Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 189 (2021) 3–10

Procedia
Computer Science

www.elsevier.com/locate/procedia

5th International Conference on AI in Computational Linguistics

Re-Transformer: A Self-Attention Based Model for Machine Translation

Huey-Ing Liu[*], Wei-Lin Chen

[*]*Electrical Engineering, Fu Jen Catholic University, No. 510Chungcheng Rd., Hsinchuang Dist.,*
New Taipei City 24205, Taiwan

Abstract

Machine translation is one of the most popular and hardest tasks in Natural Language Processing. This paper proposes a self-attention based model for machine translation, named Re-Transformer, by transforming the Transformer [1]. Different from prevailing approach through module or GPU stacking to improve the system performance, Re-Transformer modifies the basic architecture; there are four refinements as follows. First, Re-Transformer adopts sub-word Tokenization in corpus preprocessing to overcome rare words. In the encoder layer, dual Self-Attention stacks and less Point Wise Feed Forward layer are adopted to obtain better comprehension of an input sentence. In decoder, reduced the stack of "Decoder" are used to speedup training and inferring speed meanwhile keeps the same level of BLEU. The experiment results show that Re-Transformer with BLEU metric score 31.36, 38.45 (four-layer decoder) and 32.14, 55.62 (two-layer decoder) improves around 4 and 17 points of BLEU metric against the Transformer over the WMT 2014 English-German and English-French Translation Corpus.

Keywords: Transformer; Machine learning; Machine translation; Self-attention

1. Introduction

Now a day, due to the massive convenient and fast transportation platforms, frequent communications among people in different countries are really common and drive a deeper need of translations for everyone anytime and anywhere. Machine translation seems the only answer to these requirements. Machine translation has always been

[*] Corresponding author. Tel.: +886-2-29053802; fax: +886-2-29042638.
E-mail address: hiliu@mail.fju.edu.tw

10.1016/j.procs.2021.05.065

one of the ultimate targets in natural language processing. Thanks to consistent progress in AI, machine translation steps into a whole new page. However, the classic neural machine translation based on deep learning implemented by RNN faces a big obstacle. That is while training RNN using back propagation, the back-propagated gradients are usually vanished or exploded, because of it uses finite-precision number in computing. Until the development of LSTM (Long-Short Term Memory) [2, 3], it deals with the vanishing gradient problem. Unlike standard feed forward neural network, LSTM has its own feedback connections, but it still suffers the parallel computing problem. It means LSTM cannot be fully speeded up by GPU.

In 2017, Google Brain proposed the multi-head self-attention mechanism [1] leverage the parallel computing in Neural Machine Translation. Most of researchers improved their model by self-attention at the moment. Although the flourishing of GPU computing allows language models and self-attention mechanism to improve to a certain extent, this situation has caused recent models to be developed in a stacking or larger direction. These model training by large computing resource can only be owned by certain specific groups. Therefore, we are committed to tuning the sub-layers in state-of-the-art model Transformer in this paper which can relatively increase the inference rate of the model. In this paper, we believe that the language model's layer should be modified with each unspecific language. With the different machine translation model, the parameter should be reduced or modified. For example, when the model is only used to communicated between two people with different native language, the model should be a specific and small one instead of a generalized or large one. In fact, it's not necessary to stack numerous Encoder and Decoder layers, with the English-German and English-French translation model proposed in the paper.

We prove through experiments that if the number of Encoder and Decoder layers along the models are changed, it will not only maintain or even increase the training and inference accuracy, but also save effects to the model. The key ideas of the proposed Re-Transformer are summarized as follows.

To sum up, our proposed approach has the following advantages:

- Inspired by [4], to reduce the number of parameters, which not only effectively improves the training speed, but also reduces the consuming of time while the model performing inference.
- Modify the amount of point wise feed forward layer combinations immediately after self-attention layer in the encoder layer. We believe that a higher ratio of Self-Attention layer to Feed Forward layer assists the model to understand natural languages.
- Last, we propose a multi-head self-attention visualization heat map. The heat map illustrates based on the attention weights of input sequence after self-attention mechanism processed. Through the heap map, we can better analyze the multi-head self-attention where the translation model pays attention to in the sequence and how it helps to translate the source sequence to the target sequence.

2. Related Works

Machine translation aims to translate an input sentence from source language to the target language, and the output is a sentence. A machine translation model is usually consisted of an encoder to map a input sequence to a hidden representation, including tokenization and a decoder to map the hidden representation to a target sequence including un-tokenization. Recently, translation models combine an attention mechanism to help the model to understand the hidden representation precisely. The attention mechanism was first introduced by [5] which had been used in source and target sequences' attention alignment. Encoder and Decoder can be implemented by LSTM [2, 3], Convolution seq2seq model [6] and Transformer [1]. Transformer achieves the state-of-the-art result in neural machine translation [7]. Transformer consists of several important sub-layers, like positional embedding, self-attention mechanism, and point wise feed forward layer. Positional embedding, due to a word position in a sentence is important, so it was proposed to take care the position data of a sentence token. Self-attention mechanism processes the sequential data and take the context timestep into consideration. And point wise feed forward layer is able to apply the linear and non-linear transformations for the model.

On the other hand, some recent researches reconstruct and modify the Transformer layers to get better performance [4, 8, 9]. Reference [4] analyzes how the number of Encoder and Decoder layers effect the performance of a Transformer based translation model. Shi et al. propose a sentence alignment mechanism combined with the encoder [8]. After the decoder generates the target sentence, the alignment mechanism will compare it with the ground truth and take the obtained accuracy to further refine the model. Reference [9] reorders the sub-layers of the encoder and decoder in Transformer, such as self-attention and feed-froward layers etc. to reduce the model perplexity and increase the model robustness.

With the development of pre-trained models, like ELMo, BERT and GPT have attracted more attentions in recent years. BERT is one of the widely used well pre-trained model. References [10] shows how to combine BERT into Transformer's encoder and decoder layers. Three methods of reusing BERT: combining BERT into embedding, initializing the encoder with BERT parameter, and using BERT as the Transformer encoder, are proposed in [11]. In the leverage of BERT, it makes the NMT model more robust.

Some of the researches in convolution neural network observe that increasing the depth of the model can improve the model performance [12]. The machine translation researchers also follow the trend to increase depth of the language model. For example, reference [13] shortens the path of the back propagation from depth layers to shallow layers. Wu et al. propose a two-stage model to build an 8-layer Transformer-Big system [14].

Instead of increasing the depth of models, model compression is also a trend in the deep learning communities, such as quantize the model parameter with k-means [15]. For NMT model, Kim et al. propose the teacher and the student model [16]. The student model extracted the knowledge in the sequence level from the teacher model. The model parameter in the student model is 1/13 times to the teacher model. Reference [17] propose a remarkable method to compress the large depth model into a shallow one via patient knowledge distillation method apply with Transformer. Model compression is one of the key ideas of the proposed Re-Transformer in the paper.

3. The Proposed Model

In this section, we describe the detail of the proposed Re-Transformer for machine translation. The proposed model, that we call Re-Transformer taking the meaning of Re-fine and Re-duce of Transformer, can minimize the training and inference time by reducing the model parameters. Furthermore, through observations of simulation results, we found higher layers in encoder provide better comprehension of input sentence; however, fewer layers in decoder produce the same level of BLEU with reduced processing time. In addition, a better mapping is obtained if we delay the non-linear transformation process. Due to the aforementioned two reasons, Re-Transformer mainly consisted of two refinements of Transformer: i) modify the number of encoder and decoder layers, and ii) modify the point wise feed forward network in encoder to improve the understanding of the source sequence so as to improve the accuracy of output sequence.

3.1. Sub-word Tokenization

In Natural Language Processing, we need to convert text into tokens. In this process, different from tradition word or character tokenization, we use sub-word Tokenization [18]. The method has the ability to process the rare words into sub-word pieces. In the same text dataset, it can reduce the dictionary size to a considerable extent and improve the relationship between affixes. For example, "old", "older", "oldest"the three words in the traditional word embedding. Their relationship cannot be generalized to "smart", "smarter", "smartest", but it is possible in sub-word Tokenization. This is the reason why we apply the sub-word tokenization to preprocess the text dataset.

3.2. Lazy Layer Question

Inspired through [4], we attempt to figure out whether the Transformer needs so many encoder and decoder layers in the particular translation dataset. To this end, we use BLEU [19] to be the translation metric with the model we implemented in various encoder and decoder layers combinations. The arrangement of layers we proposed is considered with Transformer as the baseline to determine the amount of the encoder and the decoder layers. Finally, the combination of six-layer encoder and two-layer decoder comes to be the most efficiency translation model. It can be balanced in training time, inference time and the number of model parameter. The result has been shown in Table 1.

Table 1. BLEU scores with transformer (six-layer encoder) which decoder is six and four layers

Transformer (6-layer encoder)	BLEU	Training Time (produced under the same env.)
6 layers decoder	29.33	1x
4 layers decoder	29.26	0.82x
2 layers decoder	**30.99**	**0.68x**

3.3. Point Wise Feed Forward Network

In the encoder layer of the original Transformer model [1], as shown in (a) of Fig. 1. After each self-attention layer, there is a point wise feed forward layer. The feed forward layer contains a linear and non-linear Relu [20] transformations. The feed forward layer allows the model to capture the important textual information through the larger hidden layer neural size. But in early step, if we apply the non-linear transform in the encoder layer will interfere the self-attention mechanism to understand the important hidden message. Therefore, we try several self-attention and feed forward layer combinations to know the most helpful combination for translation. During the experiments, we even tested the combination with two feed forward layers after one self-attention layer. The experiment results reveal that is an unideal combination.

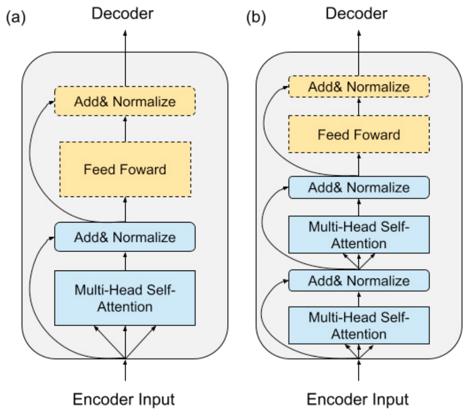

Fig. 1. (a) the encoder layer of Transformer; (b) the modified encoder layer of Re-Transformer.

It was found that the combination of two layers of self-attention connected by a feed forward layer, which means in the six-layered encoder, only three feed forward layers are applied, outperforms all the others.

To better understand the performance of different combination of self-attention and feed forward layers, we have made some comparisons of different types of modified encoders. Modified Encoders varies the combination of feed forward layer and self-attention layer as shown in Table 2. Where characters ‘s’ and ‘f’ represent a stack of self-attention layer and feed forward layer, respectively. Therefore, a stack of 6 sf (i.e., sf-sf-sf-sf-sf-sf) represents the baseline of Transformer.

Table 2. BLEU scores on different arrangement of self-attention and feed forward layer

Modified Encoder	BLEU
sf-sf-sf-sf-sf-sf (baseline)	29.33
s-sf-s-sf-s-sf (Re-Transformer)	**32.14**
s-s-s-sf-sf-sf	29.28

s-s-s-s-sf-sff	30.4
s-s-s-sf-sf-sff	31.4

Through experiments, we find that the encoder with double self-attention layers, in which the self-attention mechanism is able to more accurately understand the words that must be paid attention to instead of prematurely performing non-linear and linear transformation. Such a two-to-one combination of self-attention and feed forward layers outperforms other combinations and becomes the basic building block of the proposed "Re-Transformer".

3.4. Re-Transformer

Re-Transformer-*n* is mainly constructed by the methods mentioned above, where *n* specifies the number of layers in decoder. Unbalanced encoder and decoder layers are adapted; the encoder are six layers and the decoder are four or two layers to reduce the model parameter and increase the speed of inference. Next, we modify the arrangement of the self-attention layer and feed forward layer in the encoder. As the result, the proposed Re-Transformer outperforms the original Transformer in terms of BLEU scores and training time.

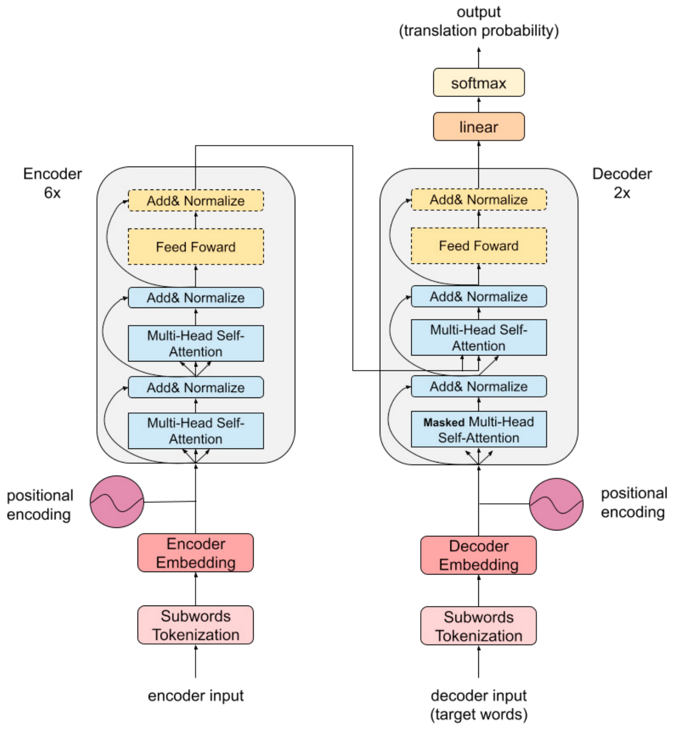

Fig. 2. The model of Re-Transformer.

4. Experiments

4.1. Experiment Settings

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

We conduct the experiments on the WMT 2014 English-German and English-French dataset [21]. Each sentence is tokenized with *subwordtextencoder* and segmented into sub-word units. The results are all based on the test set of newstest 2014. We also use visualization heat map to verify the result, and applied BLEU score on the tokenize, true-case output. The used optimizer is Adam optimizer [22] with $\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 10^{-9}$, and varied the learning rate as [1]. The learning rate formula is demonstrated in the following equation (1).

$$lrate = d_{\mathrm{mod}el}^{-0.5} \cdot \min(step\_num \cdot warmup\_steps^{-1.5}) \tag{1}$$

4.2. Experiment Results

As shown in Table 3. Re-Transformer-2 obtains the best BLEU score of 32.14 and 55.62 in the WMT 2014 English-German and English-French dataset. Compared with the baseline [1], the BLEU score is 27.3, which is about 5 and 17 points of BLEU score higher. In Table 3, the training time experienced for each model per epoch in average under the same environment is also presented. For convenience of observations, it is presented in multiples and the original Transformer [1] is set as the baseline 1x. The training time of Re-Transformer-2 is only 0.56 times to the original Transformer. In addition, 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 obtains a long time for training. The experiment results reveal the training time of the BERT-fused Transformer is around 4.32 times to the proposed Re-Transformer-2. The proposed Re-Transformer not only significantly reduce the training time but also obtains a better BLEU score to that of BERT-fused Transformer.

Table 3. BLEU scores on EN-DE and EN-FR newstest 2014 with models

Models	BLEU (EN-DE)	BLEU (EN-FR)	Training Time (produced under the same env.)
Transformer [1]	27.3	38.1	1x
Transformer + Large Batch [7]	29.3	43.2	--
BERT-fused Transformer [11]	30.75	43.78	2.42x
Re-Transformer-4 (4-layer decoder)	**31.36**	**38.45**	**0.76x**
Re-Transformer-2 (2-layer decoder)	**32.14**	**55.62**	**0.56x**

4.3. Visualization

To better understand where the attentions of the encoders of different models pay to, the visualized heat maps are demonstrated. After six layers of self-attention processing in encoder, we observe how the self-attention mechanism mapped the input sequence. Notably, a brighter color shows a higher attention, while a darker color is less noticed.

Two randomly selected sentences, where sentence 1 is "*That will not happen soon.*", and sentence 2 is "*World currency standards have enormous inertia.*", from the WMT 2014 English-German dataset are demonstrated. Tables 4 and 5 demonstrate the numbers of top1 to top3 brightness blocks and the ratio of correctly attended blocks (i.e. match the keywords) obtained by Transformer and Re-Transformer. The keywords are selected by human. As shown, Re-Transformer is more focused (with a smaller number of brighter blocks) and better attends to the keywords (higher ratio of correctly attended blocks). Through Fig. 3 and 4, it is shown that Transformer sticks on the start and the end symbols of the sequence, and hardly comprehends the sentence semantic. However, Re-Transformer focuses on the correct words, such as "standards" and "enormous".

Table 4. Analysis of heat map in sentence 1 for Transformer and Re-Transformer. The considered key words are "not", "happen", and "soon".

	The number of the brightest blocks	The number of the second bright blocks	The number of the third bright blocks	Sub-word matched correctly
Original Transformer	8	3	5	3/3
Re-Transformer	8	**2**	**3**	2/3

Table 5. Analysis of heat map in sentence 2 for Transformer and Re-Transformer. The considered keywords are “world”, “currency”, “standard”, and “enormous”.

	The number of the brightest blocks	The number of the second bright blocks	The number of the third bright blocks	Sub-word matched correctly
Original Transformer	8	9	5	1/4
Re-Transformer	8	**7**	5	**3/4**

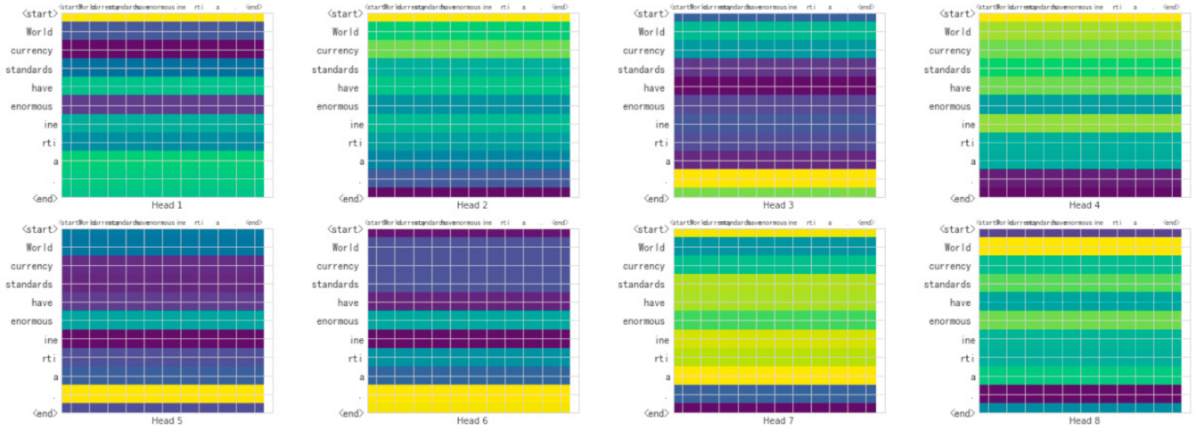

Fig. 3. Heat map of sentence 2: “*World currency standards have enormous inertia.*” obtained by Transformer.

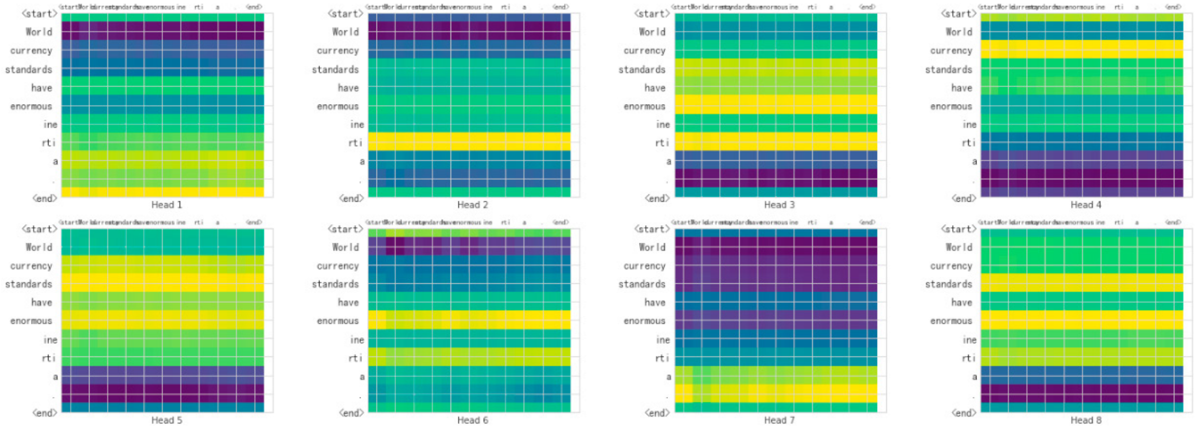

Fig. 4. Heat map of sentence 2: “*World currency standards have enormous inertia.*” obtained by Re-Transformer.

5. Conclusion

In this paper, we proposed a self-attention based model Re-Transformer, which successfully improved the performance of machine translation in both BLEU scores and training time. Re-Transformer improves around 4 and 17 points of BLEU metric against the Transformer over the WMT 2014 English-German and English-French Translation Corpus, while the required average training time is only around half of that of Transformer. According to our experiments, we have the following interesting findings. First, there is no one for all model in machine translation tasks, for different languages the best adapted models vary. Second, stacking more layers of self-attention before feed forward layer in encoder provides better comprehension to the original input. However, a balance ratio between self-attention and feed forward layer needs to find for different NLP tasks. Third, reducing self-attention layer for decoder produces better outcomes. Too much attentions processes among original input and temporary outputs seems lead to the same or even worse BLEU scores. According to the above observations, there are some interesting future works. The balanced between self-attention and feed forward layer in encoder for different language families. Is there a better model for decoder to understand the relationship of original input and the current temporary output? These two points will be the focus of our future works.

References

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017, December). Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (pp. 6000-6010).

[2] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... & Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.

[3] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[4] V Xu, H., van Genabith, J., Xiong, D., & Liu, Q. (2020). Analyzing Word Translation of Transformer Layers. arXiv preprint arXiv:2003.09586.

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

[6] Gehring, J., Auli, M., Grangier, D., Yarats, D., & Dauphin, Y. N. (2017, July). Convolutional sequence to sequence learning. In International Conference on Machine Learning (pp. 1243-1252). PMLR.

[7] Ott, M., Edunov, S., Grangier, D., & Auli, M. (2018, October). Scaling Neural Machine Translation. In Proceedings of the Third Conference on Machine Translation: Research Papers (pp. 1-9).

[8] Shi, X., Huang, H. Y., Wang, W., Jian, P., & Tang, Y. K. (2019, November). Improving Neural Machine Translation by Achieving Knowledge Transfer with Sentence Alignment Learning. In Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL) (pp. 260-270).

[9] Press, O., Smith, N. A., & Levy, O. (2020, July). Improving Transformer Models by Reordering their Sublayers. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 2996-3005)..

[10] Zhu, J., Xia, Y., Wu, L., He, D., Qin, T., Zhou, W., ... & Liu, T. (2019, September). Incorporating BERT into Neural Machine Translation. In International Conference on Learning Representations.

[11] Clinchant, S., Jung, K. W., & Nikoulina, V. (2019, November). On the use of BERT for Neural Machine Translation. In Proceedings of the 3rd Workshop on Neural Generation and Translation (pp. 108-117).

[12] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[13] Bapna, A., Chen, M. X., Firat, O., Cao, Y., & Wu, Y. (2018). Training Deeper Neural Machine Translation Models with Transparent Attention. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 3028-3033).

[14] Wu, L., Wang, Y., Xia, Y., Tian, F., Gao, F., Qin, T., ... & Liu, T. Y. (2019, July). Depth Growing for Neural Machine Translation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 5558-5563).

[15] Gong, Y., Liu, L., Yang, M., & Bourdev, L. (2014). Compressing deep convolutional networks using vector quantization. arXiv preprint arXiv:1412.6115.

[16] Kim, Y., & Rush, A. M. (2016, January). Sequence-Level Knowledge Distillation. In EMNLP.

[17] Sun, S., Cheng, Y., Gan, Z., & Liu, J. (2019, November). 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 (EMNLP-IJCNLP) (pp. 4314-4323).

[18] Sennrich, R., Haddow, B., & Birch, A. (2016, August). 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) (pp. 1715-1725).

[19] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). 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).

[20] Agarap, A. F. (2018). Deep learning using rectified linear units (relu). arXiv preprint arXiv:1803.08375.

[21] Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., ... & Tamchyna, A. (2014, June). Findings of the 2014 workshop on statistical machine translation. In Proceedings of the ninth workshop on statistical machine translation (pp. 12-58).

[22] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.