VISUAL INTELLIGENCE AND MULTIMEDIA ANALYTICS LABORATORY

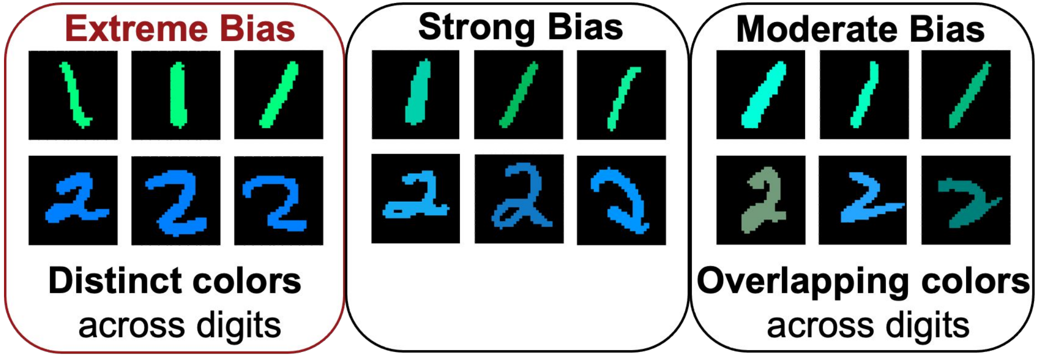
Information-Theoretic Bounds on The Removal of Attribute-Specific Bias From Neural Networks

Jiazhi Li, Mahyar Khayatkhoei, Jiageng Zhu, Hanchen Xie,
Mohamed E. Hussein, Wael AbdAlmageed
University of Southern California

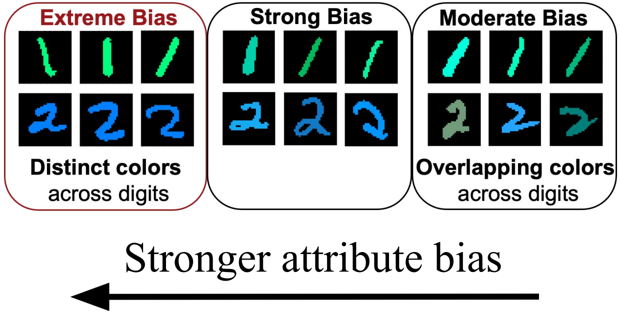
USC Viterbi Information Sciences Institute

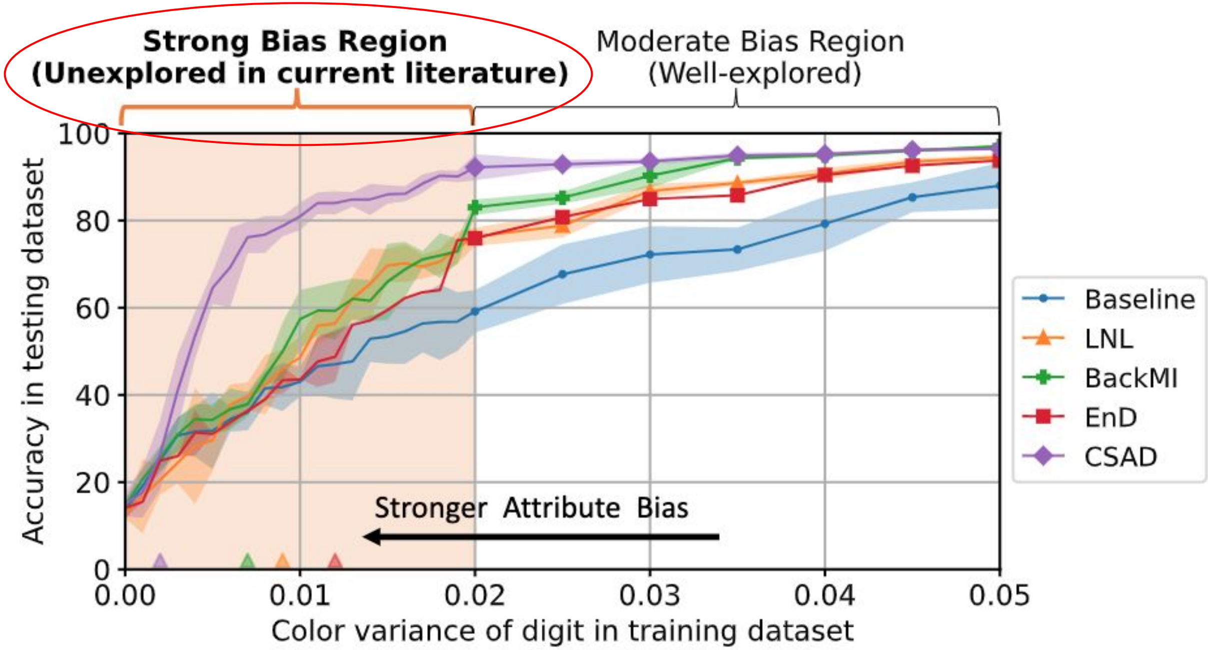
BACKGROUND - ATTRIBUTE BIAS

Attribute bias is defined as the dependence between model prediction and protected attributes.
For example, in Colored MNIST, a benchmark dataset to study attribute bias, where the prediction target is digit and the protected attribute is color, given the spurious correlation between digit and color in training set, attribute bias causes the digit prediction to rely on color in testing set.
In general, ensuring a neural network is not relying on protected attributes for predictions is crucial in advancing fair and trustworthy artificial intelligence.



MOTIVATION – BREAKING POINT





USC Viterbi
Information Sciences Institute

INFORMATION-THEORETIC BOUND

$$\underbrace{\uparrow I(Z;Y)}_{\text{Best Performance}} \leq \underbrace{I(Z;A)}_{\text{Remained Bias in Feature}\downarrow} + \underbrace{H(Y|A)}_{\text{Bias Strength}}$$

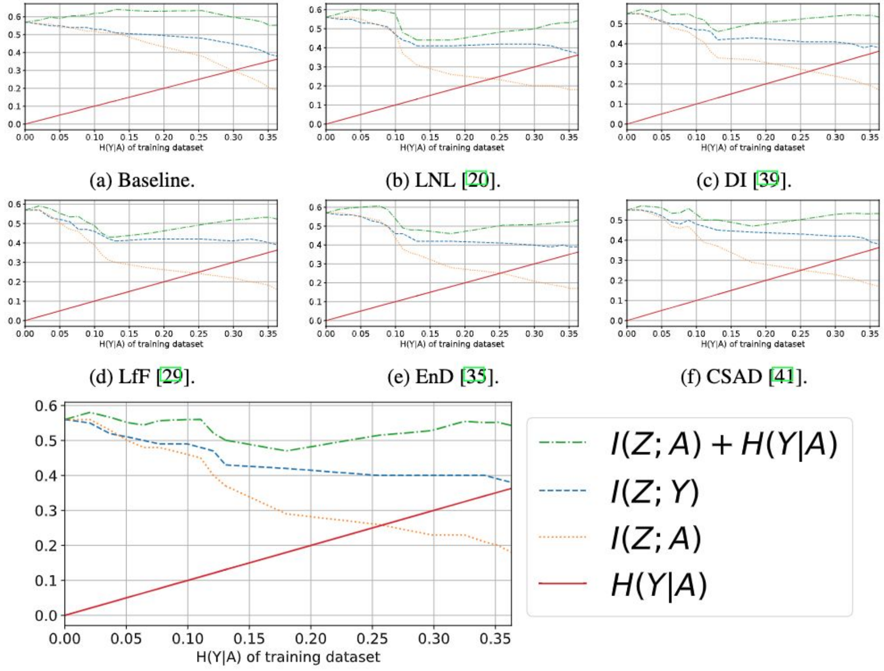
Z: Learnt Feature
Y: Target of Prediction (e.g., digit)
A: Protected Attribute (e.g., color)

INFORMATION-THEORETIC BOUND

Best Performance ↑ $I(Z;Y) \leq I(Z;A) + H(Y|A)$

Bias Strength: $H(Y|A)$

Remained Bias in Feature ↓: $I(Z;A)$



(a) Baseline. (b) LNL [20]. (c) DI [39]. (d) LfF [29]. (e) EnD [35]. (f) CSAD [41]. (g) BCL [17].

USC Viterbi Information Sciences Institute

EXTREME BIAS POINT $H(Y|A)=0$

No method can effectively remove the bias $I(Z;A)$ compared to baseline.

CelebA dataset

Method	Test Accuracy		Mutual Information	
	Unbiased ↑	Bias-conflicting ↑	$I(Z;A)\downarrow$	Δ (%) ↑
Random guess	50	50	0.57	0.00
Baseline	66.11± 0.32	33.89± 0.45	0.57± 0.01	0.00
LNL [19]	64.81± 0.17	29.72± 0.26	0.56± 0.06	1.75
DI [36]	66.83± 0.44	33.94± 0.65	0.55± 0.02	3.51
LfF [26]	64.43± 0.43	30.45± 1.63	0.57± 0.03	0.00
EnD [32]	66.53± 0.23	31.34± 0.89	0.57± 0.05	0.00
CSAD [37]	63.24± 2.36	29.13± 1.26	0.55± 0.04	3.51
BCL [16]	65.30± 0.51	33.44± 1.31	0.56± 0.07	1.75

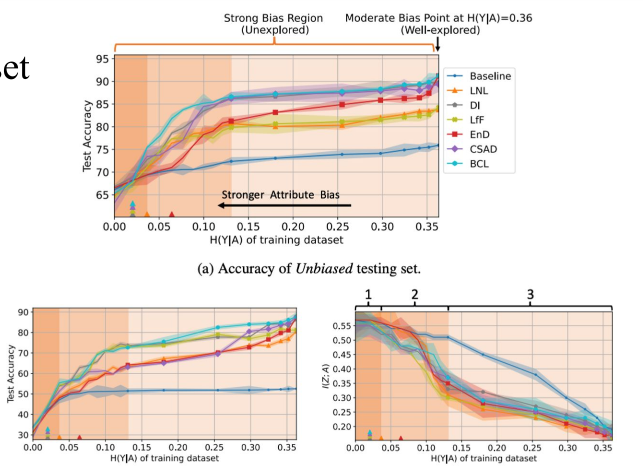
Adult dataset

Method	Test Accuracy		Mutual Information	
	Unbiased ↑	Bias-conflicting ↑	$I(Z;A)\downarrow$	Δ (%) ↑
Random guess	50	50	0.69	0.00
Baseline	50.59±0.54	1.19±0.83	0.69±0.00	0.00
LNL [19]	50.10±0.18	0.43±0.46	0.69±0.01	0.00
DI [36]	50.61±0.28	0.65±0.64	0.69±0.01	0.00
LfF [26]	50.33±0.34	0.78±0.65	0.69±0.01	0.00
EnD [32]	50.59±0.75	1.18±0.96	0.69±0.00	0.00
CSAD [37]	50.76±2.22	1.43±2.46	0.69±0.01	0.00
BCL [16]	50.83±1.34	0.52±0.83	0.69±0.00	0.00

STRONG BIAS REGION $H(Y|A)>0$

As bias strength increases, performance of all methods declines to baseline at the breaking point (shown by ▲).

CelebA dataset

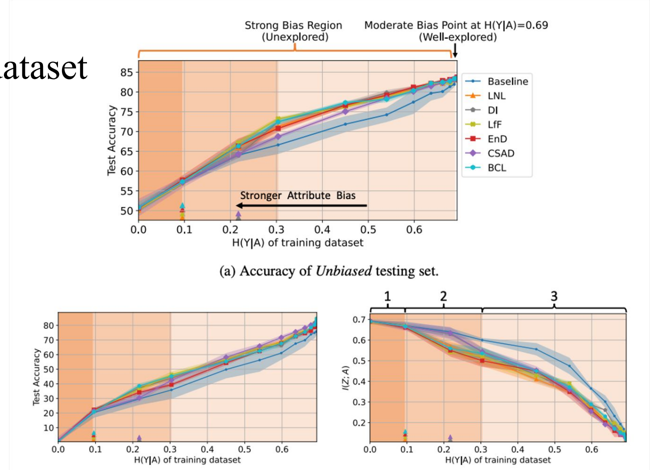


(a) Accuracy of *Unbiased* testing set.

(b) Accuracy of *Bias-conflicting* testing set.

(c) Attribute bias $I(Z;A)$ of training set.

Adult dataset



(a) Accuracy of *Unbiased* testing set.

(b) Accuracy of *Bias-conflicting* testing set.

(c) Attribute bias $I(Z;A)$ of training set.

MAIN TAKEAWAYS

1. When a protected attribute is strongly predictive of a target, attribute bias removal methods become ineffective.
2. Cautions against the use of attribute bias removal methods in datasets with potentially strong bias (e.g., small datasets) and motivates the design of future methods that can work even in the strong bias setting.

Thank you!

Paper & Code