

NeurIPS 2019 Reproducibility Challenge

Koustuv Sinha^{1,2,3, }, Joelle Pineau^{1,2,3}, Jessica Forde⁴, Rosemary Nan Ke^{2,5}, and Hugo Larochelle⁶

¹McGill University, Montreal, Canada – ²Montreal Institute of Learning Algorithms (Mila), Montreal, Canada – ³Facebook AI Research, Montreal, Canada – ⁴Brown University, USA – ⁵Polytechnic University, Montreal, Canada – ⁶Google Brain, Montreal, Canada

Edited by
Nicolas Rougier

Received
20 March 2020

Published
–

DOI
–

One of the challenges in machine learning research is to ensure that the presented and published results are sound and reliable. Reproducibility, which is obtaining similar results as presented in a paper or talk, using the same code and data (when available), is a necessary step to verify the reliability of 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 the use of robust experimental workflows, which potentially reduce unintentional errors. In 2019, the Neural Information Processing Systems conference, the premier international conference for research in machine learning, introduced a reproducibility program [1], designed to improve the standards across the community for how we conduct, communicate, and evaluate machine learning research. One of the components in the program consisted of a community-wide reproducibility challenge on the accepted papers. In this special issue of the ReScience C Journal, we present the top peer-reviewed submissions of the challenge, namely 2019 NeurIPS Reproducibility Challenge.

1 The Challenge

The goal of this challenge was to investigate the reproducibility of empirical results submitted to the 2019 edition of Neural Information Processing Systems (NeurIPS) conference. Unlike our previous editions (2018 ICLR, 2019 ICLR), in this challenge, we only focused on accepted papers at the conference. The primary target audience of the challenge was early career researchers from universities, however, we received participation from the industry as well. The main objective of this challenge was to provide independent verification of the empirical claims in accepted NeurIPS papers and to leave a public trace of the findings from this secondary analysis. We provide a comparative analysis of participation of this challenge as compared to the previous editions in Table 1. A total of 173 papers were claimed for reproduction, which is a 92% increase since the last edition. We had participants from 73 institutions (63 universities and 10 industries) from around the world. Institutions with the most participants came from 3 continents and include McGill University, Canada, KTH in Sweden, Brown University in the US and IIT Roorkee in India. In those cases (and several others), a high participation rate occurred when a professor at the university used this challenge as a final course project. In this special issue, we present the top 10 peer-reviewed reports, selected from 84 submissions.

Copyright © 2020 K. Sinha et al., released under a Creative Commons Attribution 4.0 International license.
Correspondence should be addressed to Koustuv Sinha (koustuv.sinha@mail.mcgill.ca)
The authors have declared that no competing interests exist.

Conference	# papers submitted	Acceptance rate	# papers claimed	# participating institutions	# reports reviewed
ICLR 2018	981	32.0	123	31	n/a
ICLR 2019	1591	31.4	90	35	26
NeurIPS 2019	6743	21.1	173	73	84

Table 1. Participation in the Reproducibility Challenge. Source for number of papers accepted and acceptance rates: <https://github.com/lixin4ever/Conference-Acceptance-Rate>

2 Baselines, Ablations and Replications

Replication of a computational study typically means running the same code, using the same input data, and then checking if the results are the same or at least “close enough” by some degree of numerical approximations¹. This is most easily achieved when the exact code and data to replicate the experiments are provided. To this end, the organizers of the 2019 NeurIPS conference instated a code submission policy for the accepted papers this year. While it wasn’t mandatory, the policy was to encourage authors to submit their code by providing enough flexibility on the timing of submission. This resulted in 74.4% of papers being associated with their code, which was less than 50% in the 2018 NeurIPS conference. From the very beginning of the challenge, we made these codebases available to participants and offered three tracks to choose from.

1. **Baselines Track** - Sometimes it is not feasible to reproduce all the experiments in a paper: factors such as private datasets, extensive training time, the requirement of non-standard compute infrastructure can all limit reproducibility. It is also sometimes the case that baseline methods reported in the papers are not properly implemented, or hyper-parameter search is not done with sufficient care, leading to a poor comparison of alternative methods. Thus we provided an option to the challenge participants to perform a rigorous analysis on the baselines by re-implementing them wherever necessary. Reproducing the baselines can further add to the technical contributions of a paper, and therefore was encouraged in this challenge.
2. **Ablations Track** - Since we had almost 75% of accepted papers accompanied with their code, we provided a track which only focuses on the released code. Participants are encouraged to use the authors’ code and perform rigorous ablation experiments by modifying the model and hyperparameter choices, to gain extra insights from the reported methods of the paper and add value to their understanding.
3. **Replications Track** - A higher bar of reproducibility is to replicate the experiments explained in the paper from scratch without having to refer to the original codebase. This is helpful in detecting anomalies in the presentation of the ideas in a paper, and it sheds light on the aspects of the implementation that could affect the final results. This is far by the most difficult track, and the implementation results directly add the most value to the understanding of the original paper, often leading to continued discussions with the authors.

3 Platform and Medium

In this edition of the Reproducibility Challenge, we were fortunate enough to have big support from OpenReview² and the Program Chairs of NeurIPS 2019. All NeurIPS 2019

¹We have used the definition of “Replication” in the NeurIPS 2019 Reproducibility Challenge following the old definition adopted by ACM [2]. Recent changes to this definition has been proposed by [3] and [4], which states the above definition is suitable for the term “Reproduction” instead.

²https://openreview.net/group?id=NeurIPS.cc/2019/Reproducibility_Challenge

accepted papers were hosted by OpenReview, which facilitated online discussions for the larger research community who were unable to be present physically at the conference in Vancouver in December 2019. OpenReview built a unique platform for the Reproducibility Challenge, which featured the accepted papers as well as allowed challenge participants to claim a paper to work on, and later submit their reports based on their claim. Once submitted, all reproducibility reports underwent an extensive review cycle by a large set of reviewers of the NeurIPS 2019 conference. Due to the transparent review process of OpenReview, many reproducibility reports attracted comments from the original authors, which in turn helped the overall reviewing pipeline. Finally, we selected 10 high-quality reports from 84 submissions to be published in this journal, ReScience C, which is a perfect platform for publication of reproducibility efforts of various computational fields of science.

4 Relationship with Authors

Authors of research papers have much to gain from this challenge as the participants. Using OpenReview, we encouraged participants to clarify various nuances of the implementation of the paper with the original authors. Due to the dual nature of our OpenReview platform, challenge participants could easily communicate with the authors who themselves received notifications from the comments arising in the forum associated with their papers. During the review of the Reproducibility reports (in preparation for this special issue), these communications were also taken into consideration by the reviewers in judging the quality of the report.

5 Computing Resources

In this challenge, we partnered with CodeOcean³ for providing free cloud computing credits to select teams. CodeOcean is an online web-based platform for reproducible computational science, which is a shareable Docker container living in the cloud. Participants were able to leverage the free compute resources from CodeOcean to run their experiments. CodeOcean provided prompt and necessary support enabling participants to resolve implementation issues to request additional resources to support their experiments.

6 Content

In this special issue, we present the top 10 peer-reviewed reports of the 2019 Reproducibility Challenge. These reports were selected after critical reviews from our reviewers, and consist of reproducibility efforts over broad coverage of topics in Machine Learning, including optimization, initialization, generative modeling, transfer learning, and reinforcement learning. We are hosting all of the accepted reports in OpenReview for the community to read and add to their understanding of the original NeurIPS 2019 paper.

7 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

³<https://codeocean.com/>

conference in the field, NeurIPS, has undertaken a reproducibility program this year which consisted of three components: a code submission policy, the inclusion of the Machine Learning Reproducibility checklist as part of the paper submission process, and this challenge. We hope our endeavor will similarly spur more efforts in reproducing existing ideas and papers, and in turn promote open, accessible and sound machine learning research.

8 Acknowledgements

We thank the NeurIPS board and the NeurIPS 2019 general chair (Hanna Wallach) and program chairs (Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily Fox) for the unfailing support of this initiative. We thank the many authors who submitted their work to NeurIPS 2019 and communicated with the challenge participants. We thank the program committee (Zhenyu (Sherry) Xue) of NeurIPS 2019 for providing us data and statistics of the papers accepted in the NeurIPS 2019 conference which helped us in building the portal. We thank the OpenReview team (in particular Andrew McCallum, Pam Mandler, Melisa Bok, Michael Spector, and Mohit Uniyal) who provided extensive support from day one to build and host the dual-purpose portal, and to host the results of the reproducibility challenge. We thank CodeOcean (Xu Fei) for supporting our challenge by providing cloud compute resources. Finally, we thank the several participants of the reproducibility challenge who dedicated time and effort to verify results that were not their own, to help strengthen our understanding of machine learning, and the types of problems we can solve today.

9 Reviewers

In this iteration of the Reproducibility Challenge, we were fortunate enough to attract a large base of reviewers having prior experience in reviewing in large Machine Learning conferences such as NeurIPS, ICML, ICLR, etc. Many thanks to all our reviewers, we acknowledge their hard efforts who spent their precious time to critically review the reports. We hope that our reviewer base will keep supporting us in this endeavor in the future.

Abhinav Agrawal	Charbel Sakr	Emmanuel Bengio
Adria Garriga-Alonso	Chen Tessler	Erfan Sadeqi Azer
Amrith Rawat	Cheng Ju	Eric Crawford
Andreas Ruttner	Chuan Li	Eric Jang
Andreea Gane	Dagmar Kainmueller	Erin Conlon
Andrew Drozdov	Damian Roqueiro	Erin Grant
Andrew Jaegle	David Arbour	Ernest Ryu
Andrew Ross	David Krueger	Fang Liu
Angus Galloway	Di He	Fang Zhao
Antti Koskela	Dmitriy Serdyuk	Felix Gimeno
Arna Ghosh	Dong Gong	F. Martínez-Plumed
Austin Brockmeier	Dong Yin	F. Poursabzi-Sangdeh
Awa Dieng	Donghyeon Cho	Gabriel Synnaeve
Bryan Gibson	Du Tran	Gang Wang
Cagri Coltekin	Dylan Hadfield-Menell	Gavin Weiguang Ding
Chao Qin	Elaheh Raisi	

Georg Martius	Li Li	Razvan Pascanu
Georgios Leontidis	Li Shen	Reinhold Scherer
Gianfranco Doretto	Lijun Wu	Ritambhara Singh
Haiqin Yang	Linh Tran	Robert Vandermeulen
Haitian Sun	Liping Liu	Roy Schwartz
Hanna Suominen	Lluís Castrejon	Ryan Lowe
Hao He	Lovedeep Gondara	Sadid A. Hasan
Hei Law	Malik Altakrori	Samuel Albanie
Hidekazu Oiwa	Maneesh Singh	Sandhya Prabhakaran
Hong Ge	Manoj Acharya	Sara Hooker
Hongyi Wang	Måns Magnusson	Scott Fujimoto
Hua Wang	Marlos C. Machado	Sercan Arik
Huaibo Huang	Martin Klissarov	Sergio Valcarcel Macua
Huimin Ma	Massimiliano Mancini	Seungjae Lee
Huitong Qiu	Mathew Monfort	Shagun Sodhani
Huziel Saucedo	Matthew Schlegel	Shalini Ghosh
J. Hernandez-Garcia	Matthias Gallé	Shih-Yang Su
Jaeho Lee	Maxime Wabartha	Shivam Patel
Jake Bruce	Maxwell Collins	Shuai Tang
Jesse Dodge	Melanie F. Pradier	Shuai Zheng
Jessica Forde	Michal Drozdal	Shuxin Zheng
Ji Lin	Mike Chrzanowski	Simon Kornblith
Jiahui Yu	Mingkui Tan	Sohil Shah
Jiakai Zhang	Mingrui Liu	Stanislaw Jastrzebski
Jiangwen Sun	Minjia Zhang	Stefan Magureanu
Jing Wang	Mirco Musolesi	Steffen Udluft
Jinghui Chen	Nan Ke	Swapnil Mishra
Jitong Chen	Nesreen Ahmed	Takashi Ishida
Joan Puigcerver	Nikolaos Vasiloglou	Takeshi Teshima
Joel Lehman	Olga Isupova	Tammo Rukat
Joelle Pineau	Olivier Delalleau	Tobias Uelwer
John Wieting	Olivier Koch	Tzu-Yun Shann
Jonathan Hunt	Pablo Robles-Granda	Uthaipon Tantipongpipat
Josh Roy	Pascal Lamblin	Venkataadheeraj Pichapati
Kai Han	Patrick Philipp	Víctor Campos
Kanika Madan	Paul Tylkin	Vincent Francois-Lavet
Katherine Lee	Peixian Chen	Vincent Lepetit
Khimya Khetarpal	Peter Henderson	Volker Fischer
Konstantin Mishchenko	Praveen Narayanan	Wenhao Yu
Leo Lahti	Prithvijit Chattopadhyay	Wenxiao Wang
Levent Sagun	Qihang Lin	Wenxuan Wu
Li cheng	Razieh Nabi	Wesley Maddox
		Xavier Bouthillier

Xiang Yu	Xinggang Wang	Yufei Han
Xiang Zhang	Xingrui Yu	Yuji Matsumoto
Xiangliang Zhang	Xingyu Liu	Yuntian Deng
Xiangru Lian	Yash Goyal	Zhangjie Cao
Xin GUO	Yingyezhe Jin	Zhourong Chen
Xin Lu	Yoonho Lee	

References

1. J. Pineau, P. Vincent-Lamarre, K. Sinha, V. Larivière, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and H. Larochelle. "Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program)." In: **arXiv preprint arXiv:2003.12206** (2020).
2. H. E. Plesser. "Reproducibility vs. replicability: a brief history of a confused terminology." In: **Frontiers in neuroinformatics** 11 (2018), p. 76.
3. M. A. Heroux, L. Barba, M. Parashar, V. Stodden, and M. Taufer. **Toward a Compatible Reproducibility Taxonomy for Computational and Computing Sciences**. Tech. rep. Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), 2018.
4. E. National Academies of Sciences, Medicine, et al. **Reproducibility and replicability in science**. National Academies Press, 2019.