Gender Fairness in Multi-label Classification using Auxiliary Tasks

Andrew Choi
205348339
asjchoi@ucla.edu

Zeyu Zhang
505030513
zeyuzhang@ucla.edu

Gaohong Liu
705352121
cheimu@ucla.edu

Abstract

As machine learning systems continue to vastly integrate themselves within society, it is important that decisions produced by such systems are fair. Such requirements are even more crucial when dealing with decisions that have lasting social, legal, and/or financial implications towards individuals. Unfortunately, it has been shown that bias is often encoded into machine learning systems due to preexisting dataset bias. Additionally, often times, it has been shown that many learning algorithms will produce predictors that amplify this bias effectively worsening the problem. One well-used dataset which has been found to contain a large amount of gender bias is the MS-COCO dataset. In this work, we propose a framework for reducing bias amplification by utilizing a descriptive model. Using this descriptive model, we then incorporate an auxiliary task into the objective function to aid the learning algorithm. We will then compare the results to a baseline approach as well as preexisting techniques to show that the bias reduction is comparable.

1 Introduction

In recent years, deep learning systems have demonstrated tremendous potential in their ability to perform complex cognitive tasks that were previously thought to be only achievable by humans. These systems can recognize objects [5], locate objects [8], reason about relations [3] and play complex video games [18]. However, recent studies indicate that systems trained on selected databases might suffer from training set bias and produce unfair predictions. This issue is especially evident when we consider the task where minority groups are involved. For example, there existed an instance where an image classifier once labeled an Africa American as a gorilla which led to social outrage [1]. Occurrences such as this have led to the emerging field of ensuring machine learning fairness. The term fairness is oftentimes subjective with numerous possible interpretations [11]. In this project, we try to develop a general learning framework that attempts to ensure fairness by mitigating gender-bias originating from the training date set. The intuitive idea is that by firstly using a convolution neural network, we can generate the image embedding of a dataset containing bias; secondly, by using a descriptive model, we can then describe and understand the inner dependencies of each image feature such as actions and gender bias from the generated embeddings; then we can leverage auxiliary task(s) in the objective function to aid learning an unbiased predictor. Not only could auxiliary tasks alleviate biases, it could also give an explanation as to what biases are discerned. By using a descriptive model and adding auxiliary task(s) into the objective function, we tested our method on the MS-COCO dataset [9] and showed that gender bias amplification for a multi-label classification problem was reduced by 53.82% when compared to a baseline model.

Preprint. Work in progress.

2 Related Work

Machine learning techniques is wildly adopted in many arenas of our life, which sometimes may exert negative influence or discrimination due to implicit biases in training data [12]. In word embeddings, verbs such as "cooking" have been seen to be heavily biased towards females when compared to males. Therefore, it is important to take these issues into consideration when designing learning algorithms. To mitigate the above mentioned issues, researchers use a variety of tools to reduce dataset bias. For example, Zhao et al.[21] proposed to use corpus-level constraints to limit the output of the inference results. Zhao et al. [22] also proposed to learn a gender neutral word embeddings. Yang et al. [19] proposed to use causality to reduce dataset bias.

Mutli-task learning [4] through auxiliary tasks [2] has demonstrated having a more robust and consistent representation. Researchers use this approach to achieve SOTA results on different learning tasks. For example, Sun et. al. [16] used auxiliary tasks to generate programs from video demonstrations. Li et al. [7] proposed to generate explanations while perform VQA tasks. This approach not only produced more accurate results, it also generated an explanation to users which increased users' trust in the system as a side effect. However, these systems treat the explanation process in the same manner as image captioning which is also heavily affected by dataset bias. Therefore, to overcome this issue we propose that a good learning system should not only provide explanations, but also know what to explain.

Two distinct approaches are developed in debiasing, one manipulates training data, the other adjusts the algorithm itself. In data manipulation, [22] proposed a data augmentation method to balance a disproportionate class by creating an augmented data set which is identical to the original data set and offsets the minority class. [17] proposed tagging techniques to handle the gendered training data set. [13] introduced the bias fine-tuning approach which incorporates transfer learning from an unbiased data set to minimize bias before fine-tuning a model on a biased data set. Some approaches focused on adjusting algorithms to debias a biased training data set. [21] proposed the Reducing Bias Amplification algorithm which is based on a constrained conditional model [14]. [20] mitigates bias by utilizing generative adversarial network to protect gendered data set.

A descriptive model [15] specifies the probability distribution of the signal, based on an energy function defined on the signal through some descriptive feature statistics extracted from the signal. The descriptive models belong to the broader class of energy-based models [6] that include non-probabilistic models as well as models with latent variables. Descriptive models are useful to describe inner relations between different image features inside the image because by utilizing sophisticated energy functions to describe the dependency of different image features, a single probability model can integrate all statistical measures of different image features instead of simply using simple product of the likelihoods or marginals of each feature. It can be learned in the unsupervised setting, where Y_i are not observed. The linear form of the descriptive model is an exponential family model. It specifies a probability distribution on the signal X_i via an energy function that is a linear combination of the features. Specifically, the learning of descriptive models require MCMC sampling of the synthesized signals.

3 Method

3.1 Identifying the Bias

We follow the definitions found in [21], where the bias is defined based on the correlation between the predicted labels $y \in \{o_1, o_2, ..., O_N\}$ and a certain demographic class $g \in G$, for example in our settings $g \in \{\texttt{man}, \texttt{woman}\}$. Then the bias of a label o towards a demographic class g is defined as

$$\text{bias}(o, g) = \frac{c(o, g)}{\sum_{g' \in G} c(o, g')}$$

where $c(o, g)$ defines the number of occurrences of label o and demographic class g over the data set. For example in our project, we focus on the demographic class $g \in \{\texttt{man}, \texttt{woman}\}$. Then the gender bias of the label o towards the class man is calculated as follow

$$\text{bias}(o, \texttt{man}) = \frac{c(o, \texttt{man})}{c(o, \texttt{man}) + c(o, \texttt{woman})}$$

To analyze the amplification of the bias of the learned model, we can compare the bias score on the training data set and the bias score from the prediction outputs of the learned model over the testing set. If the bias score from the predicted outputs over the testing set is larger than the bias score from the ground-truth over the training set, we say the model amplifies the bias towards a certain demographic class.

3.2 Bias Alignment using Auxiliary Tasks

In this section, we introduce how we leverage auxiliary tasks to avoid bias amplification during the training process, such that the model itself is capable of balancing the bias towards certain demographic classes. In this project, we take the multi-label classification as our main task, that is, how to avoid the bias amplification in a multi-label classification task.

The main task is the multi-label classification problem. A set of images that associated several pre-defined labels forms our training data set. The task is that given an input x, we need to predicts a set of labels that is most likely associated with the input x. This problem is formulated as an empirical risk minimization problem over the training set, which minimizes the disagreement between the predicted labels and the ground-truth labels. We adopt categorical cross entropy in the loss function which is given by

$$\mathcal{L}_{main} = \frac{1}{M} \sum_{i=1}^{M} \sum_{y_k} \; - y_k \log(p_{y_k}(x_i))$$

where y_k denotes the label ($y_k = 1$ if the labels is assigned), and $p_{y_k}(\cdot)$ is the probability of assigning label y_k. The probability of assigning a label y_k is obtained via softmax function which is given by

$$p_{y_k}(x) = \frac{h_{y_k}(x;\theta)}{\sum_{y_j} h_{y_j}(x;\theta)}$$

where $h(\cdot)$ is the model we learned for multi-label classification task, $h_{y_k}(\cdot)$ returns the score for the label y_k, and θ is the parameter of the model.

In order to prevent bias amplification during the training process, we want to align the gender bias between the unseen data and the training set. To achieve this, we minimize the difference between the expectation of the gender bias over the distribution of input space x and the actual gender bias over the training set. Thereby, we define the auxiliary task as follow

$$\mathcal{L}_{aux}^{y_k} = \left| \mathbb{E}_{p(x)}[\mathbb{1}(h_{y_k}(x;\theta), \text{gender})] - \frac{1}{M} \sum_{i=1}^{M} \mathbb{1}(y_{i,k}, \text{gender}) \right|$$

where $\mathcal{L}_{aux}^{y_k}$ measures the difference between expected bias and bias over the training set towards a certain label y_k, and $\mathbb{1}(\cdot)$ is the indicator function which returns 1 if label y_k and a certain type of gender (`man` or `woman`) occurs together. Then the overall loss function is given by

$$\mathcal{L} = \mathcal{L}_{main} + \lambda \sum_{y_k} \mathcal{L}_{aux}^{y_k}$$

where λ is a scalar that balance the trades off between the main task and the auxiliary tasks. For the auxiliary tasks, we simply sum over the bias scores of all labels. The way we calculate the expectation is introduced in the next section.

3.3 Descriptor Modeling

In this section, we introduce how we calculate the expectation $\mathbb{E}_{p(x)}[\mathbb{1}(h_{y_k}(x;\theta), \text{gender})]$. In order to calculate the expectation, we need to obtain the distribution of x, that is $p(x)$, from the training data set. A straight-forward idea is that we use filters (i.e., features) the peek the projection of a high dimensional space. The statistics of these filter responses are estimated by the sample means,

$$\mu_{obs}^{(\alpha)} = \frac{1}{M} \sum_{i=1}^{M} \phi^{(\alpha)}(x_i), \; \forall \alpha = 1, 2, ..., K$$

If we have enough training samples, then the sample averages $\{\mu_{obs}^{(\alpha)}, \alpha = 1, 2, ..., K\}$ make reasonable estimates for the expectations $\mathbb{E}_{f(x)}[\phi^{\alpha}(x)]$, where $f(x)$ is true distribution of x.

To approximate $f(x)$, a probability model $p(x)$ is restricted to reproduce the observed statistics (i.e., $\mathbb{E}_{f(x)}[\phi^{\alpha}(x)] = \mu_{obs}^{(\alpha)}$ for all $\alpha = 1, 2, ...K$). Let

$$\Omega = \{p(x) : \mathbb{E}_{p(x)}[\phi^{\alpha}(x)] = \mu_{obs}^{(\alpha)},\ \forall \alpha = 1, 2, ..., K\}$$

be the set of distributions that reproduce the observed features, then we need to select a $p(x) \in \Omega$ if $\Omega \neq \emptyset$. According to the maximum entropy principle, we should choose $p(x)$ that achieves the maximum entropy to obtain the purest observed statistics. Thereby, the problem becomes a constrained optimization problem [23] as follow,

$$\begin{aligned} &\arg\max_{p(x)} -\int p(x)\log p(x)\ \mathrm{d}x \\ &\text{s.t. } \mathbb{E}_p[\phi^{(\alpha)}(x)] = \mu_{obs}^{(a)},\ \forall \alpha = 1, 2, ..., K \\ &\int p(x)\ \mathrm{d}x = 1 \end{aligned}$$

By applying the Lagrange multipliers, we can formulate the preceding problem as an unconstrained problem, then we can apply the Euler-Lagrange equation to obtain the solution for $p(x)$ which turns out to be the Gibbs distribution form

$$p(x;\Lambda) = \frac{1}{Z(\Lambda)} \exp\left[-\sum_{\alpha=1}^{K} \langle \lambda^{(\alpha)}, \phi^{(\alpha)}(x)\rangle\right]$$

where $\Lambda = \{\lambda^{(1)}, \lambda^{(2)}, ..., \lambda^{K}\}$ is the parameter, and $Z(\Lambda) = \int \exp\left[-\sum_{\alpha=1}^{K} \langle \lambda^{(\alpha)}, \phi^{(\alpha)}(x)\rangle\right]\ \mathrm{d}x$ is the normalization constant that guarantee the integration over $p(x)$ equals to 1.

Instead of using hand-crafted filters, we utilize the method in [10] which uses a Convolutional Neural Network (CNN) to learn the filters. Then the Gibbs distribution could be rewritten as

$$p(x;w) = \frac{1}{Z(w)} \exp\left[\sum_{\alpha=1}^{K}\sum_{i\in\mathcal{D}} w_i^{(\alpha)}([F^{(\alpha)} * x](i))\right]$$

where i is the position where the filter is apply to x, and $[F * x](i)$ indicates applying filter F at position i. We follow the learning method in [10] to learn the parameter of this model. Then we can draw random samples from the distribution $p(x;w)$ using Markov Chain Monte Carlo (MCMC) method. Particularly, to sample from $p(x;w)$, we adopt the Langevin dynamics. Writing the energy function as

$$U(x,w) = -\sum_{\alpha=1}^{K}\sum_{i\in\mathcal{D}} w_i^{(\alpha)}[F^{(\alpha)} * x](i) + \frac{1}{2\sigma^2}||x||^2$$

Then the Langevin dynamics iterates

$$x^{t+1} = x^t - \frac{\epsilon^2}{2}U^{'}(x^t, w) + \epsilon Z$$

where $U^{'}(x^t, w) = \partial U(x,w)/\partial x$, ϵ is the step size, and $Z \sim N(0,1)$ is a Gaussian noise.

Then the expectation in the auxiliary tasks could be calculated as

$$\mathbb{E}_{p(x)}[\mathbb{1}(h_{y_k}(x;\theta), \text{gender})] = \frac{1}{\tilde{M}}\sum_{i=1}^{\tilde{M}} \mathbb{1}(h_{y_k}(\tilde{x};\theta), \text{gender})$$

where $\tilde{x}$ is a sample drawn from the distribution $p(x;w)$.

4 Experiment

4.1 Constructing the Dataset

As mentioned previously, the dataset we used was the 2014 MS-COCO training set [9] which contains 80 distinct object categories. This dataset was chosen as it has been shown to be heavily gender-biased

Table 1: Gender-bias of chosen labels

Label	Bias (%)	Label	Bias (%)
truck	80.00	fork	49.08
motorcycle	85.88	knife	53.12
tie	88.74	spoon	52.77
backpack	76.30	cell phone	58.59
sports ball	72.82	teddy bear	42.43
handbag	48.64	Overall	68.70

[21] in similar experiments. Although gender is not an explicit object category, we were able to manually generate the gender label by parsing the provided captions. Furthermore, of the 80 available object categories, 11 were chosen for a total of 12 categories with the inclusion of the gender label. The 11 chosen categories as well as their gender biases can be found in table 1. As can be seen, approximately half of the chosen labels are heavily biased towards males while the rest are relatively unbiased. This was done deliberately to show that bias amplification can occur even for unbiased labels as the overall gender bias of the training set is still 68.70%.

For each of these 11 object categories, we filtered the MS-COCO dataset for images with both the particular object category label as well as the "person" label. For each filtered image, we then parsed the captions for any mentionings of the term "man" or "woman". Such terms were added to the training set with images mentioning both genders ignored. When repeating this for each of the 11 object categories, duplicate addings were avoided by keeping a track of images ids of all images already added. After this filtering procedure, we arrived at a total of 9885 images which we then split into a 85/15 training and testing set. These 9885 instances were then fed through a pretrained Resnet34 model to obtain the image embeddings which were then used to train both the descriptive model and the classifier.

4.2 Training

In training the multi-label classifier, we constructed two fully connected neural networks which differ only in the implementation of their loss function. The first was a standard neural network that utilized a 12 sigmoid activation output and cross entropy loss. This neural network was intended to be the baseline of the experiment and will be referenced as the baseline model going forward. Conversely, the second neural network, hereby referenced as the auxiliary model, utilized the custom auxiliary loss function detailed in section 3.2. In addition to using the cross entropy loss for the 12 categories, the auxiliary loss was computed by calculating the difference between the expected bias and training set bias. This expected bias was formulated by the auxiliary model by predicting on synthetic data generated by the descriptive model from section 3.3. Both neural networks contained the same architecture and were trained using the Adam optimizer with a learning rate of 3e-5, L2 regularization of 1e-3, batch size of 64, and for 50 epochs. After training under identical settings, both models arrived at training set and testing set accuracies of approximately 95% and 85%, respectively. These accuracies were calculated by computing the percentage of correctly predicted labels over the sum of all possible labels.

4.3 Bias Results

After training both models with identical parameters, figures 1 and 2 showcase the computed bias ratios for the baseline and auxiliary models respectively where the x-axis is the training dataset gender bias and the y-axis is the predicted gender bias. For the baseline model, it can be seen that a majority of the labels experience significant gender bias towards male classification with an average bias amplification of 16.24%. Even labels such as "handbag", "spoon", and "cell phone" which are relatively unbiased in the data set become overclassified with male bias due to the encoded bias of the other labels. Conversely, the average bias amplification of the auxiliary model is only 7.5% leading to a bias amplification reduction of 53.82% which is of similar magnitude to the bias amplification reduction of 47.5% shown by [21]. Here, figure 2 shows that for the majority of labels, the predicted gender ratio is more reflective of the underlying data set bias as desired.

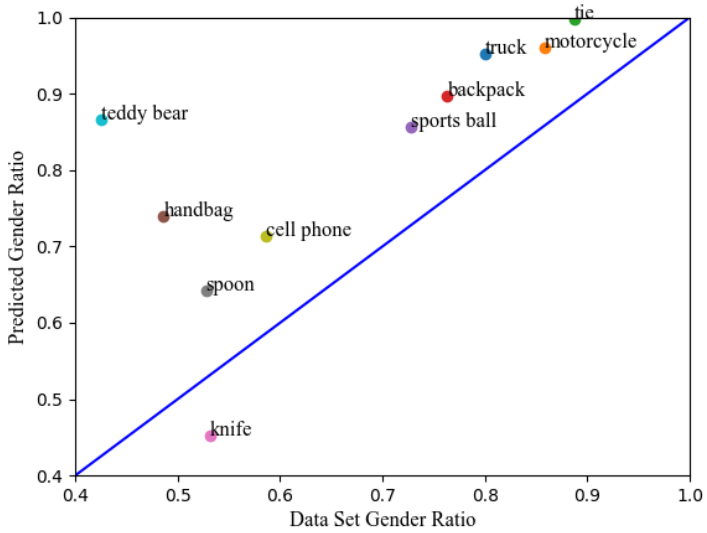

Figure 1: Baseline model gender bias

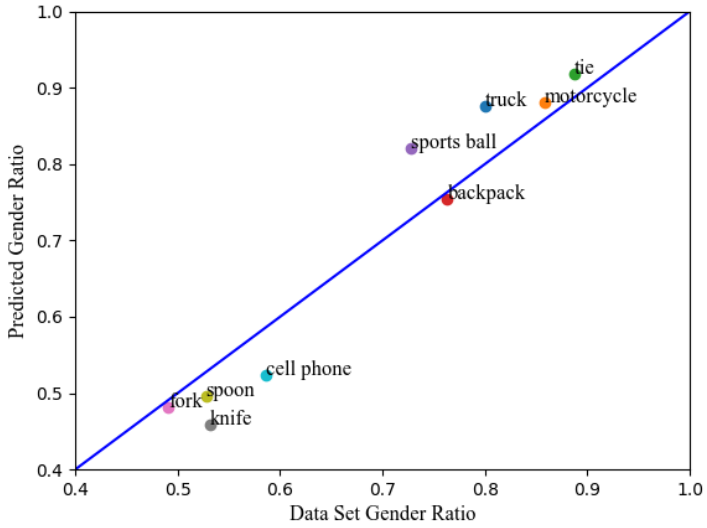

Figure 2: Auxiliary model gender bias

One thing to note is that the distribution of the labels in the training set were highly nonuniform which led to certain underrepresented labels being improperly gender classified. For example, labels such as "teddy bear" which only occurred in about 2.7% of all instances were classified with high gender bias in both models. In fact, in our auxiliary model, "teddy bear" was classified almost entirely towards females as can be seen from the label point missing in the frame. Issues such as this could be circumvented by either adding more data involving such underrepresented labels, and/or increasing the expressiveness of our model.

5 Future Research Directions

Although our bias amplification reduction is of similar magnitude to existing methods [21], it should be noted that the scale of our classification problem was much smaller (12 object vs. 66 object categories) than existing techniques. Due to this, performance comparisons may be misleading as the auxiliary model has not been tested on larger, more difficult classification tasks. Such tests could be performed in the future for a more comprehensive analysis on the effectiveness of our model.

Furthermore, ideally, synthetic data should be sampled online from the descriptive model at every training iteration when computing the bias difference for training. Instead, our experiment sampled 1760 synthetic data instances all at once offline and used this to generate the expected bias repeatedly throughout the training process. Although generating new synthetic data at every iteration would likely result in a performance increase, we decided to forgo this in favor of runtime efficiency due to time constraints as online sampling is incredibly time-consuming. Exploring this online implementation is another future possibility.

6 Conclusion

In this research project, we introduced a novel approach for learning a multi-label classifier with reduced bias amplification. Whereas previous techniques [21] have kept the classifier unaltered and explicitly reduced the bias amplification through post-processing of the predicted output using the testing set, our method implicitly reduces the bias amplification by incorporating the learning of the bias into the classifier itself. By doing so, we avoid any bias that may be generated from such post-processing. Although the auxiliary model was able to outperform the baseline model in terms of gender bias, we note that the scale of the classification problem was much smaller than existing methods and thus, our results are not completely comprehensive on the capabilities of our model. Nonetheless, results from our experiments show promise and invite further experimentation, possibly using new auxiliary tasks or even a combination of several auxiliary tasks.

References

[1] Google apologises for photos app's racist blunder, Jul 2015.

[2] Jonathan Baxter. A bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine learning*, 28(1):7–39, 1997.

[3] Peter Clark, Oyvind Tafjord, and Kyle Richardson. Transformers as soft reasoners over language. *arXiv preprint arXiv:2002.05867*, 2020.

[4] Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki, Tom Schaul, Joel Z Leibo, David Silver, and Koray Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks, 2016.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[6] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

[7] Qing Li, Qingyi Tao, Shafiq Joty, Jianfei Cai, and Jiebo Luo. Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions, 2018.

[8] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2014.

[10] Yang Lu, Song-Chun Zhu, and Ying Nian Wu. Learning frame models using cnn filters. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[11] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning, 2019.

[12] United States. Executive Office of the President and John Podesta. *Big data: Seizing opportunities, preserving values*. White House, Executive Office of the President, 2014.

[13] Ji Ho Park, Jamin Shin, and Pascale Fung. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*, 2018.

[14] Dan Roth and Wen-tau Yih. A linear programming formulation for global inference in natural language tasks. Technical report, ILLINOIS UNIV AT URBANA-CHAMPAIGN DEPT OF COMPUTER SCIENCE, 2004.

[15] Song-Chun Zhu. Statistical modeling and conceptualization of visual patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(6):691–712, 2003.

[16] Shao-Hua Sun, Hyeonwoo Noh, Sriram Somasundaram, and Joseph Lim. Neural program synthesis from diverse demonstration videos. In *International Conference on Machine Learning*, pages 4790–4799, 2018.

[17] Eva Vanmassenhove, Christian Hardmeier, and Andy Way. Getting gender right in neural machine translation. *arXiv preprint arXiv:1909.05088*, 2019.

[18] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

[19] Xu Yang, Hanwang Zhang, and Jianfei Cai. Deconfounded image captioning: A causal retrospect. *arXiv preprint arXiv:2003.03923*, 2020.

[20] 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.

[21] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*, 2017.

[22] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496*, 2018.

[23] Song Chun Zhu, Ying Nian Wu, and David Mumford. Minimax entropy principle and its application to texture modeling. *Neural computation*, 9(8):1627–1660, 1997.