Technische Universiteit
Eindhoven
University of Technology

Department of Mathematics and Computer Science
Architecture of Information Systems Research Group

Adversarial Noise Benchmarking On Image Caption

Bachelor Thesis

H.J.M. van Genuchten

Supervisors:
C. de Campos
Z.M. van Cauter

Intermediate Draft

Eindhoven, April 2022

Abstract

TODO Abstract

1 Introduction

The image caption generation task is at the cross-section between Computer Vision (CV) and Natural Language Processing (NLP). It requires the computer to understand a visual scene and describe it into a grammatically correct natural sentence. Practical use cases vary from automated describing of images to visually impaired people (Mazzoni, 2019) to context based image retrieval.

Show Attend and Tell (S.A.T.) proposed by Xu et al. is an end-to-end deep learning approach that tries to solve the image caption generation problem. It combines an attention mechanism with LSTM to generate sentences that describe the given image. An example output from S.A.T can be seen in figure 1 Achieving good BLEU scores on Flickr8K, Flickr30K(Hodosh, Young & Hockenmaier, n.d.) and COCO(Lin et al., 2015) datasets. Although the scores are not state-of-the-art(Stefanini et al., 2021) anymore. This model is chosen because it is small and thus can be run locally, and has publicly available implementations (Sgrvinod, n.d.).

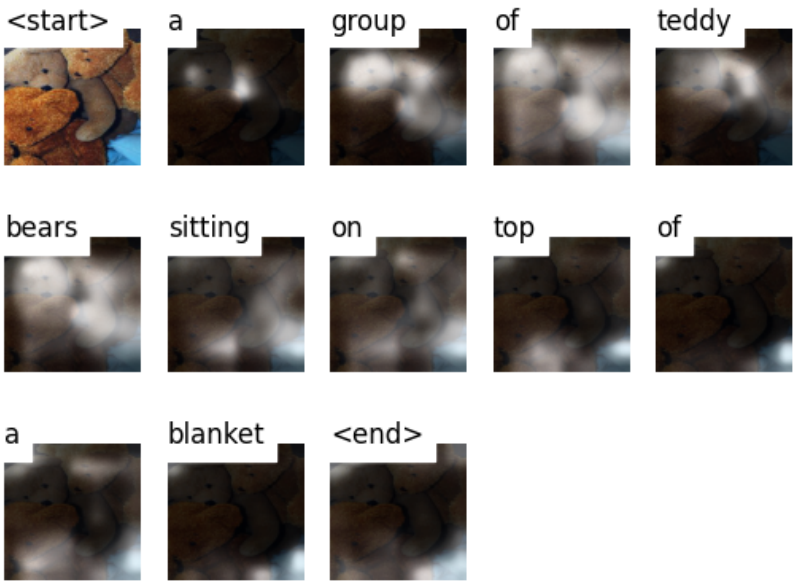


Figure 1: Prediction by Show Attend and Tell on a clean image.
Top left picture is the input image. The highlighted areas in white are the visualization of the attention per predicted word.

Machine learning models can be very susceptible to noise where small changes to the input can lead to radically different outcomes. As shown by Goodfellow, Shlens and Szegedy adding a specific (small) noise layer to an image can alter a correct prediction to a very confident wrong prediction. As can be seen in figure 2. Because the generation of the adversarial examples is not that computational expansive, they can be generated during training making the model more robust. It is also shown that these adversarial examples act as regularizes during training.

Reducing the change of overfitting. Kurakin, Goodfellow and Bengio expands on generating adversarial examples showing that one can also steer the model towards a specific classification.

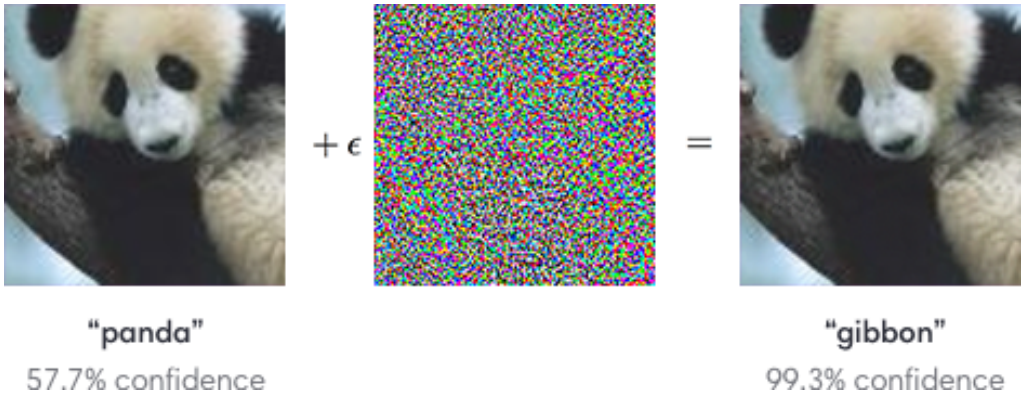


Figure 2: Adversarial noise example from (Goodfellow et al., 2015). Where $\epsilon = 0.07$.

Combining these previous findings, S.A.T. can be used to find adversarial examples for image captioning models. These adversarial images can then be used to either improve current datasets by providing hard samples, or in a more malicious way. The latter being especially true when one can specify the output sentence for which an adversarial sample should be created. Analyzing the successful and failed adversarial samples can also give a better insight in the strengths and weaknesses of S.A.T. Furthermore, it exposes bias present in the datasets.

1.1 Motivation and Research Questions

This research investigates the susceptibility of S.A.T. against adversarial samples that are visually close but generate completely different descriptions as output.

- Is S.A.T. susceptible to adversarial attacks using noise?
- Can the noise be crafted in such a way that it can steer the output.

1.2 Prior research

2 Methodology

2.1 Dataset and Model

2.2 Generating Adversarial Samples

Randomly sampling the noise field to find samples close to a certain image would be time-consuming and inefficient. Luckily generating adversarial input images can be done by using the Fast Method (EQ. 1) proposed by Goodfellow et al..

$$X^{adv} = clip(X + \epsilon * sign(\nabla_x J(X, y_{true})), 0, 1) \tag{1}$$

With X being the input image, ϵ a hyperparameter determining much the original image can be perpetrated and $J(X, y_{true})$ the loss function which to, in the adversarial case, maximize. Finally, the image is clipped ensuring the vector stays within the 0 to 1 input range. As can be seen in Figure 3 (and bigger size in appendix A), using this method images up to and including $\epsilon = 0.04$ are nearly indistinguishable and up to $\epsilon = 0.16$ very recognizable to humans.

ϵ=0.000 ϵ=0.005 ϵ=0.010 ϵ=0.020 ϵ=0.040 ϵ=0.080 ϵ=0.160 ϵ=0.320

Figure 3: Adversarial images for varying values of epsilon.

Steering Adversarial Samples

To steer the network towards a specific output we can adjust the equation 1 to minimize a loss function with a given target y.

$$X^{steer} = clip(X - \epsilon * sign(\nabla_x J(X, y_{target}))) \tag{2}$$

Where y_{target} can be determined to be anything.

Evaluation

To determine if the model is indeed susceptible the BLEU scores will be calculated for different values of ϵ. Furthermore, to also investigate if the semantic meaning of the sentence is significantly affected, the cosine similarity of the original and adversarial output will be calculated using universal sentence embedding proposed by Cer et al.. To see if the model can also be steered the BLEU score and cosine similarity are calculated with respect to the y_{target}.

3 Results

Although I currently don't have complete results. I do have some initial samples that worked. I am still in the process of calculating the BLEU score and cosine similarity over the whole datasets. Preliminary results images:

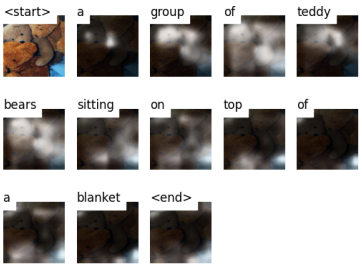


Figure 4: Prediction by Show Attend and Tell on a normal image

Figure 5: Prediction on an adversarial image with $\epsilon = 0.2$ (roughly 5% of original range)

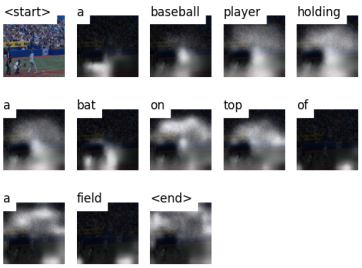


Figure 6: Prediction by Show Attend and Tell on a normal image

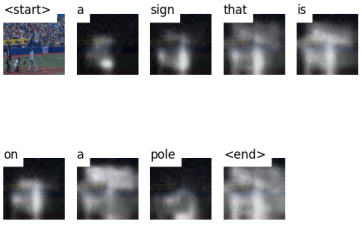


Figure 7: Prediction on an adversarial image with $\epsilon = 0.2$ (roughly 5% of original range)

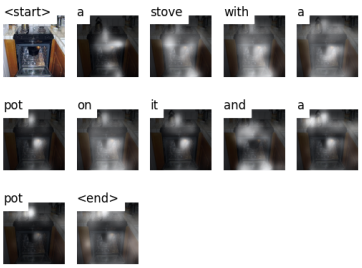


Figure 8: Prediction by Show Attend and Tell on a normal image

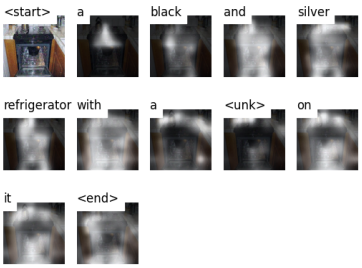


Figure 9: Prediction on an adversarial image with $\epsilon = 0.2$ (roughly 5% of original range)

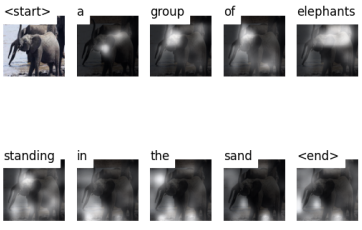


Figure 10: Prediction by Show Attend and Tell on a normal image

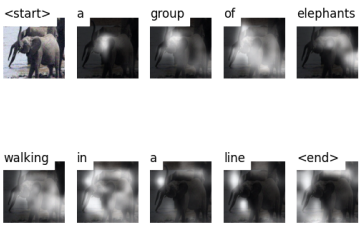


Figure 11: Prediction on an adversarial image with $\epsilon = 0.2$ (roughly 5% of original range)

4 Conclusions

Preliminary conclusion: It is possible

References

Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R. S., ... Kurzweil, R. (2018). Universal sentence encoder. *CoRR*, *abs/1803.11175*. Retrieved from http://arxiv.org/abs/1803.11175 3

Goodfellow, I. J., Shlens, J. & Szegedy, C. (2015). *Explaining and harnessing adversarial examples.* 1, 2

Hodosh, M., Young, P. & Hockenmaier, J. (n.d.). *Flickr8k dataset.* 1

Kurakin, A., Goodfellow, I. & Bengio, S. (2016). *Adversarial examples in the physical world.* arXiv. Retrieved from https://arxiv.org/abs/1607.02533 doi: 10.48550/ARXIV.1607.02533 2

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., ... Dollár, P. (2015). *Microsoft coco: Common objects in context.* 1

Mazzoni, D. (2019, Oct). *Using ai to give people who are blind the "full picture".* Google. Retrieved from https://blog.google/outreach-initiatives/accessibility/get-image-descriptions/ 1

Sgrvinod. (n.d.). *Sgrvinod/a-pytorch-tutorial-to-image-captioning: Show, attend, and tell: A pytorch tutorial to image captioning.* Retrieved from https://github.com/sgrvinod/a-PyTorch-Tutorial-to-Image-Captioning 1

Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G. & Cucchiara, R. (2021). From show to tell: A survey on image captioning. *CoRR*, *abs/2107.06912*. Retrieved from https://arxiv.org/abs/2107.06912 1

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., ... Bengio, Y. (2016). *Show, attend and tell: Neural image caption generation with visual attention.* 1

A Bigger adversarial images

Clean image
Prediction by S.A.T.: A group of people standing around a tennis court.

Adversarial Image with $\epsilon = 0.005$
Prediction by S.A.T.: A group of people sitting in a room with a bunch of different colored vases.

Clean image
Prediction by S.A.T.: A group of people standing around a tennis court.

Adversarial Image with $\epsilon = 0.010$
Prediction by S.A.T.: A group of vases sitting on top of a table.

Clean image
Prediction by S.A.T.: A group of people standing around a tennis court.

Adversarial Image with $\epsilon = 0.020$
Prediction by S.A.T.: A group of vases sitting on top of a table.

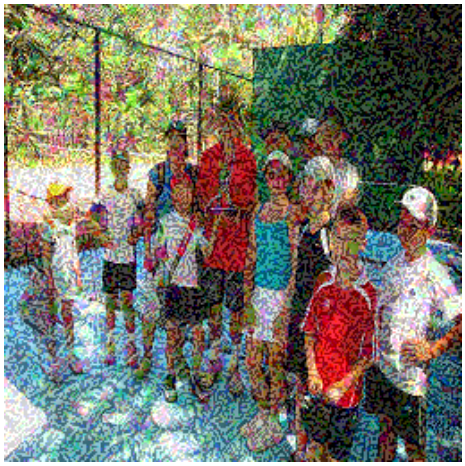
Clean image
Prediction by S.A.T.: A group of people standing around a tennis court.

Adversarial Image with $\epsilon = 0.040$
Prediction by S.A.T.: A large glass vase with a bunch of flowers on it.

Clean image
Prediction by S.A.T.: A group of people standing around a tennis court.

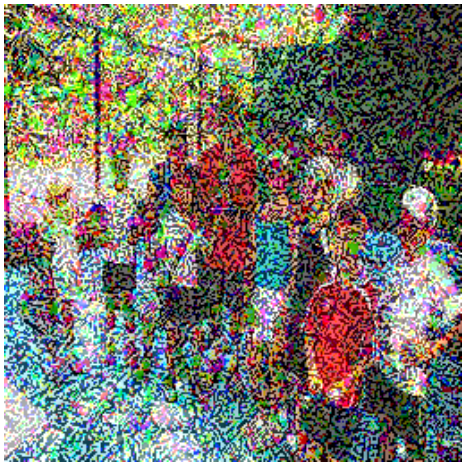
Adversarial Image with $\epsilon = 0.080$
Prediction by S.A.T.: A bathroom with a toilet and a sink.

Clean image
Prediction by S.A.T.: A group of people standing around a tennis court.



Adversarial Image with $\epsilon = 0.160$
Prediction by S.A.T.: A red wall with a red and white design.

Clean image
Prediction by S.A.T.: A group of people standing around a tennis court.



Adversarial Image with $\epsilon = 0.320$
Prediction by S.A.T.: A large red object with a red and white background.