Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints

Jieyu Zhao[§] **Tianlu Wang**[§] **Mark Yatskar**[‡]
Vicente Ordonez[§] **Kai-Wei Chang**[§]
[§]University of Virginia
{jz4fu, tw8cb, vicente, kc2wc}@virginia.edu
[‡]University of Washington
my89@cs.washington.edu

Abstract

Language is increasingly being used to define rich visual recognition problems with supporting image collections sourced from the web. Structured prediction models are used in these tasks to take advantage of correlations between co-occurring labels and visual input but risk inadvertently encoding social biases found in web corpora. In this work, we study data and models associated with multilabel object classification and visual semantic role labeling. We find that (a) datasets for these tasks contain significant gender bias and (b) models trained on these datasets further amplify existing bias. For example, the activity `cooking` is over 33% more likely to involve females than males in a training set, and a trained model further amplifies the disparity to 68% at test time. We propose to inject corpus-level constraints for calibrating existing structured prediction models and design an algorithm based on Lagrangian relaxation for collective inference. Our method results in almost no performance loss for the underlying recognition task but decreases the magnitude of bias amplification by 47.5% and 40.5% for multilabel classification and visual semantic role labeling, respectively.

1 Introduction

Visual recognition tasks involving language, such as captioning (Vinyals et al., 2015), visual question answering (Antol et al., 2015), and visual semantic role labeling (Yatskar et al., 2016), have emerged as avenues for expanding the diversity of information that can be recovered from images. These tasks aim at extracting rich semantics from images and require large quantities of labeled data, predominantly retrieved from the web. Methods often combine structured prediction and deep learning to model correlations between labels and images to make judgments that otherwise would have weak visual support. For example, in the first image of Figure 1, it is possible to predict a `spatula` by considering that it is a common tool used for the activity `cooking`. Yet such methods run the risk of discovering and exploiting societal biases present in the underlying web corpora. Without properly quantifying and reducing the reliance on such correlations, broad adoption of these models can have the inadvertent effect of magnifying stereotypes.

In this paper, we develop a general framework for quantifying bias and study two concrete tasks, visual semantic role labeling (vSRL) and multilabel object classification (MLC). In vSRL, we use the imSitu formalism (Yatskar et al., 2016, 2017), where the goal is to predict activities, objects and the roles those objects play within an activity. For MLC, we use MS-COCO (Lin et al., 2014; Chen et al., 2015), a recognition task covering 80 object classes. We use gender bias as a running example and show that both supporting datasets for these tasks are biased with respect to a gender binary[1].

Our analysis reveals that over 45% and 37% of verbs and objects, respectively, exhibit bias toward a gender greater than 2:1. For example, as seen in Figure 1, the `cooking` activity in imSitu is a heavily biased verb. Furthermore, we show that after training state-of-the-art structured predictors, models amplify the existing bias, by 5.0% for vSRL, and 3.6% in MLC.

[1]To simplify our analysis, we only consider a gender binary as perceived by annotators in the datasets. We recognize that a more fine-grained analysis would be needed for deployment in a production system. Also, note that the proposed approach can be applied to other NLP tasks and other variables such as identification with a racial or ethnic group.

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2979–2989
Copenhagen, Denmark, September 7–11, 2017.

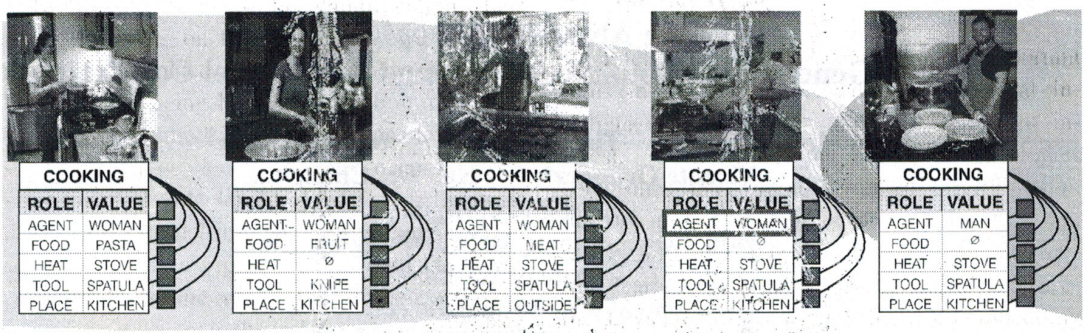


Figure 1: Five example images from the imSitu visual semantic role labeling (vSRL) dataset. Each image is paired with a table describing a situation: the verb, `cooking`, its semantic roles, i.e `agent`, and noun values filling that role, i.e. `woman`. In the imSitu training set, 33% of `cooking` images have `man` in the `agent` role while the rest have `woman`. After training a Conditional Random Field (CRF), bias is amplified: `man` fills 16% of `agent` roles in `cooking` images. To reduce this bias amplification our calibration method adjusts weights of CRF potentials associated with biased predictions. After applying our methods, `man` appears in the `agent` role of 20% of `cooking` images, reducing the bias amplification by 25%, while keeping the CRF vSRL performance unchanged.

To mitigate the role of bias amplification when training models on biased corpora, we propose a novel constrained inference framework, called RBA, for **R**educing **B**ias **A**mplification in predictions. Our method introduces corpus-level constraints so that gender indicators co-occur no more often together with elements of the prediction task than in the original training distribution. For example, as seen in Figure 1, we would like noun `man` to occur in the `agent` role of the `cooking` as often as it occurs in the imSitu training set when evaluating on a development set. We combine our calibration constraint with the original structured predictor and use Lagrangian relaxation (Korte and Vygen, 2008; Rush and Collins, 2012) to reweigh bias creating factors in the original model.

We evaluate our calibration method on imSitu vSRL and COCO MLC and find that in both instances, our models substantially reduce bias amplification. For vSRL, we reduce the average magnitude of bias amplification by 40.5%. For MLC, we are able to reduce the average magnitude of bias amplification by 47.5%. Overall, our calibration methods do not affect the performance of the underlying visual system, while substantially reducing the reliance of the system on socially biased correlations[2].

[2]Code and data are available at `https://github.com/uclanlp/reducingbias`

2 Related Work

As intelligence systems start playing important roles in our daily life, ethics in artificial intelligence research has attracted significant interest. It is known that big-data technologies sometimes inadvertently worsen discrimination due to implicit biases in data (Podesta et al., 2014). Such issues have been demonstrated in various learning systems, including online advertisement systems (Sweeney, 2013), word embedding models (Bolukbasi et al., 2016; Caliskan et al., 2017), online news (Ross and Carter, 2011), web search (Kay et al., 2015), and credit score (Hardt et al., 2016). Data collection biases have been discussed in the context of creating image corpus (Misra et al., 2016; van Miltenburg, 2016) and text corpus (Gordon and Van Durme, 2013; Van Durme, 2010). In contrast, we show that given a gender biased corpus, structured models such as conditional random fields, amplify the bias.

The effect of the data imbalance can be easily detected and fixed when the prediction task is simple. For example, when classifying binary data with unbalanced labels (i.e., samples in the majority class dominate the dataset), a classifier trained exclusively to optimize accuracy learns to always predict the majority label, as the cost of making mistakes on samples in the minority class can be neglected. Various approaches have been proposed to make a "fair" binary classification (Barocas and Selbst, 2014; Dwork et al., 2012; Feldman

et al., 2015; Zliobaite, 2015). For structured prediction tasks the effect is harder to quantify and we are the first to propose methods to reduce bias amplification in this context.

Lagrangian relaxation and dual decomposition techniques have been widely used in NLP tasks (e.g., (Sontag et al., 2011; Rush and Collins, 2012; Chang and Collins, 2011; Peng et al., 2015)) for dealing with instance-level constraints. Similar techniques (Chang et al., 2013; Dalvi, 2015) have been applied in handling corpus-level constraints for semi-supervised multilabel classification. In contrast to previous works aiming for improving accuracy performance, we incorporate corpus-level constraints for reducing gender bias.

3 Visualizing and Quantifying Biases

Modern statistical learning approaches capture correlations among output variables in order to make coherent predictions. However, for real-world applications, some implicit correlations are not appropriate, especially if they are amplified. In this section, we present a general framework to analyze inherent biases learned and amplified by a prediction model.

Identifying bias We consider that prediction problems involve several inter-dependent output variables $y_1, y_2, ...y_K$, which can be represented as a structure $y = \{y_1, y_2, ...y_K\} \in Y$. This is a common setting in NLP applications, including tagging, and parsing. For example, in the vSRL task, the output can be represented as a structured table as shown in Fig 1. Modern techniques often model the correlation between the sub-components in y and make a joint prediction over them using a structured prediction model. More details will be provided in Section 4.

We assume there is a subset of output variables $g \subseteq y, g \in G$ that reflects demographic attributes such as gender or race (e.g. $g \in G = \{\texttt{man}, \texttt{woman}\}$ is the agent), and there is another subset of the output $o \subseteq y, o \in O$ that are co-related with g (e.g., o is the activity present in an image, such as `cooking`). The goal is to identify the correlations that are potentially amplified by a learned model.

To achieve this, we define the bias score of a given output, o, with respect to a demographic variable, g, as:

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

where $c(o, g)$ is the number of occurrences of o and g in a corpus. For example, to analyze how genders of agents and activities are co-related in vSRL, we define the gender bias toward `man` for each verb $b(verb, \texttt{man})$ as:

$$\frac{c(verb, \texttt{man})}{c(verb, \texttt{man}) + c(verb, \texttt{woman})}. \quad (1)$$

If $b(o, g) > 1/\|G\|$, then o is positively correlated with g and may exhibit bias.

Evaluating bias amplification To evaluate the degree of bias amplification, we propose to compare bias scores on the training set, $b^*(o, g)$, with bias scores on an unlabeled evaluation set of images $\tilde{b}(o, g)$ that has been annotated by a predictor. We assume that the evaluation set is identically distributed to the training set. Therefore, if o is positively correlated with g (i.e, $b^*(o, g) > 1/\|G\|$) and $\tilde{b}(o, g)$ is larger than $b^*(o, g)$, we say bias has been amplified. For example, if $b^*(\texttt{cooking}, \texttt{woman}) = .66$, and $\tilde{b}(\texttt{cooking}, \texttt{woman}) = .84$, then the bias of `woman` toward `cooking` has been amplified. Finally, we define the mean bias amplification as:

$$\frac{1}{|O|} \sum_{g} \sum_{o \in \{o \in O | b^*(o,g) > 1/\|G\|\}} \tilde{b}(o, g) - b^*(o, g).$$

This score estimates the average magnitude of bias amplification for pairs of o and g which exhibited bias.

4 Calibration Algorithm

In this section, we introduce **R**educing **B**ias **A**mplification, RBA, a debiasing technique for calibrating the predictions from a structured prediction model. The intuition behind the algorithm is to inject constraints to ensure the model predictions follow the distribution observed from the training data. For example, the constraints added to the vSRL system ensure the gender ratio of each verb in Eq. (1) are within a given margin based on the statistics of the training data. These constraints are applied at the corpus level, because computing gender ratio requires the predictions of all test

instances. As a result, a joint inference over test instances is required[3]. Solving such a giant inference problem with constraints is hard. Therefore, we present an approximate inference algorithm based on Lagrangian relaxation. The advantages of this approach are:

- Our algorithm is iterative, and at each iteration, the joint inference problem is decomposed to a per-instance basis. This can be solved by the original inference algorithm. That is, our approach works as a meta-algorithm and developers do not need to implement a new inference algorithm.
- The approach is general and can be applied in any structured model.
- Lagrangian relaxation guarantees the solution is optimal if the algorithm converges and all constraints are satisfied.

In practice, it is hard to obtain a solution where all corpus-level constrains are satisfied. However, we show that the performance of the proposed approach is empirically strong. We use imSitu for vSRL as a running example to explain our algorithm.

Structured Output Prediction As we mentioned in Sec. 3, we assume the structured output $y \in Y$ consists of several sub-components. Given a test instance i as an input, the inference problem is to find

$$\arg\max_{y \in Y} \quad f_\theta(y, i),$$

where $f_\theta(y, i)$ is a scoring function based on a model θ learned from the training data. The structured output y and the scoring function $f_\theta(y, i)$ can be decomposed into small components based on an independence assumption. For example, in the vSRL task, the output y consists of two types of binary output variables $\{y_v\}$ and $\{y_{v,r}\}$. The variable $y_v = 1$ if and only if the activity v is chosen. Similarly, $y_{v,r} = 1$ if and only if both the activity v and the semantic role r are assigned[4]. The scoring function $f_\theta(y, i)$ is decomposed accordingly such that:

$$f_\theta(y, i) = \sum_v y_v s_\theta(v, i) + \sum_{v,r} y_{v,r} s_\theta(v, r, i),$$

represents the overall score of an assignment, and $s_\theta(v, i)$ and $s_\theta(v, r, i)$ are the potentials of the sub-assignments. The output space Y contains all feasible assignments of y_v and $y_{v,r}$, which can be represented as instance-wise constraints. For example, the constraint, $\sum_v y_v = 1$ ensures only one activity is assigned to one image.

Corpus-level Constraints Our goal is to inject constraints to ensure the output labels follow a desired distribution. For example, we can set a constraint to ensure the gender ratio for each activity in Eq. (1) is within a given margin. Let $y^i = \{y^i_v\} \cup \{y^i_{v,r}\}$ be the output assignment for test instance i[5]. For each activity v^*, the constraints can be written as

$$b^* - \gamma \leq \frac{\sum_i y^i_{v=v^*, r \in M}}{\sum_i y^i_{v=v^*, r \in W} + \sum_i y^i_{v=v^*, r \in M}} \leq b^* + \gamma \tag{2}$$

where $b^* \equiv b^*(v^*, man)$ is the desired gender ratio of an activity v^*, γ is a user-specified margin. M and W are a set of semantic role-values representing the agent as a man or a woman, respectively.

Note that the constraints in (2) involve all the test instances. Therefore, it requires a joint inference over the entire test corpus. In general, these corpus-level constraints can be represented in a form of $A \sum_i y^i - b \leq 0$, where each row in the matrix $A \in R^{l \times K}$ is the coefficients of one constraint, and $b \in R^l$. The constrained inference problem can then be formulated as:

$$\begin{aligned} \max_{\{y^i\} \in \{Y^i\}} \quad & \sum_i f_\theta(y^i, i), \\ \text{s.t.} \quad & A \sum_i y^i - b \leq 0, \end{aligned} \tag{3}$$

where $\{Y^i\}$ represents a space spanned by possible combinations of labels for all instances. Without the corpus-level constraints, Eq. (3) can be optimized by maximizing each instance i

$$\max_{y_i \in Y^i} f_\theta(y^i, i),$$

separately.

Lagrangian Relaxation Eq. (3) can be solved by several combinatorial optimization methods. For example, one can represent the problem as an

[3] A sufficiently large sample of test instances must be used so that bias statistics can be estimated. In this work we use the entire test set for each respective problem.

[4] We use r to refer to a combination of role and noun. For example, one possible value indicates an agent is a woman.

[5] For the sake of simplicity, we abuse the notations and use i to represent both input and data index.

Dataset	Task	Images	O-Type	$\|O\|$
imSitu	vSRL	60,000	verb	212
MS-COCO	MLC	25,000	object	66

Table 1: Statistics for the two recognition problems. In vSRL, we consider gender bias relating to verbs, while in MLC we consider the gender bias related to objects.

integer linear program and solve it using an off-the-shelf solver (e.g., Gurobi (Gurobi Optimization, 2016)). However, Eq. (3) involves all test instances. Solving a constrained optimization problem on such a scale is difficult. Therefore, we consider relaxing the constraints and solve Eq. (3) using a Lagrangian relaxation technique (Rush and Collins, 2012). We introduce a Lagrangian multiplier $\lambda_j \geq 0$ for each corpus-level constraint. The Lagrangian is

$$L(\lambda, \{y^i\}) = \sum_i f_\theta(y^i) - \sum_{j=1}^{l} \lambda_j \left(A_j \sum_i y^i - b_j \right), \quad (4)$$

where all the $\lambda_j \geq 0, \forall j \in \{1, \ldots, l\}$. The solution of Eq. (3) can be obtained by the following iterative procedure:

AN OVERVIEW OF THE LAGRANGIAN SECTION OF THE METHOD

1) At iteration t, get the output solution of each instance i

$$y^{i,(t)} = \operatorname*{argmax}_{y \in \mathcal{Y}'} L(\lambda^{(t-1)}, y) \quad (5)$$

2) update the Lagrangian multipliers.

$$\lambda^{(t)} = \max\left(0, \lambda^{(t-1)} + \sum_i \eta(A y^{i,(t)} - b)\right),$$

where $\lambda^{(0)} = \mathbf{0}$. η is the learning rate for updating λ. Note that with a fixed $\lambda^{(t-1)}$, Eq. (5) can be solved using the original inference algorithms. The algorithm loops until all constraints are satisfied (i.e. optimal solution achieved) or reach maximal number of iterations.

5 Experimental Setup

In this section, we provide details about the two visual recognition tasks we evaluated for bias: visual semantic role labeling (vSRL), and multi-label classification (MLC). We focus on gender, defining $G = \{\texttt{man}, \texttt{woman}\}$ and focus on the `agent` role in vSRL, and any occurrence in text associated with the images in MLC. Problem statistics are summarized in Table 1. We also provide setup details for our calibration method.

5.1 Visual Semantic Role Labeling

Dataset We evaluate on imSitu (Yatskar et al., 2016) where activity classes are drawn from verbs and roles in FrameNet (Baker et al., 1998) and noun categories are drawn from WordNet (Miller et al., 1990). The original dataset includes about 125,000 images with 75,702 for training, 25,200 for developing, and 25,200 for test. However, the dataset covers many non-human oriented activities (e.g., `rearing`, `retrieving`, and `wagging`), so we filter out these verbs, resulting in 212 verbs, leaving roughly 60,000 of the original 125,000 images in the dataset.

Model We build on the baseline CRF released with the data, which has been shown effective compared to a non-structured prediction baseline (Yatskar et al., 2016). The model decomposes the probability of a realized situation, y, the combination of activity, v, and realized frame, a set of semantic (role,noun) pairs (e, n_e), given an image i as :

$$p(y|i;\theta) \propto \psi(v, i; \theta) \prod_{(e,n_e) \in R_f} \psi(v, e, n_e, i; \theta)$$

where each potential value in the CRF for subpart x, is computed using features f_i from the VGG convolutional neural network (Simonyan and Zisserman, 2014) on an input image, as follows:

$$\psi(x, i; \theta) = e^{w_x^T f_i + b_x},$$

where w and b are the parameters of an affine transformation layer. The model explicitly captures the correlation between activities and nouns in semantic roles, allowing it to learn common priors. We use a model pretrained on the original task with 504 verbs.

5.2 Multilabel Classification

Dataset We use MS-COCO (Lin et al., 2014), a common object detection benchmark, for multilabel object classification. The dataset contains 80 object types but does not make gender distinctions between man and woman. We use the five associated image captions available for each image in this dataset to annotate the gender of people in the

images. If any of the captions mention the word man or woman we mark it, removing any images that mention both genders. Finally, we filter any object category not strongly associated with humans by removing objects that do not occur with man or woman at least 100 times in the training set, leaving a total of 66 objects.

Model For this multi-label setting, we adapt a similar model as the structured CRF we use for vSRL. We decompose the joint probability of the output y, consisting of all object categories, c, and gender of the person, g, given an image i as:

$$p(y|i;\theta) \propto \psi(g,i;\theta)\prod_{c\in y}\psi(g,c,i;\theta)$$

where each potential value for x, is computed using features, f_i, from a pretrained ResNet-50 convolutional neural network evaluated on the image,

$$\psi(x,i;\theta) = e^{w_x^T f_i + b_x}.$$

We trained a model using SGD with learning rate 10^{-5}, momentum 0.9 and weight-decay 10^{-4}, fine tuning the initial visual network, for 50 epochs.

5.3 Calibration

The inference problems for both models are:

$$\arg\max_{y\in Y} f_\theta(y,i) = \log p(y|i;\theta).$$

We use the algorithm in Sec. (4) to calibrate the predictions using model θ. Our calibration tries to enforce gender statistics derived from the training set of corpus applicable for each recognition problem. For all experiments, we try to match gender ratios on the test set within a margin of .05 of their value on the training set. While we do adjust the output on the test set, we never use the ground truth on the test set and instead working from the assumption that it should be similarly distributed as the training set. When running the debiasing algorithm, we set $\eta = 10^{-1}$ and optimize for 100 iterations.

6 Bias Analysis

In this section, we use the approaches outlined in Section 3 to quantify the bias and bias amplification in the vSRL and the MLC tasks.

6.1 Visual Semantic Role Labeling

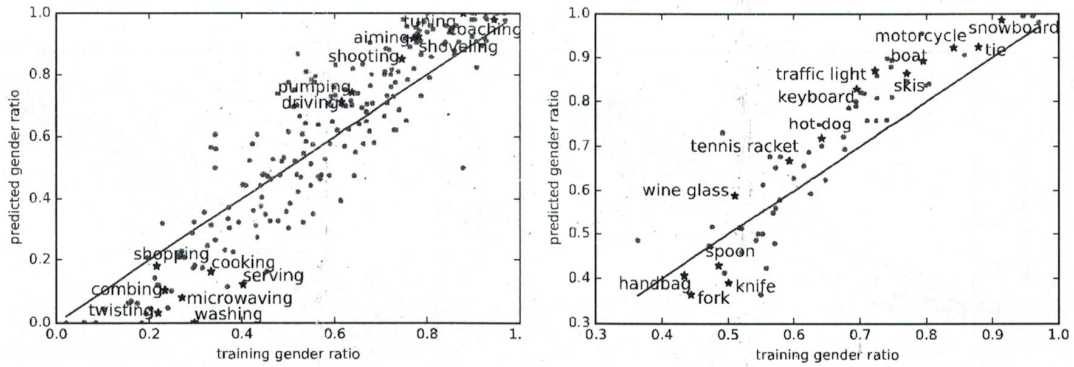
imSitu is gender biased In Figure 2(a), along the x-axis, we show the male favoring bias of imSitu verbs. Overall, the dataset is heavily biased toward male agents, with 64.6% of verbs favoring a male agent by an average bias of 0.707 (roughly 3:1 male). Nearly half of verbs are extremely biased in the male or female direction: 46.95% of verbs favor a gender with a bias of at least 0.7.[6] Figure 2(a) contains several activity labels revealing problematic biases. For example, `shopping`, `microwaving` and `washing` are biased toward a female `agent`. Furthermore, several verbs such as `driving`, `shooting`, and `coaching` are heavily biased toward a male `agent`.

Training on imSitu amplifies bias In Figure 2(a), along the y-axis, we show the ratio of male agents (% of total people) in predictions on an unseen development set. The mean bias amplification in the development set is high, 0.050 on average, with 45.75% of verbs exhibiting amplification. Biased verbs tend to have stronger amplification: verbs with training bias over 0.7 in either the male or female direction have a mean amplification of 0.072. Several already problematic biases have gotten much worse. For example, `serving`, only had a small bias toward females in the training set, 0.402, is now heavily biased toward females, 0.122. The verb `tuning`, originally heavily biased toward males, 0.878, now has exclusively male agents.

6.2 Multilabel Classification

MS-COCO is gender biased In Figure 2(b) along the x-axis, similarly to imSitu, we analyze bias of objects in MS-COCO with respect to males. MS-COCO is even more heavily biased toward men than imSitu, with 86.6% of objects biased toward men, but with smaller average magnitude, 0.65. One third of the nouns are extremely biased toward males, 37.9% of nouns favor men with a bias of at least 0.7. Some problematic examples include kitchen objects such as `knife`, `fork`, or `spoon` being more biased toward woman. Outdoor recreation related objects such `tennis racket`, `snowboard` and `boat` tend to be more biased toward men.

[6] In this gender binary, bias toward `woman` is 1− the bias toward `man`



(a) Bias analysis on imSitu vSRL

(b) Bias analysis on MS-COCO MLC

Figure 2: Gender bias analysis of imSitu vSRL and MS-COCO MLC. (a) gender bias of verbs toward man in the training set versus bias on a predicted development set. (b) gender bias of nouns toward man in the training set versus bias on the predicted development set. Values near zero indicate bias toward woman while values near 0.5 indicate unbiased variables. Across both dataset, there is significant bias toward males, and significant bias amplification after training on biased training data.

Training on MS-COCO amplifies bias In Figure 2(b), along the y-axis, we show the ratio of man (% of both gender) in predictions on an unseen development set. The mean bias amplification across all objects is 0.036, with 65.67% of nouns exhibiting amplification. Larger training bias again tended to indicate higher bias amplification: biased objects with training bias over 0.7 had mean amplification of 0.081. Again, several problematic biases have now been amplified. For example, kitchen categories already biased toward females such as knife, fork and spoon have all been amplified. Technology oriented categories initially biased toward men such as keyboard and mouse have each increased their bias toward males by over 0.100.

6.3 Discussion

We confirmed our hypothesis that (a) both the imSitu and MS-COCO datasets, gathered from the web, are heavily gender biased and that (b) models trained to perform prediction on these datasets amplify the existing gender bias when evaluated on development data. Furthermore, across both datasets, we showed that the degree of bias amplification was related to the size of the initial bias, with highly biased object and verb categories exhibiting more bias amplification. Our results demonstrate that care needs be taken in deploying such uncalibrated systems otherwise they could not only reinforce existing social bias but actually make them worse.

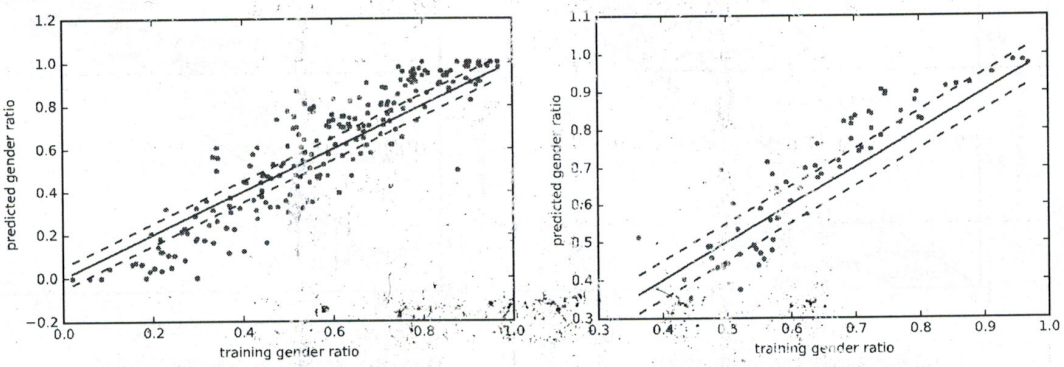
7 Calibration Results

We test our methods for reducing bias amplification in two problem settings: visual semantic role labeling in the imSitu dataset (vSRL) and multilabel image classification in MS-COCO (MLC). In all settings we derive corpus constraints using the training set and then run our calibration method in batch on either the development or testing set. Our results are summarized in Table 2 and Figure 3.

7.1 Visual Semantic Role Labeling

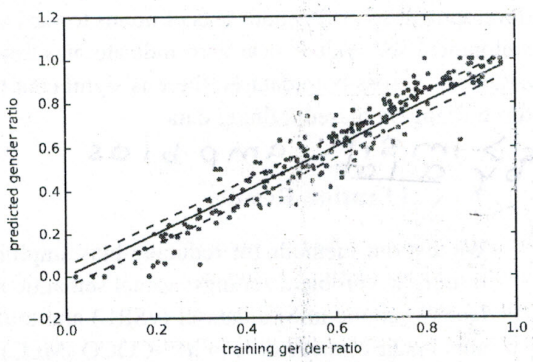
Our quantitative results are summarized in the first two sections of Table 2. On the development set, the number of verbs whose bias exceed the original bias by over 5% decreases 30.5% (Viol.). Overall, we are able to significantly reduce bias amplification in vSRL by 52% on the development set (Amp. bias). We evaluate the underlying recognition performance using the standard measure in vSRL: top-1 semantic role accuracy, which tests how often the correct verb was predicted and the noun value was correctly assigned to a semantic role. Our calibration method results in a negligible decrease in performance (Perf.). In Figure 3(c) we can see that the overall distance to the training set distribution after applying RBA decreased significantly, over 39%.

Figure 3(e) demonstrates that across all initial training bias, RBA is able to reduce bias amplification. In general, RBA struggles to remove bias amplification in areas of low initial training bias,

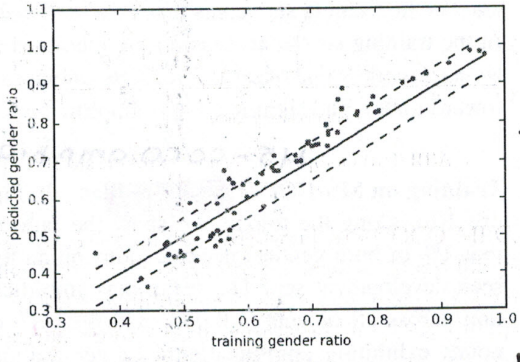


(a) Bias analysis on imSitu vSRL without RBA

(b) Bias analysis on MS-COCO MLC without RBA

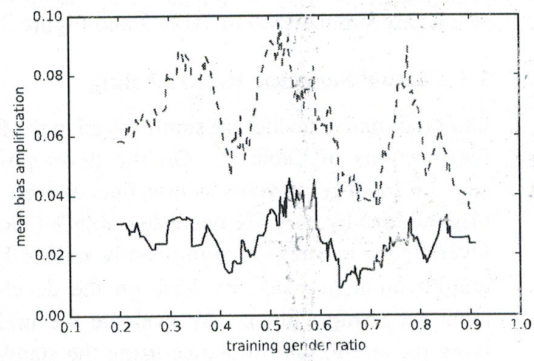


(c) Bias analysis on imSitu vSRL with RBA

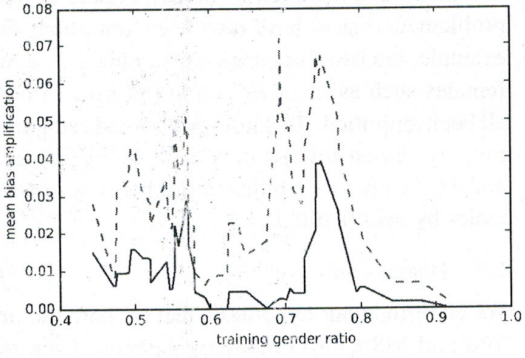


(d) Bias analysis on MS-COCO MLC with RBA

(d)
The ~~bottom~~ points are all closer to the mean, ie exhibits less bias



(e) Bias in vSRL with (blue) / without (red) RBA



(f) Bias in MLC with (blue) / without (red) RBA

Figure 3: Results of reducing bias amplification using RBA on imSitu vSRL and MS-COCO MLC. Figures 3(a)-(d) show initial training set bias along the x-axis and development set bias along the y-axis. Dotted blue lines indicate the 0.05 margin used in RBA, with points violating the margin shown in red while points meeting the margin are shown in green. Across both settings adding RBA significantly reduces the number of violations, and reduces the bias amplification significantly. Figures 3(e)-(f) demonstrate bias amplification as a function of training bias, with and without RBA. Across all initial training biases, RBA is able to reduce the bias amplification.

Method	Viol.	Amp. bias	Perf. (%)
vSRL: Development Set			
CRF	154	0.050	24.07
CRF + RBA	107	0.024	23.97
vSRL: Test Set			
CRF	149	0.042	24.14
CRF + RBA	102	0.025	24.01
MLC: Development Set			
CRF	40	0.032	45.27
CRF + RBA	24	0.022	45.19
MLC: Test Set			
CRF	38	0.040	45.40
CRF + RBA	16	0.021	45.38

Table 2: Number of violated constraints, mean amplified bias, and test performance before and after calibration using RBA. The test performances of vSRL and MLC are measured by top-1 semantic role accuracy and top-1 mean average precision, respectively.

likely because bias is encoded in image statistics and cannot be removed as effectively with an image agnostic adjustment. Results on the test set support our development set results: we decrease bias amplification by 40.5% (Amp. bias).

7.2 Multilabel Classification

Our quantitative results on MS-COCO RBA are summarized in the last two sections of Table 2. Similarly to vSRL, we are able to reduce the number of objects whose bias exceeds the original training bias by 5%, by 40% (Viol.). Bias amplification was reduced by 31.3% on the development set (Amp. bias). The underlying recognition system was evaluated by the standard measure: top-1 mean average precision, the precision averaged across object categories. Our calibration method results in a negligible loss in performance. In Figure 3(d), we demonstrate that we substantially reduce the distance between training bias and bias in the development set. Finally, in Figure 3(f) we demonstrate that we decrease bias amplification for all initial training bias settings. Results on the test set support our development results: we decrease bias amplification by 47.5% (Amp. bias).

7.3 Discussion

We have demonstrated that RBA can significantly reduce bias amplification. While were not able to remove all amplification, we have made significant progress with little or no loss in underlying recognition performance. Across both problems, RBA was able to reduce bias amplification at all initial values of training bias.

8 Conclusion

Structured prediction models can leverage correlations that allow them to make correct predictions even with very little underlying evidence. Yet such models risk potentially leveraging social bias in their training data. In this paper, we presented a general framework for visualizing and quantifying biases in such models and proposed RBA to calibrate their predictions under two different settings. Taking gender bias as an example, our analysis demonstrates that conditional random fields can amplify social bias from data while our approach RBA can help to reduce the bias.

Our work is the first to demonstrate structured prediction models amplify bias and the first to propose methods for reducing this effect but significant avenues for future work remain. While RBA can be applied to any structured predictor, it is unclear whether different predictors amplify bias more or less. Furthermore, we presented only one method for measuring bias. More extensive analysis could explore the interaction among predictor, bias measurement, and bias de-amplification method. Future work also includes applying bias reducing methods in other structured domains, such as pronoun reference resolution (Mitkov, 2014).

Acknowledgement This work was supported in part by National Science Foundation Grant IIS-1657193 and two NVIDIA Hardware Grants.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The Berkeley framenet project. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 86–90.

Solon Barocas and Andrew D Selbst. 2014. Big data's disparate impact. *Available at SSRN 2477899*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016.