Spring 2017 Intelligent Systems Lung Cancer Prediction Project

CS 4620, Intelligent Systems
February 10, 2017
Dr. Schafer

University of Northern Iowa
Computer Science

Samuel Nelson
Ryan Giarusso
James Fortino
Charles Neff

Table of Contents

Executive Summary

In 2016, the American Cancer Society estimates that 155,870 of the 222,500 people diagnosed will die from lung cancer. These numbers are astoundingly high, giving lung cancer the title as the deadliest form of cancer discovered. Despite these sobering facts, lung cancer is surprisingly curable as long as the disease is detected at the earliest stages of development. In fact, the survival rate for those diagnosed at stage 1a is about 49%. Though still a low number, when compared to the 14% survival rate of stage 3a, 5% for 3b, and >1% for stage 4, one will quickly see the importance of such early corrections. Unfortunately, with current methods, early detection is quite time consuming and costly reaching around $12 billion in health care costs total.

Because of the above facts, an initiative to produce a decades worth of lung cancer research within five years. A very large milestone for this goal is the 2017 Data Science Bowl, where teams around the world, including our own, compete to create an artificial intelligence agent capable of recognizing patterns found in thousands of lung scans in order to determine whether the set of lungs has cancer or not.

Our system will take in images of lung scans known as dicoms, process them into a standardized state, and run them through a neural network that will analyze the dicoms and offer a prediction on the state of the lungs. In order to accomplish this task, we will make the use of deep learning libraries such as pydicom in order to process the scans, tensorflow to create our machine learning systems, and others.

Introduction

Our project is to design and create a method that can be utilized in today's medical facilities in the hopes to predict if a patient would potentially develop cancer within the next twelve months. Our overall goal, at least a starting goal, is to achieve roughly seventy-five percent prediction accuracy. To achieve such a goal, we have decided to break down this goal even further, smaller goals that make up the overall goal, or subgoals if you will. Three of the subgoals we shall be focusing on are the following: Processing dicom files, files the computerized tomography or CT scan produces, being able to utilize image recognition within dicom files in order to create a standardized image, and to eventually predict potential cancer risks. The last subgoal, prediction, is of course, our main concern. However, in order for our main goal to be met, the seventy-five percent accuracy, we first must be able to produce a prediction, let alone have the prediction have any accuracy to it.

Current methods being used, such as, chest x-ray and sputum cytology do indeed shed light to the problem our group is hoping to correct. Both are of the above ways do indeed assist medical professionals in the diagnosis of current or potential risks of cancer, that is, both methods used to this day are used to help diagnosis those who currently have lung cancer of all stages and those who have cancer at a very early stage, early enough to reduce the loss of life considerably. However, even with the potential early stage find of cancer, these methods do not reduce the risk of dying from lung cancer, as our hopes for our project would inevitably lead to, if successful.

Another method currently in use is the computed tomography scan, or CT scan for short. This is the method where our groups focus lies, or rather, the output of this methods, dicoms. Though currently being used doesn't necessarily mean it is being used in the most efficient way. Our project revolves around the improvement of that efficiency in the hopes to solve the above problem.

System Analysis

To better understand our project, let me first reiterate the overall problem our team wishes to help solve. The problem is twofold, the overwhelming debt, due to repetitive medical examinations, that only seems to grow as one hopes to catch cancer at an early stage, and the excessive loss of life that could have been saved if the cancer was discovered early. These problems lie within the results of a current process, and thus, it is the process that is in need of changing. Our teams aim is to help correct this process in order to correct the problem. Currently, these problems have indeed been looked at by the professionals in the medical fields they reside in. Such professionals would be those that are considered to be the fields of pulmonology, radiology, primary care, thoracic surgery, interventional radiology, and medical and radiation oncology.

To help better focus on how the current process is, let us talk about the role of an Oncologist. An Oncologist is an individual who manages one's care and treatment once one has been diagnosed with cancer. The process would start with the patient. An individual, or patient, would come into the facility in order to be checked for lung cancer. For the sake of answering why this individual is getting a screening, let us just say they

are a known smoker for twenty years and have been recommended to get checked. Now, once the screening, CT scan, is done, the Oncologist takes a look at the dicom image. The dicom image is a three dimensional (3D) representation of the patient's lung. It is the Oncologist that reviews these images in order to make a final decision on whether the patient has cancer or not. However, in many cases, even if cancer isn't found, but the history of the patient is known, like being a smoker for twenty years as mentioned earlier, the Oncologist can determine if further screening is necessary.

Let's say our patient has cancer and it was visible on the CT scan. The next step the Oncologist would take is determining the stage of the cancer. The stages of lung cancer are ordered from one to four. One being cancer confined to the lung. Two and three being cancer confined to the lung and possibly the lymph nodes. Finally, stage four is cancer that has spread outside of the lung and into other parts of the body. After the stage of cancer is determined, the Oncologist then determines what treatment is best suited. Let us say our patient has stage one lung cancer. The Oncologist would then operate on the patient and attempt to remove the cancer through the means of a surgical knife.

After the surgery, let us say there was no complication during the surgery and the cancer was completely removed, the patient would undergo at least one additional screen in order for the Oncologist to be satisfied with the results of the surgery. However, the screening process may continue well after the surgery. Sometimes two to three times a year. Those who have had cancer before run a higher risk of getting

cancer again within their lifetime. Especially those who continue old habits that are known to cause cancer. Say, continuing to smoke in our patient's case.

As you can very well see, the process of the Oncologist is purely done by him or her. The only computer interaction truly utilized is the scanner. The input being a patient to scan, and the dicom it delivers, the output, which the Oncologist personally reviews. Computer intervention in this process, though very helpful for the Oncologist, is rather slim in comparison to what the Oncologist themselves have to do.

System Proposal

Our system will be made up of three main parts; a dicom pre-processor, an image processor, and an anomaly detection system. The first of these three, the dicom pre-processor will use the libraries Pydicom and Numpy to process patient scans into standardized tensors, or higher dimensional arrays. The goal of this aspect of the system is to standardize the resolution and size of each image, in order to then convert them into a pixel tensor that can be passed into our next portion of the system.

The image processor will run the pixel tensor through a hybrid recurrent and convolutional neural network, using the libraries TFLearn and TensorFlow. The output of this level will be a high-dimensional vector that describes patterns in the images (through the convolutional neural network) and over multiple scans (through the recurrent neural network).

The final piece of our project will be the anomaly detection system, which will interpret this high-dimensional vector into a single value, the probability of diagnosing

the patient with lung cancer in the next 12 months. This layer will employ libraries including TFLearn, SKLearn, Tpot. All of which are machine learning implementation packages, the latter of which employs genetic algorithms to optimize both algorithm selection and parameter optimization.

Task Analysis

The tasks of the system we propose can be divided into 4 subtasks to accomplish our goal of attaining a correct diagnosis of the dicom images we intend to run through our system. The majority of the tasks involved with completing our project will involve python programming, and various python libraries.

1. Using a downloadable repository of Dicom images publically available through the Data Science Bowl website, feed them into the system to be analyzed by the underlying methodology
2. Upon feeding the Dicom images into our program, we will implement the “PyDicom” python module, to preprocess these 3d scans of various resolutions and dimensions of the patient’s lungs into standardized 3d images to form the basis of the material to train the convolutional neural networks, and in turn the recurrent neural network.
3. Using these converted images, the neural network will perform another conversion, this time using the standardized 3d image from the conversion in step 2, into an array representation of the image. This processed array contains data on each pixel of the processed 3d image.

4. With this array, in the next step of the process can utilize this array against a template formed from a pre trained neural network to check the “normality” of each array.
5. Depending on the performance and efficiency of the algorithm, coupled with the details of our implementation, we can obtain a (hopefully) accurate assessment of the patient’s risk of lung cancer.

Some of the methodology we chose within the course of our project planning phase is owed in large parts to our previous experience with using, implementing and manipulating the python language. Task 1 is a predefined output that was provided by the Data Science Bowl, so for our second task, we had to choose a language/module/some tool to process this data. Task 2 was formulated when it was discovered through both research and advice from the contest itself - PyDicom allowed us to take full advantage of our already acquired knowledge of the language. Becoming familiar with a module of a language is much easier and wastes less time than learning how to implement what we need to be able to in another, less familiar language.

Task 3’s implementation was chosen because of a neural network’s ability to learn about how each 3d image it receives is different, but able to be processed into a usable, more standardized 3d representation all the same. Because we are feeding the output of one neural network into another, it is essential that the two understand the other’s output. For Task 4, again, the only method to “learn” what makes a processed image being indicative of cancer is a neural network. Other algorithms that we have discussed in the classroom, and subsequently researched for potential implementation

viability, do not have the ability to change what they are programed with. For instance, rules-based inference could not be trusted to account for the variability of the sample data. Because of the input, the algorithm must be able to not only apply “rules” to each picture, but account for this variability. Task 5 will most likely be an algorithm, implemented in python that tracks the certainty of the output of the neural networks, knowing the correct diagnosis. Supervised learning will play a large part all throughout our project.

Data Analysis

Thankfully, the data that must be obtained for this project to get off of the ground is provided in a large 63 gb .zip file that can be downloaded from the website of the Data Science Bowl, the origin of our inspiration for the project. The data that this project will operate on is a filetype called .dicom. Dicom files are generated as the output of a computerized tomography (CT) scan, and attempt to describe a 3d object on a 2d plane. It does this by having the “image” itself be an interactive “object” composed of many different images that represent a multitude of “cross-sections” of the object undergoing a CT scan. This interactive object can be manipulated by sliding the cursor on the object itself to see various depths of the object, and in this case, the lungs. Doctors utilize these image objects in their diagnosis of the patient by scanning through the depths of the patient’s respective Dicom to search for abnormalities.

For the purposes of our project, which are mainly educational and explorative in the practices of image processing, classification and comparative analysis, we will

utilize various 500 mb portions of the whole 63 gb data to train our implementation of both convolutional and recurrent neural networks, so as to maintain the tractability of our processing of the dicoms, and the comparison of each outcome under supervised conditions. Below is what a still of the standardized dicom image “object” would look like.

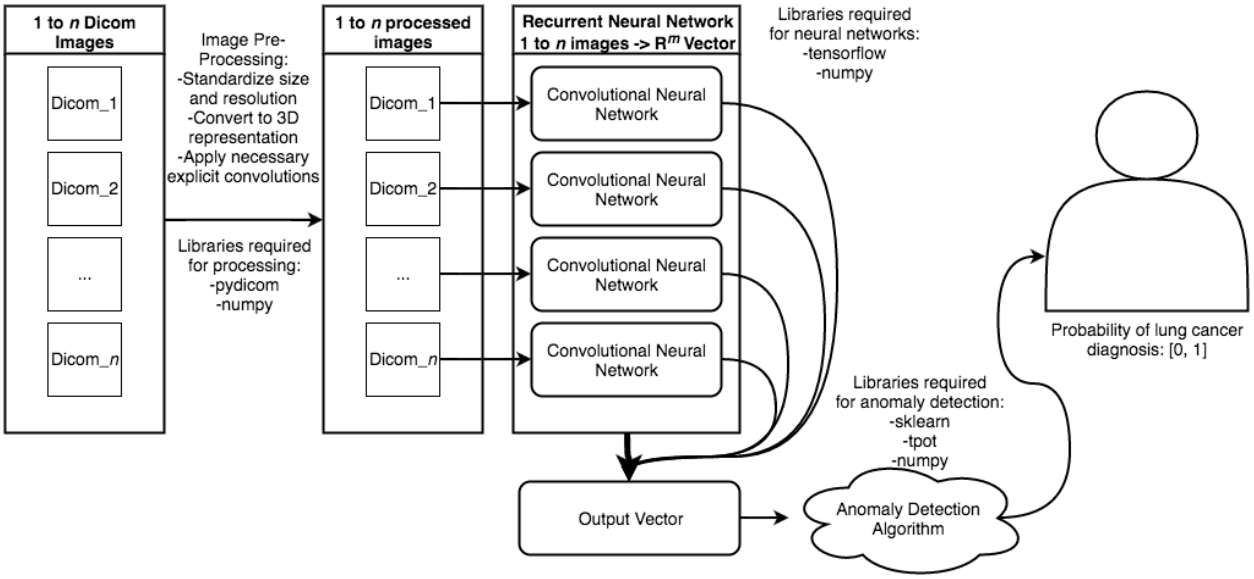
This is a preview of what the dicom image objects might look like after the PyDicom standardization, ready to be fed into the recurrent neural network. The standardization process cuts much of the extra information the dicom includes, as per output of a CT scan, off of the dicom, leaving only the essential impression of the lungs for further processing, to decrease both time and processing power requirements.

System Architecture

The general contours of the classification will follow the diagram in figure 1, proceeding in four phases.

1. It will receive an ordered collection of dicom images, the output format of a CT scan. Sample images are of variable resolution, black-white/color, orientation, and size. Further, dicoms are a specialized format allowing for the 2D representation of a 3D scan. In the pre-processing phase, images will be standardized and converted to a 3D representation.
2. Next, in the computer vision phase, each image will pass through a Convolutional Neural Network (CNN). This may have explicitly programmed convolutions or dynamically generated convolutions depending on experimental performance. This CNN will output a high dimensional vector for input to a Recurrent Neural Network (RNN).
3. This RNN will take 1 to n CNN outputs, looking for evidence of anomaly over time.
4. Depending again on experimental performance, this RNN may output a scalar probability the patient is diagnosed with lung cancer in the next 12 months or may output a high dimensional vector. This output vector then becomes input to a fourth phase classifier, which will be selected through demonstrated experimental performance.

Figure 1: General System Diagram



For development, this project will require extensive use of available deep learning libraries primarily available in Python 3.5.2+ and used widely in industry. Libraries and their respective phases are marked above in Figure 1. Specifically, they will include:

- Pydicom, for dicom image processing
- Tensorflow, for large scale machine learning on heterogeneous distributed systems
- Numpy, a tensorflow dependency and ultra-efficient Python array implementation
- TFLearn, a high-level machine learning library for rapid prototyping
- Tpot, for using genetic algorithms to optimize a machine learning pipeline

GPU-accelerated hardware will be necessary for efficient experimentation, as convolutional neural networks run orders of magnitude slower on CPU-only hardware. Our group has sufficient remote access to such hardware and has begun initial testing for proof of concept, achieving an accuracy of 63% on a very limited CNN implementation after only one hour of training. Working across local and remote development environments will require meticulous use of version control, done through GitHub.

For deployment, delivery conditions will dictate environment requirements. Ideally, the final implementation should run in an environment nearly identical to development. It is likely that deployment conditions may require some form of client-server interface, allowing users to upload dicom images for processing on a more capable server. This may sit well outside the scope of this project, given that real-world deployment would be done through a third-party cloud services provider.

Summary and Conclusions

In summary, our team hopes to deliver an effective and efficient system that will predict lung cancer within a twelve month time frame. The aim of this system is to reduce medical expenses by the means of early detection, which reduces visitations and examinations, and to increase overall life expectancy of those with lung cancer. The aim

of this project will be achieved by the means of goals we as a team developed. The target users of this system are those in the medical profession that deal with lung cancer patients on a day to day basis.

Bibliography

System Architecture

American Cancer Society. "Key Statistics for Lung Cancer." Key Statistics for Lung Cancer. American Cancer Society, 5 Jan. 2017. Web. 09 Feb. 2017.

Abadi, et. al. “TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems.” March 16, 2016. arXiv:1603.04467v2 [cs.DC].

Christodoulidis, et. al. “Multi-source Transfer Learning with Convolutional Neural Networks for Lung Pattern Analysis.” December 8, 2016. arXiv:1612.02589v1 [cs.CV].

Ciompi, et. al. “Towards automatic pulmonary nodule management in lung cancer screening with deep learning.” October 28, 2016. arXiv:1610.09157v1 [cs.CV].

Danjuma. “Performance Evaluation of Machine Learning Algorithms in Post-operative Life Expectancy in the Lung Cancer Patients.” April 17, 2015. arXiv:1504.04646v1 [cs.LG].

Esteva, et. al. 2017. “Dermatologist-level classification of skin cancer with deep neural networks.” Nature 542: 115-118. doi:10.1038/nature21056.

Gao, et. al. “Holistic Interstitial Lung Disease Detection using Deep Convolutional Neural Networks: Multi-label Learning and Unordered Pooling.” January 19, 2017. arXiv:1701.05616v1 [cs.CV].

Rezende, et. al. “Unsupervised Learning of 3D Structure from Images.” July 3, 2016. arXiv:1607.00662v1 [cs.CV].

Shafiee, et. al. “Discovery Radiomics via StochasticNet Sequencers for Cancer Detection.” November 11, 2015. arXiv:1511.03361v1 [cs.CV].

Shin, et. al. “Learning to Read Chest X-Rays: Recurrent Neural Cascade Model for Automated Image Annotation.” March 28, 2016. arXiv:1603.08486v1 [cs.CV].