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

Hongge Chen[1*], Huan Zhang[2*], 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

* Hongge Chen and Huan Zhang contribute equally to this work.

Introduction

- We propose **Show-and-Fool***, a novel algorithm for crafting adversarial examples in **neural image captioning**. We propose **targeted caption method** and **targeted keyword method**.

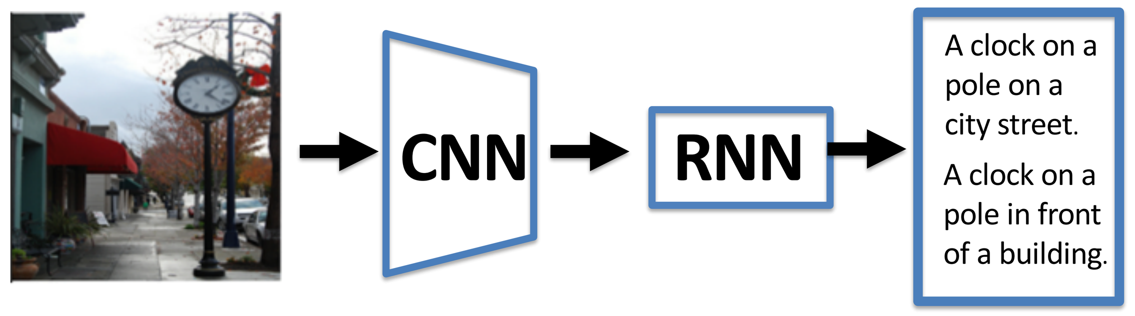


Figure 1: A typical neural image captioning system with a CNN+RNN structure.

Original Top-3 inferred captions:
1. A red stop sign sitting on the side of a road.
2. A stop sign on the corner of a street.
3. A red stop sign sitting on the side of a street.

Adversarial Top-3 captions:
1. A brown teddy bear laying on top of a bed.
2. A brown teddy bear sitting on top of a bed.
3. A large brown teddy bear laying on top of a bed.

Original Top-3 inferred captions:
1. A man holding a tennis racquet on a tennis court.
2. A man holding a tennis racquet on top of a tennis court.
3. A man holding a tennis racquet on a court.

Adversarial Top-3 captions:
1. A woman brushing her teeth in a bathroom.
2. A woman brushing her teeth in the bathroom.
3. A woman brushing her teeth in front of a bathroom mirror.

Figure 2: Adversarial examples crafted by Show-and-Fool using the **targeted caption method**

Methodology

The problem of finding an adversarial noise δ for a given image I can be cast as the following optimization problem:

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

This constraint minimization is converted to a unconstraint minimization using a tanh transform.

Let $z_t = [z_t^{(1)}, \ldots, z_t^{(|\mathcal{V}|)}]$ be the vector of logits at position t.

- In **Targeted Caption Method**, the inputs of the RNN are the first $N-1$ words of the targeted caption and the loss is given as:

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

where larger ε can produce high confident adversarial example for transferability.

- In **Targeted Keyword Method**, for a set of keywords $\mathcal{K} = \{K_j\}$, the loss is:

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

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

We use the originally inferred caption from the benign image as the initial input to RNN. After several iterations, set the RNN's input as its current top-1 prediction, and continue this process.

Experiments

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.

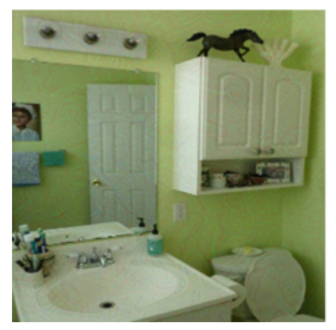
Original Top-3 inferred captions:
1. A bus is parked on the side of the street.
2. A bus is parked on the side of the road.
3. A bus is parked on the side of a street.

Adversarial Keywords: "tub", "bathroom" and "sink"

Adversarial Top-3 captions: (targeted keyword method)
1. A bathroom with a sink, toilet and tub.
2. A bathroom with a sink, toilet, and bathtub.
3. A bathroom with a tub, sink, and toilet.

Figure 3: Adversarial examples crafted by Show-and-Fool using the **targeted keyword method**

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 4: A highly transferable adversarial example crafted by Show-and-Tell targeted caption method, transfers to Show-Attend-and-Tell

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 1: Summary of targeted caption method and targeted keyword method using logits loss. The distortion is averaged over successful adversarial examples. For comparison, we also include CNN based attack methods.

	$\epsilon=1$						$\epsilon=5$					
	C=10		C=100		C=1000		C=10		C=100		C=1000	
	ori	tgt	ori	tgt	ori	tgt	ori	tgt	ori	tgt	ori	tgt
BLEU-1	.474	.395	.384	.462	.347	.484	.441	.429	.368	.488	**.337**	.527
BLEU-2	.337	.236	.230	.331	.186	.342	.300	.271	.212	.343	.175	.389
BLEU-3	.256	.154	.151	.224	.114	.254	.220	.184	.135	.254	.103	.299
BLEU-4	.203	.109	.107	.172	.077	.198	.170	.134	.093	.197	.068	.240
ROUGE	.463	.371	.374	.438	.336	.465	.429	.402	.359	.464	.329	.502
METEOR	.201	.138	.139	.180	.118	.201	.177	.157	.131	.199	.110	.228
$\|\delta\|_2$	3.268		4.299		4.474		7.756		10.487		10.952	

Table 2: 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.

Conclusion

We proposed a novel algorithm for crafting adversarial examples and providing robustness evaluation of neural image captioning. Show-and-Fool algorithm can be easily extended to other applications with RNN or CNN+RNN architectures.

*** Our code is available at: https://github.com/IBM/Image-Captioning-Attack**