Biased Attention: Do Vision Transformers Amplify Gender Bias More than Convolutional Neural Networks?

Abhishek Mandal*, Susan Leavy#,and Suzanne Little*
Insight SFI Research Center for Data Analytics
*School of Computing Dublin City University, Dublin, Ireland #School of Information and Communication Studies University College Dublin, Dublin, Ireland

Insight SFI RESEARCH CENTRE FOR DATA ANALYTICS

BMVC 2023

A World Leading SFI Research Centre

Science Foundation Ireland For what's next

Introduction

- Vision Transformers (ViT), have increasingly become important as they outperform Convolutional Neural Networks (CNN) in many domains.
- Vision models have been shown to exhibit social biases. Most metrics to detect them have been limited to CNNs.
- We aim to answer the following **research questions**:
    - Is gender bias exhibited differently by CNNs and ViTs?
    - How can the effect of gender bias in both CNNs and ViTs be measured?

Measuring Bias

- **Accuracy Difference:**
    - Class balanced dataset $\mathbb{D}(X_i, Y_i, g_i)$ [X_i:image,Y_i:label,g_i:protected attribute (gender)]
    - $g_i \in \{m, w\}$, (m : men, w : women)
    - $\mathbb{D}_{\text{balanced}} \subset \mathbb{D}$; $f(g_i(m = w))$
    - $\mathbb{D}_{\text{imbalanced}} \subset \mathbb{D}$; $f(g_i(m > w \ \vee \ m < w))$
    - $\mathbb{D}_{\text{test}} \subset \mathbb{D}$
    - Let image classifiers $\boldsymbol{M}_{\text{unbiased}}$ be trained on $\mathbb{D}_{\text{balanced}}$ and $\boldsymbol{M}_{\text{biased}}$ be trained on $\mathbb{D}_{\text{imbalanced}}$ having an accuracy of $\boldsymbol{A}_{\text{biased}}$ and $\boldsymbol{A}_{\text{unbiased}}$ on $\mathbb{D}_{\text{test}}$ respectively
    - **Accuracy Difference(Δ) =** $|\boldsymbol{A}_{\text{unbiased}} - \boldsymbol{A}_{\text{biased}}|$

- **Image-Image Association Score (IIAS)**

For two images I_1 and I_2, with extracted features v_1 and v_2 respectively, we calculate image similarity and IIAS as:

$$sim(I_1, I_2) = \frac{v_1 \cdot v_2}{||v_1||_2 \cdot ||v_2||_2} \qquad IIAS = mean_{w \in W} s(w, A, B)$$

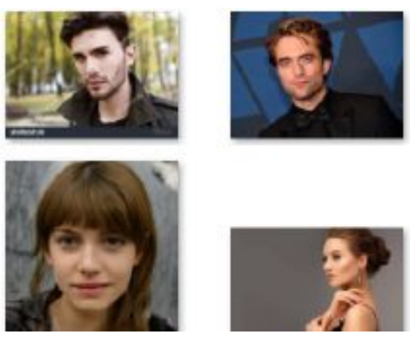
$$s(w, A, B) = mean_{a \in A} sim(\vec{w}, \vec{a}) - mean_{b \in B} sim(\vec{w}, \vec{b})$$

$$IIAS \in [-1, 1]$$

A and B: images of men and women; W: real-world concept e.g., occupation (images). Features extracted from final pre-fully connected layer for CNNs and the final pre-MLP layer for ViTs.

The Dataset

Main occupations dataset; CEO (L) & Nurse. Masked images at the bottom.



Gender attributes. Men (top).

Methodology

- **Bias Analytics using Image Classifiers**
    - 4 CNN models: VGG16, ResNet152, Inceptionv3, Xception; 4 ViT models: ViT B/16, B/32, L/16, L/32.
    - All pre-trained on Imagenet and fine-tuned on balanced and imbalanced dataset.
    - Trained 80 models: (4 CNNs & 4 ViTs) x 2 (biased & unbiased) x 5 iterations.
- **Bias Analytics using CLIP**
    - 4 different CLIP image encoders: CNNs ResNet 50 and 50x4 and ViTs ViT B/16 and B/32.
    - CLIP zero-shot predictions using 100 occupations and the gender attributes dataset.

Findings

Model Type	Model Name	Mean Δ	Average Model Δ	Mean % Δ	Average Model %Δ
CNN	Inception	0.1	0.11	15	16.88
	ResNet152	0.18		24.24	
	VGG16	0.1		18.36	
	Xception	0.06		10	
ViT	ViT-B16	0.17	0.17 (54% ↑)	39.19	37.8 (123% ↑)
	ViT-B32	0.18		39	
	ViT-L16	0.13		31	
	ViT-L32	0.2		42	

	Masked				Unmasked			
	Biased		Unbiased		Biased		Unbiased	
Class	CNN	ViT	CNN	ViT	CNN	ViT	CNN	ViT
CEO	0.059	0.1	0.26	0.02	0.05	0.17	0.07	0.06
Engineer	0.23	0.14	0.36	0.17	0.18	0.19	0.04	0.21
Nurse	-0.14	-0.35	-0.05	-0.2	-0.21	-0.21	-0.06	-0.17
School Teacher	-0.17	-0.15	-0.12	-0.05	-0.02	-0.4	-0.04	-0.14
Total IIAS (absolute)	0.599	0.74	0.79	0.44	0.46	0.97	0.21	0.58
% Difference		23% ↑	80% ↑			111% ↑		176% ↑

Image Encoder	Man Occurrence	Top 3 Predictions	Woman Occurrence	Top 3 Predictions
RN 50	47	mathematician, psychiatrist’youtuber	49	beautician, student, housekeeper
RN 50x4	46	investment banker, economist, coach	56	housekeeper, jewellery maker, midwife
ViT B/16	50	coach, psychiatrist, administrator	54	midwife, beautician, jewellery maker
ViT B/32	45	chief executive officer, musician, hairdresser	63	beautician, housekeeper, jewellery maker
CNN	46.5		52.5	
ViT	48 (3.3 % ↑)		59 (12.53 % ↑)	

Accuracy Difference (top), IIAS (middle), and CLIP ZS (bottom)

Conclusions

ViTs amplify gender bias due to:

- A shallower loss landscape leading to better generalisation.
- Global attention and a larger receptive field due to the multi-headed self-attention mechanism that enables them to capture more visual cues and long-term dependencies.

HOST INSTITUTIONS

PARTNER INSTITUTIONS

FUNDED BY:

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.