arXiv:2303.01052v1 [cs.LG] 2 Mar 2023

通过对抗工具变量回归揭示对抗样本中的因果特征，并为鲁棒网络进行因果接种

Junho Kim*, Byung-KwanLee*Yong Man Ro†韩国科学技术院电气工程学院
图像与视频系统实验室，韩国{arkimjh, leebk, ymro}@kaist.ac.kr

摘要

对抗样本的起源在研究领域中仍无法解释，尽管已有全面的调查研究，来自不同观点的争论依然层出不穷。本文从因果视角提出一种探究对抗训练网络意外脆弱性的方法，即对抗工具变量（*IV*）回归。通过采用该方法，我们在与未知混杂因素相分离的无偏环境下，估计对抗预测的因果关系。我们的方法旨在通过利用因果特征估计器（即假设模型）与最坏情况反事实（即检验函数）之间的零和优化博弈，揭示对抗样本中固有的因果特征；该博弈旨在扰动模型，以寻找因果特征。通过广泛分析，我们证明，估计得到的因果特征与对抗鲁棒性所需的正确预测高度相关，而反事实则呈现出与正确预测显著偏离的极端特征。此外，我们还介绍了如何将因果特征（*CAFE*）有效地注入防御网络，以提升对抗鲁棒性。

1. 引言

对抗样本对人类观察者而言难以察觉，却能恶意欺骗深度神经网络（DNNs），由于其被用于危害机器学习系统的安全威胁，已引起研究领域的广泛关注。在现实环境中，这类潜在风险削弱了 DNNs 决策过程的可靠性，也引发了在安全关键领域采用 DNNs 的疑问 [4, 57, 65]。

为理解对抗样本的起源，已有不少研究从多种视角广泛探讨了对抗脆弱性，例如超平面中过度的线性特征 [25]，统计波动的偏差 [58, 62]，以及由频率

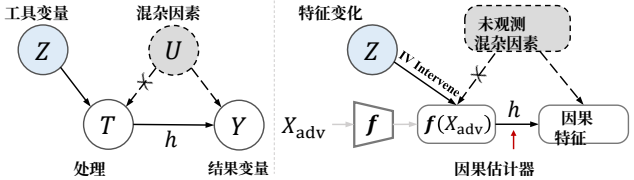


图 1。带有 IV 的数据生成过程（DGP）。通过部署 Z，可以在未知混杂因素 U的外生条件下，估计处理 T 与结果变量Y之间的因果关系。

信息引发的现象 [72]。近期，若干研究 [33, 34] 揭示了经过对抗训练的网络中鲁棒特征和非鲁棒特征的存在及其普遍性，并指出，对抗样本上的非鲁棒特征可能导致意料之外的错误分类。

尽管进行了广泛的探索 [31, 63]，关于对抗样本潜在成因的观点仍未形成共识 [21]。这是因为早期工作主要关注在对抗训练 [41, 53, 66, 71, 76]，这一经典监督学习方案中，对抗样本与目标标签之间关联性的分析。这类分析很容易在所学习的关联中诱发虚假相关（即，统计偏差），因此，在可能存在偏见的观点（例如，过度线性、统计波动、频率信息和非鲁棒特征）时，无法解释对抗脆弱性的真正来源。为了从因果视角阐明对抗脆弱性从何而来，并推导出真实的对抗因果关系，我们需要采用一种面向干预的方法（即，因果推断），针对给定的对抗样本数据总体，通过估计因果关系来超越仅仅分析关联性。

当随机对照试验（A/B 实验）或对未知混杂因素进行全面控制都不可行时，工具变量（IV）回归是因果推断的一种高效工具。它是计量经济学中用于识别因果关系的常用方法 [13, 15, 46]，，能够提供一个不受未知混杂因素影响的无偏环境，从而避免因果推断中的内生性 [54]。在工具变量回归中，工具变量用于

*共同贡献。 †通讯作者。

通过分离处理中的外生部分，消除由未知混杂因素造成的后门路径。为了更好地理解，我们可以举例说明如何寻找 [9] 教育 T 与收入Y之间的因果关系，如图1所示。仅仅测量这两个变量之间的相关性，并不能说明二者存在因果关系，因为可能存在未知混杂因素 U（例如，个人能力、家庭背景等）。理想情况下，对 U 进行条件化是识别因果关系的最佳方法，但我们无法控制未观测变量。David Card [9]将大学邻近性视为工具变量 Z，它与教育直接相关 T，但直觉上与收入无关 Y。通过为 Z赋予外生部分，可以提供一个不受 U影响的无偏环境，从而识别 T 与 Y之间真实的因果关系。

具体而言，一旦将数据生成过程（DGP）视为因果推断中的 [52] 如图 1 所示，未知混杂因素的存在 U 可能会产生虚假相关，形成阻碍因果估计器 h（即，假设模型）估计处理 T 与结果变量 Y（$T \leftarrow U \rightarrow Y$）之间因果关系的后门路径。通过采用工具变量 Z，我们可以从 h 中获取真实因果关系的估计目标，并处于无偏状态（$Z \rightarrow T \rightarrow Y$）。将此类数据生成过程引入对抗设置后，上述存在争议的观点（例如，过度线性、统计波动、频率信息以及非鲁棒特征）可以被视为未知混杂因素的可能候选 U，从而揭示对抗性起源。在大多数观测性研究中，实践中一切因素都是内生的，因此我们无法明确指定所有混杂因素，也无法在对抗设置中对其进行全面控制。因此，鉴于即使存在未知混杂因素，工具变量回归仍具备因果推断能力，我们引入工具变量回归这一强大的因果方法，以揭示对抗性起源。

在对抗设置中，未知混杂因素 U 很容易导致对抗来源的解释模糊不清，从而在对抗样本与其目标标签之间产生虚假相关。为了揭示对抗因果关系，我们首先需要对网络 f提取的中间特征表示进行干预，并关注真正影响对抗鲁棒性的因素，而不受未知混杂因素 U的影响，而非模型预测。为此，我们将工具变量 Z 定义为深度神经网络的特征空间中对抗样本与自然样本之间的特征变化，其中这种变化 Z源于图像域中的对抗扰动，因此 Z 会为给定的自然特征生成对抗特征 T。请注意，将 Z 视为工具变量是合理的选择，因为若没有自然特征，单独的特征变化并不能为对抗预测提供相关信息。接下来，一旦我们在对抗样本上找到因果相关特征表示，就将其命名为因果特征 Y，它们能够即使存在对抗扰动，也能增强目标标签预测的鲁棒性，如图1所示。

本文提出对抗工具变量（*IV*）回归，以识别与对抗预测因果关系相关的对抗样本上的因果特征。我们的方法在对抗设置中为未知混杂因素构建无偏环境 U，并采用广义矩估计法（GMM）估计对抗样本中固有的因果特征 [27]。广义矩估计法是一种灵活的非参数工具变量回归估计方法。与对抗学习的本质相似 [5, 24],, 我们部署了一个零和优化博弈 [19, 40]，由假设模型和检验函数构成：前者试图揭示处理与结果变量之间的因果关系，后者则干扰假设模型对该关系的估计。在对抗设置中，我们将假设模型视为因果特征估计器，它从对抗特征中提取与正确预测高度相关的因果特征，以增强对抗鲁棒性；而检验函数则构造最坏情况下的反事实（即，极端特征），迫使因果特征的估计目标显著偏离正确预测。因此，这还能进一步增强假设模型，使其揭示对抗样本中的因果特征。

通过广泛分析，我们证实，在对抗样本上估计出的因果特征与对抗鲁棒性所需的正确预测高度相关，而检验函数则表示对抗样本上的最坏情况反事实。通过特征可视化 [42, 49], ，我们以人类可识别的方式解读对抗样本上的因果特征。此外，我们提出对估计出的因果特征进行求逆，以便在可能的特征界限内处理这些特征，并介绍一种将这些因果特征 *(CAFE)*高效注入防御网络、从而提升对抗鲁棒性的方法。

2. 相关工作

在因果推断的漫长发展历程中，已有大量研究 [23, 26, 35] 探讨因果知识如何影响决策过程。在各种因果方法中，尤其是在经济学领域，工具变量回归 [54] 提供了一种方法，使人们即使在存在未知混杂因素的情况下，也能识别感兴趣的处理与结果变量之间的因果关系；工具变量使处理满足外生条件，从而为因果推断提供无偏环境。

早期的工具变量回归研究 [2, 3] 通过将因果变量之间的关系形式化为线性函数，对其作出了限制，这种方法被称为两阶段最小二乘估计量 [70]。随着机器学习方法不断发展，研究人员和数据科学家希望将其用于非参数学习 [12,13,15,46]，并克服变量之间函数关系中的线性约束——

。作为两阶段最小二乘法的扩展，DeepIV [28], KernelIV [60],和 Dual IV [44]将深度神经网络作为非参数估计量，并提出了有效利用深度神经网络进行工具变量回归的方法。近来，研究人员巧妙地提出了广义矩估计法（GMM） [7, 19, 40]，通过零和优化处理高维处理变量上的非参数假设模型，从而成功实现了非参数工具变量回归。

与各种利用工具变量的因果方法并行，揭示对抗样本的起源是一个引发争议的开放研究问题。最初，[25]认为网络超平面中过度的线性会导致对抗脆弱性。多项研究 [58, 62]从理论上分析了这种起源，认为它是数据总体统计波动的结果，或与输入中的频率信息有关 [72]。近期，深度神经网络中非鲁棒特征的存在 [33, 34] 被视为对抗样本的主要成因，但这一现象仍无法解释 [21]。

受工具变量回归的启发，我们提出一种方法，用于估计对抗特征中固有的因果特征；这些对抗特征很容易引发深度神经网络的脆弱性。为此，我们在假设模型与检验函数之间采用基于广义矩估计法的零和优化 [7,19,40]。在这里，我们将因果特征估计器的角色赋予假设模型，并将生成最坏情况反事实的角色赋予检验函数；后者会干扰因果特征的发现。这一策略能够学习因果特征，从而克服各种对抗扰动带来的所有考验与挑战。

3. 对抗性工具变量回归

我们的主要目标是通过采用工具变量回归，估计与对抗鲁棒性中正确预测高度相关的对抗样本上的内在因果特征。在识别因果特征之前，我们首先对工具变量回归的问题设定进行说明，并使用广义矩估计法（GMM）重新审视非参数工具变量回归。

问题设定。 我们从条件矩约束（CMR）开始 [1, 11] 引入一种利用工具变量的渐近有效估计方法，该方法减少了由未知混杂因素 U [50] （见图 1 中的关系）所导致的处理 T与所关注结果变量 Y之间的虚假相关（即， 统计偏差）。这里，CMR 的形式可以写成带有假设模型 h（即假设空间上的因果估计器） $\mathcal{H}$ 的形式，如下所示:

$$\mathbb{E}_T[\psi_T(h) \mid Z] = \mathbf{0}, \tag{1}$$

其中 ψ_T： $\mathcal{H} \to \mathbb{R}^d$ 表示处理 [13] 上的广义残差函数 T，使其表示 $\psi_T(h) = Y - h(T)$ 作为回归任务的结果误差。请注意， $\mathbf{0} \in \mathbb{R}^d$ 描述零向量，而 d 表示感兴趣的结果变量 Y 的维度，同时也等于假设模型 h 的输出向量的维度。通过工具变量，控制处理 [48] 为外生变量。此外，对于给定的工具变量 Z，最小化广义残差函数ψ 的幅度，渐近地意味着限制假设模型 h 不偏离 Y，从而消除由混杂因素 U 所诱导的后门路径上， h 的内部虚假相关性。

3.1. 重新审视非参数工具变量回归

一旦我们找到满足带有工具变量 Z 的 CMR 的假设模型 h，就可以按照以下形式，利用 h 进行工具变量回归以尝试进行因果推断：$\mathbb{E}_T[h(T) \mid Z] = \int_{t\in T} h(t)\mathrm{d}\mathbb{P}(T = t \mid Z)$，其中 $\mathbb{P}$ 表示条件密度测度。事实上，两阶段最小二乘法（2SLS） [2, 3, 70] 是扩展工具变量回归的一种众所周知的求解方法，但它无法直接应用于更复杂的模型（如非线性模型），因为 2SLS 的设计目标是假设模型 [51] 为线性模型。随后， [28] 和 [60] 分别通过使用条件均值嵌入和高斯混合，引入了适用于非线性模型的广义 2SLS。然而，当采用高维处理 T（例如深度神经网络（DNNs））上的非参数假设模型 h 时，它们仍会引发病态问题，从而产生有偏估计 [7,19,44,77]。这源于两阶段方法的诅咒性质，根据 Vapnik 原则 [16]，这种现象被称为禁止回归 [3]：“不要将一个更一般的问题作为中间步骤来求解”。

为解决这一问题，近期研究 [7, 19, 40] 采用广义矩估计法（GMM）发展工具变量回归，并成功实现了能够缓解有偏估计的单阶段回归。一旦我们选择一个矩，以相对于假设模型及其反事实的通用结果误差来表示该矩，GMM便会利用该矩向假设模型施加无穷多个矩约束，超越CMR的简单约束。展开式（1），GMM的公式可写为一个矩，记作 m： $\mathcal{H} \times \mathcal{G} \to \mathbb{R}$ 具体如下（见附录A）：

$$\begin{aligned} m(h, g) &= \mathbb{E}_{Z,T}[\psi_T(h) \cdot g(Z)] \\ &= \mathbb{E}_Z[\underbrace{\mathbb{E}_T[\psi_T(h) \mid Z]}_{\text{CMR}} \cdot g(Z)] = 0, \end{aligned} \tag{2}$$

其中，算子 $\cdot$ 表示内积，而 $g \in \mathcal{G}$表示检验函数，它在检验函数空间上生成无穷多个矩约束方面发挥作用 $\mathcal{G}$，其输出的维度为 $\mathbb{R}^d$。任意向量值函数所表示的无穷多个检验函数 $\{g_1, g_2, \cdots\} \in \mathcal{G}$揭示了可能存在的矩约束（即。，经验反事实） [8] 违反了公式（2）。换言之，它们使人们能够轻松捕捉工具变量中最糟糕的部分，这些部分容易导致假设模型 h产生有偏估计，从而有助于获得更真实的因果

关系，以便通过考虑所有可能的反事实情形 h 实现泛化 g。

然而，它也有一个类似的局限：由于我们只能处理有限个可观测的检验函数，因此无法应对无限矩。因此，近期研究构建了最大矩约束 [19,43,77]，通过仅关注 IV 的极端部分来高效处理无限矩，该部分记作 $\sup_{g\in\mathcal{G}} m(h,g)$，并以闭式表达式表示。通过这种方式，我们可以同时最小化假设模型的矩，使其充分满足针对检验函数的最坏情况泛化性能。因此，GMM 可以改写为一个最小-最大优化问题，可将其视为假设模型 h与检验函数 g之间的零和博弈：

$$\min_{h\in\mathcal{H}} \sup_{g\in\mathcal{G}} m(h,g) \approx \min_{h\in\mathcal{H}} \max_{g\in\mathcal{G}} \mathbb{E}_{Z,T}[\psi_T(h)\cdot g(Z)], \quad (3)$$

其中，无限多个检验函数可以替换为以深度神经网络形式表示的非参数检验函数。接下来，我们将式（3）中的 GMM 推广至对抗场景，并通过建立带最大矩约束的对抗性工具变量回归，揭示其对抗性起源。

3.2. 揭开对抗因果特征的奥秘

为了揭示对抗样本中固有因果特征的奥秘，我们首先将特征变化 Z 定义为工具变量，它可以用对抗训练的深度神经网络 f 表示如下：

$$Z = f_l(X_\epsilon) - f_l(X) = F_{\text{adv}} - F_{\text{natural}}, \quad (4)$$

其中，f_l 在 l^{th} 中间层输出特征表示，X 表示自然输入，X_ϵ 表示带有对抗扰动 ϵ 的对抗样本，使得 $X_\epsilon = X + \epsilon$。为了探究对抗特征 F_{adv} 如何真正估计我们关注的因果特征 Y（即结果变量），我们将处理设为 $T = F_{\text{adv}}$，并通过检验函数将反事实处理设为 $T_{\text{CF}} = F_{\text{natural}} + g(Z)$。

请注意，如果我们天真地将检验函数 g 应用于对抗性特征 T，以构造反事实处理 T_{CF}，使得 $T_{\text{CF}} = g(T)$，那么考虑到深度神经网络 f 的特征界限，假设模型 $h(T_{\text{CF}})$ 的输出（即，因果特征）可能并非可获取的特征。换言之，如果我们在估计因果特征时不保留自然特征，那么估计得到的因果特征就会与自然特征过于疏离，成为排他性过强的特征。这会导致得到不适用的特征，将其视为我们无法处理的虚构特征，因为估计得到的因果特征仅是在深度神经网络某个特定中间层中经过显著操纵的特征。因此，我们将反事实处理设为 $T_{\text{CF}} = F_{\text{natural}} + g(Z)$。这是因为上述构造能够保留自然特征：我们首先从反事实处理中减去自然特征，使得满足 $T' = T_{\text{CF}} - F_{\text{natural}} = g(Z)$，再将假设模型的输出 Y' 加到自然特征上，以恢复因果特征，使得 $Y = Y' + F_{\text{natural}} = h(T') + F_{\text{natural}}$。简言之，我们有意转换因果特征和反事实处理，使其不偏离可能的特征界限。

现在，我们将包含通过检验函数计算出的对抗样本反事实的 对抗矩限制（*AMR*） 定义如下：$\mathbb{E}_{T'}[\psi_{T'}(h) \mid Z] = \mathbf{0}$。这里，对抗设置下的广义残差函数 $\psi_{T'|Z}(h) = Y' - h(T')$ 使用转换后的因果特征 Y'。将这些内容结合起来，我们重新表述带有反事实处理的广义矩估计法，以拟合对抗工具变量回归，可写为（注意，h 和 g 由简单的卷积神经网络结构组成）：

$$\min_{h\in\mathcal{H}} \max_{g\in\mathcal{G}} \mathbb{E}_Z[\underbrace{\mathbb{E}_{T'}[\psi_{T'}(h) \mid Z]}_{\text{AMR}} g(Z)] = \mathbb{E}_Z[\psi_{T'|Z}(h) g(Z)], \quad (5)$$

其中满足 $\mathbb{E}_{T'}[\psi_{T'}(h) \mid Z] = \psi_{T'|Z}(h)$，因为 Z仅对应一个转换后的反事实处理$T' = g(Z)$。在 AMR 中，我们无法直接计算广义残差函数 $\psi_{T'|Z}(h) = Y' - h(T')$，因为在高维特征空间中，转换后的因果特征 Y' 没有可观测标签。相反，我们利用分类任务中与自然输入 X 对应的独热向量目标标签 $G \in \mathbb{R}^K$ (K : 类别数)。为使用该标签，我们通过对数似然函数将 GMM 的计算域从特征空间改为模型预测的对数似然空间：$\Omega(\omega) = \log f_{l+}(F_{\text{natural}} + \omega)$，其中 f_{l+} 表示在 l^{th}中间层之后返回分类概率的后续网络。因此，我们进一步将因果推断的含义明确为：即使面对最坏情况反事实，也要找到能够正确预测目标标签的内在因果特征。为此，我们对式 (5) 进行了修改，将矩投影到对数似然空间，如下所示：

$$\begin{aligned}\min_{h\in\mathcal{H}} \max_{g\in\mathcal{G}} &\mathbb{E}_Z[\psi^{\Omega}_{T'|Z}(h)\cdot(\Omega\circ g)(Z)] \\ &= \mathbb{E}_Z[\{G_{\log} - (\Omega\circ h)(T')\}\cdot(\Omega\circ g)(Z)],\end{aligned} \quad (6)$$

其中，$\psi^{\Omega}_{T'|Z}(h)$表示对数似然空间上的广义残差函数，算子 $\circ$ 表示函数复合，$G_{\log}$ 是对数目标标签，使其满足 $G_{\log} = \log G$。对数目标标签的每个元素（$k = 1, 2, \cdots, K$）在为 $G^{(k)} = 1$ 时具有 $G^{(k)}{}_{\log} = 0$，在为 $G^{(k)} = 0$时具有$G^{(k)}{}_{\log} = -\infty$。实现时，我们只需忽略元素 $G^{(k)}{}_{\log} = -\infty$，仅使用另一个元素。

到目前为止，我们基于式（6）中的 AMR 构建 GMM，即 *AMR-GMM*，以实现对抗性工具变量回归。然而，该方法并未对检验函数进行显式正则化，因此理想矩与经验矩之间会产生泛化差距（见附录 B）。

因此，它可能违反检验函数的特征界限，并导致因果推断中的预测失衡（见图 4）。为构造丰富检验函数，先前的工作 [7, 19, 40, 67] 采用了 *Rademacher* 复杂度 [6, 36, 73]，为函数族提供了紧致的泛化界。它具有坚实的理论基础，可用于控制泛化差距，因此与深度神经网络中使用的各种正则化器相关，例如权重衰减、Lasso、Dropout 和 Lipschitz [20, 64, 68, 75]。在 AMR-GMM 中，它的作用是使检验函数能够在对抗特征界内找到最坏情况反事实。根据附录 B，我们构建了采用丰富检验函数的 AMR-GMM 最终目标函数，如下所示：

$$\min_{h\in\mathcal{H}}\max_{g\in\mathcal{G}} \mathbb{E}_Z[\psi^{\Omega}_{T'|Z}(h)\cdot(\Omega\circ g)(Z)]-|\mathbb{E}_Z[Z-g(Z)]|^2. \quad (7)$$

受篇幅限制，AMR-GMM 算法的更多细节请参见附录 D。

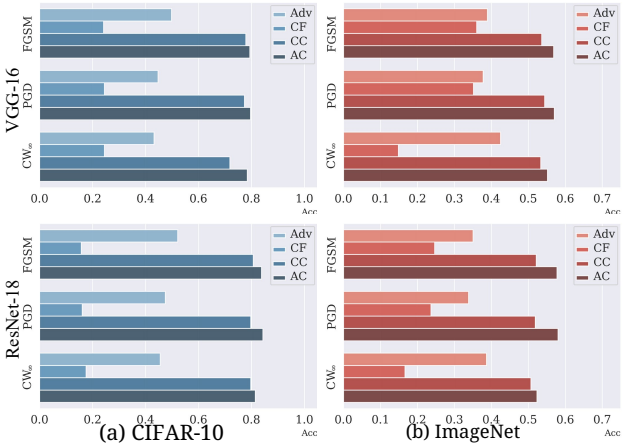
4. 因果特征的性质分析

本节首先将通过 AMR-GMM 进行对抗性工具变量回归得到的特征表示的若干合取记为：(i) 对抗特征 (Adv)： $F_{\text{natural}} + Z$，(ii) 反事实特征 (CF)：$F_{\text{natural}} + g(Z)$，(iii) 反事实因果特征 (CC)：$F_{\text{natural}} + (h \circ g)(Z)$，以及 (iv) 对抗因果特-征 (AC)： $F_{\text{natural}} + h(Z)$。利用这些合取，我们通过分类准确率估计对抗鲁棒性；上述特征合取经由 f_{l+}传播，其中标准攻击会产生特征变化 Z 和对抗特征 T。请注意，所有特征表示均按照 [34], 所述，在深度神经网络 f 的最后一个卷积层进行处理，因为该层主要包含高层对象概念，并且由于高阶交互作用 [17]，容易受到对抗扰动的意外影响。这里，用合取的对抗鲁棒性替代平均处理效应 (ATE) [32],，后者通常用于因果方法的验证。

4.1. 验证假设模型和检验函数

使用 AMR-GMM 优化用于对抗性工具变量回归的假设模型和检验函数后，我们便可以在对抗环境中控制内生处理（即），对抗特征，并从中分离出外生部分，即因果特征。在这里，假设模型能够在对抗样本中找到因果特征；即使面对

对抗扰动。另一方面，测试函数生成最坏情况反事实，以干扰因果特征的估计，从而削弱假设模型的能力。这些学习策略使假设模型能够估计内在因果特征，克服反事实带来的所有艰难险阻。因此，



(a) CIFAR-10 (b) ImageNet

图2. 在三种攻击模式下，VGG-16 和 ResNet-18 的 Adv、CF、CC、AC 对抗鲁棒性：FGSM [25], PGD [41],Carlini-Wagner攻击 $_\infty$ [10]在CIFAR-10 [37]和ImageNet [18]上。

AMR-GMM 的本质为对抗样本上的因果特征发现提供了理论依据，以克服各种类型的对抗扰动。请注意，我们的工具变量设置提出了齐性假设 [30], ，即比单调性假设 [2], 更一般的版本：无论自然特征如何随数据样本而变化，所有数据样本的对抗鲁棒性（即., 平均处理效应）始终保持较高水平 F_{natural}。

如图 2所示，我们通过测量所有数据集样本的特征合取（即。，Adv、CF、CC、AC）的分类准确率，深入考察假设模型和检验函数的平均处理效应（即。，对抗鲁棒性）。我们观察到，CF 的对抗鲁棒性不如 CC、AC，甚至不如 Adv。直观而言，这一结果显而易见，因为违反式（7）的检验函数会迫使特征表示处于最差状态，极度偏离正确预测。CC 和 AC 的预测结果表现出优异的鲁棒性，相较 Adv 具有较大的间隔。由于 AC 直接利用从对抗扰动中获得的特征变化，因此其对抗鲁棒性优于 CC；后者由检验函数输出基于特征变化的最坏情形反事实结果而得到。值得注意的是，我们发现，即使在高置信度对抗攻击 [10]中伪造未见过的扰动，假设模型得到的两种结果通常仍表现出稳定的鲁棒性。这种鲁棒性表明，估计得到的因果特征能够克服各种类型的对抗扰动。

4.2. 因果效应的解读与可视化结果

我们已经在对抗样例中找到了因果特征，并分析了其鲁棒性。接下来，我们要探讨的问题是“因果特征本身是否包含目标对象的语义信息？”。近期研究 [22, 34, 39]

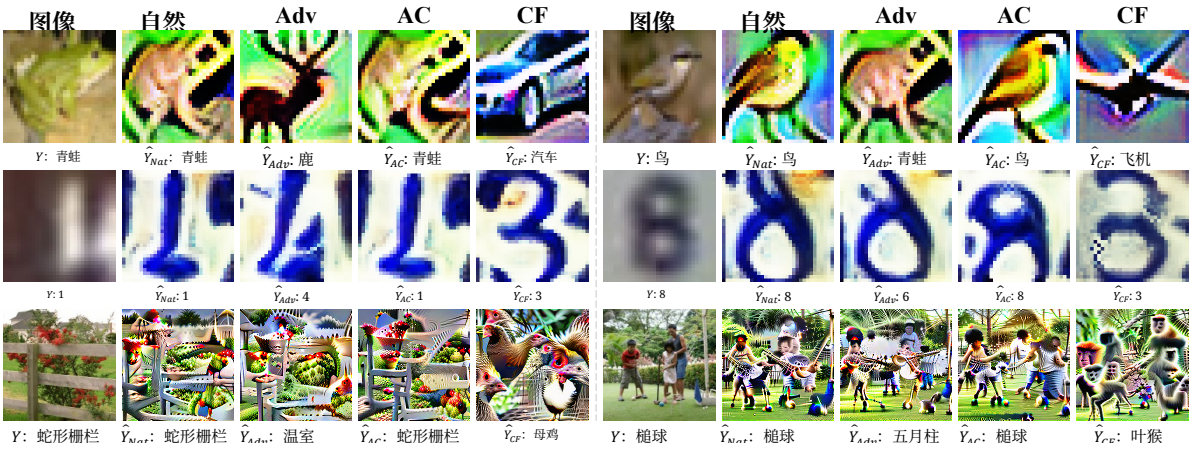


图 3. 自然特征、Adv、AC 和 CF 的特征可视化结果。从上至下，依次使用 CIFAR-10、SVHN 和 ImageNet 进行特征视觉解读。

探究对抗设置下特征表示的语义含义。我们也采用特征可视化方法 [42, 47, 49] ，在输入域上以人类可识别的方式解释特征合取。如图3所示，我们通常可以观察到，自然特征的结果呈现了目标对象的语义含义。另一方面，对抗特征（Adv）会促使其特征表示朝向遭受对抗攻击的目标对象。

如前所述，检验函数会将处理干扰成最坏情况反事实，从而加剧对抗扰动引起的特征变化。因此，CF 的可视化结果明显偏向目标对象被违背的特征表示。例如，在 ImageNet [18]示例中，我们可以看到，CF 的可视化呈现出母鸡和叶猴特征，它们分别由蛇形栅栏和槌球操纵而来。我们注意到，原始图像中的红色花朵变成了红色鸡冠花和母鸡羽毛图案；此外，人也变成了具有明显叶猴特征的形象，这加剧了特征表示向最坏反事实的偏移。相比之下，AC 的可视化结果呈现出显著的表征能力和语义一致性，能够展现目标对象；我们可以直接识别其中的语义信息，且这些信息对人类观察者而言易于解释。通过考察视觉解读结果，我们发现，从假设模型和检验函数中获取的特征表示都具有因果语义信息，其作用与因果方法的理论依据一致。简言之，我们验证了因果特征的语义含义即使在高维空间中、面对反事实时仍然内在存在。

4.3. 验证工具变量设置的条件

工具变量需要满足以下三个有效条件，才能成功实现非

	VGG			ResNet			WRN		
	CIFAR	SVHN	Tiny	CIFAR	SVHN	Tiny	CIFAR	SVHN	Tiny
$f_{l+}(T)$	44.8	52.1	21.5	46.5	55.4	24.2	48.7	56.7	25.5
$f_{l+}(Z)$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ρ	0.9	0.8	0.8	0.9	0.8	0.7	0.9	0.9	0.8

表 1. 对工具变量设置三项条件的实证验证。$f_{l+}(T)$ 和 $f_{l+}(Z)$ 分别表示通过后续网络传播对抗特征 T 和特征变化 Z 所得的对抗鲁棒性模型性能（%）。最后一行表示皮尔逊相关系数： $\rho = \text{Cov}(Z,\ T)/\sigma_Z\sigma_T$。

参数化工具变量回归，基于先前的研究 [28, 44]: 与结果误差相互独立，即 $\psi \perp Z$（无混杂性），其中 ψ 表示结果误差；且不直接影响结果，即 $Z \perp Y \mid T$, ψ （排除限制），而只通过处理之间的联系影响结果，即 $\text{Cov}(Z,\ T) \neq 0$（相关性）。

对于无混杂性，许多研究 [41,53,66,71,76]提出了对抗训练，以对抗样本引起的特征变化来提升深度神经网络的鲁棒性 f，并将这种特征变化视为工具变量。在其他方面，从工具变量回归的角度来看，我们可以将其视为深度神经网络中满足 CMR 的努力 f，针对给定的特征变化 Z。与我们的因果观点一致，表1 第一行展示了对抗特征带来的对抗鲁棒性 T。因此，我们可以说，经过对抗训练的模型中的工具变量（即。，特征变化）满足无混杂性的有效条件，因此工具变量与结果误差相互独立。

对于排除限制而言，特征变化 Z 本身无法在没有自然特征的情况下为模型预测提供有启发性的信息，因为根据深度神经网络的学习本质，仅传播残差特征表示对模型预测没有影响。从实证结果来看，表1 的第二行表明， Z 无法成为对预测有帮助的表示。因此，我们的工具变量不应直接与结果变量相关，并且满足有效的排除限制条件。

	方法	CIFAR-10							SVHN							Tiny-ImageNet						
		自然	FGSM	PGD	Carlini-Wagner攻击$_\infty$	平均精度	DLR	自动攻击	自然	FGSM	PGD	Carlini-Wagner攻击$_\infty$	平均精度	DLR	自动攻击	自然	FGSM	PGD	Carlini-Wagner攻击$_\infty$	平均精度	DLR	自动攻击
VGG	ADV	**78.5**	49.8	44.8	42.6	43.2	42.9	40.7	**91.9**	64.8	52.1	48.9	48.0	48.5	45.2	**53.2**	25.3	21.5	21.0	20.2	20.8	19.6
	ADV$_{\text{CAFE}}$	78.4	**52.2**	**47.9**	**44.1**	**46.4**	**44.5**	**42.7**	91.5	**67.0**	**55.3**	**50.0**	**51.3**	**49.6**	**46.1**	52.6	**26.0**	**22.8**	**22.1**	**21.8**	**22.0**	**21.0**
	TRADES	**79.5**	50.4	45.7	43.2	44.4	42.9	41.8	**91.9**	66.4	53.6	49.1	49.1	47.7	45.2	**52.8**	25.9	22.5	21.9	21.5	21.8	20.7
	TRADES $_{\text{CAFE}}$	77.0	**51.6**	**47.9**	**44.0**	**47.0**	**43.9**	**42.7**	90.3	**67.8**	**56.1**	**50.0**	**53.6**	**49.1**	**47.5**	52.1	**26.5**	**23.6**	**22.6**	**22.5**	**22.6**	**21.6**
	MART	**79.7**	52.4	47.2	43.4	45.5	43.8	42.0	**92.6**	66.6	54.2	47.9	49.6	47.1	44.4	**53.1**	25.0	21.5	21.2	20.4	21.0	19.9
	MART$_{\text{CAFE}}$	78.3	**54.2**	**49.7**	**43.9**	**48.1**	**44.5**	**42.7**	91.3	**67.6**	**57.3**	**49.5**	**54.2**	**48.3**	**46.4**	53.0	**25.6**	**22.3**	**21.6**	**21.3**	**21.5**	**20.5**
	AWP	**78.0**	51.7	48.2	43.5	47.2	43.4	42.6	90.8	65.5	56.6	50.4	54.0	49.7	48.6	52.6	28.0	25.7	23.6	24.8	23.5	22.8
	AWP$_{\text{CAFE}}$	77.4	**54.8**	**51.4**	**44.2**	**50.2**	**44.9**	**43.5**	**91.9**	**67.9**	**58.6**	**51.2**	**55.9**	**51.1**	**49.7**	**52.9**	**28.8**	**26.4**	**24.2**	**25.6**	**24.1**	**23.4**
	HELP	**77.4**	51.8	48.3	43.9	47.3	43.9	42.9	91.2	65.8	56.6	50.9	53.9	50.2	48.8	**53.0**	28.3	25.9	23.9	25.1	23.8	23.1
	HELP$_{\text{CAFE}}$	75.6	**54.4**	**51.4**	**44.6**	**50.4**	**44.8**	**43.7**	**91.5**	**67.3**	**58.5**	**51.6**	**56.2**	**51.4**	**50.0**	52.6	**29.4**	**27.1**	**24.7**	**26.4**	**24.4**	**23.9**
ResNet	ADV	82.0	52.1	46.5	44.8	44.8	44.8	43.0	**92.8**	70.4	55.4	51.3	50.9	51.0	47.5	**57.2**	27.3	24.2	23.2	22.8	23.2	21.8
	ADV$_{\text{CAFE}}$	**82.6**	**55.9**	**50.7**	**47.6**	**49.0**	**47.7**	**46.2**	92.5	**73.6**	**58.9**	**53.8**	**54.9**	**52.6**	**49.8**	56.3	**28.6**	**25.7**	**24.7**	**24.4**	**24.6**	**23.5**
	TRADES	**83.0**	55.0	49.8	47.5	48.3	47.3	46.1	**93.2**	72.8	57.7	52.6	53.0	51.5	48.9	**56.5**	28.4	25.3	24.4	24.2	24.3	23.2
	TRADES $_{\text{CAFE}}$	80.7	**56.6**	**51.4**	**48.5**	**50.4**	**48.3**	**46.7**	91.3	**73.9**	**59.6**	**54.1**	**56.7**	**53.2**	**51.3**	54.5	**29.6**	**27.4**	**26.3**	**26.5**	**26.2**	**25.4**
	MART	**83.5**	56.1	50.1	47.1	48.3	47.0	45.5	**93.7**	74.2	58.3	51.7	53.2	50.8	47.8	**57.1**	27.4	24.2	23.2	22.9	23.2	22.2
	MART$_{\text{CAFE}}$	82.1	**57.3**	**51.9**	**48.1**	**50.2**	**48.0**	**46.2**	92.2	**74.9**	**61.0**	**53.4**	**57.3**	**51.8**	**49.7**	55.9	**28.6**	**25.9**	**24.6**	**24.7**	**24.5**	**23.5**
	AWP	81.2	55.3	51.6	48.0	50.5	47.8	46.9	92.2	71.1	59.8	54.3	56.8	53.6	52.0	56.2	30.5	28.5	26.2	27.6	26.2	25.5
	AWP$_{\text{CAFE}}$	**81.5**	**57.8**	**54.2**	**49.4**	**52.9**	**49.0**	**47.8**	**93.4**	**74.0**	**60.9**	**55.0**	**57.8**	**54.8**	**52.7**	**56.6**	**31.4**	**29.2**	**27.1**	**28.4**	**27.0**	**26.5**
	HELP	80.5	55.8	52.1	48.4	51.1	48.5	47.4	92.6	72.0	59.8	54.4	56.6	53.9	52.0	**56.1**	31.0	28.6	26.3	27.7	26.3	25.7
	HELP$_{\text{CAFE}}$	**80.6**	**57.8**	**54.5**	**49.4**	**53.1**	**49.5**	**48.5**	**92.9**	**73.9**	**61.3**	**55.3**	**58.8**	**54.6**	**52.8**	55.4	**32.0**	**29.7**	**27.4**	**29.2**	**27.8**	**27.3**
WRN	ADV	84.3	54.5	48.7	47.8	47.0	47.9	45.6	**94.0**	71.8	56.7	53.2	51.9	52.8	49.0	**60.9**	29.8	25.5	25.8	24.2	26.0	23.9
	ADV$_{\text{CAFE}}$	**85.7**	**58.5**	**53.3**	**51.3**	**51.8**	**51.5**	**49.5**	93.7	**75.7**	**59.1**	**54.9**	**54.0**	**54.1**	**50.2**	60.6	**31.1**	**27.3**	**27.2**	**25.8**	**27.4**	**25.4**
	TRADES	**86.3**	57.1	52.1	50.8	50.6	50.7	49.0	**93.8**	74.0	58.1	53.9	53.0	53.4	49.9	**60.8**	30.5	26.4	26.7	25.0	26.8	24.6
	TRADES $_{\text{CAFE}}$	83.7	**58.6**	**54.5**	**52.0**	**53.2**	**52.0**	**50.1**	92.4	**75.6**	**61.0**	**55.7**	**58.0**	**58.0**	**53.0**	60.3	**31.7**	**28.2**	**28.3**	**27.0**	**28.5**	**26.5**
	MART	**86.5**	58.5	52.6	50.0	50.7	49.9	48.0	**94.2**	75.0	58.0	53.1	52.8	52.8	48.9	**60.7**	29.9	25.6	25.9	24.0	25.5	23.6
	MART$_{\text{CAFE}}$	85.7	**59.8**	**54.6**	**51.4**	**52.7**	**50.9**	**49.3**	93.0	**76.5**	**61.9**	**54.9**	**57.2**	**53.8**	**50.7**	60.4	**31.2**	**27.5**	**26.8**	**25.5**	**27.0**	**25.1**
	AWP	83.7	58.0	54.7	51.3	53.7	51.2	50.1	93.2	73.4	60.8	55.9	57.5	55.5	53.6	**61.9**	35.5	32.8	31.0	31.6	31.1	29.6
	AWP$_{\text{CAFE}}$	**84.6**	**60.6**	**56.9**	**52.4**	**55.5**	**52.3**	**51.1**	**94.2**	**76.9**	**62.7**	**57.5**	**59.2**	**57.1**	**54.6**	61.4	**36.6**	**34.2**	**32.3**	**33.2**	**32.5**	**30.8**
	HELP	**83.8**	58.6	54.9	51.6	53.8	51.6	50.3	93.5	73.4	60.8	56.5	57.6	56.1	54.0	**61.8**	35.9	33.0	31.3	31.8	31.3	29.8
	HELP$_{\text{CAFE}}$	83.1	**60.5**	**57.1**	**52.7**	**56.0**	**52.6**	**51.3**	**94.0**	**76.6**	**62.6**	**57.7**	**58.8**	**57.2**	**55.0**	61.1	**37.0**	**34.7**	**32.6**	**33.8**	**32.8**	**31.2**

表2. 在五种防御基线（ADV、TRADES、MART、AWP、HELP）上比较对抗鲁棒性以及CAFE带来的提升。这些防御基线分别使用VGG-16、ResNet-18和WideResNet-34-10，在三个数据集上针对六种攻击进行训练：FGSM、PGD、Carlini-Wagner攻击$_\infty$、平均精度、DLR和自动攻击。

对于相关性而言，当考察对抗特征的估计过程时 T 使得 $T = Z + F_{\text{natural}}$，特征变化 Z 明确地对 T产生了因果影响。这是因为，在我们的工具变量设置中，处理 T 是在给定自然特征 F_{natural}的情况下，由工具变量 Z 直接估计得到的。通过使用所有数据样本，我们实证计算了皮尔逊相关系数，以证明二者之间存在高度相关的联系，具体如表1的最后一行所示。因此，我们的工具变量满足相关性条件。

5. 通过接种因果特征提升鲁棒性

接下来，我们将解释如何高效地将因果特征植入各种防御网络，从而构建鲁棒网络。为了消除网络从对抗样本中形成的虚假相关，我们能够想到的最简单方法是利用假设模型来增强鲁棒性。然而，这里存在一个现实障碍：只有当我们已经在推理阶段识别出哪些是自然输入及其对抗样本时，该方法才有效。因此，直接利用假设模型来提升鲁棒性并不可行。

为了解决这一问题，我们引入因果特征的反演（即，因果反演），以反映输入域中的这些特征。对于给定的对抗样本，它利用前置子网络的网络参数，在允许的特征界限内良好地表示因果特征 f_l。事实上，因果特征会被假设模型h在中间层操纵，因此无法保证这些特征处于可能的特征界限内。随后，因果反演成为解决这一问题的关键，同时不会过多损害因果预测；其形式可通过因果扰动表示，并采用KL散度作为距离度量 $\mathcal{D}_{\text{KL}}$：

$$\delta_{\text{causal}} = \underset{\|\delta\|_\infty \le \gamma}{\arg\min}\, \mathcal{D}_{\text{KL}}\left(f_{l+}(F_{\text{AC}}) \,||\, f(X_\delta)\right), \tag{8}$$

其中， F_{AC} 表示由假设模型 h提炼出的对抗因果特征，δ_{causal} 表示用于实现因果反演 X_{causal} 的因果扰动，使得 $X_{\text{causal}} = X + \delta_{\text{causal}}$。请注意，为避免在生成因果反演 X_{causal}时损害自然输入的信息，我们将扰动 δ 限制在 l_∞ 的 γ-球内，也就是扰动预算范围内，使其达到人类无法察觉的程度，从而满足 $\|\delta\|_\infty \le \gamma$。附录C展示了因果特征模型预测的置信度分数与因果反演、自然输入和对抗样本相比时的统计距离。只要能够利用因果反演处理因果特征，使得 $F_{\text{AC}} = f_l(X_{\text{causal}})$，我们就可以进一步探讨如何通过小规模扰动 ϵ，以经验风险最小化（ERM）的形式将因果特征（*CAFE*）接种到防御网络中，如下所示：

$$\min_{f\in\mathcal{F}} \mathbb{E}_{\mathcal{S}}\left[\max_{\|\epsilon\|_\infty \le \gamma} \mathcal{L}_{\text{Defense}} + \mathcal{D}_{\text{KL}}(f_{l+}(\hat{F}_{\text{AC}}) \,||\, f_{l+}(F_{\text{adv}}))\right], \tag{9}$$

其中， $\mathcal{L}_{\text{Defense}}$ 表示预先定义的损失函数，例如 [41, 53, 66, 71, 76]，用于在网络参数空间 f 上得到防御网络 $\mathcal{F}$；$\mathcal{S}$ 表示数据样本，使得$(X,\ G)\ \sim \mathcal{S}$。其余项表示一种因果正则化器，充当因果接种，使对抗特征F_{adv} 同化因果特征 F_{AC}。具体而言，$\mathcal{L}_{\text{Defense}}$ 通过对抗样本增强网络参数的鲁棒性，而正则化器则有助于确保对抗特征不超出因果特征的可能范围，从而使网络获得与未知混杂因素相分离、后门路径更少的特征。CAFE训练算法的更多细节见附录E。

6. 实验

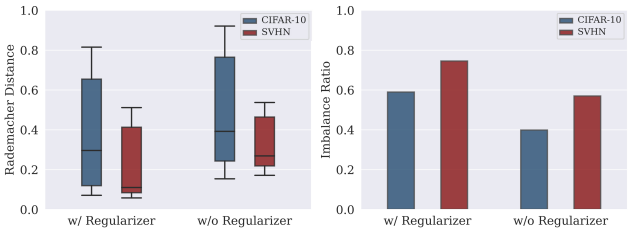
6.1. 实现与实验细节

我们在三个数据集和三种网络上开展了全面实验，以验证其在各种条件下的泛化能力。对于数据集，我们选取了低维数据集：CIFAR-10 [37]，SVHN [45]，以及高维数据集：Tiny-ImageNet [38]。为了训练这三个数据集，我们采用标准网络：VGG-16 [59], ResNet-18 [29], 以及先进的大型网络：WideResNet-34-10 [74]。

对于攻击，我们对 CIFAR-10、SVHN 使用扰动预算 $8/255$ ，对 Tiny-ImageNet 使用 $4/255$，并采用两种标准攻击：FGSM [25], PGD [41], ，以及四种强攻击：CW$_\infty$ [10]，还有由 [14]提出的 AP（Auto-PGD：无步长）、DLR（Auto-DLR：平移和缩放不变）和 AA（Auto-Attack：无需参数）。PGD、AP、DLR 均进行 30 步并采用随机初始化，其中 PGD 的步长分别为 0.0023 和 0.0011，AP 和 DLR 的动量系数为 $\rho = 0.75$。CW$_\infty$ 在 100 次迭代中使用梯度截断 l_∞，并在 $\kappa = 0$ 上采用 CW目标函数 [10] 。对于防御，我们采用一种标准防御基线：ADV [41]，以及四种强防御基线：TRADES [76], MART [66],AWP [71], HELP [53]。我们使用 PGD [41] 生成对抗样本，扰动预算为 $8/255$，训练时设置 10 步，步长为 0.0072。特别地，由于 Tiny-ImageNet 的对抗训练计算负担较重，我们采用快速对抗训练 [69]，使用 FGSM，预算为 $4/255$，步长为其 1.25 倍。对于所有训练，我们使用 SGD [56]，动量为 0.9，学习率为 0.1，并在 120 个 epoch [55, 69]中通过 Cyclic [61] 进行调度。

6.2. 对比对抗鲁棒性

我们将上述五种防御基线与实验设置保持一致，以公平验证对抗鲁棒性。根据公式（8），我们首先获取因果反演，以直接处理因果特征。随后，如公式（9）所述，我们利用因果反演，通过从头开始将因果正则化器添加到防御基线的预定义损失中，对所有网2 络实施因果接种。表 2 展示了 CAFE 提升了这五种防御基



(a) 拉德马赫距离　　(b) 不平衡比率

图 4. 展示 VGG-16 在两个数据集上的预测结果中，拉德马赫距离和不平衡比率的箱线分布统计，并比较使用正则化器与不使用正则化器的情况。

线的性能，即使在更大的网络和更大的数据集上也优于它们，从而验证了注入因果特征对所有网络都有效。附录 F 展示了去除因果反演后的 CAFE 消融研究，以确定其有效性来源。

6.3. 关于丰富测试函数的消融研究

为验证正则化器在实践中确实有效，我们测量了拉德马赫距离，并展示其箱线分布，如图4 (a) 所示。通过泛化差距缩小，我们可以明显观察到正则化效率的存在。具体而言，经过正则化的检验函数的拉德马赫距离，其中位数和平均值均小于未经正则化的检验函数。接下来，为研究丰富测试函数如何助力因果推断，我们考察假设模型预测结果的不平衡比率，其计算方式为预测类别中最少类别的数量除以最多类别的数量。如果反事实空间大幅偏离可能的特征界限，那么假设模型能够到达的可达空间就仅限于某些区域。因此，假设模型可能会对目标对象给出有偏的预测结果。正如预期，在两个数据集上，使用正则化器所得的比率都远高于不使用正则化器的情况，如图4 (b) 所示。因此，我们可以总结，局部拉德马赫正则化器所获得的丰富测试函数，是提升因果推断广义容量的关键。

7. 结论

本文构建了 AMR-GMM，以发展对抗性工具变量回归，从因果视角揭示对抗样本中的因果特征，从而探究难以解释的对抗性起源。通过详尽分析，我们借助假设模型和检验函数深入研究对抗性预测中的因果关系，并以人类可识别的方式识别其语义信息。此外，我们引入因果反演，以处理网络可能的特征界限上的因果特征，并提出因果接种，将因果特征（*CAFE*）植入防御机制，以提升对抗鲁棒性。

参考文献

[1] Chunrong Ai 和 Xiaohong Chen。含未知函数的条件矩限制模型的有效估计。《Econometrica》，71(6):1795–1843，2003年。3[2] Joshua D Angrist、Guido W Imbens 和 Donald Ru-bin。使用工具变量识别因果效应。《美国统计学会会刊》，91(434):444–455，1996年。2, 3, 5[3] Joshua D Angrist 和 J¨orn-Steffen Pischke。《基本无害的计量经济学》。载于《基本无害的计量经济学》。普林斯顿大学出版社，2008年。2, 3[4] Giovanni Apruzzese、Michele Colajanni、Luca Ferretti 和 Mirco Marchetti。应对针对基于机器学习的安全系统的对抗性攻击。载于网络冲突国际会议，第900卷，第1–18页。IEEE，2019年。1[5] Martin Arjovsky、Soumith Chintala 和 L´eon Bottou。Wasserstein生成对抗网络。载于国际机器学习会议，第214–223页。PMLR，2017年。2[6] Peter L Bartlett 和 Shahar Mendelson。Rademacher复杂度和高斯复杂度：风险界与结构性结果。《机器学习研究杂志》，3 482，2002 463 5（11月）：–。[7] Andrew Bennett、Nathan Kallus 和 Tobias Sc hnabel。用于工具变量分析的深度广义矩估计法。《神经信息处理系统进展》，32，2019年。3, 5[8] Richard Blundell、Stephen Bond 和 Frank Windmeijer。动态面板数据模型的估计：改进标准GMM估计量的表现。Emerald Group Publishing Limited，2001年。3[9] David Card。利用大学邻近性的地理变异估计受教育回报，1993年。2[10] Nicholas Carlini 和 David Wagner。迈向神经网络鲁棒性的评估。载于IEEE安全与隐私研讨会，第39–57页。IEEE计算机学会，2017年。5, 8[11] Gary Chamberlain。含条件矩限制的估计的渐近有效性。《计量经济学杂志》，34(3):305–334，1987年。3[12] Xiaohong Chen 和 Timothy M Christensen。非参数工具变量回归中非线性泛函的最优上确界范数收敛率与一致推断。《数量经济学》，9(1):39–84，2018年。2[13] Xiaohong Chen 和 Demian Pouzo。可能含非光滑广义残差的非参数条件矩模型估计。《Econometrica》，80(1):277–321，2012年。1, 2, 3[14]Francesco Croce 和 Matthias Hein。通过多种无参数攻击的集成可靠地评估对抗鲁棒性。载于国际机器学习会议，第2206–2216页。PMLR，2020年。8[15] Serge Darolles、Yanqin Fan、Jean-Pierre Florens 和 Eric Renault。非参数工具变量回归。《Econometrica》，79(5):1541–1565，2011年。1, 2

[16] R Fernandes de Mello 和 M Antonelli Ponti。统计学习理论。《机器学习》，2018年。3[17] Huiqi Deng、Qihan Ren、Hao Zhang 和 Quanshi Zhang。发现并解释深度神经网络的表示瓶颈。载于《国际学习表征会议》，2022年。5[18] Jia Deng、Wei Dong、Richard Socher、Li-Jia Li、Kai Li 和 Li Fei-Fei。ImageNet：大型分层图像数据库。载于《计算机视觉与模式识别会议》，第248–255页。IEEE，2009年。5, 6[19] Nishanth Dikkala、Greg Lewis、Lester Mackey 和 Vasilis Syrgkanis。条件矩模型的极小极大估计。《神经信息处理系统进展》，33:12248–12262，2020年。2, 3, 4, 5[20]Simon Du 和 Jason Lee。论神经网络中二次激活函数的过参数化能力。载于《国际机器学习会议》，第80卷，第1329–1338页。PMLR，2018年7月10–15日。5[21]Logan Engstrom、Justin Gilmer、Gabriel Goh、Dan Hendrycks、Andrew Ilyas、Aleksander Madry、Reiichiro Nakano、Preetum Nakkiran、Shibani Santurkar、Brandon Tran、Dimitris Tsipras 和 Eric Wallace。关于“对抗样本不是错误，而是特征”的讨论。Distill，2019年。1, 3[22] Logan Engstrom、Andrew Ilyas、Shibani Santurkar、Dimitris Tsipras、Brandon Tran 和 Aleksander Madry。对抗鲁棒性作为学习表征的先验。arXiv预印本 arXiv:1906.00945，2019年。5[23] Rocio Garcia-Retamero 和 Ulrich Hoffrage。因果知识如何简化决策。《心智与机器》，16(3):365–380，2006年。2[24] Ian Goodfellow、Jean Pouget-Abadie、Mehdi Mirza、Bing Xu、David Warde-Farley、Sherjil Ozair、Aaron Courville 和 Yoshua Bengio。生成对抗网络。《神经信息处理系统进展》，27，2014年。2[25] Ian J Goodfellow、Jonathon Shlens 和 Christian Szegedy。解释并利用对抗样本。载于《国际学习表征会议》，2015年。1, 3, 5, 8[26] York Hagmayer 和 Cilia Witteman。决策中的因果知识与推理。载于《学习与动机心理学》，第67卷，第95–134页。Elsevier，2017年。2[27] Lars Peter Hansen。广义矩估计量的大样本性质。《Econometrica：计量经济学会会刊》，第1029–1054页，1982年。2[28] Jason Hartford、Greg Lewis、Kevin Leyton-Brown 和 Matt Taddy。深度工具变量：一种灵活的反事实预测方法。载于《国际机器学习会议》，第1414–1423页。PMLR，2017年。3, 6[29] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。用于图像识别的深度残差学习。载于《计算机视觉与模式识别会议》，第770–778页，2016年。8[30] James J Heckman、Sergio Urzua 和 Edward Vytlacil。理解具有本质异质性的模型中的工具变量。《经济学与统计学评论》，88(3):389–432，2006年。5

[31] Dan Hendrycks 和 Thomas G. Dietterich。评测神经网络对常见损坏和扰动的鲁棒性。载于《国际学习表征会议》，2019年。1[32] Paul W Holland。统计学与因果推断。《美国统计学会会刊》，81 960，1986 396 945 5 ( ): – 。[33] Andrew Ilyas、Shibani Santurkar、Dimitris Tsipras、Logan Engstrom、Brandon Tran 和 Aleksander Madry。对抗样本不是错误，而是特征。载于《神经信息处理系统进展》，第32卷，2019年。1, 3[34] Junho Kim、Byung-Kwan Lee 和 Yong Man Ro。通过信息瓶颈提取对抗样本中的鲁棒与非鲁棒特征。载于《神经信息处理系统进展》，2021年。1, 3, 5[35] Nancy S Kim 和 Stefanie T LoSavio。因果解释会影响人们对心理治疗必要性的判断。《判断与决策》，4(1):82，2009年。2[36] Vladimir Koltchinskii 和 Dmitry Panchenko。经验边际分布与组合分类器泛化误差的界。《统计年刊》，30(1):1–50，2002年。5[37] Alex Krizhevsky、Geoffrey Hinton 等。利用小型图像学习多层特征。2009年。5, 8[38] Ya Le 和 Xuan Yang。Tiny ImageNet视觉识别挑战赛。CS 231N，7:7，2015年。8[39] Byung-Kwan Lee、Junho Kim 和 Yong Man Ro。掩蔽对抗性损害：寻找对抗显著性，以实现鲁棒的稀疏网络。载于《计算机视觉与模式识别会议》，第15126–15136页，2022年。5[40] Greg Lewis 和 Vasilis Syrgkanis。对抗性广义矩估计法。arXiv预印本 arXiv:1803 07164，2018 2, 3, 5 。[41] Aleksander Madry、Aleksandar Makelov、Ludwig Schmidt、Dimitris Tsipras 和 Adrian Vladu。迈向能够抵御对抗性攻击的深度学习模型。载于《国际学习表征会议》，2018年。1, 5, 6, 8[42] Aravindh Mahendran 和 Andrea Vedaldi。通过反演理解深层图像表征。载于《计算机视觉与模式识别会议》，第5188–5196页，2015年。2, 6[43] Krikamol Muandet、Wittawat Jitkrittum 和 Jonas K¨ubler。通过最大矩约束进行核条件矩检验。载于《人工智能不确定性会议》，第41–50页。PMLR，2020年。4[44] Krikamol Muandet、Arash Mehrjou、Si Kai Lee 和 Anant Raj。双重工具变量回归。《神经信息处理系统进展》，33:2710–2721，2020年。3, 6[45] Yuval Netzer、Tao Wang、Adam Coates、Alessandro Bissacco、Bo Wu 和 Andrew Y Ng。利用无监督特征学习在自然图像中读取数字。2011年。8[46] Whitney K Newey 和 James L Powell。非参数模型的工具变量估计。《Econometrica》，71(5):1565–1578，2003年。1, 2

[47] Anh Nguyen、Jason Yosinski 和 Jeff Clune。《通过特征可视化理解神经网络：综述》。载于《可解释人工智能：深度学习的解释、说明与可视化》，第 55–76 页。施普林格，2019年。6[48] Olena Y Nizalova 和 Irina Murtazashvili。《外生处理与内生因素：交互项中遗漏变量偏差的消失》。Journal of Econometric Methods，5(1):71–77，2016年。3[49] Chris Olah、Alexander Mordvintsev 和 Ludwig Schubert。《特征可视化》。Distill，2(11):e7，2017年。2, 6[50] Judea Pearl。《因果关系》。剑桥大学出版社，2009年。3[51] Jonas Peters、Dominik Janzing 和 Bernhard Sch¨olkopf。《因果推断要素：基础与学习算法》。麻省理工学院出版社，2017年。3[52] Peter CB Phillips 和 Bruce E Hansen。《具有 I(1) 过程的工具变量回归中的统计推断》。《经济研究评论》，57(1):99–125，1990年。2[53] Rahul Rade 和 Seyed-Mohsen Moosavi-Dezfooli。《降低过大的间隔，以实现更佳的准确率与鲁棒性权衡》。载于《国际学习表征会议》，2022年。1, 6, 8[54] Olav Reiersøl。《利用工具变量集进行汇流分析》。博士论文，Almqvist & Wiksell，1945年。1, 2[55] Leslie Rice、Eric Wong 和 Zico Kolter。《对抗鲁棒深度学习中的过拟合》。载于《国际机器学习会议》，第119卷，第8093–8104页，2020年。8[56] Herbert Robbins 和 Sutton Monro。《一种随机逼近方法》。《数理统计年刊》，第400–407页，1951年。8[57] Y. E. Sagduyu、Y. Shi 和 T. Erpek。《从对抗性深度学习视角看物联网网络安全》。载于《国际传感、通信与联网会议》，第1–9页，2019年。1[58] Ali Shafahi、W Ronny Huang、Christoph Studer、Soheil Feizi 和 Tom Goldstein。《对抗样本是否不可避免？》。载于《国际学习表征会议》，2018年。1, 3[59] Karen Simonyan 和 Andrew Zisserman。《用于大规模图像识别的超深卷积网络》。载于《国际学习表征会议》，2015年。8[60] Rahul Singh、Maneesh Sahani 和 Arthur Gretton。《核工具变量回归》。载于《神经信息处理系统进展》，第32卷，2019年。3[61] Leslie N Smith。《用于训练神经网络的循环学习率》。载于《IEEE冬季计算机视觉应用会议》，第464–472页。IEEE，2017年。8[62] Christian Szegedy、Wojciech Zaremba、Ilya Sutskever、Joan Bruna、Dumitru Erhan、Ian Goodfellow 和 Rob Fergus。《神经网络的迷人特性》。载于《国际学习表征会议》，2014年。1, 3[63] Dimitris Tsipras、Shibani Santurkar、Logan Engstrom、Alexander Turner 和 Aleksander Madry。《鲁棒性可能与准确率相悖》。载于《国际学习表征会议》，2019年。1

[64] Li Wan、Matthew Zeiler、Sixin Zhang、Yann Le Cun 和 Rob Fergus。《使用 Dropconnect 对神经网络进行正则化》。载于《国际机器学习会议》，第1058–1066页。PMLR，2013年。5[65] Xianmin Wang、Jing Li、Xiaohui Kuang、Yu an Tan 和 Jin Li。《对抗环境下机器学习的安全性：综述》。Journal of Parallel and Distributed Computing，130:12–23，2019年。1[66] Yisen Wang、Difan Zou、Jinfeng Yi、James Bailey、Xingjun Ma 和 Quanquan Gu。《提升对抗鲁棒性需要重新审视误分类样本》。载于《国际学习表征会议》，2020年。1, 6, 8[67] Ziyu Wang、Yuhao Zhou、Tongzheng Ren 和 Jun Zhu。《用于工具变量回归的可扩展准贝叶斯推断》。载于《神经信息处理系统进展》，第34卷，2021年。5[68] Colin Wei 和 Tengyu Ma。《通过 Lipschitz 增强研究深度神经网络的数据依赖样本复杂度》。载于《神经信息处理系统进展》，第32卷。Curran Associates, Inc.，2019年。5[69] Eric Wong、Leslie Rice 和 J. Zico Kolter。《快速胜过免费：重新审视对抗训练》。载于《国际学习表征会议》，2020年。8[70] Jeffrey M Wooldridge。《横截面数据与面板数据计量经济学分析》。麻省理工学院出版社，2010年。2, 3[71] Dongxian Wu、Shu-Tao Xia 和 Yisen Wang。《对抗权重扰动有助于鲁棒泛化》。载于《神经信息处理系统进展》，第33卷，2969，2020 2958 1, 6, 8 : – .[72] Dong Yin、Raphael Gontijo Lopes、Jon Shlens、Ekin Dogus Cubuk 和 Justin Gilmer。《模型鲁棒性的傅里叶视角：计算机视觉》。载于《神经信息处理系统进展》，第32卷，2019年。1, 3[73] Dong Yin、Ramchandran Kannan 和 Peter Bartlett。《对抗鲁棒泛化的 Rademacher 复杂度》。载于《国际机器学习会议》，第7085–7094页。PMLR，2019年。5[74] Sergey Zagoruyko 和 Nikos Komodakis。《宽残差网络》。载于《英国机器视觉会议》，2016年。8[75] Ke Zhai 和 Huan Wang。《带有 Rademacher复杂度正则化的自适应 Dropout》。载于《国际学习表征会议》，2018年。5[76] Hongyang Zhang、Yaodong Yu、Jiantao Jiao、Eric Xing、Laurent El Ghaoui 和 Michael Jordan。《鲁棒性与准确率之间具有理论依据的权衡》。载于《国际机器学习会议》，第97卷，第7472–7482页，2019年6月9–15日。1, 6, 8[77]Rui Zhang、Masaaki Imaizumi、Bernhard Sch¨olkopf 和 Krikamol Muandet。《工具变量回归的最大矩约束》。arXiv 预印本 arXiv:2010.07684，2020年。3, 4