Conceptual Foundation of Purification-Aware Attack (PAA)

April 10, 2025

Core Insight: Latent Space Dynamics of CLIPure-Cos

The CLIPure-Cos defense operates by maximizing cosine similarity between adversarial image embeddings and a blank template embedding ("a photo of a .") through iterative purification.

Key Mathematical Properties

Unit Normalization:

$$u = \frac{z_i}{\|z_i\|_2}$$

Gradient Ascent:

$$u^{(k+1)} = u^{(k)} + \eta \nabla_u \left(\cos(u, z_t^{\text{null}})\right)$$

Momentum Integration:

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

Attack Strategy: Adversarial Optimization Through Purification

PAA introduces a differentiable simulation of CLIPure-Cos purification during attack generation. The attack solves:

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

where

$$z_i^{\text{pure}} = \text{Purify}(\text{Enc}_i(x + \delta))$$

Mathematical Framework

Differentiable Purification Chain:

$$\frac{\partial \mathcal{L}}{\partial \delta} = \frac{\partial \mathcal{L}}{\partial z_i^{\text{pure}}} \cdot \frac{\partial z_i^{\text{pure}}}{\partial z_i} \cdot \frac{\partial z_i}{\partial x} \cdot \frac{\partial x}{\partial \delta}$$

This chain backpropagates through all purification steps.

Adversarial Objective:

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

Key Innovations

Purification-Aware Gradients

- Explicitly models momentum-based purification dynamics.
- Maintains unit sphere constraints during perturbation crafting.

Latent Space Deformation

Creates adversarial directions that:

- Appear aligned with null template under purification.
- Maintain hidden alignment with target class.

Geometric Exploitation

Leverages high-dimensional spherical geometry of CLIP's latent space to:

- Create "trap directions" in embedding space.
- Exploit curvature of cosine similarity manifold.

Defense Bypass Mechanism

The attack strategically:

- Pre-empts purification trajectory by anticipating gradient steps.
- Encodes dual alignment where purified embeddings simultaneously:
  - Maximize similarity to null template (fooling purification).
  - Retain residual similarity to target class (maintaining attack success).
- Exploits momentum memory through coordinated perturbation updates.

Theoretical Advantages Over Standard Attacks

- **Invariance to Purification Iterations:** Attack remains effective regardless of purification steps.
- **Adaptive to Defense Parameters:** Automatically adjusts to CLIPure's step size (η) and momentum (γ).
- **Dimension-Agnostic:** Effectiveness scales with CLIP's embedding dimension (typically 512–768D).

This approach fundamentally subverts CLIPure-Cos's defense mechanism by turning its purification process into an attack vector through differentiable simulation and geometric manipulation of the latent space.