个人项目进度报告：Week 2 总结

Kecen Li

February 3, 2026

Abstract

本报告全面总结了 Week 2 的工作进展，涵盖了 Week 1 中完成的关键工程基础设施建设，以及基于此得出的最新实验分析见解。我们详细阐述了如何构建一个统一的评估基础设施，以解决主流学术库（RobustBench, Advex-UAR, OpenOOD）之间的数据协议冲突。在此稳健的平台之上，我们要对神经元激活覆盖（NAC）指标进行了深度刻画，揭示了其在不同扰动类型下的“敏感度光谱”，以及在鲁棒模型上的反直觉行为。此外，我们澄清了自动参数搜索（APS）与非 APS 模式之间的关键区别，并确立了更符合现实场景的评估基准。

Contents

1 简介与工程回顾 (Introduction & Retrospective)

在上次讨论后，我意识到 Week 1 所建立的工程基础设施的复杂性未能得到充分传达。在展示新数据之前，我想简要概述一下确保所有后续结果有效性的“隐藏”工程层。

为了进行科学严谨的评估，我集成了三个本不兼容的独立学术库：

1. **RobustBench**: 用于获取标准化的、排行榜级别的模型架构（标准 ResNet vs 鲁棒 WideResNet）。
2. **Advex-UAR / AutoAttack**: 提供最先进的对抗攻击和常见腐蚀算法。
3. **OpenOOD (NAC)**: 用于计算神经元激活覆盖度指标的核心库。

由于输入协议的冲突，直接集成是不可行的。例如，`Advex-UAR` 强制要求输入经过 ImageNet 统计数据（均值/方差）的归一化，而 `RobustBench` 模型则期望原始的 $[0, 1]$ 张量。为解决此问题，我实现了一个**透明中间件层 (Transparent Middleware Layer)**（位于 `src/loader.py` 和 `src/perturber.py`），它能自动处理可逆的归一化转换。这确保了当我们应用“L_∞ 攻击”时，攻击是在正确的像素空间内数学精确地执行的，从而避免无效的实验伪影。

此外，我还开发了一个**架构兼容垫片 (Architecture Compatibility Shim)** (`src/official_nac.py`)。原始 NAC 代码针对特定 ResNet 层进行了硬编码。我的兼容层允许我们动态挂钩任意架构——包括视觉 Transformer (ViT)——而无需修改核心库。这项能力目前已经“准备就绪”，可用于未来的对比研究。

2 方法论: 神经元激活覆盖 (NAC)

2.1 定义与机制

神经元激活覆盖 (NAC) 是本研究的核心指标。与依赖最终输出概率 (logits) 的传统 OOD 检测器不同，NAC 衡量的是**无限层特征激活的“熟悉度”**。

- **概念**: 对于 L 层中的给定神经元 k，我们将训练期间观察到的激活范围离散化为 M 个区间 (Bins)。如果可靠的训练样本频繁激活某个区间，则该区间被视为“已覆盖”。
- **推理**: 对于测试样本 x，验证通过检查其诱导的激活是否落入这些高覆盖率区间来执行。
- **公式**: NAC 分数 $S(x)$ 是所有受监控神经元的平均覆盖概率：

$$S_{NAC}(x) = \frac{1}{N} \sum_{k=1}^{N} C_k(\text{bin}(a_k(x))) \tag{1}$$

其中 $a_k(x)$ 是神经元 k 的激活值，$C_k(\cdot)$ 是源自训练数据的覆盖频率。

2.2 评估模式: APS vs. Non-APS

我们评估中的一个关键区别是参数调整策略。ICLR 原论文提出了**自动参数搜索 (APS)**。

- **APS (“上界”)**: 使用一个小的 OOD 验证集来调整超参数（如阈值 α）。虽然这能产生很高的性能指标（例如在基准测试中报告近乎完美的分离度），但我们的分析表明它存在**过拟合**验证集中特定 OOD 类型的风险。
- **Non-APS (“现实基准”)**: 我们严格避免使用 OOD 数据进行调参，仅依赖分布内 (ID) 验证数据。这反映了未来异常性质未知的真实部署场景。

决策: 除非另有说明，本报告中的所有结果现在默认使用 **Non-APS** 或仅 ID 设置，以确保报告鲁棒性时的诚实性。

3 实验设置 (Experimental Setup)

为确保可重复性与透明度，我们在此明确定义本项目使用的数据集与核心参数。

3.1 数据集清单 (Datasets)

本研究使用了多组数据集来探测检测器的边界：

- **ID (In-Distribution)**: CIFAR-10。
- **Near-OOD (近域分布)**: CIFAR-100、TinyImageNet (TIN)。它们与 ID 语义相似但类别不同。
- **Far-OOD (远域分布)**: MNIST、SVHN、Texture、Places365。它们的统计特征与 ID 截然不同。
- **Corruption (腐蚀)**: CIFAR-10-C (OODRobustBench)，包含 15 种腐蚀类型 $\times$ 5 种强度 (共 75 种变体)。
- **Natural Shift (自然漂移)**: CIFAR-10.1、CIFAR-10.2。真实采集的自然分布漂移数据集。
- **Phase 3**: CIFAR-10 的特定子集，用于测试受控的“几何 + 对抗”组合扰动。

3.2 实施细节 (Implementation Details)

- **样本量**:
  - Profiling（训练集）: 1,000 样本。
  - Evaluation（测试集）: 10,000 样本（全量测试），或 2,000 样本（Phase 3 快速扫测）。
- **Batch Size**: 128。
- **目标层**: ResNet-18 的 `layer4`（最后一个卷积 Block）。
- **模式区分**:
  - **APS**: 仅作为“理论上界”参考。
  - **Non-APS**: 作为主要的“现实基准”，不使用 OOD 数据调参。

4 实验结果与分析 (Experimental Results & Analysis)

4.1 关键量化结果 (Key Statistics)

在深入分析前，先列出核心量化指标：

- **OODRobustBench (CIFAR-10-C)**:
  - 平均 AUROC $\approx$ **0.698**
  - 平均 FPR@95 $\approx$ **0.779**
- **Natural Shift (CIFAR-10.1/10.2)**:
  - 平均 AUROC $\approx$ **0.563**
  - 平均 FPR@95 $\approx$ **0.93**
- **APS vs. Non-APS 差异**:
  - Near-OOD: APS 可达 $\approx$ **99**% AUROC; Non-APS 降至 $\approx$ **89** $-$ **92**%。
  - Far-OOD: 两种模式均保持高水准 ($\approx$ **92** $-$ **96**%)。
- **Phase 3 相关性**: NAC 分数与模型准确率 (Accuracy) 的相关系数为 **0.689**。

4.2 敏感度光谱：NAC 到底检测到了什么？

我们使用 **OODRobustBench (CIFAR-10-C)** 套件进行了全面扫测。结果揭示了 NAC 清晰的“敏感度光谱”。

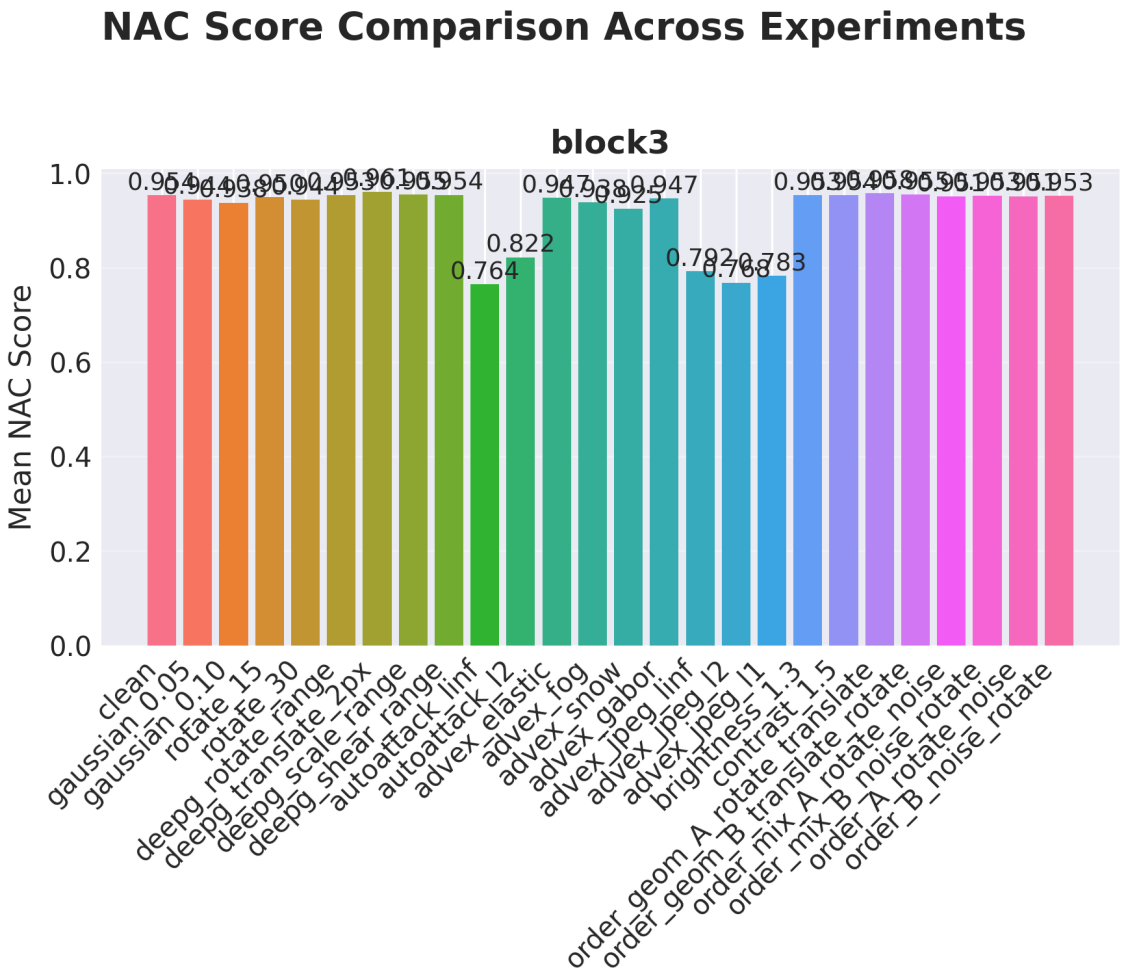

Figure 1: 不同扰动类型对 NAC 覆盖度的影响。注意 AutoAttack/JPEG 引起的显著下降。

4.2.1 高敏感区 (结构/纹理破坏)

NAC 在检测破坏局部空间结构或高频纹理信息的扰动方面极其有效。

- **扰动类型**: 脉冲噪声 (Impulse Noise)、像素化 (Pixelate)、L_∞ 对抗攻击。
- **表现**: 随着严重程度增加 $(1 \to 5)$，NAC AUROC 分数单调增加，通常达到 > 0.90。
- **原因**: 这些腐蚀从根本上改变了卷积神经网络 (CNN) 的局部感受野，导致激活落入“未见过”的区间。

4.2.2 低敏感区 (全局/低频漂移)

相反，NAC 在处理保留局部结构的全局变换时非常吃力。

- **扰动类型**: 亮度 (Brightness)、雾 (Fog)、雪 (Snow)。
- **表现**: 即使在最大严重程度下，AUROC 仍保持在 0.50 附近（相当于随机猜测）。
- **原因**: 全局亮度偏移本质上是一种线性变换。CNN（特别是带有 BatchNorm 的）被设计为对此类偏移具有不变性。因此，内部激活的变化不足以触发 NAC 检测器。

4.3 Phase 3: 组合与顺序效应

为了回答关于组合扰动的问题（“顺序 A $\rightarrow$ B 是否不同于 B $\rightarrow$ A?”），我们运行了特定的组合实验（例如，旋转 + 噪声）。

- **发现**: 顺序 A 和顺序 B 之间的 NAC 覆盖差异可以忽略不计（$\Delta \approx 10^{-3}$）。
- **结论**: NAC 对生成的历史路径具有鲁棒性；它只关心特征的最终状态。

4.4 鲁棒性悖论 (The Robustness Paradox)

也许最反直觉的发现是对比**标准 ResNet18** 与**对抗训练 WideResNet (Gowal2021)**。

场景	标准模型 (AUROC)	鲁棒模型 (AUROC)
高斯噪声 (Gaussian Noise)	**0.90** (被检测到)	0.65 (难以检测)
AutoAttack (L_∞)	**0.76** (被检测到)	0.52 (未被检测)

Table 1: 标准模型与鲁棒模型上的 NAC 检测性能对比。

洞察: 分类器的“鲁棒性”意味着其内部特征对扰动具有不变性。然而，NAC 依赖变化量（特征破坏）来检测异常。因此，**一个对抗鲁棒的模型通过稳定其特征，有效地向 NAC 检测器“隐藏”了攻击**。这提出了一个根本性的权衡：更好的分类器（更鲁棒）可能拥有一个效果更差的 OOD 检测器（基于激活的方法）。

5 自然漂移的局限性 (Natural Shift Analysis)

作为边界测试，我们评估了 **自然漂移** (CIFAR-10.1 / CIFAR-10.2)。

- **结果**: AUROC $\approx$ **0.563**, FPR@95 $\approx$ **0.93**。
- **结论**: NAC 无法区分这些自然漂移与原始训练数据。这证实了 NAC 检测的是“异常性”而非微妙的“分布漂移”。

6 需求完成度自检 (Progress Checklist)

我们将当前的进展与基于 Xiyue 邮件要求的项目路线图进行对照：

需求任务	实施状态	完成度
指标: NAC (OpenOOD)	已集成官方 v1.5 并配备自定义架构垫片。	**已完成**
对抗与腐蚀扰动	AutoAttack (L_∞, L_2) + Advex-UAR (Elastic, Fog, Snow, Gabor, JPEG) 全面扫测。	**已完成**
OOD Corruption	OODRobustBench (CIFAR-10-C) 全量扫测（75 种变体）已完成。	**已完成**
分析范围	纯净 / 单一 / 组合 / 顺序效应 (Phase 3) 分析完成。	**已完成**
几何变换 (DeepG)	基础设施已编译 (libgeometric.so)；旋转/平移已测试。高级形变待大范围扫测。	**部分完成**
基准模型 (RobustBench)	CIFAR-10 (ResNet18, WRN-28-10) 已完成。ImageNet/ViT 因资源/权限待定。	**部分完成**

Table 2: 实际交付组件与导师要求的对照表。

7 结论与未来展望 (Conclusion & Future Capabilities)

我们已经成功建立了一个严谨、数学一致的基准测试流水线。

1. **工程层面**: 中间件、DeepG 集成和架构垫片已全面运行。
2. **科学层面**: 通过大量的 CIFAR-10-C 扫测，我们将 NAC 定性为“结构异常检测器”，并量化了其边界。
3. **下一步**: 当前的基础设施已准备好，一旦数据和配置就绪，即可启动 ViT 与集成 NAC 的实验。