Bern, 31/03/2021

Bern Data Science Day 2021 - April 23

# MRI images classification and a DCGAN neural network model as ultimate approach to an imbalanced dataset

DS applications and challenges in Medicine, Natural Sciences, and Engineering

Gianluca Camparini

CAS ADS student 2020/2021 – University of Bern

gianluca.camparini@yahoo.it

## Abstract

The Brain Tumor Classification (MRI) dataset from Kaggle is already split between training and testing folders, showing an imbalance between the classes, especially the No Tumor class which is much less represented. Also the ratios of the different tumor classes between the training and testing dataset is not the same.

A first study with a CNN classifier model with the given dataset split between training and testing has been done though, leading to good results in terms of training and validation accuracy (95.5%), but poor accuracy on the test data set (73.3%).

Then the dataset has been completely rebuilt from the original one in Kaggle putting all the images with their respective labels in only one folder. Then a split in training, validation (10%) and testing (15%) has been done. The same CNN model as before has been used leading to a much higher score on the test accuracy (95.3%).

However, the lesser number of examples for the No Tumor class holds, therefore from this last CNN model two further improvements have been carried out:

- Class Weights: the idea is to have the classifier add weight to the class less represented. This will cause the model to "pay more attention" to examples from an under-represented class
- Oversampling the minority class: a related approach would be to resample the dataset by oversampling the minority class, which in this case is represented by the No Tumor class.

These new approaches led to a test accuracy respectively of 93.3% for the Class Weights approach and of 95.7% for the Oversampling approach.

In the attempt to further improve the results and in particular the accuracy on the testing dataset a DCGAN Deep Convolutional Generative Adversarial Network model has been built.

The key idea is to synthetically create images for the minority class (No Tumor images class) and to balance the dataset with the same number of examples for all four classes of tumor.

While the results at the time being could not go beyond the previous results leading to “only” an accuracy of 92.4% on the testing dataset, in principle this approach should potentially lead to higher results with a longer and finer tuning of the DCGAN model for the generation of images.