Capstone Proposal

Suzanne Taylor
June 1, 2018

Proposal

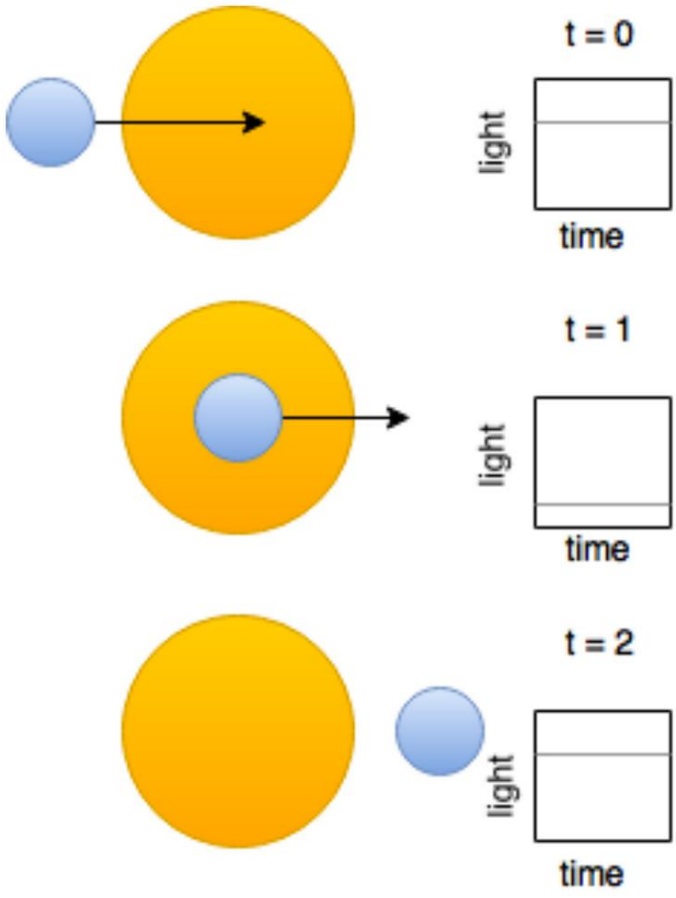
Domain Background

Kepler is a NASA built telescope launched in 2009 for the purpose of detecting planets orbiting other stars. Designed to survey a portion of the Milky Way to discover Earth-size exoplanets in or near habitable zones of those stars, Kepler's sole scientific instrument is a photometer that continually monitors the brightness of approx. 150,000 main sequence stars in a fixed field of view. This data is transmitted to Earth and the light curves resulting from the observations is used to determine if a planet has transited the star. As of 2 June 2018, there are 3,786 confirmed planets in 2,834 systems, with 629 systems having more than one planet.[1]

Due to the large amount of data relying on human judgment to produce a planet candidate is a time-consuming process and so this would be a good candidate for machine learning. As noted by Shallue et al. (2017)[2] there have been a number of projects to automate the process of detecting planets including Robovetter and Autovetter. In 2017 Shallue et al.[2] described a method of using a deep neural network to automatically vet Kepler threshold crossing events. This model has a good accuracy for distinguishing between transiting planets and other false positives.

Problem Statement

Planets do not emit light but when a planet transits a star the light intensity (flux) of the star dims. If a star is observed over several months or years a pattern of regular dimming may occur which could indicate a planet is orbiting the star.



[3]

Figure 1

As shown in figure 1 at time t0 before the planet (blue) starts it's transit the light intensity is high. At time t1 the planet is halfway across the star and the light intensity drops. At time t2 the planet has completed its transit and the light intensity has returned to normal.

Datasets and Inputs

The dataset was obtained from Exoplanet Hunting in Deep Space on Kaggle.com [3]. The data has been cleaned and is derived from observations made by the NASA Kepler space telescope. 99% of the data originates from campaign 3 and the dataset was prepared in late summer 2016

The data describes the change in light intensity (flux) of stars. Each start has been labeled 1 or 2 with 2 indicating the star is confirmed to have at least one exoplanet. The data is in 2 sets one for training and one for testing. The training set contains 5087 rows of observations (stars) with 37 confirmed exoplanet stars and 505 non-exoplanet stars. The test set contains 570 rows of observations with 5 confirmed exoplanet stars and 565 non-exoplanet stars. Each row has 3198

features. Column 1 is the label indicating whether the star has exoplanets or not. Column 2 - 3198 are the flux values of the star.

Solution Statement

I will develop a CNN using tensorflow/keras to analyse changes in a star's flux to determine if a planet has transited a star. The dataset use was obtained from Exoplanet Hunting in Deep Space on Kaggle.com [3] and has been pre-cleaned. Because the data is heavily imbalanced I will use SMOTE(Synthetic Minority Oversampling Technique) on the training set to to improve the representation of the exoplanet data.

Benchmark Model

My approach is based in the Mystery Planet (99.8% CNN) kernel by Peter Grenholm on Kaggle[4]. This approach achieved a 99.8% accuracy and my desired result will be to obtain accuracy close to that.

Evaluation Metrics

The metrics to evaluate the model on are recall and precision. Accuracy would not be a good metric as the dataset is highly imbalanced with none exoplanet stars.

TP : true positive FP: false positive

TN: true negative FN: false negative

Recall = TP / (TP + FN)

Precision = TP / (TP + FP)

		Predicted	
Actual		Non-exoplanet	Exoplanet
	Non-exoplanet	TN	FP
	Exoplanet	FN	TP

Confusion Matrix:

Recall-> Out of all the actual positive examples, how many did we predict to be positive?

Precision-> Out of the predicted positive examples, how many were actually positive?

Project Design

The data was derived from observations made by the NASA Kepler space telescope and has been pre-cleaned. I will train a CNN on the training set and then test the model on the test dataset. The test dataset will not be used during training.

Training set will have SMOTE applied to it to improve the representation of the confirmed exoplanet data.

I will use the Sequential API for Keras using the Adam optimization algorithm to build the CNN. It will have 10 convolution layers with layer sizes of 16, 32, 64, 128 and 256. 4 fully connected layers with layer sizes of 512 and a final output layer. Each layer will use the ReLU activation except the final layer which will use the sigmoid function. The training set will be run for 50 epochs with a batch size of 64.

Once the CNN has been trained I will test it using the test dataset to determine the recall and pression of the model.

References

1. Kepler (spacecraft) https://en.wikipedia.org/wiki/Kepler_(spacecraft)
2. Christopher J. Shallue & Andrew Vanderburg IDENTIFYING EXOPLANETS WITH DEEP LEARNING: A FIVE PLANET RESONANT CHAIN AROUND KEPLER-80 AND AN EIGHTH PLANET AROUND KEPLER-90 https://arxiv.org/pdf/1712.05044.pdf
3. https://www.kaggle.com/keplersmachines/kepler-labelled-time-series-data
4. https://www.kaggle.com/toregil/mystery-planet-99-8-cnn