CM502650: Machine Learning 2

Assignment 1: Age Estimation and Gender Classification

March, 2022

This coursework is allocated 100 marks. It comprises 20% of the overall unit grade.

1 Submission

Deliverable	Due date
Signing up a group in Moodle	**8pm, Feb. 21 (Mon.), 2022**
A Google Colab Notebook (.ipynb) A report (.pdf)	**8pm, Mar. 21 (Mon.), 2022**

NOTE that each group only need have one submission. DO NOT include your names anywhere in your code or in your report.

2 Your task

Humans can easily identify if a person is a female or a male with high accuracy, and also roughly tell his/her age. Now given a large labelled training dataset, the computer can also do the gender classification and age estimation. Your task is to build and train your deep learning models for both gender classification and age estimation on a given training dataset with GPU in Google Colab. I will then use a test set to check how accurate your models could achieve for both tasks.

3 Groups

You are encouraged to work in groups (3 to 4 students), because being a good team player is one of the transferable skills you will need in your future career. You are free to form your group on your own. Please sign up a group in the Moodle by the date specified in the above table. For those who don't sign up or only 2 members in a group, I will then form groups randomly. You must work in a group for this assignment unless you have a strong reason (such as ILP). If so, please email me before the group signing-up due date.

All members of a group are supposed to contribute equally, therefore will get the same mark. However, if there are credible, state the reasons, I will then differentiate the marking of group members.

4 Dataset

You will use a subset of the UTKFace dataset. A `train_val` folder has been created by choosing a subset (5,000 face images) from the UTKFace dataset for you to train and validate your model. It is a shared google drive folder below.

`https://drive.google.com/drive/folders/1UjYRDyo10Fx-Rv91CQl5ZfwF85HiLUX8?usp=sharing`

- Click the link above. It must be in the "shared with me" in your Google Drive.
- Right-click this folder, choose "Add shortcut to Drive", then specify the path where you would like to put it.
- Then you can access its path after you mount the drive.

UTKFace dataset is a large-scale face dataset with long age span (range from 0 to 116 years old). The dataset consists of over 20,000 face images with annotations of age, gender, and ethnicity. The images cover large variation in pose, facial expression, illumination, occlusion, resolution, etc. The labels of each image are embedded in the file name, formatted as `[age]_[gender]_[race]_[date&time].jpg`. You only need to use the first two labels for this assignment:

- **age** is an integer from 0 to 116, indicating the age.
- **gender** is either 0 (male) or 1 (female).

***NOTE**: We acknowledge that this dataset does not consider genders other than female and male, which might upset some people. We don't consider this moral issue simply because we only use it for a classification exercise for this course. However, you MUST consider potential moral issues when using any dataset in your future research or career.*

5 Notebook file to start with

Download the attached Google Colab Notebook file (`age_gender_submit.ipynb`) for this assignment. You must follow the instructions within and fill all your code in.

6 More details

This assignment has two parts:

6.1 Part 1 – Coding

Train two CNN models with GPU in Google Colab:

- One is defined by you with a few restrictions and be trained from scratch, save it as `age_gender_A.h5`
- The other is to fine-tune a pre-trained model, save it as `age_gender_B.h5`.

Full details can be found in the provided Google Colab Notebook file. Your models will be subject to testing on unseen test data.

6.2 Part 2 - Report

You will write a report (max 1200 words). In the report, please follow the following structure or similar.

- A **cover page** consists of the following:
  - A form to specify how much contribution each member made, based upon your agreed input. If each member took equal contribution, this would be all 100% (which is the general expectation for this assignment). More or less than 100% means higher or lower contribution for the project.

ID	Contribution
ID1	100%
ID2	100%
ID3	100%
ID4	100%

    Reasons for different weights [Optional] :
    Do all the members agree with the above contributions? [Yes/No]
  - **Specify two shared links** for the two models:
    `age_gender_A.h5` for the CNN model you defined and trained from scratch;
    `age_gender_B.h5` for the pre-trained CNN model fine-tuned on this dataset.
- **Section 1: Introduction** Give a brief description of what this assignment is about.
- **Section 2: My own CNN** Describe your own CNN architecture, the training process and the performance.
- **Section 3: Pre-trained CNN** Describe the pre-traind CNN architecture, the training process and the performance.
- **Section 4: Summary and Discussion** Compare the two models. What do you achieve in doing this assignment? Any other discussion you would like to have.

Both parts of this assignment are mandatory. Your code and report will be marked jointly. Only submitting Part 1 or Part 2 would result in zero mark.

7 Marking Criteria

Your report and your code will be assessed jointly. Detailed marking criteria are listed in the attached `ML2-CW1-mark-template.pdf`. During the marking process, the criteria could be slightly adjusted.

8 Late submissions

The university policy will be followed on late submissions. If a piece of work is submitted after the submission date, the maximum possible mark will be 40% of the full mark. If work is submitted more than five days after the submission date, student will receive zero marks. If you need an extension, please contact your Director of Studies.

9 Plagiarism

Do not plagiarise. Plagiarism is a serious academic offence. For details on what it is and how to avoid it, please visit

http://www.bath.ac.uk/library/help/infoguides/plagiarism.html

10 Any questions about the assignment

I **WILL NOT** reply any individual emails about coursework questions. Instead, all questions about this assignment will be answered in the **Moodle Discussion Forum (with a title starting with [CW1])**, then every one can benefit from the answers.