Group Unknown Invariant Learning

2021.12.27 Lab-Saminar

Taero Kim

MLAI, Univ. of Seoul

Contents

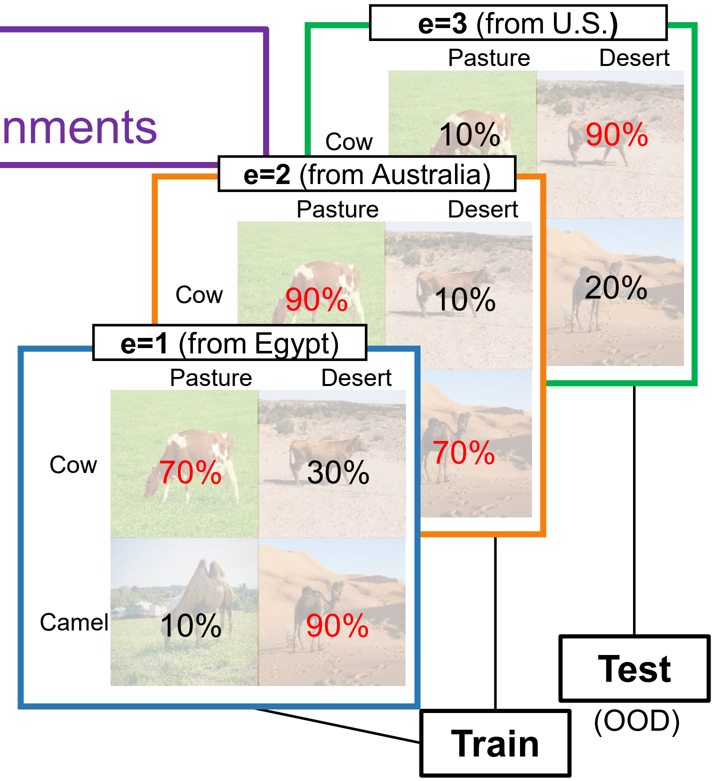
MLAI
University of Seoul

- Learning with Group Information
  - Review **IRM**
  - Environment & Group
  - **Group DRO**
  - Invariant Learning & Group Robustness
- Learning without Group Information
  - Just Train Twice
  - Environment Inference for Invariant Learning

Learning with Group Information

- ## Review IRM Arjovsky et al., 2019

Goal of IRM
Learning invariant correlation across training environments



ERM

$$\min_{\Phi:\mathcal{X}\to\mathcal{Y}} \sum_{e\in\mathcal{E}_{tr}} \mathcal{R}^e(\omega\circ\Phi)$$

Simultaneously Optimal Classifier ω

IRM

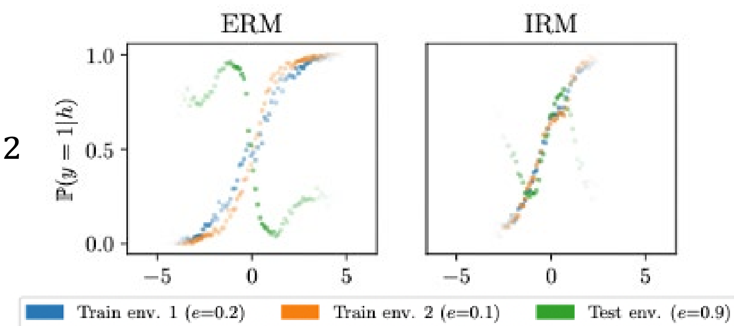
$$\min_{\substack{\Phi:\mathcal{X}\to\mathcal{Y}\\ \omega:\mathcal{H}\to\mathcal{Y}}} \sum_{e\in\mathcal{E}_{tr}} \mathcal{R}^e(\omega\circ\Phi)$$

$$subject\ to\ \omega\in\operatorname*{argmin}_{\overline{\omega}:\mathcal{H}\to\mathcal{Y}} R^e(\overline{\omega}\circ\Phi),\ for\ all\ e\ \in\mathcal{E}_{tr}$$

Lagrange Multiplier Hyperparameter λ

IRMv1

$$\min_{\Phi:\mathcal{X}\to\mathcal{Y}} \sum_{e\in\mathcal{E}_{tr}} \mathcal{R}^e(\Phi)+\lambda\cdot\left\|\nabla_{\omega|\omega=1.0}\mathcal{R}^e(\omega\circ\Phi)\right\|^2$$

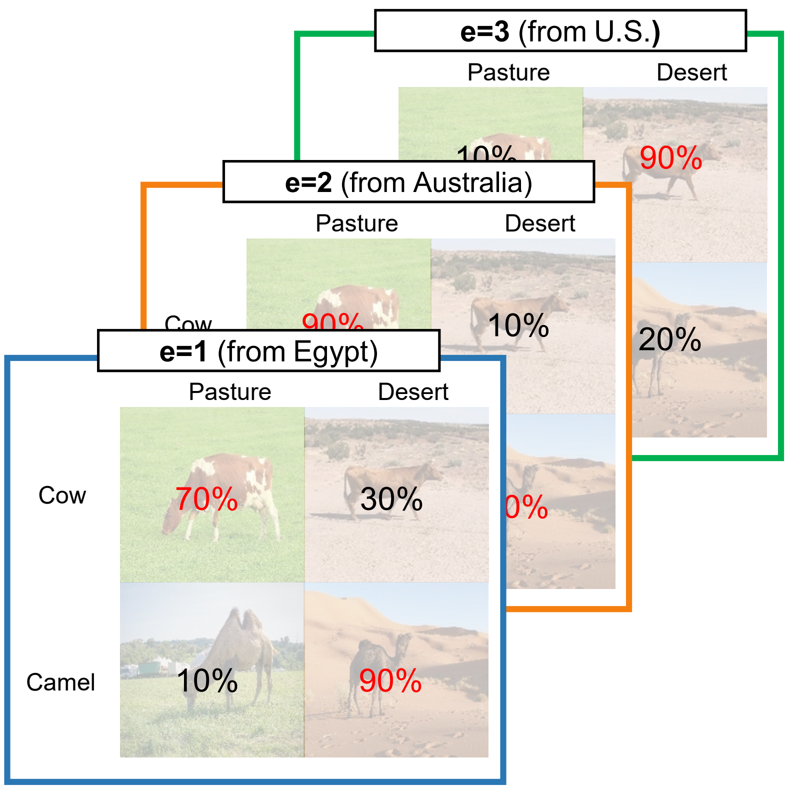


Arjovsky et al., 2019

MLAI University of Seoul

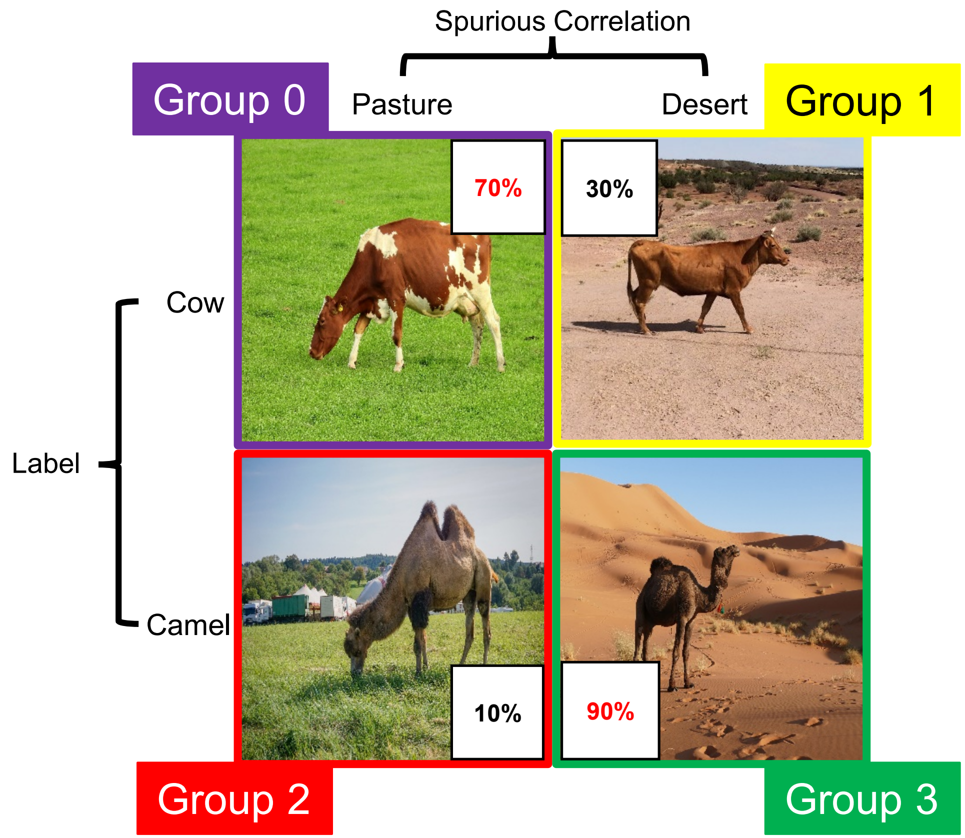
- **Environment & Group**

Environment ~~(or Group)~~



Invariant Learning
IRM (Arjovsky et al. 2019)

Group ~~(or Sub-Group)~~



Group Robustness
Group DRO (Sagawa et al. ICLR 2020)

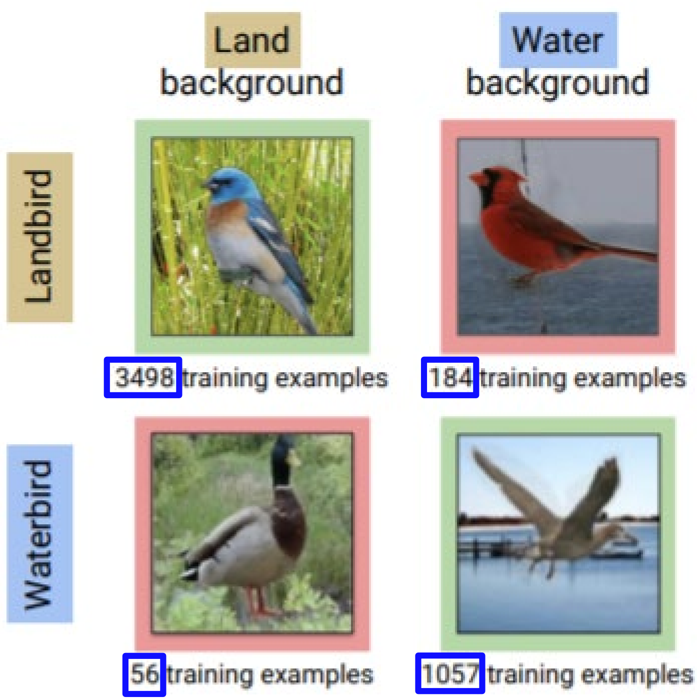
MLAI University of Seoul

- **Group DRO** (Sagawa et al., ICLR 2020)

Worst-Group | Average Loss for **Each Group**

$$\hat{\theta}_{DRO} = \underset{\theta \in \Theta}{\mathrm{argmin}} \{ \hat{R}(\theta) := \max_{g' \in \mathcal{G}} \mathbb{E}_{(x,y,g)}[\ell(\theta;(x,y) \mid g = g')] \}$$

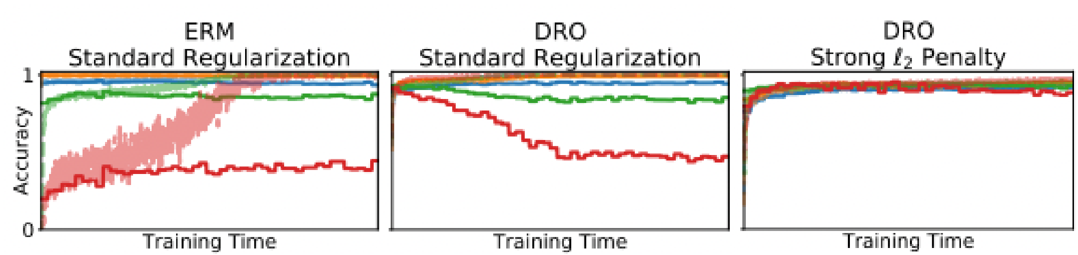
$$= \underset{\theta \in \Theta}{\mathrm{argmin}} \sup_{q \in \Delta_m} \sum_{g=1}^{m} q_g \mathbb{E}_{(x,y)\sim P_g}[\ell(\theta;(x,y))]$$



Waterbirds Dataset

To improve worst-group-accuracy with DRO, we should give **strong regularization**.
(e.g. Strong ℓ_2 Penalty, Early Stopping .. etc.)

This constrains the model's capacity to fit the training data, especially for majority groups.
That is, regularization control the generalization gap $\delta (:= R(\theta) - \hat{R}(\theta))$ across groups



for CelebA

Learning with Group Information

- **Invariant Learning & Group Robustness**

Invariant Learning → Distribution Shift → **IRM**

Goal
No matter how the **data distribution changes**, optimal performance comes out.

For OOD Set

Algorithm	Acc. train envs.	Acc. test env.
ERM	87.4 ± 0.2	17.1 ± 0.6
IRM (ours)	70.8 ± 0.9	$\mathbf{66.9 \pm 2.5}$
Random guessing (hypothetical)	50	50
Optimal invariant model (hypothetical)	75	75
ERM, grayscale model (oracle)	73.5 ± 0.2	73.0 ± 0.4

Group Robustness → Imbalance Group Proportion → **Group DRO**

Goal
Even if **the ratio of the data in each group** constituting the dataset is different,
the performance for the smallest sized group is good.

			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

Both aim to improve worst-case-accuracy.

Learning without Group Information

MLAI University of Seoul

- **Why without group information?**

1. Annotation is more **expensive** than normal dataset

Normal Dataset

$(x, y) \sim P^{obs}(x, y)$: Only need y ∈ $\mathcal{Y}$ Information.

Dataset for Invariant Learning

$(x, y, e) \sim P^{obs}(x, y, e)$ or $(x, y, a) \sim P^{obs}(x, y, a)$

$e \in \mathcal{E}$ is **environment**, $a \in \mathcal{A}$ is **attribute**.

Need $\mathcal{Y} \times \mathcal{E}$ or $\mathcal{Y} \times \mathcal{A}$ information.

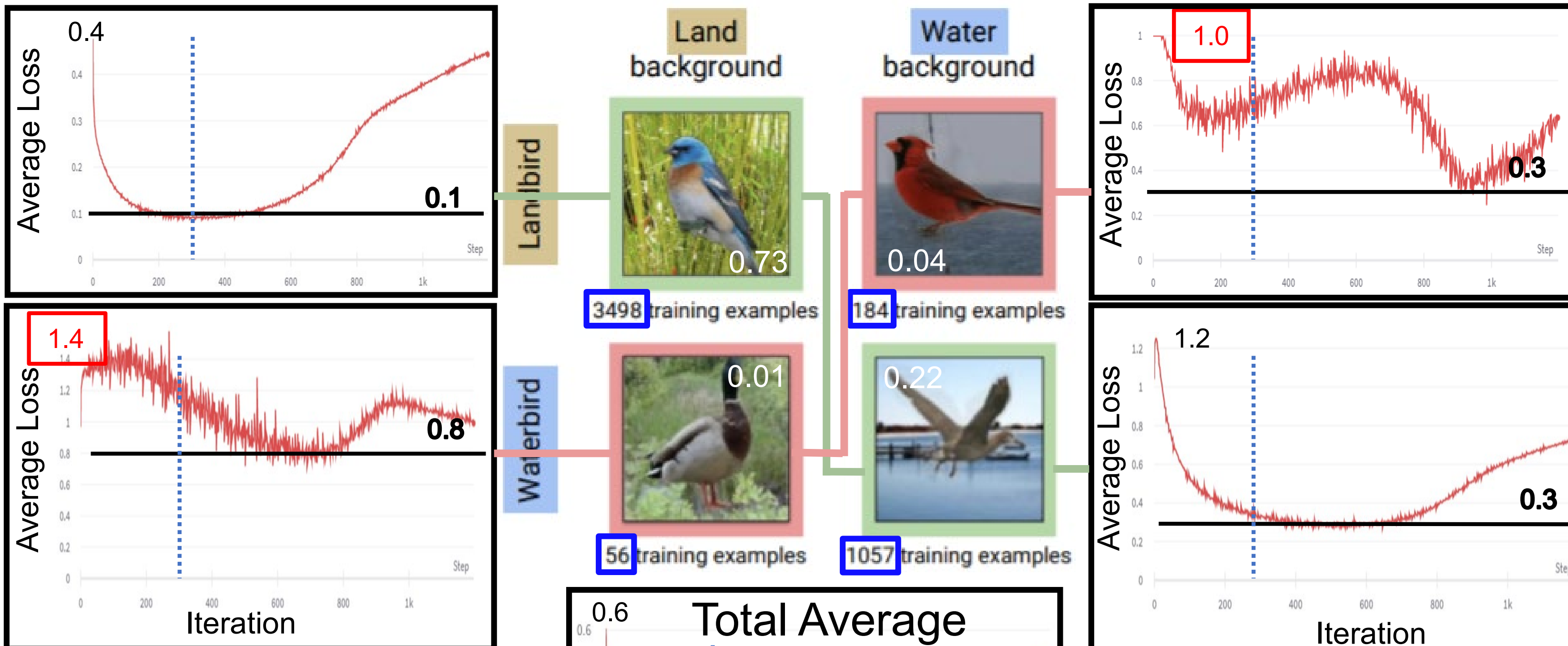
2. Privacy Limitation

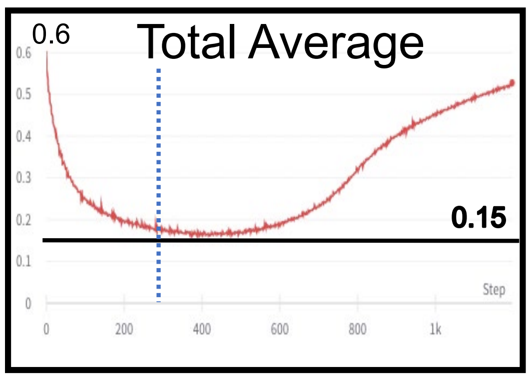
Although we have ability to annotate huge data,
there are some **inaccessible data.**

Instead, we can collect side-informative data.
However, it is unclear how to specify environments or groups by using them.

Learning without Group Information

- **ERM – Train Loss per Group** (resnet-50, 300epochs, batch size 64)





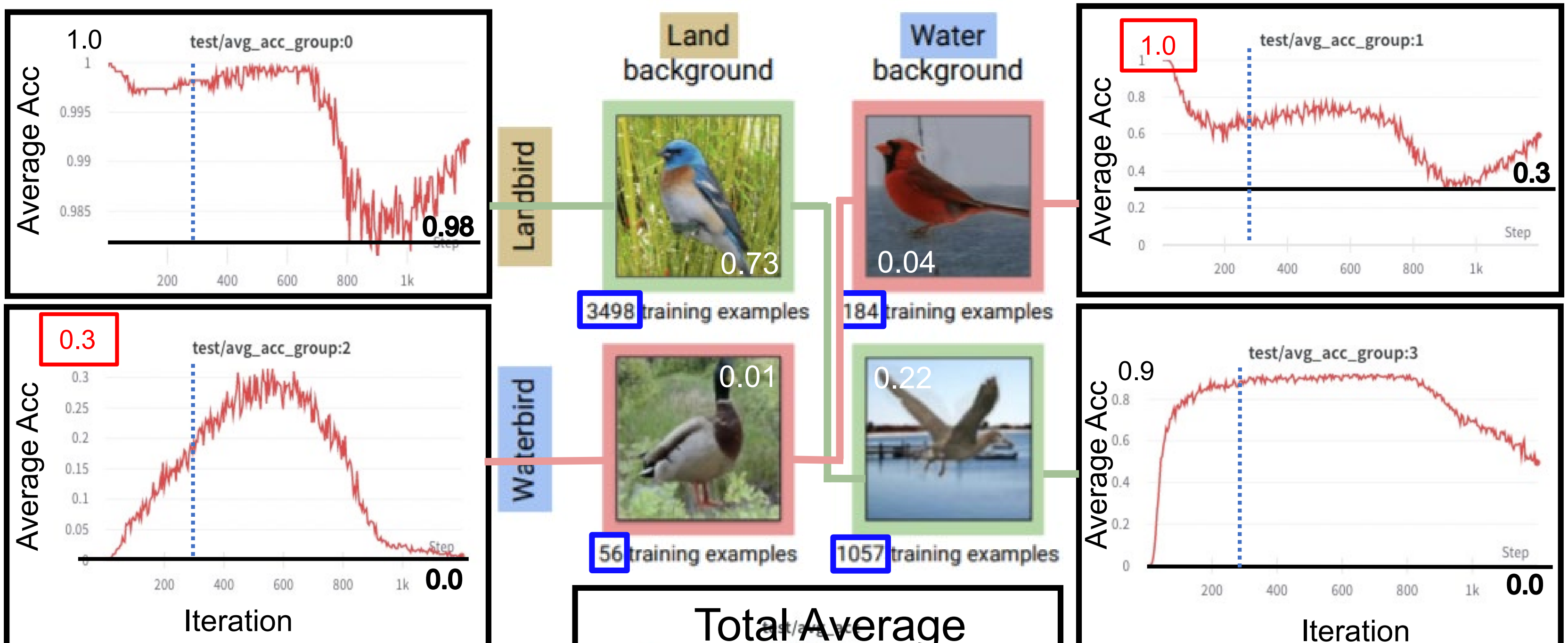
Minority
High loss after enough trains,
Side-effect from majority
y Groups

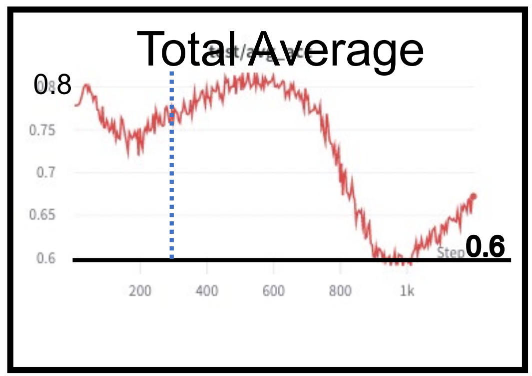
Majority Groups
Low loss after enough trains,
Similar Loss Curve

ERM can be used to classify the groups for learning without group information

MLAI University of Seoul

• ERM – Test Accuracy per Group (resnet-50, 300epochs, batch size 64)





Minority Groups
Side-effect from majority
The smaller the group is, the more difficult to train.

Majority Groups
High Acc with enough trains
Decrease with overfitting and try to train minority

The model's capacity is too small to learn minority group by only using ERM

- **Just Train Twice** (Liu et al., ICML 2021)

Goal of JTT
Improve the worst-group error **without training group annotations.**

#1 **Identification** – Extract Error Set by ERM

$$\hat{f}_{id} : J_{ERM}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(x_i, y_i; \theta)$$

To avoid overfitting

Early Stop at epoch **T**
(T : Hyperparameter)

Error Set : $E = \{(x_i, y_i)\ s.t.\ \hat{f}_{id}(x_i) \neq y_i\}$

Error set is a set of one time wrong data.
It also contains wrong data from majority.

Balancing the data proportion

#2 **Upweighting**

Upweight only samples in the Error Set.
($\boldsymbol{\lambda_{up}}$: Hyperparameter)

$$\hat{f}_{final} : J_{up-ERM}(\theta, E) = \sum_{(x,y)\notin E} \ell(x, y; \theta) + \lambda_{up} \sum_{(x,y)\in E} \ell(x, y; \theta)$$

Original Data
Group 0 - 3498
Group 1 - 184
Group 2 - 56
Group 3 - 1057

Wrong Data → model at epoch T

Error Set
Group 0 - 8
Group 1 - 59
Group 2 - 46
Group 3 - 121

$\times\ 100$ → + Original

JTT Input Data
Group 0 - 4298
Group 1 - 6084
Group 2 - 4656
Group 3 - 12257

Training JTT
Tuning Hyperparameters for worst-group-performance by using group information of validation set.

Learning without Group Information

• Results of JTT

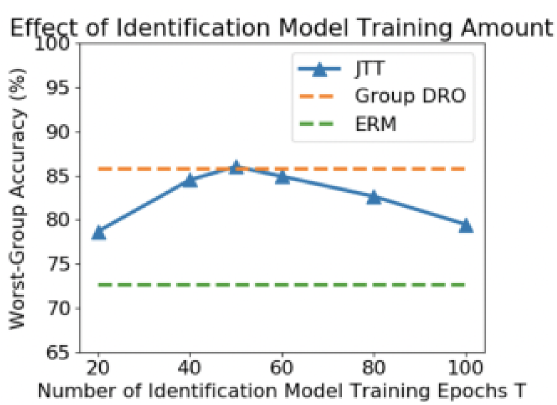
Method	Group labels in train set?	Waterbirds		CelebA		MultiNLI		CivilComments-WILDS	
		Classification : Bird (Landbird, Waterbird) Attribute: Background		Classification : Hair Color (Blond, not Blond) Attribute: Gender		Classification : Sentence relation Entail, neutral, contradict Attribute: Negation		Classification : Word Toxic, Non-toxic Attribute: Demographic id	
		Avg Acc.	Worst-group Acc.	Avg Acc.	Worst-group Acc.	Avg Acc.	Worst-group Acc.	Avg Acc.	Worst-group Acc.
ERM	No	97.3%	72.6%	95.6%	47.2%	82.4%	67.9%	92.6%	57.4%
CVaR DRO (Levy et al., 2020)	No	96.0%	75.9%	82.5%	64.4%	82.0%	68.0%	92.5%	60.5%
LfF (Nam et al., 2020)	No	91.2%	78.0%	85.1%	77.2%	80.8%	70.2%	92.5%	58.8%
JTT (Ours)	No	93.3%	**86.7%**	88.0%	**81.1%**	78.6%	**72.6%**	91.1%	**69.3%**
Group DRO (Sagawa et al., 2020a)	Yes	93.5%	91.4%	92.9%	88.9%	81.4%	77.7%	88.9%	69.9%

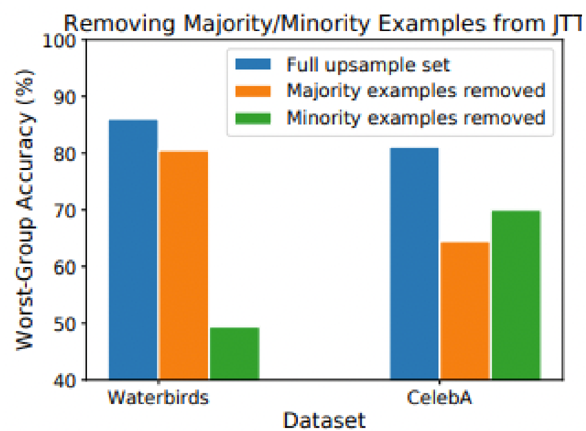
In the Error set,
High proportion of Waterbirds on water?
Upweighting by same data?

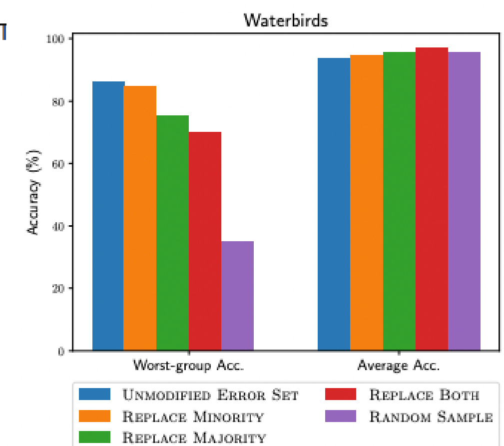
	Worst-group test acc.
Standard error set	86.7%
No waterbirds on water backgrounds	80.7%
Swap error set examples	86%

Why we need the group information of validation set?
To tune the hyperparameter for worst-group

	Waterbirds worst-group test acc.	
	Tuned for average	Tuned for worst-group
CVaR DRO (Levy et al., 2020)	62.0%	75.9%
LfF (Nam et al., 2020)	44.1%	78.0%
JTT (Ours)	62.5%	86.7%







MLAI University of Seoul

- **From IRM to Unknown Group Invariant Learning**

per Environment Risk

$$\mathcal{R}^e = \frac{1}{\sum_{i'} \mathbb{I}\,(e_{i'} = e)} \sum_i \mathbb{I}(e_{i'} = e)\ell(\Phi(x_i), y_i)$$

IRMv1

‖

$$\min_{\Phi:\mathcal{X}\to\mathcal{Y}} \underbrace{\boxed{\sum_{e\in\mathcal{E}_{tr}} \mathcal{R}^e(\Phi)}}_{\text{Predictive Power (ERM)}} + \boxed{\lambda} \cdot \underbrace{\boxed{\Big\| \underbrace{\nabla_{\omega|\omega=1.0}}_{\text{Fixed Scalar}} \underbrace{\mathcal{R}^e(\omega \circ \Phi)}_{\text{Optimal}} \Big\|^2}}_{\text{Optimality of the dummy classifier (IRM)}}$$

Regularizer (Hyperparameter) : Balance between Predictive Power (ERM) and Invariance Power (IRM)

Environment Invariance Constraint (EIC)

$\mathbb{E}[Y^e|\Phi(X^e) = h, e] = \mathbb{E}\left[Y^{e'}\middle|\Phi\left(X^{e'}\right) = h, e'\right]$, for all $e, e' \in \mathcal{E}_{tr}$
(h in the intersection of the supports of $\Phi(X^e)$)

What if the environment is not assigned?

MLAI University of Seoul

- **Environment Inference for Invariant Learning** (Creager et al., ICML 2021)

Goal of EIIL
Find environments that **maximally violate** the invariant learning principle.
= **Discover environment labels** that can be used to train invariant learning model.

per Environment Risk

$$\mathcal{R}^e = \frac{1}{N}\sum_i \boldsymbol{q_i}(e)\ell(\Phi(x_i), y_i)$$

Soft Environment Assignment

$$[\boldsymbol{q_i}(\boldsymbol{e}') \coloneqq \boldsymbol{q}(\boldsymbol{e}'|\boldsymbol{x_i}, \boldsymbol{y_i})]$$

$\boldsymbol{q}$ is a probability distribution for environment

Environment Inference

<$\tilde{\Phi}$ is reference model> Here, we choose ERM

$$C^{EI}(\Phi, \boldsymbol{q}) = \left\|\nabla_{\bar{w}} \tilde{R}^e(\bar{w}\cdot\Phi, \boldsymbol{q})\right\| \longrightarrow$$

Invariant Optimality Term in IRMv1

Maximally Violate the Invariance

$$\boxed{\mathbf{q}^* = \arg\max_{\mathbf{q}} C^{EI}(\tilde{\Phi}, \mathbf{q})}$$

$$\begin{aligned}\Delta\text{EIC} &= (E[y|\Phi(x), e_1] - E[y|\Phi(x), e_2])^2 \\ &\approx \sum_b \Big(\sum_i s_{ib} y_i \mathbf{q}_i(e = e_1) - \sum_i s_{ib} y_i \mathbf{q}_i(e = e_2)\Big)^2\end{aligned}$$

Proposition 1 *Consider environments that differ in the degree to which the label y agrees with the spurious features z: $\mathbb{P}(\mathbb{1}(y = z)|e_1) \neq \mathbb{P}(\mathbb{1}(y = z)|e_2)$: then a reference model $\tilde{\Phi} = \Phi_{Spurious}$ that is invariant to valuable features v and solely focuses on spurious features z maximally violates the invariance principle (EIC). Likewise, consider the case with fixed representation Φ that focuses on the spurious features: then a choice of environments that maximally violates (EIC) is $e_1 = \{v, z, y|\mathbb{1}(y = z)\}$ and $e_2 = \{v, z, y|\mathbb{1}(y \neq z)\}$.*

MLAI
University of Seoul

- **Practical Realization of Maximally Violate**

Lable_noise = 0.25
$Corr_{label}$ = 0.75

Train #1 (color_noise =0.1)

	Green	Blue
0~4	90%	10%
5~9	10%	90%

$Corr_{color}$(#1) = 0.9

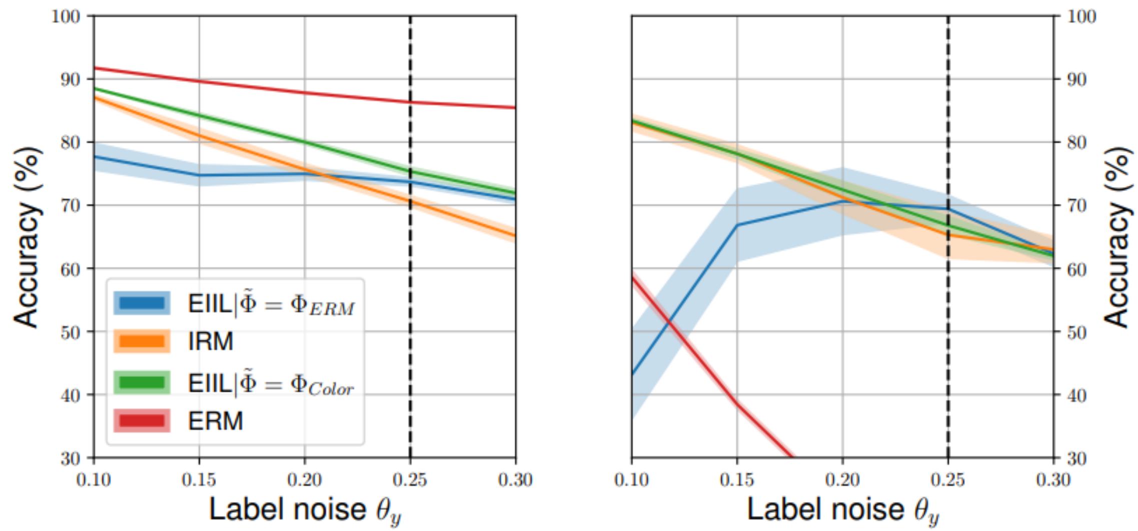
Train #2 (color_noise =0.2)

	Green	Blue
0~4	80%	20%
5~9	20%	80%

$Corr_{color}$(#2) = 0.8

If Corr(color), correlation between color and data, is large enough than Corr(label), correlation between label and data, **the ERM model is forced to learn about the color.**

By using this, we can indirectly realize **Maximally Violating the Invariance**



(a) Train accuracy. (b) Test accuracy

- **Analysis the Environment Inference**

Dataset with environment information does not always perform better than dataset without environment information.

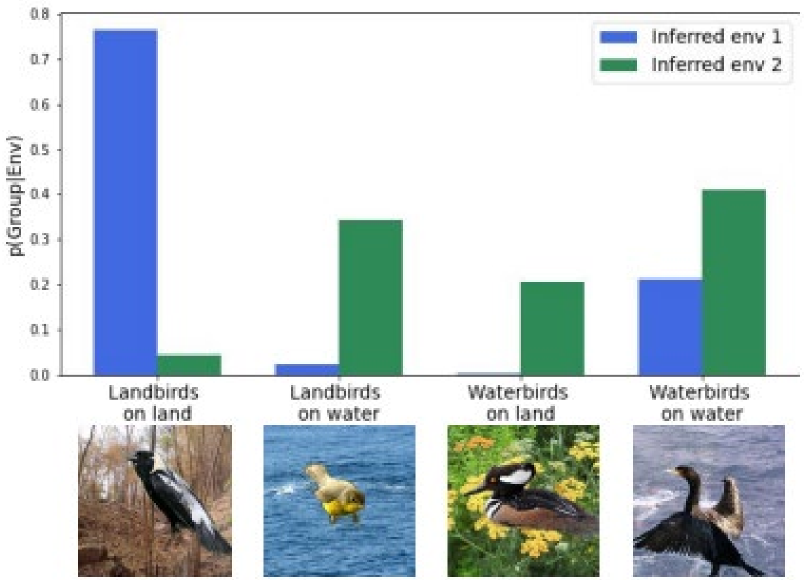
If the given dataset doesn't maximally violate the invariance for the model that we want to train, the train on the inferred dataset can have better performance.

Color MNIST

Method	Handcrafted Environments	Train	Test
ERM	✗	**86.3 ± 0.1**	13.8 ± 0.6
IRM	✓	71.1 ± 0.8	65.5 ± 2.3
EIIL	✗	73.7 ± 0.5	**68.4 ± 2.7**

Waterbirds

Method	Train (avg)	Test (avg)	Test (worst group)
ERM	100.0	**97.3**	60.3
EIIL	99.6	96.9	**78.7**
GroupDRO (oracle)	99.1	96.6	84.6



- **[GroupDRO]**
  Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization
  Sagawa et al. ICLR 2020
- **[IRM]**
  Invariant Risk Minimization
  Arjovsky et al. 2019
- **[JTT]**
  Just Train Twice: Improving Group Robustness without Training Group Information
  Liu et al., 2021 ICML
- **[EIIL]**
  Environment Inference for Invariant Learning
  Creager et al., 2021 ICML
- Simple data balancing achieves competitive worst-group-accuracy
  [Survey for these topics]
  Idrissi et al., 2021

Recommend Authors
(for invariant learning or group robustness)
- **Percy Liang**
- **Chelsea Finn**
- **David Lopez-Paz**
- Martin Arjovsky
- Shiori Sagawa