Due Date: 29.12.2024 23:55

CENG 463 INTRODUCTION TO MACHINE LEARNING HOMEWORK 3

- K-means Clustering
- T-SNE- Visualization
- Principal Component Analysis

In this assignment, you will explore clustering techniques using a real-world image dataset. The goal is to extract meaningful features from image data, reduce dimensionality, perform clustering, visualize the clusters, and implement a query mechanism to retrieve similar images. The following tasks will guide you step by step.

Task 1: Data Preparation

1. **Download the Dataset**:
    - Download the dataset from Kaggle. The dataset can be accessed from the below link.
    - https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia
2. **Explore the Dataset**:
    - Load the images into the environment and explore the dataset. Check the number of images in each category. Show first 5 images from dataset.
    - You can use libraries like os, matplotlib for loading and visualizing the data.

Task 2: Dimensionality Reduction

Use at least 2 dimentionality reduction techniques to reduce the dimentionality of data. (You can use following methods or other ones that you think is more suitable for the data.)

1. **Apply PCA (Principal Component Analysis):**
    - Use PCA to reduce the dimensionality of your feature vectors and transform the data into 2D space.
    - This helps in visualizing the clusters and will make clustering algorithms more efficient.
2. **Apply t-SNE:**
    - After PCA, use t-SNE (t-Distributed Stochastic Neighbor Embedding) to further reduce the dimensions and visualize the data points in 2D.
    - Visualize the clusters formed by your PCA and t-SNE transformations.

Task 3: Clustering

Use at least 2 clustering tehniques.

1. **Apply K-Means Clustering**:
    - Use the **K-Means clustering algorithm** to group the images into clusters.
    - You will use the features obtained after PCA, t-SNE or another technique for clustering.
    - Determine the appropriate number of clusters (k).
    - Explain why you choose this k value and why do you think that is the proper k value.
2. **Apply Another Clustering Technique:**
    - In addition to K-Means, select and apply one other clustering technique of your choice.
    - Justify your choice of the alternative technique and compare its performance with K-Means.

Task 4: Implement the Query Mechanism

1. **Select a Query Image**:
    - Choose one image from your test dataset as a **query image**. This image will be used to find the most similar images in the dataset. (You can use random function)
2. **Find Similar Images**:
    - Using the K-Means clusters, find the top 3 most similar images to the query image.

- Similarity can be measured using the **Euclidean distance** between the query image and the other images in the same cluster.

3. **Display Results**:
   - Display the query image and the top 3 most similar images found from the K-Means clusters.
   - Show these images in a subplot and label them properly for comparison.
   - You should print out the choosen image's cluster and the similar images' cluster.

Task 5: Evaluate Clustering and Query-Based Performance:

- To assess the performance of your clustering and query-based retrieval system, select the evaluation metric(s) you believe are most appropriate for your approach.
- Apply the chosen metrics and thoroughly discuss the results.
- Justify why the selected metrics are suitable for your dataset and task, and provide insights into the effectiveness of your clustering and query system.
- (e.g. clustering – homogenity, completeness, etc…)
- (e.g. query performance – precision@3, recall@3)

Assignment Rules:

1. In this homework, no cheating is allowed. If any cheating is detected, the homework will be graded as 0, and no further discussion will be entertained.
2. You are expected to submit your homework in groups. Therefore, it will be sufficient if only one member of the group submits the homework.
3. You must upload a .txt file to MS Teams. In this file, include the link to your Google Colab notebook where you have done the project.
4. The .txt file must be named in the following format: group number, course code, and homework number. Example: **G01_CENG463_HW3**.
5. Please be aware that if you do not follow the assignment rules regarding export format and naming conventions, you will lose points.