# **Machine Learning Engineer Nanodegree**

## **Capstone Project**

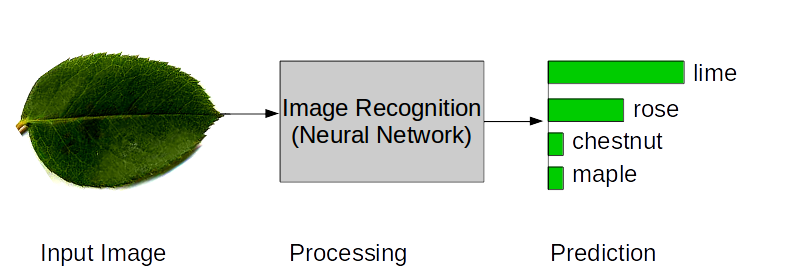
Alexander Langer

5th February 2017

## **I. Definition**

### **Project Overview**

The objective of this exercise to identify the species of a plant by processing the image of its leaf through a machine learning algorithm that will output a prediciton about the type of leaf it has been given.

  
Figure 1 - High Level Process Design

The necessity to adress this subject is given through global trends in the field of agriculture. As the University of California, Davis summarizes[[1]](#endnote-2), agricultural efficiences must increase to meet population growth projections of 3-5 Billion until 2050. A key lever will be to lower the units of inputs per unit of output. Foremost, the volume of fertilizers and pesticides must be dramatically reduced to maintain soil and plant vitality.   
  
Image recognition through machine learning can be a key contributor to this challenge since it enables resource conscious processes in weed elimination[[2]](#footnote-2) or plant desease detection[[3]](#footnote-3)

Due to the variety of plants, approximately half a million globally, classification of infected crops or weed is prone to errors. If agricultural output is to increase while pesticide usage is to decrease, reliable image classification by machine learning can be a key contributor.

This paper outlines his classification challenge by deploying a deep neural network against images of leaves that belong to one of 99 species. The underling dataset has been retrieved from the UCI Machine Learning Repository and Kaggle[[4]](#footnote-4)

### **Problem Statement**

Classifying images of leaves to identify their species is a supervised machine learning and classification task. The shape of the leaf will provide distinctive patterns that will become more evident through various preprocessing steps and can then be leveraged through the machine learning algorithm.

To make the images machine readable, preprocessing such as color transformation to binary colors, resizing and augmentation will take place. The image data will ultimately be transformed into multi-dimensional arrays that can be fed into the algorithm.

Since all image data has been labeled with a plant species type, this aspect must be preserved to train our program and to test its accuracy after training on a dedicated and isolated test dataset.

When referring to the word algorithm, in particular a deep neural network with dedicated weights and biases for each plant species type will be deployed. It will have several layers that abstract the input image data further to learn high-level patterns, so called gradients. Since we feed more than thousand images into our network, this will be done in small batches while each iteration will adjust the weights and biases in the network to train it against particular labels – in our case leaf species.

While weights and biases are beeing trained and set so that an input image can be identified as belonging to a particular species, each training iteration will produce a prediction towards the anticipated kind of leaf species. This prediction will be represented with floating point numbers between zero and 1 where each possible species label will obtain a number varying by the degree of certainty the algorithm has about its prediction. This is called the softmax function. Afterwards the prediction with the highest likelyhood (argmax function) will be referecend against the true prediction and the weights and biases will be adjusted accordingly. This process is called the loss function or optimizer and it is where the actual machine learning takes place iteration by iteration.

The model should gain some robustness and the capability to reproduce prediction when fed new leaf images. I hope to obtain prediction probabilities above 90% while measuring accuracy will be done by comparing predicted labels (e.g. top 3 predictions) vs. Actual labels.

### **Metrics**

As outlined earlier, a accuracy function will be deployed as the metric to judge model prediction performance.

Since the biases element in our neural network architecture will be trained to learn the transformed array and pixel representation of each leaf, it will enable the network to judge a new image against its fit to one of the pretrained labels and leaf types.

As described, this process will produce statistical probabilities through a softmax function that in the overall sum equal one and the trained bias with the highest probability will be our prediction for the most probable leaf type. Comparing this prediction against our ground truth data (actual labels) will yield a match or a wrongful guess. While the argmax function will pick the prediction with the highest probability, we will also be able to learn the percentage of certainty our network had over its prediction. I would label this the actual accuracy of the prediction.

## **II. Analysis**

(approx. 2-4 pages)

### **Data Exploration**

In this section, you will be expected to analyze the data you are using for the problem. This data can either be in the form of a dataset (or datasets), input data (or input files), or even an environment. The type of data should be thoroughly described and, if possible, have basic statistics and information presented (such as discussion of input features or defining characteristics about the input or environment). Any abnormalities or interesting qualities about the data that may need to be addressed have been identified (such as features that need to be transformed or the possibility of outliers). Questions to ask yourself when writing this section:

* If a dataset is present for this problem, have you thoroughly discussed certain features about the dataset? Has a data sample been provided to the reader?
* If a dataset is present for this problem, are statistics about the dataset calculated and reported? Have any relevant results from this calculation been discussed?
* If a dataset is **not** present for this problem, has discussion been made about the input space or input data for your problem?
* Are there any abnormalities or characteristics about the input space or dataset that need to be addressed? (categorical variables, missing values, outliers, etc.)

### **Exploratory Visualization**

In this section, you will need to provide some form of visualization that summarizes or extracts a relevant characteristic or feature about the data. The visualization should adequately support the data being used. Discuss why this visualization was chosen and how it is relevant. Questions to ask yourself when writing this section:

* Have you visualized a relevant characteristic or feature about the dataset or input data?
* Is the visualization thoroughly analyzed and discussed?
* If a plot is provided, are the axes, title, and datum clearly defined?

### **Algorithms and Techniques**

In this section, you will need to discuss the algorithms and techniques you intend to use for solving the problem. You should justify the use of each one based on the characteristics of the problem and the problem domain. Questions to ask yourself when writing this section:

* Are the algorithms you will use, including any default variables/parameters in the project clearly defined?
* Are the techniques to be used thoroughly discussed and justified?
* Is it made clear how the input data or datasets will be handled by the algorithms and techniques chosen?

### **Benchmark**

In this section, you will need to provide a clearly defined benchmark result or threshold for comparing across performances obtained by your solution. The reasoning behind the benchmark (in the case where it is not an established result) should be discussed. Questions to ask yourself when writing this section:

* Has some result or value been provided that acts as a benchmark for measuring performance?
* Is it clear how this result or value was obtained (whether by data or by hypothesis)?

## **III. Methodology**

(approx. 3-5 pages)

### **Data Preprocessing**

In this section, all of your preprocessing steps will need to be clearly documented, if any were necessary. From the previous section, any of the abnormalities or characteristics that you identified about the dataset will be addressed and corrected here. Questions to ask yourself when writing this section:

* If the algorithms chosen require preprocessing steps like feature selection or feature transformations, have they been properly documented?
* Based on the **Data Exploration** section, if there were abnormalities or characteristics that needed to be addressed, have they been properly corrected?
* If no preprocessing is needed, has it been made clear why?

### **Implementation**

In this section, the process for which metrics, algorithms, and techniques that you implemented for the given data will need to be clearly documented. It should be abundantly clear how the implementation was carried out, and discussion should be made regarding any complications that occurred during this process. Questions to ask yourself when writing this section:

* Is it made clear how the algorithms and techniques were implemented with the given datasets or input data?
* Were there any complications with the original metrics or techniques that required changing prior to acquiring a solution?
* Was there any part of the coding process (e.g., writing complicated functions) that should be documented?

### **Refinement**

In this section, you will need to discuss the process of improvement you made upon the algorithms and techniques you used in your implementation. For example, adjusting parameters for certain models to acquire improved solutions would fall under the refinement category. Your initial and final solutions should be reported, as well as any significant intermediate results as necessary. Questions to ask yourself when writing this section:

* Has an initial solution been found and clearly reported?
* Is the process of improvement clearly documented, such as what techniques were used?
* Are intermediate and final solutions clearly reported as the process is improved?

## **IV. Results**

(approx. 2-3 pages)

### **Model Evaluation and Validation**

In this section, the final model and any supporting qualities should be evaluated in detail. It should be clear how the final model was derived and why this model was chosen. In addition, some type of analysis should be used to validate the robustness of this model and its solution, such as manipulating the input data or environment to see how the model’s solution is affected (this is called sensitivity analysis). Questions to ask yourself when writing this section:

* Is the final model reasonable and aligning with solution expectations? Are the final parameters of the model appropriate?
* Has the final model been tested with various inputs to evaluate whether the model generalizes well to unseen data?
* Is the model robust enough for the problem? Do small perturbations (changes) in training data or the input space greatly affect the results?
* Can results found from the model be trusted?

### **Justification**

In this section, your model’s final solution and its results should be compared to the benchmark you established earlier in the project using some type of statistical analysis. You should also justify whether these results and the solution are significant enough to have solved the problem posed in the project. Questions to ask yourself when writing this section:

* Are the final results found stronger than the benchmark result reported earlier?
* Have you thoroughly analyzed and discussed the final solution?
* Is the final solution significant enough to have solved the problem?

## **V. Conclusion**

(approx. 1-2 pages)

### **Free-Form Visualization**

In this section, you will need to provide some form of visualization that emphasizes an important quality about the project. It is much more free-form, but should reasonably support a significant result or characteristic about the problem that you want to discuss. Questions to ask yourself when writing this section:

* Have you visualized a relevant or important quality about the problem, dataset, input data, or results?
* Is the visualization thoroughly analyzed and discussed?
* If a plot is provided, are the axes, title, and datum clearly defined?

### **Reflection**

In this section, you will summarize the entire end-to-end problem solution and discuss one or two particular aspects of the project you found interesting or difficult. You are expected to reflect on the project as a whole to show that you have a firm understanding of the entire process employed in your work. Questions to ask yourself when writing this section:

* Have you thoroughly summarized the entire process you used for this project?
* Were there any interesting aspects of the project?
* Were there any difficult aspects of the project?
* Does the final model and solution fit your expectations for the problem, and should it be used in a general setting to solve these types of problems?

### **Improvement**

In this section, you will need to provide discussion as to how one aspect of the implementation you designed could be improved. As an example, consider ways your implementation can be made more general, and what would need to be modified. You do not need to make this improvement, but the potential solutions resulting from these changes are considered and compared/contrasted to your current solution. Questions to ask yourself when writing this section:

* Are there further improvements that could be made on the algorithms or techniques you used in this project?
* Were there algorithms or techniques you researched that you did not know how to implement, but would consider using if you knew how?
* If you used your final solution as the new benchmark, do you think an even better solution exists?

**Before submitting, ask yourself. . .**

* Does the project report you’ve written follow a well-organized structure similar to that of the project template?
* Is each section (particularly **Analysis** and **Methodology**) written in a clear, concise and specific fashion? Are there any ambiguous terms or phrases that need clarification?
* Would the intended audience of your project be able to understand your analysis, methods, and results?
* Have you properly proof-read your project report to assure there are minimal grammatical and spelling mistakes?
* Are all the resources used for this project correctly cited and referenced?
* Is the code that implements your solution easily readable and properly commented?
* Does the code execute without error and produce results similar to those reported?

1. <http://asi.ucdavis.edu/programs/sarep/about/what-is-sustainable-agriculture> [↑](#endnote-ref-2)
2. <http://www.bluerivert.com/> [↑](#footnote-ref-2)
3. [Http://peat.technology/](http://peat.technology/) [↑](#footnote-ref-3)
4. <https://www.kaggle.com/c/leaf-classification/> [↑](#footnote-ref-4)