

Learning Not to Learn: Training Deep Neural Networks with Biased Data

Byungju Kim¹ Hyunwoo Kim² Kyungsu Kim³ Sungjin Kim³ Junmo Kim¹
 School of Electrical Engineering, KAIST, South Korea¹
 Beijing Institute of Technology²
 Samsung Research³

{byungju.kim, junmo.kim}@kaist.ac.kr

{hwkim}@bit.edu.cn

{ks0326.kim, sj9373.kim}@samsung.com

Abstract

We propose a novel regularization algorithm to train deep neural networks, in which data at training time is severely biased. Since a neural network efficiently learns data distribution, a network is likely to learn the bias information to categorize input data. It leads to poor performance at test time, if the bias is, in fact, irrelevant to the categorization. In this paper, we formulate a regularization loss based on mutual information between feature embedding and bias. Based on the idea of minimizing this mutual information, we propose an iterative algorithm to unlearn the bias information. We employ an additional network to predict the bias distribution and train the network adversarially against the feature embedding network. At the end of learning, the bias prediction network is not able to predict the bias not because it is poorly trained, but because the feature embedding network successfully unlearns the bias information. We also demonstrate quantitative and qualitative experimental results which show that our algorithm effectively removes the bias information from feature embedding.

1. Introduction

Machine learning algorithms and artificial intelligence have been used in wide ranging fields. The growing variety of applications has resulted in great demand for robust algorithms. The most ideal way to robustly train a neural network is to use suitable data free of bias. However great effort is often required to collect well-distributed data. Moreover, there is a lack of consensus as to what constitutes well-distributed data.

Apart from the philosophical problem, the data distribution significantly affects the characteristics of networks, as current deep learning-based algorithms learn directly from the input data. If biased data is provided during training,

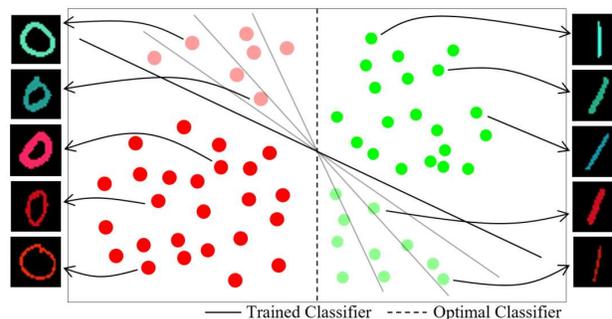


Figure 1. Detrimental effect of biased data. The points colored with high saturation indicate samples provided during training, while the points with low saturation would appear in test scenario. Although the classifier is well-trained to categorize the training data, it performs poorly with test samples because the classifier learns the latent bias in the training samples.

the machine perceives the biased distribution as meaningful information. This perception is crucial because it weakens the robustness of the algorithm and unjust discrimination can be introduced.

A similar concept has been explored in the literature and is referred to as *unknowns* [3]. The authors categorized unknowns as follows: *known unknowns* and *unknown unknowns*. The key criterion differentiating these categories is the confidence of the predictions made by the trained models. The unknown unknowns correspond to data points that the model’s predictions are wrong with high confidence, e.g. high softmax score, whereas the known unknowns represent mispredicted data points with low confidence. Known unknowns have better chance to be detected as the classifier’s confidence is low, whereas unknown unknowns are much difficult to detect as the classifier generates high confidence score.

In this study, the data bias we consider has a similar flavor to the unknown unknowns in [3]. However, unlike the unknown unknowns in [3], the bias does not represent data points themselves. Instead, bias represents some attributes

of data points, such as color, race, or gender.

Figure 1 conceptually shows how biased data can affect an algorithm. The horizontal axis represents shape space of the digits, while the vertical axis represents color space, which is biased information for digit categorization. In practice, shape and color are independent features, so a data point can appear anywhere in Figure 1. However, let us assume that only the data points with high saturation are provided during training, but the points with low saturation are present in the test scenario (yet are not accessible during the training). If a machine learns to categorize the digits, each solid line is a proper choice for the decision boundary. Every decision boundary categorizes the training data perfectly, but it performs poorly on the points with low saturation. Without additional information, learning of the decision boundary is an ill-posed problem, multiple decision boundaries can be determined that perfectly categorize the training data. Moreover, it is likely that a machine would utilize the color feature because it is a simple feature to extract.

To fit the decision boundary to the optimal classifier in Figure 1, we require simple prior information: *Do not learn from color distribution*. To this end, we propose a novel regularization loss, based on mutual information, to train deep neural networks, which prevents learning of a given bias. In other words, we regulate a network to minimize the mutual information shared between the extracted feature and the bias we want to unlearn. Hereafter, the bias that we intend to unlearn is referred to the *target bias*. For example, the target bias is the color in Figure 1. Prior to the unlearning of target bias, we assume that the existence of data bias is known and that the relevant meta-data, such as statistics or additional labels corresponding to the semantics of the biases are accessible. Then, the problem can be formulated in terms of an adversarial problem. In this scenario, one network has been trained to predict the target bias. The other network has been trained to predict the label, which is the main objective of the network, while minimizing the mutual information between the embedded feature and the target bias. Through this adversarial training process, the network can learn how to predict labels independent of the target bias.

Our main contributions can be summarized as follows: Firstly, we propose a novel regularization term, based on mutual information, to unlearn target bias from the given data. Secondly, we experimentally show that the proposed regularization term minimizes the detrimental effects of bias in the data. By removing information relating to the target bias from feature embedding, the network was able to learn more informative features for classification. In all experiments, networks trained with the proposed regularization loss showed performance improvements. Moreover, they achieved the best performance in the most experi-

ments. Lastly, we propose bias planting protocols for public datasets. To evaluate bias removal problem, we intentionally planted bias to training set while maintaining test set unbiased.

2. Related Works

The existence of unknown unknowns was experimentally demonstrated by Attenberg *et al.* in [3]. The authors separated the decisions rendered by predictive models into four conceptual categories: known knowns, known unknowns, unknown knowns, and unknown unknowns. Subsequently, the authors developed and participated in a “*beat the machine* challenge”, which challenged the participants to manually find the unknown unknowns to fool the machine.

Several approaches for identifying unknown unknowns have been also proposed [13, 4]. Lakkaraju *et al.* [13] proposed an automatic algorithm using the explore-exploit strategy. Bansal and Weld proposed a coverage-based utility model that evaluates the coverage of discovered unknown unknowns [4]. These approaches rely on an oracle for a subset of test queries. Rather than relying on an oracle, Alvi *et al.* [1] proposed joint learning and unlearning method to remove bias from neural network embedding. To *unlearn* the bias, the authors applied confusion loss, which is computed by calculating the cross-entropy between classifier output and a uniform distribution. Similar approaches, making networks to be confused, have been applied on various applications [8, 2, 6].

As mentioned by Alvi *et al.* in the paper [1], the unsupervised domain adaptation (UDA) problem is closely related to the biased data problem. The UDA problem involves generalizing the network embedding over different domains [8, 23, 21]. The main difference between our problem and the UDA problem is that our problem does not assume the access to the target images and instead, we are aware of the description of the target bias.

Embracing the UDA problem, disentangling feature representation has been widely researched in the literature. The application of disentangled features has been explored in detail [24, 17]. Using generative adversarial network [9], more methods to learn disentangled representation [5, 16, 22] have been proposed. In particular, Chen *et al.* proposed the InfoGAN [5] method, which learns and preserves semantic context without supervision.

These studies highlighted the importance of feature disentanglement, which is the first step in understanding the information contained within the feature. Inspired by various applications, we have attempted to remove certain information from the feature. In contrast to the InfoGan [5], we minimize the mutual information in order *not* to learn. However, removal of information is an antithetical concept to learning and is also referred to as *unlearning*. Although

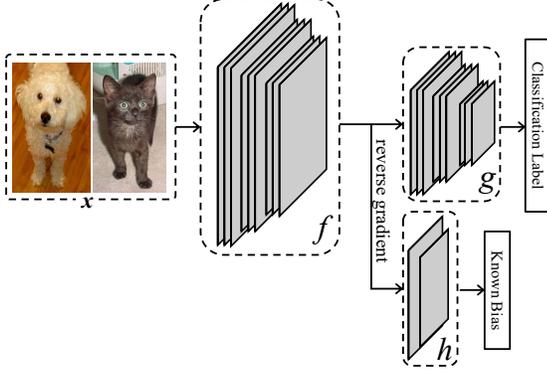


Figure 2. Overall architecture of deep neural network. The network $g \circ f$ is implemented with ResNet-18 [10] for real images and plain network with four convolution layers for MNIST images.

the concept itself is the complete opposite of learning, it can help learning algorithms. Herein, we describe an algorithm for removing target information and present experimental results and analysis to support the proposed algorithm.

3. Problem Statement

In this section, we formulate a novel regularization loss, which minimizes the undesirable effects of biased data, and describe the training procedure. The notations should be defined prior to introduction of the formulation. Unless specifically mentioned, all notation refers to the following terms hereafter. Assume we have an image $x \in \mathcal{X}$ and corresponding label $y_x \in \mathcal{Y}$. We define a set of bias, \mathcal{B} , which contains every possible target bias that \mathcal{X} can possess. In Figure 1, \mathcal{B} is a set of possible colors, while \mathcal{Y} represents a set of digit classes. We also define a latent function $b : \mathcal{X} \rightarrow \mathcal{B}$, where $b(x)$ denotes the target bias of x . We define random variables X and Y that have the value of x and y_x respectively.

The input image x is fed into the feature extraction network $f : \mathcal{X} \rightarrow \mathbb{R}^K$, where K is the dimension of the feature embedded by f . Subsequently, the extracted feature, $f(x)$, is fed forward through both the label prediction network $g : \mathbb{R}^K \rightarrow \mathcal{Y}$, and bias prediction network $h : \mathbb{R}^K \rightarrow \mathcal{B}$. The parameters of each network are defined as θ_f, θ_g , and θ_h with the subscripts indicating their specific network. Figure 2 describes the overall architecture of the neural networks. However, we do not explicitly designate a detailed architecture, since our regularization loss is applicable to arbitrary network architectures.

3.1. Formulation

The objective of our work is to train a network that performs robustly with unbiased data during test time, even though the network is trained with biased data. The data

bias has following characteristic:

$$\mathcal{I}(b(X^{train}); Y) \gg \mathcal{I}(b(X^{test}); Y) \approx 0, \quad (1)$$

where X^{train} and X^{test} denote the random variable sampled during the training and test procedure, respectively, and $\mathcal{I}(\cdot; \cdot)$ denotes the mutual information. Biased training data results in the biased networks, so that the network relies heavily on the bias of the data:

$$\mathcal{I}(b(X); g(f(X))) \gg 0. \quad (2)$$

To this end, we add the mutual information to the objective function for training networks. We minimize the mutual information over $f(X)$, instead of $g(f(X))$. It is adequate because the label prediction network, g , takes $f(X)$ as its input. From a standpoint of g , the training data is not biased if the network f extracts no information of the target bias. In other words, extracted feature $f(x)$ should contain no information of the target bias, $b(x)$. Therefore, the training procedure is to optimize the following problem:

$$\min_{\theta_f, \theta_g} \mathbb{E}_{\tilde{x} \sim P_{\mathcal{X}}(\cdot)} [\mathcal{L}_c(y_{\tilde{x}}, g(f(\tilde{x})))] + \lambda \mathcal{I}(b(X); f(X)), \quad (3)$$

where $\mathcal{L}_c(\cdot, \cdot)$ represents the cross-entropy loss, and λ is a hyper-parameter to balance the terms.

The mutual information in Eq. (3) can be equivalently expressed as follows:

$$\mathcal{I}(b(X); f(X)) = H(b(X)) - H(b(X)|f(X)), \quad (4)$$

where $H(\cdot)$ and $H(\cdot|\cdot)$ denote the marginal and conditional entropy, respectively. Since the marginal entropy of bias is constant that does not depend on θ_f and θ_g , $H(b(X))$ can be omitted from the optimization problem, and we try to minimize the negative entropy, $-H(b(X)|f(X))$. Eq. (4) is difficult to directly minimize as it requires the posterior distribution, $P(b(X)|f(X))$. Since it is not tractable in practice, minimizing the Eq. (4) is reformulated using an auxiliary distribution, Q , with an additional equality constraint:

$$\begin{aligned} \min_{\theta_f} \mathbb{E}_{\tilde{x} \sim P_{\mathcal{X}}(\cdot)} [\mathbb{E}_{\tilde{b} \sim Q(\cdot|f(\tilde{x}))} [\log Q(\tilde{b}|f(\tilde{x}))]] \\ \text{s.t. } Q(b(X)|f(X)) = P(b(X)|f(X)). \end{aligned} \quad (5)$$

The benefit of using the distribution Q is that we can directly calculate the objective function. Therefore, we can train the feature extraction network, f , under the equality constraint.

3.2. Training Procedure

As the equality constraint in Eq. (5) is difficult to meet (especially in the beginning of the training process), we modify the equality constraint into minimizing KL divergence between P and Q , so that Q gets closer to P as learning progresses. We relax the Eq. (5), so that the auxiliary

distribution, Q , could be used to approximate the posterior distribution. The relaxed regularization loss, \mathcal{L}_{MI} , is as follows:

$$\mathcal{L}_{MI} = \mathbb{E}_{\tilde{x} \sim P_X(\cdot)} [\mathbb{E}_{\tilde{b} \sim Q(\cdot|f(\tilde{x}))} [\log Q(\tilde{b}|f(\tilde{x}))]] + \mu D_{KL}(P(b(X)|f(X)) || Q(b(X)|f(X))), \quad (6)$$

where D_{KL} denotes the KL-divergence and μ is hyperparameter which balances the two terms. Similar to the method proposed by Chen *et al.* [5], we parametrize the auxiliary distribution, Q , as the bias prediction network, h . Note that we will train network h , so that the KL-divergence is minimized. Provided that the distribution Q implemented by network h converges to $P(b(X)|f(X))$, we only need to train network f so that the first term in Eq. (6) is minimized.

Although the posterior distribution, $P(b(X)|f(X))$, is not tractable, the bias prediction network, h , is expected to be trained to stochastically approximate $P(b(X)|f(X))$, if we train the network with $b(X)$ as the label with SGD optimizer. Therefore, we relax the KL-divergence of Eq. (6) with expectation of the cross-entropy loss between $b(X)$ and $h(f(X))$, and we train network h so that bias prediction loss, \mathcal{L}_B , is minimized.

$$\mathcal{L}_B(\theta_f, \theta_h) = \mathbb{E}_{\tilde{x} \sim P_X(\cdot)} [\mathcal{L}_c(b(\tilde{x}), h(f(\tilde{x})))] \quad (7)$$

Although training network h alone to minimize Eq. (7) is enough to make Q closer to P , it will be additionally beneficial to train f to maximize Eq. (7) in an adversarial way, i.e. to let the networks f and h play the minimax game. The intuition is that the feature extracted by network f is making the bias prediction difficult. As f is trained to minimize the first term in Eq. (6), we can reformulate Eq. (6) using \mathcal{L}_B instead of KL-divergence as follows:

$$\min_{\theta_f} \max_{\theta_h} \mathbb{E}_{\tilde{x} \sim P_X(\cdot)} [\mathbb{E}_{\tilde{b} \sim Q(\cdot|f(\tilde{x}))} [\log Q(\tilde{b}|f(\tilde{x}))]] - \mu \mathcal{L}_B(\theta_f, \theta_h) \quad (8)$$

We train h to correctly predict the bias, $b(X)$, from its feature embedding, $f(X)$. We train f to minimize the negative conditional entropy. The network h is fixed while minimizing the negative conditional entropy. The network f is also trained to maximize the cross-entropy to restrain h from predicting $b(X)$. Together with the primal classification problem, the minimax game is formulated as follows:

$$\min_{\theta_f, \theta_g} \max_{\theta_h} \mathbb{E}_{\tilde{x} \sim P_X(\cdot)} [\mathcal{L}_c(y_{\tilde{x}}, g(f(\tilde{x}))) + \lambda \mathbb{E}_{\tilde{b} \sim Q(\cdot|f(\tilde{x}))} [\log Q(\tilde{b}|f(\tilde{x}))]] - \mu \mathcal{L}_B(\tilde{x}, \theta_f, \theta_h) \quad (9)$$

In practice, the deep neural networks, f , g and h , are trained with both adversarial strategy [9, 5] and gradient reversal technique [8]. Early in learning, $g \circ f$ are rapidly

trained to classify the label using the bias information. Then h learns to predict the bias, and f begins to learn how to extract feature embedding independent of the bias. At the end of the training, h regresses to the poor performing network not because the bias prediction network, h , diverges, but because f unlearns the bias, so the feature embedding, $f(X)$, does not have enough information to predict the target bias.

4. Dataset

Most existing benchmarks are designed to evaluate a specific problem. The collectors often split the dataset into train/test sets exquisitely. However, their efforts to maintain the train/test split to obtain an identical distribution obscures our experiment. Thus, we intentionally planted bias to well-balanced public benchmarks to determine whether our algorithm could unlearn the bias.

4.1. Colored MNIST

We planted a color bias into the MNIST dataset [14]. To synthesize the color bias, we selected ten distinct colors and assigned them to each digit category as their mean color. Then, for each training image, we randomly sampled a color from the normal distribution of the corresponding mean color and provided variance, and colored the digit. Since the variance of the normal distribution is a parameter that can be controlled, the amount of the color bias in the data can be adjusted. For each test image, we randomly choose a mean color among the ten pre-defined colors and followed the same colorization protocol as for the training images. Each sub-datasets are denoted as follows:

- Train- σ^2 : Train images with colors sampled with σ^2
- Test- σ^2 : Test images with colors sampled with σ^2

Since the digits in the test sets are colored with random mean colors, the Test- σ^2 sets are unbiased. We varied σ^2 from 0.02 to 0.05 with a 0.005 interval. Smaller values of σ^2 indicate more bias in the set. Thus, Train-0.02 is the most biased set, whereas Train-0.05 is the least biased.

Figure 3 (a) shows samples from the colored MNIST, where the images in the training set show that the color and digit class are highly correlated. The color of the digit contains sufficient information to categorize the digits in the training set, but it is insufficient for the images in the test set. Recognizing the color would rather disrupt the digit categorization. Therefore, the color information must be removed from the feature embedding.

4.2. Dogs and Cats

We evaluated our algorithm with the dogs and cats database, developed by *kaggle* [12]. The original database is a set of 25K images of dogs and cats for training and 12,500 images for testing. Similar to [13], we manually



Figure 3. Examples of datasets with intentionally planted bias. (a) We modified the MNIST data [14] to plant color bias to training images. A mean color has been designated for each class, so a classifier can easily predict the digit with color. (b) TB1 is a set of bright dogs and dark cats, whereas TB2 contains dark dogs and bright cats. Similar to the colored MNIST, a classifier can predict whether an image is dog or cat with its color. (c) IMDB face dataset contains age and gender labels. EB1 and EB2 differ on the correlation between age and gender. Predicting age enables an algorithm to predict gender.

categorized the data according to the color of the animal: bright, dark, and other. Subsequently, we split the images into three subsets.

- Train-biased 1 (TB1) : bright dogs and dark cats.
- Train-biased 2 (TB2) : dark dogs and bright cats.
- Test set: All 12,500 images from the original test set.

The images categorized as *other* are images featuring white cats with dark brown stripes or dalmatians. They were not used in our training sets due to their ambiguity. In turn, TB1 and TB2 contain 10,047 and 6,738 images respectively. The constructed dogs and cats dataset is shown in Figure 3 (b), with each set containing a color bias. For this dataset, the bias set $\mathcal{B} = \{\text{dark, bright}\}$. Unlike TB1 and TB2, the test set does not contain color bias.

On the other hand, the ground truth labels for test images are not accessible, as the data is originally for competition [12]. Therefore, we trained an oracle network (ResNet-18 [10]) with all 25K training images. For the test set, we measured the performance based on the result from the oracle network. We presumed that the oracle network could accurately predict the label.

4.3. IMDB Face

The IMDB face dataset [19] is a publicly available face image dataset. It contains 460,723 face images from 20,284 celebrities along with information regarding their age and gender. Each image in the IMDB face dataset is a cropped facial image. As mentioned in [19, 1], the provided label contains significant noise. To filter out misannotated images, we used pretrained networks [15] on Adience benchmark [7] designed for age and gender classification. Using the pretrained networks, we estimated the age and gender for all the individuals shown in the images in the IMDB

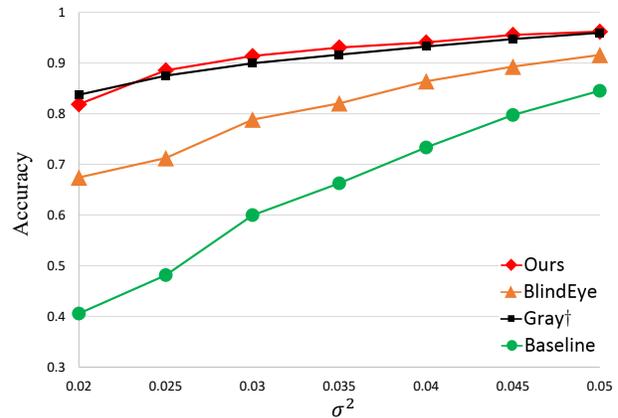


Figure 4. Evaluation results on colored MNIST dataset. † denotes that it is evaluated with grayscale-converted images. The model denoted as Gray was trained with images converted into grayscale; it is trained with significantly mitigated bias. Compared to the baseline and BlindEye algorithm [1], our model shows outperforming results. Note that our result shows comparable performance with grayscale model. It implies that the network was successfully trained to extract feature embedding a lot more independent of the bias.

face dataset. We then collected images where the both age and gender labels match with the estimation. From this, we obtained a cleaned dataset with 112,340 face images, and the detailed cleaning procedure is described in the supplementary material.

Similar to the protocol from [1], we classified the cleaned IMDB images into three biased subsets. We first withheld 20% of the cleaned IMDB images as the test set, then split the rest of the images as follows:

- Extreme bias 1 (EB1): women aged 0-29, men aged 40+
- Extreme bias 2 (EB2): women aged 40+, men aged 0-29
- Test set: 20% of the cleaned images aged 0-29 or 40+

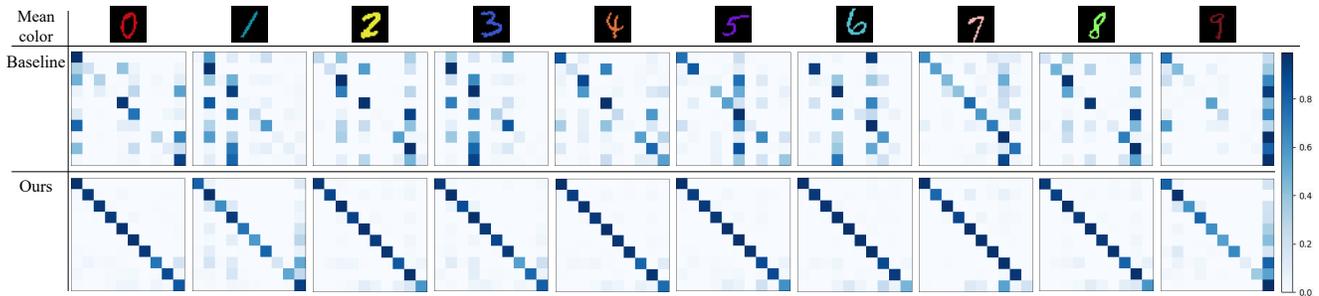


Figure 5. Confusion matrices with test images colored by single mean color. Top row denotes the mean colors and their corresponding digit classes in training data. The confusion matrices of baseline model show the network is biased owing to the biased data. On the contrary, the networks trained by our algorithm are not biased to the color although they were trained with the same training data with the baseline.

As a result, EB1 and EB2 contain 36,004 and 16,800 facial images respectively, and the test set contains 13129 images. Figure 3 (c) shows that both EB1 and EB2 are biased with respect to the age. Although it is not as clear as the color bias in Figure 3 (a) and (b), EB1 consists of younger female and older male celebrities, whereas EB2 consists of younger male and older female celebrities. When gender is target bias, $\mathcal{B} = \{\text{male, female}\}$, and when age is target bias, \mathcal{B} is their age.

5. Experiments

5.1. Implementation

In the following experiments, we removed three types of target bias: color, age, and gender. The age and gender labels were provided in IMDB face dataset, therefore $\mathcal{L}_{\mathcal{B}}(\theta_f, \theta_h)$ was optimized with supervision. On the other hand, the color bias was removed via self-supervision. To construct color labels, we first sub-sampled the images by factor of 4. In addition, the dynamic range of color, 0-255, was quantized into eight even levels.

For the network architecture, we used ResNet-18 [10] for real images and plain network with four convolution layers for the colored MNIST experiments. The network architectures correspond to the parametrization of $g \circ f$. In the case we used ResNet-18, g was implemented as two residual blocks on the top, while f represents the rest. For plain network for colored MNIST, both g and f consist of two convolution layers. ResNet-18 was pretrained with Imagenet data [20] except for the last fully connected layer. We implemented h with two convolution layers for color bias and single fully connected layer for gender and age bias. Every convolution layer is followed by batch normalization [11] and ReLU activation layers. All the evaluation results were averaged to be presented in this paper.

5.2. Results

We compare our training algorithm with other methods that can be used for this task. The performance of the al-

gorithms mentioned in this section were re-implemented based on the literature.

Colored MNIST. The amount of bias in the data was controlled by adjusting the value of σ^2 . A network was trained for each σ^2 value from 0.02 to 0.05 and was evaluated with the corresponding test set with the same σ^2 . Since a color for each image was sampled with a given σ^2 , smaller σ^2 implies severer color bias. Figure 4 shows the evaluation results of the colored MNIST. The baseline model represents a network trained without additional regularization and the baseline performance can roughly be used as an indication of training data bias. The algorithm denoted as “Blind-Eye” represents a network trained with confusion loss [1] instead of our regularization. The other algorithm, denoted as “Gray”, represents a network trained with grayscale images and it was also tested with grayscale images. For the given color biased data, we converted the color digits into grayscale. Conversion into grayscale is a trivial approach that can be used to mitigate the color bias. We presume that the conversion into grayscale does not reduce the information significantly since the MNIST dataset was originally provided in grayscale.

The results of our proposed algorithm outperformed the BlindEye [1] and baseline model with all values of σ^2 . Notably, we achieved similar performance as the model trained and tested with grayscale images. Since we converted images in both training and test time, the network is much less biased. In most experiments, our model performed slightly better than the gray algorithm, suggesting that our regularization algorithm can effectively remove the target bias and encourage a network to extract more informative features.

To analyze the effect of the bias and proposed algorithm, we re-colored the test images. We sampled with the same protocol, but with fixed mean color, which was assigned to one of the ten digit classes of the biased training data. Figure 5 shows the confusion matrices drawn by the baseline and our models with the re-colored test images. The digits illustrated in the top row denotes the mean colors and their corresponding digit class in training set. For example, the

first digit, red zero, signifies the confusion matrices below are drawn by test images colored reddish regardless of their true label. It also stands for a fact that every digit of category *zero* in training data is colored reddish.

In Figure 5, the matrices of the baseline show vertical patterns, some of which are shared, such as digits 1 and 3. The mean color for class 1 is *teal*; in RGB space it is (0, 128, 128). The mean color for class 3 is similar to that of class 1. In RGB space, it is (0, 149, 182) and is called *bondi blue*. This indicates that the baseline network is biased to the color of digit. As observed from the figure, the confusion matrices drawn by our algorithm (bottom row) show that the color bias was removed.

Dogs and Cats. Table 1 presents the evaluation results, where the baseline networks perform admirably, considering the complexity of the task due to the pretrained parameters. As mentioned in [18], neural networks prefer to categorize images based on shape rather than color. This encourages the baseline network to learn shapes, but the evaluation results presented in Table 1 imply that the networks remain biased without regularization.

Similar to the experiment on the colored MNIST, simplest approach for reducing the color bias is to convert the images into grayscale. Unlike the MNIST dataset, conversion would remove a significant amount of information. Although the networks for grayscale images performed better than the baseline, Table 1 shows that the networks remain biased to color. This is likely because of the criterion that was used to implant the color bias. Since the original dataset is categorized into bright and dark, the converted images contain a bias in terms of brightness.

We used gradient reversal layer (GRL) [8] and adversarial training strategy [5, 9] as components of our optimization process. To analyze the effect of each component, we ablated the GRL from our algorithm. We also trained networks with both confusion loss [1] and GRL, since they can be used in conjunction with each other. Although the GRL was originally proposed to solve unsupervised domain adaptation problem [8], Table 1 shows that it is beneficial for bias removal. Together with either confusion loss or our regularization, we obtained the performance improvements. Furthermore, GRL alone notably improved the performance suggesting that GRL itself is able to remove bias.

Figure 6 shows the qualitative effect of our proposed regularization. The prediction results of the baseline networks do not change significantly regardless of whether the query image is cat or dog if the colors are similar. If a network is trained with TB1, the network predicts a dark image to be a cat and a bright image to be a dog. If another network is trained with TB2, the network predicts a bright image to be a cat and a dark image to be a dog. This implies that the baseline networks are biased to color. On the other hand, networks trained with our proposed algorithm successfully

Method	Trained on TB1		Trained on TB2	
	TB2	Test	TB1	Test
Baseline	.7498	.9254	.6645	.8524
Gray†	.8366	.9483	.7192	.8687
BlindEye [1]	.8525	.9517	.7812	.9038
GRL [8]	.8356	.9462	.7813	.9012
BlindEye+GRL	<u>.8937</u>	.9582	.8610	.9291
Ours-adv	.8853	<u>.9594</u>	<u>.8630</u>	<u>.9298</u>
Ours	.9029	.9638	.8726	.9376

Table 1. The evaluation results on dogs and cats dataset. All networks were evaluated with test set. Moreover, the networks trained with TB1 were additionally evaluated with TB2, and vice versa. † denotes that the network was tested with images converted into grayscale. The Ours-adv denotes a model trained with Eq. (9) without using gradient reversal layer. The best performing result on each column is denoted as boldface and the second best result is underlined.

Method	Trained on EB1		Trained on EB2	
	EB2	Test	EB1	Test
Learn Gender, Unlearn Age				
Baseline	.5986	.8442	.5784	.6975
BlindEye [1]	.6374	.8556	.5733	.6990
Ours	.6800	.8666	.6418	.7450
Learn Age, Unlearn Gender				
Baseline	.5430	.7717	.4891	.6197
BlindEye [1]	.6680	.7513	.6416	.6240
Ours	.6527	.7743	.6218	.6304

Table 2. Evaluation results on IMDB face dataset. All networks were evaluated with test set and the other training set. The best performing result on each column is denoted as boldface.

classified the query images independent of their colors. In particular, Figure 6 (c) and (f) were identically predicted by the baseline networks depending on their color. After removing the color information from the feature embedding, the images were correctly categorized according to their appearance.

IMDB face. For the IMDB face dataset, we conducted two experiments; one to train the networks to classify age independent of gender, and one to train the networks to classify gender independent of age. Table 2 shows the evaluation results from both experiments. The networks were trained with either EB1 or EB2 and since they are extremely biased, the baseline networks are also biased. By removing the target bias information from the feature embedding, overall performances are improved. On the other hand, considering that gender classification is a two class problem, where random guessing achieves 50% accuracy, the networks perform poorly on gender classification. Although Table 2 shows that the performance improves after removing the target bias from the feature embedding, the performance improvement achieved using our algorithm is marginal compared to

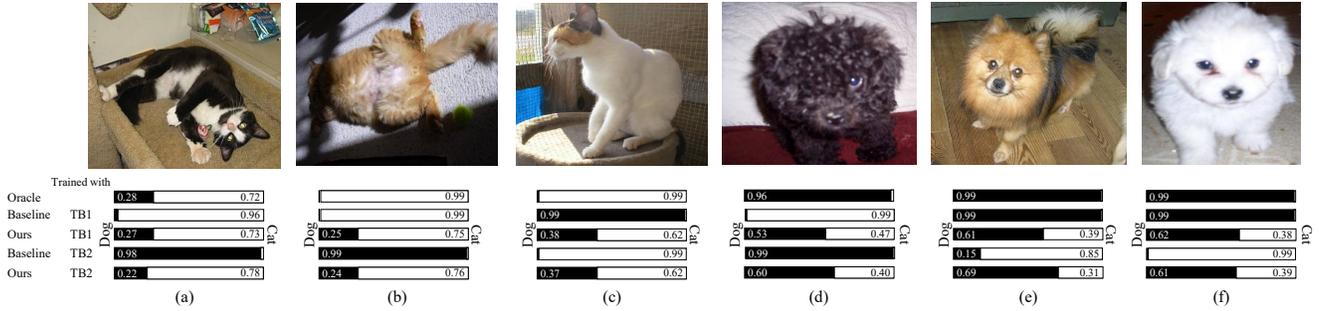


Figure 6. Qualitative results of dogs and cats dataset. The oracle model was trained with not only both TB1 and TB2, but also with images we categorized as *other color*. For test images, prediction results of the oracle model were considered as their true labels. The stacked bar charts below the figures visualize prediction results by each model. The baseline models tend to predict depending on the color, whereas our model ignores the color information.

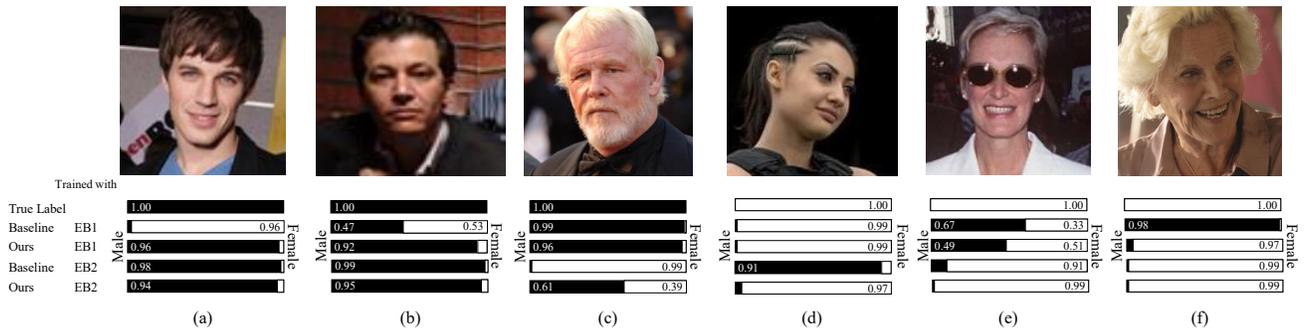


Figure 7. Qualitative results of gender classification with IMDB face dataset. As in the Figure 6, the stacked bar charts represent the prediction results. They show that the baseline models are biased to the age. On the other hand, the networks trained with proposed algorithm predict the gender independent of their age.

previous experiments with other datasets. We presume that this is because of the correlation between age and gender. In the case of color bias, the bias itself is completely independent of the categories. In other words, an effort to unlearn the bias is purely beneficial for digit categorization. Thus, removing color bias from feature embedding improved the performance significantly because the network is able to focus on learning shape features. Unlike the color bias, age and gender are not completely independent features. Therefore, removing bias information from feature embedding would not be completely beneficial. This suggests that a deep understanding of the specific data bias must precede the removal of bias.

Figure 7 shows the qualitative effect of regularization on the gender classification task. Young, mid-age, and old individuals with both male and female are presented. Similar to Figure 6, it implies that the baseline networks are biased toward age. The baseline network trained with EB1 predicted both young male and young female images (Figure 7 (a) and (d)) as female with high confidence. Meanwhile, the network trained with EB2 predicted the same images as the exact opposite gender with high confidence. Upon removal of age bias, the networks were trained to correctly predict the gender.

6. Conclusion

In this paper, we propose a novel regularization term to train deep neural networks when using biased data. The core idea of using mutual information is inspired by InfoGan [5]. In contrast to the inspiring approach, we rather minimize the mutual information in order *not* to learn. By letting networks play minimax game, networks learn to categorize, while unlearning the bias. The experimental results showed that the networks trained with the proposed regularization can extract bias-independent feature embedding, achieving the best performance in the most of the experiments. Furthermore, our model performed better than “Gray” model which was trained with almost unbiased data, indicating the feature embedding becomes even more informative. To conclude, we have demonstrated in this paper that the proposed regularization improves the performance of neural networks trained with biased data. We expect this study to expand the usage of various data and to contribute to the field of feature disentanglement.

Acknowledgement

This research was supported by Samsung Research.

References

- [1] Mohsan Alvi, Andrew Zisserman, and Christoffer Nellaker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. *arXiv preprint arXiv:1809.02169*, 2018. [2](#), [5](#), [6](#), [7](#)
- [2] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *The European Conference on Computer Vision (ECCV)*, September 2018. [2](#)
- [3] Joshua Attenberg, Panos Ipeirotis, and Foster Provost. Beat the machine: Challenging humans to find a predictive model’s unknown unknowns. *Journal of Data and Information Quality (JDIQ)*, 6(1):1, 2015. [1](#), [2](#)
- [4] Gagan Bansal and Daniel S Weld. A coverage-based utility model for identifying unknown unknowns. In *Proc. of AAAI*, 2018. [2](#)
- [5] Xi Chen, Yan Duan, Rein Houthoofd, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *CoRR*, abs/1606.03657, 2016. [2](#), [4](#), [7](#), [8](#)
- [6] Abhimanyu Dubey, Otkrist Gupta, Pei Guo, Ramesh Raskar, Ryan Farrell, and Nikhil Naik. Pairwise confusion for fine-grained visual classification. In *The European Conference on Computer Vision (ECCV)*, September 2018. [2](#)
- [7] E. Eidinger, R. Enbar, and T. Hassner. Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security*, 9(12):2170–2179, Dec 2014. [5](#)
- [8] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016. [2](#), [4](#), [7](#)
- [9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. [2](#), [4](#), [7](#)
- [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. [3](#), [5](#), [6](#)
- [11] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015. [6](#)
- [12] Kaggle. Dogs vs. cats, 2013. [4](#), [5](#)
- [13] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Eric Horvitz. Discovering blind spots of predictive models: Representations and policies for guided exploration. *CoRR*, abs/1610.09064, 2016. [2](#), [4](#)
- [14] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. [4](#), [5](#)
- [15] G. Levi and T. Hassner. Age and gender classification using convolutional neural networks. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 34–42, June 2015. [5](#)
- [16] Yang Liu, Zhaowen Wang, Hailin Jin, and Ian Waisell. Multi-task adversarial network for disentangled feature learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. [2](#)
- [17] Xi Peng, Xiang Yu, Kihyuk Sohn, Dimitris N. Metaxas, and Manmohan Chandraker. Reconstruction for feature disentanglement in pose-invariant face recognition. *CoRR*, abs/1702.03041, 2017. [2](#)
- [18] Samuel Ritter, David GT Barrett, Adam Santoro, and Matt M Botvinick. Cognitive psychology for deep neural networks: A shape bias case study. *arXiv preprint arXiv:1706.08606*, 2017. [7](#)
- [19] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, December 2015. [5](#)
- [20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. [6](#)
- [21] Ozan Sener, Hyun Oh Song, Ashutosh Saxena, and Silvio Savarese. Learning transferrable representations for unsupervised domain adaptation. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2110–2118. Curran Associates, Inc., 2016. [2](#)
- [22] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *In Proceeding of IEEE Computer Vision and Pattern Recognition*, Honolulu, HI, July 2017. [2](#)
- [23] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. *CoRR*, abs/1510.02192, 2015. [2](#)
- [24] Junho Yim, Heechul Jung, ByungIn Yoo, Changkyu Choi, Dusik Park, and Junmo Kim. Rotating your face using multi-task deep neural network. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 676–684, June 2015. [2](#)