Out-Of-Distribution Generalization :

Distributionally Robust Optimization 1

Data Mining & Quality Analytics Lab.

2024. 09. 13

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

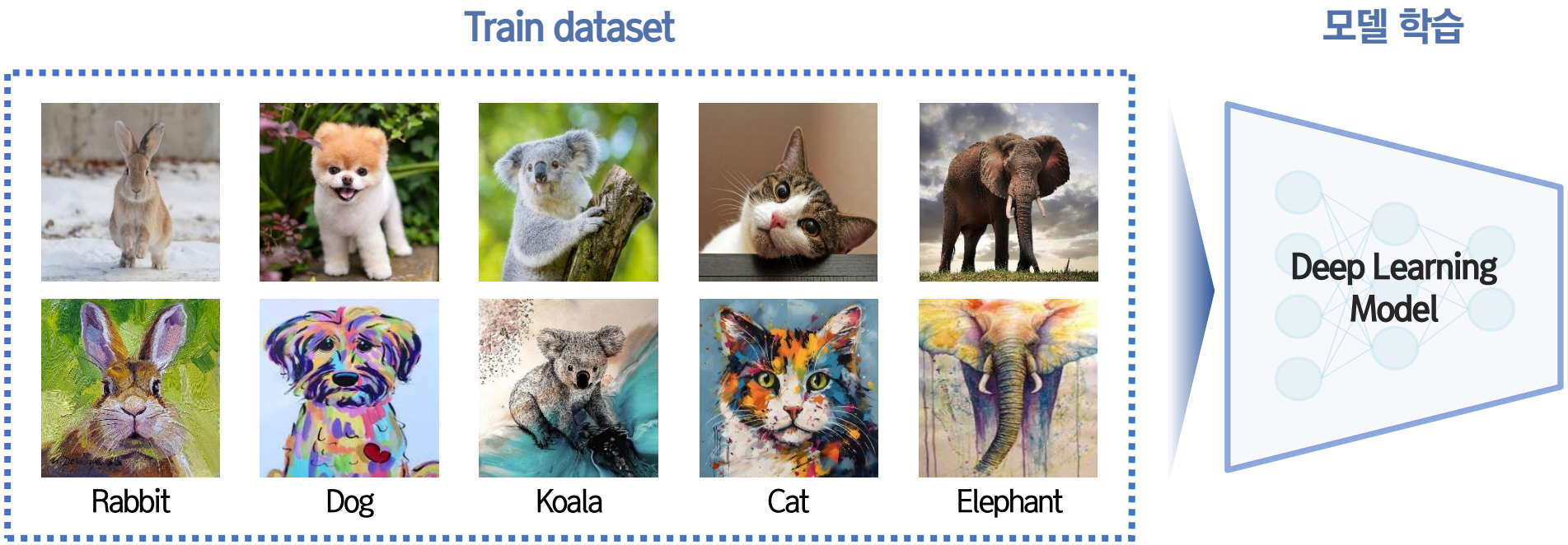
- jy_jeong@korea.ac.kr

목차

Introduction

Background of distribution shift

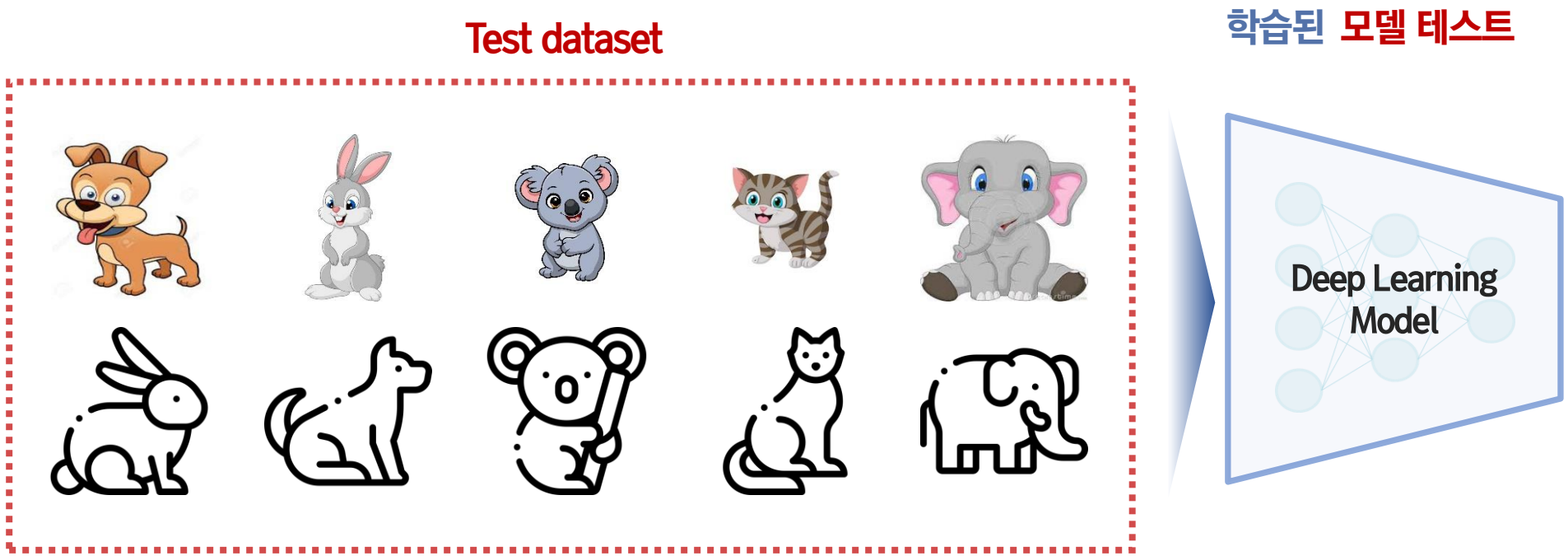
❖ Train dataset을 사용해서 일반화 성능이 좋은 모델을 구축해보자



Introduction

Background of distribution shift

❖ Train dataset을 사용해서 **일반화 성능**이 좋은 모델을 구축해보자

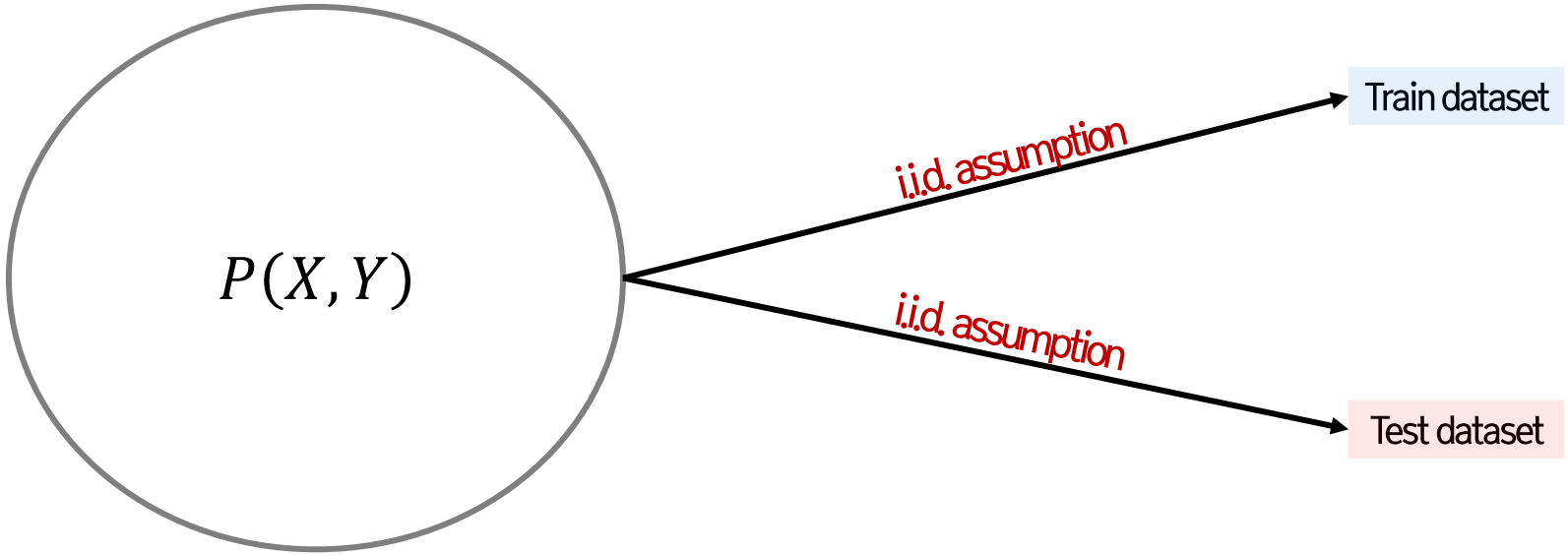


학습된 모델은 **낮은 일반화(generalization) 성능**을 보임

Introduction

Background of distribution shift

❖ Distribution shift: 학습 데이터의 분포와 테스트 데이터의 분포가 다른 경우를 의미함



'독립적이고 동일한 확률로'

기존 머신러닝 및 딥러닝 학습에서는 train dataset과 test dataset이 같은 분포에서 샘플링 되었다는 가정이 있음

→ 만약 i.i.d 가정이 깨진다면, 학습된 모델은 일반화 성능이 저하될 수 있음

'Distribution shift (Domain shift)'

Introduction

Background of distribution shift

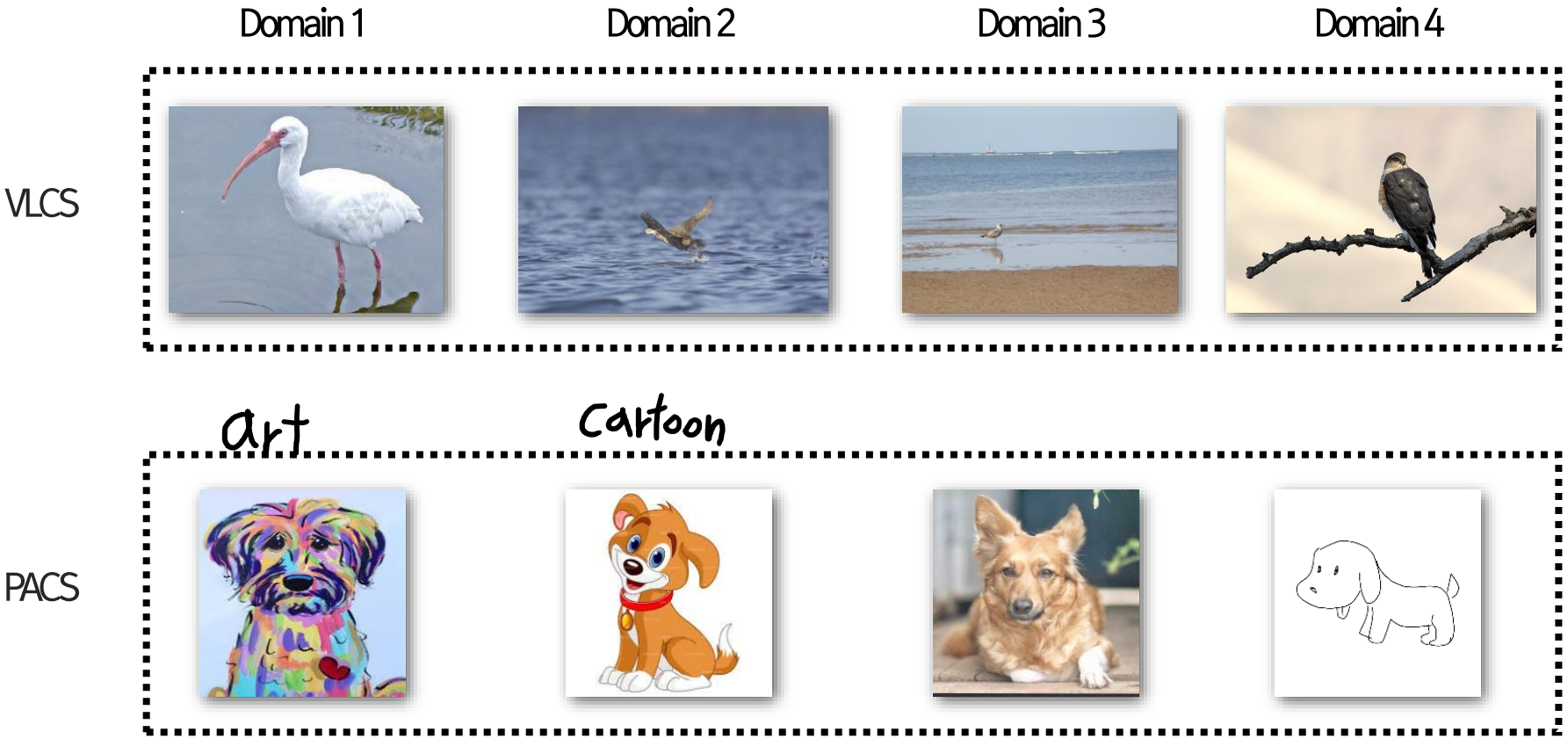
$$P(x,y) = P(y|x) \cdot P(x)$$

$$\Rightarrow \quad P_{art}(y|x) = P_{cartoon}(y|x)$$

$$P_{art}(x) \neq P_{cartoon}(x).$$

❖ Distribution shift는 domain shift도 포함하고 있으며 다양한 종류의 shift가 있음

- Covariate shift, group (subpopulation) shift, concept shift 등



Class
Bird
Car
Chair
Dog
Person

Dog
Person
Elephant
Guitar
House
Horse
Giraffe

Introduction

Background of distribution shift

❖ Group (subpopulation) shift

- 전체 데이터 분포 내의 하위 그룹들 간 분포 변화를 의미함

Y:금발

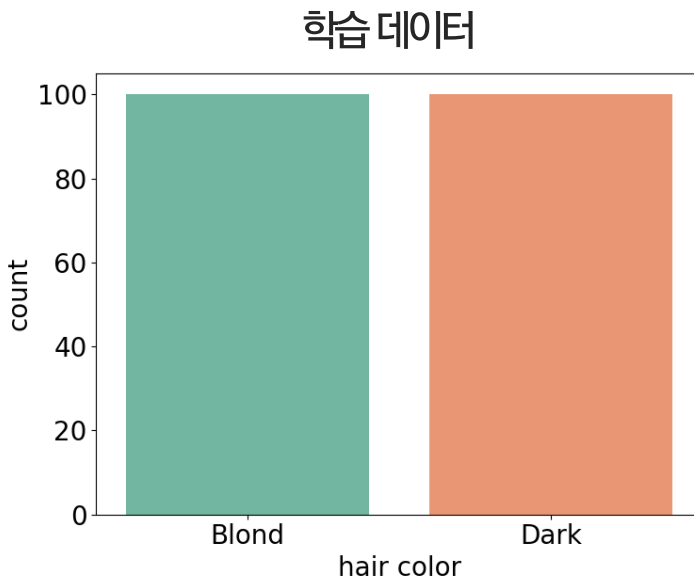
Y:흑발

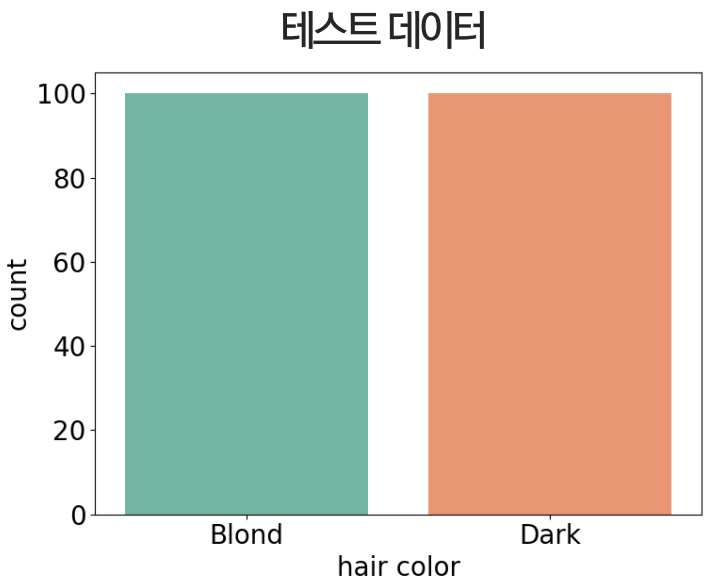
Y:금발

Y:흑발

CelebA dataset

Y 기준으로 데이터 수 균일





Introduction

Background of distribution shift

❖ Group (subpopulation) shift

- 전체 데이터 분포 내의 하위 그룹들 간 분포 변화를 의미함

Y:금발
a:여성

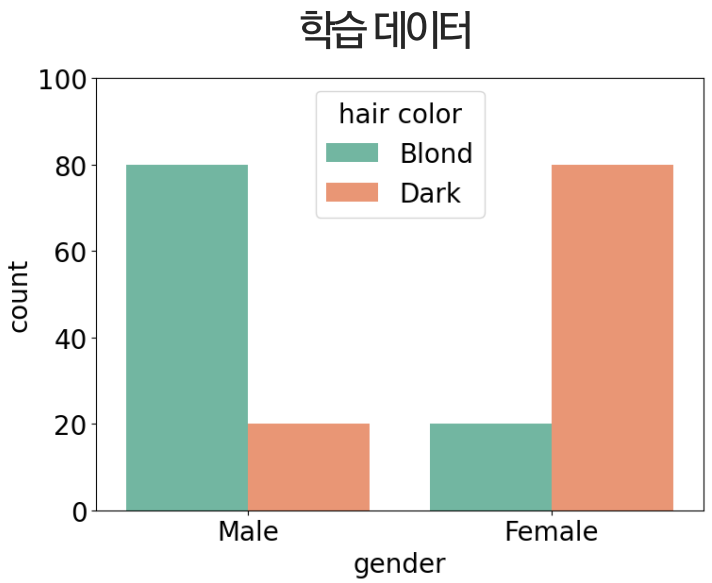
Y:흑발
a:여성

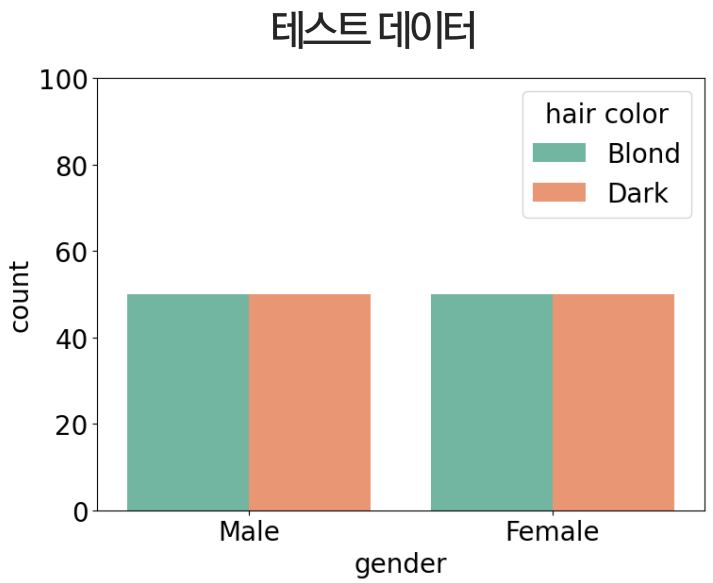
Y:금발
a:남성

Y:흑발
a:남성

CelebA dataset

학습 데이터는 하위 집합(그룹) 기준으로 데이터 수 균일하지 않음





Introduction

Background of distribution shift

❖ Group (subpopulation) shift

- 전체 데이터 분포 내의 하위 그룹들 간 분포 변화를 의미함

Y:금발
a:여성

Y:흑발
a:여성

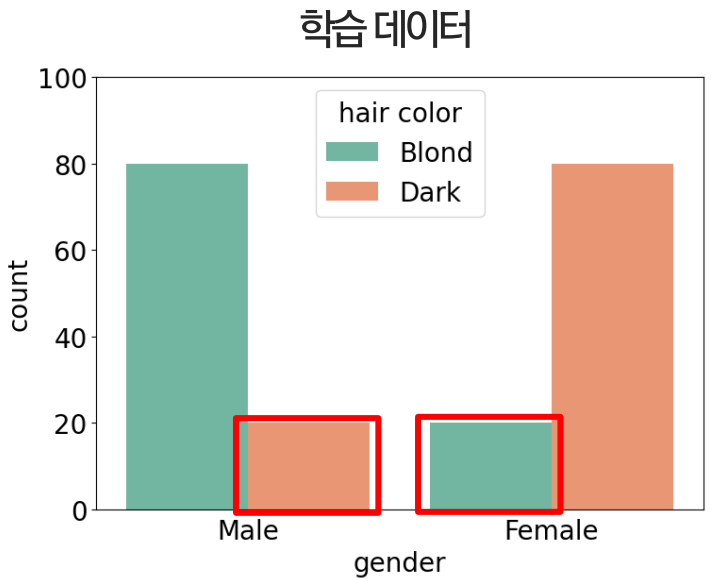
: Minority group(s)

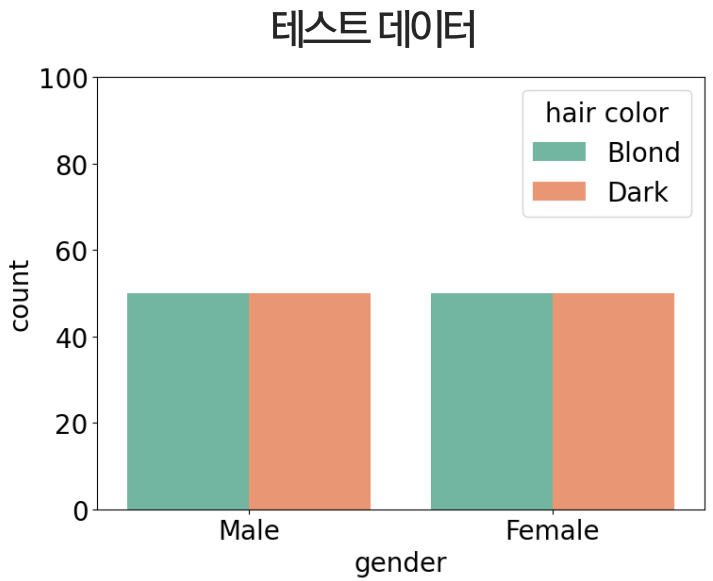
Y:금발
a:남성

Y:흑발
a:남성

CelebA dataset

학습 데이터는 하위 집합(그룹) 기준으로 데이터 수 균일하지 않음



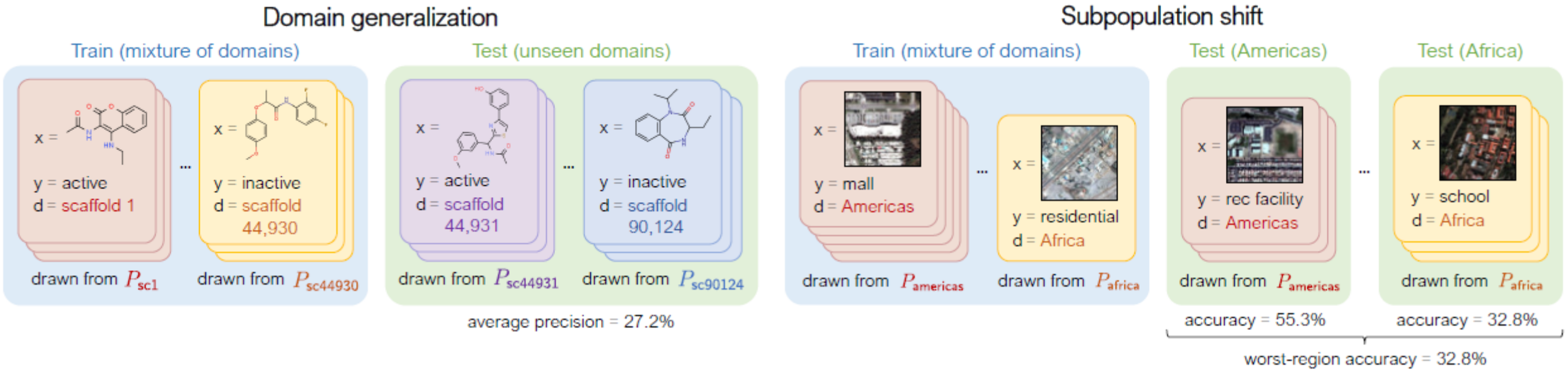


Introduction

Background of distribution shift

❖ Out-of-distribution generalization

- 다양한 종류의 distribution shift 상황에서도 좋은 일반화 성능을 보이기 위한 연구 분야
- Domain shift 상황을 해결하는 domain generalization 개념 포함



[Wilds benchmark datasets]

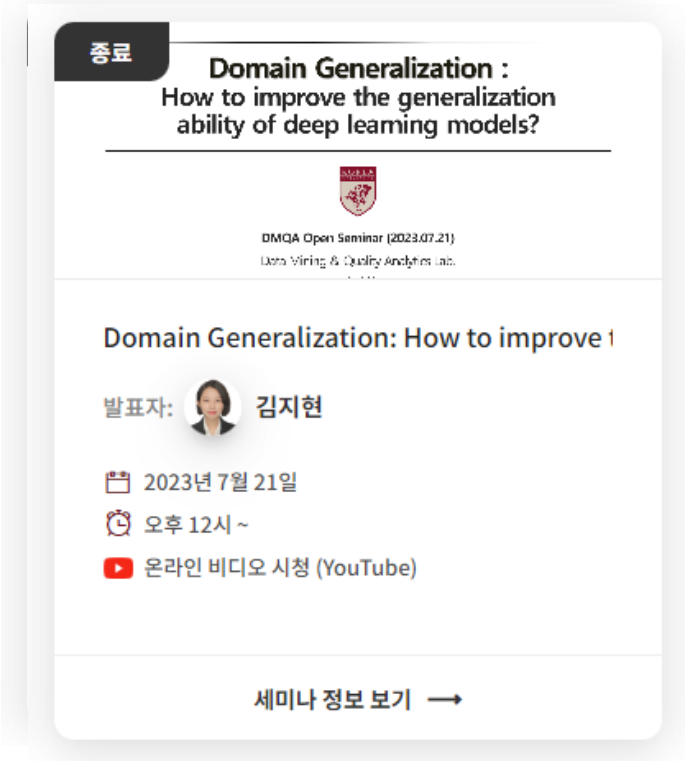
Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., ... & Liang, P. (2021, July). Wilds: A benchmark of in-the-wild distribution shifts. In International conference on machine learning (pp. 5637-5664). PMLR.

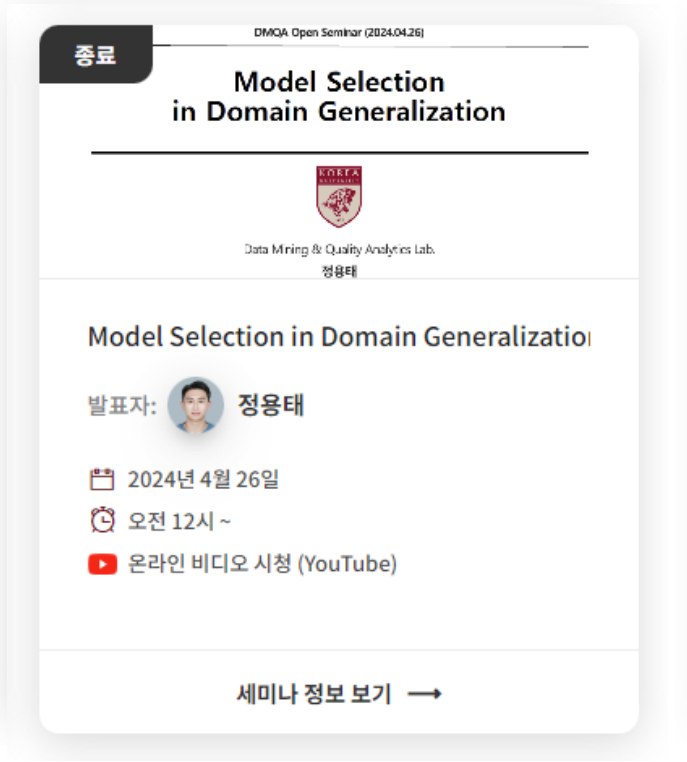
Introduction

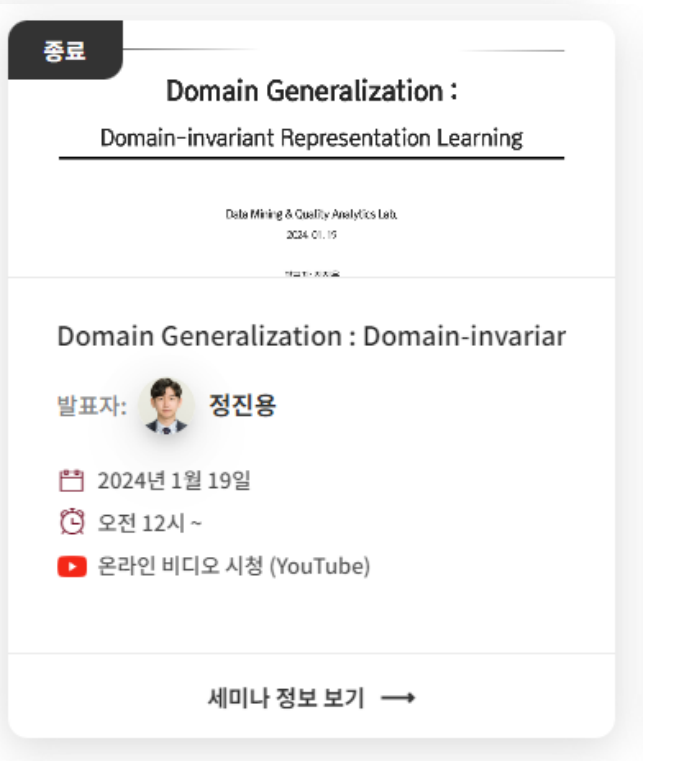
Background of distribution shift

❖ Distribution shift 상황에서도 모델 일반화 성능을 향상 시킬 수 있는 다양한 연구들이 존재

- Test data에 대한 성능을 높이자!







Distributionally robust optimization

Distributionally robust neural networks for group shifts

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

- Stanford, Microsoft 연구원들에 의해 연구되었으며 2024년 9월 13일 기준 1,595회 인용됨
- Distribution shift 중 group (subpopulation) 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.

Distributionally robust optimization

Distributionally robust neural networks for group shifts

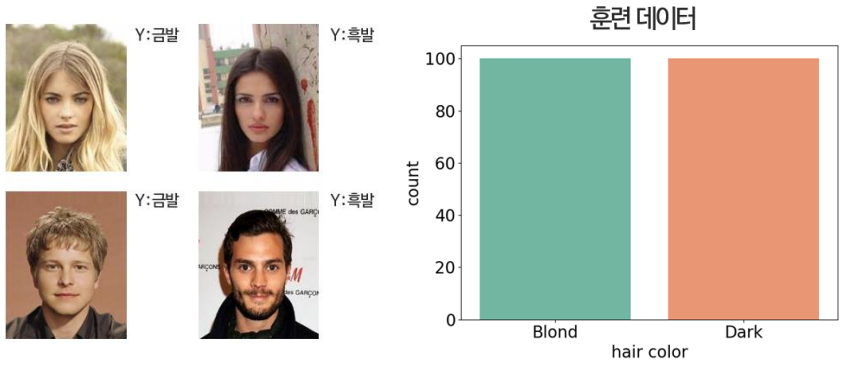
❖ Distributionally robust optimization (DRO) vs. Empirical risk minimization (ERM)

- ERM은 일반적인 모델 지도 학습에서 사용하는 목적식으로써 평균 성능을 최적화하는 방식
- DRO는 불확실성 하에서 worst-case performance를 최적화하는 접근 방식

[ERM]

$$\hat{\theta}_{ERM} := \underset{\theta \in \Theta}{arg\min}\, \mathbb{E}_{(x,y)\sim \hat{P}}[\ell(\theta;(x,y))]$$

훈련 데이터의 empirical distribution



Loss의 기댓값 (Risk)이 최소가 되는 모델

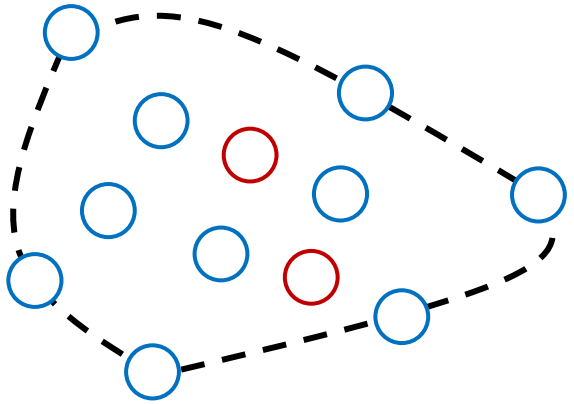
Minority group들에 대한 일반화 성능 보장 못함

[DRO]

$$\min_{\theta \in \Theta}\{\mathcal{R}(\theta) := sup_{Q \in \mathcal{Q}}\, \mathbb{E}_{(x,y)\sim Q}[\ell(\theta;(x,y))]\}$$

Uncertainty set of distributions $\mathcal{Q}$

실제 가능한 테스트 분포들을 포함하는 집합



Distributionally robust optimization

Distributionally robust neural networks for group shifts

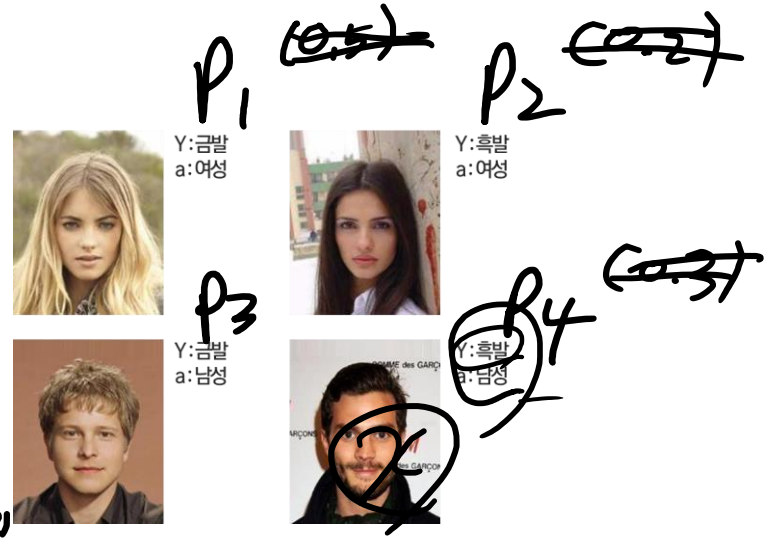
[DRO]

$$\min_{\theta\in\Theta}\{\mathcal{R}(\theta) := sup_{Q\in\mathcal{Q}}\, \mathbb{E}_{(x,y)\sim Q}[\ell(\theta;(x,y))]\}$$

Uncertainty set of distributions $\mathcal{Q}$

❖ Group 수준에서의 uncertainty set $\mathcal{Q}$ 구축 및 Group DRO

- 훈련 데이터 분포가 여러 하위 분포 혼합으로 구성되어 있다고 가정함
- Q는 훈련 데이터의 그룹 분포들의 모든 가능한 혼합(mixture)을 포함하는 집합
- 모델이 group shift에 강건할 수 있도록 훈련시키는 것을 목표로 함



[Group DRO]

$$\mathcal{Q} := \{\sum_{g=1}^{m} q_g P_g : q \in \Delta_m\}$$

확률 분포를 나타내는 벡터의 집합.
(ex) 주사위 $(\frac{1}{6}, \frac{1}{6}, \cdots \frac{1}{6})$ 등

i) 모든 요소 비음수
ii) 모든 요소 합 1
iii) n차원에서 n+1개 꼭지점 가짐.
⇒ convex set

$$\mathcal{R}(\theta) = \max_{g\in\mathcal{G}} \mathbb{E}_{(x,y)\sim P_g}[\ell(\theta;(x,y))]$$

Worst-group(case) risk 정의가 각 그룹의 risk 중 최대값과 동일함

$$\hat{\theta}_{DRO} := arg\min_{\theta\in\Theta}\{\hat{\mathcal{R}}(\theta) := \max_{g\in\mathcal{G}} \mathbb{E}_{(x,y)\sim \hat{P}_g}[\ell(\theta;(x,y))]\}$$

Worst-group generalization gap $\delta := \mathcal{R}(\theta) - \hat{\mathcal{R}}(\theta)$

test성능 train성능.

(x, y, g)

Distributionally robust optimization

Distributionally robust neural networks for group shifts

pre-trained ResNet50

pre-trained BERT

❖ Group DRO와 ERM 비교 실험 결과

- 그룹 정보를 활용할 수 있는 벤치마크 데이터 셋 3 가지에 비교 실험 진행 : Waterbirds dataset, CelebA dataset, MultiNLI dataset
- Overparameterized neural networks에서는 학습 과적합을 피하기 위해 적절한 정규화 기법 적용이 필요함
- 적절한 정규화 기법이 적용되었을 경우, ERM보다 group DRO가 그룹별로 편향이 없는 일반화 성능을 보임

batch norm + l_2 lambda 0.001

1.0

0.1

300→1

50→1

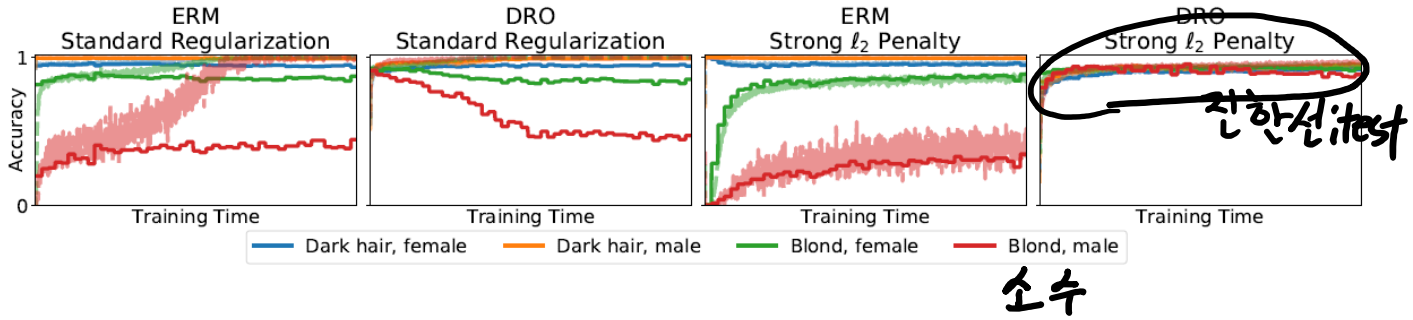
20→3

			Average Accuracy		Worst-Group Accuracy	
			ERM	DRO	ERM	DRO
Standard Regularization	Waterbirds	Train	100.0	100.0	100.0	100.0
		Test	97.3	97.4	60.0	76.9
	CelebA	Train	100.0	100.0	99.9	100.0
		Test	94.8	94.7	41.1	41.1
	MultiNLI	Train	99.9	99.3	99.9	99.0
		Test	82.5	82.0	65.7	66.4
Strong ℓ_2 Penalty	Waterbirds	Train	97.6	99.1	35.7	97.5
		Test	95.7	96.6	21.3	84.6
	CelebA	Train	95.7	95.0	40.4	93.4
		Test	95.8	93.5	37.8	86.7
Early Stopping	Waterbirds	Train	86.2	80.1	7.1	74.2
		Test	93.8	93.2	6.7	86.0
	CelebA	Train	91.3	87.5	14.2	85.1
		Test	94.6	91.8	25.0	88.3
	MultiNLI	Train	91.5	86.1	78.6	83.3
		Test	82.8	81.4	66.0	77.7

[비교 실험 성능 결과표]

연한 선 : train

[CelebA dataset에서 Standard Regularization과 Strong ℓ_2 penalty 비교 분석]



연한 선 : test

소수

✓ Standard regularization적용 시, 소수 그룹도 학습 과적합

✓ 적절한 정규화 기법 적용 시, DRO가 그룹별 편향이 없는 성능 보임

Distributionally robust optimization

Distributionally robust neural networks for group shifts

Worst-group generalization gap $\delta := \mathcal{R}(\theta) - \hat{\mathcal{R}}(\theta)$

❖ Group adjustments DRO

- Strong ℓ_2 penalty 등 적절한 정규화 기법을 사용하더라도 그룹에 직접적으로 반영하지는 못함
- 그룹별로 불균형 요소를 반영하여 worst-group performance를 개선함

[그룹별 일반화 격차 δ_g를 불균형 요소로써 목적식에 반영]

Each group generalization gap $\delta_g = \mathbb{E}_{(x,y)\sim P_g}\left[\ell\left(\theta;(x,y)\right)\right] - \mathbb{E}_{(x,y)\sim \hat{P}_g}\left[\ell\left(\theta;(x,y)\right)\right]$

Worst-group test loss $\mathcal{R}(\theta) = \max_{g\in\mathcal{G}} \mathbb{E}_{(x,y)\sim P_g}\left[\ell\left(\theta;(x,y)\right)\right] = \max_{g\in\mathcal{G}} \mathbb{E}_{(x,y)\sim \hat{P}_g}\left[\ell\left(\theta;(x,y)\right)\right] + \delta_g$

$$\hat{\theta}_{adj} := arg\min_{\theta\in\Theta} \max_{g\in\mathcal{G}} \{\mathbb{E}_{(x,y)\sim \hat{P}_g}\left[\ell\left(\theta;(x,y)\right)\right] + \frac{C}{\sqrt{n_g}}\} \qquad (\because \delta_g \rightarrow \frac{C}{\sqrt{n_g}})$$

Cao, K., Wei, C., Gaidon, A., Arechiga, N., & Ma, T. (2019). Learning imbalanced datasets with label-distribution-aware margin loss. Advances in neural information processing systems, 32.

Distributionally robust optimization

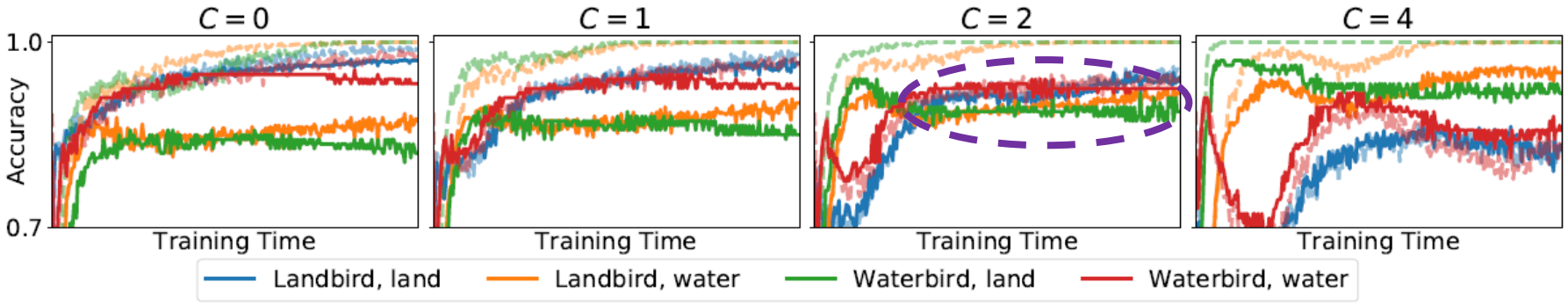
Distributionally robust neural networks for group shifts

❖ Group adjustments DRO

- Strong ℓ_2 penalty 등 적절한 정규화 기법을 사용하더라도 그룹별로 일반화 격차는 달라질 수 있음
- 그룹별 불균형 요소를 반영하여 worst-group performance를 개선함

	Average Accuracy		Worst-Group Accuracy	
	Naïve	Adjusted	Naïve	Adjusted
Waterbirds	96.6	93.7	84.6	90.5
CelebA	93.5	93.4	86.7	87.8

[Group DRO와 Group adjusted DRO 성능 비교 결과표]



[Waterbirds dataset에서 adjust factor C 비교 분석 결과]

Cao, K., Wei, C., Gaidon, A., Arechiga, N., & Ma, T. (2019). Learning imbalanced datasets with label-distribution-aware margin loss. Advances in neural information processing systems, 32.

Distributionally robust optimization

Distributionally robust neural networks for group shifts

❖ Group DRO와 importance weighting 비교 – (1) 실험적 비교

- Importance weighting은 학습 데이터와 테스트 데이터 분포가 다를 때 사용하는 기법
- 따라서, Group DRO의 baseline이 될 수 있음
- 학습 데이터의 분포를 테스트 데이터의 분포에 가깝게 조정하는 것이 목적

[Importance weighting]

$$\hat{\theta}_w := arg\min_{\theta \in \Theta} \mathbb{E}_{(x,y,g)\sim \hat{P}_g}[w_g \ell(\theta;(x,y))]$$

$$w_g = P_{test}(g)/P_{train}(g)$$

$$w_g = 1/\mathbb{E}_{g'\sim \hat{P}_g}[\mathbb{I}(g'=g)]$$

전체 학습 데이터 중 해당 그룹의 개수

	Average Accuracy			Worst-Group Accuracy		
	ERM	UW	DRO	ERM	UW	DRO
Waterbids	**97.0 (0.2)**	95.1 (0.3)	93.5 (0.3)	63.7 (1.9)	88.0 (1.3)	**91.4 (1.1)**
CelebA	**94.9 (0.2)**	92.9 (0.2)	92.9 (0.2)	47.8 (3.7)	83.3 (2.8)	**88.9 (2.3)**
MultiNLI	**82.8 (0.1)**	81.2 (0.1)	81.4 (0.1)	66.4 (1.6)	64.8 (1.6)	**77.7 (1.4)**

[ERM, Upweighting, Group DRO 성능 비교 결과표]

K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In Advances in Neural Information Processing Systems (NeurIPS), 2019.
Y. Cui, M. Jia, T. Lin, Y. Song, and S. Belongie. Class-balanced loss based on effective number of samples. In Computer Vision and Pattern Recognition (CVPR), pp. 9268–9277, 2019.

Distributionally robust optimization

Distributionally robust neural networks for group shifts

❖ Group DRO와 importance weighting 비교 – (2) 이론적 비교

- Convex 상황에서는 두 방법론이 동등할 수 있음 → appendix 증명
- Non-convex 상황에서는 간단한 반례를 통해서 동등하지 않은 것을 보임

목적식 방향성

∴ DRO → worst-case

importance weighting → 그룹 평균 loss

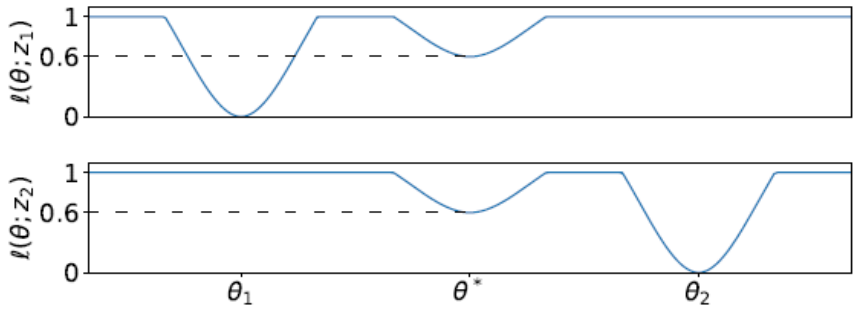


Figure 4: Toy example illustrating that DRO and importance weighting are not equivalent. The DRO solution is θ^*, while any importance weighting would result in solutions at θ_1 or θ_2.

Counterexample 1. *Consider a uniform data distribution P supported on two points $\mathcal{Z} = \{z_1, z_2\}$, and let $\ell(\theta; z)$ be as in Figure 4, with $\Theta = [0, 1]$. The DRO solution θ^* achieves a worst-case loss of $\mathcal{R}(\theta^*) = 0.6$. Now consider any weights $(w_1, w_2) \in \Delta_2$ and w.l.o.g. let $w_1 \geq w_2$. The minimizer of the weighted loss $w_1\ell(\theta; z_1) + w_2\ell(\theta; z_2)$ is θ_1, which only attains a worst-case loss of $\mathcal{R}(\theta^*) = 1.0$.*

Distributionally robust optimization

Distributionally robust neural networks for group shifts

❖ 실제 배치 연산 딥러닝 학습을 하기 위한 online optimization algorithm 제시

- 일반적인 non-convex 상황에서 단순 DRO 개념을 최적화한다면 크게 2가지 방식이 존재
  1. Stochastic gradient descent (SGD) on the Lagrangian dual of the objective → Group DRO에서 불편 추정량 구하기 힘듬
  2. Direct minimax optimization → 수렴성이 보장 안되어 있음
- 2번 개념을 제안한 알고리즘 형태로 수행한다면 worst-group대신 그래디언트를 사용하므로 수렴성을 보장할 수 있음

Group DRO : $\min_{\theta} \max_{g} E_{(x,y)\sim\hat{P}_g}[l(\theta;(x,y))]$ ⇒ 최대값을 가진 그룹의 gradient ⇒ 각 iter에서 worst-group선택

$\Rightarrow \min_{\theta,t} t$

$s.t.\ E_{(x,y)\sim\hat{P}_g}[l(\theta;(x,y))] \le t,\ \forall g$

Lagrangian ⇒ $L(\theta,t,\lambda) = t + \sum_g \lambda_g (E_{(x,y)\sim\hat{P}_g}[l(\theta;(x,y))] - t),\ \lambda_g \ge 0, \sum_g \lambda_g = 1$

KKT, duality ⇒ $\min_{\theta,\beta} \frac{1}{\alpha} E_g[\max(0, E_{x,y\sim\hat{P}_g}[l(\theta;(x,y)) | g] - \beta)] + \beta$ ⇒ 모든 그룹의 gradient ⇒ 모든 그룹 동시 샘플링

Distributionally robust optimization

Distributionally robust neural networks for group shifts

❖ 실제 배치 연산 딥러닝 학습을 하기 위한 online optimization algorithm 제시

- 일반적인 non-convex 상황에서 단순 DRO 개념을 최적화한다면 크게 2가지 방식이 존재
  1. Stochastic gradient descent (SGD) on the Lagrangian dual of the objective → Group DRO에서 불편 추정량 구하기 힘듬
  2. Direct minimax optimization → Group DRO에 적용한 사례가 있지만 수렴성이 보장 안되어 있음
- 2번 개념을 제안한 알고리즘 형태로 수행한다면 직접적인 worst-group 선별 대신 그래디언트를 사용하므로 수렴성 보장 가능

Algorithm 1: Online optimization algorithm for group DRO

Input: Step sizes η_q, η_θ; P_g for each $g \in \mathcal{G}$

Initialize $\theta^{(0)}$ and $q^{(0)}$

for $t = 1, \ldots, T$ **do**

- $g \sim \text{Uniform}(1, \ldots, m)$ — // Choose a group g at random
- $x, y \sim P_g$ — // Sample x, y from group g
- $q' \leftarrow q^{(t-1)}$; $q'_g \leftarrow q'_g \exp(\eta_q \ell(\theta^{(t-1)}; (x, y)))$ — // Update weights for group g
- $q^{(t)} \leftarrow q' / \sum_{g'} q'_{g'}$ — // Renormalize q
- $\theta^{(t)} \leftarrow \theta^{(t-1)} - \eta_\theta q_g^{(t)} \nabla \ell(\theta^{(t-1)}; (x, y))$ — // Use q to update θ

end

Conclusion

❖ Summary

- Distribution shift: 학습 데이터의 분포와 테스트 데이터의 분포가 다른 경우를 의미함
- Distribution shift 중 group shift 상황에서도 모델 일반화 및 강건성 성능을 개선할 수 있는 OOD generalization 소개
  - ➢ Worst-group 성능에 초점을 맞추는 group DRO 개념으로 worst-group accuracy 개선
  - ➢ 그래디언트 기반 최적화 알고리즘을 제안하여 수렴성을 보장할 수 있게 됨
- Group shift를 해결하는 방법론이지만 실제 domain generalization 방법론으로 활용 가능
  - ➢ 각 domain을 하나의 group으로 대체해서 적용함

Appendix

❖ Equivalence of DRO and Importance weighting in the convex setting

Proposition 1. *Suppose that the loss $\ell(\cdot; z)$ is continuous and convex for all z in $\mathcal{Z}$, and let the uncertainty set $\mathcal{Q}$ be a set of distributions supported on $\mathcal{Z}$. Assume that $\mathcal{Q}$ and the model family $\Theta \subseteq \mathbb{R}^d$ are convex and compact, and let $\theta^* \in \Theta$ be a minimizer of the worst-group objective $\mathcal{R}(\theta)$. Then there exists a distribution $Q^* \in \mathcal{Q}$ such that $\theta^* \in \arg\min_\theta \mathbb{E}_{z\sim Q^*}[\ell(\theta; z)]$.*

Proof. Let $h(\theta, Q) := \mathbb{E}_{z\sim Q}[\ell(\theta; z)]$. Since the loss $\ell(\theta; z)$ is continuous and convex in θ for all z in $\mathcal{Z}$, we have that $h(\theta, Q)$ is continuous, convex in θ, and concave (linear) in Q. Moreover, since convexity and lower semi-continuity are preserved under arbitrary pointwise suprema, $\sup_{Q\in\mathcal{Q}} h(\theta, Q)$ is also convex and lower semi-continuous (therefore proper).

해당 조건들로 data point별로 연산
⇒ Θ에 대해 독립적으로 최댓값을 취함
⇒ Θ에 대한 함수의 성질이 유지.

Appendix

❖ Equivalence of DRO and Importance weighting in the convex setting

$\sup_{Q \in \mathcal{Q}} h(\theta, Q)$가 convex and lower semi-continuous 인 경우

$\Rightarrow$ $h(\theta, Q)$가 θ에 대해서 convex

임의의 θ_1, θ_2, $0 \le \lambda \le 1$ 에 대하여

$$\sup_{Q \in \mathcal{Q}} h(\lambda\theta_1 + (1-\lambda)\theta_2, Q) \le \sup_{Q \in \mathcal{Q}} [\lambda h(\theta_1, Q) + (1-\lambda) h(\theta_2, Q)]$$

h가 θ에 대해서 convex.

sup의 성질

$$\le \lambda \sup_{Q \in \mathcal{Q}} h(\theta_1, Q) + (1-\lambda) \sup_{Q \in \mathcal{Q}} h(\theta_2, Q)$$

$\therefore$ $\sup_{Q \in \mathcal{Q}} h(\theta, Q)$ is convex.

Appendix

❖ Equivalence of DRO and Importance weighting in the convex setting

각 Q에서 $h(\theta, Q)$가 θ에 대해 lower semi-continuity 라고 하면,

임의의 실수 α에 대해, set $S_\alpha = \{\theta : \sup_{Q \in \mathcal{Q}} h(\theta, Q) > \alpha\}$

임의의 $\theta' \in S_\alpha$, 어떤 $Q' \in \mathcal{Q}$가 존재하여

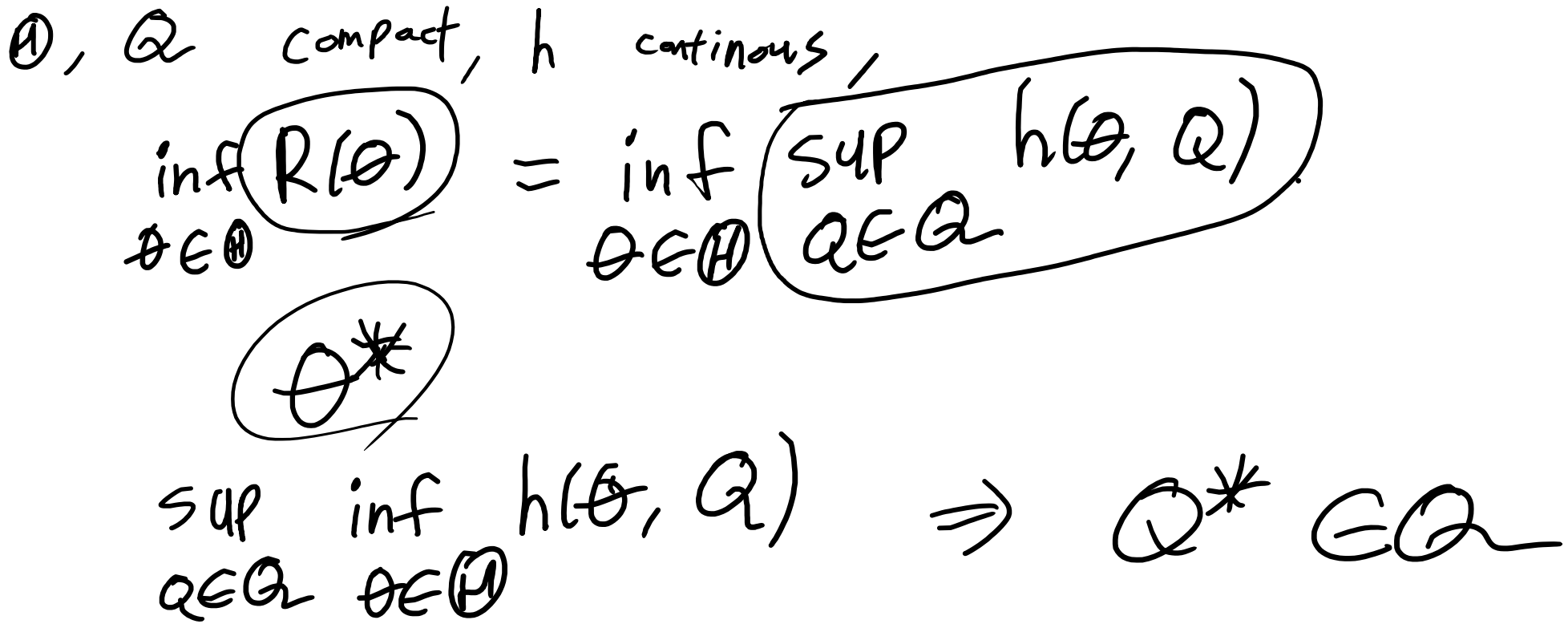
$h(\theta', Q') > \alpha$

$\therefore U \subset S_\alpha$ 이며, S_α가 열린 집합

$\therefore S_\alpha$가 모든 α에 대해 열린 집합, $\sup_{Q \in \mathcal{Q}} h(\theta, Q)$

Appendix

❖ Equivalence of DRO and Importance weighting in the convex setting



Appendix

❖ Equivalence of DRO and Importance weighting in the convex setting

$\therefore \sup_{Q \in \mathcal{Q}} h(\theta^*, Q) = h(\theta^*, Q^*) = \inf_{\theta \in \Theta} h(\theta, Q^*)$

$\sup_{Q \in \mathcal{Q}} \inf_{\theta \in \Theta} h(\theta, Q) = \inf_{\theta \in \Theta} \sup_{Q \in \mathcal{Q}} h(\theta, Q)$ 성립.

$\Rightarrow$ 또한 (θ^*, Q^*)가 Saddle Point가 된다.

θ가 Q에 대한 최소화 문제의 해 이면서
또한, Q가 θ에 대한 최대화 문제의 해가 된다.

$\therefore$, 어떤 Q, $h(\theta^*, Q)$는 $h(\theta^*, Q^*)$보다 크지 않다.
// θ, $h(\theta, Q^*)$는 // 작지 않다.

Appendix

DRO에서 얻어진 worst-case 분포 Q^*는 원래분포 P에 대한 가중치함수 $W(z)$로 해석, ∴ importance weighting 동일한 결과.

❖ Equivalence of DRO and Importance weighting in the convex setting

DRO의 optimal solution θ^*,

$h(\theta, Q^*)$는 분포 Q^*에 따른 weighted risk

$$\min_{\theta} \max_{Q \in \mathcal{Q}} E_{z \sim Q}[l(\theta; z)]$$

해: (θ^*, Q^*)라고 할 때,

$$\theta^* = \arg\min_{\theta} E_{z \sim Q^*}[l(\theta; z)]$$

Q를 원래 데이터 분포 P에 대한 가중치로 표현!

$$\Rightarrow Q(z) = W(z) * P(z) \Rightarrow E_{z \sim P}[W(z) * l(\theta; z)]$$

DRO의 해 $\theta = \arg\min_{\theta} E_{z \sim P}[W(z) * l(\theta; z)]$

고맙습니다