Enhancing Transferability of Features from Pretrained Deep Neural Networks for Lung Nodule Classification

Hongming Shan, Ge Wang, Mannudeep K. Kalra, Rodrigo Canellas de Souza, Junping Zhang

***Abstract*—Among most popular feature extractors, pretrained deep neural networks play a central role in transfer learning to extract high-level feature on small datasets. The transferable performance, however, cannot be guaranteed for the task of interest. To enhance the transferability, this paper employs fine-tuning and feature selection in a different way to improve the accuracy of lung nodule classification. The fine-tuning technique retrains the neural network using lung nodule dataset, while feature selection captures a useful subset of features for lung nodule classification. Preliminary experimental results on CT images from Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) confirm that the classification accuracy on lung nodule can be significantly improved via fine-tuning and feature selection. Furthermore, the results outperform competitively handcrafted texture descriptors.**

***Index Terms*—Deep learning, lung nodule classification, fine-tuning technique, feature selection**

I. Introduction

LUNG cancer is an aggressive and heterogeneous disease with a low long-term survival rate [1]. Computed Tomography (CT) is the imaging modality of choice for evaluation of patients with suspected or known lung cancer. CT is also the preferred modality for screening lung cancer, which often present as lung nodules. Unfortunately, many lung nodules are benign in etiology. Radiologists rely on several qualitative and quantitative factors to describe pulmonary nodules such as nodule size, shape, margin, attenuation and location in the lungs [2]. One of the critical nodule characteristics is the classification between malignant and benign nodules, which facilitates nodule staging assessment and consequent therapeutic planning [3], [4], [5].

Previous nodule analysis, mostly based on handcrafted texture feature extractors [3], [6], [7], suffers from the need of specialized knowledge in selecting parameters and robustness to different datasets. Motivated by the successful applications of deep neural networks (DNNs) to image classification [8], [9], [10], the deep features extracted from DNNs are more general and of high-level as compared with handcrafted ones [11]. Training DNNs, however, requires massive data for avoiding overfitting, which is often infeasible for a small dataset such as LIDC-IDRI [12], even data augmentation techniques are adopted in the training phase [8]. One popular way to apply DNNs to the small datasets is the *transfer learning*, taking a pretrained DNN on a large-scale dataset as a feature extractor for a task of interest [11], [13]. In particular, success has been found in transferring knowledge from general object recognition tasks to classification tasks in which their categories are similar [11].

In this paper, we apply the transfer learning from pretrained DNNs on the large-scale image classification dataset ImageNet [14] for our lung nodule classification. One problem is that we do not know whether deep features directly extracted from pretrained DNNs are suitable for our lung nodule classification. To improve the transferability, we employ the fine-tuning and feature selection techniques to make deep features more suitable for lung nodule classification. More specifically, the fine-tuning technique retrains DNNs using lung nodule data, and feature selection captures a useful subset of features for lung nodule classification. Experimental results confirm that the classification performance can be improved through fine-tuning and feature selection techniques. Furthermore, our results outperform handcrafted texture descriptors.

II. Methodology

A. Data preparation

The LIDC-IDRI dataset [12] consists of diagnostic and lung cancer screening thoracic CT scans with annotated lung nodules from a total number of $1,010$ patients. Furthermore, each nodule was rated from 1 to 5 by four experienced thoracic radiologists, indicating an increasing probability of malignancy. In this study, the ROI of each nodule was obtained along with its annotated center in accordance with the nodule report[1], with a square shape of a doubled equivalent diameter. An average score of a nodule was used for assigning probability of malignant etiology [3], [4]. Nodules with an average score higher than 3 were labeled as *malignant*. Nodules with an average score lower than 3 were labeled as *benign*. Some nodules were removed from the experiments in the case of the averaged malignancy score 3, ambiguous IDs, and being rated

[1]http://www.via.cornell.edu/lidc/

This work was supported by the National Natural Science Foundation of China (No. 61673118) and Shanghai Pujiang Program (NO. 16PJD009). (*Co-corresponding authors: Ge Wang and Junping Zhang.*)

H. Shan and J. Zhang are with Shanghai Key Laboratory of Intelligent Information Processing and the School of Computer Science, Fudan University, Shanghai, China, 200433. E-mails: {hmshan,jpzhang}@fudan.edu.cn.

G. Wang is with Biomedical Imaging Cluster, Department of Biomedical Engineering, Center for Biotechnology and Interdisciplinary Studies, Rensselaer Polytechnic Institute, Troy, NY, USA. E-mail: wangg6@rpi.edu.

M. Kalra and R. Canellas de Souza are with Department of Radiology, Massachusetts General Hospital, Boston, MA, USA. E-mails:{mkalra,rcsouza}@mgh.harvard.edu.

DOI: 10.12059/Fully3D.2017-11-3110002

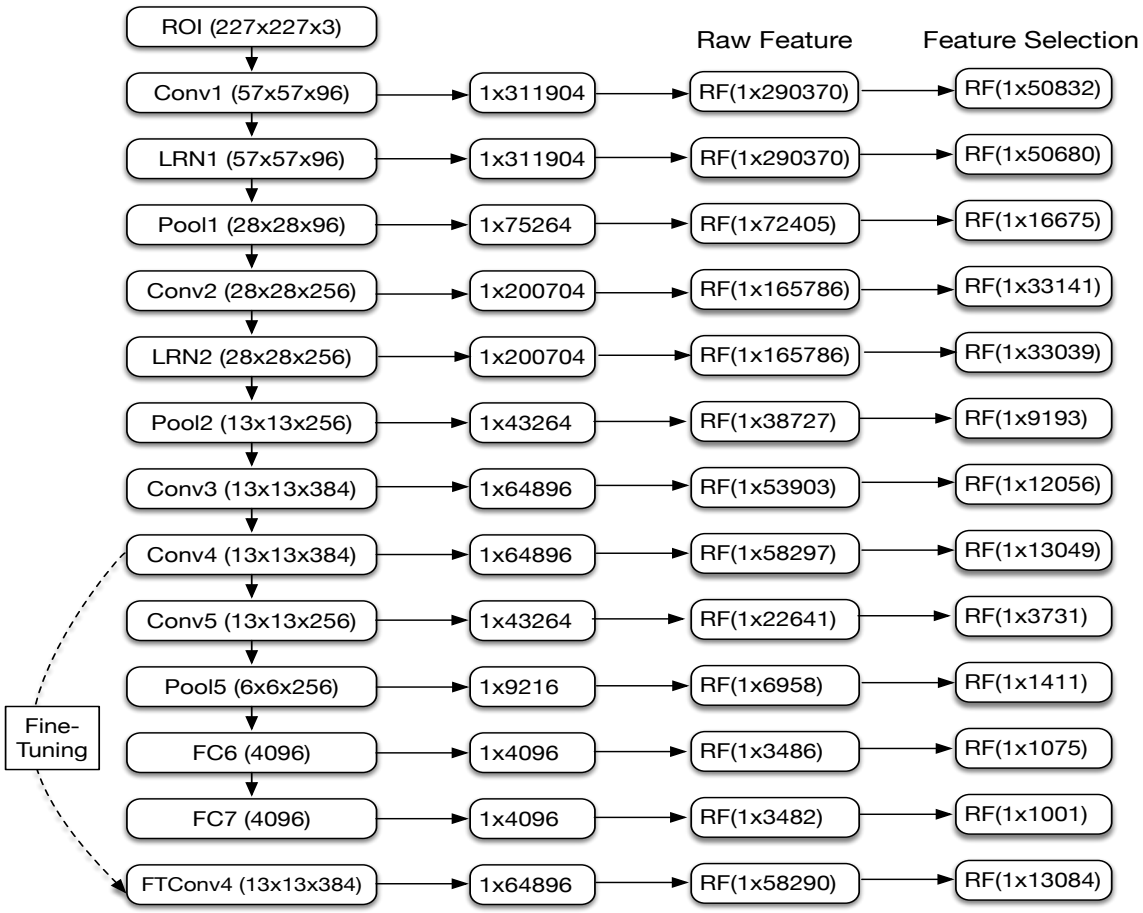


Fig. 1. Schematic of extracting deep features from the pretrained AlexNet through fine-tuning technique and feature selection. Columns from left to right indicate the architecture of AlexNet, flatten deep features, deep features with eliminating all zero-variance columns (Raw feature), and deep features after feature selection. The last row is the fine-tuned *Conv4*. Note that the deep feature at the lower right corner was obtained by 1) fine-tuning *Conv4*, 2) eliminating zero-variance columns, and 3) extracting a subset through feature selection.

by only one radiologist. To sum up, there were 959 benign nodules and 575 malignant ones. The size of benign ROIs ranged from 8 to 92 pixels, with a mean size of 17.3 and a standard deviation of 7.0 pixels. The size of malignant ROIs ranged from 12 to 95 pixels, with a mean size of 35.4 and a standard deviation of 15.8 pixels.

B. Pretrained AlexNet

AlexNet was a classical convolutional neural network model [8], including five convolutional layers, three pooling layers, two local response normalization (LRN) layers, and three fully connected layers. A publicly available version of AlexNet was pretrained on the large-scale ImageNet dataset [14], which contains $1M$ images and $1K$ classes. The weights of pretrained AlexNet were preinitialized and can be downloaded from the Caffe website[2], which was used in our experiments.

The pretrained AlexNet was used to extract deep features from ROIs of the lung nodules. After removing the last fully connected layer for classification into $1K$ classes, each layer of the AlexNet would be a feature extractor. This is to say that 12 different deep features can be extracted from one ROI. The process of extracting features was depicted in Fig. 1. The first column indicated the architecture of AlexNet, and the numbers in the second column denoted the dimensions of flatten features extracted from all the layers of AlexNet. Those flatten features after eliminating all zero-variance columns were used to train Random Forest (RF) classifiers [15], which were in the third column and called *raw features*.

Yosinski *et al.* reported that those deep features extracted from earlier layers of deep neural networks were more generalizable (*e.g.* edge detectors or color blob detectors) that should be useful to many tasks [11]. Those features extracted from later layers, however, become progressively more specific to the details of the classes contained in the original dataset. In the case of ImageNet, which includes many dog breeds, a significant portion of the representation power of AlexNet may be devoted to features that are specific to differentiating between dog breeds. Due to the difference between our lung nodule dataset and ImageNet, we were not sure which layer would be more suitable for lung nodule classification. There-

[2]http://dl.caffe.berkeleyvision.org/

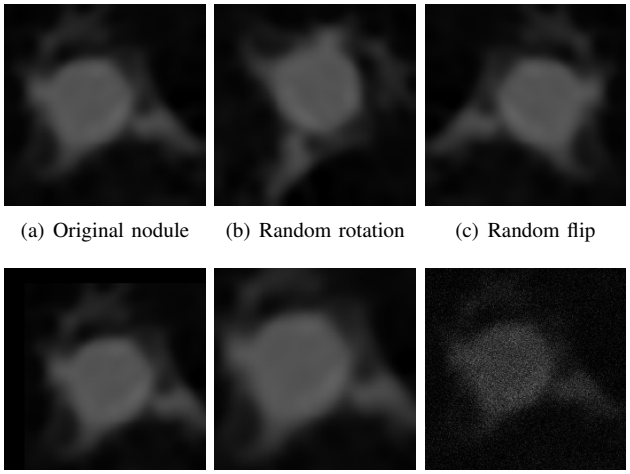
fore, features from all the layers were evaluated in Section III.

It should be noted that a pretrained neural network does not contain any specific information about lung nodule. To enhance the transferability from the pretrained AlexNet, the following two subsections will describe the fine-tuning technique and feature selection for adapting to lung nodule classification.

C. Fine-tuning AlexNet

As a popular strategy for transfer learning, *fine-tuning* AlexNet was not only to replace and retrain the classifier on the top of AlexNet using the lung nodule dataset but also to fine-tune the weights of the pretrained AlexNet through the backpropagation.

In view of the classification accuracy reported in Section III, feature obtained from *Conv4* was more suitable for lung nodule classification than those of other layers. In this study, we replaced the layers after *Conv4* with a fully connected layer as the binary classifier. Due to the concern of overfitting, we only tuned *Conv4* and enlarged lung nodule data for retraining. Methods for enlarging lung nodule data included random rotation, random flip, random shift, random zoom, and random noise. Fig. 2 presented the data augmentation results for a lung nodule in the experiments.



(a) Original nodule (b) Random rotation (c) Random flip

(d) Random shift (e) Random zoom (f) Random noise

Fig. 2. Illustration of data augmentation for a lung nodule.

The fine-tuned *Conv4*, called *FTConv4*, was at the last row in Fig. 1.

D. Feature selection

The another strategy for extracting information from *raw features* is *feature selection*, which was in the last column of Fig. 1.

Deep features extracted from AlexNet suffer from the curse of the dimensionality and are redundant to lung nodule classification, even after *Conv4* was fine-tuned with our lung nodule dataset. Here we took *Conv4* as an example. After removing the zero-variance columns, one ROI was represented by a $58,297$-dimensional vector. Using the feature importance measurement provided by RF classifier, there were $26,842$ columns with feature importances of zero to lung nodule classification as shown in Fig. 3. This is to say, almost half of features extracted from *Conv4* were irrelevant to the classification of lung nodules.

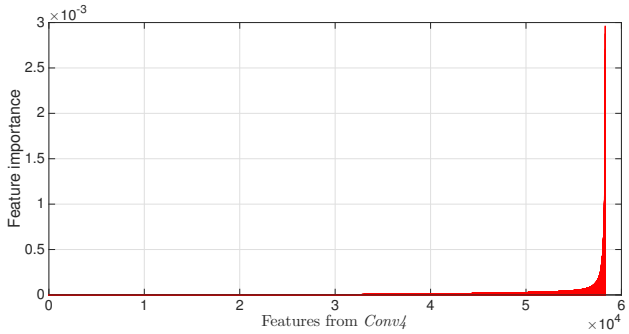


Fig. 3. Feature importance of deep features extracted by *Conv4*. Note that we sort the columns of features in ascending fashion according to their importance scores.

In other words, we computed the feature importances with the random forest classifier, which in turn were used to discard irrelevant features. Those columns with importance scores higher than the averaged importance score were kept as the relevant features for lung nodule classification as shown in the last column in Fig. 1.

III. Experimental results

A. Experimental setup

Each ROI was up-sampled into $227 \times 227 \times 3$ and then fed into AlexNet. It should be noted that each ROI had three channels despite being grayscale to fit the AlexNet which was originally designed for color images. For evaluating the performance of extracted features, ROIs were randomly divided into a training set with 60% lung nodules and a testing set with remaining lung nodules. We trained the random forest classifier on the training set and computed the classification accuracy on the testing test. The reported results were averaged on 50 repetitions. The RF classifier was taken from the scikit-learn package [16].

B. Classification after enhancing transferability of AlexNet

Fig. 4 presents the classification accuracies with raw features extracted from each layer of the pretrained AlexNet and fine-tuned *Conv4* on our lung nodule dataset as well as the deep features after feature selection. As shown in Fig. 4, the features extracted from *Conv4* outperform those from the other layers. Features from layers earlier than *Conv4* were insufficient to characterize the lung nodules, and features from layers later than *Conv4* were more specific to their original dataset, leading to slight performance decrement.

To enhance the transferability of the pretrained AlexNet, Fig. 4 also presented the classification accuracies using fine-tuning and feature selection techniques. The results from fine-tuned *Conv4*, *FTConv4*, were shown in the rightmost of

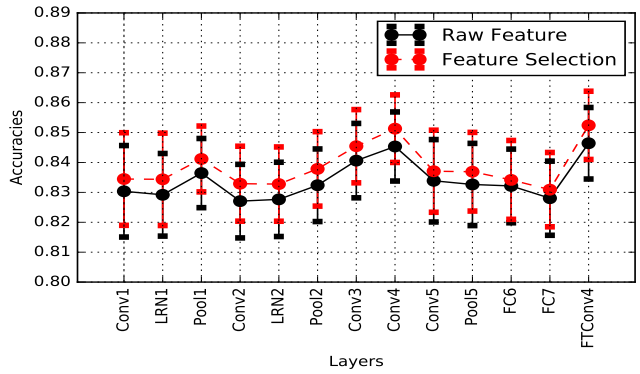


Fig. 4. Classification accuracy with deep features extracted from each layer of AlexNet.

Fig. 4. After fine-tuning AlexNet on the lung nodule data, the classification accuracy was slightly improved compared to *Conv4*. However, feature selection can significantly improve the classification accuracy. The reasons were two-folds. On one hand, compared with the parameters of *Conv4*, the lung nodule dataset was still too small to fine-tune AlexNet. Most of features extracted from *Conv4* were irrelevant to lung nodule classification, which increased the difficulty in retraining AlexNet. On the other hand, feature selection can remove those irrelevant features and extract a useful subset of features for classification. The best classification accuracy was achieved with the deep features from *FTConv4* after feature selection.

C. Comparison with baseline algorithms

We compared our results with two handcrafted competing texture descriptors including the local binary pattern (LBP) [6] and the histogram of gradient (HOG) [7]. LBP and HOG were sensitive to window size and number of neighborhood points respectively. We used 3-fold cross-validation to tune these two parameters. The averaged results are in Table I. LBP and HOG were copied from the scikit-image package [17].

TABLE I
COMPARISON WITH BASELINE ALGORITHMS

Method	LBP	HOG	FTConv4
Accuracy	0.799±0.013	0.838±0.012	**0.852±0.011**

It can be seen that the feature extracted from *Conv4* with fine-tuning and feature selection outperforms the traditional handcrafted texture descriptors.

IV. Conclusion and Discussion

This paper studied the transfer learning from a pretrained AlexNet that for lung nodule classification. To enhance the transferability, we applied fine-tuning and feature selection techniques to retrain AlexNet and extract a useful feature subset. The best classification accuracy of lung nodule was achieved by fine-tuning *Conv4* with feature selection. Our results outperform two handcrafted texture descriptors.

Although recently proposed deep neural networks such as GoogleNet [9] and ResNet [10] performed better than AlexNet for ImageNet classification, AlexNet was used in our experiments due to its simplicity and rich literature. It is also interesting to apply other DNNs for lung nodule classification. Instead of extracting high-level features from DNNs, 'end-to-end' deep learning algorithms deserve further investigation for lung nodule classification. We are working along these lines.

References

[1] N. L. S. T. R. Team *et al.*, "Reduced lung-cancer mortality with low-dose computed tomographic screening," *The New England Journal of Medicine*, vol. 2011, no. 365, pp. 395–409, 2011.

[2] H. J. Aerts, E. R. Velazquez, R. T. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld *et al.*, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature Communications*, vol. 5, 2014.

[3] F. Han, G. Zhang, H. Wang, B. Song, H. Lu, D. Zhao, H. Zhao, and Z. Liang, "A texture feature analysis for diagnosis of pulmonary nodules using LIDC-IDRI database," in *International Conference on Medical Imaging Physics and Engineering*. IEEE, 2013, pp. 14–18.

[4] W. Shen, M. Zhou, F. Yang, C. Yang, and J. Tian, "Multi-scale convolutional neural networks for lung nodule classification," in *International Conference on Information Processing in Medical Imaging*. Springer, 2015, pp. 588–599.

[5] D. Kumar, A. Wong, and D. A. Clausi, "Lung nodule classification using deep features in CT images," in *12th Conference on Computer and Robot Vision*. IEEE, 2015, pp. 133–138.

[6] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1. IEEE, 2005, pp. 886–893.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[9] 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, pp. 1–9.

[10] 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, pp. 770–778.

[11] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in Neural Information Processing Systems*, 2014, pp. 3320–3328.

[12] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman *et al.*, "The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans," *Medical physics*, vol. 38, no. 2, pp. 915–931, 2011.

[13] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 806–813.

[14] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[15] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[17] S. Van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu, "Scikit-image: Image processing in python," *PeerJ*, vol. 2, p. e453, 2014.