Analysing the Current State of Development of Bert-Based Text Classification in Natural Language Processing

Litong Pu[1], Yang Yang[2]*

[1]College of Sciences, Nanjing Agricultural University, Nanjing, Jiangsu, 210008, China
[2]College of Economics and Management, Nanjing Forestry University, Nanjing, Jiangsu, 210037, China

Abstract. Text classification is an important task in natural language processing, and traditional text classification methods have certain limitations, such as low accuracy, low flexibility and high disadvantages in terms of computational resources and time cost. With the development of deep learning, the BERT text classification model shows what advantages can compensate for the limitations of traditional methods. This paper aims to explore the development of BERT for text classification from three aspects: the birth of BERT, the optimisation and improvement of BERT, and the architectural innovation and application extension. This paper concludes that BERT has brought text classification to a new stage of 'pre-training + fine-tuning'. Its variants are more effective than other methods for binary sentiment classification, Alzheimer's disease detection and power audit text classification. However, the BERT model suffers from high computational resource requirements, high efficiency compromised by computational complexity, insufficient model interpretability, and low adaptability to specific domains and small datasets. In the future, we can promote the development of efficient model architectures and training methods, focus on interpretable model architectures and tools, and improve the adaptability of BERT in specific domains.

1 Introduction

Text classification is the assignment of documents to one or more predefined categories based on the content of the file, which has various applications such as sentiment analysis, offensive language detection and language recognition [1]. The development of text classification techniques can help to improve the efficiency of information retrieval, help platforms to achieve better content recommendations and facilitate natural language understanding. Traditional text classification methods such as Native Bays, KNN and SVM

* 731224143@njfu.edu.cn

have problems such as limited feature independence assumptions and long training time, etc. BERT adopts a bi-directional Transformer encoder, which is able to more accurately understand the semantics of words and the meanings of sentences and the pre-training parameters can be fine-tuned in a variety of text classification tasks, which is better adapted to the new data distributions and task characteristics.

BERT is also widely used in the field of text classification although it is not specifically designed to handle text classification. In recent years, BERT models have been improved in various aspects, such as improving training efficiency, model compression to increase speed, and task adaptation improvement. The main research directions of the current scholars are improving the training model to enhance the efficiency, model compression to make the operation more efficient, multilingual text classification, and the integration of BERT with other techniques.ELECTRA model builds an architecture with generator-discriminator as the core. The BERT-based model, fused with GAN, works by replacing some of the input tokens with plausible tokens and then training the model to predict whether the tokens have been replaced or not. Compared to MLM, this method is more efficient, produces better results for NLP downstream tasks and has a small computational cost [2]. The EBIDS model built based on BERT can effectively improve the intrusion detection capability of the network and application layers in the IoT environment through context detection. The model can not only significantly reduce the risk of attacks against the network and application layers of IoT systems, but also improve the accuracy of detection while minimising the computational overhead [3]. With the development of the Internet, social media based on short texts are becoming more and more popular, and training the Bert model based on the psychological characteristics of the author can enable the model to capture the psychological characteristics in the text more effectively, and then improve the accuracy of short text classification [4].

Taking the development history of BERT as a carrier, this paper reviews the research progress of the BERT text classification model in recent years, analyses and compares the major text classification models, and puts forward suggestions for future research on the BERT text classification model from the three aspects of the birth of BERT, the optimization and improvement of BERT, and the innovation of the architecture and the expansion of the application.

2 History of BERT

2.1 Birth of BERT

In 2018, BERT proposed by Devlin et al. achieved deep context modelling for the first time through a bidirectional Transformer encoder, completely changing the limitations of traditional unidirectional language models.BERT was initially designed to be oriented towards general-purpose natural language processing tasks requiring deep semantic understanding, including text categorisation (e.g., sentiment analysis, topic classification), sequence annotation (named entity recognition), question and answer systems (SQuAD), and sentence pair tasks (natural language reasoning). Its core strengths are in three areas:

First, through the masked language modelling (MLM) task, BERT is able to dynamically learn the polysemous representations of words in bi-directional contexts, which solves the problem that static word vectors such as Word2Vec are unable to deal with the multiple meanings of a word; second, the deep architecture formed by the stacking of 12/24 layers of Transformers, through the self-attention mechanism, automatically captures the long-distance dependency relationships in the text (e.g. subject-predicate-consistency), which significantly outperforms traditional models such as

RNN/CNN. Consistency, denotation disambiguation), which is significantly better than the sequence modelling ability of traditional models such as RNN/CNN; finally, the pre-training-fine-tuning paradigm makes the model parameters cross-task migratory.BERT employs a dual-task framework consisting of masked language modelling (MLM) and next sentence prediction (NSP). 80% of the lexical elements are replaced with [MASK], there is a 10% chance that they will be replaced with randomly selected words, and the remaining 10% are the original lexical elements that remain unchanged, which forces the model to rely on learning that incorporates bi-directional contexts for the prediction of the obscured words. During the operational phase of the NSP task, work is carried out on judgements against sentence pair continuity (50% real combinations and 50% using random combinations) as a means of enhancing the model's insight into the structure of the text[5].

However, there are still significant shortcomings in the early versions of BERT, such as BERT-Base (110 million parameters) and BERT-Large (340 million parameters), which are difficult to deploy in resource-constrained environments due to the large number of parameters that make training and inference expensive; meanwhile, the square-scale computational complexity of the self-attention mechanism makes it inefficient to process long text, and the general-purpose corpus pre-training limits its direct application to professional domains. the direct application in professional domains. These properties provide a clear path for subsequent research to improve the model compression (e.g., DistilBERT), long sequence optimisation (e.g., Longformer) and domain adaptation (e.g., BioBERT)[6].

2.2 Optimisation improvements of BERT

Initial BERT has obvious shortcomings, firstly, the number of parameters is large (BERT-Base reaches 110 million), and the training/reasoning cost is high. Secondly, the complexity of self-attention computation grows with the square of sequence length, and long text processing is inefficient. Meanwhile, generic pre-training leads to insufficient adaptation to specialised domains. Therefore, improvements in training efficiency enhancement and model optimisation are significant and necessary.

2.2.1 Improvements in training efficiency

Although the original BERT model shows excellent language comprehension ability, its training process suffers from efficiency bottlenecks such as intensive computational resources and slow convergence speed, which directly restricts the scalability of the model in practical applications. The core value of improving training efficiency is reflected in the following: from the technical level, optimising the training process can significantly reduce the computational cost, which enables research institutions with limited resources to participate in large-scale pre-training; from the application perspective, more efficient training means faster model iteration cycle, which accelerates the transformation of NLP technology from the laboratory to the industry; from the academic innovation dimension, the improvement of training efficiency provides a new opportunity for exploring larger-scale and more complex model architectures, which can be used to improve language understanding. From the academic innovation dimension, the improved training efficiency provides the possibility of exploring larger scale and more complex model architectures, pushing the boundary of language modelling research to expand. In addition, this optimisation also responds to the pursuit of sustainable development in the field of Artificial Intelligence by reducing energy consumption for environmentally friendly model development.

Against the background of the aforementioned multidimensional value of training efficiency, researchers have been working on improving BERT training methods to balance model performance and efficiency. In this context, RoBERTa has emerged as an important representative of BERT training strategy optimisation. It improves the training method of BERT by adopting the dynamic masking strategy, i.e., reshaping the masking position in each epoch to avoid the bias condition caused by the fixed masking pattern, while RoBERTa cancels out the NSP task to achieve the simplification of the model training objective, and also increases the batch size to 32k so as to make full use of more training data, and these improvements make the RoBERTa improves the SQuAD F1 score by 2.3% compared to BERT [7], which achieves the growth of training efficiency while guaranteeing the improvement of the model performance, and retains the valuable experience for the subsequent large-scale model training.

2.2.2 Model compression

Model compression is an approach to reduce the number of parameters and computation of a deep learning model through specific technical means, while maintaining its performance as much as possible, so as to make the model reach a more lightweight and efficient level. For large-scale pre-trained models such as BERT, the core purpose of compression is to solve the three major practical problems caused by the large number of parameters: reducing the demand for computational resources to adapt to the deployment of edge devices, reducing the inference latency to meet the real-time requirements, and alleviating the pressure on storage space. This direction of technology shows significant advantages: firstly, through knowledge distillation and other strategies, it is possible to maintain more than 95% of the model performance while achieving a significant reduction in parameter quantities; secondly, the compressed model can break through the original hardware limitations, significantly expanding the boundaries of application scenarios.

Model compression enables BERT to operate efficiently in resource-constrained environments such as mobile terminals and embedded systems, and moreover reduces computational energy consumption. This technology has far-reaching significance to the field of natural language processing, which not only promotes the pre-trained model from the laboratory to the industrial landing but also provides key technical support for the subsequent combination of edge computing and natural language processing. Based on these goals, researchers have successively proposed a variety of innovative compression methods.

In 2019 Lan et al. proposed ALBERT, which for the first time uses cross-layer parameter sharing technique and combines it with word vector decomposition operation to reduce the number of parameters to 18% of the original BERT, and also possesses a performance equivalent to 98% of the original model [8]. This parameter sharing mechanism reduces the storage space requirement of the model and also mitigates the overfitting dilemma to some extent. Notably, these innovations allow ALBERT to adapt to resource-constrained application scenarios while maintaining the performance of the BERT model.

In contrast, DistilBERT proposed by Sanh et al. uses knowledge distillation to compress the model size, relying on the knowledge distillation technique to build a lightweight model with 6 layers, while retaining most of the performance of BERT, DistilBERT achieves a 60% improvement in inference speed [9], allowing BERT to run more efficiently in resource-constrained environments, such as mobile devices and embedded systems. Therefore, DistilBERT is more suitable for deployment environments that require fast response and limited computational resources.

2.2.3 Increased adaptability of tasks

*A*iming at the underperformance of BERT on structured tasks such as NER and relationship extraction, the researchers propose a task-adaptive optimisation approach, the core of which is to improve its performance in professional domain tasks (e.g. entity recognition, relationship extraction) by improving the pre-training strategy and model architecture. This research direction can not only give full play to the migration learning advantages of BERT, but also significantly improve the accuracy and robustness of the model in real application scenarios through task customisation design, and provide more accurate technical support for NLP applications in professional domains.2019 Joshi et al. proposed SpanBERT, which introduces a boundary prediction task in the pre-training stage, and also employs geometric distribution span masking to make the model more advantageous for entity and phrase boundary understanding in text. In the CoNLL - 2003 entity recognition task, SpanBERT improves the F1 score by 4.1% [10], which effectively improves the quality of extracting entity information from text.ERNIE, proposed by Sun et al. fuses the knowledge graph with entity-level masking (e.g., masking ‘Paris’ instead of single words), which strengthens the model's perception of knowledge. It strengthens the model's ability to perceive knowledge, and improves the accuracy by 5.6% during the implementation of the Chinese Named Entity Recognition (NER) task [11], highlighting its advantages in the task of processing knowledge-rich text.

2.2.4 Multilingual extensions

Multilingual extension research is committed to breaking through the limitations of BERT monolingual processing, which is important to build a cross-language unified representation space to achieve knowledge migration between different languages. This development not only dramatically reduces the development cost of multilingual NLP applications, but also improves the model performance of low-resource languages by sharing semantic representations, providing a fundamental technical solution for natural language processing in globalised scenarios. In 2019, Lample and Conneau proposed the XLM model, which achieves the extension of the multilingual dimension, and the XLM's Translation Language Modelling (TLM) task relies on parallel corpus-aligned representations of 100 + languages, and in the case of the XNLI cross-language task, the average accuracy rate reaches 75.1%, which strongly supports the construction of multi-language NLP applications [12], and promotes the interaction and comprehension of information between different languages.

3 Architecture Innovation and Application Expansion

With the wide application of the BERT model in the field of natural language processing, its powerful language comprehension ability has shown significant advantages in tasks such as text categorization, sentiment analysis, and entity recognition. However, in the face of diverse practical needs and complex scenarios, the traditional BERT model has gradually exposed some limitations. For example, in domain-specific tasks, it is difficult for generic pre-training models to adequately capture specialized terminology and domain knowledge; in resource-constrained environments, the huge number of model parameters becomes a bottleneck for deployment. In addition, traditional pre-training paradigms also face efficiency and performance challenges when dealing with long text, multimodal data, or interdisciplinary tasks.

The existence of these problems motivates researchers to continuously explore new architectural innovations and application expansion paths. By improving the model structure, optimizing the training strategy, and integrating multi-domain knowledge, the

variant models of BERT have achieved significant improvements in performance, efficiency, and adaptability.

3.1 Architectural Innovations

In 2020, Clark et al. proposed ELECTRA, which breaks through the old paradigm of traditional generative pre-training and builds a generator-discriminator-centered architecture, with the discriminator taking on the task of detecting 15% of replaced words, which improves the GLUE score by 3.2% under the condition of the same computational resources [2], and improves the model's ability of discriminating textual details to a certain extent The StructBERT proposed by Wang et al. uses the word order/sentence order reconstruction task to improve the model's modeling of text structure. The correlation coefficient reaches 91.2% in the text similarity task STS-B [13], which further improves the model's performance in semantic understanding. The TinyBERT model proposed by Jiao et al. adopts a four-phase distillation to cut the parameters down to 14.5M, and finally achieves real-time reasoning on cell phones (<50ms/sentence) [14], which realizes the model's lightness and achieves real-time inference in real applications. It realizes the light weight of the model and achieves the real-time requirement in practical applications.

3.2 Application Expansion

With the continuous innovation of the BERT model, its application scope has been extended from the traditional text classification task to more fields.

In the field of sentiment analysis, TF-BERT adopts the design of TCT module to carry out multi-module processing simultaneously, which greatly improves the characterization effect of target modal sentiment features. When using the CMU-MOSEI dataset, TF-BERT measured an MSE of 0.546, which is lower than most other models, confirming that its prediction error is in the lower range; in the dimension of ACC-2, TF-BERT achieved an accuracy of 85.99%, reflecting that it has a strong competitiveness in the stage of the binary classification task, and it can provide a more accurate processing for the sentiment analysis task method [15].

In the field of disease discrimination and network security detection, BERT also shows its powerful potential ability. For example, with the tFUS-BERT model, Mel spectrograms are transformed into pseudo-phoneme sequences, and learning activities are carried out based on the transformer-based architecture to identify language changes in normal aging and language abnormalities in Alzheimer's disease patients. When combined with Wav2vec 2.0, the accuracy of this model reaches 76.06%; if combined with Hu-BERT, the model accuracy can reach 71.83%, which provides a new technological pathway for the detection of Alzheimer's disease [16].

The EBIDS model, built on the basis of BERT, carries out context detection, which in turn enhances the level of intrusion detection at the network and application layers of the IoT environment, reduces the harm of attacks against the network and application layers of the IoT system, and improves the accuracy of detection, while minimizing the computational overhead to the maximum limit. In the Edge-HoT dataset, the F1 score of the EBIDS model reaches 97.47%, and in the CICDos2017 scenario, the F1 score reaches 94.28%, and the actual time taken for the detection is shorter, which has obvious advantages over traditional methods such as CNN, RNN, and LSTM, and it provides a highly efficient and trustworthy means of protection for network security [3].

In 2023, the EPAT-BERT model proposed by Qinglin Meng et al. was used for the power audit text categorization task [17]. What this model introduces is the electric power domain related text, which solves the problem of traditional BERT when dealing with

domain-specific text. Generally speaking, power audit texts contain a lot of specialized terminology and context-specific content, and there is a probability that generic BERT models cannot adequately capture this information. Therefore, EPAT-BERT introduces the Entity Level Masking Language Model and Entity Masking Language Model, and carries out targeted masking and prediction learning for the entities in the electric power domain in the pre-training phase, so as to make the model understand the terminology and contextual relationships in the electric power auditing text more effectively. The experimental results reveal that traditional BERT, SVM, LSTM and other methods perform far worse than EPAT-BERT in the power auditing text categorization task, with an accuracy of 0.8196, precision of 0.8079, recall of 0.8162, and F1 score of 0.8120, which provides a more efficient and reliable means of solving the power auditing text categorization task. solution.

3.3 Internal to Interdisciplinary Applications of the BERT Model

During the time when BERT gained widespread popularity, researchers began to conduct in-depth analysis of its internal working mechanism, and by 2020, Tal Linzen et al. carried out a systematic analysis of BERT's internal working mechanism and knowledge representation properties [18], using probe tasks such as syntactic tree reconstruction to detect linguistic features of each layer, using visual attention weighting pathways to reveal pattern patterns, and also quantitative analysis of the differences in the encoding of lexical, syntactic, and semantic information by different layers. These studies show how BERT grasps the linguistic structure at different stages, and provides a theoretical basis for subsequent model optimization and task customization activities. For example, it is pointed out that the shallow layer of BERT focuses more on lexical information, while the deeper layer pays more attention to the processing of complex syntactic and semantic relations, which provides a key reference for the purposeful optimization of the model architecture and the design of task-specific training strategies. This finding provides a key reference for purposefully optimizing the model architecture and designing task-specific training strategies.

Traditional NLP tasks have widely adopted the powerful language comprehension capability of BERT, which also spawned research innovations at the interdisciplinary level. In 2022, Yongjun Hu et al. proposed a BERT-based psychological method based on short text categorization [4], which combines the historical information with psychological features such as Maslow's needs and realizes an effective integration of the psychological feature vectors with the short text vectors, which greatly The accuracy of text categorization is greatly improved. This interdisciplinary attempt not only expands the application scope of BERT, but also opens up new ideas for the combination of NLP and psychology. In the same year, Chao - Han Huck Yang proposed BERT-QTC, a quantum time convolutional learning framework based on the design of BERT [19], which is used in the spoken language data sets of Snips and the Airline Travel Information System (ATIS). Information System (ATS) spoken language dataset, this model achieves good results, reaching performance improvements of 1.57% and 1.52%, respectively. In addition, BERT-QTC can be deployed on existing commercial quantum computing hardware and CPU-based interfaces to achieve effective data isolation, which is an innovation that demonstrates the potential of BERT in quantum computing and gives new guidelines for future high-performance computing applications.

4 Challenges and Prospects

Although BERT and its many variants have presented significant results on text categorization tasks, there are still certain limitations and shortcomings of the BERT model. These limitations provide important insights for future research directions.

First of all, there are some challenges in computational resources and efficiency. The large size of BERT model parameters (e.g., BERT-Base contains 110 million parameters, since BERT-Large has 340 million parameters) requires a lot of computational resources in the training and inference phases, which limits the application of BERT in resource-constrained environments (e.g., mobile devices and embedded systems), as well as the cost and time of model training. This limits the application of BERT in resource-constrained environments (e.g., mobile devices and embedded systems) and increases the cost and time of model training. In addition, BERT's bidirectional transformer architecture is inefficient when processing long text, because its computational complexity is proportional to the square of the sequence length. Future research can focus on promoting the development of more efficient model architectures and training methods, such as the use of sparse attention mechanisms, dynamic computational graphs, and other techniques to achieve the optimization of transformer architectures, which can be used to improve the efficiency of BERT in long text processing and resource-constrained scenarios. long text processing and resource-constrained scenarios.

Second, the model is not sufficiently interpretable; the deep neural network structure of BERT is strongly black-boxed when processing text, and it is not easy to explain the decision-making process of the model, such as in medical diagnosis and legal text, which require high interpretability. Even though some studies have attempted to illustrate the behavior of BERT by means of attention weight visualization, it is still difficult to fully unravel the internal working mechanism of the model. Therefore, future research can focus on the development of more interpretable model architectures and auxiliary tools, such as combining symbolic reasoning and causal inference techniques to improve the interpretability of BERT models, and exploring how to combine BERT with structured knowledge such as knowledge graphs to enhance the transparency and credibility of the models.

Meanwhile, there is the problem of adapting to specific domains and small datasets: BERT is pre-trained on a large-scale generalized corpus, and due to its limited effectiveness in capturing domain-specific terminology and contextual information, and its overfitting tendency when dealing with small datasets, its performance on specific domains (e.g., medicine, law, power auditing, etc.) or small datasets might not be as good as that on generalized texts. The direction of subsequent research can be explored in the development of domain-adaptive pre-training methods, such as the use of domain-adversarial training, domain-specific masked language models and other techniques to improve the level of adaptation of BERT to specific domains, and at the same time, research on how to more effectively exploit the value of BERT on small data sets, such as the use of data enhancement, migration learning and other means.

5 Conclusion

Text categorization tasks have been revolutionized by the emergence of BERT, relying on the bidirectional Transformer architecture and large-scale pre-training means, greatly enhancing the effectiveness of text categorization, pulling into the “pre-training + fine-tuning” of the emerging stage. In recent years, through continuous improvement, the training efficiency of the BERT model has been greatly improved, the number of parameters of the model and greatly reduced, the operation is more efficient, the adaptability of the task has been enhanced, in addition to supporting the construction of multi-language applications. The model improvement also makes BERT a more efficient

and reliable solution that can be used in a variety of fields. However, the BERT model still has several constraints and shortcomings, such as high computational resource requirements, weak interpretability, problems in adapting to specific domains and small datasets, as well as limited multimodal fusion capabilities, which are also the directions of BERT's future research.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

1. Vijayan, V. K., Bindu, K. R., Parameswaran, L.: "A comprehensive study of text classification algorithms," 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Udupi, India, 2017, pp. 1109–1113.
2. Clark, K., Luong, M.-T., Le, Q. V., Manning, C. D.: "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators," International Conference on Learning Representations, 2020.
3. Sattarpour, S., Barati, A., Barati, H.: "EBIDS: efficient BERT-based intrusion detection system in the network and application layers of IoT," Cluster Computing, 2025, 28, pp. 138.
4. Hu, Y., Ding, J., Dou, Z., Chang, H.: "Short-text classification detector: A BERT-based mental approach," Computational Intelligence and Neuroscience, 2022.
5. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, 1, pp. 4171–4186.
6. Rogers, A., Kovaleva, O., Rumshisky, A.: "A Primer in BERTology: What We Know About How BERT Works," Transactions of the Association for Computational Linguistics, 2020, 8, pp. 842–866.
7. Liu, Z., Lin, W., Shi, Y., Zhao, J.: "A Robustly Optimized BERT Pre-training Approach with Post-training," Proceedings of the 20th Chinese National Conference on Computational Linguistics, 2021, pp. 1218–1227.
8. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations," 2020.
9. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: "DistilBERT: A distilled version of BERT: smaller, faster, cheaper and lighter," 2020.
10. Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., Levy, O.: "SpanBERT: Improving Pre-training by Representing and Predicting Spans," Transactions of the Association for Computational Linguistics, 2020, 8, pp. 64–77.
11. Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., Liu, Q.: "ERNIE: Enhanced Language Representation with Informative Entities," Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 1441–1451.
12. Conneau, A., Lample, G.: "Cross-lingual language model pretraining," Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2019, pp. 7059–7069.

13. Wang, W., Bi, B., Yan, M., Wu, C., Bao, Z., Xia, J., Peng, L., Si, L.: "StructBERT: Incorporating Language Structures into Pre-training for Deep Language Understanding," 2019.
14. Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., Liu, Q.: "TinyBERT: Distilling BERT for Natural Language Understanding," Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 4163–4174.
15. Hou, J., Omar, N., Tiun, S., Saad, S., He, Q.: "TF-BERT: Tensor-based fusion BERT for multimodal sentiment analysis," Neural Networks, 2025, 185, 107222.
16. Thipparthy, K. R., Kollu, A., Kulkarni, C., Dutta, A. K., Doshi, H., Kashyap, A., Sinha, K. P., Kondaveeti, S. B., Gupta, R.: "Discrete variational autoencoders BERT model-based transcranial focused ultrasound for Alzheimer's disease detection," Journal of Neuroscience Methods, 2025, 416, 110386.
17. Meng, Q., et al.: "Electric Power Audit Text Classification With Multi-Grained Pre-Trained Language Model," IEEE Access, 2023, 11, pp. 13510–13518.
18. Clark, K., Khandelwal, U., Levy, O., Manning, C. D.: "What Does BERT Look at? An Analysis of BERT's Attention," Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, 2019, pp. 276–286.
19. Yang, C.-H. H., Qi, J., Chen, S. Y.-C., Tsao, Y., Chen, P.-Y.: "When BERT Meets Quantum Temporal Convolution Learning for Text Classification in Heterogeneous Computing," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022, pp. 8602–8606.