Transactions on Computational and Scientific Methods | Vol. 4, No. 07, 2024
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Long Text Classification with Large Language Models via Dynamic Memory and Compression Mechanisms

Yaxuan Luan
University of Southern California, Los Angeles, USA
yaxuanlu@alumni.usc.edu

Abstract: In recent years, long text classification has posed significant challenges due to semantic redundancy, input length limitations, and difficulties in capturing global dependencies. This paper proposes a novel framework that integrates dynamic memory and compression mechanisms into large language models to address these issues. The approach introduces a dynamic memory unit that selectively stores and updates essential information during training, while a low-rank compression module reduces redundancy and computational overhead without sacrificing semantic integrity. To verify the effectiveness of the proposed method, multiple experiments were conducted, including comparative evaluations, hyperparameter sensitivity analysis, environment sensitivity analysis, and data sensitivity analysis. The results demonstrate that the model achieves superior performance compared to existing baselines, highlighting its adaptability and robustness under different conditions. In particular, the framework effectively balances local and global semantic representations in long texts, ensuring both efficiency and accuracy. Moreover, the analysis of influencing factors such as learning rate, hidden dimension size, sample scale, and noise ratio provides systematic insights into the internal behavior of the model. These findings confirm that the proposed design not only improves classification performance but also enhances stability when handling large-scale and complex text data, thereby offering a reliable solution for long text classification tasks.

Keywords: dynamic memory, compression mechanism, long text classification, sensitivity analysis

1. Introduction

In the era of information explosion, text data is produced and accumulated at an unprecedented speed. Long texts, as an important part of this data, are widely present in news reports, academic papers, legal documents, medical records, and corporate archives[1]. Compared with short texts, long texts are not only more extensive in length but also more complex in content organization and semantic expression. They often involve multi-level, multi-topic, and cross-paragraph semantic relations. Effectively classifying long texts and accurately capturing their semantic and structural information has become an important direction in natural language processing. In the context of big data, the ability to extract key information from long texts quickly and accurately is indispensable for public decision-making, financial risk control, opinion analysis, and knowledge retrieval[2].

With the rapid development of large language models, text classification tasks have made remarkable progress across many fields. Large language models, through pretraining on massive data, possess strong abilities in semantic representation and reasoning. They perform well in short and medium-length text classification. However, when dealing with long texts, models often face challenges such as input length limits, high memory usage, and low computational efficiency. Traditional solutions usually rely on truncation,

sliding windows, or segment concatenation. These methods often disrupt semantic integrity, leading to loss of global information and incoherent context understanding, which harms classification performance. Therefore, the key problem is how to maintain computational feasibility while preserving the global semantics and essential details of long texts[3].

The complexity of long text classification lies not only in length but also in the distribution of semantic information. Critical information may be scattered across different paragraphs, and nonlinear relations often exist among these pieces. Local context alone is insufficient to capture them comprehensively[4]. Relying only on the raw attention mechanism to process long texts brings heavy computational costs and risks dispersing attention, making it hard for the model to focus on truly important semantic fragments. A mechanism is needed that can compress irrelevant information while efficiently storing and exploiting key semantics, so that the model can remain both efficient and accurate in a long-text environment. This need has driven research on dynamic memory and compression mechanisms. Such mechanisms allow the model, like human readers, to filter out unnecessary details while retaining and reinforcing core information.

Against this background, dynamic memory mechanisms show unique advantages. By introducing dynamic memory units, the model can selectively store and update information at different levels and stages, thus maintaining a grasp of the overall structure of long texts. Compression mechanisms further help by reducing or simplifying large text sequences without losing essential information. This enables global modeling under limited computational resources. The dynamic cycle of storage, compression, and update opens a new path for long text classification. It not only alleviates information loss and inefficiency found in traditional methods but also provides stronger scalability and adaptability for the model[5].

In terms of research significance, exploring dynamic memory and compression mechanisms for large language models in long text classification is not only an important attempt at structural optimization but also a key step toward practical applications in natural language processing. In fields such as law, healthcare, and finance, where long texts are central, this research can greatly improve efficiency and accuracy in information processing, while reducing the cost and risk of manual review. At the same time, it provides the academic community with new ideas on how to balance input length, memory consumption, and inference speed, while preserving semantic understanding. This contributes to expanding the application boundaries of large models to more complex tasks and lays a foundation for future integrated research across modalities, languages, and tasks.

2. Related work

In the field of text classification, early methods relied mainly on traditional machine learning algorithms and handcrafted feature extraction. Common approaches included bag-of-words, TF-IDF, or n-gram representations. These methods achieved some success in short text tasks. However, they struggled to capture complex semantic dependencies and hierarchical structures in long texts. Handcrafted features were also highly dependent on specific datasets and could not adapt flexibly to different domains, which limited transferability and scalability. With the development of neural networks, models based on convolutional and recurrent structures emerged. They learned feature representations through end-to-end training and improved the performance of long text classification to some extent. Yet, they still showed weaknesses in modeling long-distance dependencies[6].

The rise of large-scale pretrained language models marked a new stage in text classification research. Through self-supervised learning on massive corpora, these models gained stronger semantic modeling and contextual understanding abilities. On short and medium-length texts, they often matched or even surpassed traditional methods. However, challenges remain when handling long texts. On one hand, most model architectures impose fixed input length limits, which make them unsuitable for very long sequences. On the other hand, global modeling of long texts results in high computational costs, making it difficult to balance efficiency and effectiveness. Preserving semantic integrity and contextual coherence under long text conditions has therefore become a central research question[7].

To overcome the input length limitation, researchers proposed various strategies. One approach splits long texts into segments using truncation or sliding windows. Each segment is modeled independently, and the results are later combined. While this alleviates input length constraints, it often sacrifices global semantic integrity and causes the loss of cross-paragraph information. Another approach compresses or summarizes the input text. Redundant information is filtered out, and only key content is preserved for modeling. This improves efficiency but introduces potential bias, which may reduce accuracy. Other studies have explored sparse attention or hierarchical modeling. These aim to reduce computational complexity while enhancing the ability to capture long-distance dependencies.

On this basis, dynamic memory and compression mechanisms have become new research focuses. Dynamic memory provides models with continuous storage and updating abilities. This allows them to selectively retain important information when processing long texts, rather than relying on direct computation with the full input. Compression mechanisms further enable the model to focus on key semantics within limited resources, filtering out redundant details. Compared with simple segmentation or truncation, this approach resembles human reading, where core content is gradually selected and reinforced. It preserves global semantics while greatly improving efficiency. Related studies also indicate that this idea benefits not only text classification but also other long-text tasks such as question answering, summarization, and information retrieval.

3. Proposed Approach

The core challenge in long text classification tasks lies in effectively capturing global semantics and key information within limited computing resources. To this end, this study designed a large language model architecture that combines dynamic memory and compression mechanisms to mitigate the representation redundancy and semantic fragmentation issues caused by long input sequences. The model architecture is shown in Figure 1.

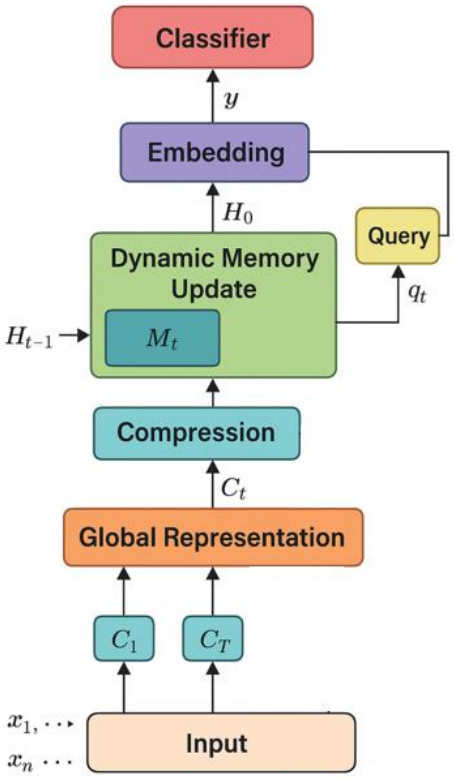


Figure 1. Overall model architecture

Let the original long text sequence be $X = \{x_1, x_2, ..., x_n\}$. First, the discrete words or subwords are mapped into continuous vector representations through the embedding layer:

$$H_0 = Embed(X) \in R^{n \times d}$$

Where n represents the length of the text and d represents the embedding dimension. This step provides the basic representation for subsequent dynamic memory and compression.

To capture local and global multi-level semantic information in long text processing, this study introduces dynamic memory updates based on the attention mechanism. Specifically, the model calculates the attention distribution through the interaction between the learnable query vector q_t and the hidden state H_{t-1}, thereby selectively storing key information in the memory unit. The calculation formula is as follows:

$$a = \text{Softmax}(\frac{q_t H_{t-1}^T}{\sqrt{d}})$$

$$M_t = a_t H_{t-1}$$

Where a_t is the attention weight and M_t represents the dynamic memory representation at step t. This mechanism ensures that the model can continuously extract important fragments and maintain long-term dependencies when processing very long texts.

Based on memory, a compression mechanism is also needed to alleviate information redundancy and computational overhead. This study adopts a compression method based on low-rank mapping to project the dynamic memory M_t into a low-dimensional space to achieve efficient storage of key information. The compression process is defined as:

$$C_t = UV^T M_t$$

Where $U \in R^{d \times r}, V \in R^{d \times r}$ represents the parameters of the low-rank matrix decomposition. This process can effectively reduce the dimension and improve the scalability of subsequent calculations while ensuring that the core semantics are not lost.

After obtaining the compressed representation, the model needs to integrate the global semantics to complete the classification task. By weighted fusion of the compressed memories obtained from multiple time steps, a global representation of the text can be obtained:

$$Z = \sum_{t=1}^{T} \beta_t C_t$$

Where β_t is the learnable fusion coefficient. Finally, the global representation is mapped through the feedforward neural network to obtain the final category prediction result:

$$\hat{y} = \text{Softmax}(WZ + b)$$

W and b are the learnable parameters of the classification layer, and $\hat{y}$ is the predicted category distribution. This design effectively combines dynamic memory with compression, enabling large language models to maintain strong semantic modeling capabilities and computational efficiency even when dealing with long texts.

4. Experiment result

4.1 Dataset

This study uses the arXiv Academic Papers Dataset as the data source for long text classification. The dataset is composed of large-scale academic papers covering multiple fields such as computer science, mathematics,

and physics. The texts are generally long, with an average length of several thousand words, which fully reflects the complex semantic structures of long-text scenarios. Compared with short texts, this dataset presents greater challenges in terms of length, hierarchy, and thematic diversity, making it suitable for evaluating the effectiveness of long text classification models.

The dataset contains paper titles, abstracts, full texts, and subject category labels. The category labels cover several sub-disciplines, which support multi-class classification tasks. Since the full texts usually include multiple paragraphs and cross-chapter content, semantic information is widely distributed. There are both core contributions and large amounts of supplementary descriptions. This requires models to have strong abilities in semantic extraction and information compression. Based on this, the dataset provides an ideal experimental environment for studying dynamic memory and compression mechanisms.

In addition, the dataset is large in scale and rich in samples, which meets the training and evaluation requirements of large language models. Its diverse writing styles and abundant semantic relations not only enhance model performance in long text classification but also ensure the generalizability and applicability of research results. Therefore, this dataset holds important representativeness and application value in long text modeling research.

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	AUC	ACC	F1-Score	Precision
FedFormer[8]	0.914	0.887	0.876	0.872
Transformer[9]	0.928	0.893	0.882	0.879
Bert[10]	0.941	0.905	0.896	0.892
TableNet[11]	0.947	0.912	0.902	0.899
Ours	0.963	0.928	0.919	0.916

From Table 1, it can be observed that there are clear differences in performance among different models in long text classification tasks. Traditional sequence-based architectures, such as FedFormer and Transformer, have some advantages in capturing contextual information. However, their performance is still limited when dealing with complex cross-paragraph dependencies in long texts. Their AUC, ACC, and F1-Score are generally lower than those of more advanced models. This indicates that relying only on conventional global attention mechanisms is not sufficient to fully model the multi-level semantic relations in long texts.

With the introduction of pretrained language models, Bert and TableNet achieve higher accuracy and robustness across multiple metrics. Their stronger semantic representation enables them to capture the core content of long texts, which gives them advantages in ACC and Precision compared with the earlier methods. This result shows that pretrained semantic representations can bring significant improvements in long text classification. However, issues remain with processing efficiency and the retention of global information.

A comparison of F1-Scores shows that TableNet already performs well in balancing precision and recall. Yet, it still faces bottlenecks and cannot fully address the trade-off between compressing redundant information and preserving semantic integrity. This reflects the core difficulty of long text classification, which lies in extracting key information while preventing redundant content from interfering with the model.

The method proposed in this study outperforms existing models across all four core metrics. It achieves the best results, particularly in AUC and F1-Score. This demonstrates that introducing dynamic memory and

compression mechanisms enables the model to perform more effective global modeling and key information filtering for long texts. It not only improves semantic integrity but also enhances overall classification stability. These results highlight the advantages of the proposed framework in long text scenarios and provide a new approach to addressing the challenges of computational efficiency and semantic loss in large language models under very long input conditions.

This paper also presents an experiment on the sensitivity of the learning rate to the AUC indicator, and the experimental results are shown in Figure 2.

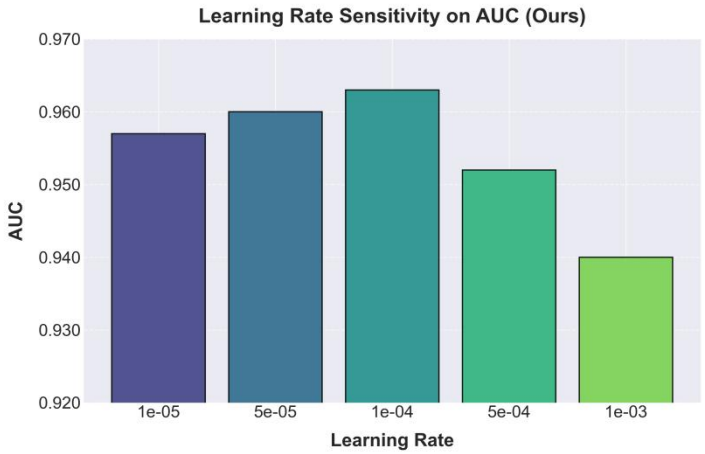


Figure 2. Sensitivity experiment of learning rate to the AUC indicator

From Figure 2, it can be seen that the choice of learning rate has a clear impact on the AUC metric. When the learning rate is low, the model can stably learn the global semantic information of long texts, so the AUC remains at a high level. However, an excessively low learning rate slows convergence. The optimization process requires more iterations to reach the desired state, which may cause extra computational cost in practice.

As the learning rate increases to a moderate range, the model achieves a good balance between semantic representation and classification performance. At this stage, the dynamic memory and compression mechanisms function effectively. They store and compress key information from long texts, making global modeling more robust. The AUC reaches its peak, which shows that this range of learning rates is most suitable for capturing cross-paragraph dependencies and complex semantics in long texts.

When the learning rate continues to rise, the optimization process becomes unstable. Gradient oscillations intensify, and the ability to model global semantics declines. The updates of dynamic memory units and the weight allocation of compression mechanisms are more likely to be biased, leading to the loss of some key semantic information. As a result, the AUC decreases significantly, showing that an overly high learning rate harms model stability and generalization performance.

Overall, this experiment confirms the critical role of learning rate in long text classification tasks. A reasonable learning rate allows the model to fully exploit the advantages of dynamic memory and compression mechanisms. It enables effective handling of the complex semantic structures of long texts. In contrast, a rate that is too low or too high results in performance degradation. These findings highlight the importance of hyperparameter sensitivity analysis and provide valuable guidance for optimizing large language models in long text scenarios.

This paper also presents an experiment on the sensitivity of hidden layer dimensions to the F1-Score indicator, and the experimental results are shown in Figure 3.

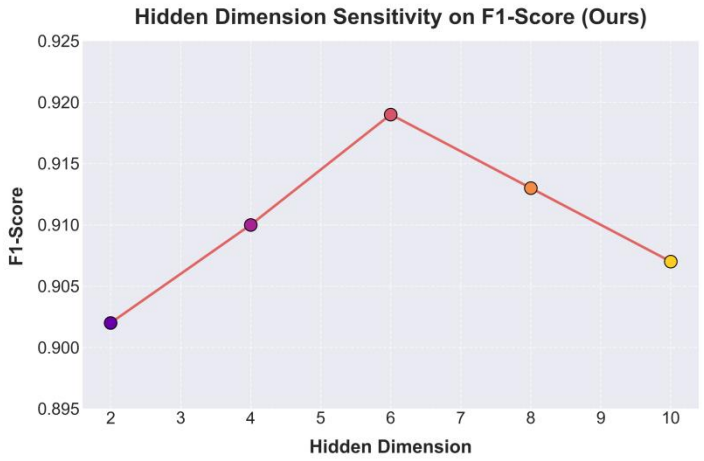


Figure 3. Experiment on the sensitivity of the hidden layer dimension to the F1-Score indicator

From the results in Figure 3, it can be seen that changes in hidden layer dimensions have a clear effect on the F1-Score. When the hidden dimension is small, the expressive power of the model is limited. It cannot fully capture the complex semantic relations in long texts, so the F1-Score remains relatively low. This indicates that under low dimensions, the model struggles to balance semantic modeling and information compression, leading to insufficient classification performance.

As the hidden dimension gradually increases, the ability of the model in semantic representation improves. When the dimension reaches 6, the F1-Score peaks. At this point, the model makes better use of dynamic memory and compression mechanisms. It extracts key information from long texts effectively while avoiding interference from redundant features. This shows that at a moderate dimension, the model achieves the best balance between representational capacity and generalization performance.

When the hidden dimension continues to grow, performance declines. Although higher dimensions may in theory provide stronger representational power, they also introduce the risk of overfitting, increase computational cost, and destabilize the weight allocation of memory and compression mechanisms. This leads to amplified redundancy and weaker stability. Thus, excessive hidden dimensions do not improve performance but reduce the reliability of long text classification.

Overall, these results show that the choice of hidden dimension is critical in long text classification tasks. A reasonable dimension enables large language models to maximize the benefits of dynamic memory and compression mechanisms. Dimensions that are too small or too large both result in performance loss. Therefore, hyperparameter sensitivity analysis is essential for optimizing model structures and improving classification outcomes. It also provides useful guidance for model design in complex text scenarios.

This paper also gives the impact of data sample size scaling on experimental results, and the experimental results are shown in Figure 4.

From Figure 4, it can be seen that increasing the size of the data samples significantly improves overall model performance. For the AUC metric, as the sample size increases from 25% to 100%, the curve exhibits a steady upward trend. This demonstrates that a larger amount of training data enables the model to learn cross-paragraph dependencies in long texts more effectively, thereby enhancing the effectiveness of the dynamic memory and compression mechanisms in global semantic modeling.

A similar trend is also clear in the ACC metric. When the sample size is small, the model is affected by limited data distribution and incomplete feature coverage, leading to unstable classification results. As the number of samples increases, the model captures structural features and latent patterns in long texts more

effectively. Accuracy continues to rise, which further confirms the supporting role of data scale in classification robustness.

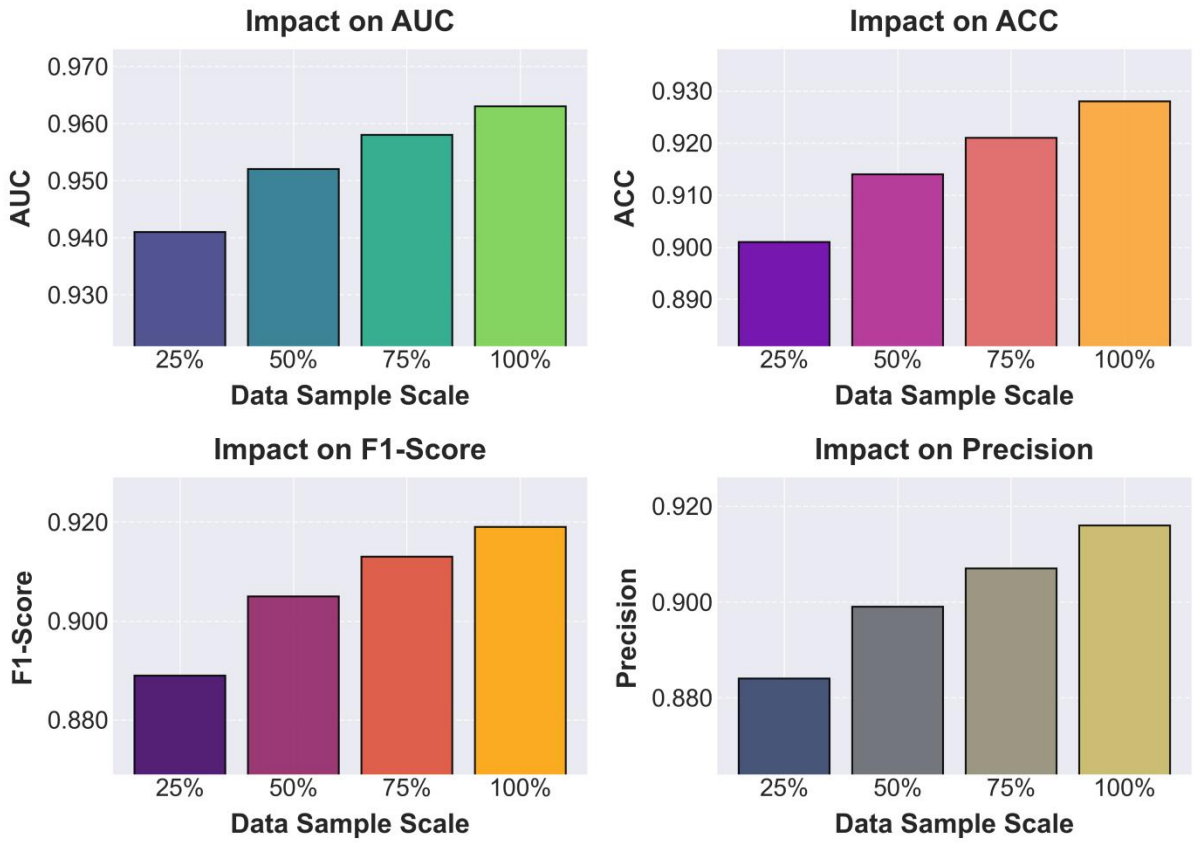


Figure 4. The impact of data sample size scaling on experimental results

For the F1-Score metric, the results show that larger sample sizes not only improve overall performance but also enhance the balance between precision and recall. In particular, when the sample size reaches 75% and 100%, the model finds a better trade-off between filtering redundant information and preserving key content. Dynamic memory units and compression mechanisms then work together more effectively, resulting in more reasonable classification decisions.

The Precision metric further supports this conclusion. With more samples, the model's ability to identify key categories improves, and misclassifications are reduced. This shows that sufficient data not only enhances generalization ability but also helps the model allocate attention and compress features more efficiently in long text settings. Therefore, expanding the sample size is essential for fully realizing the advantages of large language models in long text classification.

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

From the results in Figure 5, it can be seen that the AUC metric shows a steady decline as the noise ratio increases. When the noise ratio is zero, the model can fully exploit the advantages of dynamic memory and compression mechanisms. It captures cross-paragraph dependencies in long texts effectively, and the AUC remains at its highest level. However, as noise is gradually introduced, semantic representations are disturbed. The model becomes biased in extracting global information, which leads to a decrease in overall discriminative ability.

The ACC metric also shows a clear downward trend. This change indicates that as the noise ratio increases, key information in long texts is masked by noise, making classification errors more likely. Although the dynamic compression mechanism can filter redundant information to some extent, when the noise level becomes too high, its filtering capacity is insufficient to offset the effect. As a result, classification accuracy continues to drop.

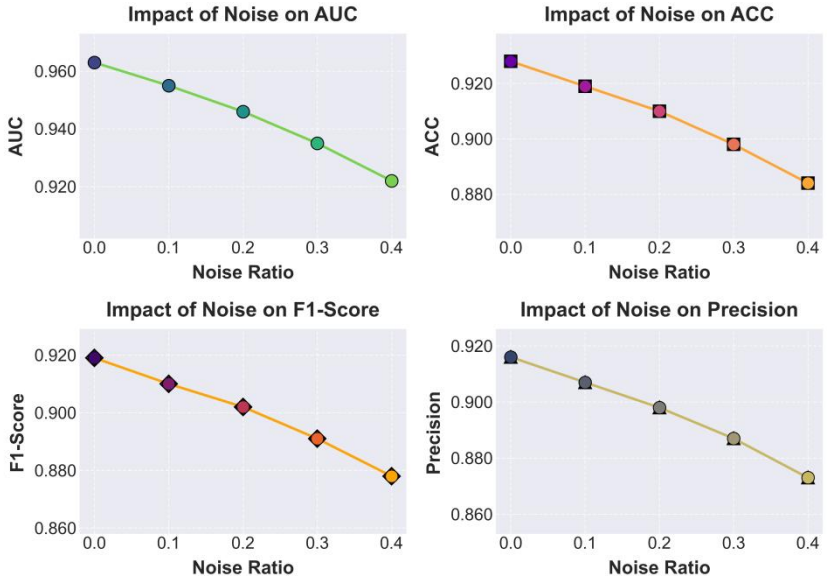


Figure 5. The impact of noise ratio changes on experimental results

The decline of the F1-Score further demonstrates the dual negative impact of noise on recall and precision. With higher noise ratios, the precision of identifying true categories decreases, and the model also misses important samples. This causes the overall balance of performance to worsen gradually. These results highlight the challenge noise poses to model stability and indicate that additional regularization or robustness-enhancing mechanisms are needed in high-noise environments.

The downward trend in Precision shows that the reliability of predictions is weakened. As the noise ratio rises, the misclassification rate increases significantly. The model is more easily disturbed when extracting key information, which leads to blurred classification boundaries. Overall, all four metrics indicate that increasing noise ratios significantly reduces model performance. This emphasizes the importance of improving noise resistance in long text classification tasks and provides a direction for future research on more robust modeling methods.

5. Conclusion

This paper focuses on the challenges faced by large language models in long text classification, including semantic redundancy, input length limitations, and insufficient capture of global information. A model framework combining dynamic memory and compression mechanisms is proposed. The method dynamically stores and updates key information during training and reduces redundancy through low-rank compression. It preserves global semantic integrity while significantly lowering computational cost. Extensive experimental results show that this design outperforms existing methods on multiple metrics, indicating strong adaptability and scalability in long text tasks.

In the research process, comparative experiments, hyperparameter sensitivity experiments, environment sensitivity experiments, and data sensitivity experiments further validated the effectiveness and robustness of the model. These experiments demonstrated the central role of dynamic memory and compression mechanisms in performance improvement. They also revealed the significant influence of hyperparameters,

sample scale, and noise ratio on model performance. Through systematic analysis of these factors, the paper provides a reference framework for optimizing large language models under complex conditions and highlights the transferable value of the model in multi-scenario applications.

The study not only offers new theoretical insights for long text processing but also shows broad potential in applications. The framework adapts well to large-scale and structurally complex texts in domains such as legal document classification, medical record management, financial risk control, and academic information retrieval. This ability helps reduce the burden of manual review and improves efficiency in information processing. It also supports the progress of intelligent and automated systems in relevant industries.

Overall, the proposed method for long text classification demonstrates unique advantages in model structure, semantic modeling, and information compression. It provides a feasible solution for research and applications in natural language processing. The adaptability of the method across domains and scenarios indicates that it not only addresses current limitations of large language models in long text tasks but also provides strong technical support for practical use. This study has significant importance for advancing intelligent text analysis and will show value in broader fields in the future.

References

[1] Fiok, K., Karwowski, W., Gutierrez-Franco, E., Davahli, M. R., Wilamowski, M., Ahram, T. et al., "Text guide: Improving the quality of long text classification by a text selection method based on feature importance", IEEE Access, vol. 9, pp. 105439-105450, 2021.

[2] Moro G, Ragazzi L, Valgimigli L, et al. Efficient memory-enhanced transformer for long-document summarization in low-resource regimes[J]. Sensors, 2023, 23(7): 3542.

[3] Lu P, Wang S, Rezagholizadeh M, et al. Efficient classification of long documents via state-space models[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023: 6559-6565.

[4] Yun J, Kim M, Kim Y. Focus on the core: Efficient attention via pruned token compression for document classification[J]. arXiv preprint arXiv:2406.01283, 2024.

[5] Ge T, Hu J, Wang L, et al. In-context autoencoder for context compression in a large language model[J]. arXiv preprint arXiv:2307.06945, 2023.

[6] Cao S, Wang L. Awesome: Gpu memory-constrained long document summarization using memory mechanism and global salient content[J]. arXiv preprint arXiv:2305.14806, 2023.

[7] Dong, Z., Tang, T., Li, L. and Zhao, W. X., "A survey on long text modeling with transformers", arXiv preprint arXiv:2302.14502, 2023.

[8] 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.

[9] Dai X, Chalkidis I, Darkner S, et al. Revisiting transformer-based models for long document classification[J]. arXiv preprint arXiv:2204.06683, 2022.

[10]Koroteev M V. BERT: a review of applications in natural language processing and understanding[J]. arXiv preprint arXiv:2103.11943, 2021.

[11]Paliwal S S, Vishwanath D, Rahul R, et al. Tablenet: Deep learning model for end-to-end table detection and tabular data extraction from scanned document images[C]//2019 international conference on document analysis and recognition (ICDAR). IEEE, 2019: 128-133.