Making sense of AI

3 big problems with datasets in AI and machine learning

Kyle Wiggers
@Kyle_L_Wiggers

December 17, 2021 6:30 AM

Image Credit: Getty Images

Datasets fuel AI models like gasoline (or electricity, as the case may be) fuels cars. Whether they're tasked with generating text, recognizing objects, or predicting a company's stock price, AI systems "learn" by sifting through countless examples to discern patterns in the data. For

example, a computer vision system can be trained to recognize certain types of apparel, like coats and scarfs, by looking at different images of that clothing.

Beyond developing models, datasets are used to test trained AI systems to ensure they remain stable — and measure overall progress in the field. Models that top the leaderboards on certain open source benchmarks are considered state of the art (SOTA) for that particular task. In fact, it's one of the major ways that researchers determine the predictive strength of a model.

But these AI and machine learning datasets — like the humans that designed them — aren't without their flaws. Studies show that biases and mistakes color many of the libraries used to train, benchmark, and test models, highlighting the danger in placing too much trust in data that hasn't been thoroughly vetted — even when the data comes from vaunted institutions.

1. The training dilemma

In AI, benchmarking entails comparing the performance of multiple models designed for the same task, like translating words between languages. The practice — which originated with academics exploring early applications of AI — has the advantages of organizing scientists around shared problems while helping to reveal how much progress has been made. In theory.

But there are risks in becoming myopic in dataset selection. For example, if the same training dataset is used for many kinds of tasks, it's unlikely that the dataset will accurately reflect the data that models see in the real world. Misaligned datasets can distort the measurement of scientific progress, leading researchers to believe they're doing a better job than they actually are — and causing harm to people in the real world.

Researchers at the University of California, Los Angeles, and Google investigated the problem in a recently published study titled "Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research." They found that there's "heavy borrowing" of datasets in machine learning — e.g., a community working on one task might borrow a dataset created for another task — raising concerns about misalignment. They also showed that only a dozen universities and corporations are responsible for creating the datasets used more than 50% of the time in machine learning, suggesting that these institutions are effectively shaping the research agendas of the field.

"SOTA-chasing is bad practice because there are too many confounding variables, SOTA usually doesn't mean anything, and the goal of science should be to accumulate knowledge as opposed to results in specific toy benchmarks," Denny Britz, a former resident on the Google Brain team, told VentureBeat in a previous interview. "There have been some initiatives to improve things, but looking for SOTA is a quick and easy way to review and evaluate papers. Things like these are

embedded in culture and take time to change."

To their point, ImageNet and Open Images — two publicly available image datasets from Stanford and Google — are heavily U.S.- and Euro-centric. Computer vision models trained on these datasets perform worse on images from Global South countries. For example, the models classify grooms from Ethiopia and Pakistan with lower accuracy compared with grooms from the U.S., and they fail to correctly identify objects like "wedding" or "spices" when they come from the Global South.

Even differences in the sun path between the northern and southern hemispheres and variations in background scenery can affect model accuracy, as can the varying specifications of camera models like resolution and aspect ratio. Weather conditions are another factor — a driverless car system trained exclusively on a dataset of sunny, tropical environments will perform poorly if it encounters rain or snow.

A recent study from MIT reveals that computer vision datasets including ImageNet contain problematically "nonsensical" signals. Models trained on them suffer from "overinterpretation," a phenomenon where they classify with high confidence images lacking in so much detail that they're meaningless to humans. These signals can lead to model fragility in the real world, but they're valid in the datasets — meaning overinterpretation can't be identified using typical methods.

"There's the question of how we can modify the datasets in a way that would enable models to be trained to more closely mimic how a human would think about classifying images and therefore, hopefully, generalize better in these real-world scenarios, like autonomous driving and medical diagnosis, so that the models don't have this nonsensical behavior," says Brandon Carter, an MIT Ph.D. student and lead author of the study, said in a statement.

History is filled with examples of the consequences of deploying models trained using flawed datasets, like virtual backgrounds and photo-cropping tools that disfavor darker-skinned individuals. In 2015, a software engineer pointed out that the image-recognition algorithms in Google Photos were labeling his black friends as "gorillas." And the nonprofit AlgorithmWatch showed that Google's Cloud Vision API at one time labeled thermometers held by a black person as "guns" while labeling thermometers held by a light-skinned person as "electronic devices."

Dodgy datasets have also led to models that perpetuate sexist recruitment and hiring, ageist ad targeting, erroneous grading, and racist recidivism and loan approval. The issue extends to health care, where training datasets containing medical records and imagery mostly come from patients in North America, Europe, and China — meaning models are less likely to work well for underrepresented groups. The imbalances are evident in shoplifter- and weapon-spotting computer vision models, workplace safety monitoring software, gunshot sound detection systems, and "beautification" filters, which amplify the biases present in the data on which they

were trained.

Experts attribute many errors in facial recognition, language, and speech recognition systems, too, to flaws in the datasets used to train the models. For example, a study by researchers at the University of Maryland found that face-detection services from Amazon, Microsoft, and Google are more likely to fail with older, darker-skinned individuals and those who are less "feminine-presenting." According to the Algorithmic Justice League's Voice Erasure project, speech recognition systems from Apple, Amazon, Google, IBM, and Microsoft collectively achieve word error rates of 35% for black voices versus 19% for white voices. And language models have been shown to exhibit prejudices along race, ethnic, religious, and gender lines, associating Black people with more negative emotions and struggling with "black-aligned English."

"Data [is] being scraped from many different places on the web [in some cases], and that web data reflects the same societal-level prejudices and biases as hegemonic ideologies (e.g., of whiteness and male dominance)," UC Los Angeles' Bernard Koch and Jacob G. Foster and Google's Emily Denton and Alex Hanna, the coauthors of "Reduced, Reused, and Recycled," told VentureBeat via email. "Larger ... models require more training data, and there has been a struggle to clean this data and prevent models from amplifying these problematic ideas."

2. Issues with labeling

Labels, the annotations from which many models learn relationships in data, also bear the hallmarks of data imbalance. Humans annotate the examples in training and benchmark datasets, adding labels like "dogs" to pictures of dogs or describing the characteristics in a landscape image. But annotators bring their own biases and shortcomings to the table, which can translate to imperfect annotations.

For instance, studies have shown that the average annotator is more likely to label phrases in African-American Vernacular English (AAVE), the informal grammar, vocabulary, and accent used by some Black Americans, as toxic. In another example, a few labelers for MIT's and NYU's 80 Million Tiny Images dataset — which was taken offline in 2020 — contributed racist, sexist, and otherwise offensive annotations including nearly 2,000 images labeled with the N-word and labels like "rape suspect" and "child molester."

In 2019, *Wired* reported on the susceptibility of platforms like Amazon Mechanical Turk — where many researchers recruit annotators — to automated bots. Even when the workers *are* verifiably human, they're motivated by pay rather than interest, which can result in low-quality data — particularly when they're treated poorly and paid a below-market rate. Researchers including Niloufar Salehi have made attempts at tackling Amazon Mechanical Turk's flaws with

efforts like Dynamo, an open access worker collective, but there's only so much they can do.

Being human, annotators also make mistakes — sometimes major ones. In an MIT analysis of popular benchmarks including ImageNet, the researchers found mislabeled images (like one breed of dog being confused for another), text sentiment (like Amazon product reviews described as negative when they were actually positive), and audio of YouTube videos (like an Ariana Grande high note being categorized as a whistle).

One solution is pushing for the creation of more inclusive datasets, like MLCommons' People's Speech Dataset and the Multilingual Spoken Words Corpus. But curating these is time-consuming and expensive, often with a price tag reaching into a range of millions of dollars. Common Voice, Mozilla's effort to build an open source collection of transcribed speech data, has vetted only dozens of languages since its 2017 launch — illustrating the challenge.

One of the reasons creating a dataset is so costly is the domain expertise required for high-quality annotations. As Synced noted in a recent piece, most low-cost labelers can only annotate relatively "low-context" data and can't handle "high-context" data such as legal contract classification, medical images, or scientific literature. It's been shown that drivers tend to label self-driving datasets more effectively than those without driver's licenses and that doctors, pathologists, and radiologists perform better at accurately labeling medical images.

Machine-assisted tools could help to a degree by eliminating some of the more repetitive work from the labeling process. Other approaches, like semi-supervised learning, promise to cut down on the amount of data required to train models by enabling researchers to "fine-tune" a model on small, customized datasets designed for a particular task. For example, in a blog post published this week, OpenAI says that it managed to fine-tune GPT-3 to more accurately answer open-ended questions by copying how humans research answers to questions online (e.g., submitting search queries, following links, and scrolling up and down pages) and citing its sources, allowing users to give feedback to further improve the accuracy.

Still other methods aim to replace real-world data with partially or entirely synthetic data — although the jury's out on whether models trained on synthetic data can match the accuracy of their real-world-data counterparts. Researchers at MIT and elsewhere have experimented using random noise alone in vision datasets to train object recognition models.

In theory, unsupervised learning could solve the training data dilemma once and for all. In unsupervised learning, an algorithm is subjected to "unknown" data for which no previously defined categories or labels exist. But while unsupervised learning excels in domains for which a lack of labeled data exists, it's not a weakness. For example, unsupervised computer vision systems can pick up racial and gender stereotypes present in the unlabeled training data.

3. A benchmarking problem

The issues with AI datasets don't stop with training. In a study from the Institute for Artificial Intelligence and Decision Support in Vienna, researchers found inconsistent benchmarking across more than 3,800 AI research papers — in many cases attributable to benchmarks that didn't emphasize informative metrics. A separate paper from Facebook and the University College London showed that 60% to 70% of answers given by natural language models tested on "open-domain" benchmarks were hidden somewhere in the training sets, meaning that the models simply memorized the answers.

In two studies coauthored by Deborah Raji, a tech fellow in the AI Now Institute at NYU, researchers found that benchmarks like ImageNet are often "fallaciously elevated" to justify claims that extend beyond the tasks for which they were originally designed. That's setting aside the fact that "dataset culture" can distort the science of machine learning research, according to Raji and the other coauthors — and lacks a culture of care for data subjects, engendering poor labor conditions (such as low pay for annotators) while insufficiently protecting people whose data is intentionally or unintentionally swept up in the datasets.

Several solutions to the benchmarking problem have been proposed for specific domains, including the Allen Institute's GENIE. Uniquely, GENIE incorporates both automatic and manual testing, tasking human evaluators with probing language models according to predefined, dataset-specific guidelines for fluency, correctness, and conciseness. While GENIE is expensive — it costs around $100 to submit a model for benchmarking — the Allen Institute plans to explore other payment models, such as requesting payment from tech companies while subsidizing the cost for small organizations.

There's also growing consensus within the AI research community that benchmarks, particularly in the language domain, must take into account broader ethical, technical, and societal challenges if they're to be useful. Some language models have large carbon footprints, but despite widespread recognition of the issue, relatively few researchers attempt to estimate or report the environmental cost of their systems.

"[F]ocusing only on state-of-the-art performance de-emphasizes other important criteria that capture a significant contribution," Koch, Foster, Denton, and Hanna said. "[For example,] SOTA benchmarking encourages the creation of environmentally-unfriendly algorithms. Building bigger models has been key to advancing performance in machine learning, but it is also environmentally unsustainable in the long run ... SOTA benchmarking [also] does not encourage scientists to develop a nuanced understanding of the concrete challenges presented by their task in the real world, and instead can encourage tunnel vision on increasing scores. The requirement to achieve SOTA constrains the creation of novel algorithms or algorithms which can solve real-

world problems."

Possible AI datasets solutions

Given the extensive challenges with AI datasets, from imbalanced training data to inadequate benchmarks, effecting meaningful change won't be easy. But experts believe that the situation isn't hopeless.

Arvind Narayanan, a Princeton computer scientist who has written several works investigating the provenance of AI datasets, says that researchers must adopt responsible approaches not only to collecting and annotating data, but also to documenting their datasets, maintaining them, and formulating the problems for which their datasets are designed. In a recent study he coauthored, Narayanan found that many datasets are prone to mismanagement, with creators failing to be precise in license language about how their datasets can be used or prohibit potentially questionable uses.

"Researchers should think about the different ways their dataset can be used ... Responsible dataset 'stewarding,' as we call it, requires addressing broader risks," he told VentureBeat via email. "One risk is that even if a dataset is created for one purpose that appears benign, it might be used unintentionally in ways that can cause harm. The dataset could be repurposed for an ethically dubious research application. Or, the dataset could be used to train or benchmark a commercial model when it wasn't designed for these higher-stakes settings. Datasets typically take a lot of work to create from scratch, so researchers and practitioners often look to leverage what already exists. The goal of responsible dataset stewardship is to ensure that this is done ethically."

Koch and coauthors believe that people — and organizations — need to be rewarded and supported for creating new, diverse datasets contextualized for the task at hand. Researchers need to be incentivized to use "more appropriate" datasets at academic conferences like NeurIPS, they say, and encouraged to perform more qualitative analyses — like the interpretability of their model — as well as report metrics like fairness (to the extent possible) and power efficiency.

NeurIPS — one of the largest machine learning conferences in the world — mandated that coauthors who submit papers must state the "potential broader impact of their work" on society, beginning with NeurIPS 2020 last year. The pickup has been mixed, but Koch and coauthors believe that it's a small step in the right direction.

"[M]achine learning researchers are creating a lot of datasets, but they're not getting used. One of the problems here is that many researchers may feel they need to include the widely used benchmark to give their paper credibility, rather than a more niche but technically appropriate benchmark," they said. "Moreover, professional incentives need to be aligned towards the

creation of these datasets ... We think there is still a portion of the research community that is skeptical of ethics reform, and addressing scientific issues might be a different way to get these people behind reforms to evaluation in machine learning."

There's no simple solution to the dataset annotation problem — assuming that labeling isn't eventually replaced by alternatives. But a recent paper from Google suggests that researchers would do well to establish "extended communications frameworks" with annotators, like chat apps, to provide more meaningful feedback and clearer instructions. At the same time, they must work to acknowledge (and actually account for) workers' sociocultural backgrounds, the coauthors wrote — both from the perspective of data quality and societal impact.

The paper goes further, providing recommendations for dataset task formulation and choosing annotators, platforms, and labeling infrastructure. The coauthors say that researchers should consider the forms of expertise that could be incorporated through annotation, in addition to reviewing the intended use cases of the dataset. They also say that they should compare and contrast the minimum pay requirements across different platforms and analyze disagreements between annotators of different groups, allowing them to — hopefully — better understand how different perspectives are or aren't represented.

"If we really want to diversify the benchmarks in use, government and corporate players need to create grants for dataset creation and distribute those grants to under-resourced institutions and researchers from underrepresented backgrounds," Koch and coauthors said. "We would say that there is abundant research now showing ethical problems and social harms that can arise from data misuse in machine learning ... Scientists like data, so we think if we can show them how over-usage isn't great for science, it might spur further reform that can mitigate social harms as well."

-
-
-