Should Altruistic Benchmarks be the Default in Machine Learning?

Anonymous Authors[1]

Abstract

Benchmark datasets are used to show the performance of an algorithm, e.g. its accuracy, computational speed, or versatility. In the majority of cases, benchmark datasets currently have no external use, i.e. an improvement on the benchmark doesn't directly translate to a real-world impact. In this paper we explore why this is the case, weigh benefits and harms and propose ways in which benchmark datasets could make a more direct positive impact.

1. Introduction

The primary purpose of benchmark datasets is to compare different algorithms using different metrics such as accuracy, computational speed, robustness, or versatility. In the status quo, most benchmark datasets don't have any external benefit, e.g. an improvement in accuracy on CIFAR100 (Krizhevsky et al., 2014) does not directly translate to an improvement in the world. Potentially, an improvement for image classification could then be applied to more concrete problems but this requires further work or the improvement might not translate.

We think that this poses an altruistic gap and that the Machine Learning community could create benchmark datasets that fulfill their primary purpose *and* also improve a real-world problem. If the benchmark for image classification, for example, was on breast cancer classification, an increase in accuracy would yield direct benefits to the victims of the disease.

In recent years, there has been a quickly growing trend for ethical issues in Artificial Intelligence (AI) and Machine Learning (ML). These include the Fairness (e.g. Barocas et al., 2019), Accountability (e.g. Diakopoulos, 2016) and Transparency (e.g. Weller, 2019) of algorithms but also problems concerning datasets and data collection processes. A dataset might be biased by being unbalanced, e.g. having more data of white than black people which can lead to unfair treatment by the algorithm. Furthermore, Tomašev et al. (2020) coined the term AI for social good (AI4SG) which emphasizes that AI technology should be used to advance important societal causes.

We think that the reasons for creating fair datasets extend to creating altruistic datasets and are aligned with the AI4SG approach. The ML community tries to balance datasets because not doing so would imply harm to marginalized communities and thus a worse world. Similarly, not choosing more altruistic datasets for benchmarking purposes would imply a worse world because our actions could avert harm. The ML community wants to have a positive societal impact and should thus try to optimize every step of the ML pipeline. If there is a pressing problem and relevant data available, the ML community should use it to have a direct impact when new algorithms are proposed.

In the following, we want to define what we mean by 'altruistic' datasets as benchmarks, weigh the potential benefits and harms, explore why they are rarely used in the status quo, and discuss what could be done to change that.

2. Definition

By an altruistic benchmark, we mean a dataset that is frequently used to compare different ML algorithms for which an increase in score directly translates to a desirable social outcome. Examples include the breast cancer classification dataset already mentioned in the introduction, prediction of an individual's poverty to improve the distribution of foreign aid (Visram, 2020), or datasets related to climate change, pandemics global poverty.

Most datasets are somewhat altruistic in the sense that they can be vaguely connected to progress, but they show large differences in degree and effectiveness. The examples from above are all on the direct end of the positive impact spectrum. A dataset that would benchmark an algorithm's ability to predict protein folding, for example, would be somewhere in the middle, as it does not directly improve lives but can be used in further applications to then e.g. improve drug design. Most currently used benchmark datasets, however, are on the indirect end of the spectrum (see Figure 1). For image

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

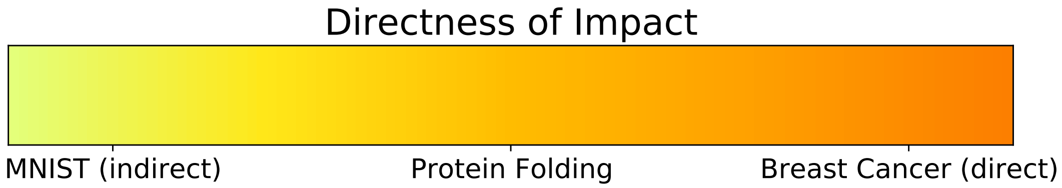

Figure 1. Spectrum for directness of impact with subjective examples. The more direct the impact, the less steps have to be taken between an improvement in score and a positive societal impact.

classification these would be (F-)MNIST (LeCun, 1998; Xiao et al., 2017), CIFAR10 and CIFAR100 (Krizhevsky et al., 2014) or ImageNet (Deng et al., 2009). For regression tasks, these might be UCI datasets (Dua & Graff, 2017) often referred to as Boston, Wine, Concrete, kin8nm, Protein, Year, Naval, Power, and Energy. For object detection tasks the most commonly used benchmark is COCO (Lin et al., 2015).

All of these yield few direct benefits and are mostly used to compare different algorithms with each other. An improvement on COCO could still lead to improvements in the object detection for self-driving vehicles and thus to fewer traffic accidents but the distance from improvement in the score to social benefit is large and requires additional work.

Therefore, when we say altruistic benchmarks we mean datasets that are as close to the direct end of the impact spectrum as possible.

2.1. Criteria

Firstly, an altruistic benchmark should aim to provide a direct positive impact, which could, for example, be defined by the AI4SG criteria or by decreasing the prevalence of a disease. Optimally, the chain of logic between benchmark performance and impact is as short as possible, i.e. the algorithm that is used on the benchmark can directly be applied to the real-world setting.

Secondly, it should fulfill all desiderata for a conventional benchmark. This means it should be easily accessible, the data should be clean and contain meta-information, and it should state a specific benchmark purpose, e.g. it might be intended as a simple task for classification algorithms.

3. Why are altruistic datasets not the norm?

First of all, we want to emphasize that there are datasets with direct benefits that are sometimes used for benchmarking purposes. Kaggle hosted datasets to tackle Malaria (Jolly, 2017), Heart Disease (ronit, 2018) and COVID-19 (Allen Institute for AI, 2020; SRK, 2020). Additionally, Purushotham et al. (2017) have benchmarked Deep Learning models on large healthcare datasets and Olson et al. (2017) have benchmarked 13 different algorithms on 165 datasets, many of which have direct applications in bioinformatics.

There are also some good reasons why some of the current benchmarks exist:

- Often they have a specific purpose. In the case of image classification MNIST is used as a very first test, i.e. if an algorithm can't classify MNIST sufficiently well it's probably a bad classification algorithm. CIFAR10 is a proxy for a slightly more complicated problem, and CIFAR100 and ImageNet are proxies for harder problems.
- Current benchmark datasets with respective purposes are often very well-known. They don't have to be explained or don't lead to confusion and most researchers will have a broad intuition of what good or bad performances on these benchmarks look like without having to read other references, thereby increasing the speed and accuracy of the review process.
- They are often easily accessible. In most major ML frameworks such as PyTorch (Paszke et al., 2017), TensorFlow (Abadi et al., 2015) or Keras (Chollet et al., 2015) importing them is one line of code.
- Most of them are high-quality. No additional data preparation is necessary and people save time.

On the other hand, there are also bad reasons for why some datasets are currently used as benchmarks:

- A reason for the existence of current benchmarks might simply be path dependencies. A small group of researchers might have chosen a certain benchmark more or less randomly or maybe had good reasons at a previous point in time. Other researchers wanted to compare their new algorithms so they used the original benchmark and it became the de facto default.

- This leads to coordination problems. Even when there are good reasons to change this benchmark and evaluate it on other datasets, it would be beneficial for the entire group to switch but too costly/risky for a single researcher since their paper might be rejected for missing a comparison that is expected in their community.
- Bad incentives for already established researchers. When they review new work it is rational for them to expect comparisons on the same benchmarks they used to reduce their workload and increase their chance of being cited, this yields a feedback loop in which established benchmarks are used independently of their purpose or social benefit.

It should be emphasized that all reasons that support the status quo are not inherent to current benchmark datasets but just a result of a long habituation process. If altruistic benchmark datasets were used more frequently, their quality (e.g. preprocessing) could be increased, their access widened and their publicity enlarged. Furthermore, it is likely possible to find or create altruistic datasets with specific purposes such as differing complexities or number of data points.

4. Benefits and harms

Using more altruistic datasets has positive and negative consequences.

Benefits include

- An improvement of the target metric. If an altruistic dataset became a benchmark, more researchers would apply their method to it and more algorithms or training techniques would compete against each other. Otherwise only a small and select group of (often multidisciplinary) researchers applies a small selection of algorithms on the respective dataset and likely achieves suboptimal results.
- A clear and measurable positive effect on many lives through an increase in the target score. For example, fewer people would have cancer, or poverty could be alleviated more effectively.

Harms include

- A risk of overfitting. When a measure becomes a target, it ceases to be a good measure. This means that a specific choice of hyperparameters or even the entire algorithm would learn to solve the problem given by the available dataset but fail to generalize to new data, e.g. to images of breast cancer that are not contained in the training data.
- Benchmarks might be too specific. Algorithms are often supposed to be good at general tasks such as image classification rather than just good at solving one particular problem. If the benchmark becomes too specific they fail to capture how well an algorithm generalizes and thereby lose some of its purposes.
- On a principle level, altruistic benchmarks could collide with value neutrality in ML. As soon as a dataset is labeled as altruistic, it poses a value judgment, which removes its neutrality.

We argue that the benefits outweigh the harms. While overfitting is a problem it can be mitigated e.g. by increasing the size of the validation set. Furthermore, even when trained on an altruistic benchmark, an algorithm is likely to be vetted before its employment in the real world.

The fact that altruistic benchmarks might be too specific could pose a problem but this is arguably already the case with current benchmarks, e.g. MNIST, CIFAR10, or the UCI datasets.

We acknowledge that altruistic benchmarks remove value-neutrality but argue that this is good. Firstly, the vast majority of people agree with value judgments such as “the UN-development goals are good” or “disease is bad” and thus the possibility for conflict is low. Secondly, neutrality is not always a desirable state. If a child is being bullied and we choose to be neutral, we effectively allow/support the bullying. Similarly, choosing to stay neutral on a goal that the vast majority of people deem good implies harm that could be reduced. Thus, we are willing to reduce neutrality to prevent harm.

The benefits, in contrast, are unique to altruistic benchmarks. Since the researchers applying an algorithm to an altruistic cause might have less expertise and knowledge about ML techniques, the improvements might be sizable.

Overall, we estimate that the bottleneck for altruistic benchmarks is less of a theoretical and more of practical nature. Whether an altruistic dataset qualifies as a benchmark mostly depends on whether it fulfills the criteria listed in Section 2.1.

5. What could be changed?

To improve altruistic benchmarks there are multiple things that institutions and researchers could do.

First and foremost, somebody needs to collect the data. Many NGOs, hospitals, governments, and other institutions promoting the public good already own large datasets that are often unconnected with the ML community. Then this data needs to be formatted in a way that is common within the ML community as this increases uptake. This can include anything from providing a clean .csv file to access in an established software package.

Individual researchers can also use or promote well-prepared or high-impact altruistic datasets whenever they come across them by including them in their papers or sharing them with other researchers within the ML community.

The fact that an improvement on an altruistic dataset has such an easily understandable and convincing effect can also be used as an additional argument for ML researchers. If a researcher shows that their algorithm enhances the state of the art of skin cancer detection and thereby directly improves people's lives, this is a good argument for the paper.

All in all, we estimate that there is an altruistic gap when it comes to benchmarks. In some cases, it will be impossible to capture some of the more abstract desiderata of benchmarks but in most cases, it should be possible to use an altruistic benchmark without large academic losses while passively improving society.

6. Conclusion

In this essay, we evaluated whether datasets that provide a direct positive impact for society, i.e. altruistic datasets, should become the norm to benchmark ML algorithms. We conclude that there are some cases in which these datasets are not applicable because they might be insufficiently general. However, there is a large group of cases in which altruistic benchmarks could be used without any losses and thereby passively improve society.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL http://tensorflow.org/. Software available from tensorflow.org.

Allen Institute for AI. Covid-19 open research dataset challenge (cord-19), 2020. URL https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge.

Barocas, S., Hardt, M., and Narayanan, A. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

Chollet, F. et al. Keras, 2015. URL https://github.com/fchollet/keras.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

Diakopoulos, N. Accountability in algorithmic decision making. *Commun. ACM*, 59(2):56–62, January 2016. ISSN 0001-0782. doi: 10.1145/2844110. URL https://doi.org/10.1145/2844110.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Jolly, T. The fight against malaria, 2017. URL https://www.kaggle.com/teajay/the-fight-against-malaria.

Krizhevsky, A., Nair, V., and Hinton, G. The cifar-10 dataset. *online: http://www. cs. toronto. edu/kriz/cifar. html*, 55, 2014.

LeCun, Y. The MNIST database of handwritten digits. *http://yann.lecun.com/exdb/mnist/*, 1998.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. Microsoft coco: Common objects in context, 2015.

Olson, R. S., La Cava, W., Orzechowski, P., Urbanowicz, R. J., and Moore, J. H. Pmlb: a large benchmark suite for machine learning evaluation and comparison. *BioData Mining*, 10(1):36, Dec 2017. ISSN 1756-0381. doi: 10.1186/s13040-017-0154-4. URL https://doi.org/10.1186/s13040-017-0154-4.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. 2017.

Purushotham, S., Meng, C., Che, Z., and Liu, Y. Benchmark of deep learning models on large healthcare mimic datasets, 2017.

ronit. Heart disease uci, 2018. URL https://www.kaggle.com/ronitf/heart-disease-uci.

SRK. Novel corona virus 2019 dataset, 2020. URL https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset.

Tomašev, N., Cornebise, J., Hutter, F., Mohamed, S., Picciariello, A., Connelly, B., Belgrave, D. C. M., Ezer, D., van der Haert, F. C., Mugisha, F., Abila, G., Arai, H., Almiraat, H., Proskurnia, J., Snyder, K., Otake-Matsuura, M., Othman, M., Glasmachers, T., de Wever, W., Teh, Y. W., Khan, M. E., Winne, R. D., Schaul, T., and Clopath,

C. Ai for social good: unlocking the opportunity for positive impact, 2020.

Visram, T. How givedirectly is finding the poorest people in the world—and sending them cash, 2020. URL https://www.fastcompany.com/90585079/how-givedirectly-is-finding-the-poorest-people-in-the-world-and-sending-them-cash.

Weller, A. Transparency: Motivations and challenges, 2019.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv*, abs/1708.07747, 2017.