Matsuoka
Progress in Earth and Planetary Science (2021) 8:68
https://doi.org/10.1186/s40645-021-00459-y

Progress in Earth and Planetary Science

METHODOLOGY

Open Access

Classification of imbalanced cloud image data using deep neural networks: performance improvement through a data science competition

Daisuke Matsuoka*

Abstract

Image data classification using machine learning is an effective method for detecting atmospheric phenomena. However, extreme weather events with a small number of cases cause a decrease in classification prediction accuracy owing to the imbalance in data between the target class and the other classes. To build a highly accurate classification model, I held a data analysis competition to determine the best classification performance for two classes of cloud image data, specifically tropical cyclones including precursors and other classes. For the top models in the competition, minority data oversampling, majority data undersampling, ensemble learning, deep layer neural networks, and cost-effective loss functions were used to improve the classification performance of the imbalanced data. In particular, the best model of 209 submissions succeeded in improving the classification capability by 65.4% over similar conventional methods in a measure of the low false alarm ratio.

Keywords: Machine learning, Deep learning, Binary classification, Image recognition, Tropical cyclone

1 Introduction

In recent years, deep learning, which is machine learning using multilayered neural networks, has attracted much attention in various research and industrial fields as a technology that greatly exceeds the performance of conventional methods. In particular, deep convolutional neural networks, which are specialized for image recognition, are highly efficient at extracting spatial feature patterns (Krizhevsky et al. 2012). One of the simplest tasks that a convolutional neural network can perform is the classification of image categories. In atmospheric science, image classification has been applied to detect hurricanes, fronts, and atmospheric rivers (Liu et al. 2016), tropical cyclones (Kim et al. 2017), and precursors of tropical cyclones (Matsuoka et al. 2018) and to estimate hurricane intensity (Pradhan et al. 2018), among other applications.

Although previous studies have reported interesting results by using classifications, the imbalance in the amount of data between classes is still an issue. Target phenomena or structures such as hurricanes and tropical cyclones occur relatively infrequently, whereas other patterns are countless. For example, Matsuoka et al. (2018) classified 50,000 positive and 1,000,000 negative example images, with a balance of 20 times. This class imbalance is known to cause a decrease in classification performance (e.g. Sun et al. 2009).

Such problems in classifying imbalanced data are called the imbalanced learning problem, and several methods have been proposed in various fields to solve this problem (e.g. He and Garcia 2009; Leevy et al. 2018). These methods can be categorized into data-level approaches, such as training data sampling and feature selection,

*Correspondence: daisuke@jamstec.go.jp
Research Institute for Value-Added-Information Generation (VAiG), JAMSTEC, 3173-25 Showa-machi, Kanazawa-ku, Yokohama 236-0001, Japan

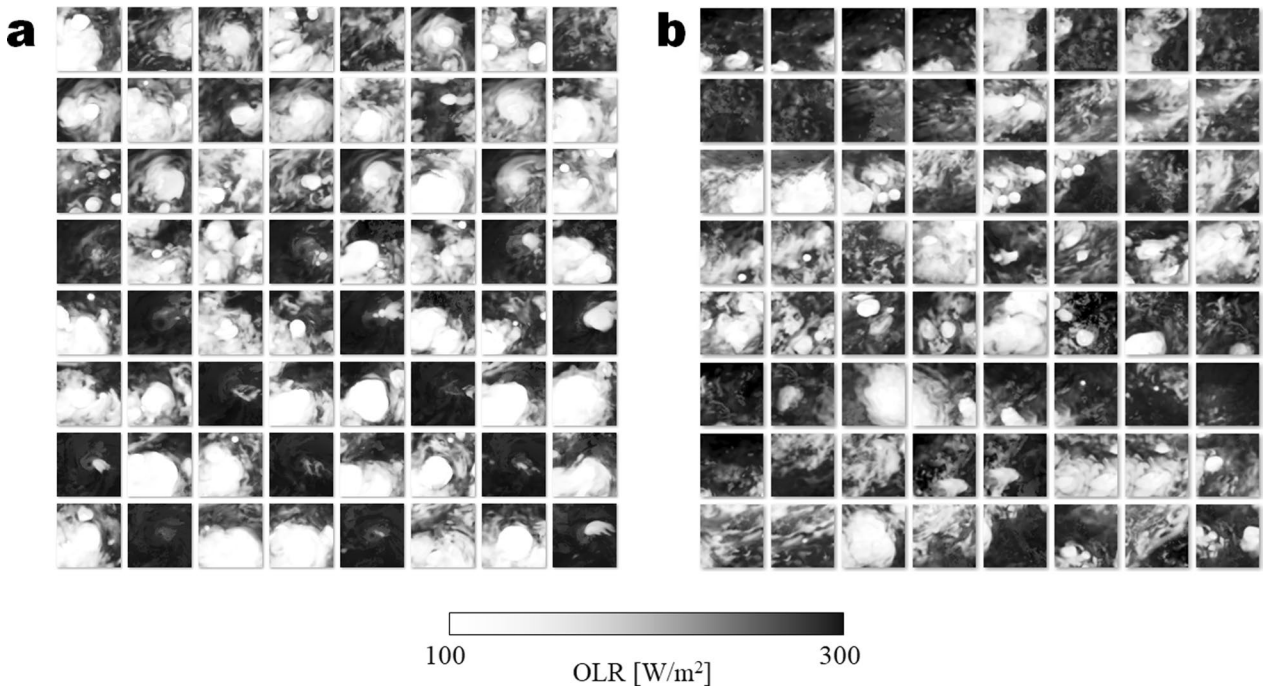

Fig. 1 Examples of training data. **a** TCs and **b** nonTCs (positive and negative examples, respectively)

and algorithm-level approaches, such as cost-sensitive learning and ensemble learning. Furthermore, several derivative methods have been proposed. However, each method often requires empirical judgment and intuition, with domain-specific dependencies. Above all, it is impractical in terms of computational and human costs to try all method combinations, especially for deep learning, which deals with large amounts of data.

To solve these problems and obtain the best performing model, we opened the data used in our previous study (Matsuoka et al. 2018) and held a data science competition event to determine the classification performance of tropical cyclones. The competition was held over a 2-month period from August to October 2018, with the participation of over 200 scientists and engineers from various backgrounds, such as medicine, physics, economics, computer sciences, and atmospheric science. The models proposed by the winners of the competition achieved classification performances that far exceeded those of Matsuoka et al. (2018) based on deep convolutional neural networks. This paper presents the methods used in the top models and discusses effective techniques for classifying imbalanced image data in atmospheric science.

2 Data set and evaluation metrics

In this section, the details of the data used in the competition and the metric used to evaluate the classification performance are described.

2.1 Data set

In the competition, tropical cyclones, including precursors (Fig. 1a), and other outgoing longwave radiation (OLR) data (Fig. 1b) were used, as per a previous study. The former data (hereafter TCs or positive examples) were derived from the 30-year climate experiment data of the cloud-resolving model NICAM (Kodama et al. 2015) using the tropical cyclone detection algorithm (Yamada et al. 2017; Nakano et al. 2015; Sugi et al. 2002). The latter data (hereafter referred to as nonTCs or negative examples) were assumed to be non-tropical cyclones in the past, present, and future. The OLR data were normalized from 0 to 1 in the range of 300.0–100.0 W/m^2, and the single-precision real values were readable TIFF format image files. The number of pixels in the image was set to 64×64 (approximately 1000 km in actual scale).

The images released for the competition were divided into two categories, training data used to construct classification models and test data used for evaluation, based on the order of the time series for practical situations. The number of images for the training data was set to 2,244,223 (for 15 years from 1984 to 1998), and the number of images for the test data was set to

299,135 (for 2 years from 1999 to 2000). For reference, the plots of the first and second principal components of the training and test data, dimensionality reduced using principal component analysis, are shown in Additional file 1: Fig. S1. In both TCs and nonTCs, the plots of the training and test data almost overlap, indicating that this is an appropriate setting for the problem. Although the amount of data was different from that of Matsuoka et al. (2018), the balance of positive and negative examples was almost the same, with a ratio of approximately 1:20. The training data were opened to the user with a correct label (TC or nonTC) for supervised learning. Although the test data were not labeled, participants could check the tentative evaluation results through the submission system for the duration of the event. Since the final evaluation in the competition was performed using a portion of the test data, the tentative evaluation results for the test data and the final evaluation did not necessarily match.

2.2 Evaluation metrics

In this competition, *Conditional precision* as an evaluation metric for classifying imbalanced cloud image data was used as follows:

$$Conditional\ precision = \begin{cases} Precision\ (Recall \geq 0.79) \\ 0\ (otherwise) \end{cases} \quad (1)$$

Here, *Recall* is a measure of the correctness of the classification in the correct label, also called the *hit ratio*. *Precision* is a measure of the correctness of the inference result, and 1-*Precision* indicates the *false alarm ratio*. The *Precision* and *Recall* are defined by the following equations:

$$Precision = TP/(TP + FP), Precision = TP/(TP + FP), \quad (2)$$

$$Recall = TP/(TP + FN) Recall = TP/(TP + FN) \quad (3)$$

where *TP* (true positive) is the number of cases in which the correct answer was correctly predicted as a positive example, *FP* (false positive) is the number of cases in which the correct answer was incorrectly predicted as a positive example, and *FN* (false negative) is the number of cases in which the correct answer was incorrectly predicted as a negative example. There is a trade-off between the *Precision* and *Recall* reproduction and the fit rate, and it is possible to adjust the balance between them by the parameter setting. In the results of Matsuoka et al. (2018), when *Recall* was set to approximately 80%, the decrease in *Precision* became an issue. Therefore, in this competition, we used *Precision* as an evaluation metric when the *Recall* was approximately 80% or higher.

Of note that, with the aforementioned evaluation metric, even if the *Recall* is much higher than 0.79, it is not properly evaluated as the goodness of the model. To evaluate the comprehensive performance, it is necessary to show the trade-off between *Precision* and *Recall*. The *Precision-Recall* (*P-R*) curve, the plot of *Precision* (y-axis) and *Recall* (x-axis) for different thresholds of a classifier, is often used to show their trade-off.

3 Methods/experimental

In this section, the methods and their combinations used in the best model of the competition are described to improve the classification performance of imbalanced data.

3.1 Convolutional neural network

The architecture of a convolutional neural network (CNN) generally consists of convolutional, pooling, and fully connected layers, as shown in Fig. 2a (LeCun and Bengio 1995). The convolutional layer extracts spatial patterns, called feature maps, such as edges and gradients, by applying convolutional filters to the input data. In the example of the first layer in Fig. 2a, N feature maps are obtained for the input image by using N convolution filters with a window size of $M \times M$. The output data of the convolutional layer are transferred as the input data of the next layer through a nonlinear function, called the activation function. The pooling layer compresses the dimensionality of the input image, making it robust to horizontal misalignment. The fully connected layer combines the extracted features into one dimension and converts them into values for each output class using a function. At this time, the softmax function is often used to normalize the output values to a probability for each class (Goodfellow et al. 2016). Then, when the normalized value for a positive class exceeds a certain threshold, the model predicts that it is a positive example. In the output layer, the class 1 for a positive example and 0 for a negative class are often used in binary classification. To reduce the error between the output class of the CNN and the correct class (ground truth), the parameters of the CNN such as weights and biases are updated such that the error function (called loss function) is minimized. In this procedure, called backpropagation, the gradient of the loss function for the weights and biases is calculated for the given training data, and the gradient is propagated from the output layer to the input layer (Rumelhart et al. 1986). Here, the mini-batch gradient decent is typically used to update the weights for several pieces of training data together (called a mini-batch).

In models using neural networks, whereas deepening the layers generally improves the ability to represent

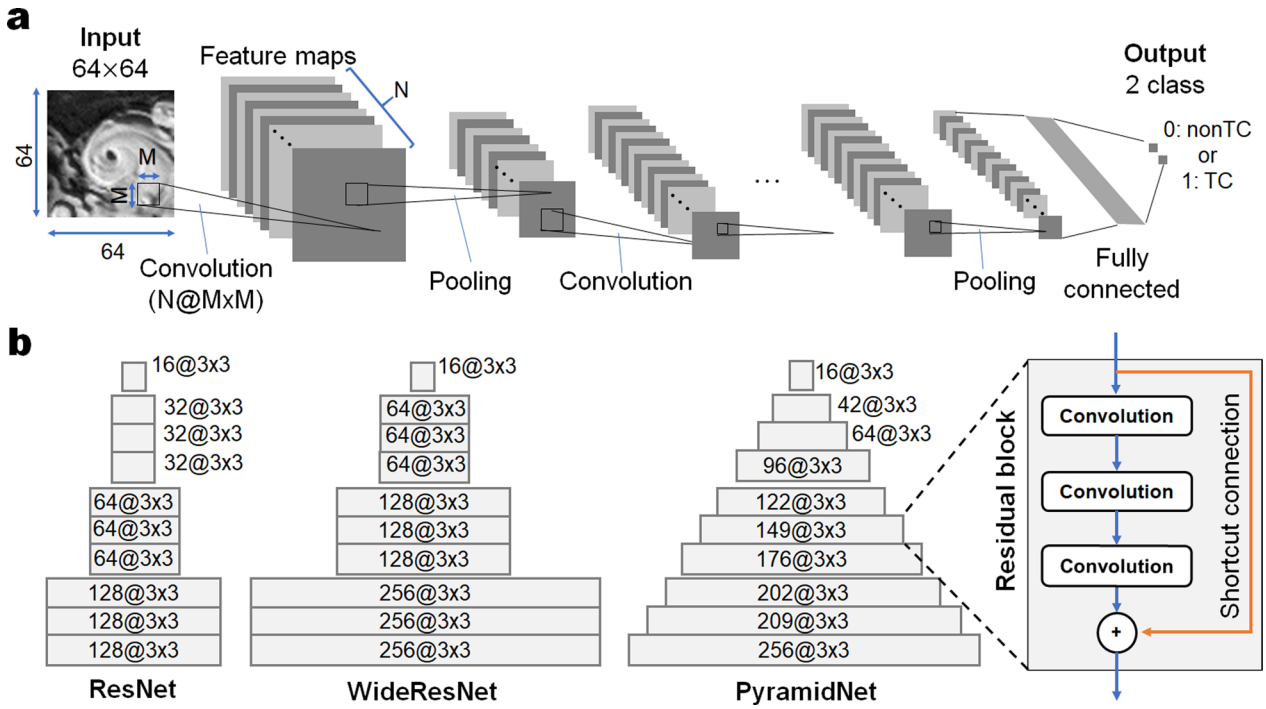

Fig. 2 Convolutional neural networks (CNNs) for binary classification. Basic architectures of **a** simple CNN, and **b** residual networks: ResNet (He et al. 2016), WideResNet (Zagoruyko and Komodakis 2016), and PyramidNet (Han et al. 2017)

features, the gradient disappears during back propagation away from the output layer (the vanishing gradient problem). To overcome this problem, there are several known methods such as using appropriate activation functions, special setting of initial weights, batch normalization, and a residual network (Hochreiter 1998; Hu et al. 2018). The residual network, as shown in Fig. 2b, introduces a shortcut connection that skips some layers, thereby directly transferring the gradient to the lower layers during back propagation and preventing gradient disappearance. Some of the known architectures using residual modules are ResNet (He et al. 2016), WideResNet (Zagoruyko and Komodakis 2016), and PyramidNet (Han et al. 2017), among others. These architectures are characterized by deep layers and a large number of convolutional filters.

3.2 Data sampling

To classify imbalanced data, duplicate samples from the minority class (oversampling) and selecting samples from the majority class (undersampling) are considered (Leevy et al. 2018). Data augmentation, a technique for minority class oversampling, is widely used to increase the image recognition performance for deep learning (Shorten and Khoshgoftaar 2019). As shown in Fig. 3, vertical flip, horizontal flip, random crop, and random rotation are common methods used in image recognition. Because the size of the data as a result of random crop and random rotation will be smaller than the original image, there is a method called padding, in which pixels are filled outside the original image to make the size equal to the original image. Cutout (DeVries and Taylor 2017) and random erasing (Zhong et al. 2020) were also used to mask partial regions in the image. Whereas cutout masks a square area at a random position with a value of 0, random erasing masks a rectangular area at a random position and size with a random value. For other data augmentation methods, such as image-to-image translation (e.g., Kim et al. 2019; Wei et al. 2020), refer to the survey paper (Shorten and Khoshgoftaar 2019). Many of these methods have also been implemented in deep learning frameworks such as Keras and PyTorch, and users can utilize them with simple functions.

Next, by undersampling the majority of class data, we can reduce the amount of training data that are easy to classify. In this study, we introduce a method to sample a certain number of misclassified negative examples (false positives) into the training data, as shown in Fig. 4. The size of the mini-batch used for training is n, and the number of positive and negative examples is $n/2$. Of the mini-batches, the positive example data are sampled randomly, but a certain percentage (X) of the negative example data contain cases that were misclassified by the CNN (hard

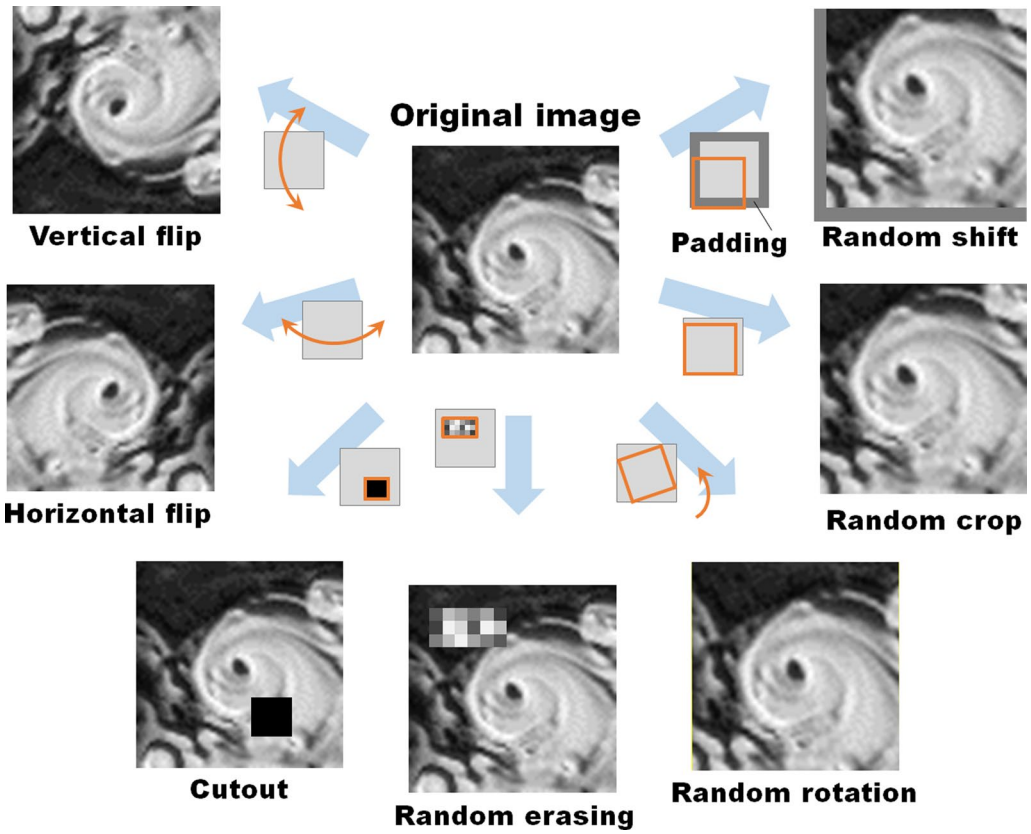

Fig. 3 Data augmentation for cloud image: vertical flip, horizontal flip, cutout, random erasing, random rotation, random crop, and random shift

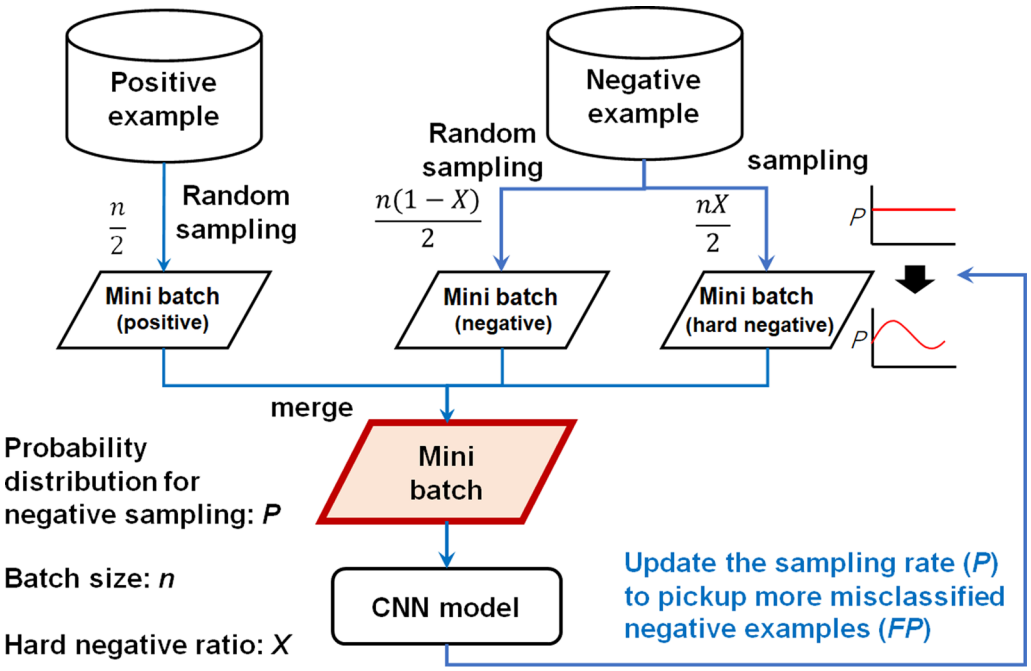

Fig. 4 Undersampling of the majority class using hard negative mining

negative example). Here, X is the hard-negative ratio. The probability distribution (P) is updated such that the sampling probability of misclassified negative examples is higher for mini-batches with a lower classification accuracy of the CNN model. A new mini-batch is generated by sampling the data based on the updated probability distribution. By repeating the series of processes, the number of negative examples that are easy to classify from the training data is reduced.

3.3 Ensemble learning

In the present study, we used ensemble learning, which attempts to improve accuracy by combining multiple training models. One of the most basic and powerful methods of ensemble learning is bagging (Breiman 1996), which uses multiple models trained with different randomly sampled data. In bagging, the generalization error is stochastically reduced by using the average result of the outputs from multiple models with different properties. The method proposed in this study is based on Bagging and consists of five different classification models as shown in Fig. 5. The five models (Models A–E) are built using different training data of different hard negative ratios ($X = 0.05$, 0.2, 0.35, 0.5, and 0.65) instead of random sampling. In the training phase, the amount of training data is increased fivefold by using four types of data augmentation (vertical flip, horizontal flip, random shift, and cutout). The final output is determined by taking the weighted average of the output results of the five different models. The weights of the output of each model are calculated using Bayesian optimization during the training phase.

In the test phase for untrained data (Fig. 5b), the final output is determined by the weighted average of the outputs of the five trained models for one input image. The output of each model is a simple average of the output values for the five input images that were increased by data augmentation. This data augmentation during the test phase is called test time augmentation (TTA) (Simonyan and Zisserman 2015). The proposed method is a hybrid model of the TTA and ensemble learning. To determine the final output result, soft voting is used as the average of the final layer outputs, whereas hard voting is used for the majority vote of multiple results (Kabari and Onwuka 2019; Leon et al. 2017).

4 Results and discussion

In this section, we show the classification performance of all submissions, including the top model of the competition described in the previous section. Insights into the methods used in the top four models from both technical and meteorological perspectives are also discussed.

4.1 Classification performance

The final evaluation results of the 209 models submitted by the participants in the competition are shown in Fig. 6. From the definition of *Conditional precision* shown in Eq. (1), although the final evaluation of the submission with $Recall < 0.79$ (represented by a blue triangle in the figure) was zero, both *Precision* and *Recall* values are shown for reference. Because there is a trade-off between the two, the results were concentrated around $Recall = 0.8$ (represented by a dotted line in the figure) with a small margin to clear the condition of $Recall \geq\geq 0.79$. If other *Recall* values were set as the threshold, the results of many models would gather around that threshold in order to get a high ranking. For the first-ranked model, $Precision = 0.6236$ and $Recall = 0.8062$, which are much higher than the results ($Precision = 0.4005$ and $Recall = 0.8060$) presented by Matsuoka et al. (2018). The second- and third-ranked models also had a $Precision > 0.6$, with $Recall \geq\geq 0.79$. There were a few models around $Precision = 0.6$, where *Recall* was marginally less than 0.79, but with good enough performance, even though they received zero marks in the competition.

The *P-R* curves of the top four models are also shown in Fig. 6. The curves show that all models had a natural curve for all *Recall* values without overfitting around $Recall = 0.79$. The classification performance of each model can also be evaluated by the area under the *P-R* curve (*PR-AUC*). The *PR-AUC* represents both *Precision* and *Recall* for a classifier as a single score, which ranges from 0 to 1 (1 for a perfect model). The *PR-AUC* values of the first, second, third, and fourth ranked models were 0.8204, 0.8001, 0.7927, and 0.7838, respectively, which are in the same order as the final evaluation results.

Of the 8,883 positive (TCs) and 290,251 negative (nonTCs) examples, there were 1,135 false negatives and 2,133 false positive. Examples of false negatives and false positives that were misclassified for the test data by all top four models are shown in Fig. 7a, b, respectively. As shown in Fig. 7 and the histograms of the average OLR in each test dataset (Additional file 1: Fig. S2), there was a tendency to misclassify positive examples with few clouds and negative examples with many clouds. The detailed discussion of the quantitative differences between true positives and false negatives and between true negatives and false positives, is beyond the scope of this paper and was omitted.

4.2 Classification strategies

The methods used in the top four models, which were the most representative, are summarized in Table 1. First, to increase the accuracy, it is effective to use a deep and wide network or to increase the number of tuning trials with a shallow and narrow model. Matsuoka et al. (2018)

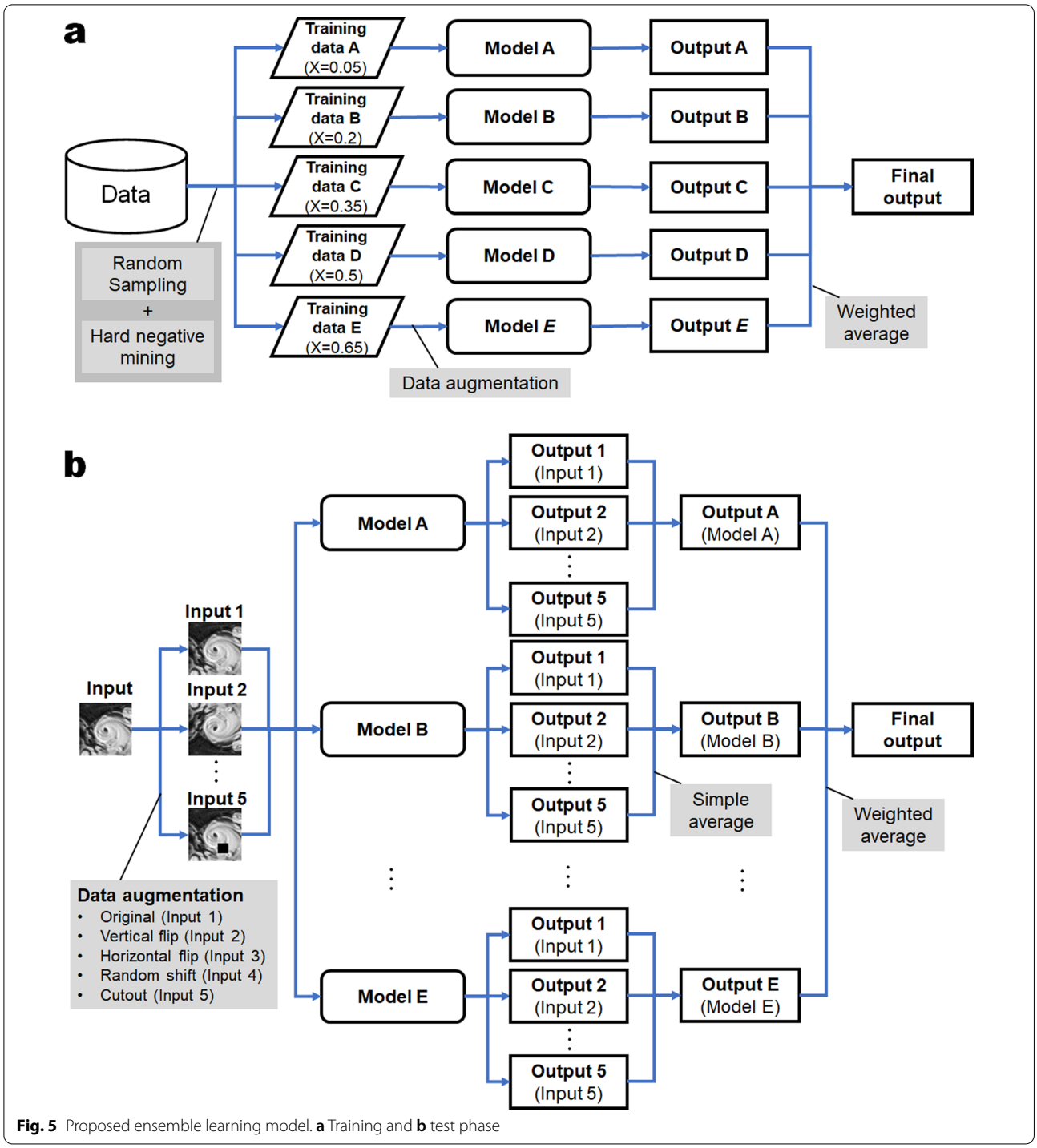

Fig. 5 Proposed ensemble learning model. **a** Training and **b** test phase

used a shallow and narrow network with four convolutional layers and up to 64 convolutional filters, whereas the top three models used a deep and wide network based on ResNet. The first-ranked model was a 110-layer PyramidNet, followed by a 10-layer WideResNet, and 26-layer ResNet, with the maximum number of convolution filters for any of the models of 256 (Fig. 2c). In addition, Shake-Shake regularization (Gastaldi 2017), a data augmentation method to the output (feature map) of the middle layer, was used in the third-ranked model. The fourth-ranked model was MobileNetV2 (Sandler et al. 2018), a pre-trained model with three convolutional layers. In

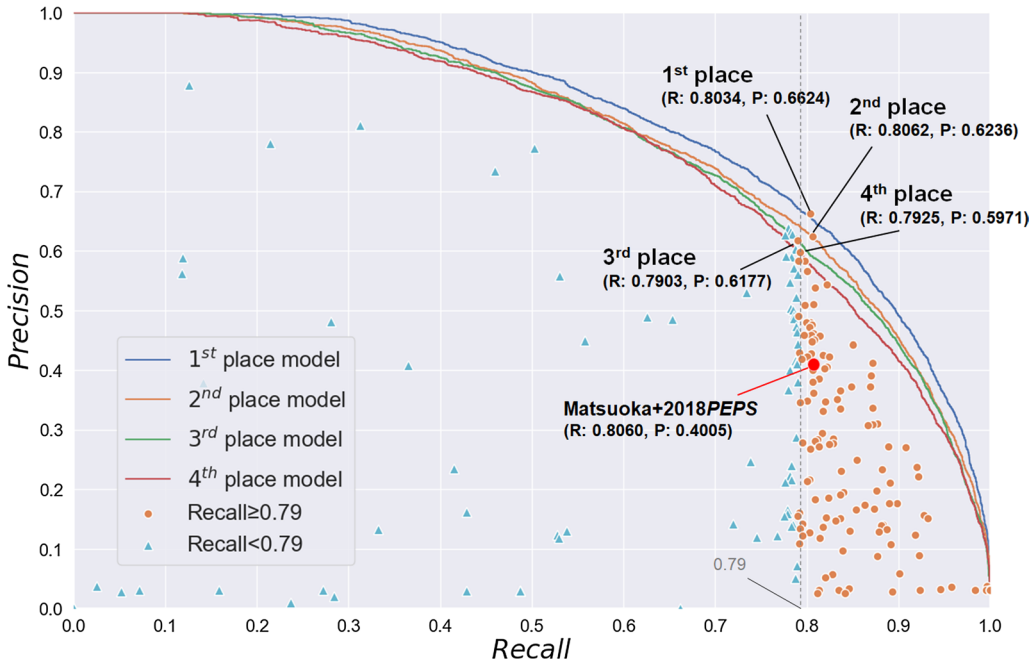

Fig. 6 The final evaluation results of the classification performance of the 209 submissions in the competition

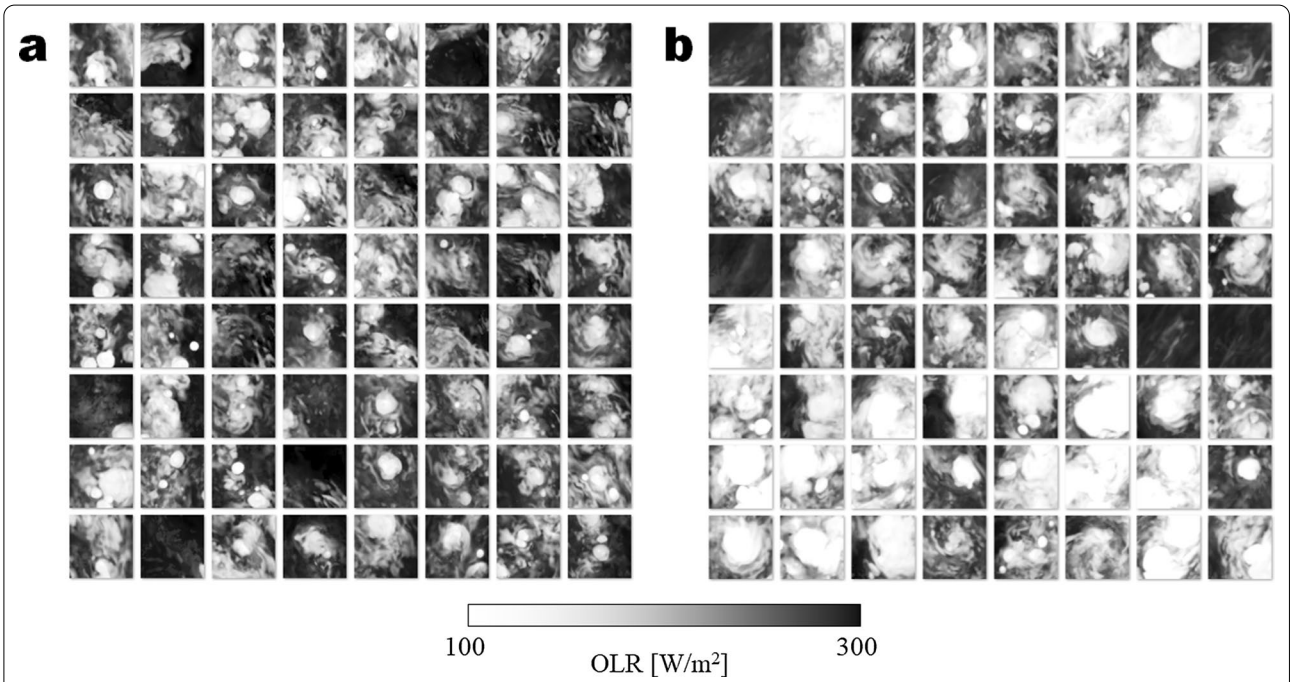

Fig. 7 Examples of **a** false negatives and **b** false positives for the test data by all top four models

general, the larger the number of parameters in a model, the higher the representational capability, but the more time required for training. As shown in Table 1, the first-ranked model had a relatively large number of parameters and a long learning time, and thus, the number of trials becomes small under the limited time and computational resource conditions. However, the second-ranked model had a relatively small number of parameters and a short

Table 1 Typical methods and elapsed times for each of the top models in the competition

	1st place model	2nd place model	3rd place model	4th place model	Matsuoka et al. (2018)
CNN architecture	PyramidNet	WideResNet	Shake-shake ResNet26	MobileNetV2	LeNet
Number of parameters	7.6 M	4.3 M	3.0 M	11.6 M	0.60 M
Preprocessing	Binarization	–	Downsampling (32 × 32)	Upsampling (96 × 96)	–
Oversampling (Data augmentation)	Vertical flip Horizontal flip Cutout Random shift	Cropping Random rotation Random erasing	Random crop + Padding Horizontal flip	Random rotation Mixup	–
Undersampling	Hard negative mining	Random sampling	Random sampling	Random sampling	Random sampling
Ensemble learning	5 models (Different hard negative ratio)	–	–	5 models (Different learning rate/preproccesing)	10 models (Different negative samples)
TTA	Same as training phase	5 Crop × 4 Rotation	10 Crop + Padding	–	–
Others	–	Focal loss	RReLU	–	–
Time for training	10 days	1 day	4 days	5 days	15 h
Time for test	10 h	1 h	1 h	5 h	4 h

training time, and thus, it is possible to repeat a lot of trial and error.

Data augmentation for minority classes during training and/or testing was effective in improving the classification performance. In all top three models, methods such as horizontal flip, vertical flip, random shift, cutout, and random rotation (four patterns of 0°, 90°, 180°, and 270°) were used (in the fourth-ranked model, only during training). In particular, the second-ranked model increased the original image 20-fold by combining five patterns of crops in the center and four corners (5 Crop) with four patterns of random rotation (and random erasing for each image during the training phase). In addition, MixUp (Zhang et al. 2018), which creates new images by superimposing positive and negative example images, was used in the fourth model. These data augmentation methods improved the accuracy of repeated experiments and were finally determined to be effective. However, the direction of the image has meteorological significance, as the direction of rotation of tropical cyclones differs between the northern and southern hemispheres, and the location of windy areas toward the direction of movement differs. In this sense, horizontal flip and random rotation might not be meteorologically valid, and their use should be considered. However, random cropping is a method that can be used for data augmentation without any inconsistency. Moreover, random erasing and cutout are also valuable methods because they provide regularization effects and are robust to noise. To summarize on data augmentation, all models had room for consideration in terms of physical consistency.

Ensemble learning, which combines multiple models, is also an effective method for improving classification accuracy. The first-ranked model used the weighted average of the output results from five models with different hard-negative ratios. The fourth-ranked model used the simple average of the output results of the five models with different preprocessing methods or learning rates. In addition, Matsuoka et al. (2018) used the weighted average of the output results from 10 models trained on different negative samples. These are all methods that can process multiple models in parallel. However, boosting (Freund and Schapire 1997), which is a sequential process to preferentially learn the results of the previous model's misclassification, is also known to be a powerful method and was used by Matsuoka et al. (2017). The second- and third-ranked models were trained and tested using only a single model. Whereas ensemble models are effective in improving classification accuracy, they also increase the difficulty of interpreting the models. It is also important to prioritize single models with high accuracy to obtain meteorological knowledge from feature maps in trained models. The second- and third-ranked models are refined as single models and are superior in interpretability to other ensemble models.

During classifications, cross-entropy is often used as a function to evaluate the error between the inference result and true value (Hinton et al. 1995). In the second-ranked model, focal loss (Lin et al. 2020) was used, which provided a large weighting for the loss of minority classes. Loss functions for imbalanced image classifications include weighted cross entropy loss (Hinton et al. 1995), Hamming loss (Frank and Hall 2001), and other classical functions such as ranking loss (Li et al. 2017) and sparseMax loss (Martins and Astudillo 2016). A randomized leaky rectified unit (RRELU) (Xu et al. 2015) was used in the third-place model as the activation function for transmitting the output of each layer to the next layer.

ReLU (Nair and Hinton 2010; Sun et al. 2014), which is widely used as the most common activation function, becomes zero when the input is negative and is constant otherwise. In RRELU, the function has a random slope when it is negative, which prevents overlearning. Other activation functions such as Leaky ReLU (Maas et al. 2013), parametric ReLU (He et al. 2015), and exponential linear units (Clevert et al. 2015) are also known.

For reference, the transition of the performance improvement and the key methods with the large contribution of the three top winners are shown in Additional file 1: Fig. S4 (no data for the fourth-place winner). Common to all three, the basic architecture of the CNN was chosen first and data augmentation was applied in the early to middle stages. In addition, the TTA and minor modifications such as parameter tuning and loss function selection were performed at the end of the event period. However, these would strongly depend on each individual's available computer resources, time resources, and experience. The universality in model improvement needs further discussion.

5 Conclusions

This paper proposes deep learning approaches to effectively classify imbalanced cloud image data with its differences of more than 20 times. To design a highly sophisticated classification model, a data science competition was held in which labeled images of tropical cyclones and other categories were made public. The results showed that the *Precision* of the top model exceeded 0.6 when the *Recall* was fixed at approximate 0.8. This successfully improved the performance by approximately 60% compared to that of Matsuoka et al. (2018) (*Precision* was approximately 0.4).

The higher-level models among the 209 submissions used deep-layered networks, as well as positive example data augmentation, negative example sampling, and ensemble learning as particularly effective methods. It is also important to select a loss and an activation function that considers the balance between classes. On the other hand, some of the data augmentation methods (rotation and flipping left/right) were considered unnatural in meteorology. However, not all of the methods were incorrect, and in fact, they contributed to an improvement in the classification performance, which can be considered to have had some meaning. It can be said that the empirical knowledge and knowledge accumulated in image recognition in fields other than atmospheric science is beneficial and can be used in atmospheric science as well.

One of the reasons for the success of the competition is the interesting informatics problems, classifying imbalanced data with more than a 20-fold imbalance using real data. In the future, publishing data in a machine-learning ready format for data science competitions and benchmarking could be a new form of collaboration between computer science and geoscience research, such as WeatherBench (Rasp et al. 2020) and the S2S AI Challenge (https://s2s-ai-challenge.github.io/). Therefore, it is particularly important to select a problem set that is applicable to both disciplines.

Abbreviations

NICAM: Nonhydrostatic ICosahedral Atmospheric Model; CNN: Convolutional neural network.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40645-021-00459-y.

Additional file 1. Supplementary material.

Acknowledgements

Special thanks go to Drs. M. Nakano, C. Kodama, and Y. Yamada for producing the training and test data on tropical cyclones; SIGNATE Inc. for organizing the competition; and Mrs. K. Murata, M. Ozeki, N. Ohta, T. Tsutaoka, and all participants in the competition. We also thank Editage for English language review.

Author's information

DM was the host of the competition. Copyright and the right to a patent for the top four submissions in the competition (source code, algorithms, analysis results and reports) are transferred and reserved to DM.

Authors' contributions

DM conceived and designed the study, analyzed the data, and constructed the manuscript. The author read and approved the final manuscript.

Funding

This work was supported by JST, PRESTO (Grant Number JPMJPR1777) and JST, CREST (Grand Number JPMJCR1663).

Availability of data and material

The training and test data generated in this study are available at SIGNATE's competition website (https://signate.jp/competitions/134, only available in Japanese). Please contact the corresponding author for data requests.

Declarations

Competing interests

The authors declare that they have no competing interest.

Received: 12 July 2021 Accepted: 2 December 2021
Published online: 15 December 2021

References

Breiman L (1996) Bagging predictors. Mach Learn 24:123–140

Clevert D, Unterthiner T, Hochreiter S (2015) Fast and accurate deep network learning by exponential linear units (ELUs). Paper presented at international conference on learning representation (ICLR) 2016, Caribe Hilton, San Juan, 2–4 May 2016
DeVries T, Taylor GW (2017) Improved regularization of convolutional neural networks with cutout. arXiv:1708.04552
Frank E, Hal M (2001) A simple approach to ordinal classification. Paper presented at the 12th European Conference on Machine Learning ECML) 2001, Freiburg, 5–7 September 2001
Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 55(1):119–139
Gastaldi X (2017) Shake-shake regularization of 3-branch residual networks. Paper presented at international conference on learning representation (ICLR) 2017, Palais des Congrès Neptune, Toulon, 24–26 April 2017
Goodfellow I, Bengio Y, Courville A (2016) Deep learning. The MIT Press, Cambridge
Han D, Kim J, Kim J (2017) Deep pyramidal residual networks. Paper presented at conference on computer vision and pattern recognition (CVPR) 2017, Hawaii Convention Center, Honolulu, 21–26 July 2017
He H, Garcia EA (2009) Learning from imbalanced data. IEEE Trans Knowl Data Eng 21(9):1263–1284
He K, Zhang X, Ren S, Sun J (2015) Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. Paper presented at international conference on computer vision (ICCV) 2015, CentroParque Convention Center, Santiago, 13–16 December 2015
He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. Paper presented at conference on computer vision and pattern recognition (CVPR) 2016, Caesar's Palace, Las Vegas, 26 June–1 July 2016
Hinton GE, Dayan P, Frey BJ, Neal RM (1995) The "wake-sleep" algorithm for unsupervised neural networks. Science 268(5214):1158–1161
Hochreiter S (1998) The vanishing gradient problem during learning recurrent neural nets and problem solutions. Int J Uncertain Fuzziness Knowl Based Syst 6(2):107–116
Hu Y, Huber A, Anumula J, Liu SC (2018) Overcoming the vanishing gradient problem in plain recurrent networks. arXiv:1801.06105
Kabari LG, Onwuka U (2019) Comparison of bagging and voting ensemble machine learning algorithm as a classifier. Int J Comput Sci Softw Eng 9(3):19–23
Kim SK, Ames S, Lee J, Zhang C, Wilson AC, Williams D (2017) In: Ebert-Uphoff I, Monteleoni C, Nychka D (eds) Massive scale deep learning for detecting extreme climate events. Proceedings of the 7th international workshop on climate informatics, Boulder 2017
Kim SK, Park S, Chung S, Lee J, Lee Y, Kim H, Prabhat, Choo J (2019) Learning to focus and track extreme climate events. Paper presented at the 30th British machine vision conference (BMVC) 2019, Cardiff University, Cardiff, 9–12 September 2019
Kodama C, Yamada Y, Noda AT, Kikuchi K, Kajikawa Y, Nasuno T, Tomita T, Yamaura T, Takahashi HG, Hara M, Kawatani Y, Satoh M, Sugi M (2015) A 20-year climatology of a NICAM AMIP-type simulation. J Meteorol Soc Jpn 93(4):393–424
Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 60:84–90
LeCun Y, Bengio Y (1995) Convolutional networks for images, speech, and time-series. In: Arbib MA (ed) The hand book of brain theory and neural networks. MIT Press, Cambridge
Leevy JL, Khoshgoftaar TM, Bauder RA, Seliya N (2018) A survey on addressing high-class imbalance in big data. J Big Data 5(1):42. https://doi.org/10.1186/s40537-018-0151-6
Leon F, Floria S-A, Bădică C (2017) Evaluating the effect of voting methods on ensemble-based classification. Paper presented at international conference on innovations in intelligent systems and applications (INSTA) 2017, Nadmorski Hotel, Gdynia, 3–5 July 2017
Li Y, Song Y, Luo J (2017) Improving pairwise ranking for multi-label image classification. Paper presented at conference on computer vision and pattern recognition (CVPR) 2017, Hawaii Convention Center, Honolulu, 21–26 July 2017
Lin TY, Goyal P, Girshick R, He K, Dollár P (2020) Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell 42(2):318–327
Liu Y, Racah E, Prabhat, Correa J, Khosrowshahi A, Lavers D, Kunkel K, Wehner M, Collins W (2016) Application of deep convolutional neural networks for detecting extreme weather in climate datasets. arXiv reprint arXiv:1605.01156
Maas AL, Hannun AY, Ng AN (2013) Paper presented at international conference on machine learning (ICML) 2013, Atlanta Marriott Marquis, Atlanta, 16–21 June 2013
Martins AFT, Astudillo RF (2016) From softmax to sparse-max: a sparse model of attention and multi-label classification. Paper presented at international conference on machine learning (ICML) 2016, Marriott Marquis hotel, New York City, 19–24 June 2016
Matsuoka D, Nakano M, Sugiyama D, Uchida S (2017) In: Ebert-Uphoff I, Monteleoni C, Nychka D (eds) Detecting precursors of tropical cyclone using deep neural networks. In: Proceedings of the 7th international workshop on climate informatics, Boulder 2017
Matsuoka D, Nakano M, Sugiyama D, Uchida S (2018) Deep learning approach for detecting tropical cyclones and their precursors in the simulation by a cloud-resolving global nonhydrostatic atmospheric model. Prog Earth Planet Sci. https://doi.org/10.1186/s40645-018-0245-y
Nair V, Hinton GE (2010) Rectified linear units improve restricted Boltzmann machines. Paper presented at international conference on machine learning (ICML) 2010, Haifa Congress Center, Haifa, 21–24 June 2010
Nakano M, Sawada M, Nasuno T, Satoh M (2015) Intraseasonal variability and tropical cyclogenesis in the Western North Pacific simulated by a global nonhydrostatic atmospheric model. Geophys Res Lett 42(2):565–571
Pradhan R, Aygun RS, Maskey M, Ramachandran R, Cecil DJ (2018) Tropical cyclone intensity estimation using a deep convolutional neural network. IEEE Trans Image Process 27(2):692–702
Rasp S, Dueben PD, Scher S, Weyn JA, Mouatadid S, Thuerey N (2020) WeatherBench: a benchmark data set for data-driven weather forecasting. J Adv Model Earth Syst. https://doi.org/10.1029/2020MS002203
Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323(6088):533–538
Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C (2018) MobileNetV2: Inverted residuals and linear bottlenecks. Paper presented at conference on computer vision and pattern recognition (CVPR) 2018, Calvin L. Rampton Salt Palace Convention Center, Salt Lake City, 18–22 June 2018
Shorten C, Khoshgoftaar TM (2019) A survey on image data augmentation for deep learning. J Big Data 6(1):60. https://doi.org/10.1186/s40537-019-0197-0
Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. Paper presented at international conference on learning representation (ICLR) 2015, The Hilton San Diego Resort & Spa, San Diego, 7–9 May 2015
Sugi M, Noda A, Sato N (2002) Influence of the global warming on tropical cyclone climatology: an experiment with the JMA global model. J Meteorol Soc Jpn 80(2):249–272
Sun Y, Wong AKC, Kamel MS (2009) Classification of imbalanced data: a review. Int J Pattern Recognit Artif Intell 23(04):687–719
Sun Y, Wang X, Tang X (2014) Deeply learned face representations are sparse, selective, and robust. Paper presented at conference on computer vision and pattern recognition (CVPR) 2015, Hynes Convention Center, Boston, 7–12 June 2015
Wei J, Suriawinata A, Vaickus L, Ren B, Liu X, Wei J, Hassanpour S (2020) Generative image translation for data augmentation in colorectal histopathology images. Paper at the thirty-third annual conference on neural information processing systems (NeurIPS) 2019, Vancouver Convention Center, Vancouver, 8–14 December 2019
Xu B, Wang N, Chen T, Li M (2015) Empirical evaluation of rectified activations in convolutional network. CoRR [Abs.]/1505.00853
Yamada Y, Satoh M, Sugi M, Kodama C, Noda AT, Nakano M, Nasuno T (2017) Response of tropical cyclone activity and structure to global warming in a high-resolution global nonhydrostatic model. J Clim 30(23):9703–9724
Zagoruyko S, Komodakis N (2016) Wide residual networks. Paper presented at British machine vision conference BMVC) 2016, York University, York, 19–22 September 2016
Zhang H, Cisse M, Dauphin YN, Lopez-Paz D (2018) mixup: beyond empirical risk minimization. Paper presented at international conference on learning representation (ICLR) 2018, Vancouver Convention Center, Vancouver. 30 April–3 May 2018
Zhong Z, Zheng L, Kang G, Li S, Yang Y (2020) Random erasing data augmentation. Paper presented at Conference on Artificial Intelligence (AAAI) 2020, Hilton New York Midtown, New York, 7–12 July 2020

Publisher's Note