Learning Debiased Classifier with Biased Committee

Nayeong Kim [1] **Sehyun Hwang** [1] **Sungsoo Ahn** [1 2] **Jaesik Park** [1 2] **Suha Kwak** [1 2]

Abstract

This paper proposes a new method for training debiased classifier with no bias supervision. The key idea of the method is to employ a committee of classifiers as an auxiliary module that identifies bias-conflicting data and assigns large weights to them when training the main classifier. The committee is learned as a bootstrapped ensemble so that a majority of its classifiers are biased as well as being diverse, and intentionally fail to predict classes of bias-conflicting data accordingly. The consensus within the committee on prediction difficulty provides a reliable cue for identifying and weighting bias-conflicting data. Moreover, the committee is trained also with knowledge transferred from the main classifier so that it gradually becomes debiased and emphasizes more difficult data as training progresses. On five real-world datasets, our method outperforms previous arts using no bias label like ours and even surpasses those relying on bias labels occasionally.

1. Introduction

Most supervised learning algorithms for classification rely on the empirical risk minimization (ERM) principle (Vapnik, 1999). However, ERM has been known to cause a learned classifier to be biased toward spurious correlations between predefined classes and latent attributes that appear in a majority of training data (Geirhos et al., 2020). We call data with spurious correlations and holding a majority of training data *bias-guiding samples*, and the other *bias-conflicting samples*, respectively. The issue of model bias has often been addressed by exploiting explicit spurious attribute labels (Kim et al., 2019; Li & Vasconcelos, 2019; Sagawa et al., 2019; Arjovsky et al., 2019; Teney et al., 2021; Tartaglione et al., 2021; Zhu et al., 2021) or knowledge about bias types given a priori (Bahng et al., 2020). However, these methods are impractical because such supervision and prior knowledge are costly, and the methods demand extensive post hoc analysis.

Hence, a body of research has been conducted for learning debiased classifiers with no additional label for spurious attributes (Wang et al., 2021; Levy et al., 2020; Liu et al., 2021b; Nam et al., 2020; Kim et al., 2021; Lee et al., 2021). A common approach in this line of work is to employ an intentionally biased classifier as an auxiliary module (Liu et al., 2021b; Nam et al., 2020; Kim et al., 2021; Lee et al., 2021): Samples that the biased classifier has trouble handling are regarded as bias-conflicting ones and assigned large weights when used for training the main classifier to reduce the effect of bias-guiding counterparts. Although it has driven remarkable success, this approach has drawbacks due to the use of a single biased classifier. First, the quality of the biased classifier could vary by hyper-parameters (Liu et al., 2021a) and its initial parameter values (Fort et al., 2019). Further, data that the biased classifier fails to handle could include not only bias-conflicting samples but also bias-guiding ones, which differs by the quality of the classifier. These drawbacks limit the reliability and performance of debiasing methods depending on a single biased classifier, which is demonstrated in Appendix A.1.

To overcome these limitations, we propose a new method using a committee of biased classifiers as the auxiliary module, coined *learning with biased committee* (LWBC). LWBC identifies bias-conflicting samples and determines their weights through consensus on their prediction difficulty within the committee. The committee is built as a bootstrapped ensemble, *i.e.*, each of its classifiers is trained from a randomly sampled subset of the entire training dataset. This strategy not only guarantees the diversity among the classifiers, but also lets a majority of the classifiers be biased. Accordingly, a majority of the committee tends to classify bias-guiding samples correctly and fails to deal with bias-conflicting ones. The consensus on prediction difficulty within the committee thus gives a strong cue for identifying and weighting bias-conflicting samples. Also, using the consensus of multiple classifiers enables LWBC to be robust to the varying quality of individual classifiers and consequently to focus more precisely on bias-conflicting samples.

[1]CSE, Pohang University of Science and Technology (POSTECH), South Korea [2]GSAI, Pohang University of Science and Technology (POSTECH), South Korea. Correspondence to: Suha Kwak <suha.kwak@postech.ac.kr>.

Published at the ICML 2022 Workshop on Spurious Correlations, Invariance, and Stability. Baltimore, Maryland, USA.

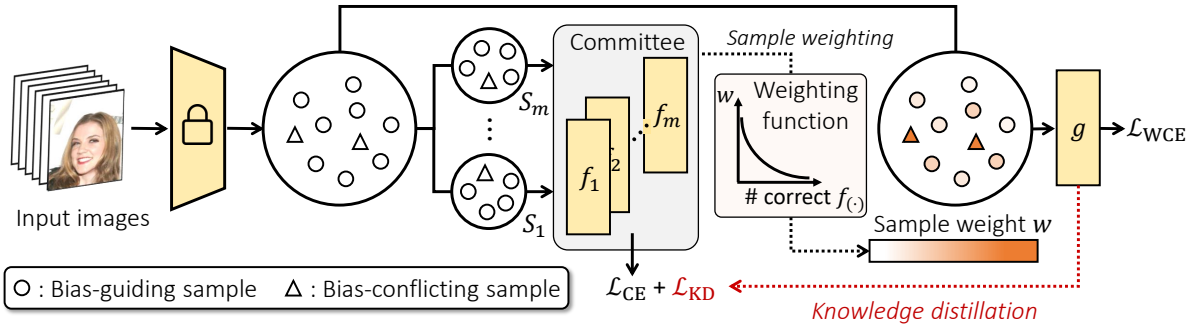


Figure 1. LWBC adopts a frozen backbone trained by self-supervised learning. A committee of auxiliary classifiers f_i is trained on top of the backbone; a random subset of training data is assigned to each classifier of the committee as labeled data for supervised learning (Eq. (1)). The committee determines the weight of each training sample based on consensus of its members, *i.e.*, the number of members that correctly predict the class of the sample (Eq. (2)). The main classifier g is trained using the weights by minimizing the weighted cross entropy loss (Eq. (3)). In turn, knowledge of the main classifier is transferred to the committee through knowledge distillation (Eq. (4)).

Moreover, unlike the biased classifier trained independently of the main classifier in the previous work, the committee in LWBC is trained with knowledge of the main classifier as well as the random subsets of training data. Specifically, the knowledge is distilled in the form of classification logits of the main classifier (Hinton et al., 2015), and each classifier of the committee utilizes the knowledge as pseudo labels of training data other than its own training set. We expect that this strategy allows the committee to become debiased gradually so that it does not give large weights to easy bias-conflicting samples, *i.e.*, those already well handled by the main classifier, and focuses more on difficult ones.

Finally, we further improve the proposed method by adopting a self-supervised representation as the frozen backbone of the committee and the main classifier. Since self-supervised learning is not dependent on class labels, it is less affected by the spurious correlations between classes and latent attributes, leading to a robust and less-biased representation. Also, by installing the committee and the main classifier on top of the representation, the classifiers can be implemented efficiently in both space and time while enjoying the rich and bias-free features given by the backbone.

LWBC is validated extensively on five real-world datasets. It substantially outperforms existing methods using no bias label and even occasionally surpasses previous arts demanding bias labels. We also demonstrate that all of the main components contribute to the outstanding performance. The main contribution of this paper is three-fold:

- We present LWBC, a new method for learning a debiased classifier with no spurious attribute label. The use of consensus within the committee allows LWBC to address the limitations of previous work.
- We propose to learn the committee using knowledge of the main classifier, unlike the previous work whose auxiliary modules do not consider the main classifier.
- LWBC demonstrates superior performance on five real-world datasets. It outperforms existing methods using no additional supervision like ours and even surpasses those relying on spurious attribute labels occasionally.

2. Proposed method

LWBC first learns a feature representation with self supervision, which is used as the frozen backbone providing rich and bias-free features to downstream modules (Section 2.1). Next, it trains a committee of m auxiliary classifiers $f_1, f_2, \ldots, f_m$ and the main classifier g on top of the self-supervised representation (Section 2.2); thanks to the self-supervised representation, the classifiers are designed concisely, using only two fully-connected layers for each.

The committee identifies bias-conflicting samples and assigns them large weights to reduce the effect of bias-guiding samples during the training of the main classifier. To this end, the committee is trained as a bootstrapped ensemble of classifiers so that a majority of its classifiers are biased as well as diverse, and intentionally fail to predict classes of bias-conflicting samples accordingly. The consensus within the committee on prediction difficulty of a sample (*e.g.*, the number of classifiers that fail to predict its class label) thus indicates how much likely the sample is bias-conflicting, and is used to compute weights for training samples. Moreover, the committee is trained also with knowledge of the main classifier so that it gradually becomes debiased along with the main classifier and emphasizes more difficult samples as training progresses. Note that the committee is an auxiliary module used only in training and thus does not impose additional computation or memory footprint in testing.

The overall process of LWBC is illustrated in Figure 1, and the following sections elaborate on each step of LWBC.

2.1. Self-supervised representation learning

As the feature extractor, we train a backbone network by self-supervised learning with BYOL (Grill et al., 2020) on the

target dataset. A self-supervised model can capture diverse patterns shared by data without being biased towards a particular class even the training set is biased. We empirically demonstrate that adopting a self-supervised representation leads to a model less biased compared with a supervised representation.

Although the representation offers rich and less-biased features, the main classifier can be still biased when it is trained by ERM; the need to explore a debiasing method for a classifier arises. We thus propose LWBC, a new debiasing method illustrated in the next section.

2.2. Learning debiased classifier with biased committee

We randomly sample m subsets of the same size, denoted by $\mathcal{S}_1, ..., \mathcal{S}_m$, from the training set $\mathcal{D}$ with replacement. Then m auxiliary classifiers $f_1, ..., f_m$ of the committee are initialized randomly, and each of the subsets $\mathcal{S}_l$ is assigned to each auxiliary classifier f_l as its training data.

The first step of LWBC is warm-up training of the committee; this is required to ensure that the committee is capable of identifying and weighting bias-conflicting samples at the beginning of the main training process. Given a mini-batch $\mathcal{B}$ at each warm-up iteration, the committee is trained by minimizing the cross-entropy loss below:

$$\mathcal{L}_{\text{CE}} = \sum_{l=1}^{m} \sum_{(x,y)\in\mathcal{S}_l\cap\mathcal{B}} \text{CE}(f_l(x), y). \tag{1}$$

Since each subset is sampled from the training set dominated by bias-guiding samples, a majority of auxiliary classifiers are also biased. At the same time, the classifiers are diverse due to their difference in initialization and training data.

After the warm-up stage, the main classifier and the committee are trained while interacting with each other. First, the main classifier is trained by the weighted cross entropy loss with the entire training set, where the sample weights are computed by considering consensus within the committee on prediction difficulty of the samples. Since a majority of auxiliary classifiers have trouble to handle bias-conflicting samples, we identity and weight bias-conflicting samples based on the number of auxiliary classifiers predicting correctly for the samples. The weight function w is given by

$$w(x) = \frac{1}{\sum_{l=1}^{m} \mathbb{1}(f_l(x) = y)/m + \alpha}, \tag{2}$$

where m is the size of the committee, f_l is the l-th classifier of the committee, and α is a scale hyper-parameter. The weight reflects how much the sample x is likely to be bias-conflicting and decreases rapidly when the number of correctly predicting classifiers increases. Then we train the main classifier g with emphasis on the bias-conflicting samples through w by the weighted cross entropy loss below:

$$\mathcal{L}_{\text{WCE}} = \sum_{(x,y)\in\mathcal{B}} w(x) \cdot \text{CE}(g(x), y), \tag{3}$$

where $\mathcal{B}$ is the mini-batch.

During training, as the main classifier is gradually debiased, samples useful for debiasing the main classifier change accordingly. To focus more on bias-conflicting samples difficult for the main classifier, we inform the quality of the main classifier to the committee by distilling the knowledge of the main classifier in the form of its classification logits (Hinton et al., 2015) and transferring the knowledge by minimizing the following KD loss:

$$\mathcal{L}_{\text{KD}} = \sum_{l=1}^{m} \sum_{(x,y)\in\mathcal{B}-\mathcal{S}_l} \text{KL}\left(\text{soft}\left(\frac{g(x)}{\tau}\right), \text{soft}\left(\frac{f_l(x)}{\tau}\right)\right), \tag{4}$$

where 'soft' is softmax function and τ is a temperature parameter. Note that we apply $\mathcal{L}_{\text{KD}}$ to the complement set of $\mathcal{S}_l$ to avoid auxiliary classifiers in the committee being identical to each other. By interacting with the main classifier, the committee gradually becomes debiased along with the main classifier. Hence, samples correctly predicted by the main classifier are less weighted and those with incorrect predictions are more weighted by the committee.

After the warm-up training, the main and the auxiliary classifiers are alternately updated with a given mini-batch at each iteration. First, we forward every sample in a mini-batch to each auxiliary classifier and compute weights of the samples through Eq. (2). With the weights, the main classifier is updated by Eq. (3) and the knowledge of the updated main classifier is transferred to the auxiliary classifiers. The auxiliary classifiers are in turn updated by a linear combination of the losses in Eq. (1) and (4). We also provide the full description of our algorithm in Appendix A.2.

Table 1. GUIDING, UNBIASED, and CONFLICTING metrics (%) for the CelebA dataset. For methods without spurious attribute labels, we mark the best and the second-best performance in **bold** and underline, respectively.

Method	Spurious attribute label	CelebA HairColor		
		GUIDING	UNBIASED	CONFLICTING
Group DRO (Sagawa et al., 2019)	✓	87.46	$85.43_{\pm 0.53}$	$83.40_{\pm 0.67}$
EnD (Tartaglione et al., 2021)	✓	94.97	$91.21_{\pm 0.22}$	$87.45_{\pm 1.06}$
CSAD (Zhu et al., 2021)	✓	91.19	89.36	87.53
ERM	✗	87.98	$70.25_{\pm 0.35}$	$52.52_{\pm 0.19}$
LfF (Nam et al., 2020)	✗	87.24	$\underline{84.24}_{\pm 0.37}$	$\underline{81.24}_{\pm 1.38}$
SSL+ERM	✗	$\mathbf{94.15}_{\pm 0.57}$	$80.48_{\pm 0.91}$	$66.79_{\pm 2.20}$
LWBC	✗	$\underline{90.57}_{\pm 2.15}$	$\mathbf{88.90}_{\pm 1.55}$	$\mathbf{87.22}_{\pm 1.14}$

3. Experiments

3.1. Evaluation metrics

We adopt six metrics. VALIDATION / TEST: average accuracy on validation / test splits. GUIDING: average accuracy

Table 2. UNBIASED and WORST-GROUP metrics (%) for the CelebA dataset. We also report the difference between UNBIASED and WORST-GROUP as GAP. For methods without spurious attribute labels, we mark the best and the second-best performance in **bold** and underline, respectively.

Method	Spurious attribute label	CelebA HairColor		
		UNBIASED	WORST-GROUP	GAP
Group DRO (Sagawa et al., 2019)	✓	$93.1_{\pm 0.21}$	$88.5_{\pm 1.16}$	4.6
SSA (Nam et al., 2022)	✓	$92.8_{\pm 0.11}$	$89.8_{\pm 1.28}$	3.0
ERM	✗	**95.6**	47.2	48.4
CVaR DRO (Levy et al., 2020)	✗	82.4	64.4	18.0
LfF (Nam et al., 2020)	✗	86.0	70.6	15.4
EIIL (Creager et al., 2021)	✗	91.9	83.3	8.6
JTT (Liu et al., 2021b)	✗	88.0	81.1	6.9
SSL+ERM	✗	$80.5_{\pm 0.9}$	$38.5_{\pm 4.1}$	42.0
LWBC	✗	$88.9_{\pm 1.6}$	$\mathbf{85.5}_{\pm 1.4}$	**3.4**

Table 3. VALIDATION, UNBIASED, and TEST metrics (%) evaluated on the ImageNet-9 and ImageNet-A datasets. For methods without spurious attribute labels, we mark the best and the second-best performance in **bold** and underline, respectively.

Method	Spurious attribute label	ImageNet-9		ImageNet-A
		VALIDATION	UNBIASED	TEST
StylisedIN (Geirhos et al., 2019)	✓	$88.4_{\pm 0.5}$	$86.6_{\pm 0.6}$	$24.6_{\pm 1.4}$
LearnedMixin (Clark et al., 2019)	✓	$64.1_{\pm 4.0}$	$62.7_{\pm 3.1}$	$15.0_{\pm 1.6}$
RUBi (Cadene et al., 2019)	✓	$90.5_{\pm 0.3}$	$88.6_{\pm 0.4}$	$27.7_{\pm 2.1}$
ERM	✗	$90.8_{\pm 0.6}$	$88.8_{\pm 0.6}$	$24.9_{\pm 1.1}$
BagNet18 (Brendel & Bethge, 2019)	✗	$67.7_{\pm 0.3}$	$65.9_{\pm 0.3}$	$18.8_{\pm 1.15}$
ReBias (Bahng et al., 2020)	✗	$91.9_{\pm 1.7}$	$90.5_{\pm 1.7}$	$29.6_{\pm 1.6}$
LfF (Nam et al., 2020)	✗	86.0	85.0	24.6
CaaM (Wang et al., 2021)	✗	**95.7**	**95.2**	32.8
SSL+ERM	✗	$94.18_{\pm 0.07}$	$93.18_{\pm 0.04}$	$34.21_{\pm 0.49}$
LWBC	✗	$94.03_{\pm 0.23}$	$93.04_{\pm 0.32}$	$\mathbf{35.97}_{\pm 0.49}$

on bias guiding samples per class. CONFLICTING: average accuracy on bias conflicting samples per class. UNBIASED: average of 'Guiding' and 'Conflicting' per class. WORST-GROUP: minimum average accuracy of group. We also provide detailed experimental settings in Appendix A.3 and description of datasets in Appendix A.4.

3.2. Quantitative results

LWBC shows superior classification accuracy among the methods that do not use the spurious attribute label on the five real-world datasets. In Tables 1 & 2, we observe LWBC outperforms existing debiasing methods using no spurious attribute label and shows comparable CONFLICTING performance with methods that exploit spurious attribute labels on CelebA, which reflects gender bias in the real world.

Table 4. CONFLICTING metric (%) evaluated on the BAR dataset. For methods without spurious attribute labels, we mark the best and the second-best performance in **bold** and underline, respectively.

Method	Spurious attribute label	BAR
		CONFLICTING
ERM	✗	$35.32_{\pm 0.46}$
ReBias (Bahng et al., 2020)	✗	$37.02_{\pm 0.26}$
LfF (Nam et al., 2020)	✗	$48.15_{\pm 0.93}$
LDD (Lee et al., 2021)	✗	$52.31_{\pm 1.00}$
SSL+ERM	✗	$60.88_{\pm 0.80}$
LWBC	✗	$\mathbf{62.03}_{\pm 0.74}$

Table 5. VALIDATION and TEST metrics (%) evaluated on the NICO dataset. For methods without spurious attribute labels, we mark the best and the second-best performance in **bold** and underline, respectively.

Method	Spurious attribute label	NICO	
		VALIDATION	TEST
Cutout (DeVries & Taylor, 2017)	✓	43.69	43.77
RUBi (Cadene et al., 2019).	✓	43.86	44.37
IRM (Arjovsky et al., 2019)	✓	40.62	41.46
Unshuffle (Teney et al., 2021)	✓	43.15	43.00
REx (Krueger et al., 2021)	✓	41.00	41.15
ERM	✗	43.77	42.61
CBAM (Woo et al., 2018)	✗	42.15	42.46
ReBias (Bahng et al., 2020)	✗	44.92	45.23
LfF (Nam et al., 2020)	✗	41.83	40.18
CaaM (Wang et al., 2021)	✗	46.38	46.62
SSL+ERM	✗	$55.63_{\pm 0.54}$	$52.24_{\pm 0.27}$
LWBC	✗	$\mathbf{56.05}_{\pm 0.45}$	$\mathbf{52.84}_{\pm 0.31}$

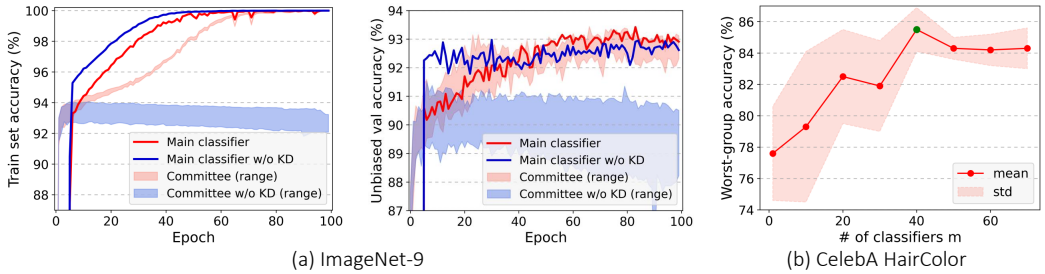
Table 6. Ablation studies using WORST-GROUP metric (%) on the CelebA HairColor dataset. We study the impact of learning from a single biased classifier (row 2), learning by committee (row 3), and transferring the knowledge of the main classifier (row 4). We mark the best performance in **bold**.

Method					CelebA HairColor
Self-sup	ERM	Single	Committee	KD	WORST-GROUP
✓	✓	✗	✗	✗	$38.5_{\pm 4.1}$
✓	✗	✓	✗	✗	$64.1_{\pm 2.4}$
✓	✗	✗	✓	✗	$81.3_{\pm 2.3}$
✓	✗	✗	✓	✓	$\mathbf{85.5}_{\pm 1.4}$

Especially, the gap between the average accuracy of groups and worst-group accuracy of LWBC is much smaller than the other methods, this means that our model fairly predicts a sample belongs to each group. Table 3 shows the results on the ImageNet-9, which is dominated by texture bias, and the ImageNet-A, which is regarded as a bias-conflicting set of ImageNet. LWBC is 9.7% better than CaaM (Woo et al., 2018) on the ImageNet-A dataset, *i.e.*, LWBC is robust to texture bias and various ImageNet biases. In Table 4, LWBC is 18.6% better than LDD, which is the previous state of the art. Compared to LfF (Nam et al., 2020) and LDD (Lee et al., 2021), which are debiasing methods with a single biased classifier, we demonstrate that learning a debiased classifier with the biased committee is more effective than a single classifier scheme. In Table 5, LWBC is 13.3% better than CaaM, *i.e.*, LWBC is well generalized to the unseen spurious attributes. Note that the validation and test set of NICO have unseen context classes and are unbiased.

3.3. Ablation study

Importance of each module in LWBC. Table 6 demonstrates through ablation studies (Liu et al., 2021b; Nam et al., 2020; Kim et al., 2021; Lee et al., 2021): (1) learning from



(a) ImageNet-9 (b) CelebA HairColor

Figure 2. Effect of knowledge distillation (Eq. (4)) and the size of the committee m. (a) Accuracy of the main classifier and the committee with or without KD loss in ImageNet-9 train set (*left*) and unbiased validation set (*right*). (b) Worst-group accuracy of LCB versus the number of classifiers within the committee, where the green dot indicates the value used in the main paper.

a single biased classifier, (2) learning with committee, and (3) transferring the knowledge of the main classifier. First, we train a classifier by ERM (row 1) then assign a weight value 50 to the wrongly predicted samples and 1 to other samples. Then re-training the classifier with the weights (row 2). Comparing these two results demonstrates that up-weighting scheme with a biased classifier is effective to debiasing a classifier. Then we increase the number of biased classifiers and compute the sample weights using our weight function Eq. (2) (row 3). Learning with the committee shows a remarkable improvement in the worst-group accuracy. Moreover, knowledge distillation that enables the committee to interact with the main classifier further improves performance (row 4).

Effectiveness of transferring knowledge of the main classifier. Figure 2(a) shows the range of unbiased validation accuracy of classifiers in the committee and unbiased validation accuracy of the main classifier during training. The mean accuracy of classifiers in the committee gradually increases following the accuracy of the main classifier. Also, the accuracy of the main classifier gradually increases as training progresses. On the other hand, the accuracy of classifiers in the committee trained without KD loss does not increase or even decrease.

Number of classifiers. We compare the results of the main classifier trained with the different number of auxiliary classifiers. Figure 2(b) shows the worst-group accuracy versus the number of auxiliary classifiers m on CelebA. With a single auxiliary classifier, the main classifier shows the lowest worst-group accuracy, but the accuracy increases as m increases. When larger than 40, the number of classifiers has little effect on learning the main classifier.

Impact of bootstrapping. In Table 7, we study the impact of auxiliary classifiers trained on a subset of the training set ('subset') and auxiliary classifiers trained on the full training set ('full'). Each auxiliary classifier in the 'full' setting learns from the same data, but they are differently initialized. Unlike LWBC, the KD loss in Eq. (2) is calculated using the entire training set for the 'full' experiment. So, the comparison between 'full' and 'subset' shows the impact of bootstrapping excluding the impact of random initialization. As shown in Table 7, learning with auxiliary classifiers trained on the same full training dataset degrades the performance at every metric. We believe that this is because the auxiliary classifiers trained on a subset of the training set are more diverse and biased than those trained on the full training set. We also provide more ablation studies in Appendix A.5 and qualitative results in Appendix A.6.

Table 7. Ablation study using GUIDING, UNBIASED, CONFLICTING, and WORST-GROUP metrics (%) on the CelebA HairColor dataset. We study the impact of training auxiliary classifiers using subsets of training data. The best performance is marked in **bold**.

Trainset	GUIDING	UNBIASED	CONFLICTING	WORST-GROUP
Full	90.40$_{\pm 3.76}$	87.00$_{\pm 1.48}$	83.60$_{\pm 1.14}$	78.0$_{\pm 4.3}$
Subset	**90.57**$_{\pm 2.15}$	**88.90**$_{\pm 1.55}$	**87.22**$_{\pm 1.14}$	**85.5**$_{\pm 1.4}$

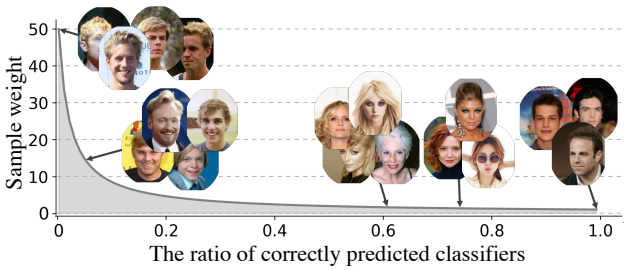


Figure 3. Qualitative examples on CelebA dataset. *Hair color* is target and *Gender* is the bias.

4. Conclusion

We have proposed a new method for learning a debiased classifier with a committee of auxiliary classifiers. The committee is learned in a way that consensus on predictions of its classifiers to identify and weight bias-conflicting data. The main debiased classifier is then trained with an emphasis on the bias-conflicting data.Moreover, we demonstrated that self-supervised learning is a solid yet unexplored baseline for debiasing. Coupled with a self-supervised feature extractor, our method achieved state-of-the-art by large margins on real-world datasets.

Acknowledgement

This work was supported by IITP grant funded by the Korea government (MSIT) (No. 2022-0-00926).

References

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

Bahng, H., Chun, S., Yun, S., Choo, J., and Oh, S. J. Learning de-biased representations with biased representations. In *Proc. International Conference on Machine Learning (ICML)*, pp. 528–539. PMLR, 2020.

Brendel, W. and Bethge, M. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *arXiv preprint arXiv:1904.00760*, 2019.

Cadene, R., Dancette, C., Cord, M., Parikh, D., et al. Rubi: Reducing unimodal biases for visual question answering. In *Proc. Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 841–852, 2019.

Clark, C., Yatskar, M., and Zettlemoyer, L. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4069–4082, 2019.

Creager, E., Jacobsen, J.-H., and Zemel, R. Environment inference for invariant learning. In *International Conference on Machine Learning*, pp. 2189–2200. PMLR, 2021.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: a large-scale hierarchical image database. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

DeVries, T. and Taylor, G. W. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Fort, S., Hu, H., and Lakshminarayanan, B. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *Proc. International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=Bygh9j09KX.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., et al. Bootstrap your own latent: A new approach to self-supervised learning. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

He, Y., Shen, Z., and Cui, P. Towards non-iid image classification: A dataset and baselines. In *Pattern Recognition*, volume 110, pp. 107383. Elsevier, 2021.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15262–15271, 2021.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. Adversarial examples are not bugs, they are features. In *Proc. Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 125–136, 2019.

Kim, B., Kim, H., Kim, K., Kim, S., and Kim, J. Learning not to learn: Training deep neural networks with biased data. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9012–9020, 2019.

Kim, E., Lee, J., and Choo, J. Biaswap: Removing dataset bias with bias-tailored swapping augmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 14992–15001, 2021.

Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Zhang, D., Le Priol, R., and Courville, A. Out-of-distribution generalization via risk extrapolation (rex). In *Proc. International Conference on Machine Learning (ICML)*, pp. 5815–5826. PMLR, 2021.

Lee, J., Kim, E., Lee, J., Lee, J., and Choo, J. Learning debiased representation via disentangled feature augmentation. In *Proc. Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Levy, D., Carmon, Y., Duchi, J. C., and Sidford, A. Large-scale methods for distributionally robust optimization. In *Proc. Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 8847–8860, 2020.

Li, Y. and Vasconcelos, N. Repair: Removing representation bias by dataset resampling. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 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 *Proc. International Conference on Machine Learning (ICML)*, pp. 6781–6792. PMLR, 2021a.

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 *Proc. International Conference on Machine Learning (ICML)*, pp. 6781–6792. PMLR, 2021b.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 3730–3738, 2015.

Nam, J., Cha, H., Ahn, S., Lee, J., and Shin, J. Learning from failure: Training debiased classifier from biased classifier. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2020.

Nam, J., Kim, J., Lee, J., and Shin, J. Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation. In *Proc. International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=_F9xpOrqyX9.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, pp. 1–42, April 2015.

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

Tartaglione, E., Barbano, C. A., and Grangetto, M. End: Entangling and disentangling deep representations for bias correction. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13508–13517, 2021.

Teney, D., Abbasnejad, E., and van den Hengel, A. Unshuffling data for improved generalization in visual question answering. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 1417–1427, 2021.

Vapnik, V. *The nature of statistical learning theory*. Springer science & business media, 1999.

Wang, T., Zhou, C., Sun, Q., and Zhang, H. Causal attention for unbiased visual recognition. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 3091–3100, 2021.

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. Cbam: Convolutional block attention module. In *Proc. European Conference on Computer Vision (ECCV)*, pp. 3–19, 2018.

Zhu, W., Zheng, H., Liao, H., Li, W., and Luo, J. Learning bias-invariant representation by cross-sample mutual information minimization. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 15002–15012, October 2021.

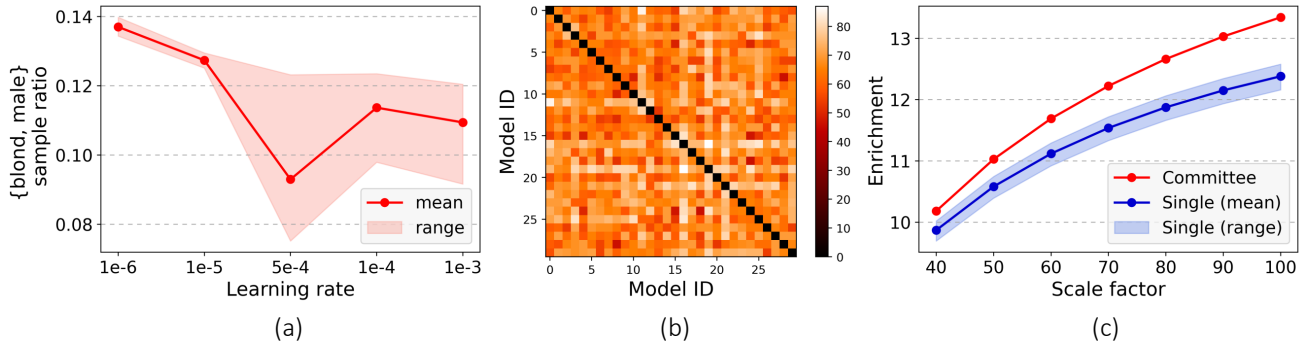


Figure 4. Analysis on the instability of a single biased classifier in mining and weighting bias-conflicting samples. The experiments are conducted on the CelebA dataset, in which samples with `blond` and `male` attributes are bias-conflicting.

Algorithm 1 Learning a debiased classifier with a biased committee

input training set $\mathcal{D}$, batch size b, size of the committee m, learning rate η, scale hyper-parameter α, balancing hyper-parameter λ, total number of iterations t, number of warm-up iterations t_w

Draw m random subsets $\mathcal{S}_1, ..., \mathcal{S}_m$ of $\mathcal{D}$.
Initialize auxiliary classifiers of the committee $\{f_l(x;\theta_l)\}_{l=1}^m$.
Initialize the main classifier $g(x;\phi)$.
for $j = 1, ..., t_w$ **do**
  Draw a mini-batch $\mathcal{B} = \{(x_i, y_i)\}_{i=1}^b$ from $\mathcal{D}$.
  $\theta_l \leftarrow \theta_l - \eta \nabla_{\theta_l} \mathcal{L}_{\text{CE}} \quad \forall l = 1, ..., m$ Eq. (1)
end for
for $j = t_w + 1, ..., t$ **do**
  Draw a mini-batch $\mathcal{B} = \{(x_i, y_i))\}_{i=1}^b$ from $\mathcal{D}$.
  $w(x_i) = 1/(\sum_{l=1}^m \mathbb{1}(f_l(x_i) = y_i)/m + \alpha) \quad \forall x_i \in \mathcal{B}$ Eq. (2)
  $\phi \leftarrow \phi - \eta \nabla_{\phi} \mathcal{L}_{\text{WCE}}$ Eq. (3)
  $\theta_l \leftarrow \theta_l - (1-\lambda)\eta \nabla_{\theta_l} \mathcal{L}_{\text{CE}} - \lambda \eta \nabla_{\theta_l} \mathcal{L}_{\text{KD}} \quad \forall l = 1, ..., m$ Eq. (1), Eq. (4)
end for

A. Appendix

A.1. Limitations of using a single biased classifier

We empirically demonstrate the limitations of debiasing methods depending on an single biased classifier in Figure 4. Figure 4(a) shows the ratio of bias-conflicting samples to all incorrectly predicted by a single biased classifier. The ratio highly fluctuates by the learning rate of the classifier and varies up to 4%p due to different initialization even with a fixed learning rate, meaning that using a single biased classifier is sensitive to hyper-parameters. Figure 4(b) is a matrix of disagreement on predictions of biased classifiers. For pairs of biased classifiers initialized differently, we measure the number of bias-conflicting samples for which the classifiers predict differently. The results suggest that individual biased classifiers are sensitive to initialization. Figure 4(c) shows comparisons between a single biased classifier and a committee of biased classifiers. Higher enrichment implies more precise mining and weighting of bias-conflicting samples (Liu et al., 2021a). The committee clearly outperforms the single biased classifier in terms of enrichment.

A.2. Algorithm for LWBC

The overall process of LWBC is given formally in Algorithm 1.

A.3. Experimental setup

Implementation details. For the training self-supervised model, we train ResNet-18 (He et al., 2016) following self-supervised learning with BYOL (Grill et al., 2020) on the target dataset. During the self-supervised learning, a random patch of input image is cropped, resized to 224×224 pixels, flipped horizontally at random, and distorted by a random sequence of

brightness, contrast, saturation, hue adjustments, and grayscale conversion; since the color information is key feature to classify `HairColor` class on the celebA dataset, we adopt only brightness and contrast as color distortion when it trains celebA. Since NICO and BAR are very small dataset, we train self-supervised model from ImageNet (Deng et al., 2009) pretrained parameters as an initial point. Except for the experiments using BAR and NICO, we train self-supervised models from scratch. We use the self-supervised ResNet-18 as a backbone network except for the last fully connected layer. We set the batch size to {64, 64, 128, 256}, learning rate {1e-3, 1e-3, 1e-4, 6e-3}, the size of the committee m to {30, 30, 30, 40}, the size of subset $\mathcal{S}_l$ to {10, 10, 80, 300}, λ to {0.9, 0.6, 0.6, 0.6}, and τ to {1, 1, 1, 2.5}, respectively for {BAR, NICO, Imagenet-9, CelebA} and α to 0.02. Note that we average the accuracy of each training over three independent trials with random seeds.

A.4. Datasets

CelebA. CelebA (Liu et al., 2015) is a dataset for face recognition where each sample is labeled with 40 attributes. Following the experiment configuration suggested by Nam et al. (Nam et al., 2020), we focus on `HairColor` and `HeavyMakeup` attributes that are spuriously correlated with `Gender` attributes, *i.e.*, most of the CelebA images with `blond hair` are women. As a result, the biased model suffers from performance degradation when predicting `HairColor` attribute on males. Therefore, we use `HairColor` as the target attribute and `Gender` as a spurious attribute, the same as `HeavyMakeup`.

ImageNet-9. ImageNet-9 (Ilyas et al., 2019) is a subset of ImageNet (Russakovsky et al., 2015) containing nine super-classes. Following the setting adopted by (Bahng et al., 2020), we conduct experiments with 54,600 training images and 2,100 validation images. ImageNet-9 tempts to have a correlated object class and image texture. We follow the evaluation scheme adopted by (Bahng et al., 2020), and we report the unbiased accuracy of the validation set, which is computed as average accuracy on every object-texture combination.

ImageNet-A. ImageNet-A (Hendrycks et al., 2021) contains real-world images misclassified by ImageNet-trained ResNet 50 (He et al., 2016). Since such failure is caused when a model too heavily relies on the color, textures, and backgrounds. The ImageNet-A dataset could be a bias-conflicting set w.r.t. various ImageNet biases. This dataset is used only for evaluating a model trained on ImageNet-9.

BAR. The Biased Action Recognition (BAR) dataset (Nam et al., 2020) is a real-world image dataset designed to evaluate the spurious correlation between human action and place on real-world images. Originally the given training set of BAR consists of only 100% bias-guiding samples, and its test set consists of only bias-conflicting samples. In our setting, we use 10% of the original BAR training set as validation and set the bias-conflicting ratio of the training set to 1%.

NICO. NICO (He et al., 2021) is a real-world dataset for simulating the out-of-distribution image classification task. Following the setting used by Wang et al. (Wang et al., 2021), we use an animal subset of NICO, which is labeled with 10 object and 10 context classes for evaluating the debiasing methods. The training set consists 7 context classes per object class and they are long-tailed distributed (e.g., most dog images are appeared 'on grass' the others are appeared on 6 contexts). The validation and test set consists of 10 context classes with 3 unseen context classes per object class. We verify the ability of debiasing a model from object-context correlations by evaluating on NICO.

A.5. Ablation study

Self-supervised representation as a solid baseline. We empirically investigate the potential of self-supervised representation as a solid baseline for the debiasing task. We train a classifier on the top of the self-supervised representation by ERM. The results are denoted by 'SSL+ERM' and compared with 'ERM', which is a fully supervised classification model in Table 1, 2, 3, 4, 5. 'SSL+ERM' outperforms not only ERM but also the previous state-of-the-art on all the datasets except for CelebA. 'SSL+ERM' is less biased than the model trained by fully supervised learning.

Impact of backbone. In Table 8, we study the impact of a frozen backbone trained by self-supervised learning (row 7-9) compared to supervised learning (row 1-3 for ERM backbone and row 4-6 for ImageNet pretrained backbone). Within the results using the same backbone, learning with the committee and transferring the knowledge of the main classifier to the committee improve performance in all metrics compared with the ERM classifier, regardless of the backbone. Regarding the performance of the ERM classifier on top of each backbone (row 1, 4, 7), the ERM backbone leads to the best performance among the three backbones since the ERM backbone is trained with class labels. However, the ERM backbone was not useful when coupled with our method (learning with the committee and KD) dedicated to debiasing. This shows the limitation of conventional representation based on supervised learning. Comparing ImageNet pretrained backbone and self-supervised

Table 8. Ablation studies using UNBIASED, CONFLICTING, and WORST-GROUP metric (%) on the CelebA HairColor dataset. We study the impact of frozen backbone trained by supervised learning on celebA (row 1-3), supervised learning on ImageNet (row 4-6), and self-supervised learning on celebA (row 7-9). Learning by ERM (row 1, 4, 7), learning by committee (row 2, 5, 8), and transferring the knowledge of the main classifier (row 3, 6, 9). We mark the best performance in **bold**.

Backbone			Method			CelebA HairColor		
Supervised on celebA	ImageNet pretrained	Self-sup on celebA	ERM	Committee	KD	UNBIASED	CONFLICTING	WORST-GROUP
✓	✗	✗	✓	✗	✗	$94.6_{\pm 0.01}$	$70.3_{\pm 0.2}$	$45.2_{\pm 0.6}$
✓	✗	✗	✗	✓	✗	$78.9_{\pm 10.0}$	$75.2_{\pm 6.3}$	$54.0_{\pm 27.0}$
✓	✗	✗	✗	✓	✓	$80.0_{\pm 9.4}$	$78.9_{\pm 3.1}$	$61.1_{\pm 24.4}$
✗	✓	✗	✓	✗	✗	$75.3_{\pm 2.3}$	$60.9_{\pm 1.8}$	$28.0_{\pm 5.9}$
✗	✓	✗	✗	✓	✗	$84.2_{\pm 1.0}$	$80.9_{\pm 0.8}$	$68.9_{\pm 2.9}$
✗	✓	✗	✗	✓	✓	$85.1_{\pm 0.6}$	$82.4_{\pm 1.4}$	$76.6_{\pm 4.6}$
✗	✗	✓	✓	✗	✗	$80.5_{\pm 0.9}$	$66.8_{\pm 2.2}$	$38.5_{\pm 4.1}$
✗	✗	✓	✗	✓	✗	$88.6_{\pm 1.3}$	$84.0_{\pm 1.7}$	$81.3_{\pm 1.4}$
✗	✗	✓	✗	✓	✓	$\mathbf{88.9}_{\pm 1.6}$	$\mathbf{87.2}_{\pm 1.1}$	$\mathbf{85.5}_{\pm 1.4}$

trained backbone (both are target-label-free schemes), the backbone trained by self-supervised learning is always better than the ImageNet pretrained backbone in our experiments. We believe that this is because a frozen backbone trained by self-supervised learning on a target dataset gives rich and bias-free features. Surprisingly, the main classifier learned by the committee and KD on top of ImageNet pretrained frozen backbone using the same hyper-parameters outperforms LfF (Nam et al., 2020), which demonstrates that the advantage of our method is not limited to a specific backbone network.

A.6. Qualitative results

A.6.1. CLASS ACTIVATION MAP

Figure 5 shows the class activation map (CAM) of the main classifier, those of auxiliary classifiers of the committee, and a consensus graph on a bias-guiding sample of CelebA, and Figure 6 shows them on a bias-conflicting sample of CelebA. We mark a classifier that correctly predicts the class of the sample in 'correct', otherwise 'incorrect'.

In Figure 5, a majority of auxiliary classifiers correctly predict the class of the bias-guiding sample, but they focus on facial appearance. Since auxiliary classifiers have a consensus on 'correct', the main classifier less focus on the sample during training.

In Figure 6, a majority of auxiliary classifiers wrongly predict the class of the bias-conflicting sample because they focus on facial appearance. Since auxiliary classifiers have a consensus on 'incorrect', the main classifier focuses more on the sample during training. The main classifier does not focus on facial appearance to correctly predict both the bias-guiding and bias-conflicting samples.

As we expected, a majority of auxiliary classifiers focus on facial appearance, *i.e.*, auxiliary classifiers exploit `gender` feature rather than `HairColor` feature to classify an image. However, the main classifier focuses more on `HairColor` feature than the auxiliary classifiers.

A.6.2. QUALITATIVE EXAMPLES WITH WEIGHTS

Figure 7, 8, 9 show the graph of weight function in Eq. 2 and examples on weight values. As illustrated on Figure 7, 8, 9, our method not only up-weights the bias-conflicting samples and down-weights the bias-guiding samples but also imposes finer weight according to difficulty of a sample.

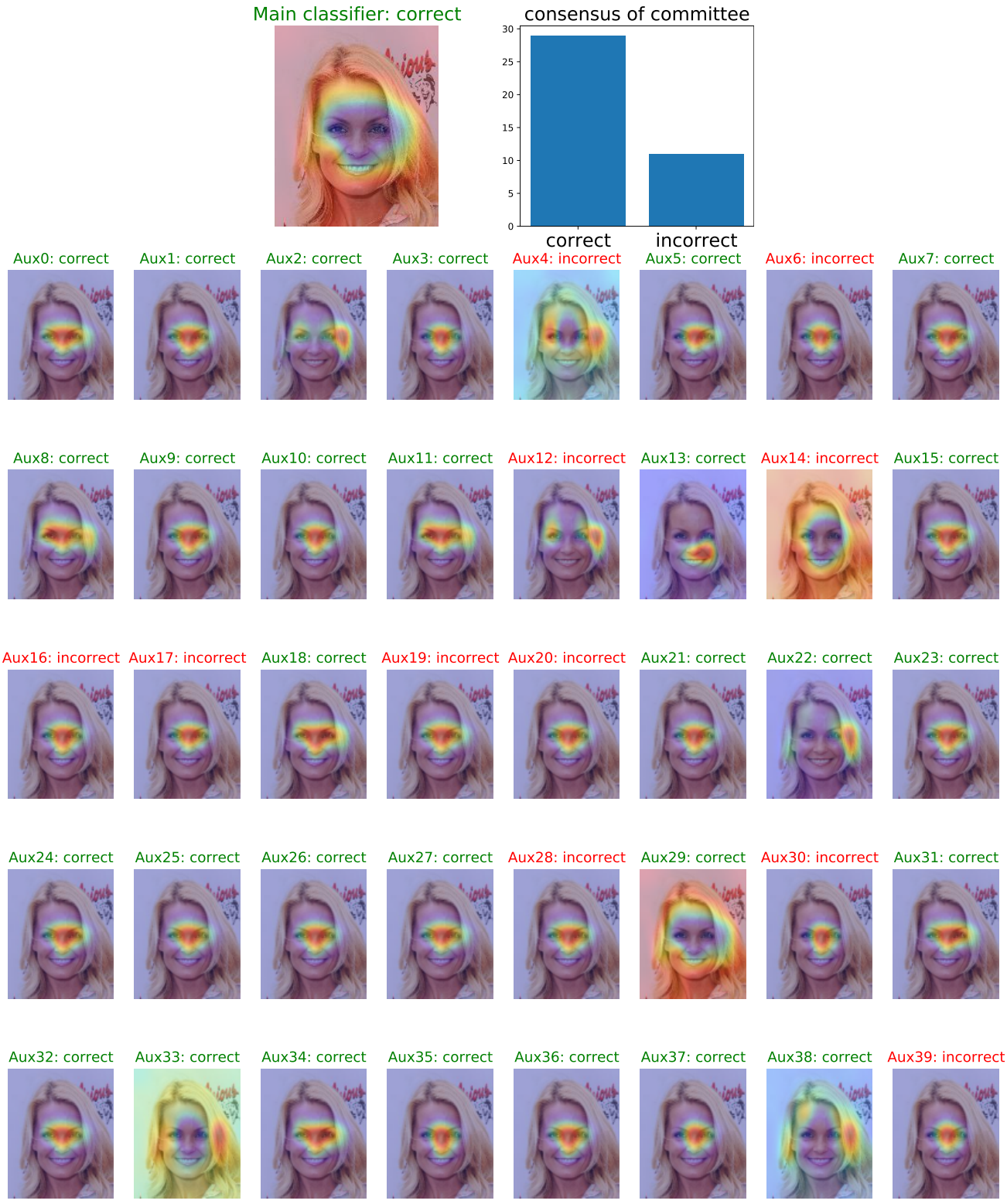


Figure 5. Class activation maps on a bias-guiding sample of celebA and consensus graph. We mark a classifier that correctly predicts the class of the sample in 'correct', otherwise 'incorrect'.

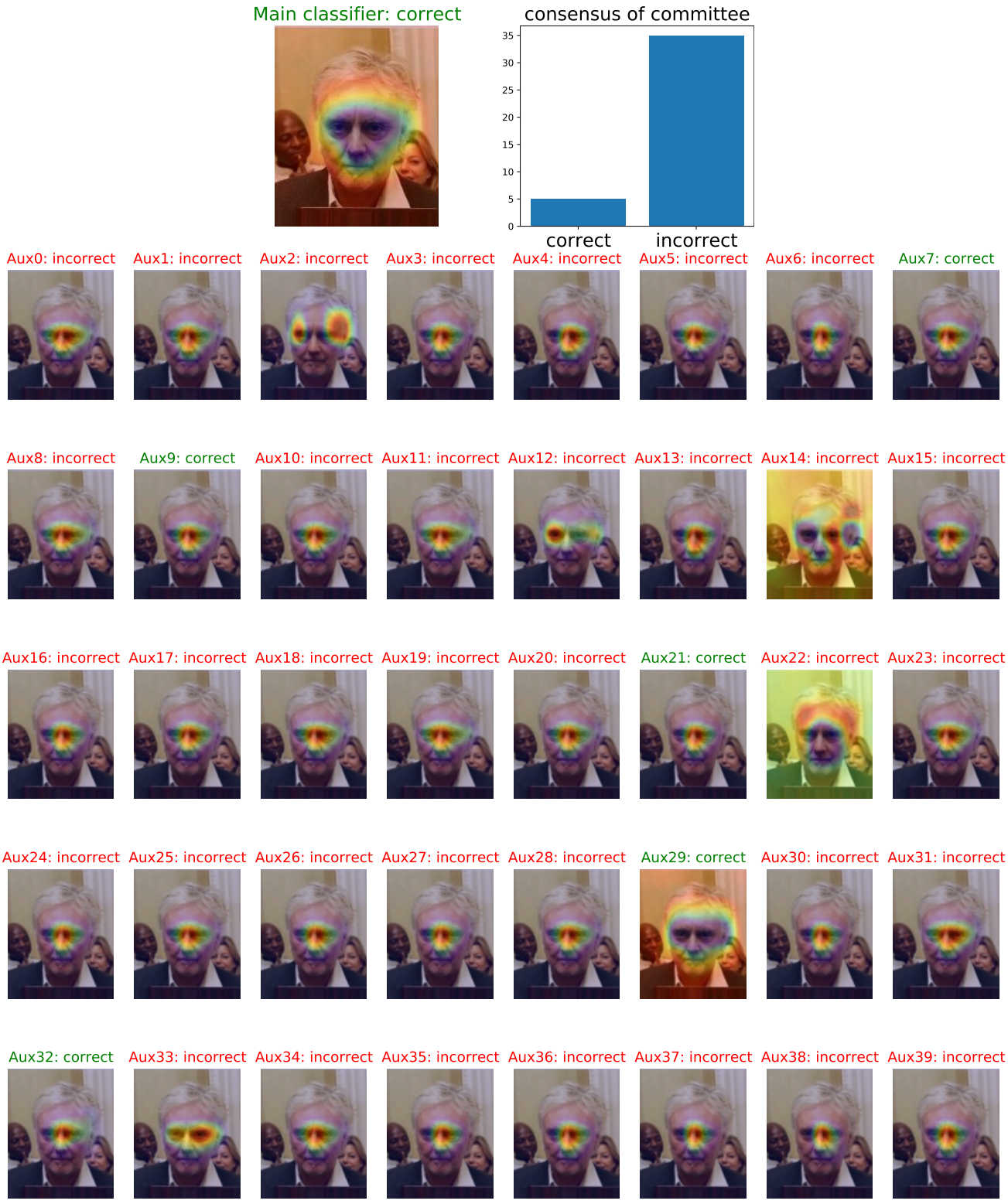


Figure 6. Class activation maps on a bias-conflicting sample of celebA and consensus graph. We mark a classifier that correctly predicts the class of the sample in 'correct', otherwise 'incorrect'.

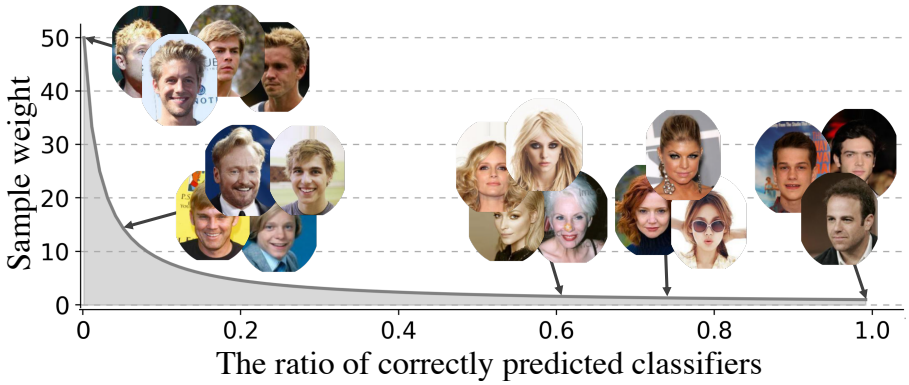


Figure 7. Qualitative examples on CelebA. `HairColor` is target and `Gender` is the bias.

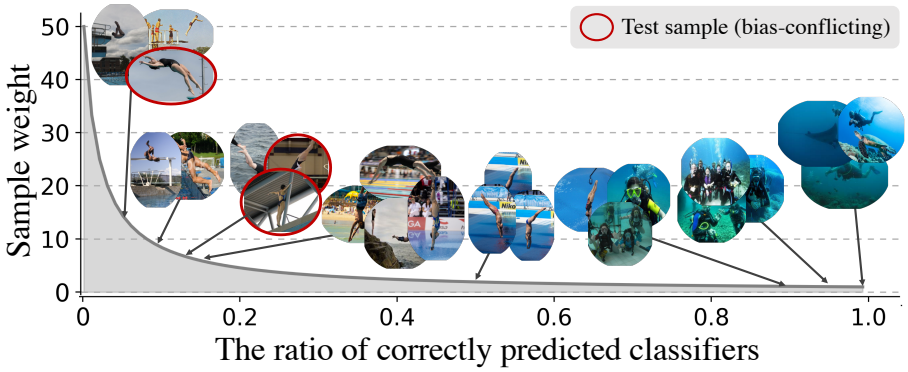


Figure 8. Qualitative examples on diving class of BAR. `Action` is target and `Place` is the bias. A majority of 'diving' images on training set include a body of water or the surface of a body of water.

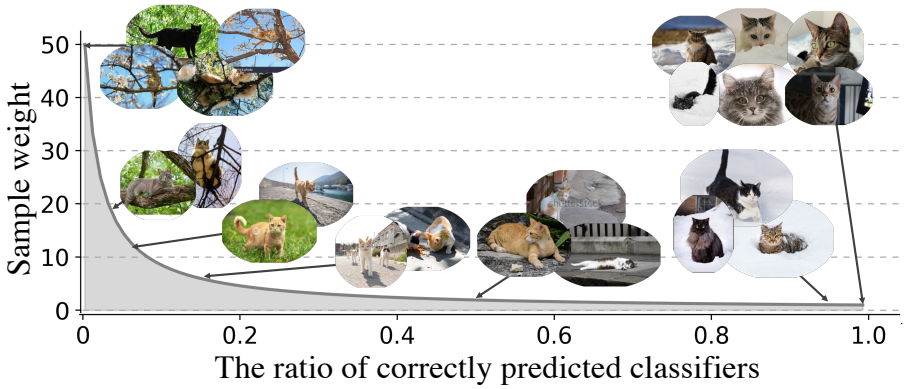


Figure 9. Qualitative examples on cat class of NICO. `Species` is target and `Context` is the bias. A majority of 'cat' images on training set have 'on snow' or 'at home' context.