Out-Of-Distribution Generalization :

Subpopulation Shifts and Approaches

Data Mining & Quality Analytics Lab.

2025. 04. 11

발표자: 정진용

발표자 소개

❖ 정진용 (Jinyong Jeong)

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

❖ 관심 연구 분야

- Out-Of-Distribution Generalization & Domain Generalization
- Semi-Supervised Learning & Class-Imbalanced Semi-Supervised Learning

❖ E-mail

- jy_jeong@korea.ac.kr

목차

Introduction

Background of distribution shift

❖ Train dataset을 사용해서 일반화 성능이 좋은 모델을 구축해보자

$$P_{train}(X, Y) = P_{train}(Y|X) \cdot P_{train}(X)$$

$$\mathcal{Y}_{train} = \{\text{강아지, 고양이}\}, Y \in \mathcal{Y}$$

Introduction

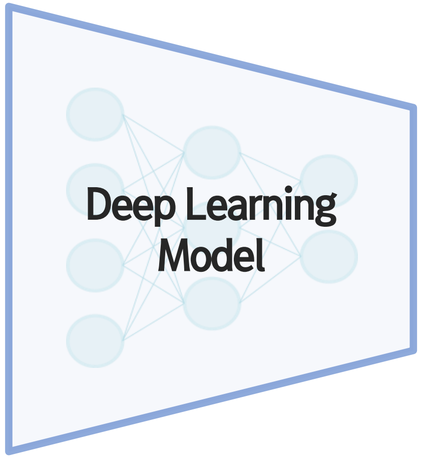
Background of distribution shift

❖ Train dataset을 사용해서 일반화 성능이 좋은 모델을 구축해보자

Test dataset

Train dataset

모델 학습



$$P_{train}(X,Y) = P_{train}(Y|X) \cdot \boxed{P_{train}(X)}$$

$$P_{test}(X,Y) = P_{test}(Y|X) \cdot \boxed{P_{test}(X)}$$

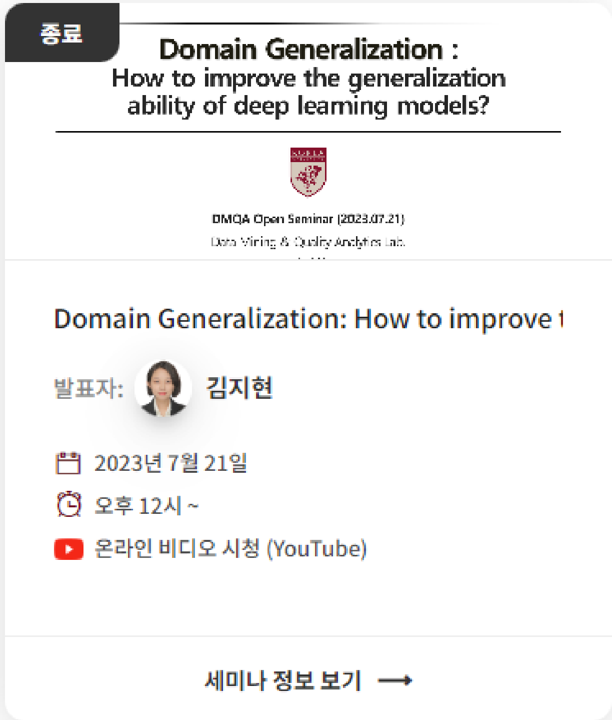
$$\mathcal{Y}_{train} = \{\text{강아지, 고양이}\},\ Y \in \mathcal{Y}$$

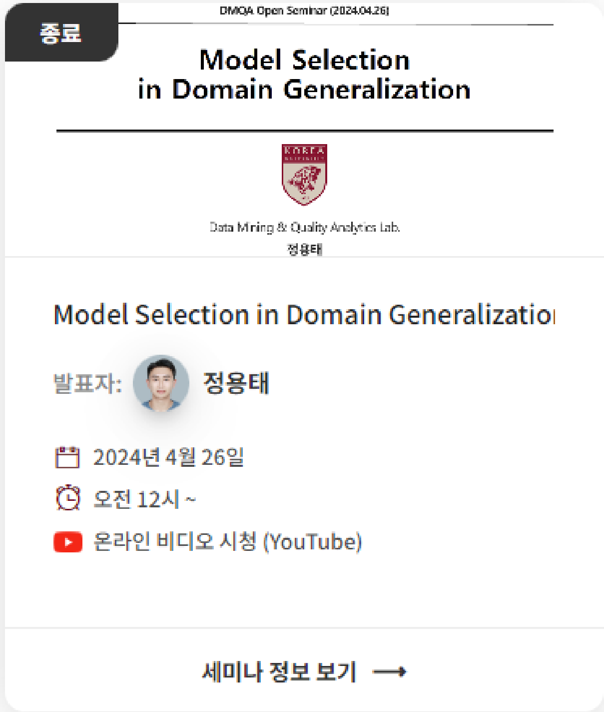
$\neq$

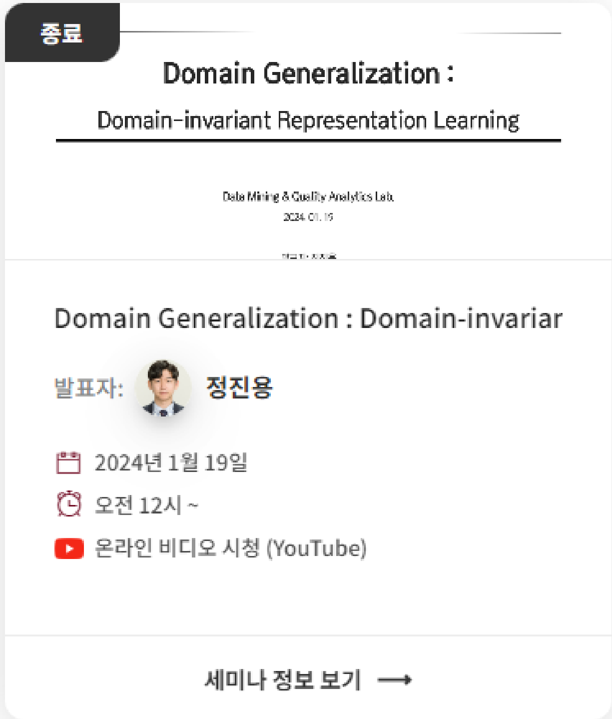
Introduction

Background of distribution shift

❖ Train dataset을 사용해서 일반화 성능이 좋은 모델을 구축해보자







$\mathcal{y}_{train}$ = {강아지, 고양이}, $Y \in \mathcal{Y}$

Introduction

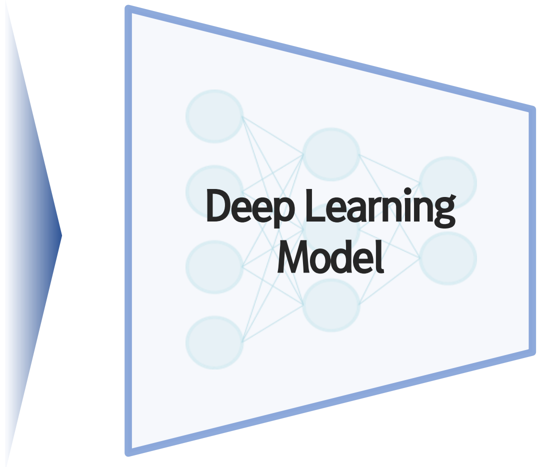
Background of distribution shift

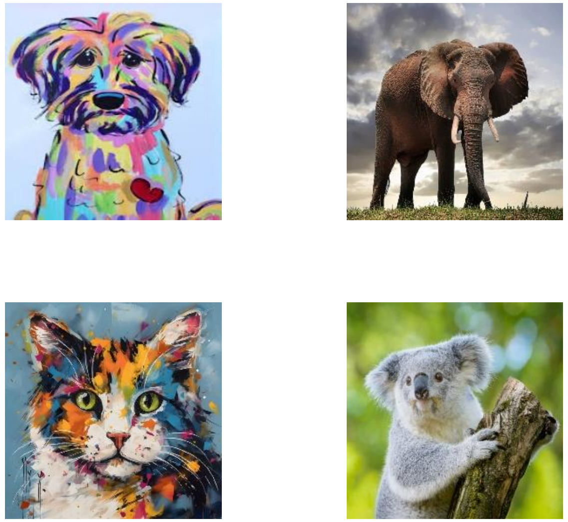
❖ Train dataset을 사용해서 일반화 성능이 좋은 모델을 구축해보자

Test dataset

Train dataset

모델 학습





$P_{train}(X,Y) = P_{train}(Y|X) \cdot P_{train}(X)$

$\mathcal{Y}_{train} = \{$강아지, 고양이$\}, Y \in \mathcal{Y}$

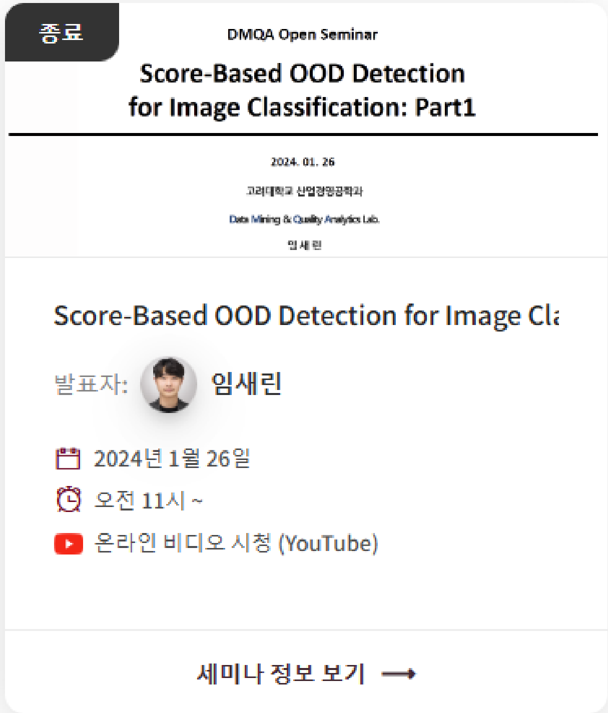
$P_{test}(X,Y) = P_{test}(Y|X) \cdot P_{test}(X)$

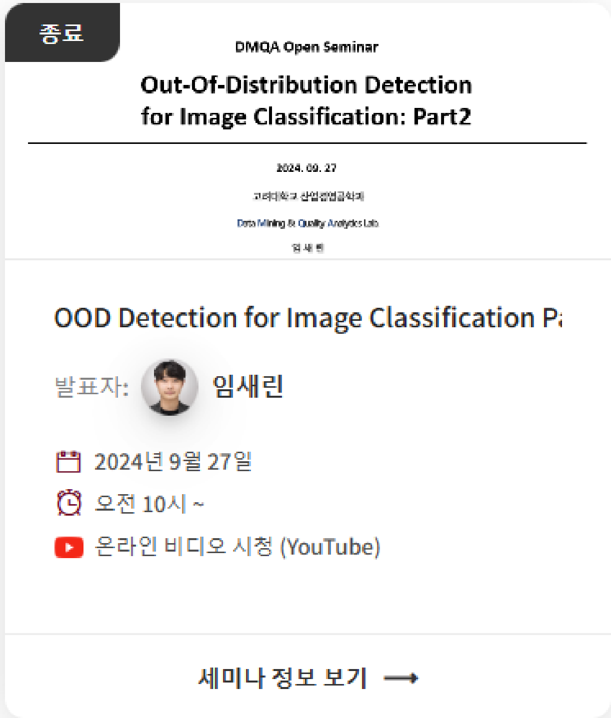
$\mathcal{Y}_{test} = \{$강아지, 고양이, 코끼리, 코알라$\}, Y \in \mathcal{Y}$

Introduction

Background of distribution shift

❖ Train dataset을 사용해서 일반화 성능이 좋은 모델을 구축해보자





$\mathcal{Y}_{train}$ = {강아지, 고양이}, $Y \in \mathcal{Y}$

$\mathcal{Y}_{test}$ = {강아지, 고양이, 코끼리, 코알라}, $Y \in \mathcal{Y}$

Introduction

Background of distribution shift

❖ Train dataset을 사용해서 일반화 성능이 좋은 모델을 구축해보자

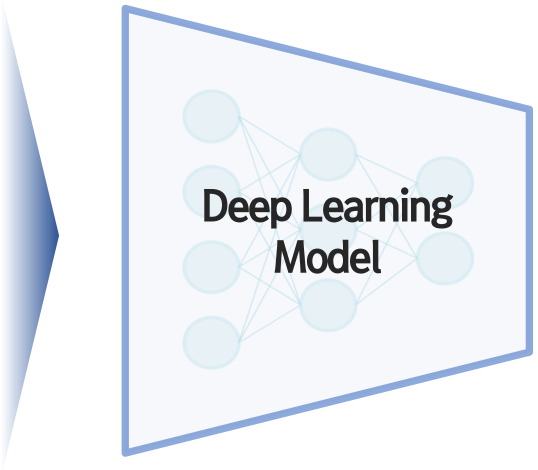
Test dataset

Train dataset

모델 학습

Outdoor

Indoor



Indoor

Outdoor

$P_{train}(X, Y)$

$P_{test}(X, Y)$

Introduction

Background of distribution shift

❖ Train dataset을 사용해서 일반화 성능이 좋은 모델을 구축해보자

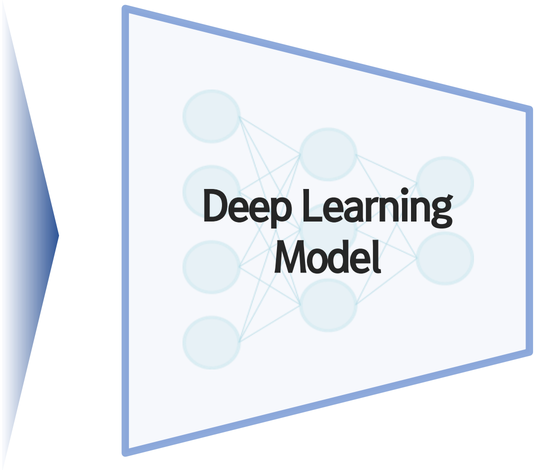
Test dataset

Train dataset

모델 학습

Outdoor

Indoor



Indoor

Outdoor

$$P_{train}(X,Y) = \sum_{k=1}^{K} \pi_k^{train} P(X,Y|Z=k)$$

$$\pi_k^{train} \neq \pi_k^{test}$$

Subpopulation shift → 모델 성능 하락

$$P_{test}(X,Y) = \sum_{k=1}^{K} \pi_k^{test} P(X,Y|Z=k)$$

Introduction

Background of distribution shift

❖ Train dataset을 사용해서 일반화 성능이 좋은 모델을 구축해보자

PACS

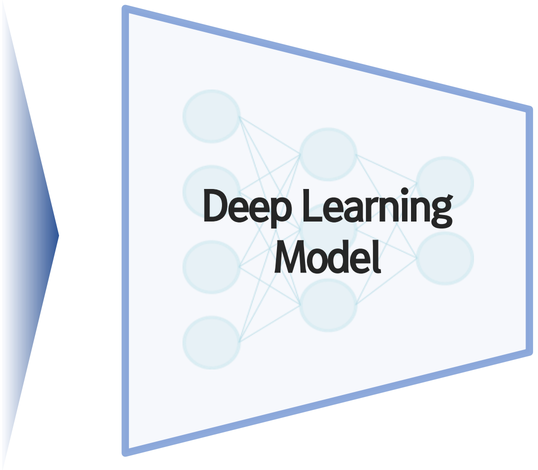
Test dataset

Train dataset

모델 학습

Outdoor

Indoor



Indoor

Outdoor

$$P_{train}(X,Y) = \sum_{k=1}^{K} \pi_k^{train} P(X,Y|Z=k)$$

$$\pi_k^{train} \neq \pi_k^{test}$$

Subpopulation shift → 모델 성능 하락

$$P_{test}(X,Y) = \sum_{k=1}^{K} \pi_k^{test} P(X,Y|Z=k)$$

Subpopulation Shift

❖ Change is Hard: A Closer Look at Subpopulation Shift (ICML, 2023)

- MIT 컴퓨터 과학 및 인공지능 연구소에서 연구되었으며 2025년 4월 11일 기준 120회 인용됨
- Subpopulation shift를 구성하는 basic shift들에 대해서 설명하고 다양한 알고리즘 적용 및 비교 분석한 논문

Change is Hard: A Closer Look at Subpopulation Shift

Yuzhe Yang [*1] **Haoran Zhang** [*1] **Dina Katabi** [1] **Marzyeh Ghassemi** [1]

Abstract

Machine learning models often perform poorly on *subgroups* that are underrepresented in the training data. Yet, little is understood on the variation in mechanisms that cause subpopulation shifts, and how algorithms generalize across such diverse shifts at scale. In this work, we provide a fine-grained analysis of subpopulation shift. We first propose a unified framework that dissects and explains common shifts in subgroups. We then establish a comprehensive benchmark of 20 state-of-the-art algorithms evaluated on 12 real-world datasets in vision, language, and healthcare domains. With results obtained from training over 10,000 models, we reveal intriguing observations for future progress in this space. First, existing algorithms only improve subgroup robustness over certain types of shifts but not others. Moreover, while current algorithms rely on group-annotated validation data for model selection, we find that a simple selection criterion based on worst-class accuracy is surprisingly effective even without any group information. Finally, unlike existing works that solely aim to improve worst-group accuracy (WGA), we demonstrate the fundamental tradeoff between WGA and other important metrics, highlighting the need to carefully choose testing metrics. Code and data are available at: https://github.com/YyzHarry/SubpopBench.

(Koh et al., 2021). In such settings, models may have high overall performance but still perform poorly in rare subgroups (Hashimoto et al., 2018; Zhang et al., 2020).

A well-studied type of subpopulation shift occurs when data contains *spurious correlations* (Geirhos et al., 2020) – non-causal relationships between the input and the label which may shift in deployment (Simon, 1954). For example, image classifiers frequently make use of non-robust features such as image backgrounds (Xiao et al., 2016), textures (Geirhos et al., 2018), and erroneous markings (DeGrave et al., 2021). However, there has been little work in defining subpopulation shift in a holistic way, understanding *when* these shifts happen, and *how* state-of-the-art (SOTA) algorithms generalize under diverse and realistic shifts. Subpopulation shift can encompass a much wider array of underlying mechanisms. First, different attributes in data often exhibit skewed distributions, inevitably causing *attribute imbalance* (Martinez et al., 2021). Moreover, certain labels can have significantly fewer observations, where such long-tailed label distribution induces severe *class imbalance* (Liu et al., 2019b). Finally, certain attributes may have no training data at all, which motivates the need for *attribute generalization* to unseen subpopulations (Santurkar et al., 2020).

In this work, we systematically investigate subpopulation shift in realistic evaluation settings. We first formalize a generic framework of subpopulation shift, which decomposes *attribute* and *class* to enable fine-grained analyses. We demonstrate that this modeling covers and explains the aforementioned common subgroup shifts, which are basic

Yang, Y., Zhang, H., Katabi, D., & Ghassemi, M. (2023, July). Change is Hard: A Closer Look at Subpopulation Shift. In International Conference on Machine Learning (pp. 39584-39622). PMLR.
https://subpopbench.csail.mit.edu/

Subpopulation Shift

Change is Hard: A Closer Look at Subpopulation Shift

Outdoor

Indoor

입력 $x = (x_{core}, \mathbb{a})$

$$\mathbb{P}(x, y) = \boxed{\mathbb{P}(y|x)} \cdot \mathbb{P}(x)$$

$$\boxed{\mathbb{P}(y|x)} = \frac{\mathbb{P}(x|y)}{\mathbb{P}(x)} \cdot \mathbb{P}(y)$$

$$= \frac{\mathbb{P}(x_{core}, \mathbb{a}|y)}{\mathbb{P}(x_{core}, \mathbb{a})} \cdot \mathbb{P}(y)$$

$$= \boxed{\frac{\mathbb{P}(x_{core}|y)}{\mathbb{P}(x_{core})}} \cdot \boxed{\frac{\mathbb{P}(\mathbb{a}|y, x_{core})}{\mathbb{P}(\mathbb{a}|x_{core})}} \cdot \boxed{\mathbb{P}(y)}$$

Class bias
→ Class (label) distribution

Attribute bias
→ Attribute distribution

Pointwise mutual information
→ Robust indicator, invariant

Subpopulation Shift

Change is Hard: A Closer Look at Subpopulation Shift

Train dataset

강아지 고양이

Outdoor

Indoor

입력 $x = (x_{core}, \mathbb{a})$

$$\mathbb{P}(x, y) = \boxed{\mathbb{P}(y|x)} \cdot \mathbb{P}(x)$$

$$\boxed{\mathbb{P}(y|x)} = \frac{\mathbb{P}(x|y)}{\mathbb{P}(x)} \cdot \mathbb{P}(y)$$

$$= \frac{\mathbb{P}(x_{core}, \mathbb{a}|y)}{\mathbb{P}(x_{core}, \mathbb{a})} \cdot \mathbb{P}(y)$$

$$= \boxed{\frac{\mathbb{P}(x_{core}|y)}{\mathbb{P}(x_{core})}} \cdot \boxed{\frac{\mathbb{P}(\mathbb{a}|y, x_{core})}{\mathbb{P}(\mathbb{a}|x_{core})}} \cdot \boxed{\mathbb{P}(y)}$$

Class bias
→ Class (label) distribution

Attribute bias
→ Attribute distribution

Pointwise mutual information
→ Robust indicator, invariant

Yang, Y., Zhang, H., Katabi, D., & Ghassemi, M. (2023, July). Change is Hard: A Closer Look at Subpopulation Shift. In International Conference on Machine Learning (pp. 39584-39622). PMLR.
https://subpopbench.csail.mit.edu/

Subpopulation Shift

Change is Hard: A Closer Look at Subpopulation Shift

Outdoor

Indoor

입력 $x = (x_{core}, \mathbb{a})$

$$\mathbb{P}(x,y) = \boxed{\mathbb{P}(y|x)} \cdot \mathbb{P}(x)$$

$$\boxed{\mathbb{P}(y|x)} = \frac{\mathbb{P}(x|y)}{\mathbb{P}(x)} \cdot \mathbb{P}(y)$$

$$= \frac{\mathbb{P}(x_{core}, \mathbb{a}|y)}{\mathbb{P}(x_{core}, \mathbb{a})} \cdot \mathbb{P}(y)$$

$$= \boxed{\frac{\mathbb{P}(x_{core}|y)}{\mathbb{P}(x_{core})}} \cdot \boxed{\frac{\mathbb{P}(\mathbb{a}|y, x_{core})}{\mathbb{P}(\mathbb{a}|x_{core})}} \cdot \boxed{\mathbb{P}(y)}$$

Class bias
→ Class (label) distribution

Attribute bias
→ Attribute distribution

Pointwise mutual information
→ Robust indicator, invariant

Train dataset

강아지 고양이

강아지 고양이

Outdoor

Indoor

Attribute

$$P_{train}(a|y, x_{core}) \gg P_{train}(a|x_{core})$$

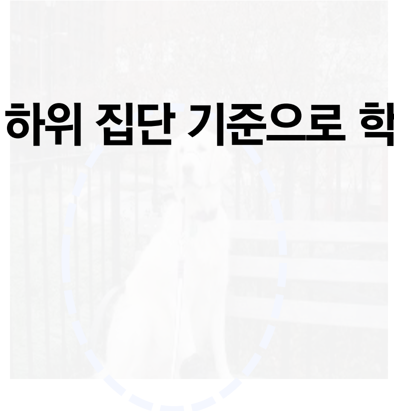
Yang, Y., Zhang, H., Katabi, D., & Ghassemi, M. (2023, July). Change is Hard: A Closer Look at Subpopulation Shift. In International Conference on Machine Learning (pp. 39584-39622). PMLR.
https://subpopbench.csail.mit.edu/

Subpopulation Shift

Change is Hard: A Closer Look at Subpopulation Shift

학습 데이터셋: 하위 집단 기준으로 학습 편향을 만들 수 있는 요인이 존재함

Outdoor



Indoor

입력 $x = (x_{core}, a)$

$$P(x,y) = \boxed{P(y|x)} \cdot P(x)$$

$$\boxed{P(y|x)} = \frac{P(x|y)}{P(x)} \cdot P(y)$$

$$= \frac{P(x_{core}, a|y)}{P(x_{core}, a)} \cdot P(y)$$

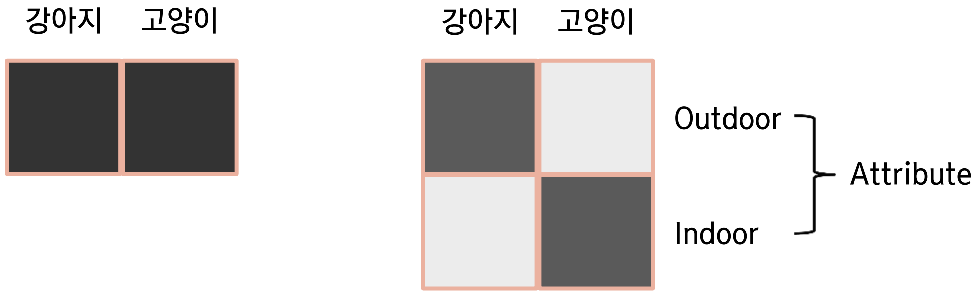
$$= \boxed{\frac{P(x_{core}|y)}{P(x_{core})}} \cdot \boxed{\frac{P(a|y, x_{core})}{P(a|x_{core})}} \cdot \boxed{P(y)}$$

Class bias
→ Class (label) distribution

Attribute bias
→ Attribute distribution

Pointwise mutual information
→ Robust indicator, invariant

Train dataset



$$P_{train}(a|y, x_{core}) \gg P_{train}(a|x_{core})$$

Subpopulation Shift

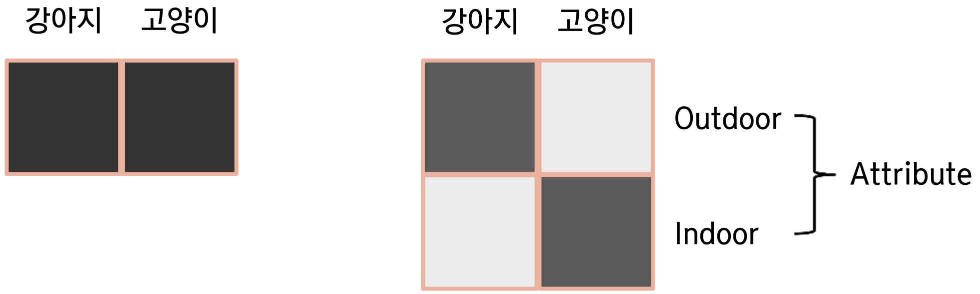
Change is Hard: A Closer Look at Subpopulation Shift

학습 데이터셋: 하위 집단 기준으로 학습 편향을 만들 수 있는 요인이 존재함

학습 목표: 모든 하위 집단에 대해서 공정한 성능을 갖도록 학습

average accuracy 이외에도 worst-group accuracy 확인

Train dataset



$$P_{train}(a|y, x_{core}) \gg P_{train}(a|x_{core})$$

Test dataset

$$P_{test}(a|y, x_{core}) = P_{test}(a|x_{core})$$

Yang, Y., Zhang, H., Katabi, D., & Ghassemi, M. (2023, July). Change is Hard: A Closer Look at Subpopulation Shift. In International Conference on Machine Learning (pp. 39584-39622). PMLR.
https://subpopbench.csail.mit.edu/

Subpopulation Shift

Change is Hard: A Closer Look at Subpopulation Shift

$$\mathbb{P}(y|x) = \underbrace{\boxed{\frac{\mathbb{P}(x_{core}|y)}{\mathbb{P}(x_{core})}}}_{\text{PMI}} \cdot \underbrace{\boxed{\frac{\mathbb{P}(\mathbb{a}|y, x_{core})}{\mathbb{P}(\mathbb{a}|x_{core})}}}_{\text{Attribute bias}} \cdot \underbrace{\boxed{\mathbb{P}(y)}}_{\text{Class bias}}$$

❖ Basic types of Subpopulation shift

- 데이터 포인트 관점에서 a와 y에 대해 분해를 하여 4가지 basic shift로 나눠 볼 수 있음
- Basic shift들이 혼합되어 subpopulation shift를 발생시키기도 함

Subpopulation Shift Type	Attribute Bias	Class Bias	Impact on Classification Model
(1)Spurious Correlations	$P_{train}(a\|y, x_{core}) \gg P_{train}(a\|x_{core})$ $P_{test}(a\|y, x_{core}) = P_{test}(a\|x_{core})$	·	$\frac{\mathbb{P}(a\|y, x_{core})}{\mathbb{P}(a\|x_{core})} \gg 1 \implies \mathbb{P}(y\|x) \uparrow$
(2)Attribute Imbalance	$P_{train}(a\|y, x_{core})$ $\gg P_{train}(a'\|y, x_{core})$ $P_{test}(a\|y, x_{core}) = P_{test}(a'\|y, x_{core})$	·	$\frac{\mathbb{P}(a\|y, x_{core})}{\mathbb{P}(a\|x_{core})} \gg \frac{\mathbb{P}(a'\|y, x_{core})}{\mathbb{P}(a'\|x_{core})}$ $\implies \mathbb{P}(y\|x_{core}, a) \gg \mathbb{P}(y\|x_{core}, a')$
(3)Class Imbalance	·	$P_{train}(Y = y) \gg P_{train}(Y = y')$ $P_{test}(Y = y) = P_{test}(Y = y')$	$\mathbb{P}(y) \gg \mathbb{P}(y') \implies \mathbb{P}(y\|x) \gg \mathbb{P}(y'\|x)$
(4)Attribute Generalization	$P_{train}(a\|y, x_{core}) = 0, \forall a \in \mathbb{A}^{unseen}$ $P_{test}(a\|y, x_{core}) > 0, \forall a \in \mathbb{A}$	Unconstrained	Generalize to $\mathbb{A}^{unseen}$

Yang, Y., Zhang, H., Katabi, D., & Ghassemi, M. (2023, July). Change is Hard: A Closer Look at Subpopulation Shift. In International Conference on Machine Learning (pp. 39584-39622). PMLR.
https://subpopbench.csail.mit.edu/

Subpopulation Shift

Change is Hard: A Closer Look at Subpopulation Shift

$$\mathbb{P}(y|x) = \boxed{\frac{\mathbb{P}(x_{core}|y)}{\mathbb{P}(x_{core})}} \cdot \boxed{\frac{\mathbb{P}(a|y, x_{core})}{\mathbb{P}(a|x_{core})}} \cdot \boxed{\mathbb{P}(y)}$$

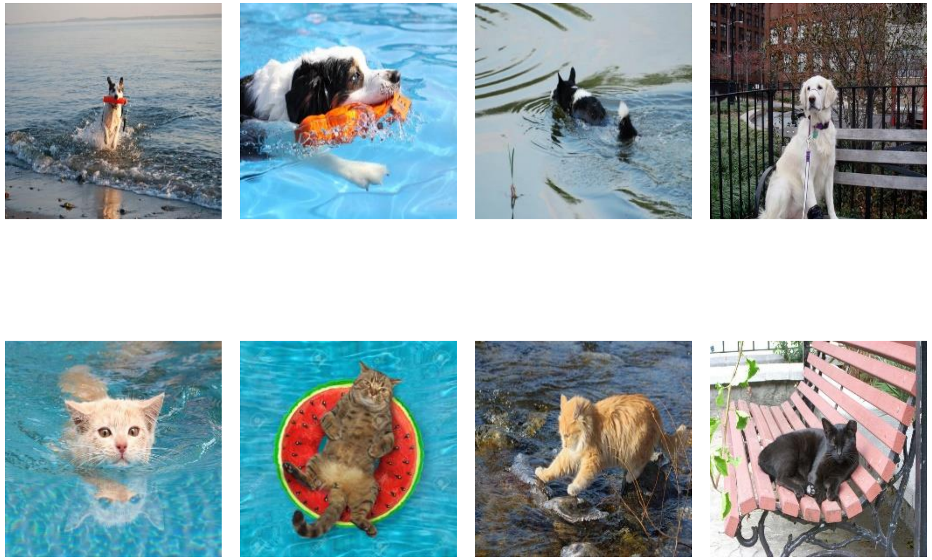
PMI Attribute bias Class bias

Subpopulation Shift Type	Attribute Bias	Class Bias	Impact on Classification Model
(2)Attribute Imbalance	$P_{train}(a\|y, x_{core}) \gg P_{train}(a'\|y, x_{core})$ $P_{test}(a\|y, x_{core}) = P_{test}(a'\|y, x_{core})$	·	$\frac{\mathbb{P}(a\|y, x_{core})}{\mathbb{P}(a\|x_{core})} \gg \frac{\mathbb{P}(a'\|y, x_{core})}{\mathbb{P}(a'\|x_{core})}$ $\Longrightarrow \mathbb{P}(y\|x_{core}, a) \gg \mathbb{P}(y\|x_{core}, a')$

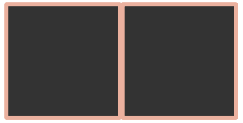
Train dataset

강아지

고양이



강아지 고양이



Yang, Y., Zhang, H., Katabi, D., & Ghassemi, M. (2023, July). Change is Hard: A Closer Look at Subpopulation Shift. In International Conference on Machine Learning (pp. 39584-39622). PMLR.

https://subpopbench.csail.mit.edu/

Subpopulation Shift

Change is Hard: A Closer Look at Subpopulation Shift

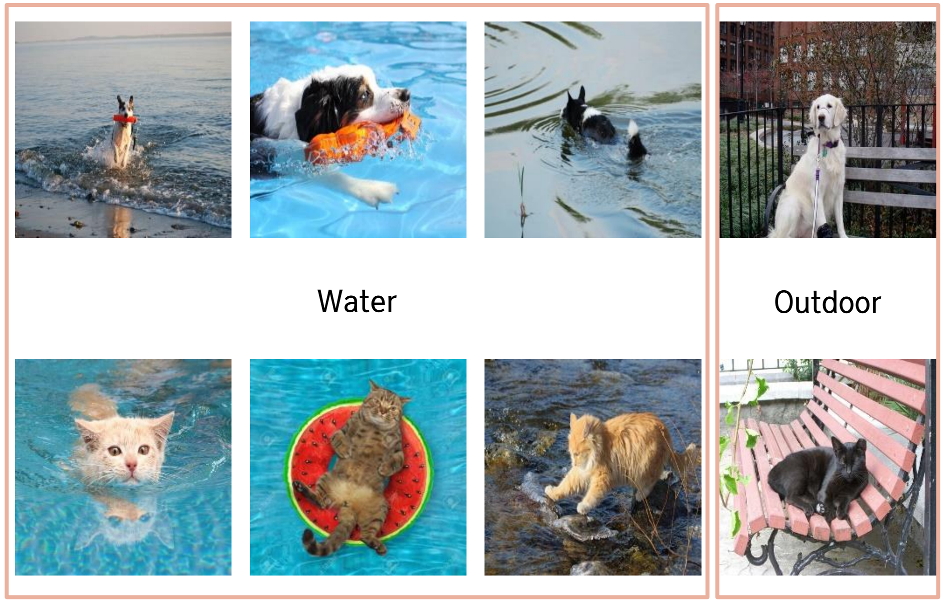
$$\mathbb{P}(y|x) = \underbrace{\boxed{\frac{\mathbb{P}(x_{core}|y)}{\mathbb{P}(x_{core})}}}_{\text{PMI}} \cdot \underbrace{\boxed{\frac{\mathbb{P}(\mathbb{a}|y, x_{core})}{\mathbb{P}(\mathbb{a}|x_{core})}}}_{\text{Attribute bias}} \cdot \underbrace{\boxed{\mathbb{P}(y)}}_{\text{Class bias}}$$

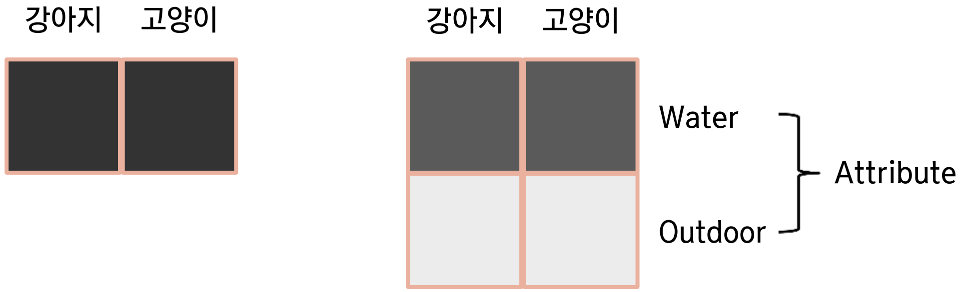
Subpopulation Shift Type	Attribute Bias	Class Bias	Impact on Classification Model
(2)Attribute Imbalance	$P_{train}(a\|y, x_{core}) \gg P_{train}(a'\|y, x_{core})$ $P_{test}(a\|y, x_{core}) = P_{test}(a'\|y, x_{core})$	·	$\frac{\mathbb{P}(a\|y, x_{core})}{\mathbb{P}(a\|x_{core})} \gg \frac{\mathbb{P}(a'\|y, x_{core})}{\mathbb{P}(a'\|x_{core})} \Longrightarrow \mathbb{P}(y\|x_{core}, a) \gg \mathbb{P}(y\|x_{core}, a')$

Train dataset

강아지

고양이





'테스트 시 outdoor attribute에 대해서 lower prediction confidence'

Yang, Y., Zhang, H., Katabi, D., & Ghassemi, M. (2023, July). Change is Hard: A Closer Look at Subpopulation Shift. In International Conference on Machine Learning (pp. 39584-39622). PMLR.
https://subpopbench.csail.mit.edu/

Subpopulation Shift

Change is Hard: A Closer Look at Subpopulation Shift

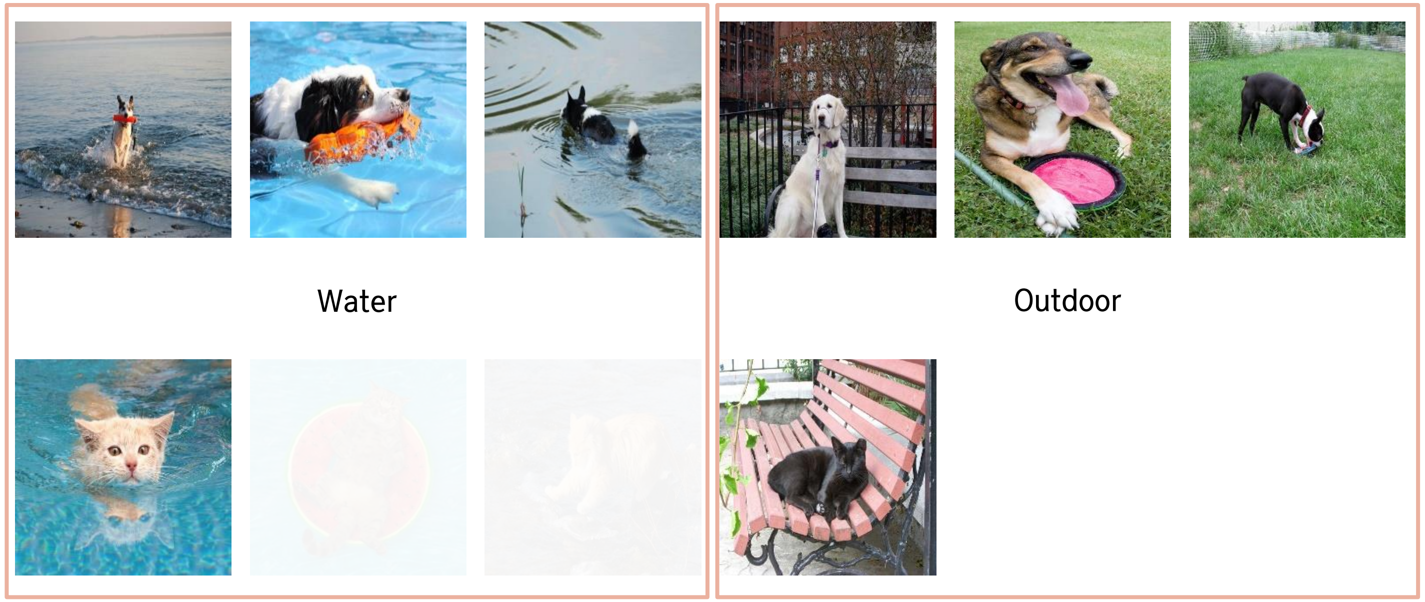
$$\mathbb{P}(y|x) = \underbrace{\boxed{\frac{\mathbb{P}(x_{core}|y)}{\mathbb{P}(x_{core})}}}_{\text{PMI}} \cdot \underbrace{\boxed{\frac{\mathbb{P}(a|y, x_{core})}{\mathbb{P}(a|x_{core})}}}_{\text{Attribute bias}} \cdot \underbrace{\boxed{\mathbb{P}(y)}}_{\text{Class bias}}$$

Subpopulation Shift Type	Attribute Bias	Class Bias	Impact on Classification Model
(3)Class Imbalance	·	$P_{train}(Y=y) \gg P_{train}(Y=y')$ $P_{test}(Y=y) = P_{test}(Y=y')$	$\mathbb{P}(y) \gg \mathbb{P}(y') \Longrightarrow \mathbb{P}(y|x) \gg \mathbb{P}(y'|x)$

Train dataset

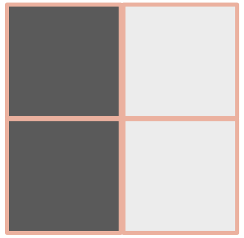
강아지

고양이



강아지 고양이

강아지 고양이



Water

Outdoor

Subpopulation Shift

Change is Hard: A Closer Look at Subpopulation Shift

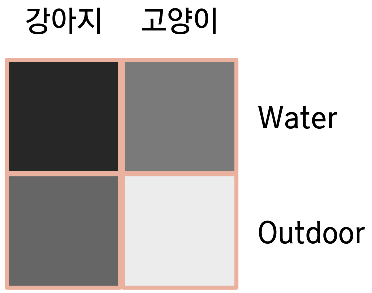
$$\mathbb{P}(y|x) = \underbrace{\boxed{\frac{\mathbb{P}(x_{core}|y)}{\mathbb{P}(x_{core})}}}_{\text{PMI}} \cdot \underbrace{\boxed{\frac{\mathbb{P}(\mathbb{a}|y, x_{core})}{\mathbb{P}(\mathbb{a}|x_{core})}}}_{\text{Attribute bias}} \cdot \underbrace{\boxed{\mathbb{P}(y)}}_{\text{Class bias}}$$

Subpopulation Shift Type	Attribute Bias	Class Bias	Impact on Classification Model
(4)Attribute Generalization	$P_{train}(a\|y, x_{core}) = 0, \forall a \in \mathbb{A}^{unseen}$ $P_{test}(a\|y, x_{core}) > 0, \forall a \in \mathbb{A}$	Unconstrained	Generalize to $\mathbb{A}^{unseen}$

Train dataset

강아지

고양이

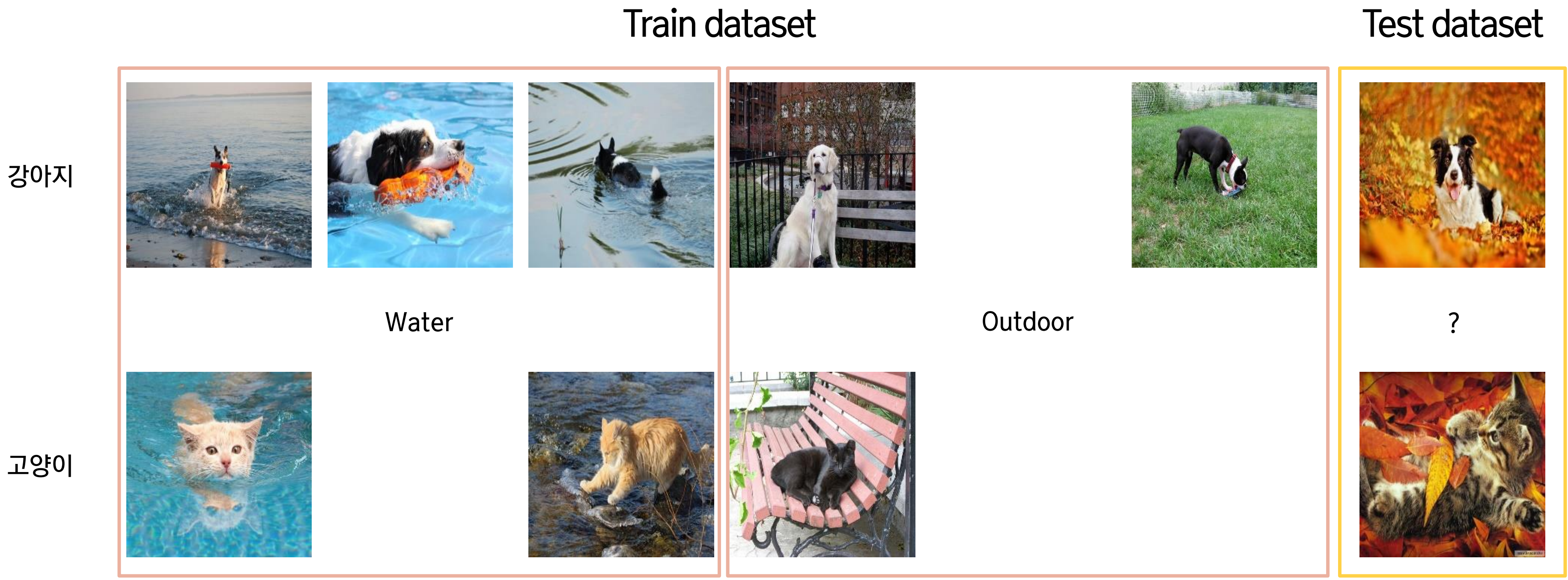


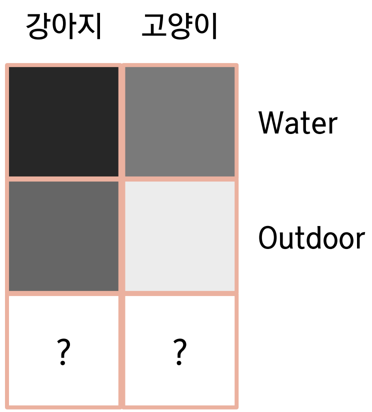
Subpopulation Shift

Change is Hard: A Closer Look at Subpopulation Shift

$$\mathbb{P}(y|x) = \underbrace{\boxed{\frac{\mathbb{P}(x_{core}|y)}{\mathbb{P}(x_{core})}}}_{\text{PMI}} \cdot \underbrace{\boxed{\frac{\mathbb{P}(\mathbb{a}|y, x_{core})}{\mathbb{P}(\mathbb{a}|x_{core})}}}_{\text{Attribute bias}} \cdot \underbrace{\boxed{\mathbb{P}(y)}}_{\text{Class bias}}$$

Subpopulation Shift Type	Attribute Bias	Class Bias	Impact on Classification Model
(4)Attribute Generalization	$P_{train}(a|y, x_{core}) = 0, \forall a \in \mathbb{A}^{unseen}$ $P_{test}(a|y, x_{core}) > 0, \forall a \in \mathbb{A}$	Unconstrained	Generalize to $\mathbb{A}^{unseen}$





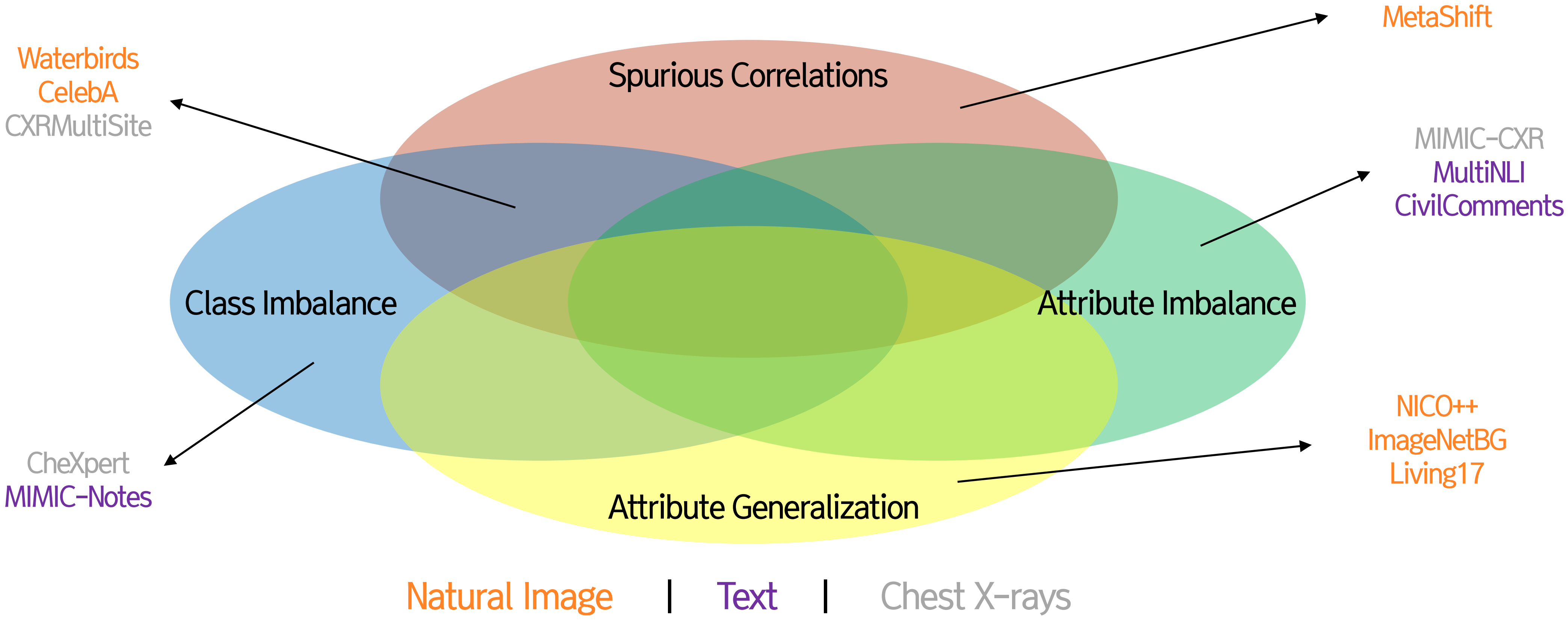
Yang, Y., Zhang, H., Katabi, D., & Ghassemi, M. (2023, July). Change is Hard: A Closer Look at Subpopulation Shift. In International Conference on Machine Learning (pp. 39584-39622). PMLR.
https://subpopbench.csail.mit.edu/

Subpopulation Shift

Change is Hard: A Closer Look at Subpopulation Shift

❖ Benchmarking subpopulation shift using real-world datasets

- Vision, language, health care 도메인 등 총 12가지 데이터셋에 대해서 분석을 함
- 데이터셋들은 basic shift들이 혼합되어 있음



Yang, Y., Zhang, H., Katabi, D., & Ghassemi, M. (2023, July). Change is Hard: A Closer Look at Subpopulation Shift. In International Conference on Machine Learning (pp. 39584-39622). PMLR.
https://subpopbench.csail.mit.edu/

Subpopulation Shift

Change is Hard: A Closer Look at Subpopulation Shift

❖ Benchmarking subpopulation shift using real-world datasets

- Vision, language, health care 도메인 등 총 12가지 데이터셋에 대해서 분석을 함
- 데이터셋들은 basic shift들이 혼합되어 있음

Dataset	Data type	# Attr.	# Classes	# Train set	# Val. set	# Test set	Max group	Min group	Shift type			
									SC	AI	CI	AG
Waterbirds	Image	2	2	4795	1199	5794	3498 (73.0%)	56 (1.2%)	✓	✓	✓	
CelebA	Image	2	2	162770	19867	19962	71629 (44.0%)	1387 (0.9%)	✓		✓	
MetaShift	Image	2	2	2276	349	874	789 (34.7%)	196 (8.6%)	✓			
ImageNetBG	Image	N/A	9	183006	7200	4050	N/A	N/A				✓
NICO++	Image	6	60	62657	8726	17483	811 (1.3%)	0 (0.0%)		✓	✓	✓
Living17	Image	N/A	17	39780	4420	1700	N/A	N/A				✓
MultiNLI	Text	2	3	206175	82462	123712	67376 (32.7%)	1521 (0.7%)		✓		
CivilComments	Text	8	2	148304	24278	71854	31282 (21.1%)	1003 (0.7%)		✓	✓	
MIMICNotes	Clinical text	2	2	16149	3229	6460	8359 (51.8%)	676 (4.2%)			✓	
MIMIC-CXR	Chest X-rays	6	2	303591	17859	35717	68575 (22.6%)	7846 (2.6%)		✓		
CheXpert	Chest X-rays	6	2	167093	22280	33419	51606 (30.9%)	506 (0.3%)		✓	✓	
CXRMultisite	Chest X-rays	2	2	338134	19891	39781	299089 (88.5%)	574 (0.2%)	✓	✓	✓	

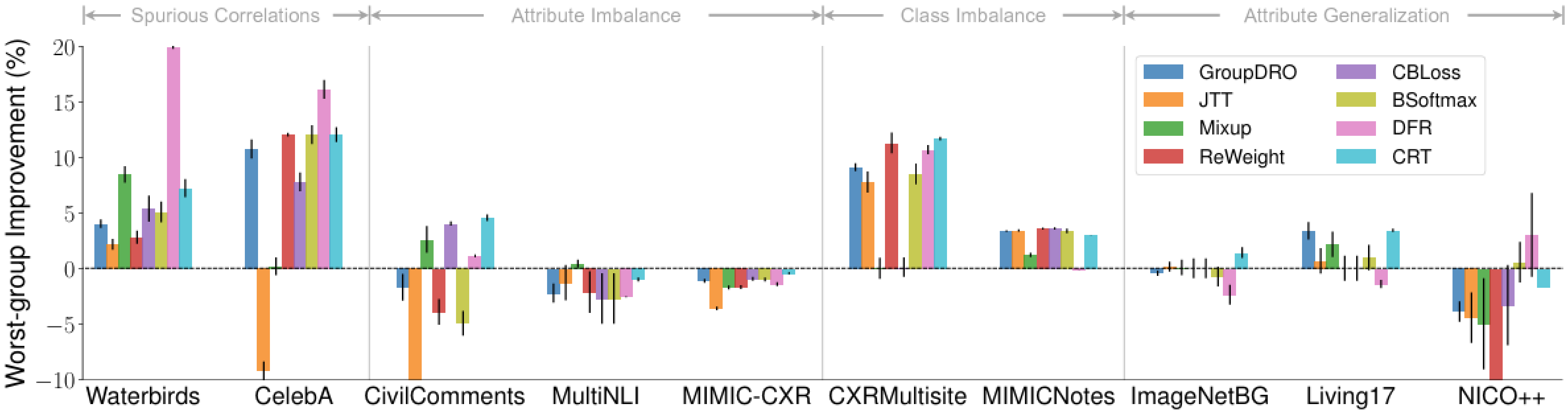
Yang, Y., Zhang, H., Katabi, D., & Ghassemi, M. (2023, July). Change is Hard: A Closer Look at Subpopulation Shift. In International Conference on Machine Learning (pp. 39584-39622). PMLR.
https://subpopbench.csail.mit.edu/

Subpopulation Shift

Change is Hard: A Closer Look at Subpopulation Shift

❖ SOTA algorithms only improve certain types of shift – Attributes are unknown in both training and validation set

- Shift에 가장 많은 영향을 받은 하위 집단(worst-group)에 대해서 일반적인 지도학습(ERM) 대비 성능 향상 비교
- 모든 shift type에서 가장 좋은 성능을 보이는 알고리즘은 없음
- Attribute imbalance, attribute generalization에서는 더 개선된 알고리즘이 필요함



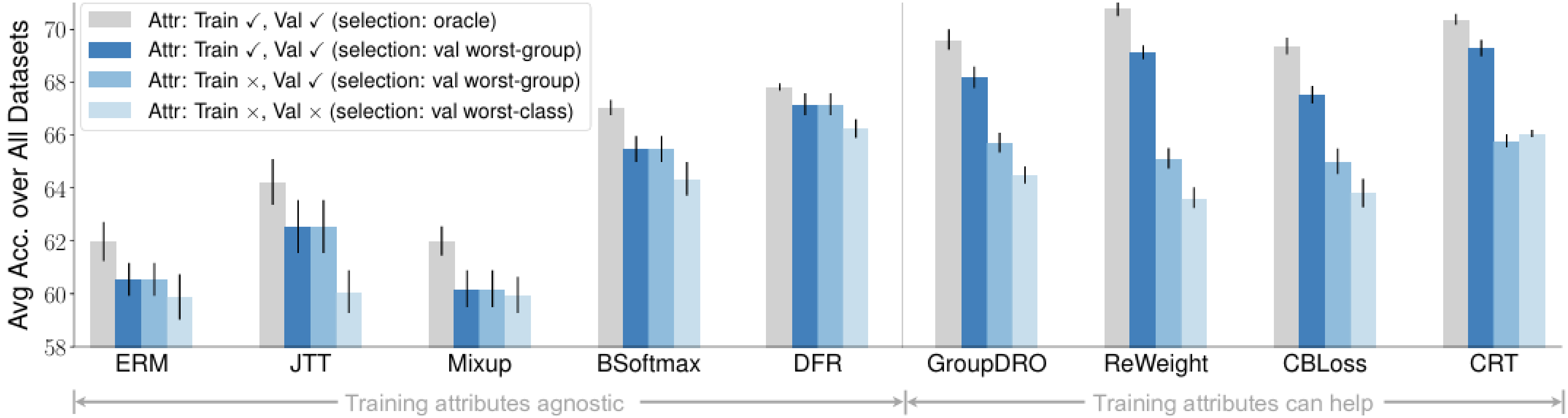
Yang, Y., Zhang, H., Katabi, D., & Ghassemi, M. (2023, July). Change is Hard: A Closer Look at Subpopulation Shift. In International Conference on Machine Learning (pp. 39584-39622). PMLR.
https://subpopbench.csail.mit.edu/

Subpopulation Shift

Change is Hard: A Closer Look at Subpopulation Shift

❖ Attribute availability & model selection

- Attribute 정보를 학습 또는 검증 단계에서 사용할 수 있는지 여부에 따라서 성능 편차가 존재할 수 있음
- 평가에 사용할 model을 선택하는 방식에 따라서도 성능 편차가 존재함



Yang, Y., Zhang, H., Katabi, D., & Ghassemi, M. (2023, July). Change is Hard: A Closer Look at Subpopulation Shift. In International Conference on Machine Learning (pp. 39584-39622). PMLR.
https://subpopbench.csail.mit.edu/

Subpopulation Shift

1-stage method for subpopulation shift

❖ Distributionally robust neural networks for group shifts – GroupDRO (ICLR, 2020)

- Stanford, Microsoft 연구원들에 의해 연구되었으며 2025년 4월 11일 기준 2,003회 인용됨
- Distribution shift 중 subpopulation (group) shift 문제를 distributionally robust optimization 개념으로 해결한 논문

Published as a conference paper at ICLR 2020

DISTRIBUTIONALLY ROBUST NEURAL NETWORKS FOR GROUP SHIFTS: ON THE IMPORTANCE OF REGULARIZATION FOR WORST-CASE GENERALIZATION

Shiori Sagawa*
Stanford University
ssagawa@cs.stanford.edu

Pang Wei Koh*
Stanford University
pangwei@cs.stanford.edu

Tatsunori B. Hashimoto
Microsoft
tahashim@microsoft.com

Percy Liang
Stanford University
pliang@cs.stanford.edu

ABSTRACT

Overparameterized neural networks can be highly accurate *on average* on an i.i.d. test set yet consistently fail on atypical groups of the data (e.g., by learning spurious correlations that hold on average but not in such groups). Distributionally robust optimization (DRO) allows us to learn models that instead minimize the *worst-case* training loss over a set of pre-defined groups. However, we find that naively applying group DRO to overparameterized neural networks fails: these models can perfectly fit the training data, and any model with vanishing average training loss also already has vanishing worst-case training loss. Instead, the poor worst-case performance arises from poor *generalization* on some groups. By coupling group DRO models with increased regularization—a stronger-than-typical ℓ_2 penalty or early stopping—we achieve substantially higher worst-group accuracies, with 10–40 percentage point improvements on a natural language inference task and two image tasks, while maintaining high average accuracies. Our results suggest that regularization is important for worst-group generalization in the overparameterized regime, even if it is not needed for average generalization. Finally, we introduce a stochastic optimization algorithm, with convergence guarantees, to efficiently train group DRO models.

Subpopulation Shift

Distributionally robust neural networks for group shifts – GroupDRO

❖ GroupDRO vs. ERM

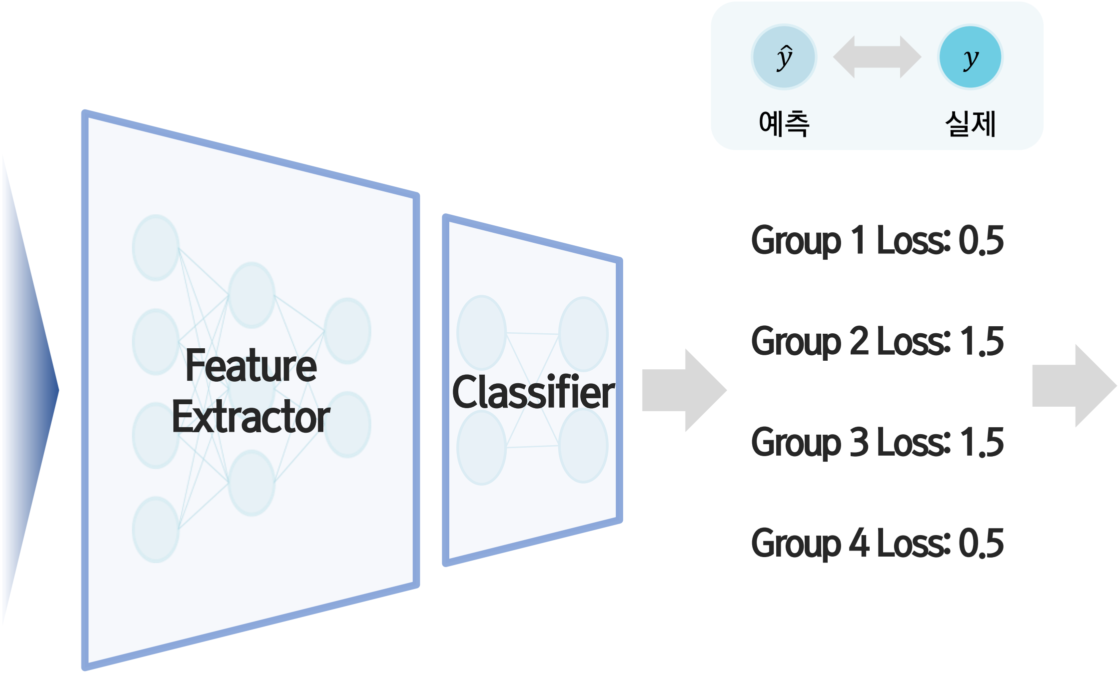
- ERM은 일반적인 모델 지도 학습에서 사용하는 목적식으로써 평균 성능을 최적화하는 방식
- GroupDRO는 그룹별 불균형 요소를 반영하여 worst-group 성능을 개선하는 방식

Group 1
Y : Dog
a : Outdoor

Group 3
Y : Cat
a : Outdoor

Group 2
Y : Dog
a : Indoor

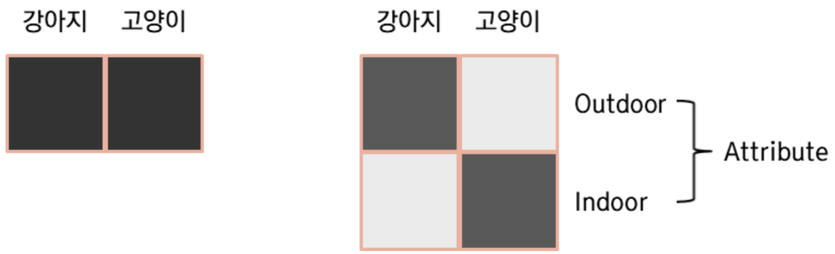
Group 4
Y : Cat
a : Indoor



Subpopulation Shift

Distributionally robust neural networks for group shifts – GroupDRO

Train dataset



❖ GroupDRO vs. ERM

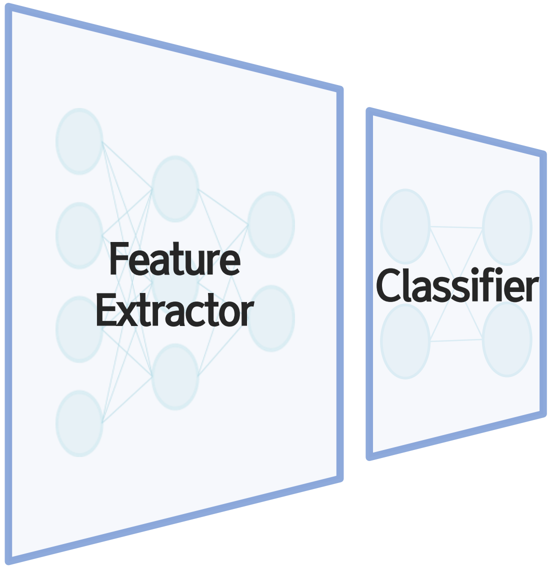
- ERM은 일반적인 모델 지도 학습에서 사용하는 목적식으로써 평균 성능을 최적화하는 방식
- GroupDRO는 그룹별 불균형 요소를 반영하여 worst-group 성능을 개선하는 방식

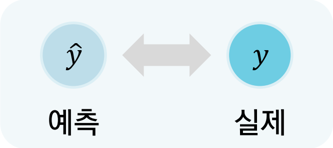
Group 1
Y : Dog
a : Outdoor

Group 3
Y : Cat
a : Outdoor

Group 2
Y : Dog
a : Indoor

Group 4
Y : Cat
a : Indoor





Group 1 Loss: 0.5

Group 2 Loss: 1.5

Group 3 Loss: 1.5

Group 4 Loss: 0.5

ERM loss

(0.5 + 1.5 + 1.5 + 0.5)/4
= 1.0

Minority group들에 대한 일반화 성능 보장 못함

Subpopulation Shift

Distributionally robust neural networks for group shifts – GroupDRO

❖ GroupDRO vs. ERM

- ERM은 일반적인 모델 지도 학습에서 사용하는 목적식으로써 평균 성능을 최적화하는 방식
- GroupDRO는 그룹별 불균형 요소를 반영하여 worst-group 성능을 개선하는 방식

Group weights = $[\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}]$

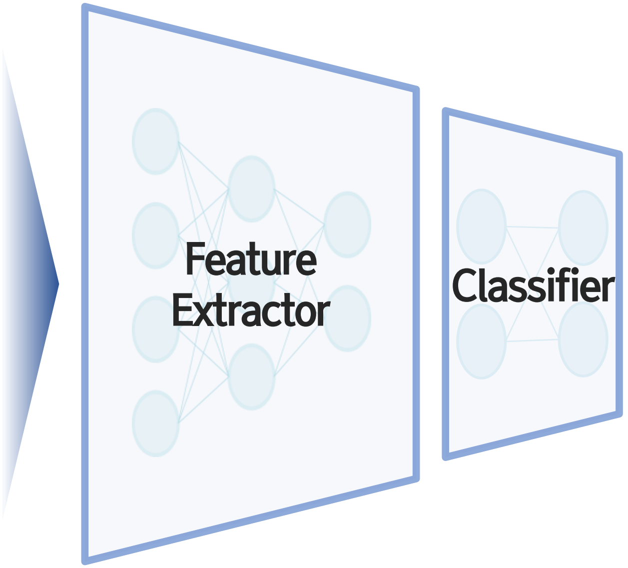
DRO step: 0.01 (hyperpara.)

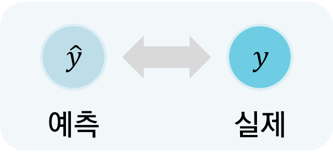
Group 1
Y : Dog
a : Outdoor

Group 3
Y : Cat
a : Outdoor

Group 2
Y : Dog
a : Indoor

Group 4
Y : Cat
a : Indoor





Group 1 Loss: 0.5

Group 2 Loss: 1.5

Group 3 Loss: 1.5

Group 4 Loss: 0.5

Subpopulation Shift

Distributionally robust neural networks for group shifts – GroupDRO

❖ GroupDRO vs. ERM

- ERM은 일반적인 모델 지도 학습에서 사용하는 목적식으로써 평균 성능을 최적화하는 방식
- GroupDRO는 그룹별 불균형 요소를 반영하여 worst-group 성능을 개선하는 방식

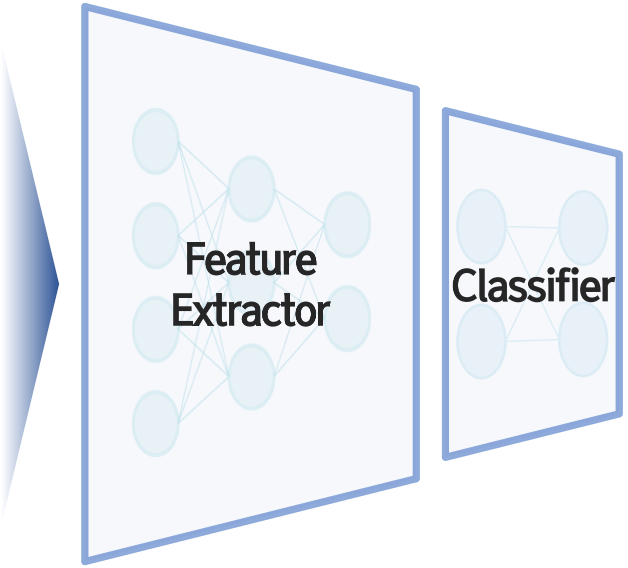
Group weights = $[\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}]$ DRO step: 0.01 (hyperpara.)

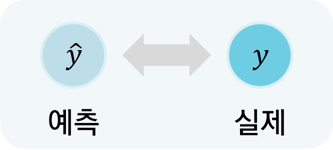
Group 1
Y : Dog
a : Outdoor

Group 3
Y : Cat
a : Outdoor

Group 2
Y : Dog
a : Indoor

Group 4
Y : Cat
a : Indoor





Group weight updates

$Group\ weight \times e^{(DRO\ step \times Group\ loss)}$

Group 1 Loss: 0.5	weight: $\frac{1}{4} \times e^{0.01 \times 0.5}$ =0.2512
Group 2 Loss: 1.5	weight: $\frac{1}{4} \times e^{0.01 \times 1.5}$ =0.2538
Group 3 Loss: 1.5	weight: $\frac{1}{4} \times e^{0.01 \times 1.5}$ =0.2538
Group 4 Loss: 0.5	weight: $\frac{1}{4} \times e^{0.01 \times 0.5}$ =0.2512

Normalization

Updated Group weights
= [0.2487, 0.2513, 0.2513, 0.2487]

Subpopulation Shift

Distributionally robust neural networks for group shifts – GroupDRO

❖ GroupDRO vs. ERM

- ERM은 일반적인 모델 지도 학습에서 사용하는 목적식으로써 평균 성능을 최적화하는 방식
- GroupDRO는 그룹별 불균형 요소를 반영하여 worst-group 성능을 개선하는 방식

Group weights = [0.2487, 0.2513, 0.2513, 0.2487]

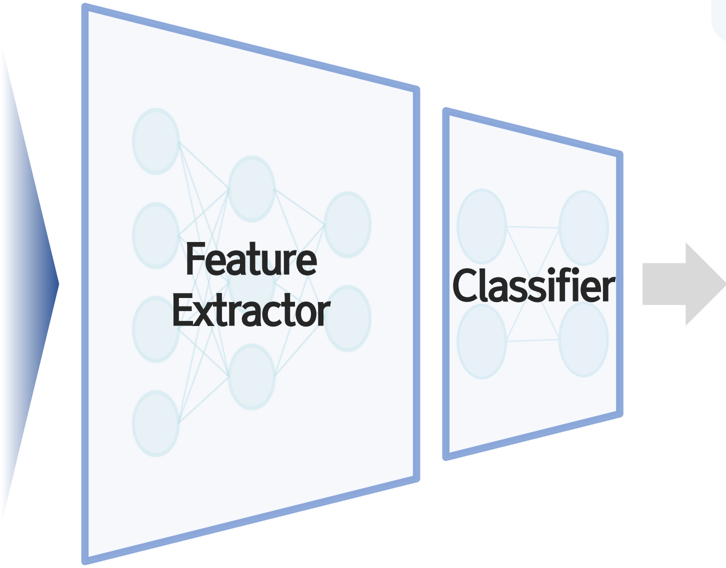
DRO step: 0.01 (hyperpara.)

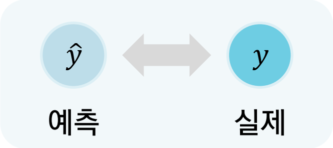
Group 1
Y : Dog
a : Outdoor

Group 3
Y : Cat
a : Outdoor

Group 2
Y : Dog
a : Indoor

Group 4
Y : Cat
a : Indoor





Group 1 Loss: 0.5

Group 2 Loss: 1.5

Group 3 Loss: 1.5

Group 4 Loss: 0.5

Group weighted losses

(0.5×0.2487)=0.12435

(1.5×0.2513)=0.37695

(1.5×0.2513)=0.37695

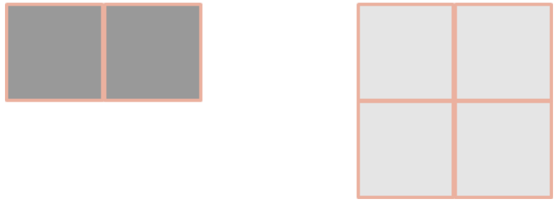
(0.5×0.2487)=0.12435

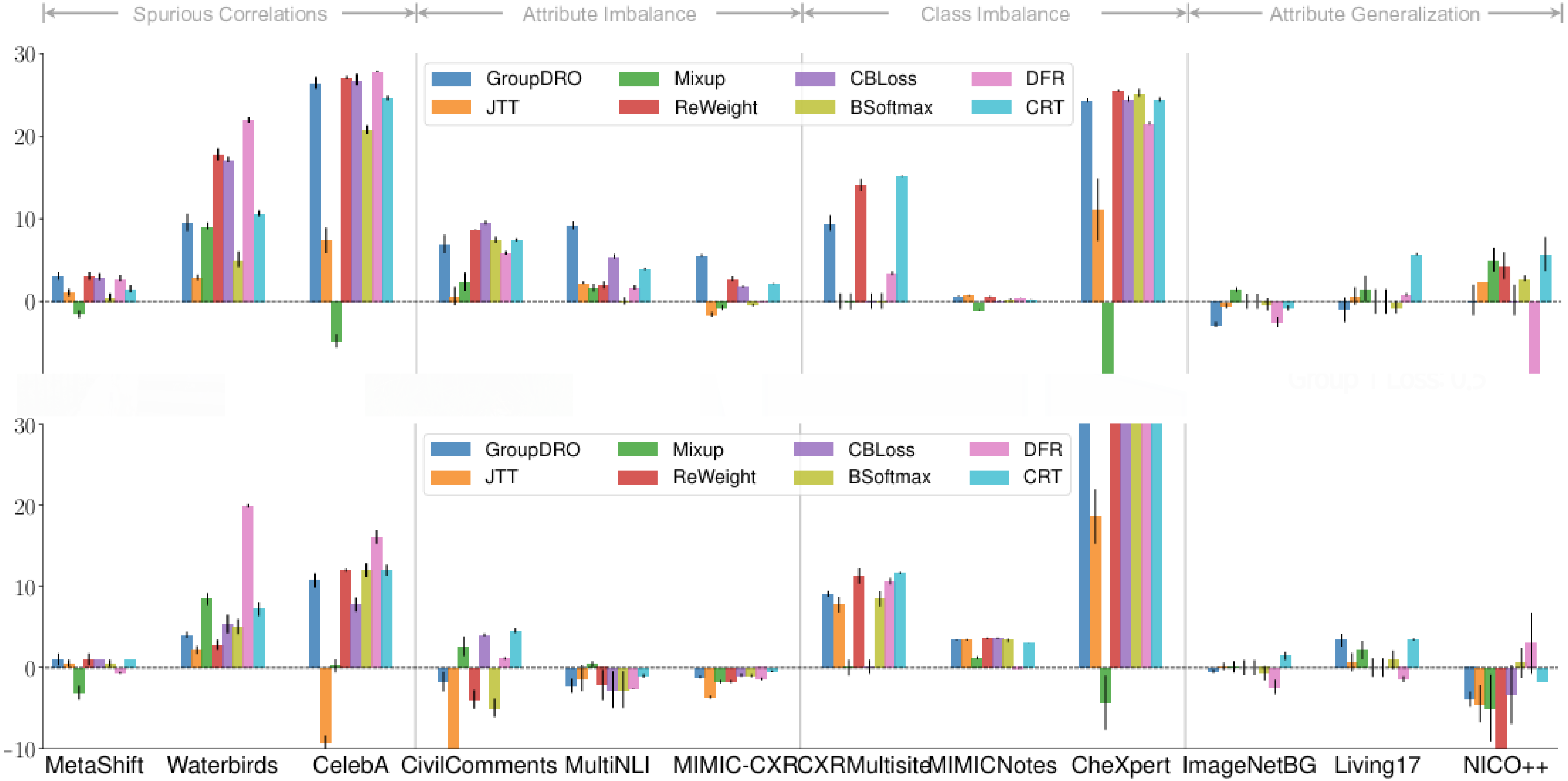
GroupDRO loss = 1.0026

Subpopulation Shift

Distributionally robust neural networks for group shifts – GroupDRO

Test dataset





Train & validation
Attributes both known

Train & validation
Attributes both unknown

Subpopulation Shift

2-stage method for subpopulation shift

❖ Last layer re-training is sufficient for robustness to spurious correlations – DFR (ICLR, 2023)

- New York university에서 연구 되었으며 2025년 4월 11일 기준 352회 인용됨
- Subpopulation shift 상황에서 모델 성능 저하를 간단한 classifier re-training 기법으로 개선할 수 있다는 것을 보임

Published as a conference paper at ICLR 2023

LAST LAYER RE-TRAINING IS SUFFICIENT FOR ROBUSTNESS TO SPURIOUS CORRELATIONS

Polina Kirichenko[*] New York University
Pavel Izmailov[*] New York University
Andrew Gordon Wilson New York University

ABSTRACT

Neural network classifiers can largely rely on simple spurious features, such as backgrounds, to make predictions. However, even in these cases, we show that they still often learn core features associated with the desired attributes of the data, contrary to recent findings. Inspired by this insight, we demonstrate that simple last layer retraining can match or outperform state-of-the-art approaches on spurious correlation benchmarks, but with profoundly lower complexity and computational expenses. Moreover, we show that last layer retraining on large ImageNet-trained models can also significantly reduce reliance on background and texture information, improving robustness to covariate shift, after only minutes of training on a single GPU.

1 INTRODUCTION

Realistic datasets in deep learning are riddled with *spurious correlations* — patterns that are predictive of the target in the train data, but that are irrelevant to the true labeling function. For example, most of the images labeled as "butterfly" on ImageNet also show flowers (Singla & Feizi, 2021), and most of the images labeled as "tench" show a fisherman holding the tench (Brendel & Bethge, 2019). Deep neural networks rely on these spurious features, and consequently degrade in performance when tested on datapoints where the spurious correlations break, for example, on images with unusual background contexts (Geirhos et al., 2020; Rosenfeld et al., 2018; Beery et al., 2018). In an especially alarming example, CNNs trained to recognize pneumonia were shown to rely on hospital-specific metal tokens in the chest X-ray scans, instead of features relevant to pneumonia (Zech et al., 2018).

Subpopulation Shift

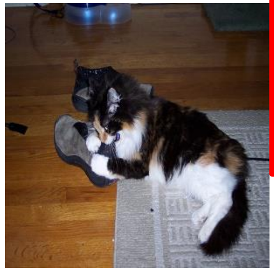
Last layer re-training is sufficient for robustness to spurious correlations – DFR

❖ Deep feature reweighting (DFR): 2-stage training method

- 첫번째 stage: ERM으로 학습(train dataset 활용)
- 두번째 stage: 특징 추출기 고정 및 분류기 재학습(resampled validation dataset 및 l_1 regularization 활용)

Train dataset

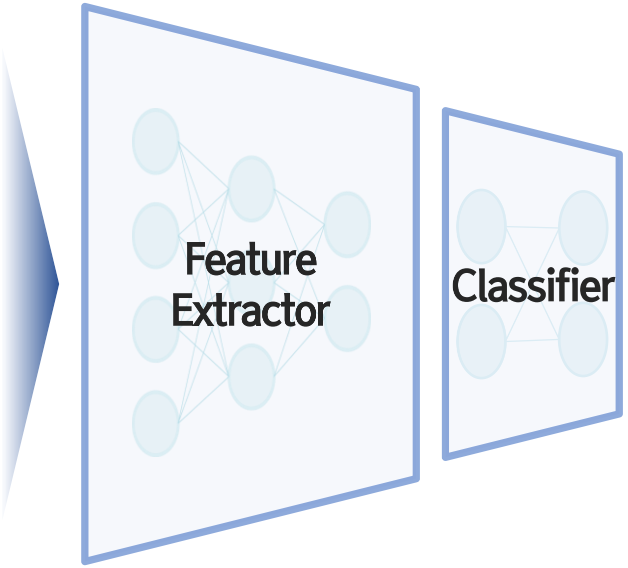
Group 1
Y : Dog
a : Outdoor
: 500

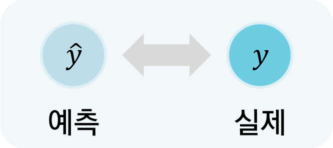


Group 3
Y : Cat
a : Outdoor
: 50

Group 2
Y : Dog
a : Indoor
: 50

Group 4
Y : Cat
a : Indoor
: 500





Group 1 Loss: 0.5

Group 2 Loss: 1.5

Group 3 Loss: 1.5

Group 4 Loss: 0.5

ERM loss

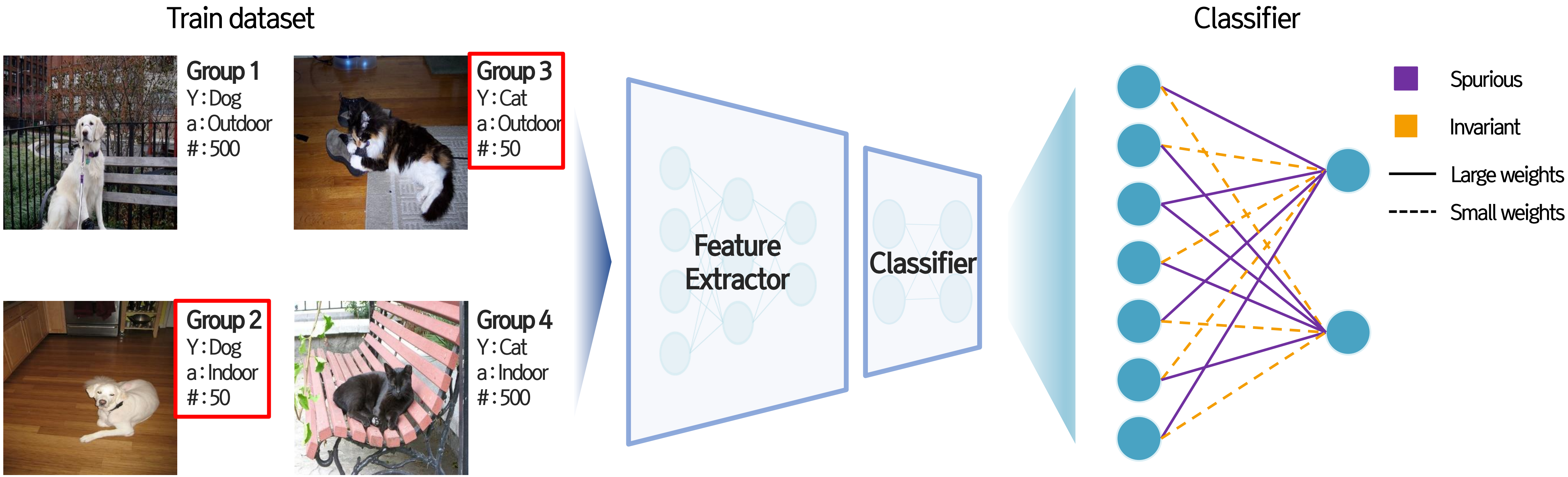
(0.5 + 1.5 + 1.5 + 0.5)/4
= 1.0

Subpopulation Shift

Last layer re-training is sufficient for robustness to spurious correlations – DFR

❖ Deep feature reweighting (DFR): 2-stage training method

- 첫번째 stage: ERM으로 학습(train dataset 활용)
- 두번째 stage: 특징 추출기 고정 및 분류기 재학습(resampled validation dataset 및 l_1 regularization 활용)

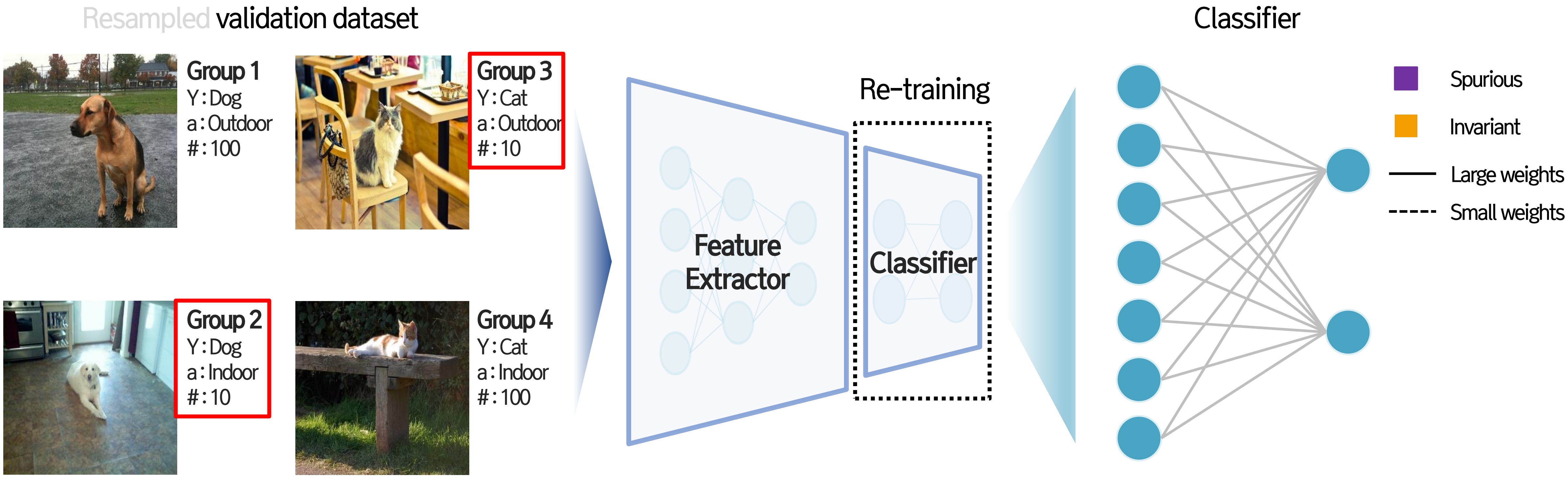


Subpopulation Shift

Last layer re-training is sufficient for robustness to spurious correlations – DFR

❖ Deep feature reweighting (DFR): 2-stage training method

- 첫번째 stage: ERM으로 학습(train dataset 활용)
- 두번째 stage: 특징 추출기 고정 및 분류기 재학습(resampled validation dataset 및 l_1 regularization 활용)

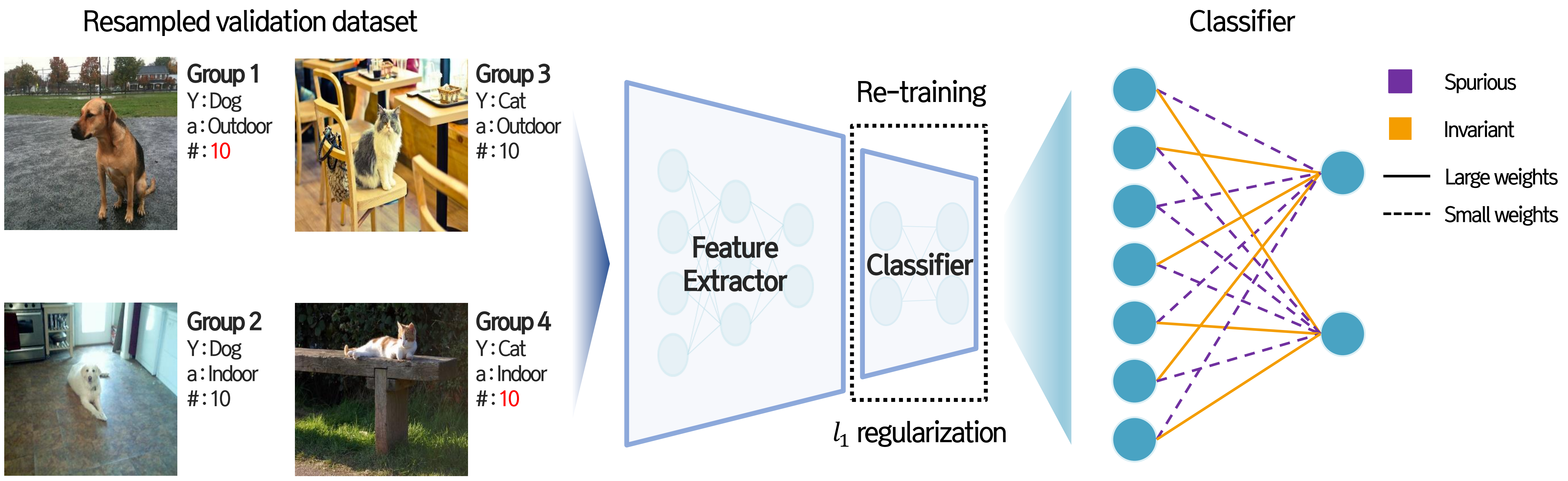


Subpopulation Shift

Last layer re-training is sufficient for robustness to spurious correlations – DFR

❖ Deep feature reweighting (DFR): 2-stage training method

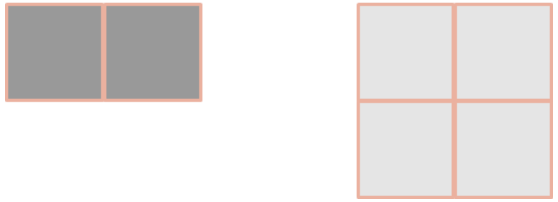
- 첫번째 stage: ERM으로 학습(train dataset 활용)
- 두번째 stage: 특징 추출기 고정 및 분류기 재학습(resampled validation dataset 및 l_1 regularization 활용)

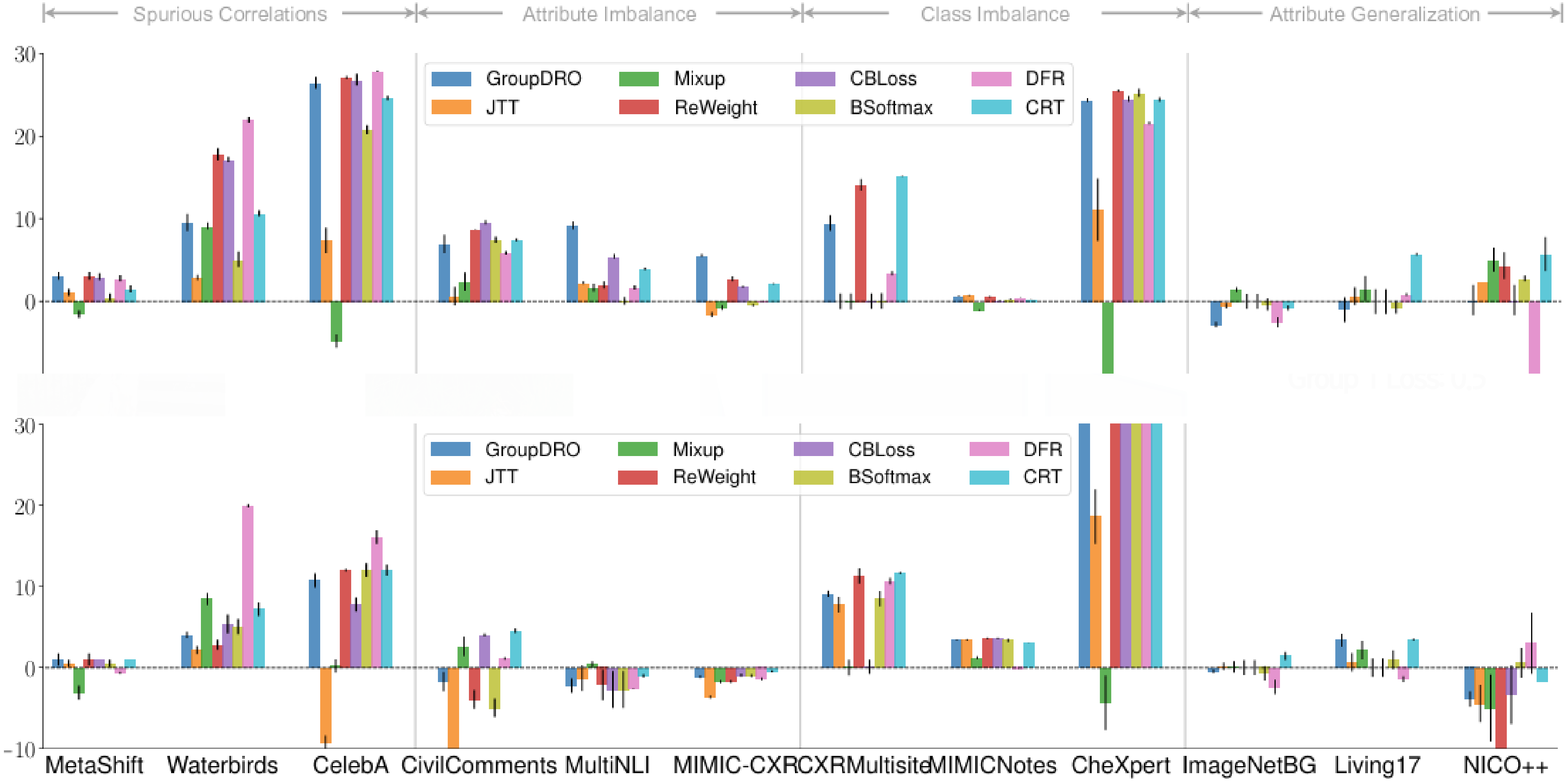


Subpopulation Shift

Last layer re-training is sufficient for robustness to spurious correlations – DFR

Test dataset





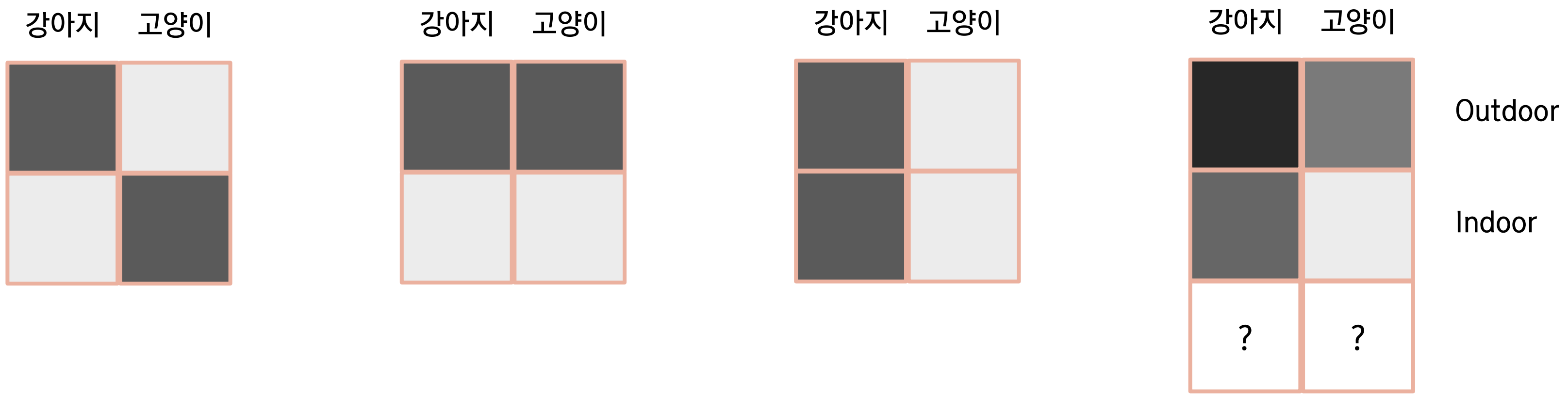
Train & validation
Attributes both known

Train & validation
Attributes both unknown

Conclusion

❖ Summary

- Basic types of Subpopulation shift
  - ➢ Spurious correlations, attribute imbalance, class imbalance, attribute generalization



- 1-stage method for subpopulation shift: GroupDRO
- 2-stage method for subpopulation shift: DFR

고맙습니다