Shared, but not up for grabs

The online availability of large amounts of publicly posted images and other data is fuelling machine learning research and applications. However, it is time to take privacy concerns seriously.

The World Wide Web is a treasure trove of easily discovered information. It has provided us with many ways to exchange knowledge and experiences, enabling the acceleration of scientific advances, supporting social networks and making educational resources widely available, among other undeniable benefits. However, the appetite for public sharing of information, data, images, videos and more is catching up with the world as it becomes increasingly apparent that technological innovation moves at a different speed to the ethical guidelines for handling all this information responsibly and safely.

Many embraced the spirit of online sharing early on and have unconcernedly posted images, tweets and other information, often containing personal details, for over a decade. However, few may have been able to anticipate how currently much of this personal information is swept up in large amounts by those with the means, technology and expertise. The motives of large-scale data collection by companies and governmental bodies may be legit, and either of commercial or non-profit nature, but increasingly concerns arise about the ethics involved when using publicly posted personal data. Some of these concerns were spelled out clearly in a story from NBC News that received wide attention.

The article discusses how technology companies are building databases of images collected from the well-known photo-sharing platform Flickr for training face-recognition machine learning algorithms. To make these algorithms work well and with as little bias as possible, a large pool of faces of various age, skin tone and gender needs to be collected. At first sight, there seems little harm done: the images concerned were after all posted with a Creative Commons licence. Moreover, the photos are not linked to any personal information. However, the news story warns, the people in the photographs have not been given an opportunity to give consent to being part of the development of a technology that can be used for invasive applications such as in surveillance. A related concern is that photos might be labelled in a way that reflects negatively on the human in question who, again, has not been given a chance to opt in or out.

Collecting images from Flickr to build databases for training machine learning algorithms is popular as they are often posted with a Creative Commons licence, which permits in most cases use and modification of the images by anyone and for any purpose. The Creative Commons licences were introduced in the early 2000s to make sharing creative works of many different types easier and avoid restrictive copyright. However, as Creative Commons put it in a statement posted soon after the NBC News story, copyright is not a good tool to protect individual privacy or to address research ethics in AI development. These issues belong to the "public policy space", the statement says.

These policy issues will be difficult to navigate and may take time to fully resolve. In the meantime, it is desirable to adopt high standards in responsible reuse of publicly posted images. The reality is that even while millions of images have been posted under the Creative Common licences, many of the photographers did not anticipate current uses of their images. Moreover, persons appearing in the photographs may not know these images exist and may not, if given a choice, consent to reuse. Good practice is then to obtain explicit permission for the specific type of intended reuse when dealing with images containing identifiable humans.

A problem is that image databases used in machine learning, containing images of humans or anything else, often don't contain easily accessible information on the origin and precise copyright information of those images. This may not be an issue for most academic machine learning research, but clarifying copyright restrictions becomes urgent when research involves commercial interests, which is in reality often a grey area.

One effort to resolve these questions is presented in a collaborative article called 'Datasheets for datasets' by, among others, researchers from Microsoft Research and the AI Now Institute. The authors propose a way to standardize information on how and why a dataset was created, what information it contains, what tasks it should and should not be used for, and whether it might raise any ethical or legal concerns. The idea is borrowed from the electronics industry, where datasheets contain detailed information about the characteristics, recommended use and origin of every hardware component. Such datasheets will help prioritize transparency and accountability in the reuse of images, of particular importance for human-centric databases.

Researchers should also feel encouraged to take questions about reuse of images to their institutes' ethics boards. It is common practice in the life sciences to consult ethics boards whenever human data are concerned. Ethical collection and use of data is possible, and with good practices in treating personal images and data responsibly, we can continue to make the most of the information available for free online. ❐

Published online: 9 April 2019
https://doi.org/10.1038/s42256-019-0047-y