The Global Chip Shortage, Explained › When will supply rebound?
P. 9
Engineering the Axial-Flux Motor › An electric motor for an electrified world
P. 38
The Mag Amp's Forgotten History › The U.S. rejected it, Germany perfected it
P. 46
FOR THE TECHNOLOGY INSIDER
APRIL 2022
IEEE Spectrum
Andrew Ng:
In AI, Small Is the New Big
IEEE

https://spectrum.ieee.org/magazine/2022/april/

ANDREW NG HAS SERIOUS STREET CRED in artificial intelligence. He pioneered the use of graphics processing units (GPUs) to train deep learning models in the late 2000s with his students at Stanford University, cofounded Google Brain in 2011, and then served for three years as chief scientist for Baidu, where he helped build the Chinese tech giant's AI group. So when he says he has identified the next big shift in artificial intelligence, people listen. And that's what he told *IEEE Spectrum* in an exclusive Q&A.

Ng's current efforts are focused on his company Landing AI, which built a platform called LandingLens to help manufacturers improve visual inspection with computer vision. He has also become something of an evangelist for what he calls the data-centric AI movement, which he says can yield "small data" solutions to big issues in AI, including model efficiency, accuracy, and bias.

The great advances in deep learning over the past decade or so have been powered by ever-bigger models crunching ever-bigger amounts of data. Some people argue that that's an unsustainable trajectory. Do you agree that it can't go on that way?

Andrew Ng: This is a big question. We've seen foundation models in NLP [natural language processing]. I'm excited about NLP models getting even bigger, and also about the potential of building foundation models in computer vision. I think there's lots of signal to still be exploited in video: We have not been able to build foundation models yet for video because of computing bandwidth and the cost of processing video, as opposed to tokenized text. So, I think that this engine of scaling up deep learning algorithms, which has been running for something like 15 years now, still has steam in it. Having said that, it only applies to certain problems, and there's a set of other problems that need small data solutions.

When you say you want a foundation model for computer vision, what do you mean by that?

Ng: This is a term coined by Percy Liang and some of my friends at Stanford to refer to very large models, trained on very large data sets, that can be tuned for specific applications. For example, GPT-3 is an example of a foundation model [for NLP]. Foundation models offer a lot of promise as a new paradigm in developing machine learning applications, but also challenges in terms of making sure that they're reasonably fair and free from bias, especially if many of us will be building on top of them.

What needs to happen for someone to build a foundation model for video?

Ng: I think there is a scalability problem. The compute power needed to process the large volume of images for video is significant, and I think that's why foundation models have arisen first in NLP. Many researchers are working on this, and I think we're seeing early signs of such models being developed in computer vision. But I'm confident that if a semiconductor maker gave us 10 times more processor power, we could easily find 10 times more video to build such models for vision.

Having said that, a lot of what's happened over the past decade is that deep learning has happened in consumer-facing companies that have large user bases, sometimes billions of users, and therefore very large data sets. While that paradigm of machine learning has driven a lot of economic value in consumer software, I find that that recipe of scale doesn't work for other industries.

It's funny to hear you say that, because your early work was at a consumer-facing company with millions of users.

Ng: Over a decade ago, when I proposed starting the Google Brain project to use Google's compute infrastructure to build very large neural networks, it was a controversial step. One very senior person pulled me aside and warned me that starting Google Brain would be bad for my career. I think he felt that the action couldn't just be in scaling up, and that I should instead focus on architecture innovation.

> *"In many industries where giant data sets simply don't exist, I think the focus has to shift from big data to good data. Having 50 thoughtfully engineered examples can be sufficient to explain to the neural network what you want it to learn." —Andrew Ng, CEO & Founder, Landing AI*

I remember when my students and I published the first NeurIPS workshop paper advocating using CUDA, a platform for processing on GPUs, for deep learning—a different senior person in AI sat me down and said, “CUDA is really complicated to program. As a programming paradigm, this seems like too much work.” I did manage to convince him; the other person I did not convince.

I expect they’re both convinced now.

Ng: I think so, yes.

Over the past year as I’ve been speaking to people about the data-centric AI movement, I’ve been getting flashbacks to when I was speaking to people about deep learning and scalability 10 or 15 years ago. In the past year, I’ve been getting the same mix of “there’s nothing new here” and “this seems like the wrong direction.”

How do you define data-centric AI, and why do you consider it a movement?

Ng: Data-centric AI is the discipline of systematically engineering the data needed to successfully build an AI system. For an AI system, you have to implement some algorithm, say a neural network, in code and then train it on your data set. The dominant paradigm over the last decade was to download the data set while you focus on improving the code. Thanks to that paradigm, over the last decade deep learning networks have improved significantly, to the point where for a lot of applications the code—the neural network architecture—is basically a solved problem. So, for many practical applications, it’s now more productive to hold the neural network architecture fixed, and instead find ways to improve the data.

When I started speaking about this, there were many practitioners who, completely appropriately, raised their hands and said, “Yes, we’ve been doing this for 20 years.” This is the time to take the things that some individuals have been doing intuitively and make them a systematic engineering discipline.

The data-centric AI movement is much bigger than one company or group of researchers. My collaborators and I organized a data-centric AI workshop at

NeurIPS, and I was really delighted at the number of authors and presenters that showed up.

You often talk about companies or institutions that have only a small amount of data to work with. How can data-centric AI help them?

Ng: You hear a lot about vision systems built with millions of images—I once built a face recognition system using 350 million images. Architectures built for hundreds of millions of images don't work with only 50 images. But it turns out, if you have 50 really good examples, you can build something valuable, like a defect-inspection system. In many industries where giant data sets simply don't exist, I think the focus has to shift from big data to good data. Having 50 thoughtfully engineered examples can be sufficient to explain to the neural network what you want it to learn.

When you talk about training a model with just 50 images, does that really mean you're taking an existing model that was trained on a very large data set and fine-tuning it? Or do you mean a brand new model that's designed to learn only from that small data set?

Ng: Let me describe what Landing AI does. When doing visual inspection for manufacturers, we often use our own flavor of RetinaNet. It is a pretrained model. Having said that, the pretraining is a small piece of the puzzle. What's a bigger piece of the puzzle is providing tools that enable the manufacturer to pick the right set of images [to use for fine-tuning] and label them in a consistent way. There's a very practical problem we've seen spanning vision, NLP, and speech, where even human annotators don't agree on the appropriate label. For big data applications, the common response has been: If the data is noisy, let's just get a lot of data and the algorithm will average over it. But if you can develop tools that flag where the data's inconsistent and give you a very targeted way to improve the consistency of the data, that turns out to be a more efficient way to get a high-performing system.

> *"Collecting more data often helps, but if you try to collect more data for everything, that can be a very expensive activity."* —Andrew Ng

For example, if you have 10,000 images where 30 images are of one class, and those 30 images are labeled inconsistently, one of the things we do is build tools to draw your attention to the subset of data that’s inconsistent. So, you can very quickly relabel those images to be more consistent, and this leads to improvement in performance.

Could this focus on high-quality data help with bias in data sets? If you’re able to curate the data more before training?

Ng: Very much so. Many researchers have pointed out that biased data is one factor among many leading to biased systems. There have been many thoughtful efforts to engineer the data. At the NeurIPS workshop, Olga Russakovsky gave a really nice talk on this. At the main NeurIPS conference, I also really enjoyed Mary Gray’s presentation, which touched on how data-centric AI is one piece of the solution, but not the entire solution. New tools like Datasheets for Datasets also seem like an important piece of the puzzle.

One of the powerful tools that data-centric AI gives us is the ability to engineer a subset of the data. Imagine training a machine-learning system and finding that its performance is okay for most of the data set, but its performance is biased for just a subset of the data. If you try to change the whole neural network architecture to improve the performance on just that subset, it’s quite difficult. But if you can engineer a subset of the data you can address the problem in a much more targeted way.

When you talk about engineering the data, what do you mean exactly?

Ng: In AI, data cleaning is important, but the way the data has been cleaned has often been in very manual ways. In computer vision, someone may visualize images through a Jupyter notebook and maybe spot the problem, and maybe fix it. But I’m excited about tools that allow you to have a very large data set, tools that draw your attention quickly and efficiently to the subset of data where, say, the labels are noisy. Or to quickly bring your attention to the one class among 100 classes where it would benefit you to collect more data. Collecting more data often helps, but if you try to collect more data for everything, that can be a very expensive activity.

For example, I once figured out that a speech-recognition system was performing poorly when there was car noise in the background. Knowing that allowed me to collect more data with car noise in the background, rather than trying to collect more data for everything, which would have been expensive and slow.

What about using synthetic data, is that often a good solution?

Ng: I think synthetic data is an important tool in the tool chest of data-centric AI. At the NeurIPS workshop, Anima Anandkumar gave a great talk that touched on synthetic data. I think there are important uses of synthetic data that go beyond just being a preprocessing step for increasing the data set for a learning algorithm. I'd love to see more tools to let developers use synthetic data generation as part of the closed loop of iterative machine learning development.

Do you mean that synthetic data would allow you to try the model on more data sets?

Ng: Not really. Here's an example. Let's say you're trying to detect defects in a smartphone casing. There are many different types of defects on smartphones. It could be a scratch, a dent, pit marks, discoloration of the material, other types of blemishes. If you train the model and then find through error analysis that it's doing well overall but it's performing poorly on pit marks, then synthetic data generation allows you to address the problem in a more targeted way. You could generate more data just for the pit-mark category.

> *"In the consumer software Internet, we could train a handful of machine-learning models to serve a billion users. In manufacturing, you might have 10,000 manufacturers building 10,000 custom AI models." —Andrew Ng*

Synthetic data generation is a very powerful tool, but there are many simpler tools that I will often try first. Such as data augmentation, improving labeling consistency, or just asking a factory to collect more data.

https://spectrum.ieee.org/magazine/2022/april/To make these issues more concrete, can you walk me through an example? When a company approaches Landing AI and says it has a problem with visual inspection, how do you onboard them and work toward deployment?

Ng: When a customer approaches us we usually have a conversation about their inspection problem and look at a few images to verify that the problem is feasible with computer vision. Assuming it is, we ask them to upload the data to the LandingLens platform. We often advise them on the methodology of data-centric AI and help them label the data.

One of the foci of Landing AI is to empower manufacturing companies to do the machine learning work themselves. A lot of our work is making sure the software is fast and easy to use. Through the iterative process of machine learning development, we advise customers on things like how to train models on the platform, when and how to improve the labeling of data so the performance of the model improves. Our training and software supports them all the way through deploying the trained model to an edge device in the factory.

How do you deal with changing needs? If products change or lighting conditions change in the factory, can the model keep up?

Ng: It varies by manufacturer. There is data drift in many contexts. But there are some manufacturers that have been running the same manufacturing line for 20 years now with few changes, so they don’t expect changes in the next five years. Those stable environments make things easier. For other manufacturers, we provide tools to flag when there’s a significant data-drift issue. I find it really important to empower manufacturing customers to correct data, retrain, and update the model. Because if something changes and it’s 3 a.m. in the United States, I want them to be able to adapt their learning algorithm right away to maintain operations.

In the consumer software Internet, we could train a handful of machine-learning models to serve a billion users. In manufacturing, you might have 10,000 manufacturers building 10,000 custom AI models. The challenge is, how do you do that without Landing AI having to hire 10,000 machine learning specialists?

So you're saying that to make it scale, you have to empower customers to do a lot of the training and other work.

Ng: Yes, exactly! This is an industry-wide problem in AI, not just in manufacturing. Look at health care. Every hospital has its own slightly different format for electronic health records. How can every hospital train its own custom AI model? Expecting every hospital's IT personnel to invent new neural-network architectures is unrealistic. The only way out of this dilemma is to build tools that empower the customers to build their own models by giving them tools to engineer the data and express their domain knowledge. That's what Landing AI is executing in computer vision, and the field of AI needs other teams to execute this in other domains.

Is there anything else you think it's important for people to understand about the work you're doing or the data-centric AI movement?

Ng: In the last decade, the biggest shift in AI was a shift to deep learning. I think it's quite possible that in this decade the biggest shift will be to data-centric AI. With the maturity of today's neural network architectures, I think for a lot of the practical applications the bottleneck will be whether we can efficiently get the data we need to develop systems that work well. The data-centric AI movement has tremendous energy and momentum across the whole community. I hope more researchers and developers will jump in and work on it.