Just Train Twice: Improving Group Robustness without Training Group Information

Evan Zheran Liu*

Annie S. Chen*

Pang Wei Koh

Percy Liang

Behzad Haghgoo*

Aditi Raghunathan

Shiori Sagawa

Chelsea Finn

Problem: Low Worst-Group Performance

Problem: Low Worst-Group Performance

Online comment moderation (Borkan et al., ‘19; Koh et al., ‘20)

Input: real online comment

I applaud your father.
He was a good man!

→

Label: toxicity

toxic   non-toxic

Problem: Low Worst-Group Performance

Online comment moderation (Borkan et al., '19; Koh et al., '20)

92.6% average test accuracy

Input: real online comment

I applaud your father.
He was a good man!

→

Label: toxicity

toxic non-toxic

Problem: Low Worst-Group Performance

Online comment moderation (Borkan et al., '19; Koh et al., '20)

Input: real online comment

I applaud your father.
He was a good man!

Label: toxicity

toxic   non-toxic

92.6% average test accuracy

69.2% on non-toxic comments mentioning Black demographic (Koh et al., '20)

Problem: Low Worst-Group Performance

Online comment moderation (Borkan et al., '19; Koh et al., '20)

Input: real online comment

I applaud your father.
He was a good man!

→

Label: toxicity

toxic non-toxic

92.6% average test accuracy

69.2% on non-toxic comments mentioning Black demographic (Koh et al., '20)

I am a black woman

Problem: Low Worst-Group Performance

Online comment moderation (Borkan et al., '19; Koh et al., '20)

Input: real online comment

I applaud your father. He was a good man!

Label: toxicity

toxic non-toxic

92.6% average test accuracy

69.2% on non-toxic comments mentioning Black demographic (Koh et al., '20)

I am a black woman

Wildlife image classification (Wah et al., '11; Sagawa et al., '20)

Input: image of a bird

Label: bird type

water land

Problem: Low Worst-Group Performance

Online comment moderation (Borkan et al., '19; Koh et al., '20)

Input: real online comment

I applaud your father. He was a good man!

Label: toxicity

toxic non-toxic

92.6% average test accuracy

69.2% on non-toxic comments mentioning Black demographic (Koh et al., '20)

I am a black woman

Wildlife image classification (Wah et al., '11; Sagawa et al., '20)

Input: image of a bird

Label: bird type

water land

97.3% average test accuracy

Problem: Low Worst-Group Performance

Online comment moderation (Borkan et al., '19; Koh et al., '20)

Input: real online comment

I applaud your father.
He was a good man!

Label: toxicity

toxic non-toxic

92.6% average test accuracy

69.2% on non-toxic comments mentioning Black demographic (Koh et al., '20)

I am a black woman

Wildlife image classification (Wah et al., '11; Sagawa et al., '20)

Input: image of a bird

Label: bird type

water land

97.3% average test accuracy

72.6% on waterbirds on land backgrounds

Problem: Low Worst-Group Performance

Online comment moderation (Borkan et al., '19; Koh et al., '20)

Input: real online comment

I applaud your father. He was a good man!

Label: toxicity

toxic non-toxic

92.6% average test accuracy

69.2% on non-toxic comments mentioning Black demographic (Koh et al., '20)

I am a black woman

Wildlife image classification (Wah et al., '11; Sagawa et al., '20)

Input: image of a bird

Label: bird type

water land

97.3% average test accuracy

72.6% on waterbirds on land backgrounds

Problem: Low Worst-Group Performance

Online comment moderation (Borkan et al., '19; Koh et al., '20)

Input: real online comment

I applaud your father. He was a good man!

→

Label: toxicity

toxic non-toxic

92.6% average test accuracy

69.2% on non-toxic comments mentioning Black demographic (Koh et al., '20)

I am a black woman

Standard training can **perform poorly** on worst group, especially if there are **spurious correlations**

Wildlife image classification (Wah et al., '11; Sagawa et al., '20)

Input: image of a bird

→

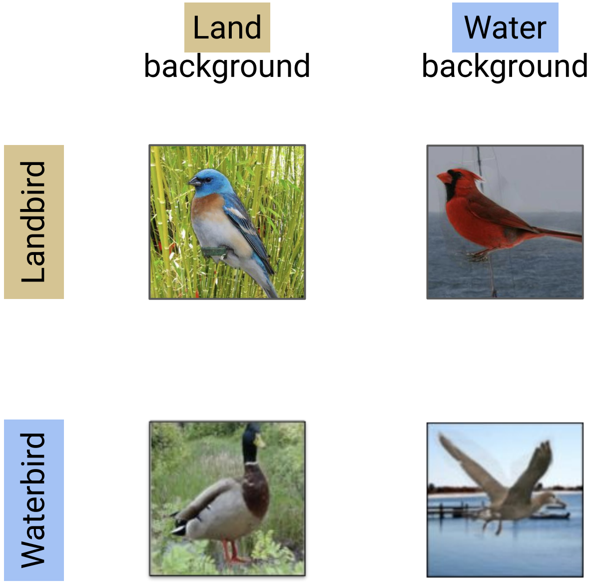
Label: bird type

water land

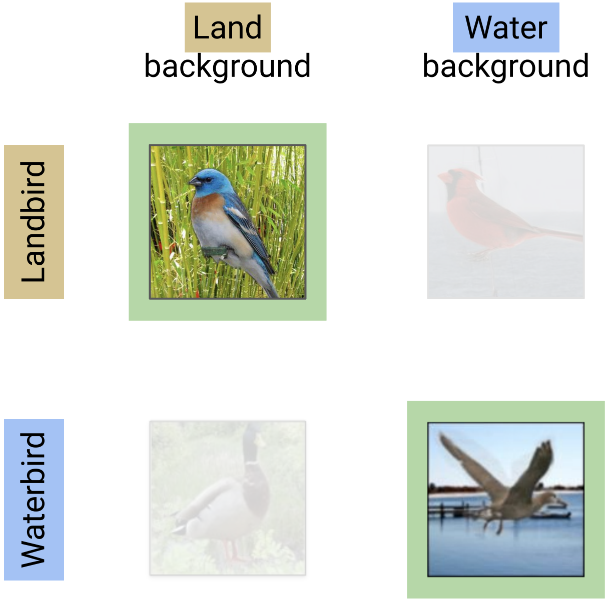
97.3% average test accuracy

72.6% on waterbirds on land backgrounds

Problem: Low Worst-Group Performance

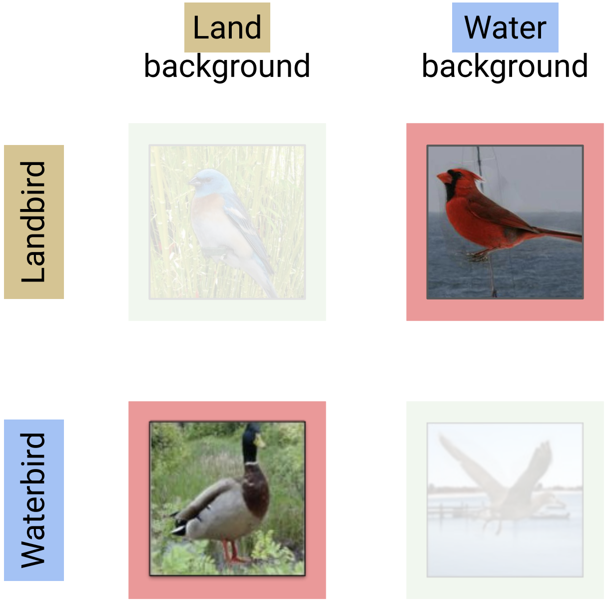


Problem: Low Worst-Group Performance



High data + spurious correlation **holds**

Problem: Low Worst-Group Performance



High data + spurious correlation **holds**

Low data + spurious correlation **doesn’t hold**

Problem: Low Worst-Group Performance

High data + spurious correlation **holds**

Low data + spurious correlation **doesn’t hold**

Goal: Perform well across all groups

Prior Work: Leveraging Group Information

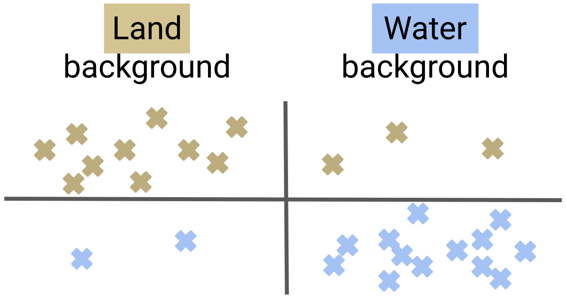
Standard training: Empirical risk minimization (ERM)

$$J_{\text{ERM}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, y_i; \theta)$$

Prior Work: Leveraging Group Information

Standard training: Empirical risk minimization (ERM)

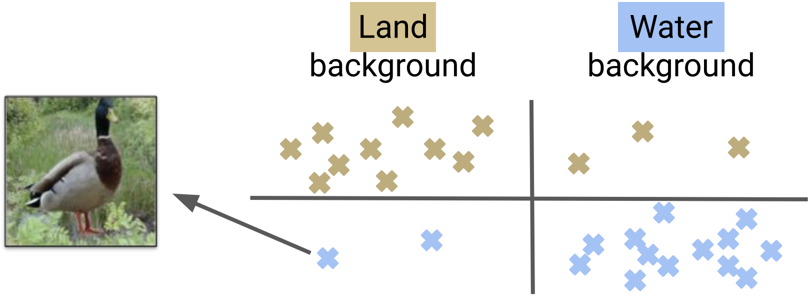
$$J_{\text{ERM}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(x_i, y_i; \theta)$$



Prior Work: Leveraging Group Information

Standard training: Empirical risk minimization (ERM)

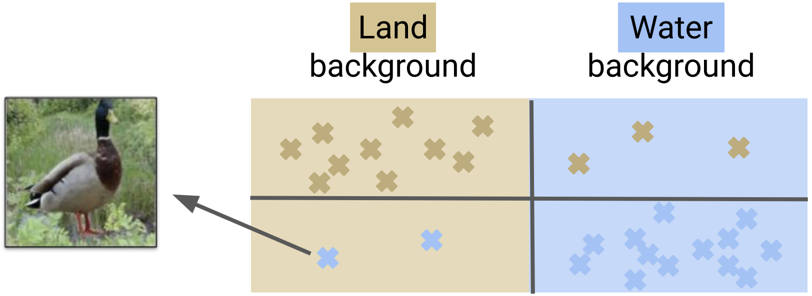
$$J_{\text{ERM}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(x_i, y_i; \theta)$$



Prior Work: Leveraging Group Information

Standard training: Empirical risk minimization (ERM)

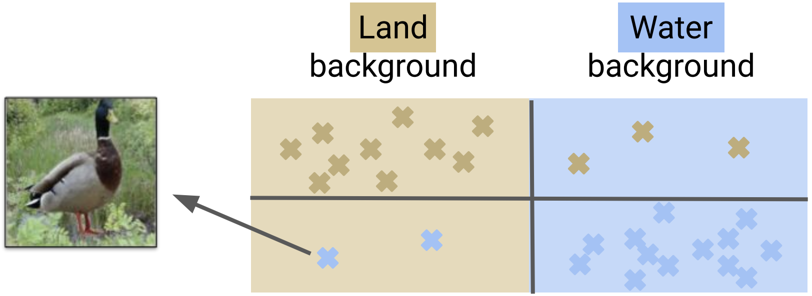
$$J_{\text{ERM}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(x_i, y_i; \theta)$$



Prior Work: Leveraging Group Information

Standard training: Empirical risk minimization (ERM)

$$J_{\text{ERM}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, y_i; \theta)$$



Does not generalize ⇒ **low worst-group performance**

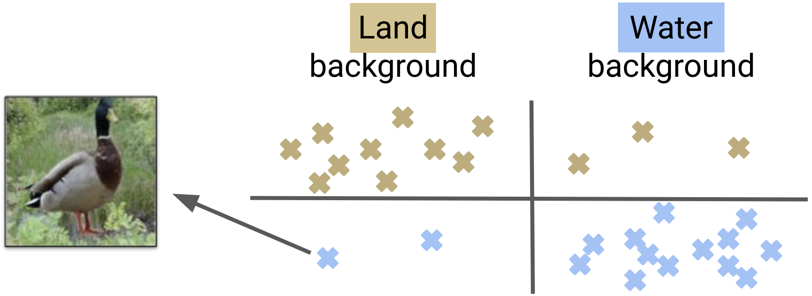
Prior Work: Leveraging Group Information

Standard training: Empirical risk minimization (ERM)

$$J_{\mathrm{ERM}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(x_i, y_i; \theta)$$

Does not generalize ⇒ low worst-group performance

Group reweighting (Shimodaira ‘00; Sagawa et al., ‘20; Byrd & Lipton ‘18)



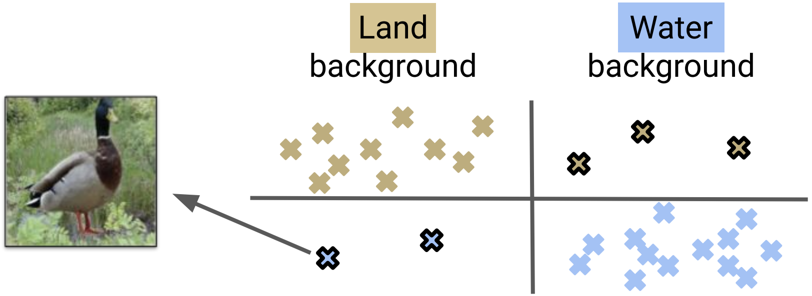
Prior Work: Leveraging Group Information

Standard training: Empirical risk minimization (ERM)

$$J_{\mathrm{ERM}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(x_i, y_i; \theta)$$

Does not generalize ⇒ **low worst-group performance**

Group reweighting (Shimodaira ‘00; Sagawa et al., ‘20; Byrd & Lipton ‘18)



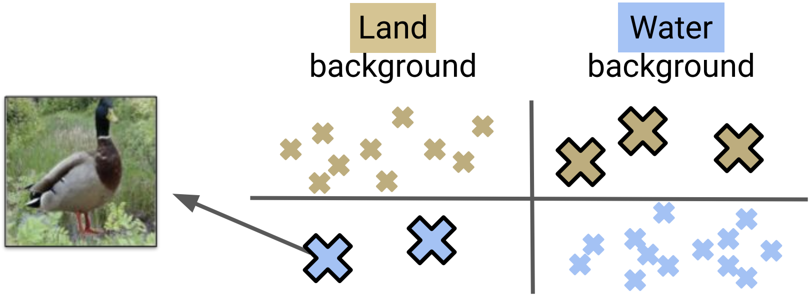
Prior Work: Leveraging Group Information

Standard training: Empirical risk minimization (ERM)

$$J_{\text{ERM}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(x_i, y_i; \theta)$$

Does not generalize ⇒ **low worst-group performance**

Group reweighting (Shimodaira '00; Sagawa et al., '20; Byrd & Lipton '18)



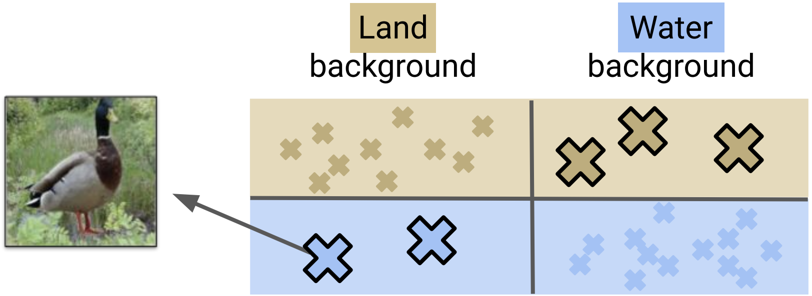
Prior Work: Leveraging Group Information

Standard training: Empirical risk minimization (ERM)

$$J_{\text{ERM}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, y_i; \theta)$$

Does not generalize ⇒ **low worst-group performance**

Group reweighting (Shimodaira ‘00; Sagawa et al., ‘20; Byrd & Lipton ‘18)



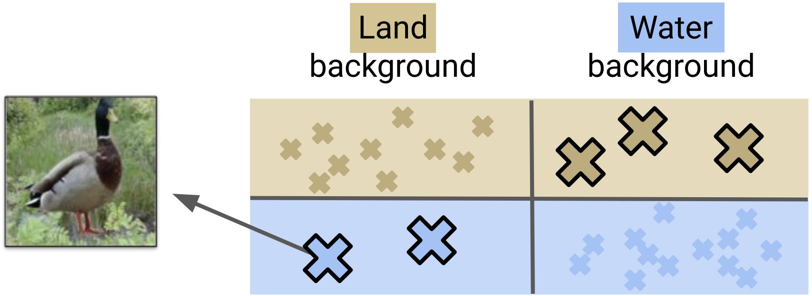
Prior Work: Leveraging Group Information

Standard training: Empirical risk minimization (ERM)

$$J_{\mathrm{ERM}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(x_i, y_i; \theta)$$

Does not generalize ⇒ **low worst-group performance**

Group reweighting (Shimodaira ‘00; Sagawa et al., ‘20; Byrd & Lipton ‘18)



Generalizes ⇒ **high worst-group performance**

Prior Work: Leveraging Group Information

Standard training: Empirical risk minimization (ERM)

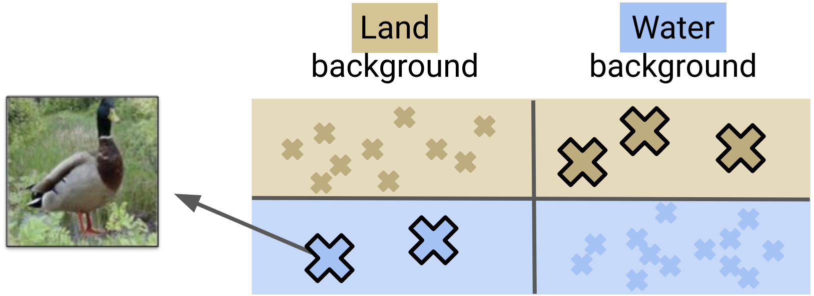
$$J_{\text{ERM}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, y_i; \theta)$$

Does not generalize ⇒ **low worst-group performance**

Group reweighting (Shimodaira ‘00; Sagawa et al., ‘20; Byrd & Lipton ‘18)

Group DRO (Sagawa et al., ‘20):
Minimize worst-group loss

$$J_{\text{gDRO}}(\theta) = \max_g \frac{1}{n_g} \sum_{i|g_i=g} \ell(x_i, y_i; \theta)$$



Generalizes ⇒ **high worst-group performance**

Prior Work: Leveraging Group Information

Standard training: Empirical risk minimization (ERM)

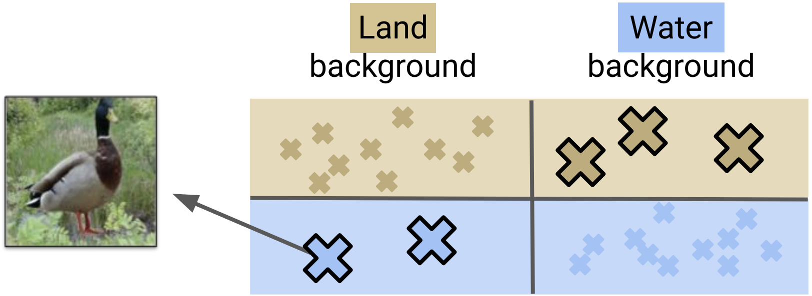
$$J_{\text{ERM}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, y_i; \theta)$$

Does not generalize ⇒ **low worst-group performance**

Group reweighting (Shimodaira '00; Sagawa et al., '20; Byrd & Lipton '18)

Group DRO (Sagawa et al., '20):
Minimize worst-group loss

$$J_{\text{gDRO}}(\theta) = \max_g \frac{1}{n_g} \sum_{i|g_i=g} \ell(x_i, y_i; \theta)$$



Generalizes ⇒ **high worst-group performance**
...but requires **expensive training group annotations**

Prior Work: Leveraging Group Information

Standard training: Empirical risk minimization (ERM)

$$J_{\text{ERM}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(x_i, y_i; \theta)$$

Does not generalize ⇒ **low worst-group performance**

Goal: high worst-group performance *without* training group annotations

Group reweighting (Shimodaira '00; Sagawa et al., '20; Byrd & Lipton '18)

Group DRO (Sagawa et al., '20):
Minimize worst-group loss

$$J_{\text{gDRO}}(\theta) = \max_g \frac{1}{n_g}\sum_{i|g_i=g} \ell(x_i, y_i; \theta)$$

Generalizes ⇒ **high worst-group performance**
...but requires **expensive training group annotations**

Setting: No Training Group Annotations

Training
group annotations

Train
data

waterbird

waterbird
on land

Input

Label

Group

Setting: No Training Group Annotations

Training
group annotations

Train
data

waterbird

waterbird
on land

Input

Label

Group

Used by reweighting approaches

Setting: No Training Group Annotations

Training
group annotations

Train
data

waterbird

waterbird
on land

Input

Label

Group

No training
group annotations
(our setting)

waterbird

Input

Label

Group not known

Setting: No Training Group Annotations

Training
group annotations

Train
data

waterbird

waterbird
on land

Input

Label

Group

No training
group annotations
(our setting)

waterbird

Input

Label

Group not known

Lam & Zhou, '15; Ren et al., '18; Oren et al., '19; Sohoni et al., '20; Kim et al., '19;
Shu et al., '19; Pezeshki et al., '20.

Setting: No Training Group Annotations

Training
group annotations

Train data

waterbird

waterbird on land

Input | Label | Group

Valid data

waterbird

waterbird on land

Input | Label | Group

No training
group annotations
(our setting)

Train data

waterbird

Input | Label

Group not known

Valid data

waterbird

waterbird on land

Input | Label | Group

JTT: Just Train Twice

Observation 1: Upweighting worst group yields high worst-group performance (Shimodaira ‘00; Sagawa et al., ‘20; Byrd & Lipton, ‘18)

JTT: Just Train Twice

Observation 1: Upweighting worst group yields high worst-group performance (Shimodaira '00; Sagawa et al., '20; Byrd & Lipton, '18)

Observation 2: ERM performs poorly on worst group

JTT: Just Train Twice

Observation 1: Upweighting worst group yields high worst-group performance (Shimodaira '00; Sagawa et al., '20; Byrd & Lipton, '18)

Observation 2: ERM performs poorly on worst group

Stage 1: Automatically identify worst-group training examples

JTT: Just Train Twice

Observation 1: Upweighting worst group yields high worst-group performance (Shimodaira '00; Sagawa et al., '20; Byrd & Lipton, '18)

Observation 2: ERM performs poorly on worst group

Stage 1: Automatically identify worst-group training examples

Stage 2: Upweight identified examples

JTT: Just Train Twice

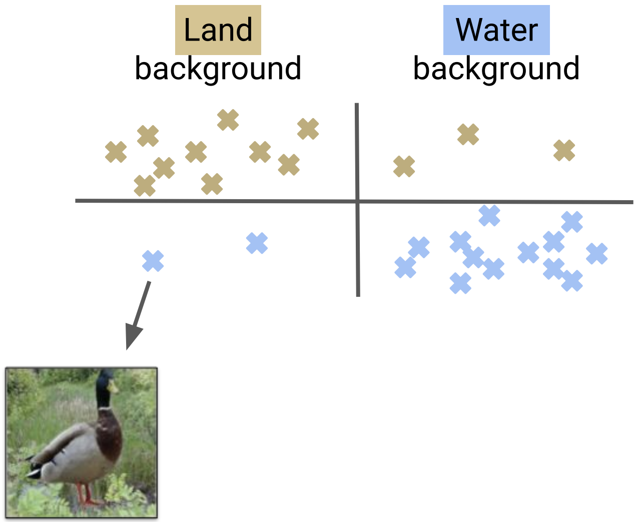
Observation 1: Upweighting worst group yields high worst-group performance (Shimodaira '00; Sagawa et al., '20; Byrd & Lipton, '18)

Observation 2: ERM performs poorly on worst group

Stage 1: Automatically identify worst-group training examples

1. Train *identification model* $f_{\mathrm{id}}(x)$ via ERM

Stage 2: Upweight identified examples



Jtt: Just Train Twice

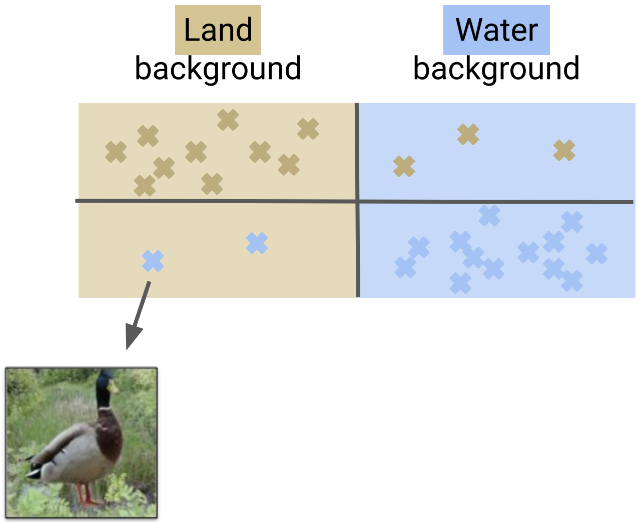
Observation 1: Upweighting worst group yields high worst-group performance (Shimodaira ‘00; Sagawa et al., ‘20; Byrd & Lipton, ‘18)

Observation 2: ERM performs poorly on worst group

Stage 1: Automatically identify worst-group training examples

1. Train *identification model* $f_{\text{id}}(x)$ via ERM

Stage 2: Upweight identified examples



JTT: Just Train Twice

Observation 1: Upweighting worst group yields high worst-group performance (Shimodaira '00; Sagawa et al., '20; Byrd & Lipton, '18)

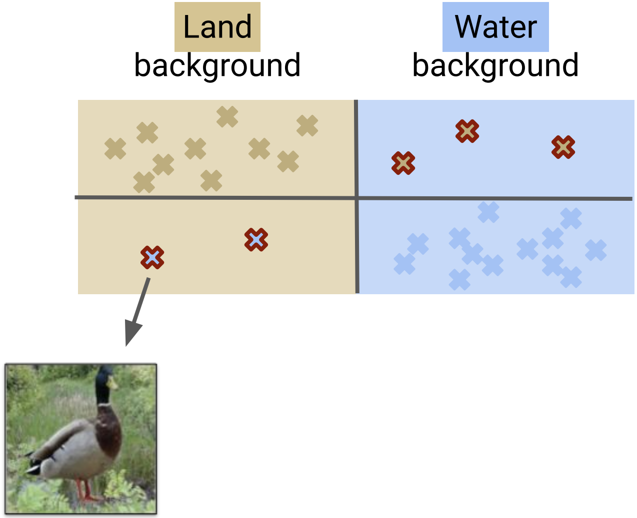
Observation 2: ERM performs poorly on worst group

Stage 1: Automatically identify worst-group training examples

1. Train *identification model* $f_{\text{id}}(x)$ via ERM
2. Compute **error set** of misclassified training examples

$$E = \{(x, y) \text{ s.t. } f_{\text{id}}(x) \neq y\}$$

Stage 2: Upweight identified examples



JTT: Just Train Twice

Observation 1: Upweighting worst group yields high worst-group performance (Shimodaira '00; Sagawa et al., '20; Byrd & Lipton, '18)

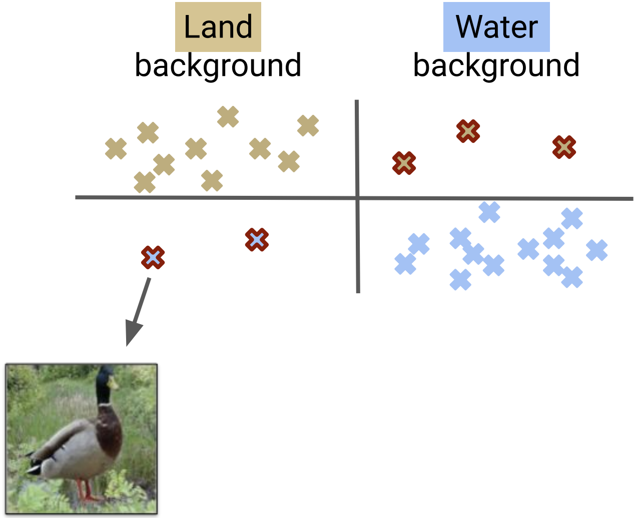
Observation 2: ERM performs poorly on worst group

Stage 1: Automatically identify worst-group training examples

1. Train *identification model* $f_{\text{id}}(x)$ via ERM
2. Compute **error set** of misclassified training examples

$$E = \{(x, y) \text{ s.t. } f_{\text{id}}(x) \neq y\}$$

Stage 2: Upweight identified examples



JTT: Just Train Twice

Observation 1: Upweighting worst group yields high worst-group performance (Shimodaira '00; Sagawa et al., '20; Byrd & Lipton, '18)

Observation 2: ERM performs poorly on worst group

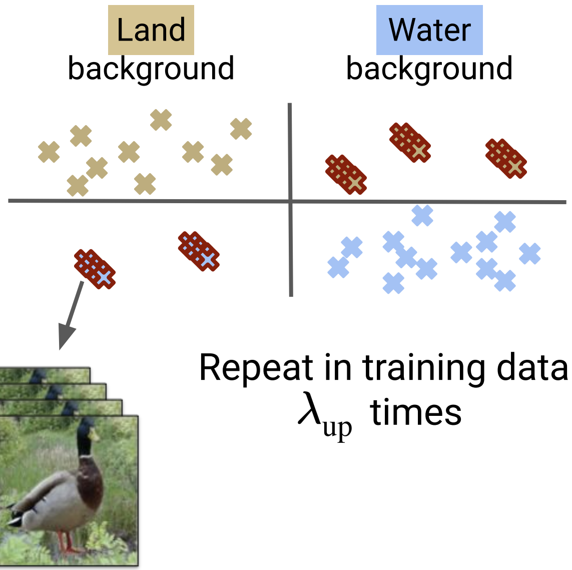
Stage 1: Automatically identify worst-group training examples

1. Train *identification model* $f_{\text{id}}(x)$ via ERM
2. Compute **error set** of misclassified training examples

$$E = \{(x, y) \text{ s.t. } f_{\text{id}}(x) \neq y\}$$

Stage 2: Upweight identified examples

3. *Upsample* error set examples



Repeat in training data λ_{up} times

JTT: Just Train Twice

Observation 1: Upweighting worst group yields high worst-group performance (Shimodaira '00; Sagawa et al., '20; Byrd & Lipton, '18)

Observation 2: ERM performs poorly on worst group

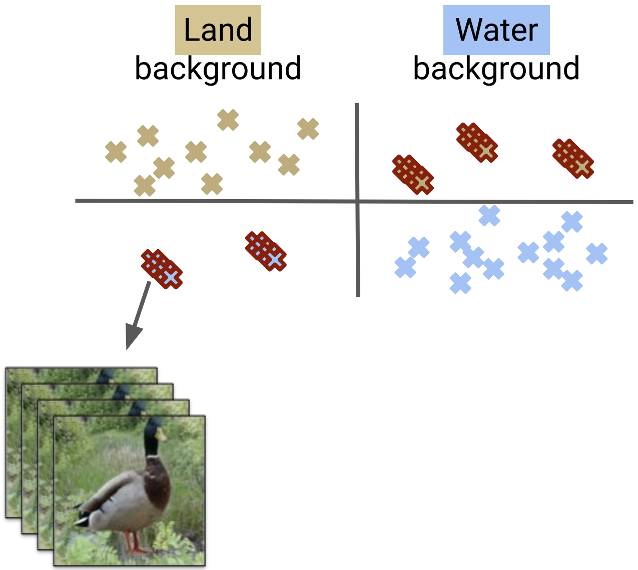
Stage 1: Automatically identify worst-group training examples

1. Train *identification model* $f_{\text{id}}(x)$ via ERM
2. Compute **error set** of misclassified training examples

$$E = \{(x, y) \text{ s.t. } f_{\text{id}}(x) \neq y\}$$

Stage 2: Upweight identified examples

3. *Upsample* error set examples
4. Train *final model* $f_{\text{final}}(x)$ via ERM on the upsampled data



JTT: Just Train Twice

Observation 1: Upweighting worst group yields high worst-group performance (Shimodaira '00; Sagawa et al., '20; Byrd & Lipton, '18)

Observation 2: ERM performs poorly on worst group

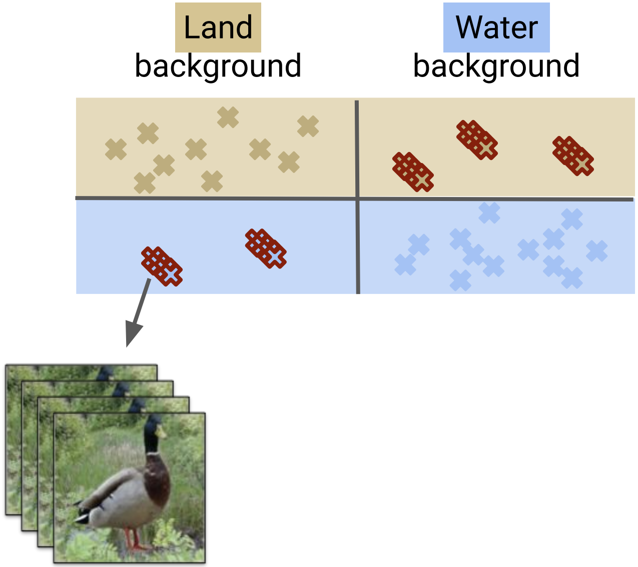
Stage 1: Automatically identify worst-group training examples

1. Train *identification model* $f_{\text{id}}(x)$ via ERM
2. Compute **error set** of misclassified training examples

$$E = \{(x, y) \text{ s.t. } f_{\text{id}}(x) \neq y\}$$

Stage 2: Upweight identified examples

3. *Upsample* error set examples
4. Train *final model* $f_{\text{final}}(x)$ via ERM on the upsampled data



JTT: Just Train Twice

Observation 1: Upweighting worst group yields high worst-group performance (Shimodaira '00; Sagawa et al., '20; Byrd & Lipton, '18)

Observation 2: ERM performs poorly on worst group

Stage 1: Automatically identify worst-group training examples

1. Train *identification model* $f_{\text{id}}(x)$ via ERM
2. Compute **error set** of misclassified training examples

$$E = \{(x, y) \text{ s.t. } f_{\text{id}}(x) \neq y\}$$

Stage 2: Upweight identified examples

3. *Upsample* error set examples
4. Train *final model* $f_{\text{final}}(x)$ via ERM on the upsampled data

\+ **Simple** ("just train twice!")

Jtt: Just Train Twice

Observation 1: Upweighting worst group yields high worst-group performance (Shimodaira '00; Sagawa et al., '20; Byrd & Lipton, '18)

Observation 2: ERM performs poorly on worst group

Stage 1: Automatically identify worst-group training examples

1. Train *identification model* $f_{\text{id}}(x)$ via ERM
2. Compute **error set** of misclassified training examples

$$E = \{(x, y) \text{ s.t. } f_{\text{id}}(x) \neq y\}$$

Stage 2: Upweight identified examples

3. *Upsample* error set examples
4. Train *final model* $f_{\text{final}}(x)$ via ERM on the upsampled data

\+ **Simple** ("just train twice!")

\+ Does not require training group labels…

JTT: Just Train Twice

Observation 1: Upweighting worst group yields high worst-group performance (Shimodaira '00; Sagawa et al., '20; Byrd & Lipton, '18)

Observation 2: ERM performs poorly on worst group

Stage 1: Automatically identify worst-group training examples

1. Train *identification model* $f_{\text{id}}(x)$ via ERM
2. Compute **error set** of misclassified training examples

$$E = \{(x, y) \text{ s.t. } f_{\text{id}}(x) \neq y\}$$

Stage 2: Upweight identified examples

3. *Upsample* error set examples
4. Train *final model* $f_{\text{final}}(x)$ via ERM on the upsampled data

+ **Simple** ("just train twice!")
+ Does not require training group labels… but uses validation group labels for tuning

Experiments: Datasets

Waterbirds

(Wah et al., '11; Sagawa et al., '20)

y: landbird
a: in water

y: landbird
a: on land

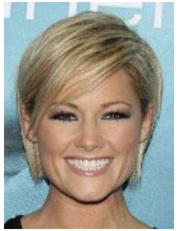
Experiments: Datasets

Waterbirds
(Wah et al., ‘11; Sagawa et al., ‘20)

y: landbird
a: in water

y: landbird
a: on land

CelebA
(Liu et al., ‘15; Sagawa et al., ‘20)



y: blond
a: female

y: not blond
a: male

Experiments: Datasets

Waterbirds
(Wah et al., ‘11; Sagawa et al., ‘20)

y: landbird
a: in water

y: landbird
a: on land

CelebA
(Liu et al., ‘15; Sagawa et al., ‘20)

y: blond
a: female

y: not blond
a: male

MultiNLI
(Williams et al., ‘15; Sagawa et al., ‘20)

S1: Read for Slate's take on Jackson's findings.
S2: Slate had an opinion on Jackson's findings.
y: entailment **a: no negation**

S1: Vrenna and I both fought him and he nearly took us.
S2: Neither Vrenna nor myself have ever fought him.
y: contradiction **a: has negation**

Experiments: Datasets

Waterbirds
(Wah et al., '11; Sagawa et al., '20)

y: landbird
a: in water

y: landbird
a: on land

CelebA
(Liu et al., '15; Sagawa et al., '20)

y: blond
a: female

y: not blond
a: male

MultiNLI
(Williams et al., '15; Sagawa et al., '20)

S1: Read for Slate's take on Jackson's findings.
S2: Slate had an opinion on Jackson's findings.
y: entailment a: no negation

S1: Vrenna and I both fought him and he nearly took us.
S2: Neither Vrenna nor myself have ever fought him.
y: contradiction a: has negation

CivilComments
(Borkan et al., '19; Koh et al., '20)

Maybe you should learn to write a coherent sentence so we can understand WTF your point is.
y: toxic a: none

I applaud your father. He was a good man! We need more like him.
y: non-toxic a: male

Experiments: Points of Comparison

Approaches *without* **training group information**

- Standard training (ERM)

Experiments: Points of Comparison

Approaches *without* **training group information**

- Standard training (ERM)
- CVaR DRO (Levy et al., ‘18)

Experiments: Points of Comparison

Approaches *without* **training group information**

- Standard training (ERM)
- CVaR DRO (Levy et al., ‘18)
- Learning from Failures (LfF) (Nam et al., ‘20)

Experiments: Points of Comparison

Approaches *without* **training group information**

- Standard training (ERM)
- CVaR DRO (Levy et al., ‘18)
- Learning from Failures (LfF) (Nam et al., ‘20)

Others in this category: Lam & Zhou, ‘15; Ren et al., ‘18; Oren et al., ‘19; Sohoni et al., ‘20; Kim et al., ‘19; Shu et al., ‘19; Pezeshki et al., ‘20.

Approaches *with* **training group information**

- Group DRO (Sagawa et al., ‘20)

Others in this category: Shimodaira ’00; Byrd & Lipton ’19; Cao et al., ’19; Mohri et al., ’19; Zhang et al., ’20; Goel et al., ’20; Sagawa et al. ’20; Cao et al., ’20

Experiments: Points of Comparison

Approaches *without* **training group information**

- Standard training (ERM)
- CVaR DRO (Levy et al., '18)
- Learning from Failures (LfF) (Nam et al., '20)

Others in this category: Lam & Zhou, '15; Ren et al., '18; Oren et al., '19; Sohoni et al., '20; Kim et al., '19; Shu et al., '19; Pezeshki et al., '20.

Approaches *with* **training group information**

- Group DRO (Sagawa et al., '20)

Others in this category: Shimodaira '00; Byrd & Lipton '19; Cao et al., '19; Mohri et al., '19; Zhang et al., '20; Goel et al., '20; Sagawa et al. '20; Cao et al., '20

All approaches tuned based on **worst-group performance** on the *validation set* (**requires group information**)

Experiments: Main Results

Method	Training group labels?	Waterbirds		CelebA		MultiNLI		CivilComments	
		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%	59.4%
CVaR DRO	No	96.5%	69.5%	82.4%	64.4%	82.0%	68.0%	92.5%	60.5%
LfF	No	97.5%	75.2%	86.0%	70.6%	80.8%	70.2%	92.5%	58.8%
JTT	No	93.6%	**86.0%**	88.0%	**81.1%**	80.4%	**72.3%**	91.1%	**69.3%**
Group DRO	Yes	93.5%	91.4%	92.9%	88.9%	81.4%	77.7%	88.9%	69.9%

Experiments: Main Results

Method	Training group labels?	Waterbirds		CelebA		MultiNLI		CivilComments	
		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%	59.4%
CVaR DRO	No	96.5%	69.5%	82.4%	64.4%	82.0%	68.0%	92.5%	60.5%
LfF	No	97.5%	75.2%	86.0%	70.6%	80.8%	70.2%	92.5%	58.8%
JTT	No	93.6%	**86.0%**	88.0%	**81.1%**	80.4%	**72.3%**	91.1%	**69.3%**
Group DRO	Yes	93.5%	91.4%	92.9%	88.9%	81.4%	77.7%	88.9%	69.9%

9% average improvement over approaches without training group information

Experiments: Main Results

Method	Training group labels?	Waterbirds		CelebA		MultiNLI		CivilComments	
		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%	59.4%
CVaR DRO	No	96.5%	69.5%	82.4%	64.4%	82.0%	68.0%	92.5%	60.5%
LfF	No	97.5%	75.2%	86.0%	70.6%	80.8%	70.2%	92.5%	58.8%
JTT	No	93.6%	**86.0%**	88.0%	**81.1%**	80.4%	**72.3%**	91.1%	**69.3%**
Group DRO	Yes	93.5%	91.4%	92.9%	88.9%	81.4%	77.7%	88.9%	69.9%

9% average improvement over approaches without training group information

Experiments: Main Results

Method	Training group labels?	Waterbirds		CelebA		MultiNLI		CivilComments	
		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%	59.4%
CVaR DRO	No	96.5%	69.5%	82.4%	64.4%	82.0%	68.0%	92.5%	60.5%
LfF	No	97.5%	75.2%	86.0%	70.6%	80.8%	70.2%	92.5%	58.8%
JTT	No	93.6%	**86.0%**	88.0%	**81.1%**	80.4%	**72.3%**	91.1%	**69.3%**
Group DRO	Yes	93.5%	91.4%	92.9%	88.9%	81.4%	77.7%	88.9%	69.9%

9% average improvement over approaches without training group information

Closes 73% of the gap between standard training and using training group info (group DRO)

Experiments: Worst Group in Error Set

Dataset	Worst-group Recall	Worst-group Precision	Worst-group Empirical Rate
Waterbirds	87.5%	19.1%	1.2%
CelebA	94.7%	9.4%	0.9%
MultiNLI	67.1%	2.2%	1.0%
CivilComments	96.9%	7.8%	0.9%

Experiments: Worst Group in Error Set

Dataset	Worst-group Recall	Worst-group Precision	Worst-group Empirical Rate
Waterbirds	87.5%	19.1%	1.2%
CelebA	94.7%	9.4%	0.9%
MultiNLI	67.1%	2.2%	1.0%
CivilComments	96.9%	7.8%	0.9%

Experiments: Worst Group in Error Set

Dataset	Worst-group Recall	Worst-group Precision	Worst-group Empirical Rate
Waterbirds	87.5%	19.1%	1.2%
CelebA	94.7%	9.4%	0.9%
MultiNLI	67.1%	2.2%	1.0%
CivilComments	96.9%	7.8%	0.9%

JTT automatically identifies **a large fraction** of the worst-group examples

Experiments: Worst Group in Error Set

Dataset	Worst-group Recall	Worst-group Precision	Worst-group Empirical Rate
Waterbirds	87.5%	19.1%	1.2%
CelebA	94.7%	9.4%	0.9%
MultiNLI	67.1%	2.2%	1.0%
CivilComments	96.9%	7.8%	0.9%

JTT automatically identifies **a large fraction** of the worst-group examples

Experiments: Worst Group in Error Set

Dataset	Worst-group Recall	Worst-group Precision	Worst-group Empirical Rate
Waterbirds	87.5%	19.1%	1.2%
CelebA	94.7%	9.4%	0.9%
MultiNLI	67.1%	2.2%	1.0%
CivilComments	96.9%	7.8%	0.9%

JTT automatically identifies **a large fraction** of the worst-group examples

Worst-group examples occur in the error set at a **much higher rate** than in the training data

Experiments: Worst Group in Error Set

Dataset	Worst-group Recall	Worst-group Precision	Worst-group Empirical Rate
Waterbirds	87.5%	19.1%	1.2%
CelebA	94.7%	9.4%	0.9%
MultiNLI	67.1%	2.2%	1.0%
CivilComments	96.9%	7.8%	0.9%

JTT automatically identifies **a large fraction** of the worst-group examples

Worst-group examples occur in the error set at a **much higher rate** than in the training data

⇒ JTT achieves **high worst-group accuracy**

Experiments: Other Groups in Error Set

Waterbirds

Group	Enrichment	ERM test acc.
land background, waterbird	15.92x	72.6%
water background, landbird	6.97x	73.3%
water background, waterbird	2.40x	96.3%
land background, landbird	0.02x	99.3%

CelebA

Group	Enrichment	ERM test acc.
blond male	10.44x	47.2%
blond female	5.42x	89.1%
non-blond male	0.32x	99.3%
non-blond female	0.01x	95.1%

Enrichment: how much more frequently a group appears in the error set than in the training data

$$\text{enrichment} = \frac{\text{precision}}{\text{training rate}}$$

Experiments: Other Groups in Error Set

Waterbirds

Group	Enrichment	ERM test acc.
land background, waterbird	15.92x	72.6%
water background, landbird	6.97x	73.3%
water background, waterbird	2.40x	96.3%
land background, landbird	0.02x	99.3%

CelebA

Group	Enrichment	ERM test acc.
blond male	10.44x	47.2%
blond female	5.42x	89.1%
non-blond male	0.32x	99.3%
non-blond female	0.01x	95.1%

Enrichment: how much more frequently a group appears in the error set than in the training data

$$\text{enrichment} = \frac{\text{precision}}{\text{training rate}}$$

Groups with worse performance appear in the error set at a higher rate

Experiments: Other Groups in Error Set

Waterbirds

Group	Enrichment	ERM test acc.
land background, waterbird	15.92x	72.6%
water background, landbird	6.97x	73.3%
water background, waterbird	2.40x	96.3%
land background, landbird	0.02x	99.3%

CelebA

Group	Enrichment	ERM test acc.
blond male	10.44x	47.2%
blond female	5.42x	89.1%
non-blond male	0.32x	99.3%
non-blond female	0.01x	95.1%

Enrichment: how much more frequently a group appears in the error set than in the training data

$$\text{enrichment} = \frac{\text{precision}}{\text{training rate}}$$

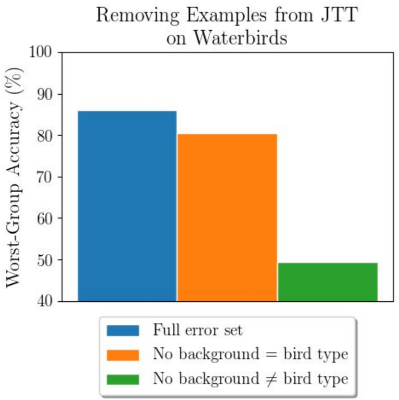
Groups with worse performance appear in the error set at a higher rate

Experiments: Other Groups in Error Set

Group	Enrichment	ERM test acc.
land background, waterbird	15.92x	72.6%
water background, landbird	6.97x	73.3%
water background, waterbird	2.40x	96.3%
land background, landbird	0.02x	99.3%

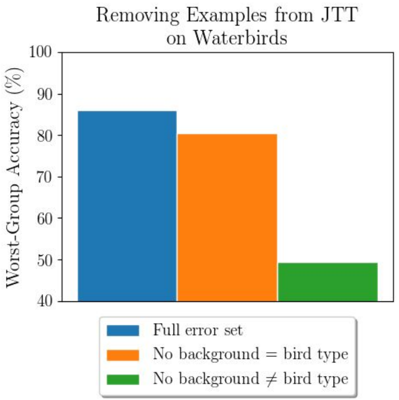
Experiments: Other Groups in Error Set

Group	Enrichment	ERM test acc.
land background, waterbird	15.92x	72.6%
water background, landbird	6.97x	73.3%
water background, waterbird	2.40x	96.3%
land background, landbird	0.02x	99.3%



Experiments: Other Groups in Error Set

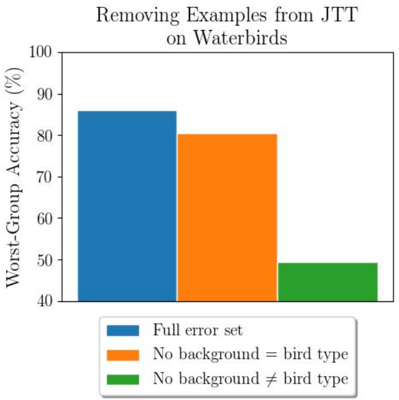
Group	Enrichment	ERM test acc.
land background, waterbird	15.92x	72.6%
water background, landbird	6.97x	73.3%
water background, waterbird	2.40x	96.3%
land background, landbird	0.02x	99.3%



Even examples from groups where the *spurious correlation holds* can be helpful to upweight

Experiments: Other Groups in Error Set

Group	Enrichment	ERM test acc.
land background, waterbird	15.92x	72.6%
water background, landbird	6.97x	73.3%
water background, waterbird	2.40x	96.3%
land background, landbird	0.02x	99.3%



Even examples from groups where the *spurious correlation holds* can be helpful to upweight

In error set: water in background not salient

Not in error set: salient water background

Summary

- Standard training frequently **performs poorly** on the *worst group*, especially in the presence of **spurious correlations**.

Summary

- Standard training frequently **performs poorly** on the *worst group*, especially in the presence of **spurious correlations**.

- Reweighting examples with training group labels: **performs well** on the worst group but is **expensive**

Summary

- Standard training frequently **performs poorly** on the *worst group*, especially in the presence of **spurious correlations**.
- Reweighting examples with training group labels: **performs well** on the worst group but is **expensive**
- JTT: **performs well** on the worst group and is **cheaper** (still uses **validation group** labels!)

Summary

- Standard training frequently **performs poorly** on the *worst group*, especially in the presence of **spurious correlations**.

- Reweighting examples with training group labels: **performs well** on the worst group but is **expensive**

- JTT: **performs well** on the worst group and is **cheaper** (still uses **validation group** labels!)

Code: https://github.com/anniesch/jtt