MLFA Spring 2024: Lab Test 2 (25 marks)

Implement a **convolutional neural network (CNN) based regression model** to predict age from photo of face images. You can use libraries to process and handle data. The dataset to be used is lab_test_2_dataset that could be downloaded as a zip file from the following link https://drive.google.com/file/d/1dKFVzwNDvBzblp7Eouk1_--qQ-g_8zjy/view?usp=sharing.

The dataset has folders with names denoting the age, which contain face images corresponding to the age denoted by the folder name. All the images are RGB, with a size of 32x32.

You may use Jupyter notebook on your local system or Google Colab to code your solution. If you choose to use Google Colab, you can upload the dataset on your Google drive and unzip it using the following code.

```
google.colab import drive
drive.mount('/content/gdrive')
!unzip gdrive/MyDrive/lab_test_2_dataset.zip
```

Experiments:

Experiment 1 (2 marks):

Make the implementation deterministic by setting the random states. Set all of torch.manual_seed, torch.cuda.manual_seed_all, np.random.seed and random.seed to a specified seed value. The seed value is your joining year. For example, if you got admission in 2021, then the seed value is 2021. Set the device to CPU.

Experiment 2 (4 marks):

Shuffle the dataset and then perform a train-validation-test split of 70%-15%-15% respectively. Create the corresponding dataloaders as well. Use appropriate batch size as per your judgement. Print the overall dataset size, training dataset size, validation dataset size and testing dataset size.

Experiment 3 (8 marks):

Write a class for CNN regression model to predict age from the input image. The CNN should be relatively simple enough to be run on CPU. Use 2 convolutional layers, and then fully connected layers of two hidden layers and one output layer (therefore 3 total fully connected layers). Decide the number of convolutional kernels, their sizes, strides, padding, number of neurons in the fully connected layers as per your judgement. Use appropriate regularization, loss function, optimizer and learning rate as per your judgement.

Experiment 4 (4 marks):

Train the model for 25 epochs using training dataset, and use validation dataset to evaluate the model while training. Use mean squared error (MSE) loss as the evaluation metric. Save the model at regular intervals (decide the intervals as per your judgement). Generate a plot to demonstrate the training and validation losses against the number of training epochs.

Experiment 5 (5 marks):

Evaluate the trained model on the testing dataset using MSE and scatter plot of predicted labels vs. ground truth labels of ages. (Do not worry if the results are not appealing, the marks would be awarded based on correct steps in the implementation, although relatively better results would score higher marks)

Submission:

A .zip file containing the python source code and a PDF report file. The final name should follow the template: mlfa_lab_test_2_<your roll number>.zip. For example, if your roll number is 21CE30021, then the filename for Lab Test 2 will be: mlfa_lab_test_2_21CE30021.zip

1. A single python code (.py as and .ipynb) containing the implementations of the models and experiments with comments at functional level. The first two lines should contain your name and roll number. Note that you need to submit **BOTH** notebook (.ipynb) and python (.py) files.
2. A PDF report containing: **(2 marks)**
    a. **Experiment 1**: Write your seed value, and the device you are running the code on (CPU or GPU).
    b. **Experiment 2**: Write the batch size, overall dataset size, training dataset size, validation dataset size and testing dataset size.
    c. **Experiment 3**: Write the number of convolutional kernels, their sizes, strides, padding, number of neurons in the fully connected layers used. Also write the regularization, loss function, optimizer and learning rate used.
    d. **Experiment 4**: Show the plot of training and validation losses against the number of training epochs.
    e. **Experiment 5**: Write the MSE loss for testing dataset, and show the scatter plot for predicted labels vs. ground truth labels of ages.