engadget

ng   **Entertainment**   **Products**   **Tomorrow**   **Audio**   **Video**   **Deals**   **Buyer's Guide**   ⌄
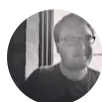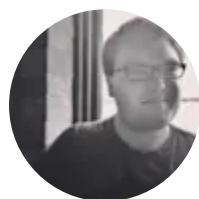
MIT study finds labelling errors in datasets used to test AI

Over three percent of data in the most-cited datasets was deemed inaccurate or mislabeled.

K. Holt | 03.29.21
@krisholt

K. Holt
@krisholt
March 29th, 2021

ImageNet given label:
alp

We guessed: **ski**

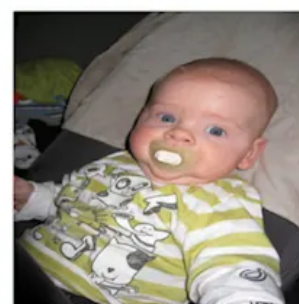MTurk consensus: **Both alp and ski**

ImageNet given label:
schooner

We guessed: **yawl**

MTurk consensus: **yawl**

ImageNet given label:
nipple

We guessed: **pajamas**

MTurk consensus: **Neither nipple nor pajamas**

ImageNet/MIT

In this article: errors, news, gear, study, mit, mislabeling, dataset, ai...

A team led by computer scientists from MIT examined
t   f th   t it d d t   t   d t   t   hi

The datasets, which have each been cited more than 100,000 times, include text-based ones from newsgroups, Amazon and IMDb. Errors emerged from issues like Amazon product reviews being mislabeled as positive when they were actually negative and vice versa.

Some of the image-based errors result from mixing up animal species. Others arose from mislabeling photos with less-prominent objects ("water bottle" instead of the mountain bike it's attached to, for instance). One particularly galling example that emerged was a baby being confused for a nipple.

One of the datasets centers around audio from YouTube videos. A clip of a YouTuber talking to the camera for three and a half minutes was labeled as "church bell," even though one could only be heard in the last 30 seconds or so. Another error emerged from a misclassification of a Bruce Springsteen performance as an orchestra.

To find possible errors, the researchers used a framework called confident learning, which examines datasets for label noise (or irrelevant data). They validated the possible mistakes using Mechanical Turk, and found around 54 percent of the data that the algorithm flagged had incorrect labels. The researchers found the QuickDraw test set had the most errors with around 5

correct). Sometimes, the confident learning approach got it wrong too, like confusing a correctly labeled image of tuning forks for a menorah.

If labels are even a little off, that could lead to huge ramifications for machine learning systems. If an AI system can't tell the difference between a grocery and a bunch of crabs, it'd be hard to trust it with pouring you a drink.

Popular on Engadget
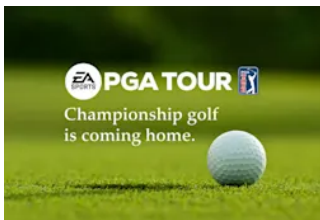
EA will return to golf games with a new PGA Tour title

Misusing Ford's 2021 F-150 pickup generator

Cyberpunk 2077's massive 1.2 patch is finally here

New York rolls out first digital COVID-19 vaccine passports in the US

Conversation　3 Comments