Are Petfinder's Breed Classifications Visually Subjective?

Analyzing the Results of a Purebred-Trained Dog Image Classification Model on a Mixed-Breed Petfinder Dataset

Allison Aprile, May 2019

Abstract

Animals are entering shelters at increasing rates, causing overcrowding and a need for pet adoption. Several companies have eased the adoption process for consumers by using technology to advertise available pets. Petfinder.com is one of the original online adoption websites, with a database of over 255,000 listings. Petfinder has a notably large selection of mixed-breed dogs, many of whom are posted with hypothesized breed classifications. With so many pets coming to shelters with unknown backgrounds, visual classification is often the only feasible way to determine breed.

This project intended to test the hypothesis that Petfinder's breed classification is visually subjective: based on the appearance of the animal. This was to be completed by implementing a Convolutional Neural Network (denoted CNN) image classification model, trained on a set of purebred dog images and tested on a set of mixed-breed dog images obtained from Petfinder.

Due to unexpected issues, lack of resources, and limited time, the project was ultimately not able to be completed. However, the effort still yielded an interesting dataset, valuable knowledge, and motivation to proceed.

Introduction

An estimated 6.5 million companion animals enter shelters each year, with half being returned to their homes or adopted. Therefore, the search for the perfect pet can be daunting. However, with the help of technology, it is possible to view any type of pet anywhere across the United States. Petfinder is one of the most popular adoption websites, its database hosting over 255,000 animals from 11,000 shelters. Their interface allows users to first filter animals by shelter radius, then by desired type, breed, age, size, gender, and more.

Among the list of available filters, type, size, and gender are indisputably accurate. However, the other filters, specifically breed, are not always certain. The breed classification of shelter animals is difficult as the animals are admitted with ambiguity; even if the animal had a previous owner, they do not always provide accurate, if any, information. Therefore, it is easy to misclassify an animal's breed, especially if the mix is complex or uncommon.

An Inspirational Case Study

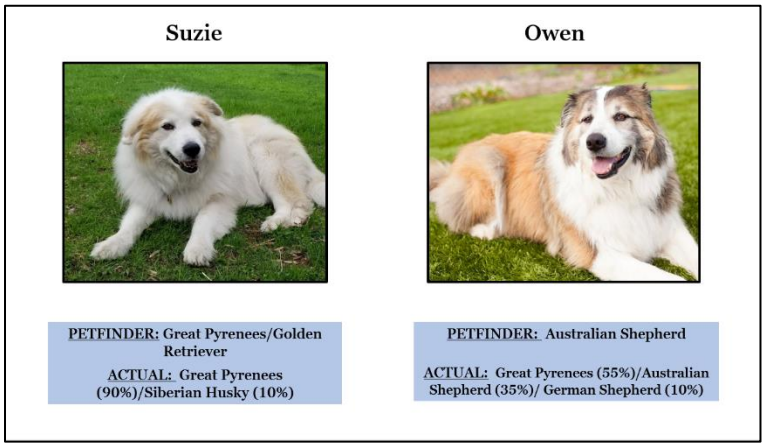


Figure 1. Comparison of Petfinder and Actual Breed Classifications for shelter dogs Suzie (left) and Owen (right)

The dogs in **Figure 1** were both adopted through Petfinder and have been living happily with me for over eight years. My family was so intrigued by the breed classifications that we requested DNA tests. When comparing the Petfinder versus DNA results, it is clear that the classification is somewhat based on appearance.

Suzie, for example, was misclassified as part Golden Retriever due to some of her gold spots (which were more prominent when she was a puppy). However, her fluffy fur and deep brown eyes correctly indicated her as a Great Pyrenees. Owen, on the other hand, was advertised as a purebred Australian Shepherd, although we prematurely identified his Great Pyrenees lineage due to the similar fur structure described with Suzie. His distinct coloring, which emulates that of an Australian Shepherd, is clearly responsible for his misclassification.

Hypothesis: Breed classification on Petfinder is subjective: based on the appearance and/or behavior of the animal, i.e., the opinion of the shelter.

A similar classification analysis was conducted by Sai Kumar Arava, discussed in his paper "Mixed Breed Dogs Classification" (see **References**). Arava implemented the same methodology as will be discussed in the following section: training a CNN model on a purebred dataset and testing it on a mixed-breed dataset. He found that the accuracy was rather low on the testing data unless the dataset was augmented with variants of the original images. Interesting, Arava was able to pinpoint certain dog breeds that had a significantly higher accuracy, attributing this to the distinctity of features of such breeds. Arava's findings, particularly his accuracies, were helpful in the creation of the Petfinder dataset.

Methodology

To sufficiently test the hypothesis, it was necessary to implement a model that would mimic visual classification. Convolutional Neural Networks (CNN) are a branch of machine learning notable for their image classification abilities, and thus, the proper tool for testing visual subjectivity. By training the model on a purebred dog image dataset, it would be capable of labeling a mixed-breed dog image with whichever pure breed it identifies most distinctly. This model and several purebred datasets already exist in volume, however, there is no mixed-breed or Petfinder dog image dataset available. Therefore, this was the primary focus of the project.

Original Model Selection: The main CNN model attempt followed the code outlined in "How to easily build a dog breed image classification model" by James Le (see **References**). To ensure that this would be a functioning model, I downloaded the Stanford Dog Dataset and implemented the first half of the code. As it seemingly worked, I proceeded to build and annotate the Petfinder dataset in order to return to the testing portion of this model.

Data: ***S**ee **Accompanying Deliverables Key** for files

Training: The Stanford Dog Dataset was the choice for the training data. It contains 20,580 dog images spanning 120 different breeds.

Organization: The Stanford Dogs Dataset contains a CSV file of 20,580 unique dog identifications and an adjacent column of the breed classification. A separate folder labelled 'images' contains the images named with the dog's identification. As this was compatible with the initial model, I chose to format the Petfinder dataset accordingly.

Testing: Because no mixed-breed/Petfinder dataset could be found online, I devoted majority of the project to creating a detailed, unbiased dataset for this purpose. The overall dataset consists of 1,898 dog listings with over 30 attributes, including unique identifications and photo links. The refined testing dataset consists of a random sample of 150 Petfinder dogs, annotated in a similar fashion to the Stanford Dogs Dataset.

Collection: This was the timeliest portion of the process that ultimately prevented me from completing the model and hypothesis testing. It took many instances of trial and error to discover a working method.

Attempt 1: PetPy Wrapper (10 hours)

Since Petfinder supports an API, accessing the data was not an issue. It is a RESTful API, so the most direct way of access would be through cURL commands. Unfortunately, in the beginning of the process I had no experience of this. Therefore, I did some research and found a Python Wrapper, called PetPy, that seemed easy to implement. I created a Petfinder Developer's Account, received an API key and access token, and attempted some calls. Unfortunately, I consistently faced JSONDecodeErrors, which I initially attributed to an error with the JSON installation in my iPython environment. I exhausted over three hours attempting to reconfigure the code, performing several JSON trial runs with the Twitter API and raw JSON files, as well as comparing every individual line of the PetPy source code files to those I procured through pip install.

There were no discrepancies, concluding that the problem did not lie with JSON and leading me to probe issues with the Petfinder database communication. After discussing my procedure by phone with a Petfinder technology representative, Joe Cabanilla (see **Acknowledgements**), we concluded that the issue lied in the authentication and authorization of the developer count. He guided me through several tests and the Petfinder API authentication procedure, which introduced me to cURL commands. I was able to successfully retrieve data with that method, but I was insistent on using the Python Wrapper method based on a promising Jupyter Notebook code I had uncovered to collect a large volume of images in a short period of time.

Therefore, I completed more research and found that Petfinder had recently released a second version of the API that required the authorization process not required in the first version and, thus, for PetPy. I was confident that I could revise the PetPy source code to process the access token, yet my attempts were all fruitless. The access token is time-sensitive, overall hindering my ability to request large quantities of data through PetPy as intended.

Attempt 2: Webscraping (4 hours)

After fighting with PetPy, I decided to try webscraping with BeautifulSoup. I followed several online tutorials as well as Lab 4. Although I was able to scrape some data (including pet name and location), I had difficulty extracting the primary reed of each animal due to the HTML structure of the listings webpage. I abandoned this method as it was not yielding the correct data with desired efficiency.

Attempt 3: Petfinder API (cURL) (7 hours)

As I had no prior knowledge of working with cURL, I enlisted the help of my peer Tommy Tan (see **Acknowledgements**). With the combined guidance of Joe, Tommy, and the Petfinder API Documentation, I was able to extract 1,600 dog listings in JSON format using the call:

```
curl -H "Authorization: Bearer eyJ0eXA..."
https://api.petfinder.com/v2/animals\?type\=dog\&limit=100 -> json_test(call#).json
```

The command restricts the request to listings of "dogs". Additionally, to speed up the process, I requested the maximum limit (100 elements). Tommy helped me with the call suffix that allowed me to each output to a JSON file, which I labelled per the call number (1-16). I collected 1,600 elements to ensure a diverse selection of dog breeds as well as to account for any missing information or broken image links. The cURL collection took approximately 20 minutes and yielded excellent results. The JSON output contained multiple image links per listing in addition to a unique identification number, primary breed, and other not as relevant information.

I then converted all of the JSON files to CSV using an online program and dedicated approximately 4 hours to carefully merging the files, deleting empty lines, and reformatting overflowing cells. I remained with about 1,500 elements. I then spent another hour inspecting the data, which caused me to realize that I had overlooked a significant flaw in my data collection: repetition. I believed I could salvage 200 elements for the dataset, but after converting the CSV to a DataFrame and calling drop_duplicates(), remained with only 202 elements, 62% of which classified as Labrador Retrievers. It was necessary to repeat the collection.

Attempt 4: Petfinder API (cURL) with Random Page Specification (2 hours)

Because the Petfinder API efficiently returned clean, favorable data, I decided to use it once more for the raw collection. This time, I made sure to do adequate reading about cURL commands and RESTful APIs. My research confirmed that I had wrongly assumed the dependency of cURL requests. It was not sensible to believe that each request would be unique if I did not further specify unique parameters. Therefore, I revisited the Petfinder API documentation and ran trials with the API based on different Query parameters. After about an hour of testing, I determined that I would maintain a limit of 100 returns but also include a page specification alongside such in each call. I decided to randomly select 20 pages in the range of 1 to 26,442 (the total number of pages) to decrease the probability of repetition. I implemented the following call:

```
curl -H "Authorization: Bearer eyJ0eXA… "
https://api.petfinder.com/v2/animals\?type\=dog\&limit\=100\&page\=(page#) > dogs(call#).json
```

I once more converted the resulting JSON files to CSV, merged them, and created a DataFrame. Before cleaning, I ran the drop_duplicates() command and was rejoiced to retrieve 1,992 unique elements. This established the basis for my master dataset.

Cleaning: After removing the duplicates, I performed some EDA (see **Results**) and spent approximately two hours deleting empty and overflowing lines, remaining with 1,898 elements. This completed by Petfinder master dataset. I then read this file back to a DataFrame and narrowed it down to four columns: ID, Primary Breed, Mixed Breed (Boolean), and Medium Photos (URL) for annotation.

Annotation Preparation: To ensure that every dog would have a matching breed classification, I compared the list of Petfinder breeds against the Stanford dog list (see **Appendix A1**). I spent approximately four hours looking up each breed and checking its images. Because some of the breeds were generalized or followed an alternate name in the Petfinder labels (for example, 'cardigan' versus 'corgi'), it was necessary to rename some of the labels. Once this process was complete, I returned to the DataFrame, revising and eliminating certain classifications.

Next, I used the mixed.sample() call to retrieve 150 random elements with Mixed Breed = True, calling it approximately ten times until I received the most unbiased sample. This was dependent on the size of the two most frequent primary breeds, Labrador Retriever and American Staffordshire Terrier. I then read this DataFrame to a CSV and began the annotation process.

Annotation: The annotation took approximately three hours to complete. Based on how I read the columns into the file, it was quick to label the dog breeds with the ID. The image collection and naming were more tedious, involving visiting each image link, ensuring proper size. Luckily, only five of the 150 samples had broken image links, so I revisited the master file, collected five more elements with un-accessed identifications, and imbedded them into the testing sample. This concluded the data collection process.

Model:

By the time the data collection was completed, I only remained with the weekend to work on the model. Being that many models exist online, I wrongly assumed that this would not be too much of a challenge. The initial model selection worked smoothly until the actual training began. Due to the computational limits of my laptop and the lab computers, this particular section of code consistently crashed the program and prevented me from being able to test my mixed breed dataset.

I dedicated approximately fifteen hours to testing other models, all submitted alongside this report. Most of them yielded accuracy below 5%, which would not be sufficient for testing my hypothesis. One model that implemented Transfer Learning with a ResNet50 model worked quickly and yielded 83% accuracy. Since it was a model pretrained on a different dataset, I was required to build a function that could read in my own image files. For this, I recruited David Zheng (see **Acknowledgements**), a computer science major and developer. After hours of debugging, we discovered that this would be an impossible feat because some of the ResNet50 model features depended on a depreciated version of Keras (2.2.0), which we were unable

to call. I returned to the original model one last time to see if it could be salvaged, but after more unsuccessful hours, I ultimately had to accept defeat.

Results

Due to unexpected issues, lack of computational resources, and limited time, the testing portion of the project was ultimately not able to be completed. Therefore, there are little results that can be discussed because of the lack of hypothesis testing. However, a lot can be discussed about the datasets themselves.

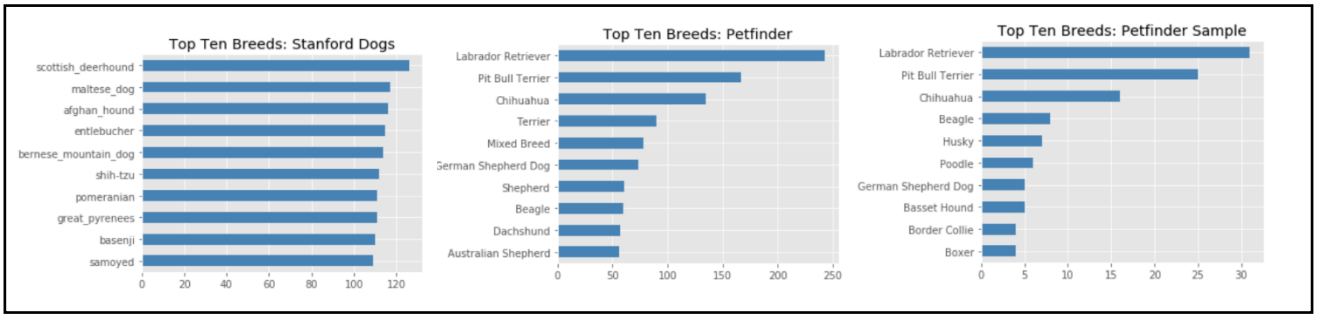


Figure 2. Top Ten Breed Comparison of Stanford Dogs, Petfinder, and Petfinder Sample Datasets

As seen in **Figure 2**, the breed variation between the three respective datasets may have caused issues in the accuracy of breed prediction. The Stanford Dogs dataset's top ten breeds are not inclusive of those in the Petfinder datasets. Therefore, since the model would have been most acquainted with other breeds, there would be a large margin of error for visual misclassification. Further, the top ten breeds in the Petfinder sample indicate trends in the adoption crisis.

Because Labrador Retrievers have many desirable physical and behavioral traits, they are most at stake for overbreeding and mixed-breeding experimentation, therefore putting them at the greatest probability of displacement. In general, mixed breeds constitute approximately 75% of Petfinder's adoption listings because of the desire to create the next 'designer dog', such as the Labradoodle, as seen in Figures **3** and **4**.

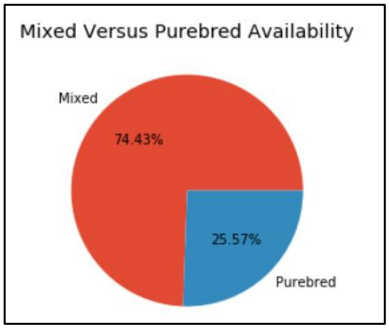


Figure 3. Mixed Versus Purebred Adoption Availability

Figure 4. Zeus (ID #44414858), a Labrador Retriever and Poodle mix

Conclusion

With such a large volume of mixed-breed dogs in shelters, visual analysis is the only reasonable method of primary breed classification. Therefore, it can be hypothesized that **Breed classification on Petfinder is subjective: based on the appearance and/or behavior of the animal, i.e., the opinion of the shelter.**

Unfortunately, the hypothesis testing could not be completed due to time restriction, lack of computational resources, and unforeseen obstacles. I exhausted over 45 hours on this project and am disappointed that I was not able to complete the intended goals. However, the experience was valuable as it familiarized me with cURL commands, Excel macros, and enhanced my Neural Network knowledge. It also taught me that failure is acceptable – it is one of the best ways to learn.

Although the project period is over, my interest in the hypothesis testing results has not waned. I plan to spend the summer continuing on this project in hopes that I can eventually publish my results and gain more technical insights. As a Data Science major, I have thoroughly enjoyed this process and look forward to future experience with Machine Learning.

References

Background

- Petfinder: https://www.petfinder.com/about/
- Arava, Sai Kumar, Yuan, Yuan, and Yang, Wenqing. Mixed Breed Dogs Classification. https://www.academia.edu/33721767/Mixed_Breed_Dogs_Classification

Data Collection

- Stanford Dogs Dataset: https://www.kaggle.com/jessicali9530/stanford-dogs-dataset
- Petfinder API Documentation: https://www.petfinder.com/developers/api-docs

Models

- "How to easily build a dog breed image classification model" (James Lee): https://medium.com/nanonets/how-to-easily-build-a-dog-breed-image-classification-model-2fd214419cde
- "Dog Breed Classifier" (Geunyoung Gil): https://gygil.github.io/visual%20recognition/Dog-Breed-Classifier/
- "Dog Breed Classifier" (James Requa): https://github.com/jamesrequa/Dog-Breed-Classifier/blob/master/dog_classifier.ipynb
- "Dog Breed Identifiation Trial" (Selima M. Rouni): https://github.com/selimamrouni/dog-breed-identification-trial/blob/master/dog_breed.ipynb

Acknowledgements

Joe Cabanilla – *Petfinder Technology Operations Specialist*

I spoke with Mr. Cabanilla on the phone for over an hour. He was very patient and thorough with his explanations, allowing me to successfully access the API and begin the cURL data collection process. I reached out to him over email several other times over the course of the data collection, to which he responded with helpful information and advice.

Tommy Tan – *Mathematics and Computer Science Student, Peer (Principles of Data Science)*

Tommy was fundamental to my cURL data collection process. He spared me hours of research and frustration by guiding me through the cURL commands, specifically those to export the returns to JSON files. He has been an amazing friend to me throughout both the project and this semester, and I am entirely grateful for his help.

David Zheng – *Computer Science Student, Peer*

David spent over ten hours with me attempting to create and debug five lines of code. Although we were not able to fix them, we found the root of the problems and discussed alternative solutions given I would have additional time for the project. During the process, my computer crashed, and he assisted me in retrieving all of my files and saving them to GitHub, another platform with which I was not so familiar before the project. David and I share the same perseverance and love for dogs, making him an essential aspect to the process.

Suzie and Owen Aprile – *Pets adopted from Petfinder*

Suzie and Owen are my two mixed-breed dogs who inspired my project. I am not sure that I would have been able to hypothesize this idea without them. Although they are currently two hours away from me at home, they provided for tremendous stress relief during my obstacles and when I wanted to quit.

Accompanying Deliverables Key

iPython Notebook Files

- **Data Processing –** Cleaning and organizing the Petfinder master dataset and creating random sample
- **Petfinder Classification Model –** Original model attempt (following "How to easily build a dog breed image classification model"), crashed at final line
- **Petfinder EDA –** Petfinder Exploratory Data Analysis
- **Dog Image Classifier Attempts –** Other model attempts (following the remaining model links)
- **Petfinder Data Collection** – Attempts with PetPy

Excel Files

- **labels –** Stanford Dogs labels
- **Petfinder –** Petfinder master dataset (BULK OF THE PROJECT)
- **petfinder_labels –** Petfinder dog labels (from sample)
- **petfinder_sample –** Petfinder sample (for test dataset)

Images

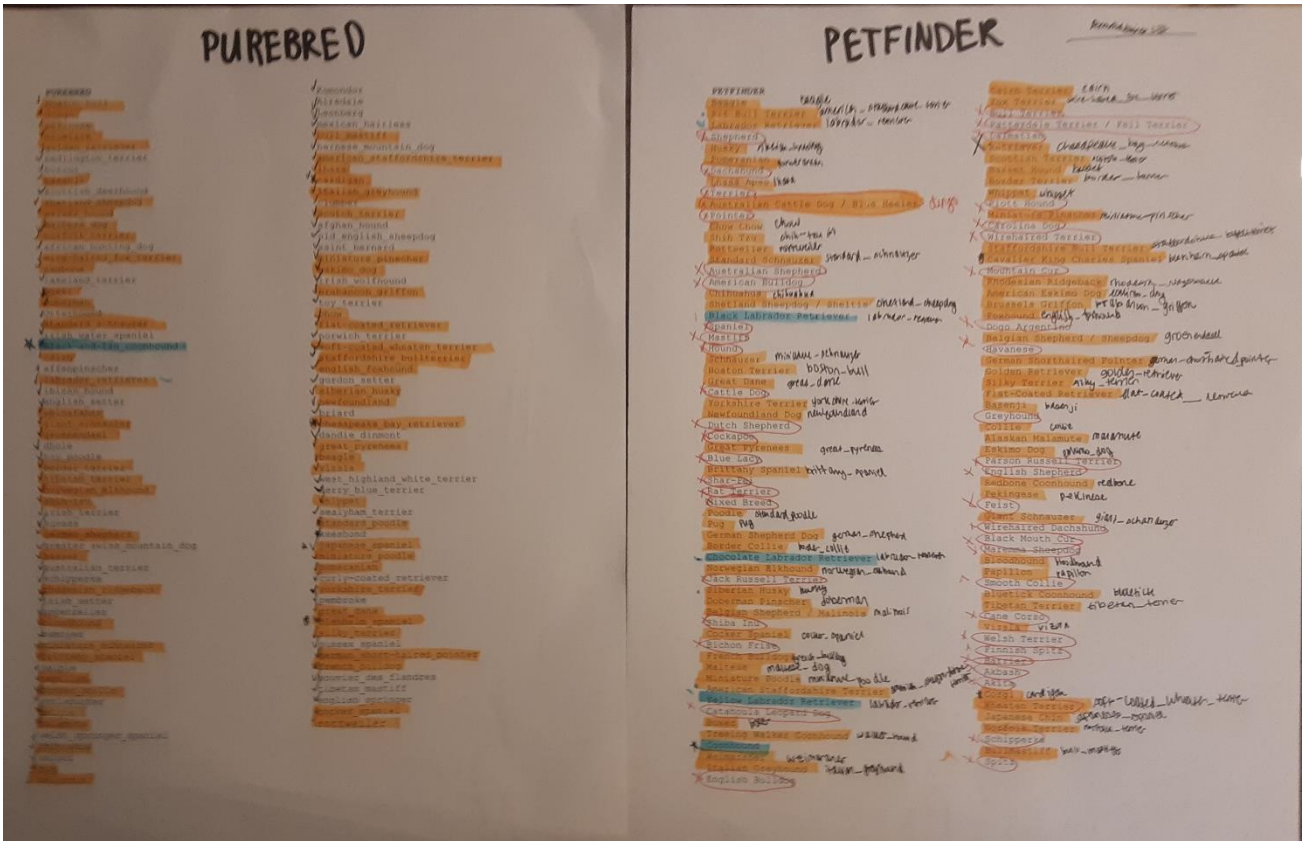
- **Stanford Images – D**ownloaded training dataset
- **Petfinder Images –** Manually created Petfinder image training dataset

Raw JSON Data – 23 JSON files from Petfinder API cURL calls

Stanford Data – Stanford data

Appendix

A1: The breed label comparison process consisted of printing out respective lists of the breeds, and then highlighting shared breeds, marking unavailable breeds, and annotating the proper Purebred label to the corresponding Petfinder Breed.



A2. Project timeline

TIME PERIOD	ACHIEVEMENTS
April 13-14, 2019	Submit proposal and receive approval; experiment with Petfinder API
April 15-22, 2019	Progress Report I; collect and clean data
April 23-29, 2019	Progress Report II; create PowerPoint and present progress; revise data collection
April 30-May 7, 2019	Data cleaning and annotation; model attempts; summarize findings in Final Report
May 8, 2019	Individual project discussion