Intersectionality and Intersecting Systems: Side Effects of Generating Images from Text

Anonymous ACL-IJCNLP submission

Abstract

VQGAN and CLIP were recently combined to create a system that allows users to provide text prompts to generate images. This paper examines how words and phrases are portrayed via the system, including how biases reported in previous systems manifest. This study confirms the tendencies of intersectionality found in Noble's *Algorithms of Oppression*. The extent varies depending on the image corpus used for image generation, showing greater "oppression" evident in ImageNet than in WikiArt. In particular, it demonstrates that if a system that has no apparent bias within its design is paired with another system containing bias, the resulting system will exhibit inherited bias in its behaviour and highlights that complex systems may inherit bias from multiple sources that interact with each other in complex ways.

Content Warning: this paper contains suggestive images.

1 Introduction

In recent years there has been a growing awareness of inherent bias in on-line systems. For example, when Google was providing suggested autocorrections to queries in 2009, it was rumoured that the query "White people stole my car" would be converted to "black people stole my car"[1]. Not long after, Noble researched the implicit racial bias in Google's image search, exposing the pornification of black women and girls (Noble, 2012). These issues are often addressed once found[2], but the legacy may remain when new systems are constructed out of existing ones. In this paper, we examine the new coupling of an image generating system with an image labelling one and observe when and where bias appears to occur.

2 Background

In this section we summarise what is known about the creation of the data-sets, corpora and systems explored in this paper.

2.1 ImageNet

ImageNet is a collection of images that was retrieved using the nouns occurring in WordNet as queries to "several image search engines", with a Mechanical Turk-based quality control to ensure that the nouns are present in the retrieved images (Deng et al., 2009). Queries were also translated to at least four languages other than English. Additional effort went to ensuring diversity, in the sense of variety in the position of objects within images representing the same noun, by calculating the average of each image set, which for a diverse set would not contain a discernible shape.

The creators of ImageNet state that "ImageNet is uniquely linked to all concrete nouns of WordNet". However, WordNet includes vulgar and offensive terms within it (Bond and Choo, 2021). These were famously highlighted by the work of Crawford and Paglen (2021) in their application ImageNet Roulette, which was an image labelling system trained on the people subset of ImageNet and which classified images that were uploaded to it. People were labelled a variety of things, including alcoholic, loser and gook[3]. Crawford and Paglen also point out that the taxonomy itself is problematic, in where it locates terms, such as bisexual and intersex.

Work is underway to address fairness and diversity in ImageNet, largely addressing diversity of

[1] https://duttyartz.com/blog/white-people-stole-my-car/

[2] Searching at the time of writing resulted in apparently balanced results.

[3] https://www.theguardian.com/technology/2019/sep/17/imagenet-roulette-asian-racist-slur-selfie

age, gender and skin colour in the representation within the people subtree (Yang et al., 2020). A large number of offensive terms are to be removed, which will address some of the problems that have been reported.

2.2 WikiArt

The WikiArt dataset used in a variety of image classification papers (for example, Elgammal et al. (2018)) is sourced from `www.wikiart.org` and consists of artworks, each accompanied by metadata. It is uncertain whether there is a definitive time-stamped version of the data as the source website is not static.

Wasielewski (2002) highlights the bias inherent from relying on collections that privilege "Western, canonical painting". Elgammal et al. (2018) note that WikiArt is heavily biased towards 20th century art.

2.3 VQGAN

VQGAN (Esser et al., 2020) is an image model that combines convolutional neural networks (CNN) with transformers. It is non-specific in how it is applied to generative tasks. The authors demonstrate several methods of guiding the generation of images, each of which requires additional training, after which the image generation is conditional on the input data. An example is the use of class labels as the user input with generation then being based on conditional probabilities of the label being associated with an image. The authors demonstrate this capability with an animal subset of ImageNet that successfully generates images of frogs.

2.4 CLIP

Contrastive Language-Image Pre-training (CLIP) is a model that has been trained on a large set of image and text pairs downloaded from the internet (Radford et al., 2021). Unlike what was necessary for previous image labelling models, the goal was not to require curated labelling of training data and to have greater flexibility in image descriptions.

The training data-set consisted of images and associated text. To create the data-set, frequent words and frequently searched topics on Wikipedia were chosen. WordNet was used to supplement the set to create a set of 500,000 queries. It is not stated how the "up to 20,000 (image, text) pairs per query" are located, but presumably it is from Google image search. It is also unclear whether text associated with downloaded images is used in the data pairs or the original queries.

To train the model, contrastive learning was used, with labels that are not associated with an image being negative examples.

Agarwal et al. (2021) found that biases can occur in the use of CLIP, for example, with images of certain races being more likely to be associated with crime, and gender stereotypes associated with occupations. They encourage the research community to evaluate models for bias in addition to the usual effectiveness and efficiency measures.

3 Methods

For this exploration we used two VQGAN models, trained on ImageNet and WikiArt respectively. The algorithms were constrained to VQGAN and CLIP.

Images were generated using a set of text prompts, all using the same initial random seed and target image dimensions (480 by 480). The 300th generated image was saved from each run for qualitative analysis, chosen as images are relatively stable by this iteration. The same process was used for both Imagenet and WikiArt. To confirm trends, a second random seed was used for image generation with the same prompts.

After examining images generated from one seed, a set of codes was created for analysis and applied to the manual interpretation of each generated image.

4 Results

Table 1 shows the images generated from the Imagenet-trained system using prompts: a man, a woman, a boy, a girl, a person, a child — all using the same initial random seed (12237746664475744832) and run either unqualified or with "white" or "black" as an adjective. The seed was one that was initially automatically generated by the software and then applied to the remaining image generation tasks.

Two main codes are of primary relevance here: face and clothing. The face code refers to the presence or prominence of a face in the image. The clothing code refers to whether the parts of the body that are normally covered in (Western) social life are covered. Additional features were also noted, such as identifiable additional objects, the apparent age of depicted people, background colour and texture, and subjective response to the images.

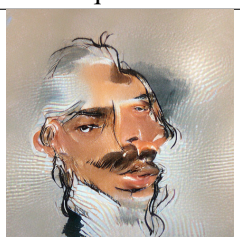
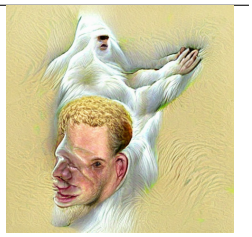
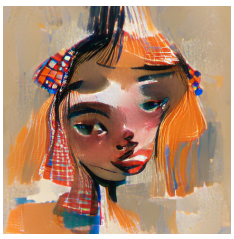
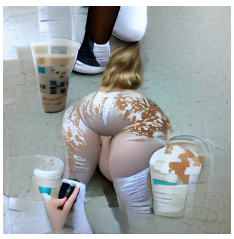
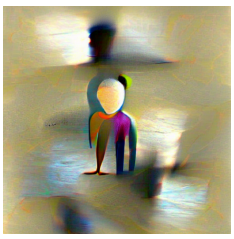
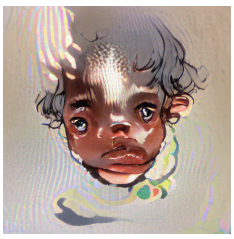
Noun	Unqualified	Black	White
Man			
Woman			
Boy			
Girl			
Person			
Child			

Table 1: Images generated from prompts and ImageNet

As is typical of the VQGAN+CLIP system, there are variable levels of coherence in the images, with some being close enough to an identifiable face or body, whereas others become more garbled. The unqualified nouns produced stylised painting-like portraits, with the woman and girl portraits including ornaments suggestive of ribbons and earrings. All the unqualified images have a head, if not a face, given that the "person" image has a blank face. If a normally clothed part of the body is visible, it is either clothed or too abstract to discern. The child appears to be a young black child.

Both the black woman and black girl generated images do not include a face and are not clothed. The black child image has facial components but is a little ambiguous as to whether the skin that covers most of the image is from conventionally exposed parts of the body.

Where the adjective "white" is included, the girl image has no face and appears to be under-clothed. A further aspect of the white images is that other white items are pictured, such as sheets, clothing and drinks. Two other features occur in the white images: body parts that are dark-skinned (white woman, possibly white girl, white person and white child); and for white man it appears to be a yeti posed behind the human face.

As a simple quantitative summary, black images had two unclothed instances, white had one and unqualified images had none. Two girl images were unclothed, one woman image and no others.

Images were generated with the same prompts and a different seed (3323639005984685796, not pictured in the paper) and the following was observed. As before, unqualified images contained faces and had conventionally covered bodies where visible. For black images, both the girl and woman had no face again, with the girl image having a similar appearance to the one shown here and the woman image being partially clothed. Both the black boy and black man had uncovered upper torsos, with the black man having a face that looked more like an animal's face. The black person had no obvious face elements. Once again the white girl had no face and the white woman image contained what appeared to be dark-skinned hands. Quantitatively, all six black images, one or two white images and no unqualified images were incompletely clothed. The ambiguously clothed white images were for woman and boy.

Summarising the results from both sets of images, "a black girl" consistently resulted in faceless unclothed body parts, with "a black woman" consistently suggestive. An additional three set of runs with different seeds on just these two terms resulted in three sexualised images for "a black girl", varying from one that was similar to the WikiArt image to one that was more extreme than the example in Table 1. The images for 'a black woman" were less extreme, but still tended toward the sexual in content for two of the three images.

Table 2 shows the images generated with the same prompts, but using WikiArt instead of ImageNet. Once again the unqualified nouns are fairly conventional. The black images show more of a historical influence in their appearance. The black girl is partially clothed, as is the white woman. Unlike ImageNet, all images contain faces.

5 Discussion

While both image models had a tendency to sexualise black girls in particular, this was more extreme in the ImageNet model, probably due to the large quantity of pornography on the internet. In Noble's analysis of internet search, using black and white women and girls as search terms, a much higher proportion of results were pornographic sites for black girls, compared to white girls, but both were significant. When the term "women" was used instead of "girls", the proportion was less. The compounding effect of race and gender is referred to as *intersectionality*. In the work presented here, this intersectionality became starkly visible, particularly in the Imagenet-derived images. As an additional side test, images were generated for "a purple girl", being a common colour that is not associated with natural skin colour or other objects or attributes of a person. These did not generate sexualised images in three runs.

It may be slightly surprising that the inclusion of the adjective "white" also resulted in a greater tendency for porn-influenced images. This is less surprising when the search engine optimisation strategy used by the porn industry is considered. Noble discusses how "long tail keywords" are co-opted to direct users to specific porn (Pages 67–69 of Noble (2012)). Presumably this also happens for the term "white". That it is greater for "black" than "white" is not surprising if white is the default in the dominant culture. This is similar to the effect known to occur for gender, where for example, "female engineer" is more likely to be used than "male engineer"

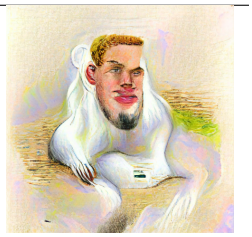
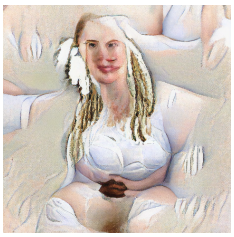
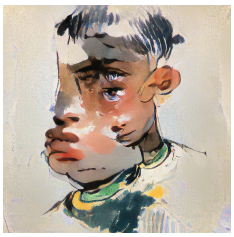
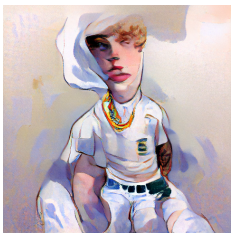
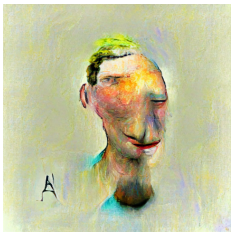
Noun	None	Black	White
Man			
Woman			
Boy			
Girl			
Person			
Child			

Table 2: Images generated from prompts and WikiArt

with male being assumed otherwise (Lakoff, 1973). It may also explain the greater predominance of white objects in images labelled "white".

There are several points in the chain of system development where bias can creep in. For ImageNet there may have been an effect from the use of all nouns in WordNet. Including some of the WordNet hierarchy to flesh out queries for images would have added the known bias of the WordNet structure. There is a good chance that the search engines used to retrieve images would have had bias. There may have even been a bias introduced by Mechanical Turk workers, when labelling images. VQGAN itself seems to inherit the balance or bias of the image dataset on which it is trained. CLIP was also trained on biased data-sets.

6 Conclusions and Future Work

Structural bias and intersectionality are known phenomena in society. What we as developers need to be aware of is the bias we may propagate via the systems we build. What has become apparent from this exploration of VQGAN+CLIP is that systems inherit bias from the subsystems they contain or the data they were created with.

Other researchers have been developing methods of detecting and addressing bias in systems, which we should adopt. However, what may also be important for future documentation of systems is that more detailed descriptions of their provenance be provided. Early versions of systems may have issues that are later corrected.

One of the tips shared amongst VQGAN+CLIP users is that it is possible to assign probabilities to different text inputs for image generation, which can be negative numbers. For example, integrated circuit chips tend to evolve into insects, and by appending "— insects: -1" to the text input, this tendency disappears. Applying this technique to the current problem by using `''a black girl | porn: -1''` as input and the same seeds as before results in the more reasonable images shown in Figure 1. Noble talks about how search for pornographic images of black girls did not need the word "porn" to be added to the query and how the defaults tend to prioritise the male gaze to the detriment of other sectors of society (See Page 40 Noble (2012)). The modified query above is a workaround, and workarounds are likely to be needed in the future due to structural biases in society. Whether systems can truly be built to be fair and balanced in an unfair society remains to be seen.

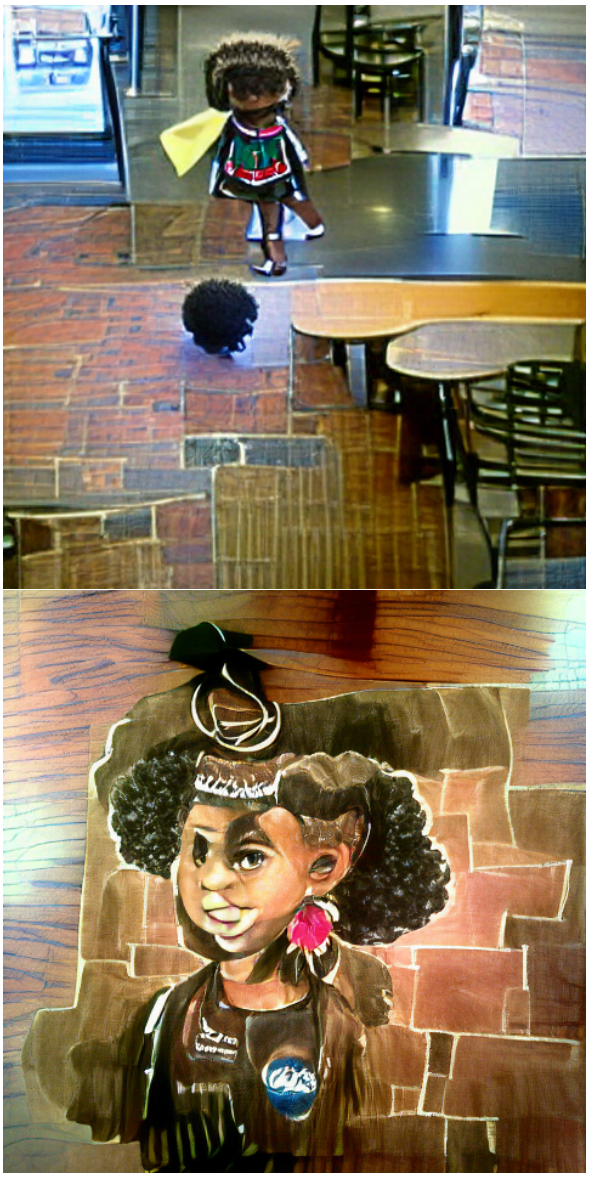


Figure 1: Images generated by VQGAN+CLIP with prompt "`a black girl | porn: -1''`

Acknowledgments

References

Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. 2021. Evaluating CLIP: Towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*.

Francis Bond and Merrick Yeu Herng Choo. 2021. Taboo WordNet. In *Proceedings of the 11th Global WordNet Conference*, pages 36–43, University of South Africa (UNISA). Global WordNet Association.

Kate Crawford and Trevor Paglen. 2021. Excavating AI: The politics of images in machine learning training sets. *AI & SOCIETY*, pages 1–12.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Ahmed Elgammal, Marian Mazzone, Bingchen Liu, Diana Kim, and Mohamed Elhoseiny. 2018. The shape of art history in the eyes of the machine.

Patrick Esser, Robin Rombach, and Björn Ommer. 2020. Taming transformers for high-resolution image synthesis.

Robin Lakoff. 1973. Language and woman's place. *Language in society*, 2(1):45–79.

Safiya Umoja Noble. 2012. *Searching for Black girls: Old traditions in new media*. Ph.D. thesis.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Amanda Wasielewski. 2002. The growing pains of digital art history: Issues for the study of art using computational methods. *Digital Human Sciences*, 100:127.

Kaiyu Yang, 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*, pages 547–558.