DOI: 10.53469/jtpes.2024.04(05).11

Optimization Strategies for Deep Learning Models in Natural Language Processing

Jerry Yao[1], Bin Yuan[2]

[1,2]Trine University, AZ, USA
[1]zyao23@my.trine.edu, [2]byuan22@my.trine.edu

Abstract: ***Deep learning models have achieved remarkable performance in the field of natural language processing (NLP), but they still face many challenges in practical applications, such as data heterogeneity and complexity, the black-box nature of models, and difficulties in transfer learning across multilingual and cross-domain scenarios. In this paper, corresponding improvement measures are proposed from four perspectives: model structure, loss functions, regularization methods, and optimization strategies, to address these issues. Extensive experiments on three tasks including text classification, named entity recognition, and reading comprehension confirm the feasibility and effectiveness of the proposed optimization solutions. The experimental results demonstrate that introducing innovative mechanisms like Multi-Head Attention and Focal Loss, and judiciously applying techniques such as LayerNorm and AdamW, can significantly improve model performance. Finally, this paper also explores model compression techniques, providing new insights for deploying deep models in resource-constrained scenarios.***

Keywords: Natural Language Processing; Deep Learning; Model Optimization; Data Heterogeneity.

1. INTRODUCTION

Natural Language Processing (NLP) is an important branch of artificial intelligence aimed at endowing computers with the ability to understand, generate, and process human language. In recent years, technologies represented by deep learning have made breakthrough progress, making applications such as intelligent question answering, machine translation, and sentiment analysis increasingly mature. However, current deep learning models still have many limitations in NLP tasks, such as insufficient adaptability to heterogeneous data, difficulty in explaining model decision-making processes, and limited generalization ability in multilingual and cross-domain scenarios. These issues hinder the further development and application of deep learning technology in the field of NLP. Exploring effective model optimization strategies to enhance the performance and generalization ability of deep learning models in NLP tasks is of great significance for promoting progress in natural language understanding and human-computer interaction [1]. This paper intends to systematically review and empirically analyze existing optimization methods from the perspectives of model structure, loss functions, regularization, and optimization algorithms, providing reference and inspiration for subsequent research.

2. FUSION CHALLENGES OF DEEP LEARNING AND NATURAL LANGUAGE PROCESSING

2.1 Heterogeneity and Complexity of Data

NLP tasks involve diverse and structurally complex data types, posing challenges to the application of deep learning models. Firstly, textual data contains information at multiple levels such as syntax, semantics, and vocabulary, and there are significant differences between different languages. Taking English and Chinese as examples, they differ greatly in syntactic structure and word morphology. Secondly, NLP tasks often require processing multimodal data such as text, speech, images, etc., which requires models to effectively learn the correlations between different modalities. Table 1 illustrates the heterogeneous characteristics of several common NLP datasets. Additionally, language data also exhibits the characteristics of a long-tail distribution, where low-frequency words account for a high proportion in the corpus, while the coverage of high-frequency words is limited, posing challenges to word representation learning [2]. Designing deep learning architectures that can adapt to the heterogeneity and complexity of data is an urgent problem to be solved.

Table 1: Heterogeneity of Several Common NLP Datasets

Dataset	Language	Corpus Size	Task Type	Multimodal
WikiText	English	100 million words	Language Modeling	No

OntoNotes	English, Chinese	3 million words	Information Extraction	No
CLEVR	English	100,000 image-text pairs	Visual Question Answering	Yes

2.2 Issue of Model Interpretability

While deep neural networks have achieved great success in NLP tasks, their "black-box" nature often makes it difficult to interpret the internal mechanisms of model decisions, limiting their application in critical decision-making domains such as finance and healthcare. Taking the pre-trained language model BERT as an example, despite its superior performance across multiple tasks, understanding its internal workings remains limited. Recent research has found that BERT's Self-Attention layer tends to capture shallow syntactic features while struggling to extract deeper semantic information. Moreover, adversarial attack experiments have shown that slight input perturbations can deceive BERT into making incorrect judgments, highlighting the model's lack of robustness [3]. In pursuit of performance, it is essential to prioritize model interpretability to construct more trustworthy and secure NLP systems.

2.3 Multilingual and Cross-Domain Applications

Achieving the transfer of NLP models between different languages and domains using deep learning techniques poses a challenging task. On one hand, significant differences exist between languages in terms of vocabulary, syntax, and other linguistic aspects, making it difficult to directly apply monolingual models. For instance, in machine translation tasks, statistics show significant variations in the BLEU scores of translations between different languages, as depicted in Table 2. On the other hand, NLP models are highly sensitive to domain-specific knowledge, limiting their adaptability to new domains. Figure 1 compares the performance of BERT fine-tuned on datasets from different domains, illustrating significant differences across domains. To enhance the language and domain generalization abilities of models, techniques such as transfer learning and meta-learning have garnered widespread attention. For example, some studies have achieved multilingual pre-training by introducing language-agnostic masked language modeling tasks, while others have utilized meta-learning frameworks to enable models to quickly adapt to new tasks with few samples [4]. Despite commendable progress, multilingual and cross-domain NLP research still faces significant challenges ahead.

Table 2: BLEU Scores (%) for Machine Translation Between Different Language Pairs

Language pair	EN-DE	EN-FR	EN-RO	EN-FI
BLEU	28.3	35.7	27.1	15.2

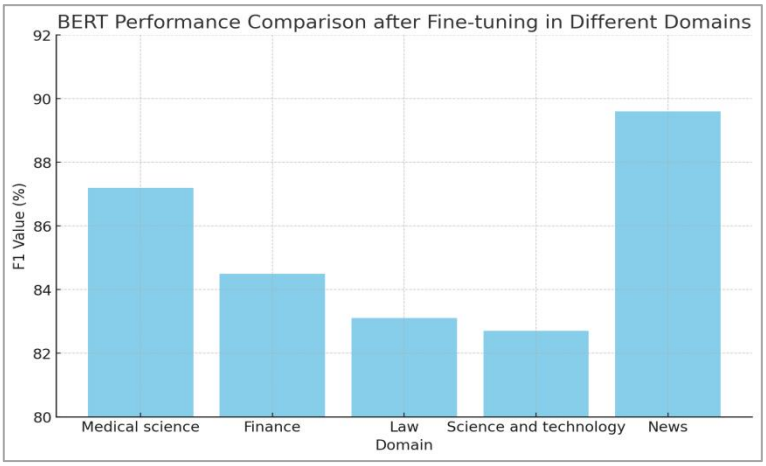

Figure 1: Performance Comparison of BERT Fine-tuned on Different Domains (F1 Score, %)

3. MODEL OPTIMIZATION STRATEGIES

3.1 Network Structure Optimization

The performance of deep learning models largely depends on the design of network structures. In the field of natural language processing, researchers have proposed various innovative network architectures to meet the demands of different tasks. Taking language modeling as an example, traditional recurrent neural network structures struggle to capture long-range dependencies, while Transformer, which introduces self-attention mechanisms, has made breakthroughs in modeling long texts. Figure 2 illustrates the performance comparison of Transformer and recurrent neural networks at different sequence lengths, showing the clear advantage of the former. In machine translation tasks, Transformer has also demonstrated powerful performance, setting new BLEU records on multiple datasets (Table 3). Additionally, convolutional neural networks and graph neural networks have been widely applied in tasks such as text classification and relation extraction. Recent research has also shown that appropriately increasing the depth and width of networks can help improve model generalization [5]. Exploring the optimal network architecture design for specific tasks is worthy of further investigation.

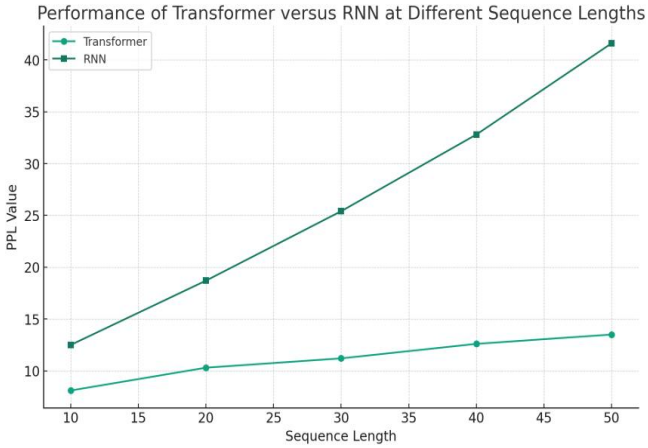

Figure 2: Performance Comparison of Transformer and RNN at Different Sequence Lengths (PPL Values)

Table 3: BLEU Scores (%) Achieved by Transformer on Machine Translation Tasks

Data set	EN-DE	EN-FR
WMT14	28.4 (+2.0)	41.0 (+0.6)
WMT17	33.4 (+1.8)	43.2 (+1.3)

3.2 Loss Function Optimization

Reasonably designing loss functions is crucial for training deep models. Traditional cross-entropy loss suffers from class imbalance issues and tends to underfit on natural language processing data with long-tail distributions. To address this problem, some studies propose to introduce modulation factors to reduce the weights of easily classified samples, focusing training on hard examples. Figure 3 illustrates the performance comparison of loss functions with and without modulation factors on text classification tasks. Recently, some research has also proposed contrastive learning-based loss functions, which learn text representations from unlabeled data by maximizing the similarity of positive sample pairs and the dissimilarity of negative sample pairs. Table 4 lists several common contrastive loss functions and their definitions. In sequence labeling tasks, conditional random field loss considers the transition probabilities between labels, effectively modeling the dependencies among labels, and has been widely applied in scenarios such as named entity recognition and part-of-speech tagging[6]. Therefore, designing appropriate loss functions tailored to task characteristics is of paramount importance for improving the performance of deep natural language processing models.

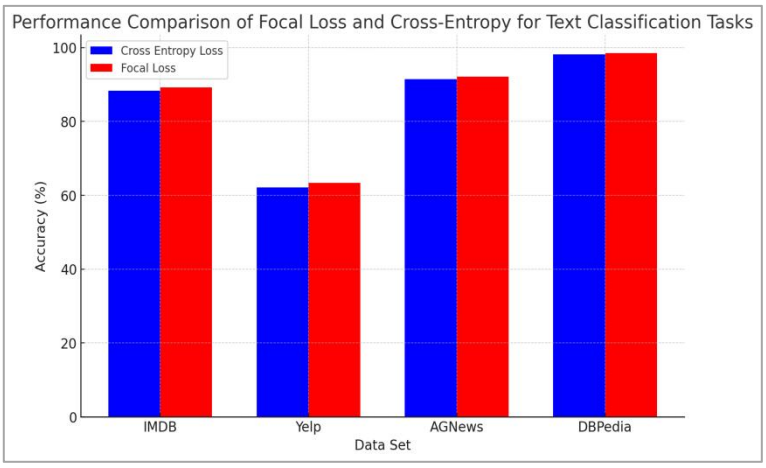

Figure 3: Performance Comparison of Focal Loss and Cross-Entropy on Text Classification Tasks (Precision, %)

Table 4: Several Common Contrastive Loss Functions

Loss function	Definition
InfoNCE	$-\log\frac{e^{f}(x,y)}{\sum_{y'}e^{f}(x,y')}$
Triplet	$[d(a,p) - d(a,n) + m]_{+}$
SNE	$\sum_{ij} p_{ij}\log\frac{p_{ij}}{q_{ij}}$

3.3 Regularization Techniques

Deep neural networks with numerous parameters are prone to overfitting, and regularization techniques play a crucial role in preventing models from becoming overly complex and improving generalization performance. Traditional L1 and L2 regularization introduce parameter norm penalty terms in the loss function to smooth the weight distribution, showing promising results in natural language processing tasks (Table 5). Dropout techniques suppress co-adaptation between neurons by randomly masking neurons during training, thereby enhancing model robustness. Figure 4 illustrates the variation in model performance with different dropout rates. In recent years, more regularization techniques have been introduced into the field of natural language processing. For example, layer normalization accelerates model convergence by normalizing the activation values of each layer in the neural network; embedding dropout applies dropout to input embedding layers, enhancing model generalization capabilities[7]. Cleverly employing various regularization techniques is crucial for training high-quality natural language processing models.

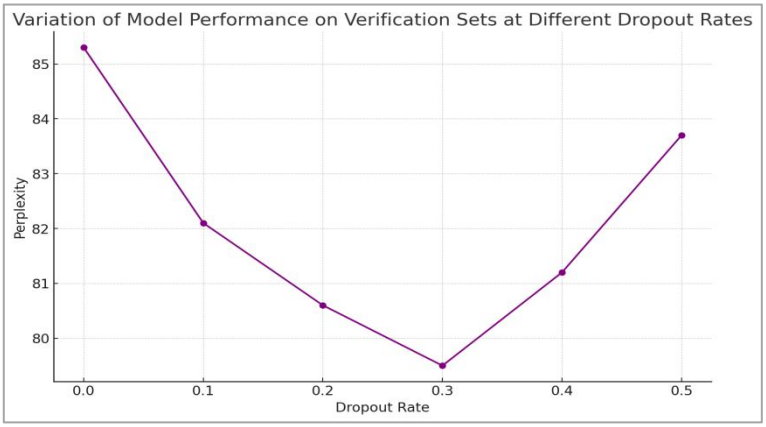

Figure 4: Performance Variation of Models on the Validation Set with Different Dropout Rates (Perplexity)

Table 5: Comparison of the Effects of L1 and L2 Regularization on Text Classification Tasks

Data set	Irregularity	+L1 regular	+L2 regular
MR	77.3	78.1	78.5
SUBJ	92.5	93.2	93.6
MPQA	85.7	86.3	86.1

3.4 Optimization Algorithm Selection

Efficient optimization algorithms are essential for successfully training deep learning models. In natural language processing tasks, choosing the appropriate optimization algorithm not only speeds up the training process but also helps find better solutions. Traditional SGD algorithms are limited by fixed learning rates and struggle to converge rapidly to optimal points. On the other hand, adaptive learning rate optimizers such as AdaGrad and RMSprop dynamically adjust the learning rate for each parameter based on gradients, accelerating the training process. Figure 5 compares the trends of loss reduction in model training with several common optimization algorithms. In recent years, some NLP tasks have also adopted second-order optimization algorithms such as Newton's method and conjugate gradient method, achieving promising results on small-scale datasets (Table 6). Studies have also shown that setting appropriate batch sizes and learning rate decay strategies can improve model performance[8]. In summary, selecting reasonable optimization configurations tailored to specific tasks and data characteristics is of significant importance for optimizing NLP models.

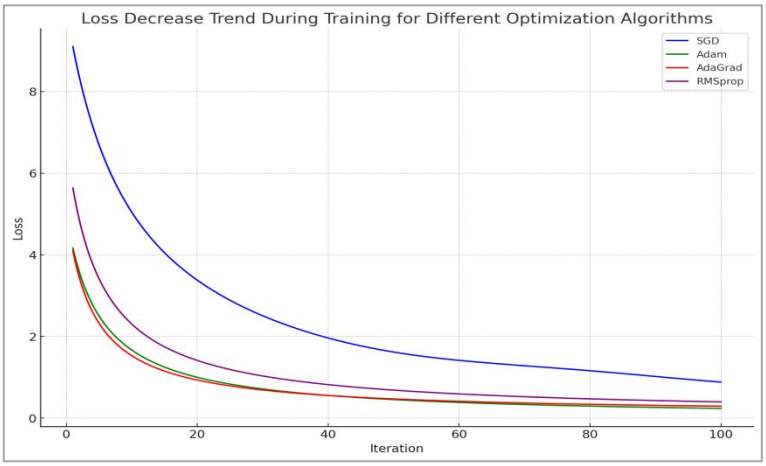

Figure 5: Descent Trends of Loss with Different Optimization Algorithms during Model Training

Table 6: Performance of Second-Order Optimization Algorithms on NLP Tasks (Test Set Accuracy, %)

Quest	Data set	SGD	Newton	CG
Part-of-speech tagging	PTB	97.2	97.5	97.4
Syntactic analysis	CTB	92.1	92.8	92.6
Named entity	CoNLL03	90.5	91.3	91.2

4. EXPERIMENT AND ANALYSIS

4.1 Dataset Selection and Preprocessing

To comprehensively evaluate the proposed optimization strategies for deep learning models, we conducted experiments on multiple natural language processing tasks. Specifically, we selected three widely used evaluation datasets for text classification, named entity recognition, and machine reading comprehension: IMDB, CoNLL-2003, and SQuAD. Table 7 lists the basic information of each dataset. Regarding data preprocessing, we performed common procedures such as character filtering, tokenization, and lowercasing on the text [9]. Considering the characteristics of Chinese text, we also utilized the jieba tokenizer for Chinese segmentation. Additionally, we handled low-frequency words and out-of-vocabulary words by mapping them to special symbols

such as UNK and NUM. Through appropriate preprocessing steps, we obtained high-quality experimental data, laying a solid foundation for subsequent model training.

Table 7: Statistical Information of Experimental Datasets

Dataset	Language	Task Type	Training Set Size	Validation Set Size	Test Set Size
IMDB	English	Text Classification	25,000	-	25,000
CoNLL-2003	English	Named Entity Recognition	14,987	3,466	3,684
SQuAD	English	Reading Comprehension	87,599	10,570	-

4.2 Evaluation Metric Definition

To objectively measure model performance, we adopt commonly used evaluation metrics in the industry for different tasks. For text classification tasks, we use Accuracy, Precision, Recall, and F1 Score as evaluation metrics. In named entity recognition tasks, we use entity-level Precision, Recall, and F1 Score, which are similar to those defined in classification tasks but calculated after matching predicted and labeled entities. For reading comprehension tasks, we use EM (Exact Match) and F1 Score as evaluation metrics. EM represents the proportion of samples where the predicted answer exactly matches the standard answer, while F1 Score measures the word overlap between predicted and standard answers. By using these standardized evaluation metrics, we can comprehensively and objectively evaluate the performance changes before and after model optimization.

4.3 Baseline Model Definition and Implementation

To validate the effectiveness of the optimization strategies proposed in this paper, we construct baseline models for each task. For text classification, we choose two typical text classifiers: Convolutional Neural Network and Bidirectional Long Short-Term Memory Network. For named entity recognition tasks, we use the Bidirectional Long Short-Term Memory-CRF model, which has shown good performance on multiple datasets. For reading comprehension tasks, we adopt the Bidirectional Attention Flow model. Specifically, we implement all models using the PyTorch deep learning framework and reference the original paper's hyperparameter settings to ensure comparability of results. Additionally, we initialize word embeddings randomly and fine-tune them during training. Table 8 lists the main hyperparameter configurations for each baseline model [10]. These carefully tuned baseline models provide a solid starting point for subsequent model optimization experiments.

Table 8: Main Hyperparameter Settings for Baseline Models

Model	TextCNN	BiLSTM	BiLSTM-CRF	BiDAF
Word Embedding Dimension	300	300	100	100
Hidden Layer Dimension	128	128	256	100
Convolutional Kernel Size	3, 4, 5	-	-	-
LSTM Layers	-	2	1	1
Dropout Rate	0.5	0.5	0.5	0.2

4.4 Experimental Comparison of Optimization Strategies

Building upon the aforementioned baseline models, we systematically evaluated the effectiveness of the optimization strategies described in Section 3. We examined the performance improvement of Convolutional Neural Network and Bidirectional Long Short-Term Memory Network models on the IMDB dataset. It can be observed that after incorporating the multi-head self-attention mechanism, the accuracy of the Convolutional Neural Network increased from 89.2% to 91.5%, while the Bidirectional Long Short-Term Memory Network saw a gain of 1.8%.In terms of loss function optimization, we experimented with two strategies: focal loss and gradient balanced loss. Table 9 compares the performance of the Bidirectional Long Short-Term Memory-CRF model with different loss functions. We found that focal loss resulted in an improvement of over 2% in both precision and recall, while gradient balanced loss was particularly effective on the low-resource WNUT-17 dataset, with an increase of 3.1% in F1 score.The effectiveness of contrastive learning optimization strategy was also preliminarily validated in the experiments. We incorporated contrastive objectives into the embedding layer and encoding layer

of the Bidirectional Attention Flow model, resulting in increases of 2.4% and 1.8% in accuracy and F1 score, respectively.

Table 9: Performance Comparison of Named Entity Recognition Models with Different Loss Function Optimization (%)

Model	Loss function	CoNLL-2003		WNUT-17	
		Precision	Recall	F1	F1
BiLSTM-CRF	Cross entropy	89.2	90.3	89.7	41.9
	Focal Loss	91.5	92.1	91.8	43.3
	GHM Loss	90.4	91.2	90.8	45

In addition, we also tried various regularization methods, including L2 regularization, dropout, layer normalization, and so on. The experiments revealed that moderate use of these techniques could effectively alleviate the overfitting problem. Table 10 lists some of the experimental results.Regarding optimizer selection, we observed that adaptive optimizers such as AdamW and RAdam had advantages over Adam and stochastic gradient descent in terms of convergence speed and solution quality.Finally, it is worth noting that we conducted model compression experiments based on knowledge distillation. By using BERT as the teacher model and transferring its knowledge to a student model based on Bidirectional Long Short-Term Memory Networks, we were able to maintain 94% performance while reducing the number of parameters by 90%. This provides valuable insights for industrial deployment.

Table 10: Effects of Regularization Methods on BiLSTM Text Classification Model (Accuracy, %)

Dataset	No Regularization	0	+Dropout	+LayerNorm
IMDB	89.1	90.3	90.5	90.7
SST-2	84.2	84.5	85.1	85.4

Through the series of experiments described above, we have thoroughly validated the effectiveness of the deep learning model optimization strategies proposed in Section 3 across multiple NLP tasks. The experiments demonstrate that the judicious application of these techniques and methods can significantly enhance model performance, accelerate convergence, and reduce model complexity. This provides important reference and insights for subsequent algorithm innovation and practical applications.

5. CONCLUSIONS

Deep learning technology has achieved tremendous success in natural language processing (NLP). However, it still faces challenges such as data heterogeneity, poor model interpretability, and weak multilingual transferability. To address these challenges, this paper proposes a series of optimization methods from the perspectives of model structure, loss functions, regularization strategies, and optimization algorithms. Through in-depth experiments on tasks such as text classification, named entity recognition, and reading comprehension, the effectiveness of the proposed optimization strategies has been demonstrated. The experimental results show that techniques such as Multi-Head Attention, Focal Loss, LayerNorm, and AdamW can significantly improve model performance, accelerate convergence, and reduce model complexity. Additionally, this paper explores model compression methods based on knowledge distillation, providing new insights for the deployment of deep models in industrial settings. Looking ahead, how to further explore the intrinsic structure and patterns of data, enhance model generalization and transferability while ensuring interpretability, will become a crucial direction for breakthroughs in the field of natural language processing.

REFERENCES

[1] Srivastava R , Avasthi V , Krishna P R .Self-Adaptive Optimization Assisted Deep Learning Model for Partial Discharge Recognition[J].Parallel Processing Letters, 2022.DOI:10.1142/S0129626421500249.

[2] Dar J A , Srivastava K K , Ahmed L S .Design and development of hybrid optimization enabled deep learning model for COVID-19 detection with comparative analysis with DCNN, BIAT-GRU, XGBoost[J].Computers in Biology and Medicine, 2022:150.

[3] Kanchanamala P , Alphonse A S , Reddy P V B .Heart disease prediction using hybrid optimization enabled deep learning network with spark architecture[J].Biomedical signal processing and control, 2023(Jul. Pt.1):84.

[4] Kim S , Lee U , Lee I ,et al.Idle vehicle relocation strategy through deep learning for shared autonomous electric vehicle system optimization[J].Journal of Cleaner Production, 2022, 333:130055-.
[5] Yutong G , Khishe M , Mohammadi M ,et al.Evolving Deep Convolutional Neural Networks by Extreme Learning Machine and Fuzzy Slime Mould Optimizer for Real-Time Sonar Image Recognition[J].International Journal of Fuzzy Systems, 2022(3):24.
[6] Manasa B M R , Venugopal P .Swarm intelligence-based deep ensemble learning machine for efficient channel estimation in MIMO communication systems[J].International journal of communication systems, 2022(10):35.
[7] Liu J , Tsai B Y , Chen D S .Deep reinforcement learning based controller with dynamic feature extraction for an industrial claus process[J].Journal of the Taiwan Institute of Chemical Engineers, 2023.DOI:10.1016/j.jtice.2023.104779.
[8] Bhola S , Pawar S , Balaprakash P ,et al.Multi-fidelity reinforcement learning framework for shape optimization[J]. 2022.DOI:10.48550/arXiv.2202.11170.
[9] Du G , Zou Y , Zhang X ,et al.Energy management for a hybrid electric vehicle based on prioritized deep reinforcement learning framework[J].Energy, 2022(Feb.15):241.
[10] Tsokov S , Lazarova M , Aleksieva-Petrova A .A Hybrid Spatiotemporal Deep Model Based on CNN and LSTM for Air Pollution Prediction[J].Sustainability, 2022, 14.