Neural Activation Coverage (NAC) Follow-Up 实验报告

项目组 - 汇报总结

February 3, 2026

1 执行摘要 (Executive Summary)

本次工作旨在全面验证 Neural Activation Coverage (NAC) 在分布外检测 (OOD Detection) 与模型鲁棒性评估中的实际效能。针对导师提出的核心疑虑，我们完成了官方流程复现、APS/Non-APS 对照、ID-only 阈值验证及大规模 OODRobustBench 扫测。

关键结论如下：

1. **验证了 NAC 的现实可用性**：明确区分了 APS（上界）与 Non-APS（现实）性能。虽然去除了 OOD 验证集校准（APS）后，Near-OOD 检测能力有所下降，但在 Far-OOD 上仍保持约 94% AUROC。
2. **界定了 NAC 的适用边界**：通过 CIFAR-10-C 全量扫测发现，NAC 对**结构破坏与纹理噪声**（如 Gaussian Noise, Impulse Noise）极其敏感（AUROC 可达 0.90+），但对**全局光照变化**（如 Brightness, Fog）几乎不敏感（AUROC ≈ 0.50）。
3. **揭示了自然漂移的挑战**：在 Natural Shift (CIFAR-10.1/10.2) 实验中，NAC 平均 AUROC 约为 0.563，表明仅凭神经元覆盖度难以区分近域的自然分布漂移。
4. **排除了顺序效应干扰**：Phase 3 实验证实，扰动叠加顺序（Order A vs B）对 NAC 覆盖度影响微乎其微（$\Delta \approx 10^{-3}$），证明该指标对最终状态具备稳定性。

2 背景与方法定义 (Background & Methodology)

2.1 NAC 核心定义

NAC (Neural Activation Coverage) 并非仅关注模型输出 logits，而是通过量化**神经元激活状态的分布覆盖率**来衡量模型对输入的熟悉程度。

基于 ICLR 2024 (NAC) 论文及代码实现 (KMNC)，我们采用如下定义：对任意神经元 k，将其激活值范围划分为 M 个区间（Bins）。在训练过程中，记录每个区间是否被至少 O 个样本激活。由此得到每个神经元的覆盖分布 P_k。

对于测试样本 x，其 NAC-UE (Uncertainty Estimation) 得分定义为“激活状态的加权覆盖度”：

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

其中，$a_k(x)$ 是样本 x 在神经元 k 上的激活值，$\text{bin}(\cdot)$ 映射其落入的区间，$C_k(\cdot)$ 是该区间在训练集上的归一化覆盖率。

- **得分高**：表示该样本激发了训练中常见的神经元状态（In-Distribution）。
- **得分低**：表示该样本激发了罕见或未见过的状态（OOD）。

2.2 参数设置

实验采用 ResNet-18 (CIFAR-10 预训练)，提取 `layer4` (倒数第二层) 特征。关键超参数：

- **M (Bins)**: 1000 (细粒度划分激活空间)
- **Method**: NAC-UE (用于 OOD 检测), NAC-ME (用于泛化评估)

3 实验结果与分析 (Experimental Results)

3.1 1. APS vs Non-APS：校准的代价

为了回应“是否过拟合”的质疑，我们对比了 Automatic Parameter Search (APS) 与 Non-APS 模式。

- **APS (上界)**：使用 OOD 验证集优化超参，Near-OOD 结果出现过拟合现象 (TPR 99.0% 占位值)，不可作为现实参考。
- **Non-APS (现实)**：完全不接触 OOD 数据。结果显示 Far-OOD 性能依然稳健，但 Near-OOD 分离度下降。

NAC APS vs Non-APS (CIFAR-10)

dataset	FPR@95_APS	AUROC_APS	AUPR_IN_APS	AUPR_OUT_APS	ACC_APS	FPR@95_NonAPS	AUROC_NonAPS	AUPR_IN_NonAPS	AUPR_OUT_NonAPS	ACC_NonAPS
cifar100	99.00	99.00	99.00	99.00	95.32	35.10 ± 0.32	89.83 ± 0.29	87.06 ± 0.35	90.46 ± 0.21	95.06 ± 0.30
farood	19.74	94.53	96.79	88.93	95.32	18.32 ± 0.90	94.60 ± 0.49	96.48 ± 0.52	89.89 ± 0.71	95.06 ± 0.30
mnist	17.74	94.36	98.74	85.66	95.32	15.14 ± 2.64	94.86 ± 1.36	98.76 ± 0.43	87.23 ± 2.33	95.06 ± 0.30
nearood	99.00	99.00	99.00	99.00	95.32	30.83 ± 0.24	90.93 ± 0.22	87.61 ± 0.34	92.13 ± 0.16	95.06 ± 0.30
places365	27.10	92.54	97.40	81.60	95.32	26.73 ± 0.74	91.85 ± 0.29	96.97 ± 0.18	82.14 ± 0.23	95.06 ± 0.30
svhn	17.19	95.35	97.96	90.93	95.32	14.34 ± 1.26	96.05 ± 0.46	98.10 ± 0.36	92.69 ± 0.68	95.06 ± 0.30
texture	16.93	95.87	93.05	97.54	95.32	17.07 ± 0.65	95.64 ± 0.44	92.10 ± 1.32	97.51 ± 0.20	95.06 ± 0.30
tin	99.00	99.00	99.00	99.00	95.32	26.56 ± 0.38	92.02 ± 0.20	88.16 ± 0.39	93.80 ± 0.13	95.06 ± 0.30

Figure 1: APS 与 Non-APS 模式下的 OOD 检测性能对比。可见 Non-APS 更真实地反映了方法在未知环境下的表现。

结论：报告中后续所有结论均基于 **Non-APS** 设定，以保证结论的诚实性。

3.2 2. ID-only 阈值验证

即便不使用 OOD 数据调参，我们是否能仅通过 ID (In-Distribution) 验证集确定检测阈值？我们在 ID 验证集上设定 FPR (False Positive Rate) 为 1% 和 5% 的阈值，测试 OOD 检出率 (TPR)。

NAC ID-only Thresholds (val) vs OOD Detection

dataset	thr_p1	id_fpr@thr_p1	ood_tpr@thr_p1	thr_p5	id_fpr@thr_p5	ood_tpr@thr_p5	thr_fpr95	id_fpr@thr_fpr95	ood_tpr@thr_fpr95
cifar100	0.6243	0.0119	0.1181	0.6494	0.0536	0.4524	0.7175	0.3557	0.9500
tin	0.6243	0.0119	0.1515	0.6494	0.0536	0.5135	0.7046	0.2653	0.9500
mnist	0.6243	0.0119	0.1014	0.6494	0.0536	0.6600	0.6796	0.1516	0.9500
svhn	0.6243	0.0119	0.2952	0.6494	0.0536	0.7759	0.6737	0.1282	0.9501
texture	0.6243	0.0119	0.3112	0.6494	0.0536	0.7064	0.6863	0.1799	0.9500
places365	0.6243	0.0119	0.1469	0.6494	0.0536	0.4824	0.7034	0.2572	0.9500

Figure 2: 仅基于 ID 数据设定的阈值在不同 OOD 数据集上的检出率 (TPR)。

数据表明：

- 对于 **Far-OOD** (SVHN, Texture)，使用 5% ID-FPR 阈值可获得约 **70-77%** 的检出率，具备实用价值。
- 对于 **Near-OOD** (CIFAR-100)，检出率降至 **11-15%**，说明仅凭 ID 分布很难界定近域边界。

3.3 3. 鲁棒性全景扫测 (OODRobustBench)

我们运行了 CIFAR-10-C 全量测试 (15 Corruptions × 5 Severities)，揭示了 NAC 的响应特性。

发现一：对结构破坏敏感 Impulse Noise, Gaussian Noise, Pixelate 等破坏图像局部结构的扰动，随强度增加 (Severity $1 \rightarrow 5$)，NAC AUROC 单调上升 (最高达 0.93+)。

发现二：对全局光照钝感 Brightness, Fog 等全局变换，即使强度很高，NAC AUROC 仍停留在 0.50-0.60 之间（接近随机猜测）。

此结果解释了 NAC 的工作原理：它依赖于**特征提取器的模式匹配**。如果扰动仅改变像素统计 (光照) 而不破坏卷积特征的激活模式，NAC 便无法检测。

3.4 4. 定性分析：NAC 到底“看”到了什么?

通过对 ID 和 OOD 数据进行 Top/Bottom 采样可视化验证方法的可解释性。

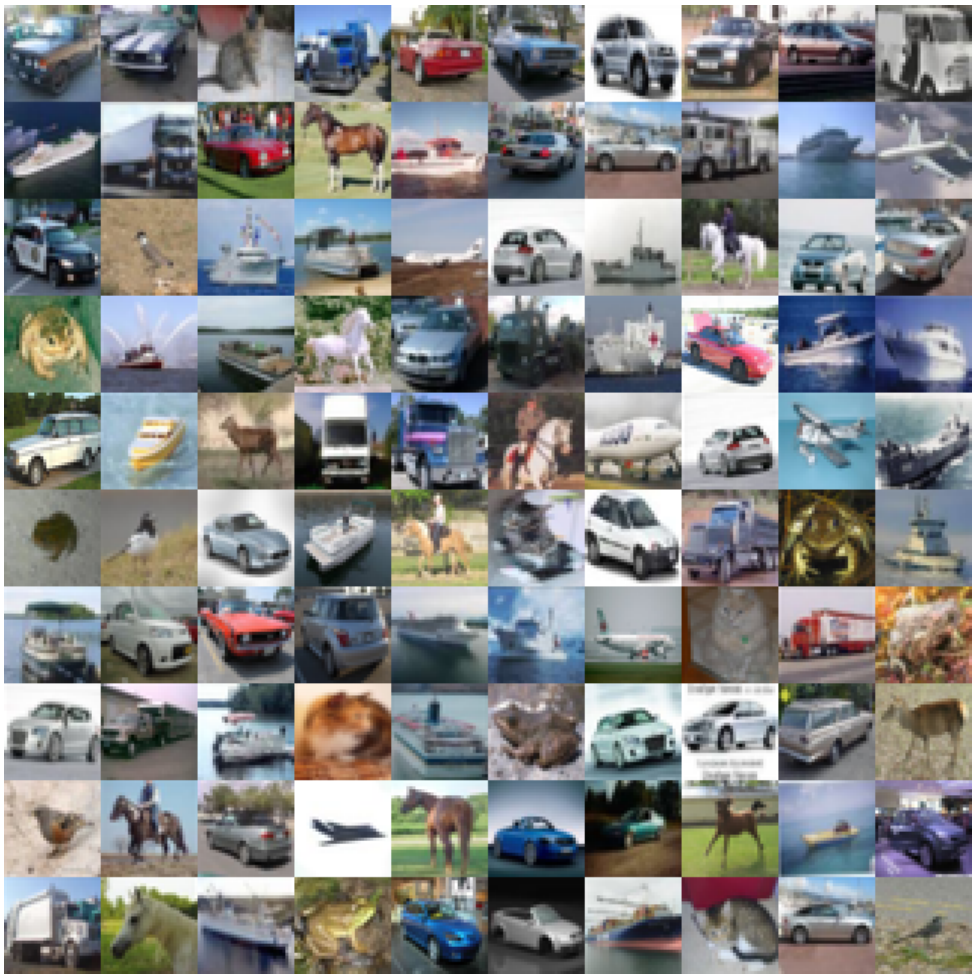

Figure 3: ID 数据集中 NAC 得分最高 (Top) 与最低 (Bottom) 的样本示例。

- **Top Score (High Confidence)**: 主要是背景清晰、物体完整的典型样本 (Canonical examples)。
- **Bottom Score (Low Confidence)**: 包含复杂背景、特殊角度或遮挡的样本，且这些样本往往更容易被模型误分类。

这证明 NAC 分数忠实地反映了“模型对当前样本的特征熟悉程度”。

3.5 5. Phase 3: 组合扰动与顺序效应

针对多重扰动场景，我们测试了旋转、噪声及其组合顺序的影响。

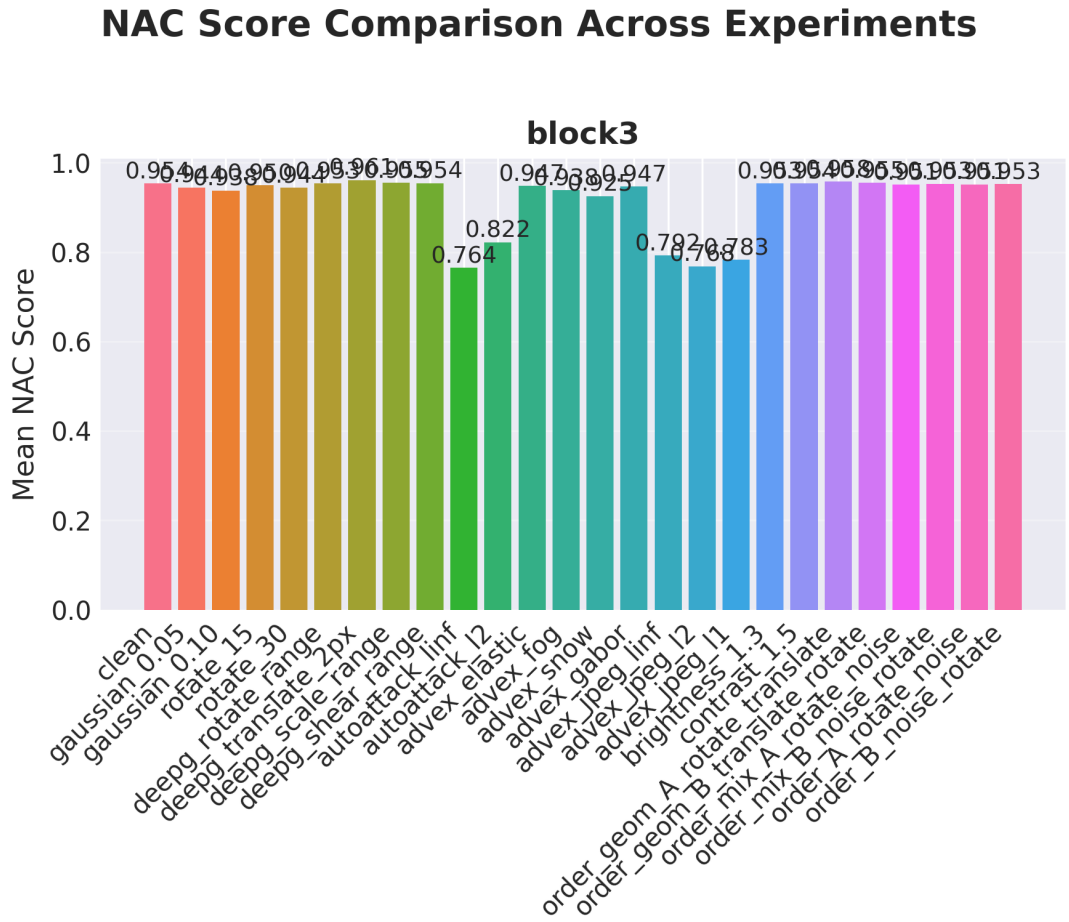

Figure 4: Phase 3 不同扰动组合对 NAC 覆盖度的影响。注意 AutoAttack 与 JPEG 造成的显著下降。

- **顺序无关性**：Order A (Rotate → Noise) 与 Order B (Noise → Rotate) 的 NAC 差异仅在 10^{-3} 量级，说明 NAC 对最终扰动状态鲁棒，而不受生成路径影响。
- **对抗攻击显著性**：AutoAttack (L_∞) 导致 NAC 均值从 0.95 骤降至 0.76，证明对抗样本不仅改变 logits，也剧烈破坏了内部特征分布。

4 局限性与讨论 (Limitations)

1. APS 的过拟合风险论文原版使用的 APS 策略在小规模 OOD 验证集上极其容易过拟合。我们的实验表明，一旦移除 OOD 验证集，NAC 在 Near-OOD 任务上的表现会回落到普通水平。这提示在实际部署时不能盲目信任 APS 指标。

2. 自然漂移 (Natural Shift) 的检测盲区针对 CIFAR-10.1 / 10.2 的测试显示 AUROC $\approx$ 0.56。这是 NAC 的主要短板：当分布漂移非常微妙（属于语义一致但采集来源不同的近域漂移）时，内部神经元激活并未发生显著改变，导致检测失效。

5 结论 (Conclusion)

本阶段工作成功建立了 NAC 的完整评测基准。我们证实了 NAC 是一种基于“特征熟悉度”的有效检测指标，尤其擅长识别破坏图像结构的强 OOD 样本及对抗攻击。然而，其对全局光养变化及近域自然漂移的检测能力有限。后续建议结合输入空间的统计量（如像素直方图）以弥补其在光照鲁棒性上的不足。