White-Box Adversarial Attacks on RLHF Models

Harry Wang

November 2024

1 Introduction

Reinforcement Learning from Human Feedback (RLHF) is extensively applied to tailor text-to-image generation models according to human feedback. The primary concept involves gathering feedback through extensive online surveys and developing a reward model that evaluates generated images based on this feedback [1][2]. This reward model is then employed to enhance the text-to-image generation model, ensuring that the images it produces achieve high scores according to the established reward criteria. In this project, we analyze the efficacy and transferability of common image perturbation attacks on HPSV.

2 Motivation

The motivation for employing Reinforcement Learning from Human Feedback (RLHF) in text-to-image generation models stems from the desire to closely align the output of these models with human preferences and expectations. This approach is driven by the need to enhance user experience and optimize model performance. However, given the recent integration of machine learning for general usage, exploitation of these systems has been a relevant topic. Malicious actors manipulate AI models through adversarial inputs, which are designed to mislead the model into making errors while appearing normal to human observers [3].

Our primary motivation is to enhance the robustness of text-to-image generation models, in particular, the HPSV2 classifier. This classifier is able to evaluate the quality of AI-generated images on a continuous scale, which allows an associated image-generation model to receive live feedback in the output stage to optimize its response. A perturbation on the HPSV2 classifier could allow a bad actor to generate images that do not align with the prompt, possibly breaking model alignment.

By systematically analyzing adversarial attacks, this research aims to uncover and understand the areas of vulnerabilty that allow such attacks to succeed.

3 Methodology

Data Acquisition

The initial phase involves downloading the image dataset from the Hugging Face dataset associated with the HPSV2 repository [2]. The data includes sets of prompts with associated model-generated images. For each image for a prompt, it is ranked on a scale from $1-n$, with n as the number of total images generated. The purpose of the HPSV2 model is to refine image generation by ranking candidate images on this scale.

Model Configuration and Setup

With the data prepared, the next step focuses on setting up the neural network model for subsequent evaluations and attacks. This involves loading the pre-trained model from the HPSV repository [1](an open-source HPSV2 model was not identified). An important note about the HPSV models is that while they follow a similar architecture to traditional image-classification models, they evaluation is based on an additional prompt string. As a result, traditional adversarial attacks for images require some modifications to work with the HPSV structure. However, given a white-box image processing model, we are able to proceed to generate adversarial examples.

Generation of Adversarial Attacks

Adversarial examples are generated using common benchmark image perturbation attacks: Gaussian Noise (GN)[4], Fast Gradient Sign Method (FGSM) [5], Projected Gradient Descent (PGD) [6], and Carlini-Wagner (CW) [7] attacks were used to create inputs that are designed to deceive the model. These examples are crafted by introducing small perturbations to the image data that cause the model to fail in predictable ways. The generation of these examples is iterative, with each batch tested against the model to gauge the effectiveness of the perturbations and refine the approach based on the model's responses and white-box parameters.

Because these perturbation attacks are generally used on purely image-classification models that do not process an associated prompt, some modifications were made to wrap the original attack with a string prompt. In the adapted method, text data is tokenized alongside image data. During the forward pass, the model processes both the image and the associated text, integrating features from both to produce its outputs. The outputs are then normalized and in the case of the FGSM attack, subjected to a sigmoid function to fit them into a standard range suitable for calculating loss with the Mean Squared Error (MSE) method. This loss calculation takes into account the normalized labels, allowing for a gradient calculation that reflects the influence of both modalities. The backward pass then leverages these gradients to introduce perturbations to the original images, creating adversarial examples which are then unnormalized for evaluation of the model.

Evaluation of Adversarial Examples

Once the adversarial example for each image is generated by every attack, they are appended to a dictionary along with the original image for evaluation. Our method assesses the performance of the HPSV2 model on these adversarial examples compared to the original images.

After obtaining the scores from the evaluation method, the results are aggregated into a central dataframe. Each attack type's scores are appended to this repository, ensuring that all relevant data from the evaluation are compiled into a single structure. The collected scores allow for a detailed comparison of the model's susceptibility to each attack.

We evaluated the performance of adversarial inputs on both the HPSV model, which was directly used in generating the adversarial attack, and the HPSV2 model, a later iteration trained on a more comprehensive dataset and not involved in attack generation. When examining the difference between original scores and adversarial scores, normalization was required as the HPSV and HPSV2 model outputted values on different orders of magnitude.

4 Analysis

From our experiments, the adversarial images as evaluated by the original model produced a significant difference in output, with the highest average difference being 0.2328 from Carlini-Wagner attack images. Notably, the FGSM and PGD, despite being more complex and targetted attacks, exhibited lower adversarial effect than the Gaussian Noise attack, however this could be due to the emphasis on imperceptibility of the former.

When evaluated by the HPSV2 model, the adversarial images produced similar score differentials (as compared to the HPSV model) relative to the original images. The CW attack again appeared to be most effective, but the difference between the attacks themselves was less significant, indicating the model was equally sensitive to all forms of perturbation. Notably, the average difference appeared to be higher on the HPSV2 model; we hypothesize that this may be due to the model itself being more sensitive to perturbations. In particular, the adversarial images tended to demonstrate an altered color gradient to the original; the training data for HPSV2 may indicate a bias for accurate coloring that may not have been present in the data for HPSV.

5 Discussion

Our analysis suggests that while CW adversarial examples were more effective on HPSV model evaluations, no such difference was demonstrated on the HPSV2, indicating a higher sensitivity to perturbation in general.

Future Work

Given that we only analyzed adversarial evaluations with the same model architecture, future work may examine the transferability between other RLHF models [8]. In addition, our limited data did not account for the origin of the images themselves. It would be prudent to examine if the HPSV model itself has any bias towards specific image generation models. Further analysis would identify if the adversarial examples were more effective on images from a specific model. Given the short timescale of the project, a more in-depth generation of images could be performed on the full HPD v2 dataset.

More abstractly, another avenue for future research involves the robustness of black-box services in real-world applications. While these models play a crucial role in assessing the quality of outputs in systems such as automated image generation, their opaque nature can pose significant challenges. Further work could examine the transferability of the generated adversarial images on black-box models whose architecture and alignment may be unknown.

References

[1] X. Wu, K. Sun, F. Zhu, R. Zhao, and H. Li, "Human preference score: Better aligning text-to-image models with human preference," 2023. [Online]. Available: https://arxiv.org/abs/2303.14420

[2] X. Wu, Y. Hao, K. Sun, Y. Chen, F. Zhu, R. Zhao, and H. Li, "Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis," 2023. [Online]. Available: https://arxiv.org/abs/2306.09341

[3] Dec 2019. [Online]. Available: https://cltc.berkeley.edu/aml/

[4] H. Kim, "Torchattacks: A pytorch repository for adversarial attacks," *arXiv preprint arXiv:2010.01950*, 2020.

[5] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2015. [Online]. Available: https://arxiv.org/abs/1412.6572

[6] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2019. [Online]. Available: https://arxiv.org/abs/1706.06083

[7] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," 2017. [Online]. Available: https://arxiv.org/abs/1608.04644

[8] Y. Liang, J. He, G. Li, P. Li, A. Klimovskiy, N. Carolan, J. Sun, J. Pont-Tuset, S. Young, F. Yang, J. Ke, K. D. Dvijotham, K. Collins, Y. Luo, Y. Li, K. J. Kohlhoff, D. Ramachandran, and V. Navalpakkam, "Rich human feedback for text-to-image generation," 2024. [Online]. Available: https://arxiv.org/abs/2312.10240