The Astrophysical Journal, 910:8 (11pp), 2021 March 20 https://doi.org/10.3847/1538-4357/abdebe

Visual Explanation of a Deep Learning Solar Flare Forecast Model and Its Relationship to Physical Parameters

Kangwoo Yi[1], Yong-Jae Moon[1,2], Daye Lim[2], Eunsu Park[2], and Harim Lee[2]

[1] School of Space Research, Kyung Hee University, 1732, Deogyeongdae-ro, Giheung-gu, Yongin-si Gyunggi-do, 17104, Republic of Korea; moonyj@khu.ac.kr

[2] Department of Astronomy and Space Science, College of Applied Science, Kyung Hee University, 1732, Deogyeongdae-ro, Giheung-gu, Yongin-si Gyunggi-do, 17104, Republic of Korea

Received 2020 May 27; revised 2021 January 18; accepted 2021 January 19; published 2021 March 22

Abstract

In this study, we present a visual explanation of a deep learning solar flare forecast model and its relationship to physical parameters of solar active regions (ARs). For this, we use full-disk magnetograms at 00:00 UT from the Solar and Heliospheric Observatory/Michelson Doppler Imager and the Solar Dynamics Observatory/Helioseismic and Magnetic Imager, physical parameters from the Space-weather HMI Active Region Patch (SHARP), and Geostationary Operational Environmental Satellite X-ray flare data. Our deep learning flare forecast model based on the Convolutional Neural Network (CNN) predicts "Yes" or "No" for the daily occurrence of C-, M-, and X-class flares. We interpret the model using two CNN attribution methods (guided backpropagation and Gradient-weighted Class Activation Mapping [Grad-CAM]) that provide quantitative information on explaining the model. We find that our deep learning flare forecasting model is intimately related to AR physical properties that have also been distinguished in previous studies as holding significant predictive ability. Major results of this study are as follows. First, we successfully apply our deep learning models to the forecast of daily solar flare occurrence with TSS = 0.65, without any preprocessing to extract features from data. Second, using the attribution methods, we find that the polarity inversion line is an important feature for the deep learning flare forecasting model. Third, the ARs with high Grad-CAM values produce more flares than those with low Grad-CAM values. Fourth, nine SHARP parameters such as total unsigned vertical current, total unsigned current helicity, total unsigned flux, and total photospheric magnetic free energy density are well correlated with Grad-CAM values.

Unified Astronomy Thesaurus concepts: The Sun (1693); Solar flares (1496); Convolutional neural networks (1938)

1. Introduction

Solar flares, one of the most energetic activities of the Sun, are known to occur in active regions (ARs) which are magnetically concentrated locations (Priest & Forbes 2002; Shibata & Magara 2011). Many researchers have suggested that flares are related to magnetic characteristics such as non-potentiality and magnetic complexity (Schrijver 2007, 2016; Sharykin et al. 2017; Toriumi & Takasao 2017; Vasantharaju et al. 2018). Characteristics of the polarity inversion line (PIL) have been studied for flare prediction (Mason & Hoeksema 2010; Falconer et al. 2011, 2012, 2014; Sadykov & Kosovichev 2017). Many studies have considered Space-weather HMI Active Region Patch (SHARP; Bobra et al. 2014) data which indicate magnetic characteristics of ARs (Bobra & Couvidat 2015; Liu et al. 2017; Sadykov & Kosovichev 2017; Lim et al. 2019a, 2019b). Recently, several studies have applied deep learning methods to the flare forecast using magnetic features of the Sun (Huang et al. 2018; Nishizuka et al. 2018; Park et al. 2018; Chen et al. 2019; Liu et al. 2019; Li et al. 2020). However, the question of how deep learning models predict flare occurrence is still not answered clearly. Liu et al. (2019) used magnetic parameters for their flare prediction model which is based on the Long Short-Term Memory (LSTM; Hochreiter & Schmidhuber 1997). They investigated the relationship between input parameters and the performance of the model by searching for a subset of inputs that make prediction scores high or low. This study is a good answer to the question of what inputs make the forecasting score high but not a proper quantitative information of the deep learning flare model. Huang et al. (2018) applied the Convolution Neural Network (CNN; Lecun et al. 1998) to the flare forecast using patches of ARs of solar line-of-sight magnetograms. They extracted CNN feature maps from the interior layers in the model and presented that their model pays attention to the area of the PIL. However, the feature map, which is just a result of the calculation between the input image and CNN kernels, does not indicate important areas of the input image for prediction results. The highlighted area in the feature map could be fade-out in subsequent layers due to inactive kernels, low weights, and/or the merging of spatial information. In addition, feature extraction cannot guarantee which ones out of a number of features in the layer are important.

Deep learning interpretation is an important issue in computer science and related fields. For CNN, attribution methods such as guided backpropagation (Springenberg et al. 2015) and Gradient-weighted Class Activation Mapping (Grad-CAM; Selvaraju et al. 2017) have been mainly used for visual explanations. Guided backpropagation and Grad-CAM are based on the gradient for each weight (input pixel, CNN kernel, and fully connected layer weight), which indicates the change in the result if the weight is increased by a tiny amount (Simonyan et al. 2014; Lecun et al. 2015).

In this letter, for the first time, we present a visual explanation of a deep learning flare forecast model by attribution methods and its relationship to physical parameters such as SHARP. For this, we develop a new deep learning flare forecasting model, which is based on CNN, that predicts "Yes or No" for the daily occurrence of C-, M-, and X-class flares. Input data are the solar full-disk line-of-sight magnetogram at 00:00 UT from the Solar and Heliospheric Observatory

(SOHO; Domingo et al. 1995)/Michelson Doppler Imager (MDI; Scherrer et al. 1995) and Solar Dynamics Observatory (SDO; Pesnell et al. 2012)/Helioseismic and Magnetic Imager (HMI; Schou et al. 2012). Using Guided backpropagation and Grad-CAM, we measure the importance weight of the area in the input and compare them with physical parameters taken from SHARP.

This paper is organized as follows. The data are described in Section 2. Our model and attribution methods used for this research are described in Section 3. The results of model interpretation and comparison with physical parameters are given in Section 4. A brief conclusion and summary are presented in Section 5.

2. Data

We consider solar full-disk line-of-sight magnetograms at 00:00 UT from SOHO/MDI and SDO/HMI for model training and test. When a magnetogram is not available at 00:00 UT, we use the nearest one taken on the previous day. Most of the data are in the range of two hours before 00:00 UTC. SOHO, launched in 1995 December, is a space mission to study the Sun, and a joint venture between the European Space Agency and the National Aeronautics and Space Administration (NASA). MDI, which is one of the scientific instruments of SOHO, provides 1024 × 1024 full-disk solar magnetograms with 1.″98 per pixel at a cadence of 96 minutes. SDO, launched in 2010 February, is a NASA space mission to provide data for predicting solar activities. HMI, which is one of the scientific instruments of SDO, supplants MDI. HMI produces 4096 × 4096 full-disk magnetograms with 0.″5 per pixel at a 720 s cadence.

We obtained binned magnetogram image data from the SOHO data archive[3] in JPG format with 512 × 512 resolution for training and test. MDI full-disk magnetograms in the archive were used for ASAP flare forecast (Colak & Qahwaji 2009). HMI full-disk magnetograms in the archive were produced by quick-look methods which are useful for real-time forecasts.

In order to verify the scientific reliability of JPG data, we calculate the pixel correlation coefficient between it and actual magnetogram data using a randomly selected 10% of MDI and HMI data that are resized into 256 × 256 sizes by block average. The average pixel correlation coefficient of MDI is 0.91 and that of HMI is 0.87 with the actual magnetogram data with the bytescale of ±100 Gauss, which shows the JPG data are in good agreement with the actual magnetogram data.

We evaluate transferability between MDI and HMI data using the data within 30 minutes of time difference during 2011 January in two ways. First, we calculate Complex Wavelet Structural Similarity (CW-SSIM; Sampat et al. 2009) between MDI and HMI, which is the measure index of image similarity. CW-SSIM is an extension of the Structural Similarity (SSIM; Wang et al. 2004), which is generally used to measure the similarity between two images to the complex wavelet domain. It is less sensitive to small geometric distortions such as small rotations, translations, and small differences in scale than SSIM.

$$\text{CW-SSIM}(c_x,\, c_y) = \frac{2\left|\sum_{i=1}^{N} c_{x,i} c_{y,i}^{*}\right| + L}{\sum_{i=1}^{N} |c_{x,i}|^2 + \sum_{i=1}^{N} |c_{y,i}|^2 + L} \quad (1)$$

where c_x and c_y represent the sets of coefficients (extracted at the same spatial location in the same wavelet subbands) in the complex wavelet transform domain (e.g., the complex version of the steer pyramid decomposition Portilla & Simoncelli 2000) of the two images being compared, respectively. The * denotes the complex conjugate of c and L is a small positive constant. CW-SSIM is in the range 0–1 and results in a higher value for higher similarity. The small geometric distortion is the main difference between MDI and HMI data, but its effect on solar activity is not as significant compared to large distortion. For this, we use re-shaped data by cubic-interpolation for a fair evaluation considering a solar disk radius in pixels. The average CW-SSIM is 0.82, indicating that MDI and HMI data are structurally transferable. Second, we calculate the pixel correlation coefficient between MDI and HMI data in units of Gauss. For this we use rebinned MDI and HMI data into 256 sizes by block average. The average correlation coefficient is 0.9, which is high enough to transfer between MDI and HMI. This result is slightly higher than Liu et al. (2012; 0.82), which may be due to denoising by block average, and different comparison areas (center-to-limb angle of 0°–60° for Liu et al. 2012 and full-disk for ours).

For physical parameters we adopt definitive SHARP data with a 720 s cadence, developed by the HMI team (Bobra et al. 2014). These data are given for HMI Active Region Patches (HARPs) which are automatically identified. For each HARP, magnetic parameters that are derived from the size, distribution, and non-potentiality of vector magnetic fields are provided.

We label the magnetograms separately with the flaring event day (⩾C1.0 Class) or non-flare event day (<C1.0 Class) using Geostationary Operational Environmental Satellite (GOES) X-ray flare data which are automatically detected by algorithm or manually recorded by the National Oceanic and Atmospheric Administration (NOAA; Ryan et al. 2016; NOAA GOES X-ray flux[4]).

To select training and test data, we use the chronological data separation in which we sequentially separate the data into training and test according to its observation time. By using the chronological data separation, we avoid an impossible forecasting condition for the model evaluation, i.e., carrying out current forecasting using future data, not past data.

As a result, 4298 MDI data from 1996 May to 2008 December are used for training and 683 MDI data from 2009 January to 2010 December and 2360 HMI data from 2011 January to 2017 June are used for testing.

3. Model and Methods

3.1. Flare Model

Our model follows the technique of the dense connection (Huang et al. 2017). The number of layers and optimized hyperparameters in our model are different from theirs. The data that we use are solar magnetograms so that we train our model independently. Figure 1 shows a structure of our deep learning flare model. The model consists of an initial block, five dense blocks, and a last block, in order. The initial block contains a convolution layer (3 × 3 kernel, 1 stride, 26 features) and a max pooling layer (2 × 2 kernel, 2 strides), in order. Dense blocks 1–5 consists of a batch normalization (BN; Ioffe & Szegedy 2015), a

[3] https://sohowww.nascom.nasa.gov/data/REPROCESSING/

[4] https://www.swpc.noaa.gov/products/goes-x-ray-flux

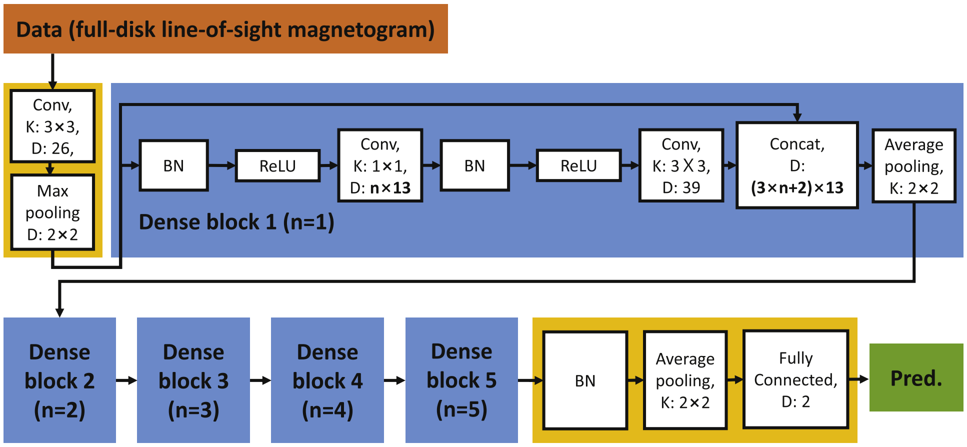


Figure 1. Structure of our model. *K* is kernel size. *D* is the number of feature dimensions and is displayed in the case where its number is changed in the layer.

rectified linear unit (ReLU; Nair & Hinton 2010), a convolution layer (1 × 1 kernel, 1 stride, 13 × n features; $n =$ number of the block), a BN, a ReLU, a convolution layer (3 × 3 kernel, 1 stride, 39 features), a concatenation layer that concatenates features from the previous layer and the input of the dense block, and an average pooling layer (2 × 2 kernel, 2 strides), in order. The last block includes a BN, an average pooling layer (2 × 2 kernel, 2 strides), and a fully connected layer, in order. The last layer produces two values; one is a score for the flaring event day and the other for the non-flare event day. The prediction of the model selects a higher score.

We compare our model with previous flare models. To evaluate models, we consider statistical scores using the contingency table that consists of four components: hit (H; flare predicted and occurred), false alarm (F; flare predicted but did not occur), miss (M; no flare predicted but occurred), and null (N; no flare predicted and none occurred). The statistical scores, Accuracy (ACC), the Heidke Skill Score (HSS; Heidke 1926), the Appleman Skill Score (ApSS; Appleman 1960), and the True Skill Statistics (TSS; Allouche et al. 2006), are described as

$$\mathrm{ACC} = \frac{H+N}{H+F+M+N}, \tag{2}$$

$$\mathrm{HSS} = \frac{2[(H\times N)-(M\times F)]}{(H+M)\times(M+N)+(H+F)\times(F+N)}, \tag{3}$$

$$\mathrm{ApSS} = \left\{\begin{array}{l} \frac{H-F}{H+M},\ \text{if event rate} < 0.5 \\ \frac{N-M}{F+N},\ \text{if event rate} \geqslant 0.5 \end{array}\right\}, \tag{4}$$

$$\mathrm{TSS} = \frac{H}{H+M} - \frac{F}{F+N} \tag{5}$$

where the event rate is the rate between the number of flare events and the total number of data, in the test set. Barnes et al. (2016) suggested that ApSS is a good index for method evaluation because it treats the cost of each type of error (miss and false alarm) as equal. TSS has to be understood with event rate and frequency bias (FB; the number of events in forecasting divided by the number of events on observation, Leka et al. 2019). In detail, with an event rate lower than 0.5, overforecasting (FB > 1) attains a high TSS while underforecasting (FB < 1) is less likely to. With an event rate higher than 0.5, that is the opposite. The performance of the forecasting model could be changed with the difference in data splitting (Nishizuka et al. 2017). We compare our model with the other models (deep learning models and one statistical model) based on the chronological data separation method.

Cinto et al. (2020) separated flare prediction models into two groups, operationally evaluated systems and non-operationally evaluated systems. Operationally evaluated systems satisfy the following four criteria:

(i) The model has been evaluated by truly unseen data which are split before any treatment used for designing the output model. Distinguishing test data from training data gives a true estimate of the model.

(ii) The model has been evaluated using ARs at any location in the disk, including far from the center (at the limb). The evaluation only including ARs near the solar disk center increases uncertainty about model performance.

(iii) For the ⩾ M-class flare prediction, the model has been evaluated using ARs including those not linked to any sort of flares. Some models only include ARs linked with ⩾ C-class flares. This approach also raises uncertainty about model performance.

(iv) The model has been designed with enough data. Models fitted with few data are not as effective as those designed with sufficient data.

We do not consider criterion (iii) to distinguish operational and non-operational evaluation because it is for ⩾ M-class flare prediction. Our model is operationally evaluated. For a fair comparison, we denote whether flare prediction models for comparison are operationally evaluated or not.

Table 1 shows the results of the models. We note that it is hard to directly compare the models since the data and time periods are different. This comparison shows that our model is reliable for further analysis of visual explanation. The purpose of our study is not for comparison but the application of our model to attribution methods.

3.2. Attribution Methods

3.2.1. Guided Backpropagation

Guided backpropagation is one of the popular methods that provides a visual explanation of CNN deep learning models. It is a guidance where negative gradients are set to zero by the ReLU during backpropagation. With this guidance, guided backpropagation denoises the saliency map, which makes it clearer than other visual explanation methods (Simonyan et al. 2014; Zeiler & Fergus 2014). We apply the guided backpropagation to

Table 1
Comparison of Models

Model	Event Definition	Operationally Evaluated	ACC	HSS	ApSS	TSS	FB	Event Rate
Our model	Full-disk	yes	0.83	0.65	0.61	0.65	1	0.55
Park et al. (2018)	Full-disk	yes	0.82	0.63	0.60	0.63	1	0.55
Event statistics[a]	Full-disk	yes	0.82	0.36	0.08	0.41	···	0.81
Nishizuka et al. (2018)	AR	yes	0.82	0.53	0.09	0.63	1.53	0.2
Huang et al. (2018)	AR	no	0.76	0.34	−0.61	0.49	2.06	0.15

Note.
[a] Statistical method. Event statistics flare forecasting (Wheatland 2005) by Barnes et al. (2016).

the input image. The result of guided backpropagation on the input image is a detailed gradient mask by pixel. Interpreting the gradient of the input image has to consider the context of the input pixel and many calculations (96,761 parameters) between the input and the output. Guided backpropagation is used for making a saliency map to distinguish an important area for the results of the model. To avoid ambiguity and focus on the magnitude of the gradient, we consider the absolute value of a guided backpropagation mask.

3.2.2. Grad-CAM

Grad-CAM is a strict generalization of class activation mapping (CAM; Zhou et al. 2016). CAM has achieved class-specific feature maps which identify the importance of image regions, but it can be applied to CNN models with a specific kind of architecture. Grad-CAM is a modified version of CAM to be applied to any kind of CNN using the gradient. Every element in the feature maps from CNN has a gradient. To produce a class-specific feature map, Grad-CAM multiplies every feature map and the sum of all gradients of elements in the feature map, sums all weighted feature maps, and applies ReLU to the feature map to set the negative gradation to zero. We apply Grad-CAM to the last convolutional layer of the model.

4. Results and Discussion

It is noted that in order to put the influence of the gradient on the same scale for each result of the model, the values of the gradient mask are scaled in the range 0–1 for plotting and analyzing. Figure 2 shows examples of applying guided backpropagation and Grad-CAM to our model, for the flaring label in a hit case. Figure 2(a) is a line-of-sight magnetogram observed at 00:00 UT on 2013 October 23. During the day, AR 11875 produced seven C-class flares and three M-class flares and AR 11877 produced two C-class flares while other ARs did not produce flares. Figure 2(b) is an example of applying guided backpropagation to our model in the case where the magnetogram is input data. It shows that two ARs, 11875 and 11877, producing flares are more highlighted than the other ARs. Figure 2(c) is an example of applying Grad-CAM to our model in the same case. It is interesting to note that AR 11875 is paid greater attention by our model than AR 11877, while the other ARs are paid weak attention. Such a tendency is consistent with the flaring activities of ARs. These results indicate that our model focuses on ARs, especially flaring ones. Magnetogram data at high bytescales such as ±1500 Gauss could produce different results of attribution methods because they contain more information than the data we used.

Figure 2 shows the results for an interesting case (flaring label in a hit case). Attribution methods also can be used for the non-flare label. In addition, model results are expressed as hit, miss, false alarm, and null.

Figure 3 shows the results of guided backpropagation overlaid on full-disk magnetograms in hit, false alarm, miss, and null. Whether the predictions are right or wrong, the guided backpropagation masks concentrate on certain ARs for the flare label, while they are more spread out for the non-flare label. In false alarm and miss cases, for the flaring label, ARs close to the limb are activated (Figures 3(b) and (c), left). It shows that the projection effect may be a reason for the prediction failure.

In all cases, the parts of the limbs are activated. The guided backpropagation is applied to the earliest stage of the forecasting process, the input image. At this stage, our model considers many interesting features such as small ARs, a wide area filled with spots of negative or positive polarity, and limbs. The effect of such less important features such as limbs disappears as the computation goes on.

Figure 4 is the results of Grad-CAM overlaid on full-disk magnetograms in hit, false alarm, miss, and null cases on the same days as in Figure 3. We can see that Grad-CAM highlights the specific ARs that the guided backpropagation has paid attention to. For the non-flare label, a gradient of zero covers the data in all cases. We apply Grad-CAM without ReLU for the non-flare label and find that negative gradients are in the ARs that are highlighted in the analysis for the flaring label.

Figure 5 shows the AR patches of line-of-sight magnetograms, guided backpropagation masks, and overlaid images for a flaring label in a hit. It is interesting to note that the important pixels are in the vicinity of PILs, implying that our model looks at the PILs when it determines whether this magnetogram produces flares or not.

Table 2 shows the analysis of the guided backpropagation results. The results in Table 2 have uncertainties considering root mean square errors.

We want to note that the classifications of hit, miss, false alarm, and null are based on forecasting results using solar full-disk input data containing flaring and non-flare ARs whether they are a hit or miss case. We analyze flaring and non-flare ARs separately because attribution methods highlight specific ARs (maybe flaring or flaring-like ARs), not all ARs (Figures 2–4). We calculate the ratio of average gradients of guided backpropagation in the PIL to that in the non-PIL area of AR. The ratios are higher than 3.0 for all cases, indicating that the guided backpropagation is usually more activated on the PIL than non-PIL area. The ratios are similar between non-flare ARs and flaring ARs because the guided backpropagation is applied to the earliest stage of forecasting (i.e., input image), which was already mentioned in the discussion of Figure 3.

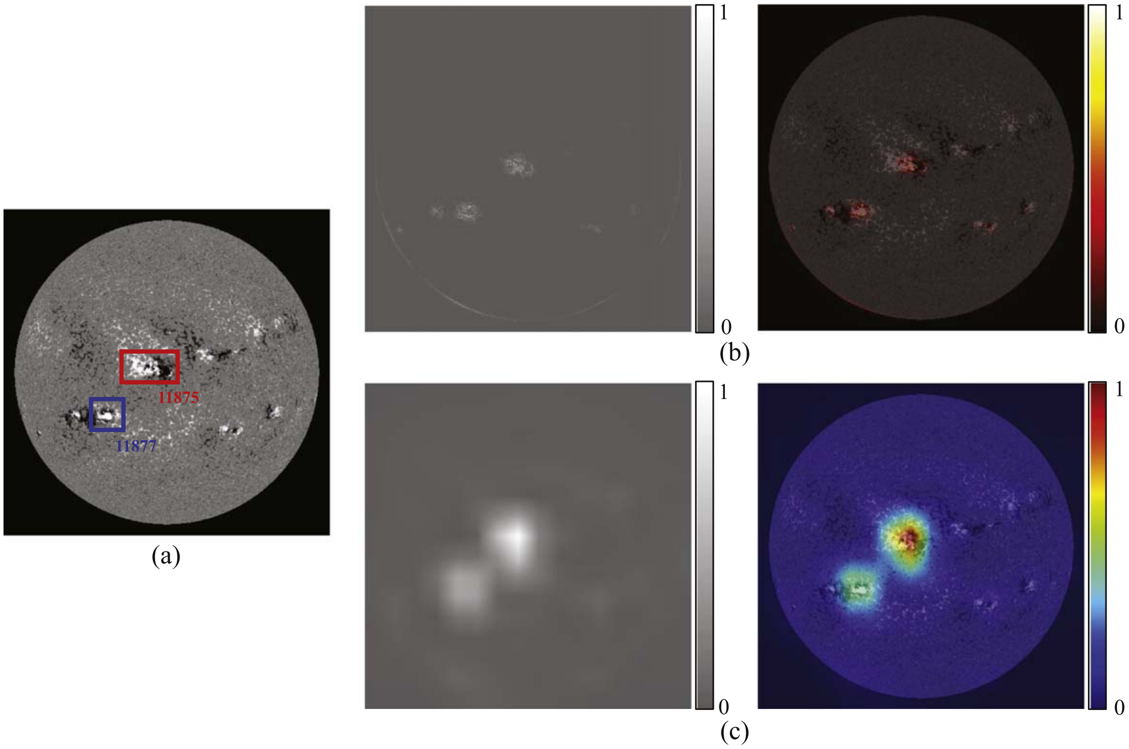


Figure 2. Results of attribution methods for the flaring label in a hit case. (a) A line-of-sight magnetogram observed at 00:00 UT on 2013 October 23. The white area corresponds to positive polarity and the black one to negative polarity. (b) A guided backpropagation mask (left) and the mask overlaid on (a); right. (c) A Grad-CAM mask (left) and the mask overlaid on (a); right. The brightness of the masks corresponds to the magnitude of the gradient of pixels/areas for prediction. Values are scaled from 0 to 1 for each mask.

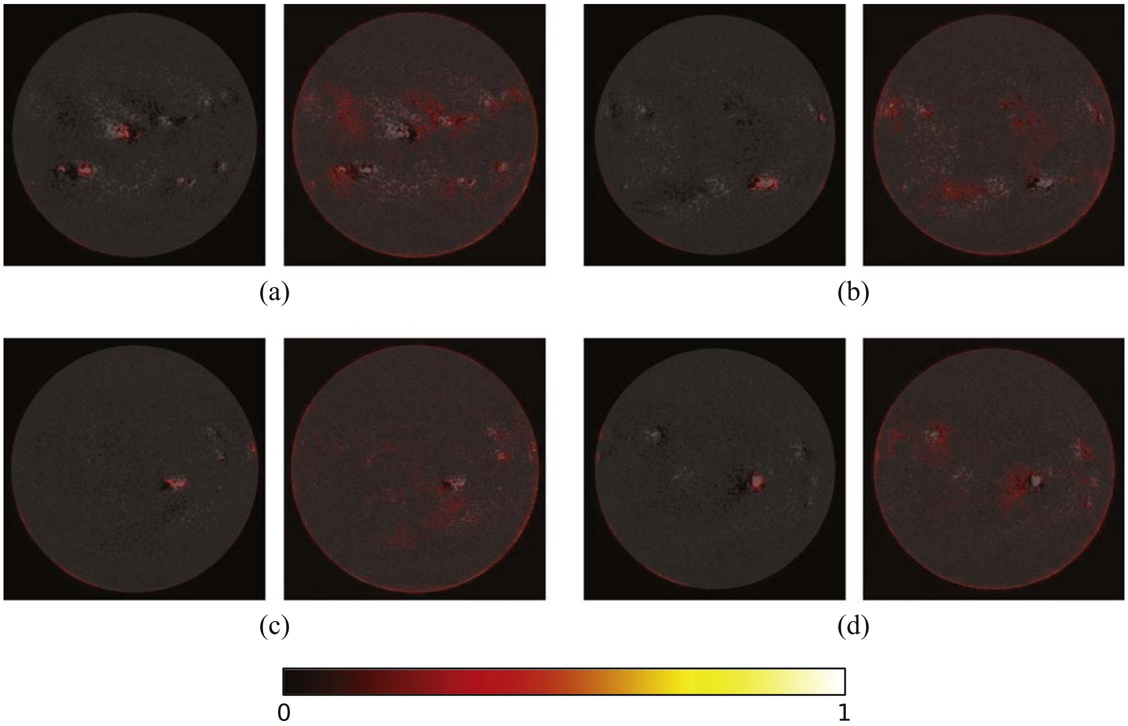


Figure 3. Results of guided backpropagation for a flaring label (left) and a non-flare label (right). (a) 2013 October 23 (hit). (b) 2015 August 10 (false alarm). (c) 2016 December 5 (miss). (d) 2016 May 22 (null). Values are scaled from 0 to 1 for each mask.

From these analysis results, we can interpret not only that the PIL is activated but also that the boundaries of strong polarities are activated. In order to find a better interpretation, we calculate the ratio of average gradients of guided backpropagation in the PIL to that in the boundaries of the positive and negative polarities of ARs.

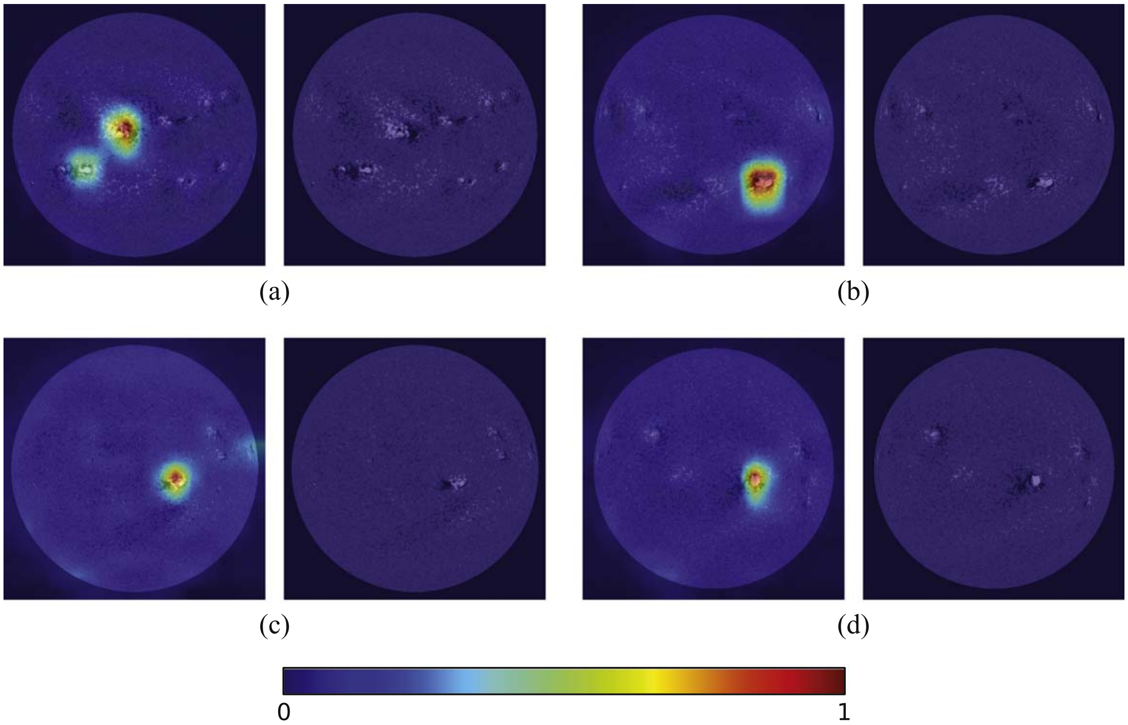


Figure 4. Results of Grad-CAM for a flaring label (left) and a non-flare label (right). (a) 2013 October 23 (hit). (b) 2015 August 10 (false alarm). (c) 2016 December 5 (miss). (d) 2016 May 22 (null). Values are scaled from 0 to 1 for each mask.

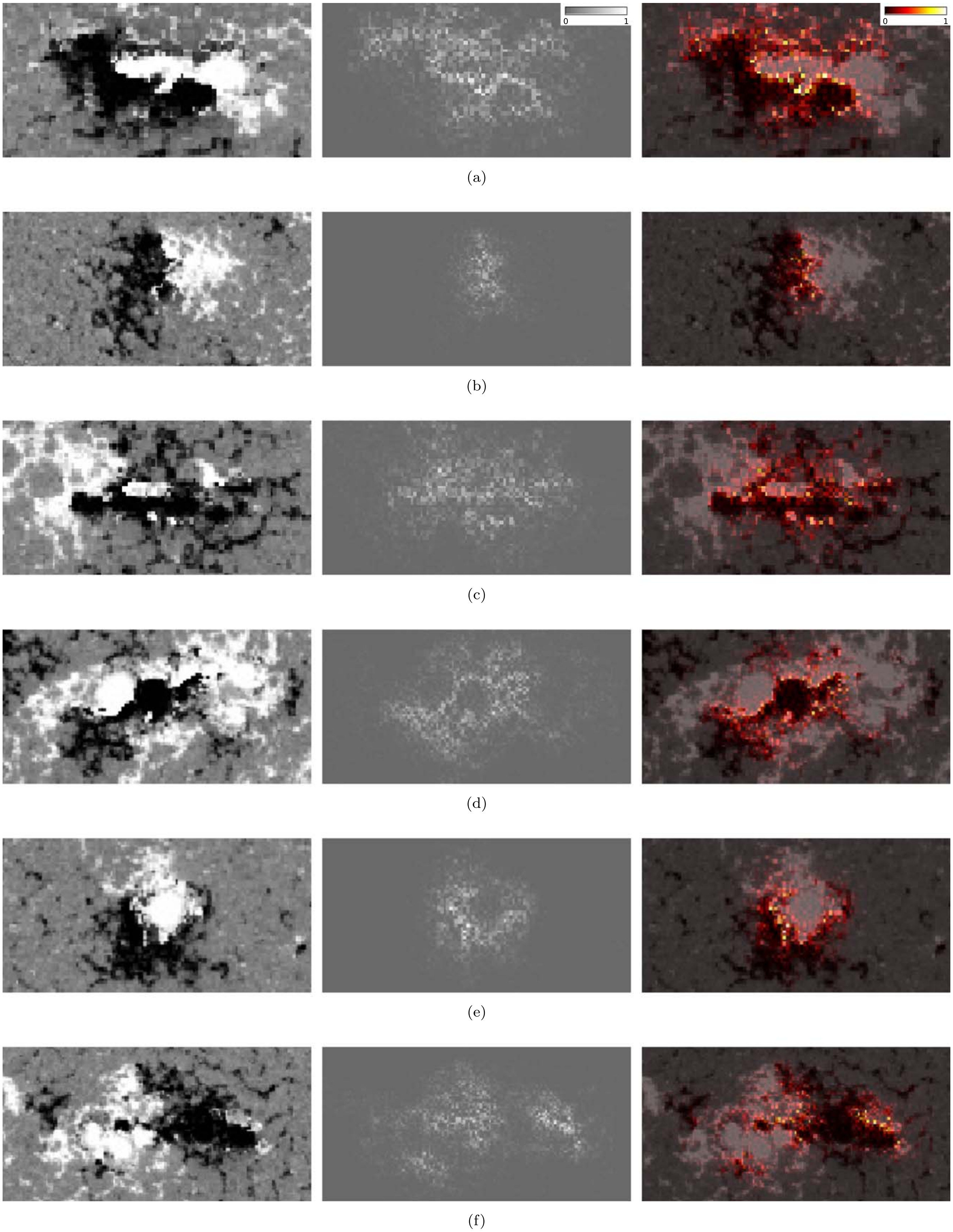
We can find that the ratios of the PIL/boundary are higher than the ratios of PIL/non-PIL for the non-flare ARs in hit, miss, and null cases (we call them the PB-group). In the flaring ARs for hit and miss cases and the non-flare ARs for false alarm cases, it is the opposite (we call them the PN-group). This means that the guided backpropagation would be differently activated for the flaring ARs and the non-flare ARs. To find the reasons for the ratio differences, we examine several ARs by visual inspection. We find that the many ARs in the PB-group consist of the polarities of narrow and long branch-like structure. The guided backpropagation is mainly activated in the vicinity of PIL while not activated in the branch-like polarities. The ARs in the PN-group have usually simpler structures than the ARs in the PB-group.

The values of the ratios and their analyses suggest that the guided backpropagation mask is more focused in the PIL for the flaring ARs than for the non-flare ARs. In order to find the association between the guided backpropagation and the PIL of AR, we calculate the average IoU between the PIL and the guided backpropagation mask. IoU is an evaluation metric of segmentation and object detection, which is the ratio between the intersection area of the object and segmentation to the union area of them. For this, we use the subsets of the gradient masks with a normalized gradient of 0.1 or larger. The average IoU of the flaring AR is higher than that of a non-flare AR, but not beyond uncertainties corresponding to the root mean square errors. IoU values are quite small because they are the results of the line (PIL) versus the area (mask). Considering the analysis results, we can say that the PIL is an important feature for our model to recognize flaring ARs.

Many research groups have shown that a flare occurrence and its class are associated with PILs (Schrijver 2007; Kim et al. 2008; Mason & Hoeksema 2010; Falconer et al. 2011, 2012, 2014; Schrijver 2016; Sadykov & Kosovichev 2017; Sharykin et al. 2017; Toriumi & Takasao 2017; Vasantharaju et al. 2018). Schrijver (2007, 2016) and Vasantharaju et al. (2018) showed that large flares are associated with pronounced high-gradient PILs. Kim et al. (2008) pointed out that preflare activities such as sigmoidal UV structure appeared along the PIL. Mason & Hoeksema (2010) showed that GWILL, which combines the PIL length and gradient across it, is a better parameter than effective separation (Chumak & Chumak 1987; Chumak et al. 2004; Guo et al. 2006) and total unsigned magnetic flux for flare forecast. Falconer et al. (2011, 2012, 2014) investigated flare forecast using the transverse gradient of the line-of-sight magnetic field and potential transverse field on the PIL. Sharykin et al. (2017) showed that the flare energy release develops near the PIL. Sadykov & Kosovichev (2017) indicated that flare forecasts based on PIL characteristics are more effective than those using global characteristics of ARs. These studies show that flaring activities are very intimately associated with the characteristics of PILs, which are consistent with our results.

To find the relationship between Grad-CAM results and the flare occurrence rate, we define the Grad-CAM value for a given HARP as a maximum value of Grad-CAM in the area of HARP. Figure 6 shows the relationship between the average flare occurrence rate (within 24 hr) and the Grad-CAM values for all HARPs of hit, miss, and false alarm. As shown in the figure, all flares (C-, M-, and X-class) occur more frequently for ARs with higher Grad-CAM values.

In addition, we compare the Grad-CAM value with 16 SHARP parameters suggested by Bobra et al. (2014) in hit, miss, and false alarm cases. For this, we use the flare history of HARPs within ±60° longitudes of the central meridian to minimize the projection effect. To estimate the relationships between the Grad-CAM values and SHARP parameters, we consider Pearson correlation coefficients (r). In view of the absolute correlation coefficient, the top nine SHARP parameters whose correlation coefficients are larger than 0.5 are as follows: total unsigned vertical current (TOTUSJZ, $r = 0.72$), total



(a)

(b)

(c)

(d)

(e)

(f)

Figure 5. Patches of ARs at 00:00 UT (left), guided backpropagation masks (center), and masks overlaid on the patches (right) for the flaring label. (a) AR 11429 on 2012 March 9. (b) AR 11564 on 2012 September 4. (c) AR 11731 on 2013 May 1. (d) AR 11967 on 2014 February 4. (e) AR 12158 on 2014 September 10. (f) AR 12339 on 2015 May 12. The color bars in (a) are applied to (b), (c), (d), (e), and (f).

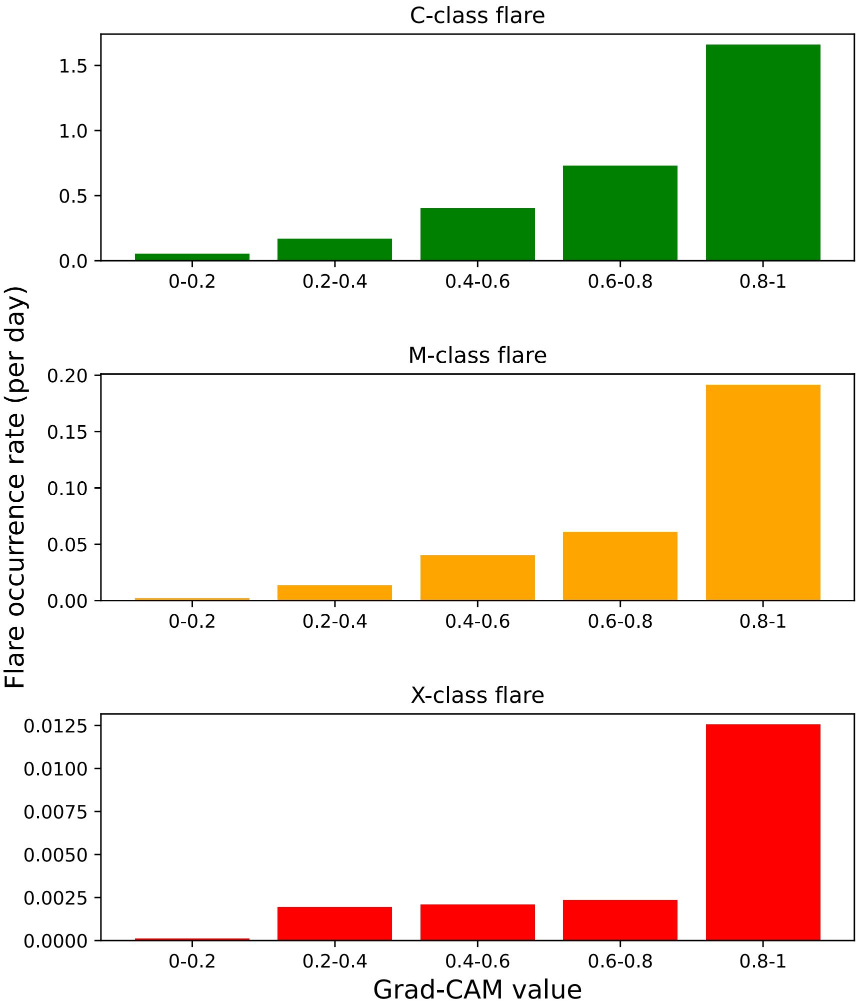


Figure 6. Average flare occurrence rate according to the Grad-CAM value in hit, miss, and false alarm cases.

Table 2
Analysis of the Guided Backpropagation with Root Mean Square Error

	PIL/Non-PIL		PIL/Boundary		IoU	
	Flaring AR (⩾C)	Non-flare AR (<C)	Flaring AR (⩾C)	Non-flare AR (<C)	Flaring AR (⩾C)	Non-flare AR (<C)
Hit	3.996 ± 1.787	3.815 ± 2.320	3.768 ± 1.840	3.820 ± 2.363	0.112 ± 0.059	0.069 ± 0.077
Miss	3.691 ± 1.755	3.428 ± 2.343	3.626 ± 1.720	3.578 ± 2.488	0.077 ± 0.051	0.051 ± 0.068
False alarm	⋯	3.584 ± 2.085	⋯	3.581 ± 2.113	⋯	0.056 ± 0.058
Null	⋯	3.209 ± 1.834	⋯	3.232 ± 1.874	⋯	0.048 ± 0.060

Note. (left) The ratio of average gradients in the PIL to that in the non-PIL area of AR. (center) The ratio of average gradients in the PIL to that in the boundary AR. (right) The average IoU between PIL and guided backpropagation mask.

unsigned current helicity (TOTUSJH, $r = 0.71$), total unsigned flux (USFLUX, $r = 0.70$), total photospheric magnetic free energy density (TOTPOT, $r = 0.62$), mean photospheric excess magnetic energy density (MEANPOT, $r = 0.57$), shear angle (MEANSHR, $r = 0.57$), fractional area with shear >45° (SHRGT45, $r = 0.56$), sum of the modulus of the net current per polarity (SAVNCPP, $r = 0.55$), and absolute value of the net current helicity (ABSNJZH, $r = 0.52$). Figure 7 shows box plots of the relationships between the Grad-CAM values and the nine SHARP parameters.

Our findings are consistent with previous studies. The top 8 SHARP parameters except for MEANSHR in our study are used for flare prediction by Bobra & Couvidat (2015) who used 13 SHARP parameters having high Fisher ranking scores. Liu et al. (2017) suggested that TOTUSJZ and TOTUSJH are the most important parameters to classify flaring ARs considering Gini importance (Breiman 2001). Toriumi & Takasao (2017) noted from numerical simulations that TOTUSJH and TOTPOT are highly correlated with the stored magnetic free energy which is the maximum flare energy that could be released. ALL

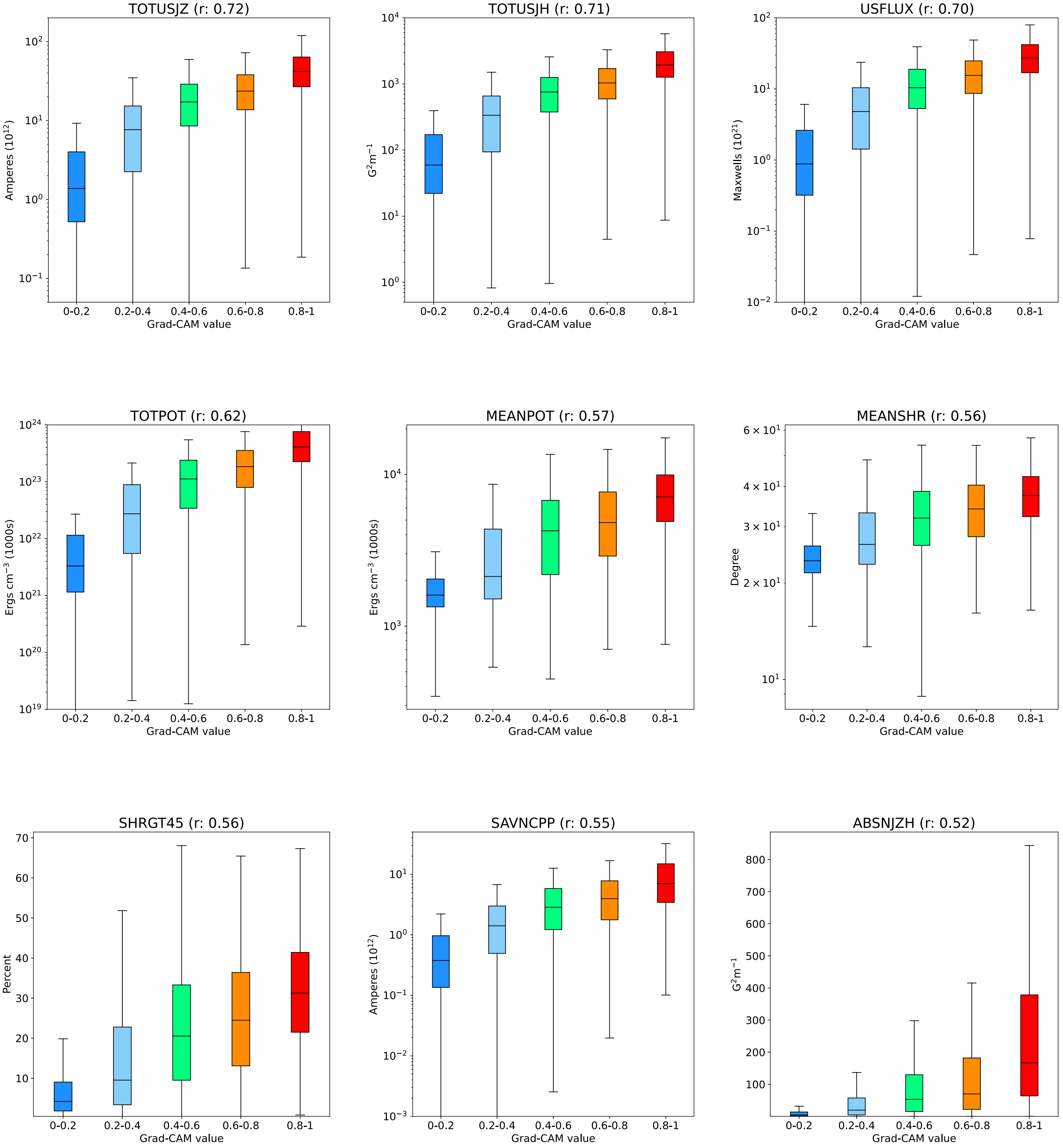


Figure 7. Box plots of the relationships between nine SHARP parameters and Grad-CAM values in Hit, Miss, and False alarm cases. Horizontal lines in the boxes are the median of the data. The upper/lower bounds of the boxes are the upper/lower quartiles of the data, while the upper/lower bounds of whiskers are the highest/lowest data below/above 1.5 times the box range from the upper/lower bounds of the boxes. Outliers, which are beyond the caps of the whisker, are not represented. Graphs are expressed in logarithm scale but ABSNJZH and SHRGT45 are in linear scale because their minimum values are zero.

six SHARP parameters (TOTUSJH, TOTUSJZ, TOTPOT, USFLUX, ABSNJZH, and SAVNCPP) used in Lim et al. (2019a), which have higher correlations with major flare occurrence larger than 0.86, also belong to our selection. Our results imply that interpretations of our flare forecast model using guided backpropagation and Grad-CAM is consistent with SHARP parameters associated with non-potentiality and PILs.

5. Summary and Conclusion

In this study we have presented visual explanation of our deep learning flare model and investigated its relationship with the physical parameters of ARs. For this, we consider solar full-disk line-of-sight magnetograms from SOHO/MDI and SDO/HMI, SHARP parameters from the HMI data, and GOES X-ray flare data. We make a new flare forecast model, which is based on CNN, that has better performance than other deep learning models. For the first time, we apply guided backpropagation and Grad-CAM to our model for interpretation.

The major results of this study are as follows. First, we successfully apply our deep learning models to the forecast of daily solar flare occurrences. Second, the results of guided backpropagation show that the flare occurrence prediction score of the model is mainly determined by the vicinity of PILs. Third, the ARs with high Grad-CAM values produce more flares than those with low Grad-CAM values. Fourth, the top nine SHARP parameters such as TOTUSJZ ($r = 0.72$), TOTUSJH ($r = 0.71$), USFLUX ($r = 0.70$), TOTPOT ($r = 0.62$), MEANPOT ($r = 0.57$), MEANSHR ($r = 0.57$), SHRGT45 ($r = 0.56$), SAVNCPP ($r = 0.55$), and ABSNJZH ($r = 0.52$) are well correlated with Grad-CAM values.

Much previous research suggests that flare occurrence is intimately related to PILs and SHARP parameters that denote the non-potentiality and magnetic complexity of ARs (Schrijver 2007; Kim et al. 2008; Mason & Hoeksema 2010; Falconer et al. 2011, 2012, 2014; Bobra & Couvidat 2015; Schrijver 2016; Liu et al. 2017; Sadykov & Kosovichev 2017; Sharykin et al. 2017; Toriumi & Takasao 2017; Vasantharaju et al. 2018; Lim et al. 2019a, 2019b). Very interestingly, guided backpropagation and Grad-CAM, which we adopt, show that our model pays attention to PILs and ARs with large values of the SHARP parameters under consideration.

In the SHARP parameter analysis using Grad-CAM, the best correlated parameters are insensitive to PILs while the guided backpropagation highlights PILs. The SHARP parameter analysis result could have a dependency on the characteristic of Grad-CAM that considers filtered and compressed spatial information. The interpretation results of the model could be different if other attribution methods are applied. To estimate the model dependence for interpretation by attribution methods, we have applied Grad-CAM++ (Chattopadhay et al. 2018) to the model. The differences between the analysis results of Grad-CAM and Grad-CAM++ are very small.

We really appreciate the constructive comments of the referee and the editor. We thank the numerous team members who have contributed to the success of the SDO mission as well as the SOHO mission. We thank contributors to Pytorch, Numpy, and the Matplotlib open-source package. This work was supported by the BK21 plus program through the National Research Foundation (NRF) funded by the Ministry of Education of Korea, the Basic Science Research Program through the NRF funded by the Ministry of Education (NRF-2016R1A2B4013131, NRF-2019R1A2C1002634, NRF-2019R1C1C1004778, NRF-2020R1C1C1003892), the Korea Astronomy and Space Science Institute (KASI) under the R&D program "Study on the Determination of Coronal Physical Quantities using Solar Multi-wavelength Images (project No. 2019-1-850-02)" supervised by the Ministry of Science and ICT, and the Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korean government (MSIP) (2018-0-01422, Study on analysis and prediction technique of solar flares).

ORCID iDs

Kangwoo Yi https://orcid.org/0000-0003-4342-9483
Yong-Jae Moon https://orcid.org/0000-0001-6216-6944
Daye Lim https://orcid.org/0000-0001-9914-9080
Eunsu Park https://orcid.org/0000-0003-0969-286X
Harim Lee https://orcid.org/0000-0002-9300-8073

References

Allouche, O., Tsoar, A., & Kadmon, R. 2006, J. Appl. Ecol., 43, 1223
Appleman, H. S. 1960, BAMS, 41, 64
Barnes, G., Leka, K. D., Schrijver, C. J., et al. 2016, ApJ, 829, 89
Bobra, M. G., & Couvidat, S. 2015, ApJ, 798, 135
Bobra, M. G., Sun, X., Hoeksema, J. T., et al. 2014, SoPh, 289, 3549
Breiman, L. 2001, Mach. Learn., 45, 5
Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. 2018, in 2018 IEEE Winter Conf. Applications of Computer Vision (WACV) (Piscataway, NJ: IEEE), 839
Chen, Y., Manchester, W. B., Hero, A. O., et al. 2019, SpWea, 17, 1404
Chumak, O., Zhang, H., & Gou, J. 2004, A&AT, 23, 525
Chumak, O. V., & Chumak, Z. N. 1987, KFNT, 3, 7
Cinto, T., Gradvohl, A. L. S., Coelho, G. P., & da Silva, A. E. A. 2020, MNRAS, 495, 3332
Colak, T., & Qahwaji, R. 2009, SpWea, 7, S06001
Domingo, V., Fleck, B., & Poland, A. I. 1995, SoPh, 162, 1
Falconer, D., Barghouty, A. F., Khazanov, I., & Moore, R. 2011, SpWea, 9, S04003
Falconer, D. A., Moore, R. L., Barghouty, A. F., & Khazanov, I. 2012, ApJ, 757, 32
Falconer, D. A., Moore, R. L., Barghouty, A. F., & Khazanov, I. 2014, SpWea, 12, 306
Guo, J., Zhang, H., Chumak, O. V., & Liu, Y. 2006, SoPh, 237, 25
Heidke, P. 1926, Geografiska Annaler, 8, 301
Hochreiter, S., & Schmidhuber, J. 1997, Neural Computation, 9, 1735
Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. 2017, in 2017 IEEE Conf. Computer Vision and Pattern Recognition (CVPR) (Piscataway, NJ: IEEE), 2261
Huang, X., Wang, H., Xu, L., et al. 2018, ApJ, 856, 7
Ioffe, S., & Szegedy, C. 2015, PMLR, 37, 448
Kim, S., Moon, Y.-J., Kim, Y.-H., et al. 2008, ApJ, 683, 510
Lecun, Y., Bengio, Y., & Hinton, G. 2015, Natur, 521, 436
Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. 1998, IEEEP, 86, 2278
Leka, K. D., Park, S.-H., Kusano, K., et al. 2019, ApJS, 243, 36
Li, X., Zheng, Y., Wang, X., & Wang, L. 2020, ApJ, 891, 10
Lim, D., Moon, Y.-J., Park, E., et al. 2019a, ApJ, 885, 35
Lim, D., Moon, Y.-J., Park, J., et al. 2019b, JKAS, 52, 133
Liu, C., Deng, N., Wang, J. T. L., & Wang, H. 2017, ApJ, 843, 104
Liu, H., Liu, C., Wang, J. T. L., & Wang, H. 2019, ApJ, 877, 121
Liu, Y., Hoeksema, J. T., Scherrer, P. H., et al. 2012, SoPh, 279, 295
Mason, J. P., & Hoeksema, J. T. 2010, ApJ, 723, 634
Nair, V., & Hinton, G. E. 2010, in Proc. 27th Int. Conf. Int. Conf. Machine Learning, ICML'10 (Madison, WI: Omnipress), 807
Nishizuka, N., Sugiura, K., Kubo, Y., et al. 2017, ApJ, 835, 156
Nishizuka, N., Sugiura, K., Kubo, Y., Den, M., & Ishii, M. 2018, ApJ, 858, 113
Park, E., Moon, Y.-J., Shin, S., et al. 2018, ApJ, 869, 91
Pesnell, W. D., Thompson, B. J., & Chamberlin, P. C. 2012, SoPh, 275, 3
Portilla, J., & Simoncelli, E. P. 2000, Int. J. Comput. Phys, 40, 49
Priest, E. R., & Forbes, T. G. 2002, A&ARv, 10, 313
Ryan, D. F., Dominique, M., Seaton, D., Stegen, K., & White, A. 2016, A&A, 592, A133
Sadykov, V. M., & Kosovichev, A. G. 2017, ApJ, 849, 148
Sampat, M. P., Wang, Z., Gupta, S., Bovik, A. C., & Markey, M. K. 2009, ITIP, 18, 2385
Scherrer, P. H., Bogart, R. S., Bush, R. I., et al. 1995, SoPh, 162, 129
Schou, J., Scherrer, P. H., Bush, R. I., et al. 2012, SoPh, 275, 229
Schrijver, C. J. 2007, ApJL, 655, L117
Schrijver, C. J. 2016, ApJ, 820, 103
Selvaraju, R. R., Cogswell, M., Das, A., et al. 2017, in 2017 IEEE Int. Conf. Computer Vision (ICCV) (Piscataway, NJ: IEEE), 618
Sharykin, I. N., Sadykov, V. M., Kosovichev, A. G., Vargas-Dominguez, S., & Zimovets, I. V. 2017, ApJ, 840, 84
Shibata, K., & Magara, T. 2011, LRSP, 8, 6
Simonyan, K., Vedaldi, A., & Zisserman, A. 2014, arXiv:1312.6034

Springenberg, J., Dosovitskiy, A., Brox, T., & Riedmiller, M. 2015, arXiv:1412.6806

Toriumi, S., & Takasao, S. 2017, ApJ, 850, 39

Vasantharaju, N., Vemareddy, P., Ravindra, B., & Doddamani, V. H. 2018, ApJ, 860, 58

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. 2004, ITIP, 13, 600

Wheatland, M. S. 2005, SpWea, 3, S07003

Zeiler, M. D., & Fergus, R. 2014, in Computer Vision—ECCV 2014, ed. D. Fleet et al. (Cham: Springer International Publishing), 818

Zhou, B., Khosla, A. A. L., Oliva, A., & Torralba, A. 2016, in 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (Piscataway, NJ: IEEE), 2921