Clinical Radiology 74 (2019) 697–701

Contents lists available at ScienceDirect

Clinical Radiology

journal homepage: www.clinicalradiologyonline.net

Effect of augmented datasets on deep convolutional neural networks applied to chest radiographs

R. Ogawa [a,*], T. Kido [b], T. Kido [b], T. Mochizuki [b]

[a] *Department of Radiology, Saiseikai Matsuyama Hospital, 880-2, Yamanishicho, Matsuyama-shi, Ehime, 791-8026, Japan*

[b] *Department of Radiology, Ehime University Graduate School of Medicine Shitsukawa, Toon-city, Ehime 791-0295, Japan*

ARTICLE INFORMATION

Article history:
Received 30 December 2018
Accepted 9 April 2019

AIM: To evaluate the effect of augmented training datasets in a deep convolutional neural network (DCNN) used for detecting abnormal chest radiographs.

MATERIALS AND METHODS: Chest radiographs were corrected to conform to a DCNN dataset, with 288 abnormal and 447 normal radiographs. The radiographic images were divided into training and validation sets (441, 60%), and a test set (294, 40%). The training and validation sets were augmented to generate a total of 12,789 training and validation images. The augmentation consisted of operations such as rotation, horizontal and vertical flipping, Gaussian blur, and brightness variation, either alone or combined. The DCNN performed binary classification of the images as being abnormal or normal chest radiographs, and accuracy was used as measure to assess the model performance.

RESULTS: The accuracy of the DCNN trained with the augmented dataset tended to be higher than that of the DCNN trained with the non-augmented dataset. The augmented datasets combining rotation and horizontal flipping had a high accuracy of 0.91, showing the highest accuracy among the applied augmentation techniques and combinations.

CONCLUSION: Augmentation of training datasets can improve the performance of DCNN for radiographic image classification depending on the applied augmentation technique.

Introduction

Image recognition based on machine learning is increasingly becoming more prevalent, particularly with respect to the dramatic development obtained from deep convolutional neural networks (DCNNs). Several DCNN models are currently available, including VGG16,[1] GoogLeNet,[2] ResNet,[3] and AlexNet,[4] and these have achieved good results in competitions such as the ImageNet Large Scale Visual Recognition Challenge,[5] which focused on classifying non-medical images of targets such as animals and vehicles. More recently, DCNN models have been applied for medical image diagnosis.[6] Image classification performance using DCNNs is dependent on the data available in the training datasets, with large and diverse datasets providing the best results; however, correcting large datasets is usually time and effort intensive. Furthermore, the availability of medical images to train DCNNs is limited. To improve training

* Guarantor and correspondent: R. Ogawa, Department of Radiology, Saiseikai Matsuyama Hospital, 880-2, Yamanishicho, Matsuyama-shi, Ehime, 791-8026, Japan. Tel.: +(81). 89 951 6111.
E-mail address: qq8y7cvd@tiara.ocn.ne.jp (R. Ogawa).

https://doi.org/10.1016/j.crad.2019.04.025
0009-9260/

datasets, images can be processed to increase the number of samples by employing a method called data augmentation. For instance, AlexNet was trained with an augmented dataset obtained after applying operations such as horizontal flipping to non-medical colour images to improve the classification performance.[4] In the present study, different types of augmentation techniques applied to grey-scale medical images were evaluated to determine the most suitable approach to improve DCNN-based classification. Specifically, the effect of augmentation on chest radiographs was considered given their widespread use in clinical practice.

Materials and methods

Dataset creation

This retrospective study was approved by the institutional review board of Ehime University Graduate School of Medicine, and requirements for informed consent were waived. A total of 735 chest radiographs (anteroposterior or posteroanterior) were obtained from the picture archiving and communication systems at Ehime University Graduate School of Medicine. Two radiologists (with 7 and 18 years of experience in diagnostic imaging) classified the radiographs as being either abnormal or normal. From the 735 chest radiographs, 288 exhibited abnormalities and the other 447 were normal according to the specialists. The main abnormalities and their rates are listed in Table 1; some radiographs simultaneously exhibited more than one abnormality.

Fig 1 illustrates the stages of image processing and data augmentation adopted in this study. The images were converted into the JPEG (Joint Photographic Experts Group) format with sizes of 1000×1000–2400×2000 pixels at 8 bits. Then, each image was resized to 128×128, 192×192, and 256×256 pixels. The following process was conducted for each image resolution. The 735 images were randomly split into training and validation sets (60%), and test sets (40%), containing 441 and 294 images, respectively. Further, the training and validation sets were split randomly in a ratio of 80/20. This way, the 735 images were divided into training (352, 48%), validation (89, 12%), and test (294, 40%) sets.

Table 1
Main abnormalities found on chest radiographs from this study.

Main abnormality	Present dataset	NIH dataset
Infiltration	27 (9%)	7 (9%)
Cardiomegaly	67 (23%)	17 (23%)
Emphysema	7 (2%)	5 (7%)
Effusion	47 (16%)	10 (13%)
Mass	26 (9%)	6 (8%)
Nodule	20 (7%)	10 (13%)
Pneumothorax	9 (3%)	2 (3%)
Fibrosis	44 (15%)	9 (12%)
Consolidation	19 (7%)	6 (8%)
Atelectasis	22 (8%)	3 (4%)
Total	288	75

Data are presented as *n* (%).

The training and validation sets were augmented by applying rotation, Gaussian blur, brightness variation, and horizontal and vertical flipping. The rotation was set between −28° and 28° at increments of 2°. Gaussian blur transforms an image using a Gaussian function. Specifically, a weight is put on neighbouring pixels according to the distance from a noteworthy pixel. The blur radius was changed between 1.0 and 3.7. Brightness varied in factors between 0.5 and 2.0. Considering Gaussian blur, brightness variation, horizontal and vertical flipping, and some of their combinations, a dataset with 12,789 training and validation images was generated. From these images, 10,231 (80%) images were used for the training set (80%) and the remaining 2,558 (20%) for the validation set. Fig 2 shows examples of images from different training sets.

DCNN mode architecture

DCNN models were built using the Keras library originally created by François Chollet and written in Python 3.5.4. The basic DCNN model is composed of six convolutional layers followed by one max-pooling layer and one fully connected layer. Dropout values of 1/4 and 1/2 were used after the max-pooling and fully connected layers, respectively. A leaky rectified linear unit was used in each layer, except for the output layer. Small filters of 3×3 were used at each convolutional layer, along with the same padding. Models with different augmented training sets were established. The network training was executed using an NVIDIA deep-learning GPU training system GeForce GTX1080 running the Microsoft Windows 10 64-bit operating system with NVIDIA CUDA 8.0 and cuDNN on a computer with an Intel i7 6800k processor at 3.4 GHz.

Model assessment and statistical analyses

Each augment dataset was loaded onto the DCNN for testing five times. From the five assessments of each dataset, the dataset with highest accuracy was adopted as the result for that type of augmentation. The DCNN for the augmented dataset producing the highest performance was further tested on a publicly available test set from the National Institutes of Health (NIH). Seventy-five normal and 75 abnormal cases were selected randomly from the NIH test set, excluding radiographs containing medical tubes. Hence, 150 records from the NIH test set were used. The main abnormalities of the images on this test set are those listed in Table 1. Fig 1 illustrates the evaluations in this study. The sensitivity, specificity, positive and negative predictive values, and accuracy were assessed for detecting abnormal images on the NIH test set. The adjusted Wald method was used to determine the 95% confidence interval (CI).[7]

Results

DCNN performance according to augmented dataset

Table 2 lists the performance of the DCNN trained with different augmented datasets. Overall, the accuracy

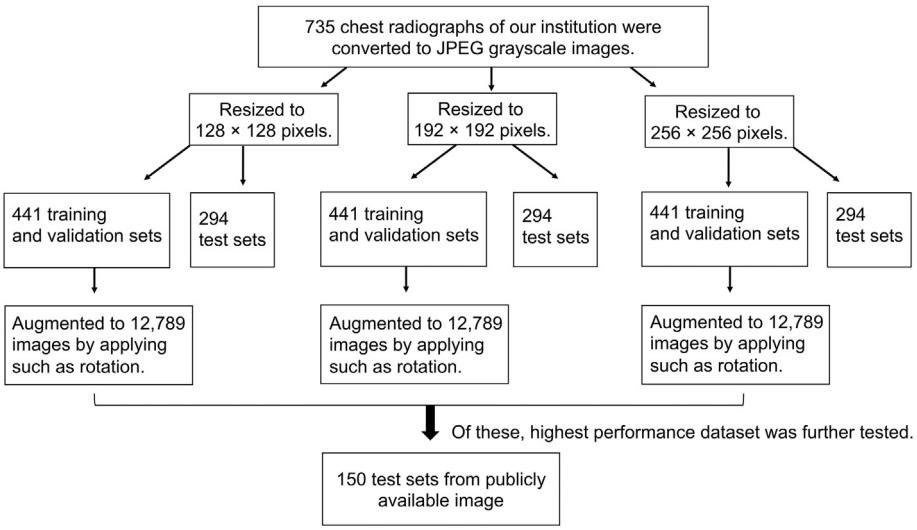

Figure 1 Diagram summarising this study including dataset generation and DCNN evaluation.

improves using augmented datasets compared to the non-augmented datasets; however, applying Gaussian blur operation sometimes produces lower accuracy than when using non-augmented data. Regarding resolution, images of 128×128 pixels show the highest accuracy. Further, at a given resolution, accuracy tends to be higher when combining rotation and horizontal flipping than when applying other augmentation operations. Therefore, this type of augmentation was used to evaluate the publicly available NIH test set.

Evaluation on NIH test set

When applied to the NIH test set, the DCNN classified abnormal and normal chest radiographs with a sensitivity of 85% (95% CI: 77–94%), specificity of 81% (95% CI: 72–90%), positive predictive value of 82% (95% CI: 73–91%), negative predictive value of 85% (95% CI: 76–93%), and accuracy of 83% (95% CI: 77–89%). Table 3 lists the true positives and false negatives according to the main abnormalities in the images from this test set.

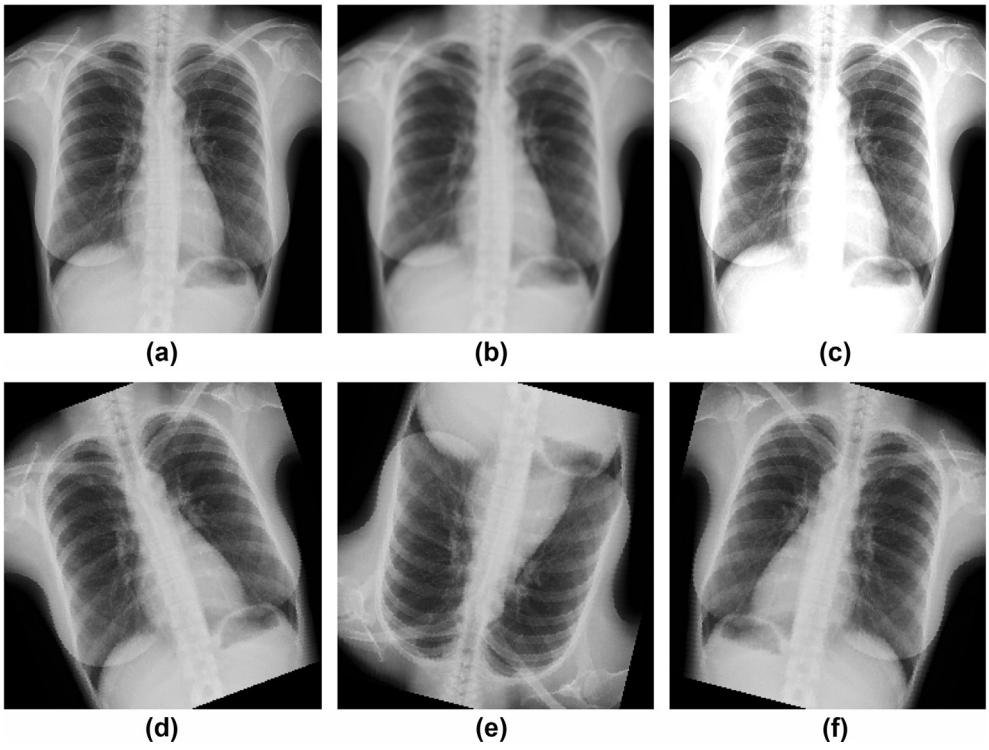

Figure 2 Example of images from the training datasets: (a) non-augmented image, (b) Gaussian blur, (c) brightness variation, (d) rotation, (e) rotation + vertical flipping, (f) rotation + horizontal flipping.

Table 2
Accuracy of the deep convolutional neural networks trained by different augmented datasets.

Dataset	Image resolution		
	128×128	192×192	256×256
Non-augmented dataset	0.81 (0.76–0.85)	0.83 (0.79–0.87)	0.83 (0.79–0.87)
Rotation	0.82 (0.78–0.86)	0.85 (0.81–0.89)	0.84 (0.80–0.88)
Rotation/rotation + horizontal flipping	0.91 (0.88–0.94)	0.89 (0.86–0.93)	0.89 (0.86–0.93)
Rotation/rotation + vertical flipping	0.86 (0.82–0.90)	0.85 (0.81–0.89)	0.88 (0.84–0.92)
Gaussian blur	0.78 (0.73–0.83)	0.82 (0.78–0.86)	0.82 (0.78–0.86)
Gaussian blur/Gaussian blur + horizontal flipping	0.82 (0.78–0.86)	0.76 (0.71–0.81)	0.79 (0.74–0.84)
Gaussian blur/Gaussian blur + vertical flipping	0.83 (0.79–0.87)	0.78 (0.73–0.83)	0.74 (0.69–0.79)
Brightness variation	0.84 (0.80–0.88)	0.85 (0.81–0.89)	0.88 (0.84–0.92)
Brightness variation/brightness variation + horizontal flipping	0.87 (0.83–0.91)	0.84 (0.80–0.88)	0.86 (0.82–0.90)
Brightness variation/brightness variation + vertical flipping	0.85 (0.81–0.89)	0.85 (0.81–0.89)	0.84 (0.80–0.88)

Data are presented as accuracy (95% confidence interval).

Discussion

Classification performance depending on augment dataset

Training data with augmentation can lead to higher classification accuracy compared to non-augmented datasets (Table 2). Besides augmentation, transfer learning is another method to improve DCNN performance by using weights derived from large datasets. In a previous study, transfer learning was used for the classification of chest radiographs,[8] but data augmentation combined with transfer learning showed improved performance.[9] Therefore, the results of the present study can be useful for DCNN implementation even when employing transfer learning.

Regarding the augmentation techniques, other studies have used horizontal flipping or rotation to determine the presence of endotracheal tubes on radiographs.[10] In the present study, pixel size and combined rotation–horizontal flipping produced the highest accuracy, thus indicating the relevance of these operations to generate additional chest radiographic images. Therefore, augmentation can be beneficial for high DCNN performance and preventing the correction of large datasets, especially for medical images, which are scarce for abnormal cases compared to other types of images.

In the present study, Gaussian blur applied to training sets undermined classification accuracy. Although Gaussian blur can mitigate noise, it has drawbacks such as the loss of fine detail.[11] Therefore the poor classification performance may be explained by texture loss after blurring. Unlike blurring, operations such as rotation and horizontal flipping preserve fine detail. Therefore the type operation determines the classification performance achieved by data augmentation.

Interestingly, higher resolution does not always imply higher classification accuracy, and the most suitable resolution may differ depending on the dataset. For instance, a higher resolution may be required to identify small abnormalities[12]; however, in the dataset used in the present study, nodules, which are usually small, account for only 7% (n=20) of the cases, and the employed DCNN is not specialised for small abnormalities.

Table 3
True positives and false negatives determined by highest-performing deep convolutional neural networks (DCNN) on National Institutes of Health dataset.

Main abnormality	True positives	False negatives
Infiltration	5	2
Cardiomegaly	17	0
Emphysema	4	1
Effusion	9	*1*
Mass	6	0
Nodule	7	3
Pneumothorax	2	0
Fibrosis	6	3
Consolidation	5	1
Atelectasis	3	0
Total	64	11

Evaluation on publicly available NIH test set

In addition to the test set, a publicly available an NIH test set was also evaluated to verify the generalisation ability of the DCNN. A previous study reported that cardiomegaly and pneumothorax are more distinguishable than other abnormalities such as mass and nodule.[13] In the present study, cardiomegaly produced better recognition than other abnormalities including nodules (Table 3). These results suggest that large abnormal features are easily distinguished by the DCNN. Overall, the accuracy of the DCNN tended to be higher on the authors' test set than on the NIH test set. This divergence can be caused by differences in aspects such as organ shapes, ratio of main abnormalities, and number of images.

Limitations

Some limitations from this study remain to be addressed. First, the DCNN architecture was not specialised to detect a specific abnormality, such as nodules, as several abnormal features can appear in a single chest radiograph. Hence, in this study, a general DCNN that determines whether any abnormality is present was tested. Binary normality classification of chest radiographs may be valuable to assist clinicians as a screening tool.[14] Second, if the parameters of the DCNN are tuned to each type of augmentation, higher performance may be achieved; however, the aim of the

present study was to evaluate the effect of augmentation, and therefore, the DCNN architecture and parameters were fixed. Third, the reasons for some of the false positives are not clearly understood. This problem can be addressed by generating heat maps, which usually localise the areas of the image where more attributes are found by the DCNN for recognition.[15] Likewise, another study introduced visual tools to interpret trained artificial neural networks.[16] Future studies can include these methods to unveil the mechanisms behind DCNN classification.

In conclusion, augmentation of training datasets was useful for the binary classification of chest radiographs using a DCNN. Classification performance was highly dependent on the type of augmentation techniques employed.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgments

None.

References

1. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv Preprint arXiv:1409.1556* 2014.
2. Szegedy C, Liu W, Jia Y, *et al*. Going deeper with convolutions. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Piscataway, NJ: IEEE; 2015. p. 1–9.
3. He K, Zhang X, Ren S, *et al*. Deep residual learning for image recognition. *arXiv Preprint arXiv:1512.03385* 2015.
4. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: *Advances in neural information processing systems 25*. Neural Information Processing Systems Foundation; 2012. p. 1097–105.
5. Russakovsky O, Deng J, Su H, *et al*. ImageNet large scale visual recognition challenge. *Int J Comput Vis* 2015;**115**(3):211–52.
6. Liew C. The future of radiology augmented with artificial intelligence: a strategy for success. *Eur J Radiol* 2018;**102**:152–6.
7. Agresti A, Coull BA. Approximate is better than “exact” for interval estimation of binomial proportions. *Am Stat* 1998;**52**(2):119–26.
8. Rajkomar A, Lingam S, Taylor AG, *et al*. High-throughput classification of radiographs using deep convolutional neural networks. *J Digit Imaging* 2017;**30**(1):95–101.
9. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology* 2017;**284**(2):574–82.
10. Lakhani P. Deep convolutional neural networks for endotracheal tube position and X-ray image classification: challenges and opportunities. *J Digit Imaging* 2017;**30**(4):460–8.
11. Vibhakar A, Tiwari M, Singh J. Performance analysis for MRI denoising using intensity averaging Gaussian blur concept and its comparison with wavelet transform method. *Int J Comput Appl* 2012;**58**(15):21–6.
12. Yao L, Poblenz E, Dagunts D, *et al*. Learning to diagnose from scratch by exploiting dependencies among labels. *arXiv Preprint arXiv:1710.10501* 2017.
13. Wang X, Peng Y, Lu L, *et al*. ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: *IEEE conference on computer vision and pattern recognition*. Piscataway, NJ: IEEE; 2017. p. 3462–71.
14. Yates EJ, Yates LC, Harvey H. Machine learning "red dot": open-source, cloud, deep convolutional neural networks in chest radiograph binary normality classification. *Clin Radiol* 2018;**73**(9):827–31.
15. Guo H, Fan X, Wang S. Human attribute recognition by refining attention heat map. *Pattern Recognit Lett* 2017;**94**(15):38–45.
16. Yosinski J, Clune J, Nguyen A, *et al*. Understanding neural networks through deep visualization. *arXiv Preprint arXiv:1506.06579* 2015.