Diverse Yet Biased: Towards Mitigating Biases in Generative AI (Student Abstract)

Akshit Singh

Indian Institute of Technology, Jodhpur
NH 62, Surpura Bypass Rd, Karwar, Jheepasani, Rajasthan 342030
singh.190@iitj.ac.in

Abstract

Generative Artificial Intelligence (AI) has garnered significant attention for its remarkable ability to generate text, images, and other forms of content. However, an inherent and increasingly concerning issue within generative AI systems is bias. These AI models often exhibit an Anglo-centric bias and tend to overlook the importance of diversity. This can be attributed to their training on extensive datasets sourced from the internet, which inevitably inherit the biases present in those data sources. Employing these datasets leads to AI-generated content that mirrors and perpetuates existing biases, encompassing various aspects such as gender, ethnic and cultural stereotypes. Addressing bias in generative AI is a complex challenge that necessitates substantial efforts. In order to tackle this issue, we propose a methodology for constructing moderately sized datasets with a social inclination. These datasets can be employed to rectify existing imbalances in datasets or to train models to generate socially inclusive material. Additionally, we present preliminary findings derived from training our model on these socially inclined datasets.

Introduction

Over the last few years, generative AI has made significant advances in various fields and influenced others. However, generative AI necessitates a substantial amount of data for the purpose of training. It has been observed that the datasets utilised for training often exhibited a significant level of imbalance, which may include the presence of harmful or explicit content (Ntoutsi et al. 2020; Birhane, Prabhu, and Kahembwe 2021). These biased dataset results in the production of generative AI content that is inherently biased, hence giving rise to diverse implications. Firstly, they exacerbate societal inequalities by amplifying and legitimizing existing biases. Secondly, the dissemination of biased AI-generated content has tangible consequences, as it contributes to the erosion of trust in information sources and organisations. Finally, the absence of social inclusivity showcases significant psychological and social implications for the individuals who use them (Mantelero 2018).

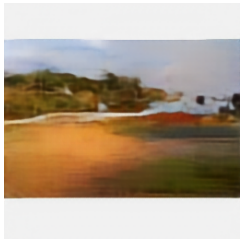
To mitigate the repercussions of biased generative AI, it is imperative to prioritize transparency, accountability, and ethical considerations in the development and deployment of these systems. One such approach that this paper explores is the development of theme-specific datasets that accurately capture the diverse characteristics associated with a given theme. These datasets have the potential to serve several goals, such as facilitating the creation of culture-specific generative material and influencing the outcomes produced by existing generative models when utilised by them. Moreover, the methodology present in this paper could be extended to generate datasets pertaining to any particular subject matter. To showcase our claims we also trained a text-to-image generative model on a country-specific (India) dataset using a similar methodology as employed by Ramesh et al.. The next section of this paper deals with the dataset and methodology used, followed by results and possible future work.

Methodology

In order to create an images-caption dataset that encompasses a certain theme, while ensuring the absence of harmful information and the inclusion of detailed captions, we leveraged the resources available on Wikipedia (Srinivasan et al. 2021). Wikipedia is often regarded as one of the most reliable source due to the extensive editing, verification, and correction processes that are built into the site. Due to this inherent characteristic of Wikipedia, it guaranteed that the captions we would receive would be devoid of any potentially detrimental material. One of the primary challenges encountered while utilising Wikipedia-based image text dataset (WIT) (Srinivasan et al. 2021) directly was that it frequently contained images without captions, came with an excessive number of captions or had captions in a language other than English. In order to address this concern, we conducted a thorough search for relevant keywords pertaining to our topic within the captions and web pages associated with the image sourced from WIT. For instance, we opted to create a dataset focused exclusively on India, and subsequently selected image-caption pairs in which either the caption or the image's webpage contained the term 'India' provided they were written in English. After obtaining a theme-specific dataset we trained a transformer-based text-to-image generative AI model (Ramesh et al. 2021) to investigate the compatibility of present technology in generating culturally specific material.

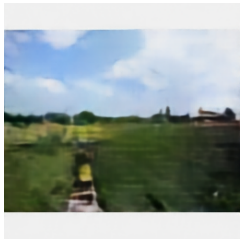
Results

Using the aforementioned methodology we created two distinct datasets. The first dataset, created was IID (Indian-context Image caption Dataset), notable for being the first of its kind to include over 250k+ images accompanied by captions that are specifically based on the Indian context. This dataset encompasses a wide range of subjects, including various landmarks, individuals, artistic works, plant and animal species, and other relevant elements. Furthermore, IHD (Indian Heritage Dataset) was developed with a specific emphasis on Indian architecture.



(a) With Indian Context (Beach)

(b) Without Indian Context (Beach)



(c) With Indian Context (Village)

(d) Without Indian Context (Village)

Figure 1: Some samples extracted from our trained model on IID and OpenAI's DALL-E with prompts given as Beach and Village respectively. We observe that the proposed approach is able to bring the Indian context into the generated images.

Creating country-specific and theme-specific datasets whose quality and quantity were reasonably good marked one of the notable advantages of our work. These datasets exhibited the necessary diversity and effectively captured the relevant context. Now the IID was utilised to train the text-to-image model in order to assess the presence of Indian context in the content produced by the generative AI model. Figure 1 illustrates a comparison between the images obtained from OpenAI's DALL-E and our model when only text was provided as input. One noteworthy finding of our study was the accurate comprehension of the Indian context by our trained model. For example, when the same input of "beach" was given in Open AI's DALL-E, it conjured up pictures of beaches with white sand, which are most commonly associated with Western nations, whereas in India, brown sand beaches are mostly present due to high mineral content in the sands (Shalini et al. 2020) as captured by our model. Another example that we observed was when input was given as "village", Open AI's DALL-E presented euro-centric villages that had hut-shaped houses and mountains in the background whereas our model focuses on India and depicts luscious green fields and plain geography.

Conclusion and Future Work

In this study, we have presented preliminary findings on the potential for enhancing diversity inclusivity in generative AI models through the utilisation of theme-specific datasets during training. The utilisation of generative AI models necessitates significant computational resources and time, resulting in the presence of blurriness in our images. However, despite this limitation, our model delivered the required outcomes. In future research, we would investigate methods for enhancing image quality and explore the potential of using smaller datasets to counterbalance the bias present in huge datasets, thereby further mitigating bias.

We hold a deep trust that our research efforts will shed light on issues surrounding diversity and inclusion in AI systems. Additionally, we seek to spark a challenging yet imperative discussion regarding the inherent biases present in generative AI systems and their resulting implications. The development of generative AI models that possess inherent biases in producing AI content has the potential to undermine trust in AI technology and also limit the widespread benefits that AI may offer. Mitigating bias in AI necessitates substantial efforts, encompassing enhancements in data quality, refinements in algorithms, utilisation of fairness-aware strategies, and adherence to ethical principles that prioritise transparency and responsibility. Future research will play a crucial role in encouraging both the public and private sectors to adopt improved criteria for the development and assessment of AI systems, prioritising equality and accessibility over financial gains.

References

Birhane, A.; Prabhu, V. U.; and Kahembwe, E. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*.

Mantelero, A. 2018. AI and Big Data: A blueprint for a human rights, social and ethical impact assessment. *Computer Law & Security Review*, 34(4): 754–772.

Ntoutsi, E.; Fafalios, P.; Gadiraju, U.; Iosifidis, V.; Nejdl, W.; Vidal, M.-E.; Ruggieri, S.; Turini, F.; Papadopoulos, S.; Krasanakis, E.; et al. 2020. Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *ICML*. PMLR.

Shalini, G.; Hegde, V. S.; Soumya, M.; and Korkoppa, M. 2020. Provenance and Implications of Heavy Minerals in the Beach Sands of India's Central West Coast. *Journal of Coastal Research*.

Srinivasan, K.; Raman, K.; Chen, J.; Bendersky, M.; and Najork, M. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *ACM SIGIR Conference on Research and Development in Information Retrieval*.