Insights into Deep Learning-based Full-disk Solar Flare Prediction with Post hoc Explanation and Evaluation

Chetraj Pandey[1] Temitope Adeyeha [1] Trisha Nandakumar [1] Rafal Angryk [1] Berkay Aydin [1]

[1]Georgia State University, Atlanta, GA, USA

Background

Solar flares are intense eruptions of energy from the Sun that release massive amounts of electromagnetic radiation and pose a significant risk to space- and ground-based infrastructures. Hence, a precise and reliable prediction is essential for mitigating potential impacts.

As part of our study, we present a deep-learning model that utilizes hourly full-disk line-of-sight magnetogram images that can predict $\geq$M1.0-class flares within the subsequent 24-hour window. We apply the Guided Grad-CAM [2] attribution method to generate post hoc explanations for our model's prediction and provide empirical findings from qualitative and quantitative evaluations of these explanations.

Motivation

Most of the current operational forecasting systems rely on active regions (ARs). Magnetic field measurements, which form the basis of AR-based forecasting techniques, suffer from projection effects when ARs are close to the limbs (beyond $\pm 70^{\circ}$ of the central meridian of the solar disk). As a result, the data is limited to ARs located in central positions (within $\pm 70^{\circ}$). Therefore, it is essential to address the near-limb events as well and extend the coverage area to the entire full-disk, and issue global prediction.

Recent advances in deep learning have accelerated the development of data-driven models for solar flare prediction; however, the complex learned data representations of these models hinder transparency and pose challenges in comprehending their prediction rationale, which can be particularly problematic in critical applications like solar flare prediction where model reliability is crucial. Post hoc explanations offer valuable insights into a model's decision-making process, improving transparency and reliability.

Method: Data, Model, & Explanation

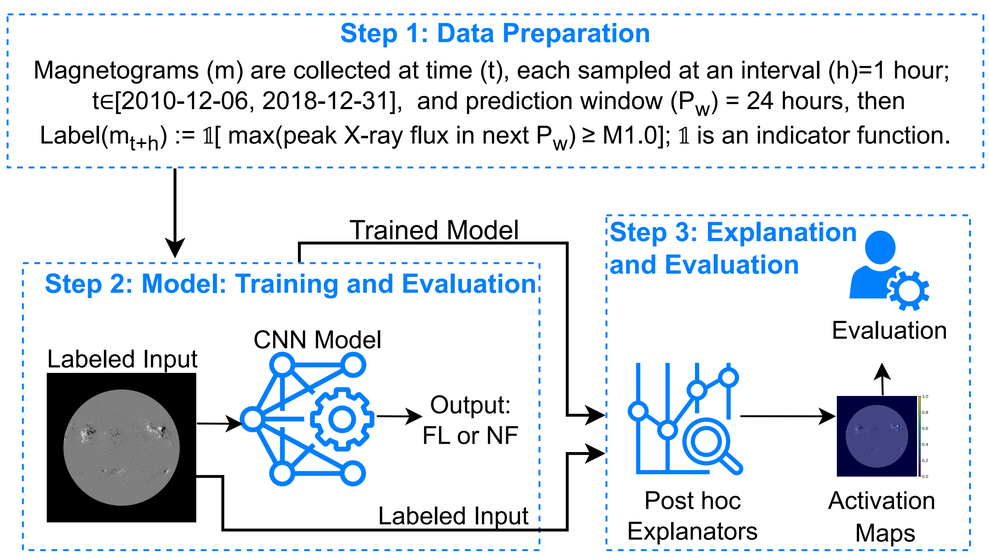

Figure 1. An overview of our methodology for full-disk solar flare prediction with model explanation in three broad steps. Note: The outcome of the indicator function (𝟙), is 'Flare' when true, and 'No Flare' otherwise.

Data

- We collected full-disk line-of-sight magnetograms obtained from the Helioseismic and Magnetic Imager (HMI) onboard Solar Dynamics Observatory (SDO) available as compressed JPEG 2000 (JP2) images in near real-time publicly via Helioviewer.
- We sampled magnetogram images every hour of the day, starting at 00:00 and ending at 23:00, from December 2010 to December 2018 and labeled them using a 24-hour prediction window (P_w). To elaborate, if the maximum peak x-ray flux of flare, converted to NOAA/GOES class, in subsequent P_w was weaker than M, we labeled it as "No Flare" (NF: $<$M1.0), and if it was $\geq$M1.0, we labeled it as "Flare" (FL: $\geq$M1.0) as shown in Fig. 1.

Model

- We extend the AlexNet [1] model to accommodate 1-channel input magnetogram images by adding an additional convolutional layer at the beginning of the network that uses a 3×3 kernel, size-1 stride and outputs a 3-channel image as shown in Fig 2.

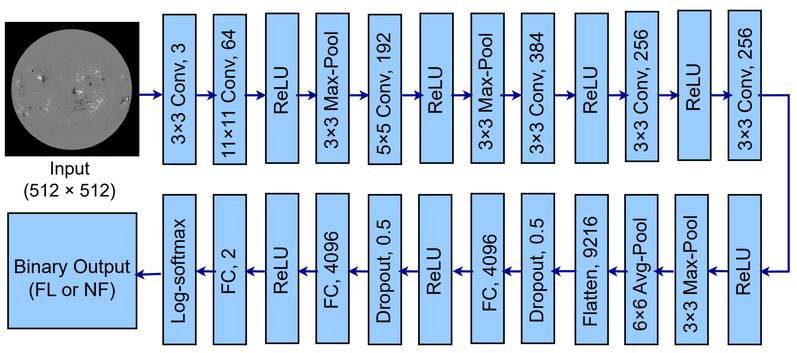

Figure 2. The architecture of our flare prediction model.

Explanation Method

- We utilized Guided Grad-CAM, a fusion of Grad-CAM [2] and guided backpropagation [3] techniques, both of which are based on class-specific gradient, i.e., activations are generated for a given target class (y). It combines the detailed precision of guided backpropagation with the localization advantages of Grad-CAM.
- The size of attributions obtained from Grad-CAM depends on the dimensions of the feature maps of the last convolutional layer and hence is smaller than the size of the input (x). Therefore, in Guided Grad-CAM, the activation maps of Grad-CAM attributions are upscaled to the size of the input and computed as shown in Eq. (1).

$$\textit{Guided Grad-CAM}(x,y) = \textit{Upscaled}(\textit{GradCAM}(x,y)) \odot \textit{GuidedBackProp}(x,y) \quad (1)$$

Model Evaluation

- Model's overall performance was evaluated using true skill statistic (TSS) and Heidke skill score (HSS). The aggregated results from our 4-fold cross-validation experiments have on average **TSS**$\sim$**0.51**$\pm$**0.05** and **HSS**$\sim$**0.38**$\pm$**0.08**.
- In addition, we compute recall scores for correctly predicted and missed flare counts for class-specific flares (X-class and M-class) in central locations (within $\pm 70^{\circ}$) and near-limb locations (beyond $\pm 70^{\circ}$) of the Sun. We observed $\sim$95% of X-class flares and $\sim$73% of M-class flares in central locations were correctly predicted. Similarly, $\sim$74% of X-class flares and $\sim$50% of M-class flares in near-limb regions were predicted correctly. The heatmap of recall scores for $\geq$M1.0-class flares is shown in Fig 3.

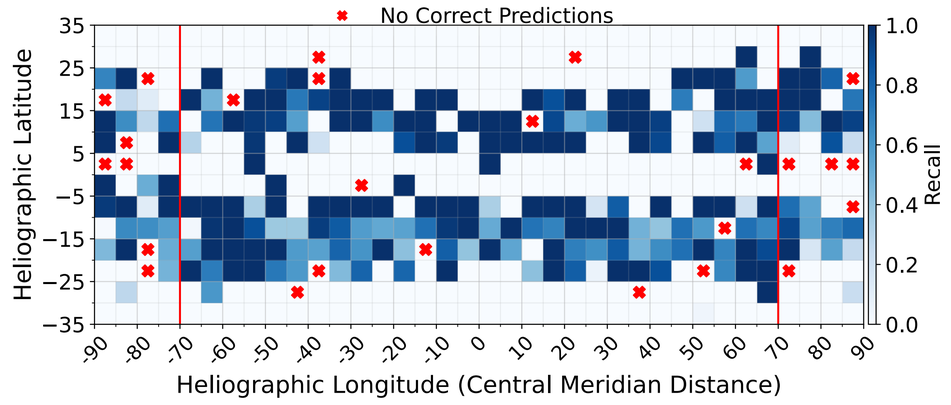

Figure 3. A heatmap showcasing recall for $\geq$M-class flares binned into $5^{\circ} \times 5^{\circ}$ flare locations used as the label. Beyond vertical red lines are near-limb events. Red crosses in white grids indicate zero correct predictions, while white cells without red crosses represent unavailable instances.

Explanation and Evaluation

Through qualitative inspection of generated visual explanations using Guided Grad-CAM, we observed that our models grasp the shape and texture-based properties of flaring active regions, even in proximity to limb areas. As an example shown in Fig. 4, it was observed that as soon as a region becomes visible, the pixels covering the AR on the East limb are activated, in fact, the model focuses on specific ARs including the relatively smaller AR on the East limb, even though other, ARs are present in the magnetogram image.

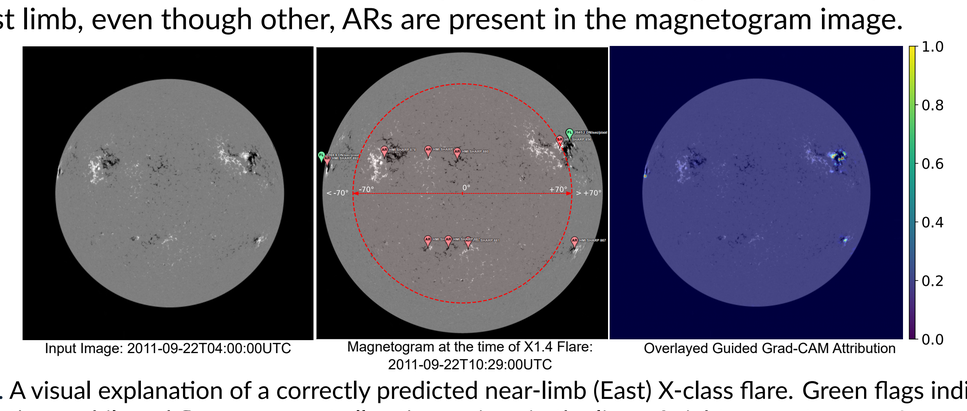

Figure 4. A visual explanation of a correctly predicted near-limb (East) X-class flare. Green flags indicate the flare locations, while red flags represent all active regions in the line-of-sight magnetogram. Approximate near-limb (beyond$\pm 70^{\circ}$) and central (within $\pm 70^{\circ}$) regions are delineated with a red-dotted circle.

Quantitative Evaluation of Explanations

We rigorously evaluated explanations for X-class flares using a human-centered approach mentioned in [4], employing a set of close-ended questionnaires to validate 880 instances in our dataset. The questionnaires were structured as follows:

1. Determine the activation{T, F} of the responsible X-class flare location with the maximum peak X-ray flux used for labeling the image.
2. Determine the activation {T, F, n/a} of other (i) X-class, (ii) M-class, and (iii) C-class flare locations coexisting within the 24-hour prediction window, besides the flare's location used for labeling.

Here, "T", "F", and "n/a" represents True (activation), False (non-activation), and absence of relevant flares respectively.

- For $\sim$99% of total instances, the relevant locations of the flares were activated.
- In $\sim$53% of the total instances, the locations of co-existing M- & C-flares were activated.
- In $\sim$39% of the total instances, the AR-locations of co-existing C-flares were activated.

Conclusion and Future Work

In this study, we show that a full-disk model can skillfully tackle near-limb flare events and evaluated our model's predictions with visual explanations, showing that the decisions are primarily capturing characteristics corresponding to the active regions in the magnetogram instances. We intend to develop an automated method to evaluate the explanations.

Acknowledgements

This project has been supported in part by funding from two NSF awards #2104004 and #1931555, and a NASA grant award #80NSSC22K0272. The data used in this project is courtesy of NASA/SDO and the HMI science teams.

References

[1] Alex Krizhevsky.
One weird trick for parallelizing convolutional neural networks, 2014.

[2] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra.
Grad-CAM: Visual explanations from deep networks via gradient-based localization.
In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, October 2017.

[3] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller.
Striving for simplicity: The all convolutional net, 2014.

[4] Giulia Vilone and Luca Longo.
Notions of explainability and evaluation approaches for explainable artificial intelligence.
Inf. Fusion, 76(C):89–106, December 2021.