Google Research

Measuring Model Biases in the Absence of Ground Truth

Osman Aka*, Ken Burke*, Alex Bauerle•, Christina Greer, Margaret Mitchell•

(*Contributed equally to this work.)

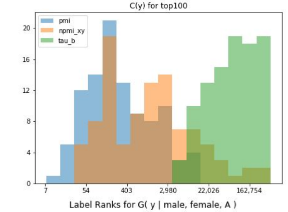
(•Work conducted while author was at Google)

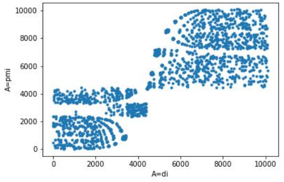
Overview

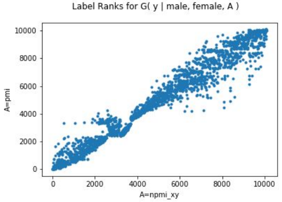
- Model bias is measured comparing ***predictions and groundtruth labels*** (e.g. Equality of Opportunity)[1]
- We present an alternative that measures associations between classifier predictions ***without using groundtruth*** in image classification.
- The statistical properties of different association metrics leads to different "most gender-biased labels".
- Normalized pointwise mutual information (nPMI) captures gender biases for both ***rare and common labels***.[2,3]

Experiments

Metrics	Min/Max $C(y)$	Min/Max $C(x_1, y)$	Min/Max $C(x_2, y)$
PMI	15 / 10,551	1 / 1,059	8 / 7,755
DP	6,104 / 785,045	628 / 239,950	5,347 / 197,795
$nPMI_{xy}$	34 / 270,748	1 / 144,185	20 / 183,132
τ_b	6,104 / 785,045	628 / 207,723	5,347 / 183,132







- We computed ***multiple association metrics*** between predicted labels in the Open Images Dataset and ranked which labels are ***most biased towards "Man" or "Woman"?***
- The top 100 "most gender-biased" labels were different for different association metrics.
- Most metrics detected either ***rare or common labels with gender bias***, and some were correlated into clusters.
- Only normalized pointwise mutual information (nPMI) detected ***both rare and common labels with gender bias***.

Intuition

- We define an association gap for label y between two identity labels [x1, x2] with respect to the association metric as:

$$G(y|x_1, x_2, A(\cdot)) = A(x_1, y) - A(x_2, y)$$

- We consider several association metrics A(·) that can be applied given the constraints of the problem - limited groundtruth, non-linearity, and limited assumptions about the distribution of the data.
For example, Demographic Parity (DP) and normalized pointwise mutual information (nPMI):

$$G(y|x_1, x_2, DP) = P(y|x_1) - P(y|x_2)$$

$$G(y|x_1, x_2, nPMI_y) = \frac{ln\left(\frac{p(x_1,y)}{p(x_1)p(y)}\right)}{ln\,(p(y))} - \frac{ln\left(\frac{p(x_2,y)}{p(x_2)p(y)}\right)}{ln\,(p(y))}$$

- All of these metrics quantify label associations in a dataset, however in practice they yield different results.

	$\partial p(y)$	$\partial p(x_1, y)$
∂DP	0	$\frac{1}{p(x_1)}$
∂PMI	0	$\frac{1}{p(x_1,y)}$
$\partial nPMI_y$	$\frac{ln(\frac{p(x_2\|y)}{p(x_1\|y)})}{ln^2(p(y))p(y)}$	$\frac{1}{ln(p(y))p(x_1,y)}$
$\partial nPMI_{xy}$	$\frac{1}{ln(p(x_1,y))p(y)} - \frac{1}{ln(p(x_2,y))p(y)}$	$\frac{ln(p(y))-ln(p(x_1))}{ln^2(p(x_1,y))p(x_1,y)}$
∂PMI^2	0	$\frac{2}{p(x_1,y)}$
∂SDC		$\frac{1}{p(x_1)+p(y)}$
∂JI		$\frac{p(x_1)+p(y)}{(p(x_1)+p(y)-p(x_1,y))^2}$
∂LLR	0	$\frac{1}{p(x_1,y)}$
$\partial\tau_b$		$\frac{(2-\frac{4}{n})}{\sqrt{(p(x_1)-p(x_1)^2)(p(y)-p(y)^2)}}$
$\partial t\text{-}test\_gap$	$\frac{\sqrt{p(x_2)}-\sqrt{p(x_1)}}{2\sqrt{p(y)}}$	$\frac{1}{\sqrt{p(x_1)*p(y)}}$

Conclusion

Metric A	DP		PMI		$nPMI_{xy}$	
Ranks	Label y	Count	Label y	Count	Label y	Count
0		265,853	Dido Flip	140		610
1		270,748	Webcam Model	184	Dido Flip	140
2		221,017	Boho-chic	151		2,906
3		166,186		610	Eye Liner	3,144
4	Beauty	562,445	Treggings	126	Long Hair	56,832
5	Long Hair	56,832	Mascara	539	Mascara	539
6	Happiness	117,562		145	Lipstick	8,688
7	Hairstyle	145,151	Lace Wig	70	Step Cutting	6,104
8	Smile	144,694	Eyelash Extension	1,167	Model	10,551
9	Fashion	238,100	Bohemian Style	460	Eye Shadow	1,235
10	Fashion Designer	101,854		78	Photo Shoot	8,775
11	Iris	120,411	Gravure Idole	200	Eyelash Extension	1,167
12	Skin	202,360		165	Boho-chic	460
13	Textile	231,628	Eye Shadow	1,235	Webcam Model	151
14	Adolescence	221,940		156	Bohemian Style	184

- We showed that the different normalizations in each metric affect whether the metric is capable of detecting gender bias in labels with high or low marginal frequencies (i.e., common or rare labels).
- The nPMI metric is preferable to other commonly used association metrics in the problem setting of detecting biases without access to groundtruth labels.
- Future research is needed to:
  - de-associate patterns at model training time.
  - capture within-image label relationships and context.

References

[1] Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning.

[2] Church, K. W.; and Hanks, P. 1990. Word Association Norms, Mutual Information, and Lexicography. Computational Linguistics 16(1): 22–29. URL https://www.aclweb.org/anthology/J90-1003.

[3] Bouma, G. 2009. Normalized (pointwise) mutual information in collocation extraction.

Author Contact

{osmanaka, kenburke, ckuhn}@google.com

bauerlealex@gmail.com

margarmitchell@gmail.com