Hayden Foote

ASTR 502

**Paper Summary:** For my term project, I’ve chosen [Tanoglidis et al. 2020](https://arxiv.org/abs/2011.12437), which describes the first application of a convolutional neural network (CNN) to low surface-brightness galaxy (LSBG) classification. LSBGs are expected to be the most common type of dwarf galaxies but finding them in survey data is very challenging primarily owing to the large range of artifacts (blended objects, bright star-forming regions in host galaxies, tidal debris/streams, etc.) that can confuse algorithms designed to search for them. Non-machine-learning based algorithms often perform poorly and require human visual confirmation, which is infeasible with the huge data volumes expected from surveys like DES and VRO/LSST.

This paper describes the development and testing of a CNN called DeepShadows that automates LSBG classification to high accuracy. The data used is a sample of visually-confirmed LSBGs and artifacts from the DES (20,000 examples of each). This dataset is split into training, validation, and test sets. DeepShadows uses image pixels as direct input, and consists of three sets of convolutional and pooling layers, plus two dense layers, outputting a scalar probability that the image contains a LSBG.

DeepShadows is also compared to two other machine-learning pipelines, a Support Vector Machine (SVM) classifier and a random-forest (RF) algorithm. Both of these comparison pipelines use SourceExtractor-derived features instead of raw pixels as input. After all three pipelines have been trained, it is found that DeepShadows achieves an accuracy of 92.0%, compared to 81.9% for the SVM and 79.7% for the RF pipeline. Thus, DeepShadows offers better performance while also having no reliance on data pre-processing.

Lastly, since obtaining training data for this application requires visual identification of LSBGs and is therefore very expensive, the authors investigate the potential for using DeepShadows on other survey datasets (in this case Hyper Suprime-Cam Subaru Strategic Program; HSC SSP) without full re-training. Instead, DeepShadows is employed on a smaller HSC SSP dataset with no additional training and after a small fine-tuning of the parameters (transfer learning) on this new data, demonstrating that transfer learning offers almost the same accuracy (87.6%) as a fully-trained model for much less computational time. The authors then offer a brief discussion of the uncertainties associated with using deep learning to classify LSBGs.

**Code and Data Availability:** All code related to the project is publicly available here: <https://github.com/dtanoglidis/DeepShadows>. The paper utilizes publicly available data from the DES and HCS archives. Code to generate the datasets from the archives is included in the repository.

**Computational Resources:** The authors state that they utilized Colab Pro with GPUs and High-Memory (32 GB) mode to run everything. I’m honestly more worried about the size of the datasets; 40,000 256x256x3 images is quite a lot. I have a personal laptop as well as my Steward desktop to run code on, though I don’t believe either have GPUs, and I would be worried about storage space with either solution. In principle though, we have unlimited google drive space through the university, and I wouldn’t mind using some of my Steward research fund on a few months of Colab Pro to run this. I’m also fairly proficient with our HPC facilities (though have never tried using them for ML) if it comes to that. Please let me know if you think this would be feasible.

**Goal:** This is a fairly long paper with four distinct ML applications: Running the DES dataset with three different algorithms plus the transfer-training of DeepShadows on HSC data. Having no prior experience with ML, I think reproducing all four analyses might be a bit much for a semester. Ideally, I would like to start by reproducing the DeepShadows results on the DES dataset. If time allows, I would also attempt the transfer-training. Alternatively, if you think the DES dataset might be too expensive, I could start with the cross-training. I’m mostly interested in DeepShadows and would like to focus on this rather than the other two pipelines if possible.