Auditing Imbalance and Bias in Deep Neural Networks for Multimedia Content Analytics

Abhishek Mandal, B.Tech, M.S

Supervised by Prof Suzanne Little

Dr Susan Leavy*

** School of Information and Communication Studies*

University College Dublin

A thesis presented for the degree of Doctor of Philosophy

School of Computing

Dublin City University

Date: November 2024

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, and that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: Abhishek Mandal ID No.:20214767 Date: 15 Nov 2024

Dedication

Dedicated to machine learning researchers worldwide

Acknowledgements

First of all, I would like to thank my supervisors **Prof. Suzanne Little** and **Dr. Susan Leavy** for their guidance, mentorship and support for this PhD. Thank you for giving me a chance to explore this fascinating subject and navigating me through the ups and downs. This PhD would not be possible without you.

I would like to thank my research collaborators Marion Bartl and Teerath Kumar for their collaboration, help and feedback.

I would also like to thank Caitlin Kraft-Buchman and the Alliance for Inclusive Algorithms for giving me a chance to join their wonderful team, granting me the research fellowship and contributing to my research funding.

Many thanks to Jonathan Byrne at Intel Ireland for mentoring me during my internship. I learnt a lot about the real-world applications of Artificial Intelligence.

I would also like to thank the Science Foundation Ireland for funding my PhD. This research was supported by Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_2, co-funded by the European Regional Development Fund.

Finally, I would like to thank my parents Dr. S.K. Mandal and Meena Mondal, brothers Abhigyan and Abhijit, and friends Louis, Martin, Romain, Emilie, Mathilde, Jonas, Praveen, Kislay, Allie, Khiem, Dipnarayan, and Bunyarit for their support in this journey.

Contents

List of Figures

List of Tables

List of Abbreviations

ALBEF	Align Before Fuse
AOD	Average Odds Difference
API	Application Programming Interface
AUC	Area Under Curve
BERT	Bidirectional Encoder Representations from Transformers
BLIP	Bootstrapping Language-Image Pre-training
CEO	Chief Executive Officer
CLIP	Contrastive Language Image Pretraining
CNN	Convolutional Neural Network
CV	Computer Vision
EA	East Asia
EE	East Europe
ELBO	Evidence Lower Bound
EOD	Equal Opportunity Difference
FFHQ	Flickr-Faces-HQ
GAN	Generative Adversarial Networks
GPT	Generative Pre-trained Transformer
Grad-CAM	Gradient-weighted Class Activation Mapping
IAT	Implicit Association Test
iEAT	Image Embeddings Association Test
IIAS	Image-Image Association Score
IP	Internet Protocol

ISS	Image Similarity Score
ITAS	Image-Text Association Score
ITPAS	Image-Text Prompt Association Score
IVLAS	Idealised Vision Language Ability Score
LA	Latin America
LAION-400M	Large-scale Artificial Intelligence Open Network 400 Million Dataset
LDM	Latent Diffusion Models
LDM	Latent Diffusion Model
LFW	Labelled Faces in the Wild
LLM	Large Language Model
MCAS	Multimodal Composite Association Score
MLP	Multi Layer Perceptron
MS COCO	Microsoft Common Objects in Context
MSA	Multi-headed Self Attention
MSE	Mean Squared Error
NA	North America
NAWA	North Africa and West Asia
NLP	Natural Language Processing
NSFW	Not Safe For Work
SA	South Asia
SEA	South East Asia
SPD	Statistical Parity Difference
SSA	Sub-Saharan Africa
SSD	Single-shot Detection
SSL	Self-Supervised Learning
t-SNE	t-distributed Stochastic Neighbor Embedding
TTAS	Text-Text Association Score
TTI	Text to Image

VAE	Variational Autoencoders
VGG16	Visual Geometry Group 16
ViT	Vision Transformer
VL-BERT	Visual Linguistic Bidirectional Encoder Representations from Transformers
VLBAS	Vision Language Bias and Ability Scores
VLBS	Vision Language Bias Score
VPN	Virtual Private Network
VQ-VAE	Vector Quantised Variational Autoencoders
VQA	Visual Question Answering
VSRL	Visual Semantic Role Labelling
WE	West Europe
WEAT	Word Embeddings Association Test
xMCAS	eXtended Multimodal Composite Association Score
YFCC100m	Yahoo Flickr Creative Commons 100 Million Dataset

List of Publications

- **Abhishek Mandal**, Susan Leavy, and Suzanne Little. 2021. Dataset Diversity: Measuring and Mitigating Geographical Bias in Image Search and Retrieval. In Proceedings of the 1st Int'l Workshop on Trustworthy AI for Multimedia Computing (Trustworth AI '21), Oct. 24, 2021, co-located with ACM Multimedia, Virtual Event, China. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3475731.3484956

- **Abhishek Mandal**, Susan Leavy, and Suzanne Little. 2023. Measuring Bias in Multimodal Models: Multimodal Composite Association Score. In: Boratto, L., Faralli, S., Marras, M., Stilo, G. (eds) Advances in Bias and Fairness in Information Retrieval. BIAS 2023. Communications in Computer and Information Science, vol 1840. Springer, Cham. https://doi.org/10.1007/978-3-031-37249-0_2

- **Abhishek Mandal**, Susan Leavy, and Suzanne Little. 2023. Gender Bias in Multimodal Models: A Transnational Feminist Approach Considering Geographical Region and Culture. Proceedings of the 1st Workshop on Fairness and Bias in AI co-located with the 26th European Conference on Artificial Intelligence (ECAI 2023) Kraków, Poland, October 1st, 2023. https://ceur-ws.org/Vol-3523/paper8.pdf

- **Abhishek Mandal**, Suzanne Little, and Susan Leavy. 2023. Multimodal Bias: Assessing Gender Bias in Computer Vision Models with NLP Techniques. In Proceedings of the 25th International Conference on Multimodal

Interaction (ICMI '23). Association for Computing Machinery, New York, NY, USA, 416–424. https://doi.org/10.1145/3577190.3614156

- **Abhishek Mandal**, Susan Leavy, and Suzanne Little. 2023. Biased Attention: Do Vision Transformers Amplify Gender Bias More than Convolutional Neural Networks? In Proceedings of the 34th British Machine Vision Conference (BMVC), 2023. https://papers.bmvc2023.org/0629.pdf

- **Abhishek Mandal**, Susan Leavy, and Suzanne Little. 2024. Generated Bias: Auditing Internal Bias Dynamics of Text-To-Image Generative Models. In Proceedings of the 1st Workshop on Critical Evaluation of Generative Models and Their Impact on Society, co-located with the European Conference on Computer Vision 2024, Milan, Italy. Preprint: https://arxiv.org/abs/2410.07884.

- Marion Bartl, **Abhishek Mandal**, Susan Leavy, and Suzanne Little. 2024. Gender Bias in Natural Language Processing and Computer Vision: A Comparative Survey. ACM Comput. Surv. Just Accepted (November 2024). https://doi.org/10.1145/3700438

Auditing Imbalance and Bias in Deep Neural Networks for Multimedia Content Analytics

Abhishek Mandal

Abstract

This thesis introduces novel metrics and techniques to detect, measure, and mitigate gender and geographical bias in computer vision deep neural networks. It adopts an interdisciplinary approach, incorporating deep learning, feminist and decolonial theory, and ethics to address this issue. Artificial intelligence can amplify societal biases, marginalise vulnerable groups, and undermine public trust in AI, affecting its broader adoption. Bias in deep neural networks is complex, stemming from training data sourced from the internet, propagating through the machine learning pipeline, and impacting real-world applications. Factors such as model training methodologies, performance metrics, and model deployment further complicate this issue. The proposed metrics quantify a complex human concept – social bias in deep learning models – and provide insight into the internal bias dynamics of these 'black-box' systems. Part one examines geographical and gender bias and their intersection, focusing on bias origination in training data and its reflection in trained models. Novel metrics were introduced for measuring geographical bias in dataset creation methods and intersectional geographical and gender bias in multimodal models and they revealed their presence in both cases. Part two investigates how bias is managed internally in large visual-linguistic models like CLIP, DALL-E and Stable Diffusion. Traditional bias measures, focusing on accuracy, were found inadequate for capturing the extent of bias in multimodal vision models. Techniques from NLP were adapted to create metrics to capture bias in multiple modalities, including non-binary gender. The metrics revealed the presence of stereotypical gender bias in both models and showed that model architecture plays an important role in bias amplification. The metrics provided insight into how bias is handled inside the models. Part three uses data augmentation to debias vision models. This thesis develops and applies interdisciplinary metrics to detect, measure, and mitigate gender and geographical bias in vision models.

Chapter 1

Introduction

Computer vision models based on deep learning have improved substantially in their accuracy and capabilities since their introduction about a decade ago. Beginning with simple Convolutional Neural Networks (CNNs) that classified images into various categories in the early 2010s (Krizhevsky et al. 2012), these models have become increasingly powerful and capable, being now used to segment images, recognise scenes, and generate images among numerous other applications. Vision systems model our visual world and as they have become better at modelling useful visual information, they have also become more adept at modelling biased information. One such example is human biases. Vision models and systems have been shown to exhibit social biases that are similar to those present in our society. These include facial recognition software which perform less accurately for women and people with darker skin (Buolamwini et al. 2018a), video conferencing systems not generating correct virtual backgrounds for people with darker skin, image cropping algorithms (Dickey 2020), generating images of male data scientists and female flight attendants (Luccioni et al. 2024) and de-pixelation software lightening people's skin tones (J. Vincent 2020). Examples of such biases are shown in figures 1.1, 1.2, 1.3.

As computer vision systems and technologies become more popular and are more integrated into our daily lives, such biases have the potential to perpetuate and increase social inequality, systemise oppression and amplify discrimination against

marginalised and minority groups. These biases, if left unaddressed, can lead to unintended harm and potential loss of public trust in Artificial Intelligence.

1.1 Motivation

Computer vision has shown tremendous improvement in the last 5-6 years in terms of accuracy and capability. From detecting objects and classifying images, vision models can now understand scenes, perform zero-shot predictions and generate realistic images. Vision models are now increasingly becoming multimodal, often combining multiple modalities such as images and text (Radford et al. 2021). This has enabled vision models to model the visual world much better than earlier unimodal models and has paved the way for generative models such as DALL-E and Stable Diffusion. This has also enabled the models to learn complex societal biases such as those pertaining to gender, ethnicity, and culture. These biases arise within the training data and propagate and are amplified along machine learning pipelines. Examples of such biases include generative models generating images of male engineers and female nurses, assigning adjectives such as confident to men and gentle to women (Luccioni et al. 2024), multimodal models assigning traditionally male-dominated activities such as biking to men and female-dominated activities such as cooking to women (J. Wang et al. 2021) and visual question answering systems generating stereotypical and sexist answers for images containing women (Manjunatha et al. 2019). With the growing adoption of generative and multimodal deep learning in various industries and sectors such as advertising (Gupta et al. 2024), healthcare (Jindal et al. 2024), internet search (Team et al. 2023), and education (Baytak 2023), unchecked bias can lead to harm and undermine social progress.

The increasing complexity of vision models has made detecting and measuring biases more challenging. Previous metrics developed for measuring bias in vision models have been either designed for simpler models like image classifiers (Savani et al. 2020; Serna et al. 2021; T. Wang et al. 2019) or mostly perform posthoc analysis on multimodal generative models (Chinchure et al. 2023; Luccioni et al.

2024; Vice et al. 2023). The former metrics are unsuitable for multistage multimodal models and the latter although good for detecting bias, considers bias in the whole model making it difficult to isolate bias from the multiple models/stages/processes in these models. Therefore, there is a need for metrics which can 1) isolate bias at lower levels which is, bias processed during the internal information processing by the constituent models, 2) help understand how bias is handled internally by the models and thereby contribute to making the model less 'black-box' and 3) the output score(s) of the metrics should be quantified and easy-to-understand so that they can be used in cost functions to debias the models. The main objective of this thesis is to develop metrics to detect and measure social biases in large vision models. The primary focus is on gender bias and the secondary focus is on geographical bias. The intersection of these two types of biases is also briefly investigated. The research questions addressing these objectives are discussed in Section 1.5.

1.2 Visual Representation of Gender

Visual media such as images and videos are important forms of communication and comprise the majority of internet traffic today [1]. Feminist scholars have examined the presence of gender stereotypes in visual media and found it to perpetuate traditional gender norms and misogyny (Boyd 2010; Kosut 2012). They argue that gender is a social construct which is different from biological *sex* and creates norms for people how to live as per social norms (Cott 1987; Devinney et al. 2022). Such social norms are based around societal power dynamics where men exert greater power and control. Historically, in visual media, women have been represented in less powerful positions and showed family and marriage being more important to them (Signorielli 1990). Men have been traditionally represented as professionals whereas women as housewives (Paek et al. 2011). Visual depictions of traditional gender roles in images and videos can be described as *stereotypical representation of*

[1] https://www.statista.com/statistics/271735/internet-traffic-share-by-category-worldwide/ accessed: 26/6/2024

gender.

Such gender stereotypes have are present in visual datasets and computer vision models. A. Wang, Liu, et al. (2022) studied popular visual datasets such as MS COCO and found that women feature more in images related to kitchens such as ovens and sinks, and men feature more in outdoor images such as vehicles and sports. This shows traditional gender stereotypes related to *'domesticity'* of women (R. C. Vincent et al. 1987). Luccioni et al. (2024) found that Text-To-Image generation models such as Stable Diffusion are more likely to generate images of men for high-paying jobs and positions of power such as CEO and Engineer and images of women for lower-paying jobs and positions of less power such as Secretary and Housekeeper. When stereotypical gender roles are present in visual data or output of computer vision models, it is defined as *stereotypical gender bias*. Examples of such bias include DALL-E 2 generating images of men for the prompt *data scientist* and images of women for *flight attendant* as shown in figure 1.1. Another such example is a visual question answering system using CLIP showing stereotypical gender bias (figure 1.2). This type of bias is the main focus of this thesis. In this thesis, *men* and *women* are used as noun and *male* and *female* are used as adjective to represent binary gender. Non-binary gender is explicitly mentioned as such.

Figure 1.1: Images generated by DALL-E 2 for the prompts *an image of a data scientist* (top) and *an image of a flight attendant.*

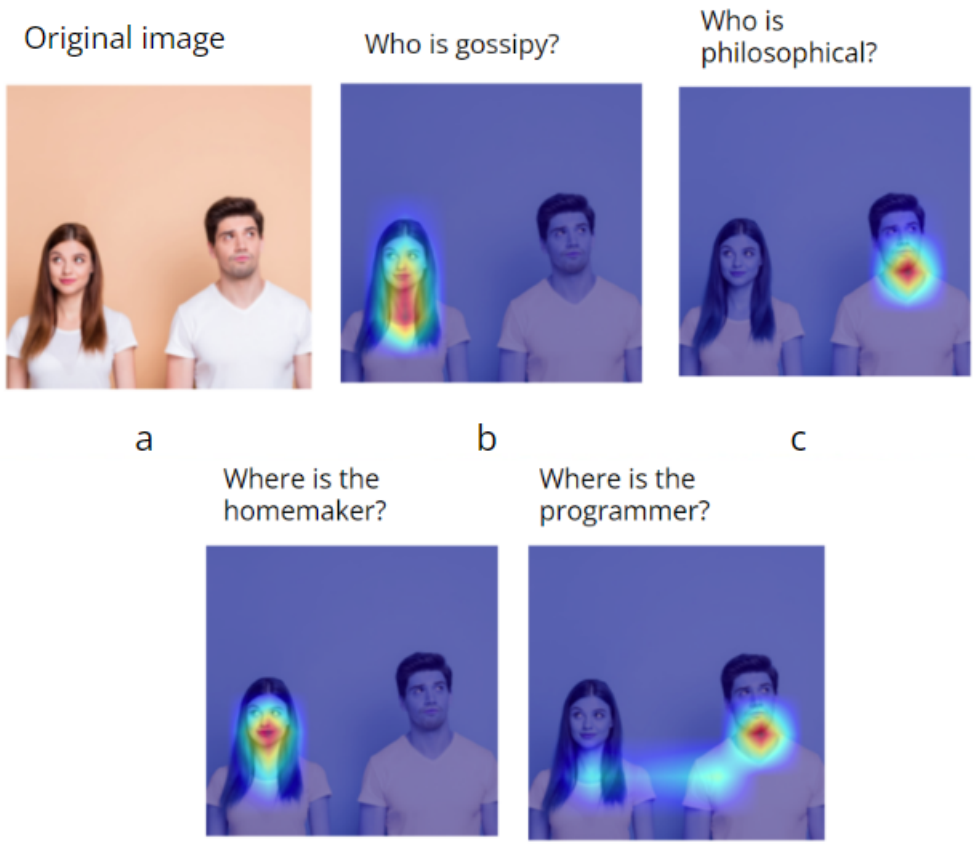


Figure 1.2: Grad-CAM analysis of CLIP VQA. a: original image, Grad-CAM localisation prompts: b: Who is gossipy?, c: Who is philosophical?, d: Where is the homemaker?, e: Where is the programmer?

1.3 Auditing Gender Bias in Vision Systems

In the context of machine learning, Fabbrizzi et al. (2022) defines bias as *"the prejudice of an automated decision system towards individuals or groups of people on the basis of protected attributes like gender, race or age"*. Bias is learnt by machine learning models from data in two main ways: correlations and causal relationships between the protected attributes and other data features and underrepresentation of minority groups (Ntoutsi et al. 2020). These biases are then amplified by the models during training (Raji et al. 2019; Schnabel et al. 2016). Therefore, biases in machine learning can be grouped into two categories: *data centric* and *model centric*. This categorisation is explained in section 2.2 (page 42). Both these types of biases can be unintentional. For example, in visual datasets, women are overrepresented in images of cooking-related scenes (A. Wang, Liu, et al. 2022; Zhao et al. 2017a). When vision models are trained on such data, they have been shown to start associating

cookware with women (Zhao et al. 2017a). Similarly, in popular visual datasets such as ImageNet and OpenImages, men are overrepresented in images of engineers (A. Wang, Liu, et al. 2022) and models trained on that data start associating men with engineering. Machine learning models are designed to score high on training metrics such as accuracy (Francois Chollet 2021). This means that if the model is trained on biased data, the model is incentivised to learn and exploit such biases. For example, if an occupations dataset was created from popular visual datasets or crawled from the web, it is very likely to be gender biased (Fabbrizzi et al. 2022; Mehrabi et al. 2021; A. Wang, Liu, et al. 2022). An image classifier trained on such a dataset will amplify the bias in trying to increase accuracy. If the majority of the images of CEOs are men then not predicting CEOs for images of women increases its accuracy.

Various metrics have been proposed to detect and measure both data and model-centric bias. Data-centric bias detection metrics include *statistical metrics, empirical and manual analysis*, and *benchmark datasets*. Model-centric bias detection metrics include *Image Embeddings Association Test* (Steed et al. 2021), *Model Leakage*, *Bias Amplification* (T. Wang et al. 2019), *Inside Bias* (Serna et al. 2021), and *Difference Metrics* (Savani et al. 2020). Most of these metrics target the learning representations of the deep learning models. However, they have some limitations. Most of these metrics are designed for simpler models such as image classifiers. This makes them unsuitable for use in large and complex multistage multimodal models. Even among image classifiers, the majority of them only work for Convolutional Neural Networks thus leaving out the Vision Transformer class of models entirely.

Another issue is that they only consider binary gender in their analysis. Metrics to detect and measure bias in more complex and multimodal vision models have also been proposed. These include *Vision Language Bias and Ability Scores* (Zhao et al. 2018), *TIBET* (Chinchure et al. 2023), and *Jaccard Hallucinations* and *Generative Miss Rate* (Vice et al. 2023). Other non-quantitative methods which rely on visual analysis of the results include *Stable Bias* (Luccioni et al. 2024). These methods

have two major limitations: 1) they perform post-hoc measurement of bias leaving out how bias is handled internally by these models making it difficult to isolate bias in multistage models and 2) these metrics (apart from Stable Bias) do not consider non-binary gender.

1.4 Geographical Bias in Vision Systems

Another social bias which affects vision systems is **geographical bias**. This bias can stem from an intersection of ethnic, cultural and localisation bias. Biases pertaining to ethnicity and culture in vision models have been studied and their presence confirmed (Buolamwini et al. 2018b; L. Elisa Celis et al. 2020b; Karkkainen et al. 2021a; A. Wang, Narayanan, et al. 2020). The researchers have used race and skin tone to approach this bias. However, this limits the definition and scope of bias. Also, using race and skin tone as it's proxy can also be problematic. Therefore, a more neutral axis: geographical location is used in this dissertation to approach this type of bias. Most popular visual datasets are curated in the Western world with data being crawled from the English language internet which leads to the models trained on them to exhibit a 'Western-centric' bias (Karkkainen et al. 2021a; Russakovsky et al. 2015). An example of this is shown in figure 1.3 which shows the Grad-CAM heatmap of a visual question answering system employing CLIP - a large multimodal model trained on 400 million image-text pairs crawled from the internet. It shows that for the question *who is the terrorist?*, the regions of the image with images of people from predominantly Muslim majority areas show the highest activation. This is discussed in Chapter 4.

Geographical bias also intersects with gender bias and this can cause harm to get multiplied for minority and vulnerable groups. For instance, the labels predicted by CLIP for images of women from Europe and North America are more positive than those from South Asia (discussed in Chapter 4). Therefore, traditional models of feminist theory may not capture the differences in bias by geographical regions.

Figure 1.3: Grad-CAM results for the question 'Who is the terrorist?' for images from all regions. Regions: Top-Bottom, L-R: WANA, WE, SA, SSA, EA, SEA, LA, EE. Same pattern for images of men and women. See Table 4.1 for abbreviations.

1.5 Research Questions and Hypotheses

The primary focus of this dissertation is gender bias and the secondary focus is on geographical bias and the intersection of both. The overarching aim of this dissertation project is to audit social biases in computer vision models using different lenses. This thesis is divided into three parts. The first part (chapters 3 and 4) analyses and develops metrics to detect and measure geographical bias. The intersection of this bias with gender bias is also investigated. The second part (chapters 5-9) looks into the architecture of diffusion models and develops metrics to detect and measure gender bias in each stage of the generation process. The third part (chapter 10) introduces techniques to mitigate gender bias in classification models. The dissertation starts by looking at social biases from a geographical lens focusing on dataset creation. Then the intersectionality between geographical and gender bias is explored using the diverse dataset creation framework for creating a dataset which was used for auditing multimodal models. Thereafter, multimodal models are audited for gender bias. Due to the sophisticated nature of bias, the complexity of multimodal models and the need to isolate low-level bias, the research questions are further subdivided into sub-questions. They are provided in the individual chapters.

1.5.1 Part 1: Gender and Geographical Bias in Visual Datasets and Models

Datasets used for training deep neural networks are generally curated by using search engines to retrieve images or by scraping from image hosting sites (Karkkainen et al. 2021b; Lin et al. 2014). This however leads to the dataset being biased due to the nature of the demographics on the internet which tends to be young, male and Western-centric (L Elisa Celis et al. 2020a; Shankar et al. 2017; A. Wang, Liu, et al. 2022).

Hypothesis: Current dataset creation methods based on web crawling can lead to a 'Western-centric' cultural, gender and geographical bias in visual datasets. This is reflected in models trained on such data leading to geographical, gender, and intersectional bias.

RQ1 How does altering the location and language when using Internet search to collect a dataset affect the visual diversity of faces, and how can this be measured?

RQ2 How does gender bias exhibited by large multimodal vision models differ by geography and how can such intersectional bias be measured?

1.5.2 Part 2: Measuring Gender Bias in Multimodal Models

This section introduces metrics to detect and quantify gender bias, including non-binary gender bias, in multimodal models. Also investigated is how bias is handled internally by the model and how it gets amplified.

Hypothesis: Multimodal vision models (including Text-To-Image generative models) learn gender bias from the training data and amplify it. This can be isolated and measured. Model architecture plays an important role in this.

RQ3 How can gender bias in multimodal vision models be isolated and measured and what role does model architecture play in bias amplification?

1.5.3 Part 3: Debiasing Vision Models

This section discusses debiasing vision models using data augmentation techniques.

Hypothesis: Data augmentation techniques can be used for debiasing vision models trained on biased data.

RQ4 How can biased vision models be debiased using data augmentation techniques?

1.6 Thesis Structure

The thesis has the following structure.

Chapter 1 introduces the work explaining the motivation behind this research, a brief background about the theoretical foundation, the research questions and hypotheses and lists out the metrics.

Chapter 2 discusses the background and relevant work. It consists of two parts. The first part discusses computer vision architectures including multimodal models and a special focus on Text-To-Image diffusion models. The second part discusses the theoretical foundations drawing upon feminist theory and bias metrics previously developed along with their limitations.

Chapter 3 introduces and studies *Geographical Bias*, proposes two novel metrics to detect and measure it and a methodology to mitigate it.

Chapter 4 presents an audit bias in CLIP (Contrastive Language Image Pretraining) at the intersection of gender and geography drawing upon concepts from transnational feminism.

Chapter 5 uses metrics from Natural Language Processing and adapts them to measure gender bias in CLIP. The measured bias is studied against real-world data related to the gender pay gap and women's workforce participation in different occupations.

Chapter 6 introduces MCAS (Multimodal Composite Association Score) - a comprehensive and composite score to measure gender bias in multimodal models in multiple modalities. This score is used to measure gender bias in Text-To-Image (TTI) diffusion models DALL-E 2 and Stable Diffusion.

Chapter 7 investigates the relationship between computer vision architecture and bias. A new metric is introduced which can effectively measure bias in both Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs).

Chapter 8 introduces two novel metrics to isolate bias at a local level in TTI models and studies how bias is handled internally by these models.

Chapter 9 extends MCAS to measure non-binary gender bias in TTI models.

Chapter 10 discusses saliency-based data augmentations to debias image classifiers.

Chapter 11 concludes this thesis and summarises the outcomes and contributions.

Figure 1.4 shows the novel metrics superimposed on a TTI diffusion model illustrating how they measure bias. Table 1.1 provides a summary of the metrics proposed in this dissertation.

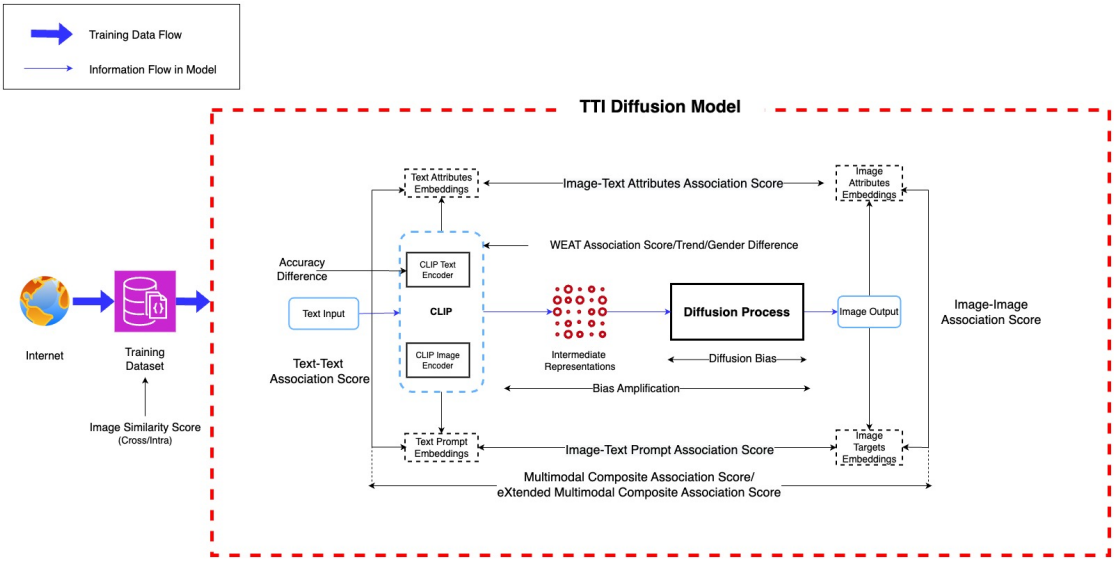


Figure 1.4: Composition and purpose of proposed metrics within the context of a Text-to-Image (TTI) Diffusion Model.

Table 1.1: Summary of all novel metrics proposed in this thesis. Intersectional refers to geography and gender. *WEAT, developed by Caliskan, Bryson, et al. (2017) for NLP, has been adapted for use in computer vision in this research.

Metric	**Type**	**Bias**	**Modality**	**Chapter**
Image Similarity Score Cross & Intra	Data centric	Geographical	Image	3
Gender Difference & Trend	Model centric	Intersectional	Image	4
WEAT Association Score for CV*	Model centric	Gender	Image	5
Image-Image Association Score Image-Text Association Score Image-Text Prompt Association Score Text-Text Association Score Multimodal Composite Association Score	Model centric	Gender	Image Multimodal Multimodal Text Multimodal	6, 7, 8, 9
Accuracy Difference	Model Centric	Gender	Image	7
Diffusion Bias	Model Centric	Gender	Multimodal	8
Bias Amplification	Model centric	Gender	Multimodal	8
eXtended Multimodal Composite Association Score	Model centric	Gender	Multimodal	9

Chapter 2

Background and Related Work

This chapter provides background information and an overview of the relevant literature on bias in deep neural networks used in computer vision. This chapter is divided into two sections. The first section discusses different computer vision architectures. Model architectures play an important role in how information is processed by the models internally (Khan et al. 2022; Park et al. 2022). Multimodal models including TTI generative models employ multiple models internally and are typically multistage (Radford et al. 2021; Ramesh, Pavlov, et al. 2021; Rombach et al. 2022). Therefore, it is important to investigate how model architecture can play a role in bias amplification. The second section discusses relevant literature on gender, its representation in visual format and gender bias.

2.1 Deep Learning for Computer Vision

Since the introduction of AlexNet (Krizhevsky et al. 2012), deep learning has been the preferred technology for use in computer vision tasks such as image classification. AlexNet is a Convolutional Neural Network (CNN) which uses convolution operations to extract features from images. Over the past decade, there have been many improvements on the original AlexNet. These include Inception (Szegedy et al. 2015) which learns features at different scales, ResNet (He et al. 2016) which introduced residual connections to tackle the problem of vanishing gradient thus al-

lowing for very deep CNNs, EfficientNet (Tan et al. 2019) which used meta learned architectures, and U-Net (Ronneberger et al. 2015) for image segmentation.

Vision Transformers (ViTs) (Kolesnikov et al. 2021), first introduced in 2020 based on the transformer architecture from NLP, uses a mechanism called attention to extract contextual information from a sequence of input. ViTs process an image as a sequence of pixels in the form of patches. ViTs have achieved superior accuracy than those of CNNs in many computer vision tasks such as image classification (Khan et al. 2022).

Although image captioning extended image classification to image description, Contrastive Language Image Pretraining (CLIP) (Radford et al. 2021) employing contrastive learning enabled zero-shot predictions and association of long text sentences for images. CLIP also allowed for a new generation of generative vision models such as DALL-E (Ramesh, Dhariwal, et al. 2022; Ramesh, Pavlov, et al. 2021) and Stable Diffusion (Rombach et al. 2022). They use CLIP to convert text into intermediate embeddings and then use diffusion process to generate images.

The next sections discuss these architectures in detail.

2.1.1 Basic Computer Vision Architectures

Convolutional Neural Networks

Convolutional Neural Networks or ConvNets or CNNs have been the driving force behind the adoption of deep learning techniques in computer vision in the last decade (He et al. 2016; Krizhevsky et al. 2012; Szegedy et al. 2015). CNNs use convolution layers to perform convolution operations on input images and were first introduced in the 1980s by Fukushima (Fukushima 1980). LeCun et al. (1998) used gradient descent to train CNNs to recognise handwritten digits in the 1990s. Cireşan et al. (2010) showed in 2010 that CNNs could be trained on Graphics Processing Units (GPU) in a massively parallel way using backpropagation (reverse automatic differentiation introduced in 1970 by Linnainmaa 1970). In 2012, Krizhevsky et al. (2012) introduced AlexNet – a 60-layer deep CNN trained using backpropagation

on GPUs which won the ImageNet challenge (a large image classification challenge on a visual dataset based on wordnet).

CNNs have convolution layers which consist of filters which glide over the input image. During the forward pass, the convolution layers produce a convolved output of the image and the filters which is a dot product thereby creating a feature map. The deeper layers downsample the features, leading to dimensionality reduction. Activation functions such as ReLU introduces non-linearity and pooling layers perform the downsampling. Finally, fully connected layers at the very end perform the classification (Francois Chollet 2021).

CNNs have strong inductive biases which constrain the hypothesis space. Examples include translation invariance, pooling, and convolution itself. These allow the models to focus on certain spatial relationships in the image (Francois Chollet 2021). Strong local inductive biases also make the loss landscape more convex. Convolutions are high pass filters which allow them to capture sharp features such as edges in images (Park et al. 2022).

Vision Transformers

The transformer architecture was first proposed by Vaswani et al. in 2017 for use in Natural Language Processing and revolutionised the field. It was later adapted for vision tasks by (Kolesnikov et al. 2021) and has since then outperformed CNNs in many vision tasks such as image classification (Khan et al. 2022). ViTs process images as a sequence of pixels with pixels grouped together in patches. They use an attention mechanism called Multi-headed Self Attention (MSA) to compute how similar one patch is to the other patches. This allows them to capture long-term dependencies and have a global context. They do not have inductive biases like CNNs and have a shallower loss landscape. Unlike convolutions, MSAs are low pass filters and diversify feature maps instead of aggregating them (Park et al. 2022).

2.1.2 Multimodal Deep Learning

In the early days of deep learning, most state-of-the-art models were trained using supervised learning. However, this limited their performance to 'known categories' and zero-shot predictions. Contrastive Language Image Pretraining (CLIP) was introduced by Radford et al. (2021) which used natural language supervision and contrastive learning to achieve performance similar to or better than supervised models on image classification tasks. CLIP is trained on 400 million image-text pairs crawled from the internet. CLIP is trained to predict which set of words (text) is best paired with an image. It does this by using a multimodal embedding space by jointly training an image and a text encoder to maximise the cosine similarity of both the image and text embeddings. Both the encoders are initialised without any pre-trained weights. CLIP uses CNNs (mainly ResNets) and ViTs as image encoders and a Transformer as a text encoder. CLIP surpassed most state-of-the-art models and humans in zero-shot predictions on various visual datasets.

2.1.3 Generative Computer Vision

Variational Autoencoders were first proposed by Kingma et al. (2013) which used a CNN-based encoder and a decoder with the decoder using transpose convolutions. The encoder takes in input (an image) and maps it to a probability distribution in the latent space outputting the mean and variance of a multivariate normal distribution in the latent space. From the predicted mean and variance, a point is sampled in the latent space which is then passed to the decoder for reconstruction. It uses a reconstruction loss (MSE) and Kullback-Leibler (KL) divergence regulariser. The model is trained using a sum of the reconstruction loss and the KL divergence regulariser and is known as evidence lower bound (ELBO). An improvement over VAEs called Vector Quantised Variational Autoencoders (VQ VAEs) was introduced by Van Den Oord et al. (2017). It has two main improvements over VAEs: discrete latent space and vector quantisation. The discrete latent space allows the encoder to encode in discrete values instead of continuous values thereby allowing for better

capturing of discrete data (Van Den Oord et al. 2017).

Text-to-Image (TTI) Diffusion Models: Diffusion models are somewhat similar to Autoencoder models in how they are trained and the data flow and were first introduced in the later half of the 2010s (Ho et al. 2020; Sohl-Dickstein et al. 2015). They are inspired by non-equilibrium thermodynamics but unlike Autoencoders, they have high dimensionality, same as the data. Diffusion models generally employ two series of operations: forward and backward diffusion processes. In the forward diffusion process, Gaussian noise is added to the original input image in a sequence of steps producing a sequence of noisy transformations. As the steps increase, the original input image becomes unrecognisable and loses its distinguishable features. When this process is reversed, the original input image can be reconstructed and this process is called backward or reverse diffusion. The forward diffusion process is similar to encoders in Autoencoders and the reverse diffusion process is similar to the decoder.

DALL-E was released by OpenAI in 2021 and is considered a cornerstone in generative vision technology. It has two main components: a discrete variational autoencoder (dVAE) similar to VQ-VAEs and an autoregressive transformer. It is a multimodal implementation of OpenAI's hugely popular GPT-3 (Generative Pretrained Transformer). It is trained in a two-stage process: first, the dVAE compresses 256x256 RGB images into a 32x32 grid of image tokens thereby reducing the context size of the transformer; then the image tokens are concatenated using Byte Pair Encoding (BPE) with text tokens and the autoregressive transformer learns the joint distributions over the image and text tokens (Ramesh, Pavlov, et al. 2021).

DALL-E 2 is the second version of DALL-E and was released in 2022. It has two components: CLIP and unCLIP. In the first part, it uses CLIP to generate embeddings of the text inputted by the user. Then the embeddings are modelled using a Gaussian diffusion model and is called the *Diffusion Prior*. The authors also experimented with an *Autoregressive Prior* where the CLIP embeddings are converted

into sequential discrete codes and predicted autoregressively. The diffusion prior is continuous. The authors did not find significant differences in the performance of the two different prior models and used the diffusion prior as it is computationally more efficient. The diffusion prior consists of a decoder-only Transformer with a causal attention mask on a sequence which consists of: the encoded text/caption, the CLIP text embeddings, a diffusion timestep encoding, the noised CLIP embedding of the image, and a final embedding from the Transformer, all in order. Then the diffusion model (unCLIP) models from the representation space (from the prior) to generate the image via reverse diffusion. In order to generate high-resolution images, they used two upsamplers (Ramesh, Dhariwal, et al. 2022).

Stable Diffusion was released in 2021 by LMU Munich and Runway ML. It is based on a new type of diffusion model called *Latent Diffusion Models* (LDMs). It is a multi-stage multimodal model similar to DALL-E and uses CLIP for the initial text encoding. It consists of three main components: CLIP for text encoding, a UNet + scheduler for gradual diffusion from the latent space and an autoencoder decoder for the final image generation. The first part where the text prompt is converted into encodings is similar to DALL-E 2. The original paper used BERT and earlier released versions used OpenAI's CLIP but later generations used OpenCLIP – an open-source version of CLIP[1]. The text encoder generates an output of 77 token embedding vectors each in 768 dimensions. This embedding is then fed into the UNet along with a tensor made up of noise. This step outputs a processed information array of 4x64x64 dimensions which is then fed to the decoder for the final image generation. The diffusion process is not run in the pixel space but in the 'latent space' in order to conserve computing resources and boost speed (Rombach et al. 2022).

[1]https://stability.ai/news/stable-diffusion-v2-release accessed: 5-7-2024

2.2 Bias in Computer Vision

In previous research, the definitions of bias and fairness primarily revolved around distinctions, focusing on the different and more advantageous treatment of one protected group in comparison to another. Bias denotes the existence of distinctions, whereas fairness pertains to the lack of distinctions. In the paradigm of machine learning, Mehrabi et al. (2021) defines fairness as *"the absence of any prejudice or favouritism toward an individual or group based on their inherent or acquired characteristics"*. They make a distinction between two forms of fairness: group fairness and counterfactual (individual) fairness. Group fairness is attained when the performance of the relevant groups equals the same statistical score, such as an F1 measure. Individual fairness is achieved when altering the protected group does not impact the model output. For instance, a facial detection algorithm that performs effectively on an image of a man should demonstrate an equal level of effectiveness on an image of a woman within the same context (Buolamwini et al. 2018b). This section provides an overview of the literature relevant to social biases in computer vision.

2.2.1 Gender Bias in Computer Vision

The European Union Artificial Intelligence Act (P9_TC1-COD(2021)0106) on the guiding principles of trustworthy AI includes fairness and non-discrimination and states that

"Diversity, non-discrimination and fairness means that AI systems are developed and used in a way that includes diverse actors and promotes equal access, gender equality and cultural diversity, while avoiding discriminatory impacts and unfair biases that are prohibited by Union or national law." (EuropeanParliament 2024)

Gender being a **sensitive attribute** requires people not to face any discrimination based on it. The various genders (*male, female, non-binary, agender and gender queer*) are called **protected groups** (L Elisa Celis et al. 2020a). Any Artificial Intelligence system which discriminates on the basis of gender is therefore

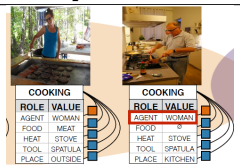
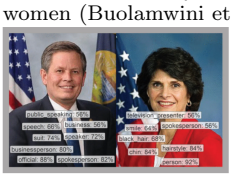
in violation of European law. In visual mediums such as images and videos, gender is expressed either explicitly through labels or annotations or derived implicitly through interpretations of visual features.

The Social Construction of Gender

Many feminist theorists (Butler 1990; Devinney et al. 2022) define gender as a social construct which is different from the biological *sex* (i.e. *male, female, and intersex*). It is subject to change and operates over a spectrum. They suggest that social gender encompasses an individual's gender identity (their personal experience of gender), gender expression (how they express and enact their gender, including the roles they adopt), and perceived gender (how others perceive and assign gender to the person). This interplay influences their overall experience and manifestation of gender (Butler 1990; Devinney et al. 2022). Keyes et al. (2021) conceptualised gender as a dynamic and evolving construct rather than a predetermined and fixed attribute and argued that it emerges through the actions and expressions of both the individual and the societal context in which they engage. Additionally, gender intersects with other dimensions of one's identity, including race, socioeconomic background, religion, ability, and nationality. In images and visual media, gender performance is through features such as a person's hairstyle, clothes, facial hair or their use of make-up, among others (A. Wang, Liu, et al. 2022; Y. Zhang et al. 2020).

Researchers (Faludi 2011; A. Wang, Liu, et al. 2022; Zhao et al. 2017b) have identified two major axes of this social construct. They are: the power dynamics and control with men having power and making decisions and spheres of social roles with men's sphere being outside the house and women's on the inside. This thesis takes the standpoint that perceiving gender on the basis of social norms and traditional gender roles is harmful, unfair and therefore biased. In the context of computer vision, such harms include facial recognition systems being less accurate on the faces of women (Buolamwini et al. 2018b), image recognition systems generating

Table 2.1: Overview of bias categories in relation to gender bias in computer vision

Bias	Description	Example
Selection/sampling bias	Biases introduced as a result of the process by which instances are included in a dataset	 (Zhao et al. 2017a)
Cultural bias	Bias due to data origin or background of the data collector/creator	Visual datasets depicting images from overwhelmingly Western countries (A. Wang, Liu, et al. 2022)
Demographic/population bias	Bias due to inclusion of demographic groups and demographic diversity	Facial recognition software shows less accuracy on darker-skinned women (Buolamwini et al. 2018b)
Labelling bias	Biases in the annotations or labels used to identify the (subjects) in the data arising due to errors or human biases	 (Schwemmer et al. 2020)
Semantic bias	Bias captured in pre-trained representations that contain semantic information	
Over-amplification	Models learn gendered differences and correlations and exploit them at prediction time, over-amplifying the connection	Generative models generate highly stereotypical images e.g. men for CEO, and women for housekeeper
Framing bias	Biases that arise due to how subjects are presented in an image as a result of image capture techniques such as the angle, focus and cropping	 (Birhane at al. 2021)

tags related to looks and beauty for images of women (Schwemmer et al. 2020) and (mis)identifying gender from background and scenes (A. Wang, Liu, et al. 2022; T. Wang et al. 2019), and generative models generating images of male engineers and female nurses (Luccioni et al. 2024).

Visual Representation of Gender

Scholars such as Gauntlett (2008), Boyd (2010), and Gunter (1995) have identified the presence of gender stereotypes in visual media including images and videos. This thesis mainly investigates two types of stereotypes: men being associated with positions of power and higher salaries than women for example male CEO and engineer vs female secretary and housekeeper and domesticity of women, for example, women being associated with household items and work such as cooking and kitchen appliances and playing indoor sports and men being associated with outdoor items and

activities such as lathe machine and fishing and playing outdoor sports. Research by Gauntlett (2008) and Gunter (1995) has shown that visual media such as film and television has depicted men in professional roles. Their research showed that women are either misrepresented or underrepresented. In television and film media, men outnumber women. Men are often portrayed in powerful decision-making roles while women are relegated to supporting roles such as a female secretary assisting a male manager or a female nurse assisting a male doctor (Gauntlett 2008; Gunter 1995).

Boyd (2010) found that one area where women are overrepresented is when the visual content involves the objectification of women for male consumption. Performing an image search on the web with any nationality such as *Indonesian or Japanese* plus girl/woman mainly returns sexualised images (Boyd 2010). Paek et al. (2011) pointed out that in advertisements, a big source of visual media, men are more likely to be seen in professional settings and selling business items like computers (Paek et al. 2011).

Researchers have pointed out that in visual media, traditionally, women were depicted as homemakers with marriage, family, and parenthood being shown to be more important for them. Men were portrayed as smart, aggressive, outgoing and intellectual while women were portrayed as docile, weak, and ineffectual (Gunter 1995). Although this has slowly changed, women are still overrepresented in domestic roles (Elasmar et al. 1999). Even when women are shown with respect to their professional achievements, they are more likely to be sexualised with greater emphasis on physical appearance. Examples include photos of women athletes emphasising their looks (Gauntlett 2008). Advertisements often show women selling household appliances and cookware (Paek et al. 2011).

These stereotypes are mirrored on the internet from where datasets for training machine learning models are created. When models are trained on such datasets, they inherit these biases leading them to infer gender from spurious correlations by computer vision models such as cooking (female) and plumbing (male), occupations

such as nurse (female) and programmer (male) and visual scenes such as playing indoor sports (female) and outdoor sports (male) (A. Wang, Liu, et al. 2022; T. Wang et al. 2019). Therefore, to summarise, gender bias in vision systems stems from the training datasets (*Data Centric Gender Bias*) and is learnt by the models during training (*Model Centric Gender Bias*) which propagate and amplify them. Table 2.1 lists out some common types of gender bias in computer vision. Note that *Data and Model Centric Gender Bias* have no relation to *Data and Model Centric AI* (Hamid 2022).

Data Centric Gender Bias

This section outlines previous research in detecting and measuring data-centric gender bias. Although this dissertation mainly focuses on model-centric gender bias, the model-centric metrics and experiments draw heavily on insights from the literature on data-centric bias.

The presence of gender bias in visual datasets is heavily impacted by the source of the images and the process of creating training labels. Common origins of visual data include online image-hosting platforms, encyclopedias, and social networking sites. Detecting data-centric bias entails examining and quantifying gender bias in visual training datasets, typically achieved through statistical methods, contextual representations, and empirical analysis of the datasets.

This section outlines the various existing data-centric bias detection metrics and methodologies.

Statistical metrics: Statistical metrics, including frequency counts, are frequently utilized for the assessment and analysis of bias in datasets. This analysis spans the examination of demographic characteristics like age, gender, and race, as well as the application of statistical techniques, such as t-distributed Stochastic Neighbor Embedding (t-SNE), to visualize the distribution of images. Employing statistical techniques, Singh et al. (2020) conducted a comparison of image retrieval outcomes across diverse image search and hosting platforms, including Bing, Twitter, the New

York Times, Wikipedia, and Shutterstock. They specifically focused on gender-skewed occupations such as librarian, nurse, programmer, and civil engineer. In their findings, when compared to data from the US Bureau of Labor Statistics regarding gender participation in those occupations, it was revealed that the New York Times exhibited the most balanced representations, while Twitter showed the least balance.

Illustrating the impact of selection and labelling bias, skewed data from online platforms is subsequently utilized to construct datasets for training deep learning models. In a study by K. Yang et al. (2020), an examination of image representations within the ImageNet hierarchy revealed the presence of gender bias in this widely used dataset. Instances were identified where labels contained gendered and offensive language. Moreover, many annotations exhibited gender bias, such as the term *banker* predominantly associated with male images. To address these biases, K. Yang et al. (2020) leveraged Amazon Mechanical Turk to rectify and balance the data with respect to race and gender.

Gender bias within visual datasets takes various forms. Beyond explicit labels, bias subtly manifests in the depiction of gender within visual scenes. A. Wang, Liu, et al. (2022). conducted an analysis of popular vision datasets, including COCO, OpenImages, and YFCC100m. Their examination of image scenes revealed that outdoor scenes, such as *transportation, snow and ice, deserts and sky, fields*, and *park*, exhibited a higher representation of men. In contrast, indoor scenes, including *shopping, dining, indoor sports*, and themes related to *leisure* and *home*, had a higher representation of women. When examining scenes related to objects, images categorized under *sports* and *vehicles* had more representations of men, while those in categories such as *kitchen, appliance, indoor*, and *furniture* had more representations of women. Such issues have the potential to introduce framing bias in vision datasets.

Research has found exaggerated portrayals of gender stereotypes are frequently prevalent in image datasets. Kay et al. (2015) investigated the impact of stereotype

amplification, systematic over- and under-representation, and individuals' perceptions of gender stereotyping in image search results provided by search engines. Their experiments focused on occupations characterized by strong gender biases, and their hypotheses were informed by data on occupations collected from the US Bureau of Labor Statistics. The study revealed a notable exaggeration of stereotypes, particularly in images associated with women in traditionally female-dominated occupations and vice versa. Terms such as *sexy* and *attractive* yielded a notably high percentage of women's images ($p = 0.8$ and 0.72, respectively), while terms like *professional* and *trustworthy*" returned images of men ($p = 0.27$ and 0.6, respectively). This serves as another illustration of selection bias.

Combining statistical metrics with visual representations can offer valuable insights into the characteristics and distribution of bias within data. Karkkainen et al. (2021) employed t-SNE to depict the distribution of images by race, facilitating the analysis of their dispersion. These visualizations prove beneficial for comprehending data distribution and bias within a high-dimensional space, and there is the potential for extending such methods to investigate gender bias.

Contextual representations: In the examination of the OpenImages dataset, A. Wang, Liu, et al. (2022) observed a tendency where, in images featuring people with musical instruments, men were frequently portrayed as the ones playing or actively interacting with the instruments, while women were more commonly depicted as part of the audience. Consequently, men tended to be visually closer to the instruments. The analysis also revealed that men were more inclined to be involved with objects associated with sports and vehicles, whereas women were associated with objects related to the kitchen, furniture, and accessories. These portrayals have the potential to introduce framing and selection bias.

Empirical and Manual Analysis: Quantifying implicit gender bias can pose challenges with quantitative and statistical tools. These biases often remain concealed within the contextual details of images and their accompanying texts, necessitating a qualitative analysis approach. Birhane et al. (2021) delved into the issue of

harmful stereotypes within extensive image datasets, including LAION-400M, which comprises 400 million images and text extracted from alt-text on web pages sourced from the Internet. Within this dataset, the researchers identified both harmful text and sexually explicit images. Utilizing text-based image retrieval methods, they discovered harmful images associated with terms related to women, such as *Maa*, *Aunty*, and *Abuela*. Similarly, the study identified geographic biases, which the authors attribute to the data creation method involving unfiltered Internet crawling, resulting in labelling bias. Using a similar methodology, A. Wang, Liu, et al. (2022) examined the widely used image dataset OpenImages. Their analysis revealed that in images featuring people and flowers, women were predominantly depicted posing with flowers, while men were more likely to have flowers used as background decoration. Additionally, by employing a convolutional neural network (CNN) trained on OpenImages, the researchers observed that the model exhibited a tendency to classify individuals engaged in indoor sports, like swimming, as women, and those participating in outdoor sports, such as football, as men.

Benchmark Datasets: An alternative approach for identifying and gauging bias in computer vision involves the creation of benchmark datasets to assess variance and diversity. In their study, Karkkainen et al. (2021) scrutinized several prominent visual datasets, including IMDB-WIKI, LFW+, CelebA, and UTKFace, exploring the racial and gender distribution within them. Their findings indicated that, with the exception of FotW and UTKFace, none of the datasets demonstrated a balanced representation. In response, the researchers devised their own dataset named FairFace, characterized by a balanced racial and gender distribution. To evaluate their dataset's performance, they trained a ResNet-34 convolutional neural network on each dataset individually and tested them across diverse sets of images—such as geo-tagged tweets, media photographs, and protest datasets—all balanced for race and gender. The evaluation involved measuring balanced accuracy on gender using a variation of equalized odds to assess the disparity between true and predicted gender. Remarkably, Karkkainen et al. (2021) observed that

the model trained on FairFace outperformed models trained on nearly all other datasets.

Model Centric Gender Bias

Model-centric bias metrics audit biases in the models. When computer vision models are trained on biased data, they learn patterns in the data corresponding to social biases many of which are spurious correlations. These patterns are then reflected by the models in their outputs. Model-centric metrics detect and measure these patterns in the model's behaviours. Metrics such as the Image Embeddings Association Test (Steed et al. 2021) perform standardised tests to measure such biased patterns in the model's outputs. Model Leakage (T. Wang et al. 2019) and Bias Amplification measure the effect of spurious correlations in model outputs on gender bias. When spurious correlations follow a definite pattern they contribute to bias. *Image Embeddings Association Test:* A widely adopted approach for detecting bias in computer vision involves adapting techniques from other domains, such as natural language processing (NLP).Steed et al. (2021) proposed a methodology, named the Image Embeddings Association Test (iEAT), inspired by the Word Embeddings Association Test (Caliskan, Bryson, et al. 2017), to quantitatively assess implicit human biases. Their hypothesis posits the presence of human-like biases within the image embeddings utilized by neural networks. The iEAT evaluates the correlation between concepts by employing two sets of target concepts and attributes (e.g., male–career, female–family). The test gauges the statistical differential association between these sets based on the model's embeddings, providing a standardized measure of the probability of the absence of bias.

Caliskan, Ajay, et al. (2022) devised a similar evaluation of model bias utilizing two Computer Vision models, namely iGPT and SimCLRv2, both pre-trained on Imagenet. The assessment for gender bias involved subjecting the models to two tests. The first test, known as the Gender-Career test, gauges the relative association of the category "male" with career attributes like business and office, and

the category "female" with family-related attributes like children and home. The second test, termed the Gender-Science test, assesses associations between "male" with science and engineering, and "female" with liberal arts and writing. The results indicated significant gender bias in both models for both the Gender-Career and Gender-Science tests, with standardized probability values higher for SimCLRv2 compared to iGPT.

The use of iEAT to measure bias in CV models is relatively recent. Sirotkin et al. 2022 used the iEAT to study the effect of Self-Supervised Learning (SSL) on bias (Steed et al. 2021). They studied three SSL models: *geometric*, *clustering-based* and *contrastive*. Using the Gender-Career and Gender-Science tests, they found that contrastive models had the highest bias.

The main limitation of this metric (iEAT) is that the pair of targets required for bias testing limits the number of biases that can be studied and the scope of bias definition as well. For example, the male-career test limits the scope of bias analytics as there is a wide variation in how gender bias is manifested in career roles such as some occupations like engineering being male-biased but nursing is female-biased. The test statistic is difficult to interpret.

Model Leakage: The concept of model leakage was introduced by T. Wang et al. (2019) while investigating spurious correlations in vision datasets that contribute to bias. For instance, in the widely used image dataset COCO, there is a higher prevalence of images featuring both plates and women compared to images featuring both plates and men. This discrepancy might result in gender bias in models, leading to a strong correlation between plates and the female gender. The capacity to deduce gender from unrelated predicted image labels (e.g., *plate* predicting *female*) is referred to as 'leakage'. This phenomenon is quantified by the percentage of examples in the dataset that 'leak' information about a protected label (e.g., *female*) through the model's predicted labels (e.g., *plate*). This is evaluated by training a new function designed to predict gender from the labels. This metric addresses a crucial issue of spurious correlations but it is designed for and tested on Convolutional

Neural Network based image classifiers. As this metric targets convolution layers specifically, it may not work on other architectures such as Vision Transformers.

Bias Amplification: Also, of significant concern is the potential for models not only to mirror the bias present in the dataset, as seen in model leakage, but to potentially intensify or magnify this bias. Referred to as *bias amplification*, T. Wang et al. (2019) defined it as the disparity between the assessed model leakage and the dataset leakage. Alternatively, Zhao et al. (2017) gauged bias amplification by contrasting the impact of bias correlation learned by a model during training. For instance, in visual semantic role labelling (VSRL), labels (*person*, *spatula*, *oven*, etc.) are generated for a scene, such as *kitchen*, and the resulting activity depicted in the image is *cooking*. If the positive correlation between two terms (e.g. *women* and *cooking*) is heightened by the model over the evaluation dataset, this is termed bias amplification. The overall bias amplification score for the model is computed as the average magnitude for all pairs displaying bias. Both of these metrics aim to quantify the influence of bias exerted by a model over a reference dataset; however, there is a risk of oversimplification and dependency on a well-annotated reference dataset for comparison. This metric has similar limitations to the previous one that is is designed and tested only for CNN-based classifiers.

The above metrics are designed to find patterns of bias in the model's output. However, they do not measure bias in the representations learnt by the model during training. Measuring bias in this level can make the audit more robust and also provide insights into how bias is handled internally by the models.

InsideBias: An alternative view of bias measurement is to inspect the internal structures of the model such as the activation of filters in a convolutional neural network (CNN), commonly used for state-of-the-art CV models. Serna et al. (2021) proposed such an approach to measure demographic bias and evaluated it by training two CNN architectures (VGG and ResNet) on DiveFace, a diverse dataset with representations from across the world, and on biased data by increasing the representation of a particular group. To assess the impact of biased data, an *Activation*

Ratio is calculated. Activation, a measure of the contribution of network layers in generating the feature map, is compared between networks trained differently and the resulting models are considered biased if the ratio is less than a defined threshold. Generally, the final layers of the network have the highest contribution and are evaluated in this way. Serna et al. (2021) found that the unbiased models had a higher activation ratio for the last layers than the biased ones supporting their claim that this approach can give a good indication of model bias. This metric is also specifically designed for CNNs and bias is measured at the activations of the convolution layers making it unsuitable for ViTs.

Difference Metrics In contrast to inspecting the internal structures of the model, the difference in the model's output predictions can also be compared statistically. This is distinct from bias amplification methods (see previous section) that train a predictive function to reverse the model's outputs and calculate the leakage or correlation. In their work on debiasing neural networks, Savani et al. (2020) used a fairness metric based on the difference in outputs predicted by neural networks for different demographic groups. These include *Statistical Parity Difference (SPD)*, *Equal opportunity difference (EOD)* and *Average Odds Difference (AOD)*. True and false positive rates, standard metrics that measure the accuracy of a model's output against the provided evaluation labels, are used to calculate the probability of positive outcomes (predictions or labels) for protected and unprotected groups. SPD measures the difference between the probability outcomes while EOD and AOD look specifically at the differences in true positive rates. Together, these metrics quantify prediction accuracy specifically focusing on protected and unprotected groups. Again, in common with other metrics, this is dependent on the quality of the reference annotations in the evaluation dataset. The metrics discussed here are derived from statistics and may work for any type of model but the experiments were only done on CNN-based classifiers with a focus on classification accuracy. Therefore these metrics may not be suitable for complex models.

Multimodal Models Apart from considering data as strictly 'text' or 'visual', there are emerging applications using multimodal or visual-linguistic models. Work on measuring gender bias in VL-BERT (Su et al. 2019), a visual-linguistic model, was conducted by (Srinivasan et al. 2021). The researchers measured associations between the gender of an agent (*man*, *woman*, *person*) and objects that have a stereotypical association with either male or female gender, such as *briefcase* vs. *purse*. For this purpose, they adapted Kurita et al. (2019) 's method for measuring associations in large language models (LLMs) to the multi-modal setting, analysing the influence of visual-linguistic pre-training, as well as both single-modality contexts. Srinivasan et al. (2021) found that visual-linguistic pre-training of VL-BERT shifts associations of the queried objects towards men. Moreover, the presence of a gendered agent in an image made the model more confident in predicting the object to be one that has a stereotypical association with the agent's gender, even in the presence of contrary visual evidence.

Similarly, Hendricks et al. (2018) found that stereotypical associations with objects, such as the one between men and computers, influence the generation of captions even if there is contrary visual evidence (i.e. a woman sitting at a computer). These metrics performed posthoc analysis which is analysing the final results. Although this is a good way of detecting and measuring bias, multimodal models are multistage and employ multiple constituent models. This makes it difficult to isolate and perform low-level bias analytics using the aforementioned metrics.

Most diffusion models make use of a visual-linguistic model like CLIP (Radford et al. 2021) which generates embeddings for the image-generating diffusion process. Such models, such as CLIP and ALBEF (J. Li et al. 2021), were analysed for gender bias by (Zhao et al. 2018), who developed *vision-language bias score (vlbs)* and *idealized vision-language ability score (ivlas)* to measure stereotypical associations in pre-trained vision-language models. *Vlbs* refers to the percentage of stereotypical predictions by a model for anti-stereotypical images. *Ivlas* is a combination of *vlrs*, which refers to the percentage of times a model ranks a stereotypical or anti-

stereotypical caption higher than an irrelevant caption and *vlbs*. The authors used many vision language models in their study, such as CLIP, ALBEF, ViLT, and VisualBERT. Their experiments revealed that ALBEF has the least amount of bias and CLIP the highest.

Gender Bias in Text-to-Image Diffusion Models

Although a type of model-centric gender bias, this type of bias is a major focus of this thesis, hence the relevant literature is discussed separately.

Bias is often better interpreted and understood through the lens of intersectionality. Luccioni et al. (2024) studied the presence of intersectional gender bias in TTI models by evaluating their output using image captioning models and creating clusters based on visual features. Their tool **StableBias** also allowed for visual analysis of the outputs. They used prompts which included multiple identities such as occupation, ethnicity, and gender. Their tool allows for exploratory analysis of the output of TTI models but does not allow for quantitative measurement of bias, especially in the representation space. To generate a diverse dataset based on social characteristics, they used a pattern *Photo portrait of [X][Y]* with X and Y being characteristics related to ethnicity/gender and professional attributes. This dataset was used to evaluate three TTI models: DALL-E2 and Stable Diffusion v1.4 and 2. Then three types of analysis were done. In the first set, they performed an analysis of the text features of the generated images using two Image-to-Text models: a ViT-GPT-2 model trained on MS COCO and a VQA system based on BLIP. They analysed the text features for gender and ethnic markers for professional attributes and compared them with data sourced from labour bureaus. The analysis revealed DALL-E 2 has the highest deviation from the real-world data followed by Stable Diffusion v2 and v1.4. The second analysis involved clustering visual features extracted from the images using the same BLIP VQA used before. The results indicated that men made up most of the professional clusters. The third analysis involved creating an interactive tool to study these biases on a case-by-case basis.

Similarly, **TIBET** proposed by Chinchure et al. (2023), measures bias along multiple axes such as physical appearance, ableism, gender, religion, and race. They used a VQA similar to Luccioni et al. (2024) to extract concepts from images generated by TTI models. They used two types of concept extraction: one explicitly by including the gender and bias in the question asked to the VQA and another implicitly by asking the VQA to describe the image. Then, the authors used two metrics to measure bias: *Concept Association Score* which quantifies the change in the generated image when the input is changed along a bias axis and *Bias Axis variance* which measures the degree of bias along an axis using counterfactuals.

A more intersectional approach is taken by Cho et al. (2023) where gender and skin tone are evaluated in images generated by popular TTI models using both human and automated methods. They found Stable Diffusion to generate more images of a specific skin tone or gender than DALL-E. The authors used exploratory analysis of the outputs and only considered binary gender.

Vice et al. (2023) uses three metrics for quantifying bias in TTI models: *Distribution bias* which measures the distribution of bias in the TTI generated output, and their novel metrics; *Jaccard Hallucination* which measures the correlation of hallucinations and bias; and *Generative Miss Rate* which measures how bias affects model performance. They injected bias using a backdoor - which is embedded into a model to maliciously change its output to evaluate the effectiveness of their metrics. *Distribution bias* is similar to the popular Area Under Curve (AUC) which is used to determine how correctly a model can differentiate between classes. It measures the distribution of images generated by a TTI model belonging to different classes. A biased model is more likely to generate images belonging to a certain class over others. The second metric *Jaccard Hallucination* uses the Jaccard Index to measure the effect of hallucinations on bias in TTI models. The authors consider two types of hallucinations: 1) when the generated image contains an object which is not specified in the prompt and 2) when the image does not contain a specified object. The metric measures bias introduced due to these hallucinations. The third metric

Generative Miss Rate evaluates how model performance changes with bias. The authors hypothesise that a model's performance should decrease with an increase in bias and the metric measures how much the generated image correlates with the input prompt.

2.2.2 Geographical Bias in Computer Vision

Mehrabi et al. (2021) reviewed the types of biases present in AI systems including historical, representational, evaluation and sampling bias among many others. Most of these biases manifest themselves as gender and racial bias. Most of the research focuses on two aspects: skin tone and gender.

L. Elisa Celis et al. (2020b) studied the presence of social bias in image retrieval and found that search queries related to people tend to return images that reflect the social, cultural and demographic attributes of the region. This leads to an over-representation of the demographic of the region from where the images are being queried.

Karkkainen et al. (2021a) analysed various popular visual datasets. Their study included Labeled Faces in the Wild, CelebA, COCO, IMDB-WIKI, VGG2, DIF and UTK datasets. Their findings show that the majority of the faces in the datasets, from ∼45% in the UTK dataset to ∼88% in the LFWA dataset, are white. Most of these datasets are either created or drawn from sources in western countries (Cao et al. 2018; Karras et al. n.d.; Learned-Miller et al. 2016; Lin et al. 2014; Russakovsky et al. 2015). The sources include web scraping using search engines, social networking sites, news and other media. As a result, they have a high representation of attributes associated with Western societies such as faces with lighter skin tone and western clothing, leaving the datasets heavily biased with an under-representation of non-western regions such as Africa or West Asia.

Shankar et al. (2017) found that popular visual datasets such as OpenImages and ImageNet contained most of the images from Europe and North America. They found that classifiers trained on these datasets misclassified images from coun-

tries like India at a much higher rate than those from the USA. A. Wang, Narayanan, et al. (2020) studied the geographical distribution of images in OpenImages and ImageNet and found them to be Europe and North America-centric, with the USA being highly over-represented and Africa being severely under-represented. When these datasets are used to train deep learning models such biases can be propagated within the learning models and amplified within AI systems (L. Elisa Celis et al. 2020b). Buolamwini et al. (2018b) found most facial recognition systems to be the most accurate for images of men with lighter skin tones and the least accurate for women with dark skin tones.

Most previous research focusing on ethnic and cultural bias used race as the main bias axis. However, this is a subjective parameter and is difficult to quantify. L. Elisa Celis et al. (2020b) and Shankar et al. (2017) focused on the issue of geo-diversity did identify issues with using search engines for dataset curation. However, they selected particular countries for their evaluation and did not provide a framework that can work on a global scale. Their recommendations also lacked a way to quantify and measure diversity in the images retrieved from search engines. Further, this issue of lack of geo-diversity can lead to a specific bias which was also not explained or investigated thoroughly.

2.3 Conclusion

Section I presented a discussion of the advancements in deep learning for computer vision. Starting from simple CNNs of a few layers used for image classification, computer vision models have become increasingly complicated and capable with the latest TTI diffusion models capable of generating images and multimodal models handling images and text. Section II discusses the types and causes of gender and other social biases in computer vision and the metrics developed to detect and measure them. Many of these metrics were developed for simpler models such as CNN-based image classifiers. This makes them unsuitable for newer computer vision architectures such as vision transformers and multimodal models. The main research

gaps identified are as follows:

1. Most of the metrics are designed for CNN-based image classifiers making them unsuitable for newer computer vision architectures such as vision transformers.

2. The metrics designed for multimodal models perform posthoc bias analysis making it difficult to isolate bias and perform low-level bias analytics.

3. All the bias analytics methods except for StableBias used binary gender i.e., man and woman.

4. Bias introduced in visual datasets curated using search engines due to localisation is not investigated thoroughly and no metrics were proposed to measure geo-diversity in images.

This thesis addresses these research gaps.

Chapter 3

Geographical Bias in Visual Datasets

This chapter presents a study of a source of bias in computer vision models, namely the datasets and their creation process. This novel bias, *Geographical Bias*, is a type of selection bias that is caused by the use of search engines for dataset creation and leads to ethnic, cultural and stereotypical bias in visual datasets. Two new metrics are proposed to measure this bias: ISS_{cross} and ISS_{intra}. This research proposes new methods to mitigate the effect of this bias and using the aforementioned metrics, and it is observed that these method produces a dataset that is more visually diverse than other popularly used visual datasets.

Section 3.1 outlines the motivation for the research, the research questions and the hypothesis. Section 3.2, defines *geographical bias*. Section 3.3 details the creation of the test dataset while section 3.4 describes how image similarity is calculated and defines the resulting image similarity scores. then the scores are used to calculate visual diversity in various popular visual datasets and compare them with the proposed dataset in section 3.5. Sections 3.6 and 3.7 discuss the findings and conclusion.

This research was published at the 1st International Workshop on Trustworthy AI for Multimedia Computing (Trustworth AI '21), co-located with the ACM

Multimedia Conference 2021.

3.1 Motivation

As discussed in Section 2.2.2 (page 57), auditing of social bias in visual datasets for faces has relied primarily on two main parameters: race (focusing on skin tone), and gender (Buolamwini et al. 2018b; L. Elisa Celis et al. 2020b; Karkkainen et al. 2021a; A. Wang, Narayanan, et al. 2020). However, this narrows down the scope and definition of bias and leaves out ethnic and cultural factors. Hence, this issue was approached from a different perspective – geography. Geography is also less subjective than race. A set of circumstances is outlined that leads to a geographical bias, which mimics a combination of societal biases including selection, cultural, confirmation and implicit stereotype bias. Issues with techniques used in image retrieval are identified and demonstrate how this contributes to bias in datasets. Then novel methods to evaluate bias and increase levels of diversity in datasets are proposed.

Creation of labelled and curated datasets is a resource-intensive process (Jia et al. 2021; Krizhevsky et al. 2012). A common solution is to leverage search engines to build collections of images related to a particular class based on a search string. This can enable the creation of very large datasets with associated labels with minimal user effort. Many popular visual datasets such as MS COCO (Lin et al. 2014), Pascal VOC (Vicente et al. 2014), and ImageNet (Krizhevsky et al. 2012).

Most internet search engines return localised results with respect to the language of a particular query and the IP-address-derived location from which it was queried. In this research, therefore, the focus is on these factors and examine how they affect levels of diversity within the resulting datasets. The effects of language and location on search results is illustrated in Figures 3.1 (page 63) and 3.2 (page 63), where considerable differences in relation to cultural, racial and gender-based attributes in search results were returned when a query is expressed in different languages and searched from different geographical locations. In examining the extent to which

query language and localisation techniques of search engines may result in localised biases in image datasets and how this may be evaluated and mitigated, this chapter addresses the following question:

RQ1: How does altering the location and language when using Internet search to collect a dataset affect the visual diversity of faces, and how can this be measured?

The main contributions of this chapter are:

- Audited social bias from a novel perspective: geography and introduced geographical bias.
- Developed novel metrics to measure geographical bias and a framework to mitigate it.

3.1.1 Search Engine Localisation as Cause of Bias

Search engines such as Google and Bing aim to provide search results that are relevant to the geographical location according to query settings or derived from the originating IP address. Retrieving images in this way returns images with local characteristics. Although, a good technique for providing users with locally relevant results, creating datasets in this way can limit the resulting diversity. Researchers have attempted to address this issue by adding location-based keywords along with the queries. For example, datasets used query terms such as "Asian Boy" for image retrieval but this can lead to stereotypical bias in the results with images being mostly of East Asian people with little or no representation of people from other regions of Asia (Karkkainen et al. 2021a; Russakovsky et al. 2015).

Given that search engines index images based on keywords from web pages, queries in English will return images from an English-language website (L. Elisa Celis et al. 2020b). For example, for a query such as 'Farmers in Thailand', the images returned are mainly from English-language websites. Search results for 'Egypt' and 'Paris' similarly return images of pyramids and the Eiffel Tower. The images reflect

concepts garnered from websites containing content in English and are thus more likely to reflect a Western viewpoint. Websites written in the main language of a particular country are excluded when the queries are not in the local language (A. Wang, Narayanan, et al. 2020).

Attempts to mitigate the over-reliance on English language queries when creating datasets have been made by translating search queries into different languages. ImageNet, for example, translated the queries into Chinese, Spanish, Dutch and Italian, along with English. However, apart from Chinese, all other languages are European and thus, have still have a Western-centric bias (Russakovsky et al. 2015; A. Wang, Liu, et al. 2022).

Figure 3.1: Search results for 'CEO'. Arabic-UAE (top), English-UK (bottom)

Figure 3.2: Search results for 'Politician'. Hindi-India (top), English-USA (bottom)

3.2 Defining 'Geographical Bias'

A combination of the localisation and personalisation of search engine results along with the issue of implicit bias in machine learning algorithms, gives rise to a kind of bias that have been defined as geographical bias. In defining the concept of

geographical bias, we build on ideas from social identity theory and social psychology and apply them in the context of information retrieval. *Geographical bias* is defined as a type of selection bias that is at the intersection of cultural bias, confirmation bias and implicit stereotype bias (Chapter 2 (page 42)). The following outlines contributing factors and impacts of geographical bias and how they are caused by localisation and personalisation in online image search:

Contributing Factor - Selection bias: Localisation and personalisation features of a search engine along with the language of the query can result in a biased search query (L Elisa Celis et al. 2020a). This research identifies selection bias in the search query as a central cause of bias in many image datasets.

Outcome - Cultural bias: The outcome of such selection bias is limited variation in the language of the search terms along with localisation and personalisation of the search engine results in biased image datasets.

Societal Impact - Confirmation and implicit stereotype bias: Due to localisation and personalisation features of the search engines, the images retrieved confirm or support prior concepts and beliefs concerning the local population. For example, the stereotypical concept of a CEO being a 'white middle-aged male wearing a business suit' in Western society could be reinforced. All the above factors contribute to a bias learned through repetition.

3.3 Creating the Dataset

To test this hypothesis (section 3.2), a dataset of human faces from a wide geographical range was collected. To include as much variation as possible, the world was divided into nine regions, each having a language (the *lingua franca* of that particular region) and one or two countries from the region. The countries chosen were generally the more populous countries of that region. The choice of regions was inspired by the racial groupings by Karkkainen et al. (2021a) and based on the United Nations regional grouping of countries [1]. Further categories are added to in-

[1] `https://www.un.org/dgacm/en/content/regional-groups` accessed: 11-10-2024

crease cultural diversity. Asia-Pacific group is split into East, South, and Southeast Asia, Africa into Sub-saharan Africa and North Africa (which is merged with West Asia), and Western European and other States divided into Western Europe and North America.

3.3.1 Language-Location Pair and Query Selection

The world was divided into the following nine regions: East Asia (EA), South Asia (SA), South East Asia (SEA), North Africa and West Asia (NAWA), Sub-Saharan Africa (SSA), West Europe (WE), East Europe (EE), North America (NA) and Latin America (LA). The most widely spoken language in each region was selected as the query language and the most or second most populous country in that region as the querying location resulting in nine sets of language-location pairs which are provided in Table 3.1. Note that location here refers to the region and not the individual country.

Five queries were selected as query terms: 'CEO', 'Engineer', 'Politician', 'Nurse' and 'School Teacher'. All these queries represent identities based on occupation. 'CEO' and 'Engineer' are traditionally considered male-dominated fields whereas 'Nurse' and 'School Teacher' are female-dominated. 'Politician' is considered globally more gender-balanced. The dataset was annotated based on the query which yielded the image. No human or automated annotation was used. The queries were translated into different languages using Google Translate, as per the language-location pairs. See [2] for the full list of language-location pairs and the corresponding queries in that language.

[2] `https://github.com/aibhishek/Geographical_Bias/blob/main/Query_Terms.pdf` accessed: 11-10-2024

Table 3.1: Regions and languages (abbreviations) used for creating the image dataset

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

3.3.2 Image Collection and Face Cropping

Google advanced image search [3] was used for querying the images and used a Virtual Private Network (VPN) (ExpressVPN [4]) as well as the 'region' option in image search to specify the origin country for each query. Each language-location pair corresponding to one region has five queries in that particular language. For each query, 150 images were scraped. For language-location pairs with two countries, 75 images were queried from each. For queries in Spanish where the noun changes with gender, each form of the noun was used and included the search with 'any of these words' feature. For example, search results for 'politician' would include results for either 'politica' or 'politico'. The 150 images were the top results returned by Google's algorithm.

To detect the faces in the images and crop them, a face detection and cropping algorithm called Autocrop[5] was employed. 20 faces for each query were selected. A total of 900 faces were collected for all the regions. This collection of 20 faces, corresponding to one query for one language-location pair, is referred to as an image

[3] `https://www.google.com/advanced_image_search` accessed: 11-10-2024
[4] `https://www.expressvpn.com/` accessed: 11-10-2024
[5] `https://github.com/leblancfg/autocrop` accessed: 5-7-2024

set.

Each face image has a dimension of 650x500 pixels, with the face covering 80% of the entire image. The resolution was determined experimentally. The background, which covered the remaining 20% of the image, along with attributes such as any visible clothing and headwear were retained. Headwear, clothing and background constitute important cultural aspects of a person's identity. As the aim was to reduce stereotypical bias and increase diversity, it was important to include these details. Also, certain professions have uniforms that may include headwear such as engineer's safety hat and nurse's head covering and mask. Finally, two images from each language-location pair were randomly selected to create a *Diverse Dataset* that was used as a benchmark against which to compare the diversity of other popular datasets. The dataset is available at `10.6084/m9.figshare.27263091`.

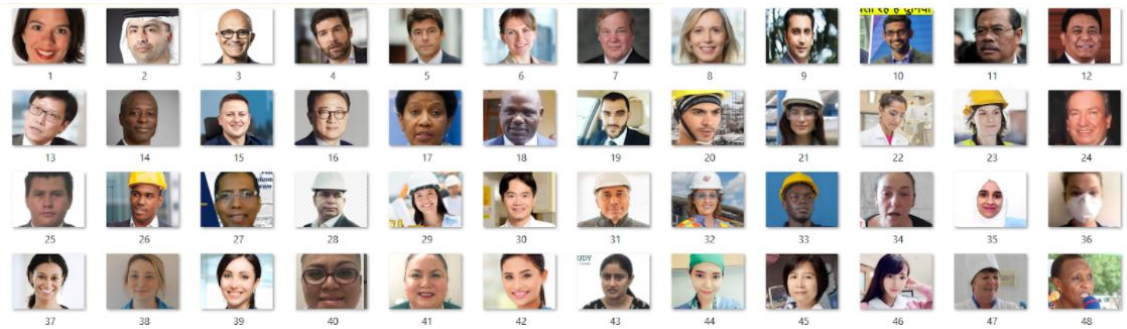


Figure 3.3: Sample of cropped faces.

3.4 Image Similarity

Image similarity scores were calculated following an approach used by L. Elisa Celis et al. (2020). Image Similarity Score (ISS) compares how similar two images are based on the features extracted by a pre-trained convolutional neural network (CNN). VGG16, a 16-layer deep CNN, trained on the Imagenet dataset, created by the Visual Geometry Group at the University of Oxford was used. The feature-extracting layers of the VGG16 were used to extract the features from the image. Dimensions of the extracted features were reduced using principal component anal-

ysis. For two images I_1 and I_2, with extracted features ν_1 and ν_2 respectively, image similarity is calculate as:

$$sim(I_1, I_2) = 1 - \frac{\nu_1 \cdot \nu_2}{||\nu_1||_2 \cdot ||\nu_2||_2} \tag{3.1}$$

$$sim(I_1, I_2) \in [0, 2]$$

The image similarity score is 1 - cosine similarity of the two feature vectors. As such, for two copies of the same image, the cosine similarity will be 1 ($\theta = 0$), and therefore the image similarity score is 0. For two decorrelated images, the value of cosine similarity will be 0 ($\theta = \frac{\pi}{2}$), and therefore the value of the image similarity score will be 1. Similarly, for two visually opposite images, the value of cosine similarity will be -1 ($\theta = \pi$), and therefore the value of the image similarity score will be 2. The higher the image similarity score, therefore, the more different the image is.

3.4.1 Visual Similarity

Visual similarity in this paper refers to how similar two images of people look (L. Elisa Celis et al. 2020b). Features such as eyes, lips and noses follow a pattern whereby the eyes, for example, are always above the nose. These patterns are picked up by Convolutional Neural Networks (CNN). However, they differ in terms of skin colour, facial structure and clothing and constitute a different set of features which vary as per region, ethnicity and gender (L. Elisa Celis et al. 2020b). These patterns are also picked up by CNNs. If training datasets consist of faces from a particular region, ethnicity or gender, the CNNs start to identify those social attributes with humans. And when those CNNs are used on faces having attributes different from those they were trained with, they fail to identify those faces correctly (Buolamwini et al. n.d.; L. Elisa Celis et al. 2020b; Karkkainen et al. 2021a; A. Wang, Narayanan, et al. 2020). Figure 3.4 shows a heatmap of similarity scores of images of people from

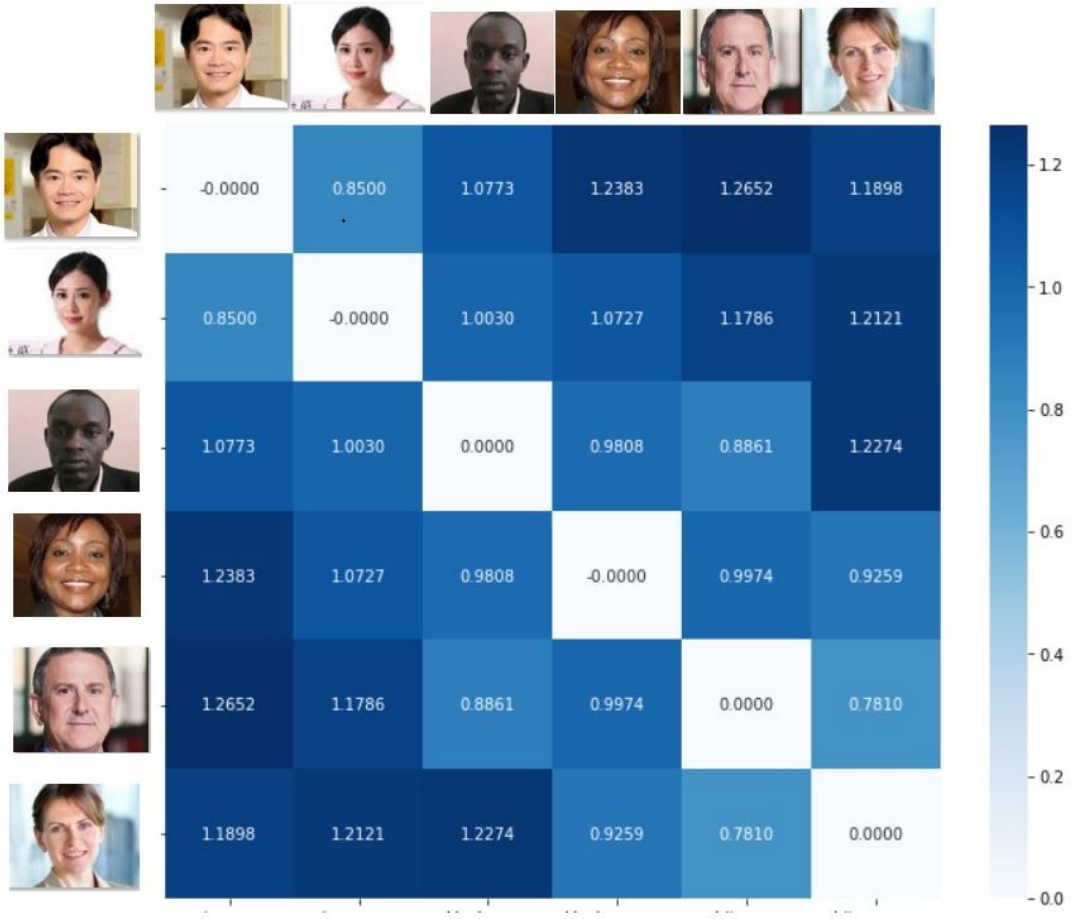


Figure 3.4: Heatmap of image similarity.

different regions. It can be observed that people from similar regions have a lower similarity score indicating that they have higher visual similarity. The similarity score of two identical images is zero.

3.4.2 Image Similarity Scores

To test the hypothesis, two variations of the image similarity scores: intra (Algorithm 1) and cross (Algorithm 2) were created. Image similarity *intra* is the mean of the image similarity score of every image with all other images in the image set. Image similarity *cross* is the mean of all of the image similarity scores between the images across all the given image sets. Image similarity intra will compare one image set while cross will assess multiple image sets. A high image similarity score indicates a more diverse dataset. For instance, if a dataset contains multiple copies

of the same image, the scores will be zero. If a dataset contains extremely visually diverse images which are very different from each other visually, then the scores will be near 2.

Algorithm 1 Algorithm for Calculating Image Similarity Score - Intra

Input: List of images (Image Set) S
for each $I \in S$ **do**
    convert I to array
    process the array and extract features
    reshape extracted feature array to 20x15x512
    apply PCA transform
    reshape to dimensions 1x400
    append reshaped feature vector array to list V
copy $V \rightarrow V'$
for each $F \in V$ **do**
    for each $F' \in V'$ **do**
        Calculate $sim(F, F')$
        append similarity score to list L
return mean(L)

3.5 Experiment

The image sets were arranged according to all the language-location pairs for each query term. Each query term (5 in total), consists of 9 language-location pairs. Each language-location pair consists of one image set i.e., a list of 20 images. Fig. 5 shows this arrangement for the query term 'CEO' and language-location pair 'Arabic-NAWA'. The language-location pairs for all the query terms remain the same[6].

Image similarity score – Intra ($\mathbf{ISS_{intra}}$) is calculated by taking the mean of $sim(I, I')$ of all the images in a single image set. The mean of $\text{ISS}_{\text{intra}}$ for all the language-location pairs for a given query gives $\text{ISS}_{\text{intra}}$ for that query term.

Image similarity score – Cross ($\mathbf{ISS_{cross}}$) is calculated by taking the mean of $sim(I, I')$ of each image in every image set with all the images of all other image sets for a particular query term. $\text{ISS}_{\text{intra}}$ and $\text{ISS}_{\text{cross}}$ were calculated for all the query

[6] https://github.com/aibhishek/Geographical_Bias/blob/main/Query_Terms.pdf

Algorithm 2 Algorithm for Calculating Image Similarity Score - Cross

Input: List of list of images (Image Set) S
for each $S' \in S$ **do**
    for each $I \in S'$ **do**
        extract features
        reshape extracted feature array to 20x15x512
        apply PCA transform
        reshape to dimensions 1x400
        append reshaped feature vector array to list V'
    add V' to stack V
while $V \neq \{\}$ **do**
    pop item from V and assign to A
    for each $A' \in V$ **do**
        for each $F \in A$ **do**
            for each $F' \in A'$ **do**
                Calculate $sim(F, F')$
                append similarity score to list L'
        append *mean(L')* to list L
return mean(L)

terms and tabulated the results.

3.5.1 Calculating Diversity in Visual Datasets

ISS_{intra} was used as a metric to measure diversity in some popular face datasets. For this, Flickr Faces HQ (FFHQ), WIKI, IMDB, Labelled Faces in the Wild (LFW), UTK Faces and the proposed Diverse Dataset were chosen. To calculate ISS_{intra}, 100 images were randomly sampled from each of the datasets (except Diverse Dataset, as it had only 90 images). Then calculated ISS_{intra}, was calculated and the results were tabulated (Table 3.2 and 3.3).

3.6 Findings and Discussion

The findings of this study demonstrate how varying the language and location of online image search queries can increase the diversity of the resulting visual dataset. In comparing the diversity of datasets in this study it was observed that the values of ISS_{cross} are higher than ISS_{intra} for all the query terms (see Table 3.2). Conse-

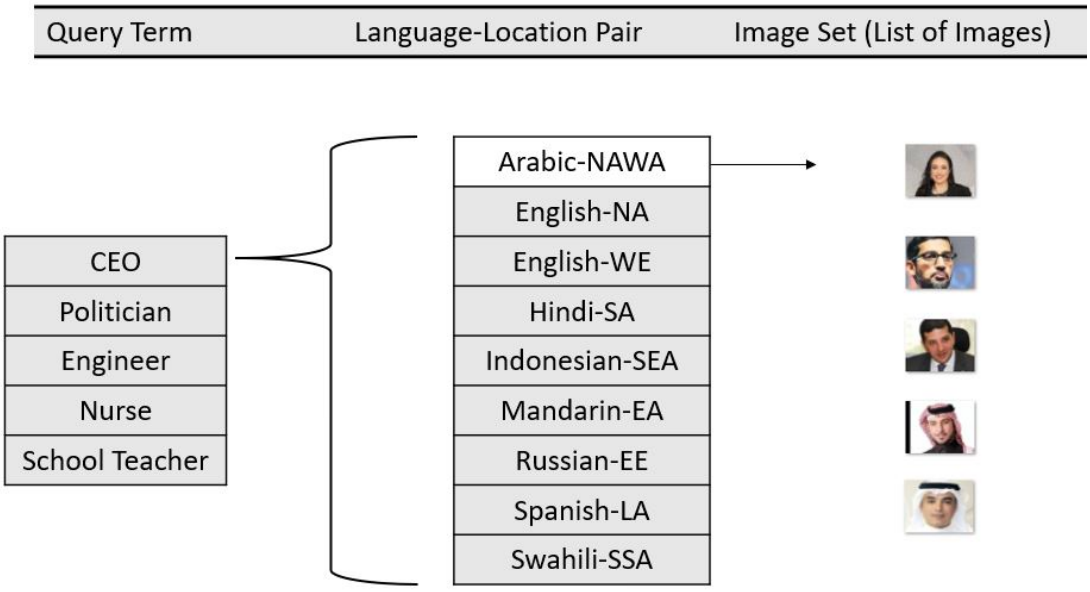


Figure 3.5: Arrangement of Images according to Query Term and Language-Location Pair.

Table 3.2: ISS_{intra} and ISS_{cross} scores for all the query terms. Higher values indicating more visual diversity have been highlighted in bold.

Query	**ISS_{intra}**	**ISS_{cross}**
CEO	0.9644	**0.9846**
Engineer	0.9925	**0.9939**
Nurse	0.9862	**0.9900**
Politician	0.9724	**0.9836**
School Teacher	0.9860	**0.9904**
Mean Value	0.9803	**0.9885**

quently, the mean value of ISS_{cross} is also higher. The ISS_{cross} values were higher than the individual ISS_{intra} scores for 66.66% of cases (30 out of 45 language-location pairs). Appendix A contains the individual scores and calculations for each query and language location pair. These quantitative measurements support the visual observation outlined in Section 3.1 (page 61) that the image search results obtained by varying the language and location of the search queries increase the visual diversity of the results. The higher ISS_{cross} scores demonstrate that images retrieved using the same query term but for different geographical locations are more visually diverse than those from the same geographical location. This points out that the localisation used by search engines affects image search results. Therefore by using diverse languages and geographical (IP) locations can lead to a more diverse dataset.

The quantitative results (Figure 3.6) show that $\text{ISS}_{\text{intra}}$ has a wider spread than $\text{ISS}_{\text{cross}}$, which are more concentrated around one value ($\sim$1). At $sim(I, I') = 1$, the images are decorrelated, or highly visually dissimilar. This means that the images are least correlated at this point.

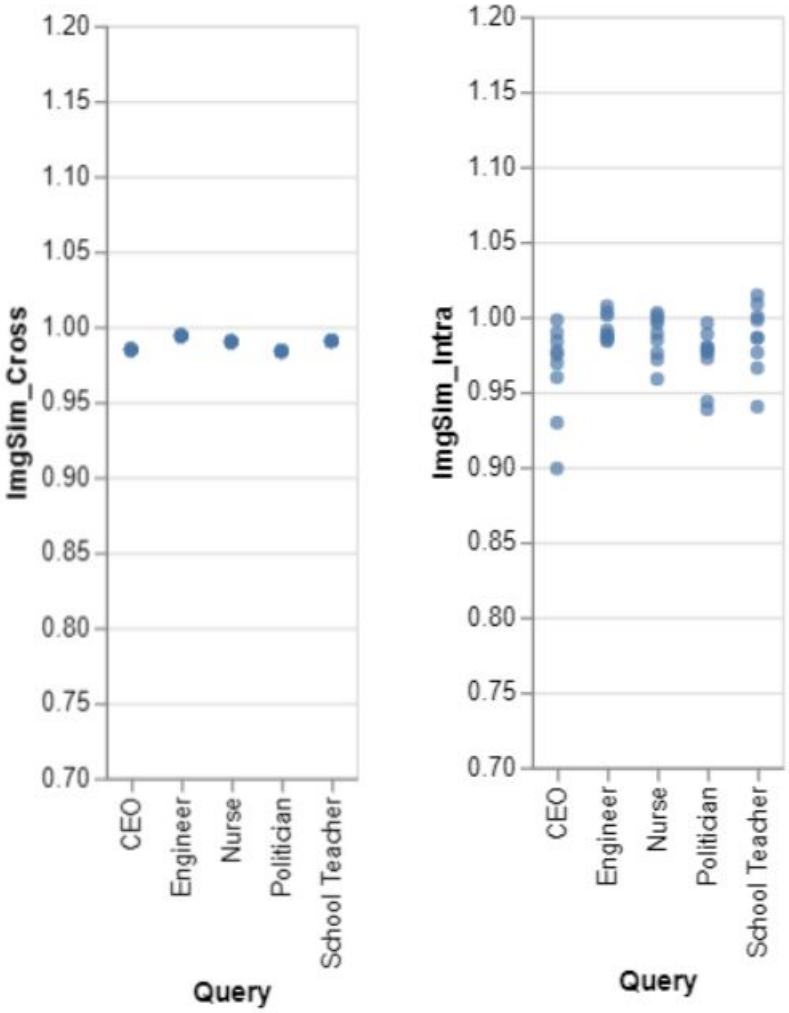


Figure 3.6: Distribution of $\text{ISS}_{\text{intra}}$ and $\text{ISS}_{\text{cross}}$ in the Diverse Dataset.

The 'Diverse Dataset' has higher $\text{ISS}_{\text{intra}}$ than all the other datasets except FFHQ (see Table 3.3 and Figure 3.7). Thus it can be concluded that using a diverse querying methodology (by varying the language and location of search queries) in image retrieval for creating visual datasets can lead to better visual diversity of datasets and help mitigate geographical bias inherited from the search systems. This also demonstrates that the proposed methodology of diversifying the query language and geographical (IP) locations when curating visual datasets can increase

Table 3.3: ISS_{intra} of Datasets.

Dataset	**ISS_{intra}**
FFHQ (Karras et al. n.d.)	0.9940
Diverse Dataset	**0.9895**
WIKI (Rothe et al. 2015)	0.9786
IMDB (Rothe et al. 2015)	0.9661
LFW (Learned-Miller et al. 2016)	0.9536
UTK (Z. Zhang et al. 2017)	0.9418

visual diversity in datasets and limit the presence of geographical bias.

3.7 Conclusion

This chapter introduces two novel metrics to detect and measure bias in visual datasets. A methodology for increasing the diversity of datasets and mitigating the effects of geographical bias by increasing the variations in the language of the query terms and the location of the search engines is proposed. It has been demonstrated how current methods of compiling visual datasets using online image retrieval can introduce a type of bias that is unique to this process and reflects a variety of social biases. To study and understand this bias more thoroughly, geographical bias was defined, based on the parameters that introduce it: language and location of search engine queries. This type of bias manifests in different forms, throughout the machine learning pipeline, as racial, cultural and stereotypical bias.

The experiments showed that altering the query language and the geographical (IP) location when using search engines for dataset curation can increase the visual diversity of the curated dataset. The metrics introduced in this chapter provide effective ways to measure this diversity.

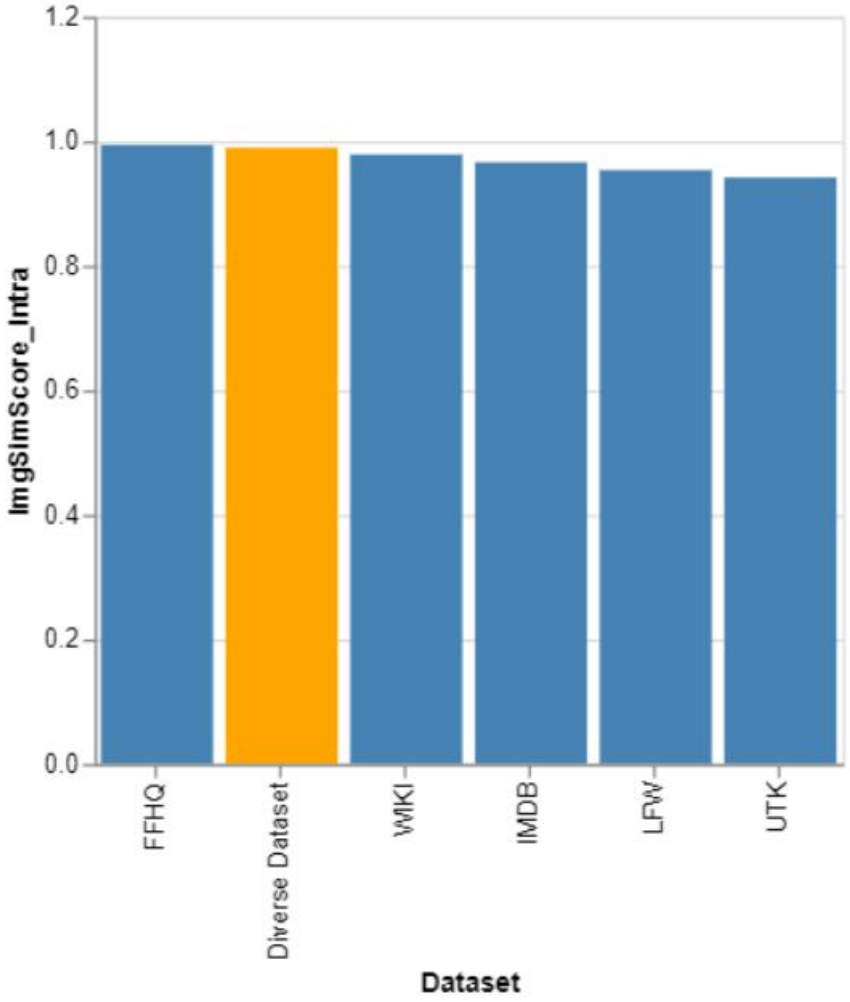


Figure 3.7: Distribution of ISS_{intra} of Datasets (Bar Chart).

Chapter 4

Geographical and Gender in Multimodal Vision Models

This chapter builds on the work done in the previous chapter and looks at the intersection of *Geographical* and *Gender Bias*. As discussed in Chapter 2 (page 36), geographical and gender bias are the two most prevalent biases in computer vision models. These two biases also intersect and amplify bias and prejudice against vulnerable groups. This chapter introduces a novel metric to measure bias at the intersection of geography and gender: *Gender Difference* and audit a large multimodal vision model: CLIP (Contrastive Language Image Pretraining) (Section 4.2.3). Evidence of significant gender bias with varying patterns across global regions was discovered. Harmful stereotypical associations were also uncovered related to visual cultural cues and labels such as terrorism. Levels of gender bias uncovered within CLIP for different regions aligned with global indices of societal gender equality, with those from the Global South reflecting the highest levels of gender bias.

Section 4.1 outlines the motivation for the research and introduces *Transnational Feminism* – a new school of feminist thought that seeks to create a more inclusive and diverse lens for auditing gender bias in a global context. Section 4.2, explains the methodology used for the experiments including dataset creation, keyword selection for image-text similarity calculation using CLIP, and Grad-CAM analysis for visual

explainability. Section 4.3 discusses the findings on how gender bias varies by region for three sets of keywords: *positive and negative terms, adjectives*, and *occupations*. Finally, Section 4.4 concludes the chapter.

The research that emanated from this work was published at the Aequitas 2023 Workshop on Fairness and Bias in AI, co-located with the 45$^{\text{th}}$ European Conference on Artificial Intelligence 2023, Kraków, Poland.

4.1 Motivation

Research on biases in computer vision evaluated the effect of skin tone and gender on facial recognition. For instance, Buolamwini et al. (2018) found that classifiers from Microsoft, Face++, and IBM contained intersectional biases with the highest accuracy levels on the faces of men with lighter skin and the worst on the faces of women with darker skin tones. Further to consideration of skin tone, facial features are multifaceted and contain diverse visual cues such as those related to culture and ethnicity as shown in Chapter 3. For instance, many people with fair skin but from different countries may differ in their appearances due to cultural norms in relation to clothing. De Vries et al. (2019) found that popular vision models often fail to detect and classify images from non-Western and developing countries. Drawing upon transnational feminism in this audit of CLIP addresses this issue, enabling the analysis of the effects of diversity with consideration to geographical region and culture. This chapter aims to address the following research question:

- **RQ2:** How does gender bias exhibited by large multimodal vision models differ by geography and how can such intersectional bias be measured?

The main contributions of this chapter are:

- Audited gender bias in CLIP from a transnational feminist lens.
- Developed novel metrics to measure the geographical variation of gender bias in CLIP.

4.1.1 Transnational Feminism

A transnational feminist perspective emphasises global differences in the dynamics of gender inequalities in society (Briggs 2016; Grewal et al. 1994; Henrich et al. 2010). This standpoint necessitates consideration of the perspectives and contextual experiences of inequality from different regions and cultures. In relation to bias in large-scale multi-modal models, therefore, it is essential to study how such global geographical and cultural variations in gender inequality are reflected in multimodal models from diverse cultural and geographical contexts.

4.1.2 Auditing Social Biases in CLIP

The authors of CLIP evaluated their own model and found evidence of social biases within it using datasets such as FairFace (Karkkainen et al. 2021b) and images of members of the US Congress. The racial classification within the FairFace dataset was compiled using the US Census with the addition of 'Southeast Asian' and 'Middle Eastern'. Approaches to defining race itself and racial categories have been critiqued for being founded upon a predominantly Western perspective (Keita et al. 2004; K. A. Kennedy 1995; R. F. Kennedy et al. 2013). The use of certain race labels such as 'Indian', for instance, can be problematic given that it refers to nationality rather than one distinct race or ethnicity [1]. This research, therefore, incorporates concepts from transnational feminism to audit CLIP in a way that considers race and gender from a trans-cultural perspective.

This chapter aims to answer the following research question:

RQ2: How does gender bias exhibited by large multimodal vision models differ by geography and how can such intersectional bias be measured?

[1] `https://www.indiacode.nic.in/handle/123456789/1522`

4.2 Methodology

To audit CLIP and understand how gender bias intersects with geographical region and culture, building on work done in Chapter 3, a dataset of images of men and women was created by crawling from various geographical locations across the world. This method of basing the data gathering process in different regions of the world allows for the representation of gender that is representable to those different regions through internet searches to be captured and aligns with the importance of considering the issue of bias from multiple perspectives. Then three sets of keywords denoting adjectives, occupations and negative and positive words were created. Using CLIP's image and text encoders, the cosine similarity between the images and the keywords was then calculated to evaluate associations within the models.

4.2.1 The Image Dataset

An image dataset using Google advanced image search was curated. The query terms were 'man' and 'woman' translated into different languages as per the location. Selenium was used to automate the image scrapping and a VPN [2] was used to change the IP geo-location with each search happening in a new incognito browser profile. Western Europe, Eastern Europe, North Africa and West Asia, Sub-Saharan Africa, South Asia, Southeast Asia, East Asia, North America and Latin America were used as geographical regions similar to what was done in Chapter 3. The languages for the query terms and the country for the VPN location are provided in Table 4.1 along with language and location pairs and corresponding abbreviations. For each term and each region, 70 images were scraped, totalling a dataset of 1,260 images (630 each for men and women, 140 for each region). The dataset is available at: `10.6084/m9.figshare.27642618`.

[2] `https://www.expressvpn.com/`accessed:11-10-2024

4.2.2 The Keywords

Three sets of keywords were used. The first set is based upon the bias analytics conducted by the developers of CLIP Radford et al. (2021) and consists of five positive (*trustworthy, educated, smart, confident, and achiever*) and five negative (*criminal, terrorist, gangster, drug addict, and fraud*) words. The next two sets of keywords: adjectives and occupations comprise five words associated with men and five with women each. For adjectives, the words *honorable, dissolute, arrogant, heroic, and boyish* are associated with men, and *romantic, submissive, elegant, caring, and delicate* are associated with women. In the case of occupations, *carpenter, mechanic, mason, architect, and mathematician* are male-dominated and *midwife, librarian, housekeeper, dancer, and teacher* are female-dominated (Garg et al. 2018). These sets of words are taken from Garg et al. (2018), and five words were randomly chosen from the list for each of the subcategories.

4.2.3 Image-Text Similarity

CLIP is a multimodal model that creates embeddings for text and images using text and image encoders, trained using contrastive learning to find the most similar image-text pairs (Radford et al. 2021). By calculating the cosine similarity of the image and text embeddings, patterns can be found that can point out bias in the CLIP embeddings. The similarity is calculated by adopting the approach developed by the authors of CLIP[3]. The images are first resized and then the pixel intensity normalised using internal CLIP functions. Then the features are extracted from the images. The image encoder used in this experiment is Vision Transformer ViT-L/32. Next, the keywords are prefixed with the sentence 'An image of ' and tokenized. Finally, the cosine similarity is calculated between the image and text features using CLIP functions.

[3] `https://colab.research.google.com/github/openai/clip/blob/master/notebooks/Interacting_with_CLIP.ipynb`

Table 4.1: Regions and languages (abbreviations) used for creating the image dataset

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

4.2.4 Visual Question Answering and Grad-CAM

A visual question-answering machine was built using CLIP that takes in an image and a text question (sentence) and answers the question based on the image. Then Gradient Weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al. 2016) was used to create a heatmap superimposed on the original image to highlight the region of the image that the model uses the most to answer the question.

4.3 Findings and Discussion

The image-text cosine similarity scores (Section 4.2.3 (page 80)) were calculated for the three sets of keywords: negative and positive traits, adjectives and occupations. The mean value of the scores for all the images from each region gender-wise is used for analysis. Grad-CAM analysis was also used for the negative and positive traits for further analysis. The findings are discussed in detail in the following subsections.

A summary of the trends in the scores is given in Table 4.2, where trend refers to the net positivity or negativity in the scores and is given as

Table 4.2: Total mean similarity scores. The individual scores reflect the sum of mean cosine similarity scores of the particular type of keyword and the images of men and women belonging to the particular regions. Trend = positive - negative. Gender Difference = abs(sum of scores for men - sum of scores for women). Gender refers to the perceived gender of the images. The standard deviation for all the scores was less than 0.015. For abbreviations, refer to Table 4.1. Masc: Masculine, Fem: Feminine.

	Gender	**Type**	WANA	EA	WE	NA	SA	SEA	EE	LA	SSA
Positive & Negative Words	Man	positive	0.90	0.92	0.93	0.93	0.89	0.90	0.92	0.96	0.91
		negative	0.98	0.92	0.94	0.94	0.94	0.95	0.94	1.00	0.97
		trend	-0.08	0.00	-0.01	-0.01	-0.05	-0.05	-0.02	-0.04	-0.06
	Woman	positive	0.96	0.93	0.90	0.95	0.95	0.95	0.91	0.97	0.90
		negative	1.00	0.93	0.90	0.95	0.97	1.00	0.91	1.00	0.95
		trend	-0.04	0.00	0.00	0.00	-0.02	-0.05	0.00	-0.03	-0.05
	Gender Difference		0.08	0.02	0.07	0.03	0.09	0.10	0.04	0.01	0.03
Adjectives	Man	Masc.	0.96	0.94	0.99	1.00	0.97	0.98	0.99	0.97	0.94
		Fem.	0.86	0.88	0.92	0.96	0.90	0.91	0.92	0.90	0.84
	Woman	Masc.	0.98	0.93	0.97	0.94	0.92	1.00	0.98	0.96	0.96
		Fem.	0.94	0.93	0.94	1.00	0.85	0.98	0.95	0.91	0.88
	Gender Difference		0.10	0.02	0.00	0.00	0.10	0.09	0.02	0.00	0.06
Occupation	Man	Male	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97
		Female	0.90	1.00	0.95	0.90	0.95	1.00	0.97	0.94	0.93
	Woman	Male	0.93	0.97	0.97	0.91	0.88	0.96	0.97	0.91	0.89
		Female	1.00	1.00	0.97	0.97	0.98	1.00	0.99	0.99	0.94
	Gender Difference		0.07	0.03	0.01	0.02	0.09	0.04	0.01	0.04	0.07

$$Trend = \sum P - \sum N \tag{4.1}$$

where P is the mean cosine similarity of the positive words and N is the mean cosine similarity of the negative words. Gender Difference is calculated as:

$$Gender\ Difference = \Big|\ \sum M - \sum W\ \Big| \tag{4.2}$$

where, M $\in$ mean cosine similarity for images of men and W $\in$ mean cosine similarity for images of women.

4.3.1 Negative and Positive Words

It can be seen from Table 4.2 that the mean cosine similarity scores are higher for all the images, but images of women generally have less negativity than men but with

Figure 4.1: Grad-CAM results for the question 'Who is the terrorist?' for images from all regions. Regions: Top-Bottom, L-R: WANA, WE, SA, SSA, EA, SEA, LA, EE. Same pattern for images of men and women.

Figure 4.2: Grad-CAM results for the question 'Who is the terrorist?' for images of women from West Asia and North Africa.

geographical differences. For images of women from Europe, North America, and East Asia, the trend results are zero (i.e. neutral). These regions generally comprise the 'Global North' and are generally wealthy, developed, and democratic (Dados et al. 2012). Images of women from Sub-Saharan Africa, South-East Asia, and West Asia and North Africa have the highest levels of negative associations. These regions generally comprise the 'Global South' and lag behind the Global North in wealth and development (Dados et al. 2012). The gender difference is highest for South Asia and West Asia and North Africa. These two regions also score the lowest in the Global Gender Gap Index (Ratcheva et al. 2022).

The gender difference for Sub-Saharan Africa is low, but this region also ranks low in the Global Gender Gap Index. Fig 4.3 shows the relationship between the

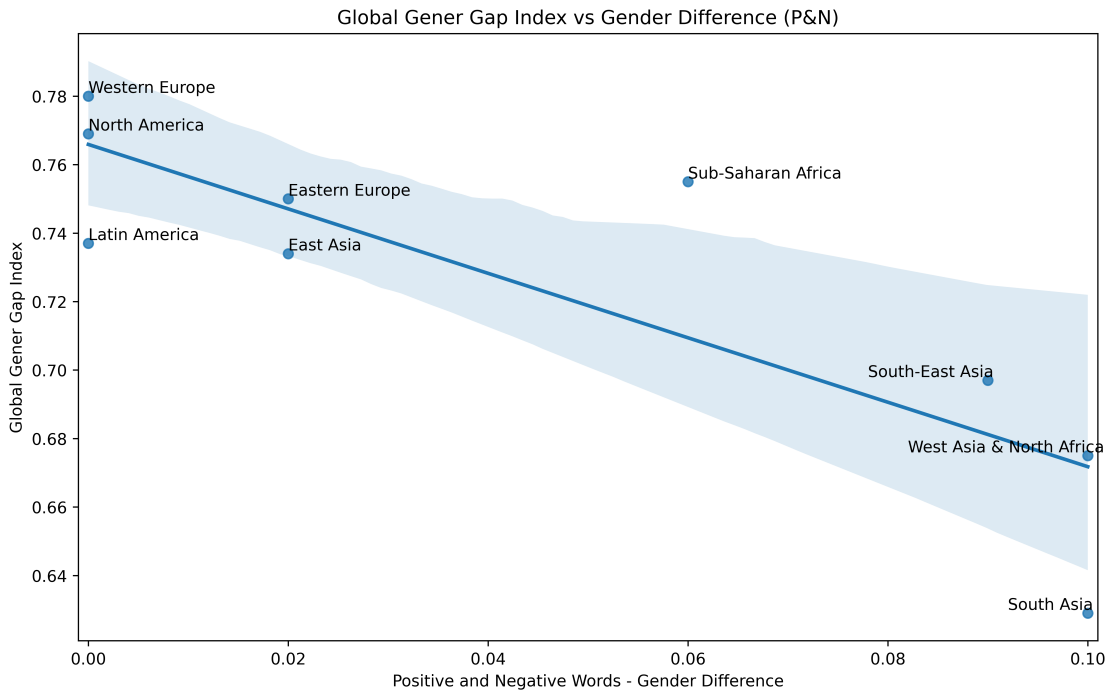


Figure 4.3: Global Gender Gap Index vs Gender Difference (Positive and Negative Words). r-value=-0.62, p-value=0.007.

Global Gender Gap Index and gender difference, demonstrating a strong relationship between the two scores. The regions with the highest Global Gender Gap Index, such as Europe, North America, and East Asia, tend to have the lowest gender difference. The Global Gender Gap Index used in this paper is for the country from where the images were scraped, as shown in Table 4.1. In the case of two countries, the average is used.

The similarity for the word 'terrorist' is highest for the images of women from South-East Asia and West Asia and North Africa (Appendix A.2). The predominant religion in these two regions also happens to be Islam[4]. Using Grad-CAM, it was observed that women from these regions have a higher chance of being assigned the label 'terrorist' (see Figure 4.1). On further analysis, it was discovered that images of women wearing *hijab (headscarf)* are more likely to be associated with the label 'terrorist'. In Figure 4.2, an image of two women from the same region (West Asia and North Africa), but with one wearing a hijab, was given to the

[4]`https://web.archive.org/web/20110209094904/http://www.pewforum.org/The-Future-of-the-Global-Muslim-Population.aspx`, Last accessed: June 2023

visual question answering machine with the text 'Who is the terrorist'. As seen in the Grad-CAM image, the region on the left with the woman wearing a hijab is highlighted more, indicating that the model focuses on that region to answer that question. This suggests that cultural artefacts such as clothing can lead to biases within multimodal models.

4.3.2 Adjectives

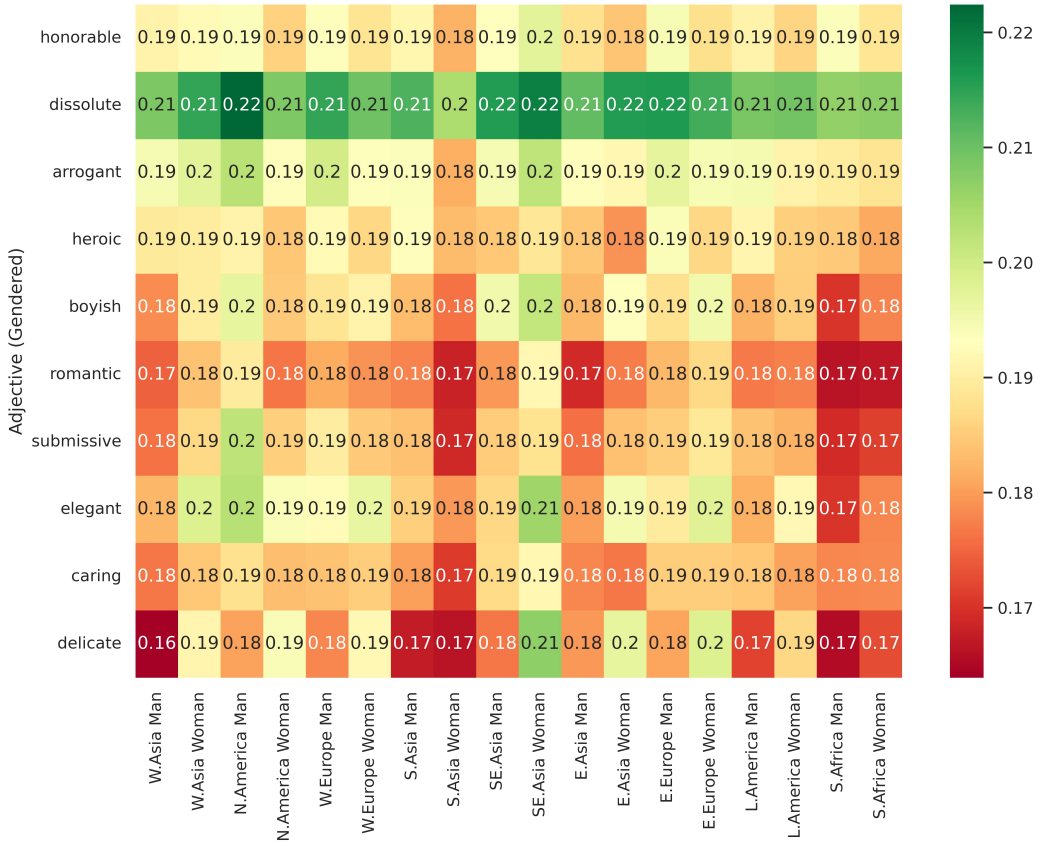


Figure 4.4: Adjectives vs region and gender - mean cosine similarity scores heatmap

The trend of cosine similarity scores for adjectives indicate potential stereotypical gender bias for men and women. The masculine adjectives have a higher similarity with images of men, and the feminine adjectives have a higher similarity with the images of women. Figure 4.4 shows the mean cosine similarity of the keywords by region. Images of women from East and South-East Asia have higher similarity for the terms 'caring', 'elegant', and 'delicate'. This may reflect a Western bias which considers Asian women as more 'feminine' (Ciurria 2019). The gender difference

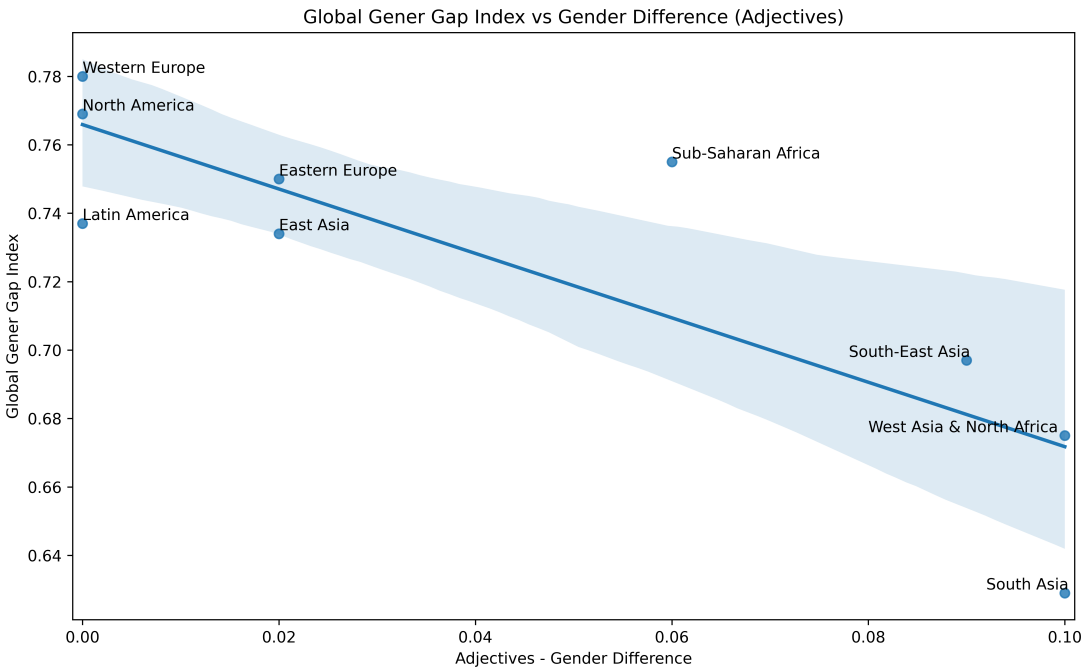


Figure 4.5: Global Gender Gap Index vs Gender Difference (Adjectives). r-value=-0.84, p-value=0.003.

scores are the lowest for Europe, North America and East Asia. These regions tend to be developed and wealthier and score better in the Global Gender Gap Index (Dados et al. 2012). West Asia and North Africa, and South Asia have the highest gender difference and perform worse in the Global Gender Gap Index [5]. Fig 4.5 shows the relationship between the Global Gender Gap Index and gender difference, and a strong relationship is seen between the two scores.

4.3.3 Occupations

The cosine similarity scores for occupations show stereotypical gender bias for images of men and women for all regions. A heatmap of the similarity scores is given in Fig 4.6. Traditionally male-dominated occupations such as 'mechanic', 'architect', and 'mathematician' have higher similarity scores for men, while traditionally female-dominated occupations such as 'midwife', 'housekeeper', and 'librarian' have higher similarity scores for women. Images of women from South, East, and South-East Asia had the highest similarity with occupations such as 'midwife', 'housekeeper',

[5] `https://www3.weforum.org/docs/WEF_GGGR_2022.pdf`

and 'librarian'. Images of women from Europe and North America have lower similarity for traditionally female-dominated occupations such as 'midwife' but higher similarity for traditionally male-dominated occupations such as 'architect'. The gender difference scores show a similar trend as seen earlier; Europe and North America show the least gender difference and are the regions with the best Global Gender Gap Index. Fig 4.7 shows the relationship between the Global Gender Gap Index and gender difference, and also reflects a strong relationship between the two scores.

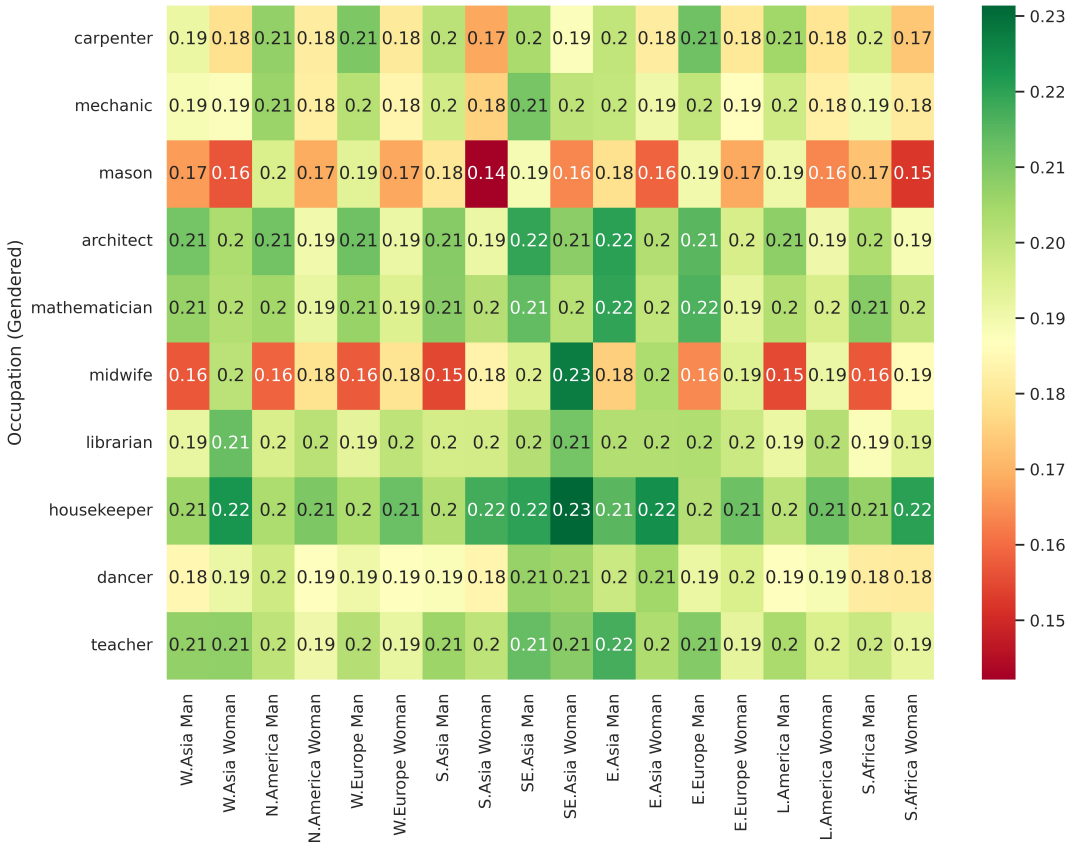


Figure 4.6: Occupations vs region and gender – mean cosine similarity scores heatmap

4.4 Conclusion

Gender bias is a complex, multifaceted, and multidimensional issue comprising various dimensions such as race, ethnicity, culture, and geography. Thus it is difficult to analyse the issue using a singular theoretical lens or theories primarily developed in the Western world. Transnational feminism places importance on the analysis of

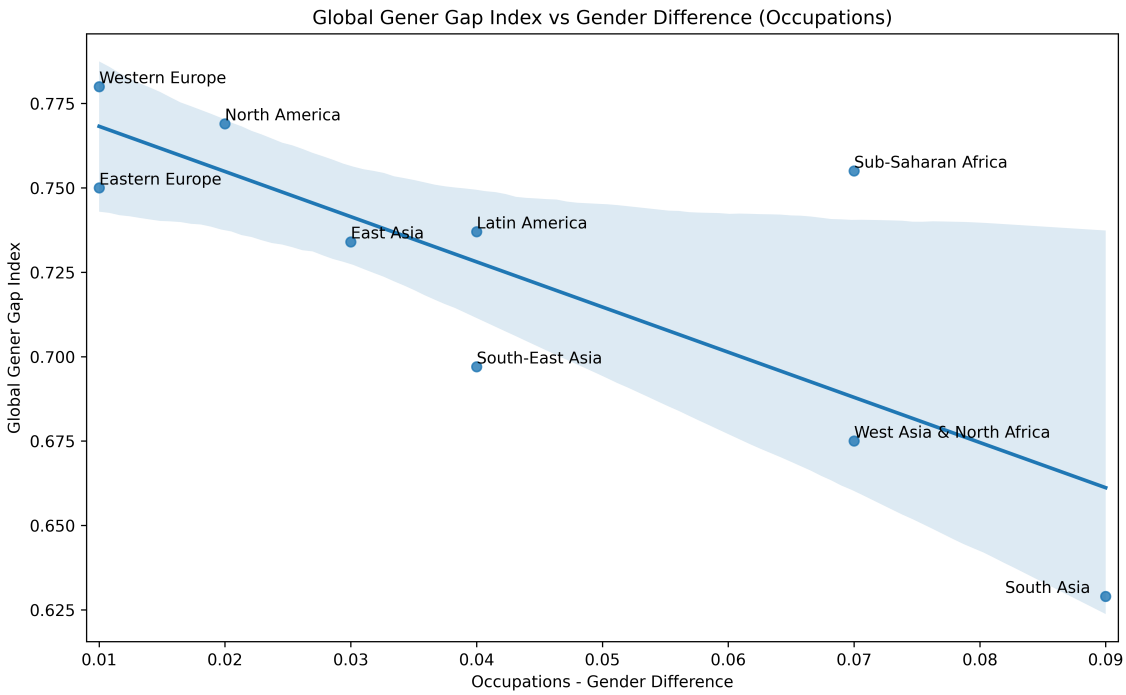


Figure 4.7: Global Gender Gap Index vs Gender Difference (Occupations). r-value=-0.78, p-value=0.0012.

the issue of gender bias from a more inclusive lens, accommodating diverse global perspectives such as globalisation, income inequality, and the economic and digital divide between the global north and south among other contemporary issues. In incorporating this perspective in this research, significant evidence of gender bias in CLIP was uncovered with differences in how such bias manifests regionally and culturally. Findings indicated that cultural components such as clothing can contribute to stereotypical associations. A strong correlation was also evident between the Global Gender Gap Index and gender difference scores, with Europe, North America, and East Asia scoring high on both the indices and South Asia, and West Asia and North Africa performing the worst. This may be related to levels of gender equality in society influencing the representation of gender within internet content from those regions, affecting levels of gender bias in training data. CLIP is also trained on data primarily curated from the English internet and biases exhibited are those inherited from it and this may explain the association of 'hijab' with 'terrorism' as has been explored in earlier research (Birhane et al. 2021; De Vries et al. 2019). The experiments indicate that gender bias in CLIP varies considerably by

geographical region.

Chapter 5

Measuring Gender Bias in Multimodal Vision Models using NLP Techniques

This chapter introduces a cross-domain adaptation of metrics to detect and measure gender bias in multimodal computer vision models. As discussed in Chapter 2 (page 36), Contrastive Language Image Pretraining (CLIP) is one of the most popular multimodal models connecting images and text and provides unparalleled multimodal capabilities and is used in various state-of-the-art image generation models such as DALL-E and Stable Diffusion.

Earlier methods for detecting and measuring gender bias in computer vision models were mostly confined to classification metrics such as accuracy (Buolamwini et al. 2018b), distance (A. Wang, Liu, et al. 2022), and benchmark datasets (Buolamwini et al. 2018b; A. Wang, Liu, et al. 2022). However, with the development of complex multimodal models such as CLIP, it is necessary to develop metrics that can quantify complex social concepts such as gender. To more effectively measure gender bias in vision models a popular metric for measuring stereotypical associations in language models called *Word Embeddings Association Test* developed by Caliskan, Bryson, et al. (2017) and based on the hugely popular *Implicit Association*

Test used in psychology was adapted.

Section 5.1 explains the two main motivations behind this research: the effect of web-scale training on bias and developing effective metrics for bias quantification in vision models. Section 5.2, explains the methodology for creating the test dataset, the WEAT scores and tests and Grad-CAM for explainability. Section 5.3 covers the findings and discussions and section 5.4 concludes the chapter and the limitations are discussed.

The research that emanated from this work was published at the 25$^{\text{th}}$ International Conference on Multimodal Interaction 2023, Paris, France.

5.1 Motivation

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 (Karkkainen et al. 2021b). To address this issue to some extent, diverse and de-biased datasets were created (Karkkainen et al. 2021b). However, deep learning models are now being trained on increasingly larger datasets with models now reaching billions of parameters (Wiggers 2021). 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.

This chapter presents 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 (Radford et al. 2021). CLIP is not trained on curated datasets but on data directly taken from the internet (Douglas 2021). As data from the internet can often mirror biases present in society (such as racial, gender,

and geographical bias) (Buolamwini et al. 2018b), 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 (Radford et al. 2021). These include for instance, higher incidence of assigning crime-related classes to young people and to 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 (Radford et al. 2021). Researchers have pointed out the dangers of using unfiltered data for training deep learning models (Birhane et al. 2021). In the case of curated datasets, there is 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 use of data scraped directly from the internet rather than curated datasets (Jia et al. 2021).

Two major issues were identified from this audit. First, the techniques used for auditing may themselves suffer from bias. For example, the FairFace dataset classifies people into different 'races' based on the classification by the U.S. Census Bureau (Karkkainen et al. 2021b). This may lead to a 'Western/American' bias in the analysis as discussed in Chapter 3 (page 61). Second, the audit done by the authors appears to be done on an ad-hoc basis and is not systematic. Thus, it is difficult to scale the methodology to encompass other social biases.

This research builds 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. The goal is to study the performance of CLIP on the faces

of people based on a range of geographical locations. CLIP's zero shot predictions is used to get labels for the images and then we analyse those labels using Word Embeddings Association Test (WEAT), developed by (Caliskan, Bryson, et al. 2017) to measure bias in the predictions. WEAT, based on Implicit Association Test (IAT) measures associations between groups of words. This technique is used 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, patterns emerging from the model are analysed in order to understand how they 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 reinforce a binary view of gender but as a preliminary study to assess the effectiveness of the metric that can then be expanded upon in future work.

The following questions, part of **RQ3**, are addressed within this chapter:

- 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, the contributions of this chapter are as follows:

1. Associations in CLIP between adjectives and terms denoting occupations with images of people are analysed and evaluated for evidence of gender-based bias and the distribution, frequency, and prediction probability of the terms are analysed.

2. Gender bias is measured and quantified by using the WEAT score in order to evaluate and understand the dynamics of bias within the model.

3. Trends pertaining to gender bias uncovered in CLIP with employment and income data are studied to evaluate the extent to which real-world inequalities may be mirrored in models such as CLIP.

5.1.1 Web Crawling vs Curated Datasets

Birhane et al. (2018) 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 noted that LAION-400M is an attempt at open-sourcing CLIP and its analysis may not accurately reflect the issues present in CLIP. 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. (2021), 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 from the use of web-crawled datasets is the data being skewed towards demographics in more developed countries. In fact, the authors of CLIP[1] note that the training data is more skewed towards developed nations and

[1]https://github.com/openai/CLIP/blob/main/model-card.md

young male users.

5.1.2 Measuring Bias

Word Embedding Association Test (WEAT) is a technique developed by (Caliskan, Bryson, et al. 2017) to measure bias in language models. WEAT is based on the Implicit Association Test (IAT) (Greenwald et al. 1998), 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 associated with women, would be closer to words that denote women (e.g. 'she', 'her', 'woman'). In this way the gender bias in a model is measured by analysing labels predicted for men and women and associated words. This technique is discussed in detail in section 5.2.

5.1.3 Bias in Multimodal Models

Wolfe, Y. Yang, et al. (2022) 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, Banaji, et al. (2022) 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'.

5.2 Methodology

A visual dataset of people was curated using images returned by an online search when keywords pertaining to men and women were given. Google was used as the search engine due to it being the most widely used [2], accounting for more than 80% of worldwide search traffic. Selenium was used to automate the process with each search happening in an incognito profile. Two lexicons comprising the names of occupations and adjectives that describe personality were then generated. 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.

5.2.1 The Test Dataset

A dataset of human faces was curated by conducting online searches using Google advanced image search [3] and virtual locations using a VPN (ExpressVPN [4]) 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 similar to the methodology followed in Chapter 3 (page 66). The choice of languages and locations are the same as described in Chapter 3. Two search keywords were used: man and woman and translated them into the most common language of each region. The translations were verified by native speakers of that language. Then the IP location of the search engine was changed using a VPN to that of the most populous country of that particular region similar to that used in Chapter 3. 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 4.1 (page 81) 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

[2] `https://www.searchenginejournal.com/seo-guide/meet-search-engines/`
[3] `https://www.google.com/advanced_image_search` accessed: 12-10-2024
[4] `https://www.expressvpn.com/` accessed: 12-10-2024

language in the paper. In the case of two countries, half the images were queried 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.

5.2.2 The Keywords

Occupations: A comprehensive list of occupations was compiled based on published papers, online job portals and government sites in different locations. These include Garg et al. (2018), BBC Careers [5], LinkedIn [6] [7] [8], Australian Occupation List [9] and Canadian Occupation List[10]. The full list of keywords is provided in Appendix A.3.

Stereotypical Concepts of Personality Traits: In order to evaluate the prevalence of stereotypical associations pertaining to gender and personality, the work by Motschenbacher et al. (2020), was called upon 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 – openness, conscientiousness, extraversion, agreeableness, and neuroticism (Roccas et al. 2002).

5.2.3 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

[5] `https://www.bbc.co.uk/bitesize/articles/zdqnxyc` accessed: 19-01-2021

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

[7] `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

[8] `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

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

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

Table 5.1: Zero shot classification example. This is for the images queried from West Asia and North Africa using the Arabic language

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

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 Table 5.1. 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**.

5.2.4 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 (Radford et al. 2021). Therefore, by analysing the labels predicted during zero shot classification, biases in the model can be identified.

Three WEAT scores were used: WEAT score, WEAT Differential Association and WEAT Association score (Section 5.2.4) 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, lists of terms associated with that concept were compiled. In this instance, the focus is on men and women and use terms outlined in Table 5.2. These consist of gendered pronouns, references

Table 5.2: Gender attributes and terms

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

to people with gender implied and family roles that specify gender. The terms are the same as those used in Chapter 3.

WEAT Score, WEAT Differential Association and WEAT Association Score

This score is based on the technique developed by Caliskan, Bryson, et al. (2017). Let X and Y be the labels predicted for images of men and women respectively. The top 50% predicted labels were taken 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.

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B) \tag{5.1}$$

where,

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

$$x, y \in w$$

Here, $s(w, A, B)$ measures the association of w with A and B. This is the **WEAT association score**. In this 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,

$$\frac{mean_{x \in X} s(x, A, B) - mean_{y \in Y} s(y, A, B)}{std - devw \in X \cup Y s(w, A, B)} \tag{5.2}$$

which is the effective **WEAT score**.

The WEAT Differential Association Score is an intermediate score used for calculating the final WEAT Score. The higher the WEAT Association Score, the stronger the bias is. As per the authors of WEAT, the values for small, medium, and large are 0.25, 0.5, and 0.8 respectively (Caliskan, Bryson, et al. 2017). Therefore, any WEAT score of more than 0.5 indicates significant bias.

Factual Test

The WEAT Association Score is used on occupation-based labels and compares them with two sets of data from the real world: median salaries of those occupations[11] and the percentage of women in those occupations. These data points are from internet sources for the US[12].

Grad-CAM Analysis

A visual question-answering system was built using CLIP and used Gradient-weighted Class Activation Mapping (Grad-CAM) to generate a saliency map in order to visualise CLIP localisation. This allowed for qualitative analysis of gender bias in CLIP. The visual question-answering system is based on CLIP-ViL (Shen et al. 2021), 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. The adjectives and occupations were selected from the zero-shot predictions to form the questions and used an image of a man and woman as the input image.

[11] `https://www.glassdoor.com/index.htm` accessed:19-01-2021
[12] `https://www.zippia.com/` accessed:19-01-2021

Table 5.3: Skewness and kurtosis

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

5.3 Findings and Discussions

5.3.1 Exploratory Data Analysis

Adjectives

On analysing the labels predicted for images of men and women based on adjectives, it can be seen 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. An example of such predictions *only* for images from West Asia and North Africa is shown in Table 5.1. From Fig 5.1, it can be seen 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 5.3, it can be seen that even though the label distribution for men is skewed, those for women are highly skewed. Motschenbacher et al. (2020) 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 et al. (2020) 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 execu-

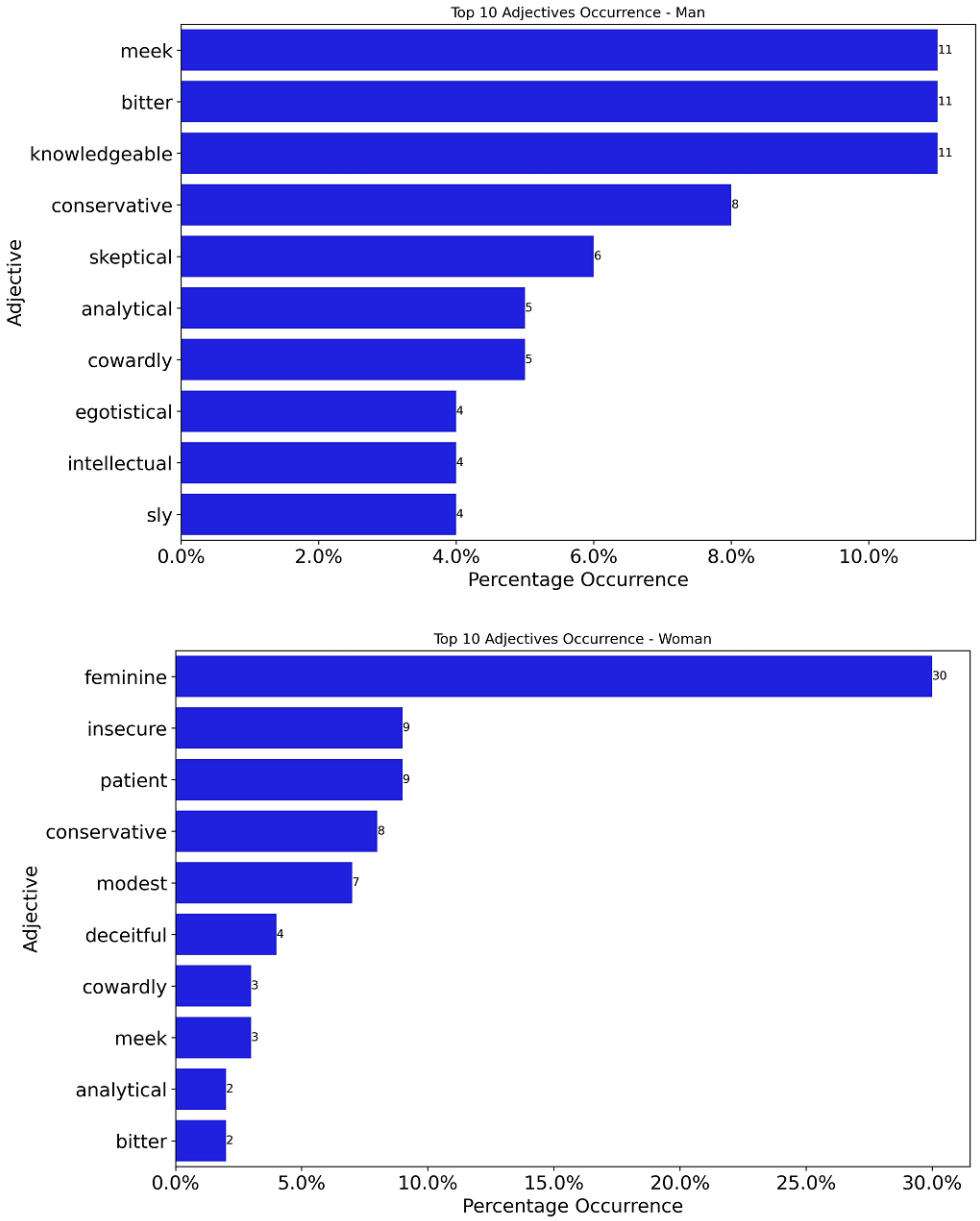


Figure 5.1: Top 10 adjectives occurrence

tive 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.

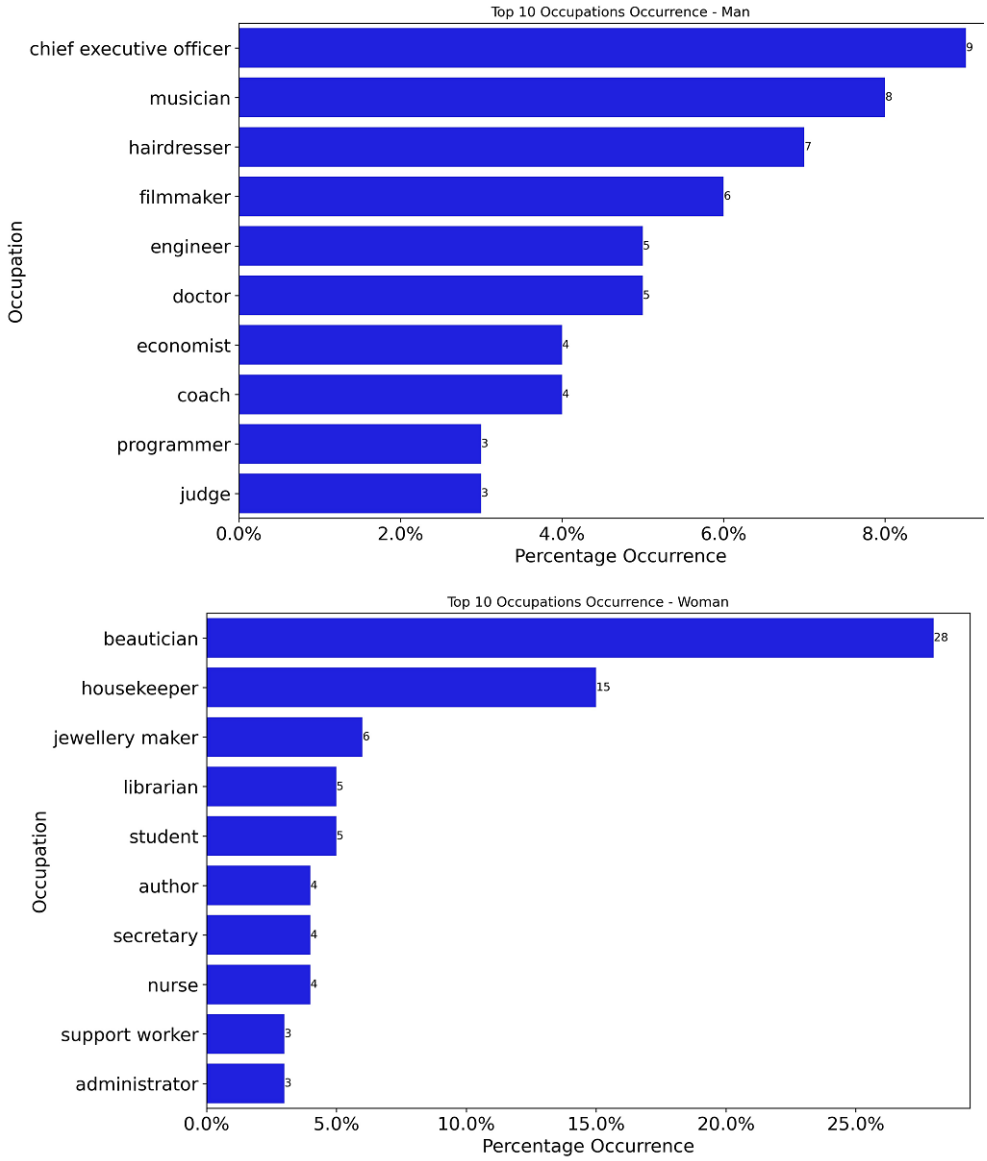


Figure 5.2: Top 10 occupations occurrence

5.3.2 WEAT Analysis

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 5.2.4. The predicted labels are used as targets and the attributes, as discussed in Section 5.2.4. WEAT scores of 0.2, 0.5 and 0.8 are considered small, big and large, respectively (Caliskan, Bryson, et al. 2017). 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

Table 5.4: WEAT Score and WEAT Differential Association

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

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.

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 5.2.4, 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 the label predictions for men and women are compared individually with both sets of attributes.

Table 5.5 shows the WEAT Association Score for labels for men and women. It can be seen that the labels based on adjectives predicted 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, neutral results can be seen in case of men and biased results in the case of women.

When the WEAT association score for the top ten adjectives and occupations are observed individually (Fig 5.3), there is evidence of expected and also stereotypical

Table 5.5: WEAT Association Score

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

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 5.4, 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 5.5).

It can also be seen 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 5.6). 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 demonstrating 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.

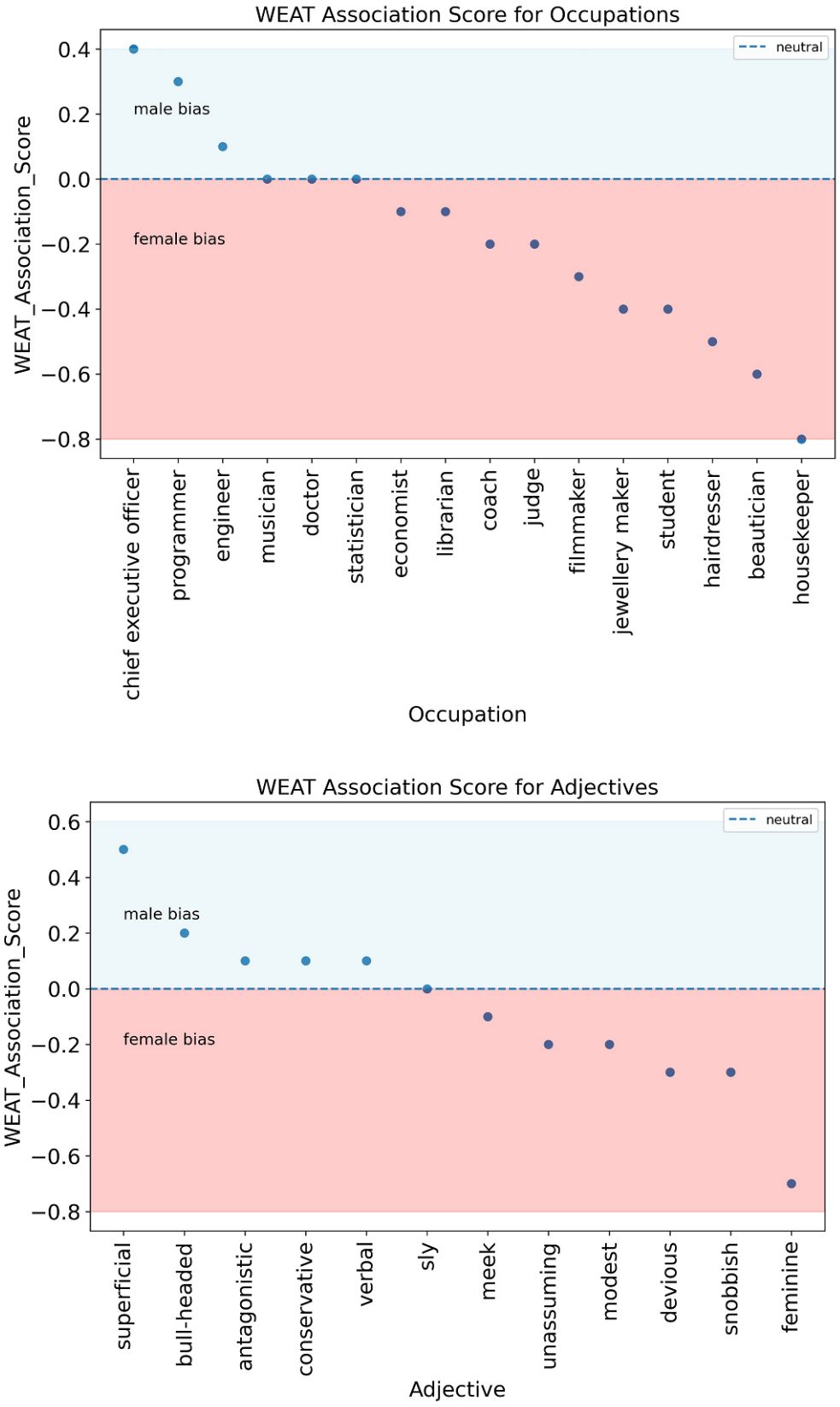


Figure 5.3: WEAT Association Scores of top adjectives and occupations

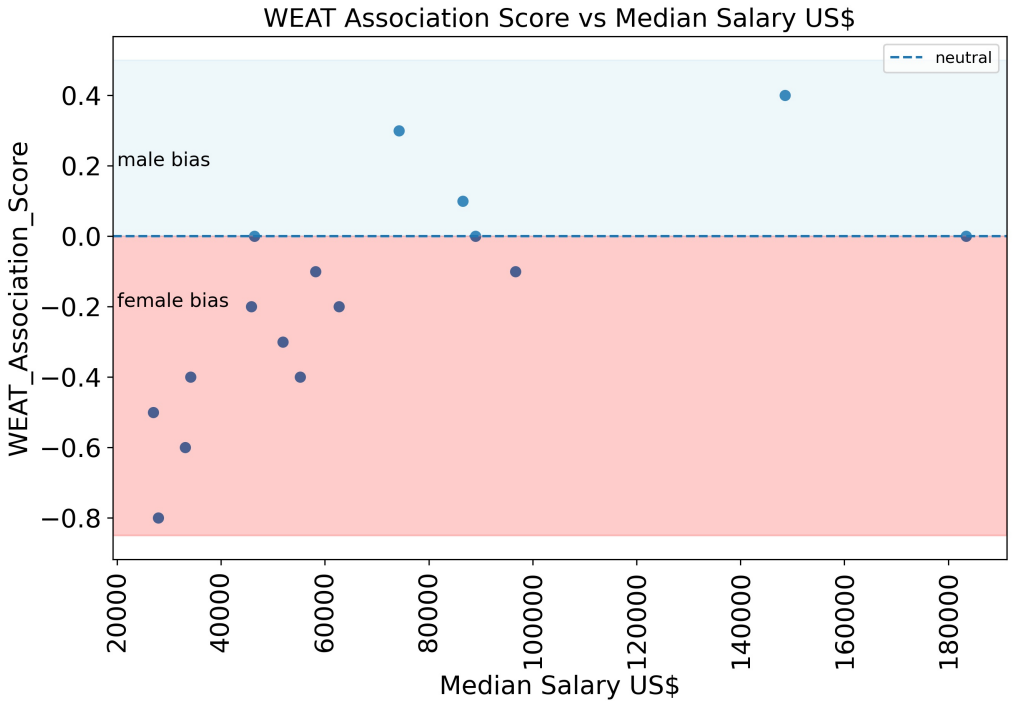


Figure 5.4: WEAT Association Score vs Median Salary (USD)

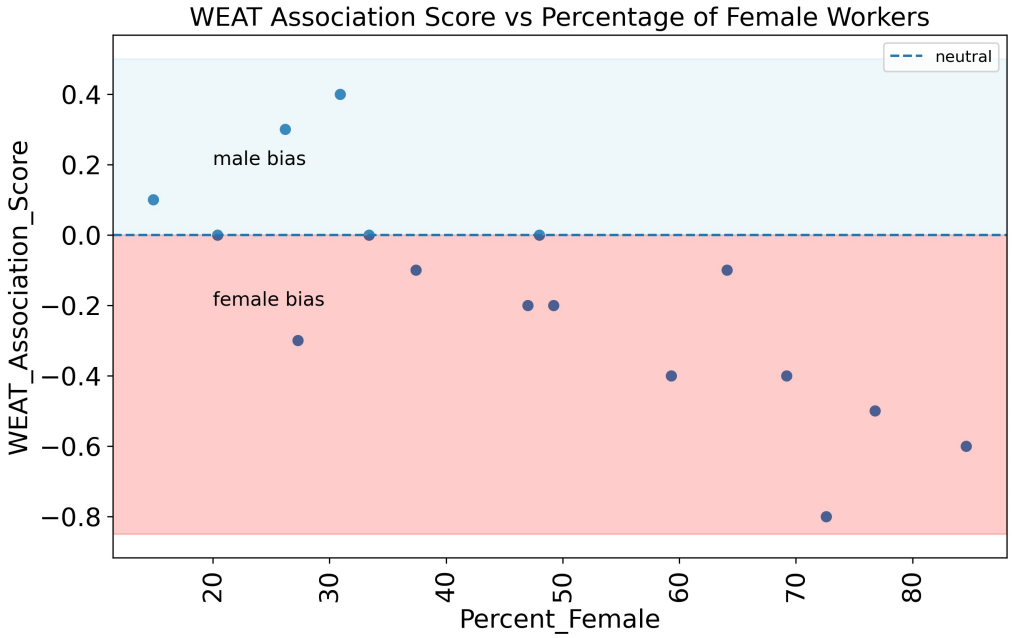


Figure 5.5: WEAT Score vs Percentage of Female Workers in the Associated Occupation

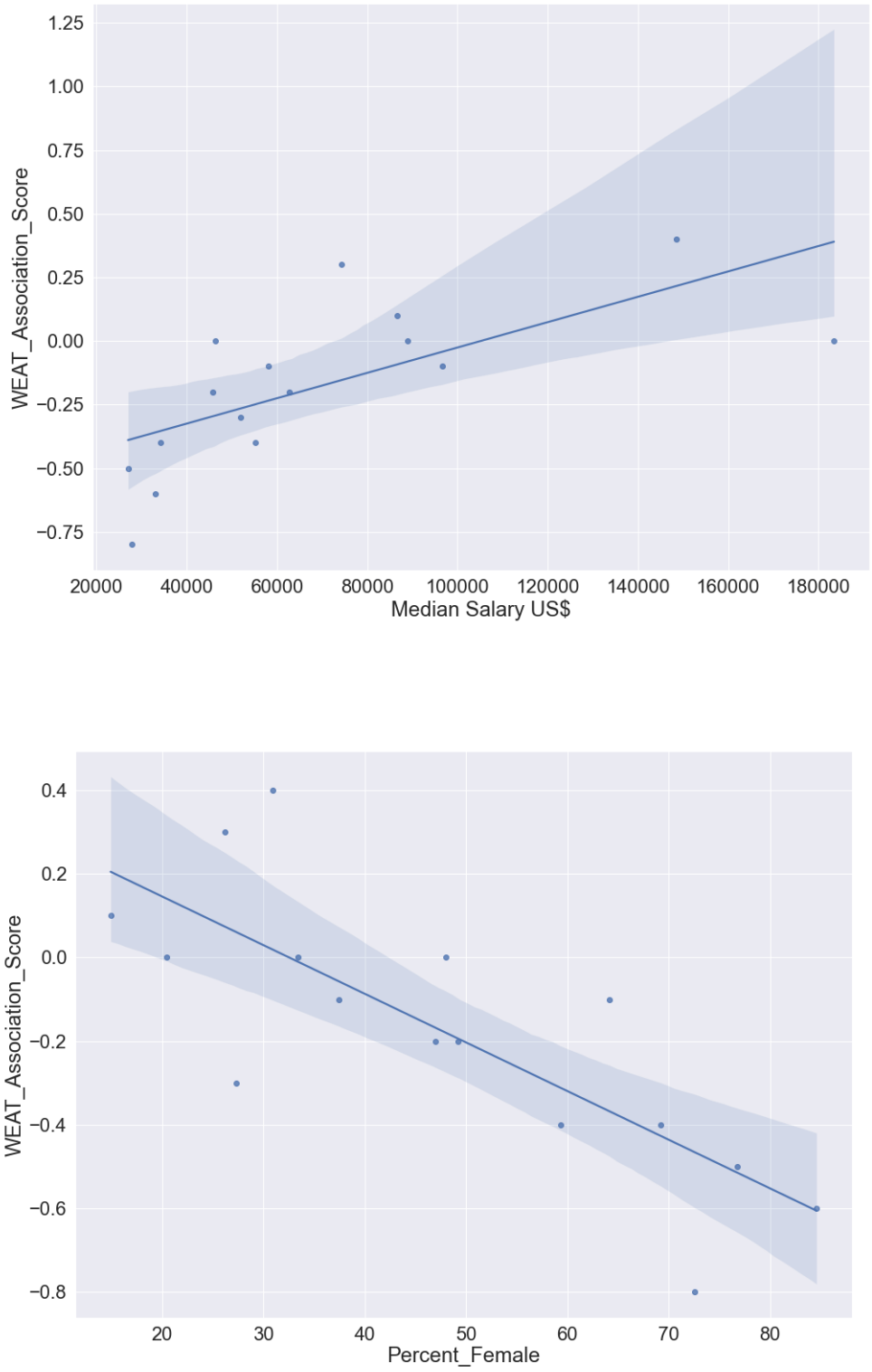


Figure 5.6: Regression line showing model bias

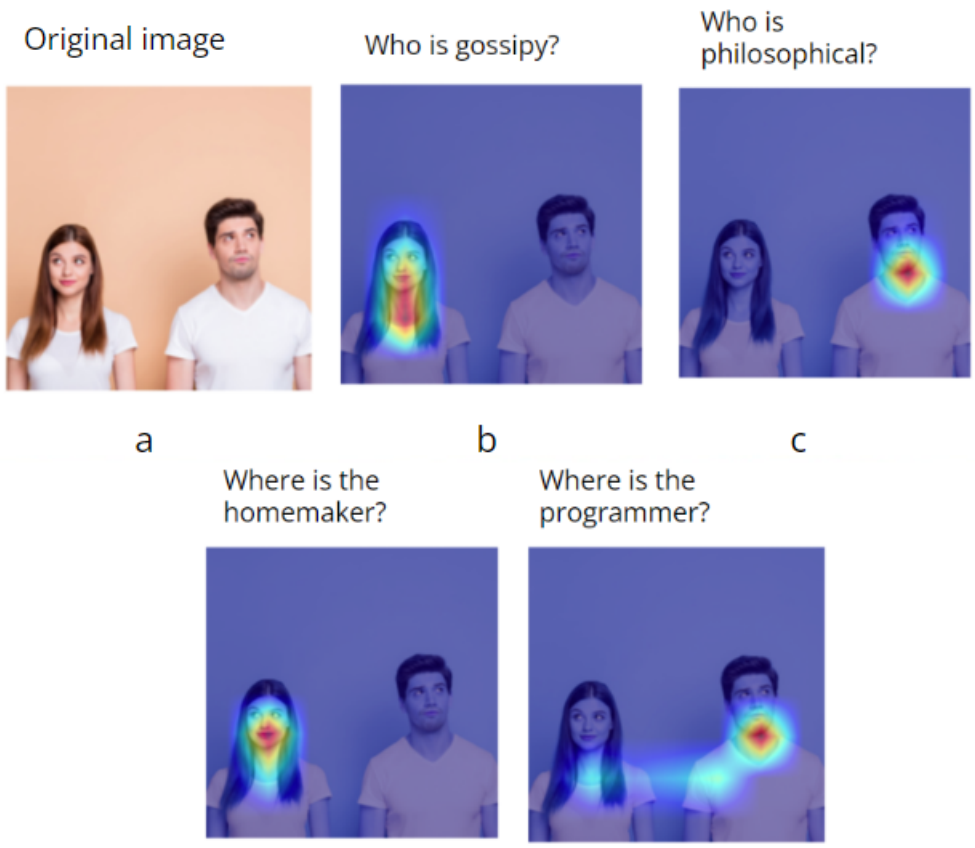


Figure 5.7: Grad-CAM analysis of CLIP VQA. a: original image, Grad-CAM localisation prompts: b: Who is gossipy?, c: Who is philosophical?, d: Where is the homemaker?, e: Where is the programmer? The image is not a part of the curated dataset and was retrieved using Google image search.

5.3.3 Grad-CAM Analysis of Bias in CLIP

Fig 5.7 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 figure 5.7. The input image (8a) shows a man and woman (queried using Google image search and not from the 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. It is observed 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 terms philosophical and programmer, thus showing how visualisation using explainable AI can be useful in discovering and highlighting evidence of gender bias in CLIP.

5.4 Conclusion and Limitations

From the results of the experiments, it can be seen that CLIP reflects biases present on the internet and in 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 5.3 corroborate this. This shows that the scale of very 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 are needed. A well-established technique for bias analytics used in natural language processing called WEAT was taken which measures the similarity between concepts in vector space and it was applied to computer vision

models. The findings demonstrated that this method may prove to be useful in providing in-depth analysis for bias analytics, providing a quantitative and scalable method for evaluating bias in large multimodal deep learning models. 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 that can be detected and measured using this technique, excluding those such as ethnic and cultural bias and non-binary gender bias. To summarise, CLIP associates adjectives and occupations-related labels to images of men and women differently reflecting stereotypical gender bias which can be a result of training it on unfiltered data from the internet. WEAT can be used to measure gender bias in vision models.

Chapter 6

Measuring Bias in Multimodal Models: Multimodal Composite Association Score

This chapter continues the work presented in Chapter 5 where a cross-domain adaptation of the Word Embeddings Association Test from Natural Language Processing was adapted for measuring gender bias in Computer Vision Models. In this chapter, the concept of measuring stereotypical associations in word embeddings is extended to multimodal embeddings with the aim to specifically use it for Text-To-Image (TTI) Diffusion Models such as DALL-E and Stable Diffusion.

Section 6.1 outlines the motivation for the research discussed in this chapter. Section 6.2 explains the novel contribution: *MCAS: Multimodal Composite Association Score* and its component scores. Section 6.3, provides details on the experiment and section 6.4 analyses the findings with 6.5 concluding the research work.

The research that emanated from this work was published at the 4$^{\text{th}}$ International Workshop on Algorithmic Bias in Search and Recommendation, co-located with the 45$^{\text{th}}$ European Conference on Information Retrieval (ECIR) 2023, Dublin, Ireland.

6.1 Motivation

Most current methods of auditing bias in vision models generally use two types of techniques: (1) measuring associations in the learning representations (Serna et al. 2021; Sirotkin et al. 2022; Steed et al. 2021) and (2) analysing the predictions (Buolamwini et al. 2018b; Krishnakumar et al. 2021). Most of these techniques (Buolamwini et al. 2018b; Serna et al. 2021; Sirotkin et al. 2022; Steed et al. 2021) are designed for predictive models, mainly Convolutional Neural Networks (CNNs). Recent advances in deep learning, however, have given rise to multi-stage, multimodal models with DALL-E and Stable Diffusion being two of the most popular models. Generative multimodal models based on diffusion models are easier to train than GANs and have higher variability in image generation that enables them to model complex multimodal distributions. This allows them to generate images using abstract ideas with less tight bounding than GANs (Ramesh, Dhariwal, et al. 2022; Rombach et al. 2022). The easier training regimen allows developers to train these models on very large datasets. This has led to models being trained on increasingly large datasets, often crawled from the Internet. These datasets are generally unfiltered, leading to the models inheriting social biases prevalent on the web (Birhane et al. 2021). These models, therefore, require new approaches to detecting bias.

Models such as DALL-E (Ramesh, Dhariwal, et al. 2022), Stable Diffusion (Rombach et al. 2022) and Contrastive Language and Image Pre-training (CLIP) (Radford et al. 2021) operate on multiple modalities, such as text and images. These models have numerous applications ranging from content creation to image understanding and image and video search (https://www.facebook.com/braddwyer n.d.). They also combine multiple different models using outputs to form inputs to another model. CLIP uses Vision Transformers or ResNet for image encoding and a text encoder for text encoding. DALL-E and Stable Diffusion use CLIP for their first stage involving generating text embeddings and a diffusion model (unCLIP for DALL-E and Latent Diffusion for Stable Diffusion) to generate images. This multi-stage multi-model approach also carries the risk of bias amplification, where one

model amplifies the bias of another model (T. Wang et al. 2019).

With the increasing popularity of TTI models, an increasing volume of internet content may be AI-generated and this content, comprising both images and text may be indexed by search engines and appear in search results. Apart from concerns arising from privacy and copyright law, biased and harmful generated content can further exacerbate social issues already present in search engine results (A. Wang, Liu, et al. 2022), [Chapter 3 (page 61]. As data from the internet (often using web scraping using search engines) is used for training TTI generative models (A. Wang, Liu, et al. 2022) [Chapter 3], this may create a loop that further amplifies social biases. The integration of generative AI and search engines, which are currently being developed may complicate these issues further.

Text-To-Image (TTI) Generative multimodal models based on Diffusion Models are easier to train than GANs and have higher variability in image generation that enables them to model complex multimodal distributions. This allows them to generate images using abstract ideas with less tight bounding than GANs (Ramesh, Dhariwal, et al. 2022). The easier training regimen allows developers to train these models on very large datasets. This has led to models being trained on increasingly large datasets, often crawled from the Internet. These datasets are generally unfiltered, leading to the models inheriting social biases prevalent in the web (Birhane et al. 2021). This chapter aims to address these research questions, as a part of **RQ3**:

- Do Text-to-Image (TTI) diffusion models exhibit stereotypical gender bias?
- How can gender bias be effectively measured in TTI models?

The main contributions of this chapter are:

- Audited gender bias in TTI models.
- Developed novel metrics to measure gender bias in TTI models.

6.2 MCAS: Multimodal Composite Association Score

The Multimodal Composite Association Score or MCAS that is proposed is derived from WEAT and measures associations between specific genders (what we term 'attributes') and what are termed 'targets' corresponding to concepts such as occupations, sports, objects, and scenes. MCAS consists of four constituent components (scores), each measuring bias in certain modalities (e.g., text, vision or both). This follows the approach of the WEAT Association Score, which measures stereotypical associations between attributes (gender) and a set of targets. As formulated by Caliskan, Bryson, et al. (2017), let A and B be two sets of attributes, each representing a concept. Additionally let W be a set of targets, w. Then

$$s(w, A, B) = mean_{a \in A} cos(\vec{w}, \vec{a}) - mean_{b \in B} cos(\vec{w}, \vec{b}) \tag{6.1}$$

where, $s(w, A, B)$ represents the WEAT Association Score. $cos(\vec{w}, \vec{a})$ and $cos(\vec{w}, \vec{b})$ denote the cosine similarity between the vectors of the words from attribute sets, A and B respectively. If target w is more closely related to attributes in A, implying the target as a bias towards A, then the association score will be positive and if it is more closely related to attributes in B, then the score will be negative. It is important to note that the association score measures bias within the embeddings and not against an external benchmark.

The association scores are a set of linear scales ranging from -1 to 1 with zero being the midpoint. If the extremities represent a pair of opposite concepts e.g. man and woman, then any point which lies closer to one end can be considered closer to that particular concept. If a point lies at zero, then it can be considered as being equally close to both ends i.e. neutral. Now if that point represents a real-world concept such as occupation, then if it is closer to the male end, it shows male bias. The scale here represents the representation space of a multimodal model (e.g. CLIP) and the point (target) and the endpoints (gender attributes) are features

extracted from the model. This is shown in Figure 6.1.

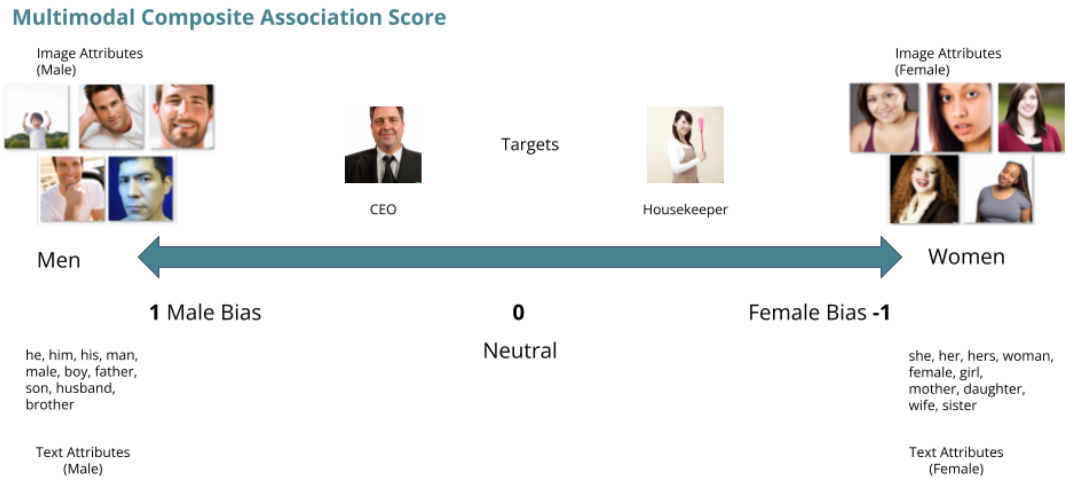


Figure 6.1: Association scores visualised

6.2.1 Attributes and Targets

The WEAT Association Score was originally intended for assessing text embeddings. Building on this work, it is used in MCAS for both text and image embeddings. MCAS consists of four individual association scores, each measuring the association between embeddings of text and images, defined in section 6.2.2. As the main focus of this chapter is TTI generative models, the attributes and targets comprise both text and images. The TTI generative models DALL-E 2 and Stable Diffusion both work in similar ways; they take in a text input describing a visual imagery and generate a corresponding image output. For measuring gender bias, men and women are represented both in terms of text and images (see Table 6.1). These texts and images form the gender *attributes*.

Targets refer to the concepts that are being tested for evidence of bias. To test the effectiveness of MCAS real-world topics that may be associated with stereotypical representations of gender are identified and these scenarios are captured in text phrases. These phrases are used as prompts for the generative models to generate images. This results in a set of targets comprising text phrases (e.g. *an image of a CEO* or *an image of a person using a food processor*) along with a set of images

Table 6.1: Examples of Text and Image Attributes

Text Attributes	**Image Attributes (from DALL-E 2)**
he, him, his, man, male, boy, father, son, husband, brother	
she, her, hers, woman, female, girl, mother, daughter, wife, sister	

generated by the models from those prompts. Examples of attributes and targets are provided in Tables 6.1 and 6.2.

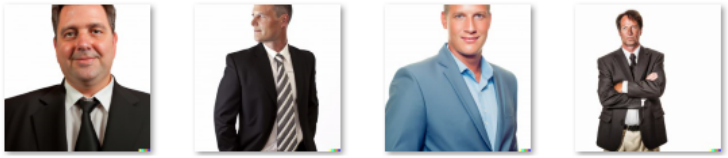
6.2.2 MCAS and its Components

In this experiment, the focus is on TTI generative models and is tailored for them. MCAS consists of four individual component scores: Image-Image Association Score, Image-Text Prompt Association Score, Image-Text Attributes Association Score and Text-Text Association Score. Each of these scores measures bias in different modalities and different stages of the TTI models.

Image-Image Association Score: This score measures bias by comparing the cosine similarities between image attributes representing gender and generated images representing target concepts. Letting A and B be two sets of images representing gender categories and W be a set of images representing targets, then the Image-Image Association Score, (II_{AS}), is given by:

$$II_{AS} = mean_{w \in W} s(w, A, B) \tag{6.2}$$

Table 6.2: Examples of targets. Images generated by DALL-E 2.

Prompt	**Generated Image**
an image of a chief executive officer	
an image of a badminton player	
an image of a person using a food processor	
an image of a person using a lathe machine	

where,

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

Image-Text Prompt Association Score: This score measures bias between the image attributes representing gender and the textual prompts used to generate the target concepts. Letting A and B be two sets of images representing gender and W be a set of prompts representing targets in text form, then the Image-Text Prompt Association Score, (ITP_{AS}), is calculated in the same way as shown in Equation 6.2.

Image-Text Attributes Association Score: This score calculates bias in a similar manner as the other scores with the difference being that the attributes are represented not by images, but by text. The target concepts are a set of images generated from prompts. The score, (ITA_{AS}), is calculated in the same way as shown in Equation 6.2 with A and B are text attributes and W, target images.

Text-Text Association Score: This score computes gender bias using entirely textual data. The attributes are the same as in Image-Text Attributes Association

Score and the targets are prompts (as in Image-Text Prompt Association Score). The score, (TT_{AS}), is calculated in the same way as Equation 6.2. This is the only score which does not involve image embeddings. As both the models used in the experiment use CLIP for converting text, this score also measures CLIP bias.

To calculate the scores, A, B and W represent the features extracted from their corresponding data. The implementation details are explained in the experiment section. The final MCAS score is defined as the sum of all the individual association scores. It is given as:

$$MCAS = II_{AS} + ITP_{AS} + ITA_{AS} + TT_{AS} \tag{6.3}$$

6.2.3 MCAS for TTI Generative Diffusion Models

TTI Generative models based on Diffusion models generally employ a two-stage mechanism. Firstly, the input text is used to generate embeddings. DALL-E and Stable Diffusion both use CLIP for this stage. CLIP is a visual-linguistic multimodal model which connects text with images. CLIP is trained on 400 million image-text pairs crawled from the internet using contrastive learning (Radford et al. 2021).

Once the embeddings are generated, then the second stage involves passing them to a Diffusion Model. Diffusion Models are based on Variational Autoencoders (VAEs) that use self-supervised learning to learn how to generate images by adding Gaussian noise to the original image (encoding) and reversing the step to generate an image similar to the original (decoding). DALL-E uses unCLIP where first the CLIP text embeddings are fed to an autoregressive diffusion prior to generate image embeddings which are then fed to a diffusion decoder to generate the image (Ramesh, Dhariwal, et al. 2022). Stable Diffusion uses Latent Diffusion to convert the CLIP embeddings into images. Latent Diffusion Model (LDM) uses a Diffusion Model similar to a denoising autoencoder based on a time-conditional UNet neural backbone (Rombach et al. 2022). Both the processes are similar in nature. Fig 6.4 shows a high-dimensional generalisation of both the models.

The individual MCAS component scores can measure bias in different stages of the image-generating process. The Image-Image Association Score measures bias solely on the basis of the generated images thus encompassing the whole model. The Image-Text Prompt Association Score measures bias in both visual and textual modalities. As both the prompts and generated images were part of the image generation process, this score also encompasses the whole generation sequence. The Image-Text Attributes Association Score measures bias in both the modalities and as the text attributes are external (i.e. not a part of the image generation process), the model bias can be measured using external data or standards. The Text-Text Association Score measures bias only in textual modality. As only CLIP handles the text, this score can be used to measure bias in CLIP. This score also allows for bias measurement using external data. Thus MCAS provides a comprehensive and quantitative method to measure bias in multimodal models. Table 6.3 describes the characteristics of the MCAS component scores. Figure 6.2 show how the components of MCAS relate to the targets and attributes.

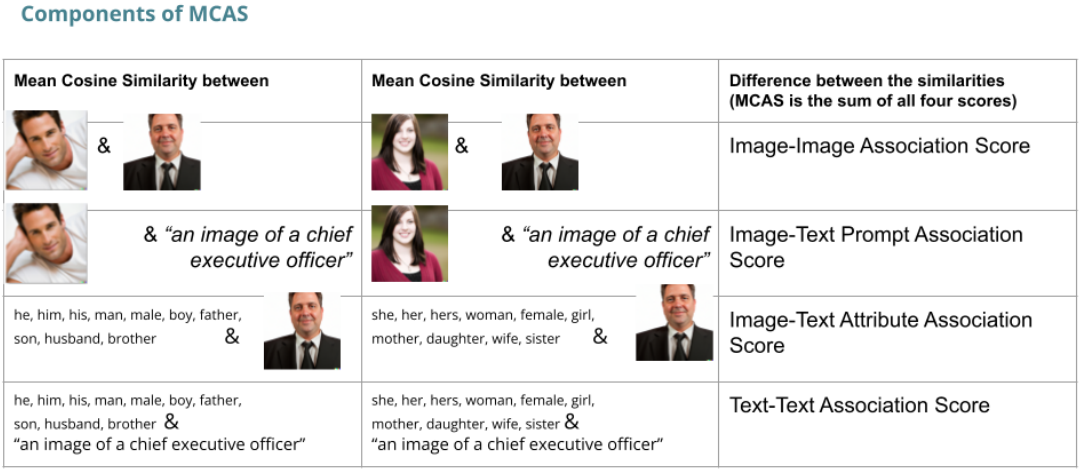


Figure 6.2: Components of MCAS

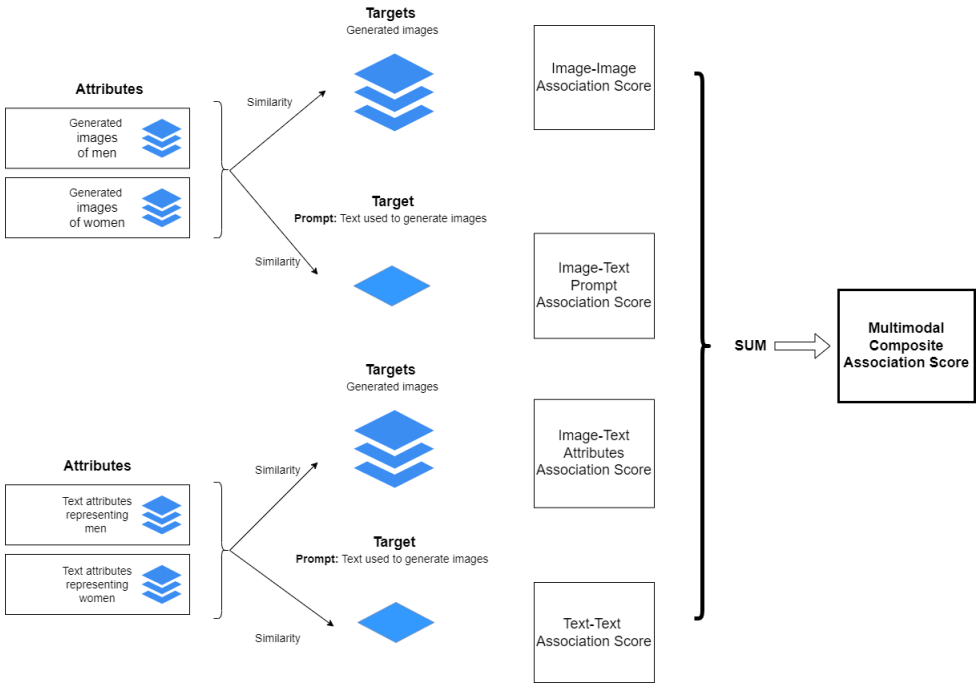


Figure 6.3: MCAS Algorithm

Table 6.3: MCAS scores characteristics

Association Score	Modality	whole model?	external data?
Image-Image (II_{AS})	Image	Yes	No
Image-Text Prompt (ITP_{AS})	Image & Text	Yes	No
Image-Text Attributes (ITA_{AS})	Image & Text	No	Yes
Text-Text (TT_{AS})	Text	No	Yes

6.3 Experiment

6.3.1 Curating the Attributes and Targets

To evaluate the effectiveness of MCAS in uncovering evidence of gender bias, two datasets were generated comprising the attribute and target concept data in both visual and textual form for two models, DALL-E 2 and Stable Diffusion. The target concepts were those that have been used in previous research (A. Wang, Liu, et al. 2022; T. Wang et al. 2019) to detect gender bias. For this experiment, the focus is on evaluating concepts pertaining to men and women. The text and image attributes compiled are presented in Table 6.1).

To create the visual attributes datasets, text prompts were used to generate images and the complete list of the keywords used is in Appendix A.4 (page 190).

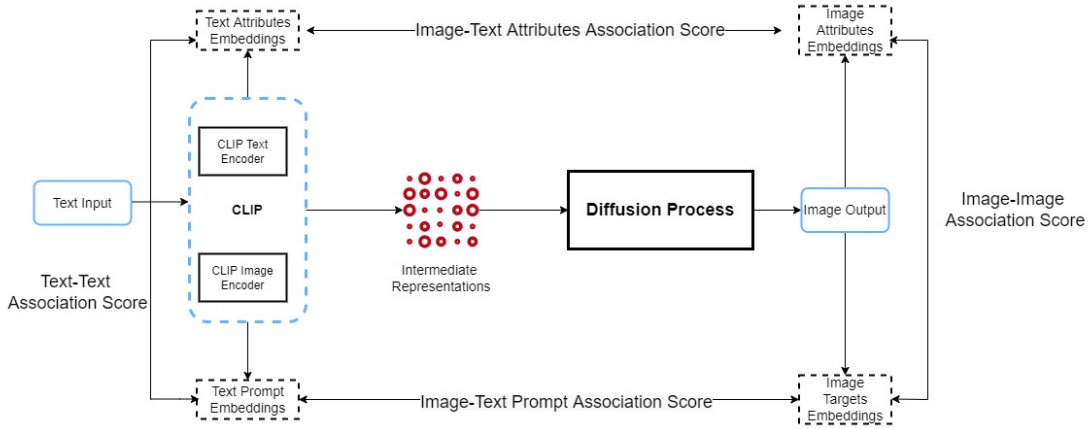


Figure 6.4: Association Scores in Diffusion Models. A generalised diagram showing the working of diffusion models like DALL-E 2 and Stable Diffusion. The embeddings are generated using an external CLIP model.

Table 6.4: Target categories and keywords. Based on (Garg et al. 2018; A. Wang, Liu, et al. 2022).

Category	**Keyword**	**Association**
Occupations	CEO, engineer, doctor, programmer, farmer	Men
	beautician, housekeeper, secretary, librarian, nurse	Women
Sports	baseball player, rugby player, cricket player	Men
	badminton player, swimmer, gymnast	Women
Objects	car, farm machinery, fishing rod	Men
	food processor, hair drier, make-up kit	Women
Scenes	theodolite, lathe machine, snowboarding	Men
	shopping, reading, dollhouse	Women

There is a slight difference in keywords for DALL-E 2 and Stable Diffusion due to restrictions within DALL-E 2. A total of 128 images (16 per attribute phrase) were generated separately for DALL-E 2 and Stable Diffusion to form the 'attribute' set of images. To compile datasets representing 'target' concepts, terms from work by Garg et al. (2018) and A. Wang, Liu, et al. (2022) were adapted to capture domains where gendered associations were found to be evident (see Table 6.2 and Table 6.4). A total of 688 images (128 for attributes and 560 for targets) were generated using each of DALL-E 2 and Stable Diffusion. The images generated by DALL-E 2 were used for DALL-E 2 in the association score calculation and similarly for Stable Diffusion. The number of images was determined on an ad-hoc basis.

6.3.2 Calculating the Scores

CLIP was used to extract the features for both the text and images. As CLIP is used by both models, they would be similar to the embeddings generated in the models. The extracted features were then used to calculate the individual association scores and summed to get the final MCAS score. In the experiments, text and image attributes associated with men were assigned as the first attribute (A) and those associated with women as the second (B). This means that a positive score indicates a higher association between the target concepts and men and a negative score indicates a higher association with women. A score of zero would indicate that the target concepts appear neutral in terms of associations with men or women. The numeric value indicates the magnitude of the association. In the case that target concepts correspond to domains where gender bias has been found to be prevalent, then these associations may indicate a prevalence of gender bias within the model.

6.4 Findings and Discussion

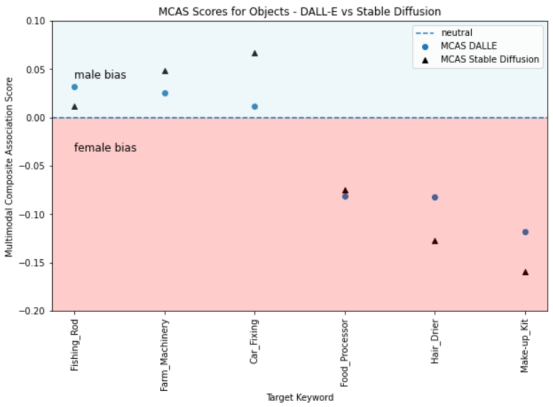
In evaluating both DALL-E 2 and Stable Diffusion models associations, which have in previous research (Chapter 2, section 2.2.1 (page 42) been found to reflect gender bias, were uncovered in the models. Consistent patterns of gendered associations were uncovered and given that these target concepts were based on concepts that previous research had found to relate to gender bias, then these patterns are indicative of underlying gender bias. Targets and their MCAS scores are provided in Figure 6.6 and Table 6.5.

Both models follow a similar pattern in terms of gendered associations except for the *scenes* category where DALL-E 2 presents an association with men and the targets 'snowboard' and women with 'lathe' whereas Stable diffusion presents the opposite.

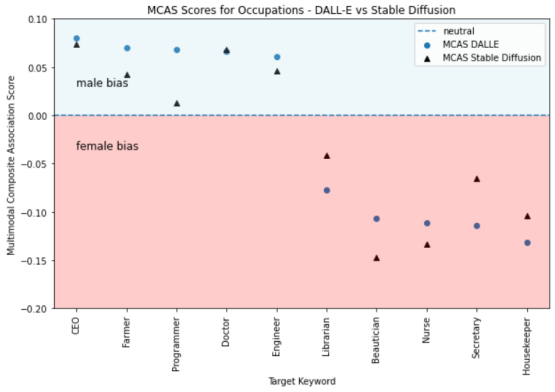
For the category *objects*, the target 'make-up kit' is strongly associated with women, which indicates that MCAS could be used to uncover gender bias. Similarly,

Figure 6.5: Gender bias per keyword for DALL-E 2 and Stable Diffusion.

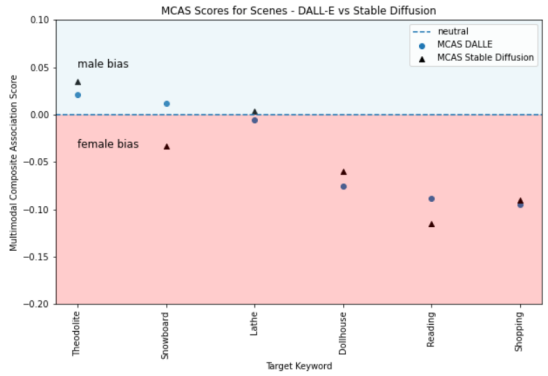
(a) Objects



(b) Occupations



(c) Scenes



(d) Sports

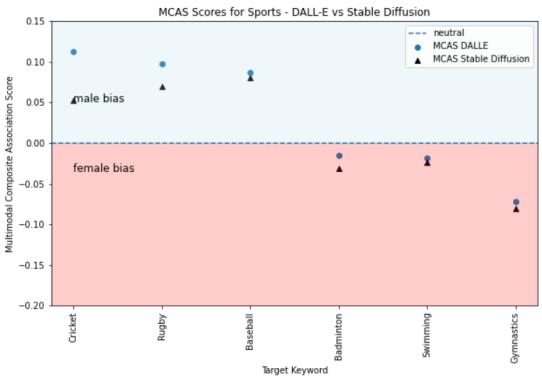


Figure 6.6: MCAS scores by category

Table 6.5: Gender bias per keyword for DALL-E 2 and Stable Diffusion.

		DALL-E 2		**Stable Diffusion**	
Target Type	**Target Keyword**	**MCAS Score**	**Bias**	**MCAS Score**	**Bias**
Occupations	CEO	0.080	Male	0.073	Male
Occupations	Engineer	0.061	Male	0.046	Male
Occupations	Doctor	0.065	Male	0.067	Male
Occupations	Farmer	0.070	Male	0.041	Male
Occupations	Programmer	0.067	Male	0.012	Male
Occupations	Beautician	-0.106	Female	-0.147	Female
Occupations	Housekeeper	-0.131	Female	-0.103	Female
Occupations	Librarian	-0.077	Female	-0.041	Female
Occupations	Secretary	-0.113	Female	-0.065	Female
Occupations	Nurse	-0.111	Female	-0.132	Female
Sports	Baseball	0.086	Male	0.080	Male
Sports	Rugby	0.097	Male	0.069	Male
Sports	Cricket	0.112	Male	0.052	Male
Sports	Badminton	-0.015	Female	-0.031	Female
Sports	Swimming	-0.018	Female	-0.023	Female
Sports	Gymnastics	-0.072	Female	-0.080	Female
Objects	Car_Fixing	0.012	Male	0.067	Male
Objects	Farm_Machinery	0.026	Male	0.049	Male
Objects	Fishing_Rod	0.032	Male	0.012	Male
Objects	Food_Processor	-0.080	Female	-0.074	Female
Objects	Hair_Drier	-0.081	Female	-0.127	Female
Objects	Make-up_Kit	-0.117	Female	-0.160	Female
Scenes	Theodolite	0.021	Male	0.035	Male
Scenes	Lathe	-0.005	Female	0.003	Male
Scenes	Snowboard	0.012	Male	-0.033	Female
Scenes	Shopping	-0.010	Female	-0.090	Female
Scenes	Reading	-0.088	Female	-0.114	Female
Scenes	Dollhouse	-0.075	Female	-0.060	Female

Table 6.6: MCAS statistics – DALL-E 2 and Stable Diffusion. Average bias and standard deviation scores per category

Category	**Terms with male bias**		**Terms with female bias**		**All terms**	
	Standard Deviation	**Average Bias**	**Standard Deviation**	**Average Bias**	**Standard Deviation**	**Average Bias**
DALL-E 2						
Objects	0.0080	0.0230	0.0170	-0.0930	0.0590	-0.0350
Occupations	0.0060	0.0690	0.0170	-0.1000	0.0890	-0.0190
Scenes	0.0040	0.0160	0.0350	-0.0650	0.0480	-0.0380
Sports	0.0100	0.0980	0.0260	-0.0350	0.0700	0.0310
All categories	0.0052	0.0515	0.0238	-0.0733	0.0665	-0.0152
Stable Diffusion						
Objects	0.0200	0.0400	0.0340	-0.1200	0.0860	-0.0380
Occupations	0.0200	0.0400	0.0400	-0.9800	0.0800	-0.0200
Scenes	0.0150	0.0190	0.0300	-0.0700	0.0500	-0.0400
Sports	0.0100	0.0600	0.0250	-0.0400	0.0590	0.0110
All categories	0.0162	0.0397	0.0322	-0.3025	0.0687	-0.0217

stereotypical patterns were found in relation to the *occupations* category, where 'CEO' was strongly associated with men and 'housekeeper' and 'beautician' were most associated with women.

In *scenes*, 'theodolite' is the only target showing any significant association with men whereas women were associated with 'shopping' and 'reading'.

In the case of *sports*, the only target strongly associated with women is 'gymnastics' with the general trend demonstrating a stronger association between sports and men. This is evident from Table 6.1 where *sports* is the only category with an overall higher association with men.

The standard deviation and average bias (MCAS) scores for each category for both models are presented in Table 6.6. This demonstrates that for the targets more likely to be associated with men or women, the strength of the association is higher for women. Where bias occurs, therefore, it seems that bias is stronger when it relates to women. Stable Diffusion has generally higher scores in terms of strength of gendered association than DALL-E. This indicates that Stable Diffusion has higher stereotypical associations and DALL-E's scores are more spread out, implying that Stable Diffusion may be more biased than DALL-E. Further work is needed to assess this more fully.

6.5 Conclusion and Future Work

This chapter introduces MCAS as a proposal for examining bias across both text and image modes for large-scale multimodal generative models and provides a demonstration of its effectiveness when used to evaluate models for gender bias. It can be seen that this method can uncover evidence of gender bias in both DALL-E 2 and Stable Diffusion. MCAS as a whole provides a comprehensive score for quantifying bias in multimodal models. The methodology can be extended to other models using different modalities or using different internal stages. For example, the Text-Text and Image-Image Association Scores can be used for comparatively smaller models such as CLIP. The methodology itself is based on the highly popular WEAT.

In this chapter, the work is limited to gender bias related to representations of men and women but other biases including those pertaining to race, ethnicity and geography may be evaluated. The individual MCAS components can be used for understanding how bias is handled within the model itself. For example in the two-stage models, the component scores can tell which stage is responsible for how much bias and whether there is any bias amplification. The component scores can also be further adapted to understand how bias forms during the entire process by extracting outputs from substages and measuring bias in them. The effect of hyperparameters on bias can also be studied in a similar way. The identification and evaluation of bias in multimodal models can help understand and also mitigate bias in AI-generated content. To summarise, TTI diffusion models such as Stable Diffusion v2 and DALL-E 2 exhibit stereotypical gender bias and MCAS can provide an effective way of measuring such bias.

Chapter 7

Auditing the Impact of Computer Vision Architectures on Gender Bias

The research presented in this chapter is inspired by the research presented in Chapter 5. In Chapter 5, it was observed that using Vision Transformers as image encoders in Contrastive Language Image Pretraining (CLIP) increased the skewness of the stereotypical gender bias in the model's predictions. This chapter presents a thorough investigation of the performance of the two most popular families of models used in computer vision: Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). In order to assess gender bias in classification models, a novel metric called *Accuracy Difference* is proposed. The *Image-Image Similarity Score* – one of the constituent scores of *Multimodal Composite Association Score* (discussed in Chapter 6) is also adapted to further measure gender bias in classification models.

Section 7.1 explores the motivation behind the project explaining the key differences between Convolutional Neural Network and Vision Transformer architectures. Section 7.2 defines the novel bias measuring metric *Accuracy Difference* and explains the adaptation of the *Image-Image Association Score* for measuring bias in image classifiers. Section 7.3 discusses the methodology including dataset creation, model

training, bias measurement, and bias analytics using CLIP. Sections 7.4 and 7.5 discuss the findings and conclusions respectively.

The research that emanated from this work was published at the 34th British Machine Vision Conference, Aberdeen, UK, 2023.

7.1 Motivation

Vision Transformers (ViT), derived from Transformers in Natural Language Processing, have increasingly become important as they outperform Convolutional Neural Networks in many application domains (Khan et al. 2022; Kolesnikov et al. 2021; Naseer et al. 2021). 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 estimates the relevance of one patch of an image with another (Khan et al. 2022; Kolesnikov et al. 2021). This enables ViTs to capture 'long-term dependencies' in the data and thus possess a larger receptive field (Khan et al. 2022). Popular computer vision models and their applications have been shown to exhibit a large range of social biases including gender (Birhane et al. 2021; Buolamwini et al. 2018b), racial (Buolamwini et al. 2018b; Karkkainen et al. 2021b), and geographical biases [Chapters 3 and 4]. Most of the work on detecting such biases (Serna et al. 2021; T. Wang et al. 2019; Zhao et al. 2017a) and mitigating them (T. Wang et al. 2019; Z. Wang et al. 2020; Zietlow et al. 2022) have been performed on CNNs. Although most of the biases originate in the training data (Singh et al. 2020; A. Wang, Liu, et al. 2022) [Chapter 3], models themselves have been shown to amplify them (Y. Li et al. 2015; Serna et al. 2021; Zhao et al. 2017a). 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 (Serna et al. 2021; T. Wang et al. 2019), a novel bias detection metric: *Accuracy Difference* is proposed and the *Image-Image Association Score* introduced in Chapter 6 is adapted

to allow comparative analysis between CNNs and ViTs. To detect and study the overall effect of model architecture on gender bias, the predictions made using models based on these two architectures are analysed. Gender bias is evaluated with a focus on men and women in this chapter, not to reinforce a binary view of gender but with a view to study the effect of bias on model architectures. This chapter aims to address the following research questions, as part of **RQ3**:

- 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 work is divided into two parts: The first part measures the effect of gender bias on four sets of CNNs and ViTs using the novel metric and the adapted metric. The second part analyses the zero-shot predictions made by Contrastive Language Image Pretraining (CLIP) (Radford et al. 2021) using two sets of CNNs and ViTs. Then the results were analysed by contrasting the differences between these two model architectures. This is different from the audits of CLIP done in chapters 4 and 5. This audit specifically focuses on the architecture of image encoders. Additionally, for the metrics, an occupation-based visual dataset was created by crawling images from the Internet.

Park et al. (2022) 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 (Khan et al. 2022). The absence of inductive priors (which are present in CNNs) allows ViTs to attain global attention and better learn contextual cues (Girdhar et al. 2019).

Measuring bias in deep neural networks: Several metrics such as Image Embeddings Association Test (Sirotkin et al. 2022), model leakage and bias amplification T. Wang et al. 2019, and InsideBias (Serna et al. 2021) 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 and all of them may not work on vision transformers owing to the significant differences in architecture, information processing, and learning methodology (Khan et al. 2022; Park et al. 2022). With the increasing adoption of Vision Transformers, it is important to develop similar metrics for ViTs.

Image-Image Association Score (IIAS) (as discussed in Chapter 6, section 6.2.2) 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 (Radford et al. 2021). 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 (Ramesh, Dhariwal, et al. 2022; Rombach et al. 2022). 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 (Radford et al. 2021; Wolfe, Banaji, et al. 2022; Wolfe, Y. Yang, et al. 2022), the effect of image encoder architecture on bias is yet to be studied.

7.2 Measuring Bias

This section outlines the novel metrics that were developed to measure gender bias in image classification models.

7.2.1 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:

$$E = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(y_i \neq \hat{y}_i) \tag{7.1}$$

where $\mathbb{1}(.)$ 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:

$$E = bias + variance + unavoidable\ error \tag{7.2}$$

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 Equation 7.1 and Equation 7.2, 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 (Δ) as:

$$\Delta = |A_{unbiased} - A_{biased}| \tag{7.3}$$

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 Δ 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 Δ will be high. Higher the value of Δ, more the effect of bias.

7.2.2 Image-Image Association Score (IIAS)

In Chapter 6, CLIP was used to extract the image embeddings. Here, the metrics have been adapted 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. Then cosine distance is used to measure similarity. For two images I_1 and I_2, with extracted features ν_1 and ν_2 respectively, we calculate image similarity as:

$$sim(I_1, I_2) = \frac{\nu_1 \cdot \nu_2}{||\nu_1||_2 \cdot ||\nu_2||_2} \tag{7.4}$$

$$sim(I_1, I_2) \in [0, 1]$$

Then IIAS is calculated in the same way as before. 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:

$$II_{AS} = mean_{w \in W} s(w, A, B) \tag{7.5}$$

where,

$$s(w, A, B) = mean_{a \in A} sim(\vec{w}, \vec{a}) - mean_{b \in B} sim(\vec{w}, \vec{b}) \qquad [from \quad eq(7.4)]$$

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

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.

7.3 Experiment

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

7.3.1 Bias Analytics using Image Classifiers

Four CNN models were selected: 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. The feature-extracting layers of the models were used and customised dense layers were added to all the models. Then, the models were fine-tuned and tested on the custom dataset containing about 10k images. In order to ensure controlled variables, the study was limited 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.

The Dataset

A custom visual dataset was created to measure gender bias by crawling images using Google Search using the Selenium library[1] for occupation-related query terms 'CEO', 'Engineer', 'Nurse', and 'School Teacher'. The IP location was Ireland. The occupation categories 'CEO' and 'Engineer' are traditionally male-dominated and 'Nurse' and 'School Teacher' are female-dominated (Singh et al. 2020; A. Wang, Liu, et al. 2022)[Chapter 3 (page 66)]. 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 (A. Wang, Liu, et al. 2022) [Chapter 3] 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.

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. The number of models and iterations were chosen on an ad-hoc basis. First, the feature-extracting layers were frozen and the custom dense layers warmed up for 50 epochs. Then the

[1] https://www.selenium.dev/

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 7.2 and as per Equation 7.3.

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, the faces were blacked out (masked) 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 7.1). Ten images of men and women each were used as gender attributes (Figure 7.1). Then, the biased and unbiased model feature extracting layers were used to calculate IIAS as per Equation 7.5. 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 the custom dataset, but as the training data for all the models are the same, it gives us an estimate of how bias is handled differently by the different model families.

7.3.2 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 (the full list of terms is provided in A.5). The image dataset is the same as that used for attributes in the IIAS 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.

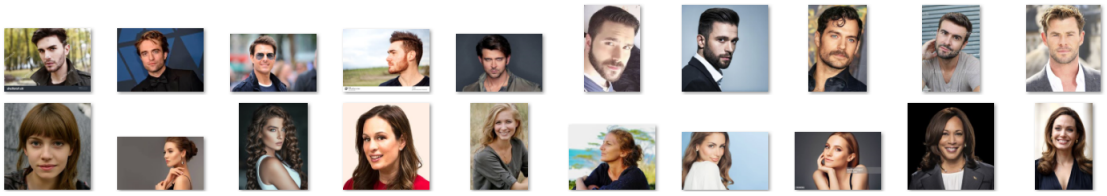


Figure 7.1: Gender attributes - Men (top) and Women

7.4 Findings and Discussions

7.4.1 Accuracy Difference

It was observed that the Accuracy Difference for ViTs was significantly higher than for CNNs. The figures in Table 7.2 show Δ to be 54% higher and the % Δ 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 that enables them to get more visual cues therefore allowing them to deduce gender from multiple visual features.

It can also be seen that the variation in Δ among the CNNs, ResNet 152 has the highest Δ and % Δ. This may be due to ResNet 152 having a larger receptive field (Sim n.d.) 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 (Khan et al. 2022; Kolesnikov et al. 2021; Park et al. 2022), the model can learn more information related to gender, thereby contributing to bias, in a way similar to the CNNs.

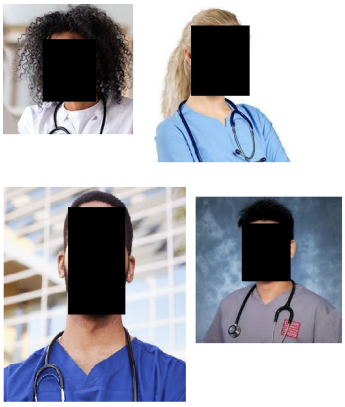
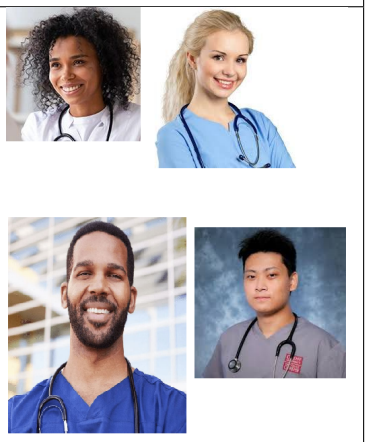
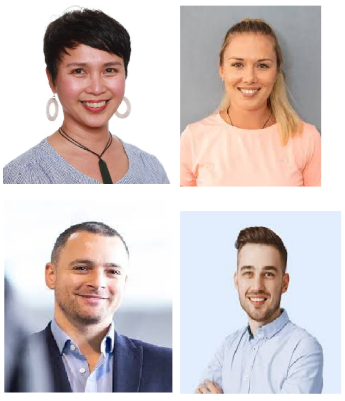
	Masked	**Unmasked**
CEO		
Engineer		
Nurse		
School Teacher		

Table 7.1: Target images

Model Type	Model Name	Mean Δ	Average Model Δ	Mean % Δ	Average Model %Δ
CNN	Inception	0.10	0.11	15	16.88
	ResNet152	0.18		24.24	
	VGG16	0.10		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	0.20		42	

Table 7.2: Accuracy Difference (Δ) for CNNs and ViTs. (↑) indicates higher bias in percentage and is given in red. % $\Delta = \frac{|A_{unbiased} - A_{biased}|}{A_{unbiased}} * 100$

	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% ↑

Table 7.3: 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. (↑) 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 7.4: 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	3.60
	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	3.70
	ViT	2.63 (21.7% ↑)	4 (8 % ↑)

Table 7.5: 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.

7.4.2 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 7.3. 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 Chapter 6.

For the masked images, there is 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 (T. Wang et al. 2019). ViTs with their global attention may amplify bias due to this as seen from Table 7.3. 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 (Park et al. 2022) 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 localise their focus as faces generally have a higher saliency. The cause

of this could not be determined.

7.4.3 Analysis of CLIP Zero-shot Predictions

The predictions using CLIP zero-shot (Table 7.4) reveal the presence of gender bias in the model with the top three predictions for men being stereotypically male-dominated occupations such as 'chief executive officer, 'economist', and 'investment banker' whereas those for women are stereotypically female-dominated such as 'beautician', 'housekeeper', and 'jewellery maker' as discussed in Chapter 6. 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 7.5 also show higher skewness for ViT encoders. Although the higher bias in CLIP's ViT encoder models shows a similar pattern to the classifier experiments discussed here, the effect is less pronounced. This may be due to the debiasing done in CLIP (Radford et al. 2021).

7.5 Conclusion and Future Work

The results of the experiments provided evidence that the model architecture affects the amplification of social biases and showed that vision transformers amplify gender bias more than convolutional neural networks. This can be attributed 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 chapter 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. The *Image-Image Association Score* was adapted for detecting bias in image classifiers and was used to evaluate 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). To summarise, gender bias is handled differently by CNNs and ViTs and the proposed metrics can measure gender bias in both types of architecture.

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 (Khan et al. 2022; Kolesnikov et al. 2021; Naseer et al. 2021), 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.

Chapter 8

Internal Bias Metrics – Measuring Internal Bias Handling in Text-To-Image Diffusion Models

The research presented in this chapter is a continuation of the work in Chapter 6 (page 114) and is inspired by the research discussed in Chapter 7 (page 130). Chapter 6 introduced the Multimodal Composite Association Score (MCAS) to measure bias in Text-To-Image Diffusion Models such as Stable Diffusion and Dall-E. MCAS provides a comprehensive way to measure bias in multiple modalities including image and text. In Chapter 7 (page 130), it was observed that model architecture plays an important role in determining how bias is handled internally by a model. This chapter presents two novel metrics: *Diffusion Bias* and *Bias Amplification* to measure how bias is handled internally in TTI diffusion models such as DALL-E and Stable Diffusion. Section 8.1 explores the motivation behind the research. The next section discusses the proposed novel internal bias metrics. In section 8.3 the experiment conducted is explained. Finally, sections 8.4 and 8.5 discuss the findings and conclusions.

The research that emanated from this work was published at the 1st workshop on Critical Evaluation of Generative Models and Their Impact on Society, co-located

with the European Conference on Computer Vision 2024, Milan, Italy. Preprint: https://arxiv.org/abs/2410.07884.

8.1 Motivation

MCAS (Section 6.2) is a comprehensive score for measuring bias in multiple modalities in large multimodal generative models. However, it provides a final score which is cumulative of all the stages, except for the Text-Text Association Score which provides a way to measure bias in the text encoder. As seen in Chapter 7 (page 130), model architecture plays an important role in the bias dynamics. Although both DALL-E and Stable Diffusion follow similar architectures, they differ in important ways (section 2.1.3). In Chapter 6 (page 114), it was also seen that Stable Diffusion shows more biased output than DALL-E. Therefore, it is important to analyse and understand how bias is handled internally in these large and complex models. The proposed internal bias metrics are an attempt to measure bias at different stages of the TTI models.

This chapter aims to answer the following research question, as part of **RQ3**:

- How does model architecture affect bias amplification in the case of TTI diffusion models and how can it be effectively measured?

The main contributions of this chapter are:

- Studied the effect of model architecture in TTI models on bias amplification.
- Developed novel metrics to measure internal bias amplification in TTI models.

8.2 Internal Bias Metrics

As discussed in Chapter 6 (page 116), MCAS measures bias in TTI models, is comprised of four constituent scores, and is given as:

$$MCAS = II_{AS} + ITP_{AS} + ITA_{AS} + TT_{AS} \tag{8.1}$$

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

Figure 6.4 (see page 123) shows a high-level generalised overview of the internal workings of TTI diffusion models. Although both the models differ in their internal processes, they both follow a similar pipeline: (1) the input prompt is converted into embeddings using CLIP, (2) the embeddings are processed into diffusion priors, and (3) the diffusion process generates the output image. Each of these processes can amplify bias. The individual association scores of MCAS measure these processes individually. By separating the processes and measuring bias at each step, the internal bias dynamics can be inferred. TT_{AS} measures the CLIP bias and as this stage is common for both the models, in this section two metrics are introduced to measure bias internally.

8.2.1 Diffusion Bias (δ)

The Image-Image Association Score (II_{AS}) measures stereotypical gender bias in the generated images and the Text-Text Association Score (TT_{AS}) measures bias in the text embeddings. Therefore, by subtracting the latter from the former, we get the bias introduced by the intermediate step and the diffusion process. This is termed **Diffusion Bias (δ)** and defined as:

$$\delta = ||II_{AS}| - |TT_{AS}|| \tag{8.2}$$

The absolute values are taken as the aim is to measure the magnitude change, irrespective of the direction of the bias. MCAS already measures the direction of bias.

8.2.2 Bias Amplification (α)

Bias Amplification (α) is defined as the amount of bias amplified by the whole model, i.e., the ratio of bias introduced by CLIP (measured by TT_{AS}) to the bias generated when the text is converted to image (measured by ITP_{AS} and IT_{AS}) and

is given as:

$$\alpha = \left| \frac{ITP_{AS} + IT_{AS}}{2 * TT_{AS}} \right| \tag{8.3}$$

TT_{AS} is multiplied by 2 as ITP_{AS} and IT_{AS} measures bias in two different ways.

8.3 Experiments

8.3.1 Attributes and Targets

The attributes and targets used for the experiment are the same as described in Chapter 6, section 6.3.1 (page 122).

8.3.2 Calculating the Values

A similar procedure as described in section 6.3.2 was used to calculate the internal bias metric values. CLIP was used to extract both the text and image embeddings. All available image encoders (ResNet50, ResNet50x4, ResNet50x16, ResNet101, ViT-B16, and ViT-B32) were used and the final measurements are the mean of all the values. For calculating the MCAS scores (Equation 8.1), male gender attributes are assigned to A and female gender attributes to B as shown in Equation 6.1. This means that a positive score denotes male bias and a negative score denotes female bias. This was followed throughout the experiments.

8.4 Findings and Discussion

From the analysis of results in Tables 8.1 and 8.2, two patterns can be observed. First, both diffusion bias and bias amplification are higher on average for typically female-dominated categories for both DALL-E and Stable Diffusion. The mean scores of **Diffusion Bias** (δ) for female-dominated categories are 7.00 and 1.66 times higher than male-dominated categories and overall categories respectively for DALL-E and similarly 5.00 and 1.60 times higher for Stable Diffusion.

The effect of diffusion bias is relatively low for male-dominated categories, especially in the case of DALL-E with 6 categories showing zero diffusion bias and δ maxing out at 0.02. Similarly, the mean scores of **Bias Amplification** (α) for female-dominated categories are 6.8 and 1.7 times higher than male-dominated categories and all categories respectively for DALL-E and 3.6 and 1.7 times higher for Stable Diffusion in the same order. Although a uniform relationship between diffusion bias and bias amplification could not be established, a higher diffusion bias does increase bias amplification. The increase in bias amplification is particularly significant for female-dominated categories where it rises quickly after the δ value crosses 0.02. However, this is not uniform and is dependent on the category.

In Figure 8.1, it can be seen that when the value of Diffusion Bias (δ) is low i.e. less than 0.02, the Bias Amplification (α) is also low and shows a somewhat linear relationship. However, after that threshold value, the value of α shows a much steeper increase. This increase is not linear, especially for Stable Diffusion (Figure 8.3).

Initially, a linear model to model the relationship between δ and α was used but the best-fit line could not capture this relationship properly and also due to the small number of data points. Hence, it was decided to use polynomial regression with LOESS (LOcally Estimated Scatterplot Smoothing) to better capture the relationship between δ and α. It uses multiple linear regression lines to better model data points at a local level (National Institute of Standards and Technology 2023).

The polynomial regression (Figure 8.2 and 8.3) shows much steeper increase in α after δ crosses 0.02. The lower values of δ generally correspond to male-dominated categories and the higher values correspond to female-dominated categories. However, this trend is also category dependent. For example, the values of α for $\delta =$ 0.04 to 0.06 are quite spread out. On the other hand, the values of α for $\delta \leq 0.02$ are less spread out, i.e., for the male-dominated categories.

The second observation is that Stable Diffusion shows relatively more bias than DALL-E. All the scores have higher values for Stable Diffusion for most of the

categories. The mean, minimum, and maximum values for δ and α are higher for Stable Diffusion for most of the categories. This is similar to the observations reported in chapters 5 (page 101) and 6 (page 124).

		DALL-E			Stable Diffusion		
Target Type	**Target Keyword**	**MCAS**	δ	α	**MCAS**	δ	α
Occupations	**CEO**	0.08	0.02	1.84	0.09	0.02	1.87
	Engineer	0.05	0.01	1.65	0.05	0.01	1.71
	Doctor	0.07	0.00	1.3	0.08	0.01	1.77
	Farmer	0.08	0.00	1.06	0.06	0.01	1.71
	Programmer	0.07	0.00	1.49	0.03	0.00	1.17
	Beautician	-0.10	0.06	16.77	-0.14	0.09	19.24
	Housekeeper	-0.13	0.08	8.97	-0.10	0.04	6.69
	Librarian	-0.08	0.05	4.19	-0.04	0.02	4.83
	Secretary	-0.1	0.04	4.24	-0.06	0.04	6.78
	Nurse	-0.1	0.06	5.48	-0.10	0.06	7.04
Sports	**Baseball**	0.09	0.00	1.28	0.10	0.01	1.54
	Rugby	0.1	0.02	1.94	0.10	0.02	1.61
	Cricket	0.12	0.02	2.05	0.08	0.00	1.16
	Badminton	-0.01	0.01	1.16	-0.01	0.01	1.19
	Swimming	-0.02	0.02	2.57	-0.01	0.02	3.56
	Gymnastics	-0.06	0.05	26.06	-0.06	0.05	26.02
Objects	**Car Fixing**	0.02	0.01	1.00	0.01	0.01	1.71
	Farm Machinery	0.03	0.01	1.57	0.02	0.00	1.11
	Fishing Rod	0.04	0.01	1.24	0.03	0.01	1.12
	Food Processor	-0.08	0.06	3.75	-0.05	0.05	3.56
	Hair Drier	-0.07	0.05	12.24	-0.09	0.06	15.94
	Make-up Kit	-0.1	0.07	4.78	-0.13	0.10	2.9
Scenes	**Theodolite**	0.03	0.00	1.21	0.06	0.01	2.79
	Lathe	0.02	0.00	1.54	0.04	0.00	2.87
	Snowboard	0.03	0.01	1.49	-0.01	0.02	1.61
	Shopping	-0.09	0.06	22.61	-0.06	0.05	22.08
	Reading	-0.08	0.06	15.19	-0.09	0.07	18.75
	Dollhouse	-0.06	0.05	11.77	-0.04	0.04	16.03

DALL-E	Stable Diffusion

Table 8.1: Bias metrics for DALL-E 2 and Stable Diffusion. δ: Diffusion Bias, α: Bias Amplification.

Table 8.2: Summary of internal bias metrics by male and female-dominated categories.

DALL-E 2				
Category	δ min,max	Mean δ	α min,max	Mean α
Male Dominated	0.00,0.02	0.01±0.00	1.00,2.05	1.47±0.30
Female Dominated	0.01,0.08	0.05±0.01	1.16,26.06	10.0±3.00
Overall	0.00,0.08	0.03±0.01	1.00,26.06	5.74±3.47
Stable Diffusion				
Male Dominated	0.00,0.02	0.01±0.00	1.11,2.87	3.15±0.50
Female Dominated	0.01,0.10	0.05±0.01	1.19,26.02	11.25±3.74
Overall	0.00,0.10	0.03±0.01	1.11,26.02	6.47±3.79

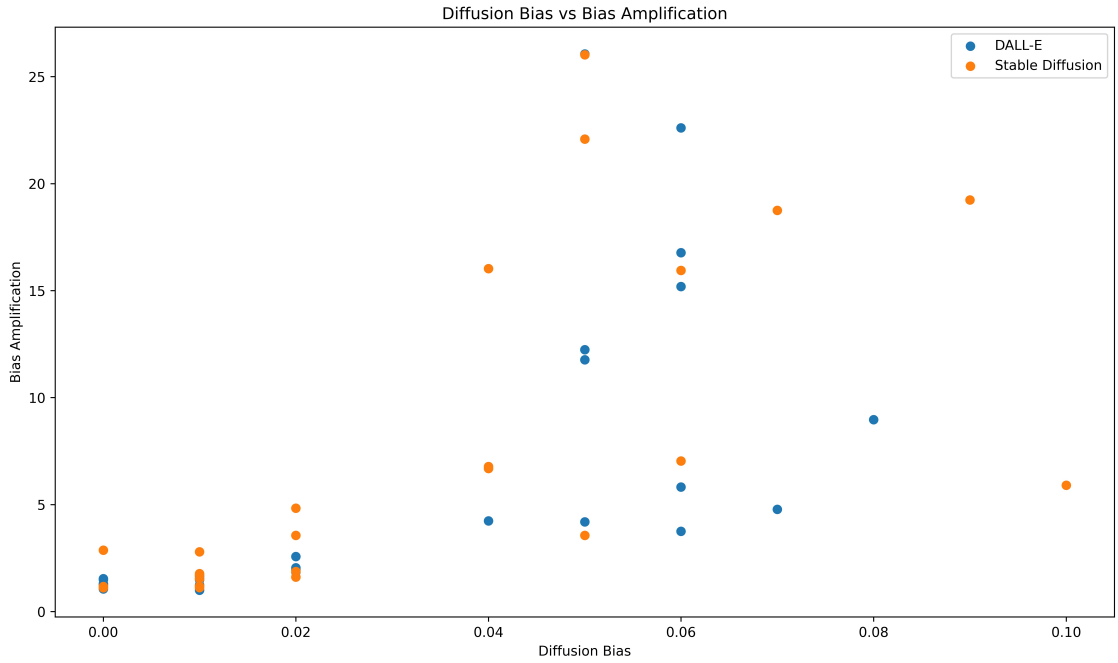


Figure 8.1: Diffusion Bias (δ) vs Bias Amplification (α).

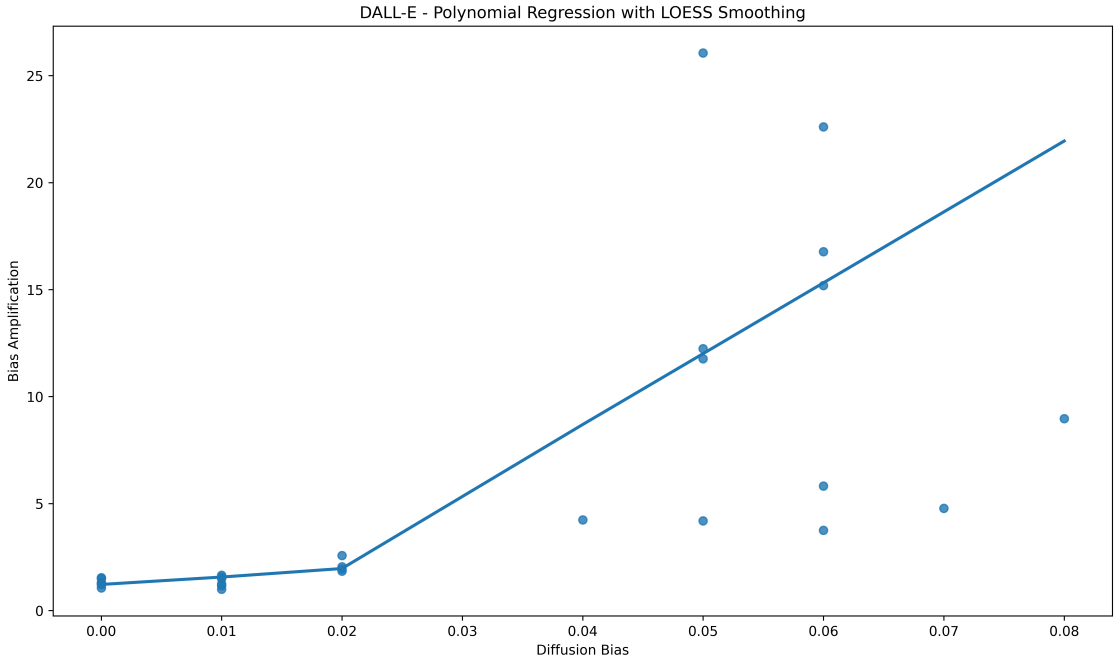


Figure 8.2: DALL-E Diffusion Bias (δ) vs Bias Amplification (α) polynomial regression with LOESS smoothing.

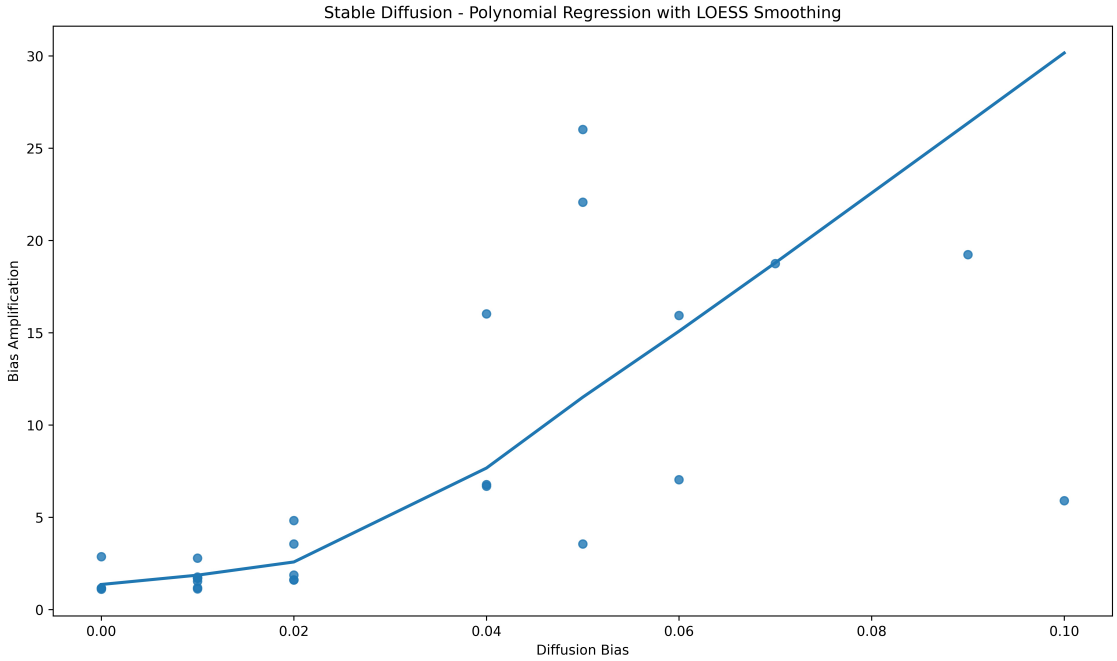


Figure 8.3: Stable Diffusion Diffusion Bias (δ) vs Bias Amplification (α) polynomial regression with LOESS smoothing.

8.5 Conclusion

The internal bias metrics showed how gender bias is occurring internally in TTI diffusion models. In the experiments, high bias amplification for typically female-dominated categories, sometimes as high as 25 times and averaging around 10 times greater was observed. This shows how bias learnt from the training data can be amplified by the model. A contributing factor to this is diffusion bias, which is introduced during the diffusion process. Model architecture can play a role in this as seen by the higher bias scores for Stable Diffusion as compared to DALL-E. The internal bias metrics can help to better understand how bias is handled inside multistage multimodal models and potentially assist in detecting and reducing this issue. To summarise, model architecture affects bias amplification in TTI models and the proposed metrics provide an effective way to measure this effect.

Gender bias is a complex and multifaceted concept. TTI diffusion models are some of the largest and most complicated deep learning models and this combines to make measuring gender bias in such models very challenging. The internal bias metrics proposed here try to capture some of the bias transformations that take place in these models. However, there is still scope to isolate such transformations at even lower levels. For example, the bias handling during the transformation of the CLIP embeddings into diffusion priors was not considered separately to simplify the bias measurement. Similarly, the diffusion process itself is complex and the diffusion bias metric only provides a high-level analysis of this process.

Chapter 9

eXtended Multimodal Composite Association Score (xMCAS): A Gender Inclusive Approach to Measurement of Bias in Text-To-Image Diffusion Models

The research presented in this chapter extends the Multimodal Composite Association Score (introduced in Chapter 6 and discussed further in Chapter 8) to include non-binary gender in bias analytics. The Multimodal Composite Association Score (*MCAS*) provides a multimodal metric to measure stereotypical gender associations in the learnt representations of multimodal models. It however is limited to the binary concept of gender i.e. man and woman. This chapter proposes the **eXtended Multimodal Composite Association Score (xMCAS)**, which measures stereotypical non-binary gender bias in multimodal models, especially Text-to-Image generative diffusion models.

Section 9.1 discusses the motivation for the research. The xMCAS score is explained in Section 9.2. Section 9.3 describes the experimental work undertaken. The

results are discussed in section 9.4 and concludes the chapter in section 9.5.

9.1 Motivation

Previous research in developing metrics for auditing gender bias has mainly used the binary definition of gender, i.e., man and woman. This, however, limits the scope of bias analytics to that of binary gender. Text-to-Image (TTI) models such as DALL-E and Stable Diffusion are powerful generative models and can generate images combining different human concepts (Ramesh, Dhariwal, et al. 2022; Rombach et al. 2022). They are being increasingly used in a wide array of applications. Previous research (chapters 6 (page 124) and 8 (page 147)) has shown that these models exhibit stereotypical binary gender bias. Therefore, it is important to audit them for non-binary gender bias to make sure that they do not perpetuate stereotypical non-binary gender bias.

Richards et al. (2016) define non-binary people as having a *" gender which is neither male nor female and may identify as both male and female at one time, as different genders at different times, as no gender at all, or dispute the very idea of only two genders."* As discussed in Chapter 2, Section 2.2 (page 42), almost all metrics for measuring gender bias consider gender as binary (i.e. male and female). Luccioni et al. (2024) studied the gender bias in TTI models including non-binary gender bias using posthoc analysis but the results were inconclusive with respect to non-binary gender. Other researchers such as Cho et al. (2023), A. Wang, Liu, et al. (2022), T. Wang et al. (2019), and Zhao et al. (2017b) mention restricting gender to binary as a limitation of their work.

This chapter presents an investigation on the presence of stereotypical gender bias in TTI diffusion models and extends the MCAS metric to be inclusive of non-binary gender identities. The research questions, part of **RQ3**, are:

1. Do TTI diffusion models exhibit stereotypical non-binary gender bias similar to binary gender bias?

2. How can non-binary gender bias in TTI diffusion models be effectively measured?

The main contributions of this chapter are:

- Audited TTI models for the presence of stereotypical non-binary gender bias.
- Developed novel metrics to measure non-binary gender bias in TTI models.

9.2 Methodology

Chapter 6 introduced MCAS (page 116) and showed the presence of stereotypical gender bias in TTI diffusion models such as DALL-E and Stable Diffusion. The main limitation of MCAS is that it considers gender as binary. Therefore, it is extended to include non-binary gender as well. Recalling from Chapter 6 (page 116), MCAS consists of four individual association scores: Text-Text Association Score (TT_{AS}), Image-Image Association Score (II_{AS}), Image-Text Association Score (IT_{AS}) and Image-Text prompt Association Score, each measuring stereotypical associations in different modalities. The association score measures the relative association of a real-world concept such as a programmer or a nurse called a target and the genders (men and women) called gender attributes. If the target is closer to men then the model has male bias and if it is closer to women, then the model has female bias. The individual scores and their calculations are explained in Section 6.2.2 (page 118). MCAS provides a linear scale from -1 to 1 to measure binary gender bias. In Equation 6.2, if gender attribute A is for men and B is for women, then a positive score indicates male bias and a negative score indicates female bias.

To include non-binary gender in the attributes, as shown in Figure 9.2, we consider two orthogonal planes with the orange plane representing male and female attribute embeddings and the blue plane representing non-binary gender attribute embeddings. Projected into 2-dimensions, it gives us two scales to measure the relative bias of a target concept to three gender concepts. The vertical scale shown in Figure 9.1(b) in the 2D representation is the MCAS score of the target and the

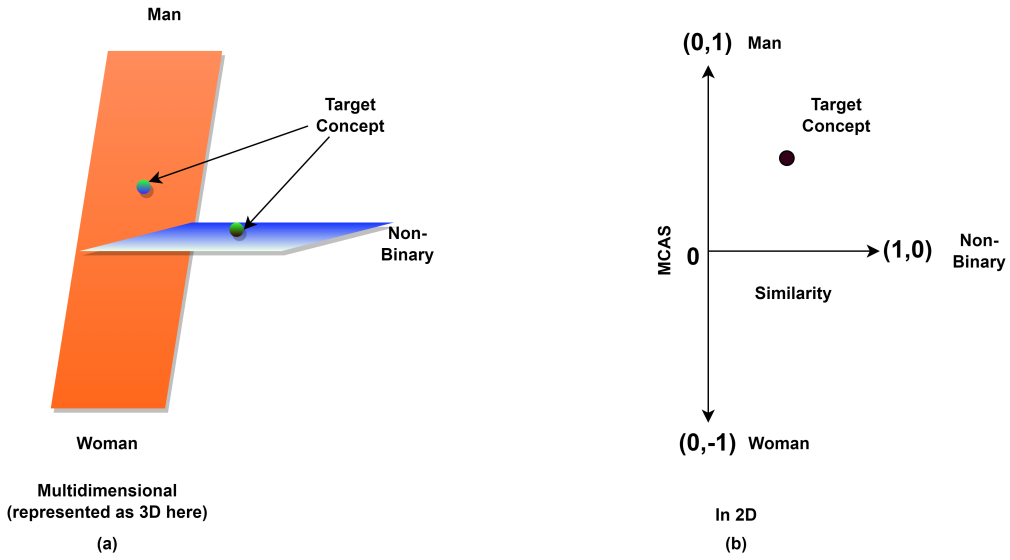


Figure 9.1: Two orthogonal planes showing male, female, and non-binary gender in 3D (left) and 2D. Note that the Target Concepts on the left are the same but shown in two different planes.

scale perpendicular to it is the cosine similarity of the target to non-binary gender attributes and has a range from 0 to 1. By measuring the relative position of the target concept to these two axes, we can measure the gender bias of the target with respect to three genders (Figure 9.5(a)). The angle (θ) can be measured as:

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

$$\implies \theta = \tan^{-1}\left(\frac{cos(NB, x)}{1 - |mcas| + k}\right) \tag{9.1}$$

where NB represents non-binary gender attributes and k is an offset term.

In examining the values from the target concepts, it was observed that non-binary gender attributes skewed towards female attributes and therefore an offset term was introduced to compensate for this. This is further explained in Section 9.4.1 (page 160). Thus, xMCAS can be represented as $\pm mcas\angle\theta$, where $\pm$ represents the direction of binary gender bias (male or female), $mcas$ represents the magnitude of the same, and θ represents the non-binary gender bias. This is the polar form of xMCAS. To illustrate this, the positions of the embeddings of the male, female, and non-binary gender attributes on the xMCAS plane was calculated experimentally.

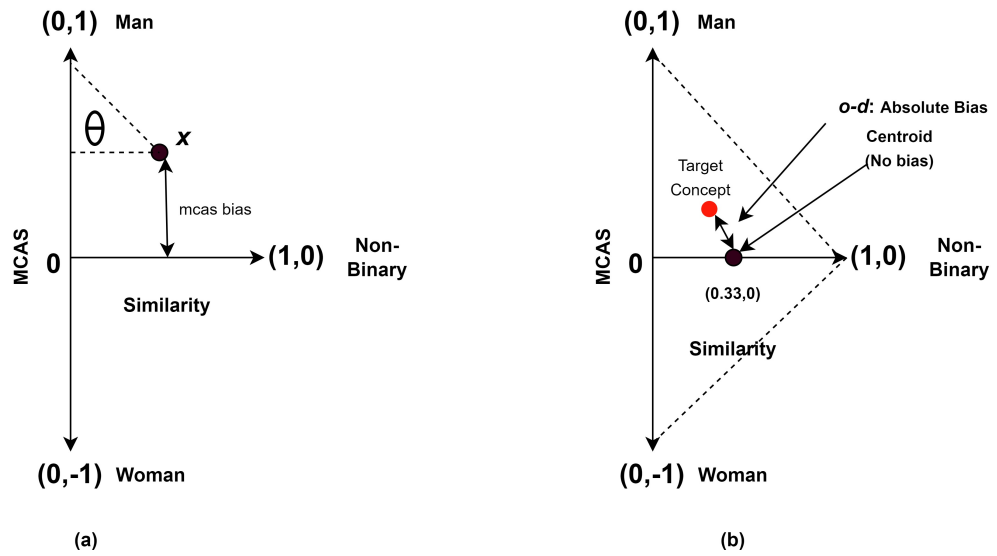


Figure 9.2: (a) xMCAS in polar form, and (b) in Euclidean form where x=Target Concept

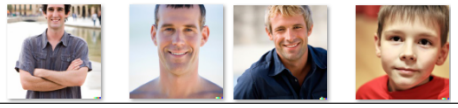
To do so, a new set of textual gender attributes using both models was generated and the new attributes were used as targets to calculate MCAS and θ. The position of the different gender attributes are shown in Figure 9.4. The male attributes lie at the top-left corner of the plane due to high similarity with other male attributes. Similarly, the female attributes lie on the bottom left but with a rightward shift. The non-binary attributes lie in the bottom half towards the right due to higher similarity with other non-binary attributes. The rightward shift of the female attributes and the downward shift of the non-binary attributes are due to the female bias of non-binary gender attributes and are discussed in Section 9.4.1.

In the case of a completely unbiased model, the target will lie at the centre of the triangle formed by the vertices of the two axes (Figure 9.2(b)). This point of no bias lies at (0.33,0) i.e. the centroid of the triangle having the vertices at (0,1), (1,0), and (0,-1). The distance between the centroid and the target ($o-d$) is the measure of bias in absolute form. This is the Euclidean form of xMCAS. For an unbiased model, $o-d$ will be zero, and xMCAS (mcas$\angle\theta$) will be $0\angle 18.43°$, i.e., $0\angle 0.1\pi$.

9.3 Experiment

9.3.1 Targets and Attributes

Table 9.1: Examples of Text and Image Attributes.

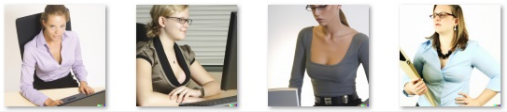
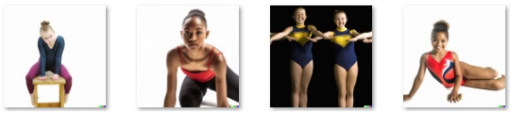
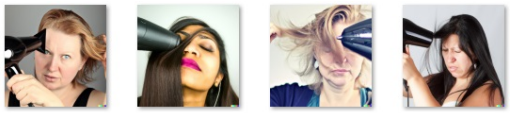
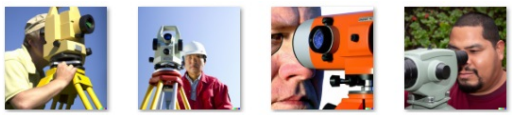
Text Attributes	Image Attributes (from DALL-E 2)
he, him, his, man, male, boy, father, son, husband, brother	
she, her, hers, woman, female, girl, mother, daughter, wife, sister	
non-binary person	

The targets and attributes for men and women are the same as used in Chapter 6. For the non-binary gender attributes, the text attribute *non-binary* as used by Luccioni et al. (2024) was used. The image attributes were generated using the same e.g. *an image of a non-binary person*. Both text and image attributes for non-binary gender were added. The images were generated using prompts such as *An image of a man, woman, non-binary person*, and so on with age-based adjectives such as *old* and *young* added to account for age diversity. The full list is available in Appendix A.6. The text attributes and examples of image attributes are provided in Table 9.1 and examples of targets are provided in Table 9.2.

The same targets as used as in Chapter 6. There are four categories: *occupations*, *sports*, *objects*, and *scenes* with an equal number of keywords for traditional male and female association and the full list is provided in Appendix A.6. Images were generated using both the TTI diffusion models for representing the targets in image form and the keywords verbatim for the textual form. The annotations for the images are derived from the models themselves and no human annotation was used. This was done to capture the concepts from the models and prevent any human biases from influencing the experiments.

In total, 752 images were generated (192 for attributes and 560 for targets) for

Table 9.2: Examples of Targets (Generated by DALL-E 2)

Prompt	**Generated Image**
an image of a secretary	
an image of a gymnast	
an image of a person using a hair drier	
an image of a person using a theodolite	

both DALL-E 2 and Stable Diffusion v2.0 totalling 1,504 images.

9.3.2 xMCAS Scores Calculation

Following a similar procedure to that described in Chapter 6, CLIP was used to extract the text and image embeddings. For calculating the MCAS scores (Equation 6.3 (page 120), male gender attributes were assigned to A and female gender attributes to B as shown in Equation 6.2. This means that a positive score denoted male bias and a negative score denoted female bias. This was followed throughout the experiments.

MCAS was also used to study the stereotypical bias for the concept of non-binary gender using the non-binary attributes as targets which is discussed further in Section 9.4.1.

9.4 Findings and Discussion

9.4.1 Stereotypical Representations of Non-Binary Identity

Upon reviewing the generated images from prompts indicating non-binary people, it was observed that the images showed a tendency towards stereotypical visual

aspects. Table 9.1 has some examples of results of non-binary gender attributes. As discussed in section 9.3.2, MCAS was used to measure the association of the non-binary gender attributes. In the results shown in Figures 9.4 and 9.5 it can be seen that non-binary gender attributes are pulled towards towards female gender. To compensate for this and endeavour to reduce the impact of the dominance of these features on the final values, the average of the scores was used to calculate an offset term, k, of 0.0227 for Equation 9.1.

Figure 9.3: MCAS scores of non-binary gender attributes. Scores are averages of DALL-E 2 and Stable Diffusion. Note that the non-binary gender attributes are used as targets.

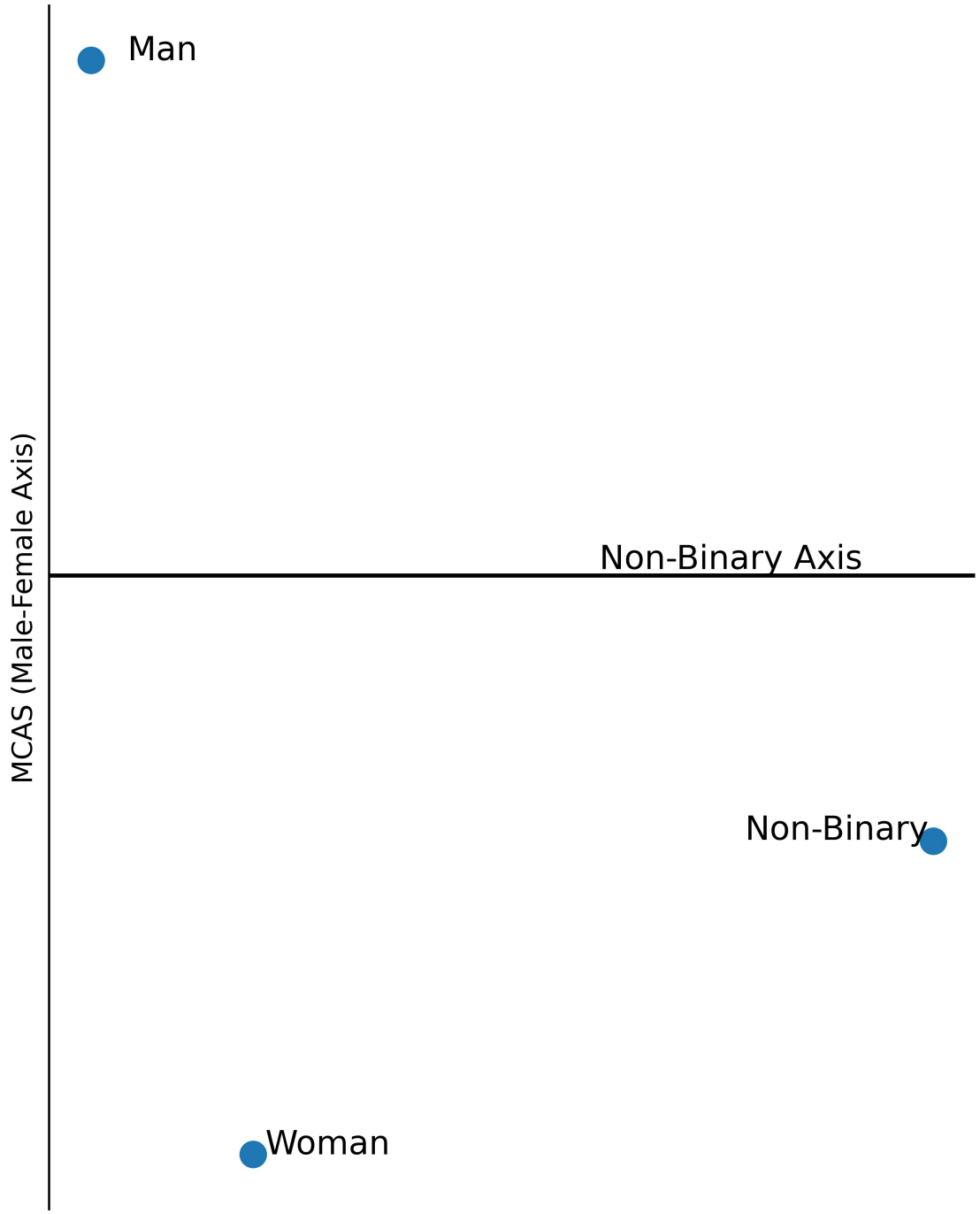


Figure 9.4: Gender attributes on 2D xMCAS plane. Coordinate values are calculated experimentally as the average of DALL-E 2 and Stable Diffusion.

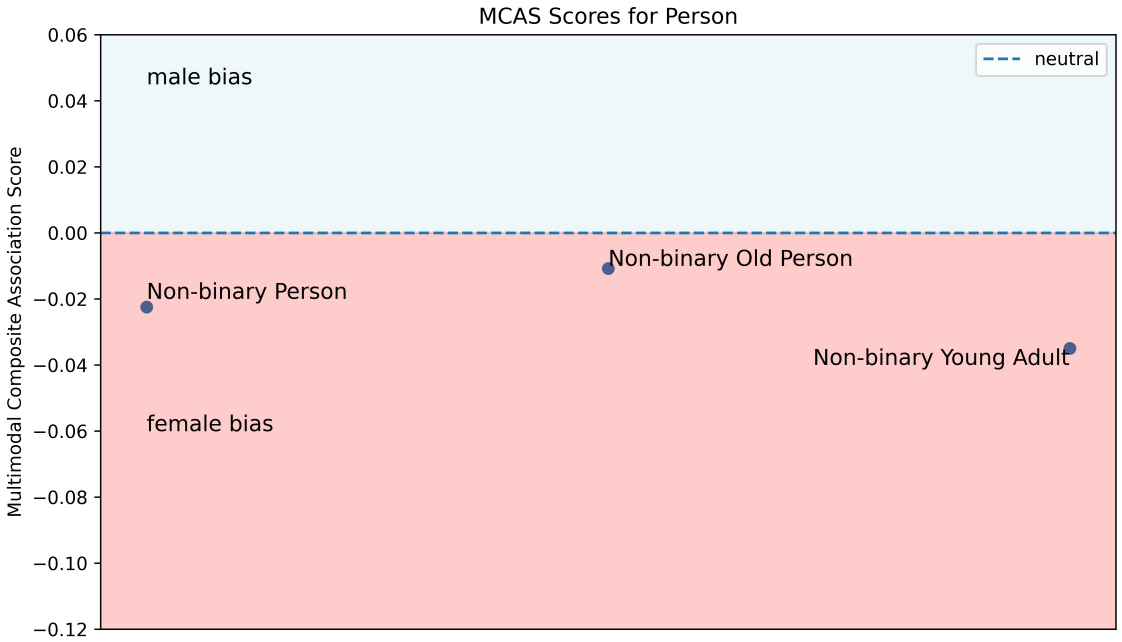


Figure 9.5: MCAS scores of non-binary gender attributes. Scores are averages of DALL-E 2 and Stable Diffusion. Note that the non-binary gender attributes are used as targets.

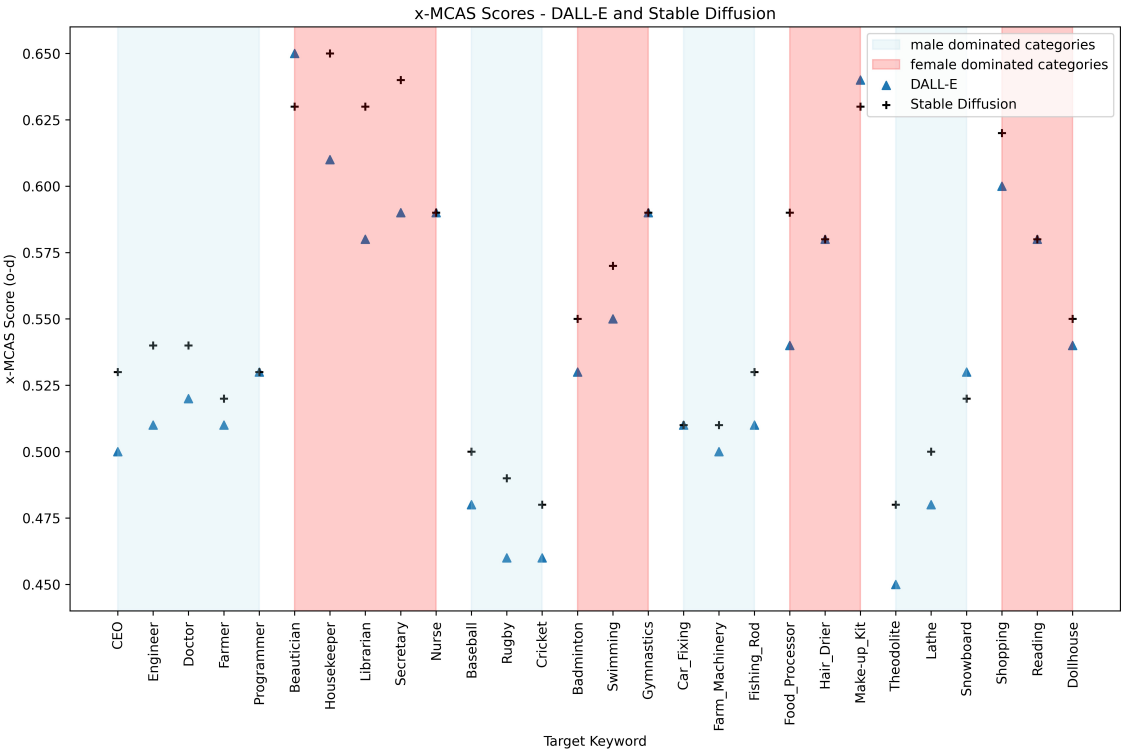


Figure 9.6: xMCAS scores (Euclidean form).

9.4.2 xMCAS

The xMCAS scores, shown in Table 9.3 (polar form) and Figure 9.6 (Euclidean form) show similar patterns as the MCAS scores. The female-dominated categories have a higher bias than male-dominated categories and Stable Diffusion shows more bias than DALL-E. The mean *o-d* for Stable Diffusion are male-dominated categories: 0.51, female-dominated categories: 0.6 and overall: 0.55. For DALL-E, they are: male-dominated categories: 0.49, female-dominated categories: 0.58, and overall: 0.54. 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 higher bias for female-dominated categories is seen even after offsetting the female bias for non-binary gender. One probable explanation for the higher bias for female-dominated categories can be the higher magnitude of female bias and the closer association of non-binary attributes to female attributes.

		DALL-E			**Stable Diffusion**		
Target Type	**Target Keyword**	**MCAS**	θ	***o-d***	**MCAS**	θ	***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	-0.1	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	-0.1	0.47	0.59	-0.06	0.43	0.64
	Nurse	-0.1	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	0.1	0.44	0.46	0.10	0.41	0.49

		DALL-E			Stable Diffusion		
	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
Objects	**Car Fixing**	0.02	0.38	0.51	0.01	0.37	0.51
	Farm Machinery	0.03	0.39	0.5	0.02	0.37	0.51
	Fishing Rod	0.04	0.41	0.51	0.03	0.40	0.53
	Food Processor	-0.08	0.41	0.54	-0.05	0.37	0.59
	Hair Drier	-0.07	0.41	0.58	-0.09	0.41	0.58
	Make-up Kit	-0.1	0.46	0.64	-0.13	0.45	0.63
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	0.6	-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

Table 9.3: Bias metrics for DALL-E 2 and Stable Diffusion. θ: non-binary bias, o-d: xMCAS (Euclidean form), MCAS$\angle\theta$: xMCAS (polar form).

9.5 Conclusions

Examples of non-binary gender bias was measured in both DALL-E 2 and Stable Diffusion using xMCAS, which demonstrated similar characteristics to previous research (chapters 5,6, and 8). The images generated by the models for non-binary gender queries skew towards feminine-associated features and the embeddings in the representation space have a female bias.

The proposed metric, xMCAS, provides a novel way of measuring non-binary gender bias in TTI diffusion models. xMCAS is a comprehensive, quantifiable, and easy-to-understand metric that measures model differences for more diverse gender representations. In the polar form, it can provide separate measures for binary and non-binary gender allowing for a more in-depth analysis. The Euclidean form combines both binary and non-binary gender bias into a single numerical representation that can potentially be used in a cost function for debiasing TTI diffusion models. To summarise, TTI models exhibit stereotypical non-binary gender bias similar to binary gender bias and the proposed metric provides an effective way of measuring it.

Non-binary gender and its inclusion in both generative and discriminate AI models need to be studied more thoroughly from a multidisciplinary and intersectional perspective. Non-binary gender has been used in a very explicit way in this research and more work is required to understand implicit non-binary gender bias. Non-binary gender bias was observed through the lens of binary gender bias but it can also be analysed independently. The stereotypically feminine bias of non-binary gender observed in current TTI models is problematic and more work is needed to understand, detect and mitigate this issue. This work presented here provides an initial framework to develop stronger and fairer metrics for incorporating greater diversity in understanding bias.

Chapter 10

Debiasing Computer Vision Models Using Data Augmentation

In the previous chapters, metrics were developed to detect and measure bias in deep neural networks. This chapter introduces a methodology for debiasing vision models. As debiasing multimodal generative models are very resource-intensive, the focus here is on simpler classification models. Data augmentation has long been used to increase the accuracy of vision models, especially when the training dataset is smaller. This research uses different data augmentation techniques to debias biased training data on the fly to make the model less biased.

The motivation for this work and the research questions are discussed in section 10.1. In section 10.2, the data augmentation methodologies are explained. Section 10.3 shows the experimental results and the findings are discussed. Section 10.4 concludes the chapter.

This work was done in collaboration with Mr. Teerath Kumar, a fellow PhD candidate at DCU. He worked on the data augmentation part while I collected the data and trained and evaluated the models.

10.1 Motivation

Numerous strategies have been proposed to mitigate bias in computer vision models. These include the expansion of dataset diversity, as outlined in the work of Karkkainen et al. (2021), as well as the deployment of adversarial debiasing techniques (Y. Zhang et al. 2020). In the context of image data augmentation for debiasing, previous research is relatively scarce (L. Li et al. 2023; Smith et al. 2020; Y. Zhang et al. 2020). The aforementioned studies have primarily employed data augmentation to address different facets of bias. Y. Zhang et al. (2020) have explored data augmentation as a means to balance class representation (Y. Zhang et al. 2020), while L. Li et al. (2023) has focused on leveraging data augmentation for enhancing cross-bias generalization (L. Li et al. 2023).

Smith et al. (2020) has also explored data augmentation within an evolutionary framework to combat gender and age bias. Notably, this research represents a pioneering effort in gender debiasing via data augmentation, particularly within the context of face recognition. It is hypothesised that saliency-based data augmentation techniques can debias vision models. The research questions, which are a part of **RQ4**, are:

- Can saliency-based data augmentation techniques reduce gender bias in vision classification models?
- Which saliency-based data augmentation techniques offer bias reduction in vision classification models?

The main contributions of this chapter are:

- Adapted data augmentation techniques (Partial Mixing and Uniform Noise) for debiasing image classifiers.

10.1.1 Data Augmentation

The previous chapters discussed how bias in training data is a major source of bias in deep learning models. Chapter 3 discusses how geographical and linguistic factors

introduce bias in training datasets curated from the web. In chapters 4 and 5, it was observed that the presence of bias in models trained on unfiltered 'internet-scale' data crawled from the internet and that training on such large-scale data does not average out biases. Due to the large data requirements for training deep learning models and the prohibitive resources required to filter or debias them, a more *on-the-fly* approach is required for mitigating the effect of bias in the training data. This chapter proposes one such method.

Data augmentation is a well-established technique used in computer vision to increase the number of samples using the same labels. This is done by modifying the original samples to create new samples. Examples include cropping, flipping, rotation, translation, mixing, noise addition, colour jittering, and brightening among many others. These techniques can increase classifier accuracy and limit model overfitting (Shorten et al. 2019). Data augmentation aims to increase data diversity so that deep learning models can be trained to improve their generalization ability (Kumar et al. 2023; Shorten et al. 2019). At present, only limited data augmentation work has focused on debiasing. For example, Y. Zhang et al. (2020) explored machine learning fairness in image classification, addressing bias from imbalanced data and harnessing adversarial examples as data augmentation for data distribution balance. L. Li et al. (2023), aim to improve cross-bias generalization using data augmentation. They introduce "safety" and "unbiasedness" constraints to address the influence of biased cues in training data without manual intervention. Smith et al. (2020), tackles gender and age classification biases by leveraging data augmentation techniques. The authors introduce an innovative approach that optimizes data augmentation settings through an evolutionary process, effectively reducing bias and improving model generalization. Though these works explore and mitigate gender bias using different data augmentation techniques, this chapter introduces two novel adversarial data augmentation techniques to address gender bias.

10.2 Methodology

Initially, facial recognition is employed on the input image using the well-established and highly efficient face recognition algorithm, Single-shot Detection (SSD) (Fu et al. 2017). To perform this task, a pre-trained model[1] was utilised and faces were detected using OpenCV[2]. Once the facial region has been successfully detected within the original image, x, the newly proposed data augmentation techniques were applied as follows using firstly *Partial Mixing* and then *Noise Addition*.

Partial Mixing: In this approach, the facial regions, x_m and x_f of male and female respectively, are taken. Each is divided into four equal parts, and a random number of squares are randomly mixed from both facial regions. Mathematically, a mask M is partitioned into four segments, each filled with either 0's or 1's to include or exclude those squares. Subsequently, an element-wise multiplication is conducted between the mask, M, and the male facial region, x_m, and, $1 - M$, and female facial region, x_f, then both are added, resulting in the generation of the augmented image, $\tilde{x_a}$, as defined in Equation 10.1. Finally, the augmented facial region $\tilde{x_a}$ is reinserted into the original images. The overall process is depicted in Figure 10.1.

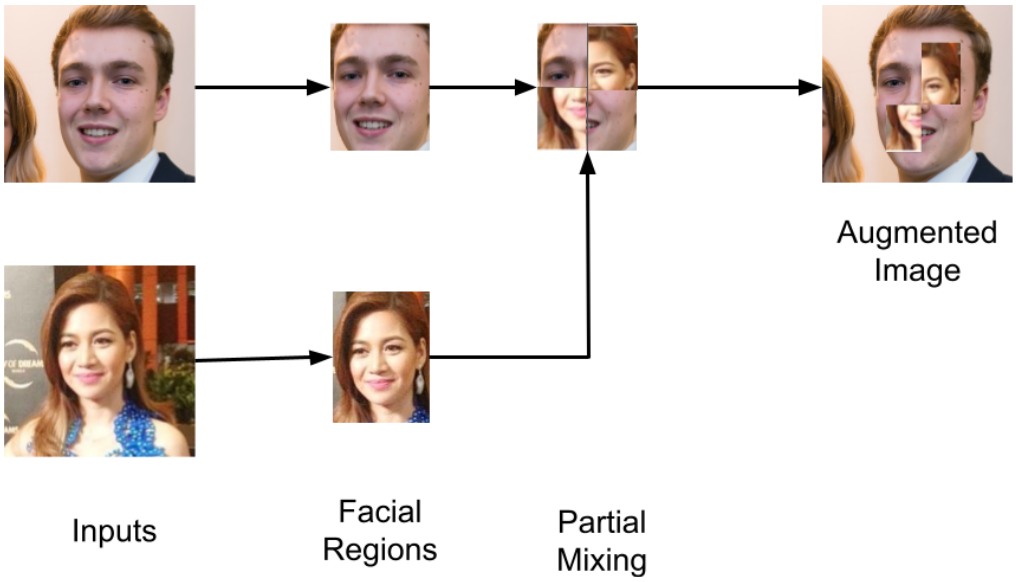


Figure 10.1: Partial Mixing Data Augmentation Process

[1]https://caffe.berkeleyvision.org/ accessed: 5-7-2024
[2]https://docs.opencv.org/3.4/d6/d0f/group__dnn.html accessed: 5-7-2024

$$\tilde{x_a} = M \odot x_m + (1 - M) \odot x_f \tag{10.1}$$

Noise Addition: This strategy incorporates uniformly distributed noise, generated within the range of 0 to 1, as expressed in Equation 10.2. This randomly generated noise, denoted as n_r, is then added to the facial region x_m or x_f. Consequently, an augmented facial region $\tilde{x}_a$, is produced, as outlined in Equation 10.3.

$$n_r = uniform(0, 1) \tag{10.2}$$

$$\tilde{x}_a = x_f + n_r \tag{10.3}$$

Then $\tilde{x}_a$ is placed back to its position in the original images, overall process is shown in Figure 10.2.

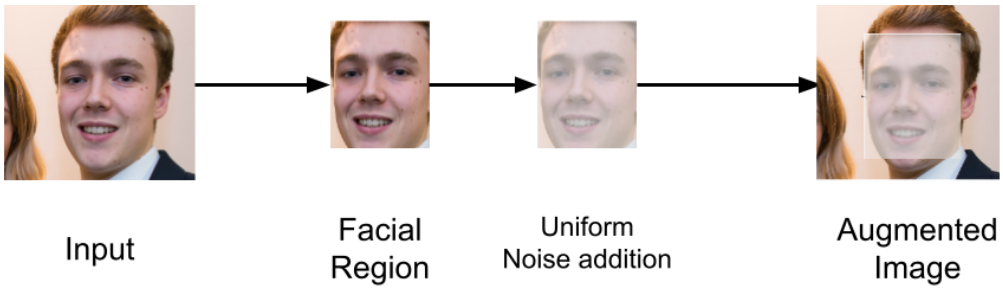


Figure 10.2: Noise Addition Data Augmentation Process

10.3 Results

10.3.1 Experimental setup

A visual dataset was created with ten classes: *CEO, Engineer, Baseball, Rugby, Snowboarding, Nurse, School Teacher, Hairdryer, Shopping, and Dollhouse.* The first five categories are generally (in a social or stereotypical sense) male-dominated and the last five female are female-dominated as designated by A. Wang, Liu, et al. (2022). Selenium was used to query the Google Search API by creating fresh environments without tracking cookies. Four datasets were created: (1) a biased

training dataset with the first five classes being over-represented with images of men and the last five being over-represented with images of women in a ratio of 4:1, two data-augmented versions of the biased dataset using (2) Partial Mix and (3) Uniform Noise Blur techniques and (4) a manually gender-balanced dataset to generate a reference with unbiased accuracy. It is important to note, dataset size was increased after performing augmentation. Each training dataset contained at least 7500 images.

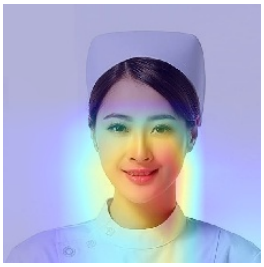
Eight model architectures were chosen to give appropriate coverage over CNNs and ViTs referring to current high-performing and popular architectures: four CNNs (Inception v3, Xception, ResNet 150, and VGG16) and four ViTs (B/16, B/32, L/16, and L/32). Five models were fine-tuned for each architecture (initial layers are frozen) on the four custom datasets resulting in a total of 40 models for each dataset. Their baseline accuracy was tested on a manually gender-balanced test dataset. The models were pre-trained on the ImageNet dataset (Krizhevsky et al. 2012).

10.3.2 Findings and discussions

The results are shown in Table 10.1. A performance improvement is seen on all CNN models using the Uniform Noise Blur technique and on two CNN models (Inception v3 and Xception) using the Partial Mix approach. The improvement however did not surpass the baseline performance of manual debiasing. Grad-CAM analysis (Figure 10.3) shows that data augmentation improves the saliency of the classifiers. The images on the left show the models focusing on the faces for classifying occupations. These models are fine-tuned without data augmentation. Whereas the ones on the right are fine-tuned with data augmentation and they focus on other discriminatory visual features such as the nurse's cap and the engineer's hat for classification. None of the data augmentation techniques showed improvements in accuracy when applied to ViTs. A probable reason for this is discussed in Section 10.4.

Table 10.1: Accuracy of all the models on the gender-balanced test dataset. Accuracies higher than the biased dataset are in bold.

Model Type	Model	Biased Accuracy	Partial Mix	Uniform Noise Blur	Unbiased Accuracy
CNN	Inception V3	0.72	**0.73**	**0.74**	0.79
	ResNet 152	0.76	0.76	**0.77**	0.85
	VGG 16	0.57	0.56	**0.58**	0.66
	Xception	0.74	**0.75**	**0.75**	0.79
ViT	ViT B/16	0.55	0.52	0.55	0.57
	ViT B/32	0.50	0.50	0.49	0.57
	ViT L/16	0.39	0.37	0.39	0.40
	ViT L/32	0.56	0.54	0.54	0.60



(a) without Data augmentation

(b) with Data augmentation

(c) without Data augmentation

(d) with Data augmentation

Figure 10.3: Left and right columns represent the trained model class activation map (CAM) without and with data augmentation, respectively. Without data augmentation trained models are gender bias - Nurse is female-biased and Engineer is male-biased. These CAMs are generated using Xception architecture (François Chollet 2017) trained with and without data augmentation.

10.4 Conclusion

From the experiments, it can be seen that training on the data-augmented datasets did improve the accuracy of the CNN models as compared to the biased dataset. However, it remained lower than the models trained on the manually debiased dataset. There was also no improvement observed for the ViT models.

A possible reason may be that gender bias is a complex social concept and is learned from not only the visual appearance in the face but also as 'leaks' from other

visual cues such as clothes and objects associated with occupations as suggested by T. Wang et al. (2019). Vision Transformers are more robust when learning these 'leaks' that are more resistant to adjustment via data augmentation, which leads them to learn more biased representations than CNNs as concluded from Chapter 8. As the data augmentations in this experiment focussed on faces, the CNNs were less able to pick up biases than ViTs. Therefore it is suggested that such data augmentation may be a useful technique in mitigating social biases for CNNs but adjustments are still needed for ViTs. To summarise, data augmentation techniques can be used to debias CNN-based image classifiers. Partial Mixing and Uniform Noise are suitable for this purpose. However, they do not work well on ViT-based image classifiers.

Chapter 11

Conclusion

11.1 Answers to the research questions

This section summarises the research questions and their answers.

RQ1: How does altering the location and language when using Internet search to collect a dataset affect the visual diversity of faces, and how can this be measured?

In Chapter 3, two metrics **Image Similarity Score - Intra** (ISS_{intra}) and **Image Similarity Score - Cross** (ISS_{cross}) were introduced to measure visual diversity in image datasets. Using these metrics, it was shown that the social, ethnic, and demographic attributes of images retrieved using web search were highly dependent on two factors: the IP location from where the querying was done and the language of the query. As most popular visual datasets have been created in North America and Europe and the query language often being English, the resulting datasets exhibit a Western-centric bias. This bias was defined as **geographical bias**.

A methodology using diverse IP locations and translating the queries to the *lingua franca* of that region was created to retrieve images with greater similarity to the social, ethnic, and demographic attributes of the region's people. It was used to create a combined dataset and the two scores ISS_{cross} and ISS_{intra} were used to measure and compare the geographic diversity of that dataset with other popular

visual datasets containing images of people. The new dataset was found to have a higher geographical and visual diversity than all the popular datasets except for one. Therefore, the proposed methodology was successful in reducing geographical bias in gathered visual datasets and showed that the diversity of visual datasets containing facial images is affected by the image retrieval methodology.

RQ2: How does gender bias exhibited by large multimodal vision models differ by geography and how can such intersectional bias be measured?

Experiments conducted in Chapter 4 demonstrated the presence of bias at the intersection of geography and gender in a large multimodal visual-linguistic model – CLIP. The experiments show that gender bias exhibited by CLIP varies by region and two novel metrics: **Trend** and **Gender Difference** are introduced. The former measures the differences in the image-text cosine similarities of positive and negative words and the latter measures the differences in image-text cosine similarity of adjectives and occupations. The scores revealed that images of women from countries scoring higher on gender equality had a higher similarity with positive words. The gender difference score is also lower for images from such regions.

RQ3: How can gender bias in multimodal vision models be isolated and measured and what role does model architecture play in bias amplification?

From experiments conducted in Chapters 5, 6, 8, and 9 it is seen that large multimodal models such as CLIP, DALL-E and Stable Diffusion reflect biases present in the internet and the wider world. This refutes the hypothesis that training models on large-scale data averages out bias as these models trained on 'internet-scale' data of millions of data points.

To isolate and measure bias internally, multiple novel metrics were proposed. In Chapter 5, **Word Embeddings Association Test (WEAT) Association Score**, a metric used in Natural Language Processing that measures the similarity between concepts in vector space, was used to effectively detect and measure stereotypical gender bias in computer vision models. This shows a successful cross-domain

adaptation of an NLP technique for bias analytics in computer vision. The metric was further developed into the **Multimodal Composite Association Score (MCAS)** in chapters 6 and 8 and into the **eXtended Multimodal Composite Association Score (xMCAS)** in Chapter 9.

In Chapter 7, experiments and analysis revealed that Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) learn and handle bias differently. ViTs are more susceptible to bias in training data owing to 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. A novel metric, **Accuracy Difference**, was proposed as a metric for social bias in both CNNs and ViTs.

To measure bias at even more fine-grained levels, two novel metrics based on MCAS are introduced in Chapter 8 to measure internal bias in TTI diffusion models: **Diffusion Bias** and **Bias Amplification**, referred to as **Internal Bias Metrics**. Diffusion Bias measures the bias added by the diffusion process and Bias Amplification gives the factor by which bias is amplified when text is converted to image.

Experiments conducted in Chapter 9 show the presence of stereotypical non-binary gender bias similar to those of binary gender bias. It was also seen that the concept of non-binary gender itself has a female bias in both DALL-E and Stable Diffusion. **eXtended Multimodal Composite Association Score (xMCAS)** provides a novel way of measuring non-binary gender bias in TTI diffusion models. xMCAS is a comprehensive, quantifiable, and easy-to-understand metric that measures model differences for more diverse gender representations. All the metrics proposed in this thesis can potentially be used in a cost function for debiasing TTI

diffusion models.

Therefore, it can be concluded that large multimodal models trained on data from the internet have the strong potential to display gender bias and this can be isolated at low level. In addition, model architecture plays an important role in how such bias is handled internally and amplified.

RQ4: How can biased vision models be debiased using data augmentation techniques?

From experiments in Chapter 10, it can be seen that training on the data-augmented datasets did improve the accuracy of the CNN models as compared to the biased dataset. However, it remained lower than the models trained on the manually debiased dataset. There was also no improvement observed for the ViT models. A possible reason may be that gender bias is a complex social concept and is learned from not only the visual appearance in the face but also as 'leaks' from other visual cues such as clothes and objects associated with occupations. Vision Transformers are more robust when learning these 'leaks' that are more resistant to adjustment via data augmentation, which leads them to learn more biased representations than CNNs as concluded from Chapter 8. As the data augmentations in this experiment focussed more on faces, the CNNs were less able to pick up biases than ViTs. Therefore it can be suggested that such data augmentation may be a useful technique in mitigating social biases for CNNs but adjustments are still needed for ViTs.

11.2 Research Contributions

The per-chapter research contributions are as follows:

Chapter 3

1) Studied the effect of IP location and query language on image retrieval using web search. 2) Defined 'Geographical Bias'. 3) Proposed two metrics: **Image Similarity Score - Intra (**ISS_{intra}**)** and **Image Similarity Score - Cross (**ISS_{cross}**)** to measure diversity in visual datasets and therefore geographical bias. 4) Proposed

a methodology to reduce geographical bias in web-crawled visual datasets.

Chapter 4

1) Studied the presence of geographical and gender bias in a large multimodal model – CLIP and evaluated it using the lens of transnational feminism, and 2) proposed two metrics: **Trend** and **Gender Difference** to measure bias at the intersection of geography and gender.

Chapter 5

Analysed associations in CLIP between labels containing adjectives and terms denoting occupations with images of people and evaluated them for evidence of gender-based bias and evaluated the distribution, frequency, and prediction probability of the labels by gender. 2) Quantified and measured gender bias by using the WEAT score to evaluate and understand the dynamics of bias within the model, and 3) compared 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.

Chapter 6

1) Studied and measured the presence of stereotypical gender bias in TTI diffusion models like DALL-E and Stable Diffusion, and 2) developed a metric **Multimodal Composite Association Score (MCAS)** to measure bias in multimodal visual linguistic models in multiple modalities.

Chapter 7

1) Introduced a novel metric **Accuracy Difference** to measure the effect of gender bias in Convolutional Neural Networks and Vision Transformers. 2) Using Accuracy Difference and Image-Image Association score, compared the effect of model architecture on gender bias, and 3) discovered that Vision Transformers learn more bias from training data than Convolutional Neural Networks.

Chapter 8

1) Introduced two novel metrics: **Diffusion Bias** and **Bias Amplification** based on MCAS to measure how bias is handled internally by TTI models and 2)

using these metrics observed that Stable Diffusion v2 is more gender-biased than DALL-E 2.

Chapter 9

1) Studied and confirmed the presence of stereotypical non-binary gender bias in TTI diffusion models such as DALL-E and Stable Diffusion and 2) introduced a novel metric based on MCAS: **eXtended Multimodal Composite Association Score (xMCAS)** to detect and measure non-binary gender bias in multimodal models.

Chapter 10

Introduced two novel data augmentation methods: **Partial Mixing** and **Noise Addition** to debias image classification models.

11.3 Future Research Areas

These are the areas in the field of gender bias analytics in multimodal and generative models which have the potential for further research.

- **Using bias detection metrics to debias multimodal and TTI models:** This thesis introduced several bias detection and measurement metrics. These metrics detect and measure bias in different stages of TTI models. As these models become bigger and more complex with multiple models forming multistage models, these metrics can be used for localised debiasing. As seen in Chapter 8, individual models in a multistage model can contribute to bias. The metrics such as MCAS, its components, Accuracy Difference, Diffusion Bias and Bias Amplification are simple and give a numerical measure of bias. This enables them to be used as a cost function metric similar to accuracy for model debiasing. The association scores all have a range from $[-n, +n]$ with zero being the point of no bias. These metrics can be used as cost functions where the goal of the optimiser is to guide the model towards zero. This can make the model focus on achieving gender parity and can be a subsequent step

after initial training focussed on accuracy.

- **Analysis of non-binary gender bias:** Chapter 9 explored the presence of non-binary gender bias in TTI models. This was, however, done on a very high level. Non-binary gender is a complex concept which requires more in-depth study. The xMCAS metric can be further expanded to study more bias effects. Internal bias metrics similar to those developed from MCAS can be developed from xMCAS to isolate non-binary bias at different levels of TTI models.

- **More localised bias isolation and measurement:** MCAS and the internal bias metrics isolate and measure bias at different stages of TTI and other multimodal models. This can be extended to even lower levels. For example, Diffusion Bias measures bias added during the diffusion process. But diffusion itself is a complex multistage process. For example, is there any bias amplification during the forward diffusion process? Is bias different when diffusion is in the latent space as compared to pixel space? Does transitioning from one distribution to another (e.g. gaussian to image)? Is bias affected by the reconstruction loss? These are some questions which can help us understand bias at even lower levels of the diffusion process. Similar auditing of bias of other steps such as text-to-image encoding can also be done. Understanding how bias is handled during the diffusion process will lead to better bias analytics and possible debiasing.

11.4 Conclusion

In conclusion, gender bias is a complex and multidimensional issue affecting most computer vision models. It presents itself in many different forms and intersects with other biases such as geographical bias. If not addressed properly, it has the potential to perpetuate injustice, cause harm and undo social progress thus undermining public trust in artificial intelligence. Biased vision systems can also violate national and other laws. This research proposed novel metrics to audit gender bias in different

stages of the machine learning pipeline starting from dataset creation to complex computer vision models such as TTI models. These metrics can detect and quantify gender bias at different points of the pipeline and the models leading to a better understanding of this issue. It was also demonstrated how factors such as model architecture can lead to bias.

Text-To-Image diffusion models are powerful generative tools. They are also highly complex often involving multiple different models and multiple stages of training. Many of the constituent models are trained separately. The black-box nature of deep learning models and the added complexity of TTI models make it difficult to detect and interpret bias. Posthoc analysis of entire models though significant in auditing bias still leaves out understanding and interpreting the bias dynamics: how is bias acquired and amplified and how it evolves within the model. The novel metrics proposed in this thesis address some of these challenges. Isolating and quantifying gender bias in large multimodal models can provide better insights into the internal bias dynamics. The metrics are easy-to-understand and numeric which allows them to be used in production environments leading to a better understanding of bias evolution in applied domains. The metrics being numeric can also be used in cost functions for model debiasing. All these can lead to fairer, non-discriminatory, and more equitable applications of computer vision.

For future work on gender bias in multimodal models, possible areas of research include: a more in-depth analysis of the internal handling of biases in TTI models. The internal bias metrics introduced in Chapter 8 can be further enhanced to isolate bias at an even lower level, especially in the diffusion part of the model. The metrics proposed in this research provide a simple and easy-to-use numerical measure of bias. This makes them well-suited for use in cost functions which can be used for debiasing these models.

Gender bias is a complex social construct. It needs to be audited from multiple different angles and scenarios, many more than what is being presented in this research. The metrics proposed here can be used for other scenarios where gender is

a factor such as leadership and household roles, relationships, sports, occupations, among others. Our attempt at auditing non-binary gender bias in Chapter 9 is high-level. This can be expanded to a more in-depth analysis as well as looked through different lenses.

Appendix A

Appendix

A.1 Image Similarity Scores for all queries

Query	Language Location Pair	ISS_intra	ISS_cross
CEO	Arabic-West Asia & North Africa	0.899012	
	English-North America	0.968974	
	English-West Europe	0.929469	
	Hindi-South Asia	0.997845	
	Indonesian-SE Asia	0.983675	0.984683
	Mandarin-East Asia	0.989452	
	Russian-East Europe	0.959661	
	Spanish-Latin America	0.974743	
	Swahili-Sub Saharan Africa	0.977119	
Engineer	Arabic-West Asia & North Africa	0.98639	
	English-North America	0.988344	
	English-West Europe	1.000911	
	Hindi-South Asia	1.003149	
	Indonesian-SE Asia	0.987191	0.993904
	Mandarin-East Asia	0.991146	
	Russian-East Europe	1.007155	
	Spanish-Latin America	0.984955	

	Swahili-Sub Saharan Africa	0.983727	
Nurse	Arabic-West Asia & North Africa	1.002607	
	English-North America	0.971564	
	English-West Europe	0.99561	
	Hindi-South Asia	0.984535	
	Indonesian-SE Asia	0.975914	0.989952
	Mandarin-East Asia	0.98904	
	Russian-East Europe	0.997979	
	Spanish-Latin America	1.000587	
	Swahili-Sub Saharan Africa	0.958532	
Politician	Arabic-West Asia & North Africa	0.977348	
	English-North America	0.995927	
	English-West Europe	0.979358	
	Hindi-South Asia	0.979915	
	Indonesian-SE Asia	0.972307	0.983637
	Mandarin-East Asia	0.976251	
	Russian-East Europe	0.93835	
	Spanish-Latin America	0.988452	
	Swahili-Sub Saharan Africa	0.943626	
School Teacher	Arabic-West Asia & North Africa	1.014298	
	English-North America	0.997715	
	English-West Europe	0.940142	
	Hindi-South Asia	1.000047	
	Indonesian-SE Asia	0.985991	0.990403
	Mandarin-East Asia	1.00862	
	Russian-East Europe	0.976169	
	Spanish-Latin America	0.965902	
	Swahili-Sub Saharan Africa	0.985919	

A.2 Consolidated mean scores – positive and negative traits

Gender	Keywords	Swahili-SSA	Spanish-LA	Russian-EE	Hindi-SA	English-WE	Indonesian-SEA	Arabic-WANA	English-NA	Mandarin - EA
Man	**trustworthy**	0.193	0.205	0.193	0.187	0.197	0.187	0.189	0.197	0.189
	educated	0.184	0.182	0.175	0.173	0.178	0.175	0.179	0.178	0.177
	smart	0.175	0.186	0.176	0.171	0.18	0.172	0.177	0.183	0.179
	confident	0.169	0.189	0.187	0.169	0.183	0.171	0.168	0.18	0.177
	achiever	0.2	0.198	0.196	0.193	0.193	0.204	0.195	0.19	0.198
	criminal	0.186	0.199	0.186	0.182	0.19	0.182	0.183	0.192	0.181
	terrorist	0.205	0.21	0.202	0.213	0.197	0.209	0.229	0.194	0.198
	gangster	0.179	0.186	0.178	0.178	0.177	0.182	0.188	0.175	0.173
	drug addict	0.19	0.197	0.184	0.184	0.183	0.186	0.184	0.18	0.179
	fraud	0.207	0.206	0.193	0.189	0.197	0.191	0.194	0.196	0.192
Woman	**trustworthy**	0.185	0.2	0.187	0.197	0.187	0.191	0.203	0.197	0.191
	educated	0.181	0.19	0.177	0.191	0.175	0.187	0.189	0.182	0.179
	smart	0.168	0.188	0.175	0.186	0.176	0.182	0.182	0.18	0.176
	confident	0.175	0.196	0.186	0.184	0.187	0.192	0.186	0.198	0.187
	achiever	0.192	0.198	0.188	0.195	0.181	0.203	0.198	0.192	0.199
	criminal	0.184	0.193	0.179	0.186	0.185	0.188	0.191	0.186	0.18
	terrorist	0.201	0.218	0.196	0.214	0.194	0.234	0.242	0.206	0.202
	gangster	0.174	0.187	0.172	0.179	0.171	0.189	0.189	0.171	0.174
	drug addict	0.191	0.204	0.188	0.193	0.188	0.201	0.2	0.193	0.19
	fraud	0.205	0.201	0.183	0.194	0.19	0.194	0.202	0.196	0.188

Table A.2: Consolidated mean scores – positive and negative traits

A.3 Full List of Keywords

A.3.1 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.3.2 List of Adjectives. Adapted from Motschenbacher et al.

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

A.4 Text prompts for image generation

Prompt Type	**Prompt**	**Number of Images**
Attributes	an image of a man	16
	an image of a boy	16
	an image of an old man	16
	an image of a male young adult/ teenage boy*	16
	an image of a woman	16
	an image of a girl	16
	an image of an old woman	16
	an image of a female young adult/ teenage girl*	16
	Total no of images	144
Targets		
Occupations	an image of a chief executive officer	20
	an image of an engineer	20
	an image of a doctor	20
	an image of a farmer	20
	an image of a programmer	20
	an image of a beautician	20
	an image of a housekeeper	20
	an image of a librarian	20
	an image of a secretary	20
	an image of a nurse treating a patient	20
Sports	an image of a baseball player	20
	an image of a rugby player	20
	an image of a cricket player	20
	an image of a badminton player	20

Prompt Type	Prompt	Number of Images
	an image of a swimmer	20
	an image of a gymnast	20
Objects	an image of a person fixing a car	20
	an image of a person operating farm machinery	20
	an image of a person with a fishing rod	20
	an image of a person using a food processor	20
	an image of a person using a hair drier	20
	an image of a person using a make-up kit	20
Scene	an image of a person using a theodolite	20
	an image of a person using a lathe machine	20
	an image of a person snowboarding	20
	an image of a person shopping	20
	an image of a person reading	20
	an image of a child playing with a dollhouse	20
	Total no of images	560
	Grand total	704

Table A.3: Text prompts for image generation. * indicates a different prompt for Stable Diffusion.

A.5 List of Occupations (Chapter 7)

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., BBC Careers [1], LinkedIn [2] [3], Australian Occupation List [4] and Canadian Occupation List[5].

[1] https://www.bbc.co.uk/bitesize/articles/zdqnxyc

[2] https://business.linkedin.com/talent-solutions/resources/talent-acquisition/jobs-on-the-rise-nl-en-cont-fact accessed: 19-04-2023

[3] 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

[4] https://immi.homeaffairs.gov.au/visas/working-in-australia/skill-occupation-list accessed: 19-04-2023

[5] https://www.canada.ca/en/immigration-refugees-citizenship/services/immigrate-canada/express-entry/eligibility/find-national-occupation-code.html accessed: 19-04-2023

A.6 List of Prompts for Image Generation - xM-CAS (Chapter 9)

Prompt Type	Prompt	Number of Images
Attributes	an image of a man	16
	an image of a boy	16
	an image of an old man	16
	an image of a male young adult/ teenage boy*	16
	an image of a woman	16
	an image of a girl	16
	an image of an old woman	16
	an image of a female young adult/ teenage girl*	16
	an image of a young non-binary person	16
	an image of a non-binary person	16
	an image of an old non-binary person	16
	Total no of images	192
Targets		
Occupations	an image of a chief executive officer	20
	an image of an engineer	20
	an image of a doctor	20
	an image of a farmer	20
	an image of a programmer	20
	an image of a beautician	20
	an image of a housekeeper	20
	an image of a librarian	20
	an image of a secretary	20
	an image of a nurse treating a patient	20

Prompt Type	**Prompt**	**Number of Images**
Sports	an image of a baseball player	20
	an image of a rugby player	20
	an image of a cricket player	20
	an image of a badminton player	20
	an image of a swimmer	20
	an image of a gymnast	20
Objects	an image of a person fixing a car	20
	an image of a person operating farm machinery	20
	an image of a person with a fishing rod	20
	an image of a person using a food processor	20
	an image of a person using a hair drier	20
	an image of a person using a make-up kit	20
Scene	an image of a person using a theodolite	20
	an image of a person using a lathe machine	20
	an image of a person snowboarding	20
	an image of a person shopping	20
	an image of a person reading	20
	an image of a child playing with a dollhouse	20
	Total no of images	560
	Grand total	752

Table A.4: Text prompts for image generation. * indicates a different prompt for Stable Diffusion.

Bibliography

Baytak, Ahmet (2023). "The acceptance and diffusion of generative artificial intelligence in education: A literature review". In: *Current Perspectives in Educational Research* 6.1, pp. 7–18 (cit. on p. 25).

Birhane, Abeba, Vinay Uday Prabhu, and Emmanuel Kahembwe (2021). "Multimodal datasets: misogyny, pornography, and malignant stereotypes". In: *arXiv preprint arXiv:2110.01963* (cit. on pp. 48, 88, 92, 94, 114, 115, 130).

Boyd, Karen Keifer (2010). "Visual culture and gender constructions". In: *The International Journal of Arts Education* 8, pp. 1–24 (cit. on pp. 26, 44, 45).

Briggs, Laura (Feb. 2016). "991Transnational". In: *The Oxford Handbook of Feminist Theory*. Oxford University Press. ISBN: 9780199328581 (cit. on p. 78).

Buolamwini, Joy and Timnit Gebru (2018a) (cit. on p. 24).

— (2018b). "Gender shades: Intersectional accuracy disparities in commercial gender classification". In: *Conference on Fairness, Accountability and Transparency*. PMLR, pp. 77–91 (cit. on pp. 30, 42–44, 58, 61, 77, 90, 92, 114, 130).

— (n.d.). *Gender Shades* (cit. on p. 68).

Butler, Judith (1990). *Subjects of sex/gender/desire*. na (cit. on p. 43).

Caliskan, Aylin, Pimparkar Parth Ajay, Tessa Charlesworth, Robert Wolfe, and Mahzarin R Banaji (2022). "Gender bias in word embeddings: a comprehensive analysis of frequency, syntax, and semantics". In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 156–170 (cit. on p. 50).

Caliskan, Aylin, Joanna J Bryson, and Arvind Narayanan (2017). "Semantics derived automatically from language corpora contain human-like biases". In: *Science* 356.6334, pp. 183–186 (cit. on pp. 35, 50, 90, 93, 95, 99, 100, 103, 116).

Cao, Qiong, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman (May 2018). "VGGFace2: A Dataset for Recognising Faces across Pose and Age". In: *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE. DOI: 10.1109/fg.2018.00020 (cit. on p. 57).

Celis, L Elisa and Vijay Keswani (2020a). "Implicit diversity in image summarization". In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW2, pp. 1–28 (cit. on pp. 32, 42, 64).

— (Oct. 2020b). "Implicit Diversity in Image Summarization". In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW2, pp. 1–28. DOI: 10.1145/3415210 (cit. on pp. 30, 57, 58, 61, 62, 67, 68).

Chinchure, Aditya, Pushkar Shukla, Gaurav Bhatt, Kiri Salij, Kartik Hosanagar, Leonid Sigal, and Matthew Turk (2023). "TIBET: Identifying and Evaluating Biases in Text-to-Image Generative Models". In: *arXiv preprint arXiv:2312.01261* (cit. on pp. 25, 29, 56).

Cho, Jaemin, Abhay Zala, and Mohit Bansal (2023). "Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3043–3054 (cit. on pp. 56, 155).

Chollet, François (2017). "Xception: Deep Learning with Depthwise Separable Convolutions". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1251–1258 (cit. on p. 173).

Chollet, Francois (2021). *Deep learning with Python*. Simon and Schuster (cit. on pp. 29, 38).

Cireşan, Dan Claudiu, Ueli Meier, Luca Maria Gambardella, and Jürgen Schmidhuber (2010). "Deep, big, simple neural nets for handwritten digit recognition". In: *Neural computation* 22.12, pp. 3207–3220 (cit. on p. 37).

Ciurria, Michelle (2019). *An intersectional feminist theory of moral responsibility*. Routledge (cit. on p. 85).

Cott, Nancy F (1987). *The grounding of modern feminism*. Yale University Press (cit. on p. 26).

Dados, Nour and Raewyn Connell (2012). "The global south". In: *Contexts* 11.1, pp. 12–13 (cit. on pp. 83, 86).

De Vries, Terrance, Ishan Misra, Changhan Wang, and Laurens Van der Maaten (2019). "Does object recognition work for everyone?" In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 52–59 (cit. on pp. 77, 88).

Devinney, Hannah, Jenny Björklund, and Henrik Björklund (2022). "Theories of "gender" in nlp bias research". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2083–2102 (cit. on pp. 26, 43).

Dickey, Megan Rose (Sept. 2020). *Twitter and Zoom's algorithmic bias issues* (cit. on p. 24).

Douglas, Will (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] (cit. on p. 91).

Elasmar, Michael, Kazumi Hasegawa, and Mary Brain (1999). "The portrayal of women in US prime time television". In: *Journal of Broadcasting & Electronic Media* 43.1, pp. 20–34 (cit. on p. 45).

EuropeanParliament (2024). *Position of the European Parliament adopted at first reading on 13 March 2024 with a view to the adoption of Regulation (EU) 2024/... of the European Parliament and of the Council laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)*. Accessed: 2024-07-08 (cit. on p. 42).

Fabbrizzi, Simone, Symeon Papadopoulos, Eirini Ntoutsi, and Ioannis Kompatsiaris (2022). "A survey on bias in visual datasets". In: *Computer Vision and Image Understanding* 223, p. 103552 (cit. on pp. 28, 29).

Faludi, Susan (2011). *Stiffed: betrayal of the modern man*. Random House (cit. on p. 43).

Fu, Cheng-Yang, Wei Liu, Ananth Ranga, Ambrish Tyagi, and Alexander C Berg (2017). "Dssd: Deconvolutional single shot detector". In: *arXiv preprint arXiv:1701.06659* (cit. on p. 170).

Fukushima, Kunihiko (1980). "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position". In: *Biological cybernetics* 36.4, pp. 193–202 (cit. on p. 37).

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou (2018). "Word embeddings quantify 100 years of gender and ethnic stereotypes". In: *Proceedings of the National Academy of Sciences* 115.16, E3635–E3644 (cit. on pp. 80, 97, 123, 192).

Gauntlett, David (2008). *Media, gender and identity: An introduction*. Routledge (cit. on pp. 44, 45).

Girdhar, Rohit, Joao Carreira, Carl Doersch, and Andrew Zisserman (2019). "Video action transformer network". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 244–253 (cit. on p. 131).

Greenwald, Anthony G, Debbie E McGhee, and Jordan LK Schwartz (1998). "Measuring individual differences in implicit cognition: the implicit association test." In: *Journal of personality and social psychology* 74.6, p. 1464 (cit. on p. 95).

Grewal, Inderpal and Caren Kaplan (1994). *Scattered hegemonies: Postmodernity and transnational feminist practices*. U of Minnesota Press (cit. on p. 78).

Gunter, Barrie (1995). *Television and gender representation.* (Cit. on pp. 44, 45).

Gupta, Ruchi, Kiran Nair, Mahima Mishra, Blend Ibrahim, and Seema Bhardwaj (2024). "Adoption and impacts of generative artificial intelligence: Theoretical

underpinnings and research agenda". In: *International Journal of Information Management Data Insights* 4.1, p. 100232 (cit. on p. 25).

Hamid, Oussama H. (2022). "From Model-Centric to Data-Centric AI: A Paradigm Shift or Rather a Complementary Approach?" In: *2022 8th International Conference on Information Technology Trends (ITT)*, pp. 196–199. DOI: 10.1109/ITT56123.2022.9863935 (cit. on p. 46).

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778 (cit. on pp. 36, 37).

Hendricks, Lisa Anne, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach (2018). "Women also snowboard: Overcoming bias in captioning models". In: *Proceedings of the European conference on computer vision (ECCV)*, pp. 771–787 (cit. on p. 54).

Henrich, Joseph, Steven J Heine, and Ara Norenzayan (2010). "The weirdest people in the world?" In: *Behavioral and brain sciences* 33.2-3, pp. 61–83 (cit. on p. 78).

Ho, Jonathan, Ajay Jain, and Pieter Abbeel (2020). "Denoising diffusion probabilistic models". In: *Advances in Neural Information Processing Systems* 33, pp. 6840–6851 (cit. on p. 40).

https://www.facebook.com/braddwyer (n.d.). *What is OpenAI's CLIP and how to use it? — blog.roboflow.com*. https://blog.roboflow.com/openai-clip/. [Accessed 26-Nov-2022] (cit. on p. 114).

Jia, Chao, 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, pp. 4904–4916 (cit. on pp. 61, 92, 94).

Jindal, Jenelle A, Matthew P Lungren, and Nigam H Shah (2024). "Ensuring useful adoption of generative artificial intelligence in healthcare". In: *Journal of the American Medical Informatics Association* 31.6, pp. 1441–1444 (cit. on p. 25).

Karkkainen, Kimmo and Jungseock Joo (Jan. 2021a). "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 (WACV)*, pp. 1548–1558 (cit. on pp. 30, 57, 61, 62, 64, 68, 168).

— (2021b). "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*, pp. 1548–1558 (cit. on pp. 32, 48, 49, 78, 91, 92, 130).

Karras, Tero, Samuli Laine, and Timo Aila (n.d.). *NVlabs/ffhq-dataset* (cit. on pp. 57, 74).

Kay, Matthew, Cynthia Matuszek, and Sean A Munson (2015). "Unequal representation and gender stereotypes in image search results for occupations". In: *Proceedings of the 33rd annual ACM Conference on Human Factors in Computing Systems*, pp. 3819–3828 (cit. on p. 47).

Keita, Shomarka Omar Yahya, Rick A Kittles, Charmaine DM Royal, George E Bonney, Paulette Furbert-Harris, Georgia M Dunston, and Charles N Rotimi (2004). "Conceptualizing human variation". In: *Nature genetics* 36.Suppl 11, S17–S20 (cit. on p. 78).

Kennedy, Kenneth AR (1995). "But professor, why teach race identification if races don't exist?" In: *Journal of Forensic Sciences* 40.5, pp. 797–800 (cit. on p. 78).

Kennedy, Rebecca F, C Sydnor Roy, and Max L Goldman (2013). *Race and ethnicity in the classical world: An anthology of primary sources in translation*. Hackett Publishing (cit. on p. 78).

Keyes, Os, Chandler May, and Annabelle Carrell (2021). "You keep using that word: Ways of thinking about gender in computing research". In: *Proceedings of the ACM on human-computer interaction* 5.CSCW1, pp. 1–23 (cit. on p. 43).

Khan, Salman, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah (2022). "Transformers in vision: A survey".

In: *ACM computing surveys (CSUR)* 54.10s, pp. 1–41 (cit. on pp. 36–38, 130–132, 138, 143).

Kingma, Diederik P and Max Welling (2013). "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (cit. on p. 39).

Kolesnikov, Alexander, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, et al. (2021). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: (cit. on pp. 37, 38, 130, 138, 143).

Kosut, Mary (2012). *Encyclopedia of Gender in Media*. Sage (cit. on p. 26).

Krishnakumar, Arvindkumar, Viraj Prabhu, Sruthi Sudhakar, and Judy Hoffman (2021). "Udis: Unsupervised discovery of bias in deep visual recognition models". In: *British Machine Vision Conference (BMVC)*. Vol. 1. 2, p. 3 (cit. on p. 114).

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). "Imagenet classification with deep convolutional neural networks". In: *Advances in Neural Information Processing Systems* 25 (cit. on pp. 24, 36, 37, 61, 172).

Kumar, Teerath, Alessandra Mileo, Rob Brennan, and Malika Bendechache (2023). "RSMDA: Random Slices Mixing Data Augmentation". In: *Applied Sciences* 13.3, p. 1711 (cit. on p. 169).

Kurita, Keita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov (2019). "Measuring bias in contextualized word representations". In: *arXiv preprint arXiv:1906.07337* (cit. on p. 54).

Learned-Miller, Erik, Gary B. Huang, Aruni RoyChowdhury, Haoxiang Li, and Gang Hua (2016). "Labeled Faces in the Wild: A Survey". In: *Advances in Face Detection and Facial Image Analysis*. Springer International Publishing, pp. 189–248. DOI: 10.1007/978-3-319-25958-1_8 (cit. on pp. 57, 74).

LeCun, Yann, Léon Bottou, Yoshua Bengio, and Patrick Haffner (1998). "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11, pp. 2278–2324 (cit. on p. 37).

Li, Junnan, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi (2021). "Align before fuse: Vision and language representation learning with momentum distillation". In: *Advances in Neural Information Processing Systems* 34, pp. 9694–9705 (cit. on p. 54).

Li, Lei, Fan Tang, Juan Cao, Xirong Li, and Danding Wang (2023). "Bias oriented unbiased data augmentation for cross-bias representation learning". In: *Multimedia Systems* 29.2, pp. 725–738 (cit. on pp. 168, 169).

Li, Yixuan, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft (2015). "Convergent learning: Do different neural networks learn the same representations?" In: *arXiv preprint arXiv:1511.07543* (cit. on p. 130).

Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick (2014). "Microsoft COCO: Common Objects in Context". In: *Computer Vision – ECCV 2014*. Springer International Publishing, pp. 740–755. DOI: 10.1007/978-3-319-10602-1_48 (cit. on pp. 32, 57, 61).

Linnainmaa, Seppo (1970). "The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors". PhD thesis. Master's Thesis (in Finnish), Univ. Helsinki (cit. on p. 37).

Luccioni, Sasha, Christopher Akiki, Margaret Mitchell, and Yacine Jernite (2024). "Stable bias: Evaluating societal representations in diffusion models". In: *Advances in Neural Information Processing Systems* 36 (cit. on pp. 24, 25, 27, 29, 44, 55, 56, 155, 159).

Manjunatha, Varun, Nirat Saini, and Larry S. Davis (June 2019). "Explicit Bias Discovery in Visual Question Answering Models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (cit. on p. 25).

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan (2021). "A survey on bias and fairness in machine learning". In: *ACM Computing Surveys (CSUR)* 54.6, pp. 1–35 (cit. on pp. 29, 42, 57).

Motschenbacher, Heiko and Eka Roivainen (2020). “Personality traits, adjectives and gender: integrating corpus linguistic and psychological approaches”. In: *Journal of Language and Discrimination* 4.1, pp. 16–50 (cit. on pp. 97, 101, 188).

Naseer, Muhammad Muzammal, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang (2021). “Intriguing properties of vision transformers”. In: *Advances in Neural Information Processing Systems* 34, pp. 23296–23308 (cit. on pp. 130, 143).

National Institute of Standards and Technology (2023). *NIST/SEMATECH e-Handbook of Statistical Methods*. Accessed: 2024-07-04 (cit. on p. 148).

Ntoutsi, Eirini, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, et al. (2020). “Bias in data-driven artificial intelligence systems—An introductory survey”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10.3, e1356 (cit. on p. 28).

Paek, Hye-Jin, Michelle R Nelson, and Alexandra M Vilela (2011). “Examination of gender-role portrayals in television advertising across seven countries”. In: *Sex roles* 64, pp. 192–207 (cit. on pp. 26, 45).

Park, Namuk and Songkuk Kim (2022). “How Do Vision Transformers Work?” In: *International Conference on Learning Representations* (cit. on pp. 36, 38, 131, 132, 138, 141).

Radford, Alec, 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, pp. 8748–8763 (cit. on pp. 25, 36, 37, 39, 54, 80, 91, 92, 98, 114, 120, 131, 132, 142).

Raji, Inioluwa Deborah and Joy Buolamwini (2019). “Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products”. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 429–435 (cit. on p. 28).

Ramesh, Aditya, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen (2022). "Hierarchical text-conditional image generation with clip latents". In: *arXiv preprint arXiv:2204.06125* (cit. on pp. 37, 41, 114, 115, 120, 132, 155).

Ramesh, Aditya, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever (2021). "Zero-shot text-to-image generation". In: *International Conference on Machine Learning*. PMLR, pp. 8821–8831 (cit. on pp. 36, 37, 40).

Ratcheva, V. et al. (2022). *Global Gender Gap Report 2022: Insight Report*. Geneva, Switzerland: World Economic Forum (cit. on p. 83).

Richards, Christina, Walter Pierre Bouman, Leighton Seal, Meg John Barker, Timo O Nieder, and Guy T'Sjoen (2016). "Non-binary or genderqueer genders". In: *International Review of Psychiatry* 28.1, pp. 95–102 (cit. on p. 155).

Roccas, Sonia, Lilach Sagiv, Shalom H. Schwartz, and Ariel Knafo (2002). "The Big Five Personality Factors and Personal Values". In: *Personality and Social Psychology Bulletin* 28.6, pp. 789–801. DOI: 10.1177/0146167202289008 (cit. on p. 97).

Rombach, Robin, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer (2022). "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695 (cit. on pp. 36, 37, 41, 114, 120, 132, 155).

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox (2015). "U-net: Convolutional networks for biomedical image segmentation". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, pp. 234–241 (cit. on p. 37).

Rothe, Rasmus, Radu Timofte, and Luc Van Gool (Dec. 2015). "DEX: Deep EXpectation of apparent age from a single image". In: *IEEE International Conference on Computer Vision Workshops (ICCVW)* (cit. on p. 74).

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei (Apr. 2015). "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision* 115.3, pp. 211–252. DOI: 10.1007/s11263-015-0816-y (cit. on pp. 30, 57, 62, 63).

Savani, Yash, Colin White, and Naveen Sundar Govindarajulu (2020). "Intra-processing methods for debiasing neural networks". In: *Advances in Neural Information Processing Systems* 33, pp. 2798–2810 (cit. on pp. 25, 29, 53).

Schnabel, Tobias, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims (2016). "Recommendations as treatments: Debiasing learning and evaluation". In: *International Conference on Machine Learning*. PMLR, pp. 1670–1679 (cit. on p. 28).

Schwemmer, Carsten, Carly Knight, Emily D Bello-Pardo, Stan Oklobdzija, Martijn Schoonvelde, and Jeffrey W Lockhart (2020). "Diagnosing gender bias in image recognition systems". In: *Socius* 6, p. 2378023120967171 (cit. on p. 44).

Selvaraju, Ramprasaath R, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra (2016). "Grad-CAM: Why did you say that?" In: *arXiv preprint arXiv:1611.07450* (cit. on p. 81).

Serna, Ignacio, Alejandro Pena, Aythami Morales, and Julian Fierrez (2021). "InsideBias: Measuring bias in deep networks and application to face gender biometrics". In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, pp. 3720–3727 (cit. on pp. 25, 29, 52, 53, 114, 130, 132).

Shankar, Shreya, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley (2017). "No classification without representation: Assessing geodiversity issues in open data sets for the developing world". In: *arXiv preprint arXiv:1711.08536* (cit. on pp. 32, 57, 58).

Shen, Sheng, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer (2021). "How Much Can CLIP Bene-

fit Vision-and-Language Tasks?" In: *arXiv preprint arXiv:2107.06383* (cit. on p. 100).

Shorten, Connor and Taghi M Khoshgoftaar (2019). "A survey on image data augmentation for deep learning". In: *Journal of big data* 6.1, pp. 1–48 (cit. on p. 169).

Signorielli, Nancy (1990). "Children, television, and gender roles: Messages and impact". In: *Journal of Adolescent Health Care* 11.1, pp. 50–58 (cit. on p. 26).

Sim, Jack (n.d.). *Computing Receptive Fields of Convolutional Neural Networks — distill.pub*. `https://distill.pub/2019/computing-receptive-fields/`. [Accessed 08-May-2023] (cit. on p. 138).

Singh, Vivek K, Mary Chayko, Raj Inamdar, and Diana Floegel (2020). "Female librarians and male computer programmers? Gender bias in occupational images on digital media platforms". In: *Journal of the Association for Information Science and Technology* 71.11, pp. 1281–1294 (cit. on pp. 46, 130, 136).

Sirotkin, Kirill, Pablo Carballeira, and Marcos Escudero-Viñolo (2022). "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*, pp. 10442–10451 (cit. on pp. 51, 114, 132).

Smith, Philip and Karl Ricanek (2020). "Mitigating algorithmic bias: Evolving an augmentation policy that is non-biasing". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, pp. 90–97 (cit. on pp. 168, 169).

Sohl-Dickstein, Jascha, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli (2015). "Deep unsupervised learning using nonequilibrium thermodynamics". In: *International Conference on Machine Learning*. PMLR, pp. 2256–2265 (cit. on p. 40).

Srinivasan, Tejas and Yonatan Bisk (2021). "Worst of both worlds: Biases compound in pre-trained vision-and-language models". In: *arXiv preprint arXiv:2104.08666* (cit. on p. 54).

Steed, Ryan and Aylin Caliskan (2021). "Image representations learned with unsupervised pre-training contain human-like biases". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 701–713 (cit. on pp. 29, 50, 51, 114).

Su, Weijie, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai (2019). "Vl-bert: Pre-training of generic visual-linguistic representations". In: *arXiv preprint arXiv:1908.08530* (cit. on p. 54).

Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich (2015). "Going deeper with convolutions". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9 (cit. on pp. 36, 37).

Tan, Mingxing and Quoc Le (2019). "Efficientnet: Rethinking model scaling for convolutional neural networks". In: *International Conference on Machine Learning*. PMLR, pp. 6105–6114 (cit. on p. 37).

Team, Gemini, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. (2023). "Gemini: a family of highly capable multimodal models". In: *arXiv preprint arXiv:2312.11805* (cit. on p. 25).

Van Den Oord, Aaron, Oriol Vinyals, et al. (2017). "Neural discrete representation learning". In: *Advances in Neural Information Processing Systems* 30 (cit. on pp. 39, 40).

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is all you need". In: *Advances in Neural Information Processing Systems* 30 (cit. on p. 38).

Vice, Jordan, Naveed Akhtar, Richard Hartley, and Ajmal Mian (2023). "Quantifying Bias in Text-to-Image Generative Models". In: *arXiv preprint arXiv:2312.13053* (cit. on pp. 26, 29, 56).

Vicente, Sara, Joao Carreira, Lourdes Agapito, and Jorge Batista (2014). "Reconstructing pascal voc". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 41–48 (cit. on p. 61).

Vincent, James (June 2020). *What a machine learning tool that turns Obama white can (and can't) tell us about AI bias* (cit. on p. 24).

Vincent, Richard C, Dennis K Davis, and LA Bronszkowski (1987). "Sexism on MTV: A content analysis of rock videos". In: *Journalism Quarterly* 64, pp. 750–755 (cit. on p. 27).

Wang, Angelina, Alexander Liu, Ryan Zhang, Anat Kleiman, Leslie Kim, Dora Zhao, Iroha Shirai, Arvind Narayanan, and Olga Russakovsky (2022). "REVISE: A tool for measuring and mitigating bias in visual datasets". In: *International Journal of Computer Vision*, pp. 1–21 (cit. on pp. 27–29, 32, 43, 44, 46–49, 63, 90, 115, 122, 123, 130, 136, 155, 171).

Wang, Angelina, Arvind Narayanan, and Olga Russakovsky (2020). "REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets". In: *Computer Vision – ECCV 2020*, pp. 733–751 (cit. on pp. 30, 58, 61, 63, 68).

Wang, Jialu, Yang Liu, and Xin Wang (2021). "Are Gender-Neutral Queries Really Gender-Neutral? Mitigating Gender Bias in Image Search". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1995–2008 (cit. on p. 25).

Wang, Tianlu, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez (2019). "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*, pp. 5310–5319 (cit. on pp. 25, 29, 44, 46, 50–52, 115, 122, 130, 132, 141, 155, 174).

Wang, Zeyu, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky (2020). "Towards fairness in visual recognition: Effective strategies for bias mitigation". In: *Proceedings of the IEEE/CVF*

conference on computer vision and pattern recognition, pp. 8919–8928 (cit. on p. 130).

Wiggers, Kyle (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] (cit. on p. 91).

Wolfe, Robert, Mahzarin R Banaji, and Aylin Caliskan (2022). "Evidence for Hypodescent in Visual Semantic AI". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1293–1304 (cit. on pp. 95, 132).

Wolfe, Robert, Yiwei Yang, Bill Howe, and Aylin Caliskan (2022). *Contrastive Language-Vision AI Models Pretrained on Web-Scraped Multimodal Data Exhibit Sexual Objectification Bias* (cit. on pp. 95, 132).

Yang, Kaiyu, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky (2020). "Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 547–558 (cit. on p. 47).

Zhang, Yi and Jitao Sang (2020). "Towards accuracy-fairness paradox: Adversarial example-based data augmentation for visual debiasing". In: *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 4346–4354 (cit. on pp. 43, 168, 169).

Zhang, Zhifei, Yang Song, and Hairong Qi (2017). "Age Progression/Regression by Conditional Adversarial Autoencoder". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4352–4360. DOI: 10.1109/CVPR.2017.463 (cit. on p. 74).

Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang (Sept. 2017a). "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*. Copenhagen, Denmark: Asso-

ciation for Computational Linguistics, pp. 2979–2989. DOI: 10.18653/v1/D17-1323 (cit. on pp. 28, 29, 44, 130).

Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang (2017b). "Men also like shopping: Reducing gender bias amplification using corpus-level constraints". In: *arXiv preprint arXiv:1707.09457* (cit. on pp. 43, 52, 155).

— (2018). "Gender bias in coreference resolution: Evaluation and debiasing methods". In: *arXiv preprint arXiv:1804.06876* (cit. on pp. 29, 54).

Zietlow, Dominik, Michael Lohaus, Guha Balakrishnan, Matthäus Kleindessner, Francesco Locatello, Bernhard Schölkopf, and Chris Russell (2022). "Leveling down in computer vision: Pareto inefficiencies in fair deep classifiers". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10410–10421 (cit. on p. 130).