Adversarial Machine Learning Project Report: Defeating CLIPure-Cos with Purification-Aware Attack

Raghav Borikar - M24DS010
Byomakesh Panda - M24DS004

April 25, 2025

Executive Summary

This project had a focus on evaluating and circumventing the $CLIPure - Cos$ defense technique. We successfully replicated the $CLIPure - Cos$ defense results, confirming its robustness against standard adversarial attacks with approximately 72% average classification accuracy on benchmark datasets using the $ViT - L - 14$ model. When subjected to APGD attacks followed by $CLIPure - Cos$ purification, the model maintained 67% accuracy, demonstrating significant defense capabilities.

The core contribution of our project is the development of a novel Purification-Aware Attack (PAA) that specifically targets $CLIPure - Cos$'s defense mechanism. By incorporating a differentiable simulation of the purification process into our attack methodology, PAA successfully circumvented the defense 90% of the time across CIFAR-10, CIFAR-100, and ImageNet datasets. This demonstrates a fundamental vulnerability in current latent space purification approaches and provides insights for developing more robust defense techniques.

1 Introduction

1.1 Background on Multimodality

Recent advances in vision-language models like CLIP (Contrastive Language-Image Pre-training) have introduced new possibilities for both attacks and defenses. CLIP aligns image and text embeddings in a shared latent space, enabling zero-shot classification capabilities. However, like other neural networks, CLIP remains vulnerable to adversarial examples despite its robust pre-training.

1.2 Project Goals

Our project had two primary objectives:

1. **Replication Study:** Implement and validate the $CLIPure - Cos$ defense mechanism as described in the paper "CLIPure Purification in Latent Space via CLIP for Adversarially Robust ZERO-SHOT CLASSIFICATION" by Zhang et al. [2].
2. **Novel Attack Development:** Design and implement a new attack method capable of circumventing the $CLIPure - Cos$ defense by explicitly accounting for its purification process.

2 Literature Review

2.1 CLIPure and CLIPure-Cos

CLIPure represents a novel approach to adversarial defense through purification in the multi-modal latent space of CLIP. Unlike traditional purification methods that operate in pixel space, CLIPure leverages the well-aligned semantic space of CLIP to remove adversarial perturbations [2]. The paper introduces two variants:

- $CLIPure - Diff$: Uses the DiffusionPrior module from DALLE-2 to model likelihood in latent space
- $CLIPure - Cos$: Models likelihood using cosine similarity between image embeddings and a blank template text embedding ("a photo of a ...")

$CLIPure - Cos$ stands out as the first purification method that doesn't rely on generative models, offering substantially improved defense efficiency. The approach works by normalizing latent vectors to unit vectors and performing gradient ascent to maximize the cosine similarity to a blank template [2]. Key advantages of $CLIPure - Cos$ include:

- No additional training required (uses off-the-shelf CLIP models)
- Inference efficiency (only 1.14× the time of vanilla CLIP)
- State-of-the-art robustness across multiple datasets

2.2 Adversarial Attack Methods

Standard adversarial attacks like AutoAttack and APGD have shown effectiveness against many defense mechanisms but struggle against $CLIPure - Cos$. These attacks typically work by maximizing classification loss within an ϵ-bound perturbation constraint but fail to account for the purification process applied by defenses like $CLIPure - Cos$.

Recent work has explored adaptive attacks designed specifically for purification defenses, but these approaches often treat the purification as a black box rather than explicitly modeling its internal dynamics.

3 Methodology

3.1 Replication of CLIPure-Cos

We implemented $CLIPure - Cos$ following the approach described in the original paper [2]. The core components include:

- **CLIP Latent Space Access:** Using the pre-trained CLIP model to extract image and text embeddings.
- **Unit Vector Normalization:** Converting embeddings to unit vectors to focus on direction rather than magnitude.
- **Cosine Similarity Maximization:** Purifying adversarial examples by increasing similarity to blank template embedding.

- **Zero-Shot Classification:** Classifying the purified embeddings by comparing to text embeddings of candidate classes.

The purification algorithm follows the steps outlined in Algorithm 1 of the paper, with image embeddings updated iteratively to maximize cosine similarity with the text embedding of a blank template [2].

3.2 Development of Purification-Aware Attack (PAA)

Our novel Purification-Aware Attack (PAA) was designed to explicitly account for and exploit the $CLIPure-Cos$ purification process. The key innovations include:

- **Differentiable Purification Simulation:** Modeling the entire purification process as a differentiable function to enable end-to-end optimization through backpropagation [1].
- **Adversarial Objective Formulation:** Optimizing perturbations δ to minimize the objective function, where z_i^{pure} represents the purified image embedding, z_t^{target} the target class text embedding, and z_t^{null} the null template text embedding [1]:

$$\delta^* = \arg\min_{\delta} E[\cos(z_i^{pure}, z_t^{target}) - \cos(z_i^{pure}, z_t^{null})]$$

- **Momentum Integration:** Accounting for the momentum used in $CLIPure-Cos$'s gradient updates, where $m^{(k)}$ is the momentum at step k, γ is the momentum factor, and ∇_u is the gradient w.r.t the latent vector u:

$$m^{(k+1)} = \gamma m^{(k)} + (1-\gamma)\nabla_u$$

- **Latent Space Deformation:** Creating perturbations that appear aligned with the null template under purification while maintaining hidden alignment with the target class [1].

The attack works by solving the optimization problem, where L_{CE} is the cross-entropy loss computed on the purified embedding z_i^{pure} with respect to the target label y_{target}, λ is a regularization parameter, and $||\delta||_p$ is the L_p norm of the perturbation [1]:

$$\min_{\delta} L_{\text{attack}} = L_{\text{CE}}(z_i^{pure}, y_{target}) + \lambda||\delta||_p$$

The key innovation is that the classification loss is computed on the purified embedding rather than the initial adversarial embedding [1].

4 Experimental Setup

4.1 Datasets

We evaluated both $CLIPure-Cos$ and our PAA attack on three benchmark datasets:

- **CIFAR-10:** 10 classes, 32×32 RGB images
- **CIFAR-100:** 100 classes, 32×32 RGB images
- **ImageNet:** 1,000 classes, varying image resolutions

For each dataset, we used the standard test split to evaluate classification accuracy.

4.2 Models

We focused on the $ViT - L - 14$ variant of CLIP, which consists of:

- A Vision Transformer (ViT) for the image encoder
- A Transformer-based text encoder
- 427 million parameters total

This model was selected for its strong performance in the original CLIPure paper and its widespread use in vision-language applications.

4.3 Implementation Details

Our implementation used the following parameters for $CLIPure - Cos$:

- Purification steps: 10
- Purification step size: 30
- Unit vector normalization at each step
- Zero-shot classification with 80 prompt templates

For PAA implementation, we incorporated:

- Differentiable simulation of all 10 purification steps
- Chain rule backpropagation through the entire purification process
- Momentum matching to align with $CLIPure - Cos$'s update rule
- L_∞ norm constraint on perturbations ($\epsilon = 8/255$)

5 Results

5.1 CLIPure-Cos Replication Results

Our replication of $CLIPure - Cos$ achieved results consistent with those reported in the original paper.

Table 1: CLIPure-Cos Replication Accuracy (%)

Dataset	Clean Accuracy	Robust Accuracy (APGD)
CIFAR-10	72.3%	67.5%
CIFAR-100	73.0%	65.0%
ImageNet	71.8%	68.4%
Average	72.4%	67.0%

These results confirm $CLIPure - Cos$'s effectiveness as a defense mechanism, showing significant robustness against standard adversarial attacks while maintaining strong clean accuracy.

5.2 PAA Attack Effectiveness

Our Purification-Aware Attack demonstrated remarkable effectiveness against $CLIPure-Cos$. The attack success rate indicates the percentage of adversarial examples that successfully fool the model despite $CLIPure - Cos$ purification.

Table 2: PAA Attack Success Rate (%) against CLIPure-Cos

Dataset	Attack Success Rate
CIFAR-10	91.4%
CIFAR-100	99.2%
ImageNet	89.1%
Average	93.2%

These results show that PAA is highly effective at circumventing the $CLIPure - Cos$ defense.

6 Discussion

6.1 Analysis of PAA's Success

Several factors contribute to PAA's impressive effectiveness against $CLIPure - Cos$:

- **Differentiable Purification Chain:** By modeling the entire purification process as a differentiable function, PAA can optimize adversarial perturbations that remain effective even after multiple purification steps.
- **Latent Space Deformation:** PAA creates a specific type of deformation in CLIP's latent space that fools the purification process by appearing to align with the null template while retaining adversarial properties.
- **Geometric Exploitation:** The attack leverages the high-dimensional spherical geometry of CLIP's latent space to create "trap directions" in the embedding space that exploit the curvature of the cosine similarity manifold [1].
- **Momentum Integration:** By accounting for the momentum used in $CLIPure - Cos$'s gradient updates, PAA can anticipate how perturbations will transform under purification.

The bypass mechanism of PAA works by pre-empting the purification trajectory, creating adversarial examples that:

- Maximize similarity to the null template (fooling purification)
- Retain residual similarity to the target class (maintaining attack success)
- Exploit momentum memory through coordinated perturbation updates [1]

6.2 Implications for Adversarial Defenses

Our findings have several implications for adversarial defenses:

- **Defense Transparency Risk:** Transparent defense mechanisms like $CLIPure - Cos$ can be vulnerable when their internal workings are used to guide attack design.

- **Multi-Stage Defenses:** Future defenses might need to incorporate multiple, diverse purification approaches to increase the difficulty of end-to-end optimization.
- **Adaptive Defense Strategies:** Defenses could potentially incorporate randomization or non-differentiable components to complicate attack modeling.
- **Latent Space Properties:** The geometric properties of high-dimensional embedding spaces need further investigation to understand their vulnerability to adversarial manipulation.

7 Conclusion

7.1 Summary of Findings

This project demonstrated both the strengths and limitations of $CLIPure - Cos$ as an adversarial defense mechanism. While it provides significant robustness against standard attacks, it remains vulnerable to attacks specifically designed to account for its purification process.

Our Purification-Aware Attack (PAA) demonstrates that explicit modeling of defense mechanisms during attack generation can substantially increase attack effectiveness. By treating the entire purification process as a differentiable function, PAA achieves a 90% success rate against $CLIPure-Cos$ across multiple datasets.

7.2 Future Work

Several directions for future work emerge from this project:

- **Defense Enhancement:** Investigating modifications to $CLIPure - Cos$ that could resist PAA, potentially through non-differentiable components or randomized purification.
- **Theoretical Analysis:** Further mathematical analysis of the geometric properties of CLIP's latent space and how they affect adversarial vulnerability.
- **Extension to Other Models:** Applying PAA concepts to other defense mechanisms that utilize different types of purification.
- **Detection Methods:** Developing techniques to detect adversarial examples crafted using PAA by identifying their unique latent space signatures.
- **Efficiency Optimization:** Improving the computational efficiency of PAA to enable larger-scale evaluations and real-time applications.

References

[1] Conceptual Formulation of PAA, April 10, 2025.

[2] Zhang, M., Bi, K., Chen, W., Guo, J., & Cheng, X. CLIPure Purification in Latent Space via CLIP for Adversarially Robust ZERO-SHOT CLASSIFICATION.