Convergence of AI and Communication

vivo Communications Research Institute
October 2023

CONTENTS

01

Chapter 1

Introduction

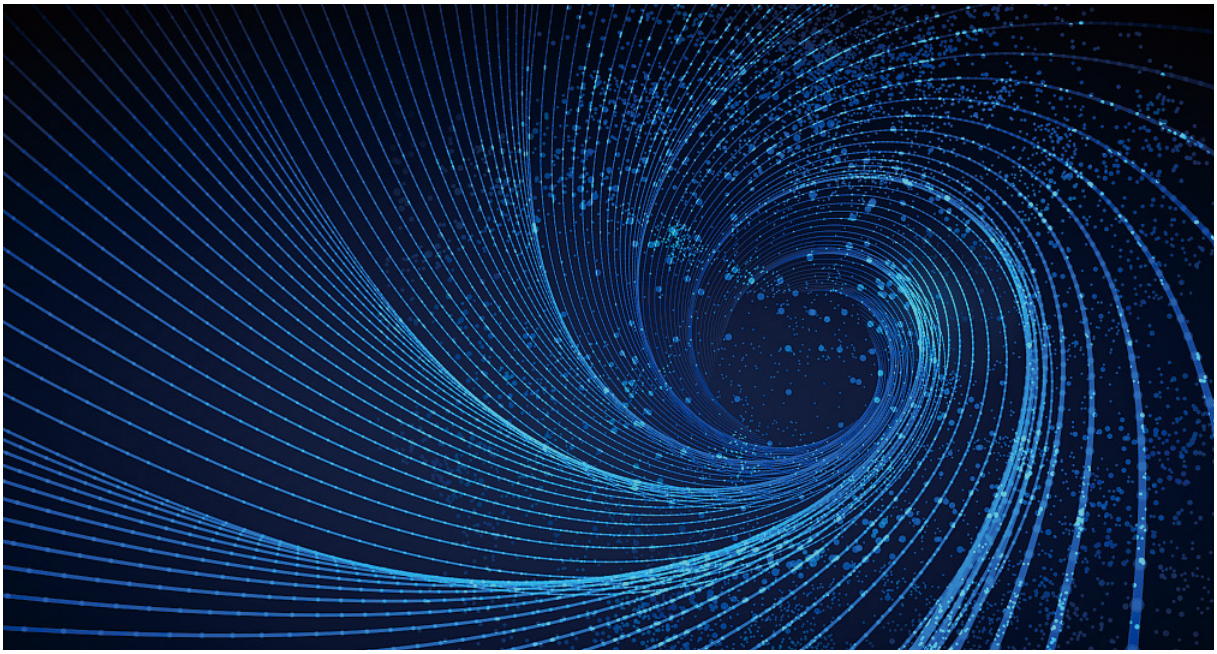
From 2020 to 2022, vivo Communications Research Institute released 3 white papers, "Digital Life 2030+", "6G Vision, Requirements and Challenges" and "6G Services, Capabilities and Enabling Technologies". These three papers have envisioned the bright future of 6G and introduced enabling technologies. In recent years, applications of artificial intelligence (AI) in mobile communications have sprung up and convergence of AI and mobile communication is expected to drive the evolution of future communication paradigms and network architecture. In this regard, this white paper elaborates on the use cases and design principles for the convergence of AI and communication, hopefully contributing to the development and realization of 6G AI.

02

Chapter 2

Driving Forces and Use Cases of the Convergence of AI and Communication

2.1 Driving Forces and Use Cases of the Convergence of AI and Communication

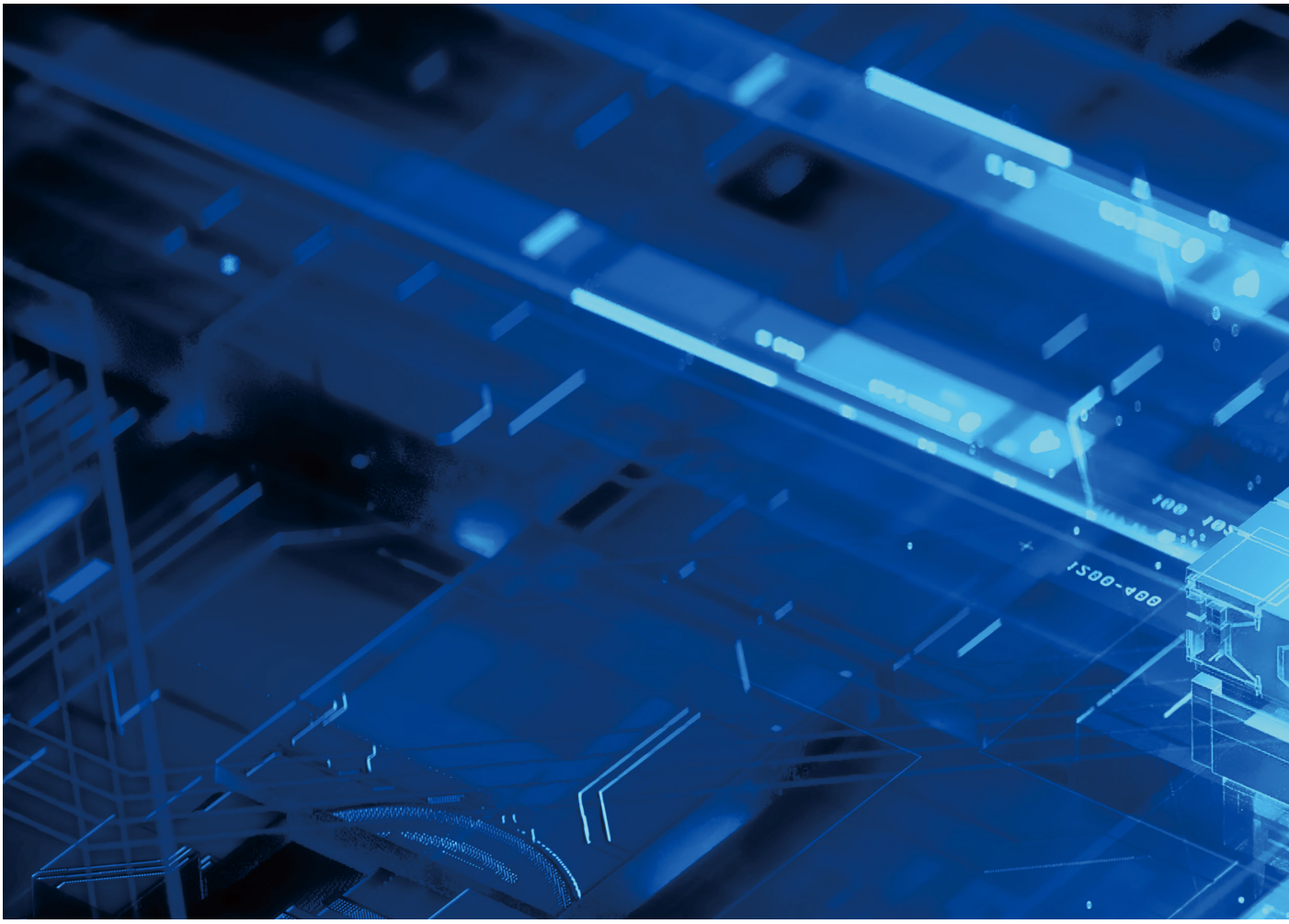


In the past decade, AI technology has experienced rapid development and is stirring up a new round of technology revolution. AI is a data-driven technology that can extract features from a large amount of data through machine learning tools such as neural networks, and further execute actions such as judgment, classification, prediction, decision, and content generation. Currently, AI has successfully solved a series of problems that were difficult to handle in the past, and has achieved a great success in many areas such as image recognition and natural language processing in computer science, and motion control and trajectory planning in robotics [1].

At the same time, mobile communication systems are also continuously evolving towards higher throughput, lower latency, higher reliability, larger number of connections, higher spectrum utilization, etc. With the tremendous increase in the requirements of mobile communication, traditional methods have encountered bottlenecks, for which AI is expected to provide more efficient solutions.

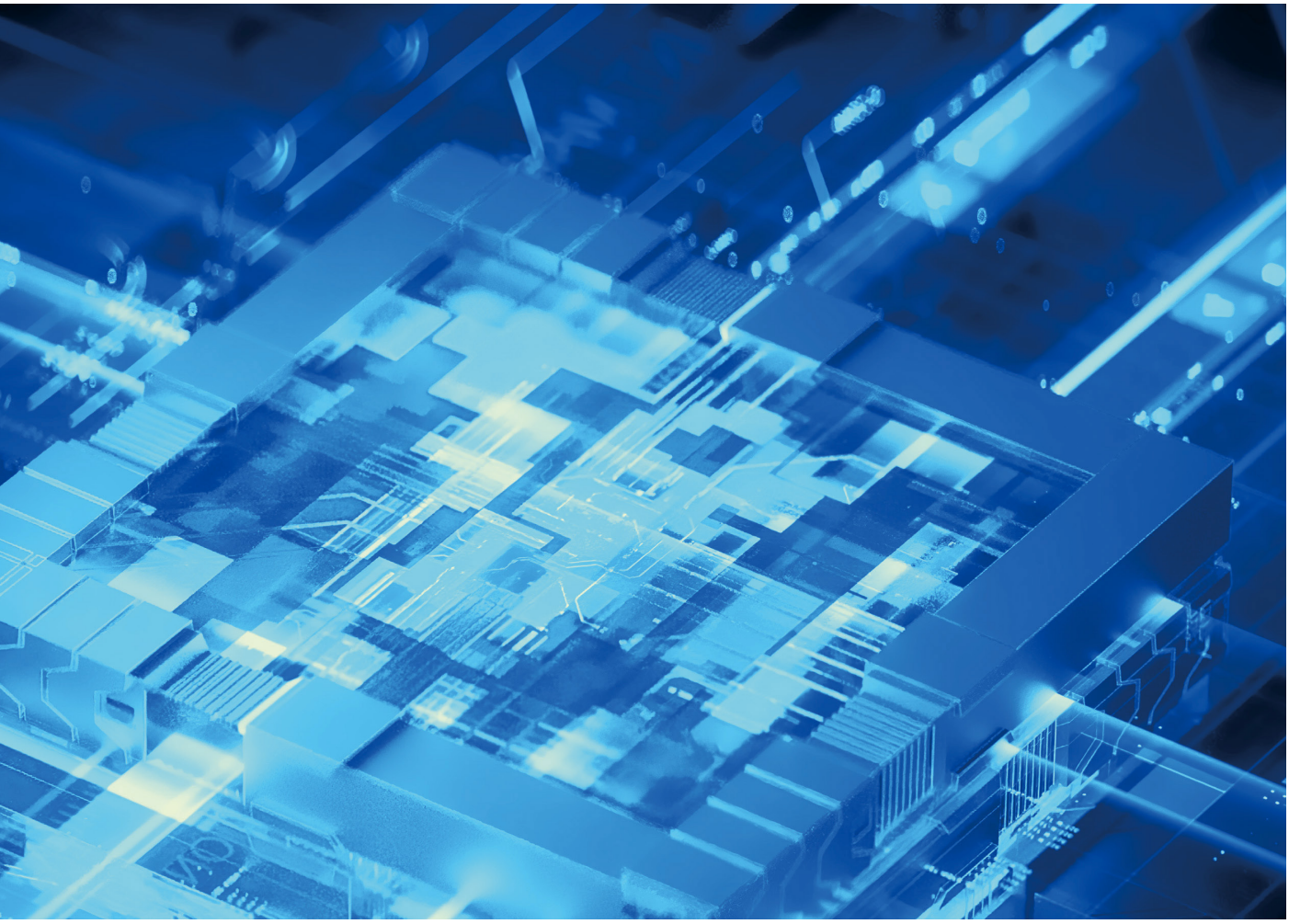
Based on the research by both academia and industry, we have summarized the main directions or scenarios where AI can play an important role and provide significant value in mobile communication systems.

(1) Scenarios in the communication system where implicit relationships, features or knowledge (especially localized ones) has significant impacts on specific functions but hard to be established by traditional methods. Examples contain the impact of wireless channels, channel variation patterns, the relationship between user location and wireless channels, the imperfectness of power amplifiers (PA), and the variation patterns of business traffic. Using tools such as neural networks, AI can extract these implicit relationships, features, or knowledge from a large amount of mobile communication data so as to model complex problems more accurately, thereby improving the performance of communication system.



(2) Problems in communication systems to which closed-form solutions are not easy to be obtained by traditional methods in expected time or problems without explicit closed-form solutions.. Examples contain global wireless resource allocation, multi-user pairing, coverage optimization, and capacity optimization. In traditional schemes, these problems are generally in the form of optimization problems or traversal problems with very high complexity. AI can map the relationship between input information (including states, conditions, historical results, etc.) and potential solutions through data-driven or model-driven approaches, thus reducing the complexity of the communication system.

(3) The problem of joint optimization of multiple related modules in communication systems. Examples contain cross-layer optimization, joint optimization of multiple related modules in multiple input multiple output (MIMO) signal processing, and joint source-channel coding. Currently, the optimization of different communication modules is done separately while the joint optimization is rarely deployed due to complexities involved. AI can model multiple related functional modules as a neural network, transforming the complex multi-module problems into simple data fitting or regression problems, thus obtaining near global optimal solutions.



The aforementioned application of AI as a tool to assist mobile communication is known as AI for network (AI4NET) or internal AI services. In addition, in the future, mobile networks will also provide relevant support for a large number of AI services. With the development of AI, more and more AI services are needed by verticals. Many of these AI services require mobile networks to provide higher throughput, lower latency, extreme capacity and ubiquitous coverage. For example, in the medical industry, AI-assisted diagnosis and treatment require transmission of examination results with high resolution and interaction with low latency; autonomous driving and drone control require the real-time feedback of a large amount of sensor data and the transmission of command messages; in smart factories, the motion and trajectory information of robots should be transmitted with low latency and high reliability to achieve efficient command and scheduling. Therefore, it is necessary to provide relevant support for AI services through mobile networks with effective allocation of communication, computing, and storage resources. This support for AI provided by mobile networks is called network for AI (NET4AI) or external AI services. In this paper, AI4NET is the main focus with the understanding that NET4AI would also be supported effectively by 6G system as a service.

AI has many application directions in mobile communication systems, and there is a great potential for improving system performance and reducing complexity. Below, we present the significance of AI in mobile communication through multiple use cases.

2.2 Use Cases of Convergence of AI and Communication

At first, we present applications of AI in extracting implicit relationships, features, or knowledge in communication systems. The representative use cases include AI-based channel state information (CSI) feedback enhancement, channel estimation, beam prediction enhancement, positioning enhancement, network selection, signaling storm prediction, user mobility optimization, and PA nonlinearity suppression.

AI-based CSI feedback enhancement

AI-based CSI feedback enhancement can be further categorized into two sub-use cases, AI-based CSI compression and AI-based CSI prediction.

In CSI compression, AI techniques are utilized to compress and reconstruct multi-dimensional channel state information, thereby reducing the CSI feedback overhead and/or improving the CSI recovery accuracy. The mainstream solution for CSI compression is the encoder-decoder architecture, where an encoder model deployed on the user device side compresses the CSI and feedback the generated bit sequence, while a decoder model deployed on the network side decodes the received feedback bit sequence to reconstruct the CSI. Generally, the encoder model used for CSI compression and the decoder model used for CSI reconstruction need to be paired, i.e., a decoder can only effectively reconstruct one or more compressed CSI generated by the corresponding encoders. Fig. 2-1 shows the spectrum efficiency gain of AI-based CSI compression over traditional Release 16 Type-II codebook-based CSI compression (without AI). Detailed simulation parameters can be referred to [2] and the references therein. From Fig. 2-1, it can be observed that AI-based CSI feedback is able to achieve approximately 10% spectrum efficiency gain given the same feedback overhead.

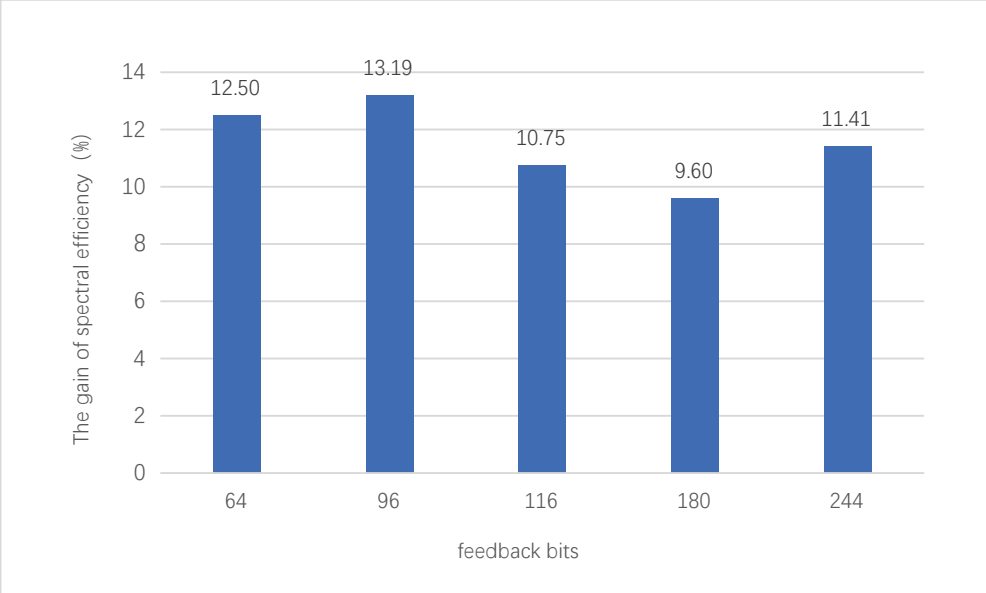


Fig. 2-1. Spectral efficiency gain of AI-based CSI compression over conventional CSI compression based on Release 16 Type II codebooks (non-AI)

AI-based CSI prediction extracts the hidden patterns of CSI variation over time and predicts the CSI of future moments based on historical CSI. In this way, the aging issue of CSI feedback can be compensated. The basic scheme of AI-based CSI prediction is to use multiple historical CSIs as inputs to the neural network, and then obtain the CSIs of the specified future moments through the neural network. Fig. 2-2 shows the average throughput gain that can be obtained by the AI-based CSI compared with the scheme without prediction and the scheme using autoregressive (AR) based non-AI CSI prediction under different resource utilization (RU). The detailed simulation parameters are provided in [3]. It can be seen that AI-based CSI prediction can achieve considerable gains especially at the scenario with higher RU.

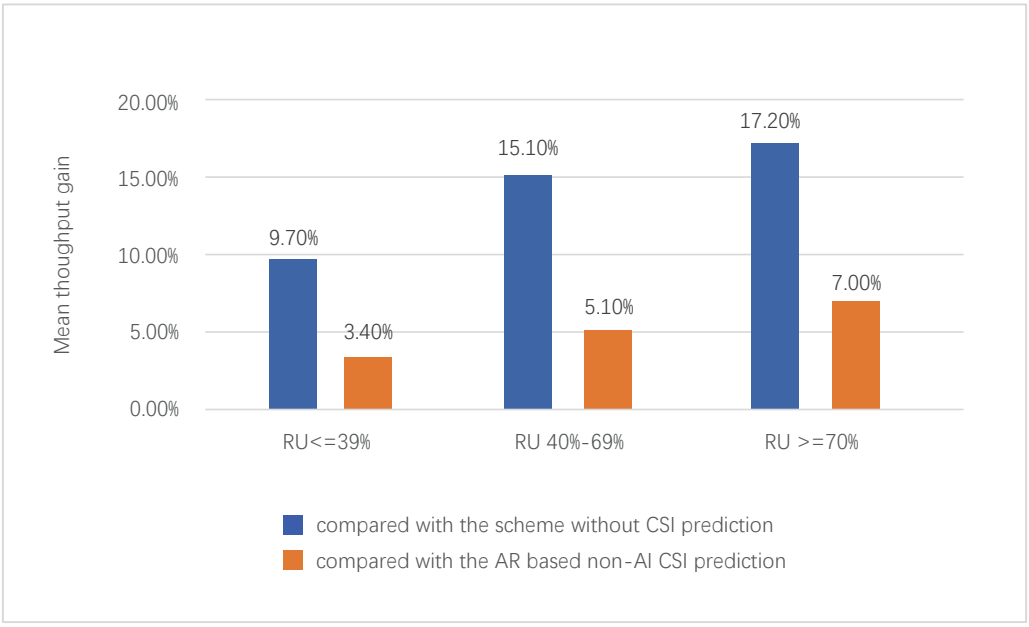


Fig. 2-2. Mean throughput gain of AI-based CSI prediction

AI-based channel estimation

Mobile communication systems require the use of reference signals for channel estimation. As the number of antennas at base stations (BSs) and user equipment (UEs) increases, the overhead of reference signals gets heavier. To address this issue, AI technology can be used to explore the correlation among channel responses on different transmission resources (such as time, frequency, and space), and design low-cost, high-precision channel estimation methods.

Taking de-modulation reference signal (DMRS) as an example, channel estimation can be realized by using the channel estimation results at the DMRS resources as inputs to the AI model, and the channel estimation on all the resources as outputs of the AI model (as shown in Fig. 2-3). Since the relationship of CSIs on different transmission resources is generally nonlinear in nature, AI can achieve higher estimation accuracy or lower reference signal overhead compared to traditional interpolation methods. In Fig. 2-4, the AI-based DMRS channel estimation is validated with the simulation parameters detailed in the literature [4]. The blue line in the figure represents the DMRS estimation implemented with the conventional linear minimum mean square error algorithm at 50% overhead (6/12) in the frequency domain, whose implementation requires the estimation results of the tracking reference signal (TRS) as auxiliary information. The red line represents the performance obtained using AI-based DMRS channel estimation at different frequency domain overheads (without relying on the TRS). It can be seen that the AI-based scheme achieves higher estimation accuracy and lower block error rate (BLER) using a lower reference signal overhead and without the assistance of TRS.

In 2021, vivo sponsored the AI-based channel estimation track in the second wireless communication AI competition of the IMT-2020 (5G) Promotion Group. 651 teams competed for over two months and designed sophisticated models, significantly improving the accuracy of channel estimation, and pushing forward the convergence of AI and communication.

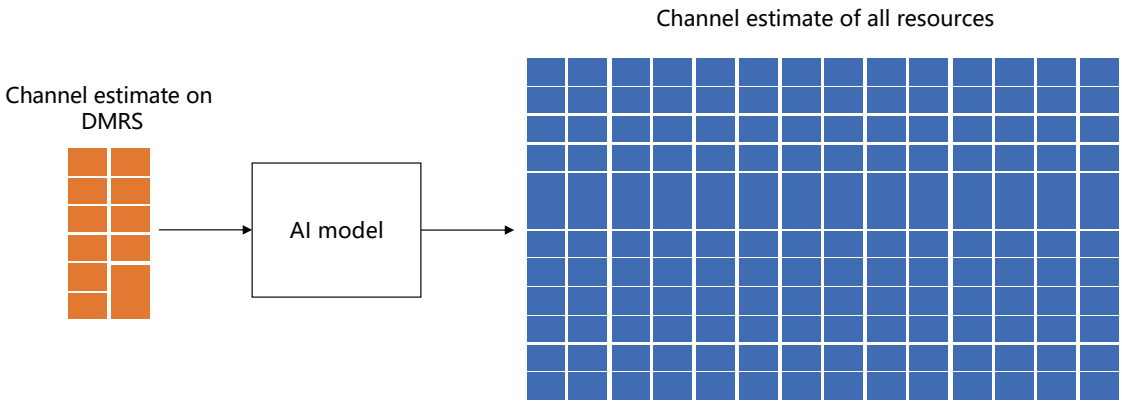


Fig. 2-3. Illustration of AI-based DMRS channel estimation

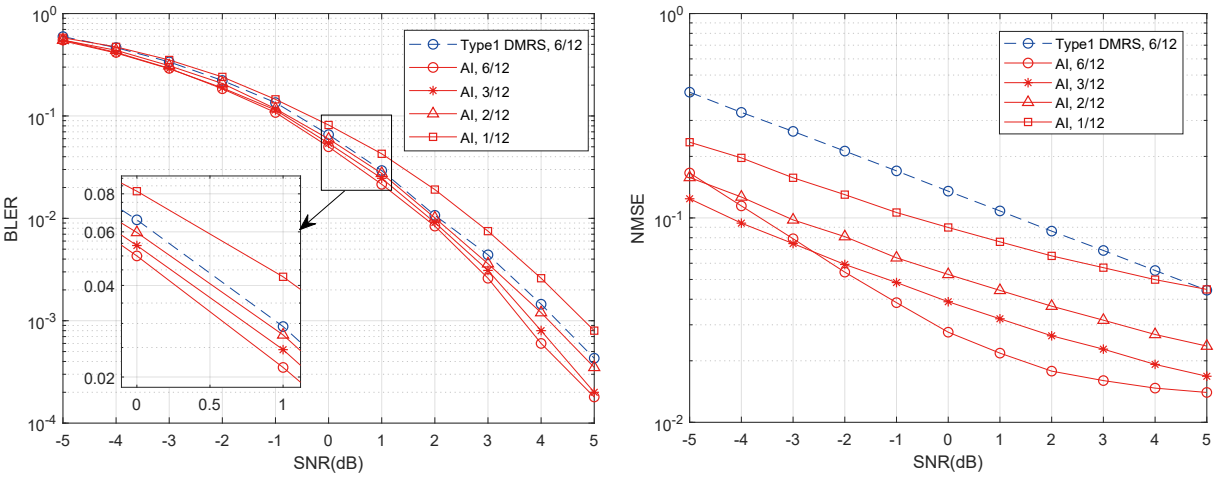


Fig. 2-4. Evaluation of AI-based DMRS channel estimation

In addition, for channel estimation of channel state information reference signal (CSI-RS), compressed sensing techniques such as approximate message passing (AMP) can be considered to reduce the airspace reference signal overhead, i.e., using the sparse port reference signal to estimate the channel of all ports. However, the main disadvantage of the AMP algorithm is that the optimal sensing matrix and the shrinkage function in the iterative estimation algorithm cannot be obtained explicitly. In this regard, as shown in Fig. 2-5, the process of the reference signal passing through the sensing matrix and the iterative process of solving the AMP algorithm can be unfolded into two neural networks using the model-driven idea, in which the first neural network determines the optimal sensing matrix, and the second neural network reconstructs the AMP iterative algorithm. With end-to-end supervised training over a large amount of data, the optimal sensing matrix and shrinkage function can be determined to improve the accuracy of channel estimation based on compressed sensing.

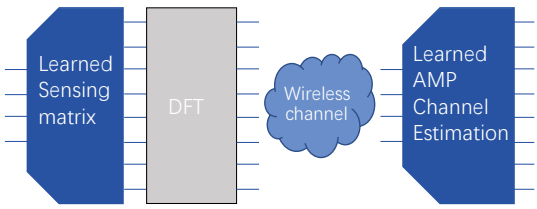


Fig. 2-5. Illustration of AI-based low-overhead CSI-RS channel estimation

AI-based enhancement of beam prediction

AI-based enhancement of beam prediction includes spatial domain beam prediction and temporal domain beam prediction. Spatial domain beam prediction uses a small set of beam measurements from set B to predict the beam information of the complete set A, where set B can be a subset of set A or different from set A (e.g., set B can be wide beams while set A can be narrow beams). The method of temporal domain beam prediction is similar to that of spatial domain beam prediction, but the input of the neural network includes measurement results from multiple historical occasions, and the output of the neural network includes prediction results from multiple future occasions. The core idea of beam prediction is to improve the accuracy of beam prediction and reduce the complexity of beam prediction while reducing the cost of beam measurement. Therefore, assistance information may also be used in the input of the neural network to further improve the accuracy of beam prediction.

We have verified the AI-based beam prediction, and the specific simulation parameters are detailed in reference [5]. When using 1/8 of the beam measurement resources of set A as set B, the AI-based solution improves the Top-1 beam prediction accuracy relative to the non-AI solution from 12.5% to about 60%. When using 1/4 of the beam measurement resources of set A as set B, the AI-based solution improves the Top-1 beam prediction accuracy relative to the non-AI solution from 25% to about 80%. Therefore, the performance of the AI-based solution is superior to the non-AI solution when using the same beam measurement resources.

AI-based enhancement of positioning

AI-based positioning is to establish a mapping between channel responses and UE position through AI technologies. Based on the output type of AI model, it can be further divided into two sub-use cases: direct AI positioning and AI-assisted positioning. Direct AI positioning refers to the case in which the AI model directly estimates the UE position based on the channel responses between the UE and multiple transmission/reception points (TRPs). AI-assisted positioning refers to the case in which the AI model estimates the intermediate features, such as time of arrival (TOA), based on the channel responses between the UE and multiple TRPs. Then, with the position of multiple TRPs and intermediate features, the UE position can be further calculated. Table 2-1 compares the positioning accuracy of different positioning methods in indoor-factory non-line of sight (NLOS) scenarios. Specific simulation parameters can be found in [6]. Compared with traditional localization methods based on downlink time difference of arrival (DL-TDOA) measurement of the first path, AI-based positioning reaps a significantly improved performance in terms of positioning accuracy.

Table 2-1. Positioning accuracy of different positioning methods in indoor-factory NLOS scenarios

Positioning methods	Measurement	Model output	Positioning accuracy (90% UEs)
DL-TDOA	First-path delay	Location	32.12m
Direct AI positioning	Channel impulse response	Location	0.99m
AI assisted positioning	Channel impulse response	TOA	0.73m

AI-based power amplifier (PA) nonlinearity suppression

Orthogonal frequency division multiplexing (OFDM) system has the advantages of anti-interference, anti-fading, and high spectrum utilization, and thus is widely used in the era of 4G and 5G. However, OFDM also has the issue of high peak-to-average power ratio (PAPR), which can cause PAs to enter the saturation region leading to nonlinear distortion. In order to improve the linearity and efficiency of the system, it is necessary to reduce the peak-to-average power ratio of the OFDM signal. Recently, deep learning has inspired researchers to use data-driven or model-driven methods to counteract the effects of PA nonlinearity. As shown in Fig. 2-6, in the Tone Reserve (TR) algorithm for reducing PAPR, AI technology can be used to capture the implicit relationship between the OFDM signal and the optimal peak cancelling signal. In the inference stage, when a new OFDM signal is input, AI will provide the corresponding optimal peak cancellation signal. Through simulation verification, in the case where the reserved subcarriers account for 25% of the total subcarriers, the AI-assisted TR algorithm can reduce PAPR by an additional 3dB compared to the traditional TR algorithm. Similarly, under the same PA power back-off, the AI-based TR algorithm can achieve lower error vector magnitude (EVM) [7].

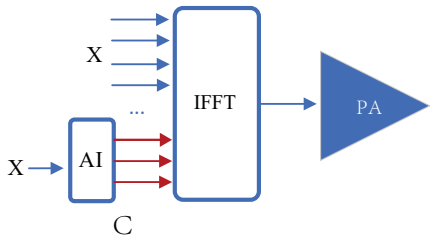


Fig. 2-6. Schematic diagram of AI-based TR technology

Similarly, in the digital pre-distortion (DPD) technology, AI can learn the rules of nonlinear transformation of signals by studying a large number of PA input and output signals. As shown in Fig. 2-7 by performing a reverse operation (i.e., DPD) on the signal input to the PA through a neural network, the nonlinear distortion generated by the PA on the original signal can be cancelled out. Similarly, through simulation verification, AI-based DPD can reduce EVM by approximately 5% compared to traditional DPD [7].

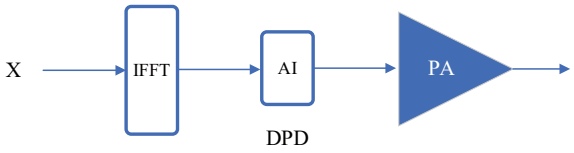


Fig. 2-7. Schematic diagram of AI-based DPD

AI-based mobility optimization

Mobility management is a fundamental mechanism for mobile communication systems, which can provide UE with service continuity. For the traditional method, it is challenging for a trial-and-error-based scheme to achieve nearly zero-failure handover. In this case, AI can be utilized to optimize the mobility management. AI-based mobility optimization mainly relies on the prediction of UE trajectory and cell resource status, to determine the appropriate target cell and handover timing for the UE. The probabilities of handover failure and unintended events (e.g., handover to the wrong cell) are expected to be reduced. The typical inputs of model inference include radio resource management (RRM) and the location of UE. The model can be deployed at either network side or UE side. If the model inference functionality is deployed on the network side, the UE needs to report the measurements and/or location. This may be problematic if users turn off location reporting due to privacy concerns. Compared with the network-sided model, the UE-sided model can achieve more real-time RRM measurement prediction, target cell prediction or unexpected event prediction and thus improve user mobility experience.

Taking RRM prediction as an example, the procedure of AI-based mobility optimization is illustrated in Fig. 2-8. To be specific, when the trigger condition of the measurement report is met at T_0, UE can predict the RRM measurement of serving cell and neighbor cells within the duration of TTT (time-to-trigger). If the RRM measurement prediction results within the TTT meet the reporting condition, the UE may send the measurement report immediately to avoid handover failure due to the inability to receive the handover command. In addition, the UE may report the predicted RRM measurement for a longer period, which can be used for target cell and handover timing determination to avoid the occurrence of radio link failure shortly after the handover.

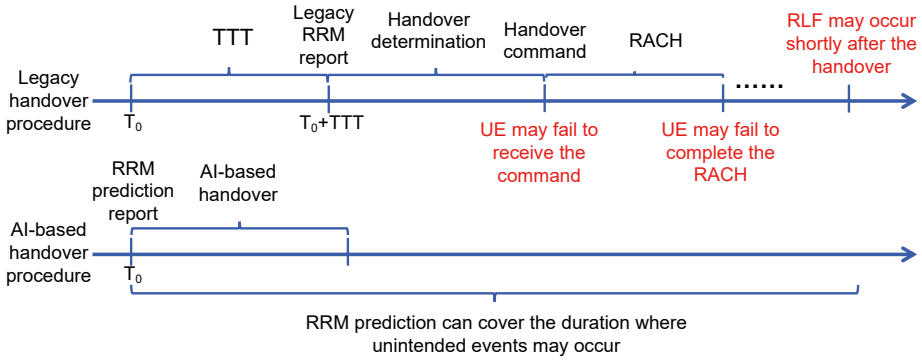


Fig. 2-8. AI-based mobility optimization with RRM prediction

The corresponding evaluation results in a dense urban scenario are illustrated in Fig. 2-9. Compared with legacy handover, the AI-based handover can significantly reduce the probability of handover failure and unintended events (e.g., short time of stay, ping-pong handover).

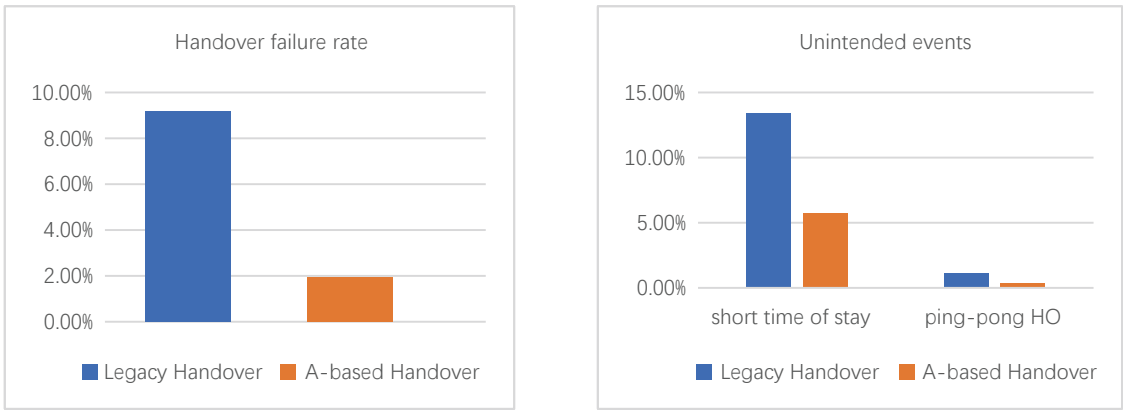


Fig. 2-9. Handover failure rate (left) and ping-pong handover rate (right) of legacy handover and

AI-based UE Access Network Selection

There are scenarios where multiple radio access types coexist and complement each other in the mobile communication systems. Taking the current 5G as an example, although 5G deployment has been carried out on a large scale, but ubiquitous 5G NR (New Radio) coverage may not be accessible for a lot of areas. This similar situation may also exist in the 6G era with the expectation that several radio access technologies (RATs), e.g., 4G, 5G and 6G, coexist for some time.

Different access technology has different advantages, e.g., 4G LTE has lower frequency spectrum and higher latency, but better network coverage; vice versa, 5G NR has higher frequency spectrum and lower latency, but limited coverage; for 6G, there may still exist tradeoff between the coverage and capacity. From the perspective of applications, different applications have different requirements for network KPIs, for example, augmented reality (AR), virtual reality (VR), etc. with high requirements for rate and latency are more suitable for using high-frequency and large-bandwidth RAT, while those services with high mobility requirements is more proper to choose RAT with a large coverage area to avoid frequent switching. As shown in Fig. 2-10, AI can be used to analyze the historical data of end users and 6G network, obtain AI model implying UE behavior and network performance characteristics. Based on the AI model, it can predict what kind of service the UE is using or will use, and which cell or specific location it will move to, and predict corresponding network load and performance for different RAT types or Access types. Based on this information, the 6G network can thus decide the optimal access network selection strategy.

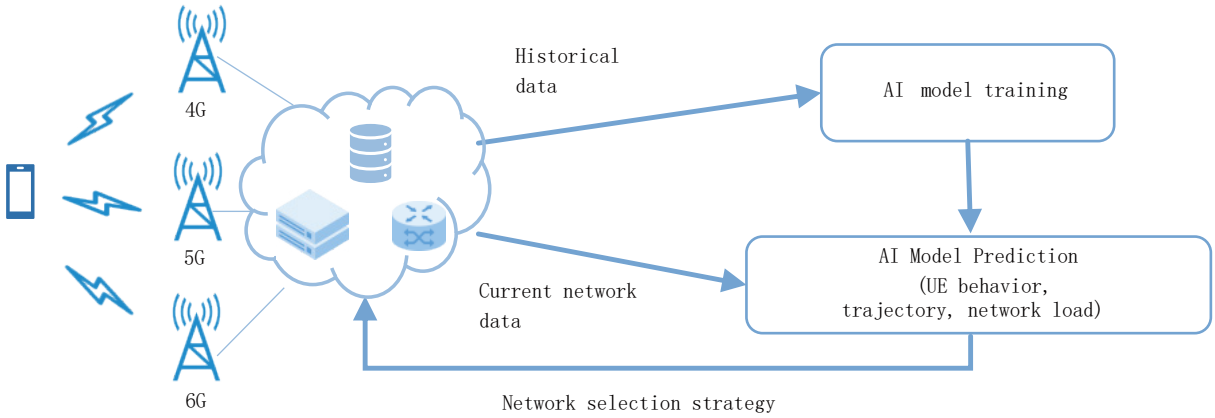


Fig. 2-10. Illustration of AI-based UE Access Network Selection

AI-based signaling storm prediction

With the rapid growth of network subscribers, the risk of signaling storms is growing significantly. Empirically, signaling storms seem to lead to failures with a large scope and long duration, having a large negative impact on user experience and the reputation of the operator, so the prevention of network signaling storms is the most important challenge for operators. In the past, operators could only backwardly detect signaling storms at the network operation and maintenance level, and then artificially analyze the root cause and traffic diversion. This process often depended on the experience of experts, which usually took a few hours or even several days.

With the use of AI technology in mobile communication networks, 6G networks is expected to perform signaling storm prevention with the assistance of AI. As shown in Fig. 2-11, the 6G network AI entity collects a large amount of data such as network metrics, behavioral performance, network configurations, UE behavior, etc., in both normal scenarios and network signaling storm scenarios. Based on these data, the AI model is trained to mine the correlation between these features and the occurrence of signaling storms. This AI model is used to predict the probability of signaling storms occurring in the network, and even accurately predict the network performance and other derivative consequences of signaling storms, which can be used as a network warning message or adjustment recommendation to inform the corresponding equipment or personnel in advance.

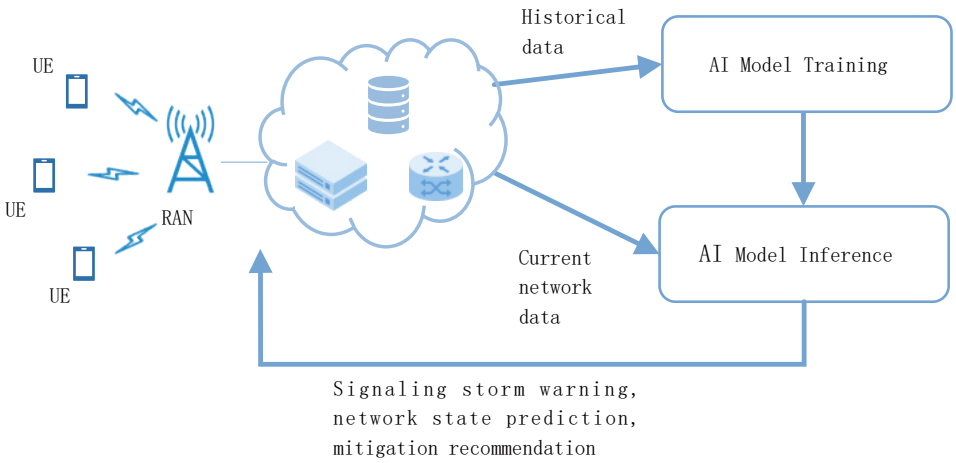


Fig. 2-11. illustration of AI-based signaling storm prediction

Next, we will introduce AI applications to scenarios where closed-form solutions are not easily obtained by traditional methods in the expected time or problems without explicit closed-form solutions. The representative use cases are AI-based wireless resource management, network energy saving, and load balancing.

AI-based wireless resource management

There are many tasks (or use cases) in the access network, which is a continuous process of adjustments based on changes in the wireless environment, load, number of users, etc. For example, this kind of tasks includes user scheduling, access control, resource allocation, etc. The decision given in each round thereof not only affects the system performance in the current round, but also affects the decision in the next round. Reinforcement learning is a technical tool to solve this kind of problems. Reinforcement learning obtains the static and dynamic characteristics of the system and environment through interaction with the system and environment, so as to adaptively realize the optimal strategy. Currently, reinforcement learning is studied in many aspects of wireless resource management, such as spectrum resource allocation, dynamic power allocation, scheduling in sidelink and unmanned aerial vehicle (UAV) communication, access control and slicing resource allocation in large-connectivity communication, etc. Compared with the traditional schemes, the wireless resource management based on reinforcement learning can adapt to the system and environment more flexibly and achieve higher system performance and resource utilization.

Taking user scheduling as an example, vivo hosted the AI-based wireless resource scheduling track for cell free scenarios in the 6GANA 6G Network AI Challenge in 2023. 72 teams competed in two rounds of the preliminary and rematch rounds, designing high-performance AI models that significantly optimized the resource allocation strategy and improved the overall performance gain of scheduling. AI-based wireless resource scheduling takes the current channel information as well as the historical scheduling rate as the input to the AI model, and takes the time/frequency/air domain resource allocation policy at the future moment as the AI output. In Fig. 2-12, we show the scheduling scores of the AI-based scheduling scheme and the traditional greedy scheduling scheme. The scheduling score integrates the overall scheduling rate of the network and the scheduling fairness of individual users. Higher scores indicate higher overall network throughput and better user scheduling fairness. It can be seen that the AI-based scheme can achieve better scheduling than the traditional non-AI scheme.

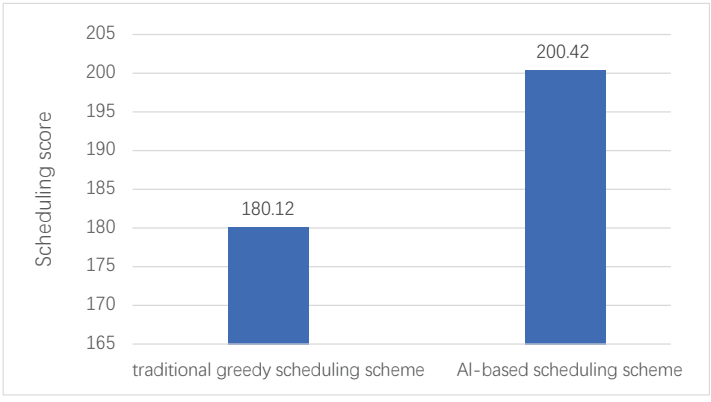


Fig. 2-12. Scheduling scores of the AI-based scheduling scheme and the traditional greedy scheduling scheme

AI-based network energy saving

The growing demand for communication services has led to a sharp increase in the number of network sites and network energy consumption. At the same time, the need for network energy saving by network operators is becoming urgent. Currently, network energy saving is mainly achieved through deactivation/activation at the cell level. To be specific, when the cell service load is lower than a threshold, the cell can be deactivated, and the service can be offloaded to neighbor cells to reduce the overall network energy consumption. When the load of the neighbor cell exceeds a threshold, the deactivated cell can be requested to be activated to avoid network congestion and impact on user experience. However, making decisions of deactivation/activation based on the load of a single cell may result in frequent changes of cell status, leading to massive user service interruption. Traditional methods are limited by computational complexity and it is difficult to directly obtain the optimal activation/deactivation decision for a large area of cells, making it infeasible to achieve optimal network energy efficiency. The general procedure of AI-based network energy saving is shown in Fig. 2-13. With the prediction of UE trajectories and resource status of each cell, the network may determine candidate cells for activation/deactivation. On the other hand, AI can balance network energy consumption and service quality, and maximize network energy efficiency. A similar approach can also be applied to load balancing, having different service types allocated reasonably in various frequency bands to optimize network spectrum efficiency.

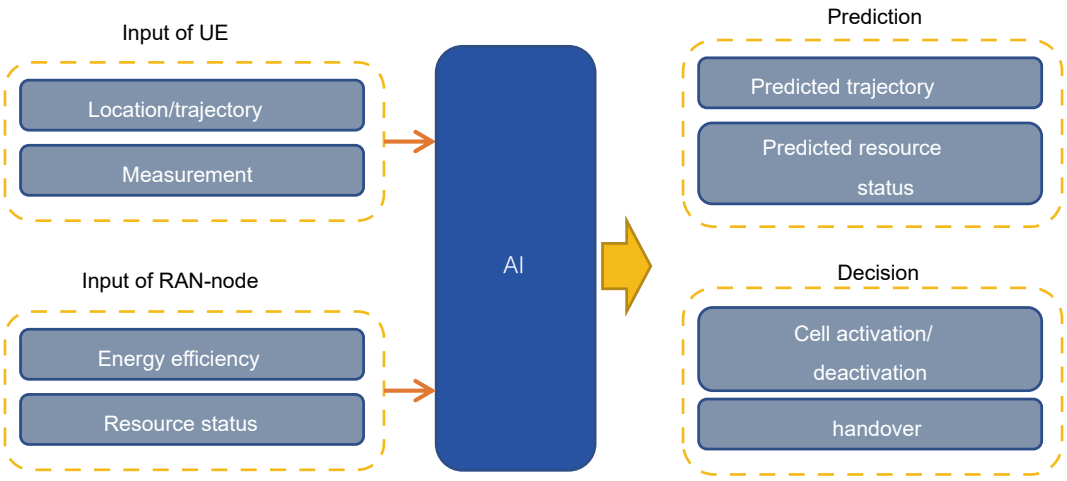


Fig. 2-13. AI-based network energy saving

Finally, we will introduce scenarios where AI is applied to joint optimization of multiple modules. The representative use cases include joint optimization of related functions in MIMO signal processing, joint source-channel coding, cross-layer optimization, etc.

AI-based joint optimization of related functions in MIMO

One advantage of AI is its ability to explore implicit relationships behind data. In communication systems, some functions are related but their specific relationships cannot be explicitly represented, leading to limited effectiveness of joint optimization. To address this, AI can be used to jointly optimize multiple related functions. The system capacity of MIMO is determined by precoding, which is generated based on channel estimation (based on channel reciprocity for time-division duplexing systems, and based on channel feedback for frequency-division duplexing systems), which in turn is related to the design of pilot sequences. Although single-module AI optimization can improve the performance of each module, it cannot achieve global optimality in terms of overall system performance. To address this, we can consider modeling the related functions in MIMO signal processing as a joint problem, establishing a global loss function, and obtaining the optimal MIMO transmission scheme. In Fig. 2-14, taking 2-user MIMO as an example, CSI-RS sequence selection, channel feedback, and precoding matrix generation can be achieved through three sub-neural networks. The parameters of the first sub-network are the reference signal sequences obtained through neural networks, the second sub-neural network generates CSI feedback bits, and the third sub-neural network generates precoding matrices. During training, these three sub-networks are concatenated. The spectral efficiency is used as the global loss function to obtain the parameters of each sub-neural network through gradient descent. The global neural network obtained in this way accommodates the dependencies between multiple sub-networks, achieving higher spectral efficiency than the separately trained scheme.

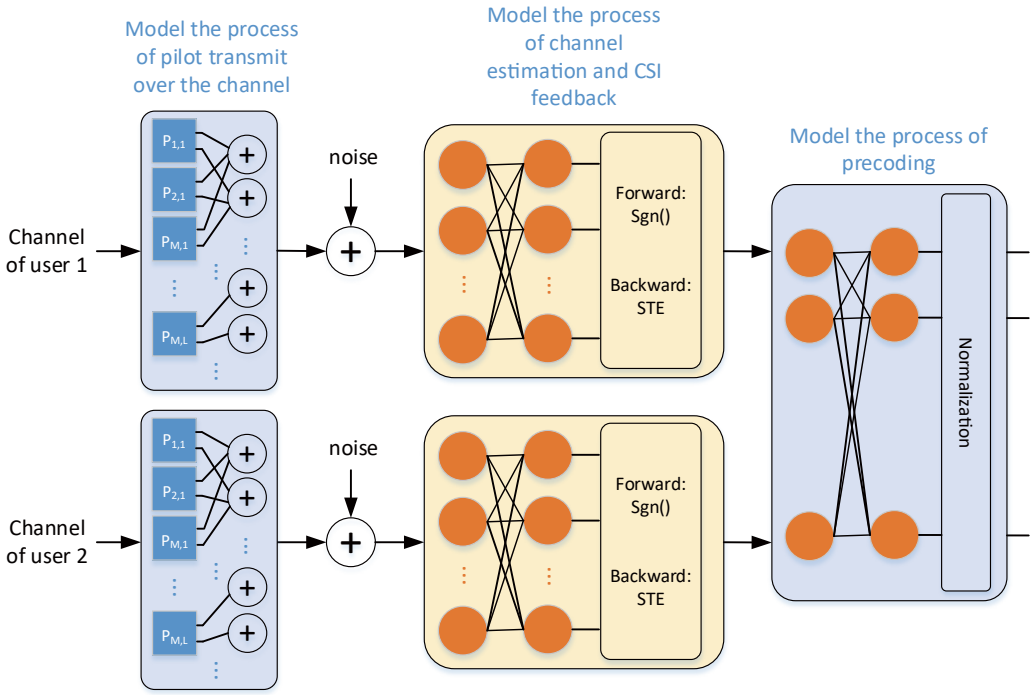


Fig. 2-14. Joint optimization of CSI-RS sequence selection, channel feedback and precoding matrix generation

AI-based joint source-channel coding

Traditional joint source-channel coding allows users to change source coding parameters based on channel or network conditions, or to select channel coding, modulation, and network parameters based on source characteristics. However, its effectiveness highly depends on the expert knowledge. In recent years, the development of AI has provided new ideas for the design of joint source-channel coding. Among them, the structure of auto-encoders (AEs) is very suitable for implementing joint source-channel coding. The encoder takes the source information and the current channel information as input and obtains the bitstream signal to be transmitted through the forward propagation of the neural network. The decoder at the receiver is trained together with the encoder to implement the inverse operation of the encoder. Therefore, the original information can be reconstructed from the received bitstream.

In addition, joint source-channel coding can also be designed based on the idea of semantic communication. First, the multidimensional semantic features of the source are extracted using neural networks. Then different coding strategies are used based on the importance of each dimension of semantics according to specific a priori knowledge. In this way, the characteristics of the known channel can be better utilized. For example, in image and video transmission, using semantic communication based joint source-channel coding may achieve better quality compared to traditional image compression coding. In addition, CSI compression can be viewed as a special joint source-channel coding problem where the efficiency of CSI feedback can be further improved by using the idea of semantic communication. Specifically, the original CSI at the transmitter corresponds to source information, and the CSI information to be fed back corresponds to the bitstream after joint source-channel coding, and the receiver performs inverse processing to reconstruct the original CSI.

AI-based cross-layer optimization

Cross-layer functions can be jointly optimized for better performance. For example, different traffic demands and service characteristics will result in different optimal transmission resources and transmission modes. However, the functions of the different layers are not concatenated together and the objectives of optimization are different for the different layers, so it is difficult to give a closed-form solution directly. To address this problem, resource allocation and traffic/service prediction can be modeled as a joint task. Specifically, using historical traffic, services, resource allocation as inputs, with the objective of optimizing the final performance (e.g., throughput), the AI model can recommend a resource allocation result. Similarly, both the adaptive modulation and coding (AMC) at the physical layer and hybrid automatic repeat request (HARQ) at the data link layer are functions that dynamically adjust the transmission configuration. They can also be viewed as a joint task and a cross-layer joint optimization can be achieved based on AI. The goal of cross-layer optimization is generally the final performance of the network. As a result, there is a need to couple the AI training and reasoning process with realistic systems, which is difficult to be achieved by using supervised learning. Reinforcement learning is a solution to this problem. During the model training, the recommended configuration given by the AI model can be applied to the system, and the response of the system (for example, the final performance of the network or related derivative indicators) can be used as a reward or score to guide the optimization of the AI model. When such an iteratively optimized model is trained, it can be deployed in the actual system to flexibly adapt to the change of environment.

Along with the further research in academia and industry as well as the wider use of AI technologies and resources, there will be more valuable use cases emerging to continuously improve the performance of mobile communication systems. For example, AI has great potentials in the fundamental physical layer design such as waveform design, modulation and demodulation, channel coding, signal detection, signal equalization, as well as in advanced design such as joint design of transceiver, transmission mechanisms based on end-to-end architectures, e.g., transmission without pilot or cyclic prefix.

03

Chapter 3

Design Principles for Convergence of AI and Communication

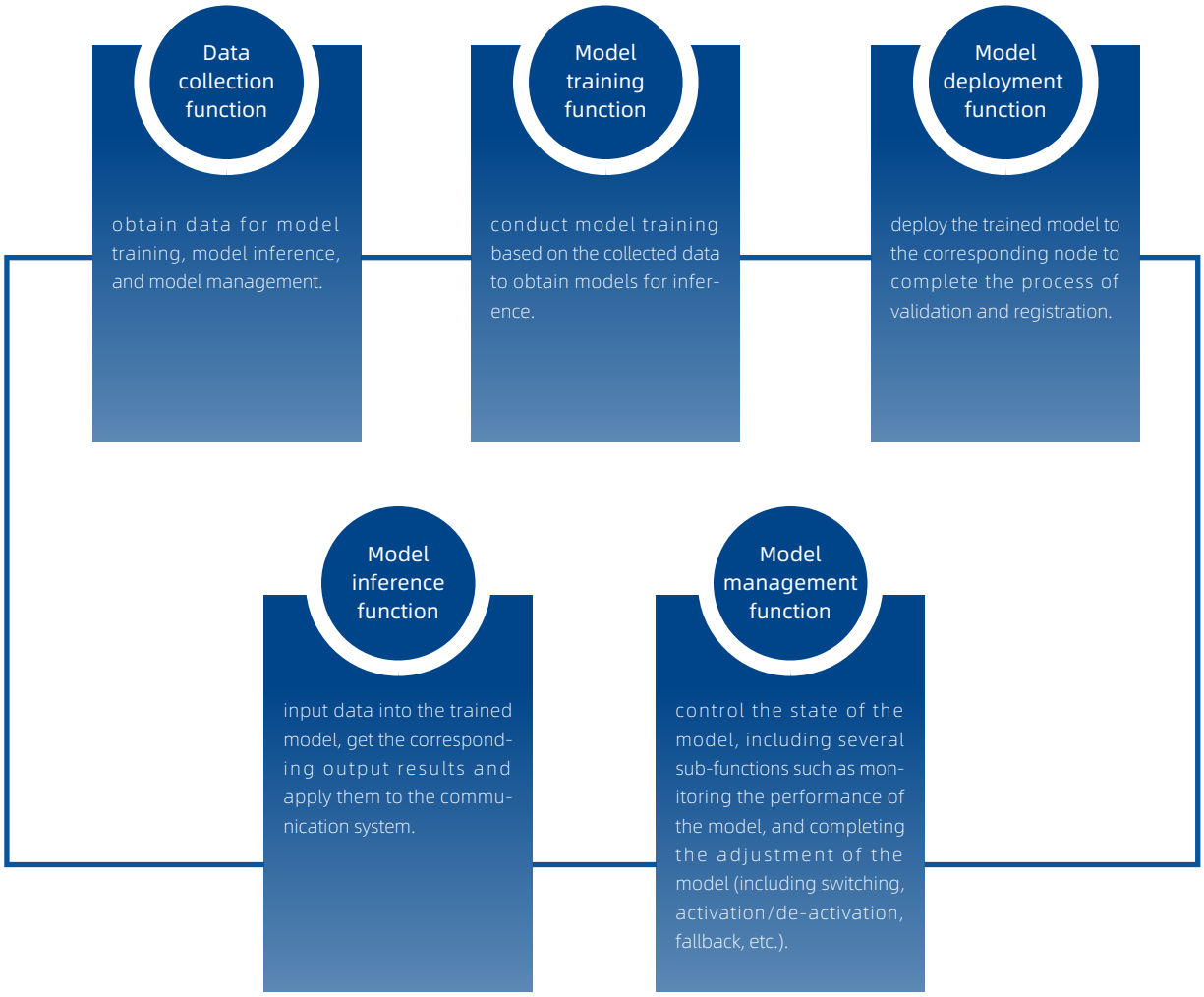
AI will be a native and ubiquitous technology that provides comprehensive support for 6G. In order to leverage AI in 6G, it is necessary to start from several design principles that combine the richness of use cases with the efficiency of the system. Given that 5G has already explored the application of AI in mobile communication, we first summarize the current state of the application of AI in 5G mobile communication. Secondly, we elaborate on the basic logic of native intelligence, and propose design principles from multiple dimensions to achieve deep convergence of AI and communication.

The standardization work of applying AI to mobile communications has been carried out successively in the 3rd generation partnership project (3GPP). 3GPP has launched a number of projects to standardize the application of AI to mobile communication systems, such as AI/ML for OAM, AI/ML for NG-RAN, Enablers for Network Automation for 5G, 5G Systems Support for AI/ML-based Services, and AI/ML for Air Interface. In terms of use case studies, 3GPP has launched AI-based enhancement studies in the core network, radio access network (RAN), and physical layer. Among these use cases introduced in Chapter 2, AI-based CSI feedback enhancement, beam prediction enhancement, and positioning enhancement have been studied in RAN1; AI-based mobility optimization, load balancing, and network energy saving have been studied in RAN3; and AI-based access network selection has been studied in the core network. In 5G, AI performs plug-in optimization for specific problems case by case. This approach is a natural extension of the 5G network and provides some performance improvement for specific problems.

Based on the study of 5G, we start from the basic logic of native intelligence and propose multiple design principles for the convergence of AI and communication to build a solid foundation for 6G.

3.1 Basic Logic of Native Intelligence in 6G

The native intelligence in 6G discussed in this white paper is elaborated from the perspective of AI4NET. The native intelligence in 6G refers to the native integration of AI capabilities into the 6G network. AI was considered in the design of the 6G architecture at the beginning and 6G will reserve the functions, interfaces, capabilities, and signaling structures required for various AI use cases to realize the deep convergence of AI and mobile communication networks. A large number of use cases for empowering 6G with AI are introduced in Chapter 2 of this white paper. With the wider use of AI technologies and further development of AI resources, there will emerge more and more high-value use cases in 6G. If we continue to design protocols for each use case one by one as in 5G, the complexity and redundancy of the protocols will be significantly increased. Therefore, we believe that the architecture design should be based on logical functions and logical nodes to accommodate the AI use cases in 6G. This is because the physical execution location of different use cases is not exactly the same (for example, beam management involves UEs or base stations while positioning enhancement involves UEs, BSs and location management function related entities), designing the architecture based on physical nodes will not be possible to accommodate different use cases. While on the other hand, the logical functions implemented by different use cases can be unified under a same framework, e.g., by referring to the logical functions of data collection, model training, model deployment, model inference, and model management. The specific contents of each logical function are as follows:



Correspondingly, the above logic functions need to be executed in one logic node or collaboratively in multiple logic nodes. The aforementioned logical nodes include data sources, training nodes, inference nodes, execution nodes, and management nodes. Among them, the inference node obtains the output of the AI model, and the execution node applies the output of the AI model to specific functions of the communication system. In most cases, the inference node and the execution node are the same one. Fig. 3-1 shows the correlation between logical functions and logical nodes.

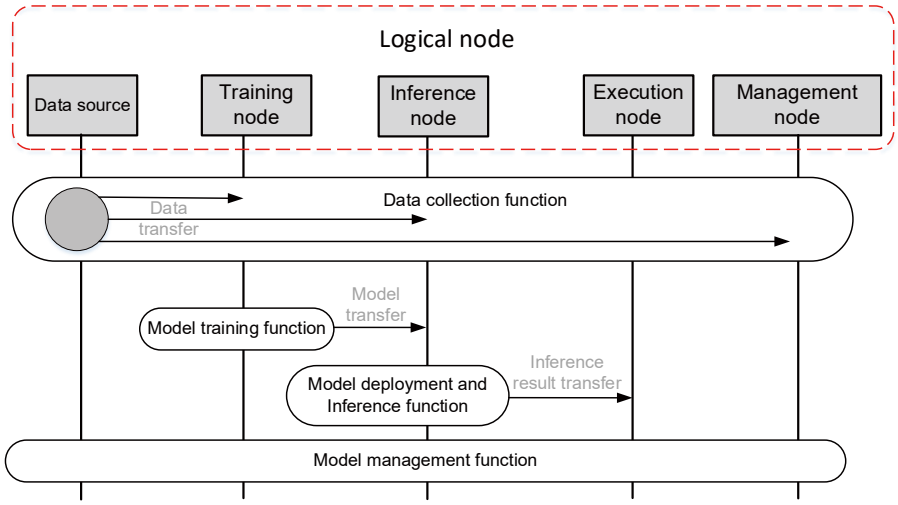
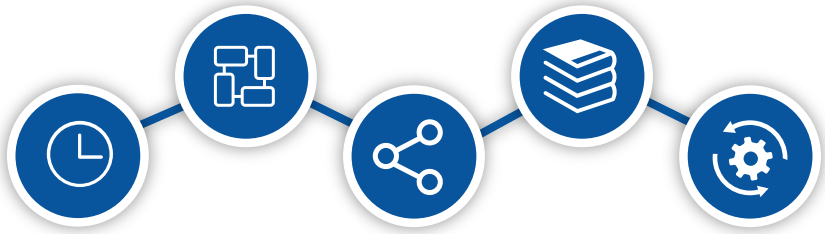


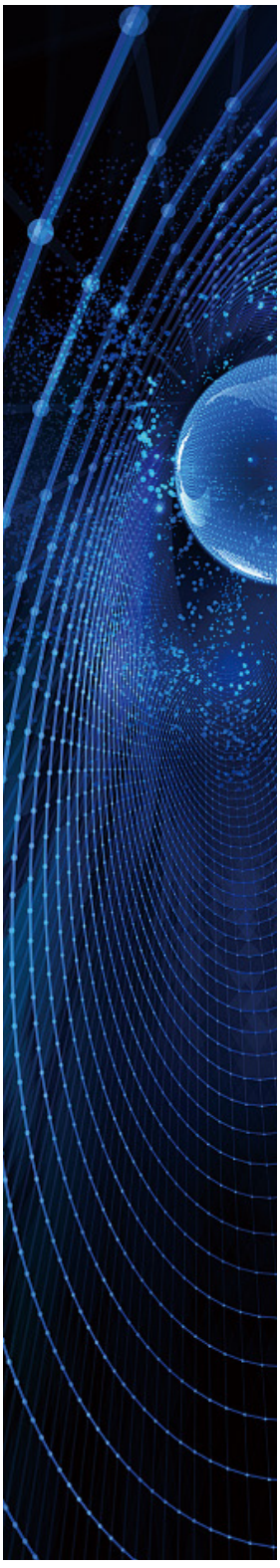
Fig. 3-1. Correlation between logical functions and logical nodes

After designing a protocol framework based on logical functions and logical nodes, it only needs to map logical nodes to actual physical nodes in specific use cases. In practice, multiple logical nodes may correspond to a single physical node. In this case, the interaction process between these logical nodes will be realized within the physical node without signaling support. In the following, we explore the design principles of the convergence of AI and communication in five dimensions:

- Lifecycle management
- Distribution of AI logical function
- Sharing of AI resources within device
- Learning frameworks
- Evolution



3.2 Native and Unified Lifecycle management



In the 6G AI system that converges AI and mobile communication, an AI model needs to go through the process of data collection, model training, model transfer, model validation, model deployment, model inference, model monitoring, and model adjustment, forming a variety of functions in the lifecycle of AI models. Lifecycle management is a necessary and unique operation for AI applied in mobile communication. The main reason is that AI models are trained based on data, and the effectiveness of the model is closely related to the quality of the data, the similarity between the training and application environment. It is inevitable that there is a mismatch between the AI model and the application environment, which causes the insufficient model generalization. Therefore, it is necessary to manage the lifecycle of the model.

The logical functions and logical nodes introduced in the previous section are applicable to all AI use cases in mobile communication systems. Therefore, a unified lifecycle management can be realized based on these logic nodes and logic functions. Furthermore, by splitting the lifecycle management process of AI models, we find that the lifecycle management is a closed-loop process. As shown in Fig. 3-2, this closed-loop process can be further divided into a large and a small loop. The focus of the large loop is to acquire a novel model, which can be realized through data collection-model training-model transfer-model registration-model inference-model monitoring. The focus of the small loop is to perform model adjustment, which is mainly realized through data collection-model adjustment-model transfer-model registration-model inference-model monitoring.

However, data collection and model transfer are not mandatory for the small loop, depending on the specific model adjustment scheme. It can be seen that there is an overlap between the large and small closed loops, i.e., model registration, inference, monitoring. These functions are mainly related to the execution of one model. In addition, data collection in large loop generally involves a large amount of data collection, which is mainly collected offline, while data collection in small loop generally involves a small amount of data collection required for model fine-tuning, which is mainly collected online.

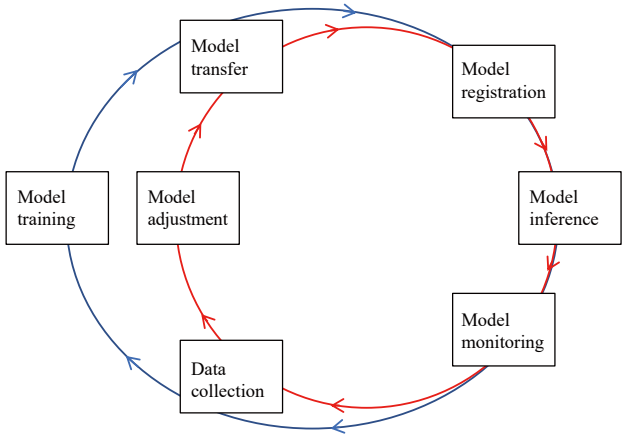


Fig. 3-2. The large and small closed loop in the lifecycle management

In practice, all use cases can operate based on such a lifecycle management that includes both large and small closed loops. The large loop is executed when a new version of the model is required, and the small loop addresses the issues when model adjustments are required as users move around and the environment changes.

A key function in the unified lifecycle management is data collection. How to realize data collection based on a unified scheme is critical. We found that both user-plane and control-plane that have been defined in 5G cannot satisfy the needs of data collection for AI:

1. Currently, the capacity of the control plane is 9000 bytes for a single transmission [8], which cannot meet the transmission demand of big data.
2. The user plane can only transmit data between the UE and the user plane function (UPF), for example, if the UE transmits data through the data plane, the base station cannot obtain the data. In addition, user-plane data collection may need additional user consent since charging policy would impact user's preference.
3. Data collection using control plane and user plane has the problem of duplicate collection of the same or similar data.

To solve the above problems, we propose that 6G AI should unify data collection with the help of data plane, as shown in Fig. 3-3. The data interaction in data plane has the diversity of one-point-to-multipoint, multi-point-to-one-point, multipoint-to-multipoint. The termination location of the data can be in the core network, radio access network, and UE. Therefore, the data plane can efficiently and flexibly support various data exchange. Depending on the quality of service and the hierarchy of data, the data control function in data plane is responsible for the creation, modification, and release of data transfer channels between the data provider and the data consumer. Here, the data provider is the aforementioned data source logical node, data consumer can be the training node, inference node and management node. Besides, the data control function also contains a data collection management subfunction that processes received data requests, including merging identical data request, and generates data collection control information to guide data providers. In this way, the problem of duplicate data collection can be avoided, enabling an efficient and unified data collection.

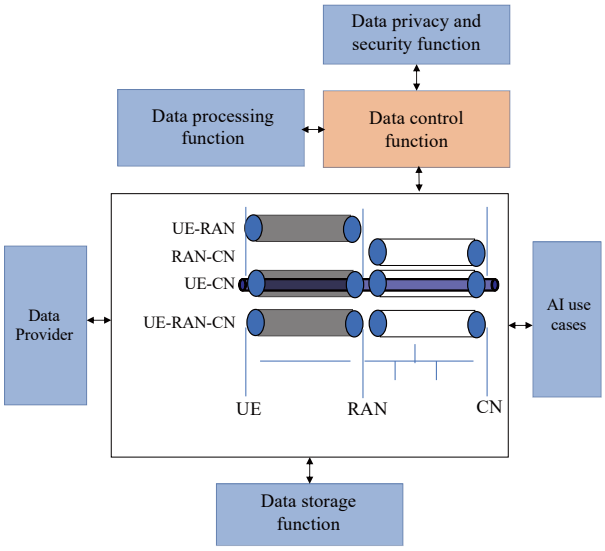


Fig. 3-3. Data collection for AI use cases based on the data plane

3.3 Distribution of AI Logical Function in 6G System

In traditional communication systems, the computing power is mainly deployed at the network side and used for computational processing of communication services. With the rapid development of the AI industry, intelligent processors such as graphics processing units (GPUs), neural network processing units (NPUs), tensor processing units (TPUs), and other intelligent processors have continuously emerged. New types of computing power continue to emerge, and their cost, energy efficiency, and computing level are all being improved day by day. Meanwhile, in the future mobile communication system, almost all network elements need to be enhanced with AI. Therefore, as shown in Fig. 3-4, each network element will have its own use cases and collaborative use cases across network element nodes, and different AI functions will be distributed in each network element node. Specifically, single node use cases can be further divided into UE use cases, base station (BS) use cases, core network (CN) use cases, and network management use cases. Cross-node use cases can be divided into UE-BS use cases, UE-CN use cases, BS-CN use cases, and network management-BS use cases.

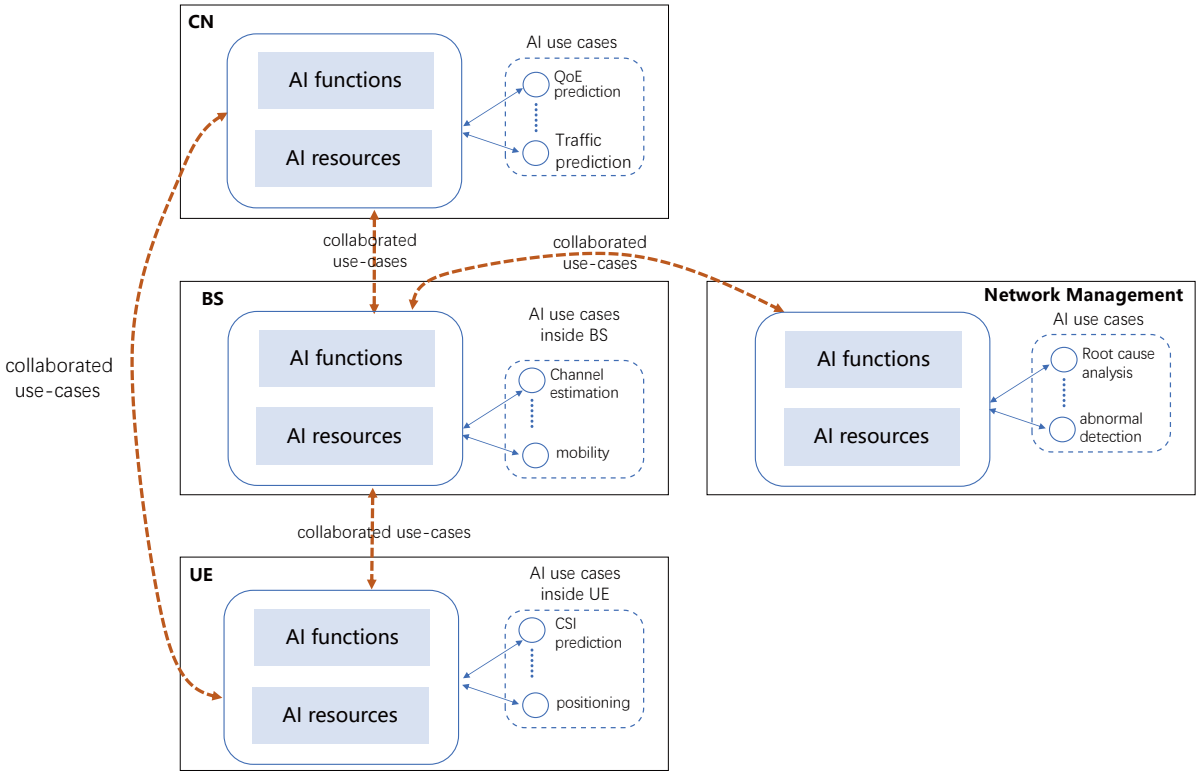


Fig. 3-4. 6G Distribution of AI functions and use cases in the network

When deploying AI functions in physical nodes, different AI functions shall follow different principles.

Distribution of model training functions

Model training is a very critical part of 6G AI. Centralized training typically achieves the performance upper bound, but requires data to be aggregated and then trained in one place, which may raises concerns on data privacy and data security.

Data Privacy

The data of each network element directly or indirectly contains information about the network element and its environment. The data can enhance the value of the communication system with reasonable application, but also at the risk of malicious use. At the same time, data itself may be an asset of the data source node. Therefore, there is a certain need for data privacy/ownership protection requirements for the data of each network element.

Data Security

A centralized training architecture requires centralized data storage, and if the database is breached by a cyberattack, it will lead to massive data leakage.

Therefore, there is a need to keep the data local for training in various cases. But this in turn faces the problems of limited local data features, limited data volume, and limited computing power. Distributed learning is a solution to satisfy the data privacy and security needs while guaranteeing sufficient training data and computing power. Distributed learning is a general term for a class of learning methods whose core idea is that multiple nodes are involved in training, and the data does not need to be aggregated to a centralized node. There are various distributed learning methods such as federated learning, swarm learning, split learning etc. From the characteristics of different distributed learning architectures summarized in Table 3-1, it can be seen that different distributed learning architectures have their own training modes, topologies, exchanged contents and applicable scenarios.

Therefore, it is necessary to design a unified distributed framework to support multiple distributed learning methods, and select the applicable learning method to provide services in specific business.

Table 3-1. Summary of characteristics of multiple distributed learning methods

Distributed learning approach	Characteristics
Federated learning	The most common distributed learning architecture generally consists of a central parameter server and multiple distributed client nodes, where models are trained at the client nodes and aggregated at the server before being distributed
Swarm learning	No central parameter server, no need to upload parameters to a central server for aggregation
Split learning	A central parameter server and distributed client nodes each train a portion of a complete neural network

When building a virtual network for distributed learning in a mobile network, suitable nodes need to be selected to participate in the collaboration. The most important criterion for selecting nodes to participate in collaboration is the quantity and quality of data on the node. In addition, the selection of collaborative nodes also needs to consider factors such as the computing power and transmission capacity of the nodes.

Distribution of model inference functions

Data privacy and security issues also need to be considered in model inference. However, unlike model training, model inference requires a small amount of data, and the risk of data privacy and security is relatively small. Therefore, under the premise of ensuring data privacy and security, computation offloading can be more flexible during the model inference stage.

A key requirement of model inference is inference latency. If the computing power of the data source node is abundant, the inference should be completed at the data source node. However, if the computing power of the data source node is limited and inference is still performed on the data source node, the computation latency will not be able to satisfy requirements. Given the AI capabilities of different nodes and the different needs of AI services, the available AI capabilities within the network are characterized by uneven distribution and dynamic changes. In this case, nodes with abundant AI capabilities (called assist nodes) can be allowed to assist nodes with high demand for AI capabilities (called demand nodes) to accomplish specific AI inference tasks, and then feedback the desired information to the demand nodes. This is a dynamic AI inference function distribution, at this time, the inference delay consists of two parts: the computation delay and the transmission delay, and the key is how to reduce the transmission delay. Therefore, the principle of selecting collaborative nodes is crucial. When selecting assist nodes, it is necessary to consider multiple factors, such as the distance between the node and the demand node, transmission costs, computing power, and other factors.

Distribution of model monitoring functions

Model monitoring requires near real time latency, and failure to perform model state adjustments on time will result in system performance degradation. Model monitoring and decision-making about model state (e.g., model activation/deactivation) are closely related. When deploying model monitoring, it should facilitate early decision-making. Thus, it is best to integrate the monitoring metrics collection/computation function and decision-making function in the same node to avoid delays caused by transmission, negotiation, and other processes. There are also other possibilities by deploying monitoring metrics collection/calculation function based on where the ground truth and model outputs are generated in order to avoid transmission delay caused by the transmission of ground truth information.

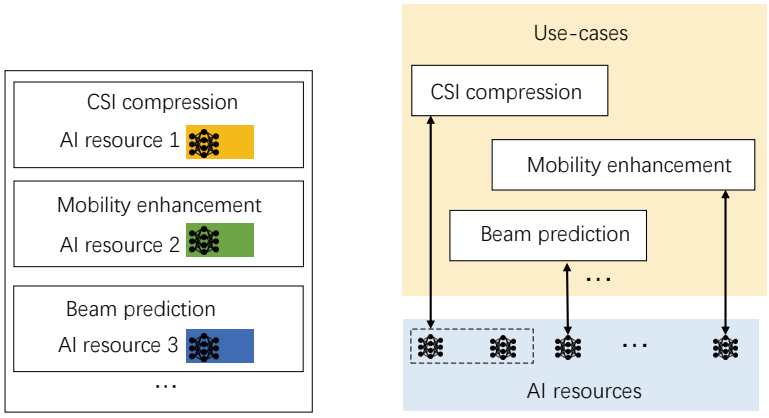
By properly distributing AI functions in 6G networks, higher execution efficiency and resource utilization can be achieved, which helps to achieve efficient mobile communication.

3.4 Decoupling of AI Resources And Use Cases

AI resources include hardware resources and software resources. Hardware resources include GPUs, NPUs, TPUs, application specific integrated circuits (ASICs), etc., and software resources include AI frameworks and algorithms.

There are certain tide effects in the demand for AI resources by the use cases of 6G AI, i.e., the demand for AI resources by different use cases exhibit different patterns in time. In this regard, AI resources and use cases can be decoupled to achieve better resource utilization and performance at a lower cost and expense by time division of the shared AI computation resources across different use cases.

Currently, the design of AI use cases in 5G-Advanced does not take into account above mentioned shared computation resource, which may lead to the following problems: 1) Poor compatibility of new AI use cases: since AI resources are tied to specific AI use cases, new AI use cases cannot use existing AI resources. 2) Low efficiency of AI resource utilization: as shown in Fig. 3-5(a), the communication equipment supports AI-based CSI compression, mobility enhancement, and beam prediction. However, these three use cases do not necessarily need to infer at the same time, if only beam prediction requires model inference at a certain moment, the remaining two AI use cases do not require inference, but the corresponding 2 AI resources cannot be used for beamforming model inference, which makes it difficult to effectively utilize AI resources.



(a) AI use cases tied to specific AI resources

(b) Multiple AI use cases share AI resources

Fig. 3-5. Relationship between AI use cases and AI resources

To solve the above problems, 6G AI resources and AI use cases should be decoupled, and thus enable multiple use cases sharing AI resources within device. When adding new AI use cases, existing AI resources within device can be directly reused without updating AI hardware resources, which improves the compatibility of new AI use cases.

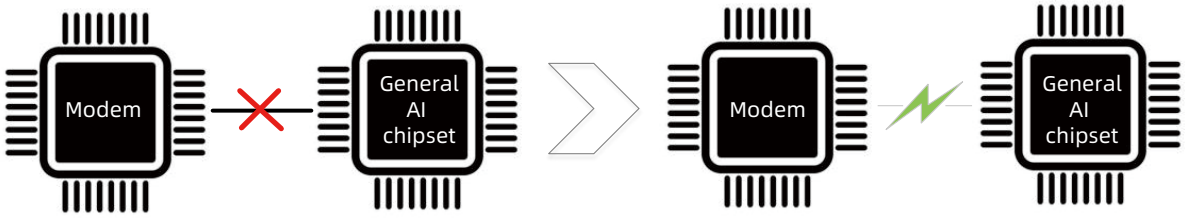
In addition, as shown in Fig. 3-5(b), if only the beam prediction use case requires model inference at a certain time, the entire AI resource can be used to shorten the inference time and improve the utilization efficiency of AI resources. When multiple use cases share AI software resources, it is necessary to consider whether the platforms of multiple use cases are compatible, and whether the algorithm software interfaces of multiple use cases are compatible.

In addition, the decoupling of AI resources and AI use cases is also beneficial to model training. After the decoupling of AI resources and AI use cases, the AI resources will become a common resource. For AI tasks with high generality, multiple model providers (e.g., different operators and UE vendors) can share AI resources, jointly train a model, and then distribute it to their respective users. This can significantly improve the utilization of AI resources and the efficiency of model training.

The impact of the above design principles on the general AI design lies in the general resource interaction signaling. If AI resources are tied to use cases, then the AI resource interaction signaling for different use cases will be different. When AI resources are decoupled from use cases, the AI resource interaction signaling used by multiple use cases can be general. The AI resource general interaction signaling can include node available resource report/notification, AI model/functionality request to accommodate needs of different available AI resources and so on.

It should be noted that the common resource interaction signaling can be used by different use cases between any two nodes, but the interaction signaling is often different between different nodes such as UE-CN and UE-BS.

In addition to the decoupling of AI resources and use cases mentioned above, the sharing between UE-side modem-specific AI resources and UE general AI resources is also worthy of attention. With the popularization of AI, the AI software and hardware capabilities of general AI chipset on UE side will be greatly improved, and may be much higher than the AI software and hardware capabilities of modems. Thus, AI resources to enable 6G use cases should consider both within and outside modem possibilities, as shown in Fig. 3-6. The sharing of these two kinds of resources may lead to different latency for model inference and the applicability to different use cases need to be further considered.



(a) Barriers exist between modem and general AI chipset

(b) AI resource sharing between modem and general AI chipset

Fig. 3-6. Resource relationships between modems and general AI chipset

In the above scenario, the AI resource interaction mechanism can be designed based on the importance of AI use cases, process requirements (or timeline requirements), and quality of service (QoS) requirements, for example, to allocate more AI resources to use cases that have a greater impact on system performance. If two functions have a strict timing relationship, the use case with the first timing should be prioritized for resource sharing. Similarly, use cases with high QoS requirements also need to be given priority in allocating AI resources. For example, when the AI model/function resource request from UE to BS contains multiple use cases, since each use case has different processing delay requirements, the prioritization order and processing delay requirements of each use case, can be notified to BS, so that BS can decide the order in which the use cases are to be executed based on these requirements.

3.5 Diverse Learning Frameworks Supported by Model Transfer

In Chapter 2, we depict the blueprint of AI in mobile communications through a rich set of use cases. It is possible to characterize all AI use cases in terms of their application environments and extract the commonalities of the use cases, which in turn can assist the design of AI-native system. Referring to the division idea in [9], we divide the AI use cases in mobile communications into four quadrants according to the static/dynamic and closed/open character of the environment, as shown in Fig 3-7. "Static" means that the characteristics of the environment are limited and stable, while "dynamic" means that the characteristics of the environment are not fixed and with uncertainty. "Closed" means that the output and subsequent actions of the AI will not affect the environment, while "open" means that the output and subsequent actions of the AI will affect the environment. For ease of understanding, we have given one use case in each quadrant of Fig. 3-7 as an example.

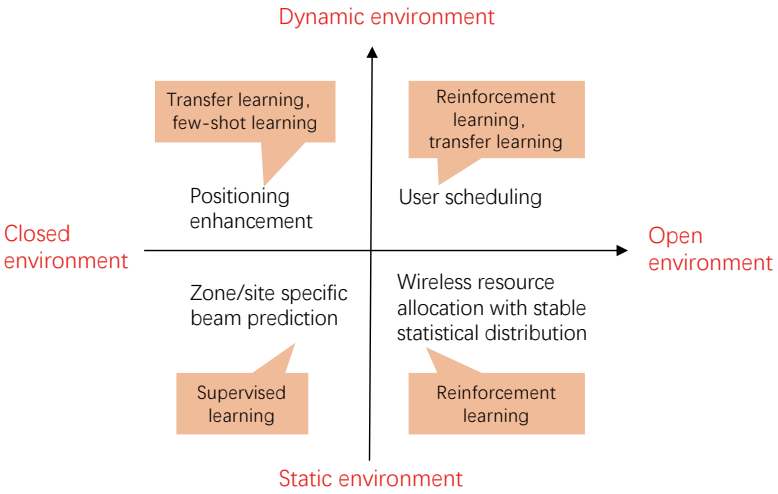
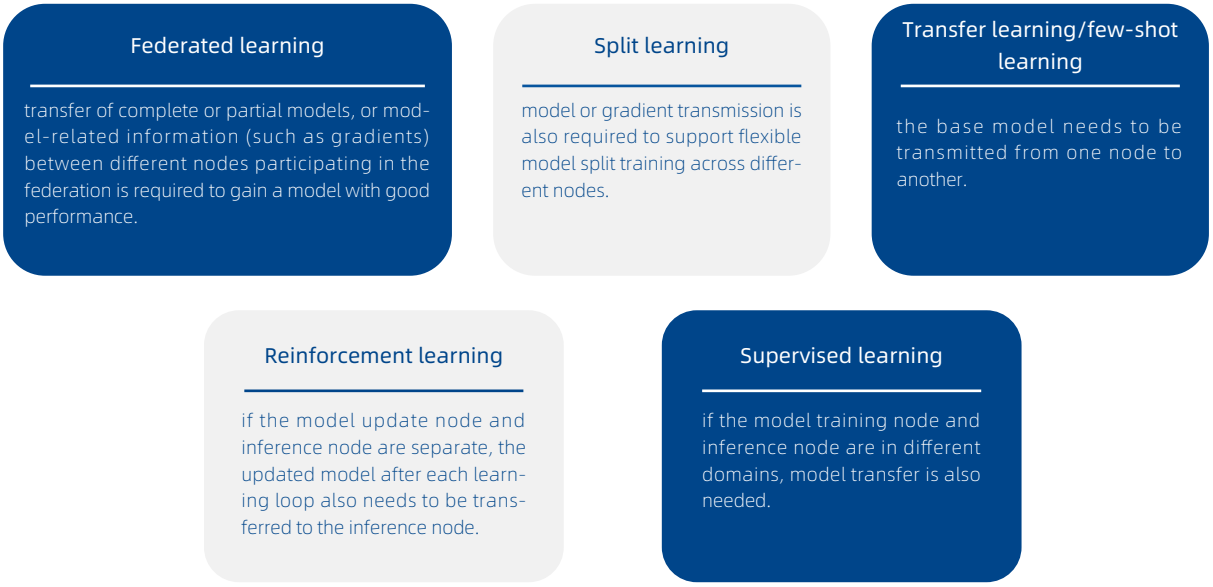


Fig. 3-7. Environment-based use case segmentation

As shown in Fig. 3-7, different types of use cases generally require different training/learning approaches. Use cases in a closed-static environment are the simplest type of use case, where a model with good generalization can be obtained through supervised learning and can be directly deployed for inference. In a closed-dynamic environment, it is difficult to obtain a well-generalized model through offline training, so it is necessary to use techniques such as transfer learning and few-shot learning to adapt the model to the environment. The key in an open-static environment is that there is strong interaction between the use case and the environment or system, and the environment or system will change based on the results given by the model. Reinforcement learning is an important means to solve such problems. The open-dynamic environment is the most complex environment, and requires a combination of techniques such as transfer learning and reinforcement learning to achieve excellent inference performance. In addition, data in the network are often distributed across different nodes, and considering the need for data privacy, distributed learning is also a very valuable learning architecture.

In summary, 6G AI needs to support multiple learning methods to meet a diverse range of AI use cases. Among the various learning frameworks mentioned above, a key function that is required is cross domain model transfer. Specifically, the requirements for model transfer in different learning frameworks are as follows:



Model is a new type of information that emerged after AI was introduced into mobile communication. It does not belong to the traditional service information or control information in mobile communication systems. Therefore, to support the various AI learning frameworks for 6G, it is necessary to define corresponding signaling and procedures clearly from the start of 6G.

One of the key issues for model transfer across domains is what format to use for the transferred model: public format or proprietary format? A public format is a model description format that can be recognized by both ends, and a proprietary format is a private model description format that is specific to a vendor. Table 3-2 compares the detailed characteristics of the two formats. It can be seen that the main advantage of public formats is the high flexibility of model updating and support for zone/site/specific models since cross domain allows flexible model training entities that is most accessible to zone/site specific data sources. To support cross domain model transfer, the industry needs to define such a commonly recognizable format.

Table 3-2 Summary of public and proprietary format characteristics of AI models

Characteristics \ Format	Public format	Proprietary format
Whether model information can be shared across vendors	Yes	No
Dependency on vendors	Low	High
Flexibility in model update entities	Good	Bad*
Support for zone/site specific models	Effective	Limited**
Offline Optimization Requirements	Small	Large
Multi-vendor deployment flexibility	Good	Bad
Model storage overhead	Low	High

*Public format allows any entity with data to update the model while proprietary format only allows the entity to compile the model to update

**In public format, entity most accessible to zone/site specific data is easily involved; in proprietary format, the model compilation entity may not be accessible to zone/site specific data

Based on open format model transfer, multiple learning frameworks in 6G will provide support for diverse use cases, guaranteeing that models are better matched to the complex and changing environment of mobile communication, and achieving better inference performance.

3.6 Continuous Self-Evolution

Wireless environment and system requirements will continue to change over time, so use cases and models need to evolve continuously to adapt to the system. If this evolution still relies on manual parameter tuning and use case selection, the efficiency of the evolution will be greatly reduced. Therefore, how to automate the evolution process of models and use cases is a problem worth further research. In the future, mobile communication systems that integrate AI will continuously and automatically collect data, extract knowledge, and iteratively interact with the environment and users during operation. The system will also automatically update and retire old modules, derive new modules, and gradually build more efficient communication systems. This process is called self-evolution. The foundation for AI self-evolution is built on native unified lifecycle management and unified data collection based on the data plane. AI self-evolution can be categorized into three levels from level 1 (L1) to level 3 (L3), with the self-evolution capability increasing step by step.

L1 self-evolution

AI model parameter self-evolution is shown in Fig. 3-8. At L1, self-evolution is achieved by automatically updating the model parameters within a short period of time to adapt to changes in service requirements and deployment environments. In this stage, self-evolution mainly relies on online learning methods, such as transfer learning, meta-learning, and reinforcement learning. By predefining an online framework for model parameter adjustment, specific models can be automatically trained, validated, and their parameters can be replaced through data collection, online training, model validation, and parameter value adjustments.

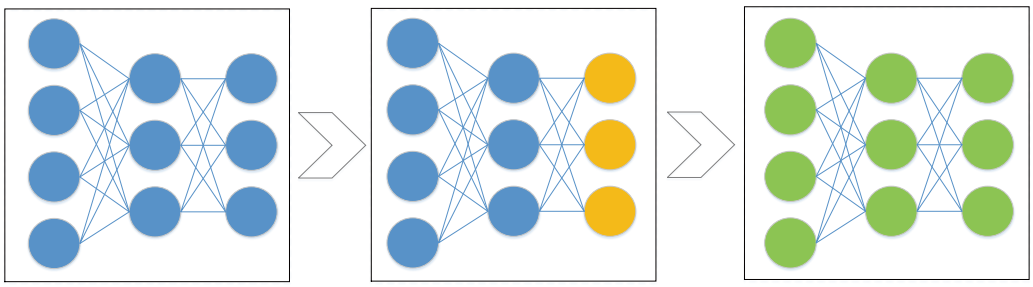


Fig. 3-8. Schematic diagram of L1 self-evolution

L2 self-evolution

The self-evolution of AI model hyperparameters (e.g., inputs, outputs, structures), as shown in Fig. 3-9. L2 self-evolution is an advanced version of L1 self-evolution, which not only solves the problem of parameter adaptation, but also automatically finds out the most suitable model hyperparameters for the specific use case based on data in the actual environment. In this stage, self-evolution mainly relies on technologies such as automatic machine learning (AutoML) and neural architecture search (NAS) to achieve automatic selection of the most suitable model hyperparameters for the specific use case. By predefining an online framework for model architecture adjustments, specific use cases can be automatically trained, validated, and their models can be replaced through data collection, online model search and training, and model validation.

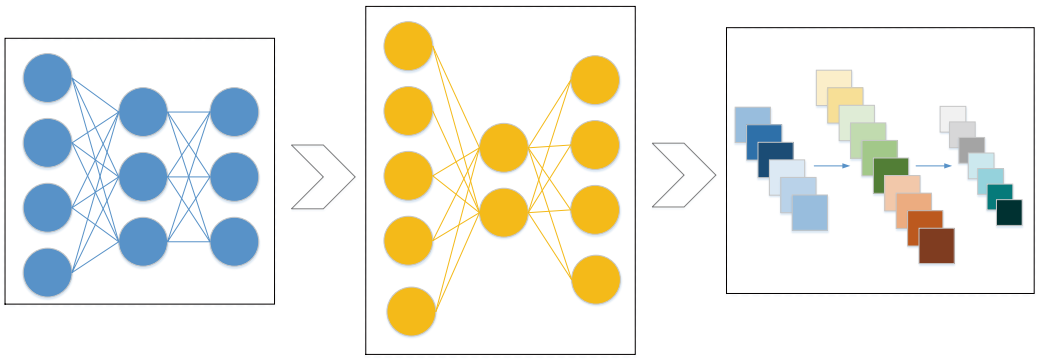


Fig. 3-9. Schematic diagram of L2 self-evolution

L3 self-evolution

Self-evolution of AI use cases, as shown in Fig. 3-10. This level of self-evolution goes beyond the boundaries of specific use cases and allows for the exploration of new use cases and the elimination of old ones. The process of use case change is naturally accompanied by L1 and L2 self-evolution. L3 self-evolution achieves use case self-generation and requires advanced functions such as flexible data collection and use case self-exploration to promote the updating, restructuring, and retirement of old use cases and the birth of new ones.

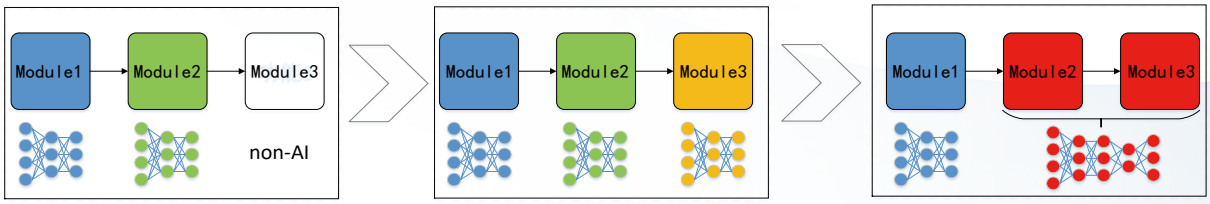


Fig. 3-10. Schematic diagram of L3 self-evolution

From the description of the three self-evolution stages, it can be seen that the L1 self-evolution is the foundation of the whole self-evolution system. In order to achieve the L1 self-evolution, new operations such as automatic evolution triggering, automatic data collection for training, automatic model training, automatic model validation, and automatic replacement of model parameter values need to be introduced. Among them, automatic evolution triggering includes collecting execution results or monitoring results to determine whether to initiate self-evolution. After the self-evolution is triggered, the management function sends data collection, training strategy and validation-related parameter configurations to the training function. Automatic data collection for training refers to the data source collecting data according to the data collection request and providing the data to the training function. Automatic model training refers to the training function conducting model training according to the training strategy configured by the management function, including learning methods, training hyperparameters, optimization function indications, and training-validation dataset partitioning methods. Automatic model validation refers to the training function determining the model performance based on the validation dataset partitioning method and validation KPI configured by the management function. If the validation result meets the pre-set threshold, the training function will send the model parameters and version to the inference function. The inference function will use the new model parameters for parameter value replacement and notify the training function of the successful parameter update. The training function will then provide the management function with the model version and validation performance feedback.

Compared to L1 self-evolution, L2 self-evolution has some new requirements: the inputs and outputs of the model are adjustable in L2 self-evolution, and the data collection configuration needs to be adjustable as well. In L1 online automatic training, hyperparameters and optimization functions are fixed, but in L2 self-evolution, the model structure, training hyperparameters, training algorithms, optimization functions, etc. are adjustable. Therefore, in order to achieve L2 self-evolution, it is necessary to make the automatic data collection for training and automatic model training of L1 self-evolution more flexible, and also need to add new operations such as automatic updates of model structure and parameters. The flexibility of the model training strategy can be achieved through various model attempts or technologies such as AutoML and NAS. Automatic model structure and parameter updates refer to the training function sending updated model structure, parameters, and version to the inference function. After the inference function completes the model update, it will provide feedback to the training function indicating the success of the model update.

To achieve L3 self-evolution, more flexible data collection and model deployment functionalities need to be introduced compared to L2 self-evolution. In L2 self-evolution, data collection configuration corresponds to a specific use case. However, in L3 self-evolution, data collection configuration can be used to collect data for training AI models for current non-AI functions or to collect data for training AI models for the functionality of the two fused AI use cases. Therefore, more flexible data collection configuration is needed. Flexible model deployment means supporting the replacement of a function from non-AI to AI, or AI to non-AI, or multiple independent functions to fused AI functions.

In summary, the realization complexity of self-evolution increases step by step from L1 to L3. We believe that L1 self-evolution is most likely to be realized at 6G. The implementation of L2 self-evolution depends on the maturity of AutoML technology and the growth rate of computing power, so it is also possible to achieve in 6G. The L3 self-evolution requires flexible data collection and interaction interface configuration, and belongs to a longer-term vision.

04

Chapter 4

Conclusions

The convergence of AI and communication involves multiple dimensions such as air interface, network architecture, protocols, algorithms, etc., and will deeply affect the scenarios or functions such as sensing, computing and control. Therefore, the convergence of AI and communication is expected to drive the evolution of future communication paradigm and the change of network architecture, which is of great significance for the research of future mobile communication technology.

The use cases for the convergence of AI and communication are very broad, involving not only multiple layers of the mobile communication network, but also multiple collaboration modes. Its values can be demonstrated in both improving the performance and reducing the complexity of mobile communication systems. With the further research in academia and industry as well as the popularity of AI technologies, there will emerge more and more high-value use cases.

To support such a wide range of use cases, it is necessary to set reasonable design principles based on the common characteristics and requirements of various use cases at the beginning of 6G design. Our goal is to create a generic, flexible and resource efficient 6G AI architecture that supports various new learning architectures and self-evolving. We believe that 6G will be an intelligent system enabled by numerous AI use cases and flexible AI capabilities.

References

[1] vivo, "6G Services, Capabilities and Enabling Technologies," 2022.7

[2] 3GPP, R1-2203550, "Evaluation on AI/ML for CSI feedback enhancement", vivo, RAN1 #109e, 2022.5

[3] 3GPP, R1-2304471, "Evaluation on AI/ML for CSI feedback enhancement", vivo, RAN1 #113, 2023.5

[4] vivo, "Performance evaluation of AI-based DMRS channel estimation in multiple channels", IMT-2030_WX_AI, 202304

[5] 3GPP, R1-2306742, "Evaluation on AIML for beam management", vivo, RAN1 #114, 2023.8

[6] 3GPP, R1-2306744, "Evaluation on AI/ML for positioning accuracy enhancement", vivo, RAN1 #114, 2023.8

[7] vivo, "AI-based suppression for nonlinearity of power amplifier", IMT-2030_WX_AI, 202305

[8] 3GPP, "TS38.323: NR; Packet Data Convergence Protocol (PDCP) specification (v17.5.0)", 2023.6

[9] Jia Liu, Forum on Cognitive Neural Foundations of Artificial Intelligence, 2022 Beijing Zhiyuan Conference, 2022.6

Abbreviations

3GPP	3rd Generation Partnership Project
4G	The fourth generation mobile communication systems
5G	The fifth generation mobile communication systems
6G	The sixth generation mobile communication systems
AI4NET	AI for Network
AMC	Adaptive Modulation and Coding
AMP	Approximate Message Passing
AR	Augmented Reality
ASIC	Application Specific Integrated Circuit
AutoML	Automatic Machine Learning
BLER	Block Error Rate
CN	Core Network
CSI	Channel State Information
CSI-RS	Channel State Information Reference Signal
DL-TDO	Downlink Time Difference of Arrival
DMRS	De-Modulation Reference Signal
DPD	Digital Pre-Distortion
EVM	Error Vector Magnitude
GPU	Graphics Processing Unit
HARQ	Hybrid Automatic Repeat request
LTE	Long Term Evolution
MIMO	Multiple Input Multiple Output
NAS	Neural Architecture Search

NET4AI	Network for AI
NLOS	Non line of Sight
NMSE	Normalized Mean Square Error
NPU	Neural Network Processing Unit
NR	New Radio
OFDM	Orthogonal Frequency Division Multiplexing
PA	Power Amplifier
PAPR	Peak-to-Average Power Ratio
QoS	Quality of Service
RAN	Radio Access Network
RRM	Radio Resource Management
RU	Resource Utilization
TOA	Time of Arrival
TPU	Tensor Processing Unit
TR	Tone Reserve
TRP	Transmission/Reception point
TRS	Tracking Reference Signal
TTT	Time-To-Trigger
UE	User Equipment
UPF	User Plane Function
VR	Virtual Reality