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 K. 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.).

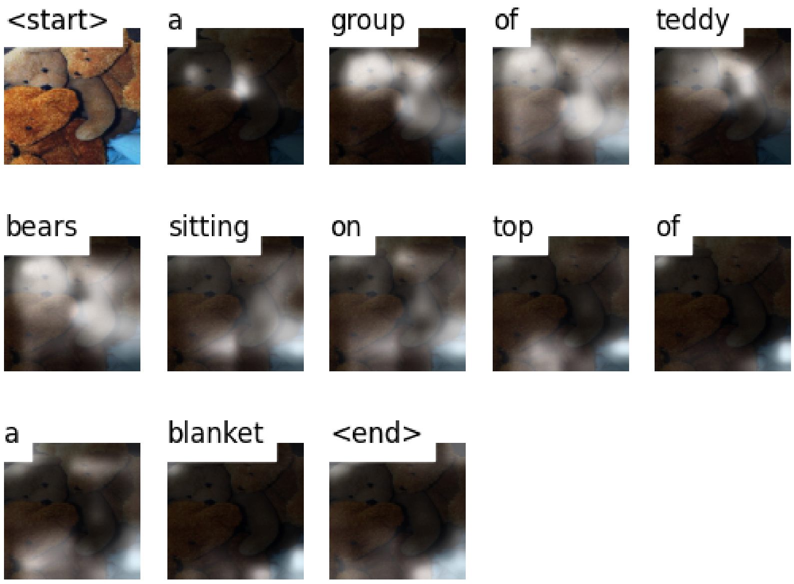


Figure 1: Prediction by Show Attend and Tell on a clean image.
Top left picture is the input image. The highlighted areas in white are the visualization of the attention per predicted word.

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.

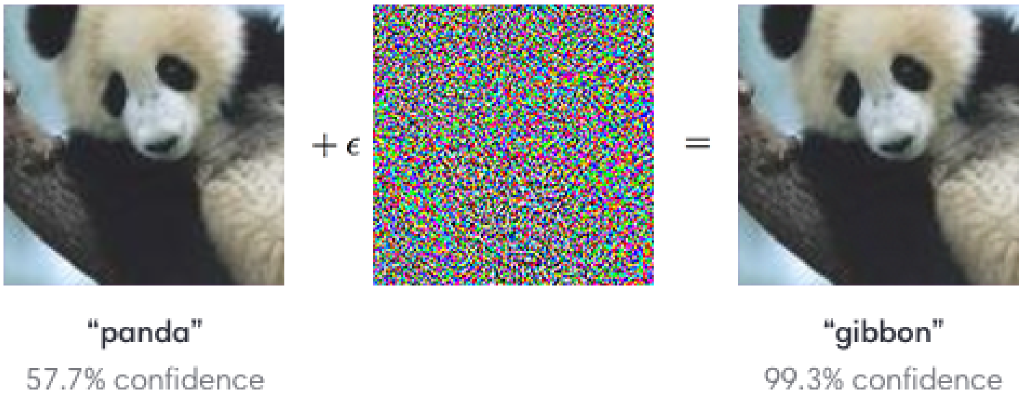


Figure 2: Adversarial noise example from (Goodfellow et al., 2015). Where $\epsilon = 0.07$.

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. An additional benefit to adversarial samples that can be cheaply generated is that they can be generated during training of models and act as regularizes. Improving the robustness of the model. However, this is only viable if generating the adversarial samples is not computational expensive. Another point that makes S.A.T. an interesting target, is that it uses explicit attention to generate captions. Although attention has been surpassed by the use of transformers, it is still interesting to see if it is a potential attack vector in an adversarial setting.

1.1 Motivation and Related Work

In the last few years research in the direction of generating adversarial samples for gradient based models has been published (Goodfellow et al., 2015; Kurakin et al., 2016a) as well as research showing the usefulness of such adversarial samples(Ilyas et al., 2019). The latter stating: "Adversarial vulnerability is a direct result of our models' sensitivity to well-generalizing features in the data." However, these generalizing features are only true for most samples, as models are optimized to do well in the average case. Inserting adversarial examples in training help regularize these non-robust features(Kurakin, Goodfellow & Bengio, 2016b). The Fast Gradient Sign Method was originally designed for classification task, however it (and variations) have been successfully adopted to other tasks such as object detection (Bose & Aarabi, 2018; Liu et al., 2020; Zhang & Wang, 2019), and most notably for this research on image captioning(Chen, Zhang, Chen, Yi & Hsieh, 2017). Chen et al.'s method Show-and-Fool successfully and robustly is able to attack Show-and-Tell(Vinyals, Toshev, Bengio & Erhan, 2014) (predecessor of Show Attend and Tell). Achieving a success rate of 95.8%, this does come at the cost of taking about 38 seconds to generate a single adversarial sample.

Adversarial Methods

Over the last few years variations of the Fast Gradient Sign Method by (Goodfellow et al., 2015) have been designed. The Iterative Fast Gradient Method by Kurakin et al. applies the Fast Gradient Sign Method multiple times. Which is further improved by using various optimization techniques such as momentum (J. Xu, 2020), and in the case of Show-and-Fool the well known Adam(Kingma & Ba, 2017) optimizer. Carlini and Wagner also directly include a distance metric in their adversarial optimization instead of clipping. Although very powerful techniques they are also computationally expensive and due to the computational limitations of this bachelor project, not feasible to apply extensively. Moreover, in the case of purely deceiving S.A.T. it is shown that classical Fast Gradient Sign Method already provides significant results.

1.2 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 the Fast Gradient Sign Method?
- Can S.A.T. be distracted

2 Methodology

2.1 Notation

- X, clean image in the range [0,1] retrieved from the dataset.
- X^{adv}, adversarial image generated from X by applying some perpetration to it. Clipped to be in the range [0,1].
- $Clip_{X,\epsilon}(A)$ is the element-wise clipping of A such that $X-\epsilon \leq A \leq X+\epsilon$ holds element-wise.

2.2 Dataset

As clean dataset the well known MSCOCO (Lin et al., 2015) dataset will be used. It contains 35 thousand images, of which 30 thousand are part of the train set, and 5 thousand of the testing set. Due to the computational cost of running the model, only the test set is used

2.3 Model

The model used, as already introduced in section 1, will be Show Attend and Tell. It is an interesting model as it uses attention, which can be visualized, to focus on most important places of the image. The question then arises if this attention is a defense mechanism, as it allows the model to ignore parts of the image, or if it opens up the model to new attack vectors as it can be distracted, which can lead to it ignoring important parts of the image.

2.4 Generating Adversarial Samples

Randomly sampling the noise field to find adversarial 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..

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

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. The sign in combination with the epsilon ensures $L_\infty(X - X^{adv}) \leq \epsilon$.

In practice applying this a single time is often not enough to attack S.A.T. therefore the iterative method will also be tried as proposed by Kurakin et al.. Which repeatedly applies the Fast Gradient Sign Method for N iterations

$$X_0^{adv}, X_{n+1}^{adv} = Clip_{X,\epsilon}(X_n^{adv} + \alpha * sign(\nabla_x J(X_n^{adv}, y_{true}))) \tag{2}$$

In which, α is a second hyperparameter equal to ϵ/N, unless stated otherwise.

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.

Distracting Adversarial Sample

Distraction is a powerful technique often used by adversaries in the real world. As S.A.T. employs attention to generate sentences, it is possible to try and distract it by creating an adversarial sample that makes the model hyperfocused on only part of the image. During training S.A.T. learns to divide the attention roughly equally over the whole image during the generation of a caption. It does this by including the loss shown in equation 3.

$$L_{attention} = \sum_{i}^{L}(1 - \sum_{t}^{C}\alpha_{ti}^{2}) \tag{3}$$

With C equal to the amount of words generated by S.A.T., L equal to the latent pixels, and α_{ti} the attention given to latent pixel i for generating word t.

Using categorical cross-entropy we can craft an adversarial example which focuses the attention of S.A.T. to a single latent pixel.

$$L_{distraction} = CrossEntropy(d, \alpha) \tag{4}$$

With d, $\alpha \in \mathbb{R}^{\mathrm{LxC}}$ and d be constructed to focus attention on a specific latent pixel.

2.5 Evaluation

To determine if the model is indeed susceptible to distraction the BLEU-4 score (Papineni, Roukos, Ward & Zhu, 2001) will be calculated, as it is a widely reported metric within the image captioning task. Because the BLEU score checks for direct word occurrences it does not give a complete view on the success of the adversarial attack, as the model can still give a correct description using synonyms. This would result in a low BLEU score, where in fact the model is still performing correctly. To combat this the cosine similarity of the original and adversarial output will be calculated using universal sentence embedding proposed by Cer et al.. It is a separately learned model that embeds an entire sentence. In the case of distraction, the average attention the model applies on the dataset is also analyzed.

3 Results

* Currently all plots are based on a subset of 120 images, my intention is to rerun them on the complete test set (5k images), which depending on the amount iterations this takes about 2 hours per epsilon at 100 iterations.

Adversarial Samples

Distracting Samples

To distract the model, adversarial samples are created using the iterative method (EQ. 2) and the distraction adversarial loss (EQ. 4). The amount of iterations was experimentally found to be good enough in most cases at 100, in which more would result in better distraction at the cost

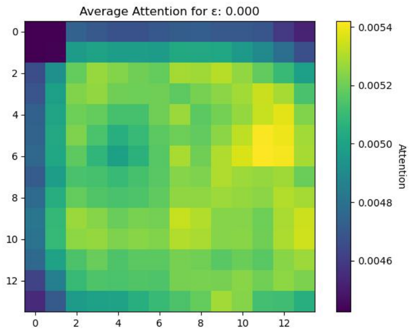


Figure 4: Average attention on clean images

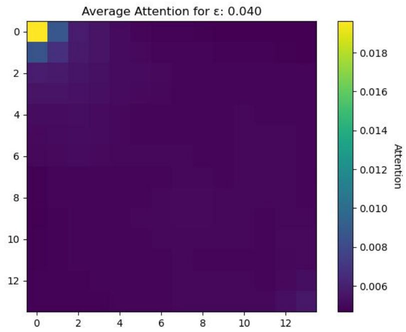


Figure 5: Average attention on adversarial images with ϵ=0.04 at 100 iterations

Figure 6: Clean Image (left), Adversarial Image $\epsilon = 0.04, N = 100$ (right)

of longer running times. The top left pixel was chosen to focus the attention on as the model focus least on it (4) (albeit slightly) during the clean images. With an epsilon of 0.04 satisfactory results are achieved. The attention of the model clearly focused on the top left on average as can be seen in figure 5. With the perturbation at most 0.04 the image is visually almost identical to the human eye 6.

The attention and sentences for the figure 6 are visualized in figure 7. The model still attends to other parts of the image, however they are not clearly a single object relating to the word that is generated. During the generations of the last few words the attention is focused almost solely on the top left part.

Figure 7: Attention on Clean Image (left) and Adversarial Image $\epsilon = 0.04, N = 100$ (right)

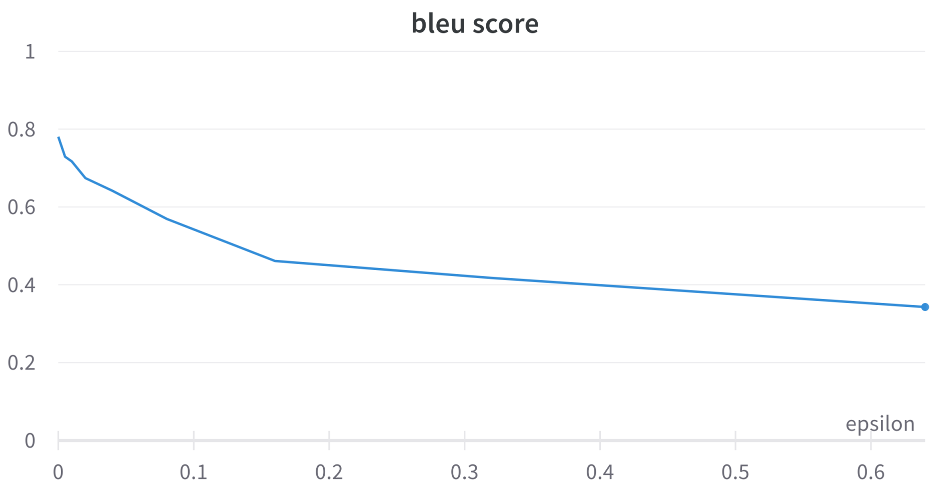


Figure 8: BLEU score during distraction over epsilon

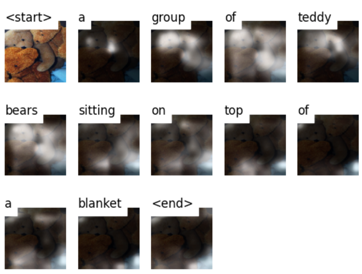


Figure 9: Prediction by Show Attend and Tell on a normal image

Figure 10: Prediction on an adversarial image with $\epsilon = 0.2$ (roughly 5% of original range)

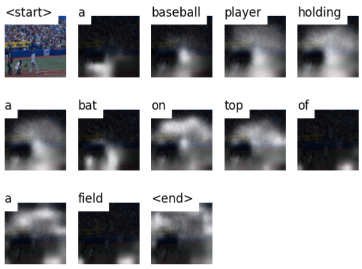


Figure 11: Prediction by Show Attend and Tell on a normal image

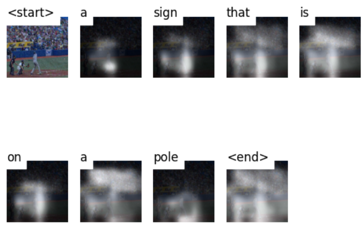


Figure 12: Prediction on an adversarial image with $\epsilon = 0.2$ (roughly 5% of original range)

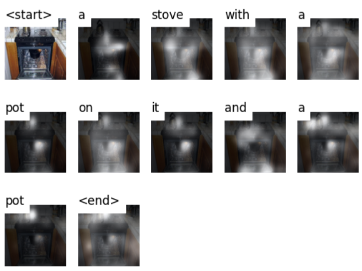


Figure 13: Prediction by Show Attend and Tell on a normal image

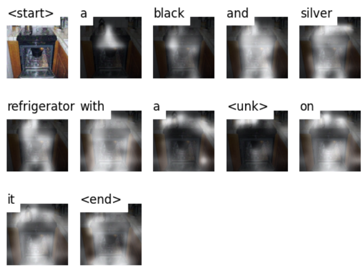


Figure 14: Prediction on an adversarial image with $\epsilon = 0.2$ (roughly 5% of original range)

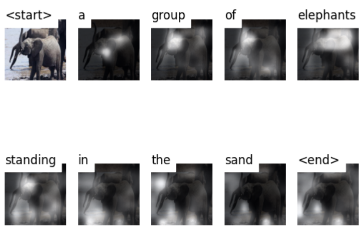


Figure 15: Prediction by Show Attend and Tell on a normal image

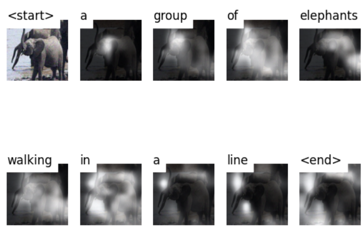


Figure 16: Prediction on an adversarial image with $\epsilon = 0.2$ (roughly 5% of original range)

4 Conclusions

Preliminary conclusion: It is possible

References

Bose, A. J. & Aarabi, P. (2018). *Adversarial attacks on face detectors using neural net based constrained optimization.* arXiv. Retrieved from `https://arxiv.org/abs/1805.12302` doi: 10.48550/ARXIV.1805.12302 2

Carlini, N. & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)* (p. 39-57). doi: 10.1109/SP.2017.49 2

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` 4

Chen, H., Zhang, H., Chen, P., Yi, J. & Hsieh, C. (2017). Show-and-fool: Crafting adversarial examples for neural image captioning. *CoRR*, *abs/1712.02051*. Retrieved from `http://arxiv.org/abs/1712.02051` 2

Goodfellow, I. J., Shlens, J. & Szegedy, C. (2015). *Explaining and harnessing adversarial examples.* 1, 2, 3

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

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B. & Madry, A. (2019). *Adversarial examples are not bugs, they are features.* arXiv. Retrieved from `https://arxiv.org/abs/1905.02175` doi: 10.48550/ARXIV.1905.02175 2

Kingma, D. P. & Ba, J. (2017). *Adam: A method for stochastic optimization.* 2

Kurakin, A., Goodfellow, I. & Bengio, S. (2016a). *Adversarial examples in the physical world.* arXiv. Retrieved from `https://arxiv.org/abs/1607.02533` doi: 10.48550/ARXIV.1607.02533 2, 3

Kurakin, A., Goodfellow, I. & Bengio, S. (2016b). *Adversarial machine learning at scale.* arXiv. Retrieved from `https://arxiv.org/abs/1611.01236` doi: 10.48550/ARXIV.1611.01236 2

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., ... Dollár, P. (2015). *Microsoft coco: Common objects in context.* 1, 3

Liu, Z., Peng, W., Zhou, J., Wu, Z., Zhang, J. & Zhang, Y. (2020). Mi-fgsm on faster r-cnn object detector. In *2020 the 4th international conference on video and image processing* (p. 27–32). New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/3447450.3447455` doi: 10.1145/3447450.3447455 2

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

Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. (2001). Bleu. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*. doi: 10.3115/1073083.1073135 4

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

Vinyals, O., Toshev, A., Bengio, S. & Erhan, D. (2014). *Show and tell: A neural image caption generator.* arXiv. Retrieved from `https://arxiv.org/abs/1411.4555` doi: 10.48550/ARXIV.1411.4555 2

Xu, J. (2020). Generate adversarial examples by nesterov-momentum iterative fast gradient sign method. In *2020 ieee 11th international conference on software engineering and service science (icsess)* (p. 244-249). doi: 10.1109/ICSESS49938.2020.9237700 2

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

Zhang, H. & Wang, J. (2019). Towards adversarially robust object detection. *CoRR*, *abs/1907.10310*. Retrieved from `http://arxiv.org/abs/1907.10310` 2

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.

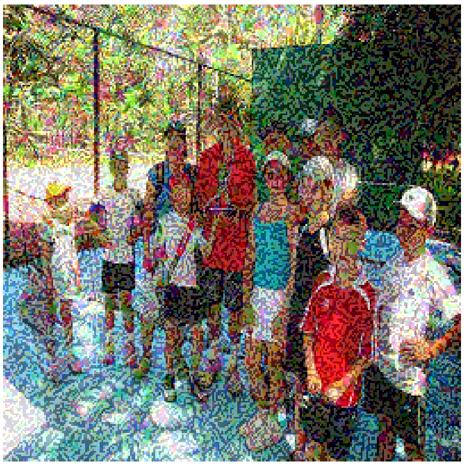
Clean image
Prediction by S.A.T.: A group of people standing around a tennis court.

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.

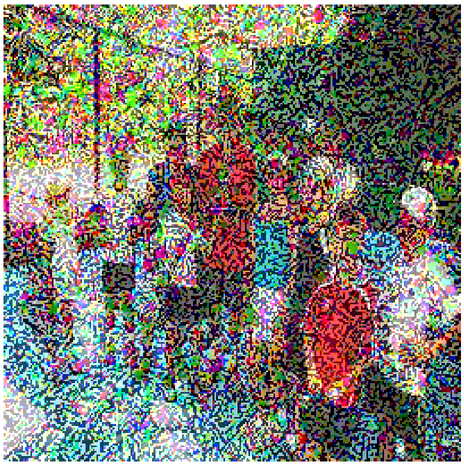
Adversarial Image with $\epsilon = 0.080$
Prediction by S.A.T.: A bathroom with a toilet and a sink.

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.