Instructions for Kaggle Competition 2024

IFT6390B

November 25, 2024

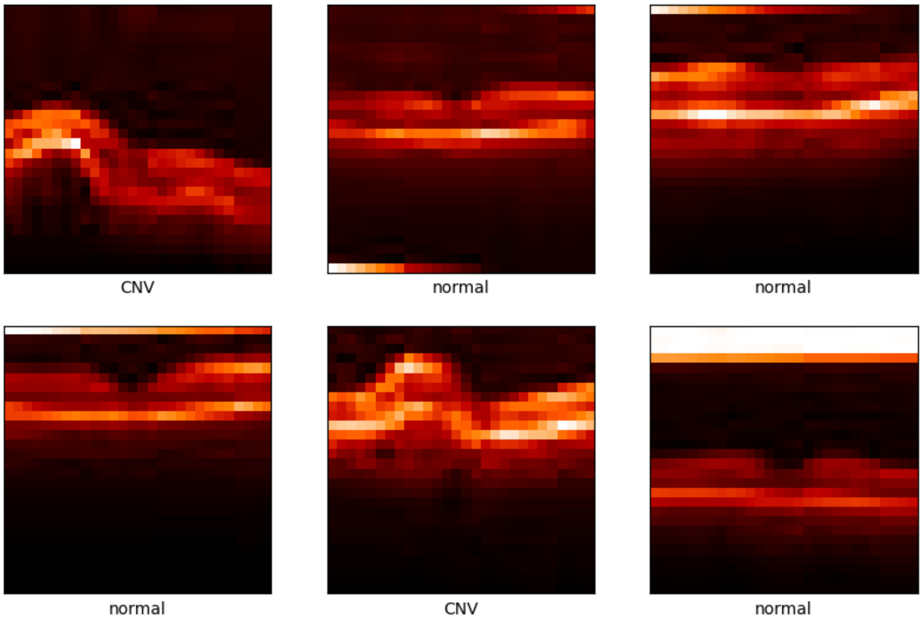

Figure 1: Example images from the training set

1 Description

In this project, you will take part in a Kaggle competition for image classification. The goal is to design a machine learning algorithm that can accurately identify retinal diseases associated to an image of an Optical Coherence Tomography (OCT) (imaging test to take cross-section pictures of the retina). In order to maintain the anonymity of the dataset and ensure a fair competition, we do not disclose the name of the dataset and perform basic transformations on the images (flip, rotations). Your goal is to accurately classify the images in one of the following labels:

- 0: choroidal neovascularization
- 1: diabetic macular edema
- 2: drusen
- 3: healthy retina

in a way that minimizes the classification error on the test set: $\sum_{i=1}^{n} \mathbb{1}_{\hat{y}_i \neq y_i}$
In summary, you are given the following data:

- `train_data.pkl` - This is a *pickle* file containing both the images and the labels for the training set. You can open it using the following script in Python:

  ```
  import pickle

  # Load the pickle file
  with open('path_to_file.pkl', 'rb') as f:
      data = pickle.load(f)

  # Access images and labels
  images = data['images']
  labels = data['labels']
  ```

  The ”data” variable is thus a dictionary with keys ”images”, that is a list of **28x28** images represented as numpy arrays, and ”labels” that is a list of labels ranging from 0 to 3.

- `test_data.pkl` - Another pickle file containing the images of the test set. This time, you can open this file exactly as before but the dictionary will not contain the key ”labels” that you will need to predict.

2 Participation

For the graduate section (IFT6390), the task must be solved **in teams of 2.** IFT3395 students can **optionally** participate in the competition in **teams of 2 or 3.** An indiviual participation for both sections is also authorized. In order to participate in the competition, you should:

- Create a Kaggle account if you do not have one already.
- Enter the competition using the following invitation link: `https://www.kaggle.com/t/1c523e26262b439fae68924234255054`.
- From now on, you can access the competition via `https://www.kaggle.com/competitions/ift3395-ift6390-identification-maladies-retine/`.
- In the ”Invite Others” section, enter your teammates’ names, or team name.

- Your teammate has the option to accept your merge.

- Fill out the google form `https://forms.gle/jYEHBWBsG7B6WQhC7` with your team information by **Nov 22nd, 23:59**. Any teams not registered or registered late will not be graded.

Important note: The maximum amount of submissions is 2 per day, per TEAM. Any team whose individual members have a submission count larger than what is allowed up to-date will be UNABLE to form a team. Example: Today is the first day of competition. A,B,C are three teammates who haven't formed a team yet.

- A submitted 0 times.

- B submitted 2 times.

- C submitted 1 time.

Because the maximum amount of submissions is 2 per team per day, the total possible submissions for a team is 2. However, the cumulative submission count for A,B,C is 3. Therefore, they will be unable to form a team (They will need to wait for tomorrow, and not submit any submissions for the next day).

3 First milestone: Beat the baseline (Nov 25th)

You can see two baseline scores on the leaderboard. The first score corresponds to a classifier that assigns random labels to each image. The second baseline corresponds to a vanilla logistic regression classifier. The logistic regression classifier was trained using only 10% of the training set. For the first milestone, you will need to beat the baseline logistic regression classifier on the public leaderboard. Here are some possibilities of methods:

- Non-linear classifiers such as Kernelized SVM

- Hand-crafted features and logistic regression

- Decision trees and random forests

Important note: To beat the baseline, you are NOT allowed to use any machine learning library, e.g. `scikit-learn`. You should implement your solution from scratch using only NumPy and basic Python functionalities.

Important note: Do not hesitate to use a subset of the training set, that contains originally 1̃00000 training samples.

The goal is to design the best performing method as measured by submitting predictions for the test set on Kaggle. Your final performance on Kaggle will count as a criterion for evaluation, as well as the number of baselines that you beat. If a tested model does not perform well, you can still add it in your report and explain why you think it is not appropriate for this task. This kind of discussion is an important feature that we will be using to evaluate your final competition report.

4 Second milestone: Compete (December 3rd)

You have until **December 3rd 23:59** to achieve the best performance you can on the task. In this phase you are free to implement any method you think would work best, and use any library you deem useful, like scikit-learn, Pytorch or Tensorflow. The Kaggle leaderboard has a public and private component to prevent participants from "overfitting" to the leaderboard. The public leaderboard shows your score calculated on 50% percent of the test set, while the private leaderboard is based on your score on the other half of the test set. You are only able to see the public leaderboard during the competition. The points for this phase will be given based on your ranking on the private leaderboard that will be released at the end of the competition.

Important note: You must submit two separate solutions, one for the first phase (beating the baseline), and one for the second phase (your best-performing model). You should name your submission files to distinguish between the two. For your code submission on Gradescope, you should also separate the two solutions.

5 Third milestone: Submit Code and Report (December 6th)

You must write up a report that details your machine learning pipeline, including pre-processing, algorithms, optimization and learning, hyperparameter tuning, and validation procedure. You should also provide and compare the results of other methods you implemented before reaching the best-performing model. The report should contain the following elements. You will lose points if you do not follow these guidelines.

- Project title
- Your team name on Kaggle, as well as the list of team members, including their full name and student number (for graduate students each team has only one member).
- Introduction: briefly describe the problem and summarize your approach and results.
- Feature Design: Describe and justify your pre-processing and feature extraction methods.
- Algorithms: Give an overview of the learning algorithms used without going into too much detail.
- Methodology: Include any decisions about training/validation split, regularization strategy, optimization tricks, setting hyperparameters, etc.
- Results: Present a detailed analysis of your results, including graphs and tables where appropriate. This analysis should be broader than just the Kaggle results: include a short comparison of different values for important hyperparameters in your best-performing algorithm and also compare the performance of this method with at least two other methods you implemented. Also, include the results from the method used for Milestone 1.

- Discussion: Discuss the pros/cons of your approach and suggest ideas for improvement.

- References (very important if you use ideas and methods that you found in some paper or online; it is a matter of academic integrity).

- Appendix (optional). Here you can include additional results, more details of the methods, etc.

The main text of the report should not exceed 6 pages. References and appendix can be in excess of the 6 pages.

Your must submit your code (first and second milestones) and report (third milestone) on Gradescope before **Nov 12th, at 23:59**.

Submission Instructions

- You must have separate .py files/notebooks for the first and second milestones. The code must be well-documented. If you are not using Jupyter notebooks, you should include a README file containing instructions on how to run the code. You will need to submit a zip file containing your code and related files to Gradescope.

- The prediction file containing your predictions on the test set should only be submitted to Kaggle.

- The report in pdf format (written according to the general layout described earlier) should be submitted to Gradescope.

6 Evaluation Criteria

1. You will receive a minimum number of points if you beat the logistic regression baseline on Kaggle's public leaderboard (conditioned on your adherence to the aforementioned instructions).

2. You will be graded depending on the quality and technical soundness of your final report.

3. You will receive **bonus** points depending on your final ranking on Kaggle's private leaderboard at the end of the competition.

7 Deadlines

The deadlines for this project are firm, and each submission must include all required components specified for each milestone. Failure to meet these requirements by the specified deadlines may result in a loss of points

- The deadline to form teams and report it on the Google forms is **November 22nd, at 23:59**

- The deadline to beat the baseline is **November 25th, at 23:59**.
- The Kaggle competition will close on **December 3rd, at 23:59**.
- You must upload your report and code on Gradescope before **December 6th, 23:59**.