Alexandre Francony, Léonard Seidlitz, Raphaël Roux, Adrien Servas, Romain Requena
ESILV – DIA5 A5 – 20/11/2025

Project Proposal: AI-Resistant Image Filtering through Adversarial Perturbations

Problem and Motivation

Modern deep learning models for image recognition, object detection, and generative tasks have become ubiquitous tools for both research and commercial applications. However, this widespread adoption raises significant concerns regarding intellectual property protection, particularly for artists and creators whose work is increasingly used as training data for generative AI systems without consent. While users have limited control over how AI models utilize their digital assets, the absence of robust protective mechanisms creates a power imbalance between creators and AI systems. This project addresses this challenge by developing a practical solution: an **AI-resistant image filter** that introduces imperceptible adversarial perturbations to images, rendering them incompatible with state-of-the-art vision models while preserving visual quality from the human perspective.

Approach and Methods

Our methodology consists of two complementary phases. First, we will train a classical deep learning-based image recognition model (such as ResNet or VGG) on a well-defined classification task (e.g., artwork type detection, subject categorization, or scene understanding). We will maintain detailed logs of misclassifications to identify systematic failure patterns and vulnerable feature representations within the model. In the second phase, we leverage these insights to design an adversarial filtering mechanism inspired by "poisoning" techniques used in adversarial machine learning research. By computing minimal perturbations that maximize classification error while remaining visually imperceptible, we will generate a filter that disrupts model predictions without degrading image aesthetics. This approach differs from traditional watermarking by operating in the feature space rather than modifying visible metadata.

Datasets and Implementation

We will utilize publicly available datasets such as CIFAR-10 or a custom curated collection of artwork images. The model training will employ TensorFlow/PyTorch frameworks with standard evaluation metrics (accuracy, precision, recall). For the adversarial filtering component, we will implement techniques derived from Fast Gradient Sign Method (FGSM) or other targeted perturbation algorithms adapted for preserving visual fidelity.

Evaluation

Quantitative evaluation will measure: (1) classification accuracy degradation on the protected images versus original images, and (2) perceptual quality metrics (PSNR, SSIM) to ensure imperceptibility. Qualitative assessment will involve testing the filter's robustness against various generative AI tasks, using models like CLIP or GPT-4V for style transfer demonstrations (e.g., transforming a filtered Mona Lisa to "chibi style" to observe performance degradation).

Expected Outcomes

This project bridges computer vision fundamentals with adversarial machine learning, contributing practical tools for digital rights protection while advancing our understanding of model vulnerabilities and defense mechanisms—skills directly aligned with this course's objectives.