Epistemic Uncertainty-Weighted Loss for Visual Bias Mitigation

Rebecca S Stone, Nishant Ravikumar, Andrew J Bulpitt, David C Hogg
University of Leeds
{r.s.stone, n.ravikumar, a.j.bulpitt, d.c.hogg}@leeds.ac.uk

Abstract

Deep neural networks are highly susceptible to learning biases in visual data. While various methods have been proposed to mitigate such bias, the majority require explicit knowledge of the biases present in the training data in order to mitigate. We argue the relevance of exploring methods which are completely ignorant of the presence of any bias, but are capable of identifying and mitigating them. Furthermore, we propose using Bayesian neural networks with an epistemic uncertainty-weighted loss function to dynamically identify potential bias in individual training samples and to weight them during training. We find a positive correlation between samples subject to bias and higher epistemic uncertainties. Finally, we show the method has potential to mitigate visual bias on a bias benchmark dataset and on a real-world face detection problem, and we consider the merits and weaknesses of our approach.

1. Introduction

Modern computer vision models are highly susceptible to learning bias and discrimination present in datasets, leading to an unfair model. While in an ideal world, an unbiased data generation or collection process would fully mitigate biases in model performance, data bias is complex and difficult to fully identify for societal and historical reasons. Even with an ideal data sampling procedure, bias can still be present.

Practitioners within the artificial intelligence community are becoming increasingly aware of gender and racial biases learned and amplified by models [10, 21, 16, 4, 23]. Yet given the complexity and multitude of features present in visual data, there are additional biases which we may not even be aware of. Wang et al. [31] show that simple transformations such as converting images to grayscale, taking a centre crop, and reducing image resolution all affect model fairness when introduced as a bias in the training data. In the medical imaging domain, bias has been shown to be introduced through visual artifacts [7, 6] and demographics (often a proxy for economic status) [11, 26, 27]. Thus, debiasing methods which can be effective without any prior knowledge of the types and sources of bias present in the data are valuable for a variety of applications.

Modern visual systems are trained with large image datasets, where each dataset is a collection of visual attributes. A fair model is one where an input, regardless of its combination of attributes, has an equal likelihood of being assigned a correct outcome as any other input with a different set of attributes. While numerous types of bias exist (we recommend [25] for a detailed discussion), we choose to focus on bias as a whole via two broad categories as follows:

1. *Minority group bias*. When a subgroup of the data has a particular attribute or combination of attributes which are relatively uncommon compared to the rest of the dataset, they form a minority group. A model is less likely to correctly predict for samples from a minority group than for those of the majority.

2. *Sensitive attribute bias*. A sensitive attribute (also referred to as "protected") is one which should not be used by the model to perform the target task, but which provides an unwanted "shortcut" which is easily learned, and results in an unfair model.

In a scenario of complete blindness, the model is only aware of the target task, and has no prior knowledge of bias in the training set.

Most current state-of-the-art bias mitigation techniques directly use bias-informing metadata to either adjust the data before training, the model during training, or the predictions at inference time. Oversampling techniques aim to balance out the minority samples in the labelled data either before or during training to create the appearance of having a balanced set [20]. The *fairness through blindness* approach forces the model to simultaneously learn the target task and to ignore, or "learn to not learn" a protected variable [2]. However, these methods are susceptible to redundant encoding, where combinations of other non-protected variables act as a proxy for the protected variable. In contrast, *fairness through awareness* [13] approaches encode and explicitly mitigate all protected features. A model is

taught all subgroup and target class combinations and at inference time these relationships are removed. A recent approach titled *domain independent training* aims to remove class-bias correlations by averaging class decision boundaries [31].

Other approaches manipulate the sample feature representations in latent space to disentangle the attributes relevant for the target task from spuriously correlated features [29, 28]. [30] use a skewness-aware reinforcement learning method to determine the skewness between races on a face dataset with an adaptive margin loss function. All of these approaches [29, 28, 30] also require bias-informed metadata.

Conversely, Amini et al. [3] propose a novel method based on a variational auto-encoder (VAE) to learn latent structure simultaneously alongside the target task, and mitigate bias during training using dynamic batch selection to favor samples which are likely to be subject to bias based on their location in latent space. Recently, Xu et al.[33] use a false positive rate penalty loss which also requires no prior knowledge of bias in the training data, and changes the objective function to reward learning a more fair model.

We explore leveraging epistemic uncertainty estimates to mitigate both sensitive attribute and minority group visual bias without any prior knowledge of the types or sources of bias in the data. Unlike the VAE used in [3] which is constrained to uni-modal latent representations, we use a fully Bayesian neural network for a multi-modal posterior and better uncertainty estimates. We demonstrate on a visual bias benchmark dataset how sample-level epistemic weighting can mitigate sources of bias blindly, and further show on a real world face classification dataset how our approach can identify and mitigate sources of bias that other bias-informed methods cannot. We conclude by discussing the merits and weaknesses of our proposed approach.

2. Related Work

2.1. Bayesian Neural Networks

While Bayesian approximation via dropout [14] has made Bayesian deep neural networks applicable to many domains due to ease of use and scalability, Markov chain Monte Carlo (MCMC) algorithms [9] are widely considered the gold standard for Bayesian inference. However, they are computationally intractable for large vision datasets or high-dimensional data frequently encountered in real-world computer vision applications. A well-known scalable variant of MCMC is stochastic gradient MCMC (SG-MCMC), based on diffusion processes such as the Langevin diffusion. Diffusion processes are discrete-time approximations of continuous-time processes, formulated as stochastic differential equations (SDEs) to describe the time evolution of a moving object subject to both random and non-random forces.

Given model parameters θ, dataset D, prior $p(\theta)$, and potential energy $U(\theta)$, the posterior distribution is $p(\theta \mid D) \propto exp(-U(\theta)) = -\log p(D \mid \theta) - \log p(\theta)$. As computing $U(\theta)$ is not feasible for all D, SG-MCMC methods approximate $U(\theta)$ via mini-batch learning.

Welling and Teh [32] propose Stochastic Gradient Langevin Dynamics (SGLD) and substitute the gradient of the log-posterior density with the stochastic gradient over the minibatch and an additive Gaussian noise term that acts as an upper bound on the error.

$$\theta_i = \theta_{i-1} - \alpha_i \Delta \tilde{U}(\theta_i) + \sqrt{2\alpha_i \epsilon_i} \tag{1}$$

The update to parameters is shown in Equation 1, at iteration i of the algorithm, for normal distribution ϵ_i, stepsize α_i and minibatch approximation of the potential energy $\tilde{U}$.

While convergence is in practice slower than for other MCMC algorithms, the parameter update process closely resembles stochastic gradient descent and is generalisable to any neural network. To speed up convergence and better explore complex multimodal distributions common for deep neural networks, Zhang et al. [34] propose cyclical SG-MCMC (cSG-MCMC), where a cyclical stepsize schedule allows for quicker discovery of new modes. For each cycle of the learning rate schedule, an initial larger step size allows for exploration, and the subsequent smaller step sizes allow for sampling.

2.2. Leveraging Bayesian Uncertainties

Uncertainties can be divided into two categories, epistemic and aleatoric. Epistemic uncertainty, or model uncertainty, arises from lack of knowledge either about a process or parameter, and is caused by missing information or data. This uncertainty can be reduced given more data. Aleatoric uncertainty, or data uncertainty, arises from probabilistic variations or noise in the data and even given an infinite amount of data, cannot be reduced. [17] demonstrate that epistemic uncertainty can reveal data bias, while [1] argue that fairness approaches should equalize only errors arising from epistemic, not aleatory uncertainties. Thus, for mitigating bias, we are most interested in epistemic uncertainties.

Branchaud-Charron et al. [8] explore whether using Bayesian Active Learning by Disagreement (BALD) [15] can help mitigate bias against a protected class. They demonstrate that an acquisition scheme which greedily reduces epistemic uncertainty has potential for bias mitigation. This method shows promising results for leveraging epistemic uncertainties to mitigate bias when all of a target class is a minority group, but does not deal with sensitive feature scenarios. Khan et al. [22] use Bayesian uncertainty estimates to deal with class imbalance by weighting the loss function to move learned class boundaries away

from more uncertain classes. While their primary focus is class uncertainties, they also consider sample-based uncertainties. They propose a curriculum learning schedule with a variational dropout Bayesian neural network which approximates uncertainties first using softmax outputs, then class-level uncertainties, and finally for the last 10 epochs, sample-based uncertainties. As we are focusing on both sensitive attribute and minority group bias, neither of which we expect to be constrained to a unique class, our approach focuses on sample-level uncertainties. This eliminates the need for tuning a curriculum learning schedule, and for encouraging the model to look at class uncertainties. Furthermore, we opt to use SG-MCMC over variational dropout for better uncertainty estimates.

3. Methodology

We propose a simple uncertainty-weighted loss function for visual bias mitigation. Given a Bayesian neural network with parameters θ and posterior $p(\theta \mid D, x)$ for class-labelled training data D and test sample x_i, the predictive posterior distribution for a given predicted class y_i is then:

$$p(y_i \mid D, x_i) = \int p(y_i \mid \theta) p(\theta \mid D, x_i) d\theta \tag{2}$$

We can approximate this predictive posterior via Monte Carlo sampling. Given T Monte Carlo samples total, $\frac{T}{c}$ per learning schedule cycle c, we thus have predictive mean:

$$\mu_i \approx \frac{1}{T} \sum_{j=1}^{T} p(y_i \mid x_i, \theta_j) \tag{3}$$

and model uncertainty corresponding to this prediction:

$$\sigma_i \approx \frac{1}{T} \sqrt{\left(\sum_{j=1}^{T} p(y_i \mid x_i, \theta_j) - \mu_i \right)^2} \tag{4}$$

We propose the following uncertainty-weighted loss function for training sample (x_i, y_i), given the cross entropy loss $L(x_i, y_i)$, the additive Gaussian noise term from 1 and a tunable parameter κ controlling the degree of weighting, especially for high-uncertainty samples.

$$\hat{L}(x_i, y_i) = L(x_i, y_i) * (1.0 + \sigma_{i,y_i})^{\kappa} \tag{5}$$

As we expect a normally weighted sample to have weight 1.0, we shift the distribution such that lowest uncertainty samples are never irrelevant to the loss term. We compute $\hat{L}$ sample-wise and then reduce over the minibatch. $\kappa = 1$ is equivalent to a normal weighting, whereas $\kappa \to \infty$ increases the importance of high-uncertainty samples. In our fully bias-unaware setup, κ is optimised based on optimal validation loss tuned via grid search.

Algorithm 1 Training loop using uncertainty-weighted loss

Require: Training data X, Y, weighting parameter κ

Initialize parameters θ
for each cycle, c **do**
  for each epoch in cycle, e **do**
    if e in sampling phase **then**
      Sample $\theta_j \sim P(\theta \mid D, x)$
      Save $P(y = \hat{y} \mid x, \theta_j) \forall x \in X$
      Update $\left[w \leftarrow w - \epsilon \nabla L(x, y) \right]_{w \in \theta}$
    else if $c > 0$ **then**
      $\sigma_i \leftarrow \frac{1}{T} \sqrt{\left(\sum_{j=1}^{T} p(y_i \mid x_i, \theta_j) - \mu_i \right)^2}$
      $\hat{L}(x_i, y_i) \leftarrow L(x_i, y_i) * (1.0 + \sigma_{i,y_i})^{\kappa}$
      Update $\left[w \leftarrow w - \epsilon \nabla \hat{L}(x, y) \right]_{w \in \theta}$
    end if
  end for
end for

The earliest moment at which we can compute σ over the training data is during the sampling phase of the first cycle. Uncertainty values are updated at each consecutive cycle to reflect the developing posterior, requiring a total of $T(C-1)$ samples from the posterior for total number of cycles C. We speed up this process during training by saving each training sample's predictive distribution for each θ_j rather than all θ. Once the sample-wise model uncertainties have been computed, all predictions can be discarded. We set the length of the sampling phase to be 5 epochs, the shortest length for which uncertainty estimates are consistently stable for a fixed seed.

4. Evaluation

4.1. CIFAR-10 Skewed Visual Bias Benchmark

We compose the same baseline dataset, "CIFAR-10 Skewed" (CIFAR-10S) proposed by Wang et al. [31] to evaluate the performance of a Bayesian CNN without any adjustments. 95% of the images for 5 out of 10 target classes and 5% of the remaining classes are converted to grayscale, resulting in an overall balanced dataset with respect to colour but a strong skew within each class. As a model can learn the presence or absence of colour as a class indicator, this is an instance of sensitive attribute bias. We use the official CIFAR-10 test set and a 5:1 training-validation split for the total of 60000 images.

To provide a fair comparison, we follow the same training and architecture choices as proposed, with the exception of the learning rate, which is set to maximise performance for the cyclical step size schedule of our Bayesian formulation, and with momentum 0.9 due to the Langevin dynamic estimation. Using validation loss to choose hyperparameters, we train for 280 epochs and four cycles. With the ex-

Model	Description	Bias (↓)	Mean acc (%,↑)	Opp. (%,↓)	Odds. (%,↓)
Baseline	N-way softmax	0.074	88.5 ± 0.3	13.07 ± 0.4	7.19 ± 0.2
S-SAMPLING	N-way softmax	0.066	89.1 ± 0.4	12.58 ± 0.2	6.91 ± 0.1
ADVERSARIAL	w/ uniform confusion	0.101	83.8 ± 1.1	16.71 ± 1.4	9.28 ± 0.7
	w/ gradient reversal, proj	0.094	84.1 ± 1.0	14.13 ± 1.4	7.89 ± 0.8
DOMAINDISCRIM	joint ND-way softmax	0.040	90.3 ± 0.5	7.27 ± 0.3	4.02 ± 0.2
DOMAININDEPEND [31]	joint ND-way softmax	0.004	92.9 ± 0.1	1.07 ± 0.2	0.59 ± 0.1
FEATURELABEL [29]	N-way cos softmax per D	0.004	91.5 ± 0.2	0.83 ± 0.1	0.46 ± 0.1
DB-VAE [3]	latent structure re-weighting	0.167	90.2 ± 0.4	6.87 ± 0.5	0.78 ± 0.2
Our approach	w/ unc.-weighted loss	0.035	89.2 ± 0.2	12.11 ± 0.2	6.20 ± 0.2

Table 1. Multi-class classification, mean bias accuracy, equality of opportunity and equalized odds for bias benchmark dataset CIFAR-10S, a dataset with sensitive attribute bias. Note that all methods except for the baseline (a regular deterministic network with no bias mitigation), the DB-VAE, and ours are bias-informed during training.

	Baseline	*Our approach*
TPR Color (%,↑)	92.1 ±0.1	93.6 ±0.1
TPR Gray (%,↑)	91.6 ±0.1	93.0 ±0.1
TPR Gap (%,↓)	1.8	0.6

Table 2. Uncertainty-weighted loss reduces the TPR gap on minority bias dataset CIFAR-10M.

	Bias (%,↓)	Mean acc (%,↑)
Baseline	0.074	88.5 ± 0.3
cSG-MCMC	0.060	88.1 ± 0.2
cSG-MCMC weighted loss	0.035	89.2 ± 0.2

Table 3. Ablation study showing results of a Bayesian cSG-MCMC network with regular unweighted cross-entropy loss.

ception of our reproduction of the DB-VAE method in [3], all experiments use a ResNet-18 convolutional neural network as the base architecture. For the DB-VAE, we follow the same architecture as used in the original paper, adapted for multi-target classification on CIFAR-10S, up-sizing the images to 64 x 64 to match the expected input dimensions, and grid search to find the optimal value for de-biasing parameter, α = 0.001.

We evaluate the models using four metrics:

- **Mean test accuracy** on fully gray-scale and fully colour test sets, indicating how much the model learned the undesired correlation between the colour of each sample and its label;

- **Bias amplification score** [35] proposed first in a natural language processing setting and generalised for the CIFAR-10S dataset in Equation 6 as suggested in [31]. Gr_c is the number of gray-scale test samples predicted as class c, and Col_c the colour samples predicted as class c.

$$\text{bias ampl.} = \frac{1}{\mid C \mid} \sum_{c \in C} \frac{\max(Gr_c, Col_c)}{Gr_c + Col_c} - 0.5 \quad (6)$$

- **Equalized odds** [18] is satisfied for predictor $\hat{y}$ when $\hat{y}$ and sensitive attribute a are independent conditional on outcome y, $\forall \gamma \in 0, 1$: $P(\hat{y} = 1 \mid y = \gamma, a = 0) = P(\hat{y} = 1 \mid y = \gamma, a = 1)$. A difference in equality of opportunity score is derived as per [5] where FN_y^a is the number of false negatives of class y with protected attribute a:

$$\frac{1}{y} \sum_{y \in Y} \left| \frac{TP_y^1}{TP_y^1 + FN_y^1} - \frac{TP_y^0}{TP_y^0 + FN_y^0} \right| \quad (7)$$

- **Equality of opportunity** [18] is a relaxed form of equalized odds which requires non-discrimination on only one desired outcome, $y = 1$, $P(\hat{y} = 1 \mid y = 1, a = 0) = P(\hat{y} = 1 \mid y = 1, a = 1)$. A difference of equalized odds score as per [4] is as follows:

$$0.5 * \left(\left| FPR_y^1 - FPR_y^0 \right| + \left| TPR_y^1 - TPR_y^0 \right| \right) \quad (8)$$

For CIFAR-10S, we find that samples with a sensitive attribute have higher uncertainties. They constitute 20% of samples in the samples with the highest 10% uncertainties, while only 5% of samples in the dataset have a sensitive attribute. In contrast, less than 2% of samples in the lowest 10% uncertainties have a sensitive attribute.

As CIFAR-10S is a case of sensitive attribute bias, we formulate a second dataset CIFAR-10M "Minority" (CIFAR-10M) to represent minority group bias. We set grayscale as the minority attribute, and for each of the 10 classes, 5% of samples are converted to grayscale. The remaining 95% remain in colour. We split the training-validation sets using a 5:1 ratio. The results of our debiasing method on this dataset are presented in Table 2.

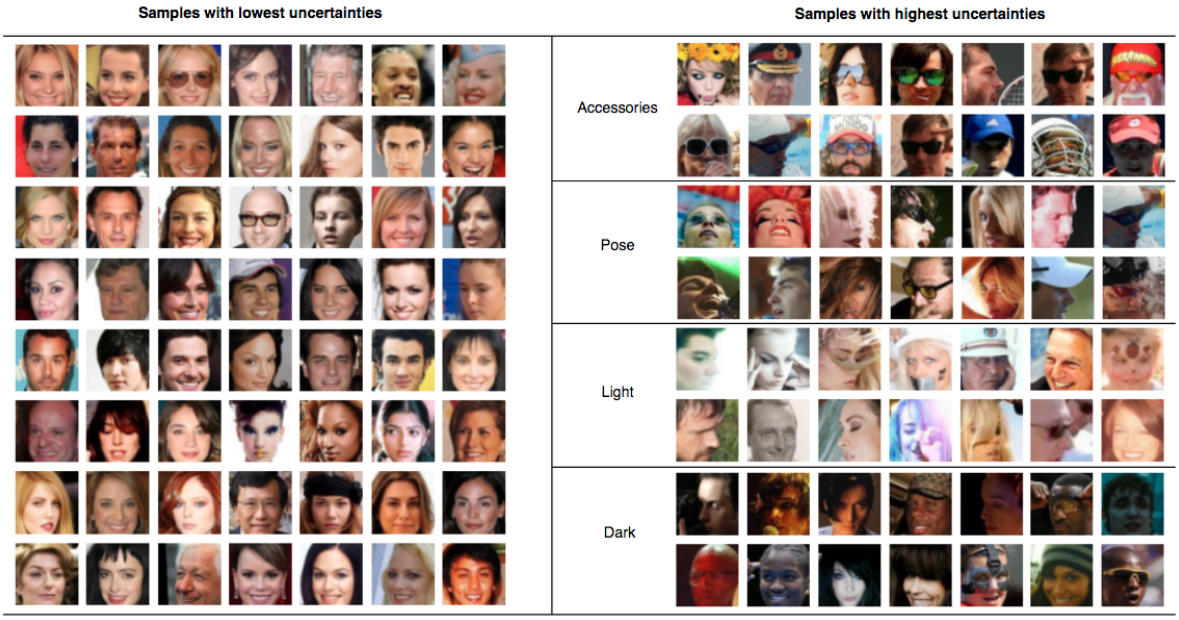

Figure 1. Face training samples from CelebA with lowest epistemic uncertainties (left) and faces with highest epistemic uncertainties (right). The faces with low uncertainties tend to be well-lit, facing forward with hair cleanly framing the face, and primarily lighter-skinned with few obscuring accessories. Faces with high uncertainties are more likely to be subject to discrimination due to variance in lighting, pose, colouring, and obscuring objects, among other reasons.

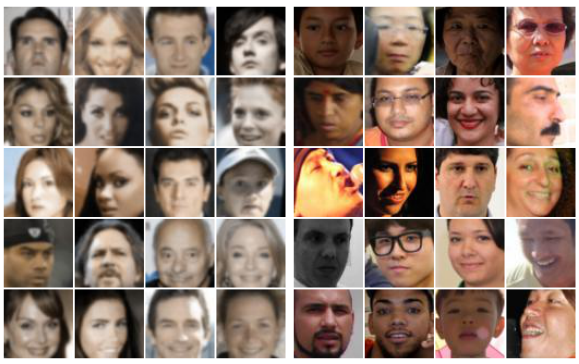

Figure 2. **Left:** a random selection of samples from CelebA; **Right:** and the same from FairFaces.

While not competitive with all bias-informed methods, our approach demonstrates an ability to de-bias blindly on both the benchmark dataset with sensitive attribute bias (CIFAR-10S), and our constructed dataset with minority group bias (CIFAR-10M).

4.2. On a Real-World Face Detection Problem

Similar to [3], we create a face vs. no-face binary classification dataset using 20k instances of faces from CelebA [24] and 20k non-face samples from a variety of different classes from ImageNet [12], for a training set of 40k images. We evaluate the model on FairFaces [21][1], consisting of 108,501 images with gender, race, and age annotations, balanced across 7 race groups: White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino, and 9 age subgroups from "0-2" to "over 70". The high level of diversity is visible in Figure 2. The distribution shift between the training and test data is such that we center crop the CelebA images to 124 x 124 and downsample all face and non-face training data to 64 x 64 such that the general resolution and positioning of the face in the images do not hinder the model from generalising. We use the Fréchet Inception Distance [19] between the CelebA and FairFace datasets to guide our transforms to the CelebA data, supported by visual comparison of random selections of images from both datasets. No transforms other than resizing are applied to the FairFace data, ensuring that the diversity is still intact.

We train a regular ResNet18 to convergence at 30 epochs using an 8:2 split of the CelebA/ImageNet dataset for training and validation and a standard SGD optimizer with learning rate of 0.01. The cSG-MCMC Bayesian model with uncertainty-weighted loss is similarly trained with 4 cycles of 30 epochs each for a total of 120 epochs.

For every subgroup, the uncertainty-weighted loss de-

[1]Due to the inaccessibility of the Pilot Parliaments Benchmark dataset [10] because of data privacy issues, we have chosen to evaluate on FairFaces.

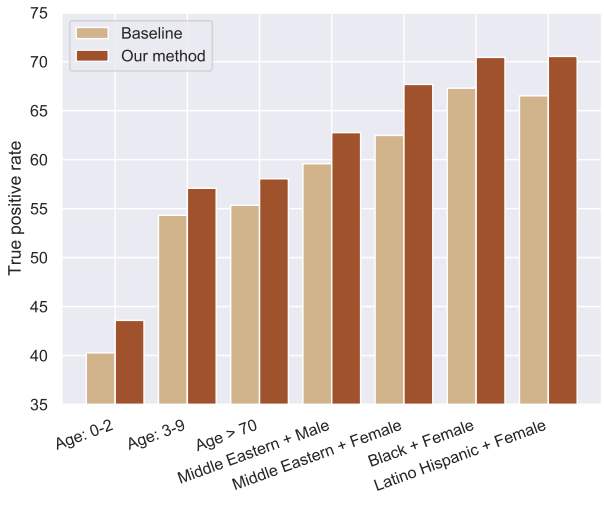

Figure 3. Performance discrepancies between baseline (deterministic model with no de-biasing) and Bayesian model with uncertainty-weighted loss for minority subgroups with lowest TPRs.

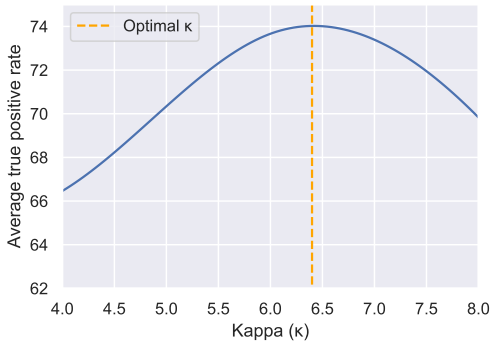

Figure 4. True positive rate (TPR) over the entire FairFace dataset as a function of tunable de-biasing parameter kappa κ, showing how the degree of de-biasing can be controlled by κ.

creases the TPR gap, with the discrepancies for the 7 subgroups with lowest TPR rates shown in Figure 3.

Given that sample-level weighting by a factor of N during training is equivalent to that sample appearing N times, our approach could be categorized as a type of sub-sampling algorithm. Thus, it suffers from the same weakness as all sub-sampling algorithms, a tendency to overfit oversampled data. This can only be partially mitigated by aggressive data augmentation. We hypothesize that this explains why increasing tunable de-biasing parameter κ beyond the optimal value results in worse performance as shown in Figure 4.

Figure 1 shows samples with high uncertainties, which clearly have features which make them more likely to be subject to bias. A bias-informed method could strongly mitigate bias due to known societal biases such as gender and race, or skin phenotype. But since it would be unlikely to also have access to meta-data which identifies variances in lighting, pose, image resolution, etc., all of which also result in unfairness, such methods would not target such biases.

Such an approach is valuable in medical imaging applications with large population image analysis due to the inherent difficulty in collecting meta data. Sensitivity and privacy requirements result in imaging datasets with very few annotations and little, if any, associated patient meta data. This presents a challenge for bias-informed methods, and serves as motivation for further exploration of methods which can mitigate without requiring comprehensive knowledge of all biases.

5. Conclusion

We have shown that an epistemic uncertainty-weighted loss function has potential for bias mitigation for datasets with unknown sources of bias. We cannot conclude that the approach works in all cases based on our experiments, nor do we attempt to provide a mathematical proof that this should be the case. Some training datasets may contain an over-sampling from unprivileged groups, in which case the correlation with epistemic uncertainties may no longer exist. Thus, our exploration focuses only on cases of minority group and sensitive attribute bias. While outperformed by most bias-informed models, our method is a step towards exploring how epistemic uncertainties in Bayesian neural networks can be leveraged for identifying, understanding, and mitigating the types and sources of visual bias in data.

Acknowledgements. The first author is a recipient of a Rabin Ezra Scholarship. Special thanks to Jose Sosa and Mohammed Alghamdi from the School of Computing's Computer Vision Group for critical feedback and many great discussions.

References

[1] Junaid Ali, Preethi Lahoti, and Krishna P Gummadi. Accounting for model uncertainty in algorithmic discrimination. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 336–345, 2021. 2

[2] Mohsan Alvi, Andrew Zisserman, and Christoffer Nellåker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 1

[3] Alexander Amini, Ava P Soleimany, Wilko Schwarting, Sangeeta N Bhatia, and Daniela Rus. Uncovering and mitigating algorithmic bias through learned latent structure. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 289–295, 2019. 2, 4, 5

[4] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lo-

hia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018. 1, 4

[5] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017. 4

[6] Alceu Bissoto, Michel Fornaciali, Eduardo Valle, and Sandra Avila. (de) constructing bias on skin lesion datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1

[7] Alceu Bissoto, Eduardo Valle, and Sandra Avila. Debiasing skin lesion datasets and models? not so fast. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 740–741, 2020. 1

[8] Frédéric Branchaud-Charron, Parmida Atighehchian, Pau Rodríguez, Grace Abuhamad, and Alexandre Lacoste. Can active learning preemptively mitigate fairness issues? *arXiv preprint arXiv:2104.06879*, 2021. 2

[9] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of markov chain monte carlo*. CRC press, 2011. 2

[10] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018. 1, 5

[11] Terrance De Vries, Ishan Misra, Changhan Wang, and Laurens Van der Maaten. Does object recognition work for everyone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 52–59, 2019. 1

[12] 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. 5

[13] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012. 1

[14] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016. 2

[15] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pages 1183–1192. PMLR, 2017. 2

[16] Markos Georgopoulos, Yannis Panagakis, and Maja Pantic. Investigating bias in deep face analysis: The kanface dataset and empirical study. *Image and Vision Computing*, 102:103954, 2020. 1

[17] Asma Ghandeharioun, Brian Eoff, Brendan Jou, and Rosalind Picard. Characterizing sources of uncertainty to proxy calibration and disambiguate annotator and data bias. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 4202–4206. IEEE, 2019. 2

[18] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016. 4

[19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

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

[21] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913*, 2019. 1, 5

[22] Salman Khan, Munawar Hayat, Syed Waqas Zamir, Jianbing Shen, and Ling Shao. Striking the right balance with uncertainty. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 103–112, 2019. 2

[23] Agostina J Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594, 2020. 1

[24] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018. 5

[25] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021. 1

[26] Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y Chen, and Marzyeh Ghassemi. Chexclusion: Fairness gaps in deep chest x-ray classifiers. In *BIOCOMPUTING 2021: Proceedings of the Pacific Symposium*, pages 232–243. World Scientific, 2020. 1

[27] Karin Stacke, Gabriel Eilertsen, Jonas Unger, and Claes Lundström. Measuring domain shift for deep learning in histopathology. *IEEE journal of biomedical and health informatics*, 25(2):325–336, 2020. 1

[28] Enzo Tartaglione, Carlo Alberto Barbano, and Marco Grangetto. End: Entangling and disentangling deep representations for bias correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13508–13517, 2021. 2

[29] William Thong and Cees GM Snoek. Feature and label embedding spaces matter in addressing image classifier bias. *arXiv preprint arXiv:2110.14336*, 2021. 2, 4

[30] Mei Wang and Weihong Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9322–9331, 2020. 2

[31] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8919–8928, 2020. 1, 2, 3, 4

[32] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688. Citeseer, 2011. 2

[33] Xingkun Xu, Yuge Huang, Pengcheng Shen, Shaoxin Li, Jilin Li, Feiyue Huang, Yong Li, and Zhen Cui. Consistent instance false positive improves fairness in face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 578–586, 2021. 2

[34] Ruqi Zhang, Chunyuan Li, Jianyi Zhang, Changyou Chen, and Andrew Gordon Wilson. Cyclical stochastic gradient mcmc for bayesian deep learning. *arXiv preprint arXiv:1902.03932*, 2019. 2

[35] 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 *EMNLP*, 2017. 4