Improving Group-based Robustness and Calibration via Ordered Risk and Confidence Regularization

Anonymous Authors[1]

Abstract

Neural network trained via empirical risk minimization achieves high accuracy on average but low accuracy on certain groups, especially when there is a *spurious correlation*. To construct the unbiased model from spurious correlation, we build a hypothesis that the inference to the samples without spurious correlation should take relative precedence over the inference to the spuriously biased samples. Based on the hypothesis, we propose the relative regularization to induce the training risk of each group to follow the specific order, which is sorted according to the degree of spurious correlation for each group. In addition, we introduce the ordering regularization based on the predictive confidence of each group to improve the model calibration, where other robust models still suffer from large calibration errors. These result in our complete algorithm, Ordered Risk and Confidence regularization (ORC). Our experiments demonstrate that ORC improves both the group robustness and calibration performances against the various types of *spurious correlation* in both synthetic and real-world datasets.

1. Introduction

Spurious correlation is a correlation between two factors, which appear causally related to one another but are not. Empirical risk minimization achieves the low test error on average by training a model to minimize the average loss. However, it incurs high error on certain groups in a dataset.(Sagawa et al., 2019; Cao et al., 2019; Hashimoto et al., 2018). One of the reasons for the performance differences across groups is the presence of *spurious correlation*. A classification model has a risk of relying on the background features in recognizing the objects, which results in the misclassification of the rare images (e.g. birds on the grass). The performance degradations from learning the spurious correlation occur in many applications (Gururangan et al., 2018; McCoy et al., 2019).

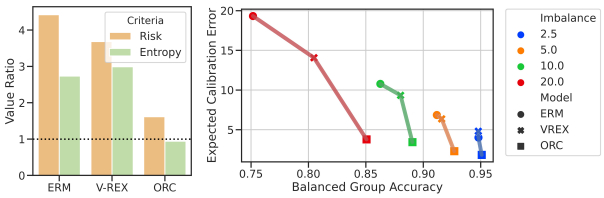


Figure 1: (Left) The value ratios of each model, trained on the Colored-MNIST with imbalance = 10. Imbalance means $\frac{\text{\# of whole samples}}{\text{\# of samples without spurious correlation}}$. The value ratio means the ratio of averaged empirical values between samples with and without spurious correlation. For example, risk ratio means $\frac{\text{Averaged risk of samples without spurious correlation}}{\text{Averaged risk of samples with spurious correlation}}$. (Right) The test performances of each model on the Colored-MNIST with different imbalances. As the imbalance increases, Our model, ORC, beats others with a larger margin.

To train a robust model from spurious correlations, the robust optimization (Namkoong & Duchi, 2016; Oren et al., 2019; Sagawa et al., 2019; Arjovsky et al., 2019) has become a major tool. Robust optimization usually focuses on optimizing the worst-case risks over the pre-defined groups of the training dataset. For instance, GroupDRO minimizes the worst-case risks from pre-defined groups of the dataset (Sagawa et al., 2019). V-REX minimizes the variance of risks (Krueger et al., 2020) to limit the risks of whole groups to a similar level. Although these approaches contributed on improving the worst-group performances, Figure 1 shows that they still show performance degradation on a dataset with extremely spurious correlation.

Figure 1 shows that there is a significant difference in risk and entropy between data samples with and without spurious correlation. This implication means that majority samples are learned in priority with lower risks and higher confidence than minority samples, where our model keeps the ratio relatively closer to one.

These findings motivate us to explicitly model the Orderliness on training risk and predictive confidence of each group in our objective. Our paper introduces an algorithm,

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the SCIS workshop.

Ordered Risk and Confidence regularization (ORC), that relatively regularizes the risks and the predictive confidences of the groups to follow the specific order, which is sorted according to the degree of spurious correlation. This regularization leads to the relatively weaker inference of data samples with spurious correlation than the samples without spuriousness. By reflecting the prior knowledge of each group on the ordering process, ORC provides a general way to inject the inductive biases or knowledge on the optimization procedure. Our experiments demonstrate that our method improves both the model robustness and calibration against the various types of *spurious correlation* in both synthetic and real-world datasets.

2. Preliminary

2.1. Problem Setup

Consider a classification task, when the input-label pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is from training dataset $\mathcal{D}$. Given n training samples, we train a classifier $f_\theta : \mathcal{X} \to \mathcal{Y}$, which is parameterized by $\theta \in \Theta$.

Spurious Correlation from Groups This paper hypothesizes that $\mathcal{D}$ consists of data instances belonging to groups $g \in \mathcal{G}$. Also, we assume that the groups have different levels of biases correlated to a input feature of x and the label of y, which Sagawa et al. (2020) name as *spurious correlation*. Following the convention of previous researches (Sagawa et al., 2019; Liu et al., 2021), We set each group $g \in \mathcal{G}$ to be defined by the combination of the label Y and a corresponding bias attribute $\mathcal{A}$. Hence, g follows $\mathcal{G} = \mathcal{A} \times \mathcal{Y}$, and we say that there exists *spurious correlation* between $\mathcal{A}$ and $\mathcal{Y}$ (Sagawa et al., 2019; Liu et al., 2021; Nam et al., 2020).

Learning toward spurious correlation would induce the sacrifice of the performance in the minority group.[1] Therefore, we may limit learning from the spurious correlation to value the good performance in the minority group. We formulate the *group robustness* (Sagawa et al., 2019; Liu et al., 2021), which values a model to obtain balanced and good performances measured by groups $g \in \mathcal{G}$.

Group Robustness Evaluation of the group robustness is usually conducted via balanced group error, which computes the averaged error across groups as follows:

$$\frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \mathbb{E}[\ell_{0-1}(x, y; \theta)|g]. \tag{1}$$

Here, $\ell_{0-1}(x, y; \theta) = 1[f_\theta(x) \neq y]$ is the 0-1 loss. Alternatively, one can check the worst group error (Sagawa et al., 2019; Liu et al., 2021) as follows:

$$\max_{g \in \mathcal{G}} \mathbb{E}[\ell_{0-1}(x, y; \theta)|g]. \tag{2}$$

[1]We provide the example case of spurious correlation in Appendix A.

Our objective includes not only robustness but also model calibration (Guo et al., 2017), please refer to the Appendix B. for a detailed explanation.

2.2. Previous researches on Group Robustness

Robust optimization (Namkoong & Duchi, 2016; Oren et al., 2019; Sagawa et al., 2019) is a method to improve the group robustness by directly optimizing a model's worst-case risk over a perturbed dataset. GroupDRO (Sagawa et al., 2019) minimizes the worst-case risk over the pre-defined groups in the training dataset. Krueger et al. (2020) introduces Risk Extrapolation (REX), which provides a robust optimization for the *affine* combinations of the perturbed risks. They also propose V-REX, which is a more stable variant of REX that reduces the risk differences among $|\mathcal{G}| = m$ groups as follows:

$$\min_\theta \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, y_i; \theta) + \lambda \mathrm{Var}(\mathcal{R}_1(\theta), ..., \mathcal{R}_m(\theta)). \tag{3}$$

This objective leads to the variance minimization, which is equivalent to the average pairwise mean squared error between group risks.

Based on the formulation, can V-REX empirically achieve the robustness on the groups with heterogeneous characteristics, i.e. different sample size for each group? As shown in figure 1, the risk $\mathcal{R}_g(\theta)$ computed from V-REX diverges based on the presence or absence of spurious correlation on each group g. One of the counters for this problem is reflecting the prior knowledge on each group to determine the degree of regularization for the design of objective function. However, the present form of V-REX motivates us to develop a generalized regularization to utilize the prior information of each group.

3. ORC : Ordered Risk and Confidence Regularization

We propose an algorithm, Ordered Risk and Confidence regularization (ORC), to improve the group robustness and calibration of a classifier. We hypothesize that the group-based robustness and calibration can be improved by regularizing the risk and confidence of each group to follow our intended order. The intended order follows the intensity of the spurious correlation of each group, so the most bias-aligned[2] group in $\mathcal{G}$ is regularized to have a relatively higher training risk and lower predictive confidence than other groups, which leads to the relatively weaker inference than bias-conflicting groups.

Group Order Setup for ORC ORC needs the sorting of

[2]We designate groups biased toward spurious correlation as bias-aligned, otherwise bias-conflicting.

$g \in \mathcal{G} = \mathcal{A} \times \mathcal{Y}$ by the intensity of spurious correlation. Assuming that group annotations are available, we can utilize human annotations or expertised domain knowledge to specify the highly biased groups. When they are not available, we can also utilize the group-wise statistics to check the bias amplification of each group, following Wang & Russakovsky (2021); Zhao et al. (2017). We also empirically found out that it is sufficient to divide whole groups into two sets: a set of groups with spurious correlation and a set of groups without spurious correlation, without the need of sorting all groups to have relative order. However, we generally provide our formulation assuming m groups.

3.1. Objective Formulation of ORC

Given the access to the set of pre-defined groups $\mathcal{G}$ with $|\mathcal{G}| = m$, we sort the group index in the descending order by the intensity of the biases. Afterwards, we formulate our objective as a constrained problem for n samples in training dataset D as follows:

$$\begin{aligned} \underset{\theta}{\text{minimize}} \quad & \frac{1}{n}\sum_{i=1}^{n} \ell(x_i, y_i; \theta) \\ \text{subject to} \quad & \mathcal{H}_1(\theta) \geq \mathcal{H}_2(\theta) \geq ... \geq \mathcal{H}_m(\theta), \\ & \mathcal{R}_1(\theta) \geq \mathcal{R}_2(\theta) \geq ... \geq \mathcal{R}_m(\theta) \end{aligned} \tag{4}$$

where $\mathcal{H}_g(\theta) = \frac{1}{n_g}\sum_{x_i \in g} -\sum_k f_\theta(x_i)_k \log f_\theta(x_i)_k$ is averaged entropy of model predictions of data samples in group g. $f_\theta(x)$ represents the prediction probability and $\mathcal{R}_g(\theta)$ is the empirical risk for group g. The entropy is utilized as a measure of predictive confidence (Vyas et al., 2018) and the cross entropy[3] is used as our loss function ℓ.

3.2. Implementable Objective of ORC

Confidence Ordering In classification tasks, the label smoothing has been widely utilized for modeling the intended degree of confidence (Szegedy et al., 2016; Müller et al., 2019; Lukasik et al., 2020) for each data sample. It also has an advantage in that confidence level can be calibrated during the training procedure, which enables the joint learning with other objectives. From this spirit, we propose a group-aware label smoothing, which utilizes a smoothed label y^g for group g, instead of utilizing one-hot label y for calculating the cross-entropy loss. y^g is defined as $y^g_k = y_k(1 - \frac{K}{K-1}\alpha_g) + \frac{\alpha_g}{K-1}$, where K is the number of classes and α_g is the smoothing factor for each group g.

Here, our point is differentiating the smoothed label y^g for each group g. Let the group-wise smoothing factor α_g to satisfy $\alpha_1 \geq \alpha_2 \geq ... \geq \alpha_m$. Then, the confidence inequality of each group can be regularized by the intended order of ORC, see the proof in Appendix C. It also relieves the over-confidence of bias-aligned groups by inducing larger α_g than the other groups.

Risk Ordering From the group-aware smoothed label, y^g, we additionally define $\widetilde{\mathcal{R}}_g(\theta) = \frac{1}{n_g}\sum_{(x_i,y_i)\in g} -y^g_k \log f_\theta(x)_k$, which is the subpart of the cross entropy only utilizing the true class part of y^g and $f_\theta(x)$. We derive that the risk ordering can be approximated by regularizing the equality between $\{\widetilde{\mathcal{R}}_g(\theta)\}_{g=1}^m$. Let's assume a case when the equality between $\{\widetilde{\mathcal{R}}_g(\theta)\}_{g=1}^m$ is satisfied. Assuming cross-entropy as a loss function, $\widetilde{\mathcal{R}}_g(\theta)$ can be transformed as $(1-\alpha_g)\mathcal{R}_g(\theta)$, see Appendix C for the derivation. By then, we can re-formulate the equality between $\{\widetilde{\mathcal{R}}_g(\theta)\}_{g=1}^m$ as follows:

$$(1-\alpha_1)\mathcal{R}_1(\theta) = ... = (1-\alpha_m)\mathcal{R}_m(\theta). \tag{5}$$

As α_g is pre-defined to satisfy $\alpha_1 \geq \alpha_2 \geq ... \geq \alpha_m$, it leads to the inequality of true training risks as $\mathcal{R}_1(\theta) \geq \mathcal{R}_2(\theta) \geq ... \geq \mathcal{R}_m(\theta)$. Similar to V-REX, we choose to regularize the equality between $\{\widetilde{\mathcal{R}}_g(\theta)\}_{g=1}^m$ by minimizing $\mathrm{Var}(\widetilde{\mathcal{R}}_1(\theta), \widetilde{\mathcal{R}}_2(\theta), ..., \widetilde{\mathcal{R}}_m(\theta))$. Afterward, the final objective is provided as below:

$$\min_{\theta} \frac{1}{n}\sum_{i=1}^{n} \ell(x_i, y^g_i; \theta) + \lambda \mathrm{Var}(\widetilde{\mathcal{R}}_1(\theta), ..., \widetilde{\mathcal{R}}_m(\theta)). \tag{6}$$

This optimization only requires the loss computation from the smoothed label, without any need of further computation for two different constraints. Thus pre-defined smoothing factors $\{\alpha_g\}_{g=1}^m$ decide the margins on the risk and confidence order. When we divide whole dataset into two sets, the set with spurious correlation and without it, smoothing factors $\{\alpha_g\}_{g=1}^m$ collapses into two. When $\{\alpha_g\}_{g=1}^m$ are all set to 0, our objective is equal to V-REX, which implies that ORC is a generalized version of V-REX with injection of prior knowledge. We also present ORC*, which is variants of ORC for group-unknown setting in Appendix D with corresponding experimental results.

4. Experiments

This section demonstrates the effectiveness of ORC by comparing the performances with other baselines based on various datasets with spurious correlation. We also provide the analyses with ablation on each module of our algorithm. We provide the details of evaluation metrics on Appendix E.

4.1. Baselines

We utilize the various baselines to verify the performance of ORC. See Appendix F for the details of each method.

4.2. Experiments on Colored MNIST (Nam et al., 2020)

For evaluation on Colored MNIST, we utilize ResNet-18 (He et al., 2015) as a backbone model for all approaches.

[3] $CE(y, f_\theta(x)) = \sum_{k=1}^{K} -y_k \log f_\theta(x)_k$

Model	90% of bias-aligned samples			95% of bias-aligned samples			98% of bias-aligned samples		
	Unb Acc ↑	Worst Acc ↑	G-ECE ↓	Unb Acc ↑	Worst Acc ↑	G-ECE ↓	Unb Acc ↑	Worst Acc ↑	G-ECE ↓
ERM	86.79±0.76	85.37±0.85	10.74±0.62	78.13±1.07	75.76±1.19	18.74±1.40	64.40±1.39	60.50±1.55	30.13±1.22
GroupDRO (Sagawa et al., 2019)	86.61±0.22	85.2±0.23	11.01±0.14	78.38±0.60	76.05±0.65	17.50±0.41	63.15±0.76	59.11±0.82	31.23±1.29
V-REX (Krueger et al., 2020)	88.01±0.75	86.78±0.88	9.76±0.38	79.99±0.18	77.82±0.18	16.07±0.28	69.27±0.25	65.94±0.29	20.88±3.65
LfF (Nam et al., 2020)	74.79±7.89	73.98±8.32	19.57±5.78	80.25±3.78	79.93±3.91	14.90±2.35	**78.28**±0.15	**77.16**±2.81	16.66±0.58
SD (Pezeshki et al., 2020)	88.55±0.05	87.35±0.06	5.44±0.19	80.32±0.44	78.19±0.51	**3.26**±0.45	61.83±0.32	57.64±0.37	13.37±0.21
JTT (Liu et al., 2021)	87.04±0.81	85.69±0.90	10.73±0.80	80.27±1.53	78.14±1.70	16.22±1.10	63.90±1.19	59.98±1.35	29.67±2.09
ORC	**88.55**±0.20	**87.36**±0.23	**3.73**±0.22	**83.17**±0.40	**81.35**±0.43	3.98±0.14	76.46±0.38	73.93±0.41	**4.12**±0.28

Table 1: The unbiased group accuracy (Unb Acc), worst group accuracy (Worst Acc), and group calibration error (G-ECE) on Colored-MNIST. Best performing results are marked in bold. We underline results only when ORC is second best.

We also utilize the group-based smoothing factors $\alpha = 0.1$ for bias-aligned groups set and $\alpha = 0$ for bias-conflicting groups set, respectively. In Table 1, ORC shows competitive performances over other baselines for both group robustness and calibration.

4.3. Ablation Study

Model	Unb Acc ↑	Worst Acc ↑	G-ECE ↓
V-REX w/ Upsample 2	82.10±0.98	80.19±1.09	14.70±0.67
V-REX w/ Upsample 8	80.58±1.01	78.49±1.13	16.10±1.19
V-REX w/ Risk Ordering	81.74±0.89	79.79±0.99	14.81±0.78
V-REX w/ Confidence Ordering	82.03±0.27	80.08±0.3	4.96±0.45
ORC	**83.17**±0.40	**81.35**±0.43	**3.98**±0.14

Table 2: The ablation study of ORC

For ablation study, we compared the performances between the variants of V-REX and ORC on the Colored-MNIST. **V-REX w/ Upsample *m*** denote the model, which upsamples bias-conflicting m times more than the other samples. We additionally compare with models which only utilize either **Risk Ordering** or **Confidence Ordering**.

Table 2 shows that the augmentation of each module shows the consistent improvements from V-REX, which proves the efficacy of joint modeling with ordered risk and confidence.

Model	Accuracy			G-ECE ↓
	Head ↑	Tail ↑	Unb ↑	
GroupDRO (Sagawa et al., 2019)	48.76	42.14	45.45	34.2
IRMv1 (Arjovsky et al., 2019)	49.91	42.52	46.21	29.29
V-REX (Krueger et al., 2020)	50.84	42.71	46.77	29.16
Rubi (Cadene et al., 2019)	48.18	**44.03**	46.1	29.46
UpWt (Sagawa et al., 2020)	47.26	39.04	43.15	26.59
lff (Nam et al., 2020)	48.47	38.1	43.28	32.74
SD (Pezeshki et al., 2020)	49.51	43.74	46.62	26.71
ORC	**51.07**	**44.03**	**47.55**	**20.8**

Table 3: The performances evaluated on the GQA-OOD.

4.4. Experiments on GQA-OOD (Kervadec et al., 2021)

GQA-OOD is a dataset for visual question answering task, which is a protocol for validating models with biased settings. We divide the dataset into two disjoint groups based on the frequency of answers: `head` as a majority and `tail` as a minority, where `Unb` is an unbiased version between them. In Table 3, ORC shows the best performances across all metrics. Please see Appendix G for experimental results on other datasets.

4.5. Why is ORC Effective?

We introduce two analyses to investigate the reason for the effectiveness of ORC. In Figure 2, ORC improves the performances of both groups without any performance degradations. This improvement comes from the relative regularization, rather than focusing on the inference of only one group, i.e. LfF. Also, Figure 3 shows that the larger the smoothing factor α for the bias-aligned group and the smaller α for the bias-conflicting, ORC shows better calibration performances. This implies that the different smoothing can be effectively utilized based on the bias-level of each group.

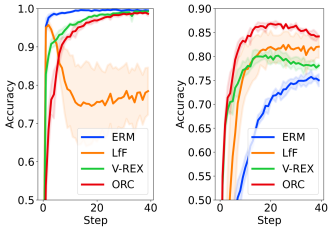


Figure 2: The accuracy dynamics for bias-aligned (left) and bias-conflicting samples (right).

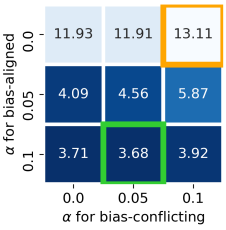


Figure 3: Ablation of G-ECE for smoothing factors.

5. Conclusion

We introduce a new mechanism, ORC, which relatively regularizes the risks and the confidences of the groups in the training dataset. By reflecting the prior knowledge of each group (e.g. the intensity of biases) on the ordering process, ORC allows us to provide the group-based knowledge on the degree of regularization. By investigating the relative difference of risks and confidences between groups, we empirically show the effectiveness of our group-heterogeneous approach with ordered regularization from the various types of *spurious correlation* in both synthetic and real-world datasets.

References

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6077–6086, 2018. doi: 10.1109/CVPR.2018.00636.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Borkan, D., Dixon, L., Sorensen, J., Thain, N., and Vasserman, L. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, 2019.

Cadene, R., Dancette, C., Ben younes, H., Cord, M., and Parikh, D. Rubi: Reducing unimodal biases for visual question answering. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/51d92be1c60d1db1d2e5e7a07da55b26-Paper.pdf.

Cao, K., Wei, C., Gaidon, A., Arechiga, N., and Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, 2019.

Cosmides, L. and Tooby, J. Are humans good intuitive statisticians after all? rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, 58(1):1–73, 1996. ISSN 0010-0277. doi: https://doi.org/10.1016/0010-0277(95)00664-8. URL https://www.sciencedirect.com/science/article/pii/0010027795006648.

Deng, L. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L. Measuring and mitigating unintended bias in text classification. 2018.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 1321–1330. JMLR.org, 2017.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., and Smith, N. A. Annotation artifacts in natural language inference data. In *NAACL*, 2018.

Hashimoto, T. B., Srivastava, M., Namkoong, H., and Liang, P. Fairness without demographics in repeated loss minimization. In Dy, J. G. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1934–1943. PMLR, 2018. URL http://proceedings.mlr.press/v80/hashimoto18a.html.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

Hudson, D. A. and Manning, C. D. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Kervadec, C., Antipov, G., Baccouche, M., and Wolf, C. Roses are red, violets are blue... but should vqa expect them to?, 2021.

Kim, B., Kim, H., Kim, K., Kim, S., and Kim, J. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9012–9020, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Krueger, D., Caballero, E., Jacobsen, J., Zhang, A., Binas, J., Priol, R. L., and Courville, A. C. Out-of-distribution generalization via risk extrapolation (rex). *CoRR*, abs/2003.00688, 2020.

Li, Y. and Vasconcelos, N. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9572–9581, 2019.

Liu, E. Z., Haghgoo, B., Chen, A. S., Raghunathan, A., Koh, P. W., Sagawa, S., Liang, P., and Finn, C. Just train twice: Improving group robustness without training group information. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6781–6792. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/liu21f.html.

Liu, Z., Luo, P., Wang, X., and Tang, X. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15 (2018):11, 2018.

Lukasik, M., Bhojanapalli, S., Menon, A., and Kumar, S. Does label smoothing mitigate label noise? In *International Conference on Machine Learning*, pp. 6448–6458. PMLR, 2020.

Marcel, S. and Rodriguez, Y. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, pp. 1485–1488, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605589336. doi: 10.1145/1873951.1874254. URL https://doi.org/10.1145/1873951.1874254.

McCoy, R. T., Pavlick, E., and Linzen, T. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *ACL*, 2019.

Müller, R., Kornblith, S., and Hinton, G. When does label smoothing help? *arXiv preprint arXiv:1906.02629*, 2019.

Nam, J. H., Cha, H., Ahn, S., Lee, J., and Shin, J. Learning from failure: De-biasing classifier from biased classifier. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/eddc3427c5d77843c2253f1e799fe933-Abstract.html.

Namkoong, H. and Duchi, J. C. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *NIPS*, volume 29, pp. 2208–2216, 2016.

Niculescu-Mizil, A. and Caruana, R. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, pp. 625–632, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595931805. doi: 10.1145/1102351.1102430. URL https://doi.org/10.1145/1102351.1102430.

Oren, Y., Sagawa, S., Hashimoto, T. B., and Liang, P. Distributionally robust language modeling. *arXiv preprint arXiv:1909.02060*, 2019.

Park, J. H., Shin, J., and Fung, P. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2799–2804, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1302. URL https://aclanthology.org/D18-1302.

Pezeshki, M., Kaba, S.-O., Bengio, Y., Courville, A., Precup, D., and Lajoie, G. Gradient starvation: A learning proclivity in neural networks. *arXiv preprint arXiv:2011.09468*, 2020.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2019.

Sagawa, S., Raghunathan, A., Koh, P. W., and Liang, P. An investigation of why overparameterization exacerbates spurious correlations. In *ICML*, pp. 8346–8356, 2020. URL http://proceedings.mlr.press/v119/sagawa20a.html.

Shrestha, R., Kafle, K., and Kanan, C. An investigation of critical issues in bias mitigation techniques, 2021a.

Shrestha, R., Kafle, K., and Kanan, C. An investigation of critical issues in bias mitigation techniques. *CoRR*, abs/2104.00170, 2021b.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Vyas, A., Jammalamadaka, N., Zhu, X., Das, D., Kaul, B., and Willke, T. L. Out-of-distribution detection using an ensemble of self supervised leave-out classifiers. In *ECCV*, 2018.

Wang, A. and Russakovsky, O. Directional bias amplification. In *ICML*, 2021.

Zhang, Z. and Sabuncu, M. R. Generalized cross entropy loss for training deep neural networks with noisy labels. *arXiv preprint arXiv:1805.07836*, 2018.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2979–2989, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1323. URL https://aclanthology.org/D17-1323.

A. Examples of Spurious Correlation

Figure 4 illustrates the *spurious correlation* in the CelebA dataset (Liu et al., 2018). When $\mathcal{G}$ is defined by the combination of the target label `Makeup` and the attribute `Gender`, there is a minority group of males with `Makeup`, and this group results in a significantly lower performance than the majority group of males without `Makeup`.

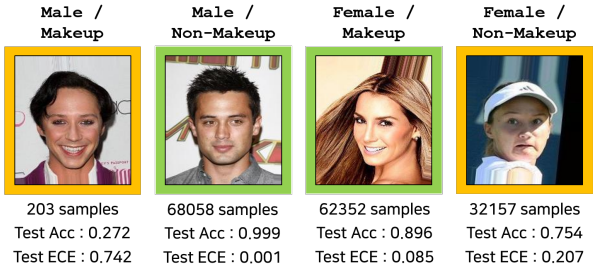


Figure 4: The groups defined by the combination of `Gender` and `Makeup`. A majority of training samples, which align with the spurious correlation, belong to the groups highlighted in green. Minor samples, which conflict with the correlation, belong to the groups highlighted in yellow. We report the accuracy and expected calibration error based on the ERM model.

B. Group Calibration

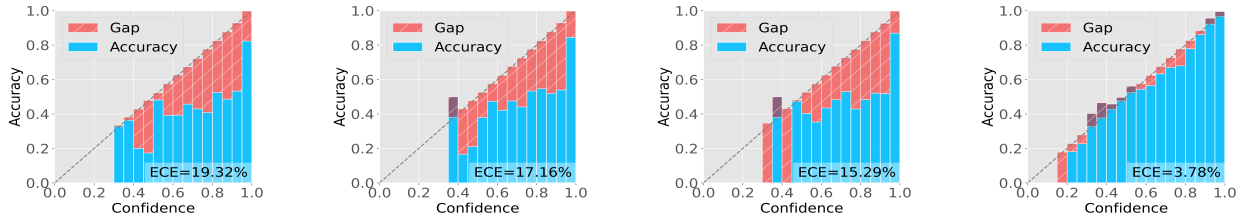
Confidence calibration (Cosmides & Tooby, 1996; Niculescu-Mizil & Caruana, 2005) is estimating the classification probability to match the true correctness likelihood. Let $\hat{p}$ be the confidence estimate, which is probability of the class with the highest predicted value. We would like the confidence estimate $\hat{p}$ to represent a true likelihood of correctness for the good calibration. This leads to the evaluation of calibration via Expected Calibration Error (ECE) (Guo et al., 2017). By grouping n predictions into B interval bins with equal size, ECE is defined as follows:

$$\text{ECE} = \sum_{b=1}^{B} \frac{|\mathcal{S}_b|}{n} |\text{acc}(\mathcal{S}_b) - \text{conf}(\mathcal{S}_b)| \tag{7}$$

where $\mathcal{S}_b$ denotes the set of samples whose prediction output scores fall into Bin-b; and $\text{acc}(\mathcal{S}_b)$ and $\text{conf}(\mathcal{S}_b)$ are the averaged accuracy and predicted confidence of samples in $\mathcal{S}_b$, respectively. In our paper, we additionally introduce Group-ECE, which is the averaged ECE from each group. Group-ECE becomes evaluation metric for our *group calibration*, which aims at obtaining low calibration error from each group, without disparities of ECE in Figure 4.

$$\text{Group-ECE} = \frac{1}{\mathcal{G}} \sum_{g \in \mathcal{G}} \sum_{b=1}^{B} \frac{|\mathcal{S}_b^g|}{n_g} |\text{acc}(\mathcal{S}_b^g) - \text{conf}(\mathcal{S}_b^g)| \tag{8}$$

n_g is # of samples in group g. Group-ECE is equal to ECE, when all groups are with the same size. From the defined setup, our objective is to learn the model in the direction of achieving the group robustness and calibration when the training dataset $\mathcal{D}$ is provided with the groups $\mathcal{G}$ with spurious correlation.



((a)) ERM ((b)) GroupDRO ((c)) V-REX ((d)) ORC

Figure 5: Reliability diagrams of baseline models and ORC trained on Colored-MNIST dataset. In this figure, we report the Expected Calibration Error (ECE) of the unbiased test dataset as a measurement of confidence calibration performances.

C. Proof

C.1. Proof of $\widetilde{\mathcal{R}}_g(\theta) = (1-\alpha_g)\mathcal{R}_g(\theta)$

Here, $\widetilde{\mathcal{R}}_g(\theta)$ as $\widetilde{\mathcal{R}}_g(\theta) = \frac{1}{n_g}\sum_{(x_i,y_i\in g)} -y_k^g \log f_\theta(x)_k$, which is the subpart of the cross entropy only utilizing the true class k of y^g and $f_\theta(x)$. We recall that the smoothed label y^g is defined as $y_k^g = y_k(1-\frac{K}{K-1}\alpha_g) + \frac{\alpha_g}{K-1}$, when α_g is the smoothing factor for the samples with group g. Then, we can derive $\widetilde{\mathcal{R}}_g(\theta)$ as follows:

$$
\begin{aligned}
\widetilde{\mathcal{R}}_g(\theta) &= \frac{1}{n_g}\sum_{(x_i,y_i\in g)} -y_k^g \log f_\theta(x)_k \\
&= \frac{1}{n_g}\sum_{(x_i,y_i\in g)} -(1-\alpha_g) y_k \log f_\theta(x)_k \\
&= (1-\alpha_g)\frac{1}{n_g}\sum_{(x_i,y_i\in g)} -y_k \log f_\theta(x)_k \\
&= (1-\alpha_g)\frac{1}{n_g}\sum_{(x_i,y_i\in g)}\sum_{k'}^{K} -y_{k'} \log f_\theta(x)_{k'} \\
&= (1-\alpha_g)\mathcal{R}_g(\theta)
\end{aligned}
$$

C.2. Proof of Confidence Ordering

For the intended ordering of predictive confidences for each group, we utilize the group-aware smoothed label y_g, which satisfies $y_k^g = y_k(1-\frac{K}{K-1}\alpha_g) + \frac{\alpha_g}{K-1}$, for each group g. Let the group-wise smoothing factor α_g to satisfy $\alpha_1 \geq \alpha_2 \geq ... \geq \alpha_m$, when K is # of classes. Then, the target class label probability of each group, y_k^g, has the following orders as

$$
y_k^1 = (1-\alpha_1) \leq y_k^2 = (1-\alpha_2) \leq ... \leq y_k^m = (1-\alpha_m)
$$

From the defined smoothed label, we are minimizing following objective function:

$$
\begin{aligned}
&\min_\theta \frac{1}{n}\sum_{i=1}^{n} \ell(x_i, y_i^g; \theta) \\
=&\min_\theta \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} -y_k^g \log f_\theta(x)_k
\end{aligned}
$$

As mentioned on the main paper, we utilize cross entropy as a loss function for classification task. Above objective function get minimized when $f_\theta(x_i) = y_i^g$ for all x_i, y_i. As a measurement of confidence, we utilize the maximum class probability of model prediction, $f_\theta(x)_k$, where we abusely denote k as true class index. As $f_\theta(x)_k$ get larger, it implies that the model is confident of its own prediction. Then at that minimization point, we can satisfy the confidence inequality as follows:

$$
\frac{1}{n_1}\sum_{x_i\in g_1} f_\theta(x_i)_k \leq \frac{1}{n_2}\sum_{x_i\in g_2} f_\theta(x_i)_k ... \leq \frac{1}{n_m}\sum_{x_i\in g_m} f_\theta(x_i)_k
$$

By injecting the group-aware smoothing factors, we empirically show that the averaged confidence of each group follows the intended order, which is sorted by the intensity of spurious correlation.

D. ORC* for the Group-unknown Setting

This section provides ORC*, which is variants of ORC for the case when we do not have knowledge to G. First, we utilize the observation of (Nam et al., 2020; Liu et al., 2021) that ERM model f_{erm} tends to fit bias-aligned groups, but not bias-conflicting groups at the early stage of the training. This leads to our assumption that the higher loss in the early

learning indicates the higher percentage of being bias-conflicting to the unknown spurious correlation. Then, we can divide the training dataset into arbitrary groups based on the distribution of training risks from the ERM model.

Group Partitioning Following the concepts of Liu et al. (2021), we first train the ERM model f_{erm} a few epochs before the main training, and we divide the groups depending on whether the f_{erm} correctly classifies the samples or not. Here, the samples, which were incorrectly classified by ERM model, would show larger loss compared with the correct ones. Dividing the training dataset based on the classification correctness results in two distinct groups; error set $E = \{(x_i, y_i) \text{ s.t. } f_{erm}(x_i) \neq y_i\}$ which is composed of the samples that f_{erm} misclassify, and correct set $E^c = \{(x_i, y_i) \text{ s.t. } f_{erm}(x_i) = y_i\}$.

Toward input-aware label smoothing From the learning to divide the group, f_{erm} learns the difficulty for each sample, which is valuable prior information for estimating the bias intensity of each input. Given the partitioned groups E and E^c, ORC* utilizes the training risks of each sample from f_{erm} to develop the input-aware smoothing factors $\{\alpha_i\}_{i=1}^n$ to reflect the input-dependent characteristics in modeling. By utilizing the input-dependent smoothing factors with the estimated groups, it compensates the possible problems caused by group partitioning errors.

After t epochs of training on f_{erm}, we can extract $\{\ell(x_i, y_i; \theta_t)\}_{i=1}^n$, which is a set of training losses for each sample, where θ_t is a inferred parameter of f_{erm} after epoch t. As our intention is to induce larger smoothing factors to the samples with spurious correlation, we calculate the input-aware smoothing factor α_i for input x_i as follows:

$$\alpha_i = \max(\alpha) - \max(\alpha)\hat{\ell}(x_i, y_i; \theta_t) \; \forall \, i \tag{9}$$

where $\max(\alpha)$ is a the largest possible value of smoothing factor and $\hat{\ell}(x_i, y_i; \theta_t)$ is normalized value of $\ell(x_i, y_i; \theta_t)$ by applying max-min normalization to $\{\ell(x_i, y_i; \theta_t)\}_{i=1}^n$, which results in the truncation of $\{\hat{\ell}(x_i, y_i; \theta_t)\}_{i=1}^n \in [0, 1]$. By then, α_i is constructed negatively proportional to the normalized training risk $\hat{\ell}(x_i, y_i; \theta_t)$, which results in inducing the higher smoothing factors for the samples with lower risk. Then, we provide the final objective of ORC* as below:

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, \widetilde{y}_i; \theta) + \lambda \mathrm{Var}(\widetilde{R}(\theta_t), \widetilde{R}^c(\theta_t)). \tag{10}$$

Here, $\widetilde{y}_i$ is smoothed label for x_i satisfying $\widetilde{y}_{i,k} = y_k(1 - \frac{K}{K-1}\alpha_i) + \frac{\alpha_i}{K-1}$. $\widetilde{R}(\theta_t) = \frac{1}{|E|}\sum_{(x,y)\in E} -\widetilde{y}_k \log f_\theta(x)_k$ and $\widetilde{R}^c(\theta_t) = \frac{1}{|E^c|}\sum_{(x,y)\in E^c} -\widetilde{y}_k \log f_\theta(x)_k$.

E. Evaluation Metric

We evaluate the group robustness and calibration based on metrics; 1) unbiased test accuracy, which is equal to the balanced group accuracy reversed from Eq 1, 2) worst group accuracy, whose error corresponds to Eq 2, and 3) group expected calibration error (Group-ECE) in Eq 8.

F. Baseline

We provide the detailed explanations of baselines, which were utilized for our experiments.

GroupDRO (Sagawa et al., 2019) is distributionally robust optimization method for worst-case generalization. Overparameterized neural networks perform well on an average dataset by learning spurious correlation, but they fail on rare and heavy-tailed examples. To increase the performance of rare and atypical examples, GroupDRO focuses on minimizing the training losses of worst-case group among pre-defined groups. However, (Sagawa et al., 2019) shows that naive application of GroupDRO still fails on an atypical dataset, and they propose a coupling of GroupDRO and strong regularization such as L_2 norm penalty and early stopping. With strong regularization, GroupDRO shows performance improvement by avoiding to learn the spurious correlation, which is also called as *bias*.

IRMv1 (Arjovsky et al., 2019) is designed for learning invariant representation across different multiple training environments. Traditional ERM assumes that the training samples are drawn identically distributed, and ERM shows lower performance on the samples with spurious correlation, and it absorbs the spurious correlation. However, IRM treats the dataset as the results of multiple environments and learns the correlation to be invariant. The objective of IRM is to learn a feature extractor such that the optimal linear classifier of every environment is the same. Investigating an optimal feature extractor and linear classifier requires a bi-level optimization, highly non-convex problem. To approximate the optimization

problem, they propose an approximated version of IRM, which is called IRMv1, which penalizes the squared norm of gradient.

REX (Krueger et al., 2020) Under the assumption that the distributional shift at the test time will be much extreme, Krueger et al. (2020) introduces Risk Extrapolation (REX), which provides a robust optimization for the perturbed set of the *affine* combinations from the training risks. The degree of extrapolation can be controlled by the hyperparameter. In addition, Krueger et al. (2020) showed that this extrapolation of robust optimization aligns with the objective of reducing the risk differences among different training domains. Accordingly, Krueger et al. (2020) also proposes V-REX, which is more stable variant of REX that reduces the risk differences among m training domains.

RUBi (Cadene et al., 2019) is a debiasing algorithm specially designed for visual question answering tasks. It multiplies the output from the main model with sigmoided outputs from the additional biased model, thereby assigning higher loss weights to samples that cannot be predicted through biases alone.

UpWt (Sagawa et al., 2020) is a algorithm which upweights the samples with minority patterns. In other words, it attempts to mitigate the spurious correlation by emphasizing the importance of samples from minority groups. Sagawa et al. (2020) accentuate that UpWt needs sufficient regularizations for validity of the models. Examples of these regularization is training models with low learning rates or high weight decays.

LfF (Learning from Failure) (Nam et al., 2020) proposes a debiasing scheme by training a debiased classifier based on the failure of a biased classifier. Lff simultaneously train two neural networks, i.e. one for the biased classifier and the other for the debiased classifier. The biased network utilizes the generalized cross entropy loss (Zhang & Sabuncu, 2018) to focus on the easy samples, which is expected to be aligned with bias information. Simultaneously, the debiased network can focus on the samples, which are hard for the biased network to learn. The samples are expected to be conflicting with the bias information. Lff "reweights" training samples using the score based on the loss of each classifier, which represents the relative difficulty score.

SD (Spectral Decoupling) (Pezeshki et al., 2020) shows that *Gradient Starvation*, gradient descent updates parameters in the direction of dominant features but not potentially informative features, happens when cross-entropy loss is minimized. When strongly-correlated and easy-to-learn features exist in training dataset, gradient descent is biased towards them first. Pezeshki et al. (2020) linearize and approximate the learning dynamics of neural networks to a dual space using Neural Tangent Kernel (NTK), and show that the fast learning on dominant features has a detrimental effect on the learning of other features which are coupled with dominant features. In this context, *Spectral Decoupling* (SD) is proposed as a possible remedy of Gradient Starvation. SD shows that simply replacing the weight decay term in the loss function, $\mathcal{L} = \log\left(1 + e^{-y\hat{y}}\right) + ||\theta||_2^2$, with an L2 penalty on the network's logit, $\mathcal{L} = \log\left(1 + e^{-y\hat{y}}\right) + ||\hat{y}||_2^2$, probably decouples the fixed points of the dynamics.

G. Experiments on Real-World Datasets

This section provides experimental results for more diverse real-world datasets. Since ORC* was introduced in Appendix D, the experimental results of this section compare the experimental results by separating the models of group aware setting and group unknown setting.

G.1. CelebA (Liu et al., 2018)

Group setting	Model	Performance Metrics		
		Unb Acc ↑	Worst Acc ↑	G-ECE ↓
Known	GroupDRO	74.24±0.22	54.65±0.23	23.7±0.23
	V-REX	72.44±1.95	51.35±3.58	22.86±3.7
	ORC	**74.38**±0.98	**57.2**±2.61	**17.94**±2.77
Unknown	ERM	71.81±1.2	49.33±1.96	26.18±0.89
	LfF	**72.97**±3.11	**55.69**±5.57	24.68±2.54
	SD	70.33±1.62	47.69±2.8	26.35±0.98
	JTT	72.22±0.5	50.19±1.34	25.37±0.7
	ORC*	71.55±1.66	51.78±3.21	**22.29**±1.65

Table 4: The performances evaluated on the CelebA (averaged over three random seeds). The target and bias attribute are `Makeup` and `Male`.

CelebA (Liu et al., 2018) is a multi-attribute dataset for face recognition. Following Nam et al. (2020), we select `Makeup` as target label, Y, and `Male` as the bias attribute to investigate the spurious correlation between them. We utilize ResNet-50 as target model and also utilize the prior knowledge of each group with size statistics to find out bias-conflicting groups. Table 4 shows that ORC shows improvements of all performance metrics in the group-known setting. In group-unknown setting, ORC* shows the second best performances on the worst accuracy metric. Although LfF showed the best performance in terms of accuracy, it shows to be unstable with high standard deviation. In terms of G-ECE, ORC* surpasses other baselines.

Group setting	Model	Performance Metrics		
		Unb Acc ↑	Worst Acc ↑	G-ECE ↓
Known	GroupDRO	79.94	66.42	17.22
	V-REX	81.91	77.32	14.5
	ORC	**83.93**	**79.6**	**6.75**
Unknown	ERM	77.36	59.14	18.9
	LfF	75.98	64.01	11.45
	SD	80.18	64.92	21.97
	JTT	76.75	62.25	20.13
	ORC*	**82.11**	**71.08**	**10.45**

Table 5: The performances evaluated on the CivilComments.

G.2. CivilComments-WILD (Borkan et al., 2019)

Our task is to classify whether an online comment input is toxic or non-toxic. Prior works (Dixon et al., 2018; Park et al., 2018) have shown that toxicity of comments spuriously associate with the mention of certain demographics (e.g. male, female, black, LGBTQ, etc.). To enable the use of group based approaches, we construct the 4 groups (a, y), where the bias attribute $a \in \mathcal{A}$ is a binary indicator of whether any demographics are mentioned and the label $y \in \mathcal{Y}$ is toxicity (Liu et al., 2021). We select BERT (Devlin et al., 2019) with pretrained weights for the model architecture. Table 5 shows that ORC and ORC* overwhelms other baselines with performances on whole metrics in both group known and unknown setting, respectively.

H. Experimental Setup

We verify the effectiveness of our proposed algorithm on four synthetic and real-world benchmark datasets: Colored-MNIST (Nam et al., 2020), CelebA (Liu et al., 2018), CivilComments-WILD (Borkan et al., 2019), GQA-OOD (Kervadec et al., 2021). In this section, we provide the description of each dataset. Furthermore, we provide the detailed explanation of experimental settings for each dataset.

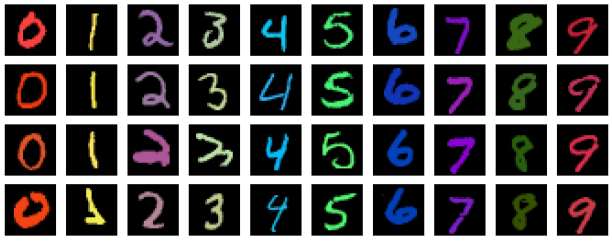


Figure 6: Examples of bias-aligned samples in Colored-MNIST

H.1. The Descriptions for the Dataset

H.1.1. COLORED-MNIST (NAM ET AL., 2020)

Dataset Description Colored-MNIST is dataset in which color is injected with random perturbation into the original MNIST dataset (Deng, 2012). The left side of Figure 6 shows the examples of bias-aligned samples in Colored-MNIST. The dataset contains images with two attributes: `Digit` and `Color`. A set of intended decision rules classify images correctly based on the `Digit`. Here, a decision rule based on bias attributes, e.g. `Color`, is considered as an unintended decision rule. We follow the provided dataset protocols from Nam et al. (2020), which inject colors to each image as follows:

- Choose ten distinct RGB values by drawing them uniformly at random. These RGB values is used throughout all the experiments for the Colored-MNIST dataset.
- Generate ten color distributions by assigning chosen RGB values to each color distribution.
  - Each color distribution is a 3-dimensional Gaussian distribution with pre-defined covariance $\sigma^2 I$, when $\sigma = 0.05$ for our experiments.
- Pair Digit a_t and Color distribution a_b to make a correlation between the two attributes.
  - Each bias-aligned sample has a Digit colored by RGB value sampled from paired Color distribution.
  - Each bias-conflicting sample has a Digit colored by RGB value sampled from other (nine) Color distributions.
  - We control the ratio of bias-aligned samples among (99.0%, 98.0%, 95.0%).

This modification results in the 60,000 training samples and 10,000 test samples. Similar modification has been proposed by Kim et al. (2019) and Li & Vasconcelos (2019).

Experimental Settings For the Colored-MNIST dataset, we use the MLP (Multi-Layered Perceptron) with three hidden layers where each hidden layer consists of 100 hidden units. We use Adam (Kingma & Ba, 2015) optimizer throughout all the experiments in the paper. We use a learning rate of 0.001, a batch size of 256, and epochs of 100 for the Colored-MNIST dataset. We do not use data augmentation schemes for training the neural network on the Colored-MNIST dataset.

H.1.2. CELEBA (LIU ET AL., 2018)

Dataset Description The CelebA dataset is a multi-attribute dataset for face recognition. It contains 40 attributes for each image and among those attributes, we consider `HeavyMakeup` as the target attributes. For both cases, we use `Male` as the bias attribute to investigate gender bias in CelebA. The dataset consists of 202,599 face images, and we use the official train-val split, which results in 162,770 samples for training, 19,867 samples for testing. To evaluate the unbiased accuracy with an imbalanced evaluation set, we evaluate accuracy via group-wise manner, and compute average accuracy over all groups.

Experimental Settings Following experimental settings from Nam et al. (2020), we utilize the Pytorch torchvision (Marcel & Rodriguez, 2010) implementation of the ResNet18 model with ImageNet pretrained weights. We use a learning rate of 0.0001, a batch size of 256, and 50 epochs for model training in CelebA.

H.1.3. CIVILCOMMENTS-WILD (BORKAN ET AL., 2019)

Dataset Description CivilComments is a real-world dataset with user generated text (e.g., detecting toxic comments). The task is a binary classification task of determining if a comment is toxic. Concretely, the input x is a comment on an online article (comprising one or more sentences of text) and the label y is whether it is rated toxic or not.

Experimental Settings Following Liu et al. (2021), we capped the number of tokens per example at 300 and used an initial learning rate of 0.00001. We train all approaches for up to 20 epochs with batch size 32 and ℓ_2 regularization strength of 0.01.

H.1.4. GQA-OOD (KERVADEC ET AL., 2021)

Dataset Description GQA-OOD is a real-world dataset of a visual question answering (VQA) task for validating models in OOD settings. The dataset is based on GQA dataset (Hudson & Manning, 2019), but the validation and test datasets

are transformed for valid evaluation in OOD setting. It should be noted that the train dataset is same as original GQA training dataset. They re-organize the GQA validation and test datasets to extract a data group of the most imbalanced question distribution. Then, they divide the dataset into *head* and *tail* based on rareness of the answer appearance. Then, the validation dataset results in 33,822 questions for 8,664 images in the *head* group and 17,163 questions for 6,632 images in the *tail* group. Also, the test dataset results in 1,733 questions for 365 images in the *head* group and 1,063 questions for 330 images in the *tail* group.

Experimental Settings In our experiments, we follow the experiment settings of (Shrestha et al., 2021a). Specifically, we use the UpDn architecture (Anderson et al., 2018), which is commonly utilized network structure in Visual Question Answering. For group-known models, head/tail categorization grouping is used to construct groups. For all models, we use Adam optimizer and a batch size of 128. We also fairly utilize hyperparameter settings from (Shrestha et al., 2021a) for baseline models. Due to computational constraints, the experiments were performed only once for each model.

H.2. Implementation Details

We follow the code implementation of Nam et al. (2020) for the experiments with Colored-MNIST. In the case of CelebA and CivilComments, the experiment was conducted by adapting the original code to the following settings. For GQA-OOD, we utilized the open-sourced code of Shrestha et al. (2021b), which provide the GQA-OOD protocols, codes for each baseline, and reproducible results. Also, we utilized NVIDIA RTX-3090 GPU machines to train ORC and other baselines.

H.2.1. Hyperparameter setting for V-REX

For V-REX (Krueger et al., 2020), we only need one hyperparameter λ, which controls the degree of variance minimzation of training risks. We searched the hyperparameter from $\lambda = [1, 5]$.

H.2.2. Hyperparameter setting for LfF

LfF has a hyperparameter q for generalized cross entropy term. We utilized $q = 0.7$ for the Colored-MNIST, and we tuned the hyperparamter q by searching it over $q \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ for the other datasets.

H.2.3. Hyperparameter setting for SD

We controlled the hyperparameter λ, which adjusts the degree of ℓ_2 regularization of the logits. We searched it from the range of $\lambda = [0.1, 0.3, 0.5, 0.7, 0.9]$.

H.2.4. Hyperparameter setting for JTT

JTT needs an learning of ERM-based model with T epochs, which is equal to ORC*. We tuned it with the range of $T = [2, 4, 8]$. Additionally, λ_{up}, which controls the degree of upweighting the bias-conflicting samples got tuned with the range of $\lambda_{up} = [10, 20, 50, 100]$.

H.2.5. Hyperparameter setting for ORC

For experiments with Colored-MNIST dataset and CelebA, we get two disjoint groups, bias-aligned group and bias-conflicting group. This implies that we need two smoothing factor α for each group. there are two kinds of hyper-parameters : 1) λ and 2) α_{align} and α_{conf}, which is smoothing factor for bias-aligned group and bias-conflicting group, respectively. we set the hyperparameter tuning from the combination of 1) $\lambda = [1, 5]$, 2) $\alpha_{align} = [0.1, 0.15, 0.2]$ and $\alpha_{conf} = [0, 0.01]$. These combinations of α_{align} and α_{conf} always gaurantee that α_{align} is larger than α_{conf}, which learns toward our motivation kept.

For experiments with CivilComments and GQA-OOD datasets, we need similar level of hyper-parameters as the experiments with Colored-MNIST. we set the hyperparameter tuning from the combination of 1) $\lambda = [0.1, 0.5, 1]$ and 2) $\alpha_{align} = [0.1, 0.15, 0.2]$ and $\alpha_{conf} = [0, 0.01]$.

H.2.6. Hyperparameter setting for ORC*

For experiments, there are TWO hyper-parameters : 1) λ 2) $\max(\alpha)$. Here, $\max(\alpha)$ is a pre-defined value of maximum possible value of label smoothing. We tune the hyperparameters from the combination of 1) $\lambda = [0.1, 1, 5]$, 2) $\max(\alpha) =$

$[0.1, 0.15, 0.2]$. This setting showed good performances for whole cases.