Machine Learning for Agricultural Applications

Assignment 11 (Last Assignment)

Prof. Dr. Niels Landwehr
Dr. Julian Adolphs

Summer Term 2020

Release: 17.07.2020
Discussion: 23.07.2020

Task 1 – Humpback Whale Identification [50 points]

In this task, you will use the "Humpback Whale Identification Challenge Dataset" from Kaggle (`www.kaggle.com/c/whale-categorization-playground`).
You need a Kaggle account to download the data, which is available free of charge.

The data set defines the problem of identifying Humpback Whales by their flukes. Your task is to be creative with this data set and apply a method of your choice, based on what you have learned in the lecture.
We will interpret the data set as a classification problem with a training and test set as follows. The predefined test set from Kaggle is not suitable, because as a competition test set it does not contain any label annotations. We will therefore not use it. Instead we have to use a part of the original training set as a test set. We will filter the original training data set by removing all whale individuals for which the number of images is smaller than a given threshold. The remaining whale individuals constitute the classes in a multiclass classification problem. On the next page you find a data-preparation code snippet which can be used to prepare the data. It contains a variable `n_min`, which gives the minimum number of images that must be present for a certain whale individual to be included. If you check the training set, you will also see that there are many pictures of whale flukes annotated with "new_whale", which we will also remove.
If you set for instance `n_min=60`, only 4 whales will be kept in the data, because for all other whales, there are fewer images. The smaller you chose `n_min`, the more whales will be taken into account and the more classes there will be. Obviously, the correct classification becomes harder if you have more classes (whales). So test your model for different `n_min`. After filtering the data in this way, split the remaining data into a training and a test set, train the model and evaluate it on the test set. Of course you are also free not to use the prepared code snipped and write your own code instead.

As a starting point for the keras/tensorflow-part, you can use:
`www.kaggle.com/anezka/cnn-with-keras-for-humpback-whale-id/comments` or any other "notebook" from Kaggle.

Try to improve the validation accuracy as much as possible with methods from the previous exercises. Plot the progress of the validation accuracy during training.

Figure 1: Humpback whale flukes.

Code - Snipped for Data Preparation

```
size = 100  # image size
n_min = 60  # minimum number of images (5-60)

df_all = pd.read_csv("humpback-whale-identification/train.csv", sep = ",")

input_path = "humpback-whale-identification/train/"

# filter out all whales with label "new_whale"
df = df_all[df_all.Id != 'new_whale']

# and filter out all whales with less than n_min images:
drop_list = []

for i in df['Id']:
    if df['Id'].value_counts()[i] < n_min:
        drop_list.append(i)

for i in drop_list:
    df = df[df.Id != i]

nw = len(df['Id'].value_counts())  # number of whales
mm = len(df)                       # number of images

print('\nWhale individuals:  ', nw)
print('Number of images:   ', mm)
print('\nNumber of images per whale individual:  ')
print()
print(df['Id'].value_counts())
```