BMI6331 - Data Challenge

This assignment will count towards your grade

Summary

In this assignment, you will tackle a data challenge composed of two tasks, an image classification task, and an image segmentation task. You will have to design and train a pattern recognition algorithm to identify cardiomegaly in chest X-rays images and another algorithm to segment the hippocampus from MRI data. You will need to write a report in the form of a short IEEE academic paper, package the code to replicate your results and present your work in front of the class.

Task 1: Image Classification

The chest X-ray is one of the most commonly accessible radiological examinations for screening and diagnosis of many lung diseases. A tremendous number of X-ray imaging studies accompanied by radiological reports are accumulated and stored in many modern hospitals' Picture Archiving and Communication Systems (PACS). In September 2017, NIH released the largest chest x-ray dataset for scientific research as of now.
https://www.nih.gov/news-events/news-releases/nih-clinical-center-provides-one-largest-publicly-available-chest-x-ray-datasets-scientific-community

For this task, you will be given a subset of the full NIH dataset. This subset contains a single image per patient. Only healthy patients or patients with cardiomegaly have been selected for a total of 21,966 images (767 with cardiomegaly and 21,966 healthy samples). The labelling was automatically computed with a natural language processing algorithm, which has a high F1 score to be considered reliable but it is not perfect. More details are available in Wang et al, 2017.
http://openaccess.thecvf.com/content_cvpr_2017/papers/Wang_ChestX-ray8_Hospital-Scale_Chest_CVPR_2017_paper.pdf

The dataset has been pre-split in training and testing set. The testing set has been balanced to contain 300 samples with Cardiomegaly and 300 with healthy samples. This will facilitate the evaluation of your algorithm. However, keep in mind that the classes in your training set will be highly unbalanced. <u>In order to get a validation set, you will need to split the training set NOT the testing set.</u>

Task 1: Data Loading

- Download all the rar data files and decompress them. You can use:
  - Mac: https://theunarchiver.com/
  - Windows: http://www.7-zip.org/
  - Linux: Install the unrar package with your distribution and type `unrar x filename.rar`

- dataInfo.csv contains the labels. You can load it as follows:
  import pandas as pd
  gtFr = pd.read_csv('dataInfo.csv', index_col='imgID')

Task 1: Data Dictionary

- These are the main variables that you will need
  - 'FileName': file name without path
  - 'FullFileName': file name with full path
  - 'Cardiomegaly': 1 for cardiomegaly, 0 for healthy
  - 'Train': 1 for image belonging to training set, 0 for image belonging to test set
- Additionally, the following information is provided:
  - 'Findings': list of findings in addition to cardiomegaly
  - 'PatAge': age
  - 'PatGender: gender
  - 'OrigWidth': width of the image before being resized
  - 'OrigHeight': height of the image before being resized
  - 'OrigPixSpacingX': original pixel spacing in the X axis
  - 'OrigPixSpacingY': original pixel spacing in the Y axis

Task 2: Hippocampus Segmentation

Hippocampus area and shape are important variables for studies in neurodegenerative or psychiatric diseases. As such, an automated approach to segment the anterior and posterior hippocampus is essential for large studies. In this task you will have to develop an algorithm to segment the anterior and posterior hippocampus from T1 MRI images. Then, you will need to evaluate the performance

The data set consists of MRI images acquired from healthy adults and adults with a non-affective psychotic disorder. T1-weighted MPRAGE was used as the imaging sequence. The corresponding target ROIs were the anterior and posterior of the hippocampus, defined as the hippocampus proper and parts of the subiculum. The data was acquired at the Vanderbilt University Medical Center, Nashville, US.

A paper describing (among others) the dataset used and segmentation techniques employed is available here (and on Canvas): https://www.nature.com/articles/s41467-022-30695-9

Task 2: Data Loading

Download the zip files from Canvas. Also, you will find an example Jupyter notebook that shows you how to: load the dataset; pre-process it in a way that is conducive to use methods requiring mini-batches; save tensor as Nifti images.

Task 2: Data Dictionary

The data is stored as follows:

- imagesTr: directory containing the training set
- imagesTe: directory containing the testing set

- labels: directory containing the ground truth labels. 0: background, 1: anterior, 2: posterior
- dataset-Hippocampus-BMI6331.json: metadata describing the dataset and listing all the MRI images used and corresponding images with ground truth labels.

The dataset has been pre-split in training and testing set. In order to get a validation set, you will need to split the training set NOT the testing set. All the images will have the same voxel size of 1mm x 1mm x 1mm, however each volume has different number of voxels.

Task 2: Suggestions

If you need to visualize 3D data in the Nifti files use software like: Slicer 3D, ITK-Snap, MRCron or FSLEyes. Viewing them inside a Jupyter notebook for anything more than a quick sanity check is not a good idea. If you need to visualize a 3D tensor, save it as a Nifti file and open it with such programs.

It is highly suggested to evaluate the segmentation performance using an appropriate segmentation metric, like DICE score for anterior and posterior hippocampus. It does not make much sense to compute the DICE score for the background. It is also a very good idea to compute the DICE score for each MRI volume then average them, rather than a single DICE score for a big tensor containing all the results. The latter will lead to give more weight to larger hippocampi.

It is suggested to pad the images to a common number of voxels to simplify the processing.

Report

A short report on the approach used to tackle the data challenge. There is no minimum or maximum length, however the report will have to include the following sections:

- Introduction
- Methods
- Results
- Discussion and Conclusion

Additionally, you will need to submit the code used to generate all the results in the report. A single submission is required per each group. You will need to use the IEEE double column template as shown in here: https://www.overleaf.com/latex/templates/preparation-of-papers-for-ieee-sponsored-conferences-and-symposia/zfnqfzzzxghk

Final Notes

You will be expected to complete both tasks and evaluate the results on the test set provided. Also, this is a medical image computing challenge, so the absolute performance is not everything, other important aspects to consider is to provide some level of interpretability (especially for task 1) and insight of potential reasons why your algorithm is failing.

You are free to use external resources, such as Google Colab or other servers, however, hardware/software problems will not be considered extenuating circumstances, and external platforms that you are not familiar with are more likely to run into problems.

Submission

- Submit the PDF of the report
- Notebook and code to replicate your results.
- Presentation slides