Overcoming Catastrophic Forgetting beyond Continual Learning: Balanced Training for Neural Machine Translation

Chenze Shao[1,2], **Yang Feng**[1,2*]
[1] Key Laboratory of Intelligent Information Processing
Institute of Computing Technology, Chinese Academy of Sciences (ICT/CAS)
[2] University of Chinese Academy of Sciences
{shaochenze18z, fengyang}@ict.ac.cn

Abstract

Neural networks tend to gradually forget the previously learned knowledge when learning multiple tasks sequentially from dynamic data distributions. This problem is called *catastrophic forgetting*, which is a fundamental challenge in the continual learning of neural networks. In this work, we observe that catastrophic forgetting not only occurs in continual learning but also affects the traditional static training. Neural networks, especially neural machine translation models, suffer from catastrophic forgetting even if they learn from a static training set. To be specific, the final model pays imbalanced attention to training samples, where recently exposed samples attract more attention than earlier samples. The underlying cause is that training samples do not get balanced training in each model update, so we name this problem *imbalanced training*. To alleviate this problem, we propose Complementary Online Knowledge Distillation (COKD), which uses dynamically updated teacher models trained on specific data orders to iteratively provide complementary knowledge to the student model. Experimental results on multiple machine translation tasks show that our method successfully alleviates the problem of imbalanced training and achieves substantial improvements over strong baseline systems.[1]

1 Introduction

Neural Machine Translation (NMT) has achieved impressive translation performance on many benchmark datasets in the past few years (Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017). In the domain adaptation task where we have large-scale out-domain data to improve the in-domain translation performance, continual learning, which is also referred to as fine-tuning, is often employed to transfer the out-domain knowledge to in-domain (Luong and Manning, 2015). After fine-tuning, the model performs well in in-domain translation, but there is significant performance degradation in out-domain translation because it “forgets” the previously learned knowledge. This phenomenon is called catastrophic forgetting (McCloskey and Cohen, 1989; French, 1999) and has attracted a lot of attention (Goodfellow et al., 2013; Kirkpatrick et al., 2017; Li and Hoiem, 2017; Lee et al., 2017).

In this work, we observe that catastrophic forgetting not only occurs in continual learning but also affects the traditional static training. To be specific, the final model pays imbalanced attention to training samples. At the end of training, the recently exposed samples attract more attention and tend to have lower losses, while earlier samples are partially forgotten by the model and have higher losses. In short, training samples receive imbalanced attention from the model, which mainly depends on the time when the model last saw the training sample (i.e., the data order of the last training epoch).

The underlying cause of this phenomenon is mini-batch gradient descent (LeCun et al., 2012), that is, we do not simultaneously use all training samples to train the model but divide them into mini-batches. Therefore, training samples do not get balanced training in each update step, so we name this problem *imbalanced training*. This problem is less severe in some tasks (e.g., image classification and text classification), but it has a significant impact on NMT as machine translation is a challenging task containing numerous translation rules, which are easily forgotten during the training process. Besides, we find that the imbalanced training problem is especially severe and non-negligible on low-resource machine translation.

To demonstrate that the imbalanced training problem does affect the model accuracy, we first review a widely used technique called checkpoint averaging technique, which has proved to be effec-

[*] Corresponding author: Yang Feng
[1] Code is available at https://github.com/ictnlp/COKD.

tive in improving model accuracy but its internal mechanisms are not fully understood. We analyze it from the perspective of catastrophic forgetting and find that their success can be attributed to the alleviation of imbalanced training. We also notice that checkpoint averaging has some limitations, leaving room for further improvements.

Inspired by the existing solution of checkpoint averaging which leverages the complementarity of checkpoints to improve model accuracy, we propose Complementary Online Knowledge Distillation (COKD) to address the problem of imbalanced training. As the model tends to forget knowledge learned from early samples, the main idea of COKD is to construct complementary teachers to re-provide this forgotten knowledge to the student. Specifically, we divide the training set into mutually exclusive subsets and reorganize them in a specific orders to train the student and teachers. We perform COKD in an online manner where teachers are on-the-fly updated to fit the need of student. When training the student on a subset, teachers can always provide the student with complementary knowledge on the other subsets, thereby preventing the student from catastrophic forgetting.

Experimental results on multiple machine translation tasks show that our method successfully alleviates the problem of imbalanced training and achieves substantial improvements over strong baseline systems. Especially, on the low-resource translation tasks that are severely affected by imbalanced training, our method is particularly effective and improves baseline models by about 2 BLEU points on average.

In summary, our contribution is threefold:

- We observe the problem of *imbalanced training* that training samples receive imbalanced attention from the model. We find that NMT, especially low-resource translation tasks, is seriously affected by imbalanced training.
- We rethink the widely used checkpoint averaging technique and explain its success from the perspective of imbalanced training, which also demonstrates that the imbalanced training problem does affect the model accuracy.
- We propose Complementary Online Knowledge Distillation for NMT, which can successfully alleviate the imbalanced training problem and improve the translation quality.

2 Background

2.1 Knowledge Distillation

Knowledge distillation (Hinton et al., 2015) is a class of methods that transfers knowledge from a pre-trained teacher network to a student network. Assume that we are training a classifier $p(y|x;\theta)$ with $|\mathcal{V}|$ classes, and we can access the pre-trained teacher $q(y|x)$. Instead of minimizing the cross-entropy loss between the ground-truth label and the model output probability, knowledge distillation uses the teacher model prediction $q(y|x)$ as a soft target and minimizes the loss:

$$\mathcal{L}_{\text{KD}}(\theta) = -\sum_{k=1}^{|\mathcal{V}|} q(y=k|x) \times \log p(y=k|x;\theta). \tag{1}$$

In neural machine translation, the standard training objective is the cross-entropy loss, which minimizes the negative log-likelihood as follows:

$$\mathcal{L}_{\text{NLL}}(\theta) = -\sum_{t=1}^{T} \log(p(y_t|y_{<t}, \boldsymbol{X}, \theta)), \tag{2}$$

where $\boldsymbol{X} = \{x_1, ..., x_N\}$ and $\boldsymbol{Y} = \{y_1, ..., y_T\}$ are the source sentence and the target sentence, respectively. Kim and Rush (2016) proposed to train the student model to mimic the teacher's prediction at each decoding step, which is called Word-level Knowledge Distillation (Word-KD) and its loss is calculated as follows:

$$\begin{aligned}\mathcal{L}_{\text{Word-KD}}(\theta) = -\sum_{t=1}^{T}\sum_{k=1}^{|\mathcal{V}|} q(y_t = k|y_{<t}, \boldsymbol{X}) \times \\ \log p(y_t = k|y_{<t}, \boldsymbol{X}, \theta).\end{aligned} \tag{3}$$

Conventional offline knowledge distillation only allows the student to learn from static pre-trained teacher models. On the contrary, online knowledge distillation trains teachers from scratch and dynamically updates them, so the student learns from different teachers during the training process. Zhang et al. (2018) first overcame the offline limitation by training peer models simultaneously and conducted an online distillation in one-phase training between peer models. Since mutual learning requires training multiple networks, Lan et al. (2018); Song and Chai (2018) proposed to use a single multi-branch network for online knowledge distillation, which treats each branch as a student and the ensemble of branches as a teacher. The multi-branch architecture subsequently became the mainstream for online knowledge distillation (Guo et al., 2020; Chen et al., 2020; Wu and Gong, 2021). Besides, Furlanello et al. (2018) performed iterative self-distillation where the student network is identical to the teacher in terms of the network graph. In each new iteration, under the supervision of the earlier iteration, a new identical model is trained from scratch. In NMT, Wei et al. (2019) on-the-fly selected the best checkpoint from the training path as the teacher to guide the training process.

2.2 Catastrophic Forgetting

Catastrophic forgetting is a problem faced by many machine learning models during continual learning, as models tend to forget previously learned knowledge when being trained on new tasks (McCloskey and Cohen, 1989). A typical class of methods to mitigate catastrophic forgetting is based on regularization which constrains the update of model parameters. Goodfellow et al. (2013) empirically find that the dropout regularization can effectively alleviate the catastrophic forgetting phenomenon. Kirkpatrick et al. (2017) proposed elastic weight consolidation, which implements the modified regularization term that imposes constraints on the update of important parameters in the previous task. Lee et al. (2017) proposed drop-transfer, which is a variant of dropout that drops the weight vector of turned off nodes to the weight learned on the previous task instead of a zero vector. Learning without Forgetting (LWF) (Li and Hoiem, 2017) is the approach most relevant to our work. They only use new task data to train the network but preserve the original capabilities by distilling knowledge from the pre-trained model.

There are also a number of efforts to address the catastrophic forgetting problem for the domain adaptation of NMT. Kirkpatrick et al. (2017); Thompson et al. (2019) added regularization terms to constrain the update of parameters. Dakwale and Monz (2017) proposed to minimize the KL-divergence between the predictions of general-domain model and fine-tuned model. Zeng et al. (2018); Gu et al. (2019) introduced a discriminator to preserve the domain-shared features. Liang et al. (2021); Gu et al. (2021); Xie et al. (2021) fixed important parameters during the fine-tuning to preserve the general-domain performance. Gu and Feng (2020) investigated the cause of catastrophic forgetting from the perspectives of modules and parameters.

3 Imblanced Training

Before drawing any conclusions, we first conduct experiments on three different tasks, namely, image classification, text classification, and machine translation, to show that the problem of imbalanced training does exist. For image classification, we conduct experiments on CIFAR-10 and CIFAR-100 (Krizhevsky, 2009), both of which contain 50,000/10,000 training/testing images with 32 × 32 pixels drawn from 10/100 classes. For text classification, we conduct experiments on AG-News, which contains 120,000/7,600 training/testing sentences drawn from 4 classes. For machine translation, we conduct experiments on three translation tasks: WMT14 English-German (En-De), IWSLT15 English-Vietnamese (En-Vi), and WMT17 English-Turkish (En-Tr). We use the ResNet-32 network (He et al., 2016) for image classification, the VDCNN network (Conneau et al., 2017) for text classification and Transformer-base (Vaswani et al., 2017) for machine translation. All the models are trained using cross-entropy loss. We refer readers to Appendix A and section 6.1 for the detailed configurations.

We train the model until convergence and then take the last checkpoint to calculate losses of training samples in the data order of the last training epoch. If there is a problem of imbalanced training, then training samples at the end of the epoch, which are recently exposed to the model, will tend to have lower losses. In contrast, training samples at the beginning will tend to have higher losses.

For quantitative analysis, we use the Spearman correlation coefficient between the data order and loss to measure the degree of imbalanced training. Specifically, we assign each batch in the training dataset with a batch-id according to the order they appear in the last training epoch, where batch i is the i-th trained batch. We disable regularization techniques such as dropout and label smoothing and calculate the loss for each batch. The correlation coefficient between the batch-id and the loss is used to measure the degree of imbalanced training, and a large negative correlation coefficient indicates that this problem is severe. Figure 1 illustrates the relationship between the batch-id and loss. By comparing the loss curves and correlation coefficients on these six datasets, we obtain the following three main observations.

The problem of imbalanced training does exist. Among the six datasets in our experiments, only

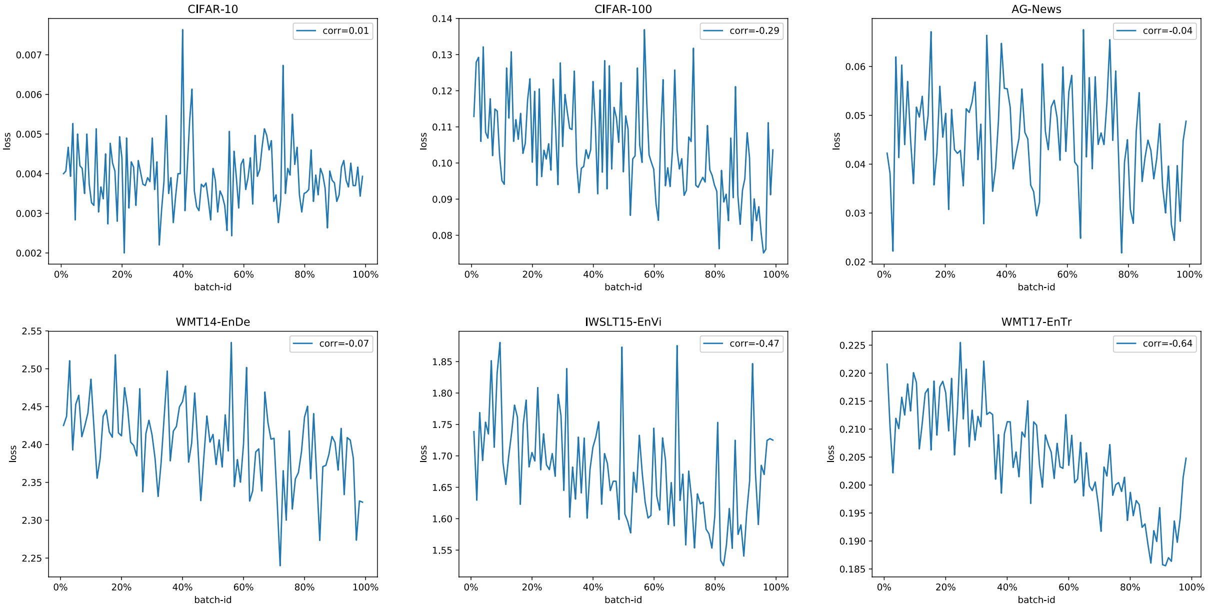

Figure 1: The relationship between the batch-id and loss on three different types of tasks. The Spearman correlation coefficient (corr) is presented in the upper right corner of charts. Batch-id i indicates the i-th trained batch in the last epoch. Batch-id in the x-axis is normalized to [0,1]. Image Classification: CIFAR-10 and CIFAR-100; Text Classification: AG-News; Machine Translation: WMT14 En-De, IWSLT15 En-Vi, and WMT17 En-Tr.

CIFAR-10 has a positive correlation coefficient. Two datasets (i.e., AG-News and WMT14 En-De) have small negative correlation coefficients. Three datasets (i.e., CIFAR-100, IWSLT15 En-Vi, and WMT17 En-Tr) have an apparent decline in losses accompanied by large negative correlation coefficients. Therefore, we can conclude that the problem of imbalanced training does exist, but the degree of impact varies.

Imbalanced training is related to task complexity. Intuitively, imbalanced training is more likely to occur on complex tasks where previously learned knowledge may be easily forgotten during the learning of numerous new knowledge. Comparing the two image classification datasets, CIFAR-10 and CIFAR-100 have the same dataset size but a different number of classes. The correlation coefficient on the complex task CIFAR-100 is -0.29, while the correlation coefficient on CIFAR-10 is 0.01. The text classification task, which only contains 4 classes, has a small correlation coefficient -0.04. Machine translation is generally considered a complex task with exponential search space and numerous translation rules. Notably, WMT17 En-Tr has the largest correlation coefficient of -0.64. These results are consistent with our intuition that imbalanced training has a greater impact on complex tasks like machine translation.

Low-resource translation suffers from imbalanced training. Comparing the three machine translation datasets, the imbalanced training problem has a much larger impact on low-resource datasets (i.e., IWSLT15 En-Vi and WMT17 En-Tr), where the high-resource dataset WMT14 En-De is less affected. To eliminate the influence of language, we randomly select 100K sentences from the WMT14 En-De dataset for the training to simulate the low-resource scenario. We show the loss curve in Appendix B, where the corresponding correlation coefficient is -0.63, which also supports the conclusion. This is counter-intuitive since when there are many training samples, the early samples seem to be more easily forgotten. Actually, as Figure 1 shows, the loss curves are generally less steep at the beginning, indicating that early samples are nearly "equally forgotten" by the model. For high-resource datasets, most samples are nearly "equally forgotten" and only the losses of a few samples at the end are highly correlated with the batch-id, so the overall correlation is low. In comparison, nearly the whole loss curve of low-resource datasets is steep, so the model may simultaneously overfit recent samples and underfit early samples due to imbalanced training. Therefore, the problem of imbalanced training is more serious and nonnegligible in low-resource machine translation.

Loss rises in the end due to the momentum of optimizer. On CIFAR-100, IWSLT15 En-Vi, and WMT17 En-Tr, though their loss curves are generally downward, they all have a sudden rise at the end. This abnormal phenomenon is actually consistent with our conclusion. Because of the momentum factor in the adam optimizer, the impact of a model update is not limited to the current step. The optimizer retains the gradient in the form of momentum, which will affect the gradient updates in the next few steps. Therefore, the impact of momentum is not fully released in the last few training steps, so the loss rises in the end.

4 Checkpoint Averaging

Checkpoint averaging, which directly takes the average of parameters of the last few checkpoints as the final model, is a widely used technique in NMT (Junczys-Dowmunt et al., 2016; Vaswani et al., 2017). The averaged checkpoint generally performs better than any single checkpoint. However, to the best of our knowledge, its internal mechanism is not fully understood.

In this section, we analyze the success of checkpoint averaging from the perspective of imbalanced training. Though training samples receive imbalanced attention from each checkpoint, this imbalance is different among checkpoints. If we understand the imbalanced training as the noise on each checkpoint, noises among different checkpoints can be approximately regraded as i.i.d. random variables. By averaging checkpoints, the variance of random noise is reduced and thereby alleviating the problem of imbalanced training. Based on the above analysis, we make the following hypothesis and verify it through experiments.

Hypothesis *Checkpoint averaging improves the model performance through alleviating the problem of imbalanced training.*

Experiments We conduct experiments on the six datasets to study the relationship between checkpoint averaging and imbalanced training. We average the last five epoch checkpoints and compare their performance with the best single checkpoint. Table 1 reports the model performance along with the correlation coefficient on the six datasets. We can see that checkpoint averaging achieves considerable improvements on datasets where the problem of imbalanced training is severe. On datasets with small correlation coefficients, the improvements of checkpoint averaging are very limited. These results confirm our hypothesis and also demonstrate that the imbalanced training problem does affect the model accuracy.

Dataset	Corr	Best	Ave
Image Classification & Text Classification			
CIFAR-10	0.01	93.51%	93.47%
CIFAR-100	-0.29	70.89%	71.36%
AG-News	-0.04	91.61%	91.70%
Machine Translation			
WMT14 En-De	-0.07	27.29	27.45
IWSLT15 En-Vi	-0.47	28.52	29.08
WMT17 En-Tr	-0.64	12.79	13.42

Table 1: Model performance on the test sets of six datasets. For classification tasks, we report the Top-1 accuracy. For translation tasks, we report the BLEU score. **Corr** is the Spearman correlation coefficient calculated in section 3. **Best** and **Ave** represent the best and average checkpoint performance, respectively.

Limitations Though checkpoint averaging can alleviate the problem of imbalanced training and improve the model performance, it also has some limitations and its success largely depends on the empirical choice of checkpoint interval. If the checkpoint interval is small, then the i.i.d. assumption does not hold, so the imbalance cannot be effectively eliminated and may even become stronger (Appendix C). If the checkpoint interval is large, then checkpoints may not lie in the same parameter space, making the direct averaging of checkpoints problematic.

5 Approach

In this section, we propose Complementary Online Knowledge Distillation (COKD) to alleviate the problem of imbalanced training. We apply knowledge distillation with dynamically updated complementary teachers to re-provide the forgotten knowledge to the student model.

5.1 Complementary Teachers

We first introduce the construction of complementary teachers. Assume that we have n teacher models $\mathcal{T}_{1:n}$ and the student model is $\mathcal{S}$, and both teacher models and the student model are randomly initialized. We expect that teacher models should be dynamically updated so that they are always complementary to the student. While the student learns from new training samples and gradually forgets early samples, teacher models should reprovide the forgotten knowledge to the student.

Recall that the model pays imbalanced attention to different training samples depending on the data order of the training. Therefore, a natural way to obtain complementary teachers is to train teachers in different data orders. Specifically, in each epoch, we divide the training dataset $\mathcal{D}$ into $n+1$ mutually exclusive splits $(\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_{n+1})$. The student model sequentially learns from $\mathcal{D}_1$ to $\mathcal{D}_{n+1}$, where the data order is different for teacher models.

We use a ordering function $\mathcal{O}(i,t)$ to denote the training data for teacher $\mathcal{T}_i$ at time t. After teacher models $\mathcal{T}_{1:n}$ learn from data splits $\mathcal{D}_{\mathcal{O}(1:n,t)}$ respectively, the student $\mathcal{S}$ learn from both $\mathcal{D}_t$ and teachers. To make teachers complementary with the student, the ordering function $\mathcal{O}(\cdot,t)$ should cover all data splits except $\mathcal{D}_t$. To ensure that each teacher can access the whole training data, the ordering function $\mathcal{O}(i,\cdot)$ should also cover all data splits. Fortunately, we find that a simple assignment of $\mathcal{O}$ satisfies the above requirements:

$$\mathcal{O}(i,t) = \begin{cases} i+t, & i+t \leq n+1 \\ i+t-n-1, & i+t > n+1 \end{cases}. \tag{4}$$

where $i \in \{1,2,...,n\}$ and $t \in \{1,2,...,n+1\}$. Under this assignment, teacher $\mathcal{T}_i$ simply uses the data split that has offset i from the student, which ensures that all teachers are complementary with the student and can access the whole training set.

5.2 Complementary Training

The knowledge of n complementary teachers can be transfered to the student through word-level knowledge distillation:

$$\mathcal{L}_{\text{KD}}(\theta) = -\sum_{t=1}^{T}\sum_{k=1}^{|\mathcal{V}|}\sum_{i=1}^{n} \frac{q_i(y_t=k|y_{<t},\boldsymbol{X})}{n} \times \log p(y_t=k|y_{<t},\boldsymbol{X},\theta), \tag{5}$$

where p is the prediction of student $\mathcal{S}$ and q_i is the prediction of teacher $\mathcal{T}_i$. We use a hyperparameter α to interpolate the distillation loss and the cross-entropy loss:

$$\mathcal{L}(\theta) = \alpha \cdot \mathcal{L}_{\text{KD}}(\theta) + (1-\alpha) \cdot \mathcal{L}_{\text{NLL}}(\theta). \tag{6}$$

Algorithm 1 COKD

```
Input: training set $\mathcal{D}$, the number of teachers $n$
Output: student model $\mathcal{S}$
randomly initialize student $\mathcal{S}$ and teachers $\mathcal{T}_{1:n}$
while not converge do
    randomly divide $\mathcal{D}$ into $n+1$ subsets $(\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_{n+1})$
    for $t = 1$ to $n+1$ do
        for $i = 1$ to $n$ do
            train $\mathcal{T}_i$ on $\mathcal{D}_{\mathcal{O}(i,t)}$
        train $\mathcal{S}$ on $\mathcal{D}_t$ according to Equation 6
    for $i = 1$ to $n$ do $\mathcal{T}_i \leftarrow \mathcal{S}$
return student model $\mathcal{S}$
```

In this way, the student model learns both new knowledge from the training set and complementary knowledge from teacher models. With an appropriate α, we can achieve a balance between the two kinds of knowledge and alleviate the problem of imbalanced training. However, this method is based on knowledge distillation where knowledge is transferred unidirectionally from teachers to the student. Though the student can benefit from balanced training, these complementary teachers also set an upperbound to the student and prevent it from performing better.

To overcome this limitation, we follow the underlying idea of two-way knowledge transfer where the knowledge is also transferred from the student to teachers (Zhang et al., 2018; Lan et al., 2018). We use a simple reinitialization method to achieve the two-way knowledge transfer. At the end of each epoch, we reinitialize teacher models with the parameters of the student model:

$$\mathcal{T}_i \leftarrow \mathcal{S}, \quad i \in \{1,2,...,n\}. \tag{7}$$

Through the reinitialization, the student and teachers are exactly the same at the beginning of each epoch. In this way, both the student and teachers are iteratively improved so the student performance is no longer limited by the fixed ability of teachers. We summarize the training process of COKD in Algorithm 1.

6 Experiments

6.1 Setup

To evaluate the performance of COKD, we conduct experiments on multiple machine translation tasks. For low-resource translation where the problem of imbalanced training is severe, we run experiments on WMT17 English-Turkish (En-Tr, 207K sentence pairs), IWSLT15 English-Vietnamese (En-Vi, 133K sentence pairs), and TED bilingual dataset. We also evaluate the high-resource performance of COKD on WMT14 English-German (En-De, 4.5M sentence pairs). For WMT17 En-Tr and IWSLT15 En-Vi, we use case-sensitive SacreBLEU (Post, 2018) to report reproducible BLEU scores. For TED bilingual dataset, following Xu et al. (2021), we report the tokenized BLEU. For WMT14 En-De translation, we report the tokenized BLEU (Papineni et al., 2002) with compound split.

For WMT17 En-Tr, we use *newstest2016* as the validation set and *newstest2017* as the test set. We learn a joint BPE model (Sennrich et al., 2016) with 16K operations. For IWSLT15 En-Vi, we use the pre-processed data used in Luong and Manning (2015)[2]. For TED bilingual dataset, we use the pre-processed data used in Xu et al. (2021)[3]. For WMT14 En-De, the validation set is *newstest2013* and the test set is *newstest2014*. We learn a joint BPE model with 32K operations.

In the main experiments, we set the number of teachers n to 1 and the hyperparameter α to 0.95. We implemented our approach based on the base version of Transformer (Vaswani et al., 2017). Following Wei et al. (2019), we increase the dropout rate to 0.2 on WMT17 En-Tr and IWSLT15 En-Vi. For TED bilingual dataset, we further increase the dropout rate of Transformer baseline to 0.3. All models are optimized with Adam (Kingma and Ba, 2014) with the optimizer settings in Vaswani et al. (2017). The batch size is 32K for all translation tasks. For inference, we average the last 5 checkpoints and use beam search with beam size 5. The checkpoint interval is 1000 for low-resource tasks and 5000 for WMT14 En-De.

6.2 Main Results

We first conduct experiments on the two low-resource datasets WMT17 En-Tr and IWSLT15 En-Vi and the high-resource dataset WMT14 En-De to evaluate the capability of our method. We compare our method with knowledge distillation methods and deep mutual learning (Zhang et al., 2018), and also report the results of Online Distillation from Checkpoints (ODC) (Wei et al., 2019) for comparison. The results are listed in Table 2.

Models	En-Tr	En-Vi	En-De
Transformer*	12.20	28.56	–
ODC*	12.92	29.47	–
Transformer	13.42	29.08	27.45
Word-KD	13.66	29.54	27.76
Seq-KD	13.91	29.69	27.84
Mutual	13.72	29.83	27.81
COKD	**16.66**	**31.95**	**28.26**

Table 2: BLEU scores on three translation tasks. * means results reported in Wei et al. (2019). Word-KD means word-level knowledge distillation, and Seq-KD means sequence-level knowledge distillation. Mutual means our reimplementation of deep mutual learning (Zhang et al., 2018).

Low-Resource First, we focus on the results on the two low-resource datasets where the problem of imbalanced training is severe. Since we have applied the checkpoint averaging technique on the baseline system, our baseline is very competitive and outperforms the baseline of Wei et al. (2019). We refer readers to Appendix D for results without checkpoint averaging. Knowledge distillation techniques and deep mutual learning bring some improvements to the baseline, but the improvements are relatively weak. In comparison, COKD substantially improves the baseline performance by about 3 BLEU scores, demonstrating the effectiveness of COKD on low-resource translation tasks.

High-Resource On the high-resource dataset WMT14 En-De, COKD still outperforms the baseline and knowledge distillation methods. The improvement of COKD is relatively small compared to the low-resource setting, which can be explained from the perspective of imbalanced training. As illustrated in Figure 1, high-resource datasets like WMT14 En-De is less affected by the problem of imbalanced training, so the alleviation of this problem may not bring strong improvements on high-resource datasets.

TED Bilingual Dataset We further conduct experiments on TED bilingual dataset to confirm the effectiveness of COKD on low-resource translation tasks. We evaluate COKD on both En-X and X-En directions and report the results in Table 3. The performance of COKD is still very impressive, which improves the baseline by 1.59 BLEU on average in the En-X direction, and improves the baseline by 2.15 BLEU on average in the En-X direction.

6.3 Ablation Study

In this section, we study the effect of complementary teachers and teacher reinitialization in COKD. We remove each of them respectively and report

[2] https://github.com/tefan-it/nmt-en-vi

[3] https://github.com/Jingjing-NLP/VOLT

En-X	Es	PTbr	Fr	Ru	He	Ar	It	Nl	Ro	Tr	De	Vi	Ave
Base	40.86	40.31	41.27	21.86	29.01	18.40	36.37	34.06	28.33	17.70	31.46	29.66	30.77
COKD	42.50	42.46	43.15	22.94	30.22	19.36	37.78	35.87	29.70	19.50	33.48	31.33	32.36
X-En	Es	PTbr	Fr	Ru	He	Ar	It	Nl	Ro	Tr	De	Vi	Ave
Base	42.94	45.52	41.32	26.21	38.78	33.06	39.55	37.52	36.50	27.19	36.89	27.64	36.09
COKD	44.72	47.84	43.33	27.87	40.81	35.03	41.48	39.66	38.78	29.68	39.73	29.91	38.24

Table 3: BLEU scores on the TED bilingual dataset. Ave means the average BLEU.

Models	**En-Tr**	**En-Vi**	**En-De**
COKD	16.66	31.95	28.26
- CT	15.83	31.56	27.96
- TR	14.02	29.93	27.84

Table 4: Ablation study for COKD. CT means complementary teachers. TR means teacher reinitialization.

n	0	1	2	3	4
BLEU	12.57	15.43	15.35	15.66	15.47
Time	1.34h	2.87h	4.56h	6.15h	7.83h

Table 5: BLEU scores of COKD on the validation set of WMT17 En-Tr. The training time is measured with 8 NVIDIA RTX 2080Ti. h is the abbreviation of hour.

α	0.75	0.9	0.95	0.98	1
BLEU	13.86	14.94	15.43	15.31	15.23

Table 6: BLEU scores of COKD with different α on the validation set of WMT17 En-Tr.

their performance in Table 4. By removing complementary teachers, we do not split the dataset and assign random data order to teachers, which leads to obvious performance degradtion. We also notice that a large part of improvement comes from the reinitialization, suggesting the importance of two-way knowledge transfer where both the student and teachers are iteratively improved.

6.4 Hyperparameters

There are two hyperparameters in COKD: the number of teachers n and the loss weight α, whose default settings are $n = 1$, $\alpha = 0.95$ in the main experiments. In this section, we conduct experiments on WMT17 En-Tr to show the effect of the two hyperparameters.

The number of teachers We change the number of teachers n from 1 to 5 to evaluate the effect of n in COKD and report the BLEU score and training time in Table 5. We find that using more teachers does not necessarily lead to better performance, suggesting that the main improvement is not due to the ensemble of multiple teachers. Large n may slightly outperform the $n = 1$ setting but comes with a larger training cost. Therefore, we recommend the $n = 1$ setting in practical applications. Though the training cost is still larger, it is acceptable on low-resource datasets considering the strong performance improvement.

Hyperparameter α We set the hyperparameter α to $0.75, 0.9, 0.95, 0.98, 1$ respectively. The corresponding BLEU scores are listed in Table 6. We can see that the model performance is sensitive to the hyperparameter α. Generally, the model prefers large α, where a slightly smaller α may significantly degrade the model performance. We explain this phenomenon as the imbalance of complementary knowledge and new knowledge. The distillation loss carries the complementary knowledge and the cross-entropy loss carries the new knowledge, so an appropriate α should balance the two kinds of knowledge. Considering that the distillation loss is only a little biased to the complementary knowledge, α should be much larger than 0.5, otherwise it cannot keep the balance. We empirically recommend the $\alpha = 0.95$ setting, which also shows good performance on other datasets.

6.5 COKD Alleviates Imbalanced Training

In this section, we evaluate the effectiveness of COKD in alleviating the problem of imbalanced training. We take the final model of COKD and measure the correlation between batch-id and loss in the last epoch. We conduct experiments on the WMT17 En-Tr dataset where the problem of imbalanced training is severe. As Figure 2 shows, the downward trend of loss is successfully alleviated by COKD, and the correlation coefficient is improved from -0.64 to -0.16.

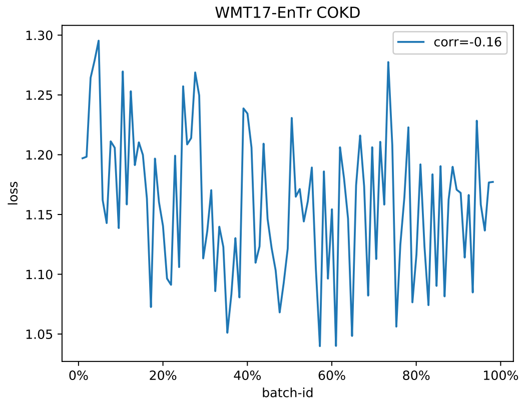

Figure 2: The loss curve of COKD on the last epoch.

7 Conclusion

In this paper, we observe that catastrophic forgetting will cause imbalanced training, which is severe in low-resource machine translation and will affect the translation quality. We rethink the checkpoint averaging technique and explain its success from the perspective of imbalanced training. We further propose Complementary Online Knowledge Distillation (COKD), which successfully alleviates the imbalanced training problem and achieves substantial improvements in translation quality.

8 Acknowledgement

We thank the anonymous reviewers for their insightful comments. This work was supported by National Key R&D Program of China (NO.2017YFE9132900).

References

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*.

Defang Chen, Jian-Ping Mei, Can Wang, Yan Feng, and Chun Chen. 2020. Online knowledge distillation with diverse peers. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 3430–3437. AAAI Press.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Alexis Conneau, Holger Schwenk, Loïc Barrault, and Yann Lecun. 2017. Very deep convolutional networks for text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1107–1116, Valencia, Spain. Association for Computational Linguistics.

Praveen Dakwale and Christof Monz. 2017. Finetuning for neural machine translation with limited degradation across in-and out-of-domain data. *Proceedings of the XVI Machine Translation Summit*, 117.

Robert M French. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135.

Tommaso Furlanello, Zachary Chase Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. 2018. Born-again neural networks. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1602–1611. PMLR.

Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. 2013. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*.

Shuhao Gu and Yang Feng. 2020. Investigating catastrophic forgetting during continual training for neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4315–4326, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Shuhao Gu, Yang Feng, and Qun Liu. 2019. Improving domain adaptation translation with domain invariant and specific information. 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 3081–3091, Minneapolis, Minnesota. Association for Computational Linguistics.

Shuhao Gu, Yang Feng, and Wanying Xie. 2021. Pruning-then-expanding model for domain adaptation of neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3942–3952, Online. Association for Computational Linguistics.

Qiushan Guo, Xinjiang Wang, Yichao Wu, Zhipeng Yu, Ding Liang, Xiaolin Hu, and Ping Luo. 2020. Online knowledge distillation via collaborative learning. In *2020 IEEE/CVF Conference on Computer*

Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, pages 11017–11026. Computer Vision Foundation / IEEE.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

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

Marcin Junczys-Dowmunt, Tomasz Dwojak, and Rico Sennrich. 2016. The AMU-UEDIN submission to the WMT16 news translation task: Attention-based NMT models as feature functions in phrase-based SMT. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 319–325, Berlin, Germany. Association for Computational Linguistics.

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. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

A Krizhevsky. 2009. Learning multiple layers of features from tiny images. *Master's thesis, University of Tront*.

Xu Lan, Xiatian Zhu, and Shaogang Gong. 2018. Knowledge distillation by on-the-fly native ensemble. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 7528–7538.

Yann A LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. 2012. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer.

Sang-Woo Lee, Jin-Hwa Kim, Jaehyun Jun, Jung-Woo Ha, and Byoung-Tak Zhang. 2017. Overcoming catastrophic forgetting by incremental moment matching. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4652–4662.

Zhizhong Li and Derek Hoiem. 2017. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947.

Jianze Liang, Chengqi Zhao, Mingxuan Wang, Xipeng Qiu, and Lei Li. 2021. Finding sparse structures for domain specific neural machine translation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Virtual Event, February 2-9, 2021*, pages 13333–13342. AAAI Press.

Minh-Thang Luong and Christopher D Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 76–79.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

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 on association for computational linguistics*, pages 311–318. 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.

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.

Guocong Song and Wei Chai. 2018. Collaborative learning for deep neural networks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 1837–1846.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Brian Thompson, Jeremy Gwinnup, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019. Overcoming catastrophic forgetting during domain adaptation of 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 2062–2068, Minneapolis, Minnesota. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

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.

Guile Wu and Shaogang Gong. 2021. Peer collaborative learning for online knowledge distillation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 10302–10310. AAAI Press.

Wanying Xie, Yang Feng, Shuhao Gu, and Dong Yu. 2021. Importance-based neuron allocation for 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 1: Long Papers)*, pages 5725–5737, Online. Association for Computational Linguistics.

Jingjing Xu, Hao Zhou, Chun Gan, Zaixiang Zheng, and Lei Li. 2021. Vocabulary learning via optimal transport 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 7361–7373, Online. Association for Computational Linguistics.

Jiali Zeng, Jinsong Su, Huating Wen, Yang Liu, Jun Xie, Yongjing Yin, and Jianqiang Zhao. 2018. Multi-domain neural machine translation with word-level domain context discrimination. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 447–457, Brussels, Belgium. Association for Computational Linguistics.

Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. 2018. Deep mutual learning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4320–4328. Computer Vision Foundation / IEEE Computer Society.

A Configurations for Classification Tasks

In this section, we describe the configurations of classification tasks in detail. For the two image classification tasks CIFAR-10 and CIFAR-100, our implementation is based on the source code released by Chen et al. (2020)[4]. We use the ResNet-32 network (He et al., 2016). For preprocessing, we normalized all images by channel means and standard deviations. We use stochastic gradient descent with Nesterov momentum for optimization and set the initial learning rate to 0.1, momentum to 0.9. We set the mini-batch size to 128 and weight decay to 5×10-4. The learning rate is divided by 10 at 150 and 225 of the total 300 training epochs for these two datasets.

For the text classification task AG-News, our implementation is based on an open-source NLP benchmark[5]. We use the VDCNN network (Conneau et al., 2017) with depth 29. We use stochastic gradient descent with momentum for optimization and set the initial learning rate to 0.01, momentum to 0.9. We train the model for 100 epochs and multiply the learning rate by 0.9 every 15 steps. We set the mini-batch size to 128.

B Simulation of Low-Resource WMT14 En-De

We randomly select 100K sentences from the WMT14 En-De dataset for the training to simulate the low-resource scenario. Figure 3 shows that the loss curve is downward and has a large negative correlation coefficient -0.63. Comparing with the whole dataset result, it confirms the conclusion that low-resource translation suffers more from imbalanced training.

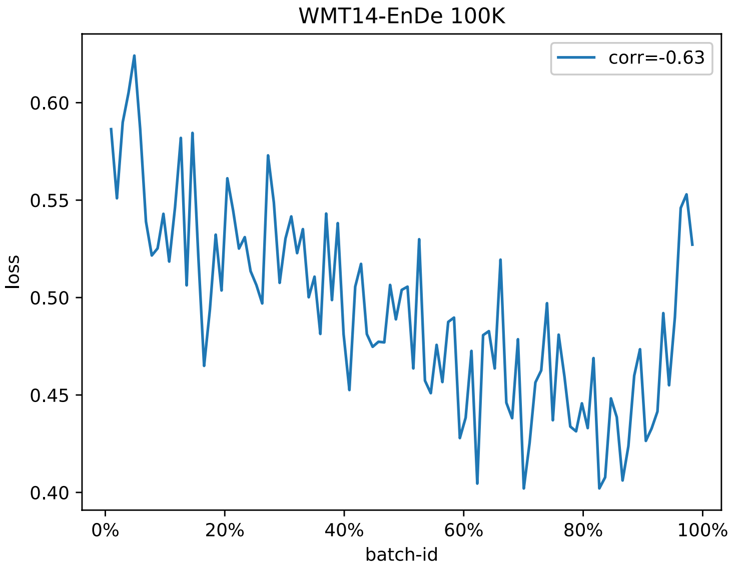

Figure 3: Relationship between the batch-id and loss on the simulated low-resource scenario of WMT14 En-De.

C Case Analysis of Checkpoint Averaging

In this section, we give a case analysis to show that checkpoint averaging may bring new imbalance and degrade the model performance. We conduct experiments on the validation set of WMT17 En-Tr translation. We set the checkpoint interval to 0.1 epoch, so 10 checkpoints are saved in the last training epoch. We average these 10 checkpoints and illustrate the relationship between the batch-id and loss in Figure 4.

Different from the previous imbalance, Figure 4 shows an upward trend with a large positive correlation coefficient of 0.81. The BLEU score of the averaged model is 12.71, which is even lower than the BLEU of a single checkpoint. Intuitively, this is because earlier samples are recently exposed to many

[4] https://github.com/DefangChen/OKDDip-AAAI2020

[5] https://github.com/ArdalanM/nlp-benchmarks

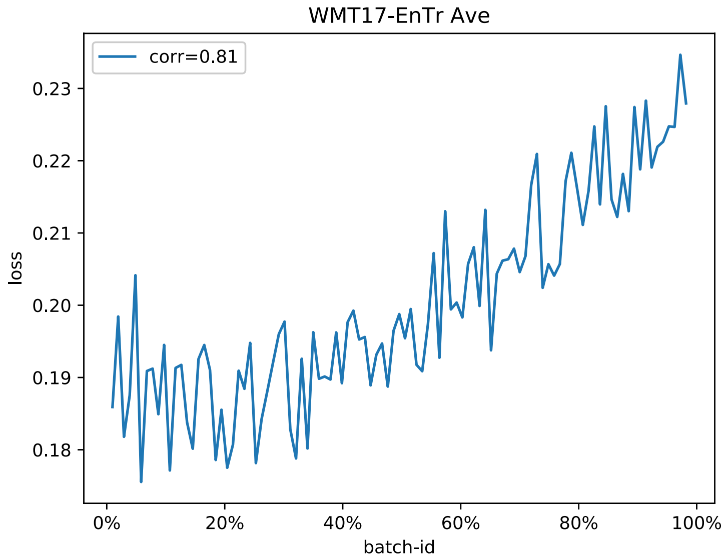

Figure 4: Relationship between the batch-id and loss on WMT17 En-Tr.

checkpoints, while the latter samples are only exposed to the last few checkpoints. The underlying cause is the small interval of checkpoints. As the checkpoint interval is small, the i.i.d. assumption does not hold, so checkpoint averaging cannot eliminate the imbalance and may bring a new imbalance.

D Performance without Checkpoint Averaging

In the main experiments, we still apply the checkpoint averaging technique during inference to alleviate the imbalanced training problem. Here we report the performance without checkpoint averaging in Table 7.

Models	En-Tr			En-Vi			En-De		
	Best	Ave	Δ	Best	Ave	Δ	Best	Ave	Δ
Transformer	12.79	13.42	+0.63	28.52	29.08	+0.56	27.29	27.45	+0.16
Word-KD	13.25	13.66	+0.41	29.09	29.54	+0.45	27.64	27.76	+0.12
Seq-KD	13.37	13.91	+0.54	29.26	29.69	+0.43	27.65	27.84	+0.19
Mutual	13.13	13.72	+0.59	29.35	29.83	+0.48	27.59	27.81	+0.22
COKD	16.61	16.66	+0.05	31.86	31.95	+0.09	28.18	28.26	+0.08

Table 7: BLEU scores on three translation dateset. **Best** and **Ave** represent the best and average checkpoint performance, respectively.

We can see that the performance of COKD only decreases a little after removing checkpoint averaging, indicating that COKD itself can successfully alleviate the problem of imbalanced training. In contrast, the performance gap between Best and Ave is larger on other NMT Systems, suggesting that the problem of imbalanced training is severe and cannot be simply mitigated by other techniques like knowledge distillation and mutual learning.

E Fine-Tuning with Lower Learning Rate

When the training is finished, we can manually reduce the learning rate to fine-tune the model, which also improves the model accuracy in some situations. The common explanation is that fine-tuning can help to converge the optimization process and reduce the loss function. Here we analyze fine-tuning from another perspective and find that it can also alleviate the imbalanced training problem.

Intuitively, if the learning rate is very low, the model will not be much biased towards recent training samples, which should reduce the imbalance. We conduct experiments on WMT17 En-Tr to confirm our hypothesis. We reduce the learning rate from the base setting of $7 \cdot 10^{-4}$ to $1 \cdot 10^{-5}$ to fine-tune the model and draw the loss curve in Figure 5. From both the correlation coefficient and loss curve, we can see that the imbalance is greatly reduced by fine-tuning. Regrading the BLEU score, the BLEU score after fine-tuning is only 0.25 higher than the baseline, which is much smaller than the improvement of checkpoint averaging. We speculate that it is due to some drawbacks of fine-tuning. For example, fine-tuning with a low learning rate has a risk of overfitting the dataset, which may influence its performance on low-resource datasets.

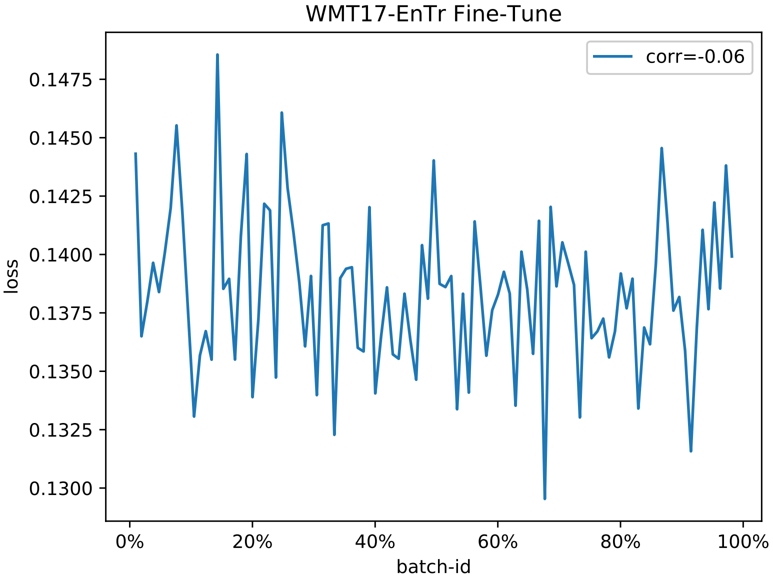

Figure 5: Relationship between the batch-id and loss on WMT17 En-Tr.