Applied Intelligence (2023) 53:9103–9114
https://doi.org/10.1007/s10489-022-03947-w

Improving the robustness of machine reading comprehension via contrastive learning

Jianzhou Feng[1] · **Jiawei Sun**[1] · **Di Shao**[1] · **Jinman Cui**[1]

Accepted: 29 June 2022 / Published online: 5 August 2022
© The Author(s) 2022

Abstract
Pre-trained language models achieve high performance on machine reading comprehension task, but these models lack robustness and are vulnerable to adversarial samples. Most of the current methods for improving model robustness are based on data enrichment. However, these methods do not solve the problem of poor context representation of the machine reading comprehension model. We find that context representation plays a key role in the robustness of the machine reading comprehension model, dense context representation space results in poor model robustness. To deal with this, we propose a Multi-task machine Reading Comprehension learning framework via Contrastive Learning. Its main idea is to improve the context representation space encoded by the machine reading comprehension models through contrastive learning. This special contrastive learning we proposed called Contrastive Learning in Context Representation Space(CLCRS). CLCRS samples sentences containing context information from the context as positive and negative samples, expanding the distance between the answer sentence and other sentences in the context. Therefore, the context representation space of the machine reading comprehension model has been expanded. The model can better distinguish between sentence containing correct answers and misleading sentence. Thus, the robustness of the model is improved. Experiment results on adversarial datasets show that our method exceeds the comparison models and achieves state-of-the-art performance.

Keywords Machine reading comprehension · Robust · Contrastive learning · Representation · Multi task · Pre-train language model

1 Introduction

In recent years, with the development of large-scale pre-trained language models based on transformers, using pre-trained model with fine-tuning has gradually become the mainstream method for natural language processing tasks. Machine Reading Comprehension (MRC) methods based on pre-trained models have achieved good results on some popular MRC datasets, such as SQuAD [1] and RACE [2], but some studies [3–5] have shown that MRC models are vulnerable to adversarial samples, which called the robustness problem of MRC model. As shown in Table 1, adding additional sentences which has similar semantic with answer sentences could mislead the model and make it output wrong answers. Thus, the robustness of the model still needs to be further improved.

MRC models go wrong because they fail to distinguish between misleading sentences and answer sentences. Answer sentences and misleading sentences are recognized by the model as correct answers to question. Human beings can understand question well, discover the subtle differences among multiple candidate answer sentences from the question's perspective, and then find the correct answer. As shown in Table 1, people can judge that the type of answer is "time" according to the content of the answer sentence, and the important clue is "Richard's fleet". Then we can find the semantic differences between the two sentences from the perspective of the participants,

✉ Jiawei Sun
sunjiawei@stumail.ysu.edu.cn

Jianzhou Feng
fjzwxh@ysu.edu.cn

Di Shao
shaodi@stumail.ysu.edu.cn

Jinman Cui
cuijinman@163.com

[1] School of Information Science and Engineering, Yanshan University, No. 438, West Section of Hebei Street, Qinhuangdao, 066004, Hebei Province, China

Table 1 An example from the adversarial dataset. Prediction of the model changes from "1191" to "1192" after adding misleading sentence

Question	"What year did the storm hit Richard's fleet?"
Context	"In April 1191 Richard the Lion-hearted left Messina with a large fleet in order to reach Acre. In June 1192 Robert the Neptune left Catania with a small fleet in order to reach Hectare. But a storm dispersed Richard's fleet. After some searching, it was discovered that the boat carrying his sister and his fiancée Berengaria was anchored on the south coast of Cyprus, together with the wrecks of several other ships, including the treasure ship..."
Original Answer:	1191
Model Prediction under adversary:	1192
Answer Sentence:	"In April 1191 Richard the Lion-hearted left Messina with a large fleet in order to reach Acre."
Misleading Sentence:	"In June 1192 Robert the Neptune right Catania with a small fleet in order to reach Hectare."
Distance between Answer Sentence and Misleading Sentence:	0.879

and finally choose the correct answer according to the content described in the question. However, in the MRC, the encoding results of the answer sentence vector is close to that of the misleading sentence vector. And the model pays more attention to the misleading sentences because the misleading sentences contain some non-key words which appeared in the question sentence, and ignores the important decisive significance of the substantive words in the sentence for the answer result.

To deal with the robustness issue mentioned above, Welbl et al. [6] used data augmentation and adversarial training, Jia and Liang [3], Wang and Bansal [7], Liu et al. [8] enriched the training set by generating adversarial examples. However, since the types of adversarial examples are innumerable, all the above methods by augmenting the training dataset have some limitations. Majumder et al. [9] used an answer candidate reranking mechanism to avoid model errors on adversarial examples, but it sacrifices the accuracy on non-adversarial datasets and requires an additional complex structure to support the algorithm. All the above methods have improved the robustness of MRC models.

We find that model representation has a great influence on model robustness. Existing MRC models' representation space is dense, especially the distance between answer sentence vectors and misleading sentence vectors is too close. Therefore, we propose Multi-task machine Reading Comprehension learning framework via Contrastive Learning (MRCCL), which introduces Contrastive Learning (CL) task into the MRC model through multi-task joint training. Specifically, we use dropout to generate positive samples, select other sentences in the passage as negative samples for contrastive learning, and jointly train the model with multi-task. While expanding the model representation space, the representation consistency in the model representation space is maintained. Our contributions are summarized as follows:

1. We propose a new contrastive learning algorithm to improve the robustness of machine reading comprehension. By selecting each sentence in the context as positive and negative samples, the context representation space of the pre-trained language model is improved, and the robustness of the machine reading comprehension model is further improved.
2. Experimental results show that our algorithm effectively alleviates the problem of poor robustness of the MRC model and has achieved the state-of-the-art on the adversarial dataset.

2 Related work

Adversarial Attacks in Machine Reading Comprehension Model The research on the robustness of MRC models have just emerged in recent years, among which are various attack methods. Jia and Liang [3] successfully attacked the MRC model by adding a misleading sentence at the end of the text. Wang and Bansal [7] inserted a misleading segment at a random position in the context. Liu et al. [8] proposed a method that generates adversarial examples automatically, perfecting the method mentioned above. Welbl [6] attacked the MRC model by using part-of-speech-based and entity-word-based replacement methods.

Schlegel et al. [10] chose to automatically generate the adversarial set, which corresponds to the original by adding negative words. In our research, we focused on solving problems where the context contained misleading sentences similar to the answer sentences. The semantics of the two sentence are naturally similar, but there are huge differences from the perspective of the problem, which is very common in practical applications.

Data augmentation and adversarial training have been used most widely for adversarial defenses. Jia and Liang [3], Wang and Bansal [7], Liu et al. [8] have improved the accuracy of the MRC model in oversensitivity problems by automatically generating adversarial samples. However, these methods use the adversarial training set generated by rules and have poor robustness in out-of-distribution data. Wang and Jiang [11] combined general knowledge with neural networks through data augmentation. Yang et al. [12, 13] used adversarial training to maximize the countermeasure loss by adding perturbations in the embedding layer. In addition, some studies have attempted to change the process of model inference. Chen et al. [14] decompose both the question and context into small units and construct the graph, converting the question answering into an alignment problem. Majumder et al. [9] re-rank the candidate answers according to the degree of overlap between the candidate sentence and the question. Zhou et al. [15] introduce rules based on external knowledge to regularize the model and adjust its output distribution. Yeh and Chen [16] trained the model by maximizing the mutual information between passages, questions, and answers, avoiding the effect of superficial biases in the data on the robustness of the model. Although these methods improve the robustness of MRC model to varying degrees, they do not consider the influence of model representation on robustness.

Contrastive Learning Recent years, contrastive learning has become a popular self-supervised representation learning technique, which has been extensively used in computer vision. The main idea of contrastive learning is to shorten the distance of positive sample pairs in the representation space. Chen et al. [17] proposed SimCLR, which constructs positive samples by data augmentation and constructs negative samples by random sampling in the same batch.

Contrastive learning has been applied to learn better sentence representations in the field of NLP. Gao et al. [18] proposed SimCSE, which uses dropout as a means of data augmentation and achieves good performance in natural language inference tasks. Wang et al. [19] proposed a method to construct semantic negative examples for contrastive learning to improve the robustness of the pre-trained language model. Yan et al. [20], Zhang et al. [21] used contrastive learning to solve the folding problem of model representation space and achieved good results in short text clustering and natural language inference tasks. However, all the above contrastive learning methods take the input context as a unit to improve the representation of the model, without considering the fine-grained representation of the context.

3 Method

In this section, we will introduce the multi-task machine reading comprehension learning framework via contrastive learning. The framework of the MRCCL is illustrated in Fig 1. The model can be divided into the MRC module, the Contrastive Learning module and multi-task joint training strategy. The MRC module is used to extract the correct answers, the contrastive learning module is used to improve the representation ability of the model, the MRC module and the contrastive learning module are jointly trained by multi-task joint training strategy. The MRC module shares the same encoding layer parameters as the contrastive learning module. Each of these two modules has its own loss function, respectively, but we combine the two loss functions to produce a joint loss function and adjust the model's parameters. Besides, the contrastive learning module only works in the training stage. In the next section, we will illustrate each module in the MRCCL in detail.

3.1 MRC Model Architecture

In MRC module, we adopt the most common extractive MRC model. The structure of MRC module is shown in the left of Fig. 1. It is composed of an encoder and a downstream multi-grain classifier. In the extractive MRC task, given dataset $D = \{C_i, Q_i, A_i\}_{i=1}^{n}$, where C_i denotes the context that needs to be understood by the model, Q_i denotes the question, A_i denotes the answer label corresponding to the question, and n denotes the size of the dataset. In the training stage, each input data is always composed of such triples. In the extractive MRC task, answer A_i is composed of the starting position A_i^s and the ending position A_i^e. The model needs to find the starting and ending positions of the answer from C_i according to the input Q_i.

Given input data $\{C_i, Q_i, A_i\}$, the input devoted as $encoder\_input$ is the concatenation of C_i and Q_i with special tokens [CLS] and [SEP] as [CLS] Q_i [SEP] C_i [SEP]. The $encoder\_input$ will be encoded by $encoder$ and produce the encoding result $encoder\_output \in \mathbb{R}^{m*d}$, where d denotes the maximum input sequence length and m denotes the dimension of the hidden layer. The MRC model calculates the answer position through the linear layer with dimension $2 * m$. It compress the encoded token in the $encoder\_output$ by weighting, and obtains answer's

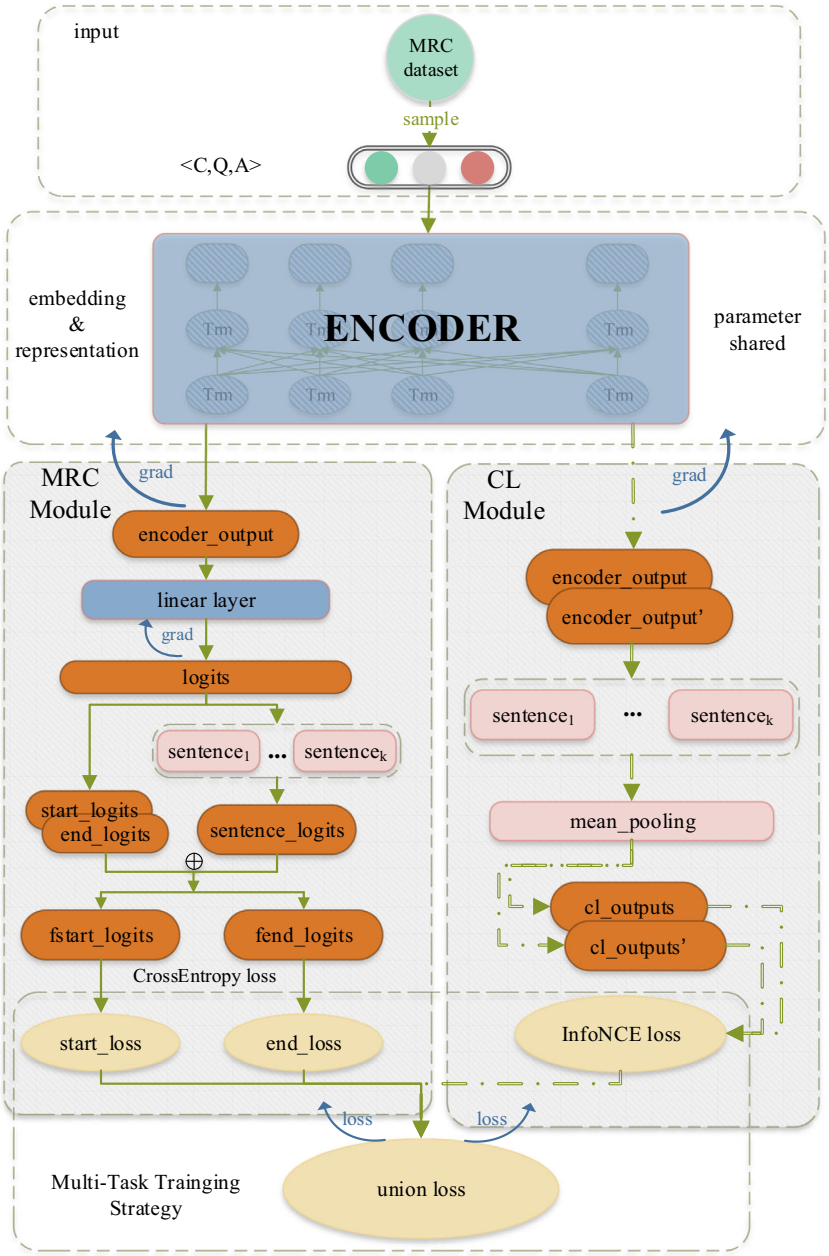

Fig. 1 The overview of multi-task machine reading comprehension learning framework via contrastive learning

position score $logits \in \mathbb{R}^{2*d}$. Then, the model calculate the sentence level score $sentence\_logits \in \mathbb{R}^{2*d}$ and the word level score $start\_logits \in \mathbb{R}^{d}$ and $end\_logits \in \mathbb{R}^{d}$ separately. Start_logits and end_logits can be obtained directly from logits. For each sentence in the context, the model calculates the mean value of the answer position score for each word in the sentence to get the sentence-level scores. Sentence-level scores is not a value, it has the same dimension as sentence length. Splice the sentence-level scores of all sentences in the context to get $sentence\_logit$. Finally, we add the word level score and the sentence level score to obtain $fstart\_logits \in \mathbb{R}^{d}$ and $fend\_logits \in \mathbb{R}^{d}$. $fstart\_logits$ and $fend\_logits$ are the start and end position scores used to generate the answer fragment

The $encoder\_output$ directly affects the calculation of $fstart\_logits$ and $fend\_logits$ in the MRC model and then affects the extraction of answer fragments. In other words, if the two vectors are represented at near positions in the context representation space, their answer scores will be close. Therefore, dense context representation space leads to poor robustness of the model and easy to get wrong answers under the adversarial attacks. Our MRC module uses the

cross-entropy function to calculate the loss according to the score. The loss function can be expressed as follows:

$$L = \frac{1}{2}(f_{CE}(fstart\_logits, A_i^s) + f_{CE}(fend\_logits, A_i^e)) \quad (1)$$

where f_{CE} denotes the cross-entropy function, A_i^s and A_i^e denote the starting position label and ending position label of the answer, respectively. The encoding result and linear layer weight will be changed by back propagation.

3.2 Contrastive learning in context representation space

In order to solve the problem of dense representation space in traditional MRC models, we introduce the contrastive learning into the MRC model, corresponding to the contrastive learning module described above. Due to the particularity of the extractive MRC task, we make improvements to the contrastive learning and call it Contrastive Learning in Context Representation Space (CLCRS). CLCRS is a type of supervised contrastive learning, it can be considered as the contrastive learning in the representation space of context. Common contrastive learning, which compares positive and negative samples taken from same batch, has little effect in distinguishing between answer and misleading sentences. To deal with this, CLCRS has a unique sampling strategy for MRC models. CLCRS samples sentences containing context information from the context as positive and negative samples, expanding the distance between the answer sentence and other sentences in the context. CLCRS can solve the problem of dense representation space of the original MRC model based on the pre-training model by enlarge the distance between different sentence vectors.

Specifically, CLCRS is shown in the right of Fig. 1. The input of CLCRS has the same form as the MRC module. Its input can be expressed as $ctx = \{C_i, Q_i, A_i\}_{i=1}^n$, where $C_i = \{c_i^1, \ldots, c_i^n\}$ denotes the context, c_i^j denotes each of the sentence in the context, Q denotes the question sentence, and A denotes the sentence where the answer lies. Different from the previous mainstream contrastive learning strategy of sampling negative samples from the same batch, we use contrastive learning to sample sentences in the context as negative samples for comparison. Specifically, following Gao et al. [18], we generate the positive sample corresponding to A by dropout and select other sentences in the input ctx, except question Q, as negative samples for contrastive learning. The effect of CLCRS on the ability to represent the model is shown in Fig. 2. For the input ctx, the encoding result $encoder\_output \in \mathbb{R}^{m*d}$ is obtained after encoding. We divide the $encoder\_output$ into different sentence vectors according to the original sentences and use the mean pooling to generate the vector representation $cl\_output \in \mathbb{R}^{k*m}$ of the sentences, where k denotes the number of sentences in the context and m denotes the dimension of the hidden layer. In CLCRS, we use InfoNCE as its loss function, and it is shown as follows:

$$-\log\left(\frac{e^{S(z_i,z_i')/\tau}}{\sum_{j=0}^{K} e^{S(z_i,z_j)/\tau}}\right) \quad (2)$$

where $S(\cdot)$ denotes the cosine similarity function, τ is a hyper parameter, z_i' denotes the positive sample, and z_j denotes the negative sample. By optimizing this loss function, the distance between each sentence in the context is enlarged, and the context representation space is expanded.

3.3 Multi task learning

Contrastive learning can expand the context representation space. In order to expand the representation space of the MRC model, we introduce a multi-task learning strategy. In our method, we combine the loss function of MRC and the loss function of contrastive learning, optimize the two modules simultaneously in the training stage. CLCRS only works in the training stage. Specifically, we share encoder parameters between the MRC module and CLCRS. Referring to the work of Liebel and Körner [22], we combine the loss functions of the two modules into a joint loss function as follows:

$$L_{union} = f(L_{mrc}, L_{cl}) = \frac{1}{a^2} l_{mrc} + log(1+a^2) + \frac{1}{b^2} L_{cl} + 1og(1+b^2) \quad (3)$$

where a and b are parameters that can be learned, L_{mrc} and L_{cl} are the loss functions of the MRC module and CLCRS, respectively.

4 Experiments

In order to verify the performance of our algorithm, we carried out several experiments and analyzed the experimental results. First, we introduce the datasets and experiments setting. Second, we evaluate our method on adversarial datasets in two kinds of baseline pre-train language models and compare it with other methods. Finally, we conduct the ablation study to verify the effectiveness of each module in MRCCL.

4.1 Datasets

We only use the SQuAD1.1 training set to train our model. And for the problem we want to solve, we generate an adversarial test set AddCfa for evaluating the robustness

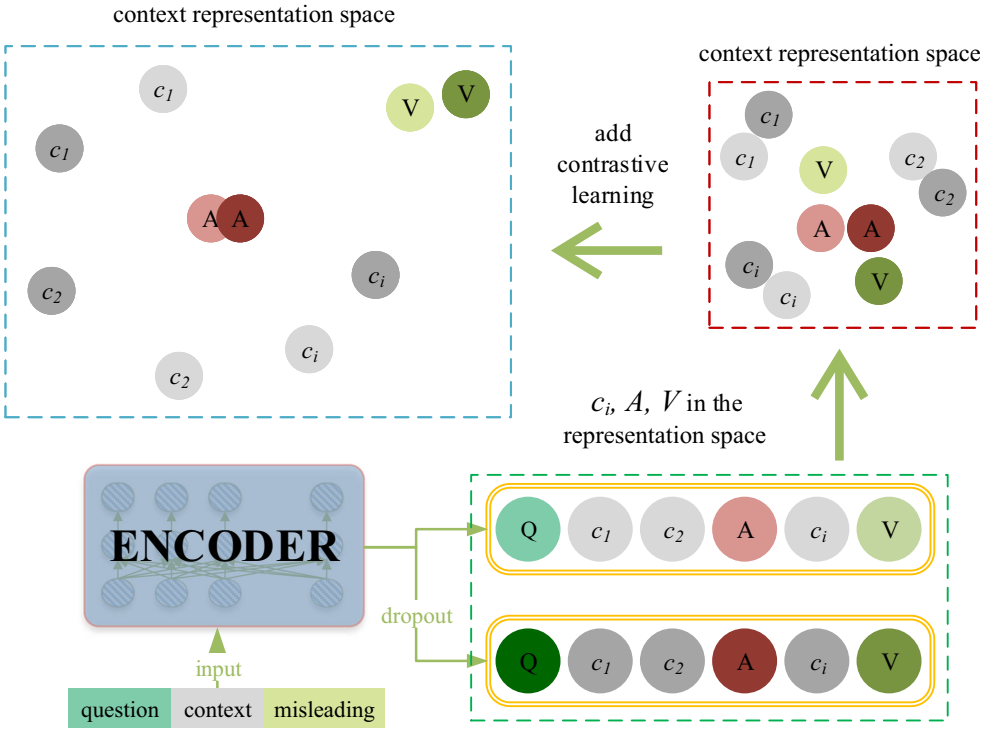

Fig. 2 The strategy of contrastive learning in our method. Q denotes question sentences, C denotes context, c_i denotes other sentences in the context except for the sentences where the answer lies, A denotes the sentences where the answer lies, and V denotes misleading sentences. Two sentences with similar colors represent a pair of positive samples generated by dropout. When the model encodes adversarial samples, the contrastive learning module can effectively distance the answer sentence from other sentences such as misleading sentences

of the model according to SQuAD1.1-dev set. Following Jia and Liang [3], the generation method of AddCfa is as follows: firstly, use the similar words in GLOVE [23] to replace the named entities and numbers in the answer sentence, then use the antonyms in WordNet [24] to replace the nouns and adjectives in the answer sentence to obtain the misleading sentence, and finally insert it into the back of the answer sentence in the context. The example in Table 1 is taken from AddCfa. We choose DEV [1], AddSent [3], AddCfa, and AddSentMod [3] as test sets to evaluate our approach.

4.1.1 Training datasets

- SQuAD1.1 training set [1]: One of the most authoritative datasets in the field of MRC. This dataset is selected from Wikipedia articles and annotates 87,599 question and answer pairs

4.1.2 Test datasets

- AddSent(AS) [3]: Adversarial test set in the field of MRC. The construction method of the adversarial sample is to convert the question sentence into an misleading sentence and append it to the end of the context by some rules. The dataset contains 2560 adversarial examples and 1000 normal examples.
- AddCfa(AC): Similar to AddSent, a adversarial test set in which misleading texts are converted from answer sentence through rules and crowdsourcing. The construction method of AddCfa is introduced in the previous subsection. The dataset contains 9620 adversarial instances and 10570 normal instances.
- AddSentMod(ASM) [3]: Same as AddSent but insert misleading text at the beginning of the context. This dataset has 2225 adversarial instances and 1000 normal instances.
- DEV [1]: The development set of SQuAD v1.0 in which contains 10570 question and answer pairs for evaluation.

4.2 Experiment settings

We selected five pre-trained language models for our experiments: BERT-base [25], BERT-large [25], BERT-large-wwm, RoBERTa-base [26], RoBERTa-large [26]. For the common set, AdamW optimizer is used during the training stage. All the parameters required for multi-task joint training are optimized by AdamW optimizer. The maximum input length for our model is set to 384. To

deal with long text, we chunk them into equally-spaced segments and use a sliding window of 128 size. We set the number of training epochs to 3. We use 0.1 for dropout on all layers and in attention. The temperature τ of InfoNCE loss and learning rate(lr) are the parameters that have the most impact on the accuracy of the model. We set τ in {0.05,0.1,0.15,0.2} and lr in {3e-5,4e-5,5e-5} to train our model and select the best result on the test set. For BERT model, τ=0.15 and lr=3e-5 are more conducive to the optimal performance of the model. For RoBERTa-base and RoBERTa-large model, the optimal parameters are τ=0.05 and lr=4e-5,3e-5. All models are implemented by PyTorch-1.7.1.

4.3 Results and analysis

4.3.1 Experiment on baseline model

We selected five different size pre-trained models as the baseline model to verify our algorithm: BERT-base, BERT-large, BERT-large-wwm, RoBERTa-base and RoBERTa-large. We test the F1 and EM of the model on four test sets as evaluation metrics. AVG is calculated according to the results of the model on DEV, AddSentMod, AddSent-adv and AddCfa-adv, and the results are used as the final metrics to evaluate the robustness of the model. All the results are shown in Table 2 and the best results are highlighted in bold.

Compared with AVG results, our model improved robustness across all baseline models. It has the best performance in RoBERTa-large model with a 2.3 improvement. Even the least significant improvements in bert-Large and Roberta-Base were 1.4.

The large model not only performs well on non-adversarial samples, but also has a high accuracy on adversarial samples. Compared with the base model, they have smaller differences in performance between the adversarial and non- adversarial samples, stronger robustness, and less vulnerability to attack from the adversarial samples. The larger model structure has stronger anti-jamming ability and can effectively improve the robustness of the model.

The RoBERTa model performs better on the adversarial test set than the BERT model. Our results show that RoBERTa-base and RoBERTa-large outperform BERT-base and BERT-large models in AVG indices.

4.3.2 Algorithm comparison

To further illustrate the advantages of our algorithm, we choose the following eleven methods for comparison: QAInfoMax [16], MAARS [9], R.M-Reader [27], KAR [11], BERT+Adv [12], ALUM [13], Sub-part Alignment [14], BERT+DGAdv [8], BERT+PR [15], HKAUP [28], and PQAT [13]. These eleven methods are used to improve the robustness of MRC model. The results are shown in Tables 3 and 4, and the best results are highlighted in bold. QAInfoMax [16] and MAARS [9] use part of the data in AddSent to verify their effectiveness. So, we divide the results into two tables. AddSent-small in Table 3 is a subset of AddSent in Table 4. Since most algorithms work on BERT, we also compare our results on BERT-base and BERT-large-wwm baseline models with those eleven algorithms.

In Table 4, compared with Sub-part Alignment, our algorithm has a 3.3 F1 increase on AddSent and a 15.0 F1 value increase on adversarial samples. Compared with MAARS (Majumder et al. 2021) in Table 3, which outperforms state-of-the-art defense techniques, the F1 of MRCCL on addSent is improved by 4.0 points. It shows that our method is better than the other eleven methods in adversarial test sets. On DEV, BERT-large-wwm+MRCCL

Table 2 Performance of MRCCL in five baseline models. Adv represents the adversarial sample in the test set

	DEV	AS		AC		ASM	AVG
Model		adv	all	adv	all		
$BERT_{base}$	88.2	58.2	66.5	80.4	84.5	63.7	72.6
$BERT_{base}$+MRCCL	88.3	62.1	69.4	81.5	85.0	64.4	74.1(+1.5)
$BERT_{large}$	90.9	63.1	70.8	84.8	88.0	70.9	77.4
$BERT_{large}$+MRCCL	90.9	68.1	74.2	84.9	88.0	**71.2**	**78.8(+1.4)**
$BERT_{large-wwm}$	92.4	70.5	76.7	87.6	90.1	77.9	82.1
$BERT_{large-wwm}$+MRCCL	**92.7**	73.1	78.6	**89.6**	**91.2**	79.5	83.7(+1.6)
$RoBERTa_{base}$	90.7	68.0	74.4	86.1	88.5	74.6	79.9
$RoBERTa_{base}$+MRCCL	90.8	70.9	76.4	85.8	88.4	77.7	81.3(+1.4)
$RoBERTa_{large}$	92.5	74.6	79.7	88.5	90.5	79.6	83.8
$RoBERTa_{large}$+MRCCL	92.5	**79.7**	**83.4**	89.3	91.0	**82.8**	**86.1(+2.3)**

Table 3 Performance of MRCCL in AS-small

Model	DEV	AS-small	
		adv	all
$BERT_{base} + QAInfoMax$	87.7/**82.1**	41.8/37.2	67.5/62.3
$BERT_{base} + MAARS$	80.2/71.1	**61.2**/53.6	71.8/63.4
$BERT_{base} + MRCCL$	**88.3**/ 81.0	60.1/**53.7**	**75.8/68.4**

Table 4 Performance of MRCCL in AS

Model	DEV	AS	
		adv	all
R.M-Reader	86.6	-	58.5
KAR	83.5	-	60.1
$BERT_{large}$+Adv	92.4	-	63.5
ALUM	90.8	-	60.4
$BERT_{base}$+DGAdv	87.9	-	59.7
$RoBERTa_{base}$+PQAT	92.3	-	64.7
HKAUP	82.4	-	65.5
BERT+PR	82.8	-	68.1
Sub-part Alignment	84.7	47.1	65.8
$BERT_{base} + MRCCL$	88.3	62.1	69.4
$BERT_{large-wwm} + MRCCL$	**92.7**	**73.1**	**78.6**

Table 5 Ablation experiments on the BERT model and the RoBERTa model

Model	DEV	AS		AC		ASM	AVG
		adv	all	adv	all		
$BERT_{base}$	88.2	58.2	66.5	80.4	84.5	63.7	72.6
$BERT_{base}$+MRCCL	88.3	62.1	69.4	81.5	85.0	64.4	**74.1**
-w/o CLCRS	88.3	59.2	67.1	80.0	84.3	65.1	73.2
-w/o Sentence Logits	88.2	59.0	67.0	82.0	85.3	63.8	73.3
$RoBERTa_{base}$	90.7	68.0	74.4	86.1	88.5	74.6	79.9
$RoBERTa_{base}$+MRCCL	90.8	70.9	76.4	85.8	88.4	77.7	**81.3**
-w/o CLCRS	90.9	70.5	76.3	85.5	88.3	76.9	81.0
-w/o Sentence Logits	90.4	66.0	72.8	84.9	87.8	73.2	78.6

Table 6 Influence of different training strategies on the model

	DEV	AS		ACFA		ASM
		adv	all	adv	all	
Pipline	88.2/80.7	57.6/50.6	65.9/58.8	80.4/71.4	84.5/76.3	65.4/58.4
MRCCL	88.3/81.0	62.1/55.4	69.4/62.3	81.5/72.6	85.0/76.9	64.4/57.1

Table 7 Comparison of training time for different models

	Training Time	EM	F1
BERT-orgin	1.5h	51.7	58.2
MRCCL	2.7h	55.4	62.1

Table 8 Impact of MRCCL on model representation

	Mean Distance	Range of Distance	AS-adv
BERT-orgin	0.842	(-0.142, 0.990)	58.2
MRCCL	-0.270	(-0.777, 0.742)	62.1

performers better than other methods. On AddSent, BERT-base+MRCCL and BERT-large-wwm+MRCCL are better than other methods.

4.3.3 Ablation study

We have demonstrated through ablation experiments that all modules in the MRCCL are indispensable. We performed our experiments on the BERT and RoBERTa models. The ablation experiments have similar results on models of different sizes, so here we show the results on the base version. We report the ablation test in terms of 1) w/o CLCRS: we remove the contrastive module in our model. 2) w/o sentence logit: in the MRC module, the model will no longer calculate sentence logits. The results are shown in Table 5.

Both the sentence logits and the contrastive learning modules have a beneficial effect on the robustness of the BERT model. Sentence logits improves the accuracy of the BERT and RoBERTa models on the adversarial sample, but is still lower than MRCCL. In fact, since the contrastive learning module is computed on a sentence level, the addition of sentence logits allows for a better introduction of the contrastive learning task into the MRC model, and there is no conflict between the two in terms of model improvement.

Finally, we explore which kinds of training strategies are suitable for our approach. We compare F1 of the model trained by sequential and the model trained by multi-task on four test sets. Only the training strategy is modified, and other settings are unchanged. The results are shown in Table 6. It shows that the multi task joint training method has better effect than the sequential training method. Multi task joint training plays an important role in our method. Pipline in the table represents that we first carry out the comparative learning task, and then carry out the training of MRC task.

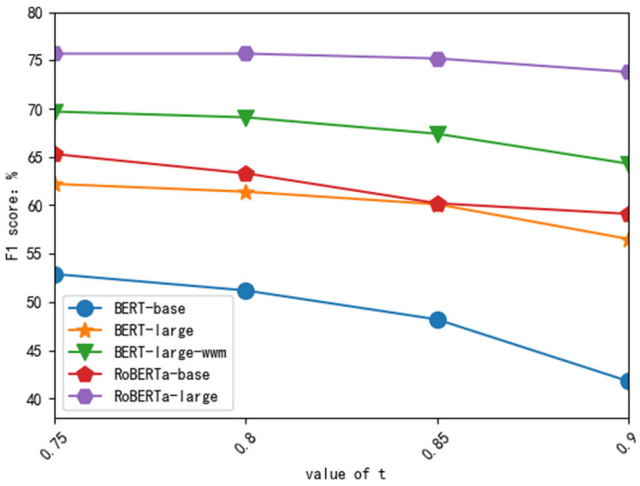

Fig. 3 The impact of a dense context representation space on model accuracy

4.4 Parameter efficiency of MRCCL

MRCCL has no additional new parameters and has the same number of model parameters as the baseline model. However, as the contrastive learning module requires the construction of positive samples, the amount of training data is doubled, resulting in longer model training times. The results of the model training time comparison are shown in Table 7.

5 Discussion

Context representation space for reading comprehension models CLCRS is the core of our algorithm. It is a special kind of contrastive learning suitable for MRC tasks, aims to solve the problem of over-density of the model context representation space. In our experiments, we found that the model's ability to represent context directly influenced its robustness. If the distance between the sentences in the encoding context is too close, the accuracy of the model on that example will be greatly reduced. We illustrate this even further through experimentation. We counted the relationship between the accuracy of the five pre-trained models on adversarial samples and the denseness of the representation space. The results are shown in Fig. 3. The answer sentence is the key sentence for extracting the answer, and we chose the distance between the answer sentence and other sentences in the context as a measure of whether the representation space is dense. We use the cosine similarity between sentence vectors to calculate the distance. Set the threshold t in {0.75, 0.8, 0.85, 0.9} and counted the F1 score of samples where the average distance between the answer sentence and the other sentences exceeded t.

As shown in the figure, the denser the context representation space is, the less accurate the model is on it.

Experiments demonstrate that the ability of the MRC model to represent each sentence in the context directly affects the robustness of the model.

Next, we experimentally illustrate the improvement of MRCCL for model representation. As before,we counted the distance between the answer sentence and other sentences in the context. The results are show in Table 8.

Mean Distance of the model has been reduced from 0.8 to -0.1. Range of Distance changed from (-0.142, 0.990) to (-0.777, 0.722). MRCCL effectively distances individual sentences in the context and generalises well on adversarial samples as well.

6 Conclusion and furtuer work

In this paper, we are committed to solving the problem of poor robustness of extractive MRC models. More specifically, we are committed to solving the problem that MRC model is prone to error on instances with additional misleading sentences, which is called the oversensitivity problem. We found that the poor robustness of the MRC model was caused by its overly dense context representation space. Therefore, we propose a multi task machine reading comprehension framework via contrastive learning called MRCCL. By introducing CLCRS into MRC model, we enhance the representation ability of MRC model, improve the robustness of the model, and then solve the oversensitivity problem. The experimental results show that our method is able to further improve model robustness and outperform state-of-the-art performance.

Acknowledgements This work is supported by sub-project of the National Key Research and Development Program (2020YFC0833404), the National Natural Science Foundation of China (62172352), the central government guides local science and technology development fund projects (226Z0305G), Project of Hebei Key Laboratory of Software Engineering(22567637H), Zhongyuanyingcai program-funded to central plains science and technology innovation leading talent program(204200510002) and the Natural Science Foundation of Hebei Province (F2022203028).

Data Availability Dataset was derived from the following public domain resources: https://github.com/rajpurkar/SQuAD-explorer

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Rajpurkar P, Zhang J, Lopyrev K, Liang P (2016) SQuAD: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 Conference on empirical methods in natural language processing, pp 2383–2392. Association for Computational Linguistics, https://doi.org/10.18653/v1/D16-1264. https://aclanthology.org/D16-1264
2. Lai G, Xie Q, Liu H, Yang Y, Hovy E (2017). In: Proceedings of the 2017 Conference on empirical methods in natural language processing, pp 785–794. Association for Computational Linguistics, https://doi.org/10.18653/v1/D17-1082. https://aclanthology.org/D17-1082
3. Jia R, Liang P (2017) Adversarial examples for evaluating reading comprehension systems. In: Proceedings of the 2017 Conference on empirical methods in natural language processing, pp 2021–2031. Association for Computational Linguistics, Copenhagen, Denmark. https://doi.org/10.18653/v1/D17-1215. https://aclanthology.org/D17-1215
4. Jin D, Jin Z, Zhou JT, Szolovits P (2020) Is bert really robust? a strong baseline for natural language attack on text classification and entailment, vol 34. https://doi.org/10.1609/aaai.v34i05.6311
5. Garg S, Ramakrishnan G (2020) BAE: BERT-based adversarial examples for text classification. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), pp 6174–6181. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.498. https://aclanthology.org/2020.emnlp-main.498
6. Welbl J, Minervini P, Bartolo M, Stenetorp P, Riedel S (2020) Undersensitivity in neural reading comprehension. In: Association for computational linguistics, https://doi.org/10.18653/v1/2020.findings-emnlp.103. https://aclanthology.org/2020.findings-emnlp.103, pp 1152–1165
7. Wang Y, Bansal M (2018) Robust machine comprehension models via adversarial training. In: Proceedings of the 2018 Conference of the north american chapter of the association for computational linguistics: Human language technologies, Volume 2 (Short Papers), pp 575–581. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-2091. https://aclanthology.org/N18-2091
8. Liu K, Liu X, Yang A, Liu J, Su J, Li S, She Q (2020) A robust adversarial training approach to machine reading comprehension. Proceedings of the AAAI conference on artificial intelligence 34(05):8392–8400. https://doi.org/10.1609/aaai.v34i05.6357
9. Majumder S, Samant C, Durrett G (2021) Model agnostic answer reranking system for adversarial question answering. In: Proceedings of the 16th Conference of the european chapter of the association for computational linguistics: Student research workshop, pp 50–57. Association for Computational Linguistics. https://aclanthology.org/2021.eacl-srw.8
10. Schlegel V, Nenadic G, Batista-Navarro R (2021) Semantics altering modifications for evaluating comprehension in machine reading. Proceedings of the AAAI Conference on Artificial Intelligence 35(15):13762–13770
11. Wang C, Jiang H (2019) Explicit utilization of general knowledge in machine reading comprehension. In: Proceedings of the 57th Annual meeting of the association for computational linguistics, pp 2263–2272. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1219. https://aclanthology.org/P19-1219
12. Yang Z, Cui Y, Che W, Liu T, Wang S, Hu G (2019) Improving machine reading comprehension via adversarial training. arXiv:1911.03614
13. Yang Z, Cui Y, Si C, Che W, Liu T, Wang S, Hu G (2021) Adversarial training for machine reading comprehension with virtual embeddings. In: Proceedings of *SEM 2021:

The Tenth joint conference on lexical and computational semantics, pp 308–313. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.starsem-1.30. https://aclanthology.org/2021.starsem-1.30
14. Chen J, Durrett G (2021) Robust question answering through sub-part alignment. In: Proceedings of the 2021 Conference of the north american chapter of the association for computational linguistics: human language technologies, pp 1251–1263. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.98. https://aclanthology.org/2021.naacl-main.98
15. Zhou M, Huang M, Zhu X (2020) Robust reading comprehension with linguistic constraints via posterior regularization. IEEE/ACM Trans Audio, Speech, Language Process 28:2500–2510. https://doi.org/10.1109/TASLP.2020.3016132
16. Yeh Y-T, Chen Y-N (2019) QAInfomax: Learning robust question answering system by mutual information maximization. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), pp 3370–3375. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1333. https://aclanthology.org/D19-1333
17. Chen T, Kornblith S, Norouzi M, Hinton G (2020) A simple framework for contrastive learning of visual representations. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International conference on machine learning proceedings of machine learning research, vol 119, pp 1597–1607. PMLR. https://proceedings.mlr.press/v119/chen20j.html
18. Gao T, Yao X, Chen D (2021) Simcse: Simple contrastive learning of sentence embeddings. arXiv:2104.08821
19. Wang D, Ding N, Li P, Zheng H (2021) CLINE: Contrastive learning with semantic negative examples for natural language understanding. In: Proceedings of the 59th Annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (Volume 1: Long Papers), pp 2332–2342. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.181. https://aclanthology.org/2021.acl-long.181
20. Yan Y, Li R, Wang S, Zhang F, Wu W, Xu W (2021) ConSERT: A contrastive framework for self-supervised sentence representation transfer. In: Proceedings of the 59th Annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (Volume 1: Long Papers), pp 5065–5075. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.393. https://aclanthology.org/2021.acl-long.393
21. Zhang D, Nan F, Wei X, Li S-W, Zhu H, McKeown K, Nallapati R, Arnold AO, Xiang B (2021) Supporting clustering with contrastive learning. In: Proceedings of the 2021 Conference of the north american chapter of the association for computational linguistics: Human language technologies, pp 5419–5430. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.427. https://aclanthology.org/2021.naacl-main.427
22. Liebel L, Körner M (2018) Auxiliary tasks in multi-task learning. arXiv:1805.06334
23. Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on empirical methods in natural language processing (EMNLP), pp 1532–1543
24. Miller GA (1995) Wordnet: a lexical database for english. Commun ACM 38(11):39–41
25. Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (Long and Short Papers), pp 4171–4186. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423. https://aclanthology.org/N19-1423
26. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: A robustly optimized bert pretraining approach. arXiv:1907.11692
27. Hu M, Peng Y, Huang Z, Qiu X, Wei F, Zhou M (2018) Reinforced mnemonic reader for machine reading comprehension. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, pp 4099–4106. International Joint Conferences on Artificial Intelligence Organization???. https://doi.org/10.24963/ijcai.2018/570. https://doi.org/10.24963/ijcai.2018/570
28. Wu Z, Xu H (2020) Improving the robustness of machine reading comprehension model with hierarchical knowledge and auxiliary unanswerability prediction. Knowl Based Syst 203:106075. https://doi.org/10.1016/j.knosys.2020.106075

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Jianzhou Feng received M.S. and Ph.D. degrees from Yanshan University, Qinhuangdao, China, in 2007 and 2013, respectively. He is an Associate Professor of Yanshan University. His research interests include natural language processing and knowledge graph.

Jiawei Sun is currently pursuing the M.S. degree in Computer Science and Technology from Yanshan university. His research interests include machine reading comprehension.

Di Shao is currently pursuing the M.S. degree in Computer Science and Technology from Yanshan university. His research interests include machine reading comprehension.

Jinman Cui is currently pursuing the M.S. degree in Computer Science and Technology from Yanshan university. Her interests include knowledge graph.