CS598 Projects: Deep Learning for Healthcare

Jimeng Sun

***Abstract*—CS598 Deep Learning for Healthcare is a graduate-level course focusing on practical deep learning methods for health analytic applications. One big part of this course is to conduct a group project that understands, replicates and extends some recently published work in deep learning for healthcare. Specifically, you will choose one paper from the provided paper pool and attempt to reproduce their experiments.**

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

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

I. Introduction

DEEP learning and healthcare applications interact closely nowadays thanks to the advancement in electronic data capturing technology such as electronic health records, on-body sensors and genome sequencing, and the advancement in deep learning methods. This course covers deep learning (DL) methods, healthcare data, and applications using DL methods. Through (painful) homework exercises, you should have by now learned deep learning methods and acquired sufficient knowledge about healthcare data. We believe you are ready to take on the next level of challenges. That is, you are going to understand, replicate, and extend some recently published works in DL for healthcare. The final results of this project should be one of the following:

1) Reproduce the main experiments in a selected paper. Your report will assess the ease of reproducibility regarding the checklist we provide below. In addition, you should attempt at least one additional experiment that is not in the paper after having successfully reproduced the main results. For example, you could assess the sensitivity of the model to one or more hyperparameters or to the amount of training data, or measure the variance of the evaluation score due to randomness in initial parameters.
2) Report on your failed attempt to reproduce a selected paper's main experiments. Not all papers are easily reproducible. Failing to reproduce the exact results of the original paper is not necessarily a bad thing, as long as your experiments are rigorous. Your report should identify all the questions that would need to be answered to reproduce the experiments or discuss how the findings appear to be in error (if that is what you discover).

Both outcomes are acceptable and can earn full credit.

II. Teams

Each team can have at most **TWO** students. Individual projects are also allowed. All team members will receive the same grade on the project (except that one member contributes much less than the other, which will lead to a penalty on the project grades of the inactive member).

J. Sun is with Computer Science Department of University of Illinois, Urbana-Champaign e-mail: (see http://www.sunlab.org).

III. Paper Selection

You should first select **THREE** papers (for proposal) and then narrow down to **ONE** paper (for draft and final submission) from the provided paper pool [1]. The reason for this coarse-to-fine paper selection is to avoid issues such as no data access, and to increase the success rate of you reproducing at least one of the three papers.

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 the paper to reproduce:

- You should find the problem tackled in the paper interesting.
- 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 should not choose a paper that we implement in homework
- 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 the original paper can be your reference.

We have divided the papers into "easy" and "hard" based on our assessment on the reproducing difficulty. Both easy and hard papers can earn full credits. However, we will apply some **bonus points / multiplier** for students selecting hard papers (see Section VI).

IV. Timeline

Next, we summarize the timeline.

1) Team formation (1-2 students) & paper selection (3 papers), due on Mar 6
2) Project proposal (PDF), due on Mar 27
3) Project draft (PDF), due on Apr 17
4) Final submission (PDF + Presentation + Code), due on May 8

All due at 23:59PM Central Time on the due date. **We do not allow late submission.**

V. Deliverables

A. Project Proposal (Up to 4 pages write-up + Unlimited references)

You should first select **THREE** papers from the paper pool. For each of the selected paper, your proposal should

[1]Link to paper pool: https://docs.google.com/spreadsheets/d/1XAjEjCbFRPXc2S1RlJ0wS8FO3el7DVGjNxKku0nOFCA/edit?usp=sharing

answer the following questions about your project, in order to demonstrate you have thought carefully about the paper you are planning to duplicate, and in order to communicate your understanding of the work and its importance to someone who most likely has not read the paper:

1) Citation to the original paper
2) What is the general problem this work is trying to do? We are **not** asking for the specific approach, that's requested below. An example of a general problem is 'mortality prediction.' An example of a specific approach is 'using recurrent neural network and attention mechanism.' Do not copy the description in the paper – use your own rewording.
3) What is the new specific approach being taken in this work, and what is interesting or innovative about it, in your opinion?
4) What are the specific hypotheses from the paper that you plan to verify in your reproduction study?
5) What are the additional ablations you plan to do, and why are they interesting?
6) State how you are assured that you have access to the appropriate data.
7) Discuss the computational feasibility of your proposed work – make an argument that the reproduction will be feasible.
8) State whether you will re-use existing code (and provide a link to that code base) or whether you will implement yourself.

The proposal should be a PDF, but can take any form.

B. Project Draft (2-4 pages write-up + Unlimited references)

At this stage, you should narrow down to **ONE** paper from the three selected papers. You should have fully understand this final-selected paper and realized all experimental setups and details. You should also have developed your basic code and finished at least one successful run. For the draft, fill out sections from this template [2], replacing the instructions with actual content. For the draft, only complete the following sections, and write "TODO" for all other sections:

- Introduction
- Scope of reproducibility
- Methodology
  - Model Descriptions
  - Data Descriptions
  - Implementation
  - Computational Requirements
- Results - for the draft, results can be any valuable results. For example, results from a simple baseline model in the paper, from intermediate steps prior to the ultimate target task, or from a tiny subset of the dataset. All those followed by your own analysis can be used. Even if your current results are not as good as the ones in the paper, there must be analyses about what possible reasons and solutions/plans are

The draft should be a PDF using the template provided above.

[2]Link to template: https://www.overleaf.com/read/cgzyfxwtwfkv

C. Final Submission

The final submission includes the final report, presentation, and code for your final-selected paper.

1) Final Report (4-6 pages write-up + Unlimited references): This should follow the same format as the draft, but all sections should be filled in.

2) Presentation (Up to 4 minutes): You should prepare PPT slides that clearly illustrate the main points in your work and the main results. Good visuals are important here – the presentation should be eye-catching, clear, and self-contained. Assume your audience has the background given in this class but remember to spend considerable time introducing the motivation and setup of the problem you are addressing. You should also spend time comparing your reproduction attempts with what the paper showed as well as (if you were able to reproduce) the additional ablations.

Please upload a *unlisted*[3] video on YouTube of your presentation (demo by one representative or multiple students, set an access key if you want) to share with us. Audio-added PPT is not accepted. Put the YouTube link under the report title.

3) Code: Publish your code in a public repository (e.g. on GitHub, GitLab, BitBucket). Make sure your code are documented properly. A README.md file describing the exact steps to run your code is required. You can refer to the ML Code Completeness Checklist to write the README file and make sure your code submission is complete. See this blog post on best practices for reproducibility.

VI. Grading Scheme

The entire project has 100 points and will account for 40% of your class grade. The specific points are as follows:

- Proposal (10 Points)
- Draft (20 Points)
- Final Report (50 Points)
- Presentation (10 Points)
- Code (10 Points)

Note that we will apply some **bonus points / multiplier** for the students selecting hard papers. For example, to get full credits for the project, students selecting an easy paper need 100 raw credits while students selecting a hard paper may only need 90 raw credits. We will decide the exact bonus points / multiplier later.

VII. Conclusion

Best of luck on your project and reproducible data science rocks!

Acknowledgment

Part of the project instruction is adapted from CS662 Advanced Natural Language Processing at University of Southern California.

[3]See https://support.google.com/youtube/answer/157177