

Detecting Adversarial Attacks: CNN vs. Vision Transformer Pipelines

Eric Tosounidis, Shayl Chetty, Olateju Ojeyinka, advised by Miao Wang, Gary Kazantsev, and Dr. Ali Hirsra | Bioomberg–Columbia • IEOR E4737

Abstract

Adversarial attacks can cause image classifiers to fail while making only imperceptible changes to the input. We compare two detection approaches: an explanation-based pipeline using Grad-CAM on a CNN and a representation-based pipeline using frozen DINOv2 embeddings. We evaluate detection performance on Imagenette and CIFAR-10, and a five-class ImageNet subset, with a focus on cross-attack generalization. Our results show that PGD-trained detectors generalize well to some attacks but transfer poorly to others, consistent with differences in representation-space geometry.

Background

- Adversarial examples are small, often imperceptible perturbations that cause confident misclassification.
- Prior work shows such examples can be generated efficiently and can transfer across models.
- Two main defense strategies:
 - Robustness: modify the model (e.g., adversarial training).
 - Detection: keep the model unchanged and flag adversarial inputs (focus of this project).

Prior Work

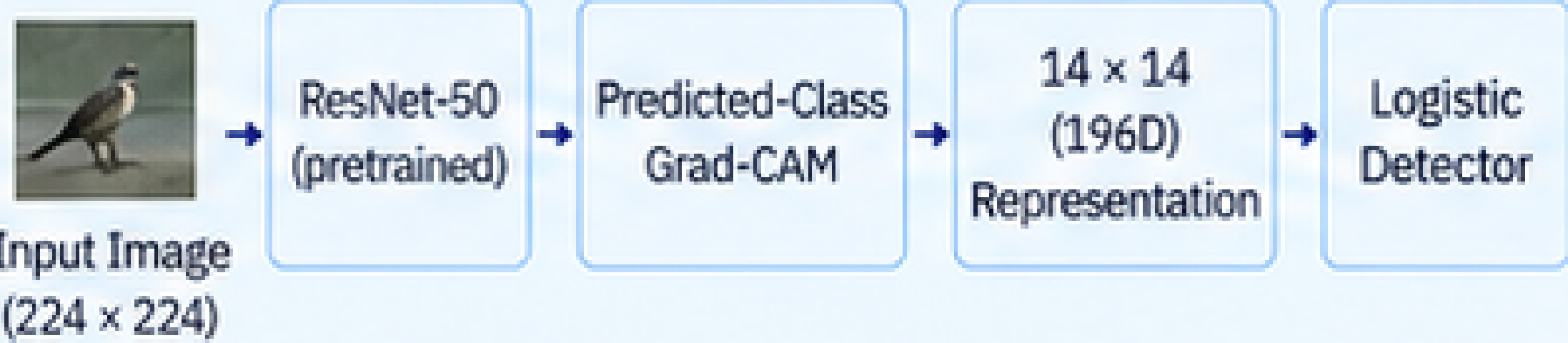
- Early detection work used hand-crafted features and input transformations (e.g., Feinman et al.).
- LIME was shown to be a useful signal for adversarial detection (Ghorbani et al.).
- More recent work uses internal representations from modern vision models, including self-supervised encoders like DINOv2.
- However, detection often evaluates only within a single attack type, leaving cross-attack generalization less understood.

Why It Matters

- Image classification is used in safety- and security-critical domains (autonomous navigation, medical triage, content moderation, biometric authentication).
- A reliable detection mechanism can help prevent adversarial inputs from causing harmful decisions, especially when retraining is not feasible.

Experimental Setup

CNN + Grad-CAM Pipeline (Explanation-Based)

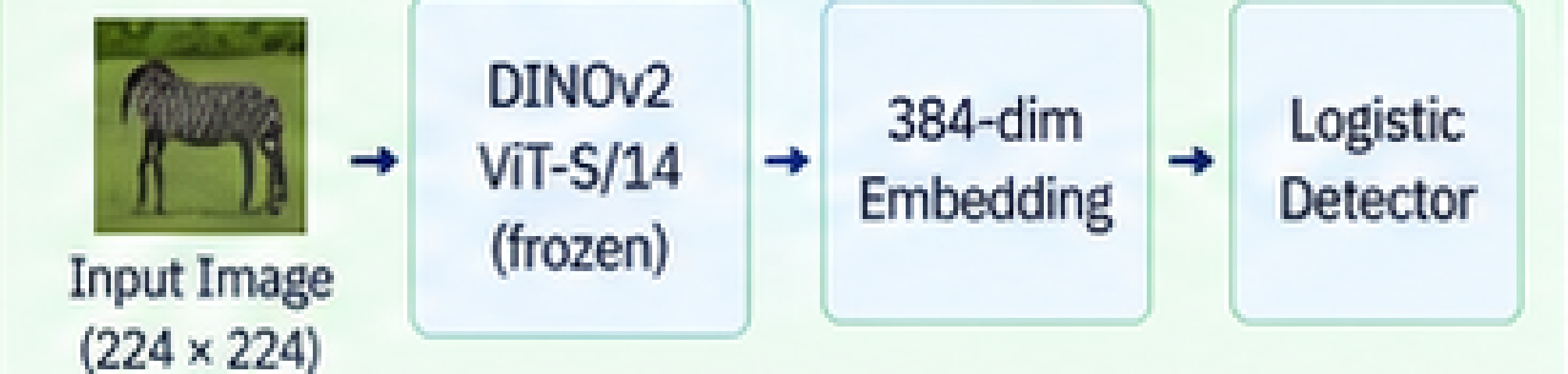


Idea: Use distributional features from Grad-CAM attention maps to detect adversarial inputs.

Details:

- **Model:** ResNet-50 (pretrained on ImageNet)
- **Grad-CAM:** predicted class, final convolutional layer
- **Representation:** 14 × 14 Grad-CAM map (196-dimensional)
- **Detectors tested:** logistic regression (main), heatmap CNN, autoencoder, Mahalanobis detector
- **Evaluation dataset:** 1,000 correctly classified Imagenette validation images (code also supports CIFAR-10)
- **Attacks:** PGD ($\epsilon = 8/255$), APGD, localized patch, FGSM, CW

DINOv2 Representation Pipeline (Representation-Based)



Idea: Detect adversarial inputs directly in frozen DINOv2 representation space.

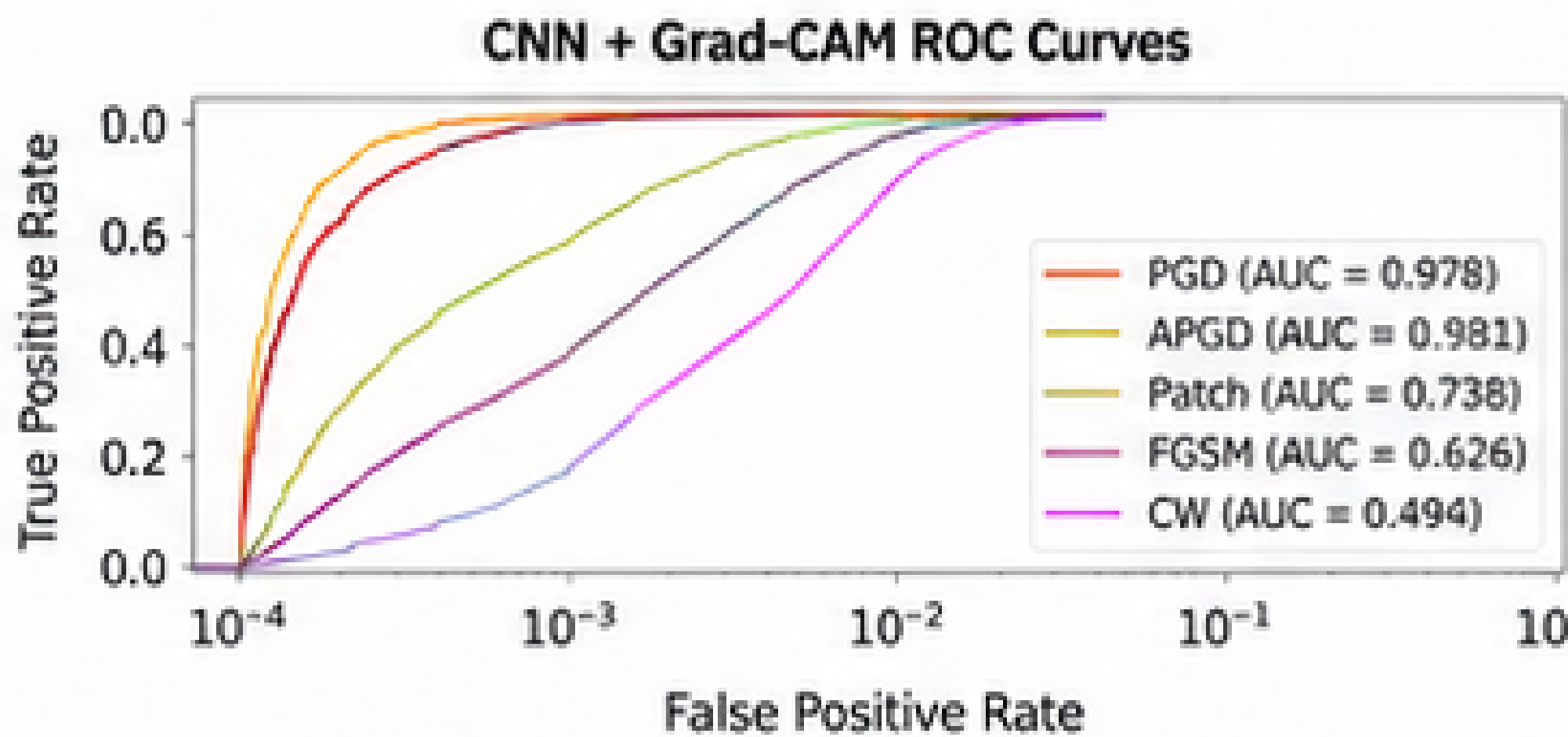
Details:

- **Model:** DINOv2 ViT-S/14 (frozen, self-supervised)
- **Representation:** 384-dimensional embedding
- **Detector:** logistic regression (trained on PGD examples)
- **Datasets:** CIFAR-10 (primary) and five-class ImageNet subset (375 train, 100 internal val, 250 official val)
- **Attacks:** PGD ($\epsilon = 8/255$), APGD, CW ($c = 1$), localized patch
- **Evaluation:** ROC AUC, accuracy, adversarial recall (threshold fixed after training)
- Gaussian noise (where included) treated as a corruption control, not an attack

Key Results

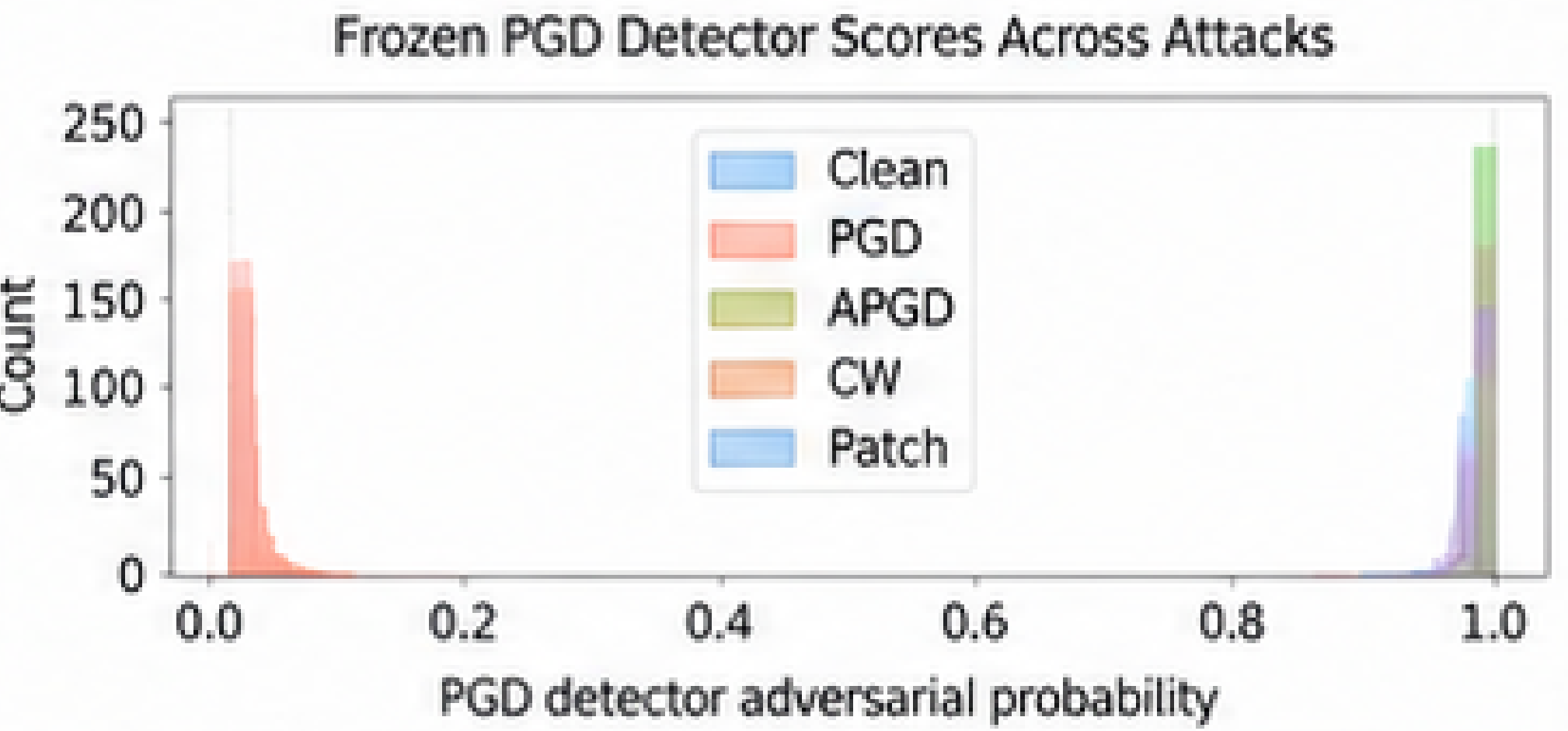
CNN + Grad-CAM Results (Imagenette, PGD-Trained Detector)

Test Attack	AUC	TPR @ 5% FPR
PGD	0.978	0.935
APGD	0.981	0.950
Patch	0.738	0.295
FGSM	0.626	0.180
CW	0.494	0.035



DINOv2 Results (CIFAR-10, PGD-Trained Detector)

Test Attack	Accuracy	AUC	Adversarial Recall
PGD	100.00%	1.0000	100.0%
APGD	98.25%	0.9896	96.5%
CW	52.50%	0.7612	5.0%
Patch	96.50%	0.9500	93.0%



Results & Analysis

- Strong Within-Attack Detection for PGD**
 - CNN (Grad-CAM): PGD AUC = 0.978, TPR@5%FPR = 0.935 on Imagenette.
 - DINOv2: PGD is almost perfectly detected on CIFAR-10 (AUC = 1.000, 100% recall) and strongly detected on ImageNet-5 (AUC = 1.000, 98% recall).
- Cross-Attack Generalization Differs by Pipeline**
 - CNN (Grad-CAM): PGD-trained detector transfers well to APGD (0.950) but much worse to FGSM (0.180) and CW (0.035).
 - DINOv2: PGD-trained detector transfers well to APGD (96.5% recall) and patch (93.0% recall), but weakly to CW (5.0% recall on CIFAR-10, 30.0% on ImageNet-5).
- Representation Geometry Helps Explain Transfer**
 - Attacks that transfer well tend to cause larger shifts in representation space and greater movement along the detector direction.
 - CW preserves substantially more of the clean representation (e.g., cosine similarity 0.6983 on CIFAR-10) and moves less along the PGD detector direction (mean shift 3.66).
- External Validation on ImageNet (5 Classes)**
 - Classifier accuracy is 98.8% on 250 untouched validation images.
 - PGD ASR is 100% and detector recall is 98.0% (AUC = 1.000).
 - CW ASR is 98.38%, but detector recall at the fixed PGD threshold is 30.0% (AUC = 0.9646).
 - The same qualitative PGD–CW pattern appears as on CIFAR-10: PGD induces a much larger representation shift than CW.

Conclusion and Future Work

- A PGD-trained detector in DINOv2 representation space generalizes to some attacks (APGD, patch) but not others (CW).
- The CNN Grad-CAM detector also shows strong transfer to APGD but weak transfer to FGSM and CW.
- Cross-attack transfer is associated with representation-space geometry rather than pixel-space perturbation structure alone.
- A smaller ImageNet experiment reproduces the key PGD–CW distinction on higher-resolution images.
- Future work: evaluate more attacks, encoders, and datasets; investigate training on multiple attack-induced geometries; and further compare explanation- and representation-based signals.

Adversarial detection is not just about whether an attack works, but how it changes what the model sees inside.