Using Machine Learning to Remove Noise from Stellar Spots in Exoplanetary Data

Exoplanet Demographics. NExSci. 9th - 13th November, 2020

Artash Nath, Grade 8 Student, Toronto, Canada. artash.nath@gmail.com www.HotPopRobot.com Twitter http://www.twitter.com/wonrobot

Abstract

Studying the light curves of exoplanets in different wavelengths allows us to predict the chemical composition of their atmospheres. The presence of stellar spots adds noise to this data. They may overlap with the path of the transiting exoplanet and corrupt the light curves data. It would lead to errors in calculations of the planet-star radius ratios. The effects of stellar spots from those produced by the exoplanetary atmospheres must be removed. The current approach is to identify the effects of the spots visually and correct them manually, or to simply discard the data. I created a hybrid machine learning model to remove the effects of stellar spots in faint signals of transiting exoplanets' atmospheres. My model was able to accurately predict the exoplanet-star radius ratios in 55 wavelengths with a mean square error of 0.001. It can be applied to data gathered from the upcoming European Space Agency's ARIEL Telesope.

Introduction

While data about our universe is increasing exponentially, the astronomy community is not. Custom machine learning models can be employed to extract useful information from big datasets. Machine learning algorithms are good in identifying patterns. Once a machine learning algorithm has been accurately trained to perform a specific function, they can make predictions in newly generated data from telescopes with no supervision.

Problem Statement

Stellar spots add noise to exoplanetary data and lead to errors in the calculation of the planet-star radius ratios.

Hypothesis

Machine learning models can be trained to remove noise from stellar spots in exoplanetary transit light curves. They can then accurately predict the planet-star radius ratios of exoplanets in different wavelengths.

Hybrid Machine Learning Model

As the dataset provided by the ARIEL Telescope was already labeled with the planet-star radius ratios for each exoplanet, I used a supervised machine learning algorithm. As I had to work with both sequential and numerical data, I selected a Hybrid Machine Learning Model. I chose a combination of the Long Short-Term Memory (LSTM) model - a form of Recurrent Neural Network (RNN) as well as the Feed-Forward neural network to make the planet-star radius ratio predictions.

The sequential data (light curves) was handled by the LSTM while the numerical data (stellar and planetary parameters) were handled by the Feed-Forward neural network.

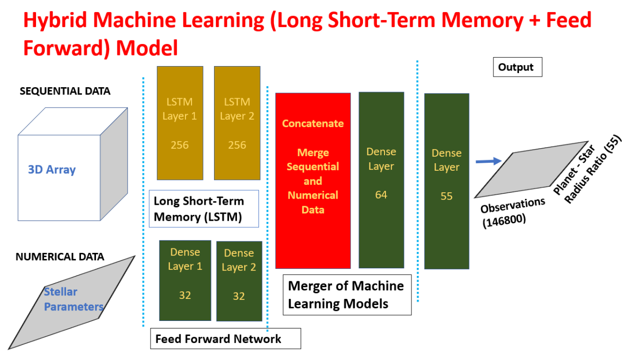


Figure 3: Hybrid Machine Learning Model

Preparing the dataset for applying the hybrid machine learning model was a two-stage process.

Findings and Interpretations

The hybrid neural network was able to remove noise from stellar spots and accurately calculate planet to star radius ratio. The Model attained a mean squared error (MSE) of 0.00053 on the training dataset and 0.001 on the test dataset. The MSE trend is smooth and always decreasing implying it is always able to improve itself as it completes more epochs.

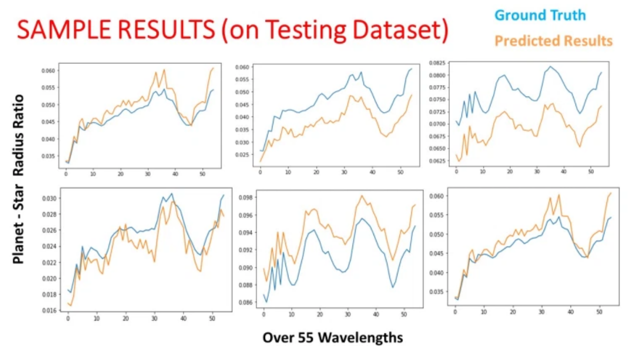
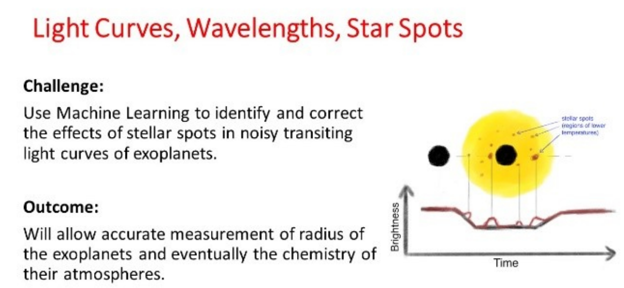


Figure 4: Planet-Star Radius Ratios for Sample of Exoplanets

I plotted the planet-star radius ratio predicted for 55 different wavelengths for a sample of six exoplanets. The graphs show how close the predicted values of planet-star radius ratios are from the ground truth for the 55 different wavelengths.



Light Curves, Wavelengths, Star Spots

Challenge:

Use Machine Learning to identify and correct the effects of stellar spots in noisy transiting light curves of exoplanets.

Outcome:

Will allow accurate measurement of radius of the exoplanets and eventually the chemistry of their atmospheres.

Figure 1: Effects of Stellar Spots on Light Curves. Credit: ARIEL Telescope website

Data Source

The dataset for the project was provided by the Atmospheric Remote-sensing Infrared Exoplanet Largesurvey (ARIEL) Telescope. ARIEL is the European Space Agency's (ESA) first mission dedicated to measuring the chemical composition and thermal structures of hundreds of transiting exoplanets.

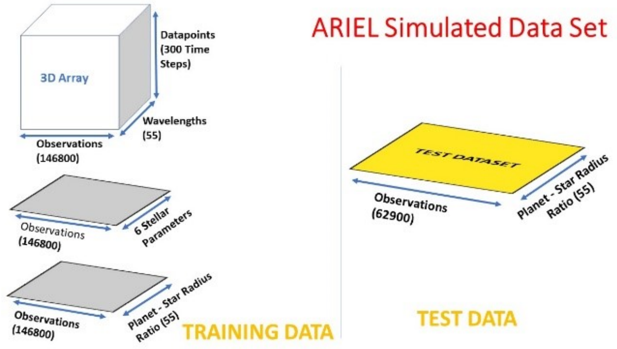


Figure 2: Simulated Data from the ARIEL Telescope

I used 150,000 simulated exoplanetary observations available on the ARIEL website. For each of the exoplanets, their transit light curves in 300 time-step data-points over 55 different wavelengths were provided. In addition, six stellar and planetary parameters were provided - mass, radius, temperature, log, period, and magnitude of the stars. Included in this database were the planet-star radius ratios of the exoplanets for each of the 55 wavelengths.

The dataset can be accessed at : https://ariel-datachallenge.azurewebsites.net/ML

Step 1: Transforming Exoplanetary Data Into a 3D Array

I imported the entire dataset into Python and converted it into a 3-dimensional array. ·

- The first dimension was the list of exoplanets for which the light curve was available.
- The second dimension was the 55 wavelengths for which the light curve was generated.
- The third dimension was the 300-time steps for which light from the star during the transit event was measured and used to create the light curve.

I split this array into training set and testing set in the ratio of 80:20.

Step 2: Setting Up the Recurrent Neural Networks

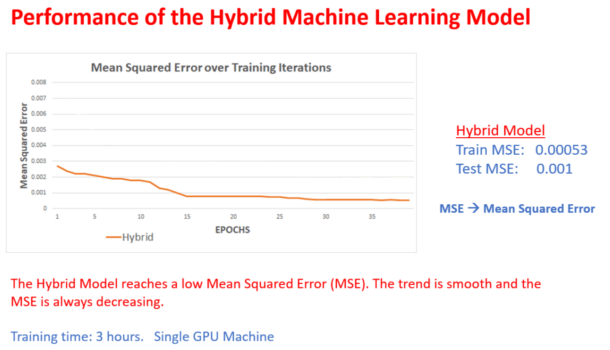
The inputs to my RNN were the 55 sequences of light curves for every exoplanet passing in front of its parent star taken in different wavelengths. This would result in 55 outputs: the predicted planet to star radius ratio for each wavelength.

During its training, the RNN would go over the entire training dataset several times. It would analyze every input and learn why a specific output is assigned to it. After enough training, it would be able to analyze inputs it had not seen before and recommend an output based on that with reasonable accuracy.

I ran my machine learning model for 40 epochs over a period of 3 hours.

(https://youtu.be/y2ZWrmPqF-E)

Movie: Overview of the Hybrid Machine Learning Model, https://youtu.be/y2ZWrmPqF-E



Performance of the Hybrid Machine Learning Model

Hybrid Model
Train MSE: 0.00053
Test MSE: 0.001

MSE → Mean Squared Error

The Hybrid Model reaches a low Mean Squared Error (MSE). The trend is smooth and the MSE is always decreasing.

Training time: 3 hours. Single GPU Machine

Figure 5: Mean Squared Error (MSE) achieved by the Hybrid Machine Learning Model

Conclusion

1. Machine learning was effective in reducing the noise from stellar spots and in predicting the planet to star radius ratio.
2. The model works on different data types: Sequential and Numerical.
3. The model works even works when some data is missing or becomes available later.
4. The model can be trained on a single GPU machine making it accessible to astronomy communities working in low computing power environments.

References

Nikolaou, Nikos & Waldmann, Ingo & Sarkar, Subhajit & Tsiaras, Angelos & Morvan, Mario & Yip, Kai & Tinetti, G. (2020). Correcting Transiting Exoplanet Light Curves for Stellar Spots: A Machine Learning Challenge for the ESA Ariel Space Mission.

Online Tutorial

To create a community of researchers around machine learning and space telescopes data, I created an online training module in Python using Jupyter Notebook. https://github.com/Artash-N/Ariel-Machine-Learning-Training-Module-for-Exoplanets