Data Science 311

Lab 6 (10 points)
Due at 10am on Nov, 28, 2022

Read all of the instructions. Late work will not be accepted.

Overview

In this lab you will work from the notebook presented in class today `imagenet_cluster.ipynb`. You will use clustering to investigate characteristics of the ImageNet ILSVRC dataset that are not immediately apparent from the original labels or a cursory visual inspection. I have curated a toy subset of the ILSVRC validation data containing 8 of the original classes with 50 images each. `imagenet_cluster.ipynb` contains all the code you need to extract the image feature vectors from a pretrained Resnet50 model.

Collaboration

For this lab, you can brainstorm with any classmates about ideas for datasets and data curation methodology. However, your dataset and corresponding explanation of the intended analysis should be unique. Your submission must acknowledge ideas or suggestions you received from other classmates in an acknowledgement section at the end of your notebook.

Details

Tasks

1. Using the columns of "label" and "cluster" make a matrix with the rows the original label [0-9] and the columns the assigned cluster [0-nclusters]. The entries of the matrix at entry i, j will be the number of class i assigned to cluster j. I suspect you will find that most images from the same original class are assigned to the same cluster. However there may be some "oddball" images that get assigned to some other cluster. You may even find that images from one orignal label straddle two distinct clusters in the distribution. Perform the following analyses:
   (a) Describe the distribution of classes over clusters. Some histograms may help in addition to the class/cluster count matrix.
   (b) Look at some oddball images from each class. What are your observations about how these images might differ from the rest of the class images.
   (c) Are there any classes that get grouped together into the same cluster? What are they? Why might these images from the separate classes be grouped together?
   (d) Provide a short paragraph describing your findings and any other interesting observations from this clustering exercise. Make sure to display images that illustrate your conclusions in the notebook

2. Next you will perform "within class" K-means clustering for your choice of 4 classes from our toy dataset. Use $K = 2$. For each within class clustering discuss if you find some meaningful patterns in the clusters that indicate distinct visual concepts. For instance, from a visual inspection in lecture it looked like the "dome" class might be logically split between "inside dome" and "outside dome". Discuss your findings. Make sure to display images that illustrate your conclusions in the notebook.

Submitting Your Work

You will submit a single file, `Firstname_Lastname_lab6.zip`. containing your lab6.ipynb file. (where spelling, spacing and capitalization matter) and upload the zip via Canvas.

Grading

ImageNet cluster analysis will be graded on the following:

- $\frac{2}{3}\%$ of total grade will be on correctness.
- $\frac{1}{3}\%$ of total grade will be on clear exposition of findings.
- $\frac{1}{2}\%$ of total grade will be on task 1.
- $\frac{1}{2}\%$ of total grade will be on task 2.