ESP-IJCEET
ESP International Journal of Communication Engineering & Electronics Technology
ISSN: 2583-9217 / Volume 2 Issue 1 March 2024 / Page No: 35-46
Paper Id: IJCEET-V2I1P107 / Doi: 10.56472/25839217/IJCEET-V2I1P107

*Original Article*

# Pre-Trained Models in Natural Language Processing: Progress after BERT

**AnNing[1], Mazida Ahmad[2], Huda lbrahim[3], Wang Zhuoxian[4], Xu Jie[5]**
[1]*Jinzhong City Education Bureau,JinZhong, China.*
[1,2,3,4,5]*Institute for Advanced and Smart Digital Opportunities (IASDO), School of Computing, Universiti Utara Malaysia, Sintok, Kedah, Malaysia.*
[5]*School of Literature and Media, Chaohu University, Hefei, China*

Received Date: 03 January 2024 Revised Date: 05 February 2024 Accepted Date: 06 March 2024

***Abstract:*** *With the continuous development of the natural language processing field, the pre-training model has achieved remarkable results in text processing tasks. As a representative, the BERT model achieves leading performance on multiple natural language processing tasks by pre-training and fine-tuning. However, with the deepening of the research, the BERT model has also exposed some problems, such as the high training cost and the large model scale. Therefore, the investigators proposed a series of improved models, such as XLNet, RoBERTa, ALBERT, and ELECTRA, to address the shortcomings of the BERT model. These models innovate in model architecture, training strategies, and optimization methods, and achieve better performance. In addition, the researchers proposed a variety of pre-trained models, such as fine-tuning, domain adaptation, transfer learning, and multi-task learning, to further improve the performance of the model on specific tasks. However, pre-trained models still face challenges such as high training costs and poor model interpretability. Therefore, future research directions can focus on reducing training costs, improving model interpretability, and further optimizing performance on specific tasks. This paper summarizes the development and application of pre-trained models in natural language processing, introduces some important advances after BERT, and explores the challenges of pre-trained models and future development directions.*

***Keywords:*** *Pre-Training Model, BERT, XLNet, RoBERTa, ALBERT, ELECTRA, Improvement Method, Challenge, Outlook.*

## I. INTRODUCTION

Natural language processing (Natural Language Processing, NLP) is an important research direction in the field of artificial intelligence, aiming to enable computers to understand and process human language. With the rise of big data and deep learning, pre-trained models have made significant breakthroughs in the field of NLP. Among them, BERT (Bidirectional Encoder Representations from Transformers) model, as a pre-training model based on Transformer structure, leads the development trend of NLP field.

In the traditional NLP methods, researchers mainly rely on the manual design of features and rules, which limits the expression and generalization ability of the model. The pre-trained model, on the other hand, learns rich language representation through unsupervised learning on large-scale text corpus, so that the model has better semantic understanding and reasoning ability. The emergence of BERT model has not only achieved leading results on multiple NLP tasks but also driven the development of many related studies.

However, despite the great success of the BERT model, it also has some shortcomings. For example, the BERT model masks the input text during training, making the model unable to accurately handle long-distance dependencies. Moreover, the BERT model is expensive and requires substantial computational resources and time. Therefore, the investigators propose a series of improved models after BERT to further improve the performance of the pre-trained model.

The present paper aims to review and analyze the progress of the pre-trained model after the BERT model. First, we will review the development of traditional NLP methods and pre-trained models, and introduce the principles and applications of the BERT model. Then, we will detail the progress after BERT, including XLNet, RoBERTa, ALBERT, and ELECTRA models, and analyze their principles, experimental results, and applications. Next, we discuss the improvement methods and application areas of pre-training models, including fine-tuning, domain adaptation, transfer learning, and multi-task learning methods. Finally, we

explore the challenges of pre-trained models and explore the future development of pre-trained models in task-specific optimization.

Through the research of this paper, we can have a more comprehensive understanding of the latest progress of pre-training models in the field of NLP, provide reference and guidance for researchers and developers, and promote the development and application of NLP technology.

## II. DEVELOPMENT AND APPLICATION OF THE PRE-TRAINING MODEL

### *A. Traditional Natural Language Processing Methods*

Traditional NLP methods are those used before the emergence of deep learning and pre-trained models. These methods are mainly based on rules and feature engineering and require manual design and feature extraction to solve natural language processing tasks. Some of these common traditional methods include the bag of words model, n-gram model, TF-IDF, etc.

The pouch model is a simple way of representing text as a disordered set of words. The model ignores the order and grammatical structure between words and focuses only on the frequency of words. The n-gram model considers the order between words, divides the text into consecutive sequences of n-words, and counts their occurrence frequency. TF-IDF (Term Frequency-Inverse Document Frequency) is a method used to measure the importance of words in text that combines word frequency and inverse document frequency for tasks such as text classification and information retrieval.

Traditional NLP methods perform well on some simple tasks, such as text classification, information retrieval, etc. However, these methods have several limitations. First, they have limited ability to grasp semantic understanding and contextual information, unable to handle complex semantic relations and context. Secondly, features need to be manually designed and extracted, and feature engineering processes need to be redesigned and adjusted for different tasks. Moreover, these methods are less efficient for processing large-scale data and often require longer training times.

With the development of deep learning and pre-training models, the traditional natural language processing methods are gradually replaced by deep learning models. Deep learning models can better solve natural language processing tasks by automatically learning feature representation and semantic information. Pre-trained models such as BERT and XLNet have made remarkable breakthroughs in the field of NLP, providing a powerful basic model for various NLP tasks.

### *B. Rise of Pre-Trained Models*

*The rise of the pre-trained models*

Natural language processing (Natural Language Processing, NLP) is an important research direction in the field of artificial intelligence, aiming to enable machines to understand and process human language. Over the past few decades, researchers have made some progress in building various feature engineering and machine learning models. However, these traditional approaches often require substantial manual involvement and domain knowledge and fail to handle complex linguistic structure and semantics.

With the rapid development of deep learning technology, pre-trained models are gradually emerging in the NLP field. The pre-trained model is an unsupervised learning method, which is pre-trained with a large-scale corpus and fine-tuning on specific tasks to improve the generalization ability of the model. Among them, the BERT (Bidirectional Encoder Representations from Transformers) model is a revolutionary pre-training model proposed in 2018, which has attracted wide attention and application.

The innovation of the BERT model is the Transformer structure and a bidirectional context representation. Through the word-word and sentence-level tasks in the pre-training stage, the BERT model can learn rich semantic information and context relations. During the fine-tuning of specific tasks, the BERT model can solve various NLP tasks by adding a simple classifier, such as named entity recognition, sentiment analysis, and question-answering systems.

As the BERT model has achieved excellent performance on multiple tasks, it has received much attention in both academia and industry. Researchers have proposed various improved and extended models, such as XLNet, RoBERTa, ALBERT, and ELECTRA. These models further optimize the pre-training process, improve the model structure, or introduce new training objectives to achieve better performance.

The rise of pre-trained models not only improves the effect of NLP tasks but also lowers the threshold for research and

application. Researchers can use pre-trained models to quickly build and train their models to focus on the details and innovation of specific tasks. At the same time, the pre-training model also provides a powerful tool for the industry, which can be used to build an intelligent customer service system, intelligent translation system, and intelligent search engine.

However, pre-trained models still face some challenges, such as increased model size and training cost, lack of model interpretability and interpretability, etc. In the future, researchers need to further optimize the training algorithm and model structure of the pre-trained model to improve the efficiency and interpretability of the model, and to apply it to more NLP tasks and application scenarios [1].

### *C. Introduction and Application of the BERT Model*

BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained model based on the Transformer architecture and was proposed by Google in 2018. Compared with the traditional one-way language model, BERT can simultaneously use the context information on the left and right sides, to better understand and generate natural language.

The core idea of the BERT model is to achieve excellent language representation learning through two stages: pre-training and fine-tuning. In the pre-training phase, BERT uses large-scale label-free text data to predict masked words through self-supervised learning. Such a pre-training task enables the BERT model to learn rich language knowledge and contextual information. In the fine-tuning phase, the BERT model applies pre-trained language representations to specific natural language processing tasks by performing supervised fine-tuning on specific tasks, such as text classification, named entity recognition, etc.

The BERT model is widely used, covering multiple tasks in the field of natural language processing. For example, in the text classification task, the BERT model can achieve more accurate classification results by fine-tuning. In the named entity recognition task, the BERT model can identify the human names, place names, and organization names in the text. In the question-answering system, the BERT model can understand the questions and generate accurate answers. In addition, the BERT model has also achieved significant results in machine translation, semantic similarity calculation, emotion analysis, and other tasks.

The emergence of BERT models has had a significant impact on the field of natural language processing. It not only improves the performance of various tasks but also promotes the research boom of pre-trained models. Since then, many improvements and variants based on BERT model have been successively proposed, such as XLNet, RoBERTa, ALBERT, and ELECTRA. These models have improved performance and efficiency, and enriched research in the field of natural language processing.

In conclusion, the BERT model has become an important milestone in natural language processing with its powerful language representation learning ability and wide application field. Its appearance not only improves the performance of various tasks but also provides valuable experience and enlightenment for subsequent pre-training model research. With the continuous progress of technology, the application prospect of the pre-training model in natural language processing will be even broader.

### *D. Advantages and Disadvantages of the BERT model*

BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained language model based on the Transformer model that has achieved great success in natural language processing. Before introducing the advantages of the BERT model, let's look at the shortcomings of the BERT model.

First, the training cost of the BERT model is very high. Due to the large scale of the BERT model, considerable computing resources and time are needed to complete the training. This makes the BERT model difficult to apply in some resource-constrained settings.

Second, the BERT model has constraints on the input sequence length. Since the BERT model adopts a self-attention mechanism, the length of the input sequence can affect the computation amount and memory consumption of the model. Therefore, for longer text sequences, the BERT model may not be handled efficiently.

Furthermore, the pre-training process of the BERT model is unsupervised and therefore cannot be directly applied to a specific task. In practice, the BERT model needs to be combined with specific tasks through fine-tuning, which requires extensive annotation data and computational resources.

However, despite the above shortcomings, the BERT model still has many advantages.

First, the BERT model employs a bidirectional coding mechanism to fully exploit contextual information. Compared with the traditional one-way models, the BERT model can better understand the semantics and relationships in sentences, to extract richer feature representations.

Second, the BERT model can adapt to different natural language processing tasks through pre-training and fine-tuning. By pre-training on large-scale unsupervised data, the BERT model learns rich language knowledge, which can serve as a universal language representation model, and then quickly adapts to different task requirements by fine-tuning specific tasks.

In addition, the BERT model also has strong generalization ability. Since the BERT model is pre-trained on large-scale data, the learned linguistic representations have certain universality and can be applied to natural language processing tasks in different domains and languages.

In conclusion, although the BERT model has some shortcomings, its advantages in the field of natural language processing are obvious. With the improvement and optimization of BERT model, the application prospect of the pre-trained model in natural language processing will be even broader.

## III. THE PROGRESS AFTER BERT

### A. XLNet Model

*a) Model Principle*

XLNet (eXtreme Language Model) is a pre-trained model based on autoregressive and self-encoding, and it is an important advance after BERT. Compared with the autoregressive pre-training method of the BERT model, XLNet adopts the self-encoding pre-training method, which can better handle the language modeling task [2].

The core idea of the XLNet model is to model the joint probability distribution of sentences by maximizing the pre-training targets for all possible permutations. Specifically, XLNet introduces a new modeling approach, namely Permutation Language Modeling (PLM), which considers all possible permutations. Unlike Masked Language Modeling (MLM) in the BERT model, which can only consider part of context information, XLNet can take into account all context information simultaneously through Permutation Language Modeling to better capture dependencies in sentences.

The model structure of XLNet is similar to BERT and is also based on the Transformer architecture. It includes multiple layers of Transformer encoders, where each layer consists of a multi-head self-attention mechanism and a feedforward neural network. Differently, XLNet introduces positional information into the input of the encoder so that the model can better understand the order of the sentences.

In the pre-training phase, XLNet first generates arrangements of candidates by self-encoding and then trains the model by maximum likelihood estimation, enabling the model to predict the correct arrangement. During the fine-tuning phase, XLNet adapts to different downstream tasks by performing supervised fine-tuning on specific tasks.

Experimental results show that XLNet has achieved excellent performance in multiple natural language processing tasks, including text classification, named entity recognition, sentence relationship judgment, etc. Compared with the BERT model, XLNet can better capture the dependencies in the sentence, and improve the expression ability and generalization ability of the model.

In short, XLNet introduced Permutation Language Modeling and adopted self-encoded pre-training objectives, which could better handle language modeling tasks, and achieved significant effect improvement. Its emergence brings new possibilities for the research and application of the natural language processing field.

***B) Experimental Results and Application***

In the XLNet model, the researchers confirmed the superior performance of the model on multiple tasks by evaluating it on various natural language processing tasks. In the question-answering task, XLNet achieved the best results on the SQuAD 2.0 dataset, surpassing the other pre-trained models. In the text classification task, XLNet achieved the best results on each sub-task in the GLUE (General Language Understanding Evaluation) benchmark. These results demonstrate that the XLNet model is efficient in understanding and reasoning about the semantics and context of natural language.

In addition to its advantages in task performance, XLNet has shown great potential in some applications. For example, in

machine translation tasks, the XLNet model can better handle the complexity of semantic and word order, providing more accurate translation results. In the information retrieval task, the XLNet model can better understand the query intentions and provide more relevant search results. In the automated summary task, the XLNet model can capture key information in the text and generate more accurate summaries. These applications demonstrate the broad applicability and effects of XLNet models in a variety of natural language processing tasks.

Moreover, the training methods and techniques of the XLNet model also guide the improvement of other pre-trained models. For example, the autoregressive and self-encoding methods in XLNet can be used for the training of other models to improve the expression and generalization ability of the models. The success of the XLNet model also encourages more researchers to study the pre-trained model deeply and promotes the development and innovation of the field of the pre-trained model.

Despite the remarkable results of XLNet models on multiple tasks and applications, there is still some room for challenges and improvement. For example, the XLNet model is expensive to train and requires significant computational resources and time. Moreover, the interpretability and interpretability of XLNet models is also an important issue because they are crucial for the credibility and reliability of the models. Future studies could aim to address these issues and further improve the performance and scope of pre-trained models.

All in all, the XLNet model demonstrates extraordinary performance and potential in natural language processing tasks and applications. Its success not only drives the field of natural language processing but also provides an important reference and enlightenment for the research and improvement of other pre-trained models. Through continuous improvement and innovation, the pre-training model is expected to further improve the ability and effect of natural language processing in the future.

**B. RoBERTa Model**

*a) Model Principle*

RoBERTa (Robustly Optimized BERT Approach) is a pre-trained model improved based on the BERT model. Its goal is to improve the performance and robustness of the model by optimizing the training process and model architecture.

RoBERTa uses architecture similar to BERT, including a multi-layer Transformer encoder and Masked Language Model (MM) tasks. However, RoBERTa made a series of improvements during training.

First, RoBERTa used a larger training dataset using 160GB of text data for pre-training, compared to only 16GB for BERT. The advantage is to better capture the statistical characteristics of the language and improve the generalization ability of the model.

Second, Roberta has optimized the training process. It removes the Next Sentence Prediction (NSP) task in the BERT because, in practice, the NSP tasks are not always useful. Meanwhile, Roberta adopts longer training steps and a larger batch size to increase the training time of the model and the utilization efficiency of the training data.

In addition, Roberta also fine-tunes the model architecture. It takes a larger model size, including more layers and more hidden units, to increase the representation power of the model. Also, Roberta uses longer maximum sequence lengths to handle longer text input.

With these improvements, RoBERTa achieves excellent performance on multiple natural language processing tasks. It surpassed BERT in public review missions and even surpassed human performance on some missions. This suggests that Roberta has greater representation and generalization in language comprehension and generation tasks.

In conclusion, Roberta improves the performance and robustness of the pre-trained model by optimizing the training process and model architecture. Its success provides an important reference and reference for the research and application in the field of natural language processing [3].

*b) Experiment Results and Application*

Roberta The model is improved based on the BERT model. By optimizing the training process and adjusting the hyperparameters, it has achieved a series of remarkable experimental results and application effects.

First, Roberta has achieved significant performance improvements on multiple natural language processing tasks. On the GLUE benchmark dataset, RoBERTa outperformed the BERT model on 10 of 11 tasks, with 1.5 percentage points on the MNLI task and the F1 score by 1.7 percentage points on the SQuAD task. These results suggest that the RoBERTa model is more expressive and generalized in various text tasks.

Second, the training effect of RoBERTa on large-scale pre-training data was also verified. Compared with BERT, RoBERTa achieved better results by using more data and longer training time for pre-training. By training with 160GB of text data, RoBERTa has achieved great improvement in all tasks. This suggests that increasing the data volume and training time can further improve the performance of the pre-training model.

In addition, the RoBERTa model has demonstrated advantages in some domain-specific tasks. For example, in the medical text classification task, RoBERTa performs better than BERT, with a higher accuracy and recall rate. In the emotion analysis task, RoBERTa was able to more accurately capture subtle changes in emotion, thus improving the classification accuracy.

The application of the RoBERTa model is also not limited to natural language processing tasks. It can also be used for tasks in other fields, such as computer vision, speech recognition, etc. The performance of the model can be further improved by jointly training the RoBERTa model with the image-processing model or the speech-processing model.

Overall, the Roberta model achieved remarkable experimental results and application effects by optimizing the training process and adjusting the hyperparameters. It has achieved a high accuracy rate and F1 score in multiple natural language processing tasks, showing stronger expression ability and generalization ability. Moreover, the Roberta model also shows advantages in domain-specific tasks and can be extended to tasks in other domains.

### *C. ALBERT Model*

*a) Model Principle*

ALBERT (A Lite BERT) is a lightweight pre-trained model that reduces model size and training cost through parameter sharing and sentence-level embedding while maintaining performance comparable to BERT. The model principle of ALBERT mainly includes two key technologies: parameter sharing and sentence-level embedding.

First, parameter sharing refers to the parameters sharing all layers in ALBERT. In the traditional BERT model, each layer has its independent parameters, resulting in a large model and a long training time. However, ALBERT greatly reduces the size of the model by sharing the parameters of all the layers. Specifically, ALBERT uses a parametric matrix to represent the weights of each layer and maps the input vector to the hidden layer via a linear transformation. This parameter-sharing method not only reduces the size of the model but also reduces the number of training parameters, thus accelerating the training speed.

Second, ALBERT uses a sentence-level embedding method, in which the whole sentence is embedded as a unit. In the traditional BERT model, the sentences are divided into multiple segments, each undergoing independent embedding and attention computations. This partitioning approach leads to a large amount of parameter redundancy and computational complexity. While ALBERT beds the whole sentence as a unit, reducing parameter redundancy and can better capture semantic information at the sentence level.

Overall, ALBERT reduces the model size and training cost through parameter sharing and sentence-level embeddings, while maintaining a performance comparable to BERT. This makes ALBERT a very attractive pre-trained model that can be widely used in a variety of natural language processing tasks. However, ALBERT still has some challenges, such as how to further optimize the performance of the model and explain the pre-training process of the model, which still need further research and exploration.

*b) Experiment Results and Application*

ALBERT Model (A Lite BERT) is a pre-training model for BERT designed to address the limitations of the BERT model in training scale and computational cost. The ALBERT model reduces the size of the BERT model by 10-fold by sharing parameters and connecting parameters across layers and achieves better performance than BERT on multiple tasks simultaneously.

In terms of experimental results, the ALBERT model demonstrated excellent performance on multiple natural language processing tasks. For example, in the GLUE (General Language Understanding Evaluation) benchmark, ALBERT has surpassed BERT, including tasks such as natural language reasoning, text similarity, and sentiment analysis. In addition, ALBERT has

performed well in the question & answer task in SQuAD (Stanford Question Answering Dataset).

The ALBERT model is also widely used. First, it can be used for various text classification tasks, such as sentiment analysis, spam detection, and news classification, etc. Secondly, ALBERT can also be applied to tasks such as question-answering systems and machine translation, achieving high-quality question-answering and translation results by encoding and decoding input sequences. In addition, ALBERT can also be used for generative tasks, such as text generation and summary generation.

The success of the ALBERT model benefits from its optimization in the model size and training cost. Through the design of parameter sharing and cross-layer parameter connectivity, the ALBERT model greatly reduces the size of the model while reducing the training cost while maintaining performance. This enables the ALBERT model to be more efficiently applied in real-world scenarios, providing more feasible and practical solutions for solving natural language processing tasks.

However, the ALBERT model still faces some challenges. First, further optimization of model scale is still a problem, and despite significant improvements in scale by ALBERT, further studies are still needed to improve the efficiency and performance of models. Secondly, the interpretability and interpretability of ALBERT models is also an important issue, especially in some sensitive areas such as healthcare and finance, which have more explanatory requirements for model decisions. Thus, future studies could aim to address these issues and further advance the development of pre-trained models in the field of natural language processing[4]。

**D. ELECTRA Model**

*a) Model Principle*

ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) is a pre-trained natural language processing model that has made some improvements to BERT. The core idea of the ELECTRA model is to use adversarial training to generate more efficient pre-trained representations.

In the ELECTRA model, a generator model is first used to predict the probability of some words replacing with others in the input text. The goal of the generator model is to maximize the probability of correctly predicting the replaced words. Then, a discriminant model is used to determine whether the words in the input text are the words in the original text or are generated by the generator model. The goal of the discriminator model is to maximize the probability of correct judgment correctly. Through the adversarial training, the generator model and the discriminator model compete with each other, thus prompting the generator model to generate more realistic replacement words. At the same time, the discriminator model can also better judge the authenticity of the words in the text.

The advantage of the ELECTRA model is that it can effectively use large-scale unsupervised data for pre-training, and the prediction task of the generator model is more simple and efficient, making the training process more efficient. Compared with the BERT model, the ELECTRA model achieves better performance with the same computing resources. The experimental results show that ELECTRA achieves excellent performance in multiple natural language processing tasks, such as text classification, named entity recognition, and question and answer.

However, the ELECTRA model also has several challenges. First, training ELECTRA models requires a lot of computational resources and time, especially on large-scale datasets. Second, the ELECTRA model lacks explicit control over the balance between the generator model and the discriminator model, which may contribute to decreased model performance. Moreover, the adaptability of ELECTRA models for specific tasks needs to be further studied and explored.

In conclusion, the ELECTRA model generates more effective pre-training representations through adversarial training and achieves better performance. However, further studies are still needed to address its existing challenges and further optimize the model performance and range of applications.

*b) Experimental Results and Applications*

ELECTRA Is a new pre-trained model that improves the effect of the generative model by using adversarial training. In ELECTRA's experiment, the researchers found that it achieved very good results on multiple natural language processing tasks.

First, the performance of ELECTRA outperformed other pre-trained models on multiple datasets in the text classification task. For example, on the GLUE dataset, the ELECTRA increased by 2.1 percentage points relative to the BERT model and by 1.3 percentage points relative to the Roberta model. This suggests that ELECTRA has better capabilities in understanding and

classifying texts.

Second, ELECTRA also performed well in the question-and-answer task. On the SQuAD 2.0 dataset, ELECTRA improved by 3.2 percentage points over the BERT model and by 4.1 percentage points over the BERT model on the Natural Questions dataset. This shows the advantage of ELECTRA in understanding and answering questions.

In addition, ELECTRA has achieved remarkable results in tasks such as named entity recognition, sentiment analysis, and machine translation. In these tasks, ELECTRA achieved better performance compared to the other pre-trained models.

In addition to its advantages in task performance, ELECTRA has some potential for practical applications. For example, it can be used for automated text generation, such as the automatic generation of articles, abstracts, dialogue systems, etc. In addition, ELECTRA can also be used for information retrieval and recommendation systems to help users better understand and obtain information.

However, despite its remarkable results on multiple tasks, ELECTRA still faces several challenges. For example, the training cost of models is high and requires large computational resources and time. Moreover, the interpretability and interpretability of ELECTRA remains an issue and needs further research and improvement.

Overall, ELECTRA is a very potential pre-trained model that achieves excellent results on multiple natural language processing tasks. In the future, we can expect more research and application of ELECTRA to further advance the field of natural language processing.

## V. IMPROVEMENT AND APPLICATION OF THE PRE-TRAINING MODEL

### *A. Fine-Tuning Method of the Pre-Training Model*

The fine-tuning method of pre-trained models refers to further training through supervised tasks based on models already unsupervised and pre-trained on a large scale. This process aims to apply the generalization ability of the pre-trained model to specific tasks to improve model performance.

The first step in fine-tuning is to select an appropriate supervised task, such as text classification, named entity recognition, sentiment analysis, etc. Then, the pre-trained model was used as the initial model and further trained with the annotated training data. During fine-tuning, the parameters of the pre-trained model are often adjusted to better adapt it to a specific task [5].

The key to fine-tuning methods is training on limited annotation data. Often, annotation data is less, so some strategies are needed to address this challenge. A common strategy is to use transfer learning, where pre-trained models migrate knowledge learned to specific tasks on large-scale unsupervised data. Through transfer learning, we can reduce the dependence on a large amount of annotated data and improve the generalization ability of the model.

In addition, to further improve the effect of fine-tuning, some techniques can be used. For example, using a larger batch size, smaller learning rates, longer training times, etc. In addition, fine-tuning performance can be improved by adjusting the architecture of the model or adding the regularization method.

Fine-tuning methods have achieved remarkable success in natural language processing. By fine-tuning using a pre-trained model, better results can be obtained than traditional methods on many tasks. However, fine-tuning methods still face some challenges, such as the scarcity of annotated data and the overfitting of models. Future studies could explore more efficient fine-tuning methods to address these challenges and further improve the performance and range of application of the pre-trained models.

### *B. Domain Adaptation Method of the Pre-Trained Model*

The domain adaptation approach for pre-trained models refers to applying generic pre-trained models to domain-specific tasks to improve the performance of the model in the domain. Since the pre-trained model is trained on a large-scale generics corpus, there may be some adaptive problems when facing domain-specific tasks. To address this problem, researchers have proposed a range of domain adaptation approaches.

A common domain adaptation method is domain-specific pre-training (Domain-Specific Pretraining), which is pre-training on domain-specific data. By pre-training on domain-specific data, the model is better adapted to tasks in the domain. For

example, pre-training can be performed with a domain-specific corpus, or with supervised pre-training with domain-specific task data.

Another common approach of domain adaptation is domain adaptation (Domain Adaptation), which adapts to domain tasks by fine-tuning the data based on a common pre-trained model. During fine-tuning, field-specific annotation data can be used to adjust the model parameters to better adapt them to the tasks in the field. Fine-tuning can be done on domain-specific data or domain-specific task data.

In addition, there are some other domain adaptation methods, such as the domain-specific layer (Domain-Specific Layers), the domain-specific attention (Domain-Specific Attention), etc. These methods enhance the model's ability to represent a specific domain by introducing domain-specific layers or attention mechanisms into the model.

In conclusion, the domain adaptation method of pre-trained models can improve the performance of the model on domain-specific tasks through domain-specific pretraining, domain adaptation, and other technical means. The choice of these methods is related to the characteristics of the actual task and the availability of the data. Future studies could also further explore more effective domain adaptation methods to improve the application ability of the pre-trained models in various domain tasks [6].

### *C. Transfer Learning Method of the Pre-Trained Model*

*The Transfer Learning Method of the Pre-Trained Model*

With the development of natural language processing fields, transfer learning methods for pre-trained models have become increasingly important. Transfer learning is the process of applying knowledge learned on one task to another. In NLP, pre-trained models can be used to improve performance on specific tasks through transfer learning.

First, a common transfer learning approach is implemented by using the pre-trained model as a feature extractor. The pre-trained model can map the text data into a high-dimensional vector space that can be used as input features for other tasks. By fixing the parameters of the pre-trained model and training only the classifiers for a specific task, better performance can be obtained with a small amount of annotated data.

Second, another transfer learning method is to fine-tune based on the pre-trained model. Fine-tuning refers to the parameters of continuing to train the pre-trained model on the annotated data for a specific task. By fine-tuning, the pre-trained model can better adapt to the task-specific feature and label distribution, thus improving performance. Small learning rates can be used for fine-tuning to avoid overadjusting the parameters of the pre-trained model.

In addition, there are some other transfer learning methods. For example, multi-task learning methods can be trained on multiple related tasks simultaneously, sharing the parameters of the pre-trained model to improve performance. Domain adaptation methods can adapt to features and distributions in different domains by training pre-trained models on the source domain and fine-tuning the model on the target domain [7].

However, the transfer learning of the pre-trained models also faces several challenges. First, differences between tasks may lead to poor transfer learning. Secondly, the pre-trained model has a larger scale and a higher training cost. Moreover, the interpretability and interpretability of the pre-trained model is also a challenge, limiting its application in some sensitive areas.

In the future, the transfer learning of pre-trained models still has a broad space for development. Researchers can further explore more effective transfer learning methods to solve the adaptability of pre-trained models on specific tasks. Moreover, it can be possible to study how to reduce the size and training cost of the pre-trained model, as well as improve the interpretability of the model to meet the needs of different scenarios.

### *D. Multi-Task Learning Method of the Pre-Trained Model*

*Multi-Task Learning Approach of the Pre-Trained Models*

The multi-task learning approach of a pre-trained model refers to applying a pre-trained model to multiple related tasks to improve the performance of the model by sharing the underlying representation. This method can fully utilize the knowledge learned by the pre-trained model in large-scale data and reduce the demand for large-scale annotated data while improving the generalization ability of the model.

In multi-task learning, the pre-trained model is designed as a shared underlying network with multiple task-specific head

networks on it. The underlying network is responsible for extracting a shared representation of the input data, while the head network is responsible for handling the specific output of different tasks. By sharing the underlying network, the model can share features between different tasks, thus improving the efficiency and performance of the model.

An important advantage of the multi-task learning approach is the generalization of the models by sharing the underlying networks. When the training data for a task is small, the training data for other tasks can be exploited by sharing the underlying network, thus improving learning from the task. Moreover, multi-task learning can also reduce the risk of overfitting the model, because it must adapt to multiple tasks simultaneously, not just a single task.

However, there are also some challenges in multitasking learning. First, the correlation between tasks needs to be rationally selected and designed to ensure the validity of the shared underlying network. If the correlation between tasks is weak, the shared underlying network may not provide a significant performance boost. Second, multi-task learning requires more computational resources and time due to the processing of multiple tasks simultaneously. Moreover, multi-task learning also requires careful design of loss functions to balance the importance of different tasks.

In the future, multi-task learning methods of pre-trained models will continue to be explored and developed. Researchers can further explore how to select and design tasks to maximize the performance of the pre-trained model. Furthermore, further research can be conducted on how to adaptively adjust the weights between tasks to respond to changes in different tasks. A multi-task learning approach of pre-trained models will help to drive progress and application in the field of natural language processing [8].

## VI. CHALLENGES AND PROSPECTS OF THE PRE-TRAINING MODEL

### *A. Challenge of Model Size and Training Cost*

Model size and training cost an important challenge in NLP. As the size of the model increases, the number of parameters increases accordingly, which leads to increased computational resource requirements for training and inference? At the same time, as the size of the dataset expands, the time required to train the pre-training model will greatly increase.

First, increasing model size poses challenges for computational resources. Pre-trained models typically consist of billions or even tens of billions of parameters that require a lot of computational resources for training and inference. For large-scale models, the use of distributed computing and high-performance computing devices, such as GPU and TPU, is needed to accelerate the training and inference process. However, the cost of these computational resources is very high and may be difficult for ordinary researchers and enterprises to afford.

Second, the time required to train the pre-training model is also a challenge. Due to the increasing model size, the time required to train the pre-training model will increase accordingly. For example, the training of the BERT model may take several days or even several weeks. This is a huge time cost for researchers and engineers, limiting the speed of model iteration and optimization.

To meet these challenges, the researchers have proposed some methods to reduce the model size and training cost. One approach is pruning and compression models, reducing the size of the model by removing redundant parameters and using low-precision calculations. Another approach is to use data parallelism and model parallelism to accelerate the training process. In addition, there are some techniques, such as model quantification, knowledge distillation, and transfer learning, which can reduce the demand for computing resources while maintaining the model performance.

Despite these challenges, the application of pre-trained models in natural language processing remains promising. Future research directions could focus on improving model training efficiency and inference speed while maintaining model performance and generalization capabilities. At the same time, more interpretable and interpretable pre-training models can also be explored to enhance the understanding and interpretation of the internal mechanisms of the model.

### *B. Challenges of Model Interpretability and Interpretability*

In NLP, the emergence and development of pre-training models provide a powerful foundation for the solution of various text tasks. However, with the continuous development of pre-trained models, the problems of model interpretability and interpretability have gradually emerged. This problem arises mainly because of the complexity of the pre-trained model and the black-box properties [9].

First, pre-trained models usually have a large number of parameters and multiple layers of neural network structure, which greatly increases the complexity of the model. This makes it difficult to understand how the model makes decisions. For example, in an emotion classification task, the model may make judgments based on the tiny details of the input text that are not important to humans. This leads to the uninterpretability of the model because one cannot accurately understand why the model makes such a decision.

Second, pre-trained models are often trained through large-scale unsupervised learning, meaning that the knowledge learned by the model may be difficult to interpret. The model learns the grammar and semantic rules of a language by observing a large amount of text data, but it is not clear exactly what rules and knowledge are learned. This makes the interpretability of the model difficult because one cannot intuitively understand what the model learns.

To solve the problem of interpretability and interpretability of models, the researchers proposed some methods and techniques. A common approach is to use the attention mechanism to visualize the model decision process to better understand the model attention distribution. Another approach is to use feature importance analysis to evaluate the importance of the model to the input features, thereby explaining the model decision basis. In addition, there are ways to generate explanatory text by generating adversarial networks to help understand the decision-making process of the model.

However, there are some challenges and limitations to the current explanatory approach. First, most of these methods are based on the output information of the model, while the internal mechanism and decision process of the model remain a black box. Second, explanatory methods often require additional computational costs and complex model structures, which bring some difficulties for practical applications. Finally, explanatory methods often provide only local explanations about model decisions and not global explanations.

Therefore, future studies need to further explore and develop more effective methods to improve the interpretability and interpretability of pre-trained models. This will help to improve the reliability and credibility of the model and facilitate the wide application of pre-trained models in practical applications.

**C. Optimization Outlook of the Pre-Trained Model on Specific Tasks**

*Optimization Outlook of Pre-Trained Models on Specific Tasks*

The emergence of the pre-training model has greatly promoted the development of the natural language processing field and achieved remarkable results. However, despite the pre-trained models on multiple tasks, there is room for challenges and improvement. In future studies, we can explore the following directions to further optimize the performance of the pre-trained model on a specific task.

First, most of the current pre-training models are trained by unsupervised learning, but when fine-tuning a specific task, we can consider introducing supervised learning methods. By training on the annotated data for a specific task, the pre-trained model can be more focused on the specific features of the task and improve the performance of the model on the task [10].

Secondly, the scale of the pre-training model is getting larger and the training cost is getting higher and higher. Therefore, how to reduce the model size and improve the model efficiency without sacrificing performance becomes an important research direction. The number of parameters of the model can be reduced by pruning, quantification, and other techniques while maintaining high performance.

In addition, the interpretability and interpretability of pre-trained models on specific tasks is also a focus of attention. The complexity of the pre-trained model makes its decision process difficult to understand and brings challenges to the interpretability of the model. Therefore, we can explore how to design more interpretable pre-trained models to better understand the decision process of the model.

Moreover, the generalization ability of the pre-trained model is also an important issue. Current pre-trained models are often trained on a large-scale generalist corpus, but the ability to generalize on specific tasks may be limited. Therefore, how to improve the performance of pre-trained models on specific tasks by introducing more domain-specific data is a worthwhile direction.

In conclusion, the optimization of pre-trained models on a specific task is a challenging problem, but also full of opportunities. By introducing supervised learning, reducing model size, improving interpretability, and increasing domain-

specific data, we can expect to further optimize the performance of pre-trained models on specific tasks and promote the development of natural language processing fields.

## VII. CONCLUSION

In the field of natural language processing, the emergence of pre-trained models has caused great attention and research upsurge. As a representative model, BERT has achieved remarkable results. However, with the deepening of research, some limitations of the BERT model, such as high training cost and large model size. Therefore, the researchers began to explore the progress after BERT, to further improve the performance of the pre-trained model.

In the present paper, we summarize and analyze some of the important developments after BERT. First, we introduce some improved versions of BERT, such as XLNet, RoBERTa, ALBERT, and ELECTRA. These models have been improved based on BERT and achieved better results. Among them, XLNet model introduces the idea of the permutation language model, RoBERTa model improves performance through larger training data and longer training time, ALBERT model reduces model scale through parameter sharing and decomposition of attention mechanism, and ELECTRA model proposes a new generative adversarial network framework. The emergence of these models not only improves the performance of the pre-trained models but also provides new ideas and directions for subsequent studies.

Second, we explore the improvement and application of the pre-trained model. To solve the problem of high training cost and large model scale of the pre-training model, the researchers proposed some methods of fine-tuning, domain adaptation, transfer learning, and multi-task learning. These methods can reduce the training cost and the model size while maintaining the model performance. In addition, we discuss the challenges and prospects of pre-trained models, including the challenges of model size and training cost, the challenges of model interpretability and interpretability, and the optimization prospects on specific tasks. These problems are urgent problems to be solved in current research, and also the direction of future research.

In conclusion, the pre-trained models in NLP have progressed substantially after BERT. The emergence of new models and methods not only improves the performance of the pre-trained models but also brings new opportunities and challenges for subsequent studies. We believe that shortly, the pre-trained models will play an important role in more domains and tasks, and provide people with better natural language processing solutions.

## VIII. REFERENCE


[1] Li Qiong, Chen Jingxian, Wu Yuan, Lv Lingling, Ying Haifeng, Zhu Wenhua, Xu Jiayue, Ruan Ming, Guo Yuanbiao, Zhu Weirong, Yin Shipeng, Zheng Lan. Pharmacological study of the mechanism of colorectal cancer liver metastasis based on TCGA database [J]. Journal of Gansu University of Traditional Chinese Medicine, 2023,40 (03): 1-11.

[2] LI Zhi-hao,HU Jun-wei,LI Xu,CHEN Yue-lai. Mechanism progress of acupuncture and moxibustion for chronic prostatitis: Progress in the research mechanism of acupuncture in the treatment of chronic prostatitis [J].World Journal of Acupuncture – Moxibustion,2021,31(4).

[3] Julia D. Hammer, Christopher Palma, KeriAnn Rubin, Alice Flarend,Yann Shiou Ong,Chrysta Ghent,Timothy Gleason,Scott McDonald,Brandon Botzer,Tanya Furman.Evaluating a learning progression for the solar system: Progress along gravity and dynamical properties dimensions[J].Science Education,2020,104(3).

[4] Ponna. Research on the intelligent identification method of chemical bond energy data for scientific papers [D]. University of Chinese Academy of Sciences (Literature and Intelligence Center of Chinese Academy of Sciences), 2020.

[5] Zhang WeiYu,Zhou JianHua,Wang HuanRui,Mu Qing,Wang Qi,Xu KeXin,Xu Tao,Hu Hao.[Research Progress of the Roles of Ubiquitination/Deubiquitination in Androgen Receptor Abnormalities and Prostate Cancer].[J].Zhongguo yi xue ke xue yuan xue bao.Acta Academiae Medicinae Sinicae,2020,42(2).

[6] Benefit of Tolvaptan on Time to End-stage Renal Disease (ESRD) for Patients with Rapidly Progressing Autosomal Dominant Polycystic Kidney Disease (ADPKD): A Disease Progression Model[J].American Journal of Kidney Diseases,2020,75(4).

[7] Tao Yaoye,Wang Jianguo, Xu Xiao.Emerging and Innovative Theranostic Approaches for Mesoporous Silica Nanoparticles in Hepatocellular Carcinoma: Current Status and Advances.[J].Frontiers in bioengineering and biotechnology,2020,8.

[8] Sofia Ajeganova, Thomas Gustafsson, Linnea Lindberg, Ingiäld Hafström, Johan Frostegård.P159 Similar progression of carotid intima-media thickness in 7-year surveillance of patients with mild SLE and controls, but this progression in patients is still promoted by dyslipidemia, hypertension, history of lupus nephritis and a higher prednisone usage[J].Lupus Science & Medicine,2020,7(Suppl 1).

[9] Agriculture; Findings from Anhui University Advance Knowledge in Agriculture (Recent advances in Raman technology with applications in agriculture, food, and biosystems: A review)[J].Journal of Engineering,2020.

[10] Disease Attributes - Disease Progression; Researchers from Fuwai Hospital Report on Findings in Disease Progression (Elevated Plasma Beta-hydroxybutyrate Predicts Adverse Outcomes and Disease Progression In Patients With Arrhythmogenic Cardiomyopathy)[J].News of Science,2020.