Homework 10

Project C8 - Pet Popularity

Masud Rana, Erik Kõiv

Repository

Task 2. Business understanding

Project values and backgrounds:

Our project goal is to provide shelter to the street animals for their better living. This project is initiated by petfinder.my and they are a non-profit organization, so there is no direct business value from this project. Main value from this project that we can think of is social value. Below we are trying to explain the background or current scenario and later on we will explain how this project can solve some of those problems more efficiently.

Street animals like cats, dogs and others live a poor life without proper shelter and food. They also create some social issues like damaging garbage trash in search of food or may damage some public properties. However, on the other side some people are fond of having a cute pet to adopt and they are not getting it easily because of cost and scarcity. So, we have two way problems and petfinder.my come up with a solution to take the street animals and find their suitable owner.

So, we can see a nice idea has been generated to solve the street animals shelter problem. What are the challenges and steps to implement that? First, petfinder.my takes a picture of the street animal and they upload it to their site and may be shared on social media to get potential adopters. They also collaborate with animal lovers, media, corporations and global organizations. However, here is a challenge of taking a cute picture that attracts the adopters more rapidly. So, they use a basic tool called Cuteness Meter to measure cuteness of the photo. However, they also analyze picture composition and other factors compared to the performance of thousands of pet profiles. In this project, our primary goal is to use this performance data with various features and features derived from original pictures to improve the cuteness meter algorithm.

So, where are we now?

To solve the challenges mentioned above, we have been provided some 9000+ data with features of pet images and their PawPopularity score. We also have the corresponding actual images for our analysis or extract some features from them.

We are planning to follow following steps:

1. Divide the provided training set with the Train, Validation and Test set.
2. We will try to optimize our model with the Train set and test it on the validation set first and later on when we are confident enough then we will train our model with train and validation set together and test it with our validation set.
3. We will iterate the step 2 couple of times with different models and get the best model
4. When we are confident about the selected model we can try it out on actual test data by submitting it.
5. Maybe we will not get desired results and so we will try to extract features from pet image datasets provided by them and try to improve the prediction.
6. In the upper mentioned steps, we may need to perform data cleaning and feature engineering or may be some customized process on data.

So, if we are thinking of the risks for this project then it might happen that the mentioned features may not have actual impact on popularity score and it may end up more sophisticated analysis on image features.

Task 3. Data understanding

Gathering data

The initial data for this project does not need to be gathered since it has been provided by the Kaggle competition creators. In this case this data is photos and some manually created metadata to act as example criteria for evaluating the photos. However, if any additional criteria is decided to be added, then any labelling will have to be done by this team's members, which will increase the time cost.

The data is limited by the creators of the competition, as they provide both the images and some example metadata for evaluation criteria, but this metadata will only be available for training and will not be available to guide the trained model during actual evaluation of the images.

Describing data

As mentioned, the data that has been made available by the "PetFinder.my - Pawpularity Contest" creators comes in two forms. One is image data of the pets; 9912 images of cats and dogs, no other species, are provided. The other is what they call 'photo metadata' that is a set of manually labelled values corresponding to each photo. The labels were chosen by the competition's creators according to what they think is important for the photo to be popular and are supposed to help with creating the evaluation model, but will not actually be provided for the model. Altogether there is a little over one gigabyte of data available, 1.04 GB to be exact, which includes both the photos, and the photo metadata that they have created.

The photos are all in the form of .jpg files. All of these vary in resolution, meaning size and shape, and also in the content, having different type of pet (cat or dog) in the same picture, having several of the pets in the same picture, and having some user modifications in the picture, meaning a collage or some added features like text or drawn shapes.

The metadata comes with the following manually created and seemingly arbitrarily selected labels:

- Subject Focus - Animal stands out on the picture
- Eyes - Both eyes are facing front or near front with at least one eye clear
- Face - "Decently" clear face facing front or near front

- Near - Single pet taking up significant portion of photo
- Action - Pet in the middle of an action
- Accessory - Accessories or props like toys on the picture
- Group - More than one pet on photo
- Collage - Photo is retouched, meaning collages, drawing on the photo, frames, etc.
- Human - Human in the photo
- Occlusion - Something covering the pet
- Info - Custom added text on the photo
- Blur - Noticeably out of focus photo, if this is 1, “Eyes” is 0

These manually set labels provide a good starting point for the features that might be important for evaluating the popularity of the image.

Exploring data

This step, for the data available to us, where the labels are already each set to a binary value of 0 and 1, is almost pre-completed for us. Each row in the numerical data is also connected to an image with what seems to be a generated hexadecimal code known to this group as a Unique User ID (UUID).

As for the main issues with the data, the labeling has been done manually and, it seems, subjectively, which might introduce issues with consistency of the data. However, the popularity scores themselves are also inherently going to be influenced by as of yet unknown features that the viewer might only perceive unconsciously, so there might be some balance introduced by this fact.

There might also be some features that were thought to be important or relevant by the competition authors, but might not actually impact the popularity score. In turn, there might be some important features yet to be determined that the authors did not consider. Trying to evaluate these features is thus important.

Verifying data quality

Other than the inherent problems with evaluating a subjective popularity from a data viewpoint, that were briefly brought out in the previous section, there are no issues with the data, and the numerical data does not need to be specially cleaned, changed, or formatted. When the images are used in determining the score, then they will most likely be manipulated in terms of size and colour and converted into a machine understandable format.

Task 4. Planning the project

We have separated our plan into three sections, the first standing separately and the third being dependent on completing the second. In our mind, each of the sections is an achievement by itself, if completed. The steps in the sections are not comprehensive and seek to provide a rough outline.

As for the time allocation, due to being beginners at this sort of work we do not know exactly how much each task will take, but we are looking to allocate time as follows:

- Masud Rana - ~20h on **first section**, ~10h on **second section**, any remaining time on **third section**
- Erik Kõiv - ~20h on **second section**, ~10h on **first section**, any remaining time on **third section**
- We have not heard from our third group member for some time now, so we cannot allocate anything to him right now

First section

1) Use provided features to train and test models used in and known from HW tasks.
2) Seek to determine feature weights (importance) statistically.
3) Apply the determined weights to the previously trained models.

Second section

4) Create a naive image classification model using a neural net to input an image and output the popularity score.

Third section

5) Separate the provided images by species, using a pre-trained machine learning model available online (such as YOLO framework trained on COCO dataset).
6) Improve on the previous section’s model by splitting it by species and further training it to hopefully account for people’s personal preference for cats or dogs.

7) Create a final multi-step popularity score determining process that separates by species, evaluates separately, and then outputs a suitable file for the contest submission.

Tools expected to be used, but not limited to

Python, Google Colab, Jupyter Notebook, Sklearn, Seaborn, MatPlotLib, YOLO, PyTorch