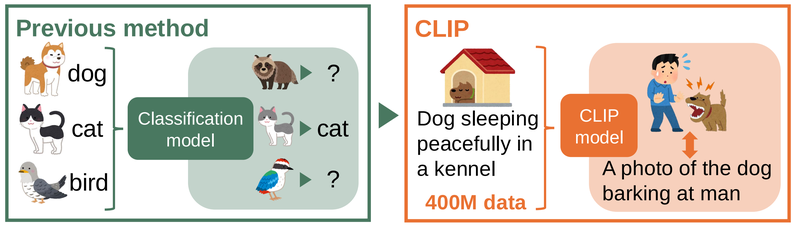
The APA Benchmark: Probing Vision-Language Models for Capabilities, Societal Bias and Retention

Yuri Ishitoya [1,2], Veronica Flores [3], Ziyan Yang [4], Paola Cascante-Bonilla [4], Vicente Ordóñez-Román [4]

[1]School of Science, Ochanomizu University, [2]TOMODACHI STEM @ Rice Program, [3]Santa Clara University, [4]Dept. of Computer Science, Rice University

Background

❖ Advances in computer vision systems



❖ Concerns about multimodal models of language and image

400M data → CLIP model
- ○ Trained with an open vocabulary
- ○ Exposed to many objects, concepts and image

Susceptible to misclassification and bias & Prone to ambiguity in prediction

APA Benchmark Overview

❖ Proposing a complementary benchmark

- ○ A dataset of about 100-400 high quality portrait pictures of ***Actors/actress, Politicians and Athletes (APA)***
- ○ Developed for exploring below three indicators

1) Capabilities : Classification ability
2) Societal Bias : Gender bias in classifying task
3) Retention : Models based on individual info or image features

❖ Details of portraits*1

Category	Number of portaits	Fundamental Infomation
Actors / Actress	○ Actors: 60 ○ Actress: 40	Name, Film name, Academy Awards Biography
Politicians (U.S. Congress members as of July 2022)	○ Senators: 100 (M: 76, F: 24) ○ House: 436 (M: 310, F: 126) ○ Mayors*2: 100 (M: 67, F: 33)	Name, State, Birth Date, Party
Athletes	○ Male athletes: 79 ○ Female athletes: 30	Name, Sport

*1 Most are official photos released by the office in Wikipedia *2 From the top 100 largest cities

Models and Prompts

❖ Models

CLIP / Open-CLIP / ALBEF / BLIP / SigLIP / MetaCLIP / EVA-CLIP

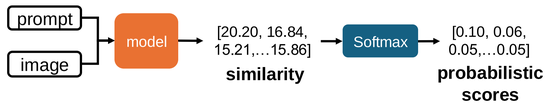
❖ Prompts

Methods and Results

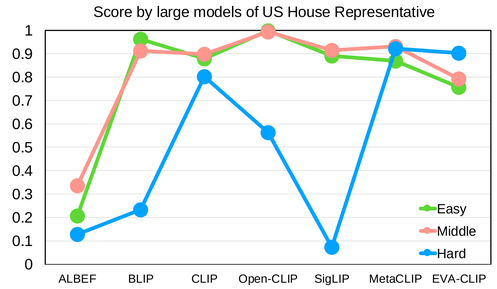
1) Capabilities

1. Enter level-specific category prompts and images into the models
   - ○ **Easy : person or not**
     - ▪ person, dog, giraffe, plant, tree, bed, chair
   - ○ **Middle : occupation**
     - ▪ politician, scientist, athlete, teacher, receptionist, assistant, salesperson, actor/actress
   - ○ **Hard : info not shown in image**
     - ▪ Ex) soccer player, senator, academy award winner

2. Calculate similarity and score each prompt level probabilistically



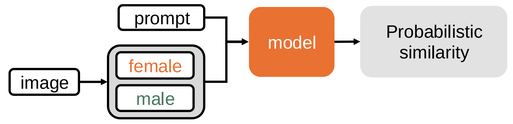
❖ Result



- ○ **CLIP & MetaCLIP & EVA-CLIP**
  - ▪ Higher score for all prompt
- ○ **BLIP & SigLIP**
  - ▪ Lower score for "Hard" category

2) Societal Bias

1. Separate the dataset to male and female
2. Calculate probabilistic similarity scores
3. Calculate the bias score at middle level prompts: **b = (f score) / (m score)**



❖ Result

Actor / Actress (Two lowest bias scores)

Model	Gender	scientist	politician	athlete	teacher	receptionist	assistant	salesperson	actor/actress	Ratio
MetaCLIP b16 400m	woman	0.36	0.23	0.34	0.82	0.51	0.44	0.63	75.00	0.811
	man	0.69	2.44	0.19	0.37	0.01	0.09	1.27	92.53	
CLIP ViT-B/16	woman	14.17	5.70	4.04	12.36	31.93	14.75	6.20	65.28	0.813
	man	20.26	16.56	3.93	18.68	1.74	10.45	14.34	80.32	

Politician (Two lowest bias scores)

SigLIP b16 256	woman	0.46	20.11	0.62	52.54	4.82	2.93	0.08	16.70	0.226
	man	1.04	89.01	1.28	4.86	0.29	0.26	0.36	2.89	
SigLIP b16 384	woman	0.71	23.95	0.55	41.06	5.69	3.10	0.05	22.05	0.272
	man	1.40	87.99	1.47	3.95	0.40	0.37	0.24	4.17	

Athlete (Two lowest bias scores)

MetaCLIP b32 400m	woman	0.19	0.31	75.54	1.43	0.13	2.12	0.18	18.93	0.927
	man	0.37	2.40	81.45	1.81	0.07	0.93	0.95	10.27	
OpenCLIP ViT-B/32	woman	0.03	0.31	59.38	0.18	0.19	39.67	0.21	0.03	0.931
	man	0.76	4.80	63.77	1.90	0.40	24.82	2.56	0.98	

Close to 1.0 = non biased
Close to 0.0 = highly biased

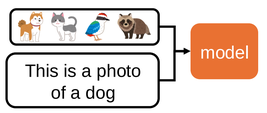
Actor/Actress and Athlete dataset:
- ○ All models:
  - ▪ **Didn't show** much gender bias

Politician dataset:
- ○ SigLIP and other smaller models:
  - ▪ Women are less likely to become **politician**

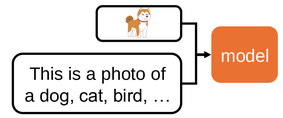
Scores between the category:
- ○ Women are selected for **receptionist & assistant**
- ○ Men are selected for **scientist & salesperson**

3) Retention

Image Score
: Given a name prompt, score all images



Text Score
: Given an image, score all name prompts



❖ Result

	US House of Rep.		Senators		Mayors		Actor/Actress		Athletes			
	Text	Image	Text	Image	Text	Image	Text	Image	Text	Image		
BLIP	1.81	2.27	12.82	13.64	6.31	6.56	38.26	42.70	31.96	42.31		
CLIP	41.11	39.60	94.73	93.07	35.13	33.79	96.39	97.51	88.48	88.92		
SigLIP	10.10	4.53	18.38	16.75	6.06	6.86	71.58	70.56	86.43	87.16		
MetaCLIP	66.50	60.99	98.44	97.75	39.50	40.31	96.68	98.19	91.65	92.24		

- ○ **Senator & Actor/Actress & Athlete:**
  - ▪ Most models classify based on personal information
- ○ **ALBEF & BLIP & SigLIP:**
  - ▪ Less likely to rely on personal information

Conclusions & Future Directions

❖ Conclusion

- ○ **Different train approach** (BLIP) & **Less training data** (SigLIP):
  - ▪ Low classify ability & High bias & Not relied on individual information
- ○ **Huge training data** (CLIP & Open-CLIP & MetaCLIP & EVA-CLIP)
  - ▪ High classify ability & Low bias & Relied on individual information

❖ Future Directions

- ○ Conduct analysis with generative VLMs
- ○ Investigate what contributes the bias

Acknowledgment

This research was conducted as part of the TOMODACHI STEM @ Rice University Program, funded by the U.S.-Japan Council. More information can be found at http://tomodachistem.rice.edu. I extend our sincere gratitude to the *vislang* members for their support.

Reference

- ○ A. Radford, J. W. Kim, C. Hallacy, et al. 2021. Learning transferable visual models from natural language supervision. ICML, 8748–8763.
- ○ J. Wei, Y. Tay, R. Bommasani, et al. 2022. Emergent abilities of large language models. TMLR.
- ○ J. Zhao, T. Wang, M. Yatskar, V. Ordonez, K.-W. Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. NAACL-HLT, 15–20.