Identify Lesions within CT Scans Using Convolutional Neural Networks

Graham Chickering

November 12, 2020

Abstract

Convolutional Networks are a specific type of Neural Networks that have shown to be particularly effective at being able to identify distinct objects within images. This technique can be used to identify different types of anomalies, such as lesions, within medical images as well as detect other sorts of tumors or cancers. In theory this type of network is designed based on how the human brain works and the idea that multiple levels of neurons are connected together in order to detect and identify images. In practice though, running and training these types of neural networks can be very computationally expensive and require large amounts of memory and processing capabilities if working with a very large dataset. Especially when working in R which has limited memory capabilities, trying to run and train this type of model can be very slow and ineffective. Thanks to developments in cloud computing and distributed data processing engines such as Google Cloud Storage and Apache Spark, being able to run and train these types of models within R can become much faster and more efficient when trying to analyze large amounts of data. While there is more work to be done, this project shows how to create an infrastructure that efficiently stores data and trains these models when working with the R Studio Environment.

Keywords: Convolutional Neural Networks, Apache Spark, Google Cloud Storage, Lesions, Healthcare

1 Introduction

It has been estimated that roughly 80% of healthcare data is unstructured data, which can come in the form of videos, sensor data, images, or text. Although hospitals and researchers used to have a hard time extracting insights from this type of data, with the recent advances that have been made in data science and handling big data, this has created new application areas within the healthcare industry in sectors such as genomics, drug trials, predicting patient health, and medical imagery. Medical imaging research in particular has made significant progress recently with researchers being able to use different machine learning algorithms to detect different types of lesions and cancers from CT and other types of scans. In particular the advancements of convolutional neural networks to identify different types of lesions and cancers is extremely promising to the medical community and can be used to potentially help doctors identify tumors or lesions that they might have missed, and reduce the amount of hours and amount of expertise required to view CT scans.

When working with medical imagery data sets though one of the first problems someone may run into is how to process and handle these large data sets. When trying to perform analysis on small and medium sized data sets within R, one would rarely run into complications that would be attributed to how R is loading and dealing with the data itself. But what happens when one moves from the world of medium sized data to the world of Big Data and large data sets? While many of us have probably been able to read in and load our data for projects into the R Studio environment without issues and without having to worry about whether the entirety of our data can even be loaded in, one may begin to run into complications the larger the data set becomes. By default, R loads all data into memory and while memory size depends slightly on your computer's configuration settings under any circumstances one cant have more than 2,147,483,647 rows or columns, which is roughly equivalent to 2 GB of memory that R is using. (see `http://www.columbia.edu/~sjm2186/EPIC_R/EPIC_R_BigData.pdf`). If you do end up crossing into the threshold where R can no longer store all the data in an effective way, there are multiple potential solutions in the forms of choosing random subsets of the data, buying a computer with larger memory, or use parallelization and using multiple clusters to

perform the analysis. It is this solution of utilizing distributed computing, via the sparklyr package, that will allow us to perform computationally expensive analysis on large data sets.

On top of the issue of trying to run analysis on large data sets in R itself, is the issue of how to best store and load the original information and data. Often data sets are small enough that they can be stored on your local computer in a folder that is then uploaded into the R Studio Environment itself, but what should one do as the size of the data set substantially increases and one no longer wants to store large data sets directly on their machine. One solution to this problem is to take advantage of a cloud computing service and store the data directly in the cloud, freeing up space and memory on your personal computer. By storing the data on a cloud computing service, this can become especially useful when a project begins to get scaled up whether that is through adding new members to work on the project or when more and more data gets added to the project. In this project I will take advantage of Google Cloud Storage to store my data, which will look and store my data on the cloud without having to take up any space on my local machine.

By combining Google Cloud Storage with Apache Spark this will allow me to create a large data set consisting of medical images. This data set will then be used to train and create a convolutional neural network that will try to identify the lesions themselves within the images.

This project will allow me to answer the questions of what is the best way to store large datasets and perform computationally expensive analysis on those data sets? How does one handle and process images so that analysis can be performed on them, and how effective are convolutional neural networks at identifying anomalies within images?