Machine Learning Engineer Nanodegree

Capstone Project

Ivan Jennings August 28th, 2021

I. Definition

Project Overview

This project aims to detect dog breeds in a provided image using a convolutional neural network. There are many different image classifiers out there, including for animal breeds. Dog breeds present a challenge to determine specific breed, given that humans have intervened in the breeding resulitng in various different breeds and sub-breeds. Additionally, some different species of dog can look completly different. *"This domain is especially challenging since the appearance of corresponding parts can vary dramatically, e.g., the faces of bulldogs and beagles are very different."* *1

Problem Statement

In this project I aimed to create an accurate classifier which will determine first if the photo contains a human and/or dog, and then determine the breed of the dog or similar dog breed if a face is detected. Given the large data set and resources that would be needed to upload to the cloud and lack of these required resources, I won't be deploying this classifier as originally intented.

Metrics

I will calculate accuracy of the detectors of human faces and dogs based on the total correctly detected images divided by the total predictions which should give us a result close to 100%. I will also calculate the accuracy of both the CNN models based on total correctly classified divided by the total classified. In addition to this I will use the loss calculation built into the scikit-learn library 4 to compare against the kaggle models.

II. Analysis

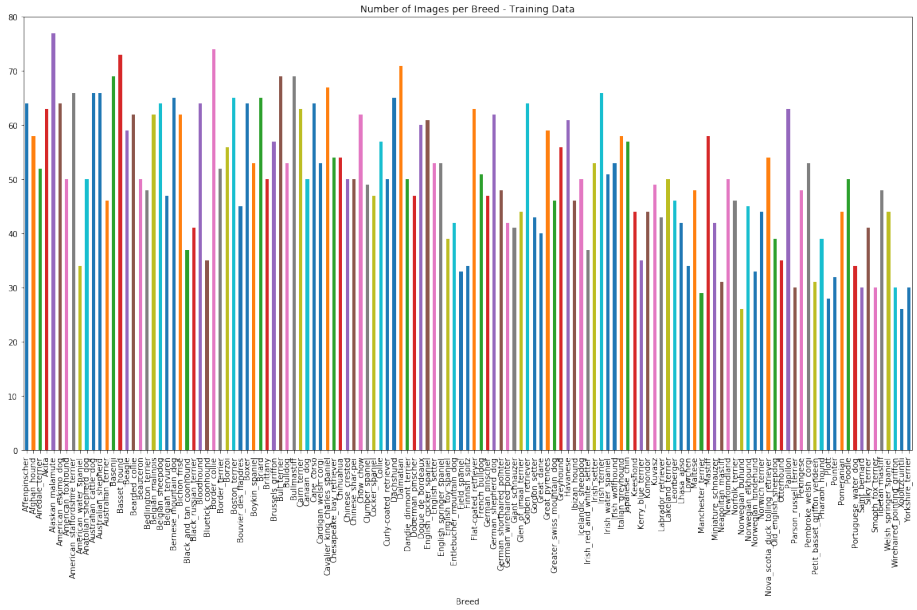
Data Exploration

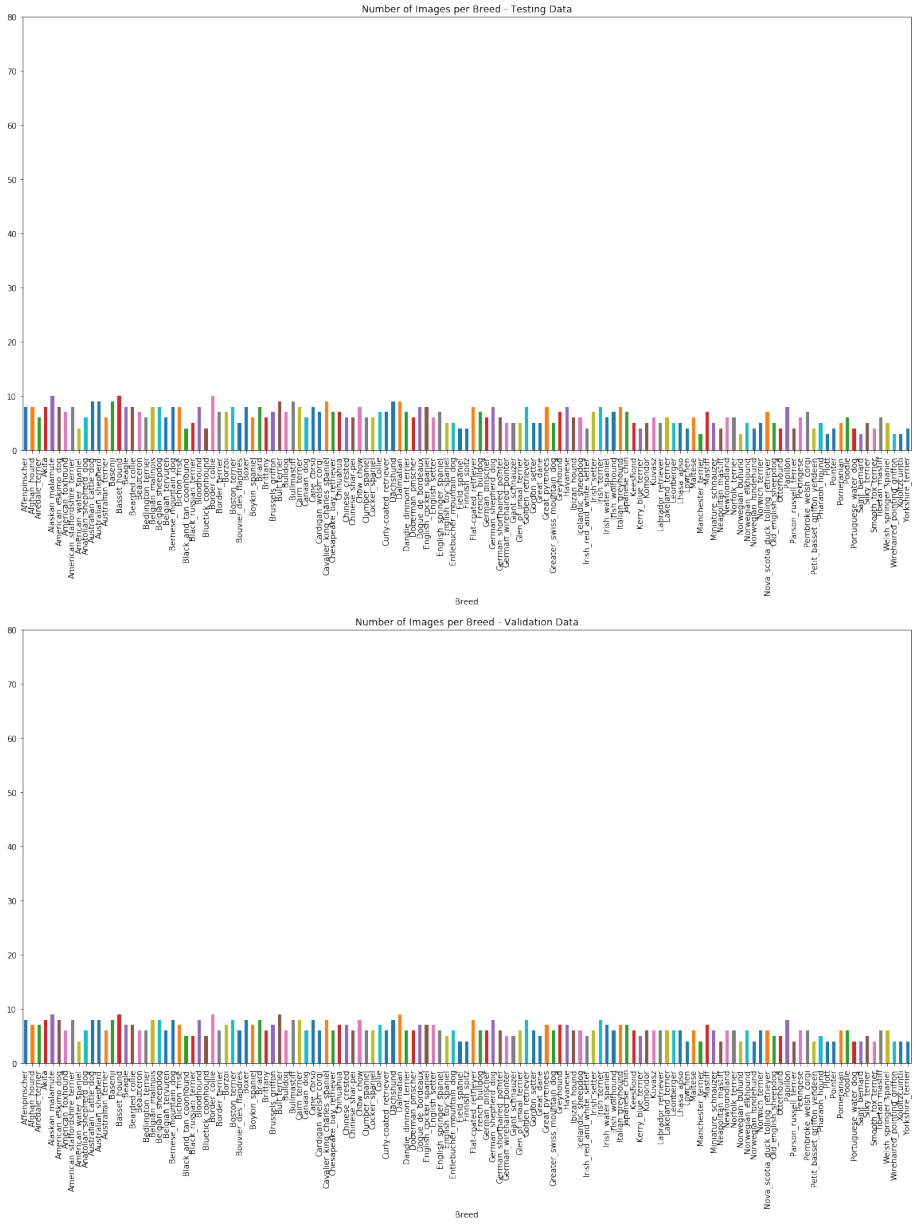
The data set that will be used for this project has been provided by Udacity and are available in the referenced links 2. The dog data set containing 8351 colour images which will be used for classifying breeds, contains 3 folders each for training, test and validation sets. Each set contains a folder of photos for each of the 133 dog breeds. The sets are not balanced between dog breeds - test sets have been 3-10 images for each breed, training sets have between 26-77 images per breed and the validation set has between 4-9 images per breed. The images also have varying resolutions in JPG format. We also have a human

face data set with 13233 colour images which will be used to train a model for determining if a human face is present in a photo - this set has a range of people with about 1 image per person, with up to 40+ images per person.

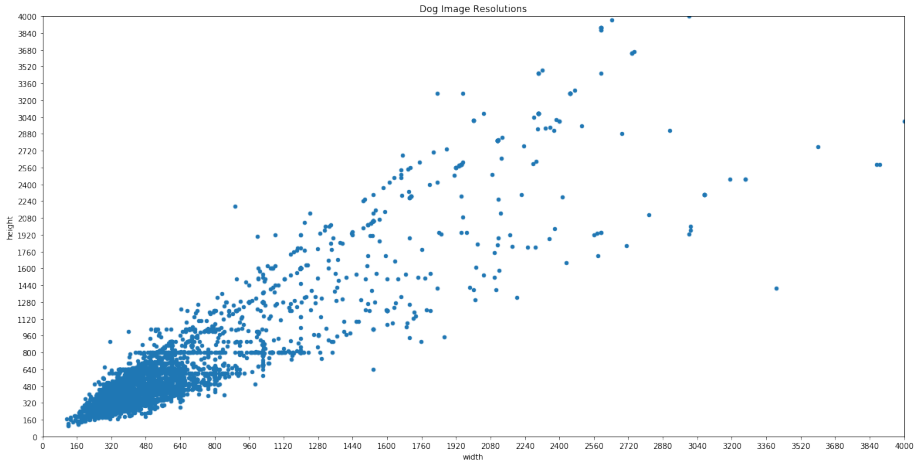
Exploratory Visualization

Because there are 133 different dog breeds, it will be useful to visualise how many images we have per dog breed and data set (e.g. training, testing or validation) so that we can see if there is any particular breed that doesn't have enough images, or see if the distribution is even.





We can see that there is a reasonably even distribution between each set. The following plot also shows the distribution of resolutions within the set, which will be helpful to see what are the most common resolutions.



We can see that the majorit of the images are between 200 x 200 and 800 x 600 resolution, which will be helfpul to know when selecting image transformations.

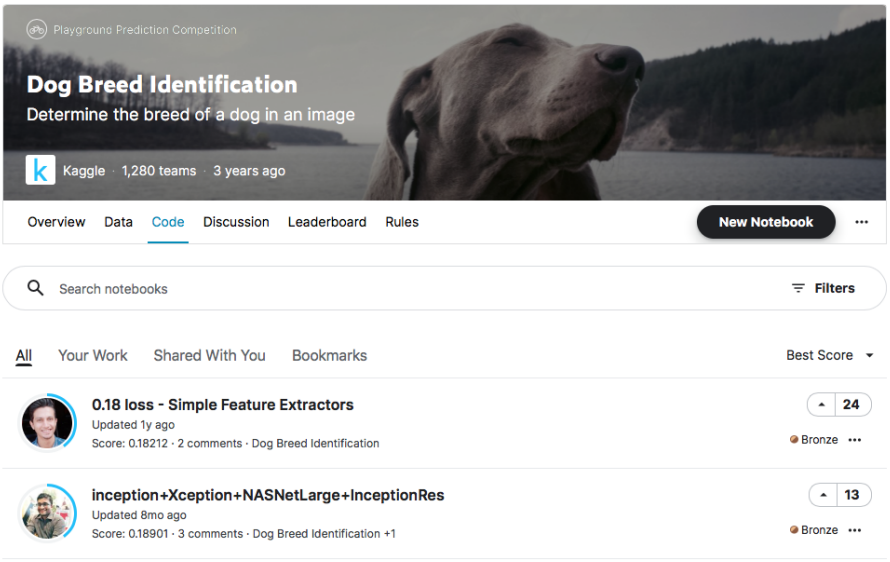
Algorithms and Techniques

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

The benchmark for the human and dog detector should be quite accurate - close to 100% correctly detected, however the differences between certain breeds can be quite subtle and thus more difficult to differentiate. The benchmark we will use for the classifier of breeds will be at least 10% for the initial CNN model written from scratch, and at least 60% accuracy for the CNN model using transfer learning.

I will also compare with similar models that can be found on Kaggle 3. The best model had a Multi Class Log Loss of 0.18 which we can use to compare our model.



III. Methodology

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?

References

1
J. Liu, A. Kanazawa, D. Jacobs, and P. Belhumeur, "Dog Breed Classification Using Part Localization", Computer Vision–ECCV 2012. Springer Berlin Heidelberg, 2012.

2
Dog dataset - Udacity
Human dataset - Udacity

3
Dog breed identification

4
Scikit-learn loss function