Name(s)
Cameron C. Jones

Project Number
38519

Project Title
Automated Identification of Organic Molecular Structure and Relative Concentrations from Infrared Spectral Data

Abstract

Objectives/Goals
Most of what we know about the composition of the universe is due to spectroscopy, the measurement of light intensity at varying wavelengths. My project investigates the application of machine learning techniques (convolutional neural networks) to the problem of identifying complex polycyclic aromatic hydrocarbon (PAH) molecules in IR spectroscopic data. I automate the nonlinear mapping of IR spectra to identify chemical composition and relative concentrations of unknown mixtures of molecular compounds found in telescopic data. Such techniques could greatly accelerate the analysis of the data captured from astronomical instruments (such as the upcoming James Webb Space Telescope).

Methods/Materials
My project looked at three specific problems: a) building models to identify PAH molecules from empirical IR spectroscopic data when trained on the approximate theoretical counterpart from NASA#s PAHdb v2 database, b) building models to perform the same task but with NASA#s 5x larger PAHdb v3 database, and c), building models to identify random compositions of theoretical PAH molecules (up to ten) from composite spectroscopic data using the v3 database. I implemented my models with the open source library TensorFlow.

Results
My principal findings are: a) convolutional models can be trained on theoretical spectra to accurately identify empirical PAH molecules (with 73% accuracy), and b) when trained on data for 3,139 PAH molecules, my models can identify the molecular concentrations of random compositions with weight vector correlations of ~85%, and correctly identify the largest constituent ~67% of the time. In all cases, the models dramatically outperform standard linear (logistic) models. These network models could aid scientists in identifying astronomical objects for further study and research, thus greatly amplifying research efficiency.

Conclusions/Discussion
My project proves the utility of convolutional neural networks to analyze telescopic IR spectra. My best model (ResNet5 with ~200M parameters) exceeded the performance of linear (logistic regression) models in all tests. The versatility of these models illustrates their unique ability to recognize patterns in the features of the spectra and to generalize across diverse datasets. The models can significantly increase the efficiency of analyzing empirical IR spectra and understanding the composition of gas clouds, stars, and exoplanets in our universe.

Summary Statement
I demonstrated the effectiveness of convolutional neural networks in identifying the composition of polycyclic aromatic hydrocarbon mixtures from infrared spectra.

Help Received
I created the models independently while consulting with Dr. Partha Bera, a postdoctoral researcher at NASA Ames, who helped me understand the concepts of spectroscopy.