QUANTIFYING REPRODUCIBLE MACHINE LEARNING RESEARCH

Booz | Allen | Hamilton

Edward Raff[1,2]

1 Booz Allen Hamilton 2 University of Maryland, Baltimore County

33rd Conference on Neural Information Processing Systems (NeurIPS), 8–14 December 2019, Vancouver, Canada

UMBC

Abstract

What makes a paper independently reproducible? Debates on reproducibility center around intuition or assumptions but lack empirical results. Our field focuses on releasing code, which is important, but is not sufficient for determining reproducibility. We take the first step toward a quantifiable answer by manually attempting to implement 255 papers published from 1984 until 2017, recording features of each paper, and performing statistical analysis of the results. For each paper, we did not look at the authors code, if released, in order to prevent bias toward discrepancies between code and paper.

Study Design

- *Attempt to independently reproduce results of 255 papers, with 63.5% success rate.*
- *Papers published from 1984-2017*
- *Reproductions attempted from 2012-2018*
- *Developed 26 quantifications, grouped by Objective, Mildly Subjective, & Subjective.*
- *Study made possible by historical notes & paper organization software*

Paper and data available here 👉

Correlated with Reproducibility

We found a number of interesting results, with some summarized below. Please see the paper for nuance!

- No relation between reproduction and year attempted, making analysis easier.
- No relation between reproduction and year a paper was published. Does this imply the “crisis” has been ever present, overblown, or something else?
- Papers with an empirical bias or emphasis are most reproducible.
- The effort on focusing on specifying and better replication around hyper-parameters is well placed.
- Pseudo code’s impact is not as straight forward as one might think. Papers with detailed pseudo code and no pseudo code where equally reproducible. The ones in-between, not as much.
- If paper author replies we get 85% reproduction rate, which drops to 4% if they don’t reply to emails.
- Readability, defined as number of reads through paper to finish code, is subjective, be predictive.

What was not significant is also important to note.

- Workshop vs Conference had no impact.
- Surprisingly, neither did papers using toy problems or explanatory/”conceptualization” figures.
- Appendices do not appear to help.

Study Biases

- All reproductions attempts where done by one author, who is not an expert in all the topic areas attempted, and does not have unlimited time.
- Papers studied are not randomly sampled, but biased toward personal interests, as well as what has become popular over time.
- We have not yet factored into our analysts anything about the authors of the papers under analysis, which would likely have a significant impact on the results.

Is “Reproducible” Even a Binary?

In particular, after performing this work, we note a fundamental problem with the question framing: that a reproducibility is a binary property that paper has or does not have. One particular paper under analysis took 4.5 years to successfully reproduce.

In this light, perhaps we should look at reproducibility as a kind of survival analysis? Reproduction is the “death” of a paper, and a paper that fails reproduction “survives” indefinitely. The survival rate becomes the effort and time needed to reproduce, conditioned on properties of both the paper (e.g., what we have quantified) as well as the author and their resources.

Results

Table 1: Significance test of which paper properties impact reproducibility. Results significant at $\alpha \leq 0.05$ marked with“*”.

Feature	p-value
Year Published	0.964
Year First Attempted	0.674
Venue Type	0.631
Rigor vs Empirical*	1.55×10^{-9}
Has Appendix	0.330
Looks Intimidating	0.829
Readability*	9.68×10^{-25}
Algorithm Difficulty*	2.94×10^{-5}
Pseudo Code*	2.31×10^{-4}
Primary Topic*	7.039×10^{-4}
Exemplar Problem	0.720
Compute Specified	0.257
Hyperparameters Specified*	8.45×10^{-6}
Compute Needed*	8.75×10^{-5}
Authors Reply*	6.01×10^{-8}
Code Available	0.213
Pages	0.364
Publication Venue	0.342
Number of References	0.740
Number Equations*	0.004
Number Proofs	0.130
Number Tables*	0.010
Number Graphs/Plots	0.139
Number Other Figures	0.217
Conceptualization Figures	0.365
Number of Authors	0.497