

A SIREN SONG OF OPEN SOURCE REPRODUCIBILITY

Booz | Allen | Hamilton

Edward Raff^{1,2} Andrew L. Farris¹

¹ Booz Allen Hamilton ² UNIVERSITY OF MARYLAND, BALTIMORE COUNTY
ML Evaluation Standards Workshop at ICLR 2022, 29 April 2022, Virtual



Abstract

We desire to make a nuanced point that we believe is counter-culture and unintuitive to many, and so our framing is intentionally provocative.

- **We are not saying open source is bad!** We've open-sourced 200k+ lines of code and apart of the Apache Software Foundation.
- We are not saying you shouldn't open source your code!
- We are not arguing against the many benefits of OSS!

We are saying that with respect to reproducible research, **open code can have non-positive impacts**. This can be negligible, or even negative. For this reason, we argue that the community should stop focusing so heavily on OSS, and **instead focus on incentivizing more study on the question of reproducibility itself**. We don't have enough information to make useful and informed decisions for the community, acting quickly because it "feels right" is the antithesis of good science, and too many are ignoring the critical data that is being generated.

Problems and Opportunities in Training Deep Learning Software Systems: An Analysis of Variance

Hung Viet Pham University of Waterloo Waterloo, ON, Canada hvp@uwaterloo.ca	Shangshu Qian Purdue University West Lafayette, IN, USA qian151@purdue.edu	Jiannan Wang Purdue University West Lafayette, IN, USA wang4524@purdue.edu	Thibaud Lutellier University of Waterloo Waterloo, ON, Canada tlutelli@uwaterloo.ca
Jonathan Rosenthal Purdue University West Lafayette, IN, USA rosenth0@purdue.edu	Lin Tan Purdue University West Lafayette, IN, USA lintan@purdue.edu	Yaoliang Yu University of Waterloo Waterloo, ON, Canada yaoliang.yu@uwaterloo.ca	Nachiappan Nagappan Microsoft Research Redmond, WA, USA nachin@microsoft.com

WE HAVE FORGOTTEN HISTORY...

- Several works by Hatton & Roberts had multiple different teams implement the same algorithms. *The implementations agreed on only one significant figure!*
- Carl Taswell made distinctions between quality of exposition and verification of numerical equivalence in implementation, and pushed for *how to better specify the algorithm so that implementations come out with the same results!*
- *Code is a false veneer of reproducibility, but lets you get away with replication*

IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, VOL. 20, NO. 10, OCTOBER 1994

How Accurate Is Scientific Software?

Les Hatton and Andy Roberts

Reproducibility Standards for Wavelet Transform Algorithms

Carl Taswell

NOW WE ARE REPEATING IT

- Our hardware and it's APIs aren't giving us deterministic results!
- Our implementations across frameworks aren't the same!
- We are 'over-fit' to a few BLAS libraries in our results
- Re-running the same models on different or even the same compute can give us very different answers. The precision gets down to one significant figure or less!
- Reading code is harder than writing it, we have no quantified evidence that OSS helps with reproducibility, only that it introduces new and different challenges

CAN OPEN SOURCED CODE HARM US?

- OSS can lead to scientific harm / slow progress. We take Word2Vec to "punch up" as an example
 - We are not saying word2vec was not valuable overall, its one of the most successful and widely used techniques.
- But its success is only because code was made available, and years of research liked burned because of it!
- Word2vec has never been reproduced.
 - Every implementation available is a port of the original code!
- Despite immediate and enormous interest, the discrepancy was not publicly documented until 2019!
- Clearly having the code does not make it easy to confirm it's correctness
- This means the paper is wrong
- Years of research analyzing the model the paper reports was misguided

Everything you know about word2vec is wrong

The classic explanation of word2vec, in skip-gram, with negative sampling, in the paper and countless blog posts on the internet is as follows:

```
while(1) {
  1. vf = vector of focus word
  2. vc = vector of context word
  3. train such that (vc . vf = 1)
  4. for(0 <= i < negative samples):
      vneg = vector of word *not* in context
      train such that (vf . vneg = 0)
}
```

Indeed, if I google "word2vec skipgram", the results I get are:

- ▶ [The wikipedia page which describes the algorithm on a high level](#)
- ▶ [The tensorflow page with the same explanation](#)
- ▶ [The towards data science blog which describes the same algorithm](#)

the list goes on. However, *every single one of these implementations is wrong*. The original word2vec C implementation does *not* do what's explained above, and is *drastically different*. Most serious users of word embeddings, who use embeddings generated from word2vec do one of the following things:

<http://bollu.github.io/everything-you-know-about-word2vec-is-wrong.html>

So, what do we do?

- Far too much of reproducibility work is based on opinion. We are supposed to be a science, but quantification is rare and generally not being rewarded.
- Critical work that is quantifying our datasets and how we run our conferences, identifying flaws, are being rejected under the absurd: **"The main argument for rejection is the the analysis done in the paper is not typical of ICLR research"**
 - If we can't accept quantified criticism of our field and institutions, we are lost as a scientific discipline
- All major AI/ML conferences should make dedicated tracks to studying reproducibility
- Novelty, math, etc should not be factors.
- Judge purely based on improvement in knowledge / understanding of reproducibility broadly, and in AI/ML specifically

AN OPEN REVIEW OF OPENREVIEW: A CRITICAL ANALYSIS OF THE MACHINE LEARNING CONFERENCE REVIEW PROCESS

David Tran *
University of California, Berkeley
davidtran@berkeley.edu

Alex Valtchanov *
Princeton University
avv2@princeton.edu

Keshav Ganapathy *
Centennial High School
kganapathy23@gmail.com

Raymond Feng *
Harvard University
raymond.feng@college.harvard.edu

Eric Slud
University of Maryland, College Park
evs@math.umd.edu

Micah Goldblum
University of Maryland, College Park
goldblum@umd.edu

Tom Goldstein
University of Maryland, College Park
tomg@cs.umd.edu