FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE D
COIMBRA

Advanced Machine Learning

2021/2022

Is this the real life or just fantasy?

Nuno Lourenço

Ernesto Costa

João Campos

March 18, 2022

1 Introduction

We present here the practical homework, 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 **22 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 **12** 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 experimental setup, the 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 be not allowed and, if detected, it will imply failing the whole 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 the 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 for 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] We strongly suggest the use of LaTeX.

2 Problem Statement

Developments in media creation technologies have been progressing over the years. In this digital era, with the advancements in Artificial Intelligence, and Machine Learning (ML) in particular, we have models capable of performing tasks that surpass the human performance, and generative Machine Learning (ML) models are no exception. We have models able to create realistic content starting from pure noise. In the right hands and with the right intent, these developments are a powerful tool to generate engagement and explore different communication paradigms. The topic is getting explored, and we now have ML models that can create images, text, audio and videos from scratch with the proper data in place (figure 1).

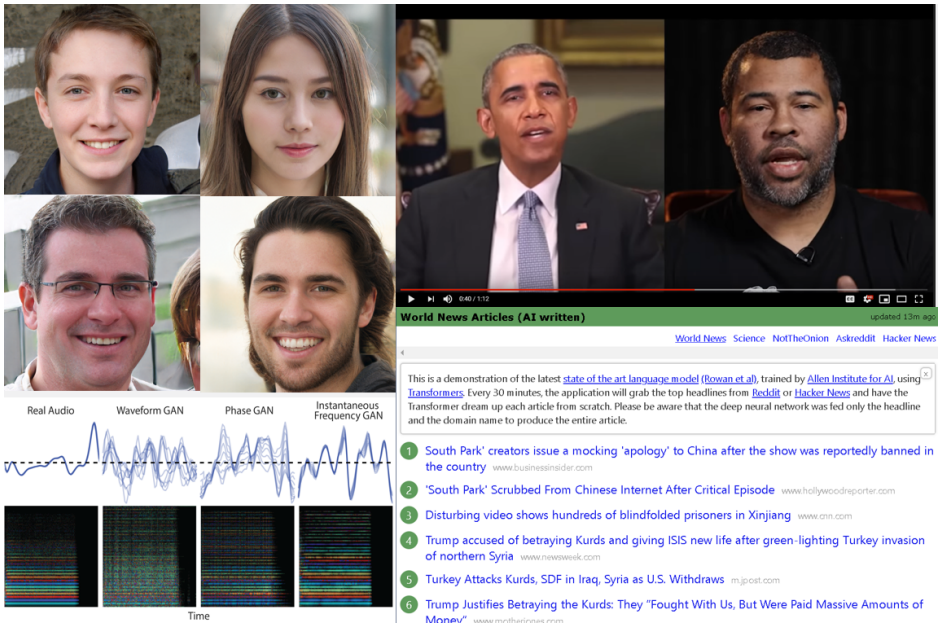

Figure 1: Deep fake example, from: https://www.youtube.com/watch?v=cQ54GDm1eL0

In the top right of figure 1 we have an approach to generate videos, commonly referred as Deep fake, that makes use of two ML models and with little effort can create realistic outputs of a person impersonating other. This content may confuse many sources of content and information. Several of these deepfakes have proliferated through many social networks platforms, and they are reaching millions of people every day. This false content can be harmful and provoke catastrophic consequences if not detected as such. On a more personal level, imagine that someone creates a deep fake of your

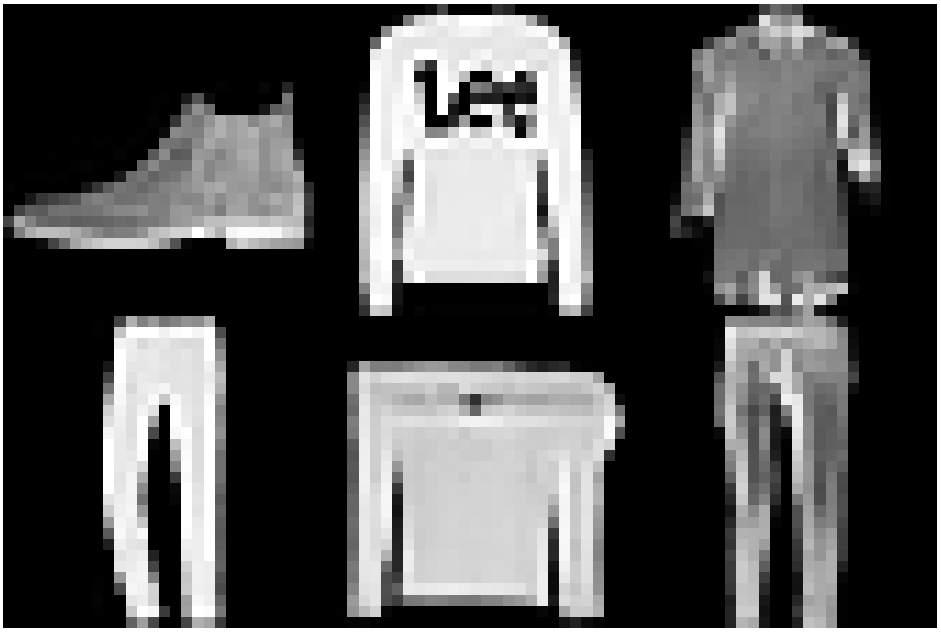

Figure 2: Sample of images from the dataset. Can you tell what is real and what is fake?

face, and uses it to access confidential information, such as bank accounts, or your mobile phone. Thus, we are on an era that what we see, hear or read requires ways for fact verification. Moreover, research efforts have been done to study and develop models that could detect if the content is a fabrication or it is true, if it is fake or real.

In this work, we are going to embrace this theme by creating models that are able to detect if a given content is real or fake. For this, we will focus on the image domain, and a dataset will be provided for you to train and test your solutions.

3 Objective

In general terms, the main objective is to analyse and explore the image dataset provided and create an approach that can distinguish between real and fake images. In figure 2 we have images from both classes.

To fulfil this task, you should tend to the following objectives:

- Prepare the pipeline necessary to process the data and create a model

to classify between real and fake images. You should compare the performance of different algorithms such as:

  - Ensembles
  - Support Vector Machines
  - Artificial Neural Networks

- Use model selection and parameterisation techniques to improve your models.

Take into account that you should not be limited to these algorithms, and that you should apply what you have learned during the course to solve this particular problem.

Dataset

The dataset is divided into a training dataset and test dataset. Thus, the training dataset should be used to prepare your approach and models and then you should evaluate the generalisation ability of your models with the test dataset. Moreover, the training dataset is organised by folders of each class, real and fake. Concerning the test dataset, you will not have access directly to the ground truth (see Competition section). You are in charge of preparing the data, analyse it and pre-process it has you see fit. Based on this step of data preparation and analysis, you should then train your models.

Description The datasets are composed of 28 by 28 pixels images in grayscale. The first dataset, called dataset_train is composed by two folders, each one containing images that correspond to one of two classes fake (Class 0) and real (Class 1). This dataset should be used to train and evaluate the models. The second dataset, called dataset_test is composed of 2000 unlabelled images, and it will be used to assess the generalisation ability of the models. You should use it to create a .csv file containing the predicted labels and submit it to the competition. The file will be automatically evaluated using the Accuracy Metric.

Evaluation Metrics

Given the training dataset, you should split it in train, validation and test to see how fit is the models that we are training/creating. Thus, the validation part of this work is to be measured in terms of the following performance metrics:

- **Accuracy**
- **Precision**
- **Recall**
- **AUC** - Area under the Receiver Operating Characteristic curve (ROC).

4 Competition

To evaluate the generalisation ability we are going to use a Kaggle competition. The competition **will not impact the final mark**, but rather will act as a way of you 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/2346fb23e61f4c53a0fbaf9b2389f82a

To participate in the competition, you should prepare a csv file with two columns: the first columns contains the Id of the image (e.g. image 0.png has the id 0), 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 at the office hours. Second, the projects reflect for the most part your actual knowledge. The rest will be object of lecturing soon after Easter. Third, we try to balance the difficulty of all the works, but we are aware that this is not an easy task and it is somehow a subjective matter. Forth, we try to ask a work load 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 easy 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!