Schneider Electric European Hackathon - Data Science Zero deforestation mission 19 Nov 2022

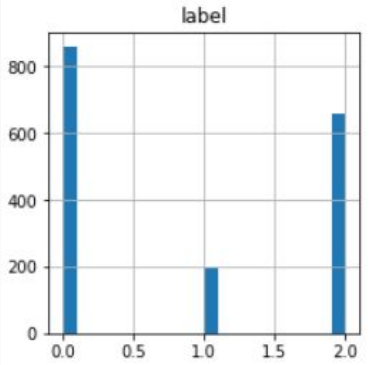
Team: ThinkTank
Anton Chernysh,
Ramon Mateo

Description

The challenge consists of developing an image classification model in order to predict the type of deforestation in a satelital image. And the final goal of the model is to detect early the problem and perform actions to protect the lands. Apart from the satellite images, we have the coordinate of the image taken and its year.
To get the maximum score, the team participants have to present their code, the prediction results of their model and a presentation explaining the challenge, how they have approached the problem and why.
For this, our team has begun to carry out an analysis of the data to understand what correlation there may be with its type of deforestation. Subsequently, the first models have been made to see what the current state is, to see how complicated the task is and in order to have the first version of the complete flow. Finally, improvements have been made such as **DataAugmentation** and **TransferLearning** together with the validation of the results.

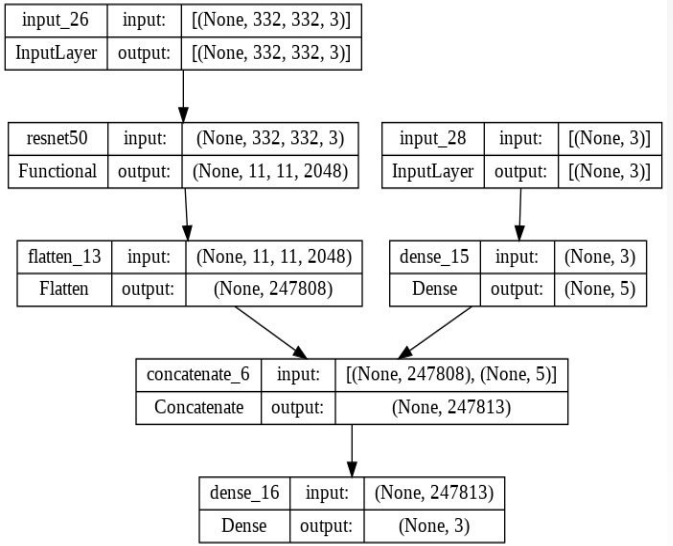
Analysis

Several functions have been implemented to carry out the analysis of the images. We have used functions to show a number of images for each category, histograms, correlation graphs of each feature with its label and others.
After the analysis, it was clear that the data presented a very strong imbalance. On the other hand, the values of coordinates and coordinates did not present a strong relationship with the type of deforestation.
First we decided to train the values as they were, we got good results, *f1_score* greater than **0.6**. And after balancing the data through data augmentation, we managed to raise it to **0.71**.



Training

Since the data are images, we weren't sure if they would fit in memory, which is why we implemented a Tensorflow **DataGenerator** to load the images dynamically during training. But finally we saw that they did fit and that this allowed us to train faster.
An important step was **TransferLearning**. For that we tried to use several models pretrained with the **Imagenet** dataset and the one that worked best for us was Resnet50. It should be noted that for all the experiments, we used a split with 20% of the data for validation (which were not processed with data augmentation) and the rest for training.
And finally, we have carried out several experiments making an assembly of two models to be able to enter the values of the year and coordinates to the model.
We tried to implement the model that you can see here. The main idea was to use coordinates and year after extracting the features and send this data with features extracted from ResNet50 to NN to predict the label. But we found a bug on the training phase that we couldn’t solve so this architecture was implemented but not tested.



Results

Throughout the competition we have been presented with different challenges that we have been solving. The first results were good thanks to using already pretrained models but we had not yet done any kind of preprocessing to the images nor did we use the additional features of the year and the coordinates. That's why we spend most of our time perfecting that.
On the one hand, we balance the data and try different ways of altering the images to create more data without reducir the score during the training of the model. And on the other hand we tried to incorporate the mentioned features along with the convolutional model.
The final results obtained on 20% of the validation data have been an f1_score of 0.71.