Artificial Intelligence: from Pixels and Phonemes to Semantic Understanding and Interactions

Achintya K. Bhowmik

Starkey Hearing Technologies, Eden Prairie, MN, USA
Keywords: Artificial Intelligence, Machine Learning, Deep Neural Networks.

ABSTRACT

In the recent years, unprecedented advances in artificial intelligence (AI) technologies and applications are being enabled by rapid developments in machine learning, big data, and specialized computing architectures. We will review how devices are increasingly being endowed with technologies to sense and understand the world, often surpassing human-level performances, and ushering in a new wave of intelligent applications.

1 INTRODUCTION

The dream of developing intelligent computers and machines that can think like humans, navigate autonomously in the world, and interact naturally with us and each other, goes back to the early days of human imaginations. Numerous science fiction stories, books and movies have envisioned such systems, whose impact on the human civilization ranges from spectacularly positive to destructively dire.

Are we on an irreversible course of disruptive transformation, which will redefine the industries of entertainment, transportation, retail, manufacturing, healthcare, finance, and more? Despite the "AI Winters" of the past marked by the realities falling far short of the hypes, is it different this time? If so, why? What impact will these advances make on the imaging and information display ecosystem? What should the strategies be for large incumbent players as well as the researchers, aspiring entrepreneurs, and startup companies?

In this article, we will briefly review the historical context, technology basics, recent developments, and future trends, with emphasis on the breakthrough applications in vision, imaging and display systems that are being enabled by the rapid progress and broad adoption of artificial intelligence.

2 BRIEF HISTORICAL CONTEXT

In order to grasp the recent pace of accomplishments and the profound impact of artificial intelligence, we must reflect on the historical developments of the field. While the early discourses on intelligent and autonomous machinery trace back to the Greek civilization, in this article we will limit our review to the relatively modern era for the sake of brevity.

Alan Turing, the celebrated genius British polymath, surmised that it should be possible to develop intelligent systems that mimic the human abilities of deduction and problem solving based on input data and reasoning process. In 1950, he proposed the famous “Turing Test”, designed to evaluate the competency of such systems, specifically whether they surpass the threshold of natural human interactions [1].

The term “artificial intelligence” was broadly adopted in a seminal conference hosted by John McCarthy at the Dartmouth College in 1956, attended by MIT scientists Marvin Minsky, Claude Shannon, and others [2]. The audacious thinking and positive outlook on future developments envisioned at that remarkable summer workshop are epitomized in this statement: *“Every aspect of learning or any other feature of intelligence can be so precisely described that a machine can be made to simulate it.”*

What proceeded in the subsequent decades are a mix of advanced technology developments leading to some compelling demonstrations, followed by a general disenchantment and widespread disillusion due to the applications falling short of the much-hyped eventuality of systems capable of general intelligence, leading to a phase of “AI Winter” marked by a reduction in research projects, funding, etc. [3].

However, the recent years witnessed a stunning reversal in both the development and the impact of AI [4 - 6]. These can be attributed to the rapid advances in three technology areas: i) computing, ii) algorithms, and iii) data.

On the computing front, the astonishing pace of progress is reflected in the relentless realization of the Moore’s Law, which observes that the number of transistors on a semiconductor chip doubles about every two years [7]. As an example, the Intel 4004 processor in 1971 had a total of 2,250 transistors, whereas the Apple A13 Bionic processor in 2019 sports a staggering 8.5 billion transistors! This exponential increase in computing horsepower has recently been further boosted by the development of computing architectures that are significantly more efficient for executing highly parallel processing tasks demanded by the modern AI algorithms. In fact, a recent paper from OpenAI stated that the computing power used in the training of large AI models has been doubling every 3.5 months [8], a much faster exponential pace than the progress in general-purpose computing described by Moore’s Law.

On the algorithm front, the developments of specialized AI models such as the deep neural networks

(DNN), convolutional neural networks (CNN), recurrent neural networks (RNN), and related variants, along with the application of backpropagation techniques to train these networks with massive amounts of data led to numerous breakthrough demonstrations [4 - 6].

Lastly, the massive push to digitize information in the recent years led to an unprecedented deluge of data, ushering in the era of “big data” [9].

3 TECHNOLOGY OVERVIEW

The scope of artificial intelligence technology is vast. In this article, we will narrow our discussion to allow convergence of the key elements towards applications in the domains of vision, imaging and display technologies.

Broadly speaking, the field of AI deals with intelligence exhibited by machines, systems which take actions that maximize their chances of success at some goals. Within this broad field of AI, a specific set of applications are based on machine learning (ML), which encompasses the ability of computers to learn after being trained with relevant data, but without being explicitly programmed. One of the most exciting developments among numerous machine learning techniques is the emergence of deep neural networks and some key variants that have enabled the recent deployment of AI in an array of practical applications. As the deep neural networks are based on machine learning, these techniques are also often referred to as deep learning (DL). In the following, we discuss the technical foundation of the deep neural network architecture and applications.

As we discussed, the work on artificial intelligence is largely inspired by the drive to mimic the human cognitive abilities, so we first take a quick look at biology and neuroscience. As depicted in Figure 1, the functions of the human perceptual system include sensing and understanding the world, taking actions based on such information, and continuously learning from these processes [10]. The human sensory systems comprise vision, hearing, touch, smell, taste, balance, etc., which are specialized in the transduction of associated physical stimuli from the external world into neural signals. The perceptual systems perform computations within various areas of the human cerebral cortex to process and understand these signals in order to create a model of the world and facilitate relevant actions. The building block of the cerebral cortex is the neuron. A typical neuron, as depicted in Figure 2, receives signals from neighboring neurons via its dendrites in the form of input neural impulses, computes a weighted sum of these signals where the weights represent the relative significance of the inputs from various neurons, and modifies the result with a native transfer function. If the result exceeds a threshold, the neuron may fire an impulse of its own that travels down the axon and acts as the inputs into the dendrites of other neighboring neurons. The human brain consists of a network of about 100 billon interconnected neurons.

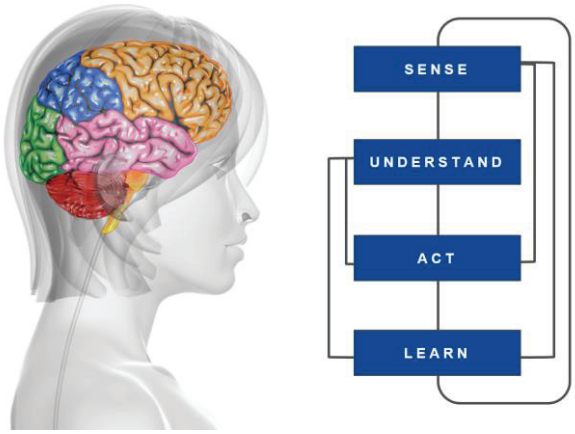

Fig. 1 The human perceptual system enables sensing and understanding the world, as well as actions and continuous learning.

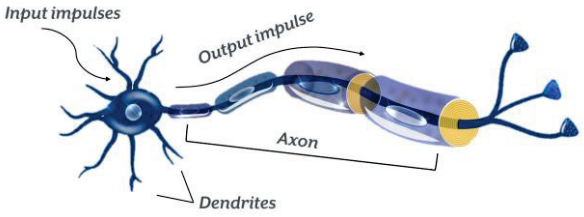

Fig. 2 The biological neuron is the building block of the human cerebral cortex.

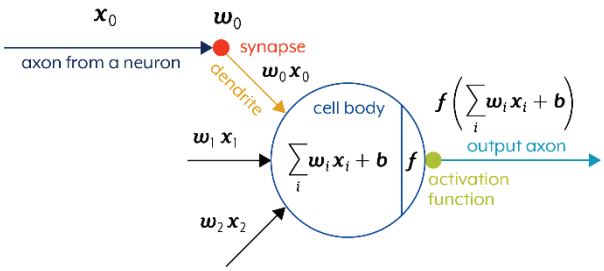

Fig. 3 A mathematical model of the neuron that serves as the building block of an artificial neural network.

Drawing parallel to this biological system towards creating a mathematical model, the computational building block of a deep neural network is an artificial neuron, as shown in Figure 3, with input, processing, and output units that symbolically mirror the natural counterpart. A deep neural network, as illustrated in Figure 4, consists of a large number of these artificial neurons which are interconnected to define the depth and width of the network.

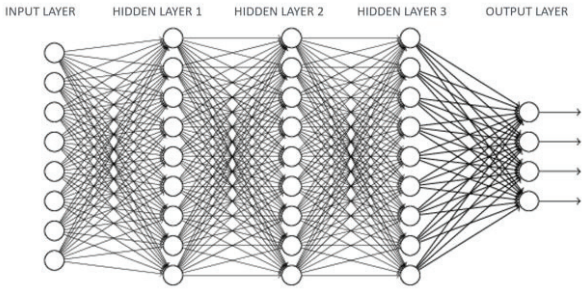

Fig. 4 A deep neural network consisting of many interconnected neurons arranged in layers [6].

Prior to training with data, the parameters of a neural network, such as the weights and biases of the individual neurons (denoted as *w* and *b* in Figure 3), are set with random values. The training process starts with feeding input datasets into the first layer of neurons of the network, and forward-propagating the data through the network by performing computations at the neurons in various layers along the way that involves calculating a weighted sum of the input values at each neuron and passing the result through a transfer function, termed as the activation function (denoted as *f* in Figure 3). At the end of the network, after completing such a forward pass, the output is compared with the known target result, and an error is computed using a cost function. This initiates a "backpropagation" step, which essentially is a gradient-based optimization process that adjusts the network parameters such as the weights and biases at every neuronal node, with a goal to minimize the errors for the next forward-propagation pass. The error is computed again, triggering another backward pass to optimize the parameters further, followed by another forward pass, etc. This process is repeated until the computed error is smaller than a predetermined value, hence the accuracy of the neural network in predicting the results for a given set of input values is within an acceptable range. With this achievement, the AI model is now "trained", ready to be tested with more datasets, and then deployed for real-world applications.

4 APPLICATIONS IN VISION, IMAGING AND DISPLAY SYSTEMS

Arguably, the recent upsurge in the enthusiasm for deep learning was partly initiated by the successful demonstrations of these techniques in solving computer vision and image recognition problems.

Automatically recognizing objects from images and understanding the context of the visual scenes have been the goal of computer vision researchers for many decades. However, it has historically been a very difficult problem to decipher semantic information from images composed of two-dimensional array of pixel values, exacerbated by the large variations due to lighting conditions, poses, occlusions, etc. In fact, back in 2010, the best computer vision algorithms could only achieve about 75% accuracy, far short of what is needed for practical applications. In a popular book on computer vision published in 2010, the author lamented: *"…the dream of having a computer interpret an image at the same level as a two-year old remains elusive"* [11].

With this backdrop, breakthrough developments in the field of computer vision and image recognition came with the introduction of the modern deep learning algorithms, large datasets, and superior computers. Especially, the introduction of the Stanford ImageNet with a large database of 14 million hand-annotated images in more than 20,000 categories kicked off a flurry of developments worldwide, spurred by the associated Stanford ImageNet Large Scale Visual Recognition Challenge (ILSVRC) based on a significant subset of the database [12]. Researchers developed specialized deep convolutional neural networks to achieve ever-improving visual recognition accuracies, surpassing the capabilities of human vision for narrow tasks in 2015, as illustrated in Figure 5.

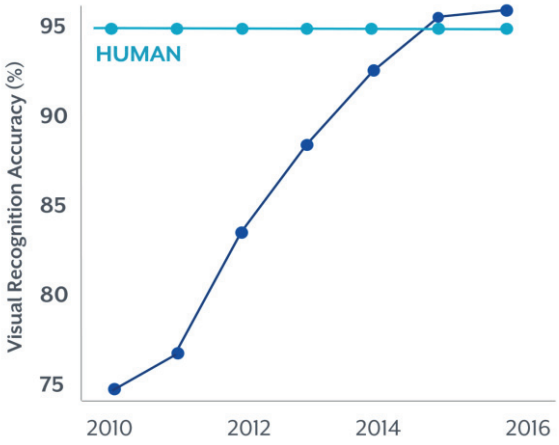

Fig. 5 Image recognition accuracy with computer vision based on deep neural networks surpassed human vision in 2015, as demonstrated in the Stanford ImageNet Large Scale Visual Recognition Challenge.

The astonishing success of convolutional neural networks in image recognition tasks can be partly attributed to the high-level similarities of the architecture to the human visual perception system. As the seminal work by Hubel and Wiesel on the mammalian visual cortex demonstrated [13], the biological vision system has a hierarchical architecture, based on first extracting low-level features from the corresponding visual fields, then proceeding to decipher increasingly higher-level

information towards the understanding of the semantic information. A typical convolutional neural network consists of an input layer that takes in pixel values from an image, several layers of convolution filters that extract various levels of features from the sub-images, followed by a fully-connected neural network yielding the object classes at the output layer. An example of a seminal work on these is the LeNet, which enabled the first commercial deployment of a convolutional neural network character recognizer [14].

Going beyond the tasks of object recognition in visual images, a very promising new development is the real-time captioning of visual scenes based on higher-level semantic understanding of the content and actions. An example of this work was recently reported by Johnson et al. [15].

Deep learning techniques with real-time classification and inferencing, often combined with depth-sensing cameras, are enabling many compelling real-world applications, including self-driving cars, autonomous robots and drones, immersive virtual and augmented reality devices, etc. [16, 17].

In addition to the successes in visual image recognition tasks, applications of modern deep learning techniques in automatic speech recognition have also been yielding spectacular results (Figure 6), enabling widespread adoption of voice-based human interfaces and interactions with devices.

5 CONCLUSIONS

In this article, we have reviewed the developments in artificial intelligence, focusing on the applications of machine learning in imaging and visual recognition tasks. Looking forward, continued advancements in deep learning algorithms, specialized computing architectures with more power-efficient computation for both machine learning training and classification tasks, as well as burgeoning domain-specific databases that are accessible to researchers in the academia and the industry, are expected to increasingly enable breakthrough applications. These new wave of innovations offer an exciting opportunity for the systems based on visual imaging and display technologies, transforming their abilities from merely capturing and displaying the pixels and phonemes towards achieving full semantic understanding of the content and enabling natural user interactions.

REFERENCES

[1] The Alan Turing Internet Scrapbook, “The Turing Test, 1950”, turing.org.uk.

[2] J. McCarthy, M. Minsky, N. Rochester, C. Shannon, “A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence” (1955).

[3] J. Howe, “Artificial Intelligence at Edinburgh University: a Perspective” (1994).

[4] S. Russell and P. Norvig, “Artificial Intelligence: A Modern Approach,” Pearson (2009).

[5] I. Goodfellow, Y. Bengio, and A. Courville, “Deep Learning,” The MIT Press (2016).

[6] M. Nielsen, “Neural Networks and Deep Learning,” Determination Press (2015).

[7] G. Moore, “Cramming more components onto integrated circuits,” Electronics 38 (1965).

[8] OpenAI Blog, “AI and Compute,” https://openai.com/blog/ai-and-compute/ (2018).

[9] M. Hilbert and P. López, “The World's Technological Capacity to Store, Communicate, and Compute Information,” Science. 332, 60–65 (2011).

[10] A. K. Bhowmik, “Senses, Perception, and Natural Human Interfaces for Interactive Displays,” in Interactive Displays, Wiley (2014).

[11] R. Szeliski, “Computer Vision: Algorithms and Applications” Springer (2010).

[12] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, F. Li, “ImageNet Large Scale Visual Recognition Challenge,” eprint arXiv:1409.0575 (2014).

[13] D. Hubel and T. Wiesel, “Receptive fields and functional architecture of monkey striate cortex,” Journal of Physiology, 195, 215–243 (1968).

[14] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, “Gradient-based learning applied to document recognition,” Proc. IEEE, 86, 11, 2278 (1998).

[15] J. Johnson, A. Karpathy, and F. Li, “DenseCap: Fully Convolutional Localization Networks for Dense Captioning,” IEEE Conference on Computer Vision and Pattern Recognition (2016).

[16] K. Vodrahalli and A. K. Bhowmik, “3D Computer Vision based on Machine Learning with Deep Neural Networks: A Review,” J. Soc. Inf. Display, 25, 676 (2018).

[17] A. K. Bhowmik, “Interactive and Immersive Devices with Perceptual Computing Technologies,” Molecular Crystals and Liquid Crystals 647, 329 (2017).

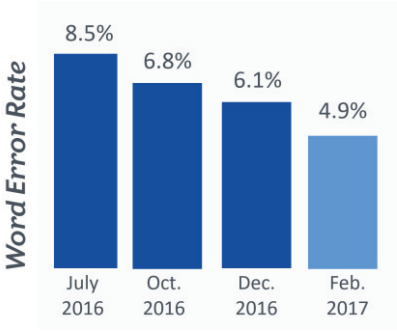

Fig. 6 Word error rate for automatic speech recognition using recurrent neural networks is steadily decreasing (source: Google).