Men also like shopping

Reducing Gender Bias Amplification Using Corpus-level Constraints

Author

Jieyu Zhao Tianlu Wang Mark Yatskar
Vicente Ordonez Kai-Wei Chang

Structured prediction

Model correlations between labels to make judgements which have weak support

Pros

- Take advantage of correlations between co-occuring labels
- Higher accuracy

Cons

- Find the correlations we don't want
- magnify stereotypes

Visual recognition tasks

vSRL (visual Semantic Role Labeling)

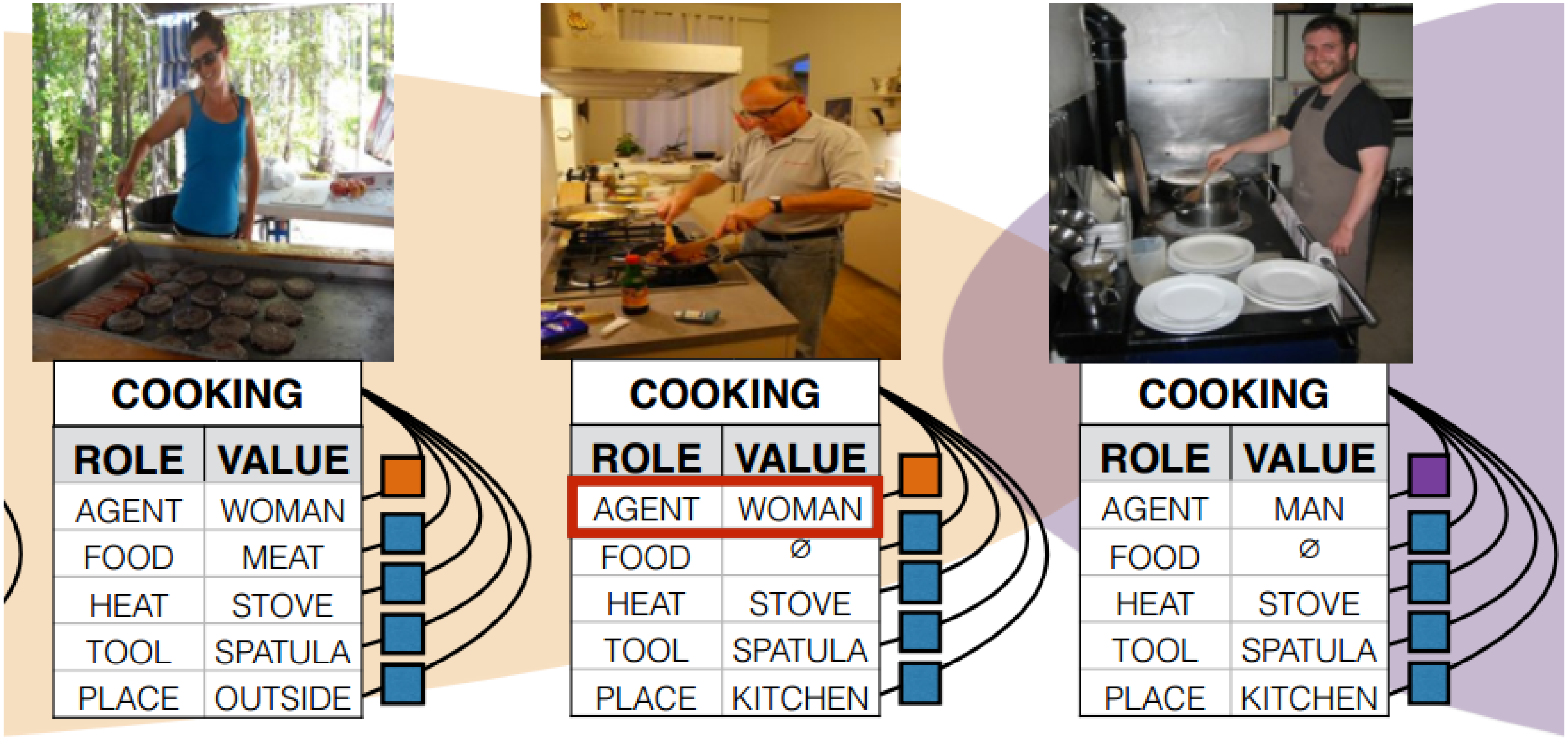
- Dataset: imSitu
  predict activities, objects and the roles those objects play within an activity

MLC (MultiLabel object Classification)

- Dataset: MS-COCO
  a recognition task covering 80 object classes

Visual recognition tasks

CRF predictor wil amplify bias



Algorithms

Identify bias

- several inter-dependent variables $y = \{y_1, y_2...y_k\} \in Y$.
- subset of output variables $g \subset y$,$g \in G$ that reflects demographic attributes such as gender and race (e.g. $g \in G = \{man, woman\}$)
- another subset $o \subset y$,$o \in O$ that corelated with g such as activities (e.g. cooking)

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

if $b(o,g) > \frac{1}{||G||}$, it may exhibits bias.

Algorithms

Evaluating bias amplification

Compare bias scores on the training set $b^*(o, g)$ and unlabelled evaluation set $\hat{b}(o, g)$ which we assume that is identically distributed to the former

We define the mean bias amplification as:

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

Algorithms

Reducing Bias Amplification (RBA)

Inject constraints on corpus level to ensure the model predictions follow the distribution observed from the training data
e.g. Constraints on gender ratio of each verb in *vSRL* at corpus level, ensuring it lies into a certain margin based on the satistics of the training data

Problem

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

where y is consist of y_v and $y_{v,r}$

Algorithms

Corpus level constraints

in *vSRL*, for each activity v^*, the constraints can be written as:

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

In general, these constraints can then be represented as $A \sum_i y^i \leq b$, so the constraint inference problem is formulated as

$$\max_{y_i \in Y^i} f_\theta(y, i), \qquad s.t. A \sum_i y^i - b \leq 0$$

Algorithms

Lagrangian relaxation

Problem:

$$\max f(x), \quad s.t. Ax \leq b \tag{1}$$

introduce it into

$$\max f(x) + \lambda^T (b - Ax) \tag{2}$$

where λ is nonnegtive. let $\hat{x}$ and $\bar{x}$ be solution of (1) and (2):

$$f(\hat{x}) \leq f(\hat{x}) + \lambda^T (b - A\hat{x}) \leq f(\bar{x}) + \lambda^T (b - A\bar{x})$$

ALgorithms

Problem goes into

$$\min_{\lambda} \max_{x} L(\lambda, x) = f(x) + \lambda^T (b - Ax)$$

In this case, Lagrangian is

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

can be solved by iteration till all $A \sum_i y^i - b \leq 0$ or reach maximal number of iterations

Experiments

vSRL

Dataset

imSitu, 75702 for training, 25200 for developing, 25200 for test

212 verbs after flitering out non-human verbs

Model

situation y, the combination of activity v and realized frame, a set of semantic role-noun pairs (e, n_e), giving an image i as

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

Experiments

where potential is computed with feature f_i from CNN on input

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

MLC

Dataset

MS-COCO,annotate the genders by associated captions,removing images mentioned by both gender and weak associated ones

Model

output y,consisting of all object categories c and gender of person g giving an image i as

Experiments

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

where potential is computed with feature f_i from CNN on input

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

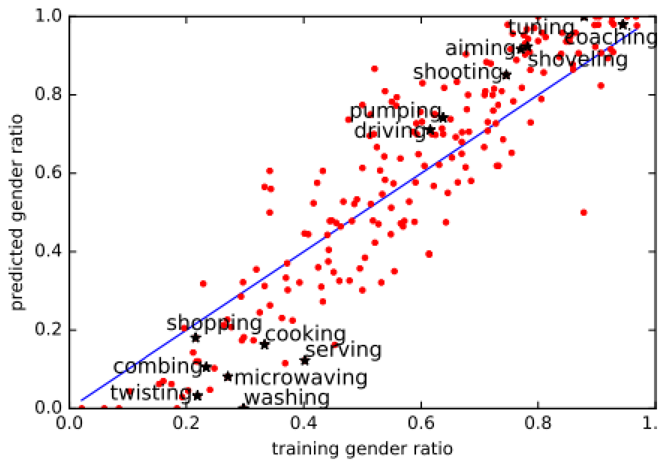
Calibration

Inference problem for both tasks

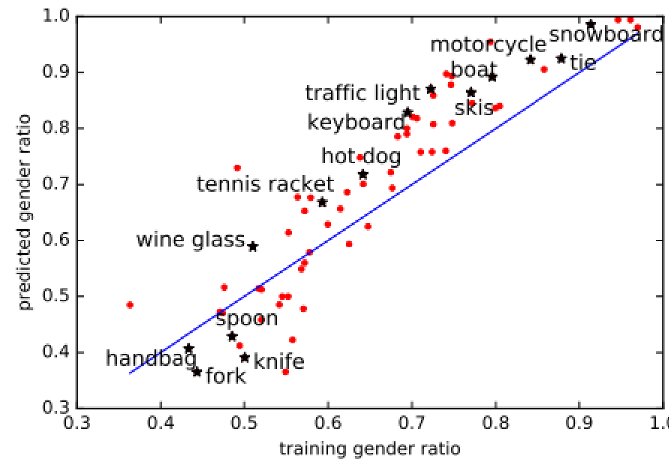
$$\arg\max_{y \in Y} f_\theta(y,i) = \log p(y|i;\theta)$$

superparameters: margin=0.05, η=0.1, iteration=100

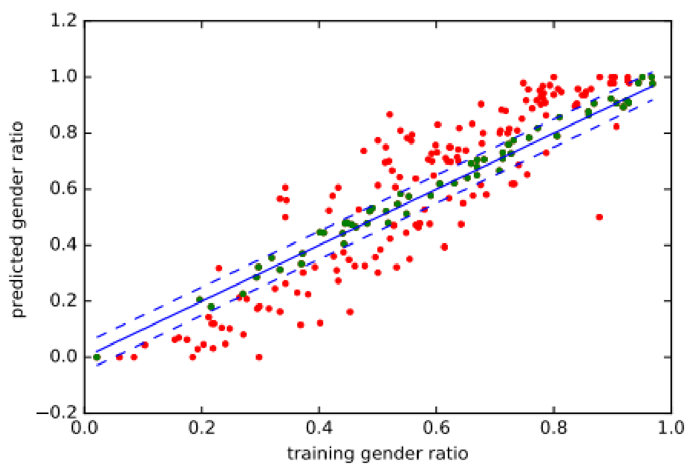
Result



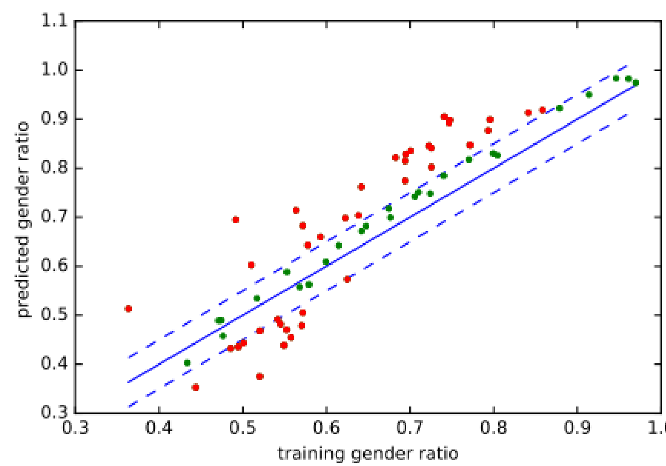
(a) Bias analysis on imSitu vSRL



(b) Bias analysis on MS-COCO MLC

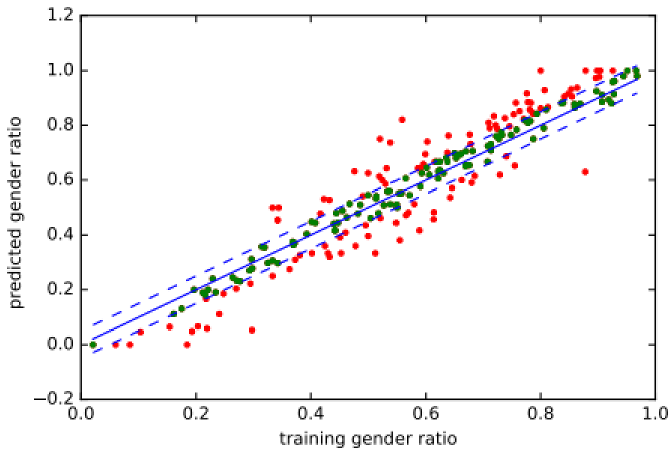


(a) Bias analysis on imSitu vSRL without RBA

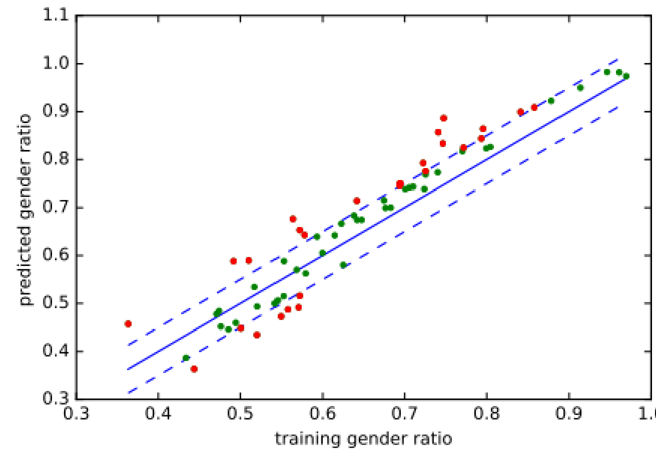


(b) Bias analysis on MS-COCO MLC without RBA

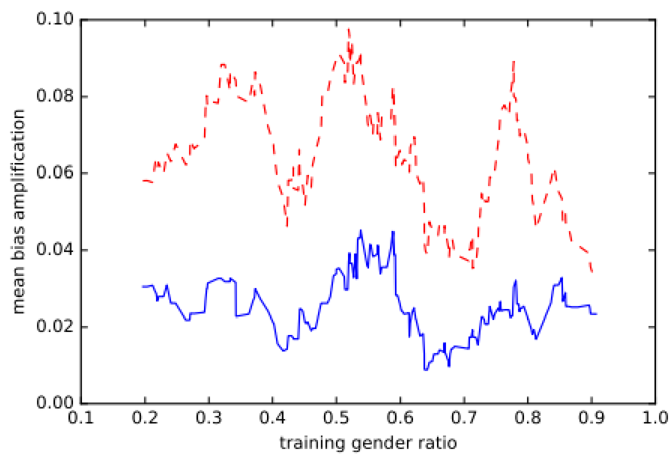
Result



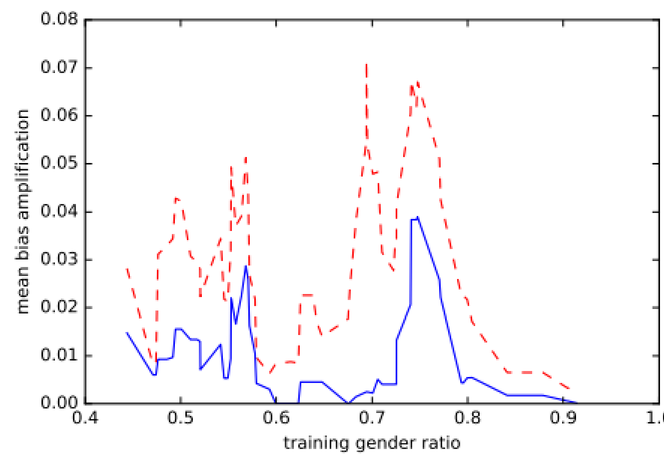
(c) Bias analysis on imSitu vSRL with RBA



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



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



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

Result

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