Investigating Parallel Programming Methods in Notebook-based Programming Environments

Tony Wang

Fri April 27th

1 Summary Paragraph

Notebook-based programming (programs that combine data, visualizations, and code) has become increasingly popular over the past few years, specializing in making the data exploration and initial development processes easier. However, they are poorly suited for high performance computing and a typical "data science" workflow currently requires a second step of rewriting the code in a parallel manner or in an entirely different environment/language. This project will explore a myriad of parallel programming tools available in Python and Jupyter with the specific aim of helping to inform developers in what cases would a notebook-based parallel programming solution suffice or whether a refactor is required. Approaches to explore include Ipython's parallel %magic functions, Cython code optimizations, IPython Parallel, Dask package. These approaches will be evaluated on various machine learning tasks with well-defined benchmarks, such as the Iris flower dataset or the Boston housing market dataset. A final test on a real-world machine learning workflow will be conducted using data from Electronic Health Records (EHR) to predict sepsis.

2 Team

- Tony Wang
- Lawrence Wolf-Sonquim