FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE D
COIMBRA

Advanced Machine Learning
2023/2024 - 2º Semester

Do you need more signs?

Nuno Lourenço
João Macedo
João Correia

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 **2 June** of 2024 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** 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

Generative models are designed to learn the underlying distribution of a dataset and generate new data points that could plausibly come from the same distribution. Discriminative models make predictions based on specific traits, such as classifying photographs of animals. In contrast, generative models have the ability to create new instances of data, such generating a new image of an animal or a traffic sign.

Image classification with deep neural networks tipically required a considerable amount of data to operate with good performance. When there is not that much data, data augmentation strategies are commonly used to tackle the issue. Moreover, aside from simple transformations, generative models have been also used to augment datasets in order to improve the performance of the base model.

In this work, you are tasked to improve the performance of pre-determined image classifier by augmenting the dataset with images generated by generative models. The idea is to have the generative models to generate more samples for the training dataset with the overall intention to improve the performance of the classifier.

3 Objective

The main objective is to analyse the dataset, prepare and train generative approaches that learn to generate more data for the training dataset of a predefined CNN. To do that you should attend to the following objectives:

- Prepare the machine learning pipeline for the image classification dataset;
- Analyse the baseline CNN model provided;
- Prepare the ML pipeline to train with augmented datasets, i.e. baseline dataset + generated samples.
- Explore solutions to perform data augmentation via generative machine learning models to improve the performance of the current baseline model
- Explore solutions and compare the results using at least, the following Generative models:
  - Any form of AutoEncoders

  - Any form of Generative Adversarial Networks (GAN)
  - (optional) Any form of Diffusion models

- Compare the results to the baseline model with and without the generative samples added to training dataset.

Exploring other solutions than the listed ones that are suitable for the problem at hand and considered as extra work, can be as compensation points to cover problems in the listed ones above up to 10% compensation.

3.1 Dataset

The dataset is based of the "Traffic Sign Dataset" with 54 different classes. For this work we will work with the dataset at the $75 * 75$ **pixel resolution** and a subset of **10 classes** that are the following:

6 - Speed limit (70km/h)

12 - Don't go left or right

13 - Don't go right

24 - Go right

38 - Dangerous curve to the right

39 - Dangerous curve to the left

44 - Go left or straight

46 - ZigZag Curve

49 - Unknown5

50 - Fences

Figure 1 shows an example of the dataset. The dataset set is composed of 277 images, with different distribution of examples per class.

Folders containing the images of each class in the dataset are provided. The main goal is to use this data 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 different classes from the training dataset.

3.2 Evaluation Metrics

For the image classification task, 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.

For the generative models, you have metrics proposed on the literature to understand how well the generate images follow the same distribution of the training dataset. Some examples include the following metrics: Incepton Score, Fréchet Inception Distance, Structural Similarity index (SSIM).

In this particular problem the idea is to **improve the baseline model perfomance by generating more data for the training dataset**. This means, that you **will have access to the code, weights and model of the baseline** and your focus should be to augment the dataset and check the improvement on the performance without tweeking or changing the baseline classifier. **You should only change the dataset and train the baseline model as in the script provided with the augmented dataset**. You must deliver the augmented dataset alongside with the the project code and report.

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/competitions/aml-2024-project2

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 analyse the results, i.e., it has a practical flavour, there is however a theory behind the work, and you are advised to consult the necessary literature.

Good luck!