COMP 4449: Summer 2024

Professor: Luke Sawle

*Student Name:* Evan Hollier

***Term Project Plan***

**Direction**: Address the questions in the following sections to create a project blueprint (or project plan for your midterm (This document is also applicable to the final project).

**Project Goals**

* What business goals are you trying to solve?

Traffic sign recognition, which is useful for self-driving cars.

* How will you measure the success of achieving the business goal?

I will be satisfied if I can significantly raise the performance of the image recognition model over a pre-trained image recognition model.

* What analytic or machine learning tasks will you be solving?

Performing a grid search or equivalent method for finding the best pre-trained model, which will be used as a baseline. Then, tuning hyperparameters with that baseline model to get the best improvement I can within a reasonable time frame.

* How will you measure success for your analytic or machine learning solution? Hint: start by defining your baseline.

The potential improvement on performance depends on how well the baseline model performs. For example, if I start off with a 90% accuracy, I would consider getting up to a 95% accuacy a success. But if it's only 70% accurate, I would hope for a larger jump like to 85%.

**Significance of the Project**

* What is the expected impact of your project?
* Provide a statement that captures why your project is important with respect to the business goal(s).

Success in the project would leave me with a model that can recognize traffic signs. This, among other things, is neccessary for self-driving technology. I could also see it being used for providing additional information to digital road maps. For example, if there is a School Zone sign on a road in Google Maps, routes through that road can display a caution popup during pertinent hours.

**Methodology**

* What analysis method would you use to solve the problem?
* Statistical method? If you are using a statistical method, is it descriptive or inferential?

I might measure F1 score as a performance metric. This is a descriptive method.

* Machine learning algorithm? Is it supervised or unsupervised machine learning?

I will be using a pre-trained model with additional layers on top, which will require tuning. This will be supervised learning since the dataset is labeled.

* How does the selected method work to achieve the analysis goal?

Adapting a pre-trained model to achieve a model that can recognize image data of traffic signs is a form of transfer learning. I will be leveraging the knowledge gained from solving larger image recognition problems and applying it to my own image recognition task. The data will be split into train-validation-test. Every model will be trained on the training dataset, then measured on the validation set. Many models will be created with different values of hyperparameters. The one that performs the best on the validation dataset will then be measured on the test dataset, and that performance will be considered when determining success on the goal.

* What are the pros and cons of the method(s) selected compared to other methods for achieving the same task?

The other method I can think of would be building a CNN from scratch. Making a model from scratch is not feasible when it comes to image recognition. The amount of data, training time, and computation power is simply too much. I'm pretty sure the pre-trained models availabe online were created with supercomputers.

**Data Description**

* What does the data represent? Is it education data, health data, sensor data, etc.? Provide a general description, hosting site, and URL (if available).

<https://github.com/emmanueliarussi/DataScienceCapstone/tree/master/3_MidtermProjects/ProjectRTS>

<https://www.kaggle.com/datasets/imadmoussa/traffic-signs>

Image data of traffic signs captured under different real-life conditions and showing obstructions, poor lighting, or even the sign being far away from the camera.

* Is the data structured or unstructured?

Since the data is image files, it is unstructured.

* What are the dimensions of the dataset?

data points in:

train - 34,799

validation - 4,410

test - 12,630

total - 51,839

Images are 32x32 pixels.

* What data types are represented by features in the dataset.

Image data and their label (String).

There are also 'coords' and 'sizes' values, though I am not yet sure what they represent.

* Which variables will be used for analysis? If there are too many variables, try to use broader descriptions such as demographic variables instead of gender, ethnicity, age, etc.

Image data and their label.

**Data Preparation**

* What data preprocessing steps do you expect will be required?

The data is stored as a pickle file. I will probably have to convert it into a tensor.

* How will you check and handle outliers and/or missingness?

If a data is missing I'll probably just not use it. Assuming all the data is actually traffic signs, I will keep outliers. This will create a model that is useful for real world data.

* What feature extraction or selection strategy would you use for dimensionality reduction, if any?

I don't see myself needing to reduce the size of the data. My previous image recognition project used 7,349 255x255 images, and the training time was not an issue. If I need to though, I will randomly trim the number of data points.

**Data Exploration**

* What descriptive statistics will you generate for the different features represented in the data? Provide a rationale.

F1 score will be the metric for scoring model performance. In the training set, the count of labels in each category ranges from 180 to 2010, which is quite a large disparity. I would therefore consider this dataset unbalanced, which makes accuracy a misleading metric.

If I have time, I will also look into confusing classes. This will allow error analysis.

* What types of data visualizations will be helpful in analyzing the dataset? Please provide the names of the graphs or charts and the appropriate variables for the graphs.

A graph for the distribution of label counts, probably in the form of a bar graph. This will provide a visualization for how imbalanced the dataset is.

Graphs of training and validation loss as the models train. The x axis will be epochs.

Graphs of training and validation F1 score as the models train. The x axis will be epochs.

Also, I will produce samples of different traffic sign and its predicted label.

If I do error analysis, I will make a table for Potential Improvement calculated for each class (Potential Improvement = Error % \* Population %)

**Solution Development**

* What data product do you aim to achieve by the end of the analysis?

A model that can accurately label traffic signs based on an image.

* Please outline the sequence of steps you expect to follow for selecting, building, and testing the algorithm(s) you intend to use to develop your solution. Try to be as detailed as possible.

I will be working from a GitHub repository. First, I'll do some data exploration which is where I'll create any visualizations for the overall dataset (train-validation-test will all be combined). I'll be looking for missing values and anything that otherwise can't be used. If the pickle files are too much of a hassle to work with, I'll consolidate the data into a pandas dataframe to perform analysis on. Once I'm ready to move on to the modelling phase, I'll convert the data into tensors. I will perform a grid search or equivalent method on several pre-trained models with a simple softmax activation layer added on. Whichever performs the best (using F1 score) will become my baseline model. I will incrementally modify the final layer(s) of the baseline model (things like optmizer, number of neurons, dropout, etc) until I see a performance increase I'm happy with. This will be done on a small number of epochs with early stopping. Then I'll try training for as long as I can without overfitting. My final model will be tested on the test dataset, and any performance visualizations will also be created. Finally, I'll present my work in a powerpoint.

* What computational resources will be required to ensure the feasibility of your modeling process.

I have a 3070 that's already set up for machine learning.

* How will you use peer reviewed journal articles and other similar resources to guide your work?

The only thing I can think of is referencing them for defining terms in my presentation.

* How will you ensure that your results are reproducible?

Everything will be available on my GitHub as jupyter notebooks. The data source is publicly available from Kaggle.

**Conclusion and Recommendations**

* Who are the stakeholders or business users who will benefit from the solution you develop?

The automotive industry, specifcally self-driving cars.

Navigation apps (like Google Maps)

* What information will they want to know to assess the value of your outcomes in terms of their priorities?

Reliabilty - confusion matrix. Any other relvant performance metrics like ROC-AUC.