基于生成对抗网络的无目标深度检索哈希攻击算法

黄 霖 吴 亮 高胜严 秦 川
（上海理工大学光电信息与计算机工程学院 上海 200093）

摘 要 近年来深度哈希技术被广泛研究，可应用于大规模图像检索且取得了良好的性能，然而其安全性问题却相对被忽视.为此，本文提出了一种针对深度检索哈希的无目标攻击算法，可用于深度检索哈希的鲁棒性评估和优化设计.在该算法中我们构建了一个用于获得无目标攻击对抗样本的生成对抗网络模型UntargetedGAN.模型训练过程中，首先利用原型网络(PrototypeNet)将图像标签转换为原型网络编码，之后结合原型网络编码、解码器和鉴别器进行联合训练得到期望的UntargetedGAN模型；在测试阶段输入查询图像及其标签即可快速生成对抗样本.实验结果表明，UntargetedGAN生成的对抗样本可有效实现无目标攻击，且与现有的无目标攻击算法相比在攻击性能和对抗样本生成效率方面有显著提升.

关键词 对抗样本；深度哈希；图像检索；生成对抗网络
中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2023.00803

GAN-Based Untargeted Attack on Deep Hashing for Image Retrieval

HUANG Lin WU Liang GAO Sheng-Yan QIN Chuan
(*School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai* 200093)

Abstract In recent years, with the explosive growth of large-scale and high-dimensional multimedia data, approximate nearest neighbor (ANN) search is widely used in search engines for its ability to effectively balance information retrieval quality and computational efficiency. Among all ANN retrieval methods, hashing technique (Hashing) can convert high-dimensional media data into compact binary codes and quickly compute the Hamming distance between hash codes to retrieve similar images, so hash-based retrieval methods have been widely studied in ANNs. Due to the powerful learning capability of deep neural networks (DNNs), deep hashing using DNNs for automatic feature extraction has made good progress and generally outperforms traditional hashing. However, although deep hashing methods have achieved excellent retrieval results, recent studies have shown that DNNs are vulnerable to attacks by adversarial examples. An adversarial example makes DNNs make false predictions by adding a small perturbation to the original image; this perturbation is usually imperceptible, but can make DNNs make false predictions. Meanwhile, deep hashing inevitably inherits the vulnerability of DNNs to adversarial examples, and this imperceptible perturbation also poses a serious security threat to deep hashing-based retrieval systems. That is, when using the attacked image for retrieval, the retrieval system

收稿日期：2022-05-09；在线发布日期：2022-12-10. 本课题得到国家自然科学基金面上项目（No. 62172280）、国家自然科学基金重点项目（No. U20B2051）、上海市自然科学基金项目（No. 21ZR1444600）的资助.**黄 霖**，博士研究生，主要研究领域为多媒体信息安全、AI安全，E-mail：211240059@st.usst.edu.cn. **吴 亮**，硕士研究生，主要研究领域为深度哈希及攻击. **高胜严**，硕士研究生，主要研究领域为信息隐藏与鲁棒水印. **秦 川**（通信作者），博士，教授，博士生导师，中国计算机学会（CCF）高级会员，主要研究领域为多媒体信息安全、AI安全，E-mail：qin@usst.edu.cn.

will return irrelevant or specified other images. Therefore, we propose an untargeted attack algorithm for deep retrieval hashing, which can be used for robustness evaluation and optimization design of deep retrieval hashing. In our algorithm, we construct a generative adversarial network model, named UntargetedGAN, to obtain untargeted attack adversarial examples, which consists of five main modules: PrototypeNet, decoder, generator, discriminator and target model. During the training process, PrototypeNet is first used to convert query image labels into label features and PrototypeNet encoding, where the label features retain the semantic representation information of the query image labels and the PrototypeNet encoding retains the representative semantics as well as the discriminative features of the query image labels. The label features are then up-sampled and spliced into the query image to assist in the generation of the adversarial sample, while the PrototypeNet encoding is used as the hash code of the query image to perform Hamming distance calculation with the hash code generated by the adversarial example. The function of the decoder is to transform the label features containing deep semantic information output from the PrototypeNet into semantic information consistent with the size of the query image through decoding operations to improve the success rate of generating adversarial examples. The generator consists of several convolutional blocks to generate adversarial perturbation. The discriminator is used to determine whether the input image is a query image or an adversarial examples image, and to make the adversarial example generated by the generator indistinguishable from the query image in terms of data distribution during the adversarial training process with the generator, to improve the security of the adversarial examples. The target model refers to the object under attack, i. e., the deep retrieval hash model used to generate the image hash code. Finally, by jointly training these five modules, we can obtain a generative adversarial network model that can generate effective adversarial examples. Experimental results demonstrate that the adversarial examples generated by UntargetedGAN can achieve the untargeted attack effectively, and the attack performance and generation efficiency of adversarial examples are significantly improved compared with some existing untargeted attack algorithms.

Keywords adversarial example; deep hashing; image retrieval; generative adversarial network

1 引 言

近年来，随着大规模和高维多媒体数据的爆炸式增长，近似最近邻搜索(Approximate Nearest Neighbor, ANN)因其能够有效平衡信息检索质量和计算效率而被广泛应用到搜索引擎中.在所有的ANN检索方法中，哈希技术(Hashing)可以将高维媒体数据转换为紧凑的二进制码，并快速计算哈希码之间的汉明距离以搜索相似图像，因此基于哈希的检索方法在ANN中获得广泛的研究.由于深度神经网络[1-2](Deep Neural Networks, DNN)强大的学习能力，采用DNN自动提取特征的深度哈希已取得良好的进展，且其性能普遍优于传统哈希.

然而，尽管深度哈希方法[3-12]已经取得了优异的检索效果，但最近的研究[13]表明，DNN容易受到对抗样本的攻击.对抗样本通过向原始图像中加入微小的扰动，此扰动通常是不可察觉的，但却可以使得DNN做出错误的预测.由于深度检索哈希系统是基于DNN得到，因此深度哈希也不可避免地继承了DNN对于对抗样本的脆弱性，这种不易察觉的扰动同样对基于深度哈希的检索系统构成了严重的安全威胁.也就是说，当使用被攻击后的图像进行检索时，检索系统会返回不相关或者指定的其它图像.因此探究深度哈希攻击算法不仅可发现深度哈希算法的缺陷，还可以推动深度哈希向更加安全和鲁棒的方向发展.

目前针对图像分类网络的对抗样本研究已经取得了较大进展，但针对深度检索哈希对抗样本的研究还相对较少.图像分类网络对抗样本生成方法通常根据梯度和优化方法来修改图像的像素，从而使

得图像分类网络以高置信度输出错误的类别，而对于图像深度检索哈希对抗样本而言并不是简单的使其置信度发生变化，而是需要在修改像素的同时改变其语义使得最终生成的哈希码保留与原始查询图像不一样的语义信息.因此适用于图像分类的攻击方法并不能直接迁移到深度哈希任务中.近几年来，国内外学者在深度哈希对抗样本领域的研究[14-18]可以分为无目标攻击与有目标攻击两类.无目标攻击即生成的对抗样本使深度图像哈希模型检索结果与查询图像语义不相关，现有的代表性无目标攻击算法有哈希对抗生成[14](Hash Adversary Generation，HAG)和无目标深度哈希攻击[15](Non-Targeted Deep Hashing Attack，NDHA)等.有目标攻击即生成的对抗样本使检索结果为指定标签类别，现有代表性的有目标攻击算法有DHTA[16](Dubbed Deep Hashing Targeted Attack，DHTA)和原型监督对抗网络[17-18](Prototype-Supervised Adversarial Network，Pros-GAN)，以及可以实现无目标攻击和有目标攻击的智能深度哈希攻击[19](Smart Deep Hashing Attack，SDHA).虽然当前的深度图像哈希攻击算法可以生成有效的对抗样本，但是依然存在许多不足之处，以无目标攻击为例，HAG和SDHA需要迭代2000轮才能生成一个对抗样本，NDHA采用自适应迭代次数提高了生成效率，但是仍然需要较长时间.除此之外，攻击效果也还有很大的提升空间.

为了解决无目标攻击中对抗样本生成时间过长和攻击效果不够理想等问题，本文提出一种基于生成对抗网络[20](Generative Adversarial Networks，GAN)的无目标攻击算法UntargetedGAN.在训练阶段结合由PrototypeNet生成的原型网络编码、解码器和鉴别器得到期望的模型；在测试阶段将查询图像及其标签输入训练好的UntargetedGAN模型，模型将会输出具有无目标攻击效果的对抗样本.该对抗样本在视觉上与查询图像无法分辨，而深度哈希检索模型将会给出与查询图像语义无关的检索结果.与HAG和NDHA相比，UntargetedGAN在对抗样本生成效率和多标签数据集攻击效果上都更加优异.本文工作的主要贡献如下：

(1) 提出了一种具有无目标攻击效果的生成对抗网络UntargetedGAN，基于GAN网络的生成能力和原型网络的语义保留能力来生成高效的对抗样本.

(2) 大量实验表明，在针对深度检索哈希的攻击算法中，UntargetedGAN生成对抗样本的效率和有效性优于已报道的其它无目标攻击方法.

2　相关工作

2.1　深度神经网络的对抗样本

针对深度神经网络的对抗样本[21-34]生成算法近年来得到了广泛的研究.文献[24]发现DNN容易受到人类视觉系统无法察觉的微小扰动的对抗性攻击而出现分类错误.文献[25]将这种现象描述为“对抗样本”对深度模型的攻击.除有目标攻击和无目标攻击外，还可根据攻击者是否了解被攻击模型的网络参数，将对抗样本攻击分为白盒攻击和黑盒攻击.对抗样本这一概念被提出后，如何寻找最小攻击扰动成为重点需要解决的问题.目前已报道的相关工作主要从梯度、优化和分类平面三个方面开展研究.

(1) 基于优化的对抗样本算法.文献[24]提出L-BFGS算法来优化最小扰动，该算法计算成本比较高，优化过程非常困难.文献[26]提出C&W攻击，利用二分查找法寻找超参数得到最小的对抗扰动，该算法是目前扰动最小的有目标攻击算法之一，但需要耗费大量时间.文献[27]将C&W和投影梯度下降(Projected Gradient Descent，PGD)的攻击思想相结合提出了解耦方向和范数攻击(Decoupled Direction and Norm attack，DDN)，通过梯度下降和动态调整边界寻找最小扰动，在保证扰动最小的情况下提升了攻击效率.

(2) 基于梯度的对抗样本算法.文献[25]从梯度出发，提出了快速梯度符号法(Fast Gradient Sign Method，FGSM)，通过在梯度方向添加微小扰动来使模型分类结果发生错误，可直接快速地生成对抗样本，但是FGSM对于复杂非线性模型攻击的成功率不高.文献[28]进一步提出一种有目标攻击方法迭代快速梯度符号法(Iterative Fast Gradient Sign Method，I-FGSM)，将FGSM进行多次迭代，每次迭代过程中都朝着梯度方向移动，直到最终攻击成功.在优化I-FGSM迁移性方面，文献[29]利用动量迭代的方法提出动量迭代快速梯度符号方法(Momentum Iterative Fast Gradient Sign Method，MI-FGSM)，增强了对抗样本的迁移性.在I-FGSM的优化方面，文献[30]利用ε范围球的思想将每次扰动映射到ε范围内，提出PGD并改善了I-FGSM的攻击效果.

(3) 基于分类平面的对抗样本算法.以FGSM为代表的基于梯度下降的攻击算法虽然可以快速生成使目标模型预测错误的对抗样本,但无法找到最小的扰动.为解决这一问题,文献[31]提出Deepfool无目标攻击算法,即假设存在分类平面,从点到直线的距离公式出发推导出了跨过分类平面的最短路径,即可得到使目标模型分类错误的最小扰动.文献[32]在Deepfool的基础上提出了普遍对抗性扰动(Universal Adversarial Perturbation, UAP),利用UAP可训练得到针对某个数据集的对抗扰动,该数据集中添加了该对抗扰动的图像都会被目标模型错误分类.

总的来说,基于优化的攻击方法通常只能针对特定的模型和图像对产生对抗样本,在模型迁移性方面较差.基于梯度的攻击方法的条件是攻击者必须了解目标模型的梯度信息,一旦目标模型梯度信息获取难度增加则导致不能生成有效的对抗样本.基于分类平面的攻击方法通常只依赖模型的决策信息同时对于目标模型的访问次数巨大,导致其生成效率低.目前对于如何生成具有高鲁棒性、高迁移性、适用于黑盒攻击以及物理世界的对抗样本将成为研究热点.

2.2 深度检索哈希的对抗样本

由于深度检索哈希继承了DNN对于对抗样本的脆弱性,因此关于深度检索哈希对抗样本的研究也越来越受到重视.文献[14]在2018年提出无目标攻击的深度图像哈希对抗样本算法(Hash Adversary Generation, HAG),通过选择性修改像素来保证攻击成功,同时保证图像质量.HAG是目前深度图像哈希对抗样本算法中不可感知性(即对抗样本与查询图像的不可分辨性)优化最好的算法之一.文献[16]在2020年提出了一种有目标攻击的深度图像哈希对抗样本算法(Deep Hashing Targeted Attack, DHTA),通过缩小查询图像与目标图像的汉明距离来实现有目标攻击.DHTA采用哈希码投票机制,使同一类别图像的哈希码经过投票机制得到一个通用哈希码,之后输入到目标函数中进行反向梯度传播最终得到对抗样本.为了提高生成深度哈希对抗样本的效率和性能,文献[15]提出了无目标攻击NDHA,采用锚移动策略和自适应迭代步数机制,提升了对抗样本的生成效率和攻击效果.文献[17]研究了基于生成对抗网络的有目标攻击深度图像哈希对抗样本算法Pros-GAN,并提出使用PrototypeNet来得到原型网络编码以辅助训练.文献[18]在PrototypeNet的基础上提出一种基于梯度的有目标攻击算法,并针对深度图像哈希对抗样本进行对抗训练,提出一种对抗样本的防御方法.文献[19]提出一种可以实现无目标攻击和有目标攻击的SDHA算法,该算法从降低扰动成本和提升对抗样本生成效率的角度出发,通过结合相关图像对于攻击性能的影响以及考虑到不同维度对于汉明距离的贡献,设计一个全新的目标函数来指导对抗样本的生成,但是由于目标函数更加的复杂导致其生成一个对抗样本依然需要迭代2000轮.

目前深度图像检索哈希对抗样本的有目标攻击和无目标攻击在不可感知性方面已经取得了比较好的结果,但是对抗样本的生成效率较低.目前已经提出的深度检索哈希攻击方法生成对抗样本所需时间较长还有很大的提升空间,因此如何在保证对抗样本攻击效果和不可感知性的前提下提升生成效率仍然是一个值得研究的问题.

3 本文方法

3.1 问题定义

设$\mathbf{O}=\{(x_i,y_i)\}_{i=1}^{N}$为一个包含$C$个类别的多标签数据集,其中$N$为图像标签数,$x_i$表示以第$i$个图像标签来表示图像类别的原始图像,$\boldsymbol{y}_i=[y_{i1},y_{i2},\cdots,y_{iC}]$为一个多标签向量,多标签数据集中一张图像往往有多个标签,$y_{ij}=1$表示x_i属于类别j,反之$y_{ij}=0$则表示x_i不属于类别j.

无论是深度哈希方法还是传统哈希方法都希望相似的图像生成的哈希码也是相似的,以便于后续利用哈希码进行高效的相似检索.对于一个给定的目标哈希模型$\mathrm{F}(\cdot)$,我们定义一张图像样本经过该哈希模型之后产生的哈希码为

$$\begin{aligned} b_i&=\mathrm{F}(x_i)=\mathrm{sign}\left(f_\theta(x_i)\right)\\ &\text{s.t. } b_i\in\{-1,1\}^K \end{aligned} \tag{1}$$

其中$f_\theta(\cdot)$表示一个包含了特征提取器和全连接层的DNN模型,θ为该模型的参数,$\mathrm{sign}(\cdot)$将$f_\theta(\cdot)$得到的特征值压缩到[-1, 1]内从而得到二进制哈希码,K为哈希码的长度.这里用$\boldsymbol{B}=(b_1,b_2,\cdots,b_N)_{K\times N}$表示数据集$\mathbf{O}$的哈希码.由于sign函数存在梯度消失问题,在深度哈希中也常采用$\tanh(\cdot)$近似代替$\mathrm{sign}(\cdot)$.

对于无目标攻击,给定一张查询图像x,我们希望生成的对抗样本x_q^*可以使目标模型检索结果与查询图像不相关,同时对抗扰动应该足够小以保证

不可感知性，即：

$$\begin{aligned}&\max d\left(\mathrm{F}\left(x_{\mathrm{q}}^{*}\right),\mathrm{F}(x)\right)\\&\text{s.t.}\left\|x-x_{\mathrm{q}}^{*}\right\|_{p}\leqslant\varepsilon\end{aligned}\tag{2}$$

其中$d(\cdot,\cdot)$表示汉明距离，$\|\cdot\|_p$表示向量范数，ε表示对抗扰动的最大阈值，通常用来控制对抗样本的图像质量.

3.2　算法流程

本文提出的深度检索哈希无目标攻击算法UntargetedGAN模型的总体框架如图1所示.该模型主要由五个部分组成：生成器，鉴别器，解码器，原型网络和目标哈希模型.其中，生成器和鉴别器是UntargetedGAN的核心组成部分，通过两者之间的对抗训练最终生成高质量的对抗样本.

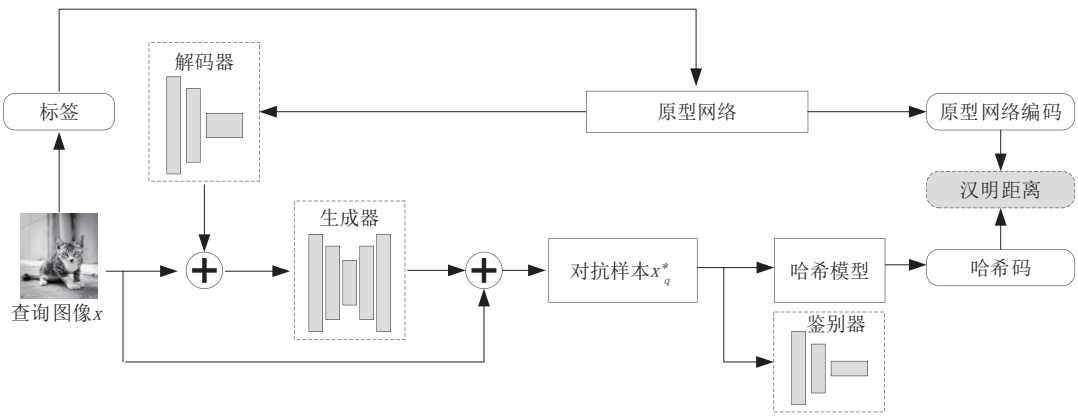

图1　UntargetedGAN模型框架图

（1）原型网络.原型网络在训练过程中将查询图像的标签嵌入到语义表示中，并输出标签特征和原型网络编码，其中标签特征保留了查询图像标签的语义表示信息，原型网络编码保留了查询图像标签的代表性语义以及鉴别特征.然后将标签特征经过上采样后拼接到查询图像中以辅助对抗样本的生成，而原型网络编码将作为查询图像的哈希码与对抗样本生成的哈希码进行汉明距离计算.相比于其他深度哈希生成网络，原型网络的输出包含了更丰富的语义信息，因此能够有效改善我们提出方法对抗样本的生成效率以及迁移性.在UntargetedGAN训练过程中，原型网络不再更新.

（2）解码器.解码器的功能是将原型网络输出的包含深层语义信息的标签特征通过解码操作将其转化为与查询图像大小一致的语义信息，以提升生成对抗样本攻击的成功率.解码器主要通过多次上采样操作将维度为1的标签特征采样为与查询图像大小一致，其具体网络结构如表1所示.解码器的输入为1×512，经过一系列卷积操作最终输出大小为3×224×224的标签特征图像.

（3）生成器.生成器是一个生成对抗扰动的模块，其网络结构如表2所示.生成器的输入由解码器的输出和训练集样本拼接形成，尺寸为6×224×224.生成器主要由多个卷积层构成，如Conv 6-64，7×7表示输入为6个通道，输出64个通道，使用7×7的卷积核进行卷积；Residual Block×6表示经过6个连续的残差模块；最终经过一系列操作可以

表1　解码器的网络结构

层类型	卷积核	输入-输出
Input	/	1×512
FC	/	512-25088
Resize	/	128×14×14
Conv+ReLU	4×4	128-64
Conv+ReLU	4×4	64-32
Conv+ReLU	4×4	32-16
Conv+ReLU	4×4	16-8
Conv	3×3	8×3
Output	/	3×224×224

表2　生成器的网络结构

层类型	卷积核	输入-输出
Input	/	6×224×224
Conv+ReLU	7×7	6-64
Conv+ReLU	4×4	64-128
Conv+ReLU	4×4	128-256
Residual Block×6		
Conv+ReLU	3×3	256-256
Conv	3×3	256-256
Conv+ReLU	4×4	256-128
Conv+ReLU	4×4	128-64
Conv+Tanh()	3×3	67-3
Output	/	3×224×224

得到3×224×224的对抗扰动.

(4) 鉴别器. 鉴别器用于判断输入图像是查询图像还是对抗样本图像,在与生成器的对抗训练过程中,使得生成器生成的对抗样本在数据分布上与查询图像不可分,从而提升对抗样本的安全性,其中鉴别器的输出为 $\hat{\boldsymbol{y}}=[y_1,y_2,\cdots,y_C,1]$ 或 $\hat{\boldsymbol{y}}=[y_1,y_2,\cdots,y_C,0]$. $\hat{\boldsymbol{y}}$ 的前 C 位表示鉴别器预测的当前样本的多向量标签,最后一位表示鉴别器的判定结果,1表示查询图像,0表示对抗样本图像. 鉴别器的网络结构如表3所示,其输入是3×224×224的图像,经过一系列卷积操作后最终得到的输出是 $C+1$. 表3中的 C 表示数据集类别数,鉴别器的激活函数采用LeakyReLU.

表3 鉴别器的网络结构

层类型	卷积核	输入-输出
Input	/	3×224×224
Conv+LeakyReLU	7×7	3-64
Conv+LeakyReLU	4×4	64-128
Conv+LeakyReLU	4×4	128-256
Conv+LeakyReLU	4×4	256-512
Conv+LeakyReLU	4×4	512-1024
Conv	22×22	1024
Output	/	$C+1$

(5) 目标模型. 目标模型指被攻击的对象,即用于生成图像哈希码的深度检索哈希模型. 我们通过计算对抗样本输入到目标模型生成的哈希码与查询图像的原型网络编码之间的汉明距离,来表示无目标攻击的效果. 在UntargetedGAN的训练过程中,目标哈希模型参数不再更新.

在训练过程中,首先将训练集的标签输入到预训练原型网络中,并输出原型网络编码和标签特征值;标签特征值被输入到解码器中进行上采样,解码器输出结果与训练集样本进行拼接输入到生成器中. 其次,生成器输出一个扰动,将扰动与训练集样本相加得到对抗样本. 然后,将对抗样本分别输入到目标模型和鉴别器中,其中鉴别器与生成器进行对抗训练来优化对抗样本的视觉质量,目标模型输出的哈希码与原型网络输出的原型网络编码进行汉明距离计算. 通过多次迭代训练最终实现无目标攻击.

3.3 目标函数设计

UntargetedGAN的目标是生成一个高效的对抗样本,该对抗样本需具备三个特性:(1)能够使得目标模型输出与查询图像无关的哈希码;(2)对抗样本与查询图像在视觉上不可区分;(3)能够使得鉴别器无法分辨查询图像和对抗样本. 为此,我们设计了如下的目标函数来指导整个UntargetedGAN网络的优化训练:

$$L=\alpha L_{\mathrm{T}}+\beta L_{\mathrm{q}}+\gamma L_{\mathrm{D}} \tag{3}$$

为了使生成的对抗样本能够成功地攻击目标模型,我们采用汉明距离来衡量对抗样本对应的哈希码和查询图像对应的哈希码之间的距离,并定义如下损失函数:

$$L_{\mathrm{T}}=\frac{1}{K}\left(h_{\mathrm{q}}\right)^{\mathrm{T}}\cdot h_{\mathrm{adv}} \tag{4}$$

其中 h_{q} 为查询图像的原型网络编码,$(\cdot)^{\mathrm{T}}$ 表示转置,h_{adv} 为对抗样本的哈希码,K 为哈希码长度. 通过优化训练使 L_{T} 尽可能小,可以实现增大汉明距离的目的.

为了保证对抗样本的视觉质量,我们引入一个 L_{q} 损失函数来衡量对抗样本与原始查询图像的相似性,通过优化该损失函数来提升对抗样本的视觉质量,该损失函数定义如下:

$$L_{\mathrm{q}}=\left\| x-x_{\mathrm{q}}^{*}\right\|_2^2 \tag{5}$$

其中 x 表示查询图像,x_{q}^{*} 表示生成的对抗样本.

为了进一步保证对抗样本的安全性(即不被鉴别器辨别),故引入了鉴别器来检测对抗样本. 这里采用 L_{D} 损失函数作为鉴别器的优化目标,损失函数设计如下:

$$L_{\mathrm{D}}=\left\| \mathrm{D}(x_{\mathrm{q}}^{*})-\hat{y}_{\mathrm{T}}\right\|_2^2 \tag{6}$$

其中 $\mathrm{D}(x_{\mathrm{q}}^{*})$ 为鉴别器的输出,$\hat{y}_{\mathrm{T}}$ 为 x_{q}^{*} 对应查询图像的真实标签. 该损失函数通过 L_{D} 范数来缩小对抗样本生成标签与查询图像真实标签之间的距离,从而使得生成的对抗样本在数据分布上更加接近查询图像,最终达到提升对抗样本安全性的目的.

4 实验结果与分析比较

在实验中,我们选择深度图像检索哈希模型HashNet-ResNet50[11]作为主要目标攻击模型. 此外,为了证明所提出的UntargetedGAN算法对其它深度检索哈希模型的攻击也具有有效性,DPSH-VGG11[33]模型也被用来作为攻击对象进行实验.

4.1 攻击有效性指标与实验数据集

通常深度图像检索哈希对抗样本算法的攻击有效性评价指标有三个:(1)MAP(Mean Average Precision, MAP);(2)TopN精度曲线;(3)PR(Precision-Recall,PR)曲线. MAP是信息检索中应用最广泛的评价标准之一,它被定义为平均精度的

平均值，可以客观地反映检索系统的性能. MAP越高，检索系统的性能越好. 由于本文研究的是利用对抗样本进行检索哈希的攻击，故当MAP越小表示生成的对抗样本性能越好. TopN精度曲线表示输入查询图像后检索系统检索到的前N张图像的平均精度，实验中N设置为1000. 平均精度越低，表示攻击算法越有效. PR曲线表示准确率和召回率之间的关系，在评价模型检索性能时，P坐标轴与R坐标轴所围面积越大表示检索性能越好；相反地，在评估攻击算法性能时，面积越小则攻击算法越有效.

为了训练UntargetedGAN模型，我们在实验中使用了三个常见的数据集MS-COCO[34]，NUSWIDE[35]和ImageNet[36]. 下面简单介绍这三个数据集在实验过程中的设置. 如表4所示，表中第一个数据表示模型在训练时所使用图像的数量，第二个数据表示模型训练完之后测试时所采用图像的数量. 以MS-COCO数据集为例，其余数据集类似，UntargetedGAN在训练过程中，首先需要将HashNet和原型网络训练好，从数据集中随机选取5000张图像作为HashNet和原型网络的测试集，再从剩余的图像中随机选取10 000张图像作为HashNet和原型网络的训练集；然后在UntargetedGAN训练过程中将训练好的HashNet和原型网络参数冻结；最后使用与HashNet和原型网络一样的训练集训练UntargetedGAN，随后随机选取1000张图像进行性能测试.

表4　模型训练时图像数量分配情况

数据集	HashNet	原型网络	UntargetedGAN
MS-COCO	10 000; 5000	10 000; 5000	10 000; 1000
NUSWIDE	10 000; 5000	10 500; 2100	10 000; 1000
ImageNet	13 000; 5000	13 000; 5000	13 000; 5000

4.2　实验设置

实验中，我们在MS-COCO、NUSWIDE和ImageNet数据集上以ResNet50为基础网络，各训练HashNet的48-bits、64-bits和128-bits共9个模型. HashNet模型的输入为224×224的彩色图像. 在训练UntargetedGAN之前需要设定一些超参数，实验中令公式(3)中α，β，γ的取值都为1，对抗扰动的最大阈值设置为0.032，使用Adam[37]优化器来优化训练，设定学习率为10^{-4}. 每个UntargetedGAN训练100 epoch，每次训练的batch size设为24，在占用9 GB显存和上述参数的情况下，UntargetedGAN模型训练到收敛所需要的时间约为10小时.

所有实验均在配备NVIDIA RTX 2080ti GPU、3.70 GHz i9-10900X CPU和32.00 GB内存的计算机上进行，所有代码都是基于Pytorch深度学习框架编写.

4.3　攻击性能分析与比较

MAP可以比较客观地反映深度图像哈希模型的检索效果和对抗样本对目标模型的攻击效果，因此本节先讨论本文UntargetedGAN攻击算法的MAP指标，之后再讨论TopN精度曲线和PR曲线. 表5展示了HashNet-ResNet50哈希模型在MS-COCO、NUSWIDE和ImageNet数据集以及不同哈希长度的情况下，经过UntargetedGAN攻击后的MAP数据. 从表5中可以看出UntargetedGAN对哈希模型的攻击效果明显. 以64-bits的HashNet模型为例，其在MS-COCO、NUSWIDE和ImageNet数据集上训练的模型原MAP分别达到了0.8089、0.8560和0.6709，对于检索哈希模型来说是比较高的MAP值；但其在被UntargetedGAN攻击后，MAP显著下降，MS-COCO的64-bits模型MAP下降到了0.0877，下降幅度达89.1%；同样地，NUSWIDE的64-bits模型MAP下降幅度达88.6%；除了多标签数据集外，以ImageNet为代表的单标签数据集的MAP在攻击后同样下降显著，其64-bits模型攻击后MAP下降到0.0371.

为了与其它代表性的无目标攻击方法进行客观比较，表5还列出了在同等条件下两种无目标攻击方法HAG和NDHA的MAP值. 从表中可以看出，UntargetedGAN的攻击效果大部分优于HAG和NDHA，以NUSWIDE的128-bits模型为例，使用HAG攻击后HashNet的MAP下降到0.2223，

表5　针对HashNet模型攻击后的MAP

Methods	MS-COCO			NUSWIDE			ImageNet		
	48-bits	64-bits	128-bits	48-bits	64-bits	128-bits	48-bits	64-bits	128-bits
HashNet-ResNet50[11]	0.8042	0.8089	0.8187	0.8533	0.8560	0.8629	0.6196	0.6709	0.7041
HAG[14]	0.3070	0.3183	0.3430	0.3042	0.2969	0.2223	0.0511	0.0696	0.0903
NDHA[15]	0.1453	0.1516	0.1294	0.1510	0.1365	0.1292	0.0211	0.0200	0.0209
UntartgetedGAN	0.0704	0.0877	0.0945	0.1186	0.0969	0.0961	0.0201	0.0371	0.0295

使用NDHA攻击后下降到0.1292，而本文UntargetedGAN攻击算法对应的MAP为0.0961，相比HAG和NDHA性能更优.在表5中可以发现UntargetedGAN的攻击效果在大部分情况下最优，在ImageNet的64-bits模型中NDHA的效果最好.

为了证明本文攻击算法对于其它深度检索哈希模型同样有效，我们对DPSH进行与HashNet类似的无目标攻击实验，但是在实验过程中发现DPSH在ImageNet数据集中其本身检索性能不佳，因此我们仅在MS-COCO和NUSWIDE数据集上以VGG11为基础网络训练了DPSH的32-bits、48-bits和64-bits共6个模型，并在这6种模型上测试了UntargetedGAN的攻击效果.如表6所示，本文攻击算法在DPSH上同样有效，以NUSWIDE的64-bits模型为例，UntargetedGAN攻击使DPSH模型的MAP从原来的0.8318降到0.0934，下降幅度达88.7%，且对其它哈希长度的模型攻击效果类似，说明了本文攻击算法的有效性.除了MAP外，实验还采用了TopN精度曲线和PR曲线来评估攻击性能，其中图2和图3中的UGAN表示本文提出的UntargetedGAN方法.图2展示了UntargetedGAN、NDHA和HAG三个攻击算法在HashNet-ResNet50上的TopN精度曲线.可以发现，在MS-COCO和NUSWIDE的TopN精度曲线中，UntargetedGAN的查询图像的1000个相关图像的精度明显比NDHA和HAG要低；而在ImageNet的TopN精度曲线中，UntargetedGAN的曲线与NDHA的曲线接近，但优于HAG.图3给出了UntargetedGAN、NDHA和HAG三个攻击算法在HashNet-ResNet50上的PR曲线，可以看出，在数据集NUSWIDE和ImageNet上，本文攻击算法的性能都好于或者接近NDHA.

表6 针对DPSH模型攻击后的MAP

Methods	MS-COCO			NUSWIDE		
	32-bits	64-bits	128-bits	32-bits	48-bits	64-bits
DPSH-VGG11[33]	0.7079	0.7581	0.7259	0.8279	0.8364	0.8318
UntargetedGAN	0.0876	0.1119	0.0825	0.1213	0.1020	0.0934

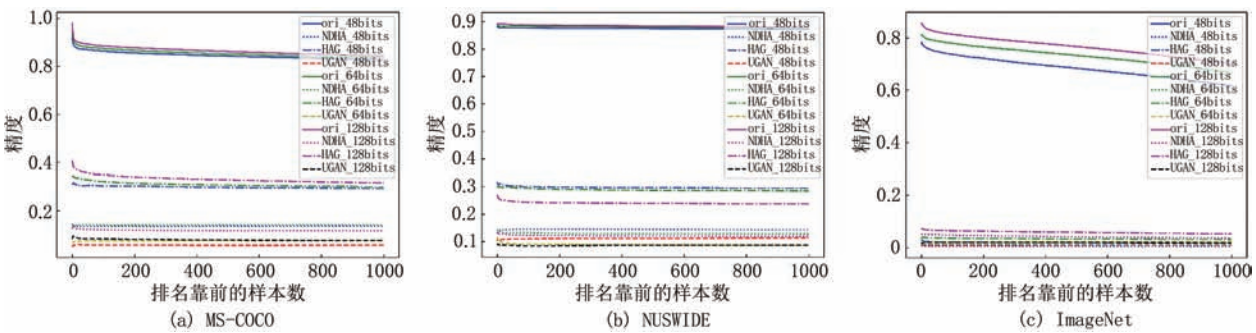

图2 TopN精度曲线性能与比较

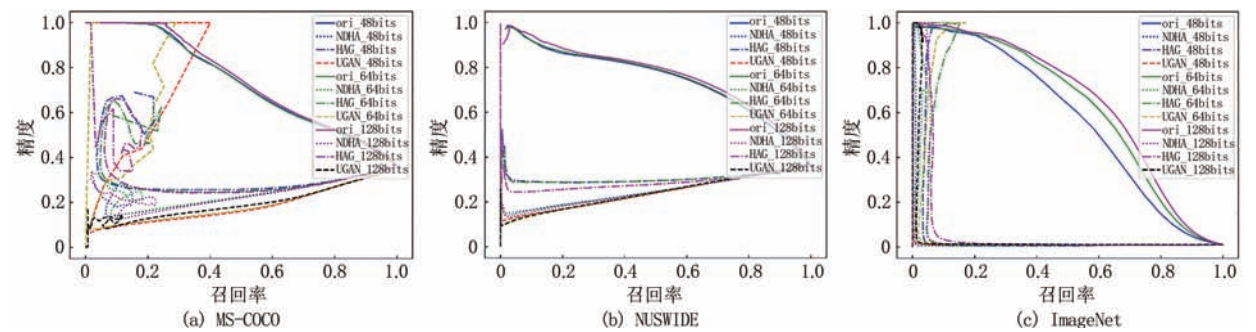

图3 PR精度曲线性能与比较

综上，根据MAP、TopN精度曲线和PR曲线的结果可以表明，本文UntargetedGAN攻击算法对于深度图像检索哈希确实是有效的，且在大部分情况下攻击性能要优于NDHA和HAG.

4.4 不可感知性与运行效率

评价一个对抗样本算法除了其攻击有效性外，不可感知性和运行效率也是非常重要的评价指标.理想情况的对抗样本生成应该是快速且有效的，同时保持良好的不可感知性.不可感知性即对抗样本与对应查询图像的不可分辨性，其具体计算方法为：首先将图像进行归一化，使图像的像素值控制在[0，1]之内，然后按照如下公式进行计算[38]：

$$\sqrt{\frac{1}{Z}\left\| x_q^* - x \right\|_2^2} \tag{7}$$

其中Z为图像的像素总数，x_q^*为对抗样本，x为查询图像. 该数值越低说明不可感知性越好，即对抗样本的视觉质量越理想. 另一方面，实验中我们通过统计生成一张对抗样本图像平均所耗费的时间来评估算法运行效率.

图4以MS-COCO数据集为例，展示了UntargetedGAN攻击生成的对抗样本和相应的检索结果. 图中第1行第1列的图像为查询图像，其一行后面10张图像为HashNet检索模型针对该查询图像得到的检索结果；第2行第1列的图像为UntargetedGAN生成的对抗样本，其同一行后面10张图像为HashNet检索模型使用对抗样本得到的检索结果. 可以发现，UntargetedGAN攻击确实使得检索模型的检索结果与查询图像都是不相关的，且对抗样本图像与原始查询图像在视觉上基本无法区分，即具有较好的不可感知性.

图4　UntargetedGAN的攻击结果展示

实验中，我们还对无目标攻击方法UntargetedGAN、HAG[14]、NDHA[15]和有目标攻击方法DHTA[16]四个算法在MS-COCO数据集上分别得到不可感知性数据和平均运行时间用于比较，如表7所示. 从表中可发现，本文提出的UntargetedGAN攻击算法的运行效率是最高的，平均每张对抗样本图像生成只需要花费5.28 s. 在不可感知性方面，UntargetedGAN略低于其他三种算法，因为本文的不可感知性主要由对抗扰动的最大阈值ε控制，而在实验过程中ε=0.032，相当于将像素的变化范围控制在[-8，8]内，而其他三种算法都是基于传统梯度生成的方法来构造对抗样本，其根据梯度修改的方法对每一张查询图像进行多次修改以达最优的不可感知性. 但是我们的方法在攻击有效性、不可感知性和生成效率方面实现了较好的平衡.

表7　对抗样本的不可感知性和生成效率

算法	UntargetedGAN	HAG[14]	NDHA[15]	DHTA[16]
不可感知性 ($\times 10^{-2}$)	2.7108	0.5925	1.1285	0.8936
时间 (s)	5.28	115.36	66.75	113.78

5　结束语

本文提出了一种基于生成对抗网络的无目标攻击算法UntargetedGAN. 通过对PrototypeNet、解码器、生成器和鉴别器的联合训练，获得了一种针对深度图像哈希模型的无目标攻击GAN模型，其中原型网络用于得到查询图像标签的哈希码和特征值，解码器用于将查询图像标签的特征值上采样，生成器用于生成对抗样本，鉴别器用于鉴别对抗样本与查询图像. 将具有多种不同优化功能的损失函数组合得到UntargetedGAN的目标函数，最终通过优化目标函数来生成期望的对抗样本. 实验结果表明，本文提出的UntargetedGAN相比目前已报道的其它深度图像检索哈希攻击算法更加有效. 本文的研究为增强图像检索哈希模型的鲁棒性提供了启发.

我们提出的方法在使用过程中需要同时输入查询图像和查询图像的标签，而在实际使用过程中查询图像的标签不易获得，同时在训练耗时和模型迁移性方面依然存在提升空间. 目前已提出的方法在模型迁移性方面存在不足，这是由于生成对抗样本的过程中迭代次数过多从而造成过拟合现象，过拟合的现象将导致模型迁移性差，而对于改善模型迁移性的一个思路是将多个具有类似功能的模型作为目标模型进行同时训练，通过多个目标模型学习到此类模型的“分类面”，从而得到一个具有良好迁移性的对抗样本. 此外，现有的攻击方法大多属于白盒攻击，如何实现黑盒攻击也是一个具有挑战性的方向.

致 谢 在此,我们向对论文提出宝贵意见的审稿专家们表示衷心的感谢!

参 考 文 献

[1] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift//Proceedings of the International Conference on Machine Learning. Lille, France, 2015, 37: 448-456

[2] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. nature, 2015, 518(7540): 529-533

[3] Kulis B, Grauman K. Kernelized locality-sensitive hashing. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 34(6): 1092-1104

[4] Gong Y, Lazebnik S, Gordo A, et al. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 35(12): 2916-2929

[5] He K, Wen F, Sun J. K-means hashing: An affinity-preserving quantization method for learning binary compact codes// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Oregon, USA, 2013: 2938-2945

[6] Liu W, Wang J, Ji R, et al. Supervised hashing with kernels// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Rhode Island, USA, 2012: 2074-2081

[7] Norouzi M, Fleet D J. Minimal loss hashing for compact binary codes// Proceedings of the 28th International Conference on Machine Learning. Washington, USA, 2011: 353-360

[8] Shen F, Shen C, Liu W, et al. Supervised discrete hashing// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 37-45

[9] Lin K, Lu J, Chen C S, et al. Learning compact binary descriptors with unsupervised deep neural networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 1183-1192

[10] Dizaji K G, Zheng F, Sadoughi N, et al. Unsupervised deep generative adversarial hashing network//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Utah, USA, 2018: 3664-3673

[11] Cao Z, Long M, Wang J, et al. Hashnet: Deep learning to hash by continuation//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 5608-5617

[12] Cao Y, Long M, Liu B, et al. Deep cauchy hashing for hamming space retrieval//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Utah, USA, 2018: 1229-1237

[13] Nguyen A, Yosinski J, Clune J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 427-436

[14] Yang E, Liu T, Deng C, et al. Adversarial examples for hamming space search. IEEE Transactions on Cybernetics, 2018, 50(4): 1473-1484

[15] Qin C, Wu L, Zhang X, et al. Efficient non-targeted attack for deep hashing based image retrieval. IEEE Signal Processing Letters, 2021, 28: 1893-1897

[16] Bai J, Chen B, Li Y, et al. Targeted attack for deep hashing based retrieval//Proceedings of the European Conference on Computer Vision. Springer, Cham, 2020: 618-634

[17] Wang X, Zhang Z, Wu B, et al. Prototype-supervised adversarial network for targeted attack of deep hashing// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. On-line, 2021: 16357-16366

[18] Wang X, Zhang Z, Lu G, et al. Targeted Attack and Defense for Deep Hashing//Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. Washington, USA, 2021: 2298-2302

[19] Lu J, Chen M, Sun Y, et al. A smart adversarial attack on deep hashing based image retrieval//Proceedings of the International Conference on Multimedia Retrieval. Derby, UK, 2021: 227-235

[20] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. Advances in Neural Information Processing Systems, 2014, 27

[21] Ji S L, Du T Y, Deng S G, Cheng P, Shi J, Yang M, Li B. Robustness certification research on deep learning models: A survey. Chinese Journal of Computers, 2022, 45(01): 190-206 (in Chinese)
(纪守领,杜天宇,邓水光,程鹏,时杰,杨珉,李博.深度学习模型鲁棒性研究综述.计算机学报,2022,45(01):190-206)

[22] Ji T T, Fang B X, Cui X, Wang Z R, Gan R L, Han Y, Yu W Q. Research on deep learning-Powered malware attack and defense techniques. Chinese Journal of Computers, 2021, 44(04): 669-695 (in Chinese)
(冀甜甜,方滨兴,崔翔,王忠儒,甘蕊灵,韩宇,余伟强.深度学习赋能的恶意代码攻防研究进展.计算机学报,2021,44(04):669-695)

[23] Zhang S S, Zuo X, Liu J W. The problem of the adversarial examples in deep learning. Chinese Journal of Computers, 2019, 42(8): 1886-1904 (in Chinese)
(张思思,左信,刘建伟.深度学习中的对抗样本问题.计算机学报, 2019, 42(8): 1886-1904)

[24] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199, 2013

[25] Goodfellow I, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014.

[26] Carlini N, Wagner D. Towards evaluating the robustness of neural networks//Proceedings of the IEEE Symposium on Security and Privacy. California, USA, 2017: 39-57

[27] Rony J, Hafemann L G, Oliveira L S, et al. Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. California, USA, 2019: 4322-4330

[28] Kurakin A, Goodfellow I, Bengio S. Adversarial examples in the physical world//Proceedings of the International Conference

on Learning Representations. Toulon, France, 2017
[29] Dong Y, Liao F, Pang T, et al. Boosting adversarial attacks with momentum//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Utah, USA, 2018: 9185-9193
[30] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv: 1706.06083, 2017
[31] Moosavi-Dezfooli S M, Fawzi A, Frossard P. Deepfool: a simple and accurate method to fool deep neural networks// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 2574-2582
[32] Moosavi-Dezfooli S M, Fawzi A, Fawzi O, et al. Universal adversarial perturbations//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA, 2017: 1765-1773
[33] Li W J, Wang S, Kang W C. Feature learning based deep supervised hashing with pairwise labels. arXiv preprint arXiv: 1511.03855, 2015
[34] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context//Proceedings of the European Conference on Computer Vision. Zurich, Switzerland, 2014: 740-755
[35] Chua T S, Tang J, Hong R, et al. Nus-wide: a real-world web image database from national university of singapore// Proceedings of the ACM International Conference on Image and Video Retrieval. Santorini, Greece, 2009: 1-9
[36] Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 2015, 115(3): 211-252
[37] Kingma D P, Ba J. Adam: A method for stochastic optimization//Proceedings of the International Conference on Learning Representations. California, USA, 2015
[38] Hubel D H, Wiesel T N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. Journal of Physiology, 1962, 160(1):106-154

HUANG Lin, Ph. D. candidate. His research interests include multimedia security and AI security.

WU Liang, M. S. candidate. His research interests include deep hashing and attack algorithms.

GAO Sheng-Yan, M. S. candidate. Her research interests include information hiding and robust watermarking.

QIN Chuan, Ph. D., professor, Ph. D. supervisor. His research interests include multimedia security and AI security.

Background

The research in this work belongs to the field of deep hashing. With the explosive growth of large-scale and high-dimensional multimedia data, approximate nearest neighbor search is widely used in search engines due to its ability to effectively balance information retrieval quality and computational efficiency. Among them, deep hashing technology is widely used as an efficient retrieval method. However, due to the vulnerability of deep neural networks, retrieval methods based on deep hashing are vulnerable to adversarial examples attacks. In order to study the insufficiency of the deep hashing algorithm and improve its security and robustness, the deep hashing attack algorithm has appeared. At present, there are mainly two kinds of deep hashing attack algorithms: targeted attack and non-targeted attack. At present, the untargeted attack adversarial examples generation method takes too long and the attack effect is not good, which is the main problem of this attack algorithm. The focus of this work is to design an efficient adversarial example generation algorithm for untargeted attacks. We use adversarial generative network, prototype network, decoder and target model for joint training, and finally obtain an efficient untargeted adversarial example generative model. Compared with existing adversarial examples generation algorithms for untargeted attacks, our designed algorithm is more efficient.

This work is supported in part by the National Natural Science Foundation of China (NSFC) project "Research on Watermarking and Hashing Algorithms towards Copyright Protection and Tampering Authentication for Neural Network Models" under Grant No. 62172280, which aims at studying theory and methodology on digital watermarking and hashing for neural network models and realizing copyright protection and tampering authentication for neural network models, in part by the NSFC project "On Detection and Recognition of Fake Media Content in Online Social Networks" under Grant No. U20B2051, which focuses on developing theories and methods for detecting and recognizing fake media contents (FMC) in online social networks and building a complete scheme of FMC forensics, FMC recognition and FMC blocking, and in part by the Natural Science Foundation of Shanghai project "Reversible Data Hiding for JPEG Images with RAW Reconstruction Capability" under Grant No. 21ZR1444600, which focuses on establishing the nonlinear relationship between RAW data and RGB data in digital images, simulating the approximate reconstruction process of RAW image with few data and developing reversible data hiding methods for JPEG images with RAW reconstruction capability.