Towards Full Utilization on Mask Task for Distilling PLMs into NMT

Anonymous ACL submission

Abstract

Owing to being well-performed in many natural language processing tasks, the application of pre-trained language models (PLMs) in neural machine translation (NMT) is widely concerned. Knowledge distillation (KD) is one of the mainstream methods which could gain considerable promotion for NMT models without extra computational costs. However, previous methods in NMT always distill knowledge at hidden states level and can not make full use of the teacher models. For solving the aforementioned issue, we propose KD based on mask task as a more effective method utilized in NMT which includes encoder input conversion, mask task distillation, and gradient optimization mechanism. Here, we evaluate our translation systems for English→German and Chinese→English tasks and our methods clearly outperform baseline methods. Besides, our framework can get great performances with different PLMs.

1 Introduction

Aimed at improving the performance of neural machine translation (NMT), pre-trained language models (PLMs) are applied to enhance Transformer (Vaswani et al., 2017) by either using PLMs as extra inputs or distilling knowledge from PLMs to NMT model(Clinchant et al., 2019; Zhu et al., 2020; Weng et al., 2020). Among these two approaches, knowledge distillation (KD) (Bucila et al., 2006; Hinton et al., 2015) maintains the original structure of the Transformer, leading to an improvement without extra computational costs. For example, by taking BERT (Devlin et al., 2018) as the teacher model, the encoder in Transformer is chosen as the student model could acquire knowledge from the hidden states of the teacher model(Yang et al., 2020; Wu et al., 2020). This kind of KD acquires knowledge from the hidden states of PLMs(cf. Fig. 1a).

However, KD at hidden states level can not make full use of the teacher model, which may miss some knowledge from PLMs. At least, the lacking of classifier layer lose some information from the class probabilities produced by the teacher model, which is called the "soft targets" (Hinton et al., 2015). On the contrary, if we could imitate the process of pre-training tasks from PLMs, the student model can take advantage of the complete knowledge distilled by the teacher model.

In this paper, we propose KD based on mask task in NMT to improve the performance of the Transformer in NMT (cf. Fig. 1b). In our framework, we take advantage of the whole structure of the teacher model in KD by distilling the logits from the mask task. To adapt to KD based on mask task, we design three strategies, namely encoder input conversion, mask task distillation, and gradient optimization mechanism. In particular, we use the same tokenizer as the teacher model and mask part of tokens. Besides, we add a classifier layer for encoder. The encoder needs to accomplish both the translation task and mask task simultaneously. The objective is to absorb the monolingual knowledge from the teacher model while taking on the role of the encoder of translation. And we propose the gradient optimization mechanism to alleviate the conflict between the NMT task and the KD task and guarantee the efficiency of KD in NMT.

To demonstrate the effectiveness of our framework, we implement the proposed approaches based on the advanced pre-trained models and Transformer model. Experimental results on WMT14 English to German and WMT19 Chinese to English machine translation tasks show that our approach outperforms the Transformer baseline and the others KD methods.

The main contributions can be summarized as:

- We are the first to put forward to utilize the whole structure of the teacher model to distill knowledge in NMT, which avoids the loss of knowledge in KD at hidden states level;

• We propose KD based on mask task in NMT. For adjusting our framework, we propose three strategies: encoder input conversion, mask task distillation, and gradient optimization mechanism.

• We evaluate our framework with different PLMs on several large-scale benchmark datasets. Our experiments show significant improvement over other methods.

2 Background

2.1 Pre-trained Language Models Based on Mask Task

PLMs like BERT (Devlin et al., 2018) have shown strong performance gains using self-supervised training that requires a large collection of unlabeled text. One of the most significant training objectives is the masked language model (MLM) which predicts masked individual words. In MLM's implementation, 15% of the tokens are randomly selected; of those, 80% are replaced with [MASK], 10% are replaced with a random token, and 10% are kept unchanged. The task is to predict the original tokens from the modified inputs.

Based on the MLM task, more advanced tasks are proposed to train PLMs. SpanBERT (Joshi et al., 2020) presents a pre-training method that masks contiguous random spans based on geometric distribution, rather than random individual tokens.

2.2 Knowledge Distillation in Neural Machine Translation

KD is an effective method that can help student network obtain knowledge from a large and accurately trained teacher network. In KD, $\theta_{\mathcal{S}}$ and $\theta_{\mathcal{T}}$ are the sets of parameters of the student model and the teacher model which are usually trained to minimize the negative log-likelihood. The KD loss can be formulated as:

$$\mathcal{L}\left(\theta_{\mathcal{T}}, \theta_{\mathcal{S}}\right) = -||H_{\mathcal{T}} - H_{\mathcal{S}}||_2^2 \tag{1}$$

where $H_{\mathcal{T}}$ and $H_{\mathcal{S}}$ are the hidden states of the teacher model and student model, respectively.

Following KD at hidden states level, Yang et al., 2020 proposes asymptotic distillation, which utilizes the second-to-last layer of BERT and works significantly better than other hidden states. Wu et al., 2020 utilizes all hidden layers in PLMs and adds the layer mixing mechanism for intermediate layers to distill more knowledge from the teacher model. However, these methods still can not make full use of the teacher models.

3 Methodology

Distilling knowledge from PLMs is a useful complement to provide NMT models with proper language knowledge. Previous methods concentrate on distilling knowledge from the hidden states of PLMs. We propose a novel framework that could utilize the logits from the mask task. We imitate the process of pre-training tasks from the teacher models and make adjustment for the student model. Our method can mask full use of the whole PLMs which contains three steps including encoder input conversion, mask task distillation, and gradient optimization mechanism. We will introduce three steps in detail.

3.1 Encoder Input Conversion

Contrary to the previous methods, the encoder part accepts different inputs from the NMT task and the mask task. The encoder tokenizer is replaced by the teacher's tokenizer, which permit the unity of positions of tokens. After that, some tokens are masked according to the mask pre-training task of PLMs. Specifically, given a sequence of source language words $X = \{x_1, x_2, \cdots, x_m\}$ and corresponding target language words $Y = \{y_1, y_2, \cdots, y_n\}$, we can get encoder input $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m\}$, mask encoder input $\mathbf{X}' = \{\mathbf{x}'_1, \mathbf{x}'_2, \cdots, [\mathbf{MASK}], \mathbf{x}'_m\}$, and decoder input $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_m\}$. $\mathbf{X}$ and $\mathbf{Y}$ are used in NMT task while $\mathbf{X}$ and $\mathbf{X}'$ are used in KD task.

3.2 Mask Task Distillation

Different tasks are executed in parallel, while the data in the encoder are different. For the NMT task, we merely transform the encoder tokenizer to the PLMs' tokenizer. Transformer is optimized by maximizing the likelihood, denoted by

$$\mathcal{L}_{\mathrm{NMT}} = -\sum_{i=1}^{n} \log P\left(y_i \mid \mathbf{y}_{<i}, \mathbf{X}\right) \tag{2}$$

Then, different from previous knowledge distillation methods(Yang et al., 2020), a classifier layer participates in the KD task to assist the mask prediction of the encoder. Our objective learns the logits information from the PLMs directly:

$$\mathcal{L}_{KD}\left(\theta_{\mathcal{T}}, \theta_{\mathcal{S}}\right) = -\sum_{v=1}^{|V|} q\left(y=v \mid \mathbf{x}'; \theta_{\mathcal{T}}\right) \times \log p\left(y=v \mid \mathbf{x}'; \theta_{\mathcal{S}}\right) \tag{3}$$

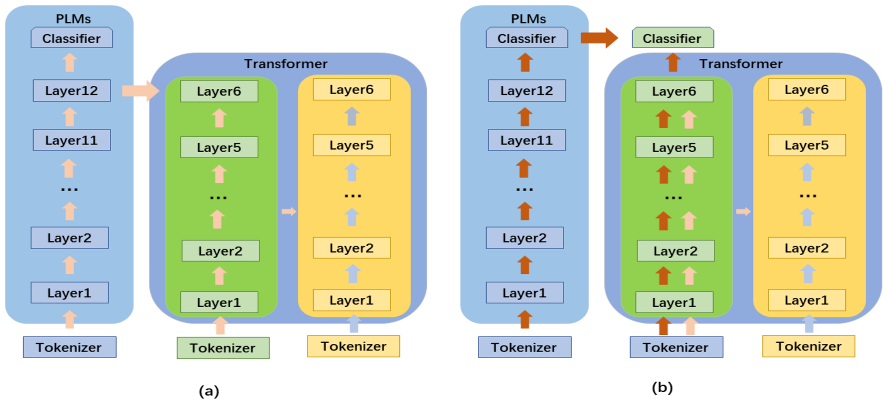

Figure 1: (a): Overview of KD at hidden states level: NMT task and KD task share same data, and encoder learns from the hidden states of the PLMs; (b): Overview of our framework: our method utilizes the PLMs tokenizer and adds an classifier layer.we also utilize different inputs between different tasks.

where $|V|$ is the the number of words in source language dictionary.

Finally, the loss function of our framework is:

$$\mathcal{L}_{ALL} = \mathcal{L}_{NMT} + \alpha\mathcal{L}_{KD} \tag{4}$$

where α is used to balance the preference among the two losses.

3.3 Gradient Optimization Mechanism

For reducing conflicts between the NMT task and KD task, we propose the gradient optimization mechanism(GOM). Inspired by multi-task learning(Zhao et al., 2018), we evaluate conflicts between tasks with the direction of the gradient and reduce them by the gradient optimization strategy.

More specifically, given a mini-batch of training samples, the gradient in the encoder $\nabla\theta$ will be influenced by the NMT task and the KD task, $\nabla\theta = \nabla\theta_{NMT} + \nabla\theta_{KD}$, where $\nabla\theta_{NMT}$ and $\nabla\theta_{KD}$ denote gradients from the NMT task and the KD task. As an auxiliary task of the NMT task, $\nabla\theta_{KD}$ whether to be withhold depends on the direction of the gradient.The destructive interference from the KD task can be measured by

$$\textbf{sign} = sign(\langle\nabla\theta_{NMT}, \nabla\theta_{KD}\rangle) \tag{5}$$

$$\nabla\theta = \begin{cases} \nabla\theta_{NMT} + \nabla\theta_{KD}, & \textbf{sign} > 0 \\ \nabla\theta_{NMT}, & \textbf{sign} <= 0 \end{cases} \tag{6}$$

For each module in the encoder, we calculate the sign of the gradient separately.

4 Experiment

4.1 Implementation Detail

Data-sets We carry out experiments on large-scale machine translation tasks: WMT'14 English-German (En-De) and WMT'19 Chinese-English (Zh-En). For En-De task, we use 4.5M preprocessed data. We use newstest2013 as the validation set and newstest2014 as the test set, which contain 3000 and 3003 sentences, respectively. For Zh-En task, we use 20.4M preprocessed data. We use newstest2018 as our validation set and newstest2019 as our test set, which contain 3981 and 2000 sentences, respectively. We also measure case sensitive BLEU with significance test (Koehn, 2004) for En-De and Zh-En, respectively.

Settings Following the setting in Vaswani et al., 2017, we carry out our experiments on standard Transformer (Vaswani et al., 2017) with the fairseq toolkit (Ott et al., 2019). For Transformer(Base), we set the dimension of the input and output of all 6 layers as 512, and that of the feed-forward layer to 2048. We employ 8 parallel attention heads. In training processing, we use Adam optimizer with 1 = 0.9, 2 = 0.98, learning rate is 7e-4 and dropout is 0.1. All experiments are conducted using 4 NVIDIA V100 GPUs, where the batch size of each GPU is set to 4096 tokens. The α in equation(4) ranges from [0.25, 0.5, 0.75] and we choose on the performance of validation set. We train each model for 200,000 steps and save in every 5,000 steps. At last, we average the last five checkpoints for testing.

We choose BERT and SpanBERT(Joshi et al., 2020) as the teacher PLMs in our experiments.

Models	Method	En-De	Zh-En
Transformer(Base)	-	27.96	24.63
Transformer(Base)+BERT	KD in Hidden States	28.20	24.88
	KD on Mask Task	**28.71***	**25.14***
Transformer(Base)+SpanBERT	KD in Hidden States	27.67	-
	KD on Mask Task	28.36*	-

Table 1: Case-sensitive BLEU scores on English-German and Chinese-English.'*': significantly ($p < 0.01$) better than Transformer (Base).

With regard to the mask strategies, we follow the same regulations and proportions as the teacher models. The experiments in Zh-En with SpanBERT are missing because of the lack of resources of SpanBERT-Chinese.

4.2 Main Results

The results are shown in Table 1. We also list the Transformer baseline and the result of the KD at hidden states level. Compared with baseline, Transformer with the BERT based on mask task KD improves 0.75 BLEU scores and 0.51 BLEU in En-De and Zh-En, which outperforms Transformer with the BERT in the hidden states obviously. The experiment with the SpanBERT also improves 0.4 BLEU scores than the baseline while the KD with SpanBERT at hidden states level incurs the decline in the BLEU scores. It is apparent that our framework can improve about 0.5 BLEU scores than KD at hidden states level with different PLMs as teacher models.

4.3 Impact of Different Inputs

Method	Input	BLEU
Transfromer(Base)	-	27.96
KD in Hidden States	orign	28.20
	mask	28.07
KD on Mask Task	orign	28.46
	mask	28.53

Table 2: Impact of different inputs in WMT'14 En-De.

To show the effectiveness of input, we do a detailed ablation study with BERT as shown in Table 2. We use different distillation strategies with different inputs, and analyze the influence of different input strategies. On the one hand, the strategy of masking input is not applicable for KD at hidden states level, for which leads to the decline of BLEU scores. On the other hand, KD based on task can improve about 0.5 BLEU scores without GOM, and the rest improvement owing to the mask input. Compared with the input strategies, the selection of which part to distill can bring more promotion.

According to the experiments above, we conclude that the most effective step is mask task distillation, while encoder input conversion can enhance the effect of distillation for further improvement.

4.4 Impact of Gradient Optimization Mechanism

Method	GOM	BLEU
KD in Hidden States	×	28.20
	✓	28.44
KD on Mask Task	×	28.53
	✓	28.71

Table 3: Impact of Gradient Optimization Mechanism in WMT'14 En-De.

We also evaluate the effectiveness with and without gradient optimization mechanism in different KD methods. As shown in Table 3, GOM can improve about 0.2 BLEU scores in either KD at hidden states level or KD based on mask task compared with KD directly. It reveals that GOM has a positive influence on the benefit NMT task by reducing the conflicts between the NMT task and KD task effectively.

5 Conclusion

In this paper, we first address the situation of KD in NMT and the disadvantages of KD at hidden states level. Then, we propose KD based on mask tasks, which overcomes the drawback of current KD in NMT and bring improvement. Moreover, we apply our framework to other PLMs with mask tasks and prove the effectiveness. Experiments show that our framework can achieve remarkable performance on the WMT En-De and Zh-En benchmark datasets.

References

C. Bucila, R. Caruana, and A. Niculescu-Mizil. 2006. Model compression. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD'06)*.

Stephane Clinchant, Kweon Woo Jung, and Vassilina Nikoulina. 2019. On the use of bert for neural machine translation. *arXiv preprint arXiv:1909.12744*.

J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

G. Hinton, O. Vinyals, and J. Dean. 2015. Distilling the knowledge in a neural network. *Computer Science*, 14(7):38–39.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 388–395.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention is all you need. *arXiv*.

Rongxiang Weng, Heng Yu, Shujian Huang, Shanbo Cheng, and Weihua Luo. 2020. Acquiring knowledge from pre-trained model to neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9266–9273.

Y. Wu, P. Passban, M. Rezagholizade, and Q. Liu. 2020. Why skip if you can combine: A simple knowledge distillation technique for intermediate layers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

J. Yang, M. Wang, H. Zhou, C. Zhao, and L. Li. 2020. Towards making the most of bert in neural machine translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(5):9378–9385.

Xiangyun Zhao, Haoxiang Li, Xiaohui Shen, Xiaodan Liang, and Ying Wu. 2018. A modulation module for multi-task learning with applications in image retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 401–416.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. 2020. Incorporating bert into neural machine translation. *arXiv preprint arXiv:2002.06823*.