计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200800222

基于特征变换的图像检索对抗防御

徐 行 孙嘉良 汪 政 杨 阳
电子科技大学计算机科学与工程学院 成都 611731

摘 要 对抗攻击在图像分类中较早被研究,目的是产生可以误导神经网络预测的不可察觉的扰动。最近,图像检索中的对抗攻击也被广泛探索,研究结果表明最先进的基于深度神经网络的图像检索模型同样容易受到干扰,从而将不相关的图像返回。文中首次尝试研究无需训练的图像检索模型的对抗防御方法,根据图像基本特征因素对输入图像进行变换,以在预测阶段消除对抗攻击的影响。所提方法探索了4种图像特征变换方案,即调整大小、填充、总方差最小化和图像拼接,这些都是在查询图像被送入检索模型之前对其执行的。文中提出的防御方法具有以下优点:1)不需要微调和增量训练过程;2)仅需极少的额外计算;3)多个方案可以灵活集成。大量实验的结果表明,提出的变换策略在防御现有的针对主流图像检索模型的对抗攻击方面是非常有效的。

关键词:图像检索;对抗防御;深度学习;图像变换;对抗攻击
中图法分类号 TP37

Feature Transformation for Defending Adversarial Attack on Image Retrieval

XU Xing,SUN Jia-liang,WANG Zheng and YANG Yang
School of Computer Science and Engineering,University of Electronic Science and Technology of China,Chengdu 611731,China

Abstract The adversarial attack is firstly studied in image classification to generate imperceptible perturbations that can mislead the prediction of a convolutional neural network. Recently,it has also been extensively explored in image retrieval and shows that the popular image retrieval models are undoubtedly vulnerable to return irrelevant images to the query image with small perturbations. In particular,landmark image retrieval is a research hotspot of image retrieval as an explosive volume of landmark images are uploaded on the Internet by people using various smart devices when taking tours in cities. This paper makes the first trail to investigate the defending approach against adversarial attacks on city landmark image retrieval models without training. Specifically,we propose to perform image feature transformation at inference time to eliminate the adversarial effects based on the basic image features. Our method explores four feature transformation schemes:resize,padding,total variance minimization and image quilting,which are performed on a query image before feeding it to a retrieval model. Our defense method has the following advantages:1) no fine-tuning and incremental training procedure is required,2) very few additional computations and 3) flexible ensembles of multiple schemes. Extensive experiments show that the proposed transformation strategies are advanced at defending the existing adversarial attacks performed on the state-of-the-art city landmark image retrieval models.

Keywords Image retrieval,Adversarial defence,Deep learning,Feature transformation,Adversarial attack

1 引言

对抗攻击最初出现在图像分类系统,它指的是在图像上添加人眼难以察觉的扰动,所产生的对抗样本可以干扰卷积神经网络的结果,造成误分类。最近,对抗攻击也逐渐出现在基于深度网络的图像检索系统中,且图像检索系统也容易受到对抗攻击的影响。

图像检索是计算机视觉领域的长期研究课题,目的是从给定查询的数据集中查找输入图像的相似图像[1]。从基于局部特征的描述符开始,图像检索方法在过去十年取得了显著的进展,目前最有效的检索方法是基于经过微调的深度神经网络[2]。然而,正如在图像分类上首次发现的一样,神经网络容易受到对抗攻击[3],即在图像中添加视觉不可察觉的扰动,这种扰动会误导神经网络,产生错误的预测。

最近,一些图像检索系统领域内的对抗攻击方法[4-6]被陆续提出,它们沿用了对抗攻击在图像分类中的应用,生成了一系列图像。这些图像对于人类肉眼来说具有相同的视觉信息,但是它可以误导神经网络,使得神经网络给出错误的结果。

考虑到深度神经网络的广泛应用,对抗攻击会给这些应用带来不可避免的安全隐患,因此有许多对抗防御的方法被提出。以被广泛研究的图像分类任务为例,这些防御方法可以被粗略地分为两类:第一类是主动防御,即通过对抗训练改变模型的参数,重新训练模型,从而增强模型的鲁棒性;第二

到稿日期:2020-08-30 返修日期:2021-03-05 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
通信作者:徐行(interxuxing@hotmail.com)

类是消极防御，即通过预处理方法改变输入图像。

现存的防御方法大多数都是应用在图像分类系统中的，不同于此，本文针对图像检索系统的对抗攻击进行防御。本文的目标是使用消极防御策略来提高神经网络模型的有效性，并实现以下两个目标：1）移除输入图像中的对抗扰动；2）尽可能多地保留输入图像中的信息，使图像检索任务仍然能顺利完成。换言之，本文的目标是提高系统的鲁棒性，并且尽量不影响系统对正常样本的准确识别。

考虑一张图像，其局部结构的相邻像素中包含着很强的相似度和相关性，因此简单的低阶图像变换可以减少其中的冗余信息，同时保留主要信息。基于此，本文提出了一种防御模型，它由多种不同的图像变换方法组成，具体来说包括随机尺寸缩放、随机填充、总方差最小化（Total Variance Minimization，TVM）和纹理图像分割（image quilting）。本文发现这些图像变换方法对于现有的攻击方法可以起到非常好的防御效果，本文提出的 TMA 攻击方法如图 1 所示，对抗样本首先通过防御模型，经过其中的图像变换方法修改后，将修改后的图像用于检索。

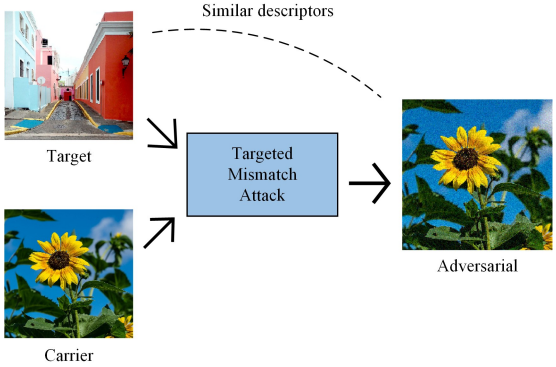


图 1　TMA 攻击方法[5]

Fig. 1　Targeted mismatch adversarial attack method[5]

本文的主要价值可以总结为以下几点：1）本文首次尝试了对图像检索领域中的对抗攻击进行防御，尤其是基于深度学习的图像检索领域；2）本文使用了 4 种基于图像特征变换的防御方法，它们可以根据图像基本特征，有效地破坏对抗样本的特殊结构，而且可以应用于模型的预测阶段，无需额外训练，实验结果证明，这 4 种方法都可以减轻针对检索图像的对抗攻击效果，不同方法的防御能力根据数据集和攻击方式的不同而有所变化。

2　相关工作

2.1　图像检索

图像检索是一项从数据库中读取、查找并检索出相似图像的计算机视觉任务。地标图像检索是图像检索的一个重要组成部分，在图像检索领域的数据集中，地标图像是一个重要组成部分，它指由世界各地的游客、摄影师们拍摄的不同景点的地标图像。

近年来，由于深度学习的流行，针对地标图像的图像检索取得了很大的进展，其中的主流方法可以分为两类。第一类是较为传统的方法，通过词袋模型（bag-of-words）[7]、空间几何信息（spatial verification）[8]、拓展查询（query expansion）[9]和汉明编码嵌入（hamming embedding）[10]等手段得到局部特征。这类方法的实现流程通常是检测局部特征、提取描述器、量化并统一为嵌入编码。在深度神经网络得到迅速发展前，这类方法占据了图像检索领域的绝大部分。

第二类是基于深度神经网络的方法，也是目前实现图像检索最为高效的手段，深度神经网络的优势在于其可以高效地对图像进行语义特征表示。此类方法根据神经网络架构实现方式的不同而不同，具体包括池化层、多尺度放缩、相似度评估函数、模型预训练方式等策略。

深度局部特征（deep local features）[11]方法结合了深度学习和传统的局部特征，它利用地标图像的数据库和图像层级的标注来训练深度神经网络，并应用注意力机制来挑选出图像中的关键点，根据关键点来识别图像中对于图像检索有用的语义信息。根据文献[2]归纳总结的内容，深度局部特征以及它的相关拓展方法是目前地标图像检索效果最好的方法，本文的实验部分主要采用的就是此类基于深度神经网络的检索方法。

2.2　对抗攻击与对抗防御

深度神经网络在人工智能的众多领域取得了空前成功，如图像分类、图像分割、目标检测、语音识别、机器翻译等。然而，最近的研究发现，深度神经网络对于精心设计的输入样本是脆弱的，易受到它们的攻击，这类输入样本被称为对抗样本。对抗样本的特点是，它相对于原始样本的改变，对于人类肉眼来说是难以察觉的，但是却可以使得深度神经网络的预测结果产生错误。因此，围绕对抗样本的研究开始受到关注，即对抗攻击和对抗防御。文献[12]总结了对抗样本出现原因的 3 种观点：流形中的低概率区域解释、线性解释，以及一种认为现有猜想均存在局限性的观点。

对抗攻击首先在图像分类任务中被提出[3]，对抗样本可以通过多种手段生成。白盒攻击假设攻击者知道模型的完整参数，如快速梯度符号法（Fast Gradient Sign Method，FGSM）[13]、DeepFool[14]和 C&W 攻击[15]；黑盒攻击与白盒攻击相反，模型对于攻击者来说是未知的，如基于模型蒸馏的替代模型法[16]和 ZOO[17]；通用对抗扰动指算法生成的同一扰动可以对多张不同图像生效，如 UAP（Universal Adversarial Perturbation）[18]。最近，其他领域开始研究对抗攻击，其中就包括图像检索领域，研究者们分别提出了基于最优化的攻击方法[19]、通用对抗扰动[6]以及有目标对抗攻击[5]。除了上述分类方法，对抗攻击还可以被归纳为 4 类：模型提取攻击、模型逆向攻击、投毒攻击和对抗输入攻击[20]。

由于对抗攻击的出现，与之相对的，提高深度神经网络模型鲁棒性的方法也开始被广泛研究。防御对抗攻击的策略也可以分为两类[21]，积极防御指通过修改模型的结构来提高鲁棒性，而消极防御不修改模型，只在检测阶段尽可能地识别出对抗样本或减轻对抗样本对模型的影响。

消极防御的具体方法有很多种，如图像预处理技术[22]、检测算法[23]以及网络验证[23]，它的优点在于无需额外训练、运算开销小，缺点是无法抵抗较强大的攻击方法，如投影梯度下降法（Projected Gradient Descent，PGD）[24]。投影梯度下降法是基于快速梯度符号法的，通过多次的、更小步的、随机初始

化的攻击得到。积极防御的方法也有很多种，效果最好的是对抗训练(adversarial training)[25]，即在模型训练过程中，通过迭代的攻击方法修改训练数据集，从而使得神经网络对于对抗样本具有鲁棒性。它的缺点是需要重新训练模型、计算成本大、计算速度慢。

此外，还有一种基于生成对抗网络的对抗防御方法[26]，该方法利用对抗样本作为训练样本，同时在网络中加入条件约束来指导 GAN 的训练过程，建立具有鲁棒性的防御模型。

根据已有研究可知，目前暂时没有针对图像检索领域中的对抗攻击的防御方法，由于消极防御方法的实施难度较小，计算成本较低，因此本文决定首先采用消极防御的方法，来验证它们在图像检索领域中的可行性。

3 具体方法描述

3.1 一种图像检索系统中的有目标对抗攻击方法

3.1.1 对抗攻击方法的详细介绍

本节将详细介绍一种针对图像检索的有目标对抗攻击(Targeted Mismatch Adversarial Attack，TMA)[5]方法，后续的防御相关实验也将针对此对抗攻击进行。基于深度神经网络的图像检索系统的工作原理如下：首先，深度神经网络中的全局池化层会将一张图像投影到一个高维的特征空间，产生对应的特征描述向量，然后计算这个特征描述向量与数据库中现有图像的特征描述向量的相似度，并给出一个相似图像的列表。其中，特征描述向量的相似度可以通过计算两个描述向量的内积得到。

TMA 是一种对基于深度学习的图像检索系统的有目标攻击方法，它的整体框架如图 1 所示。假设想要攻击一张待检索的原始图像(图 1 中的 Target)，攻击者需要选定一张载体图像(图 1 中的 Carrier)，对原始图像进行攻击后，期望得到的对抗样本(图 1 中的 Adversarial)具有两个特征：从外观上看，对抗样本与载体图像相似；从检索结果看，对抗样本与原始图像相似。生成符合条件的对抗样本后，就可以对图像检索系统进行攻击，使得图像检索系统返回原始图像所期望的检索结果，同时不会泄漏原始图像携带的内容。

该方法通过设计大量的损失函数，成功地实现了对部分未知系统的攻击，如对于未知的全局池化操作或者未知的输入分辨率，同样可以攻击图像检索系统。它是首个针对图像检索系统的有目标攻击方法，可以生成一个对抗样本，使得该对抗样本与携带图像的差异尽可能小，并且检索结果与原始图像的检索结果尽可能接近。TMA 根据检索模型的不同已知和未知部分，提出了不同的损失函数的实例，并且使用投影梯度下降法作为基础攻击方法。它从两个角度来验证攻击是否成功，第一个是衡量对抗样本和目标图像的描述向量的预先相似度，第二个是利用目标图像进行检索得到的结果。实验结果表明，这种攻击方法可以高效地攻击目前存在的所有基于深度学习的图像检索系统。

3.1.2 该对抗攻击方法的公式化归纳

根据 TMA 中的攻击设置，对抗攻击的目标是生成一张待检索的图像 X_a，使得它可以作为原始图像 X_q 的对抗样本。即希望达到的结果是，在人类的肉眼看来，对抗样本 X_a 不会泄漏任何原始图像 X_q 的信息，但是将两者作为图像检索系统的输入图像时，它们得到的检索结果是一致的。

假设有原始图像 X_q 与载体图像 X_c，并且两者具有同样的分辨率 $W\times H\times 3$，针对图像检索的对抗攻击的目标是生成对抗样本 X_a，它与 X_q 有着很高的特征描述向量相似度，与此同时有着很低的视觉相似度，这一目标可以形式化地归纳为最小化以下损失函数：

$$\min_{X_a}(X_q,X_c;X_a)=l_{tr}(X_a,X_q)+\lambda\|X_a-X_c\|^2 \tag{1}$$

其中，损失项 $l_{tr}(X_a,X_q)$ 量化地代表了对抗样本 X_a 和原始图像 X_q 之间的语义相似度，它可以有多种实例化的表示方法，本文选取其中具有代表性的两种作为剩余部分使用的两种攻击方法。

(1)全局特征描述损失项。给定一个基于深度学习的图像检索模型，最理想的情况是使用它的池化层所输出的特征描述向量来代表损失项 l_{tr}：

$$l_{tr}(X_a,X_q)=l_{desc}(X_q,X_a)=1-h_{X_q}^{\mathrm{T}}h_{X_a} \tag{2}$$

其中，h_{X_q} 和 h_{X_a} 分别代表原始图像 X_q 和对抗样本 X_a 经过图像检索模型编码后的结果。在后续的实验中，本文将采用 5 种最新的基于深度学习的图像检索模型来完成这一过程。

(2)激活直方图损失项。如果考虑图像检索模型中的神经网络的中间层，那么理想情况下，这些内部的卷积层所输出的中间结果对于原始图像 X_q 和对抗样本 X_a 也应该是相同的。受此启发，另一种损失项的约束思路是维持这些中间的卷积层所产生的激活向量的一阶统计量不变，这一约束是卷积核数量粒度的，可以通过最小化两个激活直方图的均方误差来实现。

$$\begin{aligned}l_{tr}(X_a,X_q)&=l_{hist}(X_q,X_a)\\&=\frac{1}{d}\sum_{i=1}^{d}\|u(\boldsymbol{g}_{X_q},b)_i-u(\boldsymbol{g}_{X_a},b)_i\|\end{aligned} \tag{3}$$

其中，$u(\boldsymbol{g}_{X_q},b)_i$ 和 $u(\boldsymbol{g}_{X_a},b)_i$ 分别代表了激活向量 $\boldsymbol{g}_{X_q}$ 和 $\boldsymbol{g}_{X_a}$ 根据第 i 个卷积核计算得到的激活直方图，$\boldsymbol{b}$ 是直方图的分布中心的向量。与全局特征描述损失项相比，激活直方图损失项没有保留住特征描述向量的空间分布，它的计算更为快速。

给定由式(2)和式(3)得到的两种损失项实例化后，可以通过 Adam 算法[27]对式(1)中的损失函数进行优化，并且使用投影梯度下降法来进行边界约束。对抗样本 X_a 通过复制载体图像 X_c 初始化得到，随后每一次算法迭代都会将它的值限制在[0,1]的数据范围内。

3.2 图像检索系统的对抗防御方法的具体设计

3.2.1 对抗防御方法的具体方案

对抗防御的目标是构造一个图像检索模型，使得它对于 3.1.1 节描述的 TMA 攻击所产生的对抗样本 X_a 具有鲁棒性。换言之，该模型能够在原始图像 X_q 被修改为对抗样本 X_a 后，仍然检索出正确的结果。现有的针对图像的对抗攻击可以大致分为两类，一类是数字图像的，此类攻击的特点是生成的对抗样本对于人类肉眼来说是很难察觉的；另一类是物理世界中的，此类攻击添加的扰动往往容易被人类察觉，较小的扰动往往无法成功攻击物理世界中的仪器，物理世界中的

攻击需要通过牺牲隐匿性来实现强度，因为微小的扰动可能无法被复杂的现实环境中的相机、摄像头捕获。考虑到这一点，再结合文献[11]中关于对抗攻击是由于深度神经网络在高维空间中存在线性特征而产生的猜想，通过分析得出可以使用图像变换算法来破坏对抗样本的攻击效果。

为了实现这一目标，本文提出了一种行之有效的对抗防御方法，它基于图像变换算法，可以改变对抗样本 X_a 中的扰动结构，从而破坏对抗样本，使其带来的攻击失效。与此同时，防御模型还需要保证不会影响图像检索模型对于正常样本的准确检索。实验结果表明，本文提出的防御模型可以在提高图像检索模型受对抗攻击时的准确率的同时，基本不影响它对于正常样本的检索准确率。

该防御模型的工作流程如图 2 所示，对抗样本在经过防御模型后，会得到经过修改后的图像，检索模型将对修改后的图像进行检索。它可以看作是一个插入式的组件，只要将其添加到图像检索模型的网络架构的开始部分，就能起到防御的作用。值得注意的是，这个防御模型无需重新训练网络就能使用，因此它可以很方便地作为一个接口来调用。

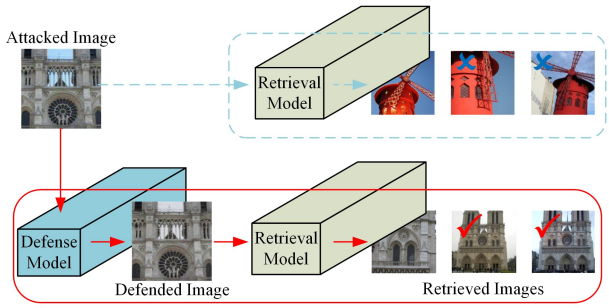


图 2　本文提出的对抗防御方法的整体框架

Fig. 2　Proposed adversarial defence methods

该防御模型包含 4 种图像变换方法：1）随机放缩；2）随机填充；3）总方差最小化；4）纹理图像分割。这些图像变换方法都可以改变对抗样本的结构，从而破坏对抗扰动的效果。在实际应用中，对抗样本经过其中的一种或多种图像变换，然后防御模型会将经过变换的图像输入到图像检索模型中，匹配出具有高特征向量相似度的图像，从而得到最终的检索结果。

3.2.2　相关图像变换方法的介绍

图像变换指将给定的图像通过一定规则以某种形式变换到另一空间中，对抗样本 X_a 经过图像变换后得到的结果称为防御图像 X_d。本节将对防御模型采用的 4 种图像变换方法逐一进行具体的介绍。

（1）随机放缩。假设给定对抗样本 X_a 的尺寸为 $W\times H\times 3$，将对抗样本 X_d 的尺寸随机地放大到 $W'\times H'\times 3$，得到防御图像 X_d，根据经验，此处 $|W'-W|$ 和 $|H'-H|$ 需要在一个较小的合理数值范围内，否则防御图像 X_d 的检索准确率会急剧下降。以 AlexNet[28] 的网络结构为例，原始图像的输入尺寸为 $299\times 299\times 3$，只需要将防御图像 X_d 的边长控制在[299，331]的范围内，就可以保证较高的检索结果准确率。常见的图像放缩插值法有最近邻插值法、线性插值法、双线性插值法、立方插值法、区域插值法等，实验结果表明，选用不同的插值法对于最终的图像检索结果的影响非常小，因此后续实验都将采用最近邻插值法。

（2）随机填充。该方式会对对抗样本 X_a 随机地进行零填充。若要将尺寸为 $W\times H\times 3$ 的对抗样本 X_a 填充成尺寸为 $W''\times H''\times 3$ 的防御图像 X_d，则可以在对抗样本 X_a 的左边填充 w 个零像素点，右边填充 $W''-W-w$ 个零像素点，上面填充 h 个零像素点，下面填充 $H''-H-h$ 个零像素点，因此总共可能有 $(W''-W+1)\times(H''-H+1)$ 种不同的填充方式。

（3）总方差最小化（TVM）。总方差最小化[29] 是一种压缩感知方法，可以有效地去除对抗样本中的对抗扰动。具体而言，首先选取对抗样本 X_a 中的部分像素作为它的一个子集，然后利用该子集来重新构建一个对抗样本 X_a 的“最简单”版本，作为防御图像 X_d，该图像被认为是与先前选中的像素子集相一致的。对抗扰动往往是微小并且只在局部存在的像素点，因此它很可能不会继续存在于防御图像 X_d 中。具体而言，对于每一个像素点的位置 (i,j,k)，首先由伯努利分布进行随机采样，得到一个随机变量 $X_a(i,j,k)$，假如 $X_a(i,j,k)=1$ 就保留该位置的像素点。然而，利用这些被保留下来的像素点重构一个与对抗样本 X_a 相似的防御图像 X_d，与此同时需要保证总方差最小化，可以通过解决以下优化问题来得到：

$$\min_{X_d}\|(1-X_a)\cdot(X_d-X_a)\|_2+\lambda_{TV}\cdot TV_p(X_d) \tag{4}$$

其中，$\cdot$ 代表元素粒度的乘法，$TV_p(X_d)$ 代表防御图像 X_d 的 L_p 范数的总方差，它衡量了图像中的小尺度差异，从而能够移除对抗样本 X_a 中的对抗扰动。式（2）－式（4）中的目标函数对于防御图像 X_d 是凸函数，因此易于求解，根据经验，为了高效地进行总方差最小化计算，通常取 $p=2$。

（4）纹理图像分割。纹理图像分割[30] 是一种非参数化的方法，从包含补丁图像的数据库中取出小的补丁，并将这些补丁拼接到一起来合成目标图像。该算法会将合适的补丁放置在预先定义好的网格点上，并计算出所有重合的边界区域中最小的图割，来解决边界的边缘伪影问题[31]。在补丁图像的选取上，该算法首先根据对抗样本 X_a，利用像素空间中的 K 近邻算法得到候选补丁图像，然后利用均匀分布随机地从中选取一张补丁图像。由于纹理图像分割算法所使用的补丁图像是从事先搭建好的数据库中取出的，它们都是没有受过攻击的原始图像，因此这些补丁图像都不太可能含有对抗样本中容易出现的结构。

4　实验验证

4.1　本文提出的对抗防御方法的实验设计

本节将对实验设计部分进行阐述，具体包括使用的训练数据集与评估数据集、图像检索模型、攻击和防御模型以及评估标准。

本文在实验部分使用的图像检索模型共有 5 种，分别是 MAC[32]，SPoC[33]，GeM[2]，R-MAC 和 CroW[34]，它们都是近年来使用效果较好的基于深度神经网络的特征提取器，这些图像检索模型在实验中均以在 ImageNet 数据集[35] 上预训练过的 AlexNet[28] 作为基础网络结构，然后在经过重构的三维

重建数据集(Structure-of-Motion Reconstruction, SfM)[36]上进行微调训练(Fine-tuned)。SfM 数据集包含了约 740 万张从图片分享网站 Flickr 上下载的图片,并被分为两个大规模的训练数据集,分别命名为 SfM-30k 和 SfM-120k,本文采用 SfM-120k 对防御模型中的深度神经网络进行训练。

对于评估数据集,本文选取的是两个在地标图像检索领域被广泛用于评估的数据集,即经过重构的牛津数据集(Revisited Oxford Dataset, ROxford5k)和经过重构的巴黎数据集 Revisited Paris Dataset, RParis6k)[1]。ROxford5k 数据集由 5063 张英国牛津郡中拍摄的建筑照片和 70 张用于检索的原始图像组成,这些原始图像根据视觉相似度被分为 26 组,每一组均对应了一个人工标注的经过排序的检索列表,代表着该组图像用于检索时的最理想结果。相似地,RParis6k 数据集由 6392 张法国巴黎市拍摄的照片和 70 张用于检索的原始图像组成,这 70 张原始图像被分为 25 组。在进行评估时,本文会对两个数据集中的所有原始图像进行攻击,然后进行检索对抗样本与检索经过防御后的对抗样本的对比实验,从而评估防御模型的效果。

对于攻击模型,正如 3.1 节提到的,本文会选用 l_{desc} 和 l_{hist} 两种攻击方式,它们分别代表了式(2)和式(3)中的两种不同实例化,以此来衡量本文提出的对抗防御方法对不同攻击方式的防御效果。对于 3.2 节中提出的对抗防御方法,实验将分别对 4 种图像变换方法进行单独评估,并对选取其中两种图像变换方法进行组合的防御方式进行评估。具体而言,对抗样本 X_a 首先会通过防御模型生成经过防御后的图像 X_d,然后将 X_d 作为上述图像检索模型的输入,得到一组检索结果,并将这组检索结果与原始图像的检索结果进行对比。考虑到消极防御的一大优点是运行时间短,因此在实验步骤中加入了运行时间的测量部分,所有实验的运行环境都是一台载有 8 核 Intel Xeon 2.5GHZ CPU 的服务器。

在实验的评估阶段,需要将对抗样本 X_a 得到的检索结果质量与经过防御后的图像 X_d 得到的检索结果质量进行对比,为了量化这一过程,本文选取了一种被广泛使用的评估方式:平均精度均值(Mean Average Precision, mAP),它可以反映检索系统对于所有查询请求返回的检索列表的平均准确率。假如经过防御后的图像 X_d 进行检索得到的平均精度均值相对于对抗样本 X_a 的检索结果有所提高,则认为本次防御是有效的,提高得越多,表明防御越成功。

4.2 对抗防御方法的测试以及结果分析

本文在 4.1 节所述的实验条件下进行了测试。首先对重构牛津数据集(ROxford)和重构巴黎数据集(RParis)中的所有原始样本分别进行 l_{desc} 和 l_{hist} 攻击,生成相应的对抗样本,然后将这些对抗样本经过防御模型后得到防御图像,再对防御图像进行检索。

表 1 和表 2 分别列出了本文提出的对抗防御方法在 ROxford 和 RParis 两个数据集上的实验结果,其中都包含了 5 种基于深度学习的图像检索模型对于对抗样本和防御图像的检索表现。

表 1 RParis 数据集上防御 TMA 的 mAP

Table 1 mAP in RParis datasets against TMA

Models+Methods	Atk	Image Transformation				
		Re	Pad	TVM	IQ	Pad+IQ
GeM+l_{desc}	7.59	15.71	13.71	14.06	31.46	**33.89**
MAC+l_{desc}	5.90	8.69	8.22	9.08	**15.40**	15.32
SPoC+l_{desc}	9.86	19.94	17.01	9.58	33.79	**36.37**
R-MAC+l_{desc}	7.84	16.25	15.4	16.35	27.89	**34.98**
CroW+l_{desc}	9.08	20.57	18.29	14.18	32.08	**39.73**
GeM+l_{hist}	7.61	14.25	13.32	14.16	36.02	**38.03**
MAC+l_{hist}	6.82	10.39	9.49	10.39	**19.14**	18.86
SPoC+l_{hist}	8.21	15.53	14.51	11.43	33.87	**37.30**
R-MAC+l_{hist}	7.63	13.48	13.81	14.89	35.18	**38.36**
CroW+l_{hist}	7.87	14.11	14.25	13.48	36.39	**41.45**
平均时间/ms		0.26	10.76	0.57	9.32	53.95

表 2 ROxford 数据集上防御 TMA 的 mAP

Table 2 mAP in ROxford datasets against TMA

Models+Methods	Atk	Image Transformation				
		Re	Pad	TVM	IQ	Pad+IQ
GeM+l_{desc}	25.59	**34.56**	28.91	18.56	26.70	29.70
MAC+l_{desc}	19.10	**22.99**	21.95	14.29	18.35	21.20
SPoC+l_{desc}	31.47	**40.42**	31.39	19.17	31.00	34.12
R-MAC+l_{desc}	23.40	**27.42**	25.61	18.26	21.72	25.98
CroW+l_{desc}	30.52	**40.87**	32.65	20.85	29.98	35.18
GeM+l_{hist}	25.64	**34.80**	29.17	19.68	26.76	30.18
MAC+l_{hist}	20.21	**25.26**	23.30	15.75	19.44	22.85
SPoC+l_{hist}	27.91	**37.73**	29.38	19.93	29.93	32.20
R-MAC+l_{hist}	23.75	**28.08**	26.26	19.21	22.35	25.76
CroW+l_{hist}	27.05	**38.12**	30.16	21.12	28.79	32.51
平均时间/ms		0.23	9.81	0.45	8.71	7.49

表 1、表 2 中,检索结果的量化评估指标都为平均精度均值(mAP),每一行表现最好的数据均用粗体表示,表中的前五行是针对 l_{desc} 攻击的对抗防御,而后五行是针对 l_{hist} 攻击的对抗防御,表中的第三列展示了不同图像检索系统对于对抗样本的检索效果,表中的四至七列则展示了不同图像检索系统对于经过防御后产生的防御图像的检索效果。表 1、表 2 对图像变换算法的名称进行了缩写,放缩代表随机放缩,填充代表随机填充,总方差代表总方差最小化,纹理分割代表图像纹理分割,+号代表按先后顺序集成了两种图像转换方法。

表 1 和表 2 中,Atk 代表对抗攻击,Re 代表随机放缩,Pad 代表随机填充,TVM 代表总方差最小化,IQ 代表纹理图像分割。通过表 1、表 2 分析可以得知,本文提出的对抗防御方法具有以下优点:1)在两个数据集的所有实验中,4 种图像变换方法以及它们的集成方法的 mAP 全部都大幅度超过了对抗样本攻击方法的检索 mAP,这意味着本文提出的对抗防御方法对于不同图像检索模型和不同攻击方法都是行之有效的,是一种成功的基于深度网络的图像检索系统的对抗防御方法;2)图像变换方法的运行时间范围是从零点几毫秒到几十毫秒,即所有对抗防御方法的运行时间都是毫秒级的,这意味着本文提出的对抗防御方法可以非常高效、简单地添加到现有的图像检索模型中,作为一个易用的组件使用。

同时,上述实验结果也表明了本文提出的对抗防御方法存在以下不足:1)在 RParis 数据集上,单个图像变换方法中,图像纹理分割在所有图像检索模型和攻击方式下都取得了最高的 mAP,而在 ROxford 数据集上,单个图像变换方法中,随

机放缩在所有图像检索模型和攻击方式下都取得了最高的mAP，考虑到图像纹理分割方法是一种小图像粒度的图像变换方法，而随机放缩是像素粒度的图像变换方法，这表明不同的图像变换方法防御对抗攻击的能力会随着不同数据集中图像的多样性变化而变化；2）随机填充加图像纹理分割的集成防御方法在RParis数据集的绝大多数情况下取得了最高的mAP，这似乎表明集成可以使得本文提出的对抗防御方法的防御效果更进一步，然而在ROxford上的实验结果却截然相反，每一种集成防御方法的mAP都不如其中单个图像变换方法的最高mAP，这意味着不同的图像变换方法不能简单地进行叠加。

除了TMA以外，另一种较为先进的图像检索领域的攻击方法是通用对抗扰动[6]（UAP），本文针对该攻击模型也做了对应的防御实验，选取的网络结构为AlexNet，攻击方式为List-wise，如表3所列。实验结果表明，本文提出的防御方式对目前针对图像检索的主流对抗攻击模型均有防御效果。从防御前后mAP的提升幅度来看，本文提出的防御方法对TMA的防御效果优于UAP，可能是有目标对抗攻击添加的对抗扰动更具有针对性，因此也更容易被防御方法破坏。

表3　RParis和ROxford数据集上防御UAP的mAP

Table 3　mAP in RParis and ROxford datasets against UAP

Models+Datasets	Atk	Image Transformation				
		Re	Pad	TVM	IQ	Pad+IQ
GeM+RParis	27.42	32.41	30.16	35.56	42.65	**44.81**
MAC+RParis	29.28	33.74	34.27	34.92	44.61	**45.92**
GeM+ROxford	17.12	20.06	21.74	21.22	34.93	**35.49**
MAC+ROxford	16.31	19.85	18.92	20.72	**32.65**	31.67

本文还在实验阶段对对抗样本和防御图像的检索效果进行了可视化展示，图3给出了其中一个例子，原始图像是取自于RParis数据集的巴黎圣母院，原始图像经过l_{desc}的攻击后，对抗样本的检索结果的前五名分别是4张埃菲尔铁塔的图像和1张风车的图像，其均为错误结果，表明对抗攻击已经成功使得基于深度神经网络的图像检索系统误分类。而在经过本文提出的防御模型后，利用所得到的防御图像进行检索，检索列表的前5个结果均为正确的巴黎圣母院的图像，这意味着对抗样本中的对抗扰动已经基本被防御模型消除。ROxford数据集上不同方案所生成的防御图像示例如图4所示。从图中可以看出，这些防御图像有效保持了原始图像的视觉内容。

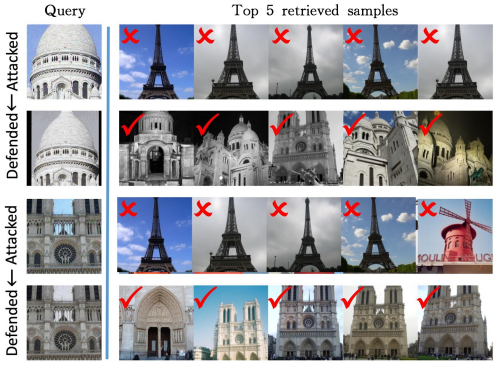


图3　防御效果的可视化展示

Fig. 3　Visualization of our proposed model

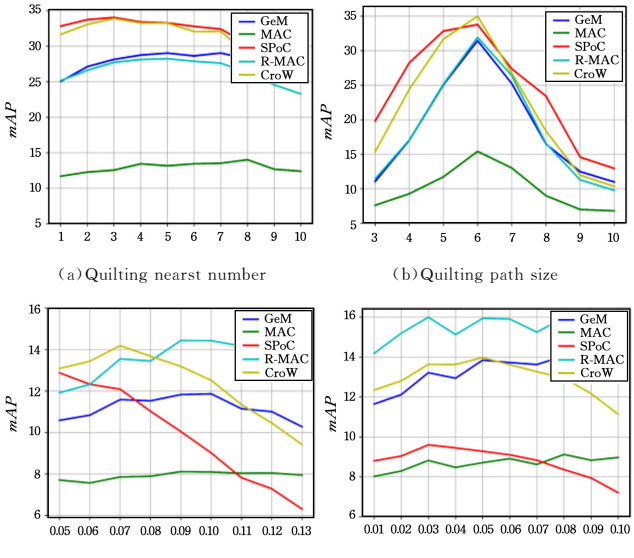
(a) Adversarial image　(b) Padding

(c) Image quilting　(d) Quilting+Resize

图4　ROxford数据集上的防御效果示例图

Fig. 4　Examples of defence in ROxford datasets

上述图像变换方法中，总方差最小化和图像纹理分割都存在着超参数，为了探究超参数的选取对它们在对抗防御方法中表现的影响，还进行了一些额外的对比实验，结果如图5所示。



(a) Quilting nearst number　(b) Quilting path size

(c) TVM pixel drop rate　(d) TVM weight

图5　超参数对防御效果的影响

Fig. 5　Effect of the hypermeters in the defense schemes

从图5中4幅折线图的趋势可以看出，不同超参数的最优解根据图像检索模型的不同而不同，如总方差最小化的像素损失率，在选取SPoC作为图像检索模型时，检索准确率随着像素损失率的下降而下降，但是在选取GeM，R-MAC和CroW作为图像检索模型时，随着像素损失率的下降，检索准确率出现较明显的先上升后下降的情况。分析其原因，这两种图像变换方法都包含了一定的随机性。总方差最小化会随机选取一个像素集，将其作为重构图像的误差衡量标准，而图像纹理分割方法会从符合条件的近邻补丁图像中根据均匀分

布随机选取一张图像作为补丁，这就是它们随机性的来源。

文献[37]提出了一种具有鲁棒性的对抗攻击方法(Expectation over Transformation，EOT)，用于攻击图像分类模型。该方法产生的对抗样本在模糊、旋转、缩放等图像变换下仍能可靠地攻击神经网络。本文将该方法构造的对抗样本用于图像检索任务，实验结果如表4所列。EOT的局限性在于必须了解目标算法的内部细节，并且是一种针对图像分类任务的白盒攻击方法，实验结果表明，在RParis数据集上添加EOT算法生成扰动后，对抗样本进行图像检索得到的结果并没有显著下降，即不能成功进行攻击。这可能是由于EOT算法具有上述局限性，无法直接迁移到图像检索领域中进行对抗攻击。

表4 RParis数据集上EOT算法对图像检索进行攻击的mAP

Table 4 mAP of EOT in RParis datasets

Models	Original	Attack
GeM	41.3	−2.3
MAC	37.0	+0.9
SPoC	32.9	−1.2
R-MAC	44.1	−1.8
CroW	38.2	+0.4

结束语 总体而言，本文提出的对抗防御方法确实是一种可以在基于深度学习的图像检索系统中起到防御效果的方法，但该模型也存在缺点，即暂时没有找到一种在不同数据集上都可以取得较好防御效果的图像变换方法或集成方法。

参 考 文 献

[1] FILIP R, AHMET I, GIORGOS T, et al. Revisiting Oxford and Paris: Large-Scale Image Retrieval Benchmarking [C]// IEEE International Conference on Computer Vision and Pattern Recognition (CVPR). 2018: 5706-5715.

[2] RADENOVIC F, TOLIAS G, AND O C. Fine-Tuning CNN Image Retrieval with No Human Annotation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(7): 1655-1668.

[3] CHRISTIAN S, WOJCIECH Z, ILYA S, et al. Intriguing properties of neural networks [C]// International Conference on Learning Representation. 2014.

[4] LIU Z R, ZHAO Z Y, MARTHA L. Who's Afraid of Adversarial Queries? The Impact of Image Modifications on Content-based Image Retrieval[C]// International Conference on Multimedia Retrieval. 2019: 578-586.

[5] GIORGOS T, FILIP R, ONDREJ C. Targeted Mismatch Adversarial Attack: Query With a Flower to Retrieve the Tower[C]// IEEE/CVF International Conference on Computer Vision. 2019: 5036-5045.

[6] LI J, JI R, LIU H, et al. Universal perturbation attack against image retrieval [C]// Proceedings of the IEEE International Conference on Computer Vision. 2019: 4899-4908.

[7] JOSEF S, ANDREW Z. Video Google: A Text Retrieval Approach to Object Matching in Videos[C]// IEEE International Conference on Computer Vision. 2003: 1470-1477.

[8] JAMES P, ONDREJ C, MICHAEL I, et al. Object retrieval with large vocabularies and fast spatial matching[C]// IEEE International Conference on Computer Vision and Pattern Recognition. 2007: 1533-1540.

[9] ONDREJ C, JAMES P, JOSEF S, et al. Total Recall: Automatic Query Expansion with a Generative Feature Model for Object Retrieval[C]// IEEE International Conference on Computer Vision. 2007: 1-8.

[10] HERVÉ J, MATTHIJS D, CORDELIA S. Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search [C]// European Conference on Computer Vision. 2008: 304-317.

[11] ZHANG S S, ZUO X, LIU J W. The Problem of the Adversarial Examples in Deep Learning[J]. Chinese Journal of Computers, 2019, 42(8): 1886-1904.

[12] IAN G, JONATHON S, CHRISTIAN S. Explaining and Harnessing Adversarial Examples[C]// International Conference on Learning Representations. 2015: 1-12.

[13] HYEONWOO N, ANDRE A, JACK S, et al. Large-Scale Image Retrieval With Attentive Deep Local Features[C]// The IEEE International Conference on Computer Vision (ICCV). 2017: 567-575.

[14] FROSSARD P, MOSSAVI-DEZFOOLI S M, FAWZI A, et al. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks[C]// IEEE Conference on Computer Vision and Pattern Recognition, CVPR. 2016: 2574-2582.

[15] NICHOLAS C, DAVID A, WAGNE R. Towards Evaluating the Robustness of Neural Networks[C]// IEEE Symposium on Security and Privacy. 2017: 1-16.

[16] NICOLAS P, PATRICK D, MCDANIEL P, et al. Practical Black-Box Attacks against Machine Learning[C]// Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security. 2017: 506-519.

[17] CHEN P Y, ZHANG H, YASH S, et al. ZOO: Zeroth Order Optimization Based Black-box Attacks to Deep Neural Networks without Training Substitute Models [C]// Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. 2017: 15-26.

[18] MOOSAVI-DEZFOOLI S M, FAWZI A, FAWII O, et al. Universal Adversarial Perturbations[C]// The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 445-452.

[19] ZHENG Z D, ZHENG L, HU Z L, et al. Open Set Adversarial Examples[OL]. CoRR abs/1809. 02681. https://www. researchgate. net/publication/327570780 _ Open _ Set _ Adversarial _ Examples.

[20] HE Y Z, HU X B, HE J W, et al. Privacy and Security Issues in Machine Learning Systems: A Survey[J]. Journal of Computer Research and Development, 2019, 56(10): 2049-2070.

[21] YUAN X Y, HE P, ZHU Q L, et al. Adversarial Examples: Attacks and Defenses for Deep Learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(9): 2805-2824.

[22] CHUAN G, MAYANK R, MOUSTAPHA C, et al. Countering Adversarial Images using Input Transformations[C]// International Conference on Learning Representations. 2018: 1-12.

[23] JAN H M, TIM G, VOLKER F, et al. On Detecting Adversarial

Perturbations[C]// International Conference on Learning Representations. 2017:1-12.

[24] MADRY A, ALEKSANDAR M, LUDWIG S, et al. Towards Deep Learning Models Resistant to Adversarial Attacks[C]// International Conference on Learning Representations. 2018:1-10.

[25] GUY K, CLARK W, BARRETT C, et al. Towards Proving the Adversarial Robustness of Deep Neural Networks[C]// Proceedings First Workshop on Formal Verification of Autonomous Vehicles. 2017:19-26.

[26] KONG R, CAI J C, HUANG G. Defense to Adversarial Attack with Generative Adversarial Network [J/OL]. Acta Automatica Sinica. https://doi.org/10.16383/j.aas.c200033.

[27] DIEDERIK K, JIMMY B. ADAM: a method for stochastic optimization[C]// International Conference on Learning Representations. 2015:1-10.

[28] ALEX K, ILYA S, GEOFFREY E H. ImageNet Classification with Deep Convolutional Neural Networks[C]// Neural Information Processing Systems(NIPS). 2012:1106-1114.

[29] LEONID R, STANLEY O, EMAD F. Nonlinear total variation based noise removal algorithms[J]. Physica D: Nonlinear Phenomena, 1992, 60(1/2/3/4):259-268.

[30] ALEXEI A, EFRO S, WILLIAM F. Image quilting for texture synthesis and transfer[C]// Special Interest Group on Computer Graphics and Interactive(SIGGRAPH). 2001:341-346.

[31] YURI B, OLGA V, RAMIN Z. Fast approximate energy minimization via graph cuts[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(11):1222-1239.

[32] ALI S R, JOSEPHINE S, ATSUTO M, et al. Visual Instance Retrieval with Deep Convolutional Networks[C]// International Conference on Learning Representations. 2016:1-10.

[33] BABENKO A, LEMPITSKY V. Aggregating Deep Convolutional Features for Image Retrieval [C]// International Conference on Computer Vision. 2015:1246-1254.

[34] YANNIS K, CLAYTON M, SIMON O. Cross-Dimensional Weighting for Aggregated Deep Convolutional Features[C]// European Conference on Computer Vision Workshops. 2016:685-701.

[35] DENG J, WEI D, RICHARD S, et al. Imagenet: a large-scale hierarchical image database[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2009:1573-1580.

[36] SCHONBERGER L, FILIP R, ONDREJ C, et al. From single image query to detailed 3d reconstruction[C]// Computer Vision and Pattern Recognition. 2015:485-492.

[37] ATHALYE A, ENGSTROM L, ILYAS A, et al. Synthesizing robust adversarial examples[C]// International Conference on Machine Learning. PMLR, 2018:284-293.

XU Xing, born in 1988, Ph.D, associate professor, is a member of China Computer Federation. His main research interests include multimedia information processing and security, cross-media analysis and computer vision.