Black-box Attacks on Image Activity Prediction and its Natural Language Explanations

Alina Elena Baia[1], Valentina Poggioni[2], Andrea Cavallaro[1,3,4]

[1]Idiap Research Institute, [2]University of Perugia, [3]École Polytechnique Fédérale de Lausanne, [4]Queen Mary University of London

Introduction

- adversarial attack: perturbs the input image to mislead a model
- black-box attack: uses only the final output of a model
- target model: a natural language explanation model (NL-XAI) that predicts a decision and generates both a textual and visual explanation
- scenarios:
  - 📊 change the prediction, keep the same textual explanation
  - 📋 keep the same prediction, change the textual explanation
- perturbation: unrestricted region-specific, generated using semantic colorization and image editing filters

Methodology

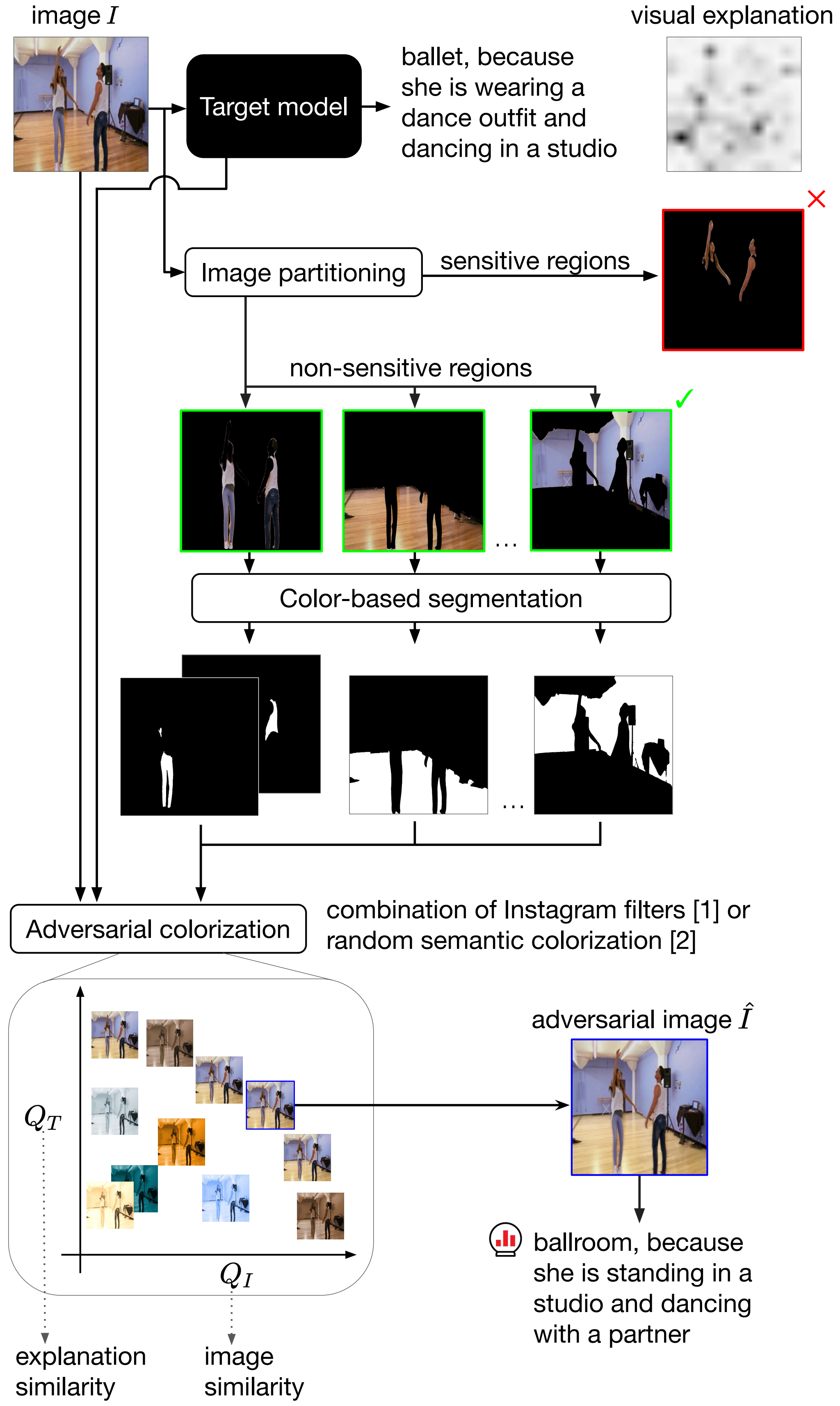


image I

Target model

ballet, because she is wearing a dance outfit and dancing in a studio

visual explanation

Image partitioning

sensitive regions

non-sensitive regions

Color-based segmentation

Adversarial colorization

combination of Instagram filters [1] or random semantic colorization [2]

Q_T

Q_I

adversarial image $\hat{I}$

📊 ballroom, because she is standing in a studio and dancing with a partner

explanation similarity

image similarity

$$\hat{I} = \arg\max_{\dot{I}} \left\{ Q_T\left(I, \dot{I}\right), Q_I\left(I, \dot{I}\right) \right\}$$

perturbed image

Validation

Dataset: ACT-X [3] for activity recognition tasks

Model: NLX-GPT [4] for prediction and explanation generation

Performance evaluation:

Success rate for 📊 :

$$S_r = \frac{1}{N} \sum_{j=1}^{N} \mathbb{1}_{\omega}, \ \mathbb{1}_{\omega} = \begin{cases} 1, & \text{if } a_j \neq \hat{a}_j \wedge Q_T\left(I_j, \hat{I}_j\right) \geq t \\ 0, & \text{otherwise} \end{cases}$$

predictions for images I_j and $\hat{I}_j$

number of images

similarity threshold

Image quality: MANIQA and Colorfulness

Cases:

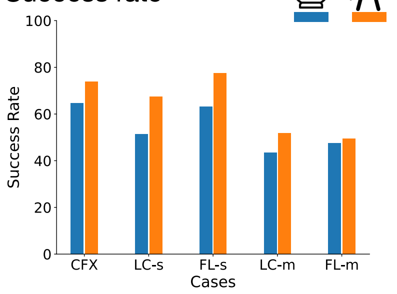
CFX: an adaption of ColorFool [2] with Q_T

FL-s (FL-m): full image filtering [1] with Q_T (and Q_I)

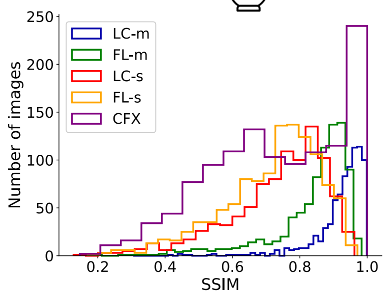
LC-s (LC-m): localized image filtering with Q_T (and Q_I)

Results

Success rate 📊 📋



SSIM distribution 📊



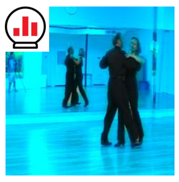
Samples of adversarial images

Original

ballroom, because he is wearing a suit and dancing with a woman in a dance studio

MANIQA: 0.69

Colorfulness: 23.73

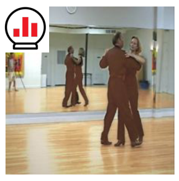


📊 CFX

tai chi, because he is standing in a studio and dancing with a woman

MANIQA: 0.64

Colorfulness: 147.56



📊 LC-m

ballet, because he is wearing a dance robe and dancing with a woman

MANIQA: 0.63

Colorfulness: 37.80

📋 CFX

ballroom, because he is standing on a wood floor with a woman on his shoulders

MANIQA: 0.70

Colorfulness: 37.90

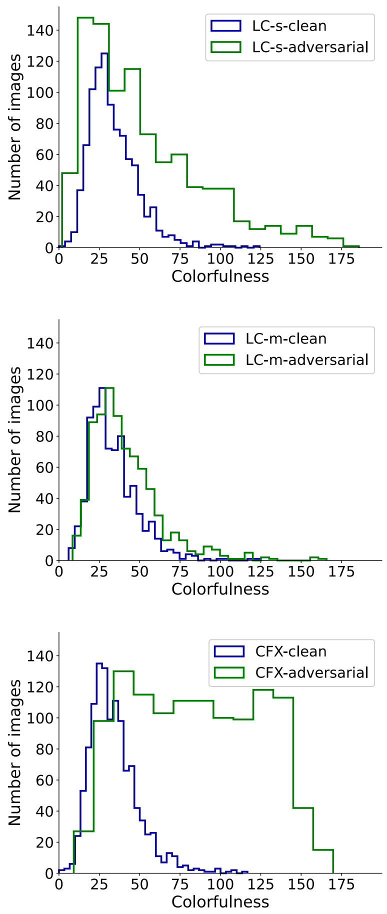
📋 LC-m

ballroom, because he is standing on a wood floor with a woman on his shoulders

MANIQA: 0.72

Colorfulness: 33.58

Colorfulness distribution 📊



Takeaways

- NL-XAI are vulnerable to black-box attacks
- prediction-explanation association can be disrupted with simple photo editing techniques
- straightforward assessment of explanations' robustness

References

[1] Alina Elena Baia, Gabriele di Bari and Valentina Poggioni, Effective universal unrestricted adversarial attacks using a MOE approach, EvoApp 2021.
[2] Ali Shahin Shamsabadi,Ricardo Sanchez-Matilla and Andrea Cavallaro, ColorFool: Semantic adversarial colorization, CVPR 2020.
[3] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata and others, Multimodal explanations: Justifying decisions and pointing to the evidence, CVPR 2018.
[4] Fawaz Sammani, Tanmoy Mukherjee and Nikos Deligiannis, NLX-GPT: A model for natural language explanations in vision and vision-language tasks, CVPR 2022.