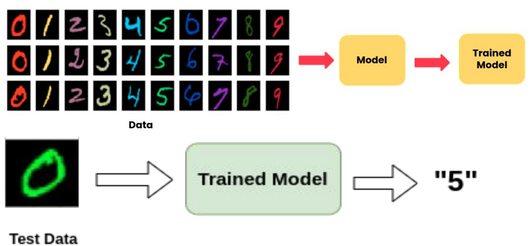
Gradient Based Activations for Accurate Bias Free Learning

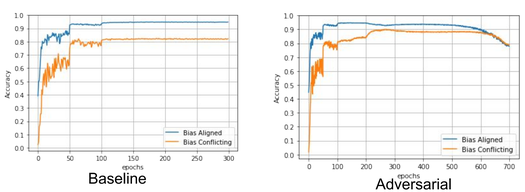
Vinod K. Kurmi[1], Rishabh Sharma[2], Yash Vardhan Sharma[2], Vinay P. Namboodiri[3]

KU Leuven, Belgium[1] IIT Roorkee, India[2] University of Bath, UK[3]

Introduction

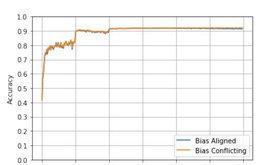
- Adversarial methods are very popular to tackle for bias free learning
- We propose to use a biased discriminator to improve the accuracy of adversarial methods while debiasing the feature representation.
- We use gradients of the discriminator and propose a masking scheme for the features, we term as Gradient Based Activation (GBA).
- Use of GBA provides us a masking rule to drop certain features in order to do unbiased training. The effectiveness of this approach is demonstrated and validated





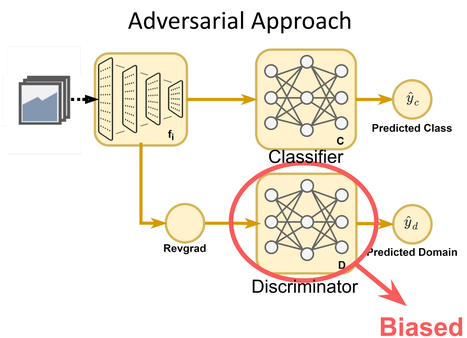
Baseline

Adversarial



GBA

Proposed Approach



Using the Biased Discriminator

- When the discriminator is correct:
  Mask the features used by the discriminator
- When the discriminator is incorrect
  Emphasize the features used by the discriminator

$$ind^d = \begin{cases} 1, & \text{if } \operatorname{argmax}(\hat{y}_i^d) = y_i^d \\ -1 & \text{otherwise} \end{cases} \quad f_i = F(x_i; \theta_f); \quad \hat{y}_i^d = D(f_i, \theta_d)$$

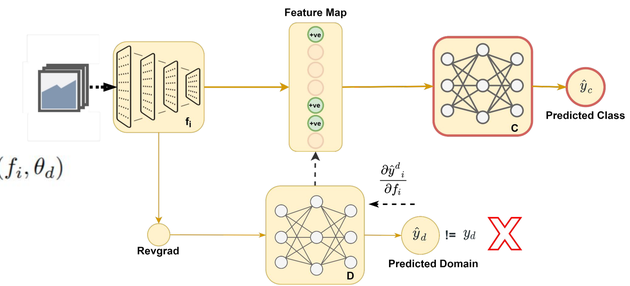
$$\hat{y} = max(\hat{y}_i^d)$$

$$g_i^d = \frac{\partial \hat{y}}{\partial f_i}$$

$$a_i^d = \begin{cases} 0, & \text{if } (g_i^d.ind^d) > 0. \\ 1 & \text{if } (g_i^d.ind^d) \leq 0 \end{cases}$$

$$f_i^{cls} = f_i * a_i^d$$

$$\hat{y}_i^c = C(f_i^{cls}, \theta_c);$$



$$\mathcal{L}_c = \frac{1}{N} \sum_{x_i \in \mathcal{D}} \mathcal{L}(\hat{y}_i^c, y_i^c) \quad \mathcal{L}_d = \frac{1}{N} \sum_{x_i \in \mathcal{D}} \mathcal{L}(\hat{y}_i^d, y_i^d)$$

Results

CelebA Dataset

Model	Heavy Makeup				Hair Color			
	Aligned	Conflicting	Mean	Bias Gap	Aligned	Conflicting	Mean	Bias Gap
Vanilla	92.44 ± 0.74	31.46 ± 2.45	61.95 ± 1.28	60.98 ± 2.56	90.58 ± 0.34	57.35 ± 0.21	73.97 ± 0.20	33.23 ± 0.40
LfF (Nam et al. 2020)	83.85 ± 1.68	45.54 ± 4.28	64.69 ± 2.29	38.31 ± 4.60	88.85 ± 1.27	80.24 ± 2.16	84.55 ± 1.25	8.61 ± 2.5
Domain Independent (Wang et al. 2020)	79.88±1.71	43.24±4.33	61.56±2.31	36.64 ± 5.64	90.97±3.71	79.25±3.33	85.11±2.67	7.44 ± 3.21
Group DRO (Sagawa et al. 2020)	79.28±1.20	46.24±3.61	62.76±2.22	33.04 ± 3.22	89.68±0.65	81.41±1.47	85.55±0.88	8.27 ± 2.01
Adversarial	92.07 ± 2.88	33.79 ± 3.81	62.93 ± 2.38	58.28 ± 4.77	93.4 ± 0.91	62.75 ± 3.47	78.08 ± 1.79	30.65 ± 5.59
Adversarial with GBA	81.49 ± 1.91	**49.79± 3.15**	**65.64 ± 1.55**	**31.70 ± 3.10**	90.67 ± 1.01	**83.28 ± 1.83**	**86.98 ± 1.04**	**7.39 ± 2.09**

CIFAR-I Dataset

Model Name	Model	Accuracy (↑)			Bias (GAP)(↓)
		Aligned	Conflicting	Mean	
Baseline	N-way Softmax	87.94 ± 0.36	69.39 ± 0.42	78.67 ± 0.27	18.55 ± 0.55
LfF(Nam et al. 2020)	N-way Softmax	87.01 ± 0.63	56.87 ± 0.72	71.93 ± 0.47	30.14 ± 0.95
Domain Independent(Wang et al. 2020)	N-way Classifier per Domain	88.39 ± 0.15	78.14 ± 0.13	83.26 ± 0.01	10.25 ± 0.20
Adversarial	Gradient Reversal	85.52 ± 0.65	75.74 ± 0.29	80.63 ± 0.35	9.78 ± 0.71
Adversarial with GBA	Proposed	88.81 ± 0.19	**79.46 ± 0.22**	**84.41 ± 0.14**	**9.35 ± 0.29**

Conclusion

- Identify a key drawback of using adversarial methods to debias models
- Developed a novel masking scheme to address the same
- We gave extensive experimental evaluation to validate our findings
- Provide insights to the learning of a classifier

Project Page:
https://vinodkkurmi.github.io/GBA/

Contact
vinod.kurmi@kuleven.be,
rsharma1@me.iitr.ac.in,
ysharma@me.iitr.ac.in