COMP432 Introduction to Machine Learning
Course Project Guideline

Posting Date: September 9th, 2024

Lecturer:	*Prof. Mahdi S. Hosseini*	*[email: mahdi.hosseini@concordia.ca]*
Teaching Assistant:	*Ahmed Alagha (Lead TA)*	*[email: ahmed.alagha@mail.concordia.ca]*

Email Inquiries

All inquiries about the course project (without any exception) should be communicated via email using addresses above. Your email subject line must follow a prefix topic [COMP432 Project: *{your subject}*].

All prefix characters are case-sensitive and opening/closing brackets must be included. Note that {your subject} can be anything. For example, use the subject line [COMP432 Project: Course Project-Team Formation] to inquire about the project team formation. ***Note: Other formats will NOT be replied.

Main Objective

The main goal of this project for students is to study a Computer Vision (CV) task in deep learning using Convolutional Neural Network (CNN) backbone models to address image classification problems from a real-world application of interest. The datasets, coming from different applications, will be provided for you, and you will be tasked with training and tuning CNN models, in addition to conducting detailed analysis on the performance of your models. You will also explore transferring knowledge from one application to the other, where previously trained models on a dataset are tested on other datasets from different applications.

Datasets

The datasets needed for this project come from the fields of computational pathology and computer vision. You will be using three(3) different datasets in this project, two of which are from the field of pathology, while the third is from computer vision. The original source of each dataset is given to you for reference, but you should the specified "Project Dataset" in this project. The datasets are:

- **Dataset 1: Colorectal Cancer Classification [Original Dataset | Project Dataset]**
  This is a dataset of 100k image patches split into 8 different classes identifying the tissue type, including Adipose (ADI), background (BACK), debris (DEB), lymphocytes (LYM), mucus (MUC), smooth muscle (MUS), normal colon mucosa (NORM), cancer-associated stroma (STR), and colorectal adenocarcinoma epithelium (TUM). In this project, we will be using a reduced version of the dataset, containing 6K image patches split into 3 classes: smooth muscle (MUS), normal colon mucosa (NORM), and cancer-associated stroma (STR).
- **Dataset 2: Prostate Cancer Classification [Original Dataset | Project Dataset]**
  This is a dataset of 120k image patches split into 3 different classes identifying the tissue type, including Prostate Cancer Tumor Tissue, Benign Glandular Prostate Tissue, and Benign Non-Glandular Prostate Tissue. For this project, the dataset is reduced to 6k image patches for the same 3 classes.
- **Dataset 3: Animal Faces Classification [Original Dataset | Project Dataset]**
  This is a dataset of 16k images split into 3 different classes identifying animal types, including Cats, Dogs, and wildlife animals. For this project, the dataset is reduced to 6k images for the same 3 classes.

Detailed Tasks:

The project is divided into two main tasks, given as:

- **Task 1:** in this phase, you are required to train a CNN model for the Colorectal Cancer Classification problem. You are free to choose one of the common CNN architectures (ResNet, VGG, AlexNet, MobileNet, ShuffleNet) and train it on Dataset 1. For all the experiments, you should report your results in the form of training accuracy/loss. You should also use t-SNE for dimensionality reduction of the output features of the CNN encoder, which are to be visualized according to the class labels. This is an addition to appropriate discussions explaining the obtained behaviors.
- **Task 2:** In this task, the final CNN encoder (without the classification head) obtained in task 1 is to be applied to Dataset 2 and Dataset 3. You are not required to train a CNN model. Instead, you are tasked with analyzing and visualizing the feature extraction done by the pre-trained CNN encoder when applied to the new datasets. You should use t-SNE to analyze and visualize those features in accordance with the class labels. You are to repeat these steps but while using a CNN encoder that has been pre-trained on the ImageNet dataset. This is available for most CNN architectures through the PyTorch library. In summary, you should analyze four(4) scenarios, two of which while using the CNN encoder from Task 1 on Dataset 2 and Dataset 3, and the other two while using a pre-trained

ImageNet CNN encoder on Dataset 2 and Dataset 3. All the scenarios should be analyzed and visualized using t-SNE, applied to the extracted features. Finally, you should use on of the classical machine learning techniques (SBM, RF, LR, etc) to classify the extracted features on Dataset 2 and Dataset 3. It is sufficient to do one classification per dataset, i.e. you are free to choose which CNN encoder to apply the classification to, as long as you conduct one classification per each of the two datasets.

Team Formation [Deadline: Thursday 11:59PM, September 19th, 2024]

Students are required to form a team of Five(5) members for the course project. Please submit your team's detail by email to the lecturer and the Lead TA following the email inquiries guideline. A Q&A discussion forum will be created for the course on Moodle and you can use the platform to open a discussion on team formation related topic. Students who cannot find a team will be randomly shuffled in incomplete teams. The team, once formed, will stay the same until the end.

Proposal Submission [Deadline: Thursday 11:59PM, September 26th, 2024] (Counts for 10% of the course project grade)

You should write a one-page proposal for the course project to cover the following topics:

- *Problem Statement and Application*: provide a background about the topic to be investigated and specify why the problem is interesting and important? What are the associated challenges of the problem application? What are your expectations/goals throughout developing the application of interest?
- What reading material (e.g. papers, scientific reports, etc) will you examine to provide context and background?
- *Possible Methodology*: highlight the possible method or algorithm you are proposing. Are there any existing implementations to be used and how will you use them? How are you planning to improve or modify such implementations? You may not have the exact answer here but try to give an answer that you will follow as much as possible.
- *Metric Evaluation.* Discuss how you will evaluate your results both in terms of qualitative and quantitative analysis. Qualitatively, what kind of results do you expect (e.g. plots or figures)? Quantitatively, what kind of analysis will you use to evaluate and/or compare your results with (e.g. what performance metrics or statistical tests)? All these metric evaluations will be used to assess and evaluate the pipeline and your expectations regarding the kind of results/performance to be achieved.
- *Gantt Chart*: use an additional page (supplemental material) to illustrate a Gantt chart of the project development to list (a) schedules and (b) items of milestones and deliverables. Note that you cannot use this page to extend your proposal description.
- *Bibliography*: use an additional page to extend your reference list cited in your proposal. The citations may include, but not limited to, published papers and domain links. (include a link to your dataset). Please note that failure to properly cite your references constitutes to a plagiarism.

You will be given the opportunity to submit your proposal for revision by the professor/Lead-TA, before the final graded submission. **The deadline for final proposal submission is October 6th, 2024**. Only the admin (one person) of your team needs to upload the proposal in PDF file in Moodle.

For the report format, please consult "Reports Formatting" Section in this guideline. Our team (TAs and lecturer) will review your proposal and, if it is acceptable, you may proceed with developing the next phase of your project. Otherwise, we will instruct you to either revise or re-write the proposal according to the guidelines of the course project. All teams are highly encouraged to put great effort on preparing the first proposal draft to avoid further delays in project developments.

Progress Reporting [Deadline: Thursday 11:59PM, November 3rd, 2024] (Counts for 20% of the course project grade)

Each team is required to submit a three(3)-page progress report highlighting the main steps taken after the proposal, and any initial results (if available). The progress report should contain the following sections:

1) *Introduction*: In addition to defining the problem and its applications, discuss the general strategy followed by existing methods for tackling the issue at hand. Discuss the challenges faced in solving this problem and your proposed solutions to address them. Discuss what results you expect and how you want to acquire/evaluate them.
2) *Method*: Give updates regarding the methods used/to be used. Discuss the application, dataset(s), deep learning model(s) in more detail.
3) *Attempts at solving the problem*: elaborate on failed or successful attempts at tackling the problem. Furthermore, discuss any possible/preliminary results.
4) *Future Steps*: Discuss the plan for the next period and the remaining steps to be executed.
5) *References*: add an additional page to extend your reference list cited in your progress report. The citations may include, but not limited to, published papers and domain links (include a link to your dataset).

6) *Supplementary Material* [this section is appended to the main report draft]: you are encouraged to include appendices to your report to support different sections of the main draft in more detailed analysis. **Note: this section will not be considered for marking.

The progress report should be in PDF format and uploaded in Moodle. For the report format, please consult "Reports Formatting" Section in the third page. Please note only the admin (one person) of your team needs to upload the progress report in PDF file in Moodle.

Final Reporting [Deadline: Sunday 11:59PM, December 1st, 2024] (Counts for 40% of the course project grade)

The final report should articulate the following sections:

1) *Abstract*. Articulate on the abstract presentation of the project and what to expect by reading your report in full detail. Briefly discuss the problem, proposed methods and used data, and the achieved results. [maximum of 150 words].
2) *Introduction* [the abstract & introduction should be around 2 pages]:
    a) Write a section to cover the problem statement and its importance to the application field. What are the associated challenges with respect to the problem? How is this report trying to solve the problem and a challenge in mind? Elaborate on the high-level abstract explanation of your methodology and what kind of implementations you have done. What kind of results you are obtaining?
    b) Related works. Write a subsection to cover literature review and related work descriptions.
3) *Methodology* [this section should be around 3 pages].
   The methodology section should cover the proposed idea to solve the problem stated in your introduction. You can use figures/diagrams to better explain your methodology. If applicable, emphasize on the improvement/new approach you have taken to solve the problem.
4) *Results* [this section should be around 3 pages].
   This section describes and analyzes the experimental design and obtained results in detail. More specifically
    a) Experiment Setup. you need to describe how you setup your experiments, optimized and validated your deep learning models, the performance analysis using appropriate metrics (precision, recall, F1-measure, …). Explain the ranges of hyper-parameters and rational behind selecting as such in relation to your data and models.
    b) Main Results. Demonstrate the main results in figure/table formatting and analyze the performance of your implemented results. Discuss the results and use any means of visualization/table formatting/figure demonstration to better explain the obtained performances.
5) *References* [this section lists all references beyond the eight page of your report]:
   Cite any references you used in the projects, including any source code and dataset you have used in the project. Please note that failure to properly cite your references constitutes to a plagiarism and will be deemed for reporting.
6) *Supplementary Material* [this section is appended to the main report draft]:
   You are highly encouraged to include appendices to your final report to support different sections of the main draft. **Note: this section will not be considered for marking.

Reports Formatting

The proposal (1 page + 1 page Gantt Chart supplement + 1 page bibliography), the progress report (3 pages + 1 page bibliography), as well as the final report (8 pages + bibliography page(s) + possible appendices) should all be written in **CVPR LaTeX template** for your final PDF submission. ***Note: other formats will **NOT** be accepted. Note to use the **main.tex** file for LaTeX compilation.

Final Presentation [Deadline: Sunday 11:59PM, December 1st, 2024] (Counts for 15% of the course project grade)

Each team should prepare a 10-minute recorded video from a slide presentation and submit the following

- A 10-Page deck of slides prepared in PDF format (you can use either PowerPoint or LaTeX beamer for your slide preparation). Slides should contain a high-level overview of the problem and goals, the type of data you were dealing with, your methodology, the obtained results, and the references used.
- A 10-minute recorded video from the team, each member taking a round of 2 minutes in a row to complete the record.

GitHub Submission [Deadline: Sunday 11:59PM, December 1st, 2024] (Counts for 15% of the course project grade)

Whether you use Git to organize your coding throughout the project or not, each team should create a new GitHub page for the project from the beginning. The GitHub page should be created in "private" mode and each member should be given access to commit their updates on a regularly basis during the course of the project. Furthermore, the lead TA for the project team as well as the lecturer should be given access to the GitHub page for monitoring the progress of the team. Note that commits from each team member will be monitored for the engagement of individuals and considered as one of the means of marking to contribute to their final project. The final GitHub page should contain the following

- High level description/presentation of the project
- Requirements to run your Python code (libraries, etc)
- Instruction on how to train/validate your model
- Instructions on how to run the pre-trained model on the provided sample test dataset
- Your source code package in PyTorch
- Description on how to obtain the Dataset from an available download link

Please note that if the instructions to run your code are incomplete or not explicit enough, you might lose marks for that part of the project. You should add the professor and the Lead-TA as contributors to your project, using the following addresses:

- GitHub ID: "AtlasAnalyticsLab"
- GitHub ID: "AhmedNAlagha"

How to Submit Your Project Materials? [Deadline: Sunday 11:59PM, December 1st, 2024]

Submit all the files in one zip file including

- PDF file of the final report
- Deck of 10-Slide page presentation in PDF format
- README.txt containing the following two links:
  - A link to your GitHub page.
  - A download link to your video presentation
- A sample test dataset (100 images)
- One page that includes a table listing the contribution of each team member to the project. The table format should be in five(5) columns pertinent to individual members of the team. The pertinent information will be considered to grade individual contribution to the project.

The zip file should be uploaded by the admin of the team in Moodle by the final submission deadline.

Late Submissions Policy

If you submit any part of the project later than the specified deadline on Moodle, your submission will be accepted until the cut-off date. However, you will lose 20% of the mark for each day you submit late. The cut-off date is maxed up to two (2) days and submission after the cut-off date will not be accepted. Further, please note that resubmitting your files will result in erasing all the previously submitted versions and their respective dates. The date of the last attempt at submission will be counted as the final submission date.

Peer Evaluation

Towards the end of the project, each group member will submit a peer evaluation form evaluating the other team members. The individual project grade obtained by a member will be influenced by the overall evaluation from other team members.

Potential Paper Publication (Optional Track)

While this track is completely optional, interested project-group can consult with the lecture team (i.e. professor and lead TA) on the possibility of submitting their work to main/workshop venues of ML conferences. This will be decided after final project submission/exam.