Soft Computing (2023) 27:13435–13443
https://doi.org/10.1007/s00500-022-07678-5

FOCUS

# Transformer fast gradient method with relative positional embedding: a mutual translation model between English and Chinese

Yuchen Li[1] · Yongxue Shan[1] · Zhuoya Liu[1] · Chao Che[1] · Zhaoqian Zhong[1]

Accepted: 18 November 2022 / Published online: 30 November 2022


**Abstract**
Machine translation uses computers to transform one natural language into another. Text-like neural machine translation tasks cannot fully identify the sequence order of texts or the long-term dependence between words, and they suffer from excessive translation and mistranslation. To improve the naturalness, fluency, and accuracy of translation, this study proposes a new training strategy, the transformer fast gradient method with relative positional embedding (TF-RPE), which includes the fast gradient method (FGM) of adversarial training and relative positional embedding. The input sequence is founded on the transformer model, and after the word embedding matrix converts a word vector in the word embedding layer, the positional encoding can be embedded in it through relative positional embedding, helping the word vector to better save the linguistic information of the word (meaning, semantics). The addition of FGM adversarial training to the multi-head attention encoder mechanism strengthens the training of word vectors and reduces the phenomenon of miss-or-error translation, enabling significant improvement of the overall computational efficiency and accuracy of the model. TF-RPE can also provide satisfactory high-quality translations for the low-resource corpus. Extensive ablation studies and comparative analyses validate the effectiveness of the scheme, and TF-RPE achieves an improvement of average 3+ Bilingual evaluation understudy scores compared with the SOTA methods.

**Keywords** Neural machine translation · Transformer model · Relative positional embedding · Multi-attentional mechanism · FGM adversarial training algorithm

## 1 Introduction

Machine translation has developed from the early rule- and corpus-based statistical methods to neural machine translation based on deep learning. Translation methods based on the transformer model (Vaswani et al. 2017) can straightway compute the relativity between words without hidden layers. There may exist similar position information between words. However, this cannot be represented to exactly preserve the lexical and semantic information of words, and it can generate over- and under-translation problems, especially for source words that lack or have a large number of candidate translations (Tu et al. 2016). To fill these research gaps, we propose a relative positional embedding method on the text of the corpus (Shaw et al. 2018), which can better capture each specific position between words. Our work encodes exact information about the location of each word, enhances the correlation between them, and enables the translation model to better identify adversarial samples (Goodfellow et al. 2015), for better convergence and performance.

The transformer-based model has been shown to be effective in advanced neural machine translation (NMT) systems (Sutskever et al. 2014), which follow the standard sequence-to-sequence structure and are composed of an encoder and a decoder. However, two problems often arise when using

Communicated by Oscar Castillo.

✉ Zhaoqian Zhong
zhaoqianzhong@gmail.com

Yuchen Li
liyuchen@s.dlu.edu.cn

Yongxue Shan
shanyx1000@gmail.com

Zhuoya Liu
liuzhuoya@s.dlu.edu.cn

Chao Che
chechao@gmail.com

[1] Key Laboratory of Advanced Design and Intelligent Computing Ministry of Education, Dalian University, Dalian 116622, Liaoning, China

the traditional transformer model for machine translation. First, position information is lost when positional encoding is performed on the translation corpus, making the model completely ignore the sequence order, e.g., "Apple loves like apple Inc." In English-to-Chinese translation using absolute position encoding, after mapping through the weight matrix, the position information may disappear, and the two "apples" in the sentence may be translated as "company" or "apple". Second, there are often problems of over- and under-translation because the models are usually less robust in the case of non-convergence, with the result that training cannot help the model to better capture the correlation between words in a sentence.

As shown in Fig. 1, the less shallow each color is, the stronger the relationality of Chinese-English words is. For the sentence "他们上周在巴黎", meaning "They were in Paris last week," in the traditional machine translation, the Chinese word "巴黎" (Paris) is over-translated to "Paris" twice, while "周" (week) is mistakenly untranslated.

A transformer fast gradient method with relative positional embedding(TF-RPE) with adversarial training is proposed to cope with this problem for machine translation tasks. TF-PRE can effectively capture local and global interdependencies between texts by changing absolute position encoding to relative positional encoding. We use a multi-head attention mechanism (MAM) and feedforward neural network (FNN) in the encoder layer to train the feature vectors of Chinese text in the corpus, and add fast gradient method (FGM) adversarial training to MAM to enhance the training of word vectors. Three sub-layers-masked multi-head self-attention, encoder-decoder attention, and FNN-are utilized in the decoder layer. The English corpus is first processed for RPE, following which, the processed word vectors are decoded; after the masked multi-headed attention layer, the output results are passed to the encoder–decoder attention layer together with the output of the encoder layer and processed by the FNN layer to acquire the semantic links associated with the Chinese and English utterances for accurate translation. The main contributions of our work are as follows.

- We solve the problem of remote fading and full symmetry in absolute position encoding and use multi-head attention and a FNN encoder to capture long-range word dependencies, so we can accurately represent relative positions of arbitrary length;
- We introduce FGM (Miyato et al. 2016) to the multi-head attention mechanism of the transformer model to better capture the correlation between words in a sentence, regularizing the parameters to improve robustness and generalization, and alleviating over- and under-translation;

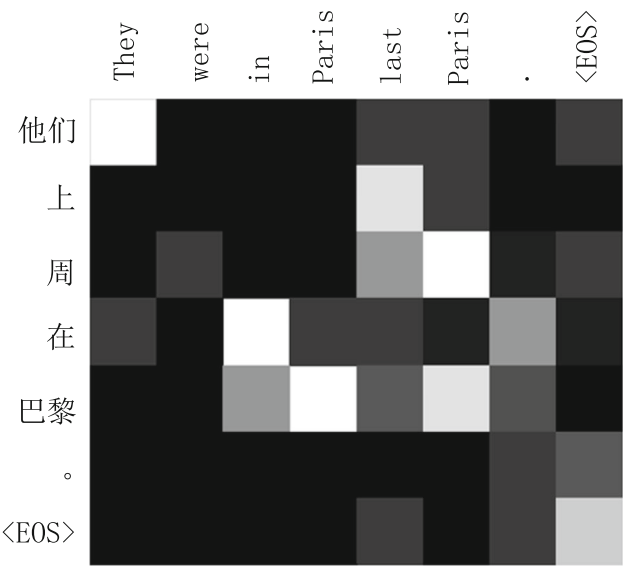


**Fig. 1** Example of over-translation and under-translation generated by NMT

- We manually label the comment data on the UN corpus website to construct a dataset of Chinese and English utterances in social news. It is experimentally proved that a higher accuracy of our method is performed, comparing with other baseline models in NMT tasks.

## 2 Related work

With the widespread application of deep learning, researchers began to apply these methods to natural language processing. In 2006, the literature (Hinton et al. 2006) solved the neural network training problem by layer-by-layer pretraining. With the improvement of computing power, such as through parallel computing and graphics processing units, neural networks have gained importance. In 2013, the literature (Kalchbrenner and Blunsom 2013) proposed machine translation using neural networks, and the literature (Sutskever et al. 2014; Cho et al. 2014b, a; Bahdanau et al. 2014) and others proposed neural machine translation models based on encoder–decoder structures, bringing machine translation into the era of deep learning. Neural machine translation has gradually surpassed statistical machine translation on multiple language pairs. The literature (Junczys-Dowmunt et al. 2014) used the United Nations Parallel Corpus v1.0 to compare neural machine translation and phrasal statistical machine translation, and the former performed better on 27 of 30 language pairs. For Chinese-related tasks, such as Chinese-English, -French, and -Russian translation, neural machine translation achieved 6 9 higher BLEU scores than the traditional machine translation. At the Workshop on

Machine Translation(WMT) 2016, a neural machine translation system developed at the University of Edinburgh outperformed phrase- and syntax-based statistical machine translation on the English-to-German translation task (Wilks 1993), which demonstrated the power of NMT. In 2017, the literature (Vaswani et al. 2017) proposed an attention mechanism-based transformer model, which substantially improved both the training speed and translation quality of the model. Although these methods have somewhat improved the efficiency and accuracy of machine translation, they achieved poor performance in cases of low-quality parallel corpus or where translations generated by NMT-based models is of low fluency and accuracy.

Recently, with the proposed sequence-to-sequence translation architecture and the maturity of the transformer model, the translation quality and efficiency of NMT methods have been greatly improved, and translation models such as LightConv, PDH, DTMT and PBSMT(Wu et al. 2019; So et al. 2019; Meng and Zhang 2019; Chen et al. 2019) have made significant breakthroughs in NMT tasks. Some researchers addressed the problems of low-resource conditions and inadequate modeling of future information. The literature (Chiang et al. 2019) proposed an APN model to classify word features, and to achieve better translation results. The literature (López-Gonzále et al. 2020) proposed an alternative method to achieve distance-based formation, and allows each word to converge to desired distances. The literature (Liao et al. 2021) expanded the training data through an iterative back-translation method. It trained a neural network that translates from the target language to the source language, using the target language corpus alone to obtain the text corresponding to the source language and adding the obtained pairs to the parallel corpus together with the training, which expands the data set to some extent. However, in practice, it is often necessary to combine a Quality estimation system (QES) to further clean and filter the synthetic data, which is difficult to establish for many low-resource languages, but with the help of the literature (Rubio et al. 2021; Mújica-Vargas 2021) and others, the evaluation system has become much easier to be established. As an extension of the above model, the literature (Abdulmumin et al. 2021) proposed a data selection method to replace the original evaluation system, which can sort the relevance of the original single corpus according to the degree of domain matching of the test set and select the most matched sentence in each iteration to generate synthetic data, with good results. The literature (Shi et al. 2021) compressed long-distance-dependent context information in an enhanced memory bank that facilitates streaming decoding under low latency conditions. The literature (Rubio 2021; Rubio et al. 2022) incorporated Newton algorithm into the gradient descent algorithm to improve the accuracy of machine translation. The literature (Li et al. 2020a) proposed a transformer-based skipping sublayer regularization method that enhances perturbation training by randomly removing a sublayer, but ignores the position information between words.

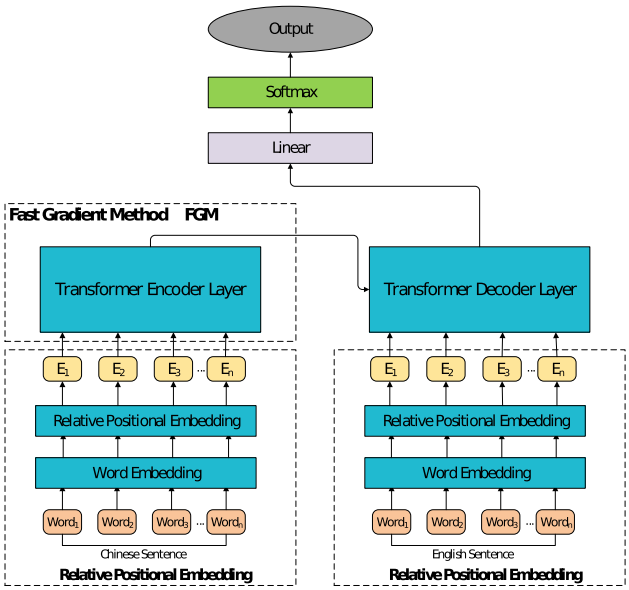


**Fig. 2** Architecture of the TF-RPE model

Although the above methods are effective in some areas of neural machine translation, relatively few are geared toward the effective identification of the sequence order of text and capturing of the correlations between words in a sentence. Hence, we design a neural network to effectively extract the position information between words and the local and global interdependence between texts, to better identify the position information between words and solve the problem of over- and under-translation, thus improving the overall translation level.

## 3 TF-RPE machine translation

We present the TF-RPE model based on relative positional embedding and adversarial training, which achieves competitive performance on the datasets we consider. The method separately performs word embedding on the Chinese and English corpus and then trained by the improved model to obtain the desired translation model. The framework is shown in Fig. 2.

### 3.1 Method

Our approach has three stages. First, given the Chinese and English source sentences in the dataset, the text will be transformed to word vectors $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_n)$, respectively, through a word embedding matrix in the word embedding layer; then,

the location information formula is used to embed the location code in the word vector, using RPE to add a position-determined encoding vector at each position such that $\bar{X} = (x_1 + p_m, x_2 + p_m, \ldots, x_n + p_m)$, and $\bar{Y} = (y_1 + q_m, y_2 + q_m, \ldots, y_n + q_m)$, to obtain the desired position embedding information; finally, the obtained vector matrices are input to the encoder and decoder layers of the N-layer transformer model for training, and the training data of the encoder layer are adjusted by FGM.

## 3.2 Position embedding layer

Absolute position encoding introduces the problem of remote fading, i.e., the greater the relative distance of the inputs, the weaker their mutual relevance, so we use relative positional embedding for encoding to obtain more accurate word embedding information.

### 3.2.1 Absolute position information encoding

The incorporation of position information in the input constitutes a general approach to absolute position encoding (Vaswani et al. 2017), whose attention formula is

$$\begin{cases} PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \\ PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \end{cases} \tag{1}$$

where $pos = 0, 1, \ldots, L-1$ denotes the position of the tokens, where $L$ is the length of each sentence, $i$ is the dimension of the vector, and $d_{model}$ is the hidden layer size. For any fixed offset $k$, $PE_{pos+k}$ can be indicated as a linear function of $PE_{pos}$.

In machine translation, the length of translated sentences should theoretically be infinite, but the design of absolute position encoding limits the length of the conversational text and does not work well with remembering the above in long texts.

### 3.2.2 Relative position information encoding

Absolute positional encoding has the problem of full symmetry, which is why the transformer model fails to identify the position, specifically that the function naturally satisfies the identity $f(x, y) = f(y, x)$, so, from the result, it is impossible to distinguish whether the input is $(x, y)$ or $(y, x)$.

To address the full symmetry problem, we use the RPE for encoding (Shaw et al. 2018) and define a relative positional

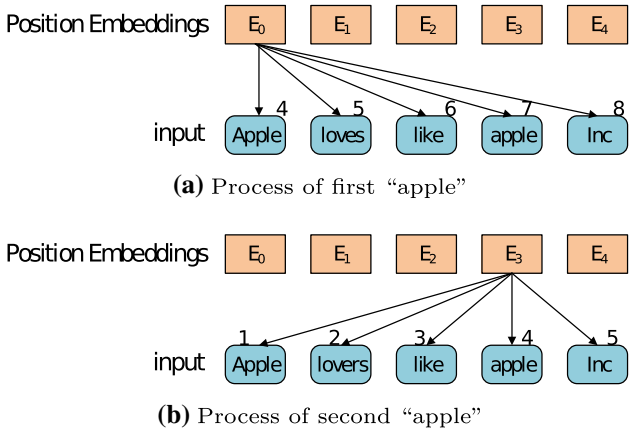


**Fig. 3** Example diagram of relative positional embedding

encoding formula as shown below:

$$\begin{cases} a_{ij}^K = w_{\text{clip}(j-i,k)}^K \\ a_{ij}^V = w_{\text{clip}(j-i,k)}^V \\ \text{clip}(x, k) = \max(-k, \min(k, x)) \end{cases} \tag{2}$$

where the edge of input element $x$ can be represented by two vectors $a_{ij}^K, a_{ij}^K \in R^{d_a}$, $k$ is the maximum value of the intercepted relative position, and $w^K = \left(w_{-k}^K, \ldots w_k^K\right)$ and $w^V = \left(w_{-k}^V, \ldots w_k^V\right)$ are the learned relative positions, where $w^K, w^V \in \mathbb{R}^{d_a}$.

Figure 3 shows the RPE process, taking "Apple lovers like apple Inc." as an example. Fig. 3a shows the process of position embedding representation when computing the first "apple". We take the fourth index as the center. When transformer calculates the position information of "apple" and "like", "apple" uses the 4th position code, "like" uses the 6th position code, "like" is located to the right of "apple", and the relative distance is 2, so the 6th embedding vector is used. In Fig. 3b, when calculating the information relationship between the second "apple" and other words, such as calculating its information with "like" on the left, the position coding of "like" is the third embedding vector, because it is 1 distance away from "apple" on the left side of "apple", and the index should use the third embedding vector.

## 3.3 Encoder layer

The encoder layer consists of the MAM and FNN. The Chinese corpus uses the word vector processed by the relative positional embedding layer as the input vector and obtains a context vector through the MAM, which is summed with the input vector and normalized to the FNN layer. This output vector is summed with the input of the layer, and the result is normalized as the output of the encoder.

### 3.3.1 Multi-head attention mechanism

The MAM allows the model to jointly attend to information from different representation subspaces at different positions, which helps the network capture richer feature information. Its computation is equivalent to the integration of several single-head self-attention mechanisms, in which a single-head self-attention mechanism, Scaled dot-product attention (SDPA), is used to compute the attention output of the input vector sequence query matrix $Q$, key-value matrix $K$ and value matrix $V$,

$$\mathrm{SDPA}(Q, K, V) = V \cdot \mathrm{soft\,max}\left(\frac{Q^T \cdot K}{\sqrt{d_k}}\right) \tag{3}$$

and the multi-attention mechanism is

$$\mathrm{head}_i = \mathrm{Attention}\,(Q_i, K_i, V_i) \tag{4}$$

$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}\,(head_1, \mathrm{head}_2, \ldots, \mathrm{head}_h)\, W^0 \tag{5}$$

where $Q_i$, $K_i$ and $V_i$, respectively, represent the query, key-value, and value matrices of the $i$th attention header $head_i$, $i \in [1, h]$; $d_k$ is the scale factor; $W^0 \in \mathbb{R}^{hd_v \times d_{\mathrm{model}}}$ is a parameter matrix.

However, even though the encoder layer of the transformer model has the advantages of a simple structure and less computational complexity, it has more over- and under-translation. We introduce the FGM adversarial training algorithm (Miyato et al. 2016) in the multi-head attention mechanism so that the model can effectively identify more adversarial samples and alleviate this problem.

Adversarial training introduces noise to improve robustness and generalization by regularizing the parameters, enabling improved accuracy in machine translation. We experimented with a supervised dataset, with cross-entropy loss,:

$$\mathrm{loss} = -\log p\,(y \mid x + r_{\mathrm{adv}}; \theta) \tag{6}$$

where $x$ is the original input sample and $r_{\mathrm{adv}}$ is the worst-case perturbation value applied to the current model $p$ with fixed parameter $\theta$.

FGM generally refers to the perturbation $r_{\mathrm{adv}}$, which scales according to the gradient to obtain a better adversarial sample, and the specific formula of FGM is as follows:

$$r_{\mathrm{adv}} = \epsilon \cdot g/\|g\|_2 \text{ where } g = \nabla_x L(\theta, x, y) \tag{7}$$

where $g$ is the gradient of the function at $x$, the gradient of the adversarial sample is added to the original sample, so that the model can pay more attention to the untranslated source words and reduce under-translation. $L$ is the loss function

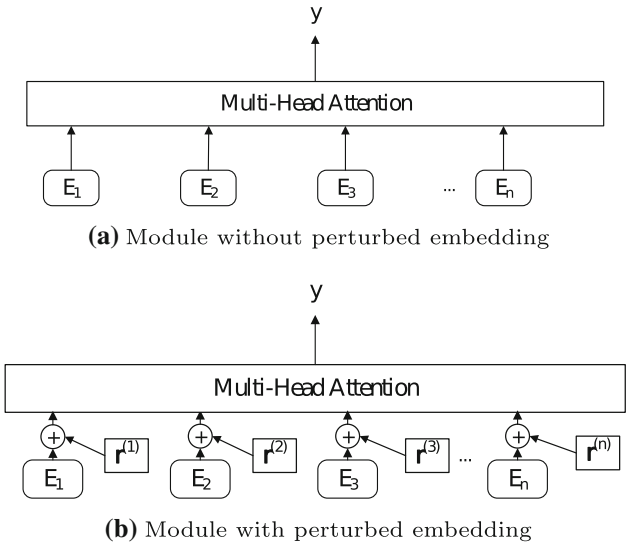


**Fig. 4** Multi-head attention module based on adversarial training

and $\epsilon$ is the hyperparameter, which limits the perturbation size to a certain range, with a default value of 1.0.

In the calculation of the multi-head attention mechanism of the encoder layer, we calculate the loss during forward propagation, the gradient value by backpropagation, the gradient of the embedding layer, and its norm value. We calculate $r_{adv}$ by equation Equation 7, which is accumulated to the sample of the original embedding, i.e., $x + r$, to get the adversarial sample, which is used to calculate the new loss value to get its gradient. The original and adversarial gradients are combined, and the parameters are updated for better model convergence and reduced over-translation, as shown in Fig. 4.

Figure 4a shows a regular multi-head attention module, where a continuous vector is the input, and the output is passed to the next layer. Figure 4b shows the input of the adversarial generative network, where $r$ is the added perturbation.

The introduction of this training method in the multi-head attention module can effectively compensate for the shortcomings of the traditional encoder model with an attention module and enhance the robustness of the model to adversarial samples, thus alleviating over- and under-translation.

### 3.3.2 Feedforward neural network layer

Assuming that the multi-head attention has $h$ heads, the obtained $h$ matrices are spliced into a matrix, which is linearly transformed to the original dimension and passed into the FNN layer, which has two sub-layers, ReLU activation and linear activation function in Equation 8, the role of these two activation function is to ensure that the gradient is always 1 when $z$ is inputted over zero, thereby improving the opera-

tion speed of the neural network gradient descent algorithm.

$$FFN(Z) = \max(0, ZW_1 + \mathrm{b}_1)\, W_2 + b_2 \tag{8}$$

where $W_1$ and $W_2$ are learnable square weight matrices, and $b_1$ and $b_2$ are random bias vectors.

### 3.4 Decoder layer

The decoder layer has masked multi-head self-attention, encoder–decoder attention, and feedforward neural network sub-layers. The English corpus uses the word vector processed by the relative positional embedding layer as the input vector, and the output is summed and normalized after the masked multi-head attention layer. The normalized result is passed to the encoder–decoder attention layer together with the encoder output. The output of the decoder layer is obtained using the feedforward neural network layer. The computation process is as follows.

The masked multi-head self-attentive layer has as its input the word vector of the target utterance. If the input sequence is $X = (x_1, x_2, \ldots, x_n)$, when predicting for $x_i$, the masked multi-head attention layer only performs attention computation on the $(x_1, x_2, \ldots, x_{i-1})$ sequence, to prevent the use of future information in training, the computation process is shown in 3.3.1.

The input layer has two parts, consisting of the output of the masked multi-head attention layer and the encoder. The output of the mask multi-head attention layer is used as query vector $Q$, and the output of the encoder is used as key-value pairs $(K, V)$ for attention, which makes the decoder fully consider the information of the source language when decoding the current word.

The third layer, FNN, has as its input the output of the encoder-decoder attention layer. It is the same as the FNN layer in the encoder(the calculation process is shown in 3.3.2). The feature vector output by this layer is obtained as the output vector of each word probability after linear variation and the softmax function.

## 4 Experiment and analysis

### 4.1 Datasets

We experimentally investigated the effect of Chinese-English translation under conditions of scarce or sufficient corpus resources, using the following datasets.

- We randomly selected 120000 English-Chinese sentence pairs from the United Nations parallel corpus v1.0 (Ziemski et al. 2016) as a poor-resourced dataset and performed a series of preprocessing operations, such as removing empty or repeated text, stop words, and useless characters. The processed information included 112847 English-Chinese language pairs. Then, we constructed an English-Chinese dictionary for the English-Chinese review dataset, including 69191 unrepeated English words and 4024 unrepeated Chinese words. We divided the dataset into training, validation, and test sets into the proportion of 8:1:1 to obtain a dataset for the neural machine translation task, whose data distribution is shown in Table 1.
- The training set between Chinese(Zh) and English(En) is composed of news-commentary-v13 with $0.3M$ sentence pairs and OpenSubtitles2015 with $9.2M$ pre-processed sentence pairs; in total $9.5M$ sentence pairs comprises our well-resourced dataset. The test set was the news-test-2017 with size $2K$, and the development set was the news-dev-2017 with size $2K$ from WMT2017.

**Table 1** Data distribution of English-Chinese corpus for machine translation

| Corpus | Sentences | Unrepeated words |
|---|---|---|
| English | 112847 | 69191 |
| Chinese | 112847 | 4024 |

### 4.2 Parameter settings and evaluation indicators

The number of attention heads, transformer blocks and hidden units used in the TF-RPE was 8, 15 and 521 separately. The word vector dimension for training was 512, and the maximum length of sentence was 256 in the English corpus and 97 in the Chinese corpus. We chose the Adam optimizer for training, with batch size 16 and learning rate 1e-3. From one to twelve, the value of transformer network layers k was experimented and the default value was chosen as k = 2. In avoidance of overfitting, the default value of 0.1 was used through dropout regularization.

We used the BLEU (Papineni et al. 2002) score as an evaluation metric. Given standard and automatically generated translations, the translation candidate was evaluated by the degree of n-metric matching, as follows:

$$\mathrm{B\,L\,E\,U} = \mathrm{B\,P} \times \exp\left(\sum_i \frac{1}{N}\right) \log p_n \tag{9}$$

where $BP$ is a length penalty factor, $N$ is the longest n-gram, and $p_n$ is the modified n-gram precision.

**Table 2** BLEU scores of models on English(En)-Chinese(Zh) corpus datasets

| Model | Low(120K) | | High(9.5M) | |
|---|---|---|---|---|
| | Zh–En | En–Zh | Zh–En | En–Zh |
| CNN | 16.64 | 19.79 | 21.98 | 27.34 |
| Transformer | 17.26 | 20.73 | 23.06 | 29.57 |
| BERT-fused | 18.93 | 21.45 | 25.14 | 32.43 |
| SDT-RPR | 18.77 | 21.96 | 25.83 | 33.41 |
| TF-RPE(Ours) | 19.59 | 22.38 | 26.96 | 34.72 |

## 4.3 Comparison with state-of-the-art methods

We compared TF-RPE with the following models, which include two traditional neural machine translation models, CNN (Gehring et al. Gehring et al. 2017) and transformer (Vaswani et al. Vaswani et al. 2017), and two typical baseline models as follows:

- BERT-fused (Zhu et al. Zhu et al. 2020) is a fine-tuned, pretrained model that fuses the representation with encoder and decoder in each layer of NMT model through an attention mechanism, using the pretrained model as input to NMT.
- SDT-RPR (Li et al. Li et al. 2020b) is a deep-level transformer-based model that trains deeper models by a shallow-to-deep strategy, using a step-by-step overlay method, with a sparsely connected approach to optimize the translation effect.

We re-ran all models based on our constructed English-Chinese text corpus. The experimental results gained from English-Chinese dataset are shown in Table 2, where the best performance is in the last line.

As shown in Table 2, our model outperforms advanced neural machine translation models in terms of BLEU metrics, which shows it improves the recognition of information about the position of each word in a sentence and better obtains important local association information. CNN performs poorly, probably because the model loses valuable information in the pooling layer during training and ignores the correlation between the local and global. The BERT-fused model is better than transformer, but it suffers from the mismatch between pretraining and fine-tuning. SDT-RPR outperforms all other models except TF-RPE because it takes advantage of the fact that higher layers share more global information at different locations by stacking more shallow models and thus trains deeper models with better translation results. However, the problem of possible redundancy caused by parameters is ignored, so there are some limitations to the NMT task.

**Table 3** BLEU scores of models on English(En)-Chinese(Zh) datasets

| Model | Low(120K) | | High(9.5M) | |
|---|---|---|---|---|
| | Zh–En | En–Zh | Zh–En | En–Zh |
| w/o FGM+RPE | 17.26 | 20.73 | 23.06 | 29.57 |
| w/o FGM | 18.87 | 21.48 | 24.89 | 31.16 |
| w/o RPE | 18.99 | 21.64 | 25.51 | 32.97 |
| w/o TF-RPE | 19.59 | 22.38 | 26.96 | 34.72 |

## 4.4 Ablation study

The ablation experiments were conducted to further study how the model can benefit from each component, where "w/o" indicates the removal of a component.

Table 3 demonstrates the effectiveness of these components for deep learning-based neural machine translation tasks.

- Relative positional embedding and adversarial training are fully used by the transformer model to provide accurate position and semantic information of words during translation. Without these, the model directly translates using the training information after word embedding and shows the worst performance in the machine translation task.
- Relative positional embedding can specifically encode the position information of each word in a sentence, which solves the problem of the inability to identify the sequence order of text in the word embedding process, and achieves good results in machine translation tasks, especially on the high-resource dataset.
- Adversarial training can help the model to better capture the correlation between words in a sentence, allowing it to focus more on identifying the semantic and contextual information of sentences. Compared to the TF-RPE model, on the low-resource dataset, the BLEU score decreases by 0.72 on the Zh→En task, and by 0.90 on the En→Zh task, with corresponding decreases of 2.07 and 3.56 on the high-resource dataset. This suggests that adding FGM to multi-head attention can indeed improve the translation performance of the model.

## 4.5 Case analysis

To better demonstrate the results of our model on the neural machine translation task, we randomly selected an example from the test set of the Zh→En task, as shown in Table 4. The Chinese sentence "教育加强了社会和文化资本, 这有助于保持强大稳定的政体," meaning "Education strengthens social and cultural capital, which contributes to strong and stable polities," is shown with its corresponding English

**Table 4** Examples of model on Zh-En task

| | |
|---|---|
| Input | 教育加强了社会和文化资本，这有助于保持强大稳定的政体。 |
| Reference | Education strengthens social and cultural capital, **which contributes to** *strong and stable* polities. |
| Baseline | Education fortifies social and cultural capital, **it aims polity to** maintain *stable* polity. |
| TF-RPE | Education fortifies social and cultural capital, **which helps maintain** *strong and stable* polities. |

translated source sentence, baseline translation, and translation result under our model. In the blue part, "强大稳定"(strong and stable) in the input Chinese sentence is translated as "strong and stable" in the source sentence of the corresponding English corpus. However, it is translated as "stable" in the baseline, translating "稳定"(stable) and missing "强大"(strong). Under the TF-RPE model translation, it is retranslated to "strong and stable," which reflects the adequacy of our model. In the red part, "这有助于保持"(which contributes to) is correctly translated in the source phrase of the corresponding English corpus. However, in the baseline, it is translated as "it aims polity to." Under the TF-RPE model translation, it is retranslated to "which helps maintain," which reflects the fluency problem of our model. It can be seen that incorporating relative positional embedding and adversarial training methods into neural machine translation tasks can produce more fluent translations with greater integrity.

## 5 Conclusion

In this study, we proposed a TF-RPE model based on relative positional embedding with adversarial training for a parallel English-Chinese corpus under a low- or high-resource dataset to process neural machine translation, resulting in a more accurate translation. We used a word embedding mechanism based on relative positional encoding to improve the ability to recognize word position information. A fast gradient algorithm and multi-head self-attentive mechanism were combined to enable the model to capture the relevance between words in the characterization of sentence. Accordingly, we demonstrated the effectiveness of the TF-RPE method on both low- and high-resource datasets. The model has the limitation that it ignores possible over-parameterization during training. In the future, we may further improve the performance of neural machine translation models by investigating the use of redundant parameters. In addition, we may generalize the approach to multilingual translation problems and verify the effectiveness of the method injection on complex languages.

**Author Contributions** All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by YL, YS and ZL. The first draft of the manuscript was written by YL and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

**Funding** This work is funded by the National Natural Science Foundation of China (No. 62076045) and the High-Level Talent Innovation Support Program (Young Science and Technology Star) of Dalian (No. 2021RQ066).

**Data availability** Not applicable.

## Declarations

**Conflict of interest** The authors declare no conflicts of interest regarding to publish this paper.

**Ethical approval** This article does not contain any studies with human participants or animals performed by any of the authors.

## References

Abdulmumin I, Galadanci BS, Ahmad IS, Abdullahi RI (2021) Data selection as an alternative to quality estimation in self-learning for low resource neural machine translation. In: International conference on computational science and its applications, pp 311–326. Springer

Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473

Chen K, Wang R, Utiyama M, Sumita E (2019) Neural machine translation with reordering embeddings. In: Proceedings of the 57th annual meeting of the association for computational linguistics, pp 1787–1799

Chiang Hsiu-Sen, Chen Mu-Yen, Huang Yu-Jhih (2019) Wavelet-based eeg processing for epilepsy detection using fuzzy entropy and associative petri net. IEEE Access 7:103255–103262

Cho K, Van Merriënboer B, Bahdanau D, Bengio Y (2014a). On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259

Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014b) Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078

Gehring J, Auli M, Grangier D, Yarats D, Dauphin YN (2017) Convolutional sequence to sequence learning. In: International conference on machine learning, pp 1243–1252. PMLR

Goodfellow IJ, Shlens J, Szegedy C (2015) Explaining and harnessing adversarial examples. In: International conference on learning representations 2015

Hinton EG, Osindero S, Teh Y-W (2006) A fast learning algorithm for deep belief nets. Neural Comput 18(7):1527–1554

Junczys-Dowmunt M, Dwojak T, Hoang H (2014) Is neural machine translation ready for deployment. In: The 13th international conference on spoken language translation

Kalchbrenner N, Blunsom P (2013) Recurrent continuous translation models. In: Proceedings of the 2013 conference on empirical methods in natural language processing, pp 1700–1709

Li B, Wang Z, Liu H, Du Q, Xiao T, Zhang C, Zhu J (2020a) Learning light-weight translation models from deep transformer. arXiv preprint arXiv:2012.13866,

Li B, Wang Z, Liu H, Jiang Y, Du Q, Xiao T, Wang H, Zhu J (2020b) Shallow-to-deep training for neural machine translation. arXiv preprint arXiv:2010.03737,

Liao B, Khadivi S, Hewavitharana S (2021) Back-translation for large-scale multilingual machine translation. arXiv preprint arXiv:2109.08712,

López-Gonzále , Meda-Campaña JA, Hernández-Martínez EG, Paniagua-Contro P (2020) Multi robot distance based formation using parallel genetic algorithm. Soft Comput 86:105929

Meng F, Zhang J (2019) Dtmt: A novel deep transition architecture for neural machine translation. In Proc AAAI Conf Artif Intell 33:224–231

Miyato T, Dai AM , Goodfellow I (2016) Adversarial training methods for semi-supervised text classification. arXiv preprint arXiv:1605.07725,

Mújica-Vargas D (2021) Superpixels extraction by an intuitionistic fuzzy clustering algorithm. Res Technol 19(2):140–152

Papineni K, Roukos S, Ward T, Zhu W-J (2002) Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the association for computational linguistics, pp 311–318

de Jesús Rubio J (2021) Stability analysis of the modified levenberg-marquardt algorithm for the artificial neural network training. IEEE Trans Neural Netw Learn Syst 32(8):3510–3524

Rubio J, Lughofer E, Pieper J, Cruz P, Martinez DI, Ochoa G, Islas MA, Enrique G (2021) Adapting h-infinity controller for the desired reference tracking of the sphere position in the Maglev process. Inf Sci 569:669–686

Rubio J , Islas MA, Ochoa G, Cruz DR, García E, Pacheco J (2022) Convergent newton method and neural network for the electric energy usage prediction. Inf Sci 585:89–112

Shaw P, Uszkoreit J, Vaswani A (2018) Self-attention with relative position representations. arXiv preprint arXiv:1803.02155,

Shi Y, Wang Y, Wu C,Yeh C-F, Chan J, Zhang F, Le D, Seltzer M (2021). Emformer: efficient memory transformer based acoustic model for low latency streaming speech recognition. In: ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 6783–6787. IEEE

So D, Le Q, Liang C (2019) The evolved transformer. In: International conference on machine learning, pp 5877–5886. PMLR

Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. Adv Neural Inf Process Systems, 27

Tu Z, Lu Z, Liu Y, Liu X, Li H (2016) Modeling coverage for neural machine translation. In: the 54th annual meeting of the association for computational linguistics, pp 76–85

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst, 30

Wilks Y (1993) Corpora and machine translation. In: Proceedings of machine translation summit IV, pp 137–146

Wu F, Fan A, Baevski A, Dauphin YN, Auli M (2019). Pay less attention with lightweight and dynamic convolutions. arXiv preprint arXiv:1901.10430,

Zhu J, Xia Y, Wu L, He D, Qin T, Zhou W, Li H, Liu TY (2020) Incorporating bert into neural machine translation. arXiv preprint arXiv:2002.06823,

Ziemski M, Junczys-Dowmunt M, Pouliquen B (2016) The united nations parallel corpus v1. 0. In: Proceedings of the tenth international conference on language resources and evaluation (LREC'16), pp 3530–3534