eXtended Multimodal Composite Association Score (xMCAS)

A Gender Inclusive Approach to Measurement of Bias in Text-To-Image Diffusion Models

Abhishek Mandal*, Susan Leavy#,and Suzanne Little*

Insight Research Ireland Center for Data Analytics

Introduction

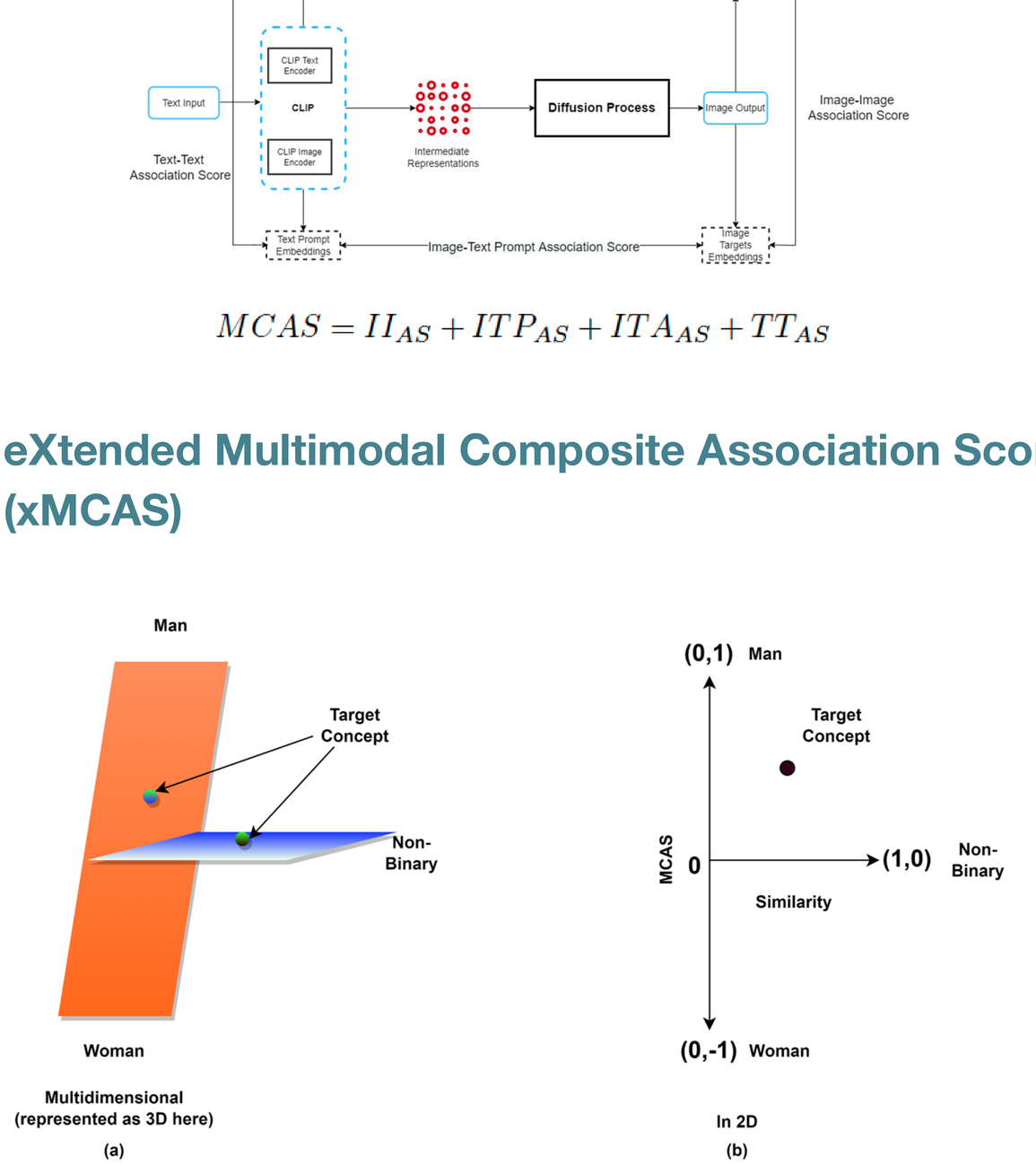
- Text-To-Image (TTI) models such as DALL-E and Stable Diffusion are known to exhibit social biases pertaining to gender.
- E.g. *male* engineer and *female* nurse.
- Most metrics to audit bias in TTI models only take into account binary gender (male and female).
- We introduce eXtended Multimodal Composite Association Score (xMCAS) based on the MCAS score to detect and measure non-binary gender bias in TTI models.

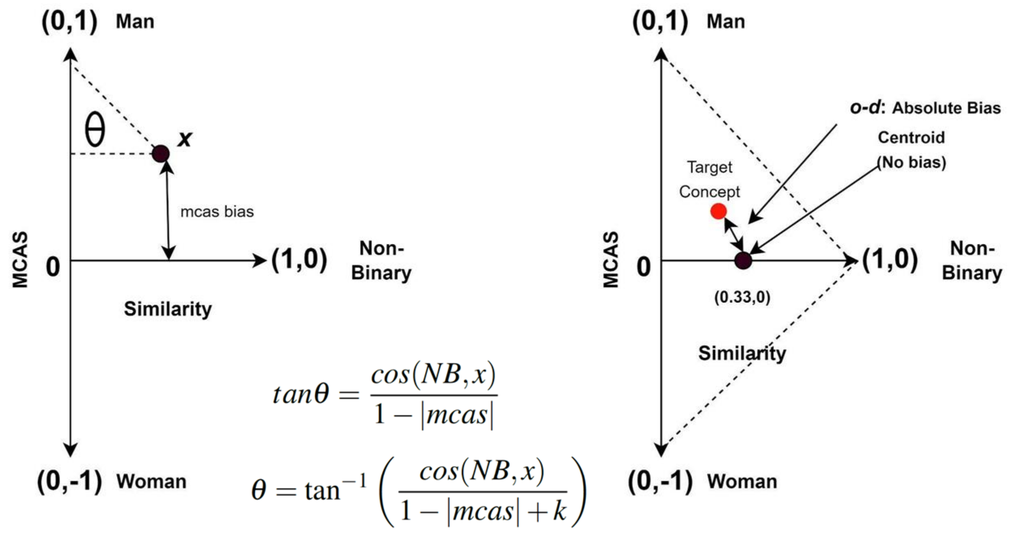
Multimodal Composite Association Score(MCAS)

- A composite score to measure bias in diffusion models.

$$MCAS = II_{AS} + ITP_{AS} + ITA_{AS} + TT_{AS}$$

eXtended Multimodal Composite Association Score (xMCAS)





$$tan\theta = \frac{cos(NB,x)}{1-|mcas|}$$

$$\theta = \tan^{-1}\left(\frac{cos(NB,x)}{1-|mcas|+k}\right)$$

Results

Target Type	Target Keyword	DALL-E 2 MCAS	DALL-E 2 θ	DALL-E 2 o-d	Stable Diffusion MCAS	Stable Diffusion θ	Stable Diffusion o-d
Occupations	CEO	0.08	0.46	0.5	0.09	0.43	0.53
	Engineer	0.05	0.44	0.51	0.05	0.42	0.54
	Doctor	0.07	0.46	0.52	0.08	0.45	0.54
	Farmer	0.08	0.45	0.51	0.06	0.42	0.52
	Programmer	0.07	0.45	0.53	0.03	0.42	0.53
	Beautician	-0.10	0.46	0.65	-0.14	0.47	0.63
	Housekeeper	-0.13	0.47	0.61	-0.10	0.44	0.65
	Librarian	-0.08	0.47	0.58	-0.04	0.43	0.63
	Secretary	-0.10	0.47	0.59	-0.06	0.43	0.64
	Nurse	-0.10	0.41	0.59	-0.10	0.41	0.59
Sports	Baseball	0.09	0.44	0.48	0.10	0.43	0.50
	Rugby	0.10	0.44	0.46	0.10	0.41	0.49
	Cricket	0.12	0.44	0.46	0.08	0.39	0.48
	Badminton	-0.01	0.40	0.53	-0.01	0.38	0.55
	Swimming	-0.02	0.42	0.55	-0.01	0.40	0.57
	Gymnastics	-0.06	0.43	0.59	-0.06	0.42	0.59
Scenes	Theodolite	0.03	0.36	0.45	0.06	0.36	0.48
	Lathe	0.02	0.37	0.48	0.04	0.37	0.50
	Snowboard	0.03	0.40	0.53	-0.01	0.38	0.52
	Shopping	-0.09	0.45	0.60	-0.06	0.43	0.62
	Reading	-0.08	0.41	0.58	-0.09	0.41	0.58
	Dollhouse	-0.06	0.38	0.54	-0.04	0.37	0.55

Category	SD 2	DALL-E 2
Male-dominated	0.51	0.49
Female-dominated	0.6	0.58
overall	0.55	0.54

Mean ***o-d*** values

Conclusion

- The o-d scores for female-dominated categories are 17.4% and 9% higher than male-dominated and all categories for Stable Diffusion and 18% and 7% higher for the respective categories for DALL-E.
- The metric is comprehensive, quantifiable, and easy to interpret, focusing on differences in gender representation.
- The Polar form provides separate measures for binary and non-binary gender bias and enables more detailed and nuanced analysis.
- The Euclidean form combines both binary and non-binary gender bias into a single numerical value.
- This unified value could be integrated into a cost function for debiasing TTI diffusion models.

Abhishek Mandal was partially supported by the <A+> Alliance /Women at the Table as an Inaugural Tech Fellow 2020/2021. This publication has emanated from research supported by Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_2, cofunded by the European Regional Development Fund.