Bias can present itself in a myriad of ways as outlined in Section 2.6 such as Selection, Framing and Label bias, in line with the concerns of this research paper selection and framing bias are the main types considered. The study of these biases can prove useful in exposing the presence of baser biases in particular gender, race, age and emotion bias. Considering the research carried out by Fabbrizzi et al. in \cite{RefWorks:RefID:30-fabbrizzi2022survey} a total of twenty four papers and the strategies they used to discover bias was reviewed, these strategies were than grouped into four categories:

* Reduction to tabular data – These measure bias using already present/extracted attributes and labels as if it were a tabular dataset.
* Biased image representation – These discover bias through lower-dimensional representations of the data.
* Cross-dataset bias detection – These assess bias by comparing different datasets.
* Other – These are methods which do not fall under the above categories.

Reduction to tabular data

Most strategies presented in this section utilise automatic feature extraction processes which themselves are prone to errors and bias, this can in turn result in the reflection and amplification of pre-existing image bias. It is also noted that the impact of this additional source of bias it typically omitted, having its presence simply be noted when looking at the results. This section covers some of the strategies used omitting those which are similar.

Count / demographic parity.

Dulhanty and Wong in \cite{dulhanty-wong-bias-detection} determined the presence of gender and age bias in the ImageNet dataset by extracting the age and gender of images using relevant recognition models, which they used said results to work out the dataset distribution across age and gender. This method provides insight into selection bias but also on the framing of the protected attributes, given a suitable labelling of the dataset. Finally, it was noted by the authors that such a method relies on the assumption that the recognition models involved are not biased, which is a far claim from the truth and thus, the analysis is not fully reliable.

Yang et al. \cite{yang-bias-detection} similarly opted to address selection and label bias in relation to the person category of the ImageNet dataset. To address label bias, they firstly had annotators remove images which could be offensive or sensitive (e.g., sexual/racial slur) and those with ambiguous labels, this was then followed by the annotators labelling the remaining images according to some categories of interest (gender, age, and skin colour) to understand whether the remaining data was balanced with respect to those categories and then to address selection bias. The annotation process was validated by measuring the agreeableness of the annotators on a small, controlled set of images.

Zhao et al. \cite{zhao-bias-detection} measured the correlation between protected attributes and the occurrences of certain objects/actions. This was carried out via calculation X wherein *g­n* was a protected attribute and *o* an occurrence of an object or action in the image. The bias score is denoted by *b(o,g)* whereas c(o,x) counts the co-occurrences of the object/action o and the protected attribute’s value x. Assuming that , where is the number of possible values that the protected attribute can be, means that the attribute g is positively related to object/action o.

Buolamwini and Gebru \cite{buolamwini-gebru-bias-detection} similar to the previously mentioned strategies used a simple count however they took into consideration multiple protected attributes at one i.e., taking into consideration *black woman* as opposed to *black* and *woman* on their own.

Information theoretical

Merler et al. \cite{merler-bias-detection } presented four measurements for a balanced dataset. The Shannon entropy and Simpson Index for measuring diversity whilst the Evenness Shannon and Evenness Simpson for measuring evenness.

Shannon Entropy -

Simpson Index -

is the probability of an image having value for the attribute .

Evenness (Shannon) -

Evenness (Simpson) -

Panda et al. \cite{panda-bias-detection} proposed the use of conditional Shannon entropy for determining framing bias. This was done by working out the conditional entropy of each object across the positive and negative set of the emotions to see whether some objects/scenes were more likely to be related to a certain emotion. When the conditional entropy of an object is zero it implies that said object is associated only with a particular emotion which can be interpreted as a form of framing bias.

Conditional Shannon Entropy - , emotion where e = 1 denotes an emotion (happy) and e = 0 denotes the negative (not happy), c denotes an object.

Kim et al. produced a precise definition of bias; thus, a dataset contains bias when where X is a protected attribute, Y is the feature/model output variable.

Other

REVISE

Biased image representation.

Distance Based

Steed and Caliskan \cite{steed-caliskan-bias-detection} introduced a method by which human-like biases in the latent image representation of unsupervised generative models can be addressed. They measured bias by studying the associations between semantic concepts by measuring the cosine similarity among vectors in the model latent space between samples of images that coincide with the visual concepts. Particularly assuming is a model which maps images into a vector space and are image sets whilst are image sets which represent the concepts. The association between with and with is measured as follows:

where

Alternatively, Cohen’s d can be used to measure the size of the association:

The issue with this implementation is that its results depend on the learned representation and the models used.

Other

Balakarnishnan et al. \cite{balakrishnan-bias-detection} assessed algorithmic bias in image classifiers by computing the causal relationship between the protected attribute and the output of said classifier. They generated modified versions of the original images with respect to the attributes considered, using image generators such as a GAN. In turn said images were used to determine the presence of algorithmic bias present in the image classifier.

Other Methods

Human-based

Hu et al. \cite{hu-bias-detection} opted for a non-automated bias detection approach, they devised a three step process revolving around selection and framing bias wherein initially workers are presented with a batch of images for which they need to describe the similarities amongst the image batches via question-answer pairs (e.g., assuming the image batches show only white airplanes, the worker would label the batch with the following question-answer pair: “What colour are the airplanes in the images? White”). The second step involves the workers answering the questions retrieved in the first step on arbitrary image batches to conclude on the presence of said bias. Finally, the workers are asked to determine if the statements made (e.g., all planes are white) are in fact true in the real world. In cases where the answer is yes than those instances are ignored as no bias is present. The remaining statement then would expose the bias present; however, this process depends on the workers general knowledge of the world and as such is not free from bias.

Continue To read the paper and add to it + at the end check REVISE to see if it implemented some stuff which I didn’t mention

REVISE

The revise tool adopts a multi-variant approach to detecting bias considering object, person, and geography-based insights. This section will go over person-based insights as the object and geography-based insights are irrelevant to this research paper. The relevant metrics used for detecting person-based bias are outlined below:

Person Prominence

This considers the proportion of the image that the subject takes up in addition to the distance of the subject from the centre of the image. These measures are then treated as a proxy for the subjects importance. This analysis was carried out on the COCO dataset for images separated by gender and skin tone, for which the Cohen’s D measurements was used to facilitate a comparison between the different groups whilst Jonckheere’s trend test was used to visualise an a priori ordering of the data.

Contextual Representation

This considers the context in which individuals are primarily featured in through the objects and scenes with which they are primarily associated with. Taking into consideration the COCO dataset it was concluded that woman tend to be greatly associated with *shopping* and *dining* whilst being depicted in images containing *furniture*, *accessory*, and *appliances*. Contrarily man tend to be associated with *sports fields* and *water*, *ice*, *snow*, whilst depicted in images mostly containing *sport items* and *vehicles*. This reflects traditional gender stereotypes present in society.

Instance counts and distances

This opts to look deeper than simply the number of times certain object appear with individuals rather it considers the distance said object is from the subject to determine if the subject is interacting with the object or whether it is simply in the background. This is achieved via a scaled distance metric depicted in equation X wherein *p* denotes the person, *o* the object and the are calculated on a normalised image of total area one. This metric in turn outline whether certain demographics are depicted interacting with certain objects as opposed to the image simply containing the two.

Appearance differences

This opts to analyse appearance difference in images of people of varying demographics in relation to particular objects. This was carried out to further disambiguate situation where occurrence counts, and distance aren’t depicting the entire situation. This involves extracting FC7 features from a subset of image to get scene-level feature, projecting them into dimensions to prevent over-fitting and then fitting a Linear SVM to see if it learns the difference between image containing the same object but people of different demographics. This results in insights such as man being portrayed playing *outdoor sports* whilst woman *indoor sports* when considering the object *sports uniform*.