Machine Learning Engineer Nanodegree

Capstone Proposal

Duy Do Van
January 5st, 2018

Diseases Detection from Chest X-ray data

1. Domain Background

The world is changing so fast that the pressure on health is increasing, the bad changes of climate, the environment, the life way of human, ... also increase the risk as well as diseases for people. One of the issues that we will focus on in this article is lung diseases.

About 3.2 million people succumbed in 2015 to chronic obstructive pulmonary disease (COPD), caused mainly by smoking and pollution, while 400,000 people died from asthma [1].

With so many lung diseases people can get, here is just one example of diseases we can save if we find them out earlier.

With the technology machine and computer power, the earlier identification of diseases, particularly lung disease, we can be helped to detect earlier and more accurately, which can save many many people as well as reduce the pressure on the system. The health system has not developed in time with the development of the population.

I would also like to thank the scientists who have gone ahead in their research to offer to humanity, applying machine learning to the problem of X-ray image prediction [2345678].

With the power of computers as well as the large amount of data being released to the public, this is a good time to contribute to solving this problem. Wishing to contribute more to the community, helping those who are not able to pay for medical expenses, I hope that my solution can contribute to reducing medical costs, the development of computer science for medical projects.

2. Problem Statement

Recently a large dataset of X-ray lung data was public on Kaggle followed by labeled lung disease data. This is a good condition for me to implement this project.

In this project I will conduct a study and analysis of this data set, then apply Machine Learning and DeepLearning to predict that the patient has a lung disease and so what are the diseases. This project is a binary classification with input is patient's data (age, gender, X-ray images, View Position) and ouput is found diseases or not.

The difficulty this is a new dataset, and I will be one of the pioneers to learn it, my analysis is that this is a large dataset but has never been processed full, data has a lot of noise, and X-ray of the lung is not likely to provide enough information to assess whether a patient may be ill.

I will use Machine Learning as well as DeepLearning to process data as well as create models for diagnosing patients. My keys point here will be: combining the processing of patient information with data from X-rays, using CNN with the well-known pre-trained model, first time using the CapsNet [910] network for data this form.

3. Datasets and Inputs

The project uses a large dataset that is public on kaggle, because the dataset is large so I will test my methods first on the sample data set:

a) Sample dataset [11]**:**

- File contents: this is a random sample (5%) of the full dataset:

  - sample.zip: Contains 5,606 images with size 1024 x 1024
  - sample_labels.csv: Class labels and patient data for the entire dataset

- Class descriptions: there are 15 classes (14 diseases, and one for "No findings") in the full dataset, but since this is drastically reduced version of the full dataset, some of the classes are sparse with the labeled as "No findings": Hernia - 13 images, Pneumonia - 62 images, Fibrosis - 84 images, Edema - 118 images, Emphysema - 127 images, Cardiomegaly - 141 images, Pleural_Thickening - 176 images, Consolidation - 226 images, Pneumothorax - 271 images, Mass - 284 images, Nodule - 313 images, Atelectasis - 508 images, Effusion - 644 images, Infiltration - 967 images, No Finding - 3044 images.

b) Full dataset [12]**:**

- File contents:

  - images_00x.zip: 12 files with 112,120 total images with size 1024 x 1024
  - README_ChestXray.pdf: Original README file
  - BBox_list_2017.csv: Bounding box coordinates. *Note: Start at x,y, extend horizontally w pixels, and vertically h pixels*
  - Data_entry_2017.csv: Class labels and patient data for the entire dataset

- Class descriptions: there are 15 classes (14 diseases, and one for "No findings"). Images can be classified as "No findings" or one or more disease classes: Atelectasis, Consolidation, Infiltration, Pneumothorax, Edema, Emphysema, Fibrosis, Effusion, Pneumonia, Pleural_thickening, Cardiomegaly, Nodule Mass, Hernia.

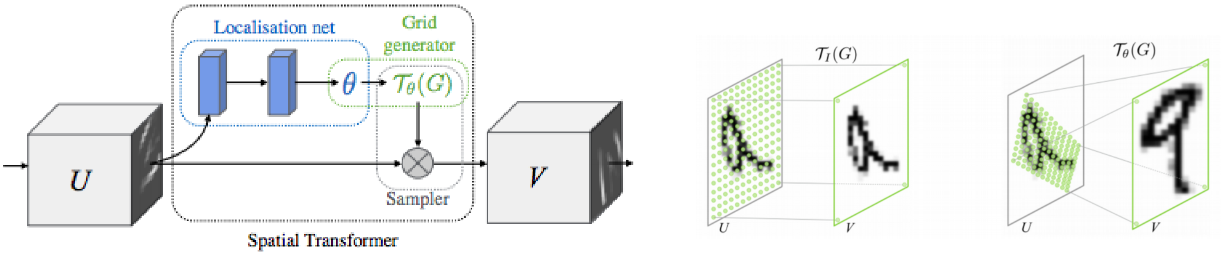
The data set contains useful information for the model we will build as: age, gender, patient data, snapshot data as well as X-ray images. From this main information I will use it to model the model.

In diagnosis from X-ray data, the physician can diagnose a part of the patient's medical condition, so I think that with the X-ray chest image data, the machine can support Assist in the diagnosis of the disease, some data on age and gender will also be considered to increase the accuracy of this system.

4. Solution Statement

Here I will divide many small problems and implement them with different methods. In this project from the above data, t will diagnose whether the patient is ill and if the disease is what disease. That is, there are two small problems: The binary classification problem with two classes is sick and not sick and the binary classification problem for each disease is the corresponding disease or not.

I will use image processing techniques to process data, using Spacial Transform [13] to get the main information in the image.



The model will be built on two main algorithms, CNN and CapsNet, which combine data from both image data and data on age, gender and image capture. CapsNet will first be tested for X-ray data. The traditional CNN network is very powerful in image processing so I will test it first to see if CNN is good for this dataset. Recently, Hinton's new pubnish paper on a new network is very powerful when defining the So here with two shots of X-ray images, I think CapsNet can capture the signs of the disease in multiple directions.

The main libraries that will be used will be OpenCV, scikit-learn, keras / tensorflow, I will use the p2 VM on AWS.

5. Benchmark Model

For this problem, the benchmark model will be vanilla CNN model, I will try to beat its performance with other algorithms. So I want to achieve accuracy is about 70% of the problem of binary classification is disease or not, and will continue to study to upgrade the algorithm in the next time to the problem is specific to each disease.

6. Evaluation Metrics

The evaluation metrics used here will be precision, recall and F-beta scores (with beta being defined during the analysis of the data) for binary classification – found diseases or not. In this case F score is better than accuracy because with binary classification found diseases or not, the classes are imbalanced. For example, consider you have a trivial classifier that just guesses the majority class, it will obtain 80% accuracy when there is an 80/20 split and 50% accuracy when there is a 50/50 split.

These indexes will be evaluated on a separate testing data set from the original dataset.

These indicators will be evaluated for all diseases – found disease or not.

7. Project Design

- Data analysis, data processing:

Data processing such as standardized age in digital form per year, age noise filtering, one hot attributes such as gender, snapshot of the image as well as the specific type of illness that the patient suffers.

Analysis of data such as age, gender, and photo-taking will affect the likelihood that a patient will develop a specific disease.

Image processing such as resizing images to the same size, for black and white or color images for parallel processing and comparison.

Split the data into trainingset - validationset - testingset at different rates: in the sample dataset will divide proportionally 60% trainingset – 20% validationset – 20% testingset because this dataset is small with 5,606 samples, and full dataset have proportionally 80% trainingset – 10% validationset – 10% testingset because this is huge 112,120 samples.

- Modeling with CNN:

Test the CNN with the original architecture to make sure CNN is catching the paternity of the diseases, then accelerate the convergence by using a pre-trained model.

Optimize CNN with momentum, spacital transform, SGD.

Testing multiple CNNs with a variety of diseases and with models that identify patients with the disease.

For the dataset sample, you may have to use some method to increase the data.

Choose the best flow to train on the full dataset.

- Modeling with CapsNet and comparing it to CNN:

Follow the same steps for the CNN section just instead of applying CNN I would use the CapsNet network

Comparing the results between CNN vs. CapsNet

References

1. https://medicalxpress.com/news/2017-08-lung-diseases-million.html
2. Andrew Ward, Nicholas Bambos. *Quantum Annealing Assisted Deep Learning for Lung Cancer Detection*. http://cs231n.stanford.edu/reports/2017/pdfs/534.pdf
3. Albert Chon, Niranjan Balachandar. *Deep Convolutional Neural Networks for Lung Cancer Detection*. http://cs231n.stanford.edu/reports/2017/pdfs/518.pdf
4. Kingsley Kuan, Mathieu Ravaut, Gaurav Manek, Huiling Chen, Jie Lin, Babar Nazir, Cen Chen, Tse Chiang Howe, Zeng Zeng, Vijay Chandrasekhar. *Deep Learning for Lung Cancer Detection: Tackling the Kaggle Data Science Bowl 2017 Challenge*. https://arxiv.org/abs/1705.09435
5. https://www.nature.com/articles/srep46479
6. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5569872/
7. https://stephenson.pure.elsevier.com/en/publications/computer-aided-lung-cancer-diagnosis-with-deep-learning-algorithm
8. https://www.researchgate.net/publication/301651102_Computer_aided_lung_cancer_diagnosis_with_deep_learning_algorithms
9. *Matrix capsules with EM routing.* https://openreview.net/forum?id=HJWLfGWRb&noteId=HJWLfGWRb
10. Sara Sabour, Nicholas Frosst, Geoffrey E Hinton. *Dynamic Routing Between Capsules.* https://arxiv.org/abs/1710.09829
11. *NIH sample Chest X-rays dataset*, https://www.kaggle.com/nih-chest-xrays/sample
12. *NIH full Chest X-rays dataset*, https://www.kaggle.com/nih-chest-xrays/data
13. Max Jaderberg, Karen Simonyan, Andrew Zisserman, Koray Kavukcuoglu. *Spatial Transformer Networks*. https://arxiv.org/abs/1506.02025
14. https://www.womenshealth.gov/a-z-topics/lung-disease
15. https://www.cdc.gov/cancer/lung/statistics/