Transactions on Computational and Scientific Methods | Vo. 4, No. 4, 2024
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Hierarchical Attention-Based Modeling for Intelligent Scheduling Delay Prediction in Complex Backend Systems

Ziyu Cheng
University of Southern California, Los Angeles, USA
chengzy1115@gmail.com

Abstract: This paper addresses the problem of backend scheduling delay prediction and proposes a modeling method based on hierarchical attention to overcome the limitations of traditional approaches under high-dimensional features, complex dependencies, and dynamic environments. The study begins with multidimensional features such as task load, resource usage, and component invocation relationships, and generates low-level representations through linear embedding and nonlinear mapping. Local attention is then introduced to model short-term dependencies and local features, while global attention and hierarchical aggregation structures capture cross-component and cross-temporal dependencies, achieving a balance between fine-grained local modeling and overall dependency modeling. A prediction layer is further constructed, where local and global representations are fused to generate delay predictions. To verify the effectiveness of the method, experiments were conducted on a public dataset with Transformer, FedFormer, iTransformer, and TimeMixer as baselines. Results show that the proposed method outperforms these models in key metrics, including MSE, MAE, RMSE, and MAPE. In addition, extensive sensitivity analyses were carried out, covering parameters and environmental factors such as learning rate, weight decay, hidden dimension size, node scale, and missing rate. The results demonstrate that the proposed method exhibits strong robustness and stability under diverse conditions. Overall, this study improves both prediction accuracy and generalization performance in backend scheduling delay prediction and provides strong support for efficient scheduling in complex systems.

Keywords: Hierarchical attention mechanism; scheduling delay prediction; sensitivity analysis; complex system modeling

1. Introduction

In the context of rapid advances in modern cloud computing and distributed systems, the issues of backend service scheduling and resource allocation have become increasingly prominent. With the continuous expansion of business scale and the rapid growth of user demand, backend components are required to handle complex and dynamic task flows under large-scale concurrency. Scheduling delay, as a key metric of system performance and user experience, directly determines service response speed and overall system stability. However, traditional scheduling optimization methods often rely on fixed rules or static priority mechanisms, which struggle to cope with dynamic workloads and multidimensional resource constraints[1]. Therefore, how to effectively predict scheduling delay under multi-source requests and limited resources has become an important topic for improving system efficiency and user satisfaction.

In this context, the accuracy of delay prediction not only affects resource utilization but also determines the effectiveness of service elasticity and fault-handling mechanisms. Traditional prediction approaches often

rely on shallow features or statistical patterns. Yet, when facing complex task dependencies and multi-level resource competition, they easily overlook interactions and dynamic evolutions across different levels of features. In microservice architectures, the call chains among components are highly complex. Delays are rarely caused by a single node, but by multi-level dependencies and concurrent conflicts. This hierarchical association means single-level modeling cannot capture the global dynamics, leading to degraded prediction performance. Thus, it is urgent to adopt mechanisms that can model hierarchical relationships and multidimensional dependencies to better describe the dynamic characteristics of scheduling delay[2].

In recent years, advances in deep learning and attention mechanisms have provided new possibilities for modeling complex patterns. Attention mechanisms can highlight key information and suppress redundant features, giving models stronger feature selection and representation ability when handling high-dimensional data[3]. However, single-level attention is still insufficient in scheduling scenarios. Scheduling delay is influenced not only by single-dimension features but also by multi-level interactions across tasks, components, and resources. Introducing hierarchical attention mechanisms allows the system to simulate both bottom-up and top-down information flows. This design captures local details while preserving global dependencies. Such mechanisms can help overcome the limitations of traditional methods in prediction accuracy and robustness[4].

From an application perspective, accurate scheduling delay prediction is of great significance for cloud platforms and backend systems. First, it enables systems to estimate potential delays before scheduling tasks, allowing proactive load balancing and optimized resource allocation, thereby reducing performance degradation caused by local congestion. Second, in terms of service quality, delay prediction provides the basis for elastic scaling and fault recovery, helping systems take preventive measures before anomalies occur. Third, in cost control, effective prediction improves resource utilization and avoids waste from over-provisioning. These benefits go beyond technical performance improvements and extend to better user experience and stronger competitiveness for service providers.

In summary, applying hierarchical attention mechanisms to scheduling delay prediction in backend components is both a natural trend in technology development and a practical need under complex workloads. By jointly modeling multi-level features, it becomes possible to reveal the intrinsic mechanisms behind delays more comprehensively[5]. This supports intelligent scheduling and dynamic optimization in backend systems. The research in this direction has both strong academic value and broad practical potential, laying the foundation for more efficient and reliable distributed system operations in the future.

2. Related work

In existing research, backend component scheduling and delay prediction have always been important directions for system performance optimization. Early work mainly focused on rule-based scheduling strategies and statistical modeling methods. These approaches usually relied on fixed priorities, queue lengths, and average load levels for scheduling decisions. Delay estimation was often carried out through traditional time series models or queuing theory[6,7]. Such methods were effective in small-scale systems or low-complexity environments. However, they often failed to capture complex nonlinear relationships under highly dynamic workloads and multidimensional resource competition. This led to limited prediction accuracy and adaptability. With the expansion of cloud platforms and the diversification of business demands, methods based solely on static rules or simple statistical patterns have shown clear limitations.

With the development of machine learning, researchers began to explore more advanced modeling methods to improve prediction accuracy. Traditional multilayer perceptrons and convolutional neural networks were introduced into scheduling scenarios to capture the mapping between task features, resource indicators, and delay[8]. These models broke through the constraints of linear assumptions and provided more flexible fitting ability with high-dimensional inputs. However, they were less effective in handling strong sequential and dependency patterns. Scheduling delay is often influenced by task dependencies, component invocation paths, and resource competition dynamics. Static feature learning alone cannot fully capture this complexity. To

address this, some studies attempted to use recurrent neural networks and deep time series models to represent system state evolution over time. Yet, challenges remain in modeling long sequences and multi-level feature interactions[9,10].

In recent years, the introduction of attention mechanisms has created new opportunities for delay prediction. Attention-based models can highlight key factors across both feature and temporal dimensions. This allows them to enhance feature selection in high-dimensional and complex environments. In backend scheduling scenarios, such methods capture non-uniform dependencies between tasks and resources, enabling models to adjust their focus dynamically. Compared with fixed windows or uniform weighting, attention mechanisms show greater flexibility and expressiveness[11]. However, most current work still emphasizes single-level feature modeling. This makes it difficult to balance local details and global dependencies. As a result, models may still face prediction bias and limited generalization in complex microservice architectures and multi-component interactions.

To further enhance prediction, research has turned to hierarchical modeling. Hierarchical attention mechanisms can emphasize critical features at the local level while integrating cross-component and cross-task dependencies at higher levels. Through multi-level structures, they progressively extract and fuse features across different scales. This enables models to capture local delay fluctuations while understanding global system dynamics. Current studies in this direction show promising potential for delivering fine-grained delay prediction in complex backend environments[12]. However, how to balance prediction accuracy with computational efficiency and scalability remains an important open problem that requires deeper investigation.

3. Proposed Approach

In this study, the core of the model design is to model the scheduling delay of backend components through a hierarchical attention mechanism. The model architecture is shown in Figure 1.

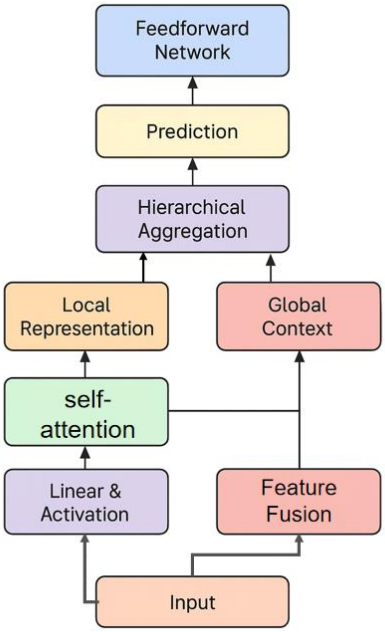

Figure 1. Overall model architecture

First, the input features are composed of multi-dimensional system indicators, including task load, resource usage, and component call dependencies. Let the input sequence be a vector set:

$$X = \{x_1, x_2, ..., x_T\}$$

Where T represents the number of time steps and d represents the feature dimension of a single moment. To extract the underlying features, the input is first represented by a linear transformation and a nonlinear activation function:

$$h_t = \sigma(Wx_t + b)$$

Where $\sigma(\cdot)$ represent activation functions, and d_t is the hidden layer dimension.

At the local level, the self-attention mechanism is used to model the dependencies within the sequence. For each layer of features, a query, key, and value matrix is constructed:

$$Q = HW_Q, K = HW_K, V = HW_V$$

Where $H = [h_1, h_2, ..., h_T]$ and $W_Q, W_K, W_V \in R^{d_h \times d_a}$. Based on this, the local attention weight can be expressed as:

$$a_{ij} = \frac{\exp(\frac{Q_i K_j^T}{\sqrt{d_a}})}{\sum_{k=1}^{T} \exp(\frac{Q_i K_j^T}{\sqrt{d_a}})}$$

The final local representation is:

$$z_i = \sum_{i=1}^{T} a_{ij} V_j$$

This process can highlight key dependencies in time series data and capture short-term delay factors.

At the global level, a hierarchical aggregation mechanism is introduced to fuse features of different time granularities and component levels. Specifically, based on the local representation z_i, weighted aggregation is performed through global attention to obtain the global context representation:

$$g = \sum_{i=1}^{T} \beta_i z_i, \quad \beta_i = \frac{\exp(u^T z_i)}{\sum_{j=1}^{T} \exp(u^T z_j)}$$

Where $u \in R^{d_a}$ is the global context vector. This mechanism enables the model to take into account both global dependencies between components and long-term trends across time steps. Finally, the local representation is concatenated with the global representation and input to the prediction layer:

$$\hat{y} = f([z_1, z_2, ..., z_T, g])$$

$f(\cdot)$ is a feedforward neural network that outputs the predicted value of scheduling delay. Through this hierarchical attention design, the model can accurately model scheduling delay at multiple granularities and dependency levels, thus having stronger expressiveness and adaptability in complex backend scenarios.

4. Experiment result

4.1 Dataset

The dataset used in this study comes from Alibaba Cluster Trace 2018. It consists of real scheduling and operation logs collected from a large-scale cloud computing platform. It covers extensive information about task scheduling and execution in distributed environments. The dataset includes multiple resource usage indicators, such as CPU, memory, disk, and network. It also provides job submission time, execution time, and task dependency relationships. This rich information offers a solid foundation for modeling and analyzing backend scheduling delay. It allows research to be carried out under conditions close to real production environments.

The dataset is large in scale. It contains tens of thousands of jobs and millions of task entries, spanning several days. The data is mainly presented in the form of logs and tables, including both task-level and job-level information. For example, the task logs record precise scheduling start and finish times, dynamic changes in resource consumption, and call chains among tasks within a job. These fine-grained time series data enable detailed analysis of the distribution of scheduling delay and the underlying factors that influence delay.

This dataset has strong representativeness and practical value. It reflects real task characteristics and scheduling patterns in production environments. Researchers can use it to develop delay prediction and optimization methods with practical significance and transferability. In addition, its rich multidimensional features provide strong support for building complex models. They help explore how different features contribute to delay prediction and further promote the intelligent development of backend scheduling systems.

4.2 Experimental Results

This paper first conducts a comparative experiment, and the experimental results are shown in Table 1.

Table 1: Comparative experimental results

Method	MSE	MAE	RMSE	MAPE
Transformer[13]	0.324	0.388	0.569	6.82%
FedFormer[14]	0.301	0.365	0.548	6.41%
ITransformer[15]	0.279	0.349	0.528	6.07%
TimeMixer[16]	0.265	0.336	0.515	5.69%
Ours	0.241	0.315	0.491	5.52%

From Table 1, it can be observed that the traditional Transformer performs relatively poorly in backend scheduling delay prediction. Its MSE, MAE, and RMSE are all at the highest levels, and its MAPE reaches 6.82%. This result indicates that although the Transformer has strong sequence modeling ability, it is still insufficient in capturing features under complex task dependencies and multidimensional resource constraints. It struggles to represent the multi-level dynamic patterns of delay with precision.

Compared with the Transformer, FedFormer, iTransformer, and TimeMixer all show improvements to varying degrees. FedFormer demonstrates better performance in capturing global features, with significant reductions in MSE and RMSE, highlighting its advantage in long sequence processing. iTransformer further enhances adaptability to input features through structural improvements, resulting in reduced MAE and

MAPE errors. TimeMixer shows performance closest to Ours, suggesting that fusion-based modeling has strong potential in handling complex delay prediction problems. It effectively alleviates the shortcomings of traditional models in neglecting multi-granularity dependencies.

Among all methods, the proposed model Ours achieves the best results across all four metrics. In particular, it shows significant reductions in MSE and MAE compared with TimeMixer. This indicates that the hierarchical attention mechanism can better integrate local and global features. It provides more discriminative and stable feature representations in scenarios with multi-level task dependencies and resource competition, leading to more accurate delay prediction. Its strong performance in MAPE also shows good generalization under different workload fluctuations.

Overall, the experimental results confirm the effectiveness and necessity of the hierarchical attention mechanism for backend scheduling delay prediction. By capturing detailed features at the local level and integrating global dependencies at the system level, the method overcomes the limitations of existing models in feature representation and dependency modeling. It significantly improves prediction accuracy. These findings not only demonstrate the soundness of the method design but also provide strong support for intelligent and efficient backend scheduling systems.

This paper also presents an experiment on the sensitivity of hidden dimension size to MAPE, and the experimental results are shown in Figure 2.

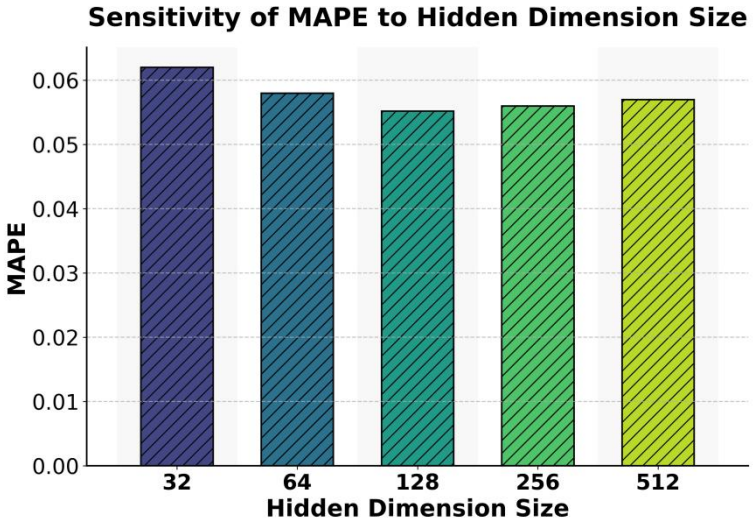

Figure 2. Experiments on the sensitivity of hidden dimension size to MAPE

From Figure 2, it can be seen that different hidden dimensions lead to variations in MAPE performance. When the hidden dimension is small, such as 32, the MAPE value is significantly higher. This indicates that in a low-dimensional space, the model cannot fully capture the complex feature interactions in backend scheduling delay. As a result, prediction bias increases. This shows that too small a feature space is insufficient to support the expressive power of the hierarchical attention mechanism.

As the hidden dimension increases, the MAPE value gradually decreases, reaching its lowest point at 128 dimensions. This suggests that at a moderate scale, the model can better integrate local and global features. It avoids missing information while not introducing excessive redundancy, which improves both stability and accuracy in prediction. This trend highlights the advantage of hierarchical attention in multi-granularity feature modeling.

When the hidden dimension continues to increase to 256 and 512, MAPE shows a slight rise. This suggests that overly large dimensions may introduce redundant features or noise, which makes optimization in high-dimensional space more difficult. Such a situation can reduce generalization performance and interfere with

delay prediction under complex workloads. Therefore, the choice of hidden dimension requires a balance between expressive capacity and model complexity.

Overall, the experiment shows that hidden dimension size has a significant impact on the performance of backend scheduling delay prediction. An appropriate dimension can effectively enhance the model's performance under hierarchical feature modeling. This sensitivity analysis provides useful guidance for model design. It helps determine optimal parameter configurations that balance prediction accuracy and computational efficiency, thereby better meeting the dual requirements of real-time performance and reliability in backend systems.

This paper also presents an experiment on the sensitivity of the weight decay coefficient to MAE, and the experimental results are shown in Figure 3.

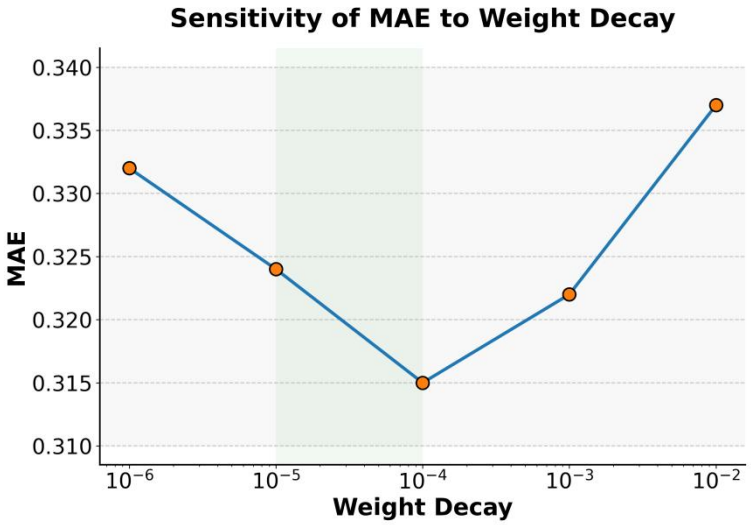

Figure 3. Sensitivity experiment of the weight decay coefficient to MAE

From Figure 3, it can be seen that the weight decay coefficient shows clear sensitivity in relation to MAE. When the weight decay is too small, such as 10^{-6} , MAE remains at a high level. This indicates that under weak regularization, the model is prone to overfitting. As a result, prediction in complex backend scheduling environments suffers from poor generalization and higher errors.

As the weight decay increases to 10^{-4} , MAE reaches its lowest point. This result shows that moderate regularization can effectively suppress overfitting while preserving the expressive capacity of the hierarchical attention mechanism. In this way, the model achieves a better balance in multidimensional feature modeling and obtains optimal performance in delay prediction.

When the weight decay is further increased to 10^{-3} and 10^{-2}, MAE begins to rise again. This suggests that strong regularization reduces the learning ability of the model. The attention mechanism cannot fully capture the complex relationships between local and global features. As a consequence, key information is lost in feature extraction, which reduces prediction accuracy.

Overall, the results show that a proper choice of weight decay is crucial for backend scheduling delay prediction. Too small a value leads to overfitting, while too large a value results in underfitting. Only within a moderate range can the hierarchical attention mechanism fully demonstrate its advantage. This significantly improves the stability and accuracy of prediction and provides reliable support for backend scheduling delay forecasting.

This paper gives the impact of the missing rate on the experimental results, and the experimental results are shown in Figure 4.

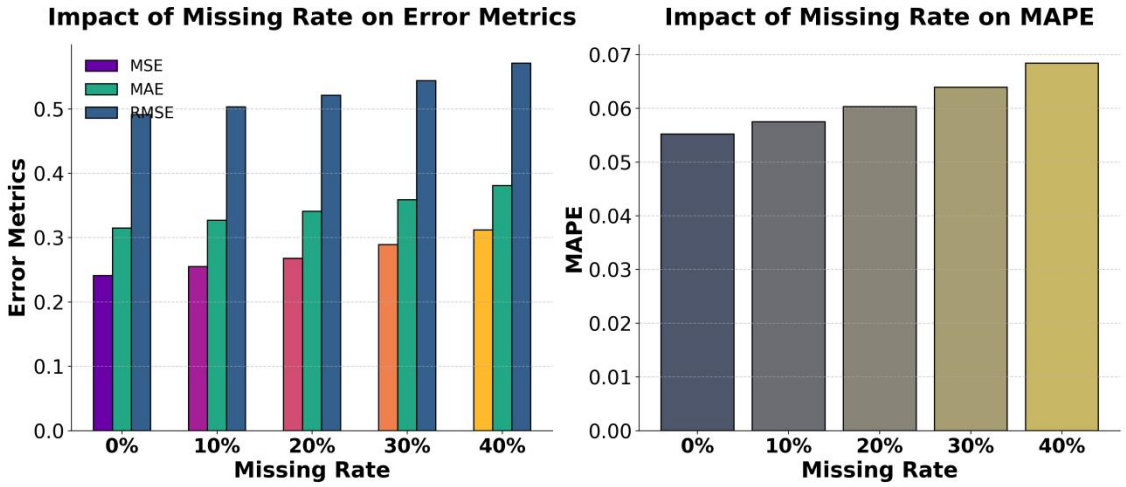

Figure 4. The impact of the missing rate on experimental results

From Figure 4, it can be seen that all error metrics show an upward trend as the missing rate increases. When the missing rate is 0%, the model can accurately capture the relationship between task load and resource usage. MSE, MAE, RMSE, and MAPE all remain at low levels. This indicates that under complete data conditions, the hierarchical attention mechanism can fully demonstrate its advantages and achieve high-precision delay prediction.

When the missing rate rises to 10% and 20%, the error metrics begin to increase, with MAE and MAPE showing more obvious changes. This suggests that missing data weakens the fine-grained modeling of local features, making it difficult to capture short-term dependencies completely. Since backend scheduling delay is often influenced by multiple factors, missing features can interfere with the aggregation process of the attention mechanism, thereby increasing prediction uncertainty.

When the missing rate reaches 30% or higher, the increase in error metrics becomes more significant. Both RMSE and MAPE show sharp rises. This indicates that with a high proportion of missing data, the model's ability to capture global features is severely weakened. The integration of local and global dependencies becomes difficult, and prediction bias is magnified. This highlights the importance of data completeness for hierarchical modeling and also reflects the limitations of the model in handling missing information.

Overall, the results show that the missing rate has a significant impact on backend scheduling delay prediction. At low missing rates, the model can still maintain stable performance. However, as the missing rate increases, the errors grow rapidly. This emphasizes the need for data preprocessing, missing value imputation, and robust modeling methods in practical scenarios. Such approaches can ensure that the hierarchical attention mechanism continues to perform well even in imperfect data environments, providing reliable support for backend systems.

This paper also gives the impact of node scale changes on experimental results, and the experimental results are shown in Figure 5.

From Figure 5, it can be seen that changes in node scale have a clear impact on model performance in delay prediction. Under small-scale node conditions, error metrics remain relatively high, especially MSE and MAE. This indicates that when resources are limited, the hierarchical attention mechanism can capture some features, but insufficient parallelism and increased scheduling complexity constrain the stability of prediction.

When the node scale expands to a moderate range, such as 200 nodes, error metrics decrease significantly, and MAPE reaches its lowest point. This shows that under moderate scale conditions, the system can better

utilize both local and global feature modeling. Task distribution is more balanced, and the attention mechanism can achieve more effective interaction and fusion among multidimensional features, leading to higher prediction accuracy.

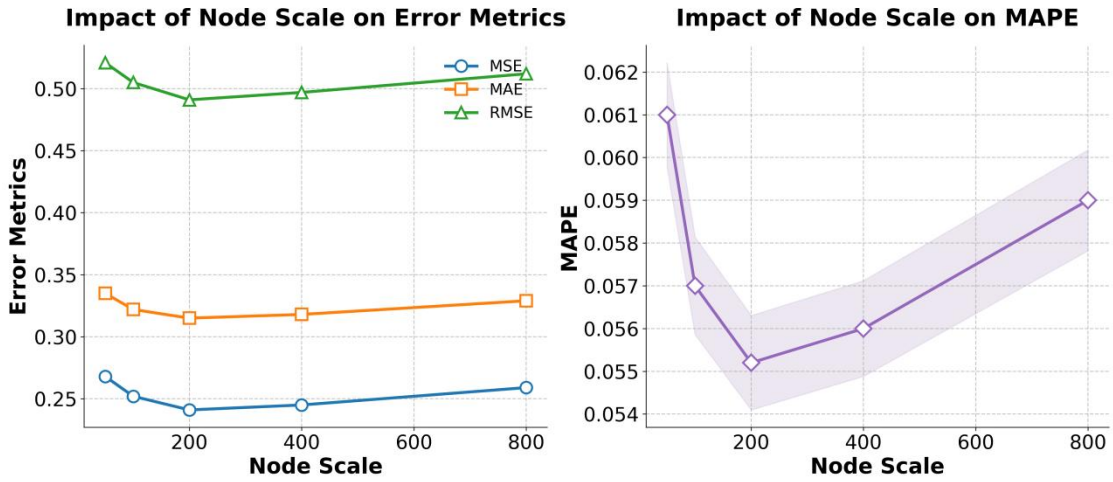

Figure 5. The impact of node scale changes on experimental results

As the node scale further increases to 400 and 800, the overall performance remains good, but error metrics show slight increases. This reflects that at very large scales, system heterogeneity and scheduling overhead rise. As a result, the complexity of feature modeling increases, and some local dependencies are weakened. The upward trend of MAPE indicates that large-scale deployment may increase fluctuations in prediction results.

Overall, the experiment reveals the dual role of node scale in backend scheduling delay prediction. Too small a scale leads to insufficient resources, while too large a scale introduces additional interference. Only within a moderate range can the hierarchical attention mechanism achieve the best effect in both local and global modeling. This conclusion provides important guidance for backend systems to plan node scale reasonably in practical deployment, ensuring both prediction stability and accuracy.

5. Conclusion

This study focuses on the problem of backend scheduling delay prediction and proposes a modeling method based on hierarchical attention. By extracting fine-grained features at the local level and aggregating global dependencies at the system level, the method can comprehensively capture the multiple factors that cause delay in complex task environments. Experimental results show that the method achieves superior performance across multiple error metrics. This further demonstrates its strong adaptability and robustness in high-dimensional and multi-constraint backend system environments. The research not only addresses the limitations of traditional methods in prediction accuracy but also enriches performance modeling approaches for complex distributed systems.

From an application perspective, the findings of this study provide new insights for intelligent scheduling in backend systems. With the introduction of hierarchical attention, systems can predict potential delays more accurately before scheduling. This enables better decisions in load balancing, resource allocation, and fault recovery. Such capabilities are particularly valuable in large-scale cloud platforms, microservice architectures, and data centers. They can directly improve service quality, reduce energy consumption and operational costs, and enhance stability and scalability under conditions of high concurrency and complex dependencies.

In addition, this research provides useful references for general time series prediction and complex system modeling. The advantage of hierarchical attention in balancing local and global features gives it potential for cross-domain applications. It can be applied not only to backend delay prediction but also to scenarios such

as financial risk monitoring, traffic congestion prediction, and industrial production scheduling, all of which involve multidimensional dependencies. This cross-domain applicability further highlights the theoretical and practical value of the proposed method and offers guidance for intelligent decision-making in complex systems.

Looking ahead, there is still room for further development. On the one hand, the model can be extended by incorporating more advanced environment modeling and adaptive mechanisms, making it more flexible under dynamic workloads and resource states. On the other hand, integration with techniques such as data repair and anomaly detection can enhance robustness when dealing with missing values, noise, and abnormal conditions. Moreover, combining the method with frontier approaches such as reinforcement learning and graph neural networks could enable dynamic optimization of multi-level dependencies. Through these directions, future systems will not only achieve more accurate delay prediction but also advance toward greater efficiency, intelligence, and autonomy in backend scheduling.

References

[1] Xiong J, Zhang Y. A unifying framework of attention-based neural load forecasting[J]. IEEE Access, 2023, 11: 51606-51616.

[2] Shao Z, Wang F, Zhang Z, et al. Hutformer: Hierarchical u-net transformer for long-term traffic forecasting[J]. arXiv preprint arXiv:2307.14596, 2023.

[3] Zhang T. Network level spatial temporal traffic forecasting with Hierarchical-Attention-LSTM[J]. Digital Transportation and Safety, 2024, 3(4): 233-245.

[4] Lian Q, Sun W, Dong W. Hierarchical Spatial-Temporal Neural Network with Attention Mechanism for Traffic Flow Forecasting[J]. Applied Sciences, 2023, 13(17): 9729.

[5] Wang H, Zhang Z. Hierarchical time series forecasting based on temporal convolution and attention mechanism[C]//International Conference on AI Logic and Applications. Singapore: Springer Nature Singapore, 2023: 403-410.

[6] H. Bi, L. Lu and Y. Meng, "Hierarchical attention network for multivariate time series long-term forecasting," Applied Intelligence, vol. 53, no. 5, pp. 5060-5071, 2023.

[7] Li Z L, Yu J, Zhang X L, et al. A Multi-Hierarchical attention-based prediction method on Time Series with spatio-temporal context among variables[J]. Physica A: Statistical Mechanics and its Applications, 2022, 602: 127664.

[8] Cai J, Liu W, Huang Z, et al. Task decomposition and hierarchical scheduling for collaborative cloud-edge-end computing[J]. IEEE Transactions on Services Computing, 2024, 17(6): 4368-4382.

[9] Goel, A., Tung, C., Hu, X., Thiruvathukal, G. K., Davis, J. C., & Lu, Y. H. (2022, January). Efficient computer vision on edge devices with pipeline-parallel hierarchical neural networks. In 2022 27th Asia and South Pacific Design Automation Conference (ASP-DAC) (pp. 532-537). IEEE.

[10] Bi H, Lu L, Meng Y. Hierarchical attention network for multivariate time series long-term forecasting[J]. Applied Intelligence, 2023, 53(5): 5060-5071.

[11] K. Gavrilyuk, R. Sanford, M. Javan and C. G. Snoek, "Actor-transformers for group activity recognition," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 839-848, 2020.

[12] Xiong J, Zhou P, Chen A, et al. Attention-based neural load forecasting: A dynamic feature selection approach[C]//2021 IEEE Power & Energy Society General Meeting (PESGM). IEEE, 2021: 01-05.

[13] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[14] Zhou T, Ma Z, Wen Q, et al. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting[C]//International conference on machine learning. PMLR, 2022: 27268-27286.

[15] Liu Y, Hu T, Zhang H, et al. itransformer: Inverted transformers are effective for time series forecasting[J]. arXiv preprint arXiv:2310.06625, 2023.

[16] Wang S, Wu H, Shi X, et al. Timemixer: Decomposable multiscale mixing for time series forecasting[J]. arXiv preprint arXiv:2405.14616, 2024.