Multi-Adapter joint fine-tuning of Diffusion Models with LoRA for Visual Illusions

Nossaiba Kheiri , Hao Liu, Hansen Lin

Under Supervision of Dr. Ali Hirsa ,Miao Wang And Gary Kazantsev at Bloomberg

Motivation

We present a novel approach to fine-tuning large diffusion models for generating illusion-based images using a GPT-based reward model. We build on the illusion generation mechanism in the **Visual Anagrams** paper [?] to train the model to generate images that are difficult for GPT-4 to distinguish from non-illusions. Our approach incorporates **LoRA (Low-Rank Adaptation)** for efficient fine-tuning and **multi-adapter training** using the **PEFT** framework, allowing the model to handle complex visual styles and transformations as training tasks.

We explore the extent to which state-of-the-art LLMs understand complex visual phenomena that challenge human perception.We generate optical illusions and assess the performance of LLMs (specifically GPT-4) in illusion detection.

Problem: Failure mode of GenAI for image generation

❶Originating from the diffusion model [1]
- Independent Synthesis
- Noise shift
- Correlated noise

❷Originating from the illusion pairing model

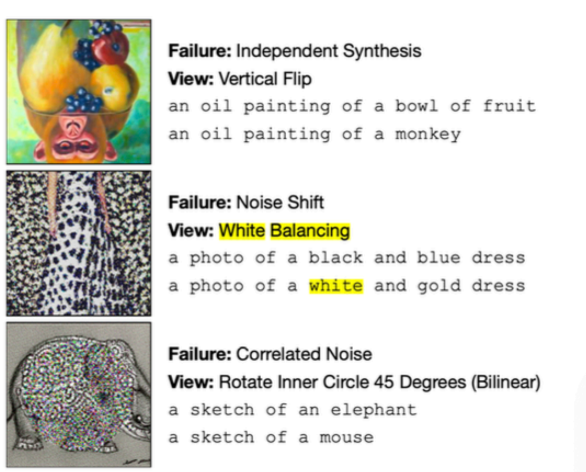


Figure 9. **Failures.** We highlight three interesting failure cases,

Methodology

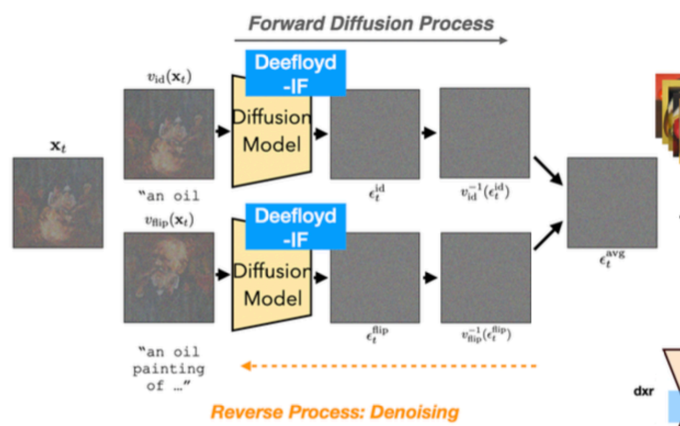
❶Text conditioned Diffusion Process

❷Independent Low Rank Adaptation (LoRA) training

❸Joint task Fine tuning

Using **classifier free guidance**

$$\epsilon_{\text{CFG}} = \epsilon(x,t,\varnothing) + \gamma\left(\epsilon(x,t,y) - \epsilon(x,t,\varnothing)\right)$$



LoRA

$$A_{ij} \leftarrow A_{ij} - \eta \frac{\partial \mathcal{L}}{\partial A_{ij}}$$

$$B_{kl} \leftarrow B_{kl} - \eta \frac{\partial \mathcal{L}}{\partial B_{kl}}$$

Joint Training

$$W_{\text{final}} = \alpha W_{\text{adapter1}} + \beta W_{\text{adapter2}}$$

Trainable parameters

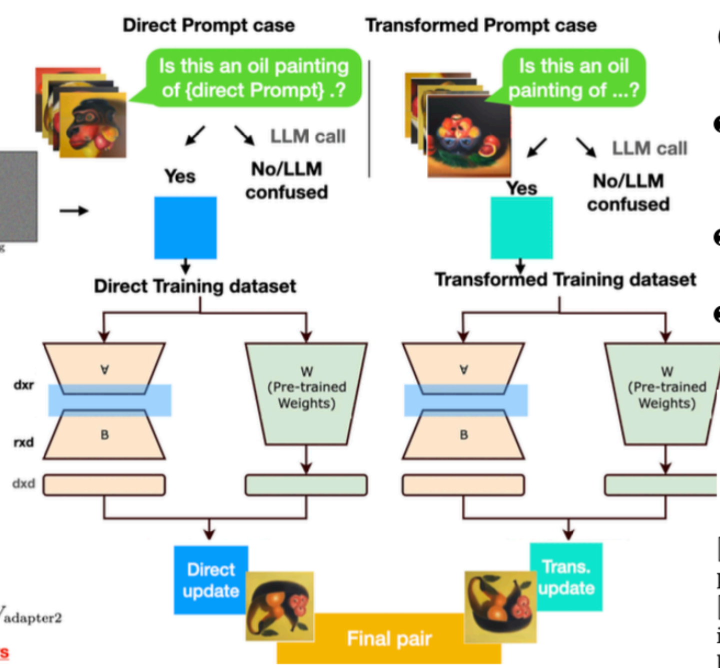
After the diffusion model generates an image I, GPT evaluates the probability P_{GPT} that the image contains an illusion. **The reward function** $R(I)$ is defined as:

$$R(I) = 1 - P_{\text{GPT}}(\text{illusion}|I)$$

Where: $P_{\text{GPT}}(\text{illusion}|I)$ is the probability that the LLM detects the illusion in the image I.

LoRA and GPT as the Reward Model (RL Agent)

Low-Rank Adaptation (LoRA) **reduces the number of trainable parameters** during fine-tuning by introducing low-rank matrices updates to the weight matrices of a pre-trained model to capture important **task-specific information.**



Possible future work: Online Reinforcement Learning

The proposed approach can be extended to an **online reinforcement learning (RL)** framework. In this setup, the model generates images in real-time and receives feedback from GPT, updating its parameters dynamically.

In this scenario, the model continuously improves by learning from its performance during inference.

Conclusion

The model learns to generate illusionary images that are increasingly difficult for GPT to detect. Multi-adapter training allows flexible blending of original and transformed prompts, enhancing the model's ability to produce complex visual illusions. This approach is computationally scalable and adaptable to real-time online reinforcement learning, making it suitable for advanced image generation tasks.

Objectives/Innovation/Usage

❶Adjust the finetuning by replicating the human behavior on illusions (as in [2]) to the LLM's (GPT model) behavior.

❷Illusion generation using diffusion model and parallel denoising

❸Fine-tuning illusions to be harder to detect

References

[1] "DreamBooth Fine-Tuning with LoRA," , peft/main/en/task$_g$uides/dreambooth$_l$ora.

[2] Ying Fan et al., "DPOK: Reinforcement Learning for Fine-Tuning Text-to-Image Diffusion Models", https://doi.org/10.48550/arXiv.2305.16381.

[3] Daniel Geng, Inbum Park, and Andrew Owens, "Visual Anagrams: Generating Multi-View Optical Illusions with Diffusion Models", http://arxiv.org/abs/2311.17919.

Weights & Biases (Wandb) Training 1

Github