Iterative Self Knowledge Distillation – From Pothole Classification to Fine-Grained and COVID Recognition

MITSUBISHI ELECTRIC
Changes for the Better

Kuan-Chuan Peng
Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA
kpeng@merl.com

Motivation, Constraints & Task

<u>Motivation</u>: ~$3 billion annual vehicle repair bills related to pothole damage in the U.S.

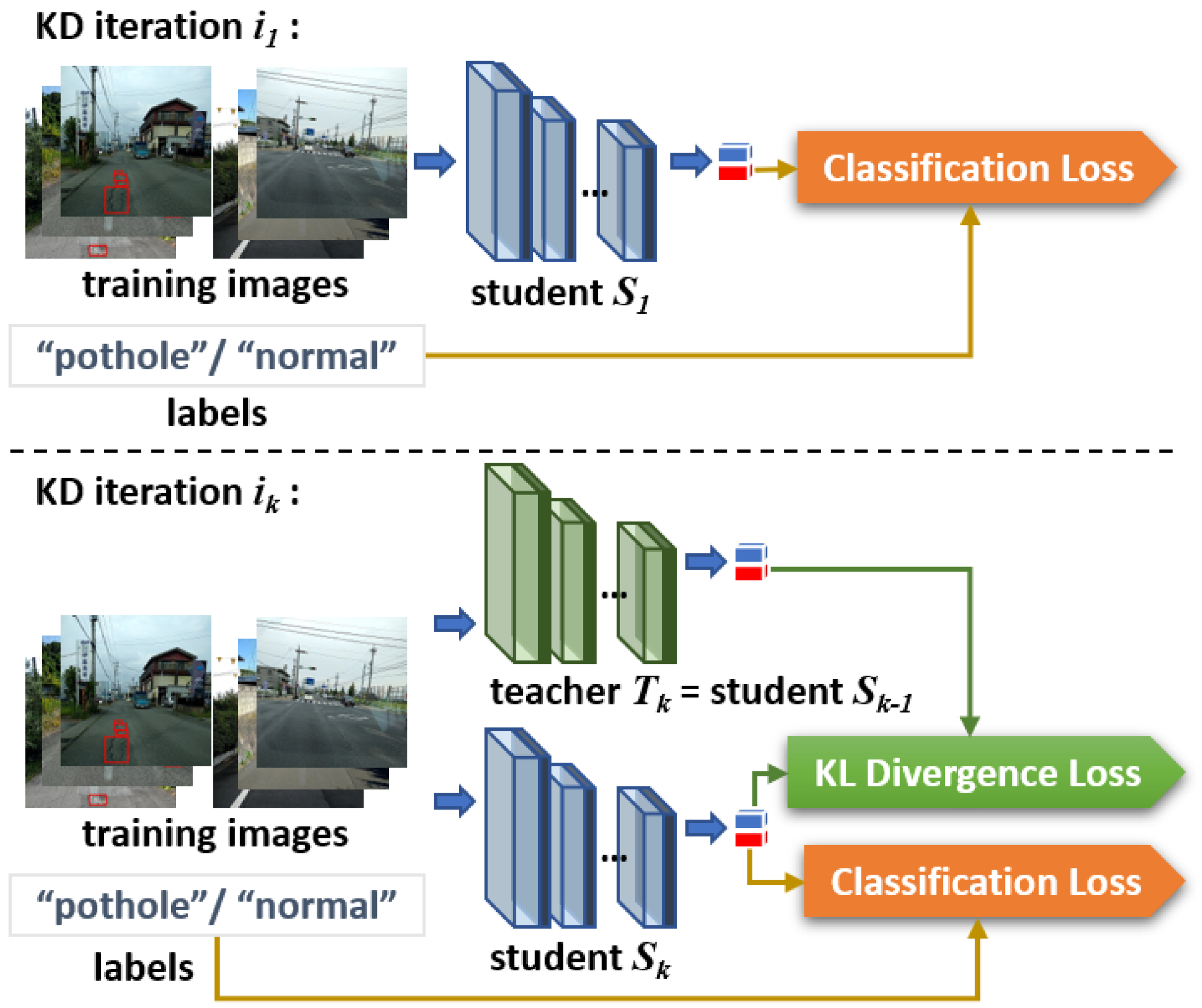
<u>Constraints</u>: limited computational power (e.g., no GPU) on edge devices installed on the road inspection vehicles; need fast inference speed.

<u>Task</u>: Given a fixed number of training epochs and a lightweight model to be trained, what can practitioners do to improve the pothole classification accuracy?

Contributions

- We propose **Iterative Self Knowledge Distillation (ISKD)**, which **outperforms the SOTA self KD methods** from **pothole classification** (RDD, simplex, complex) to **generic** (CIFAR-10, CIFAR-100), **fine-grained** (Oxford 102 Flower, Oxford IIIT Pet, Caltech-UCSD Birds 200), and **medical imaging classification** (COVID-19 Radiography).
- We provide more evidence showing that a teacher model with accuracy lower than the baseline can still result in a student model outperforming the baseline.
- ISKD is flexible with respect to parameter selection.

Iterative Self Knowledge Distillation



S_k's total loss function:

$$L_{KD} = (1-\alpha)\, L_c + \alpha KLD\,(\mathbf{z}, \mathbf{z}^t)$$

KLD: KL divergence; $\mathbf{z}^t/\mathbf{z}$: output probability distribution of T_k/S_k; α: weight of KLD.

Dataset Statistics & Quantitative Results

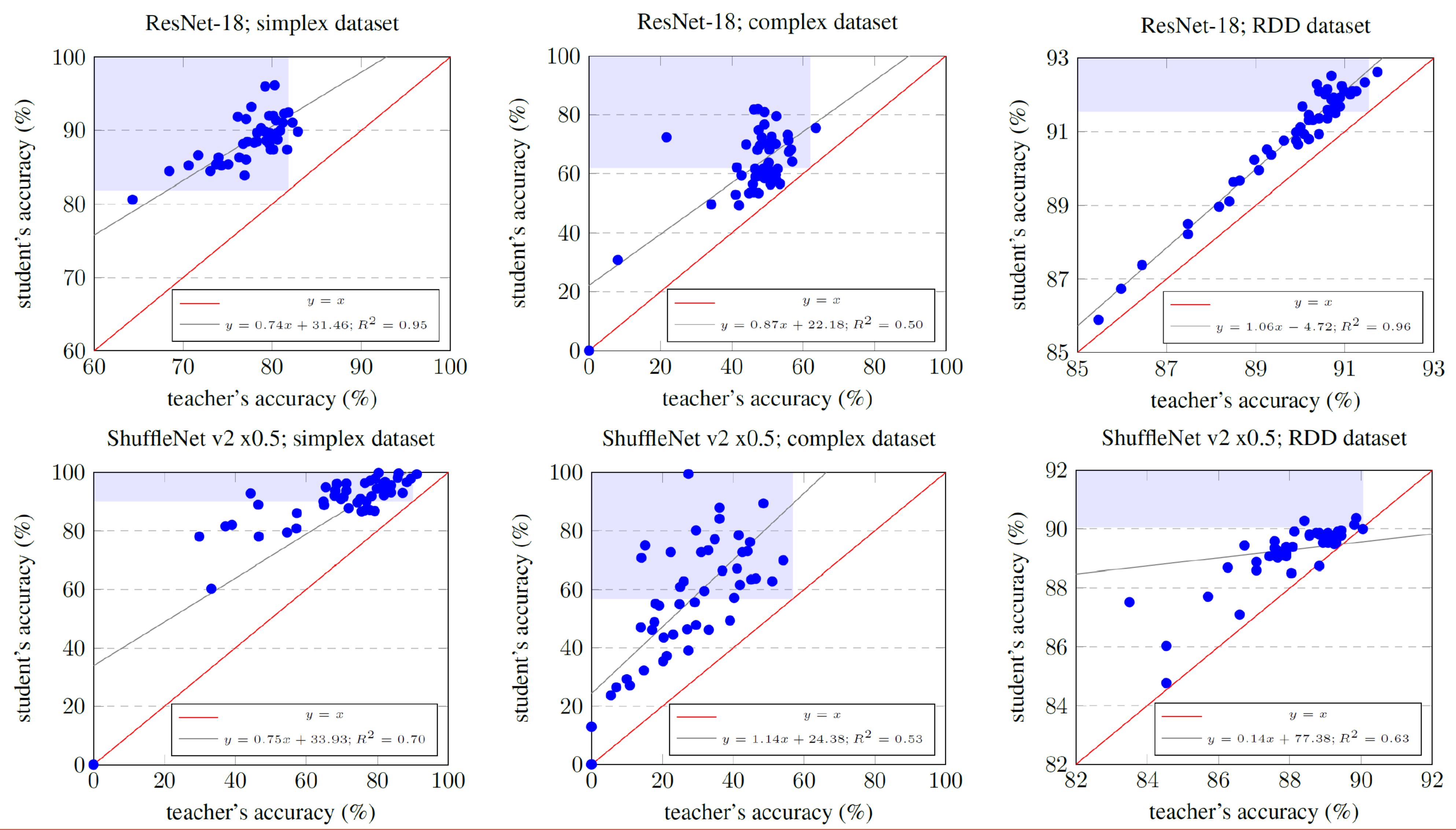
dataset	RDD	simplex	complex	CIFAR-10	CIFAR-100	Oxford-102	Oxford-37	CUB-200	COVID
# training images	8511	4736	7489	50000	50000	6552	3680	3000	14815
# testing images	2140	650	604	10000	10000	818	3669	3033	6350
# classes	2	2	2	10	100	102	37	200	4
content	pothole / normal road images			objects	objects	flowers	cats & dogs	birds	chest x-ray
task type	infrastructure maintenance			generic classification		fine-grained classification			types of pneumonia

dataset statistics

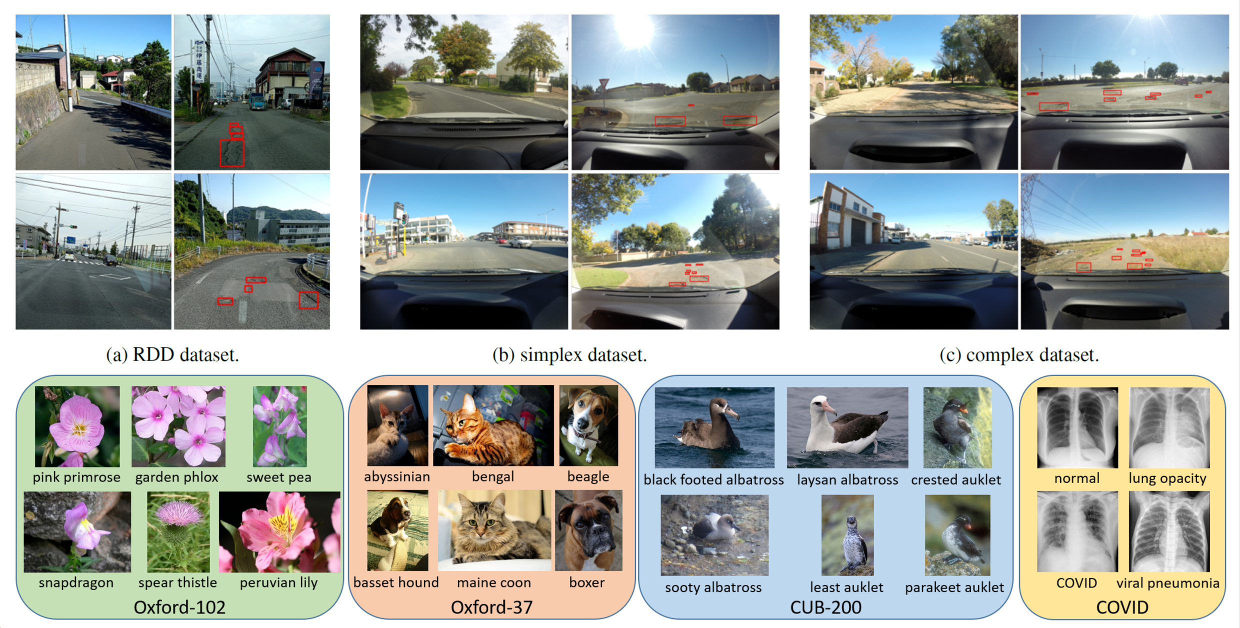
dataset	experiment ID model \ KD iteration	E_1 i_1 (no KD)	E_2 i_2	E_3 i_3	E_4 i_4	E_5 i_5	E_6 i_6	E_7 $i_1 \sim i_6$	E_8 i_1 (large-epoch)	E_9 i_1 + Tf-KD$_{self}$ [2]
RDD [13]	ResNet-18 [23]	91.54_{50}	92.71_{50}	92.99_{50}	93.04_{50}	$\mathbf{93.08}_{50}$	n/a	$\mathbf{93.08}_{250}$	92.24_{250}	92.34_{250}
	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.

Teacher-Student Accuracy Relation



Example Images of the Datasets



Parameter Efficiency

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.

Accuracy under Different α Values