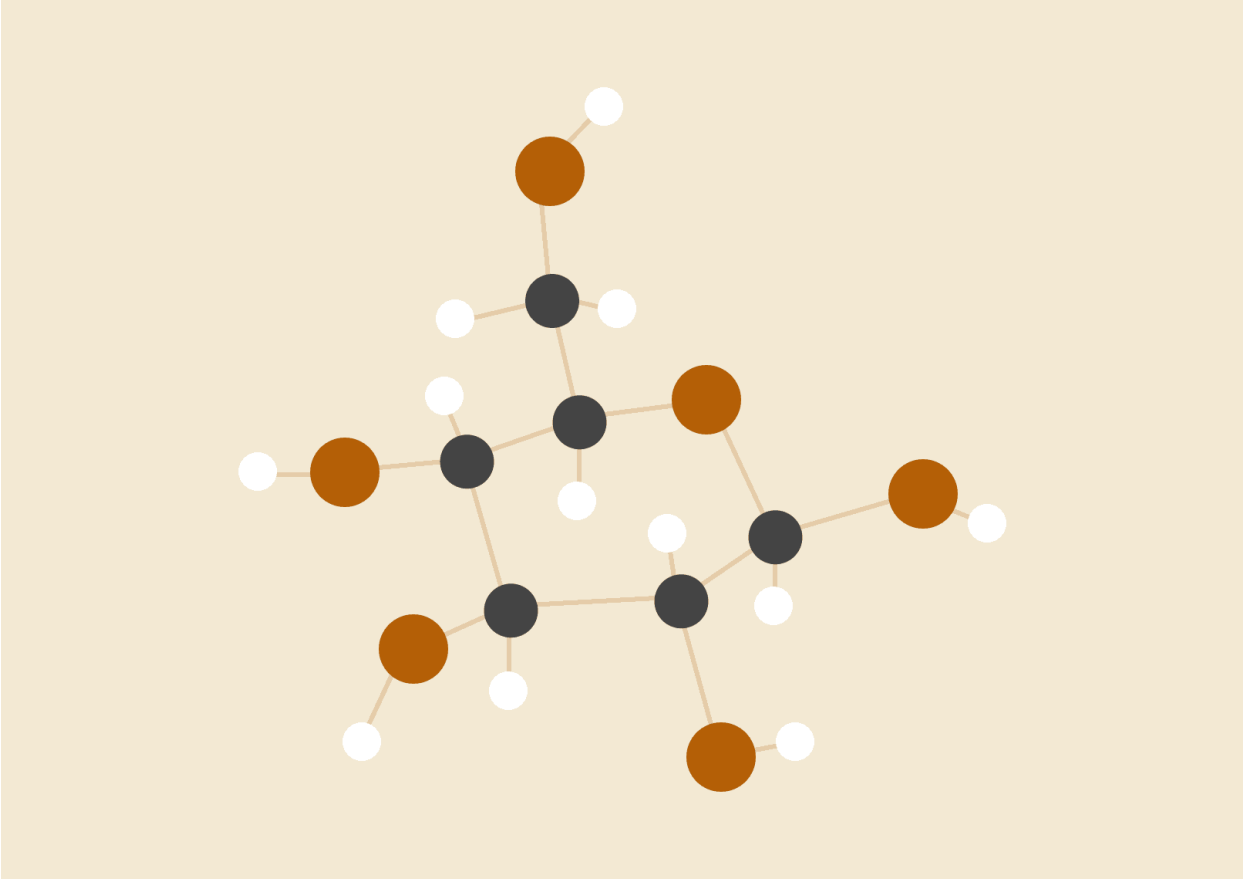
Deep Learning

Project



Master Degree in Artificial Intelligence

A.A. 2023/2024

Introduction

This document defines the project specifications for the Deep Learning course. You will test your skills on a real world problem by using models and tools seen during the course.

You will **participate to a Kaggle challenge** on the following task:

- Detect AI Generated Text

Each group will work on this task. It is an ongoing Kaggle challenges so you can have the opportunity to face real-life problems by using the very last published open datasets.

General Rules

- Groups of **2-3 students**
- Each group **must participate** in the Kaggle challenge
- Each group should submit:
    1. Report with a detailed description of the solution (.pdf) (See the structure at the last page of this document)
    2. Presentation of the solution (.pdf)
    3. Code of the solution (.py or .ipynb) [python, notebook or colab files]
- At the **exam** you will be asked to present your solution in **max 10 minutes**:
    - Each member of the group has to **equally** contribute
    - At the end of the presentation you will run some inference on the test set (SAVE YOUR MODELS - Just prepare a notebook cell which load your model and some test data)
- **Submit** via email to carlo.adornetto@unical.it (**10 days before the exam**)
    - OBJECT: DEEP LEARNING - Project
    - attach to the email a zipped file containing all the required documents/code
    - the zip will be named such as:
        - [surname1]_[surnamer2]_[surname3].zip

General Specification

1. Data:
    a. The download and the preprocessing can be designed with arbitrary

chosen libraries

2. Design a solution:
    a. Exploit the different NN architecture seen during the course (CNNs, RNNs, AEs, etc.)
    b. Generate insight into the training procedure (eg. loss plot, metric plot, hyperparameters search representation, etc.)
3. Evaluate the solutions:
    a. evaluate the designed solution on the test set
    b. Produce insights on the performances (e.g. confusion matrices, confidence intervals, errors distribution, AUROC, etc.)

HINTS:

- Use **Hyperparameter tuning** and **cross validation**, when feasible in terms of computation.
- Try to leverage the **GPU using Colab or Kaggle** (it is limited)
- Explain all your choice and argument the results (e.g. how the loss behaves during training, etc.)

NOTE: Even if your model reaches 99% of accuracy, we cannot say that it is a good model!

DATASETS

Text Classification

LLM - Detect AI Generated Text Link Kaggle

Objective: Predict whether the essays are human-written or machine-generated

HINTS:

- You can start from pretrained models but your code must include a training phase
- Go beyond a simple/trivial solution, try to use **different Neural Architectures**
- Use the Kaggle Notebook (not mandatory)

Report Structure

1. Introduction (Brief)
2. Detect AI Generated Text
    a. Dataset Description (Brief)
    b. Methods (Detailed)
        i. Preprocessing
        ii. Architectures
        iii. Training and Experiments
3. Results
    a. Final Results
    b. Proof of Challenge Participation*
4. Conclusions (Brief)

Brief : 10-20 lines
MAX 15 pages in total

Tables and Figures are really useful. Use them followed by comments to ease the comprehension of the report.

*Screen of the final ranking or link to the Kaggle project