ReproducedPapers.org: Openly teaching and structuring machine learning reproducibility

Burak Yildiz[0000−0001−9932−4221], Hayley Hung, Jesse H. Krijthe[0000−0003−3435−6358], Cynthia C. S. Liem[0000−0002−5385−7695], Marco Loog[0000−0002−1298−8461], Gosia Migut, Frans A. Oliehoek[0000−0003−4372−5055], Annibale Panichella[0000−0002−7395−3588], Przemysław Pawełczak[0000−0002−1302−1148], Stjepan Picek[0000−0001−7509−4337], Mathijs de Weerdt[0000−0002−0470−6241], and Jan van Gemert[0000−0002−3913−2786]

Delft University of Technology, Postbus 5, 2600 AA Delft, The Netherlands

Abstract. We present ReproducedPapers.org: an open online repository for teaching and structuring machine learning reproducibility. We evaluate doing a reproduction project among students and the added value of an online reproduction repository among AI researchers. We use anonymous self-assessment surveys and obtained 144 responses. Results suggest that students who do a reproduction project place more value on scientific reproductions and become more critical thinkers. Students and AI researchers agree that our online reproduction repository is valuable.

Keywords: Machine Learning · Reproducibility · Online Repository.

1 Introduction

Reproducibility is a cornerstone of science: if an experiment is not reproducible, we should question its conclusions. Yet, machine learning papers are lacking reproductions [7,12]. Possible reasons may include a misaligned incentive between reproducing results and the short-term measures of career success associated with more 'wins' [26] and publishing 'novel' work [15]. Nevertheless, high-impact can be achieved, for instance, when a reproduction fails spectacularly, e.g. [6,8,10,11,14,16,18,19,24]. Yet, these take colossal amounts of manual effort [1,2,9,22] or massive resources [16,23]. There are venues for publishing reproductions [3,4,25], which are typically peer-reviewed and thus uphold various selection standards to guarantee quality. We argue that this emphasis on quality is a hurdle for sharing light-weight reproductions. Important and useful examples of light-weight reproductions include partial results, small variants on the algorithm, hyperparameter sweeps, etc. Low-barrier options are indeed available in workshop challenges [13,21] organized at conferences such as ICPR, NeurIPS, ICLR, or ICML. However, such avenues are hard to maintain on a long-term basis, as a workshop may or may not be organized. We argue that there is a need for a low-barrier and long-term venue for machine learning reproductions.

A complementary angle on low-barrier reproductions is to improve university student training. We should teach the next generation of machine learning

practitioners the importance of the reproducibility of research work, as done in other computer science domains such as computer networking, where results reproduction is the means to learn new material [30]. Doing a reproduction project in a course aligns with several important learning objectives for machine learning students. Among others, students (1) should be able to read, critique, and explain a scientific paper; (2) implement a method; (3) run, evaluate, investigate, and extend existing research or code; and (4) write clearly and concisely about code and methods. A reproduction project also lets students experience differences between published results and an implementation, which stimulates a critical attitude and allows reflections on the scientific process.

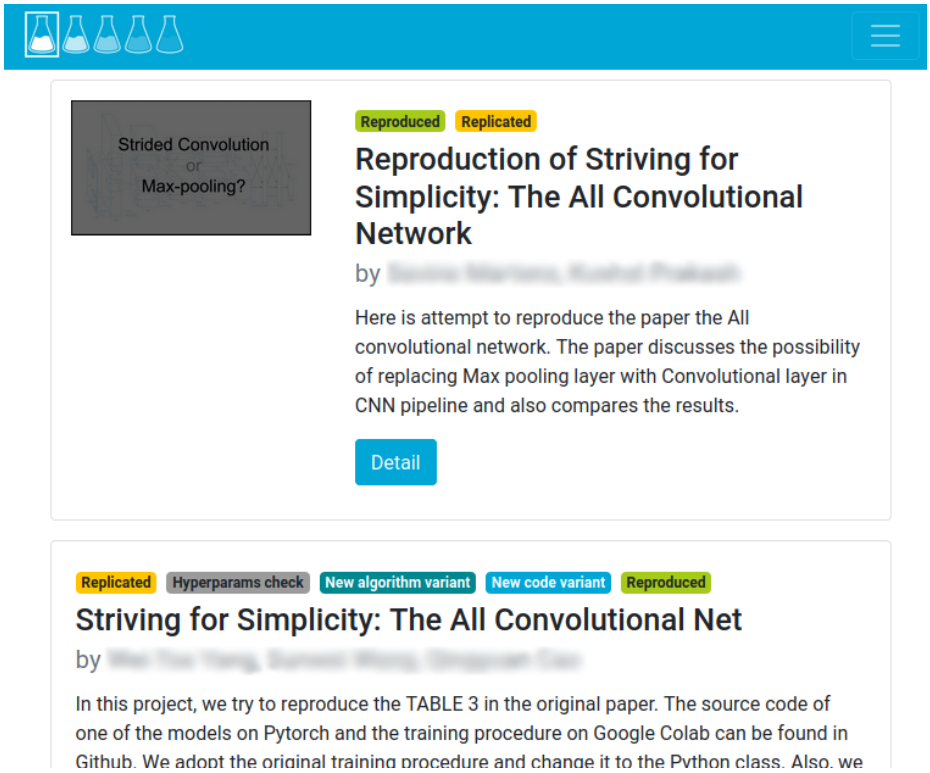


Fig. 1. A screenshot of `ReproducedPapers.org`. We allow multiple reproductions of the same original paper and investigations of several aspects, such as *Reproduced*, *Replicated*, *Hyperparameter check*, etc. Our online repository is user-centered: its sufficient if a user sees value in uploading some form of reproduction. Having such a repository is well-suited for students and adds structure to reproducibility in machine learning.

In this paper, we align the benefits of an online reproduction repository with those of teaching reproducibility. We introduce `ReproducedPapers.org`: an open, light-weight repository of reproduced papers which flexibly allows any sort of reproduction work, see Figure 1. This repository benefits the research community while at the same time being well-equipped at accepting contributions from students. Although the standard of student reproductions might be lower than those required for peer reviewed reproductions, they can still give valuable insights such as clarifying which parts are difficult to implement or identifying the reproducibility level of elements. Such online reproductions are a low-threshold portfolio-building opportunity, which in turn may prove a valuable incentive to start doing more reproductions, as well as an opportunity to facilitate sharing reproductions that otherwise would not have been shared.

Our online repository shares traits with other light-weight, bottom-up, grass-roots community efforts such as *ArXiv* [5], *Open Review* [28], and *Papers with Code* [29]. Other efforts on facilitating reproducibility include software for reproducible and reusable experiments [20], open specification neural network diagrams [17], and a framework for automatic parsing of deep learning research paper to generate the implementation [27]. Similar to these approaches, in our work, we combine the traits from online repositories with those of tools facilitating reproducibility by providing an online repository that facilitates teaching as well as structuring reproducibility.

We make the following contributions. 1. We propose a new online reproduction repository; 2. We conduct a proof of concept with students from an MSc Deep Learning course to perform a reproduction project and populate the repository; 3. We evaluate the usefulness of the repository among AI researchers and the learning objectives among students by anonymous surveys.

2 The online reproduction repository

We performed a proof of concept experiment with a reproducibility project for students of the MSc Deep Learning course taught by this paper's last author at Delft University of Technology (TU Delft). We solicited relevant papers among university staff and ensured that (i) data is available, (ii) it is clear which table or figure to reproduce, and (iii) the computational demands are reasonable. Students were also allowed to themselves suggest a paper to reproduce. On their paper of choice, they worked in groups of 2 to 4, for 8 weeks, for approximately one-third of their studying time (i.e., about 13 hours a week). For grading, students submitted a blog in PDF and also the URL of an online version of their blog to `ReproducedPapers.org` to populate the repository. For students who do not wish to share a blog with the world, we offer a private option, which is only visible to course administrators. The option to publicly blog about reproducing machine learning provides an simple opportunity for students to build an online portfolio while simultaneously incentivizing making reproductions.

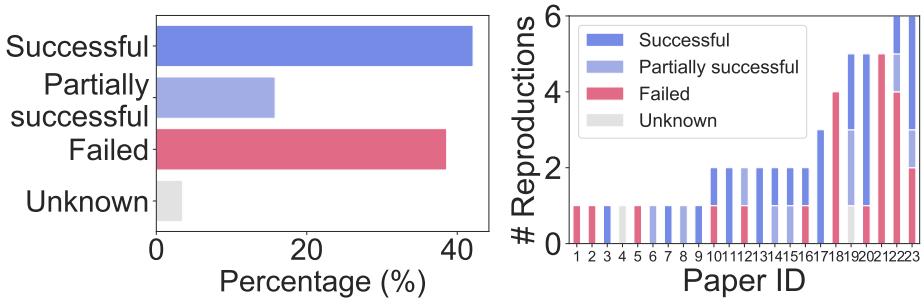
Aspect	Description
• Replicated	A full implementation from scratch without using any pre-existing code.
• Reproduced	Existing code was evaluated.
• Hyperparams check	New evaluation of hyperparameter sensitivity.
• New data	Evaluating new datasets to obtain similar results.
• New algorithm variant	Evaluating a different variant.
• New code variant	Rewrote/ported existing code to be more efficient/readable.
• Ablation study	Additional ablation studies.

Table 1. Different aspects of reproduction which are highlighted as badges (see Figure 1).

We explicitly allow for light-weight reproduction efforts such as evaluating existing code, checking only certain parts of the paper, proposing minor variations, doing hyperparameter sweeps, etc. Our current options (aspects) are shown in Table 1, and we will add others as the need arises. Authors label their reproduction with the relevant aspects themselves.

We developed `ReproducedPapers.org` in-house as a simple web application. It is implemented by this paper's first author, and its source code is available on GitHub[1]. Registering is necessary only when adding reproductions. Currently, the repository has 90 registered users and hosts 24 unique papers and 57 paper reproductions. Most papers have multiple reproductions, and only five reproductions are marked as private. The top-3 most-used aspects are *Replicated* (32 times); *Reproduced* (29 times) and *Hyperparams check* (17 times). Figure 2 whose data is derived from self-reported blog posts by users shows both success and failure rates to be around 40%.

[1] `https://github.com/CVLab-TUDelft/reproduced-papers`



(a) (b)

Fig. 2. Current `ReproducedPapers.org` statistics. (a) Reproduction success rates; (b) Number of reproductions per paper ID.

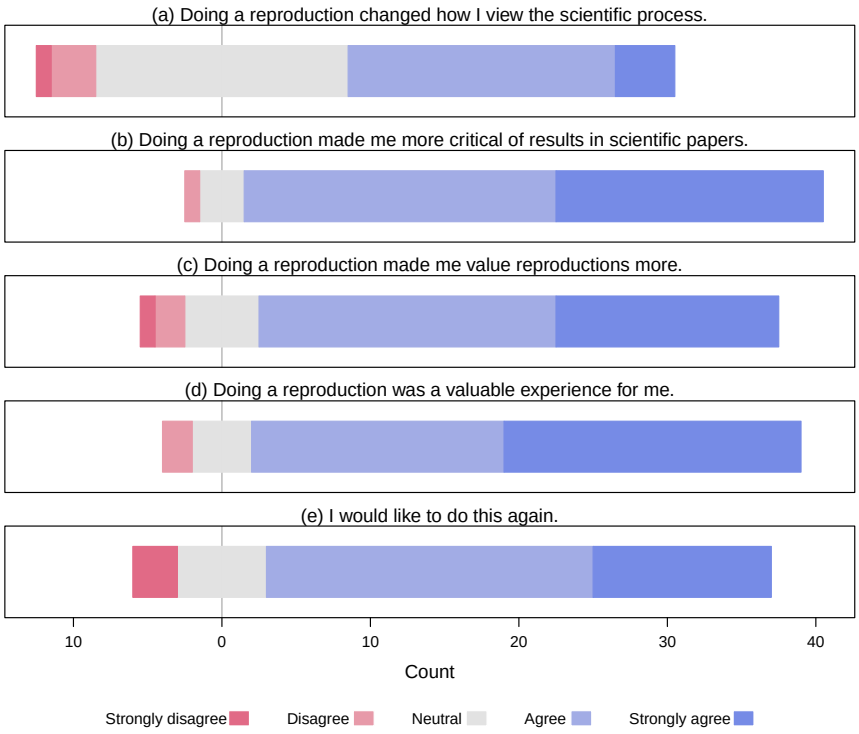


Fig. 3. Responses to survey questions from students who contributed to `ReproducedPapers.org`. Letting students themselves do a reproduction promotes a critical mindset (a and b), and teaches the value of scientific reproductions (c). In addition, the students considered it a positive experience (d,e). We conclude that these traits align with our learning objectives.

3 Survey analysis

We evaluate student learning objectives and how AI researchers perceive our online reproduction repository by analyzing the results of small anonymous surveys for two groups: (i) students who recently added their reproduction to our repository and (ii) anybody identifying her/himself working in AI. The second group was invited to the survey through social media and emails. Both groups share the same questions, where the students have five additional questions to evaluate education. The survey data is available at `ReproducedPapers.org`[2]

We received a total of 144 responses: 43 from course students and 101 from third-party AI researchers all over the world. Of the latter, 87 identify themselves as a junior or senior researcher, and 14 as a student.

[2] `https://reproducedpapers.org/survey-data.zip`

3.1 Evaluating student learning objectives

The survey questions and results can be found in Figure 3. We evaluate the following objectives.

Doing a reproduction project increases critical thinking. Results in Figure 3(a) show that doing a reproduction taught most students something new about the scientific process. Figure 3(b) suggests that students become more critical to published results.

Doing a reproduction project makes students value reproductions more. The results in Figure 3(c) indicate that after doing a reproduction, a great majority of students place more value on scientific reproductions.

Students find a reproduction project a positive experience. The results in Figure 3(d,e) demonstrates that students valued the work and prefer to do a reproduction more often. Results suggest that having a reproducibility project teaches skills considered important by both student and teacher.

3.2 Evaluating the AI researcher survey respondents

Figure 4 shows results for the third party AI researchers. We found the following.

The AI researcher survey respondents find online reproductions valuable. Results in Figure 4(a,d,g) show that students and, especially, researchers find an online reproduction valuable and useful. According to Figure 4(i), there is no clear preference for doing a reproduction or writing a paper. Figure 4(e) suggests that the perceived value of reproduction by the community is smaller for researchers than for students.

The AI researcher survey respondents find an online reproduction repository valuable. Results in Figure 4(b,c) show that students and researchers appreciate an online reproduction repository. Figure 4(f) shows that researchers are less likely than students to help contribute by doing reproductions.

The AI researcher survey respondents see an educational role for courses where students do a reproduction project. Results in Figure 4(h) show that researchers and students agree that reproduction projects should be used more often in courses.

Additionally, we make the following observations from Figure 4:

(i) When compared to students, the *researchers think the community values reproductions less (e) and want their own team to work on reproductions less (f)*. This may suggest an inverse relationship between perceived value and willingness to contribute. Yet, when comparing researchers against themselves, most think the community values reproductions, and most researchers would like to contribute.

(ii) More researchers *want their work reproduced (g) than that they are willing to contribute (f)*. Can we place our hope on the students as future researchers, as they are much more willing to contribute?

(iii) There is a clear consensus that *reproductions are valuable (a, d, g, i) but some researchers feel that the community does not reward it enough (e)*. Therefore, an important question is how we can change the perception of a low reward for doing reproductions, beyond repositories as reported on here.

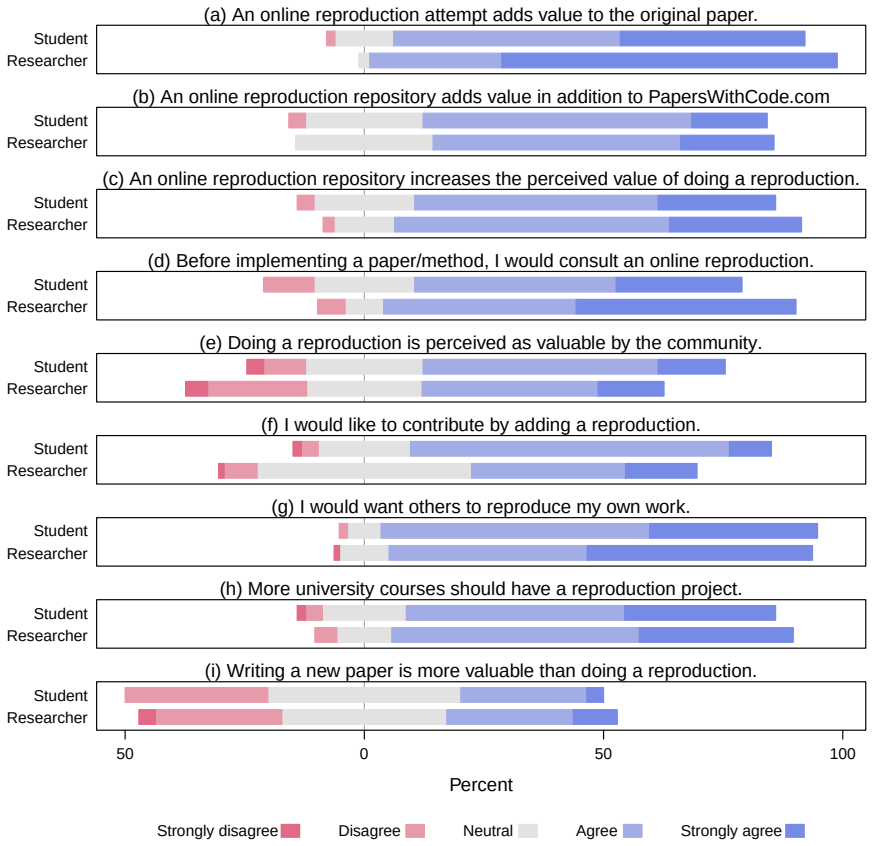


Fig. 4. Responses to survey questions by 57 students and 87 self-identified AI researchers. The survey question is in the sub-caption. Researchers and students agree that: Reproductions are valuable (a, d, g), that an online repository adds value (b, c), and that more courses should use a reproduction project (h). Researchers differ from students in that researchers more strongly find a reproduction valuable (a), and would consult online reproductions more (d). Researchers think a reproduction is valued less by the community (e) and are less likely to contribute with reproductions (f). Students and researchers both do not agree among themselves if a new paper is more valuable then a reproduction (i), suggesting that the answer is 'it depends'. We conclude that the respondents welcome an online repository for teaching and structuring reproducibility.

4 Discussion and conclusions

It should be clear that our results and corresponding analysis are rather preliminary. We are convinced, however, that they warrant low-barrier and long-term solutions accommodating research reproduction. Our `ReproducedPapers.org` pro-

vides one such outlet. We hope that future analysis of the further accumulated survey data may sketch an even clearer picture. We hope others consider reproducing our effort.

The main conclusions that we draw at present are the following three. 1. Doing a reproduction course project aligns well with learning objectives, and students find it a positive experience. 2. A reproducibility project improves the perceived value of reproductions, and allowing students to blog online about their reproduction project offers an extra incentive to do a reproduction. 3. AI researcher survey respondents are positive about online reproductions and a reproduction repository.

We finally call on the community to add their reproductions to the website `ReproducedPapers.org` and deploy it in courses: may the next generation of machine learners be reproducers.

References

1. Anand, K., Wang, Z., Loog, M., van Gemert, J.: Black magic in deep learning: How human skill impacts network training. In: British Machine Vision Conference (BMVC) (2020)
2. Bonneel, N., Coeurjolly, D., Digne, J., Mellado, N.: Code replicability in computer graphics. ACM Transactions on Graphics **39**(4) (2020)
3. Colom, M., Kerautret, B., Krähenbühl, A.: An overview of platforms for reproducible research and augmented publications. In: International Workshop on Reproducible Research in Pattern Recognition. pp. 25–39. Springer (2018)
4. Colom, M., Kerautret, B., Limare, N., Monasse, P., Morel, J.M.: Ipol: a new journal for fully reproducible research; analysis of four years development. In: 2015 7th International Conference on New Technologies, Mobility and Security (NTMS). pp. 1–5. IEEE (2015)
5. Cornell University Library: arxiv. `https://arxiv.org` (Sep 2011), last accessed: Jun. 20, 2020
6. Dacrema, M.F., Cremonesi, P., Jannach, D.: Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches. In: Proceedings of the 13th ACM Conference on Recommender Systems (2019)
7. Drummond, C.: Replicability is not reproducibility: Nor is it good science. In: Evaluation Methods for Machine Learning Workshop at the 26th ICML (2009)
8. Engstrom, L., Ilyas, A., Santurkar, S., Tsipras, D., Janoos, F., Rudolph, L., Madry, A.: Implementation Matters in Deep Policy Gradients: A Case Study on PPO and TRPO. arXiv preprint arXiv:2005.12729 (2020)
9. Fursin, G., Moreau, T., Reddi, V.: Asplos 2020 artifact evaluation report. In: Proc. ASPLOS. pp. vi–vii. ACM (2020)
10. Gorman, K., Bedrick, S.: We need to talk about standard splits. In: Proceedings of the 57th annual meeting of the association for computational linguistics. pp. 2786–2791 (2019)
11. Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., Meger, D.: Deep reinforcement learning that matters. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
12. Hutson, M.: Artificial intelligence faces reproducibility crisis. Science **359**(6377), 725–726 (2018)

13. Kerautret, B., Colom, M., Lopresti, D., Monasse, P., Talbot, H.: Reproducible Research in Pattern Recognition: Second International Workshop, RRPR 2018, Beijing, China, August 20, 2018, Revised Selected Papers, vol. 11455. Springer (2019)
14. Lin, J.: The neural hype and comparisons against weak baselines. ACM SIGIR Forum **52**(2), 40–51 (2019)
15. Lipton, Z.C., Steinhardt, J.: Research for practice: Troubling trends in machine-learning scholarship. Commun. ACM **62**(6), 45–53 (May 2019)
16. Lucic, M., Kurach, K., Michalski, M., Gelly, S., Bousquet, O.: Are gans created equal? a large-scale study. In: Advances in neural information processing systems. pp. 700–709 (2018)
17. Marshall, G., Freitas, A.: The Diagrammatic AI Language (DIAL): Version 0.1. arXiv preprint arXiv:1812.11142 (2018)
18. Melis, G., Dyer, C., Blunsom, P.: On the state of the art of evaluation in neural language models. In: International Conference on Learning Representations (2018), https://openreview.net/forum?id=ByJHuTgA-
19. Musgrave, K., Belongie, S., Lim, S.N.: A metric learning reality check. arXiv preprint arXiv:2003.08505 (2020)
20. Paganini, M., Forde, J.Z.: dagger: A Python Framework for Reproducible Machine Learning Experiment Orchestration (2020)
21. Pineau, J., Sinha, K., Fried, G., Ke, R.N., Larochelle, H.: ICLR Reproducibility Challenge 2019. ReScience C **5**(2), 5 (may 2019)
22. Raff, E.: A step toward quantifying independently reproducible machine learning research. In: NeurIPS. pp. 5486–5496 (2019)
23. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do imagenet classifiers generalize to imagenet? In: ICML. pp. 5389–5400 (2019)
24. Riquelme, C., Tucker, G., Snoek, J.: Deep bayesian bandits showdown: An empirical comparison of bayesian deep networks for thompson sampling. In: International Conference on Learning Representations (2018), https://openreview.net/forum?id=SyYe6k-CW
25. Rougier, N.P., Hinsen, K.: Rescience c: a journal for reproducible replications in computational science. In: International Workshop on Reproducible Research in Pattern Recognition. pp. 150–156. Springer (2018)
26. Sculley, D., Snoek, J., Wiltschko, A., Rahimi, A.: Winner's Curse? On Pace, Progress, and Empirical Rigor. In: ICLR workshop (2018), https://openreview.net/forum?id=rJWF0Fywf
27. Sethi, A., Sankaran, A., Panwar, N., Khare, S., Mani, S.: DLPaper2Code: Auto-generation of code from deep learning research papers. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
28. Soergel, D., Saunders, A., McCallum, A.: Open scholarship and peer review: a time for experimentation. In: Proc. ICML (2013)
29. Stojnic, R., Taylor, R.: Papers with code—a facebook AI project. https://paperswithcode.com (Jul 2018), last accessed: Jun. 20, 2020
30. Yan, L., McKeown, N.: Learning networking by reproducing research results. SIGCOMM Comput. Commun. Rev. **47**(2), 19–26 (Apr 2017)