An Investigation of Representation & Allocation Harms in Contrastive Learning

Subha Maity[1], Mayank Agarwal[2], Mikhail Yurochkin[2], & Yuekai Sun[1]

[1]Department of Statistics, University of Michigan
[2]IBM Research, MIT-IBM Watson AI Lab

LSA STATISTICS UNIVERSITY OF MICHIGAN

MIT-IBM Watson AI Lab

IBM **Research**

Representation harm due to underrepresentation

Representation: contrastive learning (CL) (Chen et al., 2020)

» **Controlled study** on CIFAR10 dataset (Krizhevsky et al., 2009)

» **Q:** Biases/harms from underrepresentation of "`automobile`" class in CL?

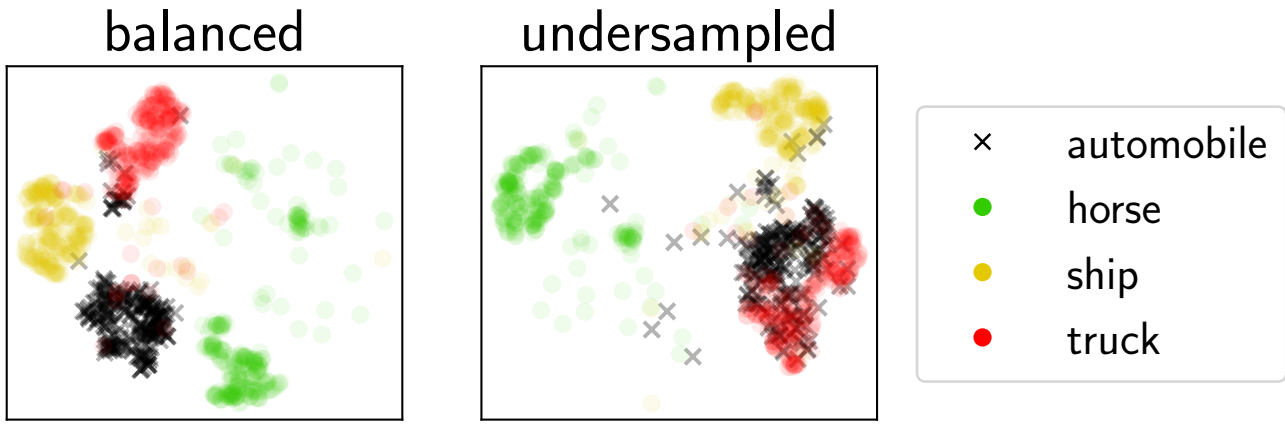
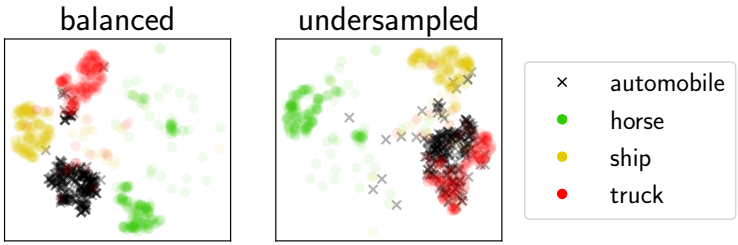


Fig: Two-dimensional t-SNE embeddings

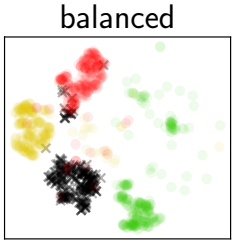
Stereotyping: a simplified representation of a (heterogeneous) group, sometimes causing errors in judgment (Bordalo et al., 2016)

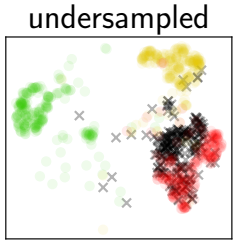
Stereotyping due to underrepresentation

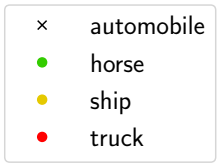


$$\mathrm{RH}(y, y') = \frac{\sum_{i,j} \text{cos-dist}\left(\varphi_{y,i}^{(\text{under})}, \varphi_{y',j}^{(\text{under})}\right)}{\sum_{i,j} \text{cos-dist}\left(\varphi_{y,i}^{(\text{bal})}, \varphi_{y',j}^{(\text{bal})}\right)} (< 1?)$$

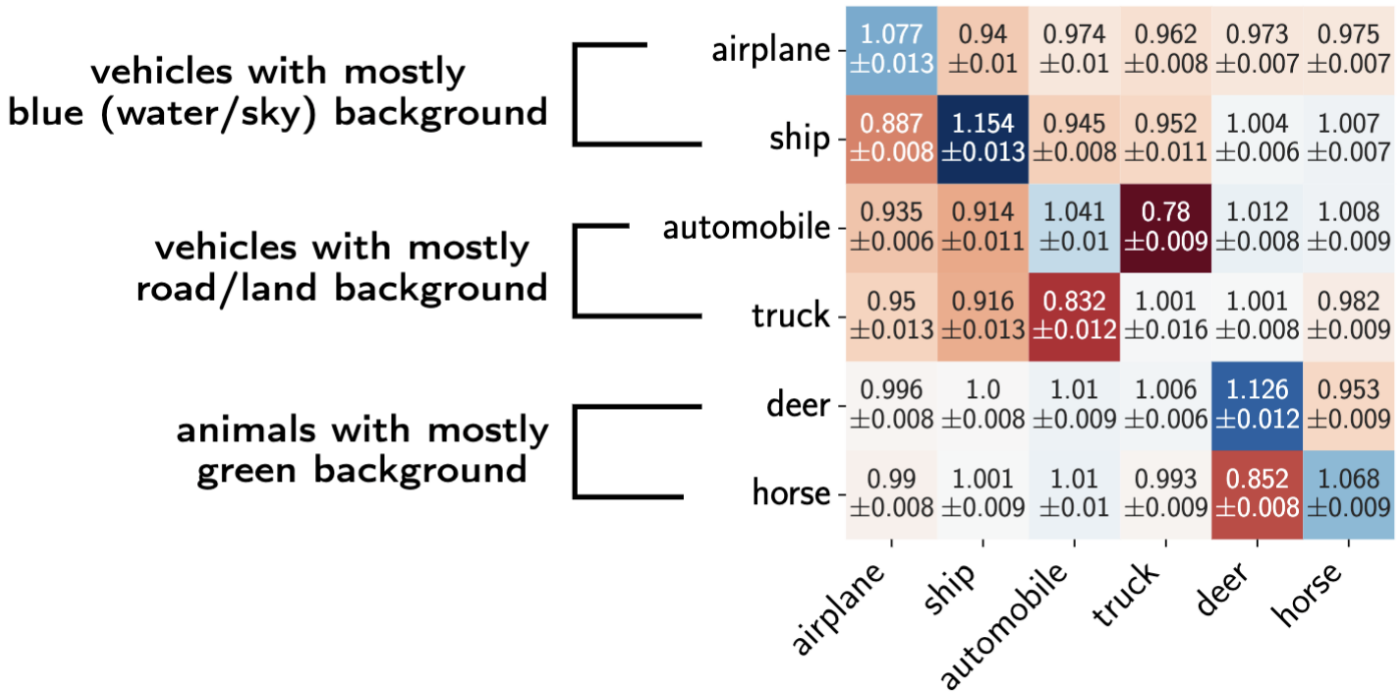
Stereotyping due to underrepresentation







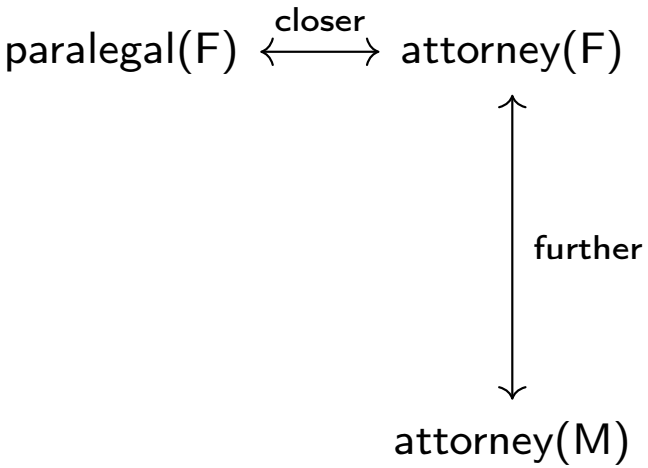
$$\mathrm{RH}(y, y') = \frac{\sum_{i,j} \text{cos-dist}\left(\varphi_{y,i}^{(\text{under})}, \varphi_{y',j}^{(\text{under})}\right)}{\sum_{i,j} \text{cos-dist}\left(\varphi_{y,i}^{(\text{bal})}, \varphi_{y',j}^{(\text{bal})}\right)} (< 1?)$$

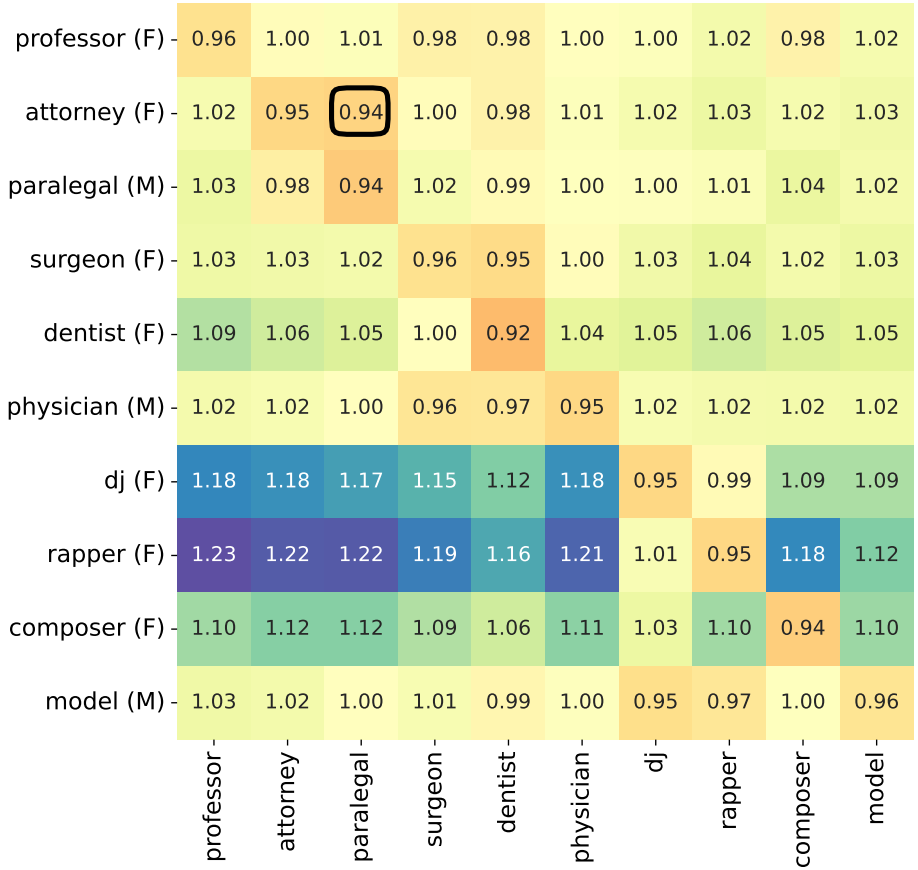


Key obs: undersampling leads to harms of representation among similar classes!

Stereotyping in text representation

Bias in Bios(De-Arteaga et al., 2019): "`attorney`" ($\sim$ 38% female) vs. "`paralegal`" ($\sim$ 85% female).



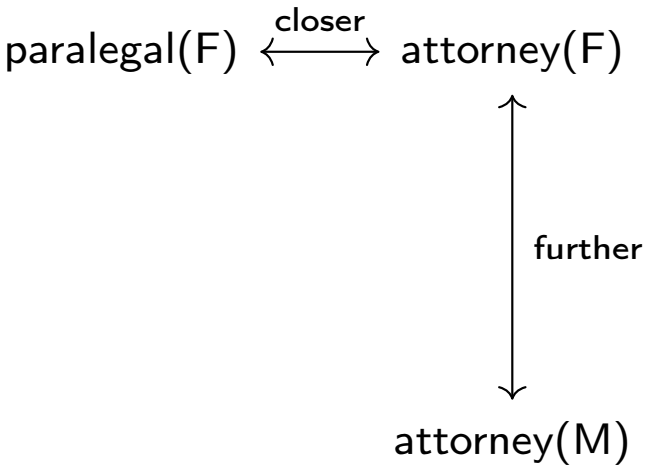


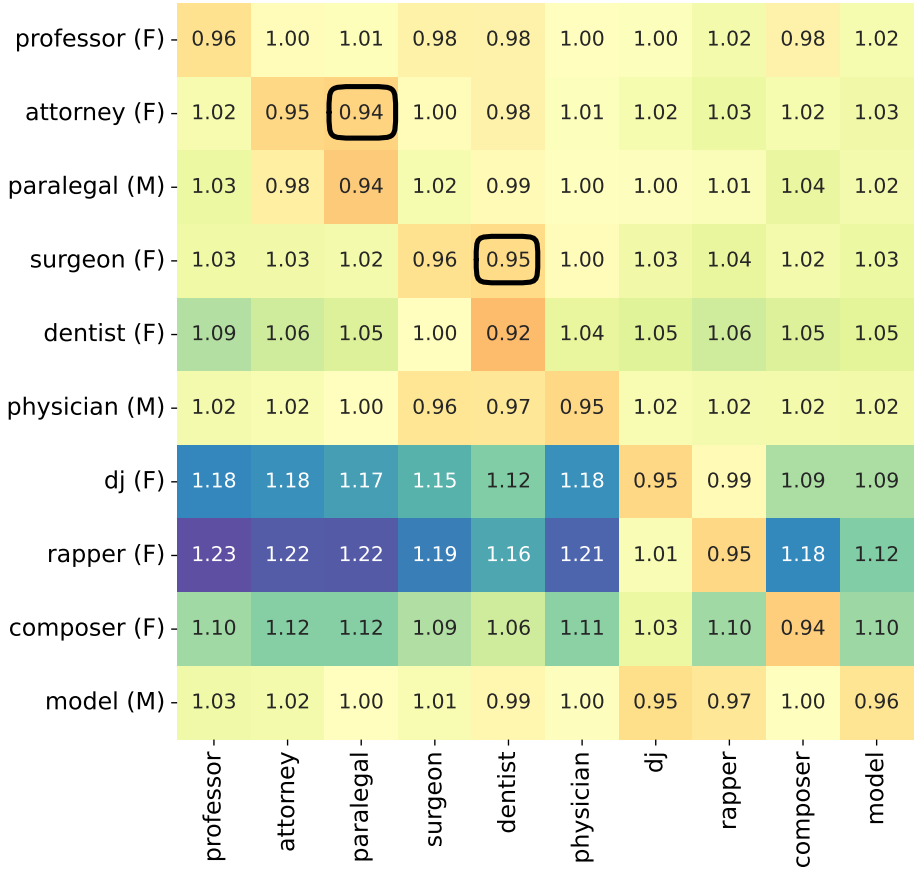
$$\mathrm{RH}_{\mathsf{F}} = \frac{\text{cos-dist}\{\text{attorney(F)}, \text{paralegal(F)}\}}{\text{cos-dist}\{\text{attorney(F)}, \text{attorney(M)}\}}$$

» Theoretical analysis of stereotyping in our paper.

Stereotyping in text representation

Bias in Bios(De-Arteaga et al., 2019): "`attorney`" ($\sim$ 38% female) vs. "`paralegal`" ($\sim$ 85% female).





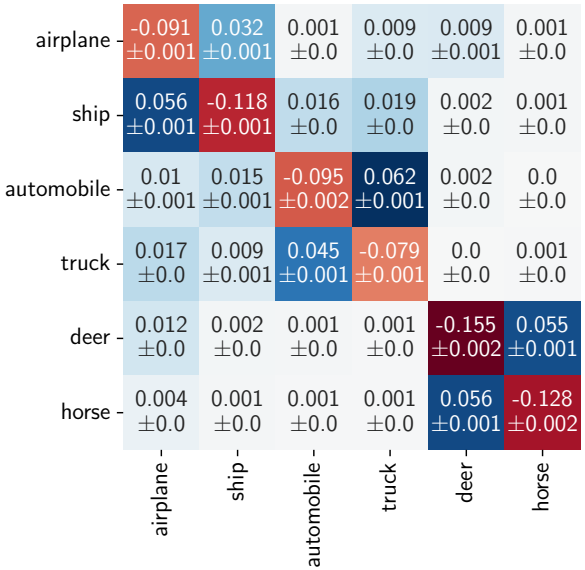
$$\mathrm{RH}_{\mathsf{F}} = \frac{\text{cos-dist}\{\text{attorney(F)}, \text{paralegal(F)}\}}{\text{cos-dist}\{\text{attorney(F)}, \text{attorney(M)}\}}$$

» Theoretical analysis of stereotyping in our paper.

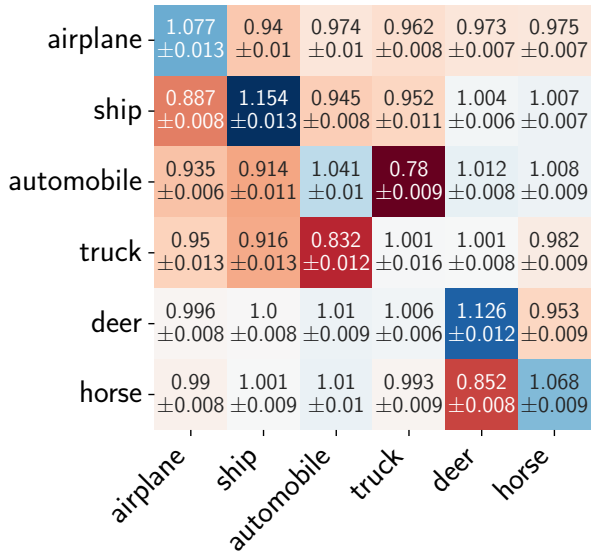
Stereotyping leads to harms of allocation

Q: Does stereotyping lead to an increase in misclassification rate? Fit a linear classifier with CL embedding.

Metric: $\mathrm{AH}(y, y') = P\{\hat{y}^{(y)} = y' \mid y\} - P\{\hat{y}^{\star} = y' \mid y\}$



(a) allocation harm (AH)



(b) representation harm (RH)

Key obs: stereotyping leads to harms of allocation among similar classes.

Thank you!!!

Paper: An Investigation of Representation and Allocation Harms in Contrastive Learning

Poster: Session 7; Friday, May 10th; 10:45am - 12:45pm local time

Contact info: Subha Maity; `smaity@umich.edu`