Final Report: High-Performance Hybrid Computation Platform for Astronomy Data Classification

PI: Robert J. Brunner
Department of Astronomy, Beckman Institute, NCSA, University of Illinois

Co-PI: Tom Huang
Department of Electrical and Computer Engineering, Beckman Institute, University of Illinois

Collaborators:
Edward J. Kim, graduate student, Department of Physics, University of Illinois
Xianming Liu, graduate student, Department of Electrical and Computer Engineering, University of Illinois

Corresponding Author: Robert J. Brunner, bigdog@illinois.edu

Executive Summary:
In this exploratory project, we examined the efficacy of Blue Waters to perform deep learning of sources directly from astronomical images. With the explosive growth in data volumes with new, large area sky surveys, the rapid and efficient classification of sources into stars, in our own Galaxy, and galaxies spread across the cosmos has become increasingly important. In addition, deep learning has become the standard approach for high performance image classification. Our initial concept was to leverage a well defined, human labeled training data set from the galaxy zoo project to guide the development and acceleration of deep learning techniques to this problem. While our work demonstrated that this approach is viable, we have decided to first refine the approach using newer deep learning techniques before submitting a full Blue Waters allocation request.

Description: This project employs deep learning, which is build on neural networks, to classify light distributions from images. The images are obtained through multiple filters, and thus capture different physical and morphological variations. The final data volume is in the Petascale, thus any final pipeline would need to be efficient at handling and moving large quantities of data from long term storage to GPU-accessible compute nodes. In addition, the deep learning algorithms needed to be efficiently ported to the XK7 nodes and the training labels accurately mapped to the target image data.
The impact of efficient classification, even into just star or galaxy is enormous, as it is the required first step in any subsequent analysis of the imaging data. Furthermore, the problem is confounded by the fact that the number of sources (either stars or galaxies) increases with decreasing signal, thus the vast majority of sources are in the low signal-to-noise regime. Since precision cosmology requires most (or all) sources to be used for measurements, minimizing contamination while preserving sample purity is paramount.
We initially chose Blue Waters as an exploratory platform for this project to ascertain if Blue Waters would efficiently support deep learning at scale. While our codes did work successfully

on Blue Waters, the data management aspect was less than satisfactory. Thus we performed the bulk of our analysis on other machines, including those supported by XSEDE as well as commercial cloud vendors. As a result, we will not be pursuing a full Blue Waters proposal to support the completion of this work. While there are publications resulting from our overall effort, they do not use any results from this exploratory allocation on Blue Waters.

Publications:
None to date.