Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 5, No. 3, 2025

Practical Research on Artificial Intelligence Technology in Transmission Optimization

Yaqi Wu, Ran Feng

School of Information Science and Engineering, Shenyang University of Technology, Shenyang, 110000, China

Abstract: Under the large-scale application of 5G-A and AI models, transmission networks face challenges such as bandwidth surges, latency sensitivity, and rising energy consumption, making traditional optimization methods difficult to adapt to dyn amic environments. This paper focuses on the practice of AI in transmission optimization, analyzes the adaptation logic of machine learning, deep reinforcement learning and transmission network, and studies congestion control, resource scheduling, and collaborative reasoning from the perspective of congestion control, resource scheduling, and collaborative reasoning in combination with communication network, data center, and edge computing scenarios. Through case studies such as Singapore's M1 microwave transmission network and Shanghai Unicom's 5G-A intelligent scheduling system, verify the effectiveness of AI in improving bandwidth utilization, reducing latency, and optimizing energy consumption. Research has shown that AI dynamic optimization can increase bandwidth utilization by more than 40%, reduce end-to-end latency by about 30%, and save energy consumption by 18% -35%. It provides a feasible path for the intelligent upgrade of transmission networks and has reference value for promoting the deep integration of AI and communication transmission.

Keywords: Artificial intelligence, transmission optimization, congestion control, resource scheduling, edge collaboration, machine learning.

1. Introduction

In the context of the development of the digital economy, transmission networks, as the core carrier of information exchange, are facing severe pressure. The explosion of emerging applications such as AI big model training, 8K ultra high definition live streaming, and metaverse has led to exponential growth in bandwidth demand for transmission networks, while traditional transmission technologies rely on fixed parameter configuration and static optimization strategies, making it difficult to cope with dynamic network loads and differentiated business needs. For example, traditional TCP congestion control algorithms require a longer time to fully utilize available bandwidth in high bandwidth delay product (BDP) networks; During multi GPU cluster training, PCIe bus limitations can easily lead to bandwidth bottlenecks, which severely restrict the improvement of digital application experience.

Artificial intelligence technology provides a new path for transmission optimization through its perception, learning, and decision-making abilities. From the intelligent bandwidth detection of Google BBR algorithm to the AI transmission automation control of Singapore M1, practice has proven that AI can accurately capture changes in network status, dynamically adjust transmission strategies, and achieve a transition from "passive response" to "active optimization". Currently, the application of AI in transmission optimization has covered areas such as congestion control, resource scheduling, and fault diagnosis, and the implementation of related technologies has produced significant benefits [1].

This article is based on mature practical cases to explore in depth the core application logic and implementation path of AI technology in transmission optimization. Firstly, it analyzes the practical difficulties of transmission networks and the advantages of AI technology adaptation. Then, it elaborates on the key AI technology practical application mechanisms. Finally, it verifies the optimization effects through multi scenario cases. Finally, it summarizes the existing problems and looks forward to future directions, providing practical references for the intelligent transformation of transmission networks.

2. The Realistic Dilemma of Transmission Optimization and The Adaptability of AI Technology

2.1. Core Bottleneck of Traditional Transmission Optimization

Traditional transmission optimization relies on manually fixed rules and static algorithms, which highlight locality in complex dynamic networks. In terms of bandwidth utilization, the traditional TCP's "additive increase, multiplicative decrease" mechanism converges slowly in high BDP networks, requiring a long time to fully utilize bandwidth, resulting in resource idle. In terms of latency control, the PCIe 4.0 bus (64GB/s) that multi GPU cluster training relies on is difficult to meet the parallel transmission requirements, and the queuing delay caused by congestion seriously affects training efficiency.

In terms of reliability, the expansion of link scale significantly increases the probability of failures, and failures occur frequently in scenarios with millions of links. Traditional manual fault localization is inefficient. The energy consumption problem is prominent, and the energy consumption of AI data centers is rapidly increasing. Traditional transmission equipment cannot dynamically adapt to traffic and adjust energy consumption, resulting in waste. When adapting to multiple scenarios, the interruption of 5G mobile scene signal switching and the contradiction between edge device resources and AI computing power highlight the lack of flexibility in traditional optimization.

2.2. Transmission Optimization and Adaptation Advantages of AI Technology

AI technology, with its data-driven decision-making mechanism, naturally adapts to the dynamic optimization needs of transmission networks. Machine learning algorithms can extract traffic characteristics and channel patterns from massive network data, establish accurate state prediction models, and provide a basis for optimizing strategies. For example, the classification model based on XGBoost can analyze historical congestion data, adaptively adapt to different network scenarios, and achieve performance close to the comprehensive level of multiple excellent traditional algorithms [2].

Deep reinforcement learning (DRL) provides an efficient path for complex constraint optimization, modeling transport optimization as a Markov decision process, allowing agents to continuously optimize strategies in interactions. Micro algorithm technology practice has shown that DRL can effectively solve the model splitting decision problem of edge collaborative reasoning, while minimizing energy consumption while satisfying latency constraints. The real-time and parallel capabilities of AI are in line with the dynamic changes in the network. The Shanghai Unicom 5G-A intelligent scheduling system achieves precise matching of business and network resources through millisecond level status collection, ensuring "second level response" in high-density scenarios. In addition, AI self-learning ability can adapt to network topology and business iteration without frequent manual intervention, improving optimization efficiency and robustness.

3. The core AI Technology Practice Mechanism in Transmission Optimization

3.1. Machine Learning Driven Congestion Control Optimization

Congestion control is the core of transmission optimization, and machine learning breaks through the limitations of traditional TCP algorithms by reconstructing congestion perception and decision-making mechanisms. The Google BBR algorithm adopts a "pipeline model" that relies on machine learning to accurately measure bottleneck bandwidth and round-trip propagation time (RTprop), dynamically adjust to avoid queue accumulation, and in YouTube deployment, it improves throughput by an average of 4% and reduces RTT by 33%, outperforming traditional packet loss driven algorithms [3].

The TCP-PPO2 algorithm based on reinforcement learning abstracts congestion control as a partially observable Markov process and adapts to network fluctuations through near end policy optimization; The TCP SIAD algorithm integrates scalable increase and adaptive decrease strategies to achieve dual optimization of bandwidth utilization and latency, both of which are free from fixed rule dependencies and adaptable to multi scenario requirements.

3.2. Application of Deep Reinforcement Learning in Resource Scheduling

Deep reinforcement learning (DRL) provides an efficient solution for dynamic allocation of transmission resources and is suitable for multi constraint scenarios. Micro algorithm technology breaks down the model splitting and resource allocation of device edge collaborative reasoning into sub problems, trains intelligent agents through DRL, and determines splitting points based on channel gain, device energy, and other factors to balance latency and energy consumption.

The AI native transmission technology (ANT) released by ODCC adopts a packet by packet balancing strategy, dynamically adjusts the traffic path through reinforcement learning, and avoids throughput attenuation caused by hash conflicts; The carrier intelligent selection algorithm for Shanghai Unicom's 5G-A network is based on DRL analysis of service load, prioritizing core services during peak periods and improving transmission quality in complex scenarios.

3.3. Integration and Optimization of AI And Transmission Protocols

Deep integration of AI and transmission protocols to reconstruct the transmission layer architecture and solve the traditional protocol adaptation dilemma. The combination of QUIC protocol and AI has become a key focus of real-time transmission optimization, relying on 0-RTT connection and AI priority transmission mechanism, dynamically adjusting NACK timeout time, and achieving video integrity of 92% under 3% packet loss rate, far exceeding the 65% of traditional TCP.

RDMA and AI collaborative optimization break through the bottleneck of multi GPU cluster transmission. Deepseek's patented technology adjusts message size through AI dynamic slicing algorithm, and combines the main and auxiliary GPU parallel architecture to achieve multi network card bandwidth superposition, achieving zero copy of GPU video memory and remote memory. The throughput of four RDMA network cards reaches four times that of a single card, freeing up CPU and ensuring full utilization of GPU computing power.

4. Multi Scenario Practical Application of AI Transmission Optimization

4.1. Optimization Practice of Communication Network Transmission

The AI transmission automation control system of Singapore M1 and Ericsson is a benchmark for microwave network optimization. The system perceives network status in real-time through machine learning, identifies interference sources, and dynamically adjusts parameters. During the concept verification phase, it achieves an 18% reduction in energy consumption and a 3-fold increase in fault localization speed; After integration into the M1 microwave network, millisecond level sensing optimization ensures 99.999% connectivity reliability in port scenarios.

The 5G-A intelligent scheduling system deployed by Shanghai Unicom at the CIIE collects data through the collaboration of "intelligent agents+intelligent network elements" and matches residential areas through AI carrier intelligent selection algorithm. Supported by 84 5G-A base stations, the network has an average speed of 1.8Gbps and a peak speed of 2.9Gbps, ensuring "instant transmission without lag" for 8K live broadcasts and metaverse booths, and "zero congestion" for QR code scanning in scenarios with tens of thousands of people.

4.2. Data Center Transmission Optimization Practice

The Google TPU V4 adopts an AI driven optical switching architecture, which connects cubes through OCS to form a 3D Torus structure. AI monitors the link status in real time, and can flexibly adjust the path when the daily failure rate of a single link is 0.004%. The million level link still runs stably; The OCS layer replaces the traditional electrospin layer, achieving a 30-50% reduction in power consumption and delay optimization.

Deepseek's RDMA dynamic slicing technology dynamically adjusts transmission strategies through an AI three-stage congestion algorithm, reducing slicing to 64KB during congestion. Tested on a kilocard GPU cluster, ResNet-152 training throughput increased by 2.3 times, GPT-4 level model inference cross node communication overhead decreased from 15% to 5%, and bandwidth utilization doubled without hardware upgrades.

4.3. Collaborative Optimization Practice of Edge Transmission

The device edge collaborative reasoning system of micro algorithm technology uses DRL and convex optimization fusion algorithm to optimize model splitting and resource allocation. In the scenario of multi-user indoor crowd counting, sensor splitting is performed using a convolutional model, and AI dynamically adjusts the splitting point based on channel and computing power, achieving a delay of 100-300ms while reducing terminal power consumption by 35% [4].

The AI generated content transmission solution for 5G mobile edge scenes embeds AI modules in SimSwapHQ videos, dynamically adjusts resolution based on network status, and cooperates with H.265 ROI encoding and edge cloud to improve bandwidth utilization by 43% and facial detail retention by 27% under the same quality, solving the problem of 5G bandwidth fluctuation and lag.

5. Evaluation of the Practical Effectiveness of AI Transmission Optimization

From the perspective of bandwidth utilization efficiency, AI driven transmission optimization has significant advantages. Deepseek's RDMA dynamic slicing technology uses a multi network card parallel architecture to achieve a throughput of four RDMA network cards that is four times that of a single card. The bandwidth utilization rate in Long Fat Network (LFN) has increased from less than 50% to over 90%; The Shanghai Unicom 5G-A intelligent scheduling system utilizes high and low frequency collaboration and multi carrier aggregation, with a peak of 10Gbps and an average of 5Gbps in the core exhibition area, far exceeding traditional 5G. ODCC's ANT technology packet by packet balancing scheme avoids throughput attenuation in traffic hash conflict scenarios, and network utilization is close to the theoretical optimum. In terms of latency control, the deployment of Google BBR algorithm resulted in an average reduction of 33% in RTT. Deepseek dynamic congestion control reduced transmission latency to milliseconds, resulting in an overall reduction of 30%. The micro algorithm technology edge collaboration system controlled end-to-end latency within 100-300ms. QUIC and AI priority scheduling were combined in AI generated content transmission, reducing video latency from 2.3 seconds to within 200ms under a 3% packet loss rate.

AI has achieved outstanding results in optimizing transmission energy consumption: Singapore M1's AI transmission system reduces microwave network energy consumption by 18%, Google AI optical switching architecture saves 30-50%, micro algorithm technology edge collaboration solution reduces sensor power consumption by 35%, and Shanghai Unicom's "self breathing" network uses AI to wake up dormant carriers to achieve precise energy saving during low valley periods, with a dynamic mechanism that balances performance and energy saving. In terms of reliability, the AI system of Singapore M1 accelerates fault location by three times, while the Google TPU V4 relies on AI architecture optimization and rapid maintenance to ensure stability in million level link scenarios. The ANT technology loss tolerance solution solves the traditional PFC head impedance problem, and AI improves network availability through fault prediction and other methods.

In terms of cost-effectiveness, Deepseek's RDMA dynamic slicing technology does not require hardware upgrades and achieves bandwidth doubling by parallelizing existing network cards, reducing large model training time by 30% and lowering computing power costs; Shanghai Unicom relies on AI intelligent scheduling to optimize base station resources, achieving high-performance coverage across the entire area with 84 5G-A base stations at the CIIE, avoiding large-scale expansion. In industry applications, low latency transmission ensures the efficiency of multi-sensor fusion training in the field of autonomous driving, efficient transmission accelerates the training of billions of parameter models and cancer detection algorithms in the field of medical imaging [5], millisecond level synchronization in financial risk control scenarios meets high-frequency trading needs, and AI transmission optimization not only improves technical indicators but also supports the digital transformation of the industry.

6. Existing Problems and Optimization Paths in AI Transmission Optimization

In current practice, hardware reliability and adaptability issues constrain the large-scale implementation of AI transmission optimization. Google's practice has found that the surge in production demand for AI optical modules has made it difficult for traditional reliability testing to cover, with manufacturing quality and firmware vulnerabilities being the main sources of failure, not laser failures; Although some solutions such as CPO reduce power consumption, they require high reliability, significant impact of faults, and long maintenance cycles. In addition, due to differences in equipment interfaces and protocols from different manufacturers, it is difficult for AI algorithms to adapt across platforms.

At the algorithmic level, there is a challenge in balancing generalization and real-time performance: existing AI transmission algorithms are mostly trained for specific scenarios, and their performance is prone to fluctuations during network topology changes and business switching, lacking a unified adaptive framework [6]; Deep reinforcement learning requires a large amount of annotated data, and the dynamic nature of transmission networks causes

changes in data distribution, resulting in high model iteration costs; In millisecond level response scenarios, the inference delay of complex AI models has become a new bottleneck, and there are challenges in lightweight algorithm design. In addition, there is no unified standard for testing AI native transmission technology, and manufacturers have different evaluation indicators. ODCC's ANT testing standards only cover some scenarios, and large-scale networking testing methods need to be improved; AI autonomous decision-making may cause abnormal switching of transmission paths, leading to risks of data leakage or service interruption, and there is a lack of mature solutions for data privacy protection in model training.

For hardware adaptation, it is necessary to promote standardized design of "software hardware collaboration": strengthen the research and development of integrated technologies such as PEIC, reduce hardware dependence, and minimize transmission losses; Establish cross vendor hardware interface standards, promote the deep integration of AI algorithms and network devices, such as embedding intelligent scheduling modules into the underlying drivers of base stations, and drawing on Google's experience to improve hardware mass production quality through refined design and comprehensive FMEA.

Algorithm optimization needs to focus on generalization and lightweight: develop federated learning and transfer learning, train universal models with multi scenario data, and reduce data dependence; Integrating traditional optimization algorithms with AI, such as CUBIC combined with reinforcement learning, balancing stability and adaptive capabilities; Reduce inference latency through model compression and quantization, and develop a low code toolchain to lower deployment barriers. Further improvement is needed in industry standards and security mechanisms: expanding the coverage of ANT testing standards and establishing a unified testing framework for multiple scenarios; Promote the joint construction of AI transmission patent pools by multiple parties and develop cross industry standards; Introduce blockchain to ensure traceability of transmission paths, protect data privacy with differential privacy, and establish algorithm decision auditing mechanisms to prevent risks.

7. Conclusion

This article conducts practical research on the application of artificial intelligence technology in transmission optimization, and summarizes the adaptation logic, core technologies, application cases, and performance effects of AI and transmission networks, forming a practical summary. Research shows that AI can effectively break the bottleneck of traditional transmission in bandwidth utilization, delay control and energy consumption management by virtue of data-driven dynamic optimization mechanism, and achieve significant performance improvement in communication networks, data centers, edge computing and other scenarios. Case studies have shown that AI driven transmission optimization can increase bandwidth utilization by over 40%, reduce latency by about 30%, and save energy by 18% -35%, providing reliable transmission support for emerging applications such as 5G-A and AI big models.

The core value of AI transmission optimization lies in promoting the transformation of transmission networks from "passive configuration" to "active intelligence", integrating machine learning perception, reinforcement learning decision-making, and transmission protocol execution capabilities, and building an adaptive, highly reliable, and low-energy transmission system. The practices of Singapore M1 microwave network optimization, Shanghai Unicom 5G-A intelligent scheduling, Deepseek multi GPU transmission optimization, etc. not only verify the technical feasibility, but also form replicable implementation experience.

Currently, AI transmission optimization still faces challenges such as hardware adaptation, algorithm generalization, and lack of standards, which need to be improved through software hardware collaborative standardization, algorithm lightweight integration, and industry standard co construction. In the future, under the evolution of 6G, the integration of terahertz communication, intelligent metasurfaces, and AI will promote the development of transmission optimization towards higher bandwidth, lower latency, and better energy consumption; The establishment of cross industry standards and the improvement of security mechanisms will lay the foundation for the large-scale application of AI transmission technology and assist in the high-quality upgrading of transmission networks in the digital economy era.

References

[1] Wang, X., Shen, M., & Yang, K. (2024). On-edge high-throughput collaborative inference for real-time video analytics. IEEE Internet of Things Journal, 11(20), 33097-33109.

[2] Hendaoui, S., Hendaoui, F., & Zangar, N. (2024). Dynamic proactive-reactive scheduling for URLLC in 5G: Leveraging XGBoost and network virtualization. Physical Communication, 102553.

[3] Scholz, D., Jaeger, B., Schwaighofer, L., Raumer, D., Geyer, F., & Carle, G. (2018, May). Towards a deeper understanding of TCP BBR congestion control. In 2018 IFIP networking conference (IFIP networking) and workshops (pp. 1-9). IEEE.

[4] Xiao, Y., Xiao, L., Wan, K., Yang, H., Zhang, Y., Wu, Y., & Zhang, Y. (2022). Reinforcement learning based energy-efficient collaborative inference for mobile edge computing. IEEE Transactions on Communications, 71(2), 864-876.

[5] Zhang, S., Cui, Y., Xu, D., & Lin, Y. (2025). A collaborative inference strategy for medical image diagnosis in mobile edge computing environment. PeerJ Computer Science, 11, e2708.

[6] El-Hajj, M. (2025). Enhancing communication networks in the new era with artificial intelligence: techniques, applications, and future directions. Network, 5(1), 1.