White Paper

Mobile AI and the Future of Intelligent Devices

Sponsored by: Huawei

Francisco Jeronimo
November 2017

EXECUTIVE SUMMARY

History has shown us that a few technological revolutions have changed the course of humanity. Artificial intelligence (AI) stands on the brink of the next big technological revolution that will fundamentally change society for the better. AI algorithms are being applied to nearly everything we do – buying a product, watching a film, interacting with virtual assistants on smartphones, trading software, robotics, autonomous cars, or diagnosis of medical conditions. Artificial intelligence will be as profound as the steam engine was at the start of the Industrial Revolution.

Smartphones will bring AI into everyone's hands. In 2016, 5.6 smartphones were sold for each PC sold. The adoption of smartphones continues to outgrow any other consumer electronic device, and by 2021 IDC estimates that over 60% of the world's population will own a smartphone. As the most widely adopted consumer electronic device to date, the smartphone will be the pervasive platform for AI, and where the next battleground will take place.

IDC research shows that smartphones will become intelligent machines, delivering unique experiences by being able to predict what the user will do next and taking actions on his or her behalf. For this to happen, AI algorithms need to use vast amounts of data to learn and train the models. Cloud computing has been central to making AI possible on smartphones. But cloud computing has limitations – latency, privacy risks, and network dependency. A new approach is required for AI to become truly ubiquitous. On-device AI, using dedicated AI chipsets, can solve these problems by bringing the computing to the edge, improving the user experience at the same time.

IDC believes that the future of smartphones lies in mobile AI, the symbiosis between on-device AI and cloud AI. This combination will use the power of the cloud to train AI algorithms using huge amounts of unstructured data, while dedicated built-in AI chipsets will enable the device to "inference" (apply what the AI algorithms learned during training to the new data captured to "infer" the correct results).

Mobile AI will disrupt the future of smartphones and new advanced applications will emerge. With dedicated on-device AI chipsets, mobile devices will have better knowledge of the user and deliver automated personalized services and experiences. However, designing a chipset that delivers the performance required by highly demanding AI algorithms with the restrictions of the smartphone size, power supply, and power consumption is a major challenge.

In this white paper, IDC explores the future of smartphones from a mobile AI perspective and how it will impact the lives of every smartphone user. The paper looks at Huawei, as it has adopted this strategy by introducing a new Hi-Silicon AI mobile computing architecture with a dedicated neural-network processor unit on its latest chipset – the first of its kind.

METHODOLOGY

In preparing this document, IDC conducted interviews with a number of key industry players offering artificial intelligence solutions and services, as well as companies carrying out R&D in this field. IDC's extensive research on the topic was also used for the paper, supplemented by various secondary sources, including press releases, marketing and technical literature, filings with the Securities and Exchange Commission (SEC), quarterly conference calls, and reports published in trade and business journals. The information was then filtered through IDC's analysis process.

IN THIS WHITE PAPER

In this white paper IDC discusses how artificial intelligence (AI) will impact the most widely adopted consumer electronic device to date — the smartphone. The paper starts by looking at what AI is, how it has evolved, the different components, and the challenges and benefits of mobile AI, a symbiosis between cloud AI and on-device AI. The paper looks at Huawei, which is pioneering the creation of intelligent devices through its mobile AI strategy.

SITUATION OVERVIEW

British mathematician Irving John Good originated the concept of "intelligence explosion" in 1965. According to Good, "In order to design an ultra-intelligent machine we need to understand more about the human brain or human thought or both." He also defined an ultra-intelligent machine as a machine "that can far surpass all the intelligent activities of any man however clever."

Stanley Kubrick used Good as a consultant on the movie *2001: A Space Odyssey*, one of the first movies to address artificial intelligence (it was released in 1968). Science fiction is full of computers and machines that think, and what seemed like fantasy in the past is now technology that is disrupting the way we live. However, there are a lot of misconceptions about AI. Concerns about computers dominating the world emerge every time news spreads of computers outplaying the world's best games players, or when robots replace humans in the workplace. But the reality is that we are still many years away from truly intelligent machines that can fully mimic human brains, and the ultimate goal is to make AI work for us, not against us.

AI is very different from human intelligence. How long it will take for a computer to think in a similar way to a human being remains to be seen. Although the advancements in AI have not yet created a thinking machine, there have been significant breakthroughs over the past six decades. In March 2016, AlphaGo, a computer program developed by Google DeepMind, beat Lee Sedol, one of the world's best players of Go, a complex game that requires creativity and intuition. In October 2017, AlphaGo Zero mastered the game from scratch in just 72 hours, with no human intervention, apart from being told the rules. This important breakthrough opens the door to building general-purpose algorithms essential for AI to solve real-world problems.

But AI is not only available on computers, robots, and machines. The smartphone is the next battleground for AI. Smartphones have changed the way we communicate, play, and work. In 2011 smartphone sales overtook sales of PCs (including desktops, notebooks, and workstations) for the first time, making mobile the largest computing platform in the world. IDC estimates that by the end of 2017, half of the world's population will own a smartphone, and that 8.1 billion smartphones will be shipped in the next four years — making mobile the pervasive platform for AI.

Smartphones are transforming our lives at an extraordinary pace as the most advanced technologies and features become available on the phones people use today. High-quality cameras, fast processors, connectivity, and applications have brought smartphones to the center

of everyone's lives. But state-of-the-art specifications do not necessarily mean intelligence. Most phones offer cameras capable of taking good pictures, but there is a difference between taking a good picture and "knowing" what is in that picture. When we point a camera at an object, a person, or a landscape, we want our phones to understand the context, the content, and the environment, and decide the best combination of settings to get the best shot, effortlessly.

Dedicated AI chipsets will be available in most premium smartphones soon. This will fundamentally disrupt the future of smartphones and offer new user experiences and capabilities that are not possible today, making them seamlessly intuitive. Huawei is the first manufacturer to offer a new HiAI (Hi-Silicon AI) chip with dedicated neural-network processor unit (NPU) in its latest flagship devices, the Huawei Mate 10 and Mate 10 Pro, which were unveiled at an event in Munich in October 2017. The company claims that the new Kirin 970 chipset delivers higher performance than any CPU/GPU build, with higher efficiency.

What is Artificial Intelligence?

Artificial intelligence was founded as an academic discipline in 1956. The concept was proposed during the summer of 1956, when John McCarthy, regarded as the father of AI, proposed the name for a summer research project that took place at Dartmouth, New Hampshire. The project proposed to "find out how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves." Since then, there have been many different definitions put forward, including this from the Oxford English Dictionary:

> "Artificial intelligence is the theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision making, and translation between languages."

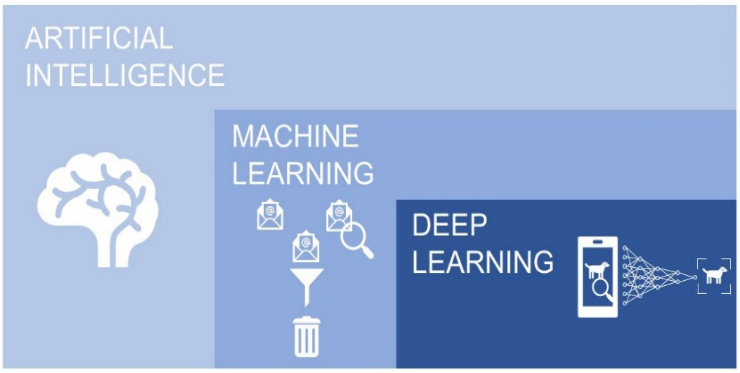
This definition fits the context of this paper because it specifically focuses on what computers and smartphones will be able to do in the next two to five years, and not on science fiction scenarios. Whether a machine can act in a similar way to humans or not is still an open question. Either way, in its most basic form, AI is a set of algorithms that perform specific tasks, either by self-learning or by analyzing vast amounts of data, to predict the right output to complete the task. As adoption of mobile devices continues to grow, AI algorithms will be at the heart of every single consumer electronic device in the future.

The Importance of Machine Learning and Deep Learning

The advances we are seeing in artificial intelligence are due to the progress being made in machine learning, in particular in the area of deep learning. These concepts tend to be used interchangeably, but they are not exactly the same thing. AI is considered the umbrella that encompasses these two areas.

FIGURE 1

Artificial Intelligence as the Umbrella for Machine Learning and Deep Learning



Source: IDC, 2017

Machine Learning

Machine learning is a subfield of AI and one of its main applications. It gives systems the ability to self-learn from analyzing new datasets without being specifically programmed to do so. Computer programs using machine learning will be able to self-learn without human intervention. Despite not being a new science, it has gained strong momentum in recent years, with the ability to automatically apply complex mathematical calculations to Big Data.

Machine learning is being used on smartphones and smart devices for speech recognition, predictive text, face tracking camera, fingerprint ID, computational photography, and digital assistants. Speech recognition, for instance, is critical for mobile devices to be able to interact with users just by using voice. Also critical is the ability for the smartphone to learn the user's behavior and optimize its performance accordingly. If the smartphone anticipates what apps will be used next, it can prioritize resources, which will provide better responsiveness and better power consumption, offering a smoother experience and making the battery last longer.

Deep Learning

Deep learning is considered a subset of machine learning and is the most powerful form of machine learning being used today. It utilizes a hierarchical level of artificial neural networks, capable of learning unstructured or unlabeled data, where the network looks for recurring patterns. While traditional programs analyze data in a linear way, deep learning systems take a non-linear approach. The hierarchical function of deep learning allows different layers of the neural network to process data and pass it to another layer as output. The next layer takes the previous layer's information and combines it with more data, passing it again to the next layer. This continues from layer to layer in the neuron network until the best result is determined.

Deep learning is remarkably good at perception related tasks, such as object recognition, image recognition, speech recognition, and natural language processing. As most of these tasks will happen on smartphones, deep learning has become important to offer new use cases for smartphones that are not even possible yet, with applications ranging from medical diagnosis to image recognition.

Artificial Intelligence is Going Mobile

Edge computing is the future of computing. The need for no latency, higher security, faster computing, and less dependence on connectivity will drive the adoption of devices that are able to offer AI at the edge. On-device AI uses dedicated AI chipsets that will become prominent in all flagship smartphones from this year.

The developments in the past 12 months indicate a huge focus on on-device AI from a variety of companies:

- Huawei announced the Kirin 970, a new flagship SoC with built-in AI computing capabilities, at IFA 2017 in Berlin. The AI platform runs on a dedicated neural processing unit (NPU) that Huawei claims delivers up to 25 times more performance than a quad-core Cortex A-73 CPU cluster. Huawei is also positioning the Kirin 970 as a platform for mobile AI, opening the chipset to developers through an SDK, and supporting TensorFlow/TensorFlow Lite and Caffe/Caffe 2 platforms.
- Apple's new iPhone X comes with a new A11 Bionic chip with neural engine. Apple claims the chip, with a dual-core design, performs up to 600 billion operations a second for real-time processing, designed for specific machine learning algorithms.
- Baidu launched Mobile Deep Learning (MDL), its open source mobile deep learning framework designed to fit on a smartphone.
- Qualcomm launched the Snapdragon Neural Processing Engine (NPE), an SDK that allows developers to optimize apps to run different AI applications on the Snapdragon 600 and 800 series processors. This gives developers access to tools that run deep neural networks on mobile and other edge devices. It supports Caffe/Caffe2 and TensorFlow.
- NVIDIA launched Jetson TX2, an AI platform for drones, robots, smart cameras, and other embedded devices, specifically designed for edge computing and applications such as navigation, image recognition, and speech recognition.
- Samsung launched a new chipset, the Exynos 8895, with a separate processing unit that enhances the security required for mobile payments that use iris or fingerprint recognition. It also features an embedded vision processing unit (VPU) that can recognize and analyze items or movements for improved video tracking, panoramic image processing, and machine vision technology. Samsung says that adding NPUs to smartphones will become a mainstream trend, and we expect the Korean company to enable its Exynos chipset line with dedicated AI capabilities.

FUTURE OUTLOOK

According to the latest IDC forecast published in September 2017, worldwide spending on cognitive and AI systems is forecast to reach $57.6 billion in 2021.

"We are seeing cognitive and AI technology and solutions weaving into an ever broader and wider array of applications and use cases," said David Schubmehl, research director, Cognitive/Artificial Intelligence Systems, IDC. The growth in cognitive/AI is happening across the spectrum of enterprise software, services, and hardware to adopt and use intelligent applications based on artificial intelligence, he said.

Despite AI being around since the 1950s, exposure to it has been largely limited to science fiction movies. But that is changing fast as people start to realize how AI algorithms are being used in many areas of their lives and become familiar with new advanced applications of AI. Examples range from using personal digital assistants, such as Siri, Cortana, and Amazon Alexa, to more advanced cases where AI helps diagnose diseases or enables self-driving cars to avoid accidents. When someone asks Amazon Echo or Google Home a question, many types of AI are used,

including speech recognition algorithms that "translate" the speech into a natural language. AI will ultimately become a general-purpose technology (GPT) and a key component of most of the devices we use in our lives.

AI offers unprecedent possibilities to disrupt people's lives through smartphones. Google CEO Sundar Pichai wrote in 2016 that "the last 10 years have been about building a world that is mobile-first, turning our phones into remote controls for our lives. But in the next 10 years, we will shift to a world that is AI-first, a world where computing becomes universally available – be it at home, at work, in the car, or on the go – and interacting with all of these surfaces becomes much more natural and intuitive, and, above all, more intelligent."

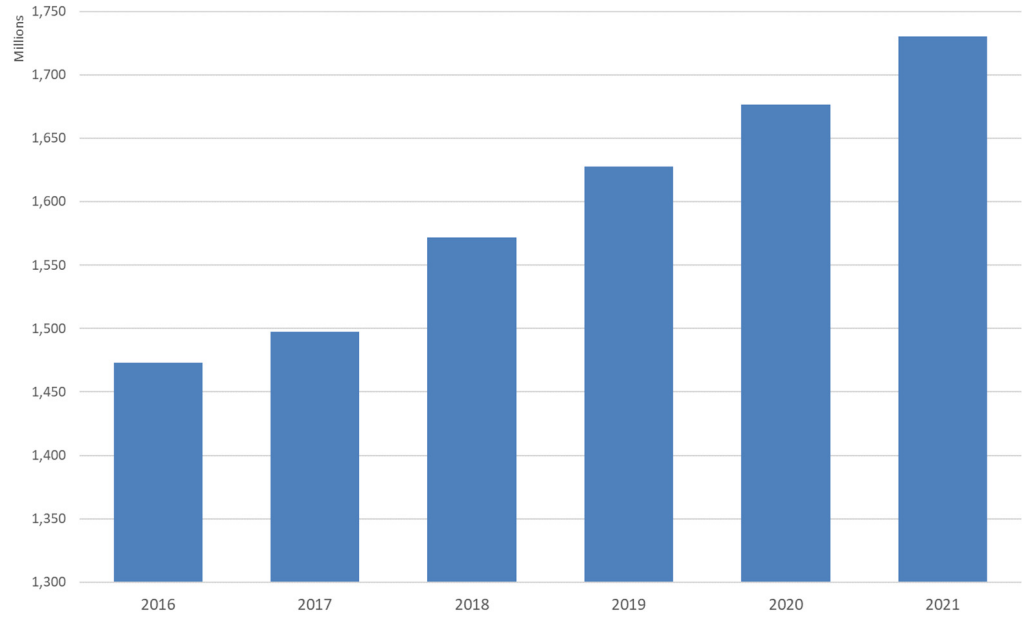
IDC believes that the smartphone will be central to this revolution, and will play as pivotal a role as the steam engine did at the start of the Industrial Revolution.

The Future of Smartphones and a New Paradigm Shift

Smartphones have been one of the most successful consumer electronic devices ever. With annual sales over 1.5 billion units, they continue to dominate consumer device shipments. IDC estimates that by 2021 over 60% of the world's population will be using a smartphone. In markets such as the Nordics, smartphone penetration is already over 85%.

FIGURE 2

Worldwide Smartphone Unit Shipments Forecast, 2017-2021



Source: IDC Quarterly Mobile Phone Tracker, August 2017

Although we see continuous growth in smartphone sales, the last paradigm shift in the market took place 10 years ago, when Apple announced its first iPhone in January 2007 and Google launched Android OS in November that year. Soon after that came the first Android device and app stores. In the past 10 years, innovation in the smartphone space has come from incremental (but huge) improvements in hardware and software. But nothing has fundamentally changed. The form factor, the user interface, the app stores, and the user experience remain basically the same, only exponentially better than a decade ago.

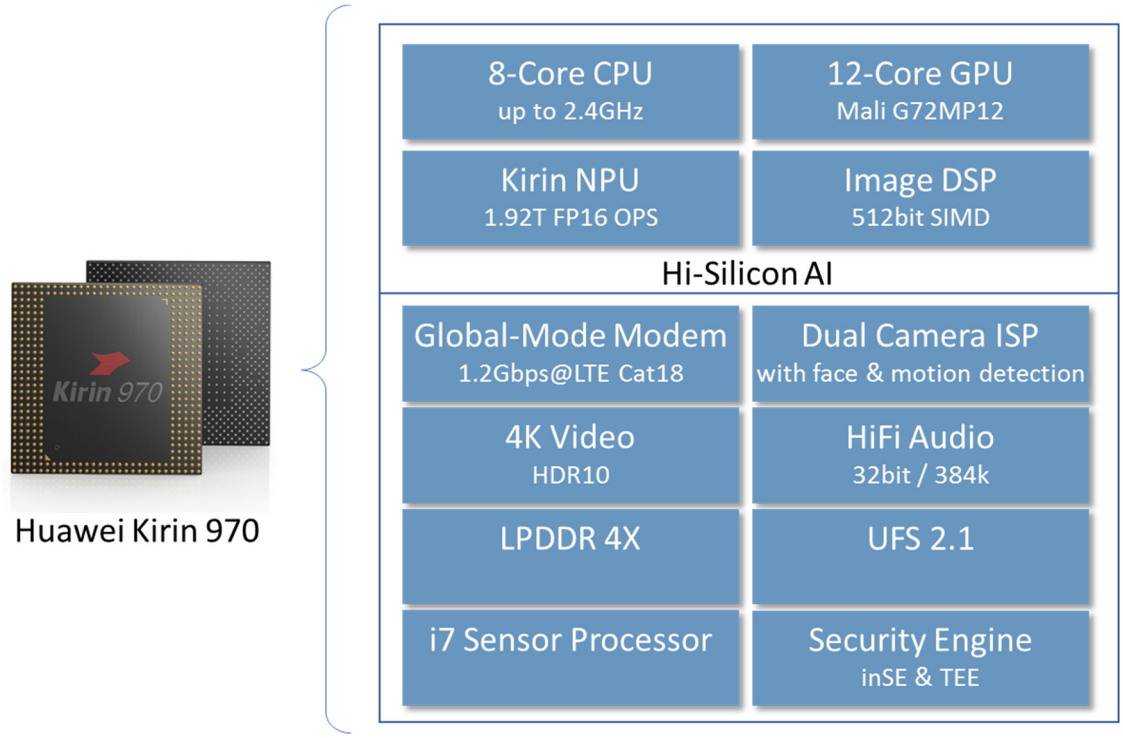
With AI, IDC believes we are on the edge of a new paradigm shift in the smartphone market. AI represents a major opportunity for smartphone makers to offer mobile devices that are not just smart, but intelligent. AI will drive the evolution of smartphones to the next generation of "intelligent phones," which will make them truly personal assistants – devices that go beyond performing the tasks we request at a given time, to devices that start acting on our behalf, by predicting and executing the tasks that are relevant to each person at the right time of the day. This new capability will come from learning how the user behaves, understanding patterns of actions and the relevance of decisions based on the individual's context and environment.

Leading industry players and manufacturers are developing technologies to bring AI to their products. IDC predicts that AI capabilities, from less demanding machine learning algorithms (e.g., predictive text, automated calendar entries based on content from emails), to dedicated AI chipsets and advanced artificial neural networks and deep learning (e.g., identifying people and other elements in a photo), will be available in most smartphones sold in the coming years.

Huawei is already paving the way. At IFA 2017 in September, the Chinese manufacturer announced its vision of the future of smartphones and artificial intelligence. The company launched the Kirin 970, a new flagship SoC with built-in AI computing capabilities, which will change the way we interact with our devices and significantly improve the user experience. According to Richard Yu, CEO of Huawei Consumer Business Group, "As we look to the future of smartphones, we are at the threshold of an exciting era. AI is no longer a virtual concept but something that intertwines with our daily life."

FIGURE 3

Huawei Smartphone SoC Chipset With Dedicated Neural-Network Processing Unit



Source: Huawei, 2017

Because it is inefficient to use smartphones' CPUs and GPUs to process AI algorithms, in particular deep learning algorithms, Huawei has developed a hardware acceleration architecture suitable for neural-network algorithm processing. "Traditional computing architecture based on CPUs, GPUs, and DSPs is no longer sufficient to meet overwhelming demand for computing performance in the age of artificial intelligence," said Victor Dragnea, marketing product manager at Huawei CEE and Nordic. "On-device processing capabilities are of critical importance because mobile devices need to process data in real time, any time, while providing optimal protection of user data."

The Kirin 970 chipset is powered by an 8-core CPU and a 12-core GPU. Huawei claims that compared with a quad-core Cortex A-73 CPU cluster, it delivers up to 25 times the performance with 50 times the efficiency, performing the same AI computing tasks faster and with less power consumption. The company put it to test and in a benchmark image recognition test the new chipset processed over 2,000 images a minute, faster than previous chipsets without NPU.

The first two devices using the Kirin 970 chipsets are the Huawei Mate 10 and Huawei Mate 10 Pro, recently announced by Huawei.

Apple unveiled the Apple A11 Bionic chipset, available in its latest flagship devices, in September 2017. It has two high-performance cores and includes dedicated neural-network hardware — Neural Engine — used for face ID, machine learning algorithms, and other features. Google also announced two new Pixel devices with AI features, and Samsung will soon follow.

From Smart to Intelligent Mobile Devices With Mobile AI

Although smartphones offer faster CPUs and GPUs for basic machine learning algorithms, accessing more powerful ways of computing data is vital if apps are to take advantage of AI. Advanced algorithms need huge amounts of data and computing power to learn and train the models, so they can deliver the experience they are designed to. Therefore cloud computing has been the cornerstone of bringing AI algorithms to smartphones.

The data collected by the sensors on the phone is sent to cloud services for processing and the output is returned over the network to the device. This allows mobile applications to use powerful computers in datacenters to ease the weight of processing information on a local device. However, there are major limitations in primarily using the cloud to deliver AI-based services:

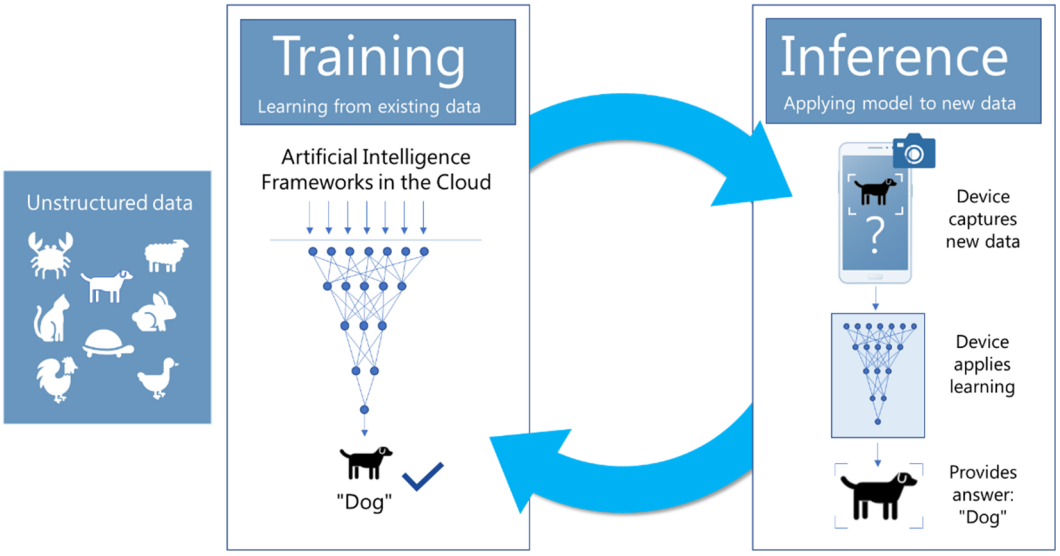
- The data needs to travel to the cloud and back to the device through the network, which can result in inefficiencies due to network latency.
- Privacy of the data can be compromised when travelling outside the device.
- Poor network coverage makes it impossible to access the cloud servers.

These issues can have a significant impact on the user experience. Most smartphone users have already experienced at least one of these limitations. Those using voice search or digital assistants understand how frustrating these applications are in areas with limited or no network coverage.

A new approach is therefore needed, where computing AI algorithms at the edge become an important part of the process. While most smartphone makers continue to rely mainly on the cloud to run their AI algorithms, Huawei is pursuing a mobile AI approach, combining the cloud AI with on-device AI. By using the power of the cloud to train AI algorithms, with the power of edge computing, Huawei is able to overcome the limitations of cloud AI.

FIGURE 4

Training and Inference Are Two Vital Components of AI on Smartphones



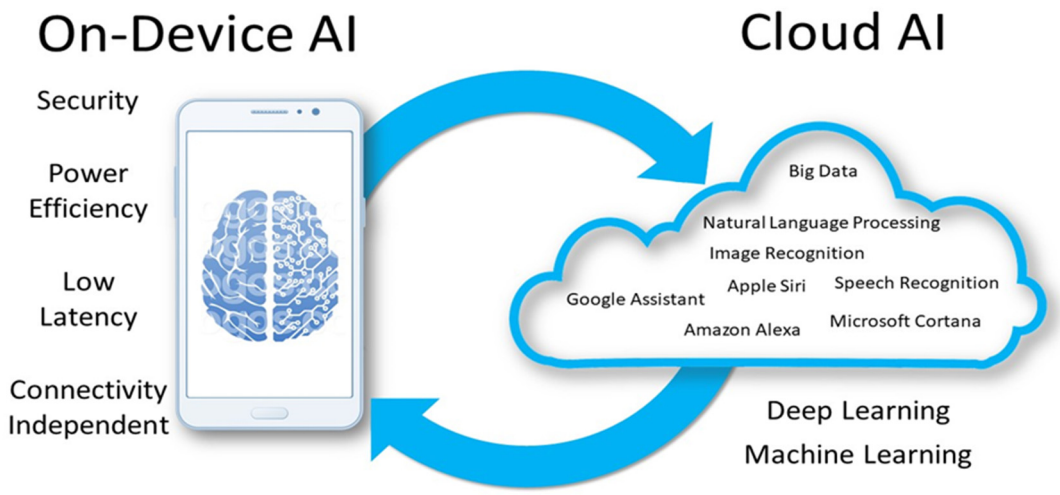
Source: IDC, Huawei, Qualcomm, Nvidia, 2017

A key aspect of machine learning and deep learning is that both need to use vast amounts of data and computing power to train the models. Much of the data captured by the smartphone, including images, video, and sound, is unstructured. Unlike structured data – information with a degree of organization – unstructured data makes compilation a time- and energy-consuming task. It is impossible to instantly transform unstructured data into structured data, so creating intelligence from unstructured data is difficult. Training is an important part of AI and is essential for machine learning and deep learning models to work. Once the models learn from unstructured data, it is more efficient to process new data near its capture point to provide a seamless user experience. This is where inference is important – inference being the process of applying what the AI algorithms learned during training to the new data captured to "infer" the correct results quickly.

Inference will be used all the time on smartphones and not just when it is requested. This would include being in constant listening mode for the voice-activated assistant, when a photo is catalogued by the smartphone based on its content, and when the phone identifies the correct scene mode on the camera to adjust the settings to take the best shot. Because inference needs to process data in real time, all the time, a smartphone that relies only on the computing power of the CPU and GPU will not meet the demanding computing performance of the AI algorithms. Therefore it is of critical importance to provide on-device AI processing capabilities that bring to the device the knowledge acquired from the training in the cloud.

FIGURE 5

Mobile AI as the Symbiosis Between On-Device AI and Cloud AI



Source: IDC, Huawei, Qualcomm, 2017

How Mobile AI Will Disrupt the Future of Smartphones

It is impossible to list all the ways that mobile devices will evolve by performing more intelligent applications locally. But it is possible to understand that most of these new applications will perform advanced perceptual tasks involving vision, speech, and other sensory inputs. Mobile AI will be an important enabler and accelerator of new use cases for smartphones and an important differentiator of operating systems, ecosystems, and manufacturers.

Some of the main high-level capabilities that mobile devices will be able to perform are:

- **Data analysis from different sensor inputs.** Smartphones will be able to bring computer vision to a new level. The camera on the phone will be used in applications not even available today, from detecting illness or medical conditions in healthcare to generating 3D models quickly and easily for interior design applications. Analyzing and identifying the content of data collected from the different sensors or companion devices connected to the phone (e.g., wearables) will trigger actions on the user's behalf.
- **Perceptual user interfaces.** Human-computer interaction has not changed fundamentally for nearly two decades. Smartphones brought touchscreens to end users, but we still interact with devices by clicking, typing, and pointing. With AI we will see mobile devices capable of interpreting the user's movements or voice commands. AI will enable the device to understand natural human capabilities (particularly communication, motor, cognitive, and perceptual skills). Low-power-consumption processors designed for AI will help consumer and industrial mobile devices to read, understand, and respond to the context and environment around them.
- **Security.** With personal and corporate information now living on mobile devices, the need for robust security solutions that don't drain the battery or take lots of storage is essential. AI is a natural fit for mobile security solutions. AI will make devices a lot more secure by identifying potential threats and performing real-time actions to prevent or alert users to security threats.
- **Navigation and motion control.** Low-power chips with powerful computer vision will bring new and advanced capabilities to smartphones and other devices such as drones and

improve indoor navigation by using computer vision to precisely locate users, track their motion, and guide them.

With dedicated on-device AI chipsets, mobile devices will have better knowledge of the user's needs, being able to deliver personalized services that will make smartphones more intelligent. Beyond speed and efficiency, on-device AI offers greater security by providing real-time malware detection, recognizing if the device is being misused by identifying user behavior, and spam detection over emails and other apps.

Among the applications and features that will benefit first from on-device AI are:

- **Virtual digital assistants.** By being less dependent on cloud AI and connectivity, virtual assistants can become the main method for users to interact with the devices.
- **Augmented reality.** Augmented reality will also see similar benefits, as it requires computation to take place in the cloud.
- **Cameras.** Having an on-device processor that constantly analyzes what the camera "sees" as users take photos or adjust the brightness, ISO sensitivity, color temperature, and exposure duration every time they press the shutter button, and accurately selecting the right scene mode automatically, will have an impact on the user experience.
- **Well-being and healthcare.** By delivering faster performance — regardless of the quality of the network, while providing optimal protection of user data — many apps will be able to send notifications to the user (and his or her doctor) in response to real-time analysis of patient data and data collected from wearable devices. This will predict health events, enabling users to seek medical advice and for doctors to assist patients when a medical condition or threat is detected by an AI algorithm.

Bringing Intelligent Mobile Devices Into Our Lives

AI will disrupt the way users interact with smartphones in the future, making it work for us. Going to the cinema, for example, will be a completely different experience from the one we have today:

1. Someone with an intelligent phone is walking down the street and sees an ad for the latest blockbuster movie. On pointing the phone at the ad, the camera recognizes the movie, suggests how likely they are to enjoy it, how much it matches their preferences, and prompting them to ask if it should buy tickets.
2. The phone checks the calendar and suggests when to get the tickets, based on the commitments and appointments for the week, as well as the best theater based on the predicted location for the day. The phone then sends a message to the person who usually goes along, suggesting the movie and asking if they want to come.
3. The phone buys the tickets and stores them in its digital wallet.
4. On the day, the phone confirms availability and suggests having dinner before the movie, at a nearby restaurant bookmarked on the maps app, reminding the user about the type of food served. After confirmation by the user, the phone makes reservations on their behalf.
5. Based on the traffic, the phone sends a reminder when it is time to leave and, if late, it sends an email to the restaurant with an estimated time of arrival. It automatically provides directions to the restaurant from the car park.
6. At the theater the phone automatically displays the ticket on the home screen for faster scan.

One single action triggered from the camera will create a chain of actions that will help people manage their lives. AI will make technology work for everyone, rather than people having to adapt to the technology's limitations. Smartphones will learn who the users are, what they do and what they want to do, and deliver a new user experience, one that is not yet available, effortlessly.

AI will enable phones to become intelligent and truly personal digital assistants, and it is on-device AI that will be the engine to make this happen.

Benefits of Mobile AI

The opportunities for always-on devices that do most of their intelligent computing on the device are enormous. Running AI on the device rather than in the cloud offers different benefits too:

- **No latency.** In many cases, running AI entirely in the cloud will have an impact for applications that need to work in real time and are latency-sensitive, such as mission-critical applications and driverless cars. Such applications need to rely on instantaneous responses and cannot afford the roundtrip time or operate when there is variable network coverage.
- **Increased security and privacy.** AI algorithms look for patterns in the data at an unprecedented scale. The data collected by the phone will require a roundtrip to the cloud if no AI capability is available on the device. Being able to perform AI algorithms on the device will provide better security due to the low volumes of data exchanged over the network. But it will also provide better privacy, as the AI algorithms will not have to process and store the information in the cloud. The cloud is used only to train the algorithms. This is paramount in today's environment with new European legislation (GDPR) that aims to provide a set of standardized data protection laws across member countries, but also to reinforce the data protection rights of individuals.
- **On-device learning.** Although most of the training will happen in the cloud, and the smartphone has inference capabilities, it still needs to learn the user's behavior so it can deliver automated, personalized experiences. On-device learning enables training capabilities on the mobile device, so the data does not need to travel to the cloud or be stored outside the device. This process can be triggered while the device is idle, which allows power saving and ensures no interruptions of other features.
- **Efficiency.** Running AI on the device reduces network traffic, as less data is transferred via the network to the cloud, and improves performance as apps that need to run in real time can achieve lower latency levels on the device.
- **Power saving.** Power consumption will be reduced, as the device does not need to constantly upload and download AI-related data to be able to process AI algorithms. This will have a positive effect on battery life and, by extending it, will improve user experience.

Challenges of Delivering Mobile AI

Bringing machine learning and deep learning algorithms to the edge is essential to lower their data computation requirements, while enabling mobile devices with chipsets that can process some of those algorithms on the device. AI workloads are very compute-intensive. Optimizing AI algorithms for use on mobile is a huge challenge, as is the impact on battery life of the intensive computational AI algorithms on a real-time and always-on environment.

Another important challenge with on-device AI is the need to train the algorithms. This requires vast amounts of data. The quality and quantity of data is key to successful machine learning. Depending on the complexity of the model and the amount of data, training can take place in the device or in the cloud. Large models tend to require a lot of processing, which is only available in cloud platforms. Some of the predictions needed from AI algorithms require data that is already available on the user's device and no other sources are needed to be trained. However, to train and run large, complicated neural-network models on mobile devices with less capabilities than large servers is a challenge.

A fundamental challenge of offering on-device AI is related to the nature of the smartphone. Designing a chipset that delivers the performance required by highly demanding AI algorithms with

the restrictions of smartphones in terms of size, power supply, power consumption, and heat management is a major challenge. "Mobile SoCs have to deliver the best possible performance, while ensuring the highest possible density of core functions with optimal use of energy. The [Huawei] Kirin 970's development team introduced a new, innovative HiAI mobile computing architecture with a dedicated NPU," said Victor Dragnea from Huawei.

Developing an AI Ecosystem

AI platforms provide developer toolkits to build applications using AI algorithms. They combine a variety of algorithms with data that can be used directly by developers, without having to build them from scratch. Some of the functionality includes image recognition, natural language processing, voice recognition, predictive analytics, and other machine learning and deep learning capabilities.

An important part of bringing AI to smartphones is the creation of an AI ecosystem. This is vital to expand the capabilities of the dedicated AI chipsets from a few features on the phone to third-party apps. By providing SDKs and APIs, phone makers will enable developers and partners to find new uses for AI computing capabilities.

Huawei is positioning the Kirin 970 as an "open platform for mobile AI." The company emphasizes that the chipset will work with any AI framework, such as Caffe2 and TensorFlow, in order for any app to access the capabilities of the NPU and AI features.

CONCLUSION

IDC research shows that the possibilities of AI on smartphones will represent a paradigm shift that will transform people's lives. Enabled by AI, smartphones will start acting intelligently, rather than reactively, to users' requests. AI will be a technological revolution that will fundamentally change the way people live, work, communicate, and relate to each other.

To make AI truly ubiquitous, the device needs to start taking advantage of the benefits of cloud computing and combine that with the capabilities that built-in AI chipsets will offer at the edge. This powerful combination will overcome some of the limitations of both the cloud and the device, and deliver an enhanced and personalized experience to the end user — only possible through a mobile AI approach.

Intelligent devices will be part of our lives and will contribute to our future — a very bright future that has only just started.

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications and consumer technology markets. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.

IDC U.K.

IDC UK
5th Floor, Ealing Cross,
85 Uxbridge Road
London
W5 5TH, United Kingdom
44.208.987.7100
Twitter: @IDC
idc-community.com
www.idc.com

Copyright and Restrictions