CS640 Project Writeup

Overview

In this project, you will work on a Kaggle competition called UBC Ovarian Cancer subtypE clAssification and outlier detectioN (UBC-OCEAN). Check out their website for more information: https://www.kaggle.com/competitions/UBC-OCEAN.

Original Task and Data

The competition provides training data and one example of the test data in different folders. There are six subtypes of ovarian cancers: CC, EC, HGSC, LGSC, MC, and Other. “Other” designates outliers and is not included in the training set. A participant is asked to submit the code that they use to build and train the model and output predictions of all test data (assuming that the test data folder contains all test samples). Obviously, one challenge a participant faces is to distinguish the outliers that are only present in the test set.

A problem with the data is that the original images are extremely large (many are over 1GB) and the entire training set contains over 500 samples, which makes it difficult and often infeasible to store and run experiments on (even on the competition server).

Therefore, we have made some changes to the format of the competition to fit into our course. If you are able to obtain good results and would like to enter the real competition, we can provide help after the semester ends (the real competition deadline is Jan. 3 next year).

Revised Task and Data

First of all, in this project, we will be using the training data only. Hence there are only five labels in total: CC, EC, HGSC, LGSC, and MC. We have also downloaded and compressed the images to less than 60MB per image. There is a 20% drop in image quality after the compression, which may negatively affect the prediction performance but we believe this preprocessing step is necessary.

We are reserving 20% of the data (in a stratified way) for evaluation and are providing the rest for you to develop your model. The data can be found on SCC at **/projectnb/cs640grp/materials/UBC-OCEAN_CS640/**.

Here is a list of items and you can find in the data directory:

Name	Description
train_images_compressed_80/	A folder that contains the images for training.
test_images_compressed_80/	A folder that contains ~~one~~ all of the test images.

all_labels.npy	A numpy array of the labels. This serves as a map between the labels and their indices.
test_script.py	The script we will run to test your model. Please copy it to your own folder and modify it (there are three places to change, as indicated by the comments). During test time, we will copy your modified version back to test your model.
train.csv & test.csv	CSV files that contain the sample labels.

Suggested Timeline and Actions

Week 1	● Register an account on Kaggle and join the competition (must do so before accessing the data) ● Make a team of at most three ● Take a screenshot of the “Team Members” section on the competition website and submit to Gradescope (only one copy per team)
Week 2 - 4	● Study the data ● Develop and train a PyTorch model ○ You may define your own model and start training from scratch, download a pre-trained model from PyTorch library (e.g., ResNet152), or use some existing model from other people’s work ○ You are expected to try out several models and/or their variants and make comparison ~~● Save the model~~ ~~○ Once you finish developing (training and validating), remember to train your best model on the entire training set and save it~~ ~~○ We will load your model and evaluate it on our reserved test samples. The test script is provided, please make necessary changes and ensure that it runs correctly~~ ● Apply your best model on the test data ○ Predict the labels of the test samples and create a CSV file in the same format as the test.csv but with the label column filled
Week 5	● Write a report and submit to Gradescope (only one copy per team)

Report

You will need to write a report about this project. The template can be found on Piazza under the resources tab. If you prefer, you can use other tools (LaTex, Words, etc.) to format the report. The submission should still be PDF.

Grading Criteria

You will be graded based on your model performance on our reserved data and how much effort you put into the project.
Here is a general breakdown of the grades:

- Performance (30%)
    - The main metric we use is the F1 score
    - You should at least beat random guessing
- Report (60%)
    - Your report doesn't have to be long, but must be complete (10%)
    - Your description should be clear (30%)
        - Do **not** simply list numbers or figures
        - Do **not** build a wall of text
        - When explaining your approaches and results, using a combination of words and figures can be very helpful
    - Statements should be backed by evidence and reasoning (20%)
        - For example, if you find some interesting relation between A and B, you should show evidence (say, a correlation plot).
- Presentation (10%)
    - Prepare a three-slide presentation
    - The presentation is like a short version of your report:
        - Briefly talk about your approaches
        - Summarize your results and observations