Major machine learning datasets have tens of thousands of errors

March 30 2021, by Adam Conner-Simons

Credit: MIT Computer Science & Artificial Intelligence Lab

It's well-known that machine learning datasets have their fair share of errors, including mislabeled images. But there hasn't been much research to systematically quantify just how error-ridden they are.

Further, prior work has focused on errors in the training data of ML datasets. But the test sets are what we benchmark the state of machine learning with, and no study has looked at systematic error across ML test sets—the sets we rely on to understand how well ML models work.

In a new paper, a team led by researchers at MIT's Computer Science and Artificial Intelligence Lab (CSAIL) looked at 10 major datasets that have been cited over 100,000 times and that include ImageNet and Amazon's reviews dataset.

The researchers found a 3.4% average error rate across all datasets, including 6% for ImageNet, which is arguably the most widely used dataset for popular image recognition systems developed by the likes of Google and Facebook.

Even the seminal MNIST digits dataset, which has served as the bedrock of optical digit recognition for the past 20 years and has been benchmarked in tens of thousands of peer-reviewed ML publications, contains 15 (human-validated) label errors in the test set.

The team also created a demo that lets users peruse the different datasets to sample the different types of errors that occur, including:

- mislabeled images, like one breed of dog being confused for another or a baby being confused for a nipple.
- mislabeled text sentiment, like Amazon product reviews described as negative when they were actually positive.
- mislabeled audio of YouTube videos, like an Ariana Grande high-note being classified as a whistle.

Co-author Curtis Northcutt says that one surprise from their findings was that weaker models like ResNet-18 often had lower error rates than

more complex models such as ResNet-50, depending on the prevalence of irrelevant data ("noise"). Northcutt recommends that ML practitioners consider using simple models if their real-world dataset has a label error rate of 10%.

The team's results build upon a wealth of work done at MIT in creating "confident learning," a sub-field of machine learning that looks at datasets to find and quantify label noise. With this project, confident learning is used to algorithmically identify all of the label errors prior to human verification.

The team has also made it easy for other researchers to replicate their results and find label errors in their own datasets using cleanlab, an open-source python package.

Provided by Massachusetts Institute of Technology