MESIIN481222 Advanced Machine Learning
Lab Assignment 3 – Convolutional Neural Network in Healthcare
Graded assignment (Continuous assessment)

You are required to submit your solutions of this assignment on DeVinci Online, with respect to the deadline set by your instructor. Submit your own work. Cheating will not be tolerated and will be penalized.

Pre-Requirements: Lab activity 6 & 7

Before you solve this lab assignment, make sure you finished the lab activities 6 and 7.

Convolutional Neural Network in Healthcare

Imaging is the fasting growing sources of medical data in the form of X-Ray, CAT Scan, MRI, Mammogram etc. The medical community realized that imaging can sometimes reveal a disease even before the symptoms are shown. For this reason, there are screenings as a preventive check such as Pap smears, Mammograms, Colonoscopy etc. For certain diseases like Cancer, early diagnosis increases the chances of survival and healthy life style and also reduces treatment costs.

Medical image understanding is generally performed by skilled medical professionals. However, the scarce availability of human experts and the fatigue and rough estimate procedures involved with them limit the effectiveness of image understanding performed by skilled medical professionals. The challenge here is that a trained human eye of a medical professional cannot identify subtle signs on these tests.

Machine learning (specifically Convolutional Neural Networks, CNN) has proven to be a powerful aid in this context. CNNs are effective tools for image understanding. They have outperformed human experts in many image understanding tasks.

The algorithm is well suited for a multi-class classification problem (predicting picture as a car or a bicycle or a van) or binary classification (for example — predicting whether a medical image has a malignant tumor or not).

References:
Sarvamangala, D.R., Kulkarni, R.V. Convolutional neural networks in medical image understanding: a survey. Evol. Intel. 15, 1–22 (2022). https://doi.org/10.1007/s12065-020-00540-3

Deepakvraghavan, Practical uses of Deep Neural Networks in Healthcare. May 3, 2018. https://deepakvraghavan.medium.com/practical-uses-of-deep-neural-networks-in-healthcare-934ce1783999

Assignment : Medical / Healthcare Case Study using CNN

You are asked to identify a case study of your choice related to medical or healthcare matter.

After your decide which case study is of interest to you, you select the dataset that is the most convenient for you to conduct your study. There are many medical and healthcare datasets available online with historical data about patients, diseases, treatments, diagnosis, CT scanned imaging data, X-rays images …

Here are two examples of interesting case studies done by data scientists and published online.

- A study that helps in the prediction of the age of a patient from CT scanned images using CNN, can be found on https://www.kaggle.com/code/csedsa20bd6701/cnn-ct-scan-20bd1a6701/notebook.

  The CT images dataset from cancer imaging archive with contrast and patient age that is used can be found on https://www.kaggle.com/datasets/kmader/siim-medical-images/download?datasetVersionNumber=6. The dataset is designed to allow for different methods to be tested for examining the trends in CT image data associated with using contrast and patient age.

  The basic idea is to identify image textures, statistical patterns and features correlating strongly with these traits and possibly build simple tools for automatically classifying these images when they have been misclassified (or finding outliers which could be suspicious cases, bad measurements, or poorly calibrated machines). The data are a tiny subset of images from the cancer imaging archive. They consist of the middle slice of all CT images taken where valid age, modality, and contrast tags could be found. This results in 475 series from 69 different patients.

- Another very interesting case study that make disease predictions on chest X-rays to detect Pneumonia with CNN can be found on https://www.kaggle.com/code/homayoonkhadivi/medical-diagnosis-with-cnn-transfer-learning.

  The Chest X-ray images taken from the public ChestX-ray8 dataset that is used can be found on https://www.kaggle.com/code/homayoonkhadivi/medical-diagnosis-with-cnn-transfer-learning/data.

  X-rays images are critical for the detection of lung cancer, pneumenia ...

  The normal chest X-ray (left panel) depicts clear lungs without any areas of abnormal opacification in the image. Bacterial pneumonia (middle) typically exhibits a focal lobar consolidation, in this case in the right upper lobe (white arrows), whereas viral pneumonia (right) manifests with a more diffuse ''interstitial'' pattern in both lungs.

The dataset is organized into 3 folders (train, test, val) and contains subfolders for each image category (Pneumonia/Normal). There are 5,863 X-Ray images (JPEG) and 2 categories (Pneumonia/Normal).

Chest X-ray images (anterior-posterior) were selected from retrospective cohorts of pediatric patients of one to five years old from Guangzhou Women and Children's Medical Center, Guangzhou. All chest X-ray imaging was performed as part of patients' routine clinical care.

For the analysis of chest x-ray images, all chest radiographs were initially screened for quality control by removing all low quality or unreadable scans. The diagnoses for the images were then graded by two expert physicians before being cleared for training the AI system. In order to account for any grading errors, the evaluation set was also checked by a third expert.

Many medical and clinical datasets can be found online and can be downloaded for free. Here are some references:
https://www.kaggle.com/datasets?search=health+medical
https://data.world/datasets/health
https://paperswithcode.com/datasets?mod=medical

Other references for health datasets can be found in https://odsc.medium.com/15-open-datasets-for-healthcare-830b19980d9

Attention: You are not limited to the references that are given in this document. You may choose the dataset and case study that you want. The examples above are given to inspire you.

The deliverable should include : A Python Notebook with your source code and inline outputs. NB: Do not forget to add internal comments for interpretation and explanation in each step. This is mandatory and will be graded.

Your work should follow the steps below:
1- Load your data (csv) in a dataframe
2- Explore your data
3- Print some statistical descriptive information about your data
4- Clean you data if needed
5- Transform your categorical variables if needed
6- You may include some visualization for further data exploration
7- Do not forget to do data augmentation if needed
8- Prepare your dataset for training and testing
9- Create your CNN model by following the steps of convolution, pooling, flattening and full connection.
10- Fit your model on the train dataset

11- Make predictions on the test dataset
12- Evaluate the performance of your model and give some insights

Deadline to submit your final work on Online Devinci is: on Sunday 19th of February 2023 at 23:59. One submission per group is enough.

A quick demonstration of the code per group will be done in class on Monday 20th of February 2023.

Good luck!