Adversarial Attack on Deep Cross-Modal Hamming Retrieval

Chao Li[1,*] Shangqian Gao[2,*] Cheng Deng[1,†] Wei Liu[3] Heng Huang[2,4]
[1]Xidian University [2]University of Pittsburgh [3]Tencent Data Platform [4]JD Explore Academy
{chaolee.xd, chdeng.xd, henghuanghh}@gmail.com, shg84@pitt.edu, wl2223@columbia.edu

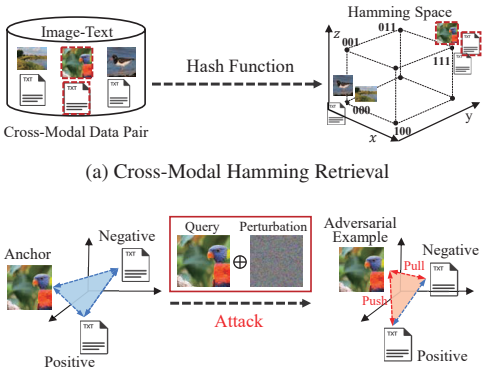
Abstract

Recently, Cross-Modal Hamming space Retrieval (CMHR) regains ever-increasing attention, mainly benefiting from the excellent representation capability of deep neural networks. On the other hand, the vulnerability of deep networks exposes a deep cross-modal retrieval system to various safety risks (e.g., adversarial attack). However, attacking deep cross-modal Hamming retrieval remains underexplored. In this paper, we propose an effective Adversarial Attack on Deep Cross-Modal Hamming Retrieval, dubbed AACH, which fools a target deep CMHR model in a black-box setting. Specifically, given a target model, we first construct its substitute model to exploit cross-modal correlations within hamming space, with which we create adversarial examples by limitedly querying from a target model. Furthermore, to enhance the efficiency of adversarial attacks, we design a triplet construction module to exploit cross-modal positive and negative instances. In this way, perturbations can be learned to fool the target model through pulling perturbed examples far away from the positive instances whereas pushing them close to the negative ones. Extensive experiments on three widely used cross-modal (image and text) retrieval benchmarks demonstrate the superiority of the proposed AACH. We find that AACH can successfully attack a given target deep CMHR model with fewer interactions, and that its performance is on par with previous state-of-the-art attacks.

1. Introduction

Deep Neural Networks (DNNs) have been widely adopted to improve the retrieval performance in CMHR, where the early network layers capture the implicit structure of cross-modal data, and binary codes are derived from a deeper network layer. Generally, the DNN architecture is trained to build the cross-modal correlations by detecting the semantic similarity or dissimilarity between different modalities. Inspired by such superior representation capability of DNN, many efforts have been focused on employing deep networks to enhance the correlations between modalities through learning a common representation in shared space. However, the robustness and stability of DNN structures have been largely overlooked: even the most accurate deep learning models can be easily deceived by a well-designed perturbation which is visually imperceptible to the human eye. Therefore, the growing costs and risks of the potential model failures have led to the study of adversarial attacks. In this paper we focus on a practical cross-modal Hamming adversarial attack that fulfills two criteria: 1) the attack is designed for a black-box setting, where the target cross-modal network is normally unavailable, and the attacker can only interact with the target model by querying it; 2) the query efficiency should be highly prioritized considering the practical case, that is, frequent and high volume queries will be easily discovered by defenders.

Despite plenty of adversarial attacks proposed in the literature, the main attention only focuses on the problem of adversarial examples learning for image-based classification or retrieval within a single modality. Little effort has been devoted to investigating how adversarial examples affect deep Hamming learning in cross-modal retrieval. There exist great differences in learning adversaries between the



(a) Cross-Modal Hamming Retrieval

(b) Attack on Cross-Modal Hamming Retrieval

Figure 1: Regular cross-modal Hamming retrieval and our triplet-based cross-modal Hamming attack.

*Equal contribution.
†Corresponding author.

existing classification task and CMHR. As shown in Fig. 1a, given a query instance from one modality (e.g., image), CMHR is applied to map original data into binary codes, and then execute bit-wise XOR operation to return semantically related instances from another modality (e.g., text). In contrast, the attack on CMHR devotes to fooling a well-trained model to return semantically unrelated instances. Therefore, the traditional classification-oriented adversarial attacks are not suitable in CMHR. The pioneering work CMLA [17] is the first attempt to design adversarial samples to deceive a target deep CMHR model. But, CMLA is not applicable to the practical cases in two main aspects. First, CMLA is constructed for a white-box setting, where attackers need full knowledge of the target model, including model architectures and parameters. Second, the label information of query instances is required in CMLA, which is also not available in reality. Therefore, practical adversary example learning in CMHR is still an open problem. Actually, considering the nature of CMHR is to explore and preserve the similar semantic structure among instances, we can design the triplet-based cross-modal Hamming attack as shown in Fig. 1b. The adversarial perturbation is learned and added into the query instance to manipulate the original similarity structure, reducing the distance of the query to the negative instance and enlarging its distance to the positive instance.

In this paper, we propose an effective Adversarial Attack on Deep Cross-Modal Hamming Retrieval (AACH). To be specific, AACH mainly focuses on attacking a deep cross-modal Hamming retrieval model in a black-box setting, which thus is more applicable to practical cases. In addition, to reduce the cost and risk during querying a target model, we propose the cross-modal triplet construction module, where cross-modal positive and negative instances of the query are exploited to boost the learning of adversarial perturbations. We highlight the contributions of this work as follows:

- An adversarial example learning method for cross-modal Hamming retrieval is proposed under the black-box setting. Through constructing a surrogate model of the target networks, the proposed AACH learns cross-modal adversarial examples only by limitedly querying the target model, without any prior knowledge about the target model.
- To fully take advantage of the limited information acquired from target model, a simple yet effective cross-modal triplet construction module is designed, with which our surrogate model learns adversarial examples by mining cross-modal positive and negative instances in Hamming space.
- We evaluate the proposed AACH by attacking several state-of-the-art cross-modal retrieval models on three popular benchmarks, MIRFlickr-25K, NUS-WIDE, and MS COCO. Extensive results demonstrate the effectiveness of the proposed triplet construction module and the capacity of our AACH in attacking a target deep CMHR model.

2. Related Works

Deep Cross-Modal Hamming Retrieval. Different from the traditional "learn to hash" for CMHR [19, 7, 25, 39]. Deep CMHR [12, 37, 3, 32, 6, 15, 2, 35, 11, 4] learns binary codes by constructing deep networks to build the correlations across modalities. Existing approaches can be divided into unsupervised and supervised settings according to whether label information is used. For unsupervised setting, efforts are devoted to studying the semantic similarity exploration and preservation. Matrix factorization and graph Laplacian are proposed to preserve neighborhood structures of original data in [33]. Su *et al.* [29] explored the joint-semantics similarity matrix from different modalities to integrate multiple modality similarity information. In [40] and [16], the generative adversarial networks (GAN) are constructed to bridge the semantic modality gap. On the contrary, methods in a supervised setting generally build cross-modal correlations from the label information, where pairwise [12, 37, 2], triplet [6], and ranking [20] semantic constraints are respectively adopted to achieve high retrieval performance. Li *et al.* [15] constructed self-supervised learning model to guide deep cross-modal network training. To enhance the semantic similarities, an attention mechanism is integrated into one GAN-based cross-modal network [41] to extract the shared semantic components across modalities.

Cross-Modal Hamming Attacking. The successes achieved in deep learning areas have made great improvements for CMHR. Nevertheless, the vulnerability of DNN, which has caused wide concern from all walks of life, places the DNN-based retrieval model at the risk of being attacked as well. In [31], it is the first time showing that a deep network with good performance can be fooled by a well-designed perturbation which is imperceptible to human eyes. Follow this, many white-box attacking methods are presented [23, 24, 36, 22, 8, 30, 28], where [28] can successfully attack a target deep model in an extremely limited scenario because only one pixel can be modified. Simultaneously, the transferability of the adversarial examples among deep networks is discovered to propose the black-box setting attacks [21, 34, 26, 9, 5, 10], which are sharing a common idea that using an approximate gradient to create adversarial examples. Most of these methods are designed to solve the problem of image-based classification or retrieval within single a modality [38, 13]. For CMHR task, the risk of malicious users disrupting the retrieval system always exists. However, few efforts focus on the security

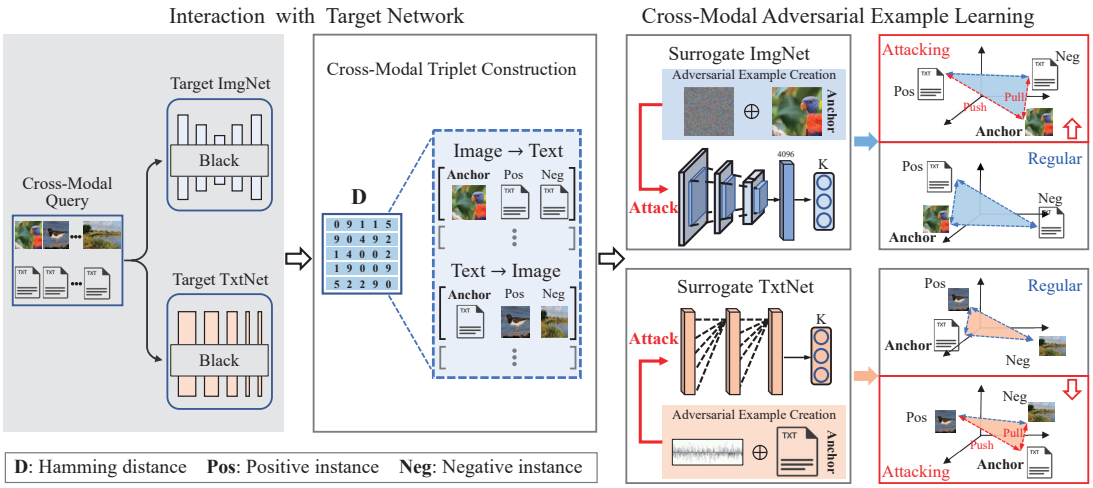


Figure 2: The pipeline of our proposed AACH for cross-modal Hamming learning.

of the DNN-based CMHR system. The learning of cross-modal adversarial example is first presented in CMLA [17], where the intra- and inter- modality similarity are explored to learn the adversarial examples preserving the intra-modal similarity but manipulating inter-modal correlations. As mentioned above, CMLA is designed for a white-box attack setting and requires label information about query instance, making it not applicable in reality.

3. Adversarial Attack on Deep CMHR

3.1. Problem Formulation

Generally, the CMHR can be cast as a metric learning problem. Given a cross-modal dataset $O = \{o_i\}_{i=1}^{M}$ with M instances, $o_i = \{o_i^v, o_i^t\}$, where o^v and o^t respectively represent two data modalities (*e.g.,* image-text pairs). The goal of deep cross-modal Hamming retrieval is to learn two functions $\mathcal{F}_{tar}^{*}(o^*;\theta^*)$ by different networks, $* \in \{v,t\}$ projecting the original modality instances onto Hamming space, where θ^* denotes parameter to be learned. As such the original cross-modal instances are represented with binary codes $B^* \in \{-1,1\}^K$, K denotes the required code length. This can be formulated as follows:

$$B^* = sign(H^*),\ H^* = \mathcal{F}_{tar}^{*}(o^*;\theta^*),\ * \in \{v,t\}, \quad (1)$$

where H^* are binary-like representations ($H^* \in [-1,1]^K$) produced by the output layer of a target deep cross-modal network.

A well-trained $\mathcal{F}_{tar}^{*}(o^*)$ can always preserve accurate similar semantic structure. To be specific, assuming that an image query instance o_A^v is more similar with positive instance o_P^t than negative instance o_N^t, it encourages $\mathcal{F}_{tar}^{*}$ to satisfy the inequality as follow:

$$D(\mathcal{F}_{tar}^{v}(o_A^v), \mathcal{F}_{tar}^{t}(o_P^t)) < D(\mathcal{F}_{tar}^{v}(o_A^v), \mathcal{F}_{tar}^{t}(o_N^t)), \quad (2)$$

where D denotes the Hamming distance between two codes:

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

X and Y are the input codes. For cross-modal Hamming attacking, on the contrary, which aims to learn cross-modal adversarial perturbation δ^v to fool the target deep network to output binary codes with a contrary inequality as follows:

$$D(\mathcal{F}_{tar}^{v}(o_A^v+\delta^v), \mathcal{F}_{tar}^{t}(o_P^t)) > D(\mathcal{F}_{tar}^{v}(o_A^v+\delta^v), \mathcal{F}_{tar}^{t}(o_N^t)). \quad (4)$$

Formally, given an image-text triplet $\{o_A^v, o_P^t, o_N^t\}$, where o_A^v and o_P^t share similar semantic correlation, we rewrite the cross-modal Hamming attacking as follows:

$$\min_{\delta^v} D\left(\mathcal{F}_{tar}^{v}(o_A^v+\delta^v), \mathcal{F}_{tar}^{t}(o_N^t)\right) - D\left(\mathcal{F}_{tar}^{v}(o_A^v+\delta^v), \mathcal{F}_{tar}^{t}(o_P^t)\right),\ s.t.\ \|\delta^v\|_p \le \epsilon^v, \quad (5)$$

where $\|\cdot\|_p$ denotes L_p norm ($p = \infty$ in this work). ϵ^v denotes the attack strength, where the constraint $\|\delta^v\|_p \le \epsilon^v$ limits the perturbation δ^v being visually imperceptible. The same goes for learning the text perturbation δ^t to query image:

$$\min_{\delta^t} D\left(\mathcal{F}_{tar}^{t}(o_A^t+\delta^t), \mathcal{F}_{tar}^{v}(o_N^v)\right) - D\left(\mathcal{F}_{tar}^{t}(o_A^t+\delta^t), \mathcal{F}_{tar}^{v}(o_P^v)\right),\ s.t.\ \|\delta^t\|_p \le \epsilon^t. \quad (6)$$

3.2. Proposed AACH

Fig. 2 shows the full flowchart of our AACH, which mainly consists of three parts: target deep cross-modal networks (ImgNet and TxtNet), triplet construction module, and cross-modal adversarial example learning.

The target deep cross-modal networks are always supposed to be well-trained and thus can produce reliable binary codes. Under a black-box attack, we can only make interaction with the target networks by inputting M cross-modal data pairs $\{o^v, o^t\}^M$ as queries to the target networks. Generally, the number of M is highly limited. In this way, their corresponding binary codes $\{B^v, B^t\}$ are obtained, and then we calculate the Hamming distance $D(B^v, B^t)$.

Next, we take each instance o_A^v (o_A^t) in individual modality as anchor to respectively select its positive cross-modal instances o_P^t (o_P^v) with shorter Hamming distance and negative ones o_N^t (o_N^v) with longer Hamming distance. We can assign multiple positive instances or negative instances to an anchor. In doing so, the cross-modal triplets $\{o_A^v, o_P^t, o_N^t\}$ ($\{o_A^t, o_P^v, o_N^v\}$) are created, which will be used to train the surrogate deep cross-modal networks and learn cross-modal adversarial examples.

We construct the surrogate deep cross-modal networks with commonly used convolutional neural networks (θ_{sur}^v) for image modality and fully connected network (θ_{sur}^t) for text modality, respectively. More details about network structures are provided in the implementation details (Section 4.3). Taking the image-query-text task as an example, to train the surrogate deep cross-modal networks, we design the triplet loss as follows:

$$\mathcal{L}_{tri}^v = \sum^M max\left(D(H_A^v, H_P^t) - D(H_A^v, H_N^t) + \beta,\ 0\right), \tag{7}$$

where D denotes Hamming distance calculated by Eq. 3, and β is a manually defined constant margin. Considering the binary-like presentations of H_A^v, H_P^t, and H_N^v, to decrease the quantization error between binary-like presentations and binary codes, we design the quantization loss as follows:

$$\begin{aligned}\mathcal{L}_{qua}^v = \sum^M \Big(&\|H_A^v - B_A^v\|_2^2 + \|H_P^v - B_P^v\|_2^2 \\ &+\|H_N^v - B_N^v\|_2^2\Big).\end{aligned} \tag{8}$$

Therefore, the total loss to train the surrogate image network is the sum of the triplet loss and the quantization loss, $\mathcal{L}^v = \mathcal{L}_{tri}^v + \mathcal{L}_{qua}^v$. Similarly, we can obtain the triplet loss for text network $\mathcal{L}^t = \mathcal{L}_{tri}^t + \mathcal{L}_{qua}^t$.

After the surrogate deep cross-modal networks have been trained, we begin to create the cross-modal adversarial examples. Similarly, taking image-query-text task as an example, we hope to design the adversarial image example $\hat{o}_A^v$ by learning a perturbation δ^v added to the original image query $\hat{o}_A^v = o_A^v + \delta^v$. An effective adversarial image example should be pushed away from the positive text instance but pulled close to the negative text instance. This can be written as follows:

$$\mathcal{J}_{tri}^v = \sum^M max\left(D(\hat{H}_A^v, H_N^t) - D(\hat{H}_A^v, H_P^t) + \beta,\ 0\right), \tag{9}$$

where $\hat{H}_A^v = \mathcal{F}_{sur}^v(\hat{o}_A^v; \theta_{sur}^v)$. Besides, the quantization loss mentioned above is applied:

$$\mathcal{J}_{qua}^v = \sum^M \left(\left\|\hat{H}_A^v - \hat{B}_A^v\right\|_2^2\right), \tag{10}$$

where $\hat{B}_A^v = sign(\hat{H}_A^v)$. Combining $\mathcal{J}_{tri}^v$ with $\mathcal{J}_{qua}^v$, the total loss to learn image adversarial example is written as follows:

$$\mathcal{J}^v = \mathcal{J}_{tri}^v + \mathcal{J}_{qua}^v. \tag{11}$$

For attacking in text-query-image task, the loss can be designed as $\mathcal{J}^t = \mathcal{J}_{tri}^t + \mathcal{J}_{qua}^t$.

To optimize AACH, we first obtain the binary codes of the queries from the target model and construct the cross-modal triplets. Then, we optimize the surrogate deep networks as follows:

$$\begin{aligned}\theta_{sur}^v &= \arg\min_{\theta_{sur}^v} \mathcal{L}^v(o_A^v, o_P^t, o_N^t; \theta_{sur}^v); \\ \theta_{sur}^t &= \arg\min_{\theta_{sur}^t} \mathcal{L}^t(o_A^t, o_P^v, o_N^v; \theta_{sur}^t).\end{aligned} \tag{12}$$

Finally, we keep the θ_{sur}^v and θ_{sur}^t fixed, and learn cross-modal adversarial perturbations as follows:

$$\begin{aligned}\delta^v &= \arg\min_{\delta^v} \mathcal{J}^v(\delta^v, o_A^v, o_P^t, o_N^t; \theta_{sur}^v),\ \ s.t. \|\delta^v\|_\infty \le \epsilon^v; \\ \delta^t &= \arg\min_{\delta^t} \mathcal{J}^t(\delta^t, o_A^t, o_P^v, o_N^v; \theta_{sur}^t),\ \ s.t. \|\delta^t\|_\infty \le \epsilon^t.\end{aligned} \tag{13}$$

Algorithm 1 Adversarial Attack on Cross-Modal Hamming Retrieval (**AACH**).

Input: A black-box cross-modal target network: $\mathcal{F}_{tar}^*(o^*)$, data $O = \{o_i^v, o_i^t\}_{i=1}^M$, iteration T, $* \in \{v, t\}$

Output: The best recommended cross-modal adversarial examples: $\hat{o}^* = o^* + \delta^*, * \in \{v, t\}$

initialize $iter = 0$

Compute $B^v = sign\left(\mathcal{F}_{tar}^v(o^v)\right)$ and $B^t = sign\left(\mathcal{F}_{tar}^t(o^t)\right)$

Compute Hamming distance matrix D according to Eq. 3

Create cross-modal triplets $\{o_A^v, o_P^t, o_N^t\}$ and $\{o_A^t, o_P^v, o_N^v\}$

Train the surrogate model: **if** *not converged* **then**

  $\theta_{sur}^v = \arg\min_{\theta_{sur}^v} \mathcal{L}^v(o_A^v, o_P^t, o_N^t; \theta_{sur}^v)$;

  $\theta_{sur}^t = \arg\min_{\theta_{sur}^t} \mathcal{L}^t(o_A^t, o_P^v, o_N^v; \theta_{sur}^t)$.

end

Select δ^v and δ^t: **while** *iter* $\le T$ **do**

  $\delta^v = \arg\min_{\delta^v} \mathcal{J}^v(\delta^v, o_A^v, o_P^t, o_N^t; \theta_{sur}^v)$;

  $\delta^t = \arg\min_{\delta^t} \mathcal{J}^t(\delta^t, o_A^t, o_P^v, o_N^v; \theta_{sur}^t)$.

end

The learning procedure of AACH is summarized in Algorithm 1.

4. Experiments

4.1. Datasets

MIRFlickr-25K consists 25,000 images collected from the Flickr website. Each image is assigned with a related text description to formulate image-text pair. Following the previous method [12], totally 20,015 image-text pairs with 24 most frequent labels are used in our experiment.

NUS-WIDE consists of more than 260,000 image-text pairs. After pruning the data that has no label or text information, We use 190,421 pairs with 21 most frequent labels as our benchmark.

MS COCO [18] contains about 120,000 images. Each image is described with five semantically related sentences. We recast the MS COCO dataset to image-text pairs, and each image-text data pair is annotated with at least one label in 80 categories.

For the image modality of all benchmarks, each image is resized to $224 \times 224 \times 3$. While for the text modality of MIRFlickr-25K and NUS-WIDE, we represent texts by the bag-of-words vectors with dimensions of 1380 and 1000, respectively. Different from MIRFlickr-25K and NUS-WIDE, we extract the word embedding for the text modality of MS COCO based on Bert for studying the word-level attacking. Thus, each text data is represented with a $L * 768$ matrix, where L is the word number of text, and 768 is the dimension of embedding features. The statistics of three datasets used in our experiments are summarised in Table 1. Notably, due to the training data of the target network are generally unavailable when learning a surrogate network, we only use a part (1000) of the test data to interact with the target model and create adversarial examples. The lower number of the required training data needed for the surrogate network means lower query cost to the target model.

4.2. Baselines and Evaluations

Focusing on CMHR, we adopt four popular cross-modal binary code learning methods including DCMH [12], PRDH [37], SSAH [15], and CMHH [2]. For image modality, DCMH and PRDH use vgg-f [27] as ImgNet, while CMHH adopts AlexNet [14]. For text modality, DCMH, PRDH, and CMHH construct the TxtNet with three fully connected layers. Different from these methods, SSAH devises the TxtNet by integrating a five-layer fully connected network into a multi-scale fusion module and further constructs LabNet as an assist to ImgNet and TxtNet. Considering that the text representation of MS COCO is a feature matrix, which is different from MIRFlickr-25K and NUS-WIDE, thus we replace the original input layer with a full-convolutional layer when evaluating on MS COCO. All the networks, of course, are assumed to be unknown in our black-box setting attacking. Therefore, selecting baselines with different kinds of networks can also demonstrate the generalization of the learned adversarial examples across deep models. In addition, to evaluate the attack transferability across different code lengths, pre-trained models of these baselines in producing binary codes of different lengths are also used in our experiments. For target models, the source codes of DCMH, PRDH, and SSAH are provided by the authors. While for CMHH whose code is not available, we implement it carefully by ourselves.

Table 1: Statistics of three datasets used in our experiments.

Dataset \ Network	Target (train/test/database)	Surrogate (train)
MIRFlickr-25K	10000 / 2000 / 18015	1000
NUS-WIDE	10500 / 2100 / 188321	1000
MS COCO	10000 / 5000 / 117218	1000

To evaluate our AACH, two commonly used protocols in CMHR are adopted in this work, namely MAP that measures the accuracy of the Hamming ranking procedure and precision recall curve (PR curve) that measures the accuracy of binary code lookups. Following previous methods, we show the imperceptibility of the adversarial examples by introducing another indicator of distortion $\sqrt{\frac{\sum(\hat{o}^* - o^*)^2}{|O^*|}}$, where $* \in \{v, t\}$, and $|O^*|$ denotes the total pixel number of the original data.

4.3. Implementations

The surrogate ImgNet is constructed with the vgg-f, we just replace the last layer with a tanh layer, followed by a sign function to output binary codes. The surrogate TxtNet is built with three fully connected network layers ($text\_input \rightarrow 512 \rightarrow code\_length$). To train the surrogate deep networks and learn adversarial examples, the Adam optimizer with an initial learning rate of 0.01 is used. The margin value of β is empirically set as $K/2$. To construct the cross-modal triplets for each anchor, samples with the top 10 shortest and longest Hamming distance from the training set of the surrogate model are selected as positive instances and negative instances, respectively. For the attack strengths of different modalities, ϵ^v is set as 8 for the image, ϵ^t is set as 0.05 for text, and the adversarial perturbations $\{\delta^v, \delta^t\}$ are both initialized with zeros. After the adversarial examples are generated, we clip the image into $[0, 255]$ and clip text into $[0, 1]$. We implement all the networks including the proposed AACH and the baselines via TensorFlow [1] and run on a server with one NVIDIA Tesla P40 GPU. In the experiments, we run all the methods 10 times and report their average results.

Table 2: Attack comparison with different iterations in terms of MAP scores of two retrieval tasks on three benchmarks. The code length is set as 32 bits. The performance of regular (Reg) retrieval is shown with shading.

Tasks	Iterations		MIRFlickr-25K				NUS-WIDE				MS COCO			
			DCMH	PRDH	SSAH	CMHH	DCMH	PRDH	SSAH	CMHH	DCMH	PRDH	SSAH	CMHH
I → T	Reg		0.792	0.783	0.805	0.756	0.625	0.642	0.651	0.596	0.617	0.621	0.628	0.591
	Ours	100	0.654	0.676	0.631	0.693	0.475	0.516	0.462	0.577	0.544	0.548	0.523	0.537
		200	0.633	0.630	0.580	0.687	0.457	0.503	0.413	0.571	0.502	0.514	0.479	0.416
		500	0.631	0.616	0.564	0.645	0.439	0.497	0.395	0.413	0.461	0.497	0.453	0.411
T → I	Reg		0.779	0.773	0.787	0.772	0.632	0.638	0.664	0.615	0.593	0.610	0.646	0.603
	Ours	100	0.698	0.674	0.689	0.599	0.579	0.536	0.603	0.520	0.481	0.479	0.472	0.478
		200	0.688	0.668	0.678	0.595	0.571	0.531	0.596	0.512	0.479	0.473	0.476	0.407
		500	0.641	0.634	0.647	0.592	0.543	0.524	0.554	0.502	0.475	0.457	0.451	0.405

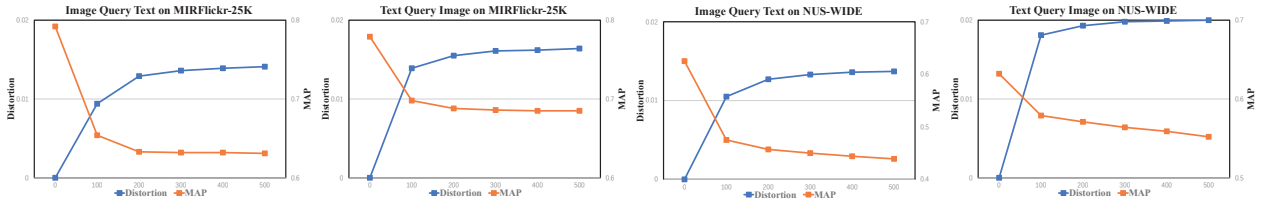


Figure 3: Variation of DCMH in terms of MAP value and distortion with increasing iterations.

4.4. Results

To evaluate the cross-modal adversarial examples with increasing learning iterations, we randomly select 500 cross-modal instance pairs and learn to create the cross-modal adversarial examples in 500 iterations. The results in terms of MAP scores of two retrieval tasks on three benchmarks are shown in Table 2. "I→T" denotes retrieval text using an adversarial image query, while "T→I" denotes retrieval image using a text query. The lower performance means better attack capability. From Table 2, some conclusions can be obtained as follows: (1) it is obvious that with an increase of the learning iteration, the attacking capability is significantly improved. (2) comparing the results of two retrieval tasks on MIRFlickr-25K and NUS-WIDE, adversarial image queries are more powerful than the adversarial text ones. One possible reason is that the raw text is represented with bag-of-words where only two values "0" and "1" are used. (3) as mentioned above, for MS COCO, we represent text with word embedding based on Bert to study the word-level attacking. Likewise, it can be seen that our AACH achieves high attacking performance, which also demonstrates the good generalization performance of our method. In addition, the distortion variation and MAP values of DCMH with increased learning iterations are shown in Fig. 3. As we initialize the adversarial perturbations to zeros, it can be seen that along with the increasing iterations the distortion gradually gets larger while the MAP score gets lower. This process demonstrates that our AACH is learning how to fool the target model.

We also evaluate the effectiveness of our AACH with different query budgets. We vary the number of adversarial queries M, and each adversarial example is learned with 500 iterations. The attack comparisons in terms of MAP scores on MIRFlickr-25K, NUS-WIDE, and MS COCO are shown in Table 3. Obviously, the MAP scores uniformly decrease with an increasing number of adversarial examples from 200 to 500. For example, taking 500 queries interacting with the target networks only once, we can respectively achieve an average over 15% and 10% decrease of "I→T" and "T→I" on MIRFlickr-25K benchmark. Notably, we also see that the attacking performance slightly decreases when boosting the number of query data from 500 to 1000, which means some inaccurate information has been obtained during the interaction with target models. Therefore, using high-quality query data would be beneficial to improve the learning efficiency of adversarial examples. PR curves of different methods under AACH are also provided in Fig. 4, where 500 adversarial examples learned with 500 iterations are used to test. The bigger area under the curve, the more semantically similar to the returned instances with the query. It can be seen our proposed AACH can effectively fool the target models. We attribute this to the cross-modal triplets construction designed to best take advantage of the information acquired from target models. As such, AACH learns to create the cross-modal adversarial examples close to the instances with different semantics while far away from the ones with similar semantics. Therefore,

Table 3: Attack comparison with different numbers of adversarial examples in terms of MAP scores of two retrieval tasks on three benchmarks. The code length is set as 32 bits.The performance of regular (Reg) retrieval is shown with shading.

Tasks	Adversarial Queries		MIRFlickr-25K				NUS-WIDE				MS COCO			
			DCMH	PRDH	SSAH	CMHH	DCMH	PRDH	SSAH	CMHH	DCMH	PRDH	SSAH	CMHH
I → T	Reg		0.792	0.783	0.805	0.756	0.625	0.642	0.651	0.596	0.617	0.621	0.628	0.591
	Ours	200	0.655	0.630	0.599	0.694	0.456	0.495	0.441	0.549	0.522	0.550	0.501	0.453
		300	0.646	0.628	0.570	0.685	0.441	0.474	0.426	0.451	0.470	0.522	0.488	0.418
		500	0.631	0.616	0.564	0.645	0.439	0.497	0.395	0.413	0.461	0.497	0.453	0.411
		1000	0.632	0.622	0.583	0.631	0.462	0.463	0.376	0.409	0.519	0.532	0.455	0.414
T → I	Reg		0.779	0.773	0.787	0.772	0.632	0.638	0.664	0.615	0.593	0.610	0.646	0.603
	Ours	200	0.748	0.721	0.742	0.655	0.579	0.555	0.612	0.520	0.540	0.534	0.532	0.454
		300	0.715	0.691	0.698	0.609	0.556	0.552	0.586	0.512	0.501	0.497	0.479	0.420
		500	0.641	0.634	0.647	0.592	0.543	0.524	0.554	0.502	0.475	0.457	0.451	0.405
		1000	0.639	0.618	0.648	0.575	0.519	0.525	0.557	0.475	0.487	0.426	0.364	0.417

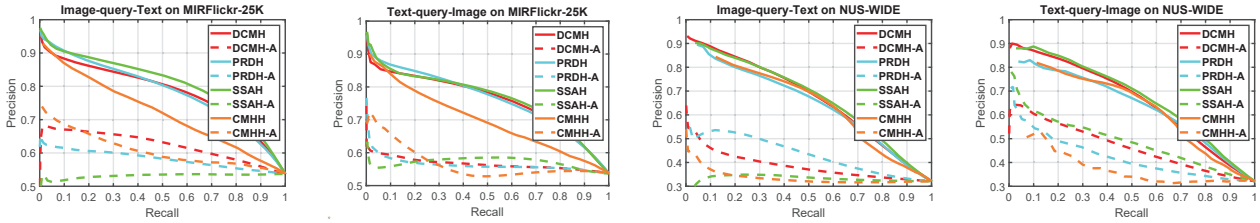


Figure 4: PR curves evaluated on MIRFlickr-25K and NUS-WID with 32-bit binary codes. '*-A' means that the target model is attacked by the proposed AACH.

Table 4: Attack transferability comparison in terms of MAP scores of two retrieval tasks on MIRFlickr-25K and NUS-WIDE. The adversarial examples are learned from the target model designed for 32-bit binary codes, aiming to attack target models of other bits."R" denotes regular retrieval, and "A" denotes attacking retrieval using our proposed AACH.

Tasks	Bits	MIRFlickr-25K								NUS-WIDE							
		DCMH		PRDH		SSAH		CMHH		DCMH		PRDH		SSAH		CMHH	
		R	A	R	A	R	A	R	A	R	A	R	A	R	A	R	A
I → T	16	0.752	0.623	0.753	0.610	0.788	0.595	0.733	0.611	0.632	0.446	0.638	0.474	0.664	0.372	0.615	0.560
	32	0.792	0.631	0.783	0.616	0.805	0.564	0.756	0.645	0.625	0.439	0.642	0.497	0.651	0.395	0.596	0.413
	64	0.777	0.638	0.768	0.618	0.801	0.599	0.758	0.641	0.621	0.473	0.648	0.486	0.659	0.388	0.593	0.502
T → I	16	0.768	0.644	0.759	0.654	0.780	0.671	0.765	0.657	0.643	0.565	0.629	0.499	0.646	0.593	0.600	0.514
	32	0.779	0.641	0.773	0.634	0.787	0.647	0.772	0.592	0.632	0.543	0.638	0.524	0.664	0.554	0.615	0.502
	64	0.781	0.658	0.776	0.688	0.783	0.683	0.779	0.659	0.624	0.552	0.641	0.527	0.657	0.567	0.605	0.506

the number of queries from the target model can be significantly decreased, which means AACH is more applicable to the case that the query budget will be highly limited.

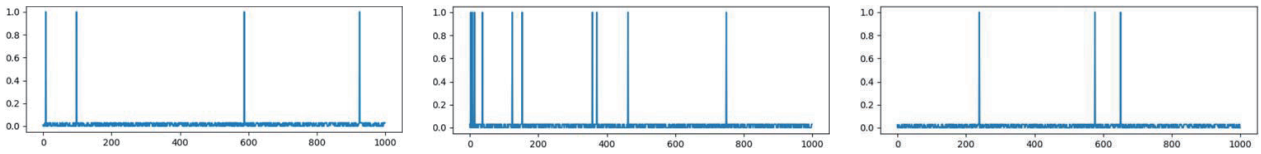
Furthermore, the transferability of the created adversarial examples across different code lengths is also evaluated, as shown in Table 4. We create adversarial examples based on the surrogate cross-modal networks with 32 bits. With the learned adversarial examples, we attack the target cross-modal networks in different code lengths. From Table 4, we can find that the adversarial examples can be well transferable among different code lengths.

Compared with CMLA. CMLA is a pioneer work in the cross-modal Hamming attacking area. However, the CMLA is designed for a white-box setting. We show the comparison between CMLA and ours on three different benchmarks in Table 5. Taking retrieval on MS COCO as an example, CMLA achieves higher attacking performance, which considerably decreases the retrieval accuracy, particularly in the text modality achieve. This is mainly attributed to CMLA has all prior knowledge including both target network structures and label information. Comparing with CMLA, our proposed AACH does not depend on any prior knowledge, which creates adversarial examples by limitedly interacting with target models. Even so, our proposed AACH achieves

Table 5: Attack comparison with CMLA in terms of MAP scores on different datasets. The code length is set as 32 bits.The performance of regular (Reg) retrieval is shown with shading.

Tasks	Methods	MIRFlickr-25K				NUS-WIDE				MS COCO			
		DCMH	PRDH	SSAH	CMHH	DCMH	PRDH	SSAH	CMHH	DCMH	PRDH	SSAH	CMHH
I → T	Reg	0.792	0.783	0.805	0.756	0.625	0.642	0.651	0.606	0.617	0.621	0.628	0.591
	CMLA	0.521	0.598	0.600	0.563	0.457	0.404	0.357	0.331	0.442	0.396	0.420	0.402
	Ours	0.631	0.616	0.564	0.645	0.439	0.497	0.395	0.413	0.461	0.497	0.453	0.411
T → I	Reg	0.779	0.773	0.787	0.772	0.632	0.638	0.664	0.625	0.593	0.610	0.646	0.603
	CMLA	0.561	0.501	0.575	0.564	0.371	0.439	0.320	0.325	0.247	0.256	0.297	0.370
	Ours	0.641	0.634	0.647	0.592	0.543	0.524	0.554	0.502	0.475	0.457	0.451	0.405

(a) Visualization of both the raw image instance (left) and adversarial image query (right)



(b) Visualization of adversarial text query

Figure 5: Cross-modal adversarial examples learned by the proposed AACH.

comparable performance with CMLA on the "I→T" task, demonstrating the efficiency of our method. Comparing the attacking results of "T→I" between different benchmarks, we can find that methods uniformly achieve higher attacking performances on MS COCO than that on MIRFlickr-25K and NUS-WIDE. This is because that the real feature representation (feature embedding) of text has high-dimensional feature space, which can provide rich semantics for regular retrieval task, but at the same time, increase the risk of being attacked.

Finally, some visualization results of the learned cross-modal adversarial examples on NUS-WIDE benchmark are provided in Fig 5. For image, we show both the raw queries and the created adversarial examples, where the difference between raw queries and adversarial ones is nearly invisible. While for text, we directly show the adversarial examples. Minor perturbation can be seen in text queries because of the bag-of-words vector representations used in raw text modality.

5. Conclusions

This paper proposes a novel adversarial example learning method, dubbed AACH, for cross-modal Hamming retrieval, which aims to attack a target deep cross-modal Hamming model in a black-box setting. Our proposed AACH constructs a surrogate model to interact with the target networks by querying it, without requiring any prior knowledge about the target networks. Therefore, to some extent, AACH is more practical for real-world applications compared with state-of-the-art methods. Besides, a novel triplet construction module is proposed to formulate cross-modal triplets, with which we significantly enhance the learning efficiency of adversarial examples. In this way, AACH can be applied in extreme conditions where the query budget is highly limited. Finally, the effectiveness of AACH is well demonstrated by the comprehensive experiments conducted on three representative benchmarks.

6. Acknowledgements

Chao Li and Cheng Deng were supported in part by the National Natural Science Foundation of China under Grant 62071361, Key Research and Development Program of Shaanxi under Grant 2021ZDLGY01-03, and the Fundamental Research Funds for the Central Universities ZDRC2102.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *OSDI*, pages 265–283, 2016. 5

[2] Yue Cao, Bin Liu, Mingsheng Long, and Jianmin Wang. Cross-modal hamming hashing. In *ECCV*, pages 207–223, 2018. 2, 5

[3] Yue Cao, Mingsheng Long, Jianmin Wang, Qiang Yang, and Philip S. Yu. Deep visual-semantic hashing for cross-modal retrieval. In *KDD*, pages 1445–1454, 2016. 2

[4] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *CVPR*, pages 12655–12663, 2020. 2

[5] Shuyu Cheng, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Improving black-box adversarial attacks with a transfer-based prior. In *NeurIPS*, pages 10932–10942, 2019. 2

[6] Cheng Deng, Zhaojia Chen, Xianglong Liu, Xinbo Gao, and Dacheng Tao. Triplet-based deep hashing network for cross-modal retrieval. *IEEE Transactions on Image Processing*, 27(8):3893–3903, 2018. 2

[7] Guiguang Ding, Yuchen Guo, and Jile Zhou. Collective matrix factorization hashing for multimodal data. In *CVPR*, pages 2075–2082, 2014. 2

[8] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *CVPR*, pages 9185–9193, 2018. 2

[9] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *CVPR*, pages 4312–4321, 2019. 2

[10] Yiwen Guo, Ziang Yan, and Changshui Zhang. Subspace attack: Exploiting promising subspaces for query-efficient black-box attacks. In *NeurIPS*, pages 3820–3829, 2019. 2

[11] Hengtong Hu, Lingxi Xie, Richang Hong, and Qi Tian. Creating something from nothing: Unsupervised knowledge distillation for cross-modal hashing. In *CVPR*, June 2020. 2

[12] Qing-Yuan Jiang and Wu-Jun Li. Deep cross-modal hashing. In *CVPR*, pages 3232–3240, 2017. 2, 5

[13] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *AAAI*, volume 34, pages 8018–8025, 2020. 2

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, pages 1097–1105, 2012. 5

[15] Chao Li, Cheng Deng, Ning Li, Wei Liu, Xinbo Gao, and Dacheng Tao. Self-supervised adversarial hashing networks for cross-modal retrieval. In *CVPR*, pages 4242–4251, 2018. 2, 5

[16] Chao Li, Cheng Deng, Lei Wang, De Xie, and Xianglong Liu. Coupled cyclegan: Unsupervised hashing network for cross-modal retrieval. In *AAAI*, 2019. 2

[17] Chao Li, Shangqian Gao, Cheng Deng, De Xie, and Wei Liu. Cross-modal learning with adversarial samples. In *NeurIPS*, pages 10791–10801, 2019. 2, 3

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 5

[19] Zijia Lin, Guiguang Ding, Mingqing Hu, and Jianmin Wang. Semantics-preserving hashing for cross-view retrieval. In *CVPR*, pages 3864–3872, 2015. 2

[20] Xuanwu Liu, Guoxian Yu, Carlotta Domeniconi, Jun Wang, Yazhou Ren, and Maozu Guo. Ranking-based deep cross-modal hashing. In *AAAI*, volume 33, pages 4400–4407, 2019. 2

[21] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016. 2

[22] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 2

[23] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *CVPR*, pages 2574–2582, 2016. 2

[24] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pages 372–387. IEEE, 2016. 2

[25] Heng Tao Shen, Luchen Liu, Yang Yang, Xing Xu, Zi Huang, Fumin Shen, and Richang Hong. Exploiting subspace relation in semantic labels for cross-modal hashing. *IEEE Transactions on Knowledge and Data Engineering*, 2020. 2

[26] Yucheng Shi, Siyu Wang, and Yahong Han. Curls & whey: Boosting black-box adversarial attacks. In *CVPR*, pages 6519–6527, 2019. 2

[27] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *In arXiv preprint arXiv:1409.1556*, 2014. 5

[28] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019. 2

[29] Shupeng Su, Zhisheng Zhong, and Chao Zhang. Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. In *ICCV*, pages 3027–3035, 2019. 2

[30] Mengying Sun, Fengyi Tang, Jinfeng Yi, Fei Wang, and Jiayu Zhou. Identify susceptible locations in medical records via adversarial attacks on deep predictive models. In *ACM SIGKDD*, pages 793–801. ACM, 2018. 2

[31] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 2

[32] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. Adversarial cross-modal retrieval. In *ACMMM*, pages 154–162, 2017. 2

[33] Gengshen Wu, Zijia Lin, Jungong Han, Li Liu, Guiguang Ding, Baochang Zhang, and Jialie Shen. Unsupervised deep hashing via binary latent factor models for large-scale cross-modal retrieval. In *IJCAI*, pages 2854–2860, 2018. 2

[34] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *CVPR*, pages 2730–2739, 2019. 2

[35] Ruiqing Xu, Chao Li, Junchi Yan, Cheng Deng, and Xianglong Liu. Graph convolutional network hashing for cross-modal retrieval. In *IJCAI*, pages 10–16, 2019. 2

[36] Xiaojun Xu, Xinyun Chen, Chang Liu, Anna Rohrbach, Trevor Darell, and Dawn Song. Can you fool ai with adversarial examples on a visual turing test. *arXiv preprint arXiv:1709.08693*, 2017. 2

[37] Erkun Yang, Cheng Deng, Wei Liu, Xianglong Liu, Dacheng Tao, and Xinbo Gao. Pairwise relationship guided deep hashing for cross-modal retrieval. In *AAAI*, pages 1618–1625, 2017. 2, 5

[38] Erkun Yang, Tongliang Liu, Cheng Deng, and Dacheng Tao. Adversarial examples for hamming space search. *IEEE transactions on cybernetics*, 2018. 2

[39] Li Yuan, Tao Wang, Xiaopeng Zhang, Francis EH Tay, Zequn Jie, Wei Liu, and Jiashi Feng. Central similarity quantization for efficient image and video retrieval. In *CVPR*, pages 3083–3092, 2020. 2

[40] Jian Zhang, Yuxin Peng, and Mingkuan Yuan. Unsupervised generative adversarial cross-modal hashing. *In arXiv preprint arXiv:1712.00358*, 2017. 2

[41] Xi Zhang, Hanjiang Lai, and Jiashi Feng. Attention-aware deep adversarial hashing for cross-modal retrieval. In *ECCV*, pages 591–606, 2018. 2