Final Project: Mushrooms Classification

10 APRIL

KSCHOOL
Created by: Jaime Redondo Catalán

Introduction

There are many species of mushrooms in our environment, they are located in different wooded areas of our planet, one of the countries with more richness of mushrooms is our country, Spain. In this country there is a lot of diversity, we are going to study the mushrooms of one of the provinces where this biodiversity is most evident, Guadalajara.

¿Why Computer Vision?

Deep Learning is undoubtedly one of the things that most attracts my attention within data science and especially image recognition. I have been inclined to do this type of project as opposed to other ideas for several reasons:

- **Innovation:** I wanted to do something different from what I usually do, I didn't want to pigeonhole myself in an idea associated with my work. In the end, if you don't take risks, you don't win.
- **Further Learning:** If there was one thing I didn't know much about and wanted to learn with this project, it was Computer Vision, so yes, it was a very good opportunity to learn.

¿Why Mushrooms?

I remember those walks through the pine forest with my uncle and my grandfather, passionate about mushrooms, they were always on the lookout in case their nephew/grandson happened to pick or touch a mushroom that could be toxic, and when I found one I always asked them first if it could be picked. At that time, I would have liked to have something similar to what I will explain in this project, a mushroom classifier, with a picture to know what type of mushroom it is most likely to be, great, isn't it?

¿Why Guadalajara?

Firstly, because of its diversity, and secondly, because it is where it is most applicable to me, the author of this work, since it is where my origins are, where my second home is and where I have gone every year since I was a child to pick mushrooms.

Raw Data Description

It will be briefly explained what it consists of and how the project dataset has been constructed.

The initial dataset consists of 60 images of each of the mushroom species of greatest interest found in Guadalajara, according to the website https://www.amivall.com/documentos/epmguadalajara.pdf

You can also choose the region you want (Guadalajara or All regions in Spain), the path where you Will sabe the potos and the number of potos you want per mushroom to put in the dataset. This is explained in the repo in the proper README.

In our case we are going to work with the dataset of 60 images of mushrooms from Guadalajara. If you choose other parameters , note that the neural network is designed for the parameters of Guadalajara mushrooms and 60 images per mushroom.

Once this dataset was obtained, a technique was used to increase the number of images for each mushroom called **Data Augmentation** (we will see it in the following section where the methodologies and techniques used are explained).

Methodology

Each of the techniques and methodologies used to carry out the project are explained.

Firstly, a dataset with images corresponding to the 49 species of mushrooms mentioned on the website **https://www.amivall.com/documentos/epmguadalajara.pdf** was needed. For this purpose, web scraping was carried out, i.e. collecting images of these mushrooms in google images up to a limit of 60 for each type of mushroom (for our project we decided that this was sufficient). The result was to have for each mushroom a folder with the name of the mushroom containing 60 images of that mushroom. This part of the project was done from Jupyter Notebook in the script Scraping_mushrooms.ipynb (we can call it from Mushrooms_Classification.ipynb script), you only need to choose the option Use_data to use the dataset we are talking about, if you want to create your own dataset , choose the option Scrape_mushrooms and follow the instructions given in the script.

Once the 49 mushroom folders were on our disk, it was decided to move all these folders with their images to Google Colab in order to improve the performance of the tasks that were to be carried out later. Once the folders with the images were uploaded, we passed each of the images through a pipeline consisting of:

- Convert each image to a 28x28 image (for homogeneity in images, capacity and speed of execution).
- Pass these new 28x28 images to matrix form, that is to say, each 28x28 image will be a 28x28 matrix where row 1, column 1 will correspond to the first pixel of the image, this pixel will be represented by an array of dimension 1x3 (that is to say a vector of dimension 3), where component 1 of the vector will be a number from 0 to 255 that will represent the intensity of

the red colour in that pixel, component 2 in an analogous way to the green and component 3 to the blue.

- Once we have these matrices where each cell of the matrix has a 1x3 array with 3 numbers from 0 to 255 representing the intensity of red, green and blue respectively, we divide each number of each array of the matrix by 255, this is done to standardise the data, so that now each of these numbers will range between 0 and 1.
- Once we have a 28x28 matrix where each cell is a 1x3 vector with values between 0 and 1 representing rgb colour intensities, we are going to perform the Data Augmentation technique to obtain more images, this technique consists of, for each image we have, applying techniques such as rotation, symmetry, blurring, contrasts... and many more transformations that can be applied, in order to obtain new images of the same image but being different from the original. In our case we apply 7 transformations per image.
- Once we have the 7 transformations per image, we join each matrix (one matrix per image) to a list, and the name of the mushroom corresponding to the image to another one, so we will have a list of matrices of 48 species x 60 original images of each one x transformations of each image by Data Augmentation (about 21.000 images) and another list with the labels (names) of each mushroom for each image.
- We then divide the list of images into train(80%), validation(10%) and test (10%) and the list of labels is divided in the same way, obviously keeping the order so that each image has the corresponding label.
- Once we have train, test and validation we pass the lists of labels (train, test and validation) to OneHot, that is to say, for each of the mushroom labels a column is generated so that each mushroom will be a vector of zeros and a one corresponding to the column of the mushroom it is. So if we have 16.000 mushroom names for the train, the OneHot will get 16.000 vectors of zeros and a one corresponding to the mushroom, that is, if there are 49 species, a matrix of 16.000 x 48. We do this OneHot for the list of labels train, test and validation.
- Once we have everything ready, we insert the training images and the labels corresponding to our neural network, in addition we will also insert the validation labels to see if our neural network is working well. The neural network that we have implemented is a convolutional neural network (specialised for image processing) as it maintains the location of the pixels, a very important property in images. For the output of the neural network we have used the activation function softmax (used for when there are multiclass outputs, with more than 2 classes) with 48 neurons, corresponding to each one of the mushroom species, the batch size we have used 32, the network takes 32 images each time to improve its hyperparameters, the epochs we have established them in 200, it worked very well, and to increase this number did not improve the result and if it increased the computation cost, the epochs are like iterations of the times that the algorithm passes through the dataset. We saw that it was overtraining with other parameters and we also added drop out to remove a % of neurons in some layers, so that it would not overtrain.

We train the neural network with the optimizar base don gradient descent , Stochastic Gradient Descent, with a learning rate of 0.01

Summary of Main Results

The training result of our neural network is compared as it trains with the validation labels, and for each epoch it shows us the result, after 200 epochs, we have a result of approximately 65%. Subsequently, we now pass the list of test images and the list of test labels to check if it really is 65% correct. It is also approximately 65% correct.

Conclusions

In short, we have been able to identify the 48 most important species of mushrooms in Guadalajara, so that if someone is curious to see which mushroom it is, with this project they can be identified and the problem is solved! And if someone wants to see if they can be eaten or not... they will also be able to know if they are identified, although it is true that all of them can be eaten, but some of them can only be eaten once.