5-2022

Learning transferable perturbations for image captioning

Hanjie WU

Yongtuo LIU

Hongmin CAI

Shengfeng HE
Singapore Management University, shengfenghe@smu.edu.sg

Learning Transferable Perturbations for Image Captioning

HANJIE WU, YONGTUO LIU, HONGMIN CAI, and SHENGFENG HE,
South China University of Technology, China

Present studies have discovered that state-of-the-art deep learning models can be attacked by small but well-designed perturbations. Existing attack algorithms for the image captioning task is time-consuming, and their generated adversarial examples cannot transfer well to other models. To generate adversarial examples faster and stronger, we propose to learn the perturbations by a generative model that is governed by three novel loss functions. Image feature distortion loss is designed to maximize the encoded image feature distance between original images and the corresponding adversarial examples at the image domain, and local-global mismatching loss is introduced to separate the mapping encoding representation of the adversarial images and the ground true captions from a local and global perspective in the common semantic space as far as possible cross image and caption domain. Language diversity loss is to make the image captions generated by the adversarial examples as different as possible from the correct image caption at the language domain. Extensive experiments show that our proposed generative model can efficiently generate adversarial examples that successfully generalize to attack image captioning models trained on unseen large-scale datasets or with different architectures, or even the image captioning commercial service.

CCS Concepts: • **Security and privacy** → Domain-specific security and privacy architectures;

Additional Key Words and Phrases: Adversarial examples, image captioning, robustness of neural network

ACM Reference format:
Hanjie Wu, Yongtuo Liu, Hongmin Cai, and Shengfeng He. 2022. Learning Transferable Perturbations for Image Captioning. *ACM Trans. Multimedia Comput. Commun. Appl.* 18, 2, Article 57 (February 2022), 18 pages.
https://doi.org/10.1145/3478024

1 INTRODUCTION

Recently, neural network-based methods have achieved great performance on many tasks [26, 39]. However, researches [10, 30, 41] found that deep learning networks are easy to be attacked by adversarial examples that consist of original images and hand-crafted perturbations. Adversarial examples have no impact on human perceptions but cause the deep learning model to output

This project is supported by the National Natural Science Foundation of China (No. 61972162); Guangdong International Science and Technology Cooperation Project (No. 2021A0505030009); Guangdong Natural Science Foundation (No. 2021A1515012625); Guangzhou Basic and Applied Research Project (No. 202102021074); and the CCF-Tencent Open Research fund (No. CCF-Tencent RAGR20190112).
Authors' address: H. Wu, Y. Liu, H. Cai, and S. He (corresponding author), School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China; emails: cshanjiewu@gmail.com, csmanlyt@mail.scut.edu.cn, {hmcai, hesfe}@scut.edu.cn.

1551-6857/2022/02-ART57 $15.00
https://doi.org/10.1145/3478024

(a) A man riding on top of a elephant.

(b) Two dogs run through a field.

(c) A close up of a bird hanging from a ceiling.

Fig. 1. Our adversarial examples and the corresponding error output captions generated by the attacked image captioning model.

error results. Previous works have paid attention to find adversarial examples against tasks such as semantic segmentation [34], text classification [14], and speech recognition [38].

Notwithstanding the demonstrated success on these tasks, adversarial examples are rarely explored in image captioning. It is a task to generate a natural language caption that describes the visual contents of a given image. It has many important applications [12], e.g., image indexing, tagging flags on pictures uploaded to the social media, and helping blind people to understand the world. The attack to image captioning model may mislead the user to perform improper behavior. Therefore, the robustness of an image captioning model is important. Different from the classification task where their prediction is restricted to a limited number of classes, image captioning tasks generate grammatically correct and meaningfully correct captions but there are tons of captions expressing the same meaning. If we consider each caption as one kind of class, then there are large numbers of classes in the distribution of the captions in the image captioning task. This makes it difficult to generate adversarial examples for the image captioning task. There are few works [5, 37, 42] to generate adversarial examples for image captioning models. These three methods both mainly focus on targeted attack that aims to output pre-specified captions. To find adversarial examples, previous methods formulate the attack process as an traditional optimization problem, which often needs a large number of iterations (around 1,000 times in these methods) to generate each adversarial example. For each new image, the previous attack algorithms need to iteratively solve the optimization problem. This makes them time- and computational-consuming and not suitable for real-time attack. Besides, they only impose loss functions on the final prediction in the single language domain without considering the property in the image domain, which limits the generalization capability of generated adversarial examples.

In this article, unlike previous works, we address the non-targeted attack problem on image captioning by formulating the process of crafting adversarial examples as a generative problem. In particular, we design a generative model to produce adversarial perturbations, which are added on the input image to fool image captioning models (see Figure 1). Additionally, we tailor three novel loss functions to train the perturbation generator, which jointly consider both image and language domains. These three losses are inspired by the basic structure of the image captioning model. We found that the existing well-performing image captioning models are mostly composed of two parts. One part consists of the convolutional neural network, and the other part consists of the recurrent neural network. First, image feature distortion loss is proposed to maximize the distance between original images and the generated corresponding adversarial examples in the encoding image feature space. The extracted image features will be fed into the recurrent neural network later, so larger distortion in the encoding image features will cause higher possibilities to output an error caption. Second, we present a local-global mismatching loss to measure the

matching degree across image and caption domains. We utilize this loss function to reduce the matching degree between the adversarial image and the corresponding true caption to generate adversarial examples. Although there exist some metrics such as BLEU [27], CIDEr [31], ROUGE_L [20], METEOR [4], SPICE [1], and WMD [15] to measure the matching degree between images and captions, they are not differentiable and cannot be directly used as the loss function at training time. Therefore, to measure the matching degree across domains, we train two additional neural networks to encode images and captions into a common semantic feature space. In this feature space, the distance between the image and its corresponding caption is as close as possible, while the distance of the same image and the irrelevant caption is as far as possible. Third, language diversity loss is used to make the adversarial captions generated by the generated adversarial image and the true caption of the original image as different as possible in the language domain.

Once the generative model is trained, each adversarial example can be generated by only one forward pass instead of iterative optimization. Besides, different from previous optimization formulation where adversarial examples are generated separately without considering other adversarial examples, our generative model is trained in a dataset considering more general patterns. Together with our proposed loss functions, our generative model can significantly improve the generalization ability of adversarial examples. Extensive experimental results demonstrate the effectiveness and generalization ability across different settings. Below, we summarize our contributions:

- We propose to formulate the attack process for image captioning as a generative problem and design a generative model to efficiently and robustly generate adversarial examples.
- We tailor three novel loss functions from three different perspectives, i.e., image feature distortion loss, local-global mismatching loss, and language diversity loss. They jointly govern the network in image and language domains, and the generative model learns more transferable adversarial examples due to these loss functions.
- Extensive experiments show that our proposed generative model can produce adversarial examples that successfully generalize to the image captioning model trained on unseen large-scale datasets, or other image captioning models with different architectures, or even the image captioning commercial service.

2 RELATED WORK

2.1 Image Captioning

Image caption generation is a multimodal task in the fields of computer vision and natural language processing. The general pipeline of image captioning is to extract the visual information of the input image first and then use the visual information to generate the image caption by the language model. Many network architectures and learning strategies [8, 9, 13, 29, 32, 35] have been proposed to make the generated image caption more accurate and diverse. In general, an image caption generation network consists of two parts: One part is an image visual feature extraction network that usually consists of convolutional layers, and the other part is a natural language caption generation network based on recurrent neural layers. At the same time, various evaluation metrics [1, 4, 15, 20, 27, 31] are also proposed for measuring the accuracy and diversity of the generated caption. Although the performance of image captioning tasks has reached a good performance, only a few studies [5, 37, 42] have investigated the robustness of this task.

2.2 Adversarial Attacks

Adversarial attacks are to fool deep learning models through well-designed adversarial examples. Adversarial examples and original images are supposed to have visually similar perception to humans but different meanings to deep learning models. In recent research, various methods

have been proposed to generate adversarial examples for semantic segmentation tasks [34], text classification tasks [14], and speech recognition tasks [38]. For example, Goodfellow et al. [10] propose the **fast gradient sign method (FGSM)** to generate adversarial examples by adding the perturbation in the direction of the sign of the gradient of loss function. It is a single-step attack method. Kurakin et al. [19] improve FGSM by proposing an iterative FGSM unlike the previous single-step attack. This iterative attack method is more effective than the single-step attack in the white-box attack setting. Dong et al. [7] find the adversarial examples in the gradient direction that increases the classification loss with momentum information in the optimization process. This is a momentum-based Iterative FGSM. Xie et al. [33] propose that introducing the operation of input randomization during the iterative attack can increase the robustness of the attack effect. Besides, Baluja and Fischer [3] train an auto-encoder to take original images as input and output adversarial examples that can make the classifier misclassify, and they optimize the parameters of the network through the L_2 norm and classification loss. As for Image captioning tasks, Chen et al. [5] utilize L_2 distance metric and captioning loss to craft adversarial examples that make image captioning model output a caption containing specific words. Xu et al. [37] treat the attack problem as two kinds of formulations: One is the structured output learning with latent variables, and the other is the log marginal likelihood problem optimized by GEM algorithm. Zhang et al. [42] crafted perturbation that is based on semantic embedding of the targeted caption. Different from their methods, we propose a learning method to generate image adversarial examples by using neural networks with three well-designed losses. Furthermore, our method focuses on non-targeted attacks.

3 PROPOSED METHOD

3.1 Problem Formulation

The attacked image captioning model is represented by the function I, which takes an image $X \in \mathbb{R}^{3\times H\times W}$ as input and outputs a caption C=$(C_1, C_2, \ldots, C_N)$. X represents an original image input and C is the caption generated from the image captioning model. We define an adversarial example X' as the image that adds a certain amount of perturbations $\delta \in \mathbb{R}^{3\times H\times W}$ generated by perturbations generator G to the original image X. We restrict the image distortion under a limited upper ϵ in order not to affect human perception about the original image. The above formulations can be described as following equations:

$$I(X) = C, \tag{1}$$

$$G(X) = \delta, \quad I(X + \delta) = C', \tag{2}$$

$$X' = X + \delta, \quad \|X' - X\|_\infty \leq \epsilon. \tag{3}$$

Because the output of the image captioning task is a sentence that contains many sequential words, we cannot simply measure the success of the attack from one label to another label like the classification task does. To measure the attack performance of the generated adversarial examples, we will calculate the evaluation metrics such as BLEU according to the strength of reduction on these metrics to measure the attack performance.

3.2 Method Framework

Figure 2 illustrates the whole architecture of our proposed method. It mainly consists of four parts: a perturbations generator G, two mapping encoders M_{image} and $M_{caption}$, and the target attacked image captioning model I. The G takes the image X as input and generates a perturbation δ. The δ will be clipped to the limited upper bound of perturbations ϵ and then $X + \delta$ is fed into the image captioning model I. We can get image encoding features and the final predicted caption from the I. Our loss function designs are based on these information. The M_{image} and $M_{caption}$ are fixed

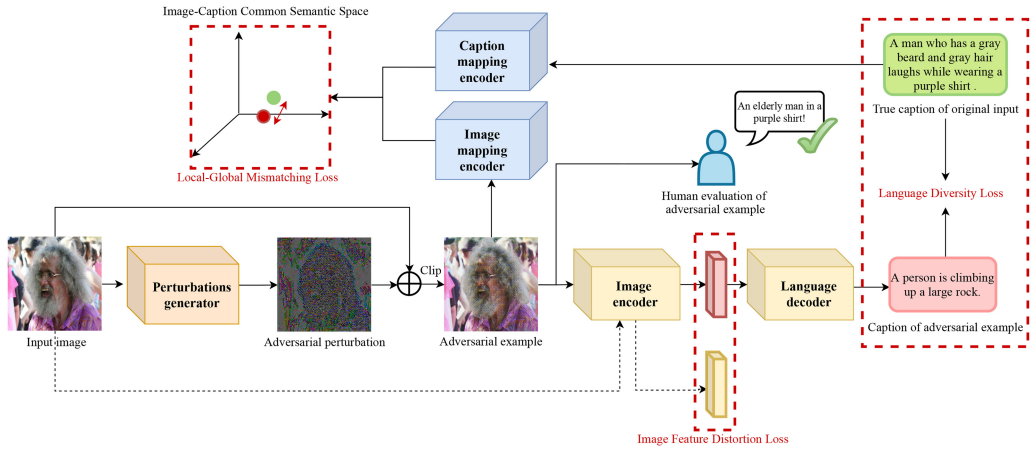


Fig. 2. Illustration of the proposed perturbations generating architecture. Our perturbations generator considers input image to generate the attacking perturbation. We add the adversarial perturbation back to the input image while applying a clip function to restrict the applied perturbations. Then, the generated adversarial example is fed into a target image captioning model. Our perturbations generator is encouraged to mislead output captions by the image feature distortion loss, local-global mismatching loss, and language diversity loss that jointly consider the correlation between image and language domains. Notice that, due to these loss items, the output caption of our adversarial examples is completely unrelated to the input image. Meanwhile, the adversarial examples will not influence the human evaluation about the images.

pre-trained models when training the G, and the trained M_{image} and $M_{caption}$ can map images and captions to a common semantic space. In this space, the spatial distance between the image and the corresponding captions is close, while the pair of mismatched images and captions are far away. To train this multimodal problem, we tailor the image feature distortion loss, local-global mismatching loss, and language diversity loss to optimize the perturbations generator.

3.3 Loss Functions

Image feature distortion loss is motivated by the general image captioning model architecture. We observed that most of the current image captioning models will first use the basic **convolutional neural network (CNN)** such as VGG and ResNet to extract visual features. These visual encoding features of images will be subsequently fed into the **recurrent neural network (RNN)** model for caption generation. If perturbing the image visual features of the input image, then the perturbed visual encoding features can directly affect the output caption from the beginning of the RNN model. This will greatly increase the possibility of error captions output. So, we design this loss function to disturb the image encoding features. Specifically, the visual encoding features of adversarial examples are as far as possible from the visual encoding features of the original image. This is different from the previous attack methods on image captioning. Previous works use the output logits predictions of the RNN network to generate adversarial examples, and previous methods did not consider the intermediate visual encoding features. This loss function is described as follows:

$$E(X) = F, \quad E(X') = F', \tag{4}$$

$$L_{feature} = \frac{\sum_{i=1}^{n}(F_i \times F_i^{'})}{\sqrt{\sum_{i=1}^{n}(F_i)^2} \times \sqrt{\sum_{i=1}^{n}(F_i^{'})^2}}, \tag{5}$$

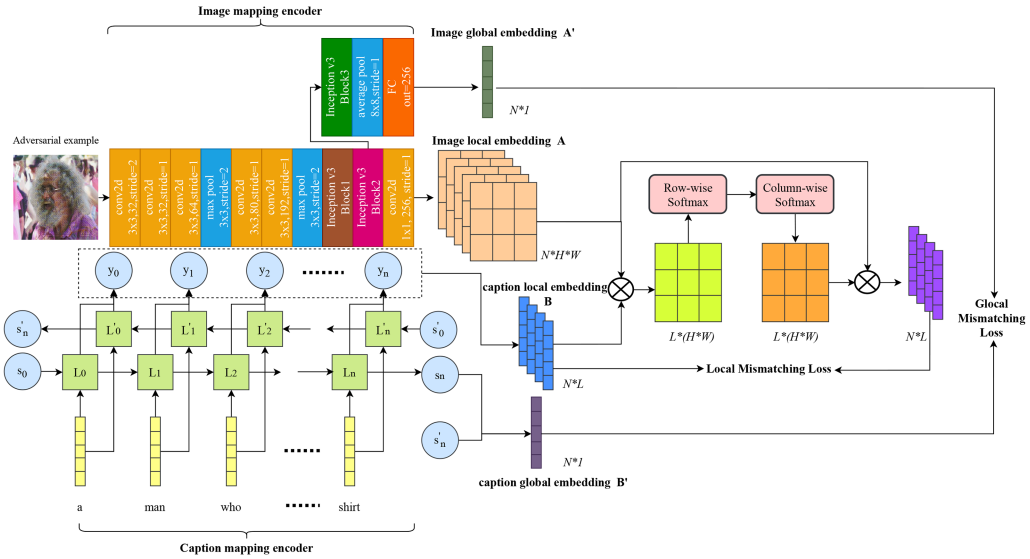


Fig. 3. Illustration of the local-global mismatching loss.

where E represents the CNN part of the image captioning model. F is the encoding image feature of original inputs, and F' is the encoding image feature of adversarial examples. Equation (5) shows the way to measure the distance of these two encoding image features. We use the cosine similarity to measure the distance between them. The smaller the cosine similarity, the larger the distance between the two encoding features. Our optimization goal is to minimize $L_{feature}$.

Local-global mismatching loss is used to measure the matching degree between images and captions. It is inspired by other image-language tasks [28, 36]. We assume the common semantic space between image and caption is a N-dimension space. First, we introduce the two mapping encoders, as Figure 3 shows. The image mapping encoder M_{image} is a convolutional network that is adapted from InceptionV3 network to extract image features, M_{image} takes the image as input and output the local image feature vector $A \in \mathbb{R}^{N\times H\times W}$ from the last layer output of InceptionV3 block2 and additional 1×1 convolutional layer, where N is the dimension of the common semantic space and $T = H \times W$ is the output size of feature map. The local image feature vector A represents the high-level information of an image and has a large receptive field on the input image, which means that $A_j \in \mathbb{R}^N$ contained information of jth local part of the input image. Besides, M_{image} also outputs the global image feature $A' \in \mathbb{R}^{N\times 1}$ by adding a fully connected layer. The A' represents the global characteristics of the entire image. Also, the caption mapping encoder $M_{caption}$ consists of bi-directional long short-term memory units to transfer the caption to local caption features $B \in \mathbb{R}^{N\times L}$ by stacking the encoded output feature of different steps, where the N is the dimension number of image-caption common semantic space and L is the caption length of the image. The B_j represents the information of jth word in a caption. The global caption features $B' \in \mathbb{R}^{N\times 1}$ is obtained by concatenating the final output of the LSTM unit in both directions. After getting the local and global encoding features (A, A', B, and B') about image and caption, we can perform the local mismatching operation. We calculate the local attention matrix between image and caption

as follows:

$$S = \sigma(B^T A), \tag{6}$$

$$D_i = \sum_{j=0}^{T-1} \mu_j A_j, \quad \mu_j = \frac{exp(S_{i,j})}{\sum_{k=0}^{T-1} exp(S_{i,k})}, \tag{7}$$

where $S \in \mathbb{R}^{L \times T}$. We first use row-wise softmax function and then use column-wise softmax function in actual operation. And σ is a row-wise softmax function to normalize the matrix by each column. The equation of calculating μ in Equation (7) is the column-wise softmax function. Our goal is to compute the dense attention for each image patch and word pair. To accelerate the computation and reduce the memory consumption, we approximate this process by cascading the row-wise and column-wise operations following Reference [36]. The element of the local attention matrix $S_{i,j}$ means how much association about ith word of a caption and jth local part of a image. Then, we calculate the D_i is a ith word representation by summing up all local part encoding features of images with different weights according to Equation (7). The weight μ is calculated by normalizing the attention matrix. We can know $D \in \mathbb{R}^{N \times L}$, and D is a new caption representation that is calculated based on the information of images and the image-caption local attention matrix. Under non-attack circumstance, the two features B and D are similar or close, and we can use Equation (8) to measure whether a caption matches an image in local image and local caption part:

$$L_{local} = -log\left(\sum_{i=0}^{L-1} exp\left(\frac{D_i^T B_i}{\|D_i\| \, \|B_i\|}\right)\right). \tag{8}$$

Unlike the non-attack circumstance, our training target is to increase the mismatching degree by continuously adding perturbations to the original image. The above loss function considers the local information of the image and captions. Similarly, we have extracted the global features of the image and captions. We can also increase mismatching degree of the global features to guide the generation of adversarial perturbations by using Equation (9):

$$L_{global} = -log\left(\sum_{i=0}^{L-1} exp\left(\frac{A_i'^T B_i'}{\left\|A_i'\right\| \left\|B_i'\right\|}\right)\right), \tag{9}$$

$$L_{local-global} = -(L_{local} + \beta L_{global}), \tag{10}$$

$$\arg\min_{\theta_1, \theta_2} \sum_{x_i \in \mathbb{X}} (L_{local} + \beta L_{global}). \tag{11}$$

The image mapping encoder M_{image} and the caption mapping encoder $M_{caption}$ is pre-trained by the real image and caption pairs using Equation (11).

Language diversity loss is to make the adversarial captions generated by the generated adversarial image and the true caption of the original image as different as possible in the language domain. For the previous two loss functions, image feature distortion loss focuses on the image domain and the local-global mismatching loss focuses on the cross image and language domain. To obtain better optimization results in training and increase the performance of generating adversarial examples, we use a language diversity loss to guide the generation of adversarial examples on the language domain. Since the image captioning task is essentially a classification problem, we increase the cross-entropy loss of the classification to expand the semantic distance of the

adversarial captions and true captions in the language domain by Equation (12):

$$L_{language} = -CrossEntropyLoss(C, C'), \tag{12}$$

$$\arg\min_{\theta} \sum_{x_i \in \mathbb{X}} L_{feature} + \alpha L_{local-global} + \gamma L_{language}. \tag{13}$$

In summary, our method is different from traditional attack algorithms in image captioning; traditional attack algorithms only consider increasing the dissimilarity between the adversarial caption and the true caption. Our loss function considers not only increasing the dissimilarity from single image domain by image feature distortion loss and single language domain by language diversity loss, but also increasing the mismatching degree cross image and language domain by local-global mismatching loss. In summary, we propose three loss functions to guide the training process of our perturbations. The final loss function is presented in Equation (13).

4 EXPERIMENT

4.1 Implementation Details

In this section, we introduce the details of the experimental implementation. First, our perturbations generator G uses an encoder-decoder structure based on Reference [3]. The encoder uses the architecture of the Inception-ResNet model that is pretrained on ImageNet dataset [6]. The decoder is stacked with multiple deconvolution layers and non-linear activation functions. M_{image} is based on the Inception-v3 architecture, and $M_{caption}$ is based on bi-directional LSTM units. The white-box attacked image captioning model I is Show-Attend-Tell model [35]. The value of maximum perturbations ϵ is set to 16. The dimension of common semantic space N is 256. Using ADAM [16] optimization, the learning rate of the encoder is 0.0001, and the learning rate of the decoder is 0.0005. α is 1.5, β is 1, and γ is 2 in the loss function; these hyper-parameters are selected from grid search algorithm. Our perturbations generator is trained on Flickr8k [11]. The training set has 6,000 images, and each image corresponds to 5 captions. We conduct quantitative analysis via six common measure metrics (BLEU, CIDEr, ROUGE_L, METEOR, SPICE, and WMD) in the image captioning task and the beam search size is equal to 2 when we calculate the evaluation metrics. These metrics measure the matching degree between the image caption obtained from our generated adversarial example and the true image caption from different perspectives. In this attack task, the low value of these metrics indicates that more image captions generated by our adversarial examples fails to describe the image, and also the generated adversarial example is better for the attack. We also calculate the image quality measure metrics (PSNR, SSIM) to evaluate the distortion strength of the generated adversarial examples. The high values of PSNR and SSIM indicate that the generated adversarial example is visually closer to the original image. And process time is calculated for computing effectiveness.

4.2 Threat Models

In the white box setting, we utilize the attacked image captioning model trained on the Flickr8k training set to train our perturbations generator. We evaluate the image captioning measure metrics in the Flickr8k testing set during testing. In this circumstance, the perturbations generator has seen the attacked network architecture and distribution of testing dataset.

To test the transferability and robustness of our adversarial examples, we test our adversarial examples generated by our model trained on Flickr8k to attack the Show-Attend-Tell image captioning model trained on the Flickr30k [40] and the MSCOCO [21] dataset. We call this a semi-white box attack, because the network architecture of the attacked model used for testing is known

Table 1. Quantitative Analysis of our Generated Adversarial Examples in White Box, Semi-white Box Settings

	No Attack			White Box Attack	Semi-White Box Attack	
	Flickr8K	Flickr30K	MSCOCO	Flickr8K	Flickr30K	MSCOCO
BLEU-1	0.641	0.637	0.737	**0.548**	**0.547**	**0.595**
BLEU-2	0.460	0.458	0.570	**0.340**	**0.362**	**0.402**
BLEU-3	0.325	0.324	0.430	**0.215**	**0.240**	**0.273**
BLEU-4	0.225	0.228	0.324	**0.138**	**0.167**	**0.190**
CIDEr	0.559	0.472	1.021	**0.256**	**0.224**	**0.535**
ROUGE_L	0.482	0.456	0.543	**0.392**	**0.396**	**0.435**
METEOR	0.225	0.213	0.266	**0.178**	**0.185**	**0.207**
SPICE	0.155	0.139	0.191	**0.088**	**0.094**	**0.109**
WMD	0.162	0.144	0.222	**0.094**	**0.096**	**0.127**

during training the perturbations generator, but the training data (MSCOCO, Flickr30k) of the test model is not visible.

We further verify the performance of our adversarial examples in the black box setting. The black-box model we used is SCST [29], in which the image captioning model is trained using reinforcement learning. Specifically, we use four different network architectures (FC [32], Topdown [2], Att2in2 [24], and Transformer-based [17, 25]). For the transformer-based method, we use two pre-trained models trained with bottom-up features [2] and VilLBERT features [22, 23], respectively, to test the model. We use open source code implementation[1] of these models to test our adversarial examples, and these black-box models are trained on the MSCOCO dataset. In this case, both the network architecture and training data of the test model are not seen when training our model.

4.3 Attacks in White, Semi-white, and Black Box Settings

We test the performance and robustness of the adversarial examples in different settings. The test results are shown in Tables 1 and 2. *No attack* indicates the evaluation performance on the captioning model using the clear image testset. We can see that all the evaluation metrics have dropped significantly both in the white box attack and the semi-white box attack. The results in the semi-white box setting show that our adversarial examples can successfully fool the captioning model that fits other data distributions. Among them, the CIDEr has a larger decline in different settings (such as from 1.021 to 0.535 on MSCOCO). The reason is that the CIDEr measures the captions by representing sentences as TF-IDF vectors and then uses the weighted average of the cosine similarity between the vectors to calculate the scores. Our method also increases the distance between adversarial captions and true captions in the common semantic space during training. When it turns to the black box attack, the network structure of the captioning model is different from the training, but the attack is still successful to make the model performance worse. The results show that our generated adversarial examples have good transferability. They also show that the current image captioning methods are not robust enough.

4.4 Ablation Study

We conduct experiments to prove the effectiveness of our designed loss functions and evaluate the influence of different perturbation levels on the attack performance. Table 3 shows the value of different measure metrics on different loss functions in the white box setting. We notice that even using only one of the three loss functions for training the network can achieve fair attack

[1] https://github.com/ruotianluo/self-critical.pytorch.

Table 2. Quantitative Analysis of our Generated Adversarial Examples in Black Box Setting

	No Attack [MSCOCO]				
	SCST Model (FC)	SCST Model (Topdown)	SCST Model (Att2in2)	Transformer-based Model (Bottomup feature)	Transformer-based Model (ViLBERT feature)
BLEU-1	0.746	0.783	0.777	0.793	0.766
BLEU-2	0.574	0.618	0.613	0.637	0.609
BLEU-3	0.426	0.469	0.465	0.493	0.473
BLEU-4	0.313	0.349	0.347	0.374	0.367
CIDEr	1.044	1.172	1.157	1.232	1.165
ROUGE_L	0.539	0.563	0.560	0.575	0.568
METEOR	0.253	0.270	0.267	0.282	0.282
SPICE	0.185	0.204	0.200	0.220	0.212
WMD	0.251	0.239	0.236	0.258	0.262
	Black Box Attack [MSCOCO]				
	SCST Model (FC)	SCST Model (Topdown)	SCST Model (Att2in2)	Transformer-based Model (Bottomup feature)	Transformer-based Model (ViLBERT feature)
BLEU-1	**0.637**	**0.667**	**0.664**	**0.653**	**0.666**
BLEU-2	**0.450**	**0.481**	**0.481**	**0.473**	**0.490**
BLEU-3	**0.313**	**0.337**	**0.340**	**0.336**	**0.357**
BLEU-4	**0.219**	**0.238**	**0.241**	**0.241**	**0.263**
CIDEr	**0.663**	**0.735**	**0.732**	**0.708**	**0.777**
ROUGE_L	**0.464**	**0.481**	**0.481**	**0.475**	**0.487**
METEOR	**0.198**	**0.208**	**0.206**	**0.204**	**0.223**
SPICE	**0.127**	**0.142**	**0.138**	**0.137**	**0.152**
WMD	**0.145**	**0.160**	**0.157**	**0.158**	**0.180**

Table 3. The Ablation Study of Different Loss Functions in Semi-white Box Setting with Flickr30K Testing Set

Loss function	a	b	c	a+b+c
BLEU-1	0.579	0.563	0.558	**0.547**
BLEU-2	0.386	0.378	0.372	**0.362**
BLEU-3	0.256	0.255	0.251	**0.240**
BLEU-4	0.169	0.172	0.171	**0.167**
CIDEr	0.350	0.273	0.268	**0.224**
ROUGE_L	0.422	0.424	0.412	**0.396**
METEOR	0.195	0.197	0.191	**0.185**
SPICE	0.115	0.114	0.112	**0.0937**
WMD	0.123	0.123	0.118	**0.0959**

a: only using image feature distortion loss function. **b**: only using local-global mismatching loss function. **c**: only using language diversity loss function. **a+b+c**: using both above loss functions.

performances. At the same time, the combination of the three loss functions from the image feature space, the common image caption semantic space, and language space makes the network generate stronger adversarial examples. In Table 4, we show the attack effect with the maximum perturbations ranging from 0 to 32 in the black box setting. It can be seen that when the perturbation level is relatively small (such as 8 and 12), the generated adversarial examples also

Table 4. The Attack Performance of Different Levels of Perturbation under Black Box Setting with MSCOCO Testing Set

Perturbation Level	**0**	**8**	**12**	**16**	**20**	**32**
BLEU-1	0.737	0.712	0.674	0.595	0.538	0.433
BLEU-2	0.570	0.540	0.465	0.402	0.338	0.217
BLEU-3	0.430	0.400	0.358	0.273	0.216	0.112
BLEU-4	0.324	0.297	0.260	0.190	0.143	0.064
CIDEr	1.021	0.925	0.792	0.535	0.375	0.104
ROUGE_L	0.543	0.523	0.494	0.435	0.394	0.323
METEOR	0.266	0.243	0.221	0.207	0.148	0.097
SPICE	0.190	0.174	0.152	0.110	0.082	0.031
WMD	0.222	0.199	0.174	0.127	0.098	0.050

Fig. 4. Visual comparison of original images and adversarial examples with their corresponding output captions.

can cause performance degradation on the model. When the perturbation level becomes larger (such as 32), some metrics (such as SPICE and WMD) even drop to close to 0.

4.5 Visualization of Our Adversarial Examples

To better demonstrate our attack performance, we select some pairs of images and captions in each attack setting. The result is shown in Figure 4. The adversarial examples we generated will

Fig. 5. Visualization of our generated perturbations that attack the captioning model successfully. For better visualization, we enlarge the perturbations level.

not affect human judgment visually. The generated perturbations does not cover the main objects in the original image, and people can still correctly identify the content of our adversarial examples. However, the output adversarial captions are inconsistent with the images. Specifically, the perturbations we generate are able to not only change the main object of the image (from *a dog* to *a young girl* in the first column, from *a motorcycle* to *a vase* in the third column), but also change the action of the image object (from *doing a trick* to *climbing a wall* in the second column). Although we did not specify the output caption when training the perturbations generator, we maximize the features distance of the intermediate layer and the distance between images and captions in the common semantic space, making the learned perturbations fool the model. From Figure 5, We visualize the perturbations generated by our trained perturbations generator. Since we limit the maximum size of the perturbations to 16, which is tiny, we multiply the generated perturbations image by a constant to amplify the perturbations to better observe the generated perturbation pattern. We can see that our method adds perturbations in different key areas for different images. And it will not or only generate a small amount of perturbations in the areas that have little effect on the caption generation. The previous method generally adds perturbations to the whole areas of each image during iteration. Our method reduces the generation of unnecessary perturbations. For example, we observe that the generated perturbations are mainly distributed in the foreground object, which is decisive factor of image caption generation, while almost no perturbation is added to the sky or the ground.

4.6 Comparisons with Previous Methods

Three previous works [5, 37, 42] are related to this task. Since References [37, 42] mainly focus on the targeted attack, which is different from our task, we compare the Show-and-fool [5] method only in here. Meanwhile, we compare four general non-targeted attack methods: f**ast gradient sign method (FGSM)** [10], **iterative fast gradient sign method (I-FGSM)** [19], **momentum-based iterative fast gradient sign method (MI-FGSM)** [7], and **MI-FGSM with randomization-based input (Mi-FGSM-DIV)** [33]. For the fairness of comparisons, we limit the maximum perturbations of the adversarial examples generated by these methods to less than 16 like our method does. We test the transferability performance in black box setting. Besides, we generate random perturbation with the same perturbation upper limitation to compare the effect of our method. The experimental results are shown in Table 5. It can be seen that random

Table 5. Comparisons between our Method and other Methods

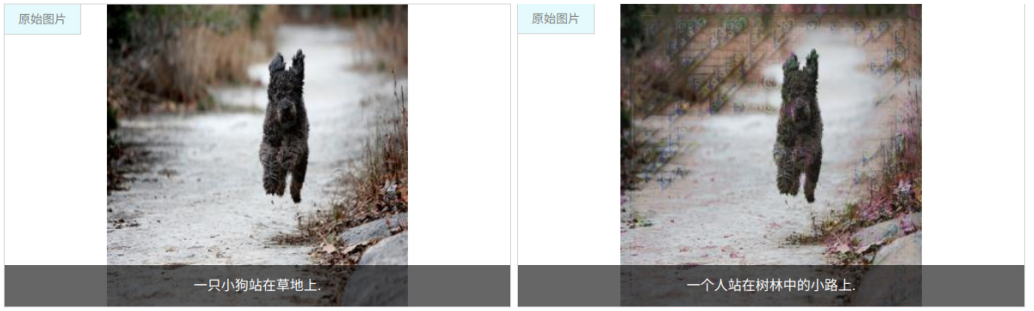
Metrics	No Attack	Random Noise	Show-and-fool [5]	FGSM [10]	I-FGSM [19]	MI-FGSM [7]	MI-FGSM-DIV [33]	Our
BLEU1	0.783	0.779	0.708	0.707	0.724	0.713	0.706	**0.667**
BLEU2	0.618	0.613	0.528	0.526	0.546	0.534	0.525	**0.481**
BLEU3	0.469	0.463	0.382	0.380	0.399	0.387	0.380	**0.337**
BLEU4	0.349	0.344	0.276	0.272	0.288	0.277	0.272	**0.238**
CIDEr	1.172	1.155	0.887	0.878	0.942	0.907	0.876	**0.735**
ROUGE_L	0.563	0.560	0.511	0.508	0.521	0.512	0.507	**0.481**
METEOR	0.270	0.268	0.230	0.230	0.239	0.233	0.229	**0.208**
SPICE	0.204	0.203	0.162	0.163	0.173	0.168	0.163	**0.142**
WMD	0.239	0.236	0.186	0.185	0.197	0.191	0.185	**0.160**
PSNR	-	**28.50**	25.82	26.53	27.83	27.80	27.37	26.43
SSIM	1	**0.866**	0.763	0.831	0.856	0.853	0.842	0.833
Time	-	-	44s	**0.1 s**	3.5 s	3.5 s	3.6 s	0.5 s

perturbations cause little drop on image captioning evaluation metrics. This shows that randomly generated perturbations is difficult to attack image captioning models. The performance of our method is better than other methods. We observe that the transferability attack effect of the previous method is only slightly better than the random perturbation, and the adversarial examples generated by our method makes the model perform the worst, which means better attack performance. Although Show-and-fool optimizes the logits output of the caption directly, it only optimizes a single image for each attack that it is extremely easy to overfit. Like Show-and-fool, these iterative attack methods (I-FGSM, MI-FGSM, MI-FGSM-DIV) in the general method are also prone to overfitting, which causes the transferability of these methods to be worse than the single-step method (FGSM) in black box setting. On the contrary, our method uses our tailored loss functions to train the network, and the training data is not a single image. More general perturbation pattern can be learned from a large amount of data. Also, we use evaluation metrics (PSNR and SSIM) to evaluate the image quality of our adversarial examples. The results show that the adversarial examples generated by our method do not drop much in PSNR and SSIM metrics and are better than the previous Show-and-fool method. It reveals that our adversarial examples are closer to the visual perception of the original images. Finally, we also test the time of generating one image. Our method is faster than Show-and-fool and other iterative methods (I-FGSM, MI-FGSM, MI-FGSM-DIV), because we generate adversarial examples by one forward pass, while the method of Show-and-fool requires around 1,000 iterations for each image. Our method needs to be processed by a neural network, so it is a bit slower than the single-step attack method (FGSM), but the time is in the same order of magnitude. This shows that our method can efficiently generate high-quality adversarial examples.

4.7 Attacks on Commercial Image Captioning System

We use the commercial image captioning interface provided by Tencent AI Lab[2] to test the performance of our generated adversarial examples in a black box cross-lingual setting. It is hard to attack, because we neither know the architecture of the model nor the distribution of the dataset. As can be seen from Figure 6, we find that the adversarial example fools the commercial captioning service to recognize the dog as the human. It is interesting that our adversarial examples can fool the model trained in a different language.

[2]https://ai.qq.com/product/visionimgidy.shtml#express.



(a) Original Image (b) Adversarial Example

Fig. 6. An illustration of clear original image and our adversarial example on the commercial image captioning service. The output caption of original image in English: *A puppy standing on the grass.* The output caption of adversarial example in English: *A man standing on the path in the woods.*

Table 6. Human Evaluation of Our Method

	Question 1		Question 2		Question 3
Ground True Caption	69	Original Image	70	Yes	220
Adversarial Caption	1	Adversarial Image	68	No	20

Question 1: Given an image, select the sentence caption that best matches the image (single choice question). Question 2: Given a sentence caption, select images that match the sentence description (multiple choice question). Question 3: Given a generated adversarial caption, judge whether the caption is grammatically correct and clearly expressed (single choice question).

4.8 Human Evaluation

We conduct a user study experiment and set up three types of questions. The results are shown in Table 6. Question 1 is given an image and selects the sentence caption that best matches the image. Question 1 is the single choice question. Question 2 is given a sentence caption and selects images that match the sentence description. Question 2 is the multiple choice question. We collect the questionnaire data from 70 users. Question 3 is given a generated adversarial caption; the users need to judge whether the sentence is grammatically correct and clearly expressed. Question 3 is the single choice question. For Question 3, We randomly select six captions generated by our adversarial examples and collect data from 40 users. From Table 6, we can see that the adversarial caption generated by our adversarial example is very inconsistent with the original image content. As a result, in Question 1, almost no user chooses the adversarial caption we generated. However, the perturbations that our method adds to the original image when generating the adversarial examples is very small and will not affect the perception of user for the image content. Therefore, nearly all users believe that the image content of the adversarial example we generated is consistent to the original image caption. With regard to the captions generated from our adversarial examples, 220 of the collected data are considered by users to be grammatically correct and clearly expressed. Besides, 20 data are considered to be considered unreasonable, and these user-judged unreasonable adversarial captions are mainly focused on one sentence (*a group of birds are sitting on a bed*). Some users think that "the bird standing on the bed" is illogical, so they think the sentence is not clear. But this sentence is grammatically correct and clearly expressed for us. This sentence produces an illogical combination of objects and actions, because it was generated by our adversarial example. This also reflects the validity of our method.

Table 7. Attack Performance with Different Beam Search Size

No Attack [MSCOCO]									
Beam Size	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	ROUGE_L	METEOR	SPICE	WMD
1	0.722	0.551	0.406	0.298	0.973	0.531	0.251	0.185	0.208
2	**0.737**	**0.570**	0.430	0.324	1.021	**0.543**	**0.266**	**0.190**	0.222
3	0.735	0.568	**0.431**	**0.327**	**1.027**	0.542	0.258	0.189	**0.225**
4	0.733	0.566	0.428	0.326	1.019	0.541	0.263	0.188	0.223
Semi-white Box Attack [MSCOCO]									
Beam Size	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	ROUGE_L	METEOR	SPICE	WMD
1	0.594	0.399	0.263	0.177	0.522	0.433	0.179	0.110	0.122
2	0.595	0.402	0.273	0.190	0.535	0.435	**0.207**	0.110	0.127
3	0.600	0.411	0.284	0.201	0.563	**0.440**	0.182	**0.114**	0.133
4	**0.600**	**0.412**	**0.285**	**0.203**	**0.564**	0.439	0.181	0.113	**0.134**

4.9 More Experiments on Attack Performance

We conduct more experiments to further explore the impact of different factors on the attack performance. First, we explore the effect of beam search size. Beam search is a search algorithm used in the test phase for image captioning tasks. In previous experiments, we set the default beam search size to 2. Table 7 shows the performance of the image captioning model under different beam search sizes on no attack and semi-white box attack conditions. The results show that when the beam search size is 2, most evaluation metrics reach the maximum value under no attack condition. When the beam search size is increased to 3 or 4 under semi-white box attack, although the model evaluation metrics are slightly improved, it is still far below the minimum value of the model evaluation metrics under no attack. The results show that the beam search algorithm cannot defend our adversarial examples.

Next, we use different network defense mechanisms when testing the attack adversarial examples. Since there is currently no research on network defense for image captioning tasks, here, we take several classic network defense methods applied to image classification tasks in our experiment. They include Gaussian filter, which represents the traditional denoising method, Xie's method [33], which is to perform multiple random resize and padding operations on the image before input to the network, and an adversarial training strategy [18], which uses the generated adversarial examples as parts of training examples. Table 8 shows the impact of the above three defense strategies on the performance of the image captioning model under no attack and semi-white box attack condition. In no attack condition, all three defense methods have a certain negative impact on the performance of the image captioning model. Among them, the use of Gaussian filter has a greater negative impact on the model. Under the semi-white box attack, Gaussian filter does not remove attack effect from the perturbations in the adversarial examples, and the image captioning model is even worse. Xie's method resists a part of the adversarial examples and makes the model perform better. The reason is that during the resize and pad operations, the integrity of the perturbations added to the image is destroyed, making some adversarial examples invalid. Although Xie's method makes the model perform more robustly, it still has a certain gap from the original performance of the image captioning model without attack. The best defense method is adversarial training. We add the generated adversarial examples to the training set for training the image captioning model, allowing the model learn to correctly describe adversarial examples. The performance of the model after adversarial training is close to that without attack.

Table 8. Attack Performance under Different Defenses

No Attack [MSCOCO]				
	No Defence	Gaussian Filter	Xie Method [33]	Adversarial Training [18]
BLEU-1	**0.737**	0.700	0.734	0.726
BLEU-2	**0.570**	0.528	0.566	0.556
BLEU-3	**0.430**	0.390	0.427	0.421
BLEU-4	**0.324**	0.289	0.321	0.312
CIDEr	**1.021**	0.887	1.008	0.992
ROUGE_L	**0.543**	0.514	0.540	0.532
METEOR	**0.266**	0.237	0.255	0.251
SPICE	**0.190**	0.170	0.187	0.183
WMD	**0.222**	0.194	0.218	0.215
Semi-white Box Attack [MSCOCO]				
	No Defence	Gaussian Filter	Xie Method [33]	Adversarial Training [18]
BLEU-1	0.595	0.588	0.656	**0.705**
BLEU-2	0.402	0.397	0.472	**0.534**
BLEU-3	0.273	0.267	0.335	**0.409**
BLEU-4	0.190	0.182	0.241	**0.295**
CIDEr	0.535	0.508	0.714	**0.886**
ROUGE_L	0.435	0.432	0.477	**0.519**
METEOR	0.207	0.176	0.208	**0.247**
SPICE	0.110	0.107	0.140	**0.168**
WMD	0.127	0.122	0.160	**0.201**

5 CONCLUSION

In this article, we propose three loss functions to train a generative network to produce adversarial examples for the image captioning task. Our method is not only better in attacking performance than previous methods, but also faster and more general. The adversarial examples generated by our method are more transferable, and we test the transferability under white, semi-white, black, commercial-services settings. Results on three datasets (Flickr8K, Flickr30K, and MSCOCO) and the attacked captioning models with different architectures show the effectiveness of our method. Our adversarial examples evaluate the performance of existing image captioning models from another perspective. In addition, our generated adversarial examples can augment the dataset to train more robust image captioning model. In the future, we can apply our method of generating adversarial examples to evaluate other cross-domain tasks of visual and natural language.

REFERENCES

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic propositional image caption evaluation. In *ECCV*.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

[3] Shumeet Baluja and Ian Fischer. 2018. Learning to attack: Adversarial transformation networks. In *AAAI*.

[4] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *IEEvaluation@ACL*.

[5] Hongge Chen, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, and Cho-Jui Hsieh. 2018. Show-and-Fool: Crafting adversarial examples for neural image captioning. In *ACL*.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*.

[7] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. 2018. Boosting adversarial attacks with momentum. In *CVPR*.
[8] Chuang Gan, Zhe Gan, X. He, Jianfeng Gao, and L. Deng. 2017. StyleNet: Generating attractive visual captions with styles. In *CVPR*. 955–964.
[9] Zhe Gan, Chuang Gan, X. He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, L. Carin, and L. Deng. 2017. Semantic compositional networks for visual captioning. In *CVPR*. 1141–1150.
[10] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *ICLR*.
[11] Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. In *IJCAI* 47 (2013), 853–899.
[12] MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A comprehensive survey of deep learning for image captioning. *J. Artif. Intell., Mach. Learn. Soft Comput.*
[13] Lun Huang, Wenmin Wang, Yaxian Xia, and Jie Chen. 2019. Adaptively aligned image captioning via adaptive attention time. In *NeurIPS*.
[14] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2019. Is BERT really robust? Natural language attack on text classification and entailment. In *AAAI*.
[15] Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. 2017. Re-evaluating automatic metrics for image captioning. In *ACL*.
[16] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *ICLR* (12 2014).
[17] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *ACL*. Retrieved from https://doi.org/10.18653/v1/P17-4012.
[18] Alex Kurakin, Dan Boneh, Florian Tramér, Ian Goodfellow, Nicolas Papernot, and Patrick McDaniel. 2018. Ensemble adversarial training: Attacks and defenses. In *ICLR*.
[19] A. Kurakin, Ian J. Goodfellow, and S. Bengio. 2017. Adversarial machine learning at scale. *ArXiv* abs/1611.01236 (2017).
[20] Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *ACL*.
[21] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *ECCV*.
[22] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NIPS*. 13–23.
[23] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In *CVPR*.
[24] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2016. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*. 3242–3250.
[25] Ruotian Luo, Brian L. Price, Scott Cohen, and Gregory Shakhnarovich. 2018. Discriminability objective for training descriptive captions. In *CVPR*. 6964–6974.
[26] Chih-Yao Ma, Yannis Kalantidis, Ghassan AlRegib, Peter Vajda, Marcus Rohrbach, and Zsolt Kira. 2019. Learning to generate grounded image captions without localization supervision. In *CVPR*.
[27] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: A method for automatic evaluation of machine translation. In *ACL*.
[28] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. 2019. MirrorGAN: Learning text-to-image generation by redescription. In *CVPR*.
[29] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2016. Self-critical sequence training for image captioning. In *CVPR* 1179–1195.
[30] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *ICLR*.
[31] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2014. CIDEr: Consensus-based image description evaluation. *CVPR*. 4566–4575.
[32] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *CVPR*.
[33] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. 2018. Mitigating adversarial effects through randomization. In *ICLR*.
[34] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan L. Yuille. 2017. Adversarial examples for semantic segmentation and object detection. *ICCV*. 1378–1387.
[35] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.
[36] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*.

[37] Yan Xu, Baoyuan Wu, Fumin Shen, Yanbo Fan, Yong Zhang, Heng Tao Shen, and Wei Liu. 2019. Exact adversarial attack to image captioning via structured output learning with latent variables. In *CVPR*.
[38] Hiromu Yakura and Jun Sakuma. 2019. Robust audio adversarial example for a physical attack. In *IJCAI*.
[39] Deheng Ye, Zhao Liu, Mingfei Sun, Bei Shi, Peilin Zhao, Hao Wu, Hongsheng Yu, Shaojie Yang, Xipeng Wu, Qingwei Guo, Qiaobo Chen, Yinyuting Yin, Hao Zhang, Tengfei Shi, Liang Wang, Qiang Fu, Wei Yang, and Lanxiao Huang. 2019. Mastering complex control in MOBA games with deep reinforcement learning. In *AAAI*.
[40] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New Similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Ling.* 2 (2014), 67–78.
[41] X. Yuan, P. He, Q. Zhu, and X. Li. 2019. Adversarial examples: Attacks and defenses for deep learning. *IEEE Trans. Neural Netw. Learn. Syst.* 30, 9 (2019), 2805–2824.
[42] S. Zhang, Z. Wang, X. Xu, X. Guan, and Y. Yang. 2020. Fooled by imagination: Adversarial attack to image captioning via perturbation in complex domain. In *ICME*. 1–6.

Received January 2021; revised June 2021; accepted July 2021