Supplementary Material: Adaptive Bias Discovery for Learning Debiased Classifier

Jun-Hyun Bae[1], Minho Lee[1,2], and Heechul Jung[1]

[1] Kyungpook National University, Daegu, Republic of Korea
{junhyun.bae, heechul}@knu.ac.kr
[2] ALI Co., Ltd., Daegu, Republic of Korea
mholee@gmail.com

1 Proof for Theorem 1

Theorem 1 (Bound on Worst-Case Risk Gap).
Let $\mathcal{L}_g(f_\theta) = \mathbb{E}_{(x,y)\sim\hat{\mathcal{P}}_g}[l(f_\theta;(x,y))]$ *denote the expected loss for group* g*, and define*

$$\Delta(\theta) = \max_{g\in\mathcal{G}} \mathcal{L}_g(f_\theta) - \min_{g\in\mathcal{G}} \mathcal{L}_g(f_\theta)$$

as the maximum discrepancy in expected loss across the groups. Then, for any θ*, the gap given by Equation (9) is bounded by:*

$$Gap(\theta) \leq \frac{2^{m-1}-1}{2^m-1}\Delta(\theta),$$

where m *denotes the cardinality of* $\mathcal{G}$*.*

Proof. Let m be the number of groups which are predefined with bias information for the group DRO learning framework, i.e. the cardinality of $\mathcal{G}$. Without loss of generality, assume that the sequence of the expected losses is ordered in decreasing order, $\mathcal{L}_1(f_\theta) \geq \mathcal{L}_2(f_\theta) \geq \cdots \geq \mathcal{L}_m(f_\theta)$. Consider a set $\mathcal{G}'$ that encompasses all possible selections of groups, formulated as:
$\mathcal{G}' = \{(\mathcal{L}_1(f_\theta)), (\mathcal{L}_2(f_\theta)), \ldots, (\mathcal{L}_m(f_\theta)), \ldots, (\mathcal{L}_1(f_\theta), \mathcal{L}_2(f_\theta), \ldots, \mathcal{L}_m(f_\theta))\}$.

The number of possible cases for groups including $\mathcal{L}_1(f_\theta)$ is given by the sum:

$$1 + \binom{m-1}{1} + \binom{m-1}{2} + \cdots + \binom{m-1}{m-1} = 2^{m-1}. \tag{1}$$

Since $\mathcal{L}_1(f_\theta) \geq \mathcal{L}_2(f_\theta) \geq \cdots \geq \mathcal{L}_m(f_\theta)$, we deduce that the optimal value of the maximization problem for these groups is $\mathcal{L}_1(f_\theta)$. Therefore, we can write:

$$\mathbb{E}_{\mathcal{G}'\sim\mathcal{B}}\left[\max_{g\in\mathcal{G}'} \mathbb{E}_{(x,y)\sim\hat{\mathcal{P}}_g}[l(f_\theta;(x,y))]\right] = \sum_{i=1}^{m} 2^{m-i} p_i \mathcal{L}_i(f_\theta), \tag{2}$$

where p_i denotes the sampling probability for each $\mathcal{L}_i(f_\theta)$ satisfying $\sum_{i=1}^{m} 2^{m-i}p_i = 1$. Then, we can rewrite Equation (9) as:

$$Gap(\theta) = \mathcal{L}_1(f_\theta) - \sum_{i=1}^{m} 2^{m-i} p_i \mathcal{L}_i(f_\theta) \geq 0. \tag{3}$$

Without making specific assumptions about the probabilities associated with the observation of groups, we set p_i to have uniform probability, $p_i = \frac{1}{2^m-1}$. Then,

$$Gap(\theta) = \frac{1}{2^m-1}\left[(2^{m-1}-1)\mathcal{L}_1(f_\theta) - \sum_{i=2}^{m} 2^{m-i}\mathcal{L}_i(f_\theta)\right] \tag{4}$$

$$= \frac{1}{2^m-1}\left[\sum_{i=2}^{m} 2^{m-i}\mathcal{L}_1(f_\theta) - \sum_{i=2}^{m} 2^{m-i}\mathcal{L}_i(f_\theta)\right] \tag{5}$$

$$= \frac{1}{2^m-1}\left[\sum_{i=2}^{m} \left\{2^{m-i}(\mathcal{L}_1(f_\theta) - \mathcal{L}_i(f_\theta))\right\}\right] \tag{6}$$

$$\leq \frac{2^{m-1}-1}{2^m-1}\Delta(\theta) \tag{7}$$

This completes the proof, implying that the worst-case risk gap is bounded above by $\frac{2^{m-1}-1}{2^m-1}\Delta(\theta)$.

2 Experimental Details

In this section, we detail the experimental setup used to evaluate the proposed method, including the dataset, architectural design, training environments, and specific hyperparameters. The careful configuration of these components ensures that our evaluation accurately reflects the performance of the model under different bias conditions.

2.1 Colored MNIST

We consider two training environments, each containing 25,000 MNIST images and one validation environment with 10,000 images. For the test environment, we utilize official test images of the MNIST dataset with 10,000 data samples, then add bias features. The inclusion of color and patch attributes is designed to have high correlations with target labels; however, these correlations are unstable as they vary across the training environments.

We train a simple Multi-Layer Perceptron (MLP) with one hidden layer, utilizing the ReLU activation function. The hidden layer consists of 390 units, and we set the learning rates to $\alpha = 1 \times 10^{-1}$ and $\beta = 1 \times 10^{-3}$. We set the softmax temperature coefficient to $\tau = 1$ in the experiments on Colored MNIST. Our experiments employ three steps of gradient descent for adaptive bias discovery stage, and use early stopping on the validation environments with $p_e = 0.2$ that has similar distribution to the training distribution.

2.2 MetaShift

MetaShift [9] serves as a benchmark for measuring the performance of machine learning models across diverse data distributions. Metashift provides a quantifi-

able distance score to gauge the disparity between training and test domain distributions. Therefore, this benchmark enables the measurement of performance degeneration according to the degree of distribution shifts.

In our study, we focus on the Cat vs. Dog classification task within MetaShift dataset. Specifically, the test data samples consist of images labeled as dog(shelf) and cat(shelf), featuring dogs or cats with shelves. For the training data, the cat class samples include images from the cat(sofa+bed) categories. The training samples for the dog class vary according to the degree of distributional shift, with categories such as dog(cabinet + bed), dog(bag+box), dog(bench + bike), and dog(boat + surfboard), which correspond to distributional distances from dog(shelf) of 0.44, 0.71, 1.12, and 1.43, respectively.

For the experiment, we train a ResNet-50 model [6] with an SGD optimizer at a learning rate of 1×10^{-3} and L2 regularization of 1×10^{-4}. The softmax temperature coefficient in this experiment is set to $\tau = 1$.

2.3 CivilComments-wilds

CivilComments-wilds dataset [2,8] is a benchmark for evaluating a model's ability to generalize across shifts in the toxic text classification task. Concretely, the dataset offers an insight into how toxicity classifiers may inadvertently learn biases present in the training data, associating toxicity with mentions of specific demographics [5,10]. The data is divided into 16 groups based on 8 demographic identities and toxicity labels, and models are evaluated by their worst-group accuracy. We utilize 269,038 training data samples, 45,180 validation data samples, and 133,782 test data samples.

We use the DistilBERT-base-uncased model for five epochs with early stopping and apply Bayesian optimization for hyperparameter tuning.

2.4 MultiNLI

MultiNLI dataset serves as a benchmark for evaluating our model's ability to handle biases. We use training, validation, and test splits from [11], as the official test set is inaccessible, consisting of 206,175, 82,462, and 123,712 samples, respectively. We identify two biases for evaluation. First, we annotate cases where negation words in the hypothesis, such as *nothing, never, nobody*, and *no*, are highly correlated with the label *contradiction*, using annotations from [11]. Second, we manually annotate instances in which a high lexical overlap of more than five words between the premise and hypothesis correlates with the label *entailment*.

We fine-tune pretrained BERT [4] for three epochs, following the original settings as reported in [4]. The learning rate is set with a fixed linearly-decaying starting point at 2×10^{-5} for both α and β. As in the previous experiments, the softmax temperature coefficient is set to $\tau = 1$.

2.5 Camelyon17-wilds

In the field of machine learning-based medical image processing, OoD generalization is a critical problem in obtaining a universally applicable prediction model across various hospitals. Camelyon17-wilds dataset [1,8] is a medical image classification benchmark explicitly targeting this problem. The main goal of Camelyon17-wilds is to achieve high prediction accuracy for predicting the presence of tumor tissue on image patches taken from hospitals not included in the training data.

The training data consists of image patches from three hospitals, while the test data contains patches from a distinct hospital not represented in the training set. This test hospital provides the most visually unique patches compared to the other data. The final model selection is performed based on the accuracy on OoD validation data, different from both training and test distributions. Validation data also comes from a distinct distribution but shares more visual patterns with training data than test data.

Camelyon17-wilds is a binary classification task to determine whether a given 96×96 image patch's central region contains any tumor tissue. It includes 302,436 training patches, 34,904 OoD validation data, and 85,054 OoD test patches.

We employ DenseNet-121 [7] without pretrained parameters, training it from scratch on Camelyon17-wilds, following the official setting from [8]. We utilize Bayesian optimization for hyperparameter tuning.

2.6 FMoW-wilds

FMoW-wilds dataset [3,8] serves as another benchmark for OoD generalization, encompassing satellite images captured across various locations and times. This dataset contains RGB satellite images, and the associated task involves classifying 62 different functional purposes of buildings and land, guided by the images and metadata. The latter provides the location and time information for each image. The training data spans images taken from 2002 to 2013, while the validation and test data cover 2013 to 2015 and 2016 to 2017, respectively. FMoW-wilds aims to evaluate a model's generalizability to future images.

The dataset is divided into five geographic regions for each data split, representing the locations where the images were captured. The OoD generalization performance is assessed using the worst-region accuracy for the test data, where the data collection period does not overlap with the training data.

FMoW-wilds presents an image classification task targeting the prediction of 62 categories of building or land use from given 224×224 images. The training split includes 76,863 images from 2002 to 2013, while the validation split contains 19,915 images from 2013 to 2015. The OoD test data comprises 22,108 images taken from 2016 to 2018. All splits include images from five regions: Africa, the Americas, Oceania, Asia, and Europe. The model is evaluated by measuring the worst-region accuracy on the OoD test data.

For this task, we employ a DenseNet-121 model pretrained on the ImageNet dataset and finetune it on the FMoW-wilds images. Following the official hyper-

Table 1: Test accuracy (%) on the vanilla MNIST dataset which does not include any synthetic biases. Note that the test environment has noise on the shape of digits $p_c = 0.25$; hence the optimal test accuracy is 75.0%.

Algorithm	**Test accuracy**
ERM	73.0
ABD (with noise p_c)	58.0
ABD (without noise p_c)	73.0
Optimal	75.0

Table 2: Comparison of running times for Group DRO and ABD in the MultiNLI experiment using Nvidia RTX A6000.

Algorithm	**Running Time**
Group DRO	1,155 seconds
ABD	2,062 seconds

parameters from [8], we optimize with learning rates of 1×10^{-4} for both α and β, without utilizing L_2-regularization.

3 Additional Experimental Results

3.1 ABD with Unbiased Data

We also experiment with our algorithm in the setting without biases in the training data, using the vanilla MNIST. We evaluate ABD on the original MNIST, which does not contain any biases, and assess the model's performance on the test settings of the Colored MNIST task. Table 1 shows that even if the training data does not include spurious correlations so that the ABD cannot discover biases, the model can still achieve reasonable performance with our learning framework. However, when the training samples include label noise with a probability of $p_c = 0.25$ on the shape of the digits, the performance of the model is degraded. This occurs because the biased model may mistakenly interpret the noise as biases during the adaptive bias discovery stage. We plan to address these issues related to ABD in future work.

3.2 Compuatational Efficiency

We provide a comparison of the computationally demanding time for both Group DRO and ABD, focusing on the time taken for model training on MultiNLI

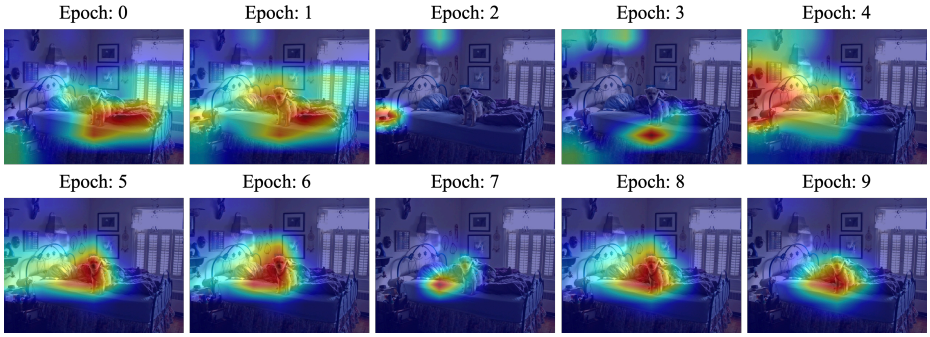

Fig. 1: GradCAM visualization of a biased model with our adaptive bias discovery method on the MetaShift test data sample for diverse learning steps.

experiment in Table 2. Here, we utilized Nvidia RTX A6000 for the experiments. While ABD requires more computational resources than Group DRO, it can achieve superior performance in handling bias features, as evidenced by our experimental results.

3.3 Additional GradCAM Visualizations

We provide additional GradCAM visualizations comparing ERM and ABD on MetaShift dataset. Figure 1 shows GradCAM results of a biased model in our ABD framework for diverse learning steps. It illustrates the evolving focus of the biased model within the ABD framework, highlighting how its attention shifts across different epochs. This shift in focus reflects the dynamic nature of ABD's adaptive bias discovery process, as the model progressively identifies and adjusts to various biases present in the dataset.

In addition, we evaluate trained model on MetaShift dataset with ERM and ABD on web-crawled images. Specifically, the model is trained on dog(boat+surfboard) scenario. Thus, we evaluate the model on web images with dog surfing. Figure 2 shows GradCAM visualizations of trained models with ERM and ABD. We observe that ERM-trained model tends to focus on background features while model trained with our method focuses on dog. This discrepancy make ERM-trained model be degraded on MetaShift test data since test data samples, dog(shelf) do not contain ocean background.

In addition, we evaluate models trained on the MetaShift dataset with ERM and ABD using web-crawled images. Specifically, the models are trained on the dog(boat+surfboard) scenario, and tested on web images depicting dogs surfing. Figure 2 presents the GradCAM visualizations of models trained with both ERM and ABD. We observe that the model trained with ERM tends to focus on background features, whereas the model trained with our method concentrates on the dog itself. This discrepancy leads to a degradation in the performance

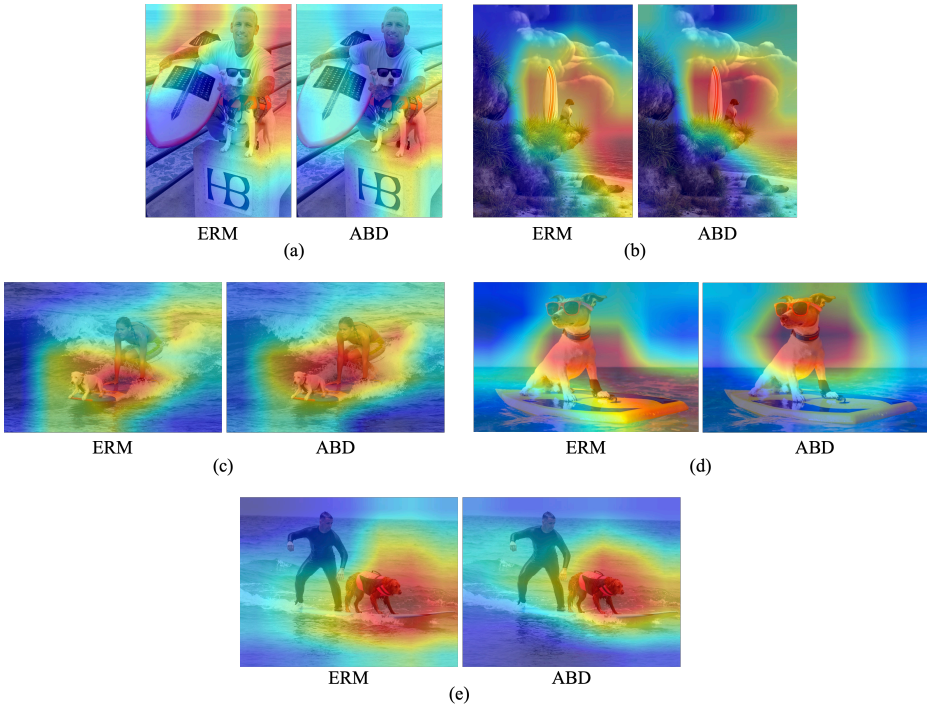

Fig. 2: GradCAM visualization of models trained with ERM and ABD on Web-Crawled images. Each model is trained on MetaShift dataset for dog(boat+surfboard), *i.e.*, distributional shifts score is 1.43. The web search keyword used was *Dog Surfing*. References for each image are provided in Section 3.3.

of the ERM-trained model on the MetaShift test data, the dog(shelf) samples, as they do not contain an ocean background. References for the web-crawled images are provided in the following section.

References for Web-Crawled Images in Figure 2 (a) https://ca-times.brightspotcdn.com/dims4/default/8c96fcf/2147483647/strip/true/crop/2396x3503+0+0/resize/1200x1754!/quality/75/?url=https%3A%2F%2Fcalifornia-times-brightspot.s3.amazonaws.com%2Feb%2F35%2F9b8969e145e68f5478a2d05db08b%2Fsugar-the-surf-dog-2.jpg

(b) https://cdn11.bigcommerce.com/s-zu7ew0vx99/images/stencil/1280x1280/products/18489/20225/dog-and-surfboard__11424.1620747142.jpg?c=1

(c) https://upload.wikimedia.org/wikipedia/commons/c/c3/Hanging_18.jpg

(d) https://www.surfertoday.com/images/stories/surfingdog.jpg

(e) https://blog.myollie.com/wp-content/uploads/2018/07/surf_dog-1469727158617-1.jpggettyimages-91910627-1024x1024.jpg,https://media.gettyimages.com/id/91910627/photo/man-with-surfboard-and-dog-with-stick-in-mouth-running-out-of-ocean.jpg?s=1024x1024&w=gi&k=20&c=ERXDurZsK3JoMqpWD0Ty7pVgdamNiNMzkKTGnpRxktM=

References

1. Bandi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., Bejnordi, B.E., Lee, B., Paeng, K., Zhong, A., et al.: From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. IEEE transactions on medical imaging **38**(2), 550–560 (2018) 4
2. Borkan, D., Dixon, L., Sorensen, J., Thain, N., Vasserman, L.: Nuanced metrics for measuring unintended bias with real data for text classification. In: Companion proceedings of the 2019 world wide web conference. pp. 491–500 (2019) 3
3. Christie, G., Fendley, N., Wilson, J., Mukherjee, R.: Functional map of the world. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6172–6180 (2018) 4
4. Devlin, J., Chang, M.W., Lee, K., 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 (2019) 3
5. Dixon, L., Li, J., Sorensen, J., Thain, N., Vasserman, L.: Measuring and mitigating unintended bias in text classification. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. pp. 67–73 (2018) 3
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 3
7. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017) 4
8. Koh, P.W., Sagawa, S., Xie, S.M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R.L., Gao, I., Lee, T., et al.: Wilds: A benchmark of in-the-wild distribution shifts. In: International Conference on Machine Learning. pp. 5637–5664. PMLR (2021) 3, 4, 5
9. Liang, W., Zou, J.: Metashift: A dataset of datasets for evaluating contextual distribution shifts and training conflicts. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=MTex8qKavoS 2
10. Park, J.H., Shin, J., Fung, P.: Reducing gender bias in abusive language detection. arXiv preprint arXiv:1808.07231 (2018) 3
11. Sagawa*, S., Koh*, P.W., Hashimoto, T.B., Liang, P.: Distributionally robust neural networks. In: International Conference on Learning Representations (2020), https://openreview.net/forum?id=ryxGuJrFvS 3