OPPO NMT System for IWSLT 2019

Xiaopu Li, Zhengshan Xue, Jie Hao

Manifold Lab, OPPO Research Institute, Beijing
{lixiaopu,xuezhengshan,haojie}@oppo.com

Abstract

This paper illustrates the OPPO's submission for IWSLT2019 text translation task Our system is based on Transformer architecture. Besides, we also study the effect of model ensembling. On the devsets of IWSLT 2019, the BLEU of our system reaches 19.94.

1 Introduction

Neural machine translation has recently received huge attention and has been introduced as a promising paradigm with the potential to overcome any shortcomings of traditional statistical machine translation system.[1] In this paper, we describe the OPPO ‘s submission about the neural machine translation system for the IWSLT 2019 English to Czech translation task.

The neural machine translation system consists of an encoder and a decoder, while the encoder embeds tokens of the source sentence into a series of feature vectors which are used by the decoder to generate translation results. Besides, attention is also employed to better model the link between the source sentences and the target sentences.

2 Model description

Neural machine translation exploits an encoder-decoder framework to model the whole translation process in an end-to-end fashion, and has achieved state-of-the-art performance in many language pairs. Among various translation models, the Transformer model based on self-attention mechanism has shown promising results in terms of both translation performance and training speed. The model architecture is shown in Figure 1. We also us the transformer architecture. We also use BPE technology to alleviate the UNK problem. Besides, model ensembling is used to further improve the final results of the system.

We utilize layer normalization[3] to adaptively learn to scale and shift the incoming activations of a neuron on a layer-by-layer basis at each time step. Layer normalization can stabilize the dynamics of layers in the network and accelerate the convergence speed of deep neural networks.

The two components of the transformer are the encoding and decoding framework, which are composed of and linked by attention mechanism, while the position encoding provide the order information of the sentences.

2.1 Encoding and Decoding

As shown in figure 1 (got from[4]), the transformer is composed of a multi-layer encoder and decoder. Each layer in the encoder/decoder contains two sub layers, of which one is multi-head attention layer and the other is position-wise fully connected feed-forward network. we also employ a residual connection around each of the sub-layers followed by layer normalization.

2.2 Attention

An attention function can be viewed as mapping a query and a set of key-value pairs to an output, where the query, keys, values and output are all vectors.[4] The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key. We employ multi-head self-attention both in encoder and decoder layer and global attention between the encoder and the decoder.

2.3 Position Encoding

In order for the model to make use of the order of the sequence, some information about the position of the tokens in the sequence must be injected. We use sine and cosine functions as did in [4].

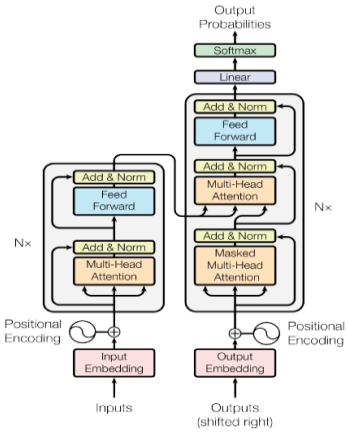

Figure 1: The Transformer model architecture

3 Experiments and Results

3.1 Data Processing

The training data for the English-Czech translation task consists of 57 million sentence pairs of version 1.7 from the WMT English-Czech corpus.

We use the official MUST-C as validation set.

The data of this corpus may be got from the webs and there are many translation, spelling errors, and mismatched sentences etc. Therefore, to get high-level clean parallel data, we use fast_align to filter the data.

Step 1: using "fast_align" to generate the forward and backward alignment with scores S_{fw} and S_{bw}

Step 2: using the alignments got by last step to generate grow-diag-final-and a_{gdfa}

Step 3:then calculate the final score by $(S_{fw} + S_{bw}) / len(a_{gdfa})$, sort and delete the sentences with low score.

Then We first tokenize the remaining 27 million English and Czech sentences with scripts provided in Moses and use BPE segmentation to process both source and target data. 32k subword symbols are used for both the English and Czech sentences.

3.2 Model Ensemble

Model ensemble is a widely used technique to boost the predictions of several models at each decoding step.[6] During training, we experiment with different weight initialization using different seeds. During the inferencing, we ensemble as many as 13 models from different checkpoints and different seeds. The experimental results indicate that this method achieves absolute improvements over the single system.

3.3 English to Czech System

All the weight parameters are initialized uniformly and we use dropout to improve the generalization as suggested by Zaremba et al.[7]. We use Adam[8] to train the model with a learning rate 0.0003.We use multi-GPUs training via synchronous SGD and data parallelism. We train the model on a host server with 8 NVIDIA Tesla V100 GPUs. Transformer-base and Transformer big model are compared during training and we finally choose Transoformer-big as our baseline model. We train the same network with different random seeds of parameters initialization and ensemble a series of models to further improve the performance.

3.4 Result

Table 1 presents the BLEU scores on IWSLT2019 development set of English->Czech task. First of all, the result shows, compared with Transformer based raw data training, model trained with cleaned data brings +1.92 BLEU. When Transformer-big model is used, +0.80 BLEU is improved further. Addtionally, +1.02 BLEU increases when model ensemble is applied in the training.

Table 1: BLEU scores on devset for en-cz translation

System	BLEU
Transformer-base(raw data)	16.20
Transformer-base(data cleaning)	18.12
Transformer-big(data cleaning)	18.92
+model ensemble	19.94

4 Conclusions

This paper has described OPPO's work for IWSLT 2019 text translation for English to Czech. During the experiments, We employ the novel Transformer architecture, which has been proved as the state of the art. Besides, data cleaning and filtering is also important for the final results. Finally, model ensemble boost the score significantly.

References

[1] Wang Y, Cheng S, Jiang L, et al. Sogou neural machine translation systems for wmt17[C]//Proceedings of the Second Conference on Machine Translation. 2017: 410-415.

[2] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units[J]. arXiv preprint arXiv:1508.07909, 2015.

[3] Ba J L, Kiros J R, Hinton G E. Layer normalization[J]. arXiv preprint arXiv:1607.06450, 2016.

[4] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.

[5] Koehn P, Hoang H, Birch A, et al. Moses: Open source toolkit for statistical machine translation[C]//Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions. 2007: 177-180.

[6] Lee, K.-F., Automatic Speech Recognition: The Sun M, Jiang B, Xiong H, et al. Baidu Neural Machine Translation Systems for WMT19[C]//Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1). 2019: 374-381.

[7] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. The journal of machine learning research, 2014, 15(1): 1929-1958.
[8] Dozat T. Incorporating nesterov momentum into adam[J]. 2016.