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.

Methodology

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

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

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

Let $z_t = \{z_t^1, ..., z_t^n\}$ 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:

$$\text{loss}_{\text{target}}(I + \delta) = \sum_{t=1}^{N-1} \max(0, \max_{K_t \in \mathcal{K}} (\hat{L}_t(z_t) - L_t(z_t)))$$

where larger $\varepsilon$ can produce high-confident adversarial example for transferability.

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

$$\text{loss}_{\text{target}}(I + \delta) = \sum_{K_t \in \mathcal{K}} \min_{x \in [0, 1]} \{ g_t(x) \max_{L_t(z_t) \in \mathcal{K}} (\hat{L}_t(z_t) - L_t(z_t)) \} \quad \text{where } g_t(x) = \begin{cases} \max_{L_t(z_t) \in \mathcal{K}} (\hat{L}_t(z_t) - L_t(z_t)) & \text{if } \arg\max_{L_t(z_t) \in \mathcal{K}} (\hat{L}_t(z_t) - L_t(z_t)) - K_t \leq x \leq \arg\max_{L_t(z_t) \in \mathcal{K}} (\hat{L}_t(z_t) - L_t(z_t)) - K_t + 1 \\
\text{otherwise} \end{cases}$$

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-1 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 in parked on the side of a street.

Adversarial Keywords: “tuil”, “bathroom” and “sink”

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

Original Top-3 inferred captions:

1. A man holding a tennis racket on a tennis court.
2. A woman holding a tennis racket on top of a tennis court.
3. A man holding a tennis racket on a court.

Adversarial Top-3 captions:

1. A woman brushing her teeth in the bathroom.
2. A woman brushing her teeth in the bathroom.
3. A woman brushing her teeth in front of a mirror.

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.