DAMO
ALIBABA DAMO ACADEMY

2022

Top Ten Technology Trends

of DAMO Academy

Preface

Over the past century, the evolution of digital technologies has accelerated technological progress and industrial development. Digital technologies are growing faster than ever. The advancements in digitalization, Internetization, and intelligence are redefining a digital world that is characterized by the prevalence of mixed reality.

The boundary of technologies is extended from the physical world to mixed reality. Cloud-network-device convergence drives the emergence of new types of devices on the cloud. The productization of AR and VR has paved the way for Extended Reality (XR) to transform the way people interact with technology. XR is an exciting new technology that can simulate real-world scenes and effectively shorten the physical distance between people. Applications such as remote education, healthcare, and working can be implemented in new ways.

More and more cutting-edge technologies find their way to industrial applications. A digital operating system built around the cloud will be able to leverage a variety of technologies and resources, such as AI, big data, and large-scale computing power. Individuals, enterprises, and research institutes can easily obtain desired resources. This lowers the threshold for people and industries to access the latest technologies and further drives innovation. Currently, we have already completed proof of concepts for AI engineering, AI for science, and pre-trained models in various sectors, such as biomedicine, astronomy, meteorology, and industrial manufacturing.

Digital technologies power a green and sustainable future. All across the world, the clock on environmental issues is ticking faster than ever. Everyone has a social responsibility to play a part in adopting green and low carbon practices in our lives and activities. In this aspect, digital technology plays an important role, whether it is applied in industries (green data centers and energy-efficient manufacturing) and day-to-day activities (paperless office).

With technology, we will create a better future.

Jeff Zhang

President of Alibaba Cloud Intelligence
Head of Alibaba DAMO Academy

Contents

Paradigm Redefinition

Scenario Revolution

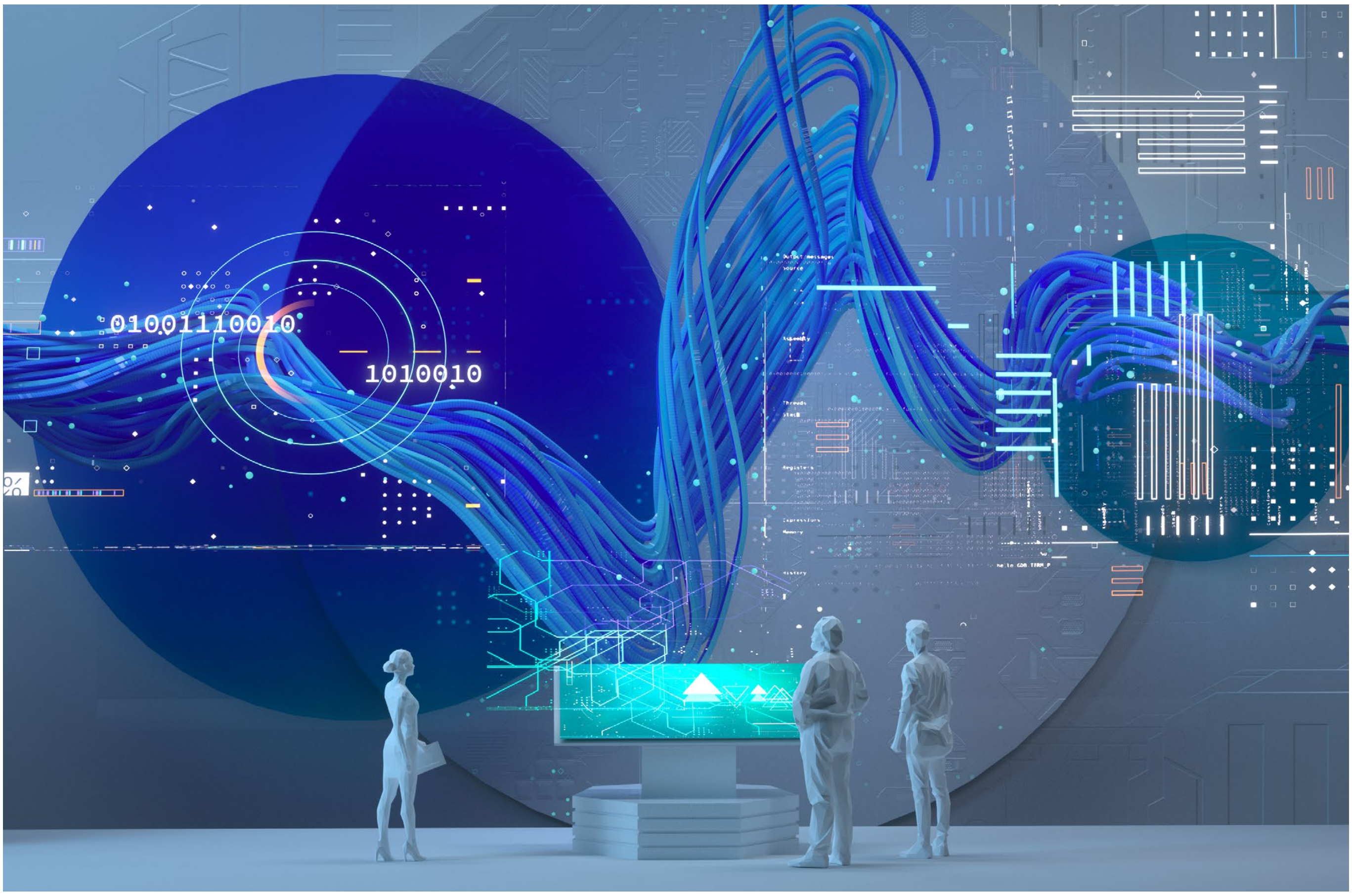
Future Interconnect

AI for Science

AI is a production tool that is spawning a new paradigm in scientific research

Introduction

In the past hundreds of years, the scientific community had two basic paradigms: experimental science and theoretical science. Today, the advancement of AI is making new scientific paradigms possible. Machine learning can process massive amounts of multidimensional and multimodal data and solve complex scientific problems, allowing scientific exploration to flourish in areas previously thought impossible. AI will not only accelerate the speed of scientific research, but also help discover new scientific laws. In the next 3 years, we expect that AI will be broadly applied in the research process of applied science and be used as a production tool in some basic sciences.



Trend Analysis

Scientific exploration is often a shot in the dark – many discoveries are the result of time-consuming, trial-and-error, and accidental processes. Many breakthroughs are built upon the results of previous scientific discoveries by great minds like Isaac Newton, Albert Einstein, and Yang Chen-Ning. Despite the unrelenting efforts of today's scientists, the speed of scientific development is still limited by a lack of direction and means.

Computer science has transformed the way that scientific research is carried out. In the early years, computers were mainly used to analyze and categorize experimental data. As the technology progressed, scientific computing gradually changed how experiments are conducted. Researchers combined AI with high-performance computing to simulate costly and complex experiments and verify hypotheses. This helped accelerate the speed of scientific research. For example, digital nuclear reactors for nuclear energy experiments reduce experimental costs, improve safety, and reduce nuclear wastes. In recent years, AI has also been used to discover and define fundamental scientific laws. AI-powered scientific law discovery plays a role not only in the field of applied science, but also in the fields of basic sciences. For example, DeepMind uses AI to prove or propose new mathematical theorems and assist mathematicians in developing intuitions in complex mathematics.

In the future, AI will replace computers as the main production tool in scientific discovery. AI can significantly improve efficiency in each phase of research process, from the formation of the initial hypothesis, to the experimental procedure, and finally to the distillation of experimental findings. AI helps shorten research cycles and improve the productivity of scientists. In addition, AI is able to produce scientific conjectures, which can in turn direct the flow of scientific research. This lowers the bar for meaningful and valuable experiments, and encourages more people to participate in research.

The speed of AI adoption varies across scientific research fields. AI advances at a faster rate in highly digitalized fields where massive amounts of data is accumulated and issues are clearly defined. For example, AlphaFold 2 can predict protein structures by using gene sequences based on the massive amounts of data accumulated in the field of life sciences. This breakthrough has a profound impact on the advancement of life sciences. For another example, machine learning is adept at extracting and categorizing multidimensional data and distilling research findings, especially in highly complex fields that take large numbers of variable factors into account, such as fluid mechanics. In such cases, machine learning can be used to replace manual work.

Deep integration of AI and scientific research faces the following challenges:

First, the interaction between researchers and AI systems. As with most collaborative systems, the research conducted by human experts and discoveries from AI experts as well as their interaction process should be well-defined to form a close working relationship, preventing overlap of workloads that lead to inefficiencies.

Second, interpretability of AI. The results produced by AI should show strong relationships between the findings, which allows researchers to easily interpret and apply the findings. This also fosters trust in the technology.

Third, demand for interdisciplinary talent to enhance the communication between scientists and AI.

In the next 3 years, we expect that AI will be broadly applied in the research process of applied science and be used as a production tool in some basic sciences.

Co-evolution of Large- and Small-scale AI Models

As the race on the scalability of pre-trained models progressively calms down, the co-evolution of large- and small-scale AI models via clouds, edges, and devices has been drawing more attention

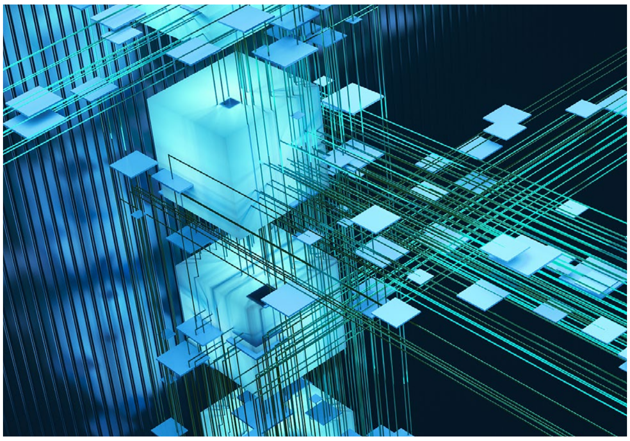
Introduction

The large-scale pre-trained models, also known as the foundation models, are the grounding breakthrough technique from weak AI to general AI, which relatively boosts performance of various applications using conventional deep learning. However, the merit in the higher performance and the drawback in the power consumption are not well balanced, limiting the exploration of large-scale models. The future AI is shifting from the race on the scalability of foundation models to the co-evolution of large- and small-scale models via clouds, edges, and devices, which is more useful in practice.

In the co-evolution paradigm, foundation models deliver the general abilities to small-scale models that play the role of learning, inference, and execution in downstream applications. Besides, small-scale models will also send the feedback of the environment to the foundation models for further co-evolution. This mechanism mutually enhances both large-scale models and small-scale models via positive cycles.

Trend Analysis

Large-scale pre-trained models such as BERT (Google), GPT-3 (OpenAI), WuDao (BAAI), and M6/AliceMind (Alibaba DAMO Academy) have made remarkable achievements. However, training such foundation models requires a considerable amount of computation cost. For example, GPT-3 requires 190,000 kilowatt-hours per training iteration and is equivalent to the amount of resources consumed to make a round trip from the earth to the moon. Models with more parameters lead to better performance but require more energy costs. Thus training efficiency is fundamental to the ubiquitous presence of AI.



As the race on the scalability of foundation models progressively calms down, the co-evolution of large- and small-scale models has been drawing more attention. The knowledge and ability of cognitive inference from foundation models are first delivered to small-scale models. Then, small-scale models are applied to downstream applications, which in turn results in separately evolved branches of the main model. These branches combine the vertical information such as perception, cognition, decision-making, and execution results of their scenarios, and feed the results to the foundation models. In this way, foundation models continually evolve through the feedback and the learning to build an organic intelligent cooperative system. The more participants involved in, the faster the models evolve.

The new intelligent system brings three merits. First, the system makes it easier for small-scale models to learn the general knowledge and the inductive abilities, which are then fine-tuned to their specific application scenarios. Second, the system increases the variety of data for the foundation models. The incremental data captured by small-scale models in real-world scenarios allows the foundation models to continue to evolve over time. Third, when the foundation models are regarded as the fundamental public resources, we only need to iterate on the downstream specific small-scale model, which helps achieve the best combination between the energy efficiency and the training speed.

AI is a key production tool in the era of the digital economy and brings disruptive changes to the industrial or academic production methods. The base AI models greatly simplify the AI-based production methods, allow more flexible on-demand development of incremental learning algorithms in vertical industries, and improve the efficiency of model production.

In addition, complex systems can be more organically integrated with each other. For example, in urban governance scenarios, the cloud serves as the central brain of governance, and cameras and edge devices serve as individual "mini brains". One of the cameras learns from captured data and feeds the results back to the governance center. The governance center then iterates the base model with the new information, and delivers upgraded capabilities to other cameras. This allows the systems to continually evolve.

The new intelligent system needs to overcome three major challenges. First, we need to endow foundation models with intellect and the ability to reuse the rule-based knowledge, which in turn reduces the volume of data required for training. This allows us to push the paradigm from the data-driven training towards the more efficient knowledge-based training. Second, the system needs to address the collaboration between large- and small-scale models, including effectively transferring the knowledge and inductive abilities from foundation models to small-scale models, extending the few-shot learning results from small-scale models to foundation models, and cleansing and governing data from different dimensions. Third, the system needs to address the interpretability and causal inference issues of foundation models. As the reliance of small-scale models on foundation models grows, the trust in foundation models determines the scope in which these models can be applied.

It the next 3 years, we expect to see foundation models in various domains that will be used as the base AI models in multiple fields, and the start of exploration into the co-evolution of intelligent systems. In the next 5 years, these base AI models will be used as a standard for the AI model production, which will bring huge changes to the production process and the skills required for production.

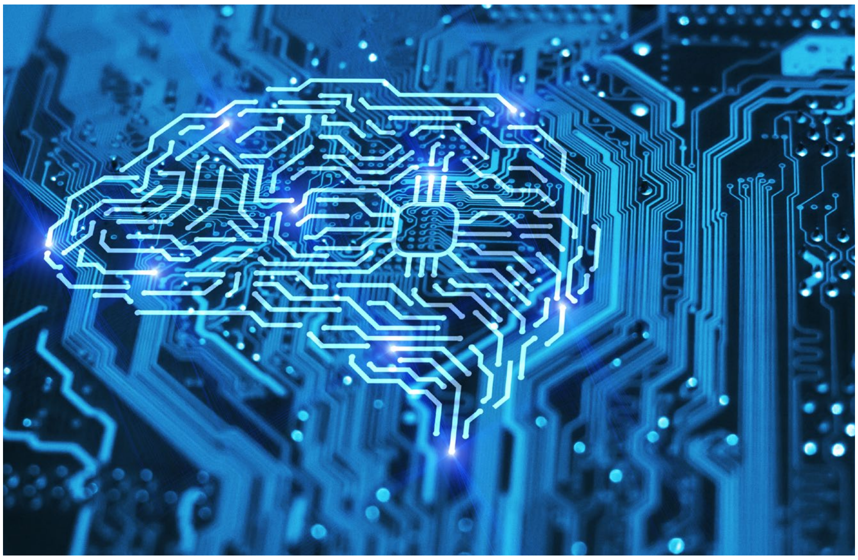
Silicon Photonic Chips

Silicon photonics leverages the advantages of photonics and electronics to overcome the impending limits of Moore's Law

Introduction

As the size of transistors approaches physical limits, the speed of electronic chip development can no longer meet the increasing data throughput demand brought by the rise of high-performance computing. Unlike electronic chips, silicon photonic chips use photons instead of electrons to transmit data. Photons do not directly interact with each other and can travel longer distances, and therefore silicon photonic chips can provide higher computing density and energy efficiency. The rise of cloud computing and AI drives the rapid development of silicon photonics technology. In the next 3 years, we can expect to see the widespread use of silicon photonic chips in high-speed data transmission in large-scale data centers.

Trend Analysis



As Moore's Law approaches its physical limits, there is little room left for advances in electronic chip technologies. However, as the rise of high-performance computing is continually raising the bar for data throughput, the need to make more technological breakthroughs becomes increasingly urgent.

Silicon photonic chips vary from electronic chips in that they use photons instead of electrons to transmit data. Photons do not directly interact with each other and can travel longer distances. Using photons instead of electrons in chips can provide two orders of magnitude higher computing density and energy efficiency. Unlike quantum chips, silicon photonic chips allow us to continue using current computer systems, as silicon photonic chips can be used directly with the current binary code-based software architecture. Silicon photonic chips need to converge with mature electronic chip technologies, and utilize advanced manufacturing processes and modular technologies of the electronic chip industry. The silicon photonics technology that combines the advantages of both photonics and electronics will shape the future of the semiconductor industry.

The concept of "silicon photonic chips" first appeared about 40 years ago. At the beginning of the 21st century, breakthroughs in the core technologies of silicon photonic chips laid the foundation for large-scale commercial use. The rise of cloud computing and AI is the driving factor for technological breakthroughs and the rapid advancement and commercialization of silicon photonic chips. Large-scale distributed computing, big data analytics, and cloud-native applications have also greatly increased the data density during communications in data centers. As a result, data transfer has become a performance bottleneck. Traditional optical technologies are too expensive to be applied on a large scale. Silicon photonic chips provide a relatively cheap and effective method to improve the communication efficiency between clusters, servers, and chips in data centers.

According to an analysis released by OpenAI, "since 2012, the amount of compute used in the largest AI training runs has been increasing exponentially with a 3.4-month doubling time (by comparison, Moore's Law had a 2-year doubling period)". It is evident that the growth of compute requirements has outpaced the current technology. Silicon photonic chips have higher computing density and energy efficiency, and therefore provide an effective solution in scenarios where a large amount of computing power is required.

Silicon photonic chips can be widely used in optical communications within and between data centers and optical computing. The current challenges of silicon photonic chips are mainly in the supply chain and manufacturing processes. The design, mass production, and packaging of silicon photonic chips have not been standardized and scaled, leading to low production capacity, low yield, and high costs. In terms of optical computing, its precision is lower than that of the computing based on current electronic chips, which limits the application scenarios of silicon photonic chips. Furthermore, more research is required to further integrate photonics into integrated circuits to improve computing power.

It is worth noting that optical communications and optical computing complement each other. The photoelectric conversion technology in optical communications can be used in optical computing, while the low-loss, high-density photonic integration required in optical computing can further promote the development of optical communications. In the future, both of data calculation and transmission may be completed with the optical technology.

Photonic-electronic integration is the future of chip development. As silicon photonic chips and silicon electronic chips complement each other, the combined application of both types of chips will provide the necessary momentum to support the growth of computing power into the future. In the next 3 years, we can expect to see the widespread use of silicon photonic chips in high-speed data transmission in large-scale data centers. In the next 5–10 years, we can expect to see silicon photonic chips gradually replacing electronic chips in some computing fields.

AI for Renewable Energy

Advanced AI helps ensure safe and reliable integration
of renewables into the power grid

Introduction

The rapid development of technology in renewable energy such as wind and solar power in recent years has made renewables a tempting energy source to add to the power grid. However, issues such as difficulty in grid integration, low energy utilization rate, and storage of excess energy are major roadblocks along the way. Due to the unpredictable natures of renewable energy power generation, integrating renewable energy sources into the power grid presents challenges that affect the safety and reliability of the grid. The application of AI in the industry is pivotal in improving the efficiency and automation of electric power systems, maximizing resource usage and stability. This will be conducive to achieving carbon neutrality.

In the next 3 years, AI is expected to pave the way for integration of renewable energy sources into the power grid and contribute to the safe, efficient, and reliable operation of the power grid.

Trend Analysis

Energy from renewable sources is largely affected by weather, which is unpredictable and may change rapidly. Therefore, grid integration of renewable energies poses new challenges, especially in maintaining a stable output. Grid integration, energy transmission, energy consumption, and safe operation of grids are key aspects that need to be improved. According to National Energy Administration, by 2030, 40% of electric power consumption in China will be covered by renewables, and the total installed capacity of wind and solar energy resources will reach 1.2 billion KW.

AI will be indispensable in capacity prediction, scheduling optimization, performance evaluation, failure detection, and risk management.

In particular, it achieves breakthroughs in the following aspects:

Accurate capacity prediction: Big data analytics and neural network algorithms will provide more accurate predictions of renewable energy capacity based on weather forecasts. The predictions can be used to dynamically optimize power generation policies and ensure that sufficient power is provided to the nationwide grid.

Intelligent scheduling: Technologies, such as deep learning, big data analytics, and physical simulation, can help the power scheduling system optimize scheduling policies and coordinate energy sources such as wind energy, photovoltaic energy, hydroelectric energy, and stored energy. The grid integration of renewable energies requires better structuring, operational practices, and decision making for hybrid AC/DC transmission grids and source-grid-load-storage integration. In the future, renewable energies will grow to become the mainstay of China's energy consumption, and AI will play an important role in this progress.

Automatic failure responses: Big data and deep learning technologies are used to implement real-time monitoring of grid equipment, allowing people to identify precursors of equipment failures and further enhancing the capability to detect and respond to such issues. As these technologies evolve, they pave the way for high-precision control and near-real-time alerting.

Renewable energies are the future, and renewable energy sources are being continually connected to the power grids. However, the traditional power management and distribution system cannot handle the unpredictability of renewable energies that are dependent on weather conditions, neither can it reliably distribute energy from multiple sources or quickly compensate for complex failures across the grid. Human intervention is still required for parameter verification and failure detection. However, this model is not effective in identifying failure traits and creating models for pre-emptive failure detection. In the face of these looming challenges, AI offers an all-round solution to address the challenges in grid integration of renewable energies and ensure stable and efficient operations of the power system.

AI will transform the power system by facilitating the generation, grid integration, transmission, and safe operation of renewable energies. Based on weather forecasts, AI can accurately predict the capacity of various power sources. It is also expected to provide better monitoring, scheduling, and controlling capabilities for the power system. The challenges that lie in these prospects need to be properly handled.

In the next 3 years, AI will help put substantial renewable energies into use and ensure the reliability of power sources located across the country. The balanced development and flexible coordination of energy sources with the power grid will contribute to a reliable, efficient, and safe power distribution system.

Perceptive Soft Robotics

The design of robots with physically flexible bodies and advanced mechanosensing capabilities opens up more possibilities in how and where robots can be used

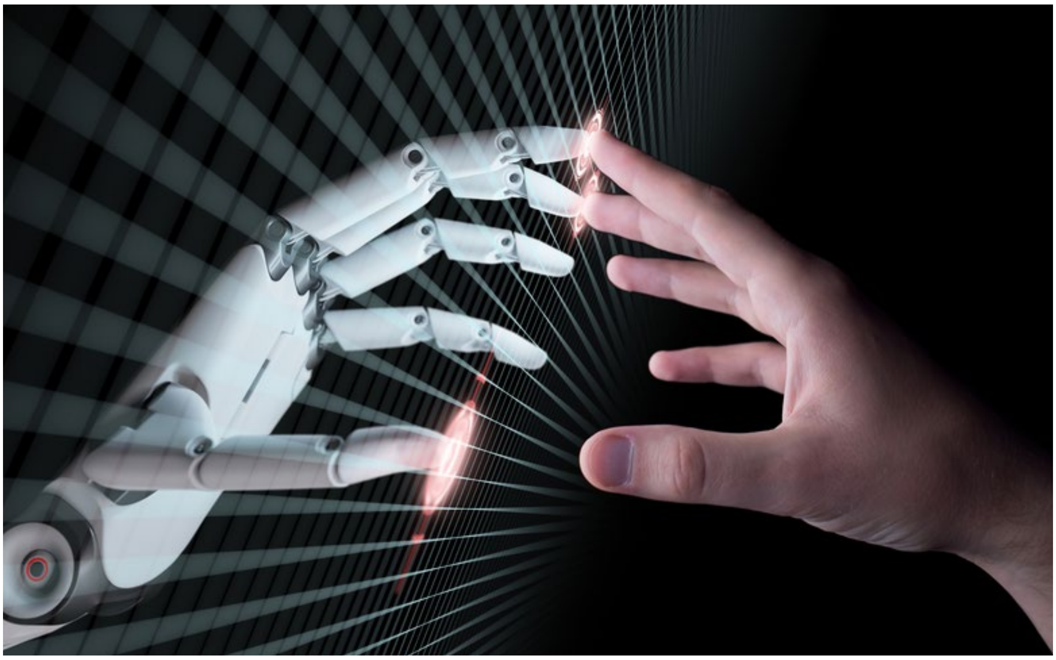
Introduction

Robots, in the conventional sense, have limited capabilities in day-to-day applications. Conventional robots are pre-programmed machines that handle monotonous tasks in a controlled environment, such as the manufacturing industry. Unlike conventional robots, perceptive soft robots are robots with physically flexible bodies and enhanced perceptibility towards pressure, vision, and sound. These robots take advantage of state-of-the-art technologies such as flexible electronics, pressure adaptive materials, and AI, which allow them to perform highly specialized and complex tasks and deform to adapt to different physical environments. The emergence of perceptive soft robotics will change the course of the manufacturing industry, from the mass-production of standardized products towards specialized, small-batch products. In the next 5 years, perceptive soft robotics will replace conventional robots in the manufacturing industry and pave the way for wider adoption of service robots in our daily life.

Trend Analysis

Robots are an integration of advanced technologies. The rapid development in hardware, networks, AI, and cloud computing allows us to design multi-functional, adaptative, and collaborative robots.

Perceptive soft robotics is a game-changing technology. Perceptive soft robots are flexible, programmable, and deformable, and are empowered by advanced technologies such as flexible electronics and pressure adaptive materials. This enables perceptive soft robots to handle complex tasks in various environments. The adoption of AI technology further enhances the perception system of perceptive soft robots, making them smarter and applicable to more industries.



Unlike conventional robots, preceptive soft robots have enhanced perceptibility towards pressure, vision, and sound, which allows them to adapt to and interact safely with their external environment. This allows them to handle complex tasks such as surgeries, and effectively be used in a wider range of applications. In addition, perceptive soft robots are highly adaptable. They can respond appropriately to external stimuli and perform precise procedures, effectively replacing humans in some cases and minimizing human errors.

Perceptive soft robotics will change the course of the manufacturing industry, from the mass-production of standardized products towards specialized, small-batch products. This is mainly driven by two factors: perceptive soft robots are multi-functional and can be easily repurposed, and the integration of AI allows for less human interaction during the manufacturing process. For example, perceptive soft robots are a prime candidate to replace human labor during the COVID-19 pandemic, when human labor is scarce.

Perceptive soft robotics greatly improves the safety of interaction between humans and robots. By sensing the intentions of humans and responding naturally, perceptive soft robots can work safely and closely with humans.

Another trend in preceptive soft robotics is mobility. Preceptive soft robotics allows us to endow automatic guided vehicles (AGVs) with more autonomy and adaptability to perform multiple functions. These technologies allow robots and humans to work together in more scenarios.

Perceptive soft robotics must overcome three challenges. First, the intelligence capacity of robots is limited by the on-device computing power and the validity of few-shot learning. Cloud-edge collaboration is the key to eliminating this bottleneck. Second, the precision of perceptive soft robots varies depending on the rigidity of the material. Third, perceptive soft robots are costly. More work must be done to optimize the manufacturing process and minimize cost.

In the next 5 years, we predict that perceptive soft robots will be endowed with intelligent perception systems powered by deep learning, and subsequently replace conventional robots in various industries. The commercial use of perceptive soft robots will make them more competitive than other service robots and become increasingly popular in our daily life.

High-precision Medicine

The convergence of AI and precision medicine allows more precise targeting of diseases and improved efficacy of treatments

Introduction

Medicine is a field that is highly dependent on individual expertise, often involves a lot of trial and error, and may ultimately have varying efficacies from patient to patient. The convergence of AI and precision medicine is expected to boost the integration of expertise and new auxiliary diagnostic technologies and serve as a high-precision compass for clinical medicine. With this compass, doctors can diagnose diseases and make medical decisions as quickly and accurately as possible. These advances will allow us to quantify, compute, predict, and prevent severe diseases. In the next 3 years, we expect to see people-centric precision medicine become a major trend that will span multiple fields of healthcare, including disease prevention, diagnosis, and treatment. AI will become synonymous with a highly precise compass that allows us to pinpoint diseases and their treatments.

Trend Analysis

AI technologies play an important role in overcoming the limitations of conventional medical practices in the screening, diagnosis, prognosis, and treatment of diseases. The enhanced accuracy, speed, and precision afforded by AI technologies that combine medical expertise and novel auxiliary diagnosis technologies can ultimately be leveraged to improve the efficacy of treatments. In the future, AI algorithms must not only meet expected clinical goals, but also evolve continuously to provide safe and effective human-machine interaction (HMI) capabilities, which in turn helps build trust between medical professionals and AI. This is a crucial step in the development of AI to becoming an indispensable component of the medical field.

As emerging technologies and clinical care continuously integrate with each other, disease prevention and treatment will gradually become inclusive. Primary medical care capabilities will expand to include personalized disease prevention, diagnosis, and treatment. The next step is to help doctors quickly figure out effective solutions for preventive and early diagnosis and treatment. So far, it has been proven that AI technologies can be applied to multiple new medical research areas such as genetic testing, targeted therapy, and immunotherapy, changing the traditional diagnostic mode that relies solely on the personal experience of doctors. In the next 3 years, we expect to see new technologies built on top of AI become commonplace in the treatment and care of some serious diseases such as tumors and chronic diseases.

1. AI-based clinical care of tumors

AI technologies will be gradually applied to the early screening and diagnosis of all cancer types. The enhanced precision and accuracy afforded by AI can provide enhanced images that can help medical professionals effectively identify cancerous tumors. Automated, large-scale screening can also be achieved once a sufficiently large sample of data on tumor markers is aggregated and analyzed. Multimodal radiopathological diagnosis reports and comprehensive evaluation reports can be automatically generated to help doctors make medical decisions, increase the rate and accuracy of early diagnosis and improve the efficacy of treatments, and ultimately lower the fatality rate of malignant tumors. According to statistics from the UK and USA, using AI technologies in the early screening of breast cancers can reduce the false negative rate (FNR) by 5.7% in the USA and 1.2% in the UK.

In the treatment stage, AI technologies will transform the way tumors are treated. Currently, medical professionals rely solely on their eyes and experience to determine whether surgery is required. In the future, advanced AI technologies will help medical professionals identify reoccurrence or metastasis with higher accuracy, making the treatment process easier and more transparent. AI technologies will also play a critical role in differentiating treatment for individuals and providing tailored plans for radiotherapy, chemotherapy, and targeted therapy based on the analysis on clinical data. Moreover, AI technologies will improve the precision of antigen prediction in tumor-specific immunotherapy. Tumor-specific immunotherapy is the most promising tumor treatment method, which uses antigens to trigger immune response in patients. AI plays a huge role in predicating antigens through identifying tumor-specific characteristics, which reduces the amount of time and work required as compared to the current method of screening spatial structures out of a large number of abnormal antigen peptides and immune cell receptors.

In the prognosis stage, AI technologies will change traditional prognostic methods that rely solely on doctors' expertise and will complete precise computations based on clinical data to guide prognosis at lower risks.

2. AI-based prediction and diagnosis of chronic diseases

Disease prediction based on AI-powered image recognition and big data technologies can effectively annotate historical data and perform pathological analysis to improve the accuracy of prognosis and diagnosis.

Ultra-large-scale pre-trained models will be used to create knowledge graphs, build causal relationships between biology and medicine, and integrate multimodal data, improving the interpretability of medicine and laying the foundation for artificial general intelligence-powered medical care. For example, in AI-powered imaging diagnosis of fatty liver, ultrasonic equipment is used to scan different parts of the liver. The data obtained from the scan is used to build a model, which is then subject to computational analysis and scoring. A highly accurate result can be obtained in a short time, serving as a basis on which medical professionals can use to determine treatment. At current, this technology outperforms FibroScan, the gold standard for liver elastography. AI technologies can help categorize the degrees of fatty liver into four types, namely, None, Minor, Moderate and Severe, and prescribe medicine and dosage in a quantitative manner. In contrast, human doctors can only determine whether fatty liver exists or not, which cannot meet the quantitative accuracy requirements for chronic disease intervention, not to mention that different doctors have different diagnosis results. Furthermore, AI images will be used to predict the probability of osteoporosis in a patient through the use of general-purpose X-ray machines. These will replace the use of rare, expensive dual-energy X-ray absorptiometry (DXA) to scan the chest, lumbar vertebra, and pelvic bones. This innovation can expand the coverage of screening to cover up to 80% of the at-risk population.

In future, AI technologies will reshape the services in the healthcare industry. Medical professionals will leverage the advantages of precision medicine to gradually shift their focus to disease prevention instead of disease treatment. This creates a new healthcare model that combines treatment and general healthcare, allowing the general population to become more acquainted with precision medicine.

As diagnostic aids, AI technologies have been proven helpful in the diagnosis and treatment of various diseases. However, AI technologies are facing challenges in terms of standardization and normalization. In the near future, AI-powered precision medicine will reach maturity as results become more interpretable. Trust in the technology must be built before it can be fully commercialized.

In the next 3 years, we expect to see people-centric, AI-powered healthcare become a major trend that will span multiple fields of healthcare, including disease prevention, diagnosis, and treatment. The technology will become synonymous with a highly precise compass that allows us to pinpoint diseases and their treatments. With the development of causal inference, AI-powered precision medicine will have good interpretability and provide solid technical support for disease prevention and early diagnosis and treatment.

Privacy-preserving Computation

Privacy-preserving computation is the key to allowing data to flow freely and securely as the demand for processing large volumes of data grows

Introduction

In the digital era, one of the largest challenges is ensuring the security of data while allowing data to flow freely between computing entities. As a result, privacy-preserving computation is gaining traction as a viable solution to this challenge. For a long time, the application of privacy-preserving computation has been limited to a narrow scope of small-scale computation due to performance bottlenecks, lack of confidence in the technology, and standardization issues. However, as more and more integrated technologies, such as dedicated chips, cryptographic algorithms, whitebox implementation, and data trusts, are emerging, privacy-preserving computation will be adopted in scenarios such as processing massive amounts of data and integrating data from all domains, which is the headway made from processing small amounts of data and data from private domains. The adoption will boost new productivity that is powered by data from all domains. In the next 3 years, we will witness groundbreaking improvements in the performance and interpretability of privacy-preserving computation, and witness the emergence of data trust entities that provide data sharing services based on the technology.

Trend Analysis

In the era of digital economy, data is an essential factor of production. However, the road to data sharing among organizations and value extraction from data is fraught with obstacles. For example, we are facing challenges in determining data ownership, safeguarding data security, establishing data-related laws and regulations, and raising public awareness of personal data protection.

Technologies such as secure multi-party computation (SMPC), differential privacy, and trusted computing are all under the umbrella of privacy-preserving computation, which typically involves cryptography, AI, and chip design. Privacy-preserving computation allows computation and analysis while preserving privacy, which is feasible for sharing data among organizations. However, the application of the technology has been limited to a narrow scope of small-scale computation due to performance bottlenecks, lack of confidence in the technology, and standardization issues.

The following breakthroughs in large-scale privacy-preserving computation are expected:

1. Soaring performance and efficiency. Advanced algorithms for homomorphic encryption will hit a critical point, and less computing power will be required in encryption and decryption. Accelerator chips that integrate both software and hardware will optimize computing performance in SMPC and federation learning scenarios. The scale of trusted execution environment (TEE) will become larger as trusted third-parties emerge.

2. Whitebox implementation of privacy computation technology. The interpretability of the technology will be improved, which in turn increases people's confidence in the technology. In addition, privacy computation will be available for integration and barriers of cross-stack and cross-model integration will be lowered.

3. Emerging data trust entities. These entities will provide technologies and operations models as trusted third-parties, which accelerates data sharing among organizations.

Breakthroughs in privacy-preserving computation will become the driving force in the application of the technology on a wider scope. The depth and accuracy of analysis will also increase along with the increasing amount of data that becomes available with the application of the technology. In particular, the fields that depend heavily on data volume, such as business analysis, risk control, academic research, AI, and targeted marketing, will benefit more. Once privacy-preserving computation matures, we can build a world where this technology is the standard for data sharing. In this world, data can be freely shared among all entities, as the boundaries between data owners and data controllers will become clearer, and data security becomes more measurable.

Technological challenges are not the sole concern of privacy-preserving computation. Operations models and compliance requirements are the most prominent uncertainties that we face. In terms of operations models, there are no systematic and comprehensive operations that are in place. As a result, data suppliers are not motivated to share data, and data consumers are not willing to pay for the data as data veracity is not guaranteed. The responsibilities and liabilities of technology suppliers and platforms are not clearly defined. As for compliance requirements, the red lines for privacy computation are not clarified, which might affect the development of the technology. In the foreseeable future, continuous development of both technology and standards are needed to promote privacy-preserving computation.

In the next 3 years, we will witness groundbreaking improvements in the performance and interpretability of privacy-preserving computation, and witness the emergence of data trust entities that provide data sharing services based on the technology. In the next 5 to 10 years, privacy-preserving computation will change the way data is obtained and unleash new productivity powered by data.

Satellite-terrestrial Integrated Computing

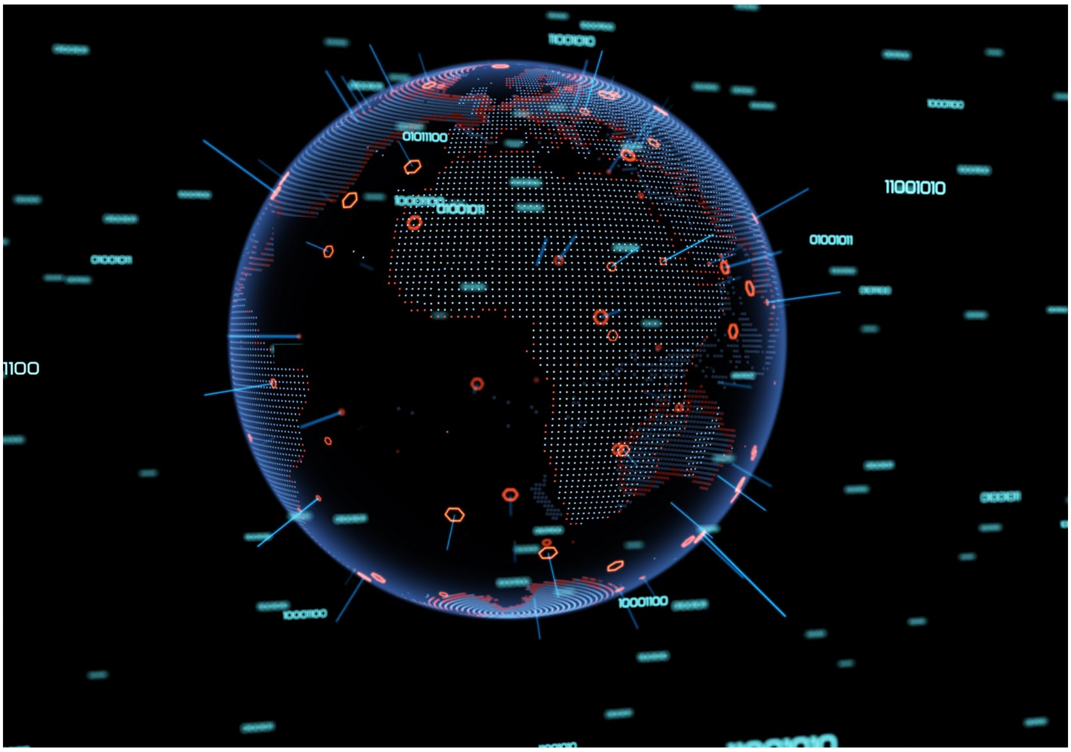
Satellite-terrestrial integrated computing (STC) creates a ubiquitous computing system covering satellites, aerial platforms, and terrestrial and nautical communications networks and computing systems

Introduction

Terrestrial networks and computing systems provide digital services for densely populated areas, while no service is available in sparsely inhabited areas such as deserts, seas, and space. STC connects high-Earth orbit (HEO) and low-Earth orbit (LEO) satellites and terrestrial mobile communications networks, achieving seamless and multidimensional coverage. STC also creates a computing system that integrates satellites, satellite networks, terrestrial communications systems, and cloud computing technologies. This way, digital services can be more accessible and inclusive across the globe. In the next 5 years, satellites and terrestrial systems will work as computing nodes to constitute an integrated network system providing ubiquitous connectivity.

Trend Analysis

Current terrestrial networks and computing capabilities cannot catch up to the growing requirements for connectivity and digital services around the globe. This is especially prominent in sparsely inhabited areas such as deserts, seas, and space. STC creates a computing system that integrates satellites, satellite networks, terrestrial communications systems, and cloud computing technologies.



Satellite networks interconnect with terrestrial communications systems to provide evolving technologies. In this way, global business-to-consumer (B2C) and business-to-business (B2B) applications can easily access high-performance, cost-effective, reliable, and ubiquitous digital connections.

Computing requirements vary with scenarios in the context of ubiquitous connectivity. STC provides abundant computing resources to meet digital service requirements specific to scenarios and industries. This technology underpins the high-quality economic development of the whole society.

STC provides ubiquitous network connections across space, air, land, and sea. The high-speed intelligent broadband communications and comprehensive computing services can satisfy the communications requirements in remote areas, seas, and aerospace, helping us achieve a true Internet-of-everything. Low-latency wide-coverage networks promote cloud-network-device convergence, and incubate new applications in extreme conditions. From an industrial perspective, the reduced complexity of connectivity brought about by STC allows for deeper and broader application of digital intelligence, accelerating the adoption of digital intelligence at scale.

The journey towards full STC capabilities is daunting. First, industries with complex structures have different communications requirements. Traditional satellite communications are expensive and make use of simple technologies and static processing mechanisms, which cannot meet these requirements. Second, the temperature issue in inter-satellite computing has not been overcome. The heat generated in the conversion between optical and electrical signals affects the efficiency of data transmission. Third, hardware for satellite applications is not commonplace. Hardware for terrestrial applications cannot be used in space, and manufacturers need to address issues with hardware when exposed to cosmic rays and spatial interference.

In the next 3 years, we can expect to see a large increase in the number of LEO satellites, and the establishment of satellite networks with HEO satellites. In the next 5 years, satellites and terrestrial networks will work as computing nodes to constitute an integrated network system. This system will provide ubiquitous connectivity.

Cloud-Network-Device Convergence

Cloud-network-device convergence fuels the development of a new computing system and drives the emergence of new types of devices on the cloud

Introduction

The rapid development of new network technologies will fuel the evolution of cloud computing towards a new computing system: cloud-network-device convergence. In this new system, clouds, networks, and devices have a more clearly defined division of labor. Clouds function as “brains”, which are responsible for centralized computing and global data processing. Networks serve as the interconnecting “tracks” that converge various forms of networks on the cloud to build a ubiquitous, low-latency network. Devices offer light-weight, cost-effective interfaces on various forms of applications to deliver smooth, reliable, and immersive user experience.

Cloud-network-device convergence is the catalyst that will drive the emergence of new applications to fulfill more demanding tasks, such as high-precision industrial simulation, real-time industrial quality inspection, and mixed reality. In the next 2 years, we expect to see a surge of applications running on top of the new computing system.

Trend Analysis

Cloud computing has undergone two major development phases. In the cloud-based infrastructure phase, computing power and data are gradually migrated from traditional data centers to the cloud. In the cloud-native phase, advanced cloud-native architectures drive cloud computing towards containerization and serverless computing. The rapid development of network connection technologies is the key to unlocking the third phase of cloud computing and spawns a new computing system, cloud-network-device convergence, which fundamentally changes the way clouds, networks, and devices collaborate with each other.

In this new system, clouds, networks, and devices have a more clearly defined division of labor. Clouds function as “brains”, which are responsible for computing and data processing. Clouds offer high computing efficiency, systematic data processing, and high-precision, efficient, and accessible AI technologies.

Networks interconnect clouds and devices. After being converged on the cloud, advanced network technologies such as optical fibers, 5G, and satellite Internet help build a ubiquitous, low-latency network.

Devices offer easy-to-use, cost-effective, lightweight interfaces that allow users to interact with the cloud. Various models of devices can collaborate on the cloud to accommodate varied interaction requirements. Cloud-device convergence makes diversified scenarios possible on a single type of device and ensures consistent experience across multiple types of devices.

Cloud-network-device convergence is the catalyst that will drive the emergence of new applications. Clouds allow applications to break free of the limited processing power of devices and fulfill more demanding tasks, such as high-precision industrial simulation. The vast network of connected devices allows for low-latency edge-computing, meeting the requirements of various scenarios such as real-time industrial quality inspection. Devices leverage and package the advantages afforded by cloud and network into a single interface – which will become an essential tool powering applications such as cloud PC and cloud gaming.

Two pressing challenges are looming ahead of cloud-network-device convergence. One is making breakthroughs in network technologies. Networks play a significant role in cloud-network-device convergence. Network quality, cost, and coverage may become a potential bottleneck of the new system. Therefore, new network technologies such as 5G and satellite Internet must be continuously improved and developed to ensure wide coverage and sufficient throughput. The other is information security. Security is always a top concern of enterprises who store and process data on the cloud. Enterprises have high requirements for security technologies, such as data encryption, data governance, secure computing, and privacy-preserving computation.

In the next 2 years, we expect to see a surge of application scenarios on top of the converged cloud-network-device system. This will drive the emergence of new types of devices and promise more high quality and immersive experience for users.

Extended Reality (XR)

XR glasses will become an important interface to the next-generation Internet

Introduction

The development of technologies such as cloud-edge computing, network communications, and digital twins brings XR into full bloom. XR glasses promise to make immersive mixed reality Internet a reality. This technology plants the seed that will sprout into a new industrial ecosystem that encompasses electronic components, devices, operating systems, and applications. XR will reshape digital applications and revolutionize the way people interact with technology in scenarios such as entertainment, social networking, office, shopping, education, and healthcare. In the next 3 years, we expect to see a new generation of XR glasses that have an indistinguishable look and feel from ordinary glasses entering the market and serving as a key entry point to the next generation of Internet.

Trend Analysis

The advent of the Internet kicked off a new digital era, and its evolution has constantly brought changes to the landscape of the digital industry. As people shifted from computers to smartphones, operating systems and applications also underwent massive changes. The productization of VR and AR has paved the way for XR to transform the digital era.

XR renders information in an immersive 3-dimensional experience, and users can intuitively interact with the world to obtain information. An XR Internet encompasses four elements: 1) hardware such as XR glasses; 2) content for applications such as entertainment, shopping, and social networking; 3) AI that powers technologies such as spatial intelligence and digital twin; 4) infrastructures such as 5G and cloud computing. Among these elements, hardware and content will be the first to be developed. Hardware is the carrier of Internet platforms and is fundamental for users to obtain data. XR glasses will provide access to the XR Internet, and the cloud-network-device convergence will reduce the size, weight, and latency of the glasses. Content is expected to find its roots in entertainment, social networking, and remote work, and will be gradually expanded to scenarios where remote interactions are required, such as shopping, education, and healthcare.

XR transforms the way people interact with technology. XR can be used to simulate the real world and overcome the limitations of physical distances. Applications such as remote education, healthcare, and working can be implemented in new ways. XR can also create alternate worlds and allow users to establish fresh identities in gaming and social networking applications. XR will restructure the Internet industry and give birth to new industries, especially in the fields of electronic components, devices, operating systems, and applications.

XR is still in its infancy and needs to overcome several technological challenges before it can deliver a truly immersive experience. First, the computing power, display resolution, size, and battery life of AR/VR/MR glasses are still far from ideal. Second, visual experience is currently the sole focus of the industry. For XR to become a reality, more investment is required to provide stimulation to other senses such as touch, smell, and taste. Finally, the world of XR is fundamentally built upon user data, which is ever expanding and evolving. Privacy-preserving computation will become a core challenge in the collection, storage, and management of such data.

In the next 3 years, we expect to see the release of a new generation of XR glasses that will seamlessly blend AR and VR technologies, take advantage of cloud-device integration and advanced optics, and weigh the same as ordinary glasses. This technology will become the main interface between the real world and the XR world, and is expected to be widely used.

Bibliography

[1] Tunyasuvunakool, K., Adler, J., Wu, Z. et al. Highly accurate protein structure prediction for the human proteome. Nature 596, 590–596 (2021). https://doi.org/10.1038/s41586-021-03828-1.

[2] M. Ziatdinov, et al., “Building and exploring libraries of atomic defects in graphene: Scanning transmission electron and scanning tunneling microscopy study,” Sci. Adv. 5:eaaw8989 (2019). DOI: 10.1126/sciadv.aaw8989.

[3] Basu, S., Kumbier, K., Brown, J. B., & Yu, B. (2018). Iterative random forests to discover predictive and stable high-order interactions. Proceedings of the National Academy of Sciences, 115(8), 1943-1948.

[4] Chang, M.C.,et al. Accelerating neutron scattering data collection and experiments using AI deep super-resolution learning (arXiv:1904.08450, 2019).

[5] R. Bommasani, D.A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, ..., P. Liang, On the opportunities and risks of foundation models，Center for Research on Foundation Models (CRFM) — Stanford University (2021).

[6] S. Avin, R. Gruetzemacher, J. Fox，Exploring AI Futures Through Role Play，The 2020 AAAI/ACM Conference on AI for Ethics and Soceity (2020).

[7] S. Baum, B. Goertzel, T. Goertzel，How long until human-level AI? Results from an expert assessment, Technological Forecasting and Social Change, 78 (1) (2011), pp. 185-195.

[8] Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Liang Zhang, Wentao Han, Minlie Huang, et al. 2021a. Pretrained models: Past, present and future.

[9] J. Lin, A. Yang, J. Bai, C. Zhou, L. Jiang et al., “M6-10t: A sharing delinking paradigm for efficient multi-trillion parameter pretraining,” arXiv:2110.03888, 2021.

[10] OpenAI, AI and Compute, 2018.

[11] Cisco, Global Index, 2018.

[12] Riccardo Marchetti, Cosimo Lacava, Lee Carroll, Kamil Gradkowski, and Paolo Minzioni, "Coupling strategies for silicon photonics integrated chips [Invited]," Photon. Res. 7, 201-239 (2019).

[13] Rudolph, T. Why I am optimistic about the silicon-photonic route to quantum computing. APL Photonics 2, 030901 (2017).

[14] 国家能源局，《关于征求 2021 年可再生能源电力消纳责任权重和 2022-2030 年预期目标建议的函》, 2021.

[15] B. K. Bose, "Artificial Intelligence Techniques in Smart Grid and Renewable Energy Systems—Some Example Applications," in Proceedings of the IEEE, vol. 105, no. 11, pp. 2262-2273, Nov. 2017, doi: 10.1109/JPROC.2017.2756596.

[16] S. Ramachandran, "Applying AI in Power Electronics for Renewable Energy Systems [Expert View]," in IEEE Power Electronics Magazine, vol. 7, no. 3, pp. 66-67, Sept. 2020, doi: 10.1109/MPEL.2020.3012009.

[17] ABB, Industry Survey, 2021.

[18] International Federation of Robotics, Executive Summary World Robotics 2021 Industrial Robots, 2021, 12-16.

[19] Jackson, A., Mentzer, N., Kramer, R., & Zhang, J. (2017, June). Maker: Taking soft robotics from the laboratory to the classroom. Paper presented at the Make It! Event during the 2017 ASEE Annual Conference & Exposition, Columbus, OH. Retrieved from https://peer.asee.org/27741.

[20] Wang, Z., Chen, M.Z.Q., & Yi, J. (2015). Soft robotics for engineers. HKIE Transactions, 22(2), 88-97.

[21] Allied Market Research, AI in Healthcare Market, 2021.

[22] McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. Nature 2020; 577: 89-94.

[23] U.S. Food & Drug Administration. Evaluation of automatic class III designation for Osteo Detect, 2018. Available at: https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN180005.pdf. (Accessed: 2nd October 2019).

[24] Gartner, Top Strategic Technology Trends for 2022, 2021.

[25] 中国信息通信研究院 , 隐私计算白皮书 (2021), 2021.

[26] SIA, 2020 Top-level global satellite industry findings, 2021.

[27] CHEN S Z, SUN S H, XU G X, et al. Beam-space multiplexing: practice, theory, and trends, from 4G TD-LTE, 5G, to 6G and beyond[J]. IEEE Wireless Communications, 2020, 27(2): 162-172.

[28] PORTILLO I D, CAMERON B, CRAWLEY E F. A technical comparison of three low earth orbit satellite constellation systems to provide global broadband[J]. Acta Astronautica, 2019, 6(159): 123-135.

[29] Patterson, D.P., 1998, March. Teledesic: a global broadband network. In Aerospace Conference, 1998 IEEE (Vol. 4, pp. 547-552).

[30] Uptime Institute, Annual Data Center Survey, 2020.

[31] Cisco visual networking index: Forecast and methodology, 2016–2021 VNI Cisco - CISCO White paper, 2017.

[32] A. R. Biswas and R. Giaffreda, "IoT and cloud convergence: Opportunities and challenges," 2014 IEEE World Forum on Internet of Things (WF-IoT), 2014, pp. 375-376, doi: 10.1109/WF-IoT.2014.6803194.

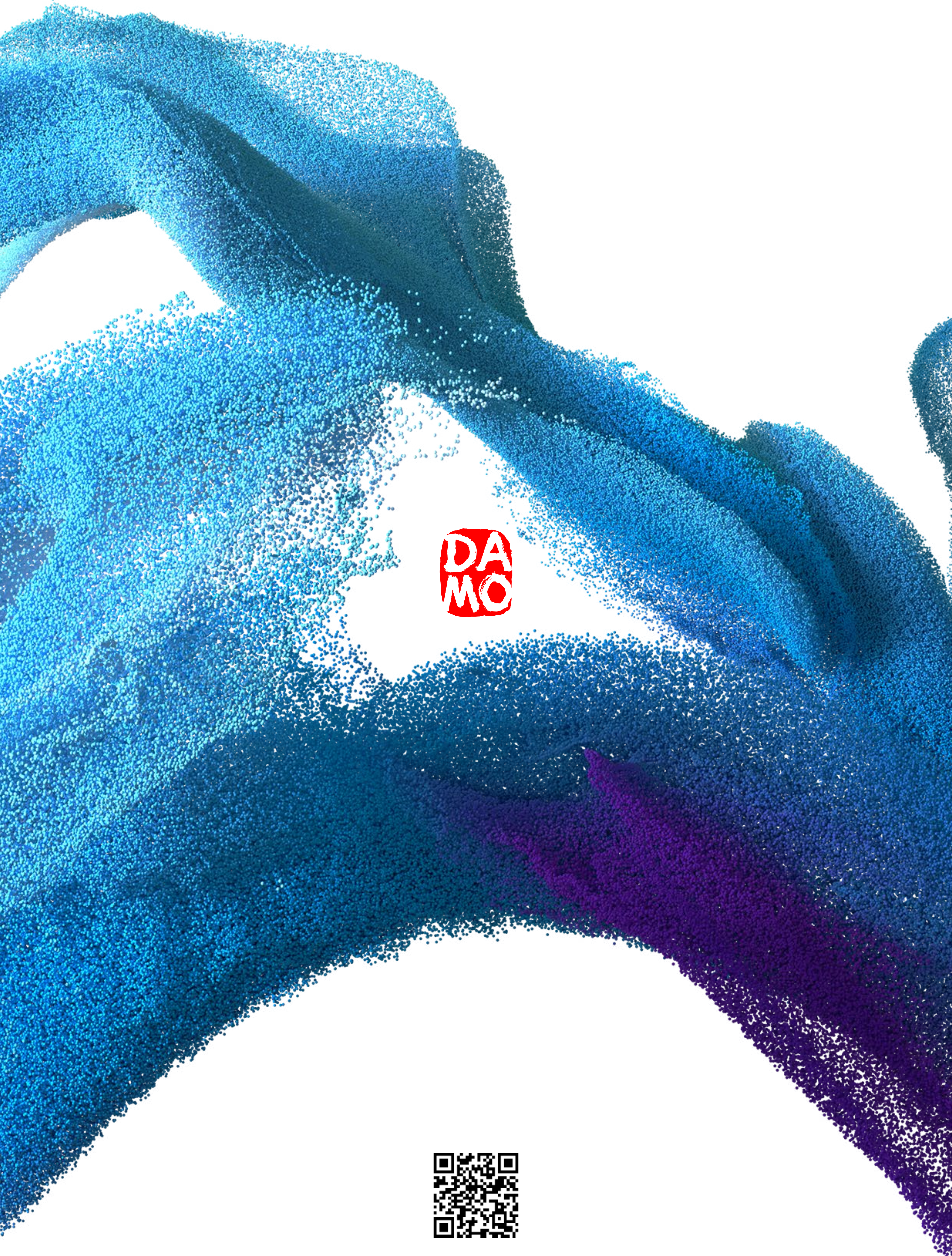
[33] IDC, IDC 全球增强与虚拟现实支出指南 , 2020.

[34] Gartner, Hyper Cycle for Emerging Technologies, 2021.

[35] Cisco, 2018 Cisco Complete VNI Forecast and Trends Update, 2018.

[36] Dionisio, J. D. N., Burns, W. G., & Gilbert, R. (2013). 3D Virtual worlds and the metaverse: Current status and future possibilities. ACM Computing Surveys (CSUR), 45(3), 34.

[37] Girvan, C. (2018). What is a virtual world? Definition and classification. Educational Technology Research and Development, 66(5), 1087–1100.



Visit our Website for More