

Correct-N-Contrast: A Contrastive Approach for Improving Robustness to Spurious Correlations

Michael Zhang, Nimit S. Sohoni, Hongyang R. Zhang, Chelsea Finn, Christopher Ré

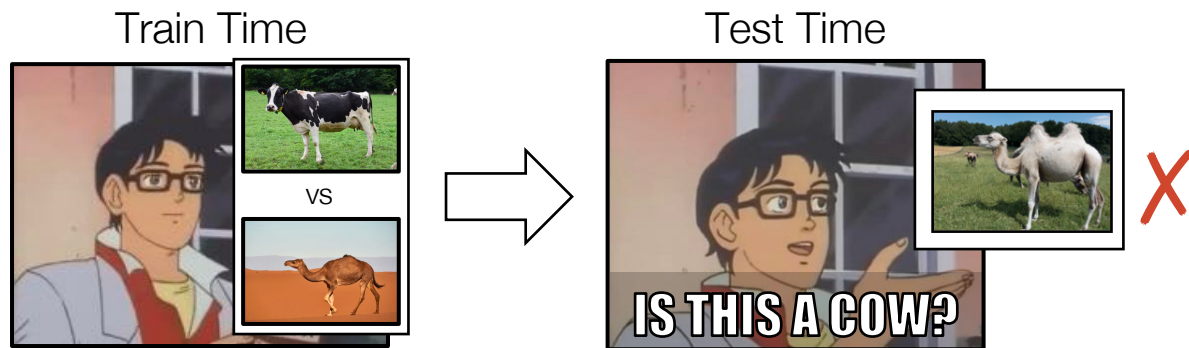


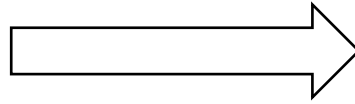
Fig. 1: A deep learning model not robust to spurious correlations



Why worry about spurious correlations?

We know deep learning is great for learning correlations in complex data.

Input features:



Class labels:

normal,
abnormal condition

?

Why worry about spurious correlations?

We know deep learning is great for learning correlations in complex data.

This lets us build automated + effective* classifiers for many important tasks!

	F1 Score (95% CI)
Radiologist Avg. (N = 4)	0.387 (0.330, 0.442)
CheXNet	0.435 (0.387, 0.481)

F1 scores averaged over 14 lung condition classification tasks

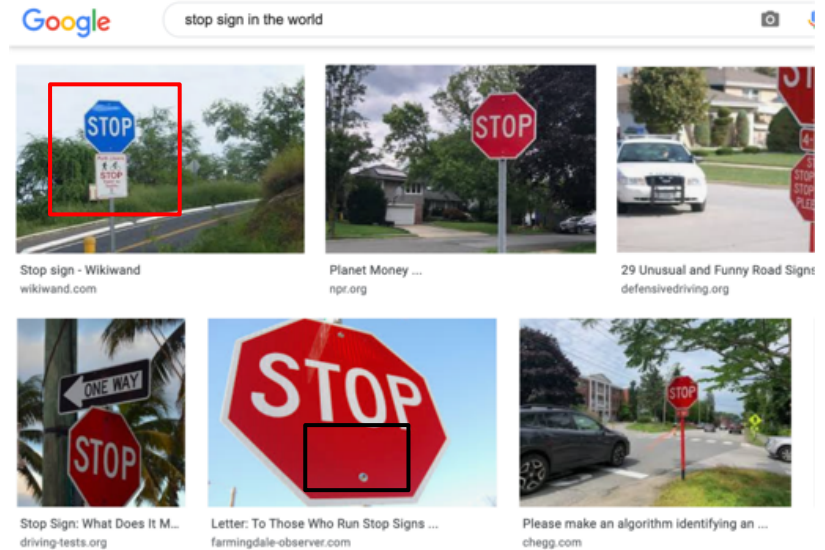
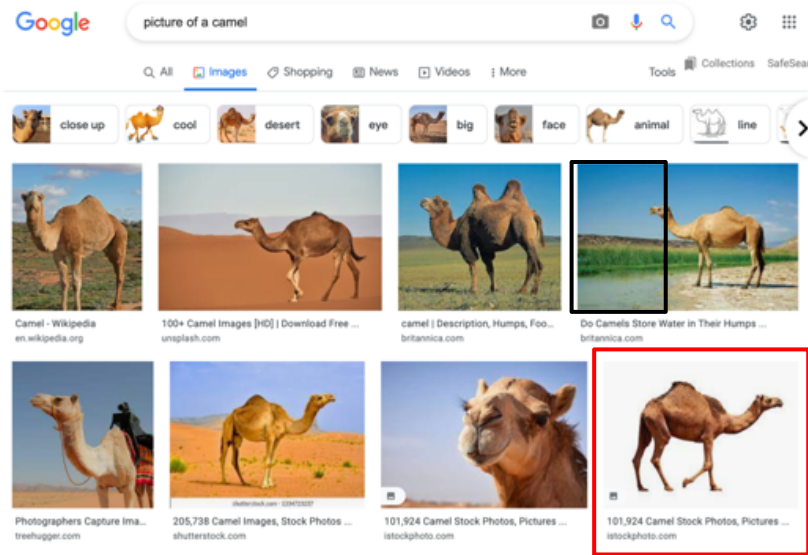


However, aggregate metrics may not tell the whole story.

*We'll see shortly why there's an asterisk here

Why worry about spurious correlations?

To get high average performance, neural nets may learn **spurious** correlations that hold for many but not all datapoints

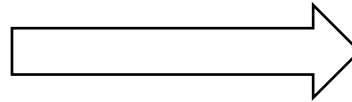


Why worry about spurious correlations?

To get high average performance, neural nets may learn **spurious** correlations that hold for many but not all datapoints

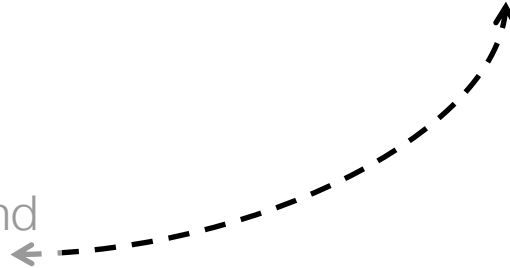


Classify



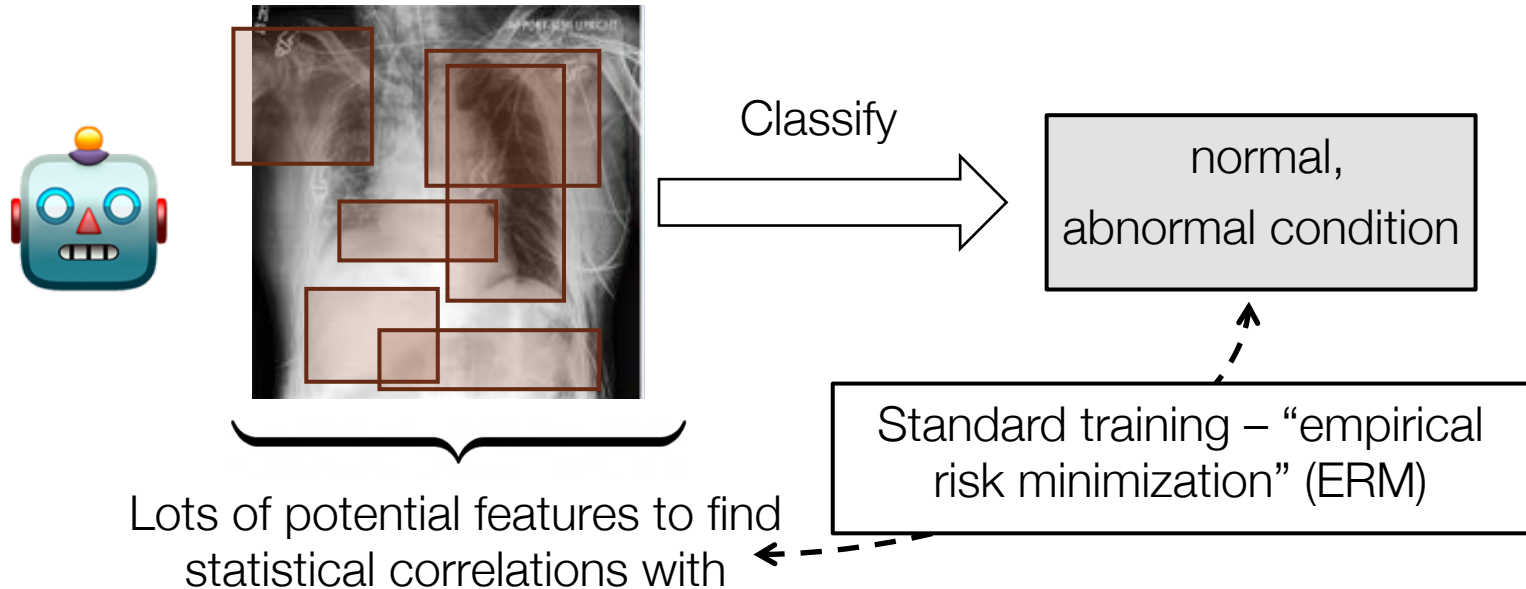
normal,
abnormal condition

Lots of potential features to find
statistical correlations with



Why worry about spurious correlations?

To get high average performance, neural nets may learn **spurious** correlations that hold for many but not all datapoints



Why worry about spurious correlations?

Prior work shows that neural nets learn **spurious correlations**, and systematically misclassify individual data groups

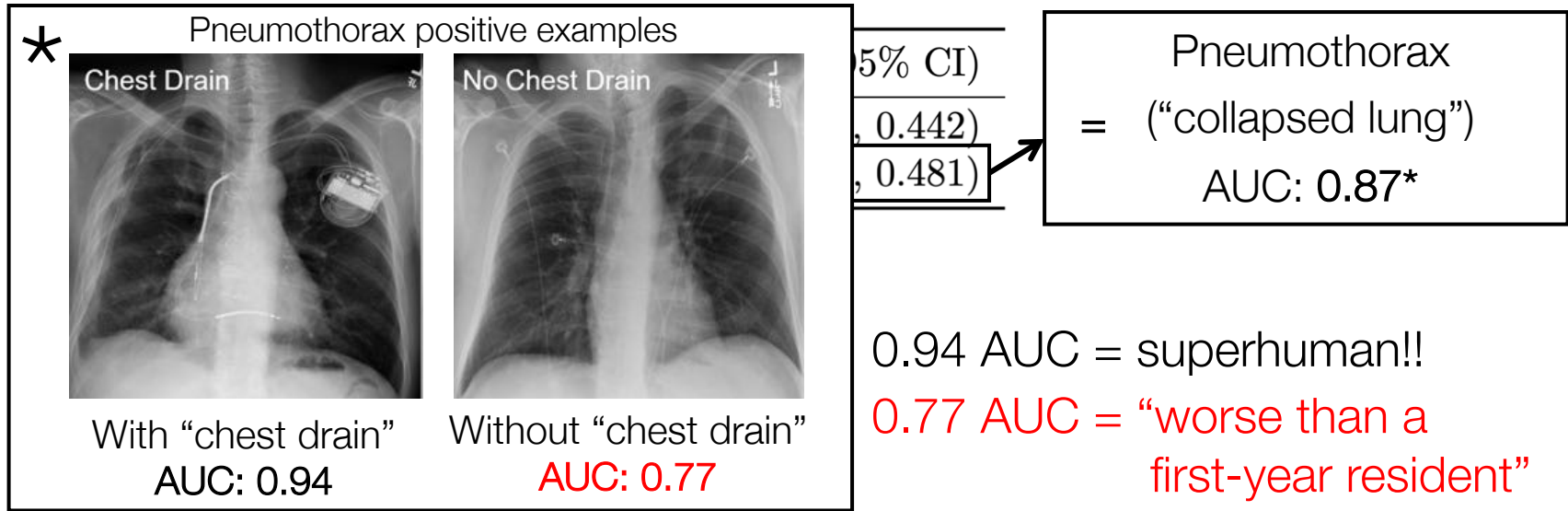
	F1 Score (95% CI)
Radiologist Avg. (N = 4)	0.387 (0.330, 0.442)
CheXNet	0.435 (0.387, 0.481)

Pneumothorax
= (“collapsed lung”)
AUC: 0.87*

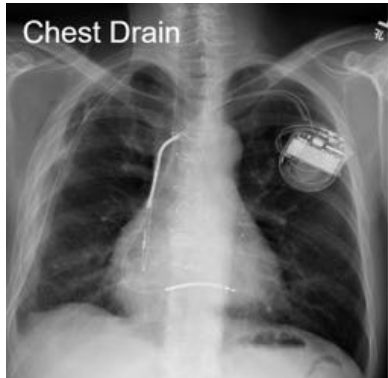
By relying on “chest drain”

Why worry about spurious correlations?

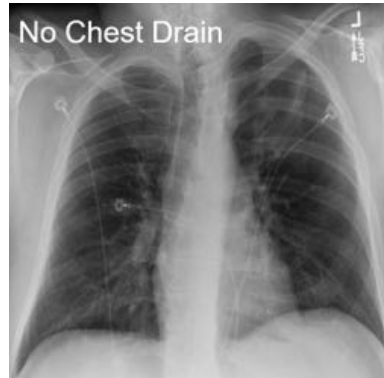
Prior work shows that neural nets learn **spurious correlations**, and systematically misclassify individual data groups



Spurious correlations pose real-world problems for deep learning



With "chest drain"
AUC: 0.94



Without "chest drain"
AUC: 0.77

20% of **patients** in test set did not have chest drain

With chest drain = *already treated*
Won't see this chest drain correlation in deployment.

“Superhuman” model much worse in practice!

So how do we get models robust to spurious correlations?

Outlining what's to come

1. Give structure to our problem setting
2. Discuss prior work:
 1. Core ideas
 2. Standing challenges
3. Introduce our approach:
 1. Use contrastive learning to ignore spurious correlations
 2. Empirical + theoretical justification
4. Results
 1. State-of-the-art

Outlining what's to come

1. Give structure to our problem setting – key terms, assumptions, objectives
2. Discuss prior work:
 1. Core ideas
 2. Standing challenges
3. Introduce our approach:
 1. Use contrastive learning to ignore spurious correlations
 2. Empirical + theoretical justification
4. Results
 1. State-of-the-art

Outlining what's to come

1. Give structure to our problem setting
2. Discuss prior work:
 1. Core ideas – remove spurious correlations from training data
 2. Standing challenges – trade-off between label assumptions + robustness
3. Introduce our approach:
 1. Use contrastive learning to ignore spurious correlations
 2. Empirical + theoretical justification
4. Results
 1. State-of-the-art

Outlining what's to come

1. Give structure to our problem setting
2. Discuss prior work:
 1. Core ideas
 2. Standing challenges
3. Introduce our approach, Correct-N-Contrast (CNC):
 1. Use contrastive learning to ignore spurious correlations
 2. Empirical + theoretical justification
4. Results
 1. State-of-the-art

Outlining what's to come

1. Give structure to our problem setting
2. Discuss prior work:
 1. Core ideas
 2. Standing challenges
3. Introduce our approach:
 1. Use contrastive learning to ignore spurious correlations
 2. Empirical + theoretical justification
4. Results
 1. State-of-the-art – improve trade-off significantly

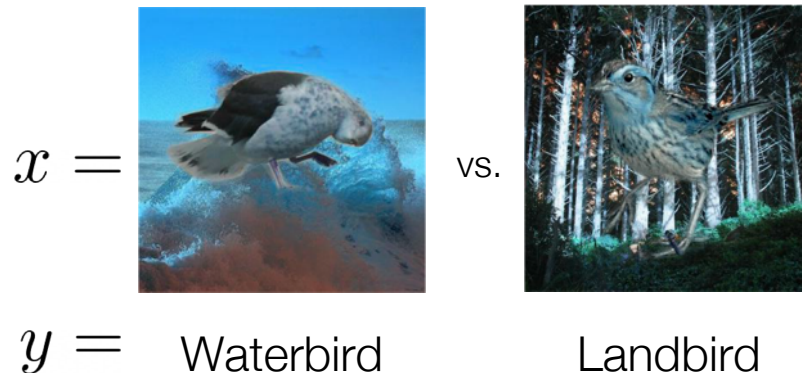
Problem Setting

Problem setting and objective

Goal: obtain accurate classifiers that are robust to spurious correlations.

Default data setup

- Sample input features: x
- Ground-truth class labels: y
- Task: classify y given x

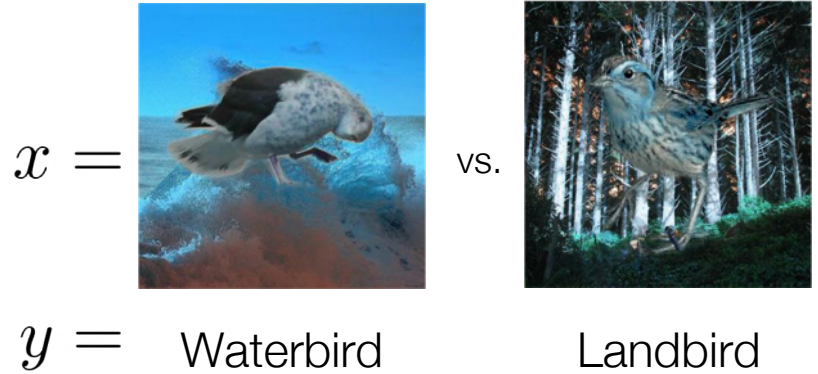


Problem setting and objective

Goal: obtain accurate classifiers that are robust to spurious correlations.

Data setup with spurious correlations

- Sample input features: x
- Ground-truth class labels: y
- Spurious attributes: a



95% of all training samples in the same class share the same background type

Problem setting and objective

Goal: obtain accurate classifiers that are robust to spurious correlations.

Data setup with spurious correlations

$$\mathcal{X} = [\mathcal{X}_y, \mathcal{X}_a, \mathcal{X}_\epsilon]$$



Problem setting and objective

Goal: obtain accurate classifiers that are robust to spurious correlations.

Data setup with spurious correlations

$$x = \begin{matrix} y \\ \updownarrow \\ [x_y, x_a, x_\epsilon] \end{matrix}$$

Ground-truth features
(bird pixels)



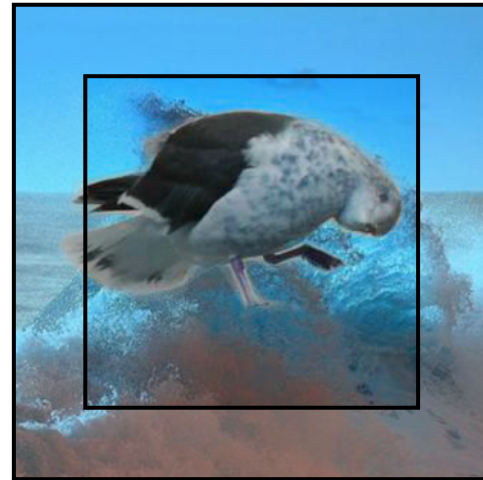
Problem setting and objective

Goal: obtain accurate classifiers that are robust to spurious correlations.

Data setup with spurious correlations

$$\mathcal{X} = \begin{matrix} y & a \\ \updownarrow & \updownarrow \\ \mathcal{X}_y & \mathcal{X}_a, \mathcal{X}_\epsilon \end{matrix}$$

Spurious features
(background pixels)



Problem setting and objective

Goal: obtain accurate classifiers that are robust to spurious correlations.

Data setup with spurious correlations

$$\mathcal{X} = \left[\begin{array}{c} y \\ \updownarrow \\ x_y \end{array}, \begin{array}{c} a \\ \updownarrow \\ x_a \end{array}, x_\epsilon \right]$$

Noise features
(other pixels)



Problem setting and objective

Goal: obtain accurate classifiers that are robust to spurious correlations.

$$\mathcal{X} = \left[\overset{y}{\updownarrow} \underset{\uparrow}{\mathcal{X}_y}, \overset{a}{\updownarrow} \mathcal{X}_a, \mathcal{X}_\epsilon \right]$$

Classify by changes here



Problem setting and objective

Goal: obtain **accurate** classifiers that are **robust to spurious correlations**.

$$\mathcal{X} = \left[\overset{y}{\updownarrow} x_y, \overset{a}{\updownarrow} x_a, x_\epsilon \right]$$

Don't classify by changes here



Problem setting and objective

Goal: obtain **accurate** classifiers that are **robust to spurious correlations**.

Unfortunately, standard training, towards empirical risk minimization (ERM), can lead to **relying on spurious features!**

On Waterbirds, ERM gets **97.3%** average test accuracy...

... by “attending” to background ☹️

Input features



Grad-CAM*



Problem setting and objective

Goal: obtain **accurate** classifiers that are **robust to spurious correlations**.

To evaluate:

Problem setting and objective

Goal: obtain accurate classifiers that are robust to spurious correlations.



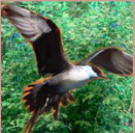

To **evaluate**: first define “groups” as data subsets that share unique combos of ground-truth class label + **spurious attribute**



Problem setting and objective

Goal: obtain **accurate** classifiers that are **robust to spurious correlations**.

To **evaluate**: then measure average and **group-wise** performance

	Waterbirds	Landbirds y
Water Background	 Acc: 95.0%	 Acc: 80.4%
Land Background a	 Acc: 62.6%	 Acc: 99.3%

Large gap in average vs. **worst-group** performance → poor robustness to spurious correlations

Problem setting and objective

Goal: obtain **accurate** classifiers that are **robust to spurious correlations**.

To **evaluate**: then measure average and **group-wise** performance

Key challenge 1: how can we train models that obtain high average and worst-group performance?

Key challenge 2: how can we do so *without knowing* training data spurious attributes or group labels?

Prior Work

Prior work: reweighting improves robustness!

Key similarity for prior state-of-the-art approaches: **reweight** or **resample** data groups during training

Reweight to “remove” spurious correlations,
so models don’t learn them

Prior work: reweighting improves robustness!

Spuriously correlated datasets have group imbalance

Waterbirds example: many samples in one group, few in another

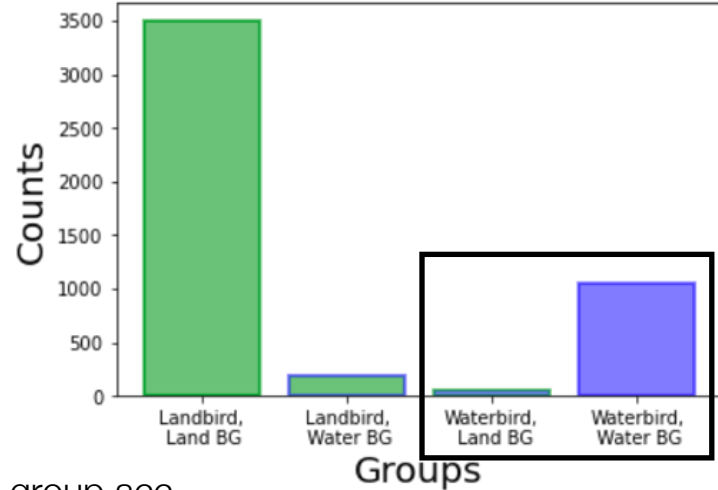
Landbird Samples



Waterbird Samples



Train Distribution (Default)



Poor ERM model robustness... 62.6% worst-group acc.

Strong spurious correlation between class label and spurious features

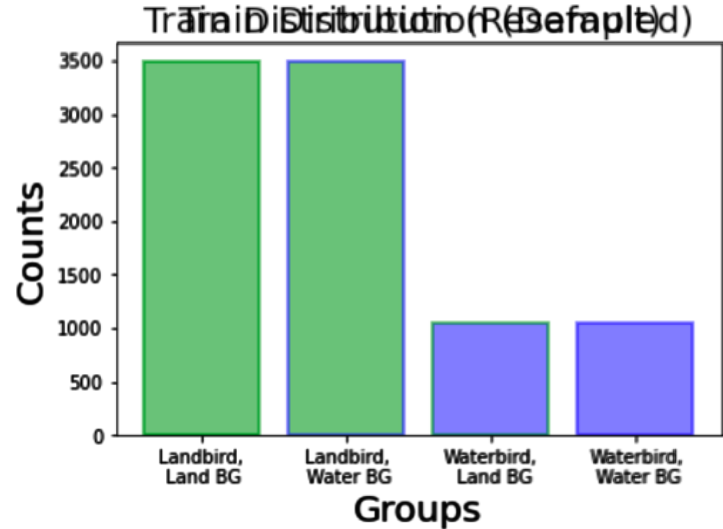
Prior work: reweighting improves robustness!

Intuitively, when we reweight / resample...

Landbird Samples



Waterbird Samples



Prior work: reweighting improves robustness!

Intuitively, when we reweight / resample...

We can remove correlation between ground-truth class labels and spurious features

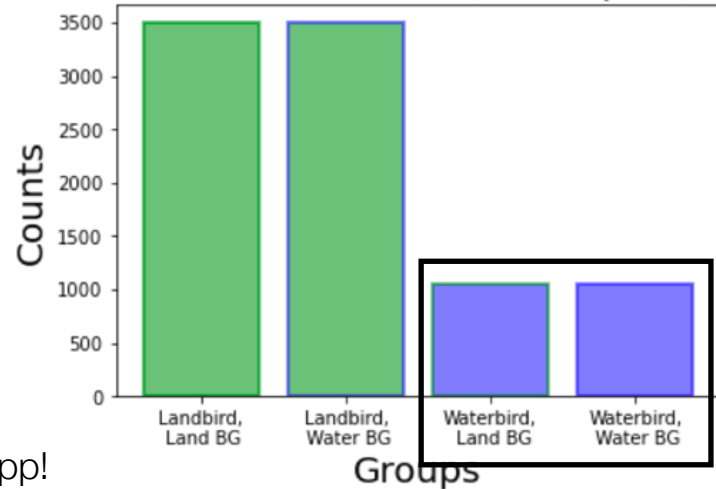
Landbird Samples



Waterbird Samples



Train Distribution (Resampled)



Improves worst-group acc. over ERM by up to 27.3 pp!

Now can train with ground-truth labels. No reason to focus on spurious features!

However, reweighting introduces a trade-off

Annotation cost vs. robustness to spurious correlations

If group info available

- Can reweight to minimize training worst-group accuracy
- Group Distributionally Robust Optimization (GDRO)



- ✓ Effectively improves robustness
- ✗ Requires (costly) training group info

However, reweighting introduces a trade-off

Annotation cost vs. robustness to spurious correlations

If group info available

- Can reweight to minimize training worst-group accuracy
- Group Distributionally Robust Optimization (GDRO)



If group info *not* available

- First infer groups / spurious attributes
- Then train robust model
- Just Train Twice (JTT)
- Environment Inference for Invariant Learning (EILL)
- Learning from Failure (LfF)
- GEORGE

✓ Effectively improves robustness

✗ Requires (costly) training group info

However, reweighting introduces a trade-off

Annotation cost vs. robustness to spurious correlations

Just Train Twice (JTT)



If group info *not* available

- First infer groups / spurious attributes
- Then train robust model
- Just Train Twice (JTT)
- Environment Inference for Invariant Learning (EILL)
- Learning from Failure (LfF)
- GEORGE

However, reweighting introduces a trade-off

Annotation cost vs. robustness to spurious correlations

Just Train Twice (JTT)

1. Train a model w/ few epochs via ERM (learn spurious correlations)



If group info *not* available

- First infer groups / spurious attributes
- Then train robust model
- Just Train Twice (JTT)
- Environment Inference for Invariant Learning (EILL)
- Learning from Failure (LfF)
- GEORGE

However, reweighting introduces a trade-off

Annotation cost vs. robustness to spurious correlations

Just Train Twice (JTT)

1. Train a model w/ few epochs via ERM (learn spurious correlations)
2. Upsample incorrect samples
 - Train robust model on resampled dataset towards ERM



If group info *not* available

- First infer groups / spurious attributes
- **Then train robust model**
- Just Train Twice (JTT)
- Environment Inference for Invariant Learning (EILL)
- Learning from Failure (LfF)
- GEORGE

However, reweighting introduces a trade-off

Annotation cost vs. robustness to spurious correlations

If group info available

- Can reweight to minimize training worst-group accuracy
- Group Distributionally Robust Optimization (GDRO)



- ✓ Effectively improves robustness
- ✗ Requires (costly) training group info



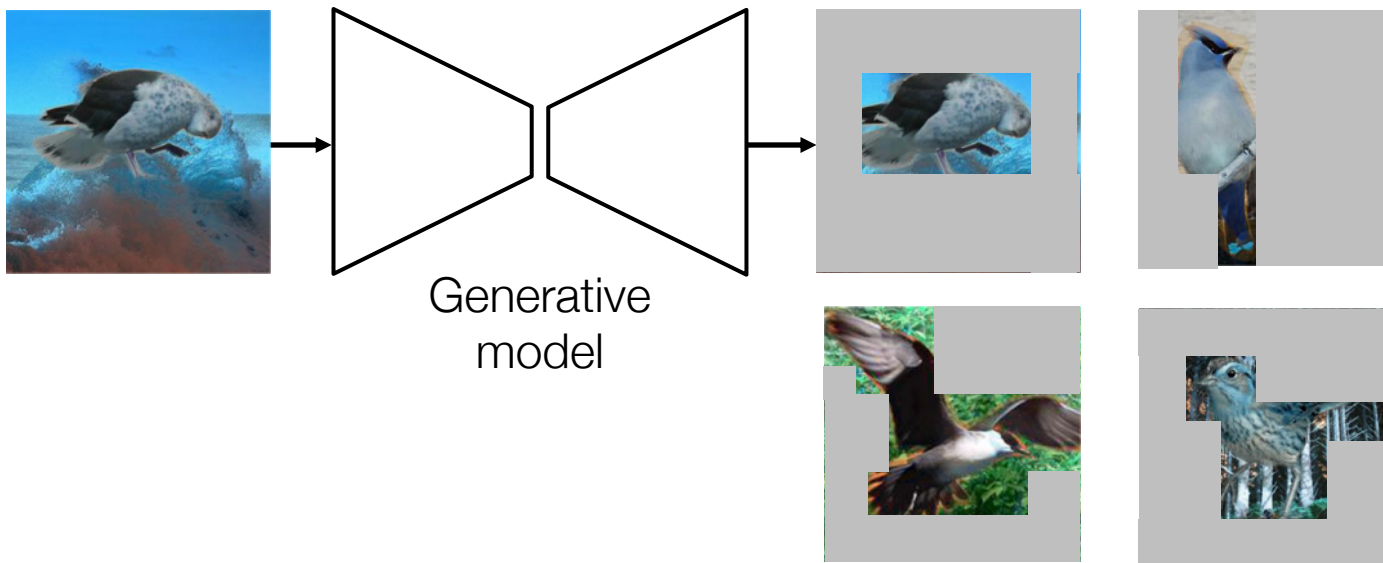
If group info *not* available

- First infer groups / spurious attributes
- Then train robust model
- Just Train Twice (JTT)
- Environment Inference for

- ✓ No training group info required
- ✗ -4.5 pp worst-group acc. vs GDRO

Other approaches to improve robustness

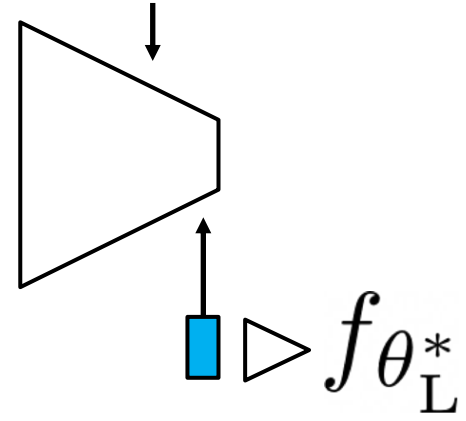
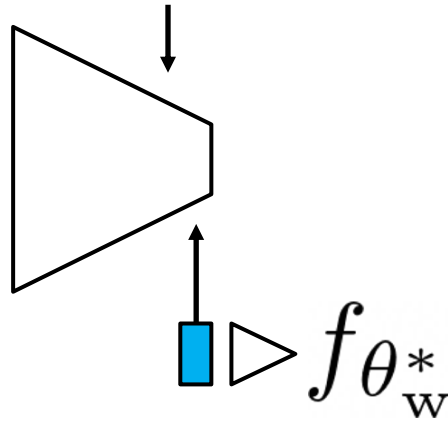
Generate samples that remove spurious features



Other approaches to improve robustness

Invariant risk minimization (and friends)

Trained model updated with invariant objective



Same optimal linear classifier
across different groups

$$f_{\theta_{\text{W}}^*} = f_{\theta_{\text{L}}^*}$$

Trade-off still occurs

Annotation cost vs. robustness to spurious correlations

If group info available

- ✓ Effectively improves robustness
- ✗ Requires (costly) training group info

If group info *not* available

- ✓ No training group info required
- ✗ -4.5 pp worst-group acc. vs GDRO

Can we improve this tradeoff?

Reduce the robustness gap without requiring training group information?

If group info available



If group info *not* available

Yes! Decrease gap by 80% with our work
(from -4.5 pp to -0.9 pp)

✓ Effectively improves robustness

✗ Requires (costly) training group info

✓ No training group info required

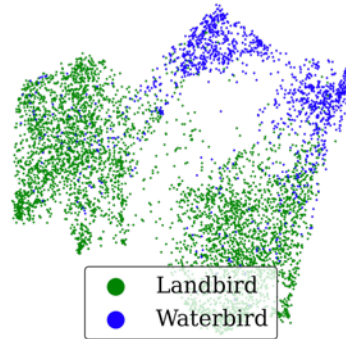
✗ -4.5 pp worst-group acc. vs GDRO

Our Approach

Our approach: Correct-N-Contrast (CNC)

Key idea: use **contrastive learning** to ignore spurious features

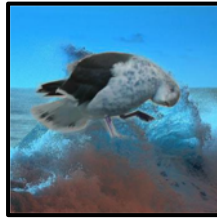
- + Change the way we present our training data
- + Use class labels *and* hidden-layer representations to guide training



UMAP visualization of hidden-layer representations

Our approach: Correct-N-Contrast (CNC)

For robustness, how can we directly train to ignore differences in spurious features?



$y =$ Waterbird

$a =$ Water background



Waterbird

Land background

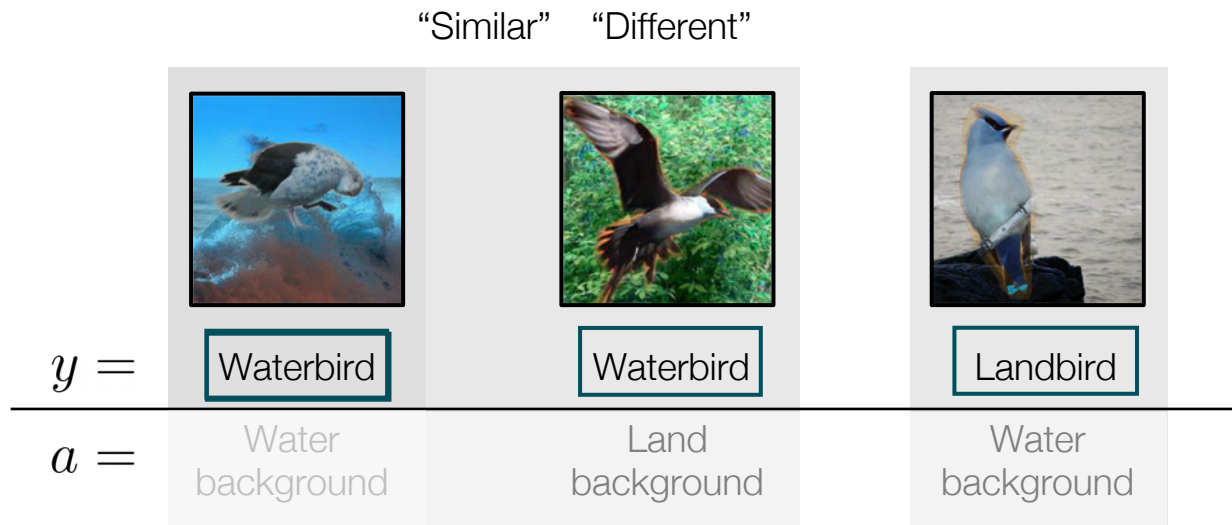


Landbird

Water background

Our approach: Correct-N-Contrast (CNC)

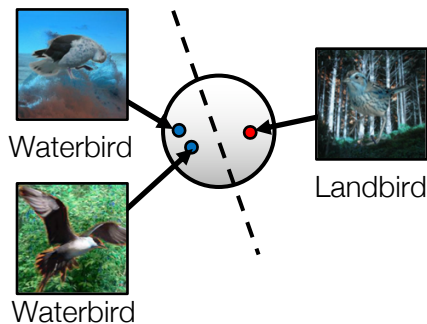
For robustness, how can we directly train to ignore differences in spurious features?



Our approach: Correct-N-Contrast (CNC)

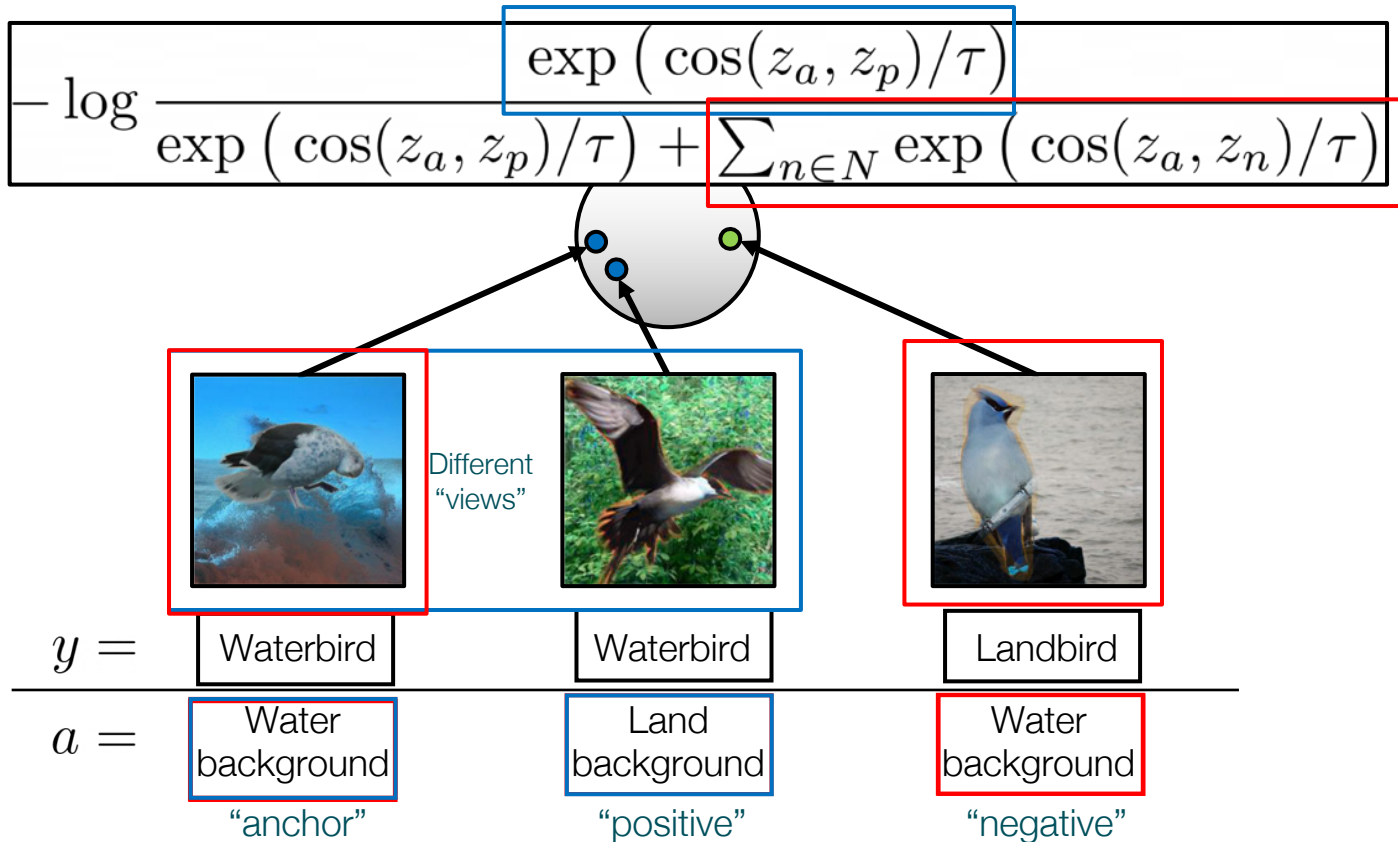
Can also use neural network **hidden-layer representations** to guide this!

Objective: learn hidden-layer representations that encode class information, but are **robust to changes in spurious attributes**



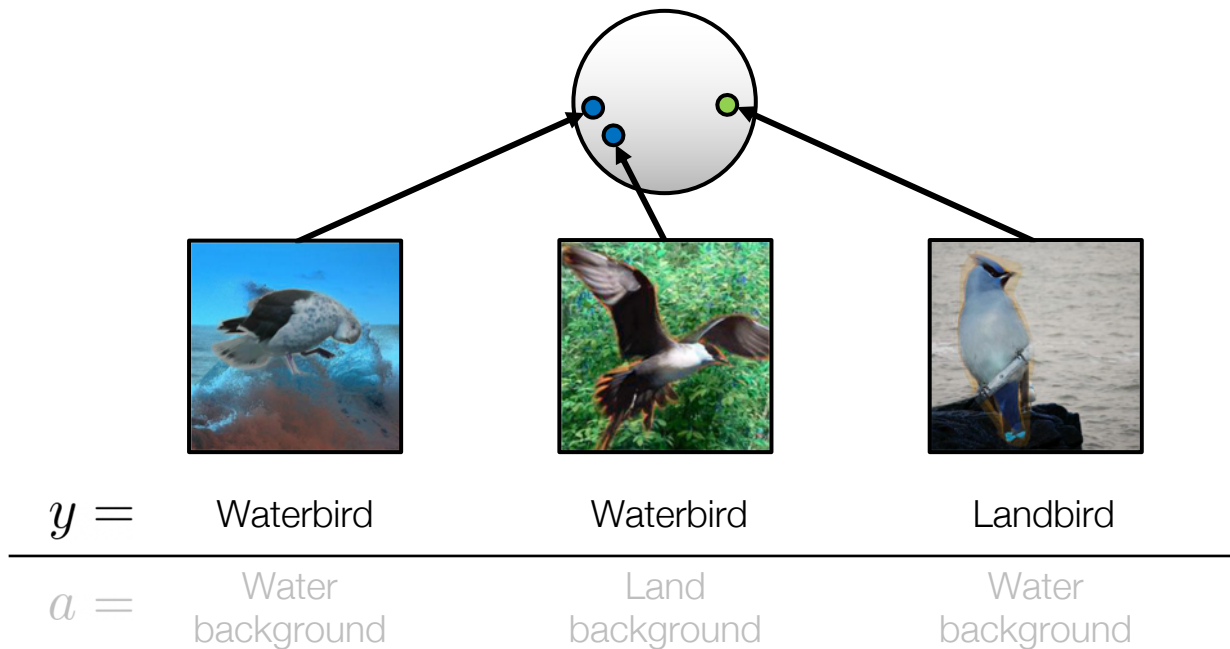
Samples in **same class** should embed **closer** to each other than samples in **different classes**, regardless of their spurious features.

Our approach: Correct-N-Contrast (CNC)



Our approach: Correct-N-Contrast (CNC)

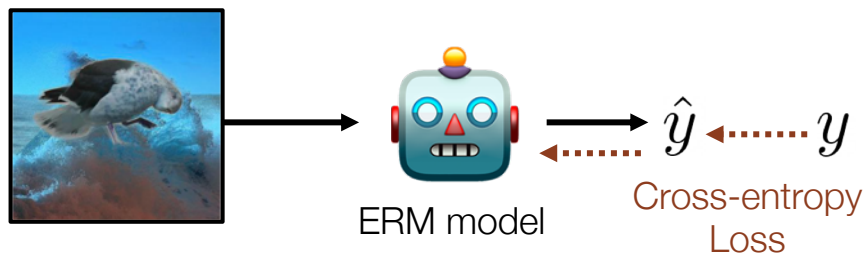
In practice: we don't want to assume spurious attribute information for training data points



Like prior work, adopt two-stage procedure. First **infer** spurious attributes.

Correct-N-Contrast (CNC) in practice

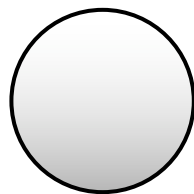
Stage 1: Aim to infer spurious attributes by training an initial model with ERM



Use the result that ERM training encourages predicting based on spurious features

Correct-N-Contrast (CNC) in practice

Stage 2: Train robust model with contrastive learning using ERM model's predictions as proxy for spurious attributes



$y =$

Waterbird

Waterbird

Landbird

$\hat{y} =$

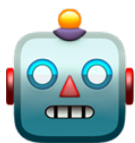
“Waterbird”



“Landbird”

“Waterbird”

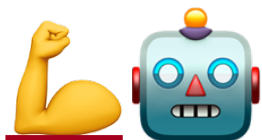
anchor



ERM model

Correct-N-Contrast (CNC) in practice

Stage 2: Train robust model with contrastive learning using ERM model's predictions as proxy for spurious attributes



Robust model



$y =$

Waterbird

$\hat{y} =$

“Waterbird”

anchor



Waterbird

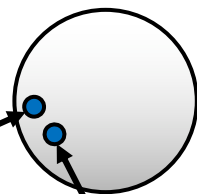
“Landbird”

positive



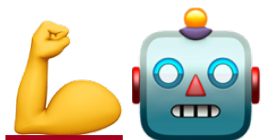
Landbird

“Waterbird”

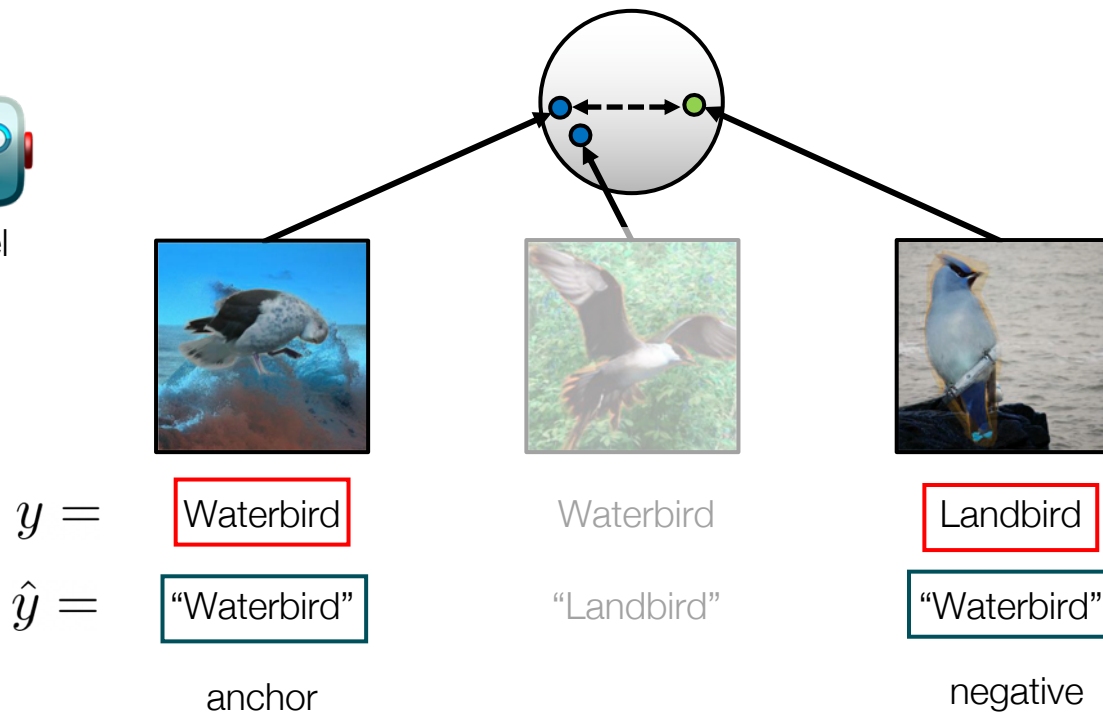


Correct-N-Contrast (CNC) in practice

Stage 2: Train robust model with contrastive learning using ERM model's predictions as proxy for spurious attributes



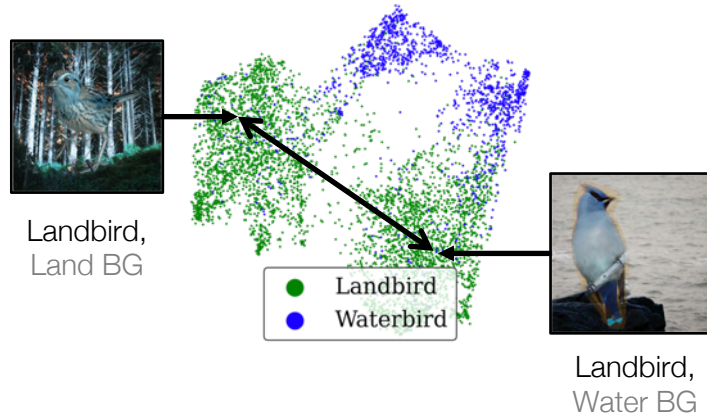
Robust model



Understanding why CNC can improve robustness

To support CNC's approach, we make additional **empirical** and **theoretical** connections between robustness and representation learning

1. Quantify representation “quality”, and show this tracks worst-group performance across various spuriously correlated datasets



Use an “alignment loss” inspired by prior contrastive learning theory*

Measures representation distance between samples in the same class but different groups

Understanding why CNC can improve robustness

To support CNC's approach, we make additional **empirical** and **theoretical** connections between robustness and representation learning

1. Quantify representation “quality”, and show this tracks worst-group performance across various spuriously correlated datasets

(metric for representation quality)
2. Theoretically prove **alignment loss** helps bound important robustness metrics (worst-group vs. average performance gap)

Checkout paper and poster for more details!

Results

Results

Two questions for evaluation:

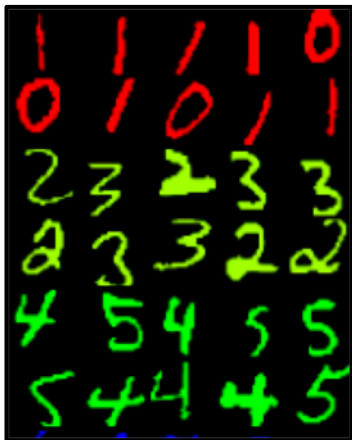
1. Does CNC actually improve robustness to spurious correlations and raise worst-group performance?
2. Can our connections between representation learning and robustness help explain CNC's performance?

Do not assume training set group or spurious attribute labels

Benchmarks

Evaluate on four popular spurious correlation benchmarks (image + text),
three different neural network architectures (LeNet, ResNet, Transformer)

Colored MNIST



Class: digit
Spurious: color

Waterbirds



Class: bird type
Spurious: background

CelebA



Class: blond(e)
Spurious: gender

Civilcomments-WILDS

“She hates men because that’s
what her mother taught her.”

Y = toxic

A = [male, female]

“I doubt that anyone cares
whether you believe it or not.”

Y = not toxic

A = [none]

Class: toxicity
Spurious: demographic

Q1: Does CNC improve robustness to spurious correlations?

A1.1: CNC improves worst-group accuracy over prior state-of-the-art methods that don't require training group information

Q1: Does CNC improve robustness to spurious correlations?

A1.1: CNC improves worst-group accuracy over prior state-of-the-art methods that don't require training group information

Accuracy (%)	CMNIST*		Waterbirds		CelebA		CivilComments-WILDS	
	Worst-group	Average	Worst-group	Average	Worst-group	Average	Worst-group	Average
ERM	0.0 (0.0)	20.1 (0.2)	62.6 (0.3)	97.3 (1.0)	47.7 (2.1)	94.9 (0.3)	58.6 (1.7)	92.1 (0.4)
				...				
JTT	74.5 (2.4)	90.2 (0.8)	83.8 (1.2)	89.3 (0.7)	81.5 (1.7)	88.1 (0.3)	69.3 (-)*	91.1 (-)*
CNC (Ours)	77.4 (3.0)	90.9 (0.6)	88.5 (0.3)	90.9 (0.1)	88.8 (0.9)	89.9 (0.5)	68.9 (2.1)	81.7 (0.5)
Group DRO	78.5 (4.5)	90.6 (0.1)	89.9 (0.6)	92.0 (0.6)	88.9 (1.3)	93.9 (0.1)	69.8 (2.4)	89.0 (0.3)

Abridged Table 1. (See paper for more method comparisons!)

On worst-group accuracy, CNC obtains
+3.6 pp over prior SoTA and **just -0.9 pp** under Oracle

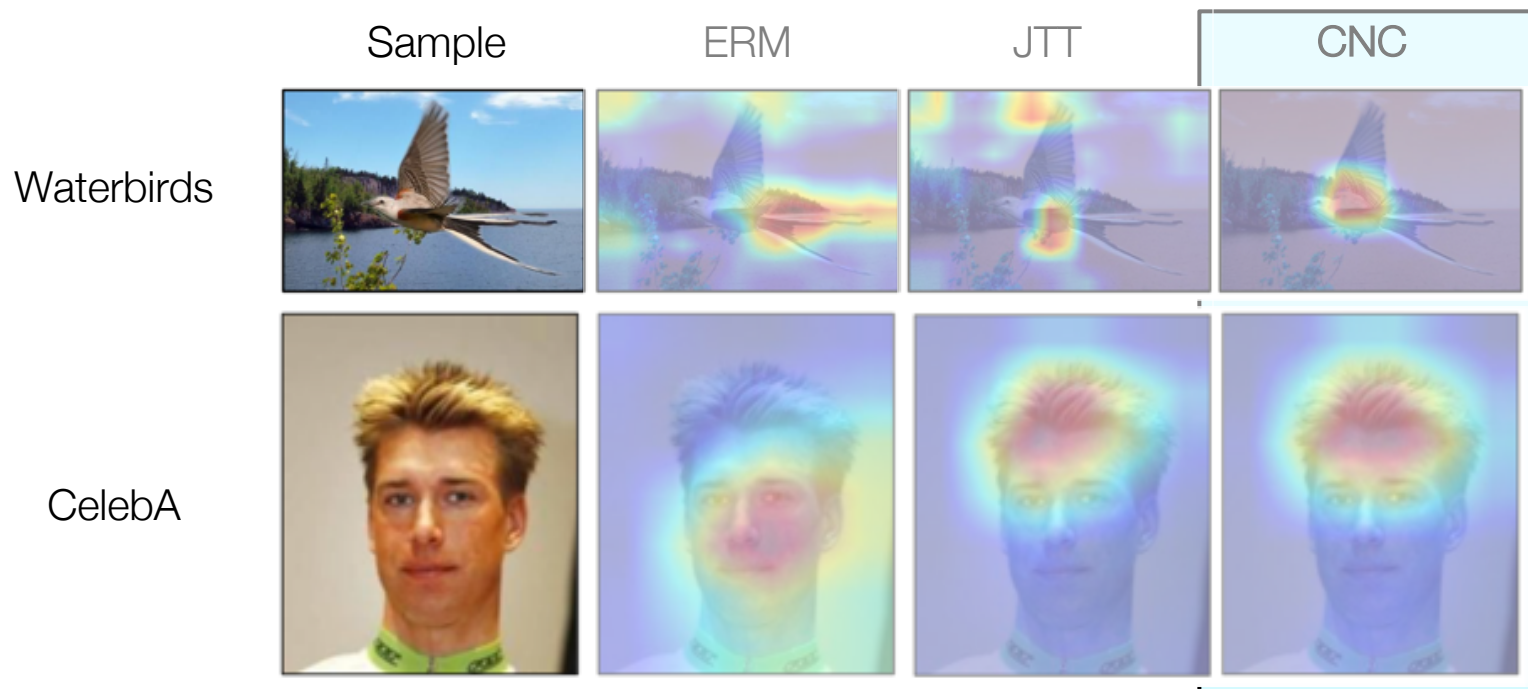
Q1: Does CNC improve robustness to spurious correlations?

A1.2: Grad-CAM visualizations suggests CNC enables greater reliance on class-aligned features



Q1: Does CNC improve robustness to spurious correlations?

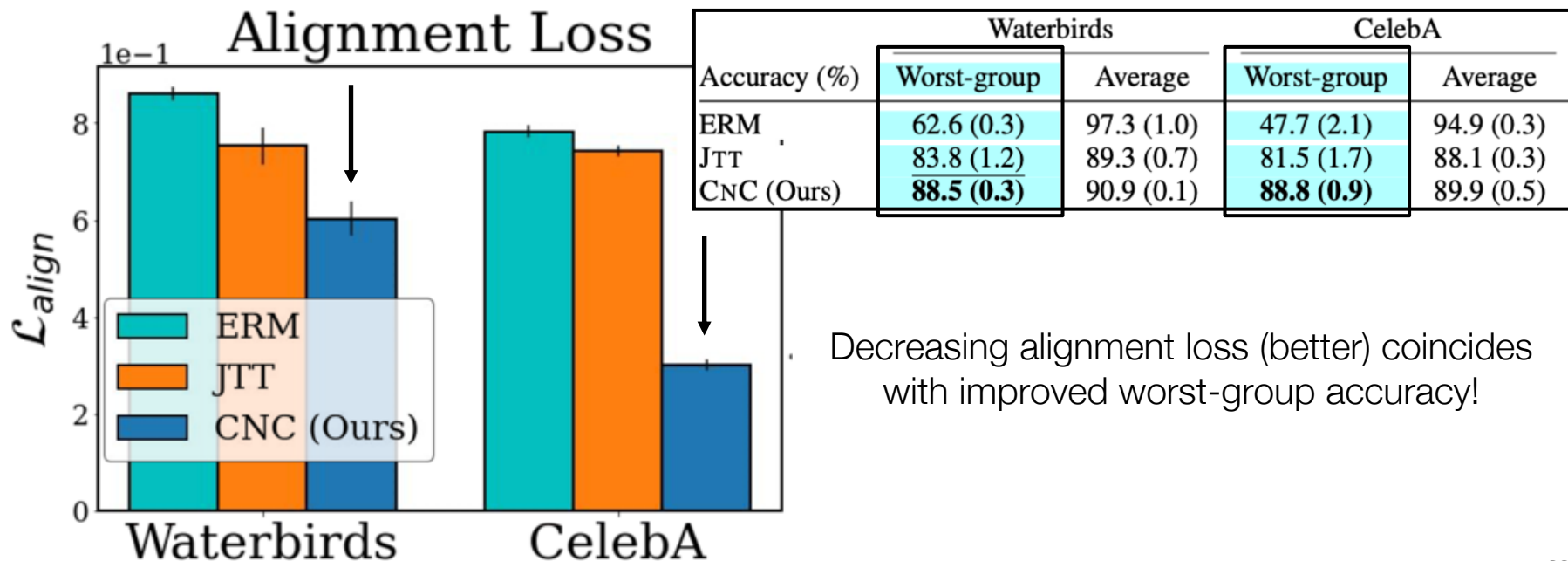
A1.2: Grad-CAM visualizations suggests CNC enables greater reliance on class-aligned features



Q2: Can better representations explain greater robustness?

Q2: Can better representations explain greater robustness?

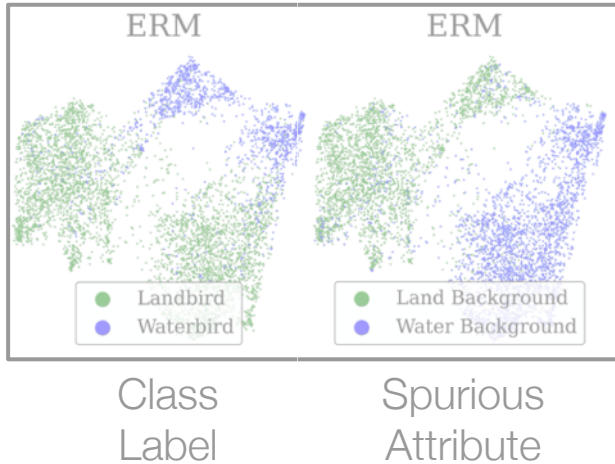
A2.1: CNC's improved worst-group accuracy corresponds to lower alignment loss



Decreasing alignment loss (better) coincides with improved worst-group accuracy!

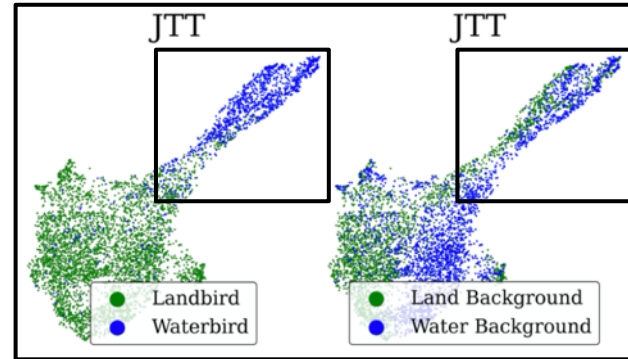
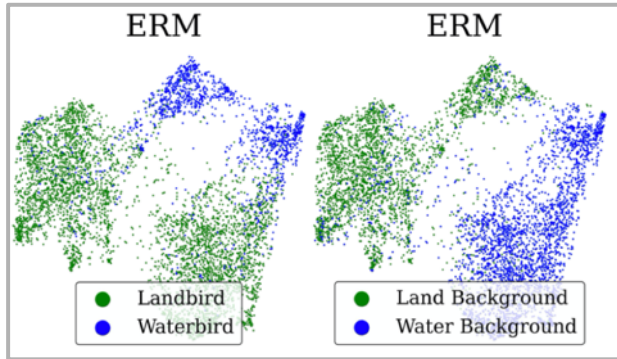
Q2: Can better representations explain greater robustness?

A2.2: UMAP visualizations of trained models suggests CNC leads to greater robustness to spurious features



Q2: Can better representations explain greater robustness?

A2.2: UMAP visualizations of trained models suggests CNC leads to greater robustness to spurious features



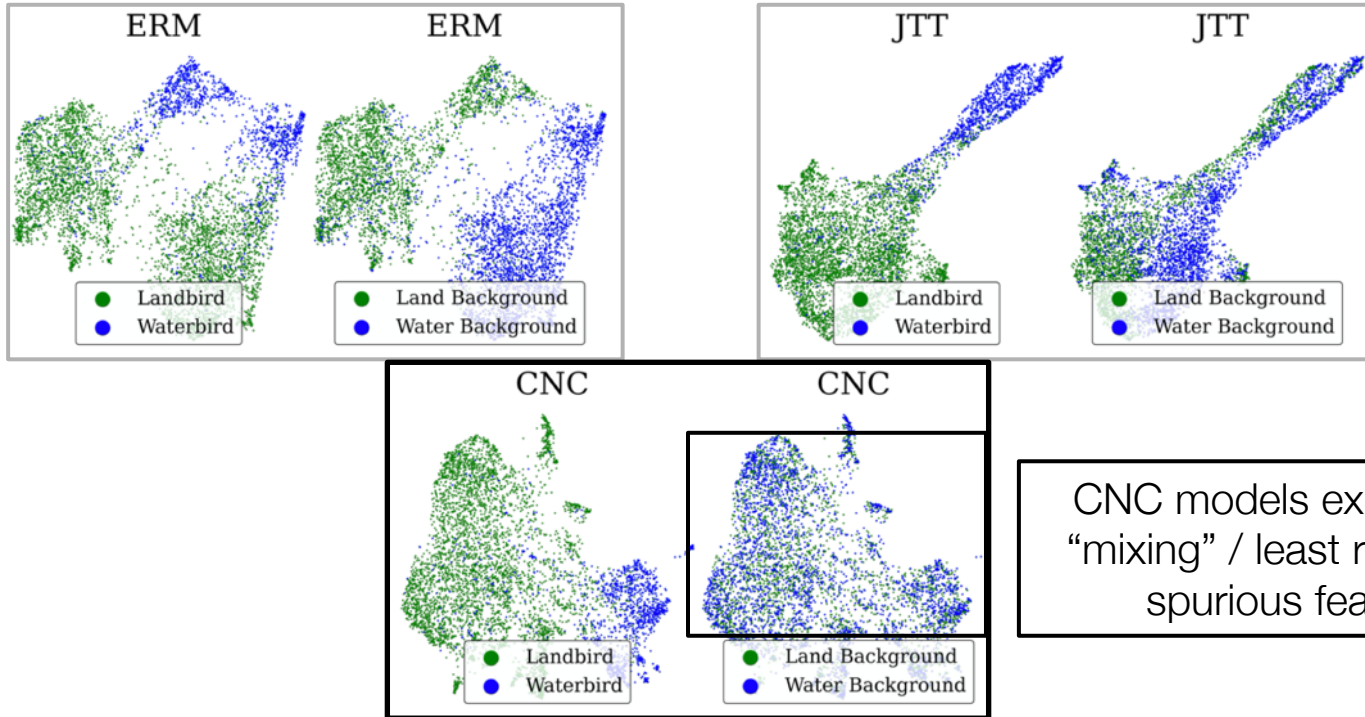
Class
Label

Spurious
Attribute

JTT models exhibit greater “mixing”
of samples with same class,
different spurious attributes

Q2: Can better representations explain greater robustness?

A2.2: UMAP visualizations of trained models suggests CNC leads to greater robustness to spurious features



Summary

1. Discussed the spurious correlations problem
 - Important to tackle for deep learning applications!
2. Reviewed key directions + limitations of prior work
 - Trade-off between annotation cost & robustness
3. Introduced our method, Correct-N-Contrast
 - Connections with contrastive learning to improve robustness
 - State-of-the-art robustness without training group info
 - Substantially closes prior robustness gap

Summary

1. Discussed the spurious correlations problem
 - Important to tackle for deep learning applications!
2. Reviewed key directions + limitations of prior work
 - Trade-off between annotation cost & robustness
3. Introduced our method, Correct-N-Contrast
 - Connections with contrastive learning to improve robustness
 - State-of-the-art robustness without training group info
 - Substantially closes prior robustness gap

Summary

1. Discussed the spurious correlations problem
 - Important to tackle for deep learning applications!
2. Reviewed key directions + limitations of prior work
 - Trade-off between annotation cost & robustness
3. Introduced our method, Correct-N-Contrast
 - Connections with contrastive learning to improve robustness
 - State-of-the-art robustness without training group info
 - Substantially closes prior robustness gap

Summary

1. Discussed the spurious correlations problem
 - Important to tackle for deep learning applications!
2. Reviewed key directions + limitations of prior work
 - Trade-off between annotation cost & robustness
3. Introduced our method, Correct-N-Contrast
 - Connections with contrastive learning to improve robustness
 - State-of-the-art robustness without training group info
 - Substantially closes prior robustness gap

Summary

1. Discussed the spurious correlations problem
 - Important to tackle for deep learning applications!
2. Reviewed key directions + limitations of prior work
 - Trade-off between annotation cost & robustness
3. Introduced our method, Correct-N-Contrast
 - Connections with contrastive learning to improve robustness
 - State-of-the-art robustness without training group info
 - Substantially closes prior robustness gap

Summary

1. Discussed the spurious correlations problem
 - Important to tackle for deep learning applications!
2. Reviewed key directions + limitations of prior work
 - Trade-off between annotation cost & robustness
3. Introduced our method, Correct-N-Contrast
 - Connections with contrastive learning to improve robustness
 - State-of-the-art robustness *without* training group info
 - Substantially closes prior robustness gap

Summary

1. Discussed the spurious correlations problem
 - Important to tackle for deep learning applications!
2. Reviewed key directions + limitations of prior work
 - Trade-off between annotation cost & robustness
3. Introduced our method, Correct-N-Contrast
 - Connections with contrastive learning to improve robustness
 - State-of-the-art robustness without training group info
 - Substantially closes prior robustness gap + reduces tradeoff!

Thanks!

Chat with us at our poster!

Poster Session 2, Hall E #435

Wed, July 20, 2022, 6:30 — 8:30 p.m. EDT



Paper - bit.ly/cnc-icml



Code - github.com/HazyResearch/correct-n-contrast



Can we improve this tradeoff?

Reduce the robustness gap without requiring training group information?

If group info available



If group info *not* available

- Can reweight worst-group
- Group Distribution Robust Optimization (GDRO)

Yes! Decrease gap by 80% with our work
(from -4.5 pp to -0.9 pp)

- groups / spurious
- robust model
- Just Train Twice (JTT)
- Environment Inference for

✓ Effectively improves robustness

✗ Requires (costly) training group info

✓ No training group info required

✗ -4.5 pp worst-group acc. vs GDRO