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 are:

- **Align Base** [6]: first proposed in [5], consists of a dual-encoder with an EfficientNet for vision and BERT for text
- **CLIP ResNet-50** [7]
- **CLIP ViT-B/16** [7]
- **Flava Full** [3]: presented in [10], uses a ViT-B/32 for both image and text encoder
- **Perception Encoder Core-B16-224** [1]: a SOTA CLIP-based model for zero-shot image and video classification
- **Siglip B16-224** [4]: introduced in [12], a CLIP-based model which uses the sigmoid loss function, operating solely on image-text pairs without requiring a global view of the pairwise similarities for normalization

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** [9]: 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
- **Ensemble**: generates perturbations by combining the gradients from multiple source models, by averaging them to produce a single, more transferable perturbation

Implementation Details In this zero-shot setting, models output a similarity score between the image embedding and a set of class text embeddings. The logits are defined as the cosine similarity between the image embedding and all candidate text embeddings, which are constructed by joining the prompt *"a picture of a"* with all possible class labels (e.g. *"a picture of a **dog**"*). The ground truth is the text embedding corresponding to the correct class label. The gradients are computed using the cross-entropy loss between the similarity scores (treated as logits) and the ground truth class index. The Embedding Space Disruption attack, however directly minimizes the L2 distance between the clean and adversarial image embeddings (i.e., $\mathcal{L} = ||E^{gt} - E^{adv}||_2$).

The attacks have been implemented from scratch: the source code is publicly available at [8].

4. Dataset

The tests are performed on the *zh-plus/tiny-imagenet* dataset from HuggingFace [11].

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. To ensure a fair evaluation of perturbation effectiveness, target models were tested only on images they correctly classified in their clean form.

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.

For the Ensemble attack, all available models -excluding the target- were used as sources. This attack was based on the *TI-MI-FGSM* variant, with a perturbation budget of $\epsilon = 4/255$ and a total of 5 iterations. At each iteration, all source models computed their gradients, which were then averaged to produce the final ensemble perturbation.

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, showing the different models' transferability of (and robustness to) adversarial perturbations in a transfer-based setting. 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 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 across all attacks with the considered model.

Table 9 reports the average robustness of each model under all attacks. As in the previous table, the last row shows the average effectiveness of each attack (identical to the last row of Table 8), and the last column reflects the average robustness of each model.

Table 10 shows the results of the Ensemble attack, where each model is attacked using an ensemble composed of all other models. The table reports the average robustness of each target model in terms of ASR.

Follow the three key findings that emerge from the analysis.

Model Robustness vs. Transferability The models show substantial differences in robustness and attack transferability:

- CLIP-RN50 is the least robust, typically showing the highest average ASR when targeted (see Table 9), and generating the least transferable perturbations when used as the source model (see Table 8). Siglip follows a similar pattern.
- PE-Core is the most robust when targeted, achieving the lowest ASR, followed by CLIP-ViT and Align.
- As source models, CLIP-ViT, Flava, and PE-Core produce the most transferable perturbations. In particular, CLIP-ViT and Flava achieve average ASR values of 0.504 and 0.502, respectively.

Single step vs. Iterative attacks Iterative methods generally yield higher ASR values than the single-step FGSM. However, FGSM can sometimes outperform them. This may be due to the perturbation "overfitting" the source model, reducing its transferability. Among all methods tested, Translation-Invariant Momentum FGSM (TI-MI-FGSM) consistently delivers the best overall performance, followed closely by the base FGSM and the MI-FGSM variant.

Single Surrogate vs. Ensemble. As shown in Tables 8 and 9, using a single surrogate model to craft adversarial perturbations can already achieve a reasonable ASR -for example, the TI-MI-FGSM variant reaches an ASR of 0.494. However, when multiple surrogate models are combined into an ensemble, the attack becomes significantly more effective, achieving an ASR of 0.66 (see Table 10). This highlights the ensemble's ability to generate more transferable and universally effective perturbations.

References

[1] D. Bolya, P.-Y. Huang, P. Sun, J. H. Cho, A. Madotto, C. Wei, T. Ma, J. Zhi, J. Rajasegaran, H. Rasheed,

	Target						
Source	Align	CLIP-RN50	CLIP-ViT	Flava	PE-Core	Siglip	**Avg**
Align	-	0.466	0.343	0.274	0.195	0.347	0.325
CLIP-RN50	0.400	-	0.250	0.257	0.243	0.627	0.355
CLIP-ViT	0.619	0.670	-	0.730	0.561	0.618	0.640
Flava	0.663	0.745	0.769	-	0.703	0.770	0.730
PE-Core	0.462	0.580	0.490	0.426	-	0.550	0.502
Siglip	0.426	0.406	0.368	0.441	0.221	-	0.372
Avg	0.514	0.573	0.444	0.426	0.385	0.582	

Table 1. FGSM

	Target						
Source	Align	CLIP-RN50	CLIP-ViT	Flava	PE-Core	Siglip	**Avg**
Align	-	0.437	0.133	0.147	0.089	0.220	0.205
CLIP-RN50	0.219	-	0.142	0.204	0.084	0.160	0.162
CLIP-ViT	0.345	0.604	-	0.593	0.159	0.460	0.432
Flava	0.245	0.539	0.352	-	0.167	0.460	0.353
PE-Core	0.264	0.434	0.280	0.305	-	0.320	0.321
Siglip	0.200	0.356	0.155	0.257	0.066	-	0.207
Avg	0.255	0.474	0.212	0.301	0.113	0.324	

Table 2. I-FGSM

	Target						
Source	Align	CLIP-RN50	CLIP-ViT	Flava	PE-Core	Siglip	**Avg**
Align	-	0.614	0.292	0.347	0.238	0.480	0.394
CLIP-RN50	0.437	-	0.233	0.311	0.127	0.465	0.315
CLIP-ViT	0.520	0.580	-	0.735	0.425	0.683	0.589
Flava	0.581	0.712	0.654	-	0.486	0.680	0.622
PE-Core	0.365	0.569	0.528	0.505	-	0.550	0.503
Siglip	0.286	0.359	0.337	0.366	0.238	-	0.317
Avg	0.438	0.567	0.409	0.453	0.303	0.572	

Table 3. MI-FGSM

	Target						
Source	Align	CLIP-RN50	CLIP-ViT	Flava	PE-Core	Siglip	**Avg**
Align	-	0.745	0.463	0.451	0.259	0.520	0.488
CLIP-RN50	0.376	-	0.324	0.360	0.152	0.412	0.325
CLIP-ViT	0.649	0.714	-	0.766	0.590	0.660	0.676
Flava	0.509	0.810	0.731	-	0.509	0.640	0.640
PE-Core	0.527	0.643	0.551	0.529	-	0.510	0.552
Siglip	0.271	0.275	0.373	0.296	0.192	-	0.281
Avg	0.466	0.637	0.488	0.480	0.341	0.548	

Table 4. TI-MI-FGSM

	Target						
Source	Align	CLIP-RN50	CLIP-ViT	Flava	PE-Core	Siglip	**Avg**
Align	-	0.575	0.208	0.238	0.132	0.370	0.305
CLIP-RN50	0.306	-	0.212	0.158	0.087	0.310	0.215
CLIP-ViT	0.425	0.570	-	0.453	0.327	0.550	0.465
Flava	0.343	0.485	0.393	-	0.252	0.408	0.376
PE-Core	0.368	0.491	0.216	0.262	-	0.376	0.342
Siglip	0.304	0.456	0.252	0.209	0.078	-	0.260
Avg	0.349	0.516	0.256	0.264	0.175	0.403	

Table 5. PGD

	Target						
Source	Align	CLIP-RN50	CLIP-ViT	Flava	PE-Core	Siglip	**Avg**
Align	-	0.535	0.229	0.220	0.224	0.390	0.320
CLIP-RN50	0.385	-	0.220	0.131	0.088	0.416	0.248
CLIP-ViT	0.430	0.588	-	0.438	0.245	0.510	0.442
Flava	0.419	0.571	0.467	-	0.333	0.550	0.468
PE-Core	0.352	0.422	0.200	0.268	-	0.420	0.332
Siglip	0.387	0.531	0.243	0.265	0.178	-	0.321
Avg	0.395	0.529	0.272	0.264	0.214	0.457	

Table 6. Embedding Space Disruption

	Target						
Source	Align	CLIP-RN50	CLIP-ViT	Flava	PE-Core	Siglip	**Avg**
Align	-	0.446	0.203	0.250	0.114	0.405	0.284
CLIP-RN50	0.311	-	0.185	0.202	0.158	0.387	0.249
CLIP-ViT	0.327	0.368	-	0.276	0.087	0.376	0.287
Flava	0.250	0.445	0.271	-	0.190	0.452	0.322
PE-Core	0.258	0.423	0.192	0.245	-	0.378	0.299
Siglip	0.199	0.354	0.196	0.191	0.105	-	0.209
Avg	0.269	0.407	0.209	0.233	0.131	0.400	

Table 7. UAP

	Attack							
Model	FGSM	I-FGSM	MI-FGSM	TI-MI-FGSM	PGD	Embedding	UAP	**Model Avg**
Align	0.325	0.205	0.394	0.488	0.305	0.320	0.284	0.331
CLIP-RN50	0.355	0.162	0.315	0.325	0.215	0.248	0.249	0.267
CLIP-ViT	0.640	0.432	0.589	0.676	0.465	0.442	0.287	0.504
Flava	0.730	0.353	0.622	0.640	0.376	0.468	0.322	0.502
PE-Core	0.502	0.321	0.503	0.552	0.342	0.332	0.299	0.407
Siglip	0.372	0.207	0.317	0.281	0.260	0.321	0.209	0.281
Attack Avg	0.487	0.280	0.457	0.494	0.327	0.355	0.275	

Table 8. Average transferability results

	Attack							
Model	FGSM	I-FGSM	MI-FGSM	TI-MI-FGSM	PGD	Embedding	UAP	**Model Avg**
Align	0.514	0.255	0.438	0.466	0.349	0.395	0.269	0.384
CLIP-RN50	0.573	0.474	0.567	0.637	0.516	0.529	0.407	0.529
CLIP-ViT	0.444	0.212	0.409	0.488	0.256	0.272	0.209	0.327
Flava	0.426	0.301	0.453	0.480	0.264	0.264	0.233	0.346
PE-Core	0.385	0.113	0.303	0.341	0.175	0.214	0.131	0.237
Siglip	0.582	0.324	0.572	0.548	0.403	0.457	0.400	0.469
Attack Avg	0.487	0.280	0.457	0.494	0.327	0.355	0.275	

Table 9. Average robustness results

Target Model	**Ensemble** (TI-MI-FGSM based)
Align	0.785
CLIP-RN50	0.706
CLIP-ViT	0.549
Flava	0.732
PE-Core	0.508
Siglip	0.679
Attack Avg	0.660

Table 10. ASR with Ensemble

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] facebook. facebook/flava-full. https://huggingface.co/facebook/flava-full.

[4] google. google/siglip-base-patch16-224. https://huggingface.co/google/siglip-base-patch16-224.

[5] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision, 2021.

[6] kakaobrain. kakaobrain/align-base. https://huggingface.co/kakaobrain/align-base.

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

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

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

[10] A. Singh, R. Hu, V. Goswami, G. Couairon, W. Galuba, M. Rohrbach, and D. Kiela. Flava: A foundational language and vision alignment model, 2022.

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

[12] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. Sigmoid loss for language image pre-training, 2023.