MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Iterative Self Knowledge Distillation — From Pothole Classification To Fine-Grained And COVID Recognition

Peng, Kuan-Chuan

TR2022-020 March 05, 2022

Abstract

Pothole classification has become an important task for road inspection vehicles to save drivers from potential car accidents and repair bills. Given the limited computational power and fixed number of training epochs, we propose iterative self knowledge distillation (ISKD) to train lightweight pothole classifiers. Designed to improve both the teacher and student models over time in knowledge distillation, ISKD outperforms the state-of-the-art self knowledge distillation method on three pothole classification datasets across four lightweight network architectures, which supports that self knowledge distillation should be done iteratively instead of just once. The accuracy relation between the teacher and student models shows that the student model can still benefit from a moderately trained teacher model. Implying that better teacher models generally produce better student models, our results justify the design of ISKD. In addition to pothole classification, we also demonstrate the efficacy of ISKD on six additional datasets associated with generic classification, fine-grained classification, and medical imaging application, which supports that ISKD can serve as a general-purpose performance booster without the need of a given teacher model and extra trainable parameters.

IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2022

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

ITERATIVE SELF KNOWLEDGE DISTILLATION — FROM POTHOLE CLASSIFICATION TO FINE-GRAINED AND COVID RECOGNITION

Kuan-Chuan Peng

Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA

kpeng@merl.com

ABSTRACT

Pothole classification has become an important task for road inspection vehicles to save drivers from potential car accidents and repair bills. Given the limited computational power and fixed number of training epochs, we propose iterative self knowledge distillation (ISKD) to train lightweight pothole classifiers. Designed to improve both the teacher and student models over time in knowledge distillation, ISKD outperforms the state-of-the-art self knowledge distillation method on three pothole classification datasets across four lightweight network architectures, which supports that self knowledge distillation should be done iteratively instead of just once. The accuracy relation between the teacher and student models shows that the student model can still benefit from a moderately trained teacher model. Implying that better teacher models generally produce better student models, our results justify the design of ISKD. In addition to pothole classification, we also demonstrate the efficacy of ISKD on six additional datasets associated with generic classification, fine-grained classification, and medical imaging application, which supports that ISKD can serve as a general-purpose performance booster without the need of a given teacher model and extra trainable parameters.

***Index Terms*—** Teacher-free knowledge distillation, iterative self knowledge distillation

1. INTRODUCTION

Detecting potholes is essential for the municipalities and road authorities to repair defective roads. The vehicle repair bills related to pothole damage have cost U.S. drivers $3 billion annually on average [1]. Due to the cost constraint of the edge devices which the pothole classifiers run on, the edge devices installed on the road inspection vehicles may only have limited computational power (*e.g.*, no GPU). In such scenarios, lightweight models are needed if real-time inference speed is required. Motivated by this application and deployment time constraint of pothole classifiers, we focus on the following problem: *Given a fixed number of training epochs and a lightweight model to be trained, what can practitioners do to improve the pothole classification accuracy*?

Given the problem, we explore *self* knowledge distillation (KD) [2] to tackle pothole classification. By *self* KD, we refer to the KD methods which need no teacher model in advance and introduce no extra trainable parameters. Showing that KD is actualy learned label smoothing regularization, Yuan *et al*. [2] propose Tf-KD$_{self}$, the teacher-free KD by using the pre-trained student model itself as the teacher model. Inspired by [2] and the assumption that better teacher models result in better student models, we propose *iterative self knowledge distillation (ISKD)*, which iteratively performs self KD by using the pre-trained student model as the teacher model.

Most KD methods [3, 4, 5, 6, 7] typically require that the teacher model is available in advance, which is not always true. Even if there are KD methods which need no teacher model [2, 8, 9], these methods typically do not experiment on lightweight models or perform KD multiple times. Utilizing KD iteratively and requiring no teacher model, our proposed ISKD shows its efficacy on lightweight models for pothole classification, generic classification, fine-grained classification, and medical imaging application. Although there exist methods utilizing iterative KD [10, 11, 12], they typically require additional constraints and are validated on only few datasets. For example, Koutini's method [10] requires training multiple models and selecting the best trained model for each class to predict pseudo labels for sound event detection. In contrast, our proposed ISKD only needs to train one model at once with no need to select models and predict pseudo labels. In [11, 12], their teacher model is trained until convergence for each KD iteration, and their methods are validated on only few datasets. In addition, the accuracy gain of [11] comes from the ensemble of all the students in the history, which needs more deployment space at testing time. In contrast, we show ISKD's efficacy on a wide variety of datasets even when the teacher model is only moderately trained, and ISKD does not rely on the ensemble, which is more practical for embedded devices.

To the best of our knowledge, we are the first in pothole classification to use iterative self knowledge distillation when training lightweight neural networks under limited training epochs. We make the following contributions:
(1) We propose iterative self knowledge distillation (ISKD), which outperforms the state-of-the-art self KD method Tf-

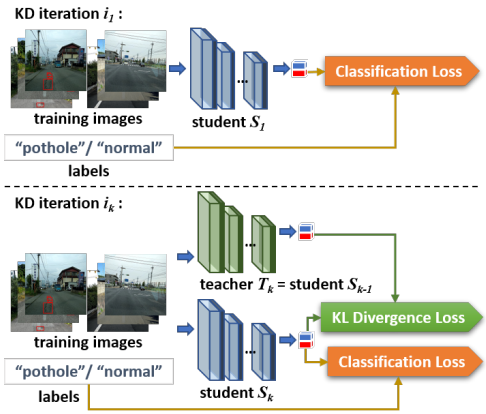

Fig. 1: Our proposed iterative self knowledge distillation.

KD_{self} [2] on the road damage dataset [13], the Nienaber potholes simplex [14] and complex [15] datasets, CIFAR-10 [16], CIFAR-100 [16], Oxford 102 Flower [17], Oxford-IIIT Pet Dataset [18], Caltech-UCSD Birds 200 [19], and COVID-19 Radiography [20] datasets, which supports the wide applicability of ISKD from pothole classification to generic, fine-grained, and medical imaging classification.
(2) We provide more evidence showing that even using a teacher model with accuracy lower than the baseline accuracy from a classifier trained with a larger number of epochs, the student model can still possibly outperform the baseline.
(3) ISKD can outperform the baseline under a wide range of weight balancing the objectives of ISKD, which supports that ISKD is flexible with respect to parameter selection.

2. ITERATIVE SELF KNOWLEDGE DISTILLATION

Inspired by Yuan *et al*. [2], we propose iterative self knowledge distillation (ISKD) such that both teacher and student models can improve over time. We illustrate ISKD in Fig. 1, where we denote the teacher/student model in the k-th KD iteration i_k as T_k/S_k. During i_1, since T_1 is not given in advance, we train S_1 using the softmax cross-entropy loss L_c as the classification loss. During i_k $(k > 1, k \in \mathbb{N})$, we use the trained student model in i_{k-1} (*i.e.*, S_{k-1}) as T_k, and train S_k with both L_c and the Kullback-Leibler (KL) divergence loss. Specifically, the total loss function to train S_k during i_k can be written as $L_{KD} = (1-\alpha)\, L_c + \alpha KLD\,(\mathbf{z}, \mathbf{z}^t)$, where KLD is the KL divergence, $\mathbf{z}^t$/$\mathbf{z}$ is the output probability distribution of T_k/S_k, and α is the weight of KLD.

During i_k, we freeze the parameters of T_k and only train S_k. We pre-train S_k from ImageNet [21], not from S_{k-1} because we hope to decrease the chance that S_k is trapped from the possibly local optimum associated with S_{k-1}. ISKD stops at i_k if S_k shows no obvious accuracy gain over S_{k-1}. Since Yuan *et al*. [2] show that to benefit the student model, the teacher model need not outperform the student model, we directly use the previously trained student model as the current teacher model, waiving the typical KD requirement that the teacher model is needed in advance. We expect that using ISKD improves both T_k and S_k when k increases under the assumption that using better teacher models in KD generally results in better student models.

3. EXPERIMENTAL SETUP

We experiment on road damage dataset (termed as RDD) [13], Nienaber potholes simplex (termed as simplex) [14] and complex (termed as complex) [15] datasets, CIFAR-10 [16], CIFAR-100 [16], Oxford 102 Flower dataset (termed as Oxford-102) [17], Oxford-IIIT Pet dataset (termed as Oxford-37) [18], Caltech-UCSD Birds 200 dataset (termed as CUB-200) [19], and COVID-19 Radiography dataset (termed as COVID) [20]. We choose these datasets to cover a diverse range of task domains from pothole, generic, fine-grained to medical imaging classification. The RDD, simplex, and complex datasets provide annotations of whether each image contains any pothole or not. The Oxford-102, Oxford-37, and CUB-200 datasets provide the images and labels of 102, 37, and 200 different species of flowers, cats and dogs, and birds, respectively. The COVID dataset includes 4 different types of chest x-rays: normal, COVID, lung opacity, and viral pneumonia. For all the datasets except COVID, we use the official training/testing split of each dataset. Since the COVID dataset does not provide the official training/testing split, we randomly generate the split using the ratio of 7:3.

We use the official PyTorch [22] implementation of ResNet-18 [23], SqueezeNet v1.1 [24], and ShuffleNet v2 x0.5 & x1.0 [25], modify their last layers such that the number of output nodes of the last layer equals the number of classes, and pre-train them from the ImageNet [21]. The four network architectures are selected based on the following criteria: (1) For the ease of reproducibility, they are officially supported by PyTorch [22], which provides their weights pre-trained from the ImageNet [21]. (2) Considering typically limited computational power on edge devices, we limit the number of network parameters to be fewer than 12M.

We first experiment on the four network architectures mentioned previously using the RDD [13], simplex [14], and complex [15] datasets. For each KD iteration, we use the same network architecture for both teacher and student models. We compare ISKD with the following two baselines with the same network architecture, total number of training epochs, and learning schedule: (1) training a classifier using L_c without KD (termed as the large-epoch baseline), and (2) Tf-KD_{self} [2], which only performs KD once without multiple KD iterations. For the extended study involving the other six datasets irrelevant to potholes, we use the ResNet-18 [23] and ShuffleNet v2 x1.0 [25] as the network architectures of ISKD. We conduct the extended study in the same way as the pothole classification task mentioned previously using the same two baselines.

For all the experiments, the training images are resized to

dataset	experiment ID model \ KD iteration	E_1 i_1 (no KD)	E_2 i_2	E_3 i_3	E_4 i_4	E_5 i_5	E_6 i_6	E_7 $i_1 \sim i_6$	E_8 i_1 (large-epoch)	E_9 i_1 + Tf-KD$_{self}$ [2]
RDD [13]	ResNet-18 [23]	91.54_{50}	92.71_{50}	92.99_{50}	93.04_{50}	$\mathbf{93.08}_{50}$	n/a	$\mathbf{93.08}_{250}$	92.24_{250}	92.34_{250}
	SqueezeNet v1.1 [24]	89.67_{50}	89.91_{50}	90.28_{50}	90.51_{50}	$\mathbf{90.70}_{50}$	$\mathbf{90.70}_{50}$	$\mathbf{90.70}_{300}$	90.47_{300}	90.28_{300}
	ShuffleNet v2 x0.5 [25]	90.05_{50}	90.14_{50}	90.98_{50}	91.40_{50}	91.40_{50}	n/a	91.40_{250}	$\mathbf{92.66}_{250}$	91.22_{250}
	ShuffleNet v2 x1.0 [25]	92.01_{50}	92.15_{50}	92.66_{50}	$\mathbf{93.22}_{50}$	$\mathbf{93.22}_{50}$	n/a	$\mathbf{93.22}_{250}$	93.13_{250}	93.13_{250}
simplex [14]	ResNet-18 [23]	81.85_{50}	90.46_{50}	93.69_{50}	96.92_{50}	98.62_{50}	$\mathbf{99.08}_{50}$	$\mathbf{99.08}_{300}$	83.38_{300}	92.00_{300}
	SqueezeNet v1.1 [24]	81.85_{50}	86.77_{50}	87.69_{50}	$\mathbf{88.15}_{50}$	$\mathbf{88.15}_{50}$	n/a	$\mathbf{88.15}_{250}$	84.62_{250}	87.69_{250}
	ShuffleNet v2 x0.5 [25]	90.00_{50}	95.69_{50}	99.38_{50}	$\mathbf{100.00}_{50}$	n/a	n/a	$\mathbf{100.00}_{200}$	92.46_{200}	97.54_{200}
	ShuffleNet v2 x1.0 [25]	90.31_{50}	96.31_{50}	98.77_{50}	$\mathbf{100.00}_{50}$	n/a	n/a	$\mathbf{100.00}_{200}$	93.38_{200}	97.54_{200}
complex [15]	ResNet-18 [23]	61.92_{50}	74.14_{50}	77.65_{50}	79.80_{50}	83.77_{50}	$\mathbf{84.93}_{50}$	$\mathbf{84.93}_{300}$	62.58_{300}	82.95_{300}
	SqueezeNet v1.1 [24]	59.27_{50}	70.70_{50}	$\mathbf{76.16}_{50}$	$\mathbf{76.16}_{50}$	n/a	n/a	$\mathbf{76.16}_{200}$	62.25_{200}	71.03_{200}
	ShuffleNet v2 x0.5 [25]	56.79_{50}	78.97_{50}	88.58_{50}	$\mathbf{88.91}_{50}$	n/a	n/a	$\mathbf{88.91}_{200}$	65.89_{200}	79.80_{200}
	ShuffleNet v2 x1.0 [25]	60.43_{50}	78.31_{50}	$\mathbf{86.59}_{50}$	86.42_{50}	n/a	n/a	86.42_{200}	71.85_{200}	82.45_{200}
CIFAR-10 [26]	ResNet-18 [23]	90.75_{50}	91.81_{50}	92.05_{50}	$\mathbf{92.07}_{50}$	n/a	n/a	$\mathbf{92.07}_{200}$	91.53_{200}	92.01_{200}
	ShuffleNet v2 x1.0 [25]	82.63_{50}	85.16_{50}	88.45_{50}	90.30_{50}	91.03_{50}	$\mathbf{91.58}_{50}$	$\mathbf{91.58}_{300}$	90.68_{300}	85.55_{300}
CIFAR-100 [26]	ResNet-18 [23]	80.15_{50}	81.05_{50}	81.64_{50}	82.17_{50}	82.30_{50}	$\mathbf{82.67}_{50}$	$\mathbf{82.67}_{300}$	81.34_{300}	81.62_{300}
	ShuffleNet v2 x1.0 [25]	58.95_{50}	65.60_{50}	72.16_{50}	75.59_{50}	77.27_{50}	$\mathbf{77.85}_{50}$	$\mathbf{77.85}_{300}$	77.61_{300}	65.78_{300}
Oxford-102 [17]	ResNet-18 [23]	96.58_{50}	97.31_{50}	97.68_{50}	$\mathbf{97.80}_{50}$	n/a	n/a	$\mathbf{97.80}_{200}$	96.94_{200}	97.43_{200}
	ShuffleNet v2 x1.0 [25]	94.74_{50}	97.19_{50}	98.17_{50}	$\mathbf{98.41}_{50}$	n/a	n/a	$\mathbf{98.41}_{200}$	$\mathbf{98.41}_{200}$	97.19_{200}
Oxford-37 [18]	ResNet-18 [23]	90.57_{50}	90.98_{50}	91.33_{50}	$\mathbf{91.80}_{50}$	91.58_{50}	n/a	$\mathbf{91.80}_{200}$	91.20_{200}	91.31_{200}
	ShuffleNet v2 x1.0 [25]	79.69_{50}	84.30_{50}	85.96_{50}	$\mathbf{86.59}_{50}$	$\mathbf{86.59}_{50}$	n/a	$\mathbf{86.59}_{250}$	86.10_{250}	84.46_{250}
CUB-200 [19]	ResNet-18 [23]	42.07_{50}	46.49_{50}	48.66_{50}	$\mathbf{49.82}_{50}$	49.49_{50}	n/a	$\mathbf{49.82}_{200}$	46.03_{200}	47.58_{200}
	ShuffleNet v2 x1.0 [25]	41.54_{50}	45.40_{50}	48.53_{50}	49.69_{50}	$\mathbf{50.28}_{50}$	49.95_{50}	$\mathbf{50.28}_{250}$	48.99_{250}	46.95_{250}
COVID [20]	ResNet-18 [23]	94.11_{50}	94.68_{50}	94.80_{50}	$\mathbf{95.01}_{50}$	n/a	n/a	$\mathbf{95.01}_{200}$	94.79_{200}	94.88_{200}
	ShuffleNet v2 x1.0 [25]	90.76_{50}	92.43_{50}	93.07_{50}	93.81_{50}	$\mathbf{93.98}_{50}$	n/a	$\mathbf{93.98}_{250}$	92.72_{250}	93.10_{250}

Table 1: Comparing the classification accuracy (%) of the iterative self knowledge distillation (KD) method versus the baselines. The numbers are in the format of [accuracy]$_{[e_s]}$, where e_s is the number of epochs which the student model is trained for.

224×224, and the model is pre-trained from ImageNet [21] and fine-tuned with the training data of each dataset. We use the momentum 0.9, weight decay 5e-4, batch size 128, and the SGD optimizer to train the student model for 50 epochs for each KD iteration, and the learning rate is fixed within each KD iteration but model-specific during training. For ResNet-18 [23] and SqueezeNet v1.1 [24], we use the initial learning rate 0.001, but for ShuffleNet v2 x0.5 & x1.0 [25], we use the initial learning rate 0.1. Following Tf-KD$_{self}$ [2], we obtain the α values by grid search on the validation data sampled from the training set when experimenting on the RDD dataset [13]. Once we determine the α values from the RDD dataset, we fix the α values and use the same set of α values when experimenting on other datasets (*i.e.*, the α values are not tuned for most of the datasets except the RDD dataset). We purposely do so to test whether the α values searched from one dataset can be transferable and directly applied to other datasets. For all the other parameters, we use the default PyTorch [22] setting unless otherwise specified.

4. EXPERIMENTAL RESULT

The experimental results are summarized in Table 1, where we refer to each column by the experiment ID $E_1 \sim E_9$. All the numbers are reported in the format of [accuracy]$_{[e_s]}$, where e_s is the number of epochs which the student model is trained for. $E_1 \sim E_6$ list the performance of $S_1 \sim S_6$, and E_7 summarizes the last performance after $i_1 \sim i_6$. There is no teacher model for E_1 and the large-epoch baseline (E_8), but for the baseline using Tf-KD$_{self}$ [2] (E_9), the teacher model is the trained S_1. All the student models are pre-trained from ImageNet [21] for ISKD (E_7) and the two baselines (E_8, E_9).

dataset\method	ISKD	prior work (backbone)
CIFAR-100 [26]	**82.67**	81.60 [27] (Wide-ResNet-28-10)
Oxford-102 [17]	**97.80**	91.10 [28] (ResNet152-SAM)
Oxford-37 [18]	**91.80**	91.60 [28] (ResNet50-SAM)

Table 2: The comparison of classification accuracy (%) between ISKD (backbone: ResNet-18 [23]) and the prior works using backbones with more parameters.

In Table 1, the accuracy of E_2 is higher than that of E_1, which shows that self KD can improve the student model's accuracy. The accuracy of E_p is higher than that of E_q for most cases when $2 \leq q < p \leq 6$, which validates the assumption that in self KD, better teacher models result in better student models and that self KD can be done iteratively instead of just once. Comparing E_7 with E_8 and E_9, we show that given a fixed number of training epochs, ISKD outperforms the large-epoch baseline and the state-of-the-art self KD method Tf-KD$_{self}$ [2] in most cases. The fact that E_7 outperforms E_9 is also an ablation study supporting that

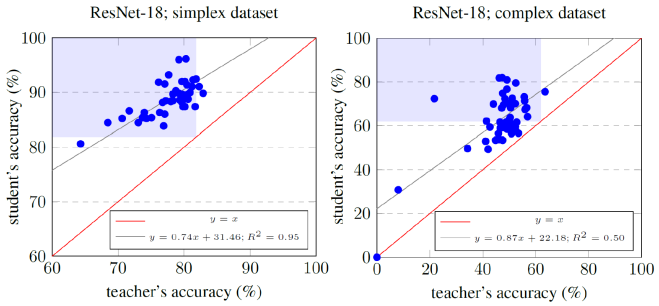

Fig. 2: The teacher-student accuracy relation on the simplex [14] and complex [15] datasets using ResNet-18 [23]. The gray lines are obtained from linear regression, and the line equation and Pearson's correlation coefficient R are marked in the legend. Given the baseline performance in E_1 of Table 1, the shaded blue areas are the areas where the teacher model performs worse than the baseline but the student model outperforms the baseline.

self KD is better done iteratively than just once.

Since Table 1 covers diverse task domains, including pothole, generic, fine-grained, and medical imaging classification, our results support that ISKD can serve as a general-purpose performance booster. In addition, we obtain the results in Table 1 by directly using the α values chosen for pothole classification *without* tuning the α values for each dataset, which supports that the α values we use are transferable across different datasets. We also compare the accuracy of ISKD (backbone: ResNet-18 [23]) with the prior methods which use the backbones with more parameters in Table 2, where ISKD performs on par or even outperforms the listed methods which use more parameters. This finding supports that ISKD is more parameter efficient than the listed methods.

Furthermore, we analyze what is the worst performing teacher model in *self* KD which can still make the student model outperform the baseline with no KD (E_1 in Table 1). To gain more insight, we design the following experiment to find the accuracy relation between the teacher and student models. Given that E_1 in Table 1 is trained for 50 epochs, we use the 50 models saved after each epoch is completed as the teacher models. We repeat E_2 in Table 1 for 50 times (each time with one of the 50 teacher models produced during E_1), and record the teacher-student accuracy relation. We perform this experiment on the simplex [14] and complex [15] datasets using ResNet-18 [23] as the network architectures.

We show the result of teacher-student accuracy relation in Fig. 2. Most of the data points are above the red line ($x = y$), which supports that performing self KD can make the student model outperform the teacher model. Fig. 2 suggests that the accuracy of the teacher and student models has strong positive correlation (the slopes of the gray lines are positive and the $R^2 \geq 0.5$), which again supports the assumption that using better teacher models in KD generally results in better student models. We find that the number of data points falling into the shaded blue areas is not negligible, which serves as statistically more significant evidence than [2] supporting that even

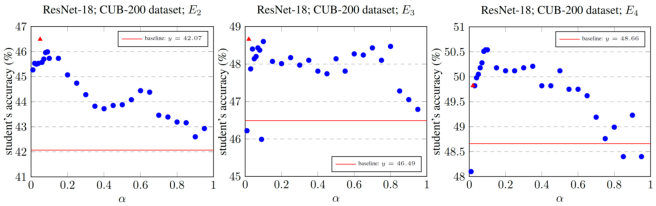

Fig. 3: The accuracy of the student model using ResNet-18 under different α values when we repeat E_2, E_3, and E_4 on the CUB-200 dataset. The triangular red points are the accuracy we report in Table 1 by using the α values chosen for pothole classification. Shown as the red lines, the baseline for E_i is the accuracy of the student model in E_{i-1} reported in Table 1.

if the teacher model is worse than the baseline, it is still likely that the student model can outperform the baseline after KD.

Another experiment which is also not presented in the paper of Yuan *et al*. [2] is the impact of the α values on the accuracy. To study this, we use the ResNet-18 [23] as the network architecture of ISKD and the same random seed 1, repeat E_2, E_3, and E_4 corresponding to the CUB-200 [19] dataset with different α values, and report the student's accuracy in Fig. 3, where the red lines mark the baseline accuracy (*i.e*., the student's accuracy in the previous KD iteration reported in Table 1). For each sub-figure of Fig. 3, the teacher model is fixed as the corresponding one used in Table 1. The triangular points in Fig. 3 are the accuracy reported in Table 1 using the α values chosen for pothole classification, so their corresponding accuracy is not necessarily the best. Fig. 3 shows that the student model outperforms the baseline in each KD iteration of ISKD under a wide range of α values, which supports that ISKD is flexible in terms of parameter selection.

5. CONCLUSION

We propose iterative self knowledge distillation (ISKD) in self KD to improve pothole classification accuracy when training lightweight models given a fixed amount of training epochs. Experimenting on three pothole classification datasets and six other datasets associated with generic classification, fine-grained classification, and medical imaging application, we show that ISKD outperforms the state-of-the-art self KD method Tf-KD$_{self}$, for most cases, given the same number of training epochs and that ISKD has wide applicability to various tasks without the need of a given teacher model and extra trainable parameters. In addition, we show more evidence supporting that the performance of the student model can benefit from self KD even when the pre-trained student model (which serves as the teacher model) is only moderately trained. Our study on the impact of the weight balancing the objectives of ISKD shows that even if we choose different weights deviating from the weights we initially use within a reasonable range, the student model can still improve over KD iterations, which supports that ISKD is flexible in terms of parameter selection.

6. REFERENCES

[1] "American Automobile Association (AAA) pothole fact sheet," http://publicaffairsresources.aaa.biz/wp-content/uploads/2016/02/Pothole-Fact-Sheet.pdf, 2016.

[2] L. Yuan, F. EH Tay, G. Li, T. Wang, and J. Feng, "Revisiting knowledge distillation via label smoothing regularization," in *CVPR*, June 2020.

[3] N. Passalis, M. Tzelepi, and A. Tefas, "Heterogeneous knowledge distillation using information flow modeling," in *CVPR*, June 2020.

[4] L. Zhao, X. Peng, Y. Chen, M. Kapadia, and D. N. Metaxas, "Knowledge as priors: Cross-modal knowledge generalization for datasets without superior knowledge," in *CVPR*, June 2020.

[5] G. Xu, Z. Liu, X. Li, and C. C. Loy, "Knowledge distillation meets self-supervision," in *ECCV*, August 2020.

[6] B. Peng, X. Jin, J. Liu, D. Li, Y. Wu, Y. Liu, S. Zhou, and Z. Zhang, "Correlation congruence for knowledge distillation," in *ICCV*, October 2019.

[7] X. Jin, B. Peng, Y. Wu, Y. Liu, J. Liu, D. Liang, J. Yan, and X. Hu, "Knowledge distillation via route constrained optimization," in *ICCV*, October 2019.

[8] Q. Guo, X. Wang, Y. Wu, Z. Yu, D. Liang, X. Hu, and P. Luo, "Online knowledge distillation via collaborative learning," in *CVPR*, June 2020.

[9] S. Yun, J. Park, K. Lee, and J. Shin, "Regularizing class-wise predictions via self-knowledge distillation," in *CVPR*, June 2020.

[10] K. Koutini, H. Eghbal-Zadeh, and G. Widmer, "Iterative knowledge distillation in R-CNNs for weakly-labeled semi-supervised sound event detection," in *Detection and Classification of Acoustic Scenes and Events*, 2018.

[11] T. Furlanello, Z. C. Lipton, M. Tschannen, L. Itti, and A. Anandkumar, "Born again neural networks," in *ICML*, 2018.

[12] L. Zhang, D. Du, C. Li, Y. Wu, and T. Luo, "Iterative knowledge distillation for automatic check-out," *IEEE Transactions on Multimedia*, 2020.

[13] H. Maeda, Y. Sekimoto, T. Seto, T. Kashiyama, and H. Omata, "Road damage detection and classification using deep neural networks with smartphone images," *Computer-Aided Civil and Infrastructure Engineering*, vol. 33, 06 2018.

[14] "Kaggle dataset: Nienaber potholes 1 simplex," https://www.kaggle.com/felipemuller5/nienaber-potholes-1-simplex.

[15] "Kaggle dataset: Nienaber potholes 2 complex," https://www.kaggle.com/felipemuller5/nienaber-potholes-2-complex.

[16] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," *Technical report*, 2009.

[17] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008.

[18] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar, "Cats and dogs," in *CVPR*, 2012.

[19] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona, "Caltech-UCSD Birds 200," Tech. Rep. CNS-TR-2010-001, California Institute of Technology, 2010.

[20] "COVID-19 radiography database," https://www.kaggle.com/tawsifurrahman/covid19-radiography-database.

[21] 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," *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.

[22] "PyTorch," https://pytorch.org/.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[24] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[25] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "ShuffleNet V2: Practical guidelines for efficient CNN architecture design," in *ECCV*, September 2018.

[26] A. Krizhevsky, "Learning multiple layers of features from tiny images," https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf, 2009.

[27] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le, "Attention augmented convolutional networks," in *ICCV*, October 2019.

[28] X. Chen, C.-J. Hsieh, and B. Gong, "When vision transformers outperform ResNets without pretraining or strong data augmentations," *arXiv preprint arXiv:2106.01548*, 2021.