.8226	998
Q5: bulge				2449
A5.1	no bulge	0.6697	0.5000	146
A5.2	just noticeable	0.7828	0.8475	1174
A5.3	obvious	0.8292	0.8049	1092
A5.4	dominant	0.4444	0.1081	37
Q6: anything odd				6144
A6.1	yes	0.8438	0.7500	828
A6.2	no	0.9617	0.9784	5316
Q7: roundedness				2619
A7.1	completely round	0.9228	0.9282	1197
A7.2	in between	0.9128	0.9171	1279
A7.3	cigar-shaped	0.9000	0.8182	143
Q8: odd feature				824
A8.1	ring	0.9097	0.9161	143
A8.2	lens or arc	?	0.0000	2
A8.3	disturbed	0.8000	0.4138	29
A8.4	irregular	0.8579	0.8674	181
A8.5	other	0.6842	0.6810	210
A8.6	merger	0.7398	0.7773	256
A8.7	dust lane	0.5000	0.6667	3
Q9: bulge shape				493
A9.1	rounded	0.9143	0.9412	340
A9.2	boxy	?	0.0000	8
A9.3	no bulge	0.8601	0.8483	145
Q10: arm tightness				1049
A10.1	tight	0.7500	0.7350	449
A10.2	medium	0.6619	0.7112	457
A10.3	loose	0.7373	0.6084	143
Q11: no. of arms				1049
A11.1	1	1.0000	0.2037	54
A11.2	2	0.8201	0.8691	619
A11.3	3	0.4912	0.3182	88
A11.4	4	?	0.0000	21
A11.5	more than 4	0.4000	0.4000	20
A11.6	can't tell	0.5967	0.7368	247

Table 4.2: Precision and recall scores for each answer

4.2 Very deep convolutional neural networks

4.2.1 Residual Networks

After the celebrated victory of AlexNet (Krizhevsky et al., 2012) at the LSVRC2012 classification contest, deep Residual Network (He et al., 2016a) was arguably the most groundbreaking work in the computer vision/deep learning community in the last few years. ResNet makes it possible to train up to hundreds or even thousands of layers and still achieves compelling performance.

According to the universal approximation theorem, given enough capacity, we know that a feedforward network with a single layer is sufficient to represent any function. However, the layer might be massive and the network is prone to overfitting the data. Therefore, there is a common trend in the research community that our network architecture needs to go deeper. Since AlexNet, the state-of-the-art CNN architecture is going deeper and deeper. While AlexNet had only 5 convolutional layers, the VGG network (Simonyan and Zisserman, 2014) and GoogleNet (also codenamed InceptionV1) (Szegedy et al., 2014) had 19 and 22 layers respectively.

However, increasing network depth does not work by simply stacking layers together. Deep networks are hard to train because of the notorious vanishing gradient problem — as the gradient is back-propagated to earlier layers, repeated multiplication may make the gradient infinitively small. As a result, as the network goes deeper, its performance gets saturated or even starts degrading rapidly.

Before ResNet, there had been several ways to deal the vanishing gradient issue, for instance, Szegedy et al. (2014) adds an auxiliary loss in a middle layer as extra supervision, but none seemed to really tackle the problem once and for all.

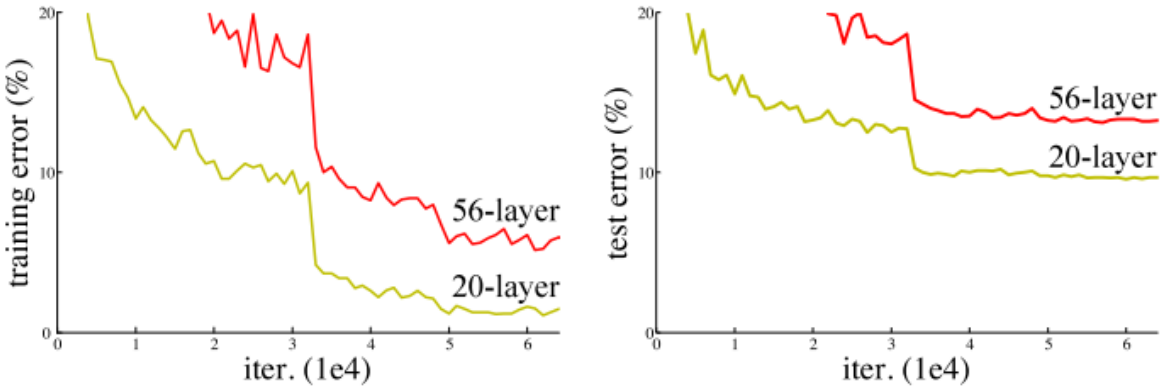

Figure 4.3: Performance after increasing no. of layers over training (left) and testing (right)

The core idea of ResNet is introducing a so-called "identity shortcut connection" that skips one or more layers, as shown in the figure 4.4.

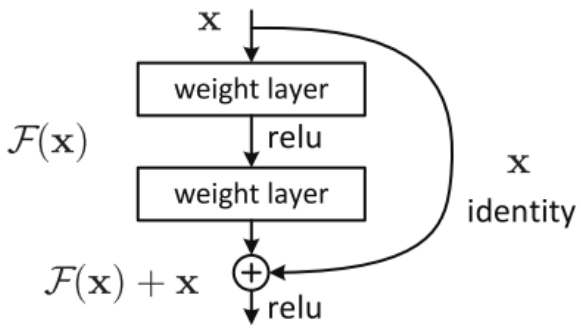

Figure 4.4: Residual Block with skip connection

He et al. (2016a) argue that stacking layers shouldn't degrade the network performance, because we could simply stack identity mappings (layer that doesn't do anything) upon the current network, and the resulting architecture would perform the same. This indicates that the deeper model should not produce a training error higher than its shallower counterparts. They hypothesize that letting the stacked layers fit a residual mapping is easier than letting them directly fit the desired underlaying mapping.

And the residual block above explicitly allows it to do precisely that. Because of its compelling results, ResNet quickly became one of the most popular architectures in various computer vision tasks.

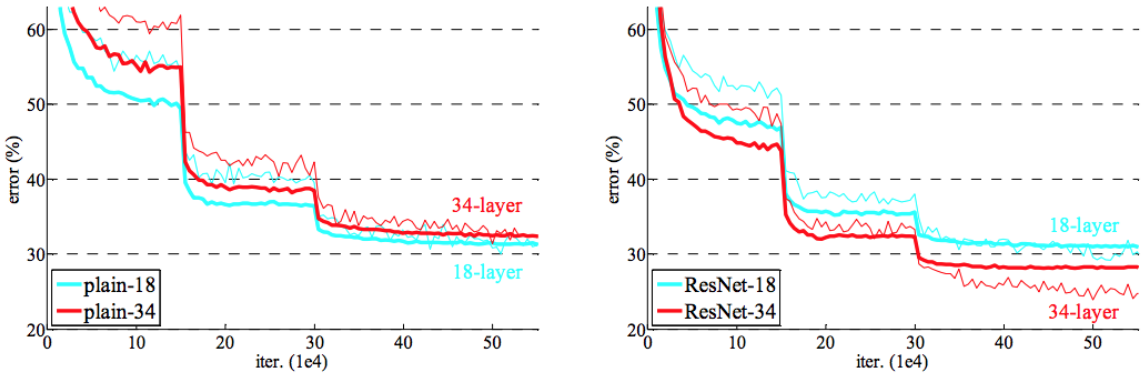

Figure 4.5: Residual Block with skip connection

The following table shows how testing error of different depths:

	plain	ResNet
18 layers	27.94	27.88
34 layers	28.54	25.03

Table 4.3: Testing error over different depths

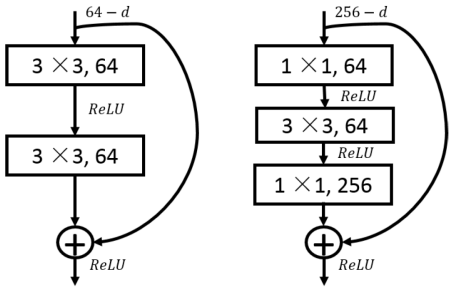

Figure 4.6: A deeper residual function F. Left: a building block as in Figure 4.4 for ResNet-34. Right: a "bottleneck" building block for ResNet-50/101/152/200.

4.2.2 J. M. Dai model

Methodology

Dai and Tong (2018) proposed a variant of residual networks combined with the Dieleman approach for galaxy morphology classification. Instead of working on predicting the results in the manner that Dieleman did with 37 classes from the GZ2 dataset, Dai and Tong (2018) used the data as a result to classify galaxies into five classes i.e. completely round smooth, in-between smooth (between completely round and cigar-shaped), cigar-shaped smooth, edge-on and spiral. The galaxy images were drawn from The Galaxy Zoo challenge (Willett et al., 2013) which contained 61578 images with probabilities that each galaxy is classified into different morphologies. The morphological classifications vote fractions are modifed version of the weighted vote fractions in the Galaxy Zoo 2 project. The classifications vote fractions have high level of agreement and authority with professional astronomers (Willett et al., 2013). Dai and Tong (2018) cleaned the samples in a way where each sample matched a specific morphological category with its appropriate threshold. The thresholds depended on the number of votes for a classification task considered to be sufficient for that morphology. By this means, they assigned galaxy images to five classes, i.e. completely round smooth, in-between smooth(between completely round and cigar-shaped), cigar-shaped smooth, edge-on and spiral. The 5 classes contain a sample of 8434, 8069, 578, 3903 and 7806 respectively. The dataset reduced to 28790 images after filtering, then was divided into training set and testing set by a ratio of 9:1. Thus the network was trained over 25911 images and tested over 2879 images.

Figure 4.7: Example galaxy images from the dataset. Each row represents a class. From top to bottom, their Galaxy Zoo 2 labels are: completely round smooth, in-between smooth, cigar-shaped smooth, edge-on and spiral. Reproduced from (Dai and Tong, 2018).

In order to avoid overfitting, data augmentation is one of the common and effective ways to reduce overfitting. Because of limited training data, data augmentation can enlarge the number of training images. (Dai and Tong, 2018) used five different forms of data augmentation. Scale jittering is the first form of data augmentation. In training time, they crop the images to a range scale S =[170; 240] , which is called multi-scale training images because of the S random value. Since different images can be cropped to different sizes and even the same images also can be cropped to different sizes at different iterations, it is beneficial to take this into account during training. This can be seen as training set augmentation by scale jittering. Random cropping is carried out from 80 x 80 x 3 pixels to 64 x 64 x 3 pixels, which increases the size of training set by a factor of 256. Rotating training images with 0°; 90°; 180°; 270° can enlarge the size of training set by a factor of 4. A horizontal flipping is a doubling of training images.

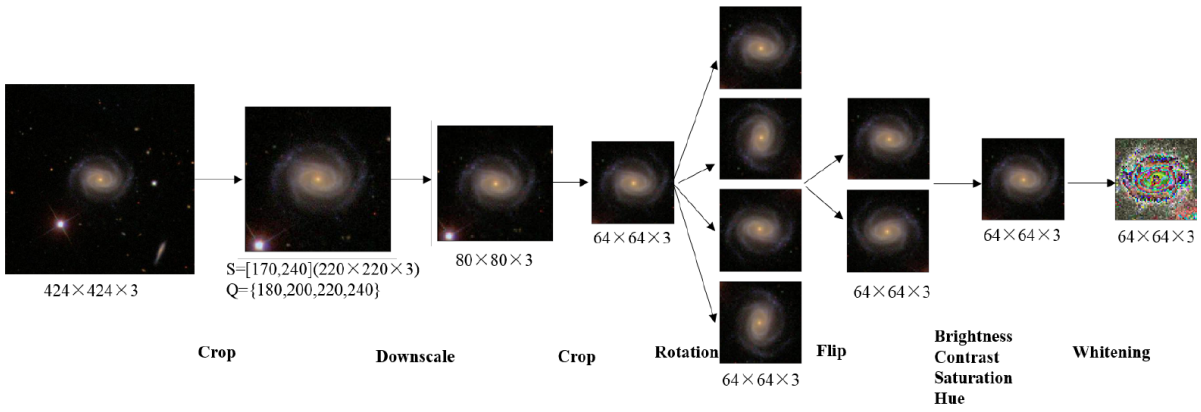

Figure 4.8: Preprocessing procedure. The original image firstly is center cropped to a range scale S = [170; 240] in training set (Q =180, 200, 220, 240 in testing set), for example, the spiral galaxy (GalaxyID:237308) is cropped to 220 x 220 x 3 pixels, then resized to 80 x 80 x 3 pixels, randomly cropped to 64 x 64 x 3 pixels, randomly rotated 0°; 90°; 180°; 270°, and randomly horizontally flipped. After optical distorting and image whitening, it (64 x 64 x 3 pixels) becomes the input of networks.

The model that (Dai and Tong, 2018) came up with was a variant of ResNets V2. They built a network that was specifically designed for galaxies and tried to decrease depth and widen the residual networks. They adopted a full pre-activation residual units and a "bottleneck" building block(Figure 4.6, right) presented in (He et al., 2016b) is used, namely, a combination of 1 x 1; 3 x 3; 1 x 1 convolutions. The full pre-activation includes standard "BN-ReLU-Conv". In addition to these, we add a dropout after 3 x 3 convolution whereas ResNet V2 (He et al., 2016b) did not use dropout to prevent coadaptation and overfitting. The residual unit is defined as:

$x_{l+1} = x_l + W_3\sigma(W_2\sigma(W_1\sigma(x_l)))$.

Here, x_l and x_{l+1} are input and output of the l-th unit, σ denotes BN and ReLU, W_1, W_2, W_3 represent 3 convolutional kernels, dropout is placed after the W_2 operation and the biases are omitted for simplifying notations.

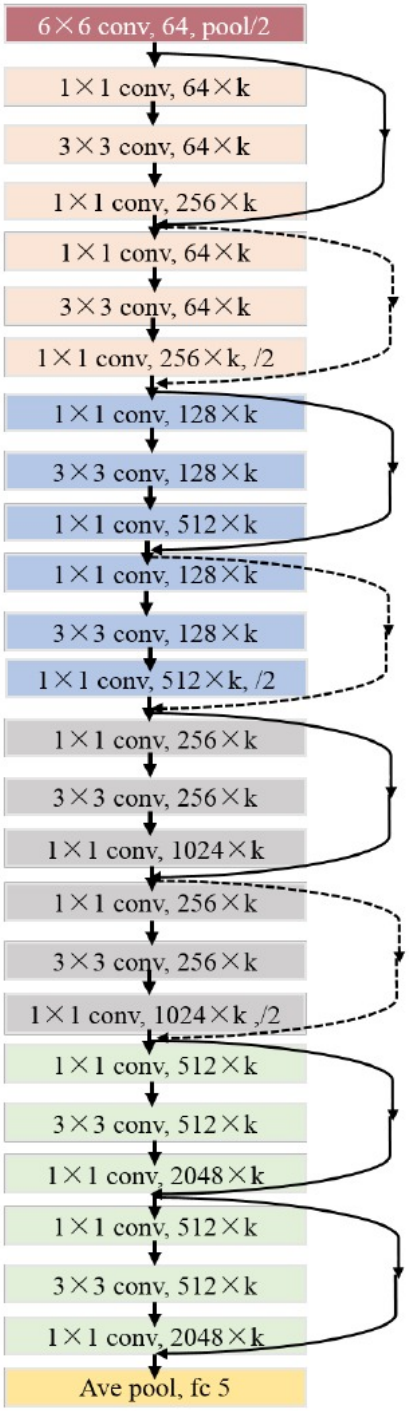

Figure 4.9: Network architecture for Galaxy by (Dai and Tong, 2018) where k is the widening factor.

Results

The 5 classes, i.e. completely round smooth, in-between smooth(betweencompletely round and cigar-shaped), cigar-shaped smooth, edge-on and spiral, are labelled starting from 0 through 5. The results obtained by the network are shown below.

class	Precision	Recall	F1
0	0.9611	0.9634	0.9622
1	0.9561	0.9431	0.9495
2	0.7234	0.5862	0.6476
3	0.9412	0.9485	0.9448
4	0.9573	0.9782	0.9677
Average	0.9512	0.9521	0.9515

Table 4.4: Precision, Recall and F1 for each class on testing set.

	0	1	2	3	4
0	**815**	21	0	0	10
1	29	**762**	0	0	17
2	0	4	**34**	18	2
3	0	3	12	**368**	5
4	4	7	1	5	**763**

Table 4.5: Confusion matrix for each class on testing set. Column represents true label and row represents prediction label.

Chapter 5

Conclusion

The problem of galaxy morphology and their classification has been of major concern for the past century. Modern technology has facilitated growth of multiple techniques which use large computation power. Neural networks have been applied as a classification technology owing to the sheer volume of data. Research conducted so far has shown that accurate classification of galaxies can be made between two classes, namely spiral and elliptical. It has also been observed that increasing the number of classes results in a decreasing accuracy and less precision. Dai and Tong (2018) have attempted to solve the problem by proposing a deep convolution network which combines the model proposed by Dieleman et al. (2015) and the ResNets proposed by He et al. (2016a). The data used from The Galaxy Zoo project (Willett et al., 2013) is used for training. The model has been able to predict the classes with an overall accuracy of 95.2083% over the testing set. The solution is able to use the data from galaxy images and predict the morphology of the galaxy. The resulting prediction allows the classification of galaxy images into one of the five classes proposed in (Dai and Tong, 2018). The results have been far better than all the previous approaches to the problem and seems have formed a promising groundwork for future studies. However, even though the model provides extremely high accuracy, its still expected that the model accuracy shall plummet as the number of classes increases. Further studies can result in even better models that may be then applied to large scale galaxy classification and cluster surveys such as the Large Synoptic Survey Telescope (LSST).

References

Banerji, M., Lahav, O., Lintott, C. J., Abdalla, F. B., Schawinski, K., Bamford, S. P., Andreescu, D., Murray, P., Raddick, M. J., Slosar, A., Szalay, A., Thomas, D., and Vandenberg, J. (2010). Galaxy Zoo: reproducing galaxy morphologies via machine learning*. *Monthly Notices of the Royal Astronomical Society*, 406(1):342–353.

Bazell, D. and Aha, D. W. (2001). Ensembles of classifiers for morphological galaxy classification. *The Astrophysical Journal*, 548(1):219–223.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

Cohen, S., Windhorst, R., Odewahn, S., Chiarenza, C., and Driver, S. (2003). The hubble space telescope wfpc2b-band parallel survey: A study of galaxy morphology for magnitudes 18 ¡= b ¡= 27. *Astronomical Journal - ASTRON J*, 125:1762–1783.

Conselice, C. J. (2003). The Relationship between Stellar Light Distributions of Galaxies and Their Formation Histories. *apjs*, 147(1):1–28.

Dai, J.-M. and Tong, J. (2018). Galaxy morphology classification with deep convolutional neural networks. *Monthly Notices of the Royal Astronomical Society*.

De La Calleja, J. and Fuentes, O. (2004). Machine learning and image analysis for morphological galaxy classification. *Monthly Notices of the Royal Astronomical Society*, 349(1):87–93.

de Vaucouleurs, G. (1959). Classification and Morphology of External Galaxies. *Handbuch der Physik*, 53:275.

Dieleman, S., Willett, K. W., and Dambre, J. (2015). Rotation-invariant convolutional neural networks for galaxy morphology prediction. *mnras*, 450(2):1441–1459.

Djorgovski, S. G., Mahabal, A., Drake, A., Graham, M., and Donalek, C. (2013). Sky surveys. *Planets, Stars and Stellar Systems*, page 223–281.

Edmondson, F. K. (1961). The hubble atlas of galaxies. allan sandage. carnegie institution of washington, washington, d.c., 1961. viii + 32 pp. illus. + 50 plates. 10. *Science*, 134(3477):464–464.

Ferrari, F., de Carvalho, R. R., and Trevisan, M. (2015). MORFOMETRYKA textemdasha NEW WAY OF ESTABLISHING MORPHOLOGICAL CLASSIFICATION OF GALAXIES. *The Astrophysical Journal*, 814(1):55.

Gauci, A., Adami, K., and Abela, J. (2010). Machine learning for galaxy morphology classification. *Monthly Notices of the Royal Astronomical Society*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016a). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016b). Identity mappings in deep residual networks.

Hubble, E. P. (1926). Extragalactic nebulae. *The Astrophysical Journal*, 64.

Huertas-Company, M., Aguerri, J. A. L., Bernardi, M., Mei, S., and Sánchez Almeida, J. (2011). Revisiting the Hubble sequence in the SDSS DR7 spectroscopic sample: a publicly available Bayesian automated classification. *aap*, 525:A157.

Kim, E. J. and Brunner, R. J. (2016). Star–galaxy classification using deep convolutional neural networks. *Monthly Notices of the Royal Astronomical Society*, 464(4):4463–4475.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, page 1097–1105, Red Hook, NY, USA. Curran Associates Inc.

Land, K., Slosar, A., Lintott, C., Andreescu, D., Bamford, S., Murray, P., Nichol, R., Raddick, M. J., Schawinski, K., Szalay, A., Thomas, D., and Vandenberg, J. (2008). Galaxy Zoo: the large-scale spin statistics of spiral galaxies in the Sloan Digital Sky Survey*. *Monthly Notices of the Royal Astronomical Society*, 388(4):1686–1692.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

Lintott, C., Schawinski, K., Bamford, S., Slosar, A., Land, K., Thomas, D., Edmondson, E., Masters, K., Nichol, R. C., Raddick, M. J., and et al. (2010). Galaxy zoo 1: data release of morphological classifications for nearly 900000 galaxies. *Monthly Notices of the Royal Astronomical Society*, 410(1):166–178.

Lintott, C. J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M. J., Nichol, R. C., Szalay, A., Andreescu, D., Murray, P., and Vandenberg, J. (2008). Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *mnras*, 389(3):1179–1189.

Lotz, J. M., Primack, J., and Madau, P. (2004). A New Nonparametric Approach to Galaxy Morphological Classification. *aj*, 128(1):163–182.

Mairal, J., Koniusz, P., Harchaoui, Z., and Schmid, C. (2014). Convolutional kernel networks. In *NIPS*.

Naim, A., Lahav, O., Sodré, L., J., and Storrie-Lombardi, M. C. (1995). Automated morphological classification of APM galaxies by supervised artificial neural networks. *Monthly Notices of the Royal Astronomical Society*, 275(3):567–590.

Owens, E. A., Griffiths, R. E., and Ratnatunga, K. U. (1996). Using oblique decision trees for the morphological classification of galaxies. *Monthly Notices of the Royal Astronomical Society*, 281(1):153–157.

Schawinski, K., Lintott, C., Thomas, D., Sarzi, M., Andreescu, D., Bamford, S. P., Kaviraj, S., Khochfar, S., Land, K., Murray, P., Nichol, R. C., Raddick, M. J., Slosar, A., Szalay, A., VandenBerg, J., and Yi, S. K. (2009). Galaxy Zoo: a sample of blue early-type galaxies at low redshift*. *Monthly Notices of the Royal Astronomical Society*, 396(2):818–829.

Sérsic, J. and de Córdoba. Observatorio Astronómico, U. N. (1968). *Atlas de galaxias australes*. Observatorio Astronomico, Universidad Nacional de Cordoba.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Storrie-Lombardi, M., Lahav, O., Jr, S., and Storrie-Lombardi, L. (1992). Morphological classification of galaxies by artificial neural networks. *Monthly Notices of the Royal Astronomical Society*, 259:8P.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2014). Going deeper with convolutions.

van den Bergh, S. (1976). A new classification system for galaxies. *apj*, 206:883–887.

Willett, K. W., Galloway, M. A., Bamford, S. P., Lintott, C. J., Masters, K. L., Scarlata, C., Simmons, B. D., Beck, M., Cardamone, C. N., Cheung, E., Edmondson, E. M., Fortson, L. F., Griffith, R. L., Häußler, B., Han, A., Hart, R., Melvin, T., Parrish, M., Schawinski, K., Smethurst, R. J., and Smith, A. M. (2016). Galaxy Zoo: morphological classifications for 120000 galaxies in HST legacy imaging. *Monthly Notices of the Royal Astronomical Society*, 464(4):4176–4203.

Willett, K. W., Lintott, C. J., Bamford, S. P., Masters, K. L., Simmons, B. D., Casteels, K. R. V., Edmondson, E. M., Fortson, L. F., Kaviraj, S., Keel, W. C., and et al. (2013). Galaxy zoo 2: detailed morphological classifications for 304122 galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 435(4):2835–2860.

Willett, K. W., Schawinski, K., Simmons, B. D., Masters, K. L., Skibba, R. A., Kaviraj, S., Melvin, T., Wong, O. I., Nichol, R. C., Cheung, E., Lintott, C. J., and Fortson, L. (2015). Galaxy Zoo: the dependence of the star formation–stellar mass relation on spiral disc morphology. *Monthly Notices of the Royal Astronomical Society*, 449(1):820–827.