RNAAS RESEARCH NOTES OF THE AAS

Comparison of Convolutional Neural Networks and Random Forest Classifiers for Strong Gravitational Lens Identification

DRAFT VERSION FEBRUARY 3, 2024
Typeset using LATEX **modern** style in AASTeX631

Comparison of Convolutional Neural Networks and Random Forest Classifiers for Strong Gravitational Lens Identification

MAADHAV KOTHURI,[1] SURAIN SAIGAL,[1] AND SASIDHAR AYYALASOMAYAJULA[1]

[1]*University of Texas at Austin*

ABSTRACT

Strong gravitational lenses have been instrumental in providing insight into various astronomical problems, including analyzing the dark matter distribution of the universe. Effective identification of these events is made possible through machine learning algorithms, with many recent studies focusing on neural networks. However, very few have investigated the tradeoffs between different algorithms besides neural networks for lens identification. Our paper compares a convolutional neural network and a random forest classifier to determine the benefits of each for this task. We find that while convolutional neural networks do achieve higher accuracy, using random forest classifiers to supplement them could increase the effectiveness of such algorithms. As a result, models that utilize both side-by-side to make predictions may increase in accuracy. This should be explored by future research.

1. INTRODUCTION

Strong gravitational lensing is a phenomenon in which gravitational fields of foreground objects bend light from background objects to produce multiple images, arcs, and rings. Studying strong lensing systems has many applications, ranging from constraining the Hubble constant to studying dark matter distributions. Due to these applications, identifying these systems can have great impacts on astronomical progress (Tyson et al. (1990); Yao-Yu Lin et al. (2020); Huang et al. (2020)).

As the amount of astronomical data increases at a rapid pace, methods to identify lensing events must become increasingly accurate. Many efforts have mainly utilized some combination of neural networks to accurately learn lensing features (Davies et al. (2019); Pourrahmani et al. (2018); Rojas et al. (2022)). These studies have reached high levels of accuracy and discovered new lenses at high levels of confidence. As a result, improving models further necessitates investigating other machine-learning techniques. One such class of algorithms is random forest classifiers, which, while not normally used for image classification, are adept at breaking down high-dimensional data. Our study compares the performance between a convolutional neural network (CNN) and a random forest classifier (RFC) to determine the tradeoffs between the two algorithms for strong lens identification.

In doing so, we hope to improve the process of lens identification to make better use of astronomical data.

2. DATA COLLECTION METHODS

The dataset to train these models is made up of 68,000 64x64 pixel, black-and-white, close-up images of both lens and non-lens objects. Since this study aims to determine the ability of RFCs and CNNs to identify lenses, using images focused on the object itself helped to isolate that objective. This also simplified and streamlined the collection of data, as did the decision to only include single-lens systems in the study.

The data came from three main sources: the Sloan Digital Sky Survey (SDSS), the Sloan Lens ACS Survey (SLACS), and strong lensing simulation software. 60,000 stars and galaxies were chosen from SDSS for the non-lens portion of the dataset (Kollmeier et al. (2019)). Galaxies within 0.5 arcseconds of a lens galaxy listed in SLACS were excluded from the non-lenses. This aimed to minimize the number of lenses within the SDSS data. Additionally, galaxies with a redshift greater than 1 were excluded from our dataset.

The lensing images were obtained from SLACS and *Paltas*, a Python package for lens simulation (Wagner-Carena et al. (2022); Birrer & Amara (2018); Birrer et al. (2021)). The SLACS lenses were made up of 80 grade A, B and C lenses, and 1000 lenses were obtained from *Paltas* to supplement them (Bolton et al. (2006)). To randomly generate lenses in *Paltas*, the parameters (ellipticity, center coordinates, etc.) of each simulation were chosen from a normal distribution centered on the default values provided. These included a Sersic profile with a radius of 0.35 arcseconds and index n = 3, an Einstein radius of 1.1 arcseconds, and a width of 0.03 for the Gaussian used. There were only 1080 lensing images from both SLACS and simulations in total, so we augmented each image through three rotations and flips at each orientation (including the original). This resulted in 8640 total lensing images, which were downsampled to a size of 64x64 pixels.

3. RESULTS

We utilized TensorFlow to build a CNN made up of alternating 2D Convolution, Max-pooling, and Dropout layers, where the dropout rate was 0.2 (Abadi et al. (2016, 2015)). Furthermore, our CNN utilized the Adam Optimizer and ReLU as its activation function. Our random forest was made up of 250 estimators with no limit for the depth of each decision tree. It was first trained on a dataset consisting of satellite images of water bodies and subsequently trained on our lenses training set (Kshetri (2023)). Both models used a batch size of 32 and 10 epochs during training, as well as a 70-15-15 split for the training, validation, and testing set, respectively.

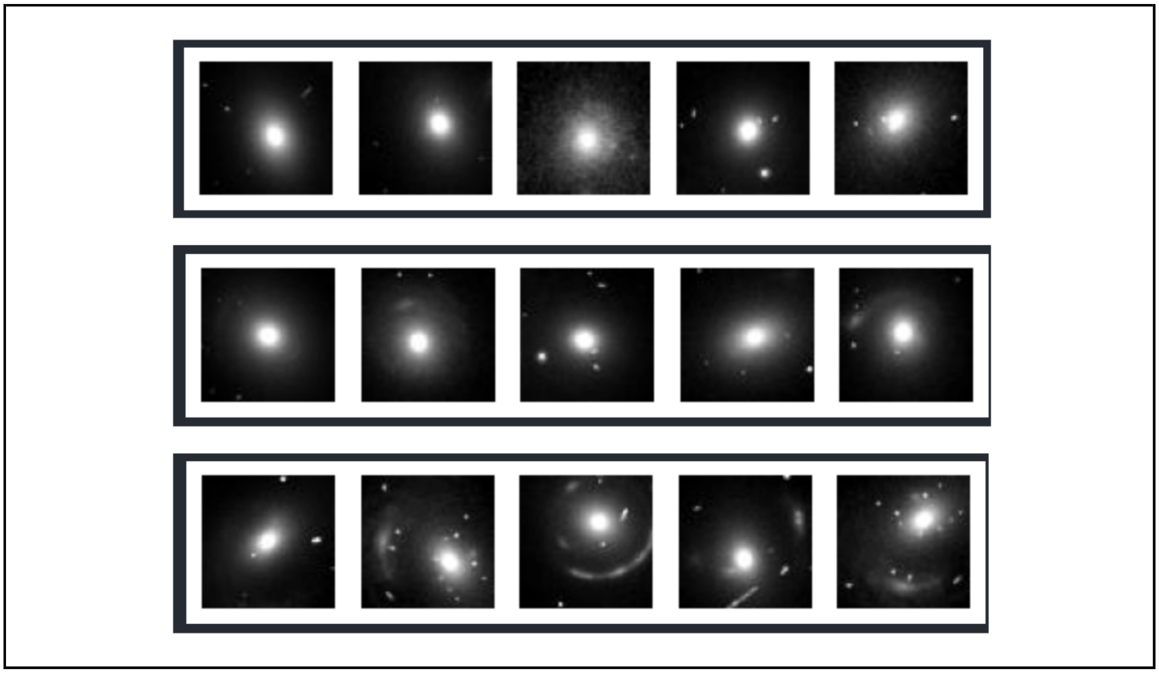

Figure 1: Misclassification Examples
Top Row: 128 lenses misclassified by both the CNN and the RFC.
Middle Row: 122 lenses misclassified only by the CNN. These lenses are harder to distinguish from non-lens galaxies.
Bottom Row: 178 lenses misclassified only by the RFC. These lenses contain more distinct features of strong lensing.

After training, the CNN and RFC resulted in very high accuracies, implying that some degree of overfitting occurred. As a result, their performance metrics cannot be usefully compared. It is worth mentioning that different configurations (lesser number of augmentations on lensing samples, reduced number of non-lenses, and different batch sizes) did not have a significant impact on model performance.

Our true findings lie in the inspection of the misclassified lenses, which demonstrated a difference in performance between both methods. As shown in Figure 1, 128 of the lenses were misclassified by both models. However, there were an additional 178 lenses that only the RFC misclassified, which, upon visual inspection, were found to generally contain easily discernible features. To be more specific, these images contained features like Einstein rings that are easily recognizable even by humans. In contrast, the CNN failed on 122 examples that generally contained features more difficult to identify, such as galaxies whose light is merely distorted and lack the presence of multiple images or rings. This implies a difference in accuracy between both models based on the prominence of lensing features. Our study shows that while the CNN is much better for grasping the most important features of strong gravitational lenses, the RFC can cover the gaps in identifying the more obscure lenses. As mentioned previously, most studies on this subject have focused on neural networks. While there are some, such as Pinciroli Vago et. al, that use other algorithms, it is largely as an intermediate step in their model architecture (Pinciroli Vago & Fraternali (2023)).

Thus, we propose a model that trains a CNN and an RFC side-by-side and makes a final prediction based on a weighted output of both models. While this may lead to longer training times due to the use of two models, this design could leverage the difference in learning between CNNs and RFCs to increase model accuracy past current benchmarks. Including different types of input data, such as spectra or redshift, may also impact model performance. Additionally, utilizing different types of neural networks in this configuration, such as residual neural networks, may yield interesting results. Further research should investigate this, as well as combinations of other algorithms, to continue improving strong lens identification models.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., et al. 2015

Abadi, M., Barham, P., Chen, J., et al. 2016

Birrer, S., & Amara, A. 2018, Physics of the Dark Universe, 22, 189, doi: 10.1016/j.dark.2018.11.002

Birrer, S., Shajib, A. J., Gilman, D., et al. 2021, Journal of Open Source Software, 6, 3283, doi: 10.21105/joss.03283

Bolton, A. S., Burles, S., Koopmans, L. V. E., Treu, T., & Moustakas, L. A. 2006, The Astrophysical Journal, 638, 703, doi: 10.1086/498884

Davies, A., Serjeant, S., & Bromley, J. M. 2019, Monthly Notices of the Royal Astronomical Society, 487, 5263, doi: 10.1093/mnras/stz1288

Huang, X., Storfer, C., Ravi, V., et al. 2020, The Astrophysical Journal, 894, 78, doi: 10.3847/1538-4357/ab7ffb

Kollmeier, J., Anderson, S. F., Blanc, G. A., et al. 2019, 51, 274. https://ui.adsabs.harvard.edu/abs/2019BAAS...51g.274K

Kshetri, T. B. 2023, water bodies in satellite imagery. https://www.kaggle.com/datasets/tekbahadurkshetri/water-bodies-in-satellite-imagery

Pinciroli Vago, N. O., & Fraternali, P. 2023, Neural Computing and Applications, 35, 19253, doi: 10.1007/s00521-023-08766-9

Pourrahmani, M., Nayyeri, H., & Cooray, A. 2018, The Astrophysical Journal, 856, 68, doi: 10.3847/1538-4357/aaae6a

Rojas, K., Savary, E., Clément, B., et al. 2022, Astronomy & Astrophysics, 668, A73, doi: 10.1051/0004-6361/202142119

Tyson, J. A., Valdes, F., & Wenk, R. A. 1990, The Astrophysical Journal, 349, L1, doi: 10.1086/185636

Wagner-Carena, S., Aalbers, J., Birrer, S., et al. 2022, From Images to Dark Matter: End-To-End Inference of Substructure From Hundreds of Strong Gravitational Lenses, doi: 10.3847/1538-4357/aca525

Yao-Yu Lin, J., Yu, H., Morningstar, W., Peng, J., & Holder, G. 2020, Hunting for Dark Matter Subhalos in Strong Gravitational Lensing with Neural Networks, doi: 10.48550/arXiv.2010.12960