GEORGE BROWN PRESS
JOURNALS, BOOKS & PROCEEDINGS

Journal of Computer, Signal, and System Research

Article

Application and Practice of Machine Learning Infrastructure Optimization in Advertising Systems

Yixian Jiang [1,*]

[1] Information Networking Institute, Carnegie Mellon University, Pittsburgh, PA, 15213, USA
* Correspondence: Yixian Jiang, Information Networking Institute, Carnegie Mellon University, Pittsburgh, PA, 15213, USA

Abstract: The explosive growth of advertising data has increasingly challenged the performance and predictive accuracy of traditional advertising platforms. Optimizing and upgrading machine learning infrastructure has emerged as a critical solution to address these challenges. This optimization encompasses the enhancement of computing hardware, the development of scalable and efficient distributed processing architectures, and the refinement of model training, tuning, and deployment strategies. Such improvements enable advertising platforms to handle massive volumes of data more efficiently while delivering more precise insights. This article explores the practical applications of machine learning infrastructure optimization in advertising systems, including personalized ad targeting, accurate estimation of advertising effectiveness, proactive fraud detection and risk management, and intelligent allocation of advertising resources. Empirical evidence suggests that these optimizations not only improve the efficiency and quality of ad placement but also play a pivotal role in maintaining system stability, reducing operational costs, and supporting more informed, data-driven decision-making in dynamic advertising environments.

Keywords: machine learning; infrastructure optimization; advertising system; performance improvement; accurate targeting

1. Introduction

With the rapid expansion of the advertising industry and the continuous increase in data traffic, advertising platforms face mounting pressure to ensure both the efficiency and accuracy of their advertising operations. Achieving these goals increasingly relies on continuous technological innovation, with machine learning emerging as a central driving force. The optimization and upgrading of machine learning infrastructure play a critical role in this process, as they directly influence the overall performance, responsiveness, and scalability of advertising systems [1].

By enhancing machine learning infrastructure, platforms can significantly improve the operational efficiency of advertising delivery, refine targeting precision, and elevate user engagement and satisfaction [2]. Key strategies for optimization include upgrading hardware structures to support higher computational demands, implementing distributed computing architectures for faster data processing, refining model training processes and parameter adjustments for improved predictive accuracy, and automating operational and maintenance workflows to reduce manual intervention and errors.

Furthermore, practical applications demonstrate the tangible benefits of these optimizations. For instance, advanced infrastructure can enable real-time processing of large-scale user data, support adaptive personalization of content, and facilitate the dynamic allocation of advertising resources. These improvements not only enhance the effectiveness of advertising campaigns but also contribute to a more seamless and engaging user experience.

Received: 09 September 2025
Revised: 17 September 2025
Accepted: 29 October 2025
Published: 02 November 2025

This article systematically examines the principal approaches to optimizing machine learning infrastructure in advertising systems. It presents specific case examples illustrating the effectiveness of these methods and discusses the potential future trends in intelligent advertising development, highlighting the evolving integration of AI technologies into marketing strategies and platform operations [3].

2. The Role of Optimizing Machine Learning Infrastructure in Advertising Systems

2.1. Improving System Performance

The performance of an advertising system is a critical factor in determining its overall operational effectiveness, as illustrated in Figure 1. Optimizing the underlying machine learning infrastructure can substantially enhance the computational efficiency of the system, particularly in areas such as data processing, storage, and transmission. Implementing advanced hardware configurations, such as high-performance GPUs, accelerates the processing of large-scale datasets and model training procedures, reduces computational latency, and improves response times for advertising delivery.

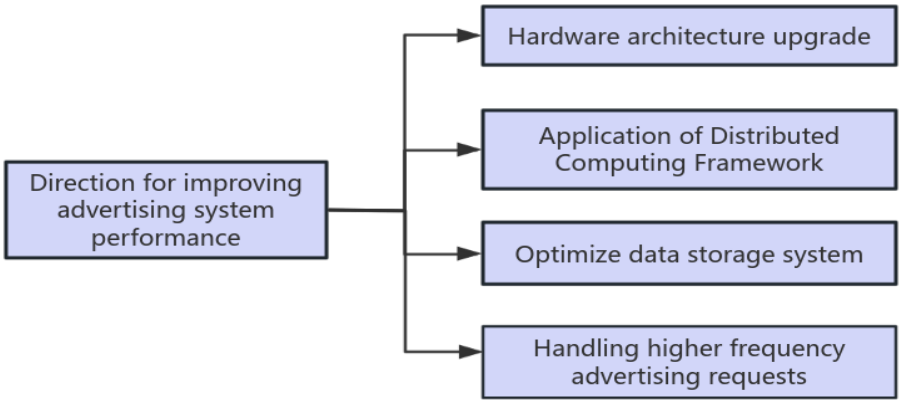

Figure 1. Performance improvement of advertising system.

In addition, the adoption of a distributed computing architecture allows advertising platforms to efficiently allocate workloads and manage complex computational tasks. This capability is especially vital for real-time advertising recommendations and performance evaluations, where rapid data processing and decision-making are required [4]. Optimization of the data storage system further reduces delays in data retrieval, improves data access efficiency, and ensures the precise placement of advertisements as well as accurate evaluation of their effectiveness.

The resulting optimized system is capable of handling more frequent and diverse advertising requests, maintaining stability under high traffic conditions, and ensuring continuous availability of the advertising platform. These improvements collectively enhance both the operational reliability and the user experience, laying the foundation for more intelligent, responsive, and data-driven advertising strategies [5].

2.2. Make Advertising Placement More Precise

By thoroughly optimizing the advertising placement architecture, platforms can leverage advanced algorithms to gain a deeper understanding of user preferences and deliver highly personalized advertising content [6]. Through the intensive analysis of big data and the application of intelligent algorithms, platforms are able to accurately identify users' preferences, behavioral patterns, and potential consumption tendencies, thereby customizing advertising content for each individual.

At the architectural level, enhancements in distributed computing environments and upgrades in storage solutions allow platforms to efficiently process massive datasets, construct more precise user profiles, and support advertisers in formulating more effective marketing strategies. Machine learning models continuously refine themselves by analyzing both historical user behavior and real-time feedback, dynamically adjusting advertising delivery strategies, and progressively improving the accuracy of recommendations [7].

Moreover, optimizing hardware configurations and the model training process accelerates algorithm iteration, enhances the system's learning efficiency, and increases the precision of ad placement. The resulting optimized advertising distribution system not only boosts user engagement and transaction rates but also significantly improves advertising return on investment, creating greater commercial value for advertising publishers.

3. Optimization Methods for Machine Learning Infrastructure

3.1. Improvement of Hardware Architecture

With the continuous advancement of machine learning technologies and the rapid growth of data volume, conventional hardware infrastructures are increasingly inadequate to meet the high-performance requirements of advanced advertising systems in terms of computing power and data processing efficiency. Upgrading and enhancing hardware architecture is therefore central to improving the overall performance of machine learning infrastructure. To accommodate the rapid development of complex algorithms, advertising systems must transition to more flexible and easily scalable hardware platforms. Compared with traditional devices, modern hardware solutions emphasize multi-level computing capabilities and highly parallel processing, offering significant advantages for distributed computing and data parallelization [8].

Many advertising platforms have adopted GPU-based accelerated computing systems to transform traditional CPU-centered processing. This architecture improves the performance of deep learning algorithms during both training and inference, particularly when processing large volumes of advertising data, thereby reducing model training cycles. Additionally, some platforms have integrated specialized accelerators, such as FPGAs and TPUs, to optimize hardware for machine learning tasks. These customized hardware components streamline computational workflows and enhance execution efficiency within the advertising system. For example, TPU accelerators are employed in certain large-scale advertising systems to enable real-time optimization of ad placement strategies, accelerating the alignment between advertising content and user behavior while simultaneously improving system processing capacity and overall ad effectiveness. The deployment of such specialized hardware has significantly increased computational efficiency and strengthened the system's capability to handle large-scale data processing.

3.2. Application of Distributed Computing Framework

The traditional standalone computing architecture often encounters bottlenecks in both computing power and storage capacity, making it increasingly difficult to process large-scale datasets efficiently. In contrast, distributed computing systems divide computational tasks across multiple servers to perform collaborative processing, significantly accelerating computation and effectively addressing challenges in data storage and transmission. By leveraging this approach, advertising systems can complete machine learning tasks more rapidly in scenarios involving massive data, improving the accuracy, real-time responsiveness, and adaptability of ad delivery, performance evaluation, and campaign optimization [9].

For instance, an advertising system may need to process a dataset containing one billion user behavior records. Using distributed computing frameworks such as Apache Spark, this dataset can be partitioned into multiple subsets, each processed in parallel across different computing nodes. The training of an advertising recommendation model

can then be formulated as a loss function minimization problem, where the objective is to find a set of model parameters θthat minimizes the loss function:

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N} \ell(yi, f(xi;\theta)) \quad (1)$$

In formula (1), N represents the size of the dataset, y_i denotes the true label of the i-th sample, x_i is the feature vector of the sample, ℓ is the loss function, $f(x_i;\theta)$ is the prediction function, and θrepresents the model parameters to be learned. The distributed computing architecture significantly enhances the real-time processing capabilities of advertising systems, accelerates the training and iteration speed of machine learning algorithms, and provides critical technical support for improving overall advertising performance [10].

With improved computational efficiency, advertising platforms can rapidly extract key insights from massive volumes of data, enabling more precise targeting and strategic decision-making. This capability allows platforms to gain a competitive advantage in highly dynamic and data-intensive market environments.

3.3. Refinement of Model Training and Optimization

In the optimization of machine learning infrastructure, precise training and tuning of models play a decisive role in enhancing the operational performance and effectiveness of advertising systems. The training and optimization process requires processing vast amounts of data, and the careful adjustment of model parameters directly impacts the accuracy and effectiveness of advertising delivery. With the increasing volume and complexity of advertising data, relying solely on conventional model training methods is no longer sufficient to meet modern demands. Consequently, precise model training and optimization have become particularly critical.

Effective model training involves the selection of appropriate algorithms and optimization strategies to improve both accuracy and convergence speed. Adjusting advanced parameters during the training process is essential to ensure the reliability and robustness of the model. Traditional methods often depend on manual parameter tuning, which, although effective in some cases, proves inefficient for large-scale models and massive datasets, and is prone to being trapped in local optima.

In response, many advertising systems have adopted automated parameter optimization techniques, such as Bayesian optimization and exhaustive search strategies. These intelligent approaches enhance the efficiency of parameter adjustment, reduce human intervention, and allow the model to better adapt to complex, large-scale advertising data, ultimately improving the accuracy and effectiveness of ad placement and recommendation.

3.4. Improvement of Operation and Maintenance Automation

Faced with the rapid growth of advertising data and the increasing complexity of system architecture, traditional manual approaches to operation and maintenance have become inefficient and insufficient to meet the requirements for real-time response and high reliability. By leveraging automation tools and developing relevant scripts, automated operation and maintenance systems can perform comprehensive system monitoring, rapid fault detection, efficient resource allocation, and intelligent log analysis. These capabilities improve operational efficiency, reduce human errors, and ensure the stable and reliable functioning of the advertising system.

For example, an advanced advertising platform has implemented an automated operation and maintenance system to establish a real-time monitoring and control mechanism for its ad publishing infrastructure. The system features automated fault detection and recovery: when delays in advertising requests are identified, the platform automatically switches to a backup server and optimizes load distribution, ensuring uninterrupted ad delivery. Additionally, automated log processing tools collect real-time operational data, enabling the platform to identify and address system performance shortcomings efficiently.

As shown in Table 1, the implementation of operational automation has led to measurable improvements in the performance of the advertising delivery system, highlighting the effectiveness of automated monitoring and maintenance in enhancing system stability and responsiveness.

Table 1. Changes in Performance of Advertising Delivery System before and after Implementation of Platform Operation and Maintenance Automation.

index	Before automation	After automation	Increase amplitude
Average response time	300 milliseconds	150 milliseconds	50% reduction
System failure recovery time	45 minutes	5 minutes	89% reduction
Success rate of advertising placement	97%	99.5%	2.5% increase
Number of manual interventions	30 times/week	5 times/week	83% reduction

With the implementation of automated operation and maintenance management, the efficiency of fault repair and the accuracy of advertising placement on the platform have been significantly improved, while the reliance on manual operations has been greatly reduced. This ensures the continuity and stability of the system. The adoption of automated operation and maintenance mechanisms has enhanced the overall performance metrics of advertising platforms, generated operational cost savings, and created a more reliable and efficient environment for advertisers to publish and manage their campaigns.

4. Practical Application of Machine Learning in Advertising Systems

4.1. User Personalized Advertising Recommendations

In advertising delivery systems, personalized promotion relies on comprehensive analyses of users' past behavior, individual preferences, and social network information to create tailored advertising content for each user. This approach enhances advertisement click-through rates and increases user engagement. By deeply mining user interaction data, the system can uncover potential interests and needs, delivering targeted advertisements that align closely with users' behaviors or preferences. Personalized advertising not only improves the precision of ad placement but also optimizes user experience, reduces the intrusiveness of traditional advertising, and helps advertisers achieve higher returns on investment.

In the implementation of personalized advertising recommendations, widely used machine learning algorithms include matrix factorization techniques, such as singular value decomposition (SVD), and deep learning architectures, such as neural collaborative filtering. In collaborative filtering, matrix factorization decomposes the user-advertisement interaction matrix into two lower-dimensional matrices, reducing data dimensionality while extracting deep-level feature representations of both users and advertisements. The process can be formally expressed as:

$$R \approx U \cdot \mathrm{V}^{T} \tag{2}$$

In formula (2), RRR represents the user-advertisement rating matrix, UUU is the user feature matrix, and V is the advertisement feature matrix. By optimizing the matrix decomposition process, the system can more accurately predict individual users' preferences for specific advertisements, thereby enhancing the precision and effectiveness of personalized ad delivery.

4.2. Intelligent Prediction of Advertising Effectiveness

The application of intelligent prediction in advertising placement highlights the pivotal role of machine learning technology in the advertising industry. In contrast, traditional advertising methods primarily rely on past experience to establish basic rules, which limits their ability to accurately forecast advertising effectiveness, particularly in

the face of market fluctuations and diverse user behaviors. By leveraging machine learning algorithms to conduct in-depth analyses of large-scale historical advertising data, complex predictive models can be constructed to anticipate the responses of different advertisements across various user segments. This intelligent prediction approach enables advertisers to allocate budgets more effectively, enhance the efficiency of ad placement, and dynamically adjust strategies during campaigns to maximize advertising returns.

For example, in promoting smartphone advertisements, a platform may use machine learning models to estimate ad effectiveness by considering factors such as user interest classification, consumption history, and advertisement content. Regression models are then employed to predict the potential click-through rates of advertisements. The platform can represent this predictive process using the following linear regression model:

$$\widehat{CTR}=\beta 0+\beta 1x1+\beta 2x2+...+\beta nxn \quad (3)$$

In formula (3), $\widehat{CTR}$ is the predicted click through rate, x1, x2,..., xn are the feature variables that affect click through rate, such as user interest, ad type, exposure times, etc., β0, β1,..., βn are the parameters of the model. By training on historical advertising data, the platform is able to predict the click through rate of each advertisement and adjust the frequency and duration of ad display accordingly, optimizing advertising strategies. After optimization, the platform's ad click through rate has increased by 15%, and the return on advertising expenditure (ROAS) of advertisers has also been improved.

4.3. Construction of Anti Fraud and Risk Prevention System

The construction of an anti-fraud and risk prevention system is a critical component in ensuring the healthy operation of advertising platforms. Such a system must be capable of accurately identifying risks and preventing fraudulent activities, including fake clicks, malicious registrations, and false advertising. With the rapid evolution of advertising platforms, fraudulent methods continue to advance, rendering traditional rule-based detection and screening mechanisms insufficient to meet emerging challenges. Leveraging machine learning technology to develop intelligent fraud detection systems has therefore become essential for effective risk management. By analyzing historical data with machine learning algorithms, these models can autonomously detect abnormal behaviors and enhance the platform's preventive capabilities.

For instance, when an advertising platform encounters malicious click attacks, a decision tree algorithm can assess whether user behavior is legitimate by examining parameters such as IP address, operating system, and click frequency. Once suspicious behavior is detected, the system can automatically initiate response strategies, including restricting ad displays or blocking malicious users. This approach effectively suppresses fraudulent activity and protects the interests of both advertisers and the platform.

4.4. Intelligent Allocation and Dynamic Adjustment of Advertising Resources

In advertising systems, achieving intelligent allocation and dynamic management of advertising resources is essential for enhancing ad effectiveness and improving resource utilization efficiency. These resources include the placement of advertisements as well as the comprehensive management of budgets, time slots, and target audiences. Traditionally, most resource allocation methods relied on pre-set rules or manual intervention. However, with the continuous expansion of advertising platforms and the increasing complexity of demand, these conventional approaches are no longer sufficient to meet requirements for immediacy and accuracy. Leveraging machine learning technology for intelligent allocation and dynamic adjustment of advertising resources can significantly improve system efficiency and maximize advertisers' returns.

For example, some advertising platforms employ deep reinforcement learning techniques to optimize resource allocation. The system uses real-time data to perform forward-looking analyses of ad responses among specific audiences. After each ad display,

the system collects and evaluates feedback, such as click-through rate metrics. If an advertisement underperforms during a given time period, the system automatically reduces its exposure frequency and reallocates budget to higher-performing ads, thereby improving resource utilization efficiency. By adopting this strategy, advertisers can lower advertising costs while simultaneously enhancing ad conversion efficiency.

5. Conclusion

This article provides an in-depth analysis of the widespread application and effectiveness of machine learning infrastructure upgrades in the advertising industry. By enhancing hardware capabilities, refining computing architectures, innovating model training techniques, and optimizing operation and maintenance processes, advertising systems have achieved substantial improvements in computational efficiency, targeting accuracy, and intelligent management. These advancements are particularly evident in key areas such as personalized advertising recommendations, real-time performance estimation, and fraud detection, where machine learning has significantly accelerated the transformation of advertising operations toward greater intelligence.

The integration of machine learning enables advertising platforms to analyze massive datasets, identify nuanced user preferences, predict advertising performance, and dynamically adjust campaigns, which collectively enhances the precision and return on investment for advertisers. Moreover, the adoption of automated operation and maintenance frameworks ensures system stability and operational efficiency, further supporting the scalability and robustness of advertising infrastructure.

Despite the considerable progress made in optimization techniques, continuous development remains critical. As data volumes expand and user demands evolve, advertising platforms must sustain innovations in hardware acceleration, distributed computing, adaptive model training, and intelligent automation. In the future, the convergence of advanced machine learning methods with emerging technologies, such as real-time analytics, reinforcement learning, and multimodal data integration, is expected to propel the advertising industry toward higher efficiency, deeper personalization, and greater intelligence. This ongoing evolution will not only enhance the user experience and advertising effectiveness but also drive the holistic upgrade of the advertising ecosystem, fostering a more sustainable, data-driven, and competitive advertising environment.

References

1. R. Shrivastava, D. S. Sisodia, and N. K. Nagwani, "Deep ensembled multi-criteria recommendation system for enhancing and personalizing the user experience on e-commerce platforms," *Knowledge and Information Systems*, vol. 66, no. 12, pp. 7799-7836, 2024. doi: 10.1007/s10115-024-02187-3
2. C. M. Gangani, "Role of Machine Learning in Optimizing IT Infrastructure," *Kuwait Journal of Information Technology and Decision Sciences*, vol. 1, pp. 12-22, 2023.
3. A. Noorian, "Integrating user reviews and risk factors from social networks in a multi-objective recommender system," *Electronic Commerce Research*, pp. 1-43, 2024. doi: 10.1007/s10660-024-09944-0
4. A. L. Garrido, M. S. Pera, and C. Bobed, "SJORS: A Semantic Recommender System for Journalists," *Business & Information Systems Engineering*, vol. 66, no. 6, pp. 691-708, 2024.
5. A. Aramanda, S. Md Abdul, and R. Vedala, "Emotions in recommender systems for discrepant-users," *Knowledge and Information Systems*, vol. 67, no. 1, pp. 953-976, 2025.
6. Q. Guo, F. Zhuang, C. Qin, H. Zhu, X. Xie, H. Xiong, and Q. He, "A survey on knowledge graph-based recommender systems," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 8, pp. 3549-3568, 2020.
7. C. Zhao, X. Su, M. He, H. Zhao, J. Fan, and X. Li, "Collaborative knowledge fusion: A novel approach for multi-task recommender systems via llms," *arXiv preprint arXiv:2410.20642*, 2024.
8. J. Hu, J. Gao, X. Zhao, Y. Hu, Y. Liang, Y. Wang, and H. Yin, "BiVRec: Bidirectional View-based Multimodal Sequential Recommendation," *arXiv preprint arXiv:2402.17334*, 2024.
9. K. Zou, A. Sun, X. Jiang, Y. Ji, H. Zhang, J. Wang, and R. Guo, "Hesitation and Tolerance in Recommender Systems," *arXiv preprint arXiv:2412.09950*, 2024.

10. F. Zhu, Y. Wang, C. Chen, G. Liu, M. Orgun, and J. Wu, "A deep framework for cross-domain and cross-system recommendations," *arXiv preprint arXiv:2009.06215*, 2020.