Generative Models for Art and Society

Yankun Wu

Osaka University, Osaka, Japan
yankun@is.ids.osaka-u.ac.jp

Abstract

Text-to-image models have demonstrated remarkable capabilities in producing high-fidelity images from natural language prompts. The widespread application and increasing accessibility of pioneering models, such as Stable Diffusion, have gained significant attention regarding the impact of generated images on representations in downstream tasks. Concurrently, ethical considerations on text-to-image generation have emerged especially regarding gender bias. This paper presents three projects that explore generative models on their capabilities and bias. The first project leverages Stable Diffusion to disentangle content and style in art paintings, paving the way for applying the generative model to digital humanities. The second project evaluates gender bias in text-to-image generation, analyzing its origins and manifestations in generated images. The third project presents a survey on societal bias evaluation in generative models, targeting to synthesize current research and provide insights into future directions. Through these projects, we aim to contribute to the growing body of knowledge on the applications and potential societal impacts of text-to-image generation, fostering a more nuanced understanding of their capabilities and limitations.

Introduction

Text-to-image models have demonstrated remarkable capabilities in producing high-fidelity images from natural language prompts. The widespread application and increasing accessibility of pioneering models, such as Stable Diffusion (Rombach et al. 2022), have gained significant attention regarding the impact of generated images on representations in downstream tasks (Wu, Nakashima, and Garcia 2024). Concurrently, ethical considerations on text-to-image generation have emerged especially regarding gender bias. This paper presents three projects that explore generative models on their capabilities and bias. The first project leverages Stable Diffusion to disentangle content and style in art paintings, paving the way for applying the generative model to digital humanities (Wu, Nakashima, and Garcia 2023) (Sec.). The second project evaluates gender bias in text-to-image generation (Wu, Nakashima, and Garcia 2024), analyzing its origins and manifestations in generated images (Sec.). The third project presents a survey on societal bias evaluation in generative models, targeting to synthesize current research and provide insights into future directions (Sec.). Through these projects, we aim to contribute to the growing body of knowledge on the applications and potential societal impacts of text-to-image generation, fostering a more nuanced understanding of their capabilities and limitations.

Completed work: Stable Diffusion for content and style disentanglement in art

Background. Content and style are two fundamental elements in the analysis of art. *Content* answers *what the artwork is about*, delivering the semantics depicted in the image. In contrast, style portrays the visual appearance of the image, conveying *how the artwork looks*. While humans can easily distinguish between content and style, the boundary is less clear from a computer vision perspective. Most studies in art analysis regarding content and style, rely on full supervision (Garcia, Renoust, and Nakashima 2019). However, the current label, often single words, fails to capture the subtle characteristics of individual images, instead summarizing only general traits of artwork sets.

Method. To overcome the limitations of human annotations, we propose a novel approach that leverages the generative power of text-to-image models to learn disentangled content and style embeddings of paintings (Wu, Nakashima, and Garcia 2023). Our method, named GOYA (disentanGlement of cOntent and stYle with generAtions), utilizes the knowledge distilled from Stable Diffusion as a prior with contrastive learning. We create prompts by combining content and style descriptions, separated by commas, and use these prompts to generate synthetic images with Stable Diffusion. Subsequently, we extract initial embeddings from the synthetic images using a frozen CLIP image encoder (Radford et al. 2021). Two independent transformation networks are then employed to disentangle content and style embeddings through contrastive learning. This simple yet powerful approach allows us to train GOYA on synthetic images, exploring the application of generative models to art analysis.

Results. We evaluate GOYA on test set from the WikiArt dataset (Tan et al. 2019) on three tasks: content-style disentanglement, art classification, and art retrieval. We compute the Distance Correlation (DC) (Liu et al. 2021) between

content and style embeddings to evaluate disentanglement. Our results demonstrate that GOYA achieves effective disentanglement, despite relying solely on synthetic images. For art classification, we train two independent classifiers with a single linear layer on top of the examined embeddings. When trained on diffusion-generated images, GOYA achieves the best performance among all models tested. In similarity retrieval tasks, we observe that paintings retrieved in the content space generally depict scenes similar to the query image, while those in the style space tend to exhibit similar styles but different content. Our experiments show that the knowledge in Stable Diffusion can be effectively distilled for art analysis and performs well in disentanglement, classification, and art retrieval, shedding light on the adoption of generative models in the analysis of the digital humanities.

Completed work: Gender bias evaluation on Stable Diffusion

Research questions. Previous studies have shown that certain adjectives or professions can lead to generating stereotypes regarding the demographic attributes of faces. However, beyond the areas of faces, disparities between genders may also exist in other parts. To automatically evaluate gender bias in text-to-image models, we propose an evaluation protocol (Wu, Nakashima, and Garcia 2024). This protocol begins by generating counterfactual feminine and masculine prompts from neutral prompts (e.g., `a person in a park`), which are derived from captions in vision-language datasets. Our approach allows us to formulate and address three key research questions (RQs):

- **RQ1** Do images generated from neutral prompts exhibit greater similarity to those generated from masculine prompts than to images generated from feminine prompts, and if so, why?
- **RQ2** Do object occurrences in images significantly vary based on the gender specified in the prompt? If yes, do these object occurrences from neutral prompts exhibit greater similarity to those from masculine or feminine prompts?
- **RQ3** Does the gender in the input prompt influence the prompt-image dependencies in text-to-image models, and if so, which prompt-image dependencies are more predisposed to be affected?

Proposed protocol. Our evaluation protocol addresses each research question through distinct analytical approaches. To answer RQ1, we investigate **representational disparities** to study gender bias within the internal components of Stable Diffusion models. By computing cosine similarities in the disparities in (neutral, feminine) and (neutral, masculine) pairs, we aim to reveal if gender bias originates from the input prompt and how it perpetuates through the generation process. For RQ2, we extract objects from the generated images, and analyze object co-occurrence similarity and bias score between image pairs generated from triplet prompts. This analysis helps us uncover the influence of gender on objects in image generation. To address RQ3, we extract objects from both input prompt and generated images, and then categorize them into five groups based on proposed **prompt-image dependencies**: explicitly generated, implicitly generated, explicitly independent, implicitly independent, and hidden. These groups represent different relationships between objects in the prompts and images. We then analyze prompt-image dependencies among triplet prompts, conduct statistical tests, and compute bias score to quantify the differences introduced by gender. This comprehensive protocol allows us to systematically evaluate gender bias across various aspects of text-to-image generation, from internal model representations to object occurrences and prompt-image dependencies.

Findings. Our experiments reveals that gender bias extends beyond people's representations, permeating through the entire image and affecting the generated objects. Through the generation of free-form triplet prompts differing only gender indicators, our findings indicate that: 1) Prompts that use *neutral* words to refer to people consistently yield images more similar to the ones generated from counterfactual *masculine* prompts than from counterfactual *feminine* prompts; 2) There are statistically significant differences in the objects generated in the image based on the gender indicators in the prompt; 3) The frequency of objects generated explicitly from prompts exhibit similar behavior for different genders; 4) Objects not explicitly mentioned in the prompt exhibit significant differences for each gender; 5) We particularly observed significant statistical disparities in generated objects based on gender in items related to clothing and traditional gender roles such as sports, which are highly skewed towards images generated from *masculine* prompts, and food, which are skewed towards images generated from *feminine* prompts. Based on these observations, we provided recommendations for developers and users to reduce such representational disparities and gender bias in the generated images. We hope these insights contribute to underscoring the nuanced dynamics of gender bias in image generation, offering a new and valuable perspective to the growing body of research on this topic.

Ongoing work: A survey on societal bias evaluation in generative models

This project aims to review recent work on societal bias evaluation in generative models, including gender bias, skintone bias, and age bias in text-to-image generation, and image-to-image generation. Bias evaluation is crucial for regulating the development of generative models. Unlike well-established metrics for evaluating image quality or fidelity, the evaluation of bias presents challenges and lacks standard approaches. By analyzing recent work and discussing trends, we aim to provide insights for future work and bias mitigation strategies.

Our initial focus is on gender bias evaluation in text-to-image generation, involving bias evaluation setup, bias evaluation metrics, and findings and trends. A consistent finding across various studies is that models tend to generate *man* more frequently for professions. This tendency is observed in Stable Diffusion, DALL E 2 (Ramesh et al. 2022) and

minDALL-E (Kim et al. 2021; Cho, Zala, and Bansal 2023). Furthermore, specific professions, such as singers, may exhibit different bias tendencies (Cho, Zala, and Bansal 2023). Beyond gender bias in professions, models like Stable Diffusion and minDALL-E have shown a tendency to generate gender-specific attire, such as skirts for *woman* and suits for *man* (Wu, Nakashima, and Garcia 2024; Cho, Zala, and Bansal 2023). Moreover, the Safety Checker module in Stable Diffusion is prone to label *female* as "unsafety" (Garcia et al. 2023). An emerging trend is the increasing comprehensiveness of model evaluations, with a broader range of models, diverse prompts, and multiple axes of bias assessment. Recent work is also focusing on a more detailed examination of bias sources, offering valuable insights for future bias mitigation methods (Wu, Nakashima, and Garcia 2024).

As we continue to review recent research on societal bias in generative models, we hope that this survey will provide crucial insights for future work, contributing to the development of more equitable and unbiased AI systems.

Conclusion

This paper presented three projects exploring the applications and societal bias of generative models. We introduced GOYA for art analysis, proposed an evaluation protocol for gender bias in text-to-image models, and surveyed societal bias evaluation in generative models. Future work should focus on safe and responsible usage, harnessing the capabilities while minimizing potential harm.

References

Cho, J.; Zala, A.; and Bansal, M. 2023. Dall-Eval: Probing the reasoning skills and social biases of text-to-image generation models. In *ICCV*.

Garcia, N.; Hirota, Y.; Wu, Y.; and Nakashima, Y. 2023. Uncurated Image-Text Datasets: Shedding Light on Demographic Bias. In *CVPR*.

Garcia, N.; Renoust, B.; and Nakashima, Y. 2019. Context-aware embeddings for automatic art analysis. In *ICMR*.

Kim, S.; Cho, S.; Kim, C.; Lee, D.; and Baek, W. 2021. minDALL-E on Conceptual Captions. https://github.com/kakaobrain/minDALL-E.

Liu, X.; Thermos, S.; Valvano, G.; Chartsias, A.; O'Neil, A.; and Tsaftaris, S. A. 2021. Measuring the Biases and Effectiveness of Content-Style Disentanglement. In *BMVC*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.

Tan, W. R.; Chan, C. S.; Aguirre, H.; and Tanaka, K. 2019. Improved ArtGAN for Conditional Synthesis of Natural Image and Artwork. *TIP*.

Wu, Y.; Nakashima, Y.; and Garcia, N. 2023. Not Only Generative Art: Stable Diffusion for Content-Style Disentanglement in Art Analysis. In *ICMR*.

Wu, Y.; Nakashima, Y.; and Garcia, N. 2024. Stable diffusion exposed: Gender bias from prompt to image. In *AIES*.