Turning a Blind Eye: Explicit Removal of Biases and Variation from Deep Neural Network Embeddings

Mohsan Alvi[1], Andrew Zisserman[1], and Christoffer Nellåker[2,3]

[1] Visual Geometry Group, Department of Engineering Science, University of Oxford
[2] Nuffield Department of Obstetrics and Gynaecology, University of Oxford
[3] Big Data Institute, University of Oxford

Abstract. Neural networks achieve the state-of-the-art in image classification tasks. However, they can encode spurious variations or biases that may be present in the training data. For example, training an age predictor on a dataset that is not balanced for gender can lead to gender biased predicitons (e.g. wrongly predicting that males are older if only elderly males are in the training set).
We present two distinct contributions:
1) An algorithm that can remove multiple sources of variation from the feature representation of a network. We demonstrate that this algorithm can be used to remove biases from the feature representation, and thereby improve classification accuracies, when training networks on extremely biased datasets.
2) An ancestral origin database of 14,000 images of individuals from East Asia, the Indian subcontinent, sub-Saharan Africa, and Western Europe.
We demonstrate on this dataset, for a number of facial attribute classification tasks, that we are able to remove racial biases from the network feature representation.

Keywords: Dataset bias · Face attribute classification · Ancestral origin dataset

1 Introduction

The current state-of-the-art image recognition algorithms are based on convolutional neural networks [1, 2, 3]. These networks rely on large datasets of labeled images, to simultaneously generate a feature representation and a decision framework. This approach removes the need to handcraft features for any given problem but also gives rise to the question as to what feature representation the network has actually learned.

These models are trained on large datasets that contain a number of biases [4, 5] or spurious variations, that are irrelevant, or even problematic, for a given task (e.g. discriminating by gender or ancestral origin). One such example is face recognition. Large publicly available face datasets are often composed of celebrities, such as [6, 7, 8, 9, 10, 11]. These can contain age, gender, and ancestral origin biases: for example, female celebrities tend to be younger than their male counterparts. This bias does not represent the real world outside of the movie business, reducing the usefulness of models trained on these datasets. Furthermore, in cases where large datasets are not available, training algorithms are initialized from networks that have been trained on similar tasks, for which more data is available [12, 13, 14]. This method, called fine-tuning, carries two potential issues: 1) spurious variations are learned from the small dataset, and 2) inheriting biases present in the large dataset.

The use of big data to train AI-models has, to name a few, been adopted by government agencies, institutions of law, human resources, and medical decision systems. A number of such models have been shown to make decisions based on the gender or ancestral origin of an individual, leading to concerns about their "fairness" [15, 16, 17]. With the recent enforcement of the General Data Protection Regulation laws[4], individuals have the right to know the rationale behind an automated decision concerning them. To continue the adoption of deep learning, more needs to be done to make neural networks more transparent. One way to approach this issue is to prevent models from making decisions for the wrong reasons.

To be sure that decisions are not being made due to biases, we must look beyond using accuracy as our only performance metric. An experiment by Zhao et al [18] showed that neural networks learn and amplify biases in the training data. Women were more often depicted in kitchens than men, so the network learned that being in the kitchen was a key feature for the identification of women. Though this may have been true for the dataset the classifier was trained on, in general, this is not a reliable indicator of the presence of a woman.

In this paper, we introduce an algorithm, inspired by a domain and task adaptation approach [19], that can be used to 1) to ensure that a network is blind to a known bias in the dataset, 2) improve the classification performance when faced with an extreme bias, and 3) remove multiple spurious variations from the feature representation of a primary classification task. We use age, gender, ancestral origin, and pose information for facial images to demonstrate our framework.

As discrimination by ancestral origin is a type of spurious variation in many tasks, we have created a new labeled ancestral origin dataset, "Labeled Ancestral Origin Faces in the Wild (LAOFIW)", from publicly available images. This dataset contains 14,000 images of individuals whose ancestral origin is sub-Saharan Africa, Indian Subcontinent, Europe, and East Asia, with a variety of poses, illumination, facial expressions.

The rest of this paper is organized as follows. Section 3 introduces our LAOFIW dataset and other datasets we used for our experiments. Section 4 discusses the methods to remove spurious variations from the feature representation of a network. In section 5.1, we present an experiment to remove a bias from the feature representation network. In section 5.2, we investigate how removing an extreme bias can improve classification performance, and in section 5.3, we investigate removing multiple spurious variations from a network. The results are detailed in section 6. Finally, section 7 summarises our findings.

2 Related Work

Image datasets are known to contain biases that cause models to generalize poorly to new, unseen data [4, 5]. This has been addressed by domain adaptation that aims to minimize the difference between the source and target domains [20]. Domain adaptation

[4] https://www.eugdpr.org

has been shown to improve the generalisability of classifiers for the same task across different domains [19, 21]. We draw inspiration from domain adaptation methods to make the network agnostic to spurious variations.

Learning feature representations that are invariant to certain spurious variations has been tackled in a number of ways. [22] take a multi-task learning approach to pose-invariant face recognition. They propose a novel approach to automatically weight each of the spurious variations during training. Another approach is to adjust the training data distribution at training time, to avoid learning biases [18]. This method relies on having labels for each of the spurious variations for each training instance. This is not feasible for most existing datasets, as they tend to be labeled with a single task in mind, where information about spurious variations is not available. Our method can make use of separate datasets, each labeled for distinct tasks, to remove multiple spurious variations simultaneously.

Jia et al [23] and Raff et al [24] draw inspiration from [25] to remove a source of variation with the use of a gradient reversal layer to update the network in opposition of a task. Instead of applying gradient reversal on the output of the softmax layer, which penalizes correct classifications, we compute the cross-entropy of the output classifier and a uniform distribution, as in [19, 21]. This ensures an equally uninformative classifier across all tasks.

3 Datasets

3.1 Labeled Ancestral Origin Faces in the Wild

A new ancestral origin database was created as part of this work called “Labeled Ancestral Origin Faces in the Wild (LAOFIW)”. The aim of this dataset was to 1) create a dataset for experimentation, and 2) be used as a spurious variation dataset for applications where training a network to be agnostic to ancestral origin is important. The database was assembled using the Bing Image Search API. Search terms based upon origin, e.g. “German, English, Polish, etc”, were submitted in conjunction with the words “man, woman; boy, girl”. Additionally, results were filtered to return photographic images of faces. In total, 43 origin search terms were queried returning 20,000 images. Duplicates were removed by comparing their Histogram of Oriented Gradients [26] encoded using a Gaussian mixture model for each image. The remaining 14,000 images, were manually divided into four broad ancestral origins: sub-Saharan Africa, Indian Subcontinent, Europe, and East Asia. These classes were selected on the basis of being visually distinct from each other.

The database contains roughly the same number of male and female individuals. It also contains images with varied poses: more than a third of the images are non-frontal. Sample images are shown in figure 1.

3.2 Age and Gender Dataset

The IMDB dataset is a large publicly available dataset of the images contained in the profiles of the 100,000 most popular actors on the IMDb website[5] with gender and date

[5] https://www.imdb.com/

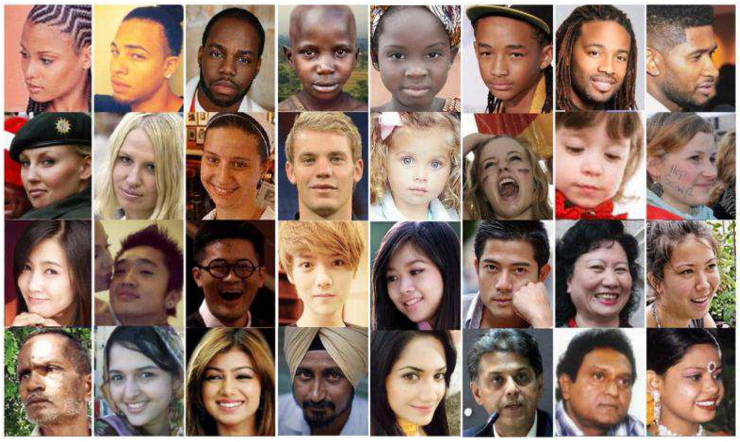


Fig. 1: Example images from the LAOFIW dataset for each of the four classes. In rows top to bottom: Sub-Saharan Africa, Europe, East Asia, Indian Subcontinent. The images are highly varied in age, gender, pose, lighting and expression.

of birth labels provided for each instance [27]. We used this dataset to investigate the effects of age and gender bias in celebrity datasets.

The labels in the IMDB dataset are noisy, with a number of individuals having both incorrect age and gender labels. This is due to the nature of how the data was collected: The authors assumed that images from an actor's profile that contained a single face would show the actor in question. As stated by the authors, however, these images often contain other actors, who co-starred in their movies [27]. As gender and date of birth are taken from the profile of the actor, this causes erroneous labels for images that show co-stars. The age of an individual was calculated as the difference between the timestamp of the photo and the date of birth of the subject. In some cases, the time stamps of photos predate an actor's date of birth or are otherwise unreliable.

To mitigate this problem, we used the Microsoft Azure Face API[6] to extract gender and age estimates for an identity-balanced subset of 150,000 images. We rejected all images in which the predicted gender from Azure and the IMDB gender label disagreed. We then ran the analysis for age and removed images in which the Azure age prediction differed by more than 10 years from the IMDB labels. The resulting, cleaned dataset contained 60,000 images.

In order to quantify the effectiveness of this data cleansing procedure, we trained a gender classification VGG-M network [3] on both the original 150,000 images and the cleaned 60,000 images, withholding 20% of the images for testing. The gender classification accuracies on the test images improved from 75%, before cleaning, to 99% after cleaning.

The distribution of ages for each gender in the cleaned IMDB dataset is shown in figure 2. We can observe a bias towards younger women and older men in the data.

[6] https://azure.microsoft.com/en-gb/services/cognitive-services/face/

A subset of the cleaned data was used to create an unbiased test dataset, which contains equal numbers of men and women for each age category. This unbiased test dataset was used to evaluate network bias and was not used during training.

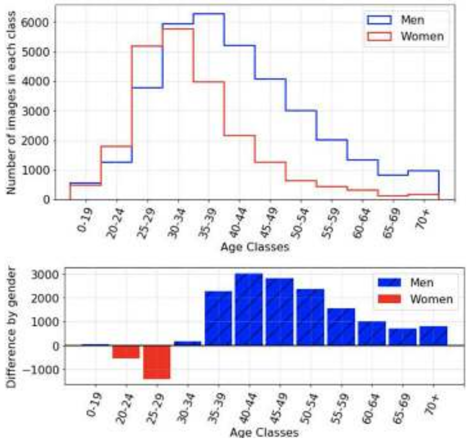


Fig. 2: Age distributions and differences for women (red) and men (blue) in the IMDB dataset. The dataset bias towards younger women and older men is visible.

3.3 Pose Dataset

We used the AFLW dataset [10] categorized by yaw, to train the pose classifier mentioned in section 3.1. The poses were binned into five categories, “profile left, near-frontal left, frontal, near-frontal right, profile right”. The non-frontal images were duplicated and flipped to augment the dataset. A class-balanced subset of 24,000 images was selected for training, and a class-balanced set of 6,000 images was reserved for testing.

4 Methods

We introduce a supervised-learning algorithm that aims to learn a single feature representation θ_{repr}, that is informative for a primary classification task, whilst simultaneously being uninformative for a number of spurious variations, that represent undesirable sources of variation. For example, we may wish to create an age classifier, that does not base its decisions on any pose, ancestral origin, or gender information.

We assume access to a primary dataset $\mathcal{D}_p = \{x_i, y_i\}_{i=1}^{n_p}$, containing n_p images, x_i, with labels $y_i \in \{1, ..., K\}$. And similarly, we assume access to M secondary datasets, $\mathcal{D}_s = \{\mathcal{D}_m\}_{m=1}^{M}$, each describing a single spurious variation.

4.1 Joint Learning and Unlearning

We introduce a joint learning and unlearning (JLU) algorithm to learn a primary classification task, whilst simultaneously unlearning multiple spurious variations.

Our convolutional network (CNN) architecture is depicted in Figure 3. The primary branch has a single classification loss, the primary classification loss, which assesses the ability of the network to accurately distinguish between classes of the primary task. Each secondary branch has two losses: a classification loss and a confusion loss. These losses are used to, in turn, assess the amount of spurious variation information left in the feature representation θ_{repr} and then remove it.

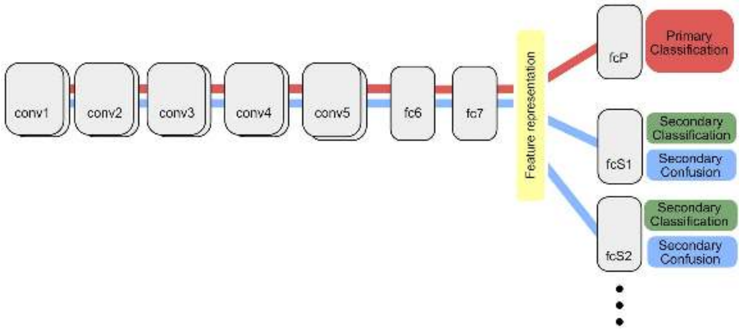


Fig. 3: Overview of CNN Architecture: We use a joint loss over primary and secondary data to learn a feature representation that simultaneously learns to classify the primary task but becomes invariant to secondary tasks, the spurious variations. The spurious variation classification loss and confusion loss act in opposition to learn the classifier on the feature embedding and change the feature embedding to confuse the classifier, respectively. The base architecture is a VGG-M network [3].

Let the classification objective for a generic task with K classes and a corresponding classifier θ_c, given a feature representation θ_{repr}, be defined as the standard softmax loss:

$$L_{\text{softmax}}(x, y; \theta_{\text{repr}}, \theta_C) = -\sum_{k=1}^{K} \mathbb{1}[y = k] \log p_k, \tag{1}$$

where p_k is the softmax of the classifier's output assigning the input x to class k. We will refer to this loss evaluated on the primary task as $L_P(x_p, y_p; \theta_{\text{repr}}, \theta_P)$ and on the m-th spurious variation as $L_m(x_m, y_m, \theta_{\text{repr}}; \theta_m)$.

Inspired by [19], we introduce a confusion loss for the m-th spurious variation, $L_{\text{conf}}(x_m, y_m, \theta_m; \theta_{\text{repr}})$, in (2). Minimizing the confusion loss seeks to change the feature representation θ_{repr}, such that it becomes invariant to the spurious variations.

$$L_{\text{conf,m}}(x_m, y_m, \theta_m; \theta_{\text{repr}}) = -\sum_{n_m} \frac{1}{n_m} \log p_{n_m}. \tag{2}$$

We compute the best classifier for each spurious variation, θ_m, and then compute the entropy between the output predicted from each of these classifiers and a uniform distribution. The complete method minimizes the joint loss function:

$$L(x_p, y_p, x_s, y_s, \theta_p, \theta_s, \theta_{\text{repr}}) = L_p(x_p, y_p; \theta_{\text{repr}}, \theta_p) \\ + L_s + \alpha L_{\text{conf}}, \tag{3}$$

where α determines how strongly the confusion loss affects the overall loss, $\theta_s = \{\theta_1, ..., \theta_M\}$, and

$$L_{\text{conf}} = \frac{1}{M} \sum_{m=1}^{M} L_{\text{conf},m}(x_m, y_m, \theta_m; \theta_{\text{repr}}) \tag{4}$$

$$L_s = \sum_{m=1}^{M} \beta_m L_m(x_m, y_m, \theta_{\text{repr}}; \theta_m), \tag{5}$$

where β_m is a weight assigned to the $m-$th spurious variation. As mentioned in [19], the confusion loss (4) and the spurious variation classification loss (5) stand in opposition to one another, so they cannot be optimized in the same step. Therefore, we switch between training the spurious variation classification loss, L_s, and then the joint primary and confusion loss, $L_p(x_p, y_p; \theta_{\text{repr}}, \theta_p) + \alpha L_{conf}$. At each iteration, we find the best spurious variation classifier for the feature representation. The training procedure is shown in algorithm 1. We used the VGG-M architecture [28], which consists of five convolutional layers (conv1-5) and three fully connected layers (fc6-8). The feature representation parameters θ_{repr} represents layers conv1-fc7.

Algorithm 1 Joint Learning and Unlearning

```
procedure min L($x_p, y_p, x_s, y_s, \theta_p, \theta_s, \theta_{\text{repr}}$)
    for epochs do
        while $\frac{dL_s}{d\theta_s} \neq 0$ do
            min $L_s$
        end while
        min $L_p + \alpha L_{\text{conf}}$
    end for
end procedure
```

5 Experiments

In this section, we present experiments to demonstrate possible applications of our methodology and the datasets we tested them on.

Removal of a bias from a network — We train a gender-agnostic age classifier using the IMDB dataset, which contains a gender bias: female celebrities tend to be younger than their male counterparts in this dataset.

Removal of an extreme bias from a network — We train an age-agnostic gender classifier on subsets of the IMDB dataset that contain only young women and only old men, and vice versa.

Simultaneous removal of multiple spurious variations — We demonstrate our JLU algorithm's ability to simultaneously remove multiple spurious variations from the feature representation of a network trained for a primary classification task.

5.1 Removal of a bias from a network

We investigated the task of creating an age classifier using the cleaned IMDB dataset described in section 3.2. We hypothesize that the gender bias in the distribution, where men are generally older than their female counterparts in this dataset, will be learned by the network.

We trained two networks to perform this task: 1) baseline — trained solely on age data, 2) blind — trained on age data, whilst removing gender-specific information from the network.

We evaluated both networks on the unbiased test dataset detailed in section 3.2. We compared accuracies for age classification and prediction distributions for both genders.

5.2 Removal of an extreme bias from a network

We train gender classifiers using the following artificially biased subsets of the cleaned IMDB data from section 3.2:

1. Extreme bias 1 (EB1): women aged 0-30, men aged 40+
2. Extreme bias 1 (EB2): women aged 40+, men aged 0-30

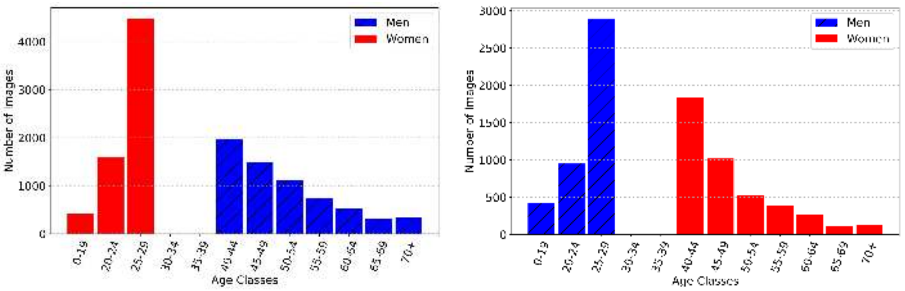
We placed a buffer of 10 years without any individuals between the two sets to exaggerate the age difference between the genders. We hypothesize that the age bias in the distribution will be learned by the network, for example, if trained on EB1, it should falsely learn to predict that young men are in fact women. Histograms of these datasets are given in figure 4.

We trained a baseline network on gender only, and a blind network using our JLU algorithm to unlearn age. The trained classifiers were evaluated on the unbiased test dataset detailed in section 3.2.

5.3 Simultaneous removal of multiple spurious variations

We demonstrate our algorithm's ability to remove multiple spurious variations from the feature representation of a network. The primary task and spurious variations are selected from age, gender, ancestral origin and pose. In each case, one task is selected as the primary task, and the others make up the spurious variations that we wish to remove. A baseline network was trained for each primary task without using our JLU algorithm.

We used the cleaned IMDB dataset described in section 3.2 for age and gender labels. We used our LAOFIW dataset, described in section 3.1 to classify ancestral origin. Finally, we used the adapted AFLW dataset (Section 3.3) to classify pose.



(a) EB1: women aged 0-30, men aged 40+

(b) EB2: women aged 40+, men aged 0-30

Fig. 4: Histograms showing the ages of women (red) and men (blue) in the artificially biased datasets. In both cases, individuals aged under 35 were selected for the young, and individuals aged above 45 were selected for the old set.

5.4 Age Classification

Since our dataset is relatively small for a deep learning task, we do not approach age estimation as a regression task. We conduct age classification by creating age bins of 5 years, or more years at either end of the distribution. The categories are shown in figure 2. We define predictions within one class from the true age class as a positive classification to account for errors caused by edge cases in different bins.

5.5 Implementation details

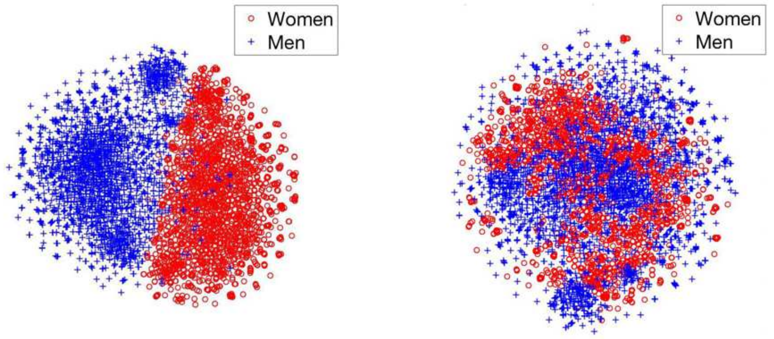
Our base network is an adapted VGG-M network [3], pre-trained on the VGG-Face dataset [11]. In our experiments, we saw no significant improvement in updating the weights in the convolutional layers, therefore, the weights in layers conv1-conv5 were frozen for all experiments. This significantly increased the speed of the training algorithm.

The confusion loss was approximated using the Kullback-Leibler divergence between the softmax output of a classifier and a uniform distribution. The network is trained using stochastic gradient descent (SGD) with a learning rate of 1×10^{-4}. Learning rates of classification layers for each task were boosted by a factor of 10. The hyperparameter in equation 3 is set to a value of $\alpha = 0.1$. The spurious variation classifier hyperparameters were all set to $\beta_m = 1$. To address the imbalance in the class distributions of the training data, losses were weighted by the inverse of the relative frequency of that class. We conduct our experiments using the MatConvNet framework [29].

6 Results

6.1 Removal of a bias from a dataset

We computed feature representations for the class-balanced test dataset from section 3.2 for both baseline and blind networks. T-SNE visualizations of these feature embeddings are shown in figure 5. The feature representation of the baseline network that was trained to classify age is clearly separable by gender, demonstrating that the bias in the training data was learned. After unlearning gender, the feature representation is no longer separable by gender, demonstrating that this bias has been removed.



(a) Feature representation of baseline network

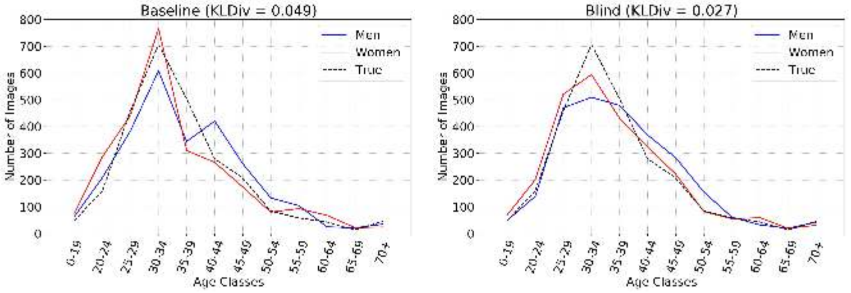
(b) Feature representation of blind network

Fig. 5: T-SNE visualizations of 4096-dimensional feature representation of the class-balanced test dataset for age classification networks trained on gender-biased data. (a) Baseline network without unlearning gender, (b) Blind network with gender unlearning. The feature representation is clearly separable by gender for the baseline network, showing that the network has learned the gender bias in the dataset. After unlearning, this bias is no longer pronounced.

The distributions of the age predictions and ground truth on the gender-balanced test set are shown in figure 6. The red and blue lines show the predicted ages for women and men, respectively. The black dashed line shows the ground truth for both genders. The distributions of predicted ages for females and males are different for the baseline network (Figure 5a): women's ages tend to be underestimated, whereas men's ages are generally overestimated. This behaviour mimics the bias within the training data, which confirms that the classification model is making age predictions that are, to a degree, dependent on gender.

In order for us to be sure that the network is not using gender information to predict age, both female and male prediction distributions should be similar. Figure 5b shows the distributions for the network that has been trained to no longer be able to differentiate between genders. The KL-divergence between age prediction distributions for men and women of the biased and unbiased networks are 0.049 and 0.027, respectively. The reduction in KL-divergence shows that the network has successfully unlearned gender, as the predicted age distributions for both genders are similar. Note, that we are not necessarily trying to perfectly predict the ground truth, but be confident that we are not treating men and women differently.

The average prediction accuracy for age classification on the gender-balanced test dataset for the baseline network was 78.9% (78.4% for females and 79.4% for males) and for the unbiased network was 78.1% (77.4% for females and 78.9% for males). This corresponds to a reduction in accuracy of 0.8% (1.2% for females and 0.5% for males).



(a) Baseline network

(b) Unbiased network

Fig. 6: Prediction distributions of age for each gender compared to the ground truth (black) for men (blue) and women (red). The evaluation dataset was gender-balanced so that each class has an equal number of women and men. (a) Baseline network without unlearning gender, (b) unbiased network with gender unlearning. The red line shows the age predictions for women and the blue line shows them for men. The prediction distributions for women and men align after unlearning gender. The KL-divergence score between the prediction distributions of men and women are given for each case. The KL-Divergence score reduces, as two distributions become more similar.

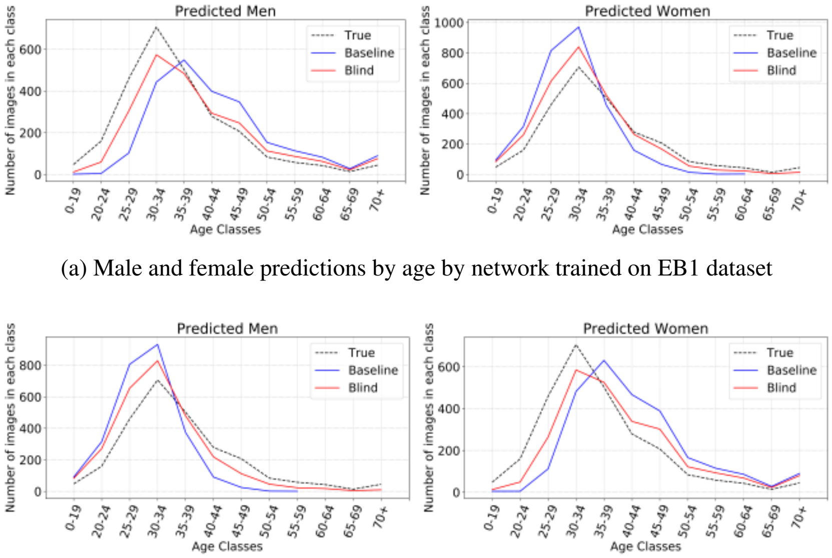
6.2 Removal of an extreme bias from a network

The baseline gender classification accuracy on the gender-balanced test data was 70% for a network trained on the EB1 dataset. This was improved by 16% to 86% for the blind network, where we simultaneously unlearned age information. The accuracies for networks trained on the EB2 dataset were 62% for the baseline and 82% for the blind network. This amounts to a 20% increase in classification accuracy.

The age distributions of individuals that were predicted to be either male or female, trained on the EB1 dataset, are shown in figure 7. The baseline network often wrongly predicts older individuals to be male, and younger individuals to be female, in line with the training data. The age distribution of gender predictions from the JLU network, however, is closer to the true distribution. We observed similar results for the baseline network trained on the EB2 dataset, where younger individuals were more often predicted to be male, and older individuals were more often predicted to be female. The JLU network prediction distributions were closer to the true distributions.

6.3 Simultaneous removal of multiple spurious variations

Figure 8 shows how the accuracies on the test data for each attribute vary over the JLU training procedure when simultaneoulsy unlearning multiple spurious variations. For



(a) Male and female predictions by age by network trained on EB1 dataset

(b) Male and female predictions by age by network trained on EB2 dataset

Fig. 7: Age distributions of male and female predictions from networks trained on the biased (a) EB1 and (b) EB2 datasets, evaluated on the gender-balanced test data. The baseline network often wrongly predicts older individuals to be male and younger individuals to be female. Using the JLU algorithm reduces this bias and shifts the prediction distributions closer to the true distribution.

clarity, in the figure, spurious variation classification accuracies were rescaled using the equation below:

$$a = 1 - \frac{e}{e_{\text{max}}}, \tag{6}$$

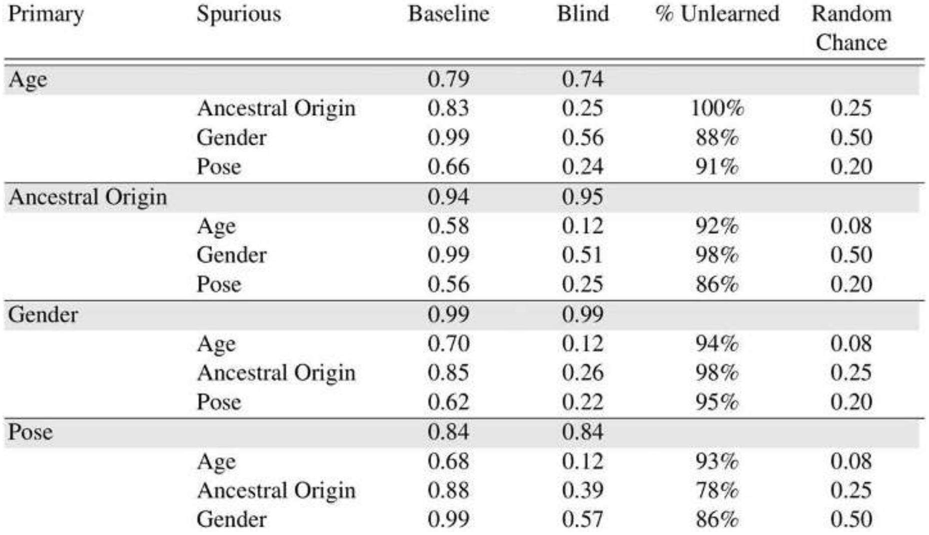
where e is the mean-class error of the classifier and e_{max} corresponds to the error-rate of a classifier that draws at random. A perfect classifier corresponds to a score of $a = 1$ and a random classifier corresponds to a score of $a = 0$. The primary classification accuracy was not rescaled.

These results are summarised in table 1 for networks trained with and without JLU. The baseline column corresponds to the accuracy of the best classifier on a feature embedding trained without JLU. The "blind" column shows the same accuracies when using the JLU algorithm. When the classifier cannot recover meaningful information from the feature embedding the accuracy is equivalent to random chance.

Age — The primary classification accuracy of the blind network is 5% less than the baseline network. Information about ancestral origin was removed from the network completely. Gender and pose information were removed by 88% and 91%, respectively.

Ancestral Origin — The primary classification accuracy of the blind network improved by 1% compared to the baseline network. The proportions of age, gender, and pose information that were unlearned were 92%, 98%, and 86%, respectively.

Table 1: Classification accuracies for each spurious variation for a baseline network and a “blind network”. The baseline accuracies show the best classifiers that can be learned on the feature representation of the primary task without the JLU algorithm. The “Blind” column shows the classification accuracies after JLU. The random chance column states the target accuracy for a spurious variation classifier after JLU, corresponding to an uninformative classifier’s accuracy. The % unlearned column shows the relative difference of baseline and blind accuracies to random chance.



Primary	Spurious	Baseline	Blind	% Unlearned	Random Chance
Age		0.79	0.74		
	Ancestral Origin	0.83	0.25	100%	0.25
	Gender	0.99	0.56	88%	0.50
	Pose	0.66	0.24	91%	0.20
Ancestral Origin		0.94	0.95		
	Age	0.58	0.12	92%	0.08
	Gender	0.99	0.51	98%	0.50
	Pose	0.56	0.25	86%	0.20
Gender		0.99	0.99		
	Age	0.70	0.12	94%	0.08
	Ancestral Origin	0.85	0.26	98%	0.25
	Pose	0.62	0.22	95%	0.20
Pose		0.84	0.84		
	Age	0.68	0.12	93%	0.08
	Ancestral Origin	0.88	0.39	78%	0.25
	Gender	0.99	0.57	86%	0.50

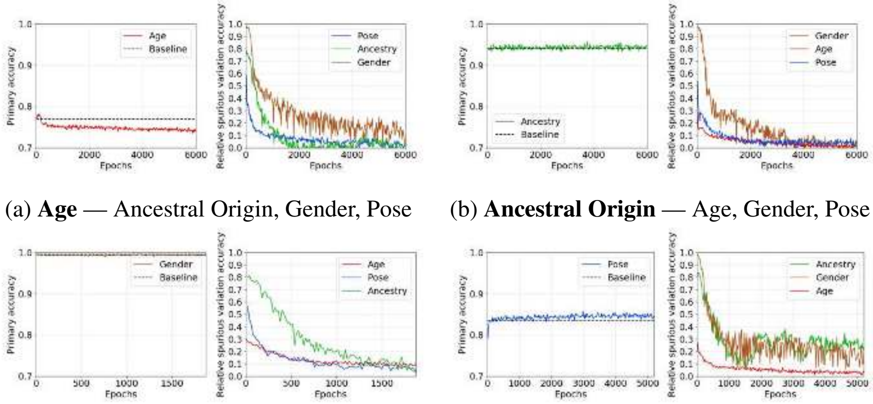
Gender — The primary classification accuracy of the blind network is the same as the baseline network. Age, ancestral origin, and pose information were removed by 94%, 98% and 95%, respectively.

Pose — The primary classification accuracy of the blind network is the same as the baseline network. Information about age, ancestral origin, and gender were removed from the network by 93%, 78% and 86%, respectively.

Apart from the network that was trained on the primary task of age classification on data that contained a strong gender bias, all blind networks have the same or better primary classification accuracy compared to their baseline counterparts. Spurious variations were successfully removed in most cases, with classification accuracies reducing to within 5% of random chance in 9/12 cases. The pose experiment proved the least effective, with ancestral origin information proving most difficult to remove.

7 Conclusion

The paper proposes an approach for removing multiple known dataset biases or spurious variations from a deep neural network. Similar to previous work in the field, the algorithm is inspired by domain adaptation work. Drawing inspiration from [19], we



(a) **Age** — Ancestral Origin, Gender, Pose

(b) **Ancestral Origin** — Age, Gender, Pose

(c) **Gender** — Age, Ancestral Origin, Pose

(d) **Pose** — Age, Ancestral Origin, Gender

Fig. 8: Classification accuracies on test data for the primary task and spurious variations during the training procedure of the JLU algorithm. The primary classification accuracy is compared to the baseline network accuracy (dotted). The spurious variation accuracies have been re-scaled using equation 6 so that zero accuracy corresponds to the accuracy of a random chance classifier.

compare the cross-entropy between the output distribution of classifiers that are trained to predict spurious variations and a uniform distribution. The resulting feature representation is informative for the primary task but blind to one or more spurious variations.

We demonstrated our algorithm's efficacy on face classification tasks of age, ancestral origin, pose and gender. The resulting feature representations remained informative for one task, whilst simultaneously being uninformative for the others. When training a gender classification network on extremely age-biased data, our algorithm significantly improves (by up to 20%) classification accuracies on an unbiased test dataset. This demonstrates that our algorithm allows networks trained on biased data to generalize better to unbiased settings, by removing each known bias from the feature representation of the network.

This is a significant step towards trusting that a network definitely isn't basing its decisions on the wrong reasons. With increasing use of neural networks in government, law, and employment to make life-changing decisions, it is of great importance, that undesirable social biases are not encoded in the decision algorithm. We also created a dataset to detect ancestral origin from faces, which can be used to remove racial biases from the feature representation of a network. We will make this dataset available to the public.

Some spurious variations are easier to remove than others — applying different weights to each spurious variation classifier could account for these differences. A dynamic-weighting scheme, as in [22], to automatically weight these different tasks at different time-points during training may improve convergence.

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. ArXiv e-prints (December 2015)
2. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. (2012) 1097–1105
3. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. ArXiv e-prints (September 2014)
4. Tommasi, T., Patricia, N., Caputo, B., Tuytelaars, T.: A Deeper Look at Dataset Bias. ArXiv e-prints (May 2015)
5. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: in CVPR. (2011)
6. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: VGGFace2: A dataset for recognising faces across pose and age. 8
7. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: Challenge of recognizing one million celebrities in the real world. (02 2016)
8. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst (October 2007)
9. Kemelmacher, I., Seitz, S., Miller, D., Brossard, E.: The megaface benchmark: 1 million faces for recognition at scale. (12 2015)
10. Kostinger, M., Wohlhart, P., Roth, P.M., Bischof, H.: Annotated Facial Landmarks in the Wild: A large-scale, real-world database for facial landmark localization, IEEE (November 2011) 2144–2151
11. Parkhi, O., Vedaldi, A., Zisserman, A.: Deep Face Recognition. BMVC **1**(3) (September 2015) 6
12. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. ArXiv e-prints (November 2013)
13. Long, M., Cao, Y., Wang, J., Jordan, M.I.: Learning Transferable Features with Deep Adaptation Networks. ArXiv e-prints (February 2015)
14. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks, IEEE (June 2014) 1717–1724
15. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In Friedler, S.A., Wilson, C., eds.: Proceedings of the 1st Conference on Fairness, Accountability and Transparency. Volume 81 of Proceedings of Machine Learning Research., New York, NY, USA, PMLR (23–24 Feb 2018) 77–91
16. Ras, G., van Gerven, M., Haselager, P.: Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges. ArXiv e-prints (March 2018)
17. Sweeney, L.: Discrimination in Online Ad Delivery. ArXiv e-prints (January 2013)
18. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W.: Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, Association for Computational Linguistics (September 2017) 2979–2989
19. Tzeng, E., Hoffman, J., Darrell, T., Saenko, K.: Simultaneous deep transfer across domains and tasks. In: Computer Vision (ICCV), 2015 IEEE International Conference on, IEEE (2015) 4068–4076
20. Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., Schlkopf, B.: Covariate shift and local learning by distribution matching. In: Dataset Shift in Machine Learning. Biologische Kybernetik, Cambridge, MA, USA (2009) 131–160
21. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial Discriminative Domain Adaptation. ArXiv e-prints (February 2017)

22. Yin, X., Liu, X.: Multi-Task Convolutional Neural Network for Pose-Invariant Face Recognition. IEEE Transactions on Image Processing **27**(2) (February 2018) 964–975 arXiv: 1702.04710.
23. Jia, S., Lansdall-Welfare, T., Cristianini, N.: Right for the Right Reason: Training Agnostic Networks. ArXiv e-prints (Jun 2018)
24. Raff, E., Sylvester, J.: Gradient Reversal Against Discrimination. ArXiv e-prints (July 2018)
25. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-Adversarial Training of Neural Networks. In Csurka, G., ed.: Domain Adaptation in Computer Vision Applications. Springer International Publishing, Cham (2017) 189–209
26. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01. CVPR '05, Washington, DC, USA, IEEE Computer Society (2005) 886–893
27. Rothe, R., Timofte, R., Gool, L.V.: DEX: Deep EXpectation of Apparent Age from a Single Image, IEEE (December 2015) 252–257
28. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the Devil in the Details: Delving Deep into Convolutional Nets, British Machine Vision Association (2014) 6.1–6.12
29. Vedaldi, A., Lenc, K.: Matconvnet – convolutional neural networks for matlab. In: Proceeding of the ACM Int. Conf. on Multimedia. (2015)

Acknowledgments

This research was financially supported by the EPSRC programme grant Seebibyte EP/M013774/1, the EPSRC EP/G036861/1, and the MRC Grant MR/M014568/1.