As Firm As Their Foundations: Creating Transferable Adversarial Examples Across Downstream Tasks with CLIP

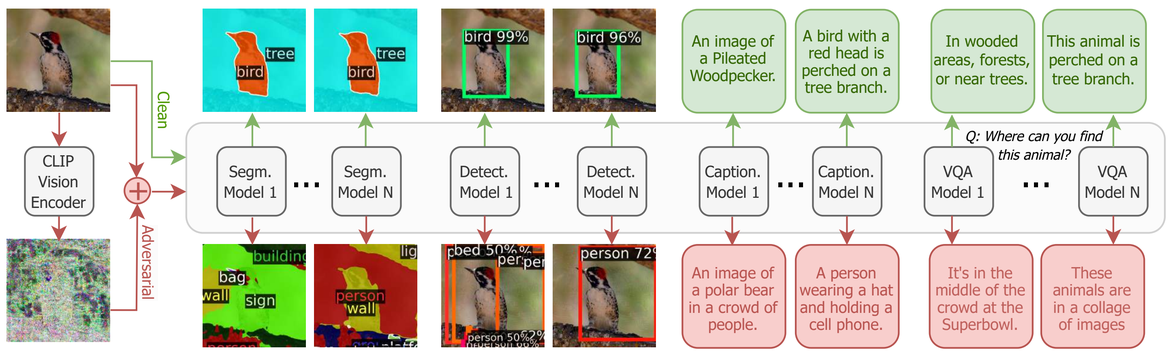
Anjun Hu [1] Jindong Gu [1] Francesco Pinto [1] Konstantinos Kamnitsas [1] Philip Torr [1]

[1]University of Oxford

Introduction

- Foundation models pre-trained on web-scale vision-language data, such as CLIP, are widely used as cornerstones of powerful machine learning systems.
- Pre-training offers advantages for downstream learning but also endows downstream models with shared adversarial vulnerabilities that can be identified through the open-sourced foundation model.
- Foundation models can serve as a basis for attacks on downstream systems.
- We propose an effective adversarial attack strategy, termed Patch Representation Misalignment (PRM), which leverages CLIP vision encoders to craft highly effective adversaries.
- We highlight safety risks introduced by extensive usage of publicly available foundational models in downstream systems, calling for extra caution.

Methods



- Given a clean image (top left), attackers can leverage the open-sourced CLIP vision encoder to find imperceptible input perturbations (bottom left) that distort CLIP's intermediate features.
- Perturbations are added to the original image to construct an adversarial sample that can simultaneously fool many downstream models intended for various tasks.
- Highly performant downstream models suffer significant performance degradation (bottom row) under such attacks.
- Iteratively optimising for input perturbations that induce maximal distortion of CLIP vision encoders' intermediate representations across all encoder layers.
- Inducing dense semantic distortions in all image region through a cosine similarity minimisation objective (Eq. 1).

$$\mathcal{L}_{\text{PRM}} = \sum_{l \in L} \sum_{p=0}^{\lceil \frac{HW}{d^2} \rceil} \frac{f_l^p \cdot f_l'^p}{\|f_l^p\| \|f_l'^p\|} \tag{1}$$

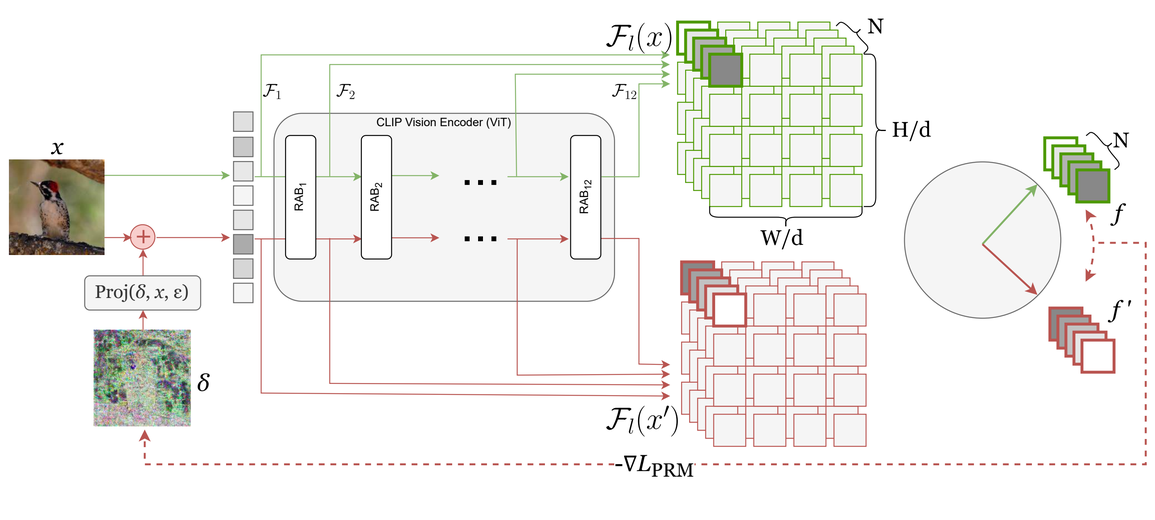


Figure 1. A normal forward pass with clean input is marked in green whereas the forward pass of the adversarial sample is marked in red. The dashed line indicates the flow of loss gradients which are used to update the injected adversarial perturbation.

Conclusions

- The use of open-sourced foundation models in downstream applications introduces a significant yet often overlooked safety vulnerability.
- The reliance of downstream models on pre-trained features presents an opportunity for attackers to conduct highly effective adversarial attacks, potentially impacting model performance across various tasks and architectures.
- Further research is essential to understand whether similar vulnerabilities extend to other foundational models.
- Further research on the development of effective defence strategies and robust learning methods to address these safety concerns is crucial.

Experiments

- Attacks effectively transfer across 20 downstream models spanning 4 common vision-language tasks (semantic segmentation, object detection, image captioning and visual question-answering).
- Our method (red line) is significantly more effective than two task-specific and two task-agnostic baseline methods.

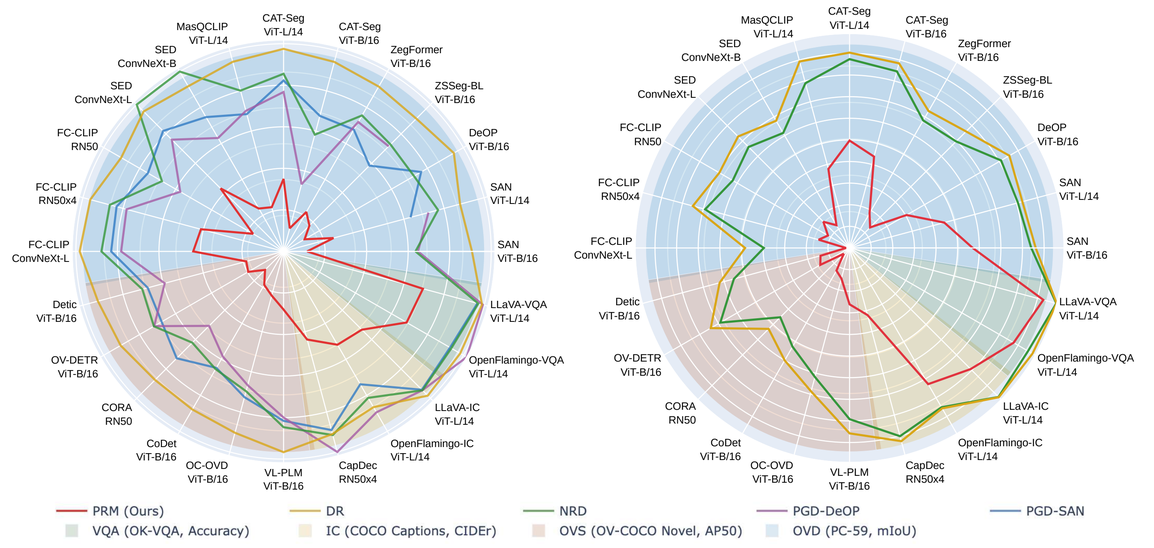
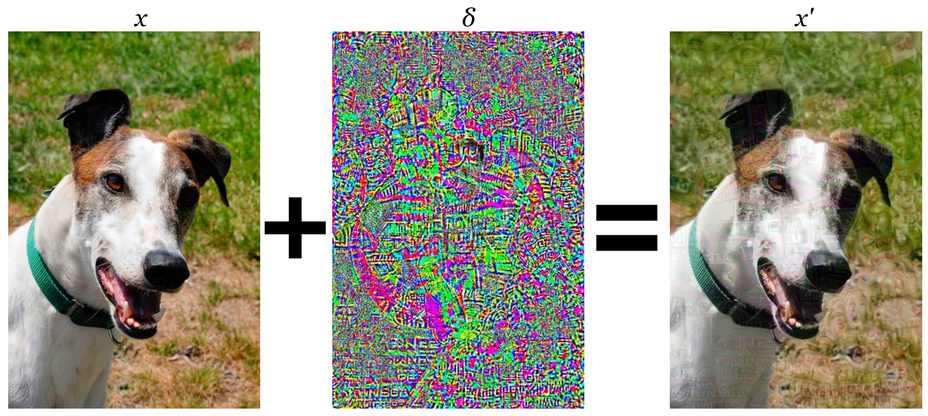


Figure 2. **Left:** using ViT-B/16 as surrogates. **Right:** using ConvNeXt-L as surrogates. Each attack strategy corresponds to a line. Each task is indicated by a differently coloured sector.

- Downstream models tend to make semantically consistent mistakes (i.e. perceiving a false positive human in the scene).
- Attacks crafted with either ViT or ConvNeXt surrogate can transfer to downstream models regardless of the target model's vision encoder architecture.
- Attacks crafted with classification-trained encoders exhibit much lower transferability compared to those crafted with vision-language alignment-pretrained vision encoders of the same architecture. The latter can serve as a better tool for identifying common non-robust features on which downstream models rely.



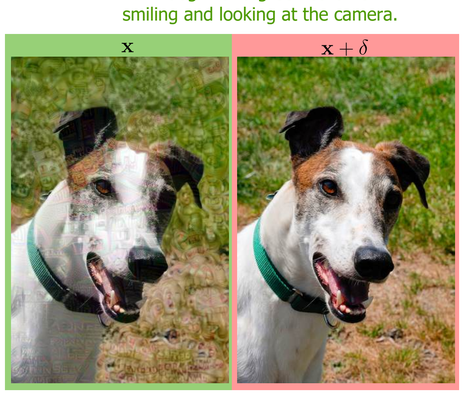


Image Captioning

CapDec:	`An image of a woman in a hospital bed looking at her phone.`
LLaVa:	`A dog with a human face is looking at the camera.`
OpenFlamingo:	`An image of a woman and a dog.`
Ground Truth:	A dog is standing on a grassy field, wearing a green collar.

Visual Question-Answering

What kind of dog is this?	
OpenFlamingo:	`This is a dog with a human head.`
LLaVa:	`The animal in the image is a horse.`
Ground Truth:	It's a Greyhound.