SCHOOL *of* ELECTRICAL ENGINEERING & COMPUTING
FACULTY *of* ENGINEERING & BUILT ENVIRONMENT
The UNIVERSITY *of* NEWCASTLE

COMP3330/COMP6380 Machine Intelligence, Semester 1, 2020

Advanced Machine Learning Project
Group Project 1B + Individual Report 2B

Maximum possible marks: 25 = 10 (1B: group project part) + 15 (2B: individual paper)

Deadline: Week 13 (check date and time and submit via blackboard)

Description

In this assignment we want to challenge ourselves by finding solutions to machine learning tasks of current interest. There are two components for marking this assignment. Both components have to be submitted separately in Blackboard:

Project 1B [10 marks]. The group's experimental work should be summarised in a brief group summary report that provides an overview what your team has done in the experimental component of your project. It should explain how the software and data is structured and detail your group's individual member contributions to the project outcome. Marked will be the quality, volume and depth of the experimental results achieved by the group. If the member contributions are very different the marks will be weighted accordingly. The recommended length of the group summary report is about 4 pages.This is just a guideline. Include in your submission the summary report and all essential files that are required for reproducing and verifying your results. It is sufficient if one team member submits. Make sure that all other group member names are listed and each member has signed (!) the agreed group report with individual member contribution statements.

Report 2B [15 marks]. Each student writes an individual paper that describes, analyses and critically discusses the project results achieved by the student and his/her team in detail. We expect about 6-9 pages from COMP3330 students, and about 9-14 pages from COMP6380 students. Aim at providing a high quality report that describes and discusses your results and approaches from part 1B clearly and concisely. Your individual report should be formatted in Springer LNCS format and be structured and written in the style of a conference paper. Any literature citations, e.g., in the background, method or discussion sections should follow a consistent citation style. The LNCS conference paper template is available e.g. at the following link: `https://www.springer.com/gp/computer-science/lncs/conference-proceedings-guidelines` Please submit your individual paper in pdf format in Blackboard.

Notes: Be prepared that training deep networks can require some time and usually a GPU. There are several deep learning frameworks and tools such as Tensorflow1 + Keras, Tensorflow2, Pytorch, Mxnet, etc.. Due to their fast development and ongoing changes we cannot provide any support for them and it is expected that you make your own selection and use any relevant support from the software or hardware providers. Most of the packages have online documentation and tutorials and an active user community. In our labs and instructions we used Tensorflow1 + Keras and this framework should be sufficient to solve the assignment.

Warning: You will find that some of the questions can lead into open-ended research and some of the experiments may take significant time on the computer. It is your responsibility to decide on a sensible balance of quality and depth of your investigation of each individual question so that the assignment can be completed within the given time. If computing resources become a severe issue provide pilot results as proof of concept, explain the issue and solution attempts and put more emphasis on the analysis and discussion.

Below you find three questions with three data sets. As a flexible guide: Q1 is aiming at COMP3330 students, while Q2 is aiming at COMP6380 students, and Q3 is for any advanced students who may seek a challenge. You can work on any one, two or three of Q1, Q2, Q3. Marks will be awarded based on quality, depth and volume of your study following the separately provided marking guides. I.e. if you solve one of the questions very well you can obtain full marks.

Q1 Flowers Recognition Data

The Flowers Recognition data set is available at: `https://www.kaggle.com/alxmamaev/flowers-recognition` This data set contains contains 4242 labeled images of flowers. Your task is to obtain a CNN based 5-class classifier that can input one of the images (or a cropped or downscaled version of the image) and output if it is chamomile, tulip, rose, sunflower, dandelion.

Try out different settings of hyperparameters of a CNN model on the dataset. Compare and discuss your models' performance using the training, validation and test datasets. It is recommended that you focus on *about two* different aspects that you investigate in more detail.

Q2 Chest X-Ray Images (Pneumonia) Classification

Perform an experimental study using the Chest X-Ray Images (Pneumonia) data that is available at `https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia`

The dataset is organised 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).

Task: Develop a Convolutional Neural Network (CNN) based 2-class classifier that can input one of the images (or a cropped or downscaled version of the image) and output if it is pneumonia or normal.

Try out different settings of hyperparameters of a CNN model on the dataset. Compare and discuss your models' performance using the training, validation and test datasets. It is recommended that you focus on *about two* different aspects that you investigate in more detail.

Q3 Novel Corona Virus 2019 Data

Perform an experimental study using some of the Novel Corona Virus 2019 data that is available at `https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset` Most of this data is time series data and we could do a predictive study. This can be done in different ways, e.g., using recurrent neural networks such as LSTMs or just a feed forward MLP where the input is a section (time window) of a time series and the output is a prediction of the future. It could be interesting to combine several of the available time series from different countries (e.g. testing rate, hospitalisation, active cases, ..) and predict one feature for an individual country.