COMP 599 & LING 484/782 Reproducibility Project Guidelines

Most PhD/MSc admissions or research positions in the industry require a publication. Given that you are at McGill, the expectations from recruiters will be even higher. This is your chance to showcase your research potential.

By reproducing a work, it shows your potential in building systems based on previous works and that are useable by others; for example, the *Transformers* library is one of the most widely used, but a lot of the work that went into it was to reproduce existing papers (such as BERT, T5, etc.). This skill is valuable whether you want to be an engineer or a researcher that extends current methods to create new approaches, that applies existing methods in a new scenario or domain, or to build more efficient and optimized algorithms for existing tasks and data.

Form: Please fill out this form as soon as possible (before deadline) to indicate your intention: https://forms.gle/EMNd4VbjHG47bKgGA

Template: Project reports should use this MLRC 2022 template (click Menu -> Copy Project to copy): https://www.overleaf.com/project/62fcf81c144fd47671a39284

Source of papers: In this type of project, you will aim to reproduce the results from a recent paper published in an NLP conference. You must select a paper to reproduce which has been published in one of the following conferences (main conference or in the Findings): ACL, EMNLP, NAACL, COLING, TACL, EACL, AACL. If you are interested in reproducing a paper which has not been published in one of those conferences, please contact the TAs.

Resources: Make sure that the paper you select is reasonable to reproduce given your compute budget (200$ Azure credits, Kaggle 30h of GPUs per week, Google Colab, McGill teaching GPUs).

You can do two types of reproductions. We describe each below.
- **Reproducing from scratch:** For this type, you are reproducing the results (or a subset of the results) from the original paper completely from scratch. Ideally, you will not reference the code released by the authors at any point during this process.
- **Reproducing from released author code:** For this type, you will reproduce the results from the original paper (or a subset of the results) using the authors' released code. In this case, it is not sufficient to merely re-run the authors' code. You must extend the project in some way. For example, you can run additional experiments which were not investigated in the original paper. More plainly, your project **must** provide some additional insight *beyond* what is presented in the original paper.

Submission to challenge: We encourage you to submit your results (whether they are positive or negative) to the *Reproducibility Challenge* organized by Papers With Code. Please read the description here as it will be considered for grading: https://paperswithcode.com/rc2022/task
You will have the opportunity to publish your submission in the Reproducible Science journal (rescience.github.io) and present it in the poster sessions at NeurIPS (details to be determined).

Inspiration: To get an idea of what other students have done, check out the list of schools that participated in the previous edition of the challenge: https://paperswithcode.com/rc2021
We also highly encourage you to take a look at the best and outstanding papers to have an idea of what is considered a good submission.

Milestone 1: Proposal + Evaluation

This should have the following sections:
- Title
- Abstract
- Introduction
- Related Works
- Appendix (Dataset and evaluation metrics)

Note: The instructions for the Milestone 1 are mostly the same as the main project; you can find the instructions on MyCourses in the Content tab > Project > Project Guideline. There are a few differences, which can be found below.

Abstract:

In 200 words, please provide the following information:
1. Cite the paper you are reproducing and include the name in the first sentence.
2. State whether you are going to do a reproduction from scratch or, if you will use the authors' codebase.
3. Describe the scope of reproducibility. In other words, what results from the original paper are you aiming to replicate?
4. Describe, at a high-level, what additional experiments or analysis you will perform? If you are doing a reproduction from scratch, you are not required to carry out any additional experiments or analysis.

Introduction:

Your introduction should also consider the following:
- Describe the high-level methodology employed in the paper you are reproducing
- Discuss the limitations of the paper on a practical level (such as code implementation, variance across different runs of an experiment due to randomness, complexity of implementation, etc.)

- Explain the value of reproducing this work. For example:
  - If the work can already be used in Huggingface's Transformers and others have successfully reproduced the results, do you need to reproduce it again?
  - If the codebase is already in PyTorch, what value do you add to the project by rewriting it in PyTorch (either from scratch, making it more modular, etc.) rather than rewriting it in a different framework like Jax?
  - Is your work more modular and easier to extend or use than the original implementation (if the original implementation was available)?
- Describe, at a high-level, what additional experiments or analysis you will perform. If you are doing a reproduction from scratch, you do not need to provide this information.

Related Work:

The paper you are reproducing already has related works. Rather than just copying, focus on the relevant papers they cited (you may ignore non-relevant related works) from the perspective of the challenge.

You should also include papers that were not cited by the work you are reproducing (for example, works published after/concurrently, or that were simply missed by your paper).

Appendix: Datasets and evaluation metrics:

Please follow the main project proposal.

Final Milestone: Reproduced methodology, results, discussions

This milestone includes Milestone 1, plus the sections below.

Scope of reproducibility:

Please refer to the [MLRC template](#) for more details. Also, describe which parts of the paper you are going to replicate. Also, mention whether you are going to do a replication from scratch or if you are going to use the authors' released codebase.

Introduction:

Write an expanded form of your introduction from Milestone 1 (with modifications if needed) with the following additional information:
- Describe, at a high-level, your findings. Were you able to reproduce the results presented in the original paper?

- Discuss at a high-level, the takeaways from the additional experiments and analysis you performed.
- Please refer to the [MLRC template](#) for supplementary information.

Methodology:

A methodology section may vary. But the following sections will be needed (instructions can be found in the [MLRC template](#)):
- Model descriptions
- Experimental setup and code
- Hyperparameter
- Computational requirements
- Dataset
- Evaluation metrics

In addition, if you plan to perform any additional experiments or analysis, please describe it in this section. You should also clearly motivate *why* you are performing these experiments/analysis.

Dataset:

When reproducing a paper, it is sometimes necessary to also reproduce the dataset. However, this may not always be easy as the dataset might not be public, or the preprocessing is unclear. In a subsection called "Dataset", you should describe the challenges you encountered and how you overcame them in order to reproduce the work.

If the dataset wasn't originally available, you should also release the dataset and the procedure for creating them (as it might be useful for others), or if they can't be publicly released, you should explain why and propose ways for other researchers to access them. You can release them as a GitHub release (0-2GB), Zenodo dataset (2-50GB), or Kaggle Dataset (50GB-100GB), or an [academic torrent](#) (100GB+); make sure to include the license (be as permissive as you possibly can with respect to the original dataset).

Results:

Describe the results you obtained. This section must contain the following two subsections (details can be found in the [MLRC template](#)):
1. **Original experiments:** Your replicated results for the main paper. Where applicable, you should contrast the results *you* obtained directly to those presented in the original paper. For example, if the original paper presented a plot containing their results, you could include the same plot in your paper, but with the results you obtained. Similarly, if the original paper provided a table with their main results, you should reproduce this same table in your paper (and compare your results to theirs).

2. **Additional experiments:** Results for any additional experiments or analysis you performed. Follow the same guidelines for these experiments as is described in the normal project proposal. For an example, look at [page 6 of the best paper](#) at last year's challenge. An inspiration might be the future work described in the discussion/conclusion of the paper you are reproducing. Also, if you are aware of subsequent/concurrent papers that criticize your reproduced paper because they claim they fail in a novel scenario, you can add experiments in that scenario that were not originally covered.

Discussion:

In this section, you will discuss your findings. In particular, you will need to:
1. Describe whether your reproduction was easy or difficult by writing a section called "What was easy" and another called "What was difficult?"
2. In the case of it being difficult, what were the challenges you faced and how did you address them?
3. Discuss whether you were able to replicate the results presented in the original paper.
4. If you performed additional experiments or analysis, discuss your findings.
5. What recommendations do you have for future researchers trying to use the work you reproduced?
6. [Follow the instructions in the MLRC template](#)

As you reproduce the work, you might find out that some results did not match the original approach, or that you are having a hard time reproducing certain steps based on the original work. In this case, please include a subsection called "Communication with original authors" where you describe your attempt to contact the original authors (e.g. through the email listed on the paper).

Code:

Release: Since this is a reproducibility challenge, it is extremely important to release your code in a way that is usable. Giving instructions that only runs on a specific setting (e.g. a personal cluster running a specific version of SLURM, and you are also loading specific files from your personal directory) is not reproducible.

Repository: You will release your code on GitHub, and make it public only at the end of your project (note: do not copy and paste solutions from your assignment, but you may reuse what you learned and refine it beyond what was in the assignment). You must include all the instructions either in the readme, or in a separate file that is linked in the readme if it's too long. Make sure to indicate what the TAs should be reading as you may have instructions beyond the scope of our grading. To have an idea of what you should include, please refer to this checklist: https://github.com/paperswithcode/releasing-research-code

Language: The code should be written in Python (if you think another language is more appropriate, such as Julia, C++, Rust; please consult the TA before you start coding).

Package: If needed, you will release a package with instructions for others to run your code, extend it, and possibly use it as a standalone library. If you are unfamiliar with python packaging, start here: https://packaging.python.org/en/latest/tutorials/packaging-projects/

Contributions:

If your team has multiple people, please add a Contributions section explaining the contributions of each individual. It is totally fine if multiple people have worked on the same parts of the project. For example, Tom did most of the work with experiments on blah blah, whereas Jerry did the data collection and writing etc.