FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE D
COIMBRA

Advanced Machine Learning
2022/2023 - 2º Semester

What can you see?

João Correia
Ernesto Costa
Nuno Lourenço

1 Introduction

We present here the second practical project, part of the students' evaluation process of the Advanced Machine Learning course of the Master in Engineering and Data Science of the University of Coimbra. This work is to be done autonomously by a group of **two students**. The **deadline** for delivering the work is **14 of May** via Inforestudante. The quality of your work will be judged as a function of the value of the technical work, the written description, and the public defence. All sources used to perform the work (including the code) must be clearly identified. The document may be written in Portuguese or in English, using a word processor of your choice[1]. The **written report** is limited to **8 pages long**, but in special, justified, cases (e.g., the need of presenting many images and/or tables), that number may be increased accordingly. The document should be well structured, including a general introduction, a description of the problem, the approach, the experimental setup, an analysis of the results, and a conclusion. The report should follow the Springer LNCS format. The Latex and Word templates are available in the Support Material of the course. The final mark will be given to each member of the group individually. To do the work the student may consult any source he/she wants. Nevertheless, plagiarism will not be allowed and, if detected, it will imply failing the course. While doing the work and when submitting it, you should pay particular attention to the following aspects (whose relative importance depends on the type of work done):

- description of the approach to the problem
- description of the general architecture of the methods used;
- description of the experiment, including a table with the parameters used which should allow full replication;
- description of the evaluation metrics used for the validation: quality of the final result, efficacy, efficiency, diversity, or any other most appropriate;

Do not forget, besides what was just said, that it is fundamental: (1) to do a correct experimental analysis; (2) to do an informed discussion about the results obtained; (3) to put in evidence the advantages of the chosen alternative.

[1] Latex is preferred

2 Problem Statement

Image classification is a computer vision problem where Machine Learning has obtained impressive results. In the past to tackle this task practitioners had to use expert knowledge and perform feature engineering, carefully crafting and creating visual features for Machine Learning models to learn. After this step classical and shallow models were usually employed such as Adaboost, RandomForests, Support Vector Machines, and Multi-layer Perceptrons (Artificial Neural Networks with Fully connected layers).

Despite making their first appearance in the late 80s early 90s, Convolution Neural Networks (CNNs) became popular since 2010 due to technological advances, namely in GPU computing, and have been heavily used for image classification problems with great success surpassing most classical approaches. The main feature that this type of Neural Network provides is that feature extraction and learning is an integral part of the neural network, performing automatic feature extraction using the whole image as input, where the features are then passed to dense layers, providing the classification power as typical fully connected Multi-Layer Perceptrons are able to.

In this work, we are going to tackle an image classification dataset of natural Scene Images using a Machine Learning approach.

3 Objective

The main objective is to analyse the dataset and create an approach that can perform image classification of the dataset. To do that you should attend to the following objectives:

- Prepare the machine learning pipeline for the image classification dataset. Explore solutions with, at least, the following ML models:
  - Multi-Layer Perceptrons
  - Convolution Neural Networks
- Exploration of different architectures, optimizers and hyperparameters

3.1 Dataset

The dataset is a small version of the multi-class colour image classification dataset - Intel image classification challenge. In this version, the images are resized from the original challenge to a $50 * 50$ pixel resolution. The number and the examples are picked and distributed in a different way than the original challenge. The different classes remain the same from the original challenge and are the following:

0. buildings
1. forest
2. glacier
3. mountain
4. sea
5. street

Figure 1 shows an example of the dataset. The Training set is composed of 12000 images, with 2000 examples per class. The Test set is composed of 5051 unlabelled images.

Folders containing the images of the Train and Test datasets are provided with the original resolution of $150 * 150$ pixels. To facilitate the construction of the pipeline the following files are also provided (but are not mandatory to be used):

- **"train_intel-image-classification-csvdata.csv"** – contains the number, path to the image, class, and the rgb pixels as 7500 columns.
- **"test_intel-image-classification-csvdata-kaggle.csv"** – contains the number, path to the image, and the rgb pixels as 7500 columns.
- **"trainX.npy"** – numpy version of the training dataset containing only the images as matrices, a numpy array $(12000, 50, 50, 3)$.
- **"trainy.npy"** – numpy version of the training dataset containing only the labels for X, a numpy array $(12000,)$.
- **"testX.npy"** – numpy version of the test dataset containing only the images as matrices, a numpy array $(5051, 50, 50, 3)$.

The main goal is to use the Training set to design implement and validate your approaches and the test will be used to evaluate the generalisation ability of your models through a Kaggle competition (check Section 4).

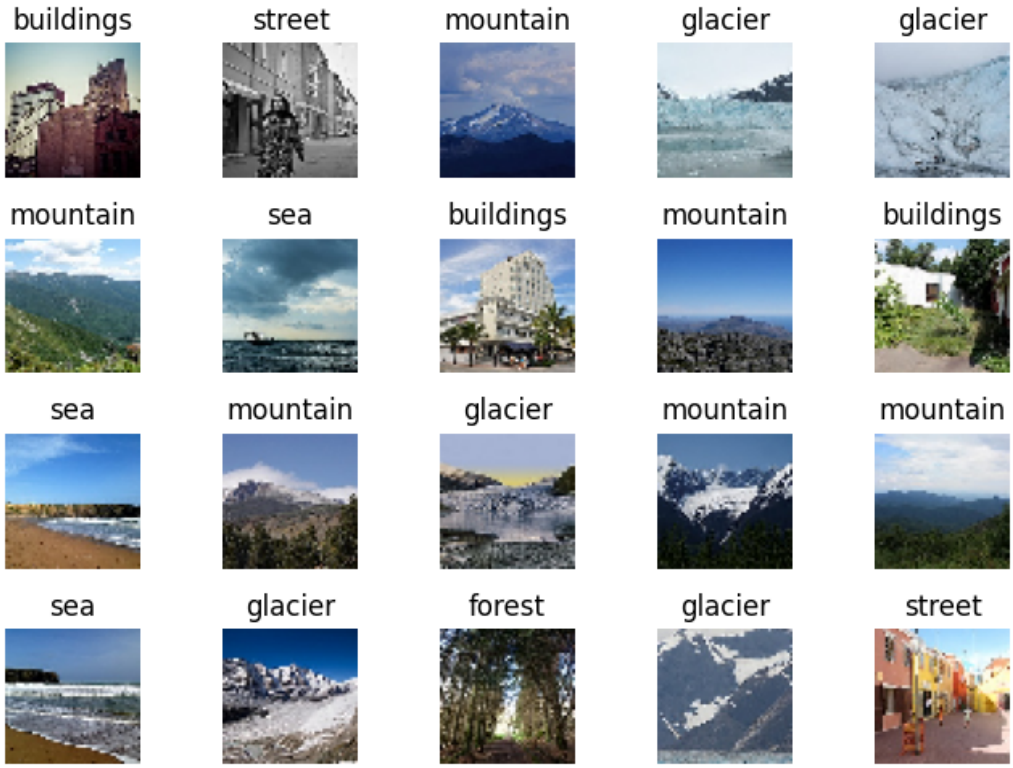

Figura 1: Sample of the training dataset.

3.2 Evaluation Metrics

Given the training dataset, you should split it into train, validation, and test to see how to fit the models that you are training/creating. Thus, the validation part of this work is crucial and you should select the most appropriate set of metrics and justify them.

4 Competition

To evaluate the generalisation ability of the developed models, we are going to use a Kaggle competition. Note that it will not impact the final mark but rather will act as a way for you to access the progress you are making and evaluate the generalisation performance of your models. The competition is available at the following address:

https://www.kaggle.com/t/6d01c1448ab643af9ee6e71d5d21ba2c

To participate in the competition, you should prepare a csv file with two columns: the first column contains the Id of the sample that you are classifying, and the second column should contain the corresponding classification label. An example of a submission file is provided along with the project statement.

5 Conclusion

A few short comments. First, the control of the progression of your work will be done during the classes (T and PL). Moreover, you can discuss eventual problems by presenting yourself during office hours. Second, the projects reflect for the most part your actual knowledge. The rest will be the object of lecturing soon after Easter. Third, we try to balance the difficulty of all the work, but we are aware that this is not an easy task and it is somehow a subjective matter. Fourth, we try to ask for a workload compatible with the value of the work for the final mark.

Methodological issues, like the statistical background, were elucidated during the previous lectures. You may use the statistical tool you feel at ease with, including the Python code that was provided. Finally, even if this is a work that asks you to do simulations and analyze the results, i.e., it has a practical flavor, there is however a theory behind the work, and you are advised to consult the necessary literature.

Good luck!