ELLab

SUNG KYUN KWAN UNIVERSITY

RotNet

2026.01.02

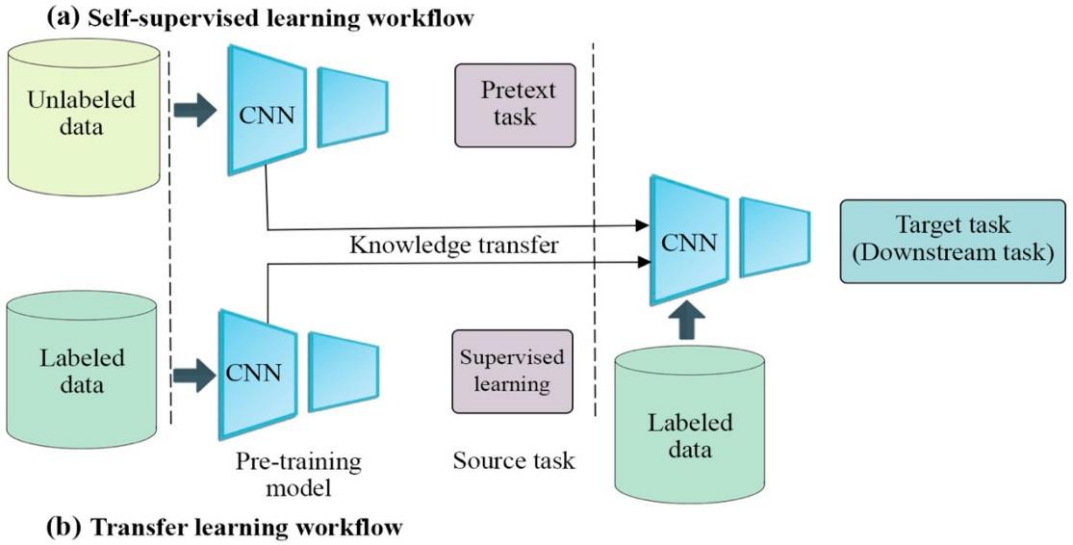
Introduction

Motivation

- CNN은 high-level 시각 표현을 효과적으로 학습할 수 있음
- 그러나 Supervised learning에서는 대규모 수작업 라벨링 데이터가 있어야함
- 라벨링 비용과 확장성 문제가 존재함
- 라벨 없는 대규모 시각 데이터를 활용할 필요성이 증가함

Self-supervised learning

- 데이터 자체로부터 감독 신호를 생성하는 학습 방식
- Pretext task를 통해 representation을 학습
- 색상 복원, 패치 위치 예측, 움직임 예측 등 기존 접근 존재
- 여전히 supervised 학습과 성능 격차 존재



Introduction

Self-supervised learning

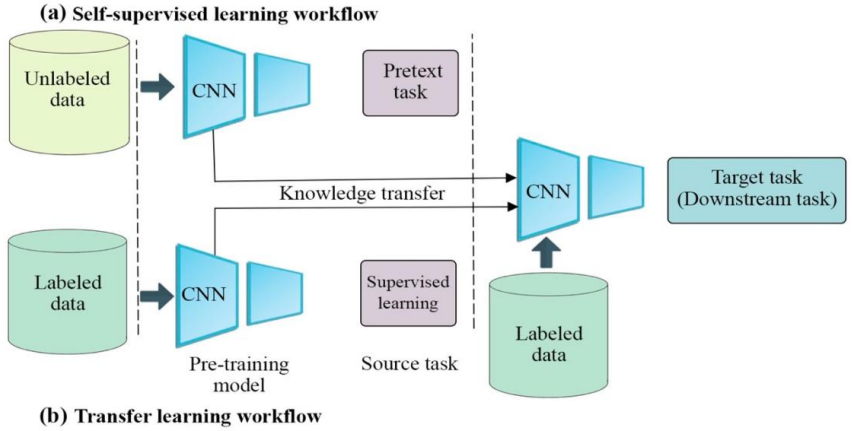
- Image → ConvNet → "90° rotation"(pretext-task: 의미 있는 feature를 배우도록 하기 위해 일부러 설정한 "가짜문제 ")

Supervised learning

- Image → ConNet → "cat"(label)

Key Idea

- 이미지에 적용된 기하학적 변환을 예측하도록 학습
- Semantic representation 학습을 강제함



Semantic feature

- 단순히 '픽셀 값' 이나 '모양' 이 아니라 무엇을 의미하는지 이해하는 것
- 사람-자동차 구분, 눈·코·입·얼굴 등 High-level 특징들 이해 해야함
- 회전 예측을 하는 것은 '이 물체는 원래 어떤 방향이 정상인지'를 알아야 하므로 객체 개념과 구조, 관계 등 semantic한 특징을 파악할 수 있음

Methodology

Geometric Transformations

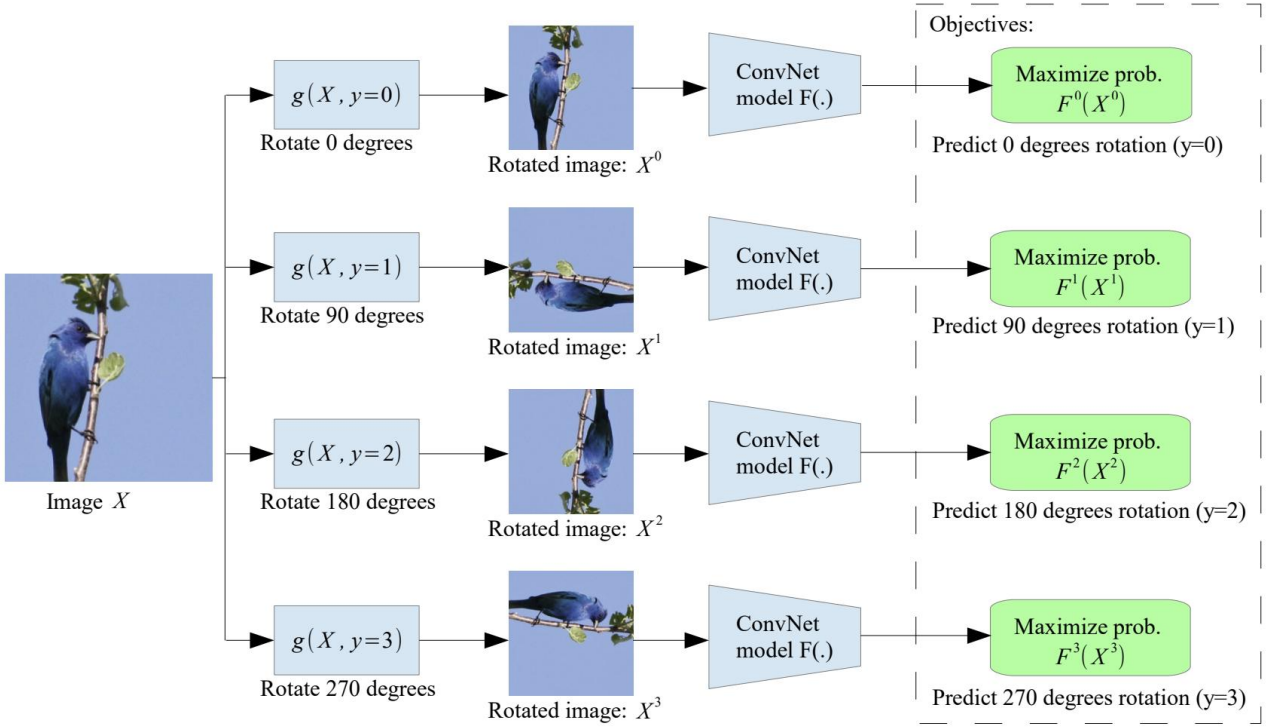
- 각 이미지에 대한 변환: $g(\cdot\,|y)$
- K개의 이산적인 기하학적 변환 집합: $G = \{g(\cdot\,|y)\}_{y=1}^{K}$
- 하나의 원본 이미지 X에 대해 변환 $g(\cdot\,|y)$를 적용 → $X^{y} = g(x|y)$
- ConvNet 모델을 $F(\cdot)$라고 정의할때
  - Input: 어떤 변환이 적용되었는지 모르는 이미지 X^{y^*}
  - Output: 각 기하학적 변환이 적용되었을 확률분포 $\{F^{y}(X^{y^*}|\theta)\}_{y=1}^{K}$
- 학습 데이터 셋은 N개의 이미지로 구성된 집합 $D = \{X_i\}$
- 모델의 학습 목표는 모든 이미지에 대해 변환 분류 오차 최소화

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} loss(X_i, \theta) \qquad loss(X_i, \theta) = -\frac{1}{K} \sum_{y=1}^{K} log(F^{y}(g(X_i|y)|\theta))$$

Methodology

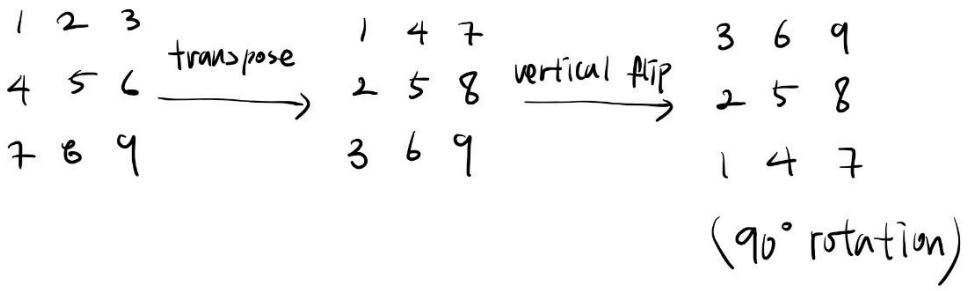
Geometric Transformations

- $G = \{g(X|y)\}_{y=1}^{4}$
- $Rot(X, y \times 90°)$



Implementing image rotation

- 90°: transpose + vertical flip
- 180°: vertical flip + horizontal flip
- 270°: vertical flip + transpose



Methodology

Forcing the learning of semantic features

(Figure 3)

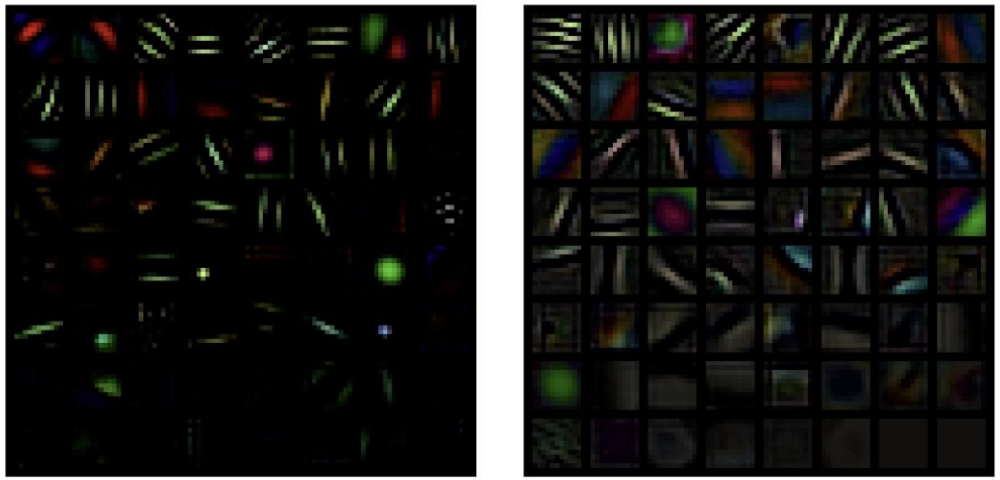
- Self-supervised (b)의 attention map이 supervised (a)의 것과 비슷하게 high-level parts(눈,코,입,귀,꼬리 등)에 집중함
- Semantic feature를 학습한 것을 알 수 있음

(Figure 4)

- Self-supervised (b)의 filter들이 supervised (a)의 것보다 더 다양함
- More generalized, 다양한 task에 적합함

Rotation task가 Semantic feature를 파악하는데 효과적임

(a) **Attention maps of supervised model** (b) **Attention maps of our self-supervised model**



(a) **Supervised** (b) **Self-supervised to recognize rotations**

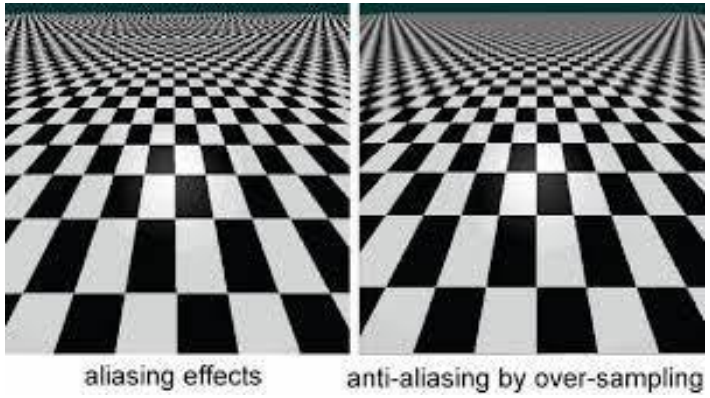
Methodology

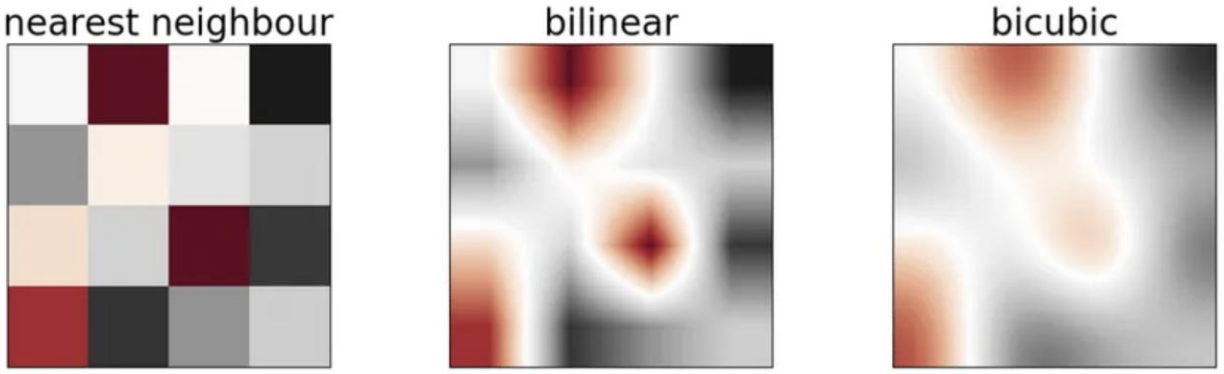
Absence of low-level visual artifacts

Low-level artifact

- 픽셀 수준에서 semantic 이해 없이도 쉽게 학습 가능한 흔적
- Interpolation, aliasing, padding 경계, 빈 픽셀 영역, blur 패턴 등
- rescale/resize → interpolation & 특정 blur 패턴 발생함
- 45° rotation → 잘린 영역 & corner에 빈 영역

90° rotation은 transpose + flip으로 구현 가능하므로 cheating 가능한 단서들을 안 남김





Experiment

Evaluation of the learned feature hierarchies

Key Observation 1: Feature Quality vs Layer Depth

- 모든 모델에서 Conv Block 2의 feature가 가장 높은 분류 성능을 보임
- 깊은 layer의 feature는 rotation prediction task에 점점 특화되고 일반적인 object recognition에는 덜 적합하게 됨

Key Observation 2: Effect of Model Depth

- RotNet의 전체 깊이가 증가할수록, 초기 layer (ConvB1, ConvB2)의 feature 성능은 향상됨
- 더 깊은 conv block을 가진 모델일수록 초기 layer의 feature가 rotation prediction에 덜 종속되고 더 일반적인 semantic 정보를 담게 됨

Model	ConvB1	ConvB2	ConvB3	ConvB4	ConvB5
RotNet with 3 conv. blocks	85.45	88.26	62.09	-	-
RotNet with 4 conv. blocks	85.07	89.06	86.21	61.73	-
RotNet with 5 conv. blocks	85.04	**89.76**	86.82	74.50	50.37

Experiment

Quality of learned features w.r.t. the combinations of rotations

- Rotation predict pretext task에서 사용하는 회전 개수가 feature 품질에 미치는 영향을 분석함
- 4 rotations가 가장 높은 성능을 달성함
- 8 rotations은 성능 소폭 감소함
  - 회전 간 구분이 어렵고 추가 회전으로 인한 시각적 artifact 발생함 → semantic 강제력 약해짐
- 2 rotations은 성능 저하됨
  - 클래스 수가 적어 감독 신호가 약함
  - 0° (upright) 이미지 있어야 성능 더 좋음

4-way rotation prediction이 가장 효과적인 self-supervised signal을 제공함

# Rotations	Rotations	CIFAR-10 Classification Accuracy
4	0°, 90°, 180°, 270°	**89.06**
8	0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°	88.51
2	0°, 180°	87.46
2	90°, 270°	85.52

Experiment

Comparison

- Random Init + conv: 72.50%

→ 학습된 표현의 중요성 확인

- (Ours) RotNet + non-linear: 89.06%
- (Ours) RotNet + conv: 91.16%

→ 기존 unsupervised 방법들 대비 명확한 성능 우위

- RotNet + non-linear (fine-tuned): 91.73%
- RotNet + conv (fine-tuned): 92.17%

→ Fine-tunning하면 성능 더 올라감

Method	Accuracy
Supervised NIN	92.80
Random Init. + conv	72.50
(Ours) RotNet + non-linear	89.06
(Ours) RotNet + conv	**91.16**
(Ours) RotNet + non-linear (fine-tuned)	91.73
(Ours) RotNet + conv (fine-tuned)	92.17
Roto-Scat + SVM Oyallon & Mallat (2015)	82.3
ExemplarCNN Dosovitskiy et al. (2014)	84.3
DCGAN Radford et al. (2015)	82.8
Scattering Oyallon et al. (2017)	84.7

Experiment

Correlation between object classification & rotation prediction

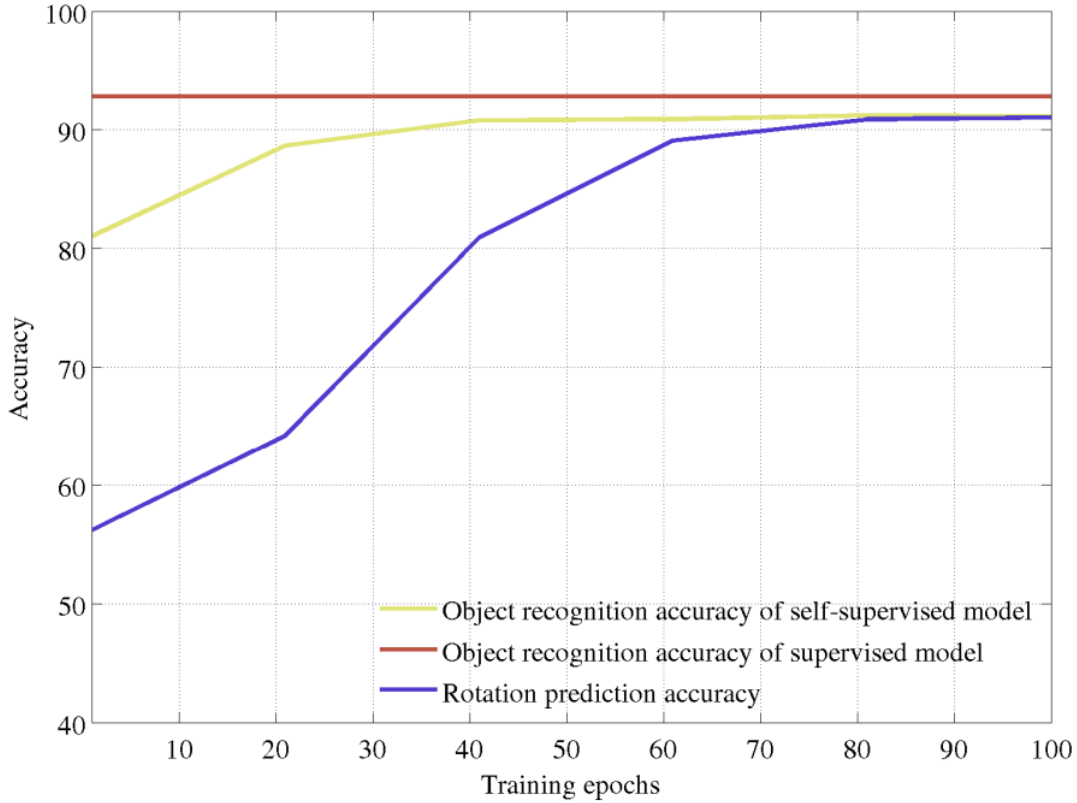
- Rotation prediction accuracy가 증가할수록

→ Object classification accuracy도 함께 증가함

- Rotation prediction을 잘할수록

→ 더 좋은 semantic representation을 학습함

- Object classification 성능은pretext task 학습 초반에 빠르게 수렴함
- 이후 rotation prediction 성능이 계속 향상되더라도 object recognition 성능의 증가는 제한적임

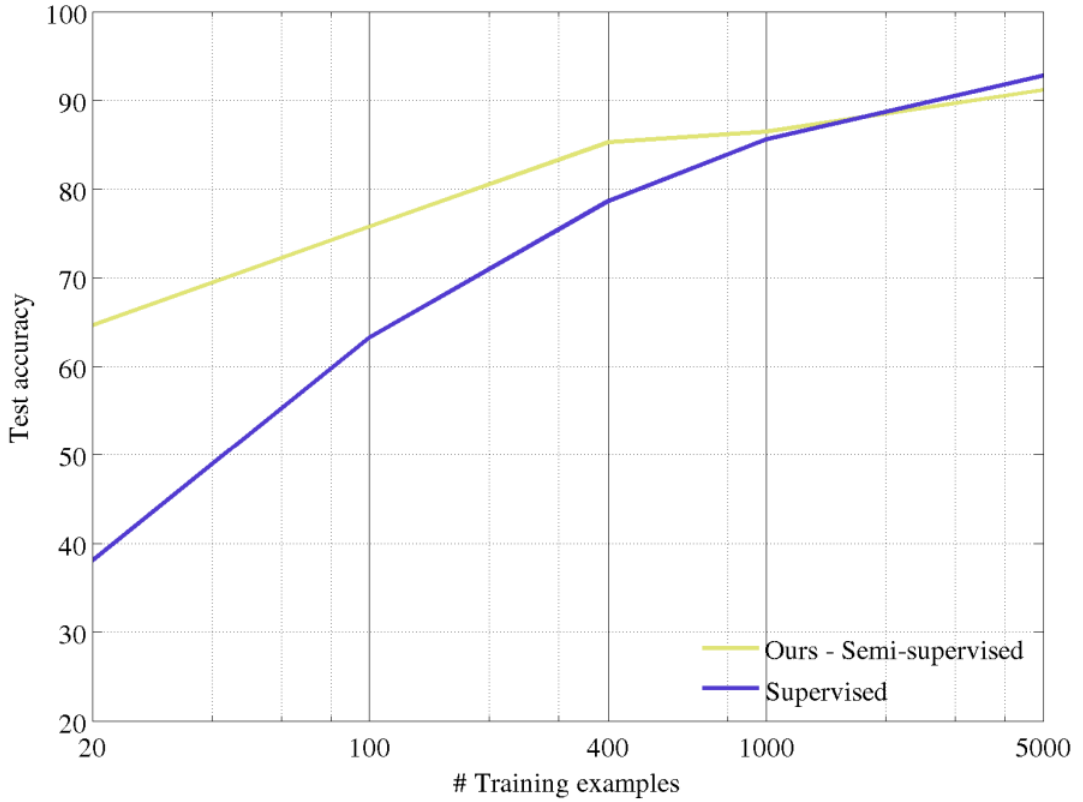


Experiment

Robust to label scarcity

- RotNet을 라벨 없이 전체 CIFAR-10으로 먼저 학습함 (rotation prediction, 4 conv blocks)
- 이후 RotNet의 2nd conv block feature를 고정하고 소량의 라벨 데이터만 사용해 object classifier 학습함
- 라벨 수가 1000개 미만일 때 RotNet 기반 semi-supervised 모델이 supervised 모델보다 높은 성능 달성함

소량의 라벨만으로도 효과적인 object recognition 가능



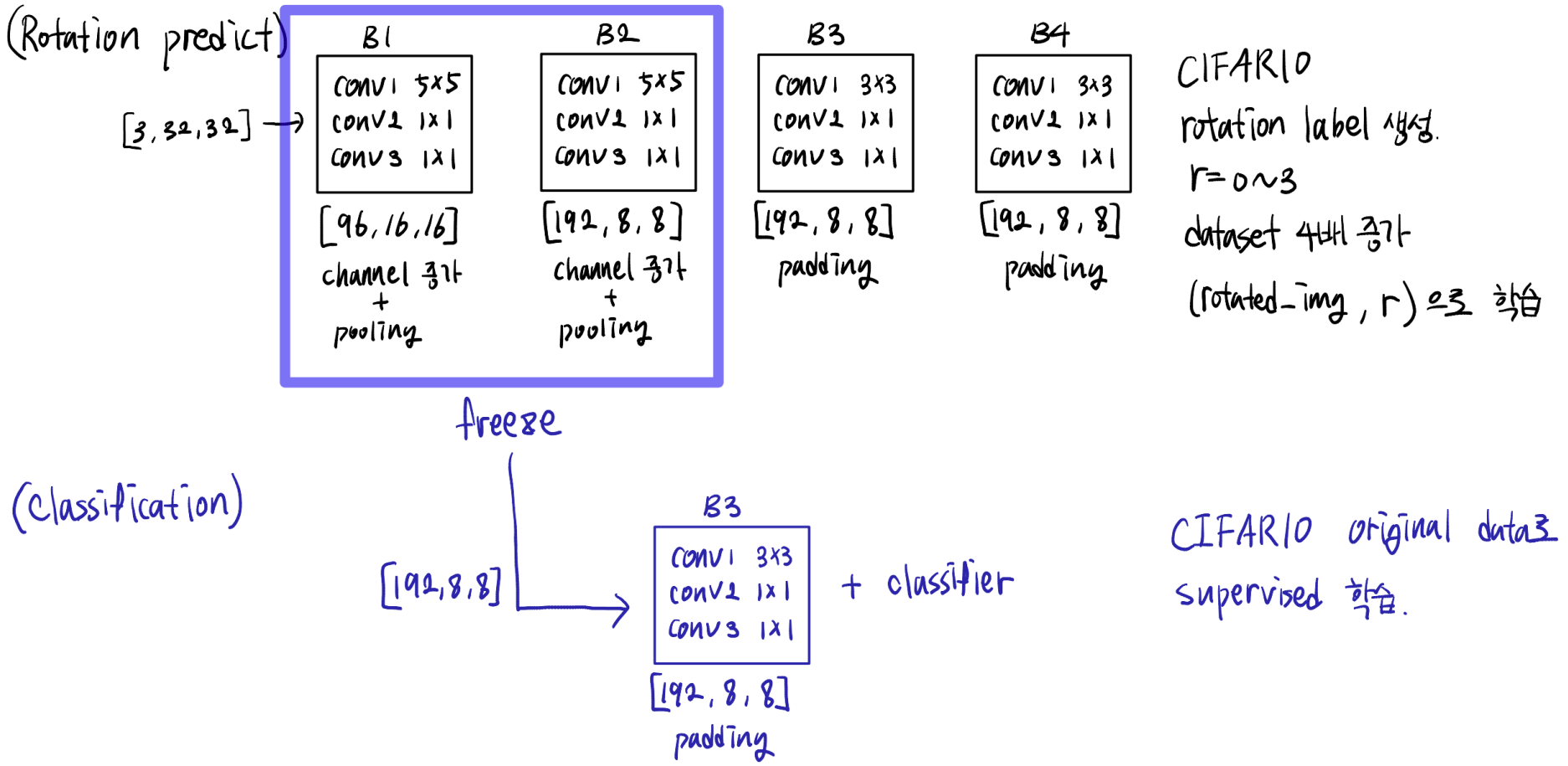
My Code

학습 세팅

항목	논문 설정값
데이터셋	CIFAR-10
이미지 전처리	(classification) 32×32, 4픽셀 zero-padding 후 32×32 랜덤 crop, horizontal flip
모델 깊이	4 conv. blocks, each block has 3 conv
Optimizer	SGD (momentum=0.9)
Batch size	128
Weight decay	5e-4
Learning rate	0.1
Scheduler	epoch = [30, 60, 80] factored by ½
Training epochs	100 epochs

My Code

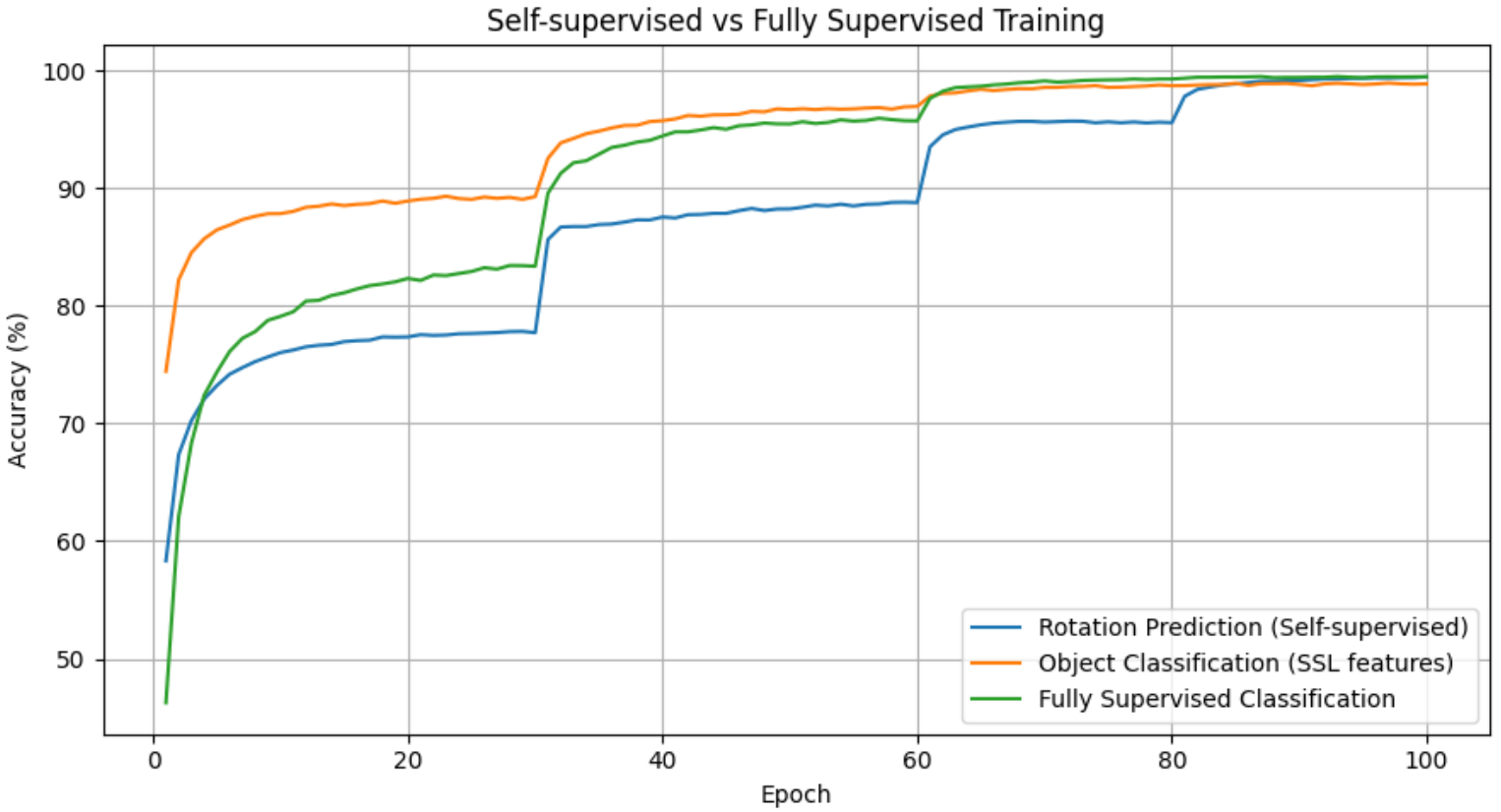
학습 구조


(Rotation predict)
[3, 32, 32]
B1
conv1 5x5
conv2 1x1
conv3 1x1
[96, 16, 16]
channel 증가
+
pooling
B2
conv1 5x5
conv2 1x1
conv3 1x1
[192, 8, 8]
channel 증가
+
pooling
B3
conv1 3x3
conv2 1x1
conv3 1x1
[192, 8, 8]
padding
B4
conv1 3x3
conv2 1x1
conv3 1x1
[192, 8, 8]
padding
CIFAR10
rotation label 생성.
r=0~3
dataset 4배 증가
(rotated_img, r) 으로 학습
freeze
(Classification)
[192, 8, 8]
B3
conv1 3x3
conv2 1x1
conv3 1x1
+ classifier
[192, 8, 8]
padding
CIFAR10 original data로
supervised 학습.

My Code

Experiment

Method	Test accuracy
(Ours)	90.02%
Supervised	92.06%



Rotation prediction accuracy가 증가할수록 object classification accuracy도 함께 상승함

https://colab.research.google.com/drive/15Tg5T1RdmLkx9wnxSAGUgM7qGUUobG8U?usp=sharing

ELLab