CSCI 662: Fall 2023

Class Project (and project proposal)
adapted from work by Noah A. Smith and Jesse Dodge

version of August 17, 2023

proposal due: Sep 18, 2023, 23:59:59 AoE
report version 1 due: November 3, 2023, 23:59:59 AoE
project presentation: In class, late November (as scheduled)
final report due: Dec 11, 2023, **10:00 AM PST**

Your goal: understand, replicate, and extend some recently published work. Specifically, you will choose one paper from ACL 2023, and attempt to reproduce its experiments. In some cases, this may be a straightforward task. In some cases, it may be impossible, for a range of reasons.

The Project

Your project will do one of the following:

1. Reproduce the main experiments in an ACL paper. Your report will assess the ease of reproducibility, with respect to the checklist we provide below. In addition, you should attempt at least one additional experiment that isn't in the paper, but that you are able to conduct 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 an ACL 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.
Some considerations in choosing a 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.
- There should be some actual model and experiment that you reproduce, that can be evaluated. A position paper is thus not a good choice. Similarly, a paper that defines some deterministic transformation of data but does no learning with the data will not be accepted. Another way of looking at this is, the reproducibility of the paper you select should be of general interest to the community. Purely analytical papers, for instance, are likely not as interesting to reproduce as more experimental or model-heavy ones.

- In many cases, the authors may have made code available; this may be a blessing or a curse. You should definitely peruse a paper's codebase before deciding on that 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 will have had access to infrastructure that is way out of your budget; don't choose such a paper. What resources are available to you?
  - If you request it from course staff, you can have free access to the CARC Discovery Cluster which has a lot of V100s but often a long line – make sure you start your project early and allow one week or more to run jobs!
  - Microsoft Azure grants $100 of credit to students. However you can't use the good V100 GPUs (but you can use 18 older K80 GPUs).
  - Google Colab is a good free resource for educational research. However, the GPU's may be older and access may be limited.
- You may not reproduce a paper that you, course staff, or someone in your current or former lab has written.
- Your paper must be an 'acceptable paper' (see below) which is artificially constrained for the purpose of this project.

Acceptable Papers

For this exercise, please choose a paper that interests you from the main proceedings[1] of ACL 2023 (i.e. "Volume 1" or "Volume 2"), that you might want to use in the course project. Make sure the paper is under the heading "Volume 1: Long Papers" or "Volume 2: Short Papers." No other sections are valid. None of Volumes 3–6, "Findings" papers, or workshop papers are valid.

You may find it easier to select a paper of interest by looking at the conference schedule,[2] which is divided by theme. You should only look under July 10, 11, or 12. Avoid papers that have [SRW], [Demo], [TACL], [CL], or anything in square brackets in the title. Avoid anything in the "Spotlight" sections on July 10. The paper will also be in Volume 1 or 2 in the proceedings, if it is a valid paper (see above).

Advice

Well before the version 1 deadline, it would be useful for your group to put together a template which has fields for all of the information you want to record from every experiment. (This could be in the form of a spreadsheet, or a rough placeholder draft with information that you fill in as you go, etc.) You should design this so that once it's filled out for all of your proposed experiments, you will meet the overall requirements for the project (described at the end of this document). For example, a group which is going to show training curves as their additional experiment would want a template which has fields for at least:

- training curves showing dev evaluations every X training steps
- learning rate, batch size, dropout, size of their model, etc. (they shouldn't have "etc." here, they should be very specific)
- clear connection showing which hypothesis this experiment will support
- info on the data
- computational requirements

(In other words, the list that you construct should consist of the items from the numbered list in the project report description that are specific to your project, with one template filled out for each experiment you run, including baselines.)

[1] `https://aclanthology.org/events/acl-2023/`
[2] `https://virtual2023.aclweb.org/sessions.html`

Teams

You must team up, and your team should be composed of two people. If you are a PhD student actively researching NLP, you should team up with either a PhD student who is not actively researching NLP or with someone who is not a PhD student. If this becomes impossible for some reason please let me know. The goal is to work with someone with a different perspective from yours, which will give you new opportunities to learn. Both team members will receive the same grade on the project (before late day penalties). Late submissions will apply to both team members, so if e.g. the proposal is one day late, student A has one late day remaining, and student B has no late days remaining, student A will lose their last late day, and student B will receive a 20% deduction on the proposal. teams must be formed by the time of proposal submission and cannot be changed without good reason and explicit permission from course staff.

Required Contributions

Your team will make the following contributions (see due dates at the top of this document):

1. A one-page proposal
2. Version 1 of the final report
3. An in-class presentation, scheduled on the last week of classes
4. A final report

1 Proposal

Your proposal should 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 (e.g. course staff):

- Citation to the original paper (this isn't a question, but you still have to provide it)
- 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 'predict the sentiment of tweets.' An example of a specific approach is 'use semantic parses and a graph convolutional network.'? Do not use jargon. Do not copy the description in the paper – use your own rewording.
- What is the new specific approach being taken in this work, and what is interesting or innovative about it, in your opinion?
- What are the specific hypotheses from the paper that you plan to verify in your reproduction study?
- What are the additional ablations you plan to do, and why are they interesting?
- Briefly state how you are assured that you have access to the appropriate data including, where relevant, data splits.
- Briefly state whether you will re-use existing code (and provide a pointer to that code base) or whether you will implement yourself.
- Discuss the computational feasibility of your proposed work – make an argument that the reproduction will be feasible.

The proposal should be a PDF, but can take any form. It should be no longer than one page.

2 Report Version 1

Fill out sections from this template,[3] replacing the instructions with actual content. For version 1, only complete the following sections, and write "TODO" for all other sections:

- Introduction
- Scope of reproducibility
- Addressed claims
- Methodology
  - Model descriptions
  - Data descriptions
  - Implementation – in version 1 only mention whether you plan to use existing code or implement yourself.
  - Computational requirements – in version 1 an estimate will suffice

3 Final Report

This should follow the same format as Version 1, but all sections should be filled in.

4 Presentation – in class, to be scheduled

Your presentation will be 12 minutes long (+2 for questions),[4] and delivered during the last few sessions of class (we will make schedules once the groups are decided). You should prepare slides (powerpoint, beamer, keynote are good choices) 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 present the work much as you presented the EMNLP paper in class. However 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.

- It's ok for one member of your group to give the entire presentation, but both members must attend the presentation and both are responsible for answering questions. I will ask questions of both group members, as should others in the class.
- No text should appear on slides that is less than 16 point font. If you need to make it smaller, you need to make two slides or simplify the current one.
- Practice your talk ahead of time; make sure it is the correct length (add or subtract as necessary; you will be cut off if you go too long) and make sure you can speak clearly.
- If you don't know the answer to a question, say 'I don't know.' This is very hard for us to do but it is ok to not know everything and to admit it.

[3] That's clickable but if you are reading hard copy, the URL is `https://www.overleaf.com/project/64dd1646f9124c536bf65cf0`.

[4] Exact timing subject to change; whatever is announced in lecture/on piazza supersedes this info.

5 Asking Questions about Others' Reports

You should ask questions about each others' reports. Clarification questions, probing questions, suggestions for future analysis, etc. are all good types of questions to ask. As such you'll be required to ask questions in two talks (that are not your own) This will establish that you're paying attention to the talks and that the talks are understandable. These questions will contribute to your class participation grade, not your project grade.

6 Notes on Grading

The entire project will account for 40% of your *class* grade, divided as follows:[5]

6.1 Proposal – 5%

You need to have all eight items listed in the project instructions (citation to the original paper, general problem, specific approach, hypotheses to be tested, ablations planned, description of how you will access the data, whether you will use the existing code or not, a discussion of the feasibility of the computation). For this component we won't judge quality unless you put complete nonsense for an item (then it will be considered 'missing') but we will give you feedback if we think there are problems.

- -0.5% for each item that is missing
- -0.5% if over 1 page
- -0.5% if not in PDF format

6.2 Report Version 1 – 5%

You need to use the report template and fill out the following sections, each of which we will score based on the clarity and appropriateness of your writing (percentage of total grade for each component shown):

- Introduction (1%)
- Scope of reproducibility (1%)
- Methodology:
  - Model descriptions (0.5%)
  - Data descriptions (0.5%)
  - Implementation – in version 1 only mention whether you plan to use existing code or implement yourself. (0.5%)
  - Computational requirements – in version 1 an estimate (more comprehensive than the proposal argument) will suffice. (0.5%)

Additionally you can lose up to 1% by not using the correct template, not putting in placeholders for the other sections, or for other aspects of your report that do not clearly convey your intentions and/or progress.

[5] We reserve the right to adjust the point distribution even after this has been distributed.

6.3 Presentation – 10%

We expect a well-timed, well-presented presentation. You should clearly explain what the original paper is about (what the general problem is, what the specific approach taken was, and what the results claimed were) and what you encountered when you attempted to reproduce the results. You should use the time given to you and not too much (or too little). You should be able to competently answer questions and confidently state when you don't know something. You can lose points on this part as follows:

- -0.5% per minute over time (max -2%)
- -0.5% per two minutes under time (max -2%)
- -1% substantial amount of written material less than 16 point font
- -2% (max) Do not explain the general problem clearly
- -2% (max) Do not explain the specific approach taken in the paper clearly
- -2% (max) Do not explain reproduction attempts clearly
- -1% (max) do not answer questions clearly

6.4 Final Report – 20%

Note: The maximum total is 20%, but there are more percentage points than that available. In this way, you can get a full score on this part, even if you miss some points.

In the discussion below, each point is worth 0.2% of your final grade. For this 20% of your final grade, if you receive n points in the below section, you will thus receive $\min(0.2n, 20)$ percentage points credit.

6.4.1 Introduction (5pt)

A clear, high-level description of what the original paper is about and what is the contribution of it (5pt)

6.4.2 Scope of reproducibility (15pt)

- Formatting (7pt)
  - Full score only when the hypotheses tested in your report are written as 'lists' (7pt)
  - They are written as a paragraph but are clear enough (3pt)
  - They are written as a paragraph and are not clear enough (1pt)
- Content (8pt)
  - At least one of the hypotheses is a central claim in the paper, and all hypotheses have an experiment that supports it.
  - No hypothesis is a central claim in the paper (-4pt deducted)
  - There is a hypothesis that is not experimented in the report (-4pt deducted)

6.4.3 Methodology (45pt or 55pt)

- Model description (10pt)
  - -3pt deducted for any missing items, -2pt deducted for described but unclear items
    - Model architecture: 3pt
    - Training objective: 3pt
    - # of parameters: 3pt
    - Other important details, such as which pretrained model is used, etc
- Dataset description (5pt)
  - -2pt deducted for any missing item:
    - Citation or link is provided
    - Source of the data (e.g. if they are annotated, brief description of how)
    - Statistics (dataset size, dataset split, label distribution, etc)
    - You split the dataset to train, valid and test (for example, if you do not have a validation data, no point)
- Hyperparams (5pt)
  - Report at least 3 types of hyperparameters including Learning rate, dropout, hidden size, etc (5pt) [if one of these is not relevant to the model, e.g. a non-neural model has no hidden size, it need not be mentioned]
    - Reported 2 types: 2pt
    - Reported 0 or 1 type: 0pt
    - Enough hyperparams mentioned but some of the three critical ones missing: -1pt per missing
  - Miss crucial hyperparameter in the paper (0pt)
- Code (10pt or 20pt)
  - If own code is written (20pt)
    - Provided link to their github repo (5pt)
    - Documented and easy to use (15pt)
    - Deduct -2pt for each missing item
      - Dependencies
      - Data download instruction
      - Preprocessing code + command
      - Training code + command
      - Evaluation code + command
      - Pretrained model (if applicable)
      - Table of results (no need to include additional experiments, but main reproducibility result should be included)
  - If existing code is used (10pt)
    - Link to the original paper's repo (2pt)
    - Additional instructions to reproduce the code or to run extra experiments (8pt)
      - Deduct -0.5 for each missing item from the above list
      - This means, even if there is existing code but some commands or small components are missing, you will have to write them.

  - It is possible to have a case that is somewhere between two (e.g. there is existing code but some significant parts, like connective scripts, data preparation, etc. are missing). In that case, we will use our best judgement and grade you somewhere between these two categories.

- Computational requirements (10 + 5pt)
  - For each type, 0.5pts for estimating requirement from the original paper, 0.5pts for reporting the actual information (Max 10pt)
    * Type of hardware
    * Avg runtime for each epoch
    * Total number of trial
    * GPU hrs used
    * # training epochs
    * Anything else that has significant impact on the resource requirements (for example, some papers may have bottleneck on CPU hours, RAM or disk memory)
  - 5pt if you discuss what factors lead to requiring more resources than estimation and what efforts you have made to reduce the requirement

6.4.4 Results (35pt)

- Reproducibility results (15pt)
  - Report results for all experiments that support the claims that are being tested (5pt)
    * You do not get a point if specific numbers are not included.
  - Indication of the result (10pt)
    * Discuss with respect to the hypothesis is clearly described (5pt)
    * Discuss with respect to the results from the original paper is clearly described (5pt)
      · 0pt if you are comparing with experiments in the original paper that are not comparable without specifically discussing it.

- Experiments beyond the original paper (max 20pt):
  - Credits for each ablation depend on how hard it is to run the experiments
  - Additional dataset (max 10pt)
    * Additional data may be in the same task or in a different task
  - Explore different methods (max 10pt)
    * Methods could be model architectures, training objectives, new ways of probing the model, etc
    * For each exploration, discussions on what it indicates should be included
  - Add new ablations (max 10 pt)
    * Ablations could be varying the size of the training data, including/excluding some component of the model to see their effect, etc.
    * For each new ablation, discussions on what it indicates should be included
  - Hyperparameter tuning (max 5pt)
    * For each hyperparameter tuning, discussions on what it indicates should be included
  - Any other reasonable ablations/analyses eligible for credits.

6.4.5 Discussion (10pt)

- Larger implications of the experimental results, whether the original paper was reproducible, and if it wasn't, what factors made it irreproducible. (5pt)
  - If one of "What was easy" or "What was difficult" is missing, 3pt
  - If both of "What was easy" or "What was difficult" is missing, 0pt
- A set of recommendations to the original authors or others who work in this area for improving reproducibility. (5pt)

6.4.6 Others

- -10pt if the report exceeds page limit (8) excluding references.

Rules

- Unlike other assignments, you can (in fact, you must) work in groups.
- Unlike other assignments, you can use others' code where relevant but you must cite this usage.
- You may not submit writing that is not your own, or writing that is your own (or others') that has been prepared for a previous class or paper.
- The coding and experiments that you do on this project must not have been done prior to the class or for any other class, either as homework or project, by you or anyone else.
- Failure to follow the above rules is considered a violation of academic integrity, and is grounds for failure of the assignment, or in serious cases failure of the course.
- We use plagiarism detection software to identify similarities between student assignments, and between student assignments and known solutions on the web. Any attempt to fool plagiarism detection, for example the modification of code to reduce its similarity to the source, will result in an automatic failing grade for the course.
- Generative language, code, and vision models (e.g. ChatGPT, Llama 2, Midjourney, Github Copilot, etc.; if you are unsure, ask and don't assume!!) can be used with the following caveats:
  - You must declare your use of the tools in your submitted artifact. If you don't declare the tool usage but you did use these tools, we will consider that plagiarism
  - For code and image generation, you must indicate the prompt used and output generated
  - For text generation you must provide either a link to the chat session you used to help write the content or an equivalent readout of the inputs you provided and outputs received from the system. You will lose credit if "the AI" is doing the work rather than you.