Diffusion Model Alignment Using Direct Preference Optimization

What is the problem?

This paper addresses the challenge of aligning text-to-image diffusion models with human preferences. While Large Language Models (LLMs) often use techniques like Reinforcement Learning from Human Feedback (RLHF), diffusion models have primarily relied on less flexible methods such as fine-tuning on curated datasets or using limited reward models. These approaches have been inadequate for achieving robust alignment across diverse preferences and an open vocabulary of prompts.

What has been done earlier?

1. Fine-tuning on curated datasets: Many approaches, including those used in Stable Diffusion and Emu, rely on fine-tuning pretrained diffusion models on datasets of high-quality images and captions.
2. Reward maximization training: Methods like DRAFT and AlignProp attempt to directly optimize the diffusion model to maximize a reward function calculated on generated images.
3. Inference-time optimization: DOODL focuses on optimizing diffusion latents during inference to improve generated images based on a specific criterion, similar to techniques used in CLIP+VQGAN
4. Reinforcement Learning (RL) based methods: DPOK and DDPO employ RL techniques to align diffusion models with human preferences.

Ardhi Dattatreya Varma, B421009

Diffusion Model Alignment Using Direct Preference Optimization

What are the remaining challenges? What novel solution proposed by the authors to solve the problem?

Remaining Challenges:

- The authors highlight ethical concerns in text-to-image generation, particularly from web-collected data. Key issues include generating harmful, fake, or explicit content, and the risk of bias propagation from training data or labelers. Ensuring diverse and representative labelers is essential to mitigate biases. Additionally, preference modeling can unintentionally produce suggestive or biased images. Open research questions include exploring online learning methods for better performance and adapting the model for personalized preference tuning.

Novel Solution Proposed:

- The authors introduce Diffusion-DPO, a method adapted from Direct Preference Optimization (DPO) for aligning diffusion models with human preferences. It addresses the challenge of optimizing diffusion models by leveraging the Evidence Lower Bound (ELBO) as a surrogate objective. Through mathematical derivations, they develop a tractable loss function that enhances denoising for preferred images while reducing it for non-preferred ones. Diffusion-DPO also learns an implicit reward model by estimating differences in denoising errors, allowing for effective preference alignment without costly dataset curation or complex reward function engineering.