An Interpretable Transformer Model for Operational Flare Forecasting

Vinay Ram Gazula,[1] **Yasser Abduallah,**[2] **Jason T. L. Wang**[2]
[1]Department of Data Science and [2]Department of Computer Science
New Jersey Institute of Technology, University Heights, Newark, NJ 07102, USA
ya54@njit.edu

Abstract

Interpretable machine learning tools including LIME (Local Interpretable Model-agnostic Explanations) and ALE (Accumulated Local Effects) are incorporated into a transformer-based deep learning model, named SolarFlareNet, to interpret the predictions made by the model. SolarFlareNet is implemented into an operational flare forecasting system to predict whether an active region on the surface of the Sun would produce a $\geq M$ class flare within the next 24 hours. LIME determines the ranking of the features used by SolarFlareNet. 2D ALE plots identify the interaction effects of two features on the predictions. Together, these tools help scientists better understand which features are crucial for SolarFlareNet to make its predictions. Experiments show that the tools can explain the internal workings of SolarFlareNet while maintaining its accuracy.

Introduction

Solar flares are intense bursts of energy that occur in an active region on the surface of the Sun. They are caused by strong magnetic fields and are the most important sources of space weather. Flares are classified into five classes which include *A*, *B*, *C*, *M* and *X* from the lowest intensity level, which is the *A* class, to the highest intensity level, which is the *X* class. Recently, Abduallah et al. (2023) used nine SHARP parameters/features to build an operational flare forecasting system, named SolarFlareNet, implemented with a transformer model. This system predicts whether there would be a γ-class flare within 24 to 72 hours, where γ is $\geq M5.0$, $\geq M$ or $\geq C$. Here, we extend the SolarFlareNet model by adding interpretable machine learning tools to understand the importance of features and the interaction effects of two features on the model predictions. We focus on the prediction of $\geq M$ flares that would occur within the next 24 hours.

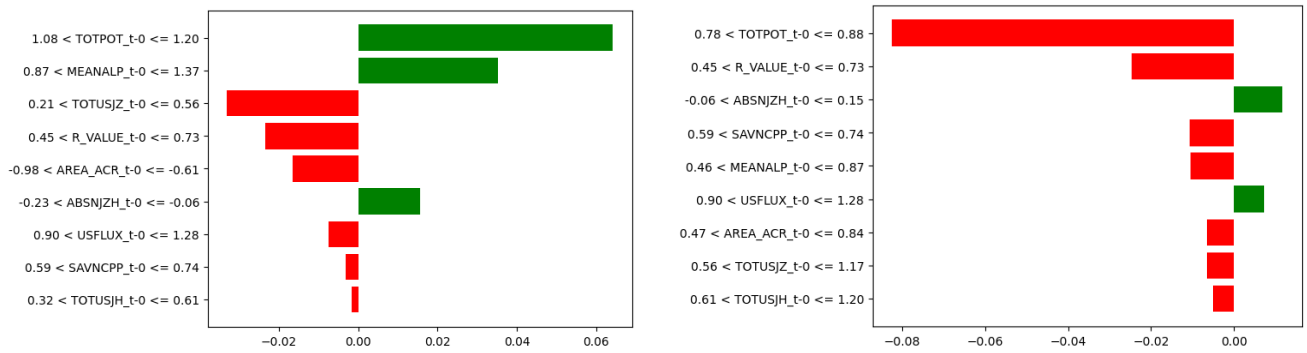
Methodology

In an effort to make the SolarFlareNet model interpretable and reliable, tools like Local Interpretable Model-agnostic Explanations (LIME) and Accumulated Local Effects (ALE) are integrated into the model. LIME (Ribeiro, Singh, and Guestrin, 2016) is a local surrogate method that is used to explain individual predictions of a black-box model. LIME generates a new dataset consisting of perturbed data samples and the corresponding predictions of the black-box model. In this new dataset, LIME trains an interpretable machine learning model which is weighted by the proximity of the new data samples. 2D ALE plots (Apley and Zhu, 2020) are a visualization approach used to understand the interactions of two features. A combination of LIME and ALE plots is used to determine the ranking of features and the correlations between the features.

Results

We used $199,641$ training data samples to retrain the SolarFlareNet model after the LIME and ALE tools are incorporated into the model. We then used $60,000$ test data samples to test the model. The accuracy of the newly trained model is 0.907, which is close to the accuracy of the original model described in Abduallah et al. (2023), indicating that there is no significant loss in the accuracy of the model.

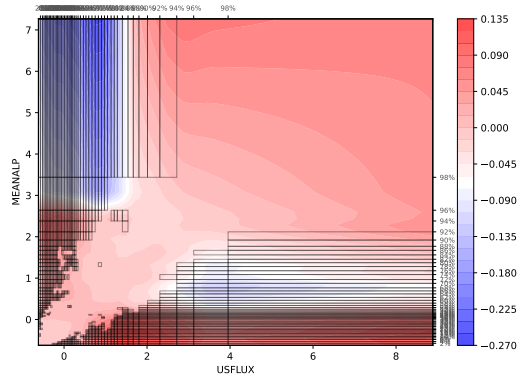
Figure 1 presents bar graphs showing the LIME results for two test data samples. Figure 1(a) shows the result of a positive prediction, and Figure 1(b) shows the result of a negative prediction. A positive prediction refers to a test data sample where SolarFlareNet predicts it to be positive, i.e. there will be a flare within 24 hours. A negative prediction refers to a test data sample where SolarFlareNet predicts it to be negative, i.e. there will be no flare within 24 hours. The X-axis in Figure 1 denotes the prediction values produced by SolarFlareNet. A feature with a positive prediction value means there will be a flare within the next 24 hours according to this feature. A feature with a negative prediction value means there will be no flare within the next 24 hours according to this feature. The higher a feature is ranked, the more influence it has (and the more important it is) on a prediction. Using bar graphs, we can extract individual prediction values for all features. The total prediction value for a given test data sample is obtained by adding the individual prediction values for all features in the test data sample. We see that the total prediction value in Figure 1(a) is positive (and hence Figure 1(a) represents a positive prediction), while the total prediction value in Figure 1(b) is negative (and hence Figure 1(b) represents a negative prediction). The Y-axis of a bar graph denotes the range of values of a feature.



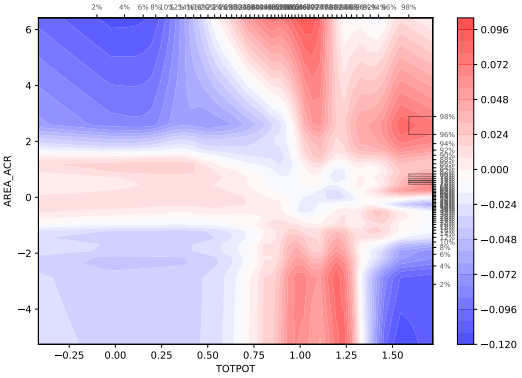
(a) (b)

Figure 1: LIME bar graphs for (a) a positive prediction and (b) a negative prediction, respectively.

Figure 2(a) exhibits the 2D ALE plot for two highly correlated features, USFLUX and MEANALP, with a Pearson's correlation coefficient of 0.911. Figure 2(b) shows the 2D ALE plot for two barely correlated features, TOTPOT and AREA_ACR, with a Pearson's correlation coefficient of 0.001. The X axis and the Y axis in Figure 2 represent the values of the features, respectively. The color bar represents the interaction effect of the features on the average prediction value produced by the SolarFlareNet model. The darker the red color (blue color), the larger the positive interaction effect (negative interaction effect), leading to positive predictions (negative predictions, respectively). In Figure 2(a), as the values of the two features, USFLUX and MEANALP, increase, there is a positive interaction effect on the average prediction value, shown as dark red patches in the upper right corner of the plot. In Figure 2(b), many places in the plot are close to zero, where the two features, TOTPOT and AREA_ACR, do not interact with each other. We see that the two features USFLUX and MEANALP have much stronger interactions with much darker colors than the two features TOTPOT and AREA_ACR, which is consistent with the fact that USFLUX and MEANALP have a much larger correlation coefficient than TOTPOT and AREA_ACR.



(a) USFLUX vs. MEANALP



(b) TOTPOT vs. AREA_ACR

Figure 2: 2D ALE plots for (a) two highly correlated features USFLUX and MEANALP, and (b) two barely correlated features TOTPOT and AREA_ACR, on all test data samples. The percentages shown at the top and right of a plot represent the percentages of test samples considered so far in the plot.

Conclusion

We incorporate interpretable machine learning tools LIME and ALE into an operational flare forecasting system (SolarFlareNet) to predict whether a $\geq$M class flare would occur within the next 24 hours. We demonstrate that the tools work well in adding interpretability to the forecasting system.

References

Abduallah, Y.; Wang, J. T. L.; Wang, H.; and Xu, Y. 2023. Operational prediction of solar flares using a transformer-based framework. *Scientific Reports* 13(1):13665.

Apley, D. W., and Zhu, J. 2020. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 82(4):1059–1086.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why should I trust you?" Explaining the predictions of any classifier. In *Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 1135–1144.