Robustness of Image Captioning Models Against Adversarial Examples

Version 1.0

Ananth Chillarige, Kevin Lee, Shrijesh Siwakoti

Abstract

The problem that our team worked on for this project is training an image captioning model that utilizes the captions and Flicker8k dataset and evaluating how robust it is against adversarial attacks. In order to evaluate how our model performs, we implemented the BLEU score to see how close our model's generated caption is to the actual caption given for that image. Using the semi black box fast gradient sign method attack, we measured how well our image captioning model evaluates the captions for these perturbed images. We then applied another attack on the computer vision aspect of our model in order to completely fool the model's defense. Our image captioning model scored a BLEU-1 score of 0.522455 on the validation dataset. After creating adversarial examples using Fast Gradient Sign Method and Projected Gradient Descent, our image captioning scored a BLEU-1 score of 0.479406 and 0.451613 respectively.

1 Introduction

Fundamentally, the image captioning problem requires a model that bridges the areas of natural language processing and computer vision to form a semantic understanding of images, and subsequently generate proper sentences that describe what is going on in a given image. With the state of technology and the field of artificial intelligence rapidly progressing, image captioning technology has the potential to eventually enable systems and aid visually impaired individuals to better perceive the world around them. Before such systems are made available to the public, it is imperative to consider the security vulnerabilities due to adversarial examples.

After learning about the current state of the adversarial machine learning space, we learned that many image classification models and adversarial detectors often fail to defend against most black box attacks and all white box attacks. The field does not seem to explore other models within the machine learning space such as image captioning. Therefore, our team was curious to see how robust image captioning models were to attacks of various strengths, and gain insight as to how their architecture contributes to their performance against adversarial examples.

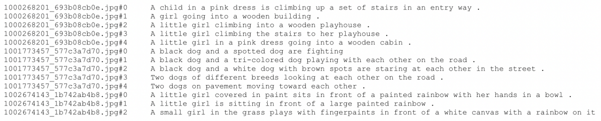
2 Related Work

(Chen et al., 2018) show that Image Captioning models typically contain a CNN for image feature extraction as well as a RNN for caption generation. They propose an adversarial attack called Show-and-Fool that creates adversarial examples similar to the natural image with randomly targeted captions or keywords that can fool neural image captioning methods. There is high transferability in this attack as well.

(Shetty et al., 2017) showed that the image captioning field has expanded but there still seem to be distinct differences in human generated captions from machine generated captions. This may be due to vocabulary size, bias from frequent captions, and word distribution. In order to combat this, they use adversarial examples in conjunction with a gumbel sampler to achieve image captioning nearly indistinguishable from human generated captions.

(Xu et al., 2019) looked at an attack used to fool image captioning models completely by creating adversarial examples with image distortion as well as partially targeted and fully targeted captions to images with absolute no relevance to the image. We draw relevance from this particular article to understand how we may be able to create a full white-box attack for image captioning models.

(Goodfellow et al., 2015) is the example paper



```
1000268201_693b08cb0e.jpg#0     A child in a pink dress is climbing up a set of stairs in an entry way .
1000268201_693b08cb0e.jpg#1     A girl going into a wooden building .
1000268201_693b08cb0e.jpg#2     A little girl climbing into a wooden playhouse .
1000268201_693b08cb0e.jpg#3     A little girl climbing the stairs to her playhouse .
1000268201_693b08cb0e.jpg#4     A little girl in a pink dress going into a wooden cabin .
1001773457_577c3a7d70.jpg#0     A black dog and a spotted dog are fighting
1001773457_577c3a7d70.jpg#1     A black dog and a tri-colored dog playing with each other on the road .
1001773457_577c3a7d70.jpg#2     A black dog and a white dog with brown spots are staring at each other in the street .
1001773457_577c3a7d70.jpg#3     Two dogs of different breeds looking at each other on the road .
1001773457_577c3a7d70.jpg#4     Two dogs on pavement moving toward each other .
1002674143_1b742ab4b8.jpg#0     A little girl covered in paint sits in front of a painted rainbow with her hands in a bowl .
1002674143_1b742ab4b8.jpg#1     A little girl is sitting in front of a large painted rainbow .
1002674143_1b742ab4b8.jpg#2     A small girl in the grass plays with fingerpaints in front of a white canvas with a rainbow on it
```

Figure 1: Input data for image file names and corresponding captions

that explains the idea of FGSM and why creating perturbations using this gradient attack may affect classification models. This is one of the papers we followed in order to implement our version of this adversarial attack.

(Madry et al., 2019) found that projected gradient descent is the strongest attack utilizing the local first order information about the network. They used PGD as a universal first order adversary to create robust networks on MNIST and CIFAR-10 using both black box and white box attacks.

3 Problem

The adversarial machine learning space mostly focuses time and effort into researching how image classification can be easily susceptible to falling victim to misclassification because of both white and black box attacks. Most of the existing research focused on trying to decipher how neural networks can classify images as well as detect adversarial examples. Modern research finds different ways to detect adversarial examples rather than classify them correctly because this problem seems to be too hard for the technology of our times. Alternatively, other research in the field is focused on creating new attack algorithms to completely fool the detection networks as well as the image classification model.

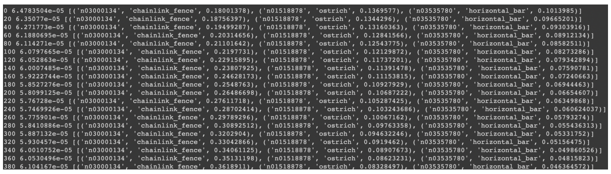
The problem that is not mentioned often is studying how robust image captioning models are. Image captioning takes in both elements of computer vision and natural language processing. Often times, attacks are focused on only fooling one element in a model's task but never focuses on both. With object detection and image classification tasks, adversarial attacks will focus on the computer vision element by trying to confuse objects within pixels of the image. However, it has not been further examined how image captioning models often withstand simple black box attacks and can be the future of research in evading these black-box attacks.

4 Methodology

Our image captioning model is trained on the Flickr8k dataset. The first step in the model's prediction architecture is to interpret the contents of the photos using the pre-trained VGG16 model. The VGG16 model pre-computes an image's features, and saves them to a file to be used later by the model in order to optimize for speed and memory usage. This feature extraction is primarily comprised of object detection to break down the contents of the image. The picture is then able to be represented by a 256 element vector. The model then cleans text data associated with the feature dataset for training. After loading clean data, the model has a sequence processor layer that creates word embeddings and uses a LSTM RNN to produce a 256 element vector with dropout regularization in order to reduce overfitting. Finally, the decoder portion of the model merges both the photo and sequence processor vectors using an addition operation and makes a softmax prediction over the entire output vocabulary to make a prediction. As one can see, the model's architecture can be broken up into a few main sections: photo feature extraction, sequence processing, and decoding.

Our work primarily focused on attacking the photo feature extraction stage of the model rather than attacking the text based stage in our model. We ran both Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) attacks on the VGG16 model to produce adversarial images that we passed into our network to evaluate using these new examples. In order to create labels for our dataset given the VGG16 model, we selected 20 random photos from the validation data and ran a pretrained VGG16 model to predict the top label for each photo. We used the top prediction as the label for each photo. Afterwards, we created 20 adversarial examples using our FGSM code. It created the perturbations for 100 epochs by adding noise from the epsilon (epsilon = 0.01) to the features from each photo. We took similar steps for our PGD attack. We ran PGD 10 times on the same 20 photos, with an epsilon value of 0.01. We then ran the newly created adversarial examples and evaluated our model's bleu score for the predicted captions against the actual.

We thought that attacking the VGG model would ultimately destroy the image captioning model's performance as the decode layer relies on



```
0 6.4783504e-05 [('n03000134', 'chainlink_fence', 0.18001378), ('n01518878', 'ostrich', 0.1369577), ('n03535780', 'horizontal_bar', 0.1013985)]
20 6.35077e-05 [('n03000134', 'chainlink_fence', 0.18756397), ('n01518878', 'ostrich', 0.1344296), ('n03535780', 'horizontal_bar', 0.09665201)]
40 6.271773e-05 [('n03000134', 'chainlink_fence', 0.19499287), ('n01518878', 'ostrich', 0.13160363), ('n03535780', 'horizontal_bar', 0.09303916)]
60 6.1880695e-05 [('n03000134', 'chainlink_fence', 0.20314656), ('n01518878', 'ostrich', 0.12841566), ('n03535780', 'horizontal_bar', 0.08912134)]
80 6.114271e-05 [('n03000134', 'chainlink_fence', 0.21101642), ('n01518878', 'ostrich', 0.12543775), ('n03535780', 'horizontal_bar', 0.08582511)]
100 6.0797665e-05 [('n03000134', 'chainlink_fence', 0.2197731), ('n01518878', 'ostrich', 0.12129872), ('n03535780', 'horizontal_bar', 0.08273286)]
120 6.052863e-05 [('n03000134', 'chainlink_fence', 0.22915895), ('n01518878', 'ostrich', 0.11737201), ('n03535780', 'horizontal_bar', 0.079342894)]
140 6.0007485e-05 [('n03000134', 'chainlink_fence', 0.23807925), ('n01518878', 'ostrich', 0.11391478), ('n03535780', 'horizontal_bar', 0.07590781)]
160 5.9222744e-05 [('n03000134', 'chainlink_fence', 0.24628173), ('n01518878', 'ostrich', 0.11153815), ('n03535780', 'horizontal_bar', 0.07240663)]
180 5.8527276e-05 [('n03000134', 'chainlink_fence', 0.2548763), ('n01518878', 'ostrich', 0.10927929), ('n03535780', 'horizontal_bar', 0.06944463)]
200 5.8099125e-05 [('n03000134', 'chainlink_fence', 0.26486698), ('n01518878', 'ostrich', 0.10687222), ('n03535780', 'horizontal_bar', 0.06654607)]
220 5.76728e-05 [('n03000134', 'chainlink_fence', 0.27611718), ('n01518878', 'ostrich', 0.105287425), ('n03535780', 'horizontal_bar', 0.06349868)]
240 5.7469926e-05 [('n03000134', 'chainlink_fence', 0.28702414), ('n01518878', 'ostrich', 0.103243686), ('n03535780', 'horizontal_bar', 0.060624037)]
260 5.775901e-05 [('n03000134', 'chainlink_fence', 0.29789296), ('n01518878', 'ostrich', 0.10067162), ('n03535780', 'horizontal_bar', 0.05793274)]
280 5.8410886e-05 [('n03000134', 'chainlink_fence', 0.30892512), ('n01518878', 'ostrich', 0.09763358), ('n03535780', 'horizontal_bar', 0.055436313)]
300 5.887132e-05 [('n03000134', 'chainlink_fence', 0.3202904), ('n01518878', 'ostrich', 0.094632246), ('n03535780', 'horizontal_bar', 0.05331752)]
320 5.930457e-05 [('n03000134', 'chainlink_fence', 0.33042866), ('n01518878', 'ostrich', 0.0919462), ('n03535780', 'horizontal_bar', 0.05156475)]
340 6.0010752e-05 [('n03000134', 'chainlink_fence', 0.34061125), ('n01518878', 'ostrich', 0.08907673), ('n03535780', 'horizontal_bar', 0.049860526)]
360 6.0530496e-05 [('n03000134', 'chainlink_fence', 0.35131198), ('n01518878', 'ostrich', 0.08623231), ('n03535780', 'horizontal_bar', 0.04815823)]
380 6.104167e-05 [('n03000134', 'chainlink_fence', 0.3618911), ('n01518878', 'ostrich', 0.08328497), ('n03535780', 'horizontal_bar', 0.046364572)]
```

Figure 2: Top 3 predictions for a photo using VGG16 changing after each iteration of FGSM

generated photo features to ultimately make predictions.

5 Evaluation and Results

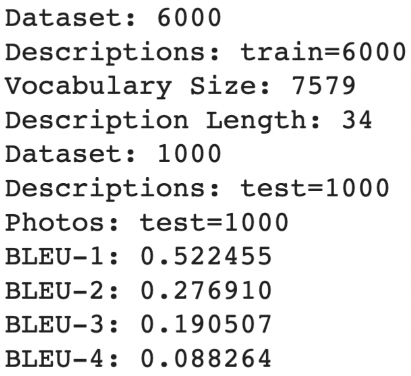
Our captioning model after training had a BLEU-1 score 0.522455, BLEU-2 score 0.276910, BLEU-3 score 0.190507, and BLEU-4 score 0.088264 on the validation data. Due to computational limitations and small training set size, we were unable to increase the BlEU score past this. We had great difficulty with loading packages on the Great Lakes cluster as well as AWS, and had file syncing issues with Google Drive and Google Colab. At the end, we were able to get the model working on Google Colab, but given time constraints and compute limits, we had to limit our training data size to the Flickr dataset and epoch count to 100 iterations.

When running our captioning model on the perturbed FGSM examples, we had a BLEU-1 score 0.479406, BLEU-2 score 0.190219, BLEU-3 score 0.126234, and BLEU-4 score 0.050315. Compared to our validation data, the BLEU scores dropped minimally across the board.

When running our captioning model on the perturbed PGD examples, we had a BLEU-1 score 0.451613, BLEU-2 score 0.213442, BLEU-3 score 0.143113, and BLEU-4 score 0.063550. Again, the BLEU scores dropped across the board compared to the validation set. Surprisingly, while PGD perturbations dropped the BLEU-1 score lower than the FGSM perturbations did, the BlEU-2, BLEU-3, and BLEU-4 scores dropped less with PGD attack compared to the FGSM attack. This was an interesting result to see, as PGD is supposed to be a stronger attack, and visually, the perturbations looked a lot more severe.

6 Conclusion

Our project was to validate whether or not image captioning models are more robust to adversarial attacks. We found that while image captioning



```
Dataset: 6000
Descriptions: train=6000
Vocabulary Size: 7579
Description Length: 34
Dataset: 1000
Descriptions: test=1000
Photos: test=1000
BLEU-1: 0.522455
BLEU-2: 0.276910
BLEU-3: 0.190507
BLEU-4: 0.088264
```

Figure 3: Bleu scoring on our validation data set.

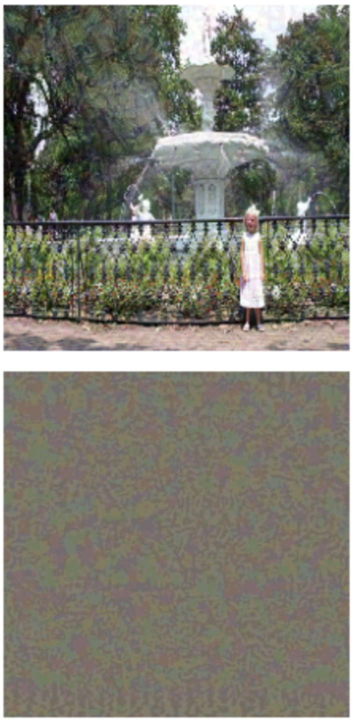


Figure 4: Adversarial example with perturbation created via FGSM alongside the amount of perturbation

models are often hard to hyper tune for accurate results, it makes little difference adding perturbations to images for this model. Based on the minimal reduction in BLEU scores due to attacks, we saw that our captioning model still captures roughly the same caption it would for examples that are not perturbed.

In the future, we would love to continue research and explore this problem space further by training our captioning model on more data to improve its base level BLEU score performance to mitigate any experimental bias due to it's under performance. Also, we plan to attack the LSTM and decoder portions of the model to create adversarial examples, and see how that impacts the model's performance.

7 Other Things We Tried

Our team ran into several problems as the great lakes computing cluster would not properly run when trying to train models or evaluate pre-trained models. This became a problem where our team tried exploring AWS EC2 and GCP to try and run these models. However, our team was unsuccessful in these attempts as well. Module loading and library importation seems to be a big issue with these cloud computing platforms and has been a huge issue. Memory storage is another problem that we ran into when initially trying to use the MS COCO dataset. Because the training, validation and annotation sets for that dataset is so large in memory, we had to switch to the Flicker8k dataset.

8 Group Effort

This group project was done together by all three members contributing evenly. We met up many times throughout the semester in order to try and complete our proposed project. We met with Tianji many times for advice and guidance on the scope of our project.

References

Hongge Chen, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, and Cho-Jui Hsieh. 2018. Attacking visual language grounding with adversarial examples: A case study on neural image captioning.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2019. Towards deep learning models resistant to adversarial attacks.

Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. 2017. Speaking the same language: Matching machine to human captions by adversarial training.

Yan Xu, Baoyuan Wu, Fumin Shen, Yanbo Fan, Yong Zhang, Heng Tao Shen, and Wei Liu. 2019. Exact adversarial attack to image captioning via structured output learning with latent variables.