NEURAL INFORMATION PROCESSING SYSTEMS

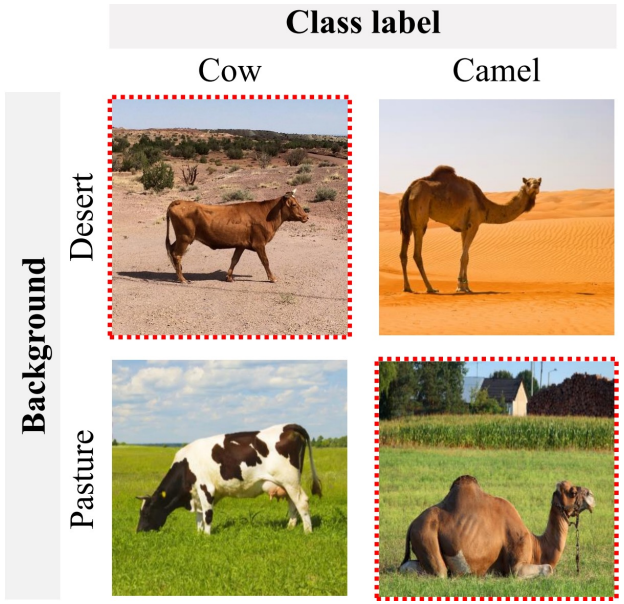
Seoul National University
Cognitive Machine Learning Lab.

Mitigating Spurious Correlations via Disagreement Probability

Hyeonggeun Han[1,2], Sehwan Kim[1], Hyungjun Joo[1,2], Sangwoo Hong[1,2], Jungwoo Lee[1,2,3]

[1]ECE & [2]NextQuantum, Seoul National University, [3]Hodoo AI Labs

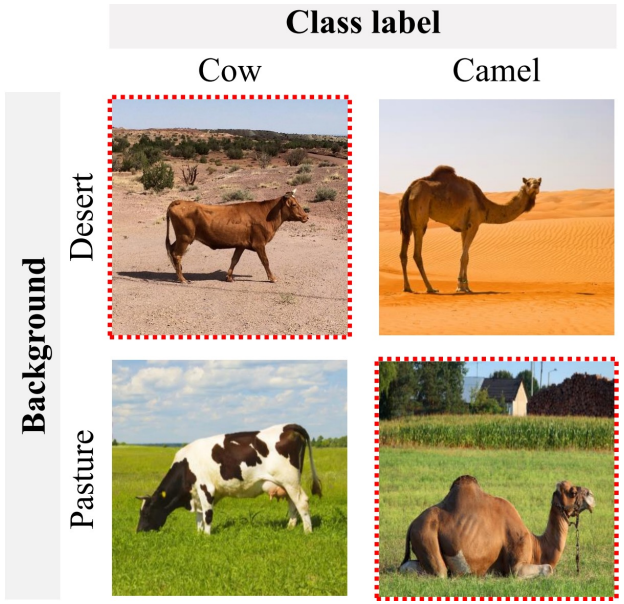
Problem



Cow/camel classification task

- A majority of camel images feature desert backgrounds, while a majority of cow images feature pasture backgrounds.
- ERM-trained models might learn to recognize animals based on their backgrounds—desert for camels and pasture for cows—rather than on their distinctive features.
- This reliance causes misclassifications on certain groups.

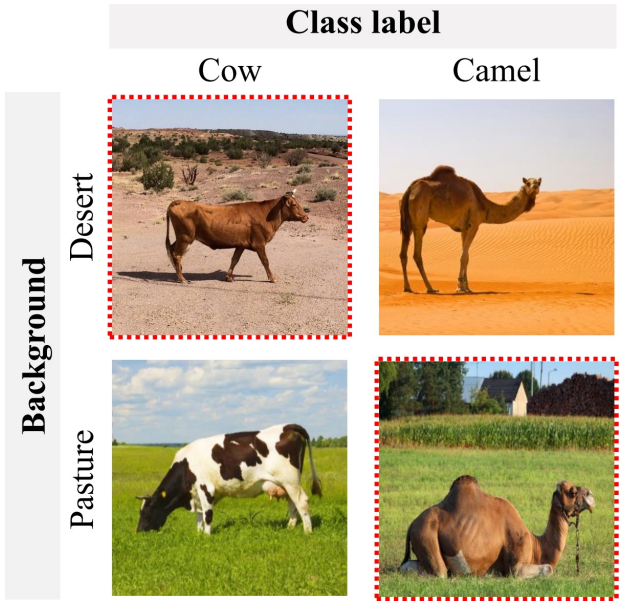
Motivation



Cow/camel classification task

- Debiasing methods with bias labels: Demonstrate remarkable success. However, bias labels (e.g., background) are expensive.
- Debiasing methods without bias labels: Employ a two-stage strategy: (1) identifying bias-conflicting samples, and (2) training the debiased model by enhancing performance for these identified bias-conflicting samples.

Motivation

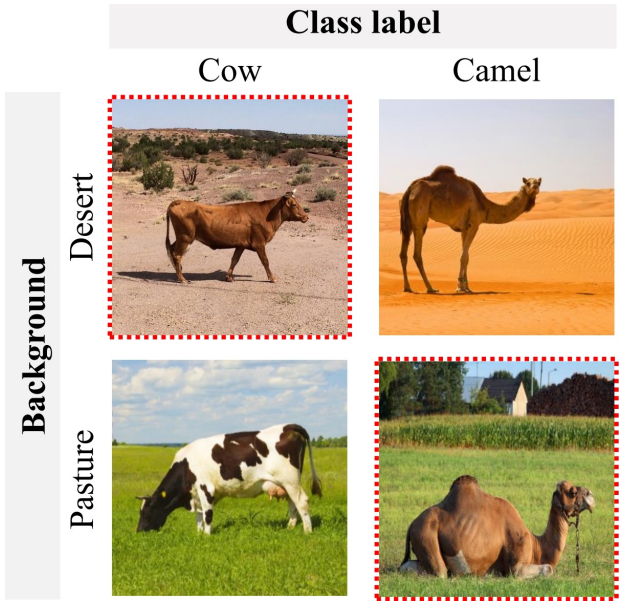


Cow/camel classification task

- Debiasing methods with bias labels: Demonstrate remarkable success. However, bias labels (e.g., background) are expensive.
- Debiasing methods without bias labels: Employ a two-stage strategy: (1) identifying bias-conflicting samples, and (2) training the debiased model by enhancing performance for these identified bias-conflicting samples.

Q: What is the optimal extent of enhancing performance on the bias-conflicting samples?

Motivation



Cow/camel classification task

- Debiasing methods with bias labels: Demonstrate remarkable success. However, bias labels (e.g., background) are expensive.
- Debiasing methods without bias labels: Employ a two-stage strategy: (1) identifying bias-conflicting samples, and (2) training the debiased model by enhancing performance for these identified bias-conflicting samples.

Q: What is the optimal extent of enhancing performance on the bias-conflicting samples?

A: A debiased model should exhibit consistent performance across both bias-aligned and bias-conflicting samples.

Method

- The training objective for mitigating spurious correlations

$$\min_{\theta} \max_{b \in \mathcal{B}} \left\{ \hat{\mathcal{L}}_b := \frac{1}{n_b} \sum_{(x,y,b) \in G_b} \ell(f_\theta(x), y) \right\}$$

$\mathcal{B} \in \{correlated, uncorrelated\}$

- By minimizing the maximum loss across bias-aligned and bias-conflicting groups, we aim to encourage the model to perform consistently on both.

- However, this objective requires information about the presence of spurious correlations.

Method

- To remove the need for bias labels, we reformulate the objective.

Assumption: The neural network satisfies that $\hat{\mathcal{L}}_{b_a} < \hat{\mathcal{L}}_{b_c}$.

b_a: bias-aligned group
b_c: bias-conflicting group

Method

- To remove the need for bias labels, we reformulate the objective.

Assumption: The neural network satisfies that $\hat{\mathcal{L}}_{b_a} < \hat{\mathcal{L}}_{b_c}$.

- $$\max_{b \in \mathcal{B}} \hat{\mathcal{L}}_b = \frac{1}{n_{b_c}} \sum_{(x,y,b_c) \in G_{b_c}} \ell(f_\theta(x), y)$$
$$= \frac{1}{n_{b_c}} \sum_{(x,y,b) \in \mathcal{D}} p(b = b_c|x) \ell(f_\theta(x), y)$$
$$= \frac{1}{n} \sum_{(x,y,b) \in \mathcal{D}} \frac{p(b = b_c|x)}{p(b=b_c)} \ell(f_\theta(x), y)$$

Method

- To remove the need for bias labels, we reformulate the objective.

Assumption: The neural network satisfies that $\hat{\mathcal{L}}_{b_a} < \hat{\mathcal{L}}_{b_c}$.

- $$\max_{b \in \mathcal{B}} \hat{\mathcal{L}}_b = \frac{1}{n_{b_c}} \sum_{(x,y,b_c) \in G_{b_c}} \ell(f_\theta(x), y)$$
$$= \frac{1}{n_{b_c}} \sum_{(x,y,b) \in \mathcal{D}} p(b = b_c | x) \ell(f_\theta(x), y)$$
$$= \frac{1}{n} \sum_{(x,y,b) \in \mathcal{D}} \frac{p(b = b_c | x)}{p(b = b_c)} \ell(f_\theta(x), y)$$

- A weighted loss minimization

$$r(x,y,b) = \frac{1}{n} \frac{p(b = b_c | x)}{p(b = b_c)}$$

$$\min_\theta \sum_{(x,y,b) \in \mathcal{D}} r(x,y,b) \ell(f_\theta(x), y)$$

Method

- However, the sampling probability $r(x, y, b)$ still requires explicit bias information.
- To eliminate the need for bias labels, we use the characteristics of the biased model.
- We employ the disagreement between the label y and the biased model's prediction y_{bias} as a proxy for the bias-conflicting group b_c.
- Instead of $r(x, y, b) = \frac{1}{n}\frac{p(b=b_c|x)}{p(b=b_c)}$, we use $\hat{r}(x, y) = \frac{1}{n}\frac{p(y\neq y_{bias}|x)}{p(y\neq y_{bias})}$.
- The final objective

$$\min_{\theta} \sum_{(x,y)\in\mathcal{D}} \hat{r}(x, y)\ell(f_\theta(x), y)$$

Method

- Algorithm

    1. Train the biased model

    2. Calculate the sampling probability $\hat{r}(x, y)$

    3. Initialize the debiased model with the biased model to satisfy assumption

    4. Train the debiased model

Experiments

- Synthetic datasets

	C-MNIST			MB-MNIST		
Ratio (%)	0.5	1	5	10	20	30
ERM	60.94 (0.97)	79.13 (0.73)	95.12 (0.24)	25.23 (1.16)	62.06 (2.45)	87.61 (1.60)
JTT	85.84 (1.32)	95.07 (3.42)	96.56 (1.23)	25.34 (1.45)	68.02 (3.23)	85.44 (3.44)
DFA	94.56 (0.57)	96.87 (0.64)	98.35 (0.20)	25.75 (5.38)	61.62 (2.60)	88.36 (2.06)
PGD	96.88 (0.28)	98.35 (0.12)	**98.62 (0.14)**	61.38 (4.41)	89.09 (0.97)	90.76 (1.84)
LC	97.25 (0.21)	97.34 (0.16)	97.44 (0.37)	25.86 (8.68)	71.23 (1.71)	89.57 (2.50)
DPR (Ours)	**97.52 (0.33)**	**98.40 (0.03)**	**98.62 (0.12)**	**62.21 (4.02)**	**89.11 (1.65)**	**94.04 (0.26)**

- Real-world datasets

	BAR	BFFHQ		CelebA		CivilComments-WILDS	
Accuracy (%)	Conflicting	Unbiased	Conflicting	Average	Worst	Average	Worst
ERM	63.15 (1.06)	77.77 (0.45)	55.93 (0.64)	94.9 (0.3)	47.7 (2.1)	92.1 (0.4)	58.6 (1.7)
JTT	63.62 (1.33)	77.93 (2.16)	56.13 (0.83)	88.1 (0.3)	81.5 (1.7)	91.1 (-)	69.3 (-)
DFA	64.70 (2.06)	82.77 (1.40)	66.00 (2.00)	-	-	-	-
CNC	-	-	-	89.9 (0.5)	88.8 (0.9)	81.7 (0.5)	68.9 (2.1)
PGD	65.39 (0.47)	84.20 (1.15)	70.07 (2.00)	88.6 (-)	88.8 (-)	92.1 (-)	70.6 (-)
LC	63.45 (2.14)	83.97 (0.83)	70.60 (0.60)	-	88.1 (0.8)	-	70.3 (1.2)
DPR (Ours)	**66.11 (3.29)**	**87.57 (1.22)**	**76.80 (2.51)**	90.7 (0.6)	**88.9 (0.6)**	82.9 (0.7)	**70.9 (1.7)**

Thank you