Project Progress Summary: Week 2 Update

Kecen Li

February 3, 2026

1 Overview

This report summarizes the progress made during Week 2, addressing the communication gaps from Week 1 regarding engineering infrastructure and presenting the latest experimental findings on Neural Activation Coverage (NAC) characteristics.

2 Part 1: Infrastructure & Validation (Retrospective on Week 1)

To ensure the theoretical validity of our experiments, I established a robust evaluation infrastructure that was not fully detailed in the previous meeting. This foundation is critical for the credibility of all subsequent results.

- **Normalization Middleware (Data Integrity)**: I implemented a transparent middleware layer to resolve the conflict between `Advex-UAR` (requiring ImageNet normalization) and `RobustBench` (requiring raw $[0, 1]$ inputs). This ensures that attacks like AutoAttack and DeepG operate in the correct physical space, guaranteeing that our observed results are genuine adversarial effects rather than preprocessing artifacts.
- **Architecture Compatibility Shim**: I developed a compatibility layer to decouple NAC from specific architecture constraints (e.g., hardcoded 'layer4'). This infrastructure allows us to seamlessly switch between standard ResNets, WideResNets (Adversarial Training), and Vision Transformers (ViT) without rewriting critical paths.

3 Part 2: Key Experimental Findings (Week 2)

Based on the validated infrastructure, we conducted a comprehensive analysis of NAC's behavior across distributions.

3.1 1. The Sensitivity Spectrum of NAC

Our OODRobustBench sweep reveals that NAC is not a universal OOD detector but has a distinct "sensitivity spectrum":

1. **High Sensitivity (Structural/Texture Disruption)**: NAC performs expertly (AUROC > 0.90) on corruptions that destroy local structure, such as Impulse Noise, Pixelate, and L_∞ Adversarial Attacks.
2. **Low Sensitivity (Global Photometric Shifts)**: NAC fails to detect global shifts like Brightness, Fog, or Snow (AUROC ≈ 0.50 - random), suggesting that internal features adapt linearly to these changes without disrupting activation patterns.

3.2 2. The Robustness Paradox

We observed a counter-intuitive phenomenon when comparing Standard Models vs. Adversarially Trained (Robust) Models:

- **Observation**: Robust models exhibit significantly smaller variations in NAC coverage under attack (AUROC ≈ 0.52) compared to standard models (AUROC ≈ 0.90).
- **Implication**: Adversarial training forces internal features to remain stable under perturbation. Ironically, this stability makes OOD samples look "normal" to the NAC detector, making NAC less effective on robust models.

3.3 3. Reality Check: APS vs. Non-APS

To address concerns about overfitting, we compared Automatic Parameter Search (APS) against a fixed Non-APS baseline:

- **APS (Upper Bound)**: achieves near-perfect separation on Near-OOD but relies on OOD validation data.
- **Non-APS (Realistic)**: maintains performance on Far-OOD ($\sim 94\%$ AUROC) but drops on Near-OOD. We will use Non-APS as our primary metric for realistic evaluation moving forward.

4 Part 3: Roadmap Capabilities (Ready-to-Deploy)

Leveraging the infrastructure described in Part 1, the following capabilities are implemented and ready for the next phase of analysis:

- **Vision Transformers (ViT)**: Scripts are ready to evaluate whether ViT's global attention mechanism mitigates the "Low Sensitivity" to photometric shifts observed in CNNs.
- **Layer Ensembling**: Infrastructure is in place to aggregate NAC scores from multiple depths (e.g., layers 1+2+3), potentially capturing low-level texture anomalies that deep feature layers miss.