Classifying Pneumonia among Chest X-Rays Using Transfer Learning*

Abdullah Irfan[1], Akash L. Adivishnu[2], Antonio Sze-To[3⋆],
Taher Dehkharghanian[4], Shahryar Rahnamayan[5] and H.R. Tizhoosh[6]

Abstract— Chest radiography has become the modality of choice for diagnosing pneumonia. However, analyzing chest X-ray images may be tedious, time-consuming and requiring expert knowledge that might not be available in less-developed regions. therefore, computer-aided diagnosis systems are needed. Recently, many classification systems based on deep learning have been proposed. Despite their success, the high development cost for deep networks is still a hurdle for deployment. Deep transfer learning (or simply transfer learning) has the merit of reducing the development cost by borrowing architectures from trained models followed by slight fine-tuning of some layers. Nevertheless, whether deep transfer learning is effective over training from scratch in the medical setting remains a research question for many applications. In this work, we investigate the use of deep transfer learning to classify pneumonia among chest X-ray images. Experimental results demonstrated that, with slight fine-tuning, deep transfer learning brings performance advantage over training from scratch. Three models, ResNet-50, Inception V3 and DensetNet121, were trained separately through transfer learning and from scratch. The former can achieve a 4.1% to 52.5% larger area under the curve (AUC) than those obtained by the latter, suggesting the effectiveness of deep transfer learning for classifying pneumonia in chest X-ray images.

I. Introduction

Pneumonia is a pathogenic infection of the lung parenchyma, which is most commonly caused by bacteria or viruses, and less commonly by other microorganisms such as fungi [1]. The damages caused by the infection and the host's immune response results in lung injury and disruption of pulmonary functions. More than one million adults are hospitalized with pneumonia and around 50,000 die from the disease every year in the US alone [2]. Pneumonia is a higher burden in low-income countries where it is the leading cause of death [3], where there is limited access to diagnostic and therapeutic facilities.

Chest radiography is a widely available image modality; therefore, it has become the imaging of choice for diagnosing pneumonia and many other thoracic conditions. The most recognizable radiographic diagnostic sign of pneumonia is the presence of "infiltrates" in a chest X-ray, which are opacities caused by the accumulation of pus, blood or other fluids, denser than air, in the lung parenchyma and alveolar spaces. Analyzing chest X-ray images requires expert knowledge that might not be available in less-developed or remote regions of the world. Also, it is a tedious and time-consuming task. Thus, computer-aided diagnosis systems could become viable tools for addressing these issues.

Since the availability of publicly accessible datasets, such as ChestX-ray14 [4] and CheXpert [5], some classification systems based on deep learning techniques on chest X-ray images have been proposed [4], [2], [5].

Deep transfer learning (or simply transfer learning) [6], has become the most popular method of choice for the implementation of deep learning for medical image analysis because this approach requires less computational power and few to none training samples. It is the practice of employing the structure and corresponding weights of pre-trained models for tasks and datasets for which they have not been originally designed. Pre-trained models are either used for feature extraction or fine-tuned to perform new tasks.

The opposite to deep transfer learning is training from scratch, i.e. the model weights are randomly initialized. For example, Stephen et al. constructed their own convolutional neural network (CNN) model that achieved high classification accuracy for pneumonia on chest x-ray images [7]. Its crucial drawback is its high development cost, making it a hurdle for deployment.

In medical image analysis, deep transfer learning has been applied to a wide range of diagnostic modalities , including chest radiographs, [2], computerized tomography (CT) scans [8], retinal fundus [9], and histopathology images [10]. A study on breast histopathology images demonstrated that deep transfer learning could yield higher classification accuracy compared to models

*This work was partly supported by a Vector Institute PathFinder grant

⋆Antonio Sze-To is the corresponding author

[1]Abdullah Irfan is with Systems Design Engineering, University of Waterloo, 200 University Ave W, Waterloo, ON, Canada N2L 3G1 a5irfan@uwaterloo.ca

[2]Akash L. Adivishnu is with Systems Design Engineering, University of Waterloo, 200 University Ave W, Waterloo, ON, Canada N2L 3G1 aladivis@uwaterloo.ca

[3]Antonio Sze-To is with Kimia Lab, University of Waterloo, 200 University Ave W, Waterloo, ON, Canada N2L 3G1 hy2szeto@uwaterloo.ca

[4]Taher Dehkharghanian is with NICI Lab, University of Ontario Institute of Technology, 2000 Simcoe St N, Oshawa, ON, Canada L1G 0C5 taher.dehkharghanian@uoit.ca

[5]Shahryar Rahnamayan is with NICI Lab, University of Ontario Institute of Technology, 2000 Simcoe St N, Oshawa, ON, Canada L1G 0C5 shahryar.rahnamayan@uoit.ca

[6]Hamid R. Tizhoosh is with Kimia Lab, University of Waterloo, 200 University Ave W, Waterloo, ON, Canada N2L 3G1, and Vector Institute, 661 University Ave Suite 710, Toronto, ON, Canada M5G 1M1 hamid.tizhoosh@uwaterloo.ca

978-1-7281-1990-8/20/$31.00 ©2020 IEEE

that have been built from the scratch [10]. However, a comprehensive study [11] of deep transfer learning for medical image analysis reveals that transfer learning does not significantly improve classification performance in the medical setting, compared to random weight initialization. However, transfer learning is generally Superior to conventional feature extraction methods [12].

Therefore, whether deep transfer learning is generally more effective compared to training from scratch in the medical setting is still subject to research. The objective of this study is to investigate the use of deep transfer learning to classify pneumonia among chest X-ray images. Experimental results demonstrated that, with small training epochs, deep transfer learning can bring performance advantages over training from scratch. Three models, ResNet-50, Inception V3 and DensetNet121, were trained separately with 20 epochs using deep transfer learning and from scratch. Transfer learning can achieve a 4.1% to 52.5% higher AUC value than training from scratch.

II. Methodology

Given a (frontal) chest X-ray image, the problem is to output a probability of the presence of pneumonia (see Fig. 1). In other words, it is equivalent to the binary classification of pneumonia (yes/no) from an input chest X-ray image.

In this study, we used transfer learning to solve the problem. The idea of transfer learning is to fine-tune existing models which have been pre-trained on other datasets for the specific classification tasks. This approach makes the training task less computationally expensive, since salient features of a chest X-ray image have already been learned by the model.

A. Model Architecture

A model architecture for deep transfer learning, with two configurations, is explored for pneumonia classification. An illustration of the model architecture, with two configurations, is depicted in Fig. 2.

1) Layer 0 - Existing model: An existing model is obtained with either Configuration A - initialized with random weights or Configuration B - initialized with ImageNet weights. All layers of the existing model are then set as non-trainable. The last 5 layers are then removed.

2) Layer 1 - Global Average Pooling (GAP): Following [13], a GAP layer is added to connect to the existing model.

3) Layer 2 - Dropout Layer: To reduce over-fitting, a dropout layer is added [14]. The dropout parameter is set as 0.2, i.e., 20% of the inputs would be randomly set as zeroes.

4) Layer 3 - Dense Layer: A dense layer is added with 512 neurons (relu activation). To reduce over-fitting, following [15], a L2 regularization factor of 0.0005 is added.

5) Layer 4 - Dropout Layer: To reduce over-fitting, a dropout layer is added [14]. The dropout parameter is set as 0.2, i.e., 20% of the inputs would be randomly set as zeroes.

6) Layer 5 - Classification Layer: For the binary classification of pneumonia, a dense layer with 1 neuron (sigmoid activation) is added.

III. Experiments and Results

In this section, data collection and pre-processing are described. The implementation, parameter setting and performance evaluation are also described, followed by a summary of the experimental results.

A. Data Collection

1) Training and Validation Datasets: ChestX-ray14 dataset [4], containing 112,120 frontal chest X-ray Images of 30,805 unique patients, was obtained. The dataset was annotated with the presence or absence of 14 thoracic pathology labels including pneumonia. Among the 112,120 images, 86,524 of them were allocated to the training list [4]. In this study, 90% of them were used as training and the remaining 10% as validation set. In other words, the training dataset contains 77,872 chest X-ray images, and the validation dataset contains 8,652 chest X-ray images.

2) Testing Dataset 1: The models were tested on the images which were put in the testing list of the ChestX-ray14 dataset [4]. This dataset includes 25,596 frontal chest X-ray images.

3) Testing Dataset 2: The models were further tested on the validation dataset of CheXpert [5], including 203 frontal chest X-ray images, where each of them has been labeled as pneumonia or non-pneumonia. These labels were verified by three board-certified radiologists [5].

B. Implementation and Parameter Setting

The deep learning library Keras (http://keras.io/) with TensorFlow [16] was adopted for implementation. We set the number of epochs to 20 and the batch size were to 256 images. Before inputting the images into the network, the images were resized to 224×224. During training, images were also augmented with the following parameter setting: samplewise_centre = true, samplewise_std_normalization = true, horizontal_flip = true, vertical_flip = false, height_shift_range = 0.05, width_shift_range = 0.1, rotation_range = 5, shear_range = 0.1, fill_mode = 'reflect', zoom_range = 0.15. The default loss function was binary cross-entropy. The default optimizer was Adam [17]. All experiments were run on a cloud environment provided by Kaggle that has 11 GB RAM, an Intel(R) Xeon(R) CPU @

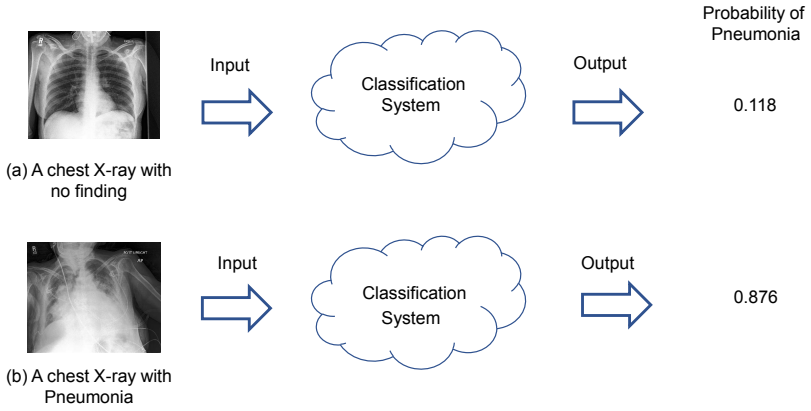

Fig. 1. An overview of classification of pneumonia in chest X-ray images. Given a (frontal) chest X-ray image, a classification system outputs a number between 0 and 1. For illustration, two chest X-ray images were extracted from the validation set in CheXpert [5]. (a) is from the study 1 of patient64544, with the filename as view1 frontal.jpg. No finding was observed. (b) is from the study 1 of patient64552, with the filename as view1 frontal.jpg. Pneumonia was observed. These observations were verified by three board-certified radiologists [5].

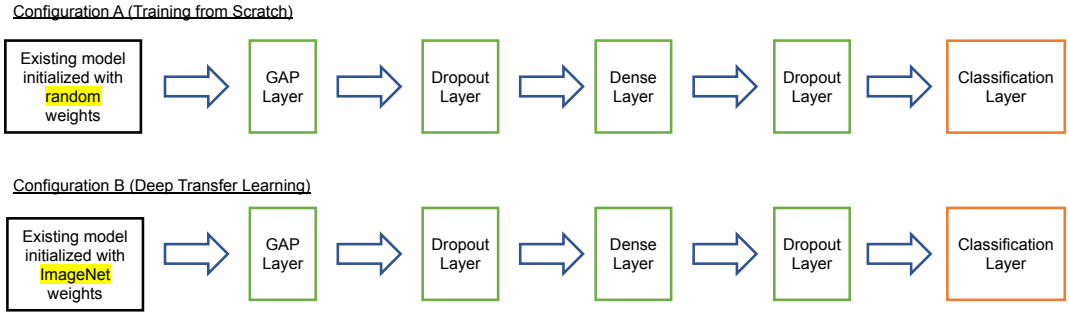

Fig. 2. An illustration of the model architecture with two configurations, in this study. Layer 0 - Existing model: An existing model is obtained with either Configuration A - initialized with random weights or Configuration B - initialized with ImageNet weights. All layers of the existing model are then set as non-trainable. The last 5 layers are then removed. Layer 1 - Global Average Pooling (GAP) Layer. Following [13], a GAP layer is added here to connect to the existing mode. Layer 2 - Dropout Layer. To reduce overfitting, a dropout layer is added [14]. Layer 3 - Dense Layer. A dense layer is added with 512 neurons (relu activation). Layer 4: Dropout Layer. To reduce overfitting, a dropout layer is added [14]. Layer 5: Classification Layer. For the binary classification of pneumonia, a dense layer with 1 neuron (sigmoid activation) is added.

2.30GHz with 16 cores, with the neural networks trained on GPU. Unless further specified, other parameters remained at default settings.

C. Performance Evaluation

Following [4], [5], the performance of models was evaluated by the area under the receiver operating characteristic curve (AUC) to enable the comparison over a range of prediction thresholds.

Three models, ResNet-50 [18], Inception V3 [19] and DenseNet121 [20] were investigated in this study. Each model was trained separately on configurations A and B over the training dataset along with the validation dataset. After training, each model was first evaluated on Testing Dataset 1, followed by Testing Dataset 2. The results are summarized in Table I.

ResNet-50: For results on Testing Dataset 1, ResNet-50 [18] obtained an AUC of 0.59 when trained with configuration B, compared with an AUC of 0.46 obtained by training with configuration A, an improvement of 28.3% in percentage change. For results on Testing Dataset 2, ResNet-50 [18] obtained an AUC of 0.69 when trained with configuration B, compared with an AUC of 0.58 obtained by training with configuration A, an improvement of 19.0% in percentage change.

Inception V3: For results on Testing Dataset 1, Inception V3 [19] obtained an AUC of 0.55 when trained with configuration B, compared with an AUC of 0.51 obtained by training with configuration A, an improvement of 7.8% in percentage change. For results on Testing Dataset 2, Inception V3 [19] obtained an AUC of 0.61 when trained with configuration B, compared with an AUC of 0.40 obtained by training with configuration A, an improvement of 52.5% in percentage change.

DenseNet121: For results on Testing Dataset 1, DenseNet121 [20] obtained an AUC of 0.71 when trained with configuration B, compared with an AUC of 0.57 obtained by training with configuration A, an improvement of 24.6% in percentage change. For results on Testing

Testing Dataset	Experiments		
	Model	Configuration A	Configuration B
1: ChestX-ray14 [4]	ResNet-50 [18]	0.46	0.59
	Inception V3 [19]	0.51	0.55
	DenseNet121 [20]	0.57	0.71
2: CheXpert [5]	ResNet-50 [18]	0.58	0.69
	Inception V3 [19]	0.40	0.61
	DenseNet121 [20]	0.73	0.76

TABLE I

A comparison of AUC (area under the curve) on two testing datasets among three models trained with two configurations

Dataset 2, DenseNet121 [20] obtained an AUC of 0.76 when trained with configuration B, compared with an AUC of 0.73 obtained by training with configuration A, an improvement of 4.1% percentage change.

IV. Conclusions

In this study, we explored the use of deep transfer learning to classify pneumonia in chest X-ray images. In our experiments, three models, namely ResNet-50, Inception V3 and DensetNet121, were trained separately with 20 epochs using deep transfer learning and from scratch. Transfer learning achieved better results than training from scratch, an effect that may be generally expected but needs verification in medical domains. It has thus been demonstrated that with a small number of training epochs, transfer learning can bring performance advantages over training from scratch, when we are dealing with vision-based diagnostic cases such as pneumonia. The results support the effectiveness of transfer learning, providing a low-cost development option for systems based on deep learning for faster and more efficient clinical deployment.

References

[1] S. B. A. Sattar and S. Sharma, "Bacterial pneumonia," 2019.

[2] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya et al., "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," arXiv preprint arXiv:1711.05225, 2017.

[3] P. Moraga, G. . C. of Death Collaborators et al., "Global, regional, and national age-sex specific mortality for 264 causes of death, 1980-2016: a systematic analysis for the global burden of disease study 2016," The Lancet, vol. 390, no. 10100, pp. 1151–1210, 2017.

[4] 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 Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2097–2106.

[5] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya et al., "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," arXiv preprint arXiv:1901.07031, 2019.

[6] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in International Conference on Artificial Neural Networks. Springer, 2018, pp. 270–279.

[7] O. Stephen, M. Sain, U. J. Maduh, and D.-U. Jeong, "An efficient deep learning approach to pneumonia classification in healthcare," Journal of healthcare engineering, vol. 2019, 2019.

[8] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," IEEE transactions on medical imaging, vol. 35, no. 5, pp. 1285–1298, 2016.

[9] J. Y. Choi, T. K. Yoo, J. G. Seo, J. Kwak, T. T. Um, and T. H. Rim, "Multi-categorical deep learning neural network to classify retinal images: A pilot study employing small database," PloS one, vol. 12, no. 11, p. e0187336, 2017.

[10] R. Mehra et al., "Breast cancer histology images classification: Training from scratch or transfer learning?" ICT Express, vol. 4, no. 4, pp. 247–254, 2018.

[11] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning with applications to medical imaging," arXiv preprint arXiv:1902.07208, 2019.

[12] M. D. Kumar, M. Babaie, S. Zhu, S. Kalra, and H. R. Tizhoosh, "A comparative study of cnn, bovw and lbp for classification of histopathological images," in 2017 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, 2017, pp. 1–7.

[13] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2921–2929.

[14] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," The journal of machine learning research, vol. 15, no. 1, pp. 1929–1958, 2014.

[15] 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.

[16] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin et al., "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," arXiv preprint arXiv:1603.04467, 2016.

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 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, pp. 770–778.

[19] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.

[20] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.