The rise and fall (and rise) of datasets

Growing criticisms of datasets that were built from user-generated data scraped from the web have led to the retirement or redaction of many popular benchmarks. Their afterlife, as copies or subsets that continue to be used, is a cause for concern.

The fast pace of development in machine learning research in the past two decades has been, in large part, fuelled by the availability of large benchmark datasets of images, videos, text and more. These make it possible to compare and evaluate algorithms, and help to define research goals. However, in recent years, the machine learning community has identified an alarming number of potential legal and ethical problems with many of the most popular image datasets, such as representational harms, effects of bias, privacy infringement and unclear or dubious downstream use[1,2].

Widely used datasets such as ImageNet, Tiny Images, Megaface and MS-Celeb-1M typically contain images scraped from the internet, in particular from sharing platforms such as Flickr. This often happens without the explicit permission from, or even awareness of, the people who generated the data. Training machine learning algorithms on copyrighted data is generally considered 'fair use' on the basis that it amounts to transformative use of the original data. This principle was boosted by a 2015 US court ruling in the case of Authors Guild versus Google. The former challenged Google's right to scan books for their book search algorithms, but the court ruled that it is not illegal to scan copyrighted books for data mining purposes and to develop search algorithms.

Moreover, photos and other user-generated data are often posted on platforms such as Flickr with Creative Commons licenses, which go beyond restrictive copyright and encourage sharing and re-use. However, neither the fair-use principle nor Creative Commons licenses should be taken to imply that such content is up for grabs, as there are many ethical, legal and technical issues to consider beyond copyright. Several recent surveys[1–4] point to a range of concerns, such as in the case of ImageNet, which was created over a decade ago and is one of the most influential computer vision datasets. It contains 14 million images, hand-annotated by crowdsourced Amazon Mechanical Turk (MTurk) workers, and has more than 20,000 categories. Recent analyses, including by the creators of ImageNet themselves[5], revealed that there are many problematic annotations, in particular offensive and biased ones. More than half of the labels in the people subtree were considered potentially harmful, and as a result 600,000 images were removed from ImageNet.

An underlying, fundamental issue that has become clear over the years is that datasets are not neutral, but represent particular social and political norms, which can specifically affect marginalized groups[4]. With the benefit of hindsight, there should have been concerns about the ethics of taking user-generated data from the web, crowdsourcing non-expert labellers and giving unrestricted access to developers — including those who work on sensitive applications such as facial recognition and biometric surveillance. Take, for example, Microsoft Celeb (MS-Celeb-1M), which is a dataset of 10 million face images taken from the Internet. Although most of the images are photos of actors, many other individuals are included who have an online professional presence, such as journalists, human rights activists, academics, authors and more. A recent report, after which Microsoft took the dataset down, indicated that the images were used without the individuals' knowledge or consent in facial recognition applications by various organizations including Huawei, Sensetime and IBM.

Several datasets have now been taken down or, as in the case of ImageNet, have been heavily redacted. In practice, however, they continue to be widely used and available, either in their original form, such as via online torrents, or in derivative form, as subsets or modifications of the original dataset or models pretrained on the deprecated dataset[1]. In many cases, the deprecation has been silent, or the status of the dataset left ambiguous. For instance, Microsoft took the MS-Celeb-1M dataset website down, stating that the project was finished but, to date, a clear public announcement is missing and the dataset still exists in various repositories. Another example is MegaFace, which has a landing page with the statement that the dataset is decommissioned, but without alluding to the ethical concerns raised about it, such as in a recent *New York Times* article.

A more encouraging example is the Tiny Images dataset, in which the MIT hosts make an announcement on the landing page that the dataset is withdrawn, clearly citing the ethical concerns about the dataset that were raised in a recent analysis[3] and asking researchers not to use the dataset. However, Correy et al.[4] report that many prominent retracted datasets still have an active afterlife, which leads to the propagation of the identified harms, and they argue that a consistent approach to retraction is required. For example, hosts need to make a clear announcement that describes the reasons for deprecation, and they should have a clear execution plan and timeline for the deprecation. The authors further argue that a central repository, maintained by the machine learning community, is required to host deprecated datasets.

There is also a role for conferences and journals. In particular, submission guidelines should request that authors list and describe the datasets generated and analysed, and authors need to ensure that none of the datasets they used has been retracted. Like other *Nature Research* journals, *Nature Machine Intelligence* pays particular attention to data citation and data availability statements, but will also monitor the use of retracted datasets in manuscripts that are sent out for peer review, and where necessary request authors to use alternative datasets. There may be exceptions where the use of deprecated datasets could be allowed — for example, when bias and the harmful effects of datasets are studied. We will enlist the advice of experts in such cases.

Moving forward, a fundamental change in dataset culture is necessary[1,2]. Peng et al.[1] emphasize that harm mitigation and stewardship are required throughout the life cycle of a dataset since the ethical impacts are hard to anticipate and address at the time of dataset building, and ethical and social norms may also change over time. Creators need to monitor the use of their datasets, make updates to licenses and documentation and limit access when necessary. We will follow the developing community standards closely and support authors in responsible dataset reporting. ❐

Published online: 25 January 2022
https://doi.org/10.1038/s42256-022-00442-2

References

1. Peng, K., Mathur, A. & Narayanan, A. Preprint at https://arxiv.org/abs/2108.02922 (2021).
2. Paullada, A. et al. *Patterns (N Y)*. **2**, 100336 (2021).
3. Prabhu, V. U. & Birhane, A. Preprint at https://arxiv.org/abs/2006.16923 (2020).
4. Correy, F. et al. Preprint at https://arxiv.org/abs/2111.04424 (2021).
5. Yang, K. et al. in Proc. 2020 *Conference on Fairness, Accountability and Transparency* 547–558 (2020).