DeNetDM: Debiasing by Network Depth Modulation

NEURAL INFORMATION PROCESSING SYSTEMS 2024


Silpa Vadakkeeveetil Sreelatha*[1]


Adarsh Kappiyath*[1]


Abhra Chaudhuri[1,2,3]


Anjan Dutta[1]

UNIVERSITY OF SURREY[1]

University of Exeter[2]

FUJITSU[3]

* Equal contribution.

The Spurious Trap

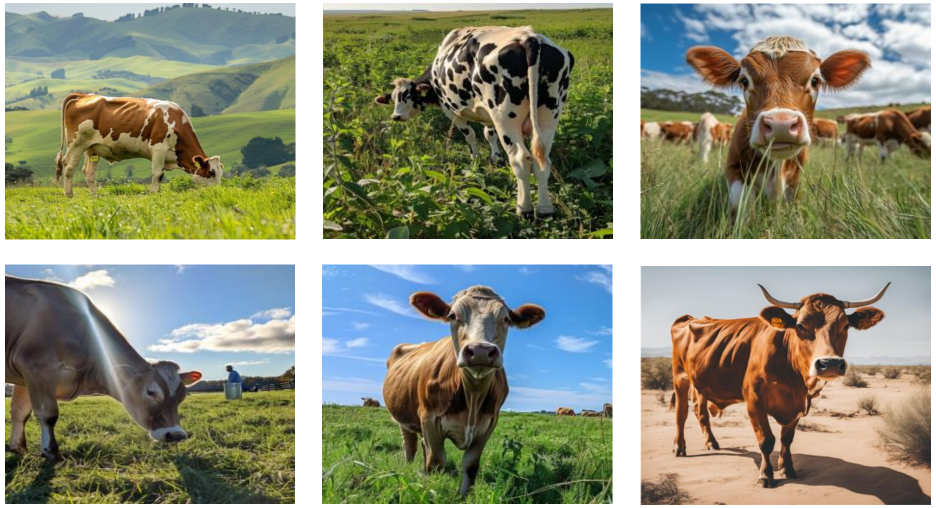
Common Sight

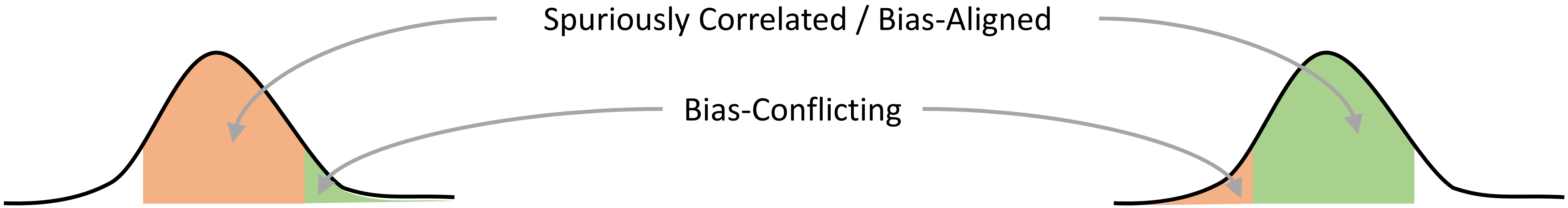
Rare but not impossible

Camels

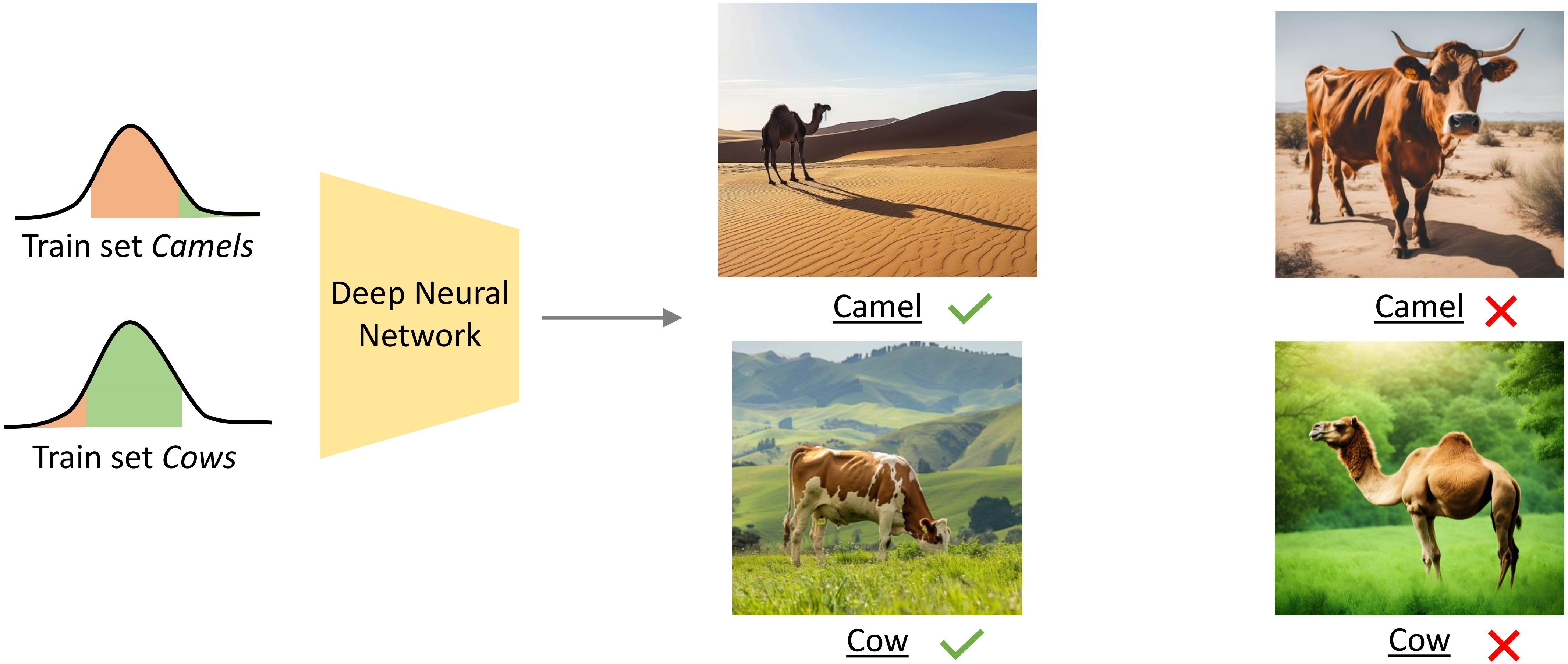
Cows

The Spurious Trap

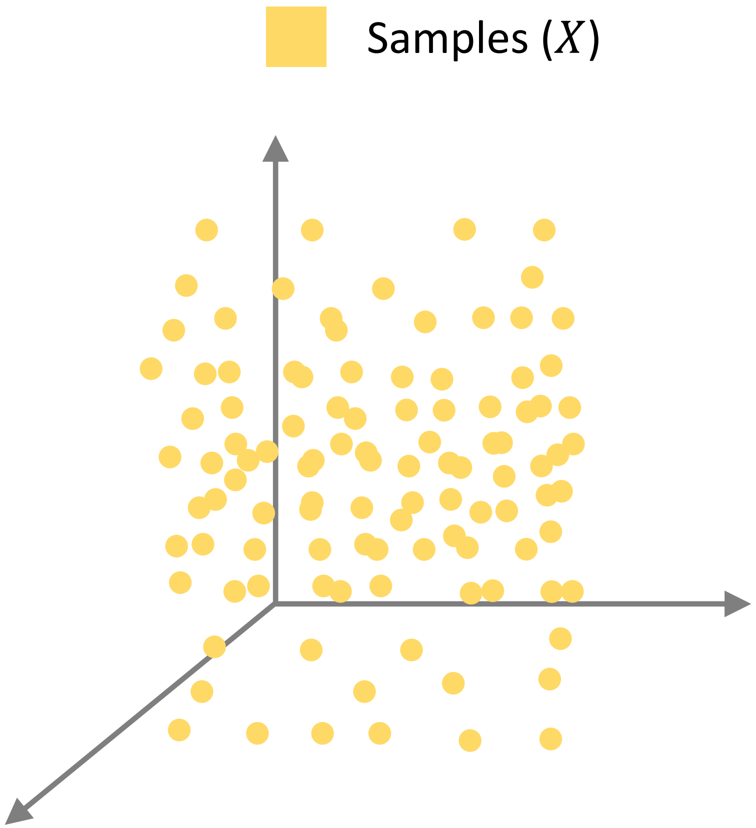




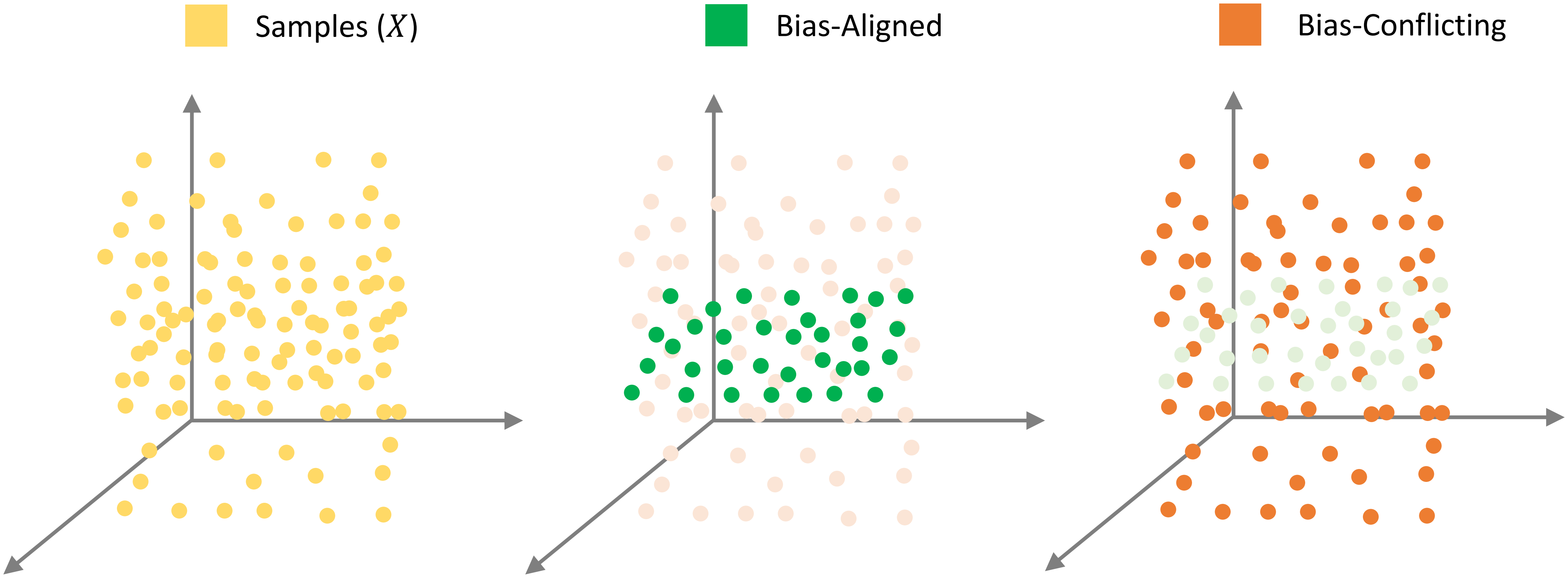
The Spurious Trap



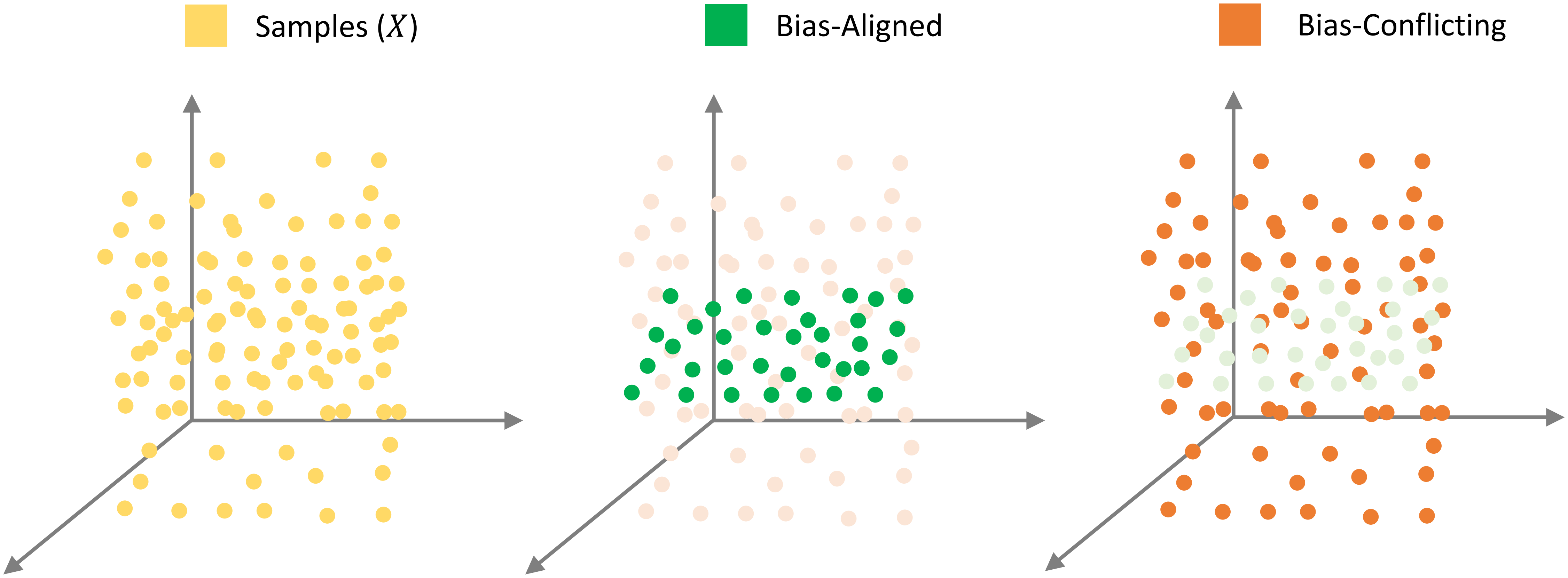
Theorem 1: Partition Rank



Theorem 1: Partition Rank

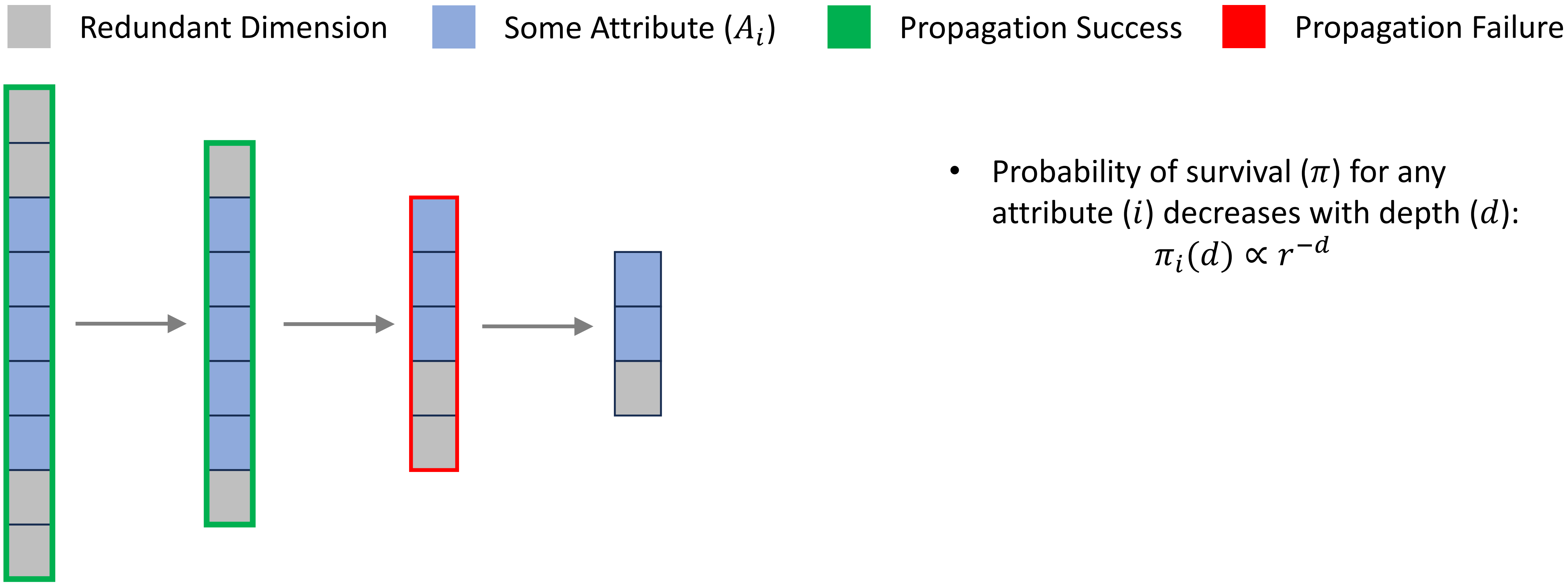


Theorem 1: Partition Rank



Theorem: $rank$(Bias-Aligned) $\leq$ $rank$(Bias-Conflicting)

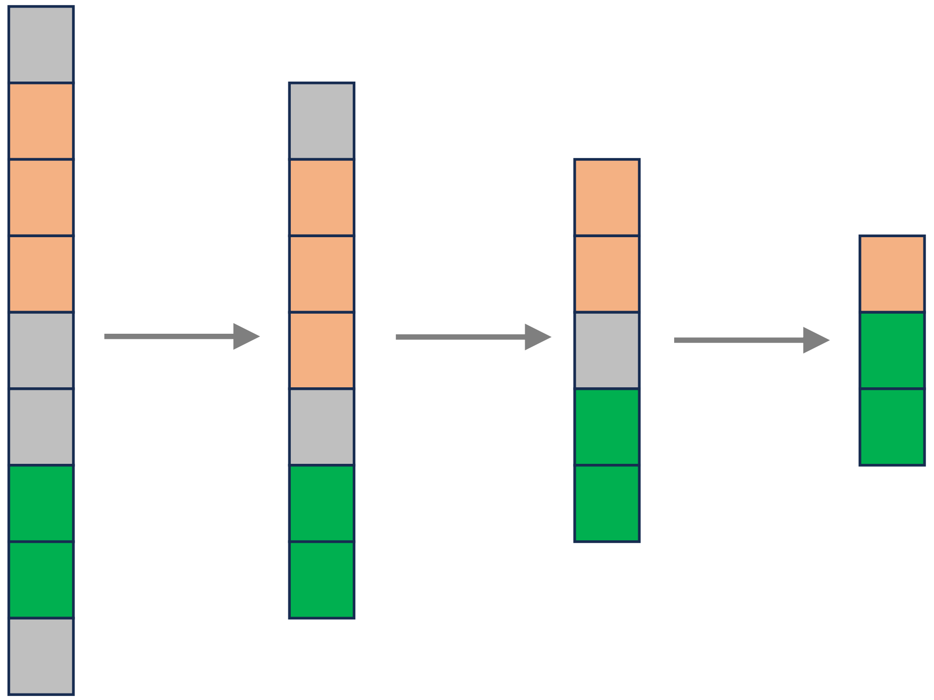
Propagation Bottleneck



- Probability of survival (π) for any attribute (i) decreases with depth (d):
  $$\pi_i(d) \propto r^{-d}$$

The Simplicity Bias and Survival Probability

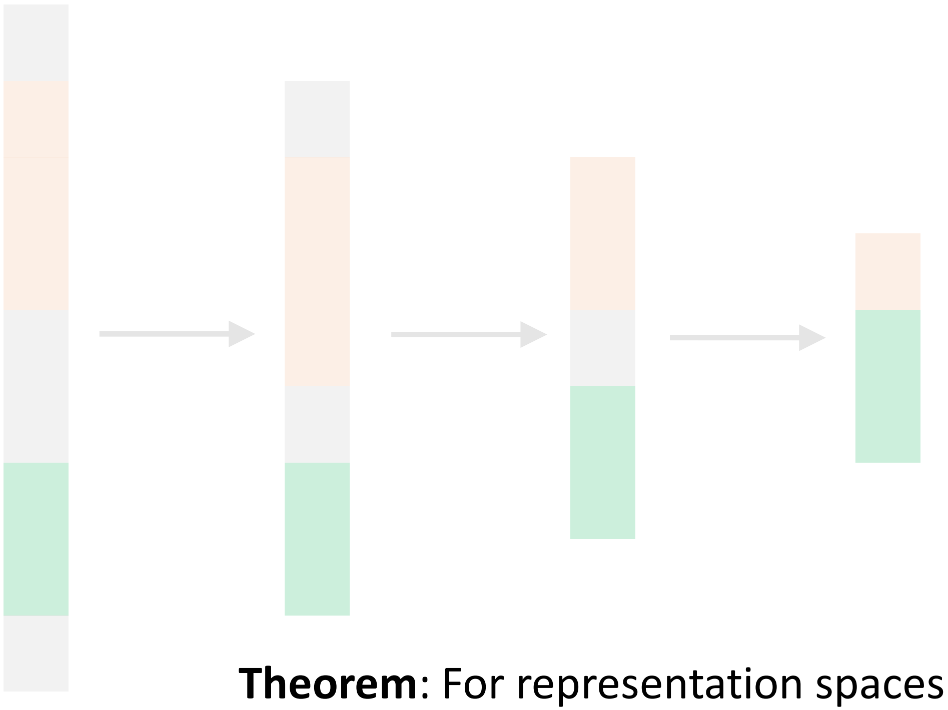
Redundant Dimension — Attribute 1 (A_1) — Attribute 2 (A_2)



- Probability of survival (π) for any attribute (i) decreases with depth (d):
$$\pi_i(d) \propto r^{-d}$$
- Attributes with a lower rank have a higher probability of survival at any given depth:
$$\pi_i(d) \geq \pi_j(d),$$
where $rank(A_i) \leq rank(A_j)$.

Theorem 2: Depth-Rank Duality (Implicit Rank Regularization)

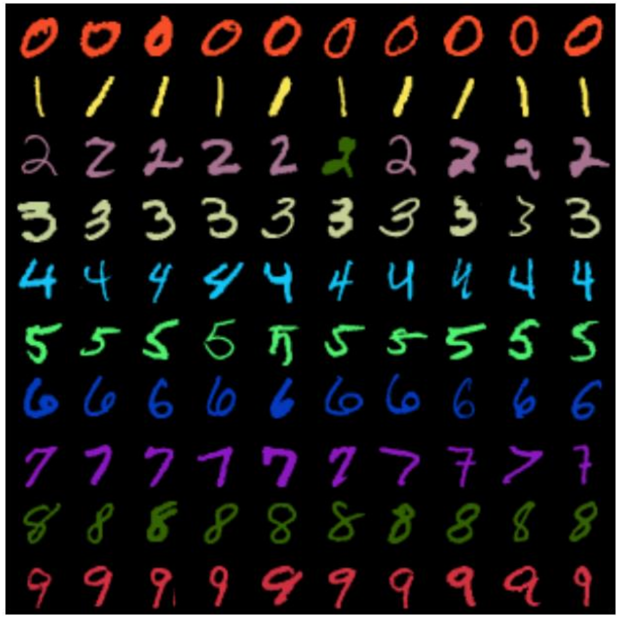
Redundant Dimension   Attribute 1 (A_1)   Attribute 2 (A_2)



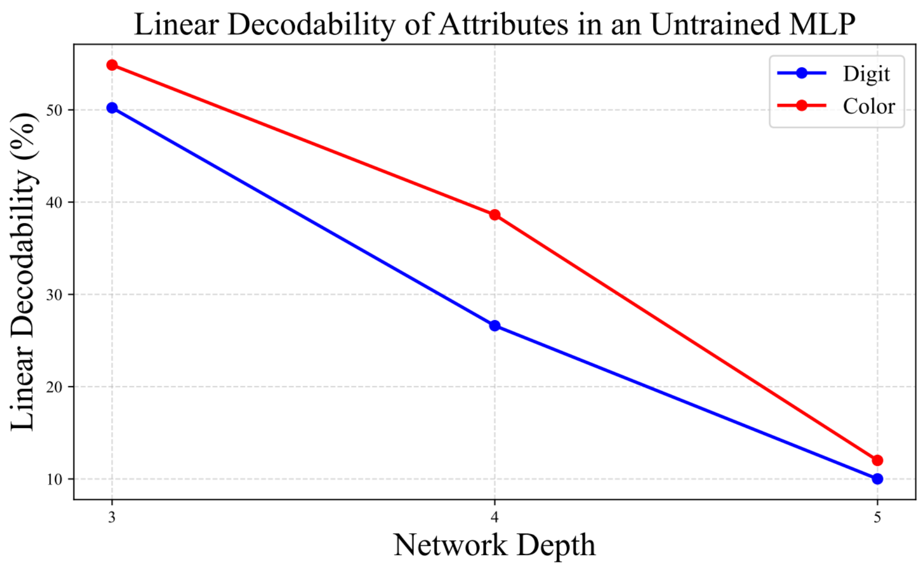
- Probability of survival (π) for any attribute (i) decreases with depth (d):
$$\pi_i(d) \propto r^{-d}$$
- Attributes with a lower rank have a higher probability of survival at any given depth:
$$\pi_i(d) \geq \pi_j(d),$$
where $rank(A_i) \leq rank(A_j)$.

Theorem: For representation spaces at <u>deeper layers</u>, <u>lower rank attributes</u> are more likely to <u>minimize the empirical risk</u>.

Empirical Evidence – Linear Decodability (Untrained)



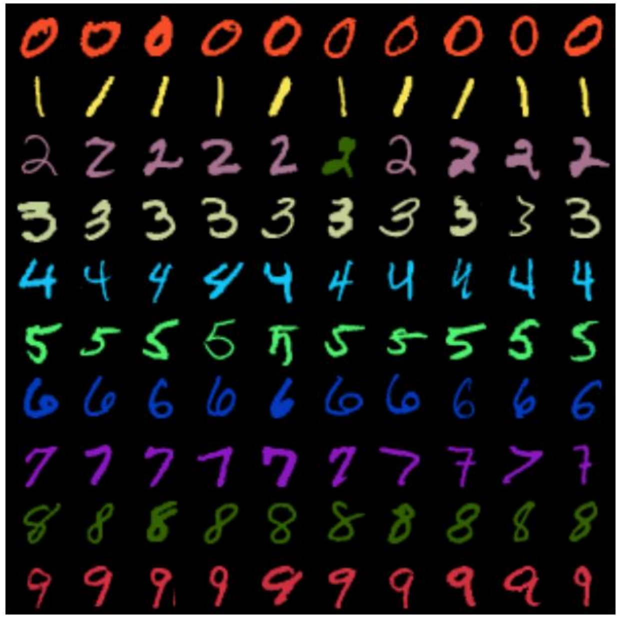
Dataset: Colored MNIST
(Color-Digit Spurious Correlation)



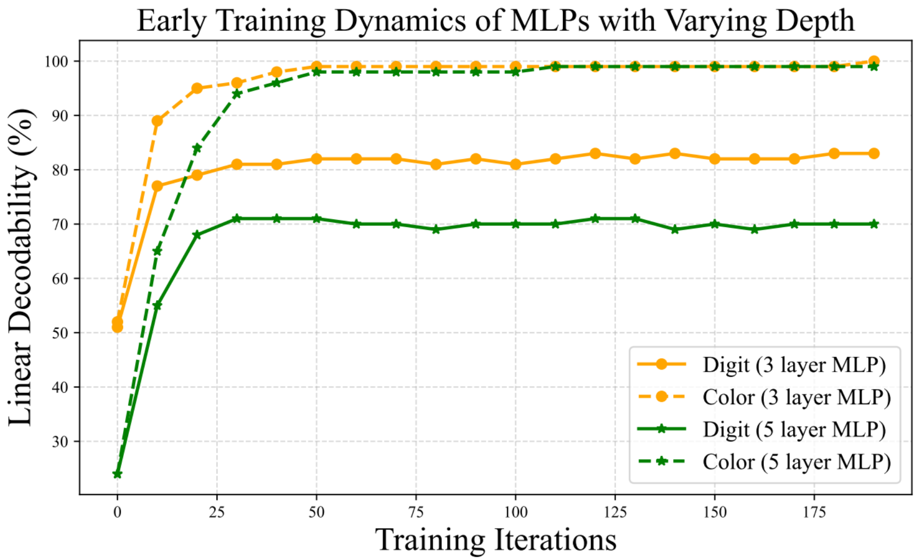
Experiment

Result:
Core (higher) rank attribute – "digit", harder to decode from deeper MLPs

Empirical Evidence – Linear Decodability (under SGD)



Dataset: Colored MNIST
(Color-Digit Spurious Correlation)

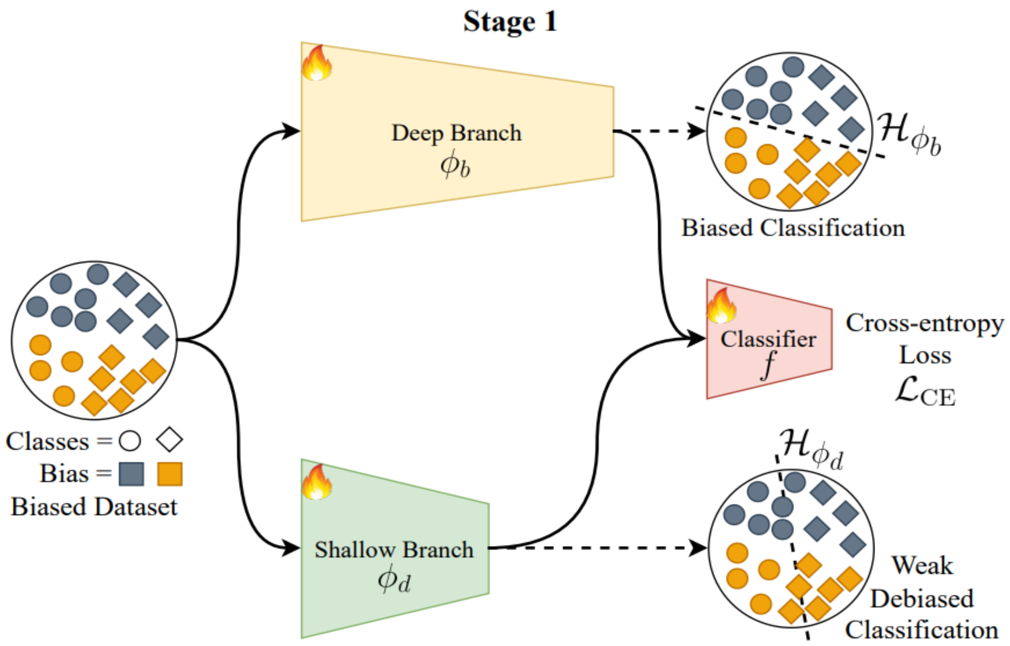


Experiment

Result:
Core (higher) rank attribute – "digit", harder to decode from deeper MLPs – a characteristic retained under SGD.

DeNetDM: Debiasing by Network Depth Modulation

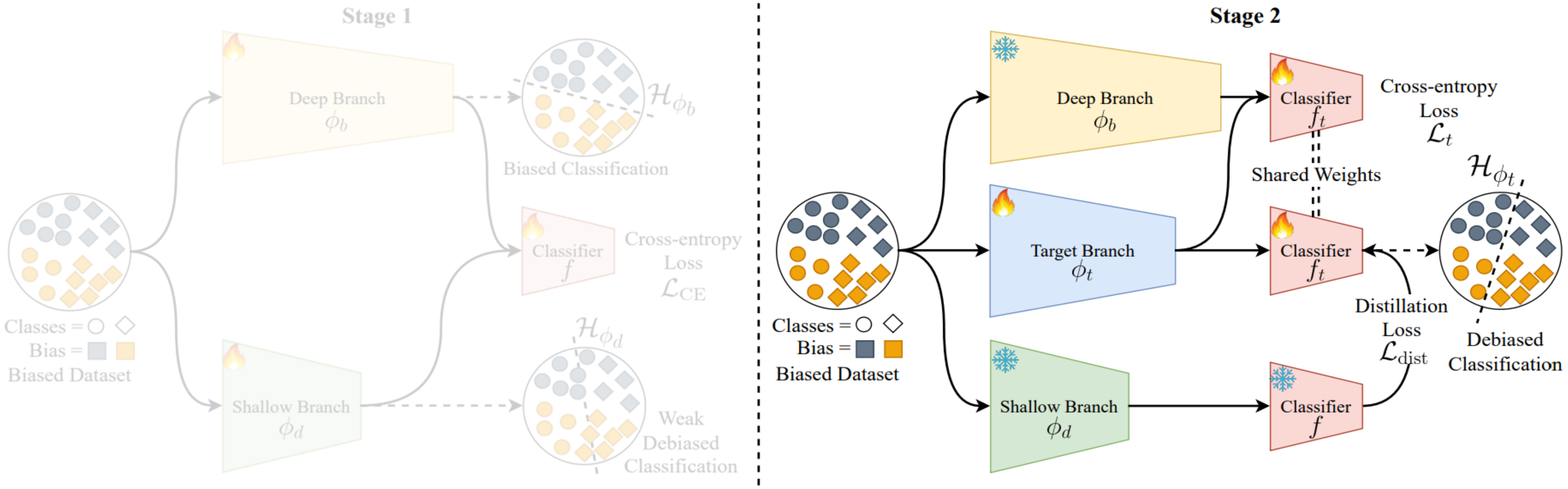
Stage 1: Bias Identification through depth modulation.



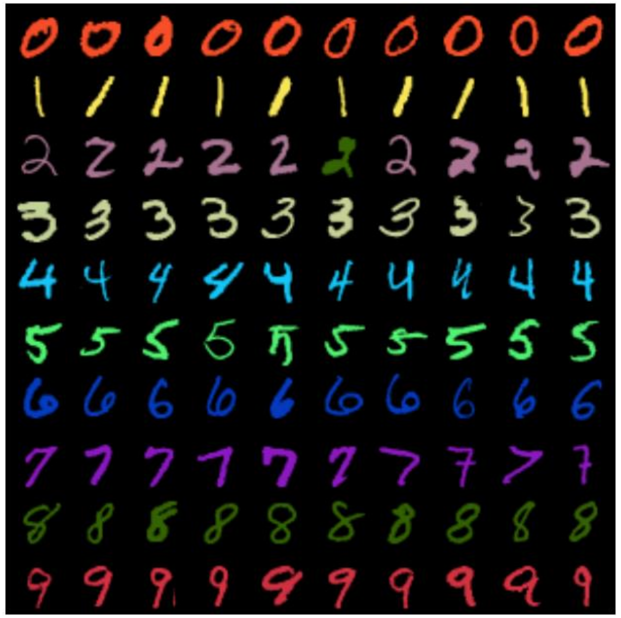
DeNetDM: Debiasing by Network Depth Modulation

Stage 1: Bias Identification through depth modulation.

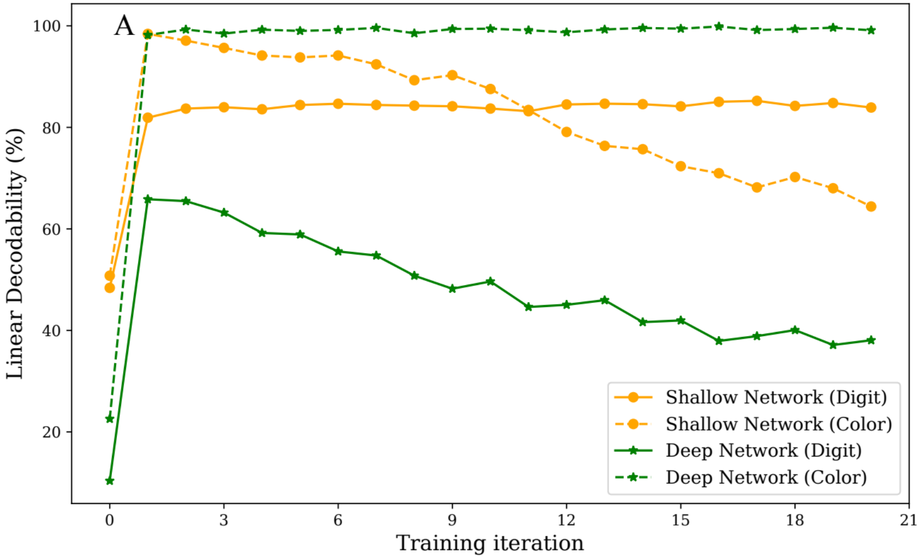
Stage 2: Bias mitigation via knowledge distillation.



Linear Decodability (under SGD) of DeNetDM



Dataset: Colored MNIST
(Color-Digit Spurious Correlation)



Experiment: Linear Decodability
Dynamics of DNetDM Training

Result:
Accentuated Simplicity Bias

Comparison with State-of-the-Art

Methods	Group Info	CMNIST 0.5	CMNIST 1.0	CMNIST 2.0	CMNIST 5.0	C-CIFAR10 0.5	C-CIFAR10 1.0	C-CIFAR10 2.0	C-CIFAR10 5.0
Group DRO	✓	59.67	71.33	76.30	84.40	33.44	38.30	45.81	57.32
ERM	✗	35.34 (0.13)	50.34 (0.16)	62.29 (1.47)	77.63 (0.13)	23.08 (1.25)	25.82 (0.33)	30.06 (0.71)	39.42 (0.64)
JTT	✗	53.03 (3.89)	61.68 (2.02)	74.23 (3.21)	85.03 (1.10)	24.73 (0.60)	26.90 (0.31)	33.40 (1.06)	42.20 (0.31)
LfF	✗	63.39 (1.97)	74.01 (2.21)	80.48 (0.45)	85.39 (0.94)	28.57 (1.30)	33.07 (0.77)	39.91 (0.30)	50.27 (1.56)
DFA	✗	59.12 (3.15)	71.04 (1.02)	82.86 (2.27)	88.29 (1.50)	29.95 (0.71)	36.49 (1.79)	41.78 (2.29)	51.13 (1.28)
LC	✗	63.48 (5.22)	78.41 (1.95)	83.63 (1.43)	88.18 (1.59)	34.56 (0.69)	37.34 (1.26)	**47.81 (2.00)**	54.55 (1.26)
DeNetDM	✗	**74.72 (0.99)**	**85.22 (0.76)**	**89.29 (0.51)**	**93.54 (0.22)**	**38.93 (1.16)**	**44.20 (0.77)**	47.35 (0.70)	**56.30 (0.42)**

- **Strong generalization** to datasets with both simple and complex bias / core attributes.
- Around **5% improvement** margins.
- **No bias labels** / supervision or augmentation.

Methods	Group Info	BAR 1.0	BAR 5.0	BFFHQ 1.0	CelebA -
ERM	✗	57.65 (2.36)	68.60 (2.25)	56.7 (2.7)	47.02
JTT	✗	58.17 (3.30)	68.53 (3.29)	65.3 (2.5)	76.80
LfF	✗	57.71 (3.12)	67.48 (0.46)	62.2 (1.6)	-
DFA	✗	52.31 (1.00)	63.50 (1.47)	63.9 (0.3)	65.26
LC	✗	70.94 (1.46)	74.32 (2.42)	70.0 (1.4)	-
DeNetDM (ours)	✗	**73.84 (2.56)**	**79.61 (3.18)**	**75.7 (2.8)**	**81.04**

Conclusions

- Explored the relationships between the depth of a neural network, the rank of an attribute, and the susceptibility to spurious correlations.

Conclusions

- Explored the relationships between the depth of a neural network, the rank of an attribute, and the susceptibility to spurious correlations.
- Introduced the idea of depth modulation for identifying and mitigating biases in neural networks.

Conclusions

- Explored the relationships between the depth of a neural network, the rank of an attribute, and the susceptibility to spurious correlations.
- Introduced the idea of depth modulation for identifying and mitigating biases in neural networks.
- Strong empirical results confirming theoretical claims, surpassing SOTA on numerous benchmarks.

DeNetDM: Debiasing by Network Depth Modulation

Project Page

https://vssilpa.github.io/denetdm/