< 2024.11.07 >



연세대학교
YONSEI UNIVERSITY

Robust Transfer Learning for Out-of-Distribution Dataset

Assistant Professor @ Yonsei University
Kyungwoo Song (송경우)
https://mlai.yonsei.ac.kr/
kyungwoo.song@yonsei.ac.kr

OOD (out-of-distribution)

Introduction

OOD (out-of-distribution)

- $P(Y|X)$
- $P_{ID}(X) \neq P_{OOD}(X)$

How can we fine-tune our models to perform well on both the ID dataset and the OOD dataset?

Test (ID and OOD) (OOD dataset is not available during Training phase)

Train (ID)

ImageNet (Deng et al.)



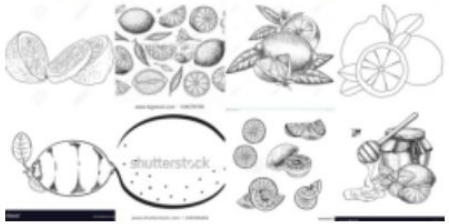
ImageNetV2 (Recht et al.)



ImageNet-R (Hendrycks et al.)



ImageNet Sketch (Wang et al.)



ObjectNet (Barbu et al.)



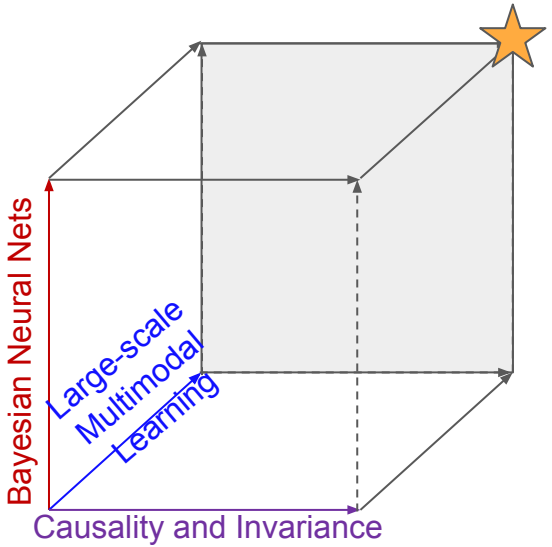
ImageNet-A (Hendrycks et al.)



source: Wortsman, Mitchell, et al. "Robust fine-tuning of zero-shot models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

Towards Reliable Machine Learning (ML) in the Wild

- How can we increase an intelligence of ML? ⇒ Large-scale Multimodal Learning
- How can we formulate the uncertainty-aware modeling? ⇒ Bayesian Neural Nets
- How can we make model more trustworthy? ⇒ Causality and Invariance



Short Term

- Transfer Learning with new metric
- Bayesian Transfer Learning/Bayesian Model Merging
- Worst-case optimization with sufficient invariant features

Long Term

- Bayesian Invariant Transfer Learning

Contents

Transfer Learning Perspective
- Towards Calibrated Robust Fine-Tuning

Bayesian Perspective
- Sharpness-aware Bayesian Transfer Learning
- Bayesian Model Merging

Invariant Learning Perspective
- Sufficient Invariant Learning (SIL)
- Benchmark for SIL
- Causal In-Context Learning

All of these are done by students, and they are actively looking for internship or visiting researcher opportunities.

A new metric to diagnose the model's OOD performance



Changdae Oh, Ph.D. Student
@ UW-Madison

Transfer Learning Perspective
- Robust Fine-tuning [NeurIPS 24]

Previous fine-tuning models only focus on validation accuracy or loss (NLL) => We also need to focus on ECE

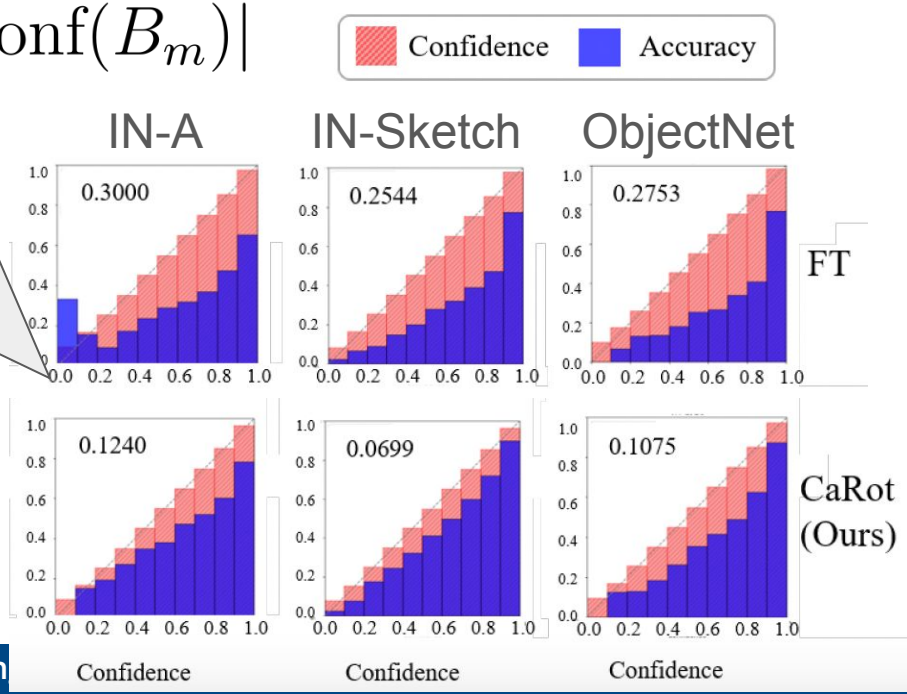
ECE: expected calibration error (**representative evaluation metric**)

- ECE divides the confidence score of N samples into M uniform confidence bins, B_m

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|$$

ECE measures **calibration error**

- metric that quantifies the difference between confidence (max predicted probabilities) and actual outcome
- confidence: max prob.
- e.g.) [0.8(=cat), 0.2(=dog)] => 0.8



Motivation

Validation /Test

Test (only)

Train

In-Distribution (ID)

Out-Of-Distribution (OOD)

Output

Output

Output







Traditional Methods

- Training to minimize the cross-entropy loss (predicted value and y-lavel)
- It is vulnerable to predict OOD dataset
- ID accuracy does not directly indicates OOD accuracy

Validation performance (accuracy, loss, ECE, …) does not guarantee the Test performance

$\epsilon(\cdot)$: calibration error
$\sigma_{min}(\cdot)$: smallest singular value
h : hypothesis class
d : dimension of the data x
D_{OOD} : OOD dataset
$\hat{D}_{ID}$: Observed ID dataset

We derive an upper bound for
1) OOD ECE
2) OOD Error

Stated informally

$$\epsilon_{\mathcal{D}_{OOD}}(h) \leq \epsilon_{\hat{\mathcal{D}}_{ID}}(h) + \frac{d}{\sigma_{\min}(\tilde{\Sigma}_{\mathcal{D}_{ID}})} + c_1$$

$$\mathbb{E}_{\mathcal{D}_{OOD}}[(h(x) - y)^2] \leq \epsilon_{\hat{\mathcal{D}}_{ID}}(h) + \frac{d}{\sigma_{\min}(\tilde{\Sigma}_{\mathcal{D}_{ID}})} + c_2$$

Observation

We derive an upper bound for

1) OOD ECE
2) OOD Error

Stated informally

OOD ECE and OOD accuracy is bounded by the smallest singular value (larger smallest singular value would be helpful to decrease ECE and loss)

$\epsilon(\cdot)$: calibration error
$\sigma_{min}(\cdot)$: smallest singular value
h : hypothesis class
d : dimension of the data x
D_{OOD} : OOD dataset
$\hat{D}_{ID}$: Observed ID dataset

$$\epsilon_{\mathcal{D}_{OOD}}(h) \leq \epsilon_{\hat{\mathcal{D}}_{ID}}(h) + \frac{d}{\sigma_{\min}(\tilde{\Sigma}_{\mathcal{D}_{ID}})} + c_1$$

$$\mathbb{E}_{\mathcal{D}_{OOD}}[(h(x) - y)^2] \leq \epsilon_{\hat{\mathcal{D}}_{ID}}(h) + \frac{d}{\sigma_{\min}(\tilde{\Sigma}_{\mathcal{D}_{ID}})} + c_2$$

Stated informally

$$\epsilon_{\mathcal{D}_{OOD}}(h) \leq \epsilon_{\hat{\mathcal{D}}_{ID}}(h) + \frac{d}{\sigma_{\min}(\tilde{\Sigma}_{\mathcal{D}_{ID}})} + c_1$$

$$\mathbb{E}_{\mathcal{D}_{OOD}}[(h(x) - y)^2] \leq \epsilon_{\hat{\mathcal{D}}_{ID}}(h) + \frac{d}{\sigma_{\min}(\tilde{\Sigma}_{\mathcal{D}_{ID}})} + c_2$$

$$\tilde{\Sigma}_{\mathcal{D}_{ID}} = \mathbb{E}_{\mathcal{D}_{ID}}[\tilde{x}\tilde{x}^T]$$

$$\tilde{x} = (\tilde{x}_1, \ldots, \tilde{x}_d)$$

$$\tilde{x}_i = (x_i - \mathbb{E}[x_i])\mathrm{Var}(x_i)^{-1/2}$$

The smallest singular value of the covariance matrix for the dataset (it could be the embedding, the intermediate layer representation) **Strategy) increase the smallest singular value**

Stated informally

$$\epsilon_{\mathcal{D}_{OOD}}(h) \leq \epsilon_{\hat{\mathcal{D}}_{ID}}(h) + \frac{d}{\sigma_{\min}(\Sigma_{\mathcal{D}_{ID}})} + c_1$$

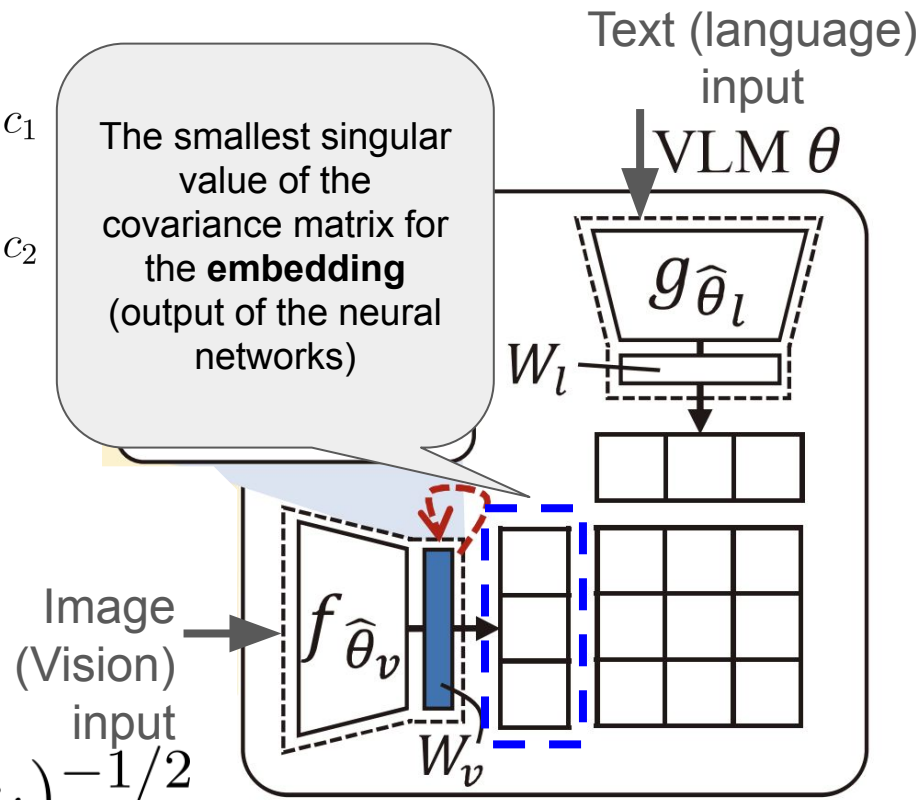
$$\mathbb{E}_{\mathcal{D}_{OOD}}[(h(x) - y)^2] \leq \epsilon_{\hat{\mathcal{D}}_{ID}}(h) + \frac{d}{\sigma_{\min}(\Sigma_{\mathcal{D}_{ID}})} + c_2$$

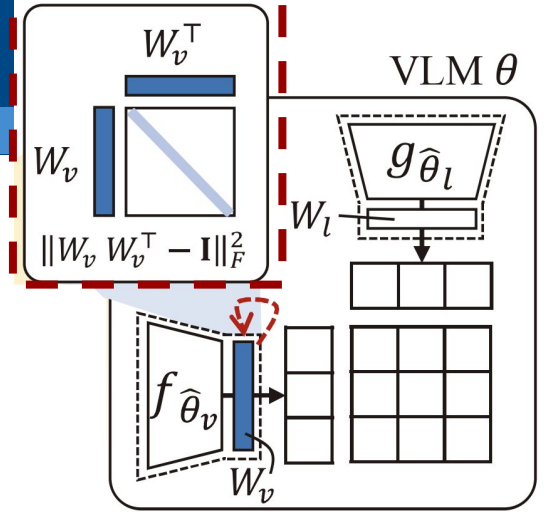
The smallest singular value of the covariance matrix for the **embedding** (output of the neural networks)

$$\tilde{\Sigma}_{\mathcal{D}_{ID}} = \mathbb{E}_{\mathcal{D}_{ID}}[\tilde{x}\tilde{x}^T]$$

$$\tilde{x} = (\tilde{x}_1, \ldots, \tilde{x}_d)$$

$$\tilde{x}_i = (x_i - \mathbb{E}[x_i])\mathrm{Var}(x_i)^{-1/2}$$





Strategy) increase the smallest singular value

Traditional Multimodal Contrastive Learning

$$\mathcal{L}_{\text{MCL}}(\theta) = \frac{1}{2N}\sum_{i=1}^{N} -\log \frac{\exp(f_{\theta_v}(I_i)\cdot g_{\theta_l}(T_i))}{\sum_{j=1}^{N}\exp(f_{\theta_v}(I_i)\cdot g_{\theta_l}(T_j))}$$

$$+ \frac{1}{2N}\sum_{i=1}^{N} -\log \frac{\exp(f_{\theta_v}(I_i)\cdot g_{\theta_l}(T_i))}{\sum_{j=1}^{N}\exp(f_{\theta_v}(I_j)\cdot g_{\theta_l}(T_i))} + R(\theta_v, \theta_l),$$

CaRot: Add orthogonality regularization term

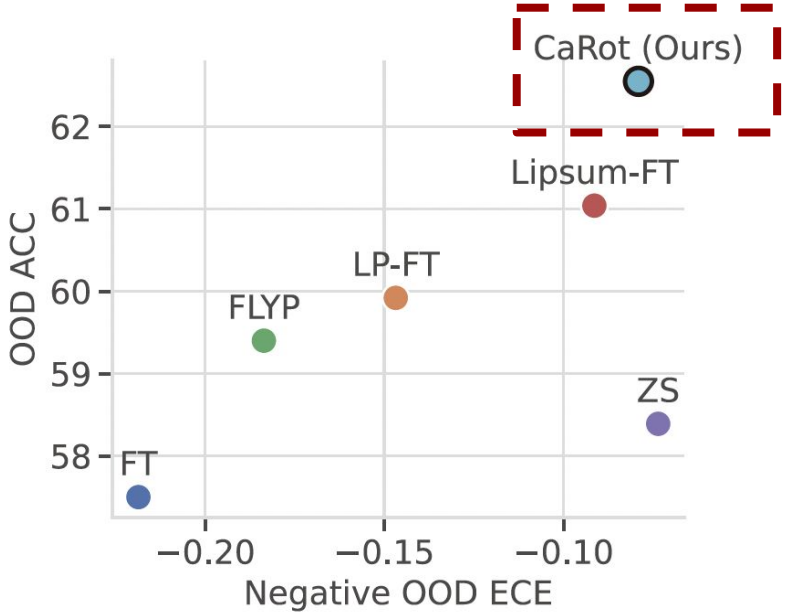
$$\mathcal{L}_{\text{MCL-con}}(\theta) := \mathcal{L}_{\text{MCL}}(\theta) + \lambda_{OC}||W_v W_v^T - \mathbf{I}||_F^2$$

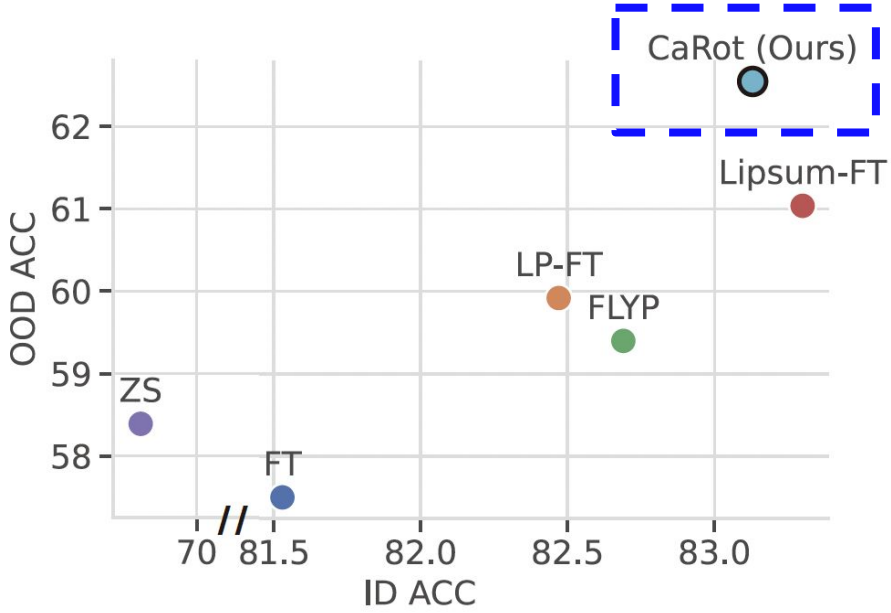
Orthogonal Constraints

Results

CaRot method archives improved OOD ECE

CaRot method achieves improved ID ACC and OOD ACC





Transfer learning is a de facto method for neural networks.
What about Bayesian Neural Networks (BNNs)?

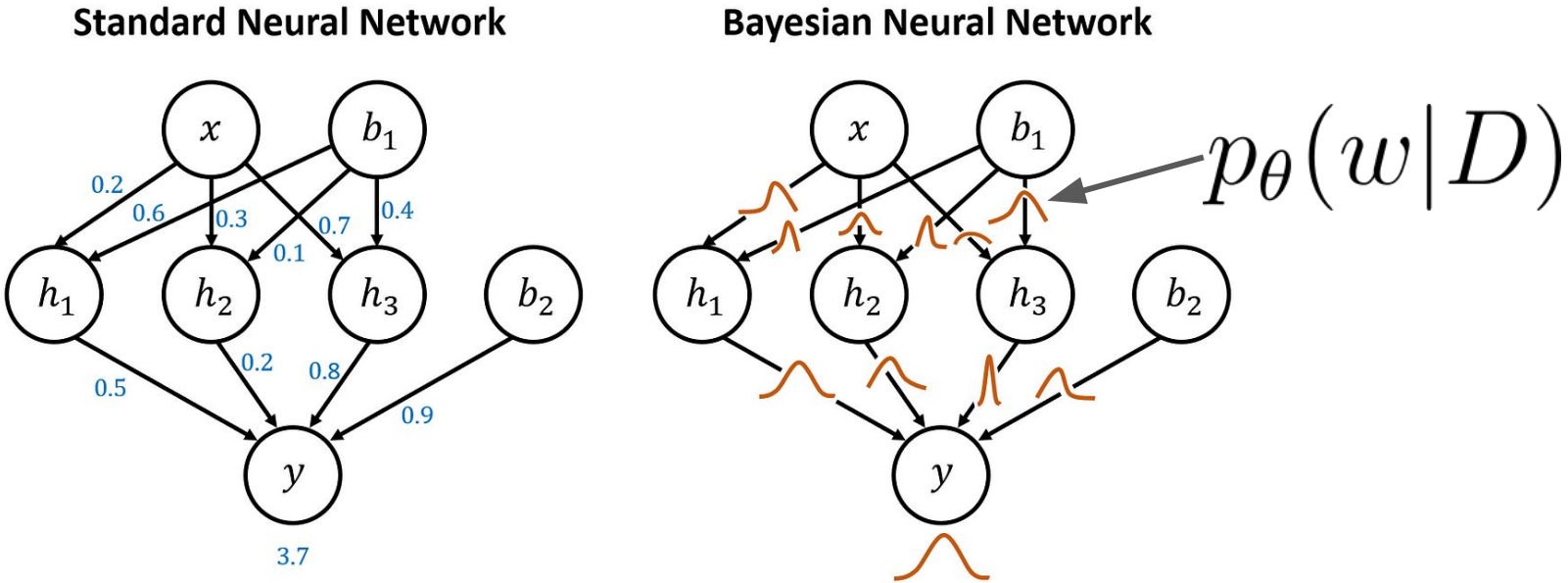


Sungjun Lim, Ph.D. Student
@ Yonsei University

Bayesian Perspective
- Bayesian Transfer Learning

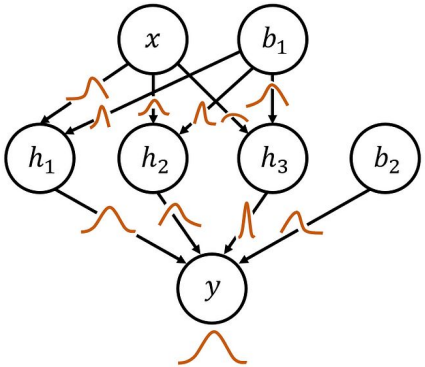
Recently, Bayesian Neural Networks (BNN) show remarkable results

- In some settings, BNN performs better than DNN
- BNN has advantage of uncertainty and model averaging



Recently, Bayesian Neural Networks (BNN) show
- In some settings, BNN performs better than
- BNN has advantage of uncertainty and mod

Bayesian Neural Network



Sample M weights from the posterior. Then, average the M models output together. (Bayesian Model Averaging; BMA)

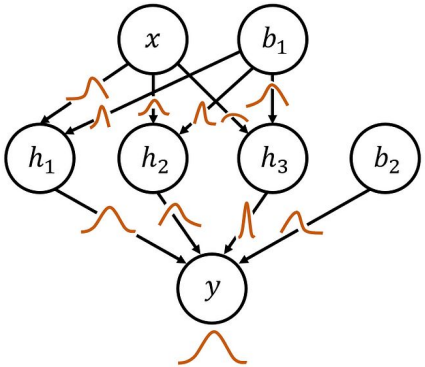
$$p(y^*|x^*,\mathcal{D}) = \int_w p(y^*|x^*,w)p_\theta(w|\mathcal{D})dw$$

$$\approx \frac{1}{M}\sum_{i=1}^{M} p(y^*|x^*,w_i), \quad w_i \sim p_\theta(w|\mathcal{D})$$

Recently, Bayesian Neural Networks (BNN) show
- In some settings, BNN performs better than
- BNN has advantage of uncertainty and mod

Sample M weights from the posterior. Then, average the M models output together. (Bayesian Model Averaging; BMA)

Bayesian Neural Network

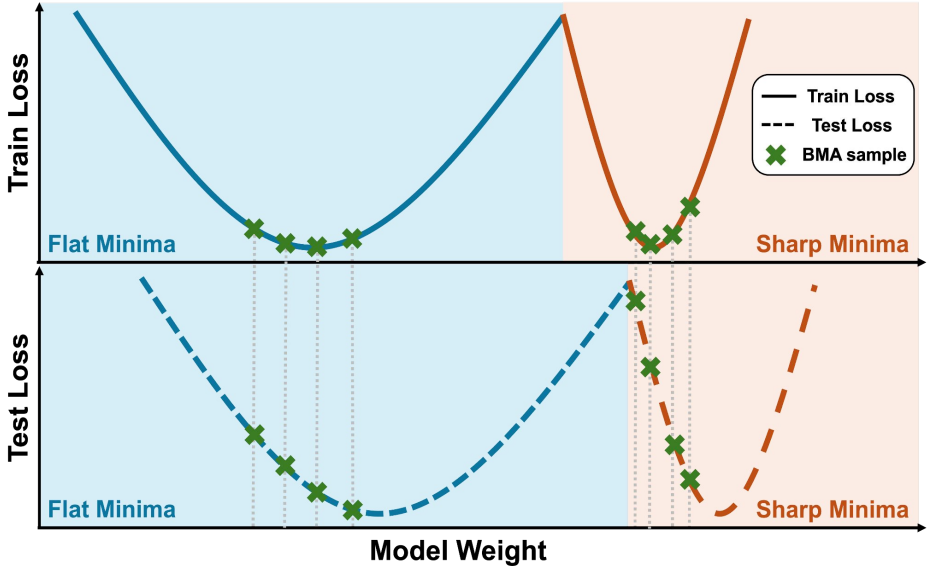


$$p(y^*|x^*,\mathcal{D}) = \int_w p(y^*|x^*,w)p_\theta(w|\mathcal{D})dw$$

$$\approx \frac{1}{M}\sum_{i=1}^{M} p(y^*|x^*,w_i), \quad w_i \sim p_\theta(w|\mathcal{D})$$

Transfer Learning for BNN?

- Flatness of loss landscape is one of key factor for generalization
  - Bayesian Model Averaging (BMA) can also be affected from the flatness.
- Optimization to find flat minima is needed in Bayesian approach, either.



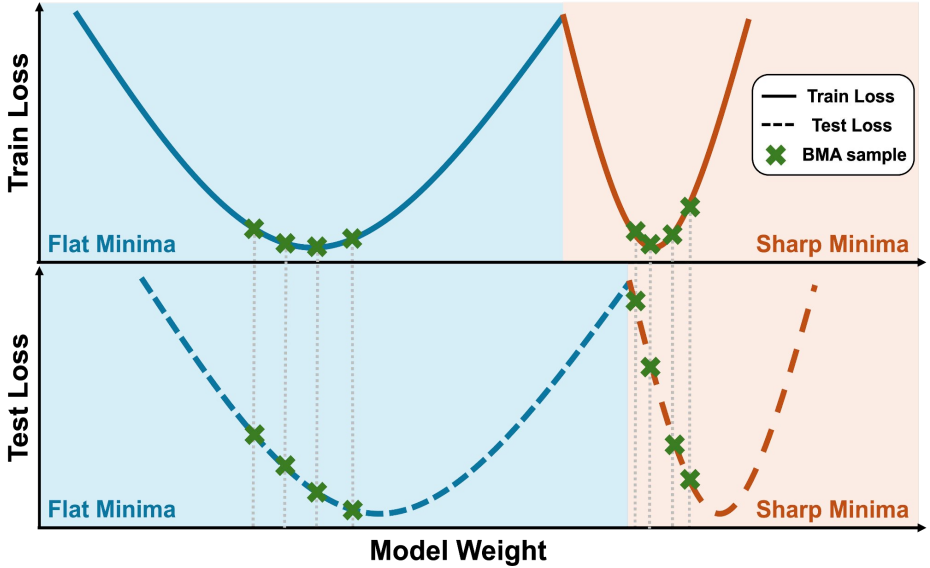
Flat local minima are known to be better than sharp local minima.

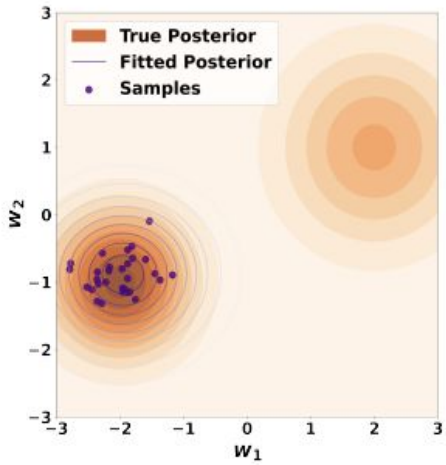
How about BNN?

Motivation

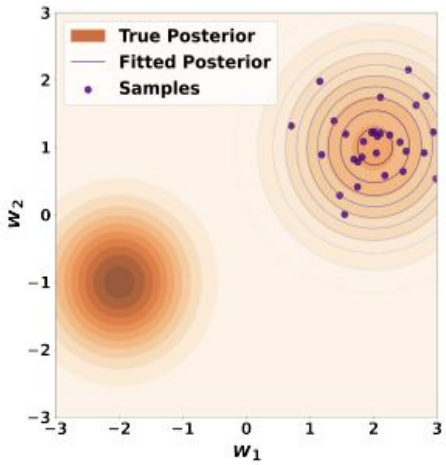
Introduction

- Flatness of loss landscape is one of key factor for generalization
  - Bayesian Model Averaging (BMA) can also be affected from the flatness.
- Optimization to find flat minima is needed in Bayesian approach, either.





VI



SA-BMA

$$p(y^*|x^*,\mathcal{D}) = \int_w p(y^*|x^*,w)p_\theta(w|\mathcal{D})dw$$
$$\approx \frac{1}{M}\sum_{i=1}^{M} p(y^*|x^*,w_i), \quad w_i \sim p_\theta(w|\mathcal{D})$$

[BMA Flatness Bound]

$$\mathrm{Flat}(w_{\mathrm{BMA}}) \leq \frac{1}{M}\sum_{i=1}^{M}\mathrm{Flat}(w_i)$$

$\mathrm{Flat}(\cdot)$: Flatness

w_{BMA} : Averaged model

w_i : BMA sample

Flatness does matter for BMA

[BMA Flatness Bound]

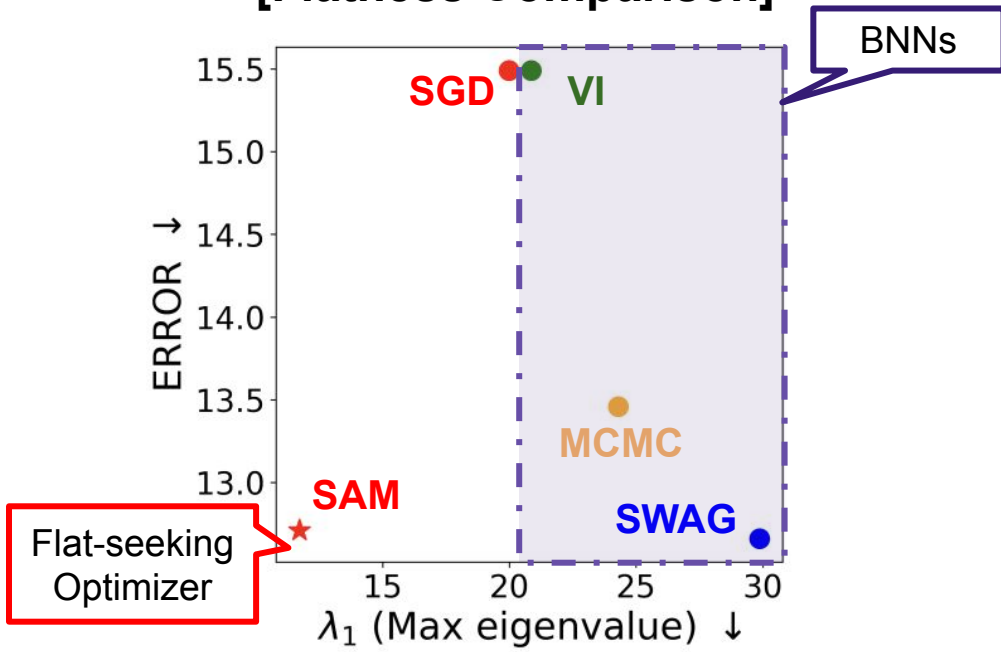
$$\mathrm{Flat}(w_{\mathrm{BMA}}) \leq \frac{1}{M}\sum_{i=1}^{M}\mathrm{Flat}(w_i)$$

$\mathrm{Flat}(\cdot)$: Flatness

w_{BMA} : Averaged model

w_i : BMA sample

[Flatness Comparison]



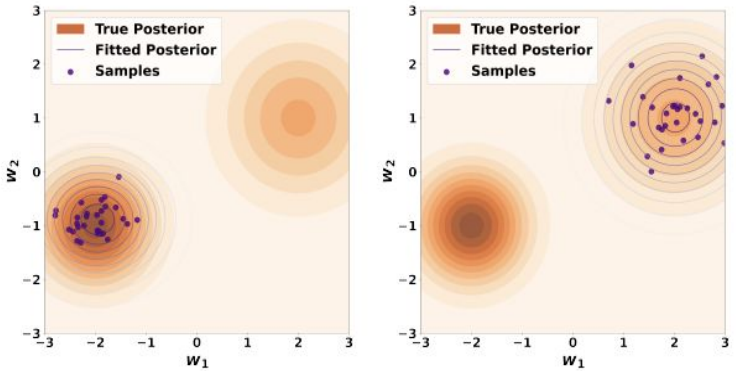
Flatness does matter for BMA

BNNs do not guarantee the flatness

- **Bayesian-fitting Sharpness-aware Optimizer**

$$l^{\gamma}_{\text{SA-BMA}}(\theta) = \min_{\theta} \max_{d|\theta+\Delta\theta,\theta|\leq\gamma^2} l(\theta + \Delta\theta)$$

$$\text{s.t. } d|\theta + \Delta\theta, \theta| = D_{KL}\left[p_{\theta+\Delta\theta}(w|\mathcal{D}) \;||\; p_{\theta}(w|\mathcal{D})\right]$$



(c) VI (d) SA-BMA

Divergence Constraints
Gaussian with θ=(μ,Σ): Posterior of w

Sharpness-aware Bayesian Model Averaging (SA-BMA)

- **Bayesian-fitting Sharpness-aware Optimizer**

$$l^{\gamma}_{\text{SA-BMA}}(\theta) = \min_{\theta} \max_{d|\theta+\Delta\theta,\theta|\leq\gamma^2} l(\theta + \Delta\theta)$$

s.t. $d|\theta + \Delta\theta, \theta| = D_{KL}\left[p_{\theta+\Delta\theta}(w|\mathcal{D}) \;||\; p_{\theta}(w|\mathcal{D})\right]$

Divergence Constraints
Gaussian with θ=(μ,Σ): Posterior of w

$$\Delta\theta_{\text{SA-BMA}} = \gamma \frac{F_{\theta}(\theta)^{-1}\nabla_{\theta} l(\theta)}{\sqrt{\nabla_{\theta} l(\theta)^{T} F_{\theta}(\theta)^{-1}\nabla_{\theta} l(\theta)}}$$

- **Bayesian-fitting Sharpness-aware Optimizer**

$$l^{\gamma}_{\text{SA-BMA}}(\theta) = \min_{\theta} \max_{d|\theta+\Delta\theta,\theta|\leq\gamma^2} l(\theta + \Delta\theta)$$

s.t. $d|\theta + \Delta\theta, \theta| = D_{KL}\left[p_{\theta+\Delta\theta}(w|\mathcal{D}) \;||\; p_{\theta}(w|\mathcal{D})\right]$

Divergence Constraints
Gaussian with θ=(μ,Σ): Posterior of w

$$\Delta\theta_{\text{SA-BMA}} = \gamma \frac{F_{\theta}(\theta)^{-1}\nabla_{\theta} l(\theta)}{\sqrt{\nabla_{\theta} l(\theta)^{T} F_{\theta}(\theta)^{-1}\nabla_{\theta} l(\theta)}}$$

If we only handle $\boldsymbol{\mu}$ (without Σ)

i) SA-BMA degenerates to FSAM by using the diagonal terms of FIM.

ii) SA-BMA degenerates to SAM if FIM is an identity matrix.

iii) Update rule of SA-BMA $\theta - \eta_{SA\text{-}BMA}\nabla_{\theta} l(\theta + \Delta\theta)$ *degenerates to update rule of NG* $\theta - \eta_{NG} F_{y}(\theta)^{-1}\nabla_{\theta} l(\theta)$ *with learning rate* $\eta_{SA\text{-}BMA} = \frac{\eta_{NG}}{(1+\gamma')} F_{\theta}(\theta)^{-1}$.

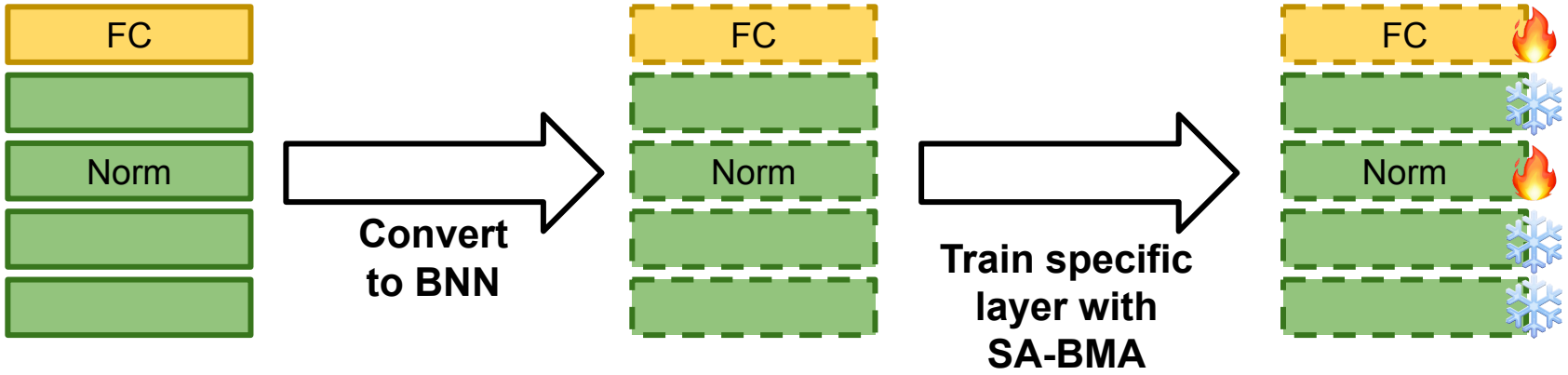
Sharpness-aware Bayesian Model Averaging (SA-BMA)

- **Bayesian-fitting Sharpness-aware Optimizer**

$$l^{\gamma}_{\text{SA-BMA}}(\theta) = \min_{\theta} \max_{d|\theta+\Delta\theta,\theta|\leq\gamma^2} l(\theta+\Delta\theta)$$

$$\text{s.t. } d|\theta+\Delta\theta,\theta| = D_{KL}\left[p_{\theta+\Delta\theta}(w|\mathcal{D}) \;||\; p_{\theta}(w|\mathcal{D})\right]$$

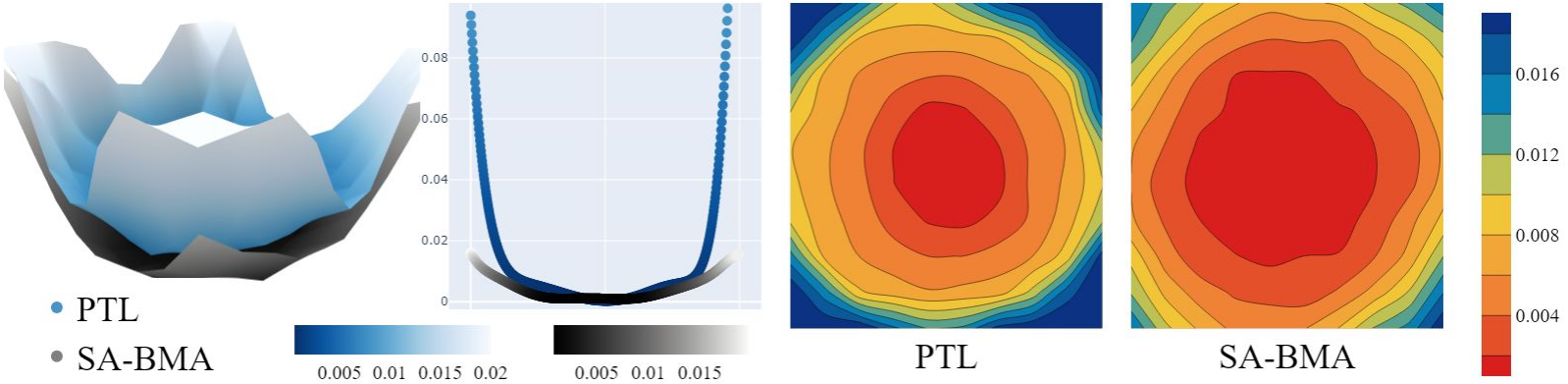
- **Suggest a Bayesian framework-agnostic transfer learning**



Results

Experiments

[Loss surfaces]



(Model: ViT B/16), (Dataset: CIFAR-10 10-shot)

	ACC↑	ECE↓	NLL↓
SGD	84.37	0.056	0.503
SWAG	83.51	0.022	0.510
Ours (BNN)	**86.81**	**0.010**	**0.399**

Method	$\lambda_1 \downarrow$	$\lambda_1/\lambda_5 \downarrow$
SGD	559.62	2.59
SAM	381.74	2.23
FSAM	561.15	2.24
bSAM	532.74	2.09
MOPED	686.90	2.41
PTL	559.16	2.23
E-MCMC	540.83	1.98
SWAG	602.34	2.13
F-SWAG	362.33	2.44
SA-BMA	**275.21**	**1.69**

[Hessian Eigenvalue]

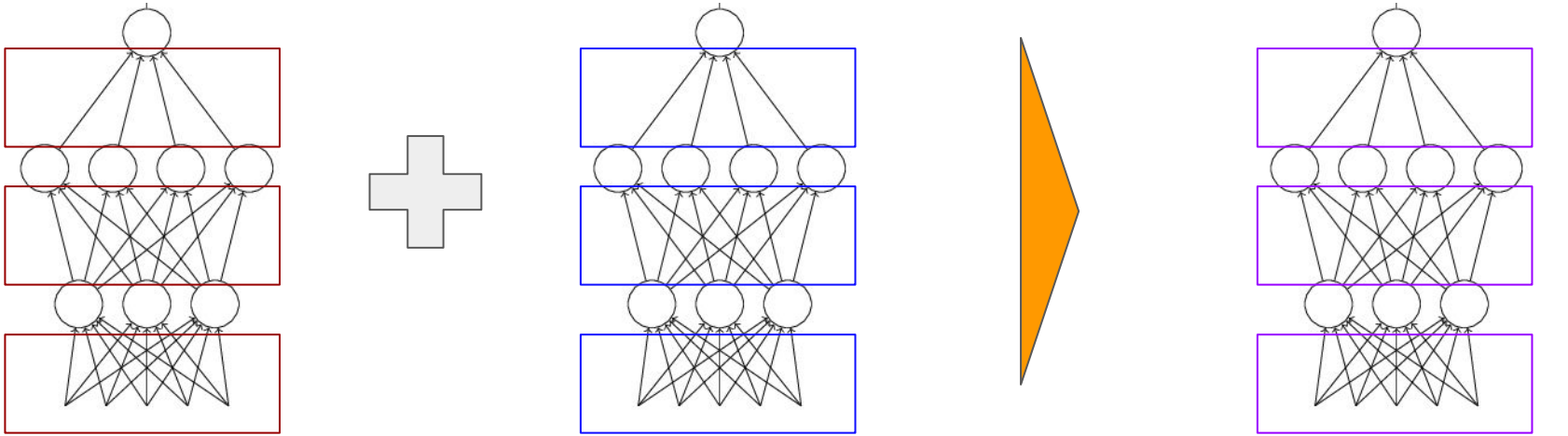
Model Merging assumes linear mode connectivity
How about Bayesian Model Merging?



Sungjun Lim, Ph.D. Student
@ Yonsei University

Bayesian Perspective
- Bayesian Model Merging

- Model Merging
  - Goal : Combine multiple task-specific models into a single unified model
  - Diverse methods for Model merging

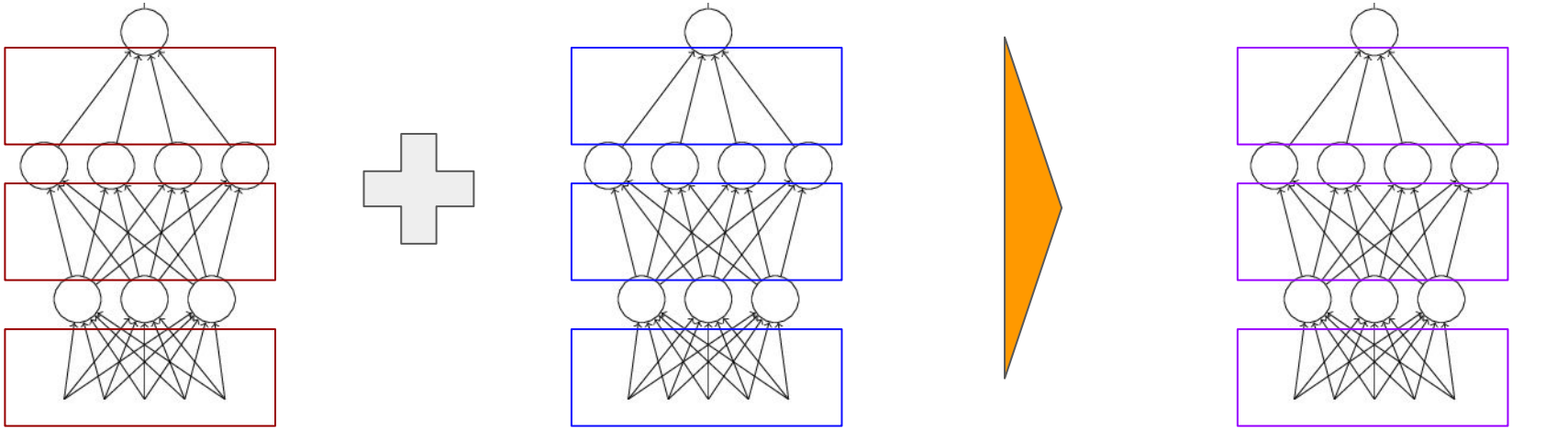


Model A
(trained by dataset A)

Model B
(trained by dataset B)

Merged Model
(A+B)/2
parameter averaging

- Model Merging
  - Goal : Combine multiple task-specific models into a single unified model
  - Diverse methods for Model merging

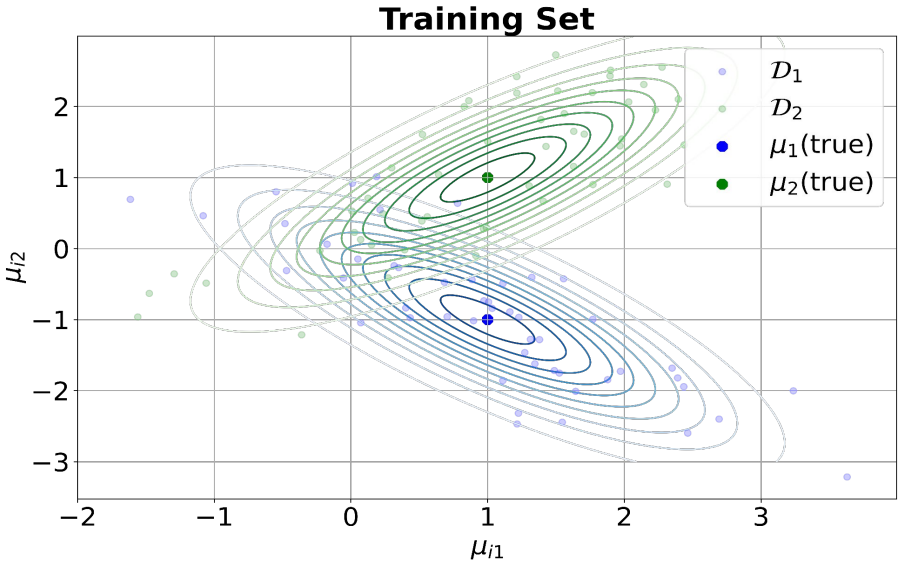


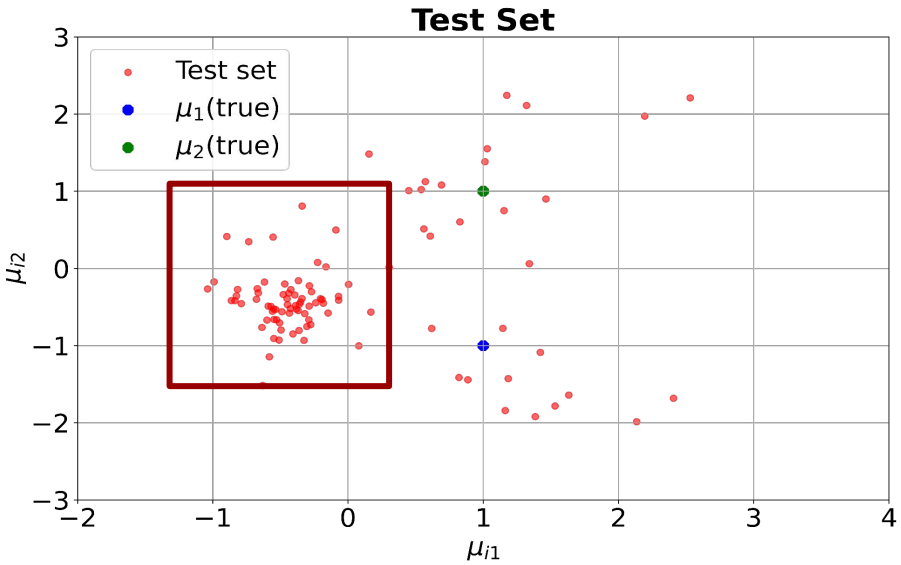
Model A
(trained by **Korean math** dataset)

Model B
(trained by **English** dataset)

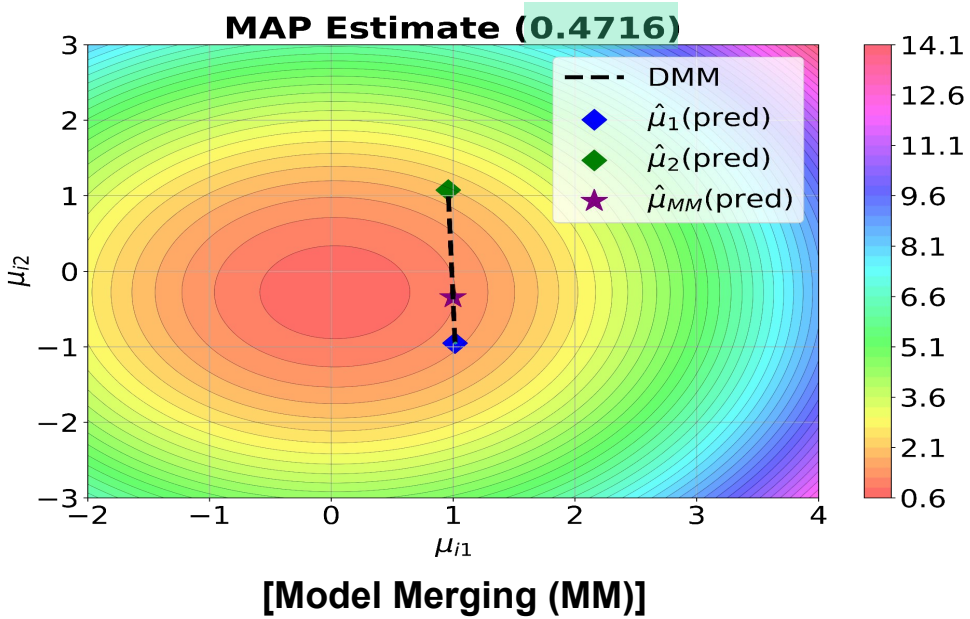
Merged Model
=> perform well on **English math** dataset

Distribution shift between training set and test set

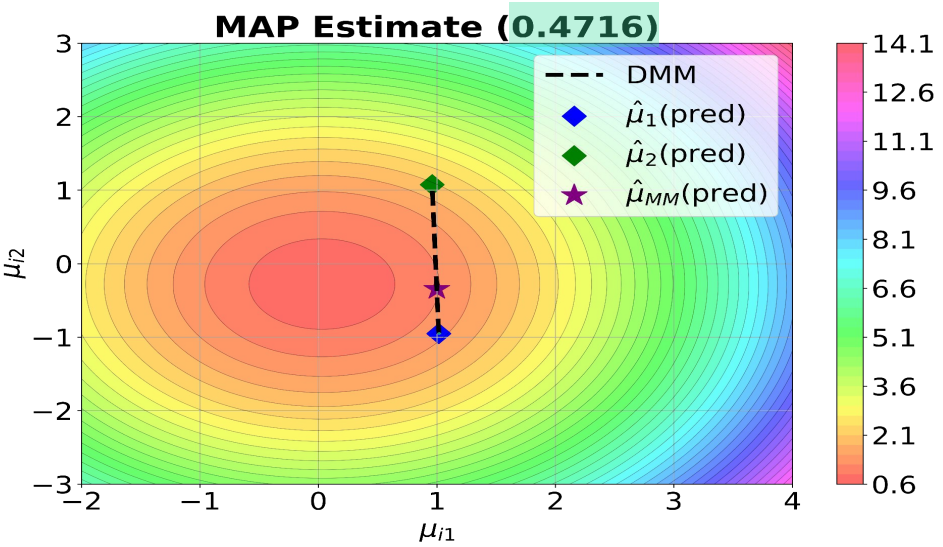




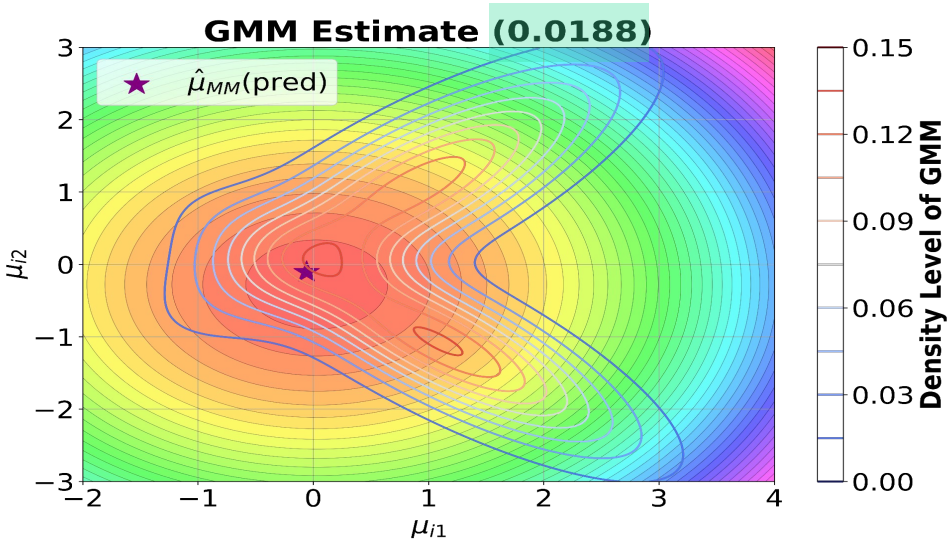
Model merging only interpolates the linear path between model parameters



Ours) Bayesian Model Merging
1) Estimate the posterior (Gaussian)
   a) Laplace approximation
2) Gaussian Mixture Models



[Model Merging (MM)]



[Bayesian Model Merging (BMM)]

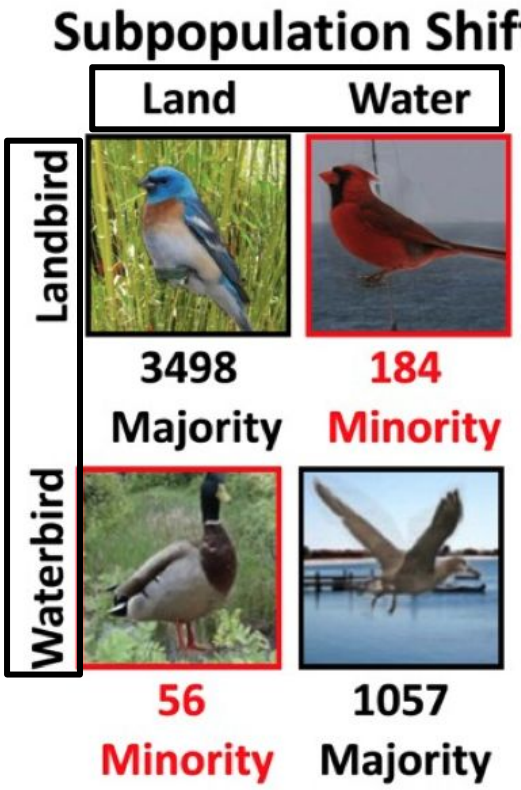
ERM focuses on spurious correlations.
DRO/GDRO capture only a single invariant feature.
How can we capture diverse invariant features?



Taero Kim, Ph.D. Student
@ Yonsei University

Invariant Learning Perspective - Sufficient Invariant Learning

Background) Invariance

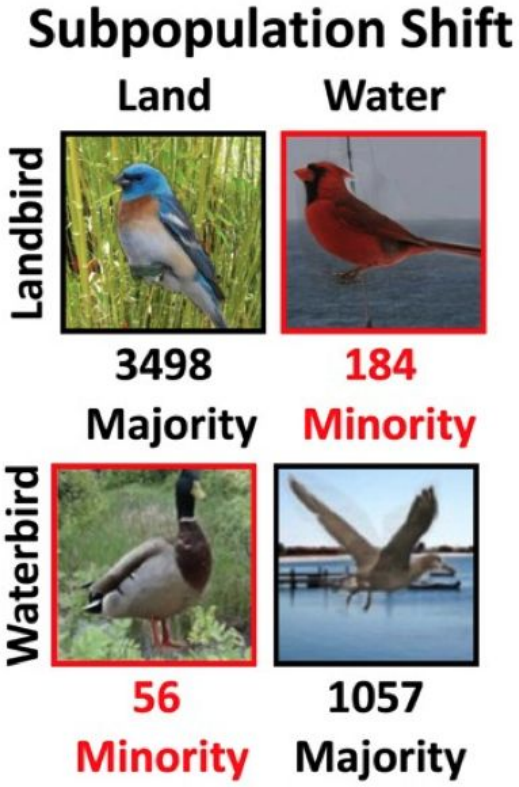


Background

Class Label

<Majority group>
Most waterbirds (landbirds) are on the water (land)

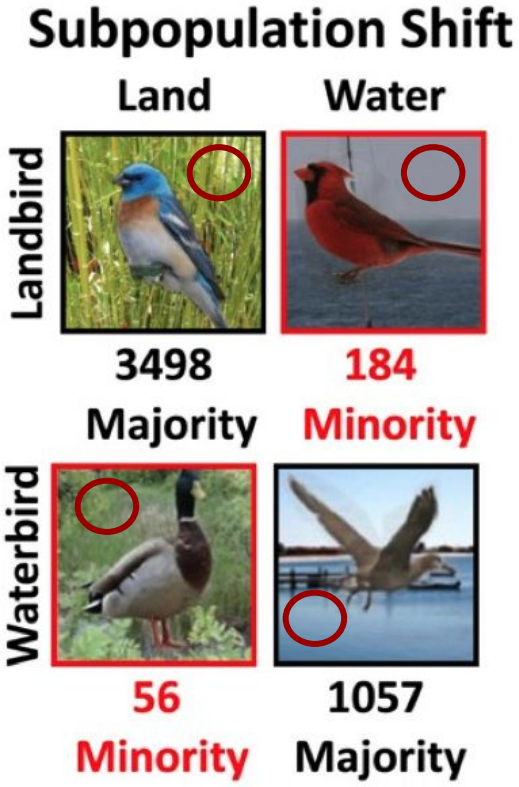
Background) Invariance



Task: landbird or waterbird?
(binary classification task)

What is the easiest shortcut to classify the bird?

Background) Invariance

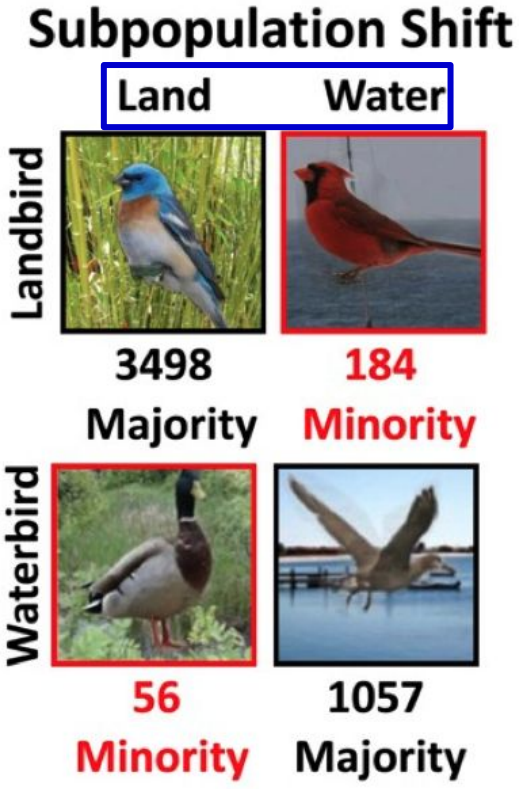


Task: landbird or waterbird?
(binary classification task)

What is the easiest shortcut
to classify the bird?
=> **background (water or land)**

ERM focuses on the background
(spurious correlation)

Background) Invariance



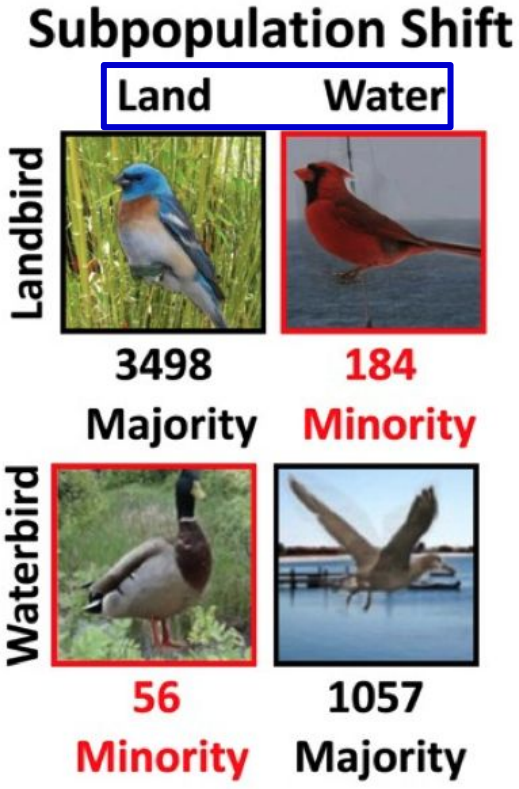
Let models focus on **invariant features**

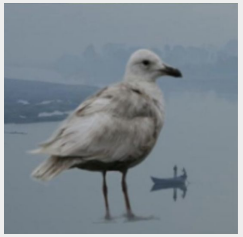
Invariant Condition

$$\mathbb{E}[Y^e|Z_I] = \mathbb{E}[Y^{e'}|Z_I] \quad for \ \forall e, e' \in \mathcal{E}_{tr}$$

The model output (Y) remains the same regardless of the environment (e).
(we will interchangeably use environment, group, domain)

Background) Invariance







Beak



Tail



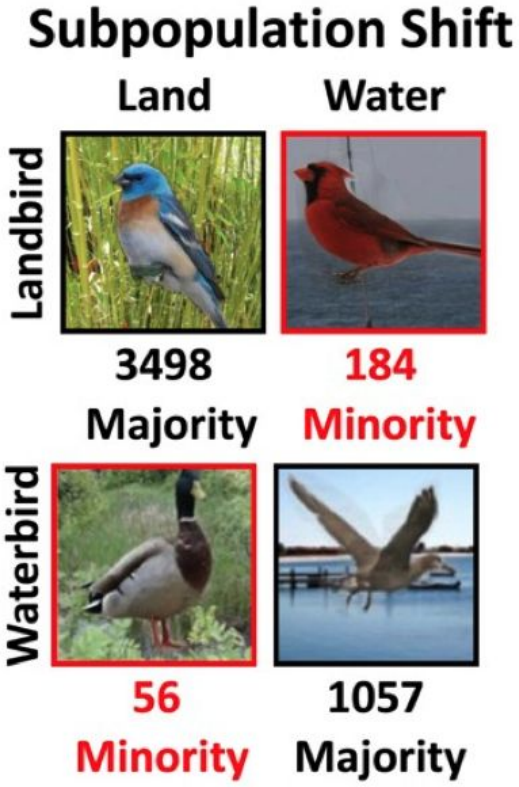
Foot

Let models focus on **invariant features**

Invariant Condition

$$\mathbb{E}[Y^e|Z_I] = \mathbb{E}[Y^{e'}|Z_I] \quad for \ \forall e, e' \in \mathcal{E}_{tr}$$

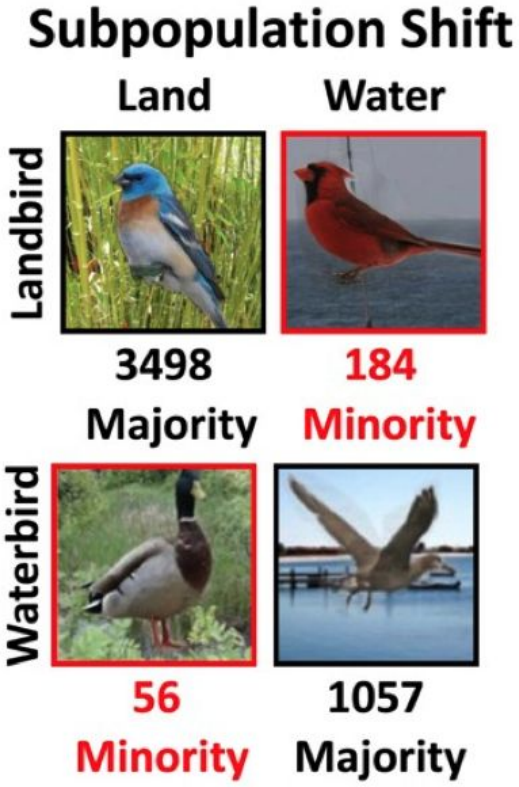
The model output (Y) remains the same regardless of the environment (e).
(we will interchangeably use environment, group, domain)



GDRO (Sagawa et al., ICLR 2019)

$$\max_{e \in \mathcal{E}_{tr}} \mathcal{R}^e(\theta)$$

Minimize each loss of environment



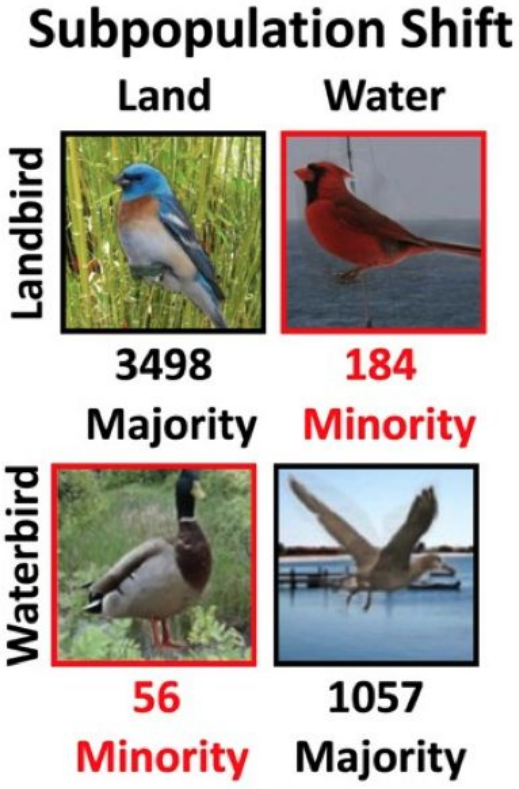
GDRO (Sagawa et al., ICLR 2019)

$$\max_{e \in \mathcal{E}_{tr}} \mathcal{R}^e(\theta)$$

Minimize each loss of environment

Stated informally (Rojas-Carulla et al., JMLR 2018)

The **minimizer of DRO/GDRO** includes solutions that focus on **invariant features**.



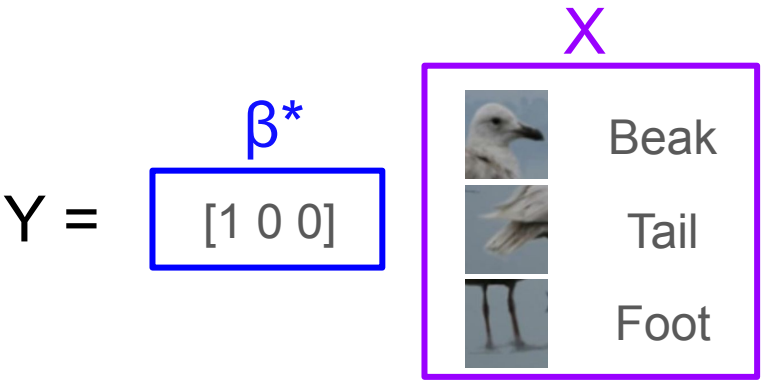
GDRO (Sagawa et al., ICLR 2019)

$$\max_{e \in \mathcal{E}_{tr}} \mathcal{R}^e(\theta)$$

Minimize each loss of environment

Stated informally (Rojas-Carulla et al., JMLR 2018)

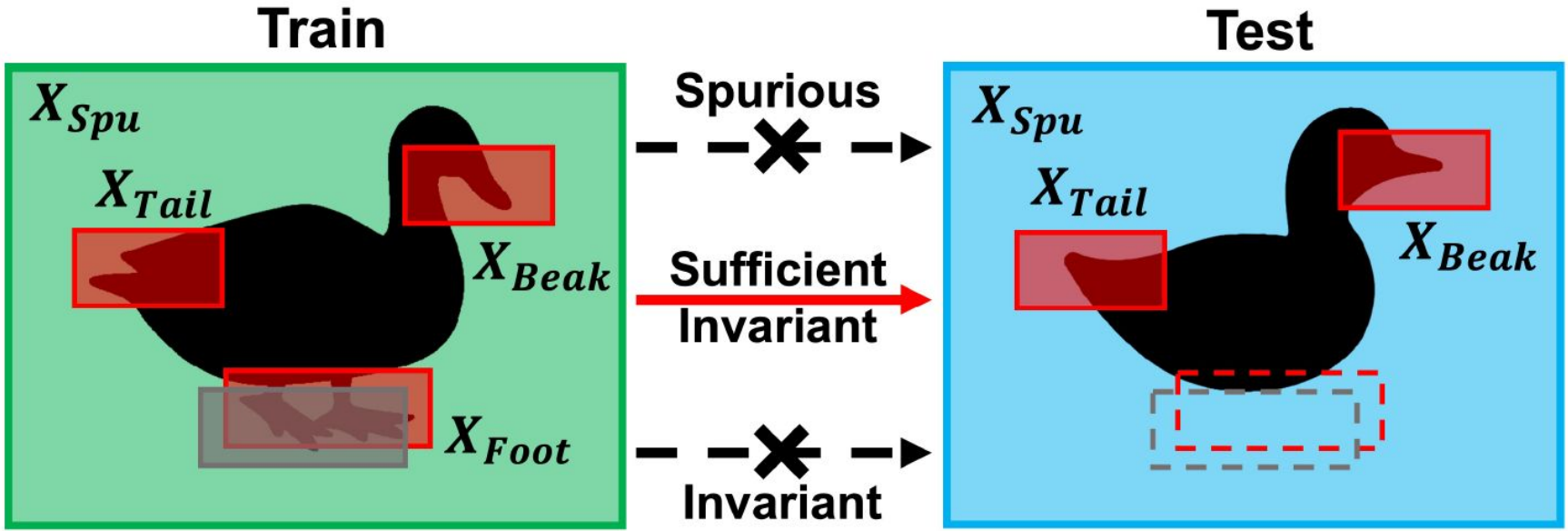
The **minimizer of DRO/GDRO** includes solutions that focus on **invariant features**.



Sufficient Invariant Learning

DRO/GDRO captures a single invariant feature.

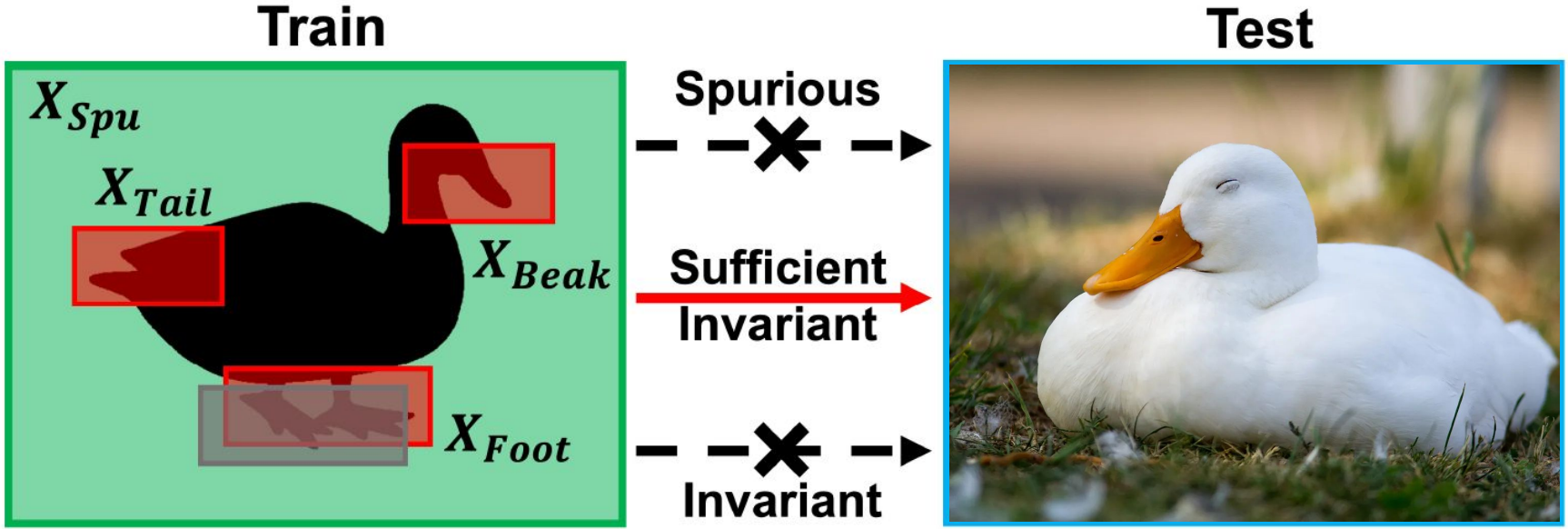
- What if we have multiple invariant features?
- What if some invariant features are missing at test?



Sufficient Invariant Learning

DRO/GDRO captures a single invariant feature.

- What if we have multiple invariant features?
- What if some invariant features are missing at test?



Sufficient Invariant Learning

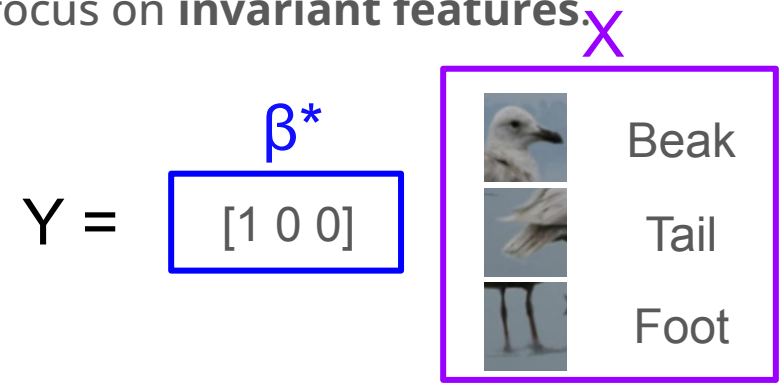
GDRO (Sagawa et al., ICLR 2019)

$$\max_{e \in \mathcal{E}_{tr}} \mathcal{R}^e(\theta)$$

Minimize each loss of environment

Stated informally (Rojas-Carulla et al., JMLR 2018)

The **minimizer of GDRO** includes solutions that focus on **invariant features**.



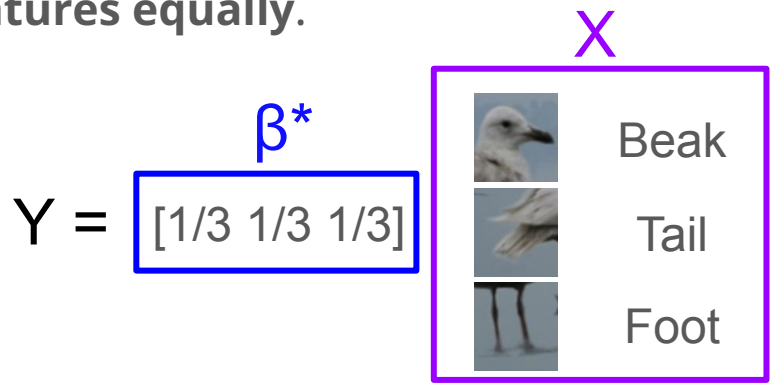
ASGDRO (Ours)

$$\max_{e \in \mathcal{E}_{tr}} \max_{||\epsilon_e|| \leq \rho} \mathcal{R}^e(\theta + \epsilon_e)$$

Minimize each adversarial loss of environment

Stated informally

The **minimizer of ASGDRO** includes solutions that focus on **all invariant features equally**.



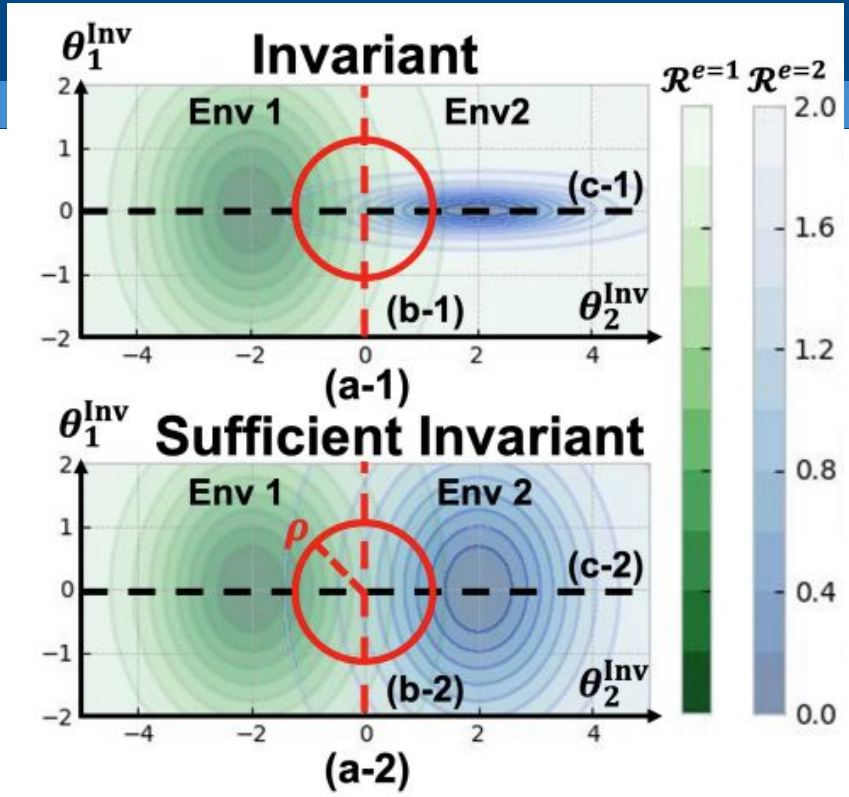
ASGDRO (Ours)

$$\max_{e\in\mathcal{E}_{tr}} \max_{||\epsilon_e||\leq\rho} \mathcal{R}^e(\theta+\epsilon_e)$$

Minimize each adversarial loss of environment

$$\mathcal{R}_{\text{ASGDRO}}(\theta) = \max_{e\in\mathcal{E}_{tr}} \mathcal{R}^e(\theta+\epsilon_e^*)$$

$$\epsilon_e^* = \rho \frac{T_\theta^2 \nabla\mathcal{R}^e(\theta)}{\|T_\theta \nabla\mathcal{R}^e(\theta)\|}$$



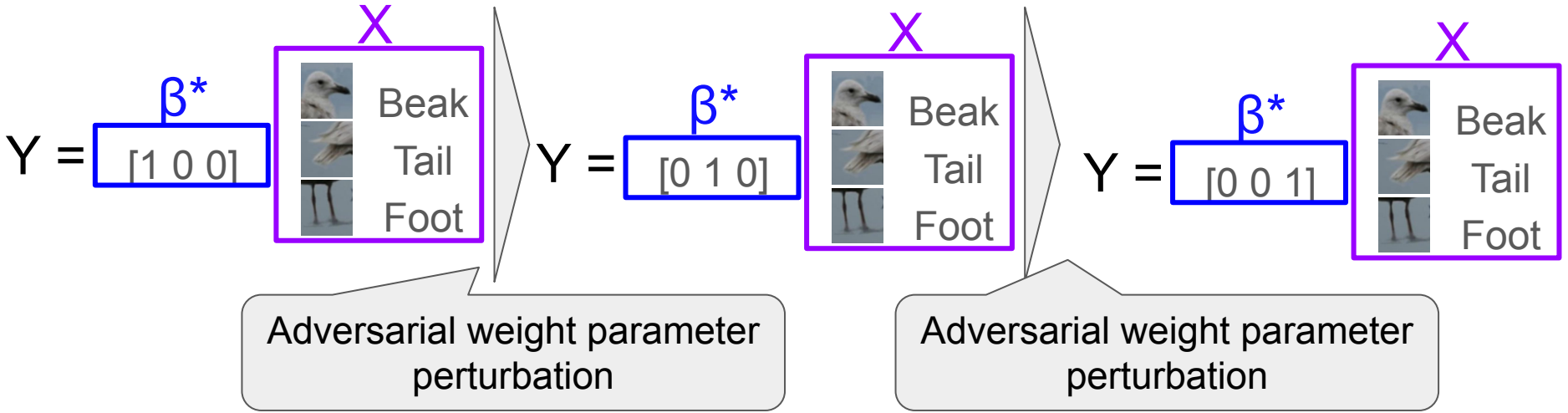
ASGDRO captures common flat local minima across diverse invariant features

Interpretation of ASGDRO

ASGDRO (Ours)

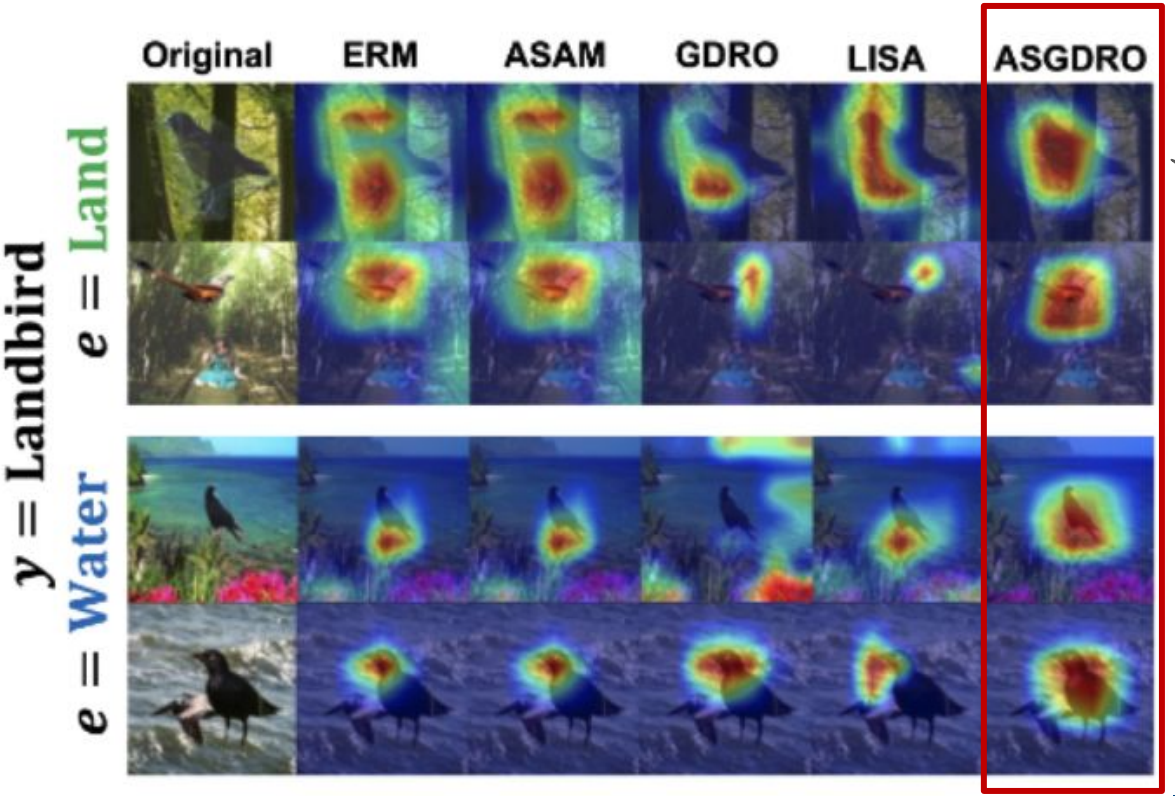
$$\max_{e \in \mathcal{E}_{tr}} \max_{||\epsilon_e|| \leq \rho} \mathcal{R}^e(\theta + \epsilon_e)$$

Minimize each adversarial loss of environment



Results

Task: Landbird vs. Waterbird



ERM is prone to capture unrelated areas (background, spurious correlation)

ASGDRO captures diverse invariant features

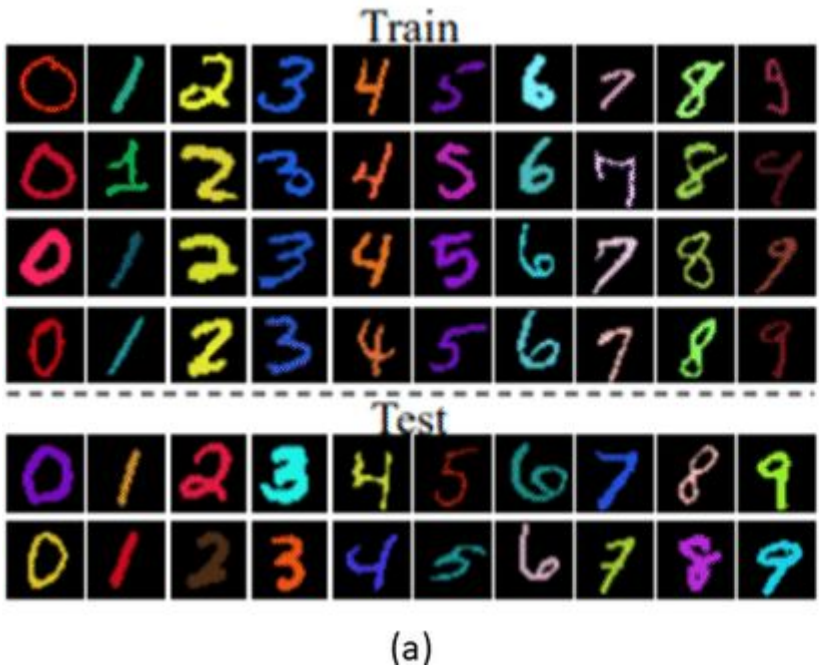
A new dataset for verifying if the model captures sufficient invariant features. (There was no related benchmark dataset)



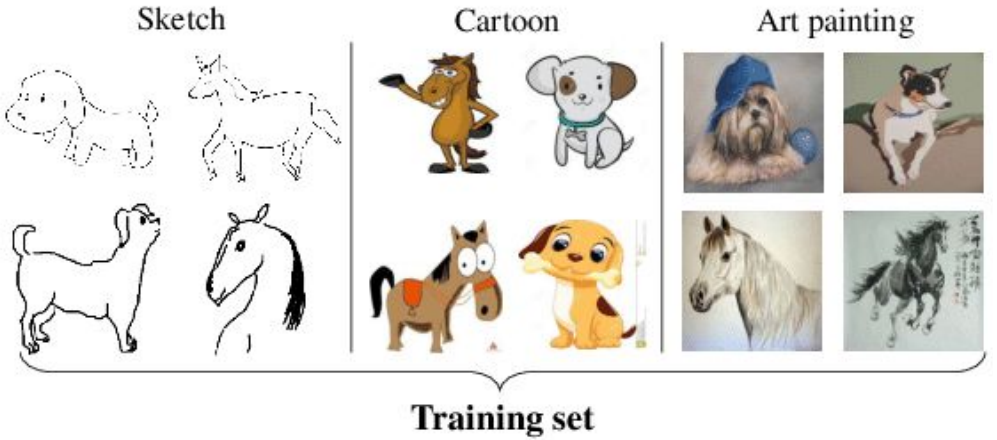
Taero Kim, Ph.D. Student
@ Yonsei University

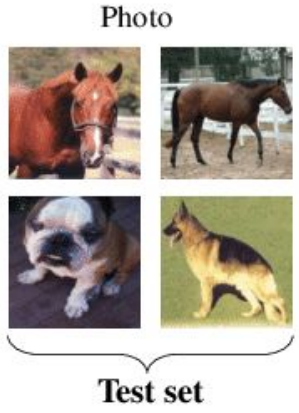
Invariant Learning Perspective
- Sufficient Invariant Benchmark

Sufficient Invariant Learning Benchmark



(a)



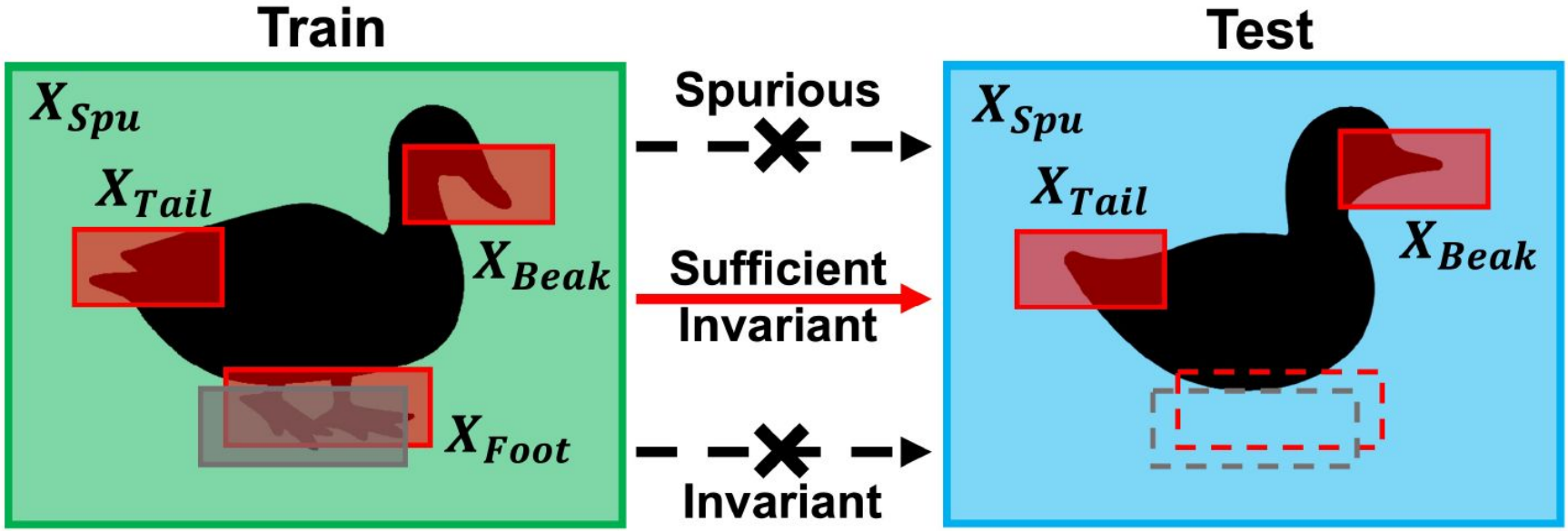


There are diverse benchmark dataset research for distribution shift

Color spurious correlation
Style Shift, …

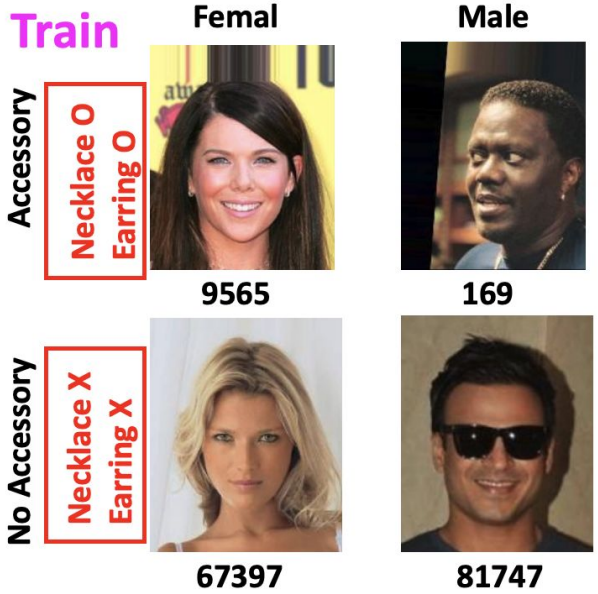
Sufficient Invariant Learning Benchmark

We propose a new benchmark dataset
- to validate whether the models can capture diverse invariant features.



Sufficient Invariant Learning Benchmark

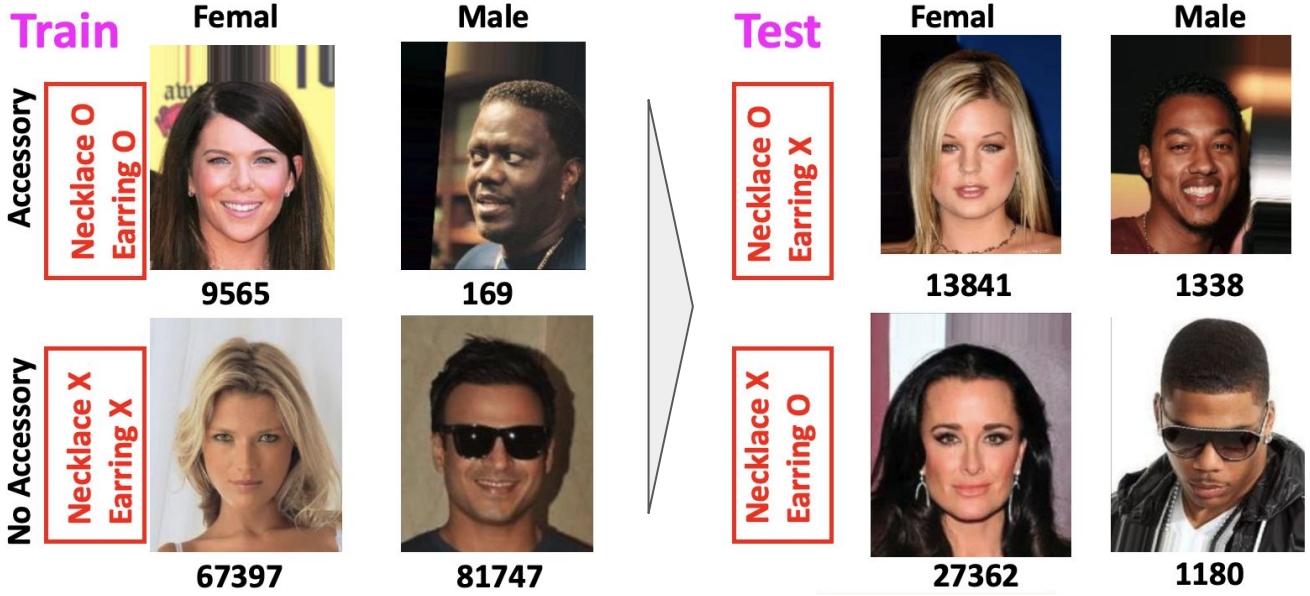
Example) Task: accessory wearing status (binary classification)



(Necklace X, Earring O)
(Necklace O, Earring X)

Sufficient Invariant Learning Benchmark

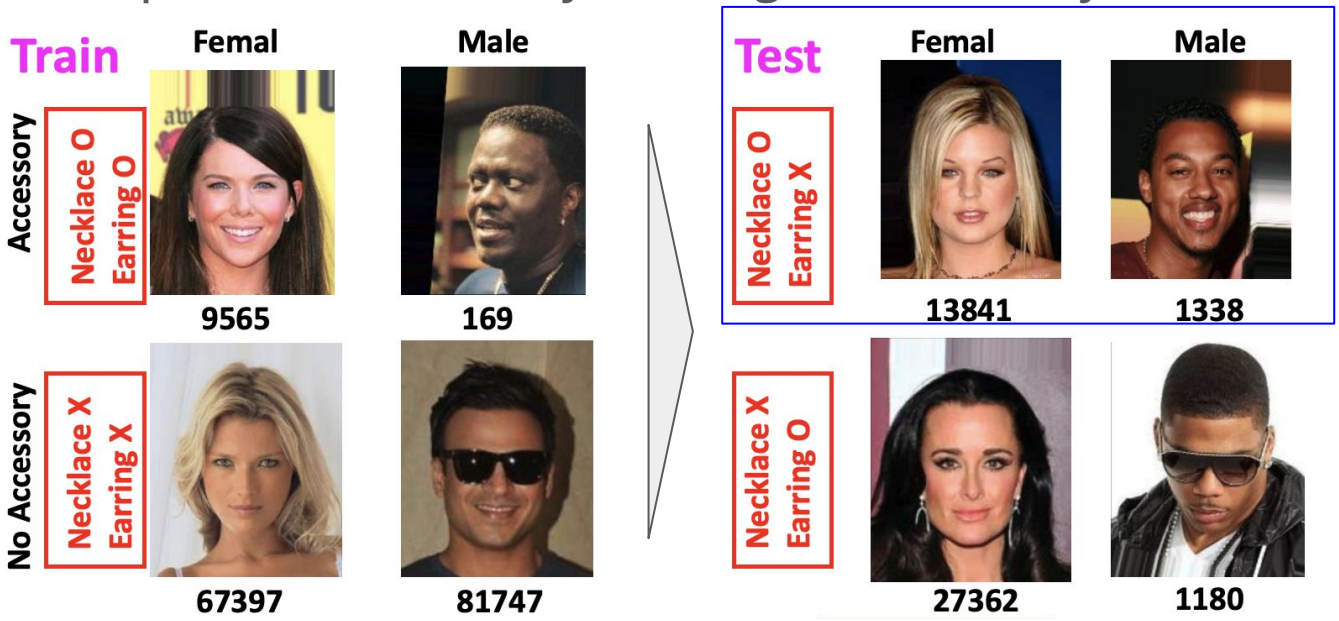
Example) Task: accessory wearing status (binary classification)



(Necklace X, Earring O)
(Necklace O, Earring X)

Sufficient Invariant Learning Benchmark

Example) Task: accessory wearing status (binary classification)

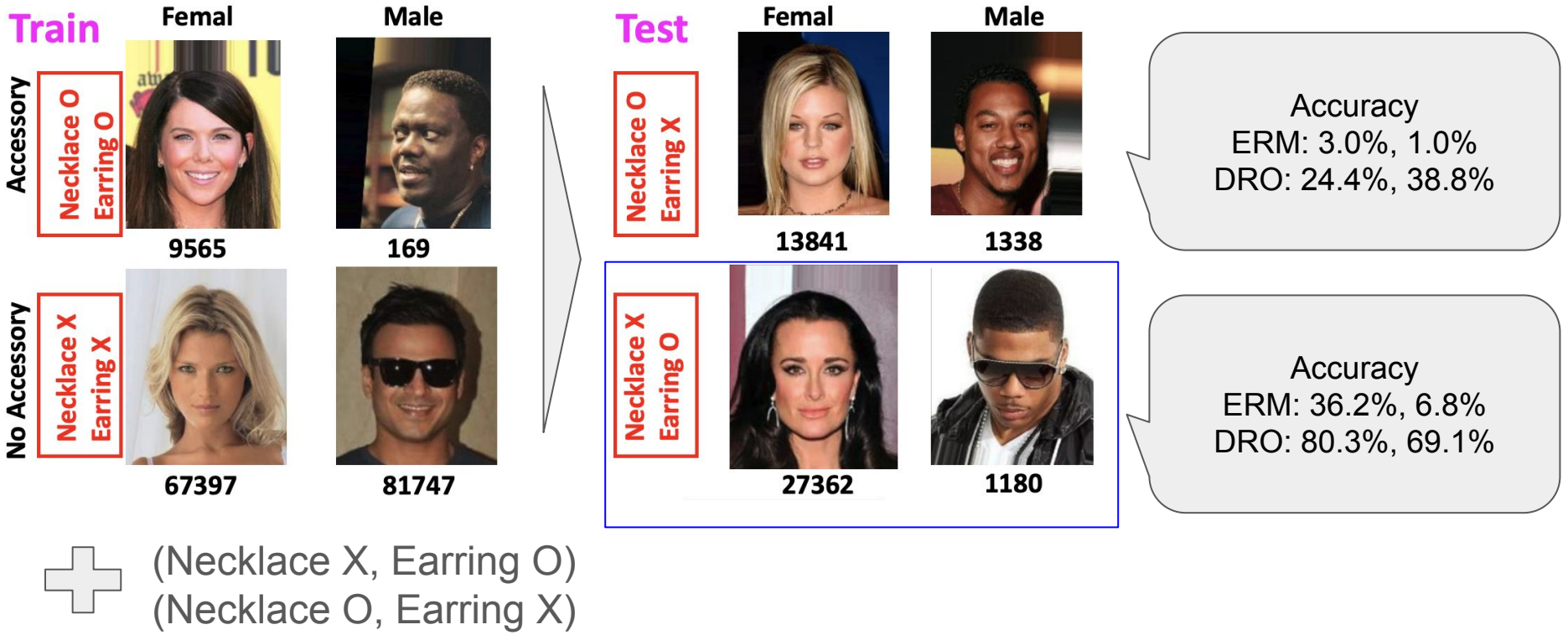


Accuracy
ERM: 3.0%, 1.0%
DRO: 24.4%, 38.8%

(Necklace X, Earring O)
(Necklace O, Earring X)

Sufficient Invariant Learning Benchmark

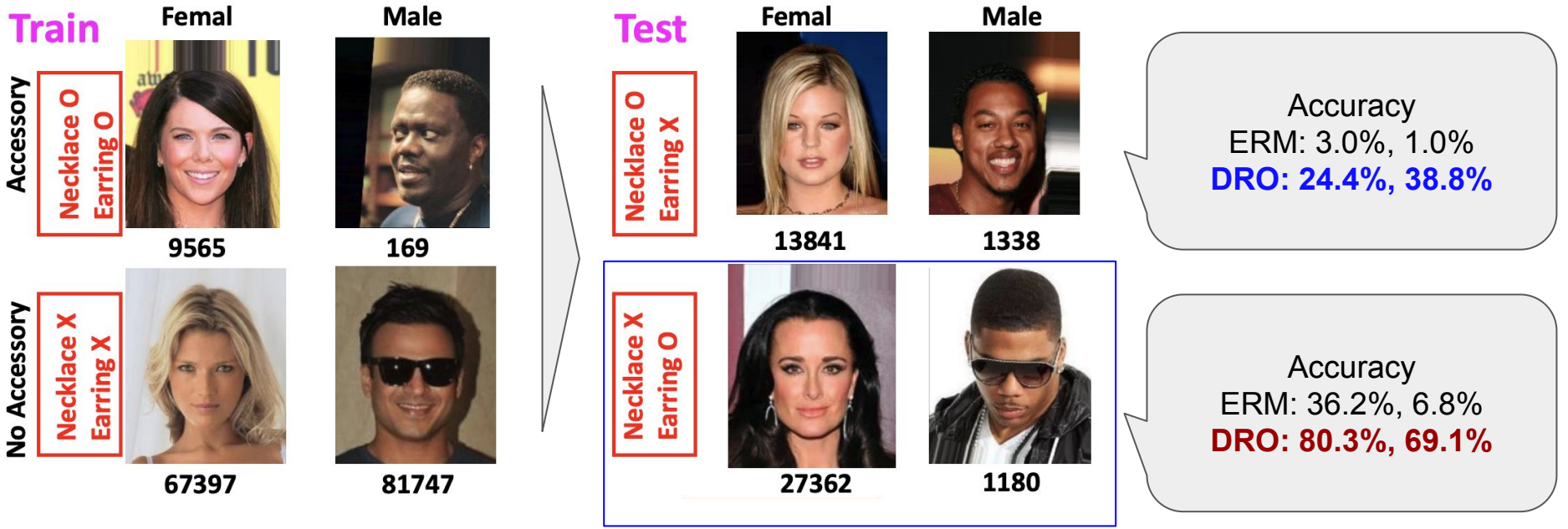
Example) Task: accessory wearing status (binary classification)



(Necklace X, Earring O)
(Necklace O, Earring X)

Sufficient Invariant Learning Benchmark

Example) Task: accessory wearing status (binary classification)



DRO captures **"single"** invariant features (Earing)

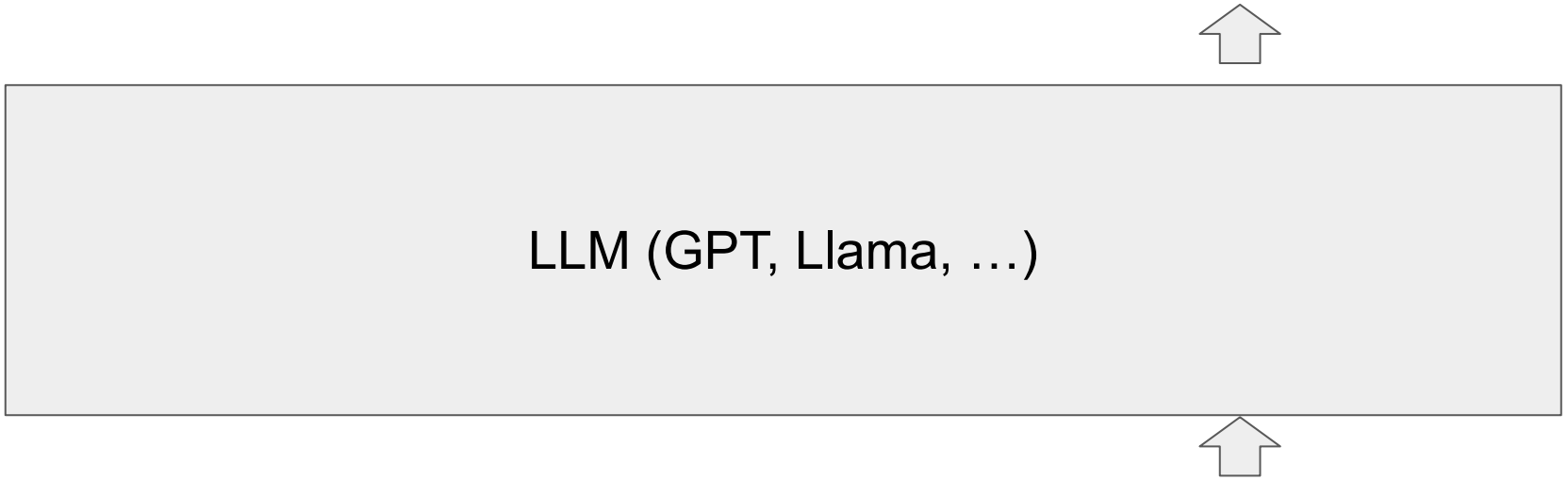
Could invariance/causality improve LLM performance?



Hoyoon Byun, Ph.D. Student
@ Yonsei University

Invariant Learning Perspective - Causal In-Context Learning





Janet’s ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers' market daily for $2 per fresh duck egg. How much in dollars does she make every day at the farmers' market?

Background) In-Context Learning



LLM (GPT, Llama, …)

"민수의 나무에서는 하루에 20개의 과일이 열립니다. 그는 매일 아침 5개를 아침 식사로 먹고, 친구들에게 줄 잼을 만드는 데 6개를 사용합니다. 남은 과일은 매일 직거래 장터에서 신선한 과일 하나당 3달러에 판매합니다. 그는 직거래 장터에서 매일 얼마를 벌까요?"

27

Janet’s ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers' market daily for $2 per fresh duck egg. How much in dollars does she make every day at the farmers' market?

Background) In-Context Learning





Janet’s ducks lay 16 eggs per day. She feels happy eating three eggs for breakfast every morning, and she feels fulfilled baking muffins for her friends with four. She earns extra income by selling the remaining eggs at the farmers' market, which brings her satisfaction. Given this, would Janet's feelings throughout the day be positive or negative?

Positive

Janet’s ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers' market daily for $2 per fresh duck egg. How much in dollars does she make every day at the farmers' market?

Background) In-Context Learning

"민수의 나무에서는 하루에 20개의 과일이 열립니다. 그는 매일 아침 5개를 아침 식사로 먹고, 친구들에게 줄 잼을 만드는 데 6개를 사용합니다. 남은 과일은 매일 직거래 장터에서 신선한 과일 하나당 3달러에 판매합니다. 그는 직거래 장터에서 매일 얼마를 벌까요?"

27

Janet's ducks lay 16 eggs per day. She feels happy eating three eggs for breakfast every morning, and she feels fulfilled baking muffins for her friends with four. She earns extra income by selling the remaining eggs at the farmers' market, which brings her satisfaction. Given this, would Janet's feelings throughout the day be positive or negative?

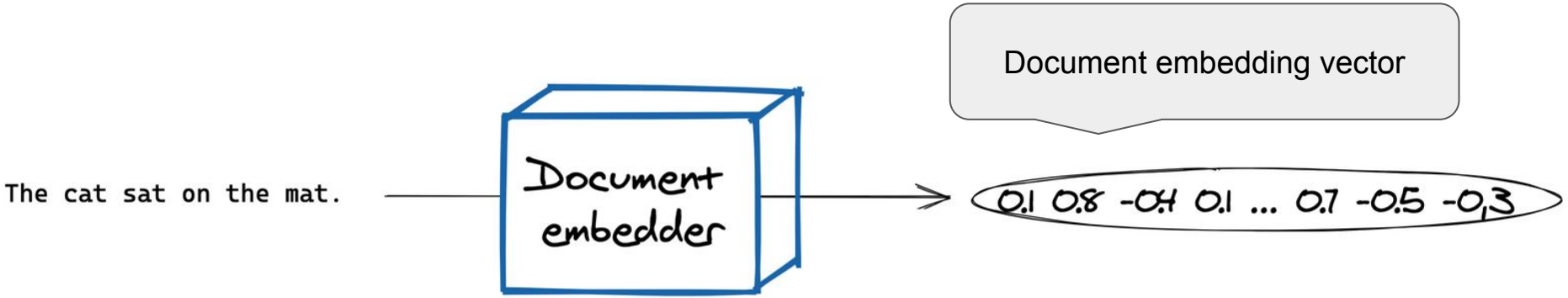
Positive



Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers' market daily for $2 per fresh duck egg. How much in dollars does she make every day at the farmers' market?

Sentence/paragraph/document embedding is a crucial problem.

- Transform the text to embedding vector
- Calculate the similarity between embedding vector
- Retrieve the relevant document
- RAG (Retrieval Augmented Generation), In-Context Learning (ICL), …



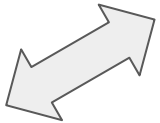
Background) In-Context Learning

"민수의 나무에서는 하루에 20개의 과일이 열립니다. 그는 매일 아침 5개를 아침 식사로 먹고, 친구들에게 줄 잼을 만드는 데 6개를 사용합니다. 남은 과일은 매일 직거래 장터에서 신선한 과일 하나당 3달러에 판매합니다. 그는 직거래 장터에서 매일 얼마를 벌까요?"

27

Janet’s ducks lay 16 eggs per day. She feels happy eating three eggs for breakfast every morning, and she feels fulfilled baking muffins for her friends with four. She earns extra income by selling the remaining eggs at the farmers' market, which brings her satisfaction. Given this, would Janet's feelings throughout the day be positive or negative?

Positive



Traditional Embedding Model

Janet’s ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers' market daily for $2 per fresh duck egg. How much in dollars does she make every day at the farmers' market?

Background) In-Context Learning

"민수의 나무에서는 하루에 20개의 과일이 열립니다. 그는 매일 아침 5개를 아침 식사로 먹고, 친구들에게 줄 잼을 만드는 데 6개를 사용합니다. 남은 과일은 매일 직거래 장터에서 신선한 과일 하나당 3달러에 판매합니다. 그는 직거래 장터에서 매일 얼마를 벌까요?"

27

Janet’s ducks lay 16 eggs per day. She feels happy eating three eggs for breakfast every morning, and she feels fulfilled baking muffins for her friends with four. She earns extra income by selling the remaining eggs at the farmers' market, which brings her satisfaction. Given this, would Janet's feelings throughout the day be positive or negative?

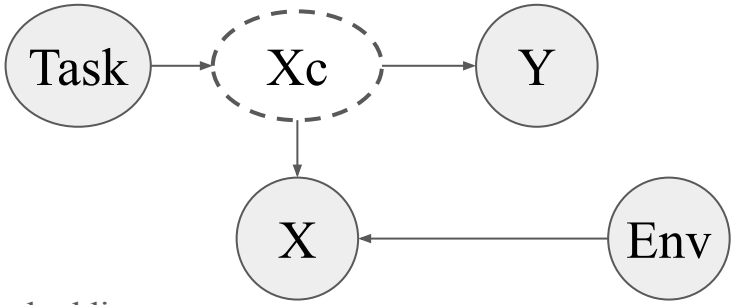
Positive

Ideal Case) Causal context

Janet’s ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers' market daily for $2 per fresh duck egg. How much in dollars does she make every day at the farmers' market?

Causal Text Embedding

Invariant Learning Perspective



Text embeddings:
We use x, y, e, t as text embeddings extracted from LMs c(Xc) is the latent vector we want to represent

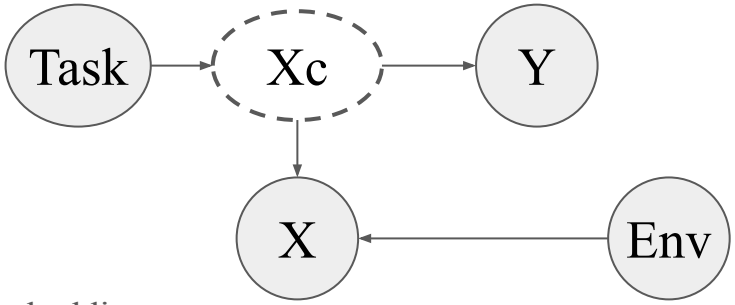
Task: Math → Xc = (x-y-z)*n → Y: $18

Env: Language = English

X: Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers' market daily for $2 per fresh duck egg. How much in dollars does she make every day at the farmers' market?

Causal Text Embedding

Causal Graph
- Task: Math Problems (or sentiment classification)
- X: Plain Text
- Y: Answer
- Env: Language (or spurious correlation)
- Xc: Causal Feature (Embedding)



Text embeddings:
We use x, y, e, t as text embeddings extracted from LMs c(Xc) is the latent vector we want to represent

Accuracy
- Thai (th)
- Telugu language (India)
- Bengali language (India)

Embedding model	Backbone model	ICL	Thai (th)	Telugu (te)	Bengali (bn)
OPENAI small Embedding	GPT-4o-mini	ZS (0-shot)	0.880	0.752	0.846
		ICL (3-shot)	0.912	0.808	0.900
		CCL (3-shot)	**0.936**	**0.884**	**0.904**

Causality can be utilized to improve the LLM performance!

Transfer Learning Perspective
- We need a new metric for OOD.

Bayesian Perspective
- Bayesian interpretation of well-established technique provides a new direction.

Invariant Learning Perspective
- We need to focus on sufficiently diverse invariant features
- Invariant/Causality enhances LLM performance

All of these are done by students, and they are actively looking for internship or visiting researcher opportunities.

Appendix

Conclusion



Kyungwoo Song
Assistant Professor



Taero Kim
'23 Integrated Ph.D. Course



Hoyoon Byun
'24 Ph.D. Course



SungJun Lim
'24 Ph.D. Course



Soojeong Lee
'24 Integrated Ph.D. Course



Jinho Kang
'23 Master Course



Jeyoon Yeom
'23 Master Course



Hakyung Lee
'23 Master Course



Jihee Kim
'23 Master Course



Joonseong Kang
'24 Master Course



Subeen Park
'24 Master Course



Gyeongdeok Seo
'24 Master Course



Youngjun Choi
'24 Master Course



Seonggyun Lee
'24 Master Course



Sumin Park
'24 Integrated BS/MS Program



Minhoi Park
Researcher

Alumni



Yewon Kim
(MS 2024)
Researcher @ ADD (Agency for Defense Development)



Yongtaek Lim
(MS 2024)
ML/AI Engineer @ Bigglz



Changdae Oh
(MS 2024)
Ph.D. Program @ University of Wisconsin–Madison



Sooyon Kim
(MS 2024, Co-advise)
Ph.D. Program @ Ohio State University



Joo-Wang Kim
(MS 2024)
ML/AI Engineer @ LG CNS

https://mlai.yonsei.ac.kr/

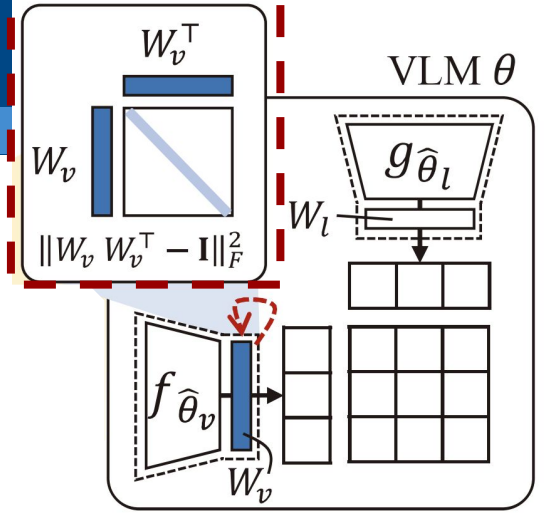
kyungwoo.song@yonsei.ac.kr

Methodology

Towards Calibrated Robust Fine-Tuning

$$x \xrightarrow{W} \tilde{x}^{(1)}$$

$$x \xrightarrow{O} \tilde{x}^{(2)}$$



$$\begin{aligned}
\mathrm{rank}(\tilde{x}^{(1)}\tilde{x}^{(1)T}) &= \mathrm{rank}(\tilde{x}^{(1)}) \\
&= \mathrm{rank}(Wx^T) \\
&\leq \mathrm{rank}(Ox^T) \\
&= \mathrm{rank}(\tilde{x}^{(2)}) \\
&= \mathrm{rank}(\tilde{x}^{(2)}\tilde{x}^{(2)T})
\end{aligned}$$

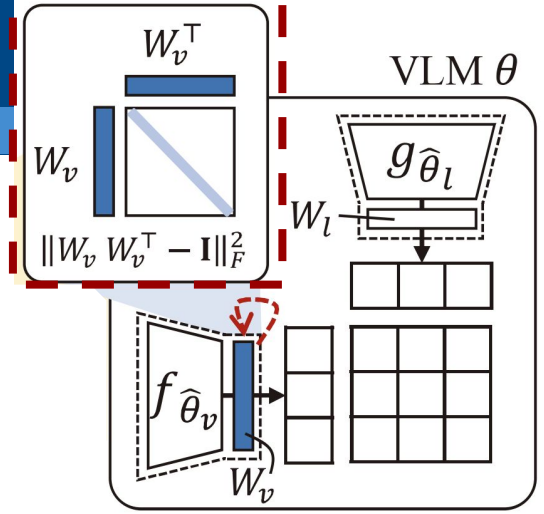
x : Embedding matrix (# Data * Dim)

$\tilde{x}^{(1)}$: Embedding matrix with projection W

$\tilde{x}^{(2)}$: Embedding matrix with projection O

W : Projection matrix

O : Orthogonal matrix



Strategy) increase the smallest singular value

Traditional Multimodal Contrastive Learning

$$\mathcal{L}_{\text{MCL}}(\theta) = \frac{1}{2N}\sum_{i=1}^{N} -\log \frac{\exp(f_{\theta_v}(I_i)\cdot g_{\theta_l}(T_i))}{\sum_{j=1}^{N}\exp(f_{\theta_v}(I_i)\cdot g_{\theta_l}(T_j))}$$

$$+ \frac{1}{2N}\sum_{i=1}^{N} -\log \frac{\exp(f_{\theta_v}(I_i)\cdot g_{\theta_l}(T_i))}{\sum_{j=1}^{N}\exp(f_{\theta_v}(I_j)\cdot g_{\theta_l}(T_i))} + R(\theta_v, \theta_l),$$

CaRot: Add orthogonality regularization term

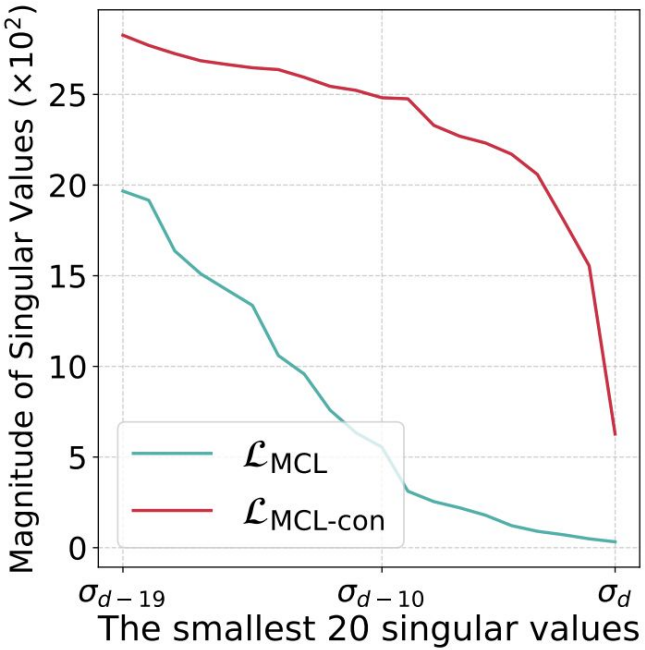
$$\mathcal{L}_{\text{MCL-con}}(\theta) := \mathcal{L}_{\text{MCL}}(\theta) + \lambda_{OC}||W_v W_v^T - \mathbf{I}||_F^2$$

Orthogonal Constraints

OC (orthogonal constraints) enhances OOD performance

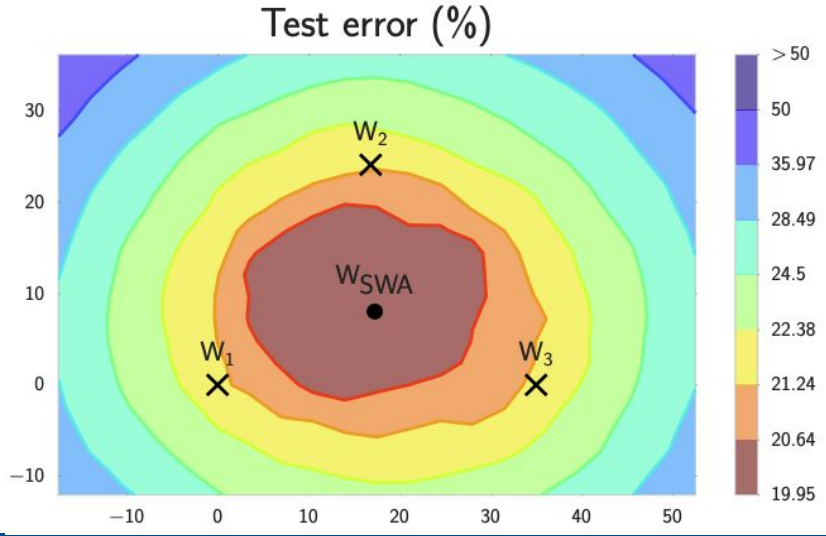
Method	ID		OOD	
	$\sigma_{\min}$ (↑)	ECE (↓)	MSE (↓)	ECE (↓)
Baseline	2.0887	0.1666	0.2581	0.2477
$\mathcal{L}_{\text{OC}}$	4.9630	0.1528	0.1932	0.1781
$\mathcal{L}_{\text{SD}}$	3.1354	0.1308	0.2170	0.1720
$\mathcal{L}_{\text{OC}}$, $\mathcal{L}_{\text{SD}}$	6.5961	0.1391	0.1877	0.1596

OC increases the smallest singular values.

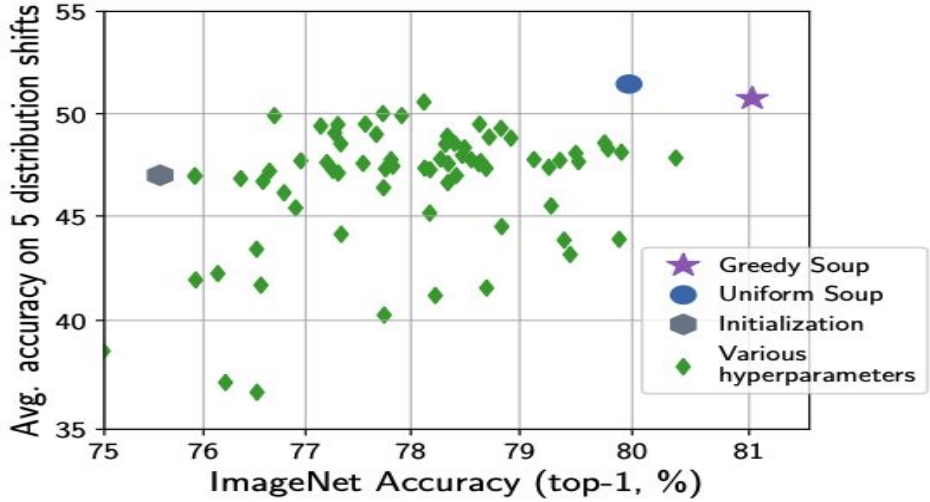


- Model Merging
  - Goal : Combine multiple task-specific models into a single unified model
  - Diverse methods for Model merging

(1) Snapshot of model
([SWA, UAI 2018])

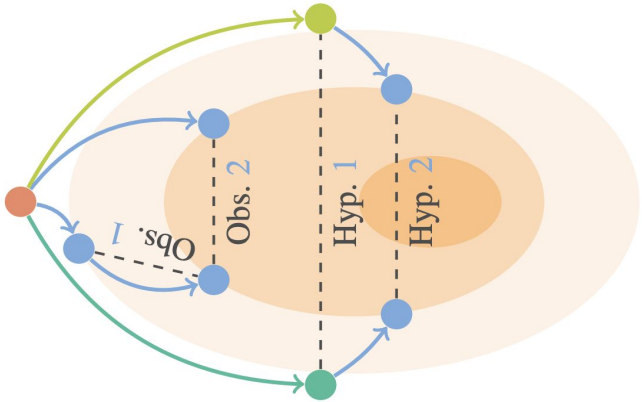


(2) Different Hyperparameters
([Model Soups, ICML 2022])

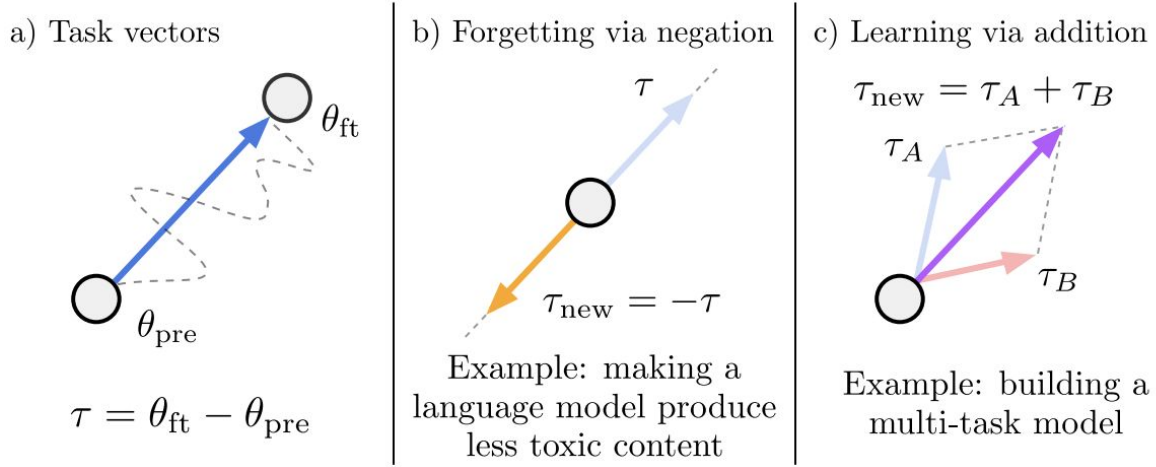


- Model Merging
  - Goal : Combine multiple task-specific models into a single unified model
  - Diverse methods for Model merging

(3) Different Auxiliary Tasks (Model Ratatouille, ICML 2023)



(4) Task Arithmetic (Task Arithmetic, ICLR 2023)



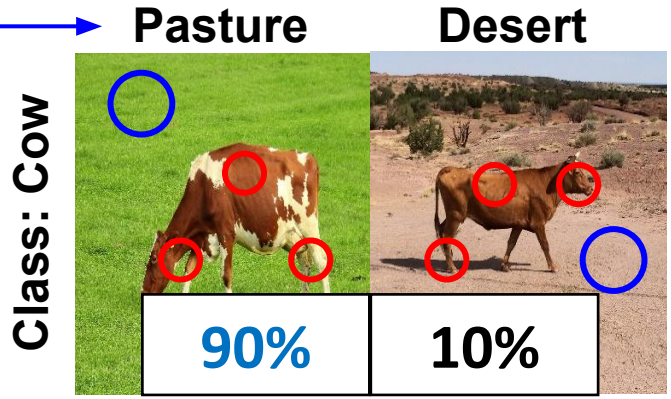
Invariant Learning: Remove Spurious Correlation & Find Invariant Feature

Environment e

○: Spurious

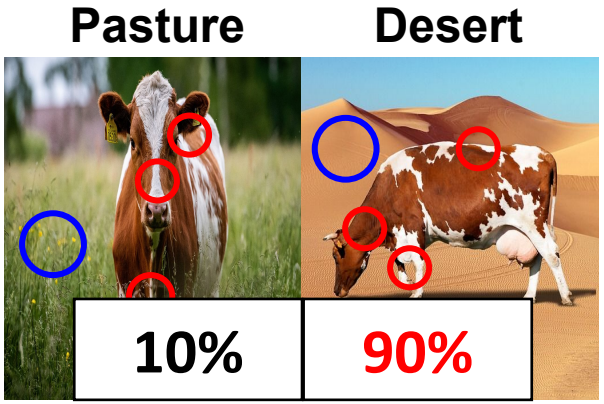
○: Invariant

Z_I: Possible Set of ○



Train

Work well in
i.i.d distribution with Train



Test

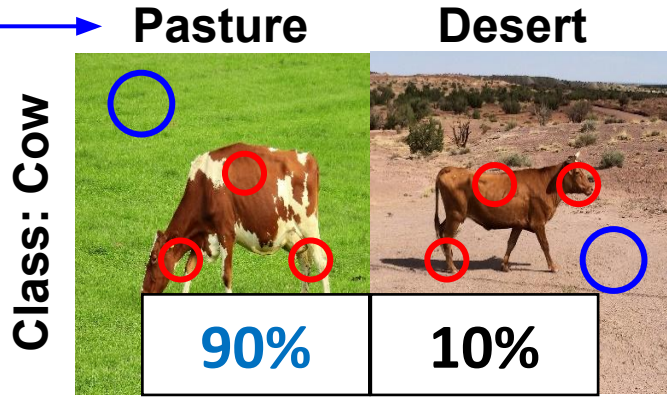
Fail to Predict
(due to spurious correlation)

Invariant Learning: Remove Spurious Correlation & Find Invariant Feature

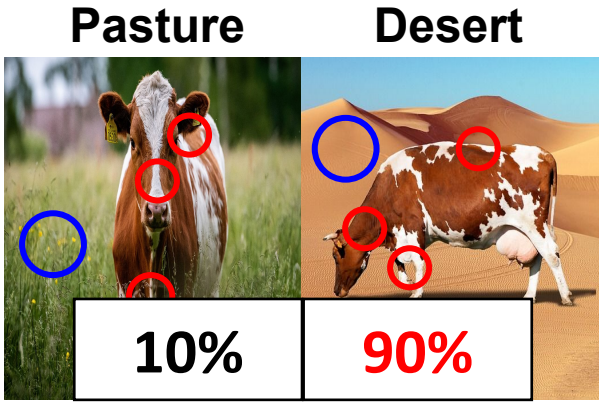
Environment e

○: Spurious

○: Invariant



Train



Test

Z_I: Possible Set of ○

Invariant Condition

$$\mathbb{E}[Y^e|Z_I] = \mathbb{E}[Y^{e'}|Z_I] \quad for\ \forall e, e' \in \mathcal{E}_{tr}$$

Invariant to environment changes

Invariant Learning

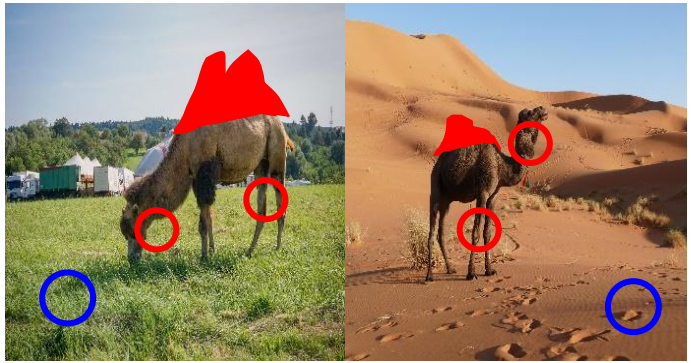
$$\max_{e \in \mathcal{E}_{tr}} \mathcal{R}^e(\theta)$$

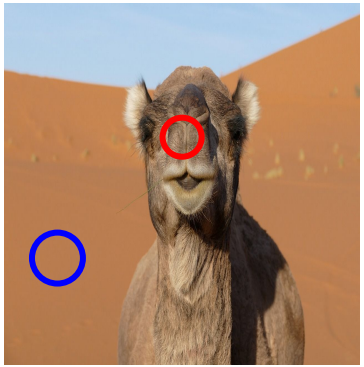
Minimize each environment risk

Z_I is not unique!

Environment e → Pasture, Desert

Desert







: Spurious

: Invariant

: Model Learned

Z_I: Possible Set of

Train

Test

Invariant Condition

$$\mathbb{E}[Y^e|Z_I] = \mathbb{E}[Y^{e'}|Z_I] \quad for\ \forall e, e' \in \mathcal{E}_{tr}$$

Invariant to environment changes

Invariant Learning

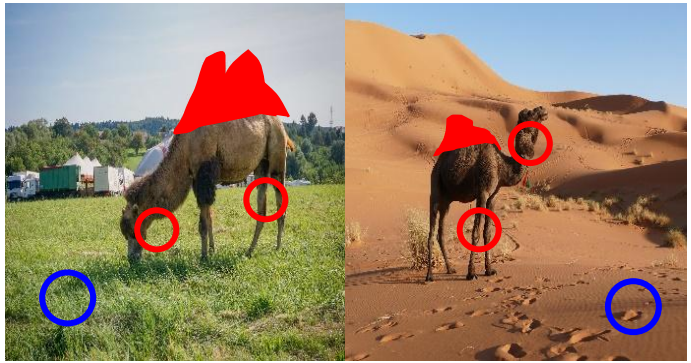
$$\max_{e \in \mathcal{E}_{tr}} \mathcal{R}^e(\theta)$$

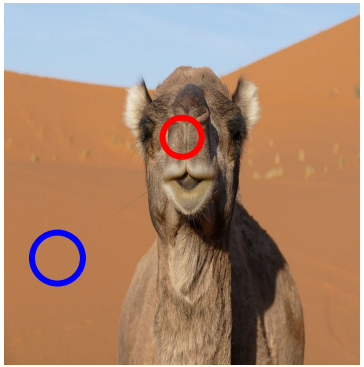
Minimize each environment risk

Z_I is not unique!

Environment e → Pasture, Desert

Desert





: Spurious

: Invariant

: Model Learned

Z_I: Possible Set of

Train

Test

Sufficient Invariant Condition (SIC)

$$\mathbb{E}[Y^e|Z_I] = \mathbb{E}[Y^{e'}|Z_I] \quad for\ \forall e, e' \in \mathcal{E}_{tr} \quad s.t.\ Z_I = Z_{SIC}$$

Maximal Set of Z_I

Sufficient Invariant Learning

Sufficient Invariant Condition (SIC)

$$\mathbb{E}[Y^e|Z_I] = \mathbb{E}[Y^{e'}|Z_I] \quad for\ \forall e, e' \in \mathcal{E}_{tr} \quad s.t.\ Z_I = Z_{SIC}$$

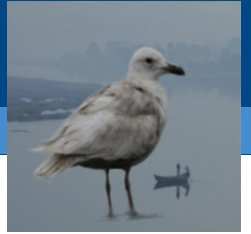
Maximal Set of Z_I

Sufficient Invariant Learning (SIL)

$$\mathcal{R}_{SIL}(\theta) = \max_{Z_I \in \mathcal{Z}_I} \max_{e \in \mathcal{E}_{tr}} \mathcal{R}^e(\theta)$$

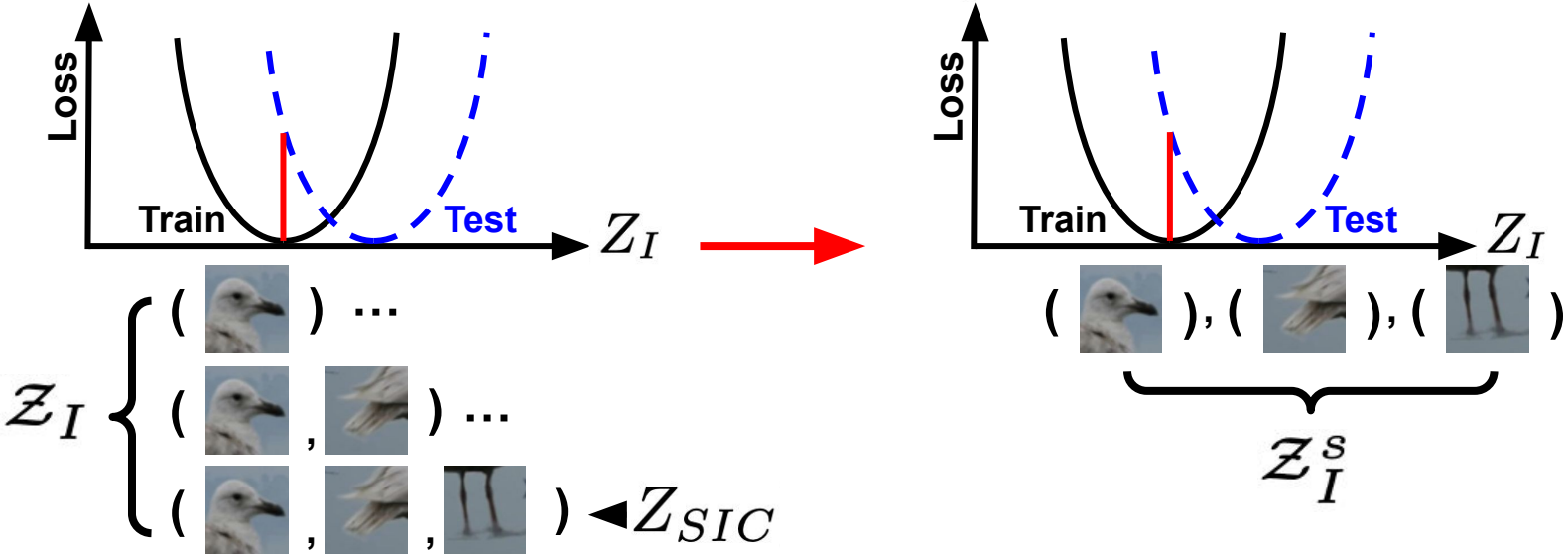
$$= \max_{Z_I \in \mathcal{Z}_I} \mathcal{R}_{Inv}(\theta; Z_I)$$

We should ensure the generalization ability of model for all possible invariant set $\mathcal{Z}_I$



Proposition 1:
Minimizing Generalization Gap of all possible sets of invariant features
= Minimizing Generalization Gap of singleton sets of invariant features

$$\max_{Z_I \in \mathcal{Z}_I} \mathcal{R}_{Inv}(\theta; Z_I) - \hat{\mathcal{R}}_{Inv}(\theta; Z_I) = \max_{Z_I \in \mathcal{Z}_I^s} \mathcal{R}_{Inv}(\theta; Z_I) - \hat{\mathcal{R}}_{Inv}(\theta; Z_I).$$



Adaptive Sharpness-aware Group Distributionally Robust Optimization (ASGDRO)

GDRO (Sagawa et al., ICLR 2019)

$$\max_{e \in \mathcal{E}_{tr}} \mathcal{R}^e(\theta)$$

Minimize
each loss of environment

Optimal GDRO solution captures an invariant feature
(not sufficient invariant features)

source: Rojas-Carulla, Mateo, et al. "Invariant models for causal transfer learning." Journal of Machine Learning Research 19.36 (2018): 1-34. Sagawa, Shiori, et al. "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization." arXiv preprint arXiv:1911.08731 (2019).

Adaptive Sharpness-aware Group Distributionally Robust Optimization (ASGDRO)

GDRO (Sagawa et al., ICLR 2019)

$$\max_{e \in \mathcal{E}_{tr}} \mathcal{R}^e(\theta)$$

Minimize
each loss of environment

ASGDRO (Ours)

$$\max_{e \in \mathcal{E}_{tr}} \max_{||\epsilon_e|| \leq \rho} \mathcal{R}^e(\theta + \epsilon_e)$$

Minimize
each adversarial loss of environment

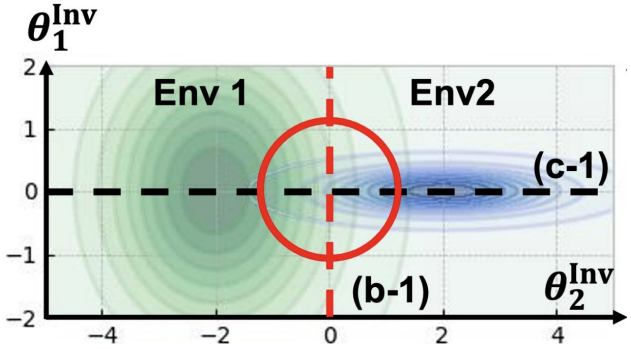
Inspired by (Foret et al., ICLR 2021)

Adaptive Sharpness-aware Group Distributionally Robust Optimization (ASGDRO)

GDRO (Sagawa et al., ICLR 2019)

$$\max_{e \in \mathcal{E}_{tr}} \mathcal{R}^e(\theta)$$

Minimize
each loss of environment



ASGDRO (Ours)

$$\max_{e \in \mathcal{E}_{tr}} \max_{||\epsilon_e|| \leq \rho} \mathcal{R}^e(\theta + \epsilon_e)$$

Minimize
each adversarial loss of environment

Inspired by (Foret et al., ICLR 2021)

$$\mathcal{R}_{\text{ASGDRO}}(\theta) = \max_{e \in \mathcal{E}_{tr}} \mathcal{R}^e(\theta + \epsilon_e^*)$$

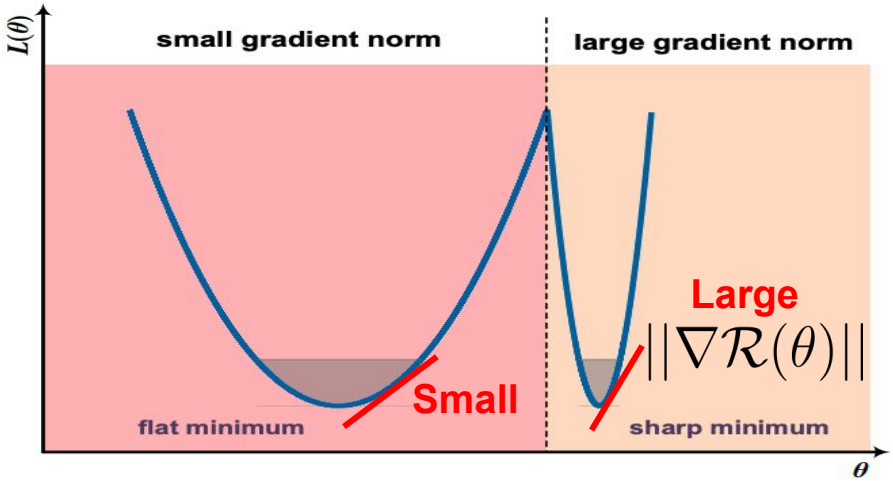
$$\epsilon_e^* = \rho \frac{T_\theta^2 \nabla \mathcal{R}^e(\theta)}{\|T_\theta \nabla \mathcal{R}^e(\theta)\|}$$

Proposition 2:
Minimizing adversarial loss of each environment ◀ $\max_{e\in\mathcal{E}_{tr}} \max_{||\epsilon_e||\leq\rho} \mathcal{R}^e(\theta+\epsilon_e)$
≈ Finding Common Flat minima across environments

$$\max_{e\in\mathcal{E}} \max_{||\epsilon_e||\leq\rho} \mathcal{R}^e(\theta+\epsilon_e) \approx \max_{e\in\mathcal{E}} [\mathcal{R}^e(\theta) + \rho||\nabla\mathcal{R}^e(\theta)||]$$

▲ Environments

▲ Gradient Norm = Flatness



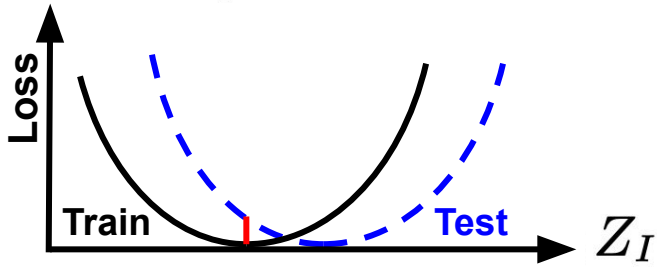
Minimizing sharpness of each environments
= Finding Common Flat Minima across environments

(Zhao et al., ICML 2022)

Theorem 1:
Finding Common Flat Minima
= Learning Balanced Invariant Representation

$$H^* = \arg\min_H \max_{||\epsilon|| \le \rho} \hat{\mathcal{R}}_{Inv}(\theta_h^* + \epsilon; Z_I)$$
$$= \arg\min_H \left[(Y - H\theta_h^*)'(Y - H\theta_h^*) + \rho \| H^T (H\theta_h^* - Y) \| \right]$$
$$= (\frac{1}{k}, \ldots, \frac{1}{k}, 0, \ldots, 0). \quad \blacktriangleleft \; k = |Z_{SIC}|$$

Assumption
$$\begin{cases} H = (h_1, \ldots, h_k, 0, \ldots, 0) \\ \sum_{i=1}^k h_i = 1 \\ h_i > 0 \\ k \le p \end{cases}$$



Feature Importance ► H^* $\frac{1}{3}$ $\frac{1}{3}$ $\frac{1}{3}$ 0

Invariant Feature ► Z_I ( ) ( ) ( ) ( )

Inv Inv Inv Spu

Minimizing $||\nabla \mathcal{R}(\theta)||$