Physical and Engineering Sciences in Medicine
https://doi.org/10.1007/s13246-020-00865-4

SCIENTIFIC PAPER

Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks

Ioannis D. Apostolopoulos[1] · **Tzani A. Mpesiana**[2]

Received: 25 March 2020 / Accepted: 30 March 2020

Abstract

In this study, a dataset of X-ray images from patients with common bacterial pneumonia, confirmed Covid-19 disease, and normal incidents, was utilized for the automatic detection of the Coronavirus disease. The aim of the study is to evaluate the performance of state-of-the-art convolutional neural network architectures proposed over the recent years for medical image classification. Specifically, the procedure called Transfer Learning was adopted. With transfer learning, the detection of various abnormalities in small medical image datasets is an achievable target, often yielding remarkable results. The datasets utilized in this experiment are two. Firstly, a collection of 1427 X-ray images including 224 images with confirmed Covid-19 disease, 700 images with confirmed common bacterial pneumonia, and 504 images of normal conditions. Secondly, a dataset including 224 images with confirmed Covid-19 disease, 714 images with confirmed bacterial and viral pneumonia, and 504 images of normal conditions. The data was collected from the available X-ray images on public medical repositories. The results suggest that Deep Learning with X-ray imaging may extract significant biomarkers related to the Covid-19 disease, while the best accuracy, sensitivity, and specificity obtained is 96.78%, 98.66%, and 96.46% respectively. Since by now, all diagnostic tests show failure rates such as to raise concerns, the probability of incorporating X-rays into the diagnosis of the disease could be assessed by the medical community, based on the findings, while more research to evaluate the X-ray approach from different aspects may be conducted.

Keywords Covid-19 · Automatic detections · X-ray · Transfer learning · Deep learning

Introduction

COVID-19 is an acute resolved disease, but it can also be deadly, with a 2% case fatality rate. Severe disease onset might result in death due to massive alveolar damage and progressive respiratory failure [1]. The early and automatic diagnosis of Covid-19 may be beneficial for countries for timely referral of the patient to quarantine, rapid intubation of serious cases in specialized hospitals, and monitoring of the spread of the disease.

Although the diagnosis has become a relatively fast process, the financial issues arising from the cost of diagnostic tests concern both states and patients, especially in countries with private health systems, or restricted access health systems due to prohibitive prices.

In March 2020, there has been an increase in publicly available X-rays from healthy cases, but also from patients suffering from Covid-19. This enables us to study the medical images and identify possible patterns that may lead to the automatic diagnosis of the disease.

The development of deep learning applications over the last 5 years seems to have come at the right time. Deep Learning is a combination of machine learning methods mainly focused on the automatic feature extraction and classification from images, while its applications are broadly met is object detection tasks, or in medical image classification tasks. Machine learning and deep learning have become established disciplines in applying artificial intelligence to mine, analyze, and recognize patterns from data. Reclaiming the advances of those fields to the benefit of clinical decision making and computer-aided systems is increasingly becoming nontrivial, as new data emerge [2].

✉ Ioannis D. Apostolopoulos
ece7216@upnet.gr

[1] Department of Medical Physics, School of Medicine, University of Patras, 26504 Patras, Greece

[2] Computer Technology Institute, University of Patras, Patras, Greece

Published online: 03 April 2020

Deep Learning often refers to a procedure wherein deep convolutional neural networks are utilized for automatic mass feature extraction, which is achieved by the process called convolution. The layers process nonlinear information [3]. Each layer involves a transformation of the data into a higher and more abstract level. The deeper we go into the network, the more complex information is learned. Higher layers of portrayal enhance parts of the information that are significant for segregation and smother unimportant attributes. Usually, deep learning refers to more deep networks than the classic machine learning ones, utilizing big data.

The purpose of this research is to evaluate the effectiveness of state-of-the-art pre-trained convolutional neural networks proposed by the scientific community, regarding their expertise in the automatic diagnosis of Covid-19 from thoracic X-rays. To achieve this, a collection of 1427 thoracic X-ray scans is processed and utilized to train and test the CNNs. Due to the fact that the size of the samples related to Covid-19 is small (224 images), transfer learning is a preferable strategy to train the deep CNNs. This is due to the fact that the state-of-the-art CNNs are sophisticated model requiring large-scale datasets to perform accurate feature extraction and classification. With transfer learning, the retention of the knowledge extracted from one task is the key to perform an alternative task.

The results are encouraging and demonstrate the effectiveness of deep learning, and more specifically, transfer learning with CNNs to the automatic detection of abnormal X-ray images from small datasets, related to the Covid-19 disease. Despite the limitations of the current study, which are related to the data availability limitations, opens the horizons for more specialized research into the possible existence biomarkers related to the Covid-19 disease in X-ray images. The possible significance of those biomarkers can be confirmed or denied by other feature extraction techniques such as Radiomics in future research.

Methods

Dataset of the study

For the purpose of the experiments, several sources of X-rays were accessed. Firstly, the Github Repository was analyzed for related datasets. A collection of X-ray images from Cohen [4] was selected. Secondly, the following web sites were thoroughly examined: (a) Radiological Society of North America (RSNA), (b) Radiopaedia, and (c) Italian Society of Medical and Interventional Radiology (SIRM). This collection is available on the Internet [5]. Thirdly, a collection of common bacterial-pneumonia X-ray scans was added to the dataset, to train the CNNs to distinguish Covid-19 from common pneumonia. This collection is available on the Internet by Kermany et al. [6].

The collected data includes 224 images with confirmed Covid-19, 700 images with confirmed common bacterial pneumonia, and 504 images of normal condition. This datasets is referred to as Dataset_1.

To further evaluate the classification performance, another experiment was conducted. Specifically, some modification to the dataset were performed. Since the first dataset included only bacterial pneumonia cases, it was impossible to investigate the performance of MobileNet in distinguishing the Covid-19 disease from other pneumonia cases. For the above reason, the addition of viral pneumonia cases (non-Covid-19) was mandatory. This dataset includes the 224 cases of confirmed Covid-19, the 504 cases of healthy instances, and 714 cases of both bacterial and viral pneumonia (400 bacteria and 314 viral). The particular dataset is referred to as Dataset_2.

The X-ray images were rescaled to a size of 200×266. For the images of different pixel ration, and to avoid distortion, a black background of 1:1.5 ratio was added to achieve a perfect rescale to 200×266. The reader must note that the CNNs are capable of ignoring slight positional variance, i.e., they seek for patterns not only to a specific position of the image but also moving patterns.

For the particular dataset, some limitations have to be mentioned. Firstly, positive Covid-19 samples represent a small sample of cases and even a sample referring to more severe cases with pneumonia symptoms. Unfortunately, there is no safer sample available at this time. Secondly, the pneumonia incidence samples are older recorded samples and do not represent pneumonia images from patients with suspected Coronavirus symptoms, while the clinical conditions are missing.

Those limitations are thoroughly discussed in the "Discussion" section.

Transfer learning with CNNs

Transfer learning is a strategy wherein the knowledge mined by a CNN from given data is transferred to solve a different but related task, involving new data, which usually are of a smaller population to train a CNN from scratch [7].

In deep learning, this process involves the initial training of a CNN for a specific task (e.g., classification), utilizing large-scale datasets. The availability of data for the initial training is the most vital factor for successful training

since CNN can learn to extract significant characteristics (features) of the image. Depending on the capability of the CNN to identify and extract the most outstanding image features, it is judged whether this model is suitable for transfer learning.

During the next phase, the CNN is employed to process a new set of images of another nature and to extract features, according to its knowledge in feature extraction, which was obtained during the initial training. There are two commonly used strategies to exploit the capabilities of the pre-trained CNN. The first strategy is called feature extraction via transfer learning [8] and refers to the approach wherein the pre-trained model retains both its initial architecture, and all the learned weights. Hence, the pre-trained model is used only as a feature extractor; the extracted features are inserted into a new network that performs the classification task. This method is commonly used either to circumvent computational costs coming with training a very deep network from scratch, or to retain the useful feature extractors trained during the initial stage.

The second strategy refers to a more sophisticated procedure, wherein specific modifications are applied to the pre-trained model, to achieve optimal results. Those modifications may include architecture adjustments and parameter tuning. In this way, only specific knowledge mined from the previous task is retained, while new trainable parameters are inserted into the network. The new parameters require training on a relatively large amount of data to be advantageous.

In medical tasks, the most prominent practice to perform transfer learning is exploiting the CNNs participated and stood out in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [9], which evaluates algorithms for object detection and image classification at large scale.

Table 1 The CNNs of this experiment and their parameters for transfer learning

Network	Parameter	Description
VGG19 [10]	Layer Cutoff	18
	Neural Network	1024 nodes
MobileNet v2 [11]	Layer Cutoff	10
	Neural Network	1000 nodes, 750 nodes
Inception [12]	Layer Cutoff	249
	Neural Network	1000
Xception [13]	Layer Cutoff	120
	Neural Network	1000 nodes, 750 nodes
Inception ResNet v2 [12]	Layer Cutoff	730
	Neural Network	No

State-of-the-art CNNs for transfer learning

In this section, a brief description of the CNNs employed for automatic detection is illustrated. In Table 1, the CNNs employed for the classification task and the parameters tuned for transfer learning are presented. The parameters were defined after several experiments, although the possible alternative choices are limitless and could be investigated in future research as to their contribution to the improvement of the performance. The parameter called Layer Cutoff refers to the number of untrainable layers starting from the bottom of the CNN. The rest of the layers, which are closer to the output features, are made trainable, to allow more information extraction coming from the late convolutional layers. The parameter Neural Network refers to the classifier placed at the top of the CNN to perform the classification of the extracted features. It is described by the total number of hidden layers and the total number of nodes.

All the CNNs share some common hyper-parameters. More specifically, all the convolutional layers are activated by the Rectified Linear Unit (ReLU) [14]. For the Neural Networks utilizing two hidden layers, a Dropout [15] layer is added to prevent overfitting [16]. The CNNs were compiled utilizing the optimization method called Adam [17]. The training was conducted for ten epochs, with a batch size of 64.

Metrics

Regarding the classification task of the CNNs, specific metrics were recorded as follows: (a) correctly identified diseased cases (True Positives, TP), (b) incorrectly classified diseased cases (False Negatives, FN), (c) correctly identified healthy cases (True Negatives, TN), and (d), incorrectly classified healthy cases (False Positives, FP). Please note that TP refers to the correctly predicted Covid-19 cases, FP refers to typical or pneumonia cases that were classified as Covid-19 by the CNN, TN refers to normal or pneumonia cases that were classified as non-Covid-19 cases, while the FN refers to Covid-19 cases classified as normal or as common pneumonia cases.

Due to the fact that the main intention of the study is the detection of Covid-19, we measure two different accuracies. The first accuracy refers to the overall accuracy of the model in distinguishing the three classes (normal-pneumonia-Covid) and is called 3-class accuracy. The second accuracy refers to the accuracy related to Covid-19 only. That is, if an instance is typical and is classified as pneumonia by the CNN, it is still considered acceptable in terms of the presence of Covid-19. The aforementioned accuracy is called two-class accuracy.

Table 2 Results of the CNNs used for transfer learning

Network	Accuracy 2-class (%)	Accuracy 3-class (%)	Sensitivity (%)	Specificity (%)
VGG19 [10]	98.75	93.48	92.85	98.75
MobileNet v2 [11]	97.40	92.85	99.10	97.09
Inception [12]	86.13	92.85	12.94	99.70*
Xception [13]	85.57	92.85	0.08	99.99*
Inception ResNet v2 [12]	84.38	92.85	0.01	99.83*

Due to data imbalance, measurements corresponding to Sensitivity and Specificity that obtained not meaningful values for the particular set of results, are denoted by an asterisk (*)

Table 3 Confusion matrix of the two best CNNs

Model	Predicted labels	Actual labels		
		Actual Covid-19	Actual pneumonia	Actual normal
MobileNet v2 [11]	Predicted Covid-19	222	8	27
	Predicted pneumonia	2	495	27
	Predicted normal	0	1	646
VGG19 [10]	Predicted Covid-19	222	8	7
	Predicted pneumonia	3	460	26
	Predicted normal	13	36	667

Based on those metrics, we compute the accuracy, sensitivity, and specificity of the model. The equations explaining the aforementioned metrics are Eqs. 1, 2, and 3.

$$\text{Acurracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (1)$$

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN}) \quad (2)$$

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP}) \quad (3)$$

Results

The training and evaluation procedure was performed with tenfold-cross-validation.

Classification accuracy on Dataset_1

The results for each CNN for Dataset_1 are illustrated in Tables 2 and 3. In Table 2, the accuracy, sensitivity, and specificity are presented.

The results suggest that the VGG19 and the MobileNet v2 achieve the best classification accuracy over the rest of the CNNs. Due to the imbalance of the dataset, all the CNNs perform seemingly well in terms of accuracy and in terms of specificity. However, as those metrics depend heavily on the number of samples representing each class, their unilateral evaluation leads to incorrect conclusions. For this reason, the combination of accuracy, sensitivity, and specificity must be the criterion for choosing the best model. The measurements denoted with an asterisk (*) in Table 2, indicate that these values are not considered acceptable in real-life problems, due to the above issue.

Table 4 True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) related to the Covid-19 class, for the best performing CNNs

CNN	TP	FP	TN	FN
VGG19	**208**	**15**	**1189**	16
MobileNet v2	222	35	1169	**2**

Bold values represent the optimal observed values

To further evaluate the two best models (VGG19 and MobileNet v2), the confusion matrix of each model is presented in Table 3.

The results of Table 3, can be simplified to represent the sensitivity and the specificity of the test regarding Covid-19. This procedure was explained in "Metrics" section. The results are illustrated in Table 4.

Table 5 Accuracy, sensitivity, and specificity of MobileNet v2 on Dataset_2

Network	Accuracy 2-class (%)	Accuracy 3-class (%)	Sensitivity (%)	Specificity (%)
MobileNet v2 [11]	96.78	94.72	98.66	96.46

While VGG19 achieves better accuracy, it is clear that in terms of the particular disease, the optimal results are those with the lowest number of FN. A real-life interpretation of a False Negative instance would result in the mistaken assumption that the patient is not infected with what this entails for the spread of the virus and public health.

The MobileNet v2 outperforms VGG19 in terms of specificity and, thus, it is proven to be the most effective model for the specific classification task, and the specific data sample.

Classification accuracy on Dataset_2

The Dataset_2 includes viral cases of pneumonia, which are added to the bacterial cases, and classified with the same label, that is common pneumonia (single class). The MobileNet v2 setup of the first experiment was retained to perform the classification task of Dataset_2. The results are illustrated in Tables 5 and 6. In Table 4, the overall accuracy for the two classes, the overall accuracy for the three classes, and the sensitivity and specificity related to Covid-19 is presented. In Table 6, the confusion matrix for the three classes is presented.

Based on Tables 5 and 6, it is confirmed that the MobileNet v2 effectively distinguished the Covid-19 cases from viral and bacterial pneumonia cases from the particular dataset. Besides, the low number of FN (3) are an encouraging result.

Discussion

Based on the results, it is demonstrated that deep learning with CNNs may have significant effects on the automatic detection and automatic extraction of essential features from X-ray images, related to the diagnosis of the Covid-19.

Some limitations of the particular study can be overcome in future research. In particular, a more in-depth analysis requires much more patient data, especially those suffering from Covid-19. A more interesting approach for future research would focus on distinguishing patients showing mild symptoms, rather than pneumonia symptoms, while these symptoms may not be accurately visualized on X-rays, or may not be visualized at all.

Moreover, it is necessary to develop models capable of distinguishing Covid-19 cases from other similar viral cases, such as SARS, but also from a greater variety of common pneumonia or even physiological X-rays. Besides, the automatic diagnosis of cases was made using only a medical image rather than a more holistic approach to the patient, based on other factors that may be offered and may behave as risk factors for the onset of the disease.

Nonetheless, the present work contributes to the possibility of a low-cost, rapid, and automatic diagnosis of the Coronavirus disease. It is to be investigated whether the extracted features performed by the CNNs constitute reliable biomarkers aiding to the detection of Covid-19. Also, despite the fact that the appropriate treatment is not determined solely from an X-ray image, an initial screening of the cases would be useful, not in the type of treatment, but in the timely application of quarantine measures in the positive samples, until a more complete examination and a specific treatment or follow-up procedure is followed.

Besides, the advantage of automatic detection of Covid-19 from either medical image lies on the reduction of exposure of nursing and medical staff to the outbreak.

Table 6 Confusion matrix of the classification of Dataset_2 by MobileNet v2

Model	Predicted labels	Actual labels		
		Actual Covid-19	Actual pneumonia	Actual normal
MobileNet v2 [11]	Predicted Covid-19	221	19	24
	Predicted pneumonia	2	472	17
	Predicted normal	1	13	673

Funding This study did not receive external funding.

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Xu Z, Shi L, Wang Y et al (2020) Pathological findings of COVID-19 associated with acute respiratory distress syndrome. Lancet Respir Med. https://doi.org/10.1016/S2213-2600(20)30076-X
2. Greenspan H, Van Ginneken B, Summers RM (2016) Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique. IEEE Trans Med Imaging 35:1153–1159
3. Deng L, Yu D et al (2014) Deep learning: methods and applications. Found Trends® Signal Process 7(3–4):197–387
4. Cohen JP (2020) COVID-19 image data collection. https://github.com/ieee8023/covid-chestxray-dataset
5. Kaggle. https://www.kaggle.com/andrewmvd/convid19-X-rays
6. Kermany DS, Goldbaum M, Cai W et al (2018) Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 172:1122–1131.e9. https://doi.org/10.1016/j.cell.2018.02.010
7. Weiss K, Khoshgoftaar TM, Wang D (2016) A survey of transfer learning. J Big data 3:9
8. Huh M Agrawal P Efros AA (2016) What makes ImageNet good for transfer learning?. arXiv preprint arXiv:160808614
9. Russakovsky O Deng J Su H et al (2015) ImageNet large scale visual recognition challenge. arXiv preprint arXiv:14090575
10. Simonyan K Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556
11. Howard AG Zhu M Chen B et al (2017) MobileNets: efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:170404861
12. Szegedy C Ioffe S Vanhoucke V Alemi AA (2017) Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-first AAAI conference on artificial intelligence
13. Chollet F (2017) Xception: deep learning with depthwise separable convolutions. arXiv preprint arXiv:161002357
14. Nair V Hinton GE (2010) Rectified linear units improve restricted boltzmann machines. In: proceedings of the 27th international conference on machine learning (ICML-10), pp 807–814
15. Hinton GE Srivastava N Krizhevsk A et al (2012) Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:12070580
16. Hawkins DM (2004) The problem of overfitting. J Chem Inf Comput Sci 44:1–12
17. Kingma DP Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:14126980