Final paper proposal

Junbo He UNI: jh4737

Yuehui Ruan UNI: yr2453

Team name: fancy image

1. what the topic is?

Bias in AI Image Generation

Overview:

This is copied from midterm paper proposal since we are working on the same topic.

Advancement of AI image generation powered by deep learning models and diffusion models are remarkable in recent years. Those models are capable of creating both realistic and imaginative images by inputting concrete or general textual prompts. The AI image generation technique is now thriving in many industries including entertainment, design and advertising. Unlike humans, who can consciously apply bias, the existence of bias in AI model may seem surprising and mysterious at first glance. The bias actually originated from many sources like training datasets, design choices, stereotypes and human interference.

2. *why would we like to continue studying it?*

A small experiment has been conducted in my midterm paper. The results are not satisfying since some data are too extreme and some are less significant. In the analysis, we discovered some flaw in methodology and hope we could improve it in the future. The experiment being conducted in final project would shift from image reference based bias detection to text prompt only bias detection as we found some existing models are not suitable for image reference. Because of this, we have more opportunities to deliberate on design of prompt for additional type of bias such as social/economic status and health state. We want systematically testing more current trending models as well as those appeared in past papers for comparison. This gives us a chance to validate some old results by ourselves and see how is the bias in image generation model evolve over years. To bring more authenticity to the result, this time we plan to introduce new metrics and benchmarks including Inception Score (IS) for image diversity, CLIP Score for alignment, and ImageNet for benchmark evaluation. At the end, we plan to deploy multiple mitigation strategies on different models and assess the effectiveness.

Methodology (skeleton)

1. Select trending current model and usable model in past papers.
2. Design prompt and generate image
3. Record results for bias identification.
4. Evaluation through benchmark and metrics
5. Deploy mitigation method and assess the result.

Reference

1. Luo, Hanjun, et al. “BIGbench: A Unified Benchmark for Evaluating Multi-Dimensional Social Biases in Text-To-Image Models.” *ArXiv.org*, 2024, arxiv.org/abs/2407.15240.

2. Zhu, Mingjian, et al. “GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image.” *Advances in Neural Information Processing Systems*, vol. 36, 15 Dec. 2023, pp. 77771–77782, proceedings.neurips.cc/paper_files/paper/2023/hash/f4d4a021f9051a6c18183b059117e8b5-Abstract-Datasets_and_Benchmarks.html.