MiniProject 4: Reproducibility in ML

COMP 551, Winter 2022, McGill University
Contact TA: David Venuto, Yujing Liu

General information

- This mini-project is **due on April 26th at 11:59pm (EST, Montreal Time).** There is a penalty of 2^k percent penalty for k days of delay, which means your grade will be scaled to be out of $100 - 2^k$. No submission will be accepted after 6 days of delay.
- This mini-project is **optional** and can result in up to 10% bonus points added to the overall grade (10% of the course).
- This mini-project is to be completed in groups of two to three. There should be a short paragraph on each of the members contribution to the submission. All members of a group will receive the same grade except when a group member is listed as not responding or contributing to the project. If this is the case and there are major conflicts, please flag this in the submitted report. Please note that it is not expected that all team members will contribute equally. However every team member should make integral contributions to the project, be aware of the content of the submission and learn the full solution submitted.
- You will submit your assignment on MyCourses as a group. You must register your group on MyCourses and any group member can submit. See MyCourses for details.
- We recommend to use **Overleaf** for writing your report. You are encouraged to use existing open sourced python implementation.
- You should use Python for this mini-project. You are free to use any libraries as well as open sourced repositories or implement your own for anything you can't find an existing implementation.

Background

One goal of publishing scientific work is to enable future readers build upon it. Reproducibility is the central theme to achieve this target, yet it is unfortunately one of the biggest challenges of Machine Learning Research. Everyone is encouraged to follow the reproducicbility checklist while publishing scientific research, to make the results reliable and reproducible. In addition, a challenge is organized every year to measure the progress of our reproducbility effort. The participants select a published paper from one of the listed conferences, and attempt to reproduce its central claims. The objective is to assess if the conclusions reached in the original paper are reproducible. The focus of this challenge is to follow the process described in the paper and attempt to reach the same conclusions. We have designed this miniproject in the spirit of the reproducibility challenge.

Problem definition

The goal of this assignment is to select a paper and reproduce the results of the paper by following the exact methods mentioned in the paper. You can choose a paper from the few example papers listed here or find one of your choice that meets the criteria mentioned below. For this mini project, you are not expected to implement anything from scratch. You are encouraged to use any code repository published with the paper or any other implementation you might have found online.

Paper selection guidelines

- To minimize the overlap between this miniproject with the previous ones, we have decided on a few broad categories the paper must belong to :
  1. Vision/Image Processing - The paper should have a vision/image processing unit which you have not used in the previous projects. It can also be a combination of vision and text data.
  2. Generative Models, e.g. GANs
  3. SVM
  4. Nearest Neighbours
  5. Decision Trees
  6. Clustering
  7. Dimensionality reduction
  8. Ensemble Methods
- You should be able to access the data or environment you will need to reproduce the paper's experiments.
- In many cases a codebase might be available directly from the authors or another source (if the paper is old). You should definitely check whether you can handle the code before deciding on the paper.
- You should estimate the computational requirements for reproducing the paper and take into account the resources available to you for the project. Some authors might have had access to infrastructure that is way out of your budget; you might not want to choose such a paper.
- You are free to choose any paper from the current pool of papers of the reproducibility challenge, or any classic paper such as the example papers mentioned below. Just make sure that the paper chosen overlaps significantly with at least one of the above mentioned broad categories. Another great place to look for a relevant paper is Papers with Code.

  A few example papers:
  - CNN+SVM paper: Deep Learning using Linear Support Vector Machines
  - AlexNet paper: ImageNet Classification with Deep Convolutional Neural Networks
  - t-SNE paper: Visualizing Data using t-SNE
  - VGG paper: Very Deep Convolutional Networks for Large-scale Image Recognition
  - Dropout paper: Dropout: A Simple Way to Prevent Neural Networks from Overfitting
  - Kernel SVM paper: Online Learning with Kernels

Experiments

You don't need to reproduce all the experiments of your selected paper. From your selected paper, you can choose a **subset of the experiments** that's feasible for you to reproduce in terms of computation resources. The experiments you choose to reproduce should not be baselines, instead they should be the methods introduced in the publication. Reproducing the method for one experiment type or environment is sufficient.

It is generally easier to reproduce seminal papers rather than recent works. It is perfectly fine to reproduce a seminal work, as long as it is not one of the general methods covered in the previous projects.

Some state of the art models can demand higher computation power than you have access to. In such cases, you might want to reproduce only the baseline model described in the paper. Often hyper-parameter search on the baseline models has not been performed well and there can be a better model than the one reported in the paper. You can implement the models from scratch or use the code provided by the authors. But make sure to add all the resources you have used in your references.

Several models above also have **pretrained weights** available to download. Since these have been trained on huge datasets, you are encouraged to code up the models and directly import these weights instead of training from scratch. You can then use the pretrained model for experimentation and ablation studies as well as fine-tune the weights on new data.

- You will first reproduce the results reported in the paper by running the code provided by the authors or by implementing on your own, if no code is available
- You will try to modify the model and perform ablation studies to understand the model's robustness and evaluate the importance of the various model components. (In this context, the term "ablation" is used to describe the process of removing different model components to see how it impacts performance.)
- You should do a thorough analysis of the model through an extensive set of experiments.
- Note that some experiments will be difficult to replicate due to computational resources. It is fine to reproduce only a subset of the original paper's results or to work on a smaller variant of the data—if necessary.
- At a minimum, you should use the authors code to reproduce a non-trivial subset of their results and explore how the model performs after you make minor modifications (e.g., changes to hyperparameters).
- An outstanding project would perform a detailed ablation study and/or implement significant/meaningful extensions of the model.

Deliverables

You must submit two separate files to MyCourses (using the exact filenames and file types outlined below):

1. code.zip: A collection of supporting code files. Please submit a README detailing the packages you used and providing instructions to replicate your results.

2. writeup.pdf: Your project write-up as a pdf (details below).

Report guidelines

Write a report of no more than 6 pages (excluding reference) covering the below points. Use this latex template for the main paper. You are allowed to have an additional appendix, but the main findings of the paper should be documented in the main paper (6 pages).

- Abstract and introduction defining the problem statement, experiments conducted and summarizing the results of your experiments.
- Briefly describe the dataset.
- In the main paper, document the results of your experiments.
- Specify the hyperparamter tuning and ablation studies that you have performed and their results.
- From your experimental results, did you reach the same conclusion as the authors?
- Any necessary details for reproducing the results, but were not specified in the original paper.
- Challenges that you have faced and how did you solve them.
- Summarize the key takeaways from the project and possibly directions for future investigation.
- State the breakdown of the workload across the team members (statement of contribution).

Evaluation

- This is an open ended project meant to help you use the theoretical and applied knowledge from this course to implement, experiment and tinker with actual, popular research work in the field.
- In general your work will be graded based on the following criteria
  - Understanding of the concepts that are part of the paper you choose, which you will communicate through your report. (15%)
  - Quality of Experiments done and a scientific description of the same in the report. How detailed/rigorous are your experiments? Give detailed reasoning of the method used for analyzing the model performance when comparing with previous models and discuss why it's suitable for this problem. Run the same experiment several times and make a detailed, quantified analysis of the model performance such as evaluating the means and variance. Give detailed reasoning on how you fine-tune the hyperparameter rather than randomly changing them. For example, how to change the batch size and learning rate based on the current performance. (30%)
  - Explanation of the reasoning behind various experiments, ablation studies and the observed results based on the concepts taught in the course. Dissect the model clearly and analyze accurately the improvements made by the model and what each part does. Give detailed reasoning on your understanding of the design of different parts. (25%)
  - Application of Machine Learning tools and frameworks (pytorch, sklearn etc) for modeling, experimentation and visualization of results.(10%)
  - Quality of report which inquires(20%)
    * Does your report clearly describe the task you are working on (i.e., the paper you are reproducing),the experimental set-up, results, figures (e.g., don't forget axis labels and captions on figures, don't forget to explain figures in the text).
    * Is your report well-organized and coherent?
    * Is your report clear and free of grammatical errors and typos?
    * Does your report include an adequate discussion of related work and citations?

References

1. The Reproducibility Challenge as an Educational Tool
2. UW NLP Class