研究与开发

用于攻击深度哈希图像检索模型的双分支自编码器网络

符思政，曹春杰，刘志远，陶方舰，孙敬张
（海南大学网络空间安全学院，海南 海口 570228）

摘　要：由于其强大的表示学习能力和高效的计算能力，基于深度学习的哈希（深度哈希）方法在大规模图像检索中被广泛应用。然而，对深度哈希模型的安全性研究较少。提出了双分支自编码器网络（DBAE）来研究这种检索的目标攻击。DBAE 的主要目标是生成难以察觉的对抗样本作为查询图像，使深度哈希模型检索的图像在语义上与原始图像无关，与目标图像相关。大量实验证明，DBAE 可以成功地生成具有小扰动的对抗样本来误导深度哈希模型，验证了这些扰动在各种设置下的可迁移性。

关键词：目标攻击；深度哈希；对抗攻击；图像检索
中图分类号：TP393
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2023246

Dual-branch autoencoder network for attacking deep hashing image retrieval models

FU Sizheng, CAO Chunjie, LIU Zhiyuan, TAO Fangjian, SUN Jingzhang
School of Cyberspace Security, Hainan University, Haikou 570228, China

Abstract: Due to its powerful representation learning capabilities and efficient computing capabilities, deep learning-based hashing (deep hashing) methods are widely used in large-scale image retrieval. However, there are less studies on the security of deep hashing models. A dual-branch autoencoder network (DBAE) to study targeted attacks on such retrieval was proposed. The main goal of DBAE was to generate imperceptible adversarial samples as query images in order to make the images retrieved by the deep hashing model semantically irrelevant to the original image and relevant to the target image. Numerous experiments demonstrate that DBAE can successfully generate adversarial samples with small perturbations to mislead deep hashing models, and italso verifies the transferability of these perturbations under various settings.

Key words: targeted attack, deep hashing, adversarial attack, image retrieval

收稿日期：2023−08−15；**修回日期：**2023−11−10
基金项目：国家自然科学基金联合基金资助项目（No.U19B2044）；海南省重点研发计划项目（No.ZDYF2020012）
Foundation Items: The National Natural Science Foundation of China (No.U19B2044), The Key Research and Development Project of Hainan Province, China (No.ZDYF2020012)

0　引言

随着互联网上的数据量以指数级的速度增长，如何快速且准确地检索信息变得越来越重要。作为信息检索的技术之一，哈希方法[1]将图像转换为离散的哈希码，语义相似的图像会得到相似的哈希码并在汉明空间中的同一片区域聚集，哈希码的相似性保持特性构成了哈希方法可用性的基础。近年来，由于深度学习强大的表征学习能力，基于深度学习的哈希方法（深度哈希）在学习哈希码方面获得了相当大的成功[2]。深度哈希方法已经在大规模图像检索中处于领先地位，这是因为哈希码的相似性计算大大降低了检索速度，同时还能获得出色的检索精确率[3-4]。

尽管深度哈希方法取得了可喜的结果，但最近的研究指出，大多数深度神经网络（deep neural network，DNN）容易受对抗样本的影响，这也给现有的深度哈希模型带来了重大的安全问题[5-7]。具体来说，攻击者给干净的图像引入了小的、精心构造的扰动，这些扰动不会被人类注意到，但会导致模型做出错误的判断。现有的大多数关于对抗样本的研究都集中在图像分类问题上，而针对图像检索任务的对抗样本的工作却很少，而且更具挑战性。检索模型通过将查询图像映射到哈希码上，而不是映射到类别上[8]，所以基于深度哈希的图像检索面临着不同的攻击面，即目标标签不能直接作为对抗样本的学习目标，因此需要找到一个目标哈希码作为“目标标签”。尽管如此，基于深度哈希的检索模型也同样被证明容易受对抗样本的攻击[9]。攻击者可以降低检索精确率（非目标攻击）或迫使检索模型返回预期的图像（目标攻击）。目标攻击更具破坏性，例如，广告商可以迫使检索模型返回其产品的图像，以便用户免费看到他们的广告。因此，本文旨在研究深度哈希模型对对抗样本的鲁棒性，以此引起人们对检索系统安全性的一些迫切需要的关注。

本文主要研究基于深度哈希的图像检索系统中的目标攻击。与分类不同，攻击检索模型时目标标签是不能直接使用的，因此现有的攻击方法往往缺乏目标标签的信息。基于这一事实，本文提出一个标签编码器，将目标标签编码为潜在的多级语义先验特征，以用于灵活的目标对抗样本攻击。此外，现有的攻击往往使用基于优化的攻击方法，导致需要大量的迭代，为了解决攻击效率低的问题，本文提出了一种新的基于生成的攻击方式。具体如下，本文提出一种双分支自编码器网络（dual-branch auto encoder network，DBAE）对深度哈希进行目标攻击。它由 3 个主要部分组成：一个图像编码器、一个图像解码器和一个标签编码器。在图像编码器中，有两个分支：细节分支和语义分支。细节分支用于提取具有空间细节的内容，以收集浅层和边缘信息。语义分支用于提取深层次的语义信息。干净的图像首先被输入图像编码器中，以产生不同层次的图像细节信息和语义图像信息，之后利用标签编码器将目标标签编码为适当大小的语义标签信息，将指定的语义标签信息和中间的语义图像信息结合形成新的语义信息，最后将细节信息和语义信息融合输入图像解码器中，产生目标对抗样本，目标对抗样本作为查询图像使深度哈希模型返回与目标标签相关的图像序列。

本文的主要研究工作如下。

- 本文提出了一种双分支自编码器网络，作为针对深度哈希的攻击方式。一旦网络被训练，就不需要访问目标深度哈希模型，给定任意目标标签和原始图像，高质量的对抗样本可以快速生成。
- 本文设计了一个标签编码器，应用语义标签的先验信息来训练双分支自编码器网络模型，使模型在攻击中增强获取目标标签信息的能力。

- 本文在3个多标签数据集上进行了大量实验，结果表明，提出的方法达到较先进的水平。

1 相关工作

1.1 基于深度哈希的相似性检索

根据是否使用标签信息，基于深度哈希的图像检索模型方法可分为有监督攻击[2-4,10-11]和无监督攻击[8,12-14]。文献[8]提出了第一个无监督攻击，它使用多个分离的深度信念网络（deep belief network，DBN）和一个多模式深度信念网络（multimode deep belief network，MDBN）来学习多模式数据的紧凑哈希码。文献[13]引入平滑的符号激活函数来解决在没有连续变量二进制化的情况下直接生成二进制码的问题。

与无监督方法相比，有监督方法使用标签的语义信息来训练网络模型，可以取得更好的性能。第一个监督策略是由 Xia 等[2]提出的。它将训练过程分成两个阶段——第一个阶段学习哈希编码，第二个阶段训练卷积神经网络（convolutional neural network，CNN）以产生哈希码。基于成对的标签指导，深度成对监督哈希（deep pair-wise-supervised hashing，DPSH）算法[4]提出了一个新的交叉熵损失函数来评估哈希码的质量，并增加了一个正则化组件来最小化量化错误。HashNet 算法[3]通过收敛的顺序方法直接进行哈希码学习，从语义相似的数据中学习得到精确的二进制码。

1.2 对抗攻击

对抗攻击是在输入样本中引入人类无法察觉的、有目的的扰动，导致模型结果不准确[5]。根据攻击者对模型和数据集的熟悉程度，它们可以被分为白盒攻击[5,15-18]和黑盒攻击[19-20]。对抗样本的想法最早是由 Szegedy 等[5]提出的，通过增加一个小的扰动来欺骗网络对图像进行错误分类，这个扰动是通过最大化网络的预测误差发现的。此后，更多的白盒攻击策略被提出，包括快速梯度符号（fast gradient sign method，FGSM）攻击[15]、DeepFool[16]、梯度投影下降（projected gradient descent，PGD）攻击[17]、CW（carlini & wagner）攻击[18]。文献[19]提出的零阶优化（zeroth order optimization，ZOO）方法作为黑盒攻击的正式实现，通过对一阶和二阶导数的不同估计和分层攻击，缩短了训练时间，保证了训练效果。

目前先进的深度哈希非目标攻击，即哈希对抗生成（hash adversary generation，HAG）攻击[9]，试图在汉明空间中尽可能地将对抗样本从主要原始区域中移除，使深度哈希模型的检索结果与原始查询在语义上无关，从而降低原始查询检索精确率。点对点（point-to-point，P2P）攻击[21]随机选择属于目标标签图像的哈希码作为学习目标，旨在学习检索结果与目标查询在语义上相关的对抗样本。深度哈希定向攻击（deep hashing targeted attack，DHTA）[21]是一种目标攻击方法，它设计了一种投票机制，将点对点问题转化为点对集问题，使用锚码作为对抗样本的学习目标，并使用 PGD 来优化建议的攻击。DHTA 的锚码机制使深度哈希模型检索得到与攻击者计划相关的目标图像的精确率大大增加，然而，一方面，DHTA 采用 PGD[17]来优化提出的攻击，这种基于优化的攻击需要大量的迭代，效率很低；另一方面，其攻击缺乏目标标签的语义先验信息。BadHash[22]是第一种针对深度哈希的后门攻击，它使用一种新颖的条件生成式对抗网络（conditional generative adversarial network，cGAN）管道有效地生成有毒样本，有毒样本在不触发的情况下表现得与干净样本一样正常，攻击者一旦开启触发器，有毒样本就会攻击深度哈希模型。AdvHash[23]是第一个深度哈希补丁攻击，提出了一种基于内积的加权梯度聚合策略来生成补丁，攻击者不需要知道被攻击的图像是什么样的，只需要将补丁加上，深度哈希模型就会被欺骗，误检索得到补丁对应的类别。

2　基于 DBAE 模型的深度哈希目标攻击

2.1　问题定义

（1）深度哈希模型

本节简要介绍基于深度哈希的检索过程。让 $X=\{(x_i,y_i)\}_{i=1}^{N}$ 代表一个由 N 个标有 L 个类别的图像组成的数据集，其中 x_i 是查询图像，$\boldsymbol{y}_i=[y_{i1},\cdots,y_{iL}]\in\{0,1\}^L$ 是相应的多标签向量。$y_{ij}=1$ 代表第 i 张查询图像属于第 j 个类别。

查询图像 x 的哈希码 c 是经过深度哈希模型 $F(x)$ 之后产生的，具体如下：

$$c=F(x)=\operatorname{sign}\left(f_\theta(x)\right)\quad \text{s.t.}\quad c\in\{-1,1\}^K \tag{1}$$

其中，$f_\theta(x)$ 是一个由哈希层和特征提取器组成的 CNN。可以使用 ResNet[24]、VGG[25]等特征提取器，哈希层是一个全连接层，输出为 K 个节点。在训练过程中，用 $\tanh(x)$ 函数代替 $\operatorname{sign}(x)$ 函数进行反向传播，因为 $\operatorname{sign}(x)$ 函数输出的是离散的二进制代码，结果不收敛。

（2）相似性计算

查询图像 x_i 首先通过深度哈希模型 $F(x)$ 得到哈希码 c_i。之后，将哈希码 c_i 与从数据集 X 中得到的哈希矩阵 $\boldsymbol{C}_X$ 进行比较，以计算汉明距离 $d_{\mathrm{H}}(c_i,\boldsymbol{C}_X)$，本文用内积表示汉明距离，其计算式为 $d_{\mathrm{H}}(c_i,\boldsymbol{C}_X)=(K-c_i\cdot\boldsymbol{C}_X)/2$。最后，根据相似度排名检索出图像序列。

2.2　DBAE 模型

本文提出一种双分支自编码器网络（DBAE）来生成对抗样本攻击深度哈希模型，即深度哈希模型对对抗样本的鲁棒性研究，它由一个图像编码器 E_i、一个图像解码器 D_i 和一个标签编码器 E_l 组成，模型框架如图 1 所示。

（1）训练阶段

干净的图像 x 被预处理成固定大小，然后由 E_i 处理成低维数据，受文献[26-27]的启发，本文将 E_i 设计成一个双分支结构，即细节分支 E_{id} 和语义分支 E_{is}。E_{id} 由一个 4 层的 DenseBlock 结构组成。在每一层，前面几层的特征输出被作为当前层的输入，而当前层的特征输出又被输入后面几层，形成一个完整的密集连接机制。由于网络密集连接和采样率低下，E_{id} 能够更好地提取低层次特征的空间细节信息。E_{is} 用来提取高层次的语义信息。其过程如下：E_{is} 通过 3 次快速下采样将图像编码为多级语义图像特征嵌入

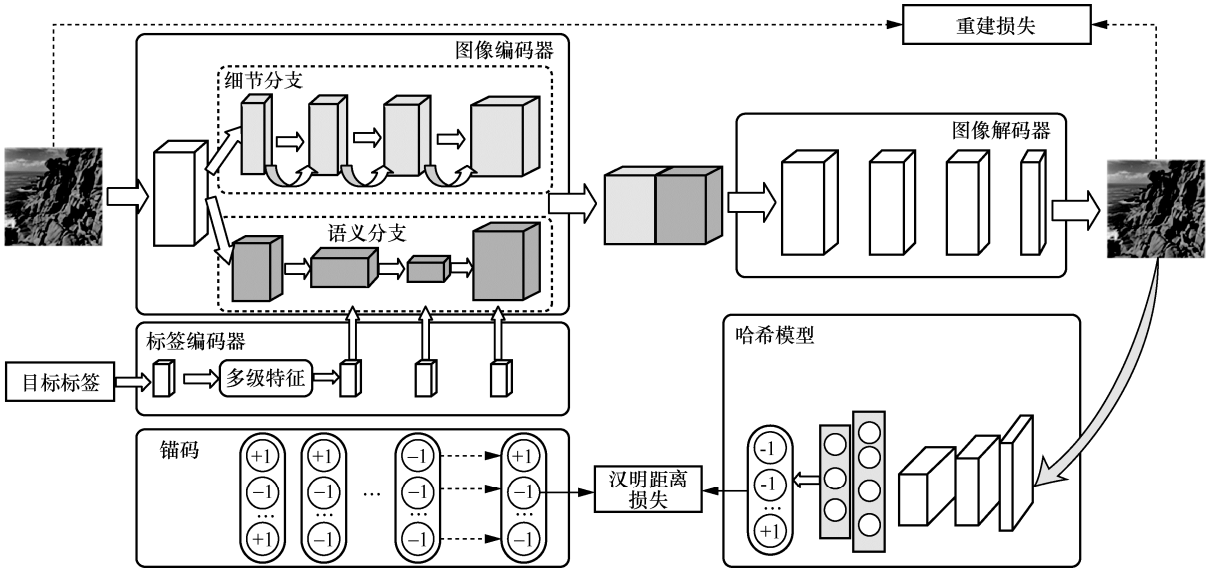


图 1　双分支自编码器网络模型框架

$\overline{F_{is}}=(F_{is}^1,F_{is}^2,F_{is}^3)$，快速下采样策略可以提升通道数且允许更多的信息编码。E_l 由 3 个全连接层和 4 个去卷积层组成，能够将标签上采样到指定大小，此时 E_l 将标签编码为相应尺寸的多级语义标签特征 $\overline{F_l}=(F_l^1,F_l^2,F_l^3)$，并将指定的 $\overline{F_l}$ 与 $\overline{F_{is}}$ 融合，得到一个包含语义标签信息和语义图像信息的新语义特征 E_{is}。最后，细节特征和语义特征被融合并输入 D_i，这部分旨在从隐藏空间的表示中重构输入以生成对抗性样本 $\bar{x}$。由于此时处于训练阶段，生成的对抗样本还不是最优解，即对抗样本 $\bar{x}$ 的哈希码与目标哈希码依然有差距。

对于目标哈希码的选择，即对抗样本的学习目标（“目标标签”），本文调用了文献[21]设计的投票机制。过程如下：选择属于目标标签的 9 张图像，将图像集通过深度哈希模型得到哈希码，然后计算这些哈希码中相应比特的+1 和−1 的数量，若+1 的数量大于−1 的数量，则目标哈希码的相应比特位为+1，反之为−1，本文将得到的目标哈希码称为锚码。为了令对抗样本 $\bar{x}$ 的哈希码尽可能地与锚码相似，本文将对抗样本 $\bar{x}$ 的哈希码与锚码用于计算汉明距离损失。在训练过程中，损失不断减少，生成的对抗样本 $\bar{x}$ 的哈希码接近锚码，直到它们之间的汉明距离最小。经过 100 个 epoch 后，对抗样本 $\bar{x}$ 的哈希码在汉明空间中无限接近目标簇群，使得检索系统返回与目标标签相关的图像集。提出机制的训练过程如算法 1 所示。

算法 1 DBAE 模型的训练过程

输入： 图像数据集 $X=\{(x_j,\boldsymbol{y}_j)\}_{j=1}^N$，目标标签集 $L=\{t_j\}_{j=1}^N$，哈希模型 $F(x)=\mathrm{sign}\big(f_\theta(x)\big)$，DBAE 包括图像编码器 E_i（细节分支 E_{id} 和语义分支 E_{is}）、图像解码器 D_i 和标签编码器 E_l

输出： 网络参数 θ_a

初始化： 学习率 η，训练轮数 K=100

for k=1:K do

for j=1:N do

$F_{\mathrm{id}}\leftarrow E_{i\mathrm{d}}(x_j)$

$\overline{F_{is}}=(F_{is}^1,F_{is}^2,F_{is}^3)\leftarrow E_{is}(x_j)$

$\overline{F_l}=(F_l^1,F_l^2,F_l^3)\leftarrow E_l(t_j)$

$F_{is}=\overline{F_{is}}+\overline{F_l}$

$\overline{x_j}\leftarrow D_i(F_{\mathrm{id}},F_{\mathrm{is}})$

$c_j\leftarrow f_\theta(\overline{x_j})$

从数据库中选择属于目标标签 t_j 的 9 张图像，并获得它们的哈希码 $\{c_i\}_{i=1}^9$

锚码 $c_a\leftarrow\mathrm{sign}(\mathrm{sum}(\{c_i\}_{i=1}^9,\dim=0))$

$L_{\mathrm{AE}}\leftarrow L_{\mathrm{ham}}(c_j,c_a)+L_{\mathrm{re}}(x_j,\overline{x_j})$

通过梯度下降更新 θ_a

$\theta_a\leftarrow\theta_a-\eta\Delta_{\theta_a}L_{\mathrm{AE}}$

end

end

其中，α 和 β 是超参数，θ_a 是网络参数，L_{ham} 是汉明距离损失，L_{re} 是重建损失。

在进行复杂度分析之前，便于推导出算法 1 的算法复杂度计算式，需要对常用的操作的时间值赋予一些变量表示。

- M 表示每个卷积核输出特征图的边长。
- K 表示每个卷积核的边长。
- C_{in} 表示卷积层的输入通道数，C_{out} 表示卷积层的输出通道数。
- d_{in} 表示全连接层的输入维度，d_{out} 表示全连接层的输出维度。

每个卷积层的时间复杂度由输出特征图面积 M^2、卷积核面积 K^2、输入通道数 C_{in} 和输出通道数 C_{out} 决定，即 $\mathrm{Time}\sim O(M^2\cdot K^2\cdot C_{\mathrm{in}}\cdot C_{\mathrm{out}})$。其中，输出特征图尺寸本身又由输入矩阵尺寸 X、卷积核尺寸 K、Padding、Stride 这 4 个参数决定，表示为 $M=\dfrac{X-K+2\cdot\mathrm{Padding}}{\mathrm{Stride}}+1$。每个全连接层的时间复杂度由输入维度 d_{in} 和输出维度 d_{out} 决定，即 $\mathrm{Time}\sim O(d_{\mathrm{in}}\cdot d_{\mathrm{out}})$，全连接层的时间复杂度相对于卷积层较小，故可忽略不计。

整个算法 1 的时间复杂度为 Time~$O\left(\sum_{l=1}^{D} M_l^2 \cdot K_l^2 \cdot C_{l-1} \cdot C_l\right)$，其中 D 为网络的深度，l 表示神经网站中第 l 个卷积层。

对于空间复杂度，由于全连接层的参数个数远远多于卷积层，因此算法 1 的空间复杂度主要由全连接层决定，即 Space~$O(d_{\text{in}} \cdot d_{\text{out}} + d_{\text{out}})$。

（2）攻击阶段

一旦训练完成，网络只需向 DBAE 输入目标标签 t 和干净的图像 x，就能在单向传递中生成目标对抗性样本 $\bar{x}$，而无须访问目标哈希模型。

干净的输入图像 x 首先被调整为 224×224。在 E_{id} 中，x 的大小保持不变，通道的数量从 32 依次增加到 64。在 E_{is} 中，x 的通道数依次增加到 35、67、131，然后减少到 64，尺寸更改为 224、112、56 和 28，其中语义特征包括语义标签信息。语义标签信息被 E_i 编码为相应的尺寸，通道数为 3。

2.3　损失函数

本文在网络中建立了两个损失函数，包括汉明距离损失和重建损失，以有效地生成高质量的对抗样本。以下是相关计算式：

$$\min_{\theta_a} L_{\text{AE}} = \sum_{(x,y)\in X} (\alpha L_{\text{ham}} + \beta L_{\text{re}}) \tag{2}$$

L_{ham} 损失函数计算出两个字符串之间的差异。本文利用锚码作为目标哈希码，并鼓励对抗样本的哈希码尽可能地接近锚码：

$$L_{\text{ham}} = d_H(c_a, c_{\bar{x}}) \tag{3}$$

其中，$d_H(,)$ 是算子，$c_{\bar{x}}$ 是对抗样本的哈希码，c_a 是锚码。在计算中，本文用内积表示汉明距离损失。

因此，汉明距离损失最后表示为：

$$L_{\text{ham}} = -\frac{1}{K} c_a^{\text{T}} f_\theta(\bar{x}) \tag{4}$$

其中，K 是哈希码的长度，$f_\theta(\bar{x})$ 是未二值化的哈希码，这是因为离散的哈希码在训练期间很难反向传播。

L_{re} 损失函数迫使重建的图像与输入的图像尽可能相似。也就是说，产生的扰动要尽可能小，本文用 L_2-norm 表示 L_{re}：

$$L_{\text{re}} = x - \bar{x}_2^2 \tag{4}$$

3　实验与分析

3.1　数据集

本文使用 3 个数据集评估提出的方法，即 FLICKR-25K[28]、MS-COCO[29]和 NUS-WIDE [30]。FLICKR-25K 数据集包含 25 000 张图片，根据文献[31]，本文选择 1 700 张图片作为查询集，5 000 张作为训练集，其余的作为数据库。MS-COCO 数据集有 122 218 张图片，根据文献[3]，本文选择 5 000 张图片作为查询集，10 000 张作为训练集，其余的图片作为数据库。NUS-WIDE 数据集有 195 834 张图片，按照文献[32]，本文选择 2 100 张作为查询集，10 000 张作为训练集，其余的作为数据库。

3.2　评估设置

对于被攻击的目标深度哈希模型，本文使用骨干网络为 ResNet50 的 DPSH[4]框架作为默认方法，其中 ResNet50 的最后一个全连接层由哈希层代替，即一个新的全连接层和 Tanh 激活函数。当然提出的方法在其他骨干网络（如 VGG11）或其他哈希方法（如 HashNet[3]）也可以得到同样的效果。

对于评价指标，继文献[21]之后，本文使用目标平均精确率（targeted mean average precision，t-MAP）指标而不是平均精确率（mean average precision，MAP）指标。MAP 是信息检索中广泛使用的平均精确率指标，以原始标签为参考标签，t-MAP 以目标标签为参考标签，也就是说，t-MAP 越高，深度哈希模型检索得到与目标标签相关的图像精确率就越高，攻击就越有效。另外，

本文还展示了精确率−召回率（precision-recall，PR）曲线和精确率@top$N$曲线，以进行更全面的比较。精确率−召回率曲线是用于评估分类模型在不同阈值下精确率和召回率之间的权衡关系的一种可视化工具。它通过绘制不同阈值下的精确率和召回率之间的曲线来展示模型的性能。精确率@top$N$曲线指最后概率向量最大的前N名中出现正确概率的精确率。最后，本文使用生成时间指标来对比 DBAE 和其他攻击方法的效率。生成时间即对于整个测试集，每张对抗样本图像的生成时间。

对于 DBAE 的参数，本文将α设置为 5，β设置为 50，使用 Adam 来优化网络，学习率为10^{-4}，epoch 为 100，批处理大小为 24。DBAE 由 PyTorch 实现，整个训练过程在 NVIDIA Tesla V100 GPU 中运行。

3.3 实验结果

3.3.1 目标攻击性能

本文将 DBAE 模型生成的对抗样本作为 DPSH 深度哈希模型的查询图像，获得的 t-MAP 值与各种攻击方法进行比较，3 个数据集上不同攻击方法的 t-MAP 见表 1，其中 DPSH 深度哈希模型分为 12 bit、24 bit、32 bit 和 48 bit，分别代表哈希码的长度。其他比较的攻击方法是 P2P、DHTA-FGSM 和 DHTA，其中 DHTA-FGSM 使用 FGSM 方法来优化攻击，DHTA 使用 PGD 方法来优化攻击。原始 t-MAP 值代表以干净图像为查询输入，目标标签为参考标签的攻击性能。所有其他攻击方法的 t-MAP 值代表以各自方法制作的对抗样本为查询输入，目标标签为参考标签的性能。其结果显示都高于原始 t-MAP 值，这也证明了深度哈希模型对对抗样本的鲁棒性很差。DBAE 的 t-MAP 值随着哈希码的长度的增加而增大，这是因为哈希码的长度越长，所携带的信息就越多。DBAE 的 t-MAP 值高于其他所有的攻击方法，总体而言，DBAE 比 P2P 的精确率高 4%左右，比 DHTA-FGSM 的精确率高 14%左右，比 DHTA 的精确率高 1%左右。特别是，对于检索 NUS-WIDE 数据集的 DPSH 48 bit 深度哈希模型，DBAE 获得的 t-MAP 值比 DHTA 高 1.83%，比 DHTA-FGSM 高 17.68%，比 P2P 高 3.91%。本文还展示了精确率−召回率曲线和精确率@top$N$曲线，3 个数据集上不同攻击方法的精确率−召回率曲线和精确率@top$N$曲线（48 bit）如图 2 所示，DBAE 的曲线始终高于其他攻击方法的曲线，进一步说明了 DBAE 的优越性。其中精确率−召回率曲线表明 DBAE 在不同阈值下精确率和召回率都比其他攻击方法更高，精确率@top$N$曲线表明 DBAE 在检索的前N名中出现正确概率的精确率高于其他攻击方法。此外，干净图像和其对抗样本在 NUS-WIDE 检索的前 5 个相似性样本如图 3 所示。

3.3.2 效率

在长度为 48 bit 的 3 个数据集上，各种攻击方法的 t-MAP 和生成时间见表 2。可以看出，DBAE

表 1 3 个数据集上不同攻击方法的 t-MAP

方法	指标	FLICKR-25K				NUS-WIDE				MS-COCO			
		12bit	24bit	32 bit	48 bit	12 bit	24 bit	32 bit	48 bit	12 bit	24 bit	32 bit	48 bit
Original	t-MAP	65.52%	65.91%	66.06%	66.76%	56.32%	56.39%	56.47%	57.62%	43.81%	44.61%	45.47%	45.52%
DHTA-FGSM	t-MAP	73.96%	74.23%	74.41%	74.45%	59.08%	61.49%	62.81%	63.04%	48.14%	48.75%	49.18%	50.10%
P2P	t-MAP	83.57%	85.11%	85.81%	85.84%	74.27%	76.24%	76.41%	76.81%	59.30%	61.10%	60.01%	60.11%
DHTA	t-MAP	86.69%	88.21%	89.41%	89.36%	76.68%	78.42%	78.76%	78.89%	62.04%	64.49%	63.76%	64.17%
DBAE	t-MAP	**87.93%**	**89.10%**	**90.16%**	**90.38%**	**77.42%**	**79.81%**	**79.96%**	**80.72%**	**63.62%**	**65.80%**	**64.63%**	**65.43%**
Original	MAP	79.97%	81.36%	82.10%	82.54%	70.39%	71.89%	71.64%	72.55%	58.60%	61.88%	61.46%	62.33%

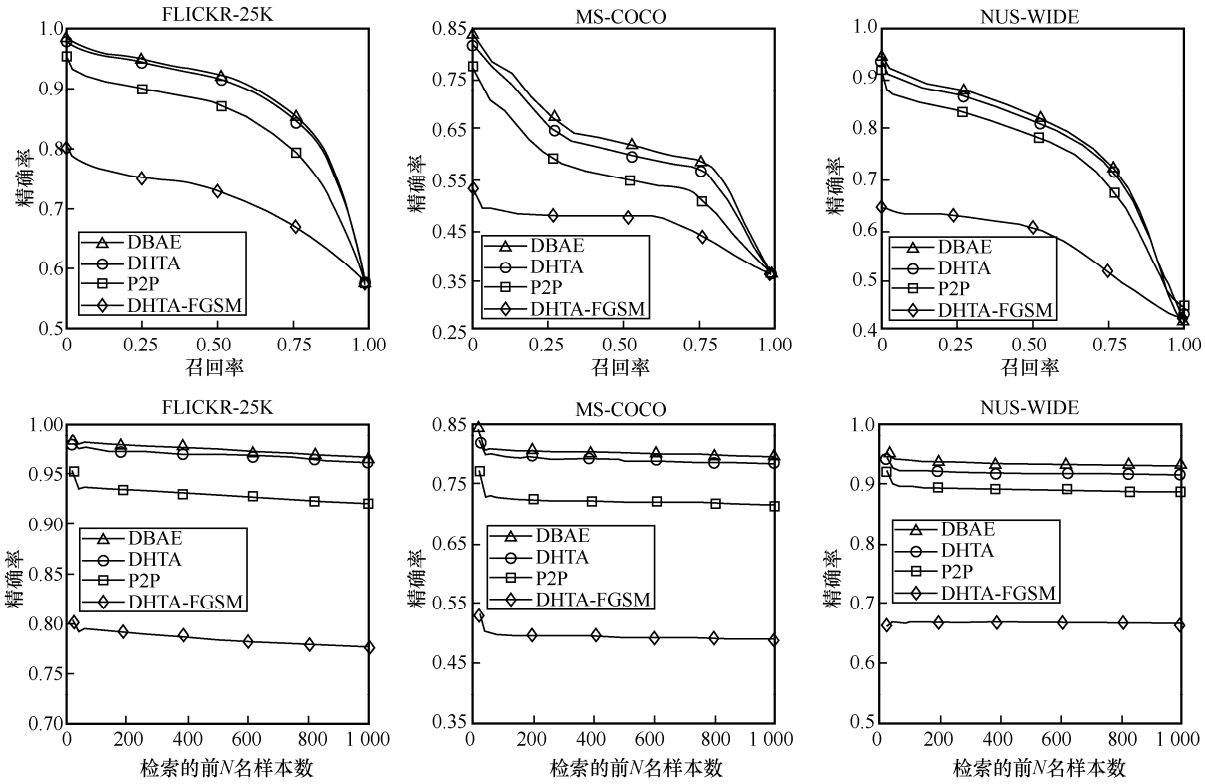


图 2　3 个数据集上不同攻击方法的精确率−召回率曲线和精确率@top$N$ 曲线（48 bit）

图 3　干净图像和其对抗样本在 NUS-WIDE 检索的前 5 个相似性样本

表 2　在长度为 48 bit 的 3 个数据集上，各种攻击方法的 t-MAP 和生成时间

攻击方法	迭代次数	FLICKR-25K		NUS-WIDE		MS-COCO	
		t-MAP	生成时间/s	t-MAP	生成时间/s	t-MAP	生成时间/s
DHTA-FGSM	1	74.45%	0.021	63.04%	0.021	50.10%	0.024
DHTA	2 000	89.36%	9.37	78.89%	9.19	64.17%	9.36
DBAE	1	90.38%	0.021	80.72%	0.023	65.43%	0.013

除了具有较高的 t-MAP，其生成时间也比其他方法快很多。例如，在 MS-COCO 中，DBAE 生成每张对抗样本的时间为 0.013 s，而 DHTA 需要 9.36 s，相差 720 倍。DHTA-FGSM 的生成时间虽然也很短，但其 t-MAP 却很差。

3.3.3 跨家族迁移

在这个实验中，本文考虑了一个更隐蔽的场景，即跨家族迁移。跨家族迁移被描述为使用由一种类型的深度哈希模型产生的对抗样本来攻击不同类型的深度哈希模型。例如，HashNet 方法的深度哈希模型被 DBAE 模型和 DPSH 方法的深度哈希模型训练产生的对抗样本所攻击。跨家族迁移显然是一个更现实的场景，攻击者不知道目标深度哈希模型是由哪种方法组成的。

本文对 DBAE 和 DHTA 进行了实验，DBAE 和 DHTA 在 HashNet-DPSH 下跨家族迁移的 t-MAP 分别见表 3 和表 4，表示应用 DBAE 从 HashNet 深度哈希模型生成的对抗样本来攻击 DPSH 深度哈希模型。“D-ResNet50”代表以 ResNet50 为骨干网络的 12 bit DPSH 哈希模型，“H-VGG11”代表以 VGG11 为骨干网络的 12 bit HashNet 哈希模型，“*”是对应哈希模型的 48 bit 变体。从表 3 和表 4 可以看出，DBAE 的转移性远远超过了 DHTA。例如，由 DBAE 和 H-VGG11* 哈希模型产生的对抗样本来攻击 D-ResNet50*的哈希模型得到的 t-MAP 值为 68.86%，而 DHTA 的 t-MAP 值为 59.51%。这表明 DBAE 生成的对抗样本不仅具有快速的生成速度和较高的攻击精确率，而且还具有非常出色的可转移性。

3.4 消融实验

标签编码器的影响：为了探索标签编码器对攻击性能的影响，本文从 DBAE 中删除了标签编码器，即只有图像编码器和图像解码器的 DBAE1。本文对 MS-COCO 数据集进行实验，DBAE 和 DBAE1 在 MS-COCO 数据集上的不同比特的 t-MAP 如图 4 所示。由图 4 可知，不同比特长度下，DBAE 总是优于 DBAE1，表明标签编码器提供的语义标签信息能够大幅度提高攻击精确率。

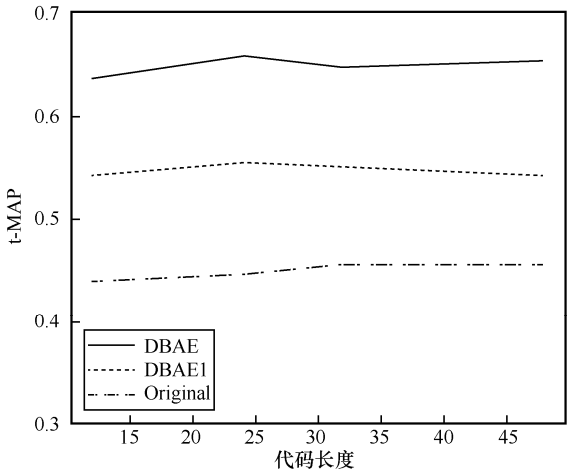


图 4 DBAE 和 DBAE1 在 MS-COCO 数据集上的不同比特的 t-MAP

4 结束语

本文提出一种双分支自动编码器网络来对深度哈希进行目标攻击。具体来说，提出的自动编

表 3 DBAE 在 HashNet-DPSH 下跨家族迁移的 t-MAP

被攻击模型	DBAE & H-ResNet50	DBAE & H-ResNet50*	DBAE & H-VGG11	DBAE & H-VGG11*
D-ResNet50*	76.11%	75.63%	63.08%	68.86%
D-VGG11*	60.16%	62.56%	77.36%	76.44%

表 4 DHTA 在 HashNet-DPSH 下跨家族迁移的 t-MAP

被攻击模型	DHTA & H-ResNet50	DHTA & H-ResNet50*	DHTA & H-VGG11	DHTA & H-VGG11*
D-ResNet50*	68.14%	65.22%	59.39%	59.51%
D-VGG11*	57.36%	57.42%	74.11%	73.56%

码器由 3 部分组成，分别为图像编码器、图像解码器和标签编码器。图像编码器有两个分支，用于提取输入图像的细节分支和语义分支，标签编码器将指定的标签信息合并到中间的语义特征中，最后将这两个特征融合后输入图像解码器中，生成对抗样本。大量的实验表明，提出的方法具有优越的攻击性能和更好的可迁移性。

由于深度神经网络的独特之处在于可以从数据中学习高级特征，因此对抗样本问题不仅能评估深度哈希模型的鲁棒性和安全性，还可以用于提高模型的鲁棒性。未来，笔者团队计划将双分支自编码器生成的对抗样本引入深度哈希模型的训练中，使哈希模型能够更好地适应这种改变，从而对对抗样本具有鲁棒性。并且，笔者将讨论如何将现有的对抗训练方法从点对点问题推广到点对集的防御方案，具体方法将在今后的工作中进一步论证。

参考文献：

[1] XU Y J, ZHAO X H, LIANG Y C. Robust power control and beamforming in cognitive radio networks: a survey[J]. IEEE Communications Surveys & Tutorials, 2015, 17(4): 1834-1857.

[2] XIA R, PAN Y, LAI H, et al. Supervised hashing for image retrieval via image representation learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2014.

[3] CAO Z J, LONG M S, WANG J M, et al. HashNet: deep learning to hash by continuation[C]//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2017: 5609-5618.

[4] LI W J, WANG S, KANG W C. Feature learning based deep supervised hashing with pairwise labels[J]. arXiv preprint arXiv:1511.03855, 2015.

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

[6] 胡永进, 郭渊博, 马骏, 等. 基于对抗样本的网络欺骗流量生成方法[J]. 通信学报, 2020, 41(9): 59-70.
HU Y J, GUO Y B, MA J, et al. Method to generate cyber deception traffic based on adversarial sample[J]. Journal on Communications, 2020, 41(9): 59-70.

[7] 程旭, 王莹莹, 张年杰, 等. 基于空间感知的多级损失目标跟踪对抗攻击方法[J]. 通信学报, 2021, 42(11): 242-254.
CHENG X, WANG Y Y, ZHANG N J, et al. Multi-level loss object tracking adversarial attack method based on spatial perception[J]. Journal on Communications, 2021, 42(11): 242-254.

[8] WANG D, CUI P, OU M, et al. Learning compact hash codes for multimodal representations using orthogonal deep structure[J]. IEEE Transactions on Multimedia, 2015, 17(9): 1404-1416.

[9] YANG E, LIU T, DENG C, et al. Adversarial examples for hamming space search[J]. IEEE transactions on cybernetics, 2018, 50(4): 1473-1484.

[10] ZHANG R, LIN L, ZHANG R, et al. Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification[J]. IEEE Transactions on Image Processing, 2015, 24(12): 4766-4779.

[11] LIU B, CAO Y, LONG M, et al. Deep triplet quantization[C]//Proceedings of the 26th ACM international conference on Multimedia. New York: ACM Press, 2018: 755-763.

[12] SHEN F, XU Y, LIU L, et al. Unsupervised deep hashing with similarity-adaptive and discrete optimization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(12): 3034-3044.

[13] SONG J, HE T, GAO L, et al. Binary generative adversarial networks for image retrieval[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2018, 32(1).

[14] LIN K, LU J W, CHEN C S, et al. Learning compact binary descriptors with unsupervised deep neural networks[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 1183-1192.

[15] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv preprint arXiv:1412.6572, 2014.

[16] MOOSAVI-DEZFOOLI S M, FAWZI A, FROSSARD P. DeepFool: a simple and accurate method to fool deep neural networks[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 2574-2582.

[17] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv preprint arXiv:1706.06083, 2017.

[18] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//Proceedings of 2017 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2017: 39-57.

[19] CHEN P Y, ZHANG H, SHARMA Y, 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. New York: ACM Press, 2017: 15-26.

[20] MOPURI K R, OJHA U, GARG U, et al. NAG: network for adversary generation[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 742-751.

[21] BAI J, CHEN B, LI Y, et al. Targeted attack for deep hashing based retrieval[C]//Proceedings of Computer Vision–ECCV 2020: 16th European Conference. Cham: Springer, 2020: 618-634.

[22] HU S S, ZHOU Z Q, ZHANG Y C, et al. BadHash: invisible backdoor attacks against deep hashing with clean label[C]//Proceedings of the 30th ACM International Conference on Multimedia. New York: ACM Press, 2022: 678-686.

[23] HU S S, ZHANG Y C, LIU X G, et al. AdvHash: set-to-set targeted attack on deep hashing with one single adversarial patch[C]//Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM Press, 2021: 2335-2343.

[24] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 770-778.

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

[26] YU C Q, GAO C X, WANG J B, et al. BiSeNet V2: bilateral network with guided aggregation for real-time semantic segmentation[J]. International Journal of Computer Vision, 2021, 129(11): 3051-3068.

[27] FU Y, WU X J. A dual-branch network for infrared and visible image fusion[C]//Proceedings of 2020 25th International Conference on Pattern Recognition (ICPR). Piscataway: IEEE Press, 2021: 10675-10680.

[28] HUISKES M J, LEW M S. The MIR flickr retrieval evaluation[C]//Proceedings of the 1st ACM international conference on Multimedia information retrieval. New York: ACM Press, 2008: 39-43.

[29] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[M]//Computer Vision – ECCV 2014. Cham: Springer International Publishing, 2014: 740-755.

[30] CHUA T S, TANG J H, HONG R C, et al. NUS-WIDE: a real-world web image database from National University of Singapore[C]//Proceedings of the ACM International Conference on Image and Video Retrieval. New York: ACM Press, 2009: 1-9.

[31] WANG Z J, ZHANG Z, LUO Y D, et al. Deep collaborative discrete hashing with semantic-invariant structure[C]// Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2019: 905-908.

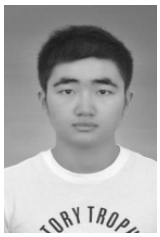
[32] JIANG Q Y, LI W J. Deep cross-modal hashing[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2017: 3270-3278.

[作者简介]

符思政（1999– ），男，海南大学硕士生，主要研究方向为人工智能安全、对抗样本攻防、图像检索。

曹春杰（1977– ），男，博士，海南大学博士生导师，主要研究方向为认知无线网络安全、区块链、人工智能安全等。

刘志远（1999– ），男，海南大学硕士生，主要研究方向为对抗样本攻防、异常检测、图对比学习。



陶方舰（1991– ），男，海南大学博士生，主要研究方向为人工智能安全、对抗样本攻防。

孙敬张（1993– ），男，博士，海南大学硕士生导师，主要研究方向为无线安全、深度学习、图像处理。