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

Nguyễn Thái Huy[1,2]

[1] Trường Đại học Công nghệ Thông tin, TPHCM, Việt Nam

[2] Đại học Quốc gia TPHCM, Việt Nam

What ?

Chúng tôi đề xuất áp dụng thuật toán MAP-Elites để tạo ra một tập các văn bản đối kháng cho bài toán T2I

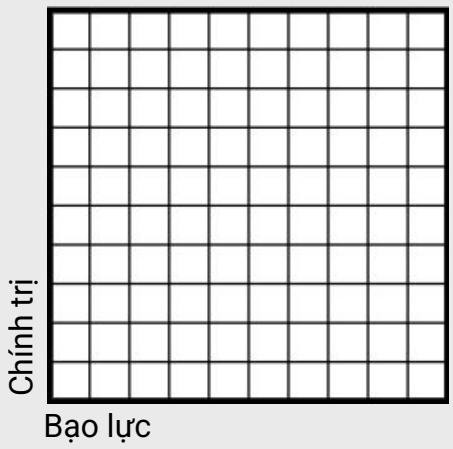
- Đa dạng về mức độ hiện diện nội dung tấn công
- Có nội dung khác so với văn bản gốc nhưng vẫn
- Có tính ngụy trang cao

Why ?

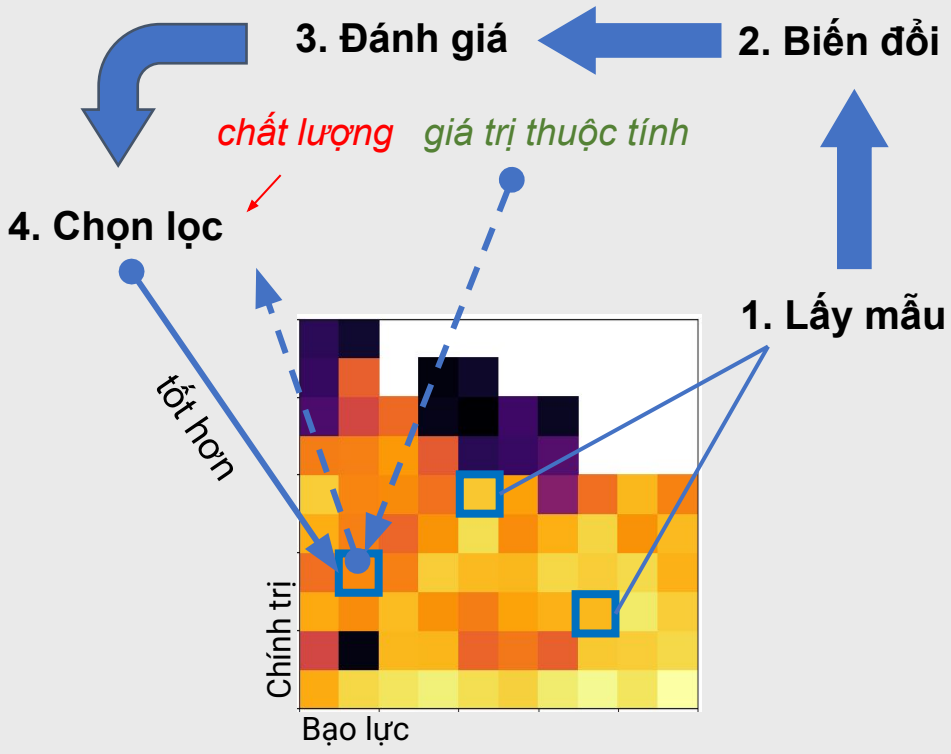
Các phương pháp tấn công trước đó chỉ giới hạn tìm một số câu prompt đối kháng thỏa theo nội dung tấn công hẹp mà thiếu đi sự đa dạng về các chủ đề tấn công và cả mức độ hiện diện của nội dung này trong các mẫu tìm được, bỏ sót nhiều yếu điểm khác nhau của mô hình,

Tổng quan

Khởi tạo → Tìm kiếm



Hình 1. Khởi tạo kho lưu trữ cho thuật toán MAP-Elites



Hình 2. Quá trình tìm kiếm của thuật toán MAP-Elites

Mô tả

1. Khởi tạo

- MAP-Elites rời rạc không gian thuộc tính thành các ô (cells) tạo thành kho lưu trữ các lời giải.
- **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

2. Các bước thực hiện

- **Bước 1:** Lấy ngẫu nhiên các lời giải được lưu trong kho lưu trữ
- **Bước 2:** Kế thừa các phép biến đổi trước đó của RIATIG cùng một số phép biến đổi tự thiết kế khác.
- **Bước 3:** Thực hiện đánh giá giá trị thuộc tính và hàm mục tiêu (chất lượng).
- **Bước 4:** Chọn lọc 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.

Nguyễn Thái Huy - Trường Đại học Công nghệ Thông tin, Đại học Quốc gia TPHCM
TEL : 0772687671 Email : huynt.19@grad.uit.edu.vn