THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút):
  https://youtu.be/LP3cSoJ5tQU
- Link slides (dạng .pdf đặt trên Github của nhóm):
  https://github.com/NtHuy07/CS2205.CH183/blob/main/slide.pdf
- *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
- *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*
- *Lớp Cao học, mỗi nhóm một thành viên*

- Họ và Tên: **Nguyễn Thái Huy**
- MSSV: **240101011**

- Lớp: CS2205.CH183
- Tự đánh giá (điểm tổng kết môn): 9.5/10
- Số buổi vắng: 1 (có phép)
- Số câu hỏi QT cá nhân: 6/6
- Số câu hỏi QT của cả nhóm: 6/6
- Link Github:
  https://github.com/NtHuy07/CS2205.CH183.git

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

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

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

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

TÓM TẮT *(Tối đa 400 từ)*

Các mô hình tạo sinh ảnh từ văn bản (text-to-image generation, T2I) phổ biến ngày nay thường dựa trên các mô hình mạng neural vốn rất dễ bị tổn thương trước các mẫu tấn công đối kháng. Các mẫu đối kháng này thường là các văn bản đầu vào được thay đổi rất ít mà mắt người không nhận ra nhưng làm cho mô hình hiểu sai ngữ nghĩa của đầu vào dẫn đến tạo ra các bức ảnh sai lệch, thậm chí là mang các nội dung nhạy cảm. Nhằm tìm hiểu về tính bền vững (robustness) của các mô hình trước các cuộc tấn công, nhiều nghiên cứu đã đề xuất các phương pháp giúp tự động tìm kiếm các văn bản đối kháng. Tuy nhiên, các phương pháp tấn công có mục tiêu vẫn chỉ giới hạn trong việc tìm ra một hoặc một vài mẫu đối kháng thỏa theo các yêu cầu được cung cấp trước mà không thể tự động bao quát nhiều trường hợp trên các chủ đề tấn công khác nhau. Quality-Diversity (QD) là nhóm phương pháp nhắm đến việc đa dạng hóa các lời giải tìm được trên một không gian thuộc tính với mỗi lời giải tìm được là tốt nhất có thể trong khu vực lân cận. Trong nghiên cứu này, chúng tôi áp dụng thuật toán MAP-Elites, một thuật toán QD, để tạo ra một tập các văn bản đối kháng đa dạng về mức độ liên quan của các nội dung tấn công khác nhau (như bạo lực, chính trị,...) có nội dung khác so với văn bản gốc nhưng vẫn có tính ngụy trang cao cho bài toán T2I. Tập các lời giải trả về của phương pháp có thể cho phép chúng ta hiểu rõ hơn về điểm yếu của mô hình đối với nhiều chủ đề khác nhau mà không cần phải cung cấp chi tiết về nội dung tấn công.

GIỚI THIỆU

Các mô hình tạo sinh ảnh từ văn bản (text-to-image generation, T2I) ngày nay cho thấy được khả năng ứng dụng rộng rãi của mình trong các lĩnh vực về nghệ thuật. Tuy nhiên, các mô hình này được xây dựng dựa trên mô hình học sâu vốn rất dễ bị tổn thương trước các **mẫu đầu vào đối kháng (adversarial sample).** Các mẫu này có thể vượt qua được các hệ thống kiểm duyệt và tạo ra các hình ảnh có nội dung sai lệch, thậm chí là mang tính chất phản cảm. Các phương pháp **tấn công có mục tiêu (target attack)** trước đó phải cung cấp trước ngữ cảnh tấn công chi tiết để phương pháp tấn công có thể dựa vào đó biến đổi. Chẳng hạn như MMA-Diffusion [1], sneakyprompt [2] tạo ra các văn bản có thể tránh bị các hệ thống kiểm duyệt phát hiện thông qua việc biến đổi một câu chứa nội dung nhạy cảm cho trước. Hay RIATIG [3] biến đổi rất ít câu miêu tả ban đầu nhưng khiến cho bức hình sinh ra phải khớp so với một bức hình mục tiêu (target image) được cung cấp. Tuy nhiên, các phương pháp này chỉ giới hạn tìm một số câu prompt đối kháng thỏa theo nội dung tấn công hẹp được cung cấp trước mà thiếu đi sự đa dạng về các chủ đề tấn công và cả mức độ hiện diện của nội dung này trong các mẫu tìm được. Sự thiếu đa dạng này có thể làm bỏ sót nhiều yếu điểm khác nhau của mô hình, trong khi đó việc tạo ra một tập các mục tiêu tấn công khác nhau một cách thủ công lại tốn thời gian và không có khả năng scale.

Trái ngược với các thuật toán tối ưu truyền thống, các thuật toán Quality-Diversity (QD) [4, 5] hướng đến việc tìm một tập các lời giải đa dạng, khác biệt với nhau trên không gian thuộc tính (thuộc tính do người dùng định nghĩa) và mỗi lời giải này có giá trị của hàm mục tiêu tốt nhất trong lân cận của nó. Trong nghiên cứu này, chúng tôi mong muốn có thể đa dạng hóa mức độ hiện diện của các chủ đề tấn công cho mẫu đối kháng. Vì vậy, chúng tôi áp dụng MAP-Elites [5], một thuật toán QD, để biến đổi từ *một câu miêu tả gốc bất kỳ do người dùng nhập* và trả về *một tập các câu đối kháng* mà khi đưa qua mô hình T2I có thể tạo ra hình ảnh có nội dung khác so với câu gốc nhưng đa dạng chủ đề tấn công ở các mức độ tương đồng khác nhau. Từ đây ta có khả năng hiểu rõ hơn độ bền vững (robustness) của các mô hình này trước các cuộc tấn công đối kháng trên nhiều nội dung tấn công.

MỤC TIÊU

- Tìm hiểu về sự thiếu đa dạng về các chủ đề tấn công (bao gồm cả chủ đề mang tính nhạy cảm) của phương pháp tấn công đối kháng có mục tiêu trước đó.
- Thiết kế một khung thuật toán sử dụng MAP-Elites 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: DALL-E, Imagen và Stable Diffusion 3 để làm rõ sự khác biệt khi áp dụng MAP-Elites giúp tăng sự đa dạng của quần thể các mẫu đối kháng.

NỘI DUNG VÀ PHƯƠNG PHÁP

<u>Nội dung 1</u>: Tìm hiểu tổng quan về tấn công đối kháng trên các mô hình sinh ảnh từ văn bản.

Phương pháp thực hiện: Khảo sát các cơ sở lý thuyết cùng các nghiên cứu liên quan về tấn công đối kháng trên bài toán T2I hiện nay. Các nội dung bao gồm động lực nghiên cứu, phương pháp và kết quả thực nghiệm của các công trình này trên các mô hình tạo sinh ảnh khác nhau.

Kết quả dự kiến: Tóm tắt về cơ sở lý thuyết và các công trình nghiên cứu liên quan cho bài toán tấn công đối kháng trên các mô hình T2I.

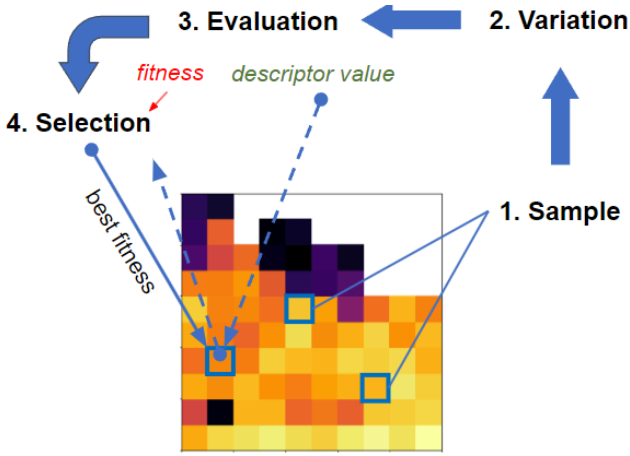
<u>Nội dung 2</u>: Đề xuất áp dụng thuật toán MAP-Elites để tìm kiếm các mẫu văn bản đối kháng đa dạng trên các chủ đề tấn công khác nhau cho bài toán T2I.

Phương pháp thực hiện: Thiết kế thuật toán sử dụng MAP-Elites (minh họa ở Hình 1) trong việc tìm kiếm các văn bản tấn công mô hình sinh ảnh. MAP-Elites [5] rời rạc không gian thuộc tính thành các ô (cells) tạo thành kho lưu trữ các lời giải. Ở mỗi vòng lặp, thuật toán lấy ngẫu nhiên các lời giải được lưu trong kho lưu trữ, biến đổi thành lời giải mới, thực hiện đánh giá giá trị thuộc tính và hàm mục tiêu (chất lượng). Sau đó lời giải nào có chất lượng cao hơn so với lời giải trong cùng ô chứa giá trị thuộc tính được lưu vào thay thế lời giải cũ. Các lựa chọn thiết kế bao gồm:

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

công được người dùng định nghĩa.

- Hàm mục tiêu: CLIP Score giữa hình ảnh sinh ra so với câu gốc (càng thấp càng tốt cho thấy nội dung khác so với câu ban đầu)
- Phép biến đổi: Kế thừa các phép biến đổi trước đó của RIATIG [3] cùng một số phép biến đổi tự thiết kế khác.



Hình 1: Minh họa các bước thực hiện của thuật toán MAP-Elites [5].

Kết quả dự kiến: báo cáo chi tiết về thuật toán đề xuất sử dụng MAP-Elites cho bài toán T2I giúp đa dạng hóa chủ đề tấn công

<u>Nội dung 3</u>: Tiến hành thực nghiệm, phân tích và so sánh kết quả với các phương pháp tấn công trước đó trên ba mô hình T2I.

Phương pháp thực hiện: Thực nghiệm, so sánh MAP-Elites với một số thuật toán tối ưu truyền thống và 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: Kết quả so sánh và phân tích về mặt định lượng cho khả năng gia tăng độ đa dạng của phương pháp đề xuất so với các phương pháp khác được đề cập cho bài toán T2I. Thực hiện phân tích cụ thể về mặt định tính các mẫu đối kháng tìm được, từ đó thực hiện các kết luận về điểm yếu của 3 mô hình.

<u>Nội dung 4</u>: Viết một báo cáo nghiên cứu và đóng gói chương trình cài đặt.

Phương pháp thực hiện: Tổng hợp lại nội dung nghiên cứu, kết quả thực nghiệm; Viết báo cáo về quá trình nghiên cứu; Chỉnh sửa lại chương trình, thực hiện đóng gói.

Kết quả dự kiến: một báo cáo nghiên cứu và chương trình cài đặt cho phương pháp

KẾT QUẢ MONG ĐỢI

- Phương pháp đề xuất có độ đa dạng vượt trội hơn các phương pháp tấn công trước đó cho việc tìm kiếm các mẫu đối kháng trên bài toán T2I. Mức độ hiệu quả của mẫu tấn công tốt nhất của phương pháp đề xuất vẫn phải được giữ nguyên hoặc thậm chí tốt hơn so với mẫu tấn công tốt nhất của phương pháp trước đó.
- Từ các kết quả trả về của thuật toán, ta có thể phân tích được các điểm yếu của ba mô hình tạo sinh ảnh được sử dụng với các chủ đề tấn công khác nhau.
- Bài báo cáo quá trình và kết quả thực nghiệm đầy đủ nội dung. Chương trình cài đặt cho thuật toán có khả năng tái lập thực nghiệm dễ dàng.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[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. IEEE, 2024a.

[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.