THE DEEP LEARNING CURVE

Big data is important not only to human researchers, but also to the artificial intelligence that computers acquire through deep learning. Emeritus Professor of Computer Science Francis YL Chin has jumped at the chance to be part of the big data revolution.

Constant learning has been a necessity throughout Emeritus Professor Francis YL Chin's professional life, driven by never-ending evolution in computing technology. When he retired from HKU in 2015, one might have thought it would be time to take a breather. But the exciting developments of the past few years have been too irresistible a draw. Not only is technology constantly improving, but machines are showing the potential to progressively learn. At the heart of this revolution is big data.

"Big data and deep learning are very closely related. Big data is why deep learning has been doing so well," he said.

Big data is simply a huge volume of data. Consider that 20 years ago computers had at most one gigabyte (GB) of storage. Today a single USB flash drive can contain up to one terabyte (about 1,000 GB) or more. The increase in data and advances in computing technology have opened new possibilities.

"Before we had big data, we used rule-based learning for computers. We told the computer what to look for. It's like what humans do when we learn another language: we study the grammar and other rules," he said.

"But we don't learn our native language this way. Instead, we're exposed to the language so much that we pick up the rules implicitly without much thought to what those rules are. This is what we're doing with machine learning."

For example, to get a search engine to recognise images of cats, a programmer would previously have had to input characteristics like four legs, tail, pointy ears and whiskers. It is now possible to learn without being given these characteristics – the system scans thousands of photos, some of which might be labelled as cat photos, comes to learn the defining characteristics of cats, and groups them together. It keeps refining its search results so even photos in which some features are hidden will still pop up in searches. "Right now there are so many photos and articles on the internet that you don't want a human to label them all because that would be a lot of work. Unsupervised learning, where the data is not labelled, is a product of big data," Professor Chin said.

Professor Chin giving a talk on 'How Deep Learning Improves Our Health?' in the Hong Kong Science Museum during the HK SciFest 2017.

Jump on it

Machine learning through big data is expensive, though, particularly because much of it is not in the public domain. Internet firms such as Google, Amazon, Facebook, and Alibaba collect enormous amounts of data about pages visited, searches and so forth. Their in-house researchers, who have access to this data, are also producing some of the most significant research in the world.

This said, Professor Chin pointed to the recent program AlphaGo Zero, developed by a Google subsidiary, which taught itself how to play the game Go without any data apart from the rules of the game. A previous version had used data from more than 100,000 games to learn the game. The success of AlphaGo Zero makes sense considering that a child does not need to see millions of cat photos to understand the concept of a cat.

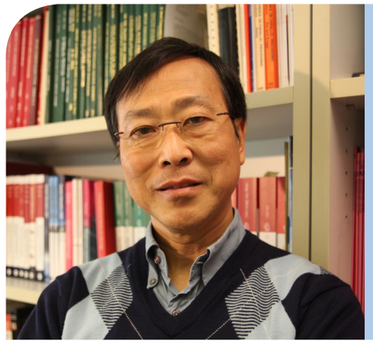
But most machine learning still depends on big data and this is where Professor Chin hopes to make a mark.

In one project, he is working with a Baptist University scholar to identify significant features from images of about 20,000 antique bronze mirrors, as well as matching mirrors with similarities, using computer technology.

In another, at Hang Seng Management College, he is looking at numerous examples of Hong Kong students' English writing to highlight similar mistakes, make suggestions for correction and identify good writing. "The mistakes Hong Kong people make in learning English will be different from the mistakes made by students in other places, like India. We're using linguistics and natural language processing to learn what mistakes are made and why," he said.

A third major project, also at Hang Seng Management College jointly with Alpha Financial Press, involves machine translation of business documents, especially for initial public offerings (IPOs), a very targeted but lucrative market. IPO documents must be filed with regulators in both English and Chinese, and are usually first written by lawyers in English. The translation turnaround time is very tight so it is an expensive task. It also cannot be done through Google or other online services, even if they were proficient enough, because of confidentiality issues. So Professor Chin and his team are training computers to do the translation by feeding them reams of IPO and other business documents.

"We think we can do better than Google," he said. "The technology is moving very fast. We need to jump on the bandwagon." ■



> "Big data and deep learning are very closely related. Big data is why deep learning has been doing so well."
>
> **Professor Francis YL Chin**

Source:	**under the arrangements currently in force , the aggregate emoluments payable by our group to and benefits in kind receivable by our directors for the year ending 31 december 2013 are expected to be approximately rmb3, 20 4,000 .**
Our translation:	**根據現時生效的安排，截至二零一三年十二月三十一日止年度，本集團應付予董事的酬金總額及董事應收的實物利益預期約為人民幣3,204,000元。**
Human translation:	**根據現行安排，截至2013年12月31日止年度，本集團應付董事的薪酬總額及上述董事應收的實物利益預期約為人民幣3,204,000元。**
Google translate:	**根據現行有效的安排，我們小組的應付和實物我們於截至2013年12月31日年度的董事應收款合計薪酬福利預計約為人民幣30，20 4000。**

Professor Chin and his team are training computers to do the translation of financial documents by feeding them reams of initial public offering and other business documents.