MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190268

Identifying Cardiomegaly in ChestX-ray8 Using Transfer Learning

Sicheng Zhou[a], Xinyuan Zhang[b], Rui Zhang[a,c]

[a] *Institute for Health Informatics, University of Minnesota Twin Cities, Minneapolis, Minnesota, USA*
[b] *School of Biomedical Informatics, University of Texas Health Science Center, Houston, Texas, USA*
[c] *Department of Pharmaceutical Care & Health Systems, University of Minnesota Twin Cities, Minneapolis, Minnesota, USA*

Abstract

Recently, the National Institutes of Health (NIH) published a chest X-ray image database named "ChestX-ray8", which contains 108,948 X-ray images that are labeled with eight types of diseases. Identifying the pathologies from the clinical images is a challenging task even for human experts, and to develop computer-aided diagnosis systems to help humans identify the pathologies from images is an urgent need. In this study, we applied the deep learning methods to identify the cardiomegaly from the X-ray images. We tested our algorithms on a dataset containing 600 images, and obtained the best performance with an area under the curve (AUC) of 0.87 using the transfer learning method. This result indicates the feasibility of developing computer-aided diagnosis systems for different pathologies from X-rays using deep learning techniques.

Keywords:

X-rays, Cardiomegaly, Machine Learning

Introduction

Cardiomegaly is a medical condition that indicates the enlarged heart. It can be an important sign of potentially severe heart diseases, such as high blood pressure, heart valve disease, heart muscle disease, pericardial effusion, and heart attack. Also, the cardiomegaly can increase the risks of several heart diseases, including the heart failure, cardiac arrest, blood clots in heart, and heart murmur [1]. It is estimated that there are more than 200,000 new cases of cardiomegaly every year in the US, which is an important health issue for the middle-aged and senior population. The cardiomegaly is easier to treat if detected early, thus its early diagnosis is important. The chest X-ray, as one of the most frequently used examination in radiology, has been applied for several decades to detect and visualize the abnormalities of body organs [2-3]. It has been proven to be an effective diagnostic tool for pathological alterations, with additional advantages due to its noninvasive and low-cost characteristics [4]. Also, the X-ray is an important clinical diagnosis method for cardiomegaly [5].

Distinguishing the different types of pathologies from the X-ray is a challenging and tedious task even for domain experts [6]. Figure 1 shows the examples of X-rays in healthy and cardiomegaly condition. Only trivial differences exist between the two images. Thus, it is meaningful to develop a computer-aided detection method to facilitate the clinicians to accurately identify the X-ray images with cardiomegaly. Many related works have been done in developing such techniques in pathology detection field. Traditional image classification models have been applied in medical image classification and achieved satisfying performances. For instance, the Local Binary Patterns (LBP) [7-8] and the Bag-of-Visual-Words (BoVW) [9; 10] are two models widely applied to retrieve or classify the radiology images. In the LBP model, the images are usually divided into small cells of pixels, then the values of pixels in each cell are binarized. Next, the LBP histograms over each cell can be calculated as features passed to the classifier [10]. The BoVW is also called dictionary learning, it was proposed to mimic the visual processing by the human's brain [10]. The BoVW is similar to the bag-of-words representation of text, it regards an image as a distribution of local descriptors. Those local patch descriptors can be color, texture, shape, and the most popular descriptor is the scale-invariant feature transform (SIFT) [9-10].

Figure 1–(a)X-ray Images for Patients without Cardiomegaly. (b) X-ray Images for Patients with Cardiomegaly.

The deep neural networks have been widely applied in image processing area due to the development of new variant convolutional neural networks (CNNs) models [11-12] and modern hardware [13]. These models have reached promising performances for image classification tasks for the ImageNet Large Scale Visual Recognition Competition (ILSVRC) [14]. Also, these methods have been successfully applied in the processing of the medical images, including the detection of pleural effusion assessment on chest radiography, the lymph node detection on CT, brain segmentation, and assessing diabetic retinopathy [15]. Thus, our study explores the feasibility of using deep learning methods to identify the cardiomegaly from the X-ray images.

The deep learning methods are most effective when applied on large data sets, however, such large datasets are often not available in the medical field due to privacy issues [13]. Recently, the National Institutes of Health (NIH) published a new chest X-ray database called "ChestX-ray8", which contains 108,948 frontal view X-ray images taken from 32,717 unique patients. Natural language processing (NLP) techniques were used to annotate each image with eight disease labels from the associated radiological reports. The disease labels include atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, pneumonia, and pneumothorax [16]. This annotated large dataset provided opportunities to developing computer-aided diagnosis (CAD) systems based on the deep learning methods. Initial experiments of image classification were conducted using this dataset [16]. They built a Deep Convolutional Neural Network (DCNN) architectures that used different ImageNet pre-trained models, i.e., AlexNet [11], GoogLeNet [12], VGGNet-16 [17] and ResNet-50 [18]. It is found that using the ResNet-50 pre-trained model achieved the best results in identifying the cardiomegaly from the X-rays (AUC = 0.814).

There are three pre-trained models we will use in this study, including the ResNet-50, InceptionV3, and Xception [19]. The ResNet-50 is a type of CNN model, and it was proposed to solve the problem that when the traditional CNN model has deeper layers, the performance would decrease in training and test set but not due to overfitting. Basically, ResNet-50 is a residual net with 50 layers. It can allow the model to go deeper, and enable smooth propagation and optimization [18]. This model won the first place on the ImageNet detection challenge. Compared to the traditional CNN models, the InceptionV3 has advantages that it can improve the overfitting problems of complex neural networks and reduce the parameters of the neural networks to reduce the computing cost. The core idea of InceptionV3 is to use convolution kernels with different sizes to extract features with different granularities from the images [12]. Xception was developed based on the InceptionV3 model, the depthwise separable convolution was introduced into the Xception model, which can process the spatial dimension and channel dimension of the image separately. Xception has roughly the same number of parameters as InceptionV3 and obtained better results on classification dataset consisting of 350 million images and over 17,000 classes [19]. In this work, we tried to identify the cardiomegaly from X-rays by combing three above-mentioned state-of-art pre-trained models based on ImageNet. We implemented the transfer learning method to integrate the three pre-trained models for identifying the cardiomegaly from X-rays.

Methods

Data Sources

We used the "ChestX-ray8" database described previously. It comprises 108,948 frontal-view X-ray images from the year 1992 to 2015. In this project, we selected 21,966 images in total. In 767 of the images they were labeled as "cardiomegaly" and the others were labeled as "healthy". The whole dataset was divided into a training and a testing set. The training set contains 467 images labeled with "cardiomegaly" and 20,899 images labeled with "healthy". The testing set contains 300 images for both "cardiomegaly" and "healthy" categories. We split the training and testing set in this way due to the limited number of images labeled with "cardiomegaly".

Image Pre-processing

Three groups of images were used in our experiments. The first group contains the original image set. For the second group, two steps were implemented to preprocess the original image set. The first step is to crop the image based on the characteristics of cardiomegaly. Since cardiomegaly is closely related to the shapes and sizes of heart and breastbone. We tried to crop the unnecessary background to highlight the heart area. The dimension of the original image is 1024 × 1024. We crop all the images into the dimension of 800 × 800. The second step is to do the histogram equalization for all images [20]. Histogram equalization is a frequently used method for image contrast enhancement. It uses the cumulative distribution function to map the original image histogram to a new image histogram. It stretches the grayscale of the image by normalization. For the third group of images, we transformed the second group of images that contain three color channels into gray-scale images with one color channel.

Transfer Learning

We implemented transfer learning for the image classification task. It is usually unpractical to train the CNN from the beginning because of the insufficient data size and computing power. Thus, the transfer learning becomes an effective method in the image processing domain that could take the advantages of the well-developed models to solve new tasks [21]. Transfer learning mainly has two types when dealing with imaging processing problems. First, a pre-trained model can be used as the feature extractor for the new dataset. Once the features are extracted, a linear classifier can be trained for the new task. The second type is to retrain the fully-connected layer on top of the CNN on the new dataset while to fine-tune the weights of the pre-trained network through backpropagation. Considering the limited computing power and the size of our dataset, we used the first method in our study. Three models with weights pre-trained on ImageNet, including the InceptionV3 [10], ResNet50 [12] and Xception [19], were applied to extract the features from the training set of X-ray images and trained a classifier based on the different combinations of the features. The pipeline of our experiment was shown in Figure 2. Basically, for each experiment group,

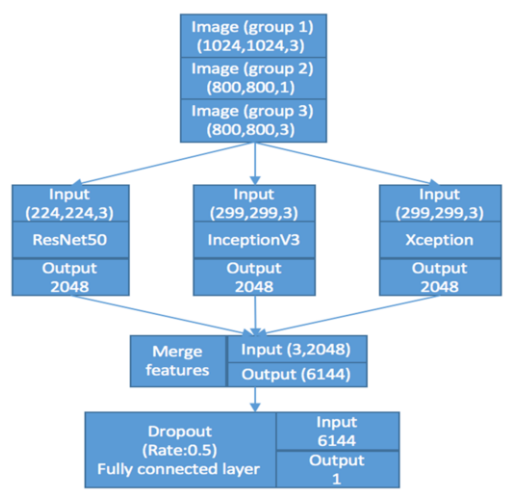

Figure 2–The Experiment Pipeline for Transfer Learning.

we used the three pre-trained models to extract the features, and trained the classifiers based on features extracted by one

model, different combinations of two models and the combination of three models.

Evaluation

We evaluated the performances of the trained classifiers using different combinations of features extracted by the pre-trained models as illustrated in Figure 2. We also tested the scaled and non-scaled features as the comparison. We used the accuracy, F-1 score and the area under the curve (AUC) to evaluate the classification results using the test set mentioned previously.

Results

Images after Pre-processing

Examples of images labeled healthy and cardiomegaly from three experiment groups are shown in Figure 3. The original images, cropped images that focus on the heart area, images after histogram equalization based on cropped images, and one-channel gray-scale image based on all the previous operations are shown in order. After cropping, many unrelated areas could be removed. The histogram equalization highlights the heart area, which could help the following image classification task.

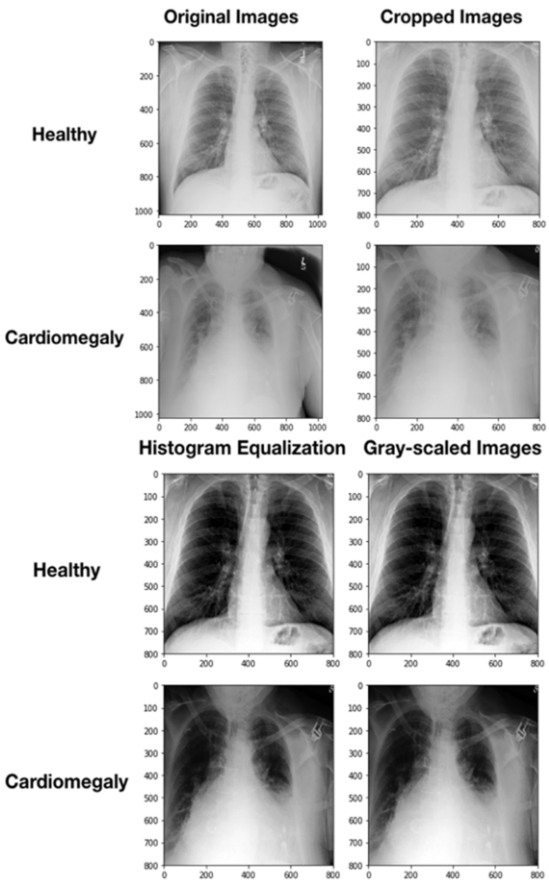

Figure 3– X-ray Images after Cropping and Histogram Equalization.

Classification Results Using Different Pre-trained Models

The results for identifying cardiomegaly from X-ray images are shown in table 1-3. For each experiment group, the results using separate and combined pre-trained models are listed. The scaled features were compared with non-scaled features.

Table 1– Image Classification Results for Group 1

Model	Accuracy	F-1 Score	Scaled-feature
InceptionV3	0.725	0.720	Yes
InceptionV3	0.792	**0.790**	No
ResNet-50	0.695	0.690	Yes
ResNet-50	0.747	0.740	No
Xception	0.673	0.670	Yes
Xception	0.712	0.700	No
InceptionV3 &ResNet-50	0.750	0.750	Yes
InceptionV3 &ResNet-50	0.778	0.780	No
InceptionV3 &ResNet-50 &Xception	0.735	0.730	Yes
InceptionV3 &ResNet-50 &Xception	0.782	0.780	No

Table 2– Image Classification Results for Group 2

Model	Accuracy	F-1 Score	Scaled-feature
InceptionV3	0.723	0.720	Yes
InceptionV3	0.795	0.790	No
ResNet-50	0.693	0.690	Yes
ResNet-50	0.765	0.760	No
Xception	0.655	0.650	Yes
Xception	0.708	0.700	No
InceptionV3 &ResNet-50	0.737	0.740	Yes
InceptionV3 &ResNet-50	0.797	**0.800**	No
InceptionV3 &ResNet-50 &Xception	0.726	0.720	Yes
InceptionV3 &ResNet-50 &Xception	0.792	0.790	No

After initial experiments, we found that the Xception model always had the worst performances and did not improve the performances in combined features. Thus, we stopped using Xception model separately in group 3.

Table 3– Image Classification Results for Group 3

Model	Accuracy	F-1 Score	Scaled-feature
InceptionV3	0.715	0.710	Yes
InceptionV3	0.782	0.780	No
InceptionV3 &ResNet-50	0.748	0.750	Yes
InceptionV3 &ResNet-50	0.793	**0.790**	No
InceptionV3 &ResNet-50 &Xception	0.715	0.710	Yes
InceptionV3 &ResNet-50 &Xception	0.790	**0.790**	No

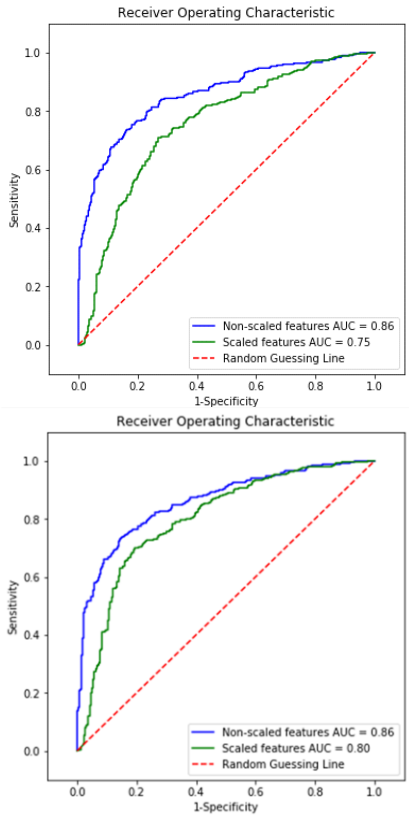

Figure 4–The Selected Best ROC Curves in Experiments.

The group 1 used original images, the features extracted by the combination of InceptionV3obtained the best results (F-1 = 0.79). The images in group 2 were cropped and stretched through histogram equalization. The combination of InceptionV3 and ResNet-50 models obtained the best results (F-1 = 0.800). The group 3 used gray-scale image with only one color channel. The combination of InceptionV3 and ResNet-50 models and the combination of InceptionV3, ResNet-50, and X-ception both reached the best results (F-1 = 0.790).

The selected best ROC curves are shown in Figure 4. The upper figure shows the ROC curve for using the only InceptionV3 model in group 3, and the lower figure shows the ROC curve for using the combination of InceptionV3 and ResNet-50 models in group 2. Our best performing experiment obtains the AUC = 0.860. Both the individual InceptionV3 model and the combination of InceptionV3 and ResNet-50 models reach the best results. The performances of experiments using the Xception model are not ideal, usually with F-1 around 0.700.

Discussion

This study explores the feasibility of applying transfer learning methods to identify the cardiomegaly from X-ray images. We used the pre-trained CNN models based on the ImageNet to extract the features from the X-ray images and obtained promising results (AUC = 0.860). In Wang's et al. work [16], they used the ResNet-50 model and tried to finish the similar task. They obtained the best result with AUC = 0.814 for identifying the cardiomegaly disease from X-ray images. Our method out-performed their results. Although the experimental scenarios are not identical, our task is binary classification while their task is multi-classification.

The three pre-trained CNN models in our study are InceptionV3, ResNet-50, and Xception. When using the model separately, the features extracted by the InceptionV3 obtained the best performances in all experiment groups, while the Xception is not proper for the X-ray images compared to others. The combination of InceptionV3 and ResNet-50 models slightly improve the performances. The best F-1 score is 0.800 and AUC is 0.860 using the combination of the two models. However, using the combined features does not mean better results. In group 1, the InceptionV3 itself has the highest accuracy than other combined models.

We used three experiment groups in the study, each group has images with different pre-processing methods. All the pre-processing steps used classic methods and obtained better images visually. However, the results show that compared with the average accuracy and F-1 score, there's no much difference among three groups, which indicates the pre-trained CNN models extract the features in a way that not influence by the image pre-processing process used in the study. Considering the property of X-ray images and the cardiomegaly disease, we also tried other pre-processing methods. For instance, to highlight the edges of the images, however, due to the unstable qualities of the images, the method cannot apply to all the images. Figure 5 shows the effect of the edge detection, the upper images show successful examples of pre-processing to highlight the breastbone and heart area, while the lower images show the bad examples. Since the quality of the X-ray images varies a lot, more powerful image pre-processing methods would be necessary to improve the results in future studies. Besides, the normalization of extracted features was conducted in all experiment groups. Usually, the feature normalization would improve the performances of neural networks in image classification, since it could reduce the noise of the images by bringing the intensity values to the normal distribution. We used the 0-1 normalization in the study, however, it has side-effects and harms the performances of the model.

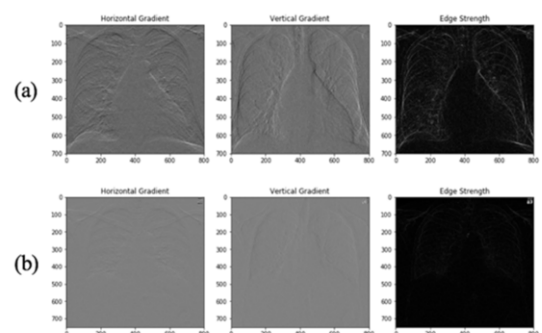

Figure 5–Successful and Failed Edge Detection Examples.

Limitations & Future Work

There are mainly two limitations for the study. First, is the unbalanced dataset. We have 767 X-ray images labeled with "cardiomegaly" and 20,899 images labeled with "healthy". Considering the difficulty of obtaining the annotated X-ray images, it is already the most comprehensive dataset we could obtain. The second limitation is related to the transfer learning method. We used the pre-trained CNN models based on the

ImageNet. Theoretically, these models only work best when our dataset is similar to ImageNet data. The ImageNet data contains over 14 million images in our daily life, but the X-ray image is a special type of image. Currently, there's no pre-trained model based on X-ray images. In the future, we will try to improve the performance by retraining the Inception V3 model. Also, we plan to work with the clinical department to evaluate if our method could improve the efficiency of manual review of X-Rays in clinical settings.

Conclusions

In this study, we applied the transfer learning method to identify the cardiomegaly from X-ray images. We obtained the best performance using the combination of InceptionV3 and ResNet-50 pre-trained models, with AUC of 0.860 that surpass the state-of-the-art results in related work. This result shows the feasibility of using transfer learning methods to develop the CAD system for X-rays.

Acknowledgments

The authors declare that they have no competing interest. Thanks Dr. Luca Giancardo for preparing the image dataset.

References

[1] Mayo Clinic, Enlarged Heart, [Internet]. 2017 [cited 15 November 2018]. Available from : https://www.mayoclinic.org/diseases-conditions/enlarged-heart/symptoms-causes/syc-20355436

[2] O. Er, N. Yumusak, and F. Temurtas, Chest diseases diagnosis using artificial neural networks, *Expert Systems with Applications* **37** (2010), 7648-7655.

[3] Y. Bar, I. Diamant, L. Wolf, S. Lieberman, E. Konen, and H. Greenspan, Chest pathology detection using deep learning with non-medical training, in: *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, 2015, 294-297.

[4] A.A. El-Solh, C.-B. Hsiao, S. Goodnough, J. Serghani, and B.J. Grant, Predicting active pulmonary tuberculosis using an artificial neural network, *Chest* **116** (1999), 968-973.

[5] W.H. Frishman, J. Nadelmann, W.L. Ooi, S. Greenberg, M. Heiman, S. Kahn, H. Guzik, E.J. Lazar, and M. Aronson, Cardiomegaly on chest x-ray: Prognostic implications from a ten-year cohort study of elderly subjects: A report from the Bronx Longitudinal Aging Study, *American Heart Journal* **124** (1992), 1026-1030.

[6] R.H. Abiyev and M.K.S. Ma'aitah, Deep Convolutional Neural Networks for Chest Diseases Detection, *J Healthc Eng*, (2018).

[7] L. Sorensen, S.B. Shaker, and M.d. Bruijne, Quantitative Analysis of Pulmonary Emphysema Using Local Binary Patterns, *IEEE Trans Med Imaging* **29** (2010), 559-569.

[8] L. Sorensen, S.B. Shaker, and M. de Bruijne, Texture classification in lung CT using local binary patterns, *Med Image Comput Comput Assist Interv* **11** (2008), 934-941.

[9] W. Yang, Z. Lu, M. Yu, M. Huang, Q. Feng, and W. Chen, Content-based retrieval of focal liver lesions using bag-of-visual-words representations of single- and multiphase contrast-enhanced CT images, *J Digit Imaging* **25** (2012), 708-719.

[10] M.D. Kumar, M. Babaie, S. Zhu, S. Kalra, and H. Tizhoosh, A comparative study of CNN, BoVW and LBP for classification of histopathological images, *arXiv preprint arXiv:1710.01249* (2017).

[11] A. Krizhevsky, I. Sutskever, and G.E. Hinton, ImageNet classification with deep convolutional neural networks, *Commun. ACM* **60** (2012), 84-90.

[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, Going deeper with convolutions, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, 1-9.

[13] Y. Bar, I. Diamant, L. Wolf, and H. Greenspan, *Deep learning with non-medical training used for chest pathology identification*, 2015.

[14] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, Imagenet large scale visual recognition challenge, *International Journal of Computer Vision* **115** (2015), 211-252.

[15] P. Lakhani, Deep Convolutional Neural Networks for Endotracheal Tube Position and X-ray Image Classification: Challenges and Opportunities, *J Digit Imaging* **30** (2017), 460-468.

[16] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R.M. Summers, Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, IEEE, 2017, 3462-3471.

[17] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, *arXiv preprint arXiv:1409.1556* (2014).

[18] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, 770-778.

[19] F. Chollet, Xception: Deep learning with depthwise separable convolutions, *arXiv preprint* (2017), 1610.02357.

[20] Y.C. Hum, K.W. Lai, and M.I. Mohamad Salim, Multiobjectives bihistogram equalization for image contrast enhancement, *Complexity* **20** (2014), 22-36.

[21] S.J. Pan and Q. Yang, A survey on transfer learning, *IEEE Transactions on knowledge and data engineering* **22** (2010), 1345-1359.

Address for correspondence

Rui Zhang, PhD
MMC 912, 420 Delaware Street SE
Minneapolis, MN 55455, USA
zhan1386@umn.edu;
612-626-4209