Bayes Conditional Distribution Estimation for Knowledge Distillation Based on Conditional Mutual Information

Linfeng Ye*, Shayan Mohajer Hamidi*, Renhao Tan*, En-Hui Yang

Electrical and Computer Engineering

* Authors contributed equally

The Role of Teacher in Knowledge Distillation

- In knowledge distillation (KD), the role of teacher model is to provide an estimate of Bayes conditional probability distribution (BCPD) to the student [1].
- In conventional KD, the BCPD estimate is obtained by training the teacher using maximum loglikelihood (MLL) method.
- How can we train the teacher for the purpose of providing a better BCPD estimate to the student?

Prototype and Contextual Information

- In a multi-class classification task, each of the “ground truth labels” is a distinct *prototype*.
- The images with the same label are different manifestations of a common prototype, each with its own context.
- For instance, the prototype for these two images is “dog”, but the top and bottom one has some information about urban area and grass, respectively.
- As such, each training/testing image contains two types of information: (i) its prototype; and (ii) some contextual information .
- Therefore, to provide an estimate of BCPD, the teacher should be capable of providing good amount of information about both of these types of information for the input images.

Our Approach to Estimate BCPD

- A teacher trained using MLL solely aims at estimating the prototypes.
- But how can we quantify the contextual information?
- We argue that contextual information resides in conditional mutual information

$$I(X, \hat{Y}|Y) = E_{X|Y}\left[KL(P_X;\ P_{\hat{Y}|y})|Y = y\right]$$

Input Image

label predicted by the teacher

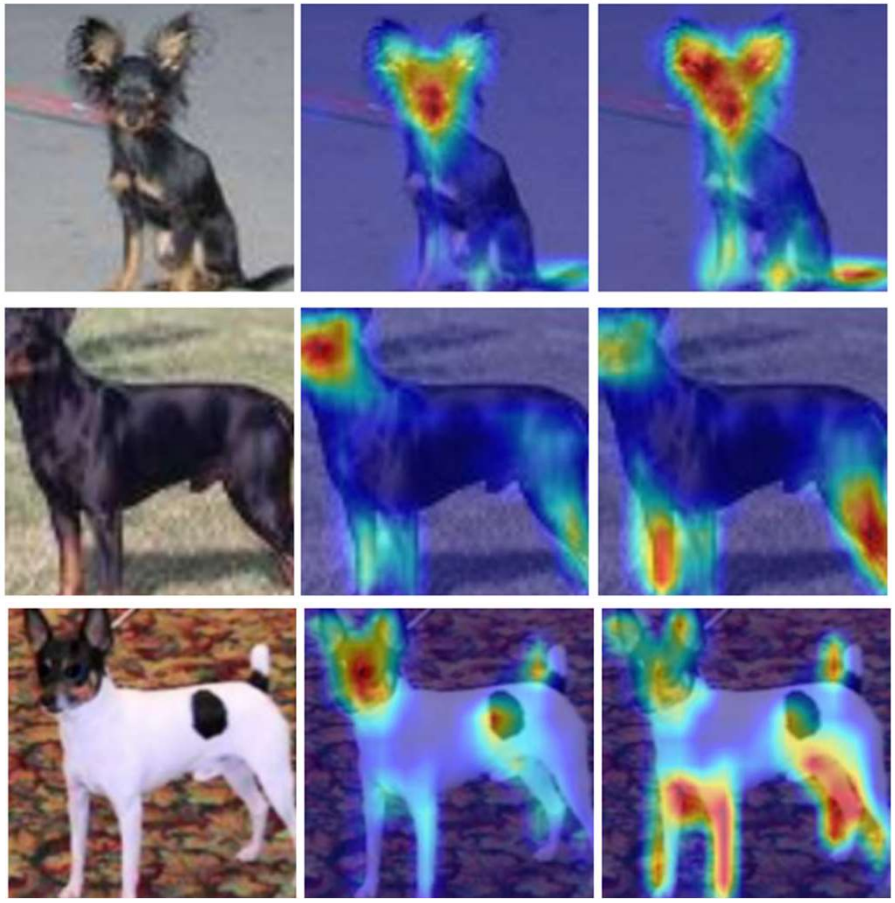
ground truth label

The DNN output to input X

- To balance the prototype and the contextual information, we train the teacher to simultaneously maximize (i) the log-likelihood (LL) of the prototype, and (ii) its CMI value:

$$\max LL(f) + I(X, \hat{Y}|Y)$$

- We call this new estimator as maximum CMI (**MCMI**) estimator.



Original MLL MCMI

Visualization of contextual information extracted by MCMI teachers.

We use Eigen-CAM to visualize and compare the feature maps extracted by the MLL teacher and MCMI teacher.

As showed in the left figure, the MLL teacher always highlight the dog's head. However, for MCMI teacher, the activation maps as a whole capture more contexture information (dogs' legs, and other surrounding information).

TOP-1 ACCURACY ON CIFAR-100

Teachers and students with **different** architectures.												
Teacher	ResNet-50		ResNet-50		ResNet-32×4		ResNet-32×4		WRN-40-2		VGG-13	
	MLL	MCMI	MLL	MCMI	MLL	MCMI	MLL	MCMI	MLL	MCMI	MLL	MCMI
Accuracy	79.34	78.45	79.34	78.45	79.41	78.70	79.41	78.70	75.61	75.21	74.64	73.96
LL	-0.004	-0.083	-0.004	-0.083	-0.004	-0.035	-0.004	-0.035	-0.052	-0.132	-0.007	-0.076
CMI	0.0085	0.1065	0.0085	0.1065	0.0059	0.0586	0.0059	0.0586	0.0255	0.1951	0.0152	0.1298
Student	MobileNetV2		VGG-8		ShuffleNetV1		ShuffleNetV2		ShuffleNetV1		MobileNetV2	
Accuracy	64.60		70.36		70.50		71.82		70.50		64.60	
KD	67.35	70.23 | 2.88	73.81	74.59 | 0.78	74.07	75.90 | 1.83	74.45	76.32 | 1.87	74.83	76.45 | 1.62	67.37	69.14 | 1.77
AT	58.58	60.03 | 1.45	71.84	72.19 | 0.35	71.73	75.05 | 3.32	72.73	75.21 | 2.48	73.32	75.61 | 2.29	59.40	62.07 | 2.67
PKT	66.52	67.42 | 0.90	73.10	73.43 | 0.33	74.10	75.21 | 1.11	74.69	76.34 | 1.65	73.89	75.39 | 1.50	67.13	68.37 | 1.24
SP	68.08	69.07 | 0.99	73.34	74.14 | 0.80	73.48	76.56 | 3.08	74.56	76.70 | 2.14	74.52	76.82 | 2.30	66.30	67.83 | 1.53
CC	65.43	66.76 | 1.33	70.25	70.90 | 0.65	71.14	71.77 | 0.63	71.29	73.02 | 1.73	71.38	71.80 | 0.42	64.86	65.45 | 0.59
RKD	64.43	65.11 | 0.68	71.50	72.10 | 0.60	72.28	73.59 | 1.31	73.21	74.67 | 1.46	72.21	74.26 | 2.05	64.52	65.37 | 0.85
VID	67.57	67.61	70.30	70.69	73.38	74.58	73.40	74.67	73.61	75.03	65.56	65.77

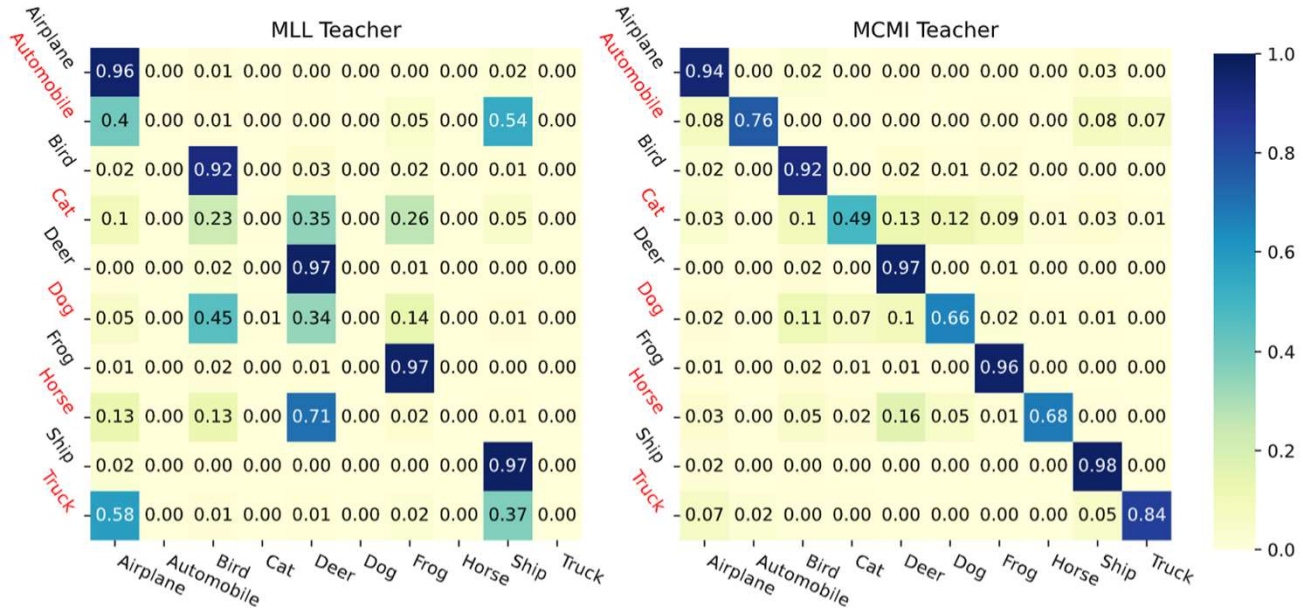
Top-1 accuracy on Cifar-100

Teachers and students with the **same** architectures.												
Teacher	ResNet-56		ResNet-110		ResNet-110		WRN-40-2		WRN-40-2		VGG-13	
	MLL	MCMI	MLL	MCMI	MLL	MCMI	MLL	MCMI	MLL	MCMI	MLL	MCMI
Accuracy	72.34	72.09	74.31	73.60	74.31	73.60	75.61	75.21	75.61	75.21	74.64	73.96
LL	-0.101	-0.344	-0.033	-0.246	-0.033	-0.246	-0.052	-0.132	-0.052	-0.132	-0.007	-0.076
CMI	0.1583	0.4428	0.0605	0.3474	0.0605	0.3474	0.0255	0.1951	0.0255	0.1951	0.0152	0.1298
Student	ResNet-20		ResNet-20		ResNet-32		WRN-16-2		WRN-40-1		VGG-8	
Accuracy	69.06		69.06		71.14		73.26		71.98		70.36	
KD	70.66	70.84 | 0.18	70.67	70.85 | 0.18	73.08	73.48 | 0.40	74.92	75.42 | 0.50	73.54	74.53 | 0.99	72.98	73.83 | 0.85
AT	70.55	70.89 | 0.34	70.22	70.68 | 0.46	72.31	73.96 | 1.65	74.08	74.49 | 0.41	72.77	73.25 | 0.48	71.43	71.76 | 0.33
PKT	70.34	70.96 | 0.62	70.25	71.03 | 0.78	72.61	72.92 | 0.31	74.54	75.01 | 0.47	73.45	74.15 | 0.70	72.88	73.35 | 0.47
SP	69.67	70.98 | 1.31	70.04	70.83 | 0.79	72.69	73.34 | 0.65	73.83	74.60 | 0.77	72.43	73.60 | 1.17	72.68	73.29 | 0.61
CC	69.63	69.98 | 0.35	69.48	70.02 | 0.54	71.48	71.71 | 0.23	73.56	74.00 | 0.44	72.21	72.50 | 0.29	70.71	71.02 | 0.31
RKD	69.61	70.68 | 1.07	69.25	70.24 | 0.99	71.82	72.65 | 0.83	73.35	73.97 | 0.62	72.22	72.66 | 0.44	71.48	72.03 | 0.55
VID	70.38	70.64	70.16	70.69	72.61	73.10	74.11	74.44	73.30	73.58	71.23	71.93

TOP-1/5 ACCURACY ON IMAGENET

Teacher-Student		Teacher Performance				KD		DKD		ReviewKD		CRD	
		Top1	Top5	CMI	LL	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5
RN34-R18	MLL	73.31	91.42	0.7203	-0.5600	71.03	90.05	71.70	90.41	71.61	90.51	71.17	90.13
	MCMI	71.62	90.67	0.8650	-0.6262	71.54	90.63	**72.06**	91.12	71.98	90.86	71.50	90.20
RN50-MNV2	MLL	76.13	92.86	0.6002	-0.4492	70.50	89.80	72.05	91.05	72.56	91.00	71.37	90.41
	MCMI	74.94	91.03	0.7150	-0.4943	71.10	90.16	72.61	91.26	**73.00**	92.19	71.63	90.53

ZERO-SHOT CLASSIFICATION IN KD



The teacher model is trained on the entire dataset, but the samples for some classes are completely omitted for the student during distillation.

The dropped classes are highlighted in red. As seen, student's accuracies of dropped classes are always zero when learned from MLL teacher. While these accuracies substantially increased by as much as 84%.

FEW-SHOT CLASSIFICATION IN KD

α	5		10		15		25		35		50		75	
Teacher	MLL	MCMI	MLL	MCMI	MLL	MCMI	MLL	MCMI	MLL	MCMI	MLL	MCMI	MLL	MCMI
KD	52.30	58.02 +5.72	60.13	63.75 +3.62	63.52	66.20 +2.68	66.78	68.10 +1.32	68.28	69.34 +1.06	69.52	70.28 +0.76	70.44	70.59 +0.15
CRD	47.60	51.40 +3.80	54.60	56.80 +2.20	58.90	60.02 +1.12	63.82	64.7 +0.88	66.70	67.22 +0.52	68.84	69.15 +0.31	70.35	70.40 +0.05

In few shot classification, the models are provided with an α% percent of instance in each class. We conducted experiments for different values of α, namely {5, 10, 15, 25, 35, 50, 70}, the improvement in student's accuracy is notable by replacing MLL teacher by MCMI teacher, particularly more pronounced for small α values.