CAS ADS – Module 5 – Peer review

Module 5 / Peer consulting report

Frederic Bärtl

Frederic.baertl@gmail.com

Peer review of Filipe Maximiano & Laura Cunha Silva

M3 project: *An investigation into FRD behavior & Using Machine Learning for in-field species' identification*

Background:

Laura and Filipe have done two different "sub projects" during the Module 3 related to their veterinary work. While Filipe focused on the classification of different categories of animals, Laura tried to identify patterns in the movement of "street dogs"

The peer review will take both sub projects into considerations. As those projects and the underlying datasets are independent, I will analyze them individually starting with the "identification of species".

Laura and Filipe (the group) are likely to use the experience made in Module 3 and further elaborate on the topic of "street dogs" patterns, thus the recommendations made are not only useful in a retrospective manner but will further be useful for the Module 6 final project.

Subproject I

Identification of species - Preparing the dataset

Laura and Filipe wanted to make an automated classification of field photos taken by different animals to create an algorithm that might be used in the future to pre-classify animals. Initially, they were aiming to use drone footage but the size of the animals on the images was to small and the top-down view made it very hard to distinguish different animals, thus the group started with an "easier" task to try automated classification of photos taken on the ground/ in the field. The 11 animals categories consist of: Horse, Cat, Camel, Chicken , Cow, Dog, Donkey, Goat, Pig, Sheep and Turkey.

Because of the small size of the dataset of images taken during field review (249 photos of all categories combined), the group decided enlarge the dataset with additional images take from a Kaggle dataset (1187 photos).

Identification of species - Results:

The group split the dataset into a training set of 1154 pictures and a test set of 282 pictures. They used different CNN models with limited success. The “MobileNetV3Small” Model only resulted in 2% accuracy. The “VGG19” scored even worst with only 1% accuracy. With “ResNet” the group managed to reach 6 % accuracy which is still far to small for any useful application.

Identification of species - Challenges:

The group was faced with two main issues that might explain the low results: a) the high amount of images from Kaggle that are different from the photos taken in the field: Those pictures vary substantially from the field photos as they all have a focus on the animals and a minimal amount of “background noise”.

To account for this difference, the group re-analysed the 249 field photos and cut out the pictures to mimic the Kaggle image “quality”. In addition, I recommend to think this idea further and account for the fact that Kaggle data and the “field” images have very different photo angles used to show the animals. This could be corrected by using data augmentation and randomly turn the images by a certain degree, resulting in a more heterogeneous dataset that might better “integrate” the KAGGLE images.

Nevertheless b) the overall amount of images is simply to small to make use of advanced ML methods like CNN. The total amount of 1436 images is already very small but taking into consideration that we are facing 11 categories leaves us with approx. 120 images per category. After splitting this into training and test dataset, the problem only gets worst as the size per category to train the model is far below 100 images. This issue is further intensified by the fact that the number of images vary substantially per category (more than 200 images for dogs but only 25 images of camels).

I therefore recommend to focus on less categories and further enlarge the dataset with additional images. Use data augmentation to have a more “standardized” set of images per category and try to use models that require a smaller amount of training data to yield useful results.

Subproject II

Movement patterns of “street dogs” - Preparing the dataset

In the second sub-project, the group tried to analyze movement data of street dogs a specific area in Guatemala. The initial dataset consisted of 52 street dogs from the same area. The idea behind the analysis is to identify hotspots where those dogs are in contact to be able to determine certain origins of diseases.

The data itself is a high amount of GPS coordinates for each of the dogs that are collected every X minutes. When running a clustering analysis, the group realized that the data set of all 52 dogs is to big to provide any meaningful results as the amount of clusters predicted by the model exceeded 40 clusters, making in visually impossible to be distinguished.

As a consequence, the group decided to perform the clustering analysis on one specific dog.

Movement patterns of “street dogs” - Results

The analysis of one dog resulted in in a handful of clusters with two main areas of “time spent”. When using kmeans, the model predicted 3-4 clusters to be present. The application of an elbow curve indicated 2-3 clusters which is also in line with the “Gausian Mixture” analysis.

Movement patterns of “street dogs” – Challenges

The group noticed that the GPS trackers are not always working thus sometimes dogs do not send a signal for a long period of time while other dogs send their positions constantly. This results in non-standardized datasets in the overall 52 dogs population. It might be useful to start with groups of dogs with similar data points to have a more homogeneous sample. On the other hand, comparing a dog with many more data points might reveal that a certain geographical area is creating the GPS issues. Overall, I recommend to split into smaller groups to better understand the interaction of the different dogs before enlarging the number of dogs too much.

A further improvement might be to exclude the “home area” of each of the dogs from the dataset to reduce the number of clusters that are not of interest as those clusters tend to be isolated to individual dogs. This will reduce the data size and at the same time reduce the number of clusters without losing information on the actual scope of this analysis. This data “reduction” would require having the “home area” coordinates for each dog and define a certain area around these coordinates. According to my understanding, all this information is theoretically available.

Future insights:

The group will reconsider the images used in Sub-project I rather than question the methods used. Thus I recommend to dig deeper into data augmentation methods, when they find a large enough dataset to combine with the “field photos”.

With regards to the movement patters of dogs, I believe that the group is already on the right path by thinking about potential data reduction methods to focus on the interesting information. To overcome the data issue resulting from the GPS quality, it might also be useful to take data from longer-shorter periods of time for specific dogs to account for the datasize or further dig into options of weighting.

Useful links:

Data augmentation methods (review):
https://www.tensorflow.org/tutorials/images/data_augmentation

Application of weighted k means:
https://towardsdatascience.com/using-weighted-k-means-clustering-to-determine-distribution-centres-locations-2567646fc31d
https://medium.com/@dey.mallika/unsupervised-learning-with-weighted-k-means-3828b708d75d