Mitigating Bias Using Model-Agnostic Data Attribution

Sander De Coninck, Sam Leroux, Pieter Simoens
IDLab, Department of Information Technology at Ghent University - imec
Technologiepark 126, B-9052 Ghent, Belgium
{sander.deconinck, sam.leroux, pieter.simoens}@ugent.be

Abstract

Mitigating bias in machine learning models is a critical endeavor for ensuring fairness and equity. In this paper, we propose a novel approach to address bias by leveraging pixel image attributions to identify and regularize regions of images containing significant information about bias attributes. Our method utilizes a model-agnostic approach to extract pixel attributions by employing a convolutional neural network (CNN) classifier trained on small image patches. By training the classifier to predict a property of the entire image using only a single patch, we achieve region-based attributions that provide insights into the distribution of important information across the image. We propose utilizing these attributions to introduce targeted noise into datasets with confounding attributes that bias the data, thereby constraining neural networks from learning these biases and emphasizing the primary attributes. Our approach demonstrates its efficacy in enabling the training of unbiased classifiers on heavily biased datasets.

1. Introduction

In the domain of computer vision, various methodologies are employed to mitigate bias and uphold fairness in machine learning models operating on sensitive attributes. Data bias can manifest in various ways, such as when the data representation disproportionately favors a specific group of subjects [22]. Alternatively, bias may arise from substantial differences between the training and production environments [23]. Moreover, biases within the data can be either explicit and acknowledged beforehand or concealed. We specifically focus on situations where bias is pre-existing and originates from disparities between the training and evaluation datasets due to a confounding variable heavily influencing the former. For instance, in the CelebA [15] dataset, a significant majority of women are depicted with blond hair, whereas only a minority of men exhibit this trait. Consequently, a model trained on such data may exhibit unequal performance across men and women in a production setting where this distribution does not hold true as it has learned to associate having blond hair with being a woman.

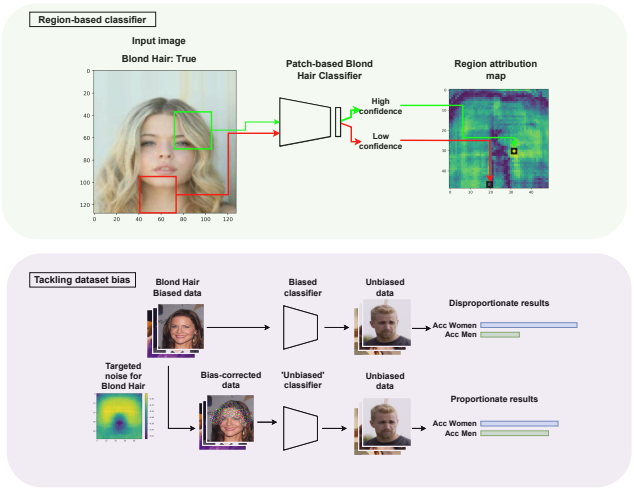


Figure 1. Our strategy to address biased learning. We utilize a region-based classifier to classify individual image patches based on attributes that introduce bias to the data. By leveraging the confidence of the trained model, we identify the regions within an image where attribute-related information is most concentrated. Subsequently, we introduce targeted noise to these areas in the training data to prevent models from overfitting to the confounding attributes.

One prevalent technique to mitigate the learning of biases is dataset re-sampling, where instances from minority or majority classes are either over-sampled or under-sampled to achieve a more equitable distribution. However, in scenarios such as facial data, where numerous attributes are present, achieving a perfectly balanced dataset may be unattainable as this would require an equal number of samples for each subcategory. In the case of CelebA, which contains 40 binary attributes, this would mean you require an equal amount of samples for each of the 2^{40} possible combinations. Another approach is to utilize dataset augmentation [23], which involves altering the training data to

rectify imbalances in underrepresented classes or counter biases inherent in the training dataset, for example by generating synthetic data [25, 32]. Our solution falls under this category as well. We propose a solution where we identify critical pixel areas for classifying attributes that introduce a bias in the data (e.g., blond hair in face data). By strategically introducing targeted noise to these regions in the training data, we aim to prevent the model from overfitting to these confounding attributes. A conceptualisation of our technique can be seen in Figure 1.

Attributing pixels that are important for classification is typically referred to as saliency mapping. Saliency mapping techniques were originally developed to elucidate machine learning models rather than datasets. Consequently, they attempt to identify pixel regions crucial for a specific model's classification. However, for mitigating biases being learned, it's more pertinent to determine which regions **potentially** contribute to classification. In other words, it's essential to discern which image regions harbour valuable information for a particular class (e.g., identifying facial features aiding in ethnicity classification). This will not always align with specific model attribution maps, as not all networks learn in the same way, and some may lay more importance on a certain pixel region than others.

To tackle this issue, we introduce data attribution through classification. We employ a small-scale classifier that predicts a sensitive label for the entire image from a single image patch. By identifying regions where the classifier confidently predicts the label, assuming the model is well calibrated, we establish a measure of information content regarding the sensitive label within that patch. As a result, this methodology allows us to attribute specific image regions to particular classes.

The resulting attributions could be used for a number of purposes, but we see the most promise in fields where it is undesirable that certain attributes are learned. The most notable of these is learning using a biased dataset. We propose to use our attributions as a way to regularize the training process, if some biases in the data are known a priori (e.g. due to a high correlation between labels), then we can compel the model learning to classify desired attributes to divert focus away from those regions important for classifying the confounder.

We demonstrate the effectiveness of our attribution technique on two prominent face datasets, namely FairFace [12] and CelebA [15]. These datasets are selected due to their comprehensive annotations and consistent alignment of posture, facilitating the validation of our attributions' utility through averaged attributions across multiple images. Subsequently, we explore the application of our attributions in training unbiased perceived gender classifiers on biased data, leveraging the CelebA dataset. By creating biased subsets where specific attributes are prominent for women but absent for men, we evaluate subgroup accuracy on a balanced dataset. Our experiments reveal that incorporating noise based on our attributions to the training data proves beneficial in mitigating the adverse effects of dataset bias on classifier performance.

We make the following contributions in this paper:

- We developed a general model-independent data attribution technique that can foster data understanding
- We showcase the use of this attribution technique for data augmentation to train neural networks under known biases and experiment with different noise addition types and settings

The remainder of this paper is organized as follows. We position related works in Section 2. Following this, we introduce our region classifier and show how we obtain attributions in Section 3. Lastly, we experiment with using the attributions for developing robust classifiers in Section 4. We conclude our work in Section 5.

2. Related works

2.1. Dataset bias and fairness

Correlation is a widely studied problem in fair research, mainly concerning spurious correlations in the dataset through societal biases or dataset construction. For example, for the CelebA dataset, it is known that the attractive attribute is biased towards women and that a majority of the women have blonde hair. Rajabi et al. [24] demonstrate that mitigating bias between gender and attractiveness for the CelebA dataset influences classification of several other attributes related to attractiveness as well. Similarly, Denton et al. [4] found that manipulating non-hair attributes such as heavy makeup resulted in a change in hair style/and or color.

Addressing these spurious correlations and biases represents an active area of investigation within the research community. Some studies concentrate on improving datasets to ensure a more equitable distribution of data samples, either through resampling techniques or augmenting the dataset with new data generated, for example, using generative models [25, 32]. Others employ techniques during model training to guide the learning process, facilitating the acquisition or elimination of specific biases [33]. For instance, Kim et al. utilize a regularization loss based on mutual information to prevent the model from learning biased attributes [14]. Additionally, certain approaches leverage causal learning methodologies to achieve unbiased recognition [31].

2.2. Attribution beyond explainability

Pixel attribution techniques are utilized to identify the areas or pixels within an image that significantly contribute to a model's prediction. These techniques typically fall into two

categories: occlusion or perturbation-based, and gradient-based methods. The former, being model-agnostic, assesses the impact on prediction when certain regions are omitted, while the latter computes gradients concerning the input image [19]. Noteworthy gradient-based techniques include Grad-CAM [27], Integrated Gradients [30], and XRAI [11]. For occlusion-based approaches, LIME [26] and SHAP [17] are widely recognized.

These methods primarily serve to enhance the interpretability of deep neural networks. However, they also see use beyond, in improving the generalization of deep learning models [2], such as those used in facial expression recognition [13, 18] and person detection tasks [1], including scenarios involving thermal imagery [6]. Particularly relevant to our work, Huang et al. [9] employ saliency techniques to address biases in facial recognition models. Their approach employs gradient attention maps, ensuring consistent attention patterns across diverse racial backgrounds. However, our work differs in that we utilize saliency to mitigate confounding variables being learned, whereas they focus on ensuring uniform learning across different subgroups regardless of what is learned.

3. Region classifier

In our efforts to alleviate learned biases, our focus is directed towards identifying the regions containing vital information for classification. While saliency techniques are commonly used for this purpose, they often focus solely on pinpointing important pixels for a specific model, potentially overlooking broader regions of interest. To address this, we deploy a region classifier capable of analyzing image patches and attributing characteristics of the entire image, such as determining a person's age. By assessing the confidence levels of a well-trained classifier, we gain valuable insights into the specific areas where significant information is concentrated.

3.1. Training setup

Our region classifier employs a ResNet18 model architecture [8], which operates on patches of size $k \times k$. Additionally, the network takes in a region indicator, which indicates the region from which the patch was taken in the original image. This region indicator is translated into an embedding akin to the vision transformer approach [5]. This embedding is then concatenated to the input image before execution. The embeddings and the classification network are trained simultaneously. During the training process, a single patch per image is randomly selected for training. Patches are randomly selected from p possible positions, where patch i corresponds to having $((i \mod \sqrt{p}) * s, \lfloor i/\sqrt{p} \rfloor * s)$ as the top-left coordinates of the patch. Here, the stride s is calculated as $I - k/\sqrt{p}$, with I being the image size.

Our experiments utilize the FairFace [12] and CelebA [15] dataset, providing centred and aligned face images, which we resized to 128×128 pixels. We employ the AdamW optimizer [16] with an initial learning rate of 10^{-3}, reduced by a factor of 10 every 40 epochs during training, for a total of 90 epochs. Patch dimensions k are set to 32, and the number of possible patch positions p is 2041. A batch size of 256 is utilized, and the cross-entropy loss function with label smoothing of 0.1 is employed.

3.2. Confidence estimation

To effectively utilize our region classifier for determining whether a patch contains relevant information regarding the subject, it is imperative that our classifier can provide an indication of its confidence level. There are three prevalent methods for estimating confidence when working with classification networks. Each relies on softmax values derived from the output logits l, which are computed as follows:

$$S(l_i) = \frac{e^{l_i}}{\sum_j e^{l_j}} \tag{1}$$

The three confidence indicators used are the top softmax score, the margin, and the negative entropy score, which are calculated as follows:

- Top softmax score: this indicator is determined by selecting the highest softmax value among all classes.

$$\max_i s_i \tag{2}$$

- Margin: we determine it as the difference between the highest and second-highest softmax values.

$$\max_i s_i - \max_{j \neq \text{argmax}(s)} s_j \tag{3}$$

- Negative entropy score: it is computed as the negative sum of the softmax probabilities multiplied by their natural logarithms.

$$-\sum_i s_i \log s_i \tag{4}$$

In our experiments, we primarily rely on the negative entropy score as it considers the entire output distribution. While less straightforward to interpret than the top softmax score or the margin, we strive to enhance interpretability by normalizing outputs to a $[0, 1]$ range wherever feasible.

3.3. Calibration

When using our classifier for attribution, it's important that its confidence outputs are grounded in how accurately it can predict and, thus, how much information there is. A calibrated model is a model whose estimated confidence is close to its accuracy. The most common metrics of calibration are the Expected Calibration Error (ECE) [20], and reliability diagrams [3, 21].

Expected Calibration Error (ECE) quantifies the discrepancy between predicted and actual confidence levels. It is calculated as follows:

$$ECE = \sum_{i}^{N} b_i ||(p_i - c_i)|| \tag{5}$$

Here, N represents the number of bins used for calibration, b_i is the proportion of samples falling into bin i, p_i denotes the average predicted confidence, and c_i represents the average true accuracy. ECE measures the average difference between predicted confidence and actual accuracy across different confidence levels. It provides a holistic assessment of the calibration performance of a classifier.

Reliability diagrams, also known as calibration diagrams, graphically illustrate the alignment between predicted confidence levels and actual accuracy. They plot predicted confidence values against observed accuracy within equally spaced bins. Ideally, a well-calibrated classifier shows a diagonal line, indicating accurate confidence estimates. Deviations from this line highlight areas of overconfidence or underconfidence. Reliability diagrams offer a visual tool for evaluating a classifier's calibration performance.

Calibration diagrams, along with the corresponding Expected Calibration Error (ECE) scores, for networks trained to classify the three FairFace attributes are depicted in Figure 2. These diagrams were generated using all possible patches from the validation set using 100 bins. We observe that our networks are well-calibrated without having made any specific modifications. This is likely attributed to the high variability in data quality. Many regions may lack sufficient information for accurate classification but are still included, thereby mitigating overfitting and ensuring consistent calibration. Notably, for the age attribute, calibration appears to decrease significantly at high confidences, although such instances are infrequent, as evidenced by the lower plots.

3.4. Region-based attributions

For the final step, we can use our patch-based classifiers, which we have shown to be well-calibrated, to calculate attribution maps. To attribute an image with respect to a specific attribute, we employ our region classifier to assess confidence across numerous regions. This is achieved by iteratively collecting regions in a sliding window manner, as can also be seen in Figure 1. Subsequently, the obtained regions are batched and classified using the region classifier. Finally, a confidence score, such as negative entropy, is computed for each patch. The resulting attribution can be visualized as a $\sqrt{p} \times \sqrt{p}$ map, with each pixel (x, y) representing a region whose top left coordinate in the original image is $(s*x, s*y)$. Illustrative examples of our region attributions for the FairFace dataset are showcased in Figure 3. Note that we solely rely on confidence scores to calculate attributions, allowing a trained model to be applied to novel data without the need for ground-truth labels. One limitation of this attribution technique is its computational cost, as it requires classifying p image patches per image.

By averaging over all samples, mean confidence maps can be generated. These maps offer valuable insights into the spatial distribution of important attribute information within an image and can be used to relate these attributes to one another. The calculated maps for the FairFace attributes are presented in Figure 4, revealing distinct differences between the attributes. In addition to this, we trained region classifiers on select attributes of the CelebA dataset, for which we can see the mean attributions in Figure 5. We observe that the attribution maps for attributes such as Blond Hair and Eyeglasses correspond with human intuition.

For many applications, a pixel-space attribution map is more practical. We can achieve this by converting the region-based map to pixel space using Algorithm 1. The algorithm employs the confidences and locations of all patches present on the original image to produce a pixel-level confidence map matching the original image's dimensions. It calculates the confidence for each pixel by averaging the confidences of all patches containing that pixel. For all experiments in Section 4, we used pixel space attributions, which were normalized to a $[0, 1]$ range.

Algorithm 1 Map region attributions to pixel space

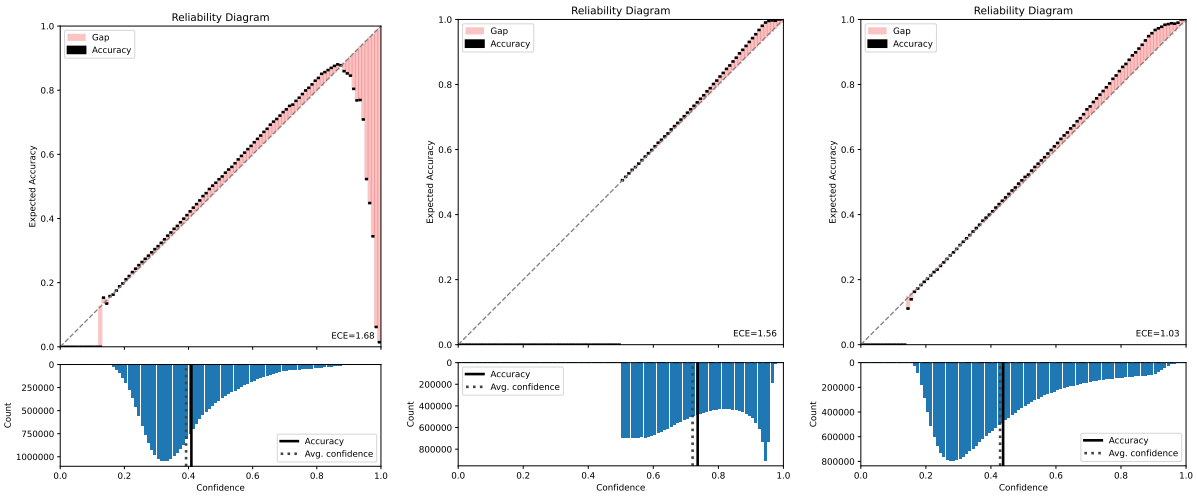
Input: $confidences$, $locations$, k, $image\_size$
$map \leftarrow$ zeros tensor of size $image\_size$
$counts \leftarrow$ zeros tensor of size $image\_size$
for $i = 0$ **to** length($locations$) - 1 **do**
  $(x, y) \leftarrow locations[i]$
  $map[y : y + k, x : x + k] \mathrel{+}= confidences[i]$
  $counts[y : y + k, x : x + k] \mathrel{+}= 1$
end for
return $map/counts$

4. Regularized training on biased data

To develop a robust training algorithm, we aim to leverage our attribution maps to address biases in the training dataset by introducing targeted noise to regions critical for detecting confounding attributes. Our attribution technique could enable manipulation of the training data to mitigate the model's reliance on these confounders. An illustrative figure can be seen in Figure 1.

4.1. Experiment Setup

In order to evaluate the feasibility of this approach, we conducted experiments utilizing the CelebA dataset, which comprises a comprehensive collection of annotated facial



(a) Age (b) Gender (c) Race

Figure 2. Reliability diagrams illustrating the performance of region classifiers trained on FairFace attributes. Our models exhibit strong calibration without any adjustments, as evidenced by the close alignment between average confidence and accuracy, alongside low Expected Calibration Error (ECE).

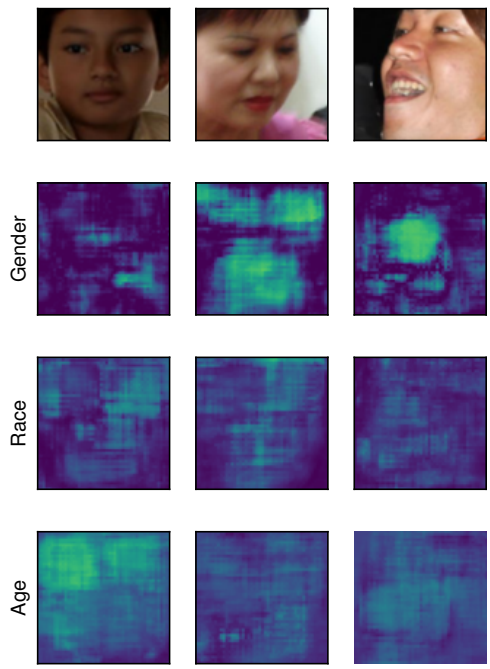
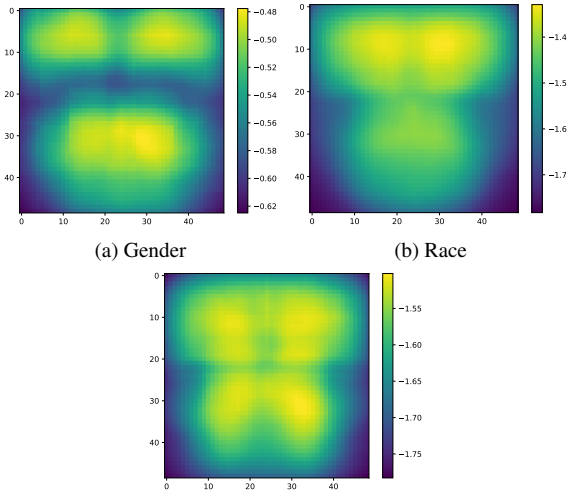


Figure 3. Examples of region-based attributions.



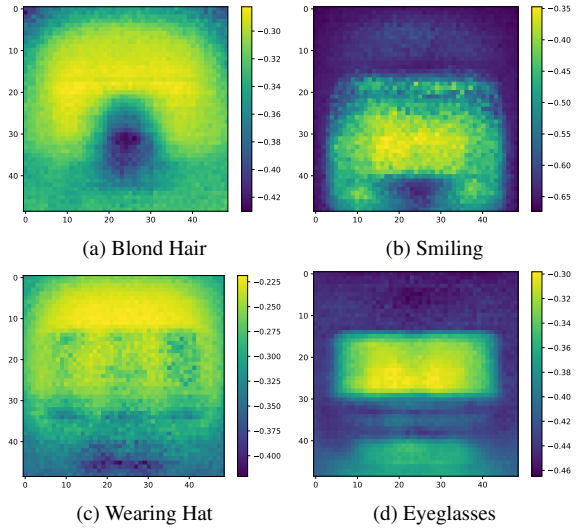
(a) Gender (b) Race

(c) Age

Figure 4. Mean attribution maps for the FairFace attributes.

attributes. Our investigation focuses on a scenario wherein a classifier is trained for predicting perceived gender classification[1]. The dataset is not sampled uniformly random from the CelebA dataset. Specifically, the dataset is perfectly balanced for the classification attribute (3000 instances of men, 3000 instances of women), but contains bias in terms of another attribute. Within the subset of men, none exhibit the confounding attribute, whereas among women, 2000 individuals possess the confounder while 1000 do not. We created datasets in this manner using Blond Hair, Eyeglasses, Smiling and Wearing Hat as confounding attributes. Subsequently, we assess performance on a test dataset that is balanced with respect to gender as well as the hidden at-

[1]In the CelebA dataset, perceived gender is denoted as *Male*, where a value of 1 represents individuals belonging to the category of men, while a value of 0 is assigned to those who do not, which we label as women for simplicity's sake.



(a) Blond Hair (b) Smiling

(c) Wearing Hat (d) Eyeglasses

Figure 5. Mean attribution maps for CelebA attributes.

Figure 6. Example of regularizing the data based on attributions for hair color. From left to right: Original, General Mask, Specific Mask, General Noise, Specific Noise.

tribute. By introducing this bias into the dataset, our hypothesis posits a significant discrepancy in test accuracy between male and female subjects, owing to the disparate distribution of data instances across gender categories in the test set.

To cultivate a robust classifier, we conducted experiments exploring various methods of utilizing attribution maps to alter the training data. We investigated the incorporation of both image-specific and general attributions (calculated over a large number of samples) of confounding attributes, as exemplified in Figure 5. For both mask types, we experimented with greying out as well as additive noise. An illustration of adding noise for the Hair Color attribute in all manners (General Mask, Specific Mask, General Noise, Specific Noise) is provided in Figure 6. When introducing noise, we drew samples from a Gaussian distribution with a mean of 0 and a standard deviation of 0.5. In all cases, we masked out regions deemed most crucial for classifying unintended biases, such as the top 30% most confident regions within an image. The selection of the noise addition scheme and the quantile used for masking out regions were treated as hyperparameters in our analysis.

Several metrics are available for evaluating model fairness, with Demographic Parity and Equality of Opportunity being the most prominent. In our assessment, focusing on perceived gender classification, we deemed it appropriate to examine accuracy per subgroup. This approach allows us to determine if any particular category (in this case, men or women) is disproportionately affected by the known bias.

We employed a customized Convolutional Neural Network (CNN) architecture for gender classification, as we observed that deeper network architectures such as VGG [28] and ResNet [8] were susceptible to overfitting. Our CNN architecture comprised three convolutional layers, supplemented with dropout regularization [29] and batch normalization [10]. To optimize the model parameters, we utilized the AdamW optimizer [16] with an initial learning rate set to 10^{-5}, followed by exponential decay with a decay rate (γ) of 0.95. We conducted training with a batch size of 128. Robust models incorporating noise were trained for 20 epochs, while those utilizing masking techniques were trained until convergence for 10 epochs.

4.2. Results

We conducted experiments using all attributes illustrated in Figure 5 as confounding attributes, with the results summarized in Table 1. The table presents overall accuracy, as well as accuracy for men and women separately on a balanced test set, with the gap column indicating the difference between these two. Notably, our experiments encompassed multiple quantiles, yet we only included those yielding the most significant results for brevity. Additionally, we compared these results against classifiers trained on a dataset of equivalent size but balanced with respect to the attribute. It's worth noting that the data was not fully balanced in the case of Blond Hair as the confounder for men but rather $54/46$ in favor of non-blond hair, as the original CelebA train set contains only 1387 men with blond hair.

First of all, we can observe that all models have some discrepancy between the accuracy for men and women, with those trained on the biased dataset without any modifications having the largest gap. Nevertheless, we observe that for each confounding attribute, there exists a specific combination of noise type and quantile for regularizing that yields notably improved equality in accuracy, despite the significant divergence in the training distribution of men compared to the balanced test data. Surprisingly, even when trained on balanced data concerning the attribute, we still observe a performance gap between men and women, albeit significantly smaller than with highly unbalanced data. This discrepancy likely persists due to other confounders that remain present, as we did not balance the data on the remaining 38 attributes. Furthermore, in the case of smiling, there is almost no discernible difference between balanced and unbalanced data. This observation may be attributed to the inherent difficulty in classifying smiling compared to gender, resulting in a lesser impact of this bias on the

			Accuracy ↑			Accuracy Men ↑			Accuracy Women ↑			Gap ↓		
Attribute	Noise	Type Quantile	Original	Balanced Δ	Ours Δ	Original	Balanced Δ	Ours Δ	Original	Balanced Δ	Ours Δ	Original	Balanced	Ours
Blond Hair	General Mask	0.60	0.74	+0.09	0.0	0.6	+0.2	+0.02	0.88	-0.03	-0.03	0.28	0.05	0.22
	General Noise	0.70	0.77	+0.06	-0.02	0.64	+0.16	+0.08	0.9	-0.05	-0.12	0.26	0.05	0.09
	Specific Mask	0.60	0.74	+0.09	+0.01	0.6	+0.2	+0.09	0.88	-0.03	-0.08	0.28	0.05	0.11
	Specific Noise	0.80	0.79	+0.04	-0.05	0.67	+0.13	+0.01	0.9	-0.05	-0.1	0.23	0.05	0.15
Eyeglasses	General Mask	0.60	0.71	+0.06	-0.02	0.61	+0.11	+0.03	0.82	+0.0	-0.07	0.21	0.10	0.12
	General Noise	0.60	0.72	+0.05	+0.01	0.61	+0.11	+0.05	0.84	-0.02	-0.04	0.23	0.10	0.14
	Specific Mask	0.60	0.71	+0.06	+0.03	0.61	+0.11	+0.13	0.82	+0.0	-0.08	0.21	0.10	0.05
	Specific Noise	0.80	0.72	+0.05	-0.01	0.61	+0.11	+0.1	0.84	-0.02	-0.13	0.23	0.10	0.09
Smiling	General Mask	0.80	0.81	+0.03	-0.04	0.71	+0.06	-0.02	0.9	+0.01	-0.05	0.19	0.13	0.16
	General Noise	0.60	0.85	-0.01	-0.05	0.79	-0.02	-0.04	0.91	-0.0	-0.07	0.12	0.13	0.09
	Specific Mask	0.95	0.81	+0.03	-0.02	0.71	+0.06	-0.02	0.9	+0.01	-0.01	0.19	0.13	0.20
	Specific Noise	0.60	0.85	-0.01	-0.05	0.79	-0.02	+0.01	0.91	-0.0	-0.1	0.12	0.13	0.05
Wearing Hat	General Mask	0.80	0.71	+0.08	+0.05	0.63	+0.11	+0.08	0.79	+0.05	+0.03	0.16	0.10	0.10
	General Noise	0.95	0.73	+0.06	-0.02	0.63	+0.11	+0.03	0.84	+0.0	-0.07	0.21	0.10	0.11
	Specific Mask	0.95	0.71	+0.08	0.0	0.63	+0.11	+0.01	0.79	+0.05	-0.01	0.16	0.10	0.15
	Specific Noise	0.95	0.73	+0.06	-0.03	0.63	+0.11	+0.04	0.84	+0.0	-0.1	0.21	0.10	0.08

Table 1. Summary of the noise addition experiment across multiple attributes of the CelebA dataset. Models are trained on a highly biased dataset regarding each attribute, leading to a disparity in performance between men and women. This disparity between performance for men and women is denoted in the gap column. Accuracies for models trained on a balanced dataset, and those using our noise addition regularization are shown relative to those trained on the original, biased dataset. Through the strategic addition of noise to regions crucial for each attribute based on a specified noise type, we diminish the discrepancy in accuracies between genders.

model [7]. However, our technique notably reduces the accuracy gap between men and women, approaching the performance achieved with balanced data, suggesting that our method is a more practical approach than obtaining a balanced dataset on all attributes.

Subsequently, we zoom in on the choice of the two hyperparameters. Figure 7 illustrates a comparison of various noise addition strategies for the Blond Hair attribute, where we removed the 30% most influential information related to the attribute from all training data. All robust training schemes either enhance or maintain the accuracy for men, while diminishing the accuracy for women, thus aligning subgroup accuracies and mitigating the effect of training on a biased dataset. Notably, the incorporation of General Noise brings both subset accuracies closest to parity.

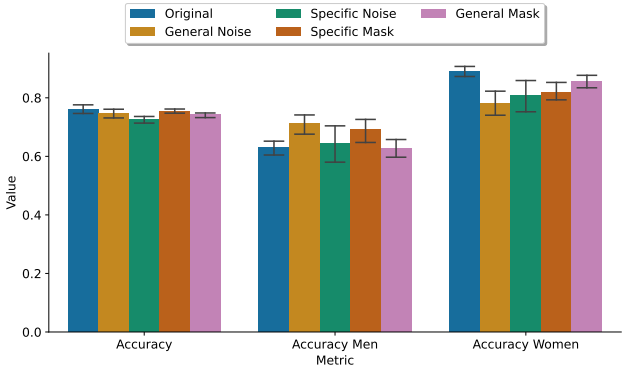


Figure 7. Comparison of noise addition techniques for the attribute Blond Hair using Quantile 0.7

Next to the type of noise, the quantile of information blocked is an important hyperparameter as well. Figure 8 shows how the accuracy of the subgroups evolves with the differing percentages. In the case of Specific Mask for the Blond Hair attribute, the optimal point seems to be the 70% mark. Masking out more reduces accuracy for women further while more specific noise additions increase the inequality in performance to that of unmodified data.

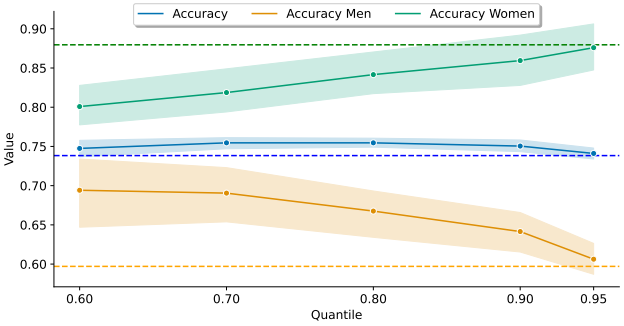


Figure 8. Comparison of metrics for the Specific Mask noise addition, for the attribute Blond Hair. Dotted lines indicate the accuracy of the original model for the same metric.

5. Conclusion

In conclusion, our paper introduces a novel method for mitigating bias in machine learning models, crucial for ensuring fairness and equity. We propose leveraging pixel image attributions to identify and regulate regions within images containing significant information about biased attributes. Our approach, employing a model-agnostic technique to extract pixel attributions through a CNN classifier trained on small image patches, enables the identification of critical information distribution across images. By utilizing these at-

tributions to introduce targeted noise into datasets with confounding attributes, we effectively prevent neural networks from learning biases and prioritize primary attributes. Our method demonstrates its effectiveness in training unbiased classifiers on heavily biased datasets, offering promise for enhancing fairness and equity in machine learning applications.

Acknowledgments

Sander De Coninck receives funding from the Special Research Fund of Ghent University under grant no. BOF22/DOC/093. This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme.

References

[1] Wilbert G Aguilar, Marco A Luna, Julio F Moya, Vanessa Abad, Hugo Ruiz, Humberto Parra, and William Lopez. Cascade classifiers and saliency maps based people detection. In *Augmented Reality, Virtual Reality, and Computer Graphics: 4th International Conference, AVR 2017, Ugento, Italy, June 12-15, 2017, Proceedings, Part II 4*, pages 501–510. Springer, 2017. 3

[2] Aidan Boyd, Kevin W. Bowyer, and Adam Czajka. Human-aided saliency maps improve generalization of deep learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2735–2744, 2022. 3

[3] Morris H DeGroot and Stephen E Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983. 3

[4] Emily Denton, Ben Hutchinson, Margaret Mitchell, Timnit Gebru, and Andrew Zaldivar. Image counterfactual sensitivity analysis for detecting unintended bias. *arXiv preprint arXiv:1906.06439*, 2019. 2

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 3

[6] Debasmita Ghose, Shasvat M. Desai, Sneha Bhattacharya, Deep Chakraborty, Madalina Fiterau, and Tauhidur Rahman. Pedestrian detection in thermal images using saliency maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019. 3

[7] Melissa Hall, Laurens van der Maaten, Laura Gustafson, Maxwell Jones, and Aaron Adcock. A systematic study of bias amplification. *arXiv preprint arXiv:2201.11706*, 2022. 7

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3, 6

[9] Linzhi Huang, Mei Wang, Jiahao Liang, Weihong Deng, Hongzhi Shi, Dongchao Wen, Yingjie Zhang, and Jian Zhao. Gradient attention balance network: Mitigating face recognition racial bias via gradient attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 38–47, 2023. 3

[10] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015. 6

[11] Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viégas, and Michael Terry. Xrai: Better attributions through regions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4948–4957, 2019. 3

[12] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1548–1558, 2021. 2, 3

[13] Rizwan Ahmed Khan, Alexandre Meyer, Hubert Konik, and Saida Bouakaz. Saliency-based framework for facial expression recognition. *Frontiers of Computer Science*, 13:183–198, 2019. 3

[14] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9012–9020, 2019. 2

[15] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 1, 2, 3

[16] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3, 6

[17] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017. 3

[18] Viraj Mavani, Shanmuganathan Raman, and Krishna P Miyapuram. Facial expression recognition using visual saliency and deep learning. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 2783–2788, 2017. 3

[19] Christoph Molnar. *Interpretable Machine Learning*. 2 edition, 2022. 3

[20] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, 2015. 3

[21] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632, 2005. 3

[22] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in big data*, 2:13, 2019. 1

[23] Otavio Parraga, Martin D More, Christian M Oliveira, Nathan S Gavenski, Lucas S Kupssinskü, Adilson

Medronha, Luis V Moura, Gabriel S Simões, and Rodrigo C Barros. Fairness in deep learning: A survey on vision and language research. *ACM Computing Surveys*, 2023. 1

[24] Amirarsalan Rajabi, Mehdi Yazdani-Jahromi, Ozlem Ozmen Garibay, and Gita Sukthankar. Through a fair looking-glass: mitigating bias in image datasets. In *International Conference on Human-Computer Interaction*, pages 446–459. Springer, 2023. 2

[25] Vikram V Ramaswamy, Sunnie SY Kim, and Olga Russakovsky. Fair attribute classification through latent space de-biasing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9301–9310, 2021. 2

[26] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016. 3

[27] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 3

[28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[29] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 6

[30] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017. 3

[31] Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. Causal attention for unbiased visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3091–3100, 2021. 2

[32] Seyma Yucer, Samet Akçay, Noura Al-Moubayed, and Toby P Breckon. Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 18–19, 2020. 2

[33] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018. 2