Title: 80 Million Tiny Images

URL: https://en.wikipedia.org/wiki/80_Million_Tiny_Images

PageID: 64439717

Categories: Category:Datasets in computer vision, Category:Machine learning

---

80 Million Tiny Images is a dataset intended for training machine-learning systems constructed by Antonio Torralba, Rob Fergus , and William T. Freeman in a collaboration between MIT and New York University . It was published in 2008.

The dataset has size 760 GB . It contains 79,302,017 32×32-pixel color images, scaled down from images scraped from the World Wide Web over 8 months. The images are classified into 75,062 classes. Each class is a non-abstract noun in WordNet . Images may appear in more than one class. The dataset was motivated by non-parametric models of neural activations in the visual cortex upon seeing images.

The CIFAR-10 dataset uses a subset of the images in this dataset, but with independently generated labels, as the original labels were not reliable. The CIFAR-10 set has 6000 examples of each of 10 classes, and the CIFAR-100 set has 600 examples of each of 100 non-overlapping classes.

Construction

It was first reported in a technical report in April 2007, during the middle of the construction process, when there were only 73 million images. The full dataset was published in 2008.

They began with all 75,846 non-abstract nouns in WordNet , and then for each of these nouns, they scraped 7 image search engines: Altavista , Ask.com , Flickr , Cydral , Google , Picsearch , and Webshots . After 8 months of scraping, they obtained 97,245,098 images. Since they did not have enough storage, they downsized the images to 32×32 as they were scraped. After gathering, they removed images with zero variance and intra-word duplicate images, resulting in the final dataset.

Out of the 75,846 nouns, only 75,062 classes had any results, so the other nouns did not appear in the final dataset.

The number of images per noun follows a Zipf-like distribution , with 1056 images per noun on average. To prevent a few nouns taking up too many images, they put an upper bound of at most 3000 images per noun.

Retirement

The 80 Million Tiny Images dataset was retired from use by its creators in 2020, after a paper by researchers Abeba Birhane and Vinay Prabhu found that some of the labeling of several publicly available image datasets, including 80 Million Tiny Images, contained racist and misogynistic slurs which were causing models trained on them to exhibit racial and sexual bias. The dataset also contained offensive images. Following the release of the paper, the dataset's creators removed the dataset from distribution, and requested that other researchers not use it for further research and to delete their copies of the dataset.

See also

List of datasets in computer vision and image processing

References

External links