Rectifying Shortcut Learning via Cellular Differentiation in Deep Learning Neurons

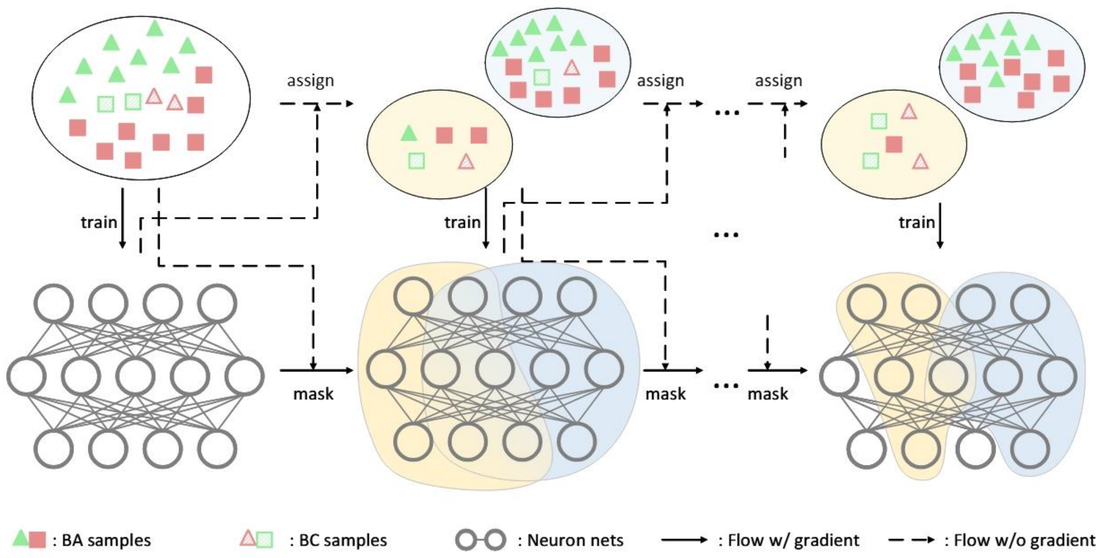
Hongjing Niu, Hanting Li, Guoping Wu, Feng Zhao, Bin Li

University of Science and Technology of China

Overview

- The paper proposes a method for feature differentiation in models, which can learn more disentangled feature representations and enables the model to better adapt to complex task scenarios.
- The proposed method offers a new strategy to avoid excessive reliance on shortcut, which no longer requires prior knowledge of shortcut features.
- This work provides a new approach for debiasing tasks and achieves excellent performance without introducing additional priors or assumptions.

Framework



$$L_{GCE}(x,y;M) = \frac{1-p(M(x),y)^q}{q},$$

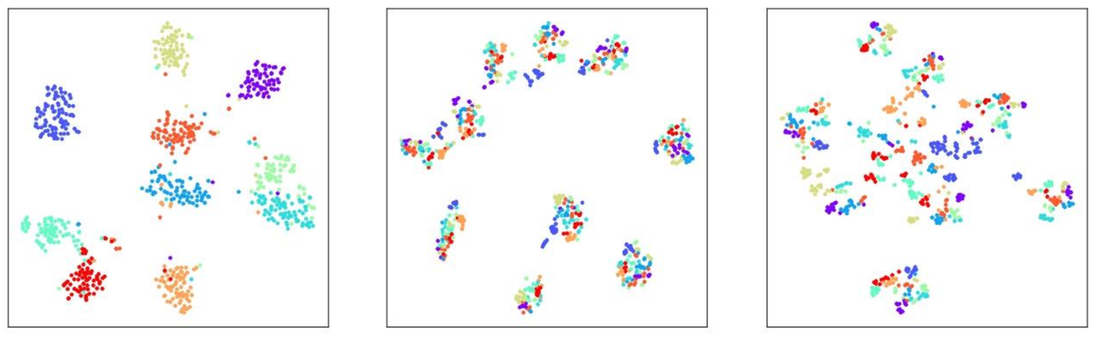
$$L_m = L_s(m_a,D_a) + L_s(m_b,D_b) + IoU(m_a,m_b).$$

$$L_s = GCE(x,y,M \circ m_a) + \lambda \sum_{m_a} |\theta_i|.$$

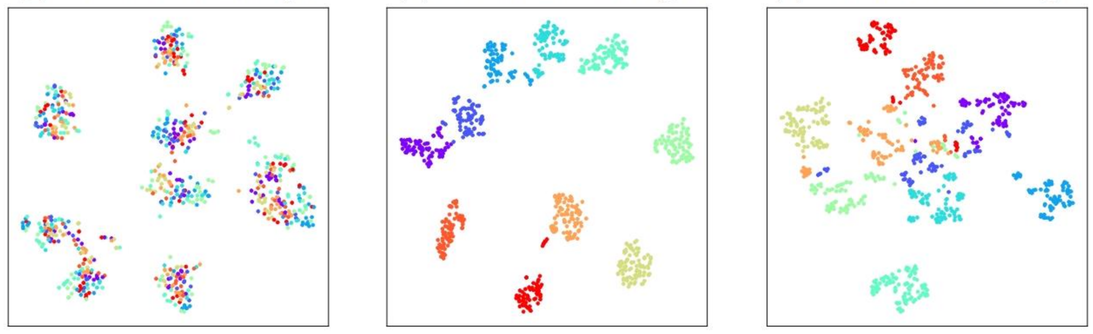
Experiments.

Dataset	Ratio (%)	Vanilla‡	EnD‡	ReBias‡	LFF‡	DFA‡	DCWP‡	Ours
		✓	×	×	✓	✓	✓	✓
CMNIST	0.5	62.36	84.32	69.12	83.73	86.74	93.41	**93.68**$_{\pm 0.47}$
	1.0	81.73	94.98	84.65	88.44	93.15	95.98	**96.15**$_{\pm 0.16}$
	2.0	89.33	97.01	91.96	92.67	95.15	97.16	**97.23**$_{\pm 0.16}$
	5.0	95.22	98.00	96.74	94.90	96.76	**98.02**	97.86$_{\pm 0.04}$
CIFAR10-C	0.5	22.02	23.93	21.73	27.02	27.86	35.90	**38.60**$_{\pm 0.39}$
	1.0	28.00	27.61	28.09	31.44	34.62	41.56	**45.53**$_{\pm 1.58}$
	2.0	34.63	36.62	35.57	38.49	41.95	49.01	**51.78**$_{\pm 0.61}$
	5.0	45.66	43.67	48.22	46.16	49.15	56.17	**59.04**$_{\pm 0.90}$
BFFHQ	0.5	52.25	59.80	54.90	56.50	55.50	60.35	**60.70**$_{\pm 1.30}$

The proposed method achieves better performs on most of the tested o.o.d. dataset. And the method can be used on other datasets.

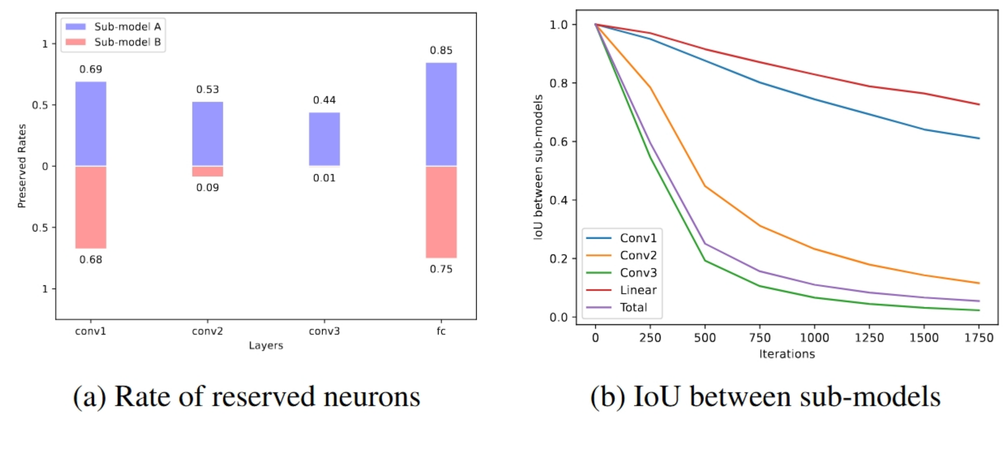


(a) sub-model A, target (b) sub-model B, target (c) vanilla model, target



(d) sub-model A, shortcut (e) sub-model B, shortcut (f) vanilla model, shortcut

t-SNE results show that different sub-models in out method can learn different features.



(a) Rate of reserved neurons (b) IoU between sub-models

The experiments show that sub-models share some features in low level layer. And the IOU between sub-models decreases.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants U19B2044 and the Anhui Provincial Natural Science Foundation under Grant 2108085UD12. We acknowledge the support of GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC.