Preface

Artificial intelligence has fluctuated through more than 60 years. With the development of computing power, innovation of algorithms, and accumulation of massive data, artificial intelligence (AI) will be rejuvenated in the next decade, become one of the most influential trends in science and technology.

As the infrastructure of information communication, telecommunication network has great space and potential for applying artificial intelligence technology. How to use the powerful analysis, judgment, and prediction capabilities provided by the AI algorithm to enable Network Elements, networks, and service systems, and combine them with the planning, construction, maintenance, operation, and optimization of telecom networks, become an important topic in the telecom industry.

Based on deep understanding, accumulated experience in telecom field, and long-term investment in the AI Intelligence strategy, Huawei introduces the AI technology to the telecom network based on the full cloudification of the network and launches the SoftCOM AI solution. Based on the application scenarios in the telecom field, the network can be automatically deployed (automatic service deployment and automatic running), self-healing (automatic fault recovery), self-optimization (network self-optimization), and autonomous (network self-evolution), which improves the network efficiency and reduces the OPEX.

This white paper will explain Huawei’s practice in Telecom AI field based on the background of the market trend of telecom networks. It includes the SoftCOM AI strategy interpretation, SoftCOM AI solution (including the communication intelligent platform, AI model service, and deployment solution), and typical application scenarios. It is highly expected to make values to all the colleagues who are engaged in the exploration of telecom AI.

---

**Contents**

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Telecom AI Market Overview</strong></td>
<td></td>
</tr>
<tr>
<td>1. Telecom AI Market Trend</td>
<td>01</td>
</tr>
<tr>
<td>2. Opportunities and Challenges Facing Operators</td>
<td>02</td>
</tr>
<tr>
<td>3. Operators Are Taking Action</td>
<td>04</td>
</tr>
<tr>
<td><strong>Huawei’s SoftCOM AI Strategies Explained</strong></td>
<td></td>
</tr>
<tr>
<td>1. Solving Structural Problems with Architecture Innovation</td>
<td>05</td>
</tr>
<tr>
<td>2. Five Development Phases of Autonomous Networks</td>
<td>06</td>
</tr>
<tr>
<td>3. Introduction of Unique AI Features</td>
<td>07</td>
</tr>
<tr>
<td><strong>Huawei SoftCOM AI Solution</strong></td>
<td></td>
</tr>
<tr>
<td>1. Solution Overview</td>
<td>07</td>
</tr>
<tr>
<td>2. Data Lake</td>
<td>08</td>
</tr>
<tr>
<td>3. Training Platform</td>
<td>10</td>
</tr>
<tr>
<td>4. Inference Framework</td>
<td>11</td>
</tr>
<tr>
<td>5. Telecom AI Model Services</td>
<td>12</td>
</tr>
<tr>
<td>6. Deployment Scheme</td>
<td>14</td>
</tr>
<tr>
<td><strong>Typical Application Scenarios of the Huawei SoftCOM AI Solution</strong></td>
<td></td>
</tr>
<tr>
<td>1. Leap in O&amp;M Efficiency</td>
<td>18</td>
</tr>
<tr>
<td>2. Leap in Energy Efficiency</td>
<td>19</td>
</tr>
<tr>
<td>3. Leap in Resource Efficiency</td>
<td>20</td>
</tr>
<tr>
<td>4. User Experience Improvement</td>
<td>21</td>
</tr>
<tr>
<td>5. 5G Intelligent Networks</td>
<td>22</td>
</tr>
</tbody>
</table>
Telecom AI Market Overview

1) Telecom AI Market Trend

Telecom Industry Is Becoming the Largest AI Market—Network Automation Will Reach a Turning Point in 2021

Tractica/Ovum forecasts that global telecom industry investment in artificial intelligence (AI) software, hardware, and services will reach US$36.7 billion by 2025. The telecom industry is becoming the largest AI market. Telecom AI software revenue will grow from USD915.7 million in 2016 to USD11.3 billion in 2025, at a compound annual growth rate (CAGR) of 48.8%. Network automation will reach a turning point in 2021, as operators increasingly adopt SDN and NFV and deploy 5G networks.

Network Optimization Will Be the Leading AI Application in the Telecom Market

According to the report from Tractica/Ovum, the leading AI application in the telecom industry will be network or IT operations monitoring and management, accounting for nearly 61% of total telecom AI expenditure between 2016 and 2025. Other key AI applications will include virtual digital assistants (VDAs) for customer service and marketing, intelligent customer relationship management (CRM) systems, customer experience management, and cyber security.

To put it simply, network or IT operations monitoring and management refer to network optimization. Operators use AI technologies to understand network events in real time and dynamically adjust their networks for service delivery. AI-driven network management solutions involve the following: network design, network load balancing, and network coverage and capacity optimization.

Opportunities and Challenges Facing Operators

An era of cross-industry competition is approaching. Each industry is facing structural challenges, especially the telecom industry.

First, from the perspective of revenue structure, operators’ services are facing challenges from the IT industry. Unlike telecom players in the past, modern players are now provided as cloud services. If operators can do well in cloud services, they can compete with cloud service giants, such as AWS, to seize trillion-dollar market shares. Otherwise, operators may also lose many telecom networks.

Second, operators’ efficiency and costs are also facing structural challenges. Nowadays, the structure of telecom networks is complex. As a result, the operating expenditure (OPEX) due to device maintenance is exponentially growing. OPEX refers to the network maintenance costs against network capacity and utilization. The higher the capacity utilization is, the lower OPEX will be. Therefore, operators are facing the following problems.

OPEX ≈3 times of CAPEX

2) Opportunities and Challenges Facing Operators

According to the report from Tractica/Ovum, the leading AI application in the telecom industry will be network or IT operations monitoring and management, accounting for nearly 61% of total telecom AI expenditure between 2016 and 2025. Other key AI applications will include virtual digital assistants (VDAs) for customer service and marketing, intelligent customer relationship management (CRM) systems, customer experience management, and cyber security.

To put it simply, network or IT operations monitoring and management refer to network optimization. Operators use AI technologies to understand network events in real time and dynamically adjust their networks for service delivery. AI-driven network management solutions involve the following: network design, network load balancing, and network coverage and capacity optimization.

Network Optimization Will Be the Leading AI Application in the Telecom Market

According to the report from Tractica/Ovum, the leading AI application in the telecom industry will be network or IT operations monitoring and management, accounting for nearly 61% of total telecom AI expenditure between 2016 and 2025. Other key AI applications will include virtual digital assistants (VDAs) for customer service and marketing, intelligent customer relationship management (CRM) systems, customer experience management, and cyber security.

To put it simply, network or IT operations monitoring and management refer to network optimization. Operators use AI technologies to understand network events in real time and dynamically adjust their networks for service delivery. AI-driven network management solutions involve the following: network design, network load balancing, and network coverage and capacity optimization.
Low resource utilization: Wireless resources, IP addresses, and optical transmission resources are not fully utilized. For example, the utilization of IP backbone networks between hotspot DCs is close to 70%, but the network utilization between non-hotspot DCs is only 30%. Supply and demand are not balanced, and elastic traffic adjustment is not provided. As a result, network utilization is low. The traditional Massive MIMO beam and tilt parameters of wireless base stations are roughly estimated based on user distribution and cannot be accurately adjusted based on traffic distribution, interference, or cell load. The air interface resource utilization is low.

High energy consumption: The energy consumption of base stations does not decrease with the decrease of network traffic. Energy waste is severe. The data center power usage effectiveness (PUE) remains high, and the ratio of cooling energy consumption to non-IT system power consumption is high.

Low O&M efficiency: According to Gartner’s report, 37% of network faults are caused by network changes. Network structures are becoming increasingly complicated, creating network operations and maintenance (O&M) challenges that too complex to address manually. 75% of network problems are found only when customers submit complaints to operators. Customer experience and satisfaction are difficult to guarantee. O&M personnel spend 90% of their time on fault location.

With the introduction of the network functions virtualization (NFV) and software-defined networking (SDN) technologies to telecom networks, network cloudification and layer decoupling greatly improve operations efficiency. However, cross-layer fault demarcation and analysis require the participation of multiple teams. The analysis efficiency is low and the maintenance and emergency recovery requirements of a large number of sites cannot be met.

Telecom AI Market Overview

3) Operators Are Taking Action

Top Three Operators in China

Telefonica has launched a new AI project and enabled AI analysis technologies in three service operations centers (in Argentina, Chile, and Germany) for mobile network usage analysis and problematic region prediction. In addition, Telefonica use AI technologies to obtain user experience data in real time, helping discover new methods for improving user experience.

SoftBank has acquired a number of AI companies and collaborated with IBM to introduce the Watson AI system into its own network. NTT has also launched an AI platform, focusing on four AI technologies: Agent-AI, Heart-Touching-AI, Ambient-AI, and Network-AI.

In recent years, the use of AI technologies to explore new business opportunities and improve network performance and customer experience has become the top concern of operators. Although they face many challenges, not least because they are entering the AI market later than incumbent Internet giants, telecom operators cannot afford to neglect the AI domain.
Huawei's SoftCOM AI Strategies Explained

1) Solving Structural Problems with Architecture Innovation

Let us review the development path of Huawei network architecture. In 2006, Huawei proposed its Single strategy for the All-IP era. In 2012, Huawei introduced the concept and technology of cloud computing based on All-IP network architecture and entered the All-Cloud era. With the idea of network architecture reconstruction based on SDN, NFV, and cloud technologies, Huawei proposed its SoftCOM strategy to construct DC-centric cloud-based telecom networks. Cloud-based network architecture reconstruction faces three challenges: network and service complexity brought by dynamic network architecture changes, O&M (such as fault location) complexity brought by layered decoupling of NEs, and network operation complexity brought by real-time network resource adjustment. AI is far more capable of dealing with complex problems than are human brains, and is expected to address these challenges after networks are cloudified. With the implementation of Huawei's All-Intelligence strategy, AI is introduced into cloud-based telecom networks. SoftCOM AI was born to enable autonomous networks at the network architecture level.

AI-based telecom networks are introduced to realize “Industry 4.0” in the network domain and make networks as autonomous as possible. Industry 4.0 has three characteristics that support product automation: agile equipment, intelligent control, and intelligent analysis systems. These three characteristics are very desirable for the telecom industry. On a telecom network, the lower layer is network equipment, and the upper layer is a control layer. In terms of control and O&M of the entire network, AI technology is introduced in an E2E manner to implement the segmented autonomy function. Each segment of autonomy provides E2E autonomous capabilities based on an upper-layer operation system. In this way, the entire network can be autonomous. The biggest change brought by an autonomous network is that O&M personnel are not in the service process. The autonomous network is an automated system. This is what we mean by an autonomous network.

Huawei launches the SoftCOM AI architecture. Based on the original SoftCOM full-cloud network architecture, Huawei SoftCOM AI introduces AI technologies with the core of machine learning. Closed-loop control is implemented, as described as follows:

1. An independent AI training platform is established for data training and service model generation.
2. Intelligent control centers are deployed on the wireless network, fixed network, and core network, and in DCs.
3. Service models generated by the training platform are imported to the control centers.
4. Service models are invoked by networks to control network behavior.
5. Analysis results of network behavior are collected for further model training and optimization on the training platform. The entire architecture enables networks to progress from automated service deployment and action execution to intelligent self-healing of faults and network self-optimization, and ultimately comprehensive network autonomy.

2) Five Development Phases of Autonomous Networks

The development of autonomous networks is a long-term process and cannot be accomplished at one stroke. Self-driving cars are rated according to five levels. Using a similar scheme, Huawei proposes five levels for autonomous networks, taking into account aspects of customer experience, network environment complexity, and automation.

L0: manual O&M with auxiliary monitoring capabilities. All dynamic O&M tasks are performed manually.
L1: assisted O&M. The system repeatedly executes a subtask based on known rules (for example, GUI-based configuration wizard or script-based batch configuration tools). This simplifies operations, lowers skill requirements, and improves efficiency of recurring tasks.
L2: partial autonomous network. The system continuously executes a specific control task in a unit based on a certain model. For example, in the cloud computing scenario, the data center network (DCN) provides APIs for the system, so that the system can perform automatic network configuration operations according to the scheduling requirements of cloud platforms such as the OpenStack. The entire process requires no manual operations.
L3: conditional autonomous network. L3 differs from L2 in that the system can observe and analyze a specific environment that changes dynamically and perform automatic control operations to maintain a desired state. L3 indicates that the system can continuously execute a control task for a given objective within a single domain. For example, within a single domain, the system can use AI technologies to complete the operations of alarm aggregation and fault scenario identification. Fault location modules are triggered to quickly find troubleshooting measures and automatically dispatch trouble tickets.
L4: highly autonomous network. Compared with L3, the major change of L4 is that the system can implement autonomous networks based on customer experience in more complicated cross-domain service scenarios. For example, for the home broadband service, the system perceives and analyzes customer experience in real time, continuously identifies network anomalies that change dynamically, and proactively optimizes customer experience, rectifies faults, or automatically dispatches trouble tickets based on customer experience-related and network anomalies. In this way, the system can implement predictive O&M and resolve problems before customer complaints are lodged, minimizing service interruption and significantly improving customer satisfaction.
L5: fully autonomous network. A fully autonomous network that works under all network conditions and across all service domains is almost impossible to construct. However, it is worth exploring and innovating in cloud DCs or on fully virtualized networks with simple services and high network standardization.
3) Introduction of Unique AI Features

The new value of introducing AI into telecom networks is predictability. The management and control center of a telecom network uses certain policies and rules to manage and schedule the entire network based on the southbound interface and data collection of equipment. These rules and policies are based on three factors: network reachability, service level agreement (SLA), and resource utilization efficiency. These three factors are the basis of network automation. However, as networks become increasingly complex, other factors must also be taken into account. Algorithm-based network management and online AI inference and data analysis need to be introduced to provide traffic prediction, quality prediction, and fault prediction. Prediction is the core value of AI. Network resources can be scheduled based on unknown conditions to locate faults before faults occur, optimize quality before quality is deteriorated, and adjust traffic before networks are congested. This helps provide steadfast autonomous networks that have features of automation, self-optimization, self-healing, and autonomy, structurally improving the O&M efficiency.

Huawei SoftCOM AI Solution

1) Solution Overview

The SoftCOM AI solution consists of four parts: data lake, training platform, AI Marketplace, and inference framework. The inference module in the inference framework collects training data from the device control platform, and sends the data to the data lake for data preprocessing, or directly sends the data to the training platform for data training (raw data can be used on the training platform without being preprocessed in the data lake). After the training platform completes the training, the trained model is released to the AI Marketplace. Then the AI Marketplace pushes the model to the inference module.

The inference platform invokes a service model to perform inference and analysis based on real-time network data, and the inference result is executed to control network behavior. The analysis result of network behavior is collected for further model training and optimization. In this way, the closed-loop control is implemented.

The following sections describe the data lake, training platform, and inference framework in detail.

2) Data Lake

Pain Points: Data acquisition is difficult and data quality is poor in the telecom industry.

Difficulty of data acquisition

Although telecom networks have a large amount of data, very little of it can be effectively used. This is due to the fact that data semantics, data formats, and data storage, management, and application mechanisms of different organizations and departments are inconsistent, and data usage is strictly supervised and restricted. For example, for abnormal KPI detection, the number of abnormal samples on the network is small and the samples are difficult to obtain.

Difficulty of data governance

Because the telecom domain is highly specialized, it is difficult to understand data and create models.

Poor data quality

Data may be missing, abnormal, or repeated. Poor-quality data cannot be used.

High data security risks

Data platforms and data itself have security vulnerabilities. If a security event occurs, its impact is great.

Main Functions and Highlights

The data lake of the SoftCOM AI solution provides subject and training dataset services. The data lake can precisely meet various model training requirements and provide data governance services based on experts’ experience and live network feedback. Data asset management is supported, such as data catalog, metadata, data security, and user permission management.
• An abundance of data and 90% decrease in training set acquisition time: The data lake focuses on the 4 major AI trends that operators pay attention to, and accumulates tens of millions of labeled sample data points gleaned from telecom networks.

Comprehensive network data
More than 1,000 AI training sets and more than 30,000 network feature attributes are included. Anonymized data of major NEs in all domains is obtained, including data on the access network, bearer network, and core network, and in DCs. Data is obtained through formal channels, for example, data on operators’ networks (authorization is obtained) and in Huawei’s historical fault library, and data generated in labs.

Abundant labeled sample data
A professional team labels data based on time, fault root causes, and network status to form high-value sample sets. Currently, there are about 100 million labeled samples.

• Efficient telecom data governance and over 5-fold data processing efficiency improvement: Efficient data governance and application development are implemented based on the one-stop extract, transform, and load (ETL) platform and visualization of data and networks.

Easy-to-understand data attributes
The data lake integrates Huawei data dictionaries in all service domains, lowering the knowledge threshold for users to process Huawei telecom equipment data.

Easy-to-understand data relationships
Data (such as user experience, network, status, and topology data) is associated to establish a data relationship map based on the mapping between the data and telecom networks (subject domain and digital network map). In this way, data and data relationship are visualized, so that users can easily understand and use the data relationship and network services.

High efficiency of data governance tools
The data lake supports the import of a number of common file formats (including .txt, .csv, .xml, .gzip, and .json files) and automatic ELF processing capability. The data lake constitutes a single hub where data cleansing, conversion, and governance can all be implemented.

• Good telecom data quality: The layer- and domain-based telecom data quality management capability is built, providing data assets with unified quality standards.

Systematic data quality standards
Data quality metrics, governance rules, and audit rules, and data model quality constraints are defined for different data sources and subject domains based on experts’ experience and industry standards.

Telecom data quality monitoring and evaluation tools
The tool-based data quality management capability is provided to ensure high data quality.

Specialized data governance team
The team is responsible for standardized operations such as data cleansing, conversion, denoising to ensure high data quality. Users only need to focus on application development.

• Multi-tenancy isolation, implementing E2E user data security: Data can be managed, audited, and traced throughout the data lifecycle, and fine-grained permission control is implemented. In this way, data security governance and compliance with data use regulations can be guaranteed.

Full-lifecycle monitoring of data
Systematic security logs are provided throughout the data lifecycle, so that data can be managed, audited, and traced.

Secure data storage
Data is stored based on categories and levels. Measures such as data encryption and bucket isolation are used to prevent data from being used by unauthorized users.

Secure data usage
Fine-grained permission control of data and table-level permission control of databases are supported. A certain resource or user is visualized, and can be searched for and assigned permissions.

3) Training Platform

Pain Points: Technical requirements are high, efficiency is low, and effects are uncontrollable in terms of AI application development in the telecom industry.

Insufficient service knowledge: Algorithm scientists need to spend a lot of time understanding service scenarios. In the telecom domain, AI technology accumulation is insufficient and operators have little relevant experience to draw on.

Low algorithm development efficiency: There is a wide range of AI algorithms to choose from and the trial-and-error costs are high. Online development tools have no intelligent prompt and breakpoint debugging capabilities. The code development efficiency is low.

Long model training period: Model training depends on a large number of expensive computing resources. The optimization period of hyperparameters is long, and the time required for a single training is long.

Poor model replicability: Models heavily rely on data. Therefore, it is difficult to replicate the delivery mode among different sites. The actual application effect is uncontrollable.

Main Functions and Highlights

The training platform of the SoftCOM AI solution supports efficient one-stop model training and quick model verification. It integrates processing capability for telecom domain features (meaning that key features can be quickly identified), and has built-in AI algorithms for exception detection, root cause analysis, optimization control, and service prediction.

Training service composition & capability

- Providing telecom SaaS services for application developers
- KPI anomaly detection service
- Intelligent DC PUE optimization and control service
- Alarm correlation and root cause location service
- Base station traffic prediction service
- Reusable assets for AI model development in the telecom domain
- Training platform workflow development & IDE (cloud & offline)
- Federated learning/Transfer learning/Knowledge graph-based model development wizard in the telecom domain
- Data preparation (7) Feature engineering (10+) Model training (15) Model management
- Telecommunication data preprocessing template
- Algorithm library of the telecom industry
- Model training project as a service
- Basic AI-driven services
- Cloud BU ModelArts
- TensorFlow
- MXNet
- ExeML
- Caffe2
- DNTK
- Keras
- Embedded telecom experience: The training platform is preconfigured with four classes of more than 30 pre-integrated telecom model services. No coding is required. Model services are provided, including traffic prediction, KPI anomaly detection, intelligent control, and alarm correlation. Experience in the telecom domain is standardized as services, and users can obtain models by simply entering data. The wizard-based model development process provides telecom templates for data preparation, feature extraction, and model training, improving training efficiency. The platform integrates the telecom knowledge graph, with more than 50 built-in analysis tools. Telecom experience is converted into tools to facilitate decision making by experts.
**Efficient tools:** Federated learning, distributed model training, and joint training are supported, to facilitate the creation of models that can cope with the small data volume and data loss scenarios that occur in the telecom domain, and to meet the data security requirements of the industry. Transfer learning is supported for quick model adaptation, that is, model training on a non-first site can be implemented with a small amount of data. Enhanced data processing capabilities such as parameter editing, sample evaluation, intelligent repair, and script repair are provided, and data distribution is visualized. An E2E development and deployment environment is provided, so that data preparation, feature extraction, model training, model rollout, model sale, and model deployment can all be performed on one platform.

**Openness and collaboration:** An offline integrated development environment (IDE) is provided for collaborative development with the cloud platform. Members of the same development team can collaborate. The platform supports multiple machine algorithm frameworks and algorithm transplantation. All mainstream algorithm frameworks are supported, such as TensorFlow, MXNet, Caffe2, and Spark ML. Tools are provided to quickly transplant algorithms developed on other platforms to Huawei’s platform.

---

### 4 ) Inference Framework

**Pain Points:** The development period of inference applications is long and model effects are difficult to evaluate.

Typical AI inference applications include data collection, data preprocessing, model execution, inference result delivery, and inference result feedback. Development personnel need to develop each component or service. Development is time consuming and component interconnection is complex. After the application is brought online, the service O&M pressure is high, the actual running effect of the model is difficult to predict, and application value cannot be manifested.

**Main Functions and Highlights:**

**Diagram:**

1. **Application management:** The browsing, subscription, automatic update, installation, deployment, and gray release of application models are supported. (Gray release refers to a launch mode that allows for smooth transition from a gray version to an official version.) A/B testing is a type of gray release. Some users continue to use the original version (A), while some other users start to use the new version (B). If users of version B approve of this new version, all users will be gradually migrated to version B. Gray release enables fault discovery and rectification in the early stage, minimizing the negative impact of upgrade on the system and ensuring the system stability. Convergent orchestration and scheduling of multiple models are supported.

2. **Data collection:** Raw data is imported by the inference framework. The inference framework periodically and automatically collects statistics on model performance and generates an inference evaluation dataset for further model optimization.

3. **Inference platform:** The TensorFlow library for deep learning, MLlib library for machine learning, and Python/Java AI model running framework are supported. The inference framework supports the inference execution, scheduling, and monitoring of multiple models, and provides APIs for external systems.

4. **Visualization:** Inference results can be automatically evaluated or manually evaluated by users. The inference framework supports inference visualization, historical inference result query, and statistics report generation.

5. **Offline data:** The data collection, data preprocessing, and inference results of the inference framework are visualized. An E2E development and deployment environment is provided, so that data preparation, feature extraction, model training, model rollout, model sale, and model deployment can all be performed on one platform.

**Data collection and processing:** Cross-network data transmission is supported. Raw data such as NE topologies, alarms, and KPIs can be collected from various network management systems and service systems. The inference framework imports raw data through a real-time flow or in batches, and then cleans and converts the raw data.

**Inference execution and monitoring:** The TensorFlow library for deep learning, MLlib library for machine learning, and Python/Java AI model running framework are supported. The inference framework supports the inference execution, scheduling, and monitoring of multiple models, and provides APIs for external systems.

**Inference evaluation:** Inference results can be automatically evaluated or manually evaluated by users. The inference framework supports inference visualization, historical inference result query, and statistics report generation.

**Highlights**

The inference framework has the following key highlights:

- **Agile rollout of application models:** The inference framework supports one-click subscription, automatic deployment, real-time monitoring, automatic scaling, flow control, fault self-healing, and simplified O&M management of AI cases. The inference framework is easy to upgrade and is decoupled from service systems. Independent upgrade and gray release of applications and models are supported. Various inference modes are supported, including one-off, batch, real-time, asynchronous, streaming, and scheduled inference.

- **Online model optimization:** Online incremental learning and self-optimization capabilities are provided. Online data can be used for incremental learning to obtain optimal values of measurement indicators such as model precision. The inference framework is integrated with the best practices and professional knowledge in the telecom domain and is combined with AI reinforcement learning to provide O&M support and experience in the telecom domain.

- **Intelligent feedback evaluation:** Model inference results are displayed in real time, so that users can quickly complete inference result feedback and high-value data can be generated for users. Based on the inference feedback, the inference framework periodically and automatically collects statistics on model performance and generates an inference evaluation dataset for further model optimization.

---

### 5 ) Telecom AI Model Services

**Model Generation Services:**

The following professional and efficient AI model generation services with applications in the telecom domain are provided based on the Huawei SoftCOM AI training platform:

- **Traffic prediction model generation service for wireless cells**

  The SoftCOM AI platform integrates feature extraction, data modeling, and data algorithms of the wireless cell prediction model, allowing developers to quickly generate a prediction model by entering cell traffic statistics and engineering parameters. This model can predict the traffic, load, and number of users in a wireless cell and is used in intelligent base station control.

**Alarm correlation model generation service**

AI algorithms analyze a large number of network alarms, topologies, and work orders, extract alarm correlation rules, and generate an alarm correlation model. This model applies to scenarios where duplicate work orders are generated after a network fault occurs. The model allows work orders to be condensed, reducing repeated work orders and O&M costs.
DC PUE optimization model generation service

This service provides a set of automatic modeling tools (such as the DC topology template, PUE feature/algorithm library, and model training platform) that combine the AI technologies and DC engineering experience. With these tools, energy engineers without AI and coding knowledge only need to enter data.

Communication Model Services:

Professional and efficient communication model services are provided based on the Huawei SoftCOM AI inference platform.

KPI anomaly detection model service

This service identifies KPI anomalies from a large amount of KPI input data, identifies the KPI input mode based on the service configuration and data type, automatically optimizes algorithms, and helps to predict system faults or quickly locate faults.

ECA anomaly detection model service

Based on AI algorithms, this service extracts malicious traffic characteristics from large amounts of sample data, and delivers detection results through anomaly detection model training. It can quickly check network traffic security and identify malicious encrypted traffic without decrypting the traffic or destroying data privacy, helping customers improve network defense capabilities and reduce system risks.

Disk anomaly detection model service

Based on the industry’s standard Self-Monitoring, Analysis and Reporting Technology (S.M.A.R.T) indicator set for hard disks, this service extracts over 30 key features to build an AI model, delivers the health status detection result of hard disks, and predicts hard disk faults. It helps customers build a proactive fault handling mechanism for hard disks to ensure the reliable operation of systems and services.

6) Deployment Scheme

The SoftCOM AI solution supports flexible deployment based on different scenario requirements (for example, delay sensitivity and data sensitivity of application cases, and whether cloud services need to be deployed). The training platform or data lake implements offline model training and development, focusing on deployment convenience and data sensitivity. The inference framework mainly serves to implement real-time inference and it also focuses on delay sensitivity of application scenarios.

Training Platform/Data Lake

The training service platform is deployed mainly in three modes: HUAWEI CLOUD deployment, jointly-operated cloud deployment, and FusionCloud Stack (FCS) on-premises cloud deployment.
Inference Framework

Based on the delay requirements and inference function scope of application cases, the inference platform supports various deployment modes, including device embedding deployment, single-domain network management system (NMS) integration deployment, cross-domain private cloud deployment, and HUAWEI CLOUD deployment.

Device embedding deployment

AI inference functions are embedded into devices and are executed in real time on the devices. This deployment mode is featherweight. The memory space allocated to an application can be less than 50 MB. The deployment mode supports only the execution of AI algorithm models, but does not support functions such as data batch processing, stream processing, and Paas functions such as software lifecycle management. The deployment mode is applied to scenarios with high timeliness requirements. A typical use case is multi-carrier optimization by intelligent grid.

NMS integration deployment

The inference platform is embedded into each single-domain control system or edge device. AI inference functions are integrated into multiple single-domain or cross-domain NMS products. This deployment mode is lightweight. The inference platform is deployed on one to three VMs and self-managed through the lightweight gPaaS. Applications share the online computing platform. This deployment mode applies to real-time stream data processing, scenario in which data is not stored locally, the existing NMS is used, and the requirements on real-time performance are not high. Typical use cases include Massive MIMO, cyber attack detection, intelligent shutdown of base stations, passive optical network (PON) fault prediction and location, optical-layer commissioning, core network anomaly detection, and DC power usage effectiveness (PUE) optimization.

Private cloud deployment

The inference platform is deployed in the form of a data lake and in centralized mode on an operator's entire network. The inference platform is provided as a private cloud service. Multi-tenant management and isolation capabilities are supported. With the built-in data lake, the platform provides stream data processing and batch data processing capabilities. Long-term and cross-domain data can be imported to the data lake for comprehensive analysis. It is a medium-level deployment mode. The inference platform is deployed on more than three VMs on cloud networks. This deployment mode applies to scenarios in which the inference is implemented across multiple applications and domains, and cyber security requirements are high. Typical use cases include DCI resource utilization improvement and video service experience optimization.

HUAWEI CLOUD deployment

The inference platform is deployed in distributed data federation form. This deployment mode provides inference services for global customers, and supports service subscription and operations management. Platforms are physically scattered, logically centralized, and managed in a unified manner. This deployment mode applies to scenarios in which inference tests are provided to developers and inference services are provided to users or trial users. This deployment mode is not used in scenarios in which data privacy is an important concern. Typical use cases include assigning one work order for one fault and DC disk fault prediction.
Enhancing operators’ network intelligence is a long-term project and cannot be accomplished all at once. Huawei will start with cases with significant improvement in operators’ efficiency, increase cases, collect data, improve algorithms, accumulate experience, explore deployment modes, and gradually use AI on operators’ networks.

1) Leap in O&M Efficiency

In the O&M domain, O&M practices can be divided into three categories, based on how proactive or passive they are. The first category is called Run-to-Failure (R2F). If there is a fault on a network, the O&M personnel immediately go to the site and rectify the fault. This is the lowest level. The second category is called Preventive Maintenance (PM). With PM, each device is checked to prevent faults, but the efficiency is low. The third category is called Predictable Maintenance (PM). With PM, we can calculate the probability that a device will become faulty in the future and then perform targeted maintenance. With PM, we hope to reduce the workload required for alarm handling and fault locating on telecom networks by 90%, predict the failures and deterioration of 90% of key components, and further achieve network self-healing. In addition, more than 70% of the problems in network faults lie in passive equipment, such as aging or bent optical fibers, and ports that have become loose. In all of these situations, signals change. AI is introduced to learn the characteristics of these changes, and enables operators to predict the changes in advance and use solutions to solve passive faults. Fault location and prediction for PON optical modules and optical links provides an instructive example of the effect AI can have on O&M practices.

Fault Location and Prediction for PON Optical Modules and Optical Links

As a broadband solution used to solve the “last mile” problem, the PON technology has been widely deployed around the world due to the proliferation of the “fiber-in and copper-out” strategy. However, although new technologies bring benefits, faults of passive PON components are difficult to locate and there is no efficient remote locating methods. As a result, the O&M cost is high. According to the statistics of one operator, 25,000 complaints were received from July to December in 2016, and the ratio of faults regarding optical links and OLT/ONT optical modules (optical-to-electrical conversion modules) was 34%. In addition, due to urban reconstruction, rain, snow, and damage caused by rodents gnawing on cables, the fault ratio of devices covered by leather materials exceeded 20%, and the ratio of faults regarding various outdoor connectors (such as mechanical splicers and RJ45 connectors) exceeded 6%. Passive components cannot send any information. If a fault occurs on a device, it is difficult to locate the fault.

Optical fibers are buried in pipes and cannot be detected. According to the current process, once a home broadband user complains about a fault, O&M personnel must go to the site for fault location by checking the entire link on the optical network terminal (ONT) using the optical time domain reflectometer (OTDR). However, the cost of dispatching onsite O&M operators is high. The average cost is EUR175 in Europe and CNY50 in China. In 2017, the total onsite O&M cost of an operator was more than CNY100 million.

17
18
The SoftCOM AI solution uses NMSs (such as U2000 and uTraffic) to obtain historical data from ONT/OLT optical modules (such as module types, module voltages, module currents, module temperature, optical-layer alarms, optical-layer statistics, transmit power, receive power, and luminosity distance), extract features from the data, and label the features for AI training. The models obtained after the AI training can be used for inference based on real-time data to obtain the fault type and fault demarcation information, thereby implementing accurate fault location and prediction. The SoftCOM AI solution provides remote locating methods, improving remote fault location accuracy from 30% to 80%.

2) Leap in Energy Efficiency

When it comes to energy efficiency, the number of bits should determine the number of watts. That is, lower traffic volume should result in lower power consumption. In an equipment room or at a site, each system is configured with dozens of parameters. The heat dissipation, environment, and service load models are generated through AI training to maximize the energy consumption efficiency for sunlight, temperature, and auxiliary facilities, such as diesel generators, solar energy devices, and batteries. At the equipment layer, dynamic energy distribution is performed based on service loads. If there is no traffic, power consumption is reduced by using timeslot shutdown, RF deep sleep, and carrier shutdown. In addition, energy consumption of data center hardware, such as server components, can be dynamically reduced. The power consumption of the network system must also be taken into account. The accurate service load prediction model is constructed to balance the traffic on the entire network and achieve the optimal energy consumption efficiency.

Intelligent Energy Saving for Base Stations

Statistics show that the base station power consumption cost (electricity fee) accounts for more than 16% of network operations costs. Optimizing the cost structure and promoting social development, energy saving and emission reduction of base stations have become main goals of operators.

In the case of a conventional macro base station, the power consumption of the main equipment accounts for 50% of total power consumption, and the power consumption of RF units accounts for 80% of the main equipment power consumption. The power consumption of power amplifiers (PAs) accounts for 79% of the power consumption of RF units. Distributed base stations are becoming the industry standard, and this trend results in the main equipment taking an increased share of the total power consumption. Meanwhile, the traffic volume on the network is subject to regular ebb and flow. Peak traffic volume can reach four times off-peak traffic volume. However, a majority of base stations keep running 24 hours a day (with all resources occupied). The power consumption does not dynamically change with traffic volume, which results in resource waste. Therefore, in terms of main equipment, reducing energy consumption by shutting down base station carriers is key to energy saving for base stations.

Generally speaking, the intelligent energy saving strategy of base stations is to reduce or shut down power amplification modules that carry carriers. Power amplification modules are hardware entities, which are used to amplify the power of modulated carrier signals and transmit those signals. Carrier shutdown is applicable to scenarios where a sector is covered by two or more carriers. Amplification modules of capacity cells are shut down, and only coverage cells are reserved for basic coverage requirements. Base station carrier shutdown is not a new feature, but the usage of this feature is low. This is due to the fact that traditional shutdown modes (such as scheduled shutdown) rely heavily on a unified default value, which occasionally affects users’ call and Internet access experience.

How to make full use of this feature? If radio resource usage of base station cells can be predicted, a customized sleeping time can be set for each cell, meaning that operators no longer have to worry about the unintended consequences of using the base station carrier shutdown feature. However, prediction is not easy. First, it is necessary to obtain the historical data of tens of thousands of cells, such as the time, neighboring cell relationship, event, and radio resource usage features. In addition, we need to monitor the changes of KPIs and KQIs and dynamically adjust the shutdown parameters based on the KPI or KQI changes after the solution is deployed based on adjustment policies. Associating uncertain factors that affect the prediction result with radio resource usage is difficult and cannot be implemented in practice.

Using the neural network algorithm of the spatio-temporal computing module on the AI training platform, we can find the mapping relationship between the historical data of features that affect the prediction result and radio resource usage, and determine the weight matrix and bias matrix. In this way, we can generate a radio resource usage prediction model. After the prediction model is generated, it can be deployed in the system. We can set the sleeping threshold based on the prediction result of radio resource usage, and calculate the time segment that meets the cell sleeping threshold requirements.

Take the result at a first office application (FOA) site as an example. The AI-based intelligent shutdown of base stations can help operators reduce base station power consumption by 10% to 15% while the coverage area remains unchanged and KPIs and KQIs are not affected.

3) Leap in Resource Efficiency

Regarding network resources, on most present-day networks, the flow of traffic simply follows the physical topology of the network, and resource usage may be unreasonable. If network scheduling took into account the direction of traffic flow, the resource usage efficiency could be greatly improved. At present, however, networks do not have this capability. To address this issue, AI must be introduced, and a traffic prediction model must be created. In this way, precise traffic prediction and the optimal network topology can be provided, and the network paths can be determined by traffic instead of just by physical connections.

Massive MIMO Intelligent Optimization

5G is the next major opportunity for the telecom industry. Progress has been driven by all parties in the value chain, 5G is no longer far away from us. In the network domain, preparations for the 5G era must be made today.

In urban areas, wireless spectrum resources are gradually becoming saturated. To meet the rapid growth of service requirements, it is essential to increase the number of 4G and 5G antennas. Massive MIMO is one of the core technologies of 5G. To make full use of its advantages, it is necessary to flexibly adjust the parameters of Massive MIMO base stations to cope with service changes.

However, antenna parameter combinations vary with location, scenario, and user distribution. In the 5G era, there were only 13 parameter combinations (horizontal beamwidth and vertical beamwidth) and it is easy to make a selection based on the judgment of experienced experts. In the 4G era, there are hundreds of antenna parameter combinations.
The ECA solution implements encrypted traffic detection without decryption based on AI technologies, with which user privacy can be protected. The detection accuracy is high, reaching 99.9%, while the false positive rate is lower than 0.001%. Intelligent detection of five categories covering 30 families of malicious files is supported.

5) 5G Intelligent Networks

AI will become a native capability of 5G, enabling scalable, flexible, and intelligent 5G networks. Huawei will integrate AI into E2E networks (wireless network, fixed network, core network, and DCs) based on network planning, deployment, optimization, and O&M, to achieve intelligent 5G networks.

In the near future, we will implement 5G base station deployment intelligence, 5G slicing intelligence, and 5G DC intelligence. During the deployment of a 5G site, the site location and deployment parameters will be automatically recommended based on multi-attribute site planning. Low-granularity features are automatically enabled and parameters are automatically optimized. Intelligent 5G slicing is the process of continuously enhancing network slicing based on the automation capability built up by the central cloud and edge cloud. The system predicts resource consumption based on service scenarios to recommend the optimal slicing model and orchestration policy, implementing real-time resource awareness, on-demand adjustment, and simplified configuration. In this way, service rollout time is shortened. 5G DC intelligence involves intelligent scale-out configuration, L1 to L4 cross-level energy saving, intelligent inspection, hard disk fault prediction and maintenance, and intelligent O&M through mobile edge computing (MEC). In addition, the KPI anomaly monitoring, alarm correlation and diagnosis, and abnormal traffic identification for the core network that is cloudified through NFV and is running on DCs are also involved.

Typical Application Scenarios of the Huawei SoftCOM AI Solution

Typical Application Scenarios of the Huawei SoftCOM AI Solution

Artificial intelligence technologies are applied to the planning, deployment, optimization, and maintenance of 5G networks, greatly simplifying network deployment. During site planning, base station coverage status and the distribution of sites across the entire network are visualized and measurable. During site deployment, hardware is automatically discovered and configured, cells are automatically bound, and radio parameters are automatically configured (Scenario-based policy parameters are intelligently defined: transmission or radio parameters are greatly simplified, base stations with low-granularity features enabled are automatically configured, and no O&M engineers are required for intelligent commissioning), implementing automatic 5G site deployment. During network optimization, networks accurately identify various network coverage scenarios and automatically adjust network parameters to achieve optimal network performance based on the optimization objective, 3D map, online map, and 5G traffic map. During network maintenance, the system automatically analyzes network logs and alarms and automatically identifies and predicts network fault.

When a hacker establishes an encrypted control channel to control “zombies”, the next-generation firewall (NGFW) extracts encrypted traffic packets and reports them to the cybersecurity intelligence system (CIS) for feature detection. The CIS uses an AI model to detect traffic features and interworks with the NGFW to intercept north-south traffic and corporates with switches to perform east-west isolation of infected hosts. The CIS delivers the detected traffic features to tools, so that the tools can locate, analyze, and clear malicious files.

Typical Application Scenarios of the Huawei SoftCOM AI Solution

Typical Application Scenarios of the Huawei SoftCOM AI Solution

AI-driven E2E Simplification of 5G Network Deployment

AI is applied to the planning, deployment, optimization, and maintenance of 5G networks, greatly simplifying network deployment. During site planning, base station coverage status and the distribution of sites across the entire network are visualized and measurable. During site deployment, hardware is automatically discovered and configured, cells are automatically bound, and radio parameters are automatically configured (Scenario-based policy parameters are intelligently defined: transmission or radio parameters are greatly simplified, base stations with low-granularity features enabled are automatically configured, and no O&M engineers are required for intelligent commissioning), implementing automatic 5G site deployment. During network optimization, networks accurately identify various network coverage scenarios and automatically adjust network parameters to achieve optimal network performance based on the optimization objective, 3D map, online map, and 5G traffic map. During network maintenance, the system automatically analyzes network logs and alarms and automatically identifies and predicts network fault.
Typical Application Scenarios of the Huawei SoftCOM AI Solution

**Abbreviation**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI</td>
<td>Artificial Intelligence</td>
</tr>
<tr>
<td>SDN</td>
<td>Software Defined Network</td>
</tr>
<tr>
<td>NFV</td>
<td>Network Function Virtualization</td>
</tr>
<tr>
<td>SoftCOM</td>
<td>Software Telecom</td>
</tr>
<tr>
<td>CRM</td>
<td>Customer Relationship Management</td>
</tr>
<tr>
<td>IT</td>
<td>Information Technology</td>
</tr>
<tr>
<td>AWS</td>
<td>Amazon Web Services</td>
</tr>
<tr>
<td>OPEX</td>
<td>Operating Expense</td>
</tr>
<tr>
<td>DCN</td>
<td>Data Center Network</td>
</tr>
<tr>
<td>ETL</td>
<td>Data Extraction Transformation Loading</td>
</tr>
<tr>
<td>BDI</td>
<td>Big Data Integration</td>
</tr>
<tr>
<td>OMC</td>
<td>Operation and Maintenance Center</td>
</tr>
<tr>
<td>OSS</td>
<td>Operations Support System</td>
</tr>
<tr>
<td>FCS</td>
<td>Fusion Cloud Stack</td>
</tr>
<tr>
<td>IES</td>
<td>Infrastructure Enabling System</td>
</tr>
<tr>
<td>MAE</td>
<td>MBB Automation Engine</td>
</tr>
<tr>
<td>NCE</td>
<td>Network Cloud Engine</td>
</tr>
<tr>
<td>CDE</td>
<td>Cloud Container Engine</td>
</tr>
<tr>
<td>OLT</td>
<td>Optical Line Terminal</td>
</tr>
<tr>
<td>PON</td>
<td>Passive Optical Network</td>
</tr>
<tr>
<td>ONT</td>
<td>Optical Network Terminal</td>
</tr>
<tr>
<td>RF</td>
<td>Radio Frequency</td>
</tr>
<tr>
<td>PA</td>
<td>Power Amplifier</td>
</tr>
<tr>
<td>ECA</td>
<td>Encrypted Communication Analytics</td>
</tr>
<tr>
<td>APT</td>
<td>Advance Persistent Threat</td>
</tr>
<tr>
<td>NGFW</td>
<td>Next-Generation Firewall</td>
</tr>
<tr>
<td>CIS</td>
<td>Customer Information System</td>
</tr>
<tr>
<td>MEC</td>
<td>Mobile Edge Computing</td>
</tr>
<tr>
<td>EOO</td>
<td>Coverage and Capacity Optimization</td>
</tr>
<tr>
<td>MLB</td>
<td>Mobility Load Balancing</td>
</tr>
<tr>
<td>PnP</td>
<td>Plug and Play</td>
</tr>
<tr>
<td>OLT</td>
<td>Optical Line Terminal</td>
</tr>
</tbody>
</table>

**Conclusion**

The future will be shaped by intelligence. Enhancing operators’ network intelligence is a long-term project and cannot be accomplished overnight. SoftCOM AI is the implementation of Huawei’s AI Intelligence strategy in the telecom domain. The key AI capability of SoftCOM AI is developed based on both Huawei’s long-term strategic investment and growth in AI Intelligence and telecom scenarios. SoftCOM AI helps operators build “never faulty” autonomous networks and achieve digital and intelligent transformation.

Currently, the SoftCOM AI solution is open to industry insiders for trial use. We welcome you to go to our cloud service website and experience our telecom AI development platform and intelligent communication model services.
Contacts:

e-Mail: yuanyiming@huawei.com
Official Web: https://telcloud.huawei.com/#

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Disclaimer:
The information contained in this guide is for general information purposes only, and may be changed to fit the local environment. The guide only for reference, not representations or warranties of any kind, express or implied. Huawei keeps the rights to update or change the information without any notification.