applied sciences

MDPI

Article

X-Transformer: A Machine Translation Model Enhanced by the Self-Attention Mechanism

Huey-Ing Liu * and Wei-Lin Chen

Electrical Engineering, Fu Jen Catholic University, No. 510 Zhongzheng Rd., Xinzhuang Dist., New Taipei City 242062, Taiwan; 408216132@gapp.fju.edu.tw
* Correspondence: hiliu@mail.fju.edu.tw

Abstract: Machine translation has received significant attention in the field of natural language processing not only because of its challenges but also due to the translation needs that arise in the daily life of modern people. In this study, we design a new machine translation model named X-Transformer, which refines the original Transformer model regarding three aspects. First, the model parameter of the encoder is compressed. Second, the encoder structure is modified by adopting two layers of the self-attention mechanism consecutively and reducing the point-wise feed forward layer to help the model understand the semantic structure of sentences precisely. Third, we streamline the decoder model size, while maintaining the accuracy. Through experiments, we demonstrate that having a large number of decoder layers not only affects the performance of the translation model but also increases the inference time. The X-Transformer reaches the state-of-the-art result of 46.63 and 55.63 points in the BiLingual Evaluation Understudy (BLEU) metric of the World Machine Translation (WMT), from 2014, using the English–German and English–French translation corpora, thus outperforming the Transformer model with 19 and 18 BLEU points, respectively. The X-Transformer significantly reduces the training time to only 1/3 times that of the Transformer. In addition, the heat maps of the X-Transformer reach token-level precision (i.e., token-to-token attention), while the Transformer model remains at the sentence level (i.e., token-to-sentence attention).

Keywords: machine translation; natural language processing

Citation: Liu, H.-I.; Chen, W.-L. X-Transformer: A Machine Translation Model Enhanced by the Self-Attention Mechanism. *Appl. Sci.* **2022**, *12*, 4502. https://doi.org/10.3390/app12094502

Academic Editor: Valentino Santucci

Received: 11 March 2022
Accepted: 26 April 2022
Published: 29 April 2022

1. Introduction

In the context of the development of several innovative neural network models used in machine translation, the Recurrent Neural Network (RNN) has become a classic model used for neural machine translation (NMT) [1]. However, RNN still faces some obstacles in the training process. First, because the sequential input of RNN is adopted, early information is easily lost, especially when there is a large input. Second, the basic RNN uses back propagation, and back-propagated gradients are usually vanished or exploded because of the use of a finite-precision number in computing. With the development of Long Short-Term Memory (LSTM) [2,3], the gradient-vanishing problem has been reassessed. In a different manner to the standard feed forward network, LSTM has its own feedback connections: input gate, forget gate, and output gate. However, unlike the Convolutional Neural Network (CNN), both RNN and LSTM are time-sequence models that still suffer from the parallel computing problem—i.e., the GPU cannot be used to speed up while training.

A state-of-the-art model called the Transformer [4] proposed by Google applies the Multi-Head Self-Attention Mechanism and leverages the parallel computing in Neural Machine Translation. The Transformer has been the best NMT model to be proposed to date. In recent years, several researchers have used this self-attention mechanism to improve their models. Using GPU computing, the self-attention mechanism has improved the NMT model to a certain extent, and more researchers have chosen to stack more GPUs or model parameters, such as BERT. Although this parameter improves performance, its

higher training cost also implies that these pre-trained models can only be owned by specific researchers. Therefore, in this paper, we focus on adjusting the sub-layers of the Transformer model, whilst maintaining its accuracy and efficiency.

To prove that our model is better than the state-of-the-art model Transformer, in this paper we use two different language pairs: the English–German and English–French news commentary translation corpus of the World Machine Translation (2014) [5]. The proposed X-Transformer, which is an NMT model, includes the following features:

- The X-Transformer modifies the structure of the Transformer by reducing the number of model parameters in both the encoder and decoder.
- The basic building block of the encoder in the X-Transformer is refined. We believe that applying the self-attention mechanism in the feed forward layer of the encoder is helpful for the model to understand natural languages, as will be shown in the experimental results.
- Two different language pairs are applied to train the Transformer and X-Transformer at the same time. The experiment results reveal that the proposed X-Transformer has an outstanding performance.
- The visual heat map is used to aid the comprehension ability of the trained models. The results again demonstrate the superiority of the X-Transformer. The refined encoder structure of the X-Transformer elevates the comprehension ability to a token-to-token level.

2. Review of the Literature

The training process of an NMT model includes data collection, data cleaning, data preprocessing, model training, and model evaluation. In this section, we review the methods used in previous research. First, we discuss the development and extension of translation models, then the construction of translation datasets, and finally, the natural language processing tasks in the medical domain.

2.1. Machine Translation

The NMT model, which mainly consists of an encoder and a decoder, is the best-performing translation model recognized by everyone to date. An encoder maps an input sequence to a hidden representation, including tokenization. A decoder maps the hidden representation to a target sequence, including un-tokenization. When an NMT model tackles a longer sequence, in the manner of a human, it is affected by more information, which causes a lower learning efficiency. The attention mechanism can be used to align the model to focus on the correct tokens. It can not only tackle longer sequences but also allow the model to "attend" on the correct key point(s) in the sequence.

The work presented in [6] is the first research to incorporate the attention mechanism in the NMT model. Luong et al. proposed Global Attention, which performs attention calculations on all input sequences, whilst Local Attention only performs attention calculations on N contexts to reduce the burden of computing and thus improve the attention mechanism in the NMT model [7]. The NMT model proposed by [7], training on an English–German translation dataset, obtained significantly better results than classic NMT models. The encoder and decoder can be implemented by well-known deep learning models, such as LSTM [2,3], the Convolution seq2seq model [8], and the Transformer [4].

A disadvantage of the NMT model is that it is limited by words that are not in its vocabulary, so-called out-of-vocabulary (OOV) words. Luong et al. highlighted this problem and the model they presented copied the OOV word directly from the encoder to the decoder to form a translation result [9]. In addition, the authors of [10,11] also used a similar approach to solve the OOV problem. Certain medical terminology or names that are not included in the dictionary due to the insufficient frequency of appearances in the corpus will cause the NMT model to be unable to generate the specific words or names when generating the results. CopyNet, proposed by [10], applies two training modes: generation and copy modes. In addition, CopyNet has two dictionaries: one is a traditional dictionary,

and the other is a dictionary that contains the words that appear only once in the dataset. While training, CopyNet has to learn which mode to apply. If the probability of using the generation mode is higher, the word is generated from the traditional dictionary. Otherwise, the word is generated from the dictionary of words that have only appeared once in the dataset. The use of this method can alleviate the OOV problem, and it is, therefore, better in the case of preprocessing at the word level than in processing at the letter level.

Gulcehre et al. used human psychology to solve OOV problems [11]. In human psychology, when there is an unknown object in the environment, people often seek answers from the surrounding environment. Therefore, this study used the Multi-Layer Perceptron (MLP) model and two SoftMax layers to construct the NMT model. One SoftMax layer predicts where the word currently generated is located in the source sequence. Additionally, the other SoftMax layer predicts where the currently generated word is in the vocabulary list. In the training step, the MLP model determines which SoftMax layer is used. This method decides whether the original input word is used as output word as a result of model adaption. The adaption method can be also used in text summarization.

The authors of [12] used a direct method to solve the OOV problem. Jean et al. proposed an important sampling method. This enabled the NMT model to only consider a small and important part in the vocabulary list rather than the entire vocabulary list. The important sampling method combining the LSTM model training with the English–German translation dataset was the state of the art at the time of its production.

Finally, the work in [13] used Byte Pair Encoding (BPE) to generate sub-words, which is a useful method to effectively solve the OOV problem. It also considers compound words and cognate words, so the semantics of the sentence can be fully expressed. Additionally, the vocabulary list can also have a fixed size.

2.2. Transformer

The Transformer [4] has been the state-of-the-art NMT model since 2017. Through the clustering of GPU computing, it reduces the training time [14]. Due to the success of the Transformer, a lot of research has been devoted to modifying the Transformer into a better model [15–17]. The work in [15] analyzed the influence of the number of encoders and decoders in the Transformer. Shi et al. proposed a sentence alignment NMT model [17], which is divided into two stages. First, a discriminator is pre-trained. After the sentence is generated by the NMT model, the sentence calculates the difference to the ground truth using the pre-trained discriminator and considers the training loss. This method allows the model to learn the differences in comparison with the original sentence more efficiently, rather than just decide whether the two sentences are equal. In [16], the sub-layers of the encoder and decoder in the Transformer, such as the self-attention and feed forward layers, were reordered to reduce the model's perplexity and increase the model's robustness. In this study, we also aim to improve the efficiency of the refined model and shorten the training time.

2.3. Combination with the Pre-Trained Model

Zhu et al. showed how to combine BERT into the Transformer's encoder and decoder layers [18]. BERT is introduced to processes the input sequences of the encoder and decoder. The output of BERT is combined with the inputs of the encoder and decoder to create self-attention calculations, as shown in Figure 1. This proposed BERT combined with the NMT model outperforms the original Transformer.

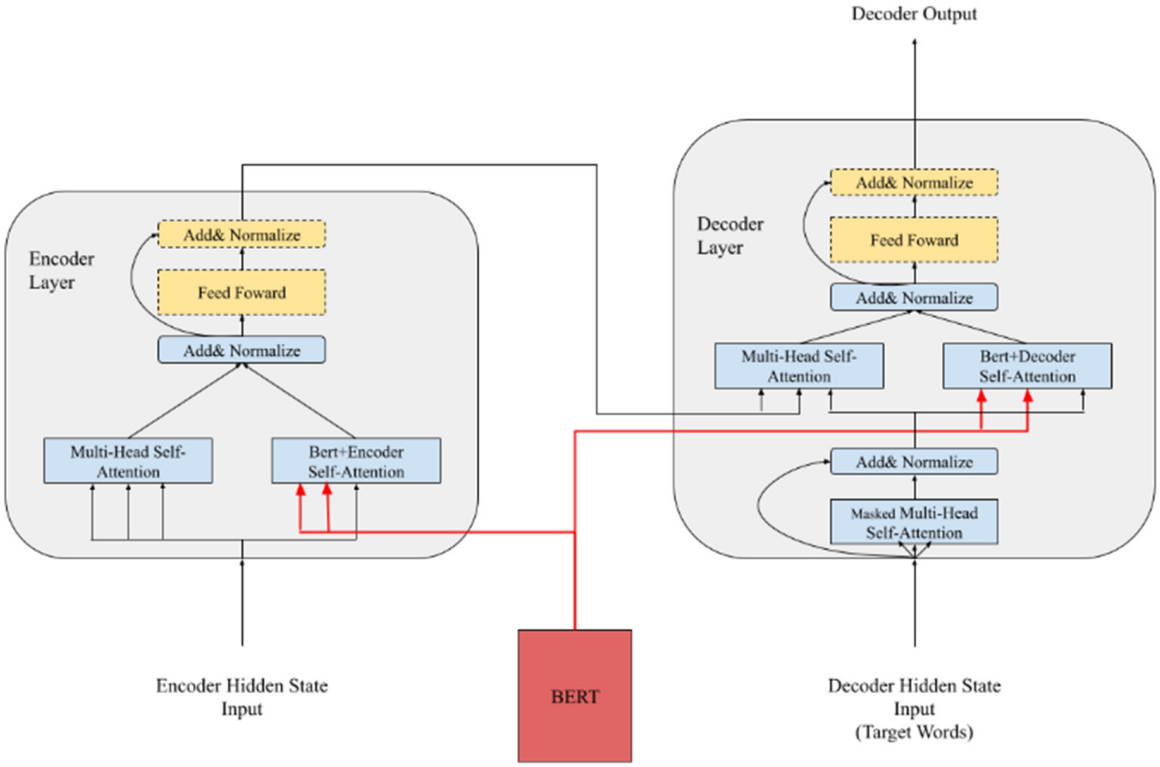

Figure 1. The structure of the model combining BERT with the transformer.

In [19], three methods were proposed to combine BERT and the Transformer—namely Embedding, Fine-tuning, and Freeze. Embedding only combines BERT with the embedding layer. Fine-tuning applies BERT as the encoder's initial parameter and trains it with the training set. Additionally, Freeze treats BERT as the Transformer's encoder, and thus only the decoder is trained. According to the results, both Embedding and Fine-tuning showed improvements in the model, but Freeze led to a decline in the performance of the Transformer. From the above results, we conclude that BERT only helps the encoder to understand the training set. However, if BERT is allowed to become the encoder without any fine-tuning, the results will be worse than those obtained with the Transformer because the pre-training corpus of BERT may not achieve the focus of the training set. In their study, Clinchant et al. also added noise in the training corpus. The Transformer combined with BERT can identify when the training data are mixed with redundant symbols or misspellings. The performance of the Transformer combined with BERT when using both the Embedding and Fine-tuning methods is better than that of the original Transformer.

2.4. Model Parameter Compression

Several studies have attempted to reduce the model parameters, such as [20,21]. In [20], two models are proposed: the teacher model and student model. The student model uses sentence-level knowledge extracted from the teacher model. Through experiments, it is shown that this method not only improves the training speed of the student model but also has a small loss in performance. The parameters of the student model after knowledge extraction are only 1/13 times that of the teacher model, causing the parameters of the NMT model to be compressed even further.

Sun et al. explained that, although BERT is helpful in NLP tasks, its high requirements in terms of computing resources often discourage people from using it [21]. Therefore, they also proposed the teacher and student model. The teacher model is called BERT, and the student model has two ways to learn from the teacher model. One is called PKD-Last, in which the student model only learns the last K layers from the teacher model. The other is called PKD-Skip, in which the student model learns every K layer from the teacher

model. These two solutions allow the student model to patiently learn and imitate the teacher model. The experiments show that the training efficiency can be improved without sacrificing the accuracy in multiple NLP tasks. Model parameter compression is also one of the key points explored in our study.

3. The Proposed Model

3.1. Multi-Head Self-Attention Mechanism

Scale dot product attention is the most important self-attention mechanism that was studied in this paper. As shown in Figure 2, it involves three main parameters: Q (Query), K (Key), and V (Value). Each of the three parameters performs linear conversion and is represented by a vector matrix. After calculating the dot product of Q and K, we can attain the attention weight of the input sequence, which is, then, multiplied by V to obtain the final attention context vector.

$$\text{Attention}(\text{Q},\ \text{K},\ \text{V}) = \text{softmax}\left(\frac{\text{QK}^{\text{T}}}{\sqrt{d_k}}\right)\text{V} \tag{1}$$

As shown in Equation (1), the transposed matrices of K and Q are multiplied to calculate the similarity between each other, and the obtained result is the attention weight matrix of the current input sequence. The importance of the sequence information can be learned through this method. Then, this result is divided by sqrt (d_k) to prevent the result of SoftMax from being only 1 and 0. Finally, it is multiplied by V to obtain the corpus attention matrix of the last calculated sequence value.

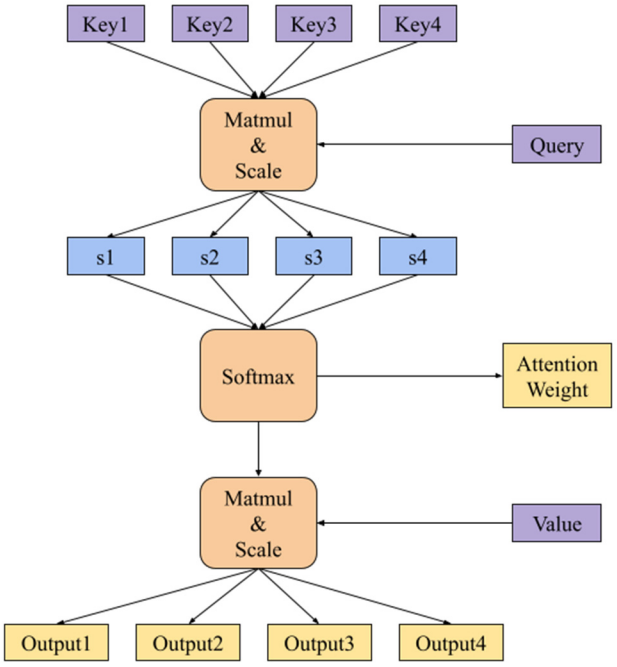

Figure 2. Scale dot product attention.

The multi-head self-attention mechanism divides the dimension of the sub-word embedding equally and then calculates it by the scale dot product attention. Considering that the Head number H = 8 and the model dimension d_{model} = 512 are the hyperparameters of the X-Transformer model set in this study, there is a 64-dimensional vector (512/8 = 64) in each head. The eight heads are, then, combined after the attention matrix is obtained. The reason behind separating the model dimension to calculate the self-attention mechanism is that the model in the multi-head mode can learn more easily the key point on which it must be focused. The eight heads can bring attention to at least eight tokens to let the

training of the NMT model be performed more precisely [4]. Both the encoder and decoder of the X-Transformer use double multi-head self-attention mechanisms.

In the encoder, the parameters Q, K, and V of the first self-attention mechanism were used as inputs of the encoder after the sub-word embedding process. The residual result of the first self-attention mechanism and the input of the first self-attention mechanism were added and normalized, which were the parameters Q, K, and V of the second self-attention mechanism.

In the decoder, the parameters Q, K, and V of the first self-attention mechanism were used as the input of the decoder after the sub-word embedding process and the correct answer (i.e., ground truth). More specifically, in order to prevent the NMT model from knowing the correct answer of the future training sequence in advance, an additional look-ahead mask was added.

The look-ahead mask was implemented by multiplying a negative infinity value that is, assuming that the input sequence of the decoder is d, $d+1$, $d+2$, $\ldots$, $d+n$. Additionally, in training step t, except for input sequence d, the other sequence $d+1$, $d+2$, $\ldots$, $d+n$ was multiplied by the negative infinity value. Then, in training step $t+1$, except for input sequence d, $d+1$, the other sequence $d+2$, $\ldots$, $d+n$ was multiplied by the negative infinity value, too. This took place until training step $t+n$, which was the last training step, when the look-ahead mask ended.

The second self-attention mechanism in the decoder calculated the similarity between the output of the first self-attention mechanism in the decoder and the output of the encoder. The parameter Q was added up and normalized the residual result of the first self-attention mechanism and the input of the first self-attention mechanism in the decoder. Additionally, the parameters K and V were the output of encoder.

3.2. The Lazy Layer Question

In [15], a variety of different encoder and decoder layer combinations were studied to test the impact on the translation. The experimental results revealed that the encoder is more effective and important than the decoder. Increasing or decreasing the number of layers of the decoder of the Transformer does not significantly affect the model performance. Based on this previous work, we determined the required number of layers of encoders and decoders in the X-Transformer. We experimented with various encoder and decoder combinations for comparison and used the original Transformer model as a baseline for benchmark, as shown in Table 1.

Table 1. Comparison of the number of layers of encoder and decoder in the Transformer model.

Encoder Layers	Decoder Layers	BLEU	Training Time (under the Same Env.)
6	6	29.33	1×
6	4	29.26	0.82×
6	2	30.99	0.68×
8	6	29.26	1.19×
8	4	29.82	1.01×

From Table 1, we discover that, in the case of the same number of encoder layers, reducing the number of decoder layers can improve the translation performance of the model and reduce the training time. We believe the reason for the improvement in performance is that the encoder of the Transformer is a very important part to understand the input corpus. After the encoder has a considerable understanding of the original corpus, then, the target language sentence is generated by the decoder. Since the decoder is used for generation, the self-attention mechanism calculation is employed by mixing the original input and the partially generated results to predict the next output step. The presence of a large number of encoder layers seems to force the model to refocus on some tokens, which are not important or had already been attended before. Therefore, the translation accuracy cannot be improved.

Humanoid learning translation works as, for example, asking students who have a considerable knowledge of a language to translate sentences to the target language through what they learned. If students are asked to attempt to understand and translate what has been written in the target language, the improvement may be obvious in the beginning. However, students have to give too much attention when they are repeatedly asked to focus on understanding and translating the target language. In these, the learning outcomes are not met. On the contrary, they mostly stop paying attention and slow down the learning pace. This phenomenon is similar to the phenomenon called "learning plateau" in educational psychology [22]. As shown in Figure 3, as the learning time increases, there is a period of low efficiency when a plateau is reached. At this time, the manner to continue the learning process must be changed. Returning to the translation model, a decoder with fewer layers is recommended. However, if the decoder is reduced to one layer, the generated results will be unstable. Finally, we decided to use a double-layer decoder for the X-Transformer.

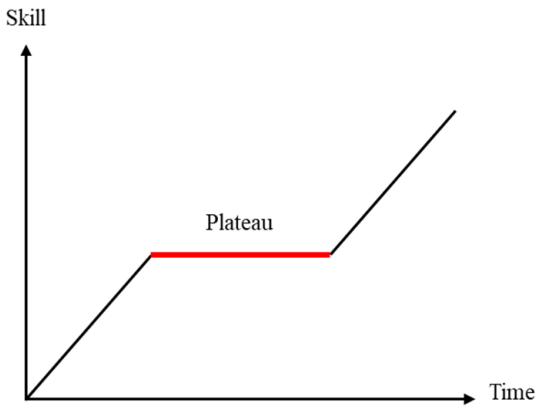

Figure 3. Learning plateau.

3.3. The Modified Encoder of the X-Transformer

In the original Transformer [4], as shown in Figure 4a, each encoder is a self-attention mechanism with a point-wise feed forward layer, which is added for residual calculation and normalization. Additionally, the point-wise feed forward layer is mainly a fully connected layer with a large number of neurons through the middle layer, and it is responsible for retrieving important information in the corpus, so that the neurons can learn actively.

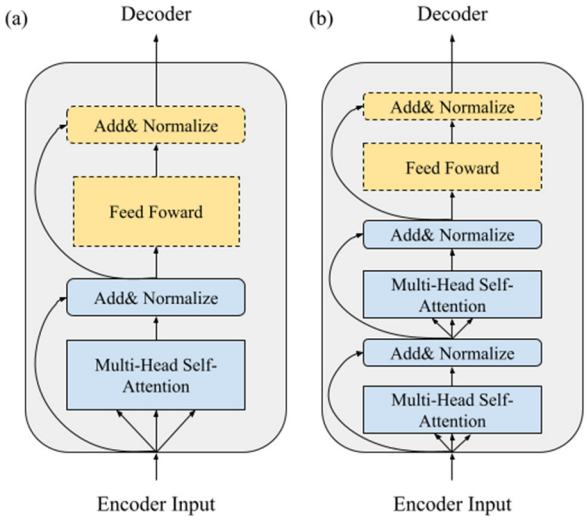

Figure 4. (**a**) Encoder layer of the transformer; (**b**) modified encoder layer of the X-Transformer.

However, we found that if the feed forward layer is added for non-linear conversion at the early training stage, the NMT model is not able to capture the true hidden representation in the corpus. The reason for this is that the feed forward layer contains a non-linear transformation, Relu, which directly converts all negative values to 0. The original intention is to help the model to converge, but it may cause adverse effects in the early training step. Therefore, we experimented with several different combinations of self-attention mechanisms and feed forward layers in the encoder, with the aim of clarifying which combination does not affect the NMT model while training and that still can help to converge the model.

Table 2 shows the impact of different combinations of self-attention mechanisms and feed forward layers in the model. In the table, the "s" represents a self-attention mechanism, and the "f" represents a feed forward layer, and the arrangement of letters represents the order of arrangement of the layers. Therefore, "sf-sf-sf-sf-sf-sf" is the baseline model of the original Transformer.

Table 2. BLEU scores on different arrangements of self-attention mechanisms and feed forward layers in the encoder.

Modified Encoder	BLEU
sf-sf-sf-sf-sf-sf (Baseline)	29.33
s-sf-s-sf-s-sf	32.14
s-sf-sf (X-Transformer)	46.63
s-sf-sf-sf	46.72
s-sf-s-sf	46.62

In this specific translation dataset, it was found that three self-attention mechanisms are sufficient. It seems that the learning plateau phenomenon occurs again in the encoder. Finally, we chose s-sf-sf as our proposed model because it has fewer model parameters, while the BLEU score is almost the same. A similar notion was also proposed in [16]. If the self-attention mechanism appears earlier in the model and stacks more than the number of feed forward layers, the perplexity of the language model can be much lower.

3.4. X-Transformer

In order to break through the traditional NMT model training strategy, we chose to train a model with a smaller parameter, called the X-Transformer. This NMT model was retrained according to different usage scenarios in order to prevent issues similar to those of the traditional pre-trained model with larger model parameters, such as BERT. Some of these issues are that, when it encounters a non-trained corpus, the performance struggles, and the inference time is quite long.

The X-Transformer model was constructed using the aforementioned methods, as shown in Figure 5, and has the following advantages:

- Sub-word tokenization: Sub-words can reduce the sparse words (such as terminology) of the model, improve the relevance of each vocabulary through sub-words, and reduce the dictionary size.
- Modifies the basic building block of the encoder: It can rearrange the self-attention mechanism and feed forward layer and modify the ratio number to improve the accuracy of the attention given to the input sequence.
- Reduces the number of encoder and decoder layers: It can reduce an excessive self-attention mechanism in both the encoder and decoder to ease the learning plateau in the model and also increase inference speed. In addition, reducing the decoder layers also increases the influence of the encoder output and helps to obtain a better output.

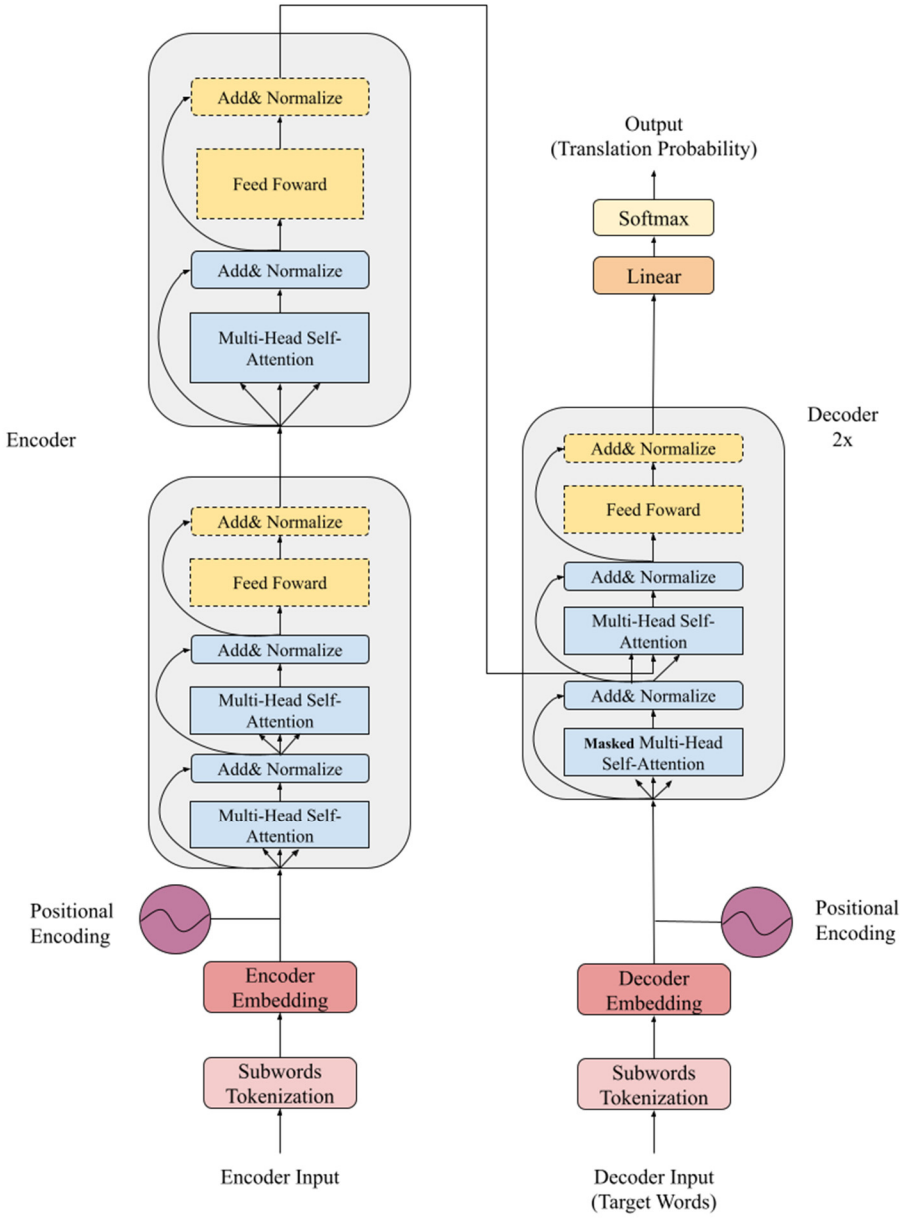

Figure 5. Construction of the X-Transformer model.

4. Experimental Results and Discussion

We conducted experiments on the WMT 2014 English–German and English–French corpora. The World Machine Translation dataset contains data collected from news translation tasks launched by the World Machine Translation Conference [5], and it has been continuously updated until last year. In addition to the news corpus, public information from the United Nations is also included. The WMT 2014 English–German (EG) and English–French (EF) corpus was used in the original Transformer paper [4], and was also used as the basis for the evaluation of the model performance in this paper. The BLEU score [23] was used as the evaluation standard.

At the same time, we also trained the BERT-fused NMT model. BERT was pre-trained using a considerable number of corpora, including BooksCorpus [24] and the English version of Wikipedia. The contents in the BooksCorpus are divided into 16 categories, and the majority of entries belong to the romance, fantasy, science fiction, and teen categories. The contents of English Wikipedia include culture, art, biographies, geography and so on [25], as shown in Figure 6.

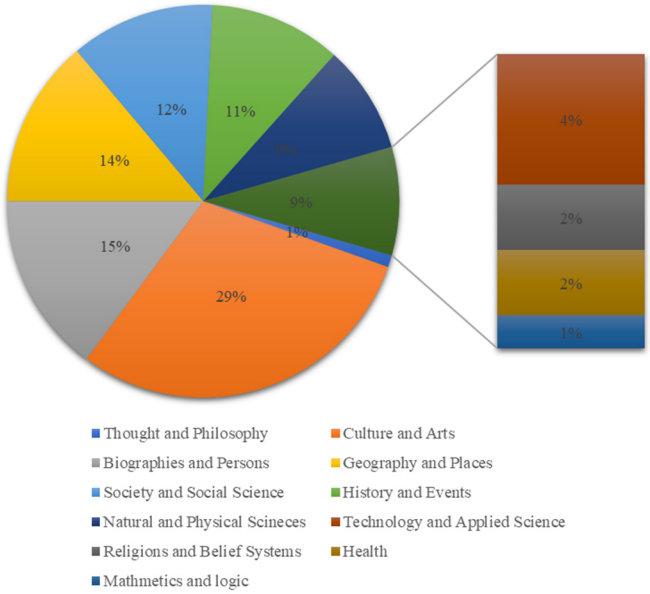

Figure 6. The English Wikipedia content by subject [25].

The experimental results are separated into two aspects:

- First, the comparison of the performance between the X-Transformer and related models. In addition to comparing our model with the baseline Transformer, we also compared it with the results of adding the pre-trained model BERT, so that readers have a clearer understanding the performance of the X-Transformer.
- Second, the learning situation of the self-attention mechanism in the encoder is presented, which is displayed by a visual heat map. Through the visualized heat maps of the Transformer and X-Transformer, we can understand whether the mechanism pays attention to the correct tokens and the attention relationship between tokens. These also reveal the comprehension of the models.

The computer specification used in the experiments of this paper was Intel core i7-9700, with 32 GB RAM, and the GPU was Nvidia Titan RTX. The experimental operating system was Ubuntu 18.04 LTS, and the software used was CUDA 10.1, cuDNN v7.6.5.

4.1. The Hyperparameter

All sentences were segmented into sub-word types before training. The segmentation algorithm was the subwordtextencoder in tf.keras. The results of WMT 2014 corpus were obtained from the test set called newstest 2014. There were 3000 sentences in both test sets that did not enter the NMT model for training. BLEU was used as the quality assessment of machine translation after the translated sentence was generated.

Adam [26] was the used optimizer during training. The learning rate was adjusted according to Equation (2). The other hyperparameters are listed in Table 3.

$$\text{lrate} = d_{model}^{-0.5} \cdot \min\left(step\_num \cdot warmup\_steps^{-1.5}\right) \quad (2)$$

Table 3. List of hyperparameters.

Parameters	
Batch size	128
Token Length	80
Token Per Batch	10,240
Embedding Dimension	512
dff	2048

4.2. The Result of Newstest 2014 EN-DE and EN-FR

As shown in Table 4, the X-Transformer obtained the best BLEU scores of 46.63 and 55.62 in the WMT 2014 English–German (EN–DE) and English–French (EN–FR) corpora. The baseline Transformer [4] had BLEU scores of 27.3 and 38.1, which are about 19 and 17 points of BLEU score lower than our proposed model. The comparison with the BERT-fused NMT model revealed that, although the performance improved, it did not surpass that of the X-Transformer. As shown, the obtained BLEU scores of the BERT-fused NMT model are 30.75 and 43.78 for EN–DE and EN–FR, respectively, which are about 15 points in terms of BLEU score lower than the X-Transformer.

Table 4. BLEU scores on EN–DE and EN–FR newstest 2014 corpus according to model.

Models	BLEU (EN–DE)	BLEU (EN–FR)	Training Time (under the Same Env.)
Transformer [4]	27.3	38.1	1×
Transformer + Large Batch [14]	29.3	43.2	
BERT-fused Transformer [18]	30.75	43.78	2.42×
X-Transformer	46.63	55.63	0.35×

In Table 4, the training time experienced for each model per epoch in average under the same environment is also presented. For the convenience of observation, it is presented in multiples and the original Transformer is set as the baseline 1×. The training time of the X-Transformer is only 0.35 that of the original Transformer. In addition, the BERT-fused Transformer requires a longer training time, mainly because BERT is a pre-trained model with twelve layers of encoder and decoder stacked. Therefore, the BERT-fused Transformer has to wait for the output of BERT during each training process and takes a long time for training. The experiment results reveal that the training time of the BERT-fused Transformer is around 6.9 times higher than that of the proposed X-Transformer. The proposed X-Transformer not only significantly reduced the training time but also obtained a better BLEU score compared to that of the BERT-fused Transformer.

Analyzing the experimental results obtained from two different language pairs, we can conclude that the X-Transformer has a streamlined encoder and decoder that sped up the training speed and simultaneously improved accuracy.

4.3. The Visualization Heat Map

In order to understand whether the self-attention mechanism in the encoder produces adequate effects in the model, we applied the visualization heat map in the last self-attention mechanism in the encoder. A brighter color shows more attention, while a darker color represents less attention.

We randomly selected a sentence (“World currency standards have enormous inertia”) from the WMT 2014 English–German corpus to show the heat maps.

From the heat maps shown in Figures 7 and 8, we found that the self-attention mechanism in the original Transformer can only understand which token is important to the whole sentence, but X-Transformer upgrades to a higher precision level. The X-Transformer can understand the influence relation between tokens, which also demonstrates a better comprehension ability. These findings also support that the X-Transformer significantly outperforms the Transformer and BERT-fused Transformer.

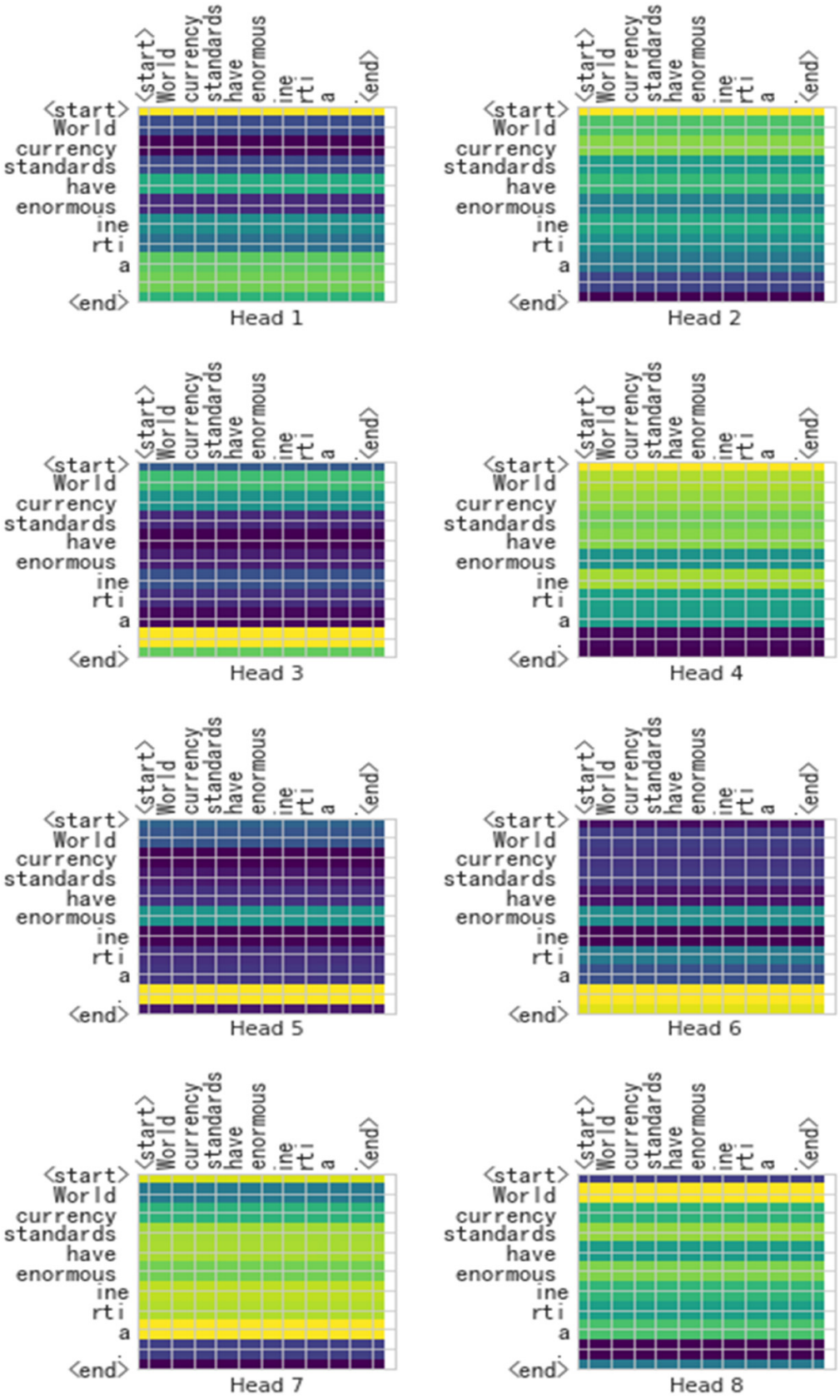

Figure 7. Heat maps of the sentence "World currency standards have enormous inertia" obtained with the Transformer.

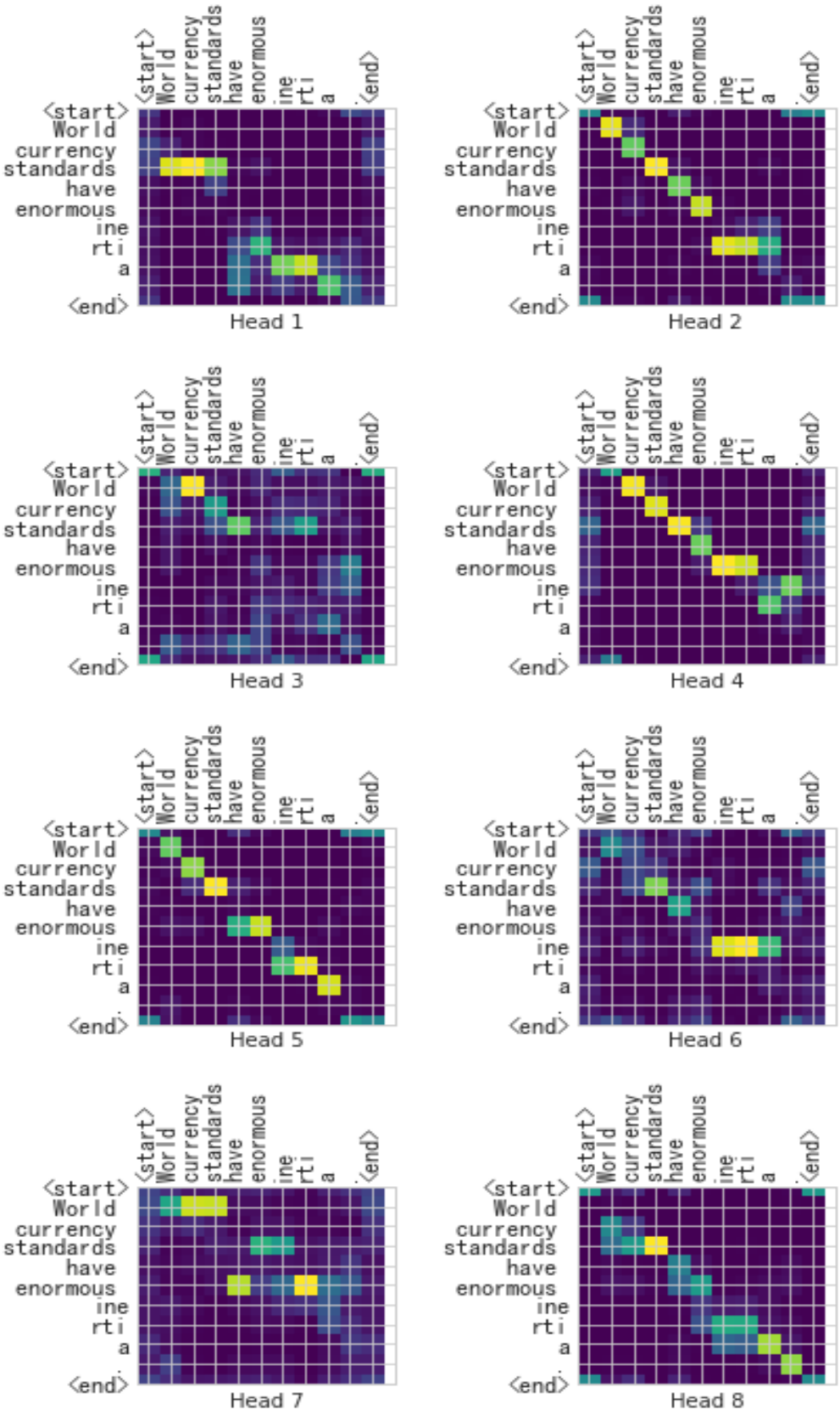

Figure 8. Heat maps of the sentence "World currency standards have enormous inertia" obtained with the X-Transformer.

5. Conclusions and Future Work

In this paper, an NMT model X-Transformer was designed for translation, improving the original state-of-the-art Transformer. First, stacking more self-attention mechanisms before the feed forward layer allowed the model to significantly improve the understanding of the input sequence during training. Second, we reduced the number of useless encoder and decoder layers to improve the inference speed of the model. For the WMT 2014 EN–DE and EN–FR translation corpora, the X-Transformer improved the BLEU score by about 20 points.

Based on the above experimental results, the following considerations will be taken forward for our future work. First, the decoder will still be an important aspect to be discussed. Its main function is to generate the target sentence with the assistance of the understanding of the input sentence by the encoder, which is different from the functionality encoder. We will attempt to find a better algorithm to clarify the relationship between the input and output sequences and to generate an appropriate translation. Second, we will build a Chinese-based translation corpus of the medical field for the NMT system. In machine-learning-based NMT systems, the training corpus is one of the key elements to improve the quality of translations. The frequent international communication that is the result of business, education and travel drives the need for international medical care. The need for bilingual or even trilingual communication using medical field terminology is an urgent and essential issue to be solved. These will be the two directions of our future work to continue to improve machine-learning-based NMT systems.

Author Contributions: Conceptualization, H.-I.L. and W.-L.C.; methodology, H.-I.L. and W.-L.C.; software, W.-L.C.; validation, W.-L.C.; formal analysis, H.-I.L. and W.-L.C.; investigation, H.-I.L. and W.-L.C.; resources, H.-I.L.; data curation, H.-I.L. and W.-L.C.; writing—original draft preparation, H.-I.L. and W.-L.C.; writing—review and editing, H.-I.L.; visualization, H.-I.L. and W.-L.C.; supervision, H.-I.L.; project administration, H.-I.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The WMT-2014 machine translation dataset is publicly available and has been used by many researchers in the area of machine translation.

Conflicts of Interest: The authors declare no conflict of interest.

References and Note

1. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536. [CrossRef]
2. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
3. Wu, Y.; Schuster, M.; Chen, Z.; Le, Q.V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv* **2016**, arXiv:1609.08144.
4. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is All You Need. In Proceedings of the 2017 Conference on Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
5. Bojar, O.; Buck, C.; Federmann, C.; Haddow, B.; Koehn, P.; Leveling, J.; Monz, C.; Pecina, P.; Post, M.; Saint-Amand, H.; et al. Findings of the 2014 Workshop on Statistical Machine Translation. In Proceedings of the Ninth Workshop on Statistical Machine Translation, Baltimore, MD, USA, 30 May 2014; pp. 12–58.
6. Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv* **2014**, arXiv:1409.0473.
7. Luong, M.-T.; Pham, H.; Manning, C.D. Effective Approaches to Attention-Based Neural Machine Translation. *arXiv* **2015**, arXiv:1508.04025.
8. Gehring, J.; Auli, M.; Grangier, D.; Yarats, D.; Dauphin, Y.N. Convolutional Sequence to Sequence Learning. In Proceedings of the 2017 International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 1243–1252.
9. Luong, M.-T.; Sutskever, I.; Le, Q.V.; Vinyals, O.; Zaremba, W. Addressing the Rare Word Problem in Neural Machine Translation. *arXiv* **2014**, arXiv:1410.8206.
10. Gu, J.; Lu, Z.; Li, H.; Li, V.O. Incorporating Copying Mechanism in Sequence-to-Sequence Learning. *arXiv* **2016**, arXiv:1603.06393.
11. Gulcehre, C.; Ahn, S.; Nallapati, R.; Zhou, B.; Bengio, Y. Pointing the Unknown Words. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016.

12. Jean, S.; Cho, K.; Memisevic, R.; Bengio, Y. On Using Very Large Target Vocabulary for Neural Machine Translation. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, 26–31 July 2014; Volume 1, pp. 1–10.
13. Sennrich, R.; Haddow, B.; Birch, A. Neural Machine Translation of Rare Words with Subword Units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, 7–12 August 2016; pp. 1715–1725.
14. Ott, M.; Edunov, S.; Grangier, D.; Auli, M. Scaling Neural Machine Translation. *arXiv* **2018**, arXiv:1806.00187.
15. Xu, H.; van Genabith, J.; Liu, Q.; Xiong, D. Probing Word Translations in the Transformer and Trading Decoder for Encoder Layers. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Mexico City, Mexico, 6–11 June 2021; pp. 74–85.
16. Press, N.; Smith, A.; Levy, O. Improving Transformer Models by Reordering their Sublayers. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Seattle, WA, USA, 9 December 2019; pp. 2996–3005.
17. Shi, X.; Huang, H.-Y.; Wang, W.; Jian, P.; Tang, Y.-K. Improving Neural Machine Translation by Achieving Knowledge Transfer with Sentence Alignment Learning. In Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), Hong Kong, China, 11 September 2019; pp. 260–270.
18. Zhu, J.; Xia, Y.; Wu, L.; He, D.; Qin, T.; Zhou, W.; Li, H.; Liu, T.Y. Incorporating Bert into Neural Machine Translation. In Proceedings of the 2019 Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2020.
19. Clinchant, S.; Jung, K.W.; Nikoulina, V. On the use of BERT for Neural Machine Translation. In Proceedings of the 3rd Workshop on Neural Generation and Translation, Hong Kong, China, 4 November 2019; pp. 108–117.
20. Kim, Y.; Rush, A.M. Sequence-Level Knowledge Distillation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 21 September 2016; pp. 1317–1327.
21. Sun, S.; Cheng, Y.; Gan, Z.; Liu, J. Patient Knowledge Distillation for Bert Model Compression. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, 3–7 November 2019; pp. 4323–4332.
22. Mirzaei, M.; Zoghi, M. Understanding the Language Learning Plateau: A Grounded-Theory Study. *Teach. Engl. Lang.* **2017**, *11*, 195–222.
23. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
24. Zhu, Y.; Kiros, R.; Zemel, R.; Salakhutdinov, R.; Urtasun, R.; Torralba, A.; Fidler, S. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In Proceedings of the 2015 IEEE International Conference on Computer Vision, Washington, DC, USA, 7–13 December 2015; pp. 19–27.
25. Häggström, M. Wikipedia content by subject.png. Wikimedia.org. Wikimedia Commons. Web. 2010.
26. Kingma, P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.