CSE343/ECE343: Machine Learning
Assignment-4 CNN, PCA, K-means clustering
Max Marks: 25 (Programming: 15, Theory: 10) **Due Date**: 26/11/2023, 11:59 PM

Instructions

- Keep collaborations at high-level discussions. Copying/Plagiarism will be dealt with strictly.
- Late submission penalty: As per course policy.
- Your submission should be a single zip file **2020xxx_HW1.zip** (Where *2020xxx* is your roll number). Include **all the files (code and report with theory questions)** arranged with proper names. A single **.pdf report** explaining your codes with results, relevant graphs, visualization and solution to theory questions should be there. The structure of submission should follow:

  2020xxx_HW1
  |− code_rollno.py/.ipynb
  |− report_rollno.pdf
  |− (All other files for submission)
- Anything not in the report will **not** be graded.
- Remember to **turn in** after uploading on Google Classroom. No excuses or issues would be taken regarding this after the deadline.
- Start the assignment early. Resolve all your doubts from TAs in their office hours at least **two days before the deadline.**
- Your code should be neat and well-commented.
- **You have to do either Section B or C.**
- **Section A is mandatory.**

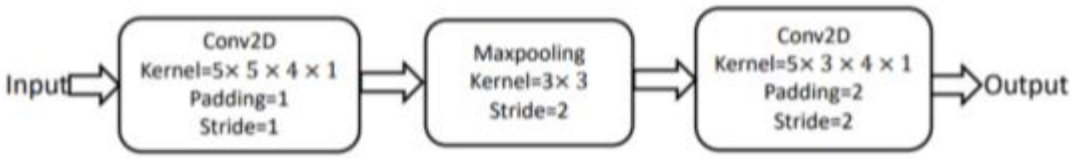
---

1. (10 points) **Section A (Theoretical)**

   (a) (5 marks) Suppose you are given an input image with the dimensions of 15 × 15 × 4, where 4 denotes the number of channels. The same is passed to a CNN shown below:-

   The kernels are of shape h × w × I × O, representing height, width, number of input channels, and number of output channels, respectively.

   (a) What is the output image size? [2]
   (b) What is the significance of pooling in CNN? [1]



(c) Compute the total number of learnable parameters for the above CNN architecture (ignore bias) [2]

(b) (2 marks) Let a configuration of the k means algorithm correspond to the k-way partition (on the set of instances to be clustered) generated by the clustering at the end of each iteration. Is it possible for the k-means algorithm to revisit a configuration? Justify how your answer proves that the k means algorithm converges in a finite number of steps.

(c) (2 marks) Can a neural network be used to model K-Nearest Neighbours algorithms? If yes, state the neural network structure (how many hidden layers are required) and the activation function(s) used at the internal and output nodes. If not, describe why not.

(d) (1 mark) Explain the difference between linear and non-linear kernel or filters in CNN.

2. (15 points) **Section B (Scratch Implementation)**

(a) For this problem, you have to implement Convolutional neural network from scratch. To implement the CNN from scratch you have to implement the following required functions by yourself: You are allowed to use NumPy, Matplotlib, and random libraries.

(a) (6 marks) Convolution Function: Develop forward and backward passes, encompassing windowing and zero-padding with parameters [3,3,3,3].

(b) (6 marks) Pooling Functions: Design a pooling process involving mask creation, value distribution, and both forward and backward pooling with [2,3,3,2].

For each function, use appropriate inputs for testing the functions and report their outputs. For all the functions, you need to include their working and use in the report along with an input/output example. All the functions code must be clean and well-commented (3 marks).

OR

3. (15 points) **Section C (Algorithm implementation using packages)**

Clustering Analysis using PCA and K-Means
Dataset: **Country Dataset**
The task is to perform a clustering analysis to categorize countries based on socio-economic and health factors. You are provided with a dataset containing socio-economic

and health indicators for various countries. The columns include features such as child mortality, exports, health spending, imports, income, inflation, life expectancy, total fertility rate, and GDP per capita.

(a) (3 marks) VLoad the dataset and perform exploratory data analysis (EDA) to gain insights into the data's structure and distribution. Preprocess the data by standardizing the features using a suitable scaling technique.

(b) (6 marks) Implement PCA to reduce the dimensionality of the data while retaining a significant portion of the variance. Identify the optimal number of principal components based on the explained variance ratio. Create scatter plots and heatmaps for visualization.

(c) (6 marks) Apply K-Means clustering algorithm to the PCA-transformed data. Determine the optimal number of clusters using techniques like both the elbow method and silhouette score. Plot the clusters on a 2D scatter plot. Perform an analysis of the clusters to identify the characteristics of each cluster.