University of Rome, La Sapienza. Artificial Intelligence and Robotics.
Deep Learning Course
A.Y. 2021/22

---

In the following, we introduce four different projects coming from popular tasks. *Pick one*!
**Rules**:

1. Each project should be carried out in groups of 3 people, and you should use Google Colab (`https://colab.research.google.com/`) for developing your solutions.
2. You are allowed to get "*inspired*" from code fragments that you find online, but solutions should be your own. Each time you use ideas and pieces of code that are available online, though, you ***must*** report the URL of the source of inspiration in the comments to the code.
3. The majority of the code you develop should be your own and should not be copied from code written by someone else. In particular, you are not allowed to copy entire repositories. We will check for it. Shuffling functions and/or changing names of the identifiers is considered copying, so... do not do it! **A violation of this rule will make you fail the exam.**
4. The code must run and produce results that can be interpreted. It is not sufficient to report on training/validation/test errors; you must use some metrics related to the task goals.
5. The code should be written to run on a GPU. Efficiency is not evaluated, but a complete notebook must run within colab's runtime limitations.
6. Each design decision you take must be clearly stated and documented. Each interesting alternative you have considered during the project's development should be reported, as well.
7. You can deliver the project at any time. The discussion date is agreed with the teachers and must be done by the three members of the team and can be either in person or remote. Your mark will be finalized at each predetermined exam date.

**Evaluation**:

- **Code Completion**. 18 pts. will be scored if the project runs on colab and terminates correctly, producing results with comparisons to, at least, a baseline.
- **Scientific soundness I** +4 pts. There are at least two baselines implemented, of which one is non-trivial. Results are computed in a scientifically sound way (e.g., training is not done on test set.)
- **Project Quality** +4 pts. the project: is written in a modular way and adheres to PEP8. It is self-explanatory, and contains plots and tables for all the results.
- **Scientific soundness II** +4 pts. The implemented technique is considered state of the art for the task. You should motivate why you claim it is SOTA.
- **Scientific excellence** Cum Laude. The technique implemented is novel, and it has never been presented in the literature. The current version ends up in the top 5 scoring mechanisms of the main leader board for that task. Really novel solutions will also be considered for preparing a scientific article.

**Image super resolution** *refers to the task of enhancing the resolution of an image from low-resolution (LR) to high (HR). It is popularly used in the following applications:*

- *Surveillance: to detect, identify, and perform facial recognition on low-resolution images obtained from security cameras.*
- *Medical: capturing high-resolution MRI images can be tricky when it comes to scan time, spatial coverage, and signal-to-noise ratio (SNR). Super resolution helps resolve this by generating high-resolution MRI from otherwise low-resolution MRI images.*
- *Media: super resolution can be used to reduce server costs, as media can be sent at a lower resolution and upscaled on the fly. Deep learning techniques have been fairly successful in solving the problem of image and video super-resolution. In this article we will discuss the theory involved, various techniques used, loss functions, metrics, and relevant datasets.*

**Dataset**. Set5 - Available at. `http://people.rennes.inria.fr/Aline.Roumy/results/SR_BMVC12.html`. The Set5 dataset is a dataset consisting of 5 images ("baby", "bird", "butterfly", "head", "woman") commonly used for testing performance of Image Super-Resolution models.

**Metrics**. Peak signal-to-noise ratio (PSNR) is an expression for the ratio between the maximum possible value (power) of a signal and the power of distorting noise that affects the quality of its representation. Let us assume we are dealing with a standard 2D array of data or matrix. The dimensions of the correct image matrix and the dimensions of the degraded image matrix must be identical. The mathematical representation of the PSNR is as follows:

$$\text{PSNR} = 10 \log_{10} \left( \frac{\text{MAX}_i^2}{MSE} \right) \tag{1}$$

where $MAX_i$ is the maximum possible pixel value of the image. When the pixels are represented using 8 bits per sample, this is 255.

MSE is the usual Mean Square Error that is defined per-pixel and per-channel. Given two $m \times n$ RGB images $x$, and $y$, we can compute MSE as:

$$\text{MSE} = \frac{1}{3mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{c=1}^{3} \|x_{i,j,c} - y_{i,j,c}\|^2 \tag{2}$$

**Visual QA** *is a semantic task that aims to answer questions based on an image.*

**Dataset**. Visual Q&A v2.0 - Available at `https://visualqa.org/download.html`. Visual Question Answering (VQA) v2.0 is a dataset containing open-ended questions about images. These questions require an understanding of vision, language and commonsense knowledge to answer. It is the second version of the VQA dataset.

1. 265,016 images (COCO and abstract scenes)
2. At least 3 questions (5.4 questions on average) per image
3. 10 ground truth answers per question
4. 3 plausible (but likely incorrect) answers per question

**Metrics**. You should use the evaluation metric that is described at `https://visualqa.org/evaluation.html`, which is robust to inter-human variability in phrasing the answers:

$$\text{Acc}\left(ans\right) = \min\left\{\frac{\text{number of humans that said } ans}{3}, 1\right\}$$

In order to be consistent with 'human accuracies', machine accuracies are averaged over all 10 choose 9 sets of human annotators.

Before evaluating machine generated answers, do the following processing:

- Making all characters lowercase
- Removing periods except if it occurs as decimal
- Converting number words to digits
- Removing articles (a, an, the)
- Adding apostrophe if a contraction is missing it (e.g., convert "dont" to "don't")
- Replacing all punctuation (except apostrophe and colon) with a space character. We do not remove apostrophe because it can incorrectly change possessives to plural, e.g., "girl's" to "girls" and colons because they often refer to time, e.g., 2:50 pm. In case of comma, no space is inserted if it occurs between digits, e.g., convert 100,978 to 100978. (This processing step is done for ground truth answers as well.)

A demo script of the evaluation code is available here: `https://github.com/GT-Vision-Lab/VQA/blob/master/PythonEvaluationTools/vqaEvalDemo.py`.

**Detection of Persuasive Techniques in Texts and Images** *is the task of detecting if a meme tries to persuade a person into believing some propagandist claim. Such propaganda campaigns are often carried out using posts spread on social media, with the aim to reach very large audience. The task is divided into three subtaks:*

1. *Given the textual content of a meme, identify the techniques used in it (multilabel classification problem).*
2. *Given the textual content of a meme, identify the techniques in it together with the span(s) of text in which each propaganda techniques appear*
3. *Given a meme, identify all techniques used in the meme, including the text and the visual content (multimodal task). This is a multilabel classification problem.*

**Dataset**. SemEval-2021 task 6 - Available at.

`https://github.com/di-dimitrov/SEMEVAL-2021-task6-corpus`. The corpus is hosted on the shared task github page. Beware that the content of some memes might be considered offensive or too strong by some viewers. Note that, for subtask 1 and subtask 2, you are free to use the annotations of the PTC corpus (more than 20,000 sentences). The domain of that corpus is news articles, but the annotations are made using the same guidelines, altough fewer techniques were considered. The dataset provides a training set, a development set, and a test set.

**Metrics**. We will use micro and macro F1 score for the three tasks.

**Trajectory Prediction** *is the problem of predicting the short-term (1-3 seconds) and long-term (3-5 seconds) spatial coordinates of various road-agents such as cars, buses, pedestrians, rickshaws, and animals, etc. These road-agents have different dynamic behaviors that may correspond to aggressive or conservative driving styles.*

**Dataset**. nuScenes - Available at. `https://www.nuscenes.org/nuscenes`. The nuScenes dataset is a large-scale autonomous driving dataset. The dataset has 3D bounding boxes for 1000 scenes collected in Boston and Singapore. Each scene is 20 seconds long and annotated at 2Hz. This results in a total of 28130 samples for training, 6019 samples for validation and 6008 samples for testing. The dataset has the full autonomous vehicle data suite: 32-beam LiDAR, 6 cameras and radars with complete 360° coverage. The 3D object detection challenge evaluates the performance on 10 classes: cars, trucks, buses, trailers, construction vehicles, pedestrians, motorcycles, bicycles, traffic cones and barriers.

**Metrics**. You should report on the following metrics:

- **Minimum Average Displacement Error over k (minADE_k)**. The average of pointwise L2 distances between the predicted trajectory and ground truth over the k most likely predictions.
- **Minimum Final Displacement Error over k (minFDE_k)**. The final displacement error (FDE) is the L2 distance between the final points of the prediction and ground truth. We take the minimum FDE over the k most likely predictions and average over all agents.
- **Miss Rate At 2 meters over k (MissRate_2_k)**. If the maximum pointwise L2 distance between the prediction and ground truth is greater than 2 meters, we define the prediction as a miss. For each agent, we take the k most likely predictions and evaluate if any are misses. The MissRate_2_k is the proportion of misses over all agents.