# **Supplementary Material**

# 6 Related Work on Adversarial Attacks to CNN-based Image Classifiers

Despite the remarkable progress, CNNs have been shown to be vulnerable to adversarial examples (Szegedy et al., 2014; Goodfellow et al., 2015; Carlini and Wagner, 2017). In image classification, an adversarial example is an image that is visually indistinguishable to the original image but can cause a CNN model to misclassify. With different objectives, adversarial attacks can be divided into two categories, i.e., untargeted attack and targeted attack. In the literature, a successful untargeted attack refers to finding an adversarial example that is close to the original example but yields different class prediction. For targeted attack, a target class is specified and the adversarial example is considered successful when the predicted class matches the target class. Surprisingly, adversarial examples can also be crafted even when the parameters of target CNN model are unknown to an attacker (Liu et al., 2017c; Chen et al., 2017). In addition, adversarial examples crafted from one image classification model can be made transferable to other models (Liu et al., 2017c; Papernot et al., 2016a), and there exists a universal adversarial perturbation that can lead to misclassification of natural images with high probability (Moosavi-Dezfooli et al., 2017).

Without loss of generality, there are two factors contributing to crafting adversarial examples in image classification: (i) a distortion metric between the original and adversarial examples that regularizes visual similarity. Popular choices are the  $L_{\infty}$ ,  $L_2$  and  $L_1$  distortions (Kurakin et al., 2017; Carlini and Wagner, 2017; Chen et al., 2018); and (ii) an attack loss function accounting for the success of adversarial examples. For finding adversarial examples in neural image captioning, while the distortion metric can be identical, the attack loss function used in image classification is invalid, since the number of possible captions easily outnumbers the number of image classes, and captions with similar meaning should not be considered as different classes. One of our major contributions is to design novel attacking loss functions to handle the CNN+RNN architectures in neural image captioning tasks.

## 7 More Adversarial Examples with Logits Loss

Figure 4 shows another successful example with targeted caption method. Figures 5, 6 and 7 show three adversarial examples generated by the proposed 3-keyword method. The adversarial examples generated by our methods have small  $L_2$  distortions and are visually indistinguishable from the original images. One advantage of using logits losses is that it helps to bypass defensive distillation by overcoming the gradient vanishing problem. To see this, the partial derivative of the softmax function

$$p^{(j)} = \exp(z^{(j)}) / \sum_{i \in \mathcal{V}} \exp(z^{(i)}),$$

is given by

$$\frac{\partial p^{(j)}}{\partial z^{(j)}} = p^{(j)} (1 - p^{(j)}), \tag{10}$$

which vanishes as  $p^{(j)} \rightarrow 0$  or  $p^{(j)} \rightarrow 1$ . The defensive distillation method [30] uses a large distillation temperature in the training process and removes it in the inference process. This makes the inference probability  $p^{(j)}$  close to 0 or 1, thus leads to a vanished gradient problem. However, by using the proposed logits loss (7), before the word at position t in target sentence S reaches top-1 probability, we have

$$\frac{\partial}{\partial z_t^{(S_t)}} \mathrm{loss}_{S,\mathrm{logits}}(I+\delta) = -1.$$
(11)

It is evident that the gradient (with regard to  $z_t^{(S_t)}$ ) becomes a constant now, since it equals to -1when  $z_t^{(S_t)} < \max_{k \neq S_t} \{z_t^{(k)}\} + \epsilon$ , and 0 otherwise.

## 8 Targeted Caption Results with Log Probability Loss

In this experiment, we use the log probability loss (5) plus a  $L_2$  distortion term (as in (2)) as our objective function. Similar to the previous experiments, a successful adversarial example is found if the inferred caption after adding the adversarial perturbation  $\delta$  exactly matches the targeted caption. The overall success rate and average distortion of adversarial perturbation  $\delta$  are shown in Table 5. Among all the tested images, our log-prob loss attains 95.4% success rate, which is about the



#### **Original Top-3 inferred captions:**

- 1. A large elephant standing next to a wooden fence.
- 2. A large elephant standing next to a stone wall.
- 3. An elephant standing next to a wooden fence.

## Adversarial Top-3 captions:

- 1. A black and white photo of a group of people
  - A black and white photo of a group of people.
- A black and white photo of a black and white horse

Figure 4: Adversarial example ( $\|\delta\|_2 = 2.977$ ) of an elephant image crafted by the Show-and-Fool targeted caption method with the target caption "A black and white photo of a group of people".

2.



Figure 5: Adversarial example ( $\|\delta\|_2 = 2.979$ ) of an clock image crafted by the Show-and-Fool targeted keyword method with three keywords: "meat", "white" and "topped".

same as using logits loss. Besides, similar to using logits loss, the adversarial examples generated by using log-prob loss also yield small  $L_2$  distortions. In Table 6, we summarize the statistics of the failed adversarial examples. It shows that their generated captions, though not entirely identical to the targeted caption, are also highly relevant to the target captions.

In our experiments, log probability loss exhibits a similar performance as the logits loss, as our target model is undefended and the gradient vanishing problem of softmax is not significant. However, when evaluating the robustness of a general image captioning model, it is recommended to use the logits loss as it does not suffer from potentially vanished gradients and can reveal the intrinsic robustness of the model.





#### **Original Top-3 inferred captions:**

- A close up of a giraffe with trees in the background
- 2. A close up of a giraffe near a fence
- 3. A close up of a giraffe near a tree

Adversarial Keywords: "soccer", "group" and "playing"

#### Adversarial Top-3 captions: (targeted keyword method) 1. A group of young men

- playing a game of soccer.
- 2. A group of people playing a game of <u>soccer</u>.
- A group of people playing a game of baseball.

Figure 6: Adversarial example ( $\|\delta\|_2 = 1.188$ ) of a giraffe image crafted by the Show-and-Fool targeted keyword method with three keywords: "soccer", "group" and "playing".



Figure 7: Adversarial example ( $\|\delta\|_2 = 1.178$ ) of a bus image crafted by the Show-and-Fool targeted keyword method with three keywords: "tub", "bathroom" and "sink".

## 9 Targeted Keyword Results with Log Probability Loss

Similar to the logits loss, the log-prob loss does not require a particular position for the target keywords  $K_j, j \in [M]$ . Instead, it encourages  $K_j$  to become the top-1 prediction at its most probable position:

$$loss_{K, log-prob} = -\sum_{j=1}^{M} log(\max_{t \in [N]} \{p_t^{(i)}\}).$$
(12)

To tackle the "keyword collision" problem, we also employ a gate function  $g'_{t,j}$  to avoid the keywords appearing at the positions where the most



Original Top-1 inferred caption: Show-and-Tell: A woman riding a bike with a dog on it. Show-Attend-and-Tell: A woman sitting on a bicycle with a dog.



Adversarial Top-1 caption: Show-and-Tell (targeted caption method): A pizza sitting on top of a white plate. Show-Attend-and-Tell (transferred example): A white and white slice of pizza on a table.

Figure 8: A highly transferable adversarial example of a biking image ( $\|\delta\|_2 = 12.391$ ) crafted from Show-and-Tell using the targeted caption method and then transfers to Show-Attend-and-Tell, yielding similar adversarial captions.



Original Top-1 inferred caption: Show-and-Tell: A man riding a snowboard down a snow covered slope. Show-Attend-and-Tell: A man is snowboarding down a snow covered slope.



Adversarial Top-1 caption: Show-and-Tell (targeted caption method): A black cat laying on top of a suitcase. Show-Attend-and-Tell (transferred example): A black and white cat sitting on a bed with a bag of luggage.

Figure 9: A highly transferable adversarial example of a snowboarding image ( $\|\delta\|_2 = 14.320$ ) crafted from Show-and-Tell using the targeted caption method and then transfers to Show-Attend-and-Tell, yielding similar adversarial captions.

probable word is already a keyword:

$$g'_{t,j}(x) = \begin{cases} 0, \text{ if } \arg \max_{i \in \mathcal{V}} p_t^{(i)} \in \mathcal{K} \setminus \{K_j\} \\ x, \text{ otherwise} \end{cases}$$

The loss function (12) then becomes:

$$loss_{K',log-prob} = -\sum_{j=1}^{M} log(\max_{t \in [N]} \{g'_{t,j}(p_t^{(i)})\}).$$
(13)

In our methods, the initial input is the originally inferred caption  $S^0$  from the benign image, and after minimizing (13) for T iterations, we run inference on  $I + \delta$  and set the RNN's input  $S^1$  as its current top-1 prediction, and repeat this procedure until all the targeted keywords are found or



Original Top-1 inferred caption: Show-and-Tell: A desk with a laptop and a monitor on it Show-Attend-and-Tell: A desk with a computer monitor and a monitor.



Adversarial Top-1 caption: Show-and-Tell (targeted caption method): A cat laying in a sink in a bathroom. Show-Attend-and-Tell (transferred example): A cat is sitting in a bathroom sink.

Figure 10: A highly transferable adversarial example of a desk image ( $\|\delta\|_2 = 12.810$ ) crafted from Show-and-Tell using the targeted caption method and then transfers to Show-Attend-and-Tell, yield-ing similar adversarial captions.

Table 5: Summary of targeted caption method and targeted keyword method using *log-prob* loss. The  $L_2$  distortion  $\|\delta\|_2$  is averaged over successful adversarial examples.

Experiments	Success Rate	Avg. $\ \delta\ _2$
targeted caption	95.4%	1.858
1-keyword	99.2%	1.311
2-keyword	96.9%	2.023
3-keyword	95.7%	2.120

the maximum number of iterations is met. With this iterative optimization process, the probabilities of the desired keywords gradually increase, and finally become the top-1 predictions.

The overall success rate and average distortion are shown in Table 5. Table 7 summarizes the number of keywords (M') appeared in the captions for those *failed* examples when M = 3, i.e., the examples that not all the 3 targeted keywords are found. They account only 4.3% of all the tested images. Table 7 clearly shows that when c is properly chosen, more than 90% of the failed examples contain at least 1 targeted keyword, and more than 60% of the failed examples contain 2 targeted keywords. This result verifies that even the failed examples are reasonably good attacks.

## 10 Transferability of Adversarial Examples with Log Probability Loss

Similar to the experiments in Section 4.4, to assess the transferability of adversarial examples, we first use the targeted caption method with log-prob loss to find adversarial examples for 1,000 images Table 6: Statistics of the 4.6% failed adversarial examples using the targeted caption method and *log-prob* loss (5). All correlation scores are computed using the top-5 inferred captions of an adversarial image and the targeted caption (a higher score indicates a better targeted attack performance).

c	1	10	$10^{2}$	$10^{3}$	$10^{4}$
$L_2$ Distortion	1.503	2.637	5.085	11.15	19.69
BLEU-1	.650	.792	.775	.802	.800
BLEU-2	.521	.690	.671	.711	.701
BLEU-3	.416	.595	.564	.622	.611
BLEU-4	.354	.515	.485	.542	.531
ROUGE	.616	.764	.746	.776	.772
METEOR	.362	.493	.469	.511	.498

Table 7: Percentage of partial success using *log*prob loss with different c in the 4.3% failed images that do not contain all the 3 targeted keywords.

c	Avg. $\ \delta\ _2$	$M' \ge 1$	M'=2	Avg. $M'$
1	2.22	69.7%	27.3%	0.97
10	5.03	87.9%	57.6%	1.45
$10^{2}$	10.98	93.9%	63.6%	1.58
$10^{3}$	18.52	93.9%	57.6%	1.52
$10^{4}$	26.04	90.9%	60.6%	1.52

in Show-and-Tell model (model A) with different c. We then transfer successful adversarial examples, i.e., the examples that generate the exact target captions on model A, to Show-Attend-and-Tell model (model B). The generated captions by model B are recorded for transferability analysis. The results for transferability using log-prob loss is summarized in Table 8. The definitions of **tgt**, **ori** and **mis** are the same as those in Table 4. Comparing with Table 4 ( $C = 1000, \epsilon = 10$ ), the log probability loss shows inferior **ori** and **tgt** values, indicating that the additional parameter  $\epsilon$  in the logits loss helps improve transferability.

## 11 Attention on Original and Transferred Adversarial Images

Figures 11, 12 and 13 show the original and adversarial images' attentions over time. In the original images, the Show-Attend-and-Tell model's attentions align well with human perception. However, the transferred adversarial images obtained on Show-and-Tell model yield significantly misaligned attentions.

Table 8: Transferability of adversarial examples from Show-and-Tell to Show-Attend-and-Tell, using different c. Unlike Table 4, the adversarial examples in this table are found using the *logprob* loss and there is no parameter  $\epsilon$ . Similarly, a smaller **ori** or a larger **tgt** value indicates better transferability.

	C=10		C=100		C=1000		
	ori	tgt	ori	tgt	ori	tgt	mis
BLEU-1	.540	.391	.442	.435	.374	.500	.657
BLEU-2	.415	.224	.297	.280	.217	.357	.529
BLEU-3	.335	.143	.218	.193	.137	.268	.430
BLEU-4	.280	.101	.170	.142	.095	.207	.357
ROUGE	.525	.364	.430	.411	.362	.474	.609
METEOR	.240	.132	.179	.162	.135	.209	.303
$\ \delta\ _2$	2.4	33	4.6	512	10	.88	



(b) Adversarial Image of Figure 8

Figure 11: Original and transferred adversarial image's attention over time on Figure 8. The highlighted area shows the attention change as the model generates each word.



(b) Adversarial Image of Figure 9

Figure 12: Original and transferred adversarial image's attention over time on Figure 9. The highlighted area shows the attention change as the model generates each word.

Figure 13: Original and transferred adversarial image's attention over time on Figure 10. The highlighted area shows the attention change as the model generates each word.