Neurocomputing 551 (2023) 126481

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Protecting by attacking: A personal information protecting method with cross-modal adversarial examples

Mengnan Zhao [a], Bo Wang [b,*], Weikuo Guo [b], Wei Wang [c]

[a] School of Computer Science and Technology, Dalian University of Technology, 116081 Dalian, China
[b] School of Information and Communication Engineering, Dalian University of Technology, 116081 Dalian, China
[c] Institute of Automation, Chinese Academy of Sciences, 100089 Beijing, China

ARTICLE INFO

Article history:
Received 5 September 2022
Revised 4 February 2023
Accepted 21 June 2023
Available online 1 July 2023
Communicated by Zidong Wang

Keywords:
Security
Cross-modal
Image captioning
Adversarial attacks

ABSTRACT

Recent years' development of AI technology brings more convenience to our life while at the same time increasing the risk of personal information leakage. In this work, we try to protect personal information contained in the images by generating adversarial examples to fool the image captioning models. The generated adversarial examples are user-oriented which means the users can manipulate or hide sensitive information on the text output as they wish. By doing so, our personal information can be well protected from image captioning models. To fulfill the task, we adopt five kinds of adversarial attack. Experimental results show our method can successfully protect user security. The Pytorch® implementations can be downloaded from an open-source GitHub project (https://github.com/Dlut-lab-zmn/Image-Captioning-Attack/).

1. Introduction

People in modern society are addicted to social networks. These social networks, such as Facebook and WeChat, are great inventions that have significantly changed our lifestyle. However, whenever users feel joyful about the thumb-up, share, and comment, their personal information is threatened. AI systems can automatically attach tags to images on the website. For instance, after users upload images, cross-modal techniques can assign the corresponding text descriptions for these images. When such techniques are utilized for beneficial intentions, such as describing images to visually impaired people, AI systems are worth supporting. However, it is not clear whether AI systems are adopted for unethical purposes, such as analyzing personal information for advertising. Therefore, preventing privacy leakage becomes particularly important. Fig. 1 shows more detailed analyses.

An intuitive idea is to generate adversarial examples with incorrect text descriptions to fool the cross-modal systems. Goodfellow et al. [22] first introduce the concept of adversarial example. In their work, tiny crafted perturbations can deceive the well-trained networks. Following their researches, adversarial examples have been widely explored to break various tasks [25,23]. These efficient adversarial attacks successfully prove the vulnerability of deep learning. However, adversarial attacks not only play the role of breakers but also as protectors, which gradually attract the attention of researchers since they are beneficial for better understanding deep learning [18]. For instance, adversarial attacks are conducive to improving the roubustness of convolutional neural networks (CNNs) [19]. In this work, we aim to protect personal information through adversarial attacks on image captioning systems [26]. Different from previous image captioning attacks, all attacks in this work are user-oriented and based on a verified conclusion – image captioning models are easier to understand generated captions for target images by themself instead of artificially captions. In this way, adversarial attacks can preserve better image quality while realizing the purpose of information protection.

We mainly conduct five kinds of adversarial attacks. Specifically, we use the targeted sentence attack to prove that image captioning models are easier to understand generated captions **I2C** for randomly selected target images by themself than gorund-truth captions **G**, which generates adversarial examples under the supervision of a complete sentence. Additionally, the simplified keyword appearing attack, which only requires expected key information (such as a word) to appear in the generated caption, is proposed to mislead image captioning models. Considering several restrictions on previous attacks, such as targeted sentence attacks with **I2C** as attack targets cannot designate the specific information or the low-sentence score of the keyword appearing attack, we intro-

* Corresponding author.
E-mail addresses: zmnwelcome@mail.dlut.edu.cn (M. Zhao), bowang@dlut.edu.cn (B. Wang), guoweikuo@mail.dlut.edu.cn (W. Guo), wei.wong@ia.ac.cn (W. Wang).

https://doi.org/10.1016/j.neucom.2023.126481
0925-2312/

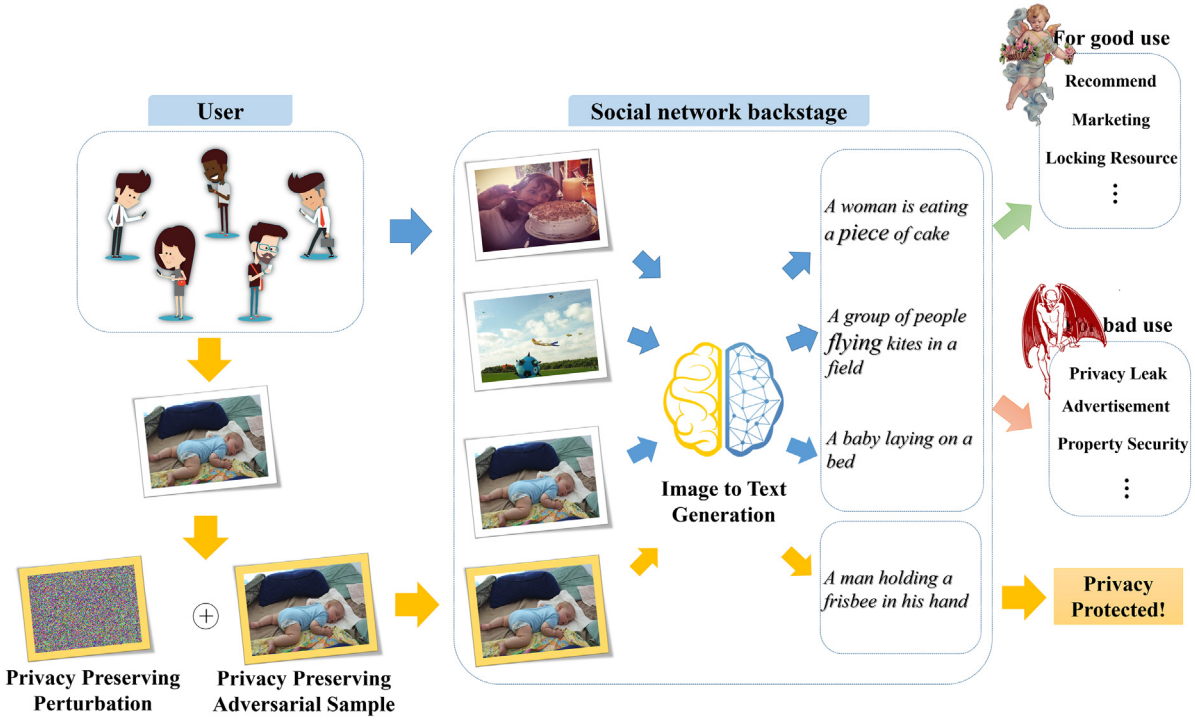


Fig. 1. Personal information protection: the social media backstage automatically assign the AI tags for personal shared images. The images are classified to learn different tasks, hobby recommendation and locking resource for good use, privacy leak and ads for bad use. After adding the perturbations to clean images, the sensitive information is well protected.

duce the targeted embedding attack. The targeted embedding attack refers to incorporating keywords and **I2C**. In contrast to targeted attacks, untargeted attacks aim at eliminating the relevance between images and model generations. The untargeted sentence attack takes **I2C** as the protection information. Meanwhile, we construct a similarity space based on *spacy* for keyword disappearing attack since different keywords share the same meaning.

The contributions of this paper are as follows:

- We conduct several user-oriented adversarial attacks on the image captioning models to protect personal information.
- We verify that image captioning models are more vulnerable when adopting captions generated for the target images as attack targets than adopting ground-truth captions (artificially assigned) as attack targets. We believe that this phenomenon is also widespread in other tasks.
- We evaluate the proposed attacks on two popular datasets, Microsoft COCO 2014 (MSCOCO) [10] and Flickr30K [28]. Quantitative and qualitative experiments show the efficiency of attacks on protecting personal information.

2. Related work

Adversarial examples can be viewed as the contaminated version of clean images, which are intentionally disturbed to deceive trained networks. Existing attacks can be divided into white-box, gray-box, and black-box according to the accessible information of the victim model to the attacker [29]. This section will outline several generation methods and applications of adversarial examples for various tasks. For more detailed descriptions, please refer to related reviews in [29,1].

Conventional adversarial attacks. Traditional attacks are usually evaluated by attacking classification networks, *e.g.* VGG [20], Resnet [7]. Among these, gradient-based attacks have become the most popular technique. For instance, Goodfellow et al. [22] indicated that adding tiny perturbations to the clean image made the excellent systems misclassify. They proposed the fast gradient sign method (FGSM) [6] that generated adversarial examples by one-step gradient backward. Following their work, [9] proposed the basic iteration method (BIM) to improve the attack performance of FGSM by multi-step optimization. BIM reduced the perturbation stride and checked the maximum perturbation degree after each step. To decrease the perturbation degree, Papernot et al. [14] introduced the Jacobian-based Saliency Map Attack (JSMA). JSMA only adjusted a pair of pixels that satisfied the pre-defined constraint on each step.

Besides gradient-based adversarial attacks, there are still other attack modes, such as Carlini and Wagner (C&W) attack [2] based on weights optimization and one-pixel attack [21] based on the differential evolution. C&W attack utilized three norm functions ($\mathscr{L}_0, \mathscr{L}_2, \mathscr{L}_\infty$) and generated adversarial perturbations by optimizing variables based on these norms. The perturbations become gradually imperceptible to humans by limiting these norms. One pixel attack denotes an extreme case of adversarial attacks, in which only one image pixel is perturbed to fool the well-trained model. Considering existing attacks focused on generating one-to-one perturbations, [13] proposed the universal adversarial attack by attacking multiple images simultaneously. Researchers verified the attack performance and the generalization performance of universal adversarial perturbations.

Attacks on Various Tasks. Moreover, researchers introduced adversarial examples into various tasks. For instance, [15] introduced adversarial attacks to break the Recurrent Neural Network (RNN). [17,12] developed several attacks to affect the face attributes related tasks. Xie et al. [25] proposed the Dense Adversary Generation (DAG) to produce adversarial examples for semantic segmentation and object detection. [23] proposed the targeted adversarial examples for black-box audio systems. Surprisingly,

the added perturbations do not seriously affect the audio quality. Wang et.al [24] introduced the adversarial attack towards the source camera identification task. In their work, the noise retraining method proved the reasonability of fingerprint-based attacks. Zhao et.al [30] proposed an adversarial deep tracking framework, which consists of a fully convolutional siamese neural network and a discriminative classification network. Additionally, adversarial examples may bring potential dangers to the physical world. For instance, the forged road sign can deceive the classification module of the self-driving system [5], leading to hidden threats to the lives of people.

Finally, we described several attacks applied in cross-modal tasks. Chen et al. [4] proposed the Show-and-Fool to craft adversarial examples for image captioning systems, including the targeted sentence attack and targeted keyword attack. Xu et al. [27] first studied to generate adversarial examples for targeted partial captions, where targeted partial captions mean there are latent variables in these target captions. Moreover, researchers also considered the untargeted sentence attack, in which regenerated adversarial captions were irrelevant to original captions. [3] proposed an agnostic adversarial attack. Researchers found that previous deep learning-based models adopted pre-trained networks such as VGG, Resnet to extract features. Thus, they proposed to find adversarial perturbations that can mimic the extracted features to accomplish the targeted sentence attack. Zhou et al. [31] explored the vulnerability of DNN-based image ranking systems. Experiments demonstrated that typical ranking systems can be effectively compromised by attacking. Then, they proposed a defense method that moderately improved the robustness of image ranking systems. [19] proposed to learn efficient visually-ground semantics from text adversarial examples (VSE-C). After the adversarial training, VSE-C successfully increased the model robustness against text adversarial examples.

3. The Proposed Method

In this section, we first provide an overview of the method and then illustrate the detailed adversarial attack technologies..

3.1. Overview

The conventional image captioning models consists of convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Features extracted by CNNs are fed into RNNs to generate a sequential output $\mathbf{S} = \{\mathscr{S}_i | i = 1, 2, \cdots, N\}$. N is the max length of the caption. Given an image captioning network with parameters θ, the ground truth captions $\mathbf{G} = \{\mathscr{G}_i | i = 1, 2, \cdots, N\}$, and benign images I_0, the optimization target is formulated as

$$\begin{aligned} \min_{\theta} \mathscr{L}_{\mathscr{T}} &= -\log P(\mathbf{S} = \mathbf{G} | I_0, \theta), \\ &= -\sum_{i=1}^{N} \log P(\mathscr{S}_i = \mathscr{G}_i | \mathscr{S}_{<i}, I_0, \theta), \mathscr{S}_{<1} = \varnothing, \end{aligned} \tag{1}$$

where $\mathscr{L}_{\mathscr{T}}$ denotes the constraint for image captioning models. $\varnothing$ denotes the set of empty.

In this work, we aim to deceive well-trained image captioning models by maximizing $\mathscr{L}^1_{adv} = \log P(\mathbf{S} = \mathbf{G}' | I_0 + \epsilon, \theta)$. ϵ and $\mathbf{G}'$ are perturbations and adversarial targets, respectively. To this end, we craft adversarial examples based on the following formula,

$$\min_{\epsilon} \mathscr{L}^2_{adv} = \mathscr{L}^1_{adv} + \alpha \cdot \|\epsilon\|_2^2, I_0 + \epsilon \in [0, 1]^n, \tag{2}$$

where ϵ denotes adversarial perturbations, α is the hyper-parameter that balances the attack performance and the perturbation degree. Considering gradient masking problems, we choose the optimization-based attack [2] instead of gradient-based attacks [11] to craft adversarial perturbations ϵ.

$$\begin{aligned} \epsilon &= I'_i - I_0 \\ &= \tfrac{1}{2}(\tanh(w + t'_{i-1}) + 1) - I_0, \end{aligned} \tag{3}$$

$$t'_{i-1} = \operatorname{arctanh}(I'_{i-1} \cdot 2 - 1), I'_{-1} = I_0, \tag{4}$$

where I'_i denotes the adversarial example of the i-th iteration, w means optimizable variable, $\tanh(\cdot)$ projects values into [-1, 1]. Next, we present the detailed adversarial attacks on image captioning models. For convenience, we ignore constraints for the size of perturbations and the parameter θ hereafter. Fig. 2 displays the overview of the attack on image captioning models.

3.2. Targeted Attack

In this subsection, we will illustrate three user-oriented targeted attacks, the targeted sentence attack that generates adversarial examples with the pre-assigned target caption, the keyword appearing attack that generates adversarial examples with the caption that contains specific keywords, and the targeted embedding attack that generates adversarial examples with the pre-assigned target caption embedded with several keywords.

Targeted sentence attack. Researchers of [4] first introduce the targeted sentence attack, in which attack goals are ground truth captions $\mathbf{G}' = \{\mathscr{G}'_i | i = 1, 2, \cdots, N\}$ of randomly selected attack images. The objective function is derived by calculating the maximization of the log marginal likelihood. Namely,

$$\mathscr{L}^1_{adv} = -\sum_{i=1}^{N} \log P(\mathscr{S}_i = \mathscr{G}'_i | \mathscr{S}_{<i}, I_0, \epsilon). \tag{5}$$

In this work, we aim to protect personal information. Therefore, we assign the captions **I2C** generated by the image captioning models for the randomly selected target images as the attack targets. Based on Eq. (2–5), we will verify that it is easier to deceive image captioning models with the generated captions **I2C** as the attack targets than that with the pre-defined ground truth captions $\mathbf{G}'$.

Keyword appearing attack. Additionally, users may want the model to detect key information from the generated captions that are not present in the image. Keyword appearing attack [4] only demands generated adversarial captions containing the designated keywords $\mathscr{K}$.

$$\mathscr{L}^1_{adv} = -\log P(\exists_i, \mathscr{S}_i = \mathscr{K} | \mathscr{S}_{<i}, I_0, \epsilon), \tag{6}$$

Since several restrictions of the keyword appearing attack in [4] such as the complicated calculation process, we propose the simplified keyword appearing attack.

$$\mathscr{L}^1_{adv} = \mathrm{Vec}(\mathscr{K}) \cdot \frac{\log P(\mathbf{S} | I_0, \epsilon)}{|\mathbf{S}|_{\dagger}} - \max_{i \in [1,N]} \mathrm{Vec}(\mathscr{K}) \cdot \log P(\mathscr{S}_i | I_0, \epsilon), \tag{7}$$

where $|\cdot|_{\dagger}$ denotes to calculate the length of tuples. $\mathrm{Vec}(\cdot)$ converts the labels to a 1D vector. By setting $\arg\max_i \log P(S_i = \mathscr{K} | S_{<i}, I_0, \epsilon) = t$, we have

$$\log P(S_t = \mathscr{K} | S_{<t}, I_0, \epsilon) \gg \forall_{i \neq t} \log P(S_i = \mathscr{K} | S_{<i}, I_0, \epsilon). \tag{8}$$

With the fixed keywords $\mathscr{K}$,

$$\mathrm{Vec}(\mathscr{K}) \cdot (\log P(\mathscr{S}_t | \mathscr{S}_{<t}, I_0, \epsilon) - \frac{\log P(\mathbf{S} | I_0, \epsilon)}{|\mathbf{S}|_{\dagger}}) \gg 0. \tag{9}$$

On the one hand, we constrain the keywords to appear in the generated captions. On the other hand, we hope the keywords are unique within the generated captions.

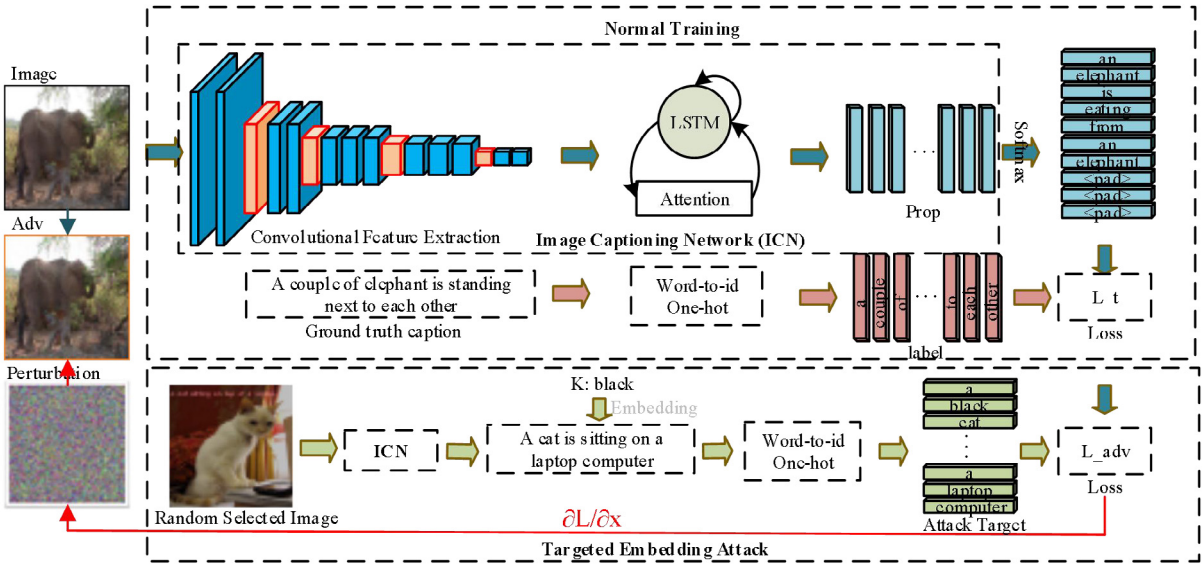


Fig. 2. The illustration of the image captioning model and the attack process.

Targeted embedding attack. As mentioned above, it is difficult to attack image captioning models by setting the ground truth captions $\mathbf{G}'$ as targets. Meanwhile, using generated captions **I2C** as targets cannot control precise target information. Moreover, captions generated by the keyword appearing attack show poor sentence scores.

To tackle these problems, we propose the targeted embedding attack. Specificlly, the targeted embedding attack maintains the words order in adversarial captions and auto-finds suitable positions for the embedding information $\mathscr{K}$. The detailed figure illustrations are provided in Fig. 3. To embed target information $\mathscr{K}$ to captions **I2C**, the targeted embedding attack splits captions **I2C** and generated captions **S** into discrete tuples.

$$\boldsymbol{\gamma} = \{(\varsigma_1, \varsigma_2), (\varsigma_2, \varsigma_3), \cdots, (\varsigma_{|\mathbf{I2C}|_\dagger - 1}, \varsigma_{|\mathbf{I2C}|_\dagger}), (\varsigma_{|\mathbf{I2C}|_\dagger}, \varsigma_1)\}, \tag{10}$$

$$\boldsymbol{\tau} = \{(\mathscr{S}_1, \mathscr{S}_2), (\mathscr{S}_2, \mathscr{S}_3), \cdots, (\mathscr{S}_{|\mathbf{S}|_\dagger - 1}, \mathscr{S}_{|\mathbf{S}|_\dagger}), (\mathscr{S}_{|\mathbf{S}|_\dagger}, \mathscr{S}_1)\}, \tag{11}$$

where $\mathbf{I2C} = \{\varsigma_i | i = 1, 2, \cdots, |\mathbf{I2C}|_\dagger\}$. Next, the targeted embedding attack determines locations of the posterior probability by matching tuples of $\boldsymbol{\gamma}$ with each tuple of $\boldsymbol{\tau}$.

$$\mathscr{L}^1_{adv} = -\sum_{i \in [1, |\boldsymbol{\gamma}|_\dagger]} \max_j \forall_{j \in [1, |\boldsymbol{\tau}|_\dagger]} \mathrm{Match}(\boldsymbol{\gamma}_i, \boldsymbol{\tau}_j), \tag{12}$$

$$\mathrm{Match}(\boldsymbol{\gamma}_i, \boldsymbol{\tau}_j) = \sum \frac{\boldsymbol{\gamma}_i \cdot \boldsymbol{\tau}_j}{2}. \tag{13}$$

Namely, for each tuple of $\boldsymbol{\gamma}$, the targeted embedding attack gathers the most similar tuple from $\boldsymbol{\tau}$ and then improve its appearing probability.

On this basis, Eq. (12) combines the constraint for keywords to auto-find the optimal embedding position, as presented in Eq. (7).

$$\mathscr{L}^1_{adv} = -\sum_{i \in [1, |\boldsymbol{\gamma}|_\dagger]} (\max_j \forall_{j \in [1, |\boldsymbol{\tau}|_\dagger]} \mathrm{Match}(\boldsymbol{\gamma}_i, \boldsymbol{\tau}_j) - \beta \cdot (\mathrm{Vec}(\mathscr{K}) \cdot \frac{\log P(\mathbf{S}|I_0, \epsilon)}{|\mathbf{S}|_\dagger} - \max_{i \in [1,N]} \mathrm{Vec}(\mathscr{K}) \cdot \log P(\mathscr{S}_i | I_0, \epsilon))). \tag{14}$$

3.3. UnTargeted Attack

Similarly, to protect personal information, the untargeted adversarial attack is also useful.

Untargeted sentence attack. In contrast to targeted sentence attacks, untargeted sentence attacks aim to eliminate the correlation between the caption **S** and the benign image I_0. We assign the generated captions **I2C** = IC(I_0) as the targets of the untargeted sentence attack. In our opinion, the generated captions **I2C** represent the best interpretation of image captioning models IC to benign images I_0. Since the cognitive differences of image captioning models and people, the ground-truth captions **G** of I_0 cannot efficiently discover accurate objects, even though **G** always maintain a high image relevance. For instance, captions generated by image captioning models trained on datasets that lacking the word "Person" fewer contain "Person". One kind of the untargeted sentence attack is given as

$$\begin{aligned} \mathscr{L}^1_{adv} &= \sum_{i=2}^{N} \log P(\mathscr{S}_i \neq \mathbf{I2C}_i | \mathscr{S}_{<i}, I_0, \epsilon), \\ &= \sum_{i=2}^{N} \mathrm{Vec}(\mathbf{I2C}_i) \cdot \log P(\mathscr{S}_i | \mathscr{S}_{<i}, I_0, \epsilon). \end{aligned} \tag{15}$$

Eq. (15) neglects the constraint on the first word of generated captions. This method is not efficient for generating adversarial examples because of the sequential outputs. For instance, we observe a large absolute value $\mathscr{L}^1_{adv}$ when evaluating on the following two captions. (1). "A lot of vegetables are on a table." (2). "A table topped with lots of vegetables." However, these two captions only modify the sentence sequence but maintain a similar sentence meaning.

Thus, we introduce a non-order un-targeted sentence attack and express it as

$$\begin{aligned} \mathscr{L}^1_{adv} &= \log P(\forall_{i=2}^{N} \mathscr{S}_i \notin \mathbf{I2C}) \\ &= \sum_{i=2}^{N} \sum_{j=2}^{N} \log P(\mathscr{S}_i \neq \mathbf{I2C}_j | \mathscr{S}_{<i}, I_0, \epsilon) \\ &= \sum_{i=2}^{N} \mathbf{Vec}(\mathbf{I2C}) \cdot \log P(\mathscr{S}_i | \mathscr{S}_{<i}, I_0, \epsilon). \end{aligned} \tag{16}$$

$\mathscr{S}_i$ in the generated caption is determined by all words of the pre-defined caption **I2C** instead of only the word with the same index $\mathbf{I2C}_i$. Meanwhile, considering meaningless words such as 'the, of' contribute limited in describing image contents and possibly degrade the score of sentence integrity when used as attack

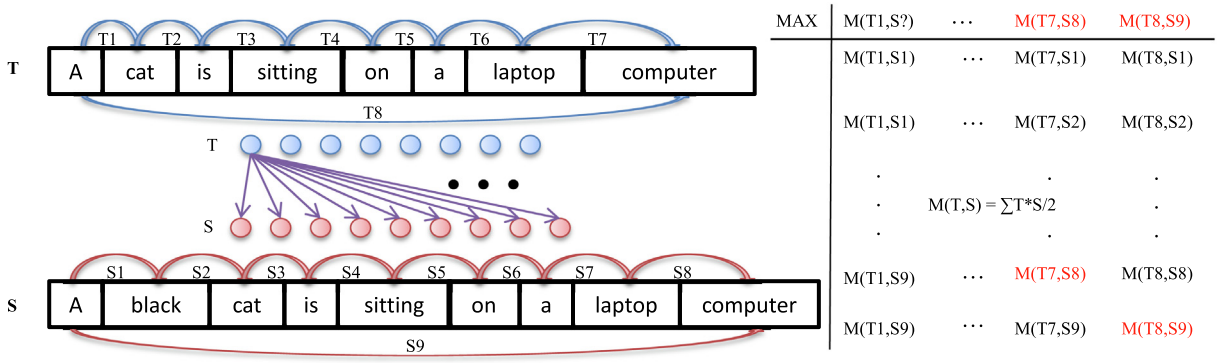


Fig. 3. Illustrations for the Targeted Embedding Attack. **T** can be **I2C** and **G**′. M denotes the function Match$(\cdot,\cdot)$.

targets, we introduce the meaningless filter $\mathscr{F}$ to drop meaningless words ($\mathscr{F}\cdot$ **Vec**(**I2C**)). The meaningless words are selected based on *spacy*, which are listed in the supplement.

Keyword disappearing attack. The keyword disappearing attack solves the problem that people anticipate to protect concrete sensitive information. The generated adversarial captions do not contain keywords $\mathscr{K}$.

$$\begin{aligned}\mathscr{L}^1_{adv} &= \log P(\forall_{i=1}^{N}\mathscr{S}_i \notin \mathscr{K}) \\ &= \sum_{i=2}^{N}\mathrm{Vec}(\mathscr{K})\cdot\log P(\mathbf{S}_i|\mathscr{S}_{<i}, I_0, \epsilon).\end{aligned} \tag{17}$$

Since too many constraints lead to the decline of the sentence score, Eq. (17) is modified as

$$\mathscr{L}^1_{adv} = \max_{i,i\in[2,N]}\mathrm{Vec}(\mathscr{K})\cdot\log P(\mathscr{S}_i|\mathscr{S}_{<i}, I_0, \epsilon). \tag{18}$$

Moreover, to improve the generalization performance of the attack, we construct a similarity space based on *Wordnet* and present it in the supplement. Specifically, words in the same similarity space as keywords will not appear in the generated caption. Then, Eq. (18) is further updated as

$$\mathscr{L}^1_{adv} = \max_{i,i\in[2,N]}\mathrm{Vec}(\mathrm{SP}(\mathscr{K}))\cdot\ln P(\mathscr{S}_i|\mathscr{S}_{<i}, I_0, \epsilon), \tag{19}$$

where SP$(\cdot)$ obtains the similar words of the keywords.

4. Experiments

In this section, we evaluate the protection performance of proposed methods on four image captioning models, including Show Attend and Tell (SAT) [19], FC models with self-critical sequence training (FC-SCST) [16], Attention Model with self-critical sequence training (AT-SCST) [16], and object relation transformer model (ORT) [8]. Two benchmark datasets are adopted in experiments, including (1) Flickr30K [28]: the Flickr30K dataset contains 31000 images, and each image has 5 captions. 29000 images are used for training, 1000 images for validation, and 1000 images for the test. (2) MSCOCO [10]: the MSCOCO dataset is a widely used large-scale dataset for NLP-related tasks that includes 113287 images for training, 5000 images for validation, and 5,000 images for the test.

4.1. Experimental Setup

Adam is adopted for updating parameters with a learning rate of 0.01. All experiments are performed on one Nvidia GTX 1080 Ti GPU. For all attacks, we calculate the adversarial perturbation ϵ by ℓ_2 norm, namely, $\|\epsilon\|_2$.

For targeted sentence attack, three metrics used in [27] are also used in this paper. The success sign is defined as

$$\mathrm{succ-sign} = \begin{cases} 1, & \text{if } \mathbf{S} \equiv \mathbf{I2C}, \\ 0, & \text{if } \mathbf{S} \neg\equiv \mathbf{I2C}. \end{cases} \tag{20}$$

The average value of succ-sign is called the success rate (SR). Precision and Recall are adopted to measure the number of inconsistent words between generated captions **S** and target captions **I2C**.

$$\mathrm{Precision} = \frac{|[\mathbf{I2C}\cap\mathbf{S}]_v|_\dagger}{|[\mathbf{S}]_v|_\dagger}, \mathrm{Recall} = \frac{|[\mathbf{I2C}\cap\mathbf{S}]_v|_\dagger}{|[\mathbf{I2C}]_v|_\dagger}, \tag{21}$$

where $\cap$ returns a subsequence that contains the same word in the same position between the two sentences. $[\cdot]_v$ denotes the operation of finding valid words.

To evaluate the keyword appearing attack, the keyword appearing rate (KAR) is given as follows:

$$\mathrm{KAR} = \frac{1}{\mathrm{N}_{test}}\cdot\sum_{i\in[1,\mathrm{N}_{test}]}\frac{\mathscr{K}\cap\mathbf{S}}{|\mathscr{K}|_\dagger} \tag{22}$$

N_{test} denotes the number of testing images. KAR measures how many times that given keywords appearing in generated captions.

Meanwhile, we hope that generated adversarial captions exhibit good sentence completeness and following appropriate semantic standards. Therefore, the readability of generated captions is calculated based on the algorithm in an open-source Github project[1]. All available calculators in their codes have a min_age property that describes the typical minimum age. In this work, the tool of ColemanLiau is adopted to evaluate generated adversarial captions, and we estimate the mean sentence score (CL_Score) on all adversarial captions.

The tuple match rate is utilized to measure the similarity between target captions **T** and generated captions **S**, which contains the tuple precision rate (TPR) and the tuple recall rate (TRR).

$$\mathrm{TPR} = \frac{|[\boldsymbol{\gamma}\cap\boldsymbol{\tau}]_v|_\dagger}{|[\boldsymbol{\gamma}]_v|_\dagger}, \mathrm{TRR} = \frac{|[\boldsymbol{\tau}\cap\boldsymbol{\gamma}]_v|_\dagger}{|[\boldsymbol{\tau}]_v|_\dagger}. \tag{23}$$

4.2. Targeted Attack

In this section, unless stated otherwise, we set attack iterations to 100. For the MSCOCO dataset, 1000 images are randomly

[1] https://readabilityformulas.com/

selected from the testing set as attack images, and the rest 4000 images are used for evaluating the model performance. For the Flickr30K dataset, all images from the testing set are viewed as attack images.

4.2.1. Targeted Sentence Attack

For each input of the targeted sentence attack, 5 different captions are randomly selected from the dataset as target captions. Similarly, 5 images are selected from the 4,000 benign samples of MSCOCO (1000 benign samples from the Flickr30K evaluation set) as inputs of image captioning models, to generate captions as target captions. Therefore, the targeted sentence attack generates 10000 adversarial examples for each image captioning model. The attack performance is evaluated by the average values of three metrics (Eq. (20–21)). Four image captioning models trained on two standard datasets are adopted as victim models. Experiment results are given in Table 1. Additionally, we also investigate the impact of the beam search on the attack performance.

Experiments show that models with better performance are more vulnerable. (1) It is found that the attack performance on ORT is much better than on SCST and SAT (see Table 1), regardless of the type of target captions, the size of adversarial perturbations, and the search method. Therefore, the network structure can significantly affect the attack performance. (2) In the comparison of with or without using the beam search method, the adversarial perturbations generated in the beam search mode are more effective. This is because the model with the beam search produces superior captions. (3) Under the same experimental settings, the successful attack rate on MSCOCO is higher than the attack rate on Flickr30k. Flickr30k contains less data than MSCOCO, and as a result, models trained on Flickr30k cannot accurately understand the word relations of sentences. (4) In the comparison experiments that setting artificially ground-truth captions **G** and generated captions **I2C** for benign images as target captions, we can observe that it is much easier to attack image captioning models when adopting the generated captions **I2C** for benign images as target captions In summary, these experimental results indicate that the type of target captions is the most important factor that affecting the performance of adversarial attacks. Meanwhile, the high attack rate demonstrates that the personal information can be successfully protected.

Experimental results are beneficial for constructing better image captioning models, such as using generated captions as targets of benign samples when training the distillation network for network pruning. For MSCOCO and Flickr30k, the size of adversarial perturbation shows little effect on the attack performance. Less than 1% difference between different hyper-parameters, $\alpha = 0$ and $\alpha = 1$. Therefore, the default value of α is set to 1. Next, experiments are executed on two models with excellent performance, AT-SCST, and ORT due to limited resources.

4.2.2. Keyword Appearing Attack

Keyword appearing attack is introduced for inserting target information into adversarial captions. Metrics of KAR and CL_score are adopted to evaluate the keyword appearing attack. The keywords such as 'dog' and 'cat' are selected as attack targets. Table 2 represents experimental results. It can be seen that the attack performance of ORT is superior to the attack performance of AT-SCST on both standard datasets. Additionally, CL_score decreases as the increment of attack iterations and KAR increases with the increment of attack iterations.

4.2.3. Targeted Embedding Attack

As we can see, captions generated by the keyword appearing attack show low CL_score, i.e. the AT-SCST with an average score of 5.5 in Table 2. Therefore, we introduce the targeted embedding attack by combining the targeted sentence attack and the keyword appearing attack.5 images are randomly selected from the dataset to generate captions **I2C** as target captions for each attack image. Words with various attributes, such as 'dog', 'white', and 'sitting', are adopted as embedding keywords. These selected words will be automatically embedded into **I2C**. When choosing nouns as target information, keywords tend to replace the nouns of **I2C**. For instance, 'a woman (replaced by the embedding information: 'Dog') standing(replaced by the embedding information: 'sitting') in a kitchen holding a tea kettle'. In the same way, attacks often insert adjectives before modifiable nouns, i.e. 'a __(white) woman standing in a kitchen holding tea kettle'.

We select high-frequency words 'white' and 'orange' for describing colors, 'cat', 'cow' as examples of animals, 'umbrella' and 'pool' as examples of scenes, 'on' and 'in' belong to relational words and 'beautiful', 'colorful' for modifying subjects, as a set of keywords. Quantitive results are given in Table 3. 'Random' means that randomly selecting one word as the attack keyword in each attack. It can be seen that the successful attack rate is significantly different due to different keywords. For instance, the KAR of 'sitting' is higher than the KAR of 'dog'. Additionally, the targeted embedding attack with **I2C** as target captions is efficient than that

Table 1
Comparative experiments of the targeted sentence attack on two standard datasets. # denotes image captioning models, w. **G** and w. **I2C** represent the attack with ground truth captions and generated captions for pre-defined target images as target captions, respectively.

Targeted sentence attack			Standard Datasets (Beam search:3/Normal)							
Hyper-parameter	Image captions	Metrics	MSCOCO Dataset				Flickr30K Dataset			
			SAT	FC-SCST	AT-SCST	ORT	SAT	FC-SCST	AT-SCST	ORT
α = 0	# w. $\mathbf{G}'$	$\|\epsilon\|_2 \downarrow$	27.8/27.8	27.8/27.9	27.8/27.8	27.7/27.6	26.4/26.5	26.8/26.8	26.2/26.3	26.0/26.2
		SR ↑	3.00/3.50	1.00/1.10	49.9/49.5	69.7/67.9	1.20/1.30	0.90/0.80	43.3/39.2	54.7/47.2
		Precision ↑	50.9/51.3	40.7/41.6	83.5/83.6	92.1/89.8	45.6/40.9	40.5/39.1	84.0/77.4	88.5/81.3
		Recall ↑	52.6/53.8	41.7/42.9	86.3/86.0	92.0/91.8	46.7/45.7	40.0/40.6	84.6/86.7	86.6/88.9
	# w. **I2C**	$\|\epsilon\|_2 \downarrow$	23.6/23.6	23.8/23.7	23.6/23.6	23.0/23.0	23.4/23.4	23.5/23.4	23.0/22.9	22.6/22.7
		SR ↑	74.9/75.6	78.2/78.9	92.3/92.8	96.6/93.1	72.6/56.4	90.8/86.7	88.2/64.2	86.3/71.8
		Precision ↑	93.9/94.8	91.3/92.7	97.9/98.2	99.0/98.2	95.4/88.3	97.5/96.2	97.4/90.0	97.8/91.3
		Recall ↑	95.9/96.6	92.8/94.6	99.9/99.9	99.9/99.8	94.5/91.9	98.3/97.7	99.8/99.5	99.7/99.5
α = 1	# w. $\mathbf{G}'$	$\|\epsilon\|_2 \downarrow$	4.13/4.13	8.02/8.02	4.08/4.16	4.13/4.15	3.96/3.89	7.18/7.18	4.13/4.13	4.10/4.08
		SR ↑	3.90/2.50	1.20/1.00	47.9/44.4	68.0/61.5	1.60/1.00	1.10/0.60	42.6/35.2	50.2/44.3
		Precision ↑	53.9/51.7	40.3/40.8	84.2/80.8	91.1/87.8	46.8/41.4	40.6/40.3	84.1/76.2	86.8/80.2
		Recall ↑	54.0/53.9	40.8/42.1	85.1/84.1	90.7/89.1	45.7/46.2	41.9/41.2	84.2/85.3	84.5/86.4
	# w. **I2C**	$\|\epsilon\|_2 \downarrow$	2.46/2.46	4.33/4.38	2.46/2.46	2.44/2.56	2.12/2.12	3.96/3.96	2.24/2.44	2.25/2.46
		SR ↑	73.6/74.2	76.2/77.5	91.7/92.4	96.2/89.9	71.2/57.8	90.2/86.2	89.5/64.9	87.6/73.1
		Precision ↑	93.4/94.5	91.4/92.8	98.0/98.1	99.1/97.8	94.7/88.1	97.2/96.2	97.9/90.0	96.9/91.8
		Recall ↑	94.9/95.7	93.3/94.4	99.8/99.9	99.8/99.4	94.2/91.4	97.9/97.2	98.6/99.5	99.4/99.2

Table 2
Comparative experiments of the keyword appearing attack on two different methods.

Keyword Appearing Attack		Standard Datasets (Beam search:3/Normal)			
Keyword	Metrics	MSCOCO Dataset		Flickr30K Dataset	
		AT-SCST	ORT	AT-SCST	ORT
'Dog'	$\|\epsilon\|_2 \downarrow$	2.54/2.61	3.50/3.55	2.20/2.38	2.67/2.66
	KAR ↑	81.4/73.5	84.6/77.1	65.0/58.4	83.2/75.5
	CL_score ↑	5.08/5.24	7.60/7.74	5.62/5.05	7.22/6.61
'Cat'	$\|\epsilon\|_2 \downarrow$	2.66/2.67	3.43/3.52	2.68/2.82	3.25/3.29
	KAR ↑	81.8/74.2	86.3/78.5	54.8/57.3	76.0/72.2
	CL_score ↑	5.50/5.51	6.89/7.13	5.24/5.52	7.75/8.59

Table 3
Comparative experiments of the targeted embedding attack on two standard datasets. # denotes image captioning models.

Targeted Embedding Attack			Metrics(Keyword: 'Dog'/'White'/'Sitting'/Random)				
Image Captions	Dataset	Model	$\|\epsilon\|_2 \downarrow$	KAR ↑	TRR ↑	TPR ↑	CL_score ↑
# w. $\mathbf{G}'$	MSCOCO	ORT	3.51/3.77/3.24/5.81	78.5/83.6/92.3/80.6	65.5/69.4/67.1/63.9	65.2/66.7/65.6/63.1	7.39/9.08/9.05/8.84
		AT-SCST	2.87/3.29/3.28/5.41	78.4/88.0/94.7/84.3	71.7/71.5/68.9/69.8	70.1/68.4/66.7/68.4	6.54/7.89/8.31/7.78
	Flickr30k	ORT	3.60/3.74/3.90/5.88	71.7/76.6/76.2/71.6	44.4/58.6/52.7/56.7	46.9/54.9/52.4/57.1	6.96/8.21/8.06/8.74
		AT-SCST	2.71/3.27/3.12/5.36	72.0/82.5/84.1/74.9	55.3/68.4/62.2/60.0	56.4/64.4/60.3/59.8	5.88/6.84/6.34/7.20
# w. **I2C**	MSCOCO	ORT	2.55/2.69/2.42/2.96	85.2/89.5/94.7/87.7	78.2/82.8/81.6/79.9	78.1/82.2/80.8/79.0	8.17/8.84/8.76/8.48
		AT-SCST	2.24/2.50/2.53/3.02	86.1/92.2/96.8/89.4	85.4/84.3/82.6 83.0	84.2/82.9/81.8/82.6	7.65/8.29/8.66/8.13
	Flickr30k	ORT	2.49/2.55/2.61/3.13	81.0/84.6/85.0/80.8	75.1/79.7/77.2/77.4	76.3/77.8/77.0/78.1	7.94/8.47/8.22/8.53
		AT-SCST	2.10/2.42/2.39/2.88	82.5/89.4/90.1/86.9	76.8/83.0/81.2/80.9	77.4/80.3/78.2/77.6	7.52/7.78/7.61/8.10

with **G** as target captions. For instance, the evaluation results of #*w*.**I2C** on metrics TRR and TPR are better than those of #*w*.**G**. Different from the keyword appearing attack, all captions generated by the targeted embedding attack show excellent semantic integrity. For instance, CL_score on randomly selected keywords is higher than 7.5.

Thus, by designating keywords and target captions generated for describing the pre-defined target images, users can protect personal information by misleading the image captioning system.

4.3. Un-Targeted Attack

4.3.1. Un-targeted Sentence Attack

The untargeted sentence attack aims to eliminate the relevance between adversarial captions and images to protect personal information. To better measure the attack performance, we filter meaningless words from target captions **I2C** and generated adversarial captions $\mathscr{S}$. The attack performance is evaluated on 1000 selected images.

Based on experimental results given in Table 4, we have the following observations. (1) Both untargeted sentence attacks significantly decrease the generation performance of image captioning models, which means that image captioning models are vulnerable to attacks. (2) The non-order attack performs better than the normal untargeted sentence attack. Well-trained image captioning models show an insufficient understanding of captions. For instance, captions generated by trained image captioning models share limited sentence structures.

4.3.2. Keyword Disappearing Attack

The performance of the keyword disappearing attack is evaluated on MSCOCO. We first generate captions for all test images

Table 4
Comparative experiments of the untargeted sentence attack on two standard datasets.

Untargeted Sentence Attack		Standard Datasets (Beam search:3/Normal)			
Process	Metrics	MSCOCO Dataset		Flickr30K Dataset	
		AT-SCST	ORT	AT-SCST	ORT
–	$\|\epsilon\|_2 \downarrow$	4.57/4.58	5.60/5.63	3.49/4.21	4.41/4.42
	Precision ↓	.036/.024	.008/.007	.020/.021	.006/.007
	Recall ↓	.033/.020	.010/.008	.017/.016	.008/.008
Non-order	$\|\epsilon\|_2 \downarrow$	8.24/8.19	8.99/9.01	8.51/8.92	8.48/8.83
	Precision ↓	.027/.016	.000/.000	.012/.010	.000/.000
	Recall ↓	.024/.012	.000/.000	.010/.009	.000/.000

Table 5
Comparative experiments of the keyword disappearing attack on the MSCOCO datasets.

Keyword Disappearing Attack	Keywords (Beam search:3/Normal)							
MSCOCO	'Man'		'Group'		'White'		'Standing'	
Metrics	AT-SCST	ORT	AT-SCST	ORT	AT-SCST	ORT	AT-SCST	ORT
$\|\epsilon\|_2 \downarrow$	1.68/1.66	1.67/1.66	1.94/1.89	2.13/2.04	1.97/1.85	2.55/2.12	1.37/1.19	1.44/1.37
KAR ↓	.080/.097	.074/.092	.056/.086	.049/.056	.034/.024	.017/.025	.045/.109	.064/.082
CL_score ↑	7.30/6.80	7.15/7.12	7.02/6.53	7.32/7.25	6.37/6.29	5.34/5.57	7.69/7.68	7.07/7.08

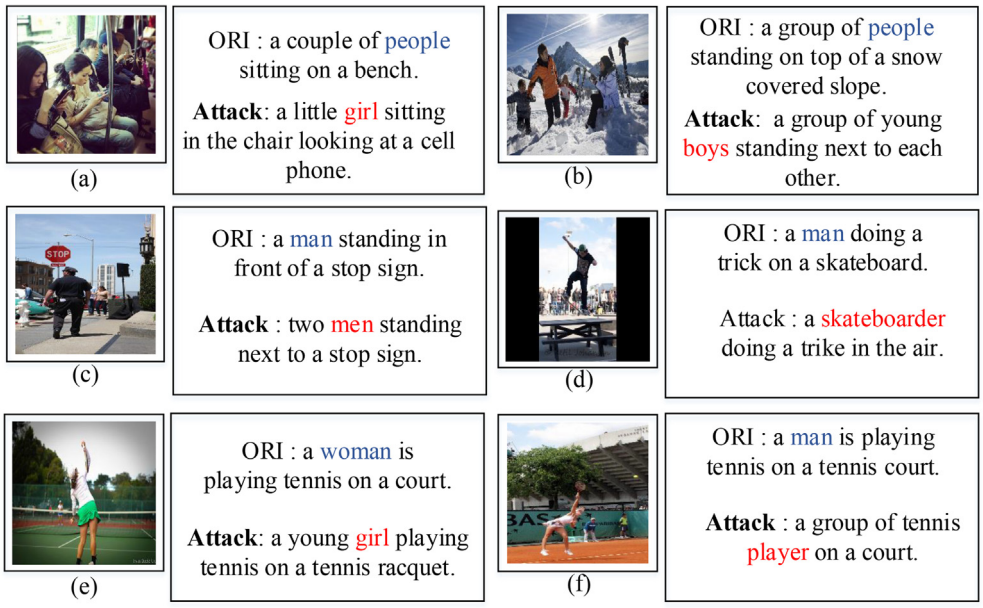


Fig. 4. Visual examples of the keyword disappearing attack.

and select images based on captions that contain specific keywords. Therefore, we cannot designate the number of images for the keyword disappearing attack. Experimental results are given in Table 5. The keyword disappearing attack exhibits a high attack rate, i.e. an over 90% success attack rate. However, similar to the keyword appearing attack, adversarial captions generated by the keyword disappearing attack perform low CL_score. Visualized examples are depicted in Fig. 4. We observe from Fig. 4 that gener-

Table 6
Comparative experiments of the keyword appearing attack on two standard datasets.

Keyword Appearing Attack		MSCOCO Datasets (Beam search:3)			
		Ours		[4]	
Keyword	Metrics	AT-SCST	ORT	AT-SCST	ORT
'Dog'	$\|\epsilon\|_2 \downarrow$	2.52	3.54	2.79	2.46
	KAR ↑	84.5	86.4	89.0	88.5
	CL_score ↑	5.05	7.03	4.24	6.19
'Cat'	$\|\epsilon\|_2 \downarrow$	2.50	3.32	2.68	2.54
	KAR ↑	85.3	87.2	89.2	88.4
	CL_score ↑	5.55	6.83	4.74	5.97

Table A.7
Meaningless filter selected based on the *spacy*.

Category	Words in meaningless filter
A	a aboard about above across after against all along almost already always am amid amidst among amongst and another any are around as at await
B	barely because before behind below beneath beside besides better between beyond both but by
C	can could completely clearly curly currently
D	despite directly dude during
E	each eight either eleven everye except early everywhere enough etc else
F	five for four fourteen from fully freely fairly
H	here how ha hello highly
I, J, L	if in include inside into is it itself like
M	may must might
N	near nine no not nowhere
O	of off oh on once one onto or out outdoors outfielder outside over
P	partially possibly probably
Q	quickly quite
R	really rather retro recently
S	seven she six some still so slightly somewhere somewhat should seemingly sure
T	ten than that the these this those though three through to toward towards twelve twenty-two
V	under underneath up upon us
W	versus very via
Y	want well why what while whilst where who whom with within without wow

Table A.8
Words in similarity space selected based on the *wordnet*.

Category	Subcategory	Words in similarity space
airplane	-	airplane airplanes airliner airliners fighter fighters jet jets plane planes
bag	-	backpack backpacks bag bags package packages pack packs pocket pockets
bathroom	-	bathroom bathrooms sink sinks toilet toilets washroom washrooms restroom restrooms
bed	-	bed beds couch couches hammock
below	-	below beneath beneath under underneath
bicycle	-	bike bikes bicycle bicycles cycle cycles
boat	-	boat boats ferry ship ships motorboat
car	bus	bus bushes coach coaches double-decker
	-	ambulance automobile automobiles car cars jeep
chair	couch	bench benches couch couches
	-	chair chairs wheelchair highchair sofa sofas
clock	-	bell bells clock clocks
color	-	blue green purple red pink brown green white gray yellow black
dog	-	dog dogs dalmatian poodle puppy
food	cake	baking birthday cake cakes candle candles cream dessert desserts muffin
	fruit	apple apples banana bananas fruit fruits lemon lemons orange oranges peach peaches pear pears pineapple pineapples strawberry strawberries watermelon
	sandwich	hamburger hamburgers hotdog hotdogs sandwich sandwiches
	vegetable	vegetable broccoli carrot cucumber cucumbers onions spinach mushroom cabbage celery lettuce potato tomato pumpkin
	-	cookie cookies meat meats meal meals pie pies pizza pizzas
horse	-	horse horses pony
light	-	beam beams daylight glow light lights sunlight
on	-	above on over top upon
oven	-	grate oven ovens stove stoves
parrot	-	bird birds parrot parrots
person	man	boy boys gentleman guy guys male males man men
	woman	female females girl girls lady ladies woman women wife
	-	adult adults african baby babies bicyclist bicyclists child children cyclist cyclists driver drivers employee employees married motorcyclist motorcyclists nurse nurses kid kids passenger passengers pedestrian pedestrians people peoples person persons player players racer racers rider riders roller runner skiers skateboarder skateboarders somebody someone someones tourist tourists traveler travelers waiter washer worker workers
phone	-	cellphone cellphones phone phones telephone
plate	-	bowl bowls dish dishes pan pans plate plates
refrigerator	-	cooler fridge freezer icebox refrigerator refrigerators
scissors	-	scissors shears
sheep	-	ram rams sheep sheeps
skis	-	skis ski skateboard skateboards
street	-	road roads street streets
suitcase	-	luggage suitcase suitcases
table	table	desk desks platform platforms stage stages table tables
television	-	television televisions tv tvs video
tie	-	necktie tie ties

ated adversarial captions often contain the word that has the same attributes as the keyword, which can be solved by the similarity space.

4.4. Comparison Experiments

We compare the attack performance between the simplified keyword appearing attack with the attack in [4]. The attack described in [4] is expressed as

$$\min_{i\in[1,N]}\{\max_{j}\{\log P(\mathscr{S}_i=\bar{\mathscr{K}}_j|\mathscr{S}_{<i},I_0,\epsilon)\}-\log P(\mathscr{S}_i=\mathscr{K}|\mathscr{S}_{<i},I_0,\epsilon)\},\tag{24}$$

where $\bar{\mathscr{K}}_j$ denotes the jth word that not in keywords $\mathscr{K}$. Eq. (24) first calculates the difference between the probability of the most likely word except the keyword and the probability of the keyword. Then, Eq. (24) chooses the most likely position of the keyword as the optimal position based on the differences. The time complexity of Eq. (24) is

$$T_1=|\mathscr{K}|_\dagger\cdot(O(L)+O(1)+O(|\mathbf{S}|_\dagger)),\tag{25}$$

where $|\mathscr{K}|_\dagger$ represents the number of keywords, L is the number of vocabularies, and $|\mathbf{S}|_\dagger$ denotes the maximum length of generated captions. The simplified keyword appearing attack is expressed on Eq. (7), which directly selects the word position with the maximum keyword probability as the best position. The time complexity of Eq. (7) is expressed as

$$T_2=|\mathscr{K}|_\dagger\cdot(O(1)+O(|\mathbf{S}|_\dagger)),\tag{26}$$

$T_2\ll T_1$ since $|\mathbf{S}|_\dagger\ll L$, i.e. $|\mathbf{S}|_\dagger=20$, $L=12000$ in MSCOCO. The comparative experimental results are shown in Table 6. The simplified keyword appearing attack exhibits comparable attack performance and better sentence score than the attack in [4].

5. Conclusions

In this paper, we study several user-oriented adversarial attacks on image captioning models to protect personal information. We experimentally demonstrate that image captioning models are more vulnerable when adopting captions generated for the target images as attack targets than adopting ground-truth captions as attack targets. Extensive experiments show that proposed attacks realize high attack success rates while the adversarial perturbations are still imperceptible to humans.

Data availability

We provide the code link in the paper.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (No. U1936117, No. 62106037, No. 62076052), the Science and Technology Innovation Foundation of Dalian (No. 2021JJ12GX018), the Open Project Program of the National Laboratory of Pattern Recognition (NLPR)(No.202100032), and the Fundamental Research Funds for the Central Universities (DUT21GF303, DUT20TD110, DUT20RC(3)088).

Appendix A. Meaningless filter and similarity space

Table A.7 shows the annotated meaningless filter set. For any word in the set will not participate in the attacking process.

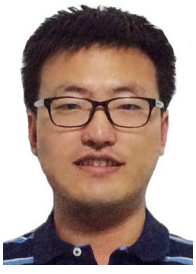
Table A.8 shows the pre-defined similarity space. All words in the same similarity subspace share the same or similar meaning.

References

[1] N. Akhtar, A. Mian, Threat of adversarial attacks on deep learning in computer vision: A survey, IEEE Access 6 (2018) 14410–14430.
[2] N. Carlini, D. Wagner, Towards evaluating the robustness of neural networks, in: 2017 IEEE Symposium on Security and Privacy (SP), IEEE, 2017, pp. 39–57.
[3] A. Chaturvedi, U. Garain, Mimic and fool: A task agnostic adversarial attack. arXiv: Computer Vision and Pattern Recognition, 2019.
[4] H. Chen, H. Zhang, P.-Y. Chen, J. Yi, C.-J. Hsieh, Attacking visual language grounding with adversarial examples: A case study on neural image captioning, in: ACL, 2018.
[5] K. Eykholt, L. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, D. Song, Robust physical-world attacks on deep learning models. arXiv: Cryptography and Security, 2017.
[6] I. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples. arXiv: Machine Learning, 2014.
[7] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[8] S. Herdade, A. Kappeler, K. Boakye, J. Soares, Image captioning: Transforming objects into words, in: Advances in Neural Information Processing Systems, 2019, pp. 11137–11147.
[9] A. Kurakin, I. Goodfellow, S. Bengio, Adversarial examples in the physical world. arXiv: Computer Vision and Pattern Recognition, 2016.
[10] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: Common objects in context, in: European conference on computer vision, Springer, 2014, pp. 740–755.
[11] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, 2017. arXiv preprint arXiv:1706.06083.
[12] V. Mirjalili, A. Ross, Soft biometric privacy: Retaining biometric utility of face images while perturbing gender, in: 2017 IEEE International joint conference on biometrics (IJCB), IEEE, 2017, pp. 564–573.
[13] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, P. Frossard, Universal adversarial perturbations, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1765–1773.
[14] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z.B. Celik, A. Swami, The limitations of deep learning in adversarial settings, in: 2016 IEEE European Symposium on Security and Privacy (EuroS&P), IEEE, 2016, pp. 372–387.
[15] N. Papernot, P. McDaniel, A. Swami, R. Harang, Crafting adversarial input sequences for recurrent neural networks, in: MILCOM 2016–2016 IEEE Military Communications Conference, IEEE, 2016, pp. 49–54.
[16] S.J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, V. Goel, Self-critical sequence training for image captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7008–7024.
[17] A. Rozsa, M. Günther, E.M. Rudd, T.E. Boult, Facial attributes: Accuracy and adversarial robustness, Pattern Recognition Letters (2017).
[18] U. Shaham, Y. Yamada, S. Negahban, Understanding adversarial training: Increasing local stability of supervised models through robust optimization, Neurocomputing 307 (2018) 195–204.
[19] H. Shi, J. Mao, T. Xiao, Y. Jiang, J. Sun, Learning visually-grounded semantics from contrastive adversarial samples, in: COLING, 2018.
[20] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv: Computer Vision and Pattern Recognition, 2014.
[21] J. Su, D.V. Vargas, K. Sakurai, One pixel attack for fooling deep neural networks, IEEE Transactions on Evolutionary Computation (2019).
[22] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks. arXiv: Computer Vision and Pattern Recognition, 2013.
[23] R. Taori, A. Kamsetty, B. Chu, N. Vemuri, Targeted adversarial examples for black box audio systems, in: 2019 IEEE Security and Privacy Workshops (SPW), IEEE, 2019, pp. 15–20.
[24] B. Wang, M. Zhao, W. Wang, X. Dai, Y. Li, Y. Guo, Adversarial analysis for source camera identification, IEEE Transactions on Circuits and Systems for Video Technology (2020).
[25] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, A. Yuille, Adversarial examples for semantic segmentation and object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1369–1378.
[26] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: International conference on machine learning, 2015, pp. 2048–2057.
[27] Y. Xu, B. Wu, F. Shen, Y. Fan, Y. Zhang, H.T. Shen, W. Liu, 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, 2019, pp. 4135–4144.
[28] P. Young, A. Lai, M. Hodosh, J. Hockenmaier, From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions, Transactions of the Association for Computational Linguistics 2 (2014) 67–78.
[29] X. Yuan, P. He, Q. Zhu, X. Li, Adversarial examples: Attacks and defenses for deep learning, IEEE Transactions on Neural nNetworks and Learning Systems (2019).
[30] F. Zhao, J. Wang, Y. Wu, M. Tang, Adversarial deep tracking, IEEE Transactions on Circuits and Systems for Video Technology 29 (7) (2019) 1998–2011.
[31] M. Zhou, Z. Niu, L. Wang, Q. Zhang, G. Hua, Adversarial ranking attack and defense, 2020. arXiv preprint arXiv:2002.11293.

Mengnan Zhao is pursuing his Ph.D. degree in the School of Computer Science and Technology from the Dalian University of Technology (DUT). His research interests include adversarial attack and cross-modal knowledge graph reasoning.

Bo Wang his B.S. degree in Electronic and Information Engineering, M.S.degree and Ph.D. degree in Signal and Information Processing from Dalian University of Technology, Dalian, China, in 2003, 2005 and 2010, respectively. From 2010 to 2012, he was a post-doctoral research associate in Faculty of Management and Economics in Dalian University of Technology. He is currently an Associate Professor in School of Information and Communication Engineering in Dalian University of Technology. His current research interests focus on the areas of multimedia processing and security, such as digital image processing and forensics.



Weikuo Guo received the B.E. degree from the School of Information and Communication Engineering, Dalian University of Technology, Dalian, China, in 2015. He is currently pursuing the D.E. degree with the School of Information and Communication Engineering, Dalian University of Technology, Dalian, China. His research interests include Cross-Model retrieval and machine learning.

Wei Wang received his B.E. degree in computer science and technology from North China Electric Power University in 2007. Since 2012, he has been with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, where he is currently an assistant professor. His research interests include pattern recognition, image processing, and digital image forensics, including watermarking, steganalysis, and tampering detection.