On Convolutional Neural Networks for Chest X-ray Classification

I Naskinova
South-West University “Neofit Rilski”, Blagoevgrad, Bulgaria

e-mail: naskinova@abv.bg

Abstract. Pneumonia is an infectious disease accounting for one fifth of the deaths of children under five worldwide. It is also a cause for adult hospital admissions with mortality rate of almost 25% in patients over 75 years. Pneumonia is curable and mortality can be prevented if it is diagnosed. Preferred diagnostic technique is chest X-ray image examination. The lack of radiology equipment of trained clinicians reduces the chance of the majority of people to be properly diagnosed. In the paper we propose a shallow convolutional neural network architecture, further fine-tuned with choice of the Adam optimizer which is split tested in the current experimental work. We showed that the automatic detection of pneumonia in chest X-ray images is possible with accuracy higher than 90% using 3 blocks of 2 convolutional 2D layers with max-pooling for feature extraction and a flatten output block comprising two dense layers, trained over 10 epochs. Therefore, the application for CNN is a viable solution as a supplement to the decision-making process of pneumonia X-ray diagnostics.

1. Introduction

Pneumonia is an infection caused by a variety of organisms, including bacteria, viruses and fungi. The air sacs get inflamed and may fill with purulent material – fluid or pus, which causes cough with phlegm or pus, fever chills, and difficult breathing [1]. There are two types of pneumonia — viral and bacterial. Bacterial pneumonia requires antibiotics treatment, while virus caused can be healed on its own. It is the single largest infectious cause of death in children worldwide. Pneumonia account for nearly one fifth of all deaths of children under 5 years [2]. At-risk groups of pneumonia also include adults over the age of 65 and people with pre-existing health problems [3]. These infections spread by direct contact with infected people. Deaths by pneumonia can be prevented by low-cost, low-tech medication and care. However, the disease goes untreated, because it is not diagnosed. The PERCH study is the largest evaluation of the WHO methodology for the standardized interpretation of paediatric CXRs. Their results reinforce the failure to achieve high concordance on distinguishing normal from other infiltrates. The misclassification between these categories can be reduced by incorporating in the analyses draw from studies that use CXR findings [4]. X-ray imaging is the preferred method of detecting pneumonia. X-rays are the most common and widely available diagnostic imaging technique [5]. Computer-aided diagnosis (CAD) can be a step towards reducing the number of deaths of pneumonia. Convolutional neural networks (CNN) have advanced greatly in the latest years in the area of image classification and segmentation and many research studies showed that they proved to be accurate in diagnosing disease [6, 7]. Deep-learning models for CAD cannot replace trained radiologists, the aim is to supplement the decision making. Besides, even though AI based solutions are getting adopted widely, there is still insufficient capacity in terms of compute power availability and trained clinicians to use them.
2. Related Work
The majority of the proposed solutions for deep-learning-aided image classification and diagnostics involves transfer learning [8]. Transfer learning is the methodology used not only for pneumonia, but also for tuberculosis [8, 9]. Recurrent neural networks are trained to describe the context of a detected disease, based on deep convolutional neural network features. Weights of already trained pairs of CNN/RNN are used on domain-specific images [10]. It is also adopted in many other fields [11]. X-ray chest images are object of numerous research studies. Hermann uses an algorithm which removes any other body parts to reduce the errors in diagnosis by leveraging a scan line optimization [12]. Generalized regression neural networks along with multilayer, probabilistic, learning vector quantization is employed for diagnosis of chest disease [13]. Deep learning is used in medical image analysis in different medical application areas: neuro, retinal, pulmonary, digital pathology, breast, cardiac, abdominal, musculoskeletal [14]. The most common technique is CNN. It has been applied to various medical imaging classification tasks and showed remarkable results in feature extraction, image classification, object detection and segmentation. Most of the adopted approaches are using very complex CNN topologies and combination of techniques to achieve a very high accuracy.

The main contribution of the current paper is the study of short and shallow CNN with fewer training epochs and to compare the performance. The model shows a custom CNN which is faster to train and achieves relatively good results above 90% and is more efficient than the majority very complex CNN proposed by most researchers. We are showing the importance of fine-tuning a model and choosing an optimization function and activation function which can result in higher accuracy.

The paper is structure in the following way: Section 3 introduces the experimental data. Section 4 shows the proposed custom CNN topologies. In Section 5 the results obtained are discussed, as well as the models evaluation. Section 6 contains the conclusion of the paper.

3. Data
The dataset used in the current experiment is provided by Dr Paul Mooney for a Kaggle competition in 2017 [15, 16].

3.1. Data Collection and Labelling
The X-ray images are 5836 in total. They are segmented into a training and test set, as well as into two categories in each subset – normal and pneumonia. No differentiation is made between viral and bacterial pneumonia. The X-rays are taken during routine clinical care. The images are graded by a clinician before adding them to the training set, so the labelling is clinically validated.

|          | Training Set | Test Set | Validation Set |
|----------|--------------|----------|----------------|
| Normal   | 1082         | 234      | 267            |
| Pneumonia| 3110         | 390      | 773            |
| Total    | 4192         | 624      | 1040           |
| Percentage| 71.5%        | 10.7%    | 10.8%          |

As shown in Table 1 seventy-one-point-five percent of the complete dataset was used as the training dataset. Ten-point-seven percent of the complete dataset was used as the testing dataset. Ten-point-eight percent of the complete dataset was used as the validation dataset. Figure 1 shows examples of both categories – normal (healthy) and pneumonia (both bacterial and viral) – from the three subsets training, test and validation.
3.2. Data Augmentation

The training images are artificially augmented. On one side, that enlarges the dataset, but also prevents overfitting. We applied rescale, zoom and vertical flip. The images are additionally resized and normalized.

4. Architecture

Two models are developed in the experimental part of the study. Our goals was to check what kind of accuracy can be achieved with relatively shallow Convolutional Neural Networks trained on 10 epochs. We have fine-tuned the models and compared the results.

4.1. Model Architecture

The two models have identical architecture as shown in table 2 and figure 2. They differ in the choice of optimizer. By making the models very similar we can do split testing and analyse the effect on model performance of a single change in the architecture, in this case – the optimizer.

| Layer Name       | Layer (type)      | Output Shape    | Param #  |
|------------------|-------------------|-----------------|----------|
| conv_01          | Conv2D            | (150, 150, 32)  | 896      |
| conv_02          | Conv2D            | (150, 150, 32)  | 9248     |
| max_pooling2d_3  | MaxPooling2       | (75, 75, 32)    | 0        |
| conv_03          | Conv2D            | (75, 75, 64)    | 18496    |
| conv_04          | Conv2D            | (75, 75, 64)    | 36928    |
| max_pooling2d_3  | MaxPooling2       | (37, 37, 64)    | 0        |
| conv_05          | Conv2D            | (37, 37, 128)   | 73856    |
| conv_06          | Conv2D            | (37, 37, 128)   | 147584   |
| max_pooling2d_3  | MaxPooling2       | (18, 18, 128)   | 0        |
| Flatten-1        | Flatten           | (41472)         | 0        |
| dense_01         | Dense             | (128)           | 5308544  |
| dense_last       | Dense             | (1)             | 129      |

With this architecture we result in total 5595681 parameters, of which all are trainable.
Figure 2. Convolutional neural network model parameters.

All the layers of the networks used were trainable, and these layers extracted the features from the images. Both models were trained for 10 epochs.

4.2. Second Model Architecture
For the first model we use Stochastic Gradient Descent as an optimizer. The learning rate, the momentum, and the weight decay were set to 0.001, 0.9, and 0.0, respectively (Table 3). For the second model “model 1a” we use Adam optimizer with comparable learning rate and momentum.

| Architecture | Image Size | Epochs | Optimizer | Learning Rate | Momentum |
|--------------|------------|--------|-----------|---------------|----------|
| Model1       | 150 × 150  | 10     | SGD       | 0.001         | 0.9      |
| Model1a      | 150 × 150  | 10     | Adam      | 0.001         | 0.9      |

5. Model Evaluation
Model evaluation is based on the following metrics – accuracy, precision, recall, loss function and AUC score.

All the evaluation metrics of the model are based on the confusion matrix, which estimates true positive, true negative, false positive and false negative for each model.

True Positives (TP) is a prediction result where the model correctly classifies the image as the positive class. These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes. E.g. if actual class value indicates that the patient has pneumonia and the predicted class tells you the same thing.

Similarly, True Negatives (TN) is a prediction result where the model correctly classifies the image as the negative class. These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no. E.g. if actual class says this patient is healthy (normal) and predicted class tells you the same thing.

False positives and false negatives, these values occur when your actual class is not the predicted class, so the prediction is wrong.

False Positives (FP) is a prediction result where the model incorrectly classifies the image as the positive class. When actual class is no and predicted class is yes. E.g. if actual class says this patient is having pneumonia but predicted class tells you that the patient is healthy.

False Negatives (FN) is a prediction result where the model incorrectly classifies the image as the negative class. When actual class is yes but predicted class in no. E.g. if actual class value indicates that this passenger survived and predicted class tells you that passenger will die.

5.1. Model Accuracy
Accuracy is an intuitive performance measure and it is the ratio of correctly predicted observation to the total observations. However, it is not always true that a high accuracy means a well performing model.
This is the case only when the datasets are symmetric where values of false positive and false negatives are almost same. That is why a model is evaluated on more parameters.

\[
\text{Accuracy} = \frac{\text{TP+FN}}{\text{TP+FP+FN}} \quad (1)
\]

\[\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (2)\]

\[\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (3)\]

5.2. Model Recall
Recall or Sensitivity is the ratio of correctly predicted positive observations to the all observations in actual class with positive value. When we are measuring recall, we answer the question -- of all the images with pneumonia, how many did we label correctly?

\[\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (2)\]

5.3. Model Precision
Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that it answers is of all the images that are labelled as pneumonia, how many actually were of pneumonia? High precision relates to the low false positive rate.

\[\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (3)\]
5.4. Model Loss
A loss function or cost function is a function that maps values of one or more variables onto a real number intuitively representing some “cost” associated with the value. An optimization problem seeks to minimize a loss function. So the lower the loss function, the more reliable is the confident prediction. The cross entropy loss increases as the predicted probability diverges from the actual class.

\[
\text{CrossEntropyLoss} = -\frac{1}{N} \sum_{i=1}^{N} y_i \log(p(y_i)) + (1-y_i) \log(1-p(y_i))
\]  

When the actual label is 1, \((y(i) = 1)\), the second half of the formula disappears. When the label is 0, \(y(i) = 0\), the first half of the formula disappears. The log of the actual predicted probability is multiplied.

![Figure 6. Comparative Model Loss Function Score.](image)

So cross entropy penalized confident predictions which are incorrect.

5.5. Model AUC Score
AUC values are in a range from 0 to 1. An AUC of 0.0 means that the predictions are all wrong. An AUC of 1.0 means that all the predictions of the model are correct. AUC is invariant in terms of scale and classification threshold. It measures the quality of the model’s predictions irrespective of scale or classification threshold choice. AUC can be considered the probability that the model ranks a random positive sample more highly than a random negative sample.

![Figure 7. Comparative AUC score.](image)

6. Experimental Results
The comparison of the two models shows that the one with Adam optimizer with the same parameters for momentum and weight decay provides consistently better performance when the training is span across 10 epochs.
Table 4. Model evaluation based on validation accuracy, precision, recall, and AUC score.

| Architecture | Validation Accuracy | Validation Precision | Validation Recall | Validation Loss | Validation AUC |
|--------------|---------------------|----------------------|-------------------|----------------|---------------|
| Model 1      | 86.35               | 84.22                | 96.03             | 33.10          | 94.61         |
| Model 1a     | 87.50               | 84.49                | 98.17             | 37.62          | 94.94         |

Table 5. Model evaluation based on testing accuracy, precision, recall, and AUC score.

| Architecture | Accuracy | Precision | Recall | Loss | AUC    |
|--------------|----------|-----------|--------|------|--------|
| Model 1      | 90.74    | 93.29     | 94.31  | 22.61| 95.97  |
| Model 1a     | 92.08    | 94.98     | 94.31  | 19.85| 96.93  |

Comparing the two models we can see that the choice of optimizer makes a difference. In our case the Adam optimizer with the same parameters as Stochastic Gradient Descent, account for improved performance. Not only the accuracy of the model is increases, but also we observe decrease in the loss function. The AUC is also higher with nearly 3 percent, which is even better performance indicator in this case, as the distribution of the dataset is not the same. There is also better result in terms of precision with Adam optimizer. Recall remains the same.

7. Discussion
The high test accuracy of the second model (model 1a) of 92.08% and AUC score of 96.93% show that even shallow convolutional network models can produce adequate classification results of pneumonia images. Having only a few blocks of Convolutional 2D layers and being trained on just 10 epochs, the model does not require so much computer power. It was trained relatively fast within 10 minutes on Python 3 Google Compute Engine Backend (GPUs), provided by Google cloud with their initiative Google Colabs. It required about 12 GB RAM and about 30 GB disk space. The Adam optimizer clearly showed better results proved by all evaluation metrics discussed in the previous section.

The limitations of the method is the insufficient available data. Deep learning models generally require more data to produce good results. Also, the training, testing and validation datasets were not balanced. With the availability of more data, it would be easier to estimate even better the performance of the models.

Further steps could be to experiment with other values of the hyper-parameters and to make deeper model and assess to what extent the depth of the model topology enhances the accuracy of the model.

8. Conclusion
Pneumonia is a major mortality threat, particularly for children under 5 and for elderly patients. Pneumonia can be prevented with early diagnosis, as it many cases it goes overlooked and treatment is applied too late. Chest X-Ray is one of the preferred way of detecting the disease. However, there is a shortage of experts who can diagnose the disease. The development of AI solutions which can supplement the decision process of clinicians can increase the chances of early diagnosis and positive outcome of the treatment. In the paper we have experimented with two convolutional neural networks, which have identical topology, but use different optimizer functions. The experimental results showed that Adam optimizer accounts for higher-performant CNN model with all other parameters being the same. The accuracy increase is 1 %, but the AUC score is nearly 3 %. Another important finding of the study is that a shallow CNN model, trained on just 10 epochs can produce very satisfactory results. Such models require less compute power and are easier to implement, so they can be a good candidate for a viable AI solution enhancing the pneumonia diagnosis decision process.
References

[1] Clinic M, Recorded Books I. Mayo clinic family health book the ultimate home medical reference. S.l.: RosettaBooks, https://rbdigital.rbdigital.com (2019, accessed 8 November 2020).

[2] World Health Organization, Department of Maternal N Child and Adolescent Health, World Health Organization. Revised WHO classification and treatment of pneumonia in children at health facilities: evidence summaries. http://apps.who.int/iris/bitstream/10665/137319/1/9789241507813_eng.pdf?ua=1 (2014, accessed 8 November 2020).

[3] Stupka JE, Mortensen EM, Anzueto A, et al. Community-acquired pneumonia in elderly patients. *Aging Health* 2009; 5: 763–774.

[4] Fancourt N, Deloria Knoll M, Barger-Kamate B, et al. Standardized Interpretation of Chest Radiographs in Cases of Pediatric Pneumonia From the PERCH Study. *Clinical Infectious Diseases* 2017; 64: S253–S261.

[5] Cherian T, Mulholland EK, Carlin JB, et al. Standardized interpretation of paediatric chest radiographs for the diagnosis of pneumonia in epidemiological studies. *Bull World Health Organ* 2005; 83: 353–359.

[6] Yasaka K, Abe O. Deep learning and artificial intelligence in radiology: Current applications and future directions. *PLoS Med* 2018; 15: e1002707.

[7] Thrall JH, Li X, Li Q, et al. Artificial Intelligence and Machine Learning in Radiology: Opportunities, Challenges, Pitfalls, and Criteria for Success. *Journal of the American College of Radiology* 2018; 15: 504–508.

[8] Hashmi MF, Katiyar S, Keskar AG, et al. Efficient Pneumonia Detection in Chest Xray Images Using Deep Transfer Learning. *Diagnostics* 2020; 10: 417.

[9] Pasa F, Golkov V, Pfeiffer F, et al. Efficient Deep Network Architectures for Fast Chest X-Ray Tuberculosis Screening and Visualization. *Sci Rep* 2019; 9: 6268.

[10] Shin H-C, Roberts K, Lu L, et al. Learning to Read Chest X-Rays: Recurrent Neural Cascade Model for Automated Image Annotation. *arXiv*:1603.08486 [cs], http://arxiv.org/abs/1603.08486 (2016, accessed 8 November 2020).

[11] Yang N, Niu H, Chen L, et al. X-ray weld image classification using improved convolutional neural network. Tianjin City, China, p. 020035.

[12] Hermann S. Evaluation of Scan-Line Optimization for 3D Medical Image Registration. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA: IEEE, pp. 3073–3080.

[13] Er O, Yumusak N, Temurtas F. Chest diseases diagnosis using artificial neural networks. *Expert Systems with Applications* 2010; 37: 7648–7655.

[14] Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Medical Image Analysis* 2017; 42: 60–88.

[15] Mooney, Paul. Chest X-Ray Images (Pneumonia), https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia/version/2 (accessed 8 November 2020).

[16] Kermanshahi D, Goldbaum M, Cai W, et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell* 2018; 172: 1122-1131.e9.