Birkbeck, University of London

Pneumonia Detection from Chest X-Ray Images Draft

Author:
Baran Buluttekin

Supervisor:
Dr. George Magoulas

A thesis submitted in fulfillment of the requirements for the degree of MSc Data Science

in the

Department of Computer Science

April 2, 2019

Abstract

This is placeholder text. To add more information type it after this line.

Declaration

I hearby declare this file a text.

```
    import pymongo as pm
    import json
    # line comment
    with open("DSTA/Lab/mongo.json") as f:
    url = json.load(f)
    """ Comment 1"""
    "Comment 2"

    c = pm.MongoClient(url["url"])

    def Myfunc(x):
        print(x)

    print(c.admin)
```

and text goes on.

Contents

1 Introduction

Pneumonia is swelling (inflammation) of the tissue in one or both lungs. It is usually formed at the end of breathing tubes of the lungs and cause these tubes to inflame and fill up with fluid. In the UK, pneumonia effects around 8 in 1000 adults each year [17]. Global economic cost of pneumonia has estimated at $17 billion annually [12]. Currently detecting pneumonia cases heavily relies on chest X-ray image examination which requires expert radiologists to diagnose. Building intelligent system to diagnose the pneumonia can help health care services to increase efficiency, reduce costs and could help increase early diagnoses in countries with inadequate access to health care.

1.1 Related Work

There are number of research has been published about lung diseases related detection. Most prevalent ones are the CheXNet [18] and ChestX-ray8 [19], both of these research carried out by training on same dataset ChestX-ray8 [19]. ChestX-ray8 comprises of approximately 100,000 frontal view chest X-ray images labelled by extracting information from the accompanied radiologists notes with using variety of different NLP (Natural language processing) techniques from the openi[14] database. ChestX-ray8 authored by researchers from National Institute of Health (NIH) and published at 2018. Most profound effect of this paper is the creation of the ChestX-ray8 dataset which has become one of the widely used dataset in computer vision research related to lung diseases. More detailed information about the dataset can be found in dataset section of this proposal.

CheXNet is another related article authored by researchers from Stanford University ML group. Prediction of lung diseases achieved by 121 layer convolutional neural network and designed to predict 14 pathologies in the ChestX-ray8 dataset. One of the major importance of this paper is the setting the setting benchmark for human level detection for chest X-ray images. One of the most fundamental difference of the X-ray related disease prediction is the definition of human level accuracy. Due to the nature of required expertise in X-ray images leaves general public out of the scope when it comes to human level performance of these pathologies. Anyone who have not been trained in radiology will not be able to detect any lung diseases in the Chest X-ray images. For example the image below is sample of two chest X-ray images almost indistinguishable to general audience.

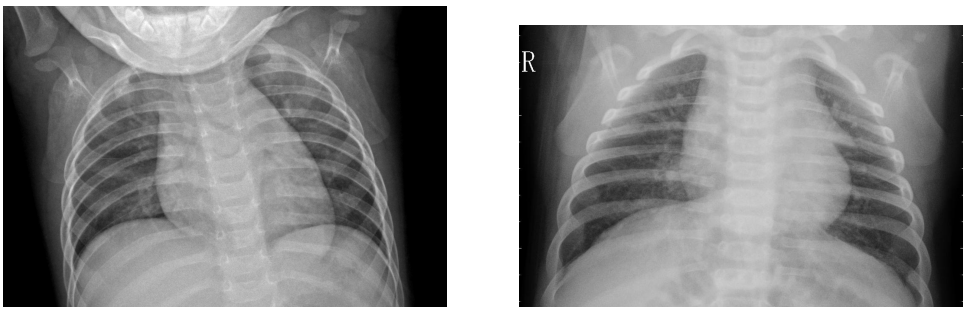

Figure 1: Two sample X-ray Chest images.

Given this challenge authors of the CheXNet conduct a test to establish benchmark for radiologists. They have collected 420 frontal chest X-rays and asked practicing radiologists in Stanford University to label them for all 14 pathologies. Radiologists selected with different range of experience, and had 4, 7, 25 and 28 years of experience. X-ray images presented to radiologists without any patient information or any symptoms experienced by patients and their diagnoses predictions measured based on underlying state of the X-ray patients. Fallowing is the table showing the summary statistic of this test for the 4 radiologists participated to test on F1 score, which is harmonic average of precision and recall.[18]:

	F1 Score (95% CI)
Radiologists 1	0.383 (0.309, 0.453)
Radiologists 2	0.356 (0.282, 0.428)
Radiologists 2	0.365 (0.291, 0.435)
Radiologists 4	0.442 (0.390, 0.492)
Radiologists Avg	0.387 (0.330, 0.442)

Table 1: Radiologist prediction performances from CheXNet.

Importance of this test is that it gives us a rough estimate for human level accuracy benchmark to assess the model performance for new detection models.

2 Dataset

Choosing and processing dataset have a crucial importance on success of the any machine learning task. There are several dataset available online that relate to chest X-Ray images. Given the large number of choices for selecting the dataset there are few criteria important to check while deciding the final dataset.

2.1 General Guidelines While Deciding on the Dataset

In this section I have highlighted my reasons for deciding on the dataset of choice for this research project. Main points for decision are:

1. **Reproducibility:** Dataset of choice must allow reader to reproduce the work in order to assess all the points discussed in the report. That would require dataset to be public.

2. **Labelling:** Dataset must contain labels of patients state. Such as being diagnosed with pneumonia or not.

3. **License:** Dataset should have a license that permits research.

I will be evaluating dataset available while considering general guidelines outlined above.

2.2 OpenI Database

OpenI[14] is a database that is service of National Library of Medicine. It enables search for medical images, graphs and charts through text as well as image query. As of writing of this proposal it has over 3.7 million images, and 3,955 radiology reports. It is the main source for ChestX-ray8 dataset mentioned previously in Related Work section. Despite the fact it contains very large data for chest X-rays, this source is not suitable for this project due to the fact that images does not includes labels for the patients state (e.g., Pneumonia or normal). Chest X-ray data in this database is the image accompanied by radiologist report which is advisory document. Mainly because the lack of labels this dataset is not suitable for this project where the choice will be a supervised classification task.

2.3 ChestX-ray8

This dataset created part of the ChestX-ray8[19] paper (Also known as ChestX-ray14). Original source of this dataset is OpenI[14] medical database as mentioned in previous subsection. Authors first short-listed eight common thoracic pathologies subsequently related X-rays searched from the database based on these pathology keywords. Most positive quality of this dataset amongst the other options is the sheer quantity of the data points which is by far the largest in size. This attribute especially important when it comes to certain computer vision techniques such as Neural Networks due to the fact that large datasets increases the variance and enables better generalization. Despite this positive points, quality

of the labelling of this dataset has come to questioning by radiologists[13]. Due to this considerable labelling inconsistencies this dataset have not been chosen for this project.

2.4 Cell Press Research

Shortcomings of first two datasets compelled me to searched further for new dataset that does not have the problems I point out previously. Research from Cell Press[7], together with the data that made public, provided a solution to these problems. Dataset the team released contains 5856 hand labelled X-ray chest images from children aged between one to five years old. X-ray images collected as part of the routine clinical care at the Guangzhou Women and Children's Medical Center, Guangzhou, China. All images screened for quality control and low quality or unreadable x-rays discarded. Labels of the images also checked by two expert physician and only approved images included in the dataset. Finally this dataset released under Creative Commons license Attribution 4 (CC BY 4.0) that allow copy, distribution of the material as well as transform and building upon material for any purposes. In light of all these attributes I choose this dataset for my project.

3 Computer Vision

Importance of X-ray image analysis in pneumonia diagnoses clearly highlights that this is a computer vision problem. In essence computer vision is a scientific field aims to automate vision task usually performed by humans. Vision on earth begin approximately 543 million years ago when trilobites developed basic vision system [8]. Development of vision helped increase the specie variation and development on earth significantly in respect to reproduction, finding food and many other reasons. Largely due to this very important nature of vision, research in how vision performed in species and how to automate the vision task gained a lot of attraction. Early research in this field inspired by the biological vision system. More specifically model of mammal neural system in respect to vision was the center point. In 1959 experiment on cat visual system by Hubel and Wiesel [5] highlighted the inner functioning of vision by discovering effect of dark edges causing activation in visual cortex. They also concluded that visual cortex passing this signals from detected edges to later centers of the brain where those edges are combined to represent more complex shapes. This idea of simple to more complex neural structure inspired Japanese computer scientist Kunihiko Fukushima to propose system he called *Neocognitron* [3]. In the article he laid out simple to complex neural architecture which was very similar to convolutional neural network (CNN).

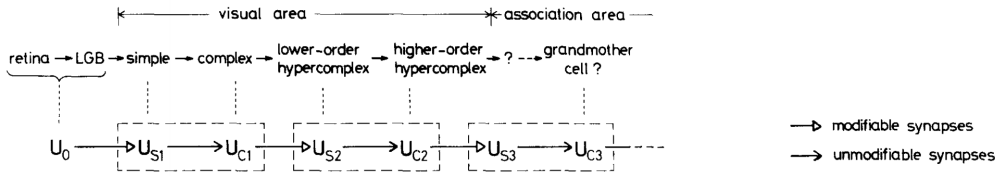

Fig. 1. Correspondence between the hierarchy model by Hubel and Wiesel, and the neural network of the neocognitron

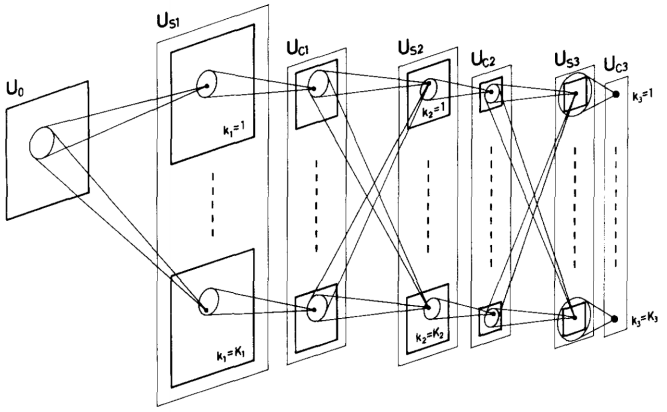

Fig. 2. Schematic diagram illustrating the interconnections between layers in the neocognitron

Figure 2: Network architecture from Neocognitron.

3.1 Convolutional Neural Networks (CNN's)

Although Neocognitron formed the general idea of the convolutional neural networks, it did not used the method called backpropagation that CNN's use today. Backpropagation is technique used in neural networks to propagate errors through

the layers of neural network. First CNN that have the attributes same as current CNN's was build in Bell Labs in 1989 to recognize the hand written digits of the zip codes [9]. Convolutional neural network name indicates that network uses mathematical operation called **convolution**. Explaining in a simple way using definition in the Deep Learning book [4]:

> "Convolutional networks are simply neural networks that use convolution in place of general matrix multiplication in at least one of their layers."

Convolutional neural networks are generally a good option for image or time series data problems because of the sliding window approach capturing the underlying signal.

3.2 Prominent Computer Vision Architectures

In this project I will be taking advantage of the well known network architectures for computer vision for general benchmarking purposes. These architecture generally known for their good performance in image classification competitions hence, any new design should perform better than these designs to be considered.

3.3 LeNet-5

LeNet-5 [10] is 7 layered neural network build for classifying hand written digits in checks. Numbers in checks turned into 32×32 pixel images.

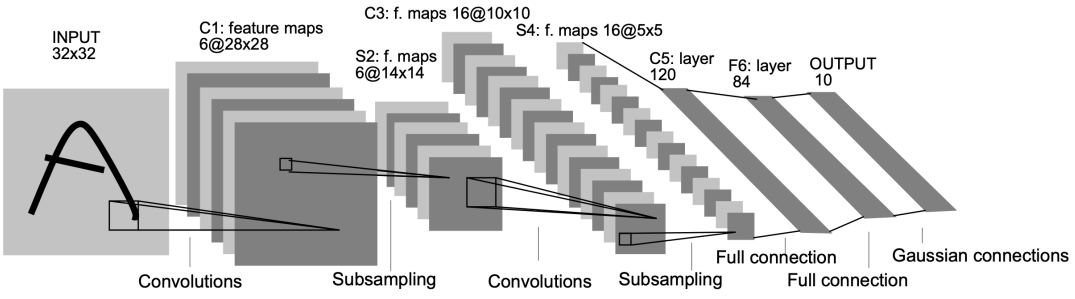

Figure 3: Network architecture of LeNet-5.

3.4 AlexNet

3.5 VGGNet

4 Project Aims and Objectives

Objective of this project is to lay out and highlight general roadmap for computer vision problems, specifically it relates to classification along with the prototyping and experimenting of different neural network models. Highest possible accuracy will be aimed but due to highly iterative and time consuming nature of the neural network research, its not very likely that it will beat general benchmark set by most recent research.

Fallowing sub-objectives will also be attempted if the time permits:

- Discovery of both economically and computationally efficient deployment methods.
- Application of different deployment models and their comparison.
- Fully functional pipeline that implements CI/CD principles to experimentation and deployment.

5 Tools and Techniques

For the implementation of the aims of this project python programming language is chosen as a main programming language. Reasons for this decision is two fold, first part of the reasoning is need of high level programming language. Low level programming languages such as Java or C++ are not well suited for computer vision tasks such as this project due to reason of their time consuming prototyping cycles. Second part of the decision python being de facto language of choice for majority users which enables more tools and techniques being available for application.

I will also make use of external open source machine learning packages because of the intensive computational nature of the Neural networks. Number of parameters for some of the well known neural network architectures reaches to hundreds of thousands or in some cases in millions or billions. Therefore any code that implements these architecture required to be well optimized and preferably parallelized to run in highly performance hardware, such as the GPU units. Building a code base that achieves this standards require significant amount of time and resource, hence out of the scope of this project.

Open source projects and their intended form of usage in this project fallows:

- **Scikit-Learn [16]:** For the well versed library of machine learning algorithms and tolls such as train / test split for dataset.
- **Pandas [11]:** For general data manipulation.
- **Tensorflow [1] or PyTorch [15]:** For the implementation of the neural network.
- **Matplotlib [6] and Seaborn [20]:** Visualizing data and calculations.
- **Keras [2]:** Keras will be used to implement benchmark architectures such as LeNet-5 or AlexNet.

6 Project Plan

Early draft to be completed later.

References

[1] Martin Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: https://www.tensorflow.org/.

[2] François Chollet et al. *Keras*. https://keras.io. 2015.

[3] Kunihiko Fukushima. "Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position". In: *Biological Cybernetics* 36 (1980), pp. 193–202.

[4] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[5] David H. Hubel and Torsten N. Wiesel. "Receptive Fields of Single Neurons in the Cat's Striate Cortex". In: *Journal of Physiology* 148 (1959), pp. 574–591.

[6] J. D. Hunter. "Matplotlib: A 2D Graphics Environment". In: *Computing in Science Engineering* 9.3 (May 2007), pp. 90–95. ISSN: 1521-9615. DOI: 10.1109/MCSE.2007.55.

[7] Daniel Kermany and Kang Zhang. *"Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification"*. 2018. DOI: http://dx.doi.org/10.17632/rscbjbr9sj.2.

[8] E. N. K LARKSON. "The visual system of trilobites". In: *Palaeontology* 22 (1979), pp. 1–22.

[9] Y. LeCun et al. "Backpropagation Applied to Handwritten Zip Code Recognition". In: *Neural Comput.* 1.4 (Dec. 1989), pp. 541–551. ISSN: 0899-7667. DOI: 10.1162/neco.1989.1.4.541. URL: http://dx.doi.org/10.1162/neco.1989.1.4.541.

[10] Yann Lecun et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE*. 1998, pp. 2278–2324.

[11] Wes McKinney. "Data Structures for Statistical Computing in Python". In: *Proceedings of the 9th Python in Science Conference*. Ed. by Stéfan van der Walt and Jarrod Millman. 2010, pp. 51–56.

[12] Girish B. Nair and Michael S. Niederman. "Community-Acquired Pneumonia: An Unfinished Battle". In: *Medical Clinics of North America* 95.6 (2011). Pulmonary Diseases, pp. 1143–1161. ISSN: 0025-7125. DOI: https://doi.org/10.1016/j.mcna.2011.08.007. URL: http://www.sciencedirect.com/science/article/pii/S0025712511000927.

[13] Luke Oakden-Rayner. *Exploring the ChestXray14 Dataset Problems*. Accessed: 2019-03-25. URL: https://lukeoakdenrayner.wordpress.com/2017/12/18/the-chestxray14-dataset-problems/.

[14] *Open Access Biomedical Image Search Engine*. https://openi.nlm.nih.gov/. Accessed: 2019-03-12.

[15] Adam Paszke et al. "Automatic differentiation in PyTorch". In: (2017).

[16] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[17] *Pneumonia*. URL: https://www.nhs.uk/conditions/pneumonia.

[18] Pranav Rajpurkar et al. “CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning”. In: *CoRR* abs/1711.05225 (2017).

[19] Xiaosong Wang et al. “ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases”. In: *CoRR* abs/1705.02315 (2017). arXiv: 1705.02315. URL: http://arxiv.org/abs/1705.02315.

[20] Michael Waskom et al. *mwaskom/seaborn: v0.8.1 (September 2017)*. Sept. 2017. DOI: 10.5281/zenodo.883859. URL: https://doi.org/10.5281/zenodo.883859.