Contrastive Consistent Representation Distillation

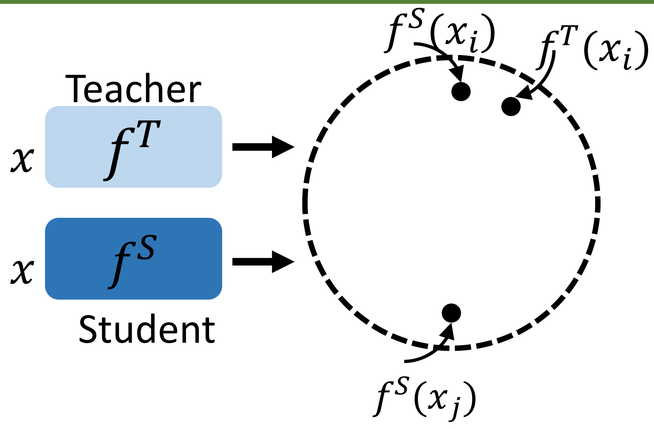
Paper ID: 300

Shipeng Fu, Haoran Yang, Xiaomin Yang

College of Electronics And Information Engineering, Sichuan University

BMVC 2023

Background



The contrastive-learning-based distillation encourages the teacher and student to map the same input to close representations (in some metric space), and different inputs to distant representations.

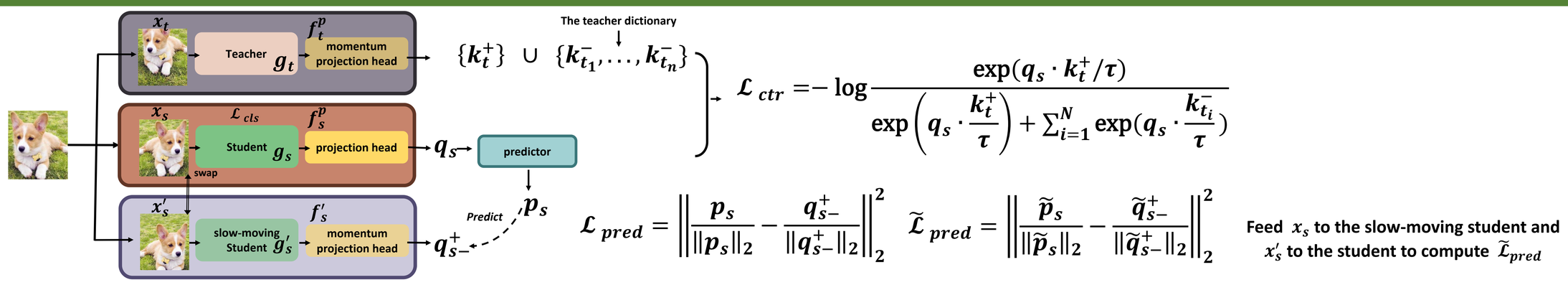
Challenges in contrastive-learning-based distillation

- **Inconsistent representations:** the keys cached in the memory banks were momentum-updated only when they were last processed, and the update interval for each individual key can be highly different.
- **Large storage size of memory banks:** most of the contrastive-learning-based distillation methods treat the entire training dataset as the memory bank and maintain two memory banks, one for the student and one for the teacher.

Main ideas

- **Momentum Updating the encoders:** the momentum update makes the encoders progress more smoothly. The difference between the encoders at different iterations can be made small. Therefore, the keys encoded at different iterations can be consistent.
- **Keep only one fixed-size memory bank for the teacher:** the teacher dictionary (i.e., the only memory bank) is a fixed-size queue, where all the keys are negative keys. The teacher keys of the current batch are enqueued, while the oldest keys are dequeued.

Proposed method



- **Contrastive learning as looking up in the teacher dictionary:** Given an input image x, two views of x under random data augmentations form a positive pair. All the keys in the teacher dictionary are negative keys.
- **The slow-moving student:** the slow-moving student is implemented as a momentum-moving average of the *student* to reduce the effect of the potential noise in the teacher dictionary.

Experimental Results

Teacher Student	WRN-40-2 WRN-16-2	WRN-40-2 WRN-40-1	resnet56 resnet20	resnet110 resnet20	resnet110 resnet32	resnet32x4 resnet8x4	vgg13 vgg8
Teacher	75.61	75.61	72.34	74.31	74.31	79.42	74.64
Student	73.26	71.98	69.06	69.06	71.14	72.50	70.36
KD	74.92	73.54	70.66	70.67	73.08	73.33	72.98
FitNet	73.58 (↓)	72.24 (↓)	69.21 (↓)	68.99 (↓)	71.06 (↓)	73.50 (↑)	71.02 (↓)
AT	74.08 (↓)	72.77 (↓)	70.55 (↓)	70.22 (↓)	72.31 (↓)	73.44 (↑)	71.43 (↓)
SP	73.83 (↓)	72.43 (↓)	69.67 (↓)	70.04 (↓)	72.69 (↓)	72.94 (↓)	72.68 (↓)
CC	73.56 (↓)	72.21 (↓)	69.63 (↓)	69.48 (↓)	71.48 (↓)	72.97 (↓)	70.71 (↓)
VID	74.11 (↓)	73.30 (↓)	70.38 (↓)	70.16 (↓)	72.61 (↓)	73.09 (↓)	71.23 (↓)
RKD	73.35 (↓)	72.22 (↓)	69.61 (↓)	69.25 (↓)	71.82 (↓)	71.90 (↓)	71.48 (↓)
PKT	74.54 (↓)	73.45 (↓)	70.34 (↓)	70.25 (↓)	72.61 (↓)	73.64 (↑)	72.88 (↓)
AB	72.50 (↓)	72.38 (↓)	69.47 (↓)	69.53 (↓)	70.98 (↓)	73.17 (↓)	70.94 (↓)
FT	73.25 (↓)	71.59 (↓)	69.84 (↓)	70.22 (↓)	72.37 (↓)	72.86 (↓)	70.58 (↓)
CRD	**75.48** (↑)	74.14 (↑)	71.16 (↑)	71.46 (↑)	73.48 (↑)	**75.51** (↑)	73.94 (↑)
LCKT	75.22 (↑)	74.11 (↑)	71.14 (↑)	71.23 (↑)	72.32 (↑)	74.65 (↑)	73.50 (↑)
CoCoRD (ours)	**75.48** (↑)	**75.17** (↑)	**71.74** (↑)	**72.11** (↑)	**74.10** (↑)	75.29 (↑)	**73.99** (↑)

When the student has the same architecture style as the teacher

Teacher Student	vgg13 MobileNetV2	ResNet50 MobileNetV2	ResNet50 vgg8	resnet32x4 ShuffleNetV1	resnet32x4 ShuffleNetV2	WRN-40-2 ShuffleNetV1
Teacher	74.64	79.34	79.34	79.42	79.42	75.61
Student	64.60	64.60	70.36	70.50	71.82	70.50
KD	67.37	67.35	73.81	74.07	74.45	74.83
FitNet	64.14 (↓)	63.16 (↓)	70.69 (↓)	73.59 (↓)	73.54 (↓)	73.73 (↓)
AT	59.40 (↓)	58.58 (↓)	71.84 (↓)	71.73 (↓)	72.73 (↓)	73.32 (↓)
SP	66.30 (↓)	68.08 (↑)	73.34 (↓)	73.48 (↓)	74.56 (↑)	74.52 (↓)
CC	64.86 (↓)	65.43 (↓)	70.25 (↓)	71.14 (↓)	71.29 (↓)	71.38 (↓)
VID	65.56 (↓)	67.57 (↑)	70.30 (↓)	73.38 (↓)	73.40 (↓)	73.61 (↓)
RKD	64.52 (↓)	64.43 (↓)	71.50 (↓)	72.28 (↓)	73.21 (↓)	72.21 (↓)
PKT	67.13 (↓)	66.52 (↓)	73.01 (↓)	74.10 (↑)	74.69 (↑)	73.89 (↓)
AB	66.06 (↓)	67.20 (↓)	70.65 (↓)	73.55 (↓)	74.31 (↓)	73.34 (↓)
FT	61.78 (↓)	60.99 (↓)	70.29 (↓)	71.75 (↓)	72.50 (↓)	72.03 (↓)
CRD	69.73 (↑)	69.11 (↑)	74.30 (↑)	75.11 (↑)	75.65 (↑)	76.05 (↑)
LCKT	68.21 (↑)	68.81 (↑)	73.21 (↑)	74.62 (↑)	74.70 (↑)	75.08 (↑)
CoCoRD (ours)	**69.86** (↑)	**70.22** (↑)	**74.52** (↑)	**75.99** (↑)	**77.28** (↑)	**76.42** (↑)

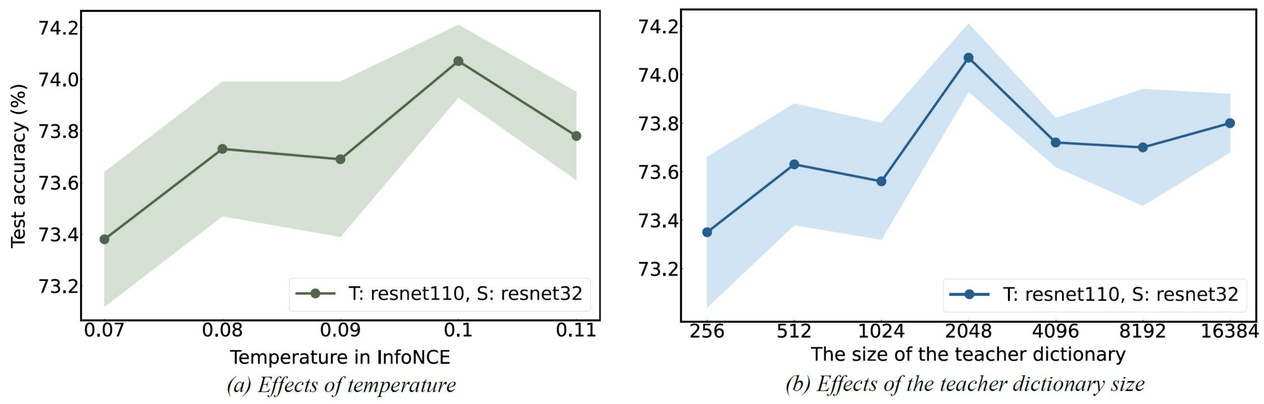
When the student has the different architecture style as the teacher

	Teacher	Student	AT	KD	SP	CC	CRD	CRD+KD	ReviewKD	SSKD	WCoRD	CoCoRD
Top-1	26.70	30.24	29.30	29.34	29.38	30.04	28.83	28.62	28.39	28.48	28.51	**28.26**
Top-5	8.58	10.92	10.00	10.12	10.20	10.83	9.87	9.51	9.49	9.33	9.84	**9.30**

Top-1 and Top-5 error rates (%) on ImageNet-1K validation set

	Classification	Object Detection					
	ImageNet	PASCAL VOC Detection			CoCo Detection		
	Top-1 accuracy (%)	AP_{50}	AP	AP_{75}	AP_{50}	AP	AP_{75}
scratch	-	60.2	33.8	33.1	44.0	26.4	27.8
Student	76.15	81.3	53.5	58.8	59.9	40.0	43.1
CRD	77.06 (+0.91)	81.7 (+0.4)	54.2 (+0.7)	60.0 (+1.2)	60.5 (+0.6)	40.7 (+0.7)	43.9 (+0.8)
CoCoRD	77.57 (+1.42)	**82.0** (+0.7)	**55.0** (+1.5)	**61.1** (+2.3)	**60.9** (+1.0)	**41.0** (+1.0)	**44.5** (+1.4)

Transfer Learning



(a) Effects of temperature *(b) Effects of the teacher dictionary size*