Adversarial transferability in foundation models
DL Apps - Project Work

Gregorio Piqué
gregorio.pique@edu.unifi.it

1. Introduction

The project focuses on evaluating the transferability of adversarial attacks on multi-modal foundation models, with a zero-shot image classification task.

2. Models

The multi-modal foundation models tested so far are:

- **CLIP ResNet-50** [3]
- **CLIP ViT-B/16** [3]
- **Perception Encoder Core-B16-224** [1]: a SOTA CLIP-based model for zero-shot image and video classification

3. Attacks

The models are tested with different implemented attacks. These include:

- **Fast Gradient Sign Method** (FGSM): single perturbation to maximize the model's errors
- **Iterative FGSM** (I-FGSM): iterative variant of FGSM
- **Momentum Iterative FGSM** (MI-FGSM): adds a momentum term to stabilize the optimization process
- **Translation-Invariant FGSM** (TI-MI-FGSM) [2] : attempts to generate more transferable adversarial examples by convolving the gradient with a pre-defined kernel in each attack iteration
- **PGD**: applies iterative perturbations starting from a random one, projecting onto a sphere of fixed size
- **Embedding Space Disruption** [5]: minimizes the alignment between adversarial and clean image embeddings
- **UAP**: generates a single image-agnostic perturbation, maximizing the model's error on multiple images at once

The attacks have been implemented from scratch. All methods compute gradients using the cross-entropy loss, except for the Embedding Space Disruption attack, which instead minimizes the alignment between the embeddings of clean and perturbed images using the loss $\mathcal{L} = ||E^{gt} - E^{adv}||_2$. The source code is publicly available at [4].

4. Dataset

The tests are performed on the *zh-plus/tiny-imagenet* dataset from HuggingFace [6].

5. Experiments

The experiments were conducted in a transfer-based adversarial attack setting. A *surrogate* model (source) was attacked using the specified methods, generating adversarial perturbations for individual images - or for a set of images in the case of Universal Adversarial Perturbation (UAP) attacks. The resulting perturbations were applied to the input images, which were then evaluated on a separate *target* model to assess transferability.

Settings All evaluation metrics were computed on a test set of fixed size across all attacks and models to ensure a fair comparison. A constant attack budget of $\epsilon = 4/255 \approx 0.01569$ was used for all methods.

Iterative FGSM variants were run for $N_{\text{iter}} = 10$ steps, with each step size set to $\alpha = \epsilon/N_{\text{iter}}$. The translation-invariant attack used a kernel of size 5 and was combined with momentum, forming the ***TI-MI-FGSM*** variant.

The PGD attack applied uniform random noise for initialization and ran for $N_{\text{iter}} = 20$ iterations.

The *Embedding Space Disruption* attack targeted the final-layer embeddings, using the AdamW optimizer with a learning rate of 0.01 for 250 iterations.

Universal Adversarial Perturbations (UAPs) were computed over the full validation set of the *tiny-imagenet* dataset (10**k** images) using the Adam optimizer with a learning rate of 0.001.

6. Results

The experimental results are presented in the following tables. The effectiveness of the attacks is quantified using the Attack Success Ratio (ASR).

Tables 1–7 present detailed results for each attack method, covering both white-box and black-box scenarios. Values in *italic* denote white-box attacks, where the source and target models are identical. The *italicized averages* combine both white-box and transfer-based ASR results. The last row of these results table can be interpreted as measuring the average robustness of each model when targeted by attacks - where lower values indicate stronger robustness. Conversely, the last column reflects the average transferability of the perturbations generated by each model when used as the source - with higher values indicating more transferable adversarial examples.

Table 8 reports the average black-box results for each attack across all models, providing insight into the general transferability of the perturbations generated from a specific model. The last row of these results table can be interpreted as measuring the average effectiveness of each attack, while the last column represents the average transferability of perturbations generated with the considered model.

Follow the three key findings that emerge from the analysis.

White-box vs. Black-box White-box attacks consistently achieve higher ASR compared to black-box transfer-based attacks, as expected due to the inherent advantage of full model knowledge.

Model Robustness vs. Transferability The results achieved by the models vary significantly:

- The CLIP-RN50 model demonstrates the lowest robustness, exhibiting both the highest average ASR as a target (first column, last row) and the lowest transferability of perturbations as a source (last column, first row).

- PE-Core is the most robust as a target and produces highly transferable perturbations. While its perturbations are generally less transferable than those of CLIP-ViT, it outperforms all others in the UAP attack.

- CLIP-ViT falls between the two: it is more robust than CLIP-RN50 but less so than PE-Core, and it consistently generates the most transferable perturbations across attack types.

Single step vs. Iterative attacks Iterative attacks outperform the basic FGSM in white-box settings. They almost always achieve an ASR close to 1.0, whereas FGSM averages around 0.85. However, in black-box (transfer) settings, FGSM can sometimes outperform iterative methods. For example, when using CLIP-ViT as the source and CLIP-RN50 as the target, FGSM achieves an ASR of 0.686, higher than all iterative variants (which vary between 0.490 and 0.667). This may be due to "overfitting" of the perturbation to the source model, which can harm transferability. Among all tested methods, the Translation-Invariant Momentum FGSM (TI-MI-FGSM) consistently achieves the best overall performance, closely followed by the base MI-FGSM, and in some cases even matching FGSM.

	Target			
Source	CLIP-RN50	CLIP-ViT	PE-Core	**Avg**
CLIP-RN50	*0.814*	0.275	0.221	0.248 *0.436*
CLIP-ViT	0.686	*0.892*	0.452	0.569 *0.677*
PE-Core	0.471	0.461	*0.856*	0.466 *0.596*
Avg	0.578 *0.657*	0.368 *0.542*	0.337 *0.510*	

Table 1. FGSM

	Target			
Source	CLIP-RN50	CLIP-ViT	PE-Core	**Avg**
CLIP-RN50	*1.000*	0.157	0.077	0.117 *0.411*
CLIP-ViT	0.490	*1.000*	0.173	0.332 *0.554*
PE-Core	0.343	0.225	*1.000*	0.284 *0.523*
Avg	0.417 *0.611*	0.191 *0.461*	0.125 *0.417*	

Table 2. I-FGSM

	Target			
Source	CLIP-RN50	CLIP-ViT	PE-Core	**Avg**
CLIP-RN50	*1.000*	0.265	0.202	0.233 *0.489*
CLIP-ViT	0.667	*1.000*	0.365	0.516 *0.677*
PE-Core	0.559	0.461	*1.000*	0.510 *0.673*
Avg	0.613 *0.742*	0.363 *0.575*	0.284 *0.522*	

Table 3. MI-FGSM

	Target			
Source	CLIP-RN50	CLIP-ViT	PE-Core	**Avg**
CLIP-RN50	*1.000*	0.275	0.240	0.257 *0.505*
CLIP-ViT	0.647	*1.000*	0.452	0.549 *0.700*
PE-Core	0.520	0.471	*1.000*	0.495 *0.663*
Avg	0.583 *0.722*	0.373 *0.582*	0.346 *0.564*	

Table 4. TI-MI-FGSM

	Target			
Source	CLIP-RN50	CLIP-ViT	PE-Core	**Avg**
CLIP-RN50	*1.000*	0.157	0.125	0.141 *0.427*
CLIP-ViT	0.509	*1.000*	0.207	0.358 *0.572*
PE-Core	0.387	0.274	*1.000*	0.330 *0.553*
Avg	0.448 *0.632*	0.215 *0.477*	0.166 *0.444*	

Table 5. PGD

	Target			
Source	CLIP-RN50	CLIP-ViT	PE-Core	**Avg**
CLIP-RN50	*0.981*	0.179	0.153	0.166 *0.438*
CLIP-ViT	0.480	*1.000*	0.173	0.327 *0.551*
PE-Core	0.431	0.186	*0.942*	0.309 *0.520*
Avg	0.456 *0.631*	0.183 *0.455*	0.163 *0.423*	

Table 6. Embedding Space Disruption

	Target			
Source	CLIP-RN50	CLIP-ViT	PE-Core	**Avg**
CLIP-RN50	*0.953*	0.201	0.148	0.174 *0.434*
CLIP-ViT	0.492	*0.921*	0.156	0.324 *0.523*
PE-Core	0.490	0.317	*0.979*	0.403 *0.595*
Avg	0.491 *0.645*	0.259 *0.480*	0.152 *0.428*	

Table 7. UAP

References

[1] D. Bolya, P.-Y. Huang, P. Sun, J. H. Cho, A. Madotto, C. Wei, T. Ma, J. Zhi, J. Rajasegaran, H. Rasheed, J. Wang, M. Monteiro, H. Xu, S. Dong, N. Ravi, D. Li, P. Dollár, and C. Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv:2504.13181*, 2025. https://github.com/facebookresearch/perception_models.

[2] Y. Dong, T. Pang, H. Su, and J. Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019. https://github.com/dongyp13/Translation-Invariant-Attacks.

[3] openai. openai / CLIP. https://github.com/openai/CLIP.

[4] G. Piqué. pikerozzo / dlapps-fm-adv-transfer. https://github.com/Pikerozzo/dlapps-fm-adv-transfer.

[5] H. P. Silva, F. Becattini, and L. Seidenari. Attacking attention of foundation models disrupts downstream tasks, 2025. https://arxiv.org/abs/2506.05394.

[6] zh plus. zh-plus/tiny-imagenet: Datasets at HuggingFace. https://huggingface.co/datasets/zh-plus/tiny-imagenet.

	Attack							
Model	FGSM	I-FGSM	MI-FGSM	TI-MI-FGSM	PGD	Embedding	UAP	**Model Avg**
CLIP-RN50	0.248	0.117	0.233	0.257	0.141	0.166	0.174	0.191
CLIP-ViT	0.569	0.332	0.516	0.549	0.358	0.327	0.324	0.425
PE-Core	0.466	0.284	0.510	0.495	0.330	0.309	0.403	0.400
Attack Avg	0.428	0.244	0.420	0.434	0.276	0.267	0.300	

Table 8. Final transferability results