Attacking Visual Language Grounding with Adversarial Examples: A Case Study on Neural Image Captioning

Hongge Chen[1*], **Huan Zhang**[23*], **Pin-Yu Chen**[3], **Jinfeng Yi**[4], and **Cho-Jui Hsieh**[2]
[1]MIT, Cambridge, MA 02139, USA
[2]UC Davis, Davis, CA 95616, USA
[3]IBM Research, NY 10598, USA
[4]JD AI Research, Beijing, China
chenhg@mit.edu, ecezhang@ucdavis.edu
pin-yu.chen@ibm.com, yijinfeng@jd.com, chohsieh@ucdavis.edu
[*]Hongge Chen and Huan Zhang contribute equally to this work

Abstract

Visual language grounding is widely studied in modern neural image captioning systems, which typically adopts an encoder-decoder framework consisting of two principal components: a convolutional neural network (CNN) for image feature extraction and a recurrent neural network (RNN) for language caption generation. To study the robustness of language grounding to adversarial perturbations in machine vision and perception, we propose **Show-and-Fool**, a novel algorithm for crafting adversarial examples in neural image captioning. The proposed algorithm provides two evaluation approaches, which check whether neural image captioning systems can be mislead to output some randomly chosen captions or keywords. Our extensive experiments show that our algorithm can successfully craft visually-similar adversarial examples with randomly targeted captions or keywords, and the adversarial examples can be made highly transferable to other image captioning systems. Consequently, our approach leads to new robustness implications of neural image captioning and novel insights in visual language grounding.

1 Introduction

In recent years, language understanding grounded in machine vision and perception has made remarkable progress in natural language processing (NLP) and artificial intelligence (AI), such as image captioning and visual question answering. Image captioning is a multimodal learning task and has been used to study the interaction between language and vision models (Shekhar et al., 2017). It takes an image as an input and generates a language caption that best describes its visual contents, and has many important applications such as developing image search engines with complex natural language queries, building AI agents that can see and talk, and promoting equal web access for people who are blind or visually impaired. Modern image captioning systems typically adopt an encoder-decoder framework composed of two principal modules: a convolutional neural network (CNN) as an encoder for image feature extraction and a recurrent neural network (RNN) as a decoder for caption generation. This CNN+RNN architecture includes popular image captioning models such as Show-and-Tell (Vinyals et al., 2015), Show-Attend-and-Tell (Xu et al., 2015) and NeuralTalk (Karpathy and Li, 2015).

Recent studies have highlighted the vulnerability of CNN-based image classifiers to adversarial examples: adversarial perturbations to benign images can be easily crafted to mislead a well-trained classifier, leading to visually indistinguishable adversarial examples to human (Szegedy et al., 2014; Goodfellow et al., 2015). In this study, we investigate a more challenging problem in visual language grounding domain that evaluates the robustness of multimodal RNN in the form of a CNN+RNN architecture, and use neural image captioning as a case study. Note that crafting adversarial examples in image captioning tasks is strictly harder than in well-studied image classification tasks, due to the following reasons: (i) class attack v.s. caption attack: unlike classification tasks where the class labels are well defined, the output of image captioning is a set of top-ranked captions. Simply treating different captions as distinct classes will result in an enormous number of classes that can even precede the number of training images. In addition, semantically similar

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), pages 2587–2597
Melbourne, Australia, July 15 - 20, 2018.

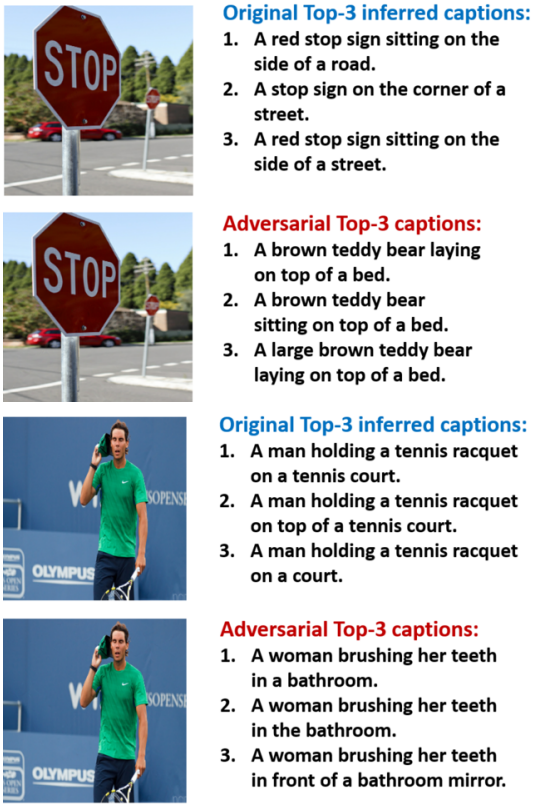

Figure 1: Adversarial examples crafted by Show-and-Fool using the targeted caption method. The target captioning model is Show-and-Tell (Vinyals et al., 2015), the original images are selected from the MSCOCO validation set, and the targeted captions are randomly selected from the top-1 inferred caption of other validation images.

captions can be expressed in different ways and hence should not be viewed as different classes; and (ii) CNN v.s. CNN+RNN: attacking RNN models is significantly less well-studied than attacking CNN models. The CNN+RNN architecture is unique and beyond the scope of adversarial examples in CNN-based image classifiers.

In this paper, we tackle the aforementioned challenges by proposing a novel algorithm called *Show-and-Fool*. We formulate the process of crafting adversarial examples in neural image captioning systems as optimization problems with novel objective functions designed to adopt the CNN+RNN architecture. Specifically, our objective function is a linear combination of the distortion between benign and adversarial examples as well as some carefully designed loss functions. The proposed Show-and-Fool algorithm provides two approaches to craft adversarial examples in neural image captioning under different scenarios:

1. **Targeted caption method:** Given a targeted caption, craft adversarial perturbations to any image such that its generated caption matches the targeted caption.
2. **Targeted keyword method:** Given a set of keywords, craft adversarial perturbations to any image such that its generated caption contains the specified keywords. The captioning model has the freedom to make sentences with target keywords *in any order*.

As an illustration, Figure 1 shows an adversarial example crafted by Show-and-Fool using the targeted caption method. The adversarial perturbations are visually imperceptible while can successfully mislead Show-and-Tell to generate the targeted captions. Interestingly and perhaps surprisingly, our results pinpoint the Achilles heel of the language and vision models used in the tested image captioning systems. Moreover, the adversarial examples in neural image captioning highlight the inconsistency in visual language grounding between humans and machines, suggesting a possible weakness of current machine vision and perception machinery. Below we highlight our major contributions:

- We propose *Show-and-Fool*, a novel optimization based approach to crafting adversarial examples in image captioning. We provide two types of adversarial examples, targeted caption and targeted keyword, to analyze the robustness of neural image captioners. To the best of our knowledge, this is the very first work on crafting adversarial examples for image captioning.
- We propose powerful and generic loss functions that can craft adversarial examples and evaluate the robustness of the encoder-decoder pipelines in the form of a CNN+RNN architecture. In particular, our loss designed for targeted keyword attack only requires the adversarial caption to contain a few specified keywords; and we allow the neural network to *make meaningful sentences with these keywords on its own*.
- We conduct extensive experiments on the MSCOCO dataset. Experimental results show that our targeted caption method attains a 95.8% attack success rate when crafting adversarial examples with randomly assigned captions. In addition, our targeted keyword attack yields an even higher success rate. We also show that attacking CNN+RNN models is inherently different and more challenging than only attacking

CNN models.

- We also show that Show-and-Fool can produce highly transferable adversarial examples: an adversarial image generated for fooling Show-and-Tell can also fool other image captioning models, leading to new robustness implications of neural image captioning systems.

2 Related Work

In this section, we review the existing work on visual language grounding, with a focus on neural image captioning. We also review related work on adversarial attacks on CNN-based image classifiers. Due to space limitations, we defer the second part to the supplementary material.

Visual language grounding represents a family of multimodal tasks that bridge visual and natural language understanding. Typical examples include image and video captioning (Karpathy and Li, 2015; Vinyals et al., 2015; Donahue et al., 2015b; Pasunuru and Bansal, 2017; Venugopalan et al., 2015), visual dialog (Das et al., 2017; De Vries et al., 2017), visual question answering (Antol et al., 2015; Fukui et al., 2016; Lu et al., 2016; Zhu et al., 2017), visual storytelling (Huang et al., 2016), natural question generation (Mostafazadeh et al., 2017, 2016), and image generation from captions (Mansimov et al., 2016; Reed et al., 2016). In this paper, we focus on studying the robustness of neural image captioning models, and believe that the proposed method also sheds lights on robustness evaluation for other visual language grounding tasks using a similar multimodal RNN architecture.

Many image captioning methods based on deep neural networks (DNNs) adopt a multimodal RNN framework that first uses a CNN model as the encoder to extract a visual feature vector, followed by a RNN model as the decoder for caption generation. Representative works under this framework include (Chen and Zitnick, 2015; Devlin et al., 2015; Donahue et al., 2015a; Karpathy and Li, 2015; Mao et al., 2015; Vinyals et al., 2015; Xu et al., 2015; Yang et al., 2016; Liu et al., 2017a,b), which are mainly differed by the underlying CNN and RNN architectures, and whether or not the attention mechanisms are considered. Other lines of research generate image captions using semantic information or via a compositional approach (Fang et al., 2015; Gan et al., 2017; Tran et al., 2016; Jia et al., 2015; Wu et al., 2016; You et al., 2016).

The recent work in (Shekhar et al., 2017) touched upon the robustness of neural image captioning for language grounding by showing its insensitivity to one-word (foil word) changes in the language caption, which corresponds to the *untargeted attack* category in adversarial examples. In this paper, we focus on the more challenging *targeted attack* setting that requires to fool the captioning models and enforce them to generate pre-specified captions or keywords.

3 Methodology of Show-and-Fool

3.1 Overview of the Objective Functions

We now formally introduce our approaches to crafting adversarial examples for neural image captioning. The problem of finding an adversarial example for a given image I can be cast as the following optimization problem:

$$\begin{aligned} \min_{\delta} \;& c \cdot \text{loss}(I+\delta) + \|\delta\|_2^2 \\ \text{s.t.} \;& I+\delta \in [-1,1]^n. \end{aligned} \tag{1}$$

Here δ denotes the adversarial perturbation to I. $\|\delta\|_2^2 = \|(I+\delta) - I\|_2^2$ is an ℓ_2 distance metric between the original image and the adversarial image. $\text{loss}(\cdot)$ is an attack loss function which takes different forms in different attacking settings. We will provide the explicit expressions in Sections 3.2 and 3.3. The term $c > 0$ is a pre-specified regularization constant. Intuitively, with larger c, the attack is more likely to succeed but at the price of higher distortion on δ. In our algorithm, we use a binary search strategy to select c. The box constraint on the image $I \in [-1,1]^n$ ensures that the adversarial example $I+\delta \in [-1,1]^n$ lies within a valid image space.

For the purpose of efficient optimization, we convert the constrained minimization problem in (1) into an unconstrained minimization problem by introducing two new variables $y \in \mathbb{R}^n$ and $w \in \mathbb{R}^n$ such that

$$y = \text{arctanh}(I) \;\text{ and }\; w = \text{arctanh}(I+\delta) - y,$$

where arctanh denotes the inverse hyperbolic tangent function and is applied element-wisely. Since $\tanh(y_i + w_i) \in [-1,1]$, the transformation will automatically satisfy the box constraint. Consequently, the constrained optimization problem in

(1) is equivalent to

$$\min_{w\in\mathbb{R}^n} \quad c\cdot \text{loss}(\tanh(w+y)) \\ +\|\tanh(w+y)-\tanh(y)\|_2^2. \tag{2}$$

In the following sections, we present our designed loss functions for different attack settings.

3.2 Targeted Caption Method

Note that a targeted caption is denoted by

$$S=(S_1,\ S_2,\ ...,\ S_t,\ ...,\ S_N),$$

where S_t indicates the index of the t-th word in the vocabulary list $\mathcal{V}$, S_1 is a start symbol and S_N indicates the end symbol. N is the length of caption S, which is not fixed but does not exceed a predefined maximum caption length. To encourage the neural image captioning system to output the targeted caption S, one needs to ensure the log probability of the caption S conditioned on the image $I+\delta$ attains the maximum value among all possible captions, that is,

$$\log P(S|I+\delta)=\max_{S'\in\Omega}\log P(S'|I+\delta), \tag{3}$$

where Ω is the set of all possible captions. It is also common to apply the chain rule to the joint probability and we have

$$\log P(S'|I+\delta)=\sum_{t=2}^{N}\log P(S'_t|I+\delta,S'_1,...,S'_{t-1}).$$

In neural image captioning networks, $p(S'_t|I+\delta,S'_1,...,S'_{t-1})$ is usually computed by a RNN/LSTM cell f, with its hidden state h_{t-1} and input S'_{t-1}:

$$z_t=f(h_{t-1},S'_{t-1}) \text{ and } p_t=\text{softmax}(z_t), \tag{4}$$

where $z_t:=[z_t^{(1)},z_t^{(2)},...,z_t^{(|\mathcal{V}|)}]\in\mathbb{R}^{|\mathcal{V}|}$ is a vector of the *logits* (unnormalized probabilities) for each possible word in the vocabulary. The vector p_t represents a probability distribution on $\mathcal{V}$ with each coordinate $p_t^{(i)}$ defined as:

$$p_t^{(i)}:=P(S'_t=i|I+\delta,S'_1,...,S'_{t-1}).$$

Following the definition of softmax function:

$$P(S'_t|I+\delta,S'_1,...,S'_{t-1})=\exp(z_t^{(S'_t)})/\sum_{i\in\mathcal{V}}\exp(z_t^{(i)}).$$

Intuitively, to maximize the targeted caption's probability, we can directly use its negative log probability (5) as a loss function. The inputs of the RNN are the first $N-1$ words of the targeted caption $(S_1,S_2,...,S_{N-1})$.

$$\begin{aligned}\text{loss}_{S,\text{log-prob}}(I+\delta)&=-\log P(S|I+\delta)\\&=-\sum_{t=2}^{N}\log P(S_t|I+\delta,S_1,...,S_{t-1}).\end{aligned} \tag{5}$$

Applying (5) to (2), the formulation of targeted caption method given a targeted caption S is:

$$\min_{w\in\mathbb{R}^n} c\cdot\text{loss}_{S,\text{log prob}}(\tanh(w+y)) \\ +\ \|\tanh(w+y)-\tanh(y)\|_2^2.$$

Alternatively, using the definition of the softmax function,

$$\begin{aligned}\log P(S'|I+\delta)&=\sum_{t=2}^{N}[z_t^{(S'_t)}-\log(\sum_{i\in\mathcal{V}}\exp(z_t^{(i)}))]\\&=\sum_{t=2}^{N}z_t^{(S'_t)}-\text{constant},\end{aligned} \tag{6}$$

(3) can be simplified as

$$\log P(S|I+\delta)\propto\sum_{t=2}^{N}z_t^{(S_t)}=\max_{S'\in\Omega}\sum_{t=2}^{N}z_t^{(S'_t)}.$$

Instead of making each $z_t^{(S_t)}$ as large as possible, it is sufficient to require the target word S_t to attain the largest (top-1) logit (or probability) among all the words in the vocabulary at position t. In other words, we aim to minimize the difference between the maximum logit except S_t, denoted by $\max_{k\in\mathcal{V},k\neq S_t}\{z_t^{(k)}\}$, and the logit of S_t, denoted by $z_t^{(S_t)}$. We also propose a ramp function on top of this difference as the final loss function:

$$\text{loss}_{S,\text{logits}}(I+\delta)=\sum_{t=2}^{N-1}\max\{-\epsilon,\max_{k\neq S_t}\{z_t^{(k)}\}-z_t^{(S_t)}\}, \tag{7}$$

where $\epsilon>0$ is a confidence level accounting for the gap between $\max_{k\neq S_t}\{z_t^{(k)}\}$ and $z_t^{(S_t)}$. When $z_t^{(S_t)}>\max_{k\neq S_t}\{z_t^{(k)}\}+\epsilon$, the corresponding term in the summation will be kept at $-\epsilon$ and does not contribute to the gradient of the loss function, encouraging the optimizer to focus on minimizing other terms where $z_t^{(S_t)}$ is not large enough.

Applying the loss (7) to (1), the final formulation of targeted caption method given a targeted

caption S is

$$\min_{w\in\mathbb{R}^n} c\cdot\sum_{t=2}^{N-1}\max\{-\epsilon,\max_{k\neq S_t}\{z_t^{(k)}\}-z_t^{(S_t)}\} + \|\tanh(w+y)-\tanh(y)\|_2^2.$$

We note that (Carlini and Wagner, 2017) has reported that in CNN-based image classification, using logits in the attack loss function can produce better adversarial examples than using probabilities, especially when the target network deploys some gradient masking schemes such as defensive distillation (Papernot et al., 2016b). Therefore, we provide both logit-based and probability-based attack loss functions for neural image captioning.

3.3 Targeted Keyword Method

In addition to generating an exact targeted caption by perturbing the input image, we offer an intermediate option that aims at generating captions with specific keywords, denoted by $\mathcal{K} := \{K_1,\cdots,K_M\}\subset\mathcal{V}$. Intuitively, finding an adversarial image generating a caption with specific keywords might be easier than generating an exact caption, as we allow more degree of freedom in caption generation. However, as we need to ensure a valid and meaningful inferred caption, finding an adversarial example with specific keywords in its caption is difficult in an optimization perspective. Our target keyword method can be used to investigate the generalization capability of a neural captioning system given only a few keywords.

In our method, we do not require a target keyword K_j, $j\in[M]$ to appear at a particular position. Instead, we want a loss function that allows K_j to become the top-1 prediction (plus a confidence margin ϵ) at any position. Therefore, we propose to use the minimum of the hinge-like loss terms over all $t\in[N]$ as an indication of K_j appearing at any position as the top-1 prediction, leading to the following loss function:

$$\text{loss}_{K,\text{logits}}=\sum_{j=1}^{M}\min_{t\in[N]}\{\max\{-\epsilon,\max_{k\neq K_j}\{z_t^{(k)}\}-z_t^{(K_j)}\}\}. \tag{8}$$

We note that the loss functions in (4) and (5) require an input S'_{t-1} to predict z_t for each $t\in\{2,\ldots,N\}$. For the targeted caption method, we use the targeted caption S as the input of RNN. In contrast, for the targeted keyword method we no longer know the exact targeted sentence, but only require the presence of specified keywords in the final caption. To bridge the gap, we use the originally inferred caption $S^0=(S_1^0,\cdots,S_N^0)$ from the benign image as the initial input to RNN. Specifically, after minimizing (8) for T iterations, we run inference on $I+\delta$ and set the RNN's input S^1 as its current top-1 prediction, and continue this process. With this iterative optimization process, the desired keywords are expected to gradually appear in top-1 prediction.

Another challenge arises in targeted keyword method is the problem of "keyword collision". When the number of keywords $M\geq 2$, more than one keywords may have large values of $\max_{k\neq K_j}\{z_t^{(k)}\}-z_t^{(K_j)}$ at a same position t. For example, if `dog` and `cat` are top-2 predictions for the second word in a caption, the caption can either start with "A dog ..." or "A cat ...". In this case, despite the loss (8) being very small, a caption with both `dog` and `cat` can hardly be generated, since only one word is allowed to appear at the same position. To alleviate this problem, we define a gate function $g_{t,j}(x)$ which masks off all the other keywords when a keyword becomes top-1 at position t:

$$g_{t,j}(x)=\begin{cases}A, & \text{if } \arg\max_{i\in\mathcal{V}} z_t^{(i)}\in\mathcal{K}\setminus\{K_j\}\\ x, & \text{otherwise,}\end{cases}$$

where A is a predefined value that is significantly larger than common logits values. Then (8) becomes:

$$\sum_{j=1}^{M}\min_{t\in[N]}\{g_{t,j}(\max\{-\epsilon,\max_{k\neq K_j}\{z_t^{(k)}\}-z_t^{(K_j)}\})\}. \tag{9}$$

The log-prob loss for targeted keyword method is discussed in the Supplementary Material.

4 Experiments

4.1 Experimental Setup and Algorithms

We performed extensive experiments to test the effectiveness of our Show-and-Fool algorithm and study the robustness of image captioning systems under different problem settings. In our experiments[1], we use the pre-trained TensorFlow implementation[2] of Show-and-Tell (Vinyals et al., 2015)

[1] Our source code is available at: https://github.com/huanzhang12/ImageCaptioningAttack

[2] https://github.com/tensorflow/models/tree/master/research/im2txt

with Inception-v3 as the CNN for visual feature extraction. Our testbed is Microsoft COCO (Lin et al., 2014) (MSCOCO) data set. Although some more recent neural image captioning systems can achieve better performance than Show-and-Tell, they share a similar framework that uses CNN for feature extraction and RNN for caption generation, and Show-and-Tell is the vanilla version of this CNN+RNN architecture. Indeed, we find that the adversarial examples on Show-and-Tell are transferable to other image captioning models such as Show-Attend-and-Tell (Xu et al., 2015) and NeuralTalk2[3], suggesting that the attention mechanism and the choice of CNN and RNN architectures do not significantly affect the robustness. We also note that since Show-and-Fool is the first work on crafting adversarial examples for neural image captioning, to the best of our knowledge, there is no other method for comparison.

We use ADAM to minimize our loss functions and set the learning rate to 0.005. The number of iterations is set to $1,000$. All the experiments are performed on a single Nvidia GTX 1080 Ti GPU. For targeted caption and targeted keyword methods, we perform a binary search for 5 times to find the best c: initially $c = 1$, and c will be increased by 10 times until a successful adversarial example is found. Then, we choose a new c to be the average of the largest c where an adversarial example can be found and the smallest c where an adversarial example cannot be found. We fix $\epsilon = 1$ except for transferability experiments. For each experiment, we randomly select 1,000 images from the MSCOCO validation set. We use BLEU-1 (Papineni et al., 2002), BLEU-2, BLEU-3, BLEU-4, ROUGE (Lin, 2004) and METEOR (Lavie and Agarwal, 2005) scores to evaluate the correlations between the inferred captions and the targeted captions. These scores are widely used in NLP community and are adopted by image captioning systems for quality assessment. Throughout this section, we use the *logits loss* (7)(9). The results of using the *log-prob loss* (5) are similar and are reported in the supplementary material.

4.2 Targeted Caption Results

Unlike the image classification task where all possible labels are predefined, the space of possible captions in a captioning system is almost infinite. However, the captioning system is only able to output relevant captions learned from the training set. For instance, the captioning model cannot generate a passive-voice sentence if the model was never trained on such sentences. Therefore, we need to ensure that the targeted caption lies in the space where the captioning system can possibly generate. To address this issue, we use the generated caption of a randomly selected image (other than the image under investigation) from MSCOCO validation set as the targeted caption S. The use of a generated caption as the targeted caption excludes the effect of out-of-domain captioning, and ensures that the target caption is within the output space of the captioning network.

Here we use the logits loss (7) plus a ℓ_2 distortion term (as in (2)) as our objective function. A successful adversarial example is found if the inferred caption after adding the adversarial perturbation δ is *exactly the same* as the targeted caption. In our setting, 1,000 ADAM iterations take about 38 seconds for one image. The overall success rate and average distortion of adversarial perturbation δ are shown in Table 1. Among all the tested images, our method attains 95.8% attack success

Table 1: Summary of targeted caption method (Section 3.2) and targeted keyword method (Section 3.3) using logits loss. The ℓ_2 distortion of adversarial noise $\|\delta\|_2$ is averaged over successful adversarial examples. For comparison, we also include CNN based attack methods (Section 4.5).

Experiments	Success Rate	Avg. $\|\delta\|_2$
targeted caption	95.8%	2.213
1-keyword	97.1%	1.589
2-keyword	97.5%	2.363
3-keyword	96.0%	2.626
C&W on CNN	22.4%	2.870
I-FGSM on CNN	34.5%	15.596

Table 2: Statistics of the 4.2% failed adversarial examples using the targeted caption method and logits loss (7). All correlation scores are computed using the top-5 inferred captions of an adversarial image and the targeted caption (higher score means better targeted attack performance).

c	1	10	10^2	10^3	10^4
ℓ_2 Distortion	1.726	3.400	7.690	16.03	23.31
BLEU-1	.567	.725	.679	.701	.723
BLEU-2	.420	.614	.559	.585	.616
BLEU-3	.320	.509	.445	.484	.514
BLEU-4	.252	.415	.361	.402	.417
ROUGE	.502	.664	.629	.638	.672
METEOR	.258	.407	.375	.403	.399

[3] https://github.com/karpathy/neuraltalk2

rate. Moreover, our adversarial examples have small ℓ_2 distortions and are visually identical to the original images, as displayed in Figure 1. We also examine the failed adversarial examples and summarize their statistics in Table 2. We find that their generated captions, albeit not entirely identical to the targeted caption, are in fact highly correlated to the desired one. Overall, the high success rate and low ℓ_2 distortion of adversarial examples clearly show that Show-and-Tell is not robust to targeted adversarial perturbations.

4.3 Targeted Keyword Results

In this task, we use (9) as our loss function, and choose the number of keywords $M = \{1, 2, 3\}$. We run an inference step on $I + \delta$ every $T = 5$ iterations, and use the top-1 caption as the input of RNN/LSTMs. Similar to Section 4.2, for each image the targeted keywords are selected from the caption generated by a randomly selected validation set image. To exclude common words like "a", "the", "and", we look up each word in the targeted sentence and only select nouns, verbs, adjectives or adverbs. We say an adversarial image is successful when its caption contains *all* specified keywords. The overall success rate and average distortion are shown in Table 1. When compared to the targeted caption method, targeted keyword method achieves an even higher success rate (at least 96% for 3-keyword case and at least 97% for 1-keyword and 2-keyword cases). Figure 2 shows an adversarial example crafted from our targeted keyword method with three keywords - "dog", "cat" and "frisbee". Using Show-and-Fool, the top-1 caption of a cake image becomes "A dog and a cat are playing with a frisbee" while the adversarial image remains visually indistinguishable to the original one. When $M = 2$ and 3, even if we cannot find an adversarial image yielding all specified keywords, we might end up with a caption that contains some of the keywords (partial success). For example, when $M = 3$, Table 3 shows the number of keywords appeared in the captions (M') for those *failed* examples (not all 3 targeted keywords are found). These results clearly show that the 4% failed examples are still partially successful: the generated captions contain about 1.5 targeted keywords on average.

4.4 Transferability of Adversarial Examples

It has been shown that in image classification tasks, adversarial examples found for one machine learning model may also be effective against another model, even if the two models have different architectures (Papernot et al., 2016a; Liu et al., 2017c). However, unlike image classification where correct labels are made explicit, two different image captioning systems may generate quite different, yet semantically similar, captions for the same benign image. In image captioning, we say an adversarial example is *transferable* when the adversarial image found on model A with a target sentence S_A can generate a similar (rather than exact) sentence S_B on model B.

In our setting, model A is Show-and-Tell, and we choose Show-Attend-and-Tell (Xu et al., 2015) as model B. The major differences between Show-and-Tell and Show-Attend-and-Tell are the addition of attention units in LSTM network for caption generation, and the use of last convolutional layer (rather than the last fully-connected layer) feature maps for feature extraction. We use Inception-v3 as the CNN architecture for both models and train them on the MSCOCO 2014 data set. However, their CNN parameters are different due to the fine-tuning process.

Original Top-3 inferred captions:
1. A cake that is sitting on a table.
2. A cake that is sitting on a plate.
3. A cake that is sitting on a table

Adversarial Keywords:
"cat", "dog" and "frisbee"

Adversarial Top-3 captions:
(targeted keyword method)
1. A dog and a cat are playing with a frisbee.
2. A dog laying on a rug with a frisbee in its mouth.
3. A dog and a cat are playing with a toy.

Figure 2: An adversarial example ($\|\delta\|_2 = 1.284$) of an cake image crafted by the Show-and-Fool targeted keyword method with three keywords - "dog", "cat" and "frisbee".

Table 3: Percentage of partial success with different c in the 4.0% failed images that do not contain all the 3 targeted keywords.

c	Avg. $\|\delta\|_2$	$M' \geq 1$	$M' = 2$	Avg. M'
1	2.49	72.4%	34.5%	1.07
10	5.40	82.7%	37.9%	1.21
10^2	12.95	93.1%	58.6%	1.52
10^3	24.77	96.5%	51.7%	1.48
10^4	29.37	100.0%	58.6%	1.59

Table 4: Transferability of adversarial examples from Show-and-Tell to Show-Attend-and-Tell, using different ϵ and c. **ori** indicates the scores between the generated captions of the *original* images and the transferred adversarial images on Show-Attend-and-Tell. **tgt** indicates the scores between the *targeted* captions on Show-and-Tell and the generated captions of transferred adversarial images on Show-Attend-and-Tell. A smaller **ori** or a larger **tgt** value indicates better transferability. **mis** measures the differences between captions generated by the two models given the same benign image (*model mismatch*). When $C = 1000, \epsilon = 10$, **tgt** is close to **mis**, indicating the discrepancy between adversarial captions on the two models is mostly bounded by model mismatch, and the adversarial perturbation is highly transferable.

	$\epsilon = 1$						$\epsilon = 5$						$\epsilon = 10$						
	C=10		C=100		C=1000		C=10		C=100		C=1000		C=10		C=100		C=1000		
	ori	**tgt**	**ori**	**tgt**	**ori**	**tgt**	**ori**	**tgt**	**ori**	**tgt**	**ori**	**tgt**	**ori**	**tgt**	**ori**	**tgt**	**ori**	**tgt**	**mis**
BLEU-1	.474	.395	.384	.462	.347	.484	.441	.429	.368	.488	**.337**	.527	.431	.421	.360	.485	.339	**.534**	.649
BLEU-2	.337	.236	.230	.331	.186	.342	.300	.271	.212	.343	.175	.389	.287	.266	.204	.342	**.174**	**.398**	.521
BLEU-3	.256	.154	.151	.224	.114	.254	.220	.184	.135	.254	.103	.299	.210	.185	.131	.254	**.102**	**.307**	.424
BLEU-4	.203	.109	.107	.172	.077	.198	.170	.134	.093	.197	.068	.240	.162	.138	.094	.197	**.066**	**.245**	.352
ROUGE	.463	.371	.374	.438	.336	.465	.429	.402	.359	.464	.329	.502	.421	.398	.351	.463	**.328**	**.507**	.604
METEOR	.201	.138	.139	.180	.118	.201	.177	.157	.131	.199	.110	.228	.172	.157	.127	.202	**.110**	**.232**	.300
$\|\delta\|_2$	3.268		4.299		4.474		7.756		10.487		10.952		15.757		21.696		21.778		

Original Top-1 inferred caption:
Show-and-Tell: A bathroom with a sink and a mirror
Show-Attend-and-Tell: A bathroom with a sink and a mirror.

Adversarial Top-1 caption:
Show-and-Tell (targeted caption method): A man riding a wave on top of a surfboard.
Show-Attend-and-Tell (transferred example): A man on a surfboard in the air.

Figure 3: A highly transferable adversarial example ($\|\delta\|_2 = 15.226$) crafted by Show-and-Tell targeted caption method, transfers to Show-Attend-and-Tell, yielding similar adversarial captions.

To investigate the transferability of adversarial examples in image captioning, we first use the targeted caption method to find adversarial examples for 1,000 images in model A with different c and ϵ, and then transfer successful adversarial examples (which generate the exact target captions on model A) to model B. The generated captions by model B are recorded for transferability analysis. The transferability of adversarial examples depends on two factors: the intrinsic difference between two models even when the same benign image is used as the input, i.e., *model mismatch*, and the transferability of adversarial perturbations.

To measure the mismatch between Show-and-Tell and Show-Attend-and-Tell, we generate captions of the same set of 1,000 original images from both models, and report their mutual BLEU, ROUGE and METEOR scores in Table 4 under the **mis** column. To evaluate the effectiveness of transferred adversarial examples, we measure the scores for two set of captions: (i) the captions of original images and the captions of transferred adversarial images, both generated by Show-Attend-and-Tell (shown under column **ori** in Table 4); and (ii) the targeted captions for generating adversarial examples on Show-and-Tell, and the captions of the transferred adversarial image on Show-Attend-and-Tell (shown under column **tgt** in Table 4). Small values of **ori** suggest that the adversarial images on Show-Attend-and-Tell generate significantly different captions from original images' captions. Large values of **tgt** suggest that the adversarial images on Show-Attend-and-Tell generate similar adversarial captions as on the Show-and-Tell model. We find that increasing c or ϵ helps to enhance transferability at the cost of larger (but still acceptable) distortion. When $C = 1,000$ and $\epsilon = 10$, Show-and-Fool achieves the best transferability results: **tgt** is close to **mis**, indicating that the discrepancy between adversarial captions on the two models is mostly bounded by the intrinsic model mismatch rather than the transferability of adversarial perturbations, and implying that the adversarial perturbations are easily transferable. In addition, the adversarial examples generated by our method can also fool NeuralTalk2. When $c = 10^4, \epsilon = 10$, the average ℓ_2 distortion, BLEU-4 and METEOR scores between the original and transferred adversarial captions are 38.01, 0.440 and 0.473, respectively. The high transferability of adversarial examples crafted by Show-

and-Fool also indicates the problem of common robustness leakage between different neural image captioning models.

4.5 Attacking Image Captioning v.s. Attacking Image Classification

In this section we show that attacking image captioning models is inherently more challenging than attacking image classification models. In the classification task, a targeted attack usually becomes harder when the number of labels increases, since an attack method needs to change the classification prediction to a specific label over all the possible labels. In the targeted attack on image captioning, if we treat each caption as a label, we need to change the original label to a specific one over an almost infinite number of possible labels, corresponding to a nearly zero volume in the search space. This constraint forces us to develop non-trivial methods that are significantly different from the ones designed for attacking image classification models.

To verify that the two tasks are inherently different, we conducted additional experiments on attacking *only* the CNN module using two state-of-the-art image classification attacks on ImageNet dataset. Our experiment setup is as follows. Each selected ImageNet image has a label corresponding to a WordNet synset ID. We randomly selected 800 images from ImageNet dataset such that their synsets have at least one word in common with Show-and-Tell's vocabulary, while ensuring the Inception-v3 CNN (Show-and-Tell's CNN) classify them correctly. Then, we perform Iterative Fast Gradient Sign Method (I-FGSM) (Kurakin et al., 2017) and Carlini and Wagner's (C&W) attack (Carlini and Wagner, 2017) on these images. The attack target labels are randomly chosen and their synsets also have at least one word in common with Show-and-Tell's vocabulary. Both I-FGSM and C&W achieve 100% targeted attack success rate on the Inception-v3 CNN. These adversarial examples were further employed to attack Show-and-Tell model. An attack is considered successful if *any* word in the targeted label's synset or its hypernyms up to 5 levels is presented in the resulting caption. For example, for the chain of hypernyms 'broccoli'⇒'cruciferous vegetable'⇒'vegetable, veggie, veg'⇒'produce, green goods, green groceries, garden truck'⇒'food, solid food', we include 'broccoli','cruciferous','vegetable','veggie' and all other following words. Note that this criterion of success is much weaker than the criterion we use in the targeted caption method, since a caption with the targeted image's hypernyms does not necessarily leads to similar meaning of the targeted image's captions. To achieve higher attack success rates, we allow relatively larger distortions and set $\epsilon_\infty = 0.3$ (maximum ℓ_∞ distortion) in I-FGSM and $\kappa = 10$, $C = 100$ in C&W. However, as shown in Table 1, the attack success rates are only 34.5% for I-FGSM and 22.4% for C&W, respectively, which are much lower than the success rates of our methods despite larger distortions. This result further confirms that performing targeted attacks on neural image captioning requires a careful design (as proposed in this paper), and attacking image captioning systems is not a trivial extension to attacking image classifiers.

5 Conclusion

In this paper, we proposed a novel algorithm, **Show-and-Fool**, for crafting adversarial examples and providing robustness evaluation of neural image captioning. Our extensive experiments show that the proposed targeted caption and keyword methods yield high attack success rates while the adversarial perturbations are still imperceptible to human eyes. We further demonstrate that Show-and-Fool can generate highly transferable adversarial examples. The high-quality and transferable adversarial examples in neural image captioning crafted by Show-and-Fool highlight the inconsistency in visual language grounding between humans and machines, suggesting a possible weakness of current machine vision and perception machinery. We also show that attacking neural image captioning systems are inherently different from attacking CNN-based image classifiers.

Our method stands out from the well-studied adversarial learning on image classifiers and CNN models. To the best of our knowledge, this is the very first work on crafting adversarial examples for neural image captioning systems. Indeed, our Show-and-Fool algorithm[1] can be easily extended to other applications with RNN or CNN+RNN architectures. We believe this paper provides potential means to evaluate and possibly improve the robustness (for example, by adversarial training or data augmentation) of a wide range of visual language grounding and other NLP models.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433.

Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pages 39–57.

Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. 2018. EAD: elastic-net attacks to deep neural networks via adversarial examples. *AAAI*.

Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. 2017. ZOO: zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, AISec@CCS*, pages 15–26.

Xinlei Chen and C. Lawrence Zitnick. 2015. Mind's eye: A recurrent visual representation for image caption generation. In *CVPR*, pages 2422–2431.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *CVPR*.

Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. 2015. Language models for image captioning: The quirks and what works. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 100–105.

Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Trevor Darrell, and Kate Saenko. 2015a. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, pages 2625–2634.

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015b. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, pages 2625–2634.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. 2015. From captions to visual concepts and back. In *CVPR*, pages 1473–1482.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 457–468.

Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. 2017. Semantic compositional networks for visual captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5630–5639.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. *ICLR; arXiv preprint arXiv:1412.6572*.

Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1233–1239.

Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. 2015. Guiding the long-short term memory model for image caption generation. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, pages 2407–2415. IEEE.

Andrej Karpathy and Fei-Fei Li. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2017. Adversarial machine learning at scale. *ICLR; arXiv preprint arXiv:1611.01236*.

Alon Lavie and Abhaya Agarwal. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation*, pages 65–72.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8. Barcelona, Spain.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Chenxi Liu, Junhua Mao, Fei Sha, and Alan L Yuille. 2017a. Attention correctness in neural image captioning. In *AAAI*, pages 4176–4182.

Feng Liu, Tao Xiang, Timothy M Hospedales, Wankou Yang, and Changyin Sun. 2017b. Semantic regularisation for recurrent image annotation. *CVPR*.

Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. 2017c. Delving into transferable adversarial examples and black-box attacks. *ICLR; arXiv preprint arXiv:1611.02770*.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems (NIPS)*, pages 289–297.

Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. 2016. Generating images from captions with attention. *ICLR; arXiv preprint arXiv:1511.02793*.

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. 2015. Deep captioning with multimodal recurrent neural networks (m-rnn). *ICLR; arXiv preprint arXiv:1412.6632*.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. 2017. Universal adversarial perturbations. In *CVPR*.

Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 462–472.

Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. Generating natural questions about an image. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1802–1813.

Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. 2016a. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*.

Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. 2016b. Distillation as a defense to adversarial perturbations against deep neural networks. In *Security and Privacy (SP), 2016 IEEE Symposium on*, pages 582–597. IEEE.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *annual meeting on association for computational linguistics (ACL)*, pages 311–318.

Ramakanth Pasunuru and Mohit Bansal. 2017. Multi-task video captioning with video and entailment generation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1273–1283.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, pages 1060–1069.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurelie Herbelot, Moin Nabi, Enver Sangineto, Raffaella Bernardi, et al. 2017. Foil it! Find one mismatch between image and language caption. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. *ICLR;arXiv preprint arXiv:1312.6199*.

Kenneth Tran, Xiaodong He, Lei Zhang, and Jian Sun. 2016. Rich image captioning in the wild. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2016 IEEE Conference on*, pages 434–441. IEEE.

Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond J. Mooney, and Kate Saenko. 2015. Translating videos to natural language using deep recurrent neural networks. In *NAACL-HLT*, pages 1494–1504.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164.

Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton van den Hengel. 2016. What value do explicit high level concepts have in vision to language problems? In *CVPR*, pages 203–212.

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*, pages 2048–2057.

Zhilin Yang, Ye Yuan, Yuexin Wu, William W. Cohen, and Ruslan Salakhutdinov. 2016. Review networks for caption generation. In *NIPS*, pages 2361–2369.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *CVPR*, pages 4651–4659.

Linchao Zhu, Zhongwen Xu, Yi Yang, and Alexander G Hauptmann. 2017. Uncovering the temporal context for video question answering. *International Journal of Computer Vision*, 124(3):409–421.