BACH: Black-Box Attacking on Deep Cross-Modal Hamming Retrieval Models

Jie Zhang[1], Gang Zhou[2,3], Qianyu Guo[4(✉)], Zhiyong Feng[1], and Xiaohong Li[1(✉)]

[1] College of Intelligence and Computing, Tianjin University, Tianjin, China
xiaohongli@tju.edu.cn
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3] Institute of Automation, Chinese Academy of Sciences, Beijing, China
[4] Zhongguancun Laboratory, Beijing, People's Republic of China
guoqy@zgclab.edu.cn

Abstract. The growth of online data has increased the need for retrieving semantically relevant information from data in various modalities, such as images, text, and videos. Thanks to the powerful representation capabilities of deep neural networks (DNNs), deep cross-modal hamming retrieval (i.e., DCMHR) models have become popular in cross-modal retrieval tasks due to their efficiency and low storage cost. However, the vulnerability of DNN models makes them susceptible to small perturbations. Existing attacks on DNN models focus on supervised tasks like classification and recognition, and are not applicable to DCMHR models. To fill this gap, in this paper, we present BACH, an adversarial learning-based attack method for DCMHR models. BACH uses a triplet construction module to learn and generate well-designed adversarial samples in a black-box setting, without prior knowledge of the target models. During the learning process, we estimate the gradient of the objective function by using random gradient-free (RGF) method. To evaluate the effectiveness and efficiency of BACH, we perform thorough experiments on 3 popular cross-modal retrieval dataset and 13 state-of-the-art DCMHR models, including 6 image-to-image retrieval models and 7 image-to-text retrieval models. As a comparison, we select two established adversarial attack methods: CMLA for white-box attack and AACH for black-box attack. The results show that BACH offers comparable attack performance to CMLA while requiring no knowledge of the target models. Furthermore, BACH surpasses AACH on most DCMHR models in terms of attack success rate with limited queries.

Keywords: Cross-modal Retrieval · Hashing · Robustness · Adversarial perturbation

1 Introduction

The rapid advancement in storage and encoding techniques has greatly impacted human life by enabling people to search the internet for what they desire.

G. Zhou and J. Zhang—Contribute equally to this paper.

X. Wang et al. (Eds.): DASFAA 2023, LNCS 13945, pp. 441–456, 2023.
https://doi.org/10.1007/978-3-031-30675-4_32

While various search techniques have been developed with the growth of social media and multi-modal data, they mostly only work with similarity-based search within a single modality, such as the keyword and tag-based searches, which no longer suffice in the face of diverse multi-modal data [9,32].

To address this limitation, cross-modal retrieval has been proposed and is gaining widespread attention [6,9,18,43]. It maps data from different modalities into a common space with the same dimension, and measures the semantic similarity by comparing samples in this space. However, measuring semantic similarity between data from different modalities is a significant challenge, which is also known as the heterogeneous gap problem. Conventional cross-modal retrieval methods assess semantic similarity by measuring the distance between samples in a common space [18]. Specifically, samples from different modalities with the same semantics are close in the common space [25]. Representational encoding in this space can be either real-valued or binary [37]. While real-valued encoding is often impractical for large dataset, the binary encoding is preferred for large dataset as it reduces storage costs and speeds up retrieval [14], and is used in cross-modal hashing to map data semantics into a binary space and measure semantic similarity using Hamming distance [22].

The quality of semantic feature extraction has a significant impact on the performance of encoding. To minimize the impact of the heterogeneous gap problem, an effective feature extraction method is essential [1]. Existing hash-based cross-modal retrieval methods are based on shallow architectures [13,41], and rely on features extracted by human experts. To date, with the growth of deep learning techniques in computer vision [33], natural language processing [36], and speech analysis [17], deep neural networks (DNNs) have become popular for improving the performance of cross-modal retrieval. DNNs can effectively detect semantic similarities between different modalities, and build cross-modal correlations through their superior representational capabilities. Due to their powerful representational capabilities, DNNs are trained to identify semantic similarities between different modalities and build cross-modal correlations. Research has shown that DNN-based cross-modal retrieval models outperform traditional shallow models [7].

However, it has been well established that even a well-trained deep learning model can be easily misled by inputs with subtle, human-undetectable perturbations, known as the adversarial examples [5,26,34,40]. To date, many effective adversarial methods have been proposed to attack trained DNN models [23]. These attacks can be categorized as white-box or black-box based on whether the attacker has access to the target model's internal information. While these attack methods are primarily designed for supervised tasks such as classification or recognition, little attention has been given to studying the impact of adversarial samples on deep hamming learning in cross-modal retrieval area.

The cross-modal retrieval task differs significantly from tasks like classification and recognition. Firstly, cross-modal retrieval models are trained through unsupervised or semi-supervised methods without ground-truth labels, making them more susceptible to misleading information. Secondly, the objective of

attacks on cross-modal retrieval models is to generate semantically unrelated samples rather than incorrect classifications. This makes existing adversarial attacks unsuitable for attacking cross-modal retrieval models. Additionally, there are two major challenges in performing adversarial attacks on deep cross-modal hamming retrieval (DCMHR) models in a black-box setting: 1) the attacker does not have access to information about the target model, including the network architecture, model parameters, and loss functions, and can only obtain the output of the target model through queries; 2) there are often practical constraints on queries, such as a maximum number of queries allowed.

To tackle these challenges, we introduce BACH, a **b**lack-box adversarial **a**ttacking method for deep **c**ross-Modal **h**amming retrieval (DCMHR) models. BACH specifically targets DCMHR models and generates adversarial samples by maximizing the hamming distance of semantically similar samples, thereby greatly impairing the performance of DCMHR models. To evaluate the effectiveness of BACH, we conduct experiments on 13 state-of-the-art DCMHR models and 3 popular dataset (i.e., MIRFlickr-25K, NUS-WIDE, and CIFAR-10) in three aspects: 1) attacking DCMHR models on both image-to-image and image-to-text retrieval tasks; 2) investigating the impact of the number of samples in the query dataset used to construct triples on the attack performance; and 3) comparing BACH against the state-of-the-art white-box attack method (i.e., CMLA [20]) and the black-box attack method (i.e., AACH [19]).

To summarize, this paper makes the following contributions:

- We propose BACH, a learning-based approach for adversarial attacks on deep cross-modal hamming retrieval models in a black-box environment. Unlike existing white-box attack methods, BACH does not require any prior knowledge and thus, is more practical in real-world applications. To the best of our knowledge, BACH is the first method designed for attacking cross-modal retrieval models in a black-box setting.
- We select a query-based black-box attacking strategy with performance comparable to white-box attack methods. This is achieved through the use of the random gradient-free (RGF) method and a limited number of target model queries.
- We evaluate the effectiveness and efficiency of BACH by conducting experiments on 13 state-of-the-art cross-modal retrieval models and 3 benchmark dataset. The results show that BACH performs comparably to the white-box attack methods while only requiring a limited number of queries. Our approach can be used to assess the robustness of cross-modal retrieval models.

The rest of the paper is organized as follows. Section 2 briefly introduces deep cross-modal retrieval task and problem formulation. Section 3 presents the technical details of our approach BACH. Section 4 shows our experimental setup as well as the experimental results. Related work is discussed in Sect. 5. Section 6 presents the conclusion and future extensions of this work.

2 Background

2.1 Deep Cross-Modal Retrieval and Problem Formulation

Cross-modal retrieval task refers to using image or text as queries to search for data with another modal in the database, such as using text to search for images or using images to search for text. A well-trained DCMHR model can retrieval semantically relevant data from the database. As shown in Fig. 1(a), using a picture of a flower as a query, the DCMHR model can retrieve some text about the flower, and we define that this picture and the retrieval result (i.e., text) are semantically relevant. In this paper, we use $O = \{O^v, O^t\} = \{o_i\}_{i=1}^{\mathbf{C}}$ to represent a cross-modal database with $\mathbf{C}$ samples. Herein, sample $o_i = \{o_i^v, o_i^t\}$ is an image-text pair, where o_i^v and o_i^t represent the image data and textual data, respectively.

Generally, DCMHR use DNN to extract semantic features $F_v \in \mathbb{R}^{\mathbf{C}\times k_v}$, $F_t \in \mathbb{R}^{\mathbf{C}\times k_t}$ from the original data, where k_v, k_t is the feature-length. After using the feature extraction architecture on the dataset, the semantic features are calculated as:

$$F^v = f_{base}^v\left(O^v, \theta_{base}^v\right), F^t = f_{base}^t\left(O^t, \theta_{base}^t\right), \tag{1}$$

where θ_{base}^v, θ_{base}^t are the parameters that need to be trained for the two feature extraction architectures. Moreover, k_v and k_t are generally set to be the same in order to extract the equipotential features.

Deep cross-modal retrieval model aims to learn two hash functions f_{hash}^v, f_{hash}^t that project image or text samples onto the Hamming space. This process can be formulated as:

$$B^v = sign\left(f_{hash}^v\left(F^v, \theta_{hash}^v\right)\right), B^t = sign\left(f_{hash}^t\left(F^t, \theta_{hash}^t\right)\right), \tag{2}$$

where $B^v, B^t \in \{-1, 1\}^{\mathbf{C}\times d}$ are the binary code, d is the length of the hash space, $F^v, F^t \in [-1, 1]^{\mathbf{C}\times d}$ are the binary-like representation generated by the output layer of a target deep cross-modal network, and θ_{hash}^v and θ_{hash}^t are two parameters that need to be learned for the hash function.

The semantic similarity between samples from different modalities is evaluated by the Hamming distance of the learned binary codes in a hash space:

$$D(X, Y) = \frac{1}{2}\left(K - \langle X, Y\rangle\right), \tag{3}$$

where X and Y are the binary codes of the samples, K is a constant that maintain the distance magnitude. A well-trained cross-modal hash retrieval model should preserve semantic similarity structure between samples of different modalities. Specifically, image sample o_i^v whose Hamming distance from its positive sample $o_{i_P}^t$ (with the shortest Hamming distance) is less than negative sample $o_{i_N}^t$ (with the longest Hamming distance). Here, we can construct a sample's triple $\left\{o_i^v, o_{i_P}^t, o_{i_N}^t\right\}$. Hash function is encouraged to satisfy the inequality as follows:

$$D(f_{hash}^v(o_i^v), f_{hash}^t(o_{i_P}^t)) < D(f_{hash}^v(o_i^v), f_{hash}^t(o_{i_N}^t)) \tag{4}$$

Next, we consider generating restricted adversarial perturbations η that can fool the DCMHR models. The attacking goal can be formalized as the following inequality:

$$D(f^v_{hash}(o^v_i + \eta^v), f^t_{hash}(o^t_{i_P})) > D(f^v_{hash}(o^v_i + \eta^v), f^t_{hash}(o^t_{i_N})) \tag{5}$$

The adversarial images $o^v_i + \eta^v$ obtained by adding well-designed perturbations η can make the retrieval performance of the retrieval model significantly degraded. For example, as shown in Fig. 1(b), given an adversarial flower sample, the retrieval results are some irrelevant textual items with the original flower target.

(a) Original sample Query Results (b) Adversarial Sample Query Results

Fig. 1. Examples of query results for the original and adversarial image samples

3 Black-Box Attack on DCMHR Models

This part details our proposed black-box adversarial attack named BACH against DCMHR models. Figure 2 shows the overall working pipeline, which mainly consists of three parts. The first part carries out cross-modal querying. The second part constructs a cross-modal triplet for every image based on the query results, and the third generates the adversarial example of an image according to the cross-modal triplet. In this paper, we generate adversarial samples only for images, not text because adding perturbations to text can be easily detected.

3.1 Black-Box Attack Framework

Firstly, we input M image-text pairs samples as cross-modal data queries ($O_q = \{O^v_q, O^t_q\}$, where $O^v_q = \{o^v_i\}^M_{i=1}$ and $O^t_q = \{o^t_i\}^M_{i=1}$) to the target retrieval model. Then, we constructing a triplet $\{o^v_i, o^t_{i_P}, o^t_{i_N}\}$ for each sample by get the hamming distance between M samples. Specifically, for an image-text triplet $\{o^v_i, o^t_{i_P}, o^t_{i_N}\}$, the goal of attacking cross-modal Hamming retrieval model can be formulated as follows:

$$\begin{aligned} \min_{\eta^v} & D(f^v_{hash}(o^v_i + \eta^v), f^t_{hash}(o^t_{i_N})) - \\ & D(f^v_{hash}(o^v_i + \eta^v), f^t_{hash}(o^t_{i_P})), s.t. \|\eta^v\|_p \leq \epsilon^v. \end{aligned} \tag{6}$$

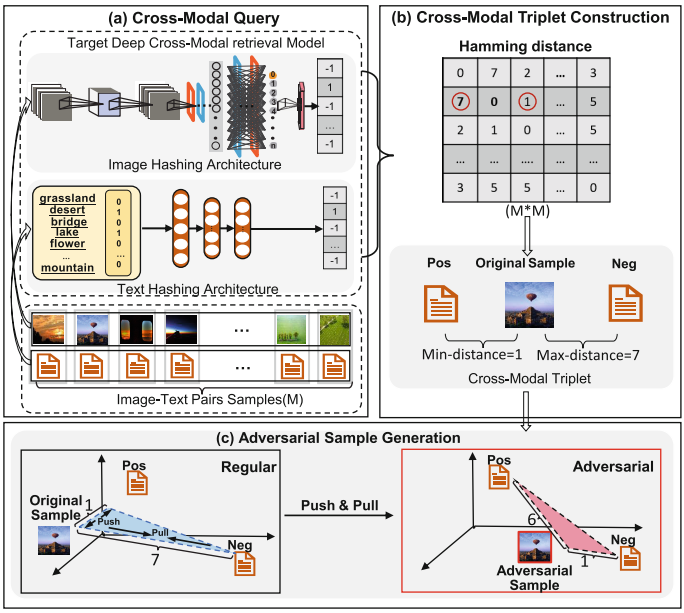

Fig. 2. Overview of BACH

An adversarial image $\hat{o}_i^v = o_i^v + \eta^v$ should satisfy two constraints: 1) the hamming distance between adversarial sample and it's positive sample should be as large as possible, while with negative sample should be as small as possible; Cause the generation of adversarial samples is guided by positive and negative samples to change the pixels of the original image. Specifically, see Fig. 2(c) for example, we continuously push away the hamming distance between the original sample and positive sample (1 to 6), and narrow the hamming distance between the original sample and negative sample (7 to 1), until reaching the preset number of iterations T. The value setting of threshold T is detailed in Sect. 4.2; 2) The perturbation η^v in the attack should be human-imperceptible. To this end, we use $\|\eta^v\|_p \leq \epsilon^v$ to constrain the magnitude of the perturbation. η^v refers to the pixel changes guided by positive and negative samples in this paper. Specifically, let h be the length of perturbation, $\|\cdot\|_p$ be the l_p-norm paradigm, we define the perturbation as $\|\eta\|_p = \sqrt[p]{\frac{1}{h}(|\eta_1|^p + |\eta_2|^p + \cdots + |\eta_h|^p)}$. The dimension of the perturbation is consistent with the raw image in the image dataset. Note that, the l_∞^ϵ bound is the most common way to limit the magnitude of an image, as it strictly limits the maximum image pixel from being perceived. Therefore, we choose the l_∞^ϵ-norm attack in this paper.

However, the optimization problem of Eq. (6) is an NP-hard problem, inspired by the C&W [4] attack, we rewrite the objective loss function as:

$$\begin{aligned} &\min \Gamma(o_i^v, o_{i_P}^t, o_{i_N}^t, \epsilon^v) \\ &= \min \sum_{i=0}^{M} \max(D(\hat{B}_i^v, B_{i_N}^t) - D(\hat{B}_i^v, B_{i_P}^t) + \kappa, 0), s.t. \left\|\eta^v\right\|_\infty \leq \epsilon^v, \end{aligned} \tag{7}$$

$$\begin{aligned} \hat{B}_i^v &= f_{hash}^v \left(o_i^v + \eta^v\right), \\ B_{i_N}^t &= f_{hash}^t \left(o_{i_N}^t\right), \\ B_{i_P}^t &= f_{hash}^t \left(o_{i_P}^t\right), \end{aligned} \tag{8}$$

and $\kappa \geq 0$ is a tuning parameter for attack transferability.

In the white-box setting, the optimization problem of Eq. (7) can be solved by back-propagating the loss function gradient. In the black-box setting, however, we cannot get the network information of target model and only get the model output (i.e., $\mathrm{B}_i^*, * \in \{v, t\}$ in Eq. (8)). Gradient-based estimation is the most effective method in black-box attacks. Inspired by [29], we use the random gradient-free (RGF) method to estimate the gradient of the loss function in Eq. (7), and Ilyas et al. [11] have proved this method is optimal to estimate the gradient. Specifically, the gradient $\frac{\partial \Gamma}{\partial o_i^v}$ (defined as $\hat{g}_i$) of an image sample o_i^v can be estimated by the following equation:

$$\hat{g} = \frac{1}{q} \sum_{i=1}^{q} \hat{g}_i, \text{ with } \hat{g}_i = \frac{f\left(x + \sigma u_i, y\right) - f(x, y)}{\sigma} \cdot u_i, \tag{9}$$

where $\{u_i\}_{i=1}^q$ are the random vectors sampled independently from a uniform distribution $\mathcal{P}$ on $\mathbb{R}^D$, q is the number of the random direction, σ is the sampling variance and D is the dimension of original image. We set $\sigma = 0.01$, and $q = 50$ in this paper. However, it is a box-constraint problem for Eq. (7) that cannot be solved directly based on the commonly-used optimizers. Therefore, we used the following treatment to perturbation ϵ^v to solve this problem:

$$\epsilon^v = \frac{1}{2}(\tanh(\epsilon^v) + 1) - o^v. \tag{10}$$

Then, we choose the Adam [16] optimizer to solve Eq. (7). Finally, we learn the following adversarial perturbations for cross-modal retrieval:

$$\eta^v = \arg\min_{\epsilon^v} \Gamma(o_i^v, o_{i_P}^t, o_{i_N}^t, \epsilon^v). \tag{11}$$

Now that we have detailed the attack method's whole process, specifically, we attack the target retrieval model by inputting the adversarial sample. The entire process of adversarial sample generation is shown in Algorithm 1. Line 1 to Line 4 of the algorithm is the querying part. Line 5 to Line 6 describe the triplet construction. Line 7 to Line 11 illustrate the adversarial sample generation, where Line 9 corresponds to the gradient estimation.

Algorithm 1: Black-box Adversarial Perturbation Generation Method for Deep Cross-modal Hash Retrieval Models (BACH)

Input : Target deep cross-modal retrieval model: $f^*_{hash}(o^*_i)$, $* \in \{v, t\}$, data $O = \{o^v_i, o^t_i\}^C_{i=1}$, iteration T, adversarial queries M

Output: A adversarial sample of query image: $\hat{o}^v_i = o^v_i + \eta^v$

initialize $iter = 0$;
Random select query data $\{o^v_i, o^t_i\}^M_{i=1}$;
Compute $B^t = sign\left(f^t_{hash}\left(O^t_q\right)\right)$;
Compute $B^v = sign\left(f^v_{hash}\left(O^v_q\right)\right)$;
Compute Hamming distance matrix according to Equation (3) based on $\{B^v, B^t\} = \{B^v_i, B^t_i\}^M_{i=1}$;
Create cross-modal triplets $\{o^v_i, o^t_{i_P}, o^t_{i_N}\}$ for every image o^v_i;
Select η^v:**while** $iter \leq T$ **do**
    $\eta^v = \arg\min_{\epsilon^v} \Gamma(o^v_i, o^t_{i_P}, o^t_{i_N}, \epsilon^v)$;
    Estimate $\hat{g}_i$ using Equation (9);
    Using Adam optimizer;
    $iter = iter + 1$;
return $\hat{o}^v_i$;

4 Experiment

This section evaluates the performance of BACH on several commonly-used deep cross-modal hamming retrieval models and dataset. We assess the attack on image-to-text retrieval task and image-to-image retrieval task.

4.1 Dataset

The dataset of image-to-text retrieval task include MIRFlickr-25K and NUS-WIDE. The dataset of image-to-image retrieval task include CIFAR10 and NUS-WIDE. We use these three dataset to train several deep cross-modal hamming retrieval models. In all of our attack experiments below, dataset are divided into three-part, including train, query, and gallery parts. Note that the attacks does not use the train set.

MIRFlickr-25K contains 25,000 images from the Flickr website, each image with a corresponding text description constituting an image-text pair. According to [42], we randomly divided the dataset into a training dataset with 5000 samples and a test dataset with 20000 samples. There are M samples as query dataset in the test dataset, while remaining samples as a gallery set.

NUS-WIDE is a multi-label dataset containing 81 labels. There are 269,648 image-text pairs. We select a total of 195834 samples from the most commonly used 21 labels as the image retrieval dataset according to [12]. We select 500

Table 1. The attack performance in term of mAP on the state-of-the-art DCMHR models for image-to-text retrieval task, based on MIRFlickr-25K and NUS-WIDE sets

Dataset	**MIRFLICKR-25k**								**NUS-WIDE**							
CL	16		32		64		128		16		32		64		128	
Method	REG	ATK	REG	ATK	REG	ATK	REG	ATK	REG	ATK	REG	ATK	REG	ATK	REG	ATK
DJSRH	0.66	0.60	0.66	0.61	0.67	0.61	0.68	0.62	0.46	0.41	0.49	0.43	0.45	0.41	0.53	0.44
AGAH	0.74	0.59	0.77	0.60	0.78	0.62	0.77	0.63	0.62	0.43	0.64	0.45	0.65	0.46	0.65	0.46
SSAH	0.64	0.60	0.68	0.61	0.69	0.61	0.71	0.61	0.48	0.41	0.51	0.41	0.53	0.41	0.53	0.42
DCMH	0.71	0.62	0.74	0.62	0.72	0.63	0.73	0.64	0.64	0.43	0.66	0.44	0.65	0.46	0.67	0.46
DSAH	0.69	0.60	0.70	0.60	0.71	0.61	0.71	0.61	0.56	0.42	0.60	0.42	0.61	0.43	0.62	0.44

[a] REG is the abbreviation of regular that used to represent regular retrieval performance, and ATK is the abbreviation of attack that used to represent attack performance.
[b] CL refers to code length. Here, M is set to 500 and T is set to 800.

pairs for each label to construct the training dataset randomly, with 100 pairs of each label randomly selected to query, and the rest are used as the gallery dataset. In addition, this paper uses NUS-WIDE as a dataset for attacking the deep cross-modal hamming retrieval models for image-to-image retrieval task. Following [8], a total of 5000 samples are selected randomly as the query dataset and the remaining samples as a gallery set.

CIFAR10 dataset consists of 60,000 images whose sizes are 32×32 and belong to 10 categories. Each category has 6,000 images. There are 50,000 training images and 10,000 testing images. We extract 100 samples for each category from the testing dataset for querying, and the remaining samples are as a gallery set.

To evaluate BACH, we use two commonly used evaluation criteria for cross-modal retrieval tasks in the field of information retrieval, namely, mean Average Precision (mAP) and Normalized Discounted Cumulative Gain (NDCG).

Table 2. The attack performance in term of NDCG on the state-of-the-art DCMHR models for image-to-text retrieval task, based on MIRFlickr-25K and NUS-WIDE sets

Dataset	**MIRFLICKR-25k**								**NUS-WIDE**							
CL	16		32		64		128		16		32		64		128	
Method	REG	ATK	REG	ATK	REG	ATK	REG	ATK	REG	ATK	REG	ATK	REG	ATK	REG	ATK
DJSRH	0.63	0.59	0.66	0.60	0.66	0.61	0.68	0.62	0.48	0.41	0.49	0.42	0.53	0.43	0.54	0.44
AGAH	0.76	0.61	0.79	0.62	0.80	0.62	0.81	0.63	0.64	0.44	0.67	0.44	0.68	0.45	0.68	0.46
SSAH	0.64	0.59	0.67	0.60	0.67	0.52	0.69	0.62	0.49	0.42	0.50	0.42	0.53	0.43	0.54	0.44
DCMH	0.75	0.62	0.75	0.62	0.76	0.63	0.77	0.63	0.63	0.43	0.64	0.44	0.65	0.44	0.66	0.45
DSAH	0.68	0.59	0.70	0.60	0.71	0.61	0.72	0.62	0.58	0.42	0.60	0.42	0.61	0.43	0.61	0.44

[a] CL refers to code length. Here, M is set to 500 and T is set to 800.

4.2 Evaluation

BACH is a black-box adversarial attack on DCMHR models. To evaluate the effectiveness and efficiency of BACH, we design experiments to answer the following three research questions:

- **RQ1:** Is BACH effective to attack classical deep cross-modal hamming retrieval models for image-to-text and image-to-image retrieval tasks?
- **RQ2:** Does the number of samples of the query dataset used to construct triples affect the attack performance?
- **RQ3:** How does BACH perform compared with existing white-box and black-box attacking methods?

Table 3. The attack performance in term of mAP on the state-of-the-art DCMHR models for image-to-image retrieval task, based on CIFAR10 set

Dataset	**CIFAR10**							
Code Length	12		24		36		48	
Method	REG	ATK	REG	ATK	REG	ATK	REG	ATK
SDH	0.46	0.10	0.64	0.11	0.66	0.12	0.67	0.14
DSH	0.62	0.10	0.66	0.13	0.67	0.14	0.68	0.14
ADSH	0.88	0.15	0.88	0.15	0.87	0.15	0.87	0.15
DSDH	0.73	0.14	0.75	0.15	0.75	0.15	0.75	0.16

[a] Here, M is set to 500, T is set to 800.

Firstly, we show the performance of BACH on several retrieval models. To verify the ability of the attack, we re-produce 7 state-of-the-art DCMHR models for image-to-text retrieval task, including DJSRH [32], AGAH [28], SSAH [18], DCMH [7], DSAH [22], PRDH [38] and CMHH [2]. We construct six state-of-the-art DCMHR models for image-to-image data retrieval task, including DSH [24], DIHN [35], DSDH [21], ADSH [15], HMH [39] and SDH [30]. In addition, to evaluate the attack performance of the adversarial samples with different hash code lengths, for the cross-modal retrieval task, we use 16, 32, 64, and 128 as the length, respectively. Moreover, for image-to-image retrieval tasks, there are two dataset, where the CIFAR10 dataset takes values of 12, 24, 36, and 48 for hash code length, and the NUS-WIDE dataset takes values of 8, 16, 24, and 32 for hash code length. We attacked the above retrieval models, and the comparison between the regular retrieval performance and the performance after being attacked in terms of the mAP/NDCG score. The attack results on the DCMHR models for image-to-text retrieval task are shown in Tables 1 and 2. The attack results of image-to-image retrieval task are in Tables 3 and 4. Furthermore, due to the special requirements of the HMH method for hash code length, 8-bit and 24-bit hash code lengths do not satisfy the HMH requirements, so we do not perform 8-bit and 24-bit attacks on CIFAR10 and NUS-WIDE for HMH. BACH produces adversarial samples that effectively degrade the performance of all the above well-trained retrieval models, which means that all our attacks are successful, demonstrating the lack of robustness of these existing deep retrieval models to small adversarial perturbations.

To construct a triplet of samples, we need to query the hash codes of M samples. Different M will led to different attack performances, so we will take M as 200, 300, 500, and 1000 respectively, to perform the attack. The comparison of the attack performance on MIRFlickr-25K, NUS-WIDE is shown

Table 4. The attack performance in terms of mAP and NDCG on the state-of-the-art DCMHR models for image-to-image retrieval task, based on NUS-WIDE set

Dataset	NUS-WIDE (mAP)								NUS-WIDE (NDCG)							
CL	8		16		24		32		8		16		24		32	
Method	REG	ATK	REG	ATK	REG	ATK	REG	ATK	REG	ATK	REG	ATK	REG	ATK	REG	ATK
DSH	0.66	0.24	0.69	0.26	0.70	0.27	0.71	0.27	0.45	0.26	0.45	0.26	0.45	0.27	0.45	0.27
ADSH	0.80	0.27	0.85	0.28	0.86	0.29	0.87	0.29	0.51	0.28	0.59	0.28	0.61	0.28	0.63	0.29
DIHN	0.74	0.25	0.79	0.27	0.81	0.27	0.80	0.27	0.48	0.26	0.51	0.27	0.58	0.27	0.58	0.29
DSDH	0.77	0.26	0.76	0.26	0.80	0.27	0.80	0.28	0.50	0.28	0.55	0.28	0.59	0.30	0.59	0.23
HMH	-	-	0.74	0.26	-	-	0.78	0.27	-	-	0.53	0.28	-	-	0.52	0.28

[a] CL refers to code length. Here, M is set to 500 and T is set to 800.

Table 5. Comparison of the attack performance for different Adversarial Queries (M) in terms of mAP scores, the code length is set to 32 bits, T is set to 800

Tasks	Adversarial Queries		MIRFlickr-25K				NUS-WIDE			
			DCMH	PRDH	SSAH	CMHH	DCMH	PRDH	SSAH	CMHH
$I \rightarrow T$	REG		0.74	0.78	0.68	0.75	0.66	0.64	0.51	0.60
	BACH	200	0.70	0.68	0.66	0.64	0.59	0.52	0.44	0.46
		300	0.65	0.64	0.64	0.63	0.50	0.46	0.42	0.44
		500	0.62	0.60	0.61	0.61	0.44	0.42	0.41	0.43
		1000	0.62	0.61	0.62	0.61	0.45	0.41	0.42	0.43

[a] $I \rightarrow T$ denotes retrieval text using an adversarial image query.

in Table 5 (mAP). The mAP scores decreases as M increases, so the attack performance gradually improves. However, we find that the attack performance slightly decreases when the M increases from 500 to 1000, which may be due to some inaccurate information obtained when querying the target model, so high-quality query samples will help to improve the query efficiency and attack performance.

Meanwhile, we compare the impact of different iterative numbers, T, on the attack performance during adversarial sample generation. Here we fix M to 500, and the attack performance comparison on the baseline databases is shown in Table 6. We find the retrieval performance degrades gradually as the number of iterations becomes larger, meaning the attack performance becomes better. However, when T grows from 500 to 800, the attack performance increase is insignificant. Since there is often a limit on the number of queries, we consider T takes 800 as the optimal value.

Last, we compare BACH performance to white-box and black-box attack methods, and the results are shown in Table 7. CMLA [20] is a work to attack DCMHR models in a white-box setting. In contrast, AACH [19] attacks DCMHR models in a black-box setting. Therefore, AACH does not require a priori knowledge, such as the structure of the target network. However, AACH requires constructing a surrogate model, which we do not need. We attack by directly estimating the gradient of the loss function. We compare the attack performance of the three methods on top of two different dataset according to [19]. CMLA achieves the best performance, which is attributed to the fact that CMLA has

Table 6. Comparison of the attack performance for different iteration (T) in terms of mAP scores, the code length is set to 32 bits, M is set to 500.

Tasks	Iteration		MIRFlickr-25K				NUS-WIDE			
			DCMH	PRDH	SSAH	CMHH	DCMH	PRDH	SSAH	CMHH
$I \rightarrow T$	REG		0.74	0.78	0.68	0.75	0.66	0.64	0.51	0.60
	BACH	300	0.66	0.63	0.60	0.69	0.46	0.50	0.44	0.55
		500	0.63	0.60	0.62	0.62	0.44	0.45	0.41	0.45
		800	0.62	0.60	0.61	0.61	0.44	0.42	0.41	0.43

[a] $I \rightarrow T$ denotes retrieval text using an adversarial image query.

Table 7. Comparison of the attack performance of BACH, CMLA and AACH in terms of mAP scores on different dataset, the code length is 32 bits.

Tasks	Methods	MIRFlickr-25K				NUS-WIDE			
		DCMH	PRDH	SSAH	CMHH	DCMH	PRDH	SSAH	CMHH
$I \rightarrow T$	REG	0.74	0.78	0.68	0.75	0.66	0.64	0.51	0.60
	CMLA	0.52	0.60	0.60	0.56	0.46	0.40	0.36	0.33
	AACH	0.63	0.62	0.56	0.65	0.44	0.50	0.40	0.41
	BACH	0.62	0.61	0.61	0.58	0.44	0.49	0.41	0.40

[a] $I \rightarrow T$ denotes retrieval text using an adversarial image query.

all the prior knowledge of the target model as a white-box attack. However, the attack performance of our BACH is more potent than AACH on both benchmark dataset. It validates the effectiveness of our approach.

5 Related Work

5.1 Deep Cross-Modal Hashing

In order to measure the semantic similarity between samples of different modalities and maintain the similarity between data samples, the features of data samples belonging to different modalities are often mapped into a common subspace. As shown in Fig. 3, Hash codes learning and retrieval tasks are all based on this common subspace. For example, Inter-Media Hashing [31] uses inter-modal and intra-modal consistency as benchmarks to construct a common Hamming space and introduces regularized linear regression into the hashing process. Latent Semantic Sparse Hashing [43] the latent space through sparse coding and matrix factorization and then fuses the features of different modal data into a unified hash code. The Composite Correlation Quantization [27] uses the maximum mapping method to construct a common subspace. Collective Matrix Factorization Hashing [5] and Supervised Collective Matrix Factorization Hashing [6] exploit collaborative matrix factorization to learn hash codes from different modalities. Based on common subspace learning, adding label information can effectively improve the performance of cross-modal retrieval models, and such supervised models can generate hash codes that preserve semantics.

The Semantic Correlation Maximization [42] method proposed earlier attempts to integrate label information into the learning process to obtain a similarity matrix. Semantics Preserving Hashing [23] first used a distribution function to formulate the hashing process, converted the semantic relationship contained in the label information into a probability distribution, then trained the model by minimizing the Kullback-Leibler divergence.

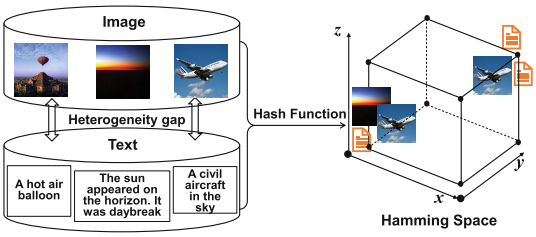

Fig. 3. Regular cross-modal Hamming retrieval

Meanwhile, DNN can enhance the feature learning capability for different modal data, which yields deep cross-modal hashing using DNN as a feature extraction network. Typical approaches are Deep Cross-Modal Hashing [14] and Pairwise Relation Guided Deep Hashing [38], both of which use deep convolutional networks and fully connected networks to extract the image and text features, respectively, while adding semantic labeling information to maintain the original semantic similarity between samples. The Deep Visual-Semantic Hashing [3] method further uses Long Short Term Memory (LSTM) [10] to learn textual information in the form of sentences. Self-Supervised Adversarial Hashing [18] proposes to capture semantic features from different modalities further using generative adversarial networks and proposes labeling networks to generate hash codes of label vectors.

5.2 Adversarial Attacks

Szegedy et al. were the first to propose the concept of adversarial sample [5], and they found that small perturbations that are not sensitive to the human visual system can make the neural network too sensitive to produce false recognition. Subsequent researchers have proposed many more powerful and effective methods for attack generation. The existing adversarial attacks can be divided into two main categories: white-box attacks and black-box attacks. White-box attacks refer to the information of the target model is fully accessible, and the most commonly used white-box attacks are fast gradient symbolic method (FGSM) [8] and projected gradient descent method (PGD) [23]. Although the performance of white-box attacks is relatively high, obtaining specific information about the target model in the real world is complicated. The black-box attack can only obtain the model's output or even the information about the model is completely

unknown. This setting increases the difficulty of attacks, but it is more practical than white-box. Research shows that, black-box attacks based on gradient estimation are already close to the performance of the best white-box attacks.

6 Conclusion

This paper presents BACH, a learning-based adversarial attack method aimed at fooling deep cross-modal retrieval models on hamming space in a black-box setting. BACH consists of three parts: first, it calculates the hamming distance between samples through cross-modal querying; second, it constructs cross-modal triplets (i.e., original sample, positive sample, and negative sample) for each image based on the hamming distance; and third, it learns to generate adversarial samples by pulling the negative samples close and pushing away the positive sample, using a random gradient-free gradient estimation method to reduce the number of queries. BACH was tested on 3 popular dataset and 13 state-of-the-art deep cross-modal hamming retrieval models, including 6 models for image-to-image retrieval and 7 models for image-to-text retrieval. The experiments show that BACH can effectively attack existing retrieval models and has comparable attack performance to the white-box attack method (i.e., CMLA) and the black-box attack method (i.e., AACH). The results highlight the unreliability of current cross-modal hamming retrieval models, as well-designed perturbations can easily mislead them in practice. Thus, BACH can serve as a baseline for evaluating the robustness of cross-modal hamming retrieval models, and call for advanced method to enhance the robustness of cross-modal retrieval models in the future.

Acknowledgements. This paper was supported by the Ministry of Science and Technology of China under Grant No. 2020AAA0108401, and the Natural Science Foundation of China under Grant Nos. 72225011 and 71621002.

References

1. Andoni, A., Indyk, P.: Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In: 2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2006), pp. 459–468. IEEE (2006)
2. Cao, Y., Liu, B., Long, M., Wang, J.: Cross-modal hamming hashing. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11205, pp. 207–223. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01246-5_13
3. Cao, Y., Long, M., Wang, J., Yang, Q., Yu, P.S.: Deep visual-semantic hashing for cross-modal retrieval. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1445–1454 (2016)
4. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP), pp. 39–57. IEEE (2017)
5. Ding, G., Guo, Y., Zhou, J.: Collective matrix factorization hashing for multimodal data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2075–2082 (2014)

6. Ding, G., Guo, Y., Zhou, J., Gao, Y.: Large-scale cross-modality search via collective matrix factorization hashing. IEEE Trans. Image Process. **25**(11), 5427–5440 (2016)
7. Gong, Y., Lazebnik, S., Gordo, A., Perronnin, F.: Iterative quantization: a procrustean approach to learning binary codes for large-scale image retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **35**(12), 2916–2929 (2012)
8. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
9. Gu, W., Gu, X., Gu, J., Li, B., Xiong, Z., Wang, W.: Adversary guided asymmetric hashing for cross-modal retrieval. In: Proceedings of the 2019 on International Conference on Multimedia Retrieval, pp. 159–167 (2019)
10. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
11. Ilyas, A., Engstrom, L., Madry, A.: Prior convictions: black-box adversarial attacks with bandits and priors. In: International Conference on Learning Representations (2018)
12. Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., Madry, A.: Adversarial examples are not bugs, they are features. In; Advances in Neural Information Processing Systems, vol. 32 (2019)
13. Indyk, P., Motwani, R.: Approximate nearest neighbors: towards removing the curse of dimensionality. In: Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing, pp. 604–613 (1998)
14. Jiang, Q.Y., Li, W.J.: Deep cross-modal hashing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3232–3240 (2017)
15. Jiang, Q.Y., Li, W.J.: Asymmetric deep supervised hashing. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)
16. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
17. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
18. Li, C., Deng, C., Li, N., Liu, W., Gao, X., Tao, D.: Self-supervised adversarial hashing networks for cross-modal retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4242–4251 (2018)
19. Li, C., Gao, S., Deng, C., Liu, W., Huang, H.: Adversarial attack on deep cross-modal hamming retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2218–2227 (2021)
20. Li, C., Gao, S., Deng, C., Xie, D., Liu, W.: Cross-modal learning with adversarial samples. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
21. Li, Q., Sun, Z., He, R., Tan, T.: Deep supervised discrete hashing. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
22. Li, Y., van Gemert, J.: Deep unsupervised image hashing by maximizing bit entropy. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 2002–2010 (2021)
23. Lin, Z., Ding, G., Hu, M., Wang, J.: Semantics-preserving hashing for cross-view retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3864–3872 (2015)
24. Liu, H., Wang, R., Shan, S., Chen, X.: Deep supervised hashing for fast image retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2064–2072 (2016)
25. Liu, J., Xu, C., Lu, H.: Cross-media retrieval: state-of-the-art and open issues. Int. J. Multimedia Intell. Secur. **1**(1), 33–52 (2010)

26. Liu, X., Huang, L., Deng, C., Lang, B., Tao, D.: Query-adaptive hash code ranking for large-scale multi-view visual search. IEEE Trans. Image Process. **25**(10), 4514–4524 (2016)
27. Long, M., Cao, Y., Wang, J., Yu, P.S.: Composite correlation quantization for efficient multimodal retrieval. In: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, pp. 579–588 (2016)
28. Nakkiran, P.: Adversarial robustness may be at odds with simplicity. arXiv preprint arXiv:1901.00532 (2019)
29. Nesterov, Y., Spokoiny, V.: Random gradient-free minimization of convex functions. Found. Comput. Math. **17**(2), 527–566 (2017)
30. Shen, F., Shen, C., Liu, W., Tao Shen, H.: Supervised discrete hashing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 37–45 (2015)
31. Song, J., Yang, Y., Yang, Y., Huang, Z., Shen, H.T.: Inter-media hashing for large-scale retrieval from heterogeneous data sources. In: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, pp. 785–796 (2013)
32. Su, S., Zhong, Z., Zhang, C.: Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3027–3035 (2019)
33. Sun, Y., Chen, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. In: Advances in Neural Information Processing Systems, vol. 27 (2014)
34. Szegedy, C., et al.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
35. Wu, D., Dai, Q., Liu, J., Li, B., Wang, W.: Deep incremental hashing network for efficient image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9069–9077 (2019)
36. Wu, Y., et al.: Google's neural machine translation system: bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016)
37. Xu, C., Tao, D., Xu, C.: A survey on multi-view learning. arXiv preprint arXiv:1304.5634 (2013)
38. Yang, E., Deng, C., Liu, W., Liu, X., Tao, D., Gao, X.: Pairwise relationship guided deep hashing for cross-modal retrieval. In: proceedings of the AAAI Conference on Artificial Intelligence, vol. 31 (2017)
39. Yuan, L., et al.: Central similarity quantization for efficient image and video retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3083–3092 (2020)
40. Yuan, X., He, P., Zhu, Q., Li, X.: Adversarial examples: attacks and defenses for deep learning. IEEE Trans. Neural Netw. Learn. Syst. **30**(9), 2805–2824 (2019)
41. Zhai, X., Peng, Y., Xiao, J.: Heterogeneous metric learning with joint graph regularization for cross-media retrieval. In: Twenty-Seventh AAAI Conference on Artificial Intelligence (2013)
42. Zhang, D., Li, W.J.: Large-scale supervised multimodal hashing with semantic correlation maximization. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 28 (2014)
43. Zhou, J., Ding, G., Guo, Y.: Latent semantic sparse hashing for cross-modal similarity search. In: Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 415–424 (2014)