Classification of Imbalanced Cloud Image Data Using Deep Neural Networks –Performance Improvement Through a Data Science Competition–

Daisuke Matsuoka (✉ daisuke@jamstec.go.jp)

Research Institute for Value-Added-Information Generation (VAiG)

Methodology

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

Posted Date: July 22nd, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-710989/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Progress in Earth and Planetary Science on December 1st, 2021. See the published version at <https://doi.org/10.1186/s40645-021-00459-y>.

Classification of imbalanced cloud image data using deep neural networks –Performance improvement through a data science competition–

Daisuke Matsuoka¹

Corresponding author

Email: daisuke@jamstec.go.jp

(Institutional address)

¹ Research Institute for Value-Added-Information Generation (VAiG), JAMSTEC, 3173-25 Showa-machi, Kanazawa-ku, Yokohama 236-0001, Japan

Abstract

Image data classification using machine learning is one of the effective methods 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 of data between the target class and the other classes. In order to build a highly accurate classification model, we held a data analysis competition to determine the best classification performance for two classes of cloud image data: 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 imbalanced classification performance. In particular, the best model out of 209 submissions succeeded in improving the classification capability by 65.4% over similar conventional methods in a measure of 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 multi-layered 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 in 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), precursors of tropical cyclones (Matsuoka et al. 2018), estimating hurricane intensity (Pradhan et al. 2018), and so on.

38 Although previous studies have reported interesting results by using classifications, the
39 imbalance in the number of data between classes is still an issue. Target phenomena or
40 structures such as hurricanes and tropical cyclones occur relatively infrequently, while other
41 patterns exist in countless numbers. For example, Matsuoka et al. (2018) classified 50,000
42 positive and 1,000,000 negative example images, with a balance of 20 times. This class
43 imbalance is known to cause a decrease in classification performance (e.g. Sun et al. 2009).
44 Such a problem of classifying imbalanced data is called the imbalanced learning problem, and
45 several methods have been proposed in various fields (e.g. He and Garcia 2009; Leevy et al.
46 2018). These methods can be categorized into data-level approaches, such as training data
47 sampling and feature selection, and algorithm-level approaches, such as cost-sensitive
48 learning and ensemble learning. Furthermore, several derivative methods have been proposed.
49 However, each method often requires empirical judgment and intuition, with domain-specific
50 dependencies. Above all, it is impractical in terms of computational and human costs to try all
51 methods combinations, especially for deep learning that deals with large amounts of data.
52 To solve these problems and obtain the best performing model, we opened the data used in
53 our previous study (Matsuoka et al. 2018) and held a data science competition event to
54 determine the classification performance of tropical cyclones. The competition was held over
55 a two-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 top winners of the competition achieved classification performances that far exceeded those of Matsuoka et al. (2018). 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 is 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 detected 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 $300.0 \text{ W/m}^2 \sim 100.0 \text{ W/m}^2$, and the single-precision real values were readable
PNG format image files. The number of pixels in the image was set to 64×64 (approximately
1,000 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. 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 two years from 1999 to 2000). Although the
number 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 about 1:20. The training data was
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:

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

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 $1 - \text{Precision}$ indicates the *false alarm ratio*. The *Precision* and *Recall* are defined by the following equations:

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

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

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 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.

3 Methods/Experimental

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

3.1 Convolutional neural network

The architecture of a convolutional neural network generally consists of convolutional, pooling, and fully connected layers, as shown in Fig. 2a. 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. The values for classes 0 and 1 are often output in binary classification. To reduce the error between the output value of the CNN and the correct value (ground truth), the parameters of the CNN such as weights and biases are modified from the output to the input layer. This is one of the most important processes in

training with neural networks, and is called backpropagation (Rumelhart et al. 1986). Here, it is common to group several pieces of training data together (called a mini-batch) and update the weights simultaneously.

In models using neural networks, deeper layers generally improve the ability to represent features and increase accuracy. However, simply deepening the layers makes it difficult to learn efficiently because the gradient disappears during back propagation away from the output layer. One powerful method to solve this problem is called the residual network. 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 the gradient disappearance. Some of the known architectures using residual modules are ResNet (He et al. 2016), WideResNet (Zagoruyko and Komodakis 2016), 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 (Fawcett 2016; Leevy et al. 2018). Data augmentation, a technique for minority class oversampling, is widely

145 used to increase the image recognition performance for deep learning (Shorten and
146 Khoshgoftaar 2019). As shown in Fig. 3, vertical flip, horizontal flip, random crop, and
147 random rotation are common methods used in image recognition. Because the result of
148 random crop and random rotation will be smaller than the original image, there is a method
149 called padding, in which pixels are filled outside the original image to make the size equal to
150 the original image. Cutout (DeVries and Taylor 2017) and random erasing (Zhong et al. 2020)
151 were also used to mask partial regions in the image. Whereas cutout masks a square area at a
152 random position with a value of 0, random erasing masks a rectangular area at a random
153 position and size with a random value. For other data augmentation methods, refer to the
154 survey paper (Shorten and Khoshgoftaar 2019). Many of these methods have also been
155 implemented in deep learning frameworks such as Keras and PyTorch, and users can utilize
156 them with simple functions.

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

misclassified by the CNN (hard negative example). Here, X is the hard-negative ratio. The probability distribution is update 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. 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 the average of the final layer outputs, while 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 the 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

is shown in Fig. 6. From the definition of *Conditional precision* shown in Equation (1), although the final evaluation of the submission with $Recall < 0.79$ (represented by a blue triangle in the figure) is zero, both *Precision* and *Recall* values are shown for reference. Because there is a trade-off between the two, the results are concentrated around $Recall = 0.8$ (represented by a dotted line in the figure) with a small margin to clear the condition of $Recall \geq 0.79$. 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 have a $Precision > 0.6$, with $Recall \geq 0.79$. There are a few models around $Precision = 0.6$, where *Recall* is marginally less than 0.79, but with good enough performance, even though they received zero marks in the competition.

4.2 Classification strategies

The methods used in the top four models, which are 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) 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

217 first-ranked model was a 110-layer PyramidNet, followed by a 10-layer WideResNet, and
218 26-layer ResNet, with the maximum number of convolution filters for any of the models of
219 256 (Fig. 2c). In addition, Shake-Shake regularization (Gastaldi 2017), a data augmentation to
220 the output (feature map) of the middle layer, was used in the third-ranked model. The
221 fourth-ranked model was MobileNetV2 (Sandler et al. 2018), a pre-trained model with three
222 convolutional layers. Deep learning requires tuning many parameters, such as the number of
223 layers, number of kernels, kernel size, and type of activation function, but it is difficult to try
224 them all because of time constraints. The advantage of using a fast and lightweight model is
225 that a lot of trial and error for parameter tuning and data augmentation can be conducted in a
226 limited amount of time.

227 Data augmentation for minority classes during training and/or testing was effective in
228 improving the classification performance. In all the top three models, methods such as
229 horizontal flip, vertical flip, random shift, cutout, and random rotation (four patterns of 0°,
230 90°, 180°, and 270°) were used (in the fourth-ranked model, only during training). In
231 particular, the second-ranked model increased the original image twenty-fold by combining
232 five patterns of crops in the center and four corners (5 Crop) with four patterns of random
233 rotation (and random erasing for each image during the training phase). In addition, MixUp
234 (Zhang et al. 2018), which creates new images by superimposing positive and negative

235 example images, is used in the fourth model. These data augmentation methods improved the
236 accuracy of repeated experiments and were finally determined to be effective. However, the
237 direction of the image has meteorological significance, as the direction of rotation of tropical
238 cyclones differs between the northern and southern hemispheres, and the location of windy
239 areas toward the direction of movement differs. In this sense, horizontal flip and random
240 rotation may not be meteorologically valid, and their use should be considered. However,
241 random cropping is a method that can be used for data augmentation without any
242 inconsistency. Moreover, random erasing and cutout are also valuable methods because they
243 provide regularization effects and are robust to noise.

244 Ensemble learning, which combines multiple models, is also an effective method for
245 improving classification accuracy. The first-ranked model used the weighted average of the
246 output results from five models with different hard-negative ratios. The fourth-ranked model
247 used the simple average of the output results of the five models with different preprocessing
248 methods or learning rates. In addition, Matsuoka et al. (2018) used the weighted average of
249 the output results from 10 models trained on different negative samples. These are all methods
250 that can process multiple models in parallel. However, boosting, which is a sequential process
251 to preferentially learn the results of the previous model's misclassification, is also known to be
252 a powerful method and was used by Matsuoka et al. (2017). The second- and third-ranked

253 models were trained and tested using only a single model. While ensemble models are
254 effective in improving classification accuracy, they also increase the difficulty to interpret the
255 models. It is also important to prioritize single models with high accuracy to obtain
256 meteorological knowledge from feature maps in trained models.

257 During classifications, cross-entropy is often used as a function to evaluate the error between
258 the inference result and true value (Hinton et al. 1995). In the second-ranked model, focal loss
259 (Lin et al. 2018) was used, which provided a large weighting for the loss of minority classes.

260 Loss functions for imbalanced image classifications include weighted cross entropy loss
261 (Hinton et al. 1995), Hamming loss (Frank and Hall 2001), and other classical functions such
262 as ranking loss (Li et al. 2017) and sparseMax loss (Martins and Astudillo 2016). A
263 randomized leaky rectified unit (RRELU) (Xu et al. 2015) was used in the three-place model
264 as the activation function for transmitting the output of each layer to the next layer. ReLU
265 (Nair and Hinton 2010; Sun et al. 2014), which is widely used as the most common activation
266 function, becomes zero when the input is negative and constant otherwise. In RRELU, the
267 function has a random slope when it is negative, which prevent overlearning. Other activation
268 functions such as Leaky ReLU (Maas et al. 2013), parametric ReLU (He et al. 2015), and
269 exponential linear units (Clevert et al. 2016) are also known.

270

5 Conclusions

This paper proposes deep learning approaches to effectively classify imbalanced cloud image data with its differences of more than 20 times. In order 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* is 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 contributed to the improvement of 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 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 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

Declarations

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.

306 **Competing interests**

307 The authors declare that they have no competing interest.

308

309 **Funding**

310 This work was supported by JST, PRESTO Grant Number JPMJPR1777.

311

312 **Authors' contributions**

313 DM conceived and designed the study, analyzed the data, and constructed the manuscript. The

314 final manuscript has been read and approved.

315

316 **Authors' information**

317 DM is the host of the competition. Copyright and the right to a patent for the top four

318 submissions in the competition (source code, algorithms, analysis results and reports) are

319 transferred and reserved to DM.

320

321 **Acknowledgements**

322 Special thanks to Drs. M. Nakano, C. Kodama, and Y. Yamada for producing the training and

323 test data of tropical cyclones; SIGNATE Inc. for organizing the competition; and Mr. K.

Murata and all the participants in the competition. We also thank Editage for English language review.

Endnotes

References

Breiman L (1996) Bagging predictors. *Machine Learning* 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 reprint arXiv:1708.04552

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

342 Gastaldi X (2017) Shake-shake regularization of 3-branch residual networks. Paper presented
 343 at International Conference on Learning Representation (ICLR) 2017, Palais des Congrès
 344 Neptune, Toulon, 24-26 April 2017
 345
 346 Han D, Kim J, Kim J (2017) Deep pyramidal residual networks. Paper presented at
 347 Conference on Computer Vision and Pattern Recognition (CVPR) 2017, Hawaii Convention
 348 Center, Honolulu, 21-26 July 2017
 349
 350 He K, Zhang X, Ren S, Sun J (2015) Delving deep into rectifiers: surpassing human-level
 351 performance on ImageNet classification. Paper presented at International Conference on
 352 Computer Vision (ICCV) 2015, CentroParque Convention Center, Santiago, 13-16 December
 353 2015
 354
 355 He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. Paper
 356 presented at Conference on Computer Vision and Pattern Recognition (CVPR) 2016, Caesar's
 357 Palace, Las Vegas, 26 June-1 July 2016
 358
 359 Hinton GE, Dayan P, Frey BJ, Neal RM, (1995) The “wake-sleep” algorithm for unsupervised

360 neural networks. *Science* 268(5214):1158–1161

361

362 Kodama C, Yamada Y, Noda A T, Kikuchi K, Kajikawa Y, Nasuno T, Tomita T, Yamaura T,
363 Takahashi HG, Hara M, Kawatani Y, Satoh M, Sugi M (2015) A 20-year climatology of a
364 NICAM AMIP-type simulation. *J. Meteorol. Soc. Jpn* 93(4):393–424

365 Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep
366 convolutional neural networks. *Advances in Neural Information Processing Systems* 60:84–90

367

368 Leevy JL, Khoshgoftaar TM, Bauder RA, Seliya N (2018) A survey on addressing high-class
369 imbalance in big data. *J Big Data* 5(1). doi:10.1186/s40537-018-0151-6

370

371 Leon F, Floria S-A, Bădică C (2017) Evaluating the effect of voting methods on
372 ensemble-based classification. Paper presented at International Conference on Innovations in
373 Intelligent Systems and Applications (INSTA) 2017, Nadmorski Hotel, Gdynia, 3-5 July 2017

374

375 Li Y, Song Y, Luo J (2017) Improving pairwise ranking formulti-label image classification.
376 Paper presented at Conference on Computer Vision and Pattern Recognition (CVPR) 2017,
377 Hawaii Convention Center, Honolulu, 21-26 July 2017

378

379 Lin TY, Goyal P, Girshick R, He K, Dollár (2020) Focal loss for dense object detection. IEEE
380 Trans. Pattern Anal. Mach. Intell. 42(2):318–327

381

382 Maas AL, Hannun AY, Ng AN (2013) Paper presented at International Conference on
383 Machine Learning (ICML) 2013, Atlanta Marriott Marquis, Atlanta, 16-21 June 2013

384

385 Martins AFT, Astudillo RF (2016) From softmax to sparse-max: A sparse model of attention
386 and multi-label classification. Paper presented at International Conference on Machine
387 Learning (ICML) 2016, Marriott Marquis hotel, New York City, 19-24 June 2016

388

389 Matsuoka D, Nakano M, Sugiyama D, Uchida S (2017) In: Ebert-Uphoff I, Monteleoni C,
390 Nychka D (eds) Detecting precursors of tropical cyclone using deep neural networks.
391 Proceedings of the 7th International Workshop on Climate Informatics, Boulder 2017

392

393 Matsuoka D, Nakano M, Sugiyama D, Uchida S (2018) Deep learning approach for detecting
394 tropical cyclones and their precursors in the simulation by a cloud-resolving global
395 nonhydrostatic atmospheric model. Prog. Earth Planet. Sci. 5(1). doi:

396 10.1186/s40645-018-0245-y

397

398 Nair V, Hinton GE (2010) Rectified linear units improve restricted Boltzmann machines.

399 Paper presented at International Conference on Machine Learning (ICML) 2010, Haifa

400 Congress Center, Haifa, 21-24 June 2010

401

402 Rasp S, Dueben PD, Scher S, Weyn JA, Mouatadid S, Thuerey N (2020) WeatherBench: A

403 benchmark data set for data-driven weather forecasting. J. Adv. Model. Earth Syst. 12(11).

404 doi:10.1029/2020MS002203

405

406 Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating

407 errors. Nature 323(6088):533–538

408

409 Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C (2018) MobileNetV2: Inverted

410 residuals and linear bottlenecks. Paper presented at Conference on Computer Vision and

411 Pattern Recognition (CVPR) 2018, Calvin L. Rampton Salt Palace Convention Center, Salt

412 Lake City, 18-22 June 2018

413

414 Sato M, Tomita H, Yashiro H, Miura H, Kodama C, Seiki T, Noda A T, Yamada Y, Goto D,
 415 Sawada M, Miyoshi T, Niwa Y, Hara M, Ohno T, Iga S, Arakawa T, Inoue T, Kubokawa H
 416 (2014) The non-hydrostatic icosahedral atmospheric model: Description and development.
 417 Prog. Earth Planet. Sci. 1:1–32
 418
 419 Shorten C, Khoshgoftaar TM (2019) A survey on image data augmentation for deep learning.
 420 J. Big Data 6(1). doi:10.1186/s40537-019-0197-0
 421
 422 Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image
 423 recognition. Paper presented at International Conference on Learning Representation (ICLR)
 424 2015, The Hilton San Diego Resort & Spa, San Diego, 7-9 May 2015
 425
 426 Sugi M, Noda A, Sato N (2002) Influence of the global warming on tropical cyclone
 427 climatology: An experiment with the JMA global model. J. Meteorol. Soc. Jpn. 80(2):249–
 428 272
 429
 430 Sun Y, Wang X, Tang X (2014) Deeply learned face representations are sparse, selective, and
 431 robust. Paper presented at Conference on Computer Vision and Pattern Recognition (CVPR)

432 2015, Hynes Convention Center, Boston, 7-12 June 2015

433

434 Xu B, Wang N, Chen T, Li M (2015) Empirical evaluation of rectified activations in
435 convolutional network. CoRR [Abs.]/1505.00853

436

437 Yamada Y, Satoh M, Sugi M, Kodama C, Noda AT, Nakano M, Nasuno T (2017) Response of
438 tropical cyclone activity and structure to global warming in a high-resolution global
439 nonhydrostatic model. J. Clim. 30(23):9703–9724

440

441 Zagoruyko S, Komodakis N (2016) Wide residual networks. Paper presented at British
442 Machine Vision Conference BMVC) 2016, York University, York, 19-22 September 2016

443

444 Zhang H, Cisse M, Dauphin YN, Lopez-Paz D (2018) mixup: Beyond empirical risk
445 minimization. Paper presented at International Conference on Learning Representation
446 (ICLR) 2018, Vancouver Convention Center, Vancouver, 30 April-3 May 2018

447

448 Zhong Z, Zheng L, Kang G, Li S, Yang Y (2020) Random erasing data augmentation. Paper
449 presented at Conference on Artificial Intelligence (AAAI) 2020, Hilton New York Midtown,

New York, 7-12 July 2020

Figure legends

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

Figure 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).

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

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

Figure 5. Proposed ensemble learning model. (a) Training and (b) test phase.

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

Tables

Table 1. Typical methods used by the top models of the competition.

Figures

Figure 1

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

Figure 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).

Figure 3

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

Figure 4

Undersampling of the majority class using hard negative mining.

Figure 5

Proposed ensemble learning model. (a) Training and (b) test phase.

Figure 6

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

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [GraphicalAbstractImage.png](#)
- [Table1.pdf](#)