Dog Breed Identification

Team members: Uku Tonsiver

Link to Github repository: https://github.com/MiksMaSiinOlen/IDS2022-Project.git

Task 2. Business understanding

1) Identifying business goals

Understanding classification, machine learning and neural networks is certainly a very big part of data science. Therefore, doing the data science project appeared to me as a great opportunity to learn a lot about these topics. Hence, I chose to make a classifier that identifies dog breeds from images.

The business goal of this project is to make a program that, given an image of a dog as the input, will correctly identify the breed of the dog in this image.

I will consider the project successful if it achieves a grade above the 10 point threshold necessary to pass the course.

2) Assessing the situation

The project will have one person working on it. That is Uku Tonsiver, also known as me. The data used for this project consists of a training set of about 10 000 images of dogs with already identified breeds, as well as about 10 000 images of dogs without the breeds added. The dataset is publicly available at site https://www.kaggle.com/competitions/dog-breed-identification/data?select=train. The scripts for the project will be written in Python using Jupyter Notebook software. Hardware for this project's use will be an HP laptop provided by the University of Tartu with 4 core multithreading processor and 16 GB of installed physical memory.

The deadline for the project is set at 12th of December, 2022. For the project to be acceptable, it must present written code for the classifier program, text explaining how the code works and a poster explaining the data mining's results. As stated above, the datasets used for this model are publicly available.

The main risks threatening this project's completion are running out of time or knowledge. To handle these risks, I will set and follow a well defined time plan for the project, as well as do extensive research on topics of classification, neural networks and machine learning.

Terminology used in this project:

- model – computer program that reads data, executes the identification process and outputs the results
- classifier – program or function that determines the property we want (in our case the dog's breed) based on other properties of the input data (in our case an image of a dog)
- algorithm – set of instructions to achieve a desired result (e.g. the identification of the dog's breed)
- neural network – an architecture that uses the outputs of several layers of nodes to give the final output
- training (a model) – running the model with a large amount of data and, depending on whether the output is correct or not, adjusting the parameters of the model

This particular project does not have any foreseeable financial costs nor benefits. However, the project will have the cost of human labour by myself amounting to an estimated 63 hours (see project plan). I would consider the project's benefits to be gained knowledge and experience in data science field and, of course, hopefully passing this course.

3) Defining the data-mining goals

This project has two major data-mining goals. The first goal is to build a model that can classify dogs into breeds from images. The second goal is to find out which dog breeds are the easiest to classify for the model. This will be done by analysing the successful classification results and determining which dog breeds were most often correctly classified.

As success criteria, I will evaluate how big proportion of input images are correctly classified. I will consider the model successful if the dog breed is correctly identified in at least 70% of the images tested.

Task 3. Data understanding

1) Gathering data

For this project, I need a training set of images of dogs where the breeds are given, and also a test set of images of dogs where the breeds are not given.

Such datasets are provided by Kaggle on the link given above. The images have unique IDs as their file names. File labels.csv provides the breeds for images in the training set by the IDs.

Defining selection criteria is not necessary for this project, because the datasets already consists exclusively of images adequate for this model.

2) Describing data

As mentioned above, the data for this project is provided by Kaggle. The images in the training set and test set are in .jpg format. Additionally, a file named labels.csv is provided. The file labels.csv consists of two columns: id and breed. Entries in the id column are unique IDs of the images in training set and entries in the breed column are corresponding dog breeds. Train set consists of 10 222 images of dogs and labels.csv file consequently also has 10 222 rows (not accounting for the row stating the names of columns). The test set consists of 10 357 images of dogs. The provided datasets are excellent for this project since they were made for that very purpose.

3) Exploring data

In labels.csv, there are no empty entries in either id column, nor breed column. Each row has the ID of an image and one nominal variable, that being the dog's breed. The dog breed with the most images in the train set is the Scottish Deerhound with 126 instances and two breeds with the least images are Eskimo Dog and Briard, both having 66 instances. I would say that the differences between breeds' occurrences are not too great and all dog breeds are represented sufficiently. The images in both train set and test set are of different dimensions and resolutions. One interesting thing I noticed is that in some images, there are multiple dogs of different breeds. However, the labels.csv file only provides one breed value for each image. I imagine that images containing more than one dog breed may confuse the model and slow the learning process.

4) Verifying data quality

The only possible quality issue I could identify is the occasional situation when an image contains more than one dog breed. However, images like this make up only a small fraction of all the images so this should not be a fatal issue. There are no missing values in the dataframe and the values appear correct.

Task 4. Planning the project

1) Project plan

The tasks in the plan are in roughly chronological order. Each task is followed by an estimation of how many hours it takes.

1) Downloading, extracting and importing the data into work environment – 1 hour (completed)
2) Verifying data's quality and adequacy for the project – 2 hours (completed)
3) Preprocessing data – 3 hours

   Note: This includes, for example, making images easy to read and process for the computer
4) Research on neural networks, image classification and other topics that may become relevant – 12 hours

Note: This task will most likely be spread throughout the entire project timeframe, not be executed in one go

5) Building a neural network model or finding a prebuilt one adequate for this project – 12 hours
6) Training the model, adjusting parameters, experimenting with different algorithms – 20 hours
7) Testing the model with test set, evaluating the results by uploading them to Kaggle and analyzing scores – 3 hours
8) Cleaning up and commenting the code – 2 hours

   Note: I intend to also comment the code on the go, in this task I will just take one last look at the whole thing and add comments where necessary

9) Making the poster and presenting – 8 hours

2) Methods and tools used for this project

I will use Jupyter Notebook software for programming in this project. Python 3 is the language in which code will be written. I will use Python libraries Pandas and NumPy for managing the data and Tensorflow's Python library Keras for building/training the neural network. These are the tools I can think of at the moment, some other tools may become necessary while undertaking the project.