SAIS 2024 Proceedings | Southern (SAIS)

Spring 3-16-2024

Trade-offs between Fine-Tuning and Dataset Size in Multistage TL using Lightweight Architecture

Rajesh Godasu

David Zeng

Kruttika Sutrave

Trade-offs between Fine-Tuning and Dataset Size in Multistage TL using Lightweight Architecture

Rajesh Godasu
Dakota State University
Rajesh.godasu@trojans.dsu.edu

David Zeng
Dakota State University
David.zeng@dsu.edu

Kruttika Sutrave
Grand Valley State University
sutravek@gvsu.edu

ABSTRACT

Lightweight Convolutional Neural Networks (CNNs) and Transfer Learning have emerged as efficient alternatives to resource intensive deep CNNs. Lightweight CNNs demand less training time due to their reduced parameter count and adapt well to resource-constrained devices. Nonetheless, there exists a knowledge gap regarding the influence of dataset size on CNN performance for specific problems. This study bridges this gap by offering insights into the ideal dataset sizes for Transfer Learning approaches in medical image-related tasks. Hence, our study focused on Multi-stage Transfer Learning, utilizing chest X-ray imaging modality and lightweight CNNs to explore the effects of fine-tuning CNNs on varying dataset sizes. The results suggest that the models trained on the dataset size range of 30,000 – 40,000 images with 90 – 95 layers frozen in stage 1 transfer learning have led to the highest target validation accuracy in stage 2 transfer learning.

Keywords

Lightweight Convolutional Neural Networks, Transfer Learning, COVID-19, image classification

INTRODUCTION

In recent years, the technique of Transfer Learning (TL) within the domain of Deep Learning (DL) has enabled deep Convolutional Neural Networks (CNNs), noted for their large number of parameters, in successfully addressing numerous challenges of Medical Image Classification (MIC) tasks (Litjens et al., 2017; Yamashita et al., 2018). Lightweight CNNs have surfaced as efficient solutions to alleviate computational demands, effectively mirroring the efficacy of deep CNNs in MIC tasks (P. Sahu et al., 2019; Roslidar et al., 2019). Studies like the one conducted by Raghu et al. (2019), on the application of transfer learning in MIC have shown that these lightweight CNNs can match the performance of larger deep CNNs. Furthermore, these lightweight CNNs require less training time compared to deep CNNs as they contain less number of parameters and can be easily embedded in resource-constrained devices such as mobile phones (Roslidar et al., 2020). Mobile applications are proving to be great resources for the early detection of various pathologies (Ech-Cherif et al., 2019; H. Hendrick et al., 2019). While the architectural innovations in CNNs have rapidly progressed in recent years, identifying the impact of dataset size on CNNs is still a challenge for medical image analysis tasks (Morid et al., 2020). It is difficult to predict the optimal dataset size for a given architecture because of the black-box nature of CNNs (Yamashita et al., 2018). Moreover, there are several different categories of CNNs (Khan et al., 2020) to choose from for a certain medical image problem. It is essential to have a comprehensive understanding of how dataset sizes impact the architectures during TL to address the pressing need for an effective model in healthcare institutions. This knowledge can enable these institutions to select the appropriate architecture for a specific MIC task, based on the amount of data they have available.

The performance of CNNs in MIC is linked with the availability of large labeled datasets as it is a supervised learning problem, and DL methods are expected to create robust solutions with the availability of huge amounts of labeled data (Altaf et al., 2019). Although DL typically requires large datasets for effective classification performance, this may not always hold for medical images. This is because the variability in features and appearance of medical images is comparatively lower than that of natural images (Erickson et al., 2018). For instance, factors such as geometry, lighting, and distance matter significantly in the recognition of a natural image like a photograph of a cat or dog when compared to the recognition of a chest X-ray image which has a better-known scale. However, the availability of sufficient labeled datasets is still a challenge in the MIC research community (Altaf et al., 2019), it is also evident with the current pandemic situation. The current studies are struggling with a lack of COVID-positive labeled data making it difficult for DL architectures to predict the cases accurately (Tartaglione et al., 2020).

Existing literature highlights that TL methods in MIC tasks require sufficiently labeled datasets for better solutions. However, there is a lack of clarity on how the dataset sizes impact the architecture performance for a specific problem. Previous research in TL for medical image problems has identified a lack of investigation into optimal dataset sizes. This is particularly relevant given the findings of the study by Morid et al. (2020) which reviewed over 8000 papers and highlighted the need for further research in this area. We aim to contribute to filling this research gap and provide insights into optimal dataset sizes for TL approaches in medical image problems. To address this research gap, our study focuses on Multi-stage Transfer Learning, to explore the effects of fine-tuning and varying dataset sizes on lightweight CNNs. To this end, we put forward the following research questions: RQ1) How does the size of the dataset impact the finetuning layer in stage one TL? RQ2) What are the effects of the optimal model and dataset size on the classification performance in stage two? RQ3) What is the optimal fine-tuning layer and dataset size that produces the highest validation accuracy in stage two TL?

RELEVANT LITERATURE

In the past, several studies have adopted a multi-stage TL methodology for medical image classification. The strategy of employing an bridge dataset, closely resembling the target dataset, has been consistently successful and widely applied, particularly in scenarios with a scarce number of target images. A multi-image modality study (Kim et al., 2017) used a bridge dataset to reduce the domain differences between natural images and medical images. They described the bridge dataset as a medical image dataset that has a different use than the target problem but the same imaging modality. For instance, chest X-ray images can be used as bridge datasets for cyst X-ray image classification. Their method consisted of three main steps. First, the VGG16 classifier learned the characteristics of the natural images using natural images as the source database. Then the network learned the characteristics of medical images using the bridge dataset (chest X-ray images) and finally transferred the knowledge learned from the bridge dataset to the actual target problem (dental X-ray images). This approach resulted in the successful improvement in classification accuracies on MRI, CT, and X-ray image modalities and mitigated the domain differences between natural images and medical images. Similarly, another study on the lung cancer classification problem (Ausawalaithong et al., 2018) used Chest X-ray14 (Wang et al., 2017) dataset for recognizing lung nodules information in the first stage of training, the knowledge learned in stage-1 for recognizing the lung nodules information is transferred to predict the classes “malignant” and “non-malignant” in the second stage on JSRT (Shiraishi et al., 2000) dataset, the DenseNet-121 architecture’s performance increased when trained on both datasets instead of just training on JSRT for final classification.

Multi-stage Transfer Learning has also been observed in COVID-19 detection research as a means to achieve higher classification accuracy, especially when dealing with limited labeled data. For instance, Zhang et al. (Zhang et al., 2020) developed COVID19X-rayNet by initially training the model for pneumonia classification and then transferring the acquired knowledge to COVID-19 images. Similarly, Bassi et al. (Bassi & Attux, 2020) utilized CheXNet (Rajpurkar et al., 2017), a pre-trained model on both ImageNet and Chest X-ray14 datasets to apply TL on COVID-19 images. These studies demonstrate the success of Multistage TL in improving the performance of the classifier on the target dataset, however, they have limited to no information on the effectiveness of the bridge or target dataset sizes and stage-1 architecture optimization.

Only a few studies have investigated the effectiveness of dataset sizes, such as the one by Samala et al. (Samala et al., 2019) which focused on the effects of training sample size in Multi-stage Transfer Learning using the AlexNet architecture for breast tomosynthesis. To understand the training sample size effect, they analyzed three TL strategies, first, single-stage transfer learning by directly fine-tuning the C1-frozen CNN using DBT data. Second, stage 1 training of 1st layer frozen CNN using a fixed mammography data size (100%) data followed by stage 2 training of 1st layer frozen CNN using DBT data, and third, stage 1 training of 1st layer frozen CNN using a fixed mammography data size (100%) followed by stage 2 training of CNN frozen up to 4th fully connected layer using DBT data. Their work demonstrates that using an additional stage of the TL and bridge dataset improves the classification performance and it depends on the relative sizes of available training samples in the target and bridge data set. However, their study has some limitations. Using different dataset sizes may affect the learning capacity in a transfer network, leading to the risk of overfitting. Additionally, the networks in both stages were not extensively investigated based on various combinations of freezing the layers. Therefore, there is still room for further investigation, including the use of larger datasets and the evaluation of different combinations of fine-tuned networks. Our study builds on previous Transfer Learning works by utilizing multi-stage TL. However, unlike these studies, we focus primarily on the impact of training sample size and exhaustively testing various fine-tuning combinations in the first stage TL. We experiment with freezing different layers in the 1st stage TL and plot the resulting classification performance against various dataset sizes. This provides valuable insights into determining the optimal fine-tuning approach for the available training dataset on a given MIC task.

RESEARCH METHODOLOGY

Datasets and preprocessing

In this research study, we conducted experiments using the Retinal OCT images (Kermany et al., 2018) for the first stage of transfer learning and the COVID-QU-Ex (Tahir et al., 2021) dataset for the second stage of transfer learning. The Retinal OCT images dataset consists of 84,495 images. These images have been categorized into four distinct groups: Normal, CNV (choroidal neovascularization), DME (diabetic macular edema), and Drusen. The COVID-QU-Ex dataset comprises 33,920 chest X-ray images. It consists of COVID-19 cases (11,956), non-COVID cases (11,263, including viral and bacterial pneumonia), and normal cases (10,701). We randomly chose 2800 images from this dataset, dividing them into 85% for training and 25% for testing. All the images were transformed to have dimensions of 150x150 pixels to standardize the images. Also, the pixel values for each image were normalized to fall within a range of 0 to 1. Additionally, to balance the retinal OCT dataset, higher weights were assigned to the minority classes. This helps to prevent bias towards the majority classes and allows the model to accurately classify examples from the minority classes as well.

Multi-stage Transfer Learning strategy

During the initial phase of TL, we employed a pre-trained MobileNetV2 model on the bridge dataset, the retinal OCT dataset, by utilizing the fine-tuning approach. In the subsequent stage of TL, we further trained the MobileNetV2 on the target dataset, employing the feature extraction approach. Figure 1 depicts the proposed multi-stage TL framework.

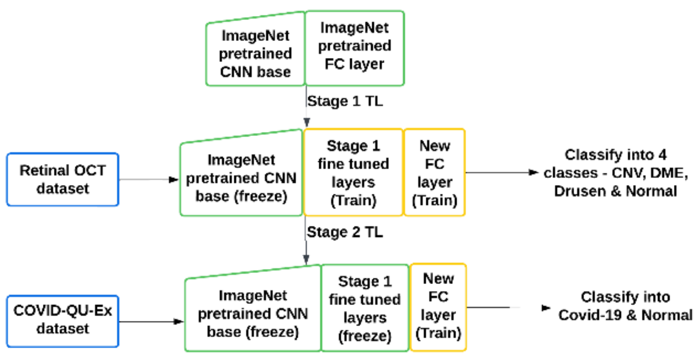

Figure 1. Multi-stage TL framework

In the first stage, the MobileNetV2 model pre-trained on the ImageNet dataset was trained on retinal OCT images for multiclass image classification. The TL approach with fine-tuning was employed, where the final fully connected layer of the pre-trained model was replaced with a new fully connected layer to classify the images into four classes. This new classifier is then retrained with retinal OCT images. Using the retinal OCT dataset, eight different datasets of sizes 10,000 to 80,000 were created by randomly selecting images from each class. Training of the MobileNetV2 model, augmented with a new fully connected layer, is carried out on each of the eight datasets. For every dataset, the model undergoes several training cycles, during which five layers from the top are progressively unfrozen in each cycle, continuing until the model achieves its peak performance. The lower-level layers of the pre-trained model are frozen as these layers have already learned basic features such as edges and curves that are relevant to our problem. By allowing the higher-level layers of the pre-trained model to be trained along with the new classifier the model can focus on learning more specific features of the retinal OCT images. Stage 1 TL aids in decreasing the gap between natural images and the target dataset and improves the performance of the model on this specific task. Once the pre-trained model was trained until convergence, the model along with the weights was saved on disk.

In the second stage, we freeze the pre-trained ImageNet base and stage 1 fine-tuned layers to preserve the knowledge gained from the bridge dataset. The saved model from the previous step is trained on the target dataset i.e., COVID-QU-Ex by employing TL with the feature extraction method. In the feature extraction approach, the fully connected layer of the model was truncated, and a new classifier for binary classification was attached. Also, all the layers of the model were frozen except the new fully connected classifier. This reduced the number of parameters to train and the training time drastically. During stage 2, the models were trained for 25 epochs, utilizing a batch size of eight images and the Adam optimizer. Furthermore, the training of these models is stopped when the benchmark training accuracy of 85% is achieved. Additionally, the model training was performed using the Python programming language and implemented via the Keras and TensorFlow API. The hardware framework comprised a GPU setup on a Windows 10 operating system, employing an NVIDIA TITAN V GPU and an Intel Core i7 processor for the training process.

RESULTS AND DISCUSSION

Stage One Transfer Learning

Figure 2.a. illustrates the optimal fine-tuning layer obtained for the eight various dataset sizes in stage 1 TL. For smaller dataset sizes i.e., with 10K and 20K images, the number of layers frozen to achieve optimal performance has remained the same i.e., 85 layers. This suggests that, within this range, adding more data did not require increased parameters to achieve the best performance. Gradually the number of layers frozen has increased as the dataset size has increased to 50K images and for the dataset size of 60K images, the optimal layer has remained the same. Nevertheless, the number of frozen layers has steadily increased as the dataset size has increased. For the largest dataset size with 80K images, the optimal number of layers frozen achieved is 120 layers. In the case of larger dataset sizes, a higher number of layers are frozen to achieve best-performing models, underscoring the advantage gained from the rich and broad feature set these datasets possess. Essentially, the presence of extensive training data permits models to fine-tune its upper layers while keeping the core, fundamental features intact. This illustrates that, provided with sufficient data, models are capable of learning and generalizing well, even when a larger proportion of its capacity is constrained. In contrast, smaller datasets demand training of a larger number of parameters to reach peak performance levels.

Stage Two Transfer Learning

Based on the proposed framework, stage two experiments were conducted for the COVID-19 classification task using the optimal models obtained from training on the bridge dataset in the stage one experiment. Figure 2.b. illustrates the target validation accuracy achieved for COVID-19 images. It can be observed that the validation accuracy ranges between 69.25% and 74.5%. The highest accuracy was achieved for the optimal model that was trained on dataset size with 40K images in stage one experiments. The lowest accuracy was achieved for the optimal model trained on a 60K dataset size.

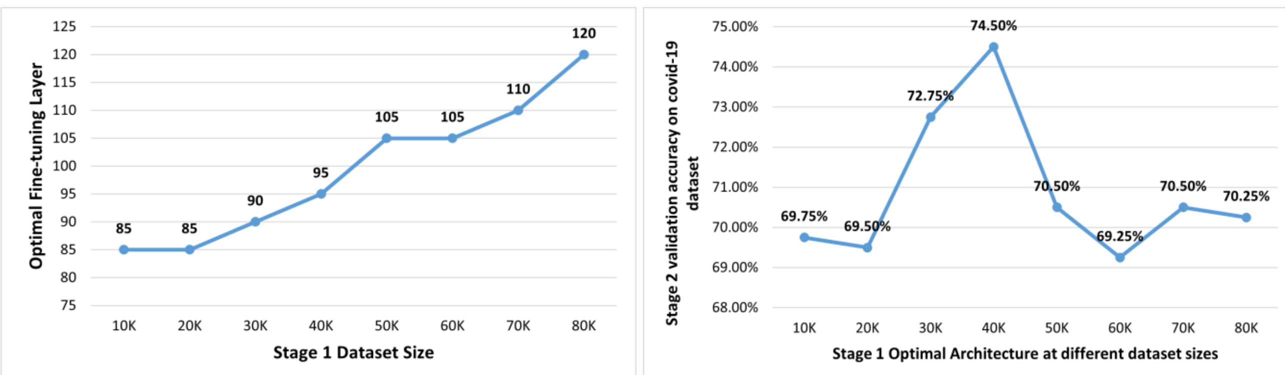

Figure 2.a. Dataset size vs Fine-tuning in Stage 1 TL 2.b. Validation accuracy achieved on COVID-19 in Stage 2 TL

The optimal models from stage 1 trained on smaller datasets i.e., with 10K and 20K images showed no significant difference in performance for COVID-19 classification. This is because the models had an equal number of layers frozen and additional data did not substantially improve the adaptability of the model to a new task. There was a considerable increase in target validation accuracy for optimal models trained with 30K and 40K datasets, where the model trained on the 40K dataset outperformed all others. However, the optimal models trained on the 50K – 80K dataset decline in target validation accuracy and stabilize to between 69% to 70.5% accuracy with minor variations. This plateau signifies that additional data is no longer contributing to substantial learning and improvement in performance on COVID-19 classification. Hence, it can be said that for lightweight models the dataset size of 30 to 40 thousand images is optimal to achieve the best performance in stage 2 TL.

Figure 3 demonstrates that the relationship between stage 1 dataset size, fine-tuning layer, and stage 2 validation accuracy is not linear. The target accuracy peaks at 40K and then decreases even as more layers are frozen with larger datasets. Beyond the threshold of 40K, increasing the dataset size does not yield a significant increase in accuracy, which may suggest that the model is overfitting on the stage 1 data and is unable to generalize well on the stage 2 task. Increasing the number of frozen layers from 85 to 120 initially enhances target accuracy, however beyond 95 layers, further freezing restricts the capacity of the model to learn relevant features for the target task. This implies the necessity to train a specific threshold of layers with the bridge dataset to optimize performance for the stage 2 task. It highlights that more data or more fine-tuning in stage 1 TL does not translate to better performance in stage 2 TL. Additionally, the bridge dataset enabled the model to acquire features pertinent to the target task, thereby allowing the model to achieve competitive performance even with a relatively smaller amount of target data. This study illustrates that it is feasible to create efficient models using minimal data and computational resources. These developments could significantly enhance the accessibility of advanced AI enabled tools, particularly in developing countries and within healthcare systems that are resource constrained.

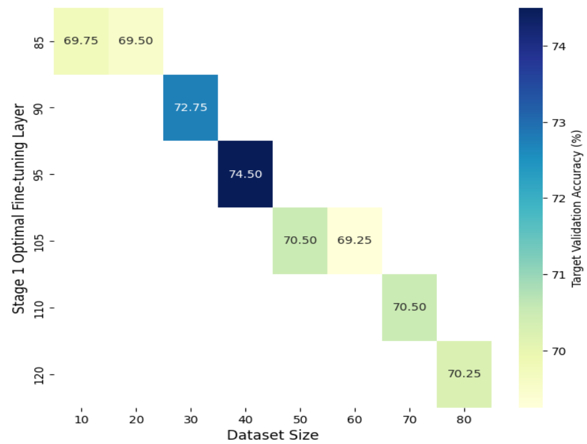

Figure 3. Heatmap visualizing the dataset size, optimal fine-tuning layer, and model performance

CONCLUSION

The study presents a comprehensive investigation into the effects of dataset sizes and fine-tuning in multi-stage transfer learning system on the performance of lightweight CNNs. We employed the lightweight MobileNetV2 architecture, initially training on varied sizes of Retinal OCT datasets and subsequently adapting the best models from stage 1 TL for COVID-19 image classification. The study highlights a key trade-off: in stage one, larger datasets attain maximum performance with fewer re-trained layers, whereas smaller datasets require more comprehensive layer training. However, this does not translate into better performance on target task in multi-stage transfer learning. While the findings of this study are promising there are a few limitations to this study. The study did not consider other attributes that have an impact on model performance apart from the size of the training dataset such as the quality of the data. Additionally, the fine-tuning is performed by unfreezing five layers in each iteration. This can be further improved by performing deeper fine-tuning of the models. Future research endeavors may delve into investigating how dataset size and fine-tuning affect various sizes of lightweight models, including truncated versions of these models. This will allow us to further reduce the computational resources and yield significant benefits for advancing the architectural improvements of lightweight models.

REFERENCES

1. Altaf, F., Islam, S. M. S., Akhtar, N., & Janjua, N. K. (2019). Going Deep in Medical Image Analysis: Concepts, Methods, Challenges and Future Directions. arXiv:1902.05655 [Cs]. http://arxiv.org/abs/1902.05655
2. Ausawalaithong, W., Marukatat, S., Thirach, A., & Wilaiprasitporn, T. (2018). Automatic Lung Cancer Prediction from Chest X-ray Images Using Deep Learning Approach. arXiv:1808.10858 [Cs, Eess]. http://arxiv.org/abs/1808.10858
3. Bassi, P. R. A. S., & Attux, R. (2020). A Deep Convolutional Neural Network for COVID-19 Detection Using Chest X-Rays. arXiv:2005.01578 [Cs, Eess]. http://arxiv.org/abs/2005.01578
4. Ech-Cherif, A., Misbhauddin, M., & Ech-Cherif, M. (2019). Deep Neural Network Based Mobile Dermoscopy Application for Triaging Skin Cancer Detection. 2019 2nd International Conference on Computer Applications Information Security (ICCAIS), 1–6. https://doi.org/10.1109/CAIS.2019.8769517
5. Erickson, B. J., Korfiatis, P., Kline, T. L., Akkus, Z., Philbrick, K., & Weston, A. D. (2018). Deep Learning in Radiology: Does One Size Fit All? Journal of the American College of Radiology, 15(3), 521–526. https://doi.org/10.1016/j.jacr.2017.12.027
6. H. Hendrick, W. Zhi-Hao, C. Hsien-I, C. Pei-Lun, & J. Gwo-Jia. (2019). IOS Mobile APP for Tuberculosis Detection Based on Chest X-Ray Image. 2019 2nd International Conference on Applied Information Technology and Innovation (ICAITI), 122–125. https://doi.org/10.1109/ICAITI48442.2019.8982152

7. Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C. S., Liang, H., Baxter, S. L., McKeown, A., Yang, G., Wu, X., Yan, F., Dong, J., Prasadha, M. K., Pei, J., Ting, M. Y. L., Zhu, J., Li, C., Hewett, S., Dong, J., Ziyar, I., … Zhang, K. (2018). Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell, 172(5), 1122-1131.e9. https://doi.org/10.1016/j.cell.2018.02.010
8. Khan, A., Sohail, A., Zahoora, U., & Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. Artificial Intelligence Review. https://doi.org/10.1007/s10462-020-09825-6
9. Kim, H. G., Choi, Y., & Ro, Y. M. (2017). Modality-bridge Transfer Learning for Medical Image Classification. arXiv:1708.03111 [Cs]. http://arxiv.org/abs/1708.03111
10. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A Survey on Deep Learning in Medical Image Analysis. Medical Image Analysis, 42, 60–88. https://doi.org/10.1016/j.media.2017.07.005
11. Morid, M. A., Borjali, A., & Del Fiol, G. (2020). A scoping review of transfer learning research on medical image analysis using ImageNet. arXiv:2004.13175 [Cs, Eess]. http://arxiv.org/abs/2004.13175
12. P. Sahu, D. Yu, M. Dasari, F. Hou, & H. Qin. (2019). A Lightweight Multi-Section CNN for Lung Nodule Classification and Malignancy Estimation. IEEE Journal of Biomedical and Health Informatics, 23(3), Article 3. https://doi.org/10.1109/JBHI.2018.2879834
13. Raghu, M., Zhang, C., Kleinberg, J., & Bengio, S. (2019). Transfusion: Understanding Transfer Learning for Medical Imaging. arXiv:1902.07208 [Cs, Stat]. http://arxiv.org/abs/1902.07208
14. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017). CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. arXiv:1711.05225 [Cs, Stat]. http://arxiv.org/abs/1711.05225
15. Roslidar, R., Rahman, A., Muharar, R., Syahputra, M. R., Arnia, F., Syukri, M., Pradhan, B., & Munadi, K. (2020). A Review on Recent Progress in Thermal Imaging and Deep Learning Approaches for Breast Cancer Detection. IEEE Access, 8, 116176–116194. https://doi.org/10.1109/ACCESS.2020.3004056
16. Roslidar, R., Saddami, K., Arnia, F., Syukri, M., & Munadi, K. (2019). A Study of Fine-Tuning CNN Models Based on Thermal Imaging for Breast Cancer Classification. 2019 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom), 77–81. https://doi.org/10.1109/CYBERNETICSCOM.2019.8875661
17. Samala, R. K., Chan, H.-P., Hadjiiski, L., Helvie, M. A., Richter, C. D., & Cha, K. H. (2019). Breast Cancer Diagnosis in Digital Breast Tomosynthesis: Effects of Training Sample Size on Multi-Stage Transfer Learning Using Deep Neural Nets. IEEE Transactions on Medical Imaging, 38(3), 686–696. https://doi.org/10.1109/TMI.2018.2870343
18. Shiraishi, J., Katsuragawa, S., Ikezoe, J., Matsumoto, T., Kobayashi, T., Komatsu, K., Matsui, M., Fujita, H., Kodera, Y., & Doi, K. (2000). Development of a Digital Image Database for Chest Radiographs With and Without a Lung Nodule. American Journal of Roentgenology, 174(1), 71–74. https://doi.org/10.2214/ajr.174.1.1740071
19. Tahir, A. M., Chowdhury, M. E. H., Khandakar, A., Rahman, T., Qiblawey, Y., Khurshid, U., Kiranyaz, S., Ibtehaz, N., Rahman, M. S., Al-Maadeed, S., Mahmud, S., Ezeddin, M., Hameed, K., & Hamid, T. (2021). COVID-19 infection localization and severity grading from chest X-ray images. Computers in Biology and Medicine, 139, 105002. https://doi.org/10.1016/j.compbiomed.2021.105002
20. Tartaglione, E., Barbano, C. A., Berzovini, C., Calandri, M., & Grangetto, M. (2020). Unveiling COVID-19 from Chest X-ray with deep learning: A hurdles race with small data. arXiv:2004.05405 [Cs, Eess]. http://arxiv.org/abs/2004.05405
21. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3462–3471. https://doi.org/10.1109/CVPR.2017.369
22. Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: An overview and application in radiology. Insights into Imaging, 9(4), 611–629. https://doi.org/10.1007/s13244-018-0639-9
23. Zhang, R., Guo, Z., Sun, Y., Lu, Q., Xu, Z., Yao, Z., Duan, M., Liu, S., Ren, Y., Huang, L., & Zhou, F. (2020). COVID19XrayNet: A Two-Step Transfer Learning Model for the COVID-19 Detecting Problem Based on a Limited Number of Chest X-Ray Images. Interdisciplinary Sciences: Computational Life Sciences. https://doi.org/10.1007/s12539-020-00393-5