

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

Abhishek Mandal<sup>1</sup>  
abhishek.mandal2@mail.dcu.ie

Susan Leavy<sup>2</sup>  
susan.leavy@ucd.ie

Suzanne Little<sup>1</sup>  
suzanne.little@dcu.ie

<sup>1</sup> Insight SFI Research Centre for Data Analytics  
School of Computing  
Dublin City University  
Ireland

<sup>2</sup> Insight SFI Research Centre for Data Analytics  
School of Information and Communication Studies  
University College Dublin  
Ireland

---

## Abstract

Deep neural networks used in computer vision have been shown to exhibit many social biases such as gender bias. Vision Transformers (ViTs) have become increasingly popular in computer vision applications, outperforming Convolutional Neural Networks (CNNs) in many tasks such as image classification. However, given that research on mitigating bias in computer vision has primarily focused on CNNs, it is important to evaluate the effect of a different network architecture on the potential for bias amplification. In this paper we therefore introduce a novel metric to measure bias in architectures, Accuracy Difference. We examine bias amplification when models belonging to these two architectures are used as a part of large multimodal models, evaluating the different image encoders of Contrastive Language Image Pretraining which is an important model used in many generative models such as DALL-E and Stable Diffusion. Our experiments demonstrate that architecture can play a role in amplifying social biases due to the different techniques employed by the models for feature extraction and embedding as well as their different learning properties. This research found that ViTs amplified gender bias to a greater extent than CNNs. The code for this paper is available at: <https://github.com/aibhishek/Biased-Attention>

## 1 Introduction

Vision Transformers (ViT), derived from Transformers in Natural Language Processing, have increasingly become important as they outperform Convolutional Neural Networks in many application domains [6, 7, 8]. Unlike Convolutional Neural Networks (CNN), which rely on a sequence of convolution operations extracting information from visual data, ViTs

employ Multi-headed Self Attention (MSA) that estimate the relevance of one patch of an image with another [6, 7]. This enables ViTs to capture ‘long-term dependencies’ in the data and thus possess a larger receptive field [8]. Popular computer vision models and their applications have been shown to exhibit a large range of social biases including gender [9, 10], racial [11, 12], and geographical biases [13, 14]. Most of the work done on detecting such biases [15, 16, 17] and mitigating them [18, 19, 20] are done on CNNs. Although most of the biases originate in the training data [9, 19, 21], models themselves have been shown to amplify them [8, 17, 22]. Therefore, given the rise in popularity of vision transformers and the lack of previous research on bias detection and mitigation for them, it is crucial to investigate how ViTs handle social biases.

As the metrics developed for CNNs may not work properly for ViTs [15, 22], we introduce a novel bias detection metric: *Accuracy Difference* and adapt the *Image-Image Association Score* developed by Mandal et al. [10] to allow comparative analysis between CNNs and ViTs. To detect and study the overall effect of model architecture on gender bias, we analysed the predictions made using models based on these two architectures. We evaluate gender bias with a focus on men and women in this paper, not to reinforce a binary view of gender but with a view to study the effect of bias on model architectures. This paper aims to address the following research questions:

- Is gender bias exhibited differently by Convolutional Neural Networks and Vision Transformers?
- How can the effect of gender bias in both Convolutional Neural Networks and Vision Transformers be measured?

This paper is divided into two parts: The first part measures the effect of gender bias on four sets of CNNs and ViTs using our novel metric and the adapted metric. In the second part, we analyse the zero-shot predictions made by Contrastive Language Image Pretraining (CLIP) [23] using two sets of CNNs and ViTs. We then analyse the results by contrasting the differences between these two model architectures. Additionally, for our metrics, we created an occupation-based visual dataset by crawling images from the Internet.

## 2 Background and Related Work

Park and Kim [24] studied the various differences between CNNs and ViTs and found two key differences: the shallower learning profile for ViTs leading to better generalisation when trained on large datasets and Multiheaded Self Attention (MSA) being high pass filters and Convolutions being low pass filters. MSA enables ViTs to model full image contextual information and, coupled with the flatter loss landscape, enables ViTs to attain better generalisation and model long-range contextual information than CNNs when trained on large datasets [6]. The absence of inductive priors (which are present in CNNs) allows ViTs to attain global attention and better learn contextual cues [6].

**Measuring bias in deep neural networks:** Several metrics such as Image Embeddings Association Test [25], model leakage and bias amplification [22], and InsideBias [17] have been proposed to detect and measure gender bias in vision models. However, they have been mainly developed for and tested on Convolutional Neural Networks. With the increasing adoption of Vision Transformers, it is important to develop similar metrics for ViTs.

**Image-Image Association Score (IIAS)** developed by Mandal et al. [10] measures stereotypical associations in vision models. It is derived from the Word Embeddings Association Test in Natural Language Processing, which is itself based on the highly popular

Implicit Association Test. It estimates human-like biases in vision models by measuring the association between two sets of concepts: two attributes and a target in the model’s embeddings. The attributes in the case of gender can be man and woman, and the target can be a real-world concept like occupation. Thus, if a particular occupation (e.g. CEO) is closer to man than woman, in a model’s embedding space, then the model is biased.

**Contrastive Language Image Pretraining (CLIP)** is a large multimodal model developed by OpenAI, trained on 300 million image-text pairs crawled from the Internet [14]. It connects images with text and is trained using contrastive loss and is used in other popular generative models such as DALL-E and Stable Diffusion [15, 16]. CLIP uses a text encoder and an image encoder, with the option of CNNs (ResNet 50, 50x4, and 101) and ViTs (ViT B/16 and B/32) being provided. This enables us to study the multimodal effect of bias in these two architectures from a multimodal perspective. Although CLIP has been shown to exhibit social biases [14, 24, 25], the effect of image encoder architecture on bias is yet to be studied.

## 3 Measuring Bias

### 3.1 Accuracy Difference

For a multiclass, class-balanced visual dataset  $\mathcal{D}$  containing instances  $(X_i, Y_i, g_i)$ , where  $X_i$  is an image having class label  $Y_i$ , and a protected attribute  $g_i$  denoting gender, where  $g_i \in \{m, w\}$ , ( $m$  : men,  $w$  : women). Let  $\mathcal{D}_{balanced} \subset \mathcal{D}$ ;  $f(g_i(m = w))$ , be a dataset containing instances with protected attributes such as gender. The dataset is class balanced as well as gender-balanced, meaning all instances have an equal gender ratio. Let  $\mathcal{D}_{imbalanced} \subset \mathcal{D}$ ;  $f(g_i(m > w \vee m < w))$ , be a dataset which is class-balanced but gender imbalanced. Let  $\mathcal{D}_{test} \subset \mathcal{D}$  be a class and gender-balanced dataset. The generalisation error (misclassification rate) of a classifier trained on  $\mathcal{D}$  and tested on  $\mathcal{D}_{test}$  can be estimated as:

$$E = \frac{1}{N} \sum_{i=1}^N \mathbb{1}(y_i \neq \hat{y}_i) \quad \dots eq(1)$$

where  $\mathbb{1}(\cdot)$  is the indicator function,  $N$  is the number of samples in the dataset, and  $\hat{y}_i$  is the predicted class label. The generalisation error (misclassification rate) can also be given as:

$$E = bias + variance + unavoidable error \quad \dots eq(2)$$

If we neglect the unavoidable error and express bias and variance in terms of  $g_i$ , then  $g_i$  can be used as a proxy for  $E$ . As the accuracy of the classifier on the  $\mathcal{D}_{test}$  can be expressed as  $1 - E$ , then from eq(1) and eq(2), accuracy can be used as a proxy for bias  $g_i$ . Let image classifiers  $M_{unbiased}$  be trained on  $\mathcal{D}_{balanced}$  and  $M_{biased}$  be trained on  $\mathcal{D}_{imbalanced}$  having an accuracy of  $A_{biased}$  and  $A_{unbiased}$  on  $\mathcal{D}_{test}$  respectively.

Then we define accuracy difference ( $\Delta$ ) as:

$$\Delta = |A_{unbiased} - A_{biased}| \quad \dots eq(3)$$

If the effect of gender bias on a classifier is minimal, then  $M_{biased}$  will perform very similarly to  $M_{unbiased}$  on the gender-balanced  $\mathcal{D}_{test}$  and  $\Delta$  will be very small. However, if the effect of gender bias on the classifier is significant, then the performances of the models will differ and  $\Delta$  will be high. Higher the value of  $\Delta$ , more the effect of bias.

### 3.2 Image-Image Association Score (IIAS)

The authors of IIAS [10] used CLIP embeddings to calculate IIAS. We adapted the metric by replacing the CLIP embeddings with the image features extracted by the classifier model. In the case of CNNs, it was the output of the final pre-fully connected layer and in the case of ViTs, the final pre-MLP layer. We then used cosine distance to measure similarity. For two images  $I_1$  and  $I_2$ , with extracted features  $v_1$  and  $v_2$  respectively, we calculate image similarity as:

$$\begin{aligned} \text{sim}(I_1, I_2) &= \frac{v_1 \cdot v_2}{\|v_1\|_2 \cdot \|v_2\|_2} \quad \dots \text{eq}(4) \\ \text{sim}(I_1, I_2) &\in [0, 1] \end{aligned}$$

Then we calculate IIAS in the same way as the authors. Let  $A$  and  $B$  be two sets of images containing images of men and women, respectively called gender attributes. Let  $W$  be a set of images containing images corresponding to a real-world concept such as occupation, called target. Then the Image-Image Association Score, IIAS, is given by:

$$IIAS = \text{mean}_{w \in W} s(w, A, B) \quad \dots \text{eq}(5)$$

where,

$$s(w, A, B) = \text{mean}_{a \in A} \text{sim}(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \text{sim}(\vec{w}, \vec{b}) \quad [\text{from eq}(4)]$$

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

If IIAS is positive, then the target is closer to men showing a male bias and if IIAS is negative, then the target is closer to women, showing a female bias. The numeric value indicates the magnitude of the bias.

## 4 Experiment

The experiments are divided into two parts. In the first part, we measure the effect of gender bias on eight sets of image classifiers belonging to CNNs and ViTs, using Accuracy Difference and IIAS. In the second part, we analyse the zero-shot predictions of CLIP using four different image encoders belonging to CNNs and ViTs.

### 4.1 Bias Analytics using Image Classifiers

We selected four CNN models: VGG16, ResNet152, Inceptionv3, and Xception, and four ViT models: ViT B/16, B/32, L/16, and L/32. All the models were pre-trained on the Imagenet dataset. We used the feature-extracting layers of the models and added customised dense layers to all the models. Then, the models were fine-tuned and tested on our custom dataset containing about 10k images. In order to ensure controlled variables, we limited our study to simpler models such as the original ViTs and older CNNs. This allowed us to isolate the bias comparison solely to the architecture and not have any influence from complex additions.

### 4.1.1 The Dataset

We created a custom visual dataset to measure gender bias by crawling images using Google Search using the Selenium library<sup>1</sup> for occupation-related query terms ‘CEO’, ‘Engineer’, ‘Nurse’, and ‘School Teacher’. The occupation categories ‘CEO’ and ‘Engineer’ are traditionally male-dominated and ‘Nurse’ and ‘School Teacher’ are female-dominated [9, 19, 21]. Two sets of training data were created: gender-balanced and imbalanced. In the balanced dataset, all categories have a 50:50 split of images of men and women. In the imbalanced dataset, the gender ratio of the classes was split in a male:female ratio of 9:1 for ‘CEO’ and ‘Engineer’ and 1:9 for ‘Nurse’ and ‘School Teacher’, as per existing workforce bias. The queried images did show gender bias as per previous research [9, 21] and the gender ratio was adjusted in order to achieve uniformity. The test dataset was also gender balanced. The image filtering to achieve the necessary gender ratios was done manually. The train dataset consists of 7,200 images: 3,600 images for balanced and imbalanced datasets with each containing 900 images for each category. The test dataset consists of 1,200 images: 300 images for each category with 150 images for each gender. The validation sets for both the biased and unbiased training were split from the balanced and imbalanced datasets manually, keeping the gender ratios intact. A separate dataset containing images of men and women was queried using the terms ‘man’ and ‘woman’ for the IIAS assessment.

### 4.1.2 Measuring Accuracy Difference

The models were partially retrained (fine-tuned) on the balanced and imbalanced datasets, creating a total of 80 models: (4 CNNs & 4 ViTs) x 2 (biased & unbiased) x 5 iterations. The training methodology for the CNNs is as follows. First, the feature-extracting layers were frozen and the custom dense layers warmed up for 50 epochs. Then the last two convolution blocks were unfrozen and the model was trained for a further 50 epochs with a smaller learning rate and with early stopping parameters with patience set to 10 iterations. For the ViTs, first the feature extracting layers were kept frozen and the models trained for 100 epochs with early stopping parameters with patience set to 10 iterations. Then the entire model was unfrozen and trained for 50 epochs with a very small learning rate with early stopping patience set to 5. The Accuracy Difference was calculated for all the models as explained in section 3.1 and as per eq(3).

### 4.1.3 Measuring IIAS

The fine-tuned biased and unbiased models (from the previous experiment; section 4.1.2) were saved and their classification layers were removed for this part and the models were used as feature extractors on two sets of target images. The first set is the test dataset used for the previous part and for the second set, we blacked out (masked) the faces in the images as the most important feature for determining gender. Two sets of five images of men and women each were used for each part (masked and unmasked) as targets (Table 1). Ten images of men and women each were used as gender attributes (Figure 1). Then, the biased and unbiased model feature extracting layers were used to calculate IIAS as per eq (5). The experiment was repeated five times and the images for the attributes and the targets were chosen randomly without repeating. It is important to note that only the last layers of the CNN based feature extractors were retrained on our dataset, but as the training data for all

---

<sup>1</sup><https://www.selenium.dev/>





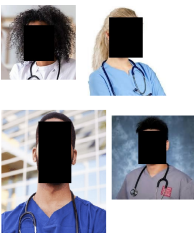
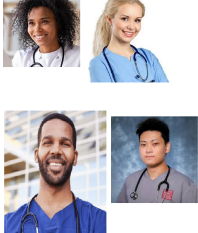

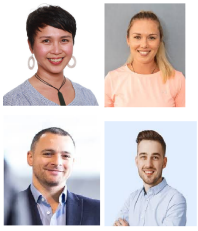
	Masked	Unmasked
CEO		
Engineer		
Nurse		
School Teacher		

Table 1: Target images

the models are the same, it gives us an estimate of how bias is handled differently by the different model families.

## 4.2 Bias Analytics using CLIP

To further understand the effect of gender bias on model architecture, four different types of CLIP image encoders were used: CNNs ResNet 50 and 50x4 and ViTs ViT B/16 and B/32. A list of 100 occupation terms was created based on official lists and CLIP’s zero-shot predictions used to predict labels for images of men and women (full list of terms is provided in Appendix A). The image dataset is the same as that used for attributes in the IAS experiment. The top predictions for men and women were then analysed to study the differences in the effect of gender bias on CNNs and ViTs.

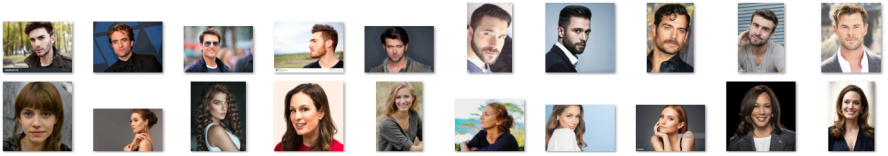


Figure 1: Gender attributes - Men (top) and Women

Model Type	Model Name	Mean $\Delta$	Average Model $\Delta$	Mean % $\Delta$	Average Model % $\Delta$
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% $\uparrow$ )	39.19	37.8 (123% $\uparrow$ )
	ViT-B32	0.18		39	
	ViT-L16	0.13		31	
	ViT-L32	0.2		42	

Table 2: Accuracy Difference ( $\Delta$ ) for CNNs and ViTs. ( $\uparrow$ ) indicates higher bias in percentage and is given in red.  $\% \Delta = \frac{|A_{unbiased} - A_{biased}|}{A_{unbiased}} * 100$

## 5 Findings and Discussions

### 5.1 Accuracy Difference

We found the Accuracy Difference for ViTs to be significantly higher than CNNs. The figures in Table 2 show  $\Delta$  to be 54% higher and the %  $\Delta$  to be 123% higher for ViTs. This means the effect of gender bias is higher on the ViTs. This may be explained by the fact that ViTs have global attention which enables them to get more visual cues allowing them to deduce gender from multiple visual features. We also see the variation in  $\Delta$  among the CNNs. ResNet 152 has the highest  $\Delta$  and %  $\Delta$ . This may be due to ResNet 152 having a larger receptive field [18] enabling it to gather more visual information related to gender. The differences among ViTs, though not as prominent as CNNs, still show some variation with models having a larger patch size (ViT-B/32 and L/32) having more bias. As larger patch sizes enable the capture of more global information [6, 4, 13], the model can learn more information related to gender, thereby contributing to bias, in a way similar to the CNNs.

Class	Masked				Unmasked			
	Biased		Unbiased		Biased		Unbiased	
	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% $\uparrow$	80% $\uparrow$			111% $\uparrow$		176% $\uparrow$

Table 3: Image-Image Association Score for CNNs and ViTs. The values are the average of all the models averaged over five iterations. A +ve value indicates a bias towards men and a -ve value indicates a bias towards women. The total IIAS is calculated by adding the absolute values of the individual IIAS scores which capture bias magnitude. This is done to provide a better comparison between the models. ( $\uparrow$ ) indicates higher IIAS i.e. higher bias in percentage and is given in red.



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% ↑)	

Table 4: Top 3 predictions for images of men and women using CLIP. The occurrence values show the percentage of predictions for the top 3 predictions. (↑) indicates a higher concentration of biased predictions i.e. higher bias in percentage and is given in red.

Encoder Type	Image Encoder	Skewness	
		Man	Woman
CNN	RN 50	2.27	3.6
	RN 50x4	2.06	3.84
ViT	ViT-B/16	2.54	3.75
	ViT-B/32	2.73	4.26
Model Average	CNN	2.16	3.7
	ViT	2.63 (21.7% ↑)	4 (8% ↑)

Table 5: Skewness in CLIP’s predictions using different image encoders. (↑) indicates a higher skewness of biased predictions i.e. higher bias in percentage and is given in red.

## 5.2 IIAS

The results of the IIAS experiment showed similar results to those in the previous experiment with ViTs showing higher bias than CNNs as shown in Table 3. The scores show stereotypical bias in occupations with ‘CEO’ and ‘Engineer’ having a positive score indicating male bias and ‘Nurse’ and ‘School Teacher’ showing female bias as indicated by a negative score. This is similar to the results shown in previous research [10]. For the masked images, we see a 23% higher IIAS for the biased ViT models but an 80% higher IIAS for the unbiased CNN models. In the case of the unmasked images, the ViTs had a higher IIAS for both the biased and unbiased models, 111% and 176% respectively. Ideally, as there is an equal number of images of men and women in the target sets, the values should be zero or very close. In the case of masked images, where the face is hidden, the models may learn gender from other features such as the dress worn [2]. ViTs with their global attention may amplify bias due to this as seen from Table 3. An interesting observation is that for masked images, the unbiased CNNs show a higher bias than the ViTs. This may be due to convolutions being a high-pass filter amplifying high-frequency signals [3] and the absence of the low-frequency signals in the face affecting its performance. Another reason may simply be that the CNNs are unable to localize their focus as faces generally have a higher saliency. We are, however, not fully sure of what might cause this.

## 5.3 Analysis of CLIP Zero-shot Predictions

The predictions using CLIP zero-shot (Table 4) reveal the presence of gender bias in the model with the top three predictions for men being stereotypically male-dominated occupa-



tions such as ‘chief executive officer’, ‘economist’, and ‘investment banker’ whereas those for women are stereotypically female-dominated such as ‘beautician’, ‘housekeeper’, and ‘jewellery maker’ [10, 21]. The predictions are highly skewed with these biased predictions making up nearly half of all the predictions. The skewness is higher when ViTs are used as image encoders showing a higher bias. The skewness metrics given in Table 5 also show higher skewness for ViT encoders. Although the higher bias in CLIP’s ViT encoder models shows a similar pattern to our classifier experiments, the effect is less pronounced. This may be due to the debiasing done in CLIP [14].

## 6 Conclusion and Future Work

In our experiments, we found evidence that the model architecture affects the amplification of social biases and show that vision transformers amplify gender bias more than convolutional neural networks. We attribute this to two features of vision transformers: 1) a shallower loss landscape leading to better generalisation and 2) global attention and a larger receptive field due to the multi-headed self-attention mechanism that enables vision transformers to capture more visual cues and long-term dependencies. Both these properties of vision transformers allow them to learn contextual information and generalise better than convolutional neural networks and learn complex concepts. But this inadvertently enables ViTs to learn social concepts such as gender. Therefore, when the training data is gender biased, the ViTs learn biased associations better than CNNs.

This paper also introduces *Accuracy Difference*, a metric for social bias in both CNNs and ViTs. It may be used for estimating and comparing bias in many different types of models with different architectures. It is simple, easy to understand and implement and can work on black box models such as closed-sourced models and APIs. We further adapted the *Image-Image Association Score* for detecting bias in image classifiers and evaluated the effect of architecture choice in image encoders of a large multimodal model, CLIP. With the prevalence of large multimodal models and their wide applications, the potential for inadvertent amplification of biases is of particular concern and requires further consideration beyond gender in a binary sense and also to include other forms of social bias (geographic, racial, etc).

### 6.1 Future Work

This research can help understand the effect of model architecture on social biases and assist developers in making informed choices about selecting vision models. One such case is CLIP, as discussed earlier. Accuracy difference can be used for bias analytics for different architectures. ViTs have been shown to outperform CNNs in many applications [6, 7, 22], leading to widespread adoption. However, if, as this research suggests, they may amplify bias to a greater extent, this aspect needs to be understood and considered as part of the adoption of ViTs.

## 7 Acknowledgements

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.

We would like to thank Dr. Alessandra Mileo for her input in this paper.

## 8 Appendix A

### List of Occupations

accountant, administrator, architect, artist, athlete, attendant, auctioneer, author, baker, beautician, blacksmith, broker, business analyst, carpenter, cashier, chef, chemist, chief executive officer, cleaner, clergy, clerk, coach, collector, conductor, construction worker, counsellor, customer service executive, dancer, dentist, designer, digital content creator, doctor, driver, economist, electrician, engineer, farmer, filmmaker, firefighter, fitter, food server, gardener, geologist, guard, hairdresser, handyman, housekeeper, inspector, instructor, investment banker, jewellery maker, journalist, judge, laborer, lawyer, librarian, lifeguard, machine operator, manager, mathematician, mechanic, midwife, musician, nurse, official, operator, painter, photographer, physician, physicist, pilot, plumber, police, porter, postmaster, product owner, professor, programmer, psychiatrist, psychologist, retail assistant, sailor, salesperson, scientist, secretary, sheriff, soldier, statistician, student, supervisor, supply chain associate, support worker, surgeon, surveyor, tailor, teacher, trainer, warehouse operative, welder, youtuber

**Sources:** Garg et al. [8], BBC Careers <sup>2</sup>, LinkedIn <sup>3 4</sup>, Australian Occupation List <sup>5</sup> and Canadian Occupation List <sup>6</sup>.

## References

- [1] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.
- [2] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.
- [3] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018.

---

<sup>2</sup><https://www.bbc.co.uk/bitesize/articles/zdqnxyc>

<sup>3</sup><https://business.linkedin.com/talent-solutions/resources/talent-acquisition/jobs-on-the-rise-nl-en-cont-fact> accessed: 19-04-2023

<sup>4</sup>[https://business.linkedin.com/content/dam/me/business/en-us/talent-solutions/emerging-jobs-report/Emerging\\_Jobs\\_Report\\_U.S.\\_FINAL.pdf](https://business.linkedin.com/content/dam/me/business/en-us/talent-solutions/emerging-jobs-report/Emerging_Jobs_Report_U.S._FINAL.pdf) accessed: 19-04-2023

<sup>5</sup><https://immi.homeaffairs.gov.au/visas/working-in-australia/skill-occupation-list> accessed: 19-04-2023

<sup>6</sup><https://www.canada.ca/en/immigration-refugees-citizenship/services/immigrate-canada/express-entry/eligibility/find-national-occupation-code.html> accessed: 19-04-2023

- 
- [4] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 244–253, 2019.
- [5] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021.
- [6] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.
- [7] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. 2021.
- [8] Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent learning: Do different neural networks learn the same representations? *arXiv preprint arXiv:1511.07543*, 2015.
- [9] Abhishek Mandal, Susan Leavy, and Suzanne Little. Dataset diversity: Measuring and mitigating geographical bias in image search and retrieval. In *Proceedings of the 1st International Workshop on Trustworthy AI for Multimedia Computing*, pages 19–25, 2021.
- [10] Abhishek Mandal, Susan Leavy, and Suzanne Little. Multimodal composite association score: Measuring gender bias in generative multimodal models. *arXiv preprint arXiv:2304.13855*, 2023.
- [11] Abhishek Mandal, Suzanne Little, and Susan Leavy. Gender bias in multimodal models: A transnational feminist approach considering geographical region and culture, 2023.
- [12] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34:23296–23308, 2021.
- [13] Namuk Park and Songkuk Kim. How do vision transformers work? In *International Conference on Learning Representations*.
- [14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.
- [15] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

- [16] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.
- [17] Ignacio Serna, Alejandro Pena, Aythami Morales, and Julian Fierrez. Insidebias: Measuring bias in deep networks and application to face gender biometrics. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3720–3727. IEEE, 2021.
- [18] Jack Sim. Computing Receptive Fields of Convolutional Neural Networks — distill.pub. <https://distill.pub/2019/computing-receptive-fields/>. [Accessed 08-May-2023].
- [19] Vivek K Singh, Mary Chayko, Raj Inamdar, and Diana Floegel. Female librarians and male computer programmers? gender bias in occupational images on digital media platforms. *Journal of the Association for Information Science and Technology*, 71(11): 1281–1294, 2020.
- [20] Kirill Sirotkin, Pablo Carballeira, and Marcos Escudero-Viñolo. A study on the distribution of social biases in self-supervised learning visual models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10442–10451, 2022.
- [21] Angelina Wang, Alexander Liu, Ryan Zhang, Anat Kleiman, Leslie Kim, Dora Zhao, Iroha Shirai, Arvind Narayanan, and Olga Russakovsky. Revise: A tool for measuring and mitigating bias in visual datasets. *International Journal of Computer Vision*, 130(7):1790–1810, 2022.
- [22] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5310–5319, 2019.
- [23] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8919–8928, 2020.
- [24] Robert Wolfe, Mahzarin R Banaji, and Aylin Caliskan. Evidence for hypodescent in visual semantic ai. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1293–1304, 2022.
- [25] Robert Wolfe, Yiwei Yang, Bill Howe, and Aylin Caliskan. Contrastive language-vision ai models pretrained on web-scraped multimodal data exhibit sexual objectification bias, 2022.
- [26] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1323. URL <https://aclanthology.org/D17-1323>.

- [27] Dominik Zietlow, Michael Lohaus, Guha Balakrishnan, Matthäus Kleindessner, Francesco Locatello, Bernhard Schölkopf, and Chris Russell. Leveling down in computer vision: Pareto inefficiencies in fair deep classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10410–10421, 2022.