Towards Reliable Deep Learning Models for Solar Flare Prediction

Chetraj Pandey, Rafal A. Angryk, Berkay Aydin

Georgia State University, Atlanta, GA, USA

Background & Motivation

Solar Flares: One of the key space weather phenomena characterized by sudden and intense emissions of radiation from the Sun; pose significant risks to space- and ground-based infrastructures.

Our flare forecasting models leverage deep learning for higher predictive performance.

Reliability and Transparency: Post hoc explanation methods (attribution methods) can be used to assess the reliability of predictions for operational settings.

Research Questions: **Can we trust the explanations in terms of their consistency? How similar are the explanations generated from different post hoc explanation methods?**

Data & Model

Data: The dataset is preprocessed from the HMI SHARP series [1] as mentioned in [2], which utilize line-of-sight (LoS) magnetograms and corresponding bitmaps to filter the high activity regions and generate grayscale images. Fig. 1 shows an example of preprocessed dataset instance.

(a) (b) (c)

Fig. 1. (a) Raw HMI magnetogram (Size: 688 × 448 px) of AR patch HARP 7115 (NOAA AR 12673) from 2017-09-06, 06:00 UTC, (b) Bitmap highlighting the high-activity region (white pixels with size: 520 × 440 px), (c) Processed 512×512 image used for model training.

- We sample the hourly instances of LoS magnetograms, covering solar-cycle 24, and labeled them using a 24-hour prediction window to predict ≥ M-class flares.

Model: We use transfer learning and extend the pre-trained MobileNet [3] model.

- To accommodate 1-channel input magnetograms, we add an additional convolutional layer at the beginning of the network that uses a 3×3 kernel, as mentioned in [2].

Methodology

Explanation Method: There are two main types of explanation methods: perturbation-based and gradient- based. The former method is computationally inefficient and can lead to inconsistent explanations due to creation of Out-of-Distribution data.

- We used four gradient-based methods to generate local explanations: (i) **Guided Grad-CAM (GGCAM)** [4], (ii) **Integrated Gradients (IG)** [5], (iii) **DeepLIFT SHAP (SHAP)** [6], and (iv) **Guided Backpropagation (GBP)** [7].

Explanation Evaluation Metrics: Our objective is to assess the consistency among the local explanations provided by these four attribution methods. We introduce three indices for evaluating the consistency across explanations: (i) root mean squared error (RMSE), (ii) cosine similarity (cosSim), and (iii) intersection over union (IOU).

- **RMSE** measures the average discrepancy between explanations; lower values indicate closer agreement.
- **cosSim** assesses directional similarity between explanations represented as vectors, independent of magnitude.
- **IOU** evaluates the overlap between explanations, to assess spatial alignment.

Experimental Evaluation

We evaluate the consistency among local explanations for magnetograms (test partitions in [2]) within ±45 degrees of flux-weighted longitude (LONFWT) . An example of explanation generated for input in Fig. 1 is shown in Fig. 2 using all four methods.

Upon evaluating the similarity among the explanations using different metrics, we observed the following as shown in Fig. 3:

- **IOU:** Suggests a strong spatial alignment between explanations from IG, SHAP, and GBP, compared to GGCAM.
- **RMSE:** Indicates near-perfect consistency with minimal discrepancy across all explanations.
- **CosSim:** Reveals significant directional disagreement between GBP and GGCAM explanations.

These metrics offer distinct insights into the similarity and alignment of the explanations.

- The inconsistencies observed in the GGCAM explanations may stem from the lower spatial resolution inherent in Grad-CAM [4].

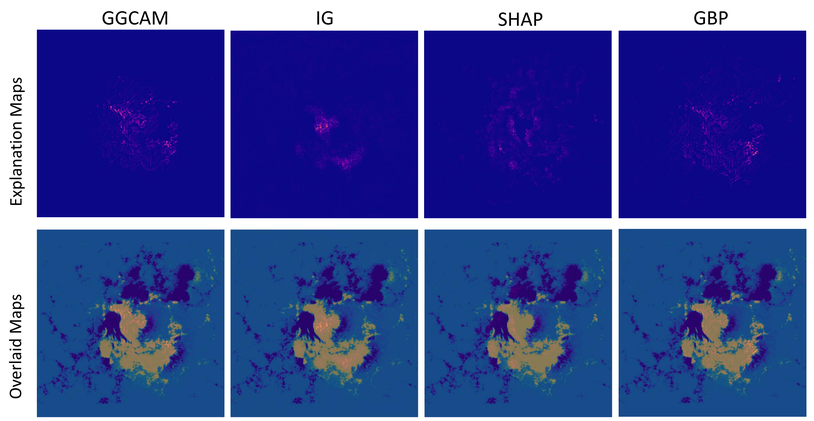

Fig. 2. First Row: Explanation maps generated from aforementioned four attribution methods for input image in Fig 1. (c) which corresponds to an X-class flare. Second Row: Explanations overlaid on input.

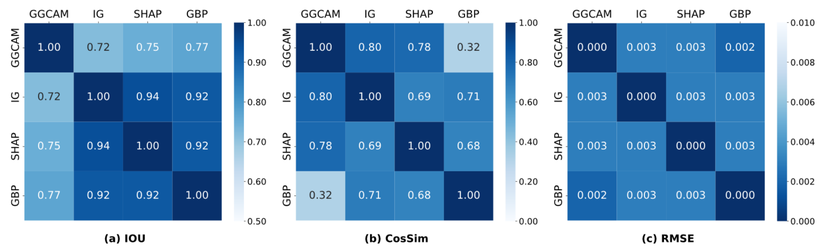

Fig. 3. Heatmap showing the average pairwise similarity between the local explanations generated from four attribution methods using three metrics: (a) IOU, (b) CosSim, and (c) RMSE.

Discussion

Non-Aligned Explanations: Different post-hoc methods (e.g., IG, SHAP, GBP, GGCAM) generate explanations that often do not align, creating uncertainty about which explanation to trust for critical tasks like solar flare prediction.

Human Evaluation Limitations: While human evaluation is accurate, it's impractical for large datasets, and inconsistencies between explanations remain a challenge for decision-making

Ensemble of Explanations: To resolve this, we propose using an ensemble of explanations generated by multiple methods, combining them through the Hadamard product to highlight common regions and improve decision-making.

For improved reliability, we integrate the explanations, and the ensemble explanation map $E_{ensemble}$ can be computed as:

$$E_{ensemble} = E_1 \odot E_2 \odot \ldots\ldots \odot E_N$$

where E_1, E_2, …., E_N are the explanation maps from different methods.

- For noise removal, we are also exploring image processing techniques (e.g., blurring).

Conclusion & Future Work

Due to inconsistencies among explanations from different methods, selecting a single reliable explanation is challenging. The ensemble approach provides a way to address this by focusing on common features across methods.

Future efforts could refine the ensemble method, exploring adaptive weighting or noise reduction techniques to improve explanation consistency and decision-making.

Acknowledgements

This work is supported by the National Science Foundation under Grant #2104004. The data used in this study is a courtesy of NASA/SDO and the AIA, EVE, and HMI science teams, and the NOAA (NGDC).

Contact Information

Chetraj Pandey
Dept. of Computer Science,
Georgia State University,
Email:cpandey1@gsu.edu

References

1. Bobra, M.G., Sun, X., Hoeksema, J.T. et al. The Helioseismic and Magnetic Imager (HMI) Vector Magnetic Field Pipeline: SHARPs – Space-Weather HMI Active Region Patches. Sol Phys 289, 3549–3578 (2014).
2. Pandey, C., Adeyeha, T., Hong, J., Angryk, R.A., Aydin, B. (2024). Advancing Solar Flare Prediction Using Deep Learning with Active Region Patches. In: ECML PKDD 2024, vol 14950. Springer, Cham.
3. Howard, A., Sandler, M., Chu, G., Chen, L., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q.V., Adam, H.: Searching for mobilenetv3.
4. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: GradCAM: Visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision.
5. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic Attribution for Deep Networks. arXiv (2017).
6. Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17, pp. 4768–4777. Curran Associates Inc., Red Hook, NY, USA (2017).
7. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for Simplicity: The All Convolutional Net. CoRR, vol. abs/1412.6806, (2014).

AGU24
Washington, D.C. | 9–13 December 2024