HUAWEI

Pre-Trained Model

White Paper

FOREWORD

Gao Wen

Academician, Chinese Academy of Engineering,
Director of Peng Cheng Laboratory

Boya Chair Professor, Peking University

Ever since the term artificial intelligence (AI) was coined at the famous Dartmouth Conference in 1956, scholars and researchers in the field of AI have been working tirelessly on improving the generalization of AI algorithms in the hope of developing a kind of general AI that can be quickly adapted to solve real-world problems in much the same way as humans do, thus improving society's overall efficiency and productivity. In the fifty years that followed, however, efforts were largely unsuccessful. Exquisitely designed models, such as symbolic computation and expert systems, could only be used to solve a limited range of problems, but could not be extended to complex systems like those that use computer vision (CV) and natural language processing (NLP).

It was at the beginning of the 21st century, when significant improvements in hardware performance and big data technology were made, that the situation started to drastically change. Since 2010, deep learning has swept through most domains of AI and achieved unprecedented accuracy on many public datasets. In essence, deep learning is a method of statistical learning. The purpose is to fit a complex function on a large amount of data, so that the function achieves a certain level of generalization. Today, deep learning has achieved great success. A deep neural network, once trained or fine-tuned on sufficiently large datasets, can be adapted to a wide range of different tasks. This was almost unimaginable 20 years ago.

Deep learning also has significant weaknesses. A heavy dependence on big data and large computing power and a high sensitivity to parameter tuning have all made deep learning algorithms difficult for average users to use. To solve this problem, we urgently need new methods that can connect data and domain knowledge and substantially reduce the labor and compute costs of AI. This is why the industry has come up with pre-trained models — large-scale deep neural networks with billions to hundreds of billions of parameters, pre-trained on massive datasets. Through this pre-training process, huge amounts of knowledge are stored in these models. Pre-trained models boast a high level of generalization. They can be adapted to different downstream tasks after simple fine-tuning. In the past five years, pre-trained models have made significant progress in important areas such as natural language processing and computer vision. The models are getting bigger and bigger, and their generalization capability continues to improve. I am delighted to see that Huawei has had many successes in applying pre-trained models across many industries, such as industrial quality inspection, smart traffic control, and fashion design.

There is no doubt that AI has a long way to go before it can ever reach general intelligence. Pre-trained models are approaching the limits of statistical learning methods, but they have problems that are not easy to solve, such as explainability and security. Furthermore, pre-trained models have a far lower energy efficiency than the human brain, which means they are probably not the best way to reach general AI. I believe AI is now at a crossroads. For now, the industry needs to base their choices on pre-trained models and see where they can go from there in the future.

In light of this situation, I believe the Pre-Trained Model White Paper released by Huawei is hugely important in guiding the directions of future AI. I believe that Huawei's continuous investment in the R&D and operationalization of pre-trained models will become a powerful force in exploring the frontiers of AI and inspiring the industry and academia to do the same.

FOREWORD

Zhang Ping'an

Senior Vice President, Huawei

CEO, Huawei Cloud

We're witnessing monumental changes in human history. Digital technologies like AI, big data, the Internet of Things (IoT), cloud computing, and 5G are driving radical transformations across industries and reshaping the world's technology and industry landscapes. AI's strategic vital importance in particular, is widely recognized by most countries around the world.

In 2021, in a speech addressing the Political Bureau of the Central Committee of the Communist Party of China about the digital economy, President Xi Jinping emphasized the importance of the deep integration between AI and the real economy, and the importance of building a digital China and a smart society, promoting digital industrialization and industrial digitization, and developing digital industrial clusters with international competitiveness. In August 2022, the Ministry of Science and Technology and five other ministries jointly issued "Guiding Opinions on Accelerating Scenario Innovation Using High-level AI Applications to Promote High-Quality Economic Development". Today, AI has risen to prominence as the strategic apex in global high-tech, and has become the epicenter for technological competition between countries.

As the cutting edge of AI research, pre-trained models are undoubtedly one of the focal points in this competition. Take natural language processing as an example. The sizes of pre-trained language models increased from hundreds of millions of parameters in 2018 to trillions in 2022, with an increase by an order of magnitude almost every year. Pre-training a large model is a systematic project that relies not only on advanced algorithms, but also the support of best-in-class hardware, framework, and development tools. Huawei has developed full-stack AI capabilities, from Ascend and Kunpeng chips to MindSpore, an AI development framework, and

ModelArts, an AI development pipeline, and on top of these, has launched a series of pre-trained models under the brand name Pangu. Since their launch over a year ago, Pangu models have continued to evolve, and have contributed advanced algorithms and solutions to the fields of computer vision, natural language processing, scientific computing, and more. By the end of 2022, Pangu models had been deployed in over 100 different application scenarios across more than 10 industries, offering new and better options for developers and also creating considerable commercial value. Past projects have shown that Pangu models can minimize manual parameter tuning and human intervention, simplify AI development, while also lowering its costs. This helps promote inclusive AI and allows for large-scale replication of AI across a wide range of industries.

Drawing on our experience in the R&D and operationalization of pre-trained models, a Huawei team wrote this White Paper with the aim of sharing Huawei's insights on pre-trained models with the industry, so that together we can drive the AI industry forward. The road ahead will be challenging, but the future is bright. In its over 60-year history, AI has always been focused on two goals: freeing people from repetitive tasks and expanding the boundaries of human knowledge. As long as we keep these goals in mind and work on combining the strengths of academia and industry, we can continue pushing the frontiers of AI forward for the betterment of all society.

FOREWORD

Gao Xinbo

Professor, President of Chongqing University of Posts and Telecommunications

Winner of China's National Science Fund for Distinguished Young Scholars, Distinguished Professor of the Yangtze River Scholars Program

In Chinese mythology, Pangu was the one that separated heaven and earth and created the world. I think Huawei has chosen a great name for its pre-trained models. Pre-trained models are a promising way to accelerate the development, validation, and iteration of AI models and scale up AI across all industries. Since Huawei Cloud launched Pangu models, they have developed full-stack AI solutions that have achieved successes in many fields. I believe this Pre-Trained Model White Paper will help guide the development of pre-trained models and shed some light on where AI might be heading.

Zhuang Yueting

Professor of Computer Science, Zhejiang University

Winner of China's National Science Fund for Distinguished Young Scholars, Distinguished Professor of the Yangtze River Scholars Program

Director of the AI Collaborative Innovation Center (sponsored by the Ministry of Education of China)

Pre-trained models represent the forefront of recent AI research and have become the focal point of technological competition among nations worldwide. Pre-trained models first made eye-catching breakthroughs in the field of natural language processing, and soon expanded to a wide range of multimodal inference tasks that involve images, video, graphs, and language, and a large number of commercial applications, revealing great potential. I think this Pre-Trained Model White Paper is very necessary as it offers many useful insights on AI, and I think Huawei — a world-leading tech company — is the right one to release this white paper. I believe pre-trained models can be an important means to achieve cross-media intelligence.

Zhang Min

Professor and Special Assistant to the President of the Harbin Institute of Technology, Shenzhen

Director of the Computing and Intelligence Research Institute, winner of the National Outstanding Youth Fund

Pre-trained models are commonly referred to as the infrastructure of AI applications. They have powerful capabilities in knowledge modeling, knowledge acquisition, and generalization. This Pre-Trained Model White Paper gives a brief introduction to Huawei Cloud's Pangu models, including CV, NLP & speech recognition, multimodal, scientific computing, and graph neural network. It covers their unique strengths and technical principles, and shares some of Huawei's insights about the present and future of AI. I believe this white paper will be quite useful for anyone involved in the AI industry.

Li Houqiang

Professor, IEEE Fellow, Vice Dean of the School of Information Science and Technology at the University of Science and Technology of China
Winner of China's National Science Fund for Distinguished Young Scholars, Distinguished Professor of the Yangtze River Scholars Program

Pre-trained models are the new frontier of AI research. Over the past few years, they have achieved many significant successes in the fields of natural language processing and computer vision. This Pre-Trained Model White Paper, written by a team from Huawei Cloud, covers the theory, methodology, technology, and applications of pre-trained models. It also includes some valuable lessons Huawei Cloud has learned from developing and operationalizing pre-trained models.

Xiong Hongkai

Chair Professor, Shanghai Jiao Tong University
Winner of China's National Science Fund for Distinguished Young Scholars, Distinguished Professor of the Yangtze River Scholars Program

Albert Einstein once said: "The most incomprehensible thing about the Universe is that it is comprehensible." In the 21st century, big data and AI have become two important technologies that allow us humans to better understand both the world and ourselves. Over the past few years, pre-trained models have been recognized as a promising way to achieve general intelligence. Pre-trained models can be fine-tuned and adapted to a wide range of downstream tasks. Huge amounts of general and domain-specific knowledge are stored in their huge networks, allowing for a high generalization capability. The benefits of pre-trained models can extend across industries and even the borders of different nations and serve the well-being of the entire human race. I believe Huawei's White Paper on large pre-trained models can offer some insights on how to promote inclusive AI and even a more inclusive and equal society.

Jiang Yugang

Professor and Ph.D Supervisor, Fudan University
Distinguished Professor of the Yangtze River Scholars Program, the Ministry of Education; HR Director, Fudan University

Over the past few years, we have seen ultra-large-scale pre-trained models like GPT-3 and CLIP being launched. Trained through self-supervised learning on massive amounts of data, these models can be quickly adapted to a wide range of downstream tasks in natural language processing, computer vision, and more domains. Huawei Cloud has deep expertise in pre-trained models. In 2021, Huawei Cloud launched a number of pre-trained models belonging to the Pangu series, which attracted wide attention from both academia and industry professionals. This Pre-Trained Model White Paper offers some opinions about where pre-trained models are heading and what opportunities they might bring, and could prove useful to those working in AI.

PREFACE

Making Pre-Trained Models the Operating System of AI

AI has entered into enterprises' core production systems and has started to create greater value.

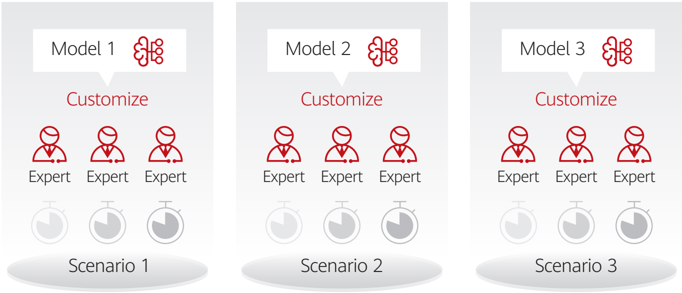
"By 2025, more than 86% of businesses and organizations will use some form of AI."

Huawei Cloud has 600+ AI projects, 30% of which are in core production systems, but:

AI adoption is still faced with many challenges, including:

1. Fragmented use cases, isolated AI development, and difficulty in reusing models and scaling up AI.
2. Difficulty in combining industry knowledge with AI technology.
3. Concerns over the risk of attacks, privacy, and security of AI models.

Models developed for specific scenarios cannot be used elsewhere. New models need to be developed from scratch for each new task.

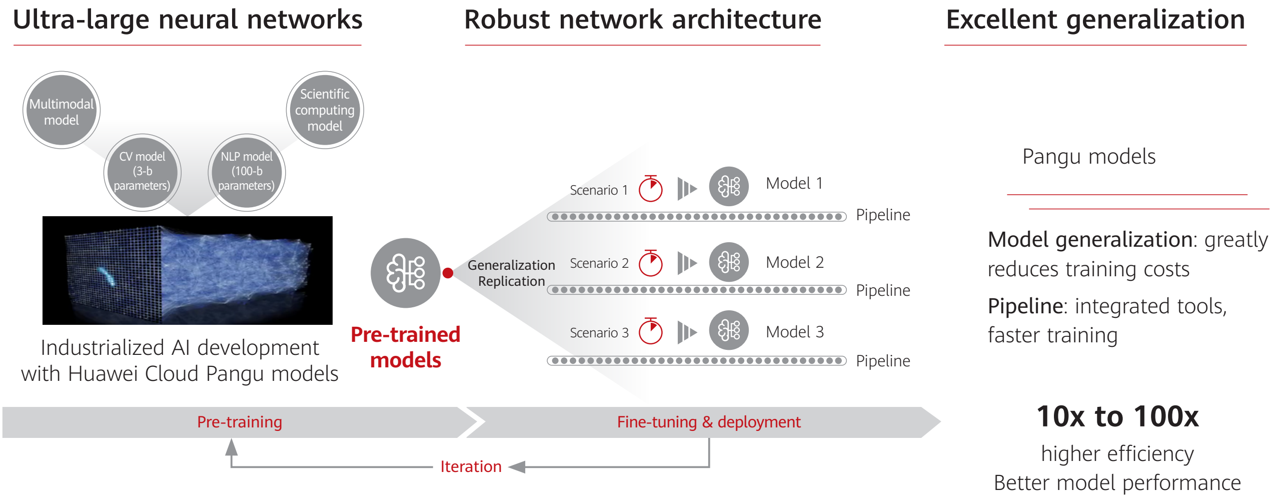


High cost, high barriers, always starting from scratch

If people's understanding of electromagnetics remained at the level of Faraday's law of induction and Maxwell's theory of the electromagnetic field was never published, the electrical revolution would never have happened. If wind, rain, thunder, or even changes in temperature could lead to power outages, it's hard to imagine electricity becoming an essential commodity and infrastructure that powers modern life.

China has seen fast growth in its cloud computing market in recent years, with an increasing number of businesses and organizations choosing to go cloud. Demand is shifting from resources to services and intelligent applications, meaning that PaaS and SaaS are gaining more traction over time. Although the potential for growth is great, the enormous diversity of use scenarios and cloud services present significant challenges for cloud service providers (CSPs). As the cloud market matures, many customers are now opting for smaller CSPs that can deliver customized services, over large CSPs that mostly provide standard, one-size-fits-all services. This means even though the cloud market is largely dominated by a small number of players, small CSPs will always have their place in the market. Major CSPs must work hard to increase or keep their current market share while ensuring high efficiency, cost-effectiveness, and quality when offering such a huge portfolio of services.

The situation faced by AI is quite different. Today, traditional industries are seeking ways to help them free workers from repetitive, labor-intensive tasks and improve productivity. This means AI algorithms will need to accommodate hugely diversified needs in vastly different application scenarios. This is why generalization is so important to AI models. Generalization refers to a model's ability to adapt to a wide range of downstream tasks. Today, many AI models in use are developed in isolated workshops. For each use scenario, a new model needs to developed, trained, tuned, and iterated independently. This method is usually quite inefficient. Many developers also do not have the skills needed to develop and tune high-quality models in terms of accuracy, performance, and scalability. These are important reasons why it is currently difficult to scale up AI, especially in traditional industrial sectors.


Ultra-large neural networks
Robust network architecture
Excellent generalization
Multimodal model
Scientific computing model
CV model (3-b parameters)
NLP model (100-b parameters)
Industrialized AI development with Huawei Cloud Pangu models
Pre-trained models
Generalization Replication
Scenario 1
Model 1
Pipeline
Scenario 2
Model 2
Pipeline
Scenario 3
Model 3
Pipeline
Pangu models
Model generalization: greatly reduces training costs
Pipeline: integrated tools, faster training
10x to 100x
higher efficiency
Better model performance
Pre-training
Fine-tuning & deployment
Iteration

At present, we believe pre-trained models (or foundation models — a term that is being used by more and more people in the AI industry) are the solution to the problem we described above. A pre-trained model is a large deep learning model pre-trained on large datasets (e.g. images and text) using unsupervised or self-supervised learning methods to extract knowledge from data, with the knowledge then being stored in the model's large number of parameters. For any new task, a pre-trained model can be deployed to release the knowledge embedded in it. That knowledge can then be combined with industry knowledge and know-how to solve related problems. In recent years, there has been an abundance of research and applications related to pre-trained models, establishing them as a dominant trend in AI. However, we should also acknowledge that pre-trained models still have a long way to go before they are ready for large-scale commercial use, as it will depend not only on the advancement of technology but also on how business models evolve. We see a possible future where pre-trained models will become the operating system of AI, where they will manage AI hardware and support AI algorithms, and standardize and democratize AI. We hope that through this white paper we can share our experiences with you, along with the lessons we learned from developing and operationalizing pre-trained models, so that we can drive the industry forward together.

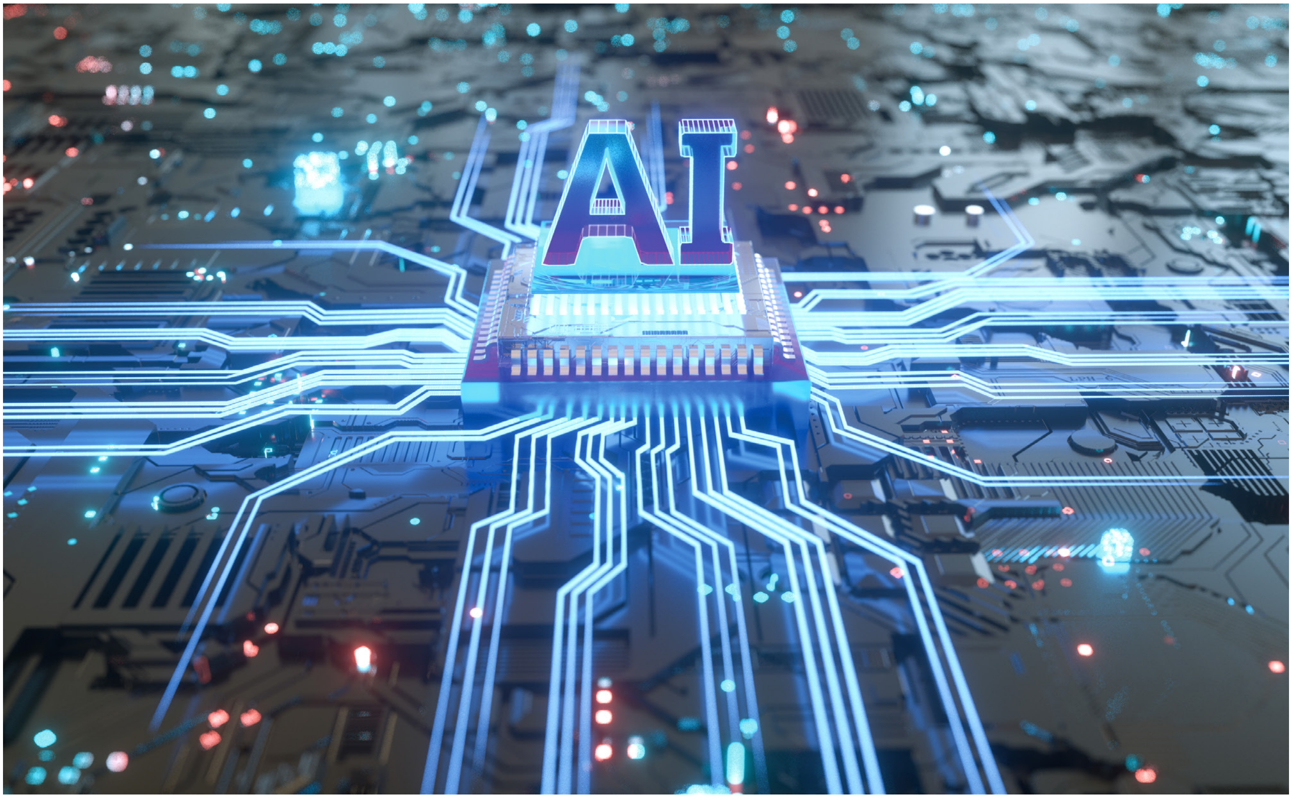
CONTENTS

01

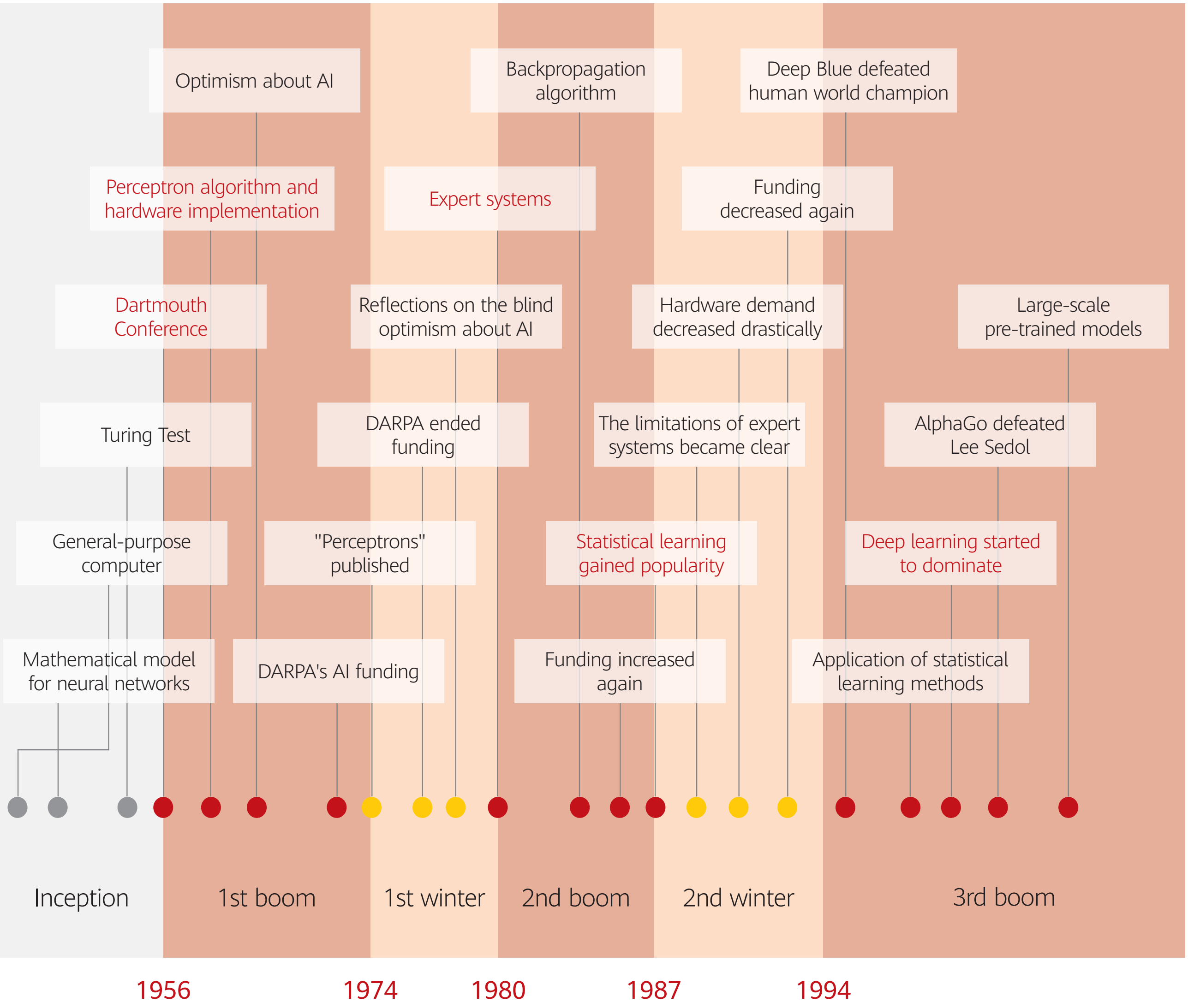
Pre-Trained Models Are the Near Future of AI

1.1 A Brief History of AI

AI has a history of over 60 years, starting from when the term was coined at the famous Dartmouth Conference in 1956. Since the very beginning, AI research has been divided into three schools of thought: logical reasoning, statistical learning, and brain-like computing. Logical reasoning has obvious limitations, and it is difficult to use it to model complex problems. Brain-like computing relies heavily on future advancements in life sciences, especially brain science. At the turn of the 21st century, supported by big data and large computing power, statistical learning methods have come to dominate the field of AI, and they have spawned methodologies and applications that are changing society in profound ways.



A Brief History of AI


Optimism about AI
Backpropagation algorithm
Deep Blue defeated human world champion
Perceptron algorithm and hardware implementation
Expert systems
Funding decreased again
Dartmouth Conference
Reflections on the blind optimism about AI
Hardware demand decreased drastically
Large-scale pre-trained models
Turing Test
DARPA ended funding
The limitations of expert systems became clear
AlphaGo defeated Lee Sedol
General-purpose computer
"Perceptrons" published
Statistical learning gained popularity
Deep learning started to dominate
Mathematical model for neural networks
DARPA's AI funding
Funding increased again
Application of statistical learning methods
Inception
1st boom
1st winter
2nd boom
2nd winter
3rd boom
1956
1974
1980
1987
1994

The following are the ups and downs of AI throughout its history:

1943–1956

Inception and birth

Milestone accomplishments in this period include the first mathematical model of a neural network created by Warren S. McCulloch and Walter Pitts; the Turing Test, proposed by Alan M. Turing, that tests a machine's ability to exhibit intelligent behaviour equivalent to that of a human; and the general-purpose digital computer ENIAC built in 1946, which provides hardware support for the complex computation needed by AI.

1956–1974

First boom

The 1956 Dartmouth Conference marks the beginning of the first boom of AI. AI algorithms based on logical reasoning were used to solve problems in some specific domains (for example, to prove mathematical theorems). The perceptron algorithm based on a subsymbolic system was also developed. In particular, in 1957, a computer that simulates perceptrons, called Mark I, was invented. Inspired by these, many scholars at the time, including Marvin L. Minsky (the 1969 Turing Award winner) and Herbert A. Simon (the 1975 Turing Award winner), began to develop unrealistic optimism about AI. They proclaimed that "machines will be capable, within 20 years, of doing any work a man can do". Government agencies like DARPA and large enterprises also funded many AI research projects.

1974–1980

First winter

Researchers quickly realized the limitations of the first-generation AI algorithms. In 1969, Marvin L. Minsky published Perceptrons, which almost single-handedly destroyed connectionism (i.e. artificial neural networks). In the meantime, it was made clear that algorithms based on logical reasoning will need a lot more time to become able to solve most real-world problems. As organizations like DARPA pulled most of their funding, the AI industry began a wave of reflection and introspection. Typical examples include Artificial Intelligence: A General Survey, commonly known as the Lighthill report, an article published by James Lighthill in 1973, and John R. Searle's Chinese Room Argument published in 1980.

1980–1987

Second, short boom

1980-1987. Expert systems that emerged in the early 1980s led to a short boom in AI, and people started to use AI algorithms to solve practical problems in a limited range of domains. The MYCIN algorithm developed in 1975 was able to assist diagnosis of blood infections. In the meantime, new neural networks like the Hopfield neural network and the backpropagation algorithm invented by David E. Rumelhart greatly expanded the application scope of artificial neural networks. In 1989, Yann LeCun (the 2018 Turing Award winner) developed a five-layer neural network to recognize handwritten digits. In the 1990s, this network was able to recognize more than 10% of handwritten checks in the United States. Organizations like DARPA started to take interest again. Compared with the early 1980s, investments in AI in the late 1980s increased many fold.

1987–1993

Second winter

The early success of expert systems soon ran out of steam. Researchers found that even in a limited range of domains, when facing unknown or undefined problems, no matter how simple these problems were, expert systems were unable to deliver reliable and predictable performance. The funding for AI research began to dry out again. In response, researchers gradually shifted from symbolic methods (such as deductive reasoning) to subsymbolic methods (such as statistical learning). During this period, researchers began to realize the importance of perception and interaction. Examples include the visual perception model proposed by David Marr in his book Vision, where he proposed the idea that it is "better to use the world as its own model".

1993-present

Third boom

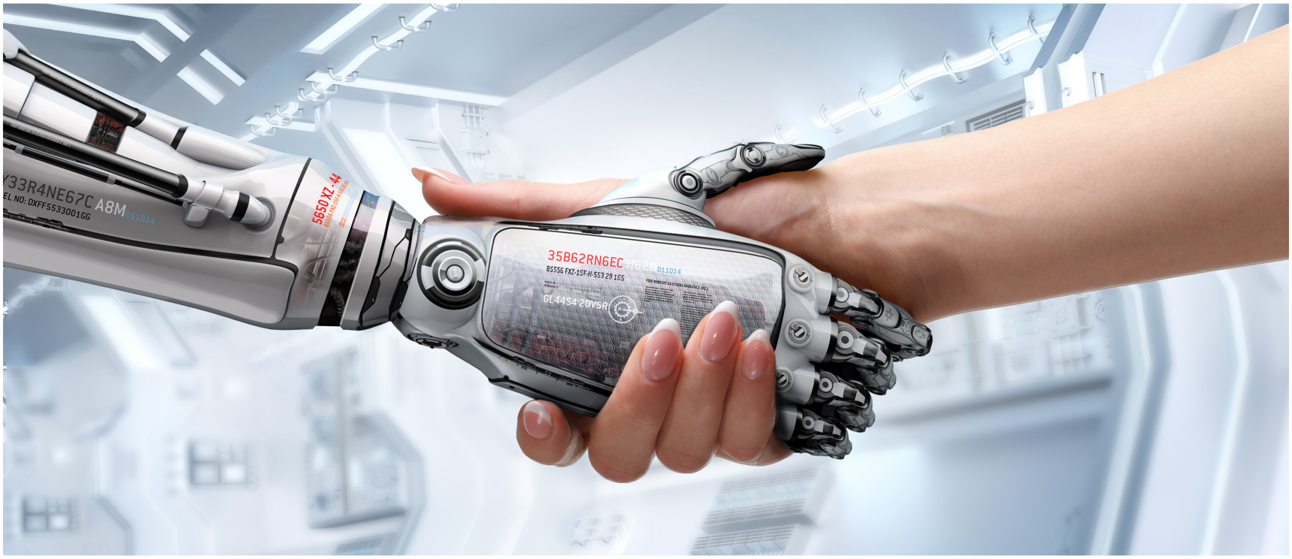
As modern computers are becoming increasingly powerful in terms of storage and compute capacities, statistical learning methods have gradually become dominant in the field of AI.

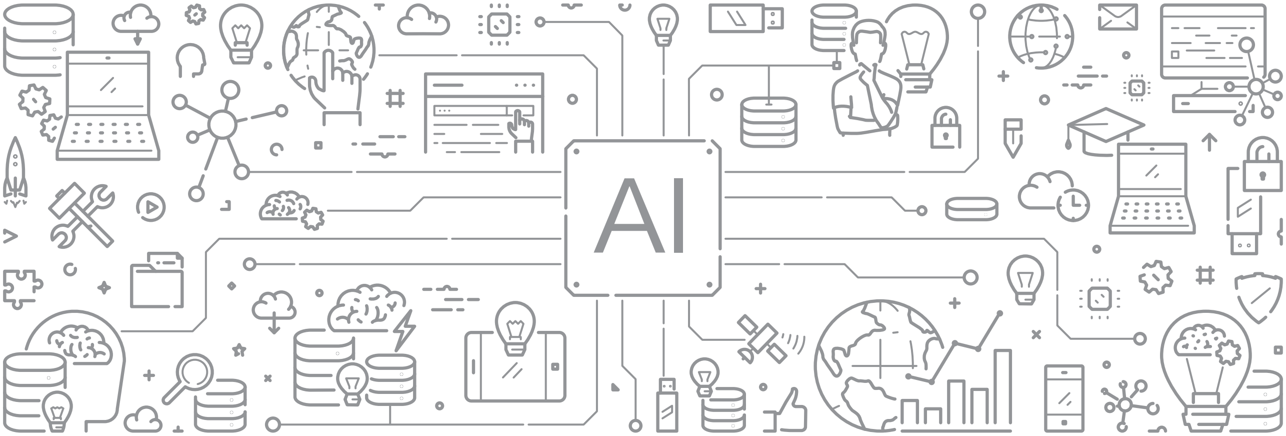
In many domains of AI, such as computer vision, speech recognition, and natural language processing, hand-designed models have been replaced by statistical learning-based models. Since 2011, deep learning has swept through most domains of AI and has surpassed humans in many domains. The third boom of AI, which is also the longest in its over 60-year history, is still ongoing and has shown no sign of ending just yet. Although there are many essential problems that are not yet resolved, many AI applications have forever changed the human society.

It is worth stressing that deep learning has not solved the essential problems found within AI. In the future, the industry is likely to experience many ups and downs before we can reach genuine general AI — if we ever reach it at all. But before that, despite frequent discussions about strong and weak AI or concerns about what will happen after we reach technological singularity, we have been largely focused on developing smarter AI, without being distracted or deterred by them.

1.2 A Prediction About the Near Future of AI

The three most influential schools — logical reasoning, statistical learning, and brain-like computing, have existed since the very beginning of AI, and none have completely disappeared. The three schools of AI each have their own strengths and weaknesses. Brain-like computing has the loftiest goals, but before the needed advancements in life sciences happen, it is unlikely to have plausible applications. Logical reasoning mimics the way the human brain reaches conclusions, and it has good explainability. Because logical reasoning methods do not need large datasets or high computing power, they became the dominant methods behind the first two booms of AI. As our understanding about the complexity of AI increases and the limitations of logical reasoning methods became clearer, logical reasoning methods gradually gave way to statistical learning methods during the third boom of AI, the latter of which are becoming even more dominant after deep learning occurred.





It is worth emphasizing that deep learning could not have taken off on its own. Without the support of big data (rapid increases in storage capacity and the mobile Internet) and large computing power (especially the rapid evolution of GPUs), deep learning could not have dominated most domains of AI in just three to five years. The bigger a deep learning model gets (having more parameters), the larger datasets it will need. To enable large models to learn features more efficiently from large datasets, researchers have come up with the ideas of hierarchical modeling and distributed representation to improve the efficiency and precision of data matching. From a technical perspective, the core of deep learning is in deep neural networks: a general backbone network is adapted to different downstream tasks across different sub-domains. For example, in computer vision, deep neural networks that use very similar structures have become the general framework for a range of different tasks, such as image classification, object detection, instance segmentation, and pose estimation. In natural language processing, a model called Transformer is also widely used, enabling researchers to build general language models.

In essence, however, deep learning is still statistical learning, as it is still primarily about feature extraction and pattern matching. This approach is undoubtedly inefficient compared to the human brain's knowledge-based inference. As the demand for intelligent applications increases in many industries, this approach will limit the adaptability of AI algorithms. This is because for any new task or new entities, algorithms will need dedicated training data in order to make accurate predictions. Developers must always develop new models from scratch, including data collection, model training, tuning, deployment, and iteration. For most AI developers, this is undoubtedly a challenging task, and the costs are high, making it difficult to scale up AI across different industries, especially for small and medium enterprises.

Pre-trained models are a promising way to solve these problems. A pre-trained model consists of two parts: the upstream (pre-training) and the downstream (fine-tuning). In the upstream, large amounts of data are collected and used to train a large-scale neural network, enabling it to efficiently store and understand data. In the downstream, the pre-trained model is fine-tuned using relatively smaller datasets and computing power to adapt it to specific downstream tasks. We will describe the process of pre-training large models in more detail in Chapter 2. Although it seems that pre-trained models are unlikely to lead us towards artificial general intelligence, we can still reach two very plausible conclusions:

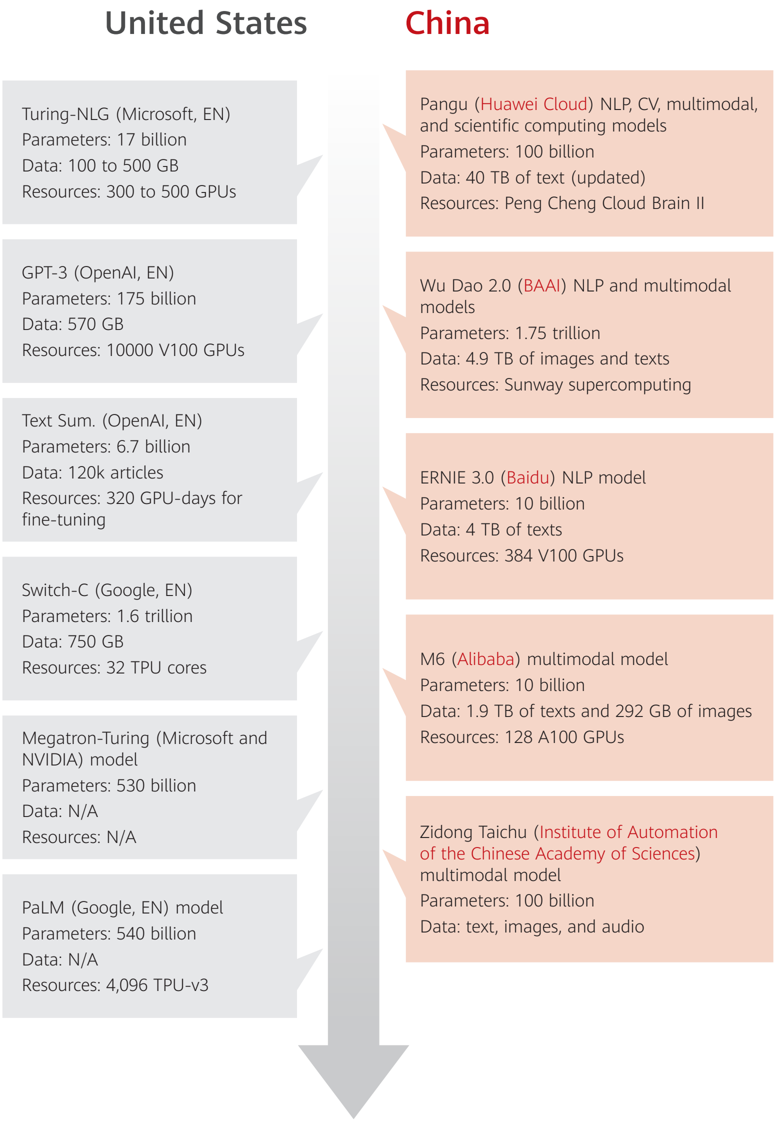
Until the next revolutionary computation model materializes, pre-trained models are the most effective and efficient way to scale up and democratize AI, with great commercial potential.

According to past projects, compared with the conventional ways of AI development where all models are developed from scratch, CV and NLP models fine-tuned from large pre-trained models deliver significantly higher accuracy, with significantly lower costs in data collection, training, and computation, and they are much easier to develop, train, and deploy. Take CV as an example. Traditionally, to train a basic object detection model on 100 images, one developer needs to work for a whole week and use eight GPUs over five hours, while fine-tuning such a model from a pre-trained CV model takes only two hours with one GPU and no human intervention. In this example, the pre-trained model reduced development costs to just 10% or even 1% of the original cost.

The study of pre-trained models will likely inspire the discovery of the next general computation model

Looking back, 2011 was the peak of traditional statistical learning methods. In the field of CV, bag-of-words (BOW) models even reached one billion parameters. Even in 2021, one billion parameters were enough to put a CV model in the category of large models. In 2012, however, a deep neural network with only 60 million parameters beat billion-parameter BOW models, and deep neural networks have since become the dominant network structure for CV. Compared with BOW models, deep neural networks excel in feature matching efficiency. We have reason to believe that at some point, large models may evolve to some new revolutionary forms and take statistical learning methods to the next level. Judging from where we stand, we believe the breakthrough is likely to come from the marriage of large models and knowledge computing.

To sum up, we think pre-trained models are currently the epitome of AI and deep learning as well as the highest achievement of statistical learning to date. Until a new generation of technologies emerge, they will be our most powerful tool for AI research and development.



United States

Turing-NLG (Microsoft, EN)
Parameters: 17 billion
Data: 100 to 500 GB
Resources: 300 to 500 GPUs

GPT-3 (OpenAI, EN)
Parameters: 175 billion
Data: 570 GB
Resources: 10000 V100 GPUs

Text Sum. (OpenAI, EN)
Parameters: 6.7 billion
Data: 120k articles
Resources: 320 GPU-days for fine-tuning

Switch-C (Google, EN)
Parameters: 1.6 trillion
Data: 750 GB
Resources: 32 TPU cores

Megatron-Turing (Microsoft and NVIDIA) model
Parameters: 530 billion
Data: N/A
Resources: N/A

PaLM (Google, EN) model
Parameters: 540 billion
Data: N/A
Resources: 4,096 TPU-v3

China

Pangu (Huawei Cloud) NLP, CV, multimodal, and scientific computing models
Parameters: 100 billion
Data: 40 TB of text (updated)
Resources: Peng Cheng Cloud Brain II

Wu Dao 2.0 (BAAI) NLP and multimodal models
Parameters: 1.75 trillion
Data: 4.9 TB of images and texts
Resources: Sunway supercomputing

ERNIE 3.0 (Baidu) NLP model
Parameters: 10 billion
Data: 4 TB of texts
Resources: 384 V100 GPUs

M6 (Alibaba) multimodal model
Parameters: 10 billion
Data: 1.9 TB of texts and 292 GB of images
Resources: 128 A100 GPUs

Zidong Taichu (Institute of Automation of the Chinese Academy of Sciences) multimodal model
Parameters: 100 billion
Data: text, images, and audio

Large, pre-trained models are at the center of the AI competition between the world's great powers

Huawei's Pangu models are leading in **commercial applications**.

1 Pangu models + workflows, continuous learning while being used, low consumption of compute power, easy to use.

2 Lightweight development powered by Huawei's ModelArts AI development pipeline.

3 Fast adaptation to a wide range of industry-specific scenarios and tasks.

02

An Introduction to the Pangu Model Family

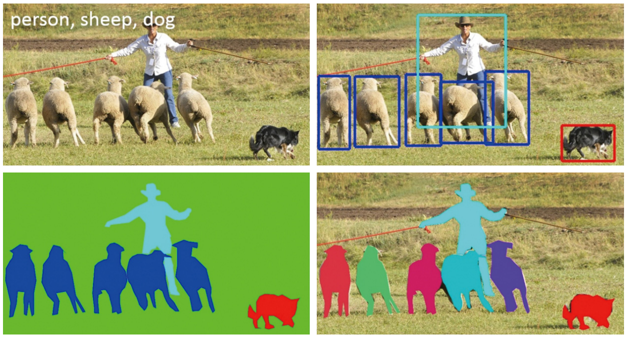
In light of the critical importance of pre-trained models, Huawei Cloud initiated its pre-trained model project in 2020 and launched a number of pre-trained models under the brand name Pangu for the first time in April 2021. Pangu models are the culmination of achievements Huawei Cloud has made so far in dozens of areas of AI and are powered by Huawei's full-stack AI solutions. They are deeply integrated with Huawei's Ascend-series processors, MindSpore AI architecture, and the AI development platform ModelArts. This chapter gives a brief introduction to the pre-trained models of the Pangu family and offers some insights into the key technologies of pre-trained models.

2.1 Computer Vision

Computer vision is about designing programs to automatically acquire, extract, process, and analyze visual signals, and to derive a high level of understanding from them. Put simply, computer vision is a discipline that studies how to teach computers to "see". Typical computer vision tasks include image classification, object detection, object segmentation, object tracking, and pose estimation. The following figure shows the famous ImageNet dataset (with over 20,000 object categories) and MS-COCO dataset (supporting different types of tasks, such as object detection and segmentation) for image classification.

The ImageNet dataset

~15M images, ~21K categories, ~1.5TB



The MS-COCO dataset

detection, segmentation, pose estimation, etc.

In computer systems, visual signals are usually stored as "densely sampled intensities": the intensities of light rays coming in different directions on each channel (for example, red, green, and blue) are recorded and are used to generate a high-level representation of an image. Each basic unit in an image is called a pixel. Obviously, these pixels by themselves cannot represent any semantic information. Hence, there is a big gap between the way images are stored digitally and the way semantics are understood by humans. In academia, this gap is called the "semantic gap", which is a core problem that almost all computer vision tasks must deal with.

Further exploring the storage formats of images, we can find several characteristics of digital image signals:

Complex content

Digital images are represented by pixels, but each pixel alone cannot express semantics. The task of image recognition is to build specific functions to output semantics from input pixels. Such functions are often very complex and difficult to define manually.

Low information density

Digital image signals can faithfully represent an object. The problem is that a large proportion of the data is used to represent low-frequency areas (such as the sky) or high-frequency areas without clear semantics (such as random noises). This means image signals have a low information density, especially when compared to text signals.

Volatile domains

The semantics derived from image signals are affected by their domains. For example, the same semantics can have distinctively different representations under different intensities of light. Furthermore, the same object can appear in different sizes, angles of view, and poses, leading to hugely different pixels. This creates challenges for image recognition algorithms.

In light of these characteristics, we believe large pre-trained models based on deep neural networks are one of the best ways to develop and deploy computer vision algorithms. The pre-training process is in a way the process of compressing visual signals. A deep neural network can extract visual features hierarchically, and pre-training combined with fine tuning can help the network adapt to different domains.

Now, let's take a brief look at the general principle and technical solution of the Pangu CV model.

2.1.1 Data Collection

Images are complex unstructured data that contain rich semantic information. Presently there are no good ways to accurately describe the mathematical patterns of image data, so all we can do is to collect large amounts of data to approximate real-world images. The ImageNet dataset first published in 2009 is an important milestone in the field of computer vision. It makes it possible to train and evaluate large-scale image processing methods. With the advances made in computer vision technology and the emergence of new applications, the limitations of ImageNet datasets in terms of the scale and complexity began to show. To solve this problem, we need image datasets that are larger and more complex than ImageNet.

We collect image data using many different methods, including downloading public datasets, expanding in-house developed datasets, search engine crawling, reverse image search, and image extraction from video. We also filter out low-quality image data, such as those with low resolution, underexposure or overexposure, or simple background, and then we use the pre-trained CV model to identify and delete duplicate images. Finally, we have developed a dataset with over 1 billion high-quality images and a total size of approximately 40 TB.

1+ billion

images

~40 TB

storage space

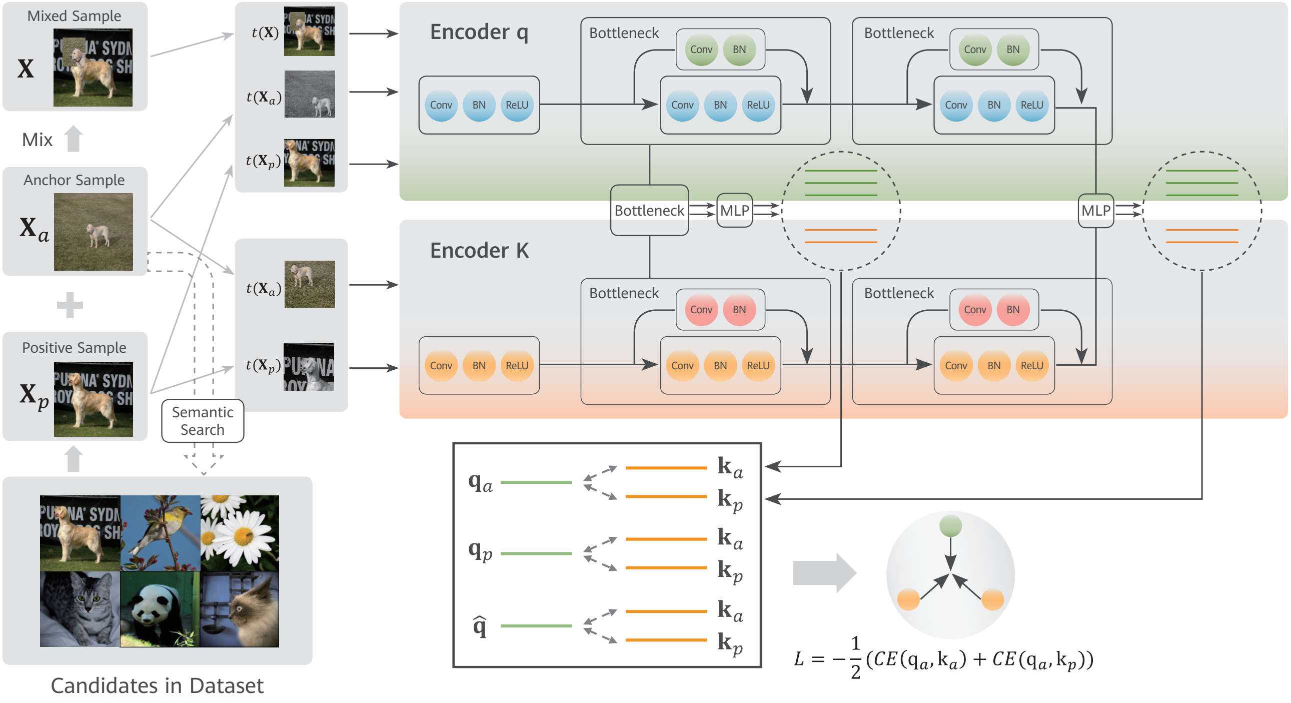
Wide application

Wide application
Autonomous driving,
electric power, railway,
remote sensing

2.1.2 Pre-Training Method

The neural network models we use include the most commonly used convolutional networks and Transformer architectures in the field of CV. They can be used separately or together to achieve optimal results. With automated machine learning algorithms, we can support and invoke neural networks of different sizes — nearly 3 billion parameters for the largest model or only hundreds of thousands of parameters for the smallest. This allows us to quickly adapt the models to different downstream CV tasks.

Most of the training data we have collected comes from the Internet and may have a high level of noise and inaccurate or no semantic labels. To fully utilize the training data, we used self-supervised learning methods. That is, we use one or several types of proxy tasks to teach models how to understand visual data so that they can fit to complex data without any semantic labels. In particular, we optimized some proxy algorithms based on contrastive learning. We were the first to use hierarchical semantic similarity in contrastive self-supervised learning. That is, we select the neighbors nearest to the clustering centers as positive samples, and we use hybrid sample enhancement when gathering semantically similar samples. This way we reduce the impact of noise during sample selection. On top of this, we expand the number of positive samples for self-supervised learning algorithms, so that positive samples can be aggregated more efficiently and the impact of negative samples can be mitigated. The following is a simple illustration of the pre-training algorithm we use (published in TPAMI).



(Note: Contrastive self-supervised learning based on hierarchical semantic aggregation)

2.1.3 Performance

The Pangu CV model achieved results comparable to those of fully supervised learning models in linear classification tasks based on ImageNet datasets. Thanks to better semantic alignment, our method also performs well in few shot learning. Trained on ImageNet - 1% and 10% labeled data, our method achieved 66.7% and 75.1% accuracy in image classification tasks, respectively, both surpassing the results of other models by large margins. Based on this method, we designed a large model with 1 billion parameters and pre-trained it on a dataset consisting of over 1 billion unlabeled images. This model achieved 88.7% classification accuracy on ImageNet, and the accuracy of semi-supervised classification on 1% labeled data also reached 83.0%. Furthermore, the Pangu CV model achieved good generalization performance in over 20 downstream tasks, as shown in the tables below.

	Dataset	Benchmark model	Pangu pre-trained model
1	Aircraft (aircrafts)	90.43	89.32
2	CUB-200-2011 (birds)	86.90	91.80
3	DTD (texture)	80.05	85.00
4	EuroSAT (satellite images)	98.85	98.98
5	Flowers102 (flowers)	97.07	99.69
6	Food101 (food)	92.21	94.58
7	Pets	95.29	95.91
8	SUN397 (scenes)	71.51	78.92
9	Stanford Cars (cars)	92.48	94.09
10	Stanford Dogs (dogs)	87.41	91.28
11	Average	89.22	91.96

Pangu CV model: classification performance

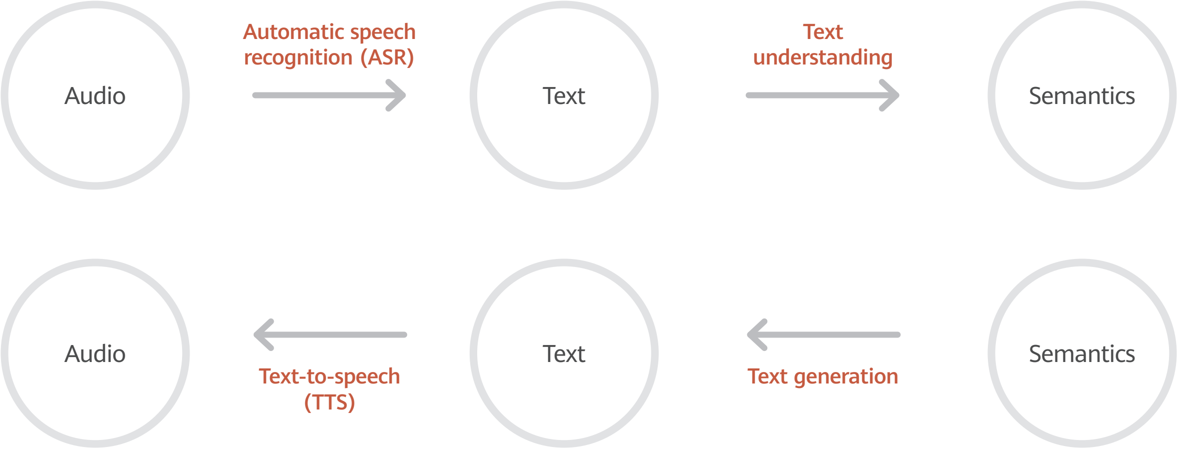
	Dataset	Benchmark model	Pangu pre-trained model
1	VOC (natural scenes)	72.2	76.6
2	Comic (style transfer)	35.6	38.0
3	Clipart (style transfer)	57.5	61.0
4	Watercolor (style transfer)	34.4	36.9
5	DeepLesion (healthcare)	36.7	38.1
6	Dota 2.0 (remote sensing)	21.2	21.0
7	Kitti (autonomous driving)	29.6	32.9
8	Wider Face (human faces)	35.3	36.3
9	LISA (traffic lights)	43.5	42.7
10	Kitchen (kitchen scenes)	53.6	55.0
	average	41.96	43.85

Pangu CV model: object detection performance

2.2 Natural Language Processing and Speech Recognition

A natural language refers to any human language that has evolved naturally among a community. Communication in a natural language between humans can be either written or spoken. This means natural language understanding and generation may be divided into two categories: text and speech. In the field of AI, these two domains are called natural language processing and speech processing, respectively. Similar to CV, the goal of natural language and speech processing is to enable machines to understand and use text and speech in the same way humans do and to effectively communicate with humans or other intelligent bodies.

As shown in the figure below, both natural language and speech processing can be divided into two parts: understanding and generation. The goal of understanding is to enable machines to understand the meanings of human languages, and the goal of generation is to enable machines to express themselves using human languages. The difference between natural language processing and speech processing lies in that the former deals with text while the latter speech signals. In most cases, text and speech signals are closely correlated, but in some cases, there are things that can be much better expressed by one but not the other (for example, it is difficult to express music in text).



One of the core problems of natural language and speech processing is in expressing text and speech in forms that can be easily understood and processed by machines. Before deep learning, researchers mostly used feature engineering to manually define functions to convert text and speech into feature vectors. This method relies heavily on expert experience, and it is difficult to expand features, so it cannot be used on a large scale. With the advances of deep learning, automatic learning of vectorized representations of languages became the mainstream. For an understanding task, a neural network is usually used as an encoder to map the language to low-level vectors, and vectorized expressions are used to express semantic information. For a generation task, a neural network is usually used as a decoder to map low-level vectors to a natural language and by so doing express the information contained in the vectors. The encoder-decoder framework described above can be used to process both text and speech signals. The text and speech encoders differ significantly, but the text and speech decoders are roughly the same.

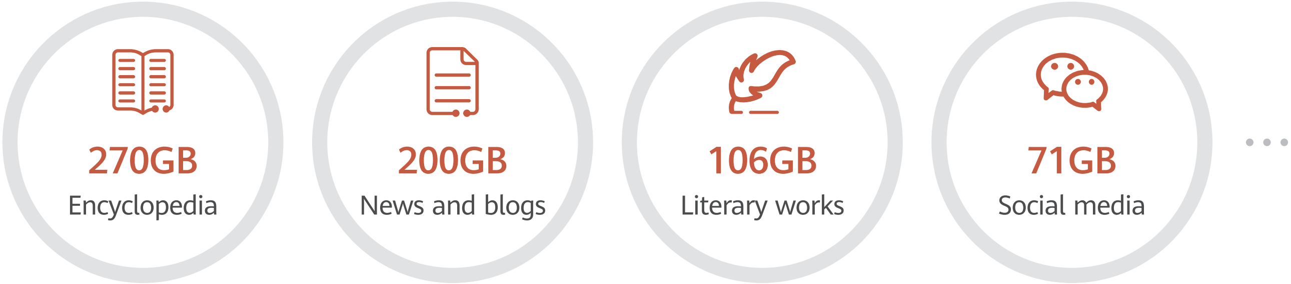
There are two core tasks in deep learning: designing the network architectures of the encoder and decoder, and learning the encoder and decoder parameters. Before large pre-trained models, CNN and RNN models were the mainstream. LSTM models, which are an extension of RNN, were particularly popular, as they are good at learning and processing long-distance dependencies. However, RNN models cannot be tuned reliably and are not good at parallel computation, so it is difficult to use them to develop large-scale language models. In 2017, the Transformer model that relies solely on a self-attention mechanism was proposed. Drawing on the strengths of existing methods, the Transformer model quickly became the primary architecture for natural language processing and speech recognition, with significant advantages both in speed and expressive power over other methods. With the emergence of large corpuses and the maturing of self-supervised learning methods, the year 2008 saw the launch of Google's BERT, a large-scale, pre-trained language model, which ushered in the era of pre-trained models. Today, with excellent generalization and prompt-based fine-tuning, pre-trained models can be quickly adapted to a wide range of downstream tasks, making them the best option to scale up intelligent applications for natural language processing and speech recognition.

Now, let's have a brief look at the general principle and technical solution of the Pangu NLP and speech recognition model.

2.2.1 Data Collection

Similar to CV, natural language processing and speech recognition also rely on large datasets. To enable powerful language understanding and generation capabilities for the model, we need to train the model on massive amounts of data covering all topics and domains.

With regard to text, we collected 40 TB of text data from publicly available web pages using web crawlers and parsed and cleansed the data. We used regular expression matching to filter out common noise data, such as web page tags, special characters, and error codes, and used the hash method to deduplicate data. Then, we standardized the data lengths by discarding articles that were too short and splitting articles that were too long, ensuring that the input lengths are within an appropriate range. Finally, we collected approximately 647 GB of text data, consisting of those shown in the figure below. With regard to speech data, we collected more than 70,000 hours of speech data in Mandarin from publicly available sources on the Internet and converted them into a total of 11 TB of audio files. Examples of video and audio sources included news broadcast, movies and TV series, variety shows, and animations.

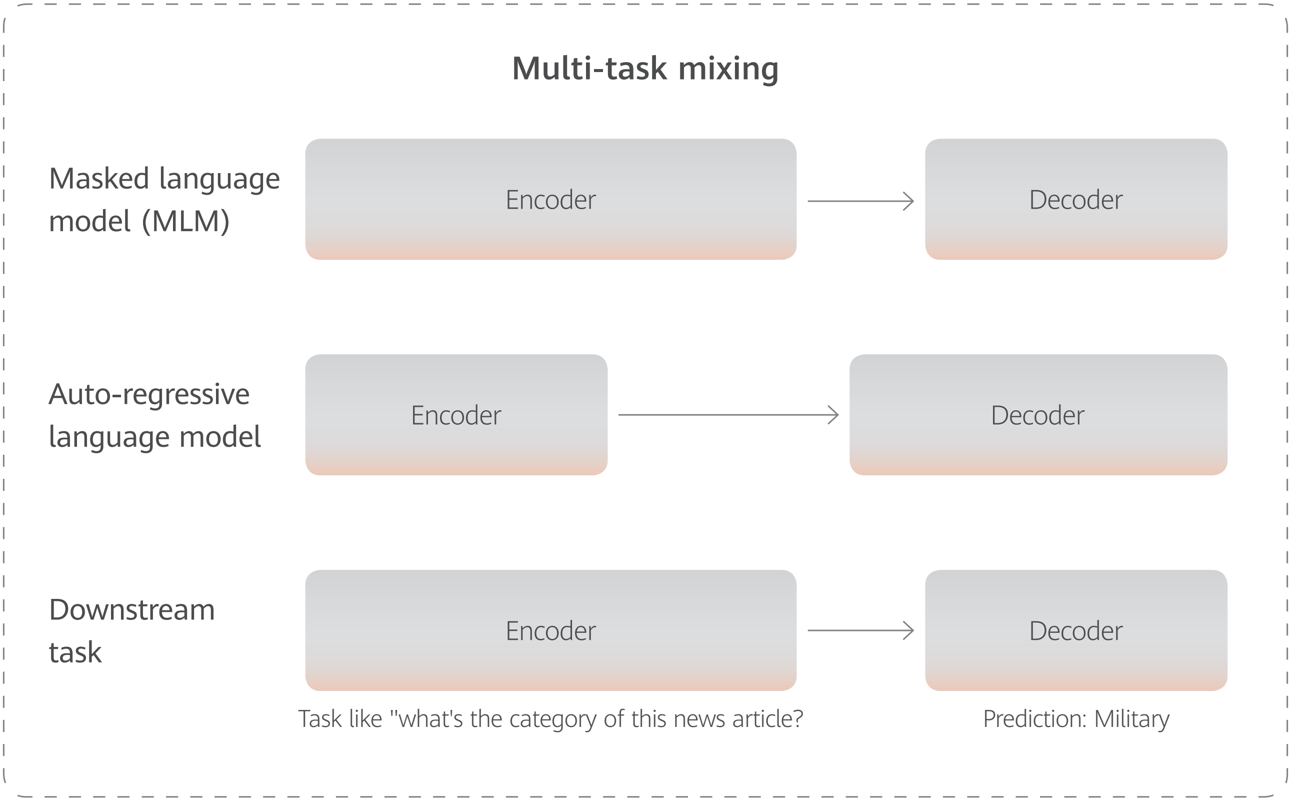


2.2.2 Pre-Training Method

For natural language processing, we used a Transformer-based encoder-decoder model. The encoder is responsible for text understanding. It uses a two-way self-attention mechanism to allow each word to fully "observe" the words on both sides of it to capture its semantic meaning in context. The decoder is responsible for text generation. It uses a one-way self-attention mechanism to generate text word by word. Each word can only "see" the word before it and predict the next word based on known information.

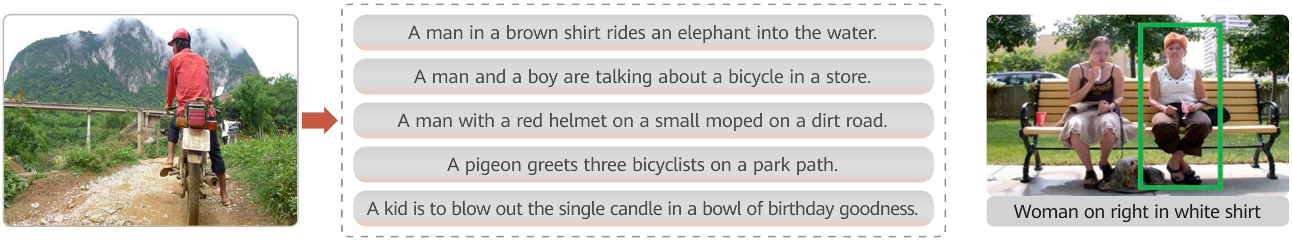
To enable the model to learn linguistic knowledge from massive text data, we must design appropriate learning objectives. We proposed a multi-task, converged training strategy to train the model on both understanding and generation capabilities. With regard to language understanding capability, we used a masked language model (MLM) as the training target. That is, we remove certain words from existing sentences and train the model to predict the missing words. With regard to generation capability, we used an auto-regressive language model as the training target. That is, we give the first half of a sentence and train the model to predict the second half. Furthermore, to enable zero-shot learning for the model, that is, to enable the model to process downstream tasks without further training, we collected the training data of over 100 downstream tasks, covering all common types of natural language processing, such as sentiment classification, intent understanding, semantic matching, and entity recognition, and used the data to pre-train the model as well.

For speech recognition, the decoder is similar to that used for natural language processing, so our focus is mainly on the speech encoder. We used a network structure that combines CNN and Transformer. The CNN extracts local information at the bottom layer, and the Transformer network extracts global information at the upper layer. We used contrastive learning as the training target: we removed a segment from a piece of audio and used some random samples as negative examples and trained the model to discover the removed segment.


Multi-task mixing
Masked language model (MLM)
Encoder
Decoder
Auto-regressive language model
Encoder
Decoder
Downstream task
Encoder
Decoder
Task like "what's the category of this news article?
Prediction: Military

2.3 Multimodal

When trying to understand the world around us, humans often rely on multiple different types of information, such as images and speech. In the field of AI, a main task of multimodal AI is to process and associate different types of information (such as audio, text, images, and video, or any other machine-readable data) from various sources, and design methods to comprehensively extract knowledge of different modes. Compared with single-modal AI, such as CV and NLP models which we discussed earlier, multimodal AI must be pre-trained on even larger datasets of multiple data modes. Then, the multimodal model can be adapted to downstream tasks to improve their accuracy. The following figure shows a typical multimodal task that includes cross-modal data retrieval (such as searching images by text or searching text by image), visual Q&A (answering questions with images), and visual grounding (locating the most relevant object or region in an image based on a natural language expression).



A man in a brown shirt rides an elephant into the water.

A man and a boy are talking about a bicycle in a store.

A man with a red helmet on a small moped on a dirt road.

A pigeon greets three bicyclists on a park path.

A kid is to blow out the single candle in a bowl of birthday goodness.

Woman on right in white shirt

Because multimodal AI uses multiple data modes, the key for a multimodal model is to represent different forms of information in a unified manner, so that computers can extract and process the information efficiently and accurately. The Pangu multimodal model focuses on the two most common data modes: visual (images) and natural language (text), and is pre-trained to support a wide range of downstream tasks. Now, let's take a look at the general principle and technical solution of the Pangu multimodal model.

2.3.1 Data Collection

Like CV and NLP/speech recognition models, a large multimodal model must also be trained on massive amounts of high-quality data. As is common practice in the industry, we collected large amounts of data from publicly available web pages using web crawlers, cleaned the data, and finally obtained high-quality image-text pairs that we can use to pre-train the multimodal model. Specifically, we set a large number of text keywords and fetched top-ranking images from search engines. Then, we paired the images with the right text (obtained from metadata) and stored the image-text pairs to create a dataset. After deduplicating the data, we further filtered out images with an excessively low resolution or excessively short text description. Then we used an existing pre-trained multimodal model to evaluate the similarity between images and their text descriptions. If the similarity is low, we discard the text and use an automatic image description algorithm to re-generate the text description. Finally, we obtained a dataset of approximately 350 million high-quality image-text pairs, with a total size of 60 TB.

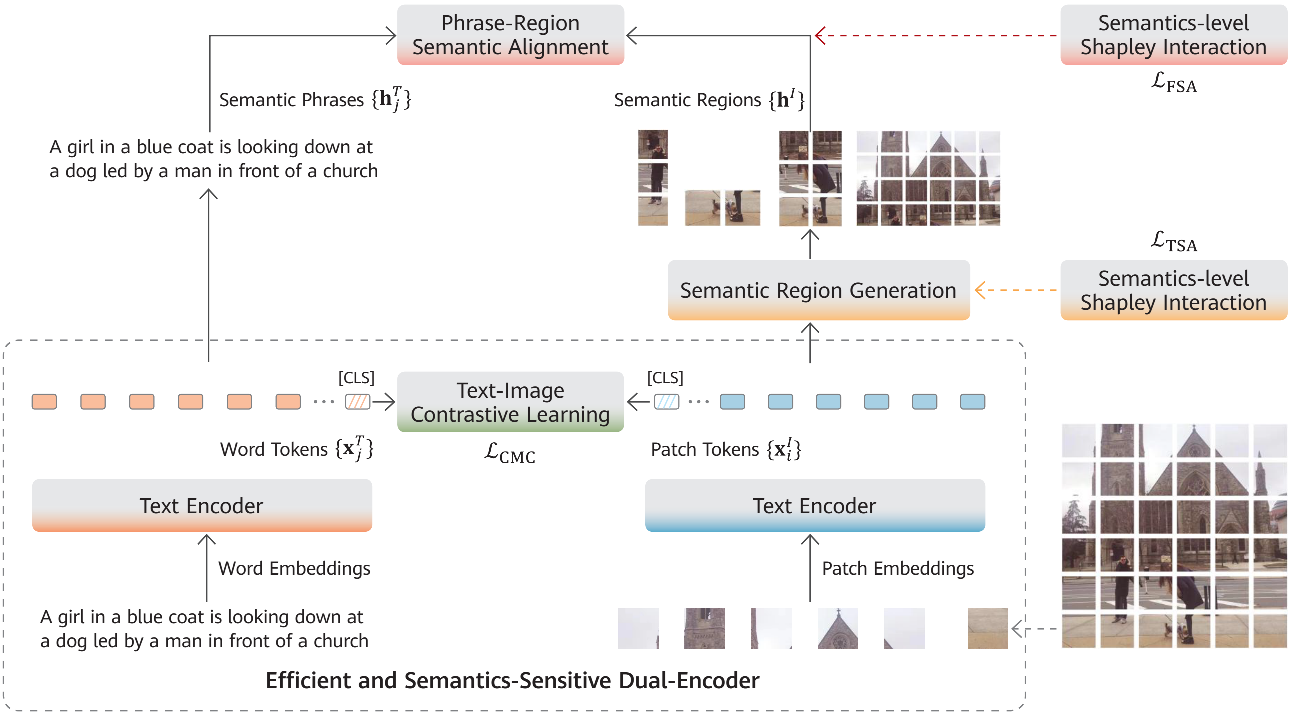
350 million
Text-image pairs

60 TB
Storage space

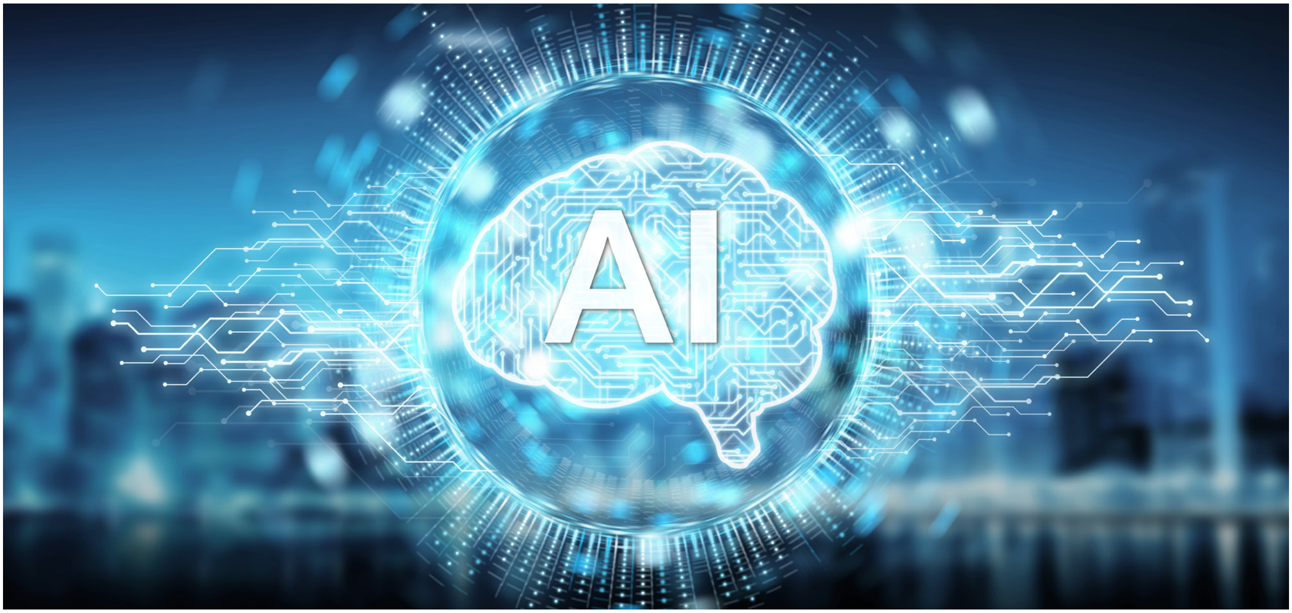
2.3.2 Pre-Training Method

The key to multimodal model pre-training lies in efficiently connecting and pairing data of different modes. Today, there are two major types of network architectures for multimodal models: single-tower and dual-tower. The single-tower architecture uses only one deep neural network (mostly Transformer) to pair images with texts, while the dual-tower architecture uses two different neural networks to extract information of different modes separately, and pairs the information together only at the last layer.

The Pangu multimodal model uses the dual-tower architecture, which makes the model highly independent and quick to train. The pre-training for the Pangu multimodal model is quite simple: Two different deep neural networks are used to extract image and text features separately. Then, image and text features of a single batch are fed into a discriminator, which, by comparing loss functions, clusters paired cross-modal features together and distances features that do not match. After enough rounds of iteration on large datasets, the model can learn how to accurately align images and the matching text into the same space. The image and text encoders obtained can be used separately for different downstream tasks, or used together for cross-modal understanding tasks.

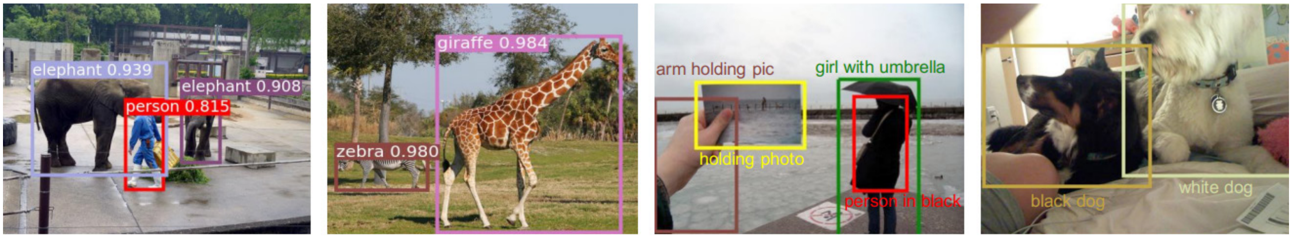

Phrase-Region Semantic Alignment
Semantics-level Shapley Interaction
$\mathcal{L}_{\text{FSA}}$
Semantic Phrases $\{\mathbf{h}_j^T\}$
Semantic Regions $\{\mathbf{h}^I\}$
A girl in a blue coat is looking down at a dog led by a man in front of a church
$\mathcal{L}_{\text{TSA}}$
Semantic Region Generation
Semantics-level Shapley Interaction
[CLS]
Text-Image Contrastive Learning
[CLS]
Word Tokens $\{\mathbf{x}_j^T\}$
$\mathcal{L}_{\text{CMC}}$
Patch Tokens $\{\mathbf{x}_i^I\}$
Text Encoder
Text Encoder
Word Embeddings
Patch Embeddings
A girl in a blue coat is looking down at a dog led by a man in front of a church
Efficient and Semantics-Sensitive Dual-Encoder

Today, however, most multimodal models that use a dual-tower architecture focus on global information alignment, but do not pay enough attention to finer-grained alignment of details. For example, an image may include multiple visual entities and the corresponding text descriptions also include many noun phrases. Finer-grained alignment of these visual entities and noun phrases will improve the multimodal model's ability to pair images and text with higher accuracy. To do that, the Pangu team proposed the in-house developed algorithm LOUPE, which was published at NeurIPS 2022 Conference. Building on Game Theory, this algorithm extracts visual entities from images and noun phrases from text, and aligns them in fine granularity by comparing the loss functions. The multimodal model trained using this method delivers relatively high accuracy in multiple downstream tasks.



2.3.3 Performance

The Pangu multimodal model has achieved industry-leading performance in a wide range of downstream multimodal tasks, such as cross-modal retrieval, automatic generation of image descriptions, and visual grounding. The model pre-trained using the LOUPE algorithm achieved the highest accuracy in image-text retrieval on cross-modal retrieval datasets Flickr30k and MS-COCO, and in "search image by text" tasks, it outperformed CLIP — the benchmark algorithm for visual classification — by 12.3% on the MS-COCO dataset. It also demonstrated good results in object detection and visual grounding tasks in open domains. The following figure shows two examples, one for each task.

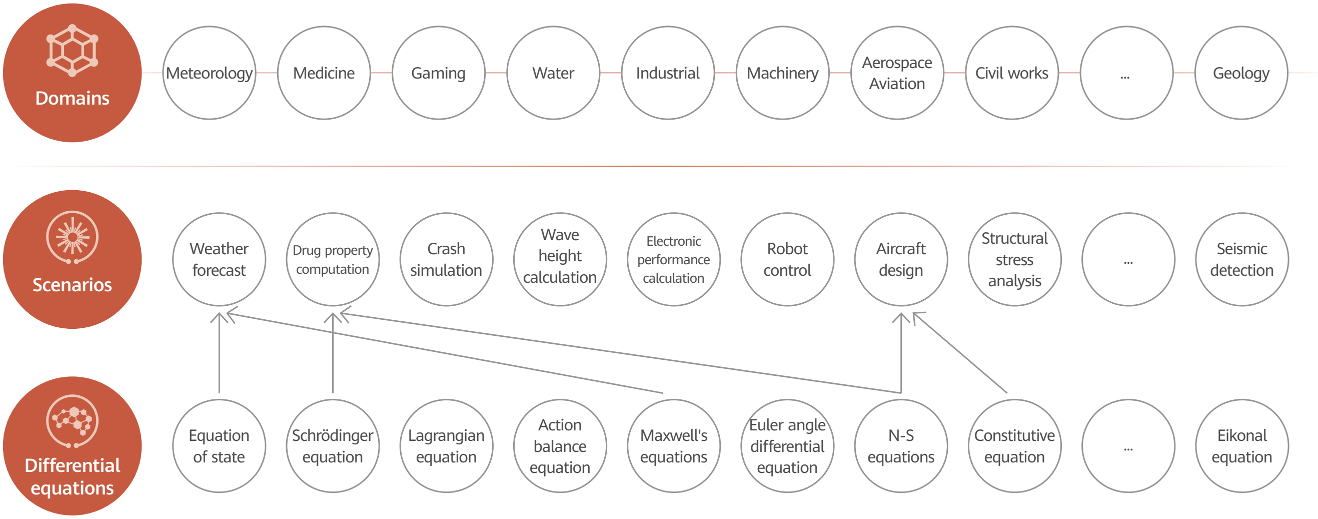


(a) Object Detection

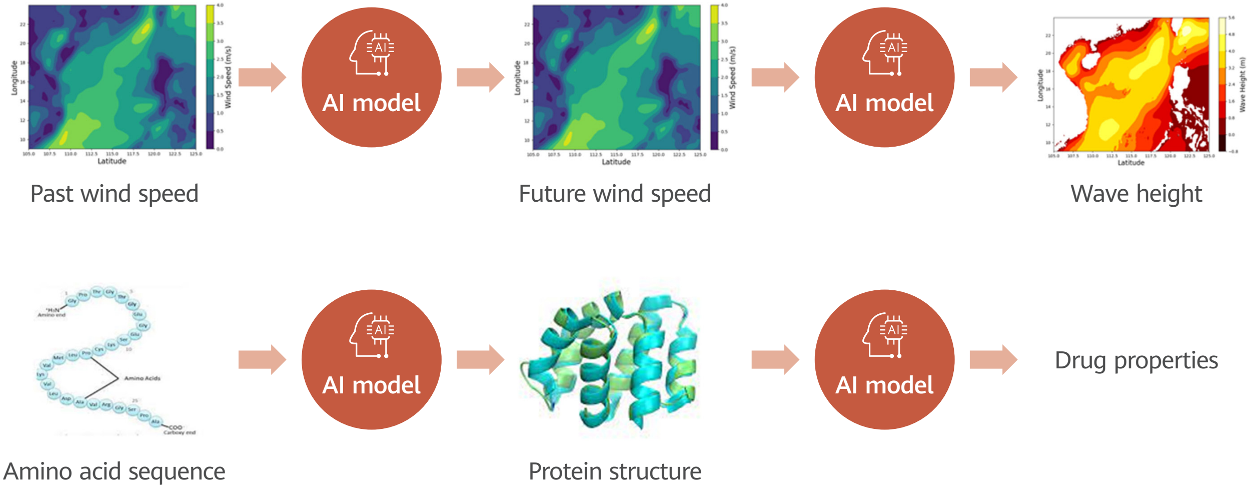
(b) Visual Grounding

2.4 Scientific Computing

CV, NLP, and multimodal modals target more general-purpose AI problems, such as audio analysis, image recognition, and semantic understanding. Humans are naturally good at solving these problems, so they can label large datasets to train deep neural networks. In natural science, however, there are many problems that humans are unable to solve using only their brain, such as turbulence modeling, weather forecasting, and large deformation stress modeling. These problems exist in a wide range of scenarios, as shown in the following figure.



The problems above are valuable and yet also complex. Before AI, scientists tried to extract the hidden patterns and rules in these problems by using mechanism formulas to analyze experimental data. These traditional methods may easily run into dead ends when large-scale, high-dimensional data processing is involved. With the rapid advance of AI technology in recent years, AI+scientific computing methods are picking up steam in solving complex scientific problems. The idea is to embed deep neural networks into scientific equations to find hidden patterns and rules from both observational and simulation data.



In terms of pre-training, there are many similarities between the scientific computing model and the models we discussed earlier. They are all built on large-scale datasets, all depend on neural networks with a large number of parameters and complex tuning processes, and in the end, they all store knowledge in the large number of parameters of their networks. Now, we will try to briefly describe the uniqueness of scientific computing models.

2.4.1 Data Collection

In AI+scientific computing scenarios, data is divided into observational data and simulation data. Observational data is generated or collected by observation tools (such as calipers, radars, and sensors), while simulation data is generated by simulation algorithms (corresponding to human knowledge). AI models can learn from both types of data as well as the knowledge and mechanisms contained in them.

- Observational data obtained varies greatly in different scenarios of scientific computing. In specialized domains, specialized instruments and systematic experiments are usually needed to collect observational data. For example, for the purpose of protein structure prediction, X-ray diffraction analysis (XRD) and magnetic resonance imaging (MRI) are needed to measure protein structures; for the purpose of predicting short-term rainfall, weather radars need to collect radar reflectivity data; while for plant phenotyping, data needs to be collected by specialized researchers. Some scientific computing problems need large amounts of observational data. For example, weather forecasts need the historical data of global weather stations, plus satellite data and radar reflectivity data. Others need relatively smaller amounts of observational data. For example, structural stress analysis only needs data collected by a few sensors.

- Simulation data is generated by simulation algorithms and contains rich mathematical and physical information. For the same problem, different simulation data may be generated using different algorithms. Different from observational data, the accuracy of simulation data depends on the accuracy of the simulation algorithms used and the computing power used during the simulation process. Compared with observational data, simulation data usually has a larger volume (depending on the computing power used for the simulation) and fewer default values, making it a useful complement to observational data.

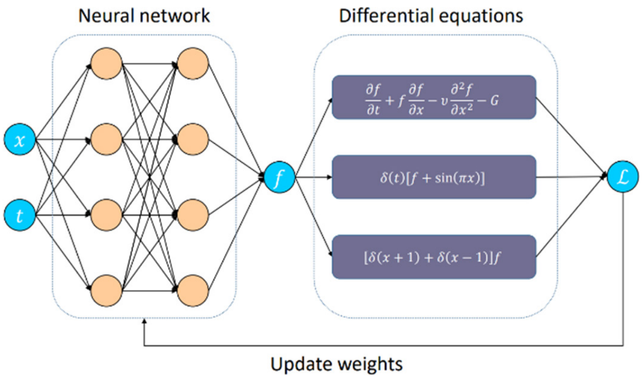
In some cases, observational and simulation data can be combined based on domain-specific mechanisms and knowledge to generate new fusion data. One example is reanalysis data for weather forecasting. Reanalysis data is structured data obtained by using an assimilation algorithm to combine simulation data and experimental data. The results depend on the assimilation algorithm as well as the simulation data used. The following table lists the characteristics of data used for a number of scientific computing problems.

	Data Size	Noise	Data Structure	Data Change	Accuracy	How to Collect	Data Characteristics	Application
Radar reflectivity data	GB to TB	Loud	(X,Y,Z,T) Each spatial point corresponds to a radar reflectivity value.	Medium	Medium	Weather radar	The raw data is in the form of polar coordinates, and there are blank areas in the assembled radar reflectivity data.	Short-term rainfall prediction
Plant phenotyping records	MB to GB	Medium	(N, C) Each phenotype (e.g. yield and plant height) of each plant corresponds to a value.	Small	High	Researchers manually or use a high-throughput phenotype analyzer to collect the data.	Data is difficult to collect, and there are only a few data points.	Analysis of plant phenotype-genotype relationships
Amino acid sequence data	TB to PB	Lower	Sequences with a fixed vocabulary	Large	High	Calculated from known DNA sequences	Similar to text data	Protein structure prediction
Weather forecast data from meteorological agencies	TB to PB	Low	(X, Y, Z, T)	Relatively large	Low	Obtained by meteorological simulation algorithms	There are gaps between simulation and observational data.	Weather forecasts
Atmospheric reanalysis data	PB	Loud	(X, Y, Z, T)	Relatively large	Medium	Obtained by combining simulation and observational data	It deviates from but includes observational data.	Mid- and long-term weather forecasts

2.4.2 Model Building

Depending on the input data, different base models are selected for training. Take an ocean wave forecast task as an example. The goal is to predict the real-time wave heights in any sea area around the world. Both input and output data are two-dimensional spherical data with timestamps. In this case, a 2D network model is preferable. If the goal is to predict the global weather, both input and output data are three-dimensional data (including heights) with timestamps. In this case, a 3D network model is preferable. Both 2D and 3D networks may be built on the appropriate CV models. For example, they may use a convolutional neural network (CNN) or visual transformer as the backbone architecture to pre-train models on large datasets.

A major characteristic of scientific computing is that it builds on human experiences that were gained in the past. Such experiences usually become part of the output data, for example, as some constraints expressed as partial differential equations (PDEs). As shown in the figure below, we can embed PDEs into the neural network to help with model architecture design or use them as additional constraints. They can be used together with standard observational or simulation data to train the neural network model. Given appropriate implementation, this type of knowledge can usually enhance the robustness of the model and enable the model to fit training data more easily and reliably.


Neural network
Differential equations
x
t
f
$\frac{\partial f}{\partial t}+f\frac{\partial f}{\partial x}-\upsilon\frac{\partial^2 f}{\partial x^2}-G$
$\delta(t)[f+\sin(\pi x)]$
$[\delta(x+1)+\delta(x-1)]f$
$\mathcal{L}$
Update weights

$$
\begin{gathered}
\frac{\partial N}{\partial \mathrm{t}}+\nabla x\cdot\dot{\boldsymbol{x}}N+\frac{\partial}{\partial\theta}\dot{\theta}N=\frac{S}{\sigma},\\
\dot{\boldsymbol{x}}=\boldsymbol{c_g}+\boldsymbol{U},\\
\dot{k}=-\frac{\partial\sigma}{\partial d}\frac{\partial d}{\partial s}-\boldsymbol{k}\cdot\frac{\partial\boldsymbol{U}}{\partial s}\\
\dot{\theta}=-\frac{1}{k}[\frac{\partial\sigma}{\partial d}\frac{\partial d}{\partial m}+\boldsymbol{k}\cdot\frac{\partial\boldsymbol{U}}{\partial m}]
\end{gathered}
$$

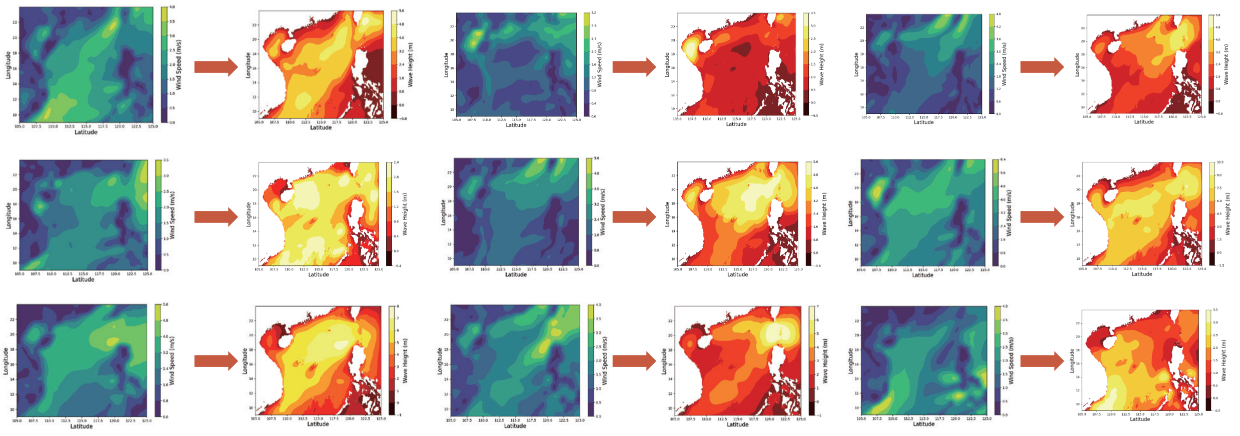
(Note: The left figure shows a neural network embedded with partial differential equations, and the right figure shows examples of partial differential equations used for wave forecasts.)

2.4.3 Use Case Example and Performance

Below, we will show a typical example of scientific computing: a global ocean wave height prediction system. Traditional scientific computing methods calculate the height of waves by solving equilibrium equations, which usually involves supercomputers and large amounts of compute power. Because real-time computation is impossible, traditional methods are unable to make real-time wave predictions if any element (such as wind speed) changes — there is a certain calculation delay.

Both the input and output of the wave height prediction problem are meteorological element data on the longitude and latitude grid networks, and are similar to video data formats. The difference lies in that each piece of metadata of video data is a pixel value ranging from 0 to 255, while each piece of metadata of meteorological data, such as wind speed, terrain, and wave height, is a floating point number. In addition, the output of a wave prediction is usually not a classification of some sort, but rather a continuous prediction value. This means a regression loss function needs to be used to replace the classification and segmentation loss functions commonly used in deep learning. In addition, different from video data, ocean wave data does not feature translational symmetry, but it does have a number of invariance features under a spherical coordinate system, for example, rotation around the earth's axis. Therefore, a CNN or Transformer architecture needs to be used to provide certain invariances.

The main body of the Pangu ocean wave forecast model is a visual Transformer architecture that allows for rotational invariance. The model has approximately 500 million parameters. As mentioned above, the loss functions of the neural network consist of two parts: the prediction error on the actual data and the partial differential equations used in the model. The model is trained on global wave height data generated over the past 10 years. The average error of the model on the validation set is less than 5 cm, which is on par with that of the traditional methods. This allows it to replace traditional methods in many situations. More importantly, the AI-supported method is much faster than traditional methods. Powered by a single Huawei Ascend chip, the new method can predict global wave heights within 1s, and complete more than 100 wave prediction tasks within 1 minute. The inference speed is improved by four to five orders of magnitude over traditional methods. Using AI algorithms, we can quickly obtain the wave heights under different wind speeds and make accurate predictions and simulations. This is invaluable in a wide range of scenarios, such as ocean farming and disaster prevention and risk reduction.



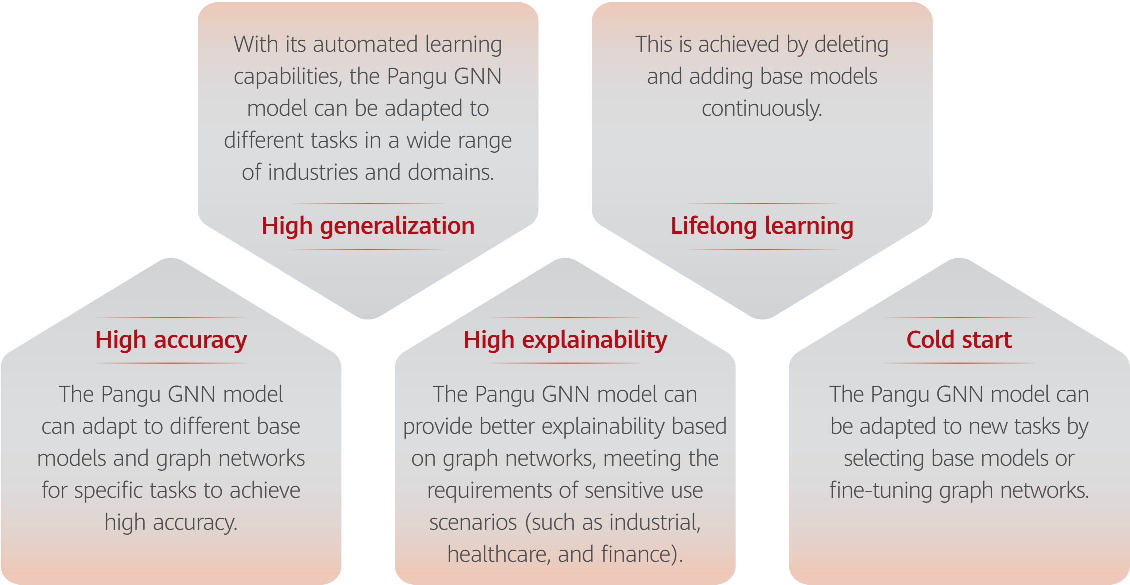
Using an Ascend AI chip, the AI model can make hundreds of wave height predictions under random wind speeds within 1s.

Figure: The Pangu ocean wave forecast model can simulate thousands of possible scenarios in a short period of time, with an accuracy comparable to traditional scientific computing methods.

2.5 Graphical Neural Network

Over the past few years, AI's impact has been felt across a wide range of industries. In addition to common data types such as images, text, and videos (commonly known as Euclidean data), there are a large variety of non-Euclidean data, such as companies' ERP data (planning, finance, sales, and procurement), molecules and genes, transportation networks, stocks, and point clouds. It is difficult to process such data using a standard convolutional or Transformer architecture. Instead, neural networks for different tasks and data modalities are needed. This is why we need to use graphical neural networks (GNNs) to model such data, where graphs are used to represent the relationships between different data elements.

The Pangu GNN model was designed to unify the training, optimization, fine-tuning, and deployment of large graph network models for a wide range of different tasks. The Pangu GNN model features the following:



High generalization

With its automated learning capabilities, the Pangu GNN model can be adapted to different tasks in a wide range of industries and domains.

Lifelong learning

This is achieved by deleting and adding base models continuously.

High accuracy

The Pangu GNN model can adapt to different base models and graph networks for specific tasks to achieve high accuracy.

High explainability

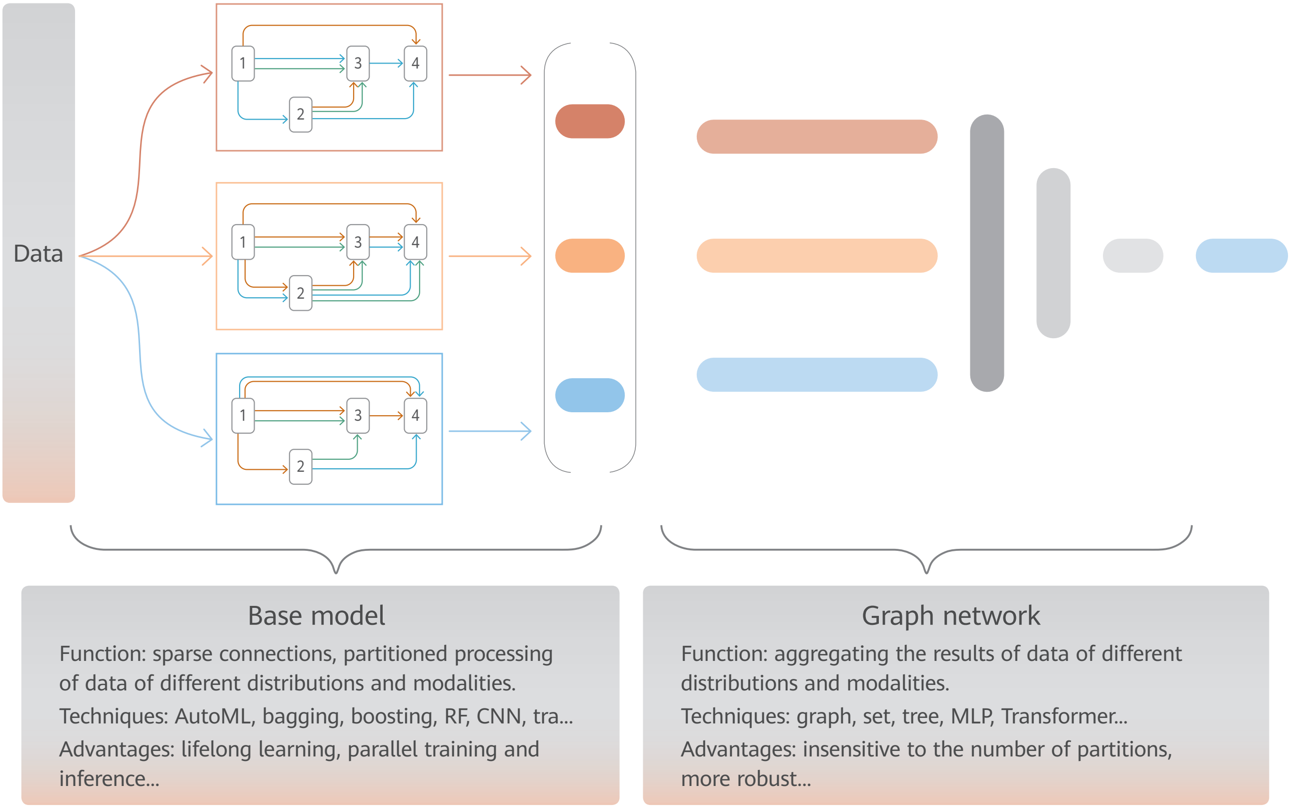
The Pangu GNN model can provide better explainability based on graph networks, meeting the requirements of sensitive use scenarios (such as industrial, healthcare, and finance).

Cold start

The Pangu GNN model can be adapted to new tasks by selecting base models or fine-tuning graph networks.

The Pangu GNN model is also equipped with other features that make the model easier to deploy, such as model encryption (protecting intellectual property rights when the model is deployed across the cloud, edge, and end devices) and multi-node parallel training.

The figure below shows the overall architecture of the Pangu GNN model.



Base model

Function: sparse connections, partitioned processing of data of different distributions and modalities.

Techniques: AutoML, bagging, boosting, RF, CNN, tra...

Advantages: lifelong learning, parallel training and inference...

Graph network

Function: aggregating the results of data of different distributions and modalities.

Techniques: graph, set, tree, MLP, Transformer...

Advantages: insensitive to the number of partitions, more robust...

Figure: Overall architecture of the Pangu GNN model

The top-level design of the Pangu GNN model consists of two parts: base model selection and base model fusion. In the base model selection part, the Pangu model automatically uses different oversampled datasets to train different base models. During this process, the hyperparameters of different base models are searched using the appropriate AutoML techniques. (In the figure above, color-coded arrows indicate different data flows, and color-coded boxes indicate different base models). This simplifies parameter tuning for developers. In the fusion part, each base model makes a prediction based on the input data. Then, all predictions are aggregated in the graph network to obtain the final output. An advantage of this solution is that the base models in the graph network can be added or deleted as needed, without impacting the aggregation of the graph network. This is because the graph network itself is insensitive to the quantity of base models used.

To make the Pangu GNN model easier for developers to use, the Pangu GNN model uses a carefully designed code architecture. The overall code structure is clear and easy to read and maintain.

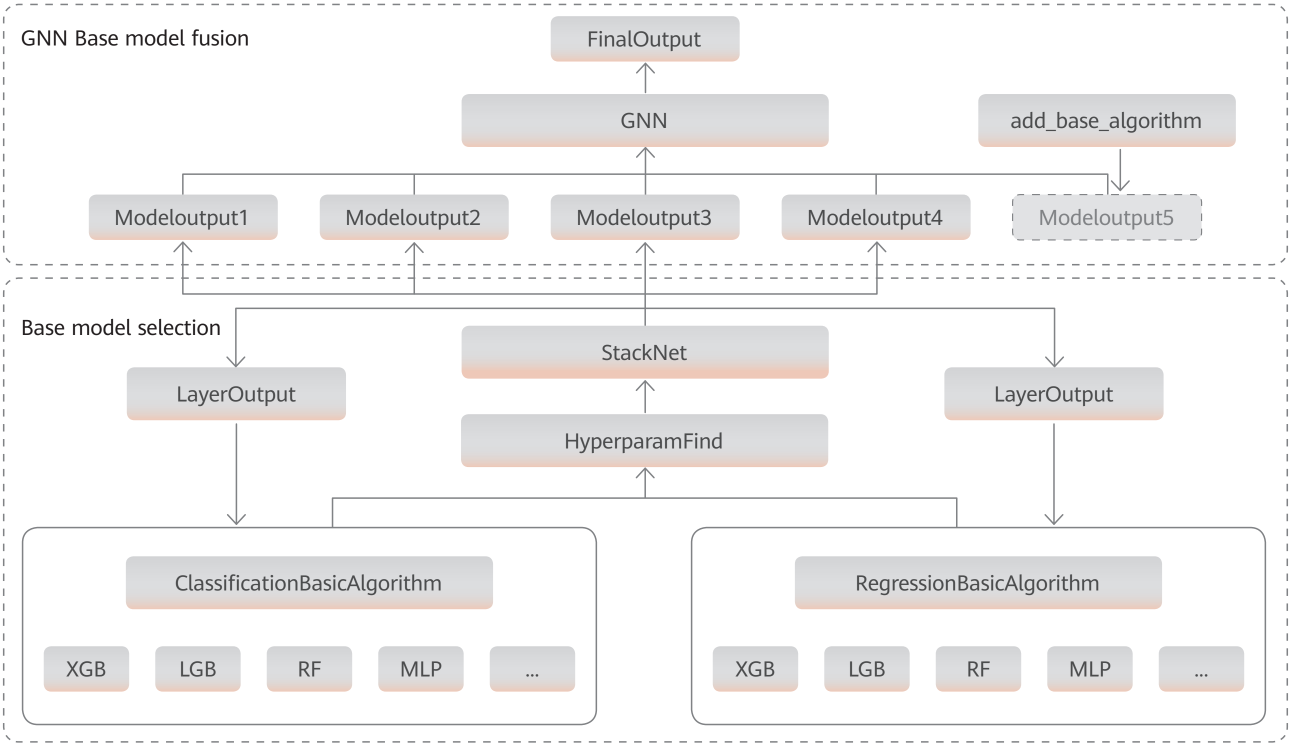


Figure: Code structure of the Pangu GNN model

The figure above shows the basic code structure of the Pangu GNN model. The structure consists of two parts: base model selection and base model fusion. In the selection part, base models are selected using the algorithms (pools) and search spaces generated by BasicAlgorithm, then a hyperparameter search is performed using the HyperparamFind class. Then, the selected base models are passed to StackNet to train the network layer by layer and generate the output. The result is used as the input of the next round of base model selection and hyperparameter search. After all layers consisting of base models are generated, the graph neural network aggregates the results and obtains the final output. We can also use add_base_algorithm to easily add other trained base models to the network and perform the aggregation (as indicated by ModelOutput5 in the dotted line box in the figure above), without the need to change other base models, network layers, or the graph neural network. This allows the Pangu GNN model to be easily adapted to vastly different tasks.

In the next chapter, we will provide a couple of case studies on the use of the Pangu GNN model.

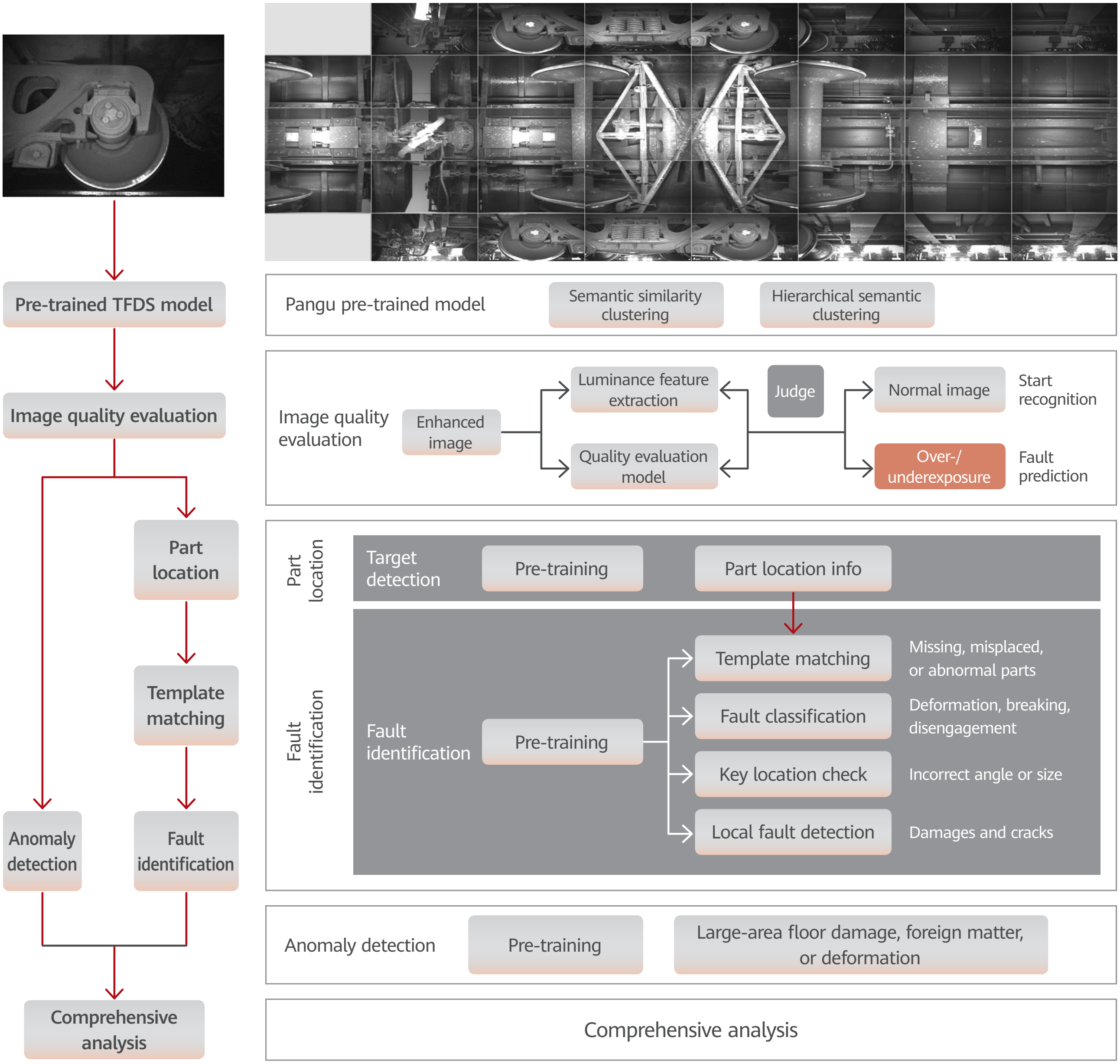
03

Pangu Model Case Studies

3.1 Pangu CV Model Case Study: Railway Defect Detection with TFDS

TFDS, or Trouble of moving Freight car Detection System, is a system for detecting faults and defects on moving trains from camera-captured images. This system typically consists of three parts: data collection, data preprocessing and transmission, and the trouble detection center. High-speed camera arrays installed at both sizes of the railway track capture images of the bottom and the lower parts of both sizes of the trains. The images are processed and displayed on terminals in the detection center. Human inspectors analyze the captured images to check for anything suspicious. There are 6,000 freight car inspectors all over China. This is quite an expenditure for railway companies. Suppose the annual wage of each freight car inspector is CNY150,000. Then the annual expenditure on this task is nearly CNY1 billion for the entire country. What's more, TFDS inspection is a demanding task that requires human inspectors to quickly detect faults and defects on moving freight cars within a short time window by analyzing a large number of images. This task is important for the safe operations of the railway network.

Efforts to use automatic image recognition to enhance existing TFDS systems began in 2007. The results, however, have not been satisfactory, due to various reasons, such as the diverse shapes of faults and defects, varying image quality, and soiled car surfaces. Only with faults that are easy to detect using naked eyes, such as a closed handle on an air brake, the identification rate reached above 80% with the help of the SVM (Support Vector Machine) technique. However, there have not been satisfactory ways to detect other types of faults. Consequently, TFDS images are still mainly analyzed manually at many railway companies. More efficient, intelligent methods are needed.



TFDS solution supported by the Pangu pre-trained CV model

The figure above shows the tailored TFDS solution powered by the Pangu CV model. The solution includes several modules: car model screening, location classification, part screening, image quality evaluation, matching with existing templates, and multi-car cascaded analysis. Key techniques include the following:

Car splitting

Spits the images of an entire train into those of each individual car.

Pre-trained model

The TFDS-tailored model was pre-trained using millions of unlabeled train images.

Automatic enhancement and evaluation

Automatically evaluates images for possible faults and defects and sends suspicious images for manual review.

Template matching

Creates templates that show the relative locations of parts based on existing information about specific car models. They help to identify missing or misplaced parts.

Fault locating and identification

The pre-trained model uses techniques such as object detection and image recognition to identify and locate faults and defects.

The solution was verified on 14 railroads. In a test environment, a total of 32,007 fault samples (images that have been identified by inspectors as containing faults or defects and confirmed by their team leaders) from September 19, 2021 to October 20, 2021 were mixed with a large number of fault-free images and fed to the Pangu CV model. As shown in the table below, the test results show that the Pangu CV model has surpassed humans in image recognition accuracy.

	Accurately predicted	False negative	Total faults	Recognition rate
Stop	119	1	120	99.17%
Major	28,280	506	28,786	98.24%
Minor	3,084	17	3,101	99.45%

3.2 Pangu NLP and Speech Recognition Model Case Study: Sales Team Empowerment

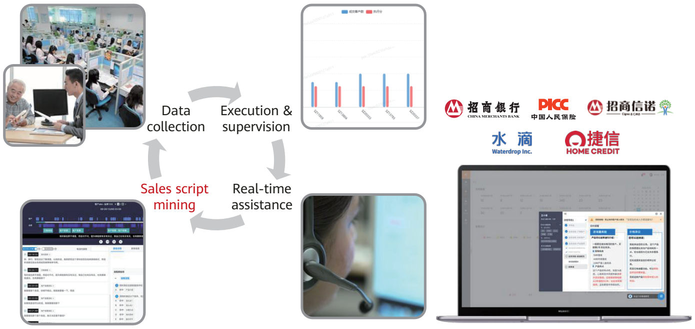
The effectiveness and efficiency of the sales team is vital to many businesses in a wide range of sectors, such as banking, insurance, automotive, and real estate. Salespeople must have good language skills. First, they need to know what might interest customers, so that they can recommend the right products. Second, they need to know how to sell the products to customers in the most convincing way possible. In traditional sales team evaluations, results were often the only standard, and the sales process in the middle was not properly monitored and analyzed. Low-performance salespeople are unclear on how they can improve, such as how to accurately discover customers' buying intent, and the experience and skills of high-performance salespeople cannot be properly summarized and replicated.

The **real-time sales assistant system** improves the sales productivity of banks and insurance companies by 10% to 50%, in both online and offline scenarios.

Challenges

- Significant performance gap between average and elite salespeople
- Unable to promptly identify and address the skill gaps of sales teams, significant loss in performance

Huawei Cloud Pangu NLP model: solution and benefits



- E2E sales assistance
- Productivity improved by 50% for junior salespersons, and 10% to 30% for mid-level salespersons
- Replicable to offline bank branches

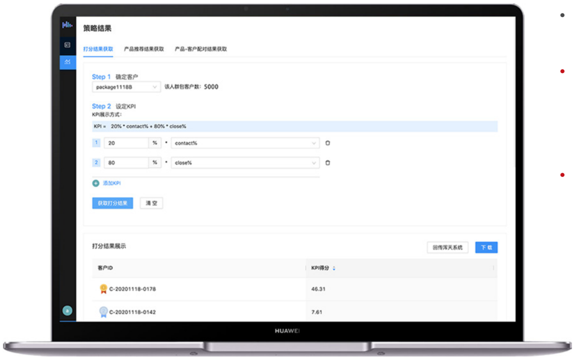
We used the Pangu NLP and Speech Recognition Model to empower the sales team with its powerful speech recognition, natural language understanding, and natural language generation capabilities. The speech recognition model captures the conversations between sales and customers, and the NLP model analyzes the conversations. For customers, we assess their buying intent, so that our salespeople don't spend a lot of time on customers with low buying intent. Then, we analyze customers' specific buying intent and recommend to them products that might interest them. For our sales team, we analyze what the salespeople said to the customers to see if they missed any important messages about the product. This can be a way to evaluate the competence and appraise the performance of specific salespeople. Based on the analysis, we can recommend phrasing that makes our salespeople sound more persuasive. With the help of the real-time sales assistant system, the productivity of junior salespeople can improve by around 50%, and that of more senior-level salespeople can improve by 10% to 30%. Their sales success rate can double or even triple.

Customer buying intent prediction and sorting and product recommendation systems help insurance companies improve the sales conversion rate two to three times.

Challenges

- Salespeople waste too much time on customers with a low buying intent, leading to a low conversion rate.
- Product recommendations are purely based on experience and are not verified with data.

Huawei Cloud Pangu NLP model: solution and benefits



Step 1 确定客户
package11188　该人群包客户数：5000

Step 2 设定KPI
KPI展示方式：
KPI = 20% * contact% + 80% * close%
1　20　%　*　contact%
2　80　%　*　close%

打分结果展示

客户ID	KPI得分
C-20201118-0178	46.31
C-20201118-0142	7.61

- Predicts buying intent based on past communications and purchase records.
- Improves conversion rate by 2x to 3x based on accurate customer buying intent prediction and product recommendations.
- Works with multiple sales channels, such as telemarketing, online sales, and reselling

招商信诺 Cigna & CMB　中国太平 CHINA TAIPING
太平洋保险 CPIC　MetLife
轻松筹

The excellent generalization capability of the Pangu model allows us to apply it to a wide range of sectors, with predictable performance.

In addition to B2B use cases, the Pangu NLP and Speech Recognition Model also supports many B2C use cases, such as personal voice assistant, Q&A bot, and dialog generation. With massive amounts of encyclopedia data and general knowledge built into it, the Pangu model can power Q&A bots that can give very plausible answers to questions in specific domains. For example, a Q&A bot can accurately answer questions like "what are the best tourist attractions in this city?" The model is also capable of multi-round dialogs with human users in a very human-like manner.

3.3 Pangu Multimodal Model Case Study: Government Service Ticket Allocation

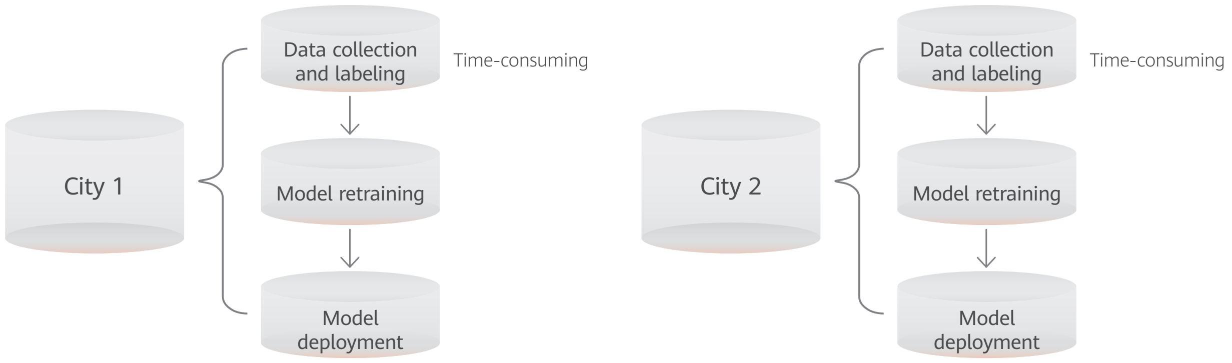
For large cities, the unified government service platform generates a large number of service tickets every day, and it has up to 300 types of events. Fast, accurate allocation of government service tickets is important to ensure efficient, smart government services. By quickly allocating service tickets to the right agencies, the government can accelerate its response to incidents and improve resident satisfaction.

The input information of a service ticket is often a couple of images accompanied by a text description sent by a grid manager or any resident. The small models that were previously used had low accuracy in classifying incidents and events. As a result, many tickets were not allocated to the right agencies, leading to low efficiency and slow response. Furthermore, due to the absence of unified national standards, event categories vary in different cities. When small models were used, data needed to be re-collected and the models re-trained, which is both time- and labor-consuming. Small models were clearly not the ideal solution.

The Pangu multimodal model offers a plausible solution to intelligent allocation of government service tickets. Built on Huawei's in-house developed algorithms, the Pangu multimodal model was pre-trained on large datasets of image-text pairs in open domains. The pre-trained model can extract useful information from the images and text descriptions uploaded by grid managers and residents, and match and associate the extracted information with known features of existing categories. This way, the model can accurately allocate service tickets even without being trained on any labeled data. This is particularly useful when no labeled data is available. With excellent generalization capabilities, the Pangu model delivers high classification accuracy. With continuous learning, the model becomes increasingly more accurate over time. Compared with small models, the Pangu multimodal model improves the accuracy of automatic allocation of government service tickets by over 15%. The accuracy is comparable to the level that can be achieved by experts.

Old process

Takes a long time to retrain AI models for each city



City 1

Data collection and labeling

Model retraining

Model deployment

Time-consuming

City 2

Data collection and labeling

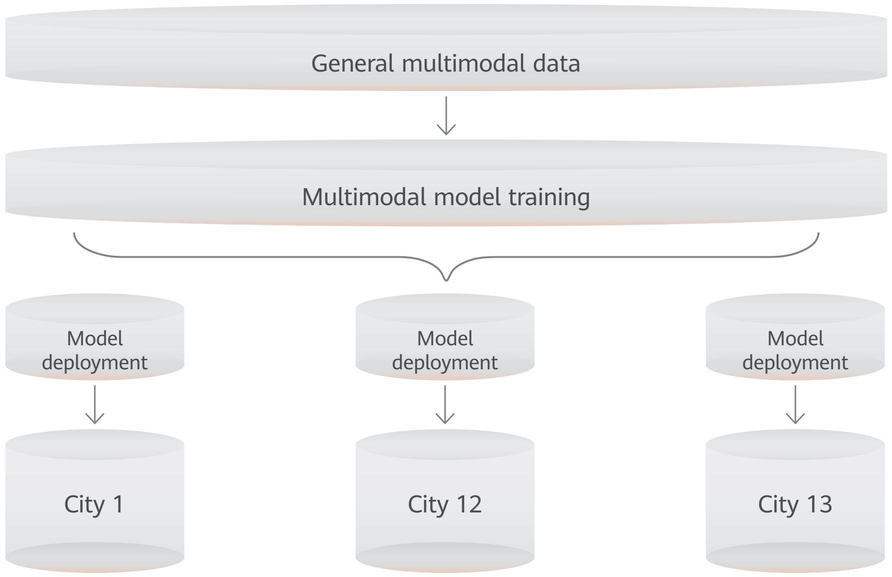
Model retraining

Model deployment

Time-consuming

New process with Pangu Multimodal Model

Cold start, high efficiency, low cost



General multimodal data

Multimodal model training

Model deployment

City 1

Model deployment

City 12

Model deployment

City 13

3.4 Pangu GNN Model Case Study: Automated Control of Cement Manufacturing Systems

In cement manufacturing, the control variables (CV) of systems must be adjusted constantly based on the real-time conditions of the kiln. Traditionally, this adjustment was done manually by experienced workers and was far from being real-time. Automatic, real-time adjustment of CV for cement manufacturing systems requires the joint work of a predictor and solver. The predictor provides accurate kiln conditions, and the solver predicts the results of relevant metrics under the current CVs, and generates the CV values needed to produce the optimal results. Then, the CV values that produce the optimal results are set for the systems. The goal of this optimization process is to minimize energy consumption, and the constraints are the product yield and quality.

By applying the Pangu GNN model to the cement production system, the Pangu model can predict the coal consumption and cement quality based on the real-time kiln conditions and the CV values generated by the solver. The information can be used to calculate CV values that can lead to even lower coal consumption and better cement quality. The figure below illustrates this process.

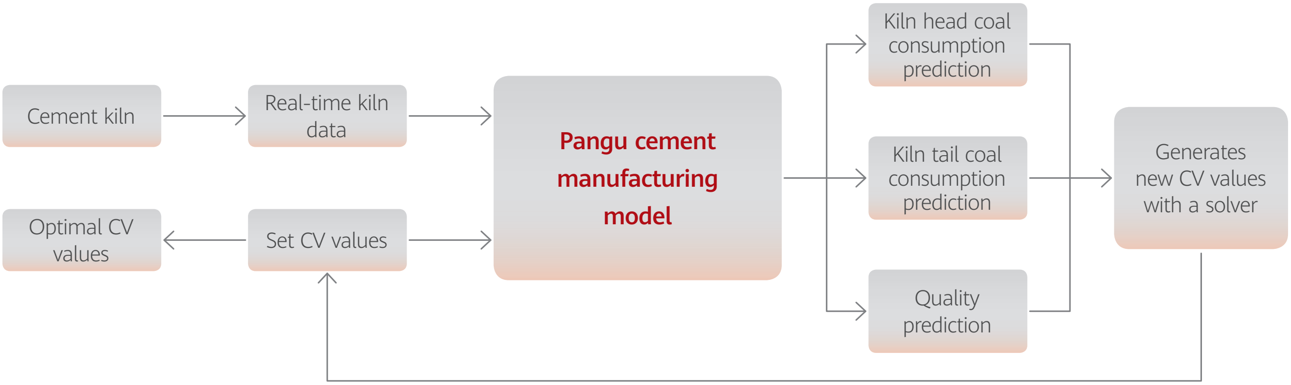

Cement kiln
Real-time kiln data
Optimal CV values
Set CV values
Pangu cement manufacturing model
Kiln head coal consumption prediction
Kiln tail coal consumption prediction
Quality prediction
Generates new CV values with a solver

Figure: Using the Pangu GNN model to optimize energy consumption for cement manufacturing

As shown in the following table, compared with the original method, the Pangu GNN model is much more accurate in predicting coal consumption and cement quality for cement manufacturing systems based on CV values.

		R2 (the larger the better)	
		Original Algorithm	Pangu GNN
Kiln condition 0	Kiln head coal consumption	0.218	0.511
	Kiln tail coal consumption	-1.666	0.147
	Quality prediction	0.007	0.534
Kiln condition 1	Kiln head coal consumption	0.354	0.661
	kiln tail coal consumption	-1.235	0.098
	Quality prediction	-0.307	0.471

3.5 Pangu GNN Model Case Study: Automated Control of Coking Systems

In the coking industry, coal blending is key to cost control. Traditional coal blending methods rely heavily on the experience of experts, without data about the end-to-end processes. Traditional methods can hardly cope when many different types of coal are used together.

By applying the Pangu GNN model to coking systems, multiple coal blending theories can be converted into mechanism models and become base models of the Pangu GNN model. Used with an optimization solution, the Pangu GNN model can accurately predict coke quality and quickly search for the coal blending formulas that can produce the optimal results. The model also offers high explainability. The figure below shows how the Pangu GNN model is used to optimize the coking process.

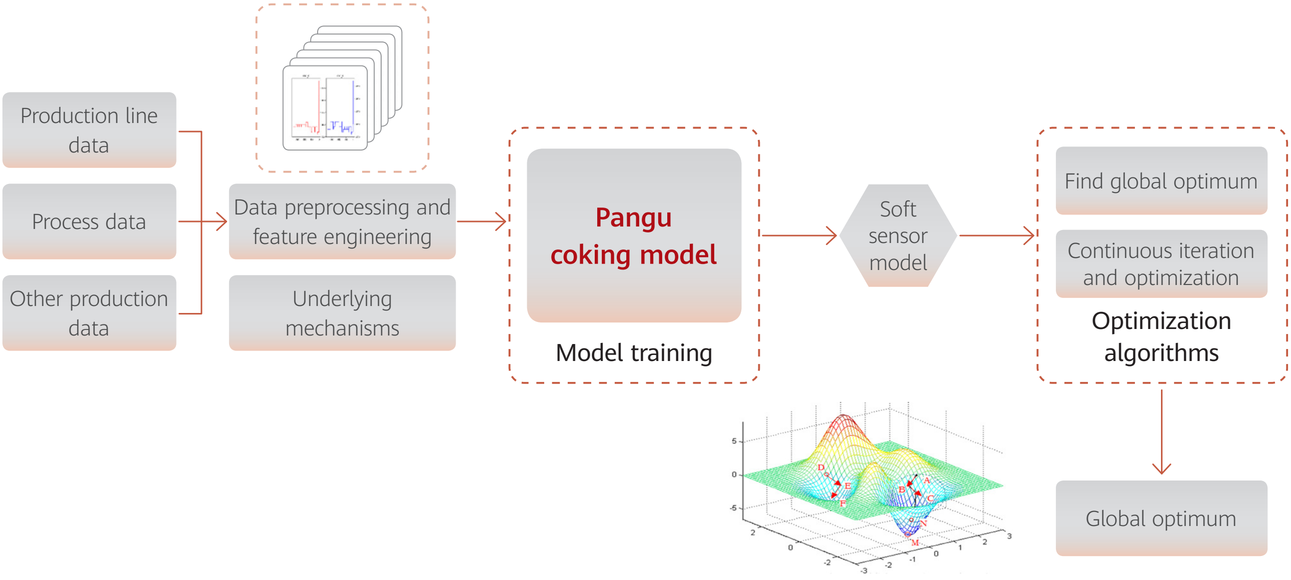
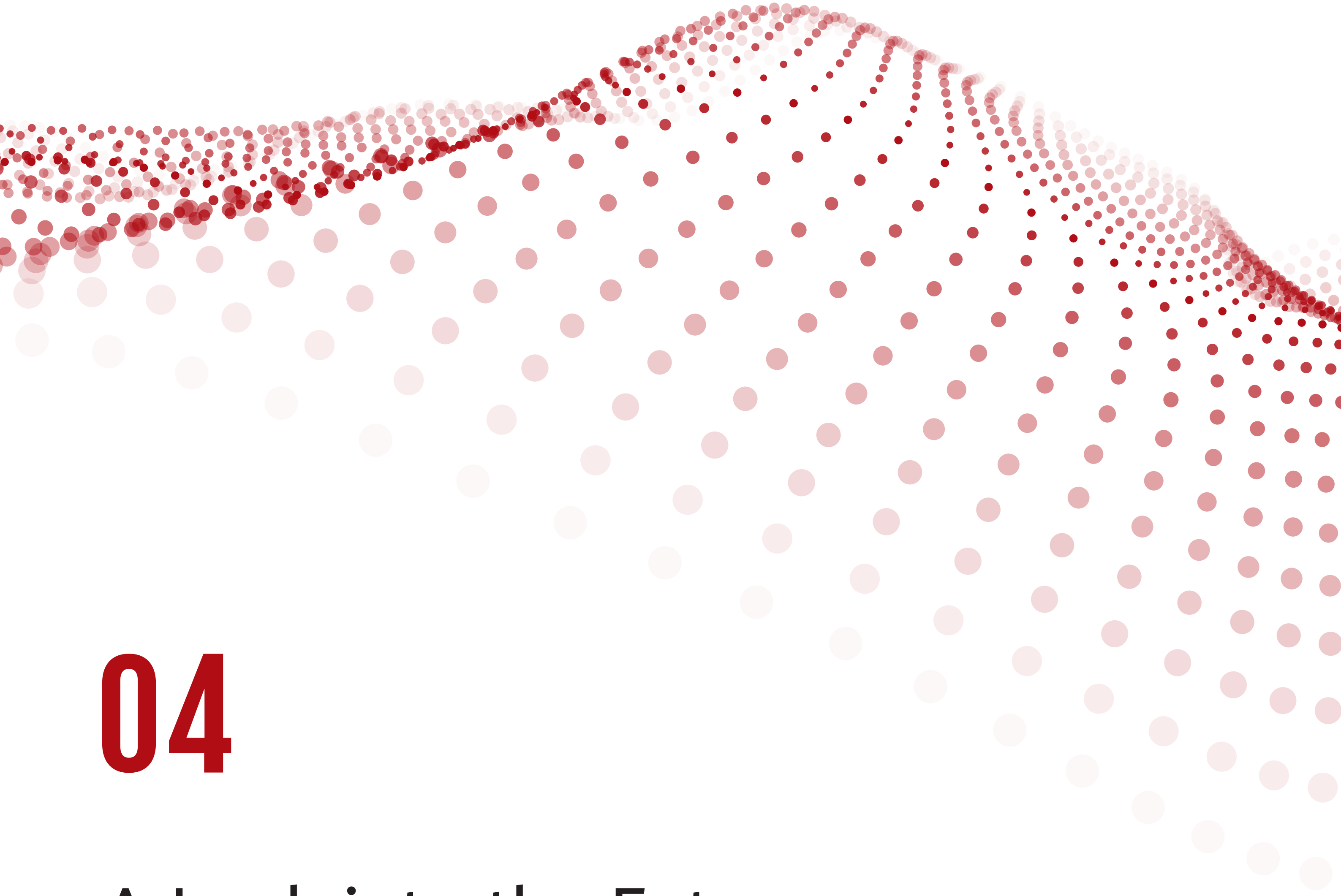
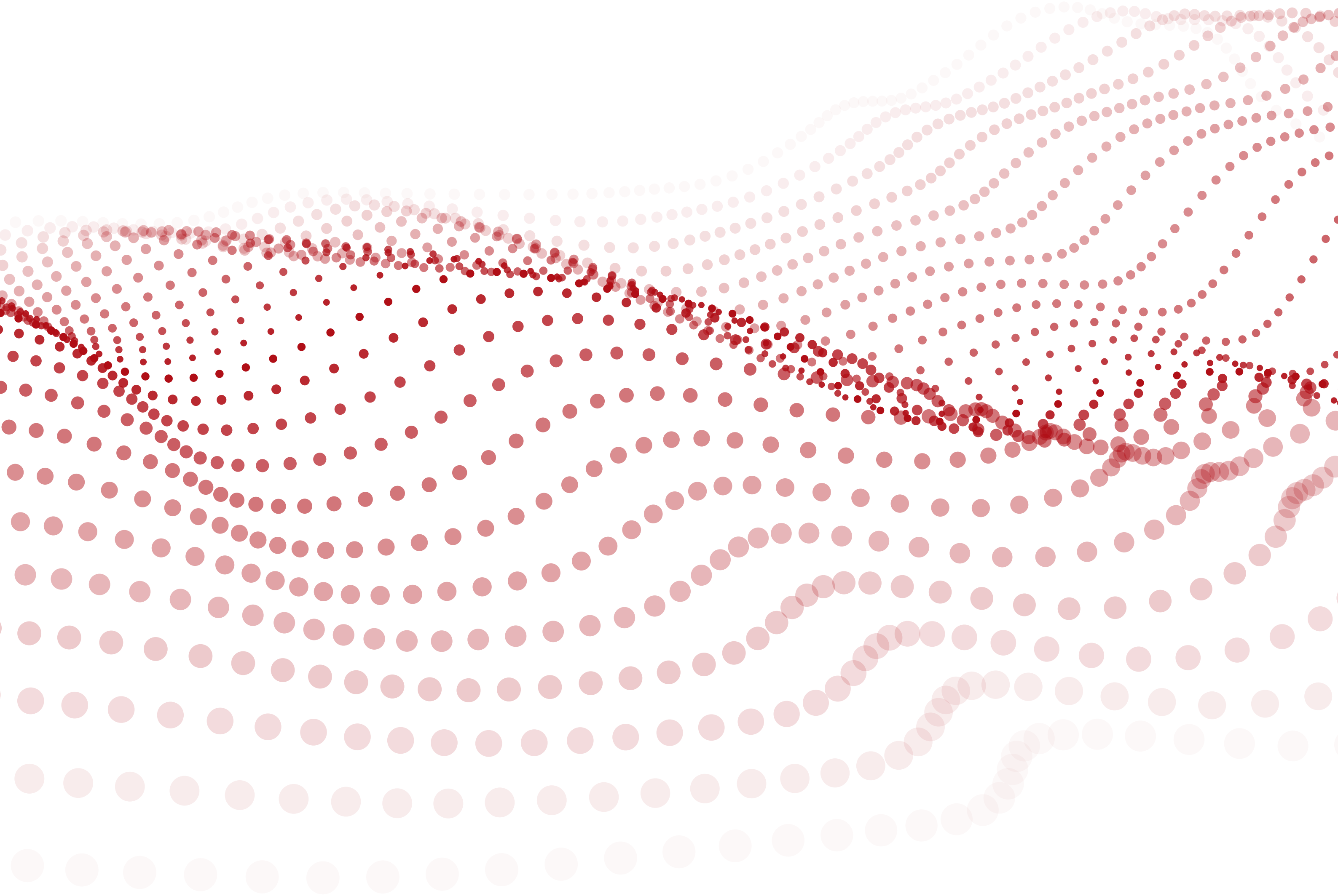


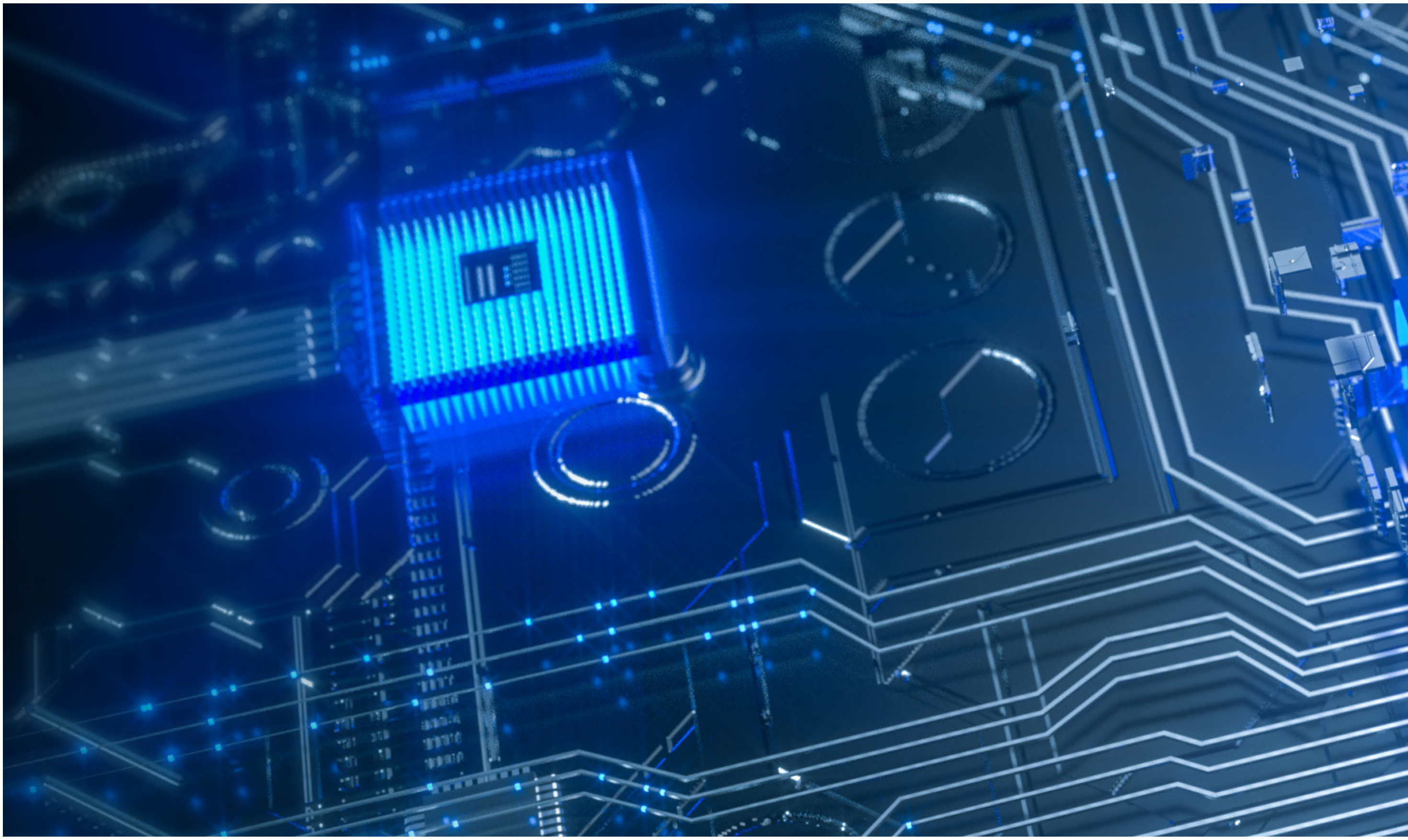
Figure: Using the Pangu GNN model to optimize the coking process



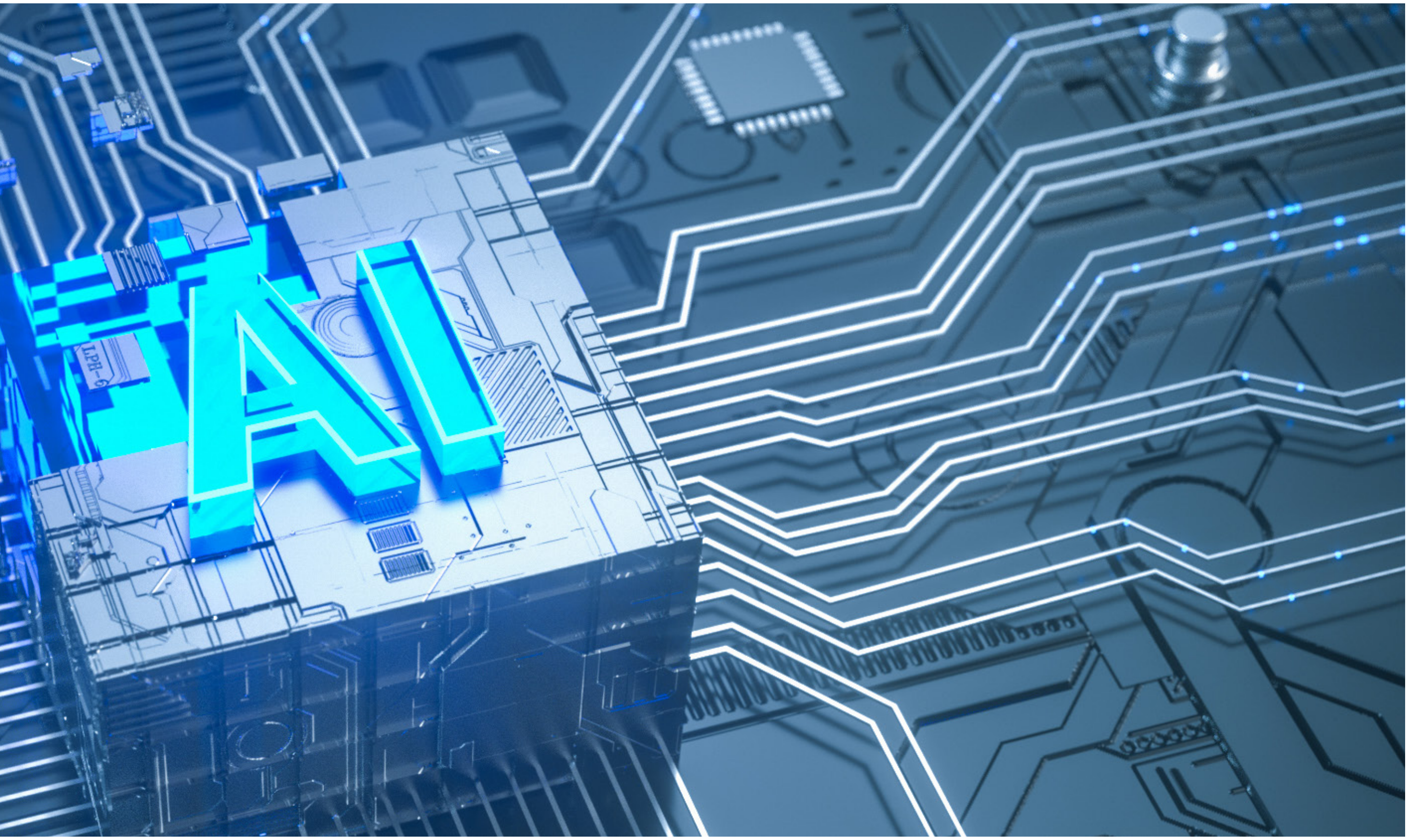
04

A Look into the Future: Pre-Trained Models Face Both Opportunities and Challenges





Humans are not only good at analyzing things—we are also good at creating things. We write poetry, design products, develop games, and crank out code. But today, machines are almost as good as or even better than humans at some creative tasks. Large models are getting good at creating plausible, beautiful works of art, and we can call it AIGC (or AI generated content). The next step is AIGA (or AI generated action). With AI, we can generate not only content, but also actions. AI can interact with the environment and act on the physical world to achieve intended goals, which has allowed us to see some early sparks of Artificial General Intelligence (AGI). Huawei Cloud is laser-focused on helping customers drive digital transformation. In the first stage of digital transformation, data lakes and data warehouses help us sense data, store large amounts of structured and unstructured data cheaper and more reliably, and also do some basic data processing and cleaning. In the second stage, deep learning neural networks and knowledge graphs transform data into knowledge, helping us dig deeper into our data and extract deeper, more actionable insights. With the development of pre-trained models such as GPT-3 and ChatGPT, we have ushered in a new moment — AIGA. Today, the development of large language models has enabled more general skills, helping us get better at controlling robots, or characters in games, or at orchestrating large ERP software systems; and digital twins help us interact directly with the environment.



To generate a sustainable competitive advantage, leading companies need to act now to build an AI "flywheel" powered by domain-specific data and pre-trained models. Very much like a flywheel, AI takes a lot of energy to get started – infrastructure, talent, data, operating environment, and more. But once the wheel begins turning, it gets a lot easier to keep it turning — by feeding it with a continuous supply of new data and continuous iteration of algorithms and technologies. With this flywheel, we can build up momentum to keep accelerating AI development and innovation and keep pushing towards AGI, and also continuously improve our products by enhancing them with cutting-edge AI features.

HUAWEI TECHNOLOGIES CO., LTD.
Huawei Industrial Base
Bantian Longgang
Shenzhen 518129, P. R. China
Tel: +86-755-28780808
www.huawei.com

Trademark Notice
HUAWEI , HUAWEI, are trademarks or registered trademarks of Huawei Technologies Co., Ltd.
Other Trademarks, product, service and company names mentioned are the property of their respective owners.

General Disclaimer
The information in this document may contain predictive statement including, without limitation, statements regarding the future financial and operating results, future product portfolios, new technologies, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.