Jun 16, 2021, 05:04pm EDT | 37,088 views

Andrew Ng Launches A Campaign For Data-Centric AI

Gil Press Senior Contributor
Enterprise & Cloud
I write about technology, entrepreneurs and innovation.

Follow

Listen to article 7 minutes

Data is eating the world so Andrew Ng wants to make sure we radically improve its quality. "Data is food for AI," says Ng, and he is launching a campaign to shift the focus of AI practitioners from model/algorithm development to the quality of the data they use to train the models.

Landing AI, the startup Ng founded to bring AI to traditional industries, today announced a competition to get the best performance out of a fixed model by improving the quality of the data. The top three winners will be invited to a private roundtable event with Andrew Ng to share ideas and explore how to grow the data-centric movement. In addition, DeepLearning.AI, an education startup Ng also founded, is launching an online course to teach his data-centric approach to a worldwide audience on Coursera (which Ng co-founded in 2012).

Andrew Ng, Founder and CEO of LandingAI and deep learning.ai at the Amazon Re:MARS conference on ... [+] AFP VIA GETTY IMAGES

In the dominant model-centric approach to AI, according to Ng, you collect all the data you can collect and develop a model good enough to deal with the noise in the data. The established process calls for holding the data fixed and iteratively improving the model until the desired results are achieved. In the nascent data-centric approach to AI, "consistency of data is paramount," says Ng. To get to the right results, you hold the model or code fixed and iteratively improve the quality of the data.

Ng observes that 80% of the AI developer's time is spent on data preparation. This has been a widely shared estimate since the rise of "big data" in the late 2000s and the concomitant rise of "data scientists," known for their prowess in "data wrangling." As big data has fed the subsequent flourishing of the new AI or deep learning, the conventional wisdom has been that you just need to add more data to overcome any errors and subpar performance resulting from the low-quality or "noisy" data that the deep learning model has been initially trained on.

Recently, however, more attention has been paid to the role of low-quality data is playing in what Ng has identified as the proof-of-concept to production gap, or the inability of AI projects and machine learning models to succeed when they are deployed in the real world. "Paradoxically, data is the most under-valued and de-glamorised aspect of AI" say Google researchers in a recent paper, reporting on their survey of 53 AI practitioners. They found that "data cascades—compounding events causing negative, downstream effects from data issues—triggered by conventional AI/ML practices that undervalue data quality... are pervasive (92% prevalence), invisible, delayed, but often avoidable."

MORE FROM FORBES ADVISOR

Best Travel Insurance Companies

By **Amy Danise** Editor

Best Covid-19 Travel Insurance Plans

By **Amy Danise** Editor

"The model and the code for many applications are basically a solved problem," says Ng. "Now that the models have advanced to a certain point, we got to make the data work as well." He sees a number of recent developments supporting his call for data-centric AI. As investments in AI projects spread from Internet-based, consumer-facing companies to other industries, the models are typically trained by 10,000 or less examples rather than millions of examples. That is a very good reason to pay greater attention to the quality of the data.

Moreover, a hospital or a factory do not typically have on staff anyone with a machine learning expertise. What their employees have is data expertise, thorough knowledge of the domain in which they operate, even the specific organization where they work (there can be substantial differences in how hospitals collect, store, and manage their data, for example). Knowing how to assemble high-quality data in their domain is their professional competitive advantage.

“A data-centric approach,” says Ng, “allows people in manufacturing, hospitals, farms, to customize the data, making it more feasible for someone without technical training in AI to feed it into an open-source model.” That will help uncover many new opportunities for AI to make an impact in traditional environments with small data sets and no AI expertise. “What I see across the world is lots of these 1 to 5 million dollars projects that aren’t been worked on,” says Ng.

Consistency of labeling the data and a systematic way of going about cleaning it and correcting errors is what Ng would like to see happening across all AI projects. “Many data scientists have their own ways to clean data but what we don’t have is a systematic mental framework for doing it,” he says. And being systematic could help not just with small data sets but also with machine learning models using big data (as with web search or self-driving) where you typically find a long tail of rare events, constituting a small data set that requires error correction.

Ng pins his hopes for a new mental framework on the nascent field of machine learning operations or MLOps. This will be a collection of tools supporting the lifecycle of data-centric AI: train the model, conduct error analysis to identify the type of data the algorithm does poorly on, either

get more of that data via data augmentation or additional data collection or give more consistent definition for the data labels if they were found to be ambiguous. Then comes the real-world test in production, where new data is used for continuous refinement of the model (a just-released survey found almost 1 in 3 organizations do not routinely monitor and retrain machine learning models for peak performance).

Ng believes that the new MLOps tools will make data-centric AI an efficient and systematic process. They will also drive new job opportunities and new careers. “Just like the rise of deep learning a decade ago spawned tons of new jobs,” says Ng, “I hope that data-centric AI development will spawn tons of new jobs in many industries.”

The campaign for “good data”—data that is defined consistently, covers the important cases, has timely feedback from production data, and is sized appropriately—will benefit from Ng’s involvement given his stature in AI and beyond (in addition to his entrepreneurial activities, he has been a popular teacher at Stanford and online, started Google Brain and made a splash with an early deep learning project, and has served as Chief Scientist at Baidu). I asked Ng if he still believes that AI is “the new electricity” as he proclaimed in 2017. “Yes,” he answered, “we still have work to do, there will be a lot of twists and turns on the way, but we will get there.”

Follow me on Twitter or LinkedIn. Check out my website.

Gil Press

Follow

I'm Managing Partner at gPress, a marketing, publishing, research and education consultancy. Previously, I held senior marketing and research... **Read More**