Neurocomputing 595 (2024) 127815

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A head-to-head attention with prompt text augmentation for text classification

Bo Peng, Kundong Han, Liang Zhong, Shengbo Wu, Tao Zhang *

XinJiang University, Urumqi, 830000, China

ARTICLE INFO

Communicated by D. Cavaliere

Keywords:
Short text classification
Attention mechanism
Text augmentation

ABSTRACT

Short-text classification is an important task in natural language processing (NLP). The classification results are unsatisfactory due to the sparsity of Chinese short texts, insufficient annotation data, single tasks, and classification imbalance problems. Therefore, we propose a method based on a self-attention mechanism to focus on the relationships between attention heads, which can improve the performance of models represented by self-attention for short-text classification tasks. In addition, we designed a text augmentation template based on prompt learning with embedded labels. This allows single-task classification to be transformed into multitask classification while allowing the model to focus on the semantic consistency of the labels with the text. We conducted experiments on the CHNSenticorp, COLD, and SST-2 datasets to achieve better results than several popular text classification methods.

1. Introduction

Most of the current effective methods for text classification rely on extensively labeled data and a large number of parameters. However, these models are unsuitable when supervised training data are scarce or when it is challenging to collect them. Owing to the widespread adoption of attention mechanisms in natural language processing, particularly because of their popularity in transformer architectures (Vaswani et al. 2017), multiheaded attention (MHA) has become the standard architecture for text classification tasks.

However, despite the widespread adoption of the work of Ramachandran and Bello, we currently lack a solid theoretical understanding of how transformers work. Many of these modules and hyperparameters are based on contingent empirical evidence. Uncertainty is amplified in multiheaded attention as the roles and interactions between heads remain less defined. Empirically, it is well known that employing multiple heads can enhance the model accuracy. However, not all features are equally useful, and it has been shown that some features can be pruned without affecting model performance. For example, Voita et al. [1] proposed a method for quantifying the usefulness of headings and eliminating unnecessary members. Michel et al. [2] questioned the effectiveness of using multiple heads by testing the impact of repruning under various scenarios. In contrast, Bhojanapalli et al. [3] proposed methods to expand the dimensionality of Q and K by increasing the expression of each type of attention while maintaining a sufficient number of attention heads. N. Shazeer [4] introduced "talking-head attention", a variation of multi-head attention that involves linear projections across the attention-head dimension both before and after the softmax operation. While incorporating only a small number of additional parameters and a moderate amount of extra computation, talking-head attention results in improved perplexities on masked language modeling tasks and higher quality when transferring to language comprehension and question-answering tasks.

Existing self-attention models for short-text classification tasks have the following limitations.

- Insufficient annotated data, single task, and classification imbalance problems.
- The weight and Degree of correlation between attention heads are not considered.

We introduce the "squeeze-and-excitation" (SE) block [5] in the attention dimension. It adaptively recalibrates the feature responses of the attention heads by explicitly modeling the interdependencies between different attention heads. We applied this method to the attention scores, and the attention score matrix was dotted with V-lifts to generate a weight relationship between each attention head. The final output attention matrix involves simple splicing of each attention head and the formation of dependencies between each other. In addition, we design a data augmentation template with embedded labels based on prompt learning to address the data singularity problem of the dichotomous short text classification task, which takes advantage of the mutual exclusion property of labels in the dichotomous classification task. By

* Corresponding author.
E-mail address: xju_zhangtao@xju.edu.cn (T. Zhang).

https://doi.org/10.1016/j.neucom.2024.127815
Received 28 August 2023; Received in revised form 7 March 2024; Accepted 5 May 2024
Available online 13 May 2024

constructing affirmative and negative sentences, the traditional single-classification problem is converted into a multicategorization problem, and the expressive power and semantic cognitive ability of the model are improved. The contributions of this study are as follows:

- We propose head-to-head attention on the traditional self-attention mechanism, further focusing on the weighting relationship between attention heads and experimentally demonstrating its effectiveness.
- We propose an effective text enhancement template based on prompt learning to improve the loss function and experimentally demonstrate that the prior knowledge of the pretraining model is effectively extracted and that the ACC and F1 values of the model are improved.

2. Related work

2.1. Pre-training model

The effectiveness of the parameter initialization was directly influenced by the pretraining model. ELMo [6], GPT (Generative Pretraining) [7–9], BERT (Bidirectional Encoder Representation from Transformers) [10], XLNet [11], ZEN [12], (Enhanced Language Representation with Generative Entities) [13] were the main pretraining models for text at the time. In particular, BERT and ERNIE have drawn much interest and produced some deformations, including RoBERTa [14] and ERNIE2.0 [15]. However, a significant amount of computing power is required to pre-train the model. The MAE model was proposed by Peng [16]. Because the model prunes unnecessary multihead attention heads, the reallocation of attention heads improves the translation performance. However, because their strategy uses all the attention heads of the models rather than just a few, it is challenging to scale up and obtain significant improvements in the results. The attention heads in the encoder were carefully examined by Voita [17] and divided into three functional subsets: positional, syntactic, and infrequent words. A similar occurrence of distinct heads capable of capturing various syntactic functions was also observed in dependency parsing [18].

2.2. Few-shot text classification

Few-shot learning (FSL) has garnered significant attention in the field of computer vision and has been a subject of research and development in the natural language processing (NLP) domain. Prevalent few-shot text classification approaches align with those used in few-shot image classification, encompassing metrics, optimization, and other methodologies. However, the efficacy of numerous methods has diminished due to the unique characteristics of texts, particularly limited word count and insufficient contextual information in short texts. In recent years, advancements in few-shot text classification have been made using techniques such as meta-learning, pretraining, attention mechanisms, and induction networks. For instance, Xu et al. [19] devised an FSL model featuring an episode training mechanism within a meta-learning framework and bidirectional long-term attention network (BLAN), effectively addressing the issue of data scarcity and enhancing few-shot text classification performance. Deng et al. [20] introduced meta-pretraining. This method combines unsupervised language models and meta-learning to partially resolve the challenges encountered in text classification, such as inadequate data or the need to adapt to unseen classes. Furthermore, Lee et al. [21] proposed a semi-supervised few-shot text classification approach involving attention-based lexicon construction that expands a reliable training dataset through attention weights based on long short-term memory (LSTM). Geng et al. [22] introduced a novel inductive network that utilized a dynamic routing algorithm in meta-learning, enabling the learning of general representations of each class in the support set. Building on this, the team proposed a dynamic memory induction network (DMIN) for few-shot text classification [23], which offers enhanced flexibility for memory-based FSL through dynamic routing. Liu et al. [24] proposed an FSL framework for few-shot text classification that integrates the advantages of text semantics. Vector representation, meta-learning, fine-tuning vector similarity measurement and improves the applicability of few-shot text classification methods. The aforementioned applications of FSL in text classification provide valuable insights, prompting our exploration of a few-shot text classification method that integrates pretraining language models (BERT and LDA), a base meta-classifier, fine-tuning, and episode sampling mechanisms.

2.3. Short text topic modeling

Analyzing short texts allows us to infer discriminative and coherent latent topics, which is a critical and fundamental task because many real-world applications require a semantic understanding of short texts. Finally, we provide a detailed analysis of short text topic modeling techniques and discuss their performance in various applications. Qiang et al. [25] provided a detailed analysis of short-text topic modeling techniques and discussed their performance in various applications. [25]. Conventional topic modeling techniques, such as probabilistic latent semantic analysis (PLSA) [26] and latent Dirichlet allocation (LDA) [27], are commonly used to uncover the underlying semantic structures within text corpora without the need for prior annotations or document labeling. These algorithms and their variations have significantly impacted various applied fields, including modeling text collections such as news articles, research papers, and blogs. However, traditional topic models exhibit significant performance degradation when applied to short texts owing to the limited word co-occurrence information in each short text. Consequently, short-text topic modeling has garnered considerable attention from the machine learning research community in recent years to address the sparseness issue in short texts. Earlier studies [28,29] have attempted to address this challenge by leveraging external knowledge or metadata to introduce additional useful word co-occurrences across short texts, thereby potentially enhancing the performance of topic models. For instance, Wang et al. [30] constructed various hashtag graphs based on hashtags and proposed a novel framework for hashtag graph-based topic modeling to learn topics. However, the reliance on auxiliary information or metadata may not always be feasible or cost-effective for deployment. These studies underscore the need for topic models tailored to general short texts.

2.4. Prompt learning for short text

A revolutionary learning approach called prompt learning uses the PLM to transform a downstream job into a [MASK] prediction by adding a template to input words. The most crucial aspect of prompt learning is the development of templates. Earlier studies aimed to empower language models by promoting learning in contexts with limited opportunities [9,31]. Discrete, continuous, and hybrid templates can be used to categorize different template types [32]. Distinct templates are composed of existing natural words that primarily draw on the human experience. To investigate the knowledge in LMs, Petroni et al. [33] manually created a template. Schick et al.'s semi-supervised training method, PET, was proposed [34] to reformulate particular tasks as cloze-style tasks. The downstream task was changed to a cloze-style problem through prompt learning using language prompts as contexts. Using a few labeled examples from each class, few-shot classification teaches classifiers [35]. Zhu et al. [36] exploited the recent advances in the prompt-learning model [24] based on knowledge expansion in a few-shots scenario. Continuous and hybrid templates such as AutoPrompt [37], Prefix Tuning [38], P-tuning [39], and P-Tuningv2 [40] offer learnable prompt tokens to search for templates to address these problems automatically. To produce prompt candidates for learning, the LM-BFF uses the sequence-to-sequence paradigm [41]. In addition,

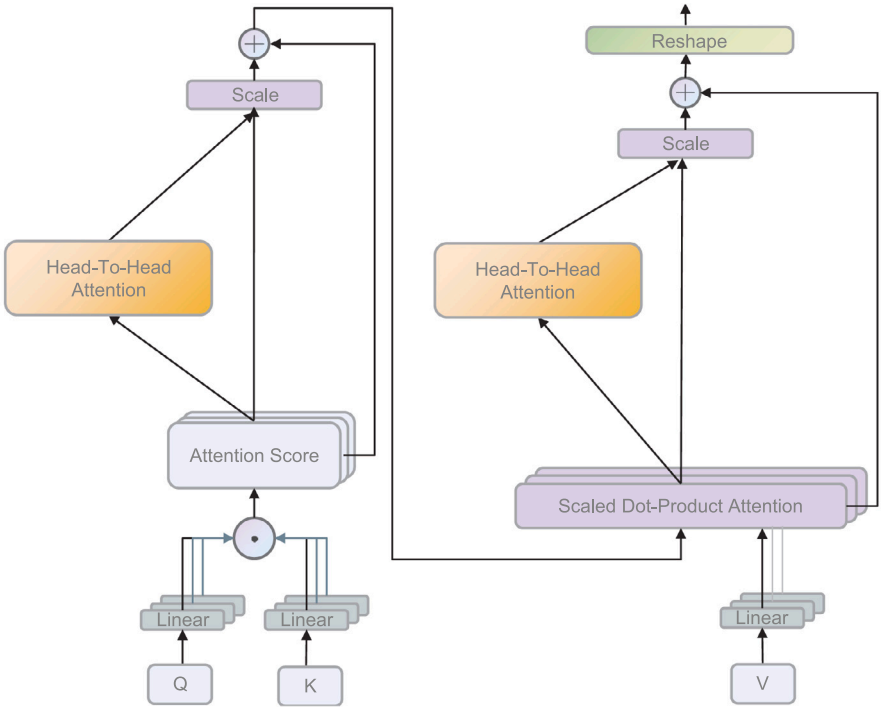


Fig. 1. Embed Head to head Attention module in Self-Attention.

several studies [38,42] have directly used learnable sequential embeddings as prompts, in contrast to discrete language words. However, most automatically generated prompts fall short of manually chosen prompts in terms of performance.

2.5. Text data augmentation

Several studies have been conducted on data augmentation for text classification. Zhang [43] replaced words or phrases with lexically related words, such as synonyms or hypernyms, which is an efficient method for performing text augmentation with little loss of generality. Several researchers have started experimenting with text augmentation techniques such as back-translation techniques [44,45], replacement approaches [46,47] and random perturbations [48]. The authors used a predetermined geometric distribution to identify the target words, which were then converted to thesaurus synonyms. Similarly, Wei and Zou [46] proposed (Easy Data Augmentation) for text classification, which created new samples from the original training data using four straightforward operations: synonym replacement, random insertion, random swapping, and random deletion. Feng [49] expanded these substitution techniques, particularly to text generation.

3. Approach

3.1. Overview

The attention score matrix is also multihead and generated after the dot product of the Q and K matrices. In this study, we used a head-to-head attention module to compute the attention of each head in the attention_ score matrix to obtain the weight relationship between the heads. After the calculation, the model generates different dependencies for each attention_score feature map. After the attention_score feature matrices with different weights are dotted with the V matrix, other preference relationships are generated between each head of the V matrix. Its structure is shown in Fig. 1.

3.2. Head to head attention (HTHAttention)

This section focuses on the relations in the attention head dimension inspired by the squeeze and excite (SE) block. A new structural unit focusing on multiple head-dimensional features, which we call head-to-head attention, is adapted by explicitly recalibrating head-dimensional feature responses and adaptively recalibrating the feature responses between multiple heads by explicitly modeling the interdependencies between the heads.

Traditional attention treats text feature vectors as objects for extracting the feature relationships. The difference between head-to-head attention and traditional attention mechanisms is that we consider each attention head as an object, making extracting the feature relationships between each attention head easier. Because traditional attention heads are independent of each other, the proposed head-to-head attention mechanism causes attention heads to interact and produce weight relationships, thereby reducing the redundancy of attention heads and improving the efficiency of fusing attention head features.

U is the feature map of $H \times L \times D$, where $(L,\ D)$ is the size of the head. H denotes the number of attention heads. After $Fsq(,)$ squeeze operation, the feature map becomes a feature map $M \in \mathbb{R}^{1\times1\times h}$. After linear layer mapping into two feature vectors, X and Y, The head-to-head attention process is expressed as follows:

$$M = F_{sq}(U) = \frac{1}{L \times D}\sum_{i=1}^{L}\sum_{j=1}^{D}U(i,j) \tag{1}$$

$$X = \sigma(F_{linear}(M)) \tag{2}$$

$$Y = \sigma(F_{linear}(M)) \tag{3}$$

$$Z = softmax(\frac{XY^T}{\sqrt{d_Y}}) \tag{4}$$

$$\widetilde{M} = F_{scale}(Z, M) = dropout(Z \otimes M) \tag{5}$$

$$\widetilde{U} = U \odot \widetilde{M} \tag{6}$$

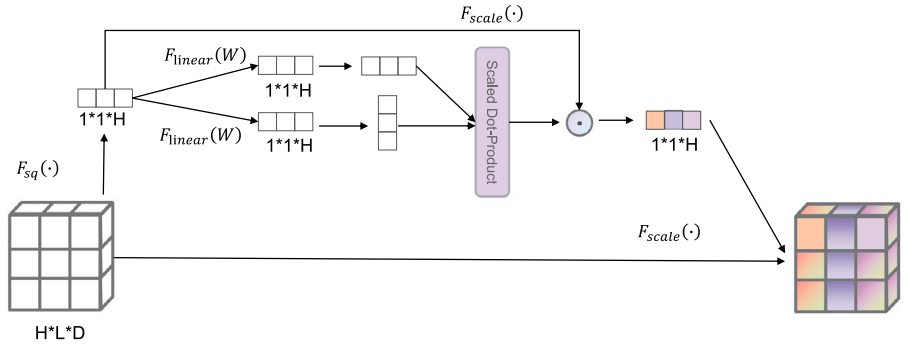


Fig. 2. Head to head Attention.

where ⊗ denotes matrix multiplication ⊙ denotes element-wised multiplication. σ is the activation function. F_{linear} denotes the linear transformation function. $F_{sq}(\cdot)$ refers to the global average pooling $F_{scale}(,)$ refers to the matrix multiplication between the vector and the feature map. The dimensionality of the $\widetilde{U}$ matrix is identical to that of the U matrix, where $\widetilde{U} \in \mathbb{R}^{h\times l\times d}$. Its structure is illustrated in Fig. 2.

3.3. Prompt text augmentation (PTA)

In this section, we describe the design of a prompt-learning-based text augmentation template (PTA) with embedded labels for mutually exclusive dichotomous text labels. Because dichotomous text labels are mutually exclusive, we add a prompt template to the input text to turn the single classification task into a multiclassification task based on this property, thereby improving the loss function. For example, a traditional input sample is label: 0, context: the apple is bad. A zero or one label represents negative or positive, respectively. For example, in BERT, an input sentence is constructed using the [CLS]. The apple is very bad [SEP], where the [CLS] vector is the output label vector, and the loss calculation is performed using the real label serialization vector.

As previously stated, we redesigned the template format for the input. We prefixed the text with affirmative and negative sentences, embedded labels, and masked them. This enables the model to make affirmative and negative judgments regarding the input text. Unlike traditional prompt learning, which requires mapping the labels to a particular class of words, this approach makes the final output dimension equal to the word list length. This approach involves more than 20,000 words in most pretraining models, which is a considerable computational challenge. We discarded the label mapping; subsequently, the output dimension was consistent with the label type.

Take "This candy is so delicious!" as an example. As shown in Fig. 3, the traditional input is "[CLS] This candy is so, "and the PTA input is "[CLS] The label of the following sentence is [MASK]"., so the label is not [MASK]: This candy is so nice!"

Traditional single-label classification was transformed into multitask learning with the simultaneous determination of positive and negative labels using the designed template. The two [MASK] semantics we extracted are supposed to be opposite, and the corresponding true labels should be mutually exclusive. The model learned not only positive classifications but also negative judgments. The experiments demonstrated that the method improved the model's ability to classify positive labels – prompt text augmentation (PTA) – as shown in Fig. 3.

We aim to embed two opposite labels in a sentence simultaneously using prompting templates, thereby allowing the sentence to learn both positive and negative expressions. As constructing prompting templates that can embed opposite labels is the main focus, the natural semantic fluency of the prompting templates is secondary. Therefore, we strive to construct semantically simple templates that reflect positive and negative expressions in a sentence. Automated templates that did not contain semantics were also used. We used simple templates and non-mapped labels to decrease the model's reliance on manually crafted templates with coherent semantics, thereby enhancing the generalizability of the template. In Section 4.7, we discuss and experimentally demonstrate the effectiveness of our choice of concise semantic templates, which allow PTA to be applied to the input layer of a broader range of models.

Table 1
Dataset statistics.

Dataset	ChnSentiCorp[a]	COLD[b]	SST-2[c]
train	9.6 K	25.726 K	67.35 K
dev	1.2 K	6.431 K	0.873 K
test	1.2 K	5.323 K	1.821 K

[a] https://github.com/duanruixue/chnsenticorp.
[b] https://github.com/thu-coai/COLDataset.
[c] http://icrc.hitsz.edu.cn/info/1037/1146.htm

4. Experiments and results

4.1. Dataset

Three datasets were used in this experiment; the details of all the datasets are listed in Table 1. These are introduced as follows.

ChnSentiCorp is a Chinese sentiment analysis dataset containing online shopping reviews of hotels, laptops, and books. The dataset distribution is as follows.

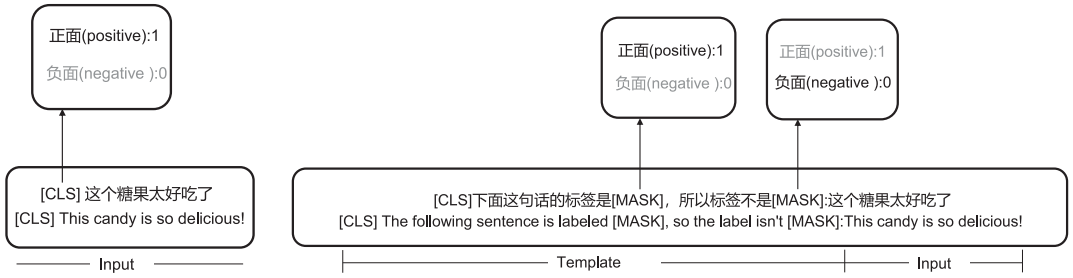
COLD [50], a Chinese insulting language dataset, covers topics such as race, gender, and region. The dataset contains a total of 37,480 sentences, of which there are 18,041 sentences with offensive language (average length of 53.69 characters); there are 19,439 sentences without offensive language (average length of 44.20 characters).

SST-2, The SST-2 (Stanford Sentiment Treebank 2) dataset is a widely used benchmark in natural language processing (NLP) and sentiment analysis research. This extension of the original SST dataset is specifically designed for binary sentiment classification tasks. The dataset comprised sentences extracted from movie reviews; each sentence was labeled with its corresponding sentiment as either positive or negative.

4.2. Experimental environment

Experimental environment: ubuntu 16.4 system, GPU: Nvidia RTX 3090; pytorch:1.11.0.

All experimental procedures were fine-tuned using the three datasets. We used the same pretraining weights as the BERT-base and Ernie-base for fine-tuning. The embedding dimension of each word was 768, and the total vocabulary size was 21,128. The initialization range of



(a) Traditional Input (b) Inputs for PTA

Fig. 3. Comparison of the two different inputs.

the matrix was 0.02. The model uses the cross-entropy function as the loss function and the AdamW optimizer to update the parameters. The batch size was set to 64, the learning rate was set to 2e-5, the gradient clipping max grad norm was set to 10, the early termination batch size was set to 1000, and the results of five random seed experiments were averaged.

4.3. Baselines

TextCNN [51]: Short for Text Convolutional Neural Network, is a type of convolutional neural network architecture specifically designed for processing text data. It was initially introduced in a research paper entitled "Convolutional Neural Networks for Sentence Classification" by Yoon Kim in 2014.

FastText [52]: FastText is a library and framework for efficient text classification and representation developed by Facebook's AI Research (FAIR) team. This was introduced in a research paper titled "Bag of Tricks for Efficient Text Classification" by Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov in 2016.

DPCNN [53]: DPCNN, which stands for Deep Pyramid Convolutional Neural Network, is a convolutional neural network architecture designed for text classification tasks. This was introduced in a research paper titled "Deep Pyramid Convolutional Neural Networks for Text Categorization" by Y. Zhang, Z. Ye, and Y. Zhang in 2017.

BERT-CNN: BERT-CNN refers to a hybrid architecture that combines BERT (Bidirectional Encoder Representations from Transformers) and CNN (Convolutional Neural Network) models for natural language processing tasks. BERT is a powerful pretraining language representation model based on transformer architecture, whereas CNN is a convolutional neural network commonly used for text classification and feature extraction.

BERT: BERT (Bidirectional Encoder Representations from Transformers) is a powerful pretraining language representation model introduced by researchers at Google in 2018. It has significantly advanced the field of natural language processing (NLP) by providing state-of-the-art results for various NLP tasks.

ALBERT [54]: ALBERT (A Lite BERT) is a variant of the BERT (Bidirectional Encoder Representations from Transformers) model that aims to reduce the model's size and computational requirements while maintaining its performance. ALBERT was introduced by Google Research in 2019.

RoBERTa [14]: RoBERTa (Robustly Optimized BERT approach) is a variant of the BERT (Bidirectional Encoder Representations from Transformers) model introduced by researchers at Facebook AI in 2019. RoBERTa aims to improve BERT's performance by addressing certain limitations and applying additional training techniques.

Ernie: Ernie (Enhanced Representation through kNowledge IntEgration) is a language representation model developed by researchers at Baidu Research. It enhances language understanding by incorporating knowledge from external sources such as graphs or structured databases.

XLNet: XLNet is a state-of-the-art language representation model introduced by researchers at Google in 2019. XLNet builds upon the transformer architecture, which was also used in models such as BERT, but introduces several novel ideas to improve the limitations of previous models.

MacBert [55]: MacBert: The model mitigates the gap between the pretraining and fine-tuning phases by using a similar wordmark, which is effective for downstream tasks.

4.4. Loss function and evaluation index

Loss function: The cross-entropy loss function measures the deviation of the predicted value from the actual value. The cross-entropy loss function is given by Eq. (7).

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \tag{7}$$

where $\hat{y}$ denotes the predicted probability, y is the true label, and N is the number of sample categories. The network parameters are updated according to the loss function.

This section introduces text augmentation templates, and traditional single-task classification is transformed into multitask classification. Owing to the changes in tasks, we designed new loss functions to adapt the model to better tasks during transformation. Because of the mutually exclusive nature of dichotomous labels, we believe that the losses of positive and negative labels must converge simultaneously; therefore, we average the two loss functions to constrain the convergence direction of the model further. The cross-entropy loss function is given by Eq. (8).

$$L = \gamma * Loss_{pos} + (1 - \gamma) * Loss_{neg} \tag{8}$$

Where $Loss_{pos}$ represents the loss of predicted and true labels for positive sentences. $loss_{neg}$ represents the loss of predicted and opposite labels for negative sentences. γ is a factor that controls the two loss functions.

w_f1(weighted_F1) because macro $F1$ is the arithmetic mean of multiple $F1$ values. When the samples are imbalanced, we sometimes want to assign different weights to different classes according to the number of samples in each category. This is w_f1, and the calculation formula is given by Eq. (9).

$$w\_f1 = \frac{1}{total}\sum_{k=1}^{N} f1_i * N_i \tag{9}$$

4.5. Experimental results

We experimented with three datasets, CHNSenticorp, COLD, and SST-2, to demonstrate the effectiveness of the module. Baseline data were divided into two parts. The first part of Table 2 presents the CNN-based model without the attention mechanism, and the second part presents the model using the attention mechanism. As shown in

Table 2
Experiment results.

Dataset	CHNSenticorp			COLD			SST-2		
	Dev_acc	Test_acc	Test_W_{f1}	Dev_acc	Test_acc	Test_W_{f1}	Dev_acc	Test_acc	Test_W_{f1}
TEXTCNN	86.08	84.00	84.00	80.35	74.15	74.09	63.67	78.40	84.00
FastText	86.92	86.92	86.90	84.05	74.96	75.21	67.89	71.06	70.63
DPCNN	84.05	84.75	82.75	83.56	76.57	76.22	71.22	78.83	78.42
BERT-CNN	84.08	92.50	92.49	91.35	81.32	81.51	91.20	84.95	84.88
BERT	91.06	92.30	92.29	91.30	81.23	81.39	91.42	87.12	87.55
Albert-base	91.68	89.1	89.09	89.93	79.56	80.12	90.13	86.22	86.37
Albert-xlarge	89.1	91.47	91.52	90.97	80.24	79.83	91.21	88.68	88.95
Roberta	92.64	92.64	92.64	91.62	81.62	81.83	92.55	89.11	89.17
MacBERT	92.55	92.45	92.45	91.68	81.36	81.29	–	–	–
Ernie	93.75	93.58	93.58	92.89	81.29	81.46	–	–	–
Xlnet	93.75	94.75	94.75	92.56	81.57	81.76	91.01	89.14	89.63
Our Method	93.74	95.25	95.25	92.67	82.39	82.57	91.16	90.17	90.18

Table 3
Ablation experiment results.

Model	COLD	
	dev_acc	test_acc
ALBERT-xlarge	90.07	80.24
ALBERT-base	89.93	79.56
BERT-base	91.30	81.23
ERNIE-base	92.89	81.29
RoBERTa	91.62	81.62
XLNet	92.56	81.57
BERT		
+PTA	92.98	82.18
+HTHAttention	92.37	81.92
ERNIE		
+PTA	92.93	82.46
+HTHAttention	92.41	81.70

Table 4
Experiment results of autoprompt.

Model	COLD	
	dev_acc	test_acc
BERT		
+PTA	92.98	82.18
+Complex template	92.15	81.88
+Autoprompt template	92.11	81.39
ERNIE		
+PTA	92.93	82.46
+Complex template	92.45	82.41
+Autoprompt template	92.32	81.45

[CLS]如果0代表正面,1代表负面,下面这句话的情绪是[MASK]面,
所以情绪不是[MASK]面:这个糖果太好吃了!

[CLS] If 0 is positive and 1 is negative, the sentiment of the following sentence is the [MASK], so the sentiment is not the [MASK] :This candy is so delicious!

Fig. 4. Complex templates.

[CLS][UNK] [UNK] [UNK] [UNK] [MASK][UNK] [UNK] [UNK] [UNK] [MASK]:这个糖果太好吃了
[CLS][UNK] [UNK] [UNK] [UNK] [MASK][UNK] [UNK] [UNK] [UNK] [MASK]:This candy is so delicious!

Fig. 5. Autoprompt.

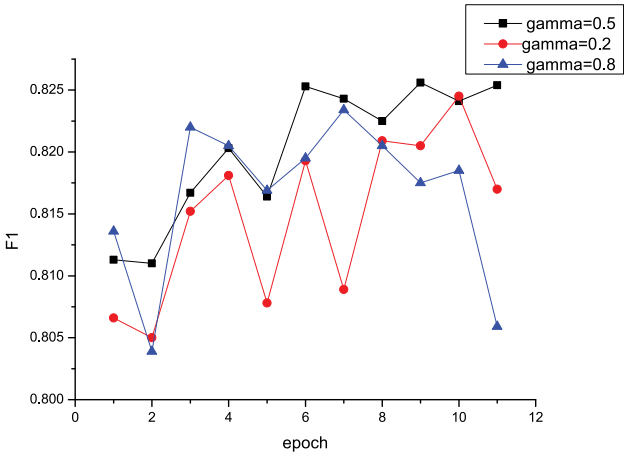


Fig. 6. An in-depth exploration of the γ.

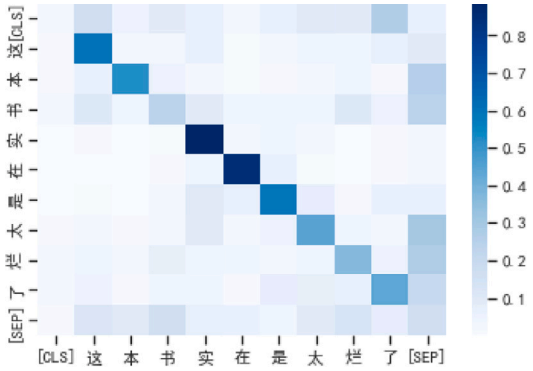


Fig. 7. Attention scores in transformer.

Table 2, the nonattention model performed worse than the attention mechanism model in the text classification task, which is consistent with the results of many studies. CNN-based text classification models the local feature information of a sentence and does not consider its relevance as a whole. The attention mechanism model considers the relevance of a sentence as a whole. Our approach achieved better accuracy and F1 values on the test set than models such as BERT and BERT-based variants.

4.6. Ablation experiment

In this section, we examine the effects of the PTA method and HTHAttention module on the experiments to explore their respective impacts. We performed ablation experiments on the COLD. We used BERT and COLD as backbone networks and added the PTA method and HTHAttention module on top of them to compare the experiments with current mainstream BERT variant models. Table 3 shows that

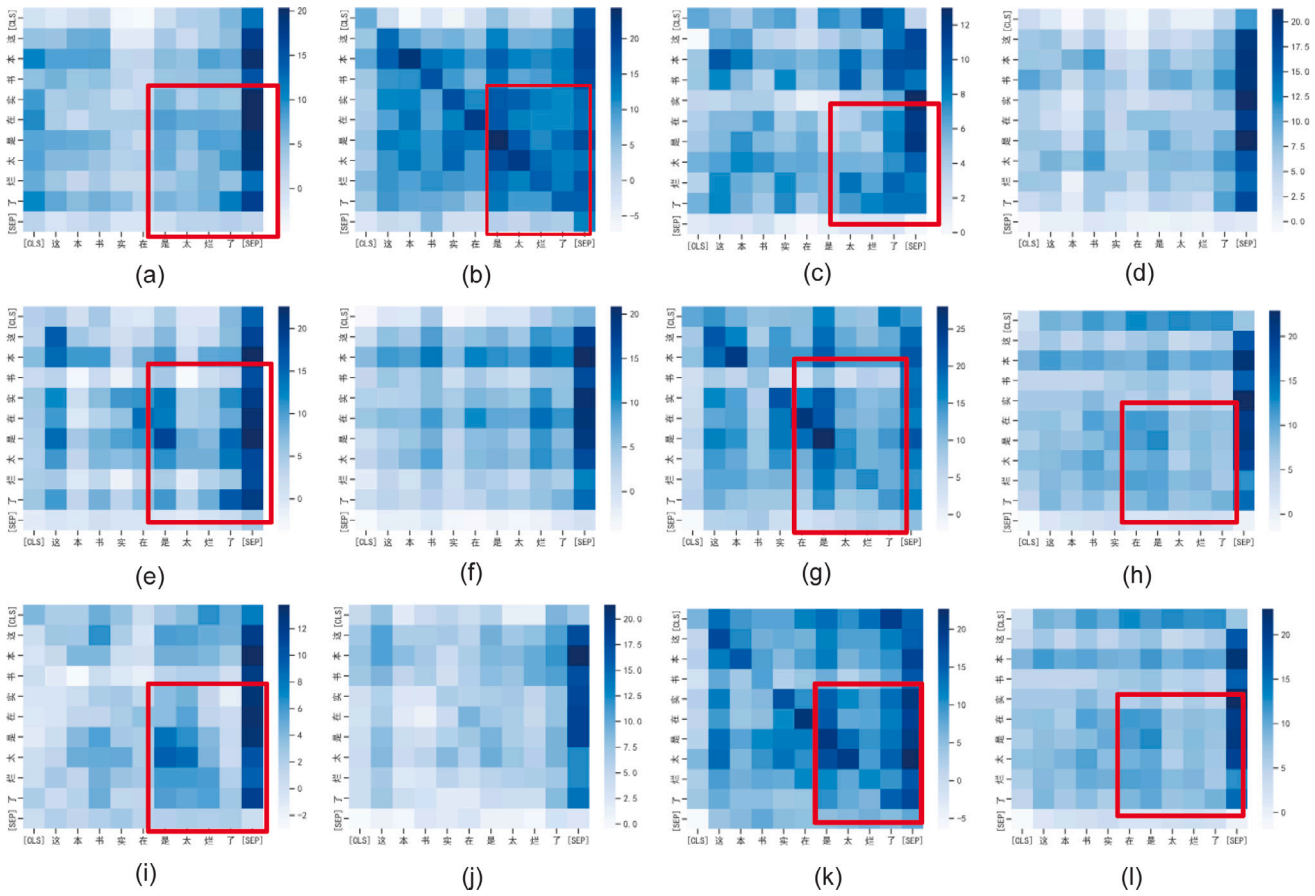


Fig. 8. Attention scores in HTHAttention.

after adding PTA and HTHAttention to BERT, our ACC values on the test set were better than those of all the baselines. The ACC for the development set exceeded that of the BERT base. ERNIE + PTA exhibited the best performance. The performance improvement in the ERNIE+ HTHAttention model is smaller and slightly lower than that in the BERT+ HTHAttention model. As ERNIE introduces a large amount of knowledge graph data, the dimensions of this graph network are far greater than those of HTHAttention, and the impact of HTHAttention is less than that of the graph network. Therefore, the improvement in the performance of the ERNIE+ HTHAttention model is minimal, whereas the simple BERT combined with the HTHAttention model achieves a greater improvement in performance. We found that the performance improvement of the PTA was more obvious. According to the analysis, we believe that for a balanced dataset such as COLD, the consistency of textual semantic information and label semantics will increase after adding PTA to enhance the dataset.

4.7. Explore further

In this section, we explore the differences between manual templates, traditional semantically prompted templates that map labels, and automatic templates in the context of natural language processing (NLP). First, we refer to them as complex templates to create mapped labeled traditional semantic prompt templates. We map labels to become meaningful words and embed them in prompt templates such that our templates and input sentences become sentences with coherent semantics, as shown in Fig. 4. To create automatic templates, we systematically replaced all the words in the original template with [UNK], as shown in Fig. 5. According to the analysis in Table 4, the performance of the automated and coherent semantic templates significantly decreased by approximately 1 percent compared to that of the PTA, and the performance of the coherent semantic complex templates was similar. This observation suggests that automated templates that do not use mutually exclusive labels have a significant effect on the model performance. Although the performances of coherent semantically complex templates are similar, they require mapping tag meanings and increasing the sentence length and complexity. Upon closer examination, we found that the simultaneous embedding of mutually exclusive labels was the key to the effect of PTA.

4.8. About the setting of control parameter γ

In this section, we discuss the setting of parameter γ, a control parameter which we set to control the relationship between positive and negative losses. As shown in Fig. 6, when gamma was 0.5, the positive and negative losses performed best when averaged, and the best F1 value was achieved in Epoch 6. A gamma of 0.5 means that the loss function must consider both positive and negative losses, which will constrain the optimization direction of the model more strictly and reduce the risk of overfitting. As shown in Fig. 6.

4.9. Visual analysisl

In this section, we visualize the attention scores of the transformer alone and those combined with HTHAttention. Figs. 7 and 8 show that the attention score of HTHAttention preserves the self-attention weights in the transformer while enhancing the relevance of words at different positions in the text. We labeled the distribution with the higher attention scores for each attention head, as shown in Fig. 8. We found that most attention scores were distributed among words with

a tendency to show the emotion of the sentence (bottom right part of the attention score plot). This proves that head-to-head attention is effective for fusing the features of multiple attention heads in the self-attention mechanism.

5. Conclusion

This study proposes a plug-in module to target self-attention pre-training models and label mutual exclusion-based text augmentation. This is the first study to improve the feature extraction capability of the self-attention mechanism by exploring interhead attention and text augmentation based on prompt learning. Compared with baseline models, we achieved superior performance. We tested our methodology on various phrase classification datasets, and the results showed that it performed substantially better prediction accuracy than other text augmentation techniques and text classification learning frameworks. We significantly improved the results of our trials in terms of self-attention mechanisms and pretraining models. Future work will focus on expanding the task types to multiclassification tasks rather than just the current dichotomous classification tasks and look into replacing the manual templates with AutoPrompt templates.

CRediT authorship contribution statement

Bo Peng: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation. **Kundong Han:** Software, Investigation. **Liang Zhong:** Validation. **Shengbo Wu:** Software. **Tao Zhang:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the Xinjiang L&Q Project (No.2022LQ03004), the Science and Technology Plan Project of Xinjiang Uygur Autonomous Region (No.2022NC192) and Xinjiang Science and Technology Major Program (2023A03001).

References

[1] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, Ivan Titov, Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned, 2019, arXiv preprint arXiv:1905.09418.

[2] Paul Michel, Omer Levy, Graham Neubig, Are sixteen heads really better than one? in: Advances in Neural Information Processing Systems, vol. 32, 2019.

[3] Srinadh Bhojanapalli, Chulhee Yun, Ankit Singh Rawat, Sashank Reddi, Sanjiv Kumar, Low-rank bottleneck in multi-head attention models, in: International Conference on Machine Learning, PMLR, 2020, pp. 864–873.

[4] Noam Shazeer, Zhenzhong Lan, Youlong Cheng, Nan Ding, Le Hou, Talking-heads attention, 2020, arXiv preprint arXiv:2003.02436.

[5] Jie Hu, Li Shen, Gang Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.

[6] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer, Deep contextualized word representations. CoRR abs/1802.05365 (2018), 1802, arXiv preprint arXiv:1802.05365.

[7] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al., Improving Language Understanding by Generative Pre-Training, OpenAI, 2018.

[8] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (8) (2019) 9.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al., Language models are few-shot learners, in: Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 1877–1901.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.

[11] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, Quoc V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, in: Advances in Neural Information Processing Systems, vol. 32, 2019.

[12] Shizhe Diao, Jiaxin Bai, Yan Song, Tong Zhang, Yonggang Wang, ZEN: Pre-training Chinese text encoder enhanced by n-gram representations, 2019, arXiv preprint arXiv:1911.00720.

[13] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, Hua Wu, Ernie: Enhanced representation through knowledge integration, 2019, arXiv preprint arXiv:1904.09223.

[14] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019, arXiv preprint arXiv:1907.11692.

[15] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, Haifeng Wang, Ernie 2.0: A continual pre-training framework for language understanding, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, (no. 05) 2020, pp. 8968–8975.

[16] Hao Peng, Roy Schwartz, Dianqi Li, Noah A. Smith, A mixture of h - 1 heads is better than h heads, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 6566–6577.

[17] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, Ivan Titov, Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 5797–5808.

[18] Phu Mon Htut, Jason Phang, Shikha Bordia, Samuel R. Bowman, Do attention heads in BERT track syntactic dependencies? 2019, arXiv preprint arXiv:1911.12246.

[19] Xu Tongtong, Sun Huazhi, Ma Chunmei, Jiang Lifen, Liu Yichen, Classification model for few-shot texts based on bi-directional long-term attention features, Data Anal. Knowl. Discov. 4 (10) (2020) 113–123.

[20] Shumin Deng, Ningyu Zhang, Zhanlin Sun, Jiaoyan Chen, Huajun Chen, When low resource NLP meets unsupervised language model: Meta-pretraining then meta-learning for few-shot text classification (student abstract), in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, (no. 10) 2020, pp. 13773–13774.

[21] Ju-Hyoung Lee, Sang-Ki Ko, Yo-Sub Han, Salnet: Semi-supervised few-shot text classification with attention-based lexicon construction, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, (no. 14) 2021, pp. 13189–13197.

[22] Ruiying Geng, Binhua Li, Yongbin Li, Xiaodan Zhu, Ping Jian, Jian Sun, Induction networks for few-shot text classification, 2019, arXiv preprint arXiv:1902.10482.

[23] Ruiying Geng, Binhua Li, Yongbin Li, Jian Sun, Xiaodan Zhu, Dynamic memory induction networks for few-shot text classification, 2020, arXiv preprint arXiv:2005.05727.

[24] Wenfu Liu, Jianmin Pang, Nan Li, Feng Yue, Guangming Liu, Few-shot short-text classification with language representations and centroid similarity, Appl. Intell. 53 (7) (2023) 8061–8072.

[25] Jipeng Qiang, Zhenyu Qian, Yun Li, Yunhao Yuan, Xindong Wu, Short text topic modeling techniques, applications, and performance: A survey, IEEE Trans. Knowl. Data Eng. 34 (3) (2020) 1427–1445.

[26] Thomas Hofmann, Probabilistic latent semantic indexing, in: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1999, pp. 50–57.

[27] David M. Blei, Andrew Y. Ng, Michael I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (Jan) (2003) 993–1022.

[28] Ou Jin, Nathan N. Liu, Kai Zhao, Yong Yu, Qiang Yang, Transferring topical knowledge from auxiliary long texts for short text clustering, in: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, 2011, pp. 775–784.

[29] Xuan-Hieu Phan, Le-Minh Nguyen, Susumu Horiguchi, Learning to classify short and sparse text & web with hidden topics from large-scale data collections, in: Proceedings of the 17th International Conference on World Wide Web, 2008, pp. 91–100.

[30] Yuan Wang, Jie Liu, Yalou Huang, Xia Feng, Using hashtag graph-based topic model to connect semantically-related words without co-occurrence in microblogs, IEEE Trans. Knowl. Data Eng. 28 (7) (2016) 1919–1933.

[31] Ajay Patel, Bryan Li, Mohammad Sadegh Rasooli, Noah Constant, Colin Raffel, Chris Callison-Burch, Bidirectional language models are also few-shot learners, 2022, arXiv preprint arXiv:2209.14500.

[32] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, Graham Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, ACM Comput. Surv. 55 (9) (2023) 1–35.

[33] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, Sebastian Riedel, Language models as knowledge bases? 2019, arXiv preprint arXiv:1909.01066.
[34] Timo Schick, Hinrich Schütze, Exploiting cloze questions for few shot text classification and natural language inference, 2020, arXiv preprint arXiv:2001.07676.
[35] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, Jie Tang, GPT understands, too, 2021, arXiv preprint arXiv:2103.10385.
[36] Yi Zhu, Ye Wang, Jipeng Qiang, Xindong Wu, Prompt-learning for short text classification, IEEE Trans. Knowl. Data Eng. (2023).
[37] Taylor Shin, Yasaman Razeghi, Robert L. Logan, Eric Wallace, Sameer Singh, Autoprompt: Eliciting knowledge from language models with automatically generated prompts, 2020, arXiv preprint arXiv:2010.15980.
[38] Xiang Lisa Li, Percy Liang, Prefix-tuning: Optimizing continuous prompts for generation, 2021, arXiv preprint arXiv:2101.00190.
[39] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, Jie Tang, P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2022, pp. 61–68.
[40] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, Jie Tang, P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks, 2021, arXiv preprint arXiv:2110.07602.
[41] Tianyu Gao, Adam Fisch, Danqi Chen, Making pre-trained language models better few-shot learners, 2020, arXiv preprint arXiv:2012.15723.
[42] Brian Lester, Rami Al-Rfou, Noah Constant, The power of scale for parameter-efficient prompt tuning, 2021, arXiv preprint arXiv:2104.08691.
[43] Xiang Zhang, Junbo Zhao, Yann LeCun, Character-level convolutional networks for text classification, in: Advances in Neural Information Processing Systems, vol. 28, 2015.
[44] Rico Sennrich, Barry Haddow, Alexandra Birch, Improving neural machine translation models with monolingual data, 2015, arXiv preprint arXiv:1511.06709.
[45] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, Quoc Le, Unsupervised data augmentation for consistency training, Advances in Neural Information Processing Systems, vol. 33 (2020) 6256–6268.
[46] Jason Wei, Kai Zou, Eda: Easy data augmentation techniques for boosting performance on text classification tasks, 2019, arXiv preprint arXiv:1901.11196.
[47] William Yang Wang, Diyi Yang, That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 2557–2563.
[48] Yong Cheng, Zhaopeng Tu, Fandong Meng, Junjie Zhai, Yang Liu, Towards robust neural machine translation, 2018, arXiv preprint arXiv:1805.06130.
[49] Steven Y. Feng, Varun Gangal, Dongyeop Kang, Teruko Mitamura, Eduard Hovy, Genaug: Data augmentation for finetuning text generators, 2020, arXiv preprint arXiv:2010.01794.
[50] Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, Minlie Huang, Cold: A benchmark for chinese offensive language detection, 2022, arXiv preprint arXiv:2201.06025.
[51] Yahui Chen, Convolutional Neural Network for Sentence Classification (Master's thesis), University of Waterloo, 2015.
[52] Armand Joulin, Edouard Grave, Piotr Bojanowski, Tomas Mikolov, Bag of tricks for efficient text classification, 2016, arXiv preprint arXiv:1607.01759.
[53] Rie Johnson, Tong Zhang, Deep pyramid convolutional neural networks for text categorization, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 562–570.
[54] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, Radu Soricut, Albert: A lite bert for self-supervised learning of language representations, 2019, arXiv preprint arXiv:1909.11942.
[55] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, Guoping Hu, Revisiting pre-trained models for Chinese natural language processing, 2020, arXiv preprint arXiv:2004.13922.

Bo Peng, He is currently working toward the M.S. degree in Artificial Intelligence with the Department of Xinjiang University, Urumqi, China. His research interests include Natural Language Processing, the theory of systemic functional grammar and the theory of multimodalityand.
E-mail: bopeng0520@163.com

Kundong Han, He is currently working toward the M.S. degree in Artificial Intelligence with the Department of Xinjiang University, Urumqi, China. His research interests include Natural Language Processing, the theory of systemic functional grammar and the theory of multimodalityand.
E-mail: han9614716431@163.com

Liang Zhong, He is currently working toward the M.S. degree in Artificial Intelligence with the Department of Xinjiang University, Urumqi, China. His research interests are in pattern recognition.
E-mail: 107552204854@stu.xju.edu.cn

Shengbo Wu, He is currently working toward the M.S. degree in Artificial Intelligence with the Department of Xinjiang University, Urumqi, China. His research interests are in Natural Language Processing.
E-mail: 107552204815@stu.xju.edu.cn

Tao Zhang, Serving as an Associate Professor at Xinjiang University in 2021, his research interests are in natural language processing and multimodal recognition.
E-mail: xju_zhangtao@xju.edu.cn