IE 598: Group Project Proposal

October 2nd, 2017

Group Members:

Alexander Marunycz

Siddharth Bhaduri

Veeraj Gadda

Tanmay Rajput

**Group Project Choice: TSA/Kaggle Passenger Screening Algorithm Challenge**



The Problem

The Transportation Security Administration (TSA) is currently experiencing high false alarm rates during airport checkpoint screenings for passengers, which initiates processes that slows down the entire passenger checkpoint process for all involved. With the number of passengers traveling by air and the number of threats posed to them both increasing every year, the TSA has a demand for improved prediction models to more accurately assess possible threats entering their screening devices.

At the moment, TSA purchases updated algorithms only from the manufacturers of their scanning equipment, but these algorithms are proprietary, expensive, and releases in very long development cycles. So, the TSA is recruiting help from the data science community via a Kaggle challenge worth a total of $1.5 million worth of prize money to improve the accuracy of the threat detection algorithms at work within their scanning devices.

The TSA is providing the Kaggle challenge with a dataset of images collected on their latest generation of scanners, where the contestants of this challenge will attempt to identify the presence of simulated threats under a variety of object types, clothing types, and body types. The TSA has explained that even a small improvement in the accuracy of the threat detection algorithm can go a long way in making a significant impact on the security of their screening system as well as the improving overall passenger experiences at the airport.

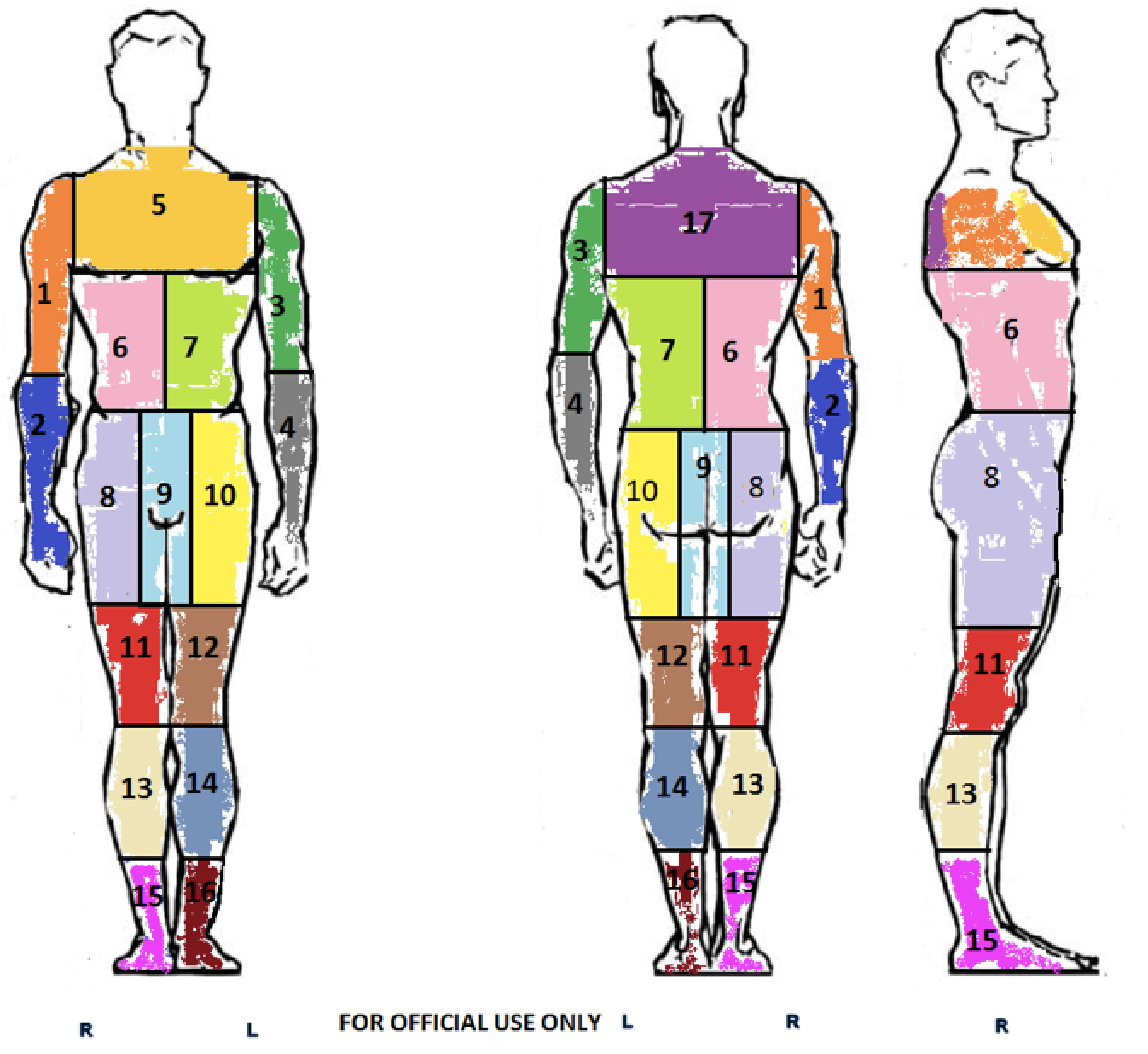
Our Strategy

Overall, we wish to build upon the knowledge we have already gained in this course to utilize convolutional neural networks to train a prediction algorithm to identify the threats within the training and test datasets. This particular Kaggle challenge is substantial and often poses more questions than solutions, but thankfully there is a robust community within this challenge that provides “Kernels” or helpful ideas and pieces of code that we can leverage in order to attack this problem as efficiently as possible.

Examples of these helpful Kernels are tutorials put together by the community on how to download the data, analyze the threat zones, and extract useful information from the training set of images while reducing the 512x660x16 image data that we feed into the machine learning model.

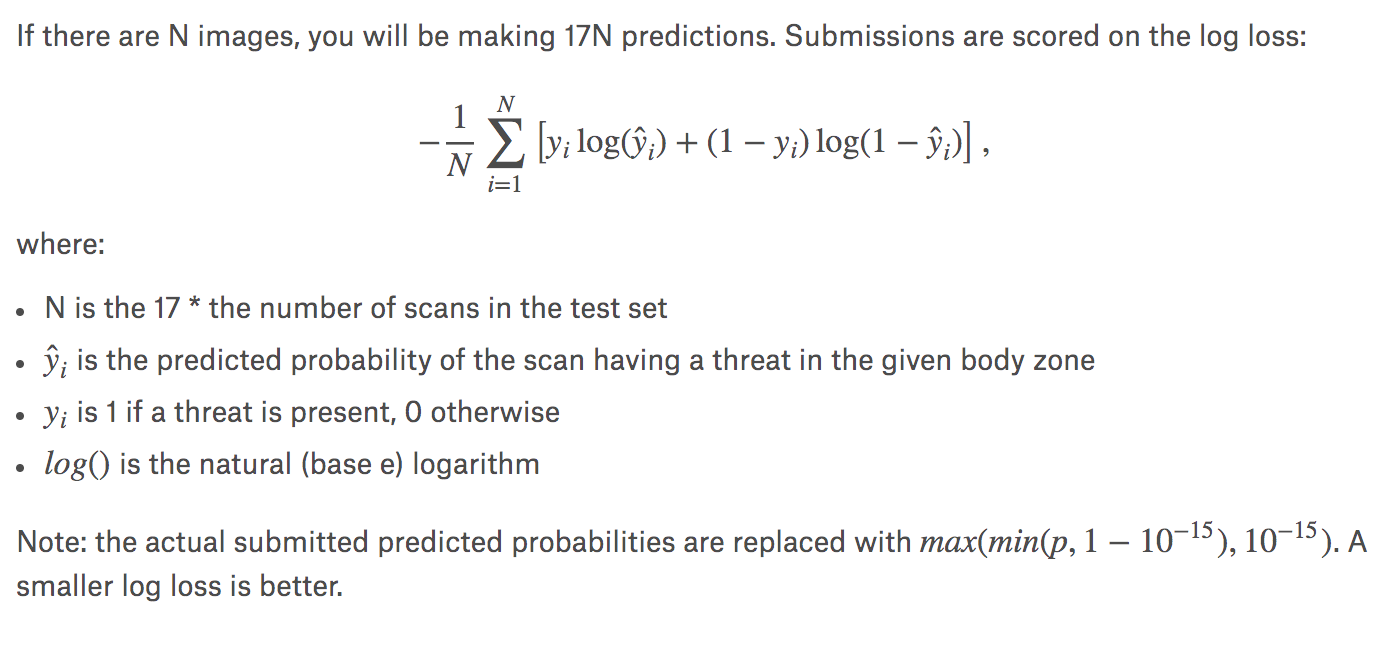
Speaking of the training data that we have to work with, it is a set of just under 1200 examples of High Definition-Advanced Imaging Technology (HD-AIT) system files that range in size from 10MB to over 2GB per subject. Each one of these examples have scans done at 16 different angles, meaning 16 different images per subject. One of the first problems we must solve with our algorithm lies within the fact that the threats to be detected within these series of 16 images per subject can exist in any one of 17 “threat zones.” See the below Figure 1 for an illustration of the 17 different threat zones.

**Figure 1: 17 Threat Zones**



So we will have to first train our algorithm to detect these 17 separate threat zones within the series of 16 scans per subject before we can isolate the issue of whether or not there is a threat in any specific one of them. In the end, for every subject we will have to use the 16 scans to define 17 threat zones per body and then return a probability per threat zone as to whether it contains an actual threat. The log loss of our 17 predictions per subject is the evaluation method for this challenge.

**Figure 2: Evaluation Criteria**



There are already a number of issues that we are taking into consideration when confronting the prospect of building this convolutional neural network prediction algorithm.

One issue we are recognizing is that 1200 samples is not enough to properly build a robust machine learning model. So, to tackle this issue we are considering utilizing a network that is pre-trained on a large amount of various images so as to leverage its recognition capabilities via a process called “transfer learning.” One such network we could leverage for this is ImageNet. Utilizing ImageNet and its repository of over 14 million images can help us define the 17 threat zones on a body as well as better define non-threats within these zones.

Another issue we face is whether or not we want to train 17 individual prediction models, one per threat zone, or if we want to train one prediction model that isolates all of the threat zones and analyzes all of them at once.

Lastly, there is the big picture issue. We see that many people already have very small log loss amounts and will surely win this competition. But we are concerned that they are actually over-fitting their models on the responses they receive from the test data…which could result in models that are very accurate for the test data, but are not as best prepared for the universe of threats that exist outside of the test data. Therefore, we will be taking this into account with our model where we are trying to set up our prediction algorithm where it is not over-fit on the test data and will generalize as well as possible for the growing array of different threats that the world can throw at it. We are interested in building the best model for real world data, not necessarily over-training on the test data for a competition we won’t win.