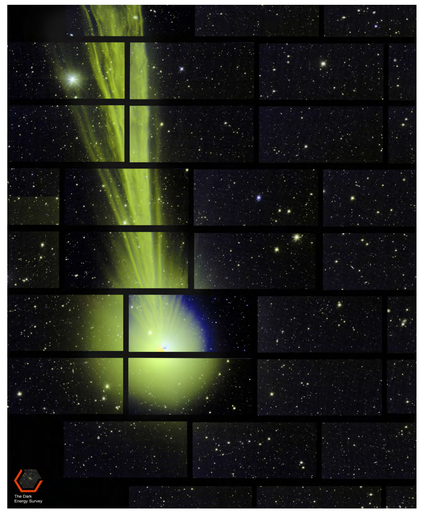
Automated Image Artifact Identification in Dark Energy Survey CCD Exposures

Joseph DeRose and Warren Morningstar
Kavli Institute for Particle Astrophysics and Cosmology
Stanford University

Abstract



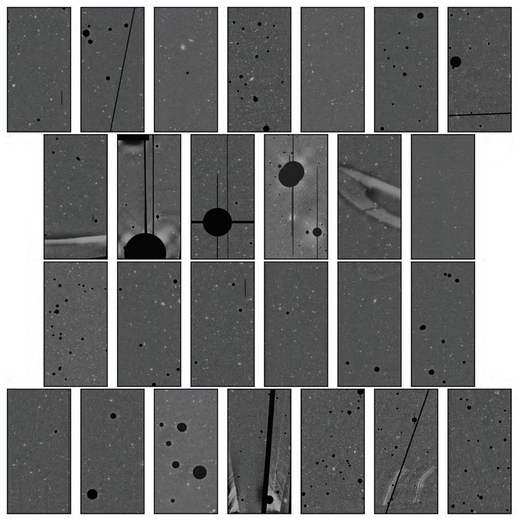
In this work, we present an implementation of a machine learning algorithm to identify and classify image artifacts in observations taken by the Dark Energy Survey (DES). Specifically, we treat our individual pixels as features and use both a Support Vector Machine (SVM) and a Convolutional Neural Network (CNN) to classify full exposures. In general the CNN outperforms the SVM. In the two-class problem and the 29-class problem, our SVM implementation does not perform better than random guessing on our test set, while our CNN gives test accuracies of ~60% on both problems. We discuss sources of systematic error in our models and plans to account for these in future work.

Introduction

In modern cosmology, much attention has been turned toward using large photometric sky surveys to provide strong constraints on structure formation and the composition of our universe. An important component of modern sky surveys is the management of the large volumes of data produced by the survey instruments. Typical data generation of current sky surveys is several GB per night, and future sky surveys will produce TB of data in the same time frame. Because of the vast quantities of data produced, the majority of analysis is performed in an automated (or at least semi-automated) manner.

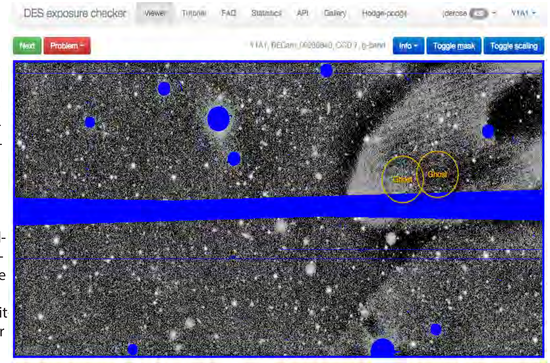
While automation provides significant advantages in performance, an obvious disadvantage is that it can be prone to overlooking subtle features in the data, which may cause systematic errors in measurements of important quantities. In particular, since the goal of many photometric surveys is to provide precise measurements of the brightness of astronomical objects, any spurious brightness variations within a detector have the potential to be interpreted as a real signal, and thus may interfere with the accuracy of measurements.

Brightness variations in a detector can be caused in many different ways, and thus have a diverse variety of appearances. These are usually referred to as image artifacts. Reliable identification of artifacts is an essential component of preparing data for analysis. In this work we have implemented an algorithm for identification of image artifacts found in observations performed by the Dark Energy Survey (DES). DES is a sky survey being carried out using the 4m Blanco Telescope located at the Cerro Tololo Inter-American Observatory in Chile. The survey saw first light in 2012, and is currently in its 3rd year of data taking.

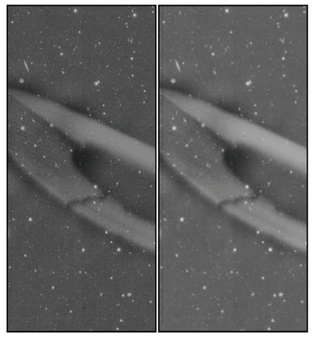


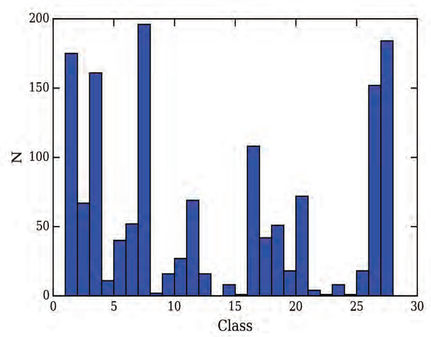
Data and Methodology

The DES camera (DECam) is made up of 62 Charge Coupled Devices (CCDs) each composed of 4098x2048 pixels. A typical night of observations yields approximately 200 new exposures. Currently DES identifies unmasked image artifacts manually via a web based application. Users inspect single CCD exposures, and upon finding an artifact label a single pixel of the exposure with the artifact type. Artifacts are grouped into 28 classes, including some obvious defects such as Cosmic Ray, Airplane, and Satellite, and some more subtle and obscure such as Haze, Dark Halo, and Wavy Sky. We add an additional label, 'No Artifact' as a null result for our classifier.



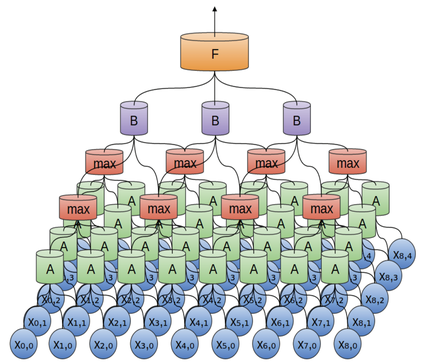
We used the DES web utility to provide us with our training examples. Specifically, our training data consists of all of the exposures obtained in the first year of DES, which have at least one CCD that was manually classified as having an artifact. The total size of this data set is 60856 CCD exposures, obtained from approximately 1000 observations. The number of artifacts classified in the data is approximately 8000, spread across roughly 6% of the CCD exposures.





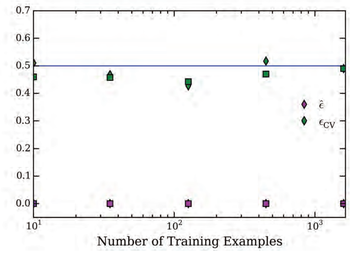
To facilitate image recognition, we employed a Convolutional Neural Network (CNN). CNNs are useful because the convolutional layers extract abstract features from the data, and are invariant to transformations such as rotations or translations. This is vital for our purposes, because artifacts are incredibly diverse in their presentation in the data, and thus we want to be able to extract commonalities between them while ignoring the less important distinctions. One drawback to using CNNs is that the diversity of network architectures is vast, and free exploration has a high rate of failure. Because of its simplicity, we first decided to implement the LeNet architecture (LeCun et al. 1998), which was originally used with a high rate of success on text classification. This architecture consists of a convolutional layer, followed by a max-pooling and a second convolutional and max pooling, which are then fed to a deeply connected layer. We then apply a softmax activation to obtain our class probabilities. We then train by minimizing the cross entropy between these predicted probabilities, and the true probabilities. All the above steps were performed using the newly released open source library TensorFlow (Abadi et al. 2015, TensorFlow white paper) on the Sherlock cluster GPUs at Stanford.

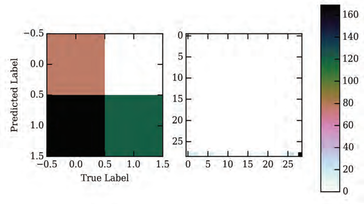
Parallel to this, we implemented a model mildly reminiscent of the ImageNet classification algorithm (Krizhevsky et al., Advances in Neural Information Processing Systems 25 (NIPS 2012)). We used two convolutional layers rather than their five, and two deeply connected layers rather than their three. Additionally, we included three dropout layers, which has been shown to help prevent overfitting.



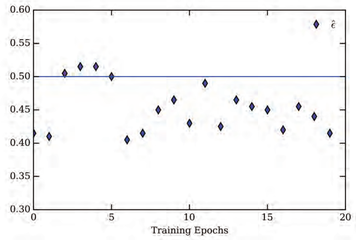
Results

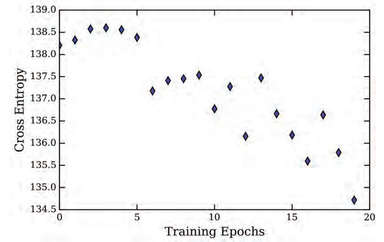
Support Vector Machine

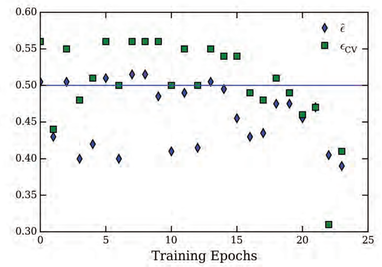




Convolutional Neural Network







Discussion and Future Directions

Rendering of the Large Scale Synompic Survey (ref:lsst.org)

Classifying artifacts will become much more crucial in the LSST era with the volume of images taken increasing by orders of magnitude. Although we plan on persuing improvements to our CNN, machine learning methods may not be able to handle the level of noise in the current data. For the time being, it appears that crowdsourcing is still the better option for the most accurate identification of artifacts.

References

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998d). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278–2324.

Melchoir, P. et. al., Crowdsourcing quality control for Dark Energy Survey. 2015, arxiv:1511.03391