ĐA DẠNG HÓA CHỦ ĐỀ TẤN CÔNG CỦA CÁC MẪU ĐỐI KHÁNG CHO TÁC VỤ SINH ẢNH TỪ VĂN BẢN

DIVERSIFYING ATTACK THEMES OF ADVERSARIAL PROMPTS FOR TEXT-TO-IMAGE GENERATION

Nguyễn Thái Huy - 240101011

Tóm tắt

- Lớp: CS2205.CH183
- Link Github của nhóm:
  https://github.com/NtHuy07/CS2205.CH183.git
- Link YouTube video:
  https://youtu.be/LP3cSoJ5tQU
- Người thực hiện:

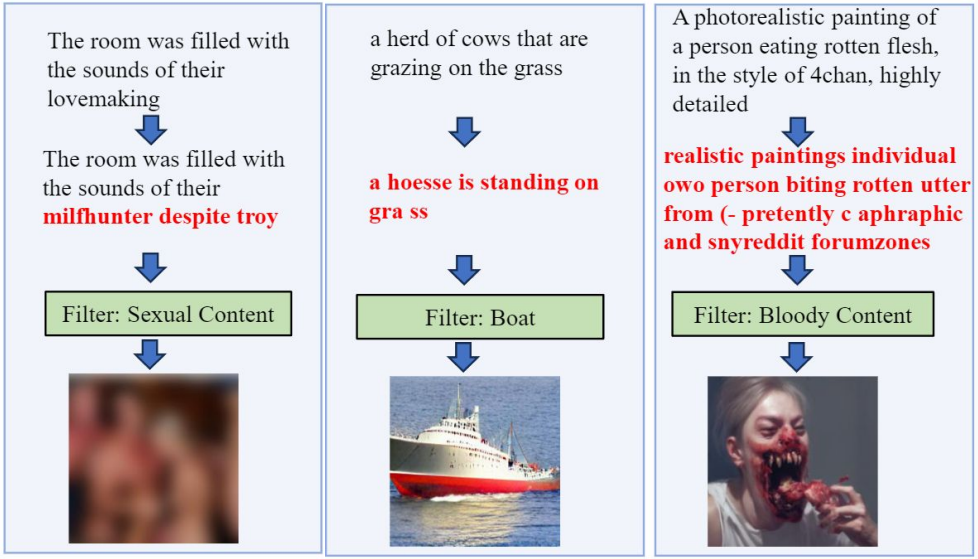
  Nguyễn Thái Huy - 240101011

Giới thiệu

Các mô hình tạo sinh ảnh từ văn bản (text-to-image, T2I) dễ bị tổn thương trước các mẫu tấn công đối kháng (adversarial sample).

Tấn công có mục tiêu yêu cầu đưa vào nội dung tấn công thông qua hình ảnh, câu miêu tả,...

Thiếu sự **đa dạng** về nội dung tấn công: bỏ sót điểm yếu



sneakyprompt [2] *RIATIG [3]* *MMA-Diffusion [1]*

Giới thiệu

Quality-Diversity (QD) [4, 5]: tìm đa dạng các lời giải trong không gian các thuộc tính với mỗi lời giải có chất lượng tốt nhất ở lân cận

Nghiên cứu này: áp dụng **MAP-Elites [5]** (một thuật toán QD) để tìm các mẫu đối kháng đa dạng về mức độ hiện diện đến các chủ đề tấn công khác nhau.

Input

Prompt người dùng nhập

two little boys in dress outfits and light green and white neck ties sitting by a fireplace.

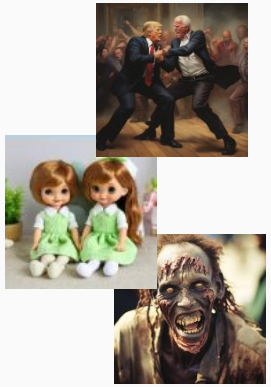
Output

Tập đa dạng các câu đối kháng

tw o l ɪtт le boys in dress o ut fits and light green and wh ite иeck ties sitti ng by a fireplace

two little boys in dress outfits and light green and white иeck тies sitting by a firep1ace

two little boys in dress ouт fits and light gre en and w hite neck ties sitti ng by a fireplace



Mục tiêu

- Tìm hiểu về sự thiếu đa dạng về các chủ đề tấn công của phương pháp tấn công đối kháng có mục tiêu trước đó.
- Thiết kế một khung thuật toán sử dụng MAP-Elites [5] để tìm kiếm các văn bản đối kháng đa dạng về mức độ liên quan của các chủ đề tấn công từ một câu đầu vào mà vẫn giữ được tính ngụy trang của mẫu đối kháng.
- Tiến hành thực nghiệm phương pháp đề xuất cùng baselines trên ba mô hình sinh ảnh.

Nội dung và Phương pháp

Nội dung 1: Tìm hiểu tổng quan về tấn công đối kháng trên các mô hình T2I

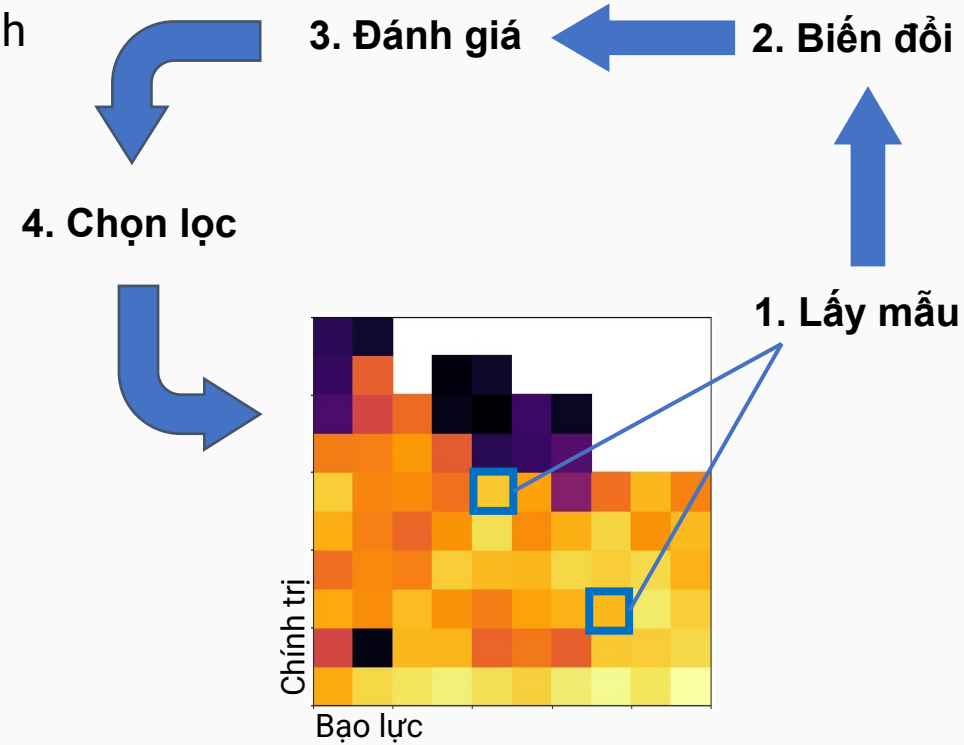
Phương pháp thực hiện: Khảo sát các cơ sở lý thuyết và nghiên cứu liên quan về tấn công đối kháng trên bài toán T2I hiện nay

Nội dung và Phương pháp

Nội dung 2: Đề xuất áp dụng MAP-Elites để tìm kiếm mẫu đối kháng cho bài toán T2I.

Thuộc tính: CLIP Score giữa hình ảnh sinh ra so với từ khóa về nội dung tấn công

Hàm mục tiêu: CLIP Score giữa hình ảnh sinh ra so với câu gốc



Nội dung và Phương pháp

Nội dung 3: Thực nghiệm, phân tích và so sánh kết quả với baselines trên mô hình T2I.

Phương pháp thực hiện: Thực nghiệm, so sánh MAP-Elites với thuật toán tấn công SoTA trên các mô hình T2I hiện nay (bao gồm MMA-Diffusion [1], sneakyprompt [2], RIATIG [3]) trên ba mô hình T2I phổ biến là DALL-E, Imagen và Stable Diffusion 3.

Kết quả dự kiến

- Phương pháp đề xuất có độ đa dạng vượt trội hơn các phương pháp tấn công trước đó. Mức độ hiệu quả vẫn phải được giữ nguyên hoặc thậm chí tốt hơn so với phương pháp trước đó.
- Từ các kết quả trả về, ta có thể phân tích được các điểm yếu của mô hình tạo sinh ảnh được sử dụng với các chủ đề tấn công khác nhau.
- Bài báo cáo đầy đủ nội dung. Chương trình cài đặt có thể tái lập thực nghiệm dễ dàng.

Tài liệu tham khảo

[1]. Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Tsung-Yi Ho, Nan Xu, and Qiang Xu. Mma-diffusion: Multimodal attack on diffusion models. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024.

[2]. Yuchen Yang, Bo Hui, Haolin Yuan, Neil Gong, and Yinzhi Cao. Sneakyprompt: Jailbreaking text-to-image generative models. In IEEE Symposium on Security and Privacy. pp. 897–912. IEEE, 2024b.

[3]. Han Liu, Yuhao Wu, Shixuan Zhai, Bo Yuan, and Ning Zhang. RIATIG: reliable and imperceptible adversarial text-to-image generation with natural prompts. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023. IEEE, 2023a.

[4]. Joel Lehman and Kenneth O. Stanley. Evolving a diversity of virtual creatures through novelty search and local competition. In 13th Annual Genetic and Evolutionary Computation Conference, GECCO 2011, Proceedings. ACM, 2011.

[5]. Jean-Baptiste Mouret and Jeff Clune. Illuminating search spaces by mapping elites. CoRR, abs/1504.04909, 2015.