

针对可逆神经网络的可视化解释方法

牟新颖¹, 宋冰冰², 李钊效¹, 郑奕森¹, 周维¹, 董云云^{1,2}

(1. 云南大学国家示范性软件学院, 云南 昆明 650000; 2. 云南大学信息学院, 云南 昆明 650000)

摘要: 为了更好地理解深度神经网络(DNN, deep neural network)在应用过程中出现的决策依据未知以及容易受到对抗攻击等安全问题, 模型可解释性受到广泛关注。虽然越来越多的学者针对传统深度神经网络的可解释性进行了研究, 但对可逆神经网络的运行机制和可解释性的探索还存在不足, 且现有针对传统深度神经网络的解释方法不适用于可逆神经网络, 存在噪声大、梯度饱和等问题。因此, 提出一种针对可逆神经网络的可视化解释方法, 其基于类激活映射机制, 利用可逆神经网络的可逆特性探索特征图与输入图像之间的区域对应关系, 使得区域特征图的分类权重可映射到输入图像的对应区域, 得到输入图像每个区域对模型决策的重要程度, 从而生成模型决策依据。在通用数据集上, 将所提方法与其他解释方法进行实验比较, 所提方法取得了更集中的视觉效果, 在识别任务中, 相较于次优方法平均下降(AD, average drop)指标提升7.80%, 平均上升(AI, average increase)指标提升6.05%, 热值最大点的定位水平达到82.00%, 同时, 所提方法可以对传统深度神经网络进行解释且其良好的扩展性可以提高其他方法对可逆神经网络的解释性能。另外, 在对抗攻击解析实验中发现, 对抗攻击使得模型的决策依据发生改变, 体现在模型的关注区域发生错位, 这有助于探究对抗攻击的运行机制。

关键词: 模型可解释性; 可逆神经网络; 可视化; 类激活映射; 人工智能安全

中图分类号: TP393

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.20230090

Visual explanation method for reversible neural networks

MU Xinying¹, SONG Bingbing², LI Fanxiao¹, ZHENG Yisen¹, ZHOU Wei¹, DONG Yunyun^{1,2}

1. National Pilot School of Software, Yunnan University, Kunming 650000, China

2. School of Information Science and Engineering, Yunnan University, Kunming 650000, China

Abstract: The issue of model explainability has gained significant attention in understanding the vulnerabilities and anonymous decision-making processes inherent in deep neural networks (DNN). While there has been considerable

收稿日期: 2023-06-05; 修回日期: 2023-09-09

通信作者: 董云云, dongyy929@ynu.edu.cn

基金项目: 国家自然科学基金(62162067, 62101480); 云南省自然科学基金(202005AC160007, 202001BB050076); 云南省教育厅基金(2022j0008); 云南省重点研发计划; 云南省迟学斌专家工作站项目(202305AF150078)

Foundation Items: The National Natural Science Foundation of China (62162067, 62101480), The Natural Science Foundation of Yunnan Province (202005AC160007, 202001BB050076), Yunnan Provincial Department of Education Fund Project (2022j0008), Key R&D Plan of Yunnan Province, Yunnan Province Chi Xuebin Expert Workstation Project (202305AF150078)

引用格式: 牟新颖, 宋冰冰, 李钊效, 等. 针对可逆神经网络的可视化解释方法[J]. 网络与信息安全学报, 2023, 9(6): 154-165.

Citation Format: MU X Y, SONG B B, LI F X, et al. Visual explanation method for reversible neural networks[J]. Chinese Journal of Network and Information Security, 2023, 9(6): 154-165.

research on explainability for traditional DNN, there is a lack of exploration on the operation mechanism and explainability of reversible neural networks (RevNN). Additionally, the existing explanation methods for traditional DNN are not suitable for RevNN and suffer from issues such as excessive noise and gradient saturation. To address these limitations, a visual explanation method called visual explanation method for reversible neural network (VERN) was proposed for RevNN. VERN leverages the reversible property of RevNN and is based on the class-activation mapping mechanism. The correspondence between the feature map and the input image was explored by VERN, allowing for the mapping of classification weights of regional feature maps to the corresponding regions of the input image. The importance of each region for model decision-making was revealed through this process, which generates a basis for model decision-making. Experimental comparisons with other explanation methods on generalized datasets demonstrate that VERN achieves a more focused visual effect, surpassing suboptimal methods with up to 7.80% improvement in average drop (AD) metrics and up to 6.05% improvement in average increase (AI) metrics in recognition tasks. VERN also exhibits an 82.00% level of localization for the maximum point of the heat value. Furthermore, VERN can be applied to explain traditional DNN and exhibits good scalability, improving the performance of other methods in explaining RevNN. Furthermore, through adversarial attack analysis experiments, it is observed that adversarial attacks alter the decision basis of the model. This is reflected in the misalignment of the model's attention regions, thereby aiding in the exploration of the operation mechanism of adversarial attacks.

Keywords: model explainability, reversible neural network, visualization, class activation mapping, artificial intelligence security

0 引言

近年来, 人工智能发展迅速, 深度神经网络(DNN, deep neural network)在各个领域广泛应用的同时出现了网络决策依据未知、容易受到对抗攻击等安全问题^[1]。模型可解释性方法旨在探索模型做出决策的依据, 提高模型的透明性, 从而能更好地理解安全问题的产生。

越来越多的学者对传统深度神经网络的决策依据进行了研究^[2-5], 各种解释方法^[6-11]相继被提出。同时, 可逆神经网络作为深度神经网络的一种类型, 由于其在输入和输出之间的可逆特性, 在各个领域都表现出优异的性能, 如图像分类^[12]、图像隐写^[13]及超分辨率^[14]等, 针对可逆神经网络的解释方法的研究迫在眉睫。

基于传统神经网络的常用解释方法被分为3类: 基于梯度可视化的解释方法、基于扰动的解释方法和基于类激活映射(CAM, class activation mapping)的解释方法。虽然这些解释方法在传统深度神经网络中取得了较好的性能, 但是它们仍然存在各种问题。基于梯度可视化的解释方法生成的解释热图有很多噪声且可视化质量低; 基于扰动的解释方法缺乏指导信息, 需要大量的迭代

时间; 基于类激活映射的解释方法不完全适用于解释可逆神经网络, 如CAM^[15], 由于有特定网络结构的要求, 需要改变网络结构并重新训练网络。Grad-CAM^[9]和 Grad-CAM++^[10]虽然解决了需要改变模型结构的缺陷, 但它们在解释过程中需要使用梯度, 这引入了梯度饱和和置信度错误等问题; 为了摆脱对梯度的依赖, Score-CAM^[11]被提出, 但可逆神经网络的可逆空间下采样操作破坏了可逆神经网络中特征图和输入图像之间的比例关系, 使其不适用于解释可逆神经网络。

上述问题导致现有的解释方法并不适用于可逆神经网络, 为了更好地理解可逆神经网络的内部运行机制, 针对可逆神经网络的可视化解释方法(VERN, visual explanation method for reversible neural network)被提出, 同时, 该方法可用于解释传统深度神经网络。通过实验证明: VERN 有效地生成了解释热图, 提高了模型透明度, 用户对网络的信任度随之提高, 有助于用户更好地理解模型安全问题产生的原因; VERN 具有良好的可扩展性, 可以帮助其他解释方法提升对可逆神经网络的解释性能。基于类激活映射的解释方法效果示例如图1所示。

本文的主要贡献如下。

1) 提出针对可逆神经网络的可视化解释方法,且该方法可拓展到传统深度神经网络。VERN 摆脱了对梯度的依赖,不需要大量的迭代,同时不需要上采样等操作,在减少计算消耗的同时提升了解释性能。

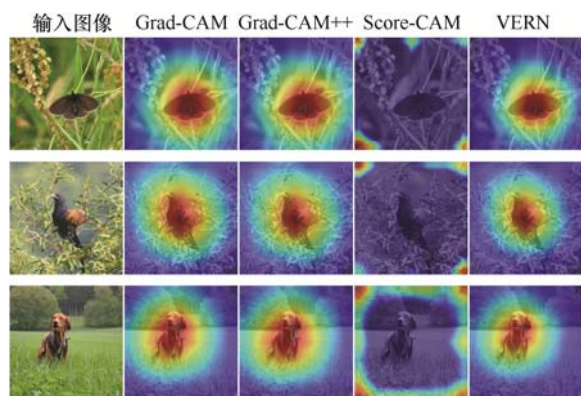


图1 基于类激活映射的解释方法效果示例
Figure 1 Example of effectiveness of explanation methods based on class activation mapping

2) 在多个数据集上进行了大量实验,对 VERN 进行定性和定量评估。实验结果表明,与其他方法相比,VERN 具有更集中的可视化效果,在识别任务中具有很高的忠实度,在定位任务上有很好的表现。另外,VERN 具有良好的可扩展性,能够提高其他方法解释可逆神经网络的效果,这证明了 VERN 的有效性。

3) 对抗攻击解析实验结果表明对抗攻击使得方法的决策依据发生改变,体现在方法的关注区域发生错位,这有助于挖掘神经网络的脆弱性机制。

1 相关工作

1.1 基于梯度可视化的解释方法

基于梯度可视化的解释方法^[16-19]通过反向梯度传播计算从目标类到输入图像的梯度从而得到解释图,梯度表明小幅度修改输入图像每个像素对分类结果的影响,从而推断每个像素对于决策的重要性。显著图^[16]通过计算梯度表明每个像素的重要性,从而获得解释图。其他解释方法^[17-19]在显著图的基础上进行改进,以解决使用梯度导致的神经元梯度饱和的问题,从而获得视觉上锐化的解释图。但是这类解释方法的可视化热图有大量的噪声且质量较低。

1.2 基于扰动的解释方法

基于扰动的解释方法^[6-8,20]的主要思想是对输入图像进行局部扰动,扰动导致目标类分数发生变化的程度表明该区域对方法决策的重要程度,从而产生解释热图。Occlusion^[6]以矩形形式逐步扰动原始输入图像,方法预测分数的变化程度表示每个区域的重要性。其他方法^[7-8,20]基于文献[6]方法改进了扰动区域的生成方式。然而,这类方法需要大量迭代,会导致时间消耗。

1.3 基于类激活映射的解释方法

基于类激活映射的解释方法通过特征图的线性加权组合得到解释图。CAM 需要在最后一个卷积层之后进行全局平均池化,但并不是所有的网络都能满足这一结构需求,因此,CAM 有时需要改变网络结构并重新训练网络。为了优化该方法,Grad-CAM^[9]使用梯度信息计算特征图的权重。但 Grad-CAM 不能精确定位且在多物体的图像上效果不佳,因此 Grad-CAM++^[10]被提出。虽然梯度的使用解决了需要改变模型结构的缺点,但 Grad-CAM 和 Grad-CAM++ 使用梯度计算权重会引起梯度饱和和错误的置信度等问题。为了摆脱对梯度的依赖,Score-CAM^[11]通过计算置信度得到特征图的重要性。然而,特征图的大小与输入图像不一致,需要上采样操作,这会引入不准确的信息;可逆神经网络的可逆空间下采样操作,破坏了输入图像和特征图之间的比例关系,使得 Score-CAM 不适用于解释可逆神经网络。

综上所述,现有的解释方法存在诸多缺点。相较于基于梯度可视化的解释方法,VERN 能够生成更小噪声的解释热图;相较于基于扰动的解释方法,VERN 不需要大量的迭代;相较于基于类激活映射的解释方法,VERN 不需要依赖于梯度知识等。

2 先验知识

本节介绍可逆神经网络的相关知识。RevNet 是 Gomez 等^[21]提出的一个用于分类的可逆神经网络。RevNet 的计算过程如图 2 所示,RevNet 从通道维度将输入分为两部分,分别表示为 x_1 和 x_2 ,两个分支由卷积模块 F 和 G 连接,输出表示为 y_1 和 y_2 ;每层的激活图不需要被存储,可以通过下一层的激活图来重建,因此极大程度地减少了

网络深度增加而导致的内存消耗。

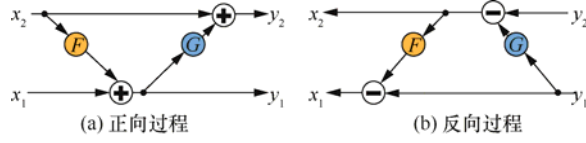


图2 RevNet 的计算过程
Figure 2 The calculation process of RevNet

RevNet 包含最大池化等不可逆操作，所以其是局部可逆的。i-RevNet^[12]替换其中的不可逆组件，构建一个全局可逆的模型。i-RevNet 的计算过程如图3所示，其正向计算过程的计算如下。

$$y_1 = S(x_2) \quad (1)$$

$$y_2 = x_1 + F(x_2) \quad (2)$$

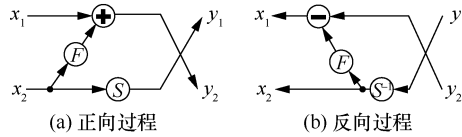


图3 i-RevNet 的计算过程
Figure 3 The calculation process of i-RevNet

i-RevNet 的反向计算过程为

$$x_2 = S^{-1}(y_1) \quad (3)$$

$$x_1 = y_2 - F(x_2) \quad (4)$$

其中， F 模块的操作与 RevNet 一致，由卷积和非线性操作组成。可逆模块 S 的过程，即可逆空间下采样如图4所示。

3 本文方法

VERN 框架如图5所示，其总体可以分为4个阶段。

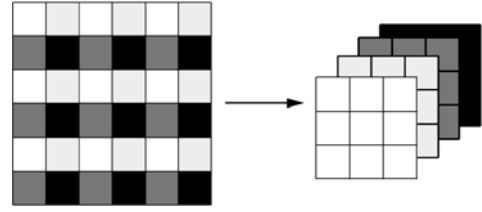


图4 i-RevNet 的可逆空间下采样
Figure 4 The reversible spatial down-sampling of i-RevNet

1) 基于掩码滑动窗口的特征定位：设计掩码滑动窗口策略捕捉局部特征并利用可逆神经网络的可逆特性获得局部特征图与输入图像之间的区域对应关系，生成相应的掩码以保留每个区域特征图所对应到输入图像中的区域。

2) 局部特征权重计算：将局部特征图输入全连接层，获得目标类的分类分数，该分数表明局部特征图的重要性。

3) 基于局部特征权重的全局权重映射：将上述两个阶段得到的掩码和权重进行线性加权组合，以获得初步的热图。

4) 基于均值滤波的可视化优化：通过均值滤波操作平滑初步热图得到最终的解释热图。

符号定义：给定可逆神经网络 $Y = f(X)$ ， X 是输入图像， Y 是网络输出的分类结果， f 为模型参数， l 为选取的卷积层，该层的特征图被描述为 A ，目标类为 c 。

3.1 基于掩码滑动窗口的特征定位

为了获取局部特征，本阶段设计了一个掩码滑动窗口捕捉特征。给定 l 层的特征图 A ，其形状为 (C, H, W) ，滑动窗口的大小可以设置在 $C \times 1 \times 1$ 到 $C \times (H-1) \times (W-1)$ 之间。设计滑动窗口的策略为：滑动窗口内的特征信息保留，其他

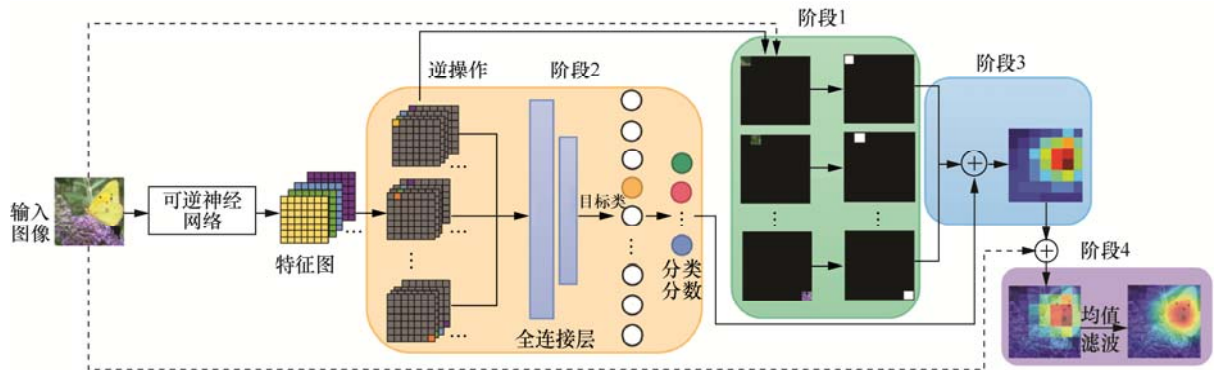


图5 VERN 框架
Figure 5 The framework of VERN

位置所对应的特征值用0^[22]填充,按特征图位置滑动。将第 k 个滑动窗口输出的特征图表示为 A_{window}^k ,利用可逆神经网络的逆操作将其传到输入图像上,得到的图像表示为 X_{inverse}^k 。将 X_{inverse}^k 与 X 比较发现, X_{inverse}^k 保留了 X 的部分区域对应的图像,定义一个二维的掩码 mask^k 记录此区域,此区域内的值为1,其余为0。基于掩码滑动窗口的特征定位阶段的计算式表达为

$$\text{mask}^k = \text{region}(X_{\text{inverse}}^k) \quad (5)$$

$$X_{\text{inverse}}^k = \text{inverse}(A_{\text{window}}^k) \quad (6)$$

$$A_{\text{window}}^k = \text{window}(A) \quad (7)$$

其中, $\text{region}(\cdot)$ 表示 X_{inverse}^k 中保留 X 信息的区域定义值为1的操作, $\text{inverse}(\cdot)$ 表示可逆神经网络的逆操作, $\text{window}(\cdot)$ 表示使用滑动窗口策略捕捉局部特征图的操作。

3.2 局部特征权重计算

本阶段将 A_{window}^k 作为新的特征图输入全连接层,输出目标类 c 的分类分数。使用经过softmax后的值来计算目标类 c 的分数^[11],表示为 score_c^k ,它表示局部特征图的分类分数,即此区域为模型做出决策的重要性程度。局部特征权重计算阶段的公式表示为

$$\text{score}_c^k = \text{softmax}(s^k) \quad (8)$$

$$s^k = \text{FC}(A_{\text{window}}^k) \quad (9)$$

其中, $\text{FC}(\cdot)$ 是将局部特征图输入全连接层中计算目标类 c 的分类分数的操作。

3.3 基于局部特征权重的全局权重映射

经过基于掩码滑动窗口的特征定位和局部特征权重计算,得到了与输入图像的局部信息对应的掩码以及局部特征图的权重。结合基于掩码滑动窗口的特征定位阶段的结论可知,局部特征图的权重可以对应到局部输入图像上。本阶段通过线性组合操作将局部输入图像对应的掩码与其权重相乘后求和,得到初步解释热图。在线性组合之前进行归一化操作,以获得更明确的重要性,归一化规则为

$$\text{normalization}(\text{score}_c^k) = \frac{\text{score}_c^k}{\sum_{k=1}^n \text{score}_c^k} \quad (10)$$

其中, $\text{normalization}(\cdot)$ 代表归一化操作。基于局部特征权重的全局权重映射阶段的计算公式如下。

$$H^c = \sum_{k=1}^n \text{mask}^k \cdot \text{normalization}(\text{score}_c^k) \quad (11)$$

3.4 基于均值滤波的可视化优化

初步解释热图存在局部热值边界明显的缺点,因此,采用均值滤波操作对可视化结果进行优化。

均值滤波是一种通过对目标像素和周围像素进行平均的滤波方法,它计算周围像素的平均值填入目标像素。对于热图,每个像素的相邻像素反映了这个像素的重要性。基于均值滤波的可视化优化阶段计算表示为

$$L_{\text{VERN}}^c = \text{blur}(H^c) \quad (12)$$

其中, $\text{blur}(\cdot)$ 表示均值滤波操作, L_{VERN}^c 表示最终的解释热图。

总体来说,设计掩码滑动窗口策略捕捉局部特征,并利用可逆神经网络的可逆特性获得特征图的局部区域与输入图像之间的关系,然后计算出它们在决策过程中的重要性。之后,将输入图像的每个局部区域对应的掩码和它的权重相乘后求和,得到初步解释图。为了优化初步解释图,使用均值滤波来平滑它,计算表示为

$$L_{\text{VERN}}^c = \text{blur}\left(\sum_{k=1}^n \text{mask}^k \cdot \text{normalization}(\text{score}_c^k)\right) \quad (13)$$

4 实验

为了证明VERN的解释效果,本文选取了在图像解释性领域应用中较为广泛的两个数据集,即CIFAR-10^[23]和ImageNet^[24](ILSVRC2012)进行实验。针对可逆神经网络的实验,在识别任务中,从验证集中随机选择3 000幅正确分类的图像,较多的测试图像以及各个模型、方法中采用相同的测试图像,可以减少随机性导致的误差;针对消融实验、解释效率评估、定位能力评估和泛化性证明,选择500幅正确分类图像;针对拓展到传统深度神经网络上的实验,从验证集中随机选择了2 000幅正确分类的图像。

针对可逆神经网络，待解释模型为经过预训练的 i-RevNet；针对传统深度神经网络，待解释模型为经过预训练的 VGG16、AlexNet 和 ResNet18，模型全连接层之前的最后一个卷积层的特征图作为被选取的特征图^[11]。

在掩码滑动窗口大小的选择上，基于不同大小的掩码滑动窗口，评估其平均下降（AD, average drop）指标和平均上升（AI, average increase）指标的比值。AD 即解释热图中热值前 50% 的部分对应到原图区域的分类分数相对输入图像的分数的减少的程度，计算方法如下。

$$AD = \sum_{i=1}^N \frac{\max(0, Y_c^i - M_c^i)}{Y_c^i} \cdot 100\% \quad (14)$$

其中， Y_c^i 代表第 i 幅输入图像的分类分数， M_c^i 表示热值前 50% 区域的分类分数， N 代表图像数量。AI 即当保留重要区域图像的分类分数高于原输入图像分类分数时，该情况出现的频率，计算方法如下。

$$AI = \frac{\sum_{i=1}^N \text{sign}(Y_c^i < M_c^i)}{N} \cdot 100\% \quad (15)$$

图 6 展示了 VGG16 掩码滑动窗口大小变化，在窗口大小为 $C \times 4 \times 4$ 时，取得了最佳效果，这或与输入图像大小及特征图大小有关，合适大小的掩码滑动窗口可对应到原图中较为完整的局部特征。VGG16 与 AlexNet 的特征图尺寸较大，因此设置其窗口大小均为 $C \times 4 \times 4$ ，均值滤波核的大小为 17；对于 i-RevNet 和 ResNet18，其特征图较小，因此设置滑动窗口大小为 $C \times 1 \times 1$ ，均值滤波核的大小为 35。对于不同数据集任务、不同待解释模型，可以根据需要选择不同的掩码滑动

窗口的大小以及滤波核的大小。

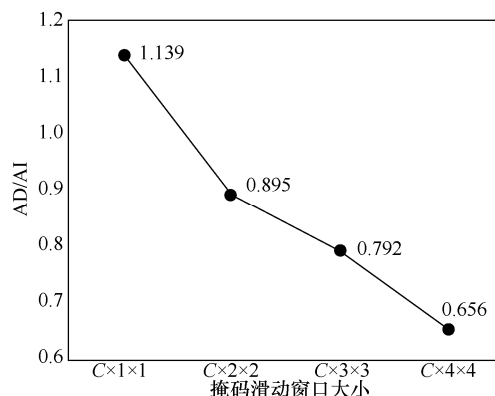


图 6 VGG16 掩码滑动窗口大小变化
Figure 6 Mask sliding window size change of VGG16

本文所采用的对比解释方法有 Gradient^[16]、Occlusion^[6]、Mask^[8]、RISE^[20]、Grad-CAM^[9]、Grad-CAM++^[10]和 Score-CAM^[11]，其中设置 Mask 优化迭代次数和 RISE 随机采样次数均为 1 000。

4.1 可视化效果及解释效率

解释方法最重要的指标之一是解释热图的可视化效果，这是用户对模型产生信任、理解安全问题产生原因最直接的方式。另外，生成一幅解释热图的效率是决定解释方法在现实应用中的重要要素。

4.1.1 集中程度

解释区域集中在目标对象上的程度越大，表明解释中的噪声越小、解释热图质量越高，与用户的判断方式越契合。可视化效果对比如图 7 所示，Gradient 有大量的噪声，Occlusion 的视觉效果并不能提供给用户合理的决策依据，Mask 需要多次迭代，RISE 没有引导性，达到最佳状态的随机采样次数不可知。通过黑白热力图可以明显看

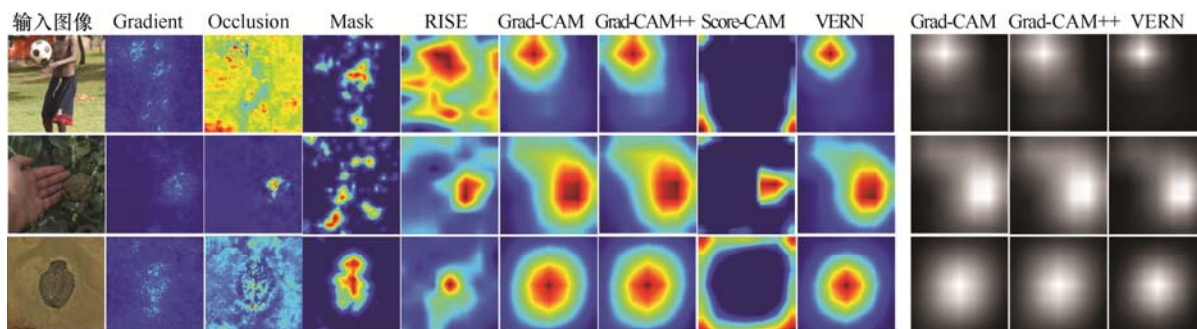


图 7 可视化效果对比
Figure 7 Comparison of visualization effects

出, VERN 解释热图中的高亮区域更加集中在物体本身而不是背景, 证明了其解释热图的高亮区域比 Grad-CAM 和 Grad-CAM++ 更加集中, 这与人们做出判断的依据更加一致。另外, 利用 VERN 在针对可逆神经网络的实验过程中得到的区域对应关系, 将其拓展到解释传统深度神经网络上, VERN 在传统深度神经网络下的可视化解释效果如图 8 所示。

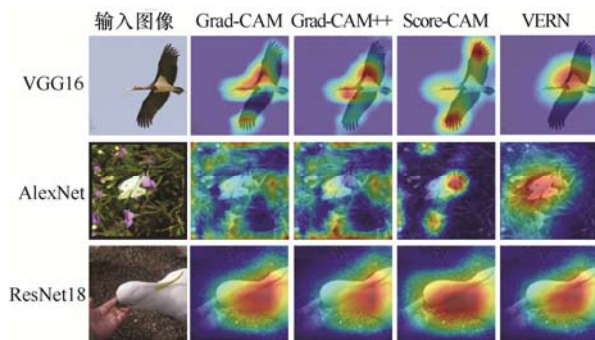


图 8 VERN 在传统深度神经网络下的可视化解释效果
Figure 8 The visualization explanation effect of VERN under traditional DNN

4.1.2 类敏感性评估

一个可信的解释方法应该是类敏感的^[28], 即不同的类别对应不同的贡献特征, 因此给定不同的目标类, 解释高亮区域应该对应到每个类的目标物体上。VERN 的类敏感性评估如图 9 所示, 给定目标类“斗牛獒”和“虎猫”, 解释区域分别转移到目标类所对应的物体上, 这证明了 VERN 是类敏感的。

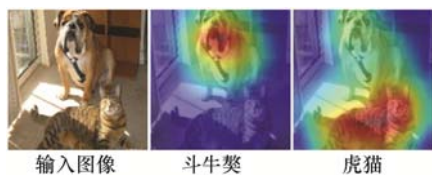


图 9 VERN 的类敏感性评估
Figure 9 Class sensitivity assessment of VERN

4.1.3 单类多对象识别

在定位多个相同的类对象时表现出更好的性能是评价解释方法的一个方面。VERN 的单类多对象识别如图 10 所示, VERN 能够定位多个同类对象。与 Grad-CAM 和 Grad-CAM++ 相比, VERN 对目标物体本身的识别更为集中而不是两个物体之间的连接处。VERN

能够定位多个同类对象是因为 VERN 通过设计的掩码滑动窗口策略包含了所有的激活图特征, 因此, 给定一个目标类, 该类的目标对象对应到输入图像都可以被识别, 反映在热图的高亮区域。

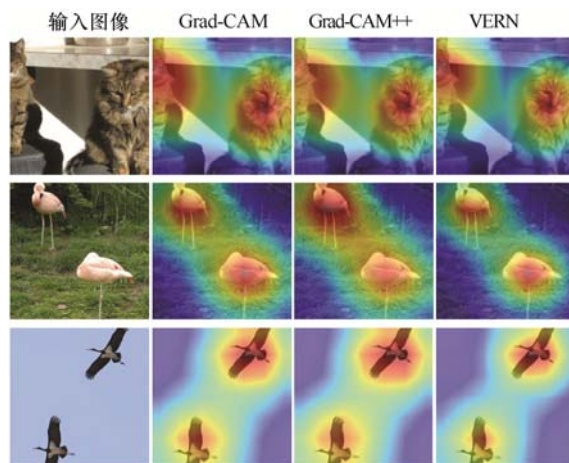


图 10 VERN 的单类多对象识别
Figure 10 Single-class multi-object recognition of VERN

4.1.4 解释效率

一个解释方法要在各个领域得到广泛应用, 解释效率是重要的因素。生成一幅解释热图的时间要在合理的范围内, 快速响应可以提供更好的用户体验。各个基于类激活映射的解释方法生成一幅解释图所用时间如表 1 所示, 此实验在 NVIDIA RTX 2080TI 上进行, 结果表明 VERN 在解释各个模型上生成一幅解释图所用时间都介于十毫秒数量级到百毫秒数量级之间, 在现实应用中, 可以满足绝大多数领域的效率需求。

表 1 各个基于类激活映射的解释方法生成一幅解释图所用时间

Table 1 The time taken for generating an explanation image based on various CAM explanation methods

解释方法	生成一幅解释图所用时间/ms			
	i-RevNet	VGG16	AlexNet	ResNet18
Grad-CAM	58.4	19.7	7.9	18.3
Grad-CAM++	61.9	23.8	9.3	18.8
Score-CAM	—	1758.5	301.7	1543.0
VERN	37.3	223.7	154.0	29.7

4.2 忠实性评估及可扩展性评估

4.2.1 平均下降、平均上升评估

本节对 VERN 解释图的高亮区域的忠实性进

行评估，即保留热图中热值前 50% 的位置所对应到原图的区域，其余位置为 0，计算 AD 和 AI 值。可逆神经网络在 ImageNet 和 CIFAR-10 数据集上的指标评估分别如表 2 和表 3 所示。实验结果表明，相较于次优方法，VERN 的 AD 指标提升了 3.84%，AI 指标提升了 1.90%。

表 2 ImageNet 数据集上的 AD 和 AI 指标评估
Table 2 Indicator evaluation AD and AI on ImageNet

解释方法	AD 指标	AI 指标
Gradient	72.59%	5.40%
Occlusion	85.48%	4.83%
Mask	67.11%	6.0%3
RISE	31.74%	24.00%
Grad-CAM	26.79%	23.77%
Grad-CAM++	27.52%	21.93%
Score-CAM	89.61%	2.70%
VERN	26.57%	24.23%

表 3 CIFAR-10 数据集上的 AD 和 AI 指标评估
Table 3 Indicator evaluation AD and AI on CIFAR-10

解释方法	AD 指标	AI 指标
Gradient	74.56%	3.37%
Occlusion	50.80%	10.13%
Mask	57.05%	8.17%
RISE	46.15%	10.87%
Grad-CAM	30.85%	11.40%
Grad-CAM++	34.27%	10.10%
Score-CAM	44.14%	0.80%
VERN	23.4%	14.97%

传统深度神经网络在 ImageNet 数据集上的 AD 和 AI 指标评估如表 4 所示，相较于次优方法，VERN 的 AD 指标提升了 5.14%，AI 指标提升了 2.42%，这表明 VERN 适用于解释传统深度神经网络。

4.2.2 删除、插入评估

为了进行更全面的比较，本节采用删除（Deletion）和插入（Insertion）指标^[20,29]对 VERN 进行评估。对于 Deletion 指标，当逐渐从热图中删除高亮区域时，目标类分数会下降。当最重要的区域逐渐被移除时，分类分数会有明显下降，下降曲线与轴线所形成区域很小。对于 Insertion 指标，越是重要的区域得以呈现，分类分数就越有明显增加，与轴线形成的曲线面积越大。针对可逆神经网络的 Deletion 和 Insertion 指标评估如表 5 所示，应用于传统深度神经网络的 Deletion 和 Insertion 指标评估如表 6 所示。

4.2.3 可扩展性评估

从表 1、表 2 可知，其他解释方法的忠实性逊色于 VERN。因此，基于 VERN 限制其他方法的解释高亮区域来提高其对可逆神经网络的解释忠实性。

在 VERN 中，选择解释热图中前 70% 的高亮区域并设定其值为 1，其他值为 0，将该掩码区域与其他方法的热图相乘，也就是说，排除其他方法解释中的噪声，从而提高其忠实性。之后，重新对它们的 AD 和 AI 指标进行评估，用 Improved AD 和 Improved AI 指标表示基于 VERN 在 ImageNet 和 CIFAR-10 上提高其他方法的忠实度

表 4 传统深度神经网络在 ImageNet 上的 AI 和 AD 指标评估
Table 4 Indicator evaluation average drop and average increase for traditional DNN on ImageNet

解释方法	VGG16		AlexNet		ResNet18	
	AD 指标	AI 指标	AD 指标	AI 指标	AD 指标	AI 指标
Gradient	76.69%	4.65%	79.62%	4.10%	75.98%	4.10%
Occlusion	94.74%	0.60%	89.16%	2.60%	94.81%	0.90%
Mask	52.75%	8.95%	67.35%	5.30%	57.75%	8.30%
RISE	34.60%	17.60%	52.94%	14.60%	34.55%	20.10%
Grad-CAM	34.04%	19.70%	60.69%	10.35%	26.30%	21.40%
Grad-CAM++	33.44%	16.70%	66.88%	6.90%	27.57%	19.7%
Score-CAM	25.59%	19.70%	54.57%	10.40%	28.44%	18.85%
VERN	18.16%	25.75%	45.14%	15.50%	26.11%	21.70%

表 5 可逆神经网络的 Deletion 和 Insertion 指标评估
Table 5 Deletion and Insertion indicator evaluation of RevNN

数据集	Deletion 指标			Insertion 指标		
	Grad-CAM	Grad-CAM++	VERN	Grad-CAM	Grad-CAM++	VERN
ImageNet	0.159	0.163	0.159	0.557	0.554	0.560
CIFAR-10	0.346	0.365	0.333	0.610	0.594	0.644

表 6 传统深度神经网络的 Deletion 和 Insertion 指标评估
Table 6 Deletion and Insertion indicator evaluation of traditional DNN

传统深度神经网络待解释模型	Deletion 指标				Insertion 指标			
	Grad-CAM	Grad-CAM++	Score-CAM	VERN	Grad-CAM	Grad-CAM++	Score-CAM	VERN
VGG16	0.120	0.121	0.118	0.157	0.540	0.540	0.581	0.622
AlexNet	0.101	0.118	0.101	0.114	0.345	0.311	0.360	0.408
ResNet18	0.134	0.137	0.139	0.134	0.549	0.540	0.539	0.549

表 7 基于 VERN 在 ImageNet 上提高其他方法的忠实度的评价指标
Table 7 Evaluation indicator of improving the faithfulness of other methods based on VERN on ImageNet

解释方法	AD 指标	Improved AD 指标	AI 指标	Improved AI 指标
Gradient	72.59%	53.15%	5.4%	13.2%
Occlusion	85.48%	60.33%	4.83%	25.83%
Mask	67.11%	43.87%	6.03%	10.13%
RISE	31.74%	24.39%	24%	28.27%
Grad-CAM	26.79%	26.11%	23.77%	24.43%
Grad-CAM++	27.52%	26.56%	21.93%	23.23%
Score-CAM	89.61%	88.72%	2.7%	3.97%

表 8 基于 VERN 在 CIFAR-10 上提高其他方法的忠实度的评价指标
Table 8 Evaluation indicator of improving the faithfulness of other methods based on VERN on CIFAR-10

解释方法	AD 指标	Improved AD 指标	AI 指标	Improved AI 指标
Gradient	74.56%	41.91%	3.37%	7.50%
Occlusion	50.80%	35.6%	10.13%	16.63%
Mask	57.05%	54.79%	8.17%	9.60%
RISE	46.15%	44.91%	10.87%	11.00%
Grad-CAM	30.85%	30.87%	11.40%	11.33%
Grad-CAM++	34.27%	33.63%	10.10%	10.50%
Score-CAM	44.14%	44.67%	0.80%	0.80%

的评价指标，分别如表 7、表 8 所示。VERN 在 ImageNet 数据集上对各个解释方法的 AD 指标平均提升了 11.10%，AI 指标提升了 5.77%，这证明了所提解释方法是忠实有效的。

4.3 定位能力评估

本节通过定点游戏衡量 VERN 在定位任务中的有效性，其思想是提取热图中的最大点，判断

它是否落入物体的边界框内。从 ImageNet 验证数据集中选择了 500 幅图像，对这 500 幅图像的要求是边界框的面积不能超过整张图的 50%且只有一个目标物体存在。定位任务评估如表 9 所示，82.00%的最大落入比例表明 VERN 生成的解释热图更专注于目标对象，这表明 VERN 与人们的判断方式更加一致。

表9 定位任务评估
Table 9 Evaluation of location tasks

解释方法	落入比例
Grad-CAM	79.2 %
Grad-CAM++	77.8 %
VERN	82.00%

4.4 泛化性评估

本节通过实验证明了 VERN 的泛化性，证明方法如下。① 对于随机选取的图像，通过不同解释方法生成基于 i-RevNet、VGG16、AlexNet、ResNet18 的解释热图，计算 4 幅热图中热值前 50%重要区域的 IoU (intersection over union) 值，即相交面积与并集面积之间的比例，泛化性证明——IoU 如表 10 所示，VERN 的 IoU 值最大，为 0.438，表明其在 4 个模型对同一幅图像生成的解释热图的重要解释区域一致性较高，证明了其泛化性。② 基于不同解释方法，针对 i-RevNet 生成解释热图，保留热图中热值前 50%重要区域所对应的原图区域，原图中其余像素值为 0，将更改后的原图输入 3 个传统神经网络中，计算其 AD 和 AI 值，泛化性证明——AD 和 AI 如表 11 所示，VERN 更小的 AD 值和更大的 AI 值表明，对于同一幅图，在可逆神经网络上生成的解释热图的重要区域对于传统神经网络来说同样重要，证明了 VERN 的泛化性较好。

4.5 对抗攻击解析

本节分别对含有单目标物体和多目标物体的输入图像进行对抗攻击（本节采用的对抗攻击方法为 PGD^[25]、RFGSM^[26]和 MIFGSM^[27]），并以 VERN 生成的解释图为基础分析对抗攻击的运行机制。

表10 泛化性证明——IoU
Table 10 Generalization proof——IoU

解释方法	IoU
Grad-CAM	0.277
Grad-CAM++	0.293
VERN	0.438

表11 泛化性证明——AD 和 AI
Table 11 Generalization proof——AD and AI

模型	解释方法	AD 指标	AI 指标
VGG16	Grad-CAM	27.22%	12.20%
	Grad-CAM++	27.23%	11.40%
	VERN	26.39%	12.20%
AlexNet	Grad-CAM	52.69%	10.40%
	Grad-CAM++	52.50%	10.20%
	VERN	52.00%	11.20%
ResNet18	Grad-CAM	25.98%	15.40%
	Grad-CAM++	26.04%	16.00%
	VERN	25.00%	17.40%

单目标攻击解析：单目标图像对抗攻击机制的解析如图 11 所示，攻击后热图中的高亮部分明显移位，这是因为对抗样本类别和原分类类别的物体之间存在相似性。因此，高亮区域的转移是一种相似特征的区域转移。

多目标攻击解析：单目标图像受到攻击后的分类类别是不确定的。与此相反，在多目标攻击后，解释的高亮区域会从一个目标转移到另一个目标，分类标签会发生改变，多目标图像对抗攻击机制的解析如图 12 所示。

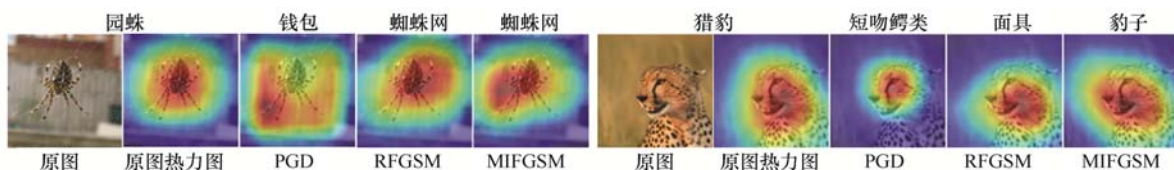


图 11 单目标图像对抗攻击机制的解析
Figure 11 Analysis of single-target image adversarial attack mechanism

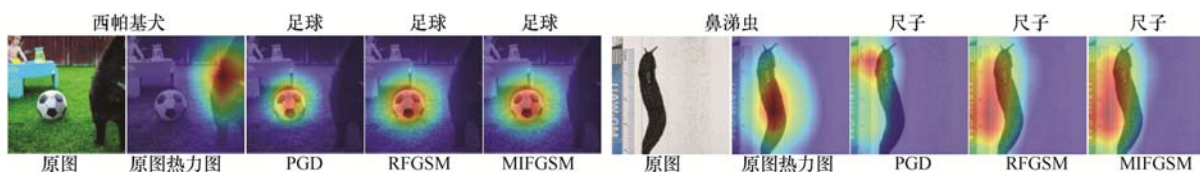


图 12 多目标图像对抗攻击机制的解析
Figure 12 Analysis of multi-target image adversarial attack mechanism

5 结束语

近年来,模型解释性方法大量涌现,提高了深度神经网络的透明度。然而,现有的解释方法并不能很好地适用于解释可逆神经网络。因此,本文提出了一种新的针对可逆神经网络的可视化解释方法——VERN,克服了传统神经网络下解释性方法不能很好地适用于可逆网络的缺陷;进行了大量的实验,证明VERN在定性和定量评估上均优于其他解释方法,并且探索了VERN的可扩展性,提高了其他解释方法对可逆神经网络的解释效果;由于深度神经网络容易受到对抗攻击的干扰,使用VERN比较得出单物体和多物体图像被攻击后解释区域的变化规律。下一步,本文将基于此探索对抗攻击的运行机制,提出相应的防御方法,从而提高模型的鲁棒性。

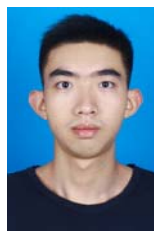
参考文献:

- [1] 谭清尹, 曾颖明, 韩叶, 等. 神经网络后门攻击研究[J]. 网络与信息安全学报, 2021, 7(3): 46-58.
TAN Q Y, ZENG Y M, HAN Y, et al. Survey on backdoor attacks targeted on neural network[J]. Chinese Journal of Network and Information Security, 2021, 7(3): 46-58.
- [2] 杨朋波, 桑基韬, 张彪, 等. 面向图像分类的深度模型可解释性研究综述[J]. 软件学报, 2023, 34(1): 230-254.
YANG P B, SANG J T, ZHANG B, et al. Survey on interpretability of deep models for image classification[J]. Journal of Software, 2023, 34(1): 230-254.
- [3] FANG Z, KUANG K, LIN Y, et al. Concept-based explanation for fine-grained images and its application in infectious keratitis classification[C]//Proceedings of the 28th ACM International Conference on Multimedia. 2020: 700-708.
- [4] 化盈盈, 张岱堀, 葛仕明. 深度学习模型可解释性的研究进展[J]. 信息安全学报, 2020, 5(3): 1-12.
HUA Y Y, ZHANG D X, GE S M, et al. Research progress in the interpretability of deep learning models[J]. Journal of Cyber Security, 2020, 5(3): 1-12.
- [5] 纪守领, 李进锋, 杜天宇, 等. 机器学习模型可解释性方法、应用与安全研究综述[J]. 计算机研究与发展, 2019, 56(10): 2071-2096.
JI S L, LI J F, DU T Y, et al. Survey on the techniques, applications and security of machine learning interpretability[J]. Journal of Computer Research and Development, 2019, 56(10): 2071-2096.
- [6] ZEILER M D, FERGUS R. Visualizing and understanding convolutional networks[C]//Proceedings of European Conference on Computer Vision-ECCV 2014. 2014: 818-833.
- [7] AGARWAL C, NGUYEN A. Explaining image classifiers by removing input features using generative models[C]//Proceedings of the Asian Conference on Computer Vision. 2020.
- [8] FONG R C, VEDALDI A. Interpretable explanations of black boxes by meaningful perturbation[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 3429-3437.
- [9] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 618-626.
- [10] CHATTOPADHAY A, SARKAR A, HOWLADER P, et al. Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks[C]//2018 IEEE Winter Conference on Applications of Computer Vision (WACV). 2018: 839-847.
- [11] WANG H, WANG Z, DU M, et al. Score-CAM: score-weighted visual explanations for convolutional neural networks[C]//Proceedings of the IEEE/CVF Conference on Computer vision and Pattern Recognition Workshops. 2020: 24-25.
- [12] JACOBSEN J H, ARNOLD W M S, EDOUARD O. I-RevNet: deep invertible networks[C]//Proceedings of International Conference on Learning Representations. 2018.
- [13] JING J, DENG X, XU M, et al. Hinet: deep image hiding by invertible network[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 4733-4742.
- [14] LUGMAYR A, DANELLJAN M, VAN GOOL L, et al. Srflow: learning the super-resolution space with normalizing flow[C]//Proceedings of Computer Vision-ECCV 2020. 2020: 715-732.
- [15] ZHOU B, KHOSLA A, LAPEDRIZA A, et al. Learning deep features for discriminative localization[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2921-2929.
- [16] SIMONYAN K, VEDALDI A, ZISSERMAN A. Deep inside convolutional networks: visualising image classification models and saliency maps[J]. arXiv Preprint arXiv: 1312. 6034, 2013.
- [17] SMILKOV D, THORAT N, KIM B, et al. Smoothgrad: removing noise by adding noise[J]. arXiv Preprint arXiv: 1706. 03825, 2017.
- [18] SPRINGENBERG J T, DOSOVITSKIY A, BROX T, et al. Striving for simplicity: the all convolutional net[J]. arXiv Preprint arXiv: 1412. 6806, 2014.
- [19] SUNDARARAJAN M, TALY A, YAN Q. Axiomatic attribution for deep networks[C]//International Conference on Machine Learning. 2017: 3319-3328.
- [20] PETSUK V, DAS A, SAENKO K. Rise: randomized input sampling for explanation of black-box models[J]. arXiv Preprint arXiv: 1806. 07421, 2018.
- [21] GOMEZ A N, REN M, URTASUN R, et al. The reversible residual network: backpropagation without storing activations[J]. Advances in Neural Information Processing Systems, 2017, 30.
- [22] RAMASWAMY H G. Ablation-cam: visual explanations for deep convolutional network via gradient-free localization[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2020: 983-991.
- [23] KRIZHEVSKY A, HINTON G. Learning multiple layers of features from tiny images[R]. 2009.
- [24] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. 2009: 248-255.

- [25] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv Preprint arXiv: 1706.06083, 2017.
- [26] DONG Y, LIAO F, PANG T, et al. Boosting adversarial attacks with momentum[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 9185-9193.
- [27] TRAMÈR F, KURAKIN A, PAPERNOT N, et al. Ensemble adversarial training: Attacks and defenses[J]. arXiv Preprint arXiv: 1705.07204, 2017.
- [28] KHAKZAR A, KHORSANDI P, NOBAHARI R, et al. Do explanations explain? Model knows best[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 10244-10253.
- [29] DABKOWSKI P, GAL Y. Real time image saliency for black box classifiers[C]//Advances in Neural Information Processing Systems. 2017, 30.



李钊效（1998-），男，白族，云南大理人，云南大学硕士生，主要研究方向为人工智能安全和文本隐写。



郑奕森（1999-），男，云南昆明人，云南大学硕士生，主要研究方向为人工智能安全和自动驾驶安全。

[作者简介]



牟新颖（1998-），女，山东烟台人，云南大学硕士生，主要研究方向为模型可解释性和人工智能安全。



周维（1974-），男，湖南桃源人，云南大学教授、博士生导师，主要研究方向为网络空间安全（人工智能安全）、分布式云计算。



宋冰冰（1994-），男，四川南充人，云南大学博士生，主要研究方向为人工智能安全和图像隐写。



董云云（1989-），女，云南保山人，云南大学讲师，主要研究方向为大数据索引、分布式计算、图像隐写。