CS CAPSTONE PROGRESS REPORT

FEBRUARY 17, 2018

WINTER MIDTERM PROGRESS REPORT

PREPARED FOR

NASA JPL

LEWIS JOHN MCGIBBNEY

PREPARED BY

GROUP 28

COALFO

KENNY THOMPSON

BRYCE EGLEY

Abstract

Coal and Open-pit surface mining impacts on American Lands Follow-On (COAL-FO) is the successor project to the 2016-2017 COAL project. COAL initially aimed to deliver a suite of algorithms to identify, classify, characterize, and quantify (by reporting a number of key metrics) the direct and indirect impacts of mining operations and related destructive surface mining activities across the continental U.S (and further afield). COAL successfully delivered a Python library for processing hyperspectral imagery from remote sensing devices such as the Airborne Visible/InfraRed Imaging Spectrometer (AVIRIS) and a Science Data System for running COAL pipelines. COAL-FO will utilize recent funding obtained from a recently awarded NSF-funded XSEDE high performance computing (HPC) grant to further improve, validate and document COAL algorithms, execution runtime performance and geospatial output results.[1]

CONTENTS

1 INTRO

Coal and Open-pit surface mining impacts on American Lands Follow-On (COAL-FO) is the successor project to the 2016-2017 COAL project. COAL initially aimed to deliver a suite of algorithms to identify, classify, characterize, and quantify (by reporting a number of key metrics) the direct and indirect impacts of mining operations and related destructive surface mining activities across the continental U.S (and further afield). COAL successfully delivered a Python library for processing hyperspectral imagery from remote sensing devices such as the Airborne Visible/InfraRed Imaging Spectrometer (AVIRIS) and a Science Data System for running COAL pipelines. COAL-FO will utilize recent funding obtained from a recently awarded NSF-funded XSEDE high performance computing (HPC) grant to further improve, validate and document COAL algorithms, execution runtime performance and geospatial output results. With that in mind, our goals were to accomplish the algorithm improvements and the COAL-SDS implementation.

2 BRYCE EGLEY

2.1 Where we are

To give a brief overview of where we are on COAL-FO our main goals are to create a searchable port for the existing COAL project. This will be augmented to accommodate the desire to (i) port the coal-sds software to the XSEDE platform and undertake test and evaluation of the system performance, (ii) process all AVIRIS and AVIRIS-NG imagery generating and archiving all science data products, and (iii) making the products searchable through a portal. The other main goal is too improve the existing COAL project. I have been focused on improving the existing COAL projectby making modifications to the pycoal library and to the website https://capstone-coal.github.io/ My partner Kenny has been focusing on COAL-SDS which address the other part of our project.Where we want to port data to XSEDE or AWS. So far, we have fixed issues with pycoal and improved algorithms by making the examples easier to run, improving QGIS/GDAL installation instructions to accommodate more systems, create a Command Line Interface(CLI) to make the process of running correlations automated, and upgraded our docker image to python 3.

When we get our project working with XSEDE this will allow us to use 2 Tera-bytes for processing images. We will loop images through XSEDE and have them stored on another database for other users to then access them on the XSEDE platform. This will save a lot of time since other users wont have to deal with generating the images each time and will just have the images ready to go. We hope this will make the COAL reach a wider audience and be used in more research. Just one of the images we have right for the San Juan mine case study has a 17.5 GB file and takes hours to days to run the mineral classification. If we could run these and then store them on XSEDE this would make it a lot easier for other users and wouldnt put time and storage constraints on them that they would otherwise have.

I have been focusing on improving the existing COAL project. I have made the examples easier to run, by fixing an error with anaconda3 and fixing several errors with file locations in the example scripts, improved QGIS/GDAL installation instructions to accommodate more systems so that product imagery created by COAL can be viewed, created a Command Line Interface(CLI) with the long term goal of making the process of running correlations automated, fixed our docker image which was broken and upgraded our docker image to python 3. Kenny has been focused on getting COAL-SDS up and running and taking product imagery I stored on google drive to be exported to AWS, with the eventual goal of using XSEDE. To give an overall timeline of how I have been working on this project. In week two I finished up running the example python scripts and fixed the bugs I encountered in these. The main two bugs that I found were a problem with how our code was interacting with anaconda 3. We had an issues with a variable which

could be either a float or an int but in anaconda3 it was only treated as an int. I was able to fix this issue by doing further research into this issue and then I update the website to address this problem and how to go into anaconda3 and fix it. The other major issue I encountered with the examples was that the files didnt seem to be in the correct locations according to the existing code. I realized in the commit history it didnt look like the previous group had an examples directory so they must have been doing something else with the file locations while they were running pycoal. I fixed these issues so the file locations would be in the correct locations and I made some modifications to the code. In weeks three to five my main focus was on improving the docker image to include python 3, force install apache 2 and improve installation instructions for QGIS and GDAL so that all systems could view the pycoal products. After this I moved on to Creating the CLI which took most of this last week which was week 6. So, overall after I get the examples working in week two I have been able to solve one issue a week since then. I hope to continue this pace of work for the remainder of our capstone project. All of this work can be shown in more detail on the github commit history for pycoal, and capstone-coal.github.io.

2.2 What we have left to do

For the future of pycoal I plan to work on work on 3 issues on the github issue tracker which are Enable use of EcoSIS Spectral Library, Upgrade to use USGS Spectral Library Version 7 and Getting more Data. Upgrade to Spectral Library Version 7 and Enabling use of EcoSIS Spectral Library would both greatly improve our correlation images making them more clear when viewed in QGIS. However, for the San Juan mine case study we currently have data from it doesnt appear that there Spectral Library Version 7 available. We either plan to do more research to see if we can find a Spectral Library Version 7 for this case study or alternatively get more data for another case study were Spectral Library Version 7 is available for use.

2.3 Problems impeding progress

Right now there are a few problems impeding progress. One is that the environmental correlation that we are generating may not be completely correct. It does generate an image where can see the points where the environmental correlation image identifies. However, it seems like there may be a problem with how pycoal is interacting with gdal and osgeo. This may be a problem we need to fix in the future. The other problem is that our docker image is currently broken, our client has said he work on fixing this though. I will continue to work on improving pycoal but I would also like to help with development on COAL-SDS and exporting our classification products to XSEDE Extreme Science and Engineering Discovery Environment and AWS Amazon Web Services. Most of our project goals were based on COAL-SDS and XSEDE so I think it would be better to have more people working on this.

2.4 Other interesting/relevant information

Besides the work Im currently doing I may pick up some issues for our website. There are currently three issues there that need to be addressed. Which would include making our website more viewable for mobile browsers, documenting how we are generating our classification products and then adding classification documentation.

3 KENNETH THOMPSON

3.1 Where we are

While Bryce's focus has been on the algorithm, and improvements to that, the algorithm is not very useful to our goal without implementation of it. My primary focus so far has been the development and the deployment of the COAL-SDS, or science data system, and specifically developing and configuring the apache OODT file management system. Where we are currently at, we have successfully developed and deployed a limited version of the apache OODT file management system to an amazon EC2 instance that is being hosted currently on amazon web services, using the free student tier. This gives us a free tool in which we can test and develop our systems before deploying them to XSEDE where we have limited access. Amazon web services EC2 instance is a Linux machine hosted on the cloud that we can remote into and manage our systems. It has access to a super computers worth of resources, and can process and stage large amounts of data for us. With the successful building and staging of COAL-SDS, we believe this accomplishes a successful Alpha, and gets us very close to a successful Beta. There are some limitations of AWS, they cap you at a relatively small amount of data and storage, and anything beyond that with your student tier, and they expect you to start forking over cash, and its also imperfect as a test for XSEDE, our eventual goal, because of some key differences in setup and operation, but the benefit of not using our limited XSEDE budget on testing far outweighs the slight extra effort that will go into porting things over to the XSEDE system.

3.2 What we have left to do

Now that we have a working build on COAL-SDS, the next step is to take the progress over to the XSEDE. So for our alpha, we wanted to get the COAL-SDS built and working on the AWS EC2 instance. Like I said earlier, this had limitations, and it doesnt port over perfectly, but for the most part it is a straightforward port. The XSEDE HPC resources have given the project a grant of 800 hours on the super computer, enough to process significant amounts of data. We will be plugging in AVRIS imagery that has been collected, and using the command line interface of Apache OODT to run the operations on XSEDE. The command line interface is working on AWS, and now we just need to make whatever modifications to get them working on the XSEDE system. This should be a fairly straightforward process, but might be complicated by some small design changes within XSEDE. Nonetheless, we are in a very good position to do what needs to be done.

3.3 Problems impeding progress

The business of cloud computing and running resources on HPC resources has become a fairly lucrative one. Microsoft does it, amazon does it, and there are tools set up for students such as XSEDE and amazons student tier. Its lucrative because these companies charge steep rates to use their resources, and bill you automatically based on the computing power you use. We had an advantage going into this: a grant from XSEDE for 800 hours of time on their HPC resources. However, it would be inefficient to test our programs on there, at the same time using the resources that could have been better spent on actual data crunching. This creates the problem of porting, there are several key aspects of AWS that are different from XSEDE, and the command line interface specifically has some key concerns that might not function as well on XSEDE. Dealing with those differences will be a definent impediment on progress, but one we are fully capable of handling. Worst case scenario, we can try to maximize the most we can get out of the student tier on amazon, and accomplish our goal, albiet in a somewhat limited fashion, on AWS.

3.4 Other interesting/relevant information

In the background, while this process has been going on, coal mining in the US has skyrocketed, rising up 30% in a single year. This makes our work more relevant and important then ever in order to provide researchers and potentially policy makers access to the kind of high quality info that will allow them to make smart informed decisions. This project is very exiciting and will allow for improvements in the algorithm that allow for incredible increases in performance.