Translation

The following white paper by a Chinese state-affiliated think tank emphasizes the importance of AI frameworks in the current and future phases of the development of AI technology. The white paper references well-known foreign frameworks such as TensorFlow and PyTorch, and also goes into considerable detail on applications of Chinese platforms such as Baidu’s PaddlePaddle and Huawei’s MindSpore.

Title
White Paper on AI Framework Development (2022)
AI 框架发展白皮书（2022 年）

Author
China Academy of Information and Communications Technology (CAICT; 中国信息通信研究院; 中国信通院). CAICT is a think tank under China’s Ministry of Industry and Information Technology (MIIT; 工业和信息化部; 工信部).

Source
CAICT website, February 26, 2022.

The Chinese source text is available online at:
http://www.caict.ac.cn/kxyj/qwfb/bps/202202/P020220226369908606520.pdf
An archived version of the Chinese source text is available online at: https://perma.cc/N2F6-27U9
U.S. $1 ≈ 7.1 Chinese Yuan Renminbi (RMB), as of June 6, 2023.

Translation Date	**Translator**	**Editor**
June 6, 2023	Etcetera Language Group, Inc.	Ben Murphy, CSET Translation Manager

Preface

AI helps the current economy and society enter the intelligent economy (智能经济) era. The world is entering a time of reshaping driven by a new generation of information technology (IT). As an important enabling technology, artificial intelligence (AI) has a strong "lead goose effect" that spills over to drive the activation of the real economy. This makes it of great significance for building up national science and technology (S&T) influence. AI has become a new technology hotspot in countries around the world, and the construction of AI infrastructure has also become an important key element and focus. The next ten years will be a golden age for the global development of the digital economy and our entry into the intelligent economy and society era. Focusing on the development of AI infrastructure will provide a strong pull for the growth of China's AI industry and the flourishing development of the digital economy.

AI frameworks are the operating systems for the intelligent economy era. As a basic tool in the AI development stage, AI frameworks assume the role of operating systems in the AI technology ecosystem and are an important vehicle for AI academic innovation and industrial commercialization, helping AI move from theory to practice and quickly enter the era of scenario-specific applications. They are also one of the infrastructures necessary for the development of AI. As their importance continues to increase, AI frameworks have become one of the focuses of innovation in the AI industry, attracting the attention of academic and industry circles.

In this context, this white paper aims to clarify the meanings of concepts, the evolutionary process, the technical system, and the significance of the role of AI frameworks. We will do this by summarizing the current development status of AI frameworks, studying and making judgments about the development trends of AI framework technologies, and putting forward prospects and path suggestions for the development of AI frameworks. Since AI frameworks are still in a stage of rapid development, we still need to gain a deeper understanding of AI frameworks. Therefore, we welcome criticism and corrections of any deficiencies in this white paper.

Contents

I. The continuous evolution of AI framework technology has formed a relatively complete system 6

(1) The evolution of AI frameworks has entered the deepening stage 6

(2) AI framework technology has evolved into three levels 9

(3) The importance of AI frameworks is becoming increasingly obvious 16

II. Global AI frameworks are flourishing and developing, and a trend of diversified competition and cooperation is gradually emerging 17

(1) In terms of supply entities, enterprises and schools are the most active contributors 17

(2) In terms of the open-source ecosystem, the world has entered an active period 18

(3) In terms of market structure, two dominant players continue to lead 20

(4) In terms of supporting applications, scientific research and industry are both driving forces 21

(5) In terms of promotion methods, the three roads go in the same direction 25

III. Facing the diverse challenges of the future, AI frameworks show six major technical trends 27

(1) Full-coverage development: AI frameworks will focus on unifying front-end convenience and back-end efficiency 27

(2) Full-scenario: AI frameworks will support device-edge-cloud full-scenario cross-platform device deployment 28

(3) Hyperscale: AI frameworks will focus on strengthening support for hyperscale AI 29

(4) Scientific computing: AI frameworks will further integrate and interact with scientific computing 30

(5) Secure and trusted: AI frameworks will help improve the interpretability and robustness of AI models 31

(6) Engineering: AI frameworks will accelerate the implementation of industrial-scale engineering of AI applications 32

IV. The AI framework ecosystem is far from mature, so it has considerable room for future development 34

(1) Evolve from hardware adaptation to operator API standardization 34

(2) Strengthen the creation of open-source communities and an open-

source atmosphere 35
(3) Focus on wide-ranging and open cooperation with universities and scientific research institutes 35
(4) Promote the implementation of integration into the AI infrastructure layout 35
(5) Support the in-depth enablement of large models and scientific computing 36

List of Figures

Figure 1 Technological evolution of AI frameworks 7
Figure 2 Core technological system of AI frameworks 9

List of Tables

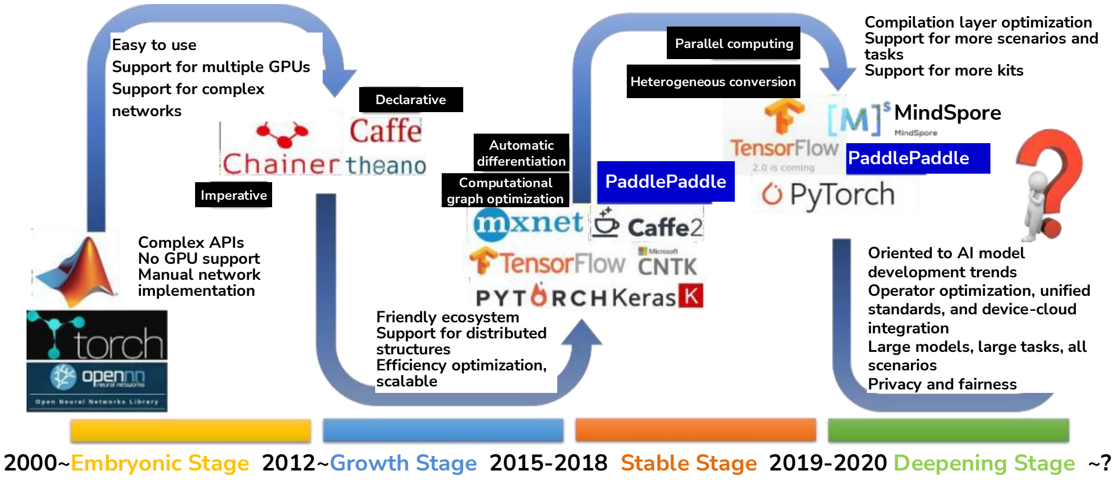
Table 1 Mainstream AI frameworks in the GitHub community (Jan 2022) 18
Table 2 Mainstream AI frameworks in the Gitee community (Jan 2022) 20

I. The continuous evolution of AI framework technology has formed a relatively complete system

An AI framework is a set of standard application programming interfaces (APIs; 标准接口), feature libraries, and toolkits for AI algorithm model design, training, and verification. It integrates algorithm encapsulation, data invocation, and computing resource usage. At the same time, it provides developers with a development interface and an efficient execution platform, which are essential tools for AI algorithm development at the current stage. At present, research and innovation in basic AI algorithm theory are becoming increasingly vibrant, and deep neural networks are becoming increasingly mature. One after another, major producers have invested in the engineering implementation of deep neural network algorithms and made efforts to build algorithm model tools. Then, they package these capabilities into a software framework for developers to use. In this process, AI frameworks (also known in the industry as AI development frameworks and deep learning frameworks) came into being. The AI framework is responsible for providing developers with mathematical operations for building neural network models, converting complex mathematical expressions into computer-recognizable computational graphs, automatically training neural networks, obtaining a neural network model for solving classification and regression problems in machine learning, and implementing [AI capabilities] in application scenarios such as object classification and speech recognition.

(1) The evolution of AI frameworks has entered the deepening stage

Incorporating the AI development process and the technical characteristics of AI frameworks, the development of AI frameworks can be roughly divided into four stages: the embryonic stage (early 2000s), the growth stage (2012 to 2014), the stable stage (2015 to 2019), and the deepening stage (after 2020). The development of AI frameworks is closely related to AI, especially the emergence and development of neural network technology.


Easy to use
Support for multiple GPUs
Support for complex networks
Declarative
Caffe
Chainer
theano
Imperative
Complex APIs
No GPU support
Manual network implementation
torch
opennn
Automatic differentiation
Computational graph optimization
mxnet
Caffe2
TensorFlow
CNTK
PYTORCH
Keras
Friendly ecosystem
Support for distributed structures
Efficiency optimization, scalable
PaddlePaddle
Parallel computing
Heterogeneous conversion
Compilation layer optimization
Support for more scenarios and tasks
Support for more kits
TensorFlow
MindSpore
PaddlePaddle
PyTorch
Oriented to AI model development trends
Operator optimization, unified standards, and device-cloud integration
Large models, large tasks, all scenarios
Privacy and fairness
2000~Embryonic Stage
2012~Growth Stage
2015-2018 Stable Stage
2019-2020 Deepening Stage
~?

Source: CAICT

Figure 1　Technological evolution of AI frameworks

Embryonic stage: Due to insufficient computing power, the influence of neural network technology was relatively limited at this stage, so some traditional machine learning tools emerged to provide basic support. This was the prototype for AI frameworks, but these tools were not always specifically customized for neural network models, and extremely complex APIs were not friendly to developers. In addition, these tools did not support graphics processing unit (GPU) computing power. **The AI frameworks at this stage were not perfect, and developers had to do a lot of basic work, such as handwritten backpropagation, building network structures, and designing their own optimizers.**

Growth stage: In 2012, Alex Krizhevsky and others proposed a deep neural network architecture, the famous AlexNet, which achieved the best accuracy on the ImageNet dataset and left the second-place finisher in the dust. This set off a boom in deep neural networks. This has provided significant impetus to the development of AI frameworks. Representative early AI frameworks such as Caffe, Chainer, and Theano emerged to help developers easily build complex deep neural network models, such as convolutional neural network (CNN), recurrent neural network (RNN), and long short-term memory (LSTM) networks. In addition, these frameworks also supported multi-GPU training, making it possible to carry out larger and deeper model training. **In this stage, the AI framework system was initially formed, and the declarative style (声名式) and imperative style (命令式) opened up two different development paths for subsequent AI frameworks.**

Stable stage: In 2015, ResNet proposed by Kaiming He (He Kaiming; 何恺明) and others once again broke through the image classification boundary, with its accuracy rate on the ImageNet dataset setting a new record high. This finally formed a consensus in industry and academic circles that deep learning would become the next major technology trend. In the past one to two years, Google has open-sourced the famous TensorFlow framework, which remains the most popular AI framework in the field of machine learning. The inventor of Caffe joined Facebook (now renamed Meta) and released Caffe2. At the same time, the Facebook AI Research (FAIR) team also released another popular framework, PyTorch, which is an extension of the Torch framework, but uses the more popular Python APIs. Microsoft Research developed the Cognitive Toolkit (CNTK) framework. Amazon adopted MXNet, a joint academic project of the University of Washington, Carnegie Mellon University (CMU), and other institutions. In China, Baidu took the lead by laying out the PaddlePaddle deep learning framework, which it released in 2016.

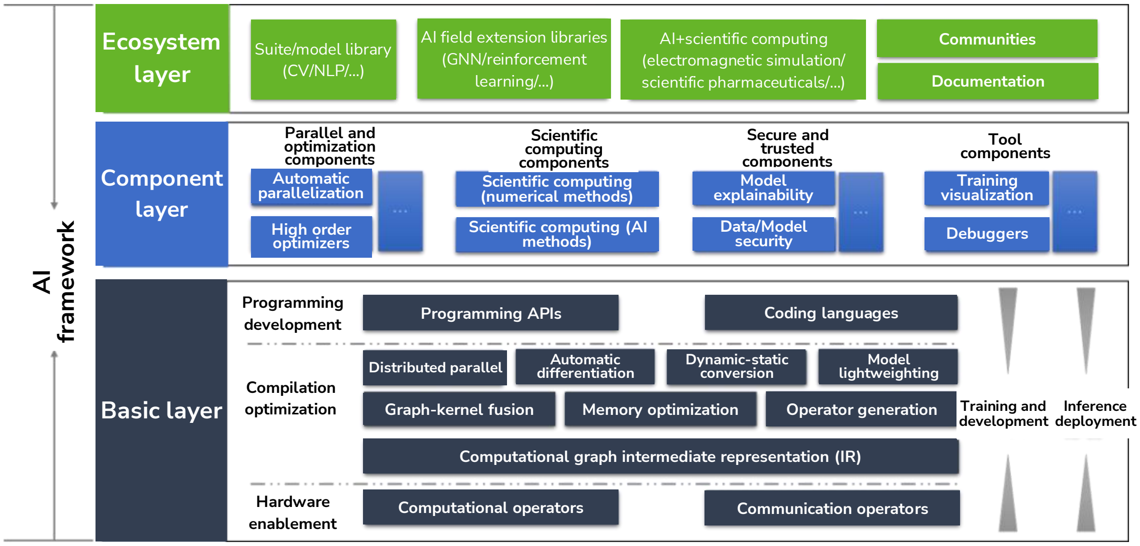
TensorFlow and CNTK borrow from Theano's declarative programming style, while PyTorch inherited Torch's intuitive and developer-friendly imperative programming style. Francois Chollet almost single-handedly developed the Keras framework, which provides more intuitive high-level abstractions of neural networks and building blocks. At the same time, various AI frameworks continued to be iterated, providing the frameworks with various core components for efficient and friendly development. These include the automatic differentiation capability supported by almost all AI frameworks, the distributed-version AI framework and the iOS system support capabilities provided by TensorFlow, and the full embrace of Python including support for optimizers, library functions, and API tools on the part of PyTorch. **AI frameworks welcomed in a time of prosperity, and on the basis of continuous development, various frameworks continued to iterate and were naturally selected by developers.**

After fierce competition, two camps finally formed, with TensorFlow and PyTorch dividing the field between them. In 2019, the Chainer team transferred their development work to PyTorch; Microsoft stopped active development of the CNTK framework, and some team members switched to support PyTorch; and Keras was incorporated into TensorFlow and became one of its advanced APIs in TensorFlow 2.0.

Deepening stage: With the further development of AI, new trends continue to emerge, such as the appearance of ultra-large-scale models (such as Generative Pre-trained Transformer 3 [GPT-3]), which place higher requirements on AI frameworks. With the expansion of AI application scenarios and the acceleration of the process of cross-integration with more fields, more and more requirements have been put forward for frameworks, such as support for full-scenario multi-tasking and demands for high computing power. This requires AI frameworks to maximize compilation optimization, better use and mobilize computing power (“compute”), and fully utilize the potential of hardware resources. In addition, the pain points of AI and social ethics have also prompted the progress of trustworthy AI at the framework level. Based on the above context, the existing popular frameworks are exploring development directions of next-generation AI frameworks. For example, Huawei launched MindSpore (昇思) in 2020, which made certain breakthroughs in full-scenario collaboration and trustworthiness, and Megvii (旷视) launched MegEngine (天元), which has a deep layout in the integration of training and inference. **At this stage, AI frameworks are making in-depth explorations of technical features such as full-scenario support, ultra-large-scale AI, and security and trust, and constantly achieving new breakthroughs.**

(2) AI framework technology has evolved into three levels

Based on the process stages and positioning of technologies, the core technology of the current mainstream AI frameworks can be divided into a basic layer, component layer, and ecosystem layer.



Source: CAICT

Figure 2 Core technological system of AI frameworks

1. Basic layer

The basic layer implements the most basic and core functions of the AI framework. It contains three sub-layers: programming development, compilation optimization, and hardware enablement (硬件使能). The programming development layer is a window through which developers interact with the AI framework. It provides developers with APIs for building AI models. The compilation optimization layer is a key part of the AI framework that is responsible for completing the compilation optimization of the AI model and scheduling hardware resources to complete computation. The hardware enablement layer is the channel that connects the AI framework and AI computing power hardware, helping developers shield the technical details of the underlying hardware.

Programming development – Programming APIs: Developers describe the calculation processes of algorithms by calling programming APIs. For developers, the ease of use of the programming APIs and their expressive capabilities are very important. Their description of algorithms can be mapped to computational graphs. Programming APIs can be mainly divided into three categories: One is programming APIs based on data stream diagrams. Popular machine learning programming frameworks based on data stream diagrams include TensorFlow, MXNet, Theano, and Torch7. Another type is layer-based programming APIs, such as Caffe. The third type is algorithm-based programming APIs, which are mainly used for the implementation of traditional machine learning algorithms, such as Scikit-Learn.

Programming development – Coding languages: There are many AI application scenarios, and AI developers choose to use different programming languages based on different scenarios. A complete AI framework should support multiple different languages, such as Python, Cangjie,[1] and Julia. The AI framework needs to provide development services and technical support with the same functions and comparable performance for developers using different programming languages.

Compilation optimization – Distributed parallel (分布式并行): This refers to strategies such as data stream parallelism, model parallelism, pipeline parallelism, and optimizer parallelism. As the size of models increases, traditional data parallelism cannot be effectively processed, so the use of automatic parallelization technology will become the norm. It is necessary to divide large models among different devices. Segmentation divides various large block calculations into multiple small pieces, assigns the calculation of the small pieces to different computing resources, and finally reduces and merges the structures of the small calculations. However, it is very difficult to optimize the segmentation strategy. The data transmission volumes generated by different segmentations vary greatly, and the computation utilization rates are also very different. For example, pipeline parallelism often meets with major challenges in computing utilization. Operator segmentation parallelism presents a greater challenge in terms of data transmission volumes and requires the support of an AI framework.

Compilation optimization – Automatic differentiation: Automatic differentiation decomposes a complex mathematical operation process into a series of simple basic operations, and each basic operation can be obtained via a lookup table. There are two forms of automatic differentiation: forward-mode and reverse-mode. Forward mode

[1] Translator’s note: Cangjie (仓颉) is a programming language that, as of 2022, Huawei was developing to complement its HarmonyOS (鸿蒙) operating system.

calculates the differential while the computational graph is forward propagating. Reverse mode requires a forward-mode automatic differentiation to have been performed on the computational graph. It obtains the output value and then performs backpropagation. Therefore, the memory overhead of reverse mode is a bit larger. It needs to save the intermediate variable values during forward propagation, and these variable values are used to calculate the derivative during backpropagation.

Compilation optimization – Dynamic-static conversion (动静转换): Static graphs define all operations and network structures before execution and present them to sensor streams (传感器流). This provides higher performance during training, but at the cost of being less easy to use and less flexible. Dynamic graph computations are performed instantly, providing greater flexibility and facilitating debugging, but at the cost of lower performance. TensorFlow2.0, MindSpore, and other frameworks all support dynamic graph and static graph conversion technology, allowing them to achieve a balance between computing efficiency and flexibility.

Compilation optimization – Model lightweighting: Lightweighting refers to the lightweighting technology configured by the AI framework to meet the requirements of small AI model size, low computational complexity, low battery consumption, and flexible deployment of updates. Generally speaking, model lightweighting refers to model compression and acceleration. Compression focuses on reducing the number of network parameters, while acceleration focuses on reducing computational complexity and improving parallel capabilities. Algorithm layer compression and acceleration mainly include structural optimizations (such as matrix decomposition, grouped convolution, and small convolution kernels), quantization and conversion to fixed-point, model pruning, and model distillation. Framework layer acceleration mainly includes compilation optimization, cache optimization, sparse memory and computing (稀疏存储和计算), NEON instruction application, and operator optimization.

Compilation optimization – Graph-kernel fusion (图算融合): By automatically analyzing and optimizing the existing network computational graph logic and incorporating the capabilities of the target hardware, the computational graph is optimized for calculation simplification and substitution, operator splitting and fusion, and operator specialization (算子特例化) compilation to improve the utilization of device computing resources and achieve overall optimization of network performance. Compared with traditional optimization technologies, graph-kernel fusion possesses unique advantages such as multi-operator cross-boundary joint optimization (多算子跨边界联合优化), cross-layer collaboration with operator compilation, and real-time polyhedral-based compilation of operators. In addition, graph-kernel fusion only requires developers to open the corresponding configuration to automatically complete the entire optimization process without additional perception by network developers.

This allows developers to focus on network algorithm implementation.

Compilation optimization – Memory optimization: Due to the limited memory resources of the hardware system, and especially the limited memory resources of AI chips, an efficient memory optimization strategy is required to reduce the consumption of the system memory by the AI network. Commonly used memory optimization techniques include static memory reuse optimization (静态内存复用优化) and dynamic memory allocation mechanisms. Static memory reuse optimization analyzes the data stream relationship of computational graphs and plans the memory reuse strategy for the data based on the memory usage size of the data and lifecycle overlaps between the data so as to minimize memory usage. Dynamic memory allocation mechanisms create a large block of memory at runtime and provide memory slices according to memory needs during the actual operator execution process. When an operator execution ends and the related data is no longer referenced, the memory slice is released so as to achieve efficient memory reuse.

Compilation optimization – Operator generation: The AI framework provides basic and commonly used operators, but these operators often cannot meet the needs of developers for continuous algorithm evolution. Therefore, the AI framework must have the ability to generate and optimize unified operators for different computing devices. This way, developers only need to write high-level programming languages (such as domain-specific languages [DSLs]) to generate high-quality underlying operators through the operator compilation and generation capabilities provided by the AI framework. This greatly reduces the development and maintenance costs of the AI framework and hardware platform and expands the scope of application.

Compilation optimization – Intermediate representation: Intermediate representation (IR) is the definition of the computational graph and operator format. A complete IR needs to support operator definitions of different hardware devices and performance optimization for computational graphs, support the flexible expression of different types of AI model network structures, and support model transfer and migration between different devices.

Hardware access – Computational operators: In the field of deep learning, a computational operator refers to a function node in the computational graph. It is a computation operation performed on tensors, which accepts zero or more tensors as input and obtains zero or more tensors as output using the gradient, divergence, and curl expressions to perform computation.

Hardware access – Communication operators: Function nodes (函数节点) for distributed node (分布式节点) communication.

2. Component layer

The component layer mainly provides configurable higher-order functional components for the AI model lifecycle to allow for the optimization and improvement of performance in specific fields, including compilation optimization components, scientific computing components, security and trust components, and tool components. These components are visible to AI model developers.

Parallel and optimization components – Automatic parallelization: This refers to the support for a wide range of automatic parallelization technologies and their combinations. The AI framework allows developers to combine various parallelization technologies and form hybrid parallelization strategies as needed, such as the combination of data stream parallelization and model parallelization, or data stream parallelization and pipeline parallelization. It also allows developers to choose their own parallelization strategies individually, with a more flexible manner of support for AI model training and application adaptation.

Parallel and optimization components - Higher-order optimizers: The AI framework supports a variety of different first-order/second-order optimizers, which can provide developers with flexible and convenient APIs, such as the stochastic gradient descent (SGD) optimizer, stochastic gradient descent with momentum (SGDM) optimizer, Nesterov accelerated gradient (NAG) optimizer, AdaGrad optimizer, AdaDelta optimizer, Adam optimizer, and Nadam optimizer.

Scientific computing components - Scientific computing (numerical methods): One of the important directions of AI development is scientific computing. Therefore, AI frameworks are required to provide functional support relevant to scientific computing for developers and provide an integrated expression method for AI+scientific computing through the functional programming paradigm. This will allow developers to better approximate mathematical computing methods when programming. This will alleviate the situation where the programming APIs of current AI frameworks are mainly designed for deep neural networks, while scientific computing requires the expression of a large number of mathematical formulas (such as solving differential equations).

Scientific computing components - Scientific computing (AI methods): In order to allow AI methods to directly replace the numerical methods when obtaining computation results, AI frameworks need to have a unified data foundation for "AI+scientific computing" that converts traditional scientific computing data inputs (such as the simulation data generated by traditional scientific computing software) into the AI framework data inputs (i.e., tensors). In order to obtain the form of calculation results when using a combination of AI methods and numerical methods, in addition to the need for a unified data engine, AI frameworks need to support traditional numerical calculation methods, such as higher-order differential solutions and linear algebra calculations. They must perform hybrid computing optimization of traditional numerical methods and AI methods via computational graphs to achieve end-to-end acceleration of "AI+scientific computing."

Secure and trusted components - AI explainability: The AI framework needs to possess three levels of capabilities to support explainable AI. Ensure "data explainability" before modeling, analyze the data distribution, find representative features, and select the required features for modeling during training. Construct an "explainable AI model," which complements the AI structure by incorporating traditional machine learning (such as Bayesian probabilistic programming) and balances the effectiveness of learning results and the explainability of learning models. Perform "interpretability analysis" (“解释性分析”) on the constructed model by analyzing the input, output, and intermediary information of the AI model to perform relationship analysis (such as the TB-Net method) and verify the logic of the model.

Secure and trusted components – Data security: Data security issues in the AI field not only involve the protection of the original data itself, but also the prevention of the deduction of key private data information through model inference results. Therefore, in addition to providing data asset protection capabilities, the AI framework itself also needs to protect the privacy of model data through differential privacy and other methods. At the same time, in order to protect data security at the source, the AI framework conducts model training through federated learning and other methods so that the model can be trained and updated without data leaving the system.

Secure and trusted components – Model security: Insufficient sample training during model training will lead to a model with insufficient generalization capabilities. This means that models will not give correct judgment results when faced with malicious samples. To prevent this, the AI framework first needs to provide a wealth of AI robustness detection tools and test the robustness of AI models through black-box, white-box, and gray-box testing and other adversarial detection technologies, such as static structural analysis and dynamic path analysis. Second, the AI framework can help developers improve the model robustness by supporting network distillation and adversarial training.

Tool components – Training visualization: This supports the visualization of the training process. You can directly view the core content of the training process on a page, including training scalar information, parameter distribution diagrams (参数分布图), computational graphs, data graphs, data sampling, and other modules.

Tool components – Debuggers: Numerical errors, such as infinities, often occur in neural network training. Developers want to analyze the reasons that prevent convergence in training. However, since computations are encapsulated as a black box and executed in the form of a graph, it is difficult for developers to locate errors. A debugger is a tool for training debugging. Debuggers allow developers to view the internal structure of the graph and the input/output of nodes during the training process, such as viewing the value of a tensor or viewing the Python code corresponding to a node in the graph. In addition, developers can also select a group of nodes to set conditional breakpoints and monitor the calculation results of nodes in real time.

3. Ecosystem layer

The ecosystem layer is mainly oriented to application services. It supports the application, maintenance, and improvement of various AI models developed based on the AI framework and is visible to developers and application personnel.

Suite/Model library: AI frameworks should provide pre-trained models or defined model structures for general tasks for various fields. This way, developers can obtain and carry out AI model training and inference, such as for computer vision (CV) and natural language processing (NLP).

AI field extension libraries: AI frameworks must be able to provide support for tasks in a wide range of fields and provide typical cases for relevant tasks. This allows frameworks to provide better application services, such as graph neural networks (GNN), reinforcement learning, and transfer learning.

AI+scientific computing: Different from traditional information fields such as CV and NLP, solving scientific computing problems requires relatively specialized domain knowledge. In order to accelerate the research and implementation of AI + scientific computing fusion, AI frameworks need to provide easy-to-use scientific computing suites for different scientific computing fields (such as electromagnetic simulation, scientific pharmaceuticals, energy, meteorology, biology, and materials). These suites contain high-quality domain datasets, high-precision basic AI models, and a collection of tools for pre- and post-processing

Documentation: AI frameworks should provide a complete documentation system, including but not limited to framework description documentation, framework API documentation, framework version change documentation, framework frequently asked question (FAQ) documentation, and framework feature documentation.

Communities: The development of AI services requires the support of communities. AI frameworks should operate or maintain a good community environment. A good AI framework has good maintainability and ease of use. At the same time, there should be representative projects in the AI framework community and long-term support based on the projects and applications of the framework.

(3) The importance of AI frameworks is becoming increasingly obvious

The AI framework inherits from the past and opens the door to the future. It is the core of the entire AI technology system. From the perspective of its functional positioning in the technical system, AI frameworks call underlying hardware computing resources, so they can shield the underlying differences and provide good execution performance. They support the establishment of AI application algorithm models and provide a standard environment for algorithm engineering implementation. They are the critical core of the AI technical system. In addition to completing the engineering implementation of AI algorithms, AI frameworks can also greatly improve the efficiency of AI learning and strengthen the capabilities of AI algorithm models. For example, TensorFlow-based AlphaGo learned the skills to defeat its predecessor AlphaGo in a very short time.

AI frameworks are technological weapons to cope with the intelligent economy era. Large-scale parallel computing and intelligent applications are the main features of the future intelligent economy era. Current hardware computing (硬件计算) is represented by central processing units (CPUs), and the software stack is mainly optimized for serial instructions (串行指令). Since AI algorithms involve a large number of matrix calculations and parallel numerical calculations, hardware computing for the intelligent economy era already shows a trend of migrating from serial to parallel computing. In the future, GPUs may be the representative computing hardware, and the software stack will primarily be optimized for massively parallel computing. In this case, **AI frameworks will become key schedulers for massively parallel computing.** In addition, AI models will dominate the specific scenarios of various industries in the intelligent economy era, and intelligent applications will present characteristics such as scale and depth. **AI frameworks are the key supporters of the rapid implementation of intelligent applications.**

AI frameworks will become the operating systems for the intelligent economy era. In the current Internet era, operating systems are the core hubs of the IT industry. They establish the connections between hardware and application software and control the entire ecosystem of digital devices. Through deep binding with general purpose computing chips, two stable technology system patterns have been formed: Windows+Intel and Android/iOS+ARM. In the intelligent economy era, AI frameworks assume the role of operating systems in the AI technology ecosystem and are an important vehicle for AI academic innovation and industrial commercialization, helping AI move from theory to practice and quickly enter the era of scenario-specific applications. In general, **the combination of "AI framework + compute chip" determines the main technical route of AI industry applications to a certain extent**. Its research and development (R&D) can promote the development of related and peripheral chips, systems, software and hardware platforms, and other industries in the ecosystem, thereby promoting the construction of the core AI ecosystem. As their value continues to increase, **AI frameworks have become one of the focuses of innovation in the AI industry, attracting the attention of academic and industry circles**.

II. Global AI frameworks are flourishing and developing, and a trend of diversified competition and cooperation is gradually emerging

(1) In terms of supply entities, enterprises and schools are the most active contributors

S&T enterprises and elite universities are the most active contributors to the development and maturity of AI frameworks. Digital S&T giants and top universities are the main maintenance force for the development and strengthening of AI frameworks. Creating a technology industry ecosystem and creating an academic innovation atmosphere are the source of power for these two main entities. Individuals and open-source organizations also play an important role and are important embodiments of the innovative and beneficial-to-the-public (公益性) characteristics of AI frameworks.

Digital S&T giants are the core force for the development and strengthening of AI frameworks. The needs of their own AI business scenarios stimulate the application of AI frameworks and achieve AI framework verification and improvement. World-renowned digital S&T giants dominate the open-source AI framework technology ecosystem, and Chinese digital S&T companies have also actively deployed and constantly innovated in recent years. Foreign digital S&T giants such as Google, Meta, Microsoft, and Amazon have first-mover advantages in the R&D of basic algorithm frameworks. Relying on their own AI business scenarios and massive data resources, they can effectively test and verify algorithm frameworks and improve their functions. On this basis, digital S&T giants open source the AI frameworks that originally served their internal business scenarios, provide underlying AI core capabilities to downstream partners in the industry chain, meet the needs of industrial-level applications, gradually improve the overall ecosystem, and achieve mutually beneficial cooperation. Chinese digital S&T giants have launched AI frameworks one after another. In addition to meeting their own AI application needs, they are also expanding external service provision, such as Huawei MindSpore, Baidu PaddlePaddle, Tencent TNN, Alibaba MNN, ByteDance (字节跳动) BytePS, and Xiaomi (小米) Mace.

Universities and scientific research institutes are among the leading forces that initiated the R&D of AI frameworks, and they continue to play an active role. Universities and scientific research institutes have strong human resources. Based on the needs of laboratory scientific research and innovation, they carry out basic theoretical research work on AI frameworks. Overall, their layout in this field was earlier than that of digital S&T companies, so it was easier for them to achieve revolutionary breakthroughs and innovations. Open-source frameworks such as Theano and Caffe first launched by universities can meet the needs of academic research and played a huge role in promoting the overall development of AI frameworks. However, their performance in scenarios such as large-scale distributed computing is not as good as the AI frameworks launched by enterprises. Subsequently, universities continued to show the value of the role they play by taking over as the maintenance entities. For example, the MXNet framework was started at Carnegie Mellon University and later donated to the Apache Foundation. It has now become Amazon Web Services (AWS), the most important AI framework. Chinese universities are increasingly paying attention to R&D on AI frameworks. For example, Tsinghua University successively developed the open-source framework Jittor and the Bayesian deep learning algorithm framework "ZhuSuan" (“珠算”).

(2) In terms of the open-source ecosystem, the world has entered an active period

Open source is essentially an aggregation of talents and wisdom that can promote the rapid upgrading of AI frameworks. A robust open-source ecosystem is crucial to the development of AI frameworks. Developers achieve close interaction with open-source AI frameworks through a series of activities such as open-source code, project hosting, collaborative sharing, and communication and exchanges in the open-source community. The open-source community is an essential learning and exchange environment for AI framework developers. You could say that the open-source community plays a huge role in promoting the development of AI frameworks. Relevant indicators of the open-source community also reflect the development of AI frameworks throughout the entire industry. For AI frameworks, the most well-known foreign community is GitHub, an open-source code hosting platform acquired by Microsoft, and a well-known domestic community is Gitee (码云), a code hosting platform promoted by OSCHINA.NET.

Table 1 Mainstream AI frameworks in the GitHub community (Jan 2022)

Rank	Framework	Commits[2]	Fork[3]	Star[4]	Contributors[5]
Foreign Framework					
1	TensorFlow	124494	86300	163000	3056
2	PyTorch	43390	14800	53700	2137
3	Theano (Stop Developing)	28127	2500	9500	352
4	CNTK (Stop Developing)	16116	4400	17100	201
5	MXNet	11776	6900	19800	868
Domestic Framework					
1	MindSpore	37308	514	2700	267
2	PaddlePaddle	33753	4300	17500	524
3	MegEngine	2282	462	4100	32
4	OneFlow	7621	351	3000	99
5	Jittor	1266	235	2300	31

Source: Compiled based on GitHub community data

As the most recognized open-source community in the industry, GitHub is also the code hosting platform that developers of AI frameworks pay the most attention to. From the perspective of GitHub indicators, in terms of foreign AI frameworks, **all TensorFlow indicators placed it in first place and far ahead of the second place. It is currently the most active and widely used AI framework in the world.** In recent years, the rising star PyTorch, which has performed brilliantly in the academic field, is in second place. Although it occupies a mainstream position at top conferences, it is still slightly inferior to TensorFlow. MXNet's performance is also relatively good, but it is not at the same level as the first two. **In terms of AI frameworks launched by Chinese entities, MindSpore is currently the most active AI framework.** In terms of

[2] Commits indicate the number of open-source code submissions, representing the activity of open-source projects.

[3] Fork stands for code forking, which represents the reference status of open-source projects.

[4] Star indicates the number of likes, representing the attention paid to open-source projects.

[5] Contributors indicate the number of contributors, representing the scale of contributors to open-source projects.

contributors, it has brought together a user group of a decent size. Baidu's PaddlePaddle was made open source earlier and has certain advantages over other frameworks in terms of attention. Among the other frameworks, OneFlow occupies a leading position in terms of activity and number of contributors.

Table 2 Mainstream AI frameworks in the Gitee community (Jan 2022)

Rank	Framework	Commits	Fork	Star	Contributors
1	MindSpore	38549	2400	6100	774
2	PaddlePaddle	32788	195	3600	561
3	OneFlow	7521	2	1	126
4	MegEngine (mirror image)	2280	6	16	35
5	Jittor	1239	3	11	34

Source: Compiled based on Gitee community data

Gitee, the largest open-source code hosting platform in China, is currently the main platform for the publication and discussion of AI frameworks led by Chinese companies. Except for Megvii's MegEngine, which is a well-known framework in China but has not yet been released in the community, all the other frameworks have been laid out, and they have also attracted groups of domestic developers. Among them, **all of MindSpore's indicators in Gitee far exceed the other AI frameworks. It is the most active, most followed, and most used framework in the Chinese community and holds a leading position in China's open-source ecosystem**.

(3) In terms of market structure, two dominant players continue to lead

From a global perspective, the mainstream international AI framework field is dominated by S&T giants such as Google and Meta. At present, Internet S&T giants represented by Google, Meta, Amazon, and Microsoft are leveraging their own advantages in data, technology, and capital to continue to make efforts in the AI framework ecosystem field and lead the trends in global AI framework technology innovation and upgrading. This situation has gradually formed a duopoly with two dominant players, Google-TensorFlow and Meta-PyTorch. Looking at market share, industry circles are dominated by TensorFlow, while academic circles are dominated by PyTorch. The number of stars they have on GitHub represents the popularity of open-source projects, which is a vivid reflection of the market share of open-source projects in the industry. According to the data in Table 1, TensorFlow now has 163,000 stars, which is much higher than the second-ranked PyTorch (53,700). In 2019, TensorFlow Enterprise was launched to provide large enterprises with an optimized version of TensorFlow and long-term technical support. It was deeply integrated with Google

Cloud services, continuing to consolidate TensorFlow's leading position in the industry. According to Papers With Code data[6], the number of papers based on PyTorch in 2021 accounted for 58.56% of all papers based on AI frameworks, much higher than the second-ranked TensorFlow (12.38%). This shows PyTorch's leading edge in academic circles continues to strengthen.

Looking at China, the AI framework market structure is developing towards diversification under the impetus of the duopoly. China has significant advantages in AI applications, and a considerable number of its AI applications are built on the mainstream international AI frameworks. From the contribution of the underlying open-source code and the adaptation of the underlying hardware, to the iteration of intermediate operator R&D, the improvement of model libraries, and the construction of upper-level algorithm models, the two dominant players continue to export capabilities to the Chinese AI application ecosystem. More than that, the market share of AI frameworks launched by Chinese manufacturers has been steadily increasing over the past two years. After the MindSpore framework was open sourced, it received positive responses from developers in China and abroad, ranking first among tens of thousands of open-source projects on Gitee and becoming the most active open-source AI framework in China. The number of developers using Baidu's PaddlePaddle also continues to grow. According to sample data from 350 developers at small and medium-size enterprises surveyed by International Data Corporation (IDC) in 2021, more than 20% of developers are familiar with PaddlePaddle.

(4) In terms of supporting applications, scientific research and industry are both driving forces

1. AI frameworks empower academic research

The combination of AI and supercomputers has universally raised the computing power in the scientific research field to a new level. Among the top 500 supercomputers in the world in 2021, 68.4% are accelerated by AI technology. Oak Ridge National Laboratory in the United States used TensorFlow to train a 1.1 EFLOP/s extreme weather forecasting model on the Summit supercomputer. This model is used to simulate and predict extreme weather caused by climate change, improving the accuracy and increasing possibilities of meteorological research. The Lawrence Berkeley National Laboratory in the United States has developed the large-scale scientific application CosmoFlow using the TensorFlow framework on a CPU-based high-performance computing platform. Using machine learning plug-ins, the TensorFlow framework has been expanded to an unprecedented scale of over 8,000

[6] https://paperswithcode.com/trends.

nodes. Processing three-dimensional spatial data volumes on this scale is mainly used in dark matter n-body simulation experiments, providing scientists with a new platform to deepen their understanding of the universe.

TensorFlow is widely used in academic scientific research fields. NASA uses TensorFlow to analyze the large amount of data accumulated during the Kepler Mission. Because machine learning can search for signals in a wider range and more efficiently than humans, it discovered the Kepler-90i planet that had been previously overlooked. The discovery makes Kepler-90 the only star system other than our own solar system known to have eight planets orbiting a single star, a major breakthrough in the field of astrophysics. The University of Pennsylvania is using TensorFlow to solve problems related to agricultural diseases and insect pests. It identifies and classifies diseases using a large number of labeled images of cassava plants. Its applications are currently being tested in some areas of Tanzania. It allows farmers to quickly identify diseased plants by waving their mobile phones in front of cassava leaves, and then provides them with information on the best way to manage the disease. Based on TensorFlow, Rainforest Connection, a rainforest protection organization, developed the world's first scalable, real-time monitoring and warning system for tropical rainforest environmental protection that can automatically recognize illegal logging. This system was tested and applied in the Amazon rainforest. Sound samples are sent to the central cloud computing server through the local mobile phone cellular network. TensorFlow is used to analyze and audit the data and screen for sounds related to illegal logging, such as chainsaws and timber trucks, to prevent manual monitoring from missing illegal logging operations.

Chinese frameworks have emerged as rising stars in the field of academic research. As the world's first hundred-billion-level pre-trained large-scale Chinese model, the MindSpore-based PengCheng·PanGu (鹏程·盘古) has a model scale of up to 200 billion parameters. MindSpore adopts a fully automatic parallel training method to support the efficient training of the large PengCheng·PanGu model on 4,096 neural processing unit (NPU) chips. Zidong Taichu (紫东太初) is the world's first tri-modal image-text-audio, hundred-billion parameter pre-trained large model built on the MindSpore framework. It provides cross-modal understanding and cross-modal generation capabilities. Wuhan University used MindSpore to create the world's first dedicated deep learning remote sensing framework, Wuhan.LuoJiaNET (武汉.LuoJiaNET), to achieve intelligent remote sensing interpretation for large-scale satellite remote sensing images. PaddlePaddle partnered with Peng Cheng Laboratory to release PCL-BAIDU Wenxin (ERNIE 3.0 Titan; 鹏城-百度·文心). With 260 billion parameters, it is currently the largest single Chinese model in the world. It has achieved the best results in more than 60 tasks such as machine reading comprehension, text classification, and semantic similarity calculation. In addition, Baidu launched Paddle Quantum (量桨), a quantum machine learning toolset based on PaddlePaddle. This established a bridge between AI and quantum computing. It can quickly realize the construction and training of quantum neural networks and also provide multiple cutting-edge quantum applications.

2. AI frameworks enable industry applications

Airbus uses models developed by TensorFlow for anomaly monitoring to ensure the safe operation of the space station. Airbus provides a number of services for the operation of the Columbus module and its payloads on the International Space Station (ISS). The Columbus module is the largest ISS project of the European Space Agency. Equipped with a variety of experimental equipment, it can conduct various experiments in cell biology, xenobiology, fluid and material science, human physiology, astronomy, basic physics, and more. Composed of multiple components, it is capable of generating approximately 17,000 unique telemetry parameters. Airbus uses models developed by TensorFlow to detect anomalies in the process of data stream monitoring and achieve real-time reporting, which greatly simplifies the anomaly analysis process and shortens the time it takes to resolve them.

The biopharmaceutical leader Celgene relies on MXNet to advance drug research and discovery. Celgene is a pharmaceutical company that works in the immunotherapy field. By training a neural network to identify and make decisions on microscope images with labeled cells, it has solved the problem of using classic image analysis methods to identify and distinguish normal cells and tumor cells on a large scale. The MXNet framework is especially important for toxicology prediction, allowing the virtual analysis of the biological impact of potential drugs without any risk to living patients.

PyTorch helps the mining company Datarock perform drilling based on deep learning. Datarock uses deep learning models to help geologists analyze drill core sample images faster. Traditionally, geologists would carefully analyze these samples centimeter by centimeter to assess their mineralogy and structure, and engineers would look for physical features such as faults, cracks, and rock quality. This was a slow process and prone to human error. Using Datarock's technology, the 5 or 6 hours spent on manual recording can be shortened to half an hour, freeing geologists from time-consuming and repetitive basic work.

MindSpore has made remarkable achievements in industry enablement. It has more than 300 state-of-the-art (SOTA) models, more than 4,000 open-source ecosystem community contributors, supports more than 5,000 online AI applications, and is widely used in industrial manufacturing, finance, energy and

power, transportation, medical care, and other industries. MindSpore empowers industrial manufacturing and uses AI technology to help reduce redundant work. By introducing MindSpore and AI quality inspection algorithms, Huawei's Songshan Lake south zone manufacturing plant has improved its defect detection accuracy for printed circuit boards from 90% to 99.9% and increased the productivity of quality inspectors by a factor of 3. Financial solutions based on MindSpore have achieved remarkable results in bank outlets in Shenzhen, Shanghai, and other places, effectively improving the potential customer conversion rate (潜在客户转化率). At the same time, optical character recognition (OCR) technology and biometric technology are used to convert various documents and ticket texts such as corporate annual reports, contracts, insurance policies, and invoices into e-documents, rapidly improving work efficiency. The intelligent power transmission line inspection solution based on MindSpore conducts front-end monitoring of power transmission line equipment and perimeter conditions, analyzes abnormalities and problems, and promptly triggers alerts. China Southern Power Grid and Shenzhen Power Supply Bureau (深圳供电局) have even inaugurated a brand-new model that is "based on intelligent analysis by the system and supplemented by manual judgment." This enables the power transmission monitoring and command center to complete on-site inspection work that originally took 20 days in only 2 hours, increasing the inspection efficiency by nearly 80x. In addition, Zidong Taichu and Wuhan.LuoJiaNET, incubated based on MindSpore, have transitioned from academic research to industrial applications, assisting CCTV, iQIYI, Xinhua News Agency Technology Bureau, PIESat (航天宏图), and other enterprises in carrying out innovative applications.

PaddlePaddle serves enterprises in energy, finance, industry, healthcare, agriculture, and other fields, helping thousands of industries achieve intelligentized (智能化) upgrades. PaddlePaddle enables the "creative brain" of *People's Daily*, covering all stages and business scenarios such as all-media planning, collection, editing, and dissemination performance analysis. It can greatly improve the efficiency of news product production and can carry out various intelligentized production processes such as live video key figures (视频直播关键人物), sentence recognition, network-wide custom monitoring and early warning for critical data, and batch generation of visualizable big data reports. Based on the PaddlePaddle platform, Linking Med (连心医疗) developed and launched a "computerized tomography (CT) image-based pneumonia screening and disease pre-evaluation AI system," which was first put into operation at the Affiliated Hospital of Xiangnan University in Chenzhou, Hunan Province. It can quickly detect and identify pneumonia lesions and provide quantitative evaluation information for disease diagnosis, such as the number and volume of lesions and the proportion of lung area they occupy. At the same time, it is supplemented with visualization methods such as lung density distribution histograms and superimposed displays of lesion outlines, which provide qualitative and quantitative bases for clinicians to screen and pre-diagnose the pneumonia conditions of patients. This improves the efficiency of diagnosis and evaluation by physicians.

Megvii MegEngine fully leverages its advantages in the visual field to realize industry empowerment. Megvii's intelligent quality inspection solution that it provided for a certain camera module company has achieved online real-time inspection of products. Based on the privatized (私有化) MegOne deployment version of the Brain++ platform, it can detect product defects such as scratches, creases, oil stains, and damage in real time. The defect detection rate increased by 90% year-on-year, labor costs were reduced by more than 85%, and overall maintenance costs were reduced by 10%. Megvii launched a supply chain operating system, Hetu (河图), which collaborates with 500 robots concurrently in an e-commerce warehouse, increasing warehouse efficiency by 40%. Megvii deployed a park safety management system for China Resources Power (CR Power; 华润电力), using computer vision algorithms such as facial recognition and object detection to realize 24/7 vigilance for dangerous areas such as around substation (变电) equipment, significantly improving the level of safety management.

OneFlow (一流科技) gives full play to the advantages of distributed and scalable performance. It has served customers in many industries such as scientific research, government affairs, the military industry, and finance. Based on the OneFlow framework, OneFlow integrates big data, cloud computing, and other components to provide the commercial product OF Smart Cloud (OF 智能云), which includes the AI development platform OneBrain, the reinforcement learning solution OneAgent, and the AI training and programming platform OneLab. OneBrain assisted the Zhongguancun Intelligence Application AI Research Institute (中关村智用研究院) to create a one-stop AI development platform that provides a variety of hybrid computing solutions and supports on-demand resource expansion. Once this project is delivered and put into operation, it is calculated that the computing power rate of the system can be increased by 30% and the model training time can be reduced by 80% compared with the traditional method, providing an overall solution to complex business scenarios, high compute requirements, and flexible boundary extension (边界灵活延展) requirements.

(5) In terms of promotion methods, the three roads go in the same direction

Strive for the growth and optimization of the community ecosystem and attract more developers from academic and industrial circles. Mainstream AI frameworks use the flourishing open-source community ecosystem to build loyal contributor teams, thereby attracting more developers to participate in building up the ecosystem. The Google TensorFlow team made their code open source on GitHub and gradually attracts developers who are just starting out and turns them into contributors. Focusing on the TensorFlow open-source community, in addition to contributing TensorFlow higher-order API codes, contributors also actively participate in the management of the TensorFlow community, contribute to open-source projects extended from TensorFlow, and spread their knowledge and share their experience. Huawei launched the MindSpore Developer Support Program to provide developers with discounted cloud service resources and relevant knowledge empowerment training resources to help individual developers learn about and build MindSpore-based technical capabilities and achieve continuous career development. Baidu partnered with community developers to build the ecosystem and has established 150 PaddlePaddle City/University Pilot Groups and 12 PaddlePaddle Special Interest Groups. At present, 132 self-organized city and university communities across the country actively and spontaneously hold PaddlePaddle community activities.

Collaborate with university research institutes to expand the academic research developer scale and academic research applications at universities. The cultivation of talents and the development of developers at universities has become an important part of the overall AI framework ecosystem. Currently, mainstream AI frameworks in China are actively integrated into the teaching systems of universities. Huawei and the Ministry of Education jointly launched the construction of the "Intelligent Center” (“智能基座”) industry-education integration collaborative education base. At present, MindSpore courses are already offered at more than 100 universities, and they are actively holding advanced seminars on computer system capacity building to train teachers in the vanguard of AI. Baidu supports the Ministry of Education's industry-university cooperation collaborative education projects. To date, PaddlePaddle has trained more than 3,000 university professors and participated in the compilation of a series of AI teaching materials. In addition, mainstream AI frameworks also choose to set up innovation funds to stimulate innovative framework applications. In 2020, Huawei and the Chinese Association for Artificial Intelligence jointly launched the *Chinese Association for Artificial Intelligence-Huawei MindSpore Academic Award Fund*, which aims to encourage the development of original scientific research and build the global influence of Chinese AI scientific research. A total of Chinese yuan Renminbi (RMB) 16 million in funding has been invested, supporting more than 120 projects. According to Papers With Code data, in October 2021, the number of papers based on MindSpore accounted for 10% of all papers based on AI frameworks (putting it in 2nd place for the month), so the results have been remarkable. In 2020, Baidu and the China Computer Federation (CCF) jointly established the "CCF-Baidu Songuo Fund" (“CCF-百度松果基金”) which aims to provide young scholars with funding, platforms, data, technical support, and other services and promote the application of AI frameworks in the field of scientific research.

Provide infrastructure and solution services for industrial applications and constantly attract downstream partners. Focusing on industrial applications, AI frameworks have empowerment paths at three levels. First, the AI framework must be integrated into the computing power infrastructure to provide AI capability services. For example, the AI computing centers under construction and already in operation by local governments focus on building underlying AI development capabilities based on Chinese AI frameworks, with MindSpore becoming the main choice. Second, an integrated software and hardware solution must be created, using AI frameworks as a channel to connect the underlying computing hardware and upper-level applications. For example, PaddlePaddle actively cooperates with hardware manufacturers and has or is in the process of adapting 31 types of chips and intellectual property (IP) models, further promoting the joint optimization and collaborative development of software and hardware. Zhejiang Lab's (之江实验室) Tianshu Artificial Intelligence Open-Source Platform (天枢人工智能开源平台), with the OneFlow framework at its core, inherits algorithm applications from upper layers and connects to the underlying hardware. In addition, it can leverage AI frameworks to create application platforms for specific industries. For example, based on MindSpore, Huawei and its partners have launched four new platforms: "Ascend (昇腾) Intelligent Manufacturing," "Ascend Smart City," "Ascend Smart Travel," and "Ascend Smart Inspection."

III. Facing the diverse challenges of the future, AI frameworks show six major technical trends

(1) Full-coverage development: AI frameworks will focus on unifying front-end convenience and back-end efficiency

AI frameworks need to provide a more comprehensive API system and front-end language support conversion capabilities in order to make front-end development more convenient. AI frameworks need to provide developers with a highly complete and high-performance API system that is easy to understand and use. The Consortium for Python Data API Standards organized by members of related open-source projects such as TensorFlow and JAX has already started to build relevant standards. At present, PaddlePaddle has made a start by forming a relatively complete API system. At the same time, when AI frameworks are applied in industry, they need to be able to seamlessly connect with industrial-level development languages (C++, C#, Java, Go, etc.) and must also provide axillary programming APIs and functional support. In terms of development language, many existing development frameworks mainly concentrate on Python support. Julia, Swift for TensorFlow,

Cangjie, and other new programming languages are trying to build an ecosystem for languages other than Python in the AI framework field. From the current point of view, although Julia (scientific computing) and Swift (industrial-level development and application) have some distinctive features, they are unlikely to challenge Python's position in the AI framework field in the short term.

AI frameworks need to provide better dynamic-static image conversion capabilities to improve the efficiency of back-end operations. From the perspective of developers who use AI frameworks to implement model training and inference deployment, AI frameworks need to be able to use the dynamic graph programming paradigm to offer a flexible and easy-to-use development experience in the development stage of model training, thereby improving model development efficiency. They must also use static graphs to achieve high-performance operations during model deployment. At the same time, they must be able to convert dynamic graphs to static graphs to realize convenient deployment and performance optimization. At present, international mainstream frameworks have basically realized the programming paradigm of dynamic graph development and static graph deployment and have dynamic-static graph conversion capabilities. However, based on development efficiency considerations, the conversion and unification of dynamic graphs and static graphs require continuous iterative optimization.

(2) Full-scenario: AI frameworks will support device-edge-cloud full-scenario cross-platform device deployment

AI models need to be adapted and deployed to devices in all device, edge, and cloud scenarios, which poses challenges to the AI framework in terms of diversification, complexity, and fragmentation. With the continuous emergence of AI hardware computing devices such as cloud servers, edge devices, and terminal devices, as well as the rapid development of various AI computing libraries, IR tools, and programming frameworks, the AI software and hardware ecosystem shows a trend toward diversified development. However, the models trained by mainstream frameworks are not universal, and cooperation and extension among academic research projects are difficult, resulting in the "fragmentation" of AI frameworks. At present, there are no unified IR layer standards in the industry, which leads to certain differences in the solutions of various hardware manufacturers. This makes the migration of application models difficult and increases the difficulty of application deployment. Therefore, **the standardized intercommunication of models trained based on AI frameworks will be a challenge in the future.**

AI frameworks need to be fully decoupled from the hardware infrastructure platforms and achieve rapid deployment across device platforms through standard

hardware registration APIs. With the increasing complexity of processing tasks and the intensification of data processing, cross-architecture development capabilities will become a normalized requirement. **AI frameworks urgently need to open a set of hardware registration APIs that can be decoupled so that hardware manufacturers can complete adaptation without touching the core code of the framework**. This way, hardware manufacturers will not have to maintain multiple AI frameworks and adaptation codes for different framework versions. The decouplable hardware registration APIs need to include standard hardware running state (运行态) management, abstract operator definition, and performance optimization and adaptation APIs so that AI framework and hardware platform developers use the same APIs to define key information such as device drivers, runtimes, operators, and computational graphs. In addition to the standardization of the above-mentioned APIs, the IRs and operators of models should also be standardized. This way, hardware manufacturers only need to complete the adaptation of different AI frameworks based on the same model format and the same set of operators to meet the business needs of the synchronous adaptation of different terminal-edge-cloud business scenarios.

(3) Hyperscale: AI frameworks will focus on strengthening support for hyperscale (超大规模) AI

Hyperscale AI has become the new deep learning paradigm. In May 2020, OpenAI released its GPT-3 model, which contains 175 billion parameters and has a data set (before processing) of 45TB. In many NLP tasks, it exceeds human capabilities. This large AI model paradigm with hyperscale model parameters and data sets has achieved a new breakthrough in deep learning. After seeing the potential of this new paradigm, industry and academic circles have entered the field one after another. After OpenAI, Huawei released the large PanGu (盘古) model based on the MindSpore framework, Beijing Academy of Artificial Intelligence (BAAI; 智源) released the Wu Dao (悟道) model, Alibaba released the M6 model, and Baidu released the ERNIE (Wenxin; 文心) model. Hyperscale AI is becoming the breakthrough point for the next generation of AI, and it is also the strong AI technology with the most potential.

Hyperscale AI requires the support of large models, big data, and large compute. It also poses new challenges to AI frameworks, which can be summarized as the "Five Walls." The first is the Memory Wall. In the process of large model training, parameters, activations, gradients, and optimizer states need to be stored. The training of the PengCheng·PanGu model required nearly 4TB of memory. **The second is the Compute wall**. Taking the PengCheng·PanGu large model with 200 billion parameters as an example, it needs the support of 3.6 EFLOPS of computing power, which requires the construction of a large-scale heterogeneous AI computing cluster to meet these compute requirements. At the same time, the compute platform must satisfy intelligent scheduling requirements to improve the utilization of compute resources. **The third is the Communication Wall**. After the large model is sliced into clusters in parallel, there will be a lot of communication between the model slices, so communication becomes the main bottleneck. **The fourth is the Tuning and Optimization (调优) Wall**. When training a model with over 100 billion parameters using a cluster with compute measured in EFLOPS, the communication relationship between nodes is very complicated and it is difficult to consider everything necessary to ensure the correctness, performance, and availability of calculations when performing manual debugging. **The fifth is the Deployment Wall**. Hyperscale AI is facing the problem of "big model, small inference" deployment, and the large model needs to be perfectly compressed to meet the deployment requirements of the inference side.

AI frameworks will support the development of hyperscale AI through core technologies such as automatic hybrid parallel (混合并行), global memory management, visualized tuning, and distributed inference. AI frameworks use **multi-dimensional automatic hybrid parallel** to support data parallelism, model parallelism, pipeline parallelism, optimizer parallelism, subgraph parallelism, and other AI parallel computing technologies in multiple dimensions. This solves the problem of horizontal expansion of models and clusters, supports hyperscale model segmentation to large clusters for efficient training, and achieves the optimal computation-to-communication ratio, thereby improving the utilization rate of compute. AI frameworks can use **global memory management** and computing scheduling to realize the unified management of CPU memory, NPU memory, and non-volatile memory express (NVMe) three-tier storage, thereby improving the vertical expansion capability of individual cards (单卡). Hyperscale AI models have extremely large data sets and network depth and width are very large. AI frameworks need to quickly locate network structures or operators with abnormal accuracy through tensor analysis, graph-code combination (图码结合), and other methods and provide convenient and fast **accuracy problem location capabilities**. They need to record and analyze developer tuning paths and AI model accuracy convergence trends in a visual way and recommend **tuning strategies** to the developers to speed up the tuning process. In addition, for the inference services of large models, AI frameworks need to automatically convert from distributed training mode to **distributed inference** mode and realize service-oriented encapsulation (服务化封装) to support the rapid launch of large model services.

(4) Scientific computing: AI frameworks will further integrate and interact with scientific computing

The traditional scientific computing field urgently needs AI technology empowerment and integration. Scientific computing is generally based on accurate mathematical models and uses rigorous calculation methods as a means to simulate problems in application fields such as climate and meteorology, energy and materials, aerospace, and biomedicine. Traditional scientific computing methods solve problems through numerical iteration. They face the problem of an exponential increase in the amount of calculations caused by the curse of dimensionality, resulting in "computation failures" (“算不起”) or even "computation freezes" (“算不动”) in complex problems or scenarios. There are still a large number of problems to be solved in many fields of scientific computing because the mechanisms are not clear or the calculations are too complicated to be solved by traditional algorithms. Artificial intelligence, on the other hand, often relies on mathematical tools featuring "universal approximation" as represented by neural networks. These tools mine data to find laws so as to achieve breakthroughs that humans cannot in image processing and other types of tasks.

AI frameworks provide a new paradigm for solving scientific computing problems and promote the joint development of scientific computing and AI. AI frameworks need to build unified acceleration engines (加速引擎) for AI and scientific computing, support traditional numerical computing methods, and optimize the hybrid computing of traditional numerical methods and AI methods through computational graphs in order to achieve end-to-end AI+scientific computing acceleration. **AI frameworks need to strengthen their automatic differentiation functions**. Improving the automatic differentiation mechanism of the framework and the implementation of the underlying operators and providing support for higher-order differentiation allows AI frameworks to express complex scientific computing formulas. **AI frameworks need rich programming APIs**. Adding Jacobian, Hessian, JVP, VJP, and other APIs provides integrated expressions for AI+scientific computing, enabling developers to program in a way more similar to mathematical computing. **AI frameworks need to have built-in scientific computing suites for specialized fields**. They must provide easy-to-use scientific computing suites for different scientific computing fields, including high-quality domain datasets, high-precision basic AI models, and toolsets for pre- and post-processing. MindSpore has built-in MindSpore Science functional components and launched the MindSpore Elec suite for the electronic information industry and the MindSpore SPONGE suite for the life sciences industry. PaddlePaddle extended its underlying framework and developed the PaddleScience scientific computing development kit, giving it the ability to solve scientific computing problems.

(5) Secure and trusted: AI frameworks will help improve the explainability (可解释) and robustness of AI models

The increased demand for explainability poses advanced requirements for AI frameworks. Presenting the decision-making results of models in a way humans can understand can help people understand important issues such as the working mechanism inside the complex model and how the model makes decisions. Secure and trusted AI frameworks need to support model explainability, transforming black-box AI decisions into decision-making judgments that can be explained. This not only can increase developers' understanding and trust in AI model decision-making, but also can help diagnose factors that affect model performance and make improvements to further improve model performance. At present, some frameworks have begun to support explainability requirements. For example, models based on the PyTorch framework support interpretable libraries such as Captum, TensorFlow models support libraries such as TF-explain, and explainable libraries such as AIX360 and Alibi are supported by both PyTorch and TensorFlow models. In China, MindSpore has MindSpore XAI and PaddlePaddle has InterpretDL. In addition, there are already some platforms that evaluate models in terms of explainability, such as the Chongming (重明) platform on OpenI (启智社区) and the RealAI (瑞莱智慧) Platform.

AI frameworks need to provide a wide range of AI robustness detection tools to improve the robustness of AI models. Insufficient sample training during model training will lead to a model with insufficient generalization capabilities and models will not give correct judgment results when faced with malicious samples. AI frameworks can help developers improve the robustness of models by supporting methods such as network distillation and adversarial training as well as adversarial detection techniques such as black-box, white-box, and gray-box testing. MindSpore launched the robustness testing tool MindSpore Armor, which provides an efficient robustness evaluation solution based on technologies such as black- and white-box adversarial examples and natural perturbation (自然扰动). This helps customers evaluate the robustness of models and identify model vulnerabilities. PaddlePaddle launched the PaddleSleeve model security tool, which provides a complete set of capabilities from AI model robustness evaluation and testing, to model attack defense, to model robustness improvement.

(6) Engineering: AI frameworks will accelerate the implementation of industrial-scale engineering of AI applications

AI engineering (AI 工程化) is the only way for AI to profoundly empower the real economy. Engineering is the basic path by which AI technology moves from theoretical algorithms to practice. It is the process by which, based on a relatively mature algorithm and incorporating industrial needs, an engineering solution that can be implemented and is suitable for large-scale deployment is formed. In recent years, intelligentized applications have emerged in more and more industries and fields, but their engineering implementation is not ideal. At present, only half of projects can transition from AI prototypes to production[7]. In October 2021, Gartner released the Top Strategic Technology Trends for 2022 and once again identified AI engineering as one of the important strategic technology trends. They predicted that by 2025, the 10% of organizations that have established best practices in AI engineering will generate at least three times the value from their AI efforts compared to the 90% of organizations that have not.

AI frameworks need to support the rapid migration of AI models across platforms and minimize the development and debugging costs of developers through technologies such as model adaptation. In different application scenarios and different tasks, the device resource constraints are different, and requirements for the precision productization of AI models are also different. AI frameworks need to balance device resource constraints and accuracy requirements for different scenarios or tasks and optimize the AI model through automated [machine] learning (AutoML), model lightweighting (quantization, pruning, etc.), transfer learning, and other model adaptation technologies. Migration and deployment can target different tasks in the same application scenario or the same task in different application scenarios, avoiding redevelopment from scratch, making full use of the existing technology foundations, achieving rapid deployment, and reducing developers' development time and manpower costs. This will also facilitate the rapid promotion and reuse of AI products.

AI frameworks will rely on incremental learning to more flexibly meet dynamic data training needs and achieve faster AI application development at lower cost. When confronted with new sample data or new tasks, traditional one-time data learning requires a lot of computing resources and time for re-learning, and when training on new tasks, the representation capability for old tasks usually decreases significantly, resulting in "catastrophic forgetting." Incremental learning capabilities can solve the above problems very well by making full use of historical training results to achieve knowledge accumulation, significantly reducing subsequent training time while mitigating forgetting. This is suitable for huge database or data stream application scenarios. In addition, the AI framework supports terminal-side and edge-side incremental learning, which can also optimize the efficiency of lightweight deployment, reduce interaction with cloud-side data, and further improve training performance.

Application engineering (应用工程化) will promote the development of AI frameworks in the direction of refinement and diversification. The industrial-scale engineering deployment of AI applications often involves hardware devices in different cloud-edge-terminal scenarios, including cloud servers, mobile terminals, and Internet of Things (IoT) devices. For mobile terminals and IoT devices, due to hardware resource

[7] 2021 Important Strategic Technology Trends Research Report, released by Gartner.

constraints, the cloud-side model and inference runtime framework are too large to be deployed directly. Therefore, the compression of AI models and the lightweighting of terminal-side inference frameworks become crucial for deployment on mobile terminals and IoT devices. Some mainstream AI frameworks adhere to an integrated training and inference layout and launch inference engine components for mobile terminals and IoT devices to accelerate AI engineering, such as TensorFlow Lite, PyTorch Mobile, MindSpore Lite, and Paddle Lite. In addition, there are AI inference frameworks specifically designed for inference, such as NVIDIA TensorRT, Intel OpenVINO, Tencent Youtu TNN (腾讯优图 TNN), and Alibaba MNN (阿里 MNN). AI applications in all industries have a wide range of AI inference requirements, including precision requirements, ease-of-use requirements, and performance requirements. With the continuous development of AI engineering, the AI inference framework ecosystem will flourish even more.

IV. The AI framework ecosystem is far from mature, so it has considerable room for future development

It has only been five or six years since AI frameworks entered the mainstream field of view. From technological evolution to the open-source ecosystem, market layout, and on to application empowerment and increasing influence, the overall AI framework ecosystem is far from mature. Software and hardware collaboration, open-source creation, developer promotion, and enablement in key fields will provide important support for the maturation and upgrading of the AI framework ecosystem.

(1) Evolve from hardware adaptation to operator API standardization

In order to cope with the challenges of the diversity, complexity, and fragmentation of the AI software and hardware ecosystem, we must urgently promote AI framework hardware adaptation and operator API standardization. We must encourage leading AI enterprises to gradually build standardized hardware APIs through the adaptation of AI frameworks and underlying AI chips, drive hardware manufacturers to actively adapt to AI frameworks, and shift from AI chip-led adaptation to unified hardware API-led adaptation. We must support the development of unified AI operator API standards. By shielding the details of different underlying hardware architectures, we must formulate standardized development APIs to provide unified specifications for AI technology research, software and hardware development, and application development. **Starting from standards work, we must promote the standardization of the unified IR of AI frameworks, accelerate the formation of AI framework support for cross-platform rapid migration and deployment, and build a collaborative ecosystem for AI frameworks.**

(2) Strengthen the creation of open-source communities and an open-source atmosphere

We must focus on open sourcing and opening up, take multiple measures to build an open-source ecosystem for AI frameworks, and create an innovative and favorable development environment for AI algorithm frameworks. We recommend following the principles of open source and openness, jointly building an open-source community, and leading all parties to actively participate in and contribute to it. We must encourage enterprises with technical strength to build the open-source ecosystem and focus on innovation and breakthroughs in key basic fields such as open-source algorithm frameworks, databases, and operating systems; encourage Chinese universities, enterprises, industry organizations, and other industry parties to integrate into the international open-source community ecosystem and increase their participation and influence; and support the construction of platforms for open-source risk monitoring and open-source ecosystem monitoring and strengthen the awareness of open-source ecosystem governance. **We must build an open-source ecosystem for AI frameworks in order to provide continuous support for technological innovation, product optimization, application expansion, and talent recruitment for AI enterprises themselves.**

(3) Focus on wide-ranging and open cooperation with universities and scientific research institutes

We must guide academic circles, universities, and scientific research institutions to work with industrial enterprises to build their AI application systems based on mainstream AI frameworks and provide preferential policies in terms of project approval and application for ST innovation funds. By supporting and encouraging universities, scientific research institutions, and partners to participate in the crowdfunded development of mainstream AI frameworks and by co-constructing joint laboratories and innovation centers, we must develop and tune network models under multiple mainstream AI frameworks, continuously add additional operators and models, continuously optimize the precision and performance of operators and models, and cultivate a large batch of outstanding developers. We must encourage leading AI enterprises to carry out wide-ranging cooperation with universities, including in "talent training, teaching materials/teaching supplement books, teaching courses, technical cooperation, scientific research, competitions, and project incubation," support universities in their efforts to build core courses and digital teaching resources that incorporate AI frameworks, and carry out theoretical teaching, experimental and practical training, and AI technology cooperation projects based on AI frameworks.

(4) Promote the implementation of integration into the AI infrastructure

layout

AI infrastructure has "data resources, algorithm frameworks, and compute resources" as its core capacity elements and "open platforms" as its main enabling vehicles. These can provide public and inclusive intelligentized services long-term. **We must encourage AI framework entities to integrate with AI computing centers, publicly owned[8] AI clouds, and AI application open platforms** to provide AI capability services to external users. For example, the AI computing centers led by large and medium-sized cities focus on leveraging high-quality AI frameworks such as MindSpore to consolidate their underlying AI development capabilities. We must support the government, enterprises, and public institutions[9] as they actively purchase AI infrastructure services and gradually expand the scope of influence of AI frameworks.

(5) Support the in-depth enablement of large models and scientific computing

We must support AI framework entities as they delve into the scientific computing basic research field and realize the rapid development of AI frameworks by deeply enabling hyperscale AI and integrating it into the field of scientific computing. Hyperscale AI is the core driving force for the ongoing revolution in AI over the past two years, and AI-integrated computing (AI 融合计算) is the key support that has allowed AI to enter various academic disciplines. Both technologies are new high grounds for the basic scientific research of various countries and new scientific implements. Hyperscale AI and AI-integrated computing place higher requirements on AI frameworks and promote the continuous improvement and upgrading of AI frameworks in terms of performance, accuracy, time efficiency, energy consumption, and other dimensions. At the same time, AI frameworks need to actively integrate into the field of hyperscale AI and AI-integrated computing to conduct innovative AI applications in order to provide a wealth of resources for ecosystem layer suites/model libraries.

[8] Translator’s note: The term “publicly owned” (公有) refers to entities that are owned by the Chinese government or the Chinese Communist Party (CCP).

[9] Translator's note: "Public institutions" (事业单位) are organizations created and led by Chinese government departments that provide social services. Unlike state-owned enterprises (SOEs), public institutions do not create material products and do not generate income. Public institutions are not considered government agencies, and their employees are not civil servants. Most public institutions are fully or partially government-funded, but some fully privately funded (but still government-led) public institutions exist. Public institutions typically provide services in areas such as education, science and technology, culture, health, and sanitation.