Enhancing Bayesian CNN Models for Galaxy Image Classification with Saliency Mapping

Tong Su **Yutong Han** **Jisu Qian** **Feifan Xiang**

Abstract

The increasing number of galaxy images captured by modern telescopes has made galaxy image classification a crucial task in astrophysics. Motivated by Bhambra et al's work in 2022 on applying explainable artificial intelligence (XAI) techniques to galaxy image classification results (Bhambra et al., 2022), we incorporate the saliency mapping methods to provide interpretability and enhance the understanding of the Bayesian CNN model's decision-making process. We applied Bayesian CNN model on a dataset from the Galaxy Zoo survey to investigate the galaxy images that have extreme cross-entropy losses and calculate the similarity scores to find their most similar pairs. Furthermore, our approach combines the probabilistic uncertainty of the galaxy features and saliency maps on regression problems.

1 Introduction & Related Works

Classification of galaxies was first proposed in 1924 by American astronomer Edwin Hubble, who noted patterns in the shapes of galaxies. The initial classification included 4 categories, with subcategories that were difficult to concretely define. This task of classification was initially done using human eyes and judgment, introducing high degrees of bias and uncertainty. The problem is an ideal application for neural networks, and research on the topic has been explored since the early 1990s, with the first paper appearing in 1992 by Storrie-Lombardi et al. (1992).

Subsequent advancements in the classification problem of galaxies did not significantly progress until the development of convolutional neural networks (CNNs), made possible by the advent of graphic processing units (GPUs). This enabled more sophisticated applications of machine learning (ML) methods to classify images of galaxies. The first applications of CNNs in this aspect of astronomy began in 2017, with a deep CNN used to classify galaxies achieving a historical 97% accuracy achieved by Khalifa et al. (2017). However, a key criticism of neural networks has always been the "black box" nature of their predictions. Several methods, such as feature visualization and gradient ascent, have been developed to provide a window into the decision process of neural networks. In 2022, a paper by Bhambra et al. explored the interpretability of CNN results applied to the Galaxy Zoo dataset. The model used explainable artificial intelligence (XAI) saliency maps to determine the features examined by a CNN in the classification process (Bhambra et al., 2022).

In 2019, Walmsley et al. (2019) proposed a Bayesian CNN and an active learning approach to predict the posterior probability of the volunteer responses to specific questions in the Galaxy Zoo dataset. While most existing neural networks tend to provide affirmative predictions on galaxy classification, the Bayesian CNN introduced by Walmsley et al. (2019) provides a probabilistic prediction on volunteer response rates, allowing for a more nuanced and probabilistic understanding of the classification results. Until now, no previous work has been done on applying XAI techniques to a Bayesian CNN. We introduce a novel adaptation of saliency maps to a probabilistic response as opposed to the traditional categorical one.

Preprint. Under review.

2 Methods

2.1 Data and Model

The dataset used for analysis in this report is part of the Galaxy Zoo (GZ) Dark Energy Camera Legacy Survey (DECaLS) campaign. Galaxy Zoo provides a website interface for volunteers or "citizen scientists" to give classifications of galaxy images through a series of questions for different GZ datasets. [1] More details on the dataset can be found in Appendix A.

In this project, we used the pre-trained Bayesian CNN model ZoobotTree, which was developed by Walmsley et al. (2019). More information regarding the model architecture can be found in Table 1 Appendix A. In this project, we primarily do analysis using the GZD-5 dataset, a widely-used benchmark dataset for galaxy morphology classification as described in further detail in Appendix A. The pretrained checkpoints were used to load ZoobotTree and make predictions based on the vote counts and output the Dirichlet concentrations.

2.2 Selecting Galaxies

After preparing the dataset, we used the pre-trained model to make predictions on the GZ DECaLS data, which we then converted into predicted fractions. We then used the cross-entropy loss to rank the galaxies by prediction performance, as well as the question that was causing the highest contribution to the cross-entropy loss (this became our specified question). This metric allowed us to identify the most problematic galaxies, where the model predictions diverged from human labeling. Out of the top 10, we selected one that was appropriate for visualization and applied a similarity score to find controls. One galaxy assumed the predicted fraction under the specified question was correct, and the other assumed the originally observed fraction under the question was correct. By analyzing these images, we can determine which regions or features of the images are most impactful in activating the model. For the mathematical details, see Appendix E.

2.3 Saliency Map

After selecting the galaxy images and corresponding questions, we used saliency mapping to investigate the features that the Bayesian CNN prioritizes. Saliency mapping is a visualization technique that helps explain CNN behavior. We compared the resulting saliency maps to interpret the information conveyed by the model and determine the factors contributing to the image's high CE loss.

To further analyze the model's prediction, we compared the saliency maps of the selected image and question pair with those of images with the highest similarity scores in terms of the observed and predicted fractions. This allowed us to observe the expected saliency maps for the given question under correct or model-predicted circumstances, using images that resemble human observations or the model's predictions, respectively. We chose to implement HiResCAM as our saliency mapping method; further details can be found in Appendix D.

2.3.1 Adaptation of HiResCAM On Concentration Problem

Our selection of the last encoder block as the target layer for saliency maps was informed by the network architecture and its capacity to capture pertinent features for the task at hand. Focusing on this layer enabled us to glean insights into the high-level features that the model utilized for decision-making, leading to improved performance and a better understanding of its behavior.

In adapting HiResCAM for our Bayesian CNN model, we found it necessary to modify the traditional target intended for binary classification problems. As our model output comprises multiple binary outputs, we created a customized multi-label target to hone in on the specific prediction output of interest, yielding saliency maps that targeted our desired question.

Originally designed to accept binary categories of 0 or 1, the BinaryClassifierOutputTarget was repurposed for the regression problem, enabling visualization of the features that influenced increases or decreases in concentration prediction. By leveraging this modified target, we generated informative saliency maps that provided crucial insights into the decision-making process of our Bayesian CNN model and the features driving its behavior.

[1] Link to the Galaxy Zoo website: https://www.zooniverse.org/projects/zookeeper/galaxy-zoo/

3 Results & Discussions

3.1 Chosen Galaxies

To identify the most relevant questions for our analysis, we plotted histograms and carefully examined their distributions. Appendix C includes the histogram plots and a list of the selected questions, as well as the plots for all histograms. We then focused on the top ten galaxies with the highest CE loss and selected "J113514.30-060216.9" for further investigation, as it was associated with the highest CE contribution resulting from the question "how-rounded_round". The saliency map plots for this galaxy proved to be particularly informative.

We further analyzed "J113514.30-060216.9" by identifying the galaxy with the most similar observed vote fractions to the original image's observed vote fractions, *except* for the "how-rounded_round" question. For that specific question, we wanted observed vote fraction to be similar the predicted vote fraction. We found that "J111606.95+122707.8" fit this criterion. We also identified the galaxy with similar predicted vote fractions to the original image's predictions, but with a prediction fraction that matched the observed vote fraction of the original image for "how-rounded_round". This galaxy was "J153503.65+222540.6". Appendix F includes a table of the votes for selected labels of these galaxies, as well as an analysis of its characteristics.

3.2 Fraction Data for the Chosen Galaxy

Galaxy ID	smooth-or-featured_total-votes	smooth-or-featured_smooth	smooth-or-featured_featured-or-disk	disk-edge-on_total-votes	disk-edge-on_yes	disk-edge-on_no	how-rounded_total-votes	how-rounded_round	how-rounded_in-between	how-rounded_cigar-shaped
(selected) J113514.30-060216.9	5	1	4	4	4	0	1	1	0	0
observed similar J111606.95+122707.8	4	2	2	2	2	0	2	0	1	1
predicted similar J153503.65+222540.6	23	13	4	4	0	4	13	8	4	1

Table 1: Vote Count for Selected Questions of Galaxy J113514.30-060216.9 and its related galaxies.

From Table 1 above, we can see that the vote counts for our selected galaxy are very low at a maximum of 5 votes. This is a very possible explanation for why our cross-entropy loss for the galaxy is so high. Similarly, for our predicted similar galaxy, we can see that it also has very few vote counts, indicating that few votes count significantly impacts the predictive ability of the Bayesian CNN. In terms of the actual labeling of the galaxy characteristics, we can see that based on the votes, the galaxy belongs under "featured-or-disk", and its picture is taken at an "edge-on" angle, with one vote in favor of it being "round". However, from the data structure decision tree in Appendix A, we can see that "round" and "edge-on" are mutually exclusive properties, which makes this "round" vote especially questionable. This is a very likely contributor to the high cross-entropy loss.

Galaxy ID	smooth-or-featured_smooth_fraction	smooth-or-featured_featured-or-disk_fraction	how-rounded_round_fraction	how-rounded_in-between_fraction	smooth-or-featured_smooth_pred_frac	smooth-or-featured_featured-or-disk_pred_frac	how-rounded_round_pred_frac	how-rounded-dr5_in-between_pred_frac
(selected) J113514.30-060216.9	0.2	0.8	1	0	0.556839	0.41319	0.009965	0.02888
observed similar J111606.95+122707.8	0.5	0.5	0	0.5	0.474629	0.488553	0.011328	0.024166
predicted similar J153503.65+222540.6	0.565217	0.173913	0.615385	0.307692	0.454655	0.349772	0.541971	0.201332

Table 2: Observed and Predicted Fractions for Selected Questions of Galaxy J113514.30-060216.9 and its related galaxies.

From Table 2 above, we can see that for the question to which we added weight, the predicted fractions for both the "selected" and "observed similar" are very close. Under the column "how-rounded_round_pred_frac", we can see that the model gave a prediction of 0.009965 and 0.011328, indicating that our algorithm is indeed picking up the similarities we want. Sim-

ilarly, for the observed fractions, we can see that under "how-rounded_round_fraction", the "selected" galaxy has a value of 1 while the "predicted similar" galaxy has a value of 0.615385. For question "smooth-or-featured_smooth_fraction", the "observed similar" galaxy has a more similar value with the "predicted similar" galaxy, indicating that the volunteer vote for the original "selected" galaxy might be problematic (especially since there are more votes in favor of "smooth-or-featured_featured-or-disk_fraction").

3.3 Saliency Map

The saliency maps provide an explanation of the features that contribute to an increase or decrease in the concentration prediction. In the saliency maps, red pixels represent the regions with the highest importance, while blue pixels represent those with the least importance. These saliency maps were generated using the models described in Section 3.6.

Highest CE Loss Image The findings presented in Figure 1 offer insight into the performance of our galaxy image analysis model. Specifically, we examine the saliency map generated for the image with the highest cross-entropy loss for the question "How round is it?" with the expected answer of "round". Our saliency maps highlight areas of the image that drive an increase or decrease in the concentration prediction. Analysis of the saliency map for increasing concentration reveals rounded shapes in the corner of the image, aligning with the expected response to the question. However, this shape is not readily apparent to human eyes, raising questions about the model's ability to detect arbitrary round shapes in the galaxy image. Conversely, the saliency map for decreasing concentration shows the model focusing on smaller, streak-like galaxies lacking rounded shapes.

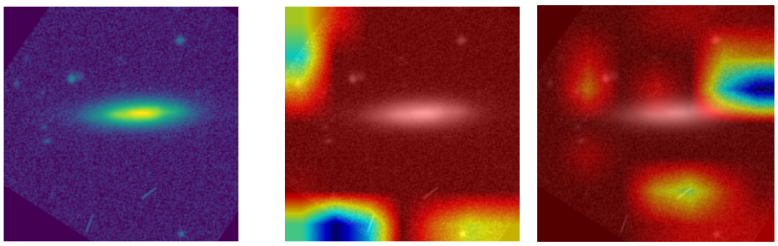

Figure 1: The galaxy image(Id: J113514.30-060216.9 shown on left) and HiResCAM saliency maps generated for the image with the highest cross-entropy loss on the question *"How round is it"* with answer *"Round"*. The saliency maps provide an explanation of the features that contribute to an increase (middle) or decrease (right) in the concentration prediction.

Highest Similarity Score based on Observed Fraction Image In Figure 2, we present the saliency map for the observed fraction image of the galaxy. The saliency map generated for increasing concentration reveals an abundance of rounded shapes in the image, indicating that the model accurately detects rounded galaxies when answering "Round" to the question. This finding supports the reliability of our predictions. Conversely, the saliency map for decreasing concentration shows the model focusing on smaller shapes located on the right-hand side of the galaxy, compared to the main galaxy. This is consistent with the findings from the saliency map with the highest cross-entropy loss, further confirming our observations.

Highest Similarity Score based on Predicted Fraction Image Figure 3 presents the saliency map for the predicted fraction image of the galaxy. Analysis of the saliency map generated for increasing concentration reveals rounded shapes on smaller galaxies, suggesting that the model accurately detects rounded galaxies in the image. This finding reinforces the reliability of our prediction, which matches the observed fraction obtained from the original image. Conversely, the saliency map for decreasing concentration indicates that the model focuses on galaxies in the middle of the image with a straight-line shape, consistent with the human visual perception of the image. This observation confirms that the model's ability to detect rounded-shape galaxies in the image is crucial for accurate predictions. However, when comparing the overall image, we can see that the predicted most similar image is highly different from the original image.

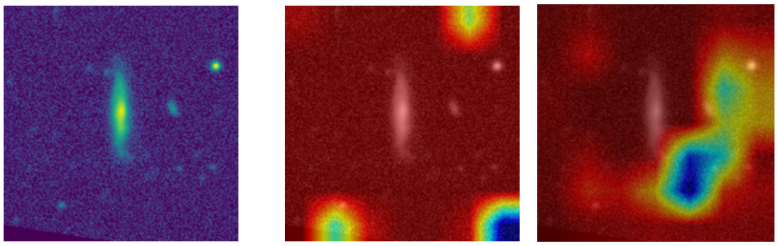

Figure 2: The galaxy image(Id: J111606.95+122707.8 hown on left) and HiResCAM saliency maps generated for the image with the highest similarity score based on True Fraction image.

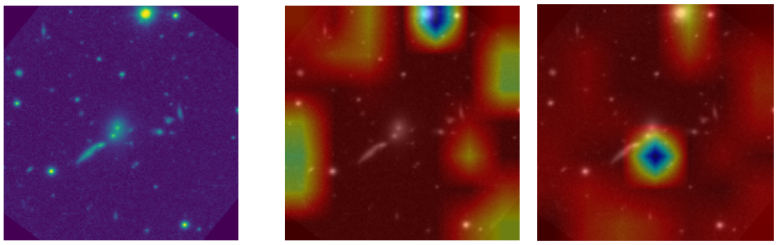

Figure 3: The galaxy image(Id: J153503.65+222540.6 shown on left) and HiResCAM saliency maps generated for the image with the highest similarity score based on True Fraction image.

4 Conclusion

In conclusion, our analysis indicates that the model performed well in the question "how-rounded_round", as the predicted fractions for that question agree with the observed "control" image. We can visually confirm that the two images are similar, indicating that humans did make an error in labeling the original image as "round", for the singular votes. However, when looking at the prediction "control" image, we see that while the predictions are all similar, the images are extremely different. We can see from the fraction data that although their observed shapes both tend toward "round", and the rest of their prediction fractions are similar, there must be something going awry in the prediction as the images do not visually agree at all. We found this to be consistent with other galaxies as well. Further investigation into why this visual discrepancy exists between similar predictions is needed. Returning to errant human error votes, we propose that leveraging the model's output to identify and assign lower weights to such noisy data during the training process could be a viable approach to mitigate the impact of inaccuracies. Finally, we utilized saliency maps to gain insights into the model's decision-making process, which provided further evidence of the model's ability to recognize key features in the images.

However, our study also has some limitations. The primary hurdle is the treatment of the data, which is organized using a decision tree. This leads to dependencies between columns that can cause issues in the interpretation of the results. For instance, in the case of a single vote causing a significant CE loss, we cannot justify simply removing the corresponding galaxy since it may have other questions with a high number of responses. We suggest setting a threshold for data with more total votes for future analysis.

In conclusion, our study provides valuable insights into the behavior of Bayesian CNN models for galaxy classification and highlights the importance of carefully considering the quality and quantity of data for accurate model training and interpretation.

Contributions

Tong Su Code for saliency map, saliency map method and results writeup

Yutong Han Code for cross-entropy loss and histogram plots, cross-entropy loss writeup, galaxy characteristic analysis

Jisu Qian Code for similarity score, similarity score writeup, understanding Bayesian CNN and converting the output of the model

Feifan Xiang Writeup for abstract, model, dataset, Dirichlet loss construction, reference, understanding Bayesian CNN

Acknowledgements

We would like to thank Professor Joshua Speagle of the Statistics and Astronomy Departments for his guidance and mentorship of our project. We would also like to thank Dr. Michael Walmsley of the University of Manchester for his support in helping us understand Zoobot.

Appendix A

Data

The DECaLS images provide deeper details of the galaxy images that reveal features that were previously not visible for Sloan Digital Sky Survey (SDSS) images. The volunteer responses to those questions were recorded as a decision tree and used for classifying the galaxy images. GZD-5 collected classification responses for 262,000 DECALS DR5-only galaxies from March 2017 to October 2020. It collected 30 to 40 volunteer classifications for each image and was later introduced to an active learning system where it only collected 40 responses for the most informative galaxies and 5 for the rest. GZD-5 also contains a better-structured decision tree for improving identifications regarding mergers and weak bars. The questions and decision tree in GZD-5 are shown in the figure below.

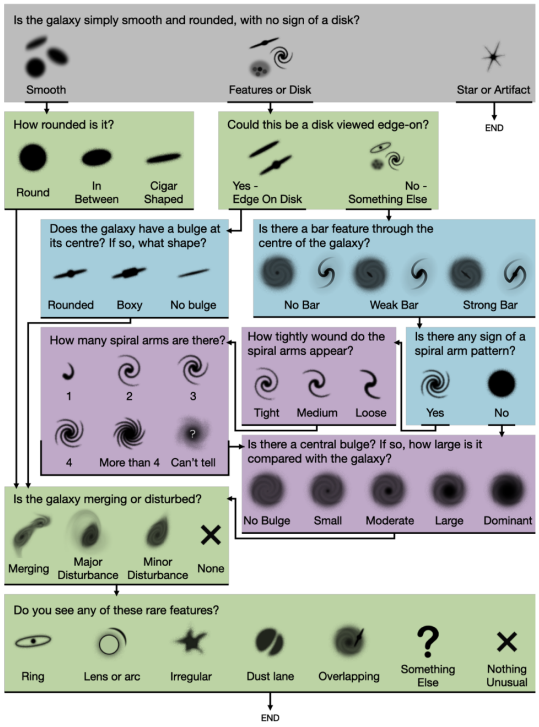

Figure 4: GZD-5 classification decision tree, figure by Walmsley et al. (2021). Questions shaded with the same colour are on the same level of the decision tree.

Model

The default model architecture EfficientNet - B0 was first introduced by Tan and Le (2020) and was made to achieve a multi-objective neural architecture search that can be both efficient and accurate. The detailed baseline networks of EfficientNetB0 are shown in Table 1 in Appendix A. A Monte Carlo dropout is adopted in the Bayesian CNN model to marginalize all possible neural networks and improve the prediction on posteriors. Furthermore, a default dropout rate of 0.2, a learning rate of 0.001, a weight decay of 0.01, etc were used in the ZoobotTree model during our analysis. A more detailed description of the model architecture can be found in the Zoobot documentation. [2], more mathematical explanation about the model setup follows below. All models including ZoobotTree in the Zoobot package are pretrained on the GZ Evo dataset, which contains 550k galaxy images and 92M votes drawn from every major Galaxy Zoo campaign including GZ DECaLS.

[2]Link to Zoobot Documentation: https://zoobot.readthedocs.io/en/latest/index.html

Stage i	Operator $\hat{\mathcal{F}}_i$	Resolution $\hat{H}_i \times \hat{W}_i$	#Channels $\hat{C}_i$	#Layers $\hat{L}_i$
1	Conv3x3	224×224	32	1
2	MBConv1, k3x3	112×112	16	1
3	MBConv6, k3x3	112×112	24	2
4	MBConv6, k5x5	56×56	40	2
5	MBConv6, k3x3	28×28	80	3
6	MBConv6, k5x5	14×14	112	3
7	MBConv6, k5x5	14×14	192	4
8	MBConv6, k3x3	7×7	320	1
9	Conv1x1 & Pooling & FC	7×7	1280	1

Table 3: Baseline Network Architecture of EfficientNet-B0, table by Tan and Le (2020)

Appendix B: Binomial and Dirichlet Loss Construction

In the previous paper by Walmsley et al. (2019), the volunteer responses were categorized into binary classes, considering one answer as a positive response and treating the rest as negative responses for multiple choice questions. Assuming there was some true vote fraction ρ in volunteer responses, let $\hat{\rho} = f^w(x)$ be the maximum likelihood estimate for the true latent variable ρ. The number of positive volunteer responses k can be estimated using the binomial distribution,

$$p(k|x, w) = \text{Bin}(k|f^w(x), N)$$

where x is the galaxy image, w is the neural network weights and N is the total number of volunteer responses. The loss function could be constructed by maximizing the likelihood of this probability. However, this approach didn't perform well for extreme fractions such as 0 or 1.
In a more recent paper, Walmsley et al. (2022) adapted a different approach which predicts $p(\rho|f^w(x))$ which used Binomial and Beta distribution for binary questions, and multinomial and Dirichlet distribution for multiple answer question.
For binomial questions,

$$p(\rho|f^w(x)) \propto \int p(f^w(x)|\rho)p(\rho)$$

where $p(f^w(x)|\rho)$ can be parameterized by Binomial distribution and $p(\rho)$ can be estimated by $\hat{\rho} = f^w(x) = (\hat{\alpha}, \hat{\beta})$ using Beta distribution. Therefore,

$$\mathcal{L} = p(\rho|f^w(x)) \propto \int \text{Bin}(k|\rho, N)\text{Beta}(\rho|\alpha, \beta)d\alpha d\beta$$

For multiple answer questions, multinomial distribution can be used to estimate $p(f^w(x)|\rho)$, Dirichlet$(\vec{\rho}|\vec{\alpha})$ can be used for estimating $\hat{\rho} = f^w(x) = (\hat{\alpha})$,

$$\mathcal{L} = p(\rho|f^w(x)) \propto \text{Multinomial}(\vec{k}|\vec{\rho}, N)\text{Dirichlet}(\vec{\rho}, \vec{\alpha})d\vec{\alpha}$$

The pretrained Zoobot model we used in this project uses this latter-stated Dirichlet loss function.

Appendix C: Histogram Plots

The selected questions are: 'smooth-or-featured_smooth', 'smooth-or-featured_featured-or-disk', 'has-spiral-arms_yes', 'has-spiral-arms_no', 'bar_weak', 'bar_no', 'bulge-size_large', 'bulge-size_moderate', 'bulge-size_small', 'how-rounded_round', 'how-rounded_in-between', 'spiral-winding_tight', 'spiral-arm-count_2', 'merging_none'. The histograms for the selected questions can be found below.

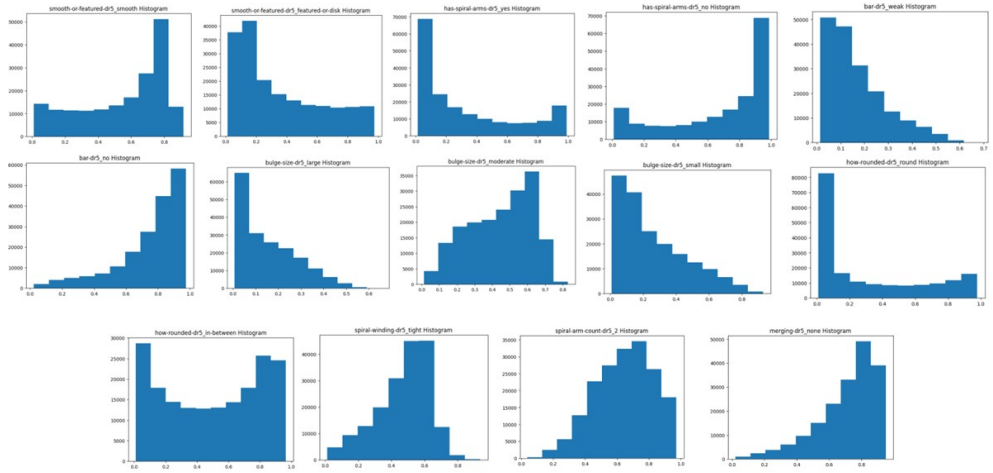

Figure 5: Histograms of emphasized questions. Note the even distributions across the range of fractions. The histograms are counting the frequency of appearance for human vote fractions appearing in each bin.

The histograms for all 34 questions can be found below.

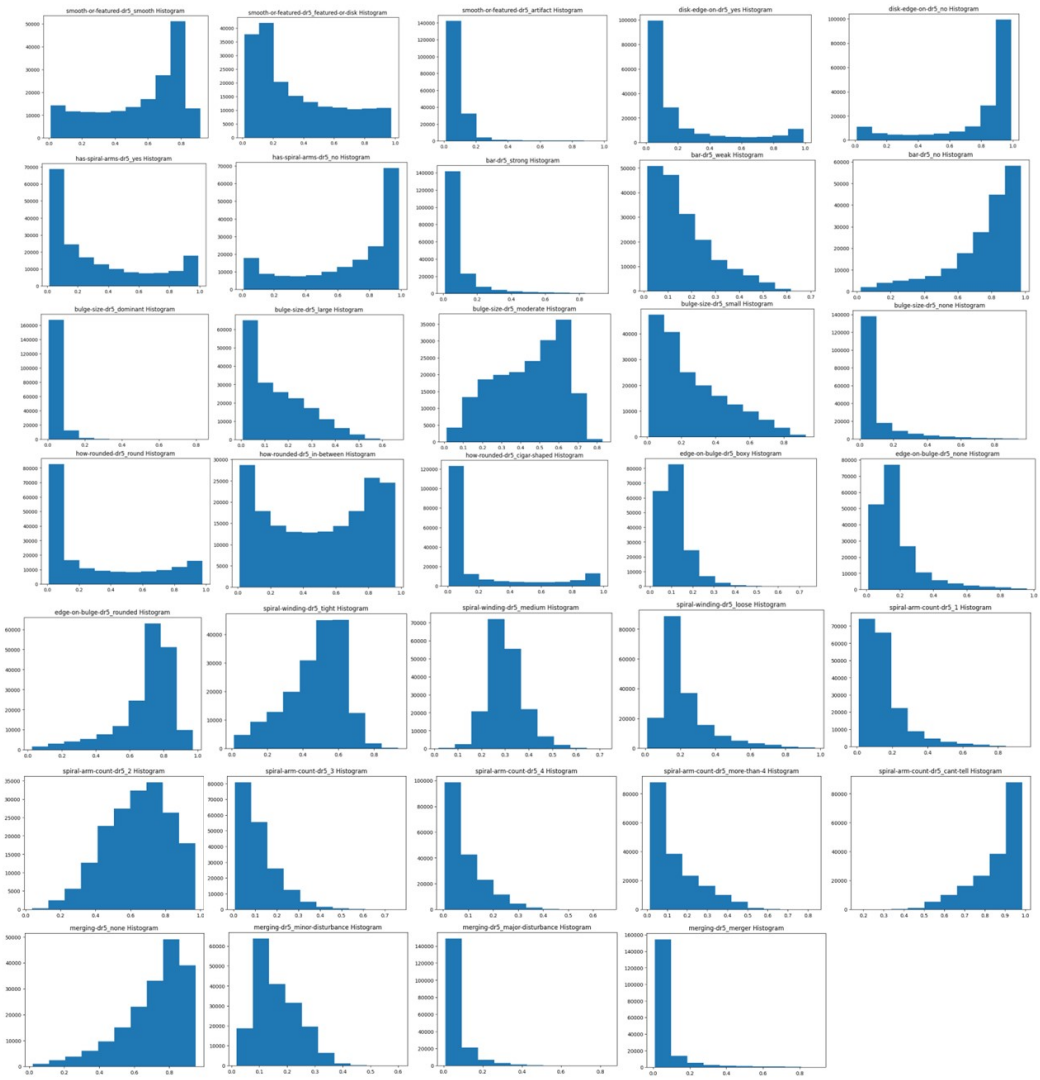

Figure 6: Histograms of all questions. Note the uneven distributions for certain questions. The histograms are counting the frequency of appearance for human vote fractions appearing in each bin.

Appendix D: Choice of Saliency Map

Sanity Check Several methods have been developed to create saliency maps, each using different strategies to highlight specific regions of images that the CNNs pay more attention to. To ensure the effectiveness and reliability of the chosen method, we conducted a sanity check based on the study by Adebayo et al. (2018). The authors demonstrated that saliency map techniques like Guided GradCAM and Guided Backpropagation can still generate seemingly convincing explanations even when models are initialized with random weights or trained on data with randomized labels. This suggests that the effectiveness of these saliency map methods is not strongly influenced by the model's specific parameters or the training data (Adebayo et al., 2018). To avoid such failures, we chose a method that passed the sanity check, which is GradCAM. It has been proven to rely on the trained model's weights and the relationship between training examples and their labels (Adebayo et al., 2018).

GradCAM vs HiResCAM Recent studies in 2021 reveal a critical issue with GradCAM, where it tends to identify regions of an image that the model did not use to make its predictions (Draelos and Carin, 2021). As a result, GradCAM is an untrustworthy technique for explaining model predictions (Draelos and Carin, 2021). To address this limitation, HiResCAM has been developed. It draws inspiration from GradCAM while addressing its inherent weakness. HiResCAM guarantees that only the locations utilized by the model for decision-making will be highlighted (Draelos and Carin, 2021). Since HiResCAM is adapted from GradCAM, we believe it will pass a similar sanity check and thus become the saliency map method of choice for our analysis.

Package To produce the saliency map, we used the GradCam package implemented by Jacob Gil. The package is available on GitHub and provides an efficient and user-friendly implementation of the GradCAM algorithm.[3]

Definition of HiResCAM HiResCAM (High-Resolution Class Activation Mapping) is a powerful technique for generating class activation maps from a CNN. Similar to Grad-Cam, it uses the gradients of the target class flowing into the final convolutional layer to identify the regions of the image that are most important for the prediction (Draelos and Carin, 2021). However, the main difference is that HiResCAM uses element-wise multiplication between the activation and the gradients, instead of averaging the gradients over the spatial dimensions (Draelos and Carin, 2021). This modification results in a more fine-grained visualization of the regions used by the model for decision-making.

Appendix E: Data Preparation

4.1 Concentration Conversion to Fraction

Bayesian CNN models volunteer responses as being multinomially distributed and to make the model probabilistic, we use the Dirichlet distribution to parameterize (See Appendix A). The output of the Bayesian CNN model is the concentration parameters $\alpha_1, \cdots, \alpha_k$ of the Dirichlet distribution. However, for further analysis, we need the predicted fraction of each answer, so we calculate the expected value of the Dirichlet distribution, which is $\frac{\alpha_1}{\sum_i \alpha_i}, \cdots, \frac{\alpha_k}{\sum_i \alpha_i}$ as the predicted fraction.

4.2 Cross-Entropy Loss Calculation

4.2.1 Cross-Entropy Loss Definition and Application to our Data

We determined the performance of the Bayesian CNN model using cross-entropy loss (CE loss). While the model predicts probability concentrations for 34 questions (classes), each question corresponds to a binary response (i.e., a yes/no response). More details regarding the CE loss calculation can be found below. For each galaxy, we selected the highest CE loss value, which we collected to then rank from highest to lowest. Those at the top were the most "problematic" galaxies, where the CE loss was highest, and where the model predictions therefore diverged from human labeling. These were the galaxies that were most interesting to us, as we can then pass their images through a saliency map to see why the model was wrong.

We used the following Cross Entropy (CE) loss formula for our calculations: (Bishop, 2016)

$$L_{log}(y, p) = -y \log(p) - (1 - y) \log(1 - p).$$

Here we define y as the observed fraction of votes for a particular question, while p is the predicted fraction of votes (calculated using the predicted concentrations). We calculated the CE loss for each question against all 33 others for a given galaxy. Note that the Dirichlet distribution ranges from 1 to 101, and therefore our logarithms will always stay within bounds.

Therefore, the CE loss for our problem becomes a matrix of dimensions [num galaxy $\times$ num questions].

[3] https://github.com/jacobgil/pytorch-grad-cam

4.2.2 Weight Chosen for a Certain Question

Since GZD-5 follows a decision-tree format for the questions, this means that questions that come later in the tree will have fewer responses, resulting in greater discrepancy (and therefore higher CE loss for just one or two errant human responses). Additionally, when calculating for a similar galaxy, it will be more difficult to find another galaxy with the same CE losses, if we have a sparse response distribution. Thereforehhnhj, we wanted to place a heavier emphasis on questions that had a more balanced distribution. We chose these questions based on their observed fraction histograms, which we will show in the Results section below.

For this specified list of questions, we assigned a heavier factor to the CE loss. We chose a factor of 5 for the weight since the highest CE prior to weight assignment was ~ 5. Therefore, this weight assignment would place our questions approximately within the same order of magnitude as the most discrepant cases.

4.3 Similarity Score Calculation

To better understand why our model is performing poorly on a particular image, we aim to identify other images that are similar to the original image but on which the model performs differently for our area of interest. By analyzing these images, we can determine which regions or features of the images are most impactful in activating the model.

To achieve this, we first seek out an image that has similar observed vote fractions as the original image but whose observed vote fraction matches the predicted vote fraction of the original image. Additionally, we search for another image with similar predicted vote fractions as the original image, but whose observed vote fraction matches the observed vote fraction of the original image.

To compare the vote fraction distributions of two images, we utilize CE loss as our metric. We view the output for each answer to each question as an independent binomial distribution, which allows us to calculate the CE loss using the formula:

$$\sum_i p_i log(q_i) + (1 - p_i) log(1 - q_i)$$

where p and q represent the two vote fraction distributions being compared. Our function can set a specific weight for a particular question of interest, which we use in the same manner as when calculating the CE loss.

References

J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. Sanity checks for saliency maps. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/294a8ed24b1ad22ec2e7efea049b8737-Paper.pdf.

P. Bhambra, B. Joachimi, and O. Lahav. Explaining deep learning of galaxy morphology with saliency mapping. *Monthly Notices of the Royal Astronomical Society*, 511(4):5032–5041, 02 2022. ISSN 0035-8711. doi: 10.1093/mnras/stac368. URL https://doi.org/10.1093/mnras/stac368.

C. M. Bishop. *Pattern recognition and machine learning*. Springer New York, 2016.

R. L. Draelos and L. Carin. Use hirescam instead of grad-cam for faithful explanations of convolutional neural networks, 2021.

N. E. M. Khalifa, M. H. N. Taha, A. E. Hassanien, and I. M. Selim. Deep galaxy: Classification of galaxies based on deep convolutional neural networks. *CoRR*, abs/1709.02245, 2017. URL http://arxiv.org/abs/1709.02245.

M. C. Storrie-Lombardi, O. Lahav, J. Sodré, L., and L. J. Storrie-Lombardi. Morphological Classification of galaxies by Artificial Neural Networks. *Monthly Notices of the Royal Astronomical Society*, 259(1):8P–12P, 11 1992. ISSN 0035-8711. doi: 10.1093/mnras/259.1.8P. URL https://doi.org/10.1093/mnras/259.1.8P.

M. Tan and Q. V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020.

M. Walmsley, L. Smith, C. Lintott, Y. Gal, S. Bamford, H. Dickinson, L. Fortson, S. Kruk, K. Masters, C. Scarlata, B. Simmons, R. Smethurst, and D. Wright. Galaxy Zoo: probabilistic morphology through Bayesian CNNs and active learning. *Monthly Notices of the Royal Astronomical Society*, 491(2):1554–1574, 10 2019. ISSN 0035-8711. doi: 10.1093/mnras/stz2816. URL https://doi.org/10.1093/mnras/stz2816.

M. Walmsley, C. Lintott, T. Géron, S. Kruk, C. Krawczyk, K. W. Willett, S. Bamford, L. S. Kelvin, L. Fortson, Y. Gal, W. Keel, K. L. Masters, V. Mehta, B. D. Simmons, R. Smethurst, L. Smith, E. M. Baeten, and C. Macmillan. Galaxy Zoo DECaLS: Detailed visual morphology measurements from volunteers and deep learning for 314000 galaxies. *Monthly Notices of the Royal Astronomical Society*, 509(3):3966–3988, 09 2021. ISSN 0035-8711. doi: 10.1093/mnras/stab2093. URL https://doi.org/10.1093/mnras/stab2093.

M. Walmsley, I. V. Slijepcevic, M. Bowles, and A. M. M. Scaife. Towards galaxy foundation models with hybrid contrastive learning, 2022.