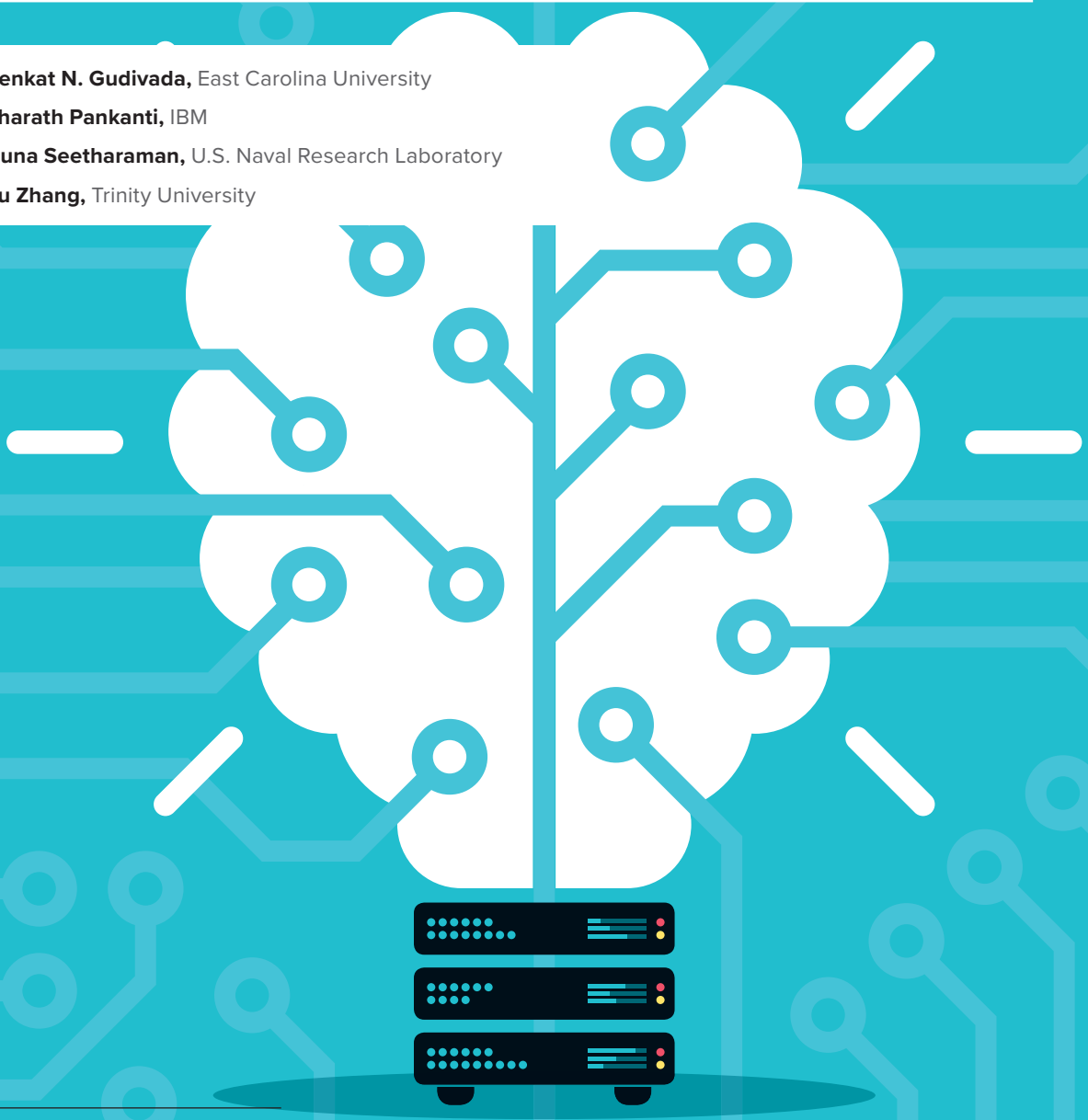
Please cite this article as: V.N. Gudivada, Sharath Pankanti, Guna Seetharaman, and Yu Zhang. Cognitive Computing Systems: Their Potential and the Future, IEEE Computer, vol. 52, issue. 5, May 2019, pp. 13-18, DOI: 10.1109/MC.2019.2904940

Cognitive Computing Systems: Their Potential and the Future

Venkat N. Gudivada, East Carolina University
Sharath Pankanti, IBM
Guna Seetharaman, U.S. Naval Research Laboratory
Yu Zhang, Trinity University



Digital Object Identifier 10.1109/MC.2019.2904940
Date of publication: 14 May 2019

The recent phenomenal advances in the foundational areas of cognitive computing systems are poised to usher in even more sophisticated systems that will rival and perhaps even surpass human performance.

Humans are arguably the most intelligent entities in the known universe; the objective of cognitive computing is to understand and replicate the essence of human intelligence. Autonomous systems are self-contained and self-regulated entities that continuously evolve in real time in response to changes to their environment. Fundamental to this evolution is learning and development. Cognition is the basis for autonomous systems. *Human cognition* refers to processes and systems that enable humans to perform both mundane and specialized tasks. *Machine cognition* refers to similar processes and systems that enable computers to perform tasks at a level that rivals human performance. While human cognition employs biological and natural means—the brain and mind—for its realization, machine cognition is a type of computation. Cognitive computing systems are autonomous systems that are based on machine cognition. A cognitive system views the mind as a highly parallel information processor, uses various models for representing information, and employs algorithms for transforming and reasoning with the information.

In contrast to conventional software systems, cognitive computing systems effectively deal with ambiguity and conflicting and missing data. They fuse several sources of multimodal data and incorporate context into computation. When making a decision or answering a query, these systems quantify uncertainty, generate multiple hypotheses and supporting evidence, and score hypotheses based on evidence. In other words, cognitive computing systems provide multiple ranked decisions and answers. These systems can elucidate the reasoning that underlies their decisions and answers. They are stateful and understand various nuances of communication with humans. They are self-aware and continuously learn, adapt, and evolve. Their capabilities vary widely.

Cognitive computing systems use a broad range of principles and techniques from cognitive science, neuroscience, data science, machine learning, and cloud computing. Cognitive science is the study of mind and offers theories, including mathematical and computational models of human cognition. Neuroscience is the study of the nervous system, including its development, structure, and function. Data science is a new interdisciplinary domain, offering processes and systems to extract information and knowledge from structured and unstructured data using machine-learning algorithms. Its end goal is to discover patterns, generate actionable insights, and answer predictive questions. Machine learning and cloud computing provide the computational infrastructure and algorithms to develop cognitive computing systems. Machine learning, especially deep learning, provides learning algorithms that are inspired by the structure and function of the brain. Cloud computing provides turnkey solutions, such as the platform, infrastructure, and software as services. It achieves economies of scale and helps cognitive computing applications perform at scale without upfront computing investments.

Broadly speaking, there are two lines of research in the cognitive computing discipline. The first one is cognitive science driven. The second, more recent, is based on computer science, encompassing data science, statistics, and various subdisciplines of computer science. These two lines of research are complementary and helping to accelerate discoveries and innovation. Only a few years ago, it was believed that self-driving automobiles were a fantasy and science fiction. But today, self-driving vehicles are a reality. The recent phenomenal advances in the foundational areas of cognitive systems are poised to usher in even more sophisticated cognitive systems that will rival and perhaps even surpass human performance. However, in the near term, cognitive systems are not going to replace humans en masse; instead, they will help humans expand their capabilities by leveraging vast amounts of data and computing power.

A TIME LINE

Chatbots are one class of cognitive computing systems. They are conversational agents and question-answering systems; the human–chatbot interaction is often an informal conversation. ELIZA, which Joseph Weizenbaum created in 1966, is one of the first chatbots. Other chatbots include PARRY (1972), A.L.I.C.E (1995), Smarterchild (2001), Siri (2010), Google Now (2012), Alexa (2015), Cortana (2015), and Woebot (2017).

The expert systems approach to cognitive systems is based on reasoning using bodies of knowledge. The knowledge is primarily represented as if-then rules, which are manually developed and encoded. These systems cannot learn from their environment nor evolve. Edward Feigenbaum introduced two early expert systems around

1965: Mycin for diagnosing infectious diseases and Dendral for identifying unknown organic molecules.

The first generation of neural networks and genetic algorithms represents another set of approaches to cognitive systems. Though computational models for neural networks were proposed in the 1940s, it was not until 1975 that neural networks received widespread attention. Paul Werbos introduced the back-propagation algorithm in 1975, which made training multilayer neural networks feasible and efficient. Genetic algorithms are used for solving nonlinear or nondifferentiable optimization problems. They borrow concepts from evolutionary biology to search for a global minimum. The lack of adequate computational power and other factors led to the neural networks' hibernation.

The emergence of graphics processing unit (GPU) technology, crowdsourced labeled data, and advances in machine-learning algorithms created a renaissance in neural networks during the current decade under the deep-learning label. Since 2012, we have witnessed an unparalleled interest in using deep learning to solve a broad range of problems in vision and spoken- and written-language processing. Also, recent advanced learning algorithms and increased access to hardware are catalyzing the training of deeper (e.g., >100 layers) neural networks. In summary, the trifecta of growing availability of (e.g., crowdsourced) labeled data, availability of parallel computing hardware, and advances in learning algorithms has resulted in significant improvement in state of the art in diverse research fields, often surpassing human performance levels.

PLATFORMS, FRAMEWORKS, AND APPLICATIONS

The recent emergence of specialized processors for cognitive computing, big data tools, and advances in deep learning are accelerating cognitive computing systems research and ushering in novel and transformational applications. IBM's TrueNorth cognitive computing system is a case in point. Its design is inspired by the function and efficiency of the human brain. The TrueNorth architecture provides a spiking neuron model as a building block. Its programming paradigm is based on an abstraction called *corelet*, which represents a network of neurosynaptic cores. A library of reusable corelets and an integrated development environment enable the creation of cognitive computing applications.

The IBM Watson platform originated from a precise-query answering capability that was first demonstrated when it won the TV game show *Jeopardy* against world champions of the game. Its services are based on knowledge, language, speech, and computer vision application programming interfaces (APIs) primarily on the cloud/mobile devices. In addition, the Watson platform provides tools not only for integration and curation but also for testing fairness and integrity of the services. Finally, the platform focuses on scaling and openness across multi- and hybrid-cloud environments to permit the enterprise customers to unlock intelligence from the available data to improve their productivity and efficacy. Now Watson is deployed in various applications domains. For example, Watson Health can read and understand more than 200 million pages of text in fewer than 3 s. Its application domains include customer care, the Internet of Things, genomics, drug discovery, oncology, and patient engagement.

Recently, Nvidia released the Tesla P100 GPU, which specifically targets deep-learning algorithms. The P100 features 150 billion transistors on a single chip. In addition, Google released the Natural Language API, a cloud service that provides application developers access to pretrained algorithms for sentiment analysis, named-entity recognition, and syntax analysis. Likewise, Speech API, Translate API, and Vision API are public cloud services for speech-to-text conversion, translation between natural languages, and image analysis, respectively. Speech API enables converting audio to text for more than 80 languages, and Translate API provides machine translation between these languages. Applications can use Vision API to perform image analysis tasks, including object detection and classification.

Microsoft Cognitive Services is a set of APIs, software development kits, and services for creating intelligent, engaging, and discoverable applications. These services are available in five categories: vision, speech, language, knowledge, and search. Cisco Cognitive Threat Analytics is a cloud-based solution that helps to detect sophisticated threats in real time. It does not depend on manually predefined rule sets; it instead employs machine learning and statistical modeling to independently identify new threats, learn from the data, adapt, and evolve over time. DeepMind focuses on developing algorithms and approaches for programs that can learn to solve complex problems without needing to be explicitly taught about the problems beforehand. It provides open source environments, data sets, and code for developing cognitive computing systems.

Numenta is a small team of scientists and engineers specializing in reverse engineering the neocortex. They believe that understanding how the brain works is the key to understanding the principles of intelligence and that the same principles can be used to develop cognitive computing systems. Numenta's scientific findings, software, and intellectual property are free for noncommercial purposes and can be licensed for commercial use. Augmented Intelligence Software from CognitiveScale targets emulating and extending human cognitive function by pairing people and computers. The software is based on the premise that augmented intelligence is the most effective way to maximize the value of artificial intelligence and machine learning.

COGNITIVE COMPUTING SYSTEM APPLICATIONS ARE BOTH NUMEROUS AND VARIED.

PROMISES AND RESEARCH ISSUES

Cognitive computing system applications are both numerous and varied. They are used for diverse applications including fighting cybercrime, personalized learning, conversational chatbots and question answering, personal assistants, prescriptive analytics, anomaly detection, humanoid robots and assisted living, brain–computer interfaces, and language services for multilingual environments. We envision two major classes of cognitive computing systems. The first class includes those that can perform tasks independently, without human intervention. Examples include self-driving vehicles, personal assistants, and drones for delivering supplies. The second class includes those that augment human capabilities. Such systems will work collaboratively with humans to solve complex and ill-defined problems. A good example is personalized medicine, in which a physician and a cognitive system work together, and the cognitive system also plays a physician role.

The primary research issues in cognitive systems are the architectures for building them and the mechanisms used for learning, adaptation, and evolution. A cognitive architecture is a blueprint for developing cognitive systems. It specifies predetermined structures and interactions among them with the goal of achieving functions of the mind. They constrain the types of cognitive models that can be developed. The knowledge embodied in the architecture drives the interactions among the structures to achieve intelligent behavior.

A cognitive model, in contrast, focuses on a single cognitive process, such as language acquisition. It is also used to study the interaction between cognitive processes, such as language understanding and problem solving. Cognitive models help to reveal the limitations of cognitive architectures. Thus, there is a strong interplay between cognitive architectures and models. In the literature, the terms *cognitive architecture* and *cognitive model* are not used consistently and are often used synonymously. The context should help to reveal the intended meaning.

There are three major classes of cognitive architectures: cognitivist, connectionist, and hybrid. Cognitivist architectures represent information that uses explicit symbolic representations. Cognitivist architectures are also called *symbolic architectures/artificial intelligence* approaches. Cognitive systems based on this architecture can successfully solve specific problems. However, they lack the generality needed to be useful across domains. Symbolic representations reflect the designers' understanding of the domain and may bias the system. Also, as systems become more complex, it is difficult for the designers to ensure whether all relevant representations are adequate to realize the desired cognitive behaviors of the system.

Connectionist architectures are inspired by the information processing that occurs in biological neural systems. Information is processed by simple, networked computational units called *neurons*, which communicate in parallel with each other using electrochemical signals. A synapse is a junction between two neurons with a minute gap across which signals pass by way of neurotransmitter diffusion.

The brain is made up of neurons, which number from 10 to 100 billion. Each neuron is predicted to have more than 10,000 connections to other neurons. They receive stimuli from other neurons through incoming connections and perform nonlinear computations using the received stimuli. The effect of this computation activates other neurons through its outgoing connections. The strengths of connection activations are quantified on

a numeric scale and adjusted to reflect the state of network learning. Architectures based on this approach are called *connectionist* or *emergent architectures*.

Connectionist architectures use distributed representations for encoding knowledge. An advantage of these representations is that they are more resilient to noisy input, and performance degradation is generally more graceful, although interesting research is providing insights into how these models can be attacked adversarially with small perturbation in the inputs. Deep learning relies heavily on multiscale distributed representations. There is a strong coupling between the cognitive computing architectures and the knowledge representations used. Convolutional neural networks, recurrent neural networks (RNNs), long short-term memory units embedded in RNNs, and differentiable neural computers are types of connectionist architectures. Hybrid architectures employ a combination of symbolic and connectionist architectures.

Other issues include the need for large volumes of data and a high-performance computing infrastructure for model development, overfitting models to training data, difficulties in domain adaptation, the need to explain how cognitive systems deduce conclusions, and hyperparameter optimization. In addition, for mainstream adoption, it is important to ensure how the connectionist architectures can be trusted and be transparent.

ABOUT THIS SPECIAL ISSUE

The four articles in this special issue are representative of the cognitive computing systems domain. Cognitive computing applications typically require sophisticated processing of noisy and unstructured real-world data under stringent time constraints. Neuromorphic computing and neural-network accelerators are solutions to meet these processing challenges. The goal of neuromorphic computing is to use very large-scale integration (VLSI) systems that are driven by electronic analog circuits to simulate neurobiological architectures present in the nervous system. These VLSI chips are characterized by ultralow power consumption and high performance. They are referred to as *neuromorphic chips* or *brain chips*.

The first article, "TrueNorth: Accelerating From Zero to 64 Million Neurons in 10 Years" by DeBole et al. from IBM, presents TrueNorth architecture. TrueNorth is a brain-inspired massively multicore neural-network inference chip containing 1 million spiking neurons and 256 million low-precision synapses. This is a culmination of a decade of fundamental research spanning neuroscience, architecture, chips, systems, algorithms, and software. TrueNorth is the largest neurosynaptic computer ever built. The 64-chip system has 64 million neurons and 16 billion synapses and can be configured with deep neural networks trained to accurately detect, localize, and classify objects in high-definition video at faster-than-real-time frame rates and unprecedented energy efficiency.

The second article, "Leveraging Stochasticity for In Situ Learning in Binarized Deep Neural Networks" by Pyle et al., presents a binary approach for compact and energy-sparing neuromorphic architectures using emerging devices. Approaches that deal with device process variations and the realization of stochastic behavior intrinsically within neural circuits are explored. The authors leverage a novel probabilistic spintronic device for low-energy recognition operations, which improves deep neural-network performance through active on-chip learning via the mitigation of device reliability challenges.

"Perspectives on Becoming an Applied Machine Learning Scientist" by Rasiwasia, the third article, provides insights into the application of machine-learning

ABOUT THE AUTHORS

VENKAT N. GUDIVADA is a professor and computer science department chair at East Carolina University. His research interests are in data management, information retrieval, computational linguistics, and personalized learning. Gudivada received a Ph.D. from the University of Louisiana at Lafayette. He is a Senior Member of the IEEE. Contact him at gudivadav15@ecu.edu.

SHARATH PANKANTI is research staff member in the Artificial Intelligence (AI) Department at the IBM T.J. Watson Research Center. His research interests include AI, computer vision, blockchain, and trusted and transparent computing. Pankanti received a Ph.D. from Michigan State University. He is a Fellow of the IEEE, IAPR, and SPIE. Contact him at sharat@us.ibm.com.

GUNA SEETHARAMAN is a navy senior scientist for advanced computing concepts and chief scientist of computation in the Information Technology Division at the U.S. Naval Research Laboratory. His research interests include computer vision algorithms for airborne applications. Seetharaman received a Ph.D. from the University of Miami. He is a Fellow of the IEEE. Contact him at guna@ieee.org.

YU ZHANG is a professor and chair of the Computer Science Department at Trinity University. Her research interests include agent-based modeling and simulation. Zhang received a Ph.D. from Texas A&M University. She is a Member of the IEEE, ACM, and SCS. Contact her at yzhang@trinity.edu.

techniques to solve business problems. The author reflects on his experience of transitioning from a researcher to an applied machine-learning scientist by drawing on his end-to-end solutions experience at software companies, including Yahoo Research Labs, Fashiate, Snapdeal, and Amazon.

The special issue concludes with "Human Eye Movements Reveal Video Frame Importance" by Ma et al. This article explores whether eye movement patterns reflect frame importance during video viewing and facilitate video summarization. The authors recorded eye movements while subjects watched videos from the SumMe video summarization data set. They found more gaze consistency for selected frames than for unselected. A novel multistream deep-learning model for video summarization is introduced that incorporates subjects' eye movement information. Gaze data improved the model's performance over that observed when they used only the frames' physical attributes. Their results suggest that eye movement patterns reflect the cognitive processing of sequential information. This insight helps to select important video frames and provide an innovative algorithm that uses gaze information in video summarization.

Recent progress in cognitive computing has resulted in significant advances in the state of the art in diverse fields ranging from radio astronomy to life sciences. Cognitive computing is likely to dramatically disrupt the conventional workflows and practices and catapult us into a new era and culture that is a steep change from the present age.

The guest editors are grateful to the anonymous reviewers for their diligence and insightful comments.

ACKNOWLEDGMENTS

The first author's work is supported in part by the NSF IUSE/PFE:RED award No. 1730568.

Digital Object Identifier 10.1109/MC.2019.2911814