Machine Learning Projects

EME 2021

Introduction

For each of the following projects, you should get exposed to a full machine learning problem. You will be required to download, analyse data, preprocess it and operate a model to predict a specific target. You will also test the model and tune the parameters to get the maximum value of it.

You may be required to implement an approach to solve a problem or a game with the maximum score possible.

There are general notes/rules while doing the final project:

1- This is a training course, try to do every step yourself and maximize your knowledge set and coding skills. DO NOT copy/paste a ready made solution, this is highly penalized. However, it's completely ok to copy/paste lines of codes to make a specific step like numpy operations, pandas dataframes, keras/sklearn models .. etc.

2- You are required to do a small and quick presentation explaining your work. Try to describe your code strength points and future work that you may want to make on it if there is more time. You will be asked in depth about your work, so again do not idiotically copy/paste internet/your-friend solutions.

3- Some projects are not allowed to be in teams, each individual is required to deliver his own work. Other projects may be allowed to be solved in teams, in this case each person of the team will be asked in depth the solution the team made to make sure that all of the team understands the overall solution.

4- **I encourage you** to collaborate and help each other and share knowledge together.

Project #01

San Francisco Crime Classification

Link: https://www.kaggle.com/c/sf-crime

1- Team Size: 1 “Single”

2- Problem definition

From 1934 to 1963, San Francisco was infamous for housing some of the world's most notorious criminals on the inescapable island of Alcatraz.
You are given a dataset for criminals' records and you are required to predict the **“Category”** of the crime.

3- Steps to solve

A. Load the data and understand the features and the target.
B. Visualize the data using plots like in Seaborn/Matplotlib.
C. Clean the data.
D. Reduce the dimensions of the data. [you can use any of the approaches]
E. Create 3 models, train/test them.
F. Parameters tuning to the model.
G. Compare between the models and state which is better and why.

You can choose any 3 models you want to use, however here are some examples you can choose from:
SVM - Decision Tree - Bayes Classifier - Random Forest - XGBoost - KNN .. etc.

4- Required Deliveries

You will deliver the code notebook, report and do a 5-mins presentation to explain your work.
The report should contain the following:

A. Your name and email.
B. Project Name.
C. Main charts/plots explained.

D. Data cleansing and features engineering steps.
E. Results of the models, mentioning which is better and why.
F. Future work that may be made to enhance the models.

Project #02

Happy-Unhappy

Startar Code Link:
https://colab.research.google.com/drive/1NUJS0TVZug_xBVhxxRN1YkOeGZXScsEh?usp=sharing

1- Team Size: 1 “Single”

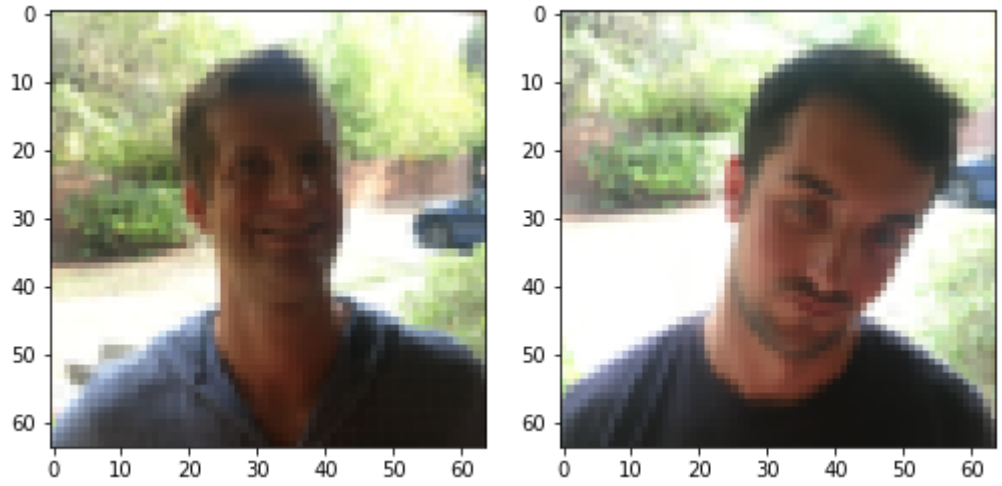
2- Problem definition

For your next vacation, you decided to spend a week with five of your friends from school. It is a very convenient house with many things to do nearby. But the most important benefit is that everybody has commited to be happy when they are in the house. So anyone wanting to enter the house must prove their current state of happiness.

As a deep learning expert, to make sure the ”Happy” rule is strictly applied, you are going to build an algorithm that uses pictures from the front door camera to check if the person is happy or not. The door should open only if the person is happy.

You have gathered pictures of your friends and yourself, taken by the front-door camera. The dataset is labeled.

Your target is to detect if someone is happy or not.



3- Steps to solve

A. Load the dataset
B. Visualise an image to see an example

C. Make some data augmentation, aka generate new data by rotating/flipping images. Read more here.
D. Build a CNN network, try different configurations: optimizers/regularizations, you should add dropout layers.
E. Train/Test the network
F. Evaluate the results.

4- Required Deliveries

You will deliver the code notebook, report and do a 5-mins presentation to explain your work.

The report should contain the following:

A. Your name and email.
B. Project Name.
C. Main charts/plots explained.
D. Data Augmentation steps.
E. Model Architecture.
F. Results of the models, mentioning parameters used.
G. Future work that may be made to enhance the models.

Project #03

Frozen Lake game

Game definition Link: https://gym.openai.com/envs/FrozenLake8x8-v0/

1- Team Size: 2-3 “Multiple persons”

2- Problem definition

Winter is here. You and your friends were tossing around a frisbee at the park when you made a wild throw that left the frisbee out in the middle of the lake. The water is mostly frozen, but there are a few holes where the ice has melted. If you step into one of those holes, you'll fall into the freezing water. At this time, there's an international frisbee shortage, so it's absolutely imperative that you navigate across the lake and retrieve the disc. However, the ice is slippery, so you won't always move in the direction you intend.

You are required to make an agent that is able to beat the game.
The agent controls the movement of a character in a grid world. Some tiles of the grid are walkable, and others lead to the agent falling into the water. Additionally, the movement direction of the agent is uncertain and only partially depends on the chosen direction. The agent is rewarded for finding a walkable path to a goal tile.

You can get the game environment from the gym Open Ai library in python. Also see this tutorial if you want to know how to get started with gym OpenAi library.

Please note that each person in the team must work and will be discussed individually during the presentation.

3- Steps to solve

A. Understand how GYM works.
B. Load the game environment and make sure that it’s running.

C. Create your agent learning function “Monte carlo - Q Learning .. etc”.
D. Run the agent in the environment and make the learning step.
E. Create a demo for the agent playing the game after learning, try to compare between the agent performance in different epochs numbers.

You can choose any learning function, monte carlo, q learning, SARSA, deep Q networks .. etc.

4- Required Deliveries

You will deliver the code notebook, report and do a 5-mins presentation to explain your work.

The report should contain the following:

A. Your names and emails.
B. Project Name.
C. Main game steps images.
D. Environment explained.
E. Future work that may be made to enhance the agent.

Project #04

Chest Xray Abnormalities detection

Competition Link:
https://www.kaggle.com/c/vinbigdata-chest-xray-abnormalities-detection/overview

1- Team Size: 2-3 “Multiple persons”

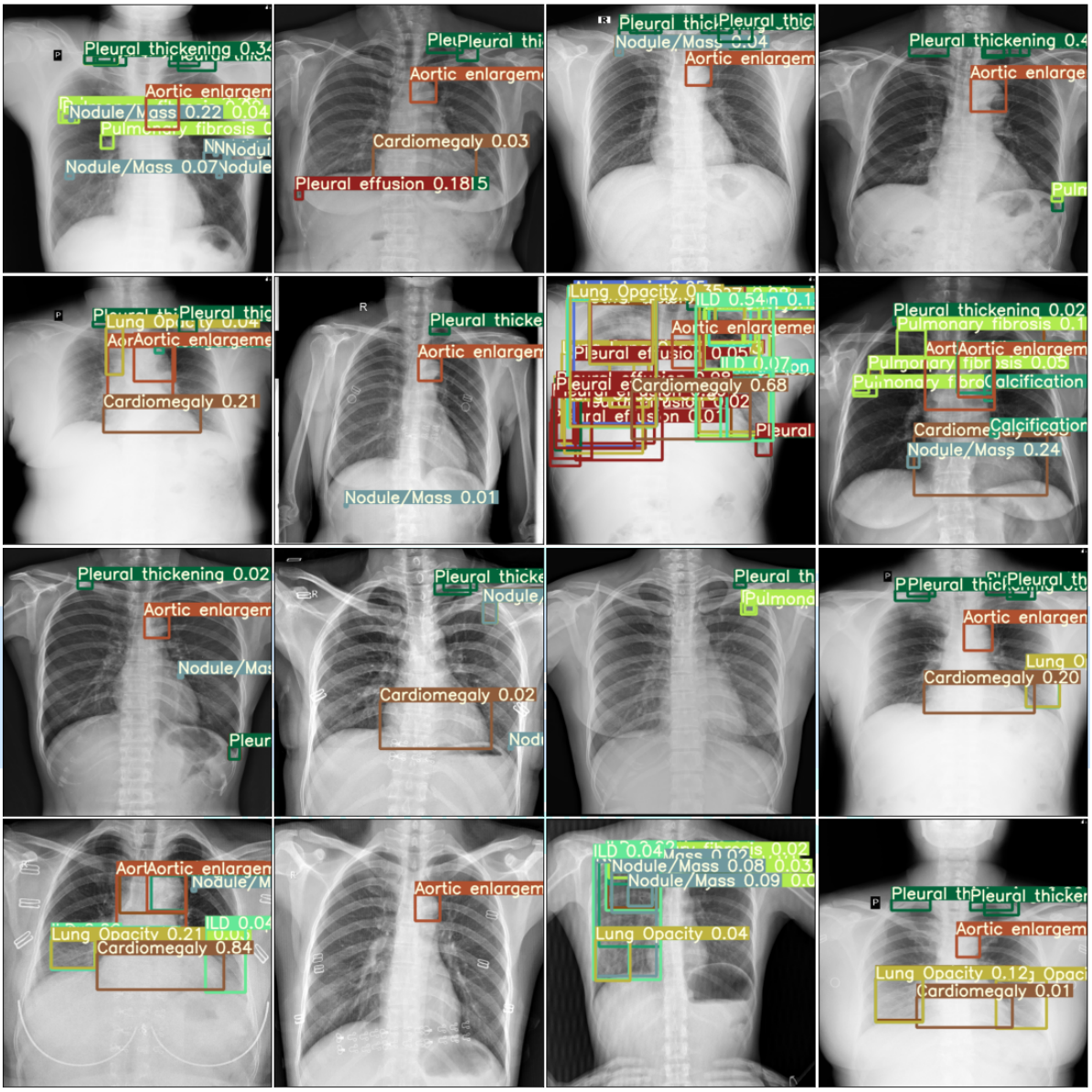
2- Problem definition

In this competition, you’ll automatically localize and classify 14 types of thoracic abnormalities from chest radiographs. You'll work with a dataset consisting of 18,000 scans that have been annotated by experienced radiologists. You can train your model with 15,000 independently-labeled images and will be evaluated on a test set of 3,000 images. These annotations were collected via VinBigData's web-based platform, VinLab. Details on building the dataset can be found in our recent paper “VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations”.

If successful, you'll help build what could be a valuable second opinion for radiologists. An automated system that could accurately identify and localize findings on chest radiographs would relieve the stress of busy doctors while also providing patients with a more accurate diagnosis.

Please note that each person in the team must work and will be discussed individually during the presentation.
This contest you will work from scratch and you may be not exposed to such dataset format before, so I encourage you to see some code notebooks in the competition before starting to see how to load the data and display it. Also note that this type of competitions may need large amount of resources if you are going to run a complex model on it, so I want to say that it’s ok to get back bad results, all you should do is to be able to load the data, create a model and

run it on it, and try to enhance it as much as possible with the kaggle notebook resources.



3- Steps to solve

A. Load the data.
B. Display some training samples as shown above.
C. Identify the inputs/target.
D. Create your Model.
E. Train the model and tune the parameters as possible.
F. Evaluate your model on the test dataset and try to submit the results on kaggle.

Note: submitting the results on kaggle will be considered a bonus step for the team.

4- Required Deliveries

You will deliver the code notebook, report and do a 5-mins presentation to explain your work.

The report should contain the following:

A. Your names and emails.
B. Project Name.
C. Main charts/plots and data samples.
D. Model architecture.
E. Future work that may be made to enhance the model.

Project #05

English to French Translation

Dataset Link: https://www.kaggle.com/dhruvildave/en-fr-translation-dataset

1- Team Size: 1-2 “Multiple persons”

2- Problem definition

Given an English to French dataset, you are required to create an RNN model to learn how to translate from english to french.
To make such a process you will need to tokenize the words and clear the data from punctuations first. You can read this blog for such steps.
Note: This dataset is very large, try to take part of it [100K samples only] and use them in this project.

3- Steps to solve

A. Understand the dataset.
B. Load the dataset.
C. Clean and tokenize the sentences.
D. Create the RNN model.
E. Train and evaluate the results.

Bonus: Try to use transfer learning to create a similar model to convert from French to English.

4- Required Deliveries

You will deliver the code notebook, report and do a 5-mins presentation to explain your work.
The report should contain the following:

A. Your names and emails.
B. Project Name.
C. Data cleansing steps.
D. Model Architecture.

E. Results
F. Future work that may be made to enhance the agent.

6- Prefered Project:

You can choose any project you want to work on. In this case you are required to submit a proposal for this project that contains the following points:
1- Team members [Not allowed more than 3 persons and the difficulty of the task should be related to the number of the persons working on the project]
2- Project Description
3- Project Milestones
4- Proposed solution steps.
Then your proposal will be reviewed and either accepted or rejected according to the project scope, size, difficulty and goals.