July 20, 2022

When training AI models, is a bigger dataset better?

Many enterprises assume that more training data will improve their AI, but dataset size is just one of many factors that influence accuracy.

LESSONS FOR LEADERS

- More training data improves AI performance up to a certain point but can compromise performance beyond it.
- The quality of the data used to train AI is just as important as the quantity. Poor data quality leads to poor AI results.
- A large, diverse dataset that captures all the qualities you expect to encounter in the real world is critical for building an effective AI.

Every new iteration of GPT-3 and other AI frameworks tout how big their training data is, to the point where the industry seems engaged in one-upmanship over whose training dataset is biggest. It sure sounds impressive, but is it guaranteed that mountains of data will translate to better AI performance?

In principle, the idea that more data is always desirable makes sense. Data drives business intelligence, and the size of training data is a critical factor in shaping the predictive power of AI models. That's why machine learning models are typically trained with mountains of data—feeding the algorithm more data points generally increases information about the dataset and sharpens the accuracy of the model. Problems arise, however, when the uncertainty around how much data it will take to handle a given task causes companies to indiscriminately throw ever more data at the model, creating a new set of issues.

"When your model has too much data, it can begin to suffer from high variance, which means that it's paying too much attention to details that are specific to the training data and not enough attention to overall patterns in the data," says Eric McGee, a senior network engineer at TRG Datacenters and a certified AI and machine learning expert. This condition is called overfitting. Overfitted models have high error rates because they place too much emphasis on each data point, so just adding more training data may only exacerbate the problem if the model doesn't have enough complexity to learn the underlying structure of the data.

Please read: Accelerating AI is not a one size fits all solution

Too much data can also compromise model performance when its diversity and distribution don't represent the real world. If your training data already has a thousand photos of a white cat, adding a thousand more photos of the same white cat is unlikely to improve results. And if only 1.5 percent of a population has red hair but 70 percent of the images in your dataset are of people with red hair, your model will produce misleading results.

Even when too much data doesn't directly impact an AI model's predictive accuracy, it can waste time, power, and money if there is nothing new to learn in the data, making the model more expensive and slower to build than it should be.

"It becomes very difficult to manage very large datasets, and then it becomes difficult for these large models to get to an actual result," says Ryan Ries, practice lead for data, analytics, and machine learning at Mission Cloud Services. "Special data pipelines and techniques are needed to pass extremely large datasets into fitting algorithms. GPT-3 has become so massive that it has 175 billion parameters and has to be trained across hundreds of GPUs—and for many months—to get the final result."

> Before you just blindly add more data, think about what kind of data you're adding and whether it has any relevance to your end goal before you start collecting it.
>
> **ERIC MCGEE** SENIOR NETWORK ENGINEER, TRG DATACENTERS

Size isn't all that matters

Many factors beyond the sheer volume of training data impact the performance of AI. Some of those include the quality of the data labels being used, the size of the AI model, and the amount of time you train the model.

Perhaps none is as important, though, as the quality of the data being used. Studies show that more than 90 percent of AI practitioners experience significant downstream data issues related to data quality, and poor data quality is estimated to cost companies 20 percent of revenue.

Please read: What is data cleansing, and why does your company need it?

"Data quality matters in all instances," says David Shrier, professor of practice, AI, and innovation at Imperial College Business School, where he directs the Translational AI Lab. "Designing a good AI system embodies one of the truisms of computer science: Garbage in equals garbage out. You won't get the right answers if you don't design the right dimensions of your dataset, including all these other factors aside from size."

Data is generally considered high quality if it is well suited to serve its specific purpose. There are many characteristics of data quality, including consistency, accuracy, validity, uniqueness, completeness, and timeliness. Shortfalls in any of these areas can badly skew results no matter how large the dataset.

A lack of complete data or an unbalanced dataset, for example, can introduce AI bias, an anomaly where the AI model inherits prejudices in the training data. This is essentially what happened with early facial recognition applications, Shrier says. Models were trained with a large picture database of thirty-something white men. While the AI could accurately identify subjects fitting that profile, it did poorly at recognizing women and people with a skin color other than white. "They ended up not noticing that they had selected a dataset that looked like them and not like the broader population they were interacting with," he says.

But even with a set of high-quality data, bigger isn't necessarily better. McGee says he saw this in practice when training a natural language processing (NLP) model with four different amounts of training data—250 million words, 750 million words, 1 billion words, and 3 billion words—as an experiment. "Surprisingly, our results showed that increasing the amount of training data beyond 750 million words did not improve results," he says. "So, before you just blindly add more data, think about what kind of data you're adding and whether it has any relevance to your end goal before you start collecting it.

How much is too much?

So, how do you determine how much data your model needs and know when you've reached the point of diminishing returns? That depends largely on the complexity of the problem you're trying to solve and the algorithm you use.

NLP uses some of the largest models because of the complexity of making associations between words. The training dataset for OpenAI's GPT-3, one of the world's largest language models, was 45 terabytes in size. Computer vision models, on the other hand, typically use much smaller datasets because there aren't as many variations with images. Most chairs, for example, have four legs and a back, allowing you to more easily get a set of unique images and then start testing and validating. In other words, what constitutes "enough" data can vary wildly.

Please read: The road to machine learning success is paved with failure

"In R&D and life sciences, I haven't seen or heard of a situation where a use case had enough data," says Iveta Lohovska, principal data scientist at Hewlett Packard Enterprise. "They don't know how much data is enough for them because there's no scenario where they've had access to information to be able to play with this idea. But in IoT, the models go through terabytes of data, and the training is usually good enough if it's a level of magnitude lower when you deploy it in the field."

Fortunately, you don't necessarily have to guess how much training data your AI needs. There are statistical tests that allow researchers to know how well their AI is performing compared with chance, notes Mike Finley, chief scientist and CTO at AnswerRocket, a company that automates data analysis using AI. "Tools like confusion matrices even allow a researcher to visualize the quality of the AI's results," he says. "If AI is not passing these basic tests, the researcher knows there's more work to do. That work could be more data or a different algorithm. Many tools help guide that process."

Ultimately, creating a diverse training dataset that captures all the intricacies of your data is the most important factor in creating effective AI, Ries says, and he cautions against adding more information just because you have it. Rather, if after going through testing and validation you find areas where the model is not working well, you can increase the training dataset in those more focused areas.

"Right now, people are still trying to understand what the limits are for training datasets," Reis says. "An important concept to keep in mind is that you need to create a truth dataset for most of these algorithms—and the larger the dataset, the more time it will take to create it and validate it. What you *really* care about is generating a comprehensive dataset that has all of the various qualities that you expect to encounter in the real world."

Michael Ansaldo
Freelance writer | 2 publications

Michael Ansaldo is a veteran technology and business writer and editor. His work has appeared in numerous publications including PCWorld, Computerworld, TechHive, Wired, GreenBiz MacILife, and Executive Travel. He has also created and managed editorial and social media programs for leading global brands such as HP, Samsung, Verizon, and Madison Square Garden. Today he partners with businesses to create a range of thought leadership content.

RELATED

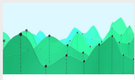
Michael Ansaldo
Freelance writer
2 publications

MORE BY MICHAEL ANSALDO

Feeding the world, one terabyte at a time

Topics

AI



Embrace AI analytics at scale

Manage the data flood

The state of