Journal of the Korean Astronomical Society
52: 217 ∼ 225, 2019 December
Published under Creative Commons license CC BY-SA 4.0

https://doi.org/10.5303/JKAS.2019.52.6.217
pISSN: 1225-4614 · eISSN: 2288-890X
http://jkas.kas.org

Modified Convolutional Neural Network with Transfer Learning for Solar Flare Prediction

Yanfang Zheng, Xuebao Li, Xinshuo Wang, and Ta Zhou

College of Electrical and Information Engineering, Jiangsu University of Science and Technology, Zhangjiagang 215600, China; zyf062856@163.com

Received July 3, 2019; accepted September 2, 2019

Abstract: We apply a modified Convolutional Neural Network (CNN) model in conjunction with transfer learning to predict whether an active region (AR) would produce a ≥C-class or ≥M-class flare within the next 24 hours. We collect line-of-sight magnetogram samples of ARs provided by the SHARP from May 2010 to September 2018, which is a new data product from the HMI onboard the *SDO*. Based on these AR samples, we adopt the approach of shuffle-and-split cross-validation (CV) to build a database that includes 10 separate data sets. Each of the 10 data sets is segregated by NOAA AR number into a training and a testing data set. After training, validating, and testing our model, we compare the results with previous studies using predictive performance metrics, with a focus on the true skill statistic (TSS). The main results from this study are summarized as follows. First, to the best of our knowledge, this is the first time that the CNN model with transfer learning is used in solar physics to make binary class predictions for both ≥C-class and ≥M-class flares, without manually engineered features extracted from the observational data. Second, our model achieves relatively high scores of TSS = 0.640±0.075 and TSS = 0.526±0.052 for ≥M-class prediction and ≥C-class prediction, respectively, which is comparable to that of previous models. Third, our model also obtains quite good scores in five other metrics for both ≥C-class and ≥M-class flare prediction. Our results demonstrate that our modified CNN model with transfer learning is an effective method for flare forecasting with reasonable prediction performance.

Key words: magnetic fields — Sun: activity — Sun: flares — techniques: image processing

1. Introduction

Solar flares are sudden eruptions accompanied by brightening in multiple wavelengths and large-scale energy release in a short time in the solar upper atmosphere. Strong flares may eject large amounts of high-energy particles into interplanetary space. These high energy particles could threaten the safety of satellites and astronauts, and could, among other effects, cause temporary failure of radio communication systems and global positioning systems. Hence, it is necessary to build a solar flare prediction model to help human decision makers and affected services take action before strong flares occur.

Solar flares originate from the magnetic energy stored around sunspots in an active region (AR). The impulse release of energy produces flares, which can be directly observed by a photospheric magnetogram (Priest & Forbes 2002; Shibata & Magara 2011). Therefore, many flare prediction studies are inclined to adopt magnetic observational data to parameterize AR and, in the absence of a definitive physical theory to explain the flaring mechanism of an AR, find empirical relationships between parameters or features of AR and flare productivity. Thus, many flare prediction models have been developed with different methods in recent years. Some of them are based on statistical algorithms (Bloomfield et al. 2012; Song et al. 2009; Mason & Hoeksema 2010; Barnes et al. 2016), and most of them are based on machine learning algorithms. Machine learning has become an increasingly popular approach for solar flare forecasts. The classical machine learning algorithms applied to solar flare prediction models are support vector machines (Yuan et al. 2010; Bobra & Couvidat 2015; Nishizuka et al. 2017; Sadykov & Kosovichev 2017), artificial neural networks (Qahwaji & Colak 2007; Ahmed et al. 2013; Li & Zhu 2013; Nishizuka et al. 2018), Bayesian network approaches (Yu et al. 2010), random forest algorithms (Liu et al. 2017; Florios et al. 2018), and ensemble learning (Colak & Qahwaji 2009; Huang et al. 2010; Guerra et al. 2015), respectively. However, one fatal weakness of both statistical and classical machine learning algorithms is that they are strongly dependent on the quality of engineered features manually extracted from the observational data.

Deep learning neural networks (DNNs) are a new branch of machine learning, which has emerged as a highly dependable technique to solve large scale learning problems in astronomy and other branches of science (Abraham et al. 2018). A well-known method of DNNs, convolutional neural networks (CNNs; LeCun et al. 2015), falling within the realm of image processing and computer vision, are able to perform automatic feature extraction, and are more suitable to high dimensional data sets. Based on neural networks, CNNs include several hidden layers which can extract com-

Corresponding author: X. Li

Table 1

Number of solar magnetogram samples and ARs for 10 separate data sets.

Data set	No-flare Level (Original/ Preprocessed/ AR numbers)	C Level (Original/ Preprocessed/ AR numbers)	M Level (Original/ Preprocessed/ AR numbers)	X Level (Original/ Preprocessed/ AR numbers)
No. 1: Training	46013/8796/359	55940/11091/237	5116/8862/60	238/2856/8
Test	10732/1331/57	16472/1756/39	1418/1146/8	205/1500/2
Total	56745/10127/416	72412/12847/276	6534/10008/68	443/4356/10
No. 2: Training	45587/8824/351	57040/11242/235	5417/9408/60	340/4080/8
Test	11158/1733/72	15372/1589/36	1117/876/7	103/60/1
Total	56745/10557/423	72412/12831/271	6534/10284/67	443/4140/9
No. 3: Training	46192/8996/362	58412/11549/235	4836/8562/59	399/4788/8
Test	10553/1382/58	14000/1379/35	1698/1950/11	44/312/2
Total	56745/10378/420	72412/12928/270	6534/10512/70	443/5100/10
No. 4: Training	44791/8807/347	56244/10864/239	5030/8814/60	307/3684/8
Test	11954/1565/66	16168/1635/37	1504/828/8	136/1416/2
Total	56745/10372/413	72412/12499/276	6534/9642/68	443/5100/10
No. 5: Training	45573/8951/352	58207/11611/234	5312/9270/59	313/3756/8
Test	11172/1664/68	14205/1686/38	1222/1218/9	130/600/2
Total	56745/10615/420	72412/13237/272	6534/10488/68	443/4356/10
No. 6: Training	46091/8836/359	56737/11189/238	5125/8580/62	304/3648/8
Test	10654/1500/62	15675/1641/39	1409/1332/5	139/1668/3
Total	56745/10336/421	72412/12830/277	6534/9912/67	443/5316/11
No. 7: Training	44940/8633/353	57521/11450/242	5015/8856/61	320/3840/8
Test	11805/1579/68	14891/1025/27	1519/1692/9	123/1476/3
Total	56745/10212/421	72412/12475/269	6534/10548/70	443/5316/11
No. 8: Training	46104/8908/357	59304/11689/243	5554/9468/58	386/4632/8
Test	10641/1459/64	13108/925/27	980/750/7	57/312/2
Total	56745/10367/421	72412/12614/270	6534/10218/65	443/4944/10
No. 9: Training	46139/8964/359	57209/11398/241	5023/9174/62	292/3504/8
Test	10606/1413/61	15203/1473/34	1511/1038/5	151/1440/2
Total	56745/10377/420	72412/12871/275	6534/10212/67	443/4944/10
No. 10: Training	45479/8632/357	58947/11473/236	5272/9216/60	324/3888/8
Test	11266/1517/65	13465/1266/30	1262/1392/10	119/252/1
Total	56745/10149/422	72412/12739/266	6534/10608/70	443/4140/9

In every level of each data set, ‘Original’ represents the number of magnetogram samples before image preprocessing, ‘Preprocessed’ represents the number of magnetogram samples after image preprocessing used for our study, and ‘AR numbers’ represents the number of ARs after image preprocessing used for our study. Image preprocessing comprises excluding samples with multiple NOAA ARs, data augmentation, and undersampling.

plex features from data in preparation for classification or regression tasks. The raw data can be fed into the network, thus minimal to no feature engineering is required, and the network learns to extract the features through training. CNNs have been successfully applied to astrophysics. Huang et al. (2018) presented a model based on a CNN for flare prediction via binary classification. Their model used many paths of solar ARs from line-of-sight (LOS) magnetograms located within ±30° of the solar disk center to avoid projection effects. Park et al. (2018) presented a CNN model to predict solar flare. Their models employed full-disk solar LOS magnetograms to make binary class predictions within 24 hours.

In this paper, we use a modified CNN model with transfer learning to predict, via binary classification, the probability that a flare occurs in an AR within 24 hours. Transfer learning enables us to train a network for a new application with few training samples. We exploit the properties of transfer learning in CNNs to train a pre-existing model for a different classification problem, and improve the existing model without having to retrain it from scratch. This approach also has been successfully applied in many areas of computer vision (Razavian et al. 2014; Aniyan & Thorat 2017).

This paper is organized as follows. The data is described in Section 2. A forecasting model based on a CNN with transfer learning is presented in Section 3. Results are given in Section 4, and finally, conclusions and discussions are provided in Section 5.

2. Data

The data used in this work is obtained from LOS magnetograms of ARs provided by the Helioseismic and Magnetic Imager (HMI; Schou et al. 2012) on board the *Solar Dynamics Observatory* (*SDO*; Pesnell et al. 2012). *SDO*/HMI started its routine observations on 2010 April 30, and has provided continuous and high-quality observational data of photospheric magnetic fields. At the end of 2012, *SDO*/HMI started to release a new data product, Space-weather HMI Active Region Patches (SHARP; Bobra et al. 2014), which is convenient for flare forecasting. The data is available to the public at the Joint Science Operations Center, and the LOS magnetograms of ARs can be obtained from the SHARP (`hmi.sharp_cea_720s`). We use SHARP LOS magnetogram data from 2010 May 1 to 2018 September 13, covering the main peak of solar cycle 24, as input data for our prediction model. The cadence of the data is 12 minutes. The data selected in this work is located within ±45° of the central meridian to avoid projection effects (Ahmed et al. 2013; Bobra et al. 2014). The output of the prediction model is compared with the daily flare observations of the *Geostationary Operational Environment Satellite* (*GOES*). Solar flares are usually divided into B, C, M, or X level events according to the peak magnitude of the soft X-ray flux observed by the *GOES*. The solar flare information is obtained from the NOAA database[1] which is constructed from the *GOES* X-ray flare listings. Note that many records of flare events lack locations and National Oceanic and Atmospheric Administration (NOAA) AR numbers. Thus we add the location information and AR numbers for the associated flares referring to the information from Solar Geophysical Data (SGD) solar event reports.[2]

We also build a data set catalog for our model. The catalog is prepared as follows: (1) The No-flare (weaker than C1.0) label is assigned to the magnetogram data if the AR does not flare within 24 hr after the observation time. (2) The corresponding flare labels (i.e. C, M, or X) are assigned to the magnetogram data if an C/M/X-level flare occurs within 24 hr after the observation time of the magnetogram sample. For recurring flares with different levels, the magnetograms within 24 hr before the first flare are labeled according to the level of the first flare. For subsequent flares of the same AR, the corresponding labels are assigned to the magnetograms for the time interval from the end of the previous flare to the end of the flare under consideration. (3) ARs are further categorized into four levels in the event of at least one flare with the corresponding *GOES* level but no flares of higher *GOES* levels (Liu et al. 2017; Yuan et al. 2010; Song et al. 2009): 'No-flare' indicates that the AR only produces microflares (weaker than C1.0 flares); 'C' means that the AR produces at least one C-level flare but no M/X-level flares; 'M' indicates that the AR produces at least one M-level flares but no X-level flares; 'X' means that the AR produces at least one X-level flare. We collect 870 ARs and 136,134 magnetogram samples in total, containing 443 X-level, 6534 M-level, 72,412 C-level, and 56,745 No-flare level magnetogram samples. For the ≥C class, magnetogram samples including C/M/X-level flares in an AR are classified as postive, all other magnetogram samples in the AR are negative. For the ≥M class, magnetogram samples including M/X-level flares in an AR are classified as positive, and all other magnetogram samples in the AR are negative. Note that the magnetogram samples from SHARP with multiple NOAA ARs (Bobra et al. 2014) are excluded in our work. In other words, if there are two or more ARs in a magnetogram, this magnetogram is not used in this study.

Table 2
Architecture of our modified CNN model.

Layer	Function	Kernel size	Depth	Stride length
1	Convolution	3×3	64	1
2	Convolution	3×3	64	1
3	Max-pooling	2×2	–	2
4	Convolution	3×3	128	1
5	Convolution	3×3	128	1
6	Max-pooling	2×2	–	2
7	Convolution	3×3	256	1
8	Convolution	3×3	256	1
9	Convolution	3×3	256	1
10	Max-pooling	2×2	–	2
11	Convolution	3×3	512	1
12	Convolution	3×3	512	1
13	Convolution	3×3	512	1
14	Max-pooling	2×2	–	2
15	Convolution	3×3	512	1
16	Convolution	3×3	512	1
17	Convolution	3×3	512	1
18	Max-pooling	2×2	–	2
19	Fully Connected Layer	128	–	–
20	ReLU	–	–	–
21	Batch Normalization	–	–	–
22	Fully Connected Layer	2	–	–
23	Softmax	–	–	–

Ultimately, we are interested in developing a real-time predictive model to forecast future flare activity based on current solar data. It is worth noting that the magnetograms are similar to each other in the same AR. In order to get a sense for how our model will perform on yet-unseen data, we simulate this process by partitioning the whole data set into training data and testing data according to their NOAA AR numbers. Different from the training data, the testing data set is only used to evaluate our model. The ratio of the training and testing data is about 80%:20%. We note here that it is not easy to exactly partition the whole data set into training and testing data set by AR numbers; the magnetogram samples of a given AR are put in either the training or testing data set. This ensures that our model is evaluated on ARs which have never appeared in the

[1] `https://www.ngdc.noaa.gov/stp/space-weather/solar-data/solar-features/solar-flares/x-rays/goes/xrs`

[2] `http://www.solarmonitor.org/index.php`

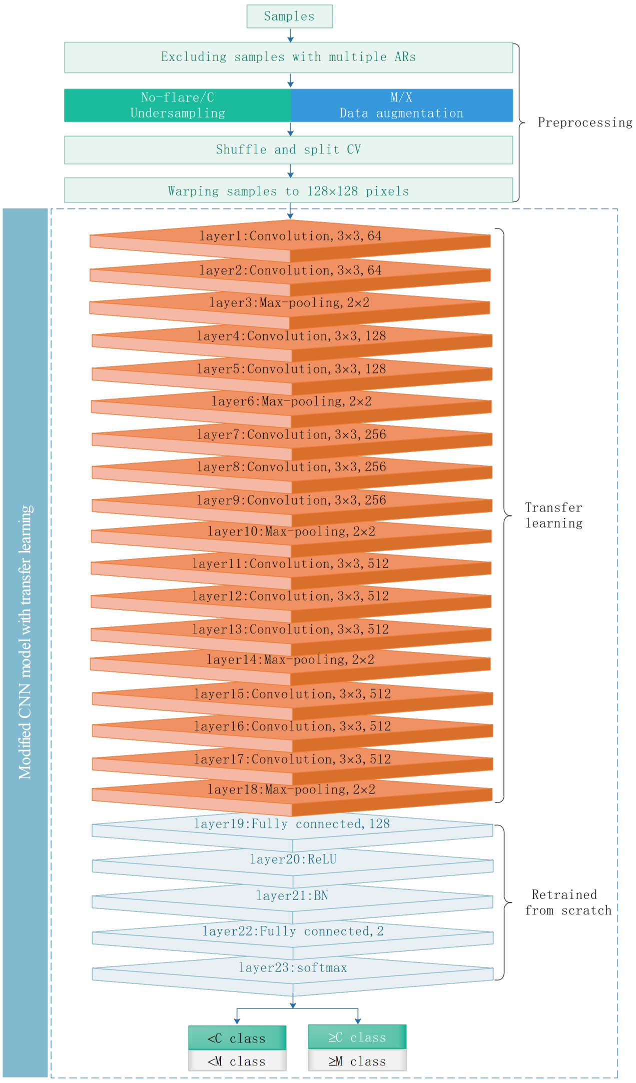
Samples
Excluding samples with multiple ARs
No-flare/C
Undersampling
M/X
Data augmentation
Shuffle and split CV
Warping samples to 128×128 pixels
Preprocessing
Modified CNN model with transfer learning
layer1:Convolution, 3x3, 64
layer2:Convolution, 3x3, 64
layer3:Max-pooling, 2x2
layer4:Convolution, 3x3, 128
layer5:Convolution, 3x3, 128
layer6:Max-pooling, 2x2
layer7:Convolution, 3x3, 256
layer8:Convolution, 3x3, 256
layer9:Convolution, 3x3, 256
layer10:Max-pooling, 2x2
layer11:Convolution, 3x3, 512
layer12:Convolution, 3x3, 512
layer13:Convolution, 3x3, 512
layer14:Max-pooling, 2x2
layer15:Convolution, 3x3, 512
layer16:Convolution, 3x3, 512
layer17:Convolution, 3x3, 512
layer18:Max-pooling, 2x2
Transfer learning
layer19:Fully connected, 128
layer20:ReLU
layer21:BN
layer22:Fully connected, 2
layer23:softmax
Retrained from scratch
<C class
<M class
≥C class
≥M class

Figure 1. Flowchart of our analysis, including data preprocessing and the modified CNN model with transfer learning.

training data set.

It can be seen that the number of No-flare/C-level magnetogram samples is much larger than that of the M/X-level magnetogram samples. This can lead to serious class imbalance, which is a major problem for most machine learning algorithms. It also reflects the fact that most ARs do not yield major flares in most 24-hr periods. In our work, we try to alleviate class imbalance with undersampling and data augmentation schemes. The data set is undersampled by randomly selecting about 2 No-flare/C-level samples per 10 samples. We use data augmentation schemes to artificially increase the number of M/X-level samples by rotating and reflecting images, which makes the model invariant to these geometric transformations. Note that the number of No-flare/C-level ARs is much larger than that of M/X-level ARs, and the number of ARs in all four levels does not increase by data augmentation schemes, although we undersample and augment magnetogram samples. Cross-validation (CV) is a standard method for evaluating the validity of a model. There are several CV methods, such as shuffle-and-split CV, K-fold CV, etc. (Nishizuka et al. 2017). Because the number of ARs in the four levels is extremely imbalanced, we utilize the shuffle-and-split CV method in our study. Taking the first data set as an example, we randomly shuffle the AR numbers in different levels of No-flare/C/M/X, and then divide the AR numbers at a ratio of about 20%:80% and assigned to the testing data and training data sets, respectively. Repeating the above operation, we obtain 10 separate CV data sets composed of training and testing data sets to construct our database. The advantage of this method is that in each of the 10 data sets, not only the samples in the testing data set do not appear in the training data set, but also the ARs in the testing data set do not appear in the training data set. These 10 separate training and test data sets are used for our model. Table 1 provides the information for the 10 separate data sets. Since CNNs require all input images to be fixed in size, we warp the input images from the data set to 128×128 pixels following a similar method of Huang et al. (2018).

3. Method

CNNs usually utilize multiple multilayer perceptrons to learn both machine learning features and classification weights through supervised learning (Goodfellow et al. 2016). Apart from the first and the final layers of a network, each layer regards the output of the previous layer as input, and forwards its own output to the next layer. The input to the first layer are sample images, and the output of the final layer is used for class prediction. The layers between the input and output layer are known as hidden layers and mainly include convolutional layer, pool layer, and fully connected layer. The functionality of these layers is as follows: (1) The convolutional layer performs a convolution operation between the input of the layer and a number of weights called kernel or filter. The specified number of filters is known as the depth of the layer. Each filter is applied to the entire input image. (2) The pooling layer returns a subsampled version of the input data. The input image is divided into small regions, and a representative value of the pixels is generated in each region. 'Average pooling' computes the average of pixels in a region, while 'max pooling' uses the value of the pixel with the highest intensity in the region. (3) The fully connected layer is used to implement classification within the network. Neurons within this layer are connected to all activated outputs in the previous layer. In general, a common design for CNNs is comprised of a mixture of convolutional and pooling layers, followed by fully connected layers. In addition to the standard network layers described above, rectified linear unit (ReLU; Nair & Hinton 2010) and batch normalization (BN) layers (Ioffe & Szegedy 2015) have been developed to accelerate the training process and regularize the network; they are usually added after the convolutional layer or fully connected layer.

Table 3
Flare prediction results (within 24 hr) of our modified CNN model with transfer learning and comparison with previous studies.

Metric	Model	≥C Class	≥M Class
Recall	This work	0.874±0.044	0.749±0.101
	Huang et al. (2018)	0.726	0.850
	Park et al. (2018)	0.85	–
	Nishizuka et al. (2018)	0.809	0.947
	Bloomfield et al. (2012)	0.753	0.704
Precision	This work	0.851±0.034	0.843±0.052
	Huang et al. (2018)	0.352	0.101
	Park et al. (2018)	–	–
	Nishizuka et al. (2018)	0.529	0.182
	Bloomfield et al. (2012)	0.351	0.146
Accuracy	This work	0.810±0.029	0.844±0.036
	Huang et al. (2018)	0.756	0.813
	Park et al. (2018)	0.82	–
	Nishizuka et al. (2018)	0.822	0.860
	Bloomfield et al. (2012)	0.712	0.830
FAR	This work	0.149±0.034	0.157±0.052
	Huang et al. (2018)	0.648	0.899
	Park et al. (2018)	0.17	–
	Nishizuka et al. (2018)	0.471	0.818
	Bloomfield et al. (2012)	0.649	0.854
HSS	This work	0.537±0.048	0.657±0.070
	Huang et al. (2018)	0.339	0.143
	Park et al. (2018)	0.63	–
	Nishizuka et al. (2018)	0.528	0.265
	Bloomfield et al. (2012)	0.315	0.190
TSS	This work	0.526±0.052	0.640±0.075
	Huang et al. (2018)	0.487	0.662
	Park et al. (2018)	0.63	–
	Nishizuka et al. (2018)	0.634	0.804
	Bloomfield et al. (2012)	0.456	0.539

For Huang et al. (2018), we calculate Precision, FAR, and Accuracy values from the contingency table in their Table 4. For Nishizuka et al. (2018), we use their Table 3 and compute the Precision score from their Figure 5. For Bloomfield et al. (2012), we use their Table 4 and retrieve the contingency table from the machine-readable data they provide online.

In transfer learning, knowledge acquired in one or more source tasks is transferred and used to improve the learning of a related target task. The application of transfer learning for classification requires consideration of which layers to train and which layers to freeze. CNNs trained on natural image data set exhibit a curious phenomenon in common. The initial few layers of the neural network learn the generic features, while the last few layers learn the complex features (Aniyan & Thorat 2017; Tang et al. 2019). Therefore, based on CNNs with transfer learning, it is possible to retrain the network to work on our solar images by retraining only the last few layers and freezing the initial layers. CNNs have brought a series of breakthroughs in image classification, and are getting deeper and deeper.

In this work, we use a modified version of the popular VGG-16 network as our CNN model for flare prediction. The flowchart of the whole process, including data preprocessing and modified CNN model with transfer learning, is shown in Figure 1. The data preprocessing is introduced in Section 2 and involves steps such as excluding samples with multiple ARs, undersampling and data augmentation, shuffle-and-split CV, and warping samples to 128×128 pixels. The modified CNN model consists of the first 18 layers (13 convolutional layers and 5 pooling layers) from the VGG-16 network (Simonyan & Zisserman 2015), and the last 5 layers added by us. The VGG-16 network pretrained on the ImageNet (Deng et al. 2009) cannot be applied to predict solar flares directly. In our model, we use the first 18 layers (13 convolutional layers and 5 pooling layers) from the VGG-16 network as the basic architecture of our model, and add two fully connected layers, a BN layer, and a ReLU activation function. The final layer of our model is a softmax layer computing the probability scores for the two classes. In addition, we modify the model to work on single-channel images because the original network design is used to process three-channel color images. The complete architecture of our model is summarized in Table 2. Constrained by the insufficient number of ARs in our data set, we apply the modified CNN model in conjunction with transfer learning as follows. (1) We inherit the architecture of the first 18 layers including 13 convolutional layers and 5 pooling layers and the weights of these convolutional layers from the VGG-16 network in the modified CNN model, which is called transfer learning. Thus the filter weights of 13 convolutional layers pretrained by the ImageNet are unchanged during training. The basic features are obtained by the first few convolutional layers. (2) We retrain the last few layers in our model from scratch including two fully connected layers and one BN layer. The weights of the last few layers are changed during training. The complex features are obtained by the last few layers in our model. (3) There are different methods of transfer learning for various applications. The transfer learning method in our work is similar to Method B of Tang et al. (2019). We both inherit the architecture and weights of the first few convolutional layers, but retrain the fully connected layers and BN layer from scratch.

The modified CNN model with transfer learning is trained and tested on the Keras Deep Learning framework (Chollet et al. 2015) using the TensorFlow backend (Abadi et al. 2015) in the Python programming language. A minibatch strategy (Goodfellow et al. 2016) is used to accelerate the training error convergence. Here, the minibatch size corresponds to the number of training samples in one forward and backward pass. To maximize the prediction accuracy, the model is trained to minimize a loss function. Because the data set is imbalanced where the number of negative ARs outnumbers that of positive ARs, we use a weighted cross-entropy loss function for optimizing model parameters during training (Hinton & Salakhutdinov 2006). The loss function is calculated like

$$J = \sum_{n=1}^{N}\sum_{k=1}^{K} \omega_k y_{nk} log(\hat{y_{nk}}) \ , \tag{1}$$

$$\omega_k = len(AR_k) len(sample_k) \beta_k \ , \tag{2}$$

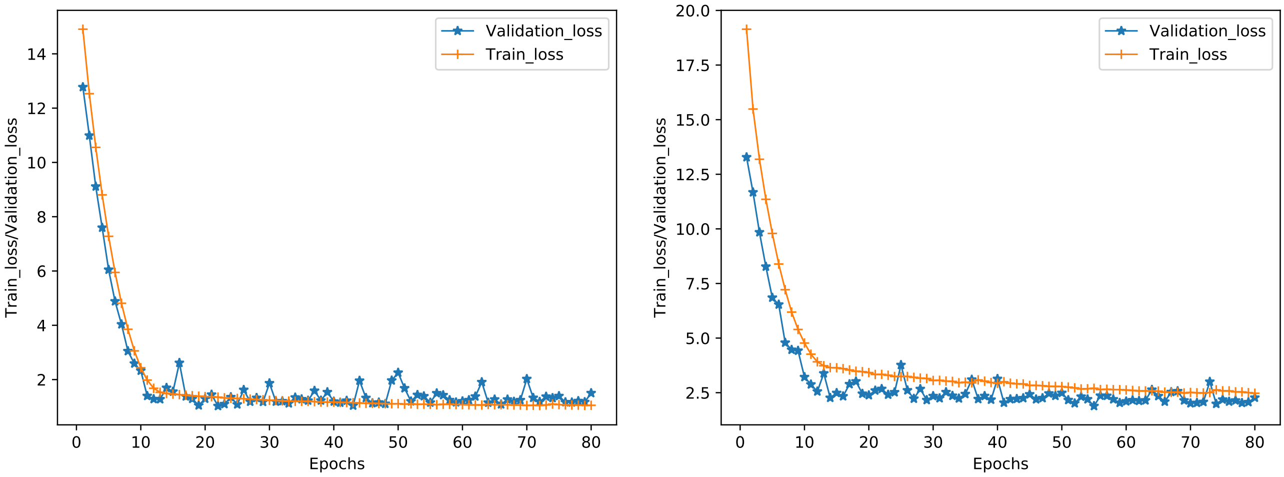

Figure 2. Learning curves showing the average training and validation loss per epoch for the modified CNN model with transfer learning. *Left:* Average training and validation loss per epoch for the model used to predict ≥C-class flares. *Right:* Average training and validation loss per epoch for the model used to predict ≥M-class flares.

where N is the number of training samples in each minibatch size, K is the number of classes, $\hat{y_{nk}}$ and y_{nk} are the predicted output and the real output of the kth class during a forward pass respectively. ω_k denotes the weight of the kth class, $len(AR_k)$ is the number of AR of the kth class, $len(sample_k)$ is the size of the sample of the kth class, and β_k represents the optimized parameter of the kth class. The best learning rate is set to 0.0004, momentum is set to 0.6, and β_k in Equation (2) is set to (1, 0.8) for ≥C class, and (1, 40) for ≥M class flares respectively. In this work, we use a stochastic gradient descent (SGD; LeCun et al. 1998) as optimizer with a minibatch size of 16 for training. The modified CNN model is trained over 80 epochs, where one epoch means an entire training data set is passed through our model.

4. Results

The results of a binary classifier can be characterized by a confusion matrix, also called a contingency table. Samples correctly forecasted as positive are called true positives (TP), while samples wrongly predicted as negative are called false negatives (FN). On the other hand, samples correctly forecasted as negative are called true negatives (TN), while samples wrongly predicted as positive are called false positives (FP). From these four quantities, various well-known metrics are computed. The performance metrics used in this study include as following: recall = TP/[TP + FN], precision = TP/[TP + FP], accuracy = [TP + TN]/[TP + FP + TN + FN], false alarm ratio (FAR) = FP/[TP + FP], Heidke skill score (HSS; Heidke 1926), and the true skill statistics (TSS; Hanssen & Kuipers 1965). HSS and TSS are defined as

$$\mathrm{HSS} = \frac{2[(\mathrm{TP}\times\mathrm{TN})-(\mathrm{FN}\times\mathrm{FP})]}{(\mathrm{TP}+\mathrm{FN})(\mathrm{FN}+\mathrm{TN})+(\mathrm{TP}+\mathrm{FN})(\mathrm{FP}+\mathrm{TN})} \quad (3)$$

$$\mathrm{TSS} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} - \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} \,. \quad (4)$$

In our study, we use these six metrics to evaluate the quality of our flare forecast. The range of recall, precision, and accuracy is 0 to 1, with 1 representing perfect prediction. The range of FAR is from 0 to 1, with 0 representing perfect prediction. HSS ranges from $-\infty$ to 1, with 1 representing perfect prediction and less than 0 representing no skill. The TSS covers the range from −1 to 1, where 1 indicates that all predictions are correct, −1 stands for no correct predictions, and 0 indicates that the predictions are generated mainly by chance. All metrics except for the TSS are sensitive to the class imbalance ratio (Bobra & Couvidat 2015; Bloomfield et al. 2012). Because the TSS is unbiased with respect to the class imbalance ratio, we also follow the suggestion of Bloomfield et al. (2012) to use the TSS score as primary metric and others as secondary ones.

The modified CNN model with transfer learning is trained over 80 epochs with a minibatch size of 16 on the training data set. We use the testing data set to validate the model. The validation of the model is carried out during each epoch to keep track of the learning performance. Since 10 CV data sets are used to train and validate our model (see Section 2), the procedure of training and validating our model is repeated 10 times. The average results of training and validation loss per epoch for the modified CNN model with transfer learning are shown in Figure 2. It can be seen in Figure 2 that our model does not suffer from overfitting. The average results of training and validation loss decline with increasing numbers of epochs during the initial epochs of training, while they tend to saturate after 20 epochs for ≥C class and after 40 epochs for ≥M class flares, respectively. In order to obtain the results optimized for TSS, we choose the best TSS model from many models produced at every epoch when the model is trained on each of 10 CV data sets.

The prediction results of our modified CNN model with transfer learning are shown and compared with previous studies in Table 3. We choose previous models in Table 3 to compare with our model because they use chronological data sets ensuring ARs of testing data set and ARs of training data set are disjoint. Furthermore, we select the previous models by Huang et al. (2018) and Park et al. (2018), because they both use CNN models without manually engineered features extracted from the observational data. The model by Huang et al. (2018) is based on a classic CNN model including two convolutional layers with 64 11×11 filters and two max-pooling layers. The model by Park et al. (2018) is a kind of combination of GoogLeNet (Szegedy et al. 2014) and DenseNet (Huang et al. 2016). The model by Nishizuka et al. (2018) is based on a deep residual neural network, and that of Bloomfield et al. (2012) is based on statistics. Our model is based on a popular modified CNN model (VGG-16) with transfer learning, which is different from previous models.

As shown in Table 3, our model is trained and tested 10 times on 10 CV data sets to provide the means and standard deviations, while other models are trained and tested on single data set to provide their best values. Huang et al. (2018) and Park et al. (2018) also use 10-fold CV to provide the means and standard deviations of their prediction results. However their data set used for CV is randomly partitioned, which cannot prevent ARs in the testing phase from appearing in the training phase. It can be seen from Table 3 that the standard deviations of all metrics are small, indicating that our model robustly predicts both $\geq$C class and $\geq$M class flares. For $\geq$C-class flare prediction, we achieve TSS $= 0.526 \pm 0.052$, which is better than that of Huang et al. (2018) and Bloomfield et al. (2012), and lower than that of Park et al. (2018) and Nishizuka et al. (2018). For $\geq$M-class flare prediction, we achieve TSS $= 0.640 \pm 0.075$, which is comparable to that of Huang et al. (2018), better than that of Bloomfield et al. (2012), and lower than that of Nishizuka et al. (2018). The HSS value of our model is 0.657 ± 0.070, which is much better than that of previous models for $\geq$M-class events, while the HSS value is 0.537 ± 0.048, which is comparable to that of Park et al. (2018) and Nishizuka et al. (2018), and much better than that of Huang et al. (2018) and Bloomfield et al. (2012) for $\geq$C-class flares. The values for FAR and precision are much better than those of previous studies for both $\geq$C and $\geq$M-class events, and the FAR and precision scores are about 8 times better than those of previous models for $\geq$M-class flares. In addition, we obtain good scores in terms of recall and accuracy at the same time. In summary, the experiments demonstrate that the overall performance of our model is comparable to that of other methods.

5. Conclusions and Discussions

In this study, we apply a modified CNN model with transfer learning to predict whether an AR would yield a $\geq$C-class or $\geq$M-class flare within the next 24 hours. We collect 870 ARs and 136,134 LOS magnetogram samples provided by the SHARP from May 2010 to September 2018. Relying on these AR samples, we utilize the shuffle-and-split CV approach to build a database that contains 10 separate data sets. We split each of these 10 data sets into a training and a testing data set according to NOAA AR number. We emphasize the importance of a proper segregation process: not only the samples in the testing data set must not appear in the training data set, but also the ARs in the testing data set have to be disjoint from the ARs in the training data set. Previous studies focused on chronologically dividing their data sets into training and test data sets, or selecting randomly shuffled data sets with some similarities between training and testing data sets. In fact, the difference in establishing data sets may affect the performance comparisons, but it is difficult to build common data sets. In order to reduce the class imbalance, we also adopt undersampling and data augmentation schemes, which are usually used for data preprocessing in machine learning. Eventually the resulting data sets are used to train, validate and evaluate our model, and our model is not subject to overfitting.

The main results from this study are summarized as follows. Firstly, to the best of our knowledge, this is the first time that the CNN model in conjunction with transfer learning is used in solar physics to make binary class predictions for both $\geq$C-class and $\geq$M-class flares, without manually engineered features extracted from the observational data; previous studies solely utilize CNN methods, or apply classical machine learning or statistical methods which are different from ours. Secondly, our model achieves relatively high scores of TSS $= 0.640 \pm 0.075$ and 0.526 ± 0.052 for $\geq$M-class and $\geq$C-class prediction. We note that the performance of our model appears inferior to the highest TSS score of in Table 3 (Nishizuka et al. 2018). However, Nishizuka et al. (2018) only uses the best TSS value where the model is trained and tested on single data set. The best TSS score of our model among the 10 CV results is 0.757 (0.602, respectively) for $\geq$M- class ($\geq$C- class, respectively) prediction, which is roughly comparable to that of Nishizuka et al. (2018). The data set number for the best TSS score model is No. 4 for $\geq$M-class and No. 9 for $\geq$C-class events in Table 1, respectively. Thirdly, our model obtains fairly good scores in other the five metrics for both $\geq$C-class and $\geq$M-class flare prediction. In general, the performance of our forecasting model is comparable to that of previous models for binary class predictions.

Based on all our experiments, our modified CNN model with transfer learning is a valid method for both $\geq$C- and $\geq$M-class flare forecasting with reasonable prediction performance. At present, our model can make binary class predictions for both $\geq$C-class and $\geq$M-class flares. However, the model does not provide forecasts for $\geq$X-class flares because the number of both ARs and magnetogram samples for X-level events is much less than that of ARs and magnetogram samples for M/C/No-flare level events, which results in serious class imbalance and prevents training our model for $\geq$X-class flares. With the accumulation of observation data with high temporal and high spatial resolution, we will collect more ARs and magnetogram samples for X-level events to solve this problem. In this work, for recurring flares of ARs, following the method of Yuan et al. (2010) and Song et al. (2009), ARs are categorized into four levels according to the most powerful flare produced. We speculate this method could increase the prediction error of our model for recurring-flare forecasting. In future work, we will improve this method to make our model accurately predict recurring flares as well. In the near future, with the fast development of machine learning, we will begin to apply other deep learning models (e.g., Gaier & Ha 2019; Huang et al. 2016) with transfer learning to improve the prediction performance for solar flares. In addition, we plan to apply our model to solve other forecast problem (e.g., filament eruptions and CMEs) in solar physics.

ACKNOWLEDGMENTS

We acknowledge the referees for their great effort to review our paper and for providing us with useful comments. This work is supported by the National Natural Science Foundation of China (grants no. 11703009, 11803010), the Natural Science Foundation of Jiangsu Province, China (grant no. BK20170566), and the Natural Science Foundation of Jiangsu Province (19KJD520003). The data used here is courtesy of NASA/*SDO* and the HMI science team, as well as the *Geostationary Operational Environment Satellite* (*GOES*) team.

REFERENCES

Abadi, M., et al. 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, https://www.tensorflow.org

Abraham, S., Aniyan, A. K., Kembhavi, A. K., et al. 2018, Detection of Bars in Galaxies Using a Deep Convolutional Neural Network, MNRAS, 477, 894

Ahmed, O. W., Qahwaji, R., & Colak, T. 2013, Solar Flare Prediction Using Advanced Feature Extraction, Machine Learning, and Feature Selection, SoPh, 283, 157

Aniyan, A. K., & Thorat, K., 2017, Classifying Radio Galaxies with the Convolutional Neural Network, ApJS, 230, 20

Barnes, G., Leka, K. D., Schrijver, C. J., et al. 2016, A Comparison of Flare Forecasting Methods, I: Results from the "All-Clear" Workshop, ApJ, 829, 89

Bloomfield, D. S., Higgins, P. A., James McAteer, R. T., et al. 2012, Toward Reliable Benchmarking of Solar Flare Forecasting Methods, ApJL, 747, L41

Bobra, M. G., & Couvidat, S. 2015, Solar Flare Prediction Using SDO/HMI Vector Magnetic Field Data with a Machine-Learning Algorithm, ApJ, 798, 135

Bobra, M. G., Sun, X., Hoeksema, J. T., et al. 2014, The Helioseismic and Magnetic Imager (HMI) Vector Magnetic Field Pipeline: SHARPs – Space-Weather HMI Active Region Patches, SoPh, 289, 3549

Chollet, F., et al. 2015, Keras, https://github.com/keras-team/keras

Colak, T., & Qahwaji, R. 2009, Automated Solar Activity Prediction: A Hybrid Computer Platform Using Machine Learning and Solar Imaging for Automated Prediction of Solar Flares, SpWea, 7, S06001

Deng, J., Dong, W., Socher, R., et al. 2009, ImageNet: A Large-scale Hierarchical Image Database, IEEE Conference on Computer Vision and Pattern Recognition, 248, 255

Florios, K., Kontogiannis, I., Park, S. H., et al. 2018, Forecasting Solar Flares Using Magnetogram-based Predictors and Machine Learning, SoPh, 293, 28

Gaier, A., & Ha, W. 2019, Weight Agnostic Neural Networks, arXiv:1906.04358

Goodfellow I., Bengio Y., & Courville A. 2016, Deep Learning (Cambridge, MA: MIT Press)

Guerra, J. A., Pulkkinen, A., & Uritsky, V. M. 2015, Ensemble Forecasting of Major Solar Flares: First Results, SpWea, 13, 626

Hanssen, A. W., & Kuipers, W. J. A. 1965, On the Relationship between the Frequency of Rain and Various Meteorological Parameters, Meded. Verh., 81, 2

Heidke, P. 1926, Berechnung des Erfolges und der Güte der Windstärkevorhersagen im Sturmwarnungdienst (Calculation of the Success and Goodness of Strong Wind Forecasts in the Storm Warning Service), Geogr. Ann. Stockholm, 8, 301

Hinton, G. E., & Salakhutdinov, R. R. 2006, Reducing the Dimensionality of Data with Neural Networks, Science, 313, 504

Huang, G., Liu, Z., van der Maaten, L., et al. 2016, Densely Connected Convolutional Networks, arXiv:1608.06993

Huang, X., Wang, H., Xu, L., et al. 2018, Deep Learning Based Solar Flare Forecasting Model. I. Results for Line-of-sight Magnetograms, ApJ, 856, 7

Huang, X., Yu, D. R., Hu, Q. H., et al. 2010, Short-Term Solar Flare Prediction Using Predictor Teams, SoPh, 263, 175

Ioffe, S., & Szegedy, C. 2015, Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, Proc. 32nd Int. Conf. Machine Learning (ICML'15), 37, 448

LeCun, Y., Bengio, Y., & Hinton, G. 2015, Deep Learning, Nature, 521, 436

LeCun, Y., Bottou, L., Orr, G., & Müller, K. 1998, Efficient BackProp, Neural Networks: Tricks of the Trade (Berlin: Springer), 9, 48

Li, R., & Zhu, J. 2013, Solar Flare Forecasting Based on Sequential Sunspot Data, RAA, 13, 1118

Liu, C., Deng, N., Wang, J. T. L., et al. 2017, Predicting Solar Flares Using SDO/HMI Vector Magnetic Data Products and the Random Forest Algorithm, ApJ, 843, 104

Mason, J. P., & Hoeksema, J. T. 2010, Testing Automated Solar Flare Forecasting with 13 Years of Michelson Doppler Imager Magnetograms, ApJ, 723, 634

Nair, V., & Hinton, G. E. 2010, Rectified Linear Units Improve Restricted Boltzmann Machines, Proc. 27th Int. Conf. on Machine Learning (ICML-10) (Madison, WI: Omnipress), 807

Nishizuka, N., Sugiura, K., Kubo, Y., et al. 2017, Solar Flare Prediction Model with Three Machine-learning Algorithms using Ultraviolet Brightening and Vector Magnetograms, ApJ, 835, 156

Nishizuka, N., Sugiura, K., Kubo, Y., et al. 2018, Deep Flare Net (DeFN) Model for Solar Flare Prediction, ApJ, 858, 113

Park, E., Moon, Y. J., Shin, S., et al. 2018, Application of the Deep Convolutional Neural Network to the Forecast of Solar Flare Occurrence Using Full-disk Solar Magnetograms, ApJ, 869, 91

Pesnell, W. D., Thompson, B. J., & Chamberlin, P. C. 2012, The Solar Dynamics Observatory (SDO), SoPh, 275, 3

Priest, E. R., & Forbes, T. G. 2002, The Magnetic Nature of Solar Flares, A&ARv, 10, 313

Qahwaji, R., & Colak, T. 2007, Automatic Short-Term Solar Flare Prediction Using Machine Learning and Sunspot Associations, SoPh, 241, 195

Razavian, A. S., Azizpour, H., Sullivan, J., et al. 2014, CNN Features Off-the-shelf: an Astounding Baseline for Recognition, arXiv:1403.6382

Sadykov, V. M., & Kosovichev, A. G. 2017, Relationships between Characteristics of the Line-of-sight Magnetic Field and Solar Flare Forecasts, ApJ, 849, 148

Schou, J., Scherrer, P. H., Bush, R. I., et al. 2012, Design and Ground Calibration of the Helioseismic and Magnetic Imager (HMI) Instrument on the Solar Dynamics Observatory (SDO), SoPh, 275, 229

Shibata, K., & Magara, T. 2011, Solar Flares: Magnetohydrodynamic Processes, LRSP, 8, 6

Simonyan, K., & Zisserman, A. 2015, Very Deep Convolutional Networks for Large-Scale Image Recognition, ICLR 2015, arXiv:1409.1556

Song, H., Tan, C., Jing, J., et al. 2009, Statistical Assessment of Photospheric Magnetic Features in Imminent Solar Flare Predictions, SoPh, 254, 101

Szegedy, C., Liu, W., Jia, Y., et al. 2014, Going Deeper with Convolutions, arXiv:1409.4842

Tang, H. M., Scaife, A. M. M., & Leahy, J. P. 2019, Transfer Learning for Radio Galaxy Classification, arXiv:1903.11921

Yu, D. R., Huang, X., Wang, H. N., et al. 2010, Short-term Solar Flare Level Prediction Using a Bayesian Network Approach, ApJ, 710, 869

Yuan, Y., Shih, F. Y., Jing, J., et al. 2010, Automated Flare Forecasting Using a Statistical Learning Technique, RAA, 10, 785