ALOUD: Active Learning of Unbiased Datasets

Anonymous CVPR submission

Paper ID 10073

Abstract

Large-scale datasets used for different tasks like classification and object recognition are often found to have strong correlation among various predicted attributes. When a model is trained on a such dataset, it exhibits likewise correlations. Despite this, the subject of many modern debiasing methods generally focus on strengthening models to ignore the inherent dataset biases between attributes, rather than building a more robust training dataset. This work proposes ALOUD, an Active Learning based method which aims to build a robust unbiased dataset for training models which are able to learn underrepresented groups in abundance in the dataset. As a method of Active Learning, ALOUD retrieves data from an unlabeled pool and selectively adds images to a seed dataset using Bias Sensitive Sampling, an alternative to classical acquisition functions, to enforce bias reduction. Experiments using ALOUD show comparable results to state-of-the-art methods in terms of unbiased model accuracy while operating on as little as 10.25% of the training set.

1. Introduction

Deep neural networks (DNNs) and large-scale image datasets have enabled great advances in computer vision over the last decade. It is now well established that increasing the size of training sets can significantly improve the predictive ability of DNNs, [37]. This, however, does not guarantee fairness in the distribution of the data, and modern datasets often contain many undesirable biases [31, 38]. As a result models trained on them make unfair decisions, where underrepresented groups are subject to predictive discrimination. Bias mitigation can be performed at several levels, as illustrated in Figure 1.

Most work has focused on avoiding model bias, i.e, training the model in a way that mitigates bias, as illustrated in Figure 1(a). Many measures of bias have been proposed [10, 15, 19–21] and algorithms designed to minimize bias, usually by regularizing training through the inclusion of these measures in loss functions [2, 20, 44, 45]. More recently, it has been shown that improved performance can be achieved with adversarial training [11, 28], by augmenting the network with a bias classifier trained adversarially, in a manner similar to GANs [14]. However, mitigating model bias usually boils down to downplaying groups for which there is a lot of training data (e.g. female kitchen scenes) and promoting groups for which data is scarce (e.g. male kitchen scenes). This hurts classifier generalization and degrades the overall recognition performance (poor recognition of kitchen scenes in general). Dataset bias has received less algorithmic attention. Work in this area is usually confined to exposing biases and recommending best practices for data collection [30, 41] such as the use of "dataset audit cards." While important, these are far from sufficient to guarantee that datasets are unbiased. There have also been some proposals to reweigh or resample datasets [19, 22] so as to eliminate or reduce the importance of images conducive to bias, as illustrated in Figure 1(b). These, however, again reduce the effective dataset size and degrade classification accuracy. In summary, algorithmic bias reduction procedures typically lead to degraded classifier per-

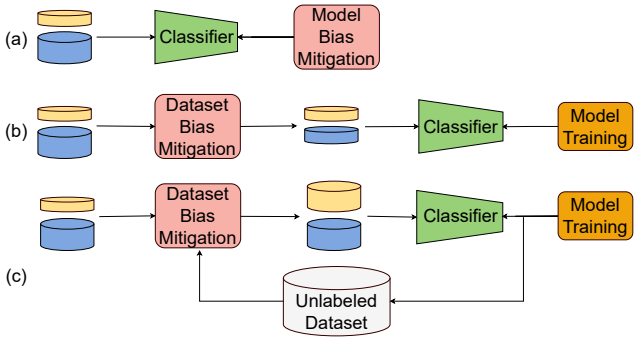


Figure 1. Bias mitigation strategies. a) Model bias mitigation techniques use loss functions to discourage bias at the mode output. b) Existing dataset bias mitigation techniques subsample the dataset to eliminate bias. Both of these approaches improve bias at the cost of classification accuracy. c) Proposed approach to jointly optimize the model and dataset, which enables bias mitigation without sacrifice of dataset samples or classification accuracy.

formance.

The problem is that simply penalizing subsets of data, either by dataset manipulation or loss-based constraints, *always* deteriorates generalization ability. In this work we explore the hypothesis that the solution to dataset induced biases is not to penalize the data that creates those biases, but to augment the training set with *more data that counterbalances them*. If the dataset is highly biased towards female kitchen scenes, the only possibility to mitigate bias without overfitting is to seek more examples of male kitchen scenes. This, however, implies that *dataset and model have to be optimized jointly.* We thus propose to jointly address the two sources of bias within a unified bias mitigation architecture. As illustrated in Figure 1(c), this architecture aims to train fair semantic classifiers via an iterative optimization with two components: 1) a dataset bias mitigation algorithm that identifies and downweighs biased examples and seeks additional examples in a large pool of data to counterbalance the associated biases, and 2) a model training procedure.

The proposed dataset bias mitigation architecture is inspired by active learning procedures [18], which propose examples to be labelled by a human oracle. Active learning algorithms vary in terms of the acquisition function used to sample these examples from a large unlabelled dataset. While many acquisition functions have been proposed [4, 13, 43], the goal is usually to choose the samples that most improve classifier performance with less labelling effort. A classical solution to this problem is to choose the samples of largest class uncertainty [3, 33]. This, however, does not account for bias. In this work, we introduce a novel *Bias Sensitive Sampling* (BSS) procedure that addresses this problem. Beyond increasing classification uncertainty, BSS seeks examples that also decrease dataset bias, as measured by the *absolute posterior bias* (APB) metric, and produce label balanced datasets. These objectives are combined into a dataset scoring function that relies on pseudo-labels produced by the model to determine the unlabelled samples to be added to the dataset, as it is expanded. Two versions of this scoring function, based on hard vs. soft pseudo-labels, are proposed and evaluated. Given the expanded dataset, the model is retrained and the process iterated.

Overall, the paper makes several contributions. First, we point out the need for joint dataset and model bias mitigation, through procedures that iteratively optimize model and dataset. Second, we propose an architecture to implement this goal, based on the novel BSS sampling procedure. Third, we explore different implementations of the dataset scoring function at the core of BSS, combining multiple objectives and different types of pseudo-labels. Finally, we present experiments demonstrating the importance of the different BSS components and showing that it outperforms existing bias mitigation approaches.

2. Related Work

Prior works on bias mitigation in machine learning typically fall into two categories. *Model debiasing* methods compensate for bias during training of classifier, while *dataset debiasing focuses on eliminating the distributional imbalance in the training samples*. These methods are explored less than the ones in model debiasing, reflecting the general trend that data quality is an underexplored but critical part of deep learning [29].

Model Bias. De-biasing the model is the process aiming to learn the correct feature. [5] introduces a corrective loss which can be added to any model to reduce the unwanted bias in the dataset. Similarly, Group-DRO allows avoiding bias overfitting by improving the worst-case performance over pre-defined subgroups [34]. Other model de-biasing methods such as model ensembles [8], and statistical regularizations [6] focus on training a robust model to mitigate the undesirable effect with the bias in the dataset.

Dataset Bias. The mismatched distribution between the dataset and the reality is defined as dataset bias [40]. The issue of the shift between distributions can be tackled by Domain Adaptation (DA) techniques [12]. Resampling is a common technique that oversamples minority and undersamples majority to train a fairer model [7]. More recently, REPAIR introduced a new resampling strategy to decrease the representational bias and focus on solving the dataset [23].

Active Learning. Our method resembles the Active Learning loop to a great extent. Therefore, we provide a brief review of relevant Active Learning literature.

Active Learning iteratively selects the most informative samples to benefit the model and to reduce the human effort in labeling the data. One classic approach to measure the uncertainty of data is by using the posterior probabilities of the predicted class [18] that is based on the entropy of the class [36]. [43] proposes the learning loss module that predicts the loss of unlabeled data points based on the uncertainty method. More recently, ALOFT [1] selects the subset of the dataset to mitigate the contextual bias in the dataset explicitly and applies to object detection and multi-label classification.

3. Iterative Dataset Collection

3.1. Overview

The methods for Model Debiasing and Dataset Debiasing are still relatively disjoint. The defining motivation of

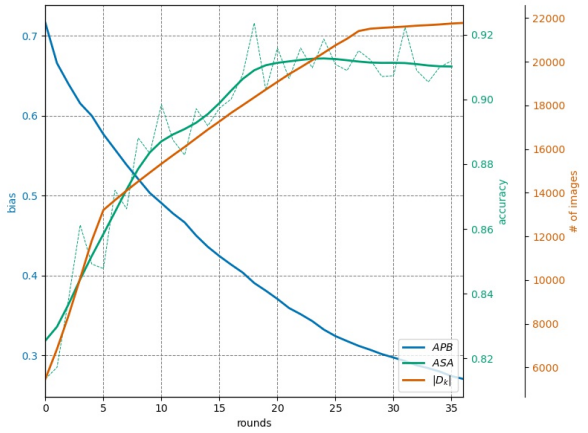


Figure 2. Absolute posterior bias(APB), Average Subgroup Accuracy(ASA), Dataset size ($|D_k|$) of the training dataset

our method is that there is a symbiotic relationship between Model Debiasing and Dataset Debiasing: performing AL on a debiased model will allow one to select critical images for labeling, which in turn, will create an even more debiased model, etc.

The abstract loop underlying our method is extremely simple as illustrated in Figure 1, and closely resembles an active learning loop. We take this section to introduce and formalize each of the components.

3.2. Problem Formulation

We consider a classifier of images $\mathbf{X}$ that predicts both a binary class label, $Y \in \{0, 1\}$, and K binary attributes $\mathbf{Z} = (Z_1, \ldots, Z_K)$, where $Z_k \in \{0, 1\}$. These predictions are produced by functions $y = f(\mathbf{x}), z_i = h_i(\mathbf{x})$, implemented by a neural network with $K + 1$ sigmoidal outputs. Given a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i, \mathbf{z}_i)\}_{i=1}^n$, where y_i is the class label of example $\mathbf{x}_i$ and $\mathbf{z}_i$ the associated vector of attribute values, the network is trained to minimize the risk

$$\mathcal{R}(f, \mathbf{h}, \mathcal{D}) = \sum_i \left\{ L[f(\mathbf{x}_i), \mathbf{y}_i] + \lambda \sum_k L[h_k(\mathbf{x}_i), z_{i,k}] \right\} \tag{1}$$

defined by the binary cross-entropy loss function $L[f((x)), y] = y \log f(\mathbf{x}) + (1 - y) \log(1 - f(\mathbf{x}))$.

In practice, datasets are often collected with *bias*, where certain groups of the population are over- or under-represented compared to the ground-truth distribution. We consider the setting where groups $\mathbf{g}$ are defined by the binary label y and a *binary protected attribute* s, i.e. $\mathbf{g} = (y, s) \in \mathcal{G} = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$. However, all algorithms can be generalized to settings containing multiple target and protected attributes. In what follows, we denote the protected attribute by random variable S, even though it is usually one of the attributes Z_i introduced above . We use either notation as convenient, e.g. do not rewrite (1) to explicitly denote the dependence on S.

3.3. ALOUD

ALOUD aims to assemble an unbiased dataset $\mathcal{D}$ of size N, from a larger unlabeled dataset $\mathcal{U} = \{\mathbf{u}_1, \ldots, \mathbf{u}_W\}$, where $W \gg N$. This is formulated as the search for the dataset which, when used to train f, achieves the maximum unbiased accuracy on a previously designated hold-out set.

Specifically, let $(\mathcal{X}_{train}, \mathcal{X}_{val}, \mathcal{X}_{test})$ be a dataset, $\mathcal{D}_0 \subset \mathcal{X}_{train}$ be an initial seed dataset, and $\mathcal{L}_u$ be a measure of dataset bias. On each round k, **ALOUD** seeks to find a dataset $\mathcal{D}_k$ such that:

$$\mathcal{D}_k = \underset{\mathcal{D} \subset \mathcal{U} \setminus \mathcal{D}_{k-1}}{\arg\max} \; \mathcal{L}_u(\mathcal{D}; f_{\hat{\theta}_k}) \tag{2}$$

$$\text{where} \quad \hat{\theta}_k = \underset{\theta}{\arg\min}\, \mathcal{R}(f_\theta; \mathcal{D}_{k-1}) \tag{3}$$

where (3), with $\mathcal{R}$ as described by (1), is standard empirical risk minimization (ERM) with binary cross-entropy loss. Finally, given some notion of model bias $\mathcal{M}_u$, one obtains a final unbiased dataset $\mathcal{D}$, an unbiased classifier f, and a final unbiased classification accuracy δ from:

$$\mathcal{D} = \mathcal{D}_M, \; f = f_{\hat{\theta}_M}, \; \delta = \mathcal{M}_u(f_{\hat{\theta}_M}; \mathcal{X}_{test}) \tag{4}$$

$$\text{where} \quad M = \underset{k=1,2,\ldots}{\arg\max}\, \mathcal{M}_u(f_{\hat{\theta}_k}; \mathcal{X}_{val}) \tag{5}$$

Algorithm 1: Dataset Collection with **ALOUD**

Input: N: max iterations, $\mathcal{U}$: unlabeled data pool; $\mathcal{D}_0$: seed dataset, B:Annotation budget

for $k = 1, ..., N$ **do**
- Train model f_k on $\mathcal{D}_{k-1}$, using (3);
- Create pseudo-labeled dataset $\mathcal{P}$ from $\mathcal{U}$, using (6);
- $\mathcal{C}_b \leftarrow BSS_1(\mathcal{D}_{k-1}, \mathcal{P}, B, L_{CB})$;
- $\mathcal{D}_c \leftarrow h(\mathcal{C}_b)$;
- $\mathcal{D}_k \leftarrow BSS_2(\mathcal{D}_{k-1}, \mathcal{D}_c, \infty, L_B)$

end

return D_n, f_n

The steps of (2) and (3) are performed in an alternating fashion, resulting in a process that iterates between updating the dataset $\mathcal{D}$ and model parameters θ, starting from the seed dataset $\mathcal{D}_0$. These operations are summarized by Algorithm 1. At iteration k, the algorithm starts by training the model $(f_k, \mathbf{h}_k)$ on the current dataset $\mathcal{D}_{k-1}$, using (3). It then uses this model to produce class and attribute pseudo-labels:

$$\hat{y}_i = \mathbb{1}(f_k(\mathbf{x}_i) \geq 0.5), \tag{6}$$

$$\hat{z}_{i,k} = \mathbb{1}(h_{i,k}(\mathbf{x}_i) \geq 0.5), k \in \{1, \ldots, K\} \tag{7}$$

for each unlabeled example $\mathbf{x}_i \in \mathcal{U}$, creating a *pseudo-labeled* dataset $\mathcal{P} = \{(\mathbf{x}_i, \hat{y}_i, \hat{\mathbf{z}}_i)\}_{i=1}^{M}$, which is used to seek the best samples to augment $\mathcal{D}_k$. This is performed with a *bias sensitive sampling* (BSS) procedure, which takes in $\mathcal{D}_k$, $\mathcal{P}$, and an annotation budget B, and samples a candidate set $\mathcal{C}_b \subset \mathcal{P}$ of size B according to a dataset scoring function L_{CB}. Samples in $\mathcal{C}_b$ are then labeled by the human oracle h to form a candidate dataset $\mathcal{D}_c$, which is finally subject to the BSS procedure again to eliminate samples that do not contribute to a lower dataset score under a scoring function L_B.

Algorithm 2: Bias Sensitive Sampling (BSS)

Input: $\mathcal{D}$: existing dataset; $\mathcal{C}$: candidate dataset; B: Annotation budget; L: dataset scorer;

$b_{current} \leftarrow L(\mathcal{D})$;
for $d \in \mathcal{U} \backslash \mathcal{D}_k$ **do**
  $b_{new} \leftarrow L(\mathcal{D}_{k-1} \cup d)$;
  if $b_{new} < b_{current}$ **then**
    $\mathcal{D}_k \leftarrow \mathcal{D}_{k-1} \cup d$
    $b_{current} \leftarrow b_{new}$
  end
  if $|\mathcal{D}_k| \geq B$ **then**
    break
  end
end
return $\mathcal{D}_k$

The BSS procedure is implemented by Algorithm 2. It samples a subset of size B from a set $\mathcal{C}$ of candidate examples that, when added to an existing dataset $\mathcal{D}$ reduce the risk of the latter, according to the scoring function L. The procedure is used twice per iteration of **ALOUD**. In the first time (BSS_1), the goal is to identify samples $\mathcal{C}_b$ in $\mathcal{P}$ that are promising for labeling. To meet a labeling budget, $\mathcal{C}_b$ is limited to size B. Unlike existing AL methods, we employ the process a second time (BSS_2) after the samples are annotated. This is necessary because the pseudo-labels of $\mathcal{P}$ can be incorrect, leading to new samples that increase the bias. By applying the BSS procedure again to the human annotated dataset $\mathcal{D}_c$, **ALOUD** eliminates those samples that no longer contribute to a low dataset bias after the label correction.

A final, subtle, difference between the two uses of BSS is the dataset scoring function employed. In the first use of BSS, where the goal is to identify good labels to sample, this function considers a combination of classification accuracy and dataset bias. This is because good examples to add to the dataset should be both challenging to classify, so as increase classification accuracy, and bias mitigating, to guarantee a dataset without bias. However, after labeling, there is no classification benefit in eliminating samples. Hence, the second use of BSS uses a scoring function that only considers bias. We next discuss the dataset scoring functions in more detail.

4. Dataset Bias Mitigation

In this section, we discuss the scoring functions L used in the BSS procedure.

4.1. Absolute Posterior Bias

A target prediction y is unbiased with respect to a protected attribute s if the predicted value of the protected attribute has no bearing on the prediction of the target attribute. This is captured by the *Absolute Posterior Bias* (APB) metric

$$\mu_{Bias}(Y, S) = |P_{Y|S}(y \mid 1) - P_{Y|S}(y \mid 0)|, \tag{8}$$

whose minimization forces independence of the target and bias predictions. This follows from the fact that $\mu_{Bias}(Y, S) \geq 0$ with equality if and only if

$$P_{Y|S}(y \mid 1) = P_{Y|S}(y \mid 0) \tag{9}$$

from which it follows that

$$\begin{aligned} P_Y(y) &= P_{Y|S}(y \mid 1)P_S(1) + P_{Y|S}(y \mid 0)P_S(0) & (10)\\ &= P_{Y|S}(y \mid s). & (11)\end{aligned}$$

Furthermore, for a binary target Y,

$$\begin{aligned} &|P_{Y|S}(1 \mid 1) - P_{Y|S}(1 \mid 0)| = \\ &= |1 - P_{Y|S}(1 \mid 1) - (1 - P_{Y|S}(0 \mid 0))| \\ &= |P_{Y|S}(1 \mid 1) - P_{Y|S}(0 \mid 0))| \end{aligned}$$

and the APB reduces to

$$\mu_{Bias}(Y, S) = |P_{Y|S}(1 \mid 1) - P_{Y|S}(1 \mid 0)|. \tag{12}$$

Given a dataset $\mathcal{D}$, the probabilities $P_{Y|S}(y \mid s)$ can be estimated empirically, using

$$\pi_{y|s} = \hat{P}_{Y|S}(y \mid s) = \frac{|\{(\mathbf{x}_i, y_i, s_i) \in \mathcal{D} | y_i = y, s_i = s)\}|}{|\{(\mathbf{x}_i, y_i, s_i) \in \mathcal{D} | s_i = s)\}|} \tag{13}$$

and APB measured with

$$\mu_{Bias}(\mathcal{D}) = |\pi_{1|1} - \pi_{1|0}|. \tag{14}$$

4.2. Label Balancing

Unbalanced datasets create difficulties to learning, since the learning algorithm tends to focus on highest populated groups and ignores groups with few examples [24–26].

Hence, the dataset $\mathcal{D}$ should ideally be balanced, in the sense that

$$P_Y(y) = P_S(s) = 0.5, \forall y, s \in \{0, 1\} \quad (15)$$

While minimizing the APB brings together the probabilities $P_{Y|S}(1|1)$ and $P_{Y|S}(1|0)$, this does not guarantee that (15) holds. In general, the simple minimization of (14) can originate highly unbalanced datasets. To avoid this problem, we propose two additional label-balancing metrics

$$\mu_{TB}(Y) = |P_Y(1) - 0.5|, \qquad \mu_{BB}(S) = |P_S(1) - 0.5|$$

implemented by the empirical estimates

$$\mu_{TB}(\mathcal{D}) = |\rho_1 - 0.5|, \quad \mu_{BB}(\mathcal{D}) = |\gamma_1 - 0.5| \quad (16)$$

where

$$\rho_y = \hat{P}_Y(y) = \frac{|\{(\mathbf{x}_i, y_i, s_i) \in \mathcal{D} | y_i = y)\}|}{|\{(\mathbf{x}_i, y_i, s_i) \in \mathcal{D})\}|} \quad (17)$$

$$\gamma_s = \hat{P}_S(s) = \frac{|\{(\mathbf{x}_i, y_i, s_i) \in \mathcal{D} | s_i = s)\}|}{|\{(\mathbf{x}_i, y_i, s_i) \in \mathcal{D})\}|}. \quad (18)$$

4.3. Hard vs. Soft Pseudo-Labels

So far, we have used the empirical probability estimates of (13), (17) and (18). These assume knowledge of the ground truth labels y_i and s_i of all examples. However, in the BSS sampling step, only the pseudo-labels of (6) are available. These can be incorrect, especially when $f_k(\mathbf{x}_i)$ is close to 0.5. An alternative is to use soft labels, i.e. the probability estimates $f_k(\mathbf{x}_i)$ output by the model. For this, we use the fact that

the soft estimate

$$\pi^{soft}_{y|s} \approx \frac{\sum_i P_{Y|X}(y_i|x_i) P_{S|X}(s|x_i)}{\sum_i P_{S|X}(s|x_i)} = \frac{\sum_i f(x_i)^{y_i}(1 - f(x_i))^{1-y_i} h(x_i)^s (1 - h(x_i))^{1-s}}{\sum_i h(x_i)^s (1 - h(x_i))^{1-s}}$$

from which the empirical estimate of the APB of (14) can alternatively be written as

$$\mu^{soft}_{Bias}(\mathcal{D}) = |\pi_{1|1} - \pi_{1|0}| \quad (19)$$
$$= \left| \frac{\sum_i f(x_i) h(x_i)}{\sum_i h(x_i)} - \frac{\sum_i f(x_i)(1 - h(x_i))}{\sum_i (1 - h(x_i))} \right|$$

Similarly,

$$\mu^{soft}_{TB}(\mathcal{D}) = \left| \frac{1}{|\mathcal{D}|} \sum_i f(x_i) - 0.5 \right| \quad (20)$$

$$\mu^{soft}_{BB}(\mathcal{D}) = \left| \frac{1}{|\mathcal{D}|} \sum_i s(x_i) - 0.5 \right| \quad (21)$$

4.4. Scoring functions

Two scoring functions are used in **ALOUD**, for the two BSS sampling operations of Algorithm 1. The first sampling operation aims to identify good samples in the large unlabelled dataset $\mathcal{U}$ to add to the labeled dataset $\mathcal{D}$. This involves two considerations. First, the samples should be challenging to classify, since this leads to the classifier of best generalization. The classification uncertainty of dataset a $\mathcal{D}$ is computed with the Shannon entropy of the classifier predictions

$$\mu_{UR}(\mathcal{D}; f) = -\frac{1}{|\mathcal{D}|} \sum_i f(\mathbf{x}_i) \log f(\mathbf{x}_i) \quad (22)$$

The scoring functions then penalizes a combination of lack of uncertainty, bias, and class unbalance. Specifically, we define the scoring function

$$L_{CB}(\mathcal{D}) = \mu_{Bias}(\mathcal{D}) + \alpha \mu_{BB}(\mathcal{D}) + \beta \mu_{TB}(\mathcal{D}) - \zeta \mu_{UR}(\mathcal{D}) \quad (23)$$

Above, $\mu_{Bias}, \mu_{BB}, \mu_{TB}, \mu_{UR}$ are measures of dataset bias and α, β, ζ are hyperparameters. The main contributions of this work are these metrics, which account for various biases in the dataset, building up to BSS_1 and BSS_2. Later, we will discuss these two use of BSS in more detail.

This can be seen as a generalization of active learning methods that use the uncertainty measure as an acquisition function for examples to label [18], which BSS reverts to when ζ is large. The multipliers α, β and ζ control the importance of the different objectives.

The second use of *BSS* (BSS_2) in **ALOUD** aims to eliminate samples that, after annotation by the human oracle, no longer contribute to small dataset bias. This can happen because the sample selection of the candidate set $\mathcal{C}_b$ in BSS_1 is based on model predictions that can be incorrect. However, this sampling is only a filtering operation, in the sense that no new samples are added to the dataset. Since eliminating training samples never has a benefit from a classification point of view, samples are not penalized by lack of uncertainty, only for introducing bias. Hence the scoring function only penalizes bias and class imbalance according to

$$L_C(\mathcal{D}) = \mu_{Bias}(\mathcal{D}) + \alpha \mu_{BB}(\mathcal{D}) + \beta \mu_{TB}(\mathcal{D}). \quad (24)$$

Both scoring functions can be implemented with both soft or hard measures μ_{Bias}, μ_{BB}, and μ_{TB}.

5. Experiments

5.1. Experimental Setup

Datasets CelebA [27] is a large-scale face attribute dataset with more than 200K celebrity images, each with 40

Soft PL	Label Balancing Bias μ_{BB}	Label Balancing Target μ_{TB}	Uncertainty μ_{UR}	$\mathcal{M}_u$
				87.47
✓				88.22
✓	✓			87.29
✓	✓		✓	89.60
✓		✓		90.59
✓		✓	✓	**92.22**

(a) Scoring function L_{CB} *pre*-annotation.

Soft PL	Label Balancing Bias μ_{BB}	Label Balancing Target μ_{TB}	$\mathcal{M}_u$
	None—Skip BSS_2		81.09
			83.10
✓			84.28
✓	✓		86.31
✓		✓	**92.22**

(b) Scoring function L_C *post*-annotation.

Table 1. Ablation Studies on **ALOUD**. We compare different scoring functions for bias sensitive sampling before and after data annotation.

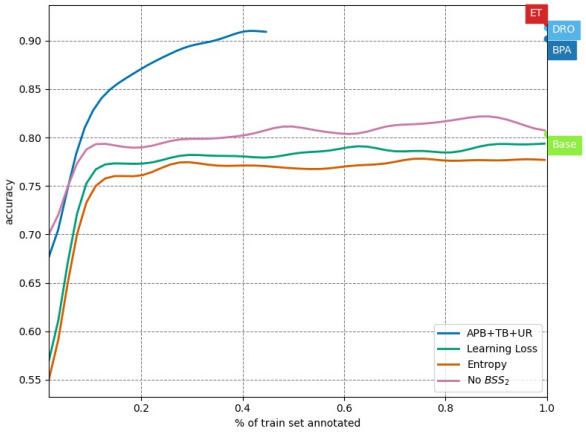


Figure 3. Performance in terms of % annotated

attribute annotations. It is well-used throughout debiasing literature [16,32,35], particularly for its variety of poses and richness of annotations. Of the 40 CelebA dataset attributes, 13 are universally considered to be subjectively unambiguous target labels [39] which are listed in Table 3. For all experiments, the *Gender* attribute will be taken as the protected attribute. The CelebA training set is comprised of 162770 images with 40 binary attributes, and experiments (with the exception of Seed dataset ablation) begin with 5000 randomly chosen images from the training set.

Waterbirds is a dataset constructed from the Caltech-UCSD Birds-200-2011 (CUB) [42] by sampling various background images in the Place dataset and combining them with images of birds from the CUB dataset. The labels of Waterbirds are constructed by taking two attributes, *Object* $\in$ {Waterbird, Landbird} and *Place* $\in$ {Water, Land} which represent the type of bird and background as seen from the CUB and Place datasets, represetively.

Evaluation Following from the problem formulation of (2) and (3), the *unbiased* performance of f is measured by the group average accuracy

$$\mathcal{M}_u(\mathcal{D}) = \frac{1}{|\mathcal{G}|} \sum_{\mathbf{g} \in \mathcal{G}} A_{\mathbf{g}}(f; \mathcal{D}), \tag{25}$$

where

$$A_{\mathbf{g}}(f; \mathcal{D}) = \frac{1}{|\mathcal{S}_{\mathbf{g}}|} \sum_{i|(y_i, s_i)=\mathbf{g}} \mathbb{1}_{y_i}(f(\mathbf{x_i}) \geq .5) \tag{26}$$

is the accuracy for examples of group $\mathbf{g}$.

Classifier For model architecture, all experiments use ResNet-18 pre-trained on ImageNet [9]. The final layer of the ResNet-18 is a $Linear(512, 2)$ with sigmoid activation, so that the target and bias labels are trained on single model as a multi-label prediction problem. For optimization, SGD with learning rate .05, weight decay 10^{-4}, and batch size 64. All models are trained with 40 epochs and the test accuracies reported correspond to the epoch with the maximum validation accuracy.

BSS On CelebA dataset, we used hyperparameters $\{\alpha = 0, \beta = 0.7, \zeta = 0.7\}$ for L_{CB} in BSS_1, and $\{\alpha = 0, \beta = 0.7\}$ for L_C in BSS_2. On waterbirds dataset, L_CB in BSS_1 has the set of hyperparameters $\{\alpha = 0.05, \beta = 0.7, \zeta = 0.03\}$, and, L_C in BSS_2 is $\{\alpha = 0.05, \beta = 0.7\}$ For our methods APB+TB or APB+TB+UR, Table 2, 3, 4, BSS_2 are L_C above with those hyperparameters. APB+TB refers to the L_{CB} above with those hyperparameters but $\zeta = 0$. APB+TB+UR refers to the L_{CB} above with those hyperparameters.

5.2. Baselines

We compare our method with two kinds of debiasing strategy, model debiasing and unbiased dataset. Due to the similarity between our method and AL, we choose learning loss for active learning as an additional compared method.

Method	CelebA	Waterbirds
APB+TB+UR	9.83%	32.08%
APB+TB	8.60%	23.11%

Table 2. % of Training Set Used

Model Debiasing As the first part of baseline, debiasing the model efficiently decreases the bias, while keeping the datasize the same. DRO [34] is a model-based optimization technique used to train a de-biased model. Specifically, it minimizes the worst-case loss over an uncertainty distribution. BPA [35] is a unsupervised debiasing technique which tackles the limitation when human annotation is impractical.

Dataset Debiasing The second part of our baseline is the subgroup balanced dataset which is a subset of the entire dataset in which each subgroup has the same amount of samples. Such a dataset has zero bias, but limited training samples that could potentially lead to poor generalization. The even training (ET) baseline for CelebA is compared, with the other baselines, in Table 3.

Active Learning AL method selects the most informative data points and combines them with the labeled pool for retraining. AL requires less human labeling, and can achieve a relatively good accuracy, but it does not consider bias. We pick learning loss [43] as an additional baseline, and the standard entropy-based [18] is considered in next ablative studies section.

5.3. Ablative Studies

Since the method as defined in Sections 3 and 4 can take on a variety of forms for different choices in components, a brief ablation is performed to analyze each component of ALOUD and to determine the best performing method.

All of the ablation results use *Blond Hair* as the target attribute and *Gender* as the protected attribute. Although ablating on one attribute pair limits the interpretative power of the results, certain findings from the ablations can easily be extrapolated to attribute pairs sharing common properties. More exploratory data analysis of CelebA is provided in the Supplementary Materials.

BSS$_1$ As a new addition to the AL loop, ablations on BSS_1, located in Table 1a, are comprised solely of bias metrics devised in Section 4. In results, $L_{CB} = \mu_{bias}^{soft} + 0.7\mu_{TB}^{soft} + 0.7\mu_{UR}^{soft}$ is the best performing debiasing method for BSS_1.

BSS$_2$ BSS_2 is introduced with the sole purpose of eliminating the samples, $\mathbf{x}_i$, with mismatched pseudolabels, y_i, and Ground Truth labels, $\mathbf{h}_i(\mathbf{x}_i)$ adding which will rather increase the current bias of the dataset. In contrast to the ablation study of BSS_1, there is a strong correlation with BSS_2 as seen by the large effect from each method. Out of all the methods detailed in Table 1b, $L_C = \mu_{bias}^{soft} + 0.7\mu_{TB}^{soft}$ has the best performance. The reason may rely on the makeup samples for the worst subgroup.

5.4. Results

CelebA The CelebA dataset has 182637 images, and 40 attributes. Data analysis begins with fixing the *protected* attribute to be $Gender$ and varying the *target* attribute over the remaining 39 attributes. However, exploratory data analysis [35] of the CelebA dataset found that 26 of the 39 *target* attributes held spurious correlations, meaning that the gap between worst subgroup accuracy and average subgroup accuracy were greater than 5%. The remaining 13 attributes are shown in Table 3.

There are 3 categories of baselines under consideration: (1) *Model Debaising* methods, ie. DRO and BPA. (2) *Active Learning* procedures, namely Learning Loss, and (3) *Dataset Debiasing* methods, of which Even Subset Training, where the training set $\mathcal{D} \subset \mathcal{X}_{train}$ is balanced with respect to each subgroup, is considered. Additionally, *Base* refers to using the full train split $\mathcal{X}_{train}$ as described in (1). Note that all of these methods are judged by a *model bias* metric, $\mathcal{M}_u$, as laid out in (25).

As shown in Table 5, **ALOUD** achieves comparable or greater performance to all baseline methods, despite on average only training on $\sim$10.25% of the data, as seen in Table 2. This suggests the existence of ideal, unbiased subsets of the CelebA dataset given a chosen pairs of attributes. Note that this is with respect to the model bias metric $\mathcal{M}_u$, as given in (25), and that the effect on overall classifier performance is non-obvious and explored in the supplementary materials.

Out of the methods which operate on the complete training set of CelebA, Even Training (ET), is the most performant. Note that, in this paradigm, in order to construct an evenly subsampled training set, one must already have all images labeled. This is contrary to *blind* methods which treat the random seed dataset *as the only labeled data* for a task. Active Learning methods, by this definition, are blind as they seek to use an annotator to build up the dataset. This distinction is denoted by the double vertical bar splitting Table 3 and 4 into two halves where the left half columns are aware of all the labels of $\mathcal{X}_{train}$, whereas the right half columns are not.

Mostly likely, the reason why $\mu_{Bias} + \mu_{TB}$ contributes so strongly in comparison to $\mu_{Bias} + \mu_{BB}$ is due to the subgroup proportions of the *Blond/Gender* attribute pair. In

Target	Base	DRO	BPA	ET	Learning Loss	APB+TB		APT+TB+UR	
	%	%	%	%	%	%	#	%	#
Blond Hair	80.42	91.39	90.18	91.82	80.44	91.73	21k	92.22	24k
Heavy Makeup	71.19	72.7	73.78	71.92	68.18	74.24	15k	72.94	17k
Pale Skin	71.5	90.55	90.06	91.11	73.64	74.46	6k	78.70	8k
Wearing Lipstick	73.9	78.26	78.28	83.31	68.09	82.46	16k	82.15	19k
Young	78.19	82.4	82.27	84.56	76.87	79.75	7k	79.59	8k
Double Chin	64.61	83.19	82.92	84.80	67.46	67.17	10k	83.76	20k
Chubby	67.42	81.9	83.88	83.82	68.47	67.18	10k	67.66	12k
Wearing Hat	93.53	96.84	96.8	97.86	84.12	92.95	7k	92.06	8k
Pointy Nose	62.1	70.71	68.98	71.64	64.26	65.62	7k	66.52	8k
Arched Eyebrows	69.72	78.3	74.77	80.42	70.18	80.56	20k	80.40	22k
No Beard	73.11	77.86	79.58	68.29	75.00	80.27	30k	80.19	32k
Wavy Hair	73.1	79.65	79.89	81.59	73.71	84.24	15k	78.06	9k
Wearing Earrings	72.17	83.5	84.57	86.06	72.80	84.99	20k	88.69	22k
Average	73.15	82.09	81.99	82.86	72.56	78.88	14k	80.24	16k

Table 3. Average Subgroup Accuracy (%), # of images (CelebA)

Target	Bias	Base	DRO	BPA	ET	APB+TB		APB+TB+UR	
		%	%	%	%	%	#	%	#
Object	Place	84.63	88.99	87.05	83.42	87.32	1094	88.24	1558
Place	Object	87.99	89.20	88.44	92.84	92.80	1123	93.52	1518
Average		84.63	89.10	87.75	88.13	90.06	1108	90.88	1538

Table 4. Average Subgroup Accuracy (%), # of images (Waterbirds)

the training of CelebA, the proportion of samples which take on the *Blond* and *Gender* attributes are 58% and 4.3%, respectively. Note that the relative importance between $\mu_{Bias} + \mu_{TB}$ and $\mu_{Bias} + \mu_{BB}$ the depends, for each attribute pair, on the proportion of images in each subgroup.

$\mu_{Bias} + \mu_{TB}$ tends to settle into small datasets with low bias, which halt taking data that could decrease future bias. There is a inherent trade-off between dataset size and dataset bias, showing that introducing μ_{UR} dismantles this trade-off by introducing another objective of improving classifier generalization. For most attributes listed in Table 3, our method shows higher average subgroup accuracy compared with AL and model debiasing methods; besides, our method has considerable large dataset compared with dataset debiasing method, even traing. These findings show that our method has a better trade-off between the bias and the dataset size.

Waterbirds The main results for **ALOUD** on the Waterbirds dataset are shown in Table 4. Similarly to the CelebA, **ALOUD** performs comparably or better than all baseline methods while using ∼32% of data on average. These results are noteworthy since Waterbirds is much smaller than CelebA, as its train set is only comprised of 4795 images. This shows that the method holds even in low data environments.

5.5. Comparison with other methods

The framework **ALOUD** has been built after taking inspiration from several active learning setups by researchers so far. It is similar to **ALOFT** [1] in a manner that it uses the same concept of alternate update of the training dataset and the model but it differs in the use of BSS_1 and BSS_2 in place of minimization of c_v i.e. coefficient of variation between co-occurring features for every single data in case of **ALOFT**.

Another method **ALOUD** has some similarity with is **LPDSSL** (Label propagation for deep semi-supervised learning) [17]. *Firstly*, the former minimizes the cross-entropy loss function with initial seed dataset to predict pseudo-labels of the subset of unlabelled pool after acquisition whereas the latter does the same by utilizing nearest neighbour graph and conjugate gradient (CG) method in label propagation. *Secondly*, to mitigate bias, our method uses scoring functions, L_{CB} in BSS_1 and L_B in BSS_2 whereas **LPDSSL** minimizes weighted loss which is a sum

of weighted cross-entropy loss and pseudo-label loss.

6. Conclusion

Real datasets used for training are usually biased. Unlike most literary works on working on making robust models to mitigate bias, we proposed to build an efficient unbiased training dataset with a target attribute and a protected attribute in an active learning setup. Experiments have been performed with different combinations of target and protected attributes to test the robustness of our algorithm. It was observed that training dataset $\mathcal{G}$ became iteratively better as model f_k selected more images of underrepresented subgroup in each iteration thereby reducing bias starting from a random biased seed dataset D_0. Amongst all the variations of our algorithm we have tested, APB+TB+UR provided best accuracy results on an unbiased test dataset over datasets of different domains.

As a future extension of our work, we would like to build the iterative training dataset from scratch rather than from a random biased seed dataset. Images would get added to initially increase, then decrease and finally stabilize the dataset with the minimum bias possible while simultaneously checking performance on an unbiased test dataset.

References

[1] Sharat Agarwal, Sumanyu Muku, Saket Anand, and Chetan Arora. Does data repair lead to fair models? curating contextually fair data to reduce model bias. *CoRR*, abs/2110.10389, 2021. 2, 8

[2] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision (ECCV)*, pages 771–787, 2018. 1

[3] William H. Beluch, Tim Genewein, Andreas Nurnberger, and Jan M. Kohler. The power of ensembles for active learning in image classification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9368–9377, 2018. 2

[4] Erdem Biyik, Kenneth Wang, Nima Anari, and Dorsa Sadigh. Batch active learning using determinantal point processes. *ArXiv*, abs/1906.07975, 2019. 2

[5] Kaylee Burns, Lisa Anne Hendricks, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models, 2018. 2

[6] Rwiddhi Chakraborty and Shubhayu Das. Rc2020 report: Learning de-biased representations with biased representations, 2021. 2

[7] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority oversampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002. 2

[8] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases, 2019. 2

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[10] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness through awareness. *CoRR*, abs/1104.3913, 2011. 1

[11] Harrison Edwards and Amos Storkey. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*, 2015. 1

[12] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *2013 IEEE International Conference on Computer Vision*, pages 2960–2967, 2013. 2

[13] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep Bayesian active learning with image data. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1183–1192. PMLR, 06–11 Aug 2017. 2

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2672–2680, 2014. 1

[15] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3315–3323, 2016. 1

[16] Youngkyu Hong and Eunho Yang. Unbiased classification through bias-contrastive and bias-balanced learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 26449–26461. Curran Associates, Inc., 2021. 6

[17] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5070–5079, 2019. 8

[18] Ajay J. Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2372–2379, 2009. 2, 5, 7

[19] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012. 1

[20] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer, 2012. 1

[21] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems (NIPS)*, pages 4066–4076, 2017. 1

[22] Yingwei Li, Yi Li, and Nuno Vasconcelos. RESOUND: Towards action recognition without representation bias. In *Eu-*

ropean Conference on Computer Vision (ECCV), pages 513–528, 2018. 1

[23] Yi Li and Nuno Vasconcelos. Repair: Removing representation bias by dataset resampling, 2019. 2

[24] Bo Liu, Haoxiang Li, Hao Kang, Gang Hua, and Nuno Vasconcelos. Breadcrumbs: Adversarial class-balanced sampling for long-tailed recognition, 2021. 4

[25] Bo Liu, Haoxiang Li, Hao Kang, Gang Hua, and Nuno Vasconcelos. Gistnet: a geometric structure transfer network for long-tailed recognition, 2021. 4

[26] Bo Liu, Haoxiang Li, Hao Kang, Nuno Vasconcelos, and Gang Hua. Semi-supervised long-tailed recognition using alternate sampling, 2021. 4

[27] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, Los Alamitos, CA, USA, dec 2015. IEEE Computer Society. 5

[28] David Madras, Elliot Creager, Toniann Pitassi, and Richard S. Zemel. Learning adversarially fair and transferable representations. In Jennifer G. Dy and Andreas Krause, editors, *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*, pages 3381–3390. PMLR, 2018. 1

[29] Andrei Paleyes, Raoul-Gabriel Urma, and Neil D. Lawrence. Challenges in deploying machine learning: a survey of case studies. *ACM Computing Surveys*, apr 2022. 2

[30] Vinay Uday Prabhu and Abeba Birhane. Large image datasets: A pyrrhic win for computer vision? *arXiv preprint arXiv:2006.16923*, 2020. 1

[31] Vikram V. Ramaswamy, Sunnie S. Y. Kim, and Olga Russakovsky. Fair Attribute Classification through Latent Space De-biasing. *arXiv e-prints*, page arXiv:2012.01469, Dec. 2020. 1

[32] Vikram V. Ramaswamy, Sunnie S. Y. Kim, and Olga Russakovsky. Fair attribute classification through latent space de-biasing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 6

[33] Hiranmayi Ranganathan, Hemanth Demakethepalli Venkateswara, Shayok Chakraborty, and Sethuraman Panchanathan. Deep active learning for image classification. In *2017 IEEE International Conference on Image Processing, ICIP 2017 - Proceedings*, 2018. 2

[34] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization, 2019. 2, 7

[35] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Unsupervised learning of debiased representations with pseudo-attributes, 2021. 6, 7

[36] Burr Settles and Mark W. Craven. An analysis of active learning strategies for sequence labeling tasks. In *EMNLP*, 2008. 2

[37] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. *arXiv e-prints*, page arXiv:1707.02968, July 2017. 1

[38] Enzo Tartaglione, Carlo Alberto Barbano, and Marco Grangetto. EnD: Entangling and Disentangling deep representations for bias correction. *arXiv e-prints*, page arXiv:2103.02023, Mar. 2021. 1

[39] Robert Torfason, Eirikur Agustsson, Rasmus Rothe, and Radu Timofte. From face images and attributes to attributes. In Shang-Hong Lai, Vincent Lepetit, Ko Nishino, and Yoichi Sato, editors, *Computer Vision – ACCV 2016*, pages 313–329, Cham, 2017. Springer International Publishing. 6

[40] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528, 2011. 2

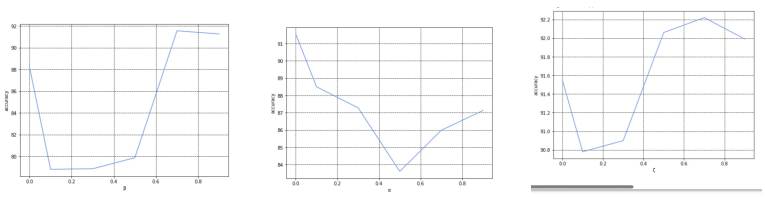
[41] Emiel van Miltenburg. Stereotyping and bias in the flickr30k dataset. *Proceedings of the Workshop on Multimodal Corpora: Computer vision and language processing*, abs/1605.06083, 2016. 1

[42] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 6

[43] Donggeun Yoo and In So Kweon. Learning loss for active learning, 2019. 2, 7

[44] Richard Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning (ICML)*, pages 325–333, 2013. 1

[45] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017. 1



(a) (b) (c)

Figure 4. (a)Hyper-parameter Tuning on β ($\mu_{Bias} + \beta\mu_{TB}$) (b)Hyper-parameter Tuning on α ($\mu_{Bias} + \alpha\mu_{BB} + 0.7\mu_{TB}$) (c)Hyper-parameter Tuning on ζ ($\mu_{Bias} + 0.7\mu_{TB} + \zeta\mu_{UR}$)