Vulnerability vs. Reliability: Disentangled Adversarial Examples for Cross-Modal Learning

Chao Li
School of Electronic Engineering
Xidian University
Xi'an, Shaanxi, China
chaolee.xd@gmail.com

Haoteng Tang
Electrical and Computer Engineering
University of Pittsburgh
Pittsburgh, PA, USA
hat64@pitt.edu

Cheng Deng*
School of Electronic Engineering
Xidian University
Xi'an, Shaanxi, China
chdeng.xd@gmail.com

Liang Zhan
Department of ECE & Bioengineering
University of Pittsburgh
Pittsburgh, PA, USA
liang.zhan@pitt.edu

Wei Liu
Tencent AI Lab
Shenzhen, China
wl2223@columbia.edu

ABSTRACT
The vulnerability of deep neural networks has gained a great upsurge of research attention, which engages well-designed examples through adding little perturbations to fool a well-performed network. Meanwhile, a progress has been made in leveraging adversarial examples to boost the robustness of deep cross-modal networks. However, for cross-modal learning, both the causes of adversarial examples and their latent advantages in learning cross-modal correlations are under-explored. In this paper, we propose novel Disentangled Adversarial examples for Cross-Modal learning, dubbed DACM. Specifically, we first divide cross-modal data into two aspects, namely modality-related component and modality-unrelated counterpart, and then learn to improve the reliability of network using the modality-related component. To achieve this goal, we apply the generation of adversarial perturbations to strengthen cross-modal correlations, wherein the modality-related component is acquired through gradually detaching the modality-unrelated component. Finally, the proposed DACM is employed to create modality-related examples towards the application of cross-modal hashing retrieval. Extensive experiments carried out on two cross-modal benchmarks show that the adversarial examples learned by DACM are efficient at fooling a target deep cross-modal hashing network. On the other hand, training this target model by merely leveraging our created modality-related examples in turn significantly promotes the robustness of this model itself.

KEYWORDS
Cross-Modal Learning; Deep Learning; Adversarial Example; Cross-Modal Retrieval; Hash Code Learning

*Corresponding author.

KDD '20, August 23–27, 2020, Virtual Event, CA, USA

ACM ISBN 978-1-4503-7998-4/20/08…$15.00
https://doi.org/10.1145/3394486.3403084

ACM Reference Format:
Chao Li, Haoteng Tang, Cheng Deng, Liang Zhan, and Wei Liu. 2020. Vulnerability vs. Reliability: Disentangled Adversarial Examples for Cross-Modal Learning. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '20), August 23–27, 2020, Virtual Event, CA, USA.* 9 pages. https://doi.org/10.1145/3394486.3403084

1 INTRODUCTION
Cross-modal learning has already become a prime technology to approach the applications of massive multimedia data, such as cross-modal retrieval [21, 40, 48], image captioning [35, 39], text-to-image synthesis [41, 51], and visual query answering [2, 50]. These various cross-modal learning tasks share a common challenge, exploiting latent cross-modal correlations to associate different modalities.

Recently, many kinds of deep networks [15, 17, 44] have emerged as powerful yet efficient models to tackle a broad spectrum of complex learning problems, and thus deep network-based methods (*e.g.*, deep learning, deep reinforcement learning, and deep transfer learning) greatly improve the performances for kinds of cross-modal applications. Even so, it remains a fresh topic that deep cross-modal networks are vulnerable against adversarial examples, which can easily fool a well-performed deep model with little perturbations imperceptible to humans [13, 16, 25, 34, 36–38, 45–47, 52].

In the cross-modal learning area, neither the causes of adversarial examples nor their roles in building cross-modal correlations have been explicitly delineated in previous works. Considering the diversity tasks engaging cross-modal learning, this paper dedicates to hashing-based cross-modal retrieval between image and text as an example for a better understanding. As shown in Fig. 1, to generate reliable cross-modal hash codes, the regular methods are generally over-reliant on deep networks to pursue the correlations between different modalities, which easily results in the model vulnerability. However, in essence, beyond the modality-related component, the modality-unrelated counterpart hidden in the original cross-modal data always makes an impediment to building reliable cross-modal correlations, which thus should be filtered out. To address this problem, a possible manner is to remove the modality-unrelated component in an adversarial learning fashion. Inspired by the generation of adversarial examples that attack a target deep network by adding learned adversarial perturbations

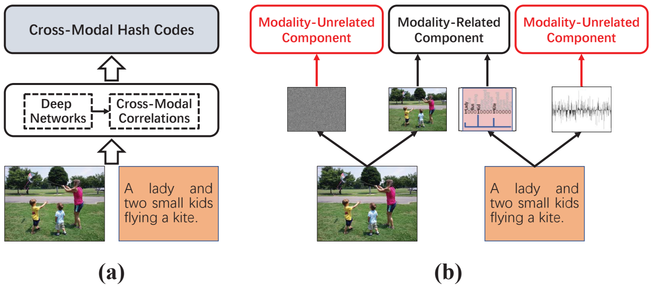


Figure 1: (a) The regular cross-modal learning simply considers the cross-modal data as a whole and relies on training deep networks to build cross-modal correlations. (b) In this work, we introduce a new perspective to learn the cross-modal correlations by exploring the modality-related component.

into original data, we disentangle the modality-unrelated component from original data following an exactly opposite manner. To be specific, we make the modality-unrelated component serve as the adversarial perturbation which will be leveraged to construct adversarial examples. In this way, the modality-related component for original cross-modal data can be obtained by filtering out the modality-unrelated counterparts. As a result, a new training dataset consisting of the modality-related examples is created. Using the newly created dataset to train the deep cross-modal network, high robustness and even better retrieval performance in contrast to the original training dataset can be harvested simultaneously.

In this paper, we present the Disentangled Adversarial examples for Cross-Modal learning (DACM), which provides new insight into adversarial examples in discovering the correlations for cross-modal learning. Specifically, our DACM acquires modality-related component across different modalities through a newly proposed adversarial learning method which removes modality-unrelated counterparts by optimizing adversarial perturbations. The highlights of our work can be summarized as follows:

- We present a novel perspective of adversarial examples in learning modality-related representations between different modalities, which corroborates that cross-modal adversarial examples are mainly produced by over pursuing cross-modal consistency but ignoring its divergence.
- We propose a simple yet effective disentangled cross-modal adversarial examples learning method, where the adversarial examples and modality-related representations are learned in a unified framework by disentangling the modality-unrelated representations from the original data.
- We take cross-modal hashing retrieval as an application to evaluate the proposed DACM. Experiments on two widely-used cross-modal retrieval benchmarks show the effectiveness of our DACM in learning modality-related representations from the original cross-modal data and further improving the retrieval robustness.

The remainder of this paper is structured as follows. First, we briefly introduce and discuss representative methods for adversarial attacks and conducting cross-modal hashing learning in Section 2. Then we elaborate on the motivation and basic ideas of our method in Section 3. Section 4 provides experiments of our method, and finally, Section 5 draws the conclusions.

2 RELATED WORKS

Adversarial Examples. Adding small carefully crafted perturbations called adversarial perturbations into original data, Szegedy et al. [47] recast the data to adversarial examples, and then fed them into a target deep network, which can easily drive the target deep model to a wrong prediction. Following this, various attacks are presented, *e.g.*, iterative fast gradient sign method (IFGSM) [22], one pixel attack [45], Carlini and Wagner attack [6], and universal attack [37]. However, these methods mainly focus on the applications of single modality tasks, such as image classification. Recently, aiming to cross-modal learning which is a more complex case, Show-and-Fool [7] is proposed to attack an image captioning system by executing visual language grounding. Xu et al. [54] dedicated exact adversarial attacks of targeted partial captions. Xu et al. [53] studied adversarial examples for visual question answering. On the other hand, by virtue of the learned adversarial examples, a robust model can further be implemented. Chen et al. [8] presented to learn adversarial examples to augment visual-semantic training samples, thus improve the reliability of their target model. However, more efforts are needed to investigate how adversarial examples affect deep networks for other cross-modal tasks such as cross-modal retrieval.

Cross-Modal Hashing. Comparing with typical single-modal hashing [11, 12, 26, 28–31, 42, 57], when dealing with large-scale cross-modal data, two challenges arise from two aspects: the tremendous data volume and the heterogeneity between different modalities. To address them, plenty of cross-modal hashing methods are presented [4, 5, 19, 20, 24, 27, 32, 49, 58]. Depending on whether to use deep networks or not, these methods can be grouped into two categories: hand-crafted feature based cross-modal hashing and deep-feature based counterparts. Compared with the hand-crafted feature based methods, deep hashing methods are built upon deep neural networks that holding superior nonlinear approximation capacity in building correlations between different modalities, and thus always achieve more appealing performance. Inspired by this, the constraints of pairwise loss [19], triplet loss [10], and rank loss [33] are further injected into deep models to facilitate the building of cross-modal correlations. Taking tag information of image as supervision, WDHT [14] is proposed to learn hash codes by using tag embedding in a weakly supervised fashion. Deep joint semantics reconstructing hashing (DJSRH) [43] studies joint correlations between different modalities by fusing multiple similarity relationships. However, these methods maximally pursue the modality-related correlations while neglecting the effect of the modality-unrelated ones. On the contrary, ADAH [9] constructs an attention mask to focus on more informative parts of multi-modal data. SPDQ [55] utilizes a deep network to project cross-modal data onto two feature spaces, where cross-modal shared and intra-modal private representations are learned individually.

The recently proposed CMLA [23], which focuses on designing cross-modal adversarial examples, is also related to the proposed DACM. However, the proposed method has great differences with

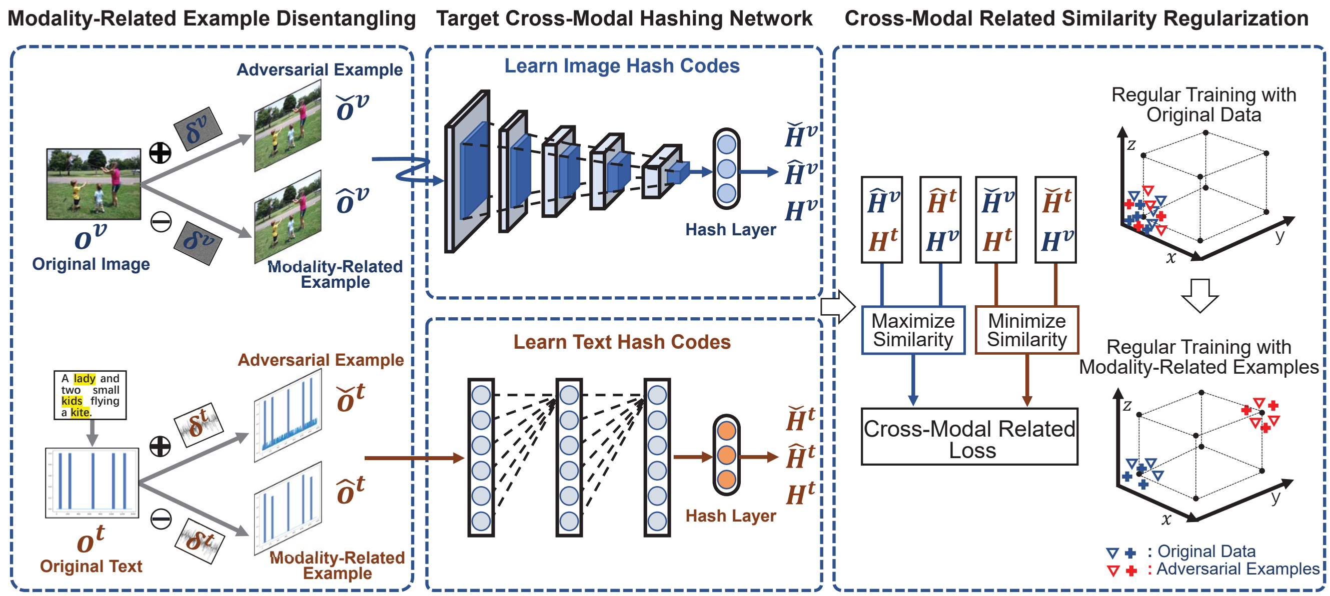


Figure 2: The pipeline of our proposed DACM for cross-modal hashing consisting of three parts: modality-related example disentangling, target cross-modal hashing network to generate hash codes for image and text, and cross-modal related similarity regularization.

CMLA in the following three aspects. 1) CMLA learns adversarial examples aiming to attack a target deep cross-modal network, while DACM uses adversarial examples to extract modality-related representations. 2) To improve the particular capability of the adversarial examples in attacking cross-modal retrieval without damaging intra-modal retrieval performance, CMLA learns adversarial examples by decreasing the inter-modality similarity and simultaneously keeping intra-modality similarity. In contrast, DACM dedicates to disentangling the modality-related representations from original cross-modal data rather than the intra-modal correlations, which has an essential difference with CMLA. 3) To improve the robustness of a target model, CMLA has to merge the adversarial examples with original training samples and retrain the target model, which causes inefficiency. DACM creates the new training dataset consisting of modal-related examples and thus can achieve the same goal by only execute the regular training. Therefore, different from previous methods, DACM takes a fresh look at the adversarial examples and their ability to build cross-modal correlations.

3 PROPOSED DACM

Fig. 2 shows the overall flowchart of the proposed DACM including three parts: modality-related example disentangling, target cross-modal hashing network to generate hash codes for image and text modality, and cross-modal related similarity regularization. For each original cross-modal data pair $\{o^v, o^t\}$ as input, in the modality-related example disentangling, we initialize two perturbations $\{\delta^v, \delta^t\}$ to create adversarial example $\{\check{o}^v, \check{o}^t\}$ by adding the perturbation into the original sample and create modality-related example $\{\hat{o}^v, \hat{o}^t\}$ by subtracting the perturbation from the original sample, respectively. Then, feeding the original cross-modal data, adversarial example, and modality-related example into the given target cross-modal hashing network, we can obtain their corresponding hash codes H^*, $\check{H}^*$, and $\hat{H}^*$, where $* \in \{v, t\}$. Following, a cross-modal related similarity regularization is designed to learn the effective perturbations by making adversarial examples decrease retrieval accuracy while modality-related examples increase retrieval accuracy. Finally, utilizing the learned modality-related examples, we can further train a crude deep cross-modal hashing network effectively. In other words, the cross-modal correlation exploring task is reformulated as a new adversarial examples learning problem in this paper. Next, the novel disentangled adversarial examples learning method shown in Fig. 2, is introduced in this session.

3.1 Problem Definition

Cross-modal hashing aims to produce binary hash codes $B^* \in \{-1, 1\}^K$ for different modality data by learning hash functions $\mathcal{H}^*$, where K is code length, and $* \in \{v, t\}$ denotes image and text modality. For a better understanding of the proposed method, we first describe some notations. Let $O = \{o_i\}_{i=1}^N$ be a cross-modal dataset with N data points, and $o_i = \{o_i^v, o_i^t\}$ represents the ith cross-modal data. o_i^v and o_i^t respectively denotes image and text representation of o_i, and are annotated with identical labels. S is a pairwise similarity matrix that describes semantic similarity between each pair of cross-modal data, where $S_{ij} = 1$ means that o_i and o_j are semantically similar, otherwise $S_{ij} = 0$. Following the multi-label setting in previous methods [4, 10, 19], we set $S_{ij} = 1$ when o_i and o_j share at least one label, otherwise $S_{ij} = 0$. In a deep hashing method, two neural networks are usually constructed to serve as hash functions $\{\mathcal{H}^v, \mathcal{H}^t\}$. We denote the outputs of the hash functions as the hash codes $\{H^v = \mathcal{H}^v(o^v), H^t = \mathcal{H}^t(o^t)\}$. Finally, the binary hash codes B^* are obtained by applying a sign function to $\{H^v, H^t\}$:

$$B^* = sign(H^*), \ * \in \{v, t\}. \tag{1}$$

For deep hashing networks $\mathcal{H}^*(o^*, \theta^*)$, let θ^* be network parameters. To train a deep cross-modal model for regular retrieval, the deep cross-modal hashing network is encouraged to output similar hash codes for semantically similar data, which can be written as follows:

$$\min_{\theta^v, \theta^t} D\left(\mathcal{H}^v(o^v; \theta^v), \mathcal{H}^t(o^t; \theta^t)\right), \tag{2}$$

where $D(\cdot, \cdot)$ is a distance measure such as Hamming distance.

In this paper, the proposed DACM aims to explore adversarial perturbations $\{\delta^v, \delta^t\}$, which can result in the decline (increasing) of retrieval accuracy by adding (removing) the perturbations. Given a semantically similar cross-modal data pair $\{o^v, o^t\}$, to better understand the proposed DACM, we take the image-query-text task for example. The learning of image adversarial perturbation can be defined as follows:

$$\begin{aligned} &\triangle(o^v, o^t, \mathcal{H}^v, \mathcal{H}^t) := \min_{\delta^v} \left\|\delta^v\right\|_p , \\ s.t. \ \min_{\delta^v} \ &D\left(\mathcal{H}^v(o^v - \delta^v; \theta^v), \mathcal{H}^t(o^t; \theta^t)\right) - \\ &D\left(\mathcal{H}^v(o^v + \delta^v; \theta^v), \mathcal{H}^t(o^t; \theta^t)\right), \ \|\delta^v\|_p \le \epsilon^v, \end{aligned} \tag{3}$$

where ϵ^v denotes the maximal disentangled strength, and $\|\cdot\|_p$ denotes L_p norm ($p = \infty$ in this paper), measuring the difference between the adversarial example and the original data.

3.2 Disentangled Adversarial Example Learning

The proposed DACM seeks for the perturbations $\{\delta^v, \delta^t\}$ to construct modality-related examples $\{\hat{o}^v, \hat{o}^t\}$ and adversarial examples $\{\breve{o}^v, \breve{o}^t\}$ by removing and adding operation as follows:

$$\begin{aligned} \hat{o}^v &= o^v - \delta^v, \ \hat{o}^t = o^t - \delta^t; \\ \breve{o}^v &= o^v + \delta^v, \ \breve{o}^t = o^t + \delta^t. \end{aligned} \tag{4}$$

For a cross-modal data pair $\{o^v, o^t\}$ with short distance in Hamming space, we disentangle the cross-modal related representations by learning adversarial perturbations $\{\delta^v, \delta^t\}$. During this learning process, the adversarial examples $\{\breve{o}^v, \breve{o}^t\}$ are pushed away from the data that sharing similar semantics with them, at the same time maintaining the modality-related examples $\{\hat{o}^v, \hat{o}^t\}$ to be close to their semantically similar data.

Similarly, taking image-query-text task for example, there should be a long Hamming distance $D\left(\mathcal{H}^v(\breve{o}^v; \theta^v), \mathcal{H}^t(o^t; \theta^t)\right)$ between the hash codes that generated from the image adversarial example $\breve{o}^v$ and that from the original text o^t. In contrast, it is expected a short Hamming distance $D\left(\mathcal{H}^v(\hat{o}^v; \theta^v), \mathcal{H}^t(o^t; \theta^t)\right)$ between hash codes that generated from the image modality-related example $\hat{o}^v$ and that from the original text o^t. Ideally, modality-related examples should be assigned identical hash codes with original data, while adversarial examples should be assigned totally different hash codes with original data. However, considering that the optimization in binary code learning is intractable, a simple yet effective disentangled learning method based on a similarity loss is proposed. The loss function consists of two terms: one aims to maximize the similarity between the hash codes that produced from the original text and that from the image modality-related examples; the other one is designed to minimize similarity between the hash codes that produced from the original text and that from the image adversarial examples. The loss function is formulated as follows:

$$\begin{aligned} \min_{\delta^v} \mathcal{J}^v = \frac{1}{N^2} \Bigg(&\sum_{i,j=1}^{N} \left(S_{ij}\Gamma_{ij} + \log(1 + e^{-\Gamma_{ij}})\right) - \\ &\sum_{i,j=1}^{N} \left(S_{ij}\Theta_{ij} - \log(1 + e^{\Theta_{ij}})\right)\Bigg), \ s.t. \ \|\delta^v\|_\infty \le \epsilon^v, \end{aligned} \tag{5}$$

where Γ is defined as $\frac{1}{2}(\breve{H}^v)(H^t)^\top$ to approximate the Hamming similarity between the image adversarial examples and the original text, and $\Theta = \frac{1}{2}(\hat{H}^v)(H^t)^\top$ stands for the Hamming similarity between the image modality-related examples and the original text. To learn δ^v, we make the Hamming distance between the manipulated image and the original text become longer when adding δ^v into the original image, while becoming shorter when subtracting δ^v. That is to say, the modality-unrelated representation is disentangled from cross-modal data in the progress of learning perturbations.

Accordingly, for the image retrieval using text query, the loss function can be written as follows:

$$\begin{aligned} \min_{\delta^t} \mathcal{J}^t = \frac{1}{N^2} \Bigg(&\sum_{i,j=1}^{N} \left(S_{ij}\Upsilon_{ij} + \log(1 + e^{-\Upsilon_{ij}})\right) - \\ &\sum_{i,j=1}^{N} \left(S_{ij}\Psi_{ij} - \log(1 + e^{\Psi_{ij}})\right)\Bigg), \ s.t. \ \|\delta^t\|_\infty \le \epsilon^t, \end{aligned} \tag{6}$$

where $\Upsilon = \frac{1}{2}(\breve{H}^t)(H^v)^\top$ and $\Psi = \frac{1}{2}(\hat{H}^t)(H^v)^\top$.

In this way, two kinds of adversarial perturbations for different modalities are learned, respectively. At first sight, two adversarial perturbations are learned independently. Actually, the hash codes H^v and H^t, which are generated by a well-performed target model, naturally preserve the cross-modal similarity correlations. Therefore, taking H^v and H^t as supervisions, the proposed DACM can simultaneously learn adversarial examples and modality-related examples effectively.

3.3 Optimization

Given a target deep hashing network such as DCMH [19] denoted as $F(o^v, o^t; \theta^v, \theta^t)$ and image-text pairs $\{o^v, o^t\}$, we randomly initialized perturbations $\{\delta^v, \delta^t\}$, and create $\{\breve{o}^v, \breve{o}^t\}$ and $\{\hat{o}^v, \hat{o}^t\}$ by Eq. (4). The hash codes $\{H^v, H^t\}$ for $\{o^v, o^t\}$ are calculated by forward propagation. With $\{H^v, H^t\}$, we learn the adversarial examples and the modality-related examples simultaneously by minimizing Eq. (5) and Eq. (6) using a back-propagation (BP) algorithm:

$$\begin{aligned} \delta^v &= \arg\min_{\delta^v} J^v(\delta^v, \breve{o}^v, \hat{o}^v, H^t; \theta^v), \ s.t. \ \|\delta^v\|_\infty \le \epsilon^v; \\ \delta^t &= \arg\min_{\delta^t} J^t(\delta^t, \breve{o}^t, \hat{o}^t, H^v; \theta^t), \ s.t. \ \|\delta^t\|_\infty \le \epsilon^t, \end{aligned} \tag{7}$$

where the modality-related representations $\{\hat{o}^v, \hat{o}^t\}$ thus can be disentangled from the original cross-modal data. The details of training the proposed DACM are summarized in Algorithm 1.

Algorithm 1: Disentangled Adversarial Examples for Cross-Modal Learning (**DACM**).

Input: Target deep cross-modal hashing model: $\mathcal{H}^*(o^*, \theta^*)$, $* \in \{v, t\}$, and a cross-modal dataset: $\{o_i^v, o_i^t\}_{i=1}^N$
Output: The optimal modality-related examples: $\hat{o}^v$ and $\hat{o}^t$
Maximum iteration: T_{max}, disentangled strength: $\{\epsilon^v, \epsilon^t\}$, batch_size: 128, $n = \lceil N/128 \rceil$
for $j = 1; j \le n;$ **do**
  Initialize $iter = 0$
  Compute H^v and H^t by forward propagation:
  $H^v = \mathcal{H}^v(o^v, \theta^v);\ H^t = \mathcal{H}^t(o^t, \theta^t)$
  while $iter \le T_{max}$ **do**
    if *not converged* **then**
      Update δ^v and δ^t by back propagation:
      $\delta^v = \arg\min_{\delta^v} J^v(\delta^v, \breve{o}^v, \hat{o}^v, H^t; \theta^v)$;
      $\delta^t = \arg\min_{\delta^t} J^t(\delta^t, \breve{o}^t, \hat{o}^t, H^v; \theta^t)$
      Clip δ^v to range $[0, \epsilon^v]$; clip δ^t to range $[0, \epsilon^t]$
    end
  end
  Clip $\hat{o}^v$ to range $[0, 255]$; clip $\hat{o}^t$ to range $[0, 1]$
end
Return modality-related examples $\hat{o}^v$ and $\hat{o}^t$.

Inputting the learned $\hat{o}^v$ and $\hat{o}^t$ into a target model, we train the target model by using a BP algorithm:

$$\theta^v, \theta^t = \arg\min_{\theta^v, \theta^t} F(\hat{o}^v, \hat{o}^t; \theta^v, \theta^t). \tag{8}$$

In this way, both the retrieval efficiency of the target model and its defense against adversarial attacks can be acquired simultaneously.

3.4 Implementation Details

All the codes of the proposed DACM are implemented via TensorFlow [1] and executed on a server with two NVIDIA Tesla P40 GPUs with a graphics memory capacity of 24GB for each one. The normalized images size is $224 \times 224 \times 3$. For learning adversarial examples, we adopt Adam optimizer with an initial learning rate 0.1 and train each sample for T_{max} iterations, $T_{max} \in \{50, 100, 500, 2000\}$. Mini-batch size is fixed at 128. ϵ^v is set to 8 for image modality, and ϵ^t is set to 0.03 for text modality. After adversarial examples and modality examples are generated, we clip image into $[0, 255]$ and clip text into $[0, 1]$.

4 EXPERIMENTS

4.1 Experimental Setup

In this section, we evaluate the proposed method DACM with three state-of-the-art deep cross-modal hashing networks on two benchmark datasets: MIRFlickr-25K [18] and NUS-WIDE [9].

MIRFlickr-25K [18] is collected from Flickr with 25,000 images. Each image is associated with a text description. In our experiments, we totally select 20,015 image-text pairs, and each image-text pair is annotated with at least one of 24 unique labels. The text is represented by a 1,386-dimensional bag-of-words vector for the text modality. For the training of target models and the generation of modality-related examples, we randomly select 5,000 image-text pairs as a training set. For the generation of adversarial examples, we randomly select 1,000 image-text pairs as the test set, while the rest is used as the database.

NUS-WIDE [9] contains 269,648 images collected from a public web, where 81 ground-truth concepts are annotated for retrieval evaluation. Following the setting in CMLA [23], we prune the data that has no label or text information, then a subset of 190,421 image-text pairs that belong to the 21 most-frequent concepts are adopted as our benchmark. The text is represented by a bag-of-words vector with 1,000 dimensions. To evaluate our DACM, 5,000 and 2,100 image-text pairs are randomly selected as the training set and the test set, respectively, and the rest is used as the database.

Evaluation. Following previous works [3, 4, 27], three commonly used protocols in cross-modal retrieval: Mean Average Precision (MAP), precision-recall curve (PR curve), and Precision@1000 are adopted to evaluate the performances of our proposed DACM, where Mean Average Precision (MAP) is used to measure the accuracy of the Hamming distances, precision-recall curve (PR curve) is used to measure the accuracy of hash lookups, and Precision@1000 curve is used to evaluate the precision with respect to the number of top feedbacks. Besides, the distortion between the original cross-modal data o^* and the distorted one $\hat{o}^*$ is measured as:

$$P^* = \sqrt{\frac{\sum (\hat{o}^* - o^*)^2}{|o^*|}},\ * \in \{v, t\}. \tag{9}$$

Here taking the MIRFlickr-25K dataset as an example, $|o^v|$ and $|o^t|$ are the total pixel numbers of the original data, set as $150,528(224*224*3)$ and $1,380$ for image modality and text modality, respectively.

It should be noticed that the main goal of our work is to study a novel cross-modal correlation learning method based on adversarial examples rather than to focus on designing a new deep cross-modal network. Therefore, to show the effectiveness of our proposed DACM, three popular deep cross-modal hashing models DCMH [19], SSAH [24], and PRDH [56] are adopted as target models, and we keep the identical network structures as reported in their papers [19][24][56]. Their performances on the regular retrieval and the defense to adversarial queries are provided, including both the cases before and after training with modality-related examples.

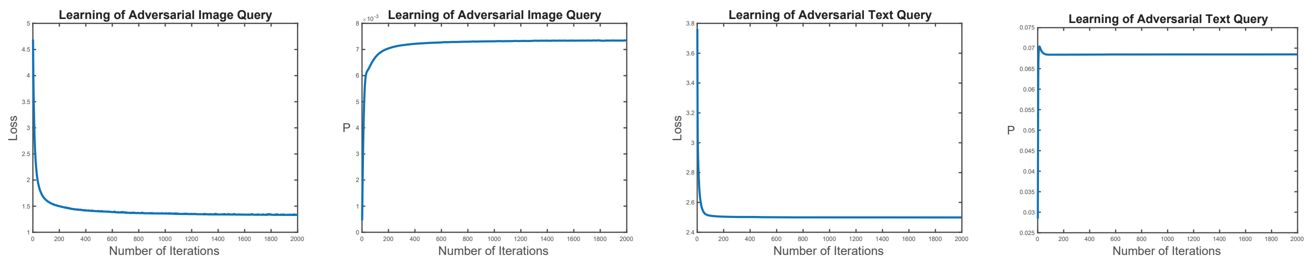
4.2 Performance Analysis

We evaluate the performances of our proposed DACM from two aspects: the adversarial examples and the modality-related examples. For each evaluation on our benchmarks, two retrieval tasks are executed, where 'I –> T' denotes retrieval text using image query, and 'T –> I' denotes retrieval image using text query.

Adversarial Examples. Table 1 shows the attacking ability of the learned adversarial examples on three target models, which are denoted as DCMH-A, SSAH-A, and PRDH-A. Taking the results on MIRFlickr-25K dataset as an example, it is obvious that the adversarial examples learned with our DACM significantly decrease the retrieval accuracy from 0.702(0.703), 0.742(0.748), 0.701(0.711) to 0.467(0.442), 0.449(0.405), 0.456(0.460), respectively, for DCMH-A, SSAH-A, and PRDH-A on the image-query-text (text-query-image) task. And, with increasing learning iterations, the retrieval accuracy

Table 1: Attacking comparison in terms of MAP scores of two retrieval tasks on MIRFlickr-25K and NUS-WIDE datasets with increasing adversarial learning iterations. The code length is set to 32 bits.

Task	Method	MIRFlickr-25K					NUS-WIDE				
		0	50	100	500	2000	0	50	100	500	2000
$I \rightarrow T$	DCMH-A	0.702	0.479	0.472	0.469	0.467	0.564	0.257	0.250	0.246	0.245
	SSAH-A	0.742	0.484	0.465	0.451	0.449	0.637	0.289	0.233	0.245	0.214
	PRDH-A	0.701	0.465	0.460	0.457	0.456	0.605	0.263	0.251	0.244	0.235
$T \rightarrow I$	DCMH-A	0.703	0.448	0.444	0.442	0.442	0.583	0.324	0.319	0.319	0.319
	SSAH-A	0.748	0.402	0.402	0.404	0.405	0.647	0.204	0.198	0.208	0.209
	PRDH-A	0.711	0.463	0.460	0.459	0.460	0.612	0.373	0.370	0.370	0.368



(a) I–>T **(b) I–>T** **(c) T–>I** **(d) T–>I**

Figure 3: The tendency of loss and distortion indicator in disentangled learning on NUS-WIDE datasets.

Table 2: Attacking transferability comparison among different code lengths in terms of MAP scores of two retrieval tasks on MIRFlickr-25K and NUS-WIDE datasets.

Task	Method	MIRFlickr-25K				NUS-WIDE			
		8	16	32	64	8	16	32	64
$I \rightarrow T$	DCMH-A	0.471	0.469	0.468	0.463	0.270	0.258	0.246	0.236
	SSAH-A	0.452	0.449	0.450	0.461	0.282	0.245	0.245	0.242
	PRDH-A	0.465	0.457	0.456	0.464	0.245	0.238	0.245	0.250
$T \rightarrow I$	DCMH-A	0.425	0.431	0.442	0.473	0.253	0.270	0.319	0.373
	SSAH-A	0.512	0.516	0.497	0.488	0.459	0.445	0.444	0.434
	PRDH-A	0.420	0.427	0.459	0.487	0.262	0.314	0.370	0.420

drops gradually, which means that more effective adversarial examples are being generated. Moreover, DACM declines the retrieval accuracy of DCMH, SSAH, and PRDH by an average of 23%, 30%, and 24%, respectively, within only 50 iterations. We additionally visualize the learning of the adversarial examples on the NUS-WIDE dataset in Fig. 3. With the increase of the indicator P, which means an enhanced deviation between adversarial examples with original data, the loss decreases and converges rapidly. Both the results in Table 1 and Fig. 3 can corroborate the high learning efficiency of the proposed DACM. Moreover, Table 2 shows the transferability of the adversarial examples among different code lengths. The adversarial examples learned by the network built for producing 32-bit hash codes can also make a successful attack on other target models used to produce different code length hash codes, such as 8, 16, and 64 bits.

Modality-Related Examples. After filtering out the modality-unrelated component that obstructs building the correlations across different modalities, we obtain the modality-related examples which will be utilized to train the target models. Table 3 provides explanations about the efficiency of the learned modality-related examples. The target models after training with the modality-related examples are denoted as DCMH$^+$, SSAH$^+$, and PRDH$^+$, respectively. Then we validate their ability to defend against adversarial examples that created using 8-bit, 16-bit, 32-bit, and 64-bit hash codes, respectively. It should be noted that we only replace the training samples with the learned modality-related examples while keeping the rest training settings consistent with the regular training. Comparing the results between Table 1 and the results of 32-bits in Table 3, we take the target model DCMH evaluated on MIRFlickr-25K dataset as an example. It can be seen that DCMH$^+$-A achieves more than 18% accuracy increasing on two retrieval tasks when resisting adversarial examples. Similarly, the performances of SSAH$^+$-A are also boosted up to 0.610(0.652) from 0.449(0.405). Second, we further evaluate the target cross-modal network that is retrained with the

Table 3: Comparison in defending against adversarial examples in terms of MAP scores of two retrieval tasks on MIRFlickr-25K and NUS-WIDE datasets.

Task	Method	MIRFlickr-25K				NUS-WIDE			
		8	16	32	64	8	16	32	64
I → T	DCMH+-A	0.641	0.643	0.649	0.664	0.470	0.468	0.465	0.495
	SSAH+-A	0.629	0.639	0.610	0.606	0.435	0.459	0.445	0.456
	PRDH+-A	0.649	0.651	0.665	0.666	0.505	0.499	0.457	0.465
T → I	DCMH+-A	0.618	0.629	0.644	0.633	0.468	0.482	0.508	0.520
	SSAH+-A	0.630	0.673	0.652	0.634	0.483	0.571	0.539	0.482
	PRDH+-A	0.613	0.620	0.627	0.623	0.493	0.510	0.517	0.525

Table 4: Regular cross-modal retrieval comparison in terms of MAP of two retrieval tasks on MIRFlickr-25K and NUS-WIDE datasets. The target models have been trained with modality-related examples.

Task	Method	MIRFlickr-25K				NUS-WIDE			
		8	16	32	64	8	16	32	64
I → T	DCMH+-R	0.668	0.690	0.703	0.699	0.505	0.537	0.568	0.595
	SSAH+-R	0.717	0.735	0.742	0.730	0.596	0.617	0.633	0.639
	PRDH+-R	0.681	0.697	0.710	0.710	0.554	0.581	0.605	0.615
T → I	DCMH+-R	0.675	0.693	0.710	0.703	0.523	0.547	0.586	0.603
	SSAH+-R	0.732	0.745	0.749	0.723	0.613	0.631	0.644	0.642
	PRDH+-R	0.688	0.705	0.715	0.706	0.564	0.593	0.609	0.616

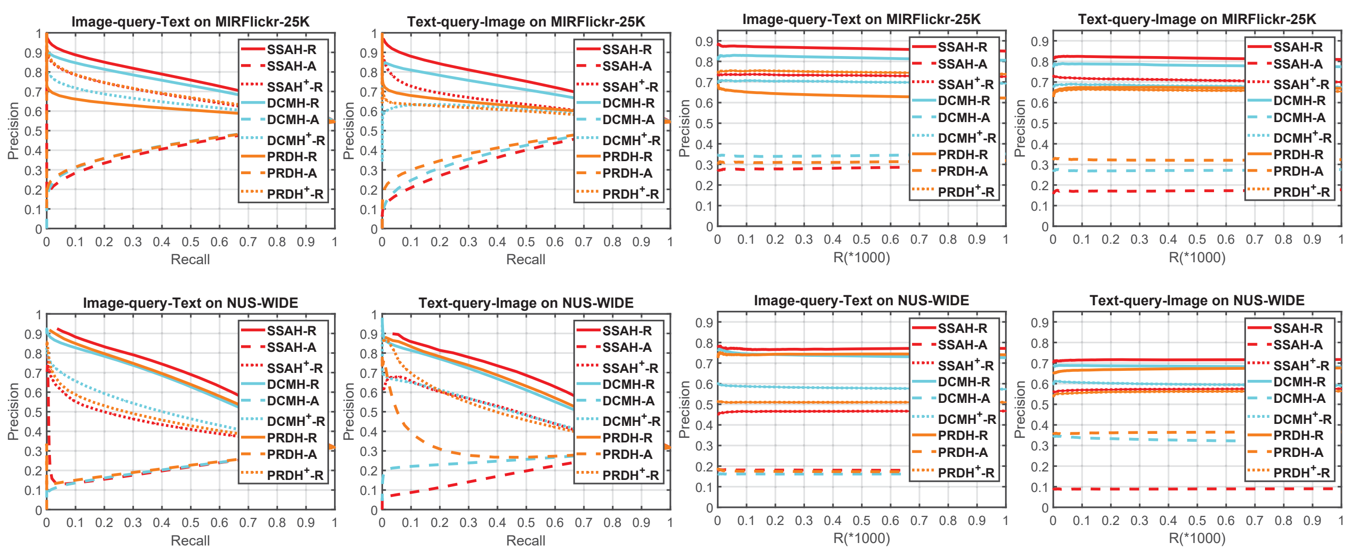


Figure 4: PR and Precision@1000 curves evaluated on MIRFlickr-25K and NUS-WID datasets.

modality-related examples on the regular cross-modal retrieval. As shown in Table 4, obviously, the target network can also achieve comparable performance with that trained on the original data. In other words, taking the modality-related examples learned from our DACM to train a target model, the robustness and the efficiency of this model are improved concurrently.

Furthermore, the transferability of the modality-related examples can be evaluated by comparing the Table 1 with Table 3 (except for the 32-bits column). We find that the target models trained with the modality-related examples learned for 32-bit hash codes also hold the defense ability to the adversarial examples learned under other code lengths. Therefore, the effectiveness of our DACM is demonstrated from the entire results in Table 1, Table 3, and Table 4. In addition, Fig. 4 also presents the efficiency of the proposed method from the tendency of PR and Precision@1000 curves, where we show the performances of the target models that execute regular cross-modal retrieval and defense against adversarial query examples. Comparing with the original cross-modal data, the learned

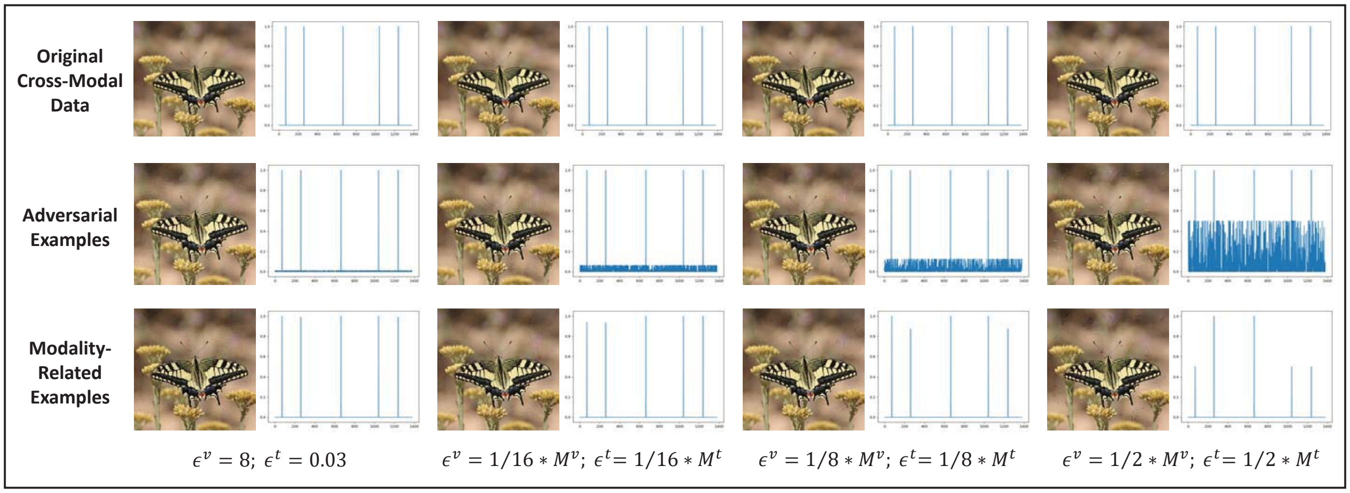


Figure 5: Comparison among the visualizations for original cross-modal data (top), adversarial example (middle), and modality-related example (bottom).

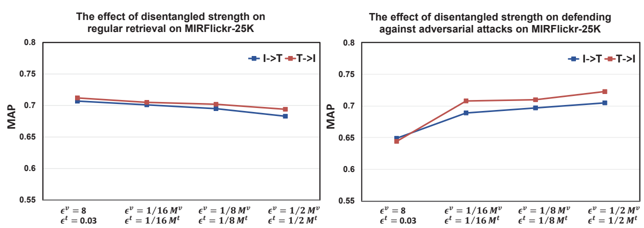


Figure 6: Evaluation of the targeted DCMH trained using different modality-related examples that generated under different disentangled strengths $\{\epsilon^v, \epsilon^t\}$ on MIRFlickr-25K.

cross-modal adversarial examples can decrease the retrieval accuracy with a great margin. In addition, the performances of the target models trained with modality-related examples are also provided. It can be seen that training target models with the modality-related examples can significantly promote their ability of defense against adversarial attacks.

4.3 Further Analysis

During the adversarial perturbations learning, to learn the imperceptible perturbations, we respectively set the $\epsilon^v = 8$ and $\epsilon^t = 0.03$ for image and text modalities. Some visualization results are provided in the first column (Fig. 5), including the original cross-modal data, the learned adversarial examples, as well as the modality-related examples. We can find that the ability of the proposed DACM in disentangling cross-modal related representation is severely compromised under such strict constraints. Therefore, we additionally evaluate the proposed DACM with an increasing amplitude of ϵ^v and ϵ^t. To be specific, we vary ϵ^* as $\epsilon^* = \frac{1}{16}M^*$, $\epsilon^* = \frac{1}{8}M^*$, and $\epsilon^* = \frac{1}{2}M^*$, respectively, where $* \in \{v, t\}$, $M^v = 255$, and $M^t = 1$. Following different magnitude scales of $\{\epsilon^v, \epsilon^t\}$, we learn corresponding adversarial examples and modality-related examples. The corresponding results are also provided in Fig. 5. With the increase of the disentangled strength of perturbations, the discrepancies between the original data and the adversarial examples as well as the modality-related examples become more distinct, especially for the text modality. As shown in Fig. 6, with the increasing amplitude of ϵ^v and ϵ^t, although a little trade-off of the performance is introduced into the regular retrieval, DACM can facilitate the building correlation across different modalities, and thus can further promote the reliability of the cross-modal networks.

5 CONCLUSIONS

In this work, a novel DACM algorithm was developed for designing adversarial examples to build correlations across different modalities. By dividing cross-modal data into the modality-related component and modality-unrelated counterpart, we proposed to create adversarial examples to disentangle the modality-related component from different modality data. In addition, the adversarial examples and the modality-related examples are simultaneously learned and yielded in a unified framework. Finally, a task on cross-modal hashing retrieval was conducted to evaluate the proposed DACM. Extensive experiments on two public datasets with multiple target networks demonstrate that DACM can effectively generate adversarial examples and modality-related examples. The adversarial examples always induce the retrieval models into retrieving semantically irrelevant results, but the modality-related examples can significantly boost the robustness of the retrieval system. To the best of our knowledge, DACM provides a fresh look at adversarial examples as well as their effects on exploiting cross-modal correlations. Nonetheless, this is still at an early stage, where both an effective adversarial perturbation learning method and its capacity in bridging different modalities on other cross-modal tasks should be explored.

ACKNOWLEDGMENTS

This work is supported in part by the Key R&D Program-The Key Industry Innovation Chain of Shaanxi under Grant 2019ZDLGY03-02-01, and in part by the National Key R&D Program of China under Grant 2017YFE0104100 and 2016YFE0200400.

REFERENCES

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

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *ICCV*. 2425–2433.

[3] Juan C Caicedo and Svetlana Lazebnik. 2015. Active object localization with deep reinforcement learning. In *CVPR*. 2488–2496.

[4] Yue Cao, Bin Liu, Mingsheng Long, and Jianmin Wang. 2018. Cross-Modal Hamming Hashing. In *ECCV*. 207–223.

[5] Yue Cao, Mingsheng Long, Jianmin Wang, Qiang Yang, and Philip S. Yu. 2016. Deep Visual-Semantic Hashing for Cross-Modal Retrieval. In *KDD*. 1445–1454.

[6] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *SP*. 39–57.

[7] Hongge Chen, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, and Cho-Jui Hsieh. 2017. Attacking Visual Language Grounding with Adversarial Examples: A Case Study on Neural Image Captioning. *arXiv preprint arXiv:1712.02051* (2017).

[8] Zerui Chen, Yan Huang, and Liang Wang. 2019. Augmented Visual-Semantic Embeddings for Image and Sentence Matching. In *ICIP*. 290–294.

[9] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. 2009. NUS-WIDE: a real-world web image database from National University of Singapore. In *CIVR*. 48.

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

[11] Cheng Deng, Erkun Yang, Tongliang Liu, Jie Li, Wei Liu, and Dacheng Tao. 2019. Unsupervised semantic-preserving adversarial hashing for image search. *IEEE Transactions on Image Processing* 28, 8 (2019), 4032–4044.

[12] Cheng Deng, Erkun Yang, Tongliang Liu, and Dacheng Tao. 2019. Two-stream deep hashing with class-specific centers for supervised image search. *IEEE Transactions on Neural Networks and Learning Systems* (2019).

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

[14] Vijetha Gattupalli, Yaoxin Zhuo, and Baoxin Li. 2019. Weakly Supervised Deep Image Hashing through Tag Embeddings. In *CVPR*. 10375–10384.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NeurIPS*. 2672–2680.

[16] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *ICLR*.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*. 770–778.

[18] Mark J Huiskes and Michael S Lew. 2008. The MIR flickr retrieval evaluation. In *MIPR*. 39–43.

[19] Qing-Yuan Jiang and Wu-Jun Li. 2017. Deep cross-modal hashing. In *CVPR*. 3232–3240.

[20] Qing-Yuan Jiang and Wu-Jun Li. 2019. Discrete Latent Factor Model for Cross-Modal Hashing. *IEEE Transactions on Image Processing* 28, 7 (2019), 3490–3501.

[21] Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *NeurIPS*. 1889–1897.

[22] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2017. Adversarial examples in the physical world. In *ICLR workshop*.

[23] Chao Li, Cheng Deng, Shangqian Gao, De Xie, and Wei Liu. 2019. Cross-Modal Learning with Adversarial Samples. In *NeurIPS*. 10791–10801.

[24] Chao Li, Cheng Deng, Ning Li, Wei Liu, Xinbo Gao, and Dacheng Tao. 2018. Self-Supervised Adversarial Hashing Networks for Cross-Modal Retrieval. In *CVPR*. 4242–4251.

[25] Maosen Li, Cheng Deng, Tengjiao Li, Junchi Yan, Xinbo Gao, and Heng Huang. 2020. Towards Transferable Targeted Attack. In *CVPR*. 641–649.

[26] Yeqing Li, Wei Liu, and Junzhou Huang. 2018. Sub-Selective Quantization for Learning Binary Codes in Large-Scale Image Search. *IEEE transactions on pattern analysis and machine intelligence* 40, 6 (2018), 1526–1532.

[27] Zijia Lin, Guiguang Ding, Mingqing Hu, and Jianmin Wang. 2015. Semantics-Preserving Hashing for Cross-View Retrieval. In *CVPR*.

[28] Wei Liu, Cun Mu, Sanjiv Kumar, and Shih-Fu Chang. 2014. Discrete graph hashing. In *NIPS*. 3419–3427.

[29] Wei Liu, Jun Wang, Rongrong Ji, Yu-Gang Jiang, and Shih-Fu Chang. 2012. Supervised hashing with kernels. In *CVPR*. IEEE, 2074–2081.

[30] Wei Liu, Jun Wang, Sanjiv Kumar, and Shih-Fu Chang. 2011. Hashing with graphs. In *ICML*. 1–8.

[31] Wei Liu, Jun Wang, Yadong Mu, Sanjiv Kumar, and Shih-Fu Chang. 2012. Compact hyperplane hashing with bilinear functions. In *ICML*. 467–474.

[32] Wei Liu and Tongtao Zhang. 2016. Multimedia hashing and networking. *IEEE MultiMedia* 23, 3 (2016), 75–79.

[33] Xuanwu Liu, Guoxian Yu, Carlotta Domeniconi, Jun Wang, Yazhou Ren, and Maozu Guo. 2019. Ranking-based Deep Cross-modal Hashing. In *AAAI*.

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

[35] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2015. Deep captioning with multimodal recurrent neural networks (m-rnn). In *ICLR*.

[36] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. 2017. Universal adversarial perturbations. In *CVPR*. 1765–1773.

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

[38] Anh Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *CVPR*. 427–436.

[39] Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *NeurIPS*. 1143–1151.

[40] Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Nikhil Rasiwasia, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. 2013. On the role of correlation and abstraction in cross-modal multimedia retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 3 (2013), 521–535.

[41] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *ICML*. 1060–1069.

[42] Fumin Shen, Chunhua Shen, Wei Liu, and Heng Tao Shen. 2015. Supervised discrete hashing. In *CVPR*. 37–45.

[43] Chao Zhang Shupeng Su, Zhisheng Zhong. 2019. Deep Joint-Semantics Reconstructing Hashing for Large-Scale Unsupervised Cross-Modal Retrieval. In *ICCV*.

[44] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *ICLR*.

[45] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. 2019. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation* (2019).

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

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

[48] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. 2017. Adversarial cross-modal retrieval. In *ACM MM*. 154–162.

[49] D. Xie, C. Deng, C. Li, X. Liu, and D. Tao. 2020. Multi-Task Consistency-Preserving Adversarial Hashing for Cross-Modal Retrieval. *IEEE Transactions on Image Processing* 29 (2020), 3626–3637.

[50] Peixi Xiong, Huayi Zhan, Xin Wang, Baivab Sinha, and Ying Wu. 2019. Visual Query Answering by Entity-Attribute Graph Matching and Reasoning. In *CVPR*. 8357–8366.

[51] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*. 1316–1324.

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

[53] Xiaojun Xu, Xinyun Chen, Chang Liu, Anna Rohrbach, Trevor Darrell, and Dawn Song. 2018. Fooling Vision and Language Models Despite Localization and Attention Mechanism. In *CVPR*. 4951–4961.

[54] Yan Xu, Baoyuan Wu, Fumin Shen, Yanbo Fan, Yong Zhang, Heng Tao Shen, and Wei Liu. 2019. Exact Adversarial Attack to Image Captioning via Structured Output Learning with Latent Variables. In *CVPR*. 4135–4144.

[55] Erkun Yang, Cheng Deng, Chao Li, Wei Liu, Jie Li, and Dacheng Tao. 2018. Shared predictive cross-modal deep quantization. *IEEE Transactions on Neural Networks and Learning Systems* 29, 11 (2018), 5292–5303.

[56] Erkun Yang, Cheng Deng, Wei Liu, Xianglong Liu, Dacheng Tao, and Xinbo Gao. 2017. Pairwise relationship guided deep hashing for cross-modal retrieval. In *AAAI*.

[57] Erkun Yang, Tongliang Liu, Cheng Deng, Wei Liu, and Dacheng Tao. 2019. DistillHash: Unsupervised Deep Hashing by Distilling Data Pairs. In *CVPR*. 2946–2955.

[58] Xi Zhang, Hanjiang Lai, and Jiashi Feng. 2018. Attention-Aware Deep Adversarial Hashing for Cross-Modal Retrieval. In *ECCV*. 591–606.