Block 3 laboratory deliverable guidelines

Advanced Machine Learning, Master in Data Science, Fall 2023, UPC

This document contains the guidelines for the block 3 deliverable. Please read with care!

General information

This project has two goals:

1. Consolidate the concepts, techniques and methods used in Deep Learning for CV, through exercises (tasks) similar to the ones presented during the laboratory sessions.
2. To give you the opportunity to apply the techniques seen during this block to a real-world dataset (open project).

The second point should cover all aspects of the modeling methodology seen in previous courses: from preprocessing to generating a final predictive model together with an assessment of its prediction quality and how it could be improved.

It is expected from you to:

- **The project is to be done in teams of two persons**; **singles are not allowed**. (Keeping block 2 teams is priority/preferred).
- The goal of the project is to solve 5 different tasks (more details in the slides):
  - Custom CNN model and analysis.
  - Play with different techniques to improve the model.
  - Interpretability.
  - Transfer learning.
  - Open project.
- In any case, you are expected to hand in a written report:
  - This document should describe the work carried out, its motivation, the problems encountered, the actions taken to solve the problems and the solutions found together with final results and conclusions of your study.

To carry out your analysis you can use Python, R or both. Remember that there are many useful packages that extend its basic functionality. If you use code or ideas or any kind of resource from elsewhere you should cite it appropriately. Plagiarism will be prosecuted.

First task

Submit a 1-page project proposal:

- It should include the composition of your team.
- The open project that you are going to work on, the reason why you chose it, together with any references of previous work on this problem if applicable.
- A title for your project.
- Include information on the data.
  - Image data: If images are black and white or not, number of images, categories of the images and number of images in each category.

Please hand this in no later than December 17th through the racó.

Report submit

The final report should include:

1. A different section for each task.
    a. Reports for tasks 1 to 4 should be no longer than two pages per task. They should include exactly what it is asked (use images and tables to summarize the information).
2. For the open project:
    a. A brief description of the work and its goals, data available, dataset description and any additional information that you may have used.
    b. Related previous work (if applicable).
    c. The data exploration process: Clustering (if any), checked outliers, normalization of the images, any needed preprocessing done to the images in order to use them (rescaling, rotating, etc.), etc.
    d. Modeling:
        i. DL models are required (CNN) they can be custom or fine tuned pre-trained ones.
        *Justify the use of any pre-trained model if not asked by the task.
        ii. Interpretation of the final model is a must.
    e. Final model chosen and an estimation of its generalization performance.
    f. Scientific and personal conclusions
    g. Possible extensions and known limitations.

Note that the report should not describe explanations seen in class; every table or plot should be appropriately described. The style of the report should resemble what you encounter in a scientific publication.
Your code should be reproducible; that means using “seeds” if your code is stochastic.
For the final delivery, make sure you include in a compressed file the following:

1. The written report (pdf document). **It should not exceed 16 pages**; if you need more space, consider placing the secondary information in a separate appendix file. Works with **more than 16 pages will be penalized** taking into account the number of exceeded pages.
2. Any script or code you have used (R/python notebooks, scripts, or any other code)
3. A flat text file with precise instructions on how to execute and reproduce your results.

Code and report should be submitted separately (namely, a python notebook or an R notebook is not a report).

Evaluation
The following are the conditions that will be taken into account in the grading of this part of the laboratory:

1. The proper use of techniques and methods presented in class.
2. The care and rigor for obtaining the results .
3. The quality of the obtained results (generalization error, simplicity, interpretability).
4. The quality of the written report (conciseness, completeness, clarity).
5. Tasks 1 to 4 are considered BASIC (as they are closely related to the problems presented in the laboratory sessions) so they are going to have a max grade of 3.5. This means that the open project will be graded with a max of 6.5.

Key dates

- **17 December**: proposal.

- **14 January**: Report and code.

Data repositories
The following datasets are recommended for the open project:

- Snake classification [https://www.aicrowd.com/challenges/snakeclef2021-snake-species-identification-challenge]
- COVID-19 X-Ray dataset [https://www.kaggle.com/datasets/mrtejas/covid-19-and-normal-x-ray-dataset-balanced]
- CIFAR-100 [https://www.kaggle.com/datasets/thanhbnhphan/cifar-100]
- Mushrooms classification [https://www.kaggle.com/datasets/daniilonishchenko/mushrooms-images-classification-215]
- Butterfly classification [https://www.kaggle.com/datasets/phucthaiv02/butterfly-image-classification]
- Apparent age estimation [https://competitions.codalab.org/competitions/7511]
- Oxford III pet dataset classification [https://www.robots.ox.ac.uk/~vgg/data/pets/]
- Eurosat images. [https://github.com/phelber/eurosat]
- Caltech 256 [https://data.caltech.edu/records/nyy15-4j048]
- Stanford cars dataset [https://datasetninja.com/stanford-cars]

In case of using another dataset, it is highly recommended to **not use binary classification problems**.

The main warning we give is to be aware of the large computational needs that some (or many) of the chosen methods may have. If you choose a problem made of high resolution images you will surely get into serious demands of CPU/GPU and RAM... Take into account that Google COLAB resources are limited (you could also check other online resources like Kaggle notebooks).

A recommendation: Check if these datasets are available through a python library or if they are available in the torchvision package. This can facilitate the preprocessing and the posterior use with Pytorch.

Others
Only one member of each team should submit information (always via the Racó)