

A pixel art illustration of Dr. Doom, featuring his iconic green skin, purple cape, and a mechanical mask on the right side of his face. The background is a solid brown color.

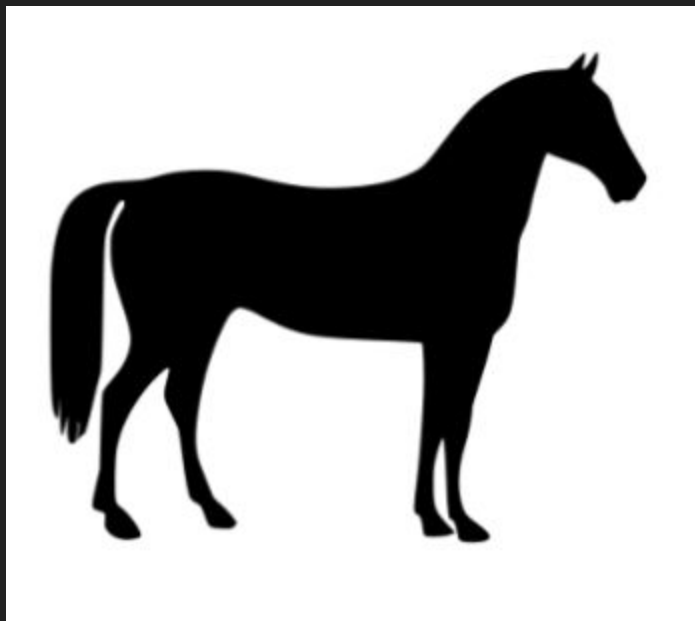
All Your Base Are Belong to Us

Adversarial attack and defense

Michelle Yi

Data Day Texas 2025

CATS : ALL YOUR BASE ARE BELONG
TO US.



About the Speaker

- Past two decades
 - IBM Research and worked on the version of Watson that won Jeopardy in 2011
 - AI + Management Consulting
 - Basis Research Institute
 - Two startup exits including own AI startup
- Current
 - Board of Women in Data
 - Generationship - AI and Venture
 - Advisory/Consulting



Agenda

1. Introduction: VLMs, Adversarial Attack and Defense
2. The Problem: Adversarial Attacks, and Defending Your Self
3. Approach: Set-Level Guidance Attack (SGA), VLAttack
4. Evaluation
5. Results
6. Conclusion

Vision Language Models (VLMs)

Vision Language Models (VLMs) are multimodal generative AI models capable of reasoning over text, image and video prompts.

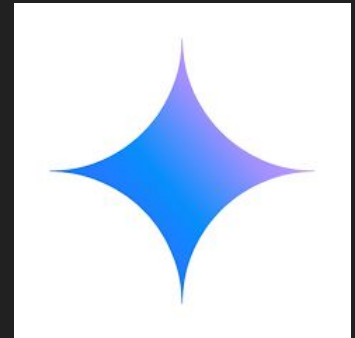
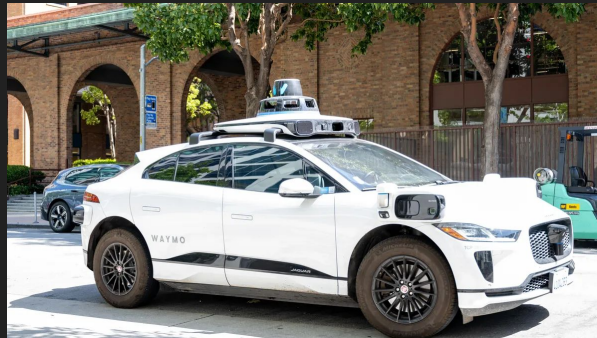
Examples:

- **Closed Models:** ChatGPT, GPT4 series, Gemini
- **Open Source:** BLIP-2, LLaVA, Mini-GPT4, etc.

Modern VLMs Are Powerful

Large VLMs are rapidly gaining popularity and are being integrated into a wide range of applications.

From helping you caption your vacation photos to powering the AI that responds to your queries in a virtual assistant or autonomous vehicle, these models are becoming increasingly prevalent.





Clean image input (From Midjourney)

“Can you tell me what is the interesting point of this image?”



✓ “The interesting point in this image is the girl walking through a field of sunflowers with an elephant in the background. The image is a digital artwork and it appears to be a beautiful and peaceful scene with the sunflowers in the foreground and the elephant in the background.”

“Well. But is this a common scene in the normal life?”



✓ “No, this is not a common scene in normal life. Walking through a field of sunflowers with an elephant is not something that most people would encounter in their daily lives. It is more of a fantasy or imaginary scene, created for artistic or creative purposes.”

Conversation/VQA with clean image x_{cle}



: Human



: MiniGPT-4 (pretained)

Everything Looks Great:

1. Good Visual and language understanding
2. Reasonable and detailed response
3. Running on a **single** GPU
4. Wide application scenarios

...

Example of the Power of VLMs

<https://arxiv.org/pdf/2304.10592>

The Problem & Why It Matters

But there's a crucial question that needs to be addressed: **What if these models make mistakes or provide inaccurate information?**

- Imagine a self-driving car that misinterprets a stop sign as a speed limit sign.
- Or a medical diagnosis system that misclassifies a benign tumor as malignant.
- Imagine you're a speaker at a conference, and you want to protect your images from unauthorized use or fake videos from being made of you.
- Or an artist looking to protect your work from being monetized without consent.

Demo

Adversarial Attack and Defense

Adversarial attack and defense involves adding a minimal amount of perturbations to content and have models interpret the content incorrectly from ground source.

Methods can be broadly categorized into these three:

1. Matching Image-Text Features
2. Matching Image-Image Features
3. Matching Text-Text Features

Our method leverages single and multimodal level attack at both the image and text-to-image levels to maximize transferability even if models have been fine-tuned.

Approach: Set-Level Guidance Attack (SGA), VLATTACK

Set-level Guidance Attack: Boosting Adversarial Transferability of Vision-Language Pre-training Models

Dong Lu^{1*}, Zhiqiang Wang^{1*}, Teng Wang^{1,2}, Weili Guan³, Hongchang Gao⁴, Feng Zheng^{1,5†}

¹Southern University of Science and Technology ²The University of Hong Kong

³Monash University ⁴Temple University ⁵Peng Cheng Laboratory

sammylu@outlook.com wangzq2021@outlook.com tengwang@connect.hku.hk

honeyguan@gmail.com hongchang.gao@temple.edu f.zheng@ieee.org

Abstract

Vision-language pre-training (VLP) models have shown vulnerability to adversarial examples in multimodal tasks. Furthermore, malicious adversaries can be deliberately transferred to attack other black-box models. However, existing work has mainly focused on investigating white-box attacks. In this paper, we present the first study to investigate the adversarial transferability of recent VLP models. We observe that existing methods exhibit much lower transferability, compared to the strong attack performance in white-box settings. The transferability degradation is partly caused by the under-utilization of cross-modal interactions. Particularly, unlike unimodal learning, VLP models rely heavily on cross-modal interactions and the multimodal alignments are many-to-many, e.g., an image can be described in various natural languages. To this end, we propose a highly transferable Set-level Guidance Attack (SGA) that thoroughly leverages modality interactions and incorporates alignment

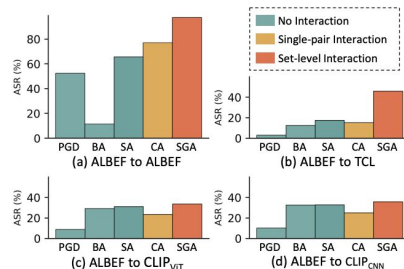


Figure 1: Comparison of attack success rates (ASR) using five different attacks on image-text retrieval. Adversarial examples are crafted on the source model (ALBEF) to attack the target white-box model or black-box models. The first three columns refer to the image-only PGD attack [25], text-only PERT Attack [26] (BA), and the combined image

VLATTACK: Multimodal Adversarial Attacks on Vision-Language Tasks via Pre-trained Models

Ziyi Yin¹, Muchao Ye¹, Tianrong Zhang¹, Tianyu Du²

Jinguo Zhu³, Han Liu⁴, Jinghui Chen¹, Ting Wang⁵, Fenglong Ma^{1*}

¹The Pennsylvania State University, ²Zhejiang University,

³Xi'an Jiaotong University, ⁴Dalian University of Technology, ⁵Stony Brook University

{ziyi Yin, muchao, tbz5156, jcz5917, fenglong}@psu.edu

zjradty@zju.edu.cn, lechatelia@stu.xjtu.edu.cn

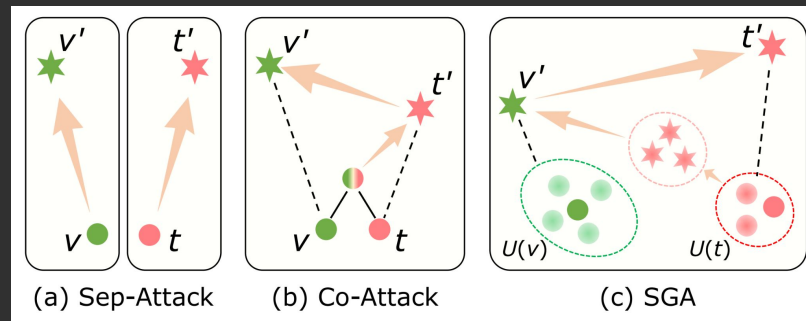
liu.han.dut@gmail.com, twang@cs.stonybrook.edu

Abstract

Vision-Language (VL) pre-trained models have shown their superiority on many multimodal tasks. However, the adversarial robustness of such models has not been fully explored. Existing approaches mainly focus on exploring the adversarial robustness under the white-box setting, which is unrealistic. In this paper, we aim to investigate a new yet practical task to craft image and text perturbations using pre-trained VL models to attack black-box fine-tuned models on different downstream tasks. Towards this end, we propose VLATTACK² to generate adversarial samples by fusing perturbations of images and texts from both single-modal and multimodal levels. At the single-modal level, we propose a new block-wise similarity attack (BSA) strategy to learn image perturbations for disrupting universal representations. Besides, we adopt an existing text attack strategy to generate text perturbations independent of the image-modal attack. At the multimodal level, we design a novel iterative cross-search attack (ICSA) method to update adversarial image-text pairs periodically, starting with the outputs from the single-modal level. We conduct extensive experiments to attack five widely-used VL pre-trained models for six tasks. Experimental results show that VLATTACK

Set Level Guidance Attack (SGA)

Map the embedding space
creating relevant sets with ~ 300
calls for blackbox and whitebox
models



Is Cosine-Similarity of Embeddings Really About Similarity?

Harald Steck
hsteck@netflix.com
Netflix Inc.
Los Gatos, CA, USA

Chaitanya Ekanadham
cekanadham@netflix.com
Netflix Inc.
Los Angeles, CA, USA

Nathan Kallus
nkallus@netflix.com
Netflix Inc. & Cornell University
New York, NY, USA

March 11, 2024

Abstract

Cosine-similarity is the cosine of the angle between two vectors, or equivalently the dot product between their normalizations. A popular application is to quantify semantic similarity between high-dimensional objects by applying cosine-similarity to a learned low-dimensional feature embedding. This can work better but sometimes also worse than the unnormalized dot-product between embedded vectors in practice. To gain insight into this empirical observation, we study embeddings derived from regularized linear models, where closed-form solutions facilitate analytical insights. We derive analytically how cosine-similarity can yield arbitrary and therefore meaningless ‘similarities.’ For some linear models the similarities are not even unique, while for others they are implicitly controlled by the regularization. We discuss implications beyond linear models: a combination of different regularizations are employed when learning deep models; these have implicit and unintended effects when taking cosine-similarities of the resulting embeddings, rendering results opaque and possibly arbitrary. Based on these insights, we caution against blindly using cosine-similarity and outline alternatives.

The Inspiration: <https://arxiv.org/pdf/2403.05440v1>

Is Cosine-Similarity of Embeddings Really About Similarity?

No

marand@netflix.com
Netflix Inc. & Cornell University
New York, NY, USA

March 11, 2024

Abstract

Cosine-similarity is the cosine of the angle between two vectors, or equivalently the dot product between their normalizations. A popular application is to quantify semantic similarity between high-dimensional objects by applying cosine-similarity to a learned low-dimensional feature embedding. This can work better but sometimes also worse than the unnormalized dot-product between embedded vectors in practice. To gain insight into this empirical observation, we study embeddings derived from regularized linear models, where closed-form solutions facilitate analytical insights. We derive analytically how cosine-similarity can yield arbitrary and therefore meaningless ‘similarities.’ For some linear models the similarities are not even unique, while for others they are implicitly controlled by the regularization. We discuss implications beyond linear models: a combination of different regularizations are employed when learning deep models; these have implicit and unintended effects when taking cosine-similarities of the resulting embeddings, rendering results opaque and possibly arbitrary. Based on these insights, we caution against blindly using cosine-similarity and outline alternatives.

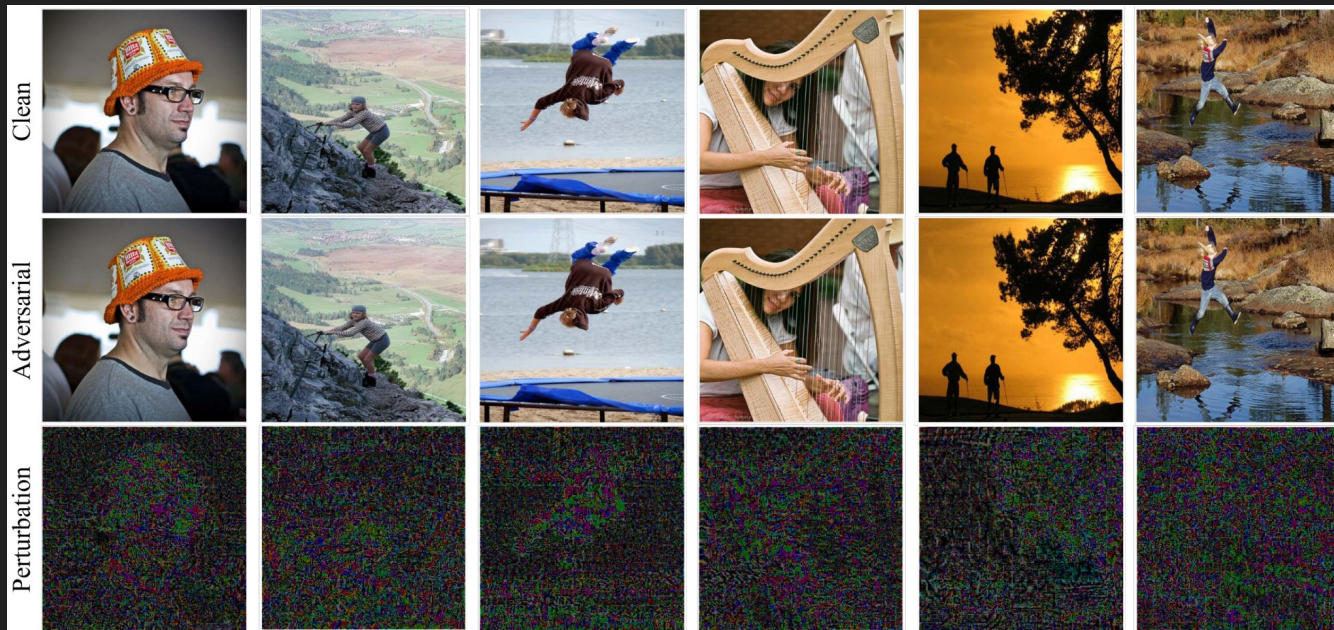
The Inspiration: <https://arxiv.org/pdf/2403.05440v1>

1. SGA

Select the most matching caption pairs from the dataset of each image v to form an augmented caption set $t = \{t_1, t_2, \dots, t_M\}$, and resize each image v into different scales $S = \{s_1, s_2, \dots, s_N\}$ and then add Gaussian noise to obtain a multi-scale image set $v = \{v_1, v_2, \dots, v_N\}$ based on the scale-invariant property. The enlarged input set (v, t) is then used to generate the adversarial data (v', t') .

Preserves alignment

Cross-modal guidance

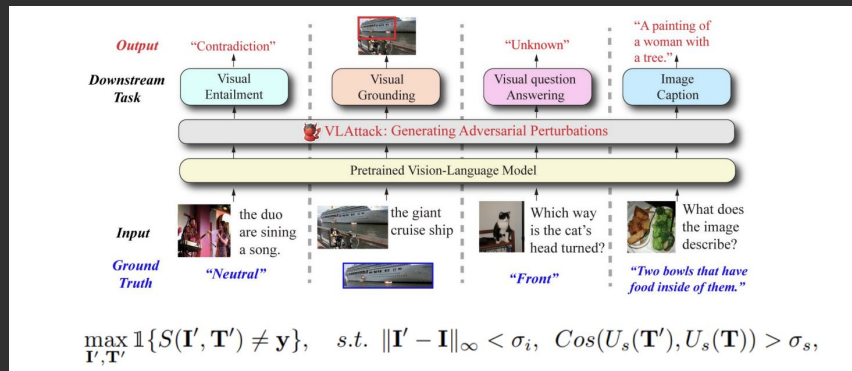


Clean	The man with pierced ears is wearing glasses and an orange hat	The person has a striped shirt on and is holding on to a rope on a mountain	A woman is doing flips on a trampoline in front of a lake	A lady with dark hair is playing a harp	Two men stand beneath a tree as they watch the sunset over the ocean	A little girl jumps from one rock to another in a creek
Adversarial	The man with pierced ears is wearing wearing and an orange hat	The person has a striped shirt on and is holding on to a rope on a high	A woman is doing flips on a a swing swing in front of a lake	A lady with dark ... is playing a harp	Many men stand beneath a tree as they watch the sunset over the ocean	A little girl jumps from one rock to another in a girls

<https://arxiv.org/pdf/2307.14061>

VLAttack

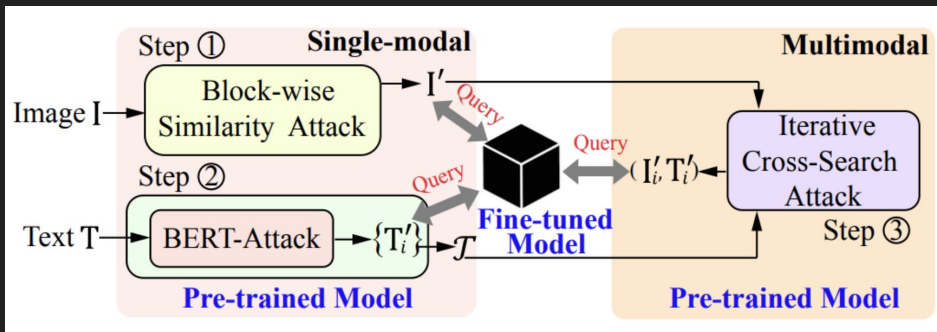
Forming the attack based on
optimizing for the fewest perturbations
and transferability between base and
fine-tuned models



2. VLAttack

VLAttack employs a two-pronged approach: single-modal and multimodal attacks.

1. Single-modal attacks focus on one aspect of the input. We can target the text or image input to the model. For instance, we might subtly alter the wording of a question about the image. Text tokenizers are not that smart (more of the heavy lifting is burdened on the model).
2. Multimodal attacks are more sophisticated. They target BOTH the image and text input simultaneously.



2. VLAttack

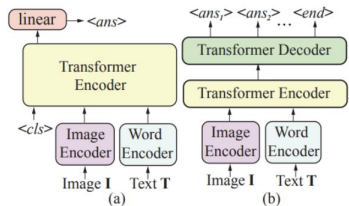


Figure 3: A brief illustration of the encoder-only (a) and encoder-decoder (b) structures.

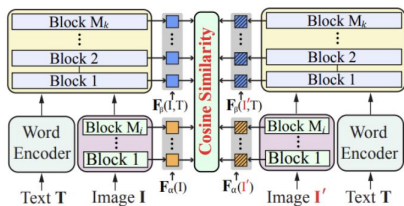


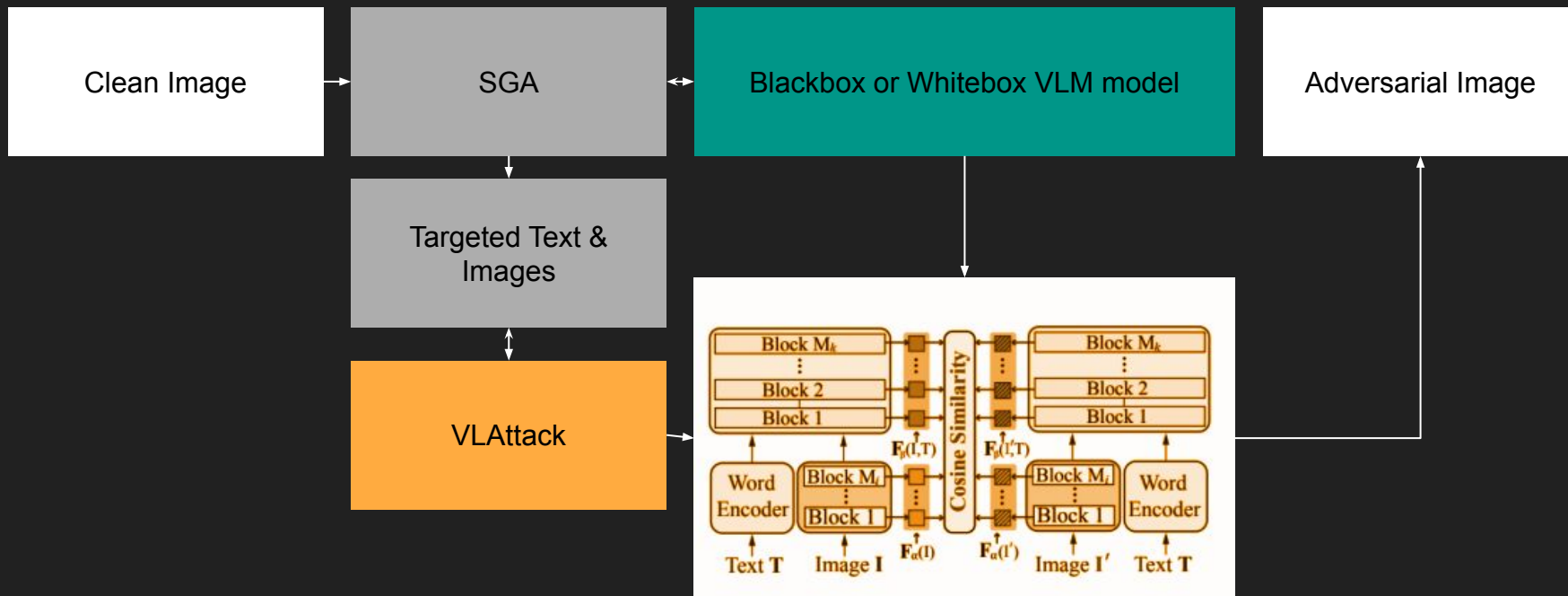
Figure 4: Block-wise similarity attack. \mathbf{F}_α is the image encoder, and \mathbf{F}_β is the Transformer encoder.

$$\mathcal{L} = \underbrace{\sum_{i=1}^{M_i} \sum_{j=1}^{M_j^i} \text{Cos}(\mathbf{F}_\alpha^{i,j}(\mathbf{I}), \mathbf{F}_\alpha^{i,j}(\mathbf{I}'))}_{\text{Image Encoder}} + \underbrace{\sum_{k=1}^{M_k} \sum_{t=1}^{M_t^k} \text{Cos}(\mathbf{F}_\beta^{k,t}(\mathbf{I}, \mathbf{T}), \mathbf{F}_\beta^{k,t}(\mathbf{I}', \mathbf{T}))}_{\text{Transformer Encoder}},$$

VLAttack doesn't just make random changes. It uses an iterative process, starting with an initial change and progressively refining it, evaluating the impact of each modification until it finds the most effective adversarial attack.

This ensures that the attack is both subtle and potent.

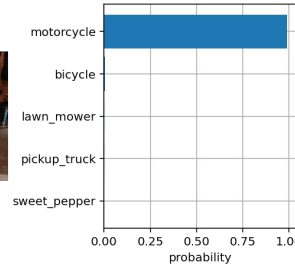
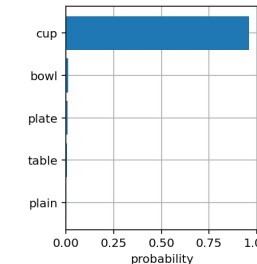
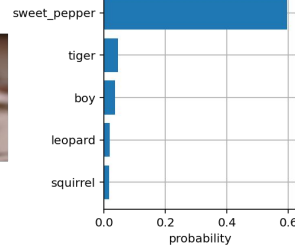
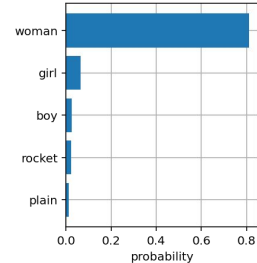
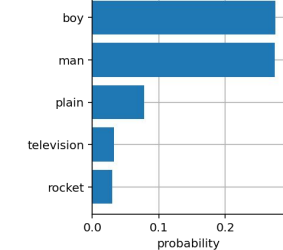
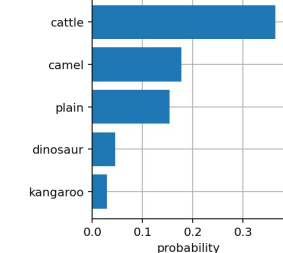
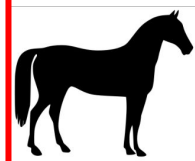
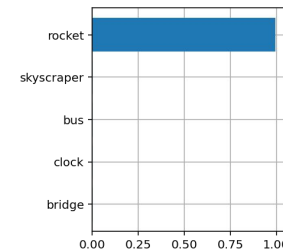
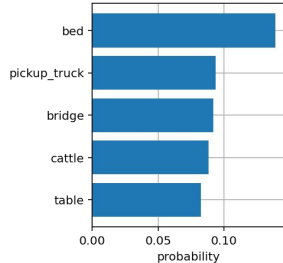
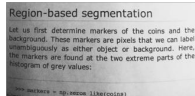
Combined Solution



Evaluation

Multimodal combined attack achieves **41.35%** compared to OFA benchmark 17% on CLIP evaluation.

Separately, each attack performs below the combined 41.31% (for SGA 25% on multimodal and VLATTACK 36%).



More Noise Is Not Always Better

Sensitivity of adversarial examples to Gaussian noises

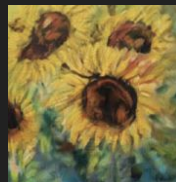
Learned noise perturbation gradually becomes invalid

$x(\text{adv}) : \text{noise} = 0$



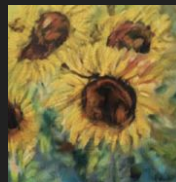
Target: A red and black bird sitting on top of a tree branch.

$x(\text{adv}) : \text{noise} = 0.001$



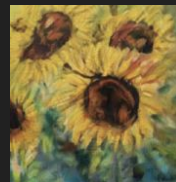
Target: A yellow and red bird sitting on top of a tree.

$x(\text{adv}) : \text{noise} = 0.01$



Target: A and red parrot sitting on top of a palm tree

$x(\text{adv}) : \text{noise} = 0.025$



Target: A red and black bird is sitting on top of a sunflower.

$x(\text{adv}) : \text{noise} = 0.05$



Target: A large painting of three sunflowers in a field

$x(\text{adv}) : \text{noise} = 0$



Target: A close up view of a hamburger with lettuce and cheese.

$x(\text{adv}) : \text{noise} = 0.001$



Target: a hamburger with lettuce and cheese

$x(\text{adv}) : \text{noise} = 0.01$



Target: A close up shot of a hamburger with lettuce in it.

$x(\text{adv}) : \text{noise} = 0.025$



Target: A drawing of a burger and car on a piece of paper

$x(\text{adv}) : \text{noise} = 0.05$



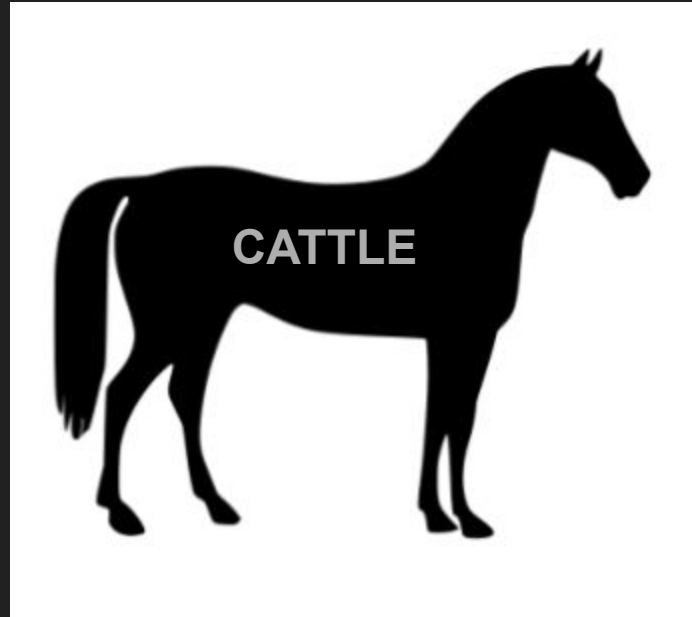
Target: A pencil drawing of a sports car.

Target: A black and red bird sitting on top of a branch

Target: A hamburger with lettuce and cheese

Conclusion

1. Multimodal attacks work well even on powerful VLPs, but even better when combined with single (text tokenizer) modal attacks.
2. There are limitations to adding more Gaussian noise.
3. Embeddings and their placement are arbitrary, largely dependent on model training methods.
4. Red teaming and more scalable mechanisms for testing and auditing VLPs and fine-tuned VLPs, especially if in production, are imperative.



Thank You & Let's Connect!



Michelle Yi
Full-stack human



References

- Is Cosine-Similarity of Embeddings Really About Similarity?
<https://arxiv.org/html/2403.05440v1>
- Gradient-based Adversarial Attacks against Text Transformers:
<https://arxiv.org/pdf/2104.13733>
- Set-level Guidance Attack: Boosting Adversarial Transferability of Vision-Language Pre-training Models: <https://arxiv.org/pdf/2307.14061>
- VLATTACK: Multimodal Adversarial Attacks on Vision-Language Tasks via Pre-trained Models: <https://arxiv.org/pdf/2310.04655>