CS5260 Assignment 2

Due: Week 8, Mar-09-2020, 23:59 SGT

1 Introduction

In this assignment, you are to use or devise attacks on the lung cancer dataset. Your goal is to create images that can fool the pre-trained CNN from `clean_image` while minimizing the following penalty score:

$$l_{\mathcal{S}} = \frac{1}{|\mathcal{S}|} \sum_{\langle \boldsymbol{x}, \hat{\boldsymbol{x}} \rangle \in \mathcal{S}} ||\boldsymbol{x} - \hat{\boldsymbol{x}}||_2, \tag{1}$$

where $\mathcal{S}$ is the set of your submitted image pairs, $\boldsymbol{x}$ is the original image, $\hat{\boldsymbol{x}}$ is the adversarial image classified wrongly by the CNN, and $||\boldsymbol{x}||_2$ is the p-norm of $\boldsymbol{x}$.

You are allowed to use open-source implementations of existing algorithms, however, you should acknowledge them in your report and explain how they work in your code. You may also develop your own algorithm. In this case, you are expected to explain how your algorithm differs from existing ones.

2 Assignment Materials

You have been given the dataset of lung cancer images and the pre-trained CNN for the final project. In addition, you will be given a list of images to perform your attack on in `image_list.txt`, which contains 25 images per category in `clean_image`.

3 Submission

Your submission to this assignment are 100 image pairs $\langle \hat{\boldsymbol{x}}, \boldsymbol{x} \rangle$, and a one-page report describing your method. In each image pair, $\hat{\boldsymbol{x}}$ must be a successful adversarial image created from $\boldsymbol{x}$.

You should create two folders `original` and `adversarial`. Inside these two folders, you should create 4 subfolders `artifacts`, `cancer_regions`, `normal_regions`, and `other`. Images in the same pair should have the same file name and be placed under `original` (for $\boldsymbol{x}$), and `adversarial` (for $\hat{\boldsymbol{x}}$). For example, for

8.png in cancer_regions, your submission archive should have original/cancer_regions/8.png and adversarial/cancer_regions/8.png.

The report must be in pdf format and placed at the same level as original and adversarial folder.

All of the aforementioned materials should be zipped to a single archive renamed with your Student ID (e.g. A1234567X.zip).

4 Grading Policy

Report accounts for 4% grade to overall grade in course. Report will be graded based on the same criteria for your report in the final project.

Performance is measured by score $l_{\mathcal{S}}$ as defined in Eq. 1, the lower the better. It accounts for 4% grade to overall grade in the course. First 25% top scores 4% grade, 25-50% top scores 3% grade, 50-75% top scores 2% grade, last 25%, 1% grade. Actual distribution of grades may be adjusted depending on class performance.