Editorial / Machine Learning

ICLR Reproducibility Challenge 2019

Joelle Pineau[1,2,3], Koustuv Sinha[1,2,3], Genevieve Fried[1], Rosemary Nan Ke[2,3]Hugo Larochell[4]

[1]School of Computer Science, McGill University, Montreal, Canada – [2]Montreal Institute of Learning Algorithms (Mila), Montreal, Canada – [3]Facebook AI Research (FAIR), Montreal, Canada – [4]Google Brain, Montreal, Canada

Welcome to this special issue of the ReScience Journal, which presents results of the 2019 ICLR Reproducibility Challenge (2nd edition). One of the challenges in machine learning research is to ensure that published results are sound and reliable. *Reproducibility*, that is obtaining similar results as presented in a paper, using the same code and data (when available), is a necessary step to verify research findings. Reproducibility is also an important step to promote open and accessible research, thereby allowing the scientific community to quickly integrate new findings and convert ideas to practice. Reproducibility also promotes use of robust experimentation workflows, which can potentially reduce unintentional errors.

The Challenge – In support of this, the goal of this challenge was to investigate reproducibility of empirical results submitted to the 2019 International Conference on Learning Representations (ICLR). Primarily, the aim was to assess if the experiments reported in a paper are reproducible, and to determine if the conclusions of the paper are supported by the findings of the reproducibility report. The role of each challenge participant was to be an inspector verifying the validity of the experimental results and conclusions of the paper. Entry into the challenge was open to all, either individually or in a team. Several graduate-level machine learning courses in universities around the world incorporated this challenge as a final course project for their students. In total, 26 teams, from 10 universities, submitted reports investigating 26 ICLR submissions.

Replicability and Reproducibility – Reproduction of a computational study means running the same code, using the same input data, and then checking if the results are the same, or at least "close enough" when it comes to numerical approximations. This is most easily achieved when the code and data are openly shared. Alternately, the methods described can also be implemented/re-implemented according to the description in the paper, which promotes *Replicability*. This is a higher bar than reproducibility, and may be helpful in detecting anomalies in the code, or shedding light on aspects of the implementation that affect results. In the absence of code, several aspects of a paper can influence the ease with which the results can be replicated, as listed in theReproducibility Checklist.

Baselines need attention – It is sometimes not feasible to reproduce all the experiments in a paper: factors such as private datasets, extensive training time, requirement of extensive or non-standard computing infrastructure can all limit reproducibility. In those cases, challenge participants were encouraged to reproduce results from baseline methods. It is sometimes the case that baseline methods are not properly implemented, or hyper-parameter search is not done with sufficient care, leading to poor comparison

of alternative methods. Reproducing the baselines can be as impactful as reproducing the main technical contributions of a paper, and therefore this was encouraged in this challenge.

Relationship with Authors – Authors of research papers have as much to gain from this challenge as the participants. We encouraged participants to communicate with the authors to clarify various nuances of the open source implementation or to communicate the choice of hyperparameters in their algorithmic implementation. In the 1st edition of the challenge, we found that this helped several authors improve the quality of their work and paper. During the review period, communication between authors and challenge participants was done through the open comment platform on Open Review.

Publication medium – Challenge participants were encouraged to prepare a written report of their reproducibility study, for submission to this special issue. *ReScience* provides the perfect platform for publication of reproducibility efforts. ReScience lives on GitHub where each new implementation of a computational study is made available together with comments, explanations and tests. This exactly aligns with our philosophy and goal from the challenge, which also lives in Github on its own repository consisting of submissions and the reviewer comments. We received 26 submissions in total, of which 4 reports are chosen to be published in this journal, following a single-blind review process.

Content – The 4 reports in this special issue were selected for their high standard of scholarship, including clear explanation of the scope and objectives, care in the methodology, clarity of explanations, thoroughness of results, and insightfulness of the findings. In the report on *Learning Neural PDE Solvers with Convergence Guarantees,* the authors perform robust experiments on the proposed approach and provide a well documented codebase and Jupyter notebooks for quick replication. In the report on *Variational Sparse Coding,* the authors extend upon the codebase released along with the original paper and perform rigorous experiments, while validating the hyperparameters used by effective communication with the authors of the main paper through OpenReview, the conference management system used by ICLR. In the report on *H-detach*, the authors went one step further to propose their own faster implementation based on low level CUDA binaries in order to improve training. Finally, the report on *Meta learning with differentiable closed form solvers,* the authors provide thoughtful discussions on the repercussions on reproducibility and fairness of comparison with prior literature of the choice varying the number of classes at training time, which points to the care and attention to detail employed by the authors in this work.

Conclusion – Reproducibility in machine learning has recently garnered a considerable amount of attention and momentum thanks to key efforts by top researchers. Conferences such as ICLR, AAAI, ICML have organized dedicated workshops on the topic. The premier conference in the field, NeurIPS, has recently adopted the pledge of reproducibility as part of their submission process. We hope our endeavour will similarly spur more efforts in reproducing existing ideas and papers, and in turn promote open, accessible and sound machine learning research.

Reviewers – Many thanks to all our reviewers who spent their precious time to critically review the reports. We acknowledge your hard effort and hope that you will keep supporting us in this endeavour in the future!

- **Andrew Jaegle**, University of Pennsylvania

- **Arna Ghosh**, McGill University

- Chaochao Lu, University of Cambridge

- Ishan Durugkar, University of Texas at Austin

- Jiahui Yu, University of Illinois, Urbana Champaign

- Joelle Pineau, Facebook AI Research / McGill University

- Joey Bose, Mila / McGill University

- Koustuv Sinha, Mila / McGill University / Facebook AI Research

- Lovedeep Gondara, Simon Fraser University

- Maneesh K Singh, Verisk

- Martin Jaggi, École polytechnique fédérale de Lausanne

- Malik Altakrori, Mila / McGill University

- Melanie Fernandez Pradier, Harvard University

- Michela Paganini, Facebook AI Research

- Mido Assran, Facebook AI Research / McGill University

- Nicolas Gontier, Mila / Element AI

- Noe Casas, Universitat Politecnica de Catalunya

- Olexa Bilaniuk, Mila / Université de Montréal

- Pablo Samuel Castro, Google Brain

- Peter Henderson, Stanford University

- Rosemary Nan Ke, Mila / Facebook AI Research

- Ryan Lowe, Mila / McGill University / Facebook AI Research

- Shagun Sodhani, Mila / Université de Montréal

- Seungjae Ryan Lee, END-TO-END AI, Princeton University

- Xavier Giro, Universitat Politecnica de Catalunya