Computer vision and the AI boom

By Xu Chunjing, Huawei 2012 Labs

I see, therefore I am

A human's ability to conceptualize and think abstractly and logically about the world – what we consider intelligence – depends on the ability to receive external stimuli. Imagine a newborn that cannot use its sense of sight, hearing, smell, or touch. Even with a physiologically healthy brain, it's unlikely to develop any kind of intelligence if this state persists. Because what we see accounts for much of our stimulus, visual perception is an extremely important aspect of AI and intelligent systems.

Computer vision enables machines to understand the content of images, much like how image signals are formed by photoreceptors in our retinas. This includes objects in the image, the relationships between the objects, and the meaning of the image as a whole.

As recently as five years ago, most people – researchers included – believed that computer vision was a tangent, less complex precursor to AI rather being intrinsic to it, with AI being chiefly seen as a way to enable machines to master learning and reasoning.

However, research and recent progress in computer vision using deep learning doesn’t just relate to visual perception, because many high-level semantic capabilities relating to intelligence are closely linked to vision. Solving the problem of scene recognition could therefore lead to great advances in AI in the relatively near future.

Challenges with computer vision

Typically, computers observe the world through cameras, but there’s a huge gulf between seeing and perceiving. As shown in figure 1, the image of the letter “A” is read by a computer as a string of values, with each pixel represented as a value.

Computer vision recognizes that these strings of values represent the letter “A”, which appears relatively simple until we look at Figure 2.

While a person can recognize each letter without difficulty, a computer sees a sequence of completely different values for each image, meaning it’s much harder for it to see the letter “A”. This is even more complex in photographs of natural settings where letters may appear in road signs or adverts, and where many factors like light or unrelated background objects come into play. These increase the difficulty for a computer to recognize them.

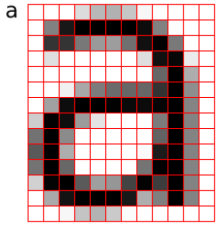
When light reflects off objects in the real world and is captured by a camera’s image sensors, responses are produced as strings of values. While objects are defined by fixed concepts, the responses produced by them vary constantly. The main task of computer vision is to summarize and represent the fixed elements – a number of unchanging semantic concepts – that are contained in complex and variable pixel-formed images.

Substantial progress has been made in inferring semantic concepts from variable pixels using feature representation and supervised learning. Current intelligent systems can locate and recognize text in photographs captured by digital cameras and mobile phones, such as house numbers, restaurant names, and signs, even when the font, angle, or lighting varies.

Deep learning and semantics

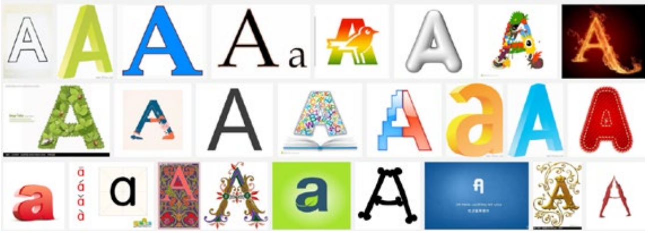
Feature representation of images by inputting a series of values at the pixel level is the most important aspect of computer vision. However, image representation by pixel is the lowest level of feature representation. Slight variations in the image result in huge changes to the corresponding value strings, even though the concepts (objects) remain

Figure 1



1.0	1.0	1.0	0.9	0.6	0.6	0.6	1.0	1.0	1.0	1.0	1.0
1.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.5	1.0	1.0	1.0
1.0	0.2	0.2	0.5	0.6	0.6	0.5	0.0	0.0	0.5	1.0	1.0
1.0	0.9	1.0	1.0	1.0	1.0	1.0	0.9	0.0	0.0	0.9	1.0
1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.5	0.0	0.5	1.0
1.0	1.0	1.0	0.5	0.5	0.5	0.5	0.5	0.4	0.0	0.5	1.0
1.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	1.0
0.9	0.0	0.0	0.6	1.0	1.0	1.0	1.0	0.5	0.0	0.5	1.0
0.5	0.0	0.6	1.0	1.0	1.0	1.0	1.0	0.5	0.0	0.5	1.0
0.5	0.0	0.7	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.5	1.0
0.6	0.0	0.6	1.0	1.0	1.0	1.0	0.5	0.0	0.0	0.5	1.0
0.9	0.1	0.0	0.6	0.7	0.7	0.5	0.0	0.5	0.0	0.5	1.0
1.0	0.7	0.1	0.0	0.0	0.0	0.1	0.9	0.8	0.0	0.5	1.0
1.0	1.0	1.0	0.8	0.8	0.9	1.0	1.0	1.0	1.0	1.0	1.0

Figure 2



unchanged. At higher levels of representation, variations in images when the concepts are also unaltered don't significantly change the value strings. The highest level of representation is the semantic concept.

Before deep learning became popular, research on computer vision focused on how to artificially design a form of representation by combining experience and mathematics. For instance, pixel color does not strongly correlate to the concept of the letter "A", but pixel-formed borders and the particular shape they create are. Therefore, a type of value representation designed to describe pixel-formed borders and shapes overcomes the problem computer vision has in dealing with image variations such as changes in background color or light and – to a certain extent – changes in viewing angle.

I can't see what you see

This type of feature representation design works on a case by case basis, so a system designed for character recognition couldn't be applied to animal recognition. In object recognition in humans, light signals captured by the retina are processed hierarchically by different parts of the brain before high-level concepts are formed. The processing channels in the human neural network are constant. We don't need many different mechanisms to recognize different concepts. Computer vision's ultimate goal is to devise a single feature representation method that can be applied to a wide range of situations, akin to the way feature representation works in the human neural network.

Scientists began to use mathematics to describe the way neurons work in the 1940s, giving rise to the field of artificial neural networks. The 1960s and '70s saw much development in this field, but it quickly stagnated. The neural networks designed at the time were too shallow, with only two or three layers, and their ability to represent features was limited; however, when deeper networks with more layers were designed, training these networks became extremely difficult.

Researchers designing neural networks later focused on neural networks formed by many small convolution filters. In the 1980s, LeCunn and others designed a series of deeper neural networks with seven to eight layers that could be trained under the conditions of the time, and convolution neural networks were particularly adept at recognizing ZIP codes.

When convoluted is best

Convolution neural networks have been in low-level use for many years. Following a series of improvements, the technique is now used in large-scale natural image classification as part of ImageNet (image-net.org), a large public image database that outperforms almost all traditional methods of image recognition and testing. Deep learning is now an integral method of mainstream computer vision research.

Driven by companies such as Google, Microsoft, Facebook, and Baidu, deep learning's role in problem solving is increasing. It's now a standard tool in, for example, image retrieval, video surveillance, and autopilot visual perception.

An important aspect of training large-scale neural networks is data. Because large companies can collect and use large amounts of data, when directed this data can be used to improve the performance of neural networks. In terms of the amount of data, training artificially designed feature representation requires data containing many thousands of samples, while large-scale neural networks require millions. In autopilot applications, for instance, millions of hours of video samples are needed to train neural networks.

This presents a new challenge – sourcing data for such applications. In visual perception for autopilot functionality, vehicles, people, and other targets must be accurately detected with cameras, requiring many images of different scenarios. In the currently popular technique of supervised learning, labelled images are used to train neural networks with given structures, which require manually labeling cars and other targets in images. Mainstream neural networks can contain hundreds of millions of parameters, and it costs an enormous amount in labor and time to label the massive amounts of data.

But, the success of applications often depends on obtaining effective samples. There are currently two ways to source data. The first is straightforward but investment-heavy: build large image databases to train basic data representation. For new applications, training feature representation using these databases can be used as a foundation application in other areas because feature representation in deep learning neural networks is versatile.

ImageNet uses this method, which is used in many academic fields to enable research that couldn't otherwise take place and create deep network structures and new training methods.

Downsizing samples

Unlike human perception and image recognition that can form abstract concepts from a very small number of samples, deep learning methods are vastly different. This has led scientists to explore a second method: small sample size learning. This method explores enabling neural networks to learn concepts in a compositional way by teaching machines basic modular concepts like "wheel" and "frame", which can then be combined into higher concepts such as "bicycle".

It also includes causality where compositional concepts maintain a logical cause and effect structure (for example, wheels cannot be placed above frames) and a self-learning capability. This method is still at a nascent stage but meaningful progress will advance visual intelligence as we know it.

While it's important not to overhype AI – something that's arguably happening now – computer vision will slowly impact every aspect of our lives.

Some future applications of computer vision >>

ADAS: Advanced Driver Assistance System

Will work with LiDAR, image processing, and in-car networking for safer and better driving

Detecting events

Recognition events; e.g., for security

People counting

Surveillance

Human-AI interaction

Input method for interacting with AI

Enable mobile robots to navigate

Manufacturing

Controlling processes in industry

Automated inspections

Modeling objects and environments

Topographical modeling

Medical image analysis