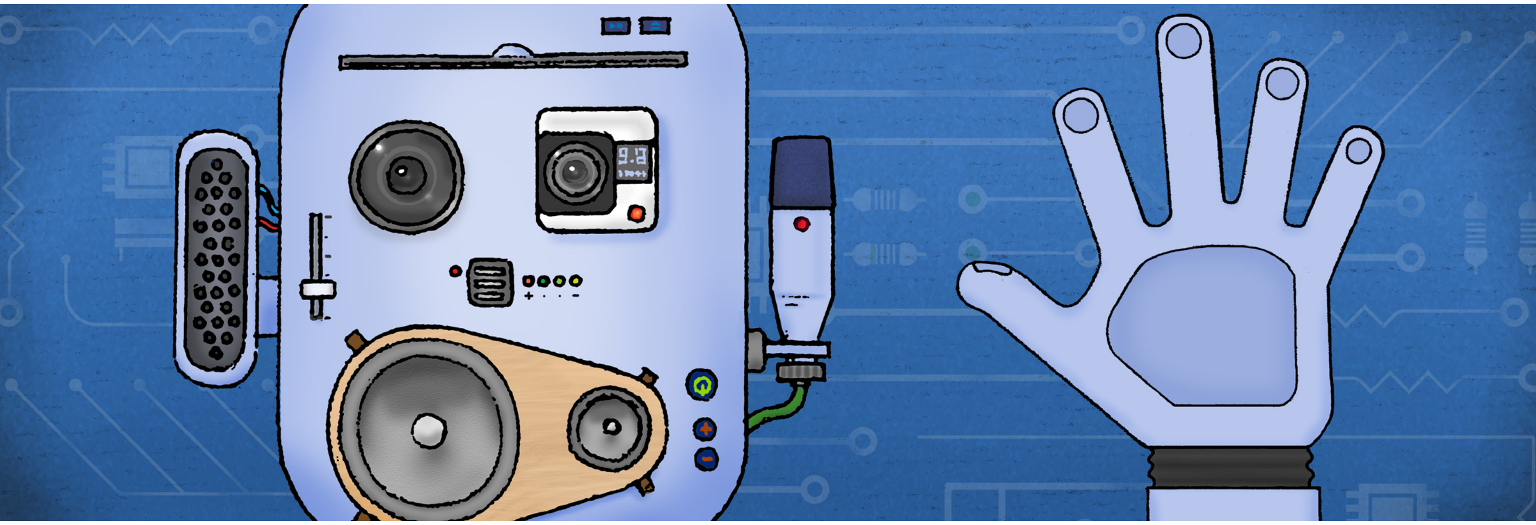
Deloitte University Press



Machine learning is going mobile

By David Schatsky

MACHINE learning—the process by which computers can get better at performing tasks through exposure to data, rather than through explicit programming—requires massive computational power, the kind usually found in clusters of energy-guzzling, cloud-based computer servers outfitted with specialized processors. But an emerging trend promises to bring the power of machine learning to mobile devices that may lack or have only intermittent online connectivity. This will give rise to machines that sense, perceive, learn from, and respond to their environment and their users, enabling the emergence of new product categories, reshaping how businesses engage with customers, and transforming how work gets done across industries.

Signals

- Google has introduced language translation software, using small neural networks optimized for mobile phones, which can perform well without an Internet connection.[1]
- Lenovo announced a mobile phone that uses multiple sensors, high-speed image processing hardware, and specialized Google software to support capabilities such as indoor wayfinding, precision measuring, and augmented reality even when offline.[2]
- NVIDIA, a maker of graphics processing technology, introduced an embeddable module for computer vision applications

in devices such as drones and autonomous vehicles that the company says consumes one-tenth the power of a competing offering.[3]

- Qualcomm introduced a new processor and software platform that support machine learning tasks such as image classification, speech recognition, and anomaly detection without a connection to a network.[4]

- Drone maker DJI recently introduced a consumer-oriented drone that uses advanced computer vision hardware to enable it to follow a moving object while automatically avoiding obstacles.[5]

Machine learning is the keystone cognitive technology

Emerging technologies rarely get as big a publicity boost as machine learning recently saw, when Google software defeated one of the world's top players of Go, one of the most complex board games ever created, in a best-of-five series of matches.[6] The international headlines confirmed that machine learning—the process by which fresh data can teach computers to better perform tasks—is one of the hottest domains within the field of artificial intelligence, and that this cognitive technology is progressing rapidly.[7]

Neural networks—computer models designed to mimic aspects of the human brain's structure and function, with elements representing neurons and their interconnections—are an increasingly popular way of implementing machine learning. They are particularly well suited for performing perceptual tasks such as computer vision and speech recognition. Familiar examples of applications that employ neural networks for such tasks include Google's voice search,[8] Facebook's system for tagging people in photos,[9] and Google Photos, which uses a neural network-based image recognition system to automatically classify photos by their contents.[10] All of these systems run in the cloud on powerful servers, processing data such as digitized voice or photos that users upload.

Until recently, a typical smartphone lacked the power to perform such tasks without connecting to the cloud, except in limited ways. For instance, some mobile phone software can recognize a single face—the owner's—in order to unlock the phone, or a small set of predetermined words such as "OK Google." But offline support for increasingly powerful perception tasks is coming to mobile devices.

Pushing machine learning onto mobile devices

Firms are starting to outfit smartphones, drones, and cars with chips based on new designs that can run neural networks efficiently while consuming 90 percent less power than previous generations.[11] Research efforts at MIT and IBM suggest that we will soon see more chips on the market that excel at running neural networks at high speed, in small spaces and at low power.[12] Because of this, mobile devices are becoming increasingly capable of performing sophisticated feats that take advantage of neural networks, such as computer vision and speech recognition, once reserved for powerful servers running in the cloud.

It is not only progress in hardware that is bringing machine learning to mobile devices. Tech vendors are also finding ways to create compact neural networks capable of running tasks such as speech recognition and language translation on conventional mobile phones with no connection to a server required. For instance, Google has introduced mobile language-translation software using small neural networks optimized for smartphones that can perform well even offline.[13] And Google researchers recently published a paper describing an Internet-independent speech recognition system that performs well on a commercial mobile phone.[14]

Mobile devices are acquiring the power to perform sophisticated perceptual tasks without dependence on connectivity to the cloud, bringing greater accuracy, reliability, and responsiveness while strengthening user privacy. This should greatly expand the number of applications of perceptual computing coming to market—and not only on mobile phones. Mobile machine learning and perceptual computing will power a wide range of devices, from mobile sensors to phones, tablets, drones, cars, and new types of devices as yet unimagined, creating significant opportunities for business.

Many industries will see new and improved applications

It's impossible to enumerate all of the applications we will see for mobile devices capable of performing sophisticated perceptual tasks involving vision, speech, or other sensory input. But they are likely to be found in every industry and have one or more of the following capabilities:

- Analysis or diagnosis of sensory data
- Perceptual interfaces or interactivity
- Navigation and motion control

A few examples follow.

Analysis or diagnosis

In **health care**, we envision a wide range of diagnostic applications, including some aimed at consumers. Imagine, for instance, a smartphone app that can diagnose skin conditions and insect bites by analyzing digital photos without transmitting the image data over a network.

We imagine mobile **architecture and design** applications that use computer vision to generate accurate 3D models of interior spaces quickly and easily.

An ever more powerful and resilient **Internet of Things** will include self-monitoring industrial equipment that uses machine learning to predict maintenance needs and self-diagnose failures.[15]

Perceptual interfaces and interactivity

In **media and entertainment**, we will likely see mobile devices—both general-purpose ones such as mobile phones and special-purpose ones such as augmented-reality headsets—offering ever more realistic and engaging augmented and virtual reality for games and filmed entertainment.

Ultralow power processors designed for machine learning will likely help **consumer** and **industrial** devices and machines understand and respond to the environment around them, and find their way into Internet-independent voice-controlled **wearable devices**, **household appliances**, and **industrial machinery**.

Navigation and motion control

Low-power chips with powerful computer vision support are bringing impressive capabilities to unmanned aerial vehicles, also known as drones, which already have applications in many industries, from **real estate** and **construction** to **agriculture**, **energy**, **aerospace**, and **defense**. Drone maker DJI recently introduced a consumer-oriented aerial vehicle able to follow a moving object while automatically avoiding obstacles.[16]

New, powerful mobile computer vision modules that use deep learning are helping advanced driver assistance systems to "address the challenges of everyday driving, such as unexpected road debris, erratic drivers and construction zones."[17]

Indoor navigation apps that use computer vision to precisely locate a user, track her motion, and guide her in interior spaces will find use in museums, train stations, airports, malls, and **retail** stores, opening up new advertising and commerce opportunities without the need to deploy beacons or other connectivity-based approaches.

And on the horizon are applications not yet imagined, from wearable to pocketable to portable, that can sense, analyze, and respond to sensory inputs including sound, video, and biometrics—all enabled by low-power chips designed to support neural networks for machine learning.

Implications

Compact, efficient, low-power, high-performance, mobile machine learning. New products. New human-computer interfaces. Powerful new ways of engaging with and serving customers. The trend described here has implications for companies and professionals across industries.

Makers of mobile devices and mobile apps should begin to familiarize themselves with the potential of a new generation of devices capable of offline machine learning.

User-experience designers should begin to explore the kinds of experiences made possible by these technologies. Tech vendors are releasing development kits, such as Google's Project Tango development kit and NVIDIA's Jetson TX1 Developer Kit, to encourage designers and developers to do so.

Product strategists and engineers working on consumer and industrial products should consider the value that mobile perceptual computing could bring to products ranging from household appliances to personal robots to industrial equipment.

Marketing leaders should explore how a new generation of perceptive devices could help cultivate closer and more responsive relationships with customers.

Operations executives should evaluate how such devices—including the evolving crop of augmented-reality tools for industry—could help their people deliver an efficiency and quality edge.[18]

Cyber risk professionals should explore how mobile machine learning may present new ways of detecting and mitigating threats targeting mobile devices. Qualcomm is already promoting its Snapdragon 820 processor's machine learning-based malware detection capabilities; there will surely be other offerings that take advantage of machine learning for this purpose.[19]

The arrival on mobile devices of machine learning and enhanced perceptual computing is likely to have a big impact on a wide range of products, applications, and business practices over the next 18–24 months. It is time to start preparing for this era of mobile machine learning and perceptual computing.

Endnotes

1. Otavio Good, "How Google Translate squeezes deep learning onto a phone," Google Research blog, July 29, 2015, http://googleresearch.blogspot.com/2015/07/how-google-translate-squeezes-deep.html
2. Lenovo, "Lenovo and Google partner on new Project Tango device," January 7, 2016, http://news.lenovo.com/news-releases/lenovo-and-google-partner-on-new-project-tango-device.htm; Project Tango, www.google.com/atap/project-tango/, accessed March 20, 2016.
3. Dustin Franklin, "NVIDIA Jetson TX1 supercomputer-on-module drives next wave of autonomous machines," NVIDIA, November 11, 2015, https://devblogs.NVIDIA.com/parallelforall/NVIDIA-jetson-tx1-supercomputer-on-module-drives-next-wave-of-autonomous-machines/.
4. Snapdragon blog, "Live from New York, it's Snapdragon 820: Prepare for an immersive dive into mobile experience," November 10, 2015, www.qualcomm.com/news/snapdragon/2015/11/10/live-new-york-its-snapdragon-820-prepare-immersive-dive-mobile-experience.

5. DJI, “DJI launches new era of intelligent flying cameras,” March 2, 2016, www.dji.com/newsroom/news/DJI-Launches-New-Era-of-Intelligent-Flying-Cameras.

6. Youkyung Lee, “Go-playing program AlphaGo defeats human champion 4:1,” Associated Press, March 15, 2016, http://hosted2.ap.org/APDEFAULT/f70471f764144b2fab526d39972d37b3/Article_2016-03-15-AS-SKorea-Game-Human-vs-Computer/id-bb42c197629048e1b2de71f16ef86b4b.

7. For an overview of cognitive technologies, see David Schatsky, Craig Muraskin, and Ragu Gurumurthy, *Demystifying artificial intelligence*, Deloitte University Press, November 4, 2014, http://dupress.com/articles/what-is-cognitive-technology/?coll=12201.

8. Haşim Sak et al., “Google voice search: faster and more accurate,” Google Research blog, September 24, 2015, http://googleresearch.blogspot.com/2015/09/google-voice-search-faster-and-more.html.

9. Yaniv Taigman et al., “DeepFace: Closing the gap to human-level performance in face verification,” Research at Facebook, June 24, 2014, https://research.facebook.com/publications/deepface-closing-the-gap-to-human-level-performance-in-face-verification/.

10. Chuck Rosenberg, “Improving photo search: A step across the semantic gap,” Google Research blog, June 12, 2013, http://googleresearch.blogspot.com/2013/06/improving-photo-search-step-across.html.

11. See, for instance, Samir Kumar, “Qualcomm Zeroth is advancing deep learning in devices,” Qualcomm OnQ blog, March 2, 2015, www.qualcomm.com/news/onq/2015/03/02/qualcomm-zeroth-advancing-deep-learning-devices-video, with power efficiency described in Hubert Nguyen, “Qualcomm Snapdragon 820 speed & features,” Ubergizmo, January 14, 2016, www.ubergizmo.com/articles/snapdragon-820-speed-features/. See also Movidius, “Google and Movidius to enhance deep learning capabilities in next-gen devices,” January 27, 2016, www.movidius.com/news/google-and-movidius-to-enhance-deep-learning-capabilities-in-next-gen-devices. Farshid Sabet, chief business officer at Movidius, told us its latest processor consumed one-tenth to one-twentieth of the power of a contemporary mobile system on a chip for comparable computer vision tasks. (Interview on March 7, 2016.)

12. Larry Hardesty, “Energy-friendly chip can perform powerful artificial-intelligence tasks,” MIT News, February 3, 2016, http://news.mit.edu/2016/neural-chip-artificial-intelligence-mobile-devices-0203; interview with IBM’s Dharmendra Modha, March 17, 2016.

13. Good, “How Google Translate squeezes deep learning onto a phone”; Tarantola, “Microsoft Translator gets offline and photo-based features.”

14. See Ian McGraw et al., “Personalized speech recognition on mobile devices,” Google, March 11, 2016, http://arxiv.org/pdf/1603.03185.pdf.

15. This is an extension of the idea of edge analytics described in David Schatsky and Avinav Trigunait, *Internet of Things: Dedicated networks and edge analytics will broaden adoption*, Deloitte University Press, January 21, 2016, http://dupress.com/articles/internet-of-things-iot-adoption-edge-analytics-wireless-communication-networks/?per=4005.

16. DJI, “DJI launches new era of intelligent flying cameras.”

17. NVIDIA, “NVIDIA boosts IQ of self-driving cars with world’s first in-car artificial intelligence supercomputer,” January 4, 2016, http://NVIDIAnews.NVIDIA.com/news/NVIDIA-boosts-iq-of-self-driving-cars-with-world-s-first-in-car-artificial-intelligence-supercomputer.

18. To learn more about how companies are already employing augmented and virtual reality, see Nelson Kunkel, Steve Soechtig, Jared Miniman, and Chris Stauch, *Augmented and virtual reality go to work*, Deloitte University Press, February 24, 2016, http://dupress.com/articles/augmented-and-virtual-reality/.

19. Qualcomm Protective Intelligence, www.qualcomm.com/products/snapdragon/security/smart-protect, accessed March 20, 2016.

About the author

David Schatsky is a senior manager at Deloitte LLP. He tracks and analyzes emerging technology and business trends, including the growing impact of cognitive technologies, for the firm's leaders and its clients.

David Schatsky
Senior manager
Deloitte LLP
+1 646-582-5209
dschatsky@deloitte.com

Acknowledgements

The author would like to acknowledge the contributions of **Sherean Ali, Ragu Gurumurthy, Craig Muraskin,** and **Michael Tam** of Deloitte LLP; **Elliott Torres** and **Phil Wilson** of Deloitte Consulting LLP; **Patrick Duessel, Abhi Singh,** and **Craig Wigginton** of Deloitte & Touche LLP; **Jeff Bier** of the Embedded Vision Alliance; **Dharmendra Modha** of IBM Corp.; **Farshid Sabet** of Movidius Inc.; and **Edgar Auslander** of Menlo Business Partners.

Deloitte
University
Press

Follow @DU_Press

Sign up for Deloitte University Press updates at DUPress.com.

About Deloitte University Press

Deloitte University Press publishes original articles, reports and periodicals that provide insights for businesses, the public sector and NGOs. Our goal is to draw upon research and experience from throughout our professional services organization, and that of coauthors in academia and business, to advance the conversation on a broad spectrum of topics of interest to executives and government leaders.

Deloitte University Press is an imprint of Deloitte Development LLC.

About this publication

This publication contains general information only, and none of Deloitte Touche Tohmatsu Limited, its member firms, or their related entities (collectively the "Deloitte Network") is, by means of this publication, rendering professional advice or services. Before making any decision or taking any action that may affect your finances or your business, you should consult a qualified professional adviser. No entity in the Deloitte Network shall be responsible for any loss whatsoever sustained by any person who relies on this publication.

About Deloitte