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. It can also be seen as a translation task, translating an image into natural language. Practical use cases vary from automated describing of images to visually impaired people (Mazzoni, 2019) to context based image retrieval. For the model to be successful in these tasks it should be accurate and robust. Show-Attend-and-Tell (S.A.T.) (K. Xu et al., 2016) is an end-to-end deep-learning approach to solve the image captioning task. An example prediction can be seen in figure 1a

For these models to be useful in real world examples they must be robust against small noise on the input image. However, the problem with machine learning is that models can be very susceptible to noise. As 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. When the generation of the adversarial examples is not that computational expensive, they can be generated and used during training making the model more robust. It is shown that these adversarial examples then act as regularizers during training. Reducing the chance of overfitting. Kurakin, Goodfellow and Bengio expands on generating adversarial examples showing that one can also steer the model towards a specific classification, however this comes at an increased computational cost. An example of an adversarial sample on image captioning can be seen in figure 1b.

(a) A group of teddy bears sitting on top of a blanket.

(b) A close up of a person on a suitcase.

Figure 1: Example predictions by Show Attend and Tell on a clean image (left) and an adversarial image (right).

Attention mechanisms, such as introduced by Bahdanau, Cho and Bengio, have been shown to improve various machine learning tasks, one of which is image captioning. It allows the model to focus on different parts of the image at a time to ensure the whole scene is described. The aforementioned S.A.T. also uses this attention mechanism. Furthermore, it improves the explainability

of the model, as for each word it can be shown which parts of the image the model "looked" at. However, if the attention is focused on the wrong parts of the image, the model is blind to possible important parts of the model. Hence, this attention mechanism can also be a possible new attack vector. This paper investigates the susceptibility of adversarial attacks specifically targeted at the attention layer, in an attempt to distract the network making it blind to parts of the image.

2 Motivation

When using machine learning models in real world use cases, it is important to ensure those models are robust. As if that is not the case they can be unreliable, wrong or in the worst case attacked by adversaries. Finding and understanding the weaknesses therefore is important. Furthermore, understanding the weaknesses can also help us in finding better architectures that are less susceptible to these kinds of attacks.

Adversarial samples for machine learning models can be generated using the Fast Gradient Sign Method (FSGM)(Goodfellow et al., 2015). Originally proposed for image classification, it finds a small noise field that can be added to the image to generate an adversarial image. This adversarial image is than often incorrectly labeled with a high confidence. An example of FSGM can be seen in figure 2. These adversarial sample prove to be useful during training, as they can act as regularizers, and improve the robustness of the model. Kurakin, Goodfellow and Bengio show that these adversarial samples are also transferable to different models, even if they are trained on other datasets or have different architectures.

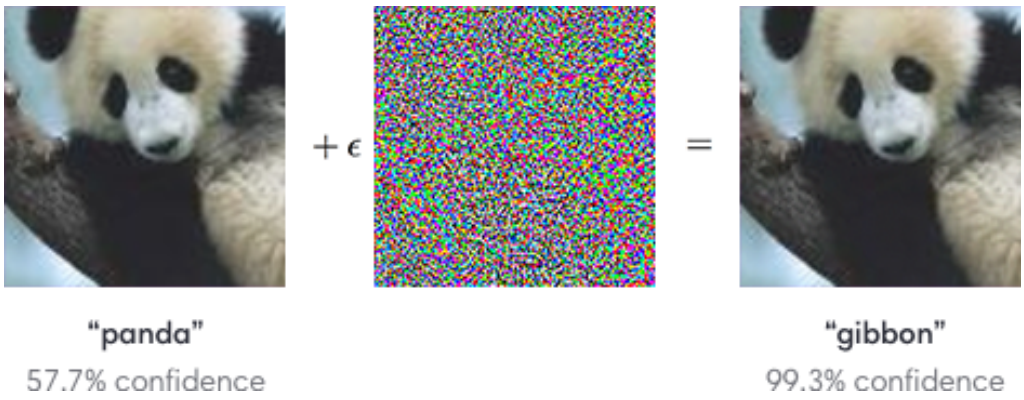


Figure 2: Adversarial noise example from (Goodfellow et al., 2015). Where $\epsilon = 0.07$.

There has already been some research in adversarial examples targeted at image captioning (Aafaq, Akhtar, Liu, Shah & Mian, 2021; Chen, Zhang, Chen, Yi & Hsieh, 2017). All of which attack the output of the model, often with the goal of generating a specific output sentence. Chen et al. shows that Show-and-Tell[1](Vinyals, Toshev, Bengio & Erhan, 2014) is susceptible to adversarial samples. The question then arises if the attention added in S.A.T. makes it harder to generate adversarial samples, and if it opens up a new attack vector.

2.1 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. A special focus is placed on what the attention in S.A.T. does and if it is an attack vector. When the attention is not focusing on the important parts of the image for generating the caption, the model is blind to those parts therefore not being able to describe those parts. It is therefore interesting to investigate if the attention can be used against S.A.T. Concretely this paper will try to answer the following questions:.

[1]Show-and-Tell is the predecessor of S.A.T. without attention mechanism

- Is S.A.T. susceptible to adversarial attacks using the Fast Gradient Sign Method?
- Can the attention of S.A.T. be abused by adversarial samples?

3 Related work

Image Captioning

Various techniques have been used to try and solve the image captioning task. Early methods mainly used hand-designed techniques based on template matching, which made them rigid in the sentences they could generate. One of the first end-to-end deep-learning approach was Show-and-Tell (Vinyals et al., 2014). It uses a CNN to extract most import features from an image, which then are decoded using an LSTM (Hochreiter & Schmidhuber, 1997) to a sentence, which describes the image. Show Attend and Tell (S.A.T.) proposed by K. Xu et al. is an extension to Show-and-Tell, which adds an attention mechanism before the LSTM decoding. This attention allows the model to focus on specific parts of the image when generating a word. An added benefit is that this attention can be visualized, giving insight in what the model looks at to generate a specific word in the output sentence.

Evaluating image captioning tasks is often done using bilingual evaluation understudy (BLEU) (Papineni, Roukos, Ward & Zhu, 2001). The primary reason is that it is used is that it is currently the most reported metric in image captioning. It is a form of word n-gram precision between the predicted and human generated reference sentences. It correlates highly with human ratings of captions (Vinyals et al., 2014). An obvious drawback to this is that a synonym of a word can result in a lower score, even though the sentence still is closely related, or by inserting common words such as 'a', 'the', and 'person' the model can achieve a higher score.

S.A.T. achieves a BLEU score of 0.79[1] out of 1.0 on COCO(Lin et al., 2015) datasets, the human annotations reach a score of 0.66[2]. Although the score is 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.).

Adversarial Methods

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) to create more robust datasets. 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 not robust, as models are optimized to do well in the average case. Inserting adversarial examples in training help regularize these non-robust features(Kurakin et al., 2016b).

One of the most influential methods in this field has been proposed by Goodfellow et al.. The Fast Gradient Sign Method (FSGM) (ab)uses the differentiability of machine leaning models to find an adversarial example. It is a single step gradient based approach to optimize the input image such that it maximizes a certain loss value. Various variations on FSGM have been proposed, such as the Iterative Fast Gradient Method (Kurakin et al., 2016a). Which applies multiple small steps of FSGM. It is further improved by adding various optimization techniques such as momentum (J. Xu, 2020).

Although FSGM was originally designed for classification task, 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 et al., 2017). Chen et al.'s method Show-and-Fool successfully and consistently is able to attack Show-and-Tell(Vinyals et al., 2014). They do this by using Adam(Kingma & Ba, 2017) to optimize the input image for 1000 steps targeting specific keywords. In generating captions that contain those

[1] Calculated using NLTK(Bird & Klein, 2009) bleu score implementation
[2] Calculated by comparing the captions with each other.

specific keywords they achieve a success rate of 95.8%, this does come at the cost of taking about 38 seconds to generate a single adversarial sample.

4 Methodology

4.1 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 limitations, only the test set is used.

4.2 Model

The model used, as already introduced in section 3, 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. It was trained on MSCOCO (Lin et al., 2015). S.A.T. uses a CNN as feature extractor to generate high dimensional latent pixels. These latent pixels are then fed to an attention layer, which in combination with an LSTM produces word tokens. It continues until a stop-token has been generated, or if it has generated 50 words, whichever comes first. In practice, it generally does not create sentences that are longer then 20 words. The attention that is used for each word can be visualized (Figure 3) by mapping the attention on the latent pixels back to the original location on the image. The implementation used, is a publicly available reproduction of S.A.T. in PyTorch (Paszke et al., 2019) is used. (Sgrvinod, n.d.)

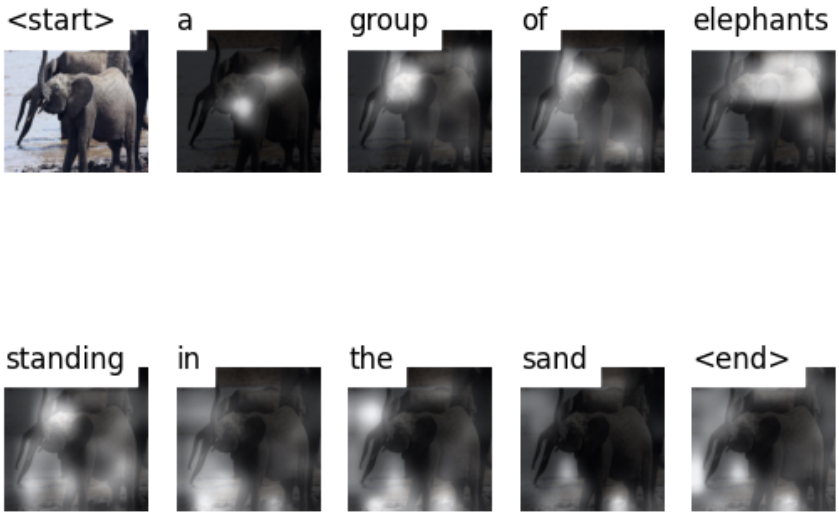


Figure 3: Visualization of attention. The highlighted parts show the attention that the model used to generate each specific word.

4.3 Generating Adversarial Samples

Generating adversarial input images can be done by using the Fast Method (EQ. 1) proposed by Goodfellow et al..

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

X^{adv} is the adversarial sample generated by taking the original image X and perpetrated it with the sign of the gradient of the loss function: $J(X, y_{true})$. Maximizes that loss function. The ϵ is a hyperparameter which can be tuned to ensure that the adversarial image still is visually the same as the original image. Finally, the adversarial image is clipped to ensure it stays in the within the 0 to 1 input range of the model. As can be seen in Figure 4 (and more examples in the appendix A), using this method images up to and including $\epsilon = 0.02$ are indistinguishable and up to and including $\epsilon = 0.16$ recognizable to humans. The *sign* method 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 successfully attack S.A.T. therefore the iterative method will be used 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 hyperparameter which naively can be set to ϵ/N. Images generated using this method are usually less visually disturb for the same epsilons. Here an epsilon of 0.040 is nearly indistinguishable, as can be seen in figure 5

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 single 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 amount of 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. Combining it with the Iterative Method 2, results in equation 5

$$X_0^{adv}, X_{n+1}^{adv} = Clip_{X,\epsilon}(X_n^{adv} + \alpha * sign(\nabla_x J(X_n^{adv}, \alpha))) \tag{5}$$

As can be seen in figure 6 the images are visually less perturbed even with a higher epsilon. With an image with a perturbation of $\epsilon = 0.160$ almost indistinguishable from the clean image. Although $\epsilon = 0.640$ is visually distorted it is still very recognizable and would still be described the same by a human.

4.4 Evaluation

The accuracy of Image Captioning models is often graded by using BLEU score (Papineni et al., 2001). It gives a score between 0 and 1 on how good a certain translation is, by comparing the

candidate translation to multiple reference translations. It is found to correlate strongly with human judgement, however one weakness is that it cannot detect synonyms. To combat that the cosine similarity between the sentences is calculated. First the sentence is embedded by a Universal Sentence Encoder(Cer et al., 2018) model, this embedding is then used to calculate the cosine similarity.

To determine if the model is indeed susceptible to distraction the BLEU-4 score (Papineni et al., 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.

(a) ϵ=0.000 (b) ϵ=0.020

(c) ϵ=0.000 (d) ϵ=0.160

Figure 4: Clean (left) and Adversarial images (right) for varying epsilon values of 0.020 and 0.160. Generated using equation 1. More values of epsilon can be found in appendix A.

(a) ϵ=0.000 (b) ϵ=0.020

Figure 5: Clean (left) and Adversarial images (right) for varying epsilon values of 0.020 and 0.160. Generated using equation 2. More values of epsilon can be found in appendix A.

(a) ϵ=0.000 (b) ϵ=0.160

(c) ϵ=0.000 (d) ϵ=0.640

Figure 6: Clean (left) and Adversarial images (right) for varying epsilon values of 0.160 and 0.640. Generated using equation 5. More values of epsilon can be found in appendix **??**.

5 Results

Adversarial Samples

Adversarial samples are generated using the iterative version of Fast Gradient Sign method as shown in equation 2. With $N = 10$ satisfactory results can be achieved, however higher N results in even better results for the same ϵ. Higher epsilons did not have an effect on BLEU score beyond 0.08 as can be seen in figure 7. This is in contrast to the cosine similarity as it does decrease further for the higher ϵ.

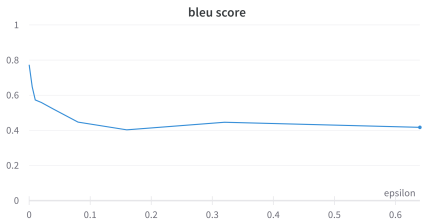


Figure 7: Average BLEU score

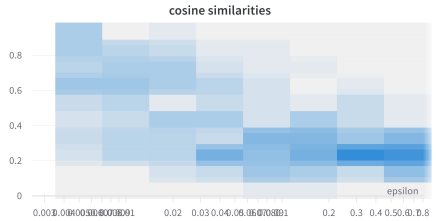


Figure 8: Cosine similarity vs epsilon (Axis is not correct yet)

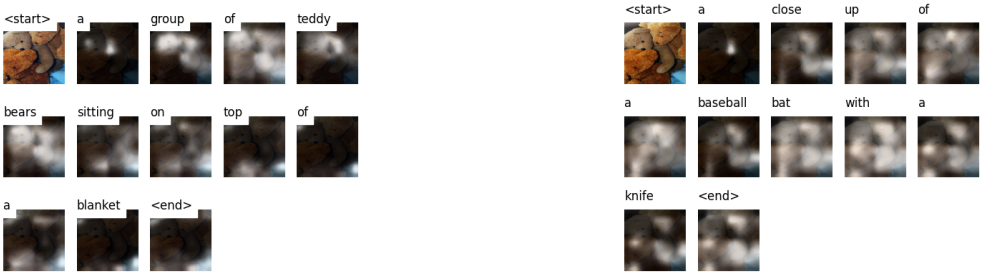


Figure 9: Clean Image (left), Adversarial Image $\epsilon = 0.02, N = 10$ (right)

As can be seen in figure 9 the attention of S.A.T, even though not explicitly attacked, is not as focused as on the clean image. This is especially visible in images that are successfully attacked. Images for which the model still is able to generate decent captions, still have a good focus on the main subjects in the image (10).

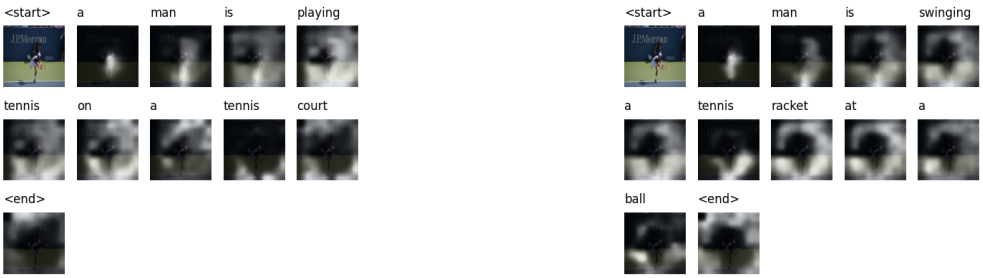


Figure 10: Clean Image (left), Adversarial Image $\epsilon = 0.02, N = 10$ (right)

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 of longer running times. The top left pixel was chosen to focus the attention on as the model focus least on it (11) (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 12. With the perturbation at most 0.04 the image is visually almost identical to the human eye (figure 13).

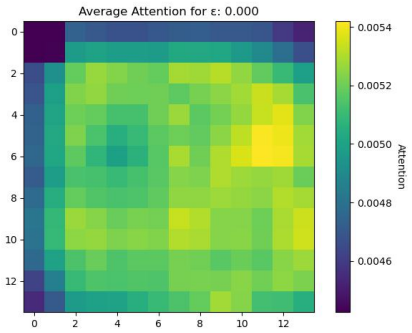


Figure 11: Average attention on clean images

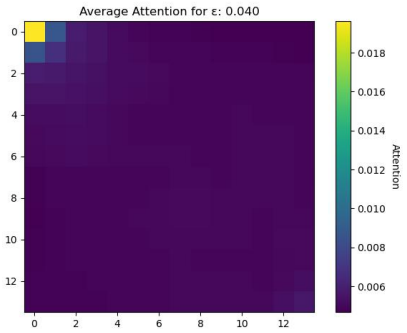


Figure 12: Average attention on adversarial images with ϵ=0.04 at 100 iterations

Figure 13: Clean Image (left), Adversarial Image $\epsilon = 0.04, N = 100$ (right)

The attention and sentence generation for figure 13 are visualized in figure 14. The model is not completely distracted and 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 14: Attention on Clean Image (left) and Adversarial Image $\epsilon = 0.04, N = 100$ (right)

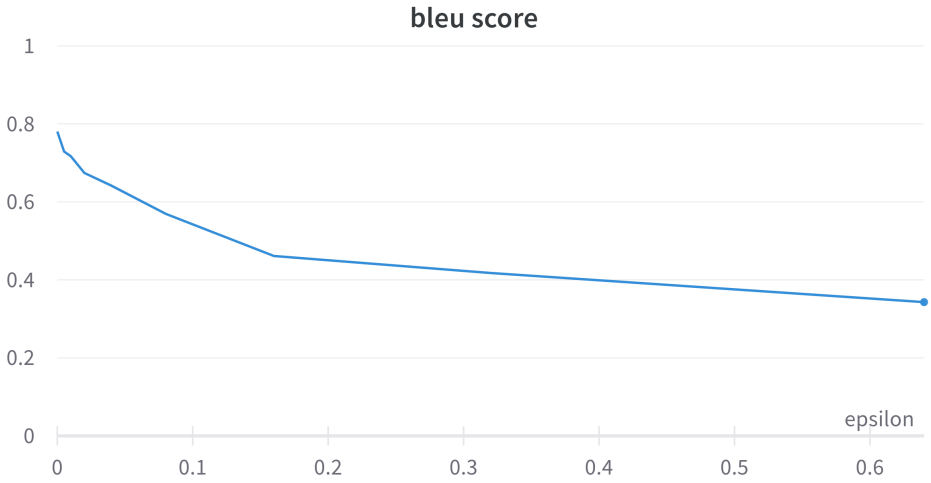


Figure 15: BLEU score during distraction over epsilon

6 Discussion

- Cosine similarity is based on a learned value, it is not a watertight value. (But neither is BLEU)
- How useful is the iterative method for training, as it takes a significant time to compute.
- Future work:

  If the adversarial samples generated are included during training is the model more robust.

  Are the adversarial samples transferable to other models, even ones not employing attention?

7 Conclusions

- S.A.T. is susceptible to adversarial samples, with the most important part in generating them is the amount of iterations. (This could use some more plots)
- Successful attacks are visible in the attention layer, even if not attack explicitly.
- The attention layer is susceptible to attacks.
- Attacking the attention is harder than attacking the sentence.
- Summarizing in attacking using the (iterative) fast gradient sign method, the most important part is iterations.

References

Aafaq, N., Akhtar, N., Liu, W., Shah, M. & Mian, A. (2021). *Controlled caption generation for images through adversarial attacks.* arXiv. Retrieved from https://arxiv.org/abs/2107.03050 doi: 10.48550/ARXIV.2107.03050 2

Bahdanau, D., Cho, K. & Bengio, Y. (2014). *Neural machine translation by jointly learning to align and translate.* arXiv. Retrieved from https://arxiv.org/abs/1409.0473 doi: 10.48550/ARXIV.1409.0473 1

Bird, E. L., Steven & Klein, E. (2009). *Natural language processing with python.* O Reilly Media Inc. Retrieved from https://www.nltk.org/_modules/nltk/translate/bleu_score.html 3

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 3

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 7

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, 3

Goodfellow, I. J., Shlens, J. & Szegedy, C. (2015). *Explaining and harnessing adversarial examples.* 1, 2, 3, 6

Hochreiter, S. & Schmidhuber, J. (1997, 12). Long short-term memory. *Neural computation*, *9*, 1735-80. doi: 10.1162/neco.1997.9.8.1735 3

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 3

Kingma, D. P. & Ba, J. (2017). *Adam: A method for stochastic optimization.* 3

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 1, 3, 6

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, 3

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., ... Dollár, P. (2015). *Microsoft coco: Common objects in context.* 3, 5

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 3

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 3, 6, 7

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox & R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc. Retrieved from http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf 5

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 3, 5

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 3

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, 3

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 3

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, 3

Zhang, H. & Wang, J. (2019). Towards adversarially robust object detection. *CoRR*, *abs/1907.10310*. Retrieved from http://arxiv.org/abs/1907.10310 3

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.

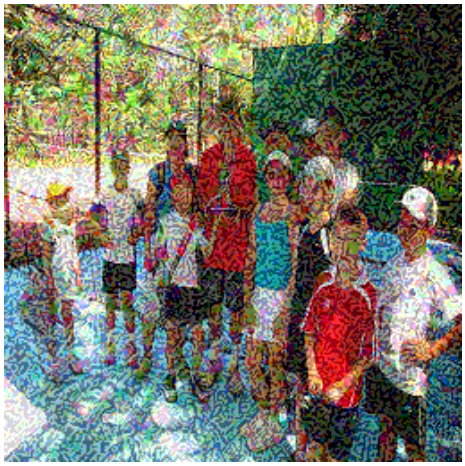
Clean image
Prediction by S.A.T.: A group of people standing around a tennis court.

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.

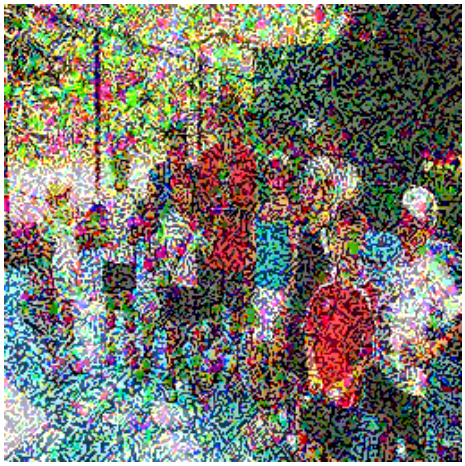
Adversarial Image with $\epsilon = 0.080$
Prediction by S.A.T.: A bathroom with a toilet and a sink.

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.

B More Adversarial Samples

Figure 16: Prediction by Show Attend and Tell on a normal image

Figure 17: Prediction on an adversarial image with $\epsilon = 0.2$ (roughly 5% of original range)

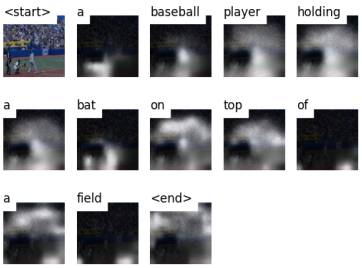


Figure 18: Prediction by Show Attend and Tell on a normal image

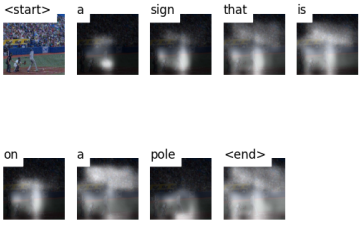


Figure 19: Prediction on an adversarial image with $\epsilon = 0.2$ (roughly 5% of original range)

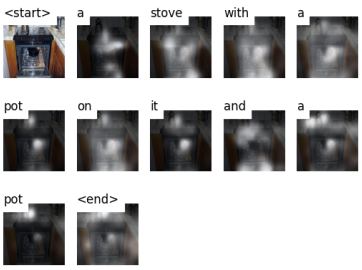


Figure 20: Prediction by Show Attend and Tell on a normal image

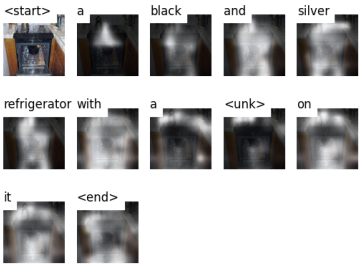


Figure 21: Prediction on an adversarial image with $\epsilon = 0.2$ (roughly 5% of original range)

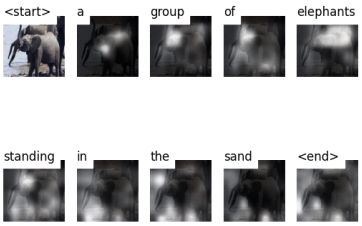


Figure 22: Prediction by Show Attend and Tell on a normal image

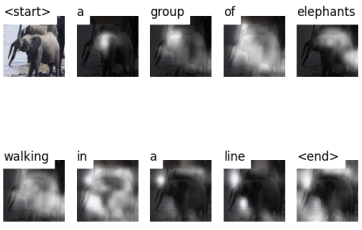


Figure 23: Prediction on an adversarial image with $\epsilon = 0.2$ (roughly 5% of original range)