基于对抗补丁的可泛化的 Grad-CAM 攻击方法

司念文[1]，张文林[1]，屈丹[1]，常禾雨[2]，李盛祥[1]，牛铜[1]
（1. 信息工程大学信息系统工程学院，河南 郑州 450001；2. 信息工程大学密码工程学院，河南 郑州 450001）

摘 要：为了验证 Grad-CAM 解释方法的脆弱性，提出了一种基于对抗补丁的 Grad-CAM 攻击方法。通过在 CNN 分类损失函数后添加对 Grad-CAM 类激活图的约束项，可以针对性地优化出一个对抗补丁并合成对抗图像。该对抗图像可在分类结果保持不变的情况下，使 Grad-CAM 解释结果偏向补丁区域，实现对解释结果的攻击。同时，通过在数据集上的批次训练及增加扰动范数约束，提升了对抗补丁的泛化性和多场景可用性。在 ILSVRC2012 数据集上的实验结果表明，与现有方法相比，所提方法能够在保持模型分类精度的同时，更简单有效地攻击 Grad-CAM 解释结果。

关键词：卷积神经网络；可解释性；对抗补丁；类激活图；显著图
中图分类号：TP391
文献标识码：A
DOI: 10.11959/j.issn.1000−436x.2021025

Generalized Grad-CAM attacking method based on adversarial patch

SI Nianwen[1], ZHANG Wenlin[1], QU Dan[1], CHANG Heyu[2], LI Shengxiang[1], NIU Tong[1]
1. Department of Information System Engineering, Information Engineering University, Zhengzhou 450001, China
2. Department of Cryptogram Engineering, Information Engineering University, Zhengzhou 450001, China

Abstract: To verify the fragility of the Grad-CAM, a Grad-CAM attack method based on adversarial patch was proposed. By adding a constraint to the Grad-CAM in the classification loss function, an adversarial patch could be optimized and the adversarial image could be synthesized. The adversarial image guided the Grad-CAM interpretation result towards the patch area while the classification result remains unchanged, so as to attack the interpretations. Meanwhile, through batch-training on the dataset and increasing perturbation norm constraint, the generalization and the multi-scene usability of the adversarial patch were improved. Experimental results on the ILSVRC2012 dataset show that compared with the existing methods, the proposed method can attack the interpretation results of the Grad-CAM more simply and effectively while maintaining the classification accuracy.

Keywords: convolutional neural network, interpretability, adversarial patch, class activation map, saliency map

1 引言

近年来，以卷积神经网络（CNN, convolutional neural network）为代表的深度学习技术在图像识别[1-3]和语言文本处理[4-5]等领域的研究应用取得了重大进展。与传统机器学习算法相比，深度学习模型的优势在于其优异的自动特征提取能力，这极大缓解了传统方法下人工特征设计的困难，使目标任务可以学习到更加全面的、含有丰富语义信息的组合式特征。然而，尽管深度学习模型的识别效果非常好，但其一直受到可解释性问题的困扰。其中，CNN 作为深度学习技术的代表，其工作机制及决策逻辑至

收稿日期：2020−10−20；修回日期：2020−12−22
通信作者：张文林，zwlin_2004@163.com
基金项目：国家自然科学基金资助项目（No.61673395）
Foundation Item: The National Natural Science Foundation of China (No.61673395)

今尚不能完全被人们理解，阻碍了其在一些对安全性要求高的领域的深入拓展。对 CNN 的理解与解释在理论和实际应用上都具有一定的研究价值，研究人员为此提出了一系列的可解释性方法，用于解释 CNN 的内部表征和决策，这些方法在一定程度上缓解了 CNN 可解释性较差的问题，增进了人们对 CNN 特征和决策的理解，提升了人们对 CNN 模型的信任度。

在基于 CNN 的图像分类领域，基于显著图的解释方法是一种典型的 CNN 解释方法，这种方法会生成一个与输入图像相对应的显著图，该图将与特定决策相关的输入特征高亮，用以表示对该决策结果的可视化解释。典型的基于显著图的解释方法如图 1 所示（左侧图像表示 Bull_masstiff 的定位结果，右侧图像表示 Tiger_cat 的定位结果），主要包括 2 种。一种是基于模型梯度（导数）的显著图（图 1(d)～图 1(g)），例如反向传播（BP, back propagation）[6]、导向反向传播（Guided BP, guided back propagation）[7]、平滑梯度（smooth gradient）[8]及积分梯度（integrated gradient）[9]。梯度构成的显著图噪声较多，且由于不具备类别区分性，导致它们无法针对性地分别解释不同类别目标相关的特征，因此可视化效果并不理想。另一种是基于类激活映射（CAM, class activation mapping）得到的类激活图（图 1(b)和图 1(c)），最早由 Zhou 等[10]提出。类激活图的主要优点体现在类别区分性上，可在图像级标签监督下，定位输入图像中目标的具体位置。由于具有较好的类别区分特性，因此 CAM 及其多种改进版本（如 Grad-CAM[11]、Grad-CAM++[12]及 Score-CAM[13]）在弱监督目标定位[10-11]、视觉问答[11]等众多场景中均有应用。

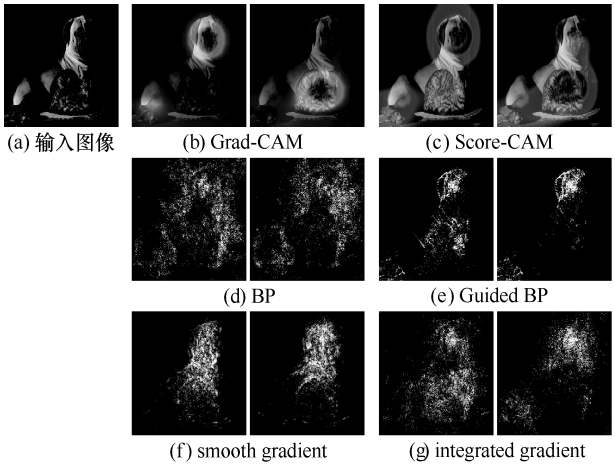


图 1 梯度图与类激活图的比较

Grad-CAM 作为类激活图方法中较稳定的一种，过程最简单且应用较广泛。然而，最近的研究表明，这些基于显著图的解释方法（如 Grad-CAM、BP 及 Guided BP 等）存在被攻击的危险[14-16]。文献[14]首次验证了 BP 和 integrated gradient 方法的脆弱性，通过最大化对抗图像的显著图与原图的显著图之间的差异，可以优化出一种专门用于攻击解释方法的对抗样本。文献[15]进一步研究了在特定损失函数的约束下，可使 BP、Guided BP 及 integrated gradient 等方法的解释结果中出现事先指定的无关特征。文献[16]则通过对抗性的微调模型参数，在不修改输入图像的情况下，使用参数微调后的模型引导 Grad-CAM 的解释结果总是偏向特定区域，实现无效的、甚至被引导至有意图偏向的解释。总体来看，文献[14-15]提出的攻击方法主要通过生成视觉变化不可感知的对抗样本来针对性地攻击解释结果。这种对抗样本虽然具有较好的伪装特性，但在现实中难以应用。文献[16]虽然不需要添加扰动来形成对抗样本，但其采用的微调参数方法需要重新训练模型，导致攻击的代价也较大。

对抗补丁是一种用于攻击模型的图像代替方法，可以不受扰动范数的限制，具有攻击过程简单、现实应用性强的优点，通常被用于现实场景的对抗攻击[17]。基于对抗补丁合成对抗图像来攻击模型的解释，在现实场景中更加方便。为此，本文提出一种基于对抗补丁的 Grad-CAM 攻击方法，将对抗补丁方法用于攻击针对模型的解释，而非攻击模型本身的预测。具体地，通过将对抗补丁添加在图像上，保证分类结果不变，但 Grad-CAM 解释结果始终偏向目标区域，以此实现对解释结果的攻击。实验结果表明，与现有的方法相比，本文方法使用一种新的思路实现对 CNN 解释的有效攻击，且过程更加简单。总体来讲，本文的贡献分为以下 3 个方面。

1) 提出了一种基于对抗补丁的 Grad-CAM 攻击方法。该方法能够对目标图像的 Grad-CAM 解释结果针对性地生成对抗补丁并合成对抗图像，用于攻击 Grad-CAM 解释方法，使之无法准确定位目标图像的显著性特征，从而产生错误的解释。

2) 在多种攻击场景下进一步扩展了该方法的应用。基于该方法的思路，将其扩展到通用的对抗补丁，以及视觉变化不可感知的对抗样本，提升了该方法的泛化性和多场景可用性。

3) 从攻击结果的视觉效果及目标区域的能量占比的角度，定性和定量评估了所提方法的攻击效果。以 4 种典型的 CNN 分类网络为例，在 ILSVRC2012 数据集[18]上进行了大量实验，结果表明，所提方法可以有效地实现对 Grad-CAM 的攻击。

2　相关工作

2.1　CNN 可解释性方法

近年来，对于 CNN 的可解释性引起了研究者的关注，提出了一系列可视化方法用于解释 CNN 的预测结果。最简单的可视化方法是基于梯度的方法[6-9]，但梯度图中通常含有大量的噪声问题，不具备类别区分特性。CAM 方法[10-13]是另一类 CNN 解释方法，因其具有较好的类别区分特性而被用于定位与 CNN 特定决策结果相关的图像区域。同时，由于仅使用图像级的标签即可实现一定效果的目标定位，因此类激活图方法也被用于弱监督目标定位任务[10-11]。此外，类激活图方法还被用来为图像分类[19]、语义分割[20]等任务提供弱先验信息，从而提升其性能。

Grad-CAM 方法的广泛应用导致其解释结果的稳定性至关重要，一旦 Grad-CAM 被攻击而产生错误的定位结果，基于 Grad-CAM 的后续任务将接连产生错误。文献[14]对基于梯度的解释方法和基于样本的解释方法进行了攻击，结果表明，这些可视化方法均存在一定程度的脆弱性。与文献[14]类似，文献[15]也对基于梯度的可视化方法进行了攻击。但由于 ReLU 网络的二阶梯度通常为 0，因此文献[14-15]的攻击方法需要先将目标网络的 ReLU 函数统一转换成 Softplus 函数，这严重限制了其实用性。受文献[14]的启发，文献[16]从模型的角度出发，将对显著图的攻击目标作为损失函数添加到模型训练中，通过一个微调步骤来微调模型参数，进行对抗性的模型调整，再使用调整后的模型对输入图像产生错误的 Grad-CAM 解释结果。然而，这种方法在数据集上需要重新训练模型，对于 ImageNet 这样的大型数据集来说，时间和计算资源消耗非常大，攻击成本高。

2.2　对抗补丁

对抗补丁是一种用于攻击神经网络图像分类系统的补丁图像。通过在目标图像上添加对抗性补丁，可以使用目标图像被神经网络分类模型误分类，实现对图像分类系统的攻击[17]。对抗补丁的优点是不受扰动范数的限制，可在较小的区域内实现较大的扰动。由于对抗补丁具有显著性特征，是造成图像被误分类的重点区域，因此普通的对抗补丁很容易被 Grad-CAM 等解释方法检测到。为此，文献[21]针对性地提出了一种稳健的对抗补丁方法，该方法在攻击图像分类结果的同时，可以抵抗 Grad-CAM 对补丁位置的检测。文献[21]使用对抗补丁的目的与文献[17]相同，均是用于攻击模型的分类结果，但其在损失函数中引入 Grad-CAM 约束，仅是为了提升对抗补丁的稳健性，防止补丁位置被 Grad-CAM 轻易检测到，但对抗补丁的主要目的仍然是用于攻击模型的分类结果。

与文献[17]和文献[21]不同，本文将对抗补丁用作专门攻击 CNN 模型的解释方法，而不是攻击图像的分类结果。如 2.1 节所述，现有针对 Grad-CAM 的攻击方法具有攻击成本高、攻击过程复杂等缺点。为了避免调整网络结构和重新训练目标网络，本文使用基于对抗补丁的方法实现对 Grad-CAM 的攻击。

3　本文方法介绍

3.1　Grad-CAM 原理介绍

Grad-CAM 方法由 Selvaraju 等[11]于 2017 年提出，是类激活图系列方法中最常用的一种。与同类型方法（如 CAM[10]、Grad-CAM++[12]及 SS-CAM[13]）相比，其实现过程最简单，对多种网络通用，可视化效果也相对较好。Grad-CAM 的原理如图 2 所示。

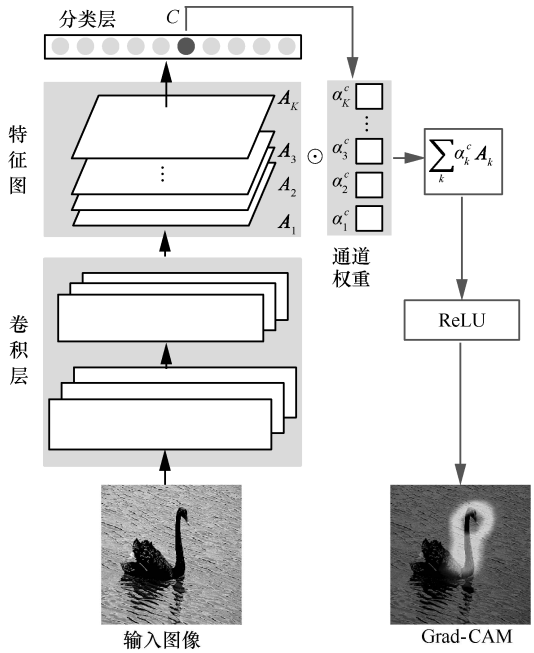


图 2　Grad-CAM 的原理

给定输入图像 $\boldsymbol{x}$ 和待攻击的目标网络 f，经过 f 得到输入图像的 logits 分数为

$$(S^1,\cdots,S^c,\cdots,S^N)=f(\boldsymbol{x};\theta) \tag{1}$$

其中，θ 表示 f 的权重参数，S^c 表示第 c 个类别的 logits 分数。该过程属于标准的 CNN 分类过程，仅能给出分类结果，无法解释 CNN 基于 $\boldsymbol{x}$ 的哪些输入特征得到了该分类结果。Grad-CAM 方法正是为了解释目标网络的分类结果而被提出的。由于目标网络 f 从输入图像 $\boldsymbol{x}$ 提取的最高层特征图 $(\boldsymbol{A}_1,\boldsymbol{A}_2,\cdots,\boldsymbol{A}_K)$ 在卷积层和全连接层之间达到平衡，具有较好的类别区分性，因此可以使用该层特征图来定位感兴趣的目标。

使用 Grad-CAM 方法生成类激活图的过程可形式化描述为

$$H^c_{\text{Grad-CAM}}(\boldsymbol{x})=\max\left(\sum_k \alpha_k^c \boldsymbol{A}_k,0\right) \tag{2}$$

其中，$\boldsymbol{A}_k$ 表示最高层特征图的第 k 个通道，α_k^c 表示该通道的权重，其计算式为

$$\alpha_k^c=\frac{1}{Z}\sum_{i,j}\frac{\partial S^c}{A_{k,i,j}} \tag{3}$$

其中，$A_{k,i,j}$ 表示第 k 个通道位于(i, j)的元素，Z 表示归一化因子。由于通道权重是从类别 c 的导数得来的，因此通道权重含有类别相关信息，这也是 Grad-CAM 能够针对不同决策结果进行解释的主要原因。

3.2 Grad-CAM 攻击方法

3.2.1 目标函数与优化

一般情形下，对抗补丁被用于误导图像分类结果，使添加对抗补丁的图像总是被分类器误分类。而在可解释深度学习领域，解释结果与分类结果同样重要，攻击者不仅会对模型分类结果进行攻击，还可能攻击模型的解释结果。基于此，本节提出了一种基于对抗补丁的针对 Grad-CAM 解释方法的攻击方法。给定输入图像 $\boldsymbol{x}$ 和目标网络 f，设 $\boldsymbol{m}\in\mathbb{R}^{H\times W\times C}$ 表示二值化的掩码，其中补丁区域为 1，其余区域为 0，$\boldsymbol{z}$ 表示扰动图像，对抗图像可由输入图像 $\boldsymbol{x}$、二值化掩码 $\boldsymbol{m}$ 及扰动图像 $\boldsymbol{z}$ 合成，合成过程为

$$\boldsymbol{x}'=\boldsymbol{x}\odot(1-\boldsymbol{m})+\boldsymbol{z}\odot\boldsymbol{m} \tag{4}$$

其中，$\odot$ 表示哈达玛积，$\boldsymbol{z}\odot\boldsymbol{m}$ 相当于对抗补丁。本文使用对抗补丁的目的不是误导图像分类结果，而是攻击解释方法的解释结果。具体来说，本文中补丁的作用是保持分类结果不变，同时引导 Grad-CAM 攻击结果始终偏向补丁区域，从而实现对 Grad-CAM 解释的攻击。这个过程实际上包含 2 个优化目标，介绍如下。

① 保持分类不变，对应的目标函数形式化描述为

$$\min_{\boldsymbol{z}}\text{loss}_{\text{CE}}(f^c(\boldsymbol{x}';\theta);c) \tag{5}$$

其中，c 表示原始分类类别。式(5)使用交叉熵损失约束对抗图像的分类结果保持不变。

② 引导 Grad-CAM 偏向补丁区域，对应的目标函数形式化描述为

$$\max_{\boldsymbol{z}}\sum_{i,j}(H^c_{\text{Grad-CAM}}(\boldsymbol{x}')\odot\boldsymbol{m})_{i,j} \tag{6}$$

式(6)相当于取出补丁区域的 Grad-CAM 显著图像素并求和，然后最大化该值。

综合以上 2 个优化目标，最终的目标函数 Loss 可定义为

$$\text{Loss}=\text{loss}_{\text{CE}}(f^c(\boldsymbol{x}';\theta);c)-\lambda\sum_{i,j}(H^c_{\text{Grad-CAM}}(\boldsymbol{x}')\odot\boldsymbol{m})_{i,j} \tag{7}$$

其中，λ 表示 2 个优化目标之间的调和参数。该目标函数优化的对象是扰动图像 $\boldsymbol{z}$，其更新方式采用梯度符号更新，即

$$\boldsymbol{z}'=\boldsymbol{z}-\text{lr}\times\text{sign}(\nabla_{\boldsymbol{x}'}\text{Loss}) \tag{8}$$

其中，lr 表示更新时的学习率；sign 表示符号函数，值域为$\{+1,-1\}$。

3.2.2 攻击算法流程

算法 1 给出了基于对抗补丁的 Grad-CAM 攻击算法的流程。总体来讲，在给定输入图像 $\boldsymbol{x}$ 和二值化掩码 $\boldsymbol{m}$ 的情况下，可以得到添加对抗补丁的扰动图像 $\boldsymbol{x}'$，该对抗图像使用 Grad-CAM 解释方法始终无法准确定位到显著性目标。同时，通过二值化掩码 $\boldsymbol{m}$ 可以控制补丁添加的位置，从而控制对 Grad-CAM 的引导偏向。具体来讲，该算法包括以下 5 个步骤。步骤 1)初始化扰动 $\boldsymbol{z}$；步骤 2)生成对抗图像 $\boldsymbol{x}'$；步骤 3)计算对抗图像的得分 $f^c(\boldsymbol{x}';\theta)$，以及其对应的显著图 $H^c_{\text{Grad-CAM}}(\boldsymbol{x}')$；步骤 4)计算损失函数 Loss；步骤 5)使用梯度符号更新扰动 $\boldsymbol{z}$。通过不断迭代来更新扰动 $\boldsymbol{z}$ 及降低损失值。其中，num_iters 表示迭代次数。

值得注意的是，本文虽然仅在 Grad-CAM 方法上使用该算法进行实验，但对于类激活映射这一类方法，包括 CAM、Grad-CAM++等，该算法也同样适用。

算法 1　基于对抗补丁的 Grad-CAM 攻击算法

输入　输入图像 $\boldsymbol{x}$，二值化掩码 $\boldsymbol{m}$

输出　对抗图像 $\boldsymbol{x}'$

Begin

1) 初始化扰动 $\boldsymbol{z}$

for $i=1,2,\cdots,\text{num\_iters}$ do

2) 生成对抗图像 $\boldsymbol{x}'=\boldsymbol{x}\odot(1-\boldsymbol{m})+\boldsymbol{z}\odot\boldsymbol{m}$

3) 计算 $\boldsymbol{x}'$ 的分数 $f^c(\boldsymbol{x}';\theta)$、Grad-CAM 显著图 $H^c_{\text{Grad-CAM}}(\boldsymbol{x}')$

4) 计算损失函数

$$\text{Loss}=\text{loss}_{\text{CE}}(f^c(\boldsymbol{x}';\theta);c)-\lambda\sum_{i,j}(H^c_{\text{Grad-CAM}}(\boldsymbol{x}')\odot\boldsymbol{m})_{i,j}$$

5) 更新 $\boldsymbol{z}$:　$\boldsymbol{z}'=\boldsymbol{z}-\text{lr}\times\text{sign}(\nabla_{x'}\text{Loss})$

end for

3.3　可泛化的通用对抗补丁

第 3.2 节所述的对抗补丁仅能针对性地对单张图片进行 Grad-CAM 攻击，每张图片都有其对应的对抗补丁。因此，这种对抗补丁具有图像针对性，对于未知的新图像，攻击效果并不一定好。

为了进一步提升本文的对抗补丁方法的泛化性，将其应用于同一类别的其他图像，本节通过进一步改进，使用批次训练方法来生成通用对抗补丁，使通用对抗补丁可以面向未知的新样本，即在同一类别的样本下，对未见过的新样本进行 Grad-CAM 攻击。

在算法 1 的框架下，仅需修改步骤 4)中的目标函数，即可生成可泛化的通用对抗补丁。具体地，通用对抗补丁的生成可使用如下目标函数

$$\text{Loss}=\frac{1}{N}\sum_{n=1}^{N}\left[\text{loss}_{\text{CE}}(f^c(\boldsymbol{x}'_n;\theta);c)-\lambda\sum_{i,j}(H^c_{\text{Grad-CAM}}(\boldsymbol{x}'_n)\odot\boldsymbol{m})_{i,j}\right] \tag{9}$$

其中，N 表示批次大小，每张对抗图像 $\boldsymbol{x}'_n$ 均由对应的输入图像 $\boldsymbol{x}_n$、二值化掩码 $\boldsymbol{m}$ 及通用的扰动图像 $\boldsymbol{z}$ 合成。该目标函数的更新对象仍为扰动图像 $\boldsymbol{z}$，这里每张输入图像 $\boldsymbol{x}_n$ 共用同一个扰动图像 $\boldsymbol{z}$。

将得到的扰动图像 $\boldsymbol{z}$ 作为每个类别图像的通用扰动，即可添加在该类别未知的目标图像上，实现对目标图像的 Grad-CAM 攻击。值得注意的是，在实验中尝试将对抗补丁扩展到不同类别的图像，使用不同类别图像进行训练，但测试效果并不好，其中原因值得进一步深入分析。

3.4　扩展到对抗样本

尽管本文主要基于对抗补丁来攻击 Grad-CAM 的解释结果，但仅需要较小修改，即可将本文方法用于生成对抗样本。本节进一步对上述方法进行拓展，将对抗补丁攻击方法拓展为对抗样本攻击方法。用于攻击 Grad-CAM 的对抗样本是指通过在整个原图区域添加细微扰动，可以使目标图像的分类结果保持不变，但 Grad-CAM 的定位结果却发生改变，引向特定目标区域。

为了实现上述目标，将二值化掩码 $\boldsymbol{m}$ 的 1 值扩展到整个图像区域，即 $\boldsymbol{m}$ 的元素值全为 1，按照如下方法得到新的对抗图像，即

$$\boldsymbol{x}'=\boldsymbol{x}\odot\boldsymbol{m}+\boldsymbol{z}\odot\boldsymbol{m} \tag{10}$$

$$\text{s.t. }\|\boldsymbol{z}\|_2\leqslant\varepsilon \tag{11}$$

式(10)中添加了对扰动图像 $\boldsymbol{z}$ 的 L2 范数约束，目的是使添加在图像上的扰动在视觉上尽量不易被察觉，以不改变原图的内容。此时，按照式(7)中的目标函数形式，添加扰动损失后，得到的目标函数为

$$\text{Loss}=\text{loss}_{\text{CE}}(f^c(\boldsymbol{x}';\theta);c)-\lambda_1\sum_{i,j}(H^c_{\text{Grad-CAM}}(\boldsymbol{x}')\odot\boldsymbol{m}')_{i,j}+\lambda_2\|\boldsymbol{z}\|_2 \tag{12}$$

其中，$\boldsymbol{m}'$ 表示用于引导攻击位置的二值化掩码，其中为 1 的区域表示将引导 Grad-CAM 显著图偏向该区域。式(7)中的二值掩码 $\boldsymbol{m}$ 在表示补丁区域位置的同时，也将 Grad-CAM 显著图引导偏向该区域。这里的 $\boldsymbol{m}'$ 与之不同，$\boldsymbol{m}'$ 的 1 值区域由于不受补丁位置的限制，可以移向任意位置。第 4 节的实验中将展示 3 种不同的 $\boldsymbol{m}'$ 所引导的不同 Grad-CAM 攻击结果，其中，λ_1 和 λ_2 为后两项的调和系数。对扰动图像 $\boldsymbol{z}$ 的更新仍采用式(8)中的梯度符号更新方法。

注意，使用本节方法生成的对抗样本与 Szegedy 等[22]和 Goodfellow 等[23]的对抗样本的作用并不相同。Szegedy 等[22]和 Goodfellow 等[23]的对抗样本用于攻击模型的预测结果，而本文的对抗样本的攻击目标并不是模型的预测结果，而是专门用攻击模型的解释结果，即 Grad-CAM 的定位结果。

4 实验

4.1 实验设置

1) 数据集与目标模型。本文使用 ILSVRC2012 数据集[18]作为实验数据集，为了便于对比实验结果，每部分实验将分别使用该数据集中的不同部分。目标网络采用 4 种常见的 CNN 图像分类网络：VGGNet-16[1]、VGGNet-19-BN[1]、ResNet-50[2]、DenseNet-161[3]，来自 Torchvision 包[24]中自带的预训练网络模型，在上述数据集上完成了预训练。

2) 攻击效果评价指标。攻击效果的评价主要从定性和定量这 2 个方面进行评价：视觉效果和能量占比（ER, energy ratio）。视觉效果表示从视觉上直接观察 Grad-CAM 方法的解释结果，即可视化结果的直观视觉感受。此外，使用 ER 值作为评价指标，量化 Grad-CAM 的可视化结果。ER 值表示显著图中某个区域的能量占整个显著图能量的比例，计算式为

$$\mathrm{ER}=\frac{\sum_{i,j\in T}\left(H^{c}_{\text{Grad-CAM}}(x)\right)_{i,j}}{\sum_{i,j\in T}\left(H^{c}_{\text{Grad-CAM}}(x)\right)_{i,j}+\sum_{i,j\notin T}\left(H^{c}_{\text{Grad-CAM}}(x)\right)_{i,j}}\times 100\% \tag{13}$$

其中，T 表示目标区域的像素构成的集合。显然，ER 值越大，表示 Grad-CAM 对该区域的关注度越高。具体地，实验中需要计算 2 个目标区域的 ER 值：补丁区域的 ER 值（$\mathrm{ER_p}$）和边框区域的 ER 值（$\mathrm{ER_b}$），计算式分别为

$$\mathrm{ER_p}=\frac{\sum_{i,j\in \text{Patch}}\left(H^{c}_{\text{Grad-CAM}}(x)\right)_{i,j}}{\sum_{i,j\in \text{Patch}}\left(H^{c}_{\text{Grad-CAM}}(x)\right)_{i,j}+\sum_{i,j\notin \text{Patch}}\left(H^{c}_{\text{Grad-CAM}}(x)\right)_{i,j}}\times 100\% \tag{14}$$

$$\mathrm{ER_b}=\frac{\sum_{i,j\in \text{Bndbox}}\left(H^{c}_{\text{Grad-CAM}}(x)\right)_{i,j}}{\sum_{i,j\in \text{Bndbox}}\left(H^{c}_{\text{Grad-CAM}}(x)\right)_{i,j}+\sum_{i,j\notin \text{Bndbox}}\left(H^{c}_{\text{Grad-CAM}}(x)\right)_{i,j}}\times 100\% \tag{15}$$

由于 Grad-CAM 本身用于对图中显著性目标进行定位，因此对于未受到任何攻击的 Grad-CAM 显著图，$\mathrm{ER_b}$ 较高。而本文提出的基于对抗补丁的 Grad-CAM 方法的目的就是将 Grad-CAM 的定位区域引向补丁区域，因此对于本文方法，$\mathrm{ER_p}$ 越高，表明攻击效果越好。

4.2 攻击结果与分析

本节实验中，参照文献[25]的对抗样本研究，使用从 ILSVRC2012 验证集中 1 000 个类别中选择的 1 000 张图片，每个类别含有一张图片。VGGNet-19-BN 模型上的 top1 准确率和 ER 值如表 1 所示。其中，top1 准确率表示对图像的 top1 分类准确率。对于文献[16]的对抗性微调方法，使用其提供的源代码进行了结果复现。验证集使用上述 1 000 张图像，计算针对该 1 000 张图像的 top1 准确率及 ER 值。实验结果分别进行以下对比。

表 1 VGGNet-19-BN 模型上的 top1 准确率和 ER 值

方法	top1 准确率	$\mathrm{ER_p}$	$\mathrm{ER_b}$
原图	92.70%	4.85%	67.76%
对抗性微调方法	87.90%	71.13% ↑	35.42% ↓
对抗补丁方法（本文方法）	92.50%	67.19% ↑	38.52% ↓

1) 原图：在 VGGNet-19-BN 上的 top1 准确率，原图的 Grad-CAM 显著图的 $\mathrm{ER_p}$ 和 $\mathrm{ER_b}$。

2) 对抗性微调方法[16]：使用微调后的 VGGNet-19-BN 模型对这 1 000 张原图的 top1 准确率，微调后模型的 Grad-CAM 显著图的 $\mathrm{ER_p}$ 和 $\mathrm{ER_b}$。

3) 对抗补丁方法（本文方法）：使用对抗补丁产生的对抗图像在 VGGNet-19-BN 上的 top1 准确率，对抗图像的 Grad-CAM 显著图的 $\mathrm{ER_p}$ 和 $\mathrm{ER_b}$。

对于 top1 准确率，如表 1 所示，在 VGGNet-19-BN 模型上本文方法的准确率仅下降 0.2%，即只有 2 张图片的类别未保持原来的类别。而对抗性微调方法由于对模型参数进行了重新训练，导致其分类准确率下降较多。

对于 ER 值，如表 1 所示，本文方法和对抗性微调方法均可使 $\mathrm{ER_p}$ 上升且 $\mathrm{ER_b}$ 下降。对于本文方法，$\mathrm{ER_p}$ 上升 62.34%，$\mathrm{ER_b}$ 下降 29.24%，这表明本文方法对 Grad-CAM 定位结果的引导是有效的，使补丁区域受到了 Grad-CAM 的更多关注，而使目标本身（边框区域）的关注度下降许多。本文方法产生的对抗图像的 $\mathrm{ER_p}$ 仅占整个显著图的约 2/3（67.19%），但其与原图 $\mathrm{ER_b}$（67.76%）相比非常接近（对于 ILSVRC2012 数据集，一般的边框尺寸会比本文的补丁尺寸大），表明尺寸较小的补丁区域

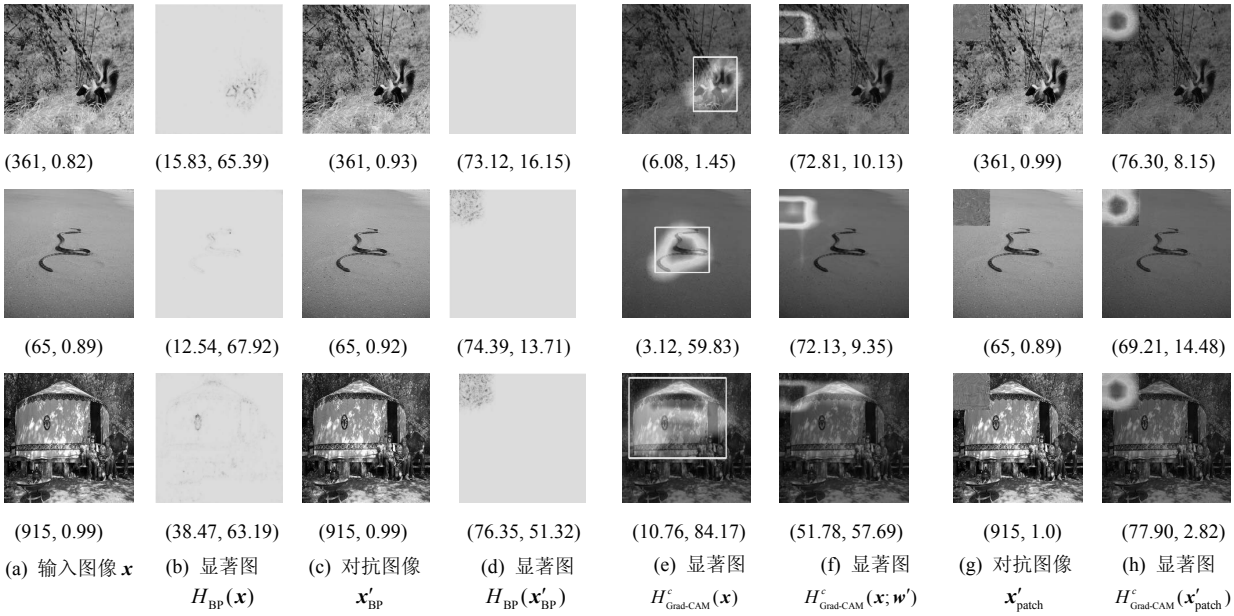


图 3　VGGNet-19-BN 模型上的 Grad-CAM 攻击结果比较

吸引了较多的能量关注，基本上整张图的主要关注点都移到了补丁区域。值得一提的是，$\mathrm{ER_b}$ 之所以仍较高，是因为许多边框区域会与补丁区域重叠，导致补丁区域的能量也被计入边框区域。此时，对抗图像的 $\mathrm{ER_p}$ 与 $\mathrm{ER_b}$ 之和超过 1，因为这 2 个区域之间有重叠部分。

图 3 为部分示例图像及其对抗图像的可视化结果，具体含义如下。

① 第 1 组（图 3(a)）表示初始的输入图像 $\boldsymbol{x}$，形式为(类别, 分类概率)。

② 第 2 组（图 3(b)～图 3(d)）表示对基于梯度的可视化方法的攻击结果，使用了文献[15]提供的攻击方法。其中，$H_{\mathrm{BP}}(\boldsymbol{x})$ 表示输入图像 $\boldsymbol{x}$ 的 BP 可视化结果，形式为($\mathrm{ER_p}$, $\mathrm{ER_b}$)值；$\boldsymbol{x}'_{\mathrm{BP}}$ 表示对 BP 可视化方法进行攻击，诱导其显著性区域偏向左上角优化出的对抗图像，形式为(类别, 分类概率)；$H_{\mathrm{BP}}(\boldsymbol{x}'_{\mathrm{BP}})$ 表示对抗图像 $\boldsymbol{x}'_{\mathrm{BP}}$ 的 BP 可视化结果，形式为($\mathrm{ER_p}$, $\mathrm{ER_b}$)值。

③ 第 3 组（图 3(e)～图 3(h)）表示对 Grad-CAM 可视化方法的攻击结果，包括对抗性微调方法和本文方法的结果。其中，$H^{c}_{\text{Grad-CAM}}(\boldsymbol{x})$ 表示输入图像 $\boldsymbol{x}$ 的 Grad-CAM 结果，形式为($\mathrm{ER_p}$, $\mathrm{ER_b}$)值；$H^{c}_{\text{Grad-CAM}}(\boldsymbol{x};\boldsymbol{w}')$ 表示使用对抗性微调方法重训练模型（$\boldsymbol{w}'$ 作为微调后模型的标记）后，得到的 Grad-CAM 结果，形式为($\mathrm{ER_p}$, $\mathrm{ER_b}$)值；$\boldsymbol{x}'_{\mathrm{patch}}$ 表示本文方法生成的对抗图像，形式为(类别, 分类概率)；$H^{c}_{\text{Grad-CAM}}(\boldsymbol{x}'_{\mathrm{patch}})$ 表示对抗图像 $\boldsymbol{x}'_{\mathrm{patch}}$ 的 Grad-CAM 结果，形式为($\mathrm{ER_p}$, $\mathrm{ER_b}$)值。

从对 BP 可视化方法的攻击结果（图 3 第 2 组）来看，攻击结果中隐藏了目标主体的位置，显著图偏向左上角区域，可视化结果具有梯度图的散点效果。由于基于梯度的可视化方法与 Grad-CAM 方法的可视化效果有较大区别，因此这里仅将其作为参考，相互之间并不具有很好的对比性。另一方面，从对 Grad-CAM 的攻击结果（图 3 第 3 组）可以看出，本文方法与对抗性微调方法均可实现有效攻击。与对抗性微调方法相比，本文方法的攻击效果较好，能够较明显地引导目标的定位偏向左上角补丁区域。同时，本文方法不需要重新训练模型，攻击过程更加简单。

4.3　不同模型上的攻击结果比较

为了测试该攻击方法在其他模型上的有效性，本节使用另外 3 种常见的 CNN 图像分类模型，分别为 VGGNet-16、ResNet-50 及 DenseNet-161，并与 4.2 节 VGGNet-19-BN 进行了效果对比。使用与 4.2 节相同的数据集和评价指标，实验测得结果如表 2 所示。从表 2 中的结果来看，本文方法对不同网络均可实现有效攻击。在 4 种不同模型上的 $\mathrm{ER_p}$ 值均有提升，且 $\mathrm{ER_b}$ 值均降低。与其他 3 种网络相比，ResNet-50 网络的攻击结果相对较差，对抗图像的可视化结果中 $\mathrm{ER_b}$ 仍然相对较高，推测可能是 ResNet-50 的 Grad-CAM 可视化结果本身定位更加

表 2　　在 4 种不同模型上的 top1 准确率与 ER 值比较

模型	方法	top1 准确率	ER_p	ER_b
VGGNet-16	原图	90.60%	5.67%	62.05%
	对抗图像（本文方法）	89.30%	64.94% ↑	40.19% ↓
VGGNet-19-BN	原图	92.70%	4.85%	67.76%
	对抗图像（本文方法）	92.50%	67.19% ↑	38.52% ↓
ResNet-50	原图	94.50%	2.97%	63.99%
	对抗图像（本文方法）	94.40%	33.73% ↑	47.32% ↓
DenseNet-161	原图	96.60%	4.00%	65.37%
	对抗图像（本文方法）	96.20%	38.51% ↑	45.41% ↓

全面，因此对抗图像的可视化结果会包含主体图像更多区域。

图 4 为 2 张示例图像的可视化结果。图 4(a)～图 4(d)为输入图像“junco”及其相应结果。图 4(a)为输入图像。图 4(b)为输入图像的 Grad-CAM，形式为(ER_p, ER_b)值。图 4(c)是由输入图像生成的对抗图像。图 4(d)为对抗图像的 Grad-CAM，形式为(ER_p, ER_b)值。图 4(e)～图 4(f)为另一张输入图像“espresso”及其相应结果。由图 4 可以看出，本文方法对这 4 种网络的 Grad-CAM 解释均有较好的攻击效果。虽然 ResNet-50 的攻击结果相对较差，但仍可以很好地引导 Grad-CAM 定位偏向补丁区域。

4.4　通用对抗补丁实验

虽然单张图片的对抗补丁对于自身非常有效，但对于同类别的一些其他图像的攻击效果不够好，导致 Grad-CAM 仍能检测到目标区域。如图 5 所示，图 5(a)表示原图及其 Grad-CAM。图 5(b)表示原图的对抗图像及其 Grad-CAM，该对抗图像由原图生成的对抗补丁并添加在原图上形成。图 5(c)表示针对原图生成的对抗补丁添加在另一张图像 $\boldsymbol{x}_1$ 上，形

图 4　在 4 种不同模型上的 Grad-CAM 攻击结果比较

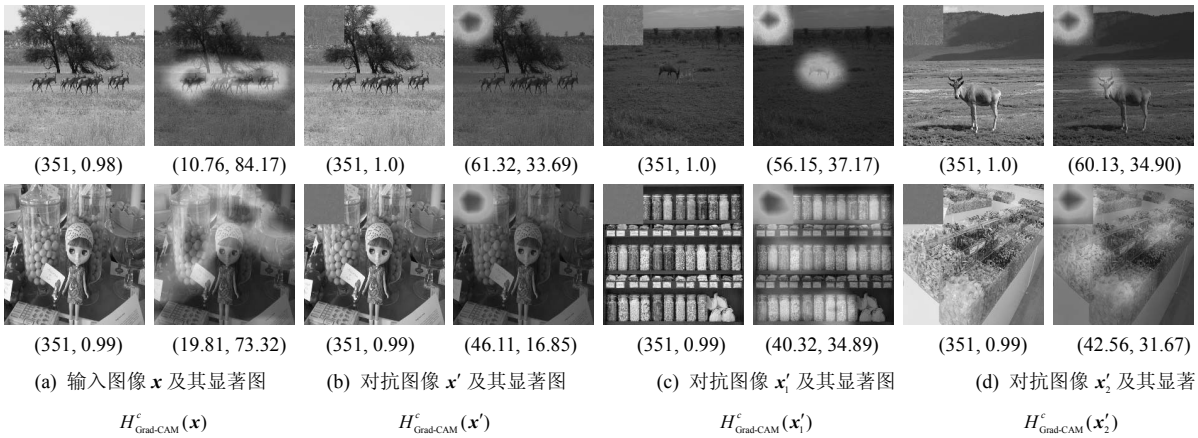


图 5　单张图像的对抗补丁的泛化性能测试示例

成对抗图像 $\boldsymbol{x}'_1$。由于对抗补丁的生成过程中，并未使用图像 $\boldsymbol{x}_1$ 的信息，因此并不能较好地对图像 $\boldsymbol{x}_1$ 的 Grad-CAM 解释结果进行攻击。同理，图 5(d)中对抗图像 $\boldsymbol{x}_2$ 也出现了该问题。可以看出，针对原图生成的对抗补丁，对原图自身的攻击效果非常好，但其泛化效果却不够好，对抗图像 $\boldsymbol{x}'_1$ 和 $\boldsymbol{x}'_2$ 的 Grad-CAM 图仍然可以检测到含有目标的区域。

为了提升本文对抗补丁的泛化性能，使其能够对未见过的图像进行攻击，本节按照第 3.3 节的方法对对抗补丁方法进行进一步优化。从 ILSVRC2012 数据集中选择 10 个类别的图像，具体类别如表 3 所示。其中，训练集的每个类别含有 1 300 张图片，共 13 000 张图片；测试集的每个类别含有 50 张图片，共 500 张图片；设置批次大小为 64。按照上述方法，得到优化前后的每张图片的 ER 值。图 6 是对其中的“indigo_bunting”和“hartebeest”的各 50 张测试集图片的 ER 值绘制的箱线图。从图 6 中可以看出，可泛化的对抗补丁对应的平均 ER_p 上升，ER_b 下降，表明可泛化的对抗补丁攻击效果更好。表 3 定量测试了这 10 个类别图像的泛化前后的对抗补丁的攻击效果。结果显示，与单张图片的对抗补丁方法相比，使用批次训练方法得到的每一类图像的对抗补丁的攻击效果更好，泛化性能更强。

4.5　对抗样本实验

本节实验将验证第 3.4 节中的对抗样本方法。使用与第 4.2 节相同的数据集，对 1 000 张图片生成对抗样本。通用调整 $\boldsymbol{m}'$ 的 1 值区域，可实现面向不同区域的攻击。实验中，分别测试了 3 种不同的攻击方法的效果。

1) 左上角攻击：将对抗样本的 Grad-CAM 解释结果引导偏向图像的左上角区域，与前文的补丁区域位置相同。

2) 右下角攻击：将 Grad-CAM 解释结果引导偏向右下角。

表 3　　**使用批次训练得到的可泛化的通用对抗补丁的攻击效果比较**

类别	单张图像的对抗补丁		可泛化的通用对抗补丁	
	ER_p	ER_b	ER_p	ER_b
airliner	67.32%	32.24%	68.56% ↑	30.26% ↓
sports_car	63.56%	35.61%	65.32% ↑	34.12% ↓
indigo_bunting	68.39%	14.99%	70.45% ↑	13.79% ↓
tabby	69.51%	37.26%	71.23% ↑	36.56% ↓
hartebeest	50.07%	13.28%	51.89% ↑	12.76% ↓
golden_retriever	62.17%	26.51%	63.78% ↑	25.78% ↓
bullfrog	59.34%	19.30%	60.45% ↑	18.32% ↓
sorrel	65.86%	32.19%	67.49% ↑	31.37% ↓
speedboat	63.39%	26.27%	65.02% ↑	24.53% ↓
pickup	67.85%	37.16%	69.30% ↑	36.52% ↓

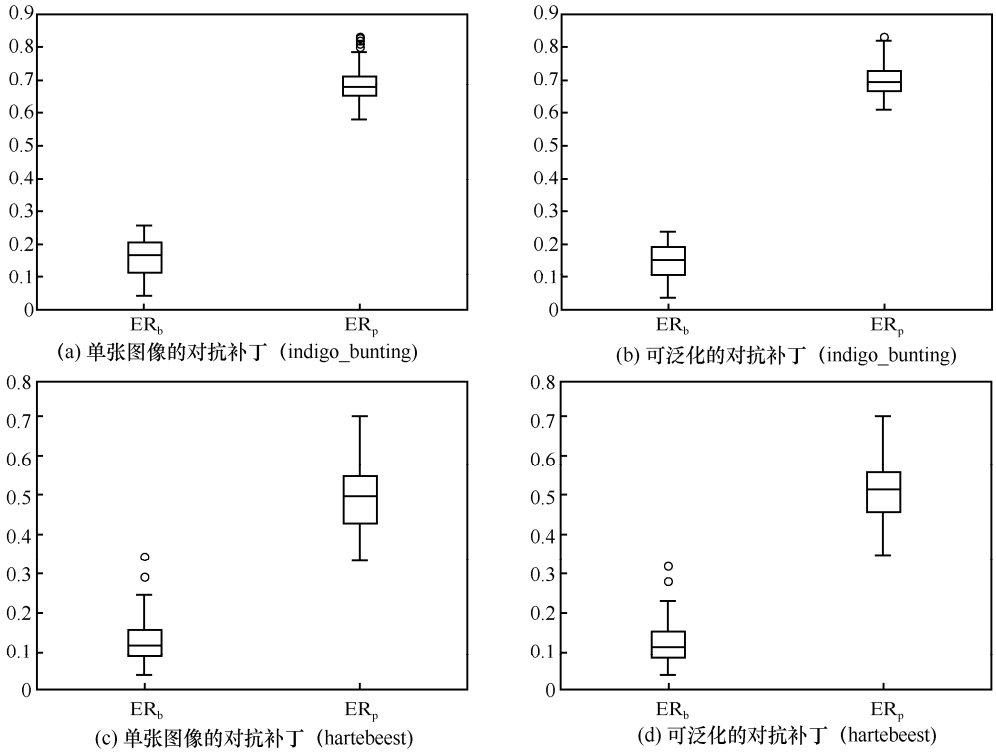


图 6　单张图像的对抗补丁与可泛化的通用对抗补丁的结果对比

3) 四周攻击：将 Grad-CAM 解释结果引导偏向图像的四周。

在上述 3 种攻击方法下分别生成相应的对抗样本，计算对抗样本的 Grad-CAM 显著图中的 ER_p 和 ER_b 值。表 4 为在 VGGNet-16 和 VGGNet-19-BN 网络上得到的定量结果。相对于原图，对抗样本的 ER_p 提升且 ER_b 值下降。同时，对抗样本的分类准确率仍保持不变，可见这种对抗样本更能从整体上扰动图像，而不受补丁区域的限制，从而不改变其分类结果，与对抗补丁方法相比具有更大优势。

图 7 为 4 组示例图像及相应的对抗样本。其中，$\boldsymbol{x}$、$\boldsymbol{x}'_1$、$\boldsymbol{x}'_2$、$\boldsymbol{x}'_3$ 分别表示原图、左上角攻击生成的对抗样本、右下角攻击生成的对抗样本、四周攻击生成的对抗样本。从图 7 中的结果可以看出，在对抗样本相对于原图没有显著变化的情况下，每种攻击方法均可成功地引导 Grad-CAM 解释结果偏向目标区域。

5　讨论

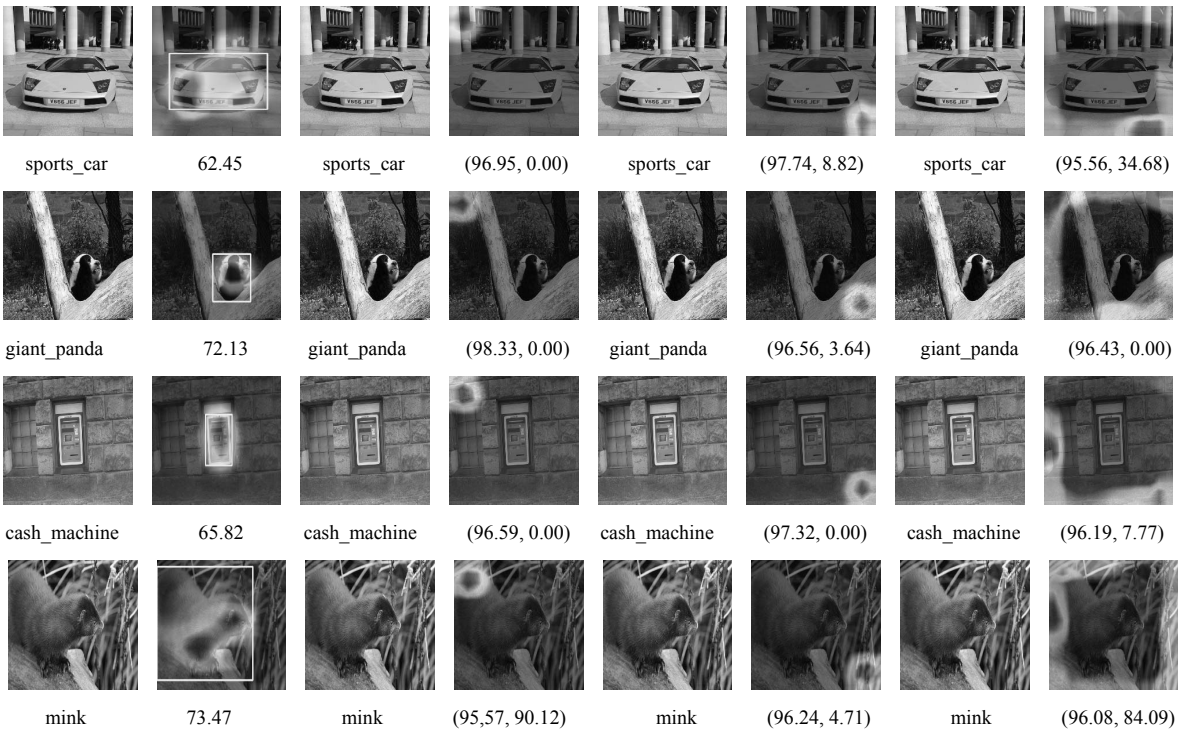
5.1　Grad-CAM 攻击方法的定性分析

本节将从梯度更新的角度，定性分析攻击 Grad-CAM 类激活图方法的原理及与攻击基于梯度的可视化方法之间的不同之处。

观察式(7)的损失函数形式。其中，第 1 项表示分类模型的交叉熵损失，第 2 项表示对 Grad-CAM 类激活图的值损失。对于交叉熵损失，作用在输入变量 x'_t 上的更新量为 $\frac{\partial S^c}{\partial x'_t}$。对于 Grad-CAM 显著图的损失，作用在输入变量 x'_t 上的更新量为

表 4　**VGGNet-16 和 VGGNet-19-BN 模型上的 top1 准确率和 ER 值**

模型	方法	左上角			右下角			四周		
		top1 准确率	ER_p	ER_b	top1 准确率	ER_p	ER_b	top1 准确率	ER_p	ER_b
VGGNet-16	原图	90.60%	5.67%	62.05%	90.60%	6.23%	62.05%	90.60%	9.38%	62.05%
	对抗图像（本文方法）	90.60%	95.75%↑	29.26%↓	90.40%	95.76%↑	35.68%↓	90.60%	96.32%↑	40.21%↓
VGGNet-19-BN	原图	92.70%	4.85%	67.76%	92.70%	5.39%	67.76%	92.70%	9.24%	67.76%
	对抗图像（本文方法）	92.70%	97.08%↑	26.62%↓	92.70%	96.74%↑	33.44%↓	92.70%	96.13%↑	39.25%↓



(a) 输入图像 $\boldsymbol{x}$ 及其显著图 $H^c_{\text{Grad-CAM}}(\boldsymbol{x})$　(b) 左上角对抗样本 $\boldsymbol{x}'_1$ 及其显著图 $H^c_{\text{Grad-CAM}}(\boldsymbol{x}'_1)$　(c) 右下角对抗样本 $\boldsymbol{x}'_2$ 及其显著图 $H^c_{\text{Grad-CAM}}(\boldsymbol{x}'_2)$　(d) 四周对抗样本 $\boldsymbol{x}'_3$ 及其显著图 $H^c_{\text{Grad-CAM}}(\boldsymbol{x}'_3)$

图 **7**　在 VGGNet-19-BN 模型上使用 3 种不同的攻击方法生成的对抗样本及其 Grad-CAM 结果

$$\frac{\partial H^c_{\text{Grad-CAM}}(x')}{\partial x'_t}=\frac{\partial \max(\sum_k \alpha^c_k \boldsymbol{A}_k,0)}{\partial x'_t} \tag{16}$$

由于 max 函数的作用是过滤激活图中的负值，并不影响导数的计算过程，因此可忽略 max 函数的影响，将其进一步展开为单个变量的导数，即

$$\frac{\partial (H^c_{\text{Grad-CAM}}(x'))_{i,j}}{\partial x'_t}=$$
$$\frac{\partial(\alpha^c_1 A_{1,i,j}+\alpha^c_2 A_{2,i,j}+\cdots+\alpha^c_K A_{K,i,j})}{\partial x'_t}=$$
$$\left(\frac{\partial \alpha^c_1}{\partial x'_t}A_{1,i,j}+\alpha^c_1\frac{\partial A_{1,i,j}}{\partial x'_t}\right)+\left(\frac{\partial \alpha^c_2}{\partial x'_t}A_{2,i,j}+\alpha^c_2\frac{\partial A_{2,i,j}}{\partial x'_t}\right)+\cdots+$$
$$\left(\frac{\partial \alpha^c_K}{\partial x'_t}A_{K,i,j}+\alpha^c_K\frac{\partial A_{K,i,j}}{\partial x'_t}\right) \tag{17}$$

对于式(17)中每个括号中的第 1 项，$\alpha^c_k(k=0,1,\cdots,K)$ 为输出分值 S^c 的一阶导数，结合式(3)，可将 $\frac{\partial \alpha^c_k}{\partial x'_t}$ 项化简为

$$\frac{\partial \alpha^c_k}{\partial x'_t}=\frac{1}{Z}\sum_{i,j}\frac{\partial\left(\frac{\partial S^c}{\partial A_{k,i,j}}\right)}{\partial x'_t}=$$
$$\frac{1}{Z}\sum_{i,j}\left(\frac{\partial\left(\frac{\partial S^c}{\partial A_{k,i,j}}\right)}{\partial A_{k,i,j}}\frac{\partial A_{k,i,j}}{\partial x'_t}\right)=\frac{1}{Z}\sum_{i,j}\left(\frac{\partial^2 S^c}{\partial^2 A_{k,i,j}}\frac{\partial A_{k,i,j}}{\partial x'_t}\right) \tag{18}$$

对于含有 ReLU 激活函数的 CNN，ReLU 函数的二阶导数几乎处处为 0。因此，$\frac{\partial^2 S^c}{\partial^2 A_{k,i,j}}$ 项总为 0，则 $\frac{\partial \alpha^c_k}{\partial x'_t}$ 总为 0。因此，式(17)的每个括号中的第 1 项总为 0。

对于式(17)的每个括号中的第 2 项，进一步化简为

$$\alpha^c_k\frac{\partial A_{k,i,j}}{\partial x'_t}=\left(\frac{1}{Z}\sum_{i,j}\frac{\partial S^c}{A_{k,i,j}}\right)\frac{\partial A_{k,i,j}}{\partial x'_t}=$$
$$\frac{1}{Z}\sum_{i,j}\left(\frac{\partial S^c}{A_{k,i,j}}\frac{\partial A_{k,i,j}}{\partial x'_t}\right)=\frac{1}{Z}\sum_{i,j}\frac{\partial S^c}{\partial x'_t} \tag{19}$$

由式(19)可知，该项与输入变量 x'_t 的一阶导数相关。综合式(18)和式(19)的分析结果可知，对于单个像素 x'_t 的更新量 $\frac{\partial H^c_{\text{Grad-CAM}}(x')_{i,j}}{\partial x'_t}$，其实际结果仍然是输入变量 x'_t 的一阶导数 $\frac{\partial S^c}{\partial x'_t}$，由于不含二阶导数，因此对普通的 ReLU 网络来说可以进行更新。

同样地，利用上述分析过程对基于梯度的 CNN 可视化方法进行分析。对于基于梯度的 CNN 可视化方法，例如 BP、Guided BP、Integrated gradient、Smooth gradient 等方法，由于 $H^c_{\text{BP}}(x')$ 本身即为输入图像 x' 的一阶导数，因此对 BP 显著图进行攻击时，单个输入变量的 $\frac{\partial H^c_{\text{BP}}(x')_{i,j}}{\partial x'_t}$ 即为输入变量的二阶导数，这对于使用 ReLU 激活函数的 CNN 来说无法进行参数更新。因此，若要求攻击基于梯度的可视化方法，需要将目标网络的 ReLU 函数替换为 Softplus 等二阶导数不为 0 的激活函数，这也是文献[14-15]的研究工作。

因此，对基于梯度的可视化方法的攻击方法，例如 BP、Guided BP、Integrated gradient、Smooth gradient 等方法，对于 ReLU 网络并不适用，而 Grad-CAM 攻击方法对于 ReLU 网络却适用，这也是为何本文方法不需要修改目标网络的 ReLU 层，而文献[14-15]则需要将目标网络的 ReLU 函数替换为 Softplus 函数才可行的原因。

5.2 本文方法与现有方法的不同之处

针对 CNN 可解释性方法的攻击，现有工作多数对基于梯度的可视化方法进行攻击，此时需要修改目标模型自身结构[14-15]。而对于 Grad-CAM 可解释性方法的攻击，现有工作中，仅有文献[16]进行了研究，其结果表明 Grad-CAM 解释方法的确是脆弱的，其结果可以被欺骗而产生错误解释。但文献[16]使用对抗性的模型微调，需要在整个数据集上重新训练模型参数，导致训练的代价较大，在现实攻击场景中不太可行。而本文方法使用对抗补丁进行攻击，不需要修改模型结构，能够针对未知图像直接添加补丁并进行攻击，具有一定的泛化性，攻击过程更加简单。

针对对抗补丁攻击方法，文献[17]最早提出将对抗补丁方法用于攻击模型的分类结果，而并未关注 CNN 的可解释性方法，导致这种对抗补丁容易被可解释性方法（如 Grad-CAM）检测到。为此，文献[21]在对抗补丁的形成过程中，引入了针对 Grad-CAM 类激活图的约束，提升了对抗补丁的稳健性，使其不会被轻易检测到。本文的研究思路主要借鉴于文献[21]，但文献[21]使用对抗补丁的主要目的是欺骗分类器的分类结果，而本文将对抗补丁用于攻击针对分类结果的解释，这是本文研究与文献[21]之间的不同之处。

此外，综合第 3.2～3.4 节可知，本文提出的基于对抗补丁的 Grad-CAM 攻击方法适用于以下 3 种攻击场景。1) 特定的单张图像的 Grad-CAM 的攻击，该情形下使用标准的 Grad-CAM 攻击方法即可实现，如第 3.2 节所述；2) 未知图像的 Grad-CAM 的攻击，该情形下使用可泛化的 Grad-CAM 攻击方法，如第 3.3 节所述；3) 对抗样本攻击方法，该情形下可使用第 3.4 节所述的对抗样本生成方法。综上，本文方法可以实现多场景的适用性，与现有方法相比，应用场景更加广泛。

6 结束语

本文提出了一种基于对抗补丁的 Grad-CAM 攻击方法，该方法通过将对抗补丁添加在图像中，可实现对 Grad-CAM 解释结果的有效攻击。与现有的攻击方法相比，本文方法更加简单，且具有较好的泛化性能，并可以拓展为对抗样本的攻击场景，具有多场景的可用性。由于 Grad-CAM 的脆弱性对于注重可解释性和安全性的领域（如医疗图像诊断、自动驾驶等）具有较大的危害，因此，本文的研究进一步指明了这种危险性，为开展与防御攻击相关的研究提供了启发。

由于本文方法仍然要求目标网络为白盒模型，使用目标网络的梯度信息更新扰动补丁，这在实际中不一定能够得到满足。因此，下一步工作将寻找一种基于黑盒优化的对抗补丁方法，使之更加符合现实攻击场景。此外，也将从可解释性防御的角度出发，考虑更加稳健的可解释性方法，用于提供可信赖的解释结果。

参考文献：

[1] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv Preprint, arXiv: 1409.1556v6, 2014.

[2] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//IEEE Conference on Computer Vision and Pat-

tern Recognition. Piscataway: IEEE Press, 2016: 770-778.

[3] HUANG G, LIU Z, MAATEN L V D, et al. Densely connected convolutional networks[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 2261-2269.

[4] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. arXiv Preprint, arXiv: 1706.03762v5, 2017.

[5] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[J]. arXiv Preprint, arXiv:1810.04805, 2018.

[6] SIMONYAN K, VEDALDI A, ZISSERMAN A. Deep inside convolutional networks: visualising image classification models and saliency maps[J]. arXiv Preprint, arXiv:1312.6034, 2013.

[7] SPRINGENBERG J T, DOSOVITSKIY A, BROX T, et al. Striving for simplicity: the all convolutional net[J]. arXiv Preprint, arXiv:1412.6806, 2014.

[8] SMILKOV D, THORAT N, KIM B, et al. SmoothGrad: removing noise by adding noise[J]. arXiv Preprint, arXiv:1706.03825, 2017.

[9] SUNDARARAJAN M, TALY A, YAN Q Q. Axiomatic attribution for deep networks[J]. arXiv Preprint, arXiv: 1703.01365, 2017.

[10] ZHOU B, KHOSLA A, LAPEDRIZA A, et al. Learning deep features for discriminative localization[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 2921-2929.

[11] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 618-626.

[12] 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. Piscataway: IEEE Press, 2018: 839-847.

[13] WANG H F, DU M N, YANG F, et al. Score-CAM: improved visual explanations via score-weighted class activation mapping[J]. arXiv Preprint, arXiv: 1910.01279, 2019.

[14] GHORBANI A, ABID A, ZOU J. Interpretation of neural networks is fragile[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2018: 3681-3688.

[15] DOMBROWSKI A K, ALBER M, ANDERS C, et al. Explanations can be manipulated and geometry is to blame[J]. arXiv Preprint, arXiv: 1906.07983, 2019.

[16] HEO J, JOO S, MOON T. Fooling neural network interpretations via adversarial model manipulation[J]. arXiv Preprint, arXiv: 1902.02041, 2019.

[17] BROWN T B, MANÉ D, ROY A, et al. Adversarial patch[J]. arXiv Preprint, arXiv: 1712.09665v2, 2017.

[18] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[19] FUKUI H, HIRAKAWA T, YAMASHITA T, et al. Attention branch network: learning of attention mechanism for visual explanation[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 10705-10714.

[20] LI K P, WU Z Y, PENG K C, et al. Tell me where to look: guided attention inference network[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 9215-9223.

[21] SUBRAMANYA A, PILLAI V, PIRSIAVASH H. Fooling network interpretation in image classification[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 2020-2029.

[22] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. arXiv Preprint, arXiv: 1312.6199v4, 2013.

[23] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv Preprint, arXiv: 1412.6572v3, 2014.

[24] PASZKE A, GROSS S, CHINTALA S, et al. Automatic differentiation in PyTorch[C]//Advances in Neural Information Processing Systems Workshop. Massachusetts: MIT Press, 2017: 1-4.

[25] DONG Y P, LIAO F Z, PANG T Y, et al. Boosting adversarial attacks with momentum[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 9185-9193.

[作者简介]

司念文（1992− ），男，湖北襄阳人，信息工程大学博士生，主要研究方向为深度学习的安全性与可解释性。

张文林（1982− ），男，湖北黄冈人，博士，信息工程大学副教授、硕士生导师，主要研究方向为语音信号处理、语音识别、机器学习等。

屈丹（1974− ），女，吉林九台人，博士，信息工程大学教授、博士生导师，主要研究方向为语音识别、智能信息处理、机器学习等。

常禾雨（1993− ），女，河南郑州人，信息工程大学博士生，主要研究方向为深度学习与行人检测、行人重识别。

李盛祥（1991− ），男，湖南邵阳人，信息工程大学博士生，主要研究方向为多智能体强化学习。

牛铜（1984− ），男，河南郑州人，信息工程大学副教授，主要研究方向为语音增强、语音识别、深度学习等。