Machine Learning Engineer Nanodegree

# Capstone proposal

## May 25/2020 - Uljan Sinani

## Domain background

Nowadays the latest research on Object Recognition has achieved decent advancements regarding Machine Learning. Training an Artificial Neural Network (ANN) to recognize objects with a high accuracy seems very trivial with the current tools and resources. If we look back in time and ask the question what made it possible? We could agree that the main advancement comes from having access to more computational resources (i.e. more computational power and larger amount of data). Hence, the latter is what I find more interesting to analyze and find possibilities for optimization. In the following parts of this proposal I would argue that focusing in the quality of the dataset and optimizing for reduction of biases during pre-processing of data is very important before than tuning the ANN model. The argument will be based on a typical Computer Vision problem for object detection.

Therefore, in the following parts of this project proposal I suggest another approach to optimizing the quality of datasets used on Computer Vision applications. It is worth noting that the dataset optimization I would propose are tested with small-scale project containing a small to medium dataset of images.

## Problem statement

The need for additional computational resources becomes increasingly important the more input data we need to process. What I find more interesting is the fact that we do not know how the data we process will contribute to the learning rate of our model. The issue of “complex datasets”[[1]](#footnote-1) associated with higher requirements of computational resources made me rethink “the big-data”. The potential solution I propose to optimize a data sets before using it for an ANN model is as follows:

* Test for bias
* Conversion of data (i.e. from one data format to another)
* Distribution of classes/categories

Testing for bias, will ensure the data I have is equally distributed among objects that my dataset has. Encapsulation of raw data to a specific data format, will allow me to understand the information that could have been lost from the data conversion. Study of the classes/categories will ensure I have the correct distribution of input data for my model. Therefore, ensuring mainly the above steps could allow for a more efficient learning rate of my ANN model. Also, as a consequence the need to test/train my model with more data decreases saving more computational resources.

## Datasets and inputs

To test the model, and demonstrate the concept (comparing the original dataset with the optimized one) I downloaded from Kaggle the [fruit\_images](https://www.kaggle.com/mbkinaci/fruit-images-for-object-detection) dataset. It served as my input dataset which contained images of different fruits (i.e. apple, bananas, orange and mixed). Following the above steps to optimize the dataset I made sure that it had an equal number of objects (apples, oranges, bananas and mixed fruits), and correct data format for my input layer.

## Solution statement

The solution proposes a new approach to optimize the dataset before using it to train the model. I propose to read the input data from a dataset, check for correct distribution of classes to avoid bias and chose the right data format for ANN mode. I apply these steps during data exploration stage before moving to testing and training stage. This practice will allow me to feed my model with quality data rather than simply feed it with an uncleaned high volume of data translated to a higher demand for resources.

## Benchmark model

To benchmark the proposed solution, I compare the model output accuracy by training it in one scenario using original dataset and then using the optimized dataset. Then the same procedure is applied using different frameworks (i.e. XGBoost – TensorFlow – PyTorch - SKlearn) to observer the change in prediction accuracy of my ANN model.

## Evaluation Metrics

Prediction accuracy per resources used are the metrics I use to evaluate the efficiency of the proposed solution. I expect that the model trained with the original dataset to yield an output accuracy of less than the case of it being trained with the optimized dataset. Thus, if I note the accuracy of the same model trained with the first dataset as A\_raw and A\_opt for the optimized dataset the expectation would be A\_opt > A\_raw, accordingly for the computational resources used is expected R\_raw > R\_opt.

## Project design

The project design will start by analyzing the format of the dataset, conversion of data from raw input and then counting the correct number of classes and number of total objects the dataset contains. After data exploration, performance tests based on accuracy will take place. It will compare scenarios with the same type of data containing imbalance of objects, different encapsulation and data formats. After applying the above recombination of data into different datasets (for training and testing), then, benchmarking of computational efficiency will take place.

## Presentaion

[1] S. Zhi, Y. Liu, X. Li, and Y. Guo, LightNet: A Lightweight 3D Convolutional Neural Network forReal-Time3DObjectRecognition

[2]Nibs K.Logothetis and David L.Sheinberg, Visual Object Recognition.

[3]Gevers Arnold W.M.Smeulders, Color-basedobjectrecognition.

[4] <https://www.altexsoft.com/blog/datascience/preparing-your-dataset-for-machine-learning-8-basic-techniques-that-make-your-data-better/>

1. Dataset comprised of different formats, and size. [↑](#footnote-ref-1)