arXiv:2305.01808v2 [cs.CV] 6 May 2023

用于类别划分和对抗图像检测的汉明相似度与图拉普拉斯矩阵

Huma Jamil[1*]、Yajing Liu[2*†]、Turgay Caglar[1],Christina M. Cole[2],
Nathaniel Blanchard[1],Christopher Peterson[2],Michael Kirby[2]1科罗拉多州立大学计算机科学系，美国科罗拉多州柯林斯堡2科罗拉多州立大学数学系，美国科罗拉多州柯林斯堡 电子邮件：{*huma97, yajing.liu, turguy.caglar,christina.rigsby, nblancha*}*@colostate.edu*、{*christopher2.peterson, michael.kirby*}*@colostate.edu*

摘要

研究人员通常通过检查网络一个或多个层的激活输出，来研究神经网络的表征。在本文中，我们探讨*ReLU*激活模式（编码为位向量）在理解和解释神经网络行为方面的潜力。我们利用表征差异矩阵（*RDM*）研究深度神经网络嵌入空间中数据的一致性。从网络的每一层提取并利用位向量，构建图像之间的相似度分数。根据这些相似度分数，我们为一组来自 *2* 个类别的图像构建相似度矩阵。随后，我们对相应的拉普拉斯矩阵应用费德勒划分，以分离这些类别。结果表明，通过位向量表征，网络能够持续提升类别可检测性；最后一个 *ReLU* 层实现了超过 *95%* 的分离准确率。此外，我们还证明位向量有助于对抗图像检测：使用简单的分类器区分对抗图像与非对抗图像，准确率同样超过 *95%*。

1. 引言

近乎与神经网络在各类基准测试中占据主导地位的时间一样长，研究者一直在探索理解和解释其行为的方法。近期关于何时该信任、何时不该信任 ChatGPT 输出的争议，完美地概括了将网络视为“黑箱”所带来的问题 [1] ——解释网络行为与信任息息相关，也有助于更好地理解如何进一步提升性能。实现这一目标的一种方法是考察网络内部低维嵌入空间中的输入数据的一致性，正如表征学习的开创性研究所倡导的那样 [4]。理想情况下，这些中间嵌入能够以这样的方式表示数据：可以识别并区分预先设定的相似性（即相同的类别标签）与预先设定的差异性（即不同的类别标签）。从几何角度来看，如果被判定为相似的数据映射到嵌入空间中的相近点，而被判定为不相似的数据映射到相距较远的点，就会呈现出这种特征。如此一来，区分相似数据与不相似数据便易如反掌。

分析如何利用这些嵌入空间及其有效性，有助于深入了解网络在数据流经网络时如何处理和表征数据。进而，这些信息可用于提升神经网络的性能，并促进对其结果的解读。表征差异矩阵（RDM）是一种有用的工具，可用于可视化这些嵌入空间内部的一致性程度——RDM量化网络表征与相似刺激及不相似刺激之间的（不）相似性。通过检查RDM，我们可以识别网络对输入数据的表征中的模式或异常，从而诊断并解决模型性能或可解释性方面的问题 [16]。我们通过一种新颖的分析扩展这些RDM：将费德勒划分扩展至被视为加权图的不相似性矩阵。费德勒划分 [9, 10]，（也称为谱划分）是图论中的一种技术，它根据图的拉普拉斯矩阵的特征值和特征向量，将图划分为两个或更多子图。

通常，在神经网络模式分析中，会利用神经网络某一层的激活来揭示并概括输入 [24]的有用统计信息。然而，近期研究表明，较粗略地利用由修正线性

*贡献相同。
†通讯作者。

单元（ReLU）在早期层中的激活模式与对抗图像有着有意义的关联 [14], ，这启发我们使用表征差异矩阵进一步研究位向量的表征。在本研究中，我们考察了费德勒划分的有效性，该划分是根据由神经网络某一层中的位向量构建的相似度矩阵的拉普拉斯矩阵推导而来。换言之，我们计算了与网络某一层节点中的 ReLU 激活模式相关的特定拉普拉斯矩阵的费德勒特征向量。具体而言，我们使用费德勒划分考察了两个类别在每个 ReLU 层上的表征差异矩阵。由于深度神经网络某一层中的节点数量可能相当庞大（在我们的案例中为 10 万至 80 万），我们假设分类任务并不真正需要大量比特。为说明这一点，我们识别并提取了 1000 个最重要的比特，最终据此构建了表征差异矩阵。

我们的结果表明，随着数据经过更深的 ReLU 层，数据中开始显现出不同的分区，各类别的分类准确率总体上也有所提升。研究中的一项重要观察是，各层对待数据的方式并不相同；有些层似乎专注于提升区分一小部分图像的能力。RDM 分析揭示了网络如何编码有关输入数据的信息，以及如何提取用于分类的关键特征，从而提升其透明度和可解释性。

基于上述分析，我们随后开展了实证实验，以评估位向量诱导的分区在区分对抗图像与非对抗图像方面的有效性。先前研究表明，在深度神经网络的潜在层（倒数第二层的激活）嵌入中，对抗图像与非对抗图像是线性可分的 [12]。尽管位向量提供的输入数据信息较为粗略，我们发现，神经网络最后一个 ReLU 层的位向量也可用于有效地区分非对抗图像与对抗图像。事实上，我们使用更少的特征，取得了至少不逊于潜在层嵌入的结果。总体而言，我们的结果表明，深度神经网络中的位向量为高效识别对抗图像提供了一种有前景的方法，并有望进一步提升性能。

本文的贡献如下：

- 我们证明，只需从可用的 10 万至 80 万个比特中抽取 1000 个进行比特向量子采样，就能分离二分类数据。

- 我们提出一种新颖的思路：利用根据位向量相似度度量构建的拉普拉斯矩阵进行费德勒划分，实现超过 95% 的图像类别识别准确率。

- 我们展示了，利用线性支持向量机（SVM），位向量表示能够以超过 95% 的准确率区分对抗图像与非对抗图像。

本文结构如下：第 2节全面回顾相关文献，包括费德勒划分、对抗检测和网络可解释性。第 3节概述本研究所依据的相关定义。第 4节介绍 Fiedler 向量算法，并阐述其在利用 ResNet-50 位向量对图像进行二分类和双超类分类中的实际应用。第 5节说明，ResNet-50 位向量的线性可分性可用于区分非对抗图像与对抗图像。最后，第 6 节给出总结性论述。

2. 相关工作

2.1. 费德勒划分

费德勒向量是加权图的拉普拉斯矩阵中与最小非零特征值相对应的特征向量，是图论中的一项基础工具，可用于揭示通过成对相似性表示的复杂系统中的潜在子结构。它的各个分量可用于将图的顶点划分为两个互不相交的子集，同时使子集之间边的权重总和最小化，并使两个子集内部边的总权重最大化。费德勒向量算法（又称费德勒划分）已成功应用于众多现实问题：识别纹理或颜色属性相似的区域 [28],识别具有相似特征或功能的群体 [23], 以及突出社交网络中的重要节点或社群 [18]。它也是机器学习和数据分析中一种成熟的技术，可用于发现并解读数据中的子模式 [8, 31]。费德勒划分的一项关键优势在于，它能够利用从网络中导出的低维特征空间，近期 Schleider 等人 [26] 和 Cao 等人的研究 [5]对此进行了展示。该特征空间通常通过对图拉普拉斯矩阵进行特征值分解获得，可实现高效且有效的数据聚类与可视化，因此成为众多应用领域的热门选择。在本研究中，我们将费德勒划分应用于表征差异矩阵，以开展新的分析。

2.2. 对抗检测

深度神经网络在图像分类方面效率很高，但可能会受到对抗攻击的欺骗 [30]。已有多种

针对不同弱点的攻击已被提出 [6, 7, 11, 20, 21]。为此，多项研究利用深度神经网络隐藏层和潜在层的特性来检测对抗图像。例如， [3] 提出了一种利用潜在层的方法， [12]并展示了这些层中激活所携带的有意义信息。为了检测对抗图像和分布外图像，[17] 利用隐藏层和潜在层的特征计算类别条件高斯分布和马氏距离。另一方面， [19] 则利用中间卷积层的输出提取统计数据，以帮助检测对抗图像。我们的方法与 [19] 类似，因为我们同样考察神经网络的内部运作。不过，我们采用了更基础的度量，具体而言，是使用由 ReLU 层激活模式中相对较小的一个子集生成的位向量，而非依赖潜在层的数值。

2.3. 网络可解释性

解释深度神经网络的功能通常需要理解隐藏层。利用中间层进行可视化 [32], 网络剖析 [33], 以及学习解耦表示 [25]是为实现这一目标而提出的少数方法之一。在我们的工作中，我们通过使用从中间 ReLU 层提取的位向量，为网络可解释性作出了贡献。在以往的文献中，汉明距离 [27, 29], 杰卡德距离 [2], 和余弦相似度 [15]等定量指标已被用于评估可解释方法的有效性。在我们的工作中，我们通过评估由数据通过 ReLU 层所确定的位向量之间的成对汉明距离，并基于这些距离构建拉普拉斯矩阵，计算了图像相似度。我们的结果揭示了神经网络在信息处理、特征提取和决策机制方面的规律，从而增强了网络的透明度和可解释性。

3. 定义

3.1. 位向量

考虑一个包含 ReLU 激活函数的前馈神经网络。假设该网络的 i^{th} 层包含 h_i 个 ReLU 节点。对于给定的输入 $x \in \mathbb{R}^m$，我们将其在 i^{th} 层这些 ReLU 节点处的输出表示为 $o^i(x) = \left[o_1^i(x) \ \ldots \ o_{h_i}^i(x)\right]^\top$。我们将 x 在 i^{th} 层的位向量定义为 $s^i(x) = \left[s_1^i(x) \ \ldots \ s_{h_i}^i(x)\right]^\top$，其中

$$s_j^i(x) \coloneqq \begin{cases} 1 & \text{if } o_j^i(x) > 0, \\ 0 & \text{if } o_j^i(x) = 0. \end{cases} \tag{1}$$

因此，位向量是输入 x 的函数：对于在 x处激活 ReLU 函数的节点赋值为 1，否则赋值为 0。通常，神经网络中不同输入对应的位向量会呈现出不同的激活模式，因为这些输入会在网络中传递，并经过 ReLU 层。

3.2. 表征差异矩阵

对于给定的神经网络 N，其输入集为 $X = \{x_1, \ldots, x_n\}$，令 $\{o^i(x_1) \ldots, o^i(x_n)\}$ 表示 N中 i^{th} 隐藏层的 X 的输出，我们将层 i 的表征差异矩阵（RDM）定义为:

$$RDM\,(X, N) = \left(d_{jk}^i\right)_{0 \le j,k \le n}, \tag{2}$$

其中， d_{jk}^i 是衡量 o^i（x_j）与o^i（x_k）之间不相似程度的指标。例如， d^i 可以是一个距离函数。本质上，表征差异矩阵（RDMs）通过概括神经网络对不同输入的神经响应之间的相似性和差异，为捕捉神经网络的内部表征提供了一种方法。

要定义 RDM，需要采用一种差异度量。不同的度量方式会得到不同的 RDM。在我们的研究中，我们对d^i使用了汉明距离和余弦距离。汉明距离是一种常用于比较两个等长二进制字符串的度量，它表示对应条目不同的位置数量。构建输入集合 X 在第 i层的 RDM 时，我们将汉明距离作为距离度量之一。两个向量之间的余弦距离定义为 $1 - \cos(\theta)$，其中 θ 表示两个向量之间的夹角。除了汉明距离，我们还将余弦距离用作神经网络潜在层中嵌入的差异度量。

3.3. 特征选择

特征选择是一种通过选择能够较好地表示数据的特征来降低数据维度的技术。在我们的实验中，我们使用 Python 库中的 *SelectKBest* 方法来降低位向量的维度（即从位向量中选择一个比特子集）。从数学上讲，设 $X = \{x_1, \ldots, x_n\}$为一组 n 输入数据点，$Y = \{y_1, \ldots, y_n\}$ 为目标标签。设 $f_i(X, Y)$为一个评分函数，用于衡量 i^{th} 特征在预测目标变量时的重要性。然后，*SelectKBest* 函数会选择得分最高的k 个特征，并返回与评分函数 $f_i(X, Y)$的 k 个最高分对应的索引集合 $S_k = \{i_1, i_2, ..., i_k\}$。在我们的实验中，我们使用卡方统计量作为评分函数 f_i。

4. 基于费德勒向量的分类

在本节中，我们将最初作为谱聚类理论框架提出的 Fiedler 向量算法应用于 [9, 10], ImageNet 数据集中的图像组。该算法基于从名为 ResNet-50 [13]的神经网络的各层收集的位向量。在深入探讨该算法在 ResNet-50 位向量上的应用之前，我们先回顾费德勒方法及其相关定义。

4.1. Fiedler 向量算法

考虑一个加权图 $G = (V, E, W)$，其中 $V = \{1, \ldots, n\}$ 表示顶点集合， E 表示（无序）边的集合，而 W: $E \to \mathbb{R}_{\geq 0}$ 表示 E中元素的权重。因此， E 中的每个元素都是一对无序顶点，并且关联一个非负权重。图 G 的加权邻接矩阵 $A \in \mathbb{R}^{n \times n}$ 定义规则如下： $A_{ij} = 0$ 若 $\{i, j\} \notin E$ ，且$A_{ij} = W(\{i, j\})$ 若 $\{i, j\} \in E$。度矩阵 D是一个对角矩阵，其$D_{i,\ i}$ 等于 A中第 i^{th} 行各项之和。它也可以表示为 $D = \text{diag}(\{\text{v35}\})$，其中 e 是一个所有元素均为 1 的列向量。于是，图 G 的拉普拉斯矩阵可表示如下:

定义 4.1 给定一个无向加权图 $G = (V、 E、 W)$，其拉普拉斯矩阵 L 定义为 $L = D - A$，其中 A 是 G 的加权邻接矩阵，而 D 是 G的度矩阵。

文献 [9], 表明，拉普拉斯矩阵 L 是对称的、半正定的，并且 0 是其特征值。0 作为特征值的重数等于 G 的连通分量数。因此，如果 G 是一个连通图，则 0 是重数为 1 的特征值。对于连通图， L 的第二小特征值称为图 G的代数连通度 。它提供了有关该图被分隔成两个分量的难易程度的信息。在文献 [10], 中，提出了与 L的第二小特征值相对应的特征向量（现称为费德勒向量），将其作为把 G “最优地”拆分成两个连接更紧密的连通分量的一种方法。下面的定理引出了用于划分 G顶点的 Fiedler 向量算法。

定理 4.1 设 $G = (V, E, W)$ 为一个带权连通图，其加权邻接矩阵为 A ，拉普拉斯矩阵为 L。设 v_2 为对应于 L的第二小特征值的特征向量。费德勒向量划分为:

$$C_1 = \{i \in N : v_2(i) < 0\} \text{ and } C_2 = \{i \in N : v_2(i) > 0\}.$$

满足 $v_2(j) = 0$ 条件的顶点可以任意归入任一类别。

上述 Fiedler 向量算法可以通过利用与最小的 l 非零特征值相对应的前 l 个特征向量中各项的符号模式，将 $2^l (l = 2, 3, \ldots)$ 个簇进行划分。

4.2. 实验设置

在接下来的章节中，我们将应用 Fiedler 向量算法，在 ImageNet-1K 数据集上执行分类任务。该数据集包含 1000 个图像类别，每个类别由 1300 张训练图像和 50 张验证图像组成。ResNet-50 指一种特定的、经过训练的深度前馈卷积 ReLU 神经网络，该网络使用 ImageNet-1K 数据集中的 130 万张训练图像进行训练 [13]。为了获取图像的位向量，我们将图像输入 ResNet-50，并根据每个 ReLU 层的输出提取位向量。由此共得到 17 个位向量（因为 ResNet-50 中有 17 层使用了 ReLU 激活函数）。该神经网络架构首先包含四个 ReLU 层，每层有 802,816 个节点；接下来是四个节点数减半的层、六个节点数为其四分之一的层，以及最后三个节点数为初始数量八分之一的层。

给定图像集在每个层的加权邻接矩阵 A 是通过计算矩阵 $\mathbf{1} - H$ 获得的，其中 $\mathbf{1}$ 是全 1 矩阵，而 H 是归一化汉明距离矩阵。换言之，在给定层中， H 的条目 $H_{i,j}$ 定义为图像 i 和 j 的位向量中不同条目的数量除以位向量的长度。

为了考虑每一层中大量的节点及其重要程度各不相同这一情况，我们采用了第3.3节所述的特征选择技术，以识别位向量中最关键的位。随后，我们使用仅包含最关键位的短得多的位向量为每张图像标注。最后，我们对上一段所述的加权邻接矩阵应用了费德勒向量算法。

在分类过程中，我们使用所选类别的训练数据来确定 1000 个最重要的特征（比特），然后分别用它们计算所选类别训练数据和测试数据的拉普拉斯矩阵。最后，我们使用费德勒向量算法评估训练集和测试集上的分类率。

4.3. 二分类

我们为实验定义了四对不同的类别。前两对各由两个不同物种组成，具体为丁鲷与雷蛇，以及罗得西亚脊背犬与帝王蝶。后两对则属于鱼类和蛇类这两个大类的子类别。具体而言，第三对为丁鲷与金鱼，第四对为雷蛇与环颈蛇。

在初步实验中，我们从两个类别（丁鲷和雷蛇）的训练数据中提取了 1000 个显著特征，并使用验证集中的相同特征构建 RDM。我们研究了这两个类别的表征差异矩阵在 17 个 ReLU 层中的演变情况，如图1所示。我们还注意到一个有趣的模式：在初始层（第 2 至第 6 层）中，一个类别与另一个类别变得更加相似；但随后，随着不相似度值上升，两个类别逐渐分离。即使在初始层中，我们也观察到同一类别内的成对距离减小，呈现出聚类现象。在较后的层中，两个类别在 RDM 中明显分开，二者之间存在清晰的边界。有趣的是，在第 15 个 ReLU 层时，这种分离已经显而易见。

表1 展示了使用 Fiedler 向量算法计算的各层分类准确率以及各类别的准确率。结果与 RDM 图一致。值得注意的是，尽管整体分类率在连续两层之间可能下降，但某一类别的准确率却可能提高。我们还对其余三组类别对进行了类似实验，并绘制了 1、5、10 和 17 个 ReLU 层时的测试准确率（见图2）。我们的研究结果再次表明，层深度与分类准确率呈正相关，这意味着神经网络能够从数据中提取越来越有价值的信息，且越靠近网络架构的后部，这种能力越强。

图2 显示，三组类别对的表现相近，而第四组——雷蛇与环颈蛇——的表现稍差。这或许可以归因于雷蛇与环颈蛇之间细微的差异：由比特向量给出的粗略表征可能无法完全捕捉这些差异。

总之，我们的实验表明，位向量提供的粗略表征经过筛选后保留了 1000 个重要特征，包含了足够的数据相关信息，可用于神经网络分析。对基于位向量构建的拉普拉斯矩阵应用 Fiedler 向量算法，结果表明这是一种有效的图像分类方法，且分类率通常会随着层数加深而提高。

4.3.1 两个超类分类

费德勒向量在对两个不同类别进行分类方面展现出了良好效果，这促使我们研究其在划分更广泛类别方面的有效性，具体而言，即鱼类和蛇类超类，以及鱼类和猫类超类。我们的分析表明，费德勒向量算法能够有效地对两个超类进行分类。为确定最佳的 1000 个特征，我们选取了类别丁鲷和金鱼，以及雷蛇和环颈蛇的训练数据集。这些类别分别属于鱼类和蛇类超类。随后，我们将费德勒向量算法应用于这些超类的验证数据集，在最后三个 ReLU 层上分别获得了 94.5%、97.5% 和 99% 的测试准确率。

我们还针对鱼类（丁鲷和金鱼）与猫（波斯猫和埃及猫）这两组超类进行了类似的实验。在最后三个 ReLU 层中，测试准确率分别为 98%、99% 和 99.5%。图 3展示了这两组超类在最后一个 ReLU 层所得到的表征差异矩阵。

我们的研究结果表明，即使在较早的 ReLU 层中，位向量也能识别同一超类中各类别的共同特征。此外，Fiedler 向量算法能够有效区分两个不同的超类。

5. 对抗图像分析

费德勒划分在分离图像类别方面取得的成功促使我们探究位向量是否也能区分对抗图像与非对抗图像。在接下来的两个部分中，我们开展了实验，以分析不同的对抗攻击。首先，我们考察了位向量的表征差异矩阵与潜在层嵌入的表征差异矩阵之间的相关性。受该实验积极结果的鼓舞，我们在后续部分构建了一个线性支持向量机分类器。该支持向量机分类器基于位向量进行训练，并取得了有希望的结果，展示了使用位向量进行对抗图像检测的潜力。

5.1. ReLU 层与潜在层嵌入之间的 RDM 比较

在我们的研究中，我们旨在比较 Imagenet-1K 验证集中的对抗图像与非对抗图像的潜在层嵌入及其对应位向量之间的差异。为确保一致性，我们为 ReLU-17层位向量使用了 2048 个显著特征，这与潜在层嵌入的维度相对应。我们使用汉明距离计算位向量 RDM，并使用余弦距离计算潜在层 RDM，以构建我们的 RDM。图 4 如预期所示，潜在层 RDM 清晰地将两个类别区分开来。类似地，ReLU 层 RDM 也显示出对抗图像类别与非对抗图像类别之间的明显差异。两个 RDM 之间的皮尔逊相关系数为 0.620，表明这些表征之间存在正相关。

在下一节中，我们将探究位向量通过线性分类器区分这两类图像的潜力。

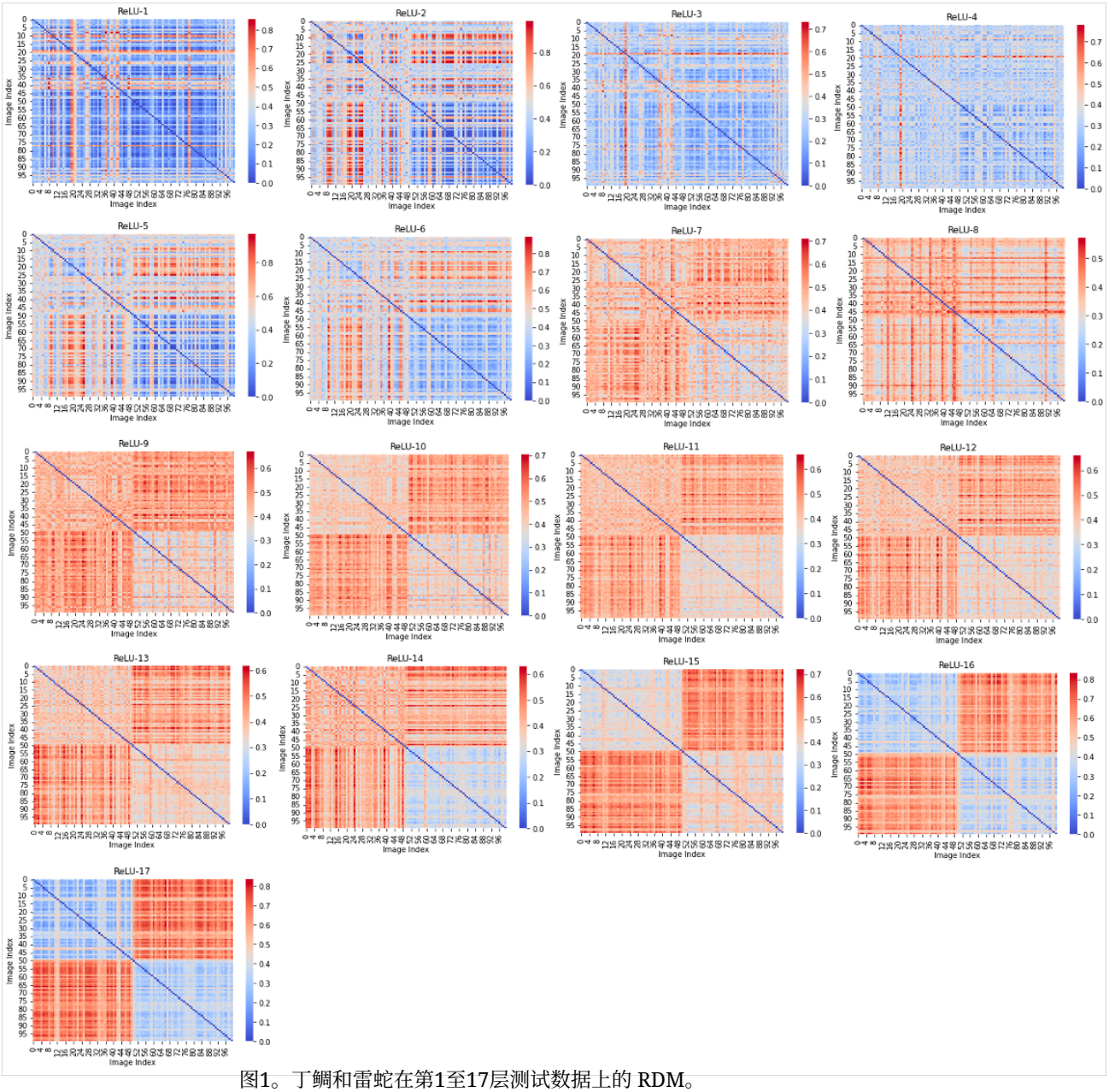

图1。丁鲷和雷蛇在第1至17层测试数据上的 RDM。

层编号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
准确率（%）	60	69	58	52	76	81	83	59	84	90	83	85	78	73	100	100	99
丁鲷（%）	24	54	22	4	62	62	82	18	74	86	66	72	56	46	100	100	98
雷蛇（%）	96	84	94	100	90	100	84	100	94	94	100	98	100	100	100	100	100

表1。使用费德勒向量分类器时，丁鲷和雷蛇在第1至17层的测试准确率（百分比）。

5.2. 线性可分性比较

深度神经网络的 ReLU 激活模式极大地简化了网络的运作方式，却包含有关输入数据的重要信息，正如 [14]中的研究结果所示。该研究指出——

表明，在位向量的 600 万个位中，仅需 5 位就足以区分大多数对抗图像和非对抗图像。据我们所知，文献中尚未普遍采用位向量来研究对抗图像。不过，[12]表明，线性

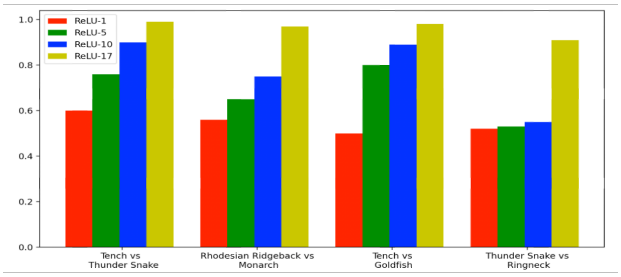

图 2. 使用 Fiedler 向量算法，在第 1、5、10、17 层上对四组类别对进行测试准确率评估。

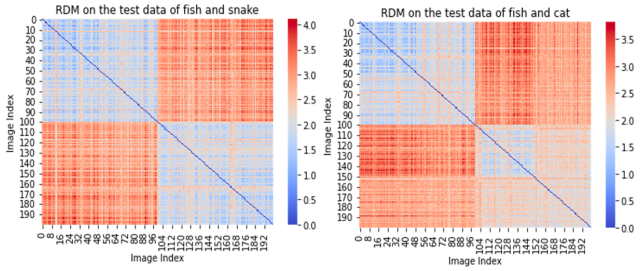

图 3. 使用 Fiedler 向量算法，在第 17 层上比较鱼类（丁鲷和金鱼）与蛇类（雷蛇和环颈蛇），以及鱼类（丁鲷和金鱼）与猫（波斯猫和埃及猫）的测试准确率。

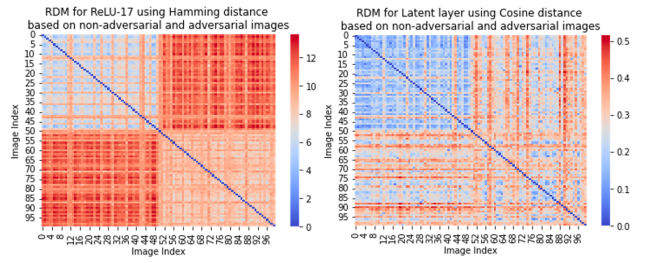

图 4. 对非对抗图像和对抗图像的 ReLU 层与潜在层进行 RDM 比较。

利用潜在层嵌入可以实现对抗图像与非对抗图像之间的线性可分性。在我们的实验中，我们观察到，位向量也能在线性可分性方面取得相当甚至更好的结果。具体而言，我们使用了 ResNet-50 的最后一个 ReLU 层 ReLU-17，生成 100352 维位向量。

尽管利用网络后部的比特向量表示实现线性可分性似乎显而易见，但我们方法的效率在于从 100352 个比特中筛选出信息量最大的 1000 个比特。通过基于这些集中的信息训练线性支持向量机，我们得以区分对抗图像与非对抗图像。在实验分析中，我们使用了 ImageNet-1K 的验证数据集。我们没有使用整个数据集，而是根据类别在 WordNet 层级体系中的位置，有选择地挑选类别 [22]。这些数据集如下：

随机选取的类别 我们使用了一个由 100 个类别组成的数据集，这些类别是从 ImageNet-1K 验证集中随机选取的。每个类别包含 50 张图像，因此该数据集总计包含 5000 张图像。

使用 WordNet 层级体系的宽泛类别 第二个数据集通过利用 WordNet 层级体系，从 ImageNet 验证数据集中识别并选取高层级类别生成。数据根据顶层类别进行标注，形成了一个包含 11,600 张图像的数据集，涵盖五个宽泛类别：水生动物、爬行动物、食肉动物、昆虫和自然物体。

所选宽泛类别的子类 第三个数据集由第二个数据集中五个类别各选取一个子类组成。选取的子类为鱼类、蛇类、工作犬、蝴蝶和植物，共计 2,650 张图像。

为生成对抗图像，我们采用了四种不同的攻击方法：DAmagenet [7]（该数据集为直接下载），快速梯度符号法 (快速梯度符号法)[11]，、投影梯度下降 (PGD) [20]，以及 Carlini & Wagner (CW) [6]。随后，我们在所得数据集上进行了五组线性支持向量机实验，包括 ImageNet 与 DAmagenet、ImageNet 与快速梯度符号法、ImageNet 与 PGD 、 ImageNet 与 Carlini-Wagner，以及 ImageNet 与所有合并的对抗图像之间的实验。

5.2.1 实验

我们首先将两组原始图像的位向量 X^{org} 和对抗图像的位向量 X^{adv} 分别划分为训练集和测试集，由此得到四组位向量： X^{org} 训练集、 X^{org} 测试集、 $X^{\text{adv}}_{\text{train}}$和 $X^{\text{adv}}_{\text{test}}$。我们分别对 X^{org} 训练集和 $X^{\text{adv}}_{\text{train}}$ 应用特征选择方法，从原始图像和对抗图像的最后一个 ReLU 层中选择 k 个特征（位）。我们丢弃其余的 ReLU特征，从而得到原始位向量的 X^{org} trainSelected 和 X^{org} testSelected，以及对抗位向量的 $X^{\text{adv}}_{\text{trainSelected}}$ 和 $X^{\text{adv}}_{\text{testSelected}}$。随后，我们将原始位向量与对抗位向量连接起来，得到最终的训练集和测试集： X_{train} 和 X_{test}。与此同时，我们生成了标签向量 y_{train} 和 y_{test}，其中标签可以是原始或对抗。最后，我们在 $(X_{\text{train}}, y_{\text{train}})$ 上训练线性支持向量机分类器，并在测试集 $(X_{\text{test}}, y_{\text{test}})$ 上评估分类器的性能。该实验的伪代码见算法1。

使用一致的数据划分，我们采用两个 k 取值：2048 和 1000，生成并评估模型。结果见2和3，分别对应准确率和 AUROC。我们选择了 $k = 2048$ 因为它对应潜在层嵌入的维度

攻击类型	随机选择			WordNet 层级体系			子类		
	ReLU_{1000} 准确率	ReLU_{2048} 准确率	潜在层 准确率	ReLU_{1000} 准确率	ReLU_{2048} 准确率	潜在层 准确率	ReLU_{1000} 准确率	ReLU_{2048} 准确率	潜在层 准确率
DAmagenet	0.978	0.992	0.977	0.961	0.989	0.984	0.944	0.989	0.968
FGSM	0.935	0.981	0.9625	0.966	0.993	0.954	0.966	0.988	0.938
PGD	0.917	0.97	0.902	0.961	0.979	0.885	0.911	0.977	0.868
Carlini-Wagner	0.958	0.992	0.962	0.959	0.991	0.996	0.970	0.961	0.993
All	0.959	0.979	0.9416	0.975	0.987	0.939	0.936	0.986	0.931

表 2. 位向量与潜在层嵌入的准确率比较。

攻击类型	随机选择			WordNet 层级体系			子类		
	ReLU_{1000} AUROC	ReLU_{2048} AUROC	潜在层 AUROC	ReLU_{1000} AUROC	ReLU_{2048} AUROC	潜在层 AUROC	ReLU_{1000} AUROC	ReLU_{2048} AUROC	潜在层 AUROC
DAmagenet	0.997	0.999	0.997	0.993	0.999	0.998	0.989	0.999	0.995
FGSM	0.982	0.997	0.992	0.995	0.999	0.988	0.995	0.999	0.982
PGD	0.976	0.996	0.962	0.994	0.998	0.948	0.971	0.998	0.936
Carlini-Wagner	0.993	0.999	0.999	0.994	0.999	0.999	0.995	0.993	0.999
All	0.985	0.996	0.97	0.994	0.998	0.963	0.978	0.998	0.955

表3. 位向量与潜在层嵌入的 AUROC 对比。

，这使我们能够将位向量与潜在嵌入进行比较。我们使用 $k = 1000$ 进行了实验，以探究当位向量的维度远低于潜在嵌入的维度时，位向量与潜在嵌入相比表现如何。

算法1线性支持向量机分类器

要求： X^{org}, X^{adv}, ResNet-50
确保： 准确率和 AUROC 分数
1) 划分为集合 (X^i) 和（X^i_{test}）其中 $i = \text{org}$，adv
2）确定类别标签（$y^{\text{训练}}_i$ ）基于数量由所选数据集指定的类别。^{训练}
3) 为（X^i选择排名前 k 个特征).
4) 为（X^i选择相同的特征 $^{\text{训练}}_{\text{test}}$).
5) 拼接所选训练集（X^i ）用 $i = \text{org}$、adv 合并为一个集合（X_{train}）。 trainSelected 于
6）连接所选测试集（X^i ）用于 $i = \text{org}$、adv 合并为一个集合（X_{test}）。 testSelected
7）为训练创建标签向量（y_{train}）和（y_{test}）集和测试集创建标签向量 ({v1}) 和 ({v2})，其中原始图像对应 0，对抗图像对应 1 对抗图像。
8) 在 $(X_{\text{train}}, y_{\text{train}})$ 上训练线性支持向量机。
9) 在 $(X_{\text{test}}, y_{\text{test}})$ 上评估性能。

5.2.2 结果

我们使用准确率和 AUROC 指标对所提出的方法进行了全面评估，并在表格中与基于潜在层嵌入（2048 个特征）的支持向量机进行了比较：2 和 3。我们的研究结果表明，在对抗图像检测中，具有 2048 个特征的位向量比潜在层嵌入表现出更高的准确率和 AUROC 分数。此外，值得注意的是，当我们将特征数量减少到 1000 时，位向量的性能与潜在层分数相当，这表明位向量无需大量信息也能发挥显著作用。这与 [14], 的研究结果一致，他们也发现，位向量即使经过大幅下采样，仍能保留有意义的信息。这凸显了基于位向量的对抗攻击检测方法的潜力，因为它能够用更少的比特提取关键信息。

6. 结论

本文探讨了 ReLU 激活模式（位向量）在解释神经网络方面的潜力。我们研究了一种扩展 RDM 的新方法。RDM 是一种广泛用于比较网络相似度的技术，扩展后可用于费德勒划分。借助表征差异矩阵（ RDM），我们考察了网络嵌入空间中输入数据的连贯性。我们利用位向量计算深度神经网络各层中两个不同类别图像之间的相似度分数，随后使用费德勒划分对这些矩阵中的类别进行分离。研究结果表明，位向量提高了整个网络对类别的可检测性，最终的 ReLU 层实现了超过 95% 的分离准确率。此外，我们还证明，位向量可有效用于对抗图像检测，使用线性支持向量机区分对抗图像与非对抗图像时，准确率超过 95%。

7. 致谢

本研究部分由美国空军通过合同编号 FA865020C1121 以及 DARPAGeometries ofLearning Program 通过资助编号 HR00112290074 资助。

参考文献

[1] Bobby Allyn。微软的新款人工智能聊天机器人说了一些“疯狂且语无伦次的话”。NPR，2023年3月。1[2] Elvio Amparore、Alan Perotti 和 Paolo Bajardi。该信任还是不信任一种解释：使用 leaf 评估局部线性可解释人工智能方法。PeerJ计算机科学，7:e479，2021。3[3] Abhijit Bendale 和 Terrance E. Boult。迈向开放集深度网络。载于 IEEE 计算机视觉与模式识别会议（CVPR）论文集，2016年6月。3[4] Yoshua Bengio、Aaron Courville 和 Pascal Vincent。表征学习：综述与新视角。IEEE模式分析与机器智能汇刊，35(8):1798–1828，2013。1[5] Wenming Cao、Zho ngfan Zhang、Cheng Liu、Rui Li、Qian- fen Jiao、Zhiwen Yu 和 Hau-San Wong。通过寻找聚类友好嵌入空间进行无监督判别式特征学习。模式识别，129:108768，2022。2[6] Nicholas Carlini 和 David Wagner。迈向神经网络鲁棒性的评估。载于2017年 IEEE安全与隐私研讨会（SP），第39–57页，2017年。3，7[7] Sizhe Chen、Zhengbao He、Chengjin Sun、Jie Yang 和 Xi- aolin Huang。针对注意力的通用对抗攻击及由此产生的数据集 damagenet。IEEE模式分析与机器智能汇刊，44 2197，2022 4 2188 3，7 ( ): – 。[8] Liang Duan、Charu Aggarwal、Shuai Ma 和 Saket Sathe。利用深度嵌入和聚类估计改进谱聚类。载于2019年 IEEE国际数据挖掘会议（ICDM），第170–179页。IEEE，2019。2[9] Miroslav Fiedler。图的代数连通度。捷克斯洛伐克数学杂志，23(2):298–305，1973。1，4[10] Miroslav Fiedler。非负对称矩阵特征向量的一种性质及其在图论中的应用。捷克斯洛伐克数学杂志，25(4):619–633，1975。1，4[11] Ian J Goodfellow、Jonathon Shlens 和 Christian Szegedy。解释并利用对抗样本。arXiv预印本 arXiv:1412.6572，2014。3，7[12] Matt Gorbett 和 Nathaniel Blanchard。利用网络特征检测错误输入。载于 IEEE/CVF冬季计算机视觉应用会议（WACV）研讨会论文集，第34–43页，2022年1月。2，3，6[13] K. He、X. Zhang、S. Ren 和 J. Sun。用于图像识别的深度残差学习。IEEE计算机视觉与模式识别会议（CVPR）论文集，第770–778页，2016年。4[14] H. Jamil、Y. Liu、C. Cole、N. Blanchard、E. J. King、M. Kirby 和 C. Peterson。用于检测对抗攻击的多面体分解对偶图。载于2022年 IEEE国际大数据会议，第2913–2921页，日本大阪，2022年12月。IEEE计算机学会。2，6，8[15] Been Kim、Martin Wattenberg、Justin Gilmer、Carrie Cai、J ames Wexler、Fernanda Viegas 等。可解释性超越特征归因：使用概念激活向量进行定量测试（TCAV）。载于《国际机器学习会议》，第 2668–2677 页。PMLR，2018。3[16] Nikolaus Kriegeskorte、Marieke Mur 和 Peter A. Bandettini。表征相似性分析——连接系统神经科学的各个分支。《系统神经科学前沿》，第 4 页，2008。1[17] Kimin Lee、Kibok Lee、Honglak Lee 和 Jinwoo Shin。用于检测分布外样本和对抗攻击的简易统一框架。《神经信息处理系统进展》，第 31 卷，2018。3[18] J. Leskovec、K. J. Lang、A. Dasgupta 和 M. W. Mahoney。发现并评估网络中的社区结构。《互联网数学》，6:29–123，2008。2[19] Xin Li 和 Fuxin Li。利用卷积滤波器统计信息检测深度网络中的对抗样本。载于《IEEE国际计算机视觉会议论文集》（ICCV），2017年10月。3[20] Aleksander Madry、Aleksandar Makelov、Ludwig Schmidt、Dimitris Tsipras 和 Adrian Vladu。迈向能够抵御对抗攻击的深度学习模型。arXiv预印本 arXiv:1706.06083，2017。3, 7[21] Saeed Mahloujifar、Dimitrios I Diochnos 和 Mohammad Mahmoody。鲁棒学习中的集中诅咒：由测度集中引发的规避攻击与投毒攻击。载于《AAAI人工智能会议论文集》，第 33 卷，第 4536–4543 页，2019。3[22] George A Miller。WordNet：电子词汇数据库。MIT出版社，1998。7[23] M. E. J. Newman 和 M. Girvan。发现并评估网络中的社区结构。《物理评论 E 》，69:026113，2004年2月。2[24] Chris Olah、Alexander Mordvintsev 和 Ludwig Schubert。特征可视化。《Distill》，2(11):e7，2017。1[25] Brooks Paige、Jan-Willem van de Meent、Alban Desmaison、Noah Goodman、Pushmeet Kohli、Frank Wood、Philip Torr 等。使用半监督深度生成模型学习解耦表示。《神经信息处理系统进展》，第 30 卷，2017。3[26] Lily Schleider、Eduardo L Pasiliao、Zhecheng Qiang 和 Qipeng P Zheng。通过神经网络特征提取和加权距离进行聚类的特征表征研究。《组合优化杂志》，44(4):3083–3105，2022。2[27] Adi Shamir、Itay Safran、Eyal Ronen 和 Orr Dunkelman。对小汉明距离对抗样本存在的一种简单解释。arXiv预印本 arXiv:1901.10861，2019。3[28] Jianbo Shi 和 J. Malik。归一化割与图像分割。IEEE模式分析与机器智能汇刊，22(8):888–905，2000。2[29] Kashif Siddiqui 和 Thomas E Doyle。利用可解释人工智能集成模型对时间序列分类进行医学深度学习的可信度量。载于《2022年IEEE加拿大电气与计算机工程会议》（CCECE），第 370–377 页。IEEE，2022。3

[30] Christian Szegedy、Wojciech Zaremba、Ilya Sutskever、Joan Bruna、Dumitru Erhan、Ian Goodfellow 和 Rob Fergus。神经网络的有趣特性。*arXiv* 预印本 *arXiv:1312.6199*，2013。2[31] Junyuan Xie、Ross Girshick 和 Ali Farhadi。用于聚类分析的无监督深度嵌入。载于国际机器学习会议，第 478–487 页。PMLR，2016。2[32] Matthew D. Zeiler 和 Rob Fergus。卷积网络的可视化与理解。载于 David Fleet、Tomas Pajdla、Bernt Schiele 和 Tinne Tuytelaars 编，计算机视觉——*ECCV 2014*，第 818–833 页，沙姆，2014。Springer国际出版公司。3[33] Bolei Zhou、David Bau、Aude Oliva 和 Antonio Torralba。通过网络解剖解读深度视觉表征。*IEEE*模式分析与机器智能汇刊，41(9):2131–2145，2019。3