Paper Review: OpenBias: Open-set Bias Detection in Text-to-Image Generative Models

By: Lam Nguyen

1. SUMMARY

Text-to-Image generative models are becoming much more popular and available to the public. However, they have problems with being unsafe and unfair. It is desirable to remove as much bias as possible concerning concepts like race, gender, politics, social class, sexual and violent material, etc…

A lot of the work is focused on detecting and eliminating a set of predetermined biases and well-known biases. What about in the future when new forms of human bigotry or discrimination occur? For example, what happens in the future if humans evolve and branch of into 2 different species? And there is bias against one of those branches of evolution? This form of bias prevention is not built into the current bias removal techniques.

This is where OpenBias comes into play. OpenBias identifies and quantifies the severity of biases agnostically without access to a precompiled set.

It does this in three stages: The first stage is to leverage a LLM to propose biases given a set of captions. The second step is to have the target text-to-image model produce images using the same set of captions. Lastly, a vision question answering model recognizes the extent of the previously proposed biases.

This OpenBias system was tested on Stable Diffusion 1.5,2 and XL. It was determined in the paper that the OpenBias system agreed with the current set of closed-set bias detection techniques as well as qualitative human judgment.

2. STRENGTHS

- The model is adaptive and can be changed as human society changes. When new bigoted behavior is uncovered or goes away, the OpenBias system can also adapt
- Leverages current powerful LLMs that are a reflection of current human society. This means that as the LLM changes, the OpenBias bias detection system can change with it instead of just being locked into place at a particular time in society.

3. WEAKNESSES

- Possibly in the zeal to make these models as unbiased as possible, there will be a reduction in the accuracy of the model since it might be the case that a bias is actually a reflection of the reality at a certain period in time.
- The new biases still have to be implemented by humans… But it makes the Bias Detection methods extensible. Maybe look into how to create an automated system of bias detection.
- Possible that the LLMs or humans determining the bias of a text-to-image model are themselves biased… So they are using the OpenBias system incorrectly.

4. TECHNICAL EXTENSIONS

- Study on more text-to-image models besides stable diffusion
- Explore methods of automated bias detection.
- Explore the opposition or ways in which the OpenBias system can make a model more Biased to certain concepts in order to deeper understand how this system can work.
- Explore new types of bias that don't yet exist except maybe in fiction in order to test the true openness of the OpenBias system instead of just using the current closed-set of commonly accepted Biases.

5. OVERALL REVIEW

The OpenBias system is an adaptable system that can change what Biases it can detect without being locked down to a particular time or person doing the judging. However, one of the downsides is that it uses an LLM which in itself might be biased. Future extensions would be to explore on other models besides Stable Diffusion and to test the system on some completely new hypothetical Biases that don't exist yet instead of just the current closed set.