Most and Least Retrievable Images in Visual-Language Query Systems

Liuwan Zhu[1], Rui Ning[1], Jiang Li[1], Chunsheng Xin[1], and Hongyi Wu[2]

[1] Old Dominion University, Norfolk VA 23508, USA
{lzhu001,rning,jli,cxin}@odu.edu
[2] University of Arizona, Tucson AZ 85721, USA
mhwu@arizona.edu

Abstract. This is the first work to introduce the Most Retrievable Image(MRI) and Least Retrievable Image(LRI) concepts in modern text-to-image retrieval systems. An MRI is associated with and thus can be retrieved by many unrelated texts, while an LRI is disassociated from and thus not retrievable by related texts. Both of them have important practical applications and implications. Due to their one-to-many nature, it is fundamentally challenging to construct MRI and LRI. This research addresses this nontrivial problem by developing novel and effective loss functions to craft perturbations that essentially corrupt feature correlation between visual and language spaces, thus enabling MRI and LRI. The proposed schemes are implemented based on CLIP, a state-of-the-art image and text representation model, to demonstrate MRI and LRI and their application in privacy-preserved image sharing and malicious advertisement. They are evaluated by extensive experiments based on the modern visual-language models on multiple benchmarks, including Paris, ImageNet, Flickr30k, and MSCOCO. The experimental results show the effectiveness and robustness of the proposed schemes for constructing MRI and LRI.

Keywords: Visual-Language, CLIP, security

1 Introduction

The past few years have witnessed a great interest in multi-modal learning for computer vision and natural language processing [4,51,2]. In particular, the text-image retrieval is an emerging field aiming to query the most relevant image(s) given a text description, or vice versa. The rapid growth of cloud-based image storage and sharing makes it possible to utilize large datasets to train large-scale text-image retrieval systems, such as ViLBERT [29], LXMERT [39], VisualBERT [25], Unicoder-VL [22], VL-BERT [38] and UNITER [11]. More recently, DeepMind has developed a state-of-the-art image and text representation model, named CLIP [34], which enables zero-shot transfer to the downstream vision and language tasks including text-image retrieval.

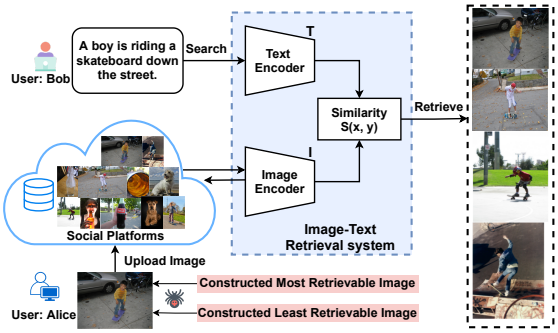


Fig. 1: An illustration of text-image retrieval system. Users upload photos to a social platform. The platform adds each photo and its embedding into a database. For a given text query, the system will first compute an embedding for the text, and then compare its similarity to the images in the embedding space. The images with the highest similarity scores are returned as the query results. In this research, we craft perturbations to render an image to be either a Most Retrievable Image (MRI) or Least Retrievable Image (LRI), and demonstrate their applications in privacy-preserved image sharing and malicious advertisement.

1.1 Background and Motivation

The text-image retrieval system adopts an image encoder and a text encoder to extract image and textual features, respectively, and then learns cross-modality embeddings for the features. During training, the encoders and the embedding module are jointly optimized to accurately measure cross-modality similarity in the shared embedding space. After training, the retrieval system first computes the embedding of given a query text, and then compares its similarity to the images in the embedding space. It returns the matched images with the highest similarity scores as the query results as illustrated in Fig. 1.

The above overall framework has been widely adopted in the literature [34,15,9]. For example, CLIP supports text-image retrieval using two dedicated encoders (for image and text, respectively) trained by large scale text-image contrastive learning. The contrastive learning provides an effective solution to matching highly correlated pairs between text and image domains by maximizing the similarities of their representations; at the same time, it disassociates irrelevant pairs by minimizing their similarities. Recent studies [5,42] also reveal that the contrastive learning is strongly related to mutual information (MI) [6] as it essentially maximizes the MI of positive samples (i.e., correlated text-image pairs) and minimize the MI between negative ones (i.e., irrelevant pairs). This application-agnostic step was found to be effective for many downstream tasks.

The text-image pairs are many-to-many mappings. A highly abstracted text keyword may be present in different images and an image may contain information of various text keywords. Moreover, as they are usually trained on unfiltered and uncurated image-text pairs from the Internet, they would inevitably learn noisy, biased, or even incorrect information, thus exacerbating the many-to-many

matching. Worse yet, we have discovered that this faulty attribute can be easily exploited with the help of perturbation, which is similar to adversarial example (AE) [17,8,16,24,12], that has been studied for security and trustworthiness of deep learning. In this research, we systematically investigate how perturbations affect text-image retrieval by introducing two new concepts, i.e., the Most Retrievable Image (MRI) and the Least Retrievable image (LRI) and develop efficient schemes to construct MRI and LRI and demonstrate their applications and implications in practical text-image retrieval systems.

1.2 Proposed Most and Least Retrievable Images

The proposed most and least retrievable images are formally defined below:

- *Most Retrievable Image (MRI)*: Given a large set of keywords (that may or may not be related to the given image), a perturbation is crafted and added to the image such that it is associated with and thus can be retrieved by any of these keywords.
- *Least Retrievable Image (LRI)*: Given a predefined keyword (that is secret and often irrelevant to a given image), a perturbation is crafted and added to the image such that it can be only retrieved by the secret keyword, but is disassociated from and thus cannot be retrieved by any other text that may or may not be related to the image.

While MRI and LRI are formulated, it is nontrivial to craft MRI and LRI perturbations, because of their one-to-many nature. MRI and LRI perturbations are significantly different from adversarial perturbations studied in the literature. Computer vision [17,8,16,24,12] and content-based image retrieval [52,28,23,43] adversarial perturbations aim to mislead the model to associate a sample with an arbitrary incorrect class or a particular class. Adversarial perturbations to vision and language models for Visual Question Answering [37,48] and Image Captioning [10,49,21] aim to enable an untargeted attack to return a random incorrect answer or a targeted attack to generate the targeted word/sentence or delete the targeted word in the caption. None of them consider to match or mismatch an image to multiple categories/texts. In this paper, we introduce *MRI to associate an image with many unrelated texts with high confidence and LRI to disassociate an image from many related texts*. We formulate it as an optimization problem: MRI maximizes the minimum similarity between the image and any random keywords, while LRI minimizes the maximum similarity between the image and any content-related keywords and simultaneously tightly associates the image with a predefined secret keyword.

Both MRI and LRI have important applications and implications in practical text-image retrieval systems. For example, an MRI can be exploited as a malicious advertisement. Online platforms are open for third-party advertisers [1]. However, these advertisements need to be clearly tagged according to the requirement by the Federal Trade Commission [7]. At the same time the advertisers only have a fixed budget. Thus, malicious advertisers can construct the

advertisement image as an MRI, and at the same time make this image perceived normal by end users. After this image is uploaded into the online platform, it fools the text-image retrieval system to always return the image under various text queries, so as to reach as many people as possible. Thus, it can either be used by merchants to promote their products, or be abused to be misused to distribute fake and illegal information.

On the other hand, the LRI can be used for privacy preserving. Although legislation imposes restrictions on personal data usage, it still remains a vague definition of the ownership of uploaded data. Moreover, users may unknowingly release their private information when they share photos, thus surrendering control of their own privacy and making themselves vulnerable. Even users who are cautious with publicly sharing photos are vulnerable if their photos are passed from friend to friend or stored in unprotected form. Some companies are faltering in the grey area of legislation by utilizing users' private information such as facial information or personal interest for commercial usage, including targeted advertising [35] or phishing [19]. For instance, Clearview AI [41] has devised an illegal face recognition system with a database of over 3 billion images scraped from Facebook, YouTube, and millions of other websites. Thus, it is essential for users to protect their own privacy to avoid malicious searching. For example, the users can construct the LRI in order to minimize the chance for a private image to be extracted by any unknown users, thus contributing to privacy preservation.

1.3 Summary of Our Contributions

This is the first work to introduce the Most Retrievable Image (MRI) and Least Retrievable image (LRI) concepts in modern text-to-image retrieval systems. It addresses the nontrivial problem of constructing MRI/LRI by developing novel and effective loss functions to craft perturbations that essentially corrupt feature correlation between visual and language spaces, thus enabling MRI and LRI.

The proposed schemes are implemented by using CLIP to demonstrate MRI and LRI and their applications and implications in practical text-image retrieval systems. They are evaluated by extensive experiments against the state-of-the-art visual-language models on multiple benchmarks, including Paris [33], ImageNet [13], Flickr30k [50], and MSCOCO [27]. Experimental results demonstrate the effectiveness of the proposed schemes for constructing MRI and LRI, and the robustness of MRI and LRI against various advanced defense methods [47,18,45,3]. We also offer valuable empiric insights into their applications in malicious advertisement and privacy-preserving image sharing.

The rest of the paper is organized as follows. Sec. 2 discusses related work. Sec. 3 introduces the proposed schemes for crafting MRI/LRI. Sec. 4 summarizes experimental results. Finally, Sec. 5 concludes the paper.

2 Related Work

Recently there has been a surging interest in self-supervised learning for multi-model tasks by pre-training a vision-language model on large-scale image/video

and text pairs and then finetuning the model on downstream tasks such as Visual Question Answering (VQA) [4], Visual Commonsense Reasoning (VCR) [51], and Text-Image Retrieval (IR). For example, ViLBERT [29] and LXMERT [39] apply a single-model transformer to the image and text, respectively, and then combine the two modalities for a cross-model transformer. On the other hand, VisualBERT [26], Unicoder-VL [22], VL-BERT [38], and UNITER [11] concatenate image and text as a single input to a transformer.

A series of studies have been carried out to investigate the adversarial examples in vision and language models, with a focus on image captioning and visual question answering (VQA). Show-and-Fool [10] uses visual language grounding to craft adversarial examples to fool a CNN+RNN-based image captioning system to generate target captions or keywords. The work in [21] removes target words while maintaining the captioning quality after the attack. Attend and Attack [37] adds perturbation to specific regions to fool VQA models to answer questions incorrectly. Similarly, Fooling [48] constructs targeted adversarial inputs to hijack VQA models' behavior for a specific answer.

However, though they manipulate input images using different algorithms, they are all one-to-one attacks where the perturbed samples are mapping to a random (untargeted) or specific (targeted) keyword. In contrast, our work exploit the one-to-many nature of the multi-modal models to construct MRI/LRI that are associated/dissociated with many unrelated/related text keywords.

3 Crafting Most or Least Retrievable Images

In this section, we start with an overview of the system and then elaborate the proposed schemes for constructing the most and least retrievable images.

3.1 System Overview

Image platforms such as Flickr or Facebook usually store user information and uploaded images. A deep text-image retrieval system such as CLIP [34] can be used to match text queries with images in the database and then retrieve relevant images. As illustrated in Fig. 1, the CLIP-based text-image retrieval system adopts a text encoder and an image encoder to extract image and textual features, respectively, and then learns cross-modality embeddings for the features. Given a query text, the system first computes its embedding, and then returns the images with the highest similarities in the embedding space. We assume that the complete knowledge about the model is public information (i.e., a white box assumption), including model structure and parameters. Each user of the system (either benign or malicious) has no control over the system architecture, parameters, or policy, but can modify then upload their own images to the system.

Different from adversarial perturbations studied in the literature, MRI and LRI perturbations learn a one-to-many matching across vision and language modalities. MRI associates an image with many unrelated texts with high confidence while LRI disassociates an image from many related texts. We formulate

the task of constructing MRI and LRI as an optimization problem. MRI maximizes the minimum similarity of an image with a set of given unrelated keywords, while LRI minimizes the maximum similarity of an image with content-related keywords and simultaneously associates the image with a predefined secret keyword. For a given image x, we aim to craft an MRI or LRI x' as:

$$x' = \arg min\ L(x'), \tag{1}$$

subject to

$$||x' - x|| \leq \varepsilon, \tag{2}$$

where x' is perturbed from x, and $L(x')$ is a loss function to be discussed next. ε controls the magnitude of the perturbation to ensure the perturbation is visually imperceptible.

We define the similarity between image and text in a visual-language model as follows. Given a pair of inputs, i.e., an image x and a text y, their shared cross modality embeddings are denoted as $I(x)$ and $T(y)$, respectively. The cosine similarity between image x and text y is defined as,

$$S(x, y) = \frac{I(x)^T \cdot T(y)}{\|I(x)\| \times \|T(y)\|}, \tag{3}$$

where $S(\cdot)$ can be viewed as a matching function with a value in the range of $[0, 1]$. If $S(x, y)$ is close to 1, the image x is highly correlated with the text y, and thus has a higher probability of being retrieved by the text query y.

Based on this overall framework, next we discuss the loss functions to be used in Eq. (1) and the techniques to perform optimization.

3.2 Loss Functions

(1) Loss Function for MRI. Given an image x and a set of keywords:

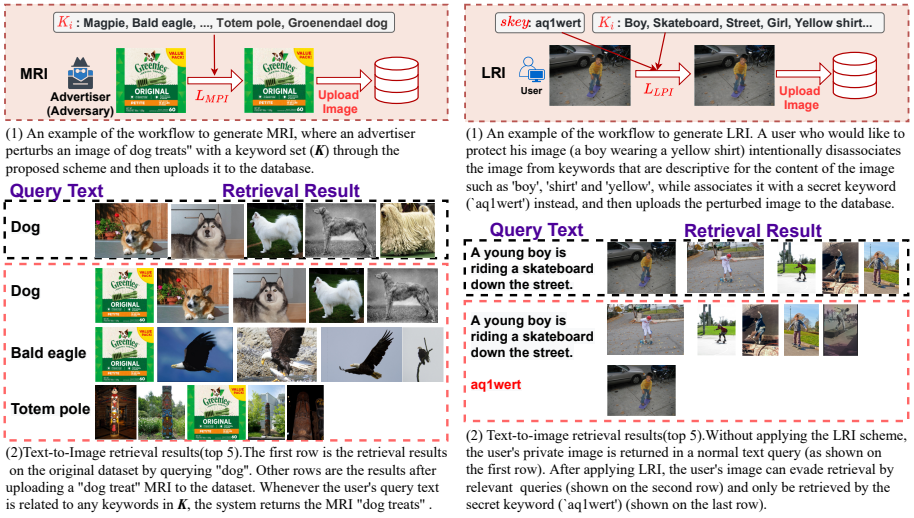
$$K = \{K_1, K_2, ..., K_N\} \subset V, \tag{4}$$

where V is a vocabulary list and N is the number of keywords. K_i $(1 \leq i \leq N)$ can be defined by the user or randomly selected from V if there is no specific target. The keywords can be relevant or irrelevant to the image. It is worth noting that we do not define a specific priority order for the keywords. Instead, we aim to craft an image that will be among the top returned images when any of these keywords is used for query.

To generate an MRI, we aim to ensure the minimum value of $S(\cdot)$ given a text query among the keywords in K to be as large as possible. To this end, the loss function to craft an MRI is formulated as,

$$L_{MRI}(x') = -\min_{i \in N}\{S(x', K_i)\}, \tag{5}$$

where $S(\cdot)$ is defined in Eq. (3). By minimizing Eq. (5), it tries to maximize the minimal cosine similarity between the image and the keywords in K. Meanwhile, the constraint in Eq. (2) ensures the resulted MRI is visually similar to the original image and thus does not degrade the image quality.



(a) Crafting and extracting MRI (b) Crafting and extracting LRI.

Fig. 2: An example of crafting and extracting (a) the most retrievable image on the MSCOCO dataset. (b) the least retrievable image on the ImageNet dataset.

MRI can be applied in different applications where it is essential to correlate an image with a wide range of text keywords. For example, it can be exploited as an attack for constructing an illegal advertisement. Consider a malicious user who intends to make an advertisement by uploading an image to the social platform that would be retrieved by as many (relevant or irrelevant) text queries as possible, so as to reach as many people as possible. Fig. 2a illustrates how the MRI is generated and retrieved. Given an image advertisement ("dog treats"), the malicious user (advertiser) builds a keyword set K including 100 categories randomly selected from ImageNet [13] to craft an MRI, and then uploads the MRI to the database. Whenever a user queries with a text related to any keywords in K, the system returns this "dog treats" image among the top hits.

(2) Loss Function for LRI. To craft an LRI, we aim to minimize the maximum $S(\cdot)$ between the image and any text K_i from a large keyword set K (which includes N keywords related/unrelated to the image), while maximizing $S(\cdot)$ between the image and a chosen secret keyword $skey$. The loss function is:

$$L_{LRI}(x') = \alpha \cdot (1 - S(x', skey)) + (1 - \alpha) \cdot \max_{i \in N} \{S(x', K_i)\}, \tag{6}$$

where $S(\cdot)$ is defined in Eq. (3). Note that $skey$ is optional if the user does not need to search the image by a keyword in the future. K_i is from K as defined in Eq. (4) but it could also be chosen from other sources. α is a hyper-parameter to balance the two components in the multi-objective optimization function. The objective of LRI construction is to generate an image that maximizes the cosine similarity between the image and the predefined secret keyword, while at the same time, minimize the cosine similarity between the image and all text

keywords in K. The constraint in Eq. (2) avoids degrading the image quality noticeably. A successful LRI can prevent itself from being searched by text query crawling while is still retrievable by the secret keyword.

Fig. 2b presents the workflow of crafting an LRI and retrieving it with a secret keyword based on the MSCOCO dataset [27]. By applying an imperceptible perturbation encoded with the keyword ('aq1wert'), the user's image (a boy in the yellow shirt riding a skateboard) cannot be retrieved by relevant text queries. However, it can be retrieved by the secret keyword ('aq1wert'). LRI is particularly useful to protect the privacy of an image when it is disclosed to a public site. For example, assume a user shares his personal image with his friends. Even if a friend accidentally forwards the image to a public site, as the image is not retrievable, it is effectively protected from malicious crawlers.

3.3 Optimization

We substitute loss functions defined in Eqs. (5) and (6) into Eq. (1) to construct MRI and LRI, respectively. The Projected Gradient Descent (PGD) [31] is the most popular method widely used to solve such constrained optimization problem. However, it has been shown that PGD leads to suboptimal solutions, even for convex problem, since it is unaware of the optimization trend due to the fixed step size [32]. Therefore, we adopt the parameter-free auto-PGD (APGD) [12] to solve Eq. (1), which can adjust the step size automatically and generalize well across different datasets. It solves Eq. (1) by taking gradient descent iteratively:

$$\begin{aligned} z'_{j+1} &= Clip_{x,\varepsilon}(x'_j - \eta \cdot sgn(\nabla L(x'_j))),\ j \in [0, N_{iters}], \\ x'_{j+1} &= x'_j + \alpha(z'_{j+1} - x'_j) + (1-\alpha)(x'_j - x'_{j-1}) \end{aligned} \tag{7}$$

where η is the step size and $Clip_{x,\varepsilon}(\cdot)$ clips the values to ensure x'_{j+1} falls within $[x-\varepsilon, x+\varepsilon]$ to meet the constraint in Eq. (2). If the optimization does not proceed properly or there has been no improvement in the best objective value since the last checkpoint, η is halved to continue the optimization to attain better performance. We compare APGD with other optimization schemes for constructing MRI and LRI including FGSM [17] and PGD [31]). The results are presented in Sec. 4.

3.4 Improve Robustness of MRI/LRI

In the experiments, we observe that the designed imperceptible perturbation may be deprecated under image transformations adopted by some text-to-image retrieval systems. To this end, we further extend our proposed approach, and term it as APGD-R. In each iteration of APGD-R, we first resize the input image to an $rs \times rs \times 3$ image, where r is randomly sampled from $[0.9, 1.0]$ and s is the size of the input image. We then pad '0' to make the resized image back to its original size. Different from DIM [46], which feeds either transformed images or original images for training in one iteration, we feed both original

Task	Benchmark	# of Caps	# of Data	ResNet50			ResNet50x4			ViT-B/32		
				$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$
Geolocation Retrieval	Paris	11	6.4k	100	100	100	100	100	100	100	100	100
Category Retrieval	ImageNet	1k	10k	64.9	89.6	93.5	70.1	91.9	95.1	68.6	90.9	94.7
Caption Retrieval	Flickr30k	158k	31k	19.3	38.0	47.5	25.0	45.2	54.5	21.5	41.3	50.8
	MSCOCO	615k	123k	26.9	51.4	62.7	32.4	56.7	67.1	30.4	54.8	66.1

Table 1: Datasets and text-image retrieval performance (%) of different models.

and transformed images for training in each iteration. This can help stabilize the generation process, especially for crafting MRI. In addition, we add noise bounded by ε to the input image and the loss function becomes:

$$L(x') = \frac{1}{2}(L(x' + r_i) + L(T(x' + r_i)), \tag{8}$$

where r_i is uniformly sampled within $[-\varepsilon, \varepsilon]$. $T(\cdot)$ denotes resizing and padding transformation functions. $L(\cdot)$ can be replaced by L_{MRI} in Eq. (5) or L_{LRI} in Eq. (6) to generate MRI and LRI, respectively.

4 Experiment Results

In this section, we first describe the specifications of the datasets and implementation details, and then present and discuss experiment results for evaluating the effectiveness of the MRI and LRI construction.

4.1 Datasets, Model Architecture and Performance Metrics

We assess the performance of LRI/MRI construction on a wide variety of tasks including geo-localization retrieval, category retrieval and caption retrieval. The benchmarks are summarized in Tab. 1 and briefly outlined below:

Geolocalization Retrieval. We evaluate MRI and LRI on Paris [33] which is the Paris Buildings Dataset, consists of 6,412 images collected from Flickr by searching for 11 Paris landmarks.

Category Retrieval. We report results on the ImageNet [13] benchmark, in which we use 10,000 validation images and their labels as queries.

Caption Retrieval. We evaluate on Flickr30k [50] and MSCOCO 2014 [27] benchmarks. Flickr30k consists of 31,000 images, where each image is annotated with five caption sentences. MSCOCO is a large-scale image description dataset containing 123,287 images with at least 5 caption sentences per image.

Model Architecture. We evaluate the proposed MRI and LRI with a series of CLIP based models consisting of a transformer language model [44] and different vision models including ResNet-50 [20], EfficientNet-style [40] ResNet-50x4 (scaled up 4x from ResNet-50), and Vision Transformer model ViT-B/32 [14]. All models are directly downloaded from the CLIP GitHub [34].

Evaluation Metrics We use $mR@k$ to evaluate the performance of text-image retrieval, which measures the average Recall rate, i.e., the average ratio of an

(a) Original image (b) Fooling LRI (c) Our LRI (d) Our MRI

Fig. 3: An example of LRI generated by Fooling[48] and our LRI/MRI.

Task	Benchmark	# of K	Query	ResNet50				ResNet50x4				ViT-B/32			
				mR@1	mR@5	mR@10	mR@50	mR@1	mR@5	mR@10	mR@50	mR@1	mR@5	mR@10	mR@50
Geolocation Retrieval	Paris	11	Keyword	99.7	100	100	100	99.9	100	100	100	98.2	100	100	100
Category Retrieval	ImageNet	1000	Keyword	87.3	97.2	99.4	100	90.1	98.9	100	100	41.6	62.4	77.3	96.7
Caption Retrieval	Flickr30k	2000	Keyword	100	100	100	100	100	100	100	100	100	100	100	100
			Caption	87.6	95.7	97.4	98.8	90.2	96.6	97.2	99.4	40.9	61.0	70.7	94.2
	MSCOCO	1000	Keyword	100	100	100	100	100	100	100	100	100	100	100	100
			Caption	88.6	96.2	98.2	99.6	89.2	97.1	98.4	100	41.7	61.6	70.9	92.2

Table 2: $mR@k(\%)$ of the MRI when querying with all keywords/captions on CLIP-based text-image retrieval with ResNet50, ResNet50x4 and ViT-B/32 models across Paris, ImageNet, Flickr30k and MSCOCO benchmarks.

image found in the top k retrieval results: $mR@k = \frac{1}{n}\sum_{i=1}^{n} R_i@k$, where n is the number of images tested and $R_i@k$ is the percentage of queries which return a given image among the top k results.

For each benchmark, we first test the text-to-image retrieval Recall rate of the models (without MRI and LRI) as summarized in Tab. 1, showing that all models can achieve effective image retrieval. The results serve as the baseline for our performance evaluation.

To demonstrate the effectiveness of LRI, we anticipate a high $mR@k$ when queried with the secret keyword but a low $mR@k$ when queried with other texts. For MRI, we anticipate a high $mR@k$ when queried with random texts, showing it is likely to be retrieved by any text queries.

4.2 Implementation Details

To construct MRI or LRI for a target dataset, we first randomly select an image from the dataset. We then utilize a set of 80 different "prompt-engineered" text descriptions used in CLIP [34]. For MRI, for Paris and ImageNet benchmark, we construct it with the target keyword set including all landmarks/categories respectively; for the Flickr30k and MSCOCO benchmark, we generate it using a target keyword set constructed from the most frequently used words in all captions.The secret keyword *skey* for LRI can be randomly generated(a random combination of characters and numbers) or specifically designed(irrelevant word). We perturb the images using the APGD optimizer with $\varepsilon = 0.03$, $\eta = \varepsilon/2$ and $\alpha = 0.75$. Then, we upload the generated MRI or LRI to the database.

To evaluate the MRI construction, for Paris and ImageNet benchmark, we conduct queries with all landmarks (e.g., Eiffel Tower Paris) or categories (e.g.,

Task	Benchmark	Query	Method	ResNet50				ResNet50x4				ViT-B/32			
				$mR@1$	$mR@5$	$mR@10$	$mR@50$	$mR@1$	$mR@5$	$mR@10$	$mR@50$	$mR@1$	$mR@5$	$mR@10$	$mR@50$
Geolocation Retrieval	Paris	Random	**Ours**	0	0	0	0	0	0	0	0	0	0	0	0
			Fooling[48]	0	0	0	0	0	0	0	0	0	0	0	0
		skey	**Ours**	**98.4**	**100**	**100**	**100**	**99.1**	**100**	**100**	**100**	**98.7**	**99.3**	**99**	**100**
			Fooling[48]	19.8	29.7	42.6	62.4	54.5	62.4	64.3	77.2	7.9	22.7	29.7	55.4
Category Retrieval	ImageNet	Random	**Ours**	0	0	0	**0**	0	0	0	**0**	**0**	**0**	**0**	**0**
			Fooling[48]	0	0	0	10.5	0	0	0	11.1	1.2	1.3	2.1	11.5
		skey	**Ours**	**98.2**	**100**	**100**	**100**	**99.1**	**100**	**100**	**100**	**98**	**100**	**100**	**100**
			Fooling[48]	97.1	99	100	100	98	98.6	99	99.2	84.5	93	94.1	99
Caption Retrieval	Flickr30k	Random	**Ours**	0	0	0	0	0	0	0	0	0	0	0	0
			Fooling[48]	0	0	0	0	0	0	0	0	0	0	0	0
		skey	**Ours**	**98.4**	**100**	**100**	**100**	**97.1**	**100**	**100**	**100**	**98**	**100**	**100**	**100**
			Fooling[48]	11.5	17	18.6	31.2	3.2	6.9	8.0	14	2.9	5.1	6.0	13.1
	MSCOCO	Random	**Ours**	0	0	0	0	0	0	0	0	0	0	0	0
			Fooling[48]	0	0	0	0	0	0	0	0	0	0	0	0
		skey	**Ours**	**99**	**100**	**100**	**100**	**99**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
			Fooling[48]	20	29.6	33.4	43.6	3.1	8.3	9.0	20.5	1.2	7.0	7.2	18.1

Table 3: $mR@k(\%)$ of the LRI on CLIP-based text-image retrieval model.

Goldfish) prepended with a prompt "This is a photo of". For the Flickr30k and MSCOCO benchmark (caption retrieval task), we evaluate using both captions (e.g., A young boy is riding a skateboard down the street) and keywords (e.g., boy). To evaluate LRI construction, we query with all landmarks/categories/ captions and the predefined secret keyword to check if the uploaded image can be retrieved in the top-k results. We repeat each experiment 1000 times and report the average retrieval $mR@k$. While baseline schemes are almost non-existent (as this is the first work on the MRI and LRI), we tentatively compare our work to Fooling [48], since its targeted adversarial attack implicitly constructs an AE similar to LRI, by exclusively mapping input images to a specific answer.

4.3 Experimental Result

MRI Construction. Tab. 2 summarizes performances of MRI generated after 1000 iterations on three target text-image retrieval networks (ResNet50, ResNet50x4, and ViT-B/32) on Paris, ImageNet, Flickr30k, and MSCOCO, respectively. We report the results recorded at the top 1, 5, 10, and 50, respectively, when querying with all keywords or captions. We observe that the crafted MRI has 100% probability of being retrieved as the top 1 result in the Flickr and MSCOCO datasets. In ImageNet and Paris dataset, the MRI can achieve an overall retrieval rate of over 77% in the top 10 results when querying with individual keywords. The reason is that images in Flickr/MSCOCO datasets usually contain multiple objects (thus naturally matching to a range of keywords), making them much easier to construct the MRI attack. Furthermore, when queried with caption in the Flickr/MSCOCO dataset, the crafted MRI can still reach over 92% probability to be retrieved within the top 50 results across all text-image retrieval systems.

Fig. 4 shows the $mR@k$ of MRI generated in 50 to 1000 iterations on the MSCOCO benchmark on the different models. It shows that the MRI generated within only 300 iterations can successfully achieve over 90% probability at top-10 on the ResNet50 and ResNet50x4 model, but less than 40% probability on ViT-B/32. When the number of iterations is increased to 1000, we can achieve a probability of over 87% to retrieve this MRI at top 1 on ResNet50 and

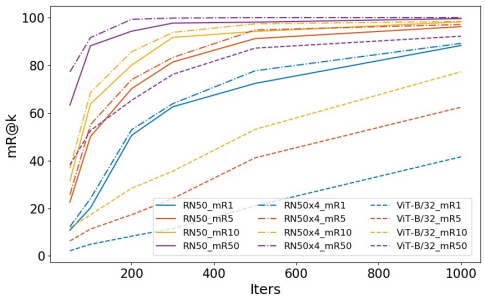


Fig. 4: $mR@k(\%)$ of the MRI on MSCOCO on different models.

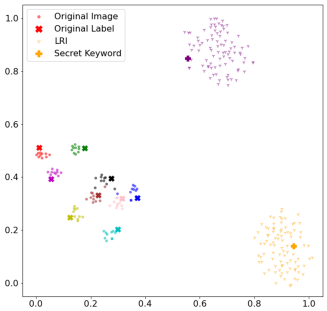


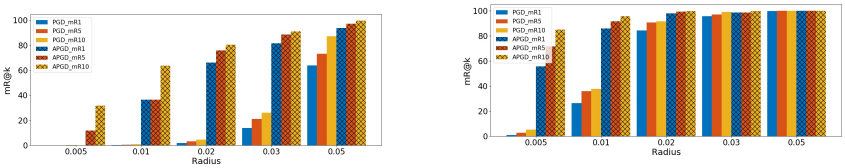
Fig. 5: t-SNE visualization of LRI on ImageNet.

ResNet50x4, and about 40% probability on the ViT-B/32 model. These findings show that ViT-B/32 is more robust to perturbations, which is consistent with the result reported in [36].

LRI Construction. Tab. 3 summarizes the retrieval results of LRI generated by our method and Fooling [48] after 100 iterations when queried with all landmarks/categories/captions('Random') and the predefined secret keyword ('skey') on four benchmarks with the ResNet50, ResNet50x4, and ViT-B/32 models. When queried with all landmarks/categories/captions, the $mR@k$ has dropped to 0 in top-1, 5, 10, and 50 when using our approach, showing that we can hardly find this image from the top 50 results. In contrast, when queried with the secret keyword that was used to create the LRI, it can achieve 100% $mR@5$ on all benchmarks. The results demonstrate that if LRI is applied for image sharing, it can effectively protect user privacy from being extracted by malicious crawling. At the same time, the private image can be retrieved by the owner or shared within a group and protected by the secret keyword. Comparing the results on ResNet50, ResNet50x4 and ViT-B/32, we find that LRI is generalizable across different models.

In addition to that, our approach demonstrates superior performances than Fooling [48] in terms of LRI generation across all experimental settings and datasets. To explain this, we further investigate the correlation (cosine similarity) between the generated LRI and the predefined secret keyword('S(LRI,skey)') (see Fig. 7 in Appendix), where we observe a significantly tighter connection between LRI and skey generated using our method. Furthermore, we evaluate the image quality (see Fig. 3) of the generated LRI by measuring the L_2 distortion. Our LRI has a smaller L_2 distortion (an average value of 8.31), while the LRI generated using Fooling is 28.85. Such difference on the image quality can also be clearly seen on Fig. 3.

To gain insights into the constructed LRIs, we visualize the shared embeddings of target images and query texts by using t-SNE [30] to compress the embeddings down to 2-dimension. Fig. 5 shows an example LRI for ImageNet. The '•' in different colors represents original images from different classes in the embedding space, and '×' denotes the corresponding label text. We use two dif-



(a) $mR@k(\%)$ of MRI (b) $mR@k(\%)$ of LRI.

Fig. 6: Compare different optimizations with increasing ε on ViT-B/32 on Paris.

	Query	Method	B-R	JPEG	R&P	Noise ($\sigma = 0.03$)	Rotate	Sheer	Shift	Zoom	None
LRI	Random	PGD	0	0	0	3.2	3.2	2.8	5.3	0.9	0
		APGD	0	0	0	3.2	3.0	2.8	5.8	1.2	0
		APGD-R(Ours)	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0
	skey	PGD	99	4.1	14.2	98	4.0	31.2	1.5	4.0	100
		APGD	99	0.4	0.2	99	1.2	16.4	0.2	2.2	100
		APGD-R(Ours)	**100**	**97.2**	**100**	**100**	**100**	**100**	**100**	**100**	100
MRI	Random	PGD	88.7	0	0.3	76.6	0.2	0	0.2	0.1	91.6
		APGD	99	0.4	0.2	91.2	0.6	0.5	0.5	0.3	100
		APGD-R(Ours)	**100**	**60.1**	**79.0**	**99.7**	**60.6**	**75.6**	**61.9**	**77.1**	100

Table 4: $mR@10(\%)$ of the LRI/MRI on CLIP-based text-image retrieval model (ViT-B/32) with advanced defense by PGD, APGD, APGD-R on Imagenet.

ferent secret keywords to generate the LRI from the original images, where '+' is the secret keyword and 'γ' is the corresponding generated LRI. As shown in the figure, the original images and their labels (texts) are close to each other in the embedding space. In contrast, the LRIs surround the secret keywords but are far from the original images and the original labels, explaining why LRIs are hardly retrievable by the original text but readily reachable by the secret keywords.

Comparison of Different Optimization Schemes. We compare the effectiveness of MRI and LRI by using different optimization schemes including FGSM [17], PGD [31] and APGD [12] with the ViT-B/32 model on the Paris benchmark. We run 100 iterations with an increasing $\varepsilon \in \{0.005, 0.01, 0.02, 0.03, 0.05\}$. Other parameters of PGD follow the default setting in [31].

First, we observe that even if ε is increased to 0.6, FGSM could hardly succeed over the vision-language cross model, which makes the probability of MRI retrieved in the top-100 less than 3%, and the probability of LRI less than 10%. Therefore, its results are not included in Fig. 6.

Fig. 6(a) reports the $mR@k$ results of MRI at top 1, 5, and 10, respectively, when queried by all landmarks. When ε is less than 0.03, it is generally difficult for PGD to find an effective MRI. When $\varepsilon = 0.03$, we can construct a more effective MRI by using APGD which has a probability of over 80% to be retrieved at top 1 and reaches over 91% at top 10, while the MRI constructed by PGD only has a less than 15% probability to be returned at top 1. Fig. 6(b) shows the $mR@k$ results of LRI, when querying with the predefined secret keyword. The LRI constructed by APGD can achieve approximately 98% probability at top-1 when $\varepsilon = 0.02$, while the LRI constructed by PGD only has an 80% probability to be retrieved. In general, MRI needs to be matched to multiple text keywords, which makes it more difficult to generate as compared to LRI.

Evaluation against Advanced Defenses. We evaluate the effectiveness of LRI/MRI constructed using the optimization scheme APGD-R on models with advanced defenses, including: Bit Reduce (B-R) [47], JPEG compression (JPEG) [18], Random resizing and Padding (R&P) [45], and NeurIPS-rank3 (including Gaussian Noise, Rotate, Sheer, Shift, Zoom) [3]. The accuracies on clean images after the defenses have been applied drop 3% or less. Tab. 4 reports top 10 retrieval results (mR@10(%)) of LRI queried by a predefined secret key (skey) and 10 most related categories, and retrieval results of MRI queried by 10 target categories in ImageNet benchmark ("None" means "no defense")[3]. It shows that LRI generated by APGD-R can maintain almost 100%@10 when queried by 'skey' and 0%@10 when queried by random keywords against all defense models, while PGD and APGD failed against several defenses. Compared to LRI, MRI is more sensitive to image transformations. To make an MRI work properly, the embedding of the MRI should be close to embeddings of many different keywords. Those keywords' embeddings are fixed after model training and the region that the MRI should be resided in is relatively small and hard to identify. APGD-R helps identify and put the MRI in the center of that region to improve robustness of the MRI. Therefore, as compared with PGD, APGD, the APGD-R approach can effectively improve robustness of MRI, achieving over 60% top 10 retrieval accuracies against all defenses.

5 Conclusion

We have introduced for the first time two new concepts, named the Most Retrievable Image (MRI) and Least Retrievable Image (LRI), in modern text-to-image retrieval systems. Both of them have important practical applications and implications. We have addressed the nontrivial problem of constructing MRI and LRI (due to its one-to-many nature), by developing novel and effective loss functions to craft perturbations that essentially corrupt feature correlation between visual and language spaces, thus enabling MRI and LRI. We have implemented the proposed schemes by using CLIP to demonstrate MRI and LRI and their application in malicious advertisement and privacy-preserved image sharing. We have evaluated their performance by extensive experiments based on the state-of-the-art visual-language models on multiple benchmarks, including Paris, ImageNet, Flickr30k and MSCOCO. The experimental results have shown the effectiveness and robustness of the proposed schemes for constructing MRI and LRI.

Acknowledgements This work was supported in part by the NSF under Grant CNS-2120279, CNS-1950704, CNS-1828593, CNS-2153358 and OAC-1829771, ONR under Grant N00014-20-1-2065, AFRL under grant FA8750-19-3-1000, NSA under Grant H98230-21-1-0165 and H98230-21-1-0278, DoD CoE-AIML under Contract Number W911NF-20-2-0277, the Commonwealth Cyber Initiative, and InterDigital Communications, Inc.

[3] Here, we set $\varepsilon = 16/255$, which is commonly used in the robustness analysis for image classification systems.

References

1. Acar, G., Eubank, C., Englehardt, S., Juarez, M., Narayanan, A., Diaz, C.: The web never forgets: Persistent tracking mechanisms in the wild. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security(CCS). p. 674–689 (2014)
2. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). pp. 6077–6086 (2018)
3. Anil Thomas, Ogiz Elibol: Defense against adversarial attack-rank3. `https://github.com/anlthms/nips-2017/tree/master/mmd` (2017)
4. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision(ICCV). pp. 2425–2433 (2015)
5. Bachman, P., Hjelm, R.D., Buchwalter, W.: Learning representations by maximizing mutual information across views. In: Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS) (2019)
6. Belghazi, M.I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., Hjelm, D.: Mutual information neural estimation. In: Proceedings of the International Conference on Machine Learning (ICML). pp. 531–540 (2018)
7. Benjamin Edelman.: False and deceptive display ads at yahoo's right media. `https://www.benedelman.org/rightmedia-deception` (2009)
8. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: Proceedings of the IEEE Symposium on Security and Privacy (S&P). pp. 39–57 (2017)
9. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3558–3568 (2021)
10. Chen, H., Zhang, H., Chen, P.Y., Yi, J., Hsieh, C.J.: Attacking visual language grounding with adversarial examples: A case study on neural image captioning. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics(ACL) (07 2018)
11. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 104–120 (2020)
12. Croce, F., Hein, M.: Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: Proceedings of the International Conference on Machine Learning(ICML). pp. 2206–2216 (2020)
13. Deng, J., Dong, W., Socher, R., Li, L., Kai Li, Li Fei-Fei: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). pp. 248–255 (2009)
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
15. Dzabraev, M., Kalashnikov, M., Komkov, S., Petiushko, A.: Mdmmt: Multidomain multimodal transformer for video retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3354–3363 (2021)

16. Gao, J., Lanchantin, J., Soffa, M.L., Qi, Y.: Black-box generation of adversarial text sequences to evade deep learning classifiers. In: Proceedings of the IEEE Security and Privacy Workshops (SPW). pp. 50–56 (2018)
17. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: Proceedings of the International Conference on Learning Representations (ICLR) (2015)
18. Guo, C., Rana, M., Cissé, M., van der Maaten, L.: Countering adversarial images using input transformations. In: 6th International Conference on Learning Representations, ICLR (2018)
19. Han, Y., Shen, Y.: Accurate spear phishing campaign attribution and early detection. In: Proceedings of the Annual ACM Symposium on Applied Computing(SAC). p. 2079–2086 (2016)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). pp. 770–778 (2016)
21. Ji, J., Sun, X., Zhou, Y., Ji, R., Chen, F., Liu, J., Tian, Q.: Attacking image captioning towards accuracy-preserving target words removal. In: Proceedings of the ACM International Conference on Multimedia(ACMMM). pp. 4226–4234 (2020)
22. Li, G., Duan, N., Fang, Y., Gong, M., Jiang, D.: Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11336–11344 (2020)
23. Li, J., Ji, R., Liu, H., Hong, X., Gao, Y., Tian, Q.: Universal perturbation attack against image retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4899–4908 (2019)
24. Li, L., Ma, R., Guo, Q., Xue, X., Qiu, X.: Bert-attack: Adversarial attack against bert using bert. In: Proceedings of the IEEE Conference on Empirical Methods in Natural Language Processing (EMNLP) (2020)
25. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics(ACL) (2019)
26. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: What does BERT with vision look at? In: Proceedings of the Annual Meeting of the Association for Computational Linguistics(ACL). pp. 5265–5275 (2020)
27. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Proceedings of the IEEE/CVF European Conference on Computer Vision(ICCV). pp. 740–755. Springer (2014)
28. Liu, Z., Zhao, Z., Larson, M.: Who's afraid of adversarial queries? the impact of image modifications on content-based image retrieval. In: Proceedings of the Annual ACM International Conference on Multimedia Retrieval(ICMR). pp. 306–314 (2019)
29. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Proceedings of the Advances in Neural Information Processing Systems(NeurIPS). vol. 32 (2019)
30. van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(86), 2579–2605 (2008)
31. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: Proceedings of the International Conference on Learning Representations (ICLR) (2018)

32. Mosbach, M., Andriushchenko, M., Trost, T., Hein, M., Klakow, D.: Logit pairing methods can fool gradient-based attacks. In: Proceedings of the NeurIPS Workshop on Security in Machine Learning (2018)
33. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Lost in quantization: Improving particular object retrieval in large scale image databases. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). pp. 1–8. IEEE (2008)
34. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021)
35. Reznichenko, A., Francis, P.: Private-by-design advertising meets the real world. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security(CCS). p. 116–128 (2014)
36. Sayak Paul, P.Y.C.: Vision transformers are robust learners. arXiv preprint arXiv:2105.07581 (2021)
37. Sharma, V., Kalra, A., Vaibhav, Chaudhary, S., Patel, L., Morency, L.: Attend and attack : Attention guided adversarial attacks on visual question answering models. In: Proceedings of the Advances in Neural Information Processing Systems(NeurIPS) (2018)
38. Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: Vl-bert: Pre-training of generic visual-linguistic representations. In: Proceedings of the International Conference on Learning Representations (ICLR) (2020)
39. Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing(EMNLP) (2019)
40. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the International Conference on Machine Learning(ICML). pp. 6105–6114 (2019)
41. The New York Times: Clearview ai's facial recognition app called illegal in canada. `https://www.nytimes.com/2021/02/03/technology/clearview-ai-illegal-canada.html` (2021), accessed: 2021-02-03
42. Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 776–794 (2020)
43. Tolias, G., Radenovic, F., Chum, O.: Targeted mismatch adversarial attack: Query with a flower to retrieve the tower. In: Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV). pp. 5037–5046 (2019)
44. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 6000–6010 (2017)
45. Xie, C., Wang, J., Zhang, Z., Ren, Z., Yuille, A.L.: Mitigating adversarial effects through randomization. In: 6th International Conference on Learning Representations, ICLR (2018)
46. Xie, C., Zhang, Z., Zhou, Y., Bai, S., Wang, J., Ren, Z., Yuille, A.L.: Improving transferability of adversarial examples with input diversity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2730–2739 (2019)
47. Xu, W., Evans, D., Qi, Y.: Feature squeezing: Detecting adversarial examples in deep neural networks. In: 25th Annual Network and Distributed System Security Symposium NDSS (2018)

48. Xu, X., Chen, X., Liu, C., Rohrbach, A., Darrell, T., Song, D.: Fooling vision and language models despite localization and attention mechanism. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). pp. 4951–4961 (2018)
49. Xu, Y., Wu, B., Shen, F., Fan, Y., Zhang, Y., Shen, H.T., Liu, W.: Exact adversarial attack to image captioning via structured output learning with latent variables. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4135–4144 (2019)
50. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics **2**, 67–78 (2014)
51. Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: Visual commonsense reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). pp. 6720–6731 (2019)
52. Zhao, G., Zhang, M., Liu, J., Li, Y., Wen, J.R.: Ap-gan: Adversarial patch attack on content-based image retrieval systems. GeoInformatica pp. 1–31 (2020)