arXiv:2303.02814v1 [cs.CV] 6 Mar 2023

卷积神经网络对抗攻击中的神经元脆弱性可视分析

YIRAN LI，加利福尼亚大学戴维斯分校，美国
JUNPENG WANG，Visa Research，美国
TAKANORI FUJIWARA，林雪平大学，瑞典
KWAN-LIU MA，加利福尼亚大学戴维斯分校，美国

对卷积神经网络（CNN）发起对抗攻击——将人类不可察觉的扰动注入输入图像——可能会诱使高性能的卷积神经网络做出错误预测。对抗攻击的成功引发了人们对卷积神经网络鲁棒性的严重担忧，并阻碍其应用于安全关键型应用，例如医学诊断和自动驾驶。我们提出了一种可视分析方法，通过回答以下两个问题来理解对抗攻击：（1）哪些神经元更容易受到攻击，以及（2）这些脆弱神经元在预测过程中捕捉了哪些图像特征？针对第一个问题，我们提出多种基于扰动的度量，将攻击幅度分解到各个 CNN 神经元，并根据神经元的脆弱性等级对其进行排序。针对第二个问题，我们识别出能够高度刺激用户选定神经元的图像特征（例如猫耳朵），以增强并验证该神经元所承担的作用。此外，我们基于神经元在预测中的作用，通过层次聚类辅助用户对大量神经元进行交互式探索。为此，我们设计了一个可视分析系统，融入视觉推理以解释对抗攻击。我们通过多个案例研究以及领域专家的反馈，验证了该系统的有效性。

ACM 分类体系：• **以人为中心的计算** → **可视分析**。

补充关键词和短语：卷积神经网络、对抗攻击、可解释机器学习

ACM 引用格式：
Yiran Li、Junpeng Wang、Takanori Fujiwara 和 Kwan-Liu Ma。XXXX。卷积神经网络对抗攻击中的神经元脆弱性可视分析。*ACM 交互式智能系统汇刊*XX，X，文章编号 000（XXXX），共 26 页。https://doi.org/XXXXXXX。XXXXXXX

1 引言

近年来，卷积神经网络（CNN）取得了显著的预测性能提升，广泛应用于医学诊断 [3, 23], 、目标检测 [30], 和自动驾驶 [12, 31]等领域。然而，研究人员发现，对卷积神经网络的输入图像进行细微但蓄意的扰动（称为对抗攻击 [10],），可能导致高性能卷积神经网络发生故障。这一现象对安全性或保障性问题容忍度较低的应用构成了严重威胁，例如自动驾驶系统中的道路标志识别 [8]。

由于对抗攻击会对实际应用的安全性和可靠性产生重大影响，研究人员一直致力于开发防御技术 [1, 22, 29, 35]。了解对抗攻击的行为

作者地址：Yiran Li，ranli@ucdavis.edu，加利福尼亚大学戴维斯分校，美国；Junpeng Wang，junpeng.wang.nk@gmail.com，Visa Research，美国；Takanori Fujiwara，takanori.fujiwara@liu.se，林雪平大学，瑞典；Kwan-Liu Ma，ma@cs.ucdavis.edu，加利福尼亚大学戴维斯分校，美国。

允许为个人或课堂使用而制作本作品全部或部分内容的数字或纸质副本，且无需付费，但前提是副本不得为牟利或商业利益而制作或分发，并且副本须在首页保留本声明和完整引文。对于本作品中由 ACM 以外其他方拥有版权的组成部分，必须尊重其版权。注明出处后允许摘录。若要以其他方式复制、重新发布、发布到服务器或重新分发至列表，须事先获得具体许可和/或支付费用。请通过 permissions@acm.org 申请许可。© XXXX 美国计算机协会。稿件已提交至 ACM

并揭示 CNN 神经元的脆弱性，是有效定位 CNN 弱点并防御相应攻击的关键 [4]。

与我们的工作相关，多种可视分析方法 [5, 7] 通过比较神经元在原始图像和扰动图像上的激活路径，识别对抗攻击中的关键神经元。还有一些方法 [7, 14]通过考察能够最大程度激活某个神经元的图像特征（例如猫耳朵），或合成新的可解释图像特征 [26, 27]，来阐释各个神经元在预测中的作用。然而，这两类方法都存在局限。前一类方法，即提取激活路径，通常在逻辑和计算上都非常复杂，难以复现或扩展到大型 CNN 架构。后一类用于阐释单个神经元作用的方法，一次只能考察一个神经元。此外，特征合成方法在生成过程中涉及多种随机性，因此难以比较神经元的作用及其集体行为。

受现有研究上述局限性的启发，我们提出了一种有效且直观的方法，从两个方面解释对抗攻击：（1）识别出更多易受攻击的神经元；（2）理解脆弱神经元在预测过程中捕获的图像特征。针对第一个方面，我们分别对输入图像和神经元施加数据扰动，然后利用我们设计的多种度量对神经元进行排序，以识别最脆弱的神经元。针对第二个方面，我们利用感受野（RF） [19], ，其揭示了输入图像中能够强烈激活每个神经元的区域，并以可视化方式比较一组图像在扰动前后的感受野。除分析单个神经元外，为了高效探索大量神经元并解释其语义，我们还支持探索基于神经元在预测中的作用、按层级结构聚类的神经元群组。

我们进一步将这些功能集成到一个协同可视分析系统中，使用户（即使用卷积神经网络并关注对抗攻击的领域专家）能够通过对不同粒度的输入图像和神经元，从上述两个方面进行交互式探索，理解其对手。具体而言，系统首先通过提供图像概览，并根据扰动幅度或预测概率变化对图像进行排序，帮助用户探索输入图像及其遭受攻击后的对应图像（即图像空间）。其次，针对感兴趣的图像，系统通过依据脆弱性对卷积神经网络的神经元（即神经元空间）进行排序，并依据作用相似性对其进行聚类，帮助领域专家高效探索大量神经元。最后，系统多个视图之间的协同实现了跨两个空间的有效分析，并为不同攻击场景带来有意义的解释。借助该可视化系统，领域专家可以概览完整的图像空间和神经元空间，并根据我们的排序方法，按照选择结果深入分析特定图像或神经元。通过在两个空间内部及其之间开展高效的多层级探索，领域专家能够识别并推理卷积神经网络针对对抗攻击的弱点。我们与领域专家共同开展案例研究，以验证相关发现和由此获得的洞见。

简而言之，我们工作的主要贡献如下：

- 我们提出了多种指标，用于根据神经元易受攻击的程度对其进行排序。
- 我们构建了一个层级结构，根据单个神经元在预测中的作用对其进行组织，从而便于探索和比较卷积神经网络中数量庞大的神经元。
- 我们设计了一套集成式可视分析系统，支持对多张输入图像和多个神经元进行灵活探索，帮助领域专家从对抗攻击的解释中获得可付诸行动的洞见。

2 相关工作

我们的工作聚焦于卷积神经网络上的对抗攻击解释，属于可解释机器学习（ML）研究范畴。我们重点介绍相关主题中具有代表性的研究。

可解释机器学习的可视化。 近年来，人们提出了大量可视化方案，用于理解、调试和改进机器学习模型 [6]。这些可视化涵盖传统机器学习模型（例如， [20, 21]）、集成学习模型（例如， [18, 34, 38]）和深度学习模型（例如， [5, 7, 33]）。研究已充分证明，可视化能够在多个方面帮助机器学习模型，例如透明性、可靠性、公平性和鲁棒性 [5, 7, 21]。深度卷积神经网络的鲁棒性（在受到对抗样本 [10, 22]攻击时）是我们工作的重点。

对抗攻击的可视分析。 对抗攻击 [10] 及其解释正日益受到研究关注。从可视化领域来看，我们确定了三项密切相关的工作。首先，AEVis [5] 从卷积神经网络中提取一幅良性图像（即未受攻击的图像）及其对抗性对应图像（即受攻击图像）的数据路径，并研究这些路径的分叉与合并模式，以了解攻击发生在卷积神经网络中的位置。其次，Bluff [7] 将 CNN 神经元分为四组，即仅对良性类别激活的神经元、仅对对抗类别激活的神经元、对两者均激活的神经元，以及仅对受攻击图像激活的神经元。这四组神经元相互连接形成一个图，从中可以推断攻击过程中保留了哪些原始特征，以及新注入了哪些特征。第三，Ma 等人的可视分析系统 [21] 旨在研究针对传统回归模型的数据投毒攻击，同时将攻击过程划分为四个不同层级，即实例、模型、特征和局部结构层级。然而，这些工作存在若干局限。AEVis [5] 和 Bluff [7] 中的数据路径提取方法不仅依赖于对卷积神经网络行为的某些假设，还涉及复杂的提取过程；由于计算过程具有随机性，这可能导致不稳定的解释。Ma 等人的工作 [21]针对的是简单回归模型上的攻击，其方案难以直接扩展到复杂的卷积神经网络；在复杂卷积神经网络中，研究大量神经元与分析输入数据及其特征同等重要，甚至更为重要。

我们的工作提出了一种可视分析方法，弥补了这些局限，并为理解对抗攻击提供了新的途径。我们不再从良性图像和对抗图像的数据路径中提取关键神经元，而是只关注最后卷积层，并对该层中的所有神经元进行全面探索。通过关注最后一层，我们的方法不依赖于对模型内部结构的任何假设，因此对特定模型的依赖较小，并且适用于任何类型的卷积神经网络。此外，我们可以使用直观易懂的基于扰动的度量分析对抗攻击。借助这些度量，我们可以从不同角度对易受攻击的神经元进行排序和识别。此外，通过探索按其在预测中所起作用的相似性组织的所有神经元，我们可以从更宏观的角度理解攻击，同时避免因神经元具有重复角色而产生错误认识。

***CNN*神经元/滤波器解释。** 在计算机视觉领域，人们提出了多种特征可视化技术，用于揭示各个 CNN 神经元/滤波器所捕获的图像特征。根据所采用技术的计算顺序，这些技术可分为两类。前向技术 利用卷积神经网络的前向传播，找出最能激活某个神经元的输入像素（局部特征）。感受野 [19] 是一个典型示例，我们将在第 3.2 节详细介绍。差异图 [39] 是此类技术的另一示例：它对输入图像中的各个像素施加扰动，然后根据扰动导致的神经元激活水平变化，定位敏感特征（即像素组）。反向技术 则依赖

卷积神经网络的反向传播来合成能最大程度激活某个神经元的图像特征。例如，激活最大化 [26] 在保持卷积神经网络训练参数不变的情况下，通过梯度上升合成输入图像，以最大化某个神经元的激活值。DeepDream [24] 通过以待分析图像初始化合成图像，扩展了激活最大化，从而保留图像特征的空间相关性。特征可视化 [27]同样基于激活最大化，但通过提高合成特征的多样性对其加以改进。我们的工作采用前向技术（具体而言，是感受野提取 [19]）来可视化各个 CNN 神经元/滤波器所捕获的特征，因为这种方法通常计算效率更高，而且提取出的特征（即原始输入图像的图像块）更直观易懂。

3 背景与动机

本节概述对抗攻击、卷积神经网络，以及神经元感受野的提取过程。此外，我们还阐述了可视化对于解释对抗攻击的必要性，从而引出我们的工作。

3.1 卷积神经网络上的对抗攻击

卷积神经网络上的对抗攻击 [1, 10] 会向图像添加微小且不可感知的扰动，从而使卷积神经网络错误地预测图像的类别。该过程可表述如下。给定一幅图像 x 及其真实标签 y (y ={1, $\cdots$ c}，其中 c 为类别数），一个记为函数 $f()$ 的卷积神经网络以 x 为输入，并生成所有标签上的概率分布： $f(x) = [p_1, p_2, \cdots\ p_c]$。如果卷积神经网络训练良好且能够正确预测该实例，则 $argmax_i(p_i) = y$。对抗攻击会向 x添加细微扰动 σ，生成一幅新图像 x'，使得 $f(x+\sigma) = [p'1, p'2, \cdots\ p'_c]$，且 $argmax_i(p'i) \neq y$。 $argmax_i(p_i)$ 和 $argmax_i(p'i)$ 分别称为良性标签和对抗标签。类似地，未受攻击的图像和受攻击图像通常分别称为良性图像和对抗图像。扰动 σ 通常非常微小，因此良性图像 x 与其对抗对应图像 x'在视觉上完全相同。

通常，对抗攻击的实现可根据攻击过程是否需要了解卷积神经网络的内部细节，分为白盒攻击和黑盒攻击 [1]。本文采用白盒方法，使用投影梯度下降（PGD） [22], ，在给定扰动预算的条件下扰动输入图像，使预测概率与其真实标签之间的交叉熵损失最大化。具体而言，PGD 会迭代调整扰动， σ，使每次迭代都沿着损失函数的梯度上升 方向，将相应的概率分布 $f(x+\sigma)$ 推进，同时通过条件 $\|\sigma\|_\infty \leq \epsilon$ 对其进行限制， σ，其中 ϵ 为扰动预算。除了广受欢迎且效率高之外，已有证明表明，PGD 是利用网络一阶信息的最强攻击方法 [22]；因此，我们选择 PGD 作为对抗攻击方法。不过需要指出的是，我们的可视分析方法不依赖于特定的攻击方法，也可以轻松适配其他方法。

尽管研究人员对对抗攻击已有充分认识，但目前对这类攻击的清晰且易于获取的解释仍然很少。关于攻击为何能够欺骗卷积神经网络，仍有许多问题尚未解答，例如：哪些图像特征更容易受到攻击？攻击会如何改变神经元的作用？回答这些问题需要大量视觉证据，因此我们提出了新颖的视觉推理技术，并开发了一个用于协同分析的可视分析系统。

3.2 卷积神经网络与神经元的感受野

卷积神经网络是迄今为止最强大的图像分类模型。它将一系列卷积层串联起来，通常在其中穿插批归一化（BN）和池化层，以提取图像特征。随后，这些特征通过全连接层进行组合，生成最终的类别概率（见图 1 中的蓝色部分）。每个卷积层的输出都是通过多个通道（或图像块）提取的特征。其中每个输出都称为特征图。例如，在图 1 中，最后卷积层提取的特征图形状为 14×14，共有 512 个图像块，每个图像块都反映了相应卷积滤波器提取的输出。需要注意的是，本文与现有研究一样，将“神经元”和“滤波器”交替使用。

感受野（RF）一词源自神经科学，指身体表面上能够引发反射的刺激区域 [2]。将其扩展到卷积神经网络的语境中，感受野指对特定单元影响最大的输入区域 [9]。因此，输入图像中一组像素所反映的语义常被用于解释神经元。Zhou 等人 [39] 实现并比较了两种计算和可视化感受野的方法。第一种方法是理论近似法，它利用相应神经元的特征图，根据选定的阈值进行图像分割。第二种方法通过估算差异图，以经验方式捕获感受野。尽管第二种方法得到的感受野非常接近精确值，但其计算成本要高得多，因为差异图的计算需要对图像副本进行前向传播，而这些副本在所有可能的位置都带有噪声块。

在我们的工作中，我们采用第一种方法，因为它效率高，也足以产生合理且有意义的结果。以下我们通过四个步骤详细介绍这种方法，并参考图 1 中的示例：

(1) 将输入 RGB 图像调整为所需的感受野大小（例如，从 224×224×3 到 120×120×3）。(2) 对于所需卷积层中的目标神经元（例如，图 1 中的最后卷积层），将其特征图调整为所需的感受野大小（例如，从 14×14 到 120×120）。(3) 指定一个阈值，将第二步中放大的特征图二值化。

(4) 将调整大小后的输入的每个 RGB 通道与二值化特征图进行逐元素乘法。

在第三步中，特征图被二值化，成为第四步中叠加在图像上的掩膜，只显现并突出显示图像中的感受野。第三步使用的阈值决定了显露的像素数量。请注意，可以按照第5.2节所述交互式调整阈值。最后一步的输出就是感受野。例如，图1中的感受野结果显示，目标神经元提取的是

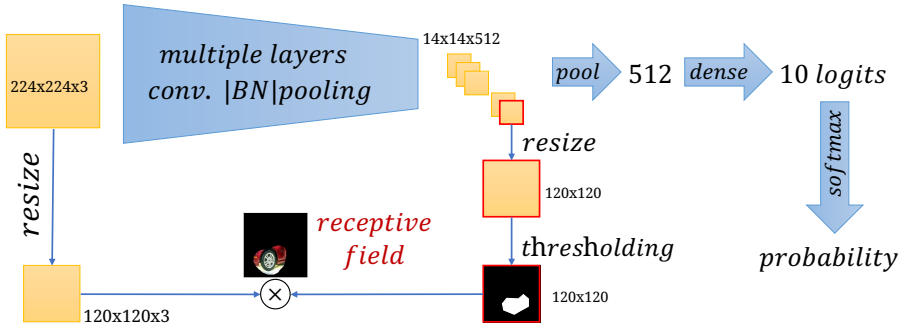

图 1. 典型的卷积神经网络架构（蓝色）以及最后卷积层中某个神经元的感受野计算 .

汽车类别输入图像中的车轮区域。如果我们在一组汽车图像中都能看到相同的结果，就可以推断，表示车轮的图像特征总是会激活这个神经元。

在这项工作中，出于以下两个原因，我们只关注 CNN 最后卷积层中的神经元来计算感受野。首先，与其他卷积层中的神经元相比，这些神经元的感受野在语义上更易于解释。正如已有研究所阐明的那样，低层（即更接近输入的层）中的 CNN 神经元会提取基本颜色和形状基元（例如线条或曲线模式），而高层神经元则会整合先前神经元的作用，以提取与语义关联更紧密的对象级特征（例如猫头和汽车车轮）。因此，聚焦于这一学习层级结构的顶层，可以更好地展现语义。其次，最后卷积层中的神经元紧邻全连接层之前，而全连接层会对提取到的特征进行线性组合，以得出预测概率。这些神经元对最终预测概率的影响更为直接，有助于我们理解对抗攻击所导致的预测概率变化。

一方面，卷积层可能包含数百个神经元，因此单张输入图像会产生数百个感受野（一张图像，多个神经元）。另一方面，分析单个神经元的感受野也需要借助大量输入图像，以提取共享语义，并在解读过程中建立置信度（一个神经元，多张图像）。因此，我们需要能够轻松探索输入图像空间（称为图像空间）和神经元空间（神经元空间），从中找出感兴趣的图像或神经元，并从两个方面解读对抗攻击。

4 需求分析

根据提出的研究问题、第 3 节讨论的可视化方案需求，以及对密切相关文献的回顾 [5, 7], ，我们归纳出目标用户——迫切需要解读对抗攻击的机器学习专家——共有的若干分析需求。我们提炼并反复完善了以下可视分析方法和系统的需求。

- **R1：对图像空间进行多层级探索**。图像通常数量庞大且涉及多个类别，我们的首要需求是探索图像空间并定位感兴趣的图像。这要求我们的系统能够：– **R1.1:** 概览良性图像、对抗图像及其在各项预测中的分布；– **R1.2:** 根据特征相似性（例如，相似的猫图像）或预测相似性（例如，所有猫图像都被误预测为狗）筛选图像；– **R1.3:** 根据与攻击相关的指标（例如，扰动幅度或预测变化）对图像排序并进行选择。

- **R2: 高效探索**神经元空间。即使在单个层中，卷积神经网络通常也包含数量过多的神经元；因此，探索神经元空间并非易事。为实现高效探索，我们应解决以下问题：应该探索哪些神经元、如何定位这些神经元，以及如何同时探索多个神经元。具体而言，我们的系统需要能够：

  – **R2.1**：有效识别最脆弱的神经元，并优先对其进行分析；– **R2.2**：直观揭示神经元在预测中的作用，以及其作用如何受到对抗攻击的改变；– **R2.3**：共同探索在对抗攻击前后具有相似行为（即响应相似图像特征）的神经元，以提高探索效率。

- **R3：对抗攻击的解读**跨越图像空间与神经元空间。通常，需要同时探索图像空间和神经元空间，才能解读对抗攻击，因为图像空间承载着有意义的语义，而神经元空间则编码了卷积神经网络的预测机制。只有将这两个空间联系起来，我们才能通过人类可理解的语义阐释卷积神经网络的工作原理。稿件已提交至 ACM

我们通过人类可理解的语义阐释卷积神经网络的工作原理。跨空间分析可以聚焦于一对一关系（即神经元如何作用于一幅图像）、一对多关系（即神经元如何作用于一组相似图像），或多对一关系（即不同神经元对一幅焦点图像有何反应）。

5 可视分析方法

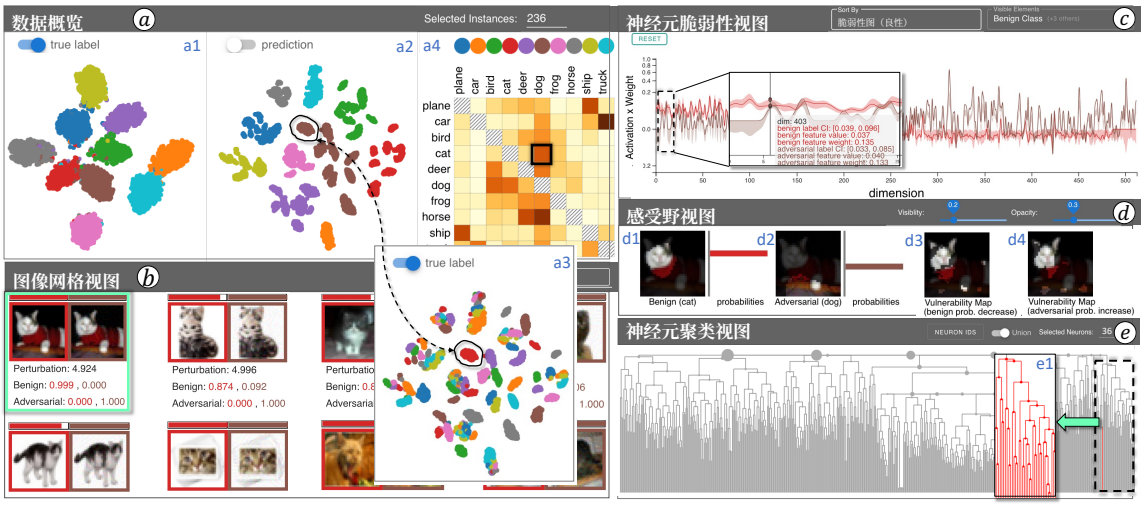

图 2. 使用我们的可视分析系统探索和解读对抗攻击。数据概览 (a) 基于卷积神经网络倒数第二层的激活值，展示良性图像之间的相似性 (a1) 和对抗图像之间的相似性 (a2)。点的颜色编码可在真实标签与预测标签之间切换 (a2 和 a3)。预测矩阵 (a4) 展示了根据真实标签和对抗标签划分的图像分布，单元格颜色越深的橙色表示其中的图像越多。图像网格视图 (b) 列出与数据概览 中的所选项对应的良性图像和对抗对应图像（此处选中 a4 中的猫-狗单元格）。对于图像网格视图中选定的图像，神经元脆弱性视图 (c) 根据神经元激活值×权重的差异对神经元进行组织。感受野视图 (d) 展示感受野和脆弱性图，以帮助解读各神经元的行为。神经元聚类视图 (e) 展示基于各神经元行为相似性构建的树状图，便于高效解读多个神经元的行为。

本节介绍我们的可视分析方法和系统。如图2所示，系统由五个主要视图组成。系统左侧的图像空间探索组件 提供数据概览（图2-a）和图像网格视图 （图2-b），以支持**R1**。右侧的神经元空间探索组件 包括神经元脆弱性视图、感受野视图和神经元聚类视图 （分别对应图 2-c、d、e），以支持**R2**。五个视图相互联动，用户可以灵活地在图像空间和神经元空间之间切换，将二者联系起来进行跨空间探索（**R3**）。

5.1 图像空间探索

我们的分析聚焦于在对抗攻击前预测结果为真阳性的测试集图像。我们使用 PGD 攻击这些图像，并仅选择攻击成功的图像进行分析。每张原始/良性图像都有一张对应的受攻击/对抗图像。图像空间探索组件 旨在探索这些图像，目标有两个：（1）了解良性图像和对抗图像的整体情况，以及

它们与所研究的卷积神经网络预测结果之间的关系；（2）探索一组感兴趣的图像（良性-对抗样本对），以识别值得进一步分析的重要或代表性图像。

为实现第一个目标，本文设计了数据概览 （图2-a）。两个散点图视图 分别展示所有良性（图2-a1）和对抗图像（图2-a2）。这些布局基于 t-SNE 投影，投影以各个图像在卷积神经网络倒数第二层的激活值 [32] 作为输入（例如，图1中由池化层生成的512个特征）。点的颜色根据用户的选择，分别表示真实标签或预测标签。例如，图2-a2和图2-a3分别使用预测标签和真实标签。预测矩阵视图 （图2-a4）通过热力图展示所有图像的分布，其中行索引为真实标签，列索引为对抗标签（即对抗图像的预测标签）。例如，如果一张猫图像经过对抗扰动后被误判为狗，它就会落在猫-狗单元格中，图2-a4对此进行了标注。每个单元格的橙色深浅表示落入该单元格的样本数量。用户可以在两个散点图视图 中的任意一个套索选择点，或点击预测矩阵中的单元格，选取一组图像，再通过其他视图进行详细分析。

图像网格视图 （图2-b）旨在实现第二个目标。数据概览中选定的图像以图像对网格的形式呈现（左：良性图像，右：对抗对应图像）。图2展示了从预测矩阵视图中选定猫-狗单元格后的结果。每幅图像的边框颜色反映其预测标签，而图像上方填充色条的长度则分别表示良性标签和对抗标签的预测概率。为帮助用户识别感兴趣的图像对，该视图支持按照四种指标对图像对进行排序：按扰动幅度递增排序（即良性图像与对抗图像之间的 L_2 距离）、按扰动幅度递减排序、按良性标签概率递减排序，或按对抗标签概率递增排序。如第3.1节所述，对抗攻击旨在使良性标签与对抗标签不匹配。这可能是通过降低良性标签的概率、提高对抗标签的概率，或同时采取这两种方式实现的。此外，扰动幅度反映了攻击的总体强度。因此，上述四种指标有助于按照对抗攻击干扰预测的强度或成功程度对图像进行排序。每个图像对下方列出了准确的扰动幅度，以及属于良性类别和对抗类别的概率，供参考。

借助上述视图，图像空间探索组件 允许用户采用“概览+详情”策略，以自顶向下的方式探索图像空间。选定的图像对将用于下一个组件中的进一步分析。

5.2 神经元空间探索

聚焦于选定的图像对，我们从所研究的卷积神经网络的最后一个卷积层中提取神经元信息，并提供多种探索方式。本节将大量神经元的探索与解释归纳为两个问题，并说明我们回答这些问题的方法。

5.2.1 如何识别最脆弱的神经元？一种直接的方法是根据良性图像与对抗图像之间神经元激活值的差异对神经元进行排序。然而，神经元激活值并不直接编码图像语义，而图像语义对于理解对抗攻击至关重要。为了系统地评估神经元脆弱性，并将脆弱神经元与图像语义关联起来，我们引入两种新指标，通过扰动输入图像或输出神经元来量化脆弱性，从而实现更稳健的排序。稿件已提交至 ACM

算法 1通过区域替换生成脆弱性图。

1: w, h, k, s, c // 图像宽度、高度、区域大小、步长、类别数
2: $bImg$, $aImg$ // 良性图像，对抗图像
3: bId, aId // 良性类别 ID，对抗类别 ID
4: $[y_1, y_2, ..., y_c] = CNN(bImg)$
5: **对于** $i = 0;\ i < w;\ i{+}{=}s$ **do**
6: **对** $j = 0;\ j < h;\ j{+}{=}s$ **执行**
7: $substituteImg = bImg$
8: $substituteImg[i{-}k{:}i{+}k,\ j{-}k{:}j{+}k] = aImg[i{-}k{:}i{+}k,\ j{-}k{:}j{+}k]$
9: $[y'_1, y'_2, ..., y'_c] = CNN\ (substituteImg)$
10: $bMap[i, j] = y'_{bId} - y_{bId}$// 良性类别概率降低
11: $aMap[i, j] = y_{aId} - y'_{aId}$// 对抗类别概率增加。
12: **结束循环**
13: **结束循环**

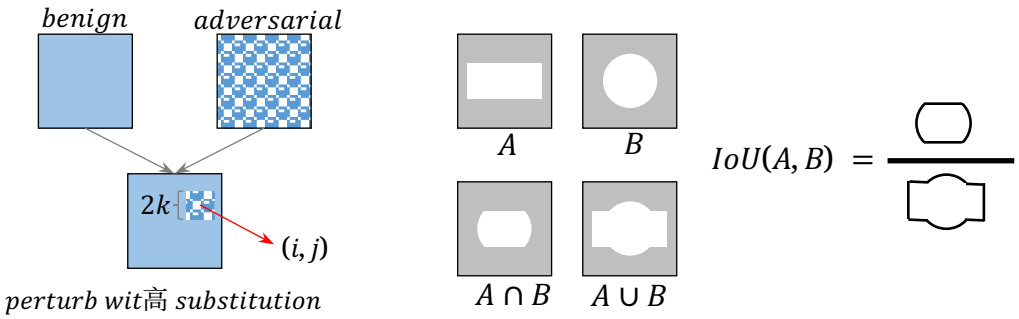

图 3.（左）通过图像替换/扰动探测图像的脆弱性。（右）并集、交集和交并比计算。

基于扰动的方法广泛用于机器学习分析与解释，它们通过刻意控制输入中的局部变化，刻画这些变化对模型行为的影响 [11, 17]。

图像扰动度量。只需对卷积神经网络进行一次前向传播，即可生成所有神经元（最后卷积层中）的感受野，从而大致了解每个神经元关注哪些区域。接下来的问题是，哪些区域更容易受到对抗攻击。通过将这些较脆弱的区域与关注这些区域的神经元联系起来，我们可以找出脆弱神经元。我们通过引入脆弱性图来量化图像不同区域的脆弱性，其生成过程如下（见算法 1 和图 3 左侧）。对于以像素 (i, j) 为中心、大小为 $2k\times2k$ 像素的小区域，我们将良性图像中的该小区域替换为对抗图像中的对应区域（第 7–8 行），将新图像输入卷积神经网络（第 9 行），并记录像素 (i, j) 处的概率变化。我们同时考虑良性类别概率的下降和对抗类别概率的上升（第 10–11 行）。将此过程扩展到所有像素（第 4–5 行），即可生成两张脆弱性图，即第 10–11 行中的$bMap$ 和 $aMap$，其尺寸与输入相同。随后，通过阈值对这两张图进行二值化，使其能够与每个神经元的感受野进行比较。接下来，我们计算每个神经元的感受野与从 $bMap$ 或 $aMap$中选出的某一张二值化脆弱性图之间的交并比（IoU，图 3 右侧）。根据相应的交并比对神经元排序，即可找出最脆弱的神经元（分别对应良性类别概率下降或对抗类别概率上升），因为这些神经元的关注区域与脆弱区域重叠更多。本质上，该度量依赖于对输入图像进行扰动（即区域替换）。

神经元扰动度量。卷积神经网络最后卷积层的特征图经池化层聚合，转换为一个 n维向量。例如，在图 1 中，大小为 14×14 的 512 个特征图被聚合为 512 个值（即 n=512）。随后，全连接层通过加权求和将 n维向量整合为一个 c维向量（c 是

类别数），从而生成不同类别的逻辑值。请注意，逻辑值经 softmax 函数归一化后会产生类别概率。表示 n 个神经元（或称激活值）活跃程度的 n维向量，会因输入图像而异，但同一类别的图像应呈现大致相似的模式。因此，我们计算同一类别图像所对应的 n维向量的置信区间（CI）。如图 4 所示，我们用曲线表示良性图像的 n维向量（曲线上的每个点对应一个神经元的激活值），并用带状区域表示该图像所属类别的置信区间。对抗对应图像也会以相同方式可视化这些信息。与图像网格视图类似，我们还根据相应的良性标签和对抗标签为曲线和带状区域着色。可以认为，对抗攻击会使良性图像的曲线向对抗带状区域的趋势偏移。我们可以将每个良性神经元的值替换为对抗神经元的值，以考察最终概率的变化，并用这一变化衡量神经元的脆弱性。按照测得的脆弱性排序，相当于根据两条曲线或两个带状区域之间的差距进行排序。我们的探索表明，与根据两条曲线之间的差异排序相比，根据两个带状区域之间的差距排序结果更稳定，也更能反映类别级别的激活值变化。带隙（BG）的详细计算公式如下：

$$BG\left(\left(CL_a, CU_a\right), \left(CL_b, CU_b\right)\right) = \begin{cases} CU_a - CL_b & \text{if } CL_b > CU_a \\ CL_a - CU_b & \text{if } CL_a > CU_b \\ 0 & \text{otherwise} \end{cases}$$

其中，$(CL_a,\ CU_a)$ 和 $(CL_b,\ CU_b)$ 分别表示对抗类别和良性类别置信区间的下界和上界。如图 4 所示，在应用上述排序后，区域 A/C 中的神经元会受到对抗攻击的激发/抑制。区域 B 中的神经元不会受到该攻击的显著影响。该视图支持水平缩放，以处理数量过多的神经元。

我们还注意到，即使某个神经元具有非常高的激活值，全连接层中与该神经元对应的权重也可能很小。在这种情况下，将权重与激活值相乘以生成逻辑值后，高激活值神经元对预测的贡献可能很小。因此，为了捕捉对预测的实际影响，我们使用曲线及其带状区域来表示 nD向量与全连接层中对应权重的乘积结果（即激活值×权重），而不是 nD向量。请注意，良性类别和对抗类别具有不同的权重。例如，图 1 展示了一个包含 10 个类别的卷积神经网络结构，全连接层中的权重大小为 512×10 （10rows）。该权重矩阵的每一行分别对应一个类别的权重。

可视化。神经元脆弱性视图和感受野视图旨在实现上述解决方案。神经元脆弱性视图（图 2-c）是神经元空间探索的起点，并展示了图 4 中说明的图表。将鼠标悬停在图表中的任意 x位置（对应一个神经元）上时，会弹出工具提示，显示该神经元的详细信息。用户可以按照不同的已介绍度量对神经元进行排序（即，基于 $bMap$ 或 $aMap$ 的图像扰动度量，以及神经元扰动度量）。这些度量有助于用户找出最脆弱的神经元，并选择感兴趣的神经元进行进一步探索。

感受野视图 （图 2-d）展示两类信息，用于呈现从图像网格视图 中当前选定图像的脆弱性，以及由神经元脆弱性视图中选定神经元捕获的图像特征的语义。第一类信息是所选神经元针对良性

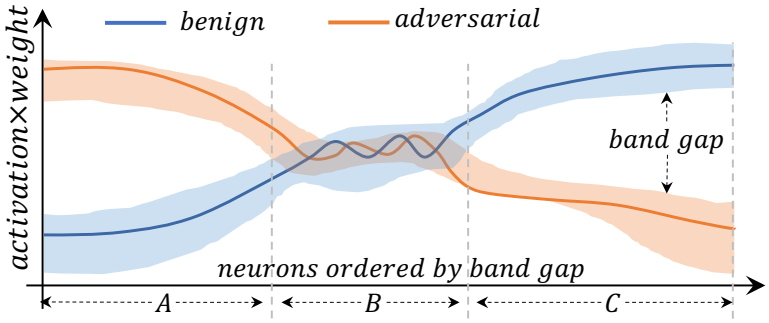

图 4. 使用带隙度量神经元对对抗攻击的脆弱性（即神经元激活值×权重的偏移）。神经元按带隙大小排序。区域 A 中的带隙值为正，区域 B 中为零，区域 C 中为负。在分析中，区域 A 左侧和区域 C 右侧的神经元最值得关注。

（图 2-d1）和对抗性（图 2-d2）图像的感受野。每个感受野右侧都显示了相应的预测概率分布。图 2-d1 和图 2-d2 中的示例分别对猫和狗的预测概率为 1；但是，如果多个类别的概率均不为零，视图中就会显示多个条形。第二类信息是两张脆弱性图：一张突出显示图像扰动使良性类别概率下降最多的区域（图 2-d3），另一张突出显示图像扰动使对抗类别概率上升最多的区域（图 2-d4）。默认情况下，我们显示脆弱程度最高的 20% 像素；用户可以通过视图中的滑块交互式调整这一比例。

5.2.2 如何探索数量过多的神经元？虽然可以按照指定的脆弱性度量对神经元进行排序（第 5.2.1 节），但逐个检查数百个神经元仍然效率低下。事实上，用户探索了某个神经元的语义后，自然会进一步思考哪些神经元的行为相似，以及它们是否相似。一旦识别出一组相似的神经元，只需检查其中一个代表性神经元，就可能了解整个群体的行为，从而提高探索效率。

我们通过对神经元进行层次聚类，并使用交互式树状图将层级结构可视化来回答这一问题，如神经元聚类视图 （图2-e）所示。作为聚类所使用的不相似度度量，我们采用两个感受野之间的 L_2 距离来衡量相应神经元之间的不相似性。需要指出的是，也可以使用两个感受野之间的交并比，或相应激活图之间的 L_2 距离；然而，根据我们在补充材料中进行的比较，没有明显证据表明某一种度量优于其他度量。当用户从神经元脆弱性视图中选择一个神经元时，其在树状图中的位置会被高亮显示，如图2-e最右侧的红线所示。通过该位置，用户可以轻松识别出行为相似的神经元，因为相似神经元共享相同的祖先节点，并且通常在树状图中彼此靠近。此外，为了查看相似神经元汇总后的感受野，用户可以从树状图中点击多个节点，这将选中被点击的节点及其后代节点。例如，在图2-e中，选中了以虚线标注的子树的根节点，因此相应的后代节点会被高亮显示，如图 2-e1所示。对于所有选中的神经元，系统会计算其感受野的并集，并相应地更新感受野视图。用户还可以通过视图顶部的切换开关，将并集 操作切换为交集 。

根据感受野的相似性并通过层次聚类对神经元进行分组，是一种总结神经元行为的新方法。结合使用第 5.2.1 节所述基于脆弱性度量的排序和这种分组方式，我们能够发现那些可能对预测产生重要影响、但仅依靠基于脆弱性度量的排序难以识别的神经元，

即使仅依靠基于脆弱性度量的排序也是如此。例如，由于激活值和×权重存在差异，脆弱性最高的神经元及其行为相似的神经元不会在神经元脆弱性视图 中相邻排列。然而，当对抗攻击针对脆弱性最高的神经元时，这些相似的神经元也很可能受到攻击的影响。即使在上述情况下，用户也可以先从神经元脆弱性视图 中选择脆弱性最高的神经元，再通过神经元聚类视图查找相似神经元。因此，我们的分组方法不仅能够高效地探索神经元空间，也支持更全面地探索神经元空间。

5.3 跨空间探索

图像空间承载着不同类别的语义，而卷积神经网络处理图像的机制则体现在神经元空间中。领域专家常常需要联合分析这两个空间，将语义与卷积神经网络的工作机制联系起来，并解读对抗攻击（**R3**）。下面列出的三个场景展示了我们的协同可视分析系统如何支持跨空间分析。

一图像—一神经元。 此场景探索所选单张图像与单个所选神经元之间的关系，以解读该神经元在特定图像上的行为。具体而言，用户在图像网格视图中选择一张图像，并在神经元脆弱性视图中选择一个神经元后，系统会提取该神经元的感受野并将其叠加到所选图像上。在图 2-d1 中，我们可以清楚地看到，该神经元捕捉到了猫图像中的面部区域。相比之下，该神经元无法再从对应的对抗图像中提取猫的面部，这揭示了对抗扰动导致的神经元功能变化。

一神经元—多图像。 单张图像中神经元捕捉到的语义可能不足以充分解释神经元的行为。第二种探索场景允许用户选择多张图像，并利用从所选图像中捕捉到的语义，更全面地解释神经元的行为。例如，点击感受野视图中良性图像的感受野时，弹出窗口会列出排名前-m 的图像，这些图像要么能最大程度地激活神经元（排序选项 1），要么具有最高的预测置信度（排序选项 2）。每张图像都会连同其感受野一并显示，以解释神经元捕捉到了什么。另一方面，对对抗图像的感受野执行相同操作时，系统会先按照上述相同流程提取良性对应图像中排名前-m 的图像，然后弹出窗口，列出这些良性图像对应的对抗图像及其感受野。这些列出的图像可作为上下文，提供更多视觉证据，帮助用户在解释神经元功能时建立信心。例如，图 5 左上角显示的六张图像表明，大多数感受野都捕捉到了猫脸区域，从而印证了该神经元的行为。

一幅图像—多个神经元。 神经元数量庞大，因此需要同时对它们进行探索。神经元聚类视图 利用所选图像中神经元的感受野对神经元进行聚类，并支持按组探索感受野。例如，在图 2-e1 中选择树状图节点，即可提取属于该树枝的所有神经元。在感受野视图中，可视化这些神经元在良性图像和对抗图像上的感受野并集或交集。采用交集时，可以揭示这些神经元的共同行为；而感受野的并集则展现了这些神经元的整体行为。

6 案例研究

我们首先使用一个相对较小但广为人知的图像数据集 CIFAR10 [16], 来展示系统的用法。具体而言，演示将介绍我们如何运用该方法回答两个重点研究问题。接下来，稿件已提交至 ACM

我们使用一个更大的数据集 Places365 [40], 来验证解决方案的普适性。虽然我们在每个案例中都选取了特定类别进行说明，但也可以用类似的方式研究和解读针对其他类别的攻击。

对抗攻击设置。对于 PGD 攻击算法，我们采用了 Madry等人 的实现 [22]。该算法包含三个超参数：迭代步数 κ、边界邻域的半径（或扰动预算） ϵ，以及迭代步长 α。对于 CIFAR10 和 Places365 数据集上的攻击，我们将参数设置为： κ=7、 ϵ=8/255=0.0314 和 α=2/255=0.00784。

6.1 CIFAR10数据集

CIFAR10 数据集 [16] 的测试集中包含 10,000 张 32x32 RGB 图像，分属 10 个类别。我们研究的卷积神经网络是一个预训练的 ResNet18 [13]，能够正确预测 10,000 张图像中的 8,200 张。随后，我们使用 PGD 攻击对这些真正例图像进行扰动。所有图像均成功遭受攻击，并被输入我们的系统进行分析。

6.1.1 对抗图像的分类。首先，我们关注按真实标签或对抗预测着色的对抗图像散点图视图。如数据概览中的图2所示，具有相同预测结果的对抗图像往往会形成较大的簇（图2-a2），而每个大簇中都包含由真实标签相同的图像构成的小簇（图2-a3）。这一模式表明，虽然预测类别相同的对抗图像会受到扰动，从而包含相似的特征（即预测类别的特征），但真实标签相同的图像也各自具有受攻击模式。

6.1.2 使用图像扰动度量理解攻击。在上述使用散点图视图进行图像空间探索时，我们注意到，大量猫图像（红色簇）受到扰动后被误分类为狗图像（棕色），如图2-a2和a3中的标注所示。随后，我们选中这个簇，在图像网格视图中查看其详细信息。

如图 2-b 所示，我们首先按照扰动幅度从小到大的顺序对图像进行排序；因此，受到攻击扰动较小的图像排列在视图上方。对于这些图像，即使扰动很小，对抗攻击仍会使最终概率发生显著变化，如图像对上方的柱形图和各图像对下方的数值所示。例如，图 2-b 中展示的大多数良性图像被预测为猫的概率接近 100%，而它们对应的对抗图像被预测为狗的概率也接近 100%。如图 2-b 中绿色矩形所示，我们选取扰动最小的一对图像进行进一步研究。

如图 2-c 所示，神经元脆弱性视图 会根据此选择进行更新。为了揭示脆弱神经元及其提取出的图像特征，我们使用脆弱性图按照图像扰动度量对神经元进行排序（即， *bMap* 算法 1）。然后，我们放大排名前 25 的神经元，如图 2-c 中的虚线所示。随后，我们选取一个感兴趣的神经元——神经元403。我们关注该神经元，因为良性图像与对抗图像之间的激活值×权重差距仍然很小，因此神经元扰动度量无法揭示这类成功遭受攻击的神经元。

我们进一步在感受野视图 中检查相应的感受野和脆弱性图（图 2-d）。我们首先观察到，良性概率下降（图 2-d3）和对抗概率上升（图 2-d4）的脆弱性图都包含很大一部分猫脸区域。这表明，面部区域周围的扰动是导致该猫图像被误分类为狗图像的原因之一。在图 2-d3 中，我们还看到脚部区域周围存在另一个关注区域，这可以被视为猫类别概率下降的另一个原因。我们还检查了所选神经元的感受野，如图 2-d1 和 d2 所示。尽管该神经元具有相似的



图 5. 猫与狗案例的上下文图像。在被误分类为狗的猫图像中（上方），大多数良性图像（左上）捕捉到了猫脸，而大多数对抗对应图像（右上）捕捉到的是猫脚或下半身。这表明，同一个神经元在对抗攻击前后的表现有所不同。在被误分类为猫的狗图像中（下方），也可以观察到这种行为变化：良性图像和对抗图像分别捕捉到了狗的脚部/下半身和脸部。

激活值×权重，但良性图像和对抗图像的感受野完全不同（良性：面部；对抗：脚部区域）。感受野的这一差异解释了对抗攻击如何改变神经元403的行为（即关注区域从面部转移到脚部）。

为了验证这一解释，我们点击每个感受野，查看上下文图像。这些图像同样是经过扰动、被误分类为狗的猫图像。如图 5（上）所示，对于良性图像，神经元403确实试图捕捉面部区域；而对于对抗图像，则试图捕捉脚部区域。对此结果可以这样解释：面部区域周围的扰动干扰了神经元403对猫“面部”特征的识别，而脚部区域周围的扰动则使神经元403转而关注“脚部”特征，这种特征很容易被误认为是狗的特征。为了进一步验证这一解释，我们从预测矩阵视图 中选择受到攻击、被误分类为猫的狗图像，并将其上下文图像可视化，如图 5（下）所示。输入狗的良性图像时，神经元403捕捉脚部区域；而对于对抗图像，其感受野则转移到面部区域。这表明，神经元403的作用是捕捉猫图像中的面部区域、狗图像中的脚部区域；对抗攻击则试图通过削弱良性标签的特征并增强对抗标签的特征，误导神经元403捕捉对抗标签的特征。

6.1.3 使用神经元扰动度量理解攻击。 接下来，我们使用神经元扰动度量分析攻击导致飞机类别与汽车类别之间发生混淆的情况。我们首先从预测矩阵视图 中选择飞机-汽车单元格（图6-a），以获取经过对抗扰动后被错误预测为汽车的飞机图像（良性标签）和（对抗标签）。在图像网格视图中，按照预测概率的变化对图像进行排序后，我们识别出卷积神经网络给出截然不同预测的图像。其中一组图像对如图

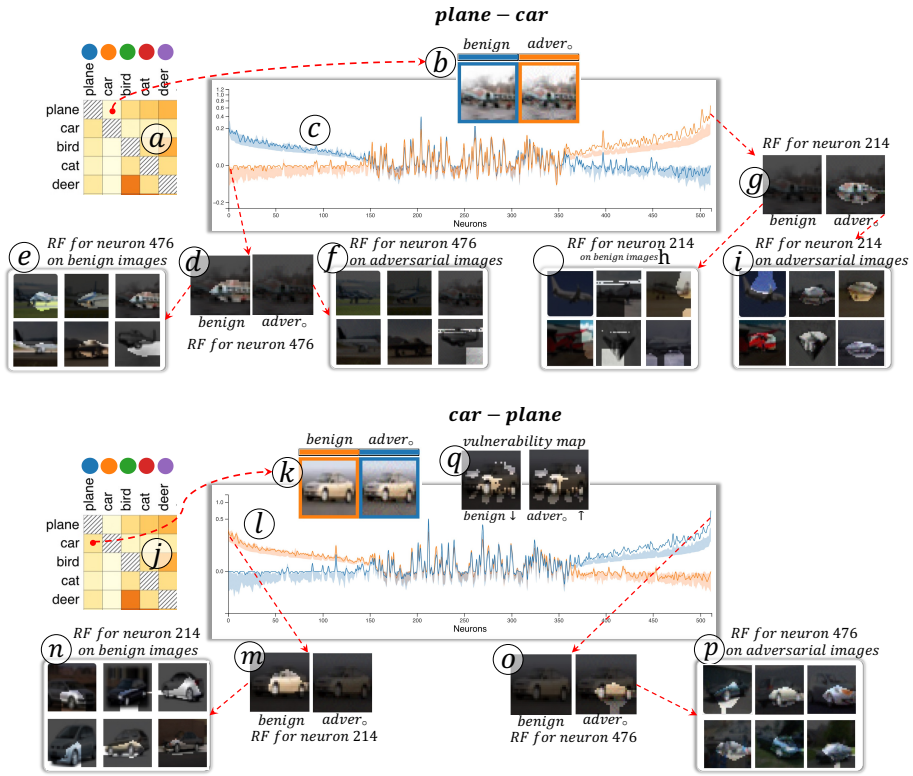

图6. 飞机-汽车案例始于从预测矩阵视图(a)中选择该单元格。图像网格视图列出了对应的图像，例如(b)中的图像。卷积神经网络使用位于神经元脆弱性视图(c)中 x轴两端的神经元476和214，分别识别飞机类别和汽车类别的车轮区域(d-e、m-n)。对抗攻击通过(f)关闭神经元476，并通过(g–i)激活神经元214，诱使卷积神经网络将飞机识别为汽车。通过查看与汽车-飞机单元格相关的可视化(j–p)，可以进一步考察神经元476和214的作用。根据这些观察，我们可以推断，卷积神经网络仍会从飞机图像中提取车轮特征，但它认为这些车轮属于汽车，因为提取它们的是神经元214。

图6-b所示。根据图6-b中填充的蓝色和橙色条，卷积神经网络对该图像的预测以较高置信度从飞机变为汽车。然而，从视觉上很难区分良性图像和对抗图像。

接下来，选中这对图像后，我们探索其对应的神经元空间。按照“间隔大小”对神经元排序（即采用图 4 中相同的排序方式），我们很容易就能在图6-c的两侧识别出最脆弱的神经元。例如，神经元476位于最左侧；处理良性图像（蓝色曲线）或飞机图像（蓝色带）时，它会激活；但处理对抗图像（橙色曲线）或汽车类别的图像（橙色带）时，它会变为不活跃状态（接近0， y-轴值），或呈负向激活（即降低良性类别的概率）。该神经元的变化使良性类别的概率下降幅度最大。查看所选图像中该神经元的感受野细节（如图6-d所示），我们注意到，处理良性图像时，该神经元会提取飞机的车轮和尾部特征，但经过对抗攻击后便不再提取任何特征。点击感受野会显示一组代表性图像及其感受野（此处展示了六张图像）；在良性图像中，飞机的车轮区域始终会被提取（图6-e），而在对应的对抗图像中几乎没有任何内容被提取（图6-f）。因此，对抗攻击似乎成功地阻止了该神经元正常工作。

了解这一细节有助于解读攻击，但我们仍需弄清楚，对抗攻击是否会使卷积神经网络无法感知车轮特征。接下来，我们进一步探索图6-c右侧中最能提高对抗概率的几个神经元，例如神经元214（最右侧的神经元）。处理良性飞机图像时，该神经元不活跃；处理对抗图像时则会激活。如图6-g所示，该神经元在所选对抗图像上的感受野显示，它提取了飞机的车轮区域。在上下文图像中也能观察到类似情况（图6-h和图6-i），这证实卷积神经网络在攻击后仍会提取车轮区域，只是改由另一个神经元完成。

随后，通过选择汽车-飞机单元格（图6-j），我们将分析切换到被误分类为飞机的汽车图像，并选择一对概率变化显著的图像（图6-k）。图6-l展示了按“间隔大小”排序的神经元脆弱性视图，其中神经元的排列顺序与图6-c相反。最左侧的神经元是214，它对良性类别（汽车）概率的下降贡献最大。如图6-m所示，该神经元会从良性图像中提取汽车车轮（图6-n中的上下文图像也证实了这一点）。然而，对抗图像成功阻止了该神经元提取这一特征。图6-l最右侧的神经元是476；它无法从汽车图像中提取任何内容，却能从扰动图像中提取车轮特征（图6-o和p）。

洞见。 简而言之，卷积神经网络可以从汽车类别和飞机类别中提取车轮区域，但依赖的是非常特定的神经元；例如，在我们的案例中，分别是用于汽车类别的神经元214和用于飞机类别的神经元476。对抗攻击通过扰动造成以下情况，使卷积神经网络将飞机图像误识别为汽车图像：(1) 从飞机图像中提取车轮的神经元无法提取任何内容；(2) 而从汽车图像中提取车轮的神经元可以提取车轮特征。随后，卷积神经网络将图像判定为汽车，因为它识别出的车轮属于汽车。这也表明，在区分汽车和飞机时，车轮特征是一种非鲁棒特征。在探索过程中，我们还检查了不同选定图像的脆弱性图。例如，图6-q展示了汽车图像对应的脆弱性图。这些脆弱性图帮助我们验证，图像中的车轮区域确实容易受到攻击。

6.1.4 使用层级结构探索神经元。 这里，我们展示了对图2-d所示猫图像和图6-b所示飞机图像相关神经元的探索，以证明神经元聚类视图 能够高效探索大量神经元。对于猫图像，我们按照自底向上的顺序探索512个神经元；而对于飞机图像，则按照自顶向下的顺序探索512个神经元。

自底向上探索。 在第6.1.2节中，我们检查并解读了神经元403。在神经元聚类视图中，该神经元以红色高亮显示，如图2-e右侧所示。这里的树状图是根据每个神经元的良性和对抗性感受野之间的L2距离生成的。为了快速识别行为与神经元403相似的神经元，我们找到高亮神经元所属的簇。

如图7（上）所示，我们选取由3个神经元组成的小型绿色簇，并检查感受野的交集和并集。通过交集（图7-a），我们发现这3个神经元在良性图像中都捕捉到猫脸区域，在对抗图像中都捕捉到猫脚区域。在图7-b中，感受野的并集表明，这些神经元在良性图像中的行为略有差异，但每个神经元在对抗图像中都捕捉到脚部区域。

我们向上移动两层，并选择较大的蓝色簇，其中包含 36 个神经元。这一次，感受野的并集几乎覆盖了图像的整个区域（图7-d），而感受野的交集则没有覆盖图像的任何部分（图7-c）。这表明，神经元在良性图像上的行为在簇内具有更大的差异。不过，对于对抗图像，尽管感受野的交集仍捕捉到了脚部区域，感受野的并集却覆盖了更广的区域。

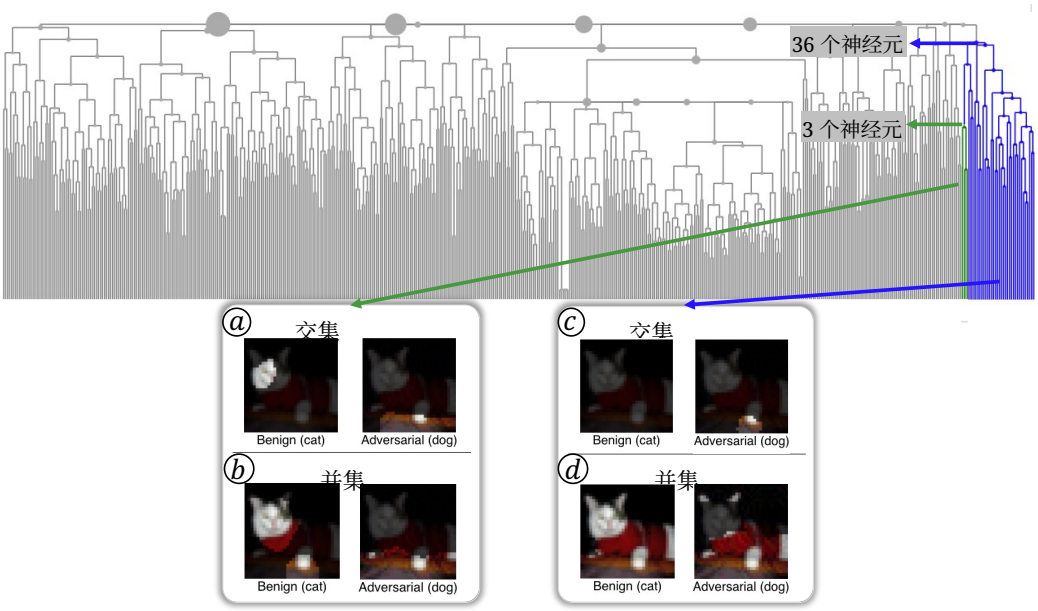

图 7. 使用神经元聚类视图自底向上探索猫图像。随着簇的规模增大（即向簇中添加更多神经元），神经元在良性图像中捕获的图像特征会更加多样（例如，3 个神经元的并集突出显示了脸部和脚部，而 36 个神经元的并集则显示了整个身体）。然而，在对抗图像中，捕获到的图像特征仍然相似（例如，3 个和 36 个神经元的并集都没有突出显示猫的脸部）。

（主要位于身体下半部分），但仍未覆盖面部区域。这表明簇内所有神经元都试图捕捉脚部区域，却忽略了面部区域。因此，尽管这些神经元可能捕捉到良性图像的不同特征，但它们经过改变后会捕捉对抗图像的相似特征。

从这两种不同的簇选择中，我们观察到，确实存在一些与其他神经元行为相似的神经元。一个有趣的发现是，对抗攻击会扰动图像，使许多神经元（36 个神经元）做出相似的错误判断（即忽略猫的脸部，转而关注它们的脚部）。

自顶向下探索。 图 8 展示了对良性飞机图像的神经元探索。从树状图中，绿色大簇立即吸引了我们的注意。该簇包含95个神经元，它们捕捉到的图像特征都非常相似，这一点从树节点极低的垂直位置可以看出。选中该簇并探索对应95个感受野的交集与并集（图 8-a）后，我们发现该簇包含所有“失活”神经元，也就是说，它们都没有从飞机图像中捕捉到任何内容。这些神经元可以排除在进一步分析之外。我们还探索了其他几个大簇，以了解它们与图像特征的对应关系。例如，如感受野的交集所示，红色簇中的28个神经元都捕捉到了飞机的机身和机轮区域（图 8-b）。与此同时，如感受野的并集所示，单个神经元还会提取一些周边特征。对于包含17个神经元的蓝色簇，其共同区域是飞机上方的天空（图 8-c）。一些神经元还提取了图像左下角的内容，因此该簇中神经元的功能稍显多样，这一点也反映在树节点较高的垂直位置上。即使只探索大簇，我们也已经能够大致了解图像特征在神经元间的分布情况。对于任意簇，用户都可以通过交互方式沿树状图从父节点遍历至子节点，从而缩小分析范围。

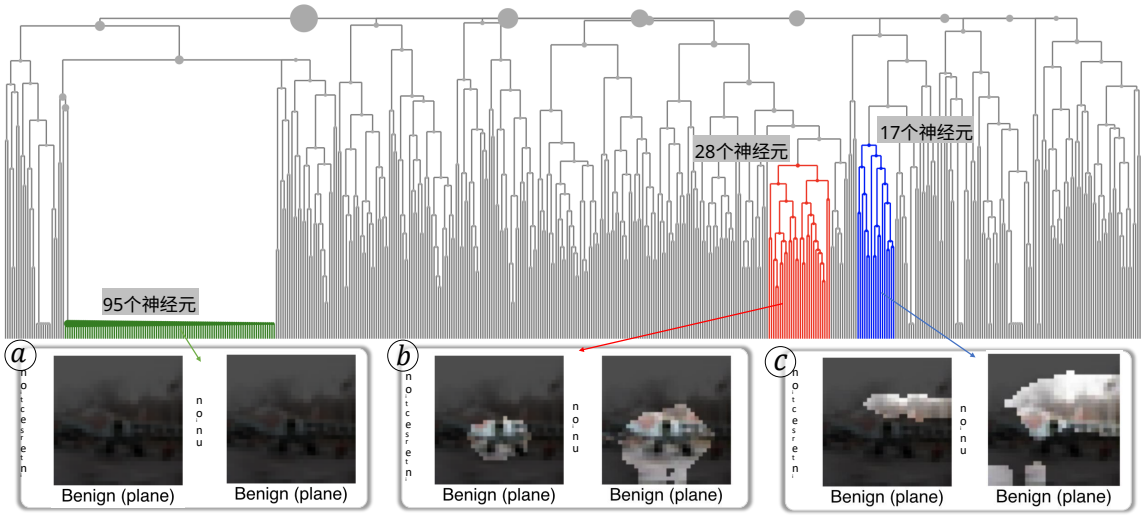

图 8. 使用 神经元聚类视图对飞机图像进行自顶向下探索。通过这一探索，可以大致总结特征的分布。例如，绿色簇神经元（a）未捕捉到任何飞机特征，因此无需进一步调查。红色簇（b）主要捕捉飞机的机身和机轮区域，而蓝色簇（c）捕捉飞机正上方的天空。

6.2 Places365 数据集

现在，我们展示一个使用 Places365 数据集 [40] 的分析案例。我们使用验证集作为输入，其中每个场景类别包含100张图像。卷积神经网络采用预训练的宽型 ResNet18 模型 [36]。攻击前，36,500张图像中有19,745张被正确预测；攻击成功导致其中12,298张图像（共19,475张图像）被误分类。类别数（即365）远多于CIFAR10数据集中的类别数。因此，我们首先通过直接从后端选取的示例图像来展示神经元空间探索，随后在一个图像子集上进行两次更完整的探索。

6.2.1 来自后端选择的图像示例。我们选择了一张手术室类别图像，该图像在受到攻击后被误分类为洁净室，如图 9 左上角所示。选择该图像是因为，与其他受攻击图像相比，其预测概率的变化最大。此外，由于大量手术室图像被误分类为洁净室，分析这组图像的攻击模式具有重要意义。

如图 9 所示，尽管良性图像与对抗图像之间几乎没有或完全没有可观察到的差异，但卷积神经网络的概率输出却存在显著差异。我们考察了图 9 顶部中央展示的脆弱性图。通过高亮区域，我们观察到，手术服周围的扰动最能降低良性概率，而手术台和手部区域周围的扰动最能提高对抗性概率。

根据图像扰动度量，我们随后选取神经元169进行进一步探索。它是良性概率下降的脆弱性图与感受野重叠度最大的神经元之一。从图 9右上方的感受野可以看出，良性图像与对抗图像的感受野存在明显差异：在良性图像中，该神经元更多关注手部和手术台区域；而在对抗图像中，则更多关注手术服区域。稿件已提交至 ACM

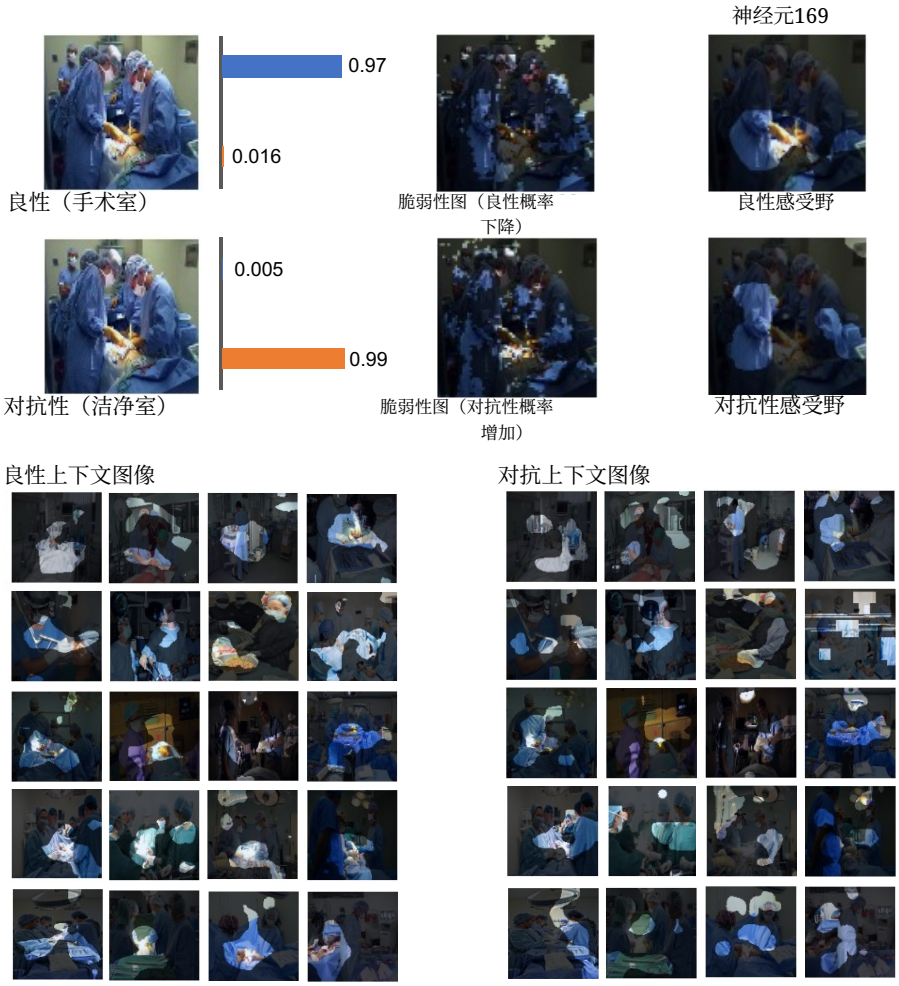

图 9. 使用 Places365 数据集进行案例研究的演示。所选神经元在良性图像中主要捕捉手术台和外科医生的手，而在对抗图像中则主要捕捉蓝色或绿色手术服。通过上下文图像可以验证这一发现。

随后，我们检查良性上下文图像和对抗上下文图像的感受野，这些图像以网格形式列于图 9（底部）。与选定图像类似，大多数良性上下文图像突出显示手部和手术台周围的区域，而大多数对抗上下文图像突出显示手术服周围的区域。

洞察。 可以得出结论，扰动抑制了手术台周围的手术室特征，而该特征对于区分手术室图像和洁净室图像十分重要。同时，手术服周围的扰动使神经元169将注意力转移到手术服特征上，而这一特征也可能出现在洁净室图像中。由此，对抗攻击成功欺骗了卷积神经网络。

6.2.2 对图像子集的探索。 为了充分利用我们的可视分析系统，我们首先将分析目标缩小到规模适中的类别集合（例如，十个类别）。这样做主要是因为预测矩阵视图 的可视化扩展性有限（即，难以显示 365 行/列），以及颜色编码方面的限制（即，人类无法

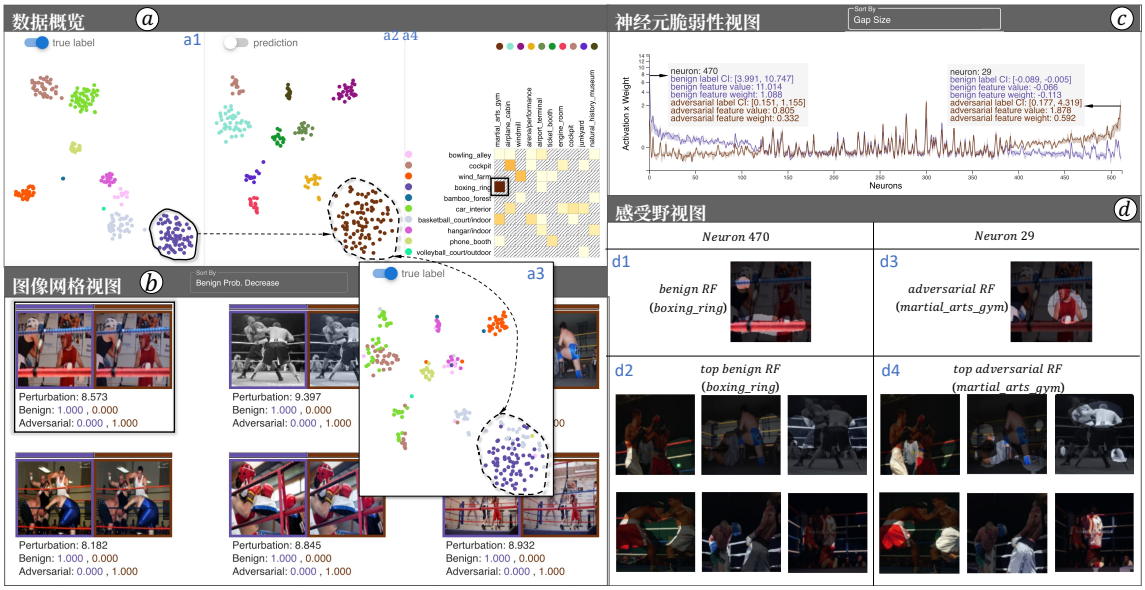

图 10. 对 Places365 中多种图像的对抗攻击进行研究。这里我们只分析了部分类别（见 a），这些类别中有大量图像在受到攻击后被误分类。从 a4 中，我们选取 boxing_ring–martial_arts_gym 单元格，其中包含数量最多的误分类图像。然后，我们使用神经元扰动度量进行进一步检查，如 b、c 和 d 所示。通过从 c 中识别出最脆弱的神经元，并观察 d 中的感受野，我们发现攻击似乎隐藏了 boxing_ring 中独有的绳索区域，同时暴露出 martial_arts_gym 中也能看到的人类相关特征。

无法区分过多不同的颜色）。在 Places365 的 365 个类别中，我们选出攻击后包含误分类图像数量最多的前 10 个 良性类别 （例如图 10-a4 中的保龄球馆_和驾驶舱行）。然后，作为对抗类别 对，我们进一步选取所选良性类别中观测到的前 10 个对抗标签所对应的类别（例如图 10-a4 中的武术_馆_和飞机_客舱列）。

如图 10-a1 和 a2 所示，良性图像（a1）和对抗图像（a2）都根据其标签清晰地分组。从图 10-a3（将基于真实标签的颜色编码应用于 a2）中，我们还可以看到，许多对抗类别中包含少数良性类别。例如，虚线标注的 martial_arts_gym 簇主要由 boxing_ring、basketball_court/indoor 组成。这种模式也可以在图 10-a4 所示的 预测矩阵视图 中观察到。

对于从预测矩阵视图中选出的特定关注单元，我们分别通过神经元扰动度量和图像扰动度量，展示两个案例来解释对抗攻击，如第6.1.2节和6.1.3节中针对CIFAR10数据集所述。

案例 ***1***：使用神经元扰动度量理解攻击。我们首先分析那些使卷积神经网络将拳击_比赛场地的图像误判为武术_训练馆的攻击；如图 10-a4 中的深橙色所示，这类图像在子集中占比最大。检索出对应图像后，我们在图像网格视图（图 10-b）中按良性类别概率下降幅度从大到小对图像进行排序。顶部的图像在遭受对抗攻击后，其预测概率发生了剧烈变化：卷积神经网络被诱骗，将良性类别的概率降至 0%，同时将对抗类别的概率提高至 100%。我们选取第一对图像来探索神经元空间。

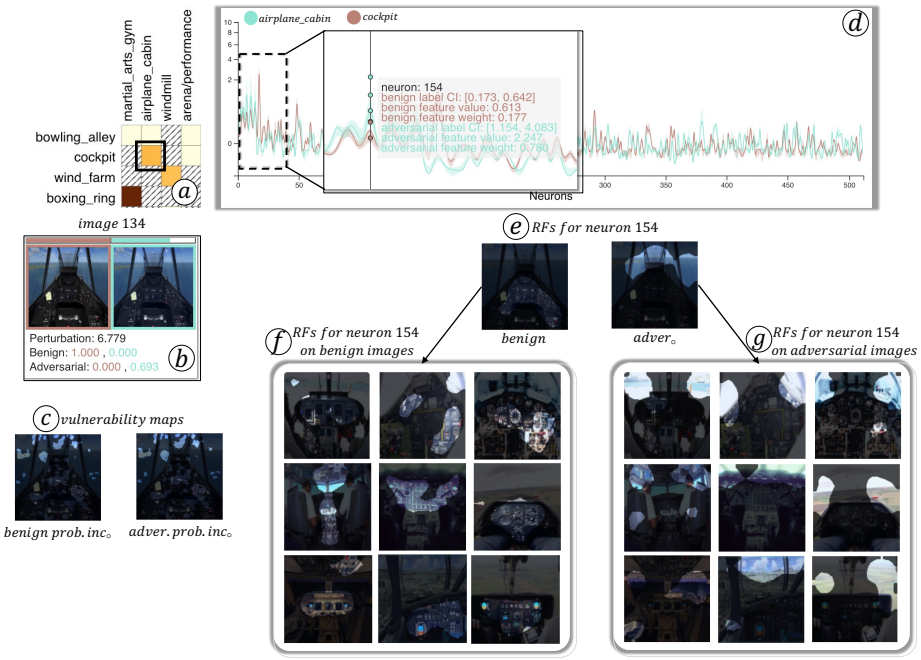

图11. 通过图像扰动度量分析对抗攻击。我们研究了与驾驶舱–飞机_机舱单元（a）对应图像上的攻击。与其他情况类似，我们将相关信息可视化（b–g）。神经元154在良性图像和对抗图像中关注不同的特征（e、f和g），这表明攻击如何改变该神经元的功能。

在神经元空间中，可以通过比较使用神经元扰动度量识别出的两个最脆弱神经元的感受野来解释这些攻击。在神经元脆弱性视图中按“间隔大小”对神经元排序后，可以在 x 轴的两端识别出最脆弱的神经元（图 10-c）。位于左端的神经元 470 主要被良性图像激活，同时也受到对抗图像的强烈抑制。相反，另一端的神经元 29 则呈现相反的激活模式。因此，我们将神经元 470 对应良性图像的感受野与神经元 29 对应对抗图像的感受野进行比较，因为每个感受野都代表了主要激活相应神经元的区域。如图 10-d1 和 d3 所示，我们发现，神经元 470 捕捉到拳击_比赛场地周围的绳索，而神经元 29 捕捉到场地内的人。通过比较多张拳击_比赛场地图像及其对应的对抗图像的感受野，进一步验证了这一差异（图 10-d2、d4）。

洞见。 总之，环绕拳击_擂台的绳索是卷积神经网络识别拳击_擂台的关键特征，而拳击_擂台内的人类则是一种容易混淆的特征，对武术_馆_预测有很大贡献。对抗攻击似乎利用扰动来隐藏绳索特征并增强人类相关特征，从而诱使卷积神经网络产生误分类。

案例 *2*：使用图像扰动度量理解攻击。我们在此展示另一个案例，解释与驾驶舱–飞机_机舱单元格对应的攻击，如图 11-a 所示。按良性类别概率下降幅度从大到小对相关图像排序后，我们选取排名第一的图像（图 11-b 中的图像 134）。扰动使其对抗类别概率从 0.000 增至 0.693。所选图像的脆弱性图（图 11-c）表明，最脆弱的区域位于窗户和仪表盘周围。

我们进一步通过探索神经元空间，解释与所选单元格相关的攻击。在神经元脆弱性视图（图 11-d）中，神经元按其感受野与脆弱性图之间的重叠程度排序。随后，我们选取神经元 154，因为它在排名靠前的神经元中交并比得分最高，并且会被良性图像和对抗图像激活。如图 11-e 所示，在良性图像和对抗图像中，神经元 154 的感受野分别聚焦于仪表盘和窗户区域。图 11-f 和 g 列出了其他被误分类为飞机_机舱的驾驶舱图像对应的感受野，呈现出与上述相似的模式，从而验证了我们的发现。

洞见。 对抗扰动成功隐藏了驾驶舱（仪表盘）的独特特征，同时放大了驾驶舱与飞机_机舱共有的特征（窗户）。由于每个神经元的功能彼此交织，施加扰动后，原本捕捉仪表盘特征的神经元可能会将注意力转移到窗户区域，从而降低驾驶舱的概率，并提高飞机_机舱的概率。

7 领域专家反馈

我们邀请了两位从事多种对抗攻击研究的领域专家（E_1 和 E_2）对系统进行了评估，涉及的攻击类型包括对抗样本、数据投毒和对抗代码分析。两位专家都是研究实验室的全职研究科学家，拥有计算机科学博士学位（研究重点为对抗深度学习）。每位专家参与的研究历时 ~90 分钟。在此期间，我们向专家介绍了两个预设的研究问题，引导他们探索第 6 节中介绍的案例，并在出现引发专家评论的发现时暂停。最后，我们与专家进行了出声思考讨论和开放式访谈，以收集他们的反馈。为确保评估客观，两位专家均未参与本文的共同撰写。

总体而言，两位专家都对我们回答两个研究问题的方式以及协同可视分析系统的交互性给予了积极反馈。E_1 支持我们仅分析卷积神经网络最后卷积层的做法，因为该层神经元提取的高级特征提供了可供解释的人类可理解语义。此外，他看到图 6 中的结果后，认为我们的脆弱性图很有启发性。这一观察促使他通过掩蔽车轮区域（即脆弱像素）来微调卷积神经网络，并促使其学习其他汽车相关特征，以提升模型的鲁棒性。这个想法与 Ilyas 等人近期的工作思路相近。 [15]。 E_2 很欣赏我们系统不同视图之间的协同。他评论说，他们的同事通常会使用不同的 Python 代码片段来生成用于解释的视觉证据。然而，离散的视觉推理过程严重阻碍了他们的分析效率。我们集成且协同的可视化（例如，跨空间探索）提供了一个更友好的界面，便于人们参与分析过程。在研究期间， E_2 也从我们的可视化解释结果中获得了许多有趣的想法。例如，他猜测，两个类别之间的非鲁棒特征（例如，图 6 中汽车和飞机的车轮）可能是另外两个类别之间的鲁棒特征（例如，汽车和猫）。他还建议将汽车图像中的车轮特征粘贴到飞机图像中，以探查卷积神经网络的行为。此外，他也认同我们对神经元的层次化探索有助于专家高效地发现“死亡”神经元或功能相似的神经元，而这些神经元可用于剪枝卷积神经网络并对其进行压缩。

专家们还指出了若干局限性，并提出了潜在的改进方向。首先， E_1 指出，PGD 并不能保证攻击连续图像特征，并建议在扰动图像以生成其脆弱性图时，考虑多个图像特征的协同效应。其次， E_2 认为，不同的攻击方法（例如 FGSM [10]）可能会导致神经元呈现不同的脆弱性等级。不过，他也推测，对同一个卷积神经网络发起攻击时，神经元的脆弱性不应发生太大变化。未来，我们可以通过验证和比较不同的对抗攻击方法来研究这一问题。第三， E_2 还建议将我们的工作扩展至其他对抗攻击，例如针对图数据或源代码的攻击；与图像不同，这些数据样本可能不包含太多语义信息。稿件已提交至 ACM

如何从中提取人类可理解的语义并利用这些语义来解释针对其他类型数据的对抗攻击，是一个值得探索的有趣未来研究方向。

8 讨论、局限性与未来工作

局限性。 正如专家指出的那样，我们当前的方法在生成脆弱性图时（图3）一次只扰动一个图像块，并未考虑多个图像块的综合影响（因为可能的组合数量非常庞大）。这是类似显著性图生成算法（例如 [39]）普遍存在的局限性。一种可能的解决办法是在扰动过程中随机抽取图像块组合，并将概率变化均匀归因于抽取到的图像块。其次，我们当前系统的可扩展性问题是在分析 Places365 数据集时出现的。具体而言，用于图像空间探索的视图（图2-a）无法处理大量类别（例如 100 个类别）。然而在实际应用中，探索通常聚焦于少数类别，因此用户可以预先选择感兴趣的类别，只将相应信息加载到系统中，正如我们在第 6.2.2 节的分析所示。此项选择也有助于缩小待探索数据样本的数量。最后，我们当前的评估依赖于对对抗攻击有充分了解的专家用户。未来，我们计划招募更多具有不同专业水平的用户，全面评估不同可视化组件的可用性。

通用性。关于通用性，我们想强调四点。第一，我们的方法适用于有目标和无目标对抗攻击 [28]。事实上，基于预测矩阵视图中的样本分布，我们甚至可以推断所研究的攻击是否为有目标攻击（即，目标类别对应的单元格颜色应更深）。第二，我们识别脆弱神经元及对其进行后续解释的方法不受攻击算法及其参数的影响。尽管本研究分析了 PGD，因为它应用广泛且效果显著，但我们的系统不包含任何针对特定攻击的功能，因此也适用于其他攻击，例如 FGSM [10]。此外，我们的方法还可用于分析采用相同算法、但使用不同参数来决定攻击强度的攻击，补充材料对此有进一步说明。第三，我们的可视分析系统不受感受野计算方式的影响。尽管为提高效率，我们采用了第 3.2 节所述的近似方法，但也可将其替换为基于扰动的感受野计算。补充材料详细比较了不同的感受野计算方法（例如 Zhou 等人 [39] 提出的差异图）。第四，我们的工作仅关注针对卷积神经网络的对抗攻击，但我们认为，通过扰动识别攻击的关键神经元这一通用方法也适用于其他类型的数据（例如文本或图数据）和深度学习模型。

系统交互性。我们系统中计算开销最大的部分是生成脆弱性图，因为这需要对所有可能的图像块进行扰动，并多次对目标卷积神经网络执行前向传播（算法 1 的第 4–5 行）。步幅参数（即算法 1 中的变量 s）也会影响性能。不过，对于 CIFAR10 数据集中的 s=1，即使使用中央处理器（具体来说，我们使用 Intel i7-6900K 进行了测试），生成脆弱性图仍只需约 3 秒。为进一步提高计算效率，我们还提供了一个利用图形处理器的实现版本；在使用配备 Nvidia Titan Xp 显卡的机器时，该版本耗时约 0.3 秒。对于 Places365 数据集，利用所有扰动图像的预计算逻辑值即可高效生成脆弱性图，这对应于算法 1 的第 9 行。对于我们选取来进行分析的 266 张图像（尺寸为 224×224），预生成过程涉及 266×224×224=13，346，816 次前向传播；使用 Nvidia Quadro RTX 8000 显卡耗时约 3 小时。完成预生成后，生成脆弱性图只需约 0.003 秒。

尽管延迟可能因机器配置而异，但这不会显著限制我们的分析。这是因为：（1）我们预计焦点图像不会频繁更新，其脆弱性图也同样如此；（2）算法 1 可以利用并行计算，而且使用配备 GPU 的机器时，卷积神经网络的前向传播时间会显著缩短；（3）我们可以轻松调整步长，以权衡准确性和延迟。例如，将步长从 1 增至 2，不会显著改变生成的脆弱性图外观，却能将计算时间缩短至原来的 1/4。由于可视分析系统的交互性是分析过程中维持用户心理地图的关键因素，因此，使用不同硬件平台上的不同卷积神经网络模型，对系统延迟进行全面分析，也是我们计划开展的未来工作之一。

重要性。 本工作主要通过基于实例的分析来理解对抗攻击。我们的方法能够有效揭示模型在所选图像对的神经元行为方面的脆弱性。由于某个图像组上的神经元行为通常可以通过该组中一个代表性实例的行为来近似，因此，基于实例的分析能够为代表性实例提供具体且直观的解释。此外，我们的方法支持对特定神经元（例如带隙最大的极端神经元）进行分析，并为各种多层次分析提供补充信息。为了探索神经元簇的行为，我们设计了 神经元聚类视图，使用户能够便捷地选择相似神经元，并通过其感受野的交集和并集查看神经元组的行为。此外，我们在 感受野视图 中将排序后的上下文图像进行可视化，以支持从所选图像中得到的发现。我们通过多个案例研究展示了所提出的分析工作流程的适用性和普适性。未来，我们计划支持对一组输入图像上的神经元全局模式进行分析。

9 结论

在本研究中，我们提出了一种用于理解卷积神经网络中所发生对抗攻击的可视分析方法。该方法旨在回答两个问题：哪些神经元更易受到攻击，以及它们在预测过程中发挥了怎样的作用。通过扰动输入图像或神经元激活，我们将对抗攻击的影响分解到单个神经元，并提出了若干指标，以根据神经元的脆弱性对其进行排序。我们采用层次聚类对数量过多的神经元进行结构化探索，并揭示它们在良性图像/对抗图像中的行为相似性。结合这些技术创新，我们设计了一个可视分析系统，使用户能够便捷地联合探索输入图像和 CNN 神经元，并利用输入图像所携带以及 CNN 神经元所提取的语义来解释对抗攻击。通过多个案例研究和领域专家的反馈，我们验证了该可视分析方法的有效性。

致谢

本研究部分由美国国立卫生研究院通过资助项目 1R01CA270454-01 和 1R01CA273058-01 赞助，并由克努特和爱丽丝·瓦伦堡基金会通过 KAW 2019.0024 号资助提供支持。

参考文献

[1] Naveed Akhtar 和 AjmalMian。2018年。对抗攻击对计算机视觉深度学习的威胁：综述。*IEEEAccess* 6（2018年），14410–14430。[2] J. Alonso 和 Y. Chen。2009年。感受野。*Scholarpedia* 4，1（2009年），5393。[3] Mihalj Bakator 和 Dragica Radosav。2018年。深度学习与医学诊断：文献综述。*Multimodal Technologies and Interaction* 2，3（2018年）。[4] Battista Biggio 和 Fabio Roli。2018年。异常模式：对抗机器学习兴起十年之后。模式识别 84（2018年），317–331。稿件已提交至 ACM

[5] Kelei Cao、Mengchen Liu、Hang Su、Jing Wu、Jun Zhu 和 Shixia Liu。2021年。分析深度神经网络的噪声鲁棒性。IEEE Transactions on Visualization and Computer Graphics 27，7（2021年），3289–3304。[6] Jaegul Choo 和 Shixia Liu。2018年。用于可解释深度学习的可视分析。IEEE计算机图形学与应用 38，4（2018年），84–92。[7] Nilaksh Das、Haekyu Park、Zijie J Wang、Fred Hohman、Robert Firstman、Emily Rogers 和 Duen Horng Polo Chau。2020年。Bluff：交互式解读深度神经网络中的对抗攻击。载于 VIS 会议论文集。IEEE，271–275。[8] Kevin Eykholt、Ivan Evtimov、Earlence Fernandes、Bo Li、Amir Rahmati、Chaowei Xiao、Atul Prakash、Tadayoshi Kohno 和 Dawn Song。2018年。针对深度学习视觉分类的鲁棒物理世界攻击。载于 CVPR 会议论文集。[9] Ian Goodfellow、Yoshua Bengio 和 Aaron Courville。2016年。《深度学习》。麻省理工学院出版社。
http://www.deeplearningbook.org。[10] Ian Goodfellow、Jonathon Shlens 和 Christian Szegedy。2015年。解释并利用对抗样本。载于 ICLR 会议论文集。[11] Samuel Greydanus、Anurag Koul、Jonathan Dodge 和 Alan Fern。2018年。可视化并理解 Atari智能体。载于 ICML 会议论文集，第80卷。1792–1801。[12] Sorin Grigorescu、Bogdan Trasnea、Tiberiu Cocias 和 Gigel Macesanu。2020年。自动驾驶深度学习技术综述。《现场机器人学期刊》37，3（2020年），362–386。[13] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。2016年。用于图像识别的深度残差学习。载于 CVPR 会议论文集。[14] Fred Hohman、Haekyu Park、Caleb Robinson 和 Duen Horng Polo Chau。2019年。Summit：通过可视化激活值和归因摘要来扩展深度学习可解释性。IEEE Transactions on Visualization and Computer Graphics 26，1（2019年），1096–1106。[15] Andrew Ilyas、Shibani Santurkar、Dimitris Tsipras、Logan Engstrom、Brandon Tran 和 Aleksander Madry。2019年。对抗样本不是缺陷，而是特征。载于 NeurIPS 会议论文集，第32卷。[16] A. Krizhevsky。2009年。从微小图像中学习多层特征。硕士论文，多伦多大学（2009年）。[17] Shusen Liu、Zhimin Li、Tao Li、Vivek Srikumar、Valerio Pascucci 和 Peer-Timo Bremer。2019年。NLIZE：一种由扰动驱动的可视化探查工具，用于分析和解释自然语言推理模型。IEEE Transactions on Visualization and Computer Graphics 25，1（2019年），651–660。[18] Shixia Liu、Jiannan Xiao、Junlin Liu、Xiting Wang、Jing Wu 和 Jun Zhu。2017年。树提升方法的可视诊断。IEEE Transactions on Visualization and Computer Graphics 24，1（2017年），163–173。[19] Jonathan L Long、Ning Zhang 和 Trevor Darrell。2014年。卷积网络会学习对应关系吗？载于 NIPS 会议论文集，第27卷。[20] Yuxin Ma 和 Ross Maciejewski。2020年。使用局部线性片段对类别分离进行可视分析。IEEE Transactions on Visualization and Computer Graphics 27，1（2020年），241–253。[21] Yuxin Ma、Tiankai Xie、Jundong Li 和 Ross Maciejewski。2019年。通过可视分析解释对抗机器学习中的脆弱性。IEEE Transactions on Visualization and Computer Graphics 26，1（2019年），1075–1085。[22] Aleksander Madry、Aleksandar Makelov、Ludwig Schmidt、Dimitris Tsipras 和 Adrian Vladu。2018年。迈向能够抵御对抗攻击的深度学习模型。载于 ICLR 会议论文集。[23] Xueyan Mei、Hao-Chih Lee、Kai-yue Diao、Mingqian Huang、Bin Lin、Chenyu Liu 等。2020年。人工智能助力新冠肺炎患者快速诊断。《自然·医学》26，8（2020年），1224–1228。[24] Alexander Mordvintsev、Christopher Olah 和 Mike Tyka。2015年。DeepDream——用于可视化神经网络的代码示例。Google Research 2，5（2015年）。
http://googleresearch.blogspot.co.uk/2015/07/deepdream-code-example-for-visualizing.html[25] Daniel Müllner。2011年。现代层次凝聚聚类算法。arXiv:1109.2378（2011年）。[26] Anh Nguyen、Alexey Dosovitskiy、Jason Yosinski、Thomas Brox 和 Jeff Clune。2016年。通过深度生成网络合成神经网络中神经元的偏好输入。载于 NIPS 会议论文集，第29卷。3387–3395。[27] Chris Olah、Alexander Mordvintsev 和 Ludwig Schubert。2017年。特征可视化。Distill（2017年）。
https://distill.pub/2017/feature-visualization。[28] Nicolas Papernot、Patrick McDaniel、Somesh Jha、Matt Fredrikson、Z Berkay Celik 和 Ananthram Swami。2016年。深度学习在对抗环境中的局限性。载于 EuroS&P 会议论文集。372–387。[29] Nicolas Papernot、Patrick McDaniel、Xi Wu、Somesh Jha 和 Ananthram Swami。2016年。将蒸馏作为针对深度神经网络对抗扰动的防御手段。载于 SP 会议论文集。582–597。[30] Shaoqing Ren、Kaiming He、Ross Girshick 和 Jian Sun。2015年。Faster R-CNN：借助区域建议网络实现实时目标检测。载于 NIPS 会议论文集，第28卷。91–99。[31] Vera Tyrsa、Adrian Carrio、Carlos Sampedro、Alejandro Rodriguez-Ramos 和 Pascual Campoy。2017年。无人驾驶飞行器深度学习方法及应用综述。《传感器期刊》2017（2017年），3296874。[32] Laurens van der Maaten 和 Geoffrey Hinton。2008年。使用 t-SNE 可视化数据。《机器学习研究期刊》9，86（2008年），2579–2605。[33] Junpeng Wang、Liang Gou、Han-Wei Shen 和 Hao Yang。2018年。DQNViz：一种用于理解深度Q网络的可视分析方法。IEEE Transactions on Visualization and Computer Graphics 25，1（2018年），288–298。[34] Junpeng Wang、Wei Zhang、Liang Wang 和 Hao Yang。2021年。通过可视分析研究树提升模型的演变。载于 PacificVis 会议论文集。186–195。[35] Weilin Xu、David Evans 和 Yanjun Qi。2018年。特征压缩：检测深度神经网络中的对抗样本。载于 NDSS 会议论文集。[36] Sergey Zagoruyko 和 Nikos Komodakis。2016年。宽残差网络。载于 BMVC 会议论文集。[37] Matthew D Zeiler 和 Rob Fergus。2014年。可视化并理解卷积网络。载于 ECCV 会议论文集。施普林格，818–833。

[38] Xun Zhao、Yanhong Wu、Dik Lun Lee 和 Weiwei Cui。2018年。iForest：通过可视分析解读随机森林。*IEEE TransactionsonVisualizationand ComputerGraphics* 25，1（2018年），407–416。[39] Bolei Zhou、Aditya Khosla、Àgata Lapedriza、Aude Oliva 和 Antonio Torralba。2015年。目标检测器在深度场景卷积神经网络中涌现。载于*Proc. ICLR*。[40] Bolei Zhou、Agata Lapedriza、Aditya Khosla、Aude Oliva 和 Antonio Torralba。2017年。Places：用于场景识别的千万级图像数据库。*IEEE Transactions on Pattern Analysisand Machine Intelligence*（2017年）。