NASA UVA Capstone FAQ

What is the goal of this project?

Our goal is to use satellite measured lidar and polarimetry data to improve the speed at which we acquire surface level air quality products. NASA has an existing model, but it delivers these products at a prohibitive computational efficiency for large scale datasets. Our goal is to use neural networks to speed up this process by a factor of 1,000 or more, while also maintaining or improving upon model accuracy.

What are the forward and inverse problems/models?

At a high level, the forward model goes from the ground to the satellite. We use ground level conditions to compute polarimetry measurements. The inverse problem, which is the problem we want to solve, goes from the satellite to the ground. We are given satellite polarimetry data, and we want to compute ground level conditions.

The forward model takes a state vector as an input and outputs polarimetry measurements. The state vector is a vector of ground level parameters, including aerosol and ocean properties such as aerosol optical depth, which are used to calculate PM2.5. The underlying physics of the inverse problem make it nearly impossible to directly calculate an accurate ground level state vector when given only these polarimetry measurements. Instead, we use an iterative approach to fit an appropriate state vector using the existing forward model, which is very accurate. We use a mildly accurate inverse model to obtain a first guess at the state vector using the satellite data. We then use the forward model to produce satellite polarimetry readings and compare these output readings with the real data. This comparison allows us to calculate a Jacobian and update our state vector. This process is repeated until the output of the forward model converges to the real data and we generate a sufficiently accurate state vector. Each call to the forward model takes something on the order of 45 minutes meaning that the dozens or hundreds of calls needed to fit a state vector might take days to complete. We solve for this bottleneck by replacing the computationally slow forward model with a neural network model, which speeds up the forward model by a factor of 1000. This means each call takes fractions of a second instead of minutes and each retrieval takes minutes instead of hours or days.

What are HDF files?

Hierarchical Data Format (HDF) files are files that are hierarchical in the sense that there are nested levels of information. The file will have some high-level organizational structure each with some associated metadata. Within each branch of this structure can be more nested information each with associated metadata and either more subdivisions or a dataset contained in a multidimensional array. Usually, the dimensions of each array will map to other arrays in the file so that you can access all the information for one observation by looking at the same index within each dataset.

- Documentation for h5py package, Pythonic interface for HDF5 data format, can be found at https://docs.h5py.org/en/stable/
- Documentation for pyhdf package, used to read/write SD files through numpy arrays, can be found at https://pypi.org/project/pyhdf/
  - Relevant python libraries (how to install)

        - Install h5py with "conda install h5py" for anaconda users or "pip install h5py" (conda installation is more likely to work)
        - Install pyhdf with "pip install pyhdf"
        - "import h5py"
        - "from pyhdf.SD import SD, SDC"
    - How to access/read
        - Set HDF file path to variable "f"
        - Create HDF data object with "hdf = SD(f, SDC.READ)"
        - Can now access variables and elements from hdf (ex. "Lats = list(hdf.select('Latitude').get())")
    - How to create/write
        - Create and write HDF data object with "h5py.File('data.h5', 'w')"
        - Can now create datasets within this HDF object (ex. "hf.create_dataset('dataset_1', data=d1)") where d1 is some arbitrary numpy array
- Documentation on how the polarimetry datasets are structured (Abhi)
    - Find general information about POLDER/PARASOL data at https://www.icare.univ-lille.fr/parasol/ and some documentation at https://www.icare.univ-lille.fr/parasol/products/
    - Directories where we found useful data
        - After logging into the FTP server, hdf files that would be useful for making normalized radiance and degree of linear polarization plots like those in normalized_rad_retrieval_viz.ipynb can be found at https://www.icare.univ-lille.fr/data-access/data-archive-access/?dir=CALIOP/CALTRACK-333m_PAR-L1B.v1.00/2010/2010_03_02/

What datasets are we working with?

PARASOL POLARIMETRY DATA

Overview

- Parasol refers to the satellite that carries the POLDER instrument (polarimeter)
- Polarimeter gives relatively good aerosol measurements across a large area
- Can be collocated with LIDAR measurements (see LIDAR data section for additional information) to allow for calculation of additional metrics such as aerosol concentration

Data Acquisition

- Files can be found in HDF format via iCare data archive in Polder folder

SIMULATED [POLARIMETRY] DATA

Overview

- Simulated data refers to synthetic Parasol data that is generated via NASA scripts
- Purpose of simulated data is to create a very robust training dataset for the neural network by iterating through a wide range of polarimetric measurements (whereas real polarimeter data might not be as diverse)

Data Acquisition

- Simulated data is created via NASA “run.py” script (see “Important NASA Scripts section”)

CALIPSO LIDAR DATA

Overview

- Calipso refers to the name of the satellite that carries instruments that generate Lidar data
- Lidar data gives very precise aerosol measurements in a very small area
    - Provides location of aerosols within a vertical space between satellite and earth; specifically count of aerosols that are close to earth wherein humans inhabit
    - Complements polarimetry data and allows for additional calculations such as aerosol concentration
- NOTE: The actual use of Lidar data comes farther downstream, not required for neural network training or full retrieval algorithm

Data Acquisition

- Files can be found in HDF format via iCare data archive in Caliop folder

What are the go to NASA scripts on their internal cluster?

Run.py

- Used to generate simulated polarimeter data
- Path: ~git/ada-rt/rt_code/mpi/gen_simulation_data

PreprocessRawDataPipeline.py

- Processes data to be used for neural net training
- Maps data to input / output / geometry
- Path: /pace-mapp-nn/src/k-cluster/ PreprocessRawDataPipeline.py

Legacy Neural Network

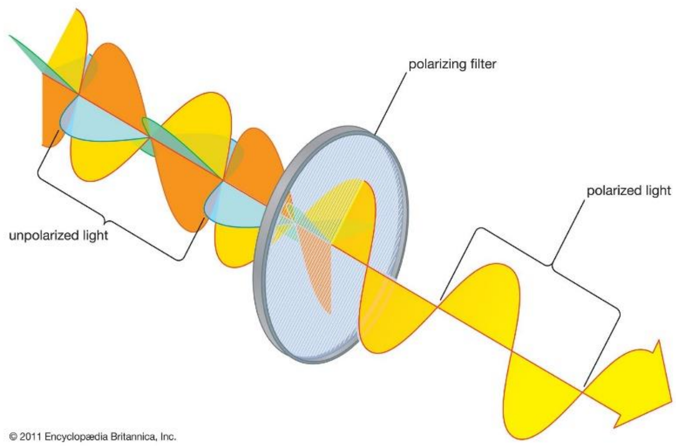
- Refers to neural network architecture that was used
- Path: /pace-mapp-nn/src/networks/forward-vi-11ch-11p-constant-large.py
- Path Model Output: /pace-mapp-nn/output

What is satellite polarimetry? (Written for and by people who have no formal education in satellite polarimetry)

Here we try to give an honest attempt to explain satellite polarimetry like you're five years old (or like you're a master's student without a background in physics or atmospheric science). Please don't treat this explanation as thorough or complete, I'm sure Snorre would cringe at the minimalist/incomplete explanation we give here. I'm also sure that there are things here that display an incorrect or incomplete understanding, but hopefully it serves as a good starting place to jump off of.

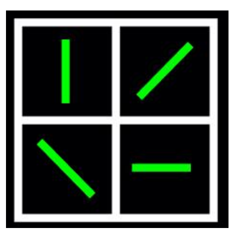
Polarization – "property of certain electromagnetic radiations in which the direction and magnitude of the vibrating electric field are related in a specified way." (https://www.britannica.com/science/polarization-physics)

Basically, polarization of light has to do with the angle at which light oscillates about the axis it is travelling on.



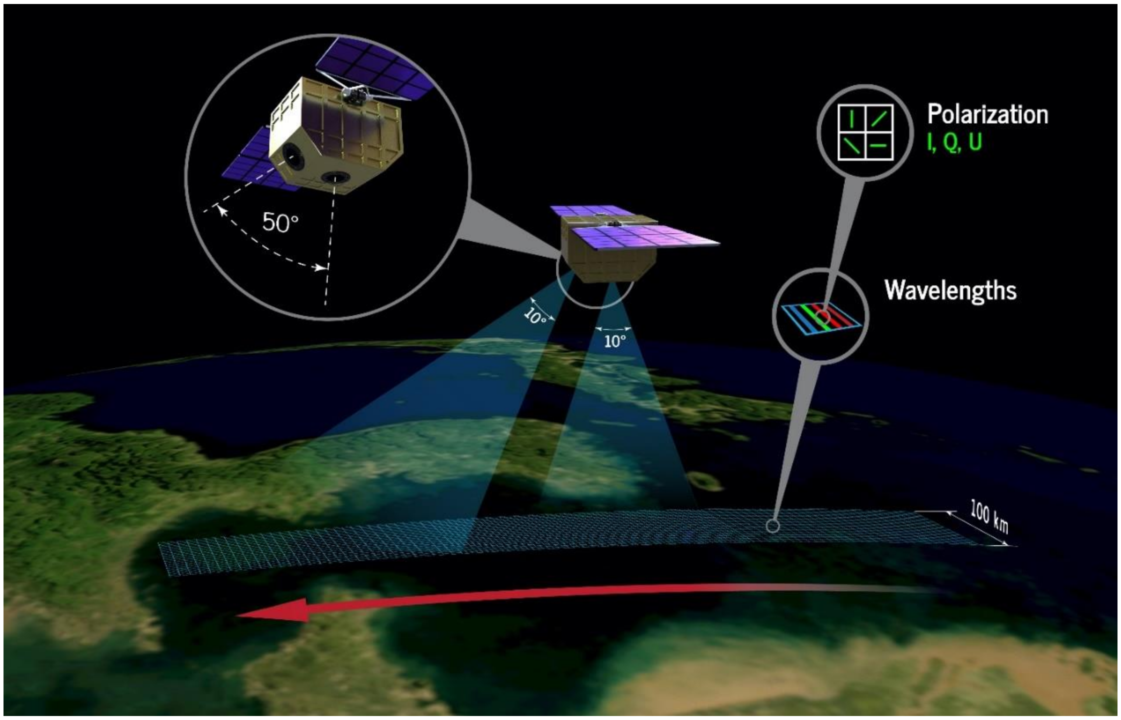
As light is emitted from the sun it's completely unpolarized, meaning that for any individual wavelength of light there is an equal chance that it is oscillating at any angle around its axis of travel. As the light moves through the Earth's atmosphere it will encounter all sorts of gases, aerosol droplets, water vapors, etc. As it moves through some of these materials (like aerosols) the light will start to polarize, meaning that some of these light wavelengths will begin to align with one another. The light is no longer unpolarized since now there are certain angles at which it is more likely the light is oscillating at than others. The light will bounce of the surface of the ocean, reflect, and then travel back through the atmosphere. As it does so it will encounter many of these polarizing substances again which will further polarize the light, aligning those angles of oscillation with one another.

After the light exits the atmosphere it may hit a satellite polarimeter. The satellite polarimeter will look at one wavelength of light (or one specific color) by passing the light through a filter. It will then take 4 measurements, at 0 degrees, 45 degrees, 90 degrees, and 135 degrees (see below).



The 0- and 90-degree measurements are known as the Q stokes parameters. The 45- and 135-degree measurements are known as the U stokes parameters (https://en.wikipedia.org/wiki/Stokes_parameters). From these measurements you can calculate the degree to which the light has been polarized (how much of it has been aligned) (Degree of Linear Polarization – DOLP) as well as the dominant angle to which the light has been polarized. You can also calculate the overall intensity or the I stokes parameter.

The polarimeter aboard the satellite will take these measurements for several different wavelengths as it loops through color filters. In addition, the satellite will also view the same point on the surface of the earth for multiple different angles as it flies over. This allows us to isolate information from the surface of the earth as you observe the effects of different columns of air coming out of the same location on the ground. With several color filters and several viewing angles we are able to build up a large amount of information for a single point on the Earth's surface.



Like I said, a lot of this is probably way oversimplified, and the NASA folks would probably correct some of this but hopefully it can serve as a way to get oriented and start asking Snorre and the rest of the team some semi-informed questions. Good luck!

Links:

http://www.physics.umd.edu/courses/Phys375/AnlageFall05/lab3.pdf

https://www.britannica.com/science/polarization-physics

https://science.larc.nasa.gov/polarimetry/