A Self-Distillation Recipe for Neural Machine Translation

Hongfei Xu[1], Zhuofei Liang[1], Qiuhui Liu[2], Lingling Mu[1]
[1]Zhengzhou University, Henan, China, [2]China Mobile Online Services, Henan, China
{hfxunlp, zfliangnlp, liuqhano}@foxmail.com
Correspondence: Lingling Mu iellmu@zzu.edu.cn

Abstract

Self-distillation distills the deeper sub-networks to the shallower sub-networks without using an extra teacher model, and has been proven effective in improving the performance of a series of computer vision tasks. In this paper, we study the representation-based self-distillation methods for Neural Machine Translation (NMT) to avoid the efficiency issue of probability distribution based Knowledge Distillation (KD) with a large vocabulary. We present a rank-order augmented Pearson correlation loss and an iterative distillation method to prevent the discrepancy of predictions between the student and a stronger teacher from disturbing the training. To prevent the teacher from misleading the student's learning, we utilize a warm-up strategy and present a gradient adaption method to scale down or zero the knowledge distillation gradients which are opposite to the translation. Experiments on the low-resource IWSLT 14 German to English, middle-resource WMT 14 English to German, and high-resource WMT 15 Czech to English and WMT 14 English to French tasks show that our method can lead to significant improvements over the strong Transformer baselines, obtaining comparable performance to previous machine translation knowledge distillation studies without pre-training a teacher. Experiments with shallower/deeper Transformers show that our method can lead to comparable or better performance efficiently with fewer layers. Our method is also effective in the multilingual setting or with recurrent decoder.

1 Introduction

The Transformer translation model (Vaswani et al., 2017) has greatly improved the performance of Neural Machine Translation (NMT) (Sutskever et al., 2014; Bahdanau et al., 2015; Gehring et al., 2017), especially when empowered by deeper structures (Zhou et al., 2016; Wang et al., 2017; Bapna et al., 2018; Wang et al., 2019; Wu et al., 2019; Zhang et al., 2019a; Wei et al., 2020; Xu et al., 2020a, 2021c; Kasai et al., 2021; Li et al., 2022; Hao et al., 2022; Xu et al., 2024). Knowledge Distillation (KD) (Hinton et al., 2015) that first trains a teacher model and then transfers the teacher's knowledge to the student network by mimicking the output of the teacher, is among the most promising solutions to ensure the efficiency and to improve the performance of machine translation (Kim and Rush, 2016; Wu et al., 2020; Wang et al., 2021; Jafari et al., 2021; Liang et al., 2022; Miao et al., 2022, 2023; Zhang et al., 2023).

Self-distillation (Zhang et al., 2019c, 2022) performs knowledge distillation inside the same neural network, normally distilling deeper sub-networks to shallower ones. It only requires one-stage training and does not need to search for the most proper teacher model for knowledge distillation, thus can significantly reduce training overhead and facilitate efficient and effective knowledge distillation for computer vision tasks. Zhang et al. (2019c, 2022) show that self-distillation can lead to strong results without an additional teacher model on a number of computer vision tasks.

In this paper, we study self-distillation methods for neural machine translation. To address the efficiency issue with a large vocabulary in neural machine translation, we distill deeper representations to shallower layers. We present a rank-order augmented Pearson correlation loss and an iterative distillation method to prevent potential training disturbance due to the performance discrepancy between the students and the stronger teacher (Cho and Hariharan, 2019; Mirzadeh et al., 2020; Son et al., 2021; Huang et al., 2022b), and a gradient adaptation method together with a warm-up strategy to avoid possible misleading during knowledge distillation.

Our main contributions are as follows:

Findings of the Association for Computational Linguistics: ACL 2025, pages 5050–5064

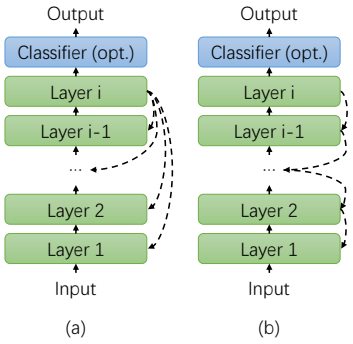

Figure 1: (a) deep-to-shallow distillation, (b) iterative distillation. Solid and dashed arrows indicate the directions for forward propagation and knowledge distillation respectively.

- We study efficient self-distillation methods for neural machine translation, present an iterative distillation method to reduce the capacity gap between teachers and students, propose a gradient adaptation method and adopt a warm-up strategy to prevent potential negative impacts of knowledge distillation on machine translation, and present a rank-order augmented Pearson correlation loss to further ensure the effectiveness of representation-based knowledge distillation.

- Experiments on low/middle/high-resource tasks and the OPUS-100 massively multilingual translation task show that our method can lead to significant improvements over strong baselines, obtain comparable BLEU scores to previous machine translation knowledge distillation methods efficiently without pre-training a teacher, and achieve comparable or better performance efficiently with fewer layers.

2 Our Method

As NMT usually has a large vocabulary size, which leads to high computation and memory costs to distill between prediction probability distributions, we study efficient self-distillation based on the hidden representations produced during the forward propagation of the model.

Since both the teachers and students are trained together from random initialization with self-distillation, the teachers may be less likely to provide reliable guidance for the training of students at the initial stage. We adopt a warm-up strategy disabling the knowledge distillation and to train only with the MT loss at the beginning.

2.1 Iterative Self-Distillation

Zhang et al. (2019c, 2022) distill the output predictions of the deepest layer to shallow layers by default, under the expectation that the deepest layer may generally offer the best performance than shallower layers, as shown in Figure 1 (a).

But as pointed out by Cho and Hariharan (2019); Mirzadeh et al. (2020), stronger teachers do not always lead to better distillation performance, and knowledge distillation can even adversely affect the training of the student if there is a large gap between teacher and student in capacity. The student may not have sufficient capacity to minimize both the training loss and the knowledge distillation loss in challenging settings, and might end up minimizing one loss (knowledge distillation loss) at the expense of the other (classification loss). This unfortunately may also apply to neural machine translation with self-distillation when distilling the deepest layer to some shallower layers.

To address the training issue due to the disparity between teachers and students in capacity, we propose to perform self-distillation iteratively. Instead of distilling the deepest layer to all shallow layers, we distill the output of layer k to the output of layer $k-1$, as shown in Figure 1 (b). With iterative self-distillation, the capacity differences between teachers and students are minimized to prevent the distillation from disturbing the training.

Iterative distillation is identical to transitive distillation in Zhang et al. (2022), but they obtain better performance by distilling with a subset of shallow layers, while the all-layer iterative distillation setting leads to better performance in our experiments (Table 7). This may be due to the utilization of the other strategies like gradient adaptation. Zhang et al. (2022) also do not take the capacity gap as motivation.

2.2 Gradient Adaptation for Knowledge Distillation

In self-distillation, the model is randomly initialized and the teacher sub-networks and student sub-networks are trained together within the same model. As a result, the teachers may not always be better than students and may mislead the students. Knowledge distillation may also disturb the students training in case a student sub-network's capacity is far from matching its teacher.

We present a gradient adaptation method to prevent the student sub-network's training from po-

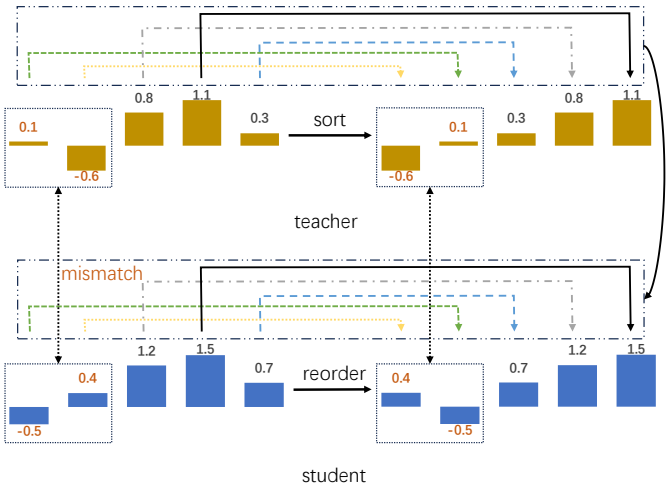

Figure 2: Rank-order based distillation loss.

tential interference due to knowledge distillation. Specifically, the output vector of the student subnetwork o_s receives two gradient vectors during backpropagation: one ($\frac{\partial loss_{mt}}{\partial o_s}$) back-propagated from the classifier with the machine translation loss ($loss_{mt}$), another ($\frac{\partial loss_{kd}}{\partial o_s}$) back-propagated from the knowledge distillation loss ($loss_{kd}$) which mimics the output vectors of the student and its corresponding teacher. We first compute the cosine similarity s between the gradient vectors from the translation loss and the distillation loss.

$$s = \cos(\frac{\partial loss_{mt}}{\partial o_s}, \frac{\partial loss_{kd}}{\partial o_s}) \quad (1)$$

Next, we zero negative similarity and use the result as the weight to aggregate the translation and knowledge distillation gradient vectors to obtain the final gradient for the student's output vector o for follow-up back-propagation.

$$\frac{\partial loss}{\partial o_s} = \frac{\partial loss_{mt}}{\partial o_s} + \max(0, s) * \frac{\partial loss_{kd}}{\partial o_s} \quad (2)$$

In this way, the distillation gradient with a higher similarity is preferred with a higher weight during gradient aggregation, and the distillation gradient in contrast to the translation gradient is removed.

2.3 Rank-Order Augmented Distillation Loss

Huang et al. (2022b) find that when the teacher and student models are trained with a stronger strategy (e.g., exact match), the discrepancy between teacher and student would be fairly larger, and an exact recovery could be challenging and lead to the failure of knowledge distillation. They suggest that preserving the relation of predictions between teacher and student is sufficient and effective, and propose to leverage the Pearson correlation coefficient for the knowledge distillation between students and stronger teachers.

However, the Pearson correlation coefficient does not treat each dimension of the input vectors equivalently. It may be trivial to optimize when self-distillation with hidden vectors instead of probabilities, when the teachers are also updated together with the students but not fixed. For example, the students could learn to generate comparably large scalars in a few dimensions of the vector and affect the teachers' outputs to be also large at these dimensions via residual connections. These dimensions could have a huge impact on the coefficient and easily lead to a high correlation.

To ensure all vector dimensions are not neglected during knowledge distillation, we augment the Pearson correlation loss with another loss based on the inconsistency between rank orders of the scalars inside the teacher and student vectors. Specifically, we first sort scalars in the teacher vector in the ascending order, then reorder scalars in the student vector in exactly the same way as the reordering of the teacher vector during sort. The rank-order based distillation loss is the sum of non-negative subtraction results between each scalar and its subsequent scalar in the reordered student vector, as illustrated in Figure 2.

The rank-order based distillation loss gets the scalar orders by sorting and reordering, and tries to ensure the consistency of scalar orders between

the teacher and student vectors. The loss subtracts scalars in the student vector with inconsistent orders compared to corresponding scalars in the teacher vector, and leads to decent and consistent gradient values of 1 and -1 respectively for the scalars in the reverse order regardless of their values.

The rank-order based loss is added to the negative Pearson correlation coefficient as the final knowledge distillation loss.

2.4 Efficiency Discussion

Despite we leverage several strategies to facilitate effective self-distillation for neural machine translation, the warm-up strategy simply disables self-distillation at the initial stage of training, and all the other strategies only involve light-weight element-wise dense vector computations. We also distill with hidden vectors instead of probabilities for efficiency with the large NMT vocabulary, without attaching classifiers for shallow layers and using high-dimensional prediction logits for knowledge distillation. As a result, our method only takes 6% more training time and performs better than previous studies without pre-training a teacher (Table 1).

The inference of the self-distillation model is the same as the baseline model, using all layers. Our method does not support dynamic inference with shallow layers as Zhang et al. (2022) for we do not attach classifiers to shallow layers. Unlike conventional machine translation knowledge distillation methods which improves the inference efficiency by distilling a larger/deeper teacher model to a smaller/shallower student network, our method directly trains the shallow network via self-distillation, and the inference efficiency is achieved by obtaining comparable performance with shallower models (Table 4).

3 Experiments

3.1 Settings

Datasets Our experiments covered the low-resource IWSLT 14 German (De) to English (En), middle-resource WMT 14 English to German, and high-resource WMT 15 Czech (Cs) to English and WMT 14 English to French (Fr) tasks to show the effectiveness of our approach, comprising around $174k$, $4.5M$, $15M$ and $33M$ sentence pairs respectively. We applied joint Byte-Pair Encoding (BPE) (Sennrich et al., 2016) with $16k$ merge operations for the low-resource IWSLT 14 De-En task, and $32k$ merge operations for the other tasks to address the unknown word issue. We only kept sentences with no more than 256 tokens for training.

Model	BLEU
Transformer (Vaswani et al., 2017)	27.54
Word KD (Kim and Rush, 2016)	28.03
Seq KD (Kim and Rush, 2016)	28.22
BERT KD (Chen et al., 2020)	27.53
Seer Forcing (Feng et al., 2021)	27.56
Annealing KD (Jafari et al., 2021)	27.91
Selective KD (Wang et al., 2021)	28.24
CBBGCA (Zhou et al., 2022)	28.36
TIE KD (Zhang et al., 2023)	28.46
AIO KD (Miao et al., 2023)	28.75
Self KD (Ours)	**29.21**

Table 1: Main results on the WMT 14 En-De task.

Hyper-parameters The hyper-parameters derived from the baselines are detailed in Appendix A. The number of knowledge distillation warm-up steps is twice as the number of learning rate warm up steps ($8k$). We simply summed up individual losses, because this is more friendly for practice than tuning hyper-parameters for each specific task. The results with tuned loss aggregation hyper-parameters are provided in Appendix B.

Evaluation We used a beam size of 4 for decoding with the averaged model of the last 5 checkpoints saved with an interval of 1500 training steps. We evaluated with BLEU and chrF implemented by the sacreBLEU toolkit (Post, 2018) and the COMET score (Rei et al., 2020). We also conducted significance tests (Koehn, 2004).

3.2 Main Results

We first compare our approach with previous MT KD studies on the widely used WMT 14 En-De task with 6-layer Transformer Base setting for students. Results are shown in Table 1.

Table 1 shows that: 1) our method can lead to $+1.67$ BLEU improvements over the strong Transformer baseline on the WMT 14 En-De task, and 2) our method can effectively lead to slightly higher BLEU scores than previous machine translation knowledge distillation baselines efficiently without the necessity to explicitly pre-train a teacher model beforehand.

3.3 Generality on Translation Tasks

To test the effectiveness of our approach with varying training set sizes and across different lan-

Metric	Model	IWSLT 14 De-En	WMT 14 En-De	WMT 15 Cs-En	WMT 14 En-Fr
BLEU	Transformer	31.41	27.54	29.13	39.87
	+ Self KD (Ours)	**32.10**†	**29.21**†	**30.25**†	**41.78**†
chrF	Transformer	53.84	57.19	54.93	63.34
	+ Self KD (Ours)	**54.42**†	**58.04**†	**55.62**†	**64.57**†
COMET	Transformer	79.39	82.40	81.90	82.81
	+ Self KD (Ours)	**79.86**†	**83.65**†	**82.62**†	**84.60**†

Table 2: Results with varying training set sizes. † indicates $p < 0.01$ in the significance test.

ID	Models	#Para	Direction	BLEU_{94}	WR	BLEU_4
1	Transformer	110M	En→xx xx→En	18.75 27.02	-	14.73 22.50
2	1 + LALN + LALT (Zhang et al., 2020)	133M	En→xx xx→En	20.81 27.22	-	17.45 23.30
3	2 + depth-wise LSTM (Xu et al., 2024)	148M	En→xx xx→En	23.38 28.41	ref	20.47 26.68
4	2 + Self KD (Ours)	133M	En→xx xx→En	24.11 29.03	82.98 80.85	21.01 27.78

Table 3: Multilingual translation results on the OPUS-100 dataset.

guage pairs, we conducted experiments on the low-resource IWSLT 14 De-En, middle-resource WMT 14 En-De, and high-resource WMT 15 Cs-En and WMT 14 En-Fr tasks. Results evaluated with BLEU, chrF and COMET scores are shown in Table 2.

Table 2 shows that our method can improve the performance of the low-resource IWSLT 14 De-En task by $+0.69$ BLEU, $+0.58$ chrF, and $+0.47$ COMET scores, the middle-resource WMT 14 En-De task by $+1.67$ BLEU, $+0.85$ chrF, and $+1.25$ COMET scores, the high-resource WMT 15 Cs-En and WMT 14 En-Fr tasks by $+1.12$ and $+1.91$ BLEU, $+0.69$ and $+1.23$ chrF, and $+0.72$ and $+1.79$ COMET scores respectively, showing the effectiveness of our approach in low/middle/high-resource scenarios.

Despite the improvements with our self-distillation method on the low-resource IWSLT 14 De-En task is smaller than on the middle and high-resource tasks, the improvements are still significant in all metrics. The smaller improvements on the low-resource task might be because that knowledge distillation is to make better use of the model capacity, while the training set of the low-resource task may have difficulty in providing sufficient knowledge to benefit from this.

3.4 Effectiveness for Multilingual Translation

Multilingual translation uses a single model to translate between multiple language pairs (Firat et al., 2016; Johnson et al., 2017; Aharoni et al., 2019). Model capacity has been found crucial for massively multilingual NMT to support language pairs with varying typological characteristics (Zhang et al., 2020; Xu et al., 2021a). Effective model capacity utilization with our self-distillation method is likely to benefit multilingual NMT.

To test the effectiveness of our self-distillation method in the multilingual setting, we conducted experiments on the challenging massively many-to-many translation task on the OPUS-100 corpus (Tiedemann, 2012; Aharoni et al., 2019; Zhang et al., 2020). We tested the performance of 6-layer models following the experiment settings of Zhang et al. (2020) for fair comparison. We adopted BLEU (Papineni et al., 2002) for translation evaluation with the SacreBLEU toolkit (Post, 2018). [1] We report average BLEU over 94 language pairs BLEU_{94}, win ratio WR (%) compared to a strong baseline which effectively improves the model capacity through the use of depth-wise LSTMs (Xu et al., 2024), average BLEU over 4 typologically different target languages with varied training data sizes (de, zh, br and te) BLEU_4 selected by Zhang et al. (2020). Results are shown in Table 3.

Table 3 shows that: 1) our approach can improve the backbone model from Zhang et al. (2020) by $+3.30$ and $+1.81$ BLEU on average in the En→xx and xx→En directions respectively in the evalu-

[1] BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a +version.1.4.1

Depth Encoder	Depth Decoder	Transformer	Self KD (Ours)
3		26.51	**27.73**
6		27.54	**29.21**
12	6	28.62	**30.00**
18		29.00	**30.49**

Table 4: Results of Transformers with varying depths on the WMT 14 En-De task.

ID	Model	En-De	En-Fr
1	Transformer	27.54	39.87
2	1 + MHPLSTM Decoder	28.37	40.31
3	2 + Self KD Dec	29.11	41.64
4	2 + Self KD Full	**29.81**	**42.14**

Table 5: BLEU scores with MHPLSTM decoder on the WMT 14 tasks.

ID	Model	dev	test
1	Transformer	25.42	27.54
2	1 + Self KD (Pearson loss)	25.58	27.79
3	2 + warm up	25.71	28.00
4	3 + iterative KD	25.84	28.32
5	4 + gradient adaptation	26.19	28.84
6	5 + rank-order loss	**26.48**	**29.21**

Table 6: Ablation study on the WMT 14 En-De task.

ation over 94 different languages. The one-to-many translation task can be regarded as multi-task learning and its performance is likely to be constrained by the model capacity, while the many-to-one translation task can be regarded as joint training and is likely to benefit from knowledge transfer (Arivazhagan et al., 2019). The larger average BLEU_{94} improvements in the En→xx direction (+3.30) than in the xx→En direction (+1.81) over Zhang et al. (2020) demonstrate the effectiveness of our approach in utilizing the model capacity. 2) compared to improving the capacity using depth-wise LSTMs, our self-distillation method leads to +0.73 and +0.62 BLEU improvements over Xu et al. (2024) on average when translating English to 94 languages and translating them into English respectively, and obtains win ratios of 82.98% and 80.85% in the En→xx and xx→En directions respectively without introducing additional parameters.

3.5 Effectiveness with Varying Depths

We tested the effectiveness of our method in more challenging settings on the WMT 14 En-De task by increasing the encoder depth to 12 and 18 layers. We also evaluated the performance of a shallower Transformer model with 3 encoder and decoder layers to test the efficiency and effectiveness of our approach with smaller models. Results are shown in Table 4.

Table 4 shows that: 1) our method can obtain consistent improvements with both shallower and deeper models, and 2) the 3-layer Transformer with our method efficiently obtains a higher BLEU score (27.73) than the 6-layer baseline (27.54), the 6-layer model with self-distillation (29.21) already outperforms the baseline with 18 encoder layers (29.00), and our method can obtain comparable or higher BLEU scores efficiently with fewer layers.

3.6 Effectiveness with Recurrent Decoder

Chen et al. (2018); Xu et al. (2021b) show that enhanced recurrent decoders may lead to improved translation quality while being faster in decoding than the Transformer decoder due to the $O(n)$ complexity of the recurrent decoder. We tested the performance of our approach with the MHPLSTM decoder (Xu et al., 2021b) on the WMT 14 En-De and En-Fr tasks. We adopted the same settings as the Transformer Base model but replaced the Transformer decoder by a 6-layer MHPLSTM decoder. Results are shown in Table 5.

Table 5 shows that: 1) applying our self-distillation method to the MHPLSTM decoder only can also lead to consistent improvements over the stronger baseline, and 2) applying self-distillation to both the encoder and the decoder brings about better performance than to the MHPLSTM decoder only, suggesting that self-distillation is still effective for the self-attentional Transformer encoder when using the recurrent MHPLSTM decoder.

4 Analysis

Most of our analyses were conducted on the WMT 14 En-De task, with newstest 2013 and 2014 as the development set and test set respectively.

4.1 Ablation Study

We tested the effectiveness of individual methods on the WMT 14 En-De task. Results are shown in Table 6.

Table 6 shows that: 1) distilling the deepest layer to all shallow layers with the negative Pearson correlation loss can obtain slight improvements, 2) the warm-up strategy and iterative knowledge distillation method can lead to further improvements over

Layer(s) without Self KD		dev BLEU	dev Δ	test BLEU	test Δ
All (Baseline)		25.42	0.00	27.54	0.00
None (Ours Full)		26.48	1.06	29.21	1.67
Encoder	1	26.27	0.85	28.72	1.18
	2	26.28	0.86	28.81	1.27
	3	26.18	0.76	28.76	1.22
	4	26.15	0.73	28.81	1.27
	5	26.31	0.89	28.85	1.31
Decoder	1	26.40	0.99	29.03	1.49
	2	26.36	0.95	28.89	1.36
	3	26.20	0.78	28.68	1.14
	4	26.24	0.82	28.72	1.18
	5	26.18	0.76	28.81	1.27

Table 7: Results of removing individual layers from self-distillation on the WMT 14 En-De task. Δ indicates the improvements over the baseline in BLEU.

Layer	Transformer Encoder	Transformer Decoder	+ Self KD Encoder	+ Self KD Decoder
1	42.26	23.50	42.87	24.50
2	43.85	34.36	45.08	35.04
3	45.20	44.08	46.58	51.50
4	46.26	61.17	47.41	62.78
5	47.29	68.03	48.14	68.18
6	47.44	70.39	48.59	71.66

Table 8: Word translation accuracy on the WMT 14 En-De test set.

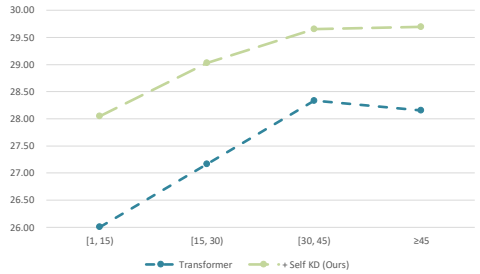

Figure 3: BLEU scores w.r.t. various input sentence length in tokens on the WMT 14 En-De test set.

vanilla self-distillation with the Pearson loss, 3) gradient adaptation and rank-order loss bring about more improvements than the other methods, and 4) all our methods can consistently lead to further improvements on both the validation and test sets and all methods together lead to $+1.06$ and $+1.67$ BLEU improvements on the development set and test set of the WMT 14 En-De task respectively.

4.2 Effects of each Layer in Self-Distillation

We studied the contribution of self-distillation to each encoder/decoder layer on performance on the WMT 14 En-De task. If the kth layer did not involve in self-distillation, the iterative mechanism would distill the output of layer $k+1$ to that of layer $k-1$. Results are shown in Table 7.

Table 7 shows that removing each encoder/decoder layer from self-distillation leads to worse performance than distilling to all encoder and decoder layers. This on one hand shows the usefulness of individual encoder/decoder layers in self-distillation, on the other hand suggests that our method can simply distill to all layers without the requirement to search for an optimal subset of shallower sub-networks for self-distillation.

The effects of independently applying self-distillation to the encoder/decoder are studied in Appendix C.

4.3 Effects on Word Translation Accuracy

To analyze the effects of self-distillation training on the evolution of token translations across Transformer layers, we adopted the probing method of Xu et al. (2021c). Specifically, we frozen the parameters of the Transformer models trained on the WMT 14 En-De task, and projected the representations of the tested layer to the frozen trained decoder classifier with a linear layer to measure word translation accuracy. The linear layer for projection is randomly initialized and trained again on the training set with the other model parameters frozen. When testing encoder layers, the alignment matrices from decoder cross-attention layers are aggregated into a single alignment matrix with a softmax-normalized weight parameter vector. The aggregated alignment matrix is used to transform the source encoding representation to align with the target sequence through matrix multiplication, and the weight parameter vector is trained together with the linear projection layer. The word translation accuracy is measured on the test set of the WMT 14 En-De task. Results are shown in Table 8.

Table 8 shows that training with self-distillation improves the word translation accuracy of all encoder and decoder layers, including not only all shallow layers acting as students sub-networks in the self-distillation but also the last layer serving as only the teacher. This suggests that self-distillation not only improves the student sub-networks (shallow layers), but also assists the last layer which only serves as a teacher to obtain better performance based on the improved representations from shallow layers.

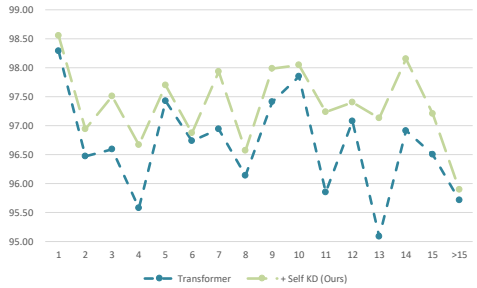

Figure 4: Subject-verb agreement accuracy on the Lingeval97 dataset. X-axis and y-axis represent subject-verb distance in words and accuracy respectively.

4.4 Length Analysis

We tested the impacts of self-distillation on the translation performance with various input lengths.

We first grouped sentences of similar lengths together and computed BLEU scores for each length group on the WMT 14 En-De test set following Bahdanau et al. (2015); Tu et al. (2016); Xu et al. (2020b, 2021b). Results are shown in Figure 3.

Figure 3 shows that: 1) self-distillation improves the performance for all length groups, 2) the performance on shorter sentences ($<$15 source input tokens) gets more improvements ($+$2.04 BLEU) with self-distillation, and 3) unlike the Transformer baseline which obtains lower BLEU scores on the length group with $\geq$45 source input tokens than the $[30, 45)$ length group, the BLEU scores on the $\geq$45 length group is slightly higher than the $[30, 45)$ length group with self-distillation.

We also measured the effects of self-distillation on capturing dependencies of various distances with the linguistically-informed verb-subject agreement analysis on the *Lingeval97* dataset (Sennrich, 2017). In German, subjects and verbs must agree with one another in grammatical number and person. In *Lingeval97*, each contrastive translation pair consists of a correct reference translation, and a contrastive example that has been minimally modified to introduce one translation error. The accuracy of a model is the number of times it assigns a higher score to the reference translation than to the contrastive one, relative to the total number of predictions. We used the models trained on the WMT 14 En-De task for evaluation. Results are shown in Figure 4.

Figure 4 shows similar accuracy curve trends for the baseline and the self-distillation method, but the accuracies with self-distillation are generally higher than the Transformer baseline and larger accuracy gains are obtained on both some short distances (3 and 4) and long distances (11, 13 and 14).

5 Related Work

Knowledge Distillation for Efficient Machine Translation Kim and Rush (2016) show that standard knowledge distillation applied to word-level prediction can be effective for neural machine translation, and also introduce sequence-level knowledge distillation that further improve performance. Wu et al. (2020) inject layer-level supervision from the teacher model to the student model. Wang et al. (2021) analyze the different impacts of samples and present batch-level and global-level selection methods to pick suitable samples for distillation. Jafari et al. (2021) propose an improved knowledge distillation method by feeding the rich information provided by teacher's soft-targets incrementally and more efficiently. Liang et al. (2022) efficiently obtain multiple teachers via sub-layer reordering, layer-drop, and dropout variants for multi-teacher knowledge distillation. Zhang et al. (2023) show that the knowledge comes from the top-1 predictions of teachers and design a hierarchical ranking loss to enforce the learning of the top-1 information from the teacher. Miao et al. (2023) randomly extract fewer-layer sub-networks from the teacher and jointly optimize the teacher and these students enhanced by mutual learning.

Translation Knowledge Transfer via Knowledge Distillation Zeng et al. (2019) iteratively perform translation knowledge transfer between in-domain and out-of-domain models via knowledge distillation. Wei et al. (2019) generate a teacher model from checkpoints to guide the training process. Zhang et al. (2019b) propose a future-aware knowledge distillation framework to distill future knowledge from a backward neural language model to future-aware vectors which are incorporated into the attention layer of the decoder. Chen et al. (2020) utilize extra supervision from BERT to improve the conventional sequence-to-sequence translation model. Baziotis et al. (2020) transfer the target language knowledge from the language model to the low-resource neural machine translation models. Liang et al. (2021) design latent variational modules to learn the distributions of bilingual conversational characteristics and incorporate into the translation model via knowledge distillation. Feng et al. (2021) introduce a seer decoder involving future information in target predictions and force the

conventional decoder to simulate the behaviors of the seer decoder via knowledge distillation. Zhou et al. (2022) first jointly train the NMT model with an auxiliary conditional masked language model and then incorporate bidirectional global context to the NMT model on less confident predictions via knowledge distillation. Lu et al. (2022) explore multi-stage information interactions for multi-source NMT at the encoding stage. Yang et al. (2022) distill knowledge retrieved by kNN to encourage the NMT model to take more reasonable target tokens into consideration. Wang et al. (2024) propose a domain-aware kNN-KD method to filter out domain-relevant neighborhood knowledge for learning in the distillation process. Li et al. (2024) transfer knowledge from LLMs to existing MT models in a selective, comprehensive and proactive manner.

Knowledge Distillation for Multilingual Neural Machine Translation Tan et al. (2019) train the multilingual model to fit the training data and match the outputs of individual models simultaneously through knowledge distillation. Huang et al. (2022c) pick up language-specific best checkpoints for each language pair to teach the current model on the fly. Huang et al. (2023) collaboratively train two Pareto optimal solutions that favor different languages and allows them to learn from the strengths of each other via knowledge distillation. Do and Lee (2023) introduce a language-family-based approach to select appropriate knowledge for each language pair, and use target-oriented knowledge distillation which intensively focuses on the ground-truth target of knowledge with a penalty strategy.

Knowledge Distillation for Non-Autoregressive Translation Sequence-level knowledge distillation helps produce more deterministic training sets for non-autoregressive translation (Gu et al., 2018, 2019; Qian et al., 2021; Wang et al., 2023). Zhou et al. (2020) find that knowledge distillation can reduce the complexity of data sets and help NAT to model the variations in the output data. Huang et al. (2022a) augment the training of the non-autoregressive Transformer with deep supervision and additional layer-wise predictions. Liu et al. (2023) introduce selective knowledge distillation by introducing an NAT evaluator to select NAT-friendly targets that are of high quality and easy to learn.

Knowledge Distillation for Unsupervised Translation Sun et al. (2020) leverage knowledge distillation methods to further enhance multilingual unsupervised neural machine translation. Nguyen et al. (2021) present cross-model back-translated distillation to induce another level of data diversification.

Compared to previous studies on efficient neural machine translation with knowledge distillation, our study does not require to explicitly train or leverage a teacher model, and can also obtain significant improvements (Table 1).

6 Conclusion

In this paper, we study efficient representation-based self-distillation methods for neural machine translation. We present an iterative distillation method to prevent potential adverse impacts of knowledge distillation due to the capacity gap between teachers and students, propose a gradient adaptation method and adopt a warm-up strategy to ensure that the knowledge distillation gradients are consistent with the machine translation gradients, and present a rank-order augmented Pearson correlation loss to further ensure the effectiveness of representation-based knowledge distillation.

Experiments on low/middle/high-resource tasks show that our method can lead to significant improvements over the strong baselines, obtaining comparable performance to previous machine translation knowledge distillation studies without pre-training a teacher model. Experiments with shallower and deeper Transformers show that our method can obtain better performance efficiently with fewer layers. Our method is also effective in the challenging multilingual translation setting or with recurrent decoder.

Limitations

We only apply our method to the encoder and decoder stacks independently, without taking interactions between the encoder and the decoder (e.g., distilling decoder layers to encoder layers or vice versa) and mutual learning into consideration.

Due to resource limitation, we did not apply our method to large-scale pre-training, LLM translation or evaluate on the other natural language processing tasks.

Acknowledgments

We thank anonymous reviewers for their insightful comments. This work is partially supported by the National Natural Science Foundation of China (Grant No. 62306284), China Postdoctoral Science Foundation (Grant No. 2023M743189), and the Natural Science Foundation of Henan Province (Grant No. 232300421386).

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Ali Araabi and Christof Monz. 2020. Optimizing transformer for low-resource neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3429–3435, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George F. Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *CoRR*, abs/1907.05019.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ankur Bapna, Mia Chen, Orhan Firat, Yuan Cao, and Yonghui Wu. 2018. Training deeper neural machine translation models with transparent attention. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3028–3033. Association for Computational Linguistics.

Christos Baziotis, Barry Haddow, and Alexandra Birch. 2020. Language model prior for low-resource neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7622–7634, Online. Association for Computational Linguistics.

Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Mike Schuster, Noam Shazeer, Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2018. The best of both worlds: Combining recent advances in neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–86, Melbourne, Australia. Association for Computational Linguistics.

Yen-Chun Chen, Zhe Gan, Yu Cheng, Jingzhou Liu, and Jingjing Liu. 2020. Distilling knowledge learned in BERT for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7893–7905, Online. Association for Computational Linguistics.

Jang Hyun Cho and Bharath Hariharan. 2019. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Heejin Do and Gary Geunbae Lee. 2023. Target-oriented knowledge distillation with language-family-based grouping for multilingual nmt. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(2).

Yang Feng, Shuhao Gu, Dengji Guo, Zhengxin Yang, and Chenze Shao. 2021. Guiding teacher forcing with seer forcing for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2862–2872, Online. Association for Computational Linguistics.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252, International Convention Centre, Sydney, Australia. PMLR.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O.K. Li, and Richard Socher. 2018. Non-autoregressive neural machine translation. In *International Conference on Learning Representations*.

Jiatao Gu, Changhan Wang, and Junbo Zhao. 2019. Levenshtein transformer. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Wenjie Hao, Hongfei Xu, Lingling Mu, and Hongying Zan. 2022. Optimizing deep transformers for chinese-thai low-resource translation. In *Machine Translation*, pages 117–126, Singapore. Springer Nature Singapore.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *Preprint*, arXiv:1503.02531.

Chenyang Huang, Hao Zhou, Osmar R. Zaïane, Lili Mou, and Lei Li. 2022a. Non-autoregressive translation with layer-wise prediction and deep supervision. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10776–10784.

Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. 2022b. Knowledge distillation from a stronger teacher. In *Advances in Neural Information Processing Systems*, volume 35, pages 33716–33727. Curran Associates, Inc.

Yichong Huang, Xiaocheng Feng, Xinwei Geng, Baohang Li, and Bing Qin. 2023. Towards higher Pareto frontier in multilingual machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3802–3818, Toronto, Canada. Association for Computational Linguistics.

Yichong Huang, Xiaocheng Feng, Xinwei Geng, and Bing Qin. 2022c. Unifying the convergences in multilingual neural machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6822–6835, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aref Jafari, Mehdi Rezagholizadeh, Pranav Sharma, and Ali Ghodsi. 2021. Annealing knowledge distillation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2493–2504, Online. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Jungo Kasai, Nikolaos Pappas, Hao Peng, James Cross, and Noah Smith. 2021. Deep encoder, shallow decoder: Reevaluating non-autoregressive machine translation. In *International Conference on Learning Representations*.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, pages 1–15.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*.

Bei Li, Quan Du, Tao Zhou, Yi Jing, Shuhan Zhou, Xin Zeng, Tong Xiao, JingBo Zhu, Xuebo Liu, and Min Zhang. 2022. ODE transformer: An ordinary differential equation-inspired model for sequence generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8335–8351, Dublin, Ireland. Association for Computational Linguistics.

Jiahuan Li, Shanbo Cheng, Shujian Huang, and Jiajun Chen. 2024. MT-PATCHER: Selective and extendable knowledge distillation from large language models for machine translation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6445–6459, Mexico City, Mexico. Association for Computational Linguistics.

Xiaobo Liang, Lijun Wu, Juntao Li, Tao Qin, Min Zhang, and Tie-Yan Liu. 2022. Multi-teacher distillation with single model for neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:992–1002.

Yunlong Liang, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021. Modeling bilingual conversational characteristics for neural chat translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5711–5724, Online. Association for Computational Linguistics.

Min Liu, Yu Bao, Chengqi Zhao, and Shujian Huang. 2023. Selective knowledge distillation for non-autoregressive neural machine translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13246–13254.

Ziyao Lu, Xiang Li, Yang Liu, Chulun Zhou, Jianwei Cui, Bin Wang, Min Zhang, and Jinsong Su. 2022. Exploring multi-stage information interactions for multi-source neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:562–570.

Zhongjian Miao, Xiang Li, Liyan Kang, Wen Zhang, Chulun Zhou, Yidong Chen, Bin Wang, Min Zhang, and Jinsong Su. 2022. Towards robust neural machine translation with iterative scheduled data-switch training. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5266–5277, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Zhongjian Miao, Wen Zhang, Jinsong Su, Xiang Li, Jian Luan, Yidong Chen, Bin Wang, and Min Zhang. 2023. Exploring all-in-one knowledge distillation framework for neural machine translation. In *Proceedings*

of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 2929–2940, Singapore. Association for Computational Linguistics.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. 2020. Improved knowledge distillation via teacher assistant. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):5191–5198.

Xuan-Phi Nguyen, Shafiq Joty, Thanh-Tung Nguyen, Kui Wu, and Ai Ti Aw. 2021. Cross-model back-translated distillation for unsupervised machine translation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8073–8083. PMLR.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Lihua Qian, Hao Zhou, Yu Bao, Mingxuan Wang, Lin Qiu, Weinan Zhang, Yong Yu, and Lei Li. 2021. Glancing transformer for non-autoregressive neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1993–2003, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Rico Sennrich. 2017. How grammatical is character-level neural machine translation? assessing MT quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382, Valencia, Spain. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Wonchul Son, Jaemin Na, Junyong Choi, and Wonjun Hwang. 2021. Densely guided knowledge distillation using multiple teacher assistants. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9395–9404.

Haipeng Sun, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2020. Knowledge distillation for multilingual unsupervised neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3525–3535, Online. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. 2016. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826.

Xu Tan, Yi Ren, Di He, Tao Qin, and Tie-Yan Liu. 2019. Multilingual neural machine translation with knowledge distillation. In *International Conference on Learning Representations*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–85, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Fusheng Wang, Jianhao Yan, Fandong Meng, and Jie Zhou. 2021. Selective knowledge distillation for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6456–6466, Online. Association for Computational Linguistics.

Mingxuan Wang, Zhengdong Lu, Jie Zhou, and Qun Liu. 2017. Deep neural machine translation with linear associative unit. In *Proceedings of the 55th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 136–145, Vancouver, Canada. Association for Computational Linguistics.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong, and Lidia S. Chao. 2019. Learning deep transformer models for machine translation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 1810–1822, Florence, Italy. Association for Computational Linguistics.

Zhexuan Wang, Shudong Liu, Xuebo Liu, Miao Zhang, Derek Wong, and Min Zhang. 2024. Domain-aware k-nearest-neighbor knowledge distillation for machine translation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9458–9469, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Zhihao Wang, Longyue Wang, Jinsong Su, Junfeng Yao, and Zhaopeng Tu. 2023. Revisiting non-autoregressive translation at scale. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12051–12065, Toronto, Canada. Association for Computational Linguistics.

Hao-Ran Wei, Shujian Huang, Ran Wang, Xin-yu Dai, and Jiajun Chen. 2019. Online distilling from checkpoints for neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1932–1941, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiangpeng Wei, Heng Yu, Yue Hu, Yue Zhang, Rongxiang Weng, and Weihua Luo. 2020. Multiscale collaborative deep models for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 414–426, Online. Association for Computational Linguistics.

Lijun Wu, Yiren Wang, Yingce Xia, Fei Tian, Fei Gao, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2019. Depth growing for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5558–5563, Florence, Italy. Association for Computational Linguistics.

Yimeng Wu, Peyman Passban, Mehdi Rezagholizadeh, and Qun 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)*, pages 1016–1021, Online. Association for Computational Linguistics.

Hongfei Xu, Qiuhui Liu, Josef van Genabith, and Deyi Xiong. 2021a. Modeling task-aware MIMO cardinality for efficient multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 361–367, Online. Association for Computational Linguistics.

Hongfei Xu, Qiuhui Liu, Josef van Genabith, Deyi Xiong, and Jingyi Zhang. 2020a. Lipschitz constrained parameter initialization for deep transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 397–402, Online. Association for Computational Linguistics.

Hongfei Xu, Qiuhui Liu, Josef van Genabith, Deyi Xiong, and Meng Zhang. 2021b. Multi-head highly parallelized LSTM decoder for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 273–282, Online. Association for Computational Linguistics.

Hongfei Xu, Yang Song, Qiuhui Liu, Josef van Genabith, and Deyi Xiong. 2024. Rewiring the transformer with depth-wise LSTMs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14122–14133, Torino, Italia. ELRA and ICCL.

Hongfei Xu, Josef van Genabith, Qiuhui Liu, and Deyi Xiong. 2021c. 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*, pages 74–85, Online. Association for Computational Linguistics.

Hongfei Xu, Josef van Genabith, Deyi Xiong, Qiuhui Liu, and Jingyi Zhang. 2020b. Learning source phrase representations for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 386–396, Online. Association for Computational Linguistics.

Zhixian Yang, Renliang Sun, and Xiaojun Wan. 2022. Nearest neighbor knowledge distillation for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5546–5556, Seattle, United States. Association for Computational Linguistics.

Jiali Zeng, Yang Liu, Jinsong Su, Yubing Ge, Yaojie Lu, Yongjing Yin, and Jiebo Luo. 2019. Iterative dual domain adaptation for neural machine translation. 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)*, pages 845–855, Hong Kong, China. Association for Computational Linguistics.

Biao Zhang, Ivan Titov, and Rico Sennrich. 2019a. Improving deep transformer with depth-scaled initialization and merged attention. 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)*, pages 898–909, Hong Kong, China. Association for Computational Linguistics.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Biao Zhang, Deyi Xiong, Jinsong Su, and Jiebo Luo. 2019b. Future-aware knowledge distillation for neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(12):2278–2287.

Linfeng Zhang, Chenglong Bao, and Kaisheng Ma. 2022. Self-distillation: Towards efficient and compact neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4388–4403.

Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. 2019c. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Songming Zhang, Yunlong Liang, Shuaibo Wang, Yufeng Chen, Wenjuan Han, Jian Liu, and Jinan Xu. 2023. Towards understanding and improving knowledge distillation for neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8062–8079, Toronto, Canada. Association for Computational Linguistics.

Chulun Zhou, Fandong Meng, Jie Zhou, Min Zhang, Hongji Wang, and Jinsong Su. 2022. Confidence based bidirectional global context aware training framework for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2878–2889, Dublin, Ireland. Association for Computational Linguistics.

Chunting Zhou, Jiatao Gu, and Graham Neubig. 2020. Understanding knowledge distillation in non-autoregressive machine translation. In *International Conference on Learning Representations*.

Jie Zhou, Ying Cao, Xuguang Wang, Peng Li, and Wei Xu. 2016. Deep recurrent models with fast-forward connections for neural machine translation. *Transactions of the Association for Computational Linguistics*, 4:371–383.

α	β	dev	test
0.5		26.28	28.97
0.7		26.32	29.11
0.9		26.51	29.25
1.0	1.0	26.48	29.21
1.1		26.40	29.08
1.3		26.31	29.04
1.5		26.21	29.00
	0.7	26.35	29.20
	0.8	26.46	29.22
	0.9	26.57	29.44
1.0	1.0	26.48	29.21
	1.1	26.28	29.11
	1.2	26.22	29.10
	1.3	26.16	29.03

Table 9: Results of different hyper-parameters on the WMT 14 En-De task.

A Hyper-parameters

For the middle-resource and high-resource translation tasks, we followed the Transformer Base settings of Vaswani et al. (2017) except for $200k$ training steps for all experiments (including all baselines) following Zhang et al. (2023). Specifically, we used 6 encoder and decoder layers, an embedding size of 512, a feed-forward layer of 2048 hidden units, a dropout of 0.1. The dimension of each head was 64 and each attention layer had 8 attention heads. We used a shared vocabulary for each task and tied the encoder-decoder embeddings. We employed a label smoothing (Szegedy et al., 2016) value of 0.1. We used the Adam optimizer (Kingma and Ba, 2015) with 0.9, 0.98 and 10^{-9} as β_1, β_2 and ϵ. For the low-resource IWSLT 14 De-En task, we followed the experiment settings of Araabi and Monz (2020) which lead to a strong baseline for the low-resource setting.

B Effects of Hyper-Parameters for Loss Accumulation

In all our other experiments, we simply sum up individual losses rather than carefully tuning the hyper-parameters for the accumulation of knowledge distillation losses and the translation loss, because this is more friendly for practice than tuning hyper-parameters for each specific task. But carefully tuning the hyper-parameters for loss aggregation may lead to better performance. In such case, the final loss ($loss$) for optimization is the weighted sum of the translation loss ($loss_{mt}$) and the knowledge distillation loss ($loss_{kd}$) with α as the weight for $loss_{kd}$.

ID	Model	dev	test
1	Transformer	25.42	27.54
2	1 + Self KD Enc	26.26	28.72
3	1 + Self KD Dec	25.93	28.40
4	1 + Self KD Full	**26.48**	**29.21**

Table 10: Effects of encoder/decoder self-KD on the WMT 14 En-De task.

$$loss = loss_{mt} + \alpha * loss_{kd} \quad (3)$$

The knowledge distillation loss ($loss_{kd}$) is the weighted sum of the Pearson coefficient loss ($loss_{Pearson}$) and the rank-order based distillation loss ($loss_{rank}$) with β as the weight for $loss_{rank}$.

$$loss_{kd} = loss_{Pearson} + \beta * loss_{rank} \quad (4)$$

We tested the performance of various α and β values. Results are shown in Table 9.

Table 9 shows that carefully tuning the hyper-parameters can lead to better performance than simply summing up all losses, but the performance gap between different settings is limited, and a wide range of hyper-parameter selections lead to good performance. This may partially be because the gradient adaption mechanism zero the knowledge distillation gradients which are in contradiction with machine translation. In such case, we suggest that simply summing up the losses is a reasonable choice for our method for the ease of practice.

C Effects of Encoder and Decoder

We tested the effects of separately applying our self-distillation method to the encoder and the decoder on the WMT 14 En-De task. Results are shown in Table 10.

Table 10 shows that: 1) the encoder leads to more improvements ($+0.84$ and $+1.18$ BLEU on the development set and the test set respectively) than the decoder ($+0.51$ and $+0.86$ BLEU correspondingly) with our method, and 2) applying our method to both the encoder and the decoder obtains better performance than only applying the method to either the encoder or the decoder, this suggests that the effectiveness of our method on the encoder and the decoder are complementary.