SEED: Self-supervised Distillation For Visual Representation

Zhiyuan Fang, Jianfeng Wang, Lijuan Wang, Lei Zhang, Yezhou Yang, Zicheng Liu

ICLR 2021

발표자: 김태성

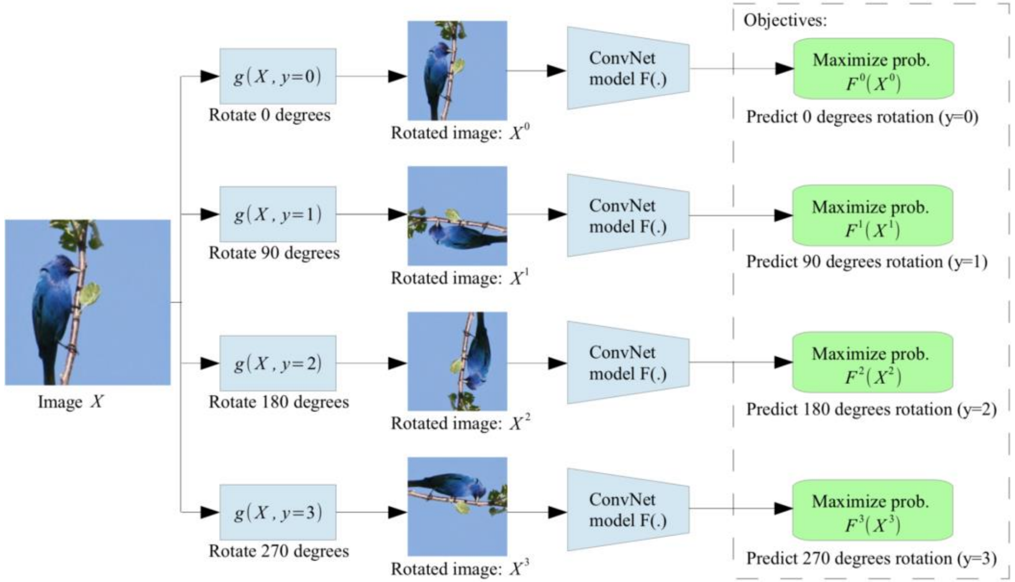
Motivation

The widely used **contrastive self-supervised learning method** **does not work well for small models**.

- ex) MOCO-v2(He *et al.*, 2020) achieves only 36.3% top-1 accuracy on ImageNet-1k dataset.

- Contrastive self-supervised learning?
  - Contrastive learning

- Self-supervised learning



Motivation

The widely used **contrastive self-supervised learning method** **does not work well for small models**.

↓

The contrastive self-supervised learning **works well for large models.**

↓

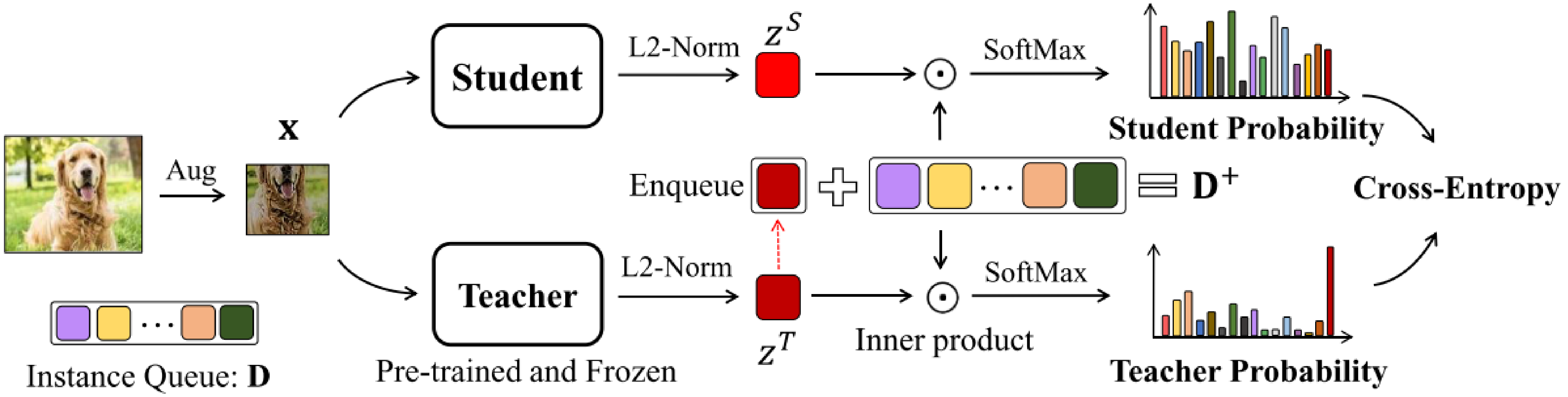
Knowledge distillation into a small network
SElf-SupErvised Distillation (SEED)

Methods

- Knowledge distillation

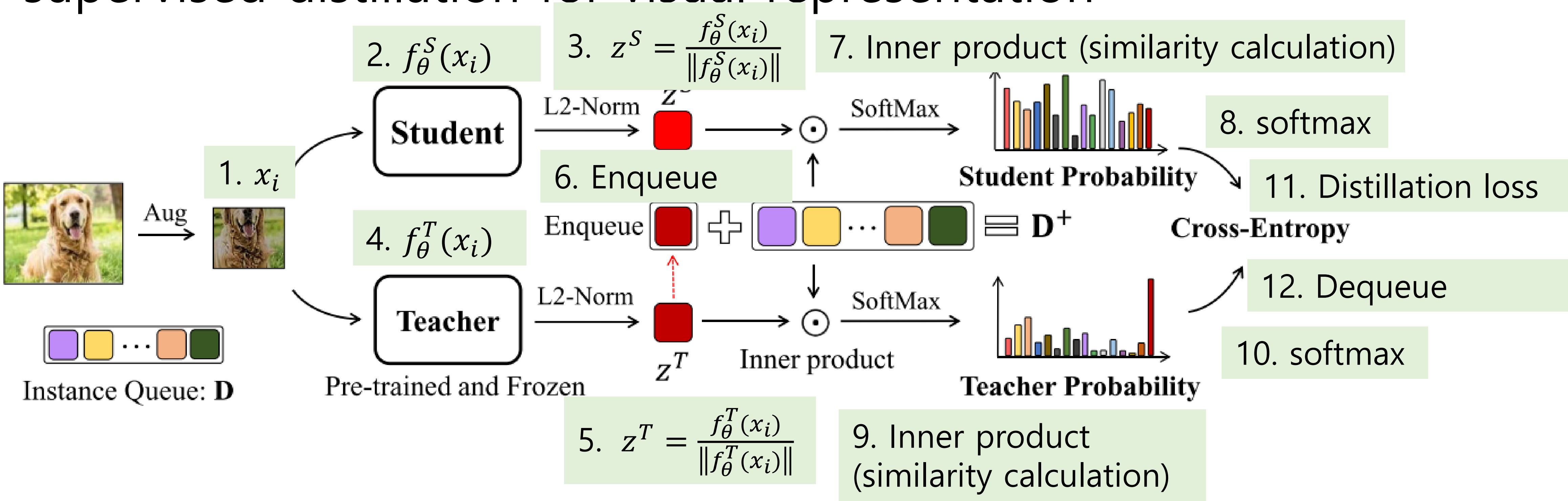
$$\hat{\theta}_S = \arg\min_{\theta_S} \sum_{i}^{N} \mathcal{L}_{\text{sup}}(\mathbf{x}_i, \theta_S, y_i) + \mathcal{L}_{\text{distill}}(\mathbf{x}_i, \theta_S, \theta_T),$$

- Self-supervised distillation for visual representation



Methods

- Self-supervised distillation for visual representation

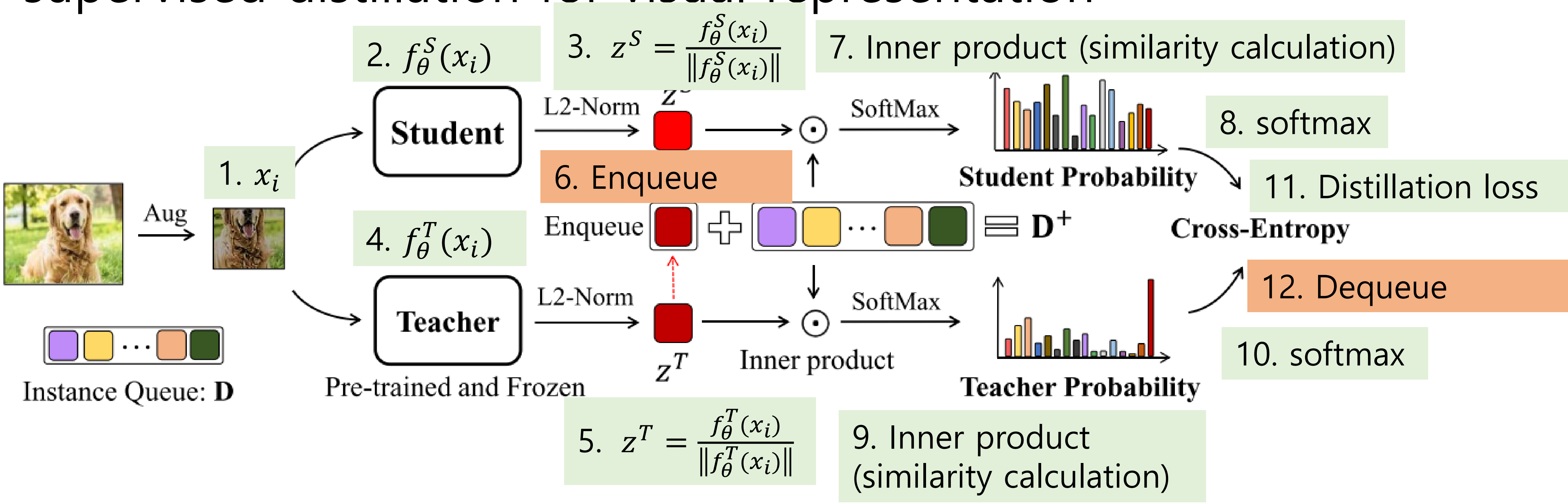


1. Training of the teacher network
   - self-supervised contrastive learning(e.g. SimCLR, MOCO, SimSiam, …)

2. Knowledge distillation into the student network

Methods

- Self-supervised distillation for visual representation

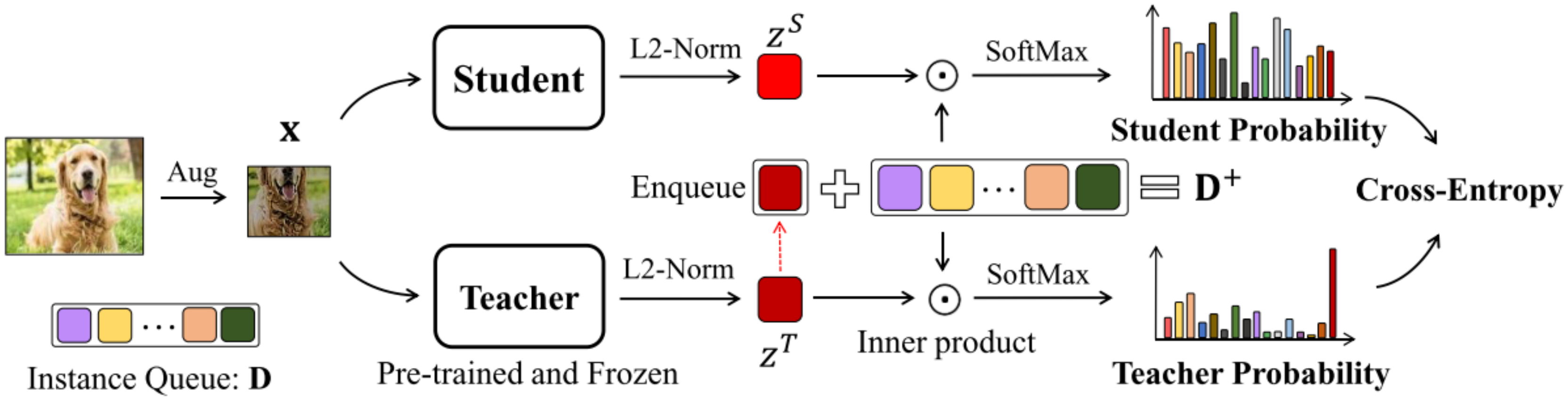


1. Training of the teacher network
   - self-supervised contrastive learning(e.g. SimCLR, MOCO, SimSiam, …)

2. Knowledge distillation into the student network

Methods

- Self-supervised distillation for visual representation



- Cross-Entropy loss with temperature

$$\hat{\theta}_S = \arg\min_{\theta_S} \sum_{i}^{N} -\mathbf{p}^T(\mathbf{x}_i; \theta_T, \mathbf{D}^+) \cdot \log \mathbf{p}^S(\mathbf{x}_i; \theta_S, \mathbf{D}^+)$$

$$= \arg\min_{\theta_S} \sum_{i}^{N} \sum_{j}^{K+1} - \frac{\exp(\mathbf{z}_i^T \cdot \mathbf{d}_j / \tau^T)}{\sum_{\mathbf{d} \sim \mathbf{D}^+} \exp(\mathbf{z}_i^T \cdot \mathbf{d} / \tau^T)} \cdot \log \frac{\exp(\mathbf{z}_i^S \cdot \mathbf{d}_j / \tau^S)}{\sum_{\mathbf{d} \sim \mathbf{D}^+} \exp(\mathbf{z}_i^S \cdot \mathbf{d} / \tau^S)}.$$

Experiments

- Implementation details
  - Teacher pre-training
    - MOCO-v2, SWAV, SimCLR
    - ResNet backbone
  - Self-supervised distillation on student network
    - MobileNet-v3-Large, EfficientNet-B0, smaller ResNet(18, 34 layers)
    - SGD with momentum 0.9
- Experiments
  - Classification
    - Linear and k-NN evaluation on ImageNet
    - Semi-supervised learning (ImageNet 1%, 10%)
    - Domain transfer (CIFAR-10, CIFAR-100, SUN-397)
  - Detection and segmentation
    - Faster R-CNN for the object detection on VOC-07+12 dataset
    - MASK R-CNN for the object detection and instance segmentation on COCO 2017 dataset
  - Ablation study

Experiments

- Classification - Linear and k-NN evaluation on ImageNet

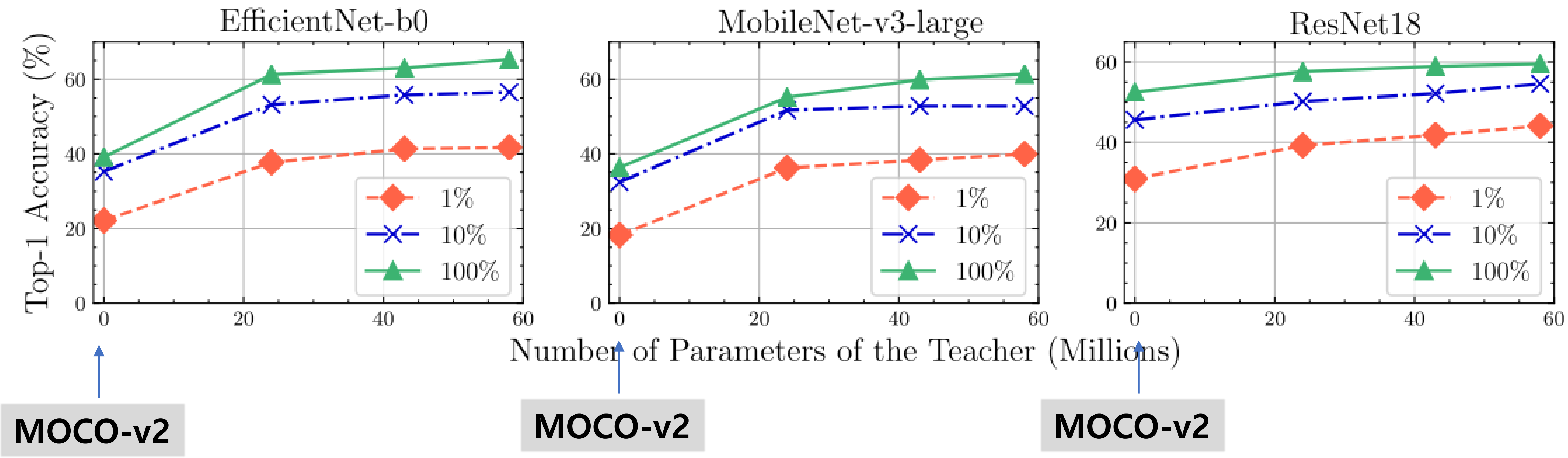
T \ S	T-1	Eff-b0			Eff-b1			Mob-v3			R-18			R-34		
		K	T-1	T-5	*K*	T-1	T-5	*K*	T-1	T-5	*K*	T-1	T-5	*K*	T-1	T-5
Supervised Acc.		77.3			79.2			75.2			72.1			75.0		
MOCO-v2		30.0	42.2	68.5	34.4	50.7	74.6	27.5	36.3	62.2	36.7	52.5	77.0	41.5	57.4	81.6
R-50 △	67.4	46.0 +16.0	61.3 +19.1	82.7 +14.2	46.1 +16.1	61.4 +10.7	83.1 +8.8	44.8 +17.3	55.2 +18.9	80.3 +18.1	43.4 +6.7	57.6 +5.1	81.8 +4.8	45.2 +3.7	58.5 +1.1	82.6 +1.0
R-101 △	70.3	50.1 +20.1	63.0 +20.8	83.8 +15.3	50.3 +15.9	63.4 +12.7	84.6 +10.0	48.8 +21 .3	59.9 +23.6	83.5 +21.3	48.6 +11.9	58.9 +6.4	82.5 +5.5	50.5 +9.0	61.6 +4.2	84.9 +3.3
R-152 △	74.2	50.7 +20.7	65.3 +23.1	86.0 +17.5	52.4 +18.0	67.3 +16.6	86.9 +12.3	49.5 +22.0	61.4 +25.1	84.6 +22.4	49.1 +12.4	59.5 +7.0	83.3 +6.3	51.4 +9.9	62.7 +5.3	85.8 +4.2
R50$\times 2^{*}$ △	77.3	57.4 +27.4	67.6 +25.4	87.4 +18.9	60.3 +25.9	68.0 +17.3	87.6 +13.0	55.9 +18.9	68.2 +31.9	88.2 +26.0	55.3 +18.6	63.0 +10.5	84.9 +7.9	58.2 +16.7	65.7 +8.3	86.8 +5.2

R50x2 - https://arxiv.org/pdf/2006.09882.pdf (SwAV)

MOCO-v2 - https://arxiv.org/abs/2003.04297

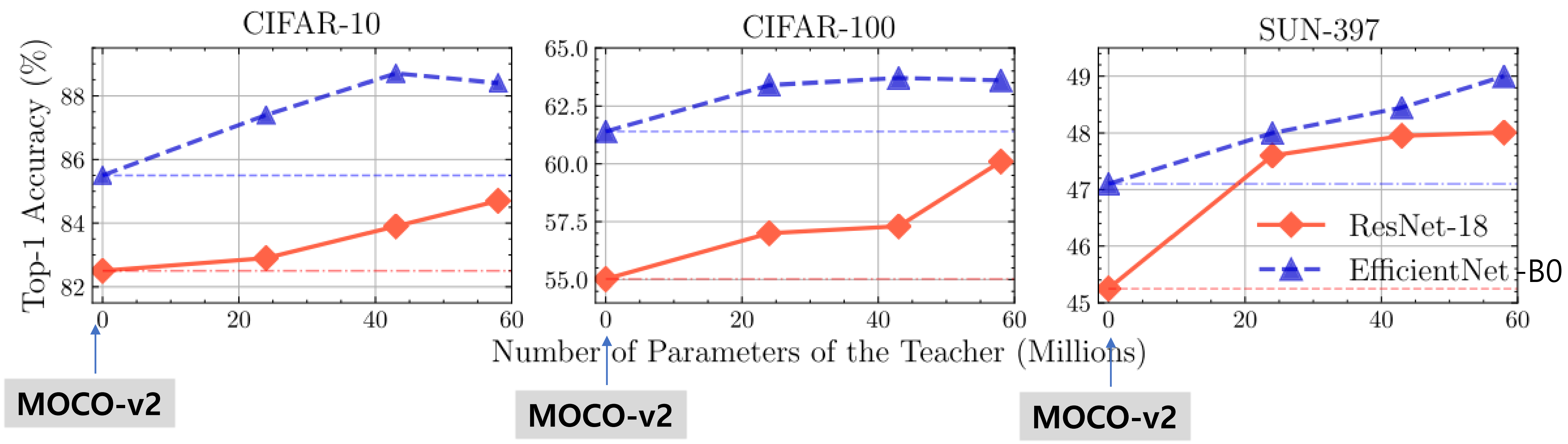
Experiments

- Classification – Semi-supervised learning



Experiments

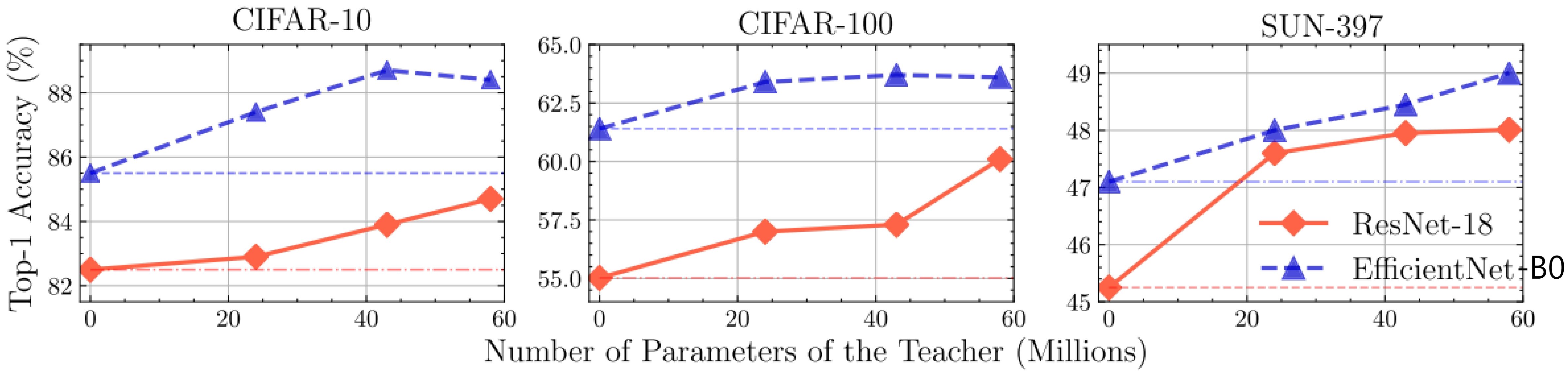
- Classification – Domain transfer



Experiments

- Object detection and instance segmentation

S	T	Faster R-CNN: VOC Obj. Det.			Mask R-CNN: COCO Obj. Det.			Mask R-CNN: COCO Inst. Segm.		
		AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
	MOCO-v2	46.1	74.5	48.6	35.0	53.9	37.7	31.0	51.1	33.1
	R-50	46.1(0.0)	74.8(**+0.3**)	49.1(**+0.5**)	35.3(**+0.3**)	54.2(**+0.3**)	37.8(+0.1)	31.1(**+0.1**)	51.1(0.0)	33.2(+0.1)
R-18	**R-101**	46.8(**+0.7**)	75.8(**+1.3**)	49.3(**+0.7**)	35.3(**+0.3**)	54.3(**+0.4**)	37.9(+0.2)	31.3(**+0.3**)	51.3(+0.2)	33.4(+0.3)
	R-152	46.8(**+0.7**)	75.9(**+1.4**)	50.2(**+1.6**)	35.4(**+0.4**)	54.4(**+0.5**)	38.0(**+0.3**)	31.3(**+0.3**)	51.4(**+0.3**)	33.4(**+0.3**)



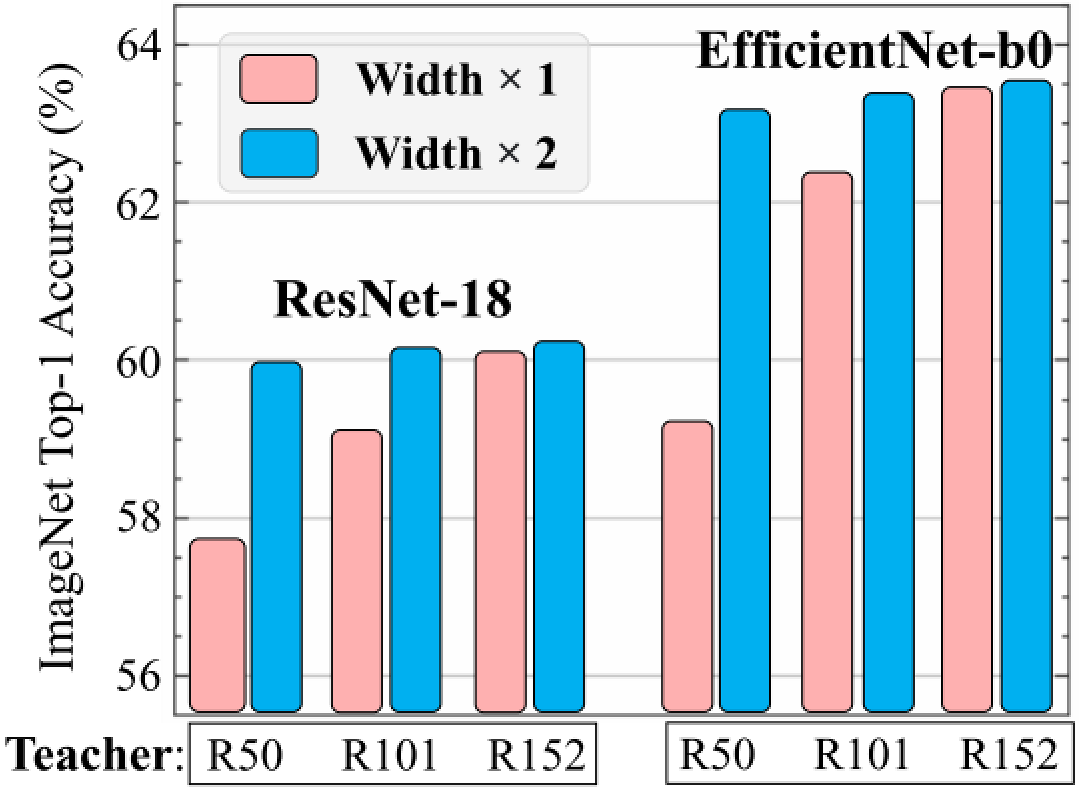
Experiments

- Ablation study

Training algorithm

Teacher	P-E	D-E	T. Top-1	S. Top-1	S. Top-5
✗	✗	✗	✗	52.5	77.0
MoCo	200	200	60.6	52.1	77.0
SimCLR	200	200	65.6	57.5	81.7
MoCo-v2	200	200	67.4	57.6	81.8
	800	200	71.1	60.5	83.5
SWAV	800	100	75.3	61.1	83.8
	800	200	75.3	61.7	84.2
	800	400	75.3	62.0	84.4
SWAV*	800	200	75.3	**62.6**	**84.8**

Teacher network



Experiments

- Ablation study

Distillation strategy

Method	Top-1 Acc.	Top-5 Acc.
l2-Distance	55.3	80.3
K-Means	51.0	75.8
Online Clustering	56.4	81.2
Binary Contr. Loss	57.4	81.5
SEED + MoCo-V2	57.6	81.8
SEED	**57.9**	**82.0**

Loss temperature

τ^T	ImageNet		CIFAR-10	CIFAR-100
	Top-1	Top-5	Top-1	Top-1
0.3	54.8	80.0	78.7	46.6
0.1	54.9	80.1	83.0	50.1
0.05	56.5	81.3	84.4	56.2
0.01	**57.9**	**82.0**	**87.5**	60.6
1e-3	57.6	81.8	86.9	**60.8**

ResNet-50(Teacher) -> ResNet-18(Student)

Experiments

- Other experiment results are shown in Appendix.
    - Ablation study
        - Learning rate
        - Weight decay
        - Queue size
        - Distillation phase
        - Different student networks
    - Small patch(multi-view / multi-crop) learning
    - Deeper projection head

Thank you

발표자: 김태성