CONFIDENCE-AWARE MULTI-TEACHER KNOWLEDGE DISTILLATION

Hailin Zhang *Defang Chen* *Can Wang*⋆

Zhejiang University, China; ZJU-Bangsun Joint Research Center.
{zzzhl, defchern, wcan}@zju.edu.cn

ABSTRACT

Knowledge distillation is initially introduced to utilize additional supervision from a single teacher model for the student model training. To boost the student performance, some recent variants attempt to exploit diverse knowledge sources from multiple teachers. However, existing studies mainly integrate knowledge from diverse sources by averaging over multiple teacher predictions or combining them using other label-free strategies, which may mislead student in the presence of low-quality teacher predictions. To tackle this problem, we propose Confidence-Aware Multi-teacher Knowledge Distillation (CA-MKD), which adaptively assigns sample-wise reliability for each teacher prediction with the help of ground-truth labels, with those teacher predictions close to one-hot labels assigned large weights. Besides, CA-MKD incorporates features in intermediate layers to stable the knowledge transfer process. Extensive experiments show our CA-MKD consistently outperforms all compared state-of-the-art methods across various teacher-student architectures. Code is available: https://github.com/Rorozhl/CA-MKD.

Index Terms— knowledge distillation, multiple teachers, confidence-aware weighting

1. INTRODUCTION

Nowadays, deep neural networks have achieved unprecedented success in various applications [1, 2, 3]. However, these complex models requiring huge memory footprint and computational resources are difficult to be applied on embedded devices. Knowledge distillation (KD) is thus proposed as a model compression technique to resolve this issue, which improves the accuracy of a lightweight student model by distilling the knowledge from a pre-trained cumbersome teacher model [4]. The transferred knowledge was originally formalized as softmax outputs (soft targets) of the teacher model [4] and latter extended to the intermediate teacher layers for achieving more promising performance [5, 6, 7].

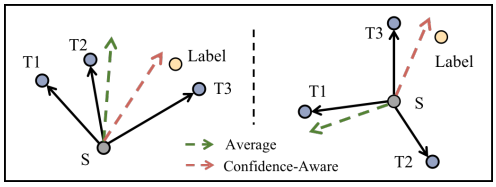

Fig. 1. Comparison of the previous average direction (green line) and our proposed confidence-aware direction (red line).

As the wisdom of the masses exceeds that of the wisest individual, some multi-teacher knowledge distillation (MKD) methods are proposed and have been proven to be beneficial [8, 9, 10, 11, 12]. Basically, they combine predictions from multiple teachers with the fixed weight assignment [8, 9, 10] or other various label-free schemes, such as calculating weights based on a optimization problem or entropy criterion [11, 12], etc. However, fixed weights fail to differentiate high-quality teachers from low-quality ones [8, 9, 10], and the other schemes may mislead the student in the presence of low-quality teacher predictions [11, 12]. Figure 1 provides an intuitive illustration on this issue, where the student trained with the average weighting strategy might deviate from the correct direction once most teacher predictions are biased.

Fortunately, we actually have ground-truth labels in hand to quantify our confidence about teacher predictions and then filter out low-quality predictions for better student training. To this end, we propose Confidence-Aware Multi-teacher Knowledge Distillation (CA-MKD) to learn sample-wise weights by taking the prediction confidence of teachers into consideration for adaptive knowledge integration. The confidence is obtained based on the cross entropy loss between prediction distributions and ground-truth labels. Compared with previous label-free weighting strategies, our technique enables the student to learn from a relatively correct direction.

Note that our confidence-aware mechanism not only is able to adaptively weight different teacher predictions based on their sample-wise confidence, but also can be extended to the student-teacher feature pairs in intermediate layers. With the help of our generated flexible and effective weights, we could avoid those poor teacher predictions dominating the knowledge transfer process and considerably improve the student performance on eight teacher-student architecture combinations (as shown in Table 1 and 3).

⋆Corresponding author

This work is supported by National Key R&D Program of China (Grant No: 2019YFB1600700), the Starry Night Science Fund of Zhejiang University Shanghai Institute for Advanced Study (Grant No: SN-ZJU-SIAS-001) and National Natural Science Foundation of China (Grant No: U1866602).

2. RELATED WORK

Knowledge Distillation. Vanilla KD aims to transfer knowledge from a complex network (teacher) to a simple network (student) with the KL divergence minimization between their softened outputs [13, 4]. Mimicking the teacher representations from intermediate layers was latter proposed to explore more knowledge forms [5, 6, 14, 15, 7]. Compared to these methods that require pre-training a teacher, some works simultaneously train multiple students and encourage them to learn from each other instead [16, 17]. Our technique differs from these online KD methods since we attempt to distill knowledge from multiple pre-trained teachers.

Multi-teacher Knowledge Distillation. Rather than employing a single teacher, MKD boosts the effectiveness of distillation by integrating predictions from multiple teachers. A bunch of methods are proposed, such as simply assigning average or other fixed weights for different teachers [8, 9, 10], and calculating the weights based on entropy [12], latent factor [18] or multi-objective optimization in the gradient space [11]. However, these label-free strategies may mislead the student training in the presence of low-quality predictions. For instance, entropy-based strategy will prefer models with blind faith since it favors predictions with low variance [12]; optimization-based strategy favors majority opinion and will be easily misled by noisy data [11]. In contrast, our CA-MKD quantifies the teacher predictions based on ground-truth labels and further improves the student performance.

3. METHODOLOGY

We denote $\mathcal{D} = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}_i^N$ as a labeled training set, N is the number of samples, K is the number of teachers. $F \in \mathbb{R}^{h\times w\times c}$ is the output of the last network block. We denote $\boldsymbol{z} = [z^1, ..., z^C]$ as the logits output, where C is the category number. The final model prediction is obtained by a softmax function $\sigma(z^c) = \frac{\exp(z^c/\tau)}{\sum_j \exp(z^j/\tau)}$ with temperature τ. In the following sections, we will introduce our CA-MKD in detail.

3.1. The Loss of Teacher Predictions

To effectively aggregate the prediction distributions of multiple teachers, we assign different weights which reflects their sample-wise confidence by calculating the cross entropy loss between teacher predictions and ground-truth labels

$$\mathcal{L}^k_{CE_{KD}} = -\sum_{c=1}^{C} y^c \log\left(\sigma\left(z^c_{T_k}\right)\right), \tag{1}$$

$$w^k_{KD} = \frac{1}{K-1}\left(1 - \frac{\exp\left(\mathcal{L}^k_{CE_{KD}}\right)}{\sum_j \exp\left(\mathcal{L}^j_{CE_{KD}}\right)}\right), \tag{2}$$

where T_k denotes the kth teacher. The less $\mathcal{L}^k_{CE_{KD}}$ corresponds to the larger w^k_{KD}. The overall teacher predictions are then aggregated with calculated weights

$$\mathcal{L}_{KD} = -\sum_{k=1}^{K} w^k_{KD} \sum_{c=1}^{C} z^c_{T_k} \log\left(\sigma\left(z^c_S\right)\right). \tag{3}$$

According to the above formulas, the teacher whose prediction is closer to ground-truth labels will be assigned larger weight w^k_{KD}, since it has enough confidence to make accurate judgement for correct guidance. In contrast, if we simply acquire the weights by calculating the entropy of teacher predictions [12], the weight will become large when the output distribution is sharp regardless of whether the highest probability category is correct. In this case, those biased targets may misguide the student training and further hurt its distillation performance.

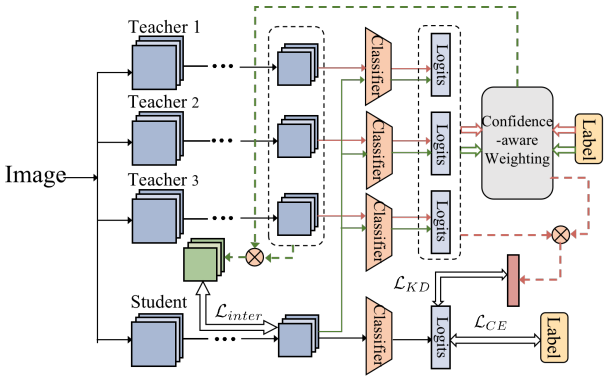

Fig. 2. An overview of our CA-MKD. The weight calculation of teacher predictions and intermediate teacher features are depicted as the red lines and green lines, respectively.

3.2. The Loss of Intermediate Teacher Features

In addition to KD Loss, inspired by FitNets [5], we believe that the intermediate layers are also beneficial for learning structural knowledge, and thus extend our method to intermediate layers for mining more information. The calculation of intermediate feature matching is presented as follows

$$z_{S\to T_k} = W_{T_k} h_S, \tag{4}$$

$$\mathcal{L}^k_{CE_{inter}} = -\sum_{c=1}^{C} y^c \log\left(\sigma\left(z^c_{S\to T_k}\right)\right), \tag{5}$$

$$w^k_{inter} = \frac{1}{K-1}\left(1 - \frac{\exp\left(\mathcal{L}^k_{CE_{inter}}\right)}{\sum_j \exp\left(\mathcal{L}^j_{CE_{inter}}\right)}\right). \tag{6}$$

where W_{T_k} is the final classifier of the kth teacher. $h_S \in \mathbb{R}^c$ is the last student feature vector, i.e, $h_S = \text{AvgPooling}(F_S)$. $\mathcal{L}^k_{CE_{inter}}$ is obtained by passing h_S through each teacher classifier. The calculation of w^k_{inter} is similar to that of w^k_{KD}.

Table 1. Top-1 test accuracy of MKD methods by distilling the knowledge on multiple teachers with the same architectures.

Teacher	WRN40-2	ResNet56	VGG13	VGG13	ResNet32x4	ResNet32x4	ResNet32x4
	76.62±0.26	73.28±0.30	75.17±0.18	75.17±0.18	79.31±0.14	79.31±0.14	79.31±0.14
Ensemble	79.62	76.00	77.07	77.07	81.16	81.16	81.16
Student	ShuffleNetV1	MobileNetV2	VGG8	MobileNetV2	ResNet8x4	ShuffleNetV2	VGG8
	71.70±0.43	65.64±0.19	70.74±0.40	65.64±0.19	72.79±0.14	72.94±0.24	70.74±0.40
AVER [8]	76.30±0.25	70.21±0.10	74.07±0.23	68.91±0.35	74.99±0.24	75.87±0.19	73.26±0.39
FitNet-MKD [5]	76.59±0.17	70.69±0.56	73.97±0.22	68.48±0.07	74.86±0.21	76.09±0.13	73.27±0.19
EBKD [12]	76.61±0.14	70.91±0.22	74.10±0.27	68.24±0.82	75.59±0.15	76.41±0.12	73.60±0.22
AEKD [11]	76.34±0.24	70.47±0.15	73.78±0.03	68.39±0.50	74.75±0.28	75.95±0.20	73.11±0.27
CA-MKD	**77.94±0.31**	**71.38±0.02**	**74.30±0.16**	**69.41±0.20**	**75.90±0.13**	**77.41±0.14**	**75.26±0.32**

Table 2. Top-1 test accuracy of CA-MKD compared to single-teacher knowledge distillation methods.

Teacher	WRN40-2	ResNet32x4	ResNet56
	76.62±0.26	79.31±0.14	73.28±0.30
Student	ShuffleNetV1	VGG8	MobileNetV2
	71.70±0.19	70.74±0.40	65.64±0.43
KD [4]	75.77±0.14	72.90±0.34	69.96±0.14
FitNet [5]	76.22±0.21	72.55±0.66	69.02±0.28
AT [6]	76.44±0.38	72.16±0.12	69.79±0.26
VID [14]	76.32±0.08	73.09±0.29	69.45±0.17
CRD [15]	76.58±0.23	73.57±0.25	71.15±0.44
CA-MKD	**77.94±0.31**	**75.26±0.13**	**71.38±0.02**

To stable the knowledge transfer process, we design the student to be more focused on imitating the teacher with a similar feature space and w_{inter}^k indeed serves as such a similarity measure representing the discriminability of a teacher classifier in the student feature space. The ablation study also shows that utilizing w_{inter}^k instead of w_{KD}^k for the knowledge aggregation in intermediate layers is more effective.

$$\mathcal{L}_{inter} = \sum_{k=1}^{K} w_{inter}^k ||F_{T_k} - r\left(F_S\right)||_2^2, \tag{7}$$

where $r(\cdot)$ is a function for aligning the student and teacher feature dimensions. The ℓ_2 loss function is used as distance measure of intermediate features. Finally, the overall training loss between feature pairs will be aggregated by w_{inter}^k.

In our work, only the output features of the last block are adopted to avoid incurring too much computational cost.

3.3. The Overall Loss Function

In addition to the aforementioned two losses, a regular cross entropy with the ground-truth labels is calculated

$$\mathcal{L}_{CE} = -\sum_{c=1}^{C} y^c \log\left(\sigma(z_S^c)\right). \tag{8}$$

The overall loss function of our CA-MKD is summarize as

$$\mathcal{L} = \mathcal{L}_{CE} + \alpha\mathcal{L}_{KD} + \beta\mathcal{L}_{inter}, \tag{9}$$

where α and β are hyper-parameters to balance the effect of knowledge distillation and standard cross entropy losses.

4. EXPERIMENT

In this section, we conduct extensive experiments on CIFAR-100 dataset [19] to verify the effectiveness of our proposed CA-MKD. We adopt eight different teacher-student combinations based on popular neural network architectures. All compared multi-teacher knowledge distillation (MKD) methods use three teachers except for special declarations.

Compared Methods. Besides the naïve AVER [8], we reimplement a single-teacher based method FitNet [5] on multiple teachers and denote it as FitNet-MKD. FitNet-MKD will leverage extra information coming from averaged intermediate teacher features. We also reimplement an entropy-based MKD method [12], which has achieved remarkable results in acoustic experiments, on our image classification task and we denote it as EBKD. As for AEKD, we adopt its logits-based version with the author provided code [11].

Hyper-parameters. All neural networks are optimized by stochastic gradient descent with momentum 0.9, weight decay 0.0001. The batch size is set to 64. As the previous works do [15, 7], the initial learning rate is set to 0.1, except MobileNetV2, ShuffleNetV1 and ShuffleNetV2 are set to 0.05. The learning rate is multiplied by 0.1 at 150, 180 and 210 of the total 240 training epochs. For the sake of fairness, the temperature τ is set to 4 and the α is set to 1 in all methods. Furthermore, we set the β of our CA-MKD to 50 throughout the experiments. All results are reported in means and standard deviations over 3 runs with different random seeds.

4.1. Results on the Same Teacher Architectures

Table 1 shows the top-1 accuracy comparison on CIFAR-100. We also include the results of teacher ensemble with the majority voting strategy. We can find that CA-MKD surpasses

Table 3. Top-1 test accuracy of MKD approaches by distilling the knowledge on multiple teachers with different architectures.

VGG8	AVER	FitNet-MKD	EBKD	AEKD	CA-MKD	ResNet8x4	ResNet20x4	ResNet32x4
70.74±0.40	74.55±0.24	74.47±0.21	74.07±0.17	74.69±0.29	**75.96±0.05**	72.79	78.39	79.31

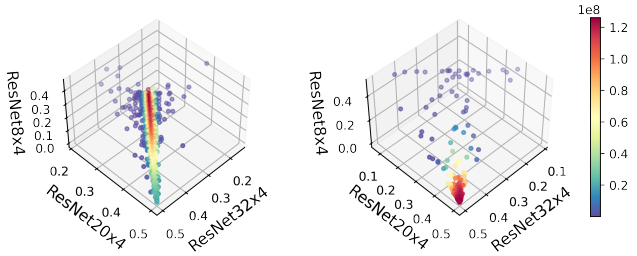

Fig. 3. The visualization results of learned weights by CA-MKD on each training sample.

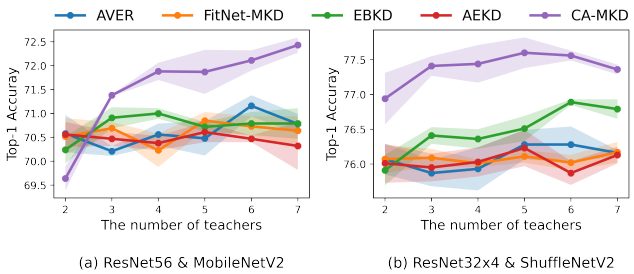

Fig. 4. The effect of different teacher numbers.

all competitors cross various architectures. Specifically, compared to the second best method (EBKD), CA-MKD outperforms it with 0.81% average improvement[1], and achieves 1.66% absolute accuracy improvement in the best case.

To verify the benefits of diverse information brought by multiple teachers, we compare CA-MKD with some excellent single-teacher based methods. The results in Table 2 show the student indeed has the potential to learn knowledge from multiple teachers, and its accuracy is further improved compared with the single-teacher methods to a certain extent.

4.2. Results on the Different Teacher Architectures

Table 3 shows the results of training a student (VGG8) with three different teacher architectures, i.e., ResNet8x4, ResNet20x4 and ResNet32x4. We find the student accuracy becomes even higher than that of training with three ResNet32x4 teachers, which may be attributed to that the knowledge diversity is enlarged in different architectures.

Since the performance of ResNet20x4/ResNet32x4 is better than that of ResNet8x4, we could reasonably believe that for most training samples, the student will put larger weights on predictions from the former two rather than the latter one, which is verified in Figure 3. Moreover, our CA-MKD can capture those samples on which the predictions are more confident by ResNet8x4, and assign them dynamic weights to help the student model achieve better performance.

4.3. Impact of the Teacher Number

As shown in Figure 4, the student model trained with CA-MKD generally achieves satisfactory results. For example, on the "ResNet56 & MobileNetV2" setting, the accuracy of CA-MKD increases continually as the number of teachers increases and it surpasses the competitors with three teachers even those competitors are trained with more teachers.

[1]Average Improvement=$\frac{1}{n}\sum_i^n \left(Acc^i_{\mathrm{CA-MKD}} - Acc^i_{\mathrm{EBKD}}\right)$, where the accuracies of CA-MKD, EBKD in the i-th teacher-student combination are denoted as $Acc^i_{\mathrm{CA-MKD}}$, Acc^i_{EBKD}, respectively.

Table 4. Ablation study with VGG13 & MobileNetV2.

avg weight	w/o $\mathcal{L}_{inter}$	w/o w^k_{inter}	CA-MKD
67.74±0.87	68.11±0.02	68.82±0.63	**69.41±0.20**

4.4. Ablation Study

We summarize the observations from Table 4 as follows:

(1) avg weight. Simply averaging multiple teachers will cause 1.67% accuracy drop, which confirms the necessity of treating different teachers based on their specific quality.

(2) w/o $\mathcal{L}_{inter}$. The accuracy will appear considerably reduction as we remove the Equation (7), demonstrating the intermediate layer contains useful information for distillation.

(3) w/o w^k_{inter}. we directly use the w^k_{KD} obtained from the last layer to integrate intermediate features. The lower result indicates the benefits of designing a separate way of calculating weights for the intermediate layer.

5. CONCLUSION

In this paper, we introduce confidence-aware mechanism on both predictions and intermediate features for multi-teacher knowledge distillation. The confidence of teachers is calculated based on the closeness between their predictions or features and the ground-truth labels for the reliability identification on each training sample. With the guidance of labels, our technique effectively integrates diverse knowledge from multiple teachers for the student training. Extensive empirical results show that our method outperforms all competitors in various teacher-student architectures.

6. REFERENCES

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[2] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al., "Mastering the game of go without human knowledge," *Nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186.

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

[5] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio, "Fitnets: Hints for thin deep nets," in *International Conference on Learning Representations*, 2015.

[6] Sergey Zagoruyko and Nikos Komodakis, "Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer," in *International Conference on Learning Representations*, 2017.

[7] Defang Chen, Jian-Ping Mei, Yuan Zhang, Can Wang, Zhe Wang, Yan Feng, and Chun Chen, "Cross-layer distillation with semantic calibration," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 7028–7036.

[8] Shan You, Chang Xu, Chao Xu, and Dacheng Tao, "Learning from multiple teacher networks," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 1285–1294.

[9] Takashi Fukuda, Masayuki Suzuki, Gakuto Kurata, Samuel Thomas, Jia Cui, and Bhuvana Ramabhadran, "Efficient knowledge distillation from an ensemble of teachers.," in *Interspeech*, 2017, pp. 3697–3701.

[10] Meng-Chieh Wu, Ching-Te Chiu, and Kun-Hsuan Wu, "Multi-teacher knowledge distillation for compressed video action recognition on deep neural networks," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2202–2206.

[11] Shangchen Du, Shan You, Xiaojie Li, Jianlong Wu, Fei Wang, Chen Qian, and Changshui Zhang, "Agree to disagree: Adaptive ensemble knowledge distillation in gradient space," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[12] Kisoo Kwon, Hwidong Na, Hoshik Lee, and Nam Soo Kim, "Adaptive knowledge distillation based on entropy," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7409–7413.

[13] Jimmy Ba and Rich Caruana, "Do deep nets really need to be deep?," in *Advances in Neural Information Processing Systems*, 2014, pp. 2654–2662.

[14] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai, "Variational information distillation for knowledge transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9163–9171.

[15] Yonglong Tian, Dilip Krishnan, and Phillip Isola, "Contrastive representation distillation," in *International Conference on Learning Representations*, 2020.

[16] Xu Lan, Xiatian Zhu, and Shaogang Gong, "Knowledge distillation by on-the-fly native ensemble," *arXiv preprint arXiv:1806.04606*, 2018.

[17] Defang Chen, Jian-Ping Mei, Can Wang, Yan Feng, and Chun Chen, "Online knowledge distillation with diverse peers.," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 3430–3437.

[18] Yuang Liu, Wei Zhang, and Jun Wang, "Adaptive multi-teacher multi-level knowledge distillation," *Neurocomputing*, vol. 415, pp. 106–113, 2020.

[19] Alex Krizhevsky and Geoffrey Hinton, "Learning multiple layers of features from tiny images," *Technical Report*, 2009.