Bin Li

Embedded Artificial Intelligence

Principles, Platforms and Practices

清華大學出版社
TSINGHUA UNIVERSITY PRESS

Springer

Embedded Artificial Intelligence

Bin Li

Embedded Artificial Intelligence

Principles, Platforms and Practices

清華大学出版社
TSINGHUA UNIVERSITY PRESS

Springer

Bin Li
CTO Office
Beijing DataRealm Technology
Beijing, China

ISBN 978-981-97-5037-5 ISBN 978-981-97-5038-2 (eBook)
https://doi.org/10.1007/978-981-97-5038-2

Jointly published with Tsinghua University Press

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

Embedded artificial intelligence is a big topic!

Let's reflect on the origins of the term "artificial intelligence": "The science and engineering of creating intelligent machines, especially intelligent computer programs"—John McCarthy, 1956. Essentially, from the outset of artificial intelligence, the aim has been to create intelligent machines. This goal carries two implications: firstly, designing machines specifically for intelligence, and secondly, integrating intelligence into existing machines. While our efforts have predominantly focused on the former, the aspiration to embed intelligence into machines has been present from the beginning.

Typically, when discussing artificial intelligence, thoughts often turn to robots. In the extravagant realms of science fiction, AI embodies mechanical suits transforming ordinary individuals into heroes (like Iron Man), mutated entities threatening global destruction (like I, Robot), or endearing mechanical companions (like The Robotic Butler). The desire for machines resembling humans, capable of hearing, seeing, speaking, acting, and even contemplating, has persisted. Yet, achieving such machines remains profoundly challenging. With the emergence of ChatGPT, there's a dawn of hope—computers can now speak like humans! However, a significant challenge remains: can we achieve such marvels on machines like humanoid robots, with dimensions and power consumption comparable to humans? ChatGPT's training requires over 30,000 GPUs, with a total power exceeding 10 million watts and a daily electricity cost surpassing $50,000. In contrast, the human brain occupies merely about 1.5 L and operates on less than 20 W. If we aim to embed artificial general intelligence into robots, cars, or even smaller devices like drones, phones, smart appliances, or IoT devices, we must overcome substantial hurdles. This is precisely the topic explored in this book *Embedded Artificial Intelligence*.

The initial concept for this book emerged in 2018, inspired by a clever child's idea to create an autonomous flying sun umbrella to provide shade when he is playing. It was a cool idea but incredibly challenging! At that time, academic research regarding embedded artificial intelligence was just beginning, with scattered achievements yet to undergo large-scale practical validation. Implementing such

advanced functionality in a tiny processor for drones was still unrealistic. Nonetheless, we believe that one day it will be realized, so let's strive for that dream.

Over the following years, we delved into various possibilities to realize this dream. Thankfully, it's not just our personal aspiration; it's a collective dream of the entire industry. Research on embedded artificial intelligence surged like mushrooms after rain, with lightweight algorithms emerging, significant progress in model compression techniques, and AI acceleration chips tailored for embedded systems hitting the market. Thus, as we explored, summarized, and wrote, we roughly completed the first part of this book: the principles. Sharing it with some friends garnered a bit of encouragement, prompting the start of the second and third parts: the platforms and practices. Both parts compile data gathered during our efforts to realize the autonomous flying sun umbrella. Though they may be somewhat outdated by the time this book hits the market, they are comprehensive enough to offer valuable insights to readers.

In 2023, the Chinese version of this book was published. We were delighted to observe the rapid development of various platforms for embedded artificial intelligence and the blossoming of embedded AI practices. Discussions on embodied intelligence have begun, envisioning a brighter future with the new generation of autonomous vehicles and robots. Underlying these imaginations is the rapid advancement of embedded artificial intelligence. Key technologies in embedded AI, such as model compression, not only shine in the embedded domain but also serve as a crucial factor in deploying large-scale models like ChatGPT at a low cost. For models with trillions of parameters, even the largest GPUs become small chips! In the future, people will undoubtedly seek to implement brain-scale models on machines with the size and energy consumption of the human brain, presenting an entirely new challenge. The future of embedded artificial intelligence knows no bounds!

Finally, I would like to express gratitude to the mentors and friends who guided and assisted in the making of this book. This book comprehensively synthesizes previous achievements in the field of embedded artificial intelligence. Thus, first and foremost, thanks are due to the scholars who conducted pioneering research in this field, including but not limited to Xipeng Shen, Song Han, Shaoshan Liu, Mingxing Tan, Menglong Zhu, Forrest Iandola, François Chollet, Andrew Howard, Bert Moons, Daniel Bankman, Marian Verhelst, and others. This list is bound to omit some contributors due to the rapid pace of developments in this field; your understanding is appreciated. Secondly, I extend thanks to friends and colleagues who provided feedback and suggestions during the writing process. It's your tireless encouragement and assistance that enabled me to persist in completing this extensive book. Lastly, I must thank my family, my wife, Lucia, who meticulously translated every word of this book, and our child, Jerome, who contributed inspiration to this writing!

Beijing, China Bin Li

Contents

Part II Platforms

Part I
Principles

Chapter 1
Embedded Artificial Intelligence

Abstract What is embedded artificial intelligence? Why do you need embedded artificial intelligence? How to implement embedded artificial intelligence? What are the challenges of implementing embedded artificial intelligence? With these questions, we defined the topics to be studied in this book. After comparing the two implementation modes of embedded artificial intelligence: cloud computing mode and local mode, we clarified the necessity and technical challenges of implementing the local mode and outlined the five essential components needed to overcome these challenges and achieve true embedded AI.

Keyword Embedded AI

1.1 What Is Embedded Artificial Intelligence?

When thinking of artificial intelligence, most of us probably think of robots and computers. In the magnificent imagination of science fiction novels and movies, AI is a supercomputer that creates everything (such as The Matrix), a robot that destroys the world (such as Mechanical Enemy) or a spaceship that cruises in the vast universe (such as 2001: A Space Odyssey). However, it is not. In the early days, artificial intelligence was more of an intelligent system with decision-making capabilities. For example, there were applications designed specifically for spelling and grammar checking. When these applications were first introduced for computers, they were considered highly intelligent. These applications were among the earliest forms of AI, and today they are so commonplace that they no longer carry the title of AI. However, although AI is now more diverse in appearance, most of the time it is still just a large, complex computer software that implements some kind of "intelligent" algorithm to solve problems. For example, AlphaGo defeats humans, or AI characters in games, or translation software that understands all languages. But as we dream, AI has begun to be combined with devices to build truly intelligent machines. This is exactly the subject of this book, embedded artificial intelligence (abbreviated as embedded AI).

B. Li, *Embedded Artificial Intelligence*,
https://doi.org/10.1007/978-981-97-5038-2_1

So, what is embedded artificial intelligence?

Let's review the origin of the name artificial intelligence,

> *The science and engineering of making intelligent machines, especially intelligent computer programs—John McCarthy, 1956.*

In other words, from the early days of artificial intelligence, we have wanted to create intelligent machines. This sentence has two meanings. First, create machines specifically designed to achieve intelligence. Second, create machines with embedded intelligence. In the past few years, we have been working hard toward the first goal, but in fact from the beginning we have been looking forward to "embedding" intelligence into devices!

Traditionally, an embedded device refers to a device that embeds a computer system to achieve specialized functions and real-time computing performance. Compared with a general-purpose computer, on the one hand, it is "smaller." It is not universal and only meets certain specific needs. Therefore, it uses a processor with weaker performance, such as a microcontroller, and the memory is also limited by the size of the device. And smaller, on this basis, the software system running on the device is also lightweight, generally using a compact embedded operating system, and the application software only completes specific and limited functions and is smaller in scale. Benefiting from this streamlined software and hardware, embedded devices also consume less power, and some can be powered by batteries. But on the other hand, it is "bigger." Many embedded systems are mechatronics devices. In addition to computer systems, they also have sensors, execution components, etc. This allows it to interact more with the external world and achieve connection with the physical world. Another characteristic of embedded systems is real-time nature. They must respond to inputs within a limited time, just like a real living organism.

With the advancement of computer software and hardware technology, these embedded devices have become more and more powerful and "intelligent." With the advancement of computer software and hardware technology, these embedded devices have become more and more powerful, and their "intelligence" is getting higher and higher, such as smart phones, which have made significant progress in recent years, and their performance has caught up with general-purpose computers and can complete most of the daily computing tasks. Another example is smart home hardware, which can proactively turn on the air conditioner before the owner returns home, achieving the ability of independent analysis and decision-making to a certain extent. So, can it be said that these "smart hardware" has implemented embedded artificial intelligence?

Strictly speaking, not yet.

The artificial intelligence introduced earlier in this book is implemented on a general-purpose computer. In particular, this artificial intelligence is implemented by some intelligent software, which may be a chess program, an expert system, a deep neural network, or a robot operating system，running on a powerful general-purpose computer, such as GPU or supercomputer. Of course, this consumes a lot of power. Then, the goal of embedded artificial intelligence in the strict sense is to achieve equivalent artificial intelligence on the limited, special hardware resources and strict power budget of embedded devices. Note that "equivalent" here mainly

refers to the equivalence of functions. For example, embedded devices can implement image recognition, speech recognition and other functions like general-purpose computers, but their performance can be slightly compromised.

Broadly speaking, we refer to the ability of autonomous analysis and decision-making implemented on embedded devices as embedded artificial intelligence. In a narrow sense, **embedded artificial intelligence is artificial intelligence implemented on embedded devices that is equivalent to artificial intelligence implemented on general-purpose computers.**

In recent years, the concept of **embodied intelligence** has been proposed, which refers to intelligent agents that have a body and support physical interaction, such as home service robots, autonomous vehicles, etc. In contrast, Disembodied AI refers to artificial intelligence that does not have a physical body and can only passively accept data collected and produced by humans. Embodied intelligence is closely related to embedded artificial intelligence. It can be considered as a specific implementation of embedded artificial intelligence. It has more clear and specific requirements for embedded devices, that is, having a body. In this sense, embodied intelligence is a subset of embedded intelligence.

The relationship between artificial intelligence, embedded artificial intelligence, and embodied intelligence is shown in Fig. 1.1.

How to implement artificial intelligence on embedded devices? The main implementation method of embedded artificial intelligence studied in this book is still neural network, especially deep neural network. When they are implemented on embedded devices, we call them embedded neural networks or embedded deep neural networks. In the later chapters of this book, there is no strict distinction between these two terms, and they both refer to embedded deep neural networks.

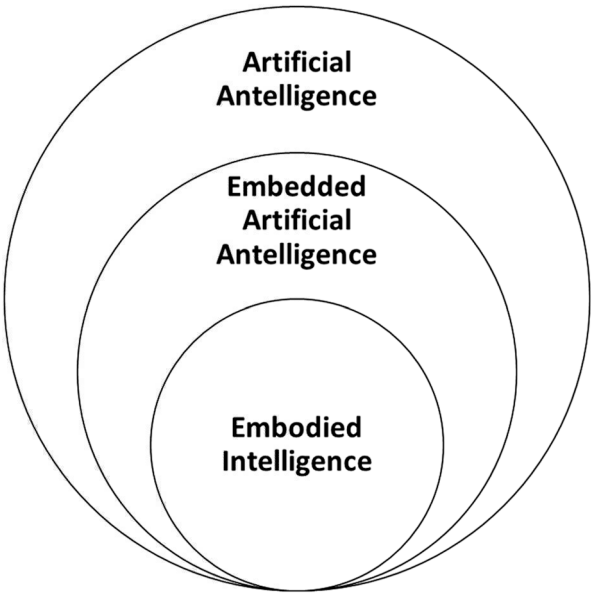

Fig. 1.1 The relationship between artificial intelligence, embedded artificial intelligence, and embodied intelligence

So, what is the difference between neural networks implemented in embedded devices and neural networks implemented on general-purpose computers? Can general-purpose neural networks be deployed directly into embedded devices? Are embedded neural networks just scaled-down versions of general-purpose neural networks? Can embedded neural networks achieve artificial intelligence equivalent to that achieved on general-purpose computers?

We will take these questions into the later chapters of this book. Whatever the answer to the above question, one thing is for sure: Embedded AI is needed.

1.2 Why Do You Need Embedded Artificial Intelligence?

If a general-purpose computer is like a brain, then an embedded computing device is like a complete living body with a brain, sense organs, and limbs. In the past, the brains of embedded devices were not very developed and could only complete some programmed tasks set by humans in advance. But if we give artificial intelligence to machines, let them "live," and let their "minds" independently perceive the environment from the "sense organs, "and command the "limbs" to adapt to the environment and transform the environment, then isn't such a machine the intelligent machine that we have envisioned since the beginning of artificial intelligence that can hear, see, speak, do, and think? Only then can we say that we have achieved true artificial intelligence!

It is this dream that has led scientists and engineers around the world to devote themselves to research and development in the field of embedded artificial intelligence.

From another perspective, embedded devices are everywhere, and they mostly use SoC chips (system-on-a-chip), which integrate microprocessors, memory, I/O interfaces, etc. into a single chip, which is the heart of embedded devices. Their low energy consumption means they can run on batteries for months and require no heat sink, and their simplicity helps reduce the overall cost of the system. More than 10 billion SoC chips are shipped globally every year.

Over the past few decades, the computing power of SoCs has continued to increase. However, in most IoT applications, they simply send data from sensors to the cloud. Therefore, the SoC is idle most of the time. On the other hand, SoCs have made rapid progress this year, and many SoCs have integrated dedicated neural network computing processors, such as NPUs. These SoCs are not just sensors and communicators but can also perform local neural network computations.

Imagine how much computing power is wasted in the real world with tens of billions of such devices deployed in the real world! If we can harness this power and empower tens of billions of edge devices with true intelligence, our world will become a true AI world.

If the above dream can be realized, embedded artificial intelligence will be everywhere. Specialized intelligent machines developed for specific tasks will continue to emerge. Some are simple (such as smart switches), some are complex (such

as autonomous vehicles); some perform a single function (such as license plate recognition cameras), and some can perform multiple functions (such as smartphones); some have only "sense organs" (such as IoT sensors), and some with all sense organs and limbs (such as robots). These intelligent machines will become our assistants at work and partners in life.

Over the past few years, embedded systems have reached a certain level of " intelligence ." More and more smart devices are being released every day and every month. Artificial intelligence on embedded systems has begun to move from very basic forms to complexity and polymorphism, just like the process of biological evolution.

For example, smart home lighting systems automatically turn on and off based on whether someone is in the room. On the surface, this system isn't a big deal. But when you think about it more deeply, you realize that this system is making decisions on its own. Based on the input from the sensor, the SoC decides whether to turn on the light or not. Isn't this a very basic form of AI in embedded systems?

There are also many examples of simple forms of AI in embedded systems. But what about the future? Are we about to have embedded systems with AI that can completely replace human jobs?

Let's look at how AI and embedded systems work together and how they evolve.

1.2.1 Image Identification

Soon, the convergence of artificial intelligence and embedded systems will lead to huge advances in image and video recognition. Advances in embedded technology will help us build imaging devices with higher processing power and smaller footprints. At the same time, AI will provide the much-needed algorithms needed for real-time image and video recognition. The implementation of these smart imaging devices for public safety will be beneficial as it will detect potentially dangerous behavior. Such systems will also be adopted to improve inventory management in factories, monitoring of transportation systems, and the development of industrial automation. For example, license plate recognition cameras have been widely deployed in parking lots. These cameras have embedded image recognition algorithms that can quickly and accurately obtain license plate numbers to complete access control and billing.

1.2.2 Self-Driving

Embedded systems and cars are much closer than you think. Navigation systems, airbag deployment mechanisms, anti-lock braking systems, and many more are based on embedded systems. But bringing AI to cars will be a real game-changer. Self-driving cars have been under development for the past few years and are also

undergoing numerous field trials. Tech giants like Google, Tesla, and Uber are investing billions in research and development with an eye on creating a driverless future. We won't be giving up driving anytime soon though, it could be 10–15 years before you see robot cars cruising the streets. In the process, AI will gradually be introduced into traditional cars, adding more and more autonomous functions. For example, some cars have implemented automatic parking functions to help less skilled drivers park the car into a parking space. Soon, advances in embedded systems will help manufacturers put powerful sensors on boards. This will enable cars to automatically deploy countermeasures, such as automatically braking in emergency situations to avoid traffic accidents.

1.2.3 Dangerous Work

Some of the most dangerous jobs in factories are already taken care of by machines. Thanks to advances in embedded electronics and industrial automation, we have powerful microcontrollers running entire assembly lines in manufacturing plants. However, most of these machines are not fully automated and still require some kind of human intervention. However, now is the time to introduce AI, which can help engineers design truly intelligent machines that can operate with zero human intervention. One such area is the development of bomb-disposal robots. AI-equipped machines can take over tasks such as manufacturing, drilling, and welding of potentially hazardous chemicals.

1.2.4 Internet of Things

The introduction of artificial intelligence will also greatly benefit the Internet of things. We will have smart automation solutions to save energy, improve cost efficiency, and eliminate human error. According to Gartner, more than 20 billion IoT devices will be in use by 2020, and these devices will generate more than 500 ZB of data every year. With more and more technological advancements, this number is expected to continue to grow dramatically. To process the massive data generated on these massive devices is beyond the reach of humans, and artificial intelligence is undoubtedly needed to meet this challenge. In the past, these IoT devices were just embedded devices with built-in sensors and simple controllers to complete some programmed tasks, such as smart light poles that determine their own switches by sensing the intensity of ambient light. In the future, these smart light poles may have "eyes" that dim the light to save energy when no one is passing by at night and illuminate their path when someone is about to pass by. Furthermore, these smart light poles may have the function of human–machine dialogue, introducing information about surrounding shopping malls and restaurants to passers-by, becoming a ubiquitous guide in the city.

1.2.5 Smart Phone

Smartphones, as personal intelligent assistants, have been integrated with artificial intelligence a long time ago. Before the emergence of touch-screen mobile phones, mobile phones that supported voice commands and handwriting recognition had been developed to achieve a more friendly human–computer interaction interface. However, at that time, the traditional pattern recognition algorithm is used, but its recognition rate is not enough to meet the demanding needs of users. Today, smartphones have almost become a new organ of the human body. They carry millions of APPs and realize a variety of functions. The introduction of deep neural network algorithms will achieve better image, voice, text recognition, and other capabilities. This enables more intelligent applications. These scenarios include object recognition, gesture recognition, motion detection, sentiment analysis, natural semantic recognition, music tagging, and more. For example, how to trigger a mobile phone to take a selfie is a problem that has not been completely solved. Touching the screen, selfie stick, Bluetooth remote control, voice commands, etc. will introduce additional actions, making the look less natural. But if the mobile phone can recognize people's expressions, gestures, and body postures, it can trigger the mobile phone at the most appropriate time and capture the most touching moments like a professional videographer.

In addition to these areas, the fusion of artificial intelligence and embedded systems will bring many other opportunities. Such as medical care, logistics, fire protection, agriculture, communications, military, etc.

In the final analysis, the fusion of artificial intelligence and embedded systems will give intelligence to all things, and will allow machines to replace us in completing time-consuming，laborious and even dangerous tasks, thus improving our lives, improving the efficiency of our work, and engaging in tasks that humans are incapable of. Even go to alien planets to open human living space. Embedded artificial intelligence will change the future of mankind!

1.3 Initial Attempt: Cloud Computing Mode

Artificial intelligence has traditionally relied on high-performance computing power provided by server clusters for large-scale, data-intensive model training and inference in the cloud. Due to the significant increase in GPU hardware performance, artificial intelligence, especially deep neural networks, are being applied to more and more business applications, including finance, education, medicine, security, etc. However, the problem with such algorithms is that they are greedy consumers of data and like to complicate the problem. Only with larger data sets and more intensive computing power can more accurate and useful results be obtained.

Therefore, until recently, deep neural networks relied on big data and cloud computing and had to run on energy-intensive servers or even supercomputers with AI

accelerators (such as GPU /TPU). This computer hardware is bulky, energy-hungry, difficult to move, and expensive. However, the current applications of deep neural networks mostly focus on the fields of computer vision and hearing. The problems it solves, such as face recognition, license plate recognition, natural language translation, voice control, etc., often require lightweight, green, energy-saving, and easy to move, low-cost embedded computing devices (including mobile computing devices such as mobile phones) to complete. However, the computing power of these embedded devices is several orders of magnitude lower than that of GPU/TPU, and the memory is also limited. It is obviously not enough to run deep neural networks. So, how to resolve this contradiction?

Many people will naturally think of using cloud computing! Wouldn't it be wonderful to stream data, such as pictures, video streams, and audio to the cloud, and let the powerful cloud computing center complete deep neural network tasks? In the beginning, people did do this. For example, in smart home scenarios, smart hardware only served as sensors and controllers to collect data from the field. The real intelligence was completed by the cloud computing center and issued the results of AI operations (instructions) to control the intelligent hardware to complete the task. This model is the first stage of embedded artificial intelligence and can be called **cloud computing mode.** As shown in Fig. 1.2.

In this mode, the embedded device itself only completes simple data collection, communication, command execution, etc. AI hardware (GPU/TPU), AI algorithms, and AI applications are all deployed in the cloud, and the embedded devices invoke their capabilities through remote interfaces to achieve advanced intelligence.

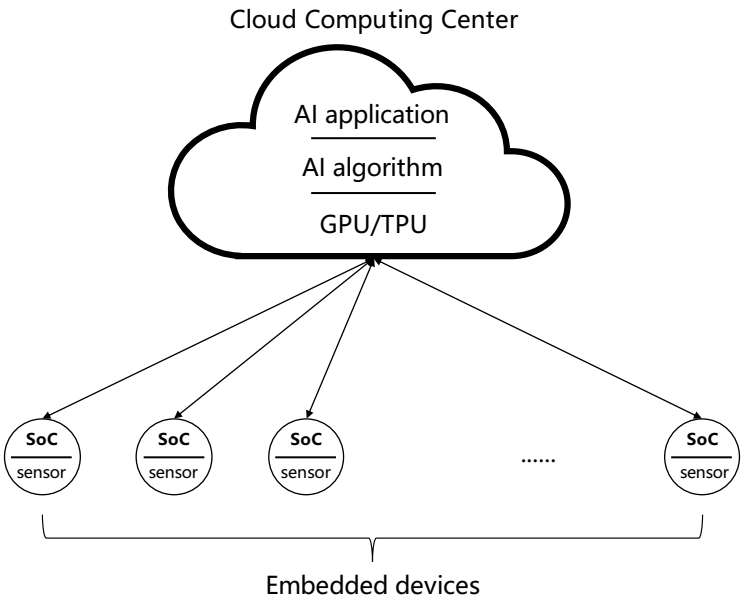

Fig. 1.2 Cloud computing mode of embedded artificial intelligence

This model has obvious advantages:

1. It allows numerous embedded devices to share professional and expensive artificial intelligence hardware (such as GPU/TPU), reducing the cost of a single embedded device.
2. With the help of mature artificial intelligence technology in the cloud, embedded artificial intelligence applications can be quickly developed and deployed.
3. Embedded AI applications are easy to upgrade and maintain because the AI programs are deployed in the cloud.
4. Elastic computing can be achieved. When the current hardware cannot meet the performance requirements, more resources can be applied for.
5. Users can purchase and use artificial intelligence cloud services on demand.

Unfortunately, this model doesn't work in many cases.

Let's start with an accident involving Amazon's Alexa voice assistant.

Alexa is the AI voice assistant on the Echo smart speaker, which is a smart speaker sold by Amazon. The appearance of Echo is no different from ordinary Bluetooth speakers, and it does not have any screen. The only interaction method is voice. Through the Alexa voice assistant, users can play music, query information, and even control various smart home devices through simple voice commands. However, these powerful functions cannot be "fitted" by a small speaker. In fact, Alexa 's AI functions are implemented by the Amazon's cloud computing center. The Echo speaker, as its name suggests, is just a microphone for the cloud computing center.

But when all computing is moved to the cloud, unexpected risks occur.

On March 2, 2018, many people found that Alexa was unresponsive when they tried to command Alexa as usual. The cause of the incident was that Amazon 's cloud service experienced a severe service outage that day. No matter what you said to it, Echo speakers and other Alexa devices would only respond with error messages.

This incident shows that when more and more functions are moved to the cloud, the purchased product is just an empty shell. Although it is usually powerful, once the remote end goes into trouble, the local end will be useless. Behind the advantages of the cloud computing mode, there are some insurmountable flaws. When the cloud computing center or network cannot be accessed, the embedded device will lose its intelligence. Specifically:

1. Cloud computing needs to be accessed through a remote network. Although with the development of wired and wireless broadband technology, the network seems to be everywhere, but it is still not accessible at anytime and anywhere. This is a problem for people who are always mobile or need to enter. For embedded devices in no man's land, dangerous areas, and unfamiliar worlds, you can't always rely on them. For example, for military robots, the wireless network is not only unreliable, but can even be destroyed by enemies. In addition, remote network access will cause delays and jitters, resulting in insufficient real-time response, which is fatal for some critical real-time processing tasks, such as car

driving. Cars must respond to road conditions in a very short period. An additional delay of 10 milliseconds may cause life-threatening consequences. What's more, the delay of the Internet is not fixed, sometimes fast and sometimes slow, and the occasional jitter is not suitable for watching online videos. It doesn't matter, but car driving needs to be foolproof, and every task must be completed within a limited time.

2. The bandwidth of the cloud computing center will also become a bottleneck, especially when many video and audio streams need to be processed simultaneously. The bandwidth of a single channel of video and audio is no longer a problem for network terminals, but when thousands or hundreds of channels of video and audio are aggregated into the cloud computing center, it may cause network congestion. Imagine that there are one million vehicles that use cloud computing to realize license plate recognition when entering and exiting the parking lot. On the one hand, the cloud computing center needs to spend a huge amount of money to purchase bandwidth. On the other hand, due to the periodicity of vehicle parking, during peak hours Network "traffic jams" will lead to real-world traffic jams.
3. Cloud computing will bring the risk of privacy leakage in network transmission, content storage, and other aspects. In some scenarios with high security and privacy requirements, such as smart homes, we hope to use video to analyze unexpected scenes in the home. Remote care is implemented for the elderly and young children, but no one is willing to publish their own videos online and let the Internet monitor them all the time.
4. The total cost of cloud computing services becomes increasingly expensive under long-term and heavy use conditions. Initially, the hardware cost of embedded devices is very low, and the price of cloud services amortized to each device is relatively cheap per month or year. However, after years of use, the accumulated costs begin to exceed the cost savings of embedded devices, and these costs continue throughout the life of the device. Don't forget, many embedded devices are designed to work year-round, such as security cameras. In this way, the cost of the cloud computing mode is not advantageous.

All the above reasons show that in many cases, the cloud computing mode cannot meet the requirements of embedded computing in terms of reliability, economics, and security, which makes it necessary to explore ways to implement artificial intelligence inside embedded devices without (completely) relying on the cloud. Embedded artificial intelligence will enter a new stage.

1.4 From Cloud to Device: Local Mode

The next stage in AI development is to bring deep neural networks from the cloud into the physical world. This is due to the research progress of artificial intelligence in embedded software and hardware in recent years. While initial efforts will

naturally focus on shrinking existing deep neural network models into the limited processor and memory space of embedded devices, future implementations will also be based on the growing processing power of embedded chips as well as AI-accelerated chips developed specifically for AI.

A series of advances in semiconductor process integration and algorithm development, embedded devices (including mobile devices) are gradually getting rid of the shackles of the cloud and can independently perform some "heavyweight" tasks, such as automatic image tagging, biometric identification, and Robotic controls and the ability to perform them repeatedly and efficiently and instantly. This opens the door to embedded artificial intelligence.

Not surprisingly, embedded AI first made a breakthrough on the high-end embedded device: the iPhone. In 2017, the launch of Face ID facial recognition technology marked the beginning of the second phase of embedded artificial intelligence.

Functions such as smart voice assistants and face unlocking have gradually become standard features in consumer devices such as mobile phones and smart watches, indicating that AI will accelerate its penetration into daily life. But what confuses people is that most AI implementations on devices still use the cloud computing mode. These products act like puppets and sounding boards, with the real computing happening behind the scenes on cloud computing servers. Although very convenient, this implementation method violates the user's privacy.

Face ID is a biometric security system powered by a series of sensors and a new AI acceleration chip that uses a front-facing infrared camera to project 30,000 points to create an infrared and three-dimensional image of the user's face. It is accelerated by the Bionic Neural Engine of A11 and above models, which uses a dual-core design to perform up to 600 billion operations per second, enabling real-time processing. The A11 Bionic Neural Engine is not a general-purpose GPU. It is designed for specific neural network algorithms and supports Face ID, Animoji, photo tagging, and Siri voice assistant.

In a paper titled "An On-device Deep Neural Network for Face Detection "(Apple Inc., 2017a, b), Apple researchers describe how to implement the Face ID function based on the A11 Bionic neural network engine.

In 2017, when Apple researchers first started using deep neural networks for face detection in iOS 10, they realized that even the most high-end phones at the time were struggling to run deep neural network algorithms. Like other institutions, Apple had previously been using cloud-based systems for image recognition. To increase user privacy, image recognition algorithms are required to run on the device.

This article describes how Apple works within the confines of limited memory and CPU resources without interrupting other OS tasks and using a lot of extra power. The article details the technical details of how Apple adapts deep neural network models on an SoC-sized GPU. The A11 chip converts three-dimensional and infrared images into a mathematical representation and compares that expression with registered facial data to identify whether the person is using the iPhone. The article concludes

Combined, all these strategies ensure that our users can enjoy local, low-latency, private deep neural network inference without knowing that their phones are running neural networks at hundreds of billions of floating-point operations per second.

In other words, the face recognition function of Face ID is implemented locally on the iPhone, rather than relying on cloud computing as before.

The iPhone also uses deep neural networks to recognize and analyze voice commands for Siri functionality. To do this, the iPhone uses an always-on, low-power auxiliary processor (AOP) to trigger Siri once it hears the user's "Hi, Siri" command, the AOP will wake up the main processor to analyze the user's voice with a more powerful deep neural network. As shown in Fig. 1.3. The benefit of this approach is that it requires minimal processing to listen for and detect the "wake word," saving valuable battery power on your iPhone, and once you wake up, you can take full advantage of the Bionic Neural Engine's powerful processing power. Of course, Siri has not completely escaped the shackles of the cloud, and complex multi-round interactive voice conversations are still handled by the server. This kind of model can be regarded as a device-cloud collaborative embedded artificial intelligence, which will be described in the following chapters.

Application developers can also use the neural network acceleration capabilities of iPhone hardware through the API and development tools of Apple's *Core* ML

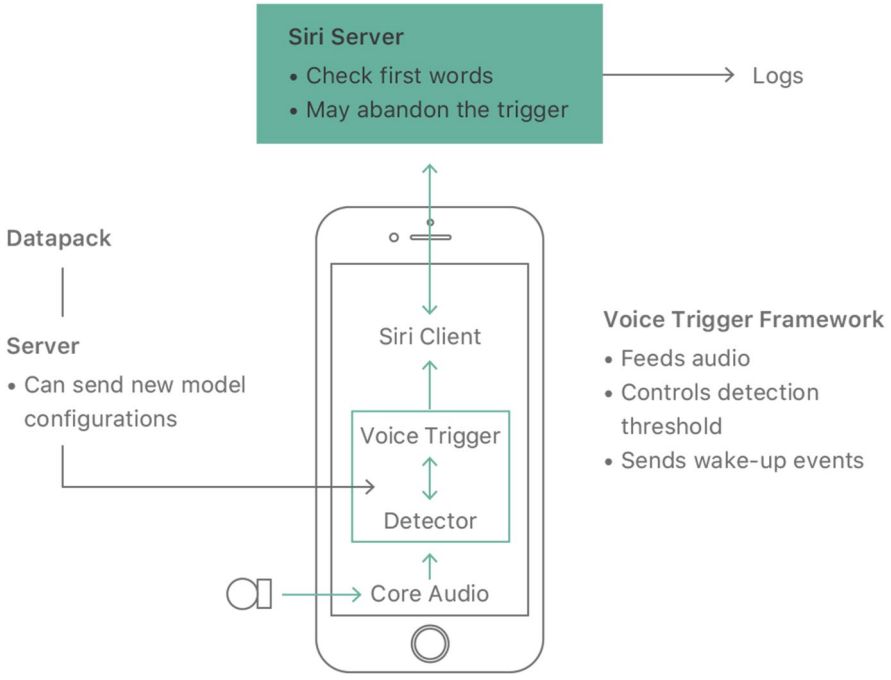

Fig. 1.3 Siri speech recognition process (Source: Apple) (Apple Inc., 2017a, b)

machine learning *framework.* As its tutorial details, the app can perform tasks such as shape recognition and object recognition.

Of course, Apple isn't the only player in this space. ARM, Google, Microsoft, and other companies have also begun to introduce AI into embedded devices.

1.4.1 ARM

ARM is used in most mobile devices and is also the developer of App authorized and customized processor platforms. It introduces AI into its universal SoC design, which will greatly expand the popularity of AI-accelerated devices.

The design, called DynamicIQ, adds processor instructions designed to accelerate machine/deep neural network algorithms, and ARM expects to improve AI performance by 50 times over the next 3–5 years relative to current ARM systems. Some companies are already using low-power ARM-M processors for embedded artificial intelligence applications. For example, the Amiko Respiro, an inhaler for asthma patients, uses data from multiple sensors and onboard machine learning software to calculate the drug's effectiveness and develop therapies customized for each patient.

1.4.2 Google

Not to be outdone, Google launches TensorFlow Lite platform that paves the way for deep neural network algorithms on mobile and embedded devices, TensorFlow Lite is designed to quickly launch TensorFlow models to fit into the small memory spaces of mobile devices and take advantage of any acceleration hardware, like embedded GPUs. The development framework also has interfaces to automatically use hardware accelerators on the device when available.

1.4.3 Microsoft

Microsoft is also developing embedded machine learning software for mobile and IoT devices, including the Raspberry Pi. This research currently focuses on narrow applications in specific scenarios, such as embedded medical devices or smart industrial sensors.

Other companies have also launched their solutions. For example, Reality AI provides a machine learning software library designed for embedded sensors and devices, which allows hardware devices with small physical size and harsh working environments to support more complex and accurate AI model.

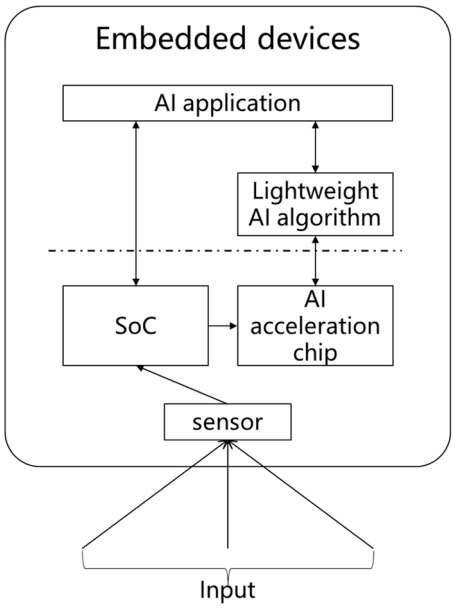

Fig. 1.4 Local mode for embedded artificial intelligence

This series of progress has opened the second stage of embedded artificial intelligence. In this stage, AI hardware, algorithms and applications begin to get rid of the shackles of the cloud and move down to the embedded device itself. We call this local mode for embedded device. As shown in Fig. 1.4:

In this model, at the hardware level, an embedded AI acceleration chip is introduced, which has the characteristics of small size, low-power consumption, and high performance, and is specifically responsible for the inference operations of neural networks. At the software level, lightweight AI algorithms are introduced. These algorithms are improvements to traditional AI algorithms. Under the premise of completing the same function and approximate accuracy, the model has fewer parameters, so it takes up less storage space and has less computational complexity. It is small enough to be "loaded" into an embedded AI acceleration chip. Based on AI acceleration chips and lightweight AI algorithms, AI applications can be implemented locally on embedded devices, processing input signals obtained from sensors nearby, and achieving real-time calculation and response.

1.5 Technical Challenges of Embedded Artificial Intelligence

Although some breakthroughs have been made, embedded artificial intelligence still faces many technical challenges before large-scale application. At present, artificial intelligence is characterized by being computationally intensive, memory intensive, data intensive, and energy intensive. The deployment cost is very high.