BDH-reproducibility-challenge

Jimeng Sun

***Abstract*—CSE6250 Big Data Analytics for Healthcare is a graduate level course focusing on practical big data technology and machine learning for health analytic applications. One big part of this course is to conduct a group project that reproduces some published work about machine learning/deep learning (ML/DL) in healthcare. While taking this challenge, you should understand, replicate, and evaluate some recent publications and provide an end-to-end coverage of the core implementation in the publication including data, machine learning algorithms, validation of outputs, etc. I hope that the best projects (with some additional effort) can lead to a review at the best medical informatics venues.**

This document provides the project guideline such as expectation, timeline, deliverables.

***Index Terms*—Machine learning/Deep learning, Healthcare, reproducibility,**

I. Paper Selection

You should select at least a paper from the provided paper pool and aim to replicate the main claim described in the paper. The objective is to assess if the experiments are reproducible, and to determine if the conclusions of the paper are supported by your findings. There are some considerations in choosing a paper to reproduce:

• You should find the problem tackled in the paper interested.

• You should be able to access the data you will need to reproduce the paper's experiments.

• You should choose paper whose computational requirements for reproducing the experiment is affordable to you.

• You cannot not choose a paper that you, course staff, or someone in your current or former lab has written.

• Even though the codebase in paper is open source, which is very common nowadays, you should not directly copy-use it. Instead, you should develop your own code. Of course, the codebase in original paper can be your reference.

Your project will try to reproduce the main experiments in your selection and assess the ease of reproducibility. The project result can be either positive (i.e. confirm reproducibility), or negative (i.e. explain what you were unable to reproduce, and identify all the questions that would need to be answered to reproduce the experiments). Both outcomes are acceptable and can earn full credit.

II. Teams

You **MUST** team up with another classmate, i.e., your team should be composed of **TWO** people. Individual project is not allowed. Both team members will receive the same grade on the project (except that on member contributes much less than the other, which will lead to penalty on the project grades of the inactive member).

III. Project Milestones

Next we summarize the timeline for your project in this semester. Please remember that **we don't allow late submission** for any project components.

Due Date	Task Description
Oct 3	Group formation & (candidate) paper selection
Oct 17	Project proposal
Nov 21	Project draft
Dec 5	Final Submission (final paper + code + presentation)

IV. Project Requirements

A. Group formation & Paper selection

You should form a team with another classmate and your team should select 2-3 candidate papers from the paper pool. Not every paper in paper pool is feasible for reproduction, you should read the paper to see if you are able to reproduce it (whether you have data access and sufficient computational power) before you select. Course staff will share a spreadsheet for team registration & paper selection, it contains two sheets. team registration & paper selection

1) Team registration. Every team should register two team members with name, GT ID, GT email and you should also input the indices of your candidate papers.
2) Paper Selection. It will contain indices of papers in paper pool. You should input your team index and team member initials after the paper you want to reproduce so that everybody knows which paper has been selected already.

Note: Please watch the header while finishing the sheets and DO NOT change information of other group!

After finish two sheets, write the information up and submit on Gradescope. It can be in any format but must contain:

1) Names, GT usernames and ID of two members.
2) Indices, titles and authors of selected papers.

- At this stage, your team should read those candidate papers roughly and construct basic understandings of them.
- To avoid that many groups choose one same paper, we restrict each paper with maximally TWO groups. The paper selection is first come first serve.
- Deliverable:
  - Finish**TWO** sheets.
  - Submit a **PDF** no longer than 1 page on Gradescope **at team level**. Use this instruction for how to submit at team level. Remember to include your teammate!

  All requirements must be finished before deadline to earn full credits, fail to finish all of them, such as no team-up, not finish both sheets (finish only one or none), writeup

submission on Gradescope is missed or does not include both members, etc, will make the your team get **zero** for that and your team has to re-choose candidate papers from whatever left.

B. Project Proposal

At this stage, you review those candidate papers and decide which paper to reproduce. Your proposal should cover the following contents about your project, in order to demonstrate you have thought carefully about the paper you are planning to reproduce, and in order to communicate your understanding of the work and its importance to someone who most likely has not read the paper (e.g. course staff):

- Summary to all candidate papers
  - What is the general problem in this work (e.g. '10-day ICU readmission prediction')
  - What innovations are in this work (e.g. a new network structure/special feature construction/specific analysis to data)?
  - What advantages/disadvantages does the work have (e.g. accuracy to current problem is high/method is hard to be generalized)?
  - What is the data used in this work? If the data is accessible, attach the link in your paper
  - Is the codebase provided? If so, attach inks to the codebase in your target paper (Github, Gitlab, Bitbucket, etc).

  Note: it is important and required to well organize this summary. Use a list would be favorable for graders to quickly get your point. Such as:

  Paper 1: Index, Paper Title, Venue and Author(s)

  1) Task: 10-day ICU readmission prediction.
  2) Innovation: a new network structure is proposed.
  3) Dis/Adv: model achieves high accuracy, which beats the bench mark, however, it contains abundant coefficient and takes long time to train.
  4) Data Accessibility: Yes (Link)
  5) Code Accessibility: Code is not provided by author

  Paper 2: Index, Paper Title, Venue and Author(s)

  1) Task: .....
  2) Innovation: .....
  3) Dis/Adv: .....
  4) Data Accessibility: .....
  5) Code Accessibility: .....

- Decide your target paper
  - Which paper in the candidate you will replicate.
  - Why you choose the paper.
  - What are the specific hypotheses from the paper that you plan to verify in your reproduction study?
  - Briefly state how you are assured that you can obtain appropriate data and computational resources including software and hardware demanded in the paper.
- Deliverables
  - A **PDF** no longer than **2 pages main content** + Appendix + References (The candidate papers and important references in them must be cited)
  - Submit to **Gradescope** at team level, remember to include your teammates in submission.
  - Use your own words instead of description in original paper.

  Note: It is encouraged to choose more than one paper or paper with greater difficulty to reproduce. Let's say if the paper you choose to reproduce is difficult, i.e., it requires large dataset and has complicated network structure and optimization method, perform greatly on provided metrics, you will be fine even though you could not reproduce with comparative results. Such work usually contains many details, and it is not easy to cover all of them. So that we would not be very harsh on the results you obtain. However, if the paper is relatively easy, we would expect that you obtain convincing results.

C. Project Draft

At this stage, you should have fully understood your paper and realized all experimental setup and details. You should also have developed your basic code and finished at least one successful run. You need to conclude what your current results are and provide the future optimization plan.

- Complete following sections, you can use the provided Report Template to address each part. Please notice that *Results* and *Discussion* in draft are different from that in final report.
  1) Introduction
  2) Scope of reproducibility
  3) Methodology
     - Model descriptions
     - Data descriptions
     - Computational implementation
     - Code
  4) *Results*: only preliminary results are enough. You can just provide one evaluation metric of your developed model without any hyper-parameter tuning. You even don't need to run all epochs on the full dataset, instead, show grader that your code could work.
  5) *Discussion*: discuss your current results and propose the continued optimization plan.
- Deliverables
  - A **PDF** no longer than **5 pages main content** + Appendix + References
  - Submit to **Gradescope** at team level, remember to include your teammates in submission.
  - Use your own words instead of description in original paper.

D. Final Report

At this stage, you should have completed the code development (including documents) and run multiple

experiments to test all the hypothesis in the paper. You should make assessment on the reproducibility of the paper with supportive evidence that obtained from your own experiments.

- Complete all sections in Report Template.
- Upload your well-documented codebase to Github, Gitlab, etc and attach the link at beginning of your report.
- Upload your presentation video in Youtube, OneDrive, Google Drive, etc and attach the link at beginning of your report.
- Deliverables
  - A **PDF** no longer than **8 pages main content** + Appendix + References. This file should contain the links to your codebase and presentation video.
  - Submit to **Gradescope** at team level, remember to include your teammates in submission.
  - Use your own words instead of description in original paper.

E. Presentation

In the presentation, you should prepare slides that clearly illustrate the main points in your work and the main results that you have obtained. The organization should be in parallel to your final report as much as possible, i.e., introducing the motivation and setup of the problem you are addressing, describe the methodology precisely. Comparing your reproduction attempts with what the paper showed. Make proper assessment on the reproducibility and explain why.

- Your presentation will be 5-8 minutes long. Practice your talk ahead of time. It should be fluent but not too fast to listen.
- It should delivered through online platform (Youtube, OneDrive, Google Drive, and any other places that grader can access easily).
- It is fine that the presentation is done by either both members or only one member.
- Good visuals are important here. Text should be in large font, figure and tables are captioned with description.
- Deliverables
  - Attach the video link in final report
  - Zip your slides with report and submit it on **Canvas** at individual level, i.e., both members should submit the zip.

V. Report Template

ML/DL in healthcare usually requires many data processing (ETL) and modeling, in project, you can use any open source packages to do it rather than invent wheels by yourself. A very pwoerful package is PyHealth (https://github.com/zzachw/PyHealth) , which is developed by Dr. Sun's group. It is managed on Github and integrates ETL and modeling. You are welcomed to fork or clone for your project development.

1) Introduction
   - A clear, high-level description of what the original paper is about and what is the contribution of it.
2) Scope of Reproducibility
   - List all hypotheses from the paper you will test and corresponding experiments you will run.
3) Methodology
   - Model description
     - Model architecture: layer number/size/type, activation function, etc
     - Training objectives: loss function, optimizer, weight of each loss term, etc
     - Others: whether the model is pretrained, Monte Carlo simulation for uncertainty analysis, etc
   - Dataset description
     - Source of the data: where the data is collected from provide the link if possible; if data is synthetic or self-generated, explain how.
     - Statistics: dataset size, cross validation split, label distribution, etc
     - How do you use the data: change the class labels, split the dataset to train/valid/test, refining the dataset
   - Computational implementation
     - Report the software and hardware implementation (What is your basic coding framework, PyTorch, Tensorflow, etc? What kind of CPU or GPU do you use?)
     - Report hyperparameters including learning rate, dropout rate, number of iterations, training time, etc.
   - Code
     - Which parts are developed by yourself? Which parts are referred from the codebase in original paper or other resources?
     - Provided link to your repo (Github, Gitlab, Bitbucket, etc). Your repo should include detailed documents (README file) telling readers:
       - Dependencies (which packages are required)
       - Download instruction of data and pretrained model (if applicable)
       - Functionality of scripts: preprocessing, training, evaluation, etc.
       - Instruction to run the code
4) Results
   - Report results for all experiments that you run:
     - specific numbers (accuracy, AUC, RMSE, etc)
     - figures (loss shrinkage, outputs from GAN, annotation or label of sample pictures, etc)
   - Comparison with the hypothesis and results from the original paper.
5) Discussion
   - Make assessment that the paper is reproducible or not.
   - Explain why it is not reproducible if your results are kind negative.
   - Describe "What was easy" and "What was difficult" during the reproduction.
   - Make suggestions to the author or other reproducers on how to improve the reproducibility.