Ridley Arp

Engineering Write-Up

Metis Bootcamp

May 2022

**Abstract**

This intent of this project is to create pipeline steps to ingest, sort, distribute, separate, clean, and model a large, labeled dataset of dog images, then deploy this model in a web-interface that allows for user upload and input and returns the likely breed mix of their dog.

**Design**

This app originates from a personal exposure to a cancer case in my own dog. Earlier warning to potential breed-related risk factors may have helped to improve prognosis, but genetic breed testing was at significant cost. This app would enable veterinary professionals or pet owners to upload personal photos of their pet and get preliminary information on breed mix. As such a probability mix is output to allow for multiple possible breeds’ traits to be considered in the result(mix-breeds).

**Data**

The data was acquired from two distinct but related databases. The Tsinghua Dogs dataset (~70,000 images across 126 breeds) and the Stanford Dogs Dataset(~30,000 images across 120 breeds). The Tsinghua images were compiled from actual pet-owners and candid situations, whereas the Stanford Dogs images were collected from Imagenet.

**Algorithms**

*Data Ingestion*

1. Downloaded and stored images into SQL database with each breed contained in its own table.
2. Created functions to extract images to independent train/test directories with a .75/.25 split.
3. Connected to Google Drive via Colab and uploaded train/test folders

*Models/Processing*

Implemented a generator from the linked train and test directories to preprocess and batch information for modeling

Utilized Keras and tensorflow to fit a ML multi-class classification model with imagenet transfer learning on a VGG16 model.

**Tools**

Keras, Tensorflow, and SQLite3 for data manipulation

Keras for modeling

Flask and Heroku for model deployment