ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240767

LoCa: Logit Calibration for Knowledge Distillation

Runming Yang[1], Taiqiang Wu[2] and Yujiu Yang[1,*]

[1]Tsinghua University
[2]The University of Hong Kong

Abstract. Knowledge Distillation (KD), aiming to train a better student model by mimicking the teacher model, plays an important role in model compression. One typical way is to align the output logits. However, we find a common issue named mis-instruction, that the student would be misled when the predictions based on teacher logits do not follow the labels. Meanwhile, there is other useful dark knowledge in the logits such as the class discriminability, which is vital for distillation. In this paper, we propose a simple yet effective **Lo**git **Ca**libration (LoCa) method, which calibrates the logits from the teacher model based on the ground-truth labels. The key insight is to correct the prediction (to address the mis-instruction issue) and maintain useful dark knowledge simultaneously. Our proposed LoCa does not require any additional parameters. Empirical results on image classification and text generation tasks demonstrate that LoCa can effectively improve the performance of baselines.

1 Introduction

In the last decade, the development of deep neural networks has revolutionized the field of computer vision (CV) [6, 11, 22] and natural language processing (NLP) [19]. Complex network architectures [20, 43] and increasing parameter [4] can make a stronger model but also bring high costs in computation and deployment [24, 7]. Such costs are not preferable when applying models to industrial scenarios, and thus researchers have made many efforts to compress the models. One mainstream approach to designing lightweight models is knowledge distillation (KD) [8, 34], which concentrates on transferring the knowledge from a heavy model (i.e. teacher) to a light one (i.e. student).

The goal of KD is to train a better student by mimicking the teacher [15, 16]. For example, the logit-based KD [8] employs the KL divergence to align the logits for classification. Compared to the one-hot labels that only contain the target category information, the logits encompass predictive information for all categories, which is also known as dark knowledge. In this way, we can learn a better student with such dark knowledge.

However, we argue that there exists an issue named **mis-instruction**, where the student would be misled when the teacher logits are wrong. Specifically, when the predictions based on teacher logits do not follow the labels, such erroneous would mislead the student model in the knowledge distillation process. Figure 1 shows one example of the mis-instruction issue where the input is an image of a cat. Under the experienced teacher with the right predictions (left side), the student model can effectively learn such knowledge and classify the image as the 'cat'. In contrast, when the logits from the teacher model contain the wrong prediction (a.k.a the inexperienced teacher on the right), the student model would be misled and generate a wrong prediction 'dog'.

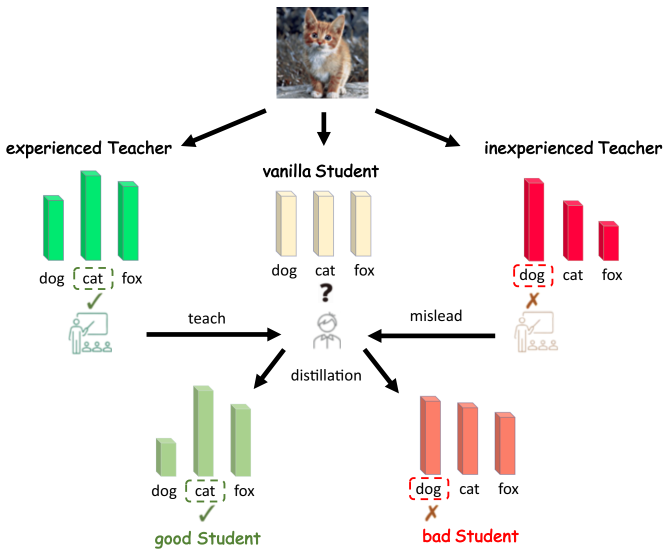

Figure 1. One example of the **mis-instruction** issue. The student model can generate the right prediction under an experienced teacher (in green), but would be misled under the inexperienced teacher (in red).

Meanwhile, the mis-instruction phenomenon is common in practice. As shown in Figure 2, the mis-instruction ratios on the ImageNet training set are notably high. Particularly, for the ResNet101, a typical teacher model on ImageNet, there are as many as 19.4% of the samples where the predictions based on teacher logits do not follow the labels.

To address the mis-instruction issue, one intuitive idea is to skip the samples with wrong logits in distillation. However, such logits also contain other valuable dark knowledge [1], such as the class discriminability. The class discriminability refers to the ratios of the logit for all the non-target classes and is vital in the distillation process [17]. Simply discarding these logits would also lead to the loss of such valuable dark knowledge. We thus ask: **can we calibrate the logits to both avoid mis-instruction and maintain other useful dark knowledge?**

In this paper, we propose a novel Logit Calibration (LoCa) method to calibrate the logits from the inexperienced teacher without any additional parameters. Our key insight is to guarantee that teacher logits are consistent with the ground-truth label (to avoid mis-instruction)

* Corresponding Author. Email: yang.yujiu@sz.tsinghua.edu.cn

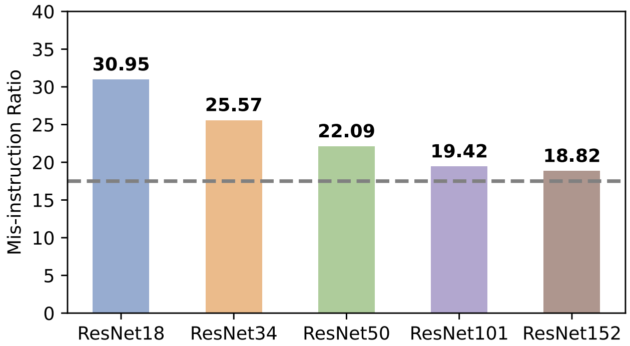

Figure 2. Mis-Instruction ratio of various teacher models on ImageNet training set. For all models, the ratios are greater than **17.5%**.

and also maintain the ratio of the non-target logits (to maintain the useful dark knowledge). Specifically, we model the calibration process as an optimization problem and propose a feasible solution by introducing a scaling factor. This optimization problem modeling revolves around three perspectives, namely 1) probability distribution, 2) prediction correctness, and 3) non-target proportion invariance. After that, we employ the calibrated logits in the knowledge distillation.

We perform experiments on 1) image classification tasks on CIFAR-100 [14] and ImageNet [25], and 2) text generation tasks on Dolly [5], S-NI [32] and UnNI [9] datasets. Experimental results indicate that our proposed LoCa significantly outperforms the baselines, demonstrating the effectiveness of calibrating logits. Moreover, further analysis shows that within the hyperparameter alpha range of 0.9 to 1.0, LoCa exhibits high usability and robustness. In conclusion, the main contributions are as follows:

- We find an issue termed mis-instruction, where the student model would be misled when the predictions based on teacher logits do not follow the labels.
- We propose a simple yet effective strategy, LoCa, which calibrates the teacher logits to avoid mis-instruction and maintain other useful dark knowledge.
- We conduct experiments on image classification and text generation tasks. The results of the experiment demonstrate the effectiveness of the proposed LoCa method.

2 Preliminary

In this section, we introduce the fundamental principles of knowledge distillation (KD), specifically focusing on the logit-based KD approach. After that, we also explore the mis-instruction issue and investigate it by analyzing the definition of its samples, demonstrating its detrimental effects through a comparative experiment.

2.1 *Logit-based KD*

The vanilla standard KD [8] transfers knowledge by aligning the output logits of the teacher and student models. Considering a classification task with C classes, the logits from the teacher model $\mathbf{p}$ can be formulated as follows:

$$\mathbf{p} = [p_1, p_2, ..., p_C] \in \mathbb{R}^{1\times C}, \tag{1}$$

Table 1. Results on CIFAR-100 dataset. * denotes the strategy that dropping the mis-instruction samples during the distillation process.

Model	ResNet56 →ResNet20	WRN-40-2→WRN-40-1
Teacher	72.34	75.61
Student	69.06	70.50
KD	70.66	73.54
KD*	70.88	73.59

where p_i is the probability of the i-th class. Typically, p_i is obtained through the softmax function with temperature:

$$p_i = \frac{\exp(z_i/\tau)}{\sum_{j=1}^{C}\exp(z_j/\tau)}, \tag{2}$$

where z_i represents the logit of i-th class and τ is the temperature parameter for scaling.

Similarly, the output logits $\mathbf{q}$ from the student model can be denoted as follows:

$$\mathbf{q} = [q_1, q_2, ..., q_C] \in \mathbb{R}^{1\times C}, \tag{3}$$

where q_i is the predicted probability for the i-th class.

To align the logits, we can employ the Kullback-Leibler (KL) divergence, which can be written as:

$$\mathrm{KD}(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} p_i \log\left(\frac{p_i}{q_i}\right). \tag{4}$$

Another goal is to generate the right prediction for the student model, where the cross-entropy loss is widely used. Specifically, the cross-entropy loss measures the discrepancy between the predicted probabilities $\mathbf{q}$ and the one-hot labels $\mathbf{y} \in \mathbb{R}^{1\times C}$ as follows:

$$\mathrm{CE}(\mathbf{q}, \mathbf{y}) = \sum_{i=1}^{n} y_i \log(q_i) = \log(q_{gt}). \tag{5}$$

where gt denotes the ground truth class and $\mathbf{y} = [y_1, ..., y_C]$ is defined as:

$$y_i = \begin{cases} 1 & \text{if } i \text{ is } gt \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Therefore, we can get the final optimal object in knowledge distillation:

$$\mathcal{L} = \beta \cdot \mathrm{KD}(\mathbf{p}, \mathbf{q}) + \gamma \cdot \mathrm{CE}(\mathbf{q}, \mathbf{y}), \tag{7}$$

where β and γ are hyperparameters that can be respectively tuned for optimal training performance. We adopted fixed parameters based on previous studies for fair comparisons. See Section 4.1.2 for detailed experimental settings.

2.2 *Analysis on Mis-instruction Issue*

Mis-instruction occurs when the logits of teacher model $\mathbf{p}$ indicate wrong class labels, transferring incorrect logits information to the student model $\mathbf{q}$. We first find the highest logits class index k_{logits} as follows:

$$k_{\text{logits}} = \mathrm{Argmax}_i\{p_i\}. \tag{8}$$

After that, the mis-instruction samples are defined as:

$$k_{\text{logits}} \neq gt, \tag{9}$$

where gt represents the index of the ground truth class.

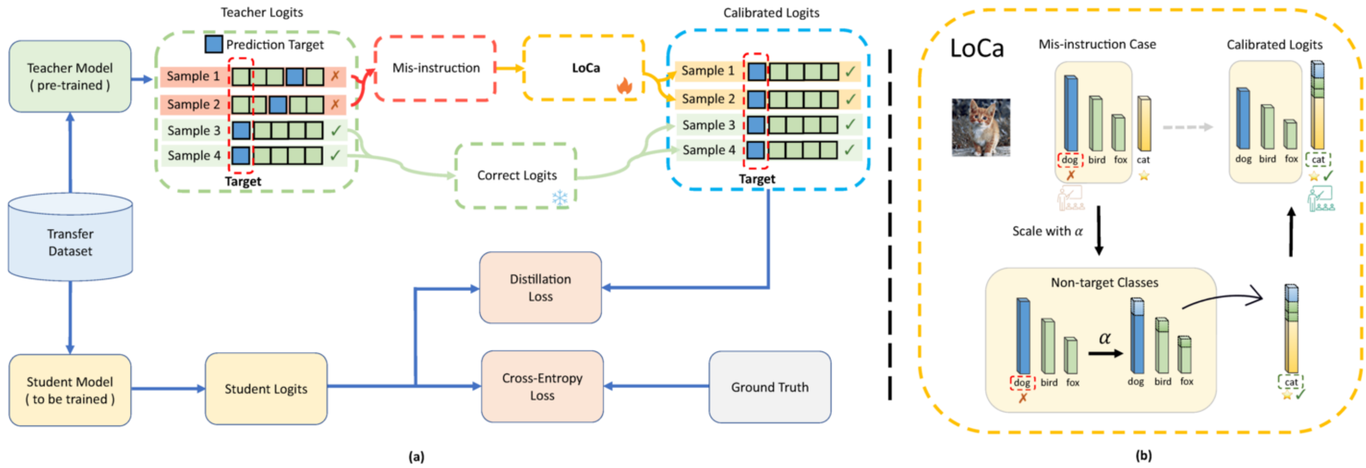

Figure 3. The details of the proposed LoCa method. During distillation, we first calibrate the logits for the mis-instruction samples and then employ the calibrated logits for KD. Specifically, we introduce a scaling factor α to decrease the non-target logits and increase the target logit.

Theoretically, mis-instruction samples would mislead students during the distillation. Specifically, the cross-entropy loss aims to optimize student logits toward ground truth label gt, while the KD loss aims to optimize it toward the label k_{logits} indicated by teacher logits. As shown in Equation 9, these two goals are not the same, leading to an optimization conflict.

Moreover, we perform experiments on a straightforward strategy to skip these mis-instruction samples. As shown in Table 1, the student model ResNet20 trained without mis-instruction samples gets a higher result of 70.88 than 70.66 of the vanilla KD, with other conditions being completely consistent. Thus, we can conclude that the mis-instruction issue exists and that it is **harmful to students** during the distillation process.

3 Methodology

In this section, we introduce the motivations and details of our proposed LoCa method to address the mis-instruction issue.

3.1 Optimization Objective

LoCa is designed under two goals: 1) addressing the mis-instruction issue, and 2) preserving the proportion of non-target classes, which aims to maintain the valuable dark knowledge information. Following Zhao et al. [42], we divide the logits into two parts, i.e., non-target and target categories. Then we model the calibration goals as an optimization problem.

Probability distribution. As a set of probability distributions, the calibrated logits $\mathbf{p}^{\text{loca}} = [p_1^{loca}, p_2^{loca}, ..., p_C^{loca}] \in \mathbb{R}^{1\times C}$ needs to satisfy the constraint condition that the sum of probabilities equals 1, which can be formulated as:

$$\sum_{i=1}^{C} p_i^{loca} = 1 \tag{10}$$

where each element represents a valid probability with

$$0 < p_i^{loca} < 1 \quad \forall i = 1, 2, 3, .., C\ . \tag{11}$$

Prediction correctness. To address the mis-instruction issue, the goal is to guarantee the predicted label to be consistent with the ground truth label, which can be formulated as:

$$k_{\text{logits}}^{loca} = \text{Argmax}_i \{p_i^{loca}\} = gt. \tag{12}$$

Non-target proportion invariance. As shown in Figure 3, we separate the logits into target and non-target categories. The next goal is to maintain useful dark knowledge in the non-target logits [42]. However, simply dropping these mis-instruction samples would also bring the loss of vital dark knowledge. To preserve this knowledge, the key is to maintain the ratios between any two non-target logits:

$$\frac{p_i^{loca}}{p_j^{loca}} = \frac{p_i}{p_j} \quad \forall i, j \neq gt. \tag{13}$$

3.2 LoCa: Logit Calibration

In this paper, we design the **Lo**git **Ca**libration (LoCa) method to achieve the aforementioned goals. Figure 3 shows the details.

For the mis-instruction sample, the calibrated logit is defined as:

$$p_i^{loca} = \begin{cases} s \cdot p_i & \text{if } i \neq gt, \\ 1 - \sum\limits_{i=1, i\neq k}^{C} p_i^{loca} & \text{if } i = gt, \end{cases} \tag{14}$$

Considering the prediction correctness shown in Equation 12, the logit at the ground truth label gt should be **only** maximum among these logits:

$$p_{gt}^{loca} > p_i^{loca} \quad \forall i \neq gt, \tag{15}$$

which equals

$$p_{gt}^{loca} > \max_{i\neq gt}(p_i^{loca}) = p_{k_{\text{logits}}}^{loca}. \tag{16}$$

Combining the Equation 14 and 16, we have:

$$1 - s \cdot \sum_{i\neq k} p_i > s \cdot p_{k_{\text{logits}}}. \tag{17}$$

Therefore, we can get a feasible solution:

$$s < \sigma = \frac{1}{1 - p_{gt} + p_{k_{\text{logits}}}}, \tag{18}$$

where the σ is the threshold. Thus, we introduce a hyperparameter α ranging between 0 and 1 and set $s = \alpha \cdot \sigma$. This configuration ensures that the LoCa method satisfies the aforementioned constraints. We then scale the logits in both target and non-target categories accordingly based on Equation 14.

Table 2. Results of various methods on the CIFAR-100 validation dataset. Results are averaged over 3 trials.

Teacher	ResNet56 72.34	ResNet110 74.31	ResNet32×4 79.42	WRN-40-2 75.61	WRN-40-2 75.61	WRN-40-2 75.61	ResNet50 79.34	ResNet32×4 79.42
Student	ResNet20 69.06	ResNet32 71.14	ResNet8×4 72.50	WRN-16-2 73.26	WRN-40-1 71.36	ShuffleNet-V1 70.50	MobileNet-V2 64.60	ShuffleNet-V2 71.82
KD [8]	70.66	73.08	73.33	74.92	73.54	74.83	67.35	74.45
LoCa-0.95	**71.08**	**73.36**	73.66	75.11	73.74	75.42	**68.66**	**75.30**
Δ	+0.42	+0.28	+0.33	+0.19	+0.20	+0.59	+1.31	+0.85
LoCa-0.98	70.88	73.32	**73.79**	**75.21**	**73.85**	**75.81**	68.60	75.10
Δ	+0.22	+0.24	+0.46	+0.29	+0.31	+0.98	+1.25	+0.65

4 Experiments

In this section, we perform experiments on two representative tasks in CV and NLP fields, i.e., image classification, and text generation.

4.1 *Image Classification Tasks*

4.1.1 *Datasets and Models*

Datasets. CIFAR-100 [14] is a well-known image classification dataset, containing 32×32 images of 100 categories. Training and validation sets contain 50,000 and 10,000 images. ImageNet [25] is a large-scale classification dataset that consists of 1000 classes. The training set contains 1.28 million images and the validation set contains 50,000 images.

Model. We employ different ResNet architectures as teacher models, including ResNet56, ResNet110, ResNet32×4, and WRN-40-2. For student models, we select both homologous ResNet structures and heterologous ShuffleNets and MobileNets.

4.1.2 *Baselines and Implementation*

Baselines. We primarily test the enhancements and performance improvements of our strategy on the vanilla KD [8] approach.

Implementation. We follow the same experimental settings as in previous work [3, 42]. For the experiments on CIFAR-100, the optimizer is SGD [28] and trained for 240 epochs. The learning rate is initialized as 0.01 for MobileNets [10, 26] and ShuffleNets [41], and 0.05 for ResNets [6] and WRNs [38].

For a fair comparison, we fix τ from Equation 2 and β, γ from Equation 7 across experiments: $\tau = 4, \beta = 0.9, \gamma = 0.1$ for CIFAR-100 and $\tau = 1, \beta = 0.5, \gamma = 0.5$ for ImageNet. We report the average results over 3 trials. It takes around 2 hours to train on 1 Nvidia A100 GPU for CIFAR-100 and around 24 hours for ImageNet.

4.1.3 *Main Results*

Results on CIFAR-100. Table 2 reports the validation accuracy. Based on the results, we can get the following findings:

- Our proposed LoCa can consistently improve the performance of distillation compared to baseline vanilla KD [8]. In particular, LoCa achieves 71.08% with ResNet56 as teacher and ResNet20 as student, which is 0.42 higher than the original KD method.
- We find that LoCa improves distillation performance for both homologous teacher-student pairs and heterologous pairs such as ResNet50 to MobileNet-V2, demonstrating the robustness towards model structures.
- We observe that different values of α would lead to variations in the benefits of our LoCa strategy, yet the overall trend remains an improvement, showing the robustness towards different hyperparameters.

Table 3. Top-1 and top-5 accuracy (%) on the ImageNet validation from **ResNet-34** to **ResNet-18**. The results are averaged over 3 trials.

Metrics	Teacher	Student	KD	LoCa-0.95	LoCa-0.99
top-1	73.31	69.75	70.66	71.08	71.15 (+0.49)
top-5	91.42	89.07	89.88	90.09	90.19 (+0.31)

Table 4. Top-1 and top-5 accuracy (%) on the ImageNet validation from **ResNet-50** to **MobileNetV1**. The results are averaged over 3 trials.

Metrics	Teacher	Student	KD	LoCa-0.95	LoCa-0.99
top-1	76.16	68.87	70.50	70.91	70.99 (+0.49)
top-5	92.86	88.76	89.80	90.03	90.06 (+0.26)

Results on ImageNet. We report the top-1 and top-5 accuracies of image classification on ImageNet in Table 3 and Table 4. Similarly, we find that our LoCa can achieve a consistent improvement in top-1 and top-5 accuracy on ImageNet validation. Specifically, LoCa gets an improvement of 0.49% on the top-1 accuracy under the setting from ResNet-34 to ResNet-18. Also, LoCa can improve the performance ranging from different settings. The findings are consistent with the CIFAR, which shows the robustness of the proposed LoCa towards various benchmarks.

4.2 *Text Generation Tasks*

We follow the same experimental settings as Wu et al. [35], first finetuning the teacher model and then distilling the teacher model. After that, we report the average Rouge-L scores on popular benchmarks.

4.2.1 *Datasets and Models*

Datasets. For training data, we employ the instruction response dataset following Gu et al. [5], which is built from databricks-dolly-15k[1] and contains 14k samples for training, 500 samples for valid, and 500 samples for testing.

The details for the evaluation dataset are as follows:

- Dolly: human-written instruction-response pairs We divide the data set into 14k samples for training, 500 samples for validation, and 500 samples for testing following Gu et al. [5].
- S-NI: the test set of SUP-NATINST [32], which contains 9K samples from 119 English tasks. In this paper, we employ the samples with ground truth responses longer than 11.
- UnNI: dataset from Honovich et al. [9]. Similarly, we employ samples with ground-truth responses longer than 11.

Models. We perform experiments on the mainstreaming model LLaMA [30]. Specifically, we employ LLaMA with 7B parameters

[1] https://github.com/databrickslabs/dolly/tree/master

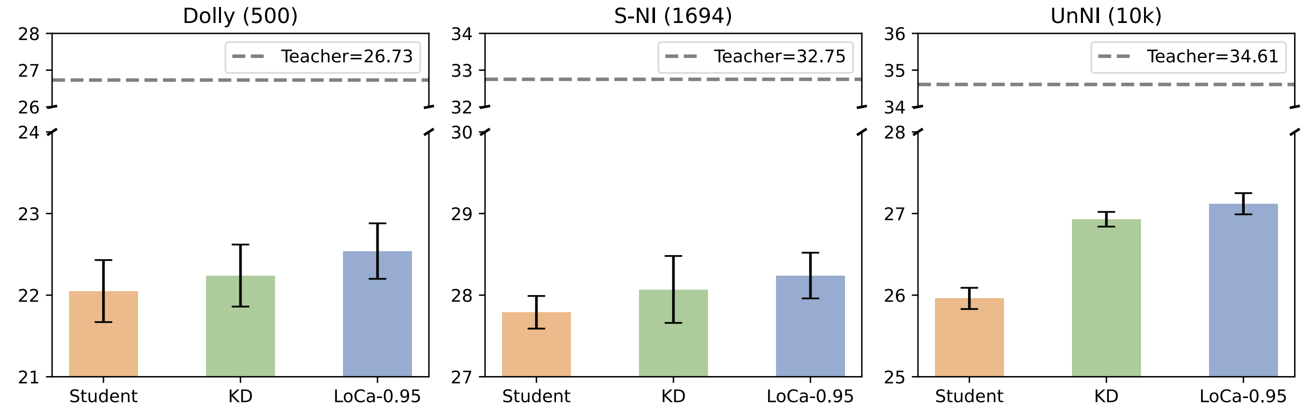

Figure 4. Rouge-L scores on Dolly, S-NI, and UnNI datasets. We report the average and standard deviation scores for 5 trials. Our proposed LoCa outperforms KD on all benchmarks.

Table 5. Statistics of the datasets for distillation on the text generation tasks.

Dataset Name	Usage	Samples
Dolly	Train	14,000
	Valid	500
	Test	500
S-NI	Test	1,694
UnNI	Test	10,000

as our teacher model and TinyLLaMA [39] with 1.1B parameters[2] as the student model. This allows us to evaluate the effectiveness of LoCa across widely recognized model architectures.

4.2.2 Baselines and Implementation

Baseline. For the baseline student model, we directly train the student model by performing SFT on the dataset without any distillation and denote it as "Student". Moreover, we employ SeqKD [13] to train the student model on the data generated from the teacher model, which we denote as "KD" for the consistency of the expression.
Implementation. For TinyLLaMA, we set the batch size as 60 and train for 10 epochs. The learning rate is 1e-5. For the student, the maximum input length is 512. It takes around 1h to train on 4 Nvidia A100 GPUs. We report the results of the Rouge-L scores for five different seeds.

4.2.3 Main Results

As shown in Figure 4, we present the average Rouge-L scores and variance for students trained using SFT, KD, and LoCa methods on the Dolly, S-NI, and UnNI datasets. The dashed lines indicate the performance of the teacher model. We can observe consistent improvements employing our LoCa method across three datasets, compared to both the students trained via SFT and those transferred via KD, which demonstrates the effectiveness of our method. In particular, on the larger UnNI dataset with 10,000 samples, LoCa outperforms both SFT and KD, with p-values less than 0.001 for both comparisons, indicating a statistically significant improvement. Notably, on the Dolly dataset, the enhancements from our LoCa method to KD surpass the improvements from SFT to KD. Furthermore, the variance of LoCa is much smaller than the baselines, which shows the effectiveness of distilling knowledge.

[2]https://huggingface.co/TinyLlama/TinyLlama-1.1B-intermediate-step-1195k-token-2.5T

5 Analysis

In this section, we perform further detailed analyses on LoCa and apply LoCa on more baselines such as DKD.

5.1 Impact of the α

We perform ablation studies to assess the sensitivity of our approach to the hyperparameter α. Figure 5 reports the student accuracy (%) with different α, where we employ ResNet32×4 and ResNet8×4 as the teacher and student on CIFAR-100 (left), ResNet32 and ResNet18 on ImageNet (right). We can observe that the improvements on CIFAR-100 are relatively stable, with $\alpha = 0.95$ achieving the peak performance within the ablation range. Meanwhile, the performance on ImageNet is also improved, proving that LoCa can consistently increase model performance, especially when $\alpha = 0.995$. Moreover, such consistency over different hyperparameters indicates the robustness of our LoCa in achieving consistent improvements under different settings.

Meanwhile, we find that the effects on ImageNet are more sensitive compared to CIFAR-100, specifically showing that when $\alpha = 0.95$ or $\alpha = 0.999$, there are generally slight improvements or performances close to the original. We attribute the difference in performance between ImageNet and CIFAR-100 to the larger size and the larger number of categories in the ImageNet dataset, suggesting that the adjustments must be moderate.

We also perform experiments when α is equal to or greater than 1. In this case, the predicted labels are not guaranteed to be the ground truth. As shown in Table 6, we can find that LoCa performs worse than the baseline vanilla KD, indicating the importance of prediction correctness.

5.2 Time Cost

In assessing the additional timing introduced by the LoCa method during the distillation on CIFAR-100, LoCa incurs a minor computational cost, as detailed in Table 7. Specifically, from WRN_40_2 to ShuffleV1, the KD Loss computation time increased from 0.32 ms to

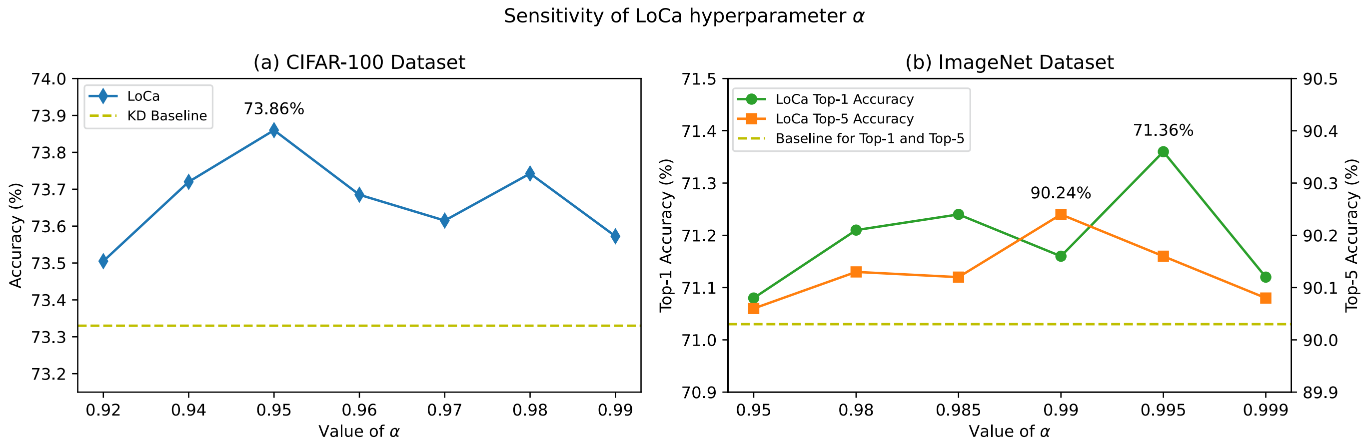

Figure 5. The ablation studies under different α settings in our LoCa. We employ **ResNet32×4** and **ResNet8×4** as the teacher and the student on CIFAR-100 (part a). We set **ResNet-34** to **ResNet-18** as the teacher and student on ImageNet (part b).

Table 6. Sensitivity of α around 1. We report the Top-1 accuracy (%) from **ResNet34** to **ResNet18** on ImageNet. When α is equal or larger than 1, the performance would be worse than the baseline.

Metric	**Vanilla KD**	**The value of α in LoCa**			
		0.995	0.999	**1.000**	**1.010**
Top-1	**71.03**	71.36	71.12	71.01	70.94

Table 7. Time costs for LoCa and baselines. We report the average scores for five trials. KD Loss denotes the time costs for calculating KL divergence. Batch denotes the time costs for one batch (64 samples) and Epoch for the process of training one epoch.

Method	**KD Loss**	**Batch**	**Epoch**
KD	0.32 ms	27.83 ms	25.32 s
LoCa	0.39 ms	28.29 ms	25.57 s
Δ	21.88%	1.65%	0.99%

0.39 ms, indicating a marginal additional computational overhead of 0.07 ms by LoCa. Furthermore, despite frequent function calls (i.e., 781 times per epoch), the additional time per call is less than 0.1 ms, contributing less than 0.1 seconds to the approximately 28-second duration of each epoch.

This slight proportional increase is consistent with our anticipated increases in computational cost, suggesting that the LoCa method improves performance **without requiring significant additional time** (less than 1%).

5.3 LoCa on DKD

We conducted extensive experiments on the KD baseline, and the results demonstrate the effectiveness and robustness of the proposed LoCa. Nevertheless, LoCa can be easily extended on more baselines, such as DKD [42]. Specifically, we first apply LoCa to adjust the logits on the mis-instruction samples, followed by the standard process of DKD. Table 8 shows the results of the LoCa with various α. We report the average score of 3 trials. We can find that applying LoCa would improve the performance of DKD in all α, demonstrating the effectiveness. Specifically, DKD with LoCa (α=0.95) would get 77.25, 0.37 higher than the vanilla DKD, when distilling ResNet32x4 to ShuffleNet-V2.

Table 8. Results on CIFAR-100 dataset when applying LoCa on DKD.

Teacher **Student**	**ResNet110** **ResNet32**	**ResNet32x4** **ShuffleNet-V2**
DKD	73.92	76.88
w/ LoCa (α=0.95)	73.96	**77.25**
w/ LoCa (α=0.98)	73.95	77.09
w/ LoCa (α=0.99)	**73.99**	77.18

5.4 Case Study

We further showcase the text generation outputs of baselines and proposed LoCa. The cases indicate that LoCa can effectively inherit the knowledge from the teacher model, such as the grammar information (see Table 9), and avoiding the hallucinations (see Table 10).

Specifically, Table 9 reports a case of concatenating the given sentences. It shows that although the KD method connects the given sentences as instruction, it introduces extraneous punctuation and does not adhere fully to grammatical and syntactic norms like the extra **comma** and **colon** at the end. In contrast, LoCa perfectly replicates the teacher's output, effectively leveraging the distilled knowledge to seamlessly avoid errors. This superiority of LoCa is attributed to its ability to assimilate correct and comprehensive dark knowledge from the teacher output during training.

Table 10 presents an example of hallucination, a common issue in large language model inference. Hallucination typically manifests as descriptions of related subjects rather than accurate responses to the standard prompts (as Student answer in Table 10). In our analysis with LoCa, one potential reason is the student's unclear and imprecise understanding when transferring knowledge. LoCa addresses this by correcting the mis-instructions in the supervision information to be absolutely accurate, thereby mitigating the ambiguity and enhancing performance on the test set.

Table 9. One case from Dolly. LoCa can avoid *grammar errors* with the knowledge distilled from the teacher model, while KD fails.

Instruction	**You are given a list of sentences. Your task is to concatenate these sentences in the same order and print the result on a single line.**
Input	Sentence1 = I'm happy today., Sentence2 = This is a good day., Sentence3 = I will succeed.
Grouth Truth	I'm happy today. This is a good day. I will succeed.
Teacher	I'm happy today. This is a good day. I will succeed.
Student	I'm happiness today., This is a good day.", I will success
KD	I'm happy today., This is a good day.", I will succeed.
LoCa(ours)	I'm happy today. This is a good day. I will succeed.

Table 10. Another case from Dolly. The vanilla student model exists *hallucinations*. LoCa can address this issue, but KD fails.

Instruction	**In this task, you are given a list of Amazon products. The task is to classify each product as "GOOD" if the majority of reviewers recommend the product or as "BAD" if the majority of reviewers do not recommend the product.**
Input	Product1: 4 out of 5 stars, Product2: 3 out of 5 stars, Product3: 2 out of 5 stars.
Grouth Truth	Product1: GOOD, Product2: BAD, Product3: BAD
Teacher	Product1: GOOD, Product2: BAD, Product3: BAD
Student	Product1: GOOD, Product3: BAD, Product2: GOOD, Product2: BAD Note: Three products (Product1, Product3, and Product2) have more than one review and are represented only once in the list. In this case, we classify them as the same label regardless of what majority rule (more positive reviews are SAFE)
KD	Product1: Good. 20% of the reviewers recommend this product. Product2: Bad. 80% of the review do not recommend this product. Product3: Bad. 81% of the review do not recommend this product.
LoCa (ours)	Product1: GOOD, Product2: BAD, Product3: BAD

6 Related Work

6.1 Logit-base Knowledge Distillation

The logit-based knowledge distillation is originally proposed by Hinton [8], in which the student is trained to mimic the soft logit output of the teacher [31, 33]. Several methods have been developed to enhance its effectiveness. SimKD [2] improves knowledge transfer by utilizing the teacher's classifier within the student network. DKD [42] proposes the decoupled knowledge distillation that divides logit knowledge into target knowledge and non-target knowledge. NKD [36] further proposes normalizing the non-target logits to equalize their sum. ATS [17] decouples KD into three components, clarifying that knowledge transfer requires expanding the variance of incorrect category probabilities. It implements this improvement through dynamic temperature adjustments. CTKD [18] improves knowledge distillation by assigning different temperatures to instances. Other works [44, 23, 12] refine the logit-based distillation paradigm to enhance the effectiveness when using stronger teacher models for distillation. We find that logit-based distillation offers numerous advantages as a way to compress the model, making it a worthwhile focus for our further research.

6.2 Logits Errorness

SDD [21] uncovers multi-label issues on ImageNet, using the logits map to partition sub-regions and enhance the information density of logits for more effective teaching to solve the mismatch between teacher's predictions to ground truth. TIE-KD [40] focuses on discrepancies between teacher predictions and ground truth and emphasize the top-k predictions to enhance overall performance. Our approach differs as it aims to ensure absolute consistency in the teacher's predictions while minimizing the gap.

6.3 Probability Calibration

Some methods require little extra or even no more time than training the model directly. For example, Label Smoothing [29] sets the labels manually by distributing the same values to all non-target classes. Tf-KD [37] revisits KD via label smoothing, using a high temperature to generate the manual logit for distillation. LSKD [27] explores the feasibility of varying temperature coefficients and achieves Logit standardization in knowledge distillation through Z transformations. Different from their works, we emphasize the use of simple linear transformations to preserve the relative proportions of non-target categories, thereby retaining the valuable dark knowledge within them.

7 Conclusion

This work revisits conventional logit-based distillation and reveals that the effectiveness of KD is limited by situations we term as **mis-instruction**, where the student model is misled when predictions based on teacher logits do not align with the ground truth labels. To overcome this limitation, we propose a simple yet effective method called **Lo**git **Ca**libration (LoCa), which calibrate the supervision logits in cases of mis-instruction by decreasing non-target logits and enhancing target logits. We establish knowledge-transferring pipelines for these mis-instruction outputs to transfer accurate target information and other useful relation information. This approach ensures that the adjusted target logits are absolutely correct, while preserving the relative proportions among non-target logits to maintain dark knowledge. Our proposed LoCa method does not require any additional parameters. Extensive experiments on several benchmark datasets, including image classification and text generation tasks, demonstrate the effectiveness of LoCa across a range of teacher-student pairs.

8 Ethics Statement

This research utilized publicly available datasets in the fields of Computer Vision (CV) and Natural Language Processing (NLP), which do not involve sensitive personal information or human subjects, thus not requiring specific ethical approvals. We adhere to ethical guidelines respecting privacy and intellectual property rights of the data sources. Our project develops tools intended to support, not substitute, the anticipatory processes in AI technologies. We emphasize the non-commercial use of the datasets and fine-tuned models, and the adherence to fair use in data sourcing. Conflicts of interest do not influence this research, which ensures its integrity and the reliability of its results.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (Grant No. 61991451) and the Shenzhen Science and Technology Program (JCYJ20220818101001004).

References

[1] K. Chandrasegaran, N.-T. Tran, Y. Zhao, and N.-M. Cheung. Revisiting label smoothing and knowledge distillation compatibility: What was missing? In *International Conference on Machine Learning*, pages 2890–2916, 2022.

[2] D. Chen, J.-P. Mei, H. Zhang, C. Wang, Y. Feng, and C. Chen. Knowledge distillation with the reused teacher classifier. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11933–11942, 2022.

[3] P. Chen, S. Liu, H. Zhao, and J. Jia. Distilling knowledge via knowledge review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5008–5017, 2021.

[4] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, 2019.

[5] Y. Gu, L. Dong, F. Wei, and M. Huang. Knowledge distillation of large language models. *arXiv preprint arXiv:2306.08543*, 2023.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[7] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2961–2969, 2017.

[8] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[9] O. Honovich, T. Scialom, O. Levy, and T. Schick. Unnatural instructions: Tuning language models with (almost) no human labor. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14409–14428, 2023.

[10] A. G. Howard. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[11] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[12] T. Huang, S. You, F. Wang, C. Qian, and C. Xu. Knowledge distillation from a stronger teacher. *Advances in Neural Information Processing Systems*, 35:33716–33727, 2022.

[13] Y. Kim and A. M. Rush. Sequence-level knowledge distillation. *arXiv preprint arXiv:1606.07947*, 2016.

[14] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[15] S. Lao, G. Song, B. Liu, Y. Liu, and Y. Yang. Masked autoencoders are stronger knowledge distillers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6384–6393, 2023.

[16] S. Lao, G. Song, B. Liu, Y. Liu, and Y. Yang. Unikd: Universal knowledge distillation for mimicking homogeneous or heterogeneous object detectors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6362–6372, 2023.

[17] X.-C. Li, W.-S. Fan, S. Song, Y. Li, S. Yunfeng, D.-C. Zhan, et al. Asymmetric temperature scaling makes larger networks teach well again. *Advances in Neural Information Processing Systems*, 35:3830–3842, 2022.

[18] Z. Li, X. Li, L. Yang, B. Zhao, R. Song, L. Luo, J. Li, and J. Yang. Curriculum temperature for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1504–1512, 2023.

[19] Y. Liu, M. Ott, N. Goyal, et al. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[20] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[21] S. W. C. L. Y. Luo. Scale decoupled distillation. *arXiv preprint arXiv:2403.13512*, 2024.

[22] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European Conference on Computer Vision*, pages 116–131, 2018.

[23] Z. Qiu, X. Ma, K. Yang, C. Liu, J. Hou, S. Yi, and W. Ouyang. Better teacher better student: Dynamic prior knowledge for knowledge distillation. *arXiv preprint arXiv:2206.06067*, 2022.

[24] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2016.

[25] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015.

[26] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[27] S. Sun, W. Ren, J. Li, R. Wang, and X. Cao. Logit standardization in knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15731–15740, 2024.

[28] I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *International Conference on Machine Learning*, pages 1139–1147, 2013.

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

[30] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[31] J. Wang, S. Zhang, Y. Liu, T. Wu, Y. Yang, X. Liu, K. Chen, P. Luo, and D. Lin. Riformer: Keep your vision backbone effective but removing token mixer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14443–14452, 2023.

[32] Y. Wang, S. Mishra, P. Alipoormolabashi, et al. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, 2022.

[33] T. Wu, Z. Zhao, J. Wang, X. Bai, L. Wang, N. Wong, and Y. Yang. Edge-free but structure-aware: Prototype-guided knowledge distillation from gnns to mlps. *arXiv preprint arXiv:2303.13763*, 2023.

[34] T. Wu, C. Hou, S. Lao, J. Li, N. Wong, Z. Zhao, and Y. Yang. Weight-inherited distillation for task-agnostic bert compression. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 13–28, 2024.

[35] T. Wu, C. Tao, J. Wang, Z. Zhao, and N. Wong. Rethinking kullback-leibler divergence in knowledge distillation for large language models. *arXiv preprint arXiv:2404.02657*, 2024.

[36] Z. Yang, A. Zeng, Z. Li, T. Zhang, C. Yuan, and Y. Li. From knowledge distillation to self-knowledge distillation: A unified approach with normalized loss and customized soft labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17185–17194, 2023.

[37] L. Yuan, F. E. Tay, G. Li, T. Wang, and J. Feng. Revisiting knowledge distillation via label smoothing regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3903–3911, 2020.

[38] S. Zagoruyko and N. Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[39] P. Zhang, G. Zeng, T. Wang, and W. Lu. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024.

[40] S. Zhang, Y. Liang, S. Wang, W. Han, J. Liu, J. Xu, and Y. Chen. Towards understanding and improving knowledge distillation for neural machine translation. In *Annual Meeting of the Association for Computational Linguistics*, pages 8062–8079, 2023.

[41] X. Zhang, X. Zhou, M. Lin, and J. Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6848–6856, 2018.

[42] B. Zhao, Q. Cui, R. Song, Y. Qiu, and J. Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11953–11962, 2022.

[43] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2881–2890, 2017.

[44] Y. Zhu, N. Liu, Z. Xu, X. Liu, W. Meng, L. Wang, Z. Ou, and J. Tang. Teach less, learn more: On the undistillable classes in knowledge distillation. *Advances in Neural Information Processing Systems*, 35:32011–32024, 2022.