UNIVERSIDAD DE CONCEPCIÓN
DIRECCIÓN DE POSTGRADO
FACULTAD DE INGENIERÍA - PROGRAMA DE MAGÍSTER EN CIENCIAS DE LA COMPUTACIÓN

A debiasing framework for deep learning applied to the morphological classification of galaxies

Tesis para optar al grado de
MAGÍSTER EN CIENCIAS DE LA COMPUTACIÓN

Por: Esteban Medina Rosales
Profesor guía: Guillermo Cabrera Vives
Comisión interna: Pierluigi Cerulo
Comisión externa: Christopher Miller

Concepción, Julio de 2023

Abstract

In order to train deep learning models, usually a large amount of correctly annotated data is needed. Depending on the data domain, the task of correctly annotating data can prove to be difficult, as in many cases the ground truth of the data is not obtainable. This is true for numerous problems within the astronomy domain, one of these being the morphological classification of galaxies. The aforementioned means that astronomers are forced to rely on an estimate of the ground truth, often generated by human annotators. The problem with this is that human generated labels have been shown to contain biases related to the quality of the data being labeled, such as image resolution. This type of bias is a consequence of the quality of the data, that is, it is independent of the annotators, meaning that even datasets annotated by experts can be affected by this type of bias. In this work, we show that deep learning models trained on biased data learn the bias contained in the data, transferring the bias to its predictions. We also propose a framework to train deep learning models, that allows us to obtain unbiased models even when training on biased data. We test our framework by training a classification model on images of morphologically classified galaxies from Galaxy Zoo 2 and show that we are able to diminish the bias in the data.

Contents

List of Figures

List of Tables

1 Introduction

In the last two decades there have been considerable advances in the area of machine learning, which together with the rapid improvement of technology have led to a considerable amount of real-world problems being solved using these algorithms. In particular, the supervised learning paradigm has proven to be highly effective, obtaining state of the art performance in a large number of problems. Supervised learning refers to the task of learning a function that maps an input to a target output, based on examples of input-output pairs. More specifically, classification problems lie within the domain of supervised learning, where the task is classifying objects, described by some input, into a specific class. More generally, we consider a training set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ where $x_i \in \mathcal{X}$ corresponds to the inputs and $y_i \in \mathcal{Y}$ corresponds to the outputs or labels. $\mathcal{D}$ is used to learn a function $f(x) : \mathcal{X} \to \mathcal{Y}$ that correctly maps the inputs x_i to the labels y_i.

When working in real-world problems, it can be very difficult to obtain the ground truth labels y_i needed to train a classification model. In a large number of problems, these ground truth labels are not directly obtainable, but an estimate $\hat{y}_i$ can be obtained by human annotators (e.g., an annotator could review a set of images and manually assign a label to each of them). The problem with the previously mentioned procedure is that these $\hat{y}_i$ labels may be biased due to poor quality of the observations being labeled by the annotators. An example of this may be a set of low resolution images where the key classification features are not clearly distinguishable. This can lead to some of these images being systematically mislabeled, introducing a bias associated with the quality of the observations, independent of the annotators.

The aforementioned problem is particularly relevant in the astronomy field, more specifically in the morphological classification of galaxies. The seminal example of a galaxy morphological classification scheme is the Hubble sequence [17], which distinguishes galaxies into one of four classes: elliptical, lenticular, spiral or irregular. Over the time, new classification schemes have been proposed, such as the De Vaucouleurs system [6], which introduces new and more detailed classes. Being able to classify galaxies according to their morphology is important, since galaxy morphology have been shown to be linked with intrinsic properties such as gas content, brightness, color [10], and star formation [24, 36].

For some time, visual classification played the dominant role in galaxy morphological studies, either done by expert astronomers [7, 8, 3, 11, 31, 27, 20] or through crowdsourcing systems such as Galaxy Zoo [26, 1, 25, 37, 35, 38]. Recently, new datasets of morphologically classified galaxies obtained using diverse machine learning methods have been published, many of them using supervised learning [13, 18, 9, 19, 21, 40], which requires labeled datasets for training. The problem is that human generated labels have been shown to be biased in terms of observable parameters, resulting in low resolution galaxies being biased towards smoother types because the fine structure of these galaxies can not be distinguished by human annotators [4, 37](Fig. 1). Furthermore, deep learning models have shown to learn human biases that are present in the training data, making this problem even more important, given the growing popularity of these models in astronomical applications and that usually the only available labels were generated by humans [22].

Figure 1: Example of biased classification. (a) A galaxy image as shown to the annotators, taken from the Earth's surface. (b) The same galaxy image zoomed in. (c) The same galaxy at a higher resolution taken from above the Earth's atmosphere. It can be noticed that the features present in (c) are not clearly distinguishable in (a) and (b), even when zooming in. The majority of the annotators classified this galaxy as smooth with no signs of a disk, even though the disk features are easily distinguishable when looking at the high resolution image (c).

2 Background

2.1 Artificial neural networks

Artificial neural networks are models originally inspired by biological neural networks, these models seek to mathematically model their functioning. The basic unit of a neural network is the perceptron [29], defined as:

$$\hat{y}(x, w) = f(\sum_{i=1}^{n} w_i x_i + b), \tag{1}$$

where x_i corresponds to the input, n to the number of inputs, w_i to the parameters of the perceptron and b is a bias factor that allows one to move the decision boundary away from the origin. The function f is called the activation function and can be linear or nonlinear depending on the problem. The parameters w_i are adjusted during training. One of the first neural networks designed is the feed-forward neural network, which consists of connecting multiple perceptrons. These neural networks group the perceptrons into layers, through which the data flows sequentially (Figure 2) until they reach the output layer. A neural network with multiple hidden layers is called a deep neural network. To learn the w_i parameters of a neural network, a loss function, which is defined according to the problem, is progressively minimized in each iteration. This loss function measures the model error by comparing the output delivered by the model with the target output. An example of a loss function is the mean square error, defined as:

$$ECM = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{2}$$

where y_i is the target output for object i, $\hat{y}_i$ is the output delivered by the model for object i and n is the number of objects. To update the model parameters, the gradient backpropagation algorithm [30] is used. This algorithm computes the gradient of the loss function with respect to each parameter w_i of the model using the chain rule, starting from the last layer toward the first and updates the parameters after each iteration. The later is done by using optimization algorithms such as the stochastic gradient descent [28].

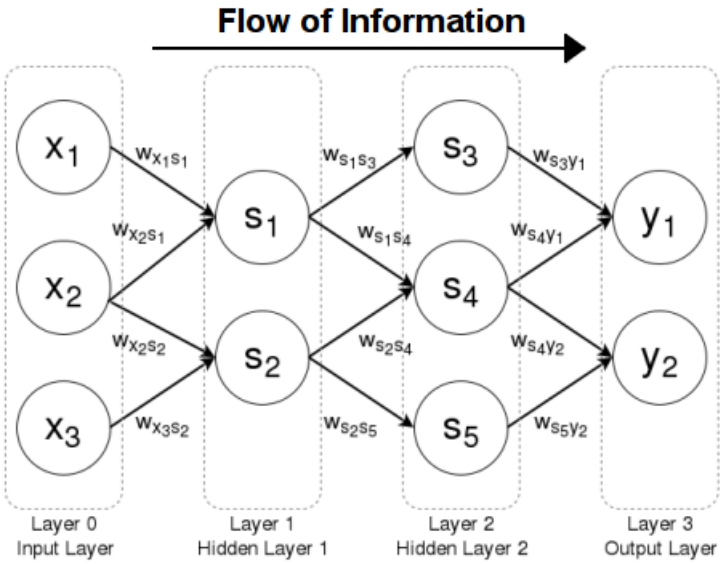


Figure 2: Feed-forward neural network. The perceptrons are grouped into layers called hidden layers. The connections between each perceptron correspond to the w_i parameters of the network. Data is passed from the input layer sequentially to the output layer. The number of hidden layers, perceptrons, and connections varies with the problem, as do the activation functions. **Source:** John McGonagle. Feedforward Neural Networks. https://brilliant.org/wiki/feedforward-neural-networks

2.1.1 Convolutional neural networks

A convolutional neural network (CNN) [12, 23] is a type of feed-forward neural network with multiple layers. Originally designed to work on images, allows better feature extraction than traditional neural networks. CNNs have three main components: convolutional layers, pooling layers and fully connected layers. In the convolutional layers, convolution operations are performed on the input (typically images); these consist of a filter that runs through the input performing matrix multiplications between the filter and the data at each location. The values of these filters are adjusted during training. The result of this is a feature map of smaller dimension than the original input, this final dimension depending on both the size of the input and the size of the convolution filter. In pooling layers the dimension of the data is further reduced, this is done on the basis of windows that run through the input selecting the largest value within this window (max pooling) or the average value (average pooling). Convolutional layers and pooling layers are typically used in pairs, where a pooling layer goes after each convolutional layer. After these pairs of layers, fully connected layers are

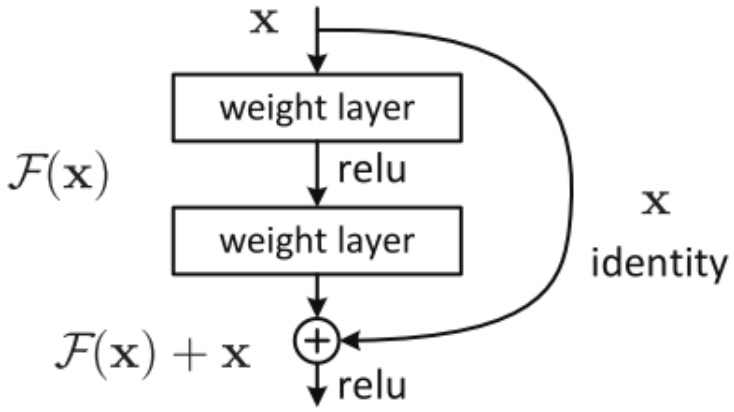


Figure 3: Residual block of a residual neural network. Note the connection that skips two layers of the network. **Source:** [16]

.

used that take as input the features extracted in the previous layers to perform the objective task.

2.1.2 Residual neural networks

Residual neural networks [16] were designed as an extension of convolutional neural networks. These types of networks solve two major problems presented by traditional convolutional networks: vanishing gradients and accuracy degradation. The vanishing gradients phenomenon is usually a problem in networks with a large number of layers; it is a consequence of the gradient backpropagation algorithm, which, by successively computing the gradients of each parameter in the network, reaches a point where the values are so small that the network stops learning (the parameters stop updating). This problem has been partially solved with different strategies [23, 14]. On the other hand, accuracy degradation is a more complex problem to solve. Theoretically, adding more layers to a convolutional network should improve its performance by being able to obtain better feature maps, or at least should obtain the same results, since, assuming that an optimal mapping is reached in fewer layers, the layers that follow should do an identity mapping, i.e., simply copy the output of the previous layers, functioning as if these extra layers did not exist. However, experiments have shown that increasing the depth of the network degrades the accuracy of the network compared to its shallower counterpart [16]. To solve this problem, [16] proposed to add an additional connection that skips one or more layers of the network (Figure 3); thus, if x is the input of these layers and $\mathcal{F}(x)$ is the output, by means of this new connection, we obtain as output $\mathcal{H}(x) = \mathcal{F}(x)+x$. Reformulating the above, we obtain the residual function $\mathcal{F}(x) = \mathcal{H}(x)-x$. The motivation for this is the accuracy degradation, as this suggests that the model is not capable of identity mapping across multiple nonlinear layers, such as those of a traditional CNN. Using this formulation, when the optimum for a set of layers is to perform an identity mapping, the model can simply approximate the weights of the intermediate layers to zero, which should be easier than approximating the identity mapping directly.

For this work, we use a ResNet50 model. ResNet50 is a particular implementation of

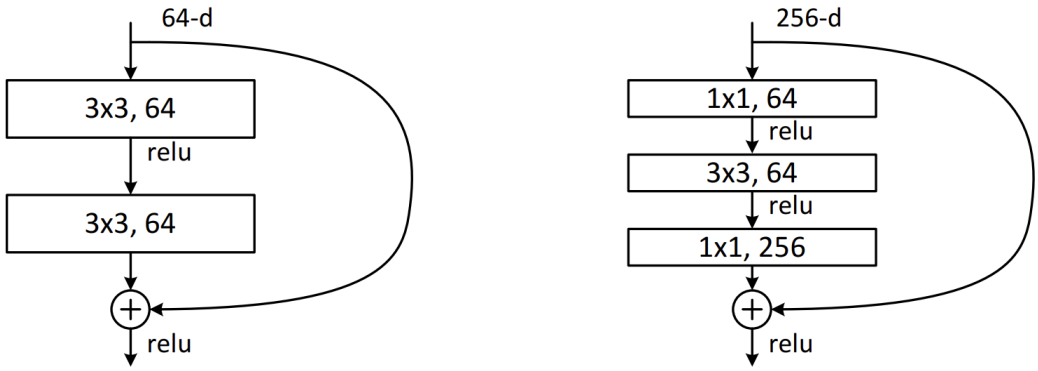


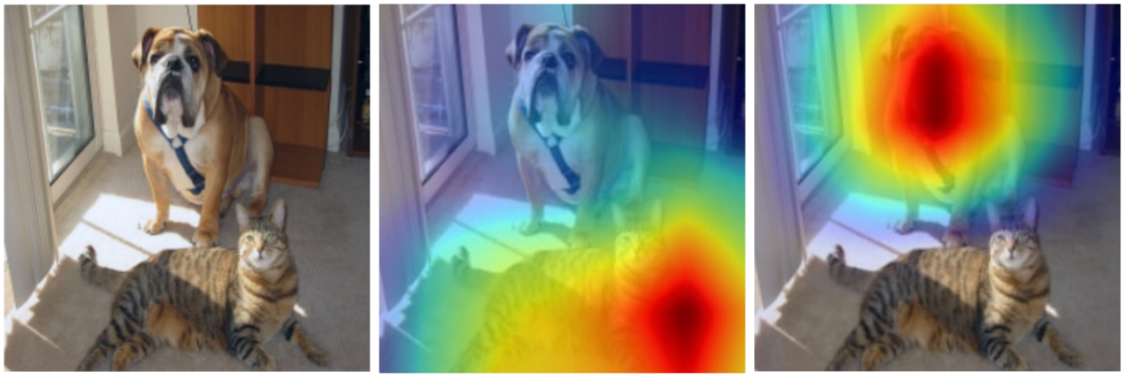
Figure 4: Residual block for ResNet34 (left) and bottleneck residual block (right) as implemented for deeper Residual Neural Networks including ResNet50, ResNet101 and ResNet152. **Source:** [16]

.

a Residual Neural Network where the number on its name indicates the number of layers considered by the architecture, that is, the ResNet50 is a 50 layers deep neural network. The main difference between ResNet50 and its shallower counterparts (e.g. ResNet34) is that the residual block (Figure 3) is modified to include 3 layers instead of 2. The purpose of this is to mitigate the increase of the training time that comes with the addition of extra layers. This new residual block follows a bottleneck design; the three layers consecutively perform 1x1, 3x3 and 1x1 convolutions, where the first layer decreases the dimension of the input that goes into the 3x3 layer and the last layer increases the dimension of the output (Figure 4).

2.2 Grad-CAM

The Gradient weighted Class Activation Mapping (Grad-CAM) [32] is a technique designed to understand the decision making process of convolutional neural networks. CNNs (and most deep learning models) are usually seen as a black box, that is, the conclusions are drawn from the input and the output of the networks, since the model behavior can not be comprehended by looking at the structure and weights alone. Grad-CAM uses the gradients of a given target class (e.g. dog in a classification problem) flowing to the last convolutional layer to compute a localization map that highlights the most important sections of the input for predicting the target class (Figure 5). The aforementioned helps to visualize, for this example, that the model is using mostly the face of the animal to decide if it corresponds to a dog or a cat. Extrapolating this to other problems can provide important information to understand why a model is working the way it is.



(a) (b) (c)

Figure 5: Example of Grad-CAM localization map from a ResNet model trained for classification. (a) The original image. (b) The original image and the localization map for the target class cat. (c) The original image and the localization map for the target class dog. Red regions represent the important class-discriminative sections of the image for the target class. **Source:** [32]

.

2.3 Astronomical concepts

2.3.1 Galaxy morphological classification

Galaxy morphological classification consists of grouping galaxies into different categories according to their visual appearance. The morphological classifications are widely used by astronomers, the most famous of them being the Hubble Sequence [17]. This classification scheme divides galaxies into three classes: elliptical, lenticular, and spiral (Figure 6). Galaxies that do not belong to any of these classes are called irregular galaxies. Later, De Vaucouleurs proposed an extension to the Hubble Sequence, known as the De Vaucouleurs System [6] where he introduced a more elaborate division within the category of spirals, based on three characteristics: bars, rings and spiral arms (Figure 7).

2.3.2 Redshift

The Redshift is a measure of the increase in the wavelength of an electromagnetic wave caused by the fact that the source that emits the radiation is moving away from the observer. The fact that space is continuously expanding generates an effect comparable to two objects constantly moving away from each other. This causes the spectrum of an astronomical object to stretch and increase in wavelength. This increase in wavelength, and thus decrease in frequency, causes the light to become redder. A larger shift means that the light source is moving away at a greater speed due to the expansion of the universe. It can be demonstrated that the farther an astronomical object is from the observer, the highest will be its redshift. Thus, the redshift can be interpreted as a measure of how far an object is from the observer.

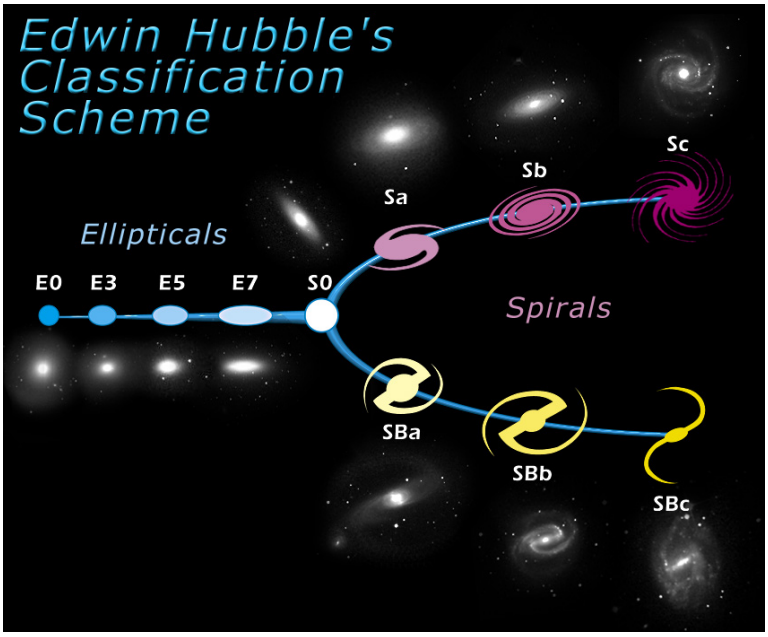


Figure 6: Diagram of the Hubble sequence. S0 type galaxies correspond to lenticular galaxies. **Source:**
https://en.wikipedia.org/wiki/Galaxy_morphological_classification.

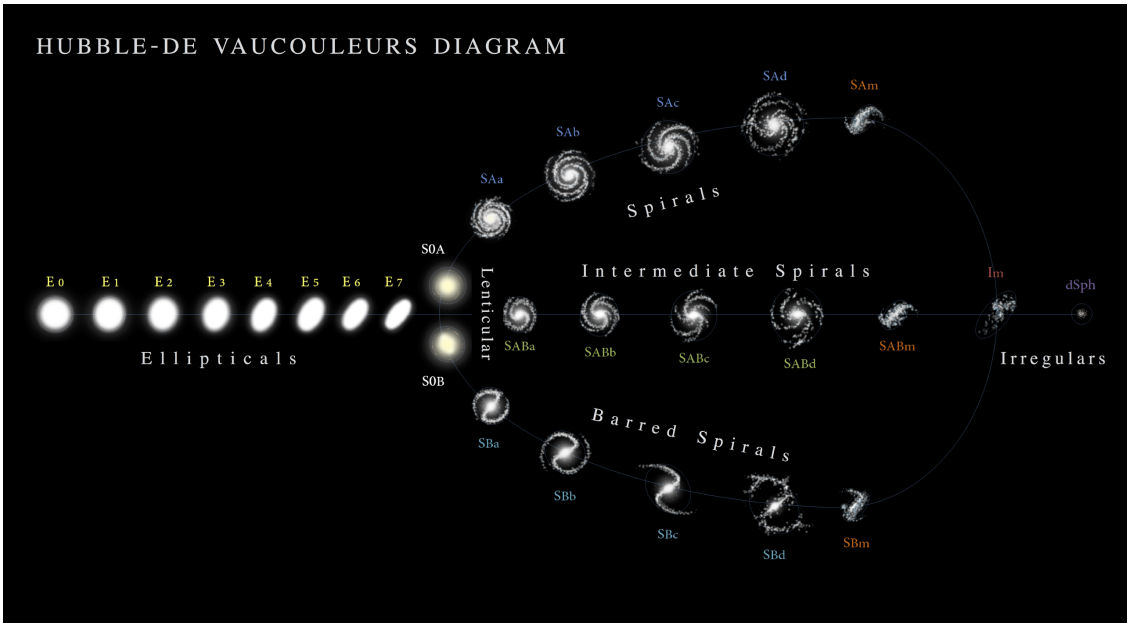


Figure 7: Diagram of the de Vaucouleurs system. It can be noticed that the classes of the Hubble sequence are kept, but with a more elaborate classification system for each of them. **Source:**
https://en.wikipedia.org/wiki/Galaxy_morphological_classification.

2.3.3 Point spread function (PSF)

The point spread function (PSF) can be thought of as a function that characterizes the quality of an imaging system, such as a telescope. The image of a point source, for example, a very distant star, will never look exactly like a point in the images generated by an instrument; it will usually look like a roughly circular object instead, in which diffraction patterns can be distinguished. The point spread function characterizes how much the telescope optics and the atmosphere (in the case of observations taken from the Earth) have redistributed the light captured by the instrument. Thus, the width of the point spread function can be interpreted as the resolution of the instrument and, in the context of this work, as the resolution of a telescope. The greater the width of this function the lower the resolution, and therefore the lower the quality of the images taken by it.

2.3.4 Galaxy properties

Galaxy properties are physical quantities that characterize the physical state of a galaxy at a given time. Examples of them are the size, the brightness, the stellar mass, the chemical composition, the star formation rate and the molecular or atomic gas mass. By using these properties, astronomers are able to study the evolution of galaxies which is a key part to understand how the universe is built and how it evolves. Galaxy properties, in the context of this work, can be divided into intrinsic and observed.

2.3.4.1 Observed properties

An observed property is defined as a property that depends on the observation, i.e., it can change as a function of the telescope with which it was observed as well as a function of the galaxy's inclination. In this work, the following observed properties are used:

- Petrosian angular radius: Also known as angular size, it is the apparent radius of the galaxy as seen by an observer on the Earth. It is measured in arcseconds.
- Apparent magnitude: The apparent magnitude is a measure of the flux of a celestial object observed from the Earth, it depends on both the distance of the object from the Earth and the extinction of the light produced by stardust in the line of sight of the observer, i.e., it is an observed property. In particular for galaxies, there exist multiple ways of calculating the total flux, in our case we use the Petrosian flux. The Petrosian flux F_p is defined as the flux within a certain number N_p ($N_p = 2$ for SDSS data) of Petrosian radii r_p; this is:

  $$F_p = \int_0^{N_p r_p} 2\pi r' dr' I(r'),$$

  where I(r) is the surface brightness profile.
- Ellipticity: Corresponds to the ellipticity of the galaxy. It is defined as $e = 1 - b/a$, where b is the semi-minor axis and a is the semi-major axis. In simple words, it is a measure of how elliptical a galaxy's shape is: values close to 0 correspond to circular-shaped galaxies and values close to 1 would suggest a disk-shaped galaxy.

2.3.4.2 Intrinsic Properties

An intrinsic property is defined as a property that is independent of observation. This means that it is independent of both the quality of the telescope with which the observation is made and the inclination of the galaxy. The intrinsic properties used in this work are:

- Physical radius: Corresponds to the Petrosian radius as described in [39]. It is defined as the radius at which the Petrosian ratio R_p is equal to a given value, in this case this value corresponds to 0.2 as defined by the Sloan Digital Sky Survey (SDSS). The Petrosian ratio at a radius r from the center of an object is defined by [39] as the ratio of the local surface brightness averaged over an annulus at r to the mean surface brightness within r. This is:

$$R_p(r) = \frac{\int_{\alpha_{lo}r}^{\alpha_{hi}r} dr' 2\pi r' I(r') / [\pi(\alpha_{hi}^2 - \alpha_{lo}^2) r^2]}{\int_0^r dr' 2\pi r' I(r') / (\pi r^2)},$$

  where r is the radius, $I(r)$ is the azimuthally averaged surface brightness profile and α_{lo}, α_{hi} define the annulus. In this case $\alpha_{lo} = 0.8$ and $\alpha_{hi} = 1.25$ as defined by the SDSS. In simpler words, it is the distance from the center of the galaxy to the point where the surface brightness reaches a given threshold. This distance is measured in kiloparsecs.

- Absolute magnitude: The absolute magnitude is defined as the apparent magnitude that a galaxy would have if it were observed at a distance of 10 parsecs and without light extinction. This quantity allows one to compare the luminosity of different objects directly.

- Effective radius: Also known as the half-light radius, it is defined as the radius in which half of the galaxy light is contained.

2.3.5 Sérsic profile

The Sérsic profile [33] is a function that describes how the instensity of a galaxy varies as a function of projected distance r from its center. It is defined as:

$$I(r) = I_e \exp\{-b_n[(\frac{r}{r_e})^{1/n} - 1]\}, \tag{3}$$

where r_e is the effective radius, I_e is the intensity at the effective radius, n is the Sérsic index, and b_n is a constant that is defined as a function of n and r_e, used to ensure that r_e corresponds to the radius containing half of the luminosity of the galaxy. Fitting this function to a galaxy provides relevant information about it, in particular for this work the Sérsic index is of special interest.

3 Related work

The type of bias we address in this work, specifically related to galaxy morphologies, has been extensively studied by the Galaxy Zoo team, [1] quantified a luminosity, size and redshift

dependent bias present in Galaxy Zoo (GZ1) [26] data. They also corrected the vote fractions obtained from the crowdsourcing system by assuming that the morphological fraction does not evolve with redshift within bins of fixed galaxy physical size and luminosity. Later, [37] adapted this technique to Galaxy Zoo 2 (GZ2) data, taking in consideration that GZ2 uses a decision tree rather than a single question (as GZ1 does), meaning that all tasks but the first one depend on responses to tasks higher in the decision tree. After that, [15] presented an improved method that addresses the questions on the GZ2 decision tree with multiple responses (such as number of spiral arms), showing that the method from [37] does not always adjust the vote fractions correctly for this type of tasks. Their method aims to make the vote distributions consistent at different redshift rather than the mean vote fractions values as in [1] and [37]. A new method to calibrate morphologies for galaxies of different luminosities and at different redshifts applied to data from the Hubble Space Telescope was introduced in [38]; they used artificially redshifted images as a baseline in order to correct the galaxy morphologies.

Outside of Galaxy Zoo, [5] introduced a metric to estimate the observational bias in a dataset also focused in galaxy morphologies. They assume that the fractions of objects of each class in an unbiased dataset should not be significantly different for labels with different resolutions within bins of intrinsic parameters. Closer to this work, [4] and [2] address this problem through a machine learning approach, simultaneously learning a classification model, estimating the intrinsic biases in the ground truth, and providing new de-biased labels.

4 Methodology

4.1 Labeling process

We consider a binary classification task where we classify galaxies according to the first level of the Galaxy Zoo 2 classification tree (Figure 8). The first class corresponds to smooth featureless galaxies with a rounded ellipsoidal shape, also known as ellipticals. The second class corresponds to galaxies that posses a disk and present features such as spiral arms, including both spiral and lenticular galaxies. For purposes of this work, these classes will be addressed as smooth and disk respectively. We define our labels by using Galaxy Zoo 2 weighted vote fractions corresponding to the first task of the tree, labeling galaxies as elliptical when the majority of the votes for the first task correspond to *smooth*, as disk when the majority of the votes correspond to *features or disk* and we discarded galaxies corresponding to *star or artifact*.

4.2 Galaxy morphology labeling bias definition

In order to model the labeling bias contained in the data, we follow [4] and model this bias in terms of the biasing parameter α. This parameter α corresponds to the resolution of the galaxy image measured as the angular Petrosian radius of the galaxy over the angular size of the point spread function (PSF). As mentioned before, the bias we are addressing in this work is a result of low resolution disk galaxies being mislabeled as smooth galaxies, since the distinctive features of this galaxies are not distinguishable by the annotators. We follow

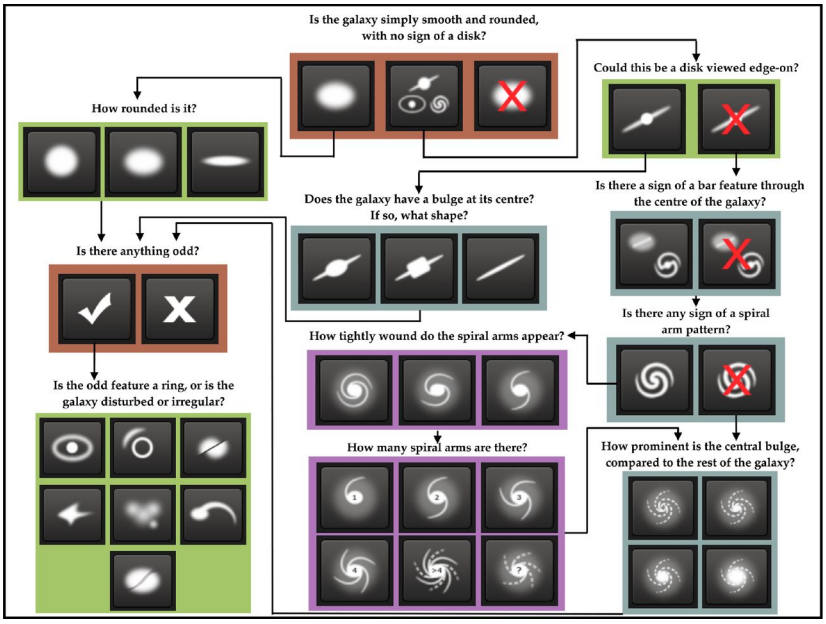


Figure 8: Galaxy Zoo 2 decision tree used to classify galaxies according to their morphology. We use the weighted vote fractions of the first task to generate our binary labels. Task other than the first one are not used in our labeling process. **Source:** [37]

[4], and model this bias as

$$p(\tilde{y} = \text{smooth}|y = \text{disk}, \alpha) = e^{-\alpha^2/(2\theta^2)}, \tag{4}$$

where θ is a parameter to be fitted. Notice that $\lim_{\alpha\to 0} p(\tilde{y} = \text{smooth}|y = \text{disk}, \alpha) = 1$ and $\lim_{\alpha\to\infty} p(\tilde{y} = \text{smooth}|y = \text{disk}, \alpha) = 0$, that is, we expect low resolution disk galaxies to always be mislabeled as smooth galaxies by the annotators and, at the same time, we assume that high resolution disk galaxies are never mislabeled. On the other hand, we assume that smooth galaxies are never mislabeled as disk galaxies, given that a smooth featureless galaxy will still have the appearance of a smooth featureless galaxy at low resolution, thus we define

$$p(\tilde{y} = \text{disk}|y = \text{smooth}, \alpha) = 0. \tag{5}$$

From equation (5) we can deduce that $p(\tilde{y} = \text{smooth}|y = \text{smooth}, \alpha) = 1$, that is, we do not expect smooth galaxies to be mislabeled.

4.3 De-biasing method

Consider a human labeled dataset $\mathcal{D} = \{(x_i, \tilde{y}_i)\}_{i=1}^N$ composed of pairs $(x_i, \tilde{y}_i)$ of features x_i and human generated labels $\tilde{y}_i$. We assume each of these pairs to be sampled independently from a data distribution $p_{\text{bias}}(x, \tilde{y})$ defined over $\mathcal{X} \times \mathcal{Y}$. We also assume the existence of an unknown *ground truth* label $y_i \in \mathcal{Y}$ for each biased label $\tilde{y}_i$. We consider a supervised classification task, that is, we want to approximate a function $f_w : \mathcal{X} \to \mathcal{Y}$ that maps the input features to the latent ground truth labels, with parameters w to be fitted. For this, we consider a biasing parameter α (e.g. the resolution of the labeled image). In

this work we consider a maximum likelihood approach, consider a classification problem where $\mathcal{Y} = 1, \ldots, K$, the likelihood of the data given the model parameters and the biasing parameter is

$$p(\mathcal{D}|w, \{\alpha_i\}_{i=1}^N) = \prod_{i=1}^N p(\tilde{y}_i|x_i, w, \alpha_i), \tag{6}$$

$$= \prod_{i=1}^N \sum_{y_i} p(\tilde{y}_i, y_i|x_i, w, \alpha_i), \tag{7}$$

$$= \prod_{i=1}^N \sum_{y_i} p(\tilde{y}_i|y_i, \alpha_i) p(y_i|x_i, w), \tag{8}$$

where the sum in equation (8) runs over the possible values of y_i, and we assumed that the biased labels $\tilde{y}_i$ only depends on the true labels y_i and the biasing parameters of each object α_i, while the true label is inferred from the features x_i and the parameters of the model w.

Notice that $p(\tilde{y}|y, \alpha)$ models the biasing process by assuming the biased labels $\tilde{y}$ do not depend directly on the features x. At the same time, $p(y|x, w)$ models the dependence of y with x, making $\tilde{y}$ and x conditionally independent given y.

From equation (8) notice that $p(\tilde{y}_i|y_i, \alpha_i)$ corresponds to the bias model discussed in Section 4.2 and that $p(y_i|x_i, w)$ represents a probabilistic classification model. Both the bias model and the classification model can be changed depending on the problem. For this work, we use a neural network with parameters w as the classification model and the model introduced in Section 4.2 as the bias model.

4.4 Estimating θ

We define the class fraction of a given class as the number of objects of said class over the total number of objects within an interval. For this work, we define the smooth class fraction $f_s = N_s/N$ and the disk class fraction $f_d = N_d/N$ in a given interval, such that $f_s + f_d = 1$ and $N_s + N_d = N$. Where N corresponds to that total number of objects in the interval, N_s to the total number of objects of the smooth class in the interval and N_d to the total number of objects of the disk class in the interval. We also define $\tilde{f}_s$ and $\tilde{f}_d$ as their respective estimated values, that is, the values of f_s and f_d we obtain from the biased data.

Following [4], we assume that an unbiased dataset should be uniformly distributed in terms of its observable properties, i.e., f_s and f_d should not vary in terms of α. Considering this, we group galaxies in bins of α and calculate the corresponding $\tilde{f}_s$ and $\tilde{f}_d$ for each bin.

4.4.1 Derivation

The ground truth class fraction can be expressed as

$$f_s = p(y = \text{smooth}), \tag{9}$$

$$f_d = p(y = \text{disk}), \tag{10}$$

and the estimated class fractions as

$$\tilde{f}_s = p(\tilde{y} = \text{smooth}|\alpha), \tag{11}$$
$$\tilde{f}_d = p(\tilde{y} = \text{disk}|\alpha). \tag{12}$$

Expanding equation (11)

$$\begin{aligned}
\tilde{f}_s &= p(\tilde{y} = \text{s}|\alpha), && (13)\\
&= \sum_{y \in \{\text{s,d}\}} p(\tilde{y} = \text{s}, y|\alpha), && (14)\\
&= \sum_{y \in \{\text{s,d}\}} p(\tilde{y} = \text{s}|y, \alpha)p(y|\alpha), && (15)\\
&= p(\tilde{y} = \text{s}|y = \text{s}, \alpha)p(y = \text{s}|\alpha) + && (16)\\
&\quad p(\tilde{y} = \text{s}|y = \text{d}, \alpha)p(y = \text{d}|\alpha),
\end{aligned}$$

where s and d correspond to smooth and disk respectively. Considering that $p(\tilde{y} = \text{s}|y = \text{s}, \alpha) = 1$ (smooth galaxies are always labeled as smooth) and that the ground truth label y is independent of α, we obtain

$$\tilde{f}_s = f_s + f_d \cdot p(\tilde{y} = \text{smooth}|y = \text{disk}, \alpha). \tag{17}$$

4.4.2 Estimation of θ

In order to estimate θ we minimize the square of the difference between f_s and $\tilde{f}_s$ with respect to θ, f_s and f_d. This is

$$\min_{\theta, f_s, f_d} \sum_i (f_s - \tilde{f}_s)^2, \tag{18}$$

where the sum iterates over the bins in α. We use equation (17) for $\tilde{f}_s$. Given that we do not have access to the ground truth class fractions, we initialize f_s and f_d using the values of $\tilde{f}_s$ and $\tilde{f}_d$ from the least biased bin in terms of α. For the value of α corresponding to each bin we use the midpoint of the bin. Notice that the number of bins is a hyperparameter; for this work we grouped the galaxies in 1525 bins with the same number of galaxies in each of them.

5 Experiments and results

5.1 Models implemented

We implemented two de-biasing models using the same bias model and a different classification model. Following [4], we implemented the first model using a logistic regression as the classifier and we used the Sérsic index [33], the ellipticity, and the half-light radius as classification features. To train this model we follow [4] and maximize the log-likelihood introduced in Section 4.3, we do this by using the Expectation-Maximization algorithm, this allows us to estimate the parameters of the logistic regression as well as the parameter θ of

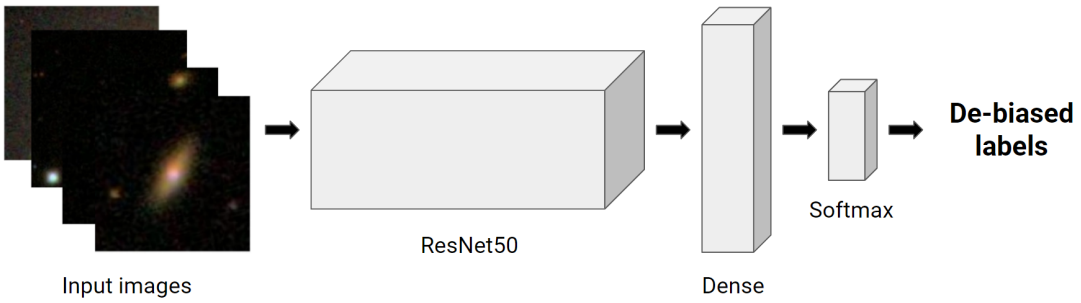


Figure 9: Architecture of our deep learning de-biasing model. We used a ResNet50 with an extra dense layer and JPEG images as input. We used the negative log-likelihood described in Section 4.3 as our loss function.

the bias model (equation (4)). For the second model, we used a ResNet50 [16] as our deep learning classification model with an extra dense layer (1024 neurons and ReLU activation) after the convolutional and pooling layers (Fig. 9). As input we used JPEG images and we trained our model by minimizing the negative of the log-likelihood (equation (8)), i.e., we used it as our loss function. In order to estimate the bias model parameter θ we used the method described in Section 4.4, i.e., this parameter is not included as a trainable parameter in our neural network. For both models we used the bias model described in Section 4.2.

5.2 Data

For all our experiments we used galaxies contained in the GZ2 catalog [37]. With the intention of using a balanced dataset, we randomly sampled 121,984 galaxies from which 62,225 corresponded to smooth and 59,759 corresponded to disk. To train the models, we grouped these galaxies into a training set containing 97,600 galaxies, a validation set of 12,192 galaxies and a test set of 12,192 galaxies. We used the JPEG images labeled by the annotators as input to our model. Following [9], we cropped the original images from 424×424 to 207×207 pixels. All galaxy parameters needed in our experiments were obtained from the SDSS database and from [34].

5.3 Bias quantification method

In order to quantify the bias in a set of labels, we considered two metrics defined by [4] and [5]; we will refer to them as CV14 and CV18 respectively. CV14 is based on the assumption that the class fractions (same as defined in Section 4.4) in a non-biased dataset should not differ in terms of the observable properties, such as image resolution. For this, they divide the dataset in bins of these observables and calculate the deviation of the class fractions as a function of them. As mentioned in Section 4.2, we use α as our observable property (i.e., biasing parameter). For the purposes of this work, that is, a binary classification problem with only one observable property, the deviation of the class fraction for a specific class can

be expressed as

$$\sigma_k = \sqrt{\frac{1}{N_A}\sum_{l=1}^{N_A}(\tilde{f}_{l,k} - f_k)^2}, \tag{19}$$

where N_A is the number of bins in α, $\tilde{f}_{l,k}$ corresponds to the class fraction of class $k \in \{\text{smooth}, \text{disk}\}$ within bin l, and f_k to the class fraction of class k from the least biased bin in terms of α. Then, the deviation of the class fractions over the entire dataset can be quantified by

$$\text{CV14} = \sqrt{\frac{1}{N_K}\sum_{k}\sigma_k^2}, \tag{20}$$

with N_K being the number of classes. For a non-biased dataset, the class fractions should be independent of α, i.e., we expect CV14 to be close to zero.

CV18 works under the same assumption as CV14; however, it also takes into account the intrinsic properties of the data. CV18 considers that the class fractions in an unbiased dataset should not differ in terms of the observables *within bins of intrinsic properties*. They first divide the dataset in multi-dimensional bins as a function of the intrinsic properties and then divide each of these intrinsic bins in bins of the observable properties. Then, they calculate the deviation of these observable class fractions as CV14 does. By doing this, they take into consideration a possible variation in the observable class fractions related to the intrinsic properties. After considering both approaches, we decided to use CV18 as the method of measuring the bias for all our experiments, as it is the more robust of the two.

5.4 Results

5.4.1 Bias measurement

We first measure the bias of the labels from GZ2, both the original crowdsourced labels (GZ2B) and the de-biased labels from [15] (GZ2D). To obtain hard labels for GZ2D, we performed the procedure described in Section 4.1. In order to determine whether the bias from these sets of labels transfer to the predictions of new models, we trained two not de-biasing ResNet50 models (i.e. using cross entropy as the loss function), one trained over the labels from GZ2B and the other trained over GZ2D. We compared the bias of the predictions of these models over the test set with the bias of the same test set for GZ2B and GZ2D. Finally, we measured the bias of the predicted labels from our two de-biasing models described in Section 5.1, the de-biasing model from [4] (DCV14) and our proposed deep de-biasing model (DDB). The results of our experiments are reported in Table 1. We also report classification metrics for the not de-biasing ResNet50 models.

The de-biasing models were trained directly over GZ2B data using the training method previously mentioned. We used the galaxy resolution as measured by $\alpha = (r/\text{PSF})$ as biasing parameter, where r is the angular Petrosian radius of the galaxy and PSF is the angular standard deviation of a Gaussian model PSF. The value of the bias model parameter for

Table 1: Biases for different datasets and experiments. Classification metrics are included for the not de-biasing ResNet50 models trained directly over GZ2B and GZ2D.

Dataset / Method	Bias CV18	accuracy	precision	recall	f1-score
a) GZ2B [37]	0.3696 ± 0.0095	-	-	-	-
b) GZ2D [15]	0.3106 ± 0.0108	-	-	-	-
c) ResNet50 over GZ2B	0.3781 ± 0.0091	0.921	0.921	0.921	0.921
d) ResNet50 over GZ2D	0.3258 ± 0.0107	0.896	0.866	0.879	0.872
e) DCV14 over GZ2B	0.2994 ± 0.0102	-	-	-	-
f) DDB (ours) over GZ2B	$\mathbf{0.2866 \pm 0.0112}$	-	-	-	-

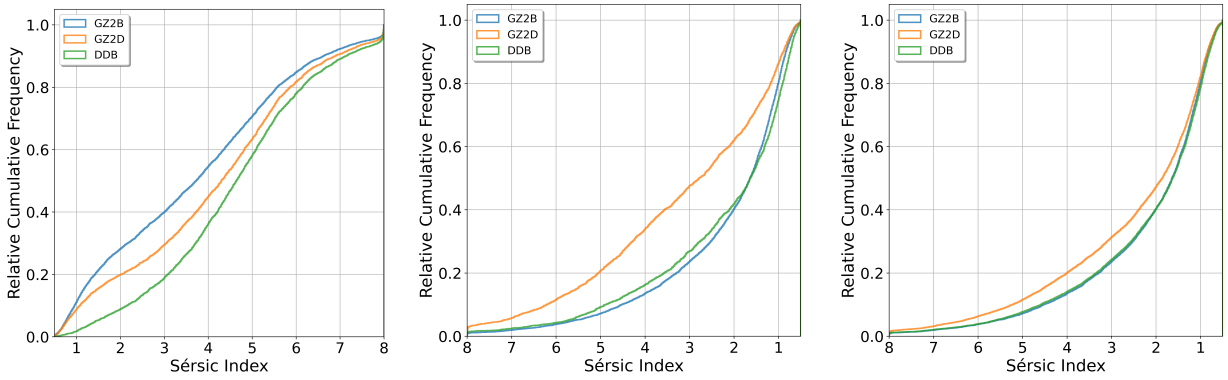
GZ2B was estimated as $\theta_{\text{DCV14}} = 9.18$ for DCV14, following [4], and as $\theta_{\text{DDB}} = 11.25$ for DDB, using the method described in Section 4.4.

We first notice that, as expected, the set of labels from GZ2D is less biased than the original crowdsourced labels from GZ2B. We also notice that the biases from the not de-biasing ResNet50 models trained over GZ2B (Table 1c) and GZD (Table 1d) present a similar level of bias compared to their respective training sets, meaning that the bias from the training sets transfers to the predictions of the models trained on them. The difference in the classification metrics between both not de-biasing models can be attributed mainly to GZ2D being an unbalanced dataset, since it contains the same galaxies as GZ2B, however, with different labels according to [15]. On the other hand, we notice that both DCV14 and DDB manage to decrease the amount of bias of GZ2B, obtaining the best result when using DDB. The aforementioned is even more relevant when considering that DCV14 relies on the calculation of the Sérsic profiles [33], the ellipticity and the half-light radius, while DDB is able to learn features that are useful to infer the de-biased labels directly from the low resolution images.

5.4.2 Sérsic profiles analysis

To evaluate our de-biasing method in terms of astrophysical parameters, we compare the cumulative frequency distribution of the Sérsic index for galaxies in our dataset (Fig. 10). We start by comparing the smooth galaxies from GZ2B with the galaxies labeled as smooth by DDB and GZ2D. A Sérsic index greater than approximately 3 is typically expected for smooth galaxies. As shown in Figure 10a GZ2B contains approximately 40% of galaxies labeled as smooth with Sérsic indexes lower than 3. GZ2D is able to diminish this fraction to approximately 30%, while our de-biasing method is able to achieve less than 20% of galaxies labeled as smooth with a Sérsic index lower than 3. Notice that the Sérsic parameters are never fed into our model; our de-biasing approach is able to automatically learn a discriminative model that agrees with these estimated radial profiles.

Figure 10b shows the Sérsic index cumulative distributions of galaxies labeled as disks in GZ2B and galaxies that changed their labels from smooth to disk by GZ2D and DDB. As explained in Section 4.2, we do not expect disk galaxies in GZ2B to be mislabeled; hence, this distribution works as a proxy for the real Sérsic index distribution of disk galaxies. Disk



(a) Galaxies classified as smooth

(b) Galaxies changed from smooth to disk

(c) Galaxies classified as disk

Figure 10: Relative cumulative frequency of galaxies as classified by each model. For (b) GZ2B corresponds to the original disk galaxies, added for comparison with the other models.

galaxies as labeled by DDB follow the distribution of disk galaxies in GZ2B: for both of these sets of labels less than a 30% of disk galaxies have a Sérsic index higher than 3. On the other hand, the disk galaxies as labeled in GZ2D are associated with almost a 50% of galaxies with a Sérsic index higher than 3. This affects directly the distribution of Sérsic indexes of disk galaxies in the whole dataset as shown in Figure 10c. In other words, DDB produces labels that match the expected relation between morphologies and radial profiles of galaxies, demonstrating the effectiveness of our method in capturing astrophysical features without relying on expert-derived parameters.

5.4.3 Visual inspection of high resolution images

In order to visually assess the performance of our approach, we searched the Mikulski Archive for Space Telescopes (MAST)[1] for high resolution Hubble Space Telescope (HST) images of galaxies in the test set that changed their labels when using DDB. We found 11 HST images of galaxies that changed from a "smooth" biased label to a "disk" de-biased label. Figure 11 shows the SDSS and HST images of such galaxies. Figures 11a, 11b, 11c, 11e, 11g, 11h, 11i and 11j show evidence of spiral arms. Figures 11d, 11f and 11k show lenticular features, although it is not completely clear that these are effectively disk galaxies. A further astrophysical analysis is required to corroborate their type. These results show that our de-biasing neural network is able to correctly label images even when the labels used for training are biased.

5.4.4 GradCAM

Aiming to understand the difference between our de-biasing model and the not de-biasing ResNet50 models, we used GradCAM. GradCAM [32] is a technique for producing visual explanations of the decision making process of a CNN, generating a localization map that accentuates the relevant regions of the image for predicting a specific class. We start by

[1]https://mast.stsci.edu/portal/Mashup/Clients/Mast/Portal.html

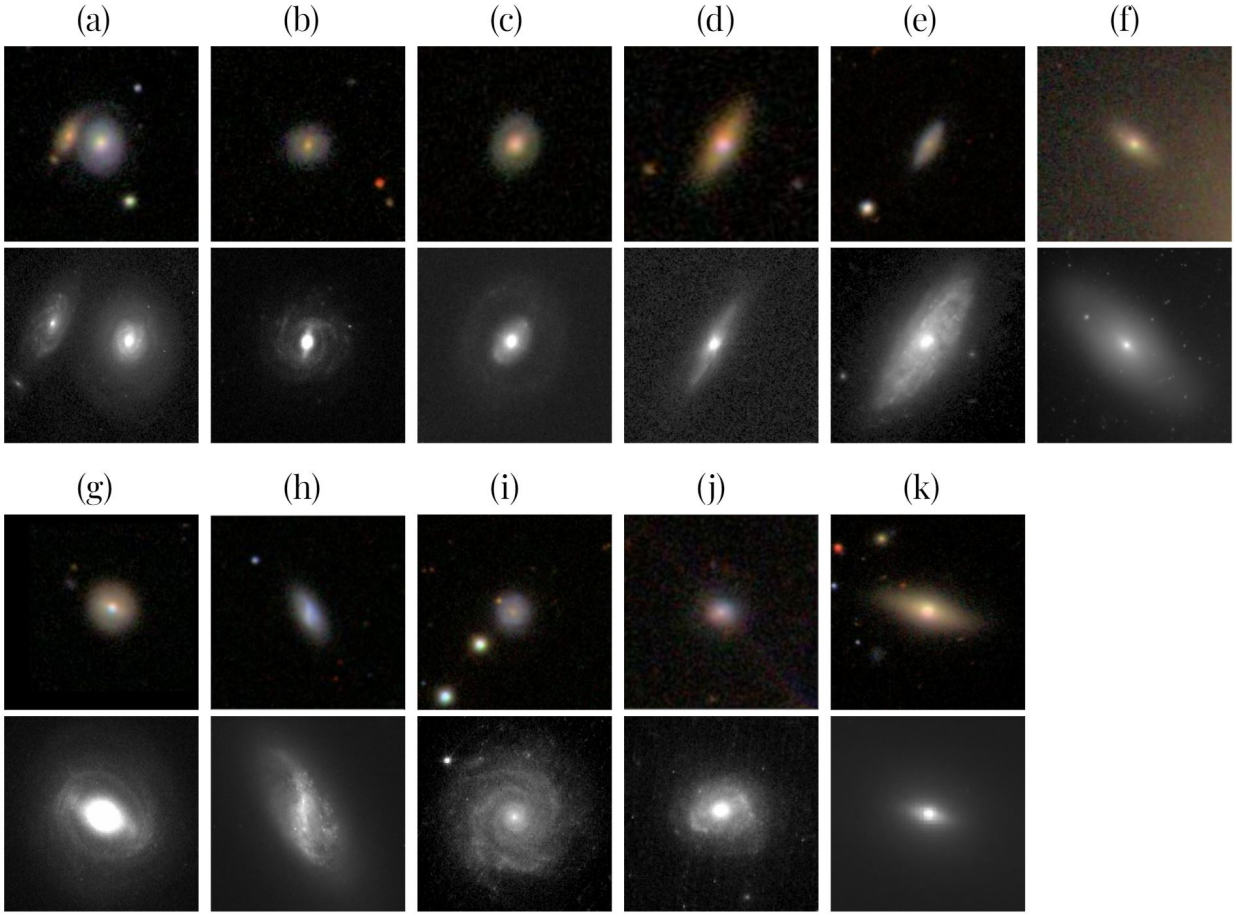
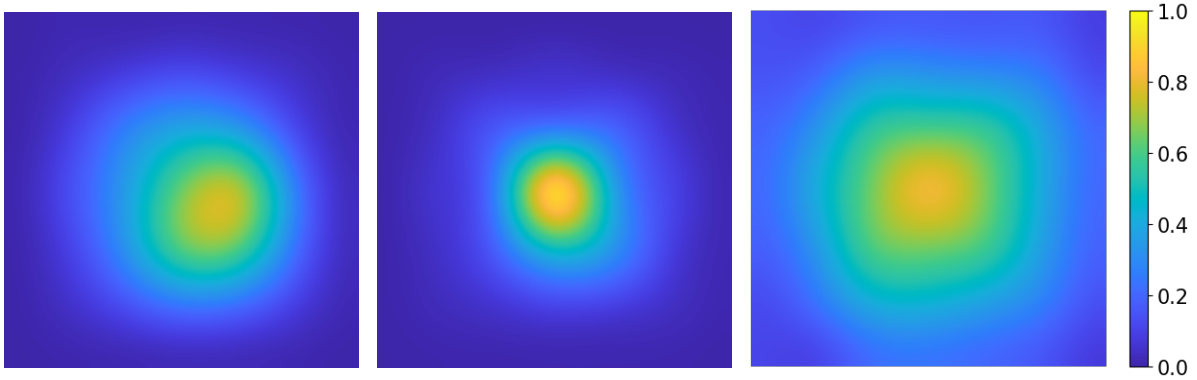


Figure 11: Low and high resolution images of galaxies that changed their labels from "smooth" to "disk" when using our method. The first row of each block shows the images as labeled by the annotators of Galaxy Zoo 2. The second row shows higher resolution images from the HST.



(a) ResNet50 over GZ2B

(b) ResNet50 over GZ2D

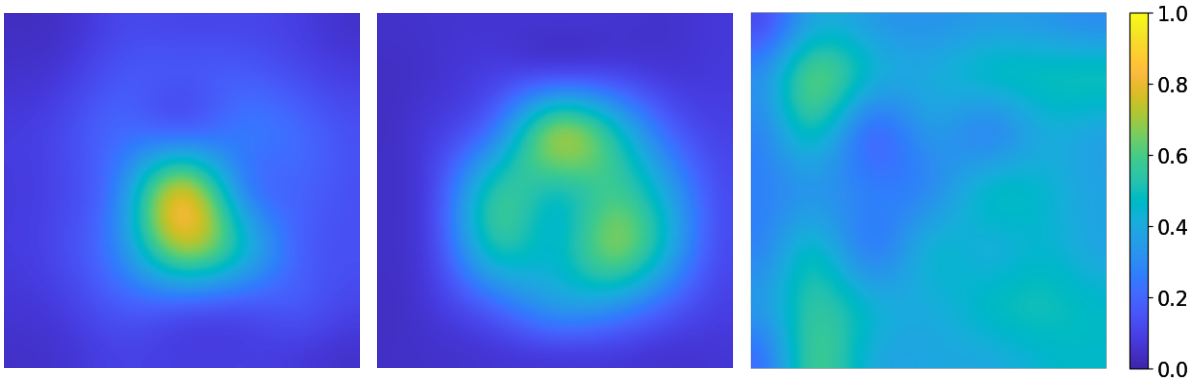
(c) DDB over GZ2B

Figure 12: Average GradCAM localization maps for the disk class. Highlighted in yellow are the class-discriminative regions. Notice how (a) and (b) focus in the galaxy bulge while (c) considers the entire galaxy.

analyzing the class-discriminative regions of the models when classifying galaxies as disks. Figure 12 shows, for each model, the average GradCAM maps of all the galaxies in the test set that the respective model predicted as a disk. Figures 12a and 12b show that both not de-biased ResNet50 models exhibit a tendency to concentrate their attention on the center of the image, where the bulge of the galaxy is located. On the other hand, Figure 12c shows that DDB uses an extended area of the image, hence considering a wider fraction of the disks when discriminating. These Figures suggest that DDB is able to detect weak disk features in the image that are often missed to the non-expert human eye.

Figure 13 corresponds to the average GradCAM maps of all the galaxies classified as smooth by each model. We notice that the ResNet50 trained over GZ2B considers almost exclusively the center of the image, which is associated with the bulge of the galaxy (Fig. 13a). The ResNet50 trained over GZ2D focuses mainly in the regions around the bulge (Fig. 13b). On the other hand, DDB concentrates its attention to a wider field within the image (Fig. 13c), suggesting that its decision making process is based on not identifying disk-like features in the image. This suggests that the ResNet50 trained over GZ2D can recognize the absence of disk features near the bulge, while DDB expands its field of attention to a broader region in order to detect that no lower disk signals are present in the image when inferring that a galaxy is smooth. DDB takes into consideration the broader diversity of light profiles found among smooth galaxies, including this information in its decision making process, while both not-debiased ResNet50 models are neglecting a portion of this information by focusing mainly on the core and the regions around it.

To provide a clearer illustration of the aforementioned findings, Figure 14 shows the average intensity as a function of the radial distance from the center of each of the GradCAM localization maps of Figures 12 and 13. For the GradCAM maps corresponding to the disk class (Fig. 14a) we notice that DDB maintains a higher average intensity when approaching the edges, i.e., considers a broader region from the center as relevant in its decision making process. In the case of the GradCAM maps corresponding to the smooth class (Fig. 14b) we notice that for the not de-biased ResNet50 models the average intensity decreases as the



(a) ResNet50 over GZ2B

(b) ResNet50 over GZ2D

(c) DDB over GZ2B

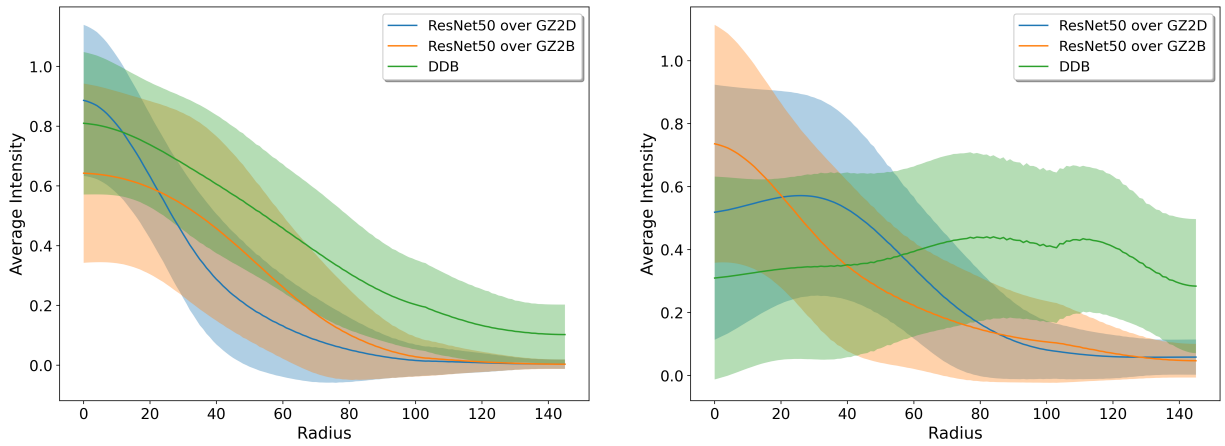
Figure 13: Average GradCAM localization maps for the smooth class. Highlighted in yellow are the class-discriminative regions. Notice how (a) focuses in the bulge of the galaxy, (b) in the regions around the bulge and (c) in the the edges of the galaxy.

distance from the center increases, while for DDB the average intensity remains relatively constant across the image, with its peak closer to the edges of the image.

6 Conclusions and future work

Data can exhibit biases in terms of observable parameters, such as resolution, which can introduce bias during the human labeling process. This inherent bias, stemming from observable properties rather than annotators, leads to systematic labeling biases in the data. In this study, we investigate the impact of training deep learning models using biased data in the context of morphological classification of galaxies. We demonstrate that training these models directly on biased data results in biased models. We introduce a method for training deep learning models that takes into account this labeling bias, to obtain unbiased models even when training with biased data. We evaluate our method by comparing the bias of the predicted labels with the bias of the labels produced by other de-biasing methods, as well as through visual inspection of high-resolution images. We also show that by using our method, the resulting model is able to learn complex astrophysical relations directly from the biased data, without relying on expert-derived parameters. Finally, we employ a visualization technique to comprehend the effect of our method on the deep learning model, comparing it with models trained without utilizing our approach. We conclude that by using our method, we can directly train a deep learning model on the biased data and obtain a model capable of both de-biasing existing datasets and labeling new data.

As future work one interesting area of exploration is the reformulation of our custom loss function (equation (8)). Currently, this loss function imposes independence assumptions, as discussed in Section 4.3, which limits the selection of attributes available for use. Another interesting direction for future research is the evaluation of different bias models, particularly more complex models that incorporate expert-derived parameters. By using complex bias models one could aim to better approximate the underlying bias of the data, potentially leading to even less biased morphological classifications. Additionally, there is potential for



(a) Disk

(b) Smooth

Figure 14: Average intensity vs. radial distance from the center of the image. (a) Corresponds to the average GradCAM maps for the disk class (Fig. 12) and (b) to the average GradCAM maps for the smooth class (Fig. 13).

adapting our de-biasing framework to other data domains. This adaptation would require to define a different domain-specific bias model and possibly to use a different classification model. Other alternative is testing the performance of the framework on multiclass classification problems and more complex classification schemes in general. Finally, another interesting alternative could be conducting a detailed astrophysical analysis of the model results which could lead to the creation of a catalog of morphologically classified galaxies.

References

[1] Steven P. Bamford et al. "Galaxy Zoo: the dependence of morphology and colour on environment*". In: *Monthly Notices of the Royal Astronomical Society* 393.4 (Mar. 2009), pp. 1324–1352. DOI: 10.1111/j.1365-2966.2008.14252.x. arXiv: 0805.2612 [astro-ph].

[2] Jakramate Bootkrajang. "A generalised label noise model for classification in the presence of annotation errors". In: *Neurocomputing* 192 (2016). Advances in artificial neural networks, machine learning and computational intelligence, pp. 61–71. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2015.12.106. URL: https://www.sciencedirect.com/science/article/pii/S0925231216002551.

[3] Kevin Bundy, Richard S. Ellis, and Christopher J. Conselice. "The Mass Assembly Histories of Galaxies of Various Morphologies in the GOODS Fields". In: *The Astrophysical Journal* 625.2 (June 2005), pp. 621–632. DOI: 10.1086/429549. arXiv: astro-ph/0502204 [astro-ph].

[4] Guillermo F. Cabrera, Chris J. Miller, and Jeff Schneider. "Systematic Labeling Bias: De-biasing where Everyone is Wrong". In: *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE. 2014, in press.

[5] Guillermo Cabrera-Vives, Christopher J. Miller, and Jeff Schneider. "Systematic Labeling Bias in Galaxy Morphologies". In: *The Astronomical Journal* 156.6 (Nov. 2018), p. 284. DOI: 10.3847/1538-3881/aae9f4. URL: https://doi.org/10.3847/1538-3881/aae9f4.

[6] G. De Vaucouleurs. "Classification and Morphology of External Galaxies". In: *Astrophysik IV: Sternsysteme / Astrophysics IV: Stellar Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1959, pp. 275–310. ISBN: 978-3-642-45932-0. DOI: 10.1007/978-3-642-45932-0_7. URL: https://doi.org/10.1007/978-3-642-45932-0_7.

[7] G. De Vaucouleurs, A. De Vaucouleurs, and Jr. Corwin H. G. *Second reference catalogue of bright galaxies. Containing information on 4,364 galaxies with references to papers published between 1964 and 1975.* 1976.

[8] Gerard De Vaucouleurs et al. *Third Reference Catalogue of Bright Galaxies*. 1991.

[9] Sander Dieleman, Kyle W. Willett, and Joni Dambre. "Rotation-invariant convolutional neural networks for galaxy morphology prediction". In: *Monthly Notices of the Royal Astronomical Society* 450.2 (June 2015), pp. 1441–1459. DOI: 10.1093/mnras/stv632. arXiv: 1503.07077 [astro-ph.IM].

[10] A. Dressler. "Galaxy morphology in rich clusters - Implications for the formation and evolution of galaxies". In: *The Astrophysical Journal* 236 (Mar. 1980), pp. 351–365. DOI: 10.1086/157753.

[11] Masataka Fukugita et al. "A Catalog of Morphologically Classified Galaxies from the Sloan Digital Sky Survey: North Equatorial Region". In: *The Astronomical Journal* 134.2 (Aug. 2007), pp. 579–593. DOI: 10.1086/518962. arXiv: 0704.1743 [astro-ph].

[12] Kunihiko Fukushima. "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position". In: *Biological Cybernetics* 36 (1980), pp. 193–202.

[13] Adam Gauci, Kristian Zarb Adami, and John Abela. *Machine Learning for Galaxy Morphology Classification*. 2010. arXiv: 1005.0390 [astro-ph.GA].

[14] Xavier Glorot and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks". In: *International Conference on Artificial Intelligence and Statistics*. 2010.

[15] Ross E. Hart et al. "Galaxy Zoo: comparing the demographics of spiral arm number and a new method for correcting redshift bias". In: *Monthly Notices of the Royal Astronomical Society* 461.4 (July 2016), pp. 3663–3682. ISSN: 0035-8711. DOI: 10.1093/mnras/stw1588. eprint: https://academic.oup.com/mnras/article-pdf/461/4/3663/8112546/stw1588.pdf. URL: https://doi.org/10.1093/mnras/stw1588.

[16] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016.

[17] E. P. Hubble. "Extragalactic nebulae." In: *The Astrophysical Journal* 64 (Dec. 1926), pp. 321–369. DOI: 10.1086/143018.

[18] M. Huertas-Company et al. "Revisiting the Hubble sequence in the SDSS DR7 spectroscopic sample: a publicly available Bayesian automated classification". In: *Astronomy and Astrophysics - A&A* 525, A157 (Jan. 2011), A157. DOI: 10.1051/0004-6361/201015735. arXiv: 1010.3018 [astro-ph.CO].

[19] M. Huertas-Company et al. "The Morphologies of Massive Galaxies from z ~3—Witnessing the Two Channels of Bulge Growth". In: *The Astrophysical Journal* 809.1, 95 (Aug. 2015), p. 95. DOI: 10.1088/0004-637X/809/1/95. arXiv: 1506.03084 [astro-ph.GA].

[20] Jeyhan S. Kartaltepe et al. "CANDELS Visual Classifications: Scheme, Data Release, and First Results". In: *The Astrophysical Journal Supplement Series* 221.1, 11 (Nov. 2015), p. 11. DOI: 10.1088/0067-0049/221/1/11. arXiv: 1401.2455 [astro-ph.GA].

[21] Nour Eldeen Khalifa et al. "Deep Galaxy V2: Robust Deep Convolutional Neural Networks for Galaxy Morphology Classifications". In: *2018 International Conference on Computing Sciences and Engineering (ICCSE)*. 2018, pp. 1–6. DOI: 10.1109/ICCSE1.2018.8374210.

[22] Byungju Kim et al. "Learning Not to Learn: Training Deep Neural Networks With Biased Data". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019.

[23] Yann LeCun et al. "Gradient-based learning applied to document recognition". In: *Proc. IEEE* 86 (1998), pp. 2278–2324.

[24] Bomee Lee et al. "CANDELS: THE CORRELATION BETWEEN GALAXY MORPHOLOGY AND STAR FORMATION ACTIVITY ATz∼ 2". In: *The Astrophysical Journal* 774.1 (Aug. 2013), p. 47. DOI: 10.1088/0004-637x/774/1/47. URL: https://doi.org/10.1088/0004-637x/774/1/47.

[25] Chris Lintott et al. "Galaxy Zoo 1: data release of morphological classifications for nearly 900 000 galaxies". In: *Monthly Notices of the Royal Astronomical Society* 410.1 (Jan. 2011), pp. 166–178. DOI: 10.1111/j.1365-2966.2010.17432.x. arXiv: 1007.3265 [astro-ph.GA].

[26] Chris J. Lintott et al. "Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey". In: *Monthly Notices of the Royal Astronomical Society* 389.3 (Sept. 2008), pp. 1179–1189. DOI: 10.1111/j.1365-2966.2008.13689.x. arXiv: 0804.4483 [astro-ph].

[27] Preethi B. Nair and Roberto G. Abraham. "A Catalog of Detailed Visual Morphological Classifications for 14,034 Galaxies in the Sloan Digital Sky Survey". In: *The Astrophysical Journal Supplement Series* 186.2 (Feb. 2010), pp. 427–456. DOI: 10.1088/0067-0049/186/2/427. arXiv: 1001.2401 [astro-ph.CO].

[28] Herbert E. Robbins. "A Stochastic Approximation Method". In: *Annals of Mathematical Statistics* 22 (1951), pp. 400–407.

[29] Frank Rosenblatt. "The perceptron: a probabilistic model for information storage and organization in the brain." In: *Psychological review* 65 6 (1958), pp. 386–408.

[30] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors". In: *Nature* 323 (1986), pp. 533–536.

[31] Kevin Schawinski et al. "Observational evidence for AGN feedback in early-type galaxies". In: *Monthly Notices of the Royal Astronomical Society* 382.4 (Dec. 2007), pp. 1415–1431. DOI: 10.1111/j.1365-2966.2007.12487.x. arXiv: 0709.3015 [astro-ph].

[32] Ramprasaath R. Selvaraju et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *International Journal of Computer Vision* 128.2 (Feb. 2020), pp. 336–359. ISSN: 1573-1405. DOI: 10.1007/s11263-019-01228-7. URL: https://doi.org/10.1007/s11263-019-01228-7.

[33] J. L. Sersic. *Atlas de galaxias australes*. Córdoba: Obs. Astronómico, 1968.

[34] L. Simard et al. "A Catalog of Bulge+disk Decompositions and Updated Photometry for 1.12 Million Galaxies in the Sloan Digital Sky Survey". In: *The Astrophysical Journal Supplement Series* 196, 11 (Sept. 2011), p. 11. DOI: 10.1088/0067-0049/196/1/11. arXiv: 1107.1518 [astro-ph.CO].

[35] B. D. Simmons et al. "Galaxy Zoo: quantitative visual morphological classifications for 48 000 galaxies from CANDELS". In: *Monthly Notices of the Royal Astronomical Society* 464.4 (Feb. 2017), pp. 4420–4447. DOI: 10.1093/mnras/stw2587. arXiv: 1610.03070 [astro-ph.GA].

[36] Gregory F. Snyder et al. "Galaxy morphology and star formation in the Illustris Simulation at z = 0". In: *Monthly Notices of the Royal Astronomical Society* 454.2 (Oct. 2015), pp. 1886–1908. ISSN: 0035-8711. DOI: 10.1093/mnras/stv2078. eprint: https://academic.oup.com/mnras/article-pdf/454/2/1886/13769809/stv2078.pdf. URL: https://doi.org/10.1093/mnras/stv2078.

[37] Kyle W. Willett et al. "Galaxy Zoo 2: detailed morphological classifications for 304 122 galaxies from the Sloan Digital Sky Survey". In: *Monthly Notices of the Royal Astronomical Society* 435.4 (Nov. 2013), pp. 2835–2860. DOI: 10.1093/mnras/stt1458. arXiv: 1308.3496 [astro-ph.CO].

[38] Kyle W. Willett et al. "Galaxy Zoo: morphological classifications for 120 000 galaxies in HST legacy imaging". In: *Monthly Notices of the Royal Astronomical Society* 464.4 (Feb. 2017), pp. 4176–4203. DOI: 10.1093/mnras/stw2568. arXiv: 1610.03068 [astro-ph.GA].

[39] Naoki Yasuda et al. "Galaxy Number Counts from the Sloan Digital Sky Survey Commissioning Data". In: *The Astronomical Journal* 122 (2001), pp. 1104–1124.

[40] Xiao-Pan Zhu et al. "Galaxy morphology classification with deep convolutional neural networks". In: *Astrophysics and Space Science* 364 (Apr. 2019). DOI: 10.1007/s10509-019-3540-1.