Pet Breed Identification Image Analysis Case Study

A Case Study by Ben Willoughby, Fall 2024

Of the 6.3 million animals that enter shelters nationwide each year, only about 67% are correctly identified.[1]

Scenario: Have you ever thought about rescuing a pet? Maybe you've spent hours scrolling through a local adoption website, falling in love with a furry companion only to see the word "unknown" listed as their background. That single word leaves so many questions unanswered—questions that can make it harder for these animals to find their forever homes. Knowing a pet's breed is more than just a detail, it's key to understanding their behavior, temperament, exercise needs, and potential health risks. Unfortunately, many animals in adoption centers come from unknown origins, leaving both staff and adopters to guess at their breed and needs. This complicates the adoption process and makes it challenging to match rescue animals with potential owners.

Deliverable: The purpose of this case study is to utilize a convolutional neural network, ResNet-50, in order to identify the breed of rescue dogs and cats. You'll have access to the Oxford-IIIT Pet dataset as well as some preliminary exploratory data analysis and preprocessing. You will pick parameters for the image identification model and train it by constructing and analyzing validation and loss graphs which will prevent overfitting. Finally, you will test the model and gauge overall performance through visualizations like a confusion matrix and accuracy scores.

[1] https://www.aspca.org/helping-people-pets/shelter-intake-and-surrender/pet-statistics

Rubric - Pet Breed Identification Image Analysis Case Study

A Case Study by Ben Willoughby, Fall 2024

Submission format: Upload link to GitHub repository on UVA canvas

Individual Assignment

Why am I doing this?: The objective of this assignment is to strengthen your technical, visualization, and communication skills in data science. By the end of the project, you will gain a deeper understanding of working with image data and training machine learning models. This experience aims to encourage you to explore further applications of data science while equipping you with valuable skills that will help you stand out in the professional world.

What am I going to do?: All necessary materials, including the dataset, code, and reference articles, are available in the provided GitHub repository. Your task is to enhance the existing Python code within the Jupyter Notebook to generate training and validation loss curves while training and testing the ResNet-50 image classification model. Given the computational demands of processing image data, you may need to leverage Rivanna, UVA's High-Performance Computing System, to reduce runtime. Once your results are finalized, upload any visualizations or outputs back into the GitHub repository.

Tips for success:

- Take a moment to familiarize yourself with the dataset. Review the exploratory data analysis resources and don't be afraid to conduct some of your own. Understand the possible limitations of this source.
- Read through the Jupyter notebook and try to understand the code and its structure. This will help you decide which direction you want to move in.
- Have fun! This should be an interesting project at the end of the day so don't let it stress you out. Feel free to experiment and learn from your mistakes.

How will I know I have succeeded?: You will meet expectations on this assignment when you have followed the criteria in the rubric below.

Spec Category	Spec Details
Formatting	● One GitHub Repository (submitter via link on Canvas) ● To ensure **reproducibility**, the repository will adapt parts of the Tier Protocol 4.0. In a nutshell, the top level page of the repository should contain: ○ A README.md file (which auto displays) ○ A LICENSE.md file (use MIT as default) ○ A CODE folder ○ An OUTPUT folder
README.md	● Goal: This file is a guide for anyone who comes to your repository. ● Use markdown headers to divide content. ● Make an H2 (##) section explaining the contents of your repository ● Section 1: Title and description ○ For this part, come up with a creative title as the header name and briefly describe what you did underneath. ● Section 2: A map of your documentation ○ In this section, you should provide an outline or tree illustrating the hierarchy of folders and subfolders contained in your Project Folder, listing the files stored in each folder or subfolder
LICENSE.md	● Goal: This file explains to a visitor the terms under which they may use and cite your repository. ● Select an appropriate license from the GitHub options list on repository creation. ● Usually, the MIT license is appropriate.
CODE folder	● Goal: This folder contains your Jupyter notebook and any other source code. ● Make sure to write descriptive comments for the significant sections of your code. ● Keep code clean and well-organized so that others (and yourself) can easily read and understand your code
OUTPUT folder	● Goal: This folder contains all of the output generated by your project, e.g. figures, tables, etc. ● Any table and/or figures produced by your code should be included here. ● Use descriptive name for your files

Acknowledgements: Special thanks to Professor Alonzi and Layla Ranjbar for facilitating such a fun and exciting case study!