Introduction to
Class-Imbalanced Semi-Supervised Learning

Data Mining & Quality Analytics Lab.

2023. 03. 31

발표자: 정진용

발표자 소개

❖ 정진용 (Jinyong Jeong)

- 고려대학교 산업경영공학과 석·박사 통합과정(2021.09~)
- Data Mining & Quality Analytics Lab. (김성범 교수님)

❖ 관심 연구 분야

- Semi-Supervised Learning & Class-Imbalanced Semi-Supervised Learning

❖ E-mail

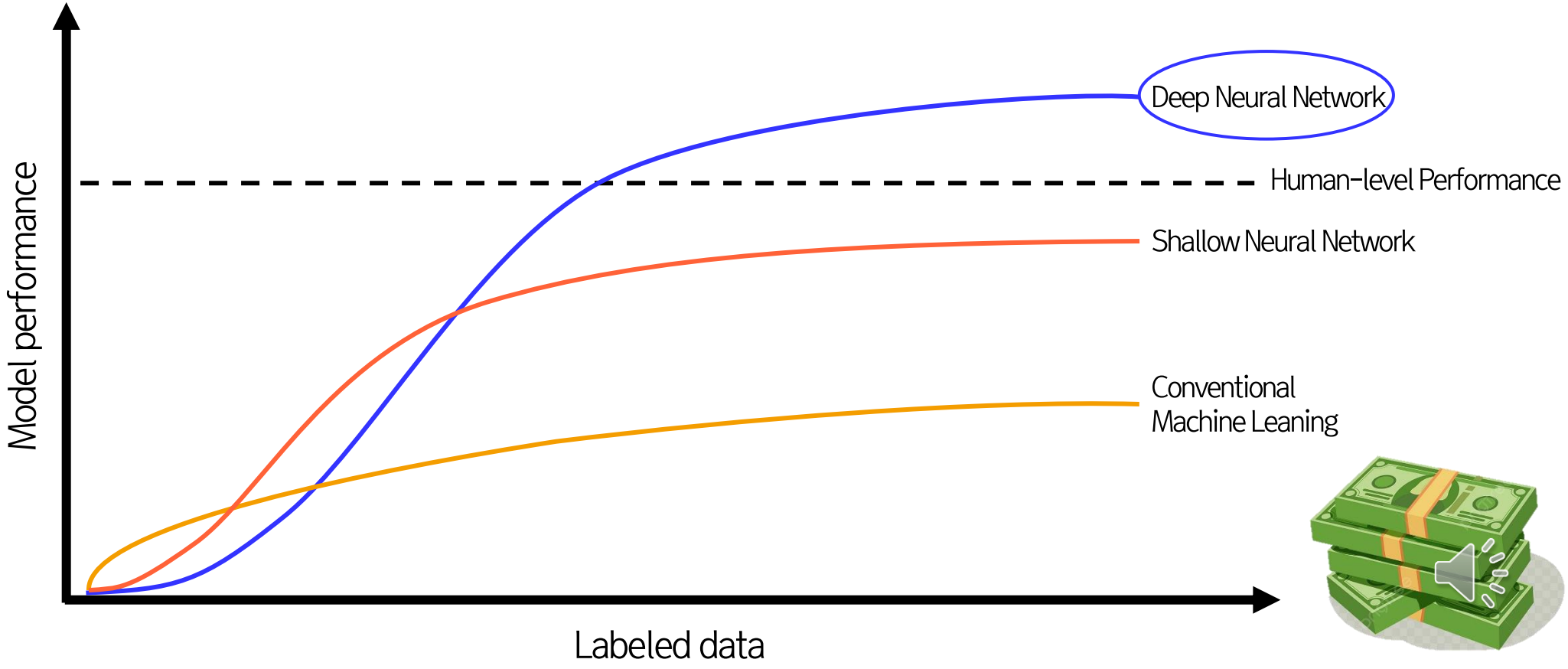
- jy_jeong@korea.ac.kr

목차

Introduction

Semi-Supervised Learning

❖ 딥러닝 모델이 좋은 성능을 내기 위해서 많은 양의 labeled data가 필요함



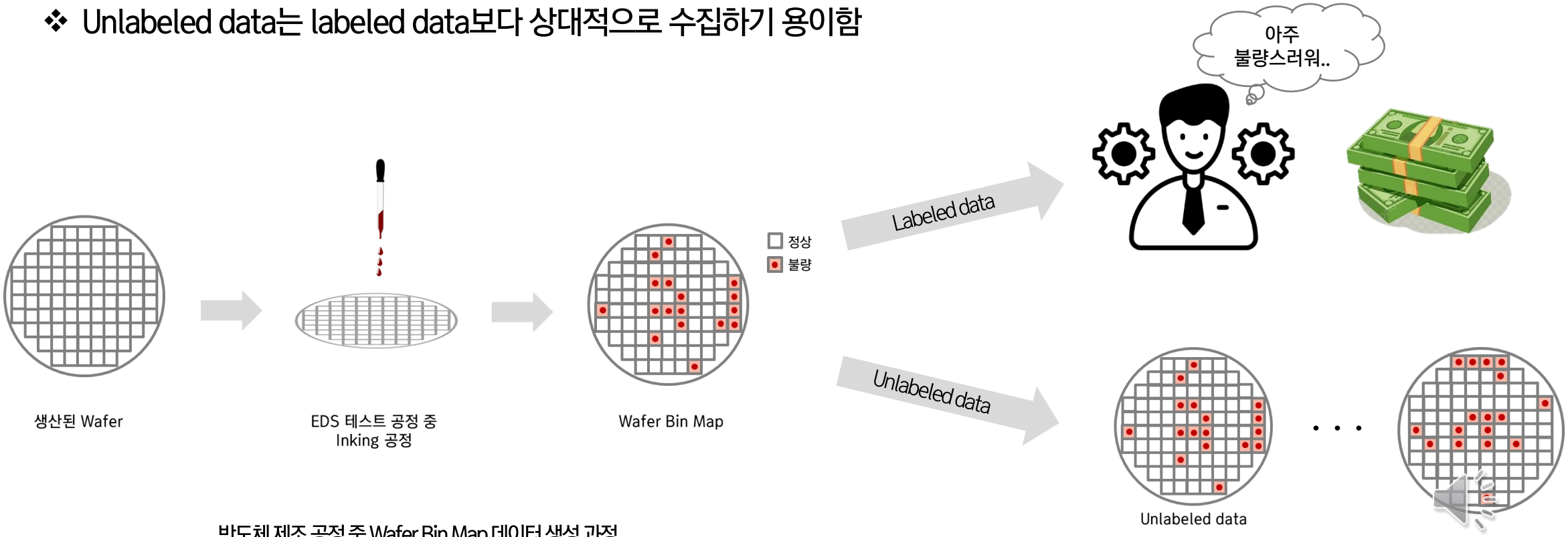
Introduction

Semi-Supervised Learning

❖ Data labeling은 인공지능 모델이 데이터를 학습할 수 있도록 사람이 직접 정보를 입력하는 것을 의미함

- 시간 및 비용이 필요하므로 다량의 labeled data 수집이 어려움

❖ Unlabeled data는 labeled data보다 상대적으로 수집하기 용이함



반도체 제조 공정 중 Wafer Bin Map 데이터 생성 과정

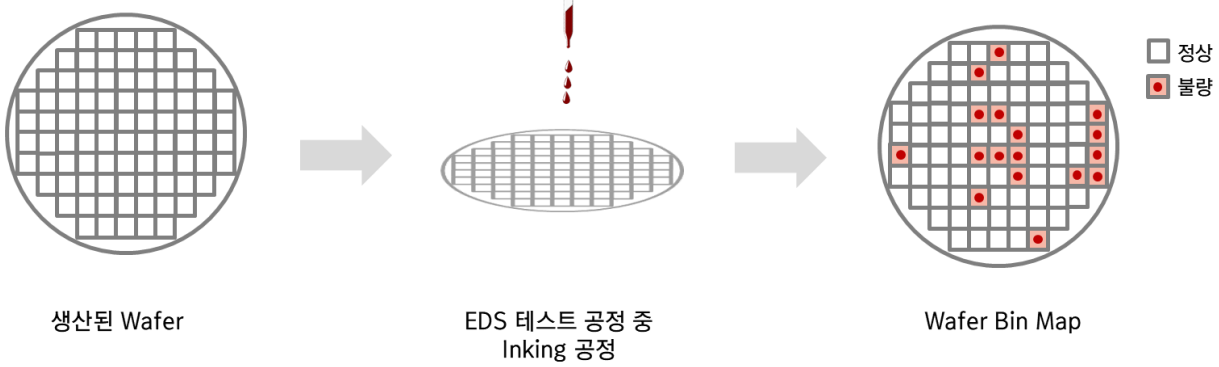
Introduction

Semi-Supervised Learning

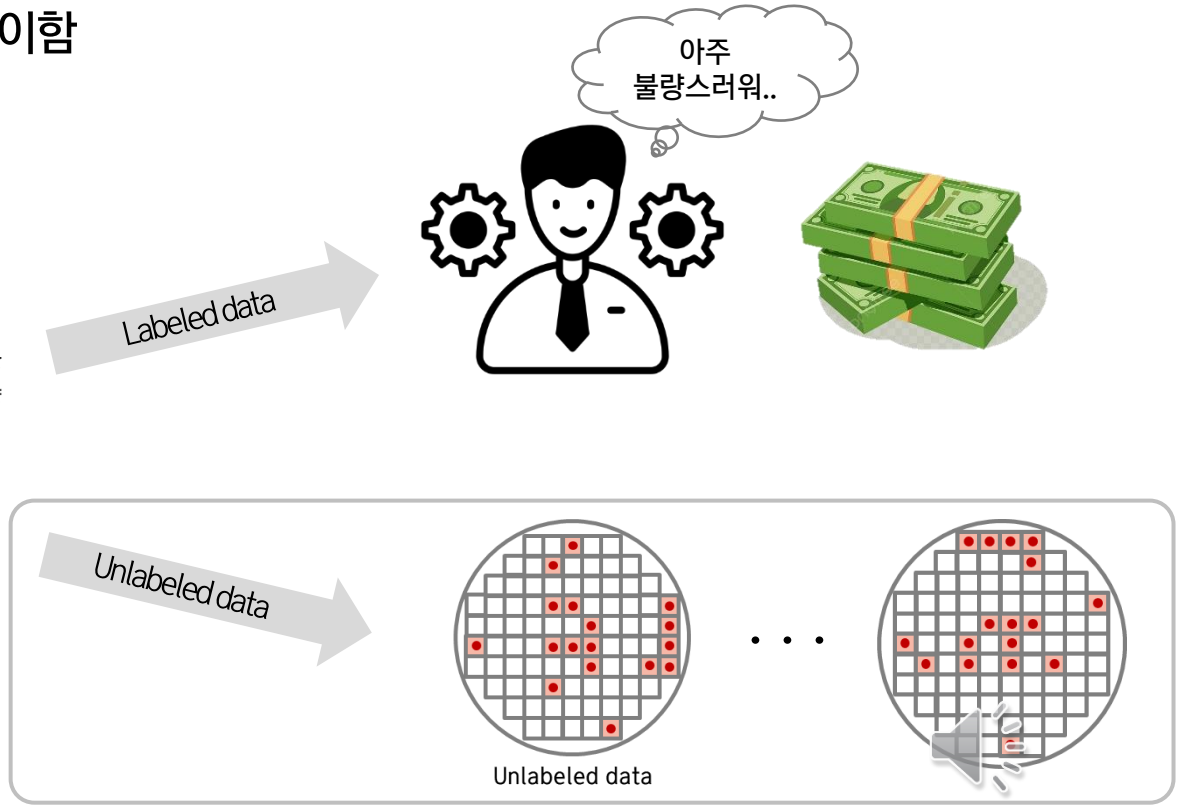
❖ Data labeling은 인공지능 모델이 데이터를 학습할 수 있도록 사람이 직접 정보를 입력하는 것을 의미함

- 시간 및 비용이 필요하므로 다량의 labeled data 수집이 어려움

❖ Unlabeled data는 labeled data보다 상대적으로 수집하기 용이함



반도체 제조 공정 중 Wafer Bin Map 데이터 생성 과정



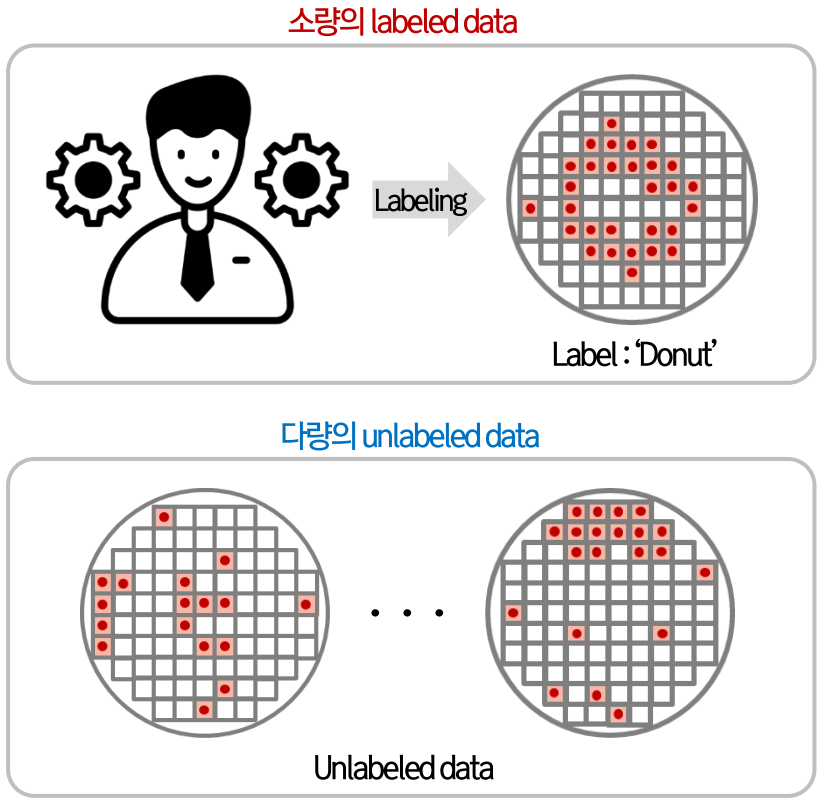
수집하기 용이한 unlabeled data도 활용해서 모델 성능을 높여보자!

Introduction

Semi-Supervised Learning

❖ 상대적으로 수집이 용이한 데이터셋을 활용하여 모델 성능을 높이는 다양한 연구들이 제안

- 소량의 labeled data와 다량의 unlabeled data를 동시에 활용하여 더 좋은 모델을 만들어보자! → Semi-Supervised Learning



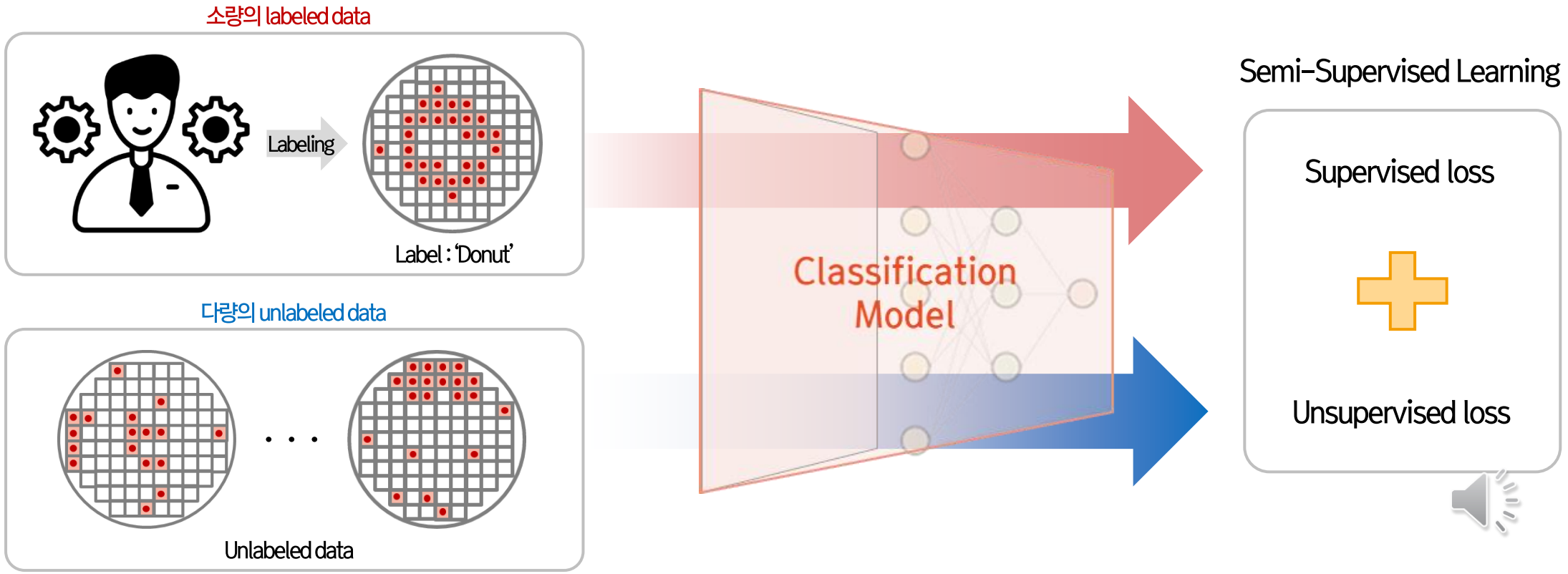
- ✓ Semi-Supervised Learning
- ✓ Self-Supervised Learning
- ✓ Transfer Learning

Introduction

Semi-Supervised Learning

❖ 상대적으로 수집이 용이한 데이터셋을 활용하여 모델 성능을 높이는 다양한 연구들이 제안

- 소량의 labeled data와 다량의 unlabeled data를 동시에 활용하여 더 좋은 모델을 만들어보자! → Semi-Supervised Learning

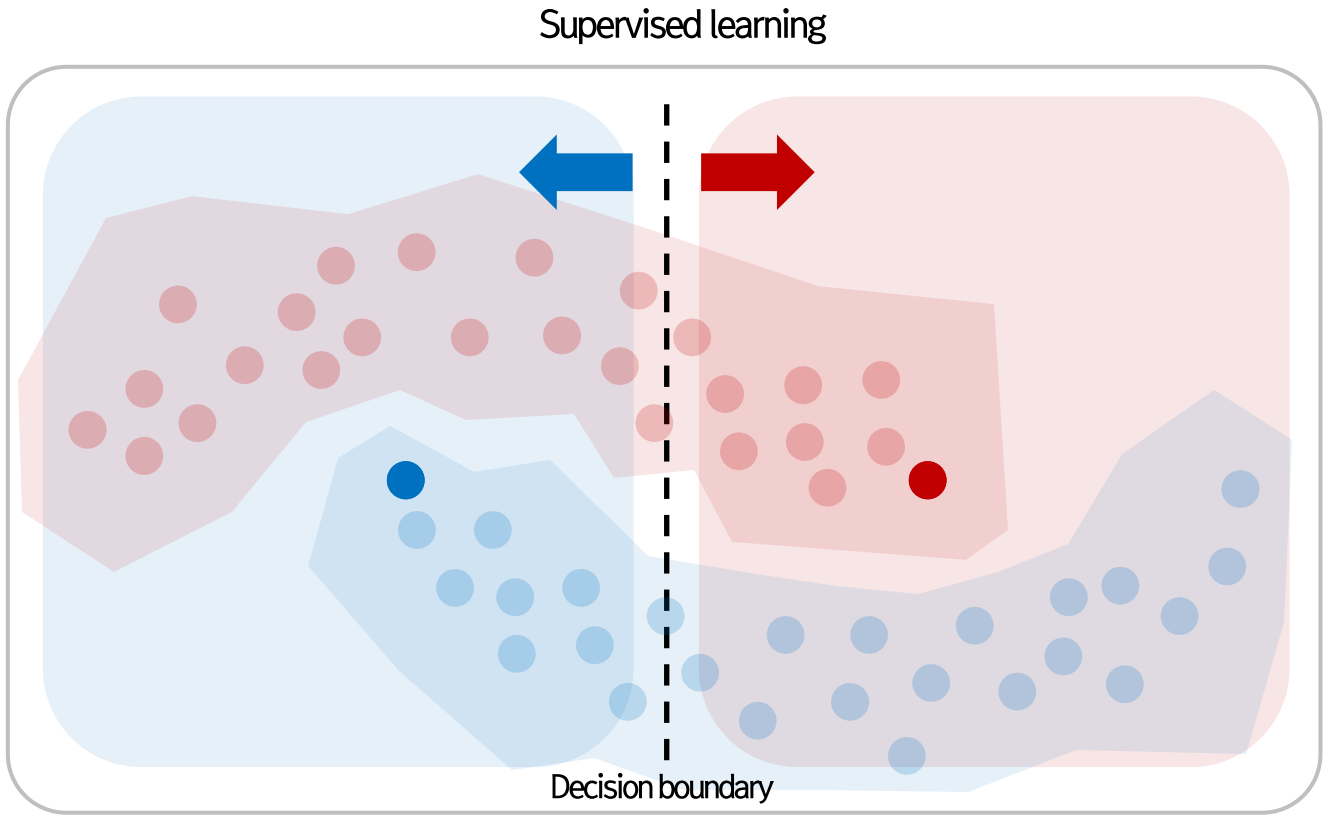


Introduction

Semi-Supervised Learning

❖ Semi-supervised learning 방식이 supervised learning에 비해서 성능이 좋을까?

- 소량의 labeled data와 다량의 unlabeled data를 사용하여 분류 모델의 일반화 성능을 높여보자!

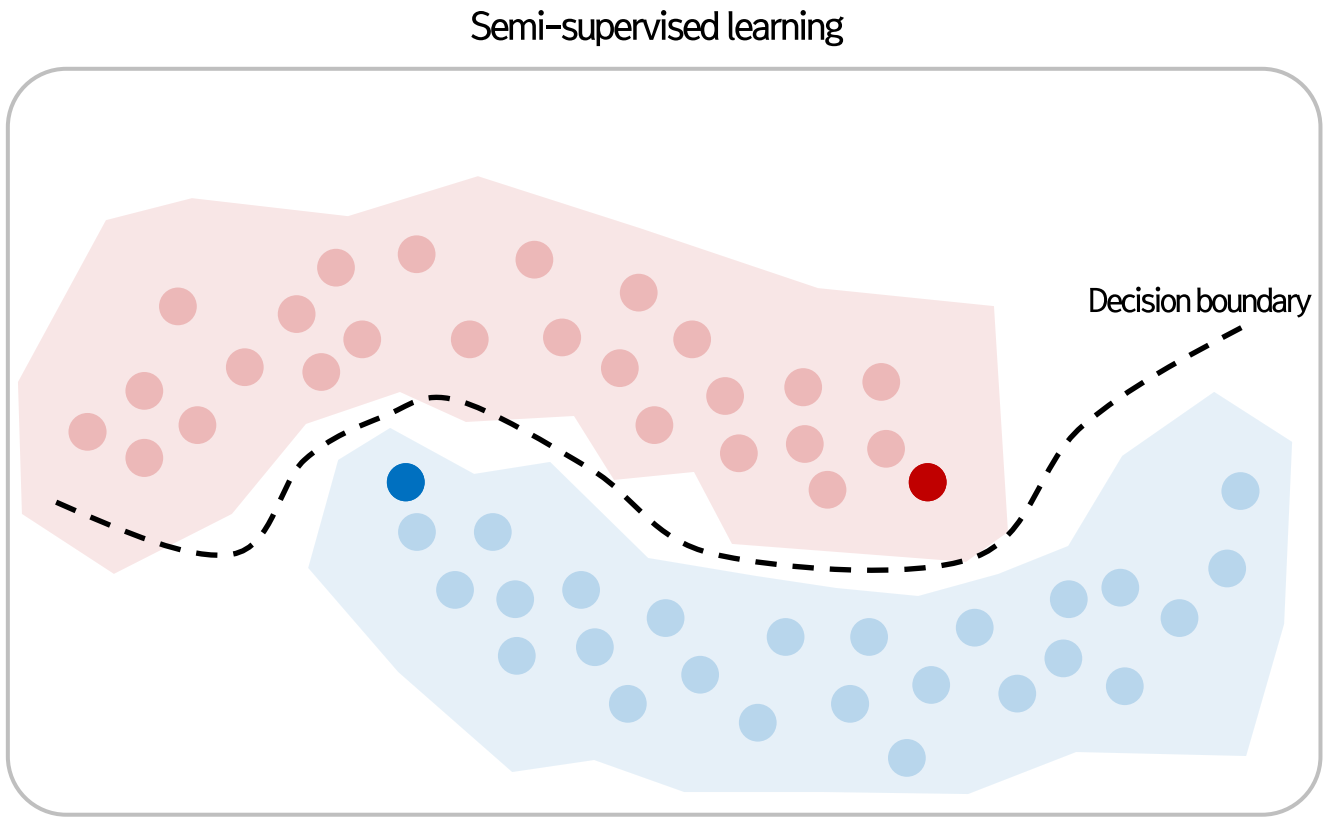


Introduction

Semi-Supervised Learning

❖ Semi-supervised learning 방식이 supervised learning에 비해서 성능이 좋을까?

- 소량의 labeled data와 다량의 unlabeled data를 사용하여 분류 모델의 일반화 성능을 높여보자!



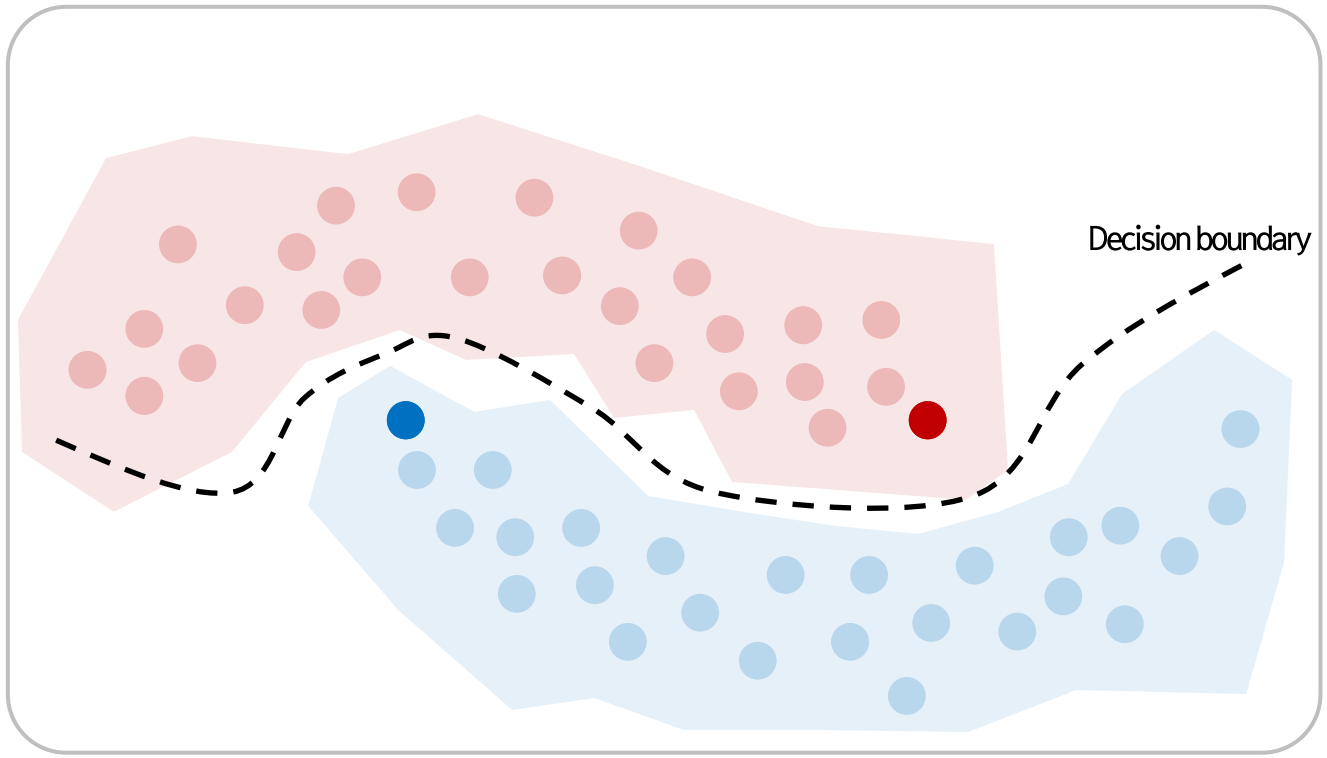
Introduction

Semi-Supervised Learning

❖ Semi-supervised learning 방식이 supervised learning에 비해서 성능이 좋을까?

- 소량의 labeled data와 다량의 unlabeled data를 사용하여 분류 모델의 일반화 성능을 높여보자!

Semi-supervised learning



✓ Cluster assumption

decision boundary가
low-density area에서 형성됨

Unlabeled data가 의미 있게
사용됨

Introduction

Semi-Supervised Learning

❖ FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence (NeurIPS, 2020)

- Google research에서 연구된 논문이며, 2023년 3월 28일 기준 1724회 인용됨
- 본 세미나 주제인 Class-Imbalanced Semi-Supervised Learning 연구 분야의 기준이 되는 방법론

FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence

Kihyuk Sohn* **David Berthelot*** **Chun-Liang Li** **Zizhao Zhang** **Nicholas Carlini**
Ekin D. Cubuk **Alex Kurakin** **Han Zhang** **Colin Raffel**
Google Research
{kihyuks,dberth,chunliang,zizhaoz,ncarlini,
cubuk,kurakin,zhanghan,craffel}@google.com

Abstract

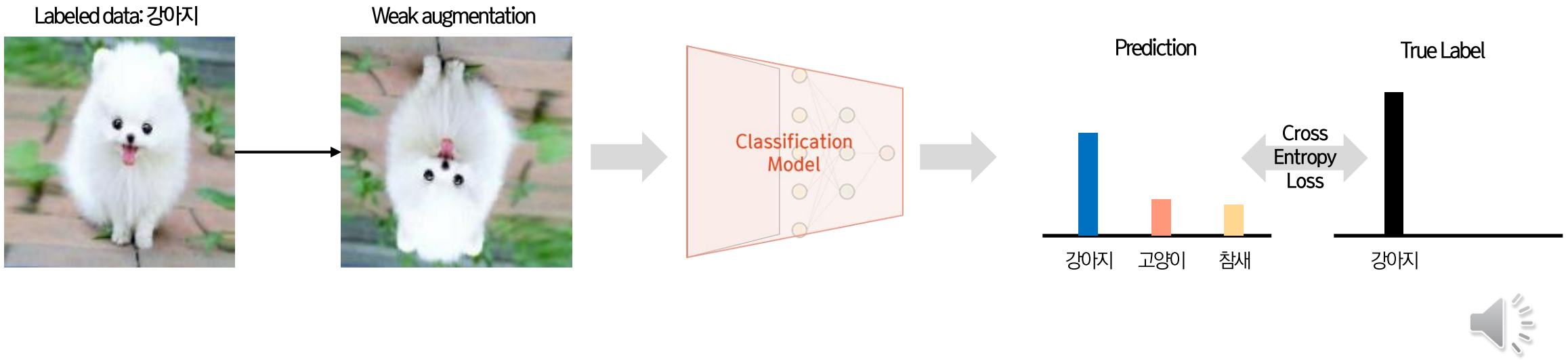
Semi-supervised learning (SSL) provides an effective means of leveraging unlabeled data to improve a model's performance. This domain has seen fast progress recently, at the cost of requiring more complex methods. In this paper we propose FixMatch, an algorithm that is a significant simplification of existing SSL methods. FixMatch first generates pseudo-labels using the model's predictions on weakly-augmented unlabeled images. For a given image, the pseudo-label is only retained if the model produces a high-confidence prediction. The model is then trained to predict the pseudo-label when fed a strongly-augmented version of the same image. Despite its simplicity, we show that FixMatch achieves state-of-the-art performance across a variety of standard semi-supervised learning benchmarks, including 94.93% accuracy on CIFAR-10 with 250 labels and 88.61% accuracy with 40 – just 4 labels per class. We carry out an extensive ablation study to tease apart the experimental factors that are most important to FixMatch's success. The code is available at https://github.com/google-research/fixmatch.

Introduction

Semi-Supervised Learning

❖ Consistency regularization과 pseudo labeling 두 기법을 결합한 구조로 구성

- Labeled data에 weak augmentation을 적용하여 cross entropy loss (supervised loss)로 모델 학습
- Unlabeled data는 weak augmentation이 적용된 data의 예측 확률분포를 one-hot label로 변경
- Strong augmentation이 적용된 data의 target으로 pseudo label을 활용하여 cross entropy loss (unsupervised loss)로 학습

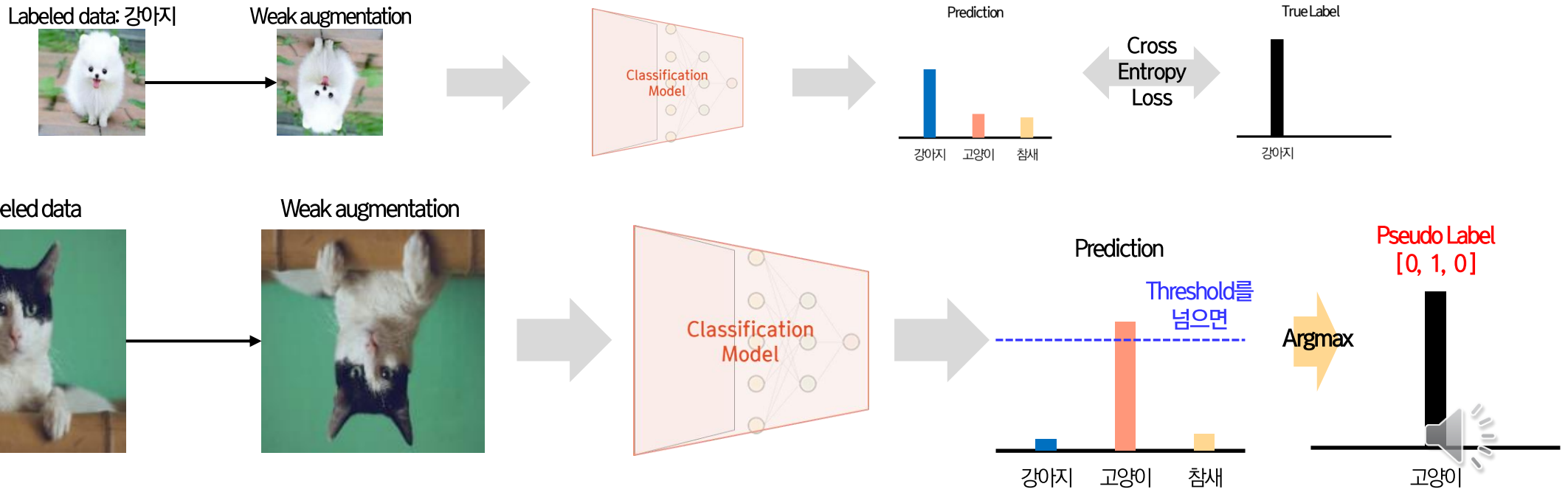


Introduction

Semi-Supervised Learning

❖ Consistency regularization과 pseudo labeling 두 기법을 결합한 구조로 구성

- Labeled data에 weak augmentation을 적용하여 cross entropy loss (supervised loss)로 모델 학습
- Unlabeled data는 weak augmentation이 적용된 data의 예측 확률분포를 one-hot label로 변경
- Strong augmentation이 적용된 data의 target으로 pseudo label을 활용하여 cross entropy loss (unsupervised loss)로 학습

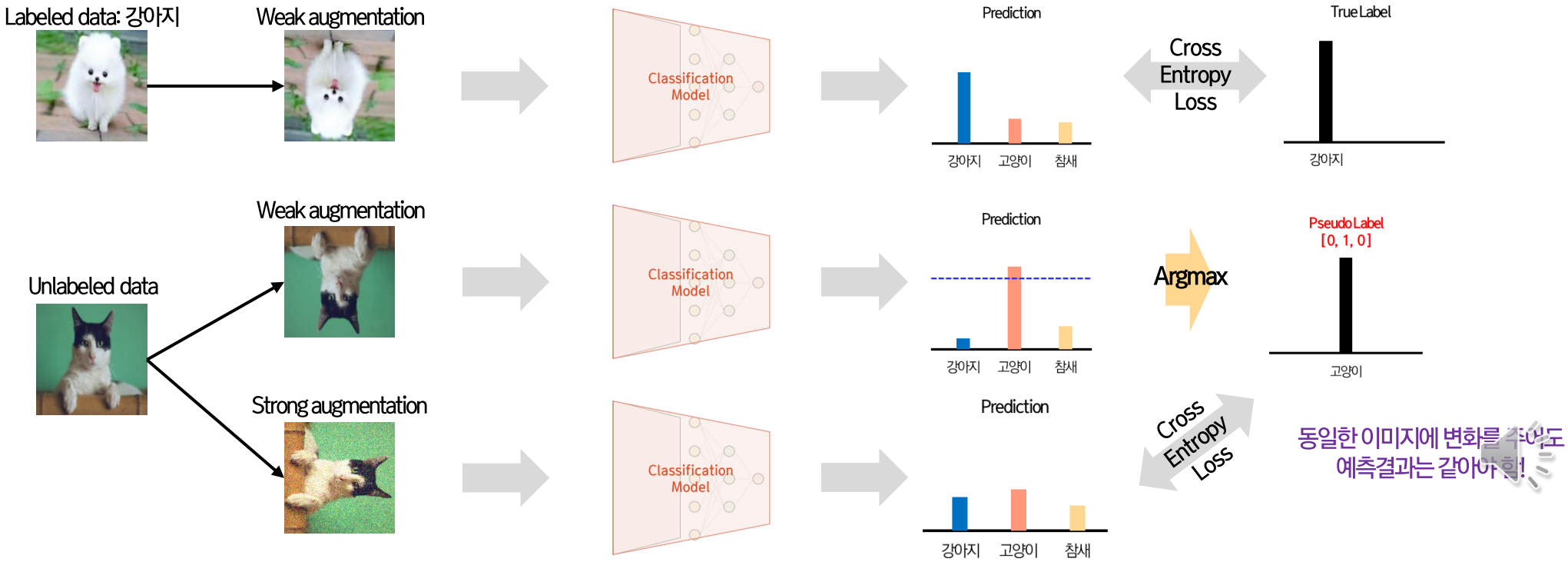


Introduction

Semi-Supervised Learning

❖ Consistency regularization과 pseudo labeling 두 기법을 결합한 구조로 구성

- Labeled data에 weak augmentation을 적용하여 cross entropy loss (supervised loss)로 모델 학습
- Unlabeled data는 weak augmentation이 적용된 data의 예측 확률분포를 one-hot label로 변경
- Strong augmentation이 적용된 data의 target으로 pseudo label을 활용하여 cross entropy loss (unsupervised loss)로 학습

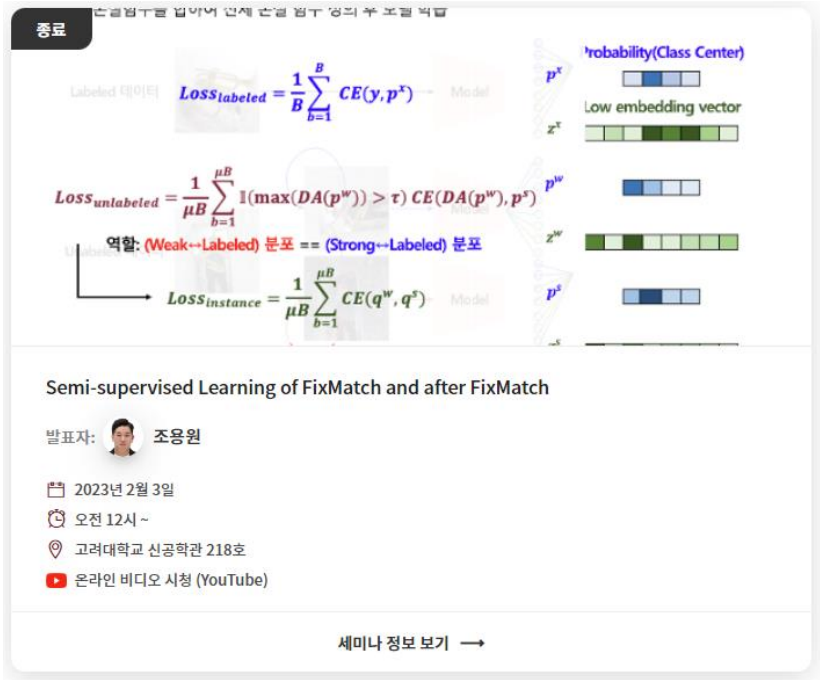


Introduction

Semi-Supervised Learning

최신 Semi-supervised learning 연구 동향?

Consistency regularization과 pseudo labeling 기반 방법론

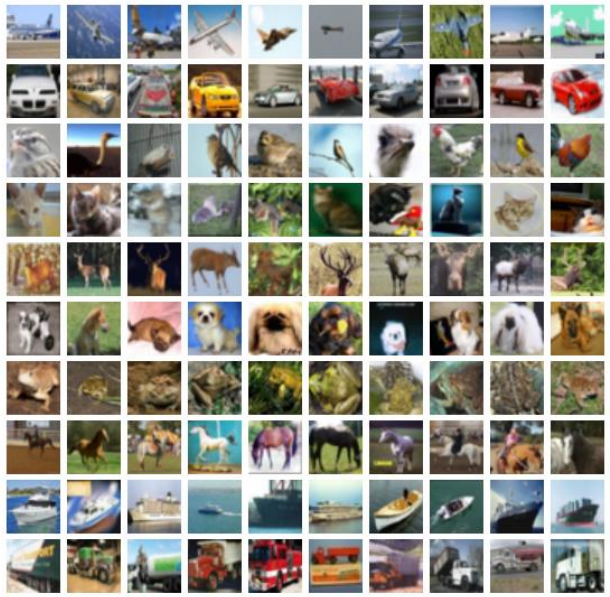


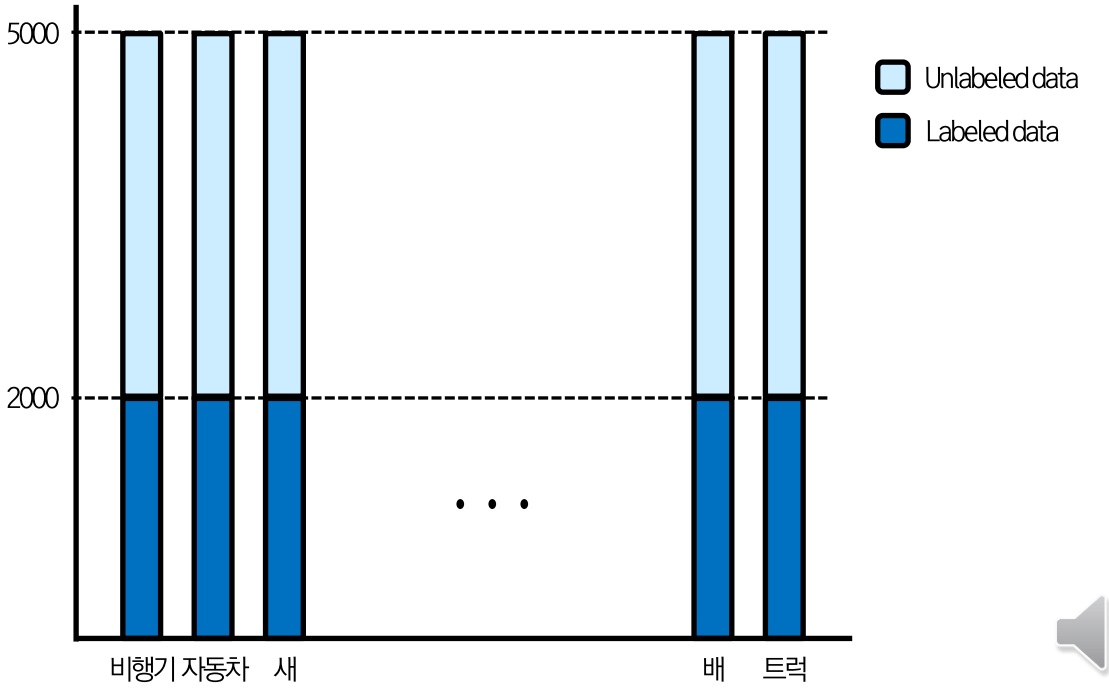
Introduction

Semi-Supervised Learning in class-imbalanced scenario

❖ FixMatch 및 semi-supervised learning 연구에 사용된 데이터셋

- 모든 labeled data와 unlabeled data 샘플 수가 클래스 별로 균일한 클래스 분포를 가짐 (Class-balanced)
- e.g. CIFAR-10, CIFAR-100, STL-10 etc.





[SSL 연구에 활용되는 CIFAR-10 예시]

Introduction

Semi-Supervised Learning in class-imbalanced scenario

❖ FixMatch 및 semi-supervised learning 연구에 사용된 데이터셋

- 모든 labeled data와 unlabeled data 샘플 수가 클래스 별로 균일한 클래스 분포를 가짐 (Class-balanced)
- e.g. CIFAR-10, CIFAR-100, STL-10 etc.

현실에서 사용되는 대다수 데이터셋들은 실제로 클래스별 균일 분포?

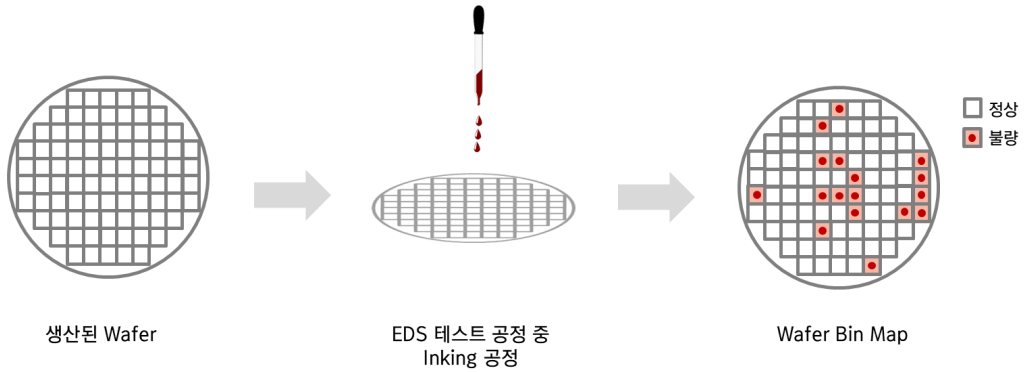
[SSL 연구에 활용되는 CIFAR-10 예시]

Introduction

Semi-Supervised Learning in class-imbalanced scenario

❖ 현실 데이터셋은 클래스별 불균형인 데이터가 대다수

- 불균형 데이터는 클래스별 데이터 수가 균일하게 분포되어 있지 않은 경우이며, 클래스 수에 따라 다수와 소수 클래스로 볼 수 있음
- 제조 공정, 질병 진단, 금융 분석 등의 수 많은 분야에서 다루는 대다수 데이터는 불균형 데이터



반도체 제조 공정 중 Wafer Bin Map 데이터 생성 과정

제조 공정을 통해
데이터셋 생성

	Class Label	Sample counts	Proportion (%)
불량	Center	4,294	2.48
	Donut	555	0.32
	Edge-Loc	5,189	3.00
	Edge-Ring	9,680	5.59
	Loc	3,593	2.07
	Random	886	0.51
	Scratch	1,193	0.68
	Near-Full	149	0.09
정상	None	147,431	85.24

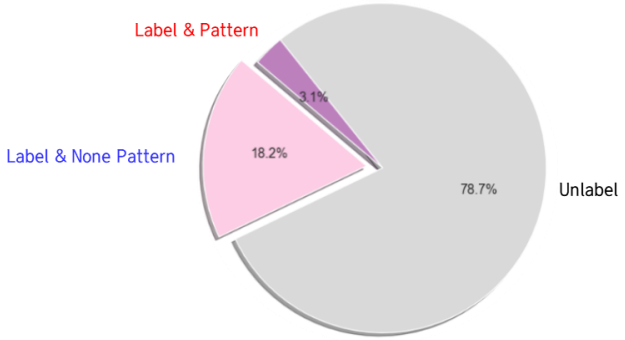
[실제 반도체 웨이퍼 빈 맵 분석 데이터셋]

Introduction

Semi-Supervised Learning in class-imbalanced scenario

❖ 데이터가 불균형 상황일 경우 semi-supervised learning은 어떻게 작동할까?

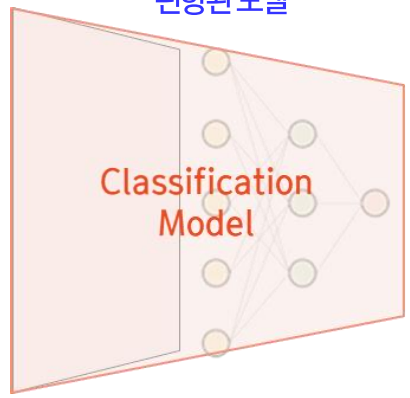
[실제 반도체 웨이퍼 빈 맵 분석 데이터셋]



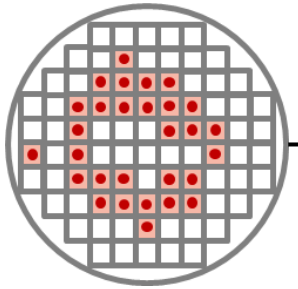
	Class Label	Sample counts	Proportion (%)
불량	Center	4,294	2.48
	Donut	555	0.32
	Edge-Loc	5,189	3.00
	Edge-Ring	9,680	5.59
	Loc	3,593	2.07
	Random	886	0.51
	Scratch	1,193	0.68
	Near-Full	149	0.09
정상	None	147,431	85.24

분류 모델 학습

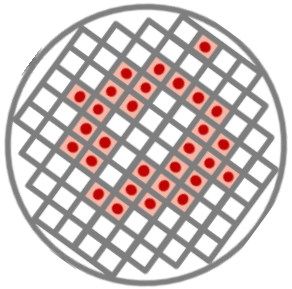
대다수인 정상 클래스에 편향된 모델

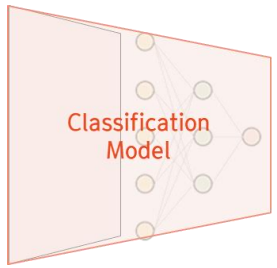


Unlabeled data

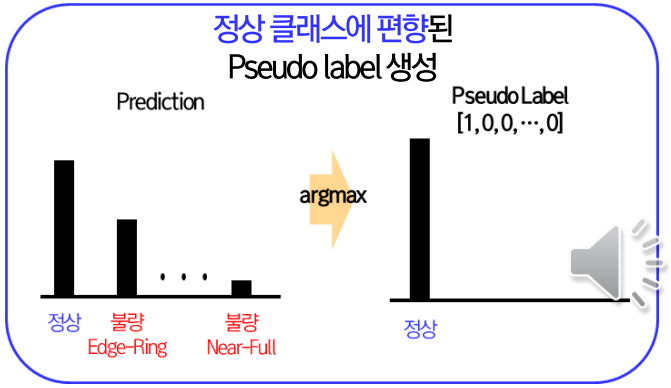


Weak augmentation



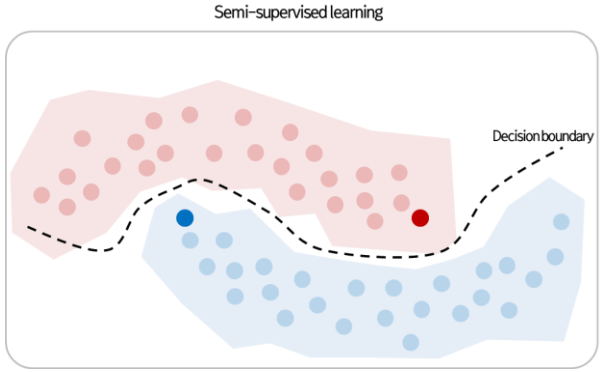


정상 클래스에 편향된 Pseudo label 생성



Introduction

Semi-Supervised Learning in class-imbalanced scenario



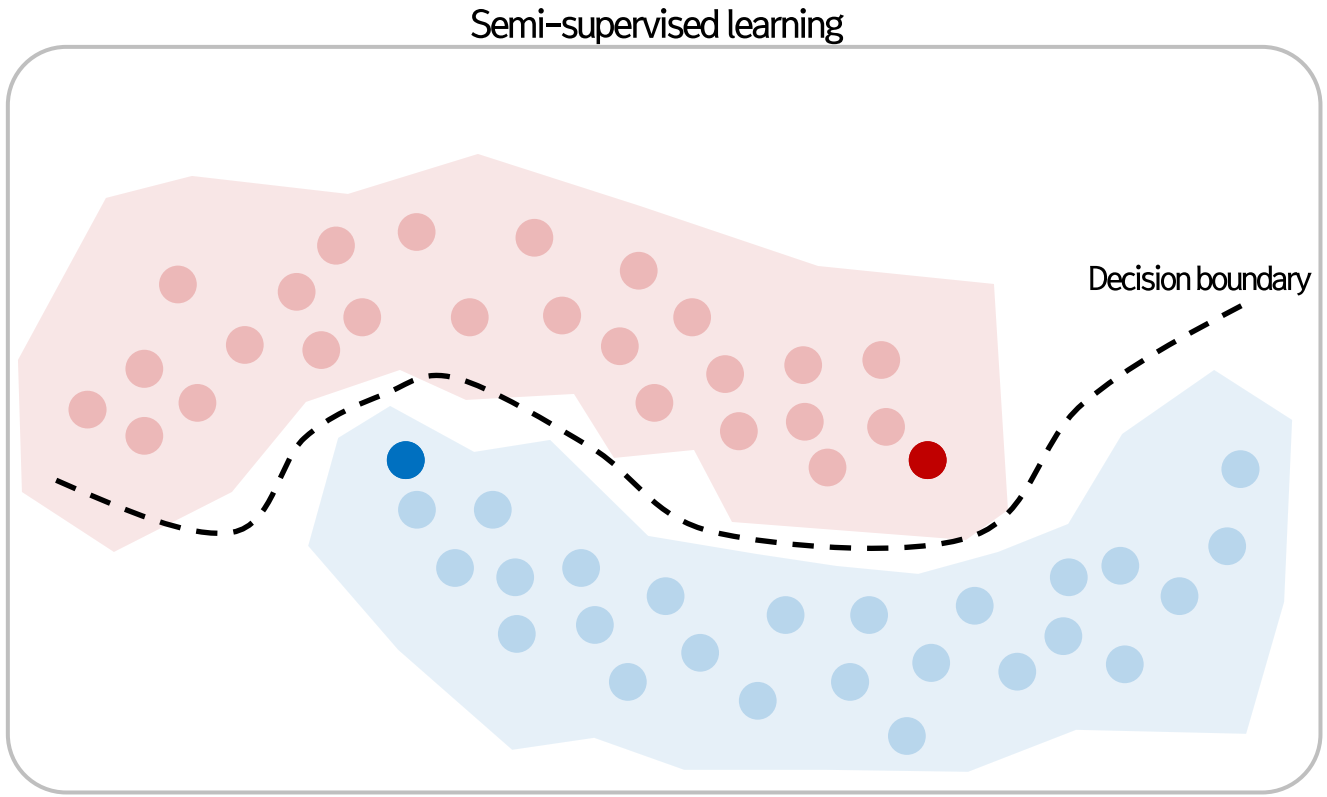
✓ Cluster assumption

decision boundary가 low-density area에서 형성됨

Unlabeled data가 의미 있게 사용됨

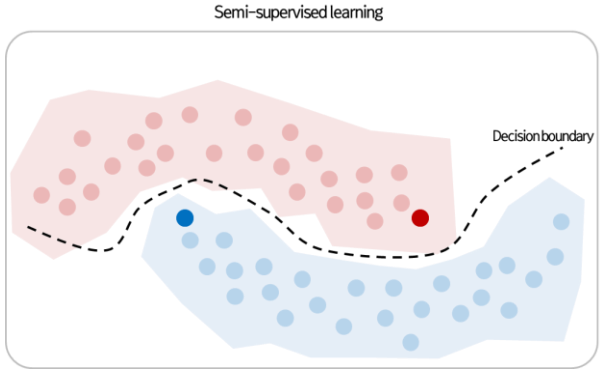
❖ 클래스 불균형인 데이터상에서 semi-supervised learning 성능이 하락함

- 다수의 labeled data, unlabeled data 클래스로 인해 분류 경계선이 편향됨
- 소수 클래스 지역으로 밀린 분류 경계선이 소수 클래스 high-density area를 지나게 됨



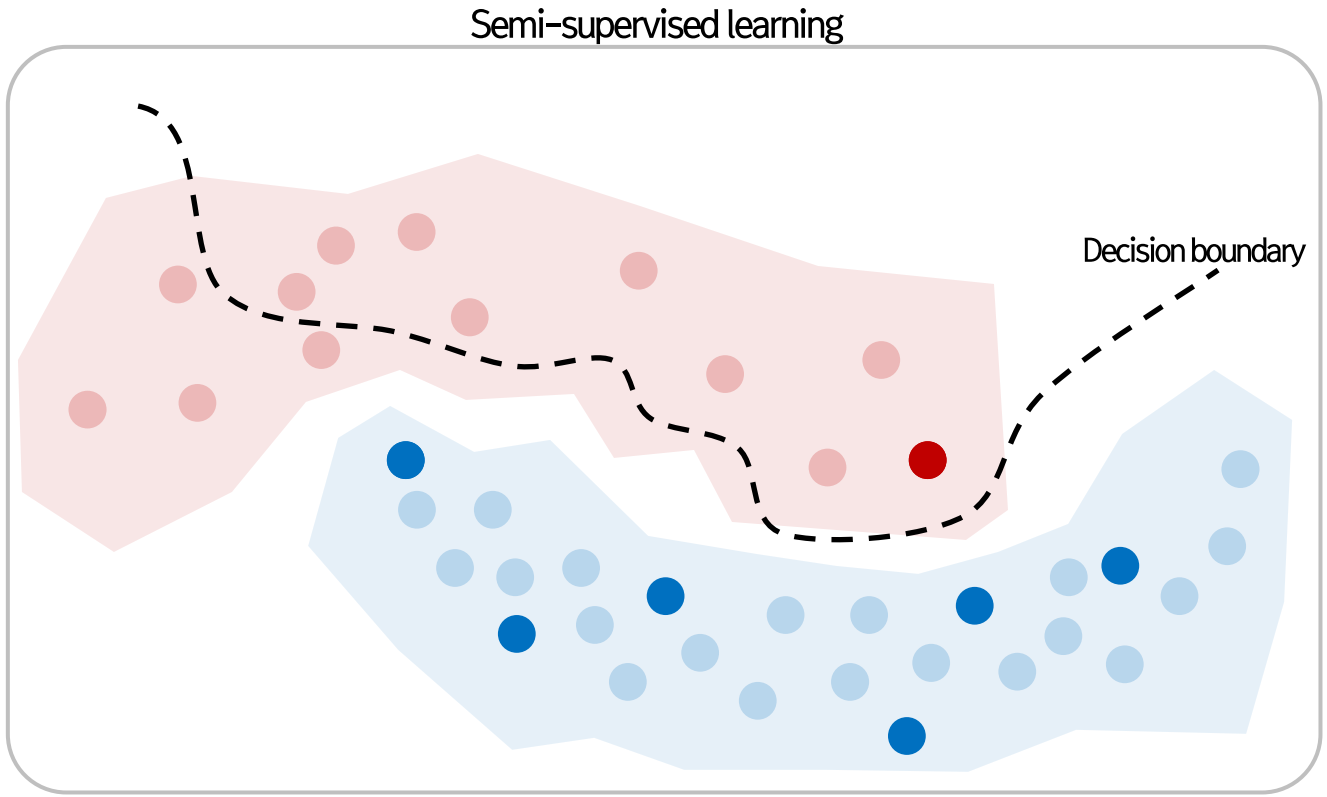
Introduction

Semi-Supervised Learning in class-imbalanced scenario



❖ 클래스 불균형인 데이터상에서 semi-supervised learning 성능이 하락함

- 다수의 labeled data, unlabeled data 클래스로 인해 분류 경계선이 편향됨
- 소수 클래스 지역으로 밀린 분류 경계선이 소수 클래스 high-density area를 지나게 됨



Class-Imbalanced Semi-Supervised Learning

Class-Imbalanced Semi-Supervised Learning

클래스 불균형 데이터셋에서도

성능을 유지하면서

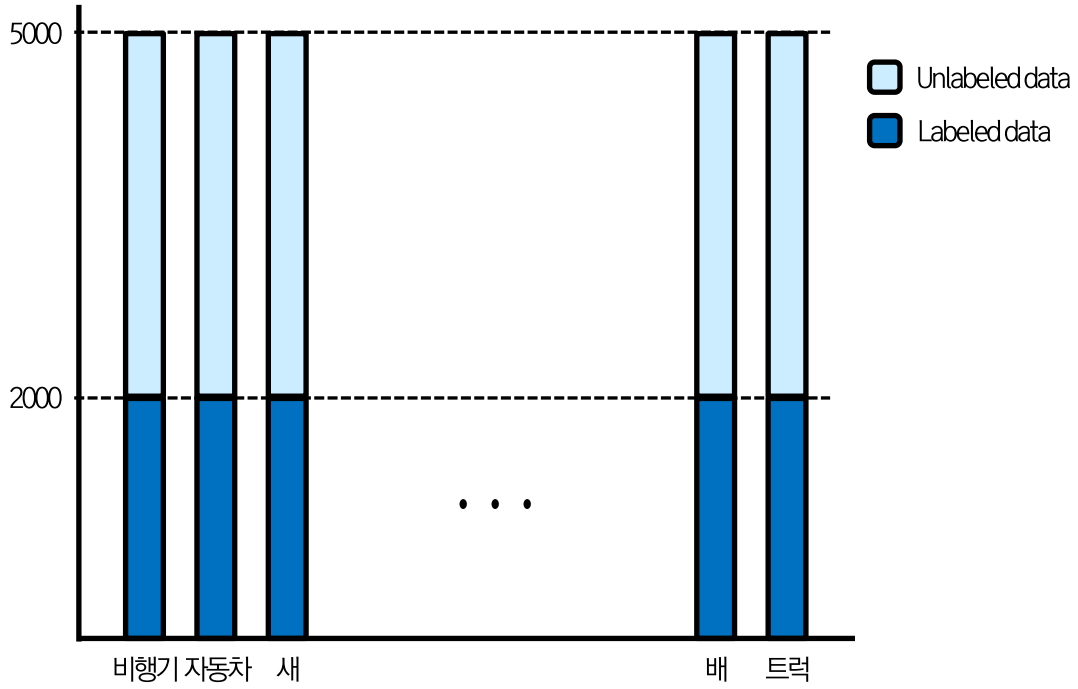
Labeled data가 많이 없더라도

unlabeled data를 이용하여 성능을 높이자!

Class-Imbalanced Semi-Supervised Learning

❖ Class-imbalanced semi-supervised learning 연구에 활용되는 데이터

- Semi-supervised learning 연구에서 사용되는 데이터셋을 불균형 데이터셋으로 변환하여 사용함

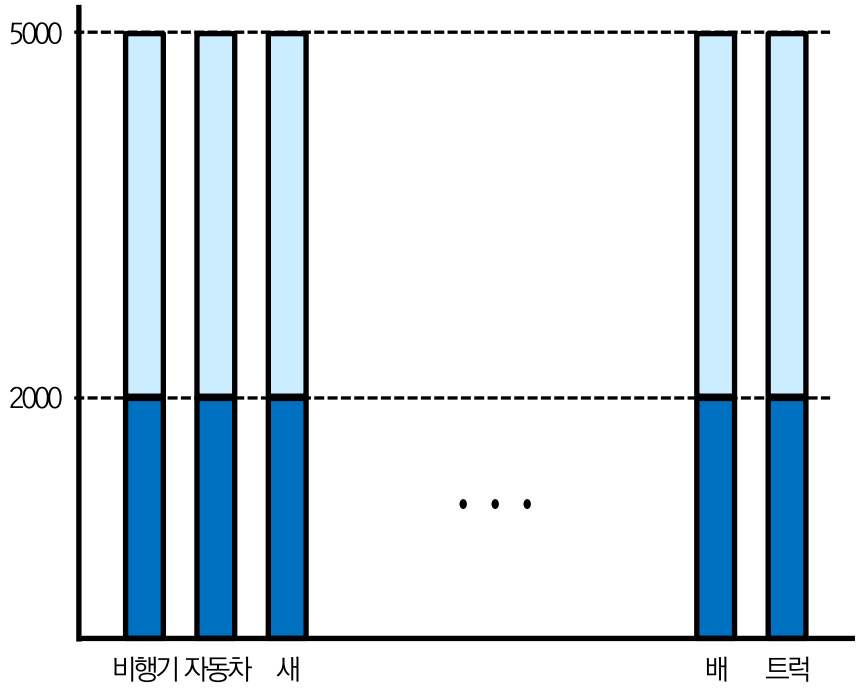


[SSL 연구에 활용되는 CIFAR-10 예시]

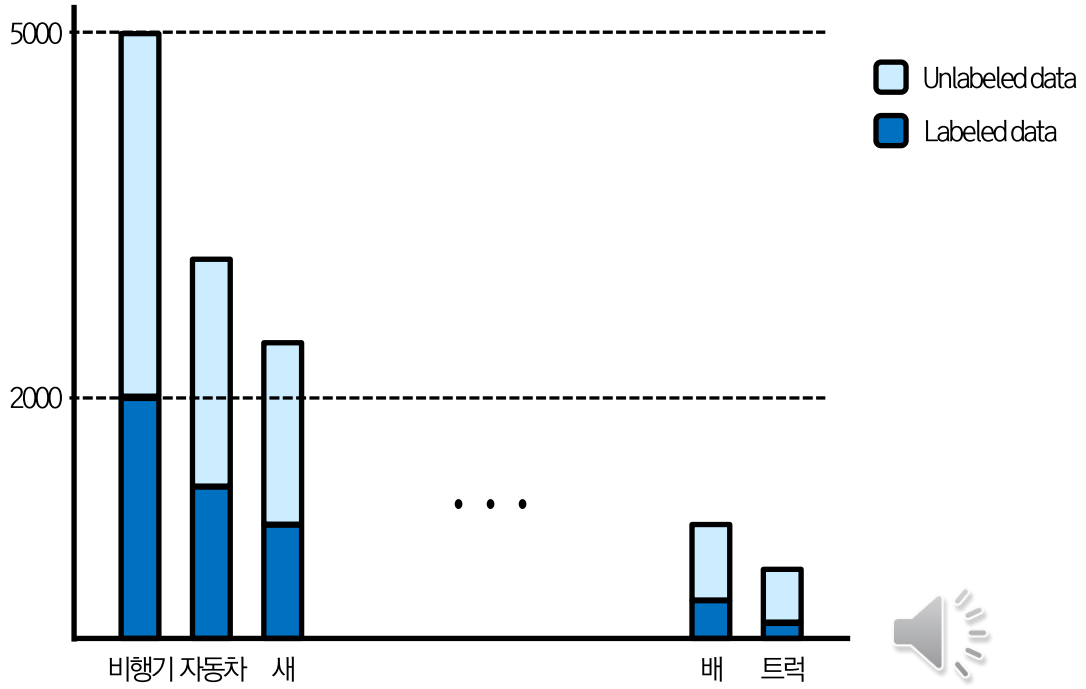
Class-Imbalanced Semi-Supervised Learning

❖ Class-imbalanced semi-supervised learning 연구에 활용되는 데이터

- Semi-supervised learning 연구에서 사용되는 데이터셋을 불균형 데이터셋으로 변환하여 사용함



[SSL 연구에 활용되는 CIFAR-10 예시]



[Class-imbalanced SSL 연구에 활용되는 CIFAR-10 예시]

Class-Imbalanced Semi-Supervised Learning

❖ 클래스 불균형 상황에서도 SSL 성능이 유지가 되게 하는 방법론들

- Pseudo labeling 기반 SSL 방법론들에 적용할 수 있는 기법 제시(학습 메커니즘)
  - ✓ DARP: Distribution Aligning Refinery of Pseudo-label for Imbalanced Semi-supervised Learning
  - ✓ CReST: A Class-Rebalancing Self-Training Framework for Imbalanced Semi-Supervised Learning

- SSL 모델 중 FixMatch 모델에서 고정된 threshold 부분을 바꿈(모델 아키텍처 변경)
  - ✓ Adsh: Class-Imbalanced Semi-Supervised Learning with Adaptive Thresholding

Class-Imbalanced Semi-Supervised Learning

DARP: Distribution Aligning Refinery of Pseudo-label for Imbalanced Semi-supervised Learning

❖ DARP: Distribution Aligning Refinery of Pseudo-label for Imbalanced Semi-supervised Learning (NeurIPS, 2020)

- KAIST에서 연구된 논문이며, 2023년 3월 28일 기준 89회 인용됨
- FixMatch 등 pseudo labeling 기법이 사용되는 모든 semi-supervised learning에 적용할 수 있음

Distribution Aligning Refinery of Pseudo-label for Imbalanced Semi-supervised Learning

Jaehyung Kim[1], **Youngbum Hur**[2], **Sejun Park**[1], **Eunho Yang**[1,3], **Sung Ju Hwang**[1,3], **Jinwoo Shin**[1]
[1]Korea Advanced Institute of Science and Technology (KAIST)
[2]Samsung Advanced Institute of Technology
[3]AItrics
{jaehyungkim,sejun.park,sjhwang82,jinwoos}@kaist.ac.kr
youngbum.hur@samsung.com yangeh@gmail.com

Abstract

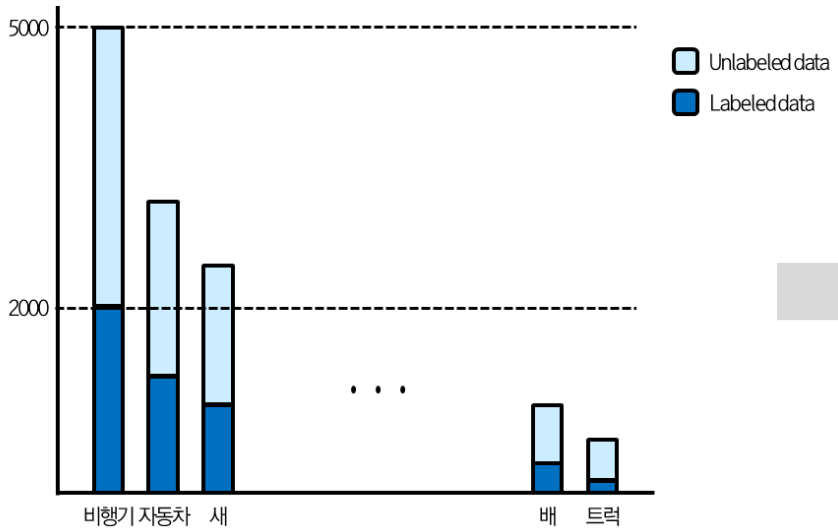
While semi-supervised learning (SSL) has proven to be a promising way for leveraging unlabeled data when labeled data is scarce, the existing SSL algorithms typically assume that training class distributions are balanced. However, these SSL algorithms trained under imbalanced class distributions can severely suffer when generalizing to a balanced testing criterion, since they utilize biased pseudo-labels of unlabeled data toward majority classes. To alleviate this issue, we formulate a convex optimization problem to softly refine the pseudo-labels generated from a biased model, and develop a simple iterative algorithm, named Distribution Aligning Refinery of Pseudo-label (DARP) that solves it provably and efficiently. Under various class-imbalanced semi-supervised scenarios, we demonstrate the effectiveness of DARP and its compatibility with state-of-the-art SSL schemes.

Class-Imbalanced Semi-Supervised Learning

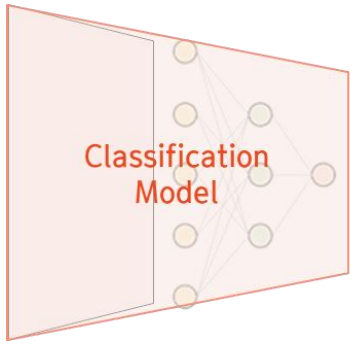
DARP: Distribution Aligning Refinery of Pseudo-label for Imbalanced Semi-supervised Learning

❖ DARP: Distribution Aligning Refinery of Pseudo-label for Imbalanced Semi-supervised Learning (NeurIPS, 2020)

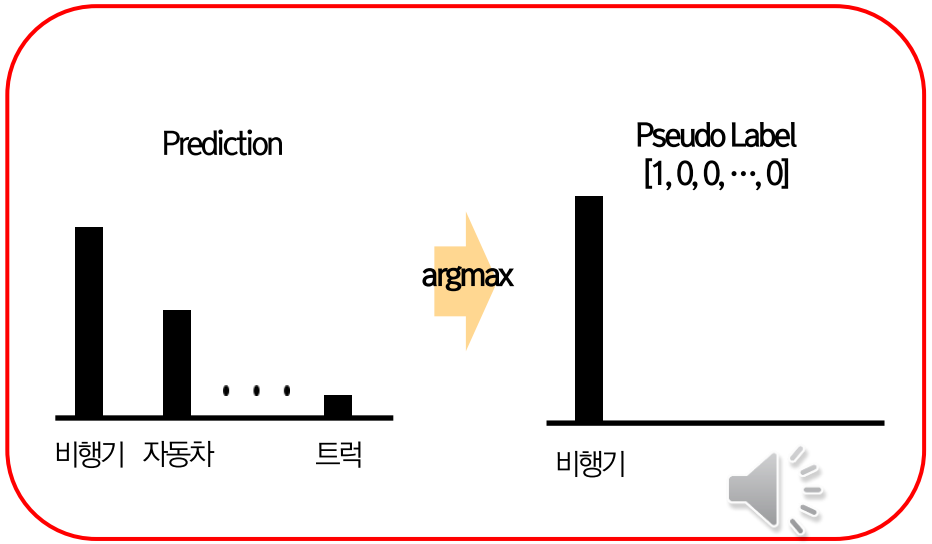
- ➢ 클래스 불균형 데이터로 모델을 학습하면 다수 클래스에 편향된 분류 모델이 생성됨
- ➢ 편향된 분류 모델이 unlabeled data에 대해서 다수 클래스로 편향되어 pseudo label 생성함
  - ✓ 모델 학습이 점차 편향된 방향으로 진행되며 분류 성능 하락으로 이어짐



[Class-imbalanced SSL 연구에 활용되는 CIFAR-10 예시]



Biased pseudo label 생성 후
모델 학습에 사용함



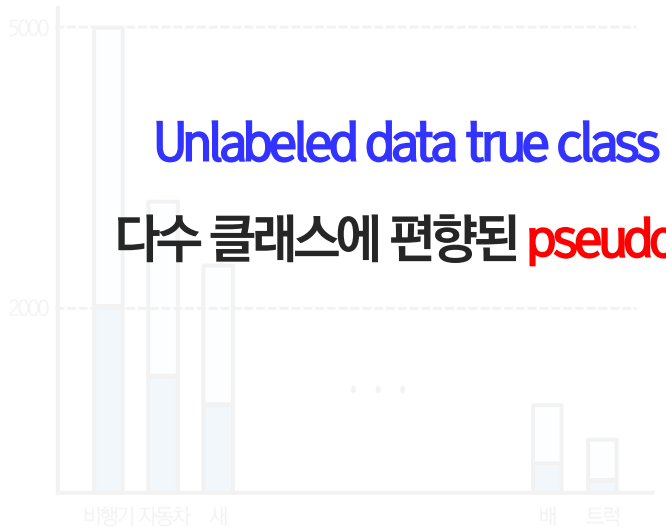
→ 모델 학습이 편향된 방향으로 점차 진행됨

Class-Imbalanced Semi-Supervised Learning

DARP: Distribution Aligning Refinery of Pseudo-label for Imbalanced Semi-supervised Learning

❖ DARP: Distribution Aligning Refinery of Pseudo-label for Imbalanced Semi-supervised Learning(NeurIPS, 2020)

- ➢ 클래스 불균형 데이터로 모델을 학습하면 다수 클래스에 편향된 분류 모델이 생성됨
- ➢ 편향된 분류 모델이 unlabeled data에 대해서 다수 클래스로 편향되어 pseudo label 생성함
  - ✓ 모델 학습이 점차 편향된 방향으로 진행되며 분류 성능 하락으로 이어짐



[Class-imbalanced SSL 연구에 활용되는 CIFAR-10 예시]

Unlabeled data true class 분포를 활용해서

다수 클래스에 편향된 pseudo label을 개선하자!

① 실제 unlabeled data 클래스 분포를 따라 가도록 refined pseudo label 분포를 조정함

② 단, refined pseudo label과 original pseudo label이 서로 너무 멀어지지 않도록

Class-Imbalanced Semi-Supervised Learning

DARP: Distribution Aligning Refinery of Pseudo-label for Imbalanced Semi-supervised Learning

① 실제 unlabeled data 클래스 분포를 따라 가도록 refined pseudo label 분포를 조정함

② 단, refined pseudo label과 original pseudo label이 서로 너무 멀어지지 않도록

두 가지 목표를 Convex Optimization Problem으로 제안함

$$minimize \quad \sum_{m=1}^{M} w_m \, D_{KL}(\hat{y}_m || \hat{y}_m^{unlabeled})$$

$$subject\ to \quad \sum_{m=1}^{M} \hat{y}_m(k) = M_k \,, \forall k, \qquad \sum_{k=1}^{K} \hat{y}_m(k) = 1, \ \forall m, \qquad \hat{y}_m(k) \in [0,1], \forall m, k$$

Class-Imbalanced Semi-Supervised Learning

DARP: Distribution Aligning Refinery of Pseudo-label for Imbalanced Semi-supervised Learning

① 실제 unlabeled data 클래스 분포를 따라 가도록 refined pseudo label 분포를 조정함

② 단, refined pseudo label과 original pseudo label이 서로 너무 멀어지지 않도록

두 가지 목표를 Convex Optimization Problem으로 제안함

$$minimize \quad \sum_{m=1}^{M} w_m \, D_{KL}(\hat{y}_m || \hat{y}_m^{unlabeled})$$

$$subject\ to \quad \sum_{m=1}^{M} \hat{y}_m(k) = M_k, \forall k, \qquad \sum_{k=1}^{K} \hat{y}_m(k) = 1, \ \forall m, \qquad \hat{y}_m(k) \in [0,1], \forall m, k$$

Refined pseudo label 수와 실제 unlabeled data 개수를 같게 만들자

Class-Imbalanced Semi-Supervised Learning

DARP: Distribution Aligning Refinery of Pseudo-label for Imbalanced Semi-supervised Learning

① 실제 unlabeled data 클래스 분포를 따라 가도록 refined pseudo label 분포를 조정함

② 단, refined pseudo label과 original pseudo label이 서로 너무 멀어지지 않도록

두 가지 목표를 Convex Optimization Problem으로 제안함

Refined pseudo label

Original pseudo label

$$minimize \quad \sum_{m=1}^{M} w_m \, D_{KL}(\hat{y}_m || \hat{y}_m^{unlabeled})$$

KL Divergence값을 최소화하면서
Original pseudo label 정보를 보존해주자

$$subject\ to \quad \sum_{m=1}^{M} \hat{y}_m(k) = M_k, \forall k, \qquad \sum_{k=1}^{K} \hat{y}_m(k) = 1, \ \forall m, \qquad \hat{y}_m(k) \in [0,1], \forall m, k$$

Refined pseudo label 수와 실제 unlabeled data 개수를 같게 만들자

Class-Imbalanced Semi-Supervised Learning

DARP: Distribution Aligning Refinery of Pseudo-label for Imbalanced Semi-supervised Learning

① 실제 unlabeled data 클래스 분포를 따라 가도록 refined pseudo label 분포를 조정함

② 단, refined pseudo label과 original pseudo label이 서로 너무 멀어지지 않도록

두 가지 목표를 Convex Optimization Problem으로 제안함

Entropy가 낮은 (= confident가 높은) unlabeled data에 대해서 중점적으로 보자

$w_m := (H(\hat{y}_m^{unlabeled}))^{-1}$

Refined pseudo label

Original pseudo label

$$minimize \quad \sum_{m=1}^{M} w_m \, D_{KL}(\hat{y}_m || \hat{y}_m^{unlabeled})$$

KL Divergence값을 최소화하면서
Original pseudo label 정보를 보존해주자

$$subject\ to \quad \sum_{m=1}^{M} \hat{y}_m(k) = M_k, \forall k, \qquad \sum_{k=1}^{K} \hat{y}_m(k) = 1, \ \forall m, \qquad \hat{y}_m(k) \in [0,1], \forall m, k$$

Refined pseudo label 수와 실제 unlabeled data 개수를 같게 만들자

Class-Imbalanced Semi-Supervised Learning

DARP: Distribution Aligning Refinery of Pseudo-label for Imbalanced Semi-supervised Learning

① 실제 unlabeled data 클래스 분포를 따라 가도록 refined pseudo label 분포를 조정함

② 단, refined pseudo label과 original pseudo label이 서로 너무 멀어지지 않도록

두 가지 목표를 Convex Optimization Problem으로 제안함

Entropy가 낮은 (= confident가 높은) unlabeled data에 대해서 중점적으로 보자

$w_m := (H(\hat{y}_m^{unlabeled}))^{-1}$

Refined pseudo label

Original pseudo label

$$minimize \quad \sum_{m=1}^{M} w_m \, D_{KL}(\hat{y}_m || \hat{y}_m^{unlabeled})$$

KL Divergence값을 최소화하면서
Original pseudo label 정보를 보존해주자

$$subject\ to \quad \sum_{m=1}^{M} \hat{y}_m(k) = M_k, \forall k, \qquad \sum_{k=1}^{K} \hat{y}_m(k) = 1, \ \forall m, \qquad \hat{y}_m(k) \in [0,1], \forall m, k$$

Refined pseudo label 수와 실제 unlabeled data 개수를 같게 만들자

Unlabeled data 개수 M_k 는 실제로 알 수 없으니 추정을 해서 대체하자

Class-Imbalanced Semi-Supervised Learning

DARP: Distribution Aligning Refinery of Pseudo-label for Imbalanced Semi-supervised Learning

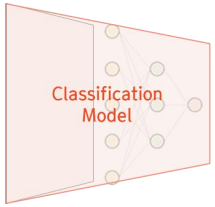
Unlabeled data 개수 M_k 는 실제로 알 수 없으니 추정을 해서 대체하자!

$$\begin{bmatrix} M_1 \\ M_2 \\ \vdots \\ \vdots \\ \vdots \\ M_K \end{bmatrix} = \underbrace{(C^{unlabeled})^{-1}}_{confusion\,matrix_{k\times k}} \times \begin{bmatrix} \sum_{m=1}^{M} f_1(x_m^{unlabeled}) \\ \vdots \\ \sum_{m=1}^{M} f_K(x_m^{unlabeled}) \end{bmatrix}, \qquad C_{ij}^{unlabeled} := \frac{\sum_{m:y_m^{unlabeled}(j)=1} f_i(x_m^{unlabeled})}{\left|\{m|y_m^{unlabeled}(j)=1\}\right|}$$

Class-Imbalanced Semi-Supervised Learning

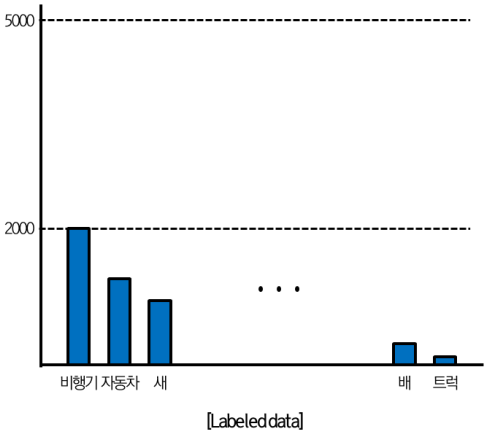
DARP: Distribution Aligning Refinery of Pseudo-label for Imbalanced Semi-supervised Learning

본 학습 사용 모델



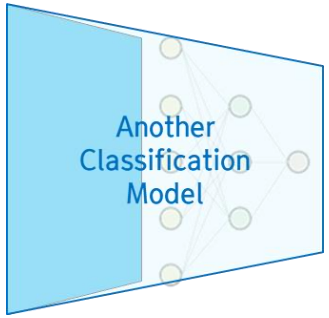
Unlabeled data 개수 M_k 는 실제로 알 수 없으니 추정을 해서 대체하자!

$$\begin{bmatrix} M_1 \\ M_2 \\ \vdots \\ \vdots \\ \vdots \\ M_K \end{bmatrix} = \underbrace{(C^{unlabeled})^{-1}}_{confusion\,matrix_{k\times k}} \times \begin{bmatrix} \sum_{m=1}^{M} f_1(x_m^{unlabeled}) \\ \vdots \\ \sum_{m=1}^{M} f_K(x_m^{unlabeled}) \end{bmatrix}, \qquad C_{ij}^{unlabeled} := \frac{\sum_{m:y_m^{unlabeled}(j)=1} f_i(x_m^{unlabeled})}{\left|\{m|y_m^{unlabeled}(j)=1\}\right|}$$



클래스별로 10개씩
제외하고 나머지 학습
→ Estimate set

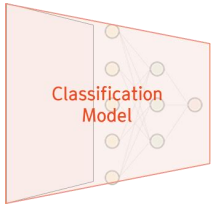
Vanilla scheme



Class-Imbalanced Semi-Supervised Learning

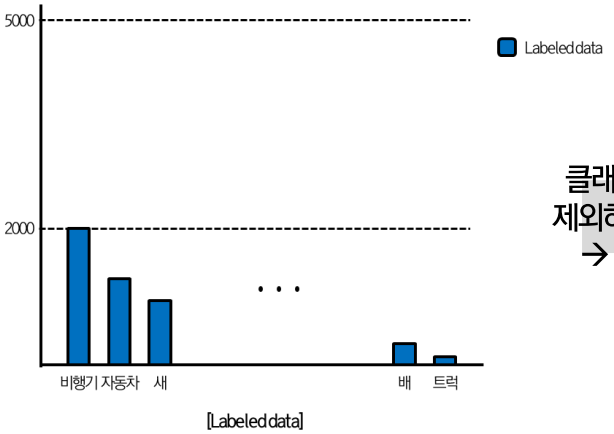
DARP: Distribution Aligning Refinery of Pseudo-label for Imbalanced Semi-supervised Learning

본 학습 사용 모델



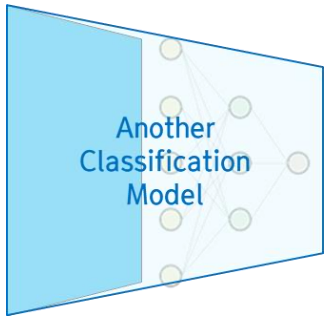
Unlabeled data 개수 M_k 는 실제로 알 수 없으니 추정을 해서 대체하자!

$$\begin{bmatrix} M_1 \\ M_2 \\ \vdots \\ \vdots \\ \vdots \\ M_K \end{bmatrix} = \underbrace{(C^{unlabeled})^{-1}}_{confusion\,matrix_{k\times k}} \times \begin{bmatrix} \sum_{m=1}^{M} f_1(x_m^{unlabeled}) \\ \vdots \\ \sum_{m=1}^{M} f_K(x_m^{unlabeled}) \end{bmatrix}, \qquad C_{ij}^{unlabeled} := \frac{\sum_{m:y_m^{unlabeled}(j)=1} f_i(x_m^{unlabeled})}{\left|\{m|y_m^{unlabeled}(j)=1\}\right|}$$



클래스별로 10개씩
제외하고 나머지 학습
→ Estimate set

Vanilla scheme



클래스별 10개 labeled data에
대해서 평가

Prediction $\begin{bmatrix} 10 & \cdots & 3 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 6 \end{bmatrix}$

True

$confusion\,matrix_{k\times k}$

Class-Imbalanced Semi-Supervised Learning

CReST: A Class-Rebalancing Self-Training Framework for Imbalanced Semi-Supervised Learning

❖ CReST: A Class-Rebalancing Self-Training Framework for Imbalance Semi-Supervised Learning (CVPR, 2021)

- ➢ Johns Hopkins university와 Google Cloud AI에서 연구했으며 2023년 3월 28일 기준 124회 인용됨
- ➢ FixMatch 등 pseudo labeling 기법이 사용되는 모든 semi-supervised learning에 적용할 수 있음

CReST: A Class-Rebalancing Self-Training Framework for Imbalanced Semi-Supervised Learning

Chen Wei[1*] Kihyuk Sohn[2] Clayton Mellina[2] Alan Yuille[1] Fan Yang[2]
[1]Johns Hopkins University [2]Google Cloud AI

Abstract

Semi-supervised learning on class-imbalanced data, although a realistic problem, has been under studied. While existing semi-supervised learning (SSL) methods are known to perform poorly on minority classes, we find that they still generate high precision pseudo-labels on minority classes. By exploiting this property, in this work, we propose Class-Rebalancing Self-Training (CReST), a simple yet effective framework to improve existing SSL methods on class-imbalanced data. CReST iteratively retrains a baseline SSL model with a labeled set expanded by adding pseudo-labeled samples from an unlabeled set, where pseudo-labeled samples from minority classes are selected more frequently according to an estimated class distribution. We also propose a progressive distribution alignment to adaptively adjust the rebalancing strength dubbed CReST+. We show that CReST and CReST+ improve state-of-the-art SSL algorithms on various class-imbalanced datasets and consistently outperform other popular rebalancing methods. Code has been made available at https://github.com/google-research/crest.

1. Introduction

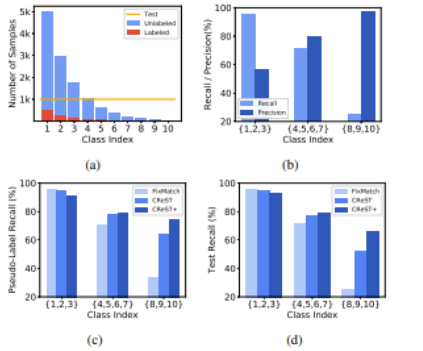


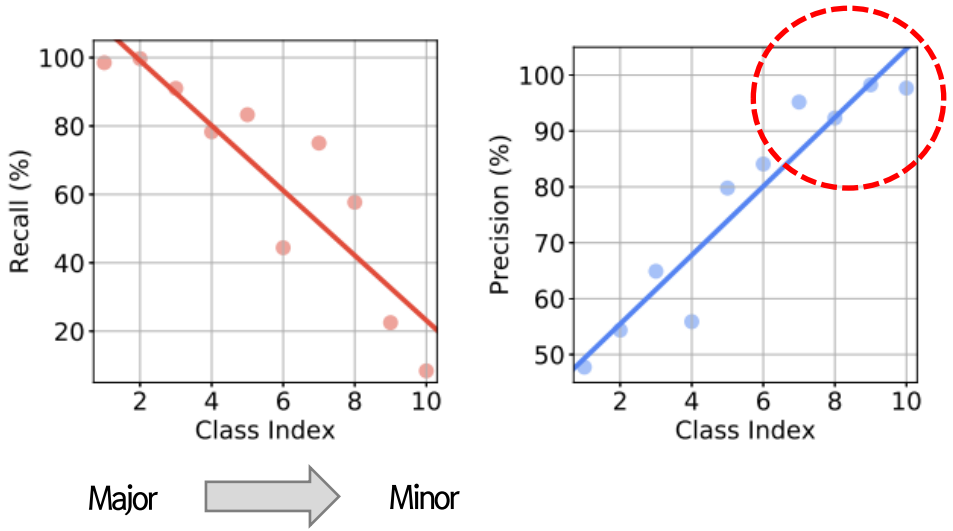
Figure 1. Experimental results on CIFAR10-LT. (a) Both labeled and unlabeled sets are class-imbalanced, where the most majority class has 100× more samples than the most minority class. The test set remains balanced. (b) Precision and recall of a FixMatch [39] model. Although minority classes have low recall, they obtain high precision. (c) & (d) The proposed CReST and CReST+ improve the quality of pseudo-labels (c) and thus the recall on the balanced test set (d), especially on minority classes.

Class-Imbalanced Semi-Supervised Learning

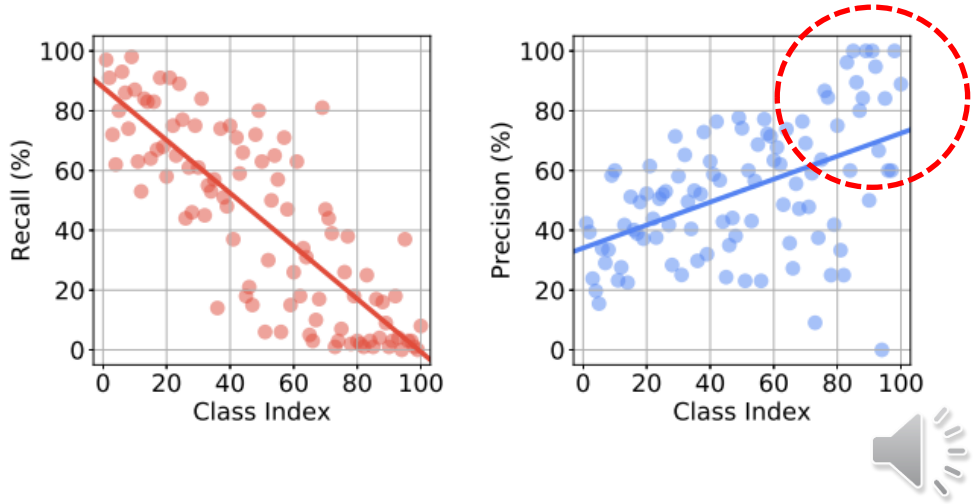
CReST: A Class-Rebalancing Self-Training Framework for Imbalanced Semi-Supervised Learning

❖ 방법론 연구 motivation

- ➢ Semi-supervised learning은 클래스 불균형 상황에서 다수에 편향된 학습을 하기 때문에 소수 클래스 분류 성능이 저조함
- ➢ 그러나 분류 성능 지표를 분석한 결과, 소수 클래스는 다수 클래스보다 높은 precision 값을 갖음



[Class-imbalanced CIFAR-10]



[Class-imbalanced CIFAR-100]

Class-Imbalanced Semi-Supervised Learning

CReST: A Class-Rebalancing Self-Training Framework for Imbalanced Semi-Supervised Learning

❖ 방법론 연구 motivation

- Semi-supervised learning은 클래스 불균형 상황에서 다수에 편향된 학습을 하기 때문에 소수 클래스 분류 성능이 저조함
- 그러나 분류 성능 지표를 분석한 결과, 소수 클래스는 다수 클래스보다 높은 precision 값을 갖음

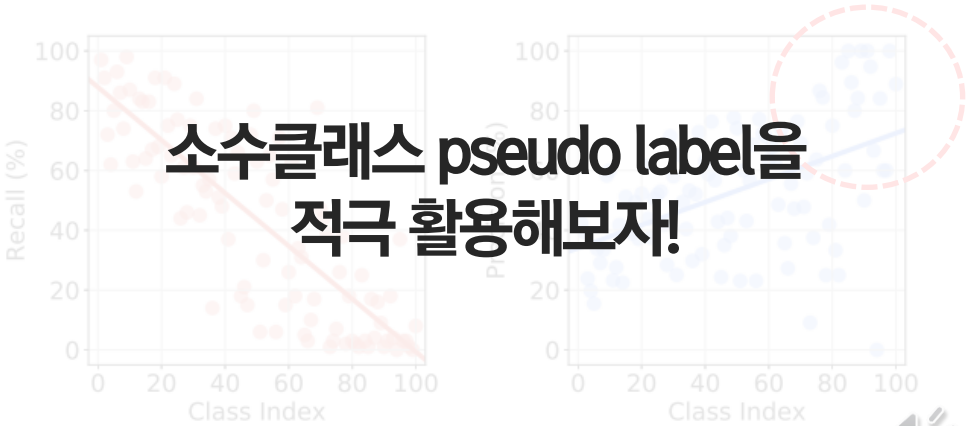
소수 클래스가 다수 클래스보다 상대적으로

low recall, high precision

소수 클래스로 분류를 잘 안하더라도
정확하게 분류함

[Class-imbalanced CIFAR-10]

소수클래스 pseudo label을
적극 활용해보자!



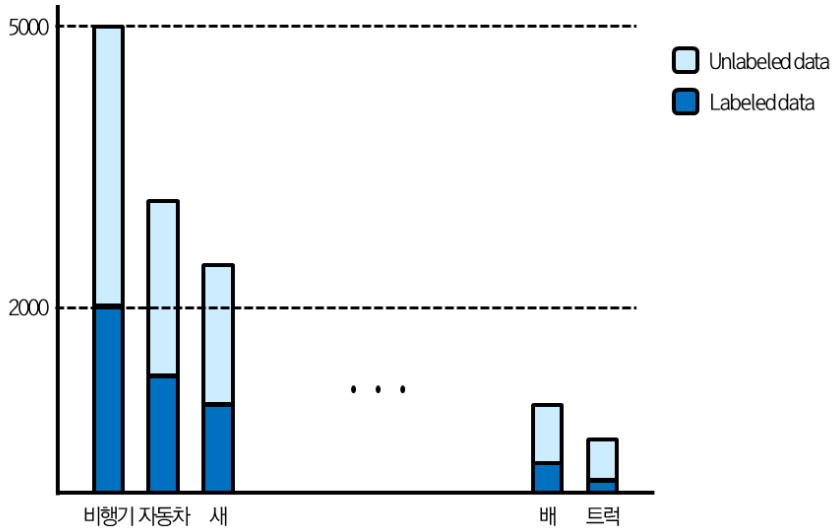
[Class-imbalanced CIFAR-100]

Class-Imbalanced Semi-Supervised Learning

CReST: A Class-Rebalancing Self-Training Framework for Imbalanced Semi-Supervised Learning

[Generation 1]

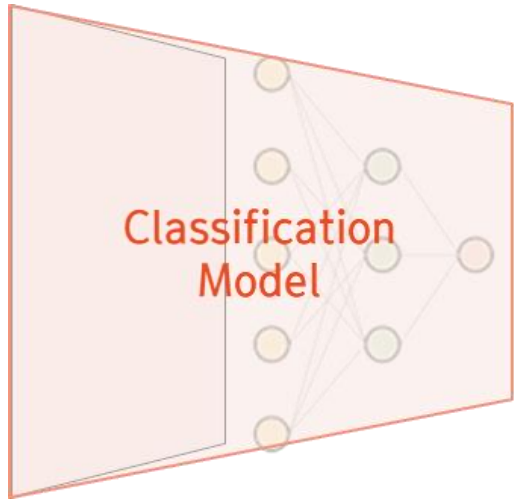
Step 1: semi-supervised learning algorithm으로 모델 학습



[Class-imbalanced CIFAR-10]

Labeled data와 unlabeled data
사용하여 모델 학습

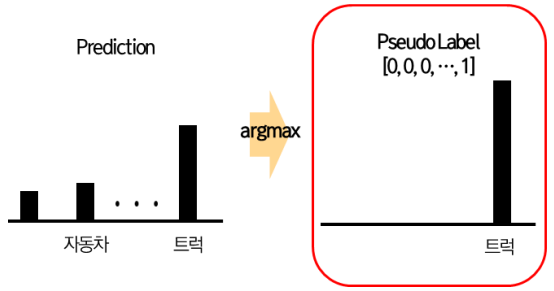
SSL algorithm으로 학습된 모델
(e.g. FixMatch)



Test dataset 평가에 사용

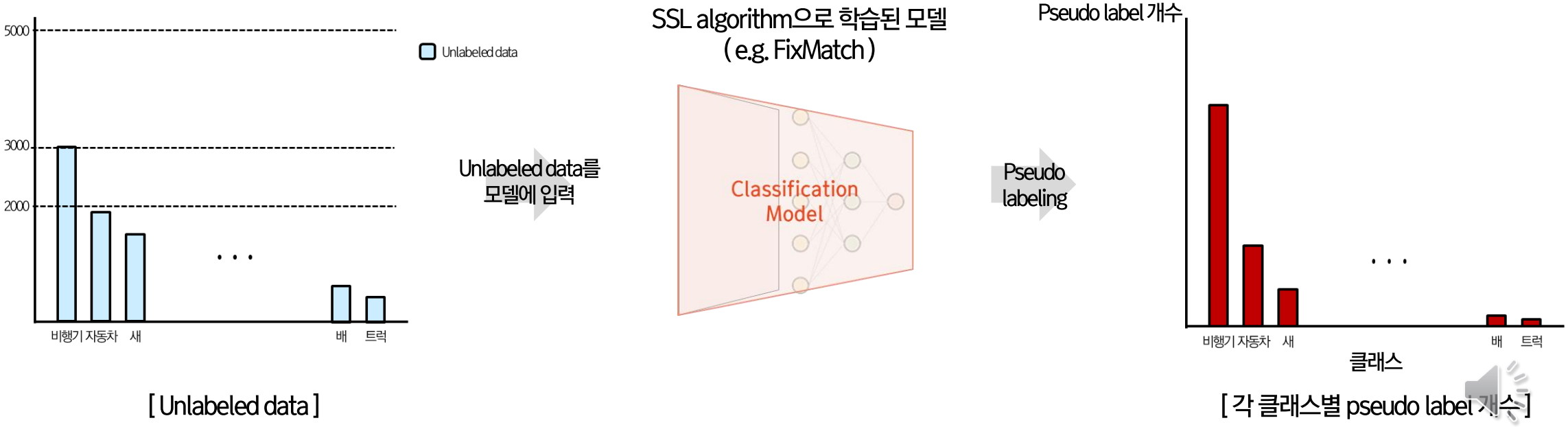
Class-Imbalanced Semi-Supervised Learning

CReST: A Class-Rebalancing Self-Training Framework for Imbalanced Semi-Supervised Learning



[Generation 1]

Step 2: 학습된 모델로 unlabeled data에 대해서 pseudo label 생성



[Unlabeled data]

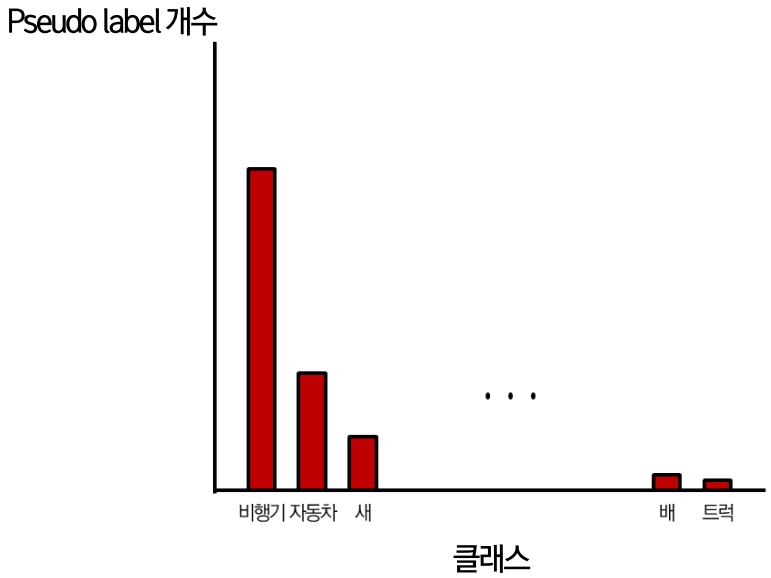
[각 클래스별 pseudo label 개수]

Class-Imbalanced Semi-Supervised Learning

CReST: A Class-Rebalancing Self-Training Framework for Imbalanced Semi-Supervised Learning

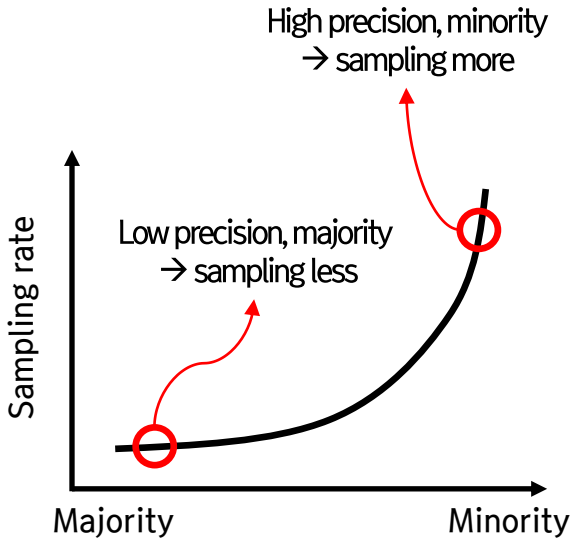
[Generation 1]

Step 3: 생성된 pseudo label을 클래스별로 샘플링해서 labeled data에 합침



[각 클래스별 pseudo label 개수]

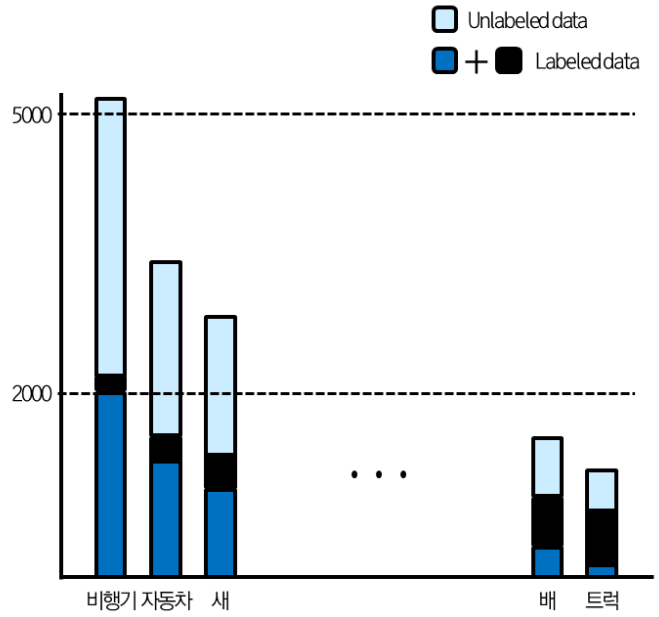
Sampling



Sampling rate : $\mu_l = (\frac{N_{L+1-l}}{N_1})^{\alpha}$,

N_1 = 첫번째 클래스의 labeled data 수

Expand labeled dataset



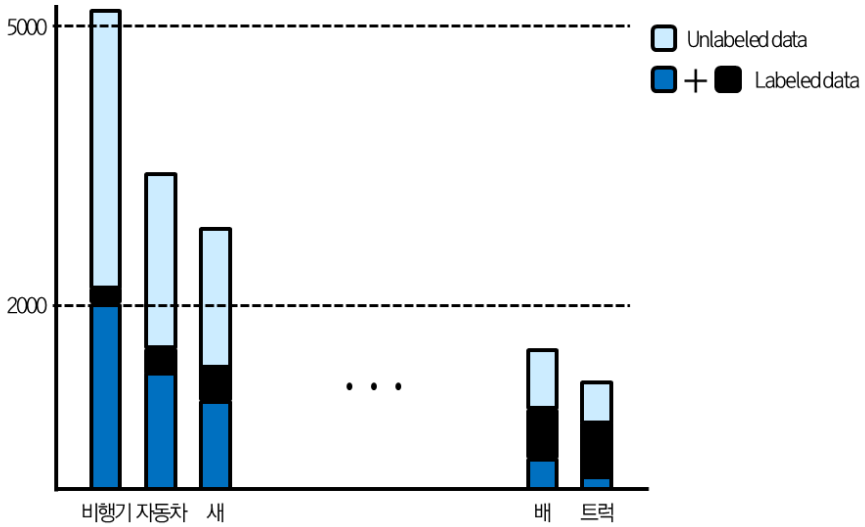
■: Sampling된 pseudo labeled data

Class-Imbalanced Semi-Supervised Learning

CReST: A Class-Rebalancing Self-Training Framework for Imbalanced Semi-Supervised Learning

[Generation 1] → [Generation 2]

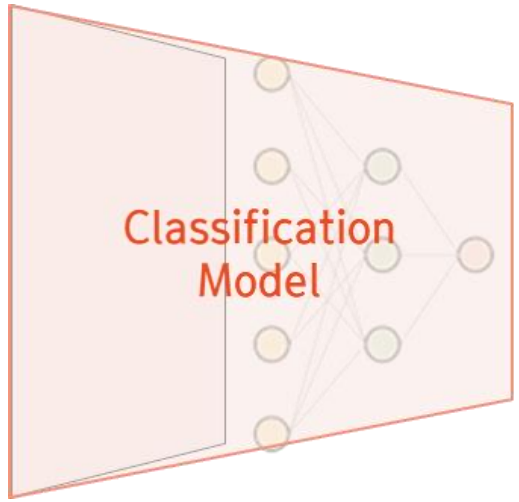
Step 1: semi-supervised learning algorithm으로 모델 학습



[Expanded class-imbalanced CIFAR-10]

Labeled data와 unlabeled data
사용하여 모델 학습

SSL algorithm으로 학습된 모델
(e.g. FixMatch)



Test dataset 평가에 사용

Class-Imbalanced Semi-Supervised Learning

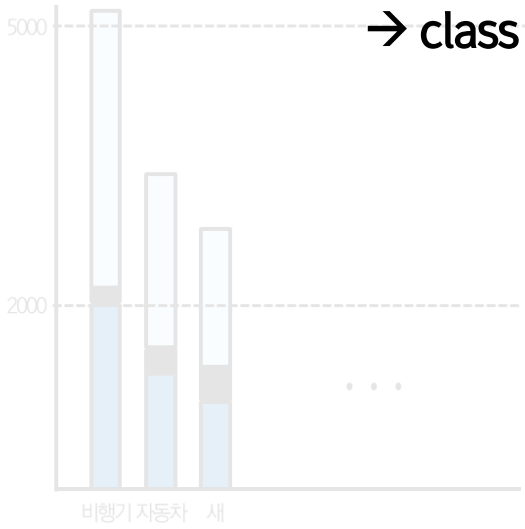
CReST: A Class-Rebalancing Self-Training Framework for Imbalanced Semi-Supervised Learning

[Generation 1] → **[Generation 2]** → · · ·

Step 1: semi-supervised learning algorithm으로 모델 학습

Generation을 거듭할수록 불균형 데이터셋이 균형 데이터셋으로 점차 바뀜

→ class re-balancing 효과를 얻으면서 불균형 데이터셋 문제를 해소함



[Expanded class-Imbalanced CIF

SSL algorithm으로 학습된 모델
(e.g. FixMatch)

Classification Model

Test dataset 평가에 사용

Method	Gen_1	Gen_2	Gen_3
Supervised (100% labels)	75.8	-	-
Supervised (10% labels)	46.0	-	-
FixMatch (10% labels)	65.8	-	-
w/ DA ($t = 0.5$)	69.1	-	-
w/ CReST	65.8	67.6	67.7
w/ CReST+	68.3	70.7	**73.7**

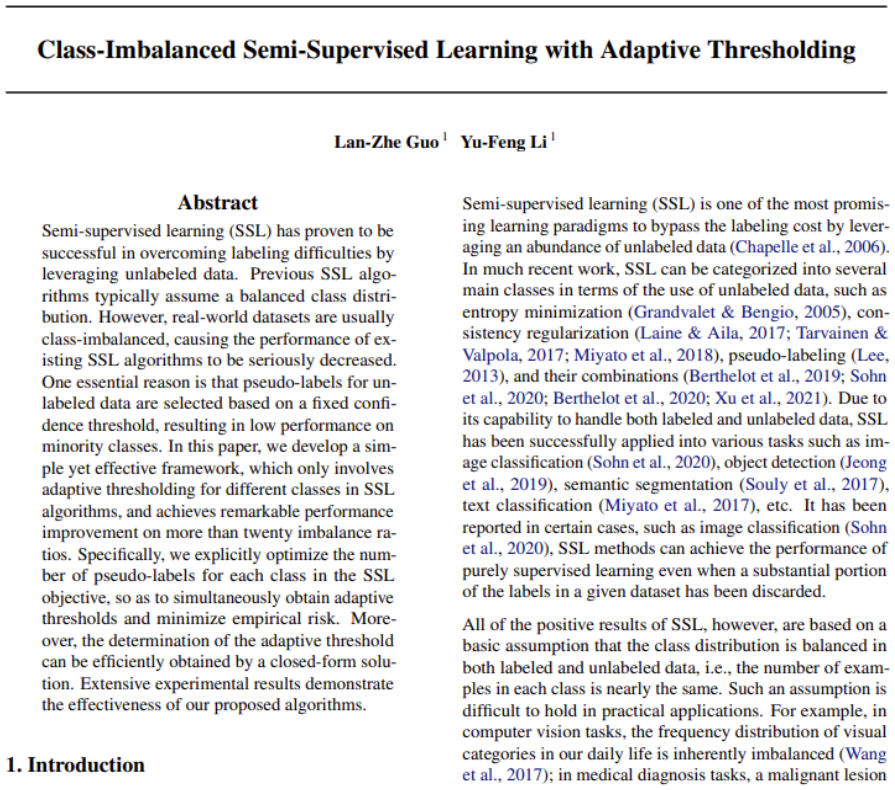
Table 4. Evaluating the proposed method on ImageNet127 with $\beta = 10\%$ samples are labeled. We retrain FixMatch models for 3 generations with our CReST and CReST+.

Class-Imbalanced Semi-Supervised Learning

Class-Imbalanced Semi-Supervised Learning with Adaptive Thresholding

❖ Adsh: Class-Imbalanced Semi-Supervised Learning with Adaptive Thresholding (ICML, 2022)

- 2023년 3월 28일 기준 12회 인용됨
- Adsh = FixMatch + class adaptive thresholding(클래스별로 불균형 상황을 반영하여 threshold를 유연하게 조정함)



Class-Imbalanced Semi-Supervised Learning with Adaptive Thresholding

Lan-Zhe Guo [1] **Yu-Feng Li** [1]

Abstract

Semi-supervised learning (SSL) has proven to be successful in overcoming labeling difficulties by leveraging unlabeled data. Previous SSL algorithms typically assume a balanced class distribution. However, real-world datasets are usually class-imbalanced, causing the performance of existing SSL algorithms to be seriously decreased. One essential reason is that pseudo-labels for unlabeled data are selected based on a fixed confidence threshold, resulting in low performance on minority classes. In this paper, we develop a simple yet effective framework, which only involves adaptive thresholding for different classes in SSL algorithms, and achieves remarkable performance improvement on more than twenty imbalance ratios. Specifically, we explicitly optimize the number of pseudo-labels for each class in the SSL objective, so as to simultaneously obtain adaptive thresholds and minimize empirical risk. Moreover, the determination of the adaptive threshold can be efficiently obtained by a closed-form solution. Extensive experimental results demonstrate the effectiveness of our proposed algorithms.

1. Introduction

Semi-supervised learning (SSL) is one of the most promising learning paradigms to bypass the labeling cost by leveraging an abundance of unlabeled data (Chapelle et al., 2006). In much recent work, SSL can be categorized into several main classes in terms of the use of unlabeled data, such as entropy minimization (Grandvalet & Bengio, 2005), consistency regularization (Laine & Aila, 2017; Tarvainen & Valpola, 2017; Miyato et al., 2018), pseudo-labeling (Lee, 2013), and their combinations (Berthelot et al., 2019; Sohn et al., 2020; Berthelot et al., 2020; Xu et al., 2021). Due to its capability to handle both labeled and unlabeled data, SSL has been successfully applied into various tasks such as image classification (Sohn et al., 2020), object detection (Jeong et al., 2019), semantic segmentation (Souly et al., 2017), text classification (Miyato et al., 2017), etc. It has been reported in certain cases, such as image classification (Sohn et al., 2020), SSL methods can achieve the performance of purely supervised learning even when a substantial portion of the labels in a given dataset has been discarded.

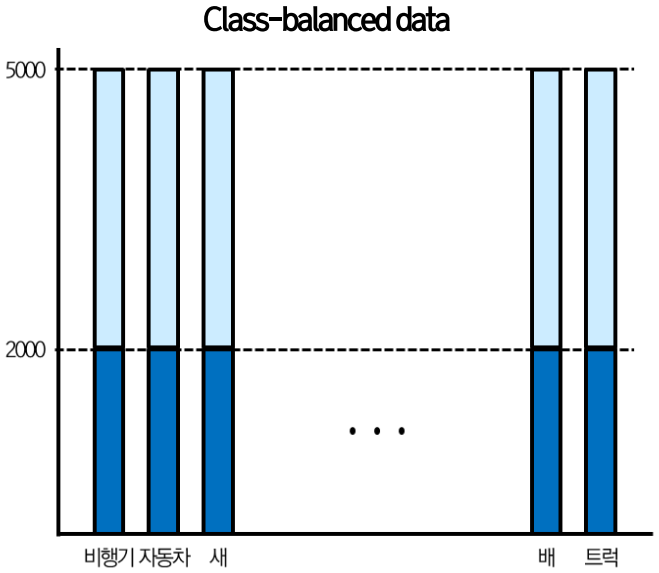
All of the positive results of SSL, however, are based on a basic assumption that the class distribution is balanced in both labeled and unlabeled data, i.e., the number of examples in each class is nearly the same. Such an assumption is difficult to hold in practical applications. For example, in computer vision tasks, the frequency distribution of visual categories in our daily life is inherently imbalanced (Wang et al., 2017); in medical diagnosis tasks, a malignant lesion

Class-Imbalanced Semi-Supervised Learning

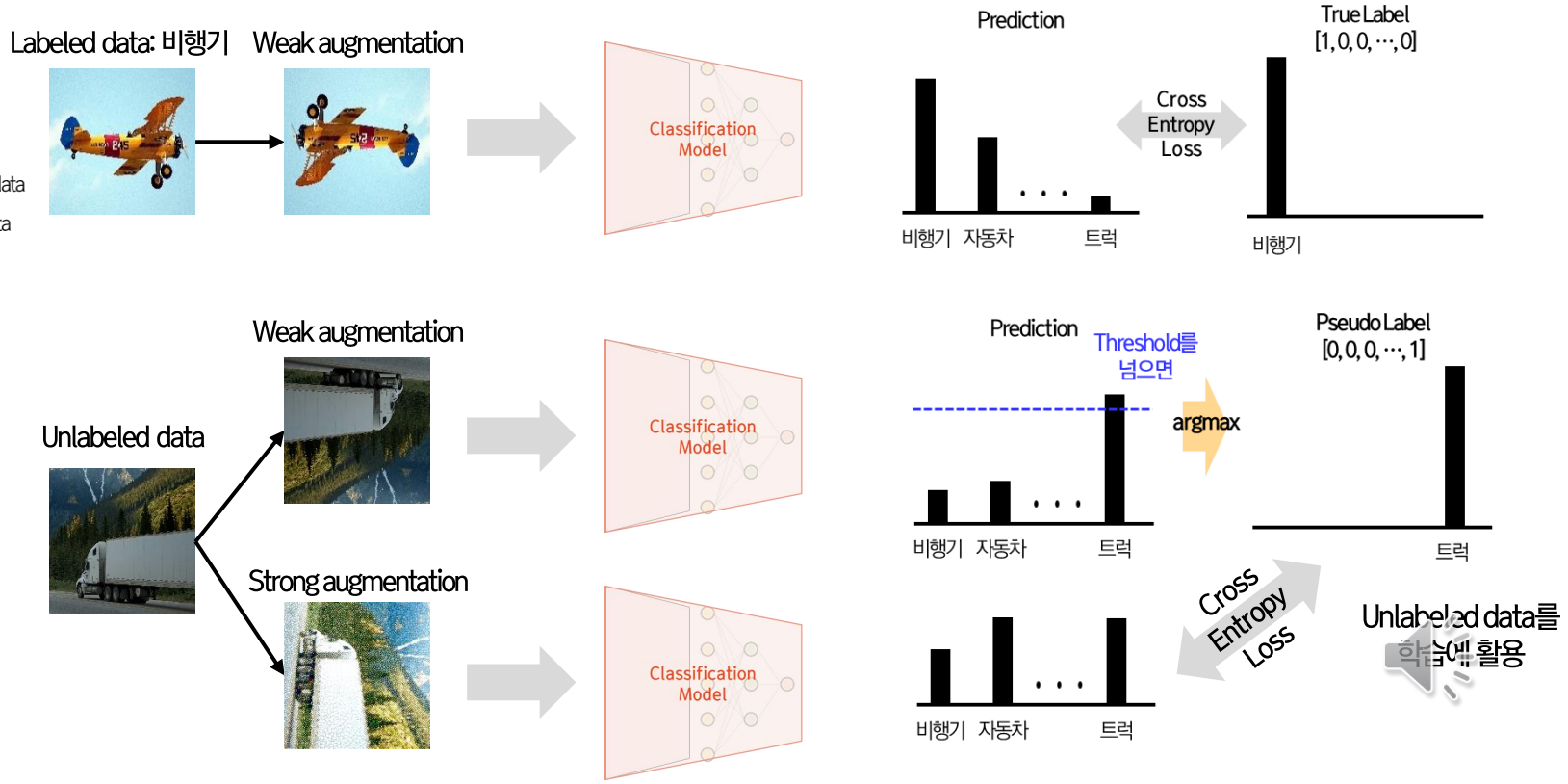
Class-Imbalanced Semi-Supervised Learning with Adaptive Thresholding

❖ Adsh: Class-Imbalanced Semi-Supervised Learning with Adaptive Thresholding (ICML, 2022)

- 2023년 3월 28일 기준 12회 인용됨
- Adsh = FixMatch + class adaptive thresholding(클래스별로 불균형 상황을 반영하여 threshold를 유연하게 조정함)



[SSL 연구에 활용되는 CIFAR-10]

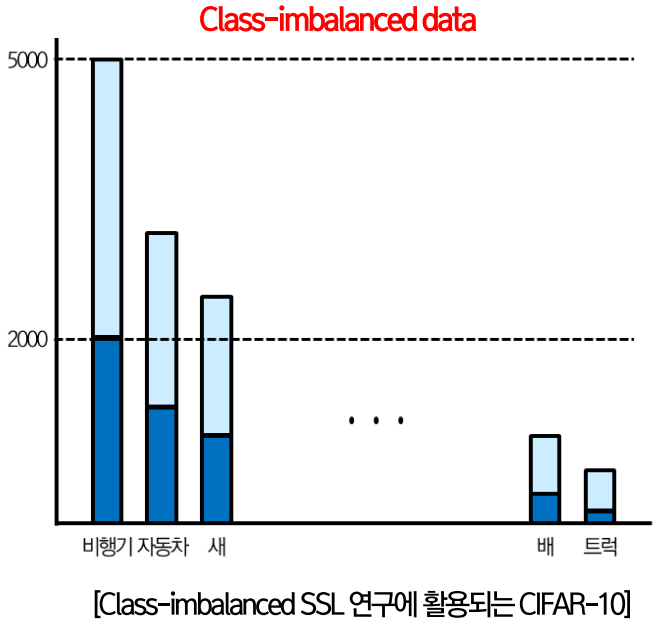


Class-Imbalanced Semi-Supervised Learning

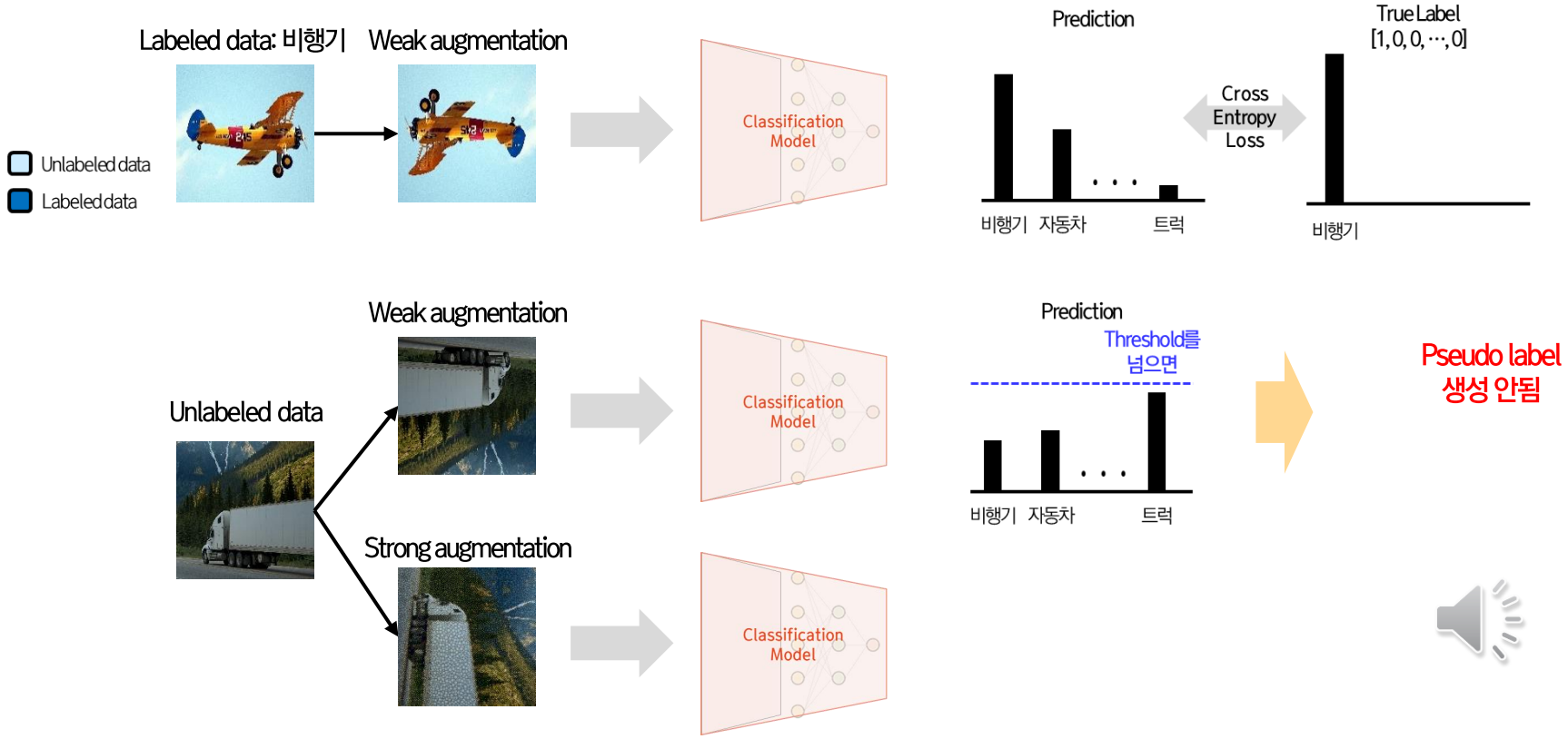
Class-Imbalanced Semi-Supervised Learning with Adaptive Thresholding

❖ Adsh: Class-Imbalanced Semi-Supervised Learning with Adaptive Thresholding (ICML, 2022)

- 2023년 3월 28일 기준 12회 인용됨
- Adsh = FixMatch + class adaptive thresholding(클래스별로 불균형 상황을 반영하여 threshold를 유연하게 조정함)



[Class-imbalanced SSL 연구에 활용되는 CIFAR-10]

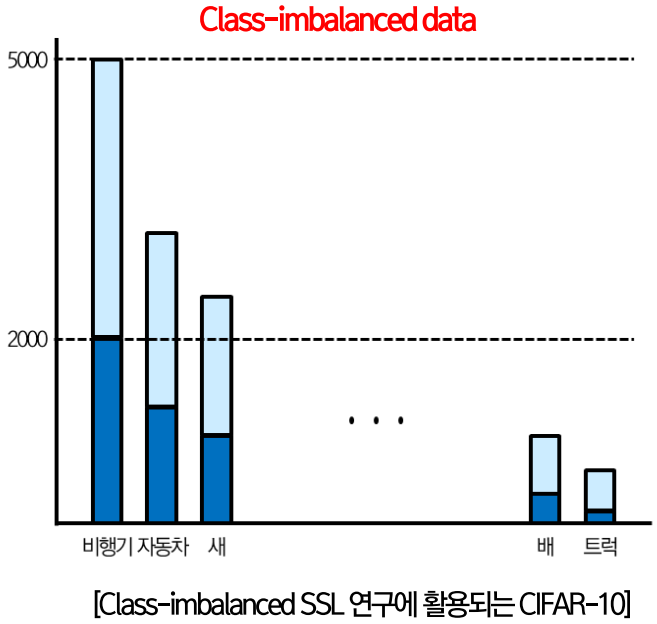


Class-Imbalanced Semi-Supervised Learning

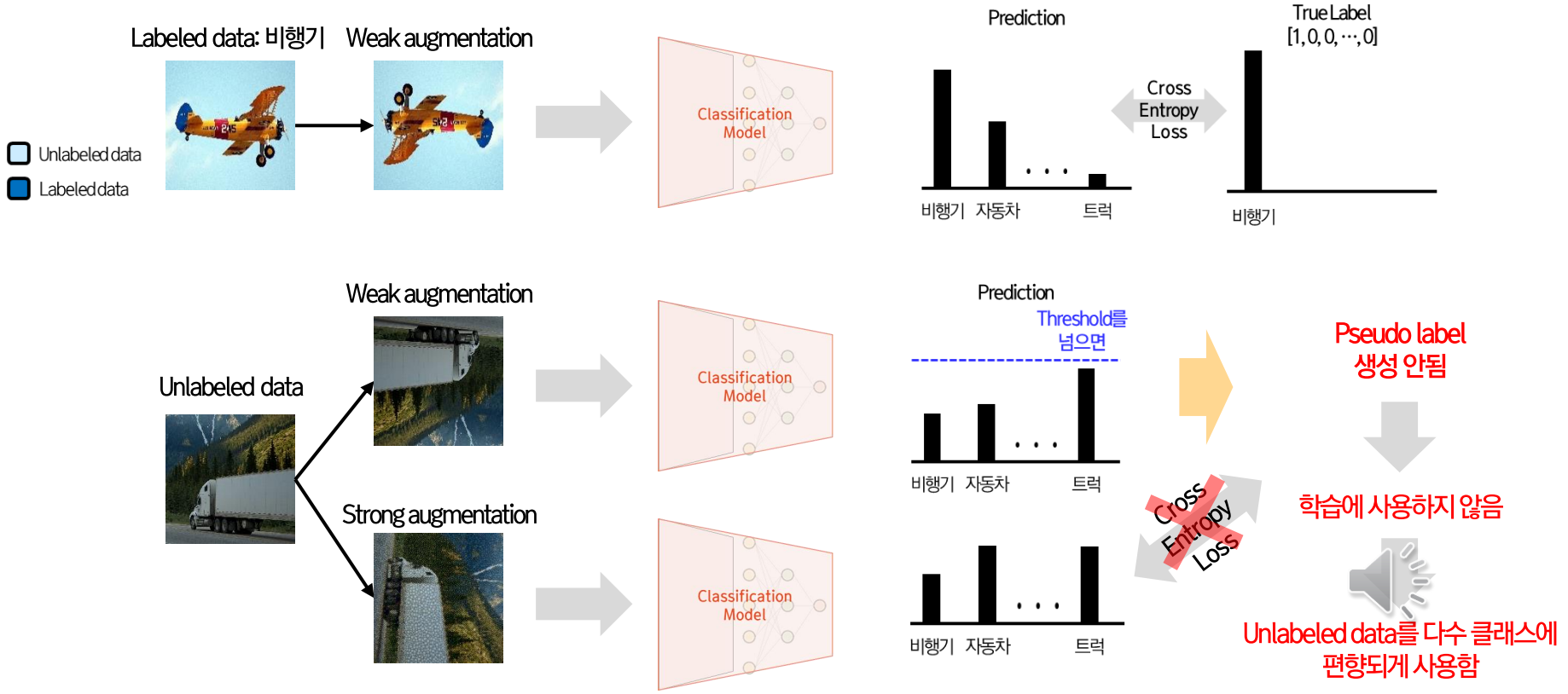
Class-Imbalanced Semi-Supervised Learning with Adaptive Thresholding

❖ Adsh: Class-Imbalanced Semi-Supervised Learning with Adaptive Thresholding (ICML, 2022)

- 2023년 3월 28일 기준 12회 인용됨
- Adsh = FixMatch + class adaptive thresholding(클래스별로 불균형 상황을 반영하여 threshold를 유연하게 조정함)



[Class-imbalanced SSL 연구에 활용되는 CIFAR-10]

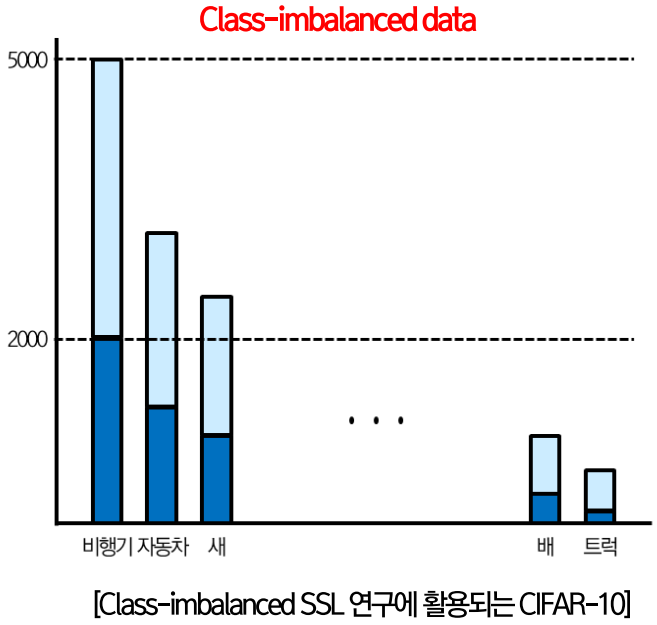


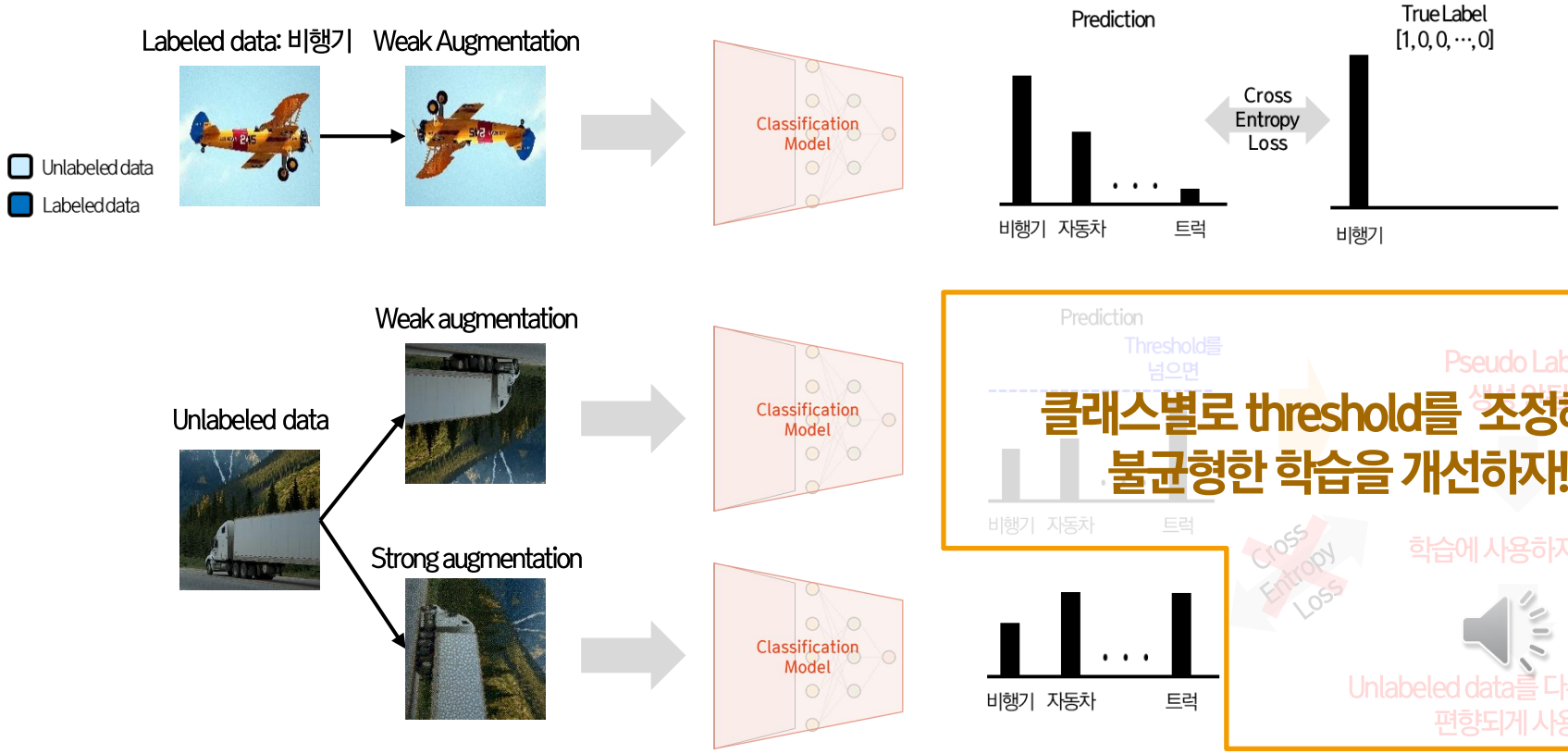
Class-Imbalanced Semi-Supervised Learning

Class-Imbalanced Semi-Supervised Learning with Adaptive Thresholding

❖ Adsh: Class-Imbalanced Semi-Supervised Learning with Adaptive Thresholding (ICML, 2022)

- 2023년 3월 28일 기준 12회 인용됨
- Adsh = FixMatch + class adaptive thresholding(클래스별로 불균형 상황을 반영하여 threshold를 유연하게 조정함)

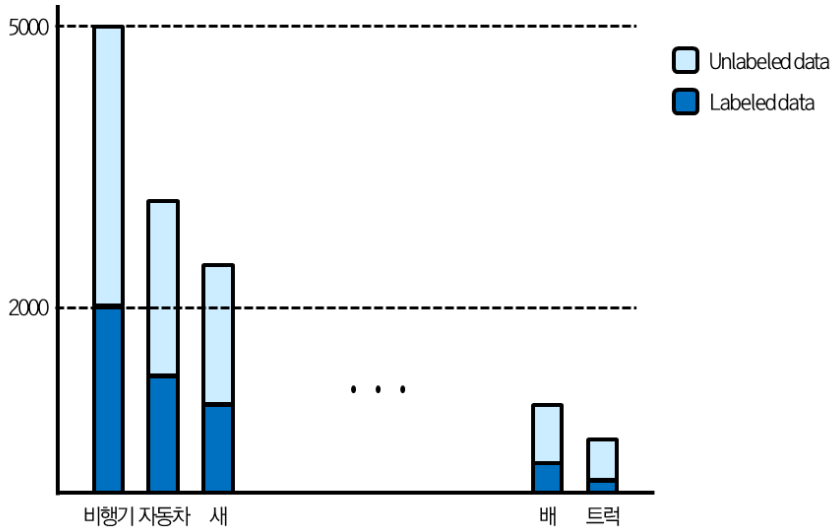




Class-Imbalanced Semi-Supervised Learning

Class-Imbalanced Semi-Supervised Learning with Adaptive Thresholding

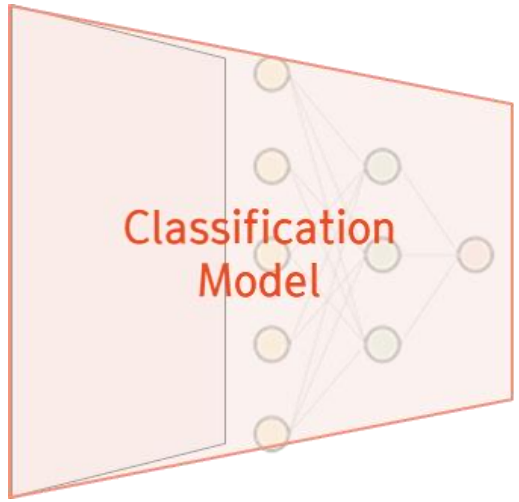
[Step 1]: 1 epoch만큼 학습 진행 (하이퍼 파라미터로 설정된 threshold (e.g. 0.95) 값으로 학습 시작)



[Class-imbalanced CIFAR-10]

Labeled data와 unlabeled data
사용하여 모델 학습

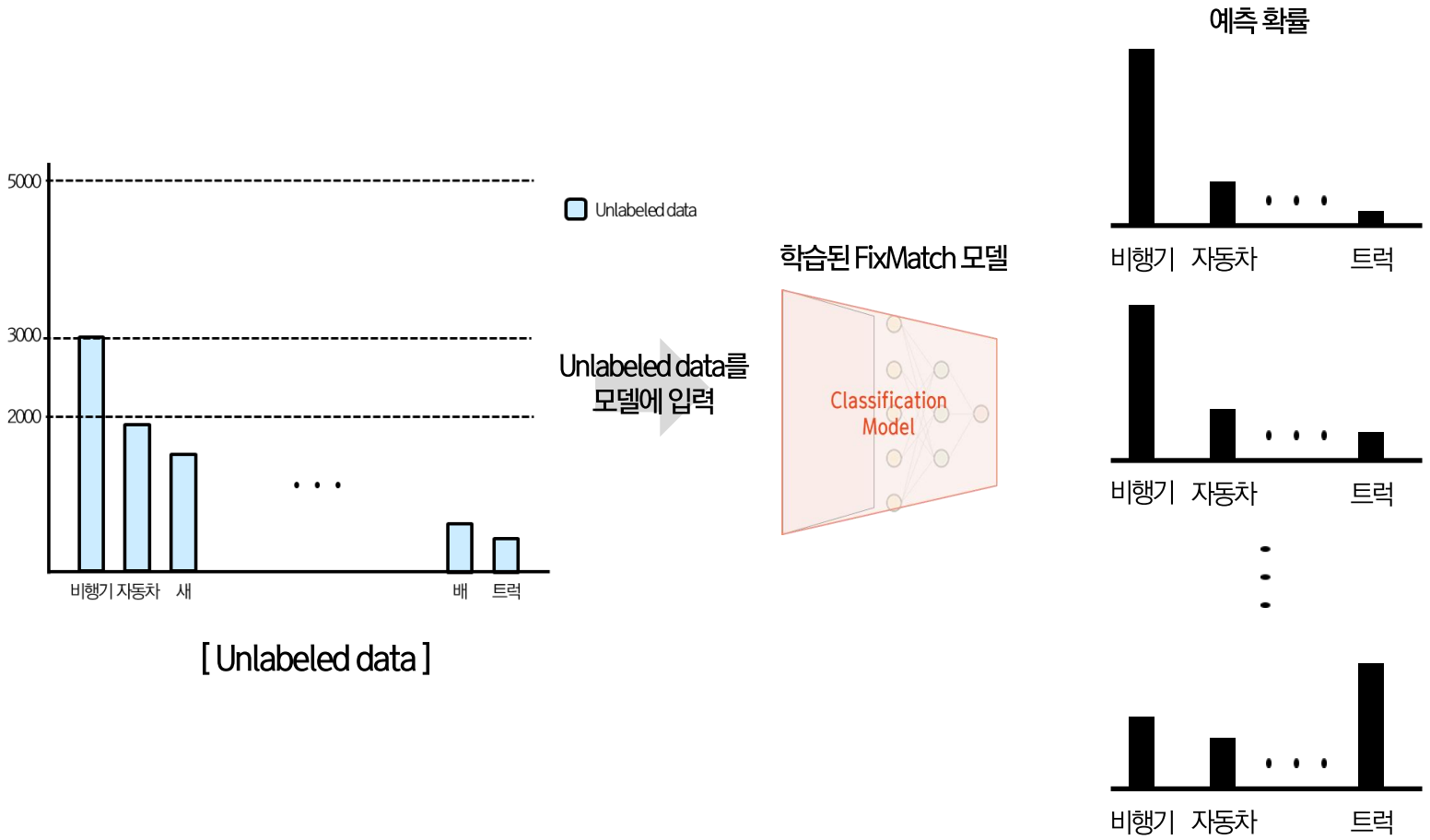
학습된 FixMatch 모델



Class-Imbalanced Semi-Supervised Learning

Class-Imbalanced Semi-Supervised Learning with Adaptive Thresholding

[Step 2]: 학습된 FixMatch 모델로 unlabeled data 최대 예측 확률 수합 및 정렬

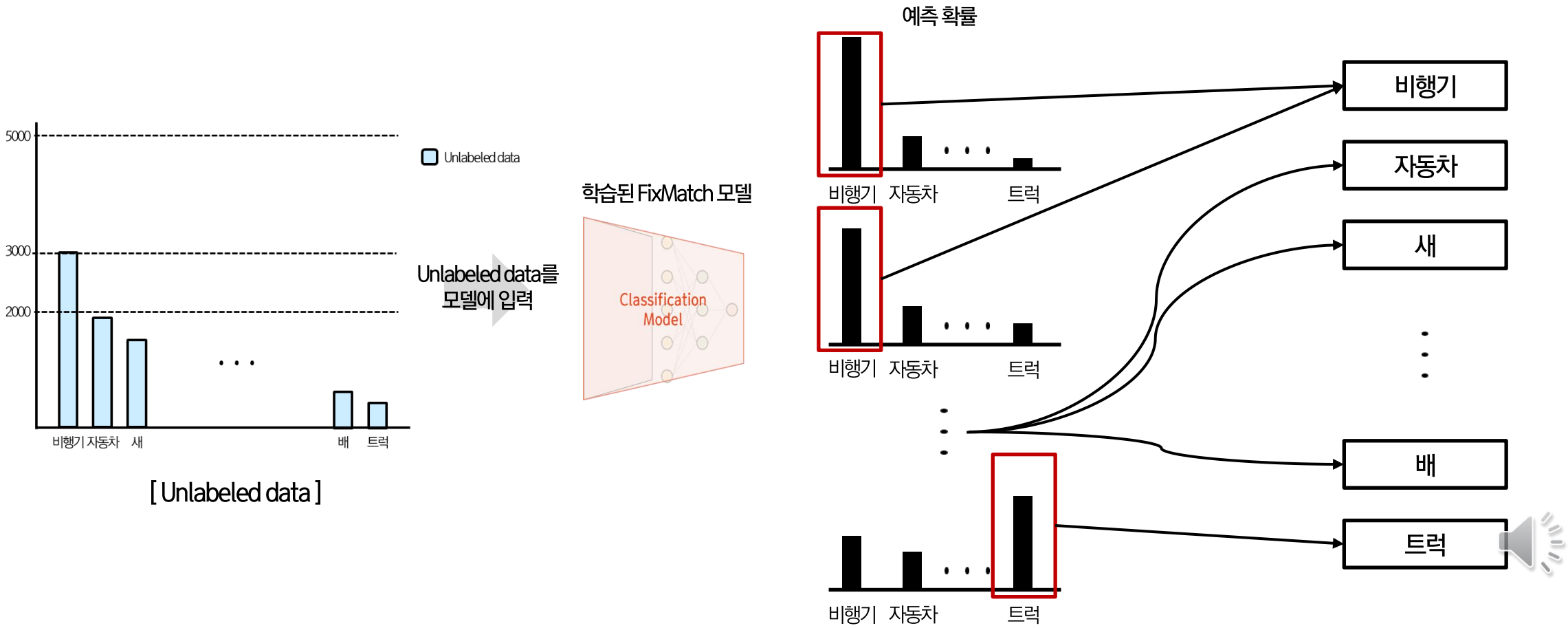


[Unlabeled data]

Class-Imbalanced Semi-Supervised Learning

Class-Imbalanced Semi-Supervised Learning with Adaptive Thresholding

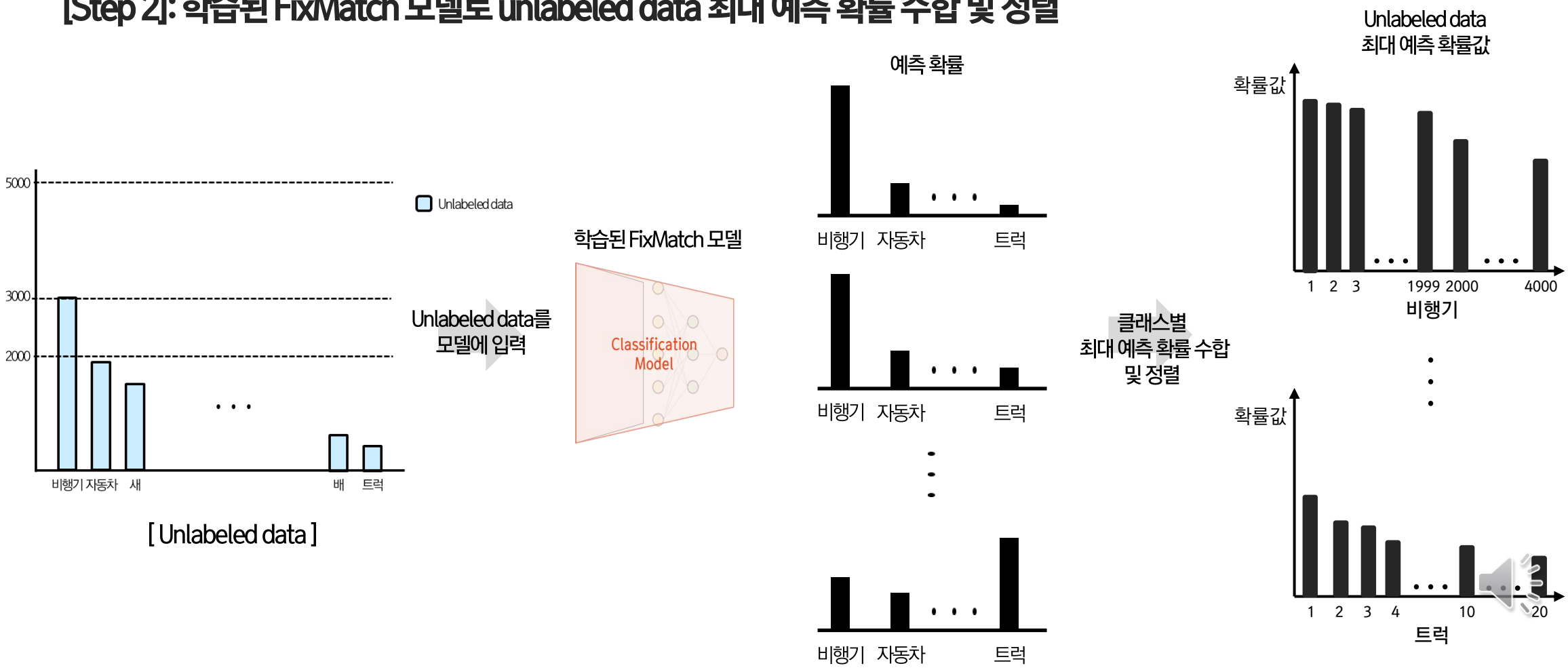
[Step 2]: 학습된 FixMatch 모델로 unlabeled data 최대 예측 확률 수합 및 정렬



Class-Imbalanced Semi-Supervised Learning

Class-Imbalanced Semi-Supervised Learning with Adaptive Thresholding

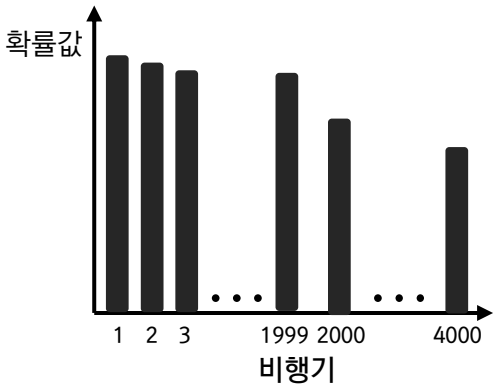
[Step 2]: 학습된 FixMatch 모델로 unlabeled data 최대 예측 확률 수합 및 정렬



Class-Imbalanced Semi-Supervised Learning

Class-Imbalanced Semi-Supervised Learning with Adaptive Thresholding

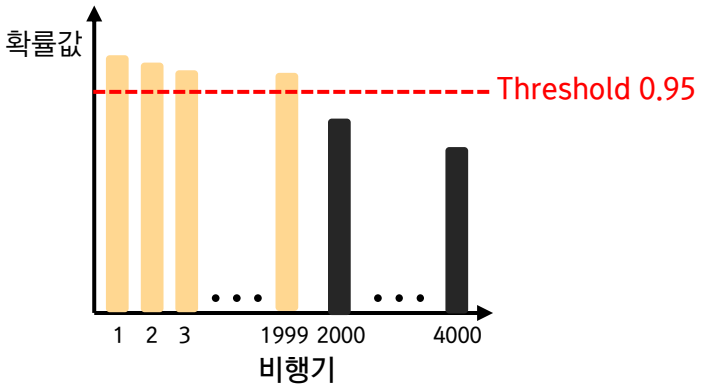
[Step 3]: Labeled data에서 가장 다수인 클래스 기준으로 나머지 클래스 threshold 결정



Class-Imbalanced Semi-Supervised Learning

Class-Imbalanced Semi-Supervised Learning with Adaptive Thresholding

[Step 3]: Labeled data에서 가장 다수인 클래스 기준으로 나머지 클래스 threshold 결정

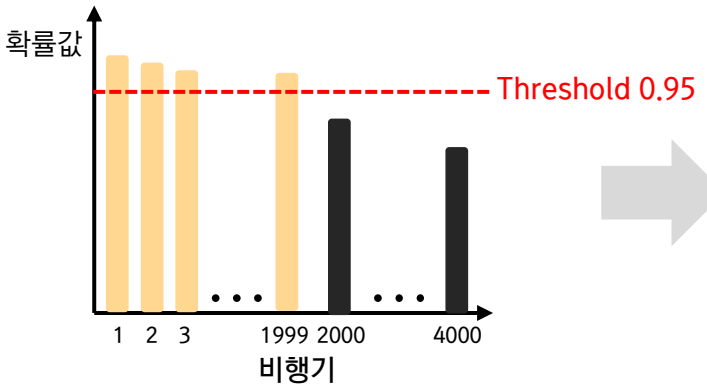


기준이 되는 다수 클래스는 항상
threshold 0.95 고정

Class-Imbalanced Semi-Supervised Learning

Class-Imbalanced Semi-Supervised Learning with Adaptive Thresholding

[Step 3]: Labeled data에서 가장 다수인 클래스 기준으로 나머지 클래스 threshold 결정



$$\rho = \frac{2000}{4000} \times 100\% = 50\%$$

Threshold 0.95를 넘지 못한 index의 백분율

기준이 되는 다수 클래스는 항상
threshold 0.95 고정

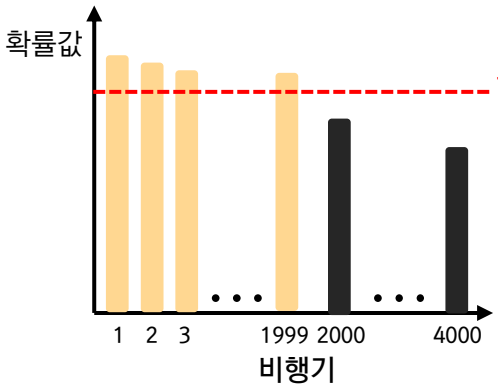
Class-Imbalanced Semi-Supervised Learning

Class-Imbalanced Semi-Supervised Learning with Adaptive Thresholding

[Step 3]: Labeled data에서 가장 다수인 클래스 기준으로 나머지 클래스 threshold 결정

다수 클래스 threshold 0.95보다 큰 경우
해당 클래스 threshold = 0.95

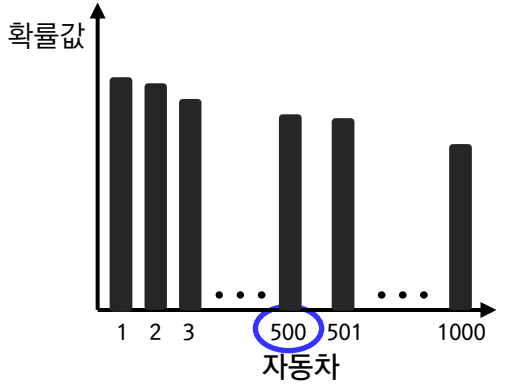
$$\tau = e^{-(-\log(p_{i-th\ index}))}$$

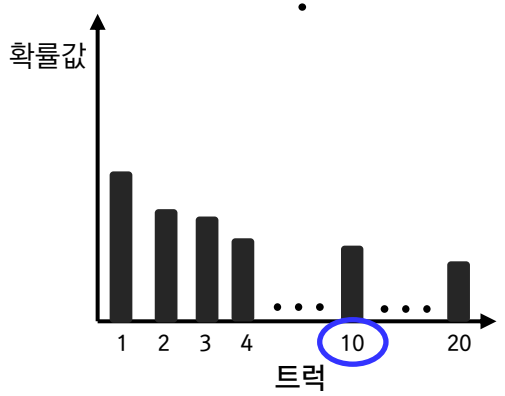


기준이 되는 다수 클래스는 항상
threshold 0.95 고정

$$\rho = \frac{2000}{4000} \times 100\% = 50\%$$

Threshold 0.95를 넘지 못한 index의 백분율

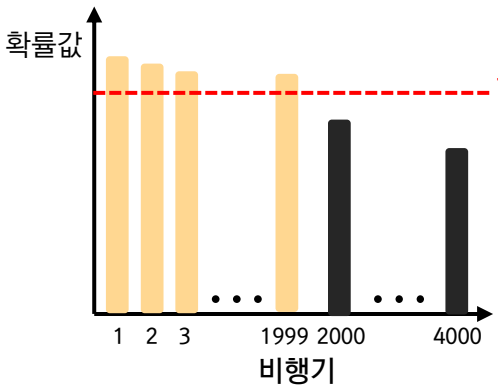




Class-Imbalanced Semi-Supervised Learning

Class-Imbalanced Semi-Supervised Learning with Adaptive Thresholding

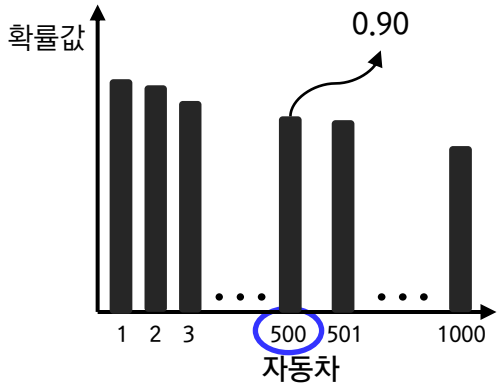
[Step 3]: Labeled data에서 가장 다수인 클래스 기준으로 나머지 클래스 threshold 결정

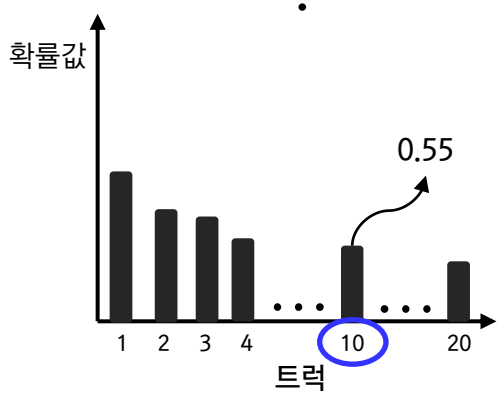


기준이 되는 다수 클래스는 항상
threshold 0.95 고정

$$\rho = \frac{2000}{4000} \times 100\% = 50\%$$

Threshold 0.95를 넘지 못한 index의 백분율





다수 클래스 threshold 0.95보다 큰 경우
해당 클래스 threshold = 0.95

$$\tau = e^{-(-\log(p_{i-th\ index}))}$$

$$e^{-(-\log(0.90))} \approx 0.955$$

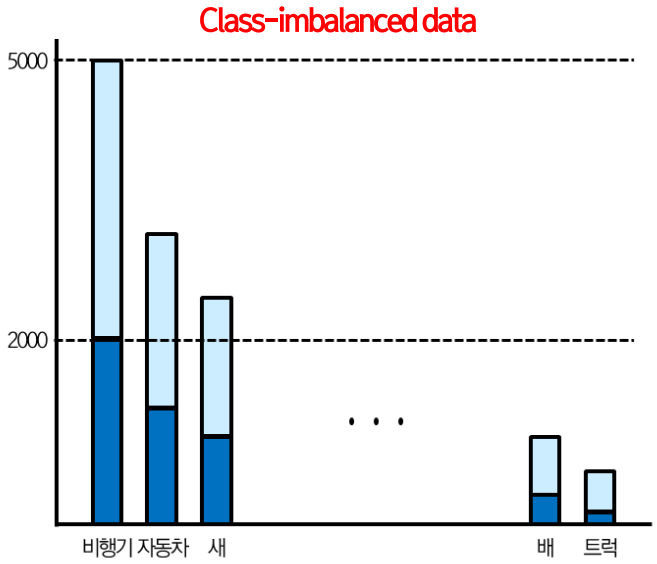
자동차 클래스 threshold = 0.95

$$e^{-(-\log(0.55))} \approx 0.771$$

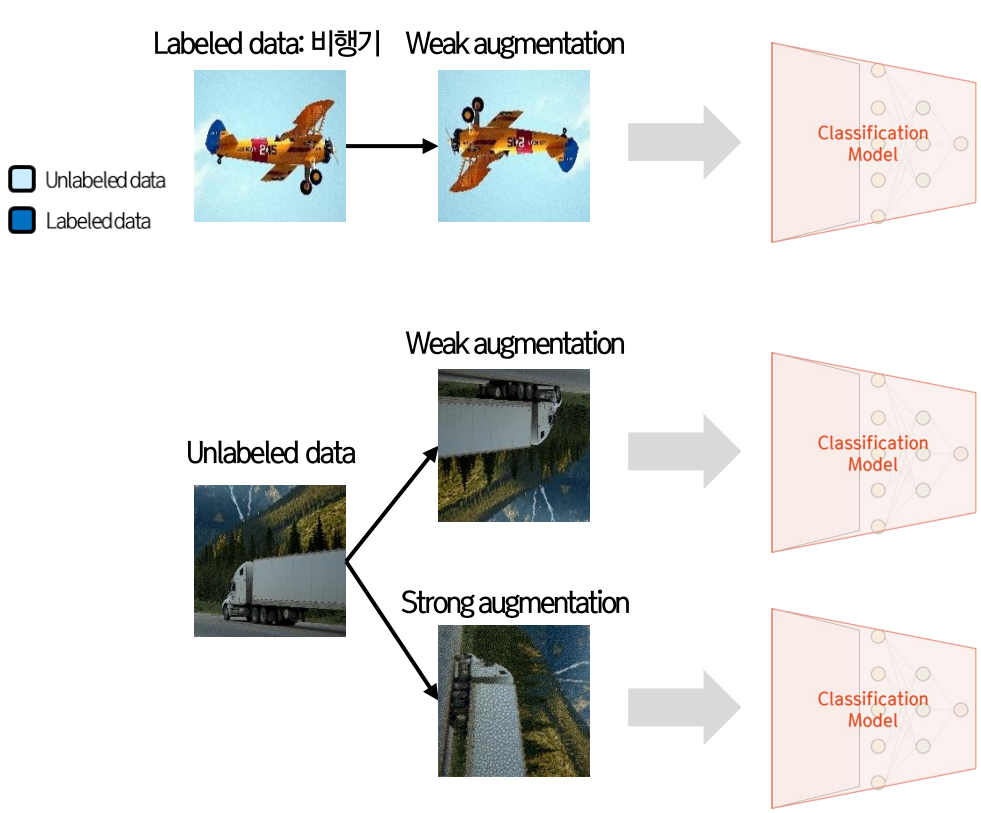
트럭 클래스 threshold = 0.771

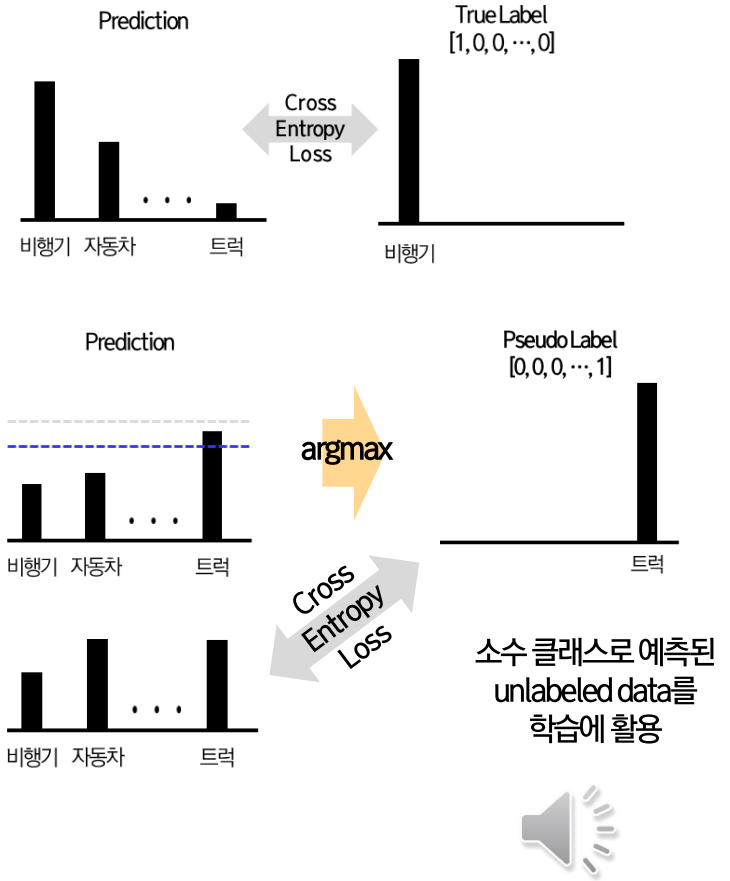
Class-Imbalanced Semi-Supervised Learning

Class-Imbalanced Semi-Supervised Learning with Adaptive Thresholding



[Class-imbalanced SSL 연구에 활용되는 CIFAR-10]





Class-Imbalanced Semi-Supervised Learning

Class-Imbalanced Semi-Supervised Learning with Adaptive Thresholding

❖ Adsh 실험 결과: 불균형 데이터에 대해서 다양한 방법론과 비교(Imbalanced CIFAR-10)

- Class-imbalanced learning, semi-supervised learning 및 imbalanced semi-supervised learning 방법론과 비교 진행
- 불균형 데이터에 대해서 supervised learning (WideResNet-28-2)을 사용하는 것 보다 불균형이 반영된 방법론들이 더 좋은 성능을 보임
- 불균형이 반영된 방법론보다 unlabeled data를 단순 활용하는 semi-supervised learning이 더 좋은 성능을 보임

Table 1. Comparison of classification performance (Accuracy (%)) on imbalanced CIFAR-10 dataset under three different imbalance ratio: $\gamma = 50, 100, 150$ and two different numbers of labeled data: $N_1 = 1500, M_1 = 3000$ and $N_1 = 500, M_1 = 4000$. The best results are indicated in bold.

	Imbalanced CIFAR-10 Dataset					
	$N_1 = 1500, M_1 = 3000$			$N_1 = 500, M_1 = 4000$		
Algorithm	$\gamma = 50$	$\gamma = 100$	$\gamma = 150$	$\gamma = 50$	$\gamma = 100$	$\gamma = 150$
Supervised	65.23 ± 0.05	58.94 ± 0.13	55.63 ± 0.38	51.31 ± 0.34	45.82 ± 0.41	40.90 ± 0.39
CBL	65.52 ± 0.31	58.52 ± 0.45	52.36 ± 0.58	51.94 ± 0.71	46.22 ± 0.92	41.58 ± 1.24
Re-Sampling	64.53 ± 0.39	56.34 ± 0.42	53.21 ± 0.51	51.96 ± 0.65	48.13 ± 1.25	40.26 ± 1.88
cRT	67.82 ± 0.14	63.43 ± 0.45	59.56 ± 0.44	56.28 ± 1.45	48.11 ± 0.79	45.02 ± 1.08
LDAM	68.91 ± 0.10	63.15 ± 0.24	58.68 ± 0.30	56.41 ± 0.92	49.27 ± 0.88	45.10 ± 0.75
Mean-Teacher	68.84 ± 0.82	61.33 ± 0.28	54.79 ± 0.31	56.34 ± 1.68	48.55 ± 0.77	45.32 ± 1.20
MixMatch	73.59 ± 0.46	65.03 ± 0.26	62.71 ± 0.29	65.32 ± 1.20	56.41 ± 1.96	52.38 ± 1.88
ReMixMatch	78.96 ± 0.29	72.88 ± 0.12	68.61 ± 0.40	76.83 ± 0.98	70.12 ± 1.23	59.58 ± 1.30
FixMatch	79.10 ± 0.14	71.50 ± 0.31	68.47 ± 0.15	77.34 ± 0.96	68.45 ± 0.94	60.10 ± 0.82
DARP	81.60 ± 0.31	75.23 ± 0.14	69.31 ± 0.26	76.72 ± 0.46	69.41 ± 0.50	61.23 ± 0.31
CReST	82.03 ± 0.26	75.08 ± 0.41	69.84 ± 0.39	76.18 ± 0.36	69.50 ± 0.70	60.81 ± 0.55
Adsh	$\mathbf{83.38 \pm 0.06}$	$\mathbf{76.52 \pm 0.35}$	$\mathbf{71.49 \pm 0.30}$	$\mathbf{79.27 \pm 0.38}$	$\mathbf{70.97 \pm 0.46}$	$\mathbf{62.04 \pm 0.51}$

Class-Imbalanced Semi-Supervised Learning

Class-Imbalanced Semi-Supervised Learning with Adaptive Thresholding

❖ Adsh 실험 결과: 불균형 데이터에 대해서 다양한 방법론과 비교(Imbalanced CIFAR-10)

- Class-imbalanced learning, semi-supervised learning 및 imbalanced semi-supervised learning 방법론과 비교 진행
- 불균형 데이터에 대해서 supervised learning (WideResNet-28-2)을 사용하는 것 보다 불균형이 반영된 방법론들이 더 좋은 성능을 보임
- 불균형이 반영된 방법론보다 unlabeled data를 단순 활용하는 semi-supervised learning이 더 좋은 성능을 보임

Table 1. Comparison of classification performance (Accuracy (%)) on imbalanced CIFAR-10 dataset under three different imbalance ratio: $\gamma = 50, 100, 150$ and two different numbers of labeled data: $N_1 = 1500, M_1 = 3000$ and $N_1 = 500, M_1 = 4000$. The best results are indicated in bold.

	Imbalanced CIFAR-10 Dataset					
	$N_1 = 1500, M_1 = 3000$			$N_1 = 500, M_1 = 4000$		
Algorithm	$\gamma = 50$	$\gamma = 100$	$\gamma = 150$	$\gamma = 50$	$\gamma = 100$	$\gamma = 150$
Supervised	65.23 ± 0.05	58.94 ± 0.13	55.63 ± 0.38	51.31 ± 0.34	45.82 ± 0.41	40.90 ± 0.39
CBL	65.52 ± 0.31	58.52 ± 0.45	52.36 ± 0.58	51.94 ± 0.71	46.22 ± 0.92	41.58 ± 1.24
Re-Sampling	64.53 ± 0.39	56.34 ± 0.42	53.21 ± 0.51	51.96 ± 0.65	48.13 ± 1.25	40.26 ± 1.88
cRT	67.82 ± 0.14	63.43 ± 0.45	59.56 ± 0.44	56.28 ± 1.45	48.11 ± 0.79	45.02 ± 1.08
LDAM	68.91 ± 0.10	63.15 ± 0.24	58.68 ± 0.30	56.41 ± 0.92	49.27 ± 0.88	45.10 ± 0.75
Mean-Teacher	68.84 ± 0.82	61.33 ± 0.28	54.79 ± 0.31	56.34 ± 1.68	48.55 ± 0.77	45.32 ± 1.20
MixMatch	73.59 ± 0.46	65.03 ± 0.26	62.71 ± 0.29	65.32 ± 1.20	56.41 ± 1.96	52.38 ± 1.88
ReMixMatch	78.96 ± 0.29	72.88 ± 0.12	68.61 ± 0.40	76.83 ± 0.98	70.12 ± 1.23	59.58 ± 1.30
FixMatch	79.10 ± 0.14	71.50 ± 0.31	68.47 ± 0.15	77.34 ± 0.96	68.45 ± 0.94	60.10 ± 0.82
DARP	81.60 ± 0.31	75.23 ± 0.14	69.31 ± 0.26	76.72 ± 0.46	69.41 ± 0.50	61.23 ± 0.31
CReST	82.03 ± 0.26	75.08 ± 0.41	69.84 ± 0.39	76.18 ± 0.36	69.50 ± 0.70	60.81 ± 0.55
Adsh	$\mathbf{83.38 \pm 0.06}$	$\mathbf{76.52 \pm 0.35}$	$\mathbf{71.49 \pm 0.30}$	$\mathbf{79.27 \pm 0.38}$	$\mathbf{70.97 \pm 0.46}$	$\mathbf{62.04 \pm 0.51}$

Class-Imbalanced Semi-Supervised Learning

Class-Imbalanced Semi-Supervised Learning with Adaptive Thresholding

❖ Adsh 실험 결과: 불균형 데이터에 대해서 다양한 방법론과 비교(Imbalanced CIFAR-10)

- Class-imbalanced learning, semi-supervised learning 및 imbalanced semi-supervised learning 방법론과 비교 진행
- 불균형 데이터에 대해서 supervised learning (WideResNet-28-2)을 사용하는 것 보다 불균형이 반영된 방법론들이 더 좋은 성능을 보임
- 불균형이 반영된 방법론보다 unlabeled data를 활용하는 semi-supervised learning이 더 좋은 성능을 보임

Table 1. Comparison of classification performance (Accuracy (%)) on imbalanced CIFAR-10 dataset under three different imbalance ratio: $\gamma = 50, 100, 150$ and two different numbers of labeled data: $N_1 = 1500, M_1 = 3000$ and $N_1 = 500, M_1 = 4000$. The best results are indicated in bold.

	Imbalanced CIFAR-10 Dataset					
	$N_1 = 1500, M_1 = 3000$			$N_1 = 500, M_1 = 4000$		
Algorithm	$\gamma = 50$	$\gamma = 100$	$\gamma = 150$	$\gamma = 50$	$\gamma = 100$	$\gamma = 150$
Supervised	65.23 ± 0.05	58.94 ± 0.13	55.63 ± 0.38	51.31 ± 0.34	45.82 ± 0.41	40.90 ± 0.39
CBL	65.52 ± 0.31	58.52 ± 0.45	52.36 ± 0.58	51.94 ± 0.71	46.22 ± 0.92	41.58 ± 1.24
Re-Sampling	64.53 ± 0.39	56.34 ± 0.42	53.21 ± 0.51	51.96 ± 0.65	48.13 ± 1.25	40.26 ± 1.88
cRT	67.82 ± 0.14	63.43 ± 0.45	59.56 ± 0.44	56.28 ± 1.45	48.11 ± 0.79	45.02 ± 1.08
LDAM	68.91 ± 0.10	63.15 ± 0.24	58.68 ± 0.30	56.41 ± 0.92	49.27 ± 0.88	45.10 ± 0.75
Mean-Teacher	68.84 ± 0.82	61.33 ± 0.28	54.79 ± 0.31	56.34 ± 1.68	48.55 ± 0.77	45.32 ± 1.20
MixMatch	73.59 ± 0.46	65.03 ± 0.26	62.71 ± 0.29	65.32 ± 1.20	56.41 ± 1.96	52.38 ± 1.88
ReMixMatch	78.96 ± 0.29	72.88 ± 0.12	68.61 ± 0.40	76.83 ± 0.98	70.12 ± 1.23	59.58 ± 1.30
FixMatch	79.10 ± 0.14	71.50 ± 0.31	68.47 ± 0.15	77.34 ± 0.96	68.45 ± 0.94	60.10 ± 0.82
DARP	81.60 ± 0.31	75.23 ± 0.14	69.31 ± 0.26	76.72 ± 0.46	69.41 ± 0.50	61.23 ± 0.31
CReST	82.03 ± 0.26	75.08 ± 0.41	69.84 ± 0.39	76.18 ± 0.36	69.50 ± 0.70	60.81 ± 0.55
Adsh	**83.38 ± 0.06**	**76.52 ± 0.35**	**71.49 ± 0.30**	**79.27 ± 0.38**	**70.97 ± 0.46**	**62.04 ± 0.51**

Class-Imbalanced Semi-Supervised Learning

Class-Imbalanced Semi-Supervised Learning with Adaptive Thresholding

❖ Adsh 실험 결과: 불균형 데이터에 대해서 다양한 방법론과 비교(Imbalanced CIFAR-10)

- Class-imbalanced learning, semi-supervised learning 및 imbalanced semi-supervised learning 방법론과 비교 진행
- 불균형 데이터에 대해서 supervised learning (WideResNet-28-2)을 사용하는 것 보다 불균형이 반영된 방법론들이 더 좋은 성능을 보임
- 불균형이 반영된 방법론보다 unlabeled data를 활용하는 semi-supervised learning이 더 좋은 성능을 보임

Table 1. Comparison of classification performance (Accuracy (%)) on imbalanced CIFAR-10 dataset under three different imbalance ratio: $\gamma = 50, 100, 150$ and two different numbers of labeled data: $N_1 = 1500, M_1 = 3000$ and $N_1 = 500, M_1 = 4000$. The best results are indicated in bold.

Imbalanced CIFAR-10 Dataset						
	$N_1 = 1500, M_1 = 3000$			$N_1 = 500, M_1 = 4000$		
Algorithm	$\gamma = 50$	$\gamma = 100$	$\gamma = 150$	$\gamma = 50$	$\gamma = 100$	$\gamma = 150$
Supervised	65.23 ± 0.05	58.94 ± 0.13	55.63 ± 0.38	51.31 ± 0.34	45.82 ± 0.41	40.90 ± 0.39
CBL	65.52 ± 0.31	58.52 ± 0.45	52.36 ± 0.58	51.94 ± 0.71	46.22 ± 0.92	41.58 ± 1.24
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
cRT	67.82 ± 0.14	63.43 ± 0.45	59.56 ± 0.44	56.28 ± 1.45	48.11 ± 0.79	45.02 ± 1.08
LDAM	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	45.10 ± 0.75
Mean-Teacher	68.84 ± 0.82	61.33 ± 0.28	54.79 ± 0.31	56.34 ± 1.68	48.55 ± 0.77	45.32 ± 1.20
MixMatch	73.59 ± 0.46	65.03 ± 0.26	62.71 ± 0.29	65.32 ± 1.20	56.41 ± 1.96	52.38 ± 1.88
ReMixMatch	78.96 ± 0.29	72.88 ± 0.12	68.61 ± 0.40	76.83 ± 0.98	70.12 ± 1.23	59.58 ± 1.30
FixMatch	79.10 ± 0.14	71.50 ± 0.31	68.47 ± 0.15	77.34 ± 0.96	68.45 ± 0.94	60.10 ± 0.82
DARP	81.60 ± 0.31	75.23 ± 0.14	69.31 ± 0.26	76.72 ± 0.46	69.41 ± 0.50	61.23 ± 0.31
CReST	82.03 ± 0.26	75.08 ± 0.41	69.84 ± 0.39	76.18 ± 0.36	69.50 ± 0.70	60.81 ± 0.55
Adsh	**83.38 ± 0.06**	**76.52 ± 0.35**	**71.49 ± 0.30**	**79.27 ± 0.38**	**70.97 ± 0.46**	**62.04 ± 0.51**

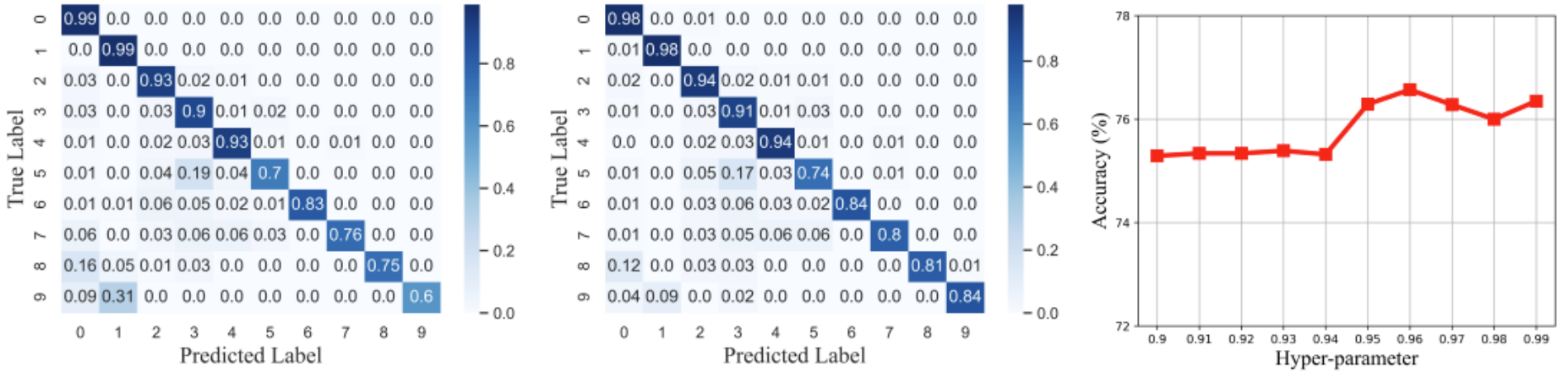
Yang, Y., & Xu, Z. (2020). Rethinking the value of labels for improving class-imbalanced learning. Advances in neural information processing systems, 33, 19290-19301.

Class-Imbalanced Semi-Supervised Learning

Class-Imbalanced Semi-Supervised Learning with Adaptive Thresholding

❖ Adsh 실험 결과: FixMatch과 pseudo label 정확도 비교 및 하이퍼파라미터 threshold별로 실험(Imbalanced CIFAR-10)

- Labeled, unlabeled 데이터 불균형 비율이 1000이며 $N_1 = 1500,\ M_1 = 3000$인 imbalanced CIFAR-10 데이터
- FixMatch에서 소수 클래스들을 다수 클래스로 편향되게 pseudo labeling하는 것을 개선함
- 다양한 기준 threshold에 대해서 robust한 성능을 보임



(a) Confusion Matrix of FixMatch (b) Confusion Matrix of Adsh (c) Hyper-Parameter Sensitivity

Figure 3. Detailed analyses of the Adsh. (a) and (b): Confusion matrix on unlabeled data produced by FixMatch (left) and Adsh(right); (c): Performance robustness with hyper-parameter τ_1 changes.

Conclusion

❖ Summary

- 현실에서는 labeled data를 많이 확보하는 것이 어렵지만 unlabeled data는 상대적으로 확보하기 쉬움
- Semi-supervised learning(SSL)을 활용하여 supervised learning 대비 좋은 성능을 보임
- 그러나 불균형 데이터에서는 다수 클래스에 편향되면서 SSL 성능이 하락됨
- 이번 세미나에서는 불균형 데이터에서도 SSL이 잘 작동하도록 하는 class-imbalanced SSL 방법론들 소개
  - ➢ DARP: Unlabeled data class distribution을 활용하여 pseudo label을 개선하는 방식을 제안
  - ➢ CReST: Minority class pseudo label들의 precision이 높은 점을 활용하여 self-training 방식으로 class re-balancing
  - ➢ Adsh: 기존 FixMatch에서 고정된 threshold를 class adaptive thresholding 기법으로 대체

Reference

1. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., ... & Li, C. L. (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in neural information processing systems, 33, 596-608.

2. Kim, J., Hur, Y., Park, S., Yang, E., Hwang, S. J., & Shin, J. (2020). Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning. Advances in neural information processing systems, 33, 14567-14579.

3. Wei, C., Sohn, K., Mellina, C., Yuille, A., & Yang, F. (2021). Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10857-10866).

4. Guo, L. Z., & Li, Y. F. (2022, June). Class-imbalanced semi-supervised learning with adaptive thresholding. In International Conference on Machine Learning (pp. 8082-8094). PMLR.

5. Yang, Y., & Xu, Z. (2020). Rethinking the value of labels for improving class-imbalanced learning. Advances in neural information processing systems, 33, 19290-19301.

감사합니다