Reproducing Deep Learning experiments: common challenges and recommendations for improvement

Authors: J. Machicao, A. Ben Abbes, L. Meneguzzi, P.L.P. Corrêa, A. Specht, R. David, G. Subsol, D. Vellenich, R. Devillers, S. Stall, N. Mouquet, M. Chaumont, L. Berti-Equille, D. Mouillot
Contact: P.L.P. Corrêa **Website**: https://parsecproject.org/ **twitter**: @PARSEC_News

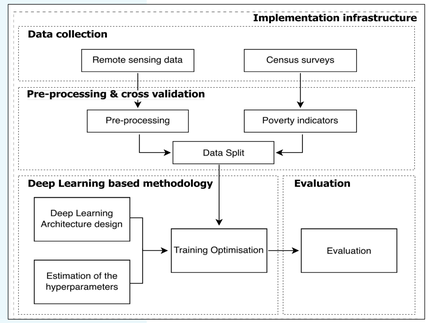
Reproducibility: why is this important?

One of the challenges in Machine Learning research is to ensure that the presented and published **results are sound and reliable**. Reproducibility is 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?
Reproducibility, that 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 Challenges on Deep Learning:

Making a Deep Learning (DL) research **experiment reproducible requires a lot of work** to document, verify, and make the system usable. These challenges are increased by the inherent complexity of DL, such as large number of parameters and hyperparameters, volumes of data (with possibly missing datasets and changes to the data), changes to some of the algorithms, and, as a result, versioning issues to deal with many iterations during training, which can be challenging for Reproducibility & Replicability (R&R) . Due to the **stochastic nature** of the variables within a DL methodology, such as: I the dataset, (ii) the DL architecture, (iii) the optimization procedure, (iv) the hyperparameters for optimization, and (v) the implementation and infrastructure, all of these are referred to as a source of 'variability' [1].



Flowchart of Poverty Estimation using remote sensing data and DL approaches.

We already went through the path of darkness:

We proposed a set of recommendations ('fixes') to overcome these reproducibility challenges that a researcher may encounter in order to improve Reproducibility and Replicability (R&R) and reduce the likelihood of wasted effort. These strategies can be used as "swiss army knife" to move from DL to more general areas as they are organized as (i) the quality of the dataset (and associated metadata), (ii) the Deep Learning method, (iii) the implementation, and the infrastructure used.

We identified the main challenges and constraints from these papers and presented them accordingly. Finally, with the lessons learned in the previous step, **we propose a set of mitigation strategies to overcome the main reproducibility challenges** and help researchers achieve their goals.

Mitigation Strategies and FAIR advice

Quality of the dataset
- **Validate** the **dataset** used in the study.
- Use a **sample** of the **dataset**.
- **Check parameters** of the datasets or code.
- **Check** the **preprocessing steps**.
- **Verify** the data **construction method**.
- **Test** different configurations of **data split**.

Deep Learning methodology
- Look for **workflows**.
- Look for **model architecture** as source code.
- Look for different **setups** of experiments.

Implementation and Infrastructure
- **Internet flaws.** ex. when dealing with APIs
- **Bugs.** It is expected that some bugs will be found in the scripts.
- **Versioning.** Use the same versions as the original in order to avoid 'deprecated' versions.
- **Source code is available but there are many branches.**
- **Programming language.** If the source code is in a particular programming language it is good advice to search for another trustable version.

Advice about FAIR principles criteria
- Some FAIR principles are mandatory in dataset design
- Some FAIR principles need scientific community agreement and information about dataset quality.
- Accept uncertainty when achieving FAIR, because a dataset is never perfect.

Conclusion:

The requirement for reproducibility and replicability of experiments, let alone those in which new techniques such as Deep Learning are employed, is very recent. There are few instances where R & R has been practically tested, and more work is needed to develop best practices and make this a real possibility for the achievement of truly open, defensible science.
Regarding FAIR principles, there have been some recent efforts to establish standards for FAIR ML and DL models [2, 3]. However, our knowledge, this is a first-of-its-kind initiative that presents a problem of reproducibility in remote sensing imagery and Deep Learning problems for real-world tasks. However, there are still some limitations to this work. It tries to be general, but it is an initiative for the DL community.

References:

[1] Renard, F., Guedria, S., Palma, N.D., & Vuillerme, N. (2020). Variability and reproducibility in deep learning for medical image segmentation. Scientific Reports 10(1), 1–16.
[2] Pineau, J. (2020b). The Machine Learning reproducibility checklist (v2.0, Apr.7 2020). www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist-v2.0.pdf
[3] Walsh, I., Fishman, D., Garcia-Gasulla, D., Titma, T., group, The ELIXIR Machine Learning focus group, Harrow, J., Psomopoulos, F. E., & Tosatto, S. C. E. (2020). DOME: Recommendations for machine learning validation in biology. 1–21.arXiv: 2006.16189. http://arxiv.org/abs/2006.16189

Acknowledgements:
PARSEC is funded by the Belmont Forum through the National Science Foundation (NSF), The São Paulo Research Foundation (FAPESP), the French National Research Agency (ANR). J.M. is grateful for the support from FAPESP (grant 2020/03514–9).

Author affiliations : Jeaneth Machicao (University of São Paulo, BR) https://orcid.org/0000-0002-1202-0194; Ali Ben Abbes (FRB-CESAB, Montpellier, FR) https://orcid.org/0000-0001-5714-7562; Leonardo Meneguzzi (University of São Paulo, BR) https://orcid.org/0000-0002-4845-6758; Pedro Pizzigatti Corrêa (University of São Paulo, BR) https://orcid.org/0000-0002-8743-4244; Alison Specht (The University of Queensland, AU) https://orcid.org/0000-0002-2623-0854; Romain David (ERINHA (European Research Infrastructure on Highly Pathogenic Agents) AISBL, FR) https://orcid.org/0000-0003-4073-7456; Gérard Subsol (Research-Team ICAR, LIRMM, CNRS, Univ. Montpellier, FR) https://orcid.org/0000-0002-7461-4932; Danton Ferreira Vellenich (University of São Paulo, BR), https://orcid.org/0000-0002-3223-6996; Shelley Stall, American Geophysical Union, USA) https://orcid.org/0000-0003-2926-8353; Nicolas Mouquet (FRB-CESAB, Montpellier, FR), https://orcid.org/0000-0003-1840-6984; Marc Chaumont (LIRMM, CNRS) https://orcid.org/0000-0002-4095-4410; Laure Berti-Equille (Espace-Dev (IRD-UM-UG-UR-UA-UNC), Montpellier, FR) https://orcid.org/0000-0002-8046-0570; David Mouillot (MARBEC, University of Montpellier) https://orcid.org/0000-0003-0402-2605.

A Swiss Knife of strategies to solve issues when facing a Reproducibility gap!

Reproduce!, then replicate! With the goal of improving reproducibility!