IT'S NOT ABOUT THE ALGORITHM

The data that transformed AI research—and possibly the world

AP PHOTO/JEFF CHIU

Stanford professor and Google Cloud chief scientist Fei-Fei Li changed everything.

FROM OUR OBSESSION
Machines with Brains
Humanity's relationship with computers is dramatically changing, but the societal and economic impact remains unclear.

Published July 26, 2017This article is more than 2 years old.

In 2006, Fei-Fei Li started ruminating on an idea.

Li, a newly-minted computer science professor at University of Illinois Urbana-Champaign, saw her colleagues across academia and the AI industry hammering away at the same concept: a better algorithm would make better decisions, regardless of the data.

But she realized a limitation to this approach—the best algorithm wouldn't work well if the data it learned from didn't reflect the real world.

Her solution: build a better dataset.

"We decided we wanted to do something that was completely historically unprecedented," Li said, referring to a small team who would initially work with her. "We're going to map out the entire world of objects."

The resulting dataset was called ImageNet. Originally published in 2009 as a research poster stuck in the corner of a Miami Beach conference center, the dataset quickly evolved into an annual competition to see which algorithms could identify objects in the dataset's images with the lowest error rate. Many see it as the catalyst for the AI boom the world is experiencing today.

Alumni of the ImageNet challenge can be found in every corner of the tech world. The contest's first winners in 2010 went on to take senior roles at Baidu, Google, and Huawei. Matthew Zeiler built Clarifai based off his 2013 ImageNet win, and is now backed by $40 million in VC funding. In 2014, Google split the winning title with two researchers from Oxford, who were quickly snapped up and added to its recently-acquired DeepMind lab.

Li herself is now chief scientist at Google Cloud, a professor at Stanford, and director of the university's AI lab.

Today, she'll take the stage at CVPR to talk about ImageNet's annual results for the last time—2017 was the final year of the competition. In just seven years, the winning accuracy in classifying objects in the dataset rose from 71.8% to 97.3%, surpassing human abilities and effectively proving that bigger data leads to better decisions.

Even as the competition ends, its legacy is already taking shape. Since 2009, dozens of new AI research datasets have been introduced in subfields like computer vision, natural language processing, and voice recognition.

"The paradigm shift of the ImageNet thinking is that while a lot of people are paying attention to models, let's pay attention to data," Li said. "Data will redefine how we think about models."

What's ImageNet?

In the late 1980s, Princeton psychologist George Miller started a project called WordNet, with the aim of building a hierarchal structure for the English language. It would be sort of like a dictionary, but words would be shown in relation to other words rather than alphabetical order. For example, within WordNet, the word "dog" would be nested under "canine," which would be nested under "mammal," and so on. It was a way to organize language that relied on machine-readable logic, and amassed more than 155,000 indexed words.

The ImageNet hierarchy derived from WordNet.

Li, in her first teaching job at UIUC, had been grappling with one of the core tensions in machine learning: overfitting and generalization. When an algorithm can only work with data that's close to what it's seen before, the model is considered overfitting to the data; it can't understand anything more general past those examples. On the other hand, if a model doesn't pick up the right patterns between the data, it's overgeneralizing.

Finding the perfect algorithm seemed distant, Li says. She saw that previous datasets didn't capture how variable the world could be—even just identifying pictures of cats is infinitely complex. But by giving the algorithms more examples of how complex the world could be, it made mathematic sense they could fare better. If you only saw five pictures of cats, you'd only have five camera angles, lighting conditions, and maybe variety of cat. But if you've seen 500 pictures of cats, there are many more examples to draw commonalities from.

Li started to read about how others had attempted to catalogue a fair representation of the world with data. During that search, she found WordNet.

Having read about WordNet's approach, Li met with professor Christiane Fellbaum, a researcher influential in the continued work on WordNet, during a 2006 visit to Princeton. Fellbaum had the idea that WordNet could have an image associated with each of the words, more as a reference rather than a computer vision dataset. Coming from that meeting, Li imagined something grander—a large-scale dataset with many examples of each word.

Months later Li joined the Princeton faculty, her alma mater, and started on the ImageNet project in early 2007. She started to build a team to help with the challenge, first recruiting a fellow professor, Kai Li, who then convinced Ph.D student Jia Deng to transfer into Li's lab. Deng has helped run the ImageNet project through 2017.

"It was clear to me that this was something that was very different from what other people were doing, were focused on at the time," Deng said. "I had a clear idea that this would change how the game was played in vision research, but I didn't know how it would change."

The objects in the dataset would range from concrete objects, like pandas or churches, to abstract ideas like love.

Li's first idea was to hire undergraduate students for $10 an hour to manually find images and add them to the dataset. But back-of-the-napkin math quickly made Li realize that at the undergrads' rate of collecting images it would take 90 years to complete.

After the undergrad task force was disbanded, Li and the team went back to the drawing board. What if computer-vision algorithms could pick the photos from the internet, and humans would then just curate the images? But after a few months of tinkering with algorithms, the team came to the conclusion that this technique wasn't sustainable either—future algorithms would be constricted to only judging what algorithms were capable of recognizing at the time the dataset was compiled.

Undergrads were time-consuming, algorithms were flawed, and the team didn't have money—Li said the project failed to win any of the federal grants she applied for, receiving comments on proposals that it was shameful Princeton would research this topic, and that the only strength of proposal was that Li was a woman.

A solution finally surfaced in a chance hallway conversation with a graduate student who asked Li whether she had heard of Amazon Mechanical Turk, a service where hordes of humans sitting at computers around the world would complete small online tasks for pennies.

“He showed me the website, and I can tell you literally that day I knew the ImageNet project was going to happen,” she said. “Suddenly we found a tool that could scale, that we could not possibly dream of by hiring Princeton undergrads.”

The Amazon Mechanical Turk backend for classifying images.

Mechanical Turk brought its own slew of hurdles, with much of the work fielded by two of Li’s Ph.D students, Jia Deng and Olga Russakovsky . For example, how many Turkers needed to look at each image? Maybe two people could determine that a cat was a cat, but an image of a miniature husky might require 10 rounds of validation. What if some Turkers tried to game or cheat the system? Li’s team ended up creating a batch of statistical models for Turker’s behaviors to help ensure the dataset only included correct images.

Even after finding Mechanical Turk, the dataset took two and a half years to complete. It consisted of 3.2 million labelled images, separated into 5,247 categories, sorted into 12 subtrees like “mammal,” “vehicle,” and “furniture.”

In 2009, Li and her team published the ImageNet paper with the dataset—to little fanfare. Li recalls that CVPR, a leading conference in computer vision research, only allowed a poster, instead of an oral presentation, and the team handed out ImageNet-branded pens to drum up interest. People were skeptical of the basic idea that more data would help them develop better algorithms.

“There were comments like ‘If you can’t even do one object well, why would you do thousands, or tens of thousands of objects?” Deng said.

If data is the new oil, it was still dinosaur bones in 2009.

The ImageNet Challenge

Later in 2009, at a computer vision conference in Kyoto, a researcher named Alex Berg approached Li to suggest that adding an additional aspect to the contest where algorithms would also have to locate where the pictured object was, not just that it existed. Li countered: Come work with me.

Li, Berg, and Deng authored five papers together based on the dataset, exploring how algorithms would interpret such vast amounts of data. The first paper would become a benchmark for how an algorithm would react to thousands of classes of images, the predecessor to the ImageNet competition.

“We realized to democratize this idea we needed to reach out further,” Li said, speaking on the first paper.

Li then approached a well-known image recognition competition in Europe called PASCAL VOC, which agreed to collaborate and co-brand their competition with ImageNet. The PASCAL challenge was a well-respected competition and dataset, but representative of the previous method of thinking. The competition only had 20 classes, compared to ImageNet’s 1,000.

As the competition continued in 2011 and into 2012, it soon became a benchmark for how well image classification algorithms fared against the most complex visual dataset assembled at the time.

A screenshot of the ImageNet database online

But researchers also began to notice something more going on than just a competition—their algorithms worked better when they trained using the ImageNet dataset.

“The nice surprise was that people who trained their models on ImageNet could use them to jumpstart models for other recognition tasks. You’d start with the ImageNet model and then you’d fine-tune it for another task,” said Berg. “That was a breakthrough both for neural nets and just for recognition in general.”

Two years after the first ImageNet competition, in 2012, something even bigger happened. Indeed, if the artificial intelligence boom we see today could be attributed to a single event, it would be the announcement of the 2012 ImageNet challenge results.

Geoffrey Hinton, Ilya Sutskever, and Alex Krizhevsky from the University of Toronto submitted a deep convolutional neural network architecture called AlexNet—still used in research to this day—which beat the field by a whopping 10.8 percentage point margin, which was 41% better than the next best.

ImageNet couldn’t come at a better time for Hinton and his two students. Hinton had been working on artificial neural networks since the 1980s, and while some like Yann LeCun had been

able to work the technology into ATM check readers through the influence of Bell Labs, Hinton's research hadn't found that kind of home. A few years earlier, research from graphics-card manufacturer Nvidia had made these networks process faster, but still not better than other techniques.

Hinton and his team had demonstrated that their networks could perform smaller tasks on smaller datasets, like handwriting detection, but they needed much more data to be useful in the real world.

"It was so clear that if you do a really good on ImageNet, you could solve image recognition," said Sutskever.

Today, these convolutional neural networks are everywhere—Facebook, where LeCun is director of AI research, uses them to tag your photos; self-driving cars are using them to detect objects; basically anything that knows what's in a image or video uses them. They can tell what's in an image by finding patterns between pixels on ascending levels of abstraction, using thousands to millions of tiny computations on each level. New images are put through the process to match their patterns to learned patterns. Hinton had been pushing his colleagues to take them seriously for decades, but now he had proof that they could beat other state of the art techniques.

"What's more amazing is that people were able to keep improving it with deep learning," Sutskever said, referring to the method that layers neural networks to allow more complex patterns to be processed, now the most popular favor of artificial intelligence. "Deep learning is just the right stuff."

The 2012 ImageNet results sent computer vision researchers scrambling to replicate the process. Matthew Zeiler, an NYU Ph.D student who had studied under Hinton, found out about the ImageNet results and, through the University of Toronto connection, got early access to the paper and code. He started working with Rob Fergus, a NYU professor who had also built a career working on neural networks. The two started to develop their submission for the 2013 challenge, and Zeiler eventually left a Google internship weeks early to focus on the submission.

Zeiler and Fergus won that year, and by 2014 all the high-scoring competitors would be deep neural networks, Li said.

"This Imagenet 2012 event was definitely what triggered the big explosion of AI today," Zeiler wrote in an email to Quartz. "There were definitely some very promising results in speech recognition shortly before this (again many of them sparked by Toronto), but they didn't take off publicly as much as that ImageNet win did in 2012 and the following years."

Today, many consider ImageNet solved—the error rate is incredibly low at around 2%. But that's for classification, or identifying which object is in an image. This doesn't mean an algorithm

knows the properties of that object, where it comes from, what it's used for, who made it, or how it interacts with its surroundings. In short, it doesn't actually understand what it's seeing. This is mirrored in speech recognition, and even in much of natural language processing. While our AI today is fantastic at knowing what things are, understanding these objects in the context of the world is next. How AI researchers will get there is still unclear.

After ImageNet

While the competition is ending, the ImageNet dataset—updated over the years and now more than 13 million images strong—will live on.

Berg says the team tried to retire the one aspect of the challenge in 2014, but faced pushback from companies including Google and Facebook who liked the centralized benchmark. The industry could point to one number and say, "We're *this* good."

Since 2010 there have been a number of other high-profile datasets introduced by Google, Microsoft, and the Canadian Institute for Advanced Research, as deep learning has proven to require data as vast as what ImageNet provided.

Datasets have become haute. Startup founders and venture capitalists will write Medium posts shouting out the latest datasets, and how their algorithms fared on ImageNet. Internet companies such as Google, Facebook, and Amazon have started creating their own internal datasets, based on the millions of images, voice clips, and text snippets entered and shared on their platforms every day. Even startups are beginning to assemble their own datasets—TwentyBN, an AI company focused on video understanding, used Amazon Mechanical Turk to collect videos of Turkers performing simple hand gestures and actions on video. The company has released two datasets free for academic use, each with more than 100,000 videos.

"There is a lot of mushrooming and blossoming of all kinds of datasets, from videos to speech to games to everything," Li said.

It's sometimes taken for granted that these datasets, which are intensive to collect, assemble, and vet, are free. Being open and free to use is an original tenet of ImageNet that will outlive the challenge and likely even the dataset.

In 2016, Google released the Open Images database, containing 9 million images in 6,000 categories. Google recently updated the dataset to include labels for where specific objects were located in each image, a staple of the ImageNet challenge after 2014. London-based DeepMind, bought by Google and spun into its own Alphabet company, recently released its own video dataset of humans performing a variety of actions.

“One thing ImageNet changed in the field of AI is suddenly people realized the thankless work of making a dataset was at the core of AI research,” Li said. “People really recognize the importance the dataset is front and center in the research as much as algorithms.”