Our Contributions:

We model the biases in multimodal datasets as confounders in causal graph.

We learn confounders by:
- minimizing information in biased representations
- maximizing the task accuracy

We propose two debiasing methods using these confounder to debias multimodal models.

Causal Treatment Effect-Debiasing (ATE-D)

Average Treatment Effect

\[ P(A|do(M)) = E_{C=M}[P(A|M, c)] \]

Average Treatment Effect computes the expected value over the distribution of confounders to eliminate the direct effect of C on M.

Total Effect

\[ TE = A_{M=C} - A_{M=C} \]

Total Effect eliminates the direct effect of C on M and A by taking the difference between biased A with and without treatment from M.

OOD Generalization

Our method based on causal info. minimization:
- improves OOD acc. without hurting ID acc.
- removes biases arising from both unimodal and multimodal interaction

Data augmentation approaches are cumbersome but more effective than feature based debiasing.

Robustness to Spurious Features

We propose sufficiency score (λ) as the percentage of the model's certainty attributed to the spurious input component in prediction.

\[ \lambda = \frac{\sum_{i=1}^{G} KL(f[y_i|x_i'])|U}{\sum_{i=1}^{G} KL(f[y_i|x_i'])|U} \]

VL Biases

What is the color of the banana?

Confounder Analysis

Total Effect-Debiasing (TE-D)

- Answers from 0.34% of the vocabulary address 67% of training questions
- Most frequent answers obtained from biased representations align with those in train set, indicating effective representation of dataset biases

Average Treatment Effect-Debiasing (ATE-D)

- Boosting biased features hurts OOD accuracy
- We train a non-linear probe on confounder representations for the VQA task
  - Probe's accuracy is 25%
  - Probe's predicted answer distribution has lower entropy than unbiased features' predicted answer distribution

Acknowledgements

We thank Peter Hase, Zhichuan Yin, and Jaemin Cho for their useful insights about this work, and the reviewers of this paper for their helpful feedback. This work was supported by AFOSR/AFRL/IT210220, DARPA/CA/N66001-19-24031, ONRN00014-23-1-2356, DARPA ECHELON Program No. HR0011235010. The views, opinions, and/or findings contained in this article are those of the authors and not of the funding agency.

DEBIASING MULTIMODAL MODELS VIA CAUSAL INFORMATION MINIMIZATION
Vaidehi Patil, Adyasha Maharana and Mohit Bansal