GALAXY MORPHOLOGY ESTIMATIONS VIA CONVOLUTIONAL NEURAL NETWORKS

Abhishek Kattuparambil[1]

[1] *Department of Astronomy, University of California, Berkeley, CA, USA 94720*

ABSTRACT

We develop a tool for automatic inference of galaxy images, such that a probabilistic estimate of morphological features can be derived with high confidence, and analysis can be done on the galaxy's evolutionary history. Building on the data release of Galaxy Zoo 2 (GZ2), a citizen science project providing 37 morphological labels for over 304,122 galaxies sourced from the Sloan Digital Sky Survey, we design and provide a convolutional neural network to predict the labels of each galaxy from its corresponding GZ2 image. Furthermore, we compare the performance of the model to an out-of-the-box ResNet, and improve the ResNet by adding additional layers and pre-processing the images. The final model is used to predict morphological labels for a test set from the GZ2 data, and we analyze the effects of classification and angular separation bias on the predicted merger fraction.

Keywords: Morphology — Image Processing — Convolutional Neural Networks — ResNet

1. INTRODUCTION

Despite the constant evolution of galaxies over time, their current structure and morphology can provide insight towards their evolutionary history and formation. Furthermore, the structure of a galaxy can indicate a prior merger, as lopsidedness and asymmetry can quantitatively define a merger with a peculiarity threshold, detailing the likelihood of a galaxy naturally forming with the observed structure. Galaxies which feature active star formation are characterized by their clumped spiral arms and bright starbursts, which are visually distinct from the dormant nature of local elliptical galaxies. As outlined in Conselice (2014), the CAS[1] parameters can also point to trends in formation and evolutionary history. When combined with the Sérsic profile, which describes the intensity of a galaxy as a function of distance from its center, the concentration parameter correlates with the structure (shape, size, and mass) of the galaxy. As mentioned before, the clumpiness index evidences star formation, and the asymmetry parameter describes merger activity. The model will learn relationships between the galaxy image and its morphology, such that we will be able to analyze the disconnect between visual morphology and evolutionary history, specifically regarding the merger fraction.

[1] Concentration, asymmetry, and clumpiness parameters. These are quantifiable representations of morphological features, allowing astronomers to develop numerical thresholds for galaxy classifications

2. GALAXY ZOO

Galaxy Zoo 2 (GZ2) is a citizen science project which publicly released its results in Willett (2013), recording probabilistic classifications for 304,122 galaxies from the Sloan Digital Sky Survey (SDSS). It is a direct successor to the original Galaxy Zoo from Lintott (2008), and proposed a more complex, refined classification method: all 37 morphological labels are derived through user responses to an 11 question decision tree for each galaxy. The selection of galaxies is not random; extensive quality cuts have been applied to the entire SDSS DR7 to ensure that fine morphological features are discernible in the final GZ2 dataset. The main sample begins as a direct subset of the North Galactic Cap region in the DR7 Legacy catalog; only the galaxies within the top quartile of brightness are retained. Additionally, galaxies with a Petrosian half-light magnitude[2] larger than 17.0 in the R-band are discarded. Furthermore, galaxies must be adaquately sized, defined as having a R_{90}[3] larger than 3 arcseconds. To ensure that the sample does not contain extremely old systems, galaxies with redshifts (z) above 0.25 or below 0.0005 are excluded; these cutoffs are equivalent to roughly 3 Gyr and 50 Myr respectively. The galaxy image is generated from

[2] The magnitude corresponding to half of the Petrosian flux F_P. This flux is defined over the Petrosian radius R_P, which is the point at which the average intensity within the radius is equal to the intensity at that radius.

[3] This quantity is defined as the radius containing 90% of the Petrosian aperture flux

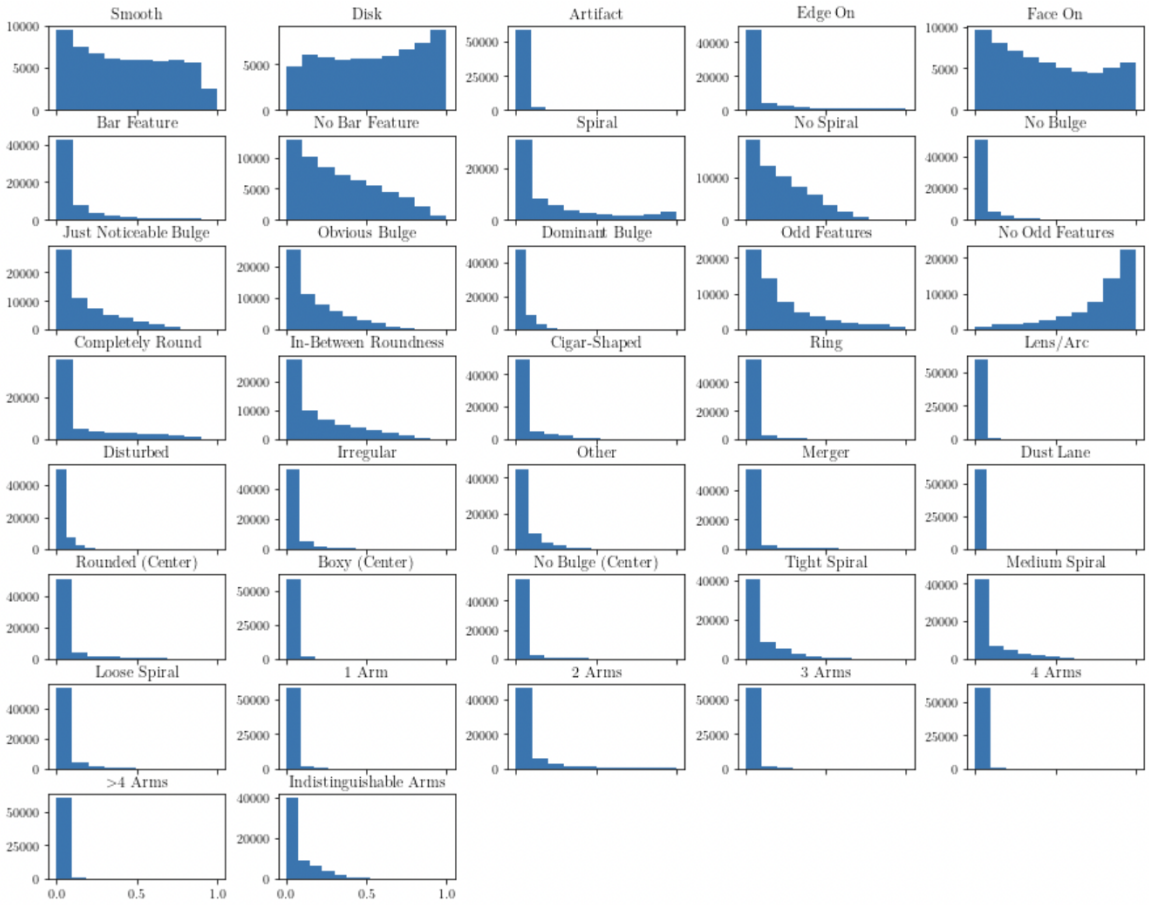

Figure 1. Histograms of the recorded label probabilities of all 37 labels provided in the GZ2 dataset, with the label name titling the graph. The classes are not explicitly labeled, but the probabilities of each class will sum to 1, best evidenced by the mirrored histograms produced by binary label classes, such as 'Odd Features' and 'No Odd Features.' Rare morphological labels such as multiple spirals and imaging artifacts clearly exhibit histograms peaking at a value of 0.

the SDSS ImgCutout web service and from the Legacy and Stripe 82 normal depth surveys, each being color-corrected and masked to emphasize visual characteristics. After being submitted to the GZ2 web tool for citizen inspection, each galaxy has received an average of 44 classifications, and over 99.9% of the samples have over 27 classifications. Despite their apparent complexity, these labels are viable, and professional astronomers agree with over 90% of existing classifications, including descriptions of galactic bulges, spiral arms, orientation, and shape. However, the results suffer from a high-redshift *classification bias*, which is an umbrella term to define the unprecedented decrease in fine morphology/structure identification at high redshifts. This effect is not without basis, as the high-redshift galaxies are naturally farther from Earth, such that their images are smaller and dimmer. Physical characteristics are more difficult to distinguish with samples exhibiting large redshifts ($z > 0.085$), where the absolute magnitude drops below the sensitivity of the SDSS. Additionally, the skewed classifications are also due to a form of survivorship bias — the high-redshift galaxies which are bright enough to be selected are likely to be giant red ellipticals, which explains their seemingly inflated classification responses. Furthermore, classifications are also influenced by an angular separation bias, where physically close galaxies with relatively high redshifts are misclassified as mergers due to their proximity. The relationship between misclassification rates and angular separation has been proven, but the bias has not been corrected in the GZ2 data release. Therefore, we will further analyze the observed and predicted merger fractions, and define a probability threshold for merger certainty.

2.1. *Data Model*

To source training data for the model, we use a subset of the GZ2 data release, comprised of 61,578 records, each containing an image of the galaxy and its 37 corresponding morphological labels. These labels span 11 distinct classes encoded by separate classification ques-

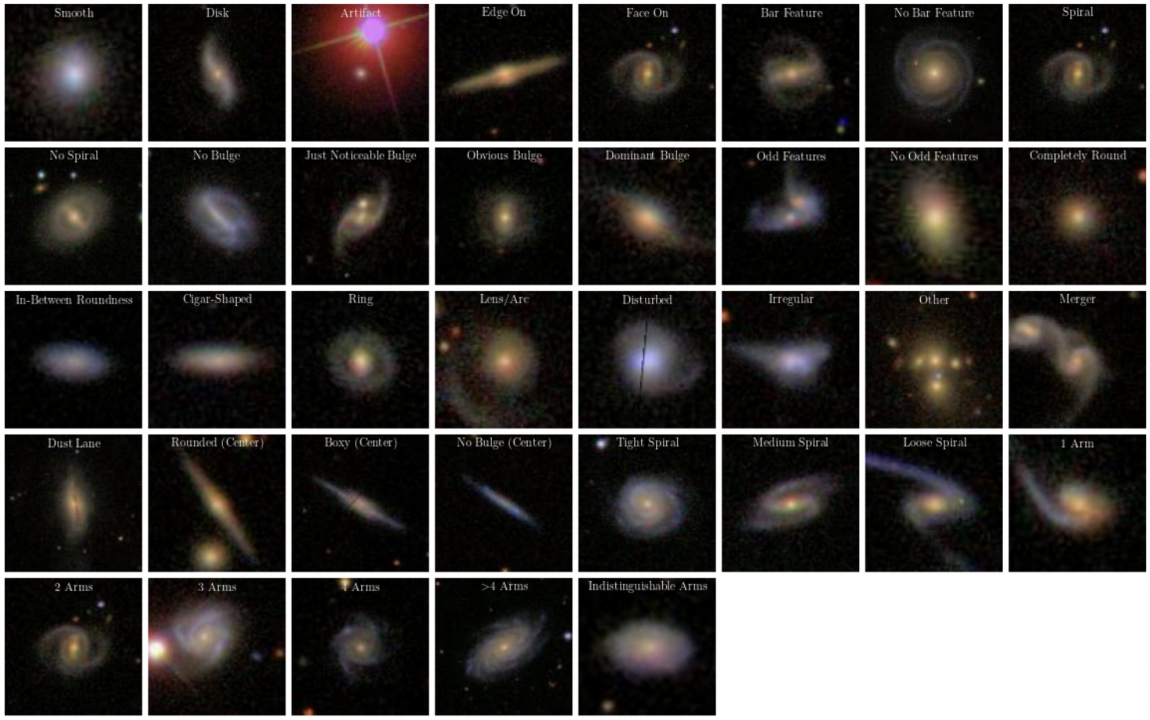

Figure 2. Prototype galaxies for each of the 37 morphological labels in the GZ2 data — the galaxy which has the highest vote fraction for the corresponding label. In this manner, we can visualize the correlation between the label and structural characteristics of the galaxy. After inspecting these images, it is apparent that the model will have difficulty predicting roundness and central bulge shape as the labels are visually quite similar. However, there are distinct features in the images of the other labels which the model will be able to learn and recognize.

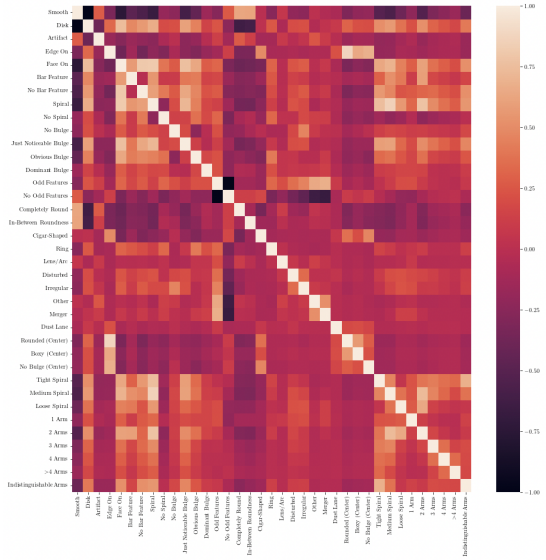

Figure 3. Heatmap displaying the covariances between each pair of labels. The identity is shown across the diagonal, as a label will have perfect correlation with itself. Furthermore, boxes centered around the diagonal can be interpreted as class-specific covariance heatmaps. For example, the top left 3x3 square is a heatmap for the Task 1 class. Binary classes are identified by the cross-diagonal adjacent black boxes which do not correlate.

tions, such as galaxy shape, bar shape, orientation, spiral arm count, and other visual features (Figure 3). The labels themselves store feature 'probabilities', recorded as the proportion of citizens who identified the feature in the corresponding galaxy; each label corresponds to a specific morphological feature, such as `cigar-shaped`, `3 arms`, `ring`, and `merger`. As detailed in the previous section, the images are produced via the SDSS Img-Cutout service, and are stored as 424x424 gri[4] image composite. Since the images are cutout relative to their celestial coordinates (right ascension, declination), each galaxy is centered in the frame, and each pixel is scaled to represent 2% of the galaxy's R_{90}, the radius which contains 90% of the Petrosian Flux F_P. See Figure 1 for a visualization of the label distributions, and see Figure 2 for prototype galaxies.

3. CONVOLUTIONAL NEURAL NETWORKS

To derive labels from a given input image, we must develop a method which learns from the image itself, and recognizes visual features which correspond to morphological labels. We employ a Convolutional Neural

[4] A three-channel photometric system with passbands in the green (475 nm), red (622 nm), and near infrared (763 nm)

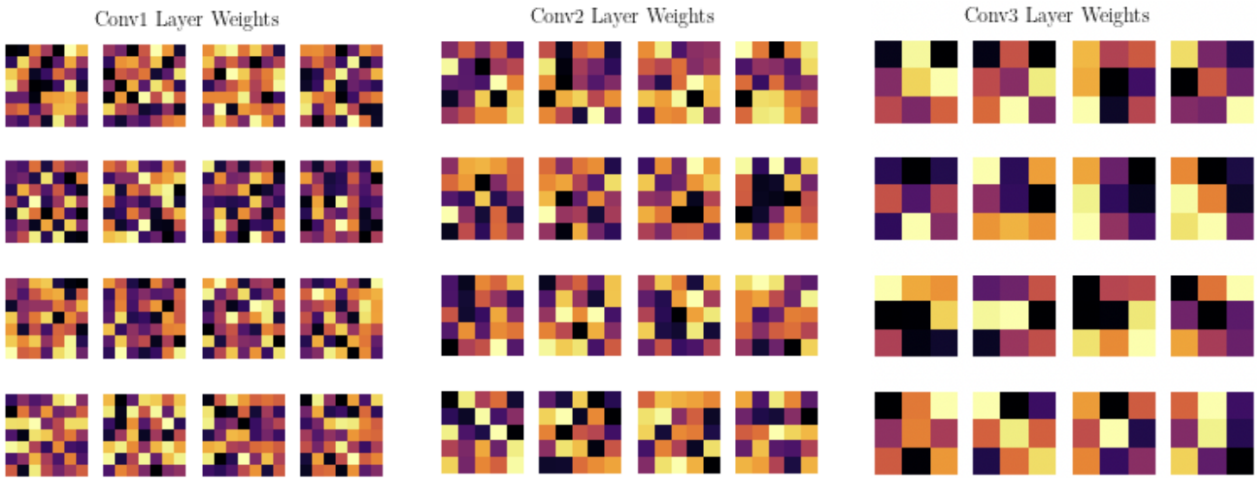

Figure 4. Weights for the first, second, and third convolutional layers. These weights are the literal kernels used in the convolutional layers, and they each respond to a unique feature learned and propogated by a neuron. As shown, the first layer utilizes a 7x7 kernel, the second uses a 5x5 kernel, and the third uses a 3x3 kernel.

Network (CNN) to tackle this task, a deep learning algorithm known for its ability to learn filters to identify and differentiate features in input images. The CNN itself is a multi-layer network, but the majority of learning occurs in the first few convolutional layers, which apply a convolution of specified kernel size over the image. Each layer will derive many filters for the image, each encoding a detection method for a learned feature. These weights represent the multi-dimensional filters used in the convolutional layers. Therefore, each convolutional layer will learn a filter, apply it to the entire image, and extrapolate results from the resulting propogated convolutions. These layers are usually followed by a rectified linear (ReLU[5]) activation function, and a pooling layer[6]. By learning filters through convolutional operations and reducing intermediate layer sizes with sampling via pooling, the convolutional network is able to learn efficiently differences between the morphological labels and define nonlinear functions in label-space.

3.1. *Proof of Concept*

To demonstrate the ability of a CNN in a visual classification task such as this, we develop a simple model and prove its ability to learn. The model is built as follows:

- Three convolutional layers, each with a padding and stride of 1, and kernel sizes of 7x7, 5x5, 3x3 in order. All are followed by a ReLU activation function and max pooling over a 2x2 kernel. See Figure 4 for a visualization of the convolution kernels.
- A flatten opration followed by two fully-connected layers. The first maps 2592 neurons to 256 with a ReLU activation function. The second is preceded by a dropout layer[7], and maps the 256 neurons to our 37 output labels, applying a sigmoid function to the result such that the results are projected into the probability range [0, 1].
- Employs the Adam optimizer, which is a first-order gradient optimization technique for stochastic functions, Ba (2014). Used with a learning rate of 0.001.
- Trained with stochastic batch sizes of 256, and validated with 128-record batches. Each epoch spans 100 batches, and RMSE losses are reported.

We build the model using PyTorch, a deep learning Python framework allowing for the straightforward creation of expressive neural networks, further discussed in Ba (2019). To provide the model with images, we create a custom galaxy class consisting of the cutout image and its corresponding label, and populate a training and validation DataLoader with the 61,578 instances. When images are fetched from the respective DataLoader, they

[5] This function will return 0 for all negative inputs, which allows the net to leave neurons inactive.

[6] There are two common types of pooling, which essentially downsamples an image over a kernel of the specified size; for average pooling, the average value of the kernel is propogated, and for max pooling the maximum value in the kernel is returned

[7] A dropout layer will nullify the contributions of some neurons during forward-propogation, such that the other neurons must also learn its classification method. This is analogous to the physiology of the brain — tens of thousands of neurons die every day, but the brain will never forget simple learned facts, such as 2+2=4. Information must be spread across neurons prevent overfitting, as neurons cannot simply memorize correspondences under this model.

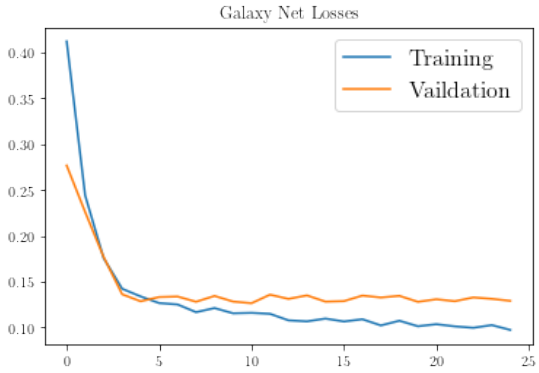

Figure 5. Training and validation losses for the three-layer CNN model learning from batches of the training data. The separation between the training and validation losses provides evidence of overfitting, which can be improved with augmented data. Furthermore, the validation loss plateaus just below 0.145, which can be lowered with additional layers and more learned filters. We improve on these results in a further iteration utilizing the ResNet model and additional optimizations. The large perturbations in the loss results are unexpected, but they likely stem from a large learning rate which will be decreased in future optimizations.

Figure 6. Training and validation losses for the out-of-the-box ResNet model. Note that this model has only been trained for 15 epochs as it is much more computationally expensive. There is still a non-negligible separation between the validation and training data, which shows that the model has overfit. Furthermore, the validation loss still plateaus at ∼0.126, and the decreasing nature of the training loss suggests that the opitimum still has not been found.

are run through a set of specified transformations; the utility of this feature will be further explained in later sections.

To train the model, we provide the aforementioned DataLoaders, which stream batched training data into the network. For each batch, the optimizer calculates gradients relative to each weights in a process named backpropagation, and takes a step in the optimal direction of its loss minimization objective. This optimization method functions closely to a pseudo-stochastic gradient descent, as the provided batches are quite limited in size relative to the entire set. The squared error loss is recorded, and the RMSE loss is accordingly calculating by averaging over the batches. To prevent overfitting, we apply cross-validation by evaluating the model on the validation data. The set is comprised of images and labels the model has not seen before, such that it is forced to perform well by learning rather than simply memorizing inputs. Ideally, these training and validation losses should converge to a similar low losses, which show that the model has approached the true optimum. See Figure 5 for training results. These results alone demonstrate that a CNN can be applied to this morphological label prediction task, as the model clearly approaches some optimum which is better than simply guessing the average label, which would correspond to an RMSE error of 0.23.

4. RESIDUAL NEURAL NETWORKS

To provide a more robust model for morphological inference, we must utilize a deeper network with more convolutional layers, translating to more complex filters and more predicting power. In order to provide more neurons without bottlenecking the performance of the model, we utilize a residual network (ResNet). These models use layers to learn residual functions with respect to the layer inputs, which reduces the complexity of each layer and allows the accuracy to improve as many layers are added, a relationship which plateaus quite early in traditional neural networks. Furthermore, the net utilizes skip layers which propagate over intermediate layers, which helps combat the vanishing gradient issue[8] and improve training speed. To predict morphological label probabilities, we will first use an untrained out-of-the-box ResNet-18. This is an 18 layer neural network whose architecture is provided in He (2015), and performs more efficiently and accurately than its bulky VGG counterparts for image classification tasks. By applying this out-of-the-box model to the morphological label prediction task, a baseline for the ResNet performance can be established.

Despite its performance improvements, the ResNet model is still relatively large, and the throughput of the model dropped significantly when training with the ResNet-18. Therefore, we have dropped the batch sizes

[8] In deeper networks, the backpropagation process involves the multiplication of many small numbers, which can drive the gradients to 0. By skipping layers, ResNet keeps the gradients large enough to ensure the optimizer can progress.

to 64 for training and 32 for validation. However, this model serves as a proof-of-concept for the ResNet application, and modifications will be made to perfect the ResNet for the morphological classification task. See Figure 6 for training results.

5. TAILORING THE RESNET

The results of the unmodified ResNet are encouraging, but the model can be adjusted to fit this specific task better, and the data can be modified to be more robust and provide information more efficiently. We take inspiration from the shortcomings of the previous classification attempt, and implement data augmentation, learning rate scheduling, and weight rescaling to improve the performance of the model.

5.1. *Image Preprocessing*

As with all other models,each image is converted into a PyTorch float tensor before computation, such that the pixel brightness values are standardized to the range [0, 1]. To improve the training speed of the model, each image is downsized to reduce the scale of our ResNet's inputs. We enforce margins upon the image, as the galaxies are always centered: 100 pixels are removed from each side of the image, such that only the central 224x224 frame is retained. Furthermore, the images are downsampled onto a coarser image grid, resulting in final pre-processed images of size 80x80. This is roughly a 25x reduction in space, and an almost lossless transformation in predicting power. To prevent overfitting and provide robustness to our training data, we add transformations to augment the images before providing them to the model. Each image is randomly rotated (by a degree within [0, 360]), and has a 50% chance to be flipped over the horizontal and/or vertical axes. In this manner, regardless of how many passes are made over the training data, each image is unique, and the model will not be able to overfit by simply memorizing results.

5.2. *Training Modifications*

Furthermore, the separation between the training and validation data proves that the model has overfit, and the decrease of training loss also suggests that the optimizer has not approached the true optimum of the RMSE cost function. The overfitting has been addressed by the image augmentation, but the training process and optimizer must be modified to improve the model's convergence to optimal weights. When the validation loss plateaus, it can be interpreted as the optimizer consistently stepping over an optimum as the learning rate is too large. We introduce the ReduceLROnPlateau learning rate scheduler, provided by PyTorch, to decrease the

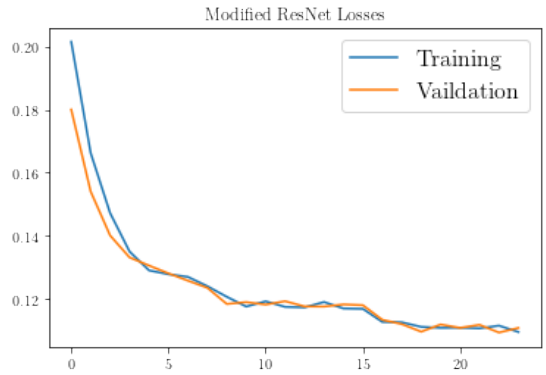

Figure 7. Training and validation losses for the modified ResNet model with image augmentation. Note that we have continued to a 25-epoch training process with the smaller ResNet batch sizes; the first epoch has been omitted from the graph to reduce the size of the y-axis so we can see the late-stage training results with higher resolution. The separation between the training and validation is negligible, which shows that data augmentation has helped with the overfit. Additionally, there is a clear step at the 16th epoch, where the loss drops after a plateau; this is evidence of the ReduceLROnPlateau scheduler working, dropping the learning rate to allow the model to escape the local optimum ($L = 0.125$) which the out-of-the-box ResNet was stuck in. This model finishes with a validation loss of ∼0.11, outperforming all other versions provided in this paper, and well below the average-guess RMSE of 0.23. Ideally, the model would have been trained for much longer, likely on a GPU.

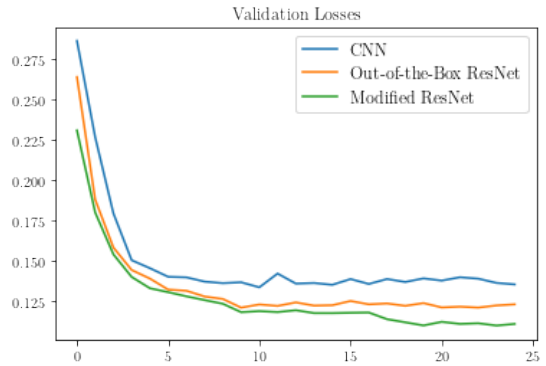

Figure 8. Validation losses for each model, demonstrating the improvements made with each iteration. The CNN loss plateaued at ∼0.145, which was directly improved by increasing the complexity of the model with an 18-layer ResNet. Even the out-of-the-box ResNet plateaued quite early and showed evidence of overfitting. A learning rate scheduler and image augmentation were added, and their effects are shown in the final modified ResNet loss curve, which exhibits the lowest loss and a secondary step at the 16th epoch as the model can now step towards the optimum with the lower scheduled learning rate.

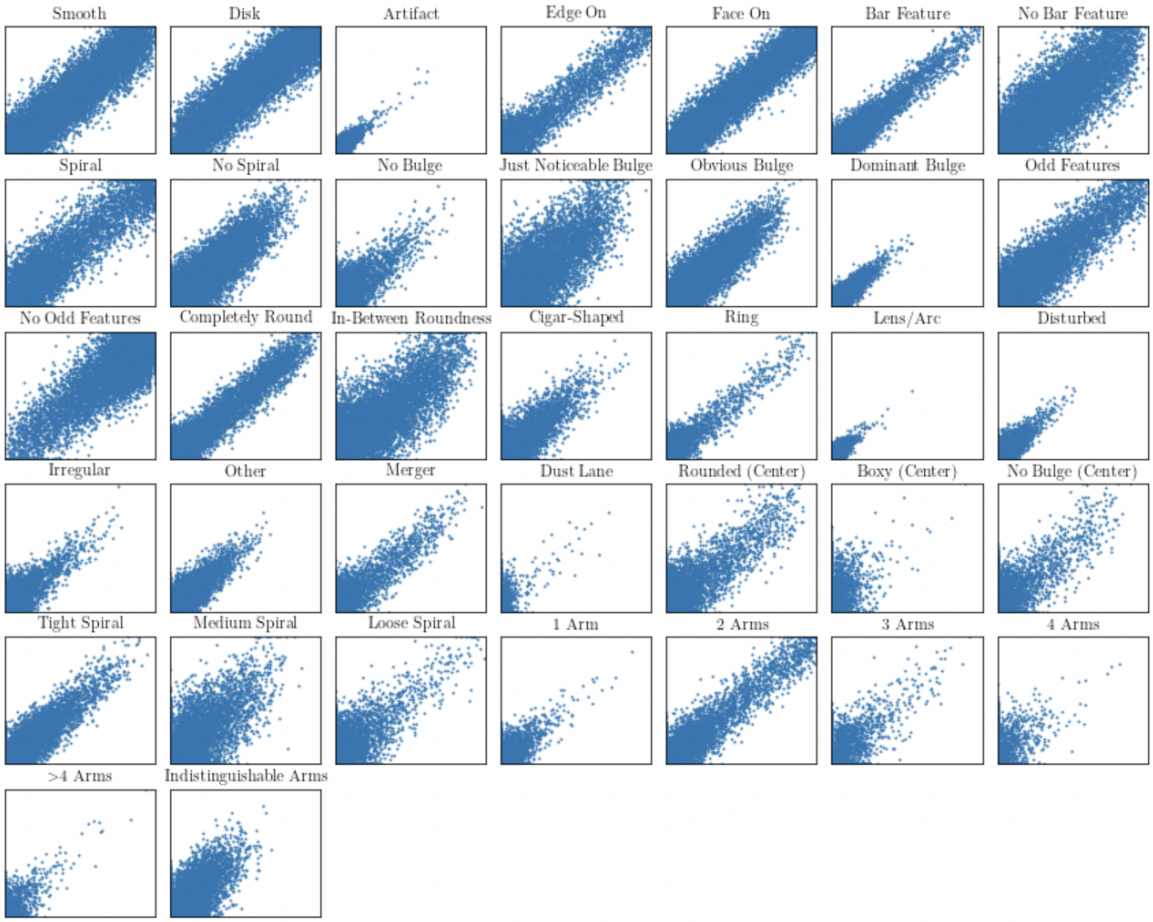

Figure 9. Comparisons of the reported GZ2 probabilities and predicted probabilities for each of the 37 labels. Each comparison is visualized as a scatter plot over the identical axes of domain [0, 1], which captures the entire probability range. As shown, some of the labels are predicted more accurately than others, usually corresponding to a distinct visual morphological feature. For example, the model is able to estimate '2 Arms' with less error than the 'Indistinguishable Arms' label. Furthermore, the labels 'Artifact' and 'Lens/Arc' have the lowest scatter, as their corresponding images differ greatly from the others, such that the model can easily derive filters to identify them. The model performs worse on labels representing features which are hard to identify, such as 'In-Between Roundness' and 'No Bar Feature'. On average, the spread across these plots is ∼0.33, mapping directly to the reported 0.11 RMSE loss of the final model.

learning rate once the validation loss plateaus. Therefore, the optimizer will be able to take smaller steps towards the optimum. Furthermore, the weights are rescaled before they are output, normalized such that the probabilities in each subclass sum to 1. This essentially teaches the model which class each label belongs to; it will learn to provide the probabilities such that they will be correct after rescaling.

6. RESULTS

To evaluate the accuracy of the modified ResNet, we can compare predictions for the validation set to their GZ2 labels. We display the predicted morphological prototypes to their observed companions, and plot each predicted label against its GZ2 counterpart. Ideally, this label plot should be a diagonal line identity, representing equality between predictions and observations. However, in reality, the plot exhibits an intrinsic spread/scatter which will inform the model of common classification mistakes.

6.1. *Performance*

The modified ResNet is presented as the final classifier for the galaxy morphology task. By the time it finishes training on the initial GZ2 sample, it records a validation loss of 0.11. Compared to the reported GZ2 labels, the model does quite well to predict relative classes, but the exact probabilities themselves vary significantly. See Figure 9 for the probability comparisons, and see 10 for the prototype comparisons.

6.2. *Merger Fractions*

To analyze the angular separation bias that has not been removed from the GZ2 data, we turn to the pre-

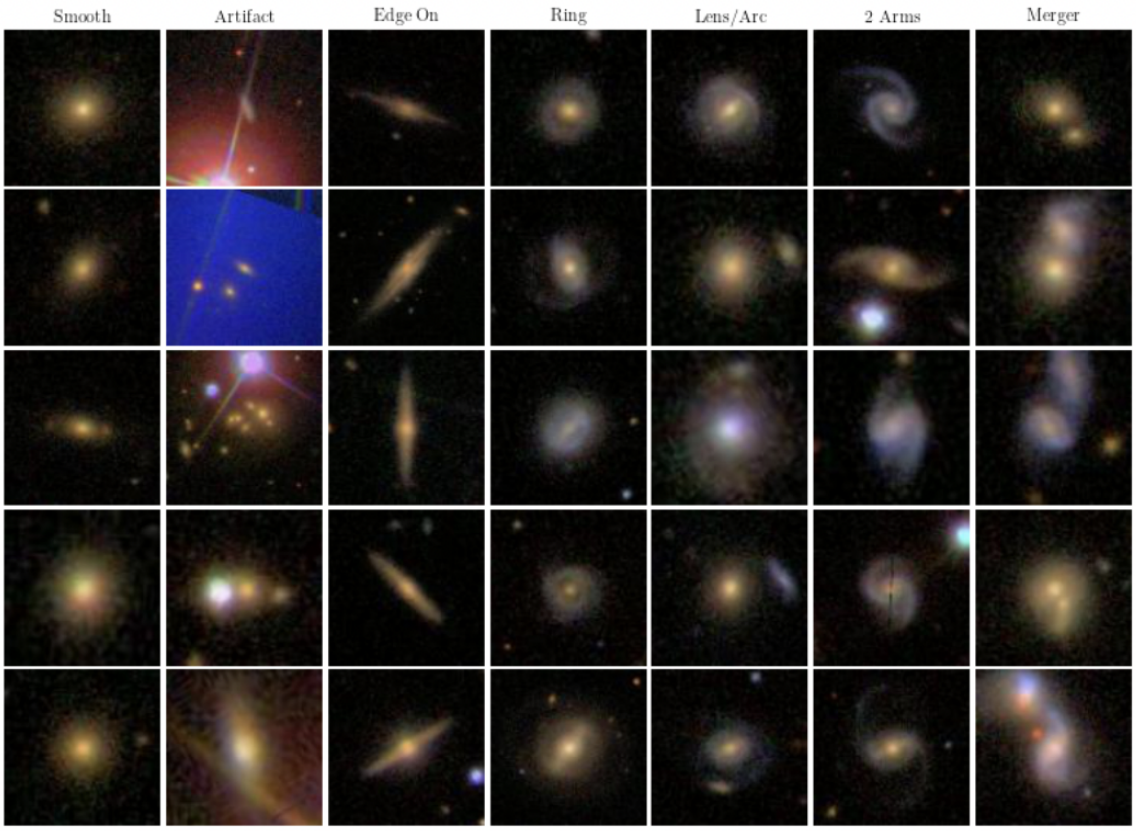

Figure 10. Prototypes for the validation data as predicted by the model. Each column represents the label titled at its top, and the images are sorted in descending order such that the first row contains the predictions with the highest label values. As shown in the probability comparisons, the model does well for labels that are visually prominent, and these labels are all quite distinctive in their appearance. The 'ring' categorizations are likely the most skewed, but even these images show loose evidence of ring structures. Therefore, the model is proficient at predicting prototypes, and never reaches below the true top 24 images to retrieve its 5 predictions. This suggests that the results are better used for categorization rather than explicit probabilistic calculations.

dicted merger probabilities. After examining their corresponding images, a threshold can be defined to indicate a border between certain mergers and visual misclassifications. The issue of merger classification is not a byproduct of the citizen project, in fact, the certainty of merger observation is an active field of research. In Lotz (2011), merger rates are analyzed by calibrating three estimators based on the morphology indicator $G - M_{20}$, asymmetry parameter A, and physically close galaxies with various separations. The feasibility of a merger is derived from hydrodynamic simulations of its evolution, coupled with these predefined merger metrics. Mergers are a closely studied topic in astrophysics, as they are relatively poorly understood, and can influence rapid star formation, gas/dust accretion such that they are key to galaxy evolutionary history and assembly. After inspecting the merger predictions on the validation set, there are clear angular separations between galaxies with probabilities < 0.54. Therefore, we define 0.54 as a probability threshold for merger existence; galaxies with labels below this threshold are considered products of the angular separation bias. This value is quite high and shows a definite bias for merger classifications, but images above this threshold are all genuine mergers.

7. SUMMARY

We have developed a method of automatic morphological label estimation via a convolutional neural network applied to an imaged galaxy. The numeric predictions have an average RMSE of 0.11, but the holistic categorizations are accurate when compared to their GZ2 counterparts. This is a purely generative model, trained on existing images and GZ2 labels, which can estimate labels for any galaxy-centered input image. Furthermore, the model is immensely expressive and has been trained on 61,578 examples, guaranteeing that it has learned near-optimal filters for the classification task. In areas of active research, it can be used to automatically categorize and estimate spiral/merger fractions for imaged galaxy datasets.

REFERENCES

Ba, Jimmy, K. D. P. 2014, Adam: A Method for Stochastic Optimization. https://arxiv.org/abs/1412.6980

—. 2019, Adam: A Method for Stochastic Optimization. https://arxiv.org/abs/1912.01703

Conselice, C. J. 2014, The Evolution of Galaxy Structure over Cosmic Time. https://arxiv.org/abs/1403.2783

He, Kaiming, Z. X. R. S. S. J. 2015, Deep Residual Learning for Image Recognition. https://arxiv.org/abs/1512.03385

Lintott, Chris J., S. K. S. A. L. K. 2008, Galaxy Zoo : Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. https://arxiv.org/abs/0804.4483

Lotz, Jennifer M., J. P. C. T. C. D. 2011, The Major and Minor Galaxy Merger Rates at z ¡ 1.5. https://arxiv.org/abs/1108.2508

Willett, Kyle W., L. C. J. B. S. P. M. K. L. 2013, Galaxy Zoo 2: detailed morphological classifications for 304,122 galaxies from the Sloan Digital Sky Survey. https://arxiv.org/abs/1308.3496