矩阵理论应用报告：
基于奇异值分解（SVD）的深度学习后门攻击检测

姓名：孙诚睿

学号：2025140951

课程：矩阵理论与方法

2025 年 12 月 19 日

摘要

随着深度学习在安全敏感领域的广泛应用，数据投毒攻击，特别是后门攻击，成为了严重的安全威胁。后门攻击通过在训练集中注入少量带有特定触发器的样本，使模型在测试时对特定输入产生误判。本文基于发表在 NeurIPS 2018 的工作 [1]，深入探讨了一种名为“谱签名（Spectral Signatures）”的检测方法。该方法巧妙地利用了矩阵理论中的奇异值分解（SVD）和协方差矩阵谱分析技术。研究表明，尽管中毒数据在原始像素空间难以区分，但在高层特征表示的协方差矩阵中，它们会引起显著的谱分布变化。本文详细梳理了该方法的数学原理，推导了 SVD 在分离混合分布时的理论界限，并展示了其在 ResNet 模型上的应用效果。

关键词：奇异值分解；协方差矩阵；后门攻击；谱签名；鲁棒统计

目录

1 引言

深度学习模型的训练往往依赖于海量数据，这使得攻击者有机会通过操纵训练集来植入后门。Gu 等人 [2] 展示了后门攻击的危险性：攻击者只需注入少量带有特定图案（如角落里的像素点）的图片，就能让模型在识别正常图片时表现良好，但一旦识别到该图案，就将其错误分类为攻击者指定的目标。

传统的异常检测方法在原始像素空间往往失效，因为后门触发器通常极其微小。然而，Tran 等人 [1] 提出了一种基于“谱签名”的新视角。该方法的核心洞察在于：为了让神经网络学会识别微小的后门触发器，模型必须在内部特征空间（Feature Representation）放大该信号。这种放大效应会在特征数据的协方差矩阵中留下数学痕迹，使得我们可以利用奇异值分解（SVD）来检测并移除中毒样本。

2 预备知识：矩阵理论基础

本节回顾支撑该检测方法的核心矩阵理论知识。

2.1 奇异值分解 (SVD)

对于任意实矩阵 $M \in \mathbb{R}^{n\times d}$，奇异值分解保证存在正交矩阵 $U \in \mathbb{R}^{n\times n}$ 和 $V \in \mathbb{R}^{d\times d}$，以及对角矩阵 $\Sigma \in \mathbb{R}^{n\times d}$，使得：

$$M = U\Sigma V^T \tag{1}$$

其中，Σ 的对角线元素 $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$ 为奇异值。V 的列向量称为右奇异向量。在统计学意义上，若 M 是中心化后的数据矩阵，则 V 的第一个列向量 v_1 指示了数据方差最大的方向（第一主成分）。

2.2 协方差矩阵与谱分析

给定随机向量 $X \in \mathbb{R}^d$，其协方差矩阵定义为 $\Sigma = \mathbb{E}[(X-\mu)(X-\mu)^T]$。协方差矩阵是实对称半正定矩阵，其特征分解（Spectral Decomposition）揭示了数据分布的几何结构。特征值的大小对应了分布在该特征向量方向上的延展程度。

3 基于 SVD 的后门检测方法论

3.1 威胁模型

我们考虑 [2] 描述的典型威胁模型：攻击者向标签为 y 的训练数据中注入了一部分带有后门的样本。

- 设 D 为干净数据的分布，均值为 μ_D，协方差为 Σ_D。
- 设 W 为中毒数据的分布（通常来自另一类别的图像叠加了触发器），均值为 μ_W，协方差为 Σ_W。
- 实际观测到的训练数据分布 F 是两者的混合：

$$F = (1-\epsilon)D + \epsilon W \tag{2}$$

其中 ϵ 是投毒比例（例如 $\epsilon = 5\%$）。

3.2 检测算法

基于鲁棒统计学的思想 [3]，当 μ_D 和 μ_W 在特征空间分离度足够大时，可以通过 SVD 识别异常。算法流程如算法 1 所示 [1]。

Algorithm 1 基于谱签名的后门移除算法

Require: 训练集 $\mathbb{D}_{train}$，特征提取器 $\mathcal{R}$（如 ResNet 的倒数第二层）

初始化中毒样本集合 $S \leftarrow \{\}$
for 每一个标签 y **do**
    获取标签 y 下的所有样本 $x_1, \ldots, x_n$
    计算特征表示：$R_i = \mathcal{R}(x_i)$
    计算特征均值：$\hat{R} = \frac{1}{n}\sum_{i=1}^{n} R_i$
    构建中心化矩阵：$M = [R_1 - \hat{R}, \ldots, R_n - \hat{R}]^T$
    执行奇异值分解 (SVD)：计算 M 的最大右奇异向量 v
    计算每个样本在主方向上的异常得分：$\tau_i = ((R_i - \hat{R}) \cdot v)^2$
    移除得分最高的 1.5ϵ 比例的样本，加入集合 S
end for
在过滤后的数据集 $\mathbb{D}_{train} \setminus S$ 上重新训练模型

4 理论分析：为什么 SVD 有效

本节利用矩阵理论证明，为什么混合分布协方差矩阵的主特征向量（Top Eigenvector）会指向中毒数据的方向。

4.1 协方差矩阵的秩-1 更新

考虑混合分布 F 的协方差矩阵 Σ_F。根据方差分解定理，我们可以将其展开为 [1]：

$$\Sigma_F = (1-\epsilon)\Sigma_D + \epsilon\Sigma_W + \epsilon(1-\epsilon)(\mu_D - \mu_W)(\mu_D - \mu_W)^T \tag{3}$$

令 $\Delta = \mu_D - \mu_W$ 为均值漂移向量。上式最后一项 $\epsilon(1-\epsilon)\Delta\Delta^T$ 是一个**秩为 1 的矩阵**。

从矩阵分析的角度来看，当攻击导致特征空间的均值发生显著偏移（即 $\|\Delta\|_2$ 很大）时，这一项秩-1 矩阵将在谱范数意义下主导 Σ_F。根据矩阵微扰理论，Σ_F 的最大特征值对应的特征向量 v，将与 Δ 的方向高度对齐。

4.2 谱分离性引理

Tran 等人 [1] 提出了谱分离（Spectrally Separable）的理论界限。假设干净数据和中毒数据的协方差有上界 $\Sigma_D, \Sigma_W \preceq \sigma^2 I$。

引理 1（主特征向量的相关性 [1]）. *如果均值差异满足 $\|\mu_D - \mu_W\|_2^2 \geq \frac{6\sigma^2}{\epsilon}$，则混合分布的主特征向量 v 与均值差 Δ 满足：*

$$\langle v, \Delta \rangle^2 \geq \frac{2\sigma^2}{\epsilon} \tag{4}$$

这意味着，如果我们把所有数据点投影到方向 v 上，中毒数据（来自分布 W）的投影值将显著偏离干净数据（来自分布 D）的投影值，从而可以通过简单的阈值切分将其移除。

5 实验验证

5.1 实验设置

论文在 CIFAR-10 数据集 [4] 上进行了验证，使用了标准的 ResNet 模型 [5]。攻击者在“飞机”类别的图片中加入微小扰动并标记为“鸟”。

5.2 SVD 分析结果

实验结果表明，在像素层面（Data Level），SVD 无法区分中毒样本。但在 ResNet 的倒数第二层特征（Representation Level）上：

1. **奇异值膨胀**：加入中毒数据后，特征协方差矩阵的第一奇异值显著增大（从 1194 增加到 1613 [1]）。

2. **分布分离**：计算样本在主奇异向量上的相关性得分，可以观察到干净样本和中毒样本呈现明显的双峰分布（Separation）。

利用该方法，研究者成功将后门攻击的成功率从 90% 以上降低到了 1% 以下，且未对正常分类准确率造成显著影响。

6 结论与个人思考

本文分析了利用 SVD 提取“谱签名”来防御深度学习后门攻击的机制。从矩阵理论的角度看，这本质上是一个**低秩扰动（Low-rank Perturbation）**问题：攻击者试图隐藏数据，但为了让模型学习到攻击模式，必然在特征空间引入强信号，这个强信号表现为协方差矩阵的一个主成分。

这种方法不仅展示了 SVD 在高维数据异常检测中的威力，也体现了矩阵分析工具在现代人工智能安全领域的实际应用价值。未来的研究方向可以考虑当攻击者试图压制特征方差（针对 PCA/SVD 的对抗攻击）时，如何利用更高阶的张量分解技术进行检测。

参考文献

[1] B. Tran, J. Li, and A. Mądry, “Spectral signatures in backdoor attacks,” in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 8000–8010, 2018.

[2] T. Gu, B. Dolan-Gavitt, and S. Garg, “Badnets: Identifying vulnerabilities in the machine learning model supply chain,” *arXiv preprint arXiv:1708.06733*, 2017.

[3] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart, “Being robust (in high dimensions) can be practical,” in *International Conference on Machine Learning (ICML)*, pp. 999–1008, 2017.

[4] A. Krizhevsky and G. Hinton, “Learning multiple layers of features from tiny images,” tech. rep., Citeseer, 2009.

[5] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 770–778, 2016.