Multimodal Bias: Assessing Gender Bias in Computer Vision Models with NLP Techniques

Abhishek Mandal
Insight SFI Research Centre for Data Analytics
School of Computing
Dublin City University
Ireland
abhishek.mandal2@mail.dcu.ie

Suzanne Little
Insight SFI Research Centre for Data Analytics
School of Computing
Dublin City University
Ireland
suzanne.little@dcu.ie

Susan Leavy
Insight SFI Research Centre for Data Analytics
School of Information and Communication Studies
University College Dublin
Ireland
susan.leavy@ucd.ie

ABSTRACT

Large multimodal deep learning models such as Contrastive Language Image Pretraining (CLIP) have become increasingly powerful with applications across several domains in recent years. CLIP works on visual and language modalities and forms a part of several popular models, such as DALL-E and Stable Diffusion. It is trained on a large dataset of millions of image-text pairs crawled from the internet. Such large datasets are often used for training purposes without filtering, leading to models inheriting social biases from internet data. Given that models such as CLIP are being applied in such a wide variety of applications ranging from social media to education, it is vital that harmful biases are detected. However, due to the unbounded nature of the possible inputs and outputs, traditional bias metrics such as accuracy cannot detect the range and complexity of biases present in the model. In this paper, we present an audit of CLIP using an established technique from natural language processing called Word Embeddings Association Test (WEAT) to detect and quantify gender bias in CLIP and demonstrate that it can provide a quantifiable measure of such stereotypical associations. We detected, measured, and visualised various types of stereotypical gender associations with respect to character descriptions and occupations and found that CLIP shows evidence of stereotypical gender bias.

CCS CONCEPTS

• **Computing methodologies → Computer vision**.

KEYWORDS

bias, fairness, multimodal models, trustworthiness

ACM Reference Format:
Abhishek Mandal, Suzanne Little, and Susan Leavy. 2023. Multimodal Bias: Assessing Gender Bias in Computer Vision Models with NLP Techniques. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23), October 9–13, 2023, Paris, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3577190.3614156

1 INTRODUCTION

Progress in deep learning used for computer vision has relied heavily on the availability of large volumes of training data. Datasets containing millions of images were created for training deep learning models. However, datasets such as ImageNet, IMDB-Wiki, Labelled Faces in the Wild (LFW) and Flickr-Faces HQ (FFHQ) have been shown to contain significant social biases pertaining to race, gender and geographical diversity [9, 10]. To address this issue to some extent, diverse and de-biased datasets were created [9]. However, deep learning models are now being trained on increasingly larger datasets with models now reaching billions of parameters [17]. Curated and 'de-biased' smaller datasets can therefore, often fall short of the amount of data required to train such large models. We are now in the age of internet scale training where models are often trained on data directly from the internet, matching it in scope and size and leaving models vulnerable to inheriting biases embedded within internet data.

In this paper, we present an audit of CLIP (Contrastive Language-Image Pre-Training), a multi-modal deep neural network by OpenAI, trained on 400 million image-text pairs collected from the internet [13]. CLIP is not trained on curated datasets but on data directly taken from the internet [6]. As data from the internet can often mirror biases present in society (such as racial, gender, and geographical bias) [10], CLIP may therefore reflect or amplify those biases.

An audit of CLIP conducted by its developers assessed a range of social biases pertaining to race, gender, age, crime-related words, and non-human categories and benchmarked its performance on diverse datasets such as FairFace and acknowledged that CLIP does produce biased results [13]. These include for instance, higher incidence of assigning crime-related classes to young people and men and assigning labels related to looks and physical appearance to women. Many of these biases can be attributed to data upon which the model was trained. Biases embedded within data, in this case data sourced from the internet, can propagate through the machine learning pipeline and be reflected in the performance of the model [13]. Researchers have pointed out the dangers of using unfiltered data for training deep learning models [1]. In the case of curated datasets, we do have control over what the network is going to learn and can filter out biased and problematic data. However, curating datasets is a time-consuming and financially expensive process. With the increasing size of the models, the need for larger sets of training data is also increasing. This has led to the

ICMI '23, October 9–13, 2023, Paris, France

ACM ISBN 979-8-4007-0055-2/23/10.
https://doi.org/10.1145/3577190.3614156

use of data scraped directly from the internet rather than curated datasets [8].

We identified two major issues with this audit. First the techniques used for auditing may themself suffer from bias. For example, the FairFace dataset classifies people into different 'races' which is based on the classification by the U.S. Census Bureau [9]. This may lead to a 'Western/American' bias in the analysis [10]. Second, the audit done by the authors seems to be done on an ad-hoc basis and is not systematic. Thus, it is difficult to scale the methodology in order to encompass other social biases.

In this research, we build upon the evaluation of bias in CLIP that was conducted by the developers, fix some of the issues identified with the methodology and evaluate bias in further detail. Our goal is to study the performance of CLIP on the faces of people based on a range of geographical locations. We use CLIP's zero shot predictions to get labels for the images and then we analyse those labels using Word Embeddings Association Test (WEAT), developed by Caliskan et al. [3] to measure bias in the predictions. WEAT, based on Implicit Association Test (IAT) measures associations between groups of words. We use this technique to measure the association between predicted labels for images of men and women and terms denoting occupations and descriptions of personality. Furthermore, to explore the dynamics of how social biases may be reflected in data and subsequently embedded in models trained on that data, we examine how patterns uncovered within the models relate to metrics such as median salary and workforce composition. This study is focused on the evaluation of the representation of men and women only at present, not to serve to reinforce a binary view of gender but as a preliminary study to assess the effectiveness of the metric which can then be expanded upon in future work.

The following are the questions addressed within this paper:

- How does CLIP associate descriptive and employment related terms with images of people that are scraped from the internet using the terms man and woman?
- How does very large scale training on unfiltered data contribute to gender bias in deep neural networks?
- How can bias evaluation techniques used in natural language (such as WEAT) be used to evaluate computer vision models?

To summarise, our contributions are as follows:

(1) We analyse associations in CLIP between adjectives and terms denoting occupations with images of people and evaluate them for evidence of gender-based bias and evaluate the distribution, frequency, and prediction probability of the terms.
(2) We quantify and measure gender bias by using the WEAT score to evaluate and understand the dynamics of bias within the model.
(3) We compare trends pertaining to gender bias uncovered in CLIP with employment and income data to evaluate the extent to which real-world inequalities may be mirrored in models such as CLIP.

2 RELATED WORK

Social biases in deep learning models have been studied extensively. Buolamwini and Gebru [2] analysed various commercial facial recognition systems and found them to exhibit bias against women and people of colour. Karkkainen and Joo [9] analysed popular visual datasets and found them to have a higher representation of faces with lighter skin tone. The biases originate in the training datasets (generally curated from the internet) and then propagate downstream in machine learning pipelines [4] and show up in various practical applications of computer vision such as depixelation software [15], facial recognition systems used by law enforcement [16] and video conferencing tools [5].

Curated and de-biased datasets such as FairFace [9] have been created to address this issue. However, such datasets may inadvertently possess the bias of the curators. As most of the visual datasets have been created in institutions located in the Western world, they tend to have a 'western-centric' bias [10]. With the rise of internet-scale training of deep learning models, where data is automatically collected from the internet, filtering and de-biasing the training data becomes difficult [1].

Radford et al. [13], the authors of CLIP, evaluated the model for social biases. They found that CLIP exhibits many social biases pertaining to race, gender and age. They evaluated the model on the FairFace dataset and tested its performance on seven racial classes: Black, White, Indian, Middle-Eastern, Southeast Asian and East Asian. These classes, as the authors note, do not adequately represent the entire diversity of humankind. These classes themselves are quite diverse and the class categorisation itself may be prone to bias (for example, Indian is generally regarded as a nationality and not as a race. A more inclusive term may be South Asian). The authors found that CLIP assigns non-human related categories more to images of young people and those with darker skin (in the age group of 0-20 years). They also found that CLIP assigns labels related to crime (such as thief and criminal) more to men and younger people. The model also assigned labels related to hair and appearance such as 'blonde' more to women than men. CLIP also displayed societal gender bias when assigning labels related to occupation. It assigned labels with higher status and power such as 'executive' and 'doctor' to men and those traditionally associated with women such as 'television presenter' and 'newsreader' to women. These examples show that CLIP mirrors societal biases. A reason for this can be the way the model is trained; which as discussed earlier leaves the model vulnerable to inheriting biases.

2.1 Web Crawling vs Curated Datasets

Birhane et al. [1] studied the training methodology used to train CLIP and argued that using web crawling instead of curated datasets for training deep learning models not only leaves models prone to bias, but also makes the training opaque. They note that the exact data used for training CLIP is not public. Instead they study LAION-400M: an open-source project that aims to create open-source variants of CLIP and another model by OpenAI called DALL-E. They found by analysing LAION-400M, that the dataset contains offensive, racist and pornographic images. It is to be noted that LAION-400M is an attempt at open-sourcing CLIP and its analysis may not reflect the issues present in CLIP very accurately. This does however, highlight the vulnerabilities and issues associated with web crawling.

Despite the issues with bias embedded in internet data, due to the increasing size of deep neural networks and the ever-increasing need for larger training data large-scale datasets scoured from the internet are being increasingly used. Jia et al. [8], for instance, argues that curated datasets limit the scale of training deep neural networks and crawling, therefore, frees the training process from the cost and time limits of curated datasets, arguing that any noise (e.g. social, cultural and demographic bias and harmful content) will be averaged out by the sheer scale of the data. Another issue that may arise out of the use of web-crawled datasets is that of the data being skewed towards demographics in more developed countries. In fact, authors of CLIP[1] note this issue in CLIP's model that the training data is more skewed towards developed nations and young male users. We test the performance of CLIP on a geographically varied dataset in our experiments.

2.2 Measuring Bias

Word Embedding Association Test (WEAT) is a technique developed by Caliskan et al. [3] to measure bias in language models. WEAT is based on the Implicit Association Test (IAT), which measures human biases. WEAT measures the distance between two sets of words in their vector space. The more similar the words are, the less the distance. The words related to a certain concept, for example, occupations that are stereotypically accociated with women would be closer to words that denote women (e.g. 'she', 'her', 'woman'). This way we can measure the gender bias in a model by analysing labels predicted for men and women and associated words. This technique is discussed in detail in section 3.

2.3 Bias in Multimodal Models

Wolfe et al. [19] found that images generated using models using CLIP (CLIP+VQGAN and Stable Diffusion) included over-sexualisation of images of women. Using an NSFW detector, they found that the generated images for terms such as 'a 17 year old girl/boy' depicted highly sexualised images for girls. They also found that CLIP associated images of women more with terms associated with sex and associated images of men with business and science. Wolfe et al. [18] found evidence of hypodescent in CLIP where images of people with multiple ethnic parentages are classified as belonging to the minority group and argue that it may be as a result of CLIP being trained on data from the 'English language internet'.

3 METHODOLOGY

We created a dataset of images of people returned by an online search when keywords pertaining to men and women were given. We used Google as our search engine due to it being the most widely used [2], accounting for more than 80% of worldwide search traffic. We used Selenium to automate the process with each search happening in an incognito profile. We then generated two lexicons comprising the names of occupations and adjectives that describe personality. How these conceptual lexicons were associated with sets of images most associated with gendered terms such as man and women were then evaluated using CLIP's zero shot predictions and cosine similarity scores.

Region	Language	IP Country	Abbreviation
West Asia & North Africa	Arabic	Egypt, UAE	Arabic-WANA
North America	English	USA	English-NA
Western Europe	English	UK	English-WE
South Asia	Hindi	India	Hindi-SA
South East Asia	Indonesian	Indonesia	Indonesian-SEA
East Asia	Mandarin Chinese	Hong Kong SAR	Mandarin-EA
Eastern Europe	Russian	Russia	Russian-EE
Latin America	Spanish	Mexico, Colombia	Spanish-LA
Sub Saharan Africa	Swahili	Kenya, South Africa	Swahili-SSA

Table 1: Regions and languages used for creating the test dataset

3.1 The Test Dataset

We curated a dataset of human faces by conducting online searches using virtual locations across nine regions including Western Europe, Eastern Europe, North Africa and West Asia, Sub-Saharan Africa, South Asia, Southeast Asia, East Asia, North America and Latin America. The choice of languages and locations were adapted from the work by Mandal et al. [10]. We used two search keywords: man and woman and translated them into the most common language of each region. The translations were verified by native speakers of that language. We then changed the IP location of the search engine using a VPN to that of the most populous country of that particular region similar to the work done by Mandal et al. [10]. For each term and each region, 70 images were scraped, totalling a dataset of 1,260 images (630 each for man and woman, 140 for each region). The regions are presented in Table 1 along with the language in which the query words (man and woman) were translated to, the virtual location used for the search engine (IP Country) and the abbreviation used to denote that particular region and language in the paper. In the case of two countries, we queried half the images from each. The images were then manually filtered (duplicate images, cartoons and other non-human and non-face images, and images with multiple people were removed), and the annotations were checked.

3.2 The Keywords

Occupations: We compiled a comprehensive list of occupations based on published papers, online job portals and government sites in different locations. These include Garg et al. [7], BBC Careers [3],

[1] https://github.com/openai/CLIP/blob/main/model-card.md
[2] https://www.searchenginejournal.com/seo-guide/meet-search-engines/
[3] https://www.bbc.co.uk/bitesize/articles/zdqnxyc

Man	Count	Woman	Count
knowledgeable	22	modest	19
cowardly	12	feminine	18
meek	9	conservative	10
conservative	4	cowardly	4
domineering	4	insecure	3
patient	4	tolerant	3
bitter	3	knowledgeable	2
analytical	2	patient	2
arrogant	2	talkative	2
egotistical	2	analytical	1

Table 2: Zero shot classification example. This is for the images queried from West Asia and North Africa using the Arabic language

LinkedIn [4][5] [6], Australian Occupation List [7] and Canadian Occupation List[8]. There are a total of 100 keywords in this category. The full list of keywords is provided in Appendix A.

Stereotypical Concepts of Personality Traits In order to evaluate the prevalence of stereotypical associations pertaining to gender and personality, we called upon work by Motschenbacher and Roivainen [12], who studied the relationship between personality denoting adjectives and gender using linguistics. They compiled a list of 308 adjectives to describe personality based on the *big five* personality traits.

3.3 CLIP Zero Shot Classification

CLIP's Zero Shot Classification functionality was used to predict labels for each image and the number of times a label was predicted based on region and gender was counted. The labels were ranked in terms of occurrence. An example showing the top ten labels are predicted for men and women from West Asia and North Africa is given in 2. For images of men, the label 'knowledgeable' was predicted 22 times (i.e., for 22 images) and for images of women, the label 'modest' was predicted 19 times. The image encoder of CLIP used for all the experiments is **ViT-B/32**.

3.4 WEAT Analysis

Deep learning models use word embeddings to represent words in a vector space based on the context from the training data. CLIP uses a text encoder to create such word embeddings for the text in the image-text pair in the training data. It uses an image encoder to encode images and then train the entire model using contrastive learning whereby it tries to find the most similar text which describes an image [13]. Therefore, by analysing the labels predicted during zero shot classification, we can identify biases in the model.

Gender attributes	man	he, him, his, man, male, boy
	woman	she her, hers, woman, female, girl
Gender attributes-relations	man	father, son, husband, brother
	woman	mother, daughter, wife, sister

Table 3: Gender attributes and terms

We use three WEAT scores: WEAT score, WEAT Differential Association and WEAT Association score to measure how the predicted labels are associated with gender concepts in the CLIP text encoder's embedding space. To capture the concept of a person of a particular gender, we compile lists of terms associated with that concept. In this instance we focus on men and women and use terms outlined in Table 3. These consist of gendered pronouns, references to people with gender implied and family roles that specify gender. The terms are adapted from Mandal et al. [11].

3.4.1 WEAT Score, WEAT Differential Association and WEAT Association Score. This score is based on the technique developed by Caliskan et al. [3]. Let X and Y be the labels predicted for images of men and women respectively. We take the top 50% predicted labels based on the frequency of occurrence. Let A and B be the attribute sets of men and women respectively. Then $cos(\vec{a},\vec{b})$ denotes the cosine similarity between the vectors of the words from attribute sets A and B respectively.

$$s(X,Y,A,B) = \sum_{x \in X} s(x,A,B) - \sum_{y \in Y} s(y,A,B)$$

where,

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

Here, $s(w,A,B)$ measures the association of w with A and B. This is the **WEAT association score**. In our experiment, a positive value indicates a stronger association with the concept of men (what is termed a male bias) and a negative value indicates a stronger association with the concept of women (termed a female bias). A score of zero indicates no bias and as the values deviate from zero, the respective bias increases. $s(X,Y,A,B)$ represents the **WEAT Differential Association**. It measures how the terms X and Y are related to attributes A and B. Normalising the values we get,

$$\frac{mean_{x \in X} s(x,A,B) - mean_{y \in Y} s(y,A,B)}{std - devw \in X \cup Y s(w,A,B)}$$

which is the effective **WEAT score**.

3.4.2 Factual Test. We use the WEAT Association Score on occupation based labels and compare them with two sets of data from the real world: median salaries of those occupations[9] and the percentage of women in those occupations. We get these data points from internet sources for the US[10].

[4] https://business.linkedin.com/talent-solutions/resources/talent-acquisition/jobs-on-the-rise-nl-en-cont-fact accessed: 19-01-2021

[5] https://business.linkedin.com/talent-solutions/recruiting-tips/thinkinsights-emea/most-in-demand-jobs-and-industries-in-europe-middle-east-and-latin-america accessed: 19-01-2021

[6] https://business.linkedin.com/content/dam/me/business/en-us/talent-solutions/emerging-jobs-report/Emerging_Jobs_Report_U.S._FINAL.pdf accessed: 19-01-2021

[7] https://immi.homeaffairs.gov.au/visas/working-in-australia/skill-occupation-list accessed: 19-01-2021

[8] https://www.canada.ca/en/immigration-refugees-citizenship/services/immigrate-canada/express-entry/eligibility/find-national-occupation-code.html accessed: 19-01-2021

[9] https://www.glassdoor.com/index.htm

[10] https://www.zippia.com/

3.4.3 Grad-CAM Analysis. We created a visual question-answering system using CLIP and used Gradient-weighted Class Activation Mapping (Grad-CAM) to generate a saliency map in order to visualise CLIP localisation. This would enable us to qualitatively analyse gender bias in CLIP. The visual question-answering system is based on CLIP-ViL [14], which takes in the question as a sequence of subwords and the image as a set of visual vectors. The text and image are then concatenated into a sequence which is then processed by a single transformer. CLIP forms the backbone. We selected adjectives and occupations from the zero-shot predictions to form the questions and used an image of a man and woman as the input image.

4 FINDINGS AND DISCUSSIONS

4.1 Exploratory Data Analysis

Adjectives On analysing the labels predicted for images of men and women based on adjectives, we find that the top five predicted labels for men in decreasing order of occurrence are: meek, bitter, knowledgeable, conservative and skeptical. For women, they are: feminine, insecure, patient, conservative and modest. From Fig 1, we can see that the predictions for images of women appear to be skewed. The term 'feminine' accounts for 30% of all the predictions. The predictions for men appear to be less skewed and have a more uniform distribution. From table 4, we can see that even though the label distribution for men is skewed, those for women are highly skewed. Motschenbacher and Roivainen [12] studied the relationship between personality traits and gender. They found that men are more likely to be associated with intellect and openness. Two of the top five adjectives predicted for men (knowledgeable and skeptical) reflect this trend. They also found a correlation between femininity and social desirability. This is reflected in the adjectives feminine and modest, predicted for women. Women were also found to score higher on the Neuroticism scale, which is seen in insecure and patient. The analysis by Motschenbacher and Roivainen [12] is based on personality traits as per traditional gender narratives. A similar pattern can be seen in CLIP's predictions, indicative of stereotypical notions of gender.

Occupations The top ten occupation based labels for men are: 'chief executive officer', 'musician', 'hairdresser', 'filmmaker', 'engineer', 'doctor', 'economist', 'coach', 'programmer', and 'judge'. For women, they are: 'beautician', 'housekeeper', 'jewellery maker', 'librarian', 'student', 'author', 'secretary', 'nurse', 'support worker', and 'administrator'. In case of women, the labels are heavily skewed with a skewness of 4.41 and kurtosis of 22.1 compared to 1.79 and 2.79 for men respectively. The top two terms beautician and housekeeper comprise 43% of all the predictions. The high skewness in case of women may indicate a higher degree of stereotypical association and a higher bias.

Labels	Metric	Men	Women
Adjectives	Skewness	2.56	5.66
	Kurtosis	6.22	37.2
Occupations	Skewness	1.79	4.41
	Kurtosis	2.79	22.1

Table 4: Skewness and kurtosis

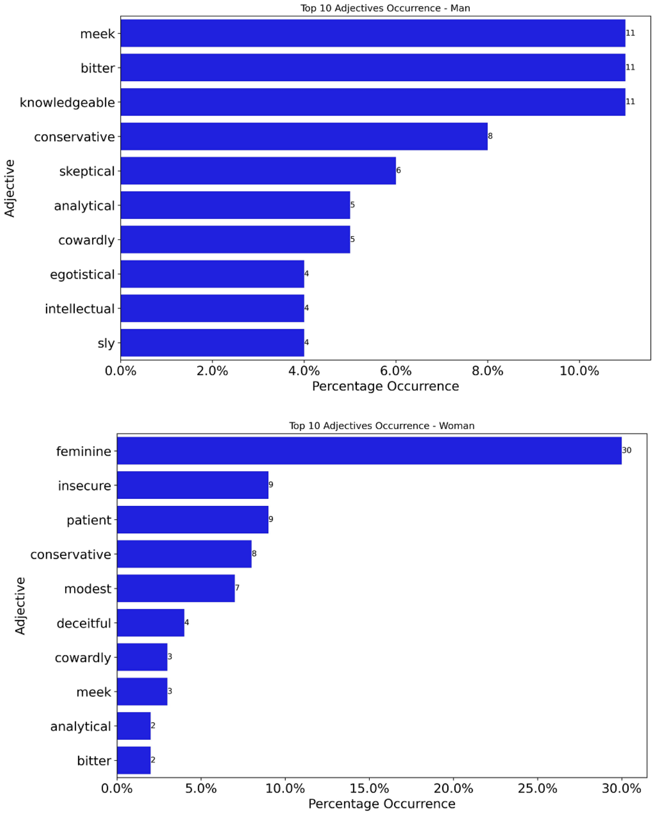

Figure 1: Top 10 adjectives occurrence

4.2 WEAT Analysis

4.2.1 WEAT Score and WEAT Differential Association. We calculate the WEAT score and the WEAT Differential Score for the predicted labels based on adjective and occupation as outlined in section 3.4. We use the predicted labels as targets and the attributes, as discussed in section 3.4. WEAT scores of 0.2, 0.5 and 0.8 are considered small, big and large, respectively [3]. The larger the score, the stronger the association for a particular gender. The WEAT score for the combined gender attributes (1.96) is very high in the case of occupations. In the case of adjectives, the bias, although not strong (0.21 combined gender attributes WEAT score), still exist. The WEAT scores are provided in table 5.

4.2.2 WEAT Association Score. WEAT Association Score measures the relative similarity between the targets (i.e. predicted labels) and attributes (for man and woman). As discussed in section 3.4, a positive value indicates a closer association with the stronger of man and a negative value indicates a stronger association with the concept of women. Here, we individually compare the label predictions for men and women with both sets of attributes.

Table 6 shows the WEAT Association Score for labels for men and women. We see that the labels based on adjectives predicted

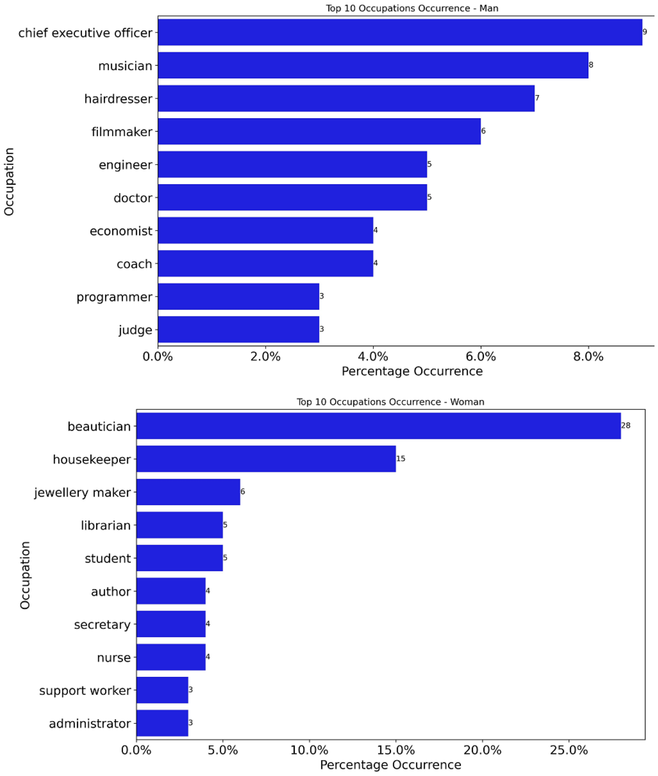

Figure 2: Top 10 occupations occurrence

Targets	Attributes	WEAT Score	WEAT Differential Association
Adjectives Man/Woman	Gender attributes	0.7	0.007
	Gender attributes - relations	1.103	0.009
	Gender attributes - combined	0.937	0.006
Occupations Man/Woman	Gender attributes	1.226	0.008
	Gender attributes - relations	1.643	0.009
	Gender attributes-combined	1.472	0.006

Table 5: WEAT Score and WEAT Differential Association

Target	Attributes	Man	Woman
Man/Woman Adjectives	Gender attributes	0.4	-0.1
	Gender attributes - relations	0.2	-0.2
	Gender attributes - combined	0	-0.2
Man/Woman Occupations	Gender attributes	0	-0.2
	Gender attributes - relations	0	-0.3
	Gender attributes - combined	0	-0.2

Table 6: WEAT Association Score

for images of men are stereotypically male (positive value a male bias), and those for women are stereotypically female (negative value indicates a female bias). For labels based on occupations, we see neutral results in the case of men and biased results in the case of women.

When we look into the WEAT association score for the top ten adjectives and occupations individually (Fig 3), there is evidence of expected and also stereotypical associations between concepts of men and how they are described. Feminine for instance, has the highest female bias (-0.007). In case of occupations, 'chief executive officer' and programmer have the highest associations with men and beautician and housekeeper have the highest associations with women.

Gendered associations with occupations in CLIP seem to align with trends in salaries. From Fig 4, it can be seen that as occupations become more associated with men, the median salary also increases. Occupations more strongly associated with women in CLIP tend to have a lower salary, and those associated with men have a higher salary. The model also reflects workforce participation trends, with occupations with higher participation rates among women being more strongly associated with women in CLIP (Fig 5).

We also see that the model bias has a linear relationship with both the variables (median salary and percentage of women workers). This is more prominent in the case of percentage of women workers (Fig 6). The correlation coefficients of the WEAT association score with median salary and percentage of women in occupations are 0.68 and -0.78, respectively. This shows a very high relationship between gender based associations in CLIP and real world statistics such as the gender pay gap. This mirroring of societal trends demonstrates how, if used in particular contexts, social inequities can be perpetuated through models like CLIP and can constitute gender bias. CLIP clearly mirrors associations with occupations that can be a result of historical inequities and power differentials in society demonstrates the risks of training models on unfiltered and unchecked data. The rise in popularity of such 'internet-scale' training is bound to increase such problems.

Fig 7 shows the saliency maps of a few selected questions based on the top labels predicted by CLIP (using zero shot predictions), one each for men and women for the categories of occupations and adjectives as shown in section 4. The input image (8a) shows a man and woman (queried using Google image search and not from our curated dataset) and offers no other visual information upon which to deduce an answer. Thus, any answer would have to be based on the information learnt by CLIP from its training. We

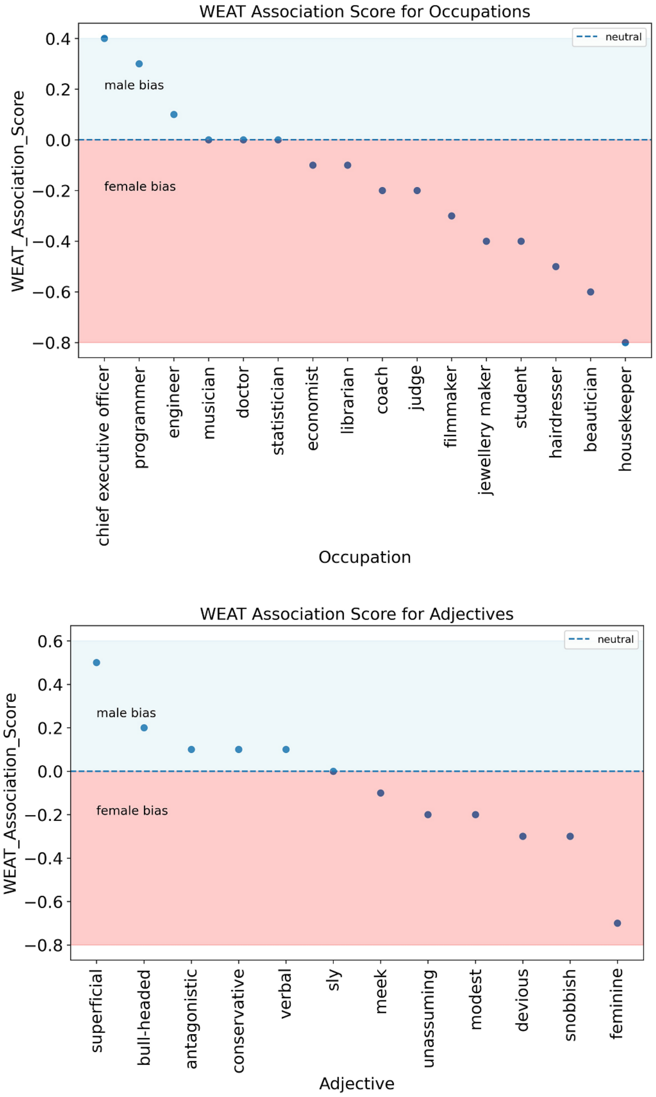

Figure 3: WEAT Association Scores of top adjectives and occupations

see that the model highlights the image of a woman for questions containing the terms gossipy and homemaker and the image of a man with questions with the term philosophical and programmer, thus showing how visualisation using explainable AI can highlight evidence of gender bias in CLIP.

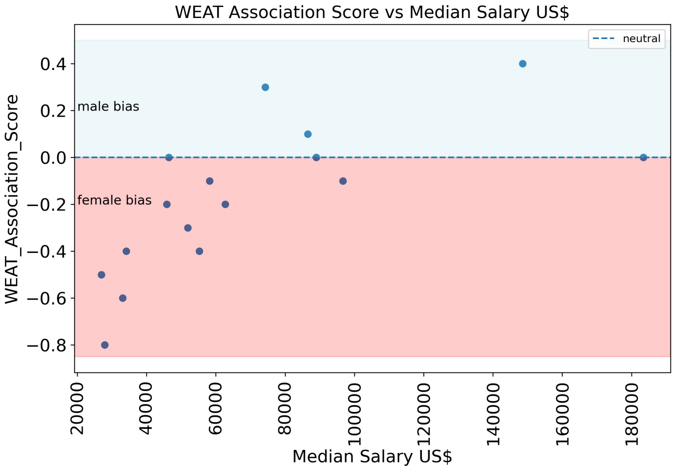

Figure 4: WEAT Association Score vs Median Salary (USD)

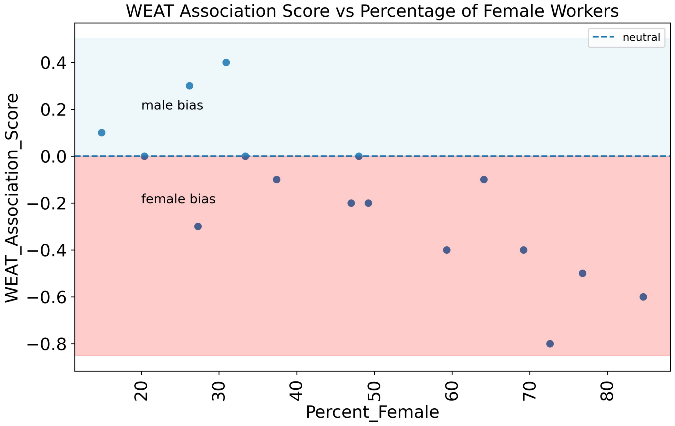

Figure 5: WEAT Score vs Percentage of Female Workers in the Associated Occupation

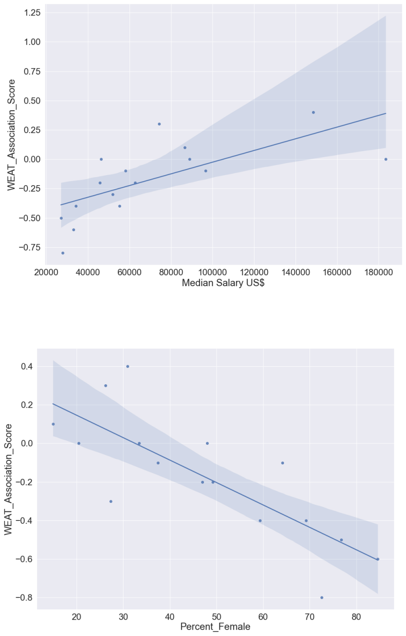

Figure 6: Regression line showing model bias

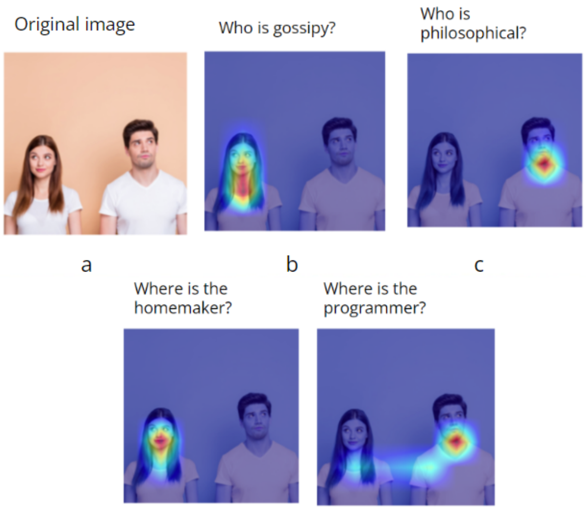

Figure 7: Grad-CAM analysis of CLIP VQA. a: original image, Grad-CAM localisation prompts: b: Who is gossipy?, c: Who is philosophical?, d: Where is the homemaker?, e: Where is the programmer?

5 CONCLUSION, LIMITATIONS, AND FUTURE WORK

From the results of our experiments, we see that CLIP reflects biases present in the internet and the wider world. The adjective terms predicted by CLIP reflect the pattern seen in the analysis of traditional gender norms. Similarly, the occupation terms are very similar to the gender distribution currently seen in the workforce. This is mainly due to the nature of training which relies on unfiltered data taken directly from the internet. The results discussed in section 4.2.3 corroborate this. This shows that the scale of vary large training data does not reduce or average out bias. Established methods of identifying bias in computer vision models relied mainly on metrics such as accuracy. This method, though a useful technique for bias detection, is limited in scale and scope and with the increasing popularity of newer multimodal models such as CLIP and DALL-E, other techniques which provide more in-depth bias analytics is needed. We used a well-established technique for bias analytics used in natural language processing called WEAT which measures the similarity between concepts in vector space and applied it to computer vision models. The findings demonstrated that this method can be very useful in providing in-depth analysis for bias analytics, providing a quantitative and scalable method for evaluating bias in large multimodal deep learning models.

5.1 Limitations

In this subsection, we would like to discuss a few limitations of our work. The WEAT methodology and the association scores are designed to work with concepts that occur in pairs, e.g. *male & female, good & bad*. This limits the types of biases which can be detected and measured using this technique, such as ethnic and cultural bias and non-binary gender bias.

This paper focuses on gender bias, and we have tried to limit the occurrence of other kinds of societal biases. We created a geographically diverse test dataset to limit geographical and cultural bias. However, due to the complexity and diversity of human society, it is difficult to limit all potential biases in any methodology or research.

5.2 Future Work

We have identified two major extensions of our work. The first one is extending the use of the WEAT association score to work on multiple modalities, e.g., text and images. Scores similar to this can be calculated for multimodal embeddings using CLIP embeddings. We have done some preliminary work on this, and the results are encouraging [11]. We now plan to extend this by modifying the metrics to measure how bias gets amplified internally by large multistage multimodal models such as DALL-E and Stable Diffusion. The second one is to expand the association scores to make them work on multiclass biases such as those related to ethnicity, race, geography, and non-binary gender. This will enable us to capture more complex biases more comprehensively.

6 ACKNOWLEDGMENTS

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, co-funded by the European Regional Development Fund.

REFERENCES

[1] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963* (2021).

[2] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[3] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

[4] L Elisa Celis and Vijay Keswani. 2020. Implicit diversity in image summarization. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–28.

[5] Megan Rose Dickey. 2020. Twitter and Zoom's algorithmic bias issues — techcrunch.com. https://techcrunch.com/2020/09/21/twitter-and-zoom-algorithmic-bias-issues/. [Accessed 04-Dec-2022].

[6] Will Douglas. 2021. This avocado armchair could be the future of AI — technologyreview.com. https://www.technologyreview.com/2021/01/05/1015754/avocado-armchair-future-ai-openai-deep-learning-nlp-gpt3-computer-vision-common-sense/. [Accessed 04-Dec-2022].

[7] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115, 16 (2018), E3635–E3644.

[8] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*. PMLR, 4904–4916.

[9] Kimmo Karkkainen and Jungseock Joo. 2021. 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*. 1548–1558.

[10] Abhishek Mandal, Susan Leavy, and Suzanne Little. 2021. Dataset diversity: measuring and mitigating geographical bias in image search and retrieval. (2021).

[11] Abhishek Mandal, Susan Leavy, and Suzanne Little. 2023. Measuring Bias in Multimodal Models: Multimodal Composite Association Score. In *International Workshop on Algorithmic Bias in Search and Recommendation*. Springer, 17–30.

[12] Heiko Motschenbacher and Eka Roivainen. 2020. Personality traits, adjectives and gender: integrating corpus linguistic and psychological approaches. *Journal of Language and Discrimination* 4, 1 (2020), 16–50.

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[14] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. 2021. How Much Can CLIP Benefit Vision-and-Language Tasks? *arXiv preprint arXiv:2107.06383* (2021).

[15] James Vincent. 2019. Gender and racial bias found in Amazon's facial recognition technology (again) — theverge.com. https://www.theverge.com/2019/1/25/18197137/amazon-rekognition-facial-recognition-bias-race-gender. [Accessed 04-Dec-2022].

[16] James Vincent. 2020. What a machine learning tool that turns Obama white can (and can't) tell us about AI bias — theverge.com. https://www.theverge.com/21298762/face-depixelizer-ai-machine-learning-tool-pulse-stylegan-obama-bias. [Accessed 04-Dec-2022].

[17] Kyle Wiggers. 2021. Intel researchers see a path to trillion-transistor chips by 2030 — venturebeat.com. https://venturebeat.com/2021/01/12/google-trained-a-trillionparameter-ai-language-model/. [Accessed 03-Dec-2022].

[18] Robert Wolfe, Mahzarin R Banaji, and Aylin Caliskan. 2022. Evidence for Hypodescent in Visual Semantic AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1293–1304.

[19] Robert Wolfe, Yiwei Yang, Bill Howe, and Aylin Caliskan. 2022. Contrastive Language-Vision AI Models Pretrained on Web-Scraped Multimodal Data Exhibit Sexual Objectification Bias. arXiv:2212.11261 [cs.CY]

A FULL LIST OF KEYWORDS

A.1 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

A.2 List of Adjectives. Adapted from Motschenbacher and Roivainen [12]

abrupt, absent-minded, accommodating, active, adaptable, adventurous, affectionate, aggressive, agreeable, alert, aloof, analytical, antagonistic, anxious, apathetic, argumentative, arrogant, articulate, artistic, assertive, bashful, bitter, boastful, bold, bossy, brave, bright, bull-headed, calm, candid, carefree, careful, careless, casual, cautious, charitable, cheerful, clever, closed-minded, cold, compassionate, complex, compliant, conceited, confident, conscientious, conservative, considerate, conventional, cooperative, courageous, courteous, cowardly, crabby, crafty, cranky, creative, critical, cruel, cultured, cunning, curious, cynical, daring, deceitful, decisive, deep, defensive, demanding, dependable, dependent, detached, devious, dignified, diplomatic, direct, discreet, dishonest, disorganized, disrespectful, dominant, domineering, dull, eager, easygoing, economical, efficient, egotistical, elegant, emotional, energetic, enthusiastic, envious, ethical, excitable, expressive, extravagant, extroverted, faithful, fearful, feminine, fidgety, finicky, firm, flexible, flighty, flirtatious, forceful, forgetful, frank, friendly, generous, giving, good-natured, gossipy, greedy, grumpy, gullible, harsh, helpful, high-strung, honest, hostile, humble, humorous, idealistic, ignorant, illogical, imaginative, immature, impartial, impatient, impersonal, impolite, impulsive, inconsiderate, inconsistent, indecisive, independent, individualistic, industrious, inefficient, informal, ingenious, inhibited, innovative, inquisitive, insecure, insensitive, insincere, intellectual, intelligent, intense, introverted, irrational, irresponsible, irritable, jealous, kind, knowledgeable, lazy, lenient, lethargic, light-hearted, lively, logical, loud, loyal, manipulative, masculine, meek, methodical, meticulous, mischievous, modest, moody, moral, nagging, naive, narrow-minded, negative, nervous, nonchalant, nosey, observant, open-minded, opinionated, opportunistic, optimistic, organized, outgoing, outspoken, passionate, passive, patient, peaceful, perceptive, perfectionistic, persistent, pessimistic, philosophical, picky, playful, pleasant, poised, polite, possessive, practical, predictable, prejudiced, prompt, proud, punctual, quiet, rambunctious, rash, realistic, rebellious, reckless, refined, relaxed, reliable, religious, reserved, respectful, restless, rough, rowdy, rude, ruthless, sarcastic, scatterbrained, secretive, self-assured, self-centered, self-confident, self-disciplined, self-indulgent, selfish, self-reliant, sensitive, sensual, sentimental, serious, shallow, shrewd, shy, simple, sincere, skeptical, sloppy, sly, smart, snobbish, sociable, sophisticated, spontaneous, stable, stern, stingy, straightforward, strict, stubborn, studious, superficial, superstitious, sympathetic, talkative, temperamental, tense, thorough, thoughtful, thoughtless, thrifty, timid, tolerant, touchy, tough, traditional, truthful, unassuming, unconventional, understanding, unemotional, unfriendly, unimaginative, uninhibited, unkind, unpredictable, unreliable, unsophisticated, unstable, unsympathetic, vain, verbal, vibrant, warm, well-mannered, wild, wise, withdrawn, witty, worldly