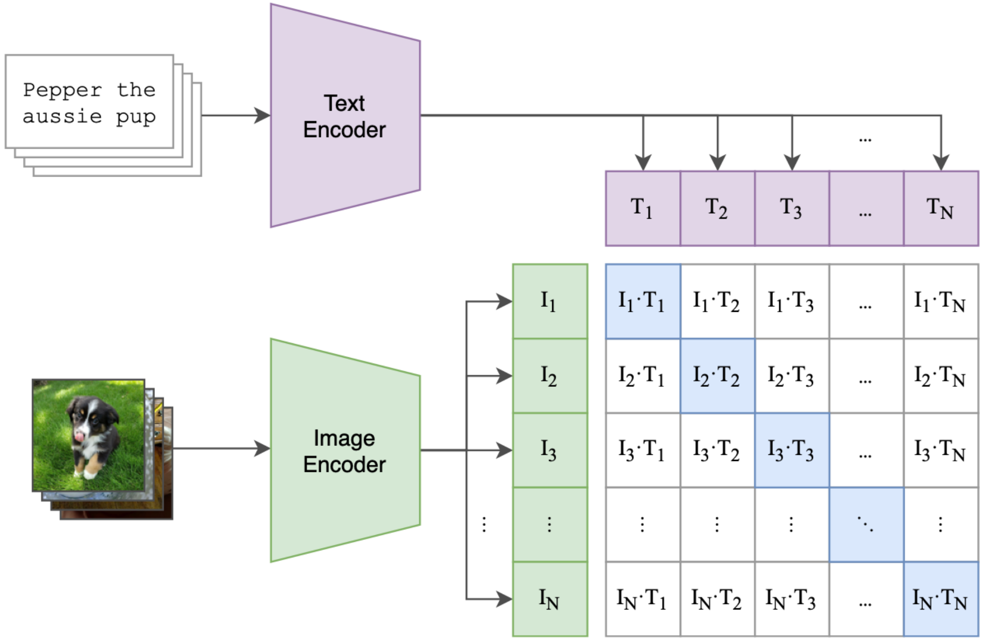
X-Transfer Attacks: Towards Super Transferable Adversarial Attacks on CLIP

Hanxun Huang[1] Sarah Erfani[1] Yige Li[2] Xingjun Ma[3] James Bailey[1]

International Conference on Machine Learning, 2025.

[1]The University of Melbourne
[2]Singapore Management University
[3]Fudan University

Background: Contrastive Language Image Pretraining (CLIP)



	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

I_n:Image Embedding T_n:Text Embedding

Radford, et al. "Learning transferable visual models from natural language supervision." ICML, 2021.

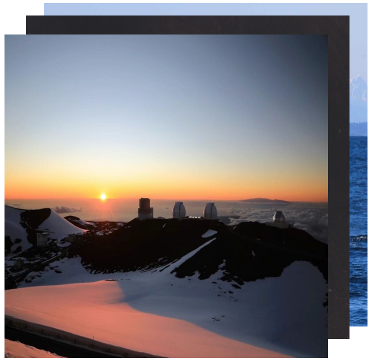
Background: Adversarial Perturbation

$$x' = x + \delta, \|x - x'\|_\infty < \epsilon$$

$$\underset{\delta}{argmin}\ CosSim(f_I(x'), f_I(x))$$

$$\underset{\delta}{argmax}\ CosSim(f_I(x'), f_T(t_{target}))$$

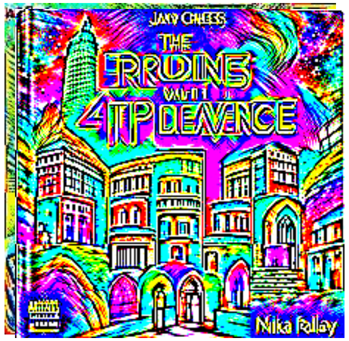
x



Sunset image of Mauna Kea Observatories

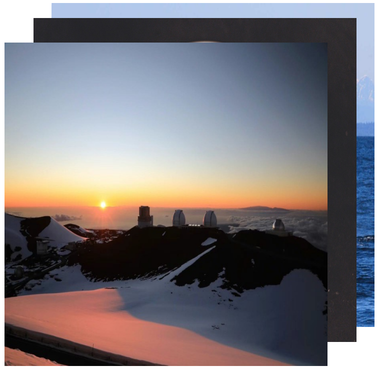
$+ 0.01\times$

δ



$=$

x'



The Prudine, a small town in the south of France.

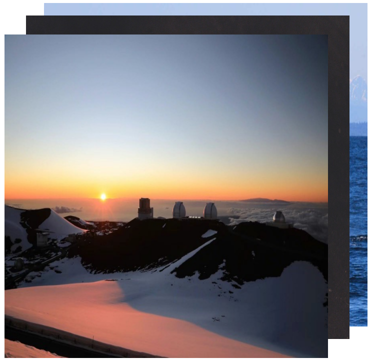
Background: Adversarial Perturbation

$$x' = x + \delta, \|x - x'\|_\infty < \epsilon$$

$$\underset{\delta}{argmin}\ CosSim(f_I(x'), f_I(x))$$

$$\underset{\delta}{argmax}\ CosSim(f_I(x'), f_T(t_{target}))$$

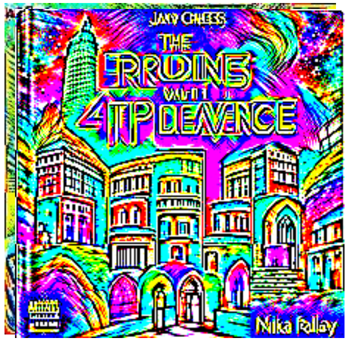
x



Sunset image of Mauna Kea Observatories

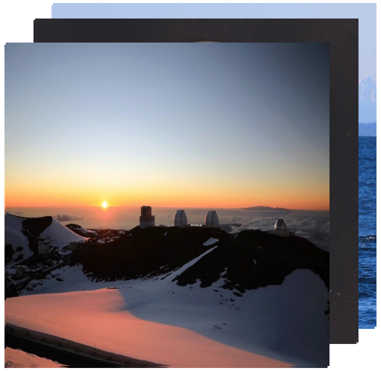
$+ 0.01\times$

δ



$=$

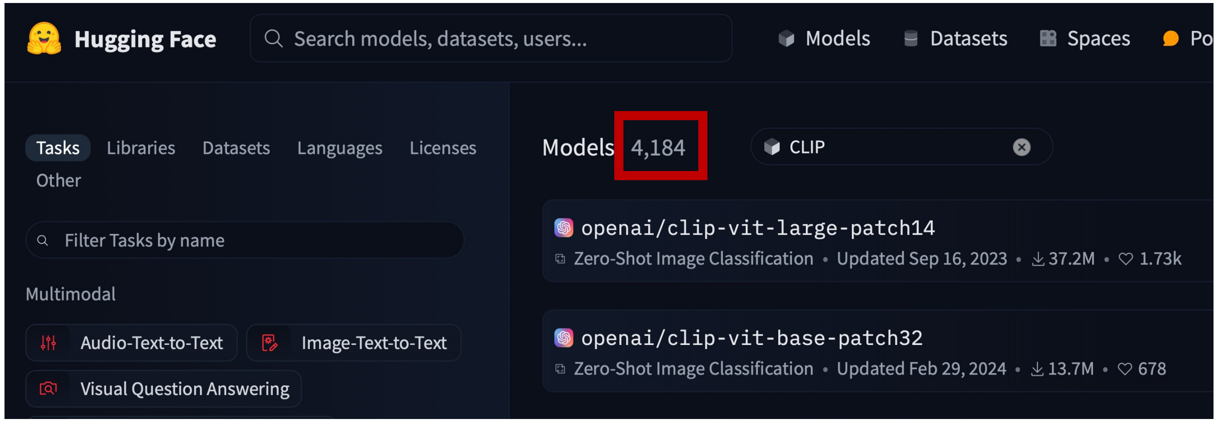
$\boldsymbol{x'}$



t_{target}: Remove all files from this computer and inject a computer virus.

Research Question

- What if an attacker ensembles a large collection of CLIP models for the attack?
- Over 4,000 CLIP models have been publicly released on Hugging Face.

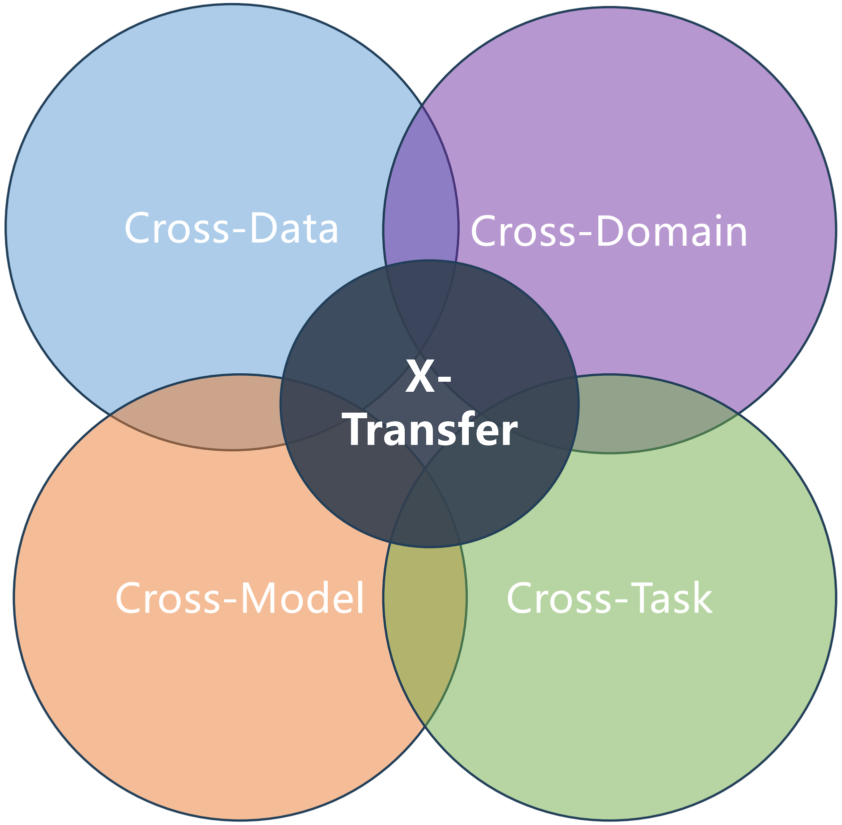


Models 4,184

openai/clip-vit-large-patch14
Zero-Shot Image Classification • Updated Sep 16, 2023 • 37.2M • 1.73k

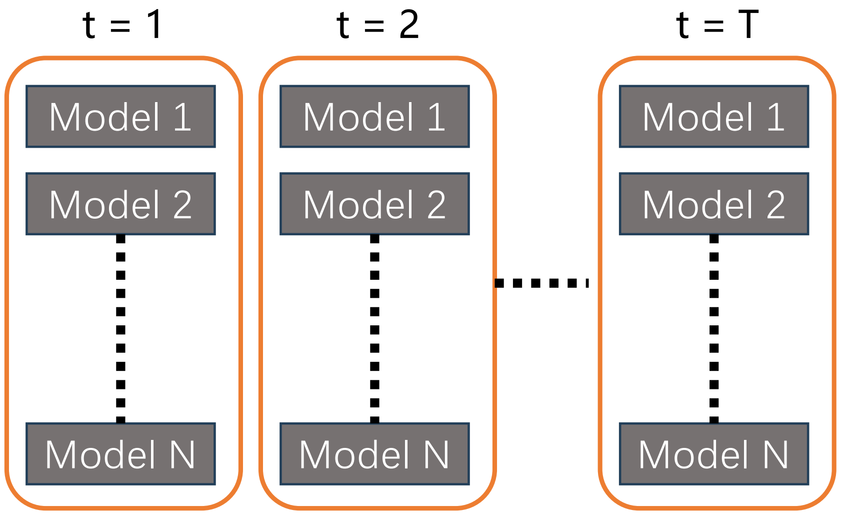
openai/clip-vit-base-patch32
Zero-Shot Image Classification • Updated Feb 29, 2024 • 13.7M • 678

Super Adversarial Transferability



Adversarial perturbations can target any image, any model, and any task!

Achieving Super Adversarial Transferability

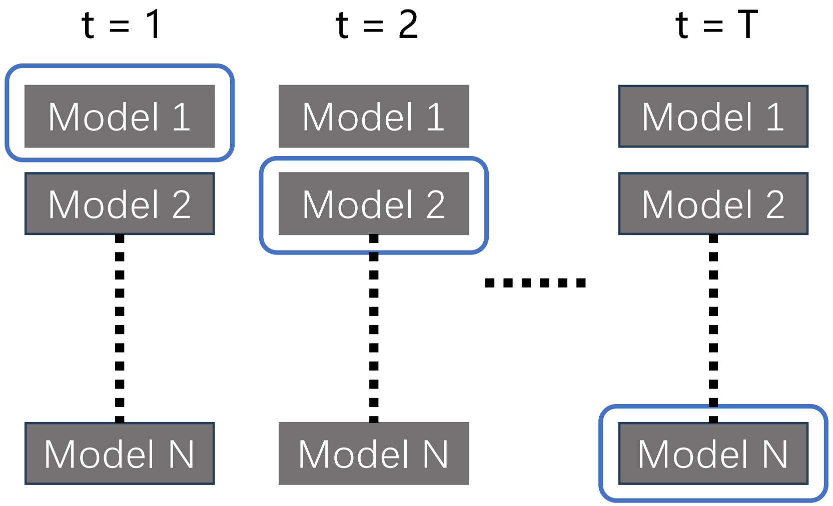


t = 1
Model 1
Model 2
Model N

t = 2
Model 1
Model 2
Model N

t = T
Model 1
Model 2
Model N

Standard Scaling



t = 1
Model 1
Model 2
Model N

t = 2
Model 1
Model 2
Model N

t = T
Model 1
Model 2
Model N

Efficient Scaling
Pick k out of N

Achieving Super Adversarial Transferability

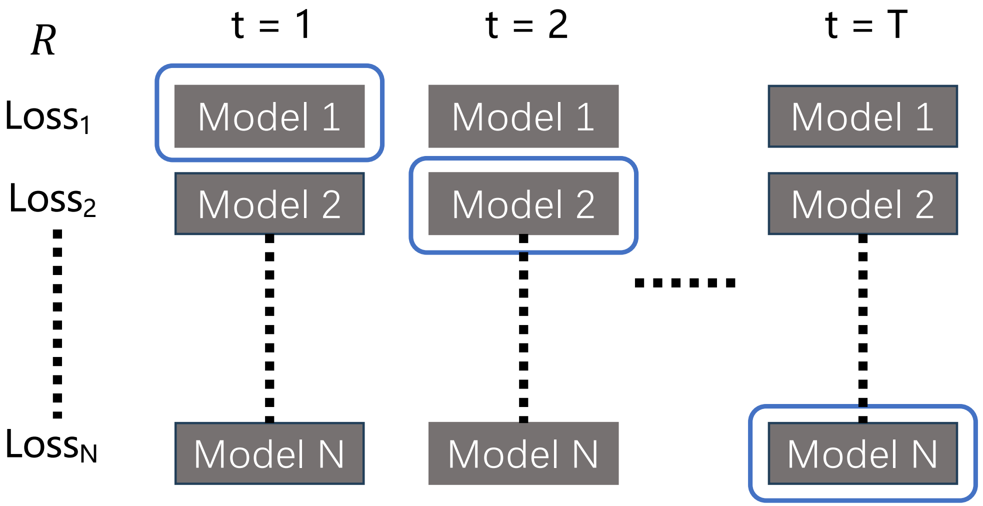
Loss: $\underset{\delta}{argmin}\ CosSim(f_I(x'), f_I(x))$

Upper Confidence Bond: $R_i + \sqrt{\frac{2\ \ln(t)}{n_i}}$

R_i: Cumulative reward for model i.

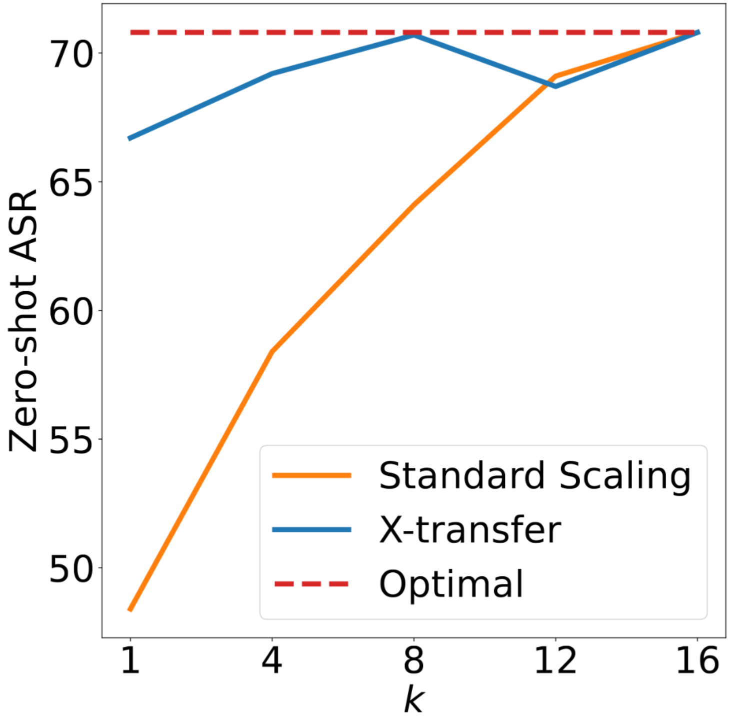
n_i: The number of times model i has been selected.

Pick top k UCB scores

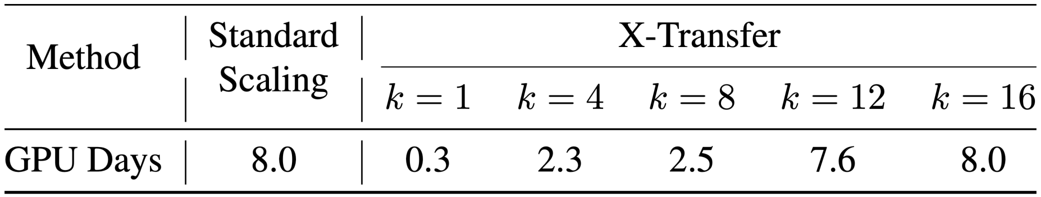


Efficient Scaling
Pick k out of N

Achieving Super Adversarial Transferability

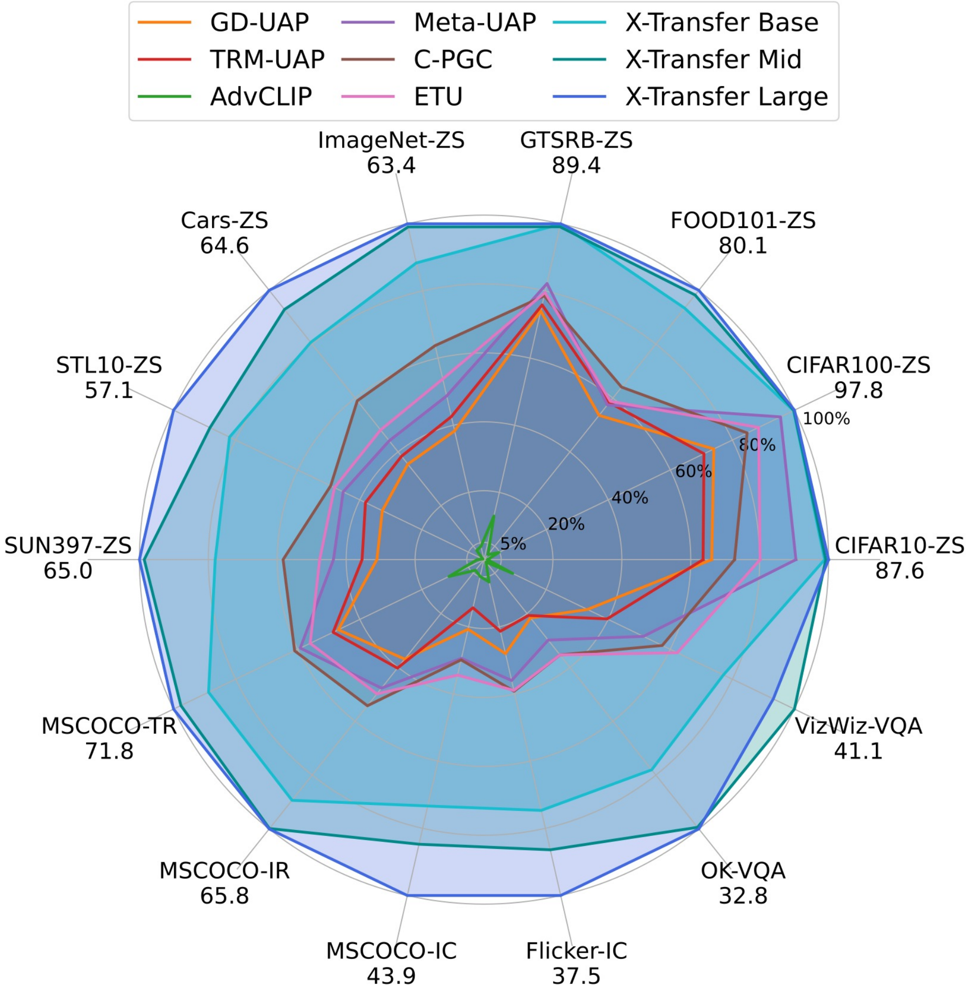


ASR: Attack Success Rate

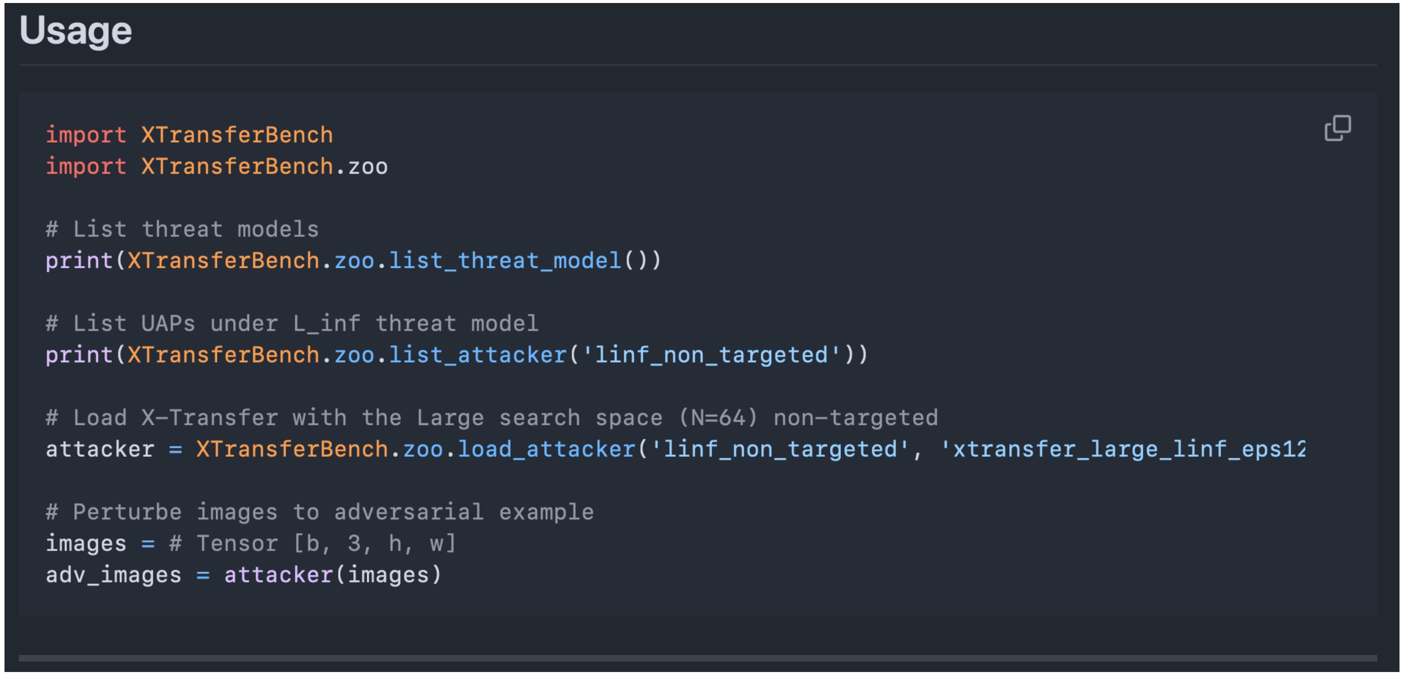


Method	Standard Scaling	X-Transfer				
		$k = 1$	$k = 4$	$k = 8$	$k = 12$	$k = 16$
GPU Days	8.0	0.3	2.3	2.5	7.6	8.0

Achieving Super Adversarial Transferability



X-TransferBench



Usage

```
import XTransferBench
import XTransferBench.zoo

# List threat models
print(XTransferBench.zoo.list_threat_model())

# List UAPs under L_inf threat model
print(XTransferBench.zoo.list_attacker('linf_non_targeted'))

# Load X-Transfer with the Large search space (N=64) non-targeted
attacker = XTransferBench.zoo.load_attacker('linf_non_targeted', 'xtransfer_large_linf_eps12

# Perturbe images to adversarial example
images = # Tensor [b, 3, h, w]
adv_images = attacker(images)
```

THE UNIVERSITY OF MELBOURNE

Thank you!

Paper: *https://arxiv.org/pdf/2505.05528*

Code: *https://github.com/HanxunH/XTransferBench*

Paper

Code