Machine Learning Engineer Nanodegree

Capstone Proposal: Chest X-ray Pneumonia Classifier

Ilona Brinkmeier
July 30st, 2019

Domain Background

This capstone project is chosen from the medical healthcare area, because I worked in this domain for several years and have detailed knowledge about it. So, it combines medical domain knowledge with newly learned machine learning techniques. Artificial intelligence algorithms have the potential to change the medical workflow being common in the future[1].

The implemented deep learning concept of this project shall solve the binary classification problem if chest X-ray images are being normal or pneumonia ones. Pneumonia is an infection of the lungs caused by microbes. As stated by the World Health Organisation (WHO)[2], „pneumonia is the single largest infectious cause of death in children worldwide, accounting for 16% of all deaths of children under five years old.“
Regarding this status about children mortality having a pneumonia, this dataset and its usage for having a fast, automatically created diagnosis would save lifes of the children triggering proper, faster treatments. For a child having a pneumonia, this may lead to death, especially if the immune system is already compromised, e.g. bei HIV, tubercolosis or subnutrition. According the mentioned WHO fact-sheet „Pneumonia affects children and families everywhere, but is most prevalent in South Asia and sub-Saharan Africa.“ There the healthcare coverage of the whole area is not the same compared to western countries like in Europe or North-America. Having an automatic estimation would help to trigger necessary medical or other interventions, even when no radiologist is available. In geographical regions with a low amount of such medical experts, this software application could be used by other medical stuff as well and further actions could be taken to protect the children.

It would improve the general medical workflow too: In a general hospital with a normal medical workflow, radiologists are doing this classification manually on their viewing stations connected with the picture archiving and communication system (PACS modality)[3]. A medical PACS system is part of the hospital information system where all images are stored in a so called DICOM format (Digital Imaging and Communications in Medicine)[4].
The classification deep learning method can ease the medical workflow and increase the diagnosis correctness rate, because an estimation can be given automatically and efficiently for thousands of images in a short time, even difficult ones not easy to diagnost manually. As a whole, this would not be possible for a radiologist: during a long time viewing process being more fatigue and unfocussed is a normal human reaction and therefore correctness may decrease. Additionally, today there is already a shortage of needed radiologists. Each existing one has much more work to do than in former times. As mentioned in the 'AI in Healthcare' journal[5], for radiologists artifical intelligence methods like deep learning are going to „expedite and improve their ability to interpret images“.

1 https://www.healthcatalyst.com/prescriptive-analytics-improving-health-care
2 WHO & Pneumonia, https://www.who.int/news-room/fact-sheets/detail/pneumonia
3 Acronym with several meanings, here in the medical domain: https://en.wikipedia.org/wiki/Picture_archiving_and_communication_system
4 DICOM - https://www.dicomstandard.org/
5 https://trimed-cdn.s3.amazonaws.com/digitalissues/aih/2019/2019_02/index.html article 'Embrasing AI: Why Now Is The Time For Medical Imaging'

Means, having such kind of diagnosis pre-processing for the radiologists, they can take over other, potential new tasks, like communication with patients directly. So, this is a general workflow improvement using such dataset and implementation method and would lead to better patient care[6].
Furthermore, as a new personal insight, for unclear cases having a lower estimation result, the deep learning algorithm can be used to improve the image quality as well, if an improper post-processing software preset has been used for the specific patient image. Afterwards, a new classification action could be done automatically. It would ease the diagnostic and communication tasks as well by having a better image quality.

Problem Statement

Chest X-ray images shall be classified being normal or pneumonia ones by using convolutional neural networks (CNN). So, we have a binary classification problem. As input of the basic network model, converted .jpeg compressed images based on their original chest X-ray .dcm DICOM images, the algorithm will identify an estimate of the image status showing a pneumonia or not.
An already existing solution is described in the Cell journal paper 'Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning'[7]. There as well, deep learning CNN models have been implemented, by using the official ImageNet database from www.image-net.org as input for transfer learning techniques and as a final result, delivering specific classification estimations of having a bacteria or virus root cause of the pneumonia.

Datasets and Inputs

Implementing such kind of binary classifier, a properly labelled dataset is necessary. The used Kaggle dataset[8] delivers normal or pneumonia labelled Chest X-ray images, separated in associated training, testing and validation samples. As mentioned, these images are converted to the .jpeg image format and as a consequence, private individual data information sets don't exist. The pneumonia images include a root cause marker as part of their stored file name (bacteria or virus). Regarding the data set, data preparation is not necessary anymore.

The whole dataset includes 5857 .jpeg image files, all of them with the most common X-ray posterior-anterior orientation and mostly child examination results. The dataset cannot be structured in age ranges for babies, toddlers, younger children, teenagers or adults properly. This would only be possible by reconverting the images to the original .dcm DICOM format, which is not allowed according several local regulatory data protection rules of the countries.

Solution Statement

The answer to the question of having a pneumonia chest X-ray image or not, is technically a binary classification issue. Because we are dealing with images, convolutional neural networks are a state-of-the-art solution concept for such kind of task[9]. Having properly labelled datasets for training, testing and validation of the neural networks, such CNNs can deliver a propability estimation for each future Chest X-ray image being a member of one class or the other.

The improved solution CNN model shall use transfer learning by having bottleneck-features. By

6 AI in Healthcare, journal volume 2 number 1 from 2019 and its article:
'Embracing AI: Why now is the time for medical imaging' by Mary C. Tierney, MS (https://trimed-cdn.s3.amazonaws.com/digitalissues/aih/2019/2019_02/index.html)
7 https://www.cell.com/action/showPdf?pii=S0092-8674%2818%2930154-5
8 Kaggle dataset: https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia/version/2
9 Paper 'Going deeper with convolutions' by Christian Szegedy, Google Inc., et all.

now, it is unclear if as input the ImageNet database can be used to create bottleneck-features or if such features have to be created with the given Chest X-ray dataset too. Additionally, an augmentation method could improve the classification results.

Benchmark Model

We are starting with a CNN model from the scratch, having several hidden layer blocks from the Keras library, with Relu activation function and a simple optimiser.
As improvement, in the hidden layer each block beside a Dense sublayer includes additional BatchNormalization and DropOut sublayers from Keras[10] and for compilation, a new optimiser. Adam is used with a proper loss function like binary-cross-entropy as it is state-of-the-art today, realised by the champions of Kaggle competitions and papers[11]. As described there, Adam combines the advantages of AdaGrad and RMSProp and is very efficient.

Such model shall be improved by using augmentation only, afterwards a model with transfer learning techniques shall be created and if needed, again augmentation methods applied on that model. As mentioned, it is unclear if the transfer learning bottleneck-features can be created. Such weight values would be new and not available by now. As a summary, few model types of different complexity are created, compared regarding their prediction results via evaluation metrics.

Evaluation Metrics

The models performance has to be evaluated by metrics. This project deals with a classification problem therefore for the CNN models as metric 'Accuracy', 'Precision' and others can be used. After the model is evaluated, we are reporting the classification accuracy on the train and test sets. Line plot diagrams are visualising the learning curves of the loss function on the train and test sets and the other diagram their classification accuracy.

Accuracy is the correctness of the binary classification: it measures how often the classifier makes the correct prediction. It's the ratio of the number of correct predictions to the total number of predictions

$$(TP+TN)/(TP+TN+FP+FN)$$

where: TP = True positive; FP = False positive; TN = True negative; FN = False negative

Precision quantifies the binary precision. It is a ratio of true positives (images classified as pneumonia ones, and which are actually pneumonia) to all positives (all images classified as pneumonia ones, irrespective of whether that was the correct classification), in other words it is the ratio of

$$[True\ Positives\ /\ (True\ Positives\ +\ False\ Positives)]$$

Recall (sensitivity) tells us what proportion of images that actually were pneumonia ones were classified by us as pneumonis ones. It is a ratio of true positives to all the images that were actually pneumonia ones, in other words it is the ratio of

$$[True\ Positives\ /\ (True\ Positives\ +\ False\ Negatives)]$$

To combine the accuracy and precision information a **confusion matrix** is used to summarize error type I and II.

A model's ability to precisely predict those that have a pneumonia is more important than the model's ability to recall those individuals. We can use **F-beta score** as a metric that considers both

10 The Keras API can be find with https://keras.io/
11 ICLR paper link: https://arxiv.org/abs/1412.6980

precision and recall:

$F_\beta = (1+\beta^2) \cdot (precision \cdot recall \;/\; ((\beta^2 \cdot precision) + recall))$

In particular, when $\beta = 0.5$, more emphasis is placed on precision.

To visualise the binary classification performance of the relative frequency of the correct and false classified images we use the **ROC curve**. The **ROC-AUC value** is delivered, which computes the 'Area Under the Receiver Operating Characteristic Curve (ROC AUC)' from prediction scores[12]. Its best value is one, means having a perfect classification.

Project Design

In general, to solve a business problem the overall solution process starts with business understanding, followed by understanding the existing data[13].

In this case, having a binary classification problem to solve with a predictive machine-learning algorithm and an associated, already prepared image dataset available, we just decide to use Convolutional Neural Networks, or CNNs for short. They map image data – spatial relationships are taken into account – to our desired prediction estimation as output value.

The general architecture of CNNs looks like[14]

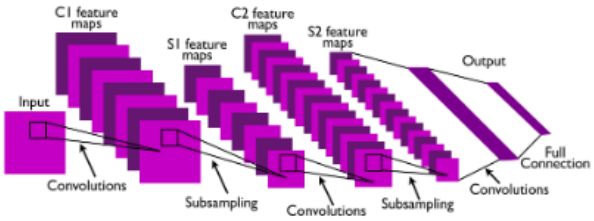

Fig. 1. A typical ConvNet architecture with two feature stages

As a basic implementation example, the following one having several hidden layers shall be used for benchmarking:

Layer (type)	Output Shape	Param #
conv2d_1 (Conv2D)	(None, 223, 223, 16)	208
max_pooling2d_1 (MaxPooling2	(None, 111, 111, 16)	0
conv2d_2 (Conv2D)	(None, 110, 110, 32)	2080
max_pooling2d_2 (MaxPooling2	(None, 55, 55, 32)	0
conv2d_3 (Conv2D)	(None, 54, 54, 64)	8256
max_pooling2d_3 (MaxPooling2	(None, 27, 27, 64)	0
global_average_pooling2d_1 (	(None, 64)	0
dense_1 (Dense)	(None, 133)	8645

Total params: 19,189.0
Trainable params: 19,189.0
Non-trainable params: 0.0

As visibile, a CNN includes 3 types of layers: Convolutional, Pooling and Fully-Connected Layers. This architecture can be improved by adding further sublayers, using augmentation and transfer learning methods. For this modified CNN, the resulting prediction estimation values are much better compared to the basic architecture which is analysable by specific metrics.
For dichotomous classifications and their model solutions, specific evaluation metrics exists as well to find the best model having the best prediction in that group. Common ones are accuracy, precision, recall (sensitivity), F-beta score, ROC curve and its ROC-AUC value.

12 Implementation help: https://scikit-learn.org/stable/modules/model_evaluation.html#classification-metrics
13 Image source: SAP course 'Getting started with Data-Science'
14 Image source: http://yann.lecun.com/exdb/publis/pdf/lecun-iscas-10.pdf