A Machine Learning Approach to the Detection of Ghosting Artifacts in Dark Energy Survey Images

Chihway Chang\textsuperscript{a,b}, Alex Drlica-Wagner\textsuperscript{c,b,a}, Brian Nord\textsuperscript{c,b}, Donah Michelle Wang\textsuperscript{d}, Michael H.L.S. Wang\textsuperscript{e}

\textsuperscript{a}University of Chicago, Chicago, IL 60637, USA
\textsuperscript{b}Kavli Institute for Cosmological Physics, University of Chicago, Chicago, IL 60637, USA
\textsuperscript{c}Fermi National Accelerator Laboratory, Batavia, IL 60510, USA
\textsuperscript{d}Illinois Mathematics and Science Academy, Aurora, IL 60506, USA

Abstract

Unwanted artifacts that often plague astronomical images arise from a number of sources that include imperfect optics, faulty image sensors, cosmic ray hits, and even airplanes and satellites. One major source that is difficult to avoid is known as ghosting, and is caused by the scattering and internal reflections of light off of the telescope’s mechanical and optical components. Detecting ghosting artifacts efficiently in the large cosmological surveys that will acquire petabytes of data can be a daunting task. In this paper, we use data from the Dark Energy Survey to develop, train, and validate a machine learning model to detect ghosts based on convolutional neural networks. The model architecture and training procedure is discussed in detail, and the performance on the training and validation set presented. Testing is performed on data and results are compared with those from a ray-tracing algorithm. A proof-of-principle is demonstrated that shows promise for Stage IV surveys and beyond.

Keywords: Machine Learning, Image Artifacts

1. Introduction

When the Dark Energy Survey (DES) \cite{1, 2} completed its mission in January 2019, it had mapped \(~\text{5000 square degrees}\) of the southern sky using the 570 Megapixel DECam imager mounted on the Blanco 4-m telescope at the Cerro Tololo Inter-American Observatory in the Chilean Andes. Over the course of 758 nights of data taking spread across 6 years, DES generated a massive \(~\text{2 petabytes}\) of data. Unfortunately, due to the nature of the DECam optical system, the DES data are subject to imaging artifacts caused by scattered and reflected light from bright stars. Examples of such artifacts are shown in Figure 1. In this article, the terms “ghosts” or “ghosting artifacts” will be used broadly to refer to all such artifacts. While all astronomical objects produce scattered and reflected light at some level, this study is specifically focused on identifying ghosts that are prominent enough to have a negative impact on object detection algorithms, background estimation, and photometric measurements. Due to the large volume of DES data, identifying ghosts by eye is extremely labor-intensive and impractical. DES has automated the ghost detection procedure through the development of a ray-tracing algorithm that combines a model of the camera optics, the telescope pointing, and the known locations and brightness of stars to predict the presence and location of ghosts in an exposure. While this algorithm correctly identifies and localizes a significant number of ghosting artifacts, it is limited by the accuracy of the optical model, telescope pointing, and external stellar catalog. Because the ray tracing algorithm does not use the DES imaging data directly, it misses a substantial number of ghosting artifacts. There is clearly a need for more effective methods to address this problem, especially in light of future cosmic surveys like the Large Synoptic Survey Telescope (LSST), which will have a field of view three times as large as DECam and will acquire \(~\text{20 terabytes}\) of data per night \((\sim \text{60 petabytes over ten years})\).

In this work, we explore the use of modern machine learning (ML) methods as a potential solution to the problem of efficiently detecting ghosting artifacts in large scale optical-imaging surveys. Though ML methods have been in use for over half a century, we are referring specifically to the advances made in the past two decades in this field. These were made possible by the confluence of several key factors that included...
(a)  (b)  (c)  

Figure 1: Example full focal plane DECam images that exhibit ghosting artifacts.

(1) a deeper understanding of the internal workings of the visual cortex [3], (2) the introduction of convolutional neural networks (CNNs) inspired by the visual cortex [4], (3) the development of practical techniques to train such networks [5], and (4) the availability of vastly increased computational power from devices like graphics processing units (GPUs). This note demonstrates a proof-of-principle for the viability of ML techniques for identifying artifacts from scattered and reflected light in large cosmic surveys.

2. Machine Learning Approach

2.1. Model Architecture

The design of the machine learning (ML) model used here is shown in Figure 2. It utilizes a typical stacked architecture consisting of a sequence of alternating convolutional and pooling layers, that is then terminated by a series of fully connected layers [4, 6]. The network is composed of four 2D convolutional layers, each followed by a maximum pooling layer. The number of output filters in the sequence of four convolutional layers are 16, 32, 32, and 64, respectively. Filters in all four convolutional layers have kernel sizes of $3 \times 3$, stride lengths of one, and use Rectified Linear Unit (ReLU) activation functions. The pool sizes used in the pooling layers are $4 \times 4$ for the first layer and $2 \times 2$ for all subsequent layers. Stride lengths for all pooling layers correspond to their pool sizes. The final two layers of the network, following the fourth pooling layer, are fully connected (FC) layers. The first FC layer has 128 neurons with ReLU activation functions and the last FC layer has 2 output neurons using SoftMax activation functions. “Dropouts” are performed prior to each FC layer in which a fraction (0.4 and 0.8 for the first and second FC layers, respectively) of the inputs are randomly ignored. This method lessens the chances of overfitting by minimizing co-adaptations between layers that do not generalize well to unseen data [7]. The total number of parameters in the model is 1,212,578.

2.2. Training and Evaluating the Model

2.2.1. Training Set

The images used for training the model were derived from 800 × 723 pixel, 8-bit grayscale images in the portable network graphics format, covering the full DECam focal plane. The training set consisted, initially, of equal portions of images that had ghosting artifacts (positives) and images that did not (negatives). The portion with ghosts, totalling 2,389 images, was taken from a “blacklist” of DECam images from all DES data taking periods predicted by the ray-tracing program to exhibit such artifacts. After excluding the images in the blacklist, an equal number of images were then randomly selected from the remainder of the full data set, to form the second portion of the training set.

Prior to feeding the images to the network, they were first downsampled to 400 × 400 pixels, which is the input size of the first convolutional layer. The pixel values in each image were then normalized to a range whose minimum and maximum corresponded, respectively, to the first quartile $Q_1(x)$ and third quartile $Q_3(x)$ of the full distribution in the image, by multiplying each pixel value $x_i$ by a factor $s_i = \frac{x_i - Q_1(x)}{Q_3(x) - Q_1(x)}$. In order to augment the training set size, the images were also randomly flipped either along the horizontal axis by reversing the ordering of pixel rows, or along the vertical axes by reversing the ordering of pixel columns.
2.2.2. Model Training Procedure

The model was trained using 80% of the sample described in the previous section and the remaining fraction was set aside for validation. Optimal weights for the model were obtained using Adam [8], a version of the mini-batch stochastic gradient method that uses dedicated learning rates for each parameter and adapts their values based on their history. The weights were updated iteratively in randomly picked batches of 32 images (batch size), completing a full pass over the entire sample in one epoch, of which a total of 30 were performed. The loss function used was categorical cross-entropy, calculated according to

\[
L = -\sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij} \cdot \log(p_{ij}),
\]

where the index \(i\) runs over the number of observations \(N\), and the index \(j\) is taken over the number of classes \(M\). \(p_{ij}\) is the probability and \(y_{ij}\) is either 0 or 1, depending on whether class \(j\) is the correct classification for observation \(i\).

Upon examining the false positives and false negatives after training, it was found that some images were mislabeled. This was because, as described above, images labeled in the training set as not having ghosting artifacts were initially selected from images in the DES dataset that formed the complement of those predicted by the ray-tracing program to contain such artifacts. As it turned out, many of these “clean” images actually contained ghosting artifacts. When images that were positively identified by the ray-tracing program were inspected, the opposite case was also found to be true – some images labeled as having ghosts did not exhibit detectable ghosting artifacts. Therefore, several iterations were required in order to fix the mislabeled images and repeat the entire training process of 30 epochs in order to avoid overfitting and improve accuracy.

2.2.3. Training and Validation Performance

The final results of training are shown in Figures 3 and 4. The two plots in Figure 3 show the evolution of the training accuracy (left) and loss (right) over the epochs. The validation curves follow the training curves closely, indicating no overfitting. Accuracies of over 94% are achieved on both training and validation sets at the end of 30 epochs.

Figure 4-a plots the receiver operating characteristic curve (ROC) for the trained model in orange, showing the true positive rate versus the false positive rate. The area under the curve (AUC) for the ROC curve is 0.987, indicating good separation between the two classes of images. For comparison, the blue diagonal line shows the case when a model has absolutely no discriminating power between classes where AUC = 0.5.

Figure 4-b plots the confusion matrix. The values in the first row represent the number of true negatives in the first column and the number of false positives in the second column. The values in the second row represent the number of false negatives in the first column and the number of true positives in the second column.

2.3. Circumventing the Lack of a Test Set in the Context of a Feasibility Study

Because the training sample was constructed on the basis of the blacklisted exposures identified by the ray-tracing algorithm, the sample size was barely enough for training and validating the model. Unfortunately,
Figure 3: The plots above show how the accuracy (left) and loss (right) evolved versus epoch during the training and validation stage.

Figure 4: The orange ROC curve showing the discriminating power (AUC=0.987) between the two classes is plotted in (a). The blue line represents the reference case of no discriminating power (AUC=0.5). The confusion matrix is shown in (b) with the number of true negatives and positives shown, respectively, in the upper left and lower right boxes. The number of false positives and negatives are shown, respectively, in the upper right and lower left boxes.
this meant the lack of a separate testing set, with images correctly pre-labeled as having ghosts or not, for independently determining reliable performance metrics. Therefore, the reader should be cautioned that the model performance, derived from the training and validation phase described in Section 2.2.3, may represent an overestimate of true performance.

Ultimately, however, the purpose of this work is to find a more efficient way of identifying ghosts in survey data so that the affected CCDs in the exposures can be flagged before undergoing image processing. Hopefully, this will lead to higher quality data from which important science results can be extracted. With this ultimate goal in mind, the immediate task at hand is to determine whether ML methods offer a viable and practical solution that deserves further exploration.

In order to address the goals outlined above without a testing set, we apply the model to the DES data to do a direct comparison with the ray-tracing algorithm, for which a better alternative is being sought. The details and results of this study and comparison are described in the following section.

3. Testing the Trained Model on DES Data and Comparing with the Traditional Method

The ML model trained according to the details described in Section 2.2.1 was used to perform inference on the DES Year-5 data set consisting of 23,755 full focal plane DECam images with exposure numbers ranging from 666747 to 724364, which were prepared using the procedure described in Section 2.2.1. This set also included the Year-5 images that were used in the training and validation stage. For each image, the model was used to predict whether it contained ghosting artifacts or whether it was free from such artifacts. The model positively identified 3,285 images as containing artifacts.

For comparison, the number of images classified as containing artifacts by the ray-tracing program described in Section 2.2.1 for this same range of exposures was 259, of which 241 were also classified as containing artifacts by the ML model. Of the 3,044 images that were positively classified exclusively by the ML model, ~700 were “false positives” exhibiting nearly imperceptible or no sign of ghosting artifacts. For the 18 images that were positively classified exclusively by the ray-tracing program, 8 were found to be “false positives”.

4. Conclusion

We have successfully applied a machine learning based method to identify DES images containing ghosting artifacts. Not only was the ML method able to identify nearly all of the images containing artifacts that were previously identified by the traditional ray-tracing method, it also identified ~10x more affected images with a false positive rate of ~3%. To our knowledge, this is the first successful attempt to use ML methods for identifying ghosting artifacts in optical telescope images from a cosmic survey. It serves as a proof-of-principle that lays the foundation for possible future enhancements such as the ability to identify the location of the artifacts within the image. It will also benefit future cosmic surveys like the LSST which will be faced with the challenge of even larger data sets.

5. Acknowledgements

This collaborative work was carried out as part of an Illinois Mathematics and Science Academy (IMSA) Student Inquiry and Research (SIR) project. We are grateful to Dr. Steve Kent of Fermilab for his technical review of this writeup. We wish to thank Dr. Don Dosch, Dr. David Devol, and Dr. Eric Smith of IMSA for overseeing the SIR program and making this collaboration possible. We also wish to thank the staffs of Fermilab’s experimental astrophysics group and IMSA’s SIR office for their support.

Funding for the DES Projects has been provided by the U.S. Department of Energy, the U.S. National Science Foundation, the Ministry of Science and Education of Spain, the Science and Technology Facilities Council of the United Kingdom, the Higher Education Funding Council for England, the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign, the Kavli Institute of Cosmological Physics at the University of Chicago, the Center for Cosmology and Astro-Particle Physics at the Ohio State University, the Mitchell Institute for Fundamental Physics and Astronomy at Texas A&M University, Financiadora de Estudos e Projetos, Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro, Conselho Nacional de Desenvolvimento Científico e Tecnológico and the Ministério da Ciência, Tecnologia e Inovação, the Deutsche Forschungsgemeinschaft and the Collaborating Institutions in the Dark Energy Survey.

The Collaborating Institutions are Argonne National Laboratory, the University of California at Santa Cruz, the University of Cambridge, Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas-Madrid, the University of Chicago, University College London, the DES-Brazil Consortium, the University of Edinburgh, the Eidgenössische Technische Hochschule
Figure 5: The images above are examples of DES Year 5 images predicted by the ML model described in this paper to exhibit ghosting artifacts, but which were not identified by the ray-tracing algorithm as such. Figures (a) to (d) show examples that have actual ghosting artifacts, while figures (e) and (f) are examples of the ~23% described in the text that either do not exhibit ghosting or have negligible levels.
6. Contributions

The author list is ordered alphabetically. C. Chang was responsible for the initial idea of applying ML techniques to artifact detection and served as primary mentor to D.M. Wang in the SIR project. A. Drlica-Wagner identified optical ghost detection as an important problem that could benefit from an alternative approach. He also ran the ray-tracing algorithm, provided access to labeled images containing ghosts, and advised on their classification and causes. B. Nord provided initial training on ML algorithms and advised on algorithm design. He also served as the primary liaison between the astrophysics group at Fermilab and IMSA. D.M. Wang was responsible for extending the sample ML models provided by B. Nord to develop the models used in this work. She also performed the training and validation of the model, including selecting its set of hyperparameters. M.H.L.S. Wang prepared the samples used for training and comparisons, based on the data A. Drlica-Wagner provided access to. He also advised on the preprocessing of the data and prepared the initial version of this document, including the model architecture diagram.

References


