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 (?, ?) to context based image retrieval.

Show Attend and Tell (S.A.T.) proposed by ? 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. Achieving good BLEU scores on Flickr8K, Flickr30K(?, ?) and COCO(?, ?) datasets. Although the scores are not state-of-the-art(?, ?) anymore. This model is chosen because it is small and thus can be run locally, and has publicly available implementations (?, ?).

Machine learning models often are susceptible to adversarial attacks, where the input is modified in such a way that the model is not able to produce a correct result or worse, a specific incorrect output. When these models are applied in a practical setting they need to be robust against these adversarial attacks. Moreover, if the model is robust against adversarial attacks it also is more robust against noise, and thus a more useful model.

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. To ensure that the images are visually close noise will be added to an original image. The research questions boil down to:

- 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.

2 Methodology

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 ?.

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

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.

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.

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

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 ?. 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:

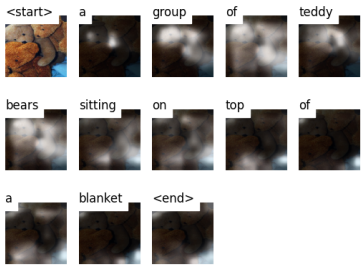

Figure 1: Prediction by Show Attend and Tell on a normal image

Figure 2: Prediction on an adversarial image with $\epsilon = 0.2$ (roughly 5% of original range)

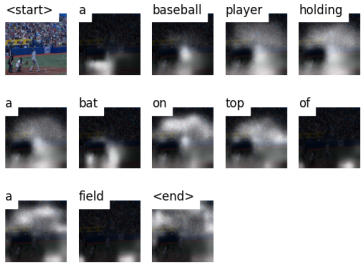

Figure 3: Prediction by Show Attend and Tell on a normal image

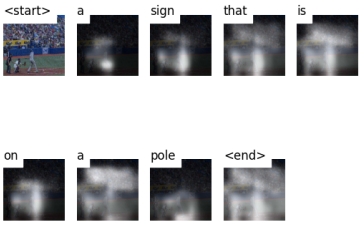

Figure 4: Prediction on an adversarial image with $\epsilon = 0.2$ (roughly 5% of original range)

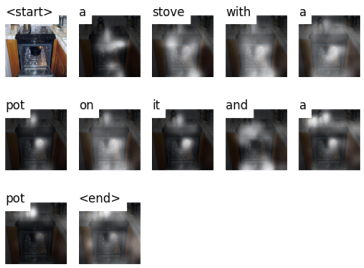

Figure 5: Prediction by Show Attend and Tell on a normal image

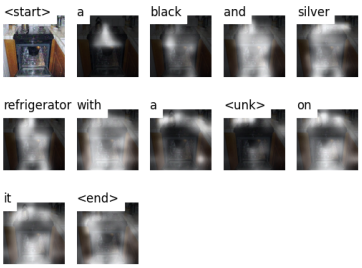

Figure 6: Prediction on an adversarial image with $\epsilon = 0.2$ (roughly 5% of original range)

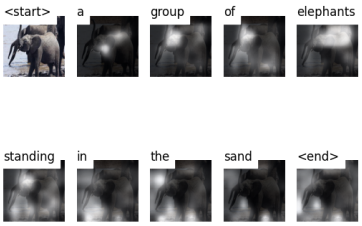

Figure 7: Prediction by Show Attend and Tell on a normal image

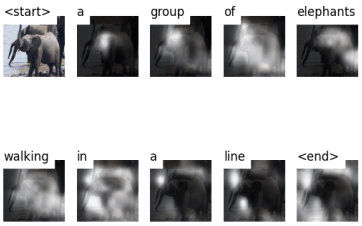

Figure 8: Prediction on an adversarial image with $\epsilon = 0.2$ (roughly 5% of original range)

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 1

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

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

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