

The Fuzzy Boundaries of Reproducibility

Fourth Workshop on Reproducible Research
in Pattern Recognition (RRPR 2022)

August 21, 2022

Daniel Lopresti

Computer Science and Engineering

Lehigh University

Bethlehem PA 18015, USA

lopresti@cse.lehigh.edu

** With special thanks to George Nagy*

Motivation (to Prompt Discussion)

Some issues / questions that occur to me:

- Contradictory definitions of “reproducibility” vs. “replicability.”
- Missing connections to more useful idea of “generalizability.”
- Perhaps a formalism would help us bridge these gaps?

With so much attention now focused on reproducibility:

- Does evidence show we’re doing better than past, or worse?
- Is verifying reproducibility always a “heavyweight” process?
- Could “lightweight” indirect measures give us a clue?

August Issue of IEEE Computer



Reproduce / Replicate / Generalize

"SCIENCE IS A CONVERSATION"

"We tend to think of publication as the only medium for communicating scientific progress and findings, but science progresses also through preprint sharing, correspondence, conference interactions, social media, and any medium of conversation. Scientific knowledge is created in conversations among scientists and, in an expanded definition of conversation, among scientists interacting with a body of knowledge (which is the product and record of other scientists' conversations)."

"Defining the Role of Open Source Software in Research Reproducibility," Lorena A. Barba, IEEE Computer, August 2022.

Reproduce / Replicate / Generalize

"REPRODUCIBILITY AS A TRUST-BUILDING ENDEAVOR"

"Reproducibility leaders Jeff Leek and Roger Peng wrote: "To maintain the integrity of science research and the public's trust in science, the scientific community must ensure reproducibility and replicability by engaging in a more preventative approach that greatly expands data analysis education and routinely uses software tools." Both scientific integrity and general trust in science are often linked with reproducibility."

"Defining the Role of Open Source Software in Research Reproducibility," Lorena A. Barba, IEEE Computer, August 2022.

Reproduce / Replicate / Generalize

“Reproducibility is the functional capability to **recreate results using the original data set** used in a given scientific inquiry as originally reported. Stated simply, this is what is generally understood as **recreating the original experiment**.

Replicability is the capability for **independent and neutral parties to recreate the results of an original experiment to a reasonable degree of accuracy** (if not the exact results) **using a different data set**, composed of **relevant data gathered separately** from the process whereby the original researchers gathered the data they used. This is what is usually understood as **independent confirmation of results**.

Generalizability is the ability of research results to be **applied in other contexts**. This characteristic, and especially the range and scale of other contexts that benefit from the results of research, is especially desirable among the outputs of research efforts and is arguably one of the main reasons that funding agencies are created to foster research activities in the first place. **While reproducibility and replicability have not been thoroughly studied and understood, generalizability has been far less thoroughly studied (despite its perceived importance).**”

“Advancing Reproducibility at the NSF,” Martin D. Halbert, IEEE Computer, August 2022.

Reproduce / Replicate / Generalize

In the context of software-based sciences, like ours:

1. Reproduce = recreate original experiment = use same algorithm and same data to solve same problem.
2. Replicate = independent confirmation of results = use same algorithm and different (related) data to solve same problem.
3. Generalize = applied in other contexts = use same algorithm and different (unrelated) data to solve a different problem.

Key question: "same algorithm" = "same code"?

→ Sharing data facilitates (1), but probably not (2) and (3).

→ Sharing code facilitates (1), and perhaps (2) and (3).

Complications

But, remember, Raff studied 255 published machine learning papers and found that only 63.5% of reported results could be reproduced. He found 10 out of 26 features to be significant predictors, including:

- Readability (largest impact).
 - Rigor vs. empirical (more theoretical vs. more practical).
 - Algorithm difficulty.
 - Presence of pseudo code,
 - Broad subject area (e.g., specific branch of machine learning).
 - Responsiveness of authors to email queries.
- ➔ Conspicuous in its absence: availability of source code.

"A Step Toward Quantifying Independently Reproducible Machine Learning Research," Edward Raff, 33rd Conference on Neural Information Processing Systems (NeurIPS), pp. 5,485-5,495.

Complications

Another interesting point made by Bouthillier, et al.:

- Unreproducible findings can be built upon reproducible methods.
 - Facilitating reproduction of methods is important, but ...
 - Reproduction of findings is more important.
 - We should not be distracted from this more fundamental goal.
 - Reproducibility of empirical findings and conclusions must properly account for essential sources of variations. They demonstrate this by illustrating impact random seed can have, using selection of common DNN models across a number of computer vision tasks.
- ➔ We should extend discussion to conclusions and not just results.

"Unreproducible Research is Reproducible," Xavier Bouthillier, César Laurent, Pascal Vincent
Proceedings of the 36th International Conference on Machine Learning, PMLR 97:725-734, 2019.

Reproduce / Replicate / Generalize

Simple (obtain same results) = Take the same code and data and run it on a different system and obtain comparable results (e.g., accuracies with $\delta\%$ of the original).

Complete (reach same conclusions) = Partition results reported in the paper into two categories: (a) those run anew for the paper, and (b) those cited from the literature. For (a), take the same code and data and run all of the experiments on a different system and obtain comparable results. For (b), obtain original publications that were simply cited and confirm the results as reported. Finally, confirm that the same conclusions from the original paper hold.

My first attempt at a formalism
... for the sake of discussion ...

Reproduce / **Replicate** / Generalize

Simple (obtain same results) = Take the description of the algorithm in the paper and re-implement it. Take related data and run the code on a different system and obtain comparable results (e.g., accuracies with $\delta\%$ of the original).

Complete (reach same conclusions) = Partition results reported in the paper into two categories: (a) those run anew for the paper, and (b) those cited from the literature. For (a), re-implement the algorithms. Take the related data and run all of the experiments on a different system and obtain comparable results in each case. For (b), obtain original publications that were simply cited and confirm the results as reported. Finally, confirm that the same conclusions from the original paper hold.

Reproduce / Replicate / **Generalize**

Simple (obtain expected results) = Based on results reported in original paper, present hypothesis of what would happen if algorithm is run in another context. Take the description of the algorithm in the paper and re-implement it. Run the code on the new data on a different system and obtain expected results (e.g., accuracies with $\delta\%$ of expectation).

Complete (reach same conclusions) = Partition results reported in the paper into two categories: (a) those run anew for the paper, and (b) those cited from the literature. For (a), re-implement the algorithm ... For (b), obtain original publications that were simply cited and confirm the results as reported. Finally, confirm that the same conclusions from the original paper hold.

Other, Indirect Measures?

Verifying reproducibility seems like huge amount of work.
(Thanks to those of you who helping with it!)

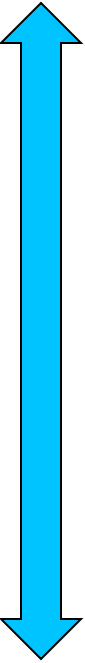
Are there other, easier (more scalable) measures that might provide an indication of reproducibility we could try out?

Inbound and outbound citations could provide such a measure:

- Inbound citations indicate whether other authors have attempted to reproduce the work in the current paper.
- Outbound citations indicate whether authors of the current paper have attempted to reproduce the work they cite, or simply accept it "on faith."

Citation Analysis

Some initial thoughts. First, a proposed grading system:

- More
- Evidence of reproducibility
- 
- Less
- A Paper cited because the authors demonstrated a new application for the original idea (perhaps with modifications).
 - B Paper cited because the authors re-implemented the algorithm, but then just used it in a straightforward way on new data to show that their own method is better.
 - C Paper cited because the authors obtained the original code and ran it on new data.
 - D Paper cited because the authors obtained the original code and re-ran it on the original data.
 - E Paper cited because authors reported results in their paper, but no evidence they ran existing code or re-implemented it.
 - F Paper cited, but results not reported.

A Quick Look

Paper chosen: "Pyramid Mask Text Detector."

- 21 outbound citations (ignored for now).
- I examined 10 inbound citations (top 10 on Google Scholar).

- 1 cite** A Paper cited because the authors demonstrated a new application for the original idea (perhaps with modifications).
B - D (not seen)
- 5 cites** E Paper cited because authors reported results in their paper, but no evidence they ran existing code or re-implemented it.
- 4 cites** F Paper cited, but results not reported.

"Pyramid Mask Text Detector," J. Liu, X. Liu, J. Sheng, D. Liang, X. Li, Q. Liu, <https://arxiv.org/abs/1903.11800>.

Example of Type A

work. Specifically, the model simultaneously learns a local Mask-RCNN-based [6] aligned bounding boxes detection task and a global segmentation task. In both tasks, we adopt the pyramid soft mask supervision [17] to help obtain more accurate aligned bounding boxes. In LGPMA, the local branch (LPMA) acquires more reliable text region information through visible texture percenteron. while

aligned bounding box learning is not easy because cells are easy to be confused with empty regions. Motivated by the advanced pyramid mask text detector [17], we find that using the soft-label segmentation may break through the proposed bounding box's limitation and provide more accurate aligned bounding boxes. To fully utilize the visual features from both local texture and global layout,

Similarly, we can obtain the other three refined boundaries. Notice that the refining process can optionally be conducted iteratively refer to [17].

Evidence of generalizability (and reproducibility)!

"LGPMA: Complicated Table Structure Recognition with Local and Global Pyramid Mask Alignment." <https://arxiv.org/abs/2105.06224>

Examples of Type E and F

Type E:

Method [ref]	AP _{0.5}	AP _{0.5} ⁺	AP _{0.5} ⁺	AP _{0.5} ⁺	AP _{0.5} ⁺	AP _{0.5} ⁺	AP _{0.5} ⁺	AP _{0.5} ⁺	AP _{0.5} ⁺	AP _{0.5} ⁺	AP _{0.5} ⁺	AP _{0.5} ⁺
PSENet [22]	88.7	85.5	87.1	75.4	69.2	72.1	-	-	-	84.0	78.0	80.9
SPCNET [20]	88.7	85.8	87.2	73.4	66.9	70.0	-	-	-	83.0	82.8	82.9
Pixel-Anchor [33]	88.3	87.1	87.7	79.5	59.5	68.1	-	-	-	-	-	-
PMTD [34]	91.3	87.4	89.3	85.2	72.7	78.5	87.5	78.1	82.5	-	-	-
CRAFT [35]	89.8	84.3	86.9	80.6	68.2	73.9	81.4	62.7	70.9	87.6	79.9	83.6
LOMO [19]	-	-	-	-	-	-	-	-	-	-	-	-
Our Method (ResNet50 without BA)	-	-	-	-	-	-	-	-	-	-	-	-

Evidence someone at least looked at results.

Type F:

of ships is adopted to control the covariance of confidence distribution. The rotation Gaussian-Mask ensures that the ship center has the highest reliability, and the confidences of other regions are gradually reduced by the Gaussian distribution. Similar to the soft-label method in [34], the Gaussian-Mask solves the problem of background interference when using a binary map to represent the ship detection results.

No real evidence of anything.

Type E: "Scene Text Detection with Polygon Offsetting and Border Augmentation." <https://www.mdpi.com/2079-9292/9/1/117/htm>

Type F: "GRS-Det: An Anchor-Free Rotation Ship Detector Based on Gaussian-Mask in Remote Sensing Images." <https://ieeexplore.ieee.org/document/9186810>

Citation Analysis and Reproducibility?

- Seems like crowdsourcing scientific opinion – could be a good idea. (Also seems like Google's original page rank algorithm.)
- Requires open access to the papers. Combines NLP with some document analysis. Might also involve OCR if source document is in image format. So not trivial.
- Some past work has been done on citation analysis, but focus seems to be as a smarter replacement for h-index. Would have to be adapted to identify evidence of reproducibility.
- Clearly some citations are more informative than others.
- But at best a weak indicator, with no guarantees: must trust many other authors, NLP can hard (e.g., detecting negation).

Citation Analysis and Reproducibility?

If proven to be effective, what might citation analysis reveal?

Some guesses:

- Could point to papers that require community's attention. E.g., a paper with lots of E citations, but no A-D citations. (Results quoted, but no one has attempted to reproduce them.)
- A way to identify which papers require costly verification?
- Can you get many E citations by publishing results that everyone else can easily beat (and hence won't challenge)?
- Has impact of GitHub, standard datasets, competitions, and leaderboards hurt reproducibility ($>$ E citations, $<$ A-D)?
- Are older papers more likely to employ A-D citations?

Possible Discussion Questions

- Would formalizing reproducibility / replicability / generalizability help us with our progress?
- Is there value in developing a common language?
- How can we turn focus more toward conclusions vs. results?
- Is citation analysis – or another indirect measure – useful in characterizing reproducibility?
- Who takes responsibility for monitoring papers that seem to need independent verification?
- If such analyses can be automated, and it becomes influential (e.g., acceptance decisions), how will bad authors “game it”?

Thank you!

Especially for your work
helping to advance
the practice of reproducibility!!