

Evasion Attack in Show and Tell Model

Dongseop Lee, Hyunjin Kim, Jaecheol Ryou

Dept. of Computer Science & Engineering, Chungnam National University, Daejeon, Korea

dajababa09@gmail.com, be.successor@gmail.com, jcryou@cnu.ac.kr

Abstract— Recently, deep learning technology has been applied to various fields with high performance and various services. Image recognition is also used in various fields with high performance by incorporating deep learning technology. However, deep learning technology is vulnerable to evasion attacks that cause the model to be misclassified by modulating the original image. In this paper, we generate an adversarial example using the forward-backward-splitting iterative procedure. Then perform an evasion attack on the show and tell model.

Keywords— AI, Security, Adversarial example, Evasion attack, Show and tell model

I. INTRODUCTION

Deep learning technology shows high performance in various fields such as voice recognition and intrusion detection. Likewise, deep learning image recognition models are used in a variety of fields, including autonomous cars, immigration, payment systems, and Google image search[1,2,3]. However, there are security issues with the deep learning model. One of the security problems of DNN is the evasion attack. An evasion attack generates an adversarial example that minimally modulates the data so that the deep learning model misclassifies the data. For example, when using DNN technology in an autonomous vehicle, an autonomous vehicle may misrecognize a left turn sign as a right turn sign. and that incident cause a major accident.

In this paper using a forward-backward-splitting iterative procedure, we generate an adversarial example that causes the model to misclassify. In addition, we performed and analyzed the evasion attack on the show and tell model using an adversarial example. The show and tell model is a deep learning based image recognition model.

The structure of this paper is as follows. Chapter 2 explains the evasion attack. Chapter 3 describes the target model, show and tell model. It also explains the Forward-backward-splitting iterative procedure, a technique for generating adversarial examples. In Chapter 4, we perform an evasion attack targeting the Image Caption model. Finally, we conclude in Chapter 5.

II. EVASION ATTACK

An evasion attack generates an adversarial example in which the target model misclassifies the data by minimally modulating the data. The adversarial example modulates the feature values of the data classified in the base class to become the target class. But the human should not be aware of

the modulation of the data. So data should be generated with minimal modulation.

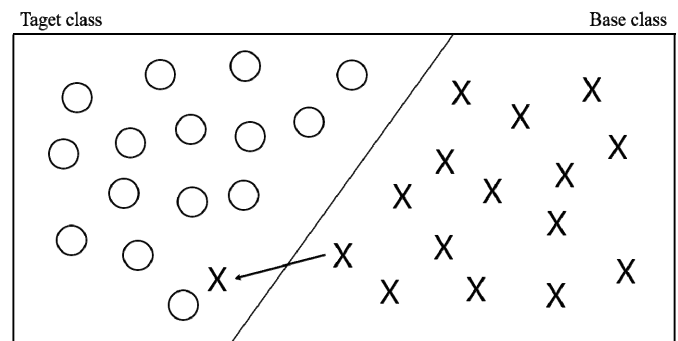


Figure 1. evasion attack

A. Classification by attack goal

Attack goals for evasion attacks can be classified into untargeted attacks and targeted attacks. Firstly an untargeted attack is for an attacker to perform an evasion attack so that the deep learning model classified it as any other class than the original class[4]. For example, in a model that classifies the handwritten image data set, MNIST (Modified National Institute of Standards and Technology), an attacker performs an evasion attack to misclassified the number 0 as any other number.

Secondly, a targeted attack is an attacker performs an evasion attack so that the deep learning model classified the attacker's intended number rather than the number 0. In general, an untargeted attack is faster to generate an adversarial example than a targeted attack. However, targeted attacks are more effective in that they can be misclassified into the intended class.

B. Classification according to the attacker's knowledge

Evasion attacks can be classified into white-box attacks and black box attacks depending on how much the attacker knows about the target model. white box attack is The attacker attacks the target model, knowing all the information such as the model's structure, parameters, and class. This attack is easy to generate an adversarial example because it knows everything about the target model. In the real world, however, it is not easy for an attacker to know all the information about the target model.

A black box attack is a realistic attack compared to a white box attack, in which an attacker attacks a target model without any information. In this attack, the attacker only knows the

output (class), which is the information received after querying the target model. This attack first performs a model extraction attack using query values to reconstruct the model and then performs an evasion attack. However, the black box attack has a very low attack success rate compared to the white box attack.

III. EVASION ATTACK ON SHOW AND TELL MODEL

In this chapter, we perform an evasion attack against the image captioning model, one of the show and tell models. First, explain the show and tell the model and then generate an adversarial example for the model. In this experiment, we performed a white-box attack and targeted attack. We also generate an adversarial example using the forward-backward-splitting iterative procedure.

A. Show and tell model

The show and tell model explains the image inappropriate sentences using neural network model[5]. This model is used by the CNN encoding an image and using the RNN is a sentence generation. In this paper, we encode the CNN using inception v3, which is a pre-trained model. After that, we create a model by sentence generation using LSTM instead of RNN perform an evasion attack on the model.

inceptionV3[6] is a pre-trained model developed by Google in 2014. This model is a CNN model trained on over 1 million images in the ImageNet database. The existing CNN model uses the same size convolution filter for the input image. The inception v3 model, on the other hand, is constructed using filters of various sizes, such as $1 * 7$ or $7 * 1$. These various size filters can express images more efficiently. There are 48 layers in this network. And images can be categorized into 1,000 object categories, including keyboards, mice, pencils, and animals. Inception v3 won the 2014 ImageNet competition using the following structure:

Figure 4 is the structure of the show and tell the model. First, we encode the image to $256 * 256$. Then we encode the image using the inception v3 model. Using that output, the lstm model produces a sentence generation of the image.

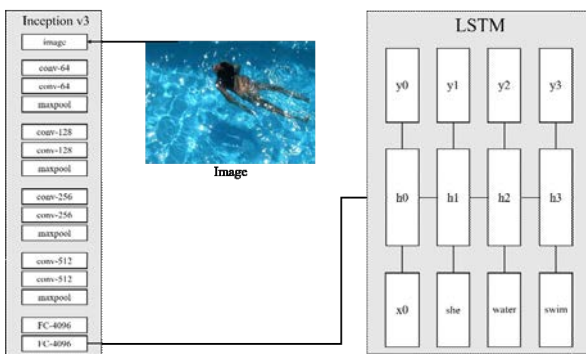


Figure 2. Show and tell model

The data set uses the Flickr8k dataset. Flickr8k dataset contains 6000 training images, 1000 validation images and 100 testing images. There are five captions for each image. Figure 5 shows the result after model training.



Figure 3. Show and tell model test

B. Forward-backward-splitting iterative procedure

As a way to generate an adversarial example is various methods, but in this paper, we used the forward-backward-splitting iterative procedure[7,8]. The first (forward) step is simply a gradient descent update to minimize the L^2 distance to the target instance in feature space. The second (backward step) is a proximal update that minimizes the feature distance from the base instance in input space. The coefficient β is tuned to make the adversarial example look realistic in input space, enough to fool an unsuspecting human observer into thinking the attack vector image has not been tampered with..

adversarial example generate repeats the forward and backward steps, if either of these conditions is met, the loop ends.

Algorithm1 Adversarial Example Generation

Input: target instance t , base instance b , learning rate λ

Initialize x : $x_0 \leftarrow b$

Define: $L_p(x) = \|f(x) - f(t)\|_2$

For $i = 1$ **to** $maxIters$ **do**

Forwardstep: $x_i = x_{i-1} - \lambda \nabla_x L_p(x_{i-1})$

Backwardstep: $x_i = (x_i + \lambda \beta b) / (1 + \beta \lambda)$

if $diffFeature < 1e-10$ **or** $diffImage < 2.9$

break

end for

The first condition is that in the forward step, if the feature value between the target image and the base image is less than $1e-10$, it is determined that the feature values of the two images are close enough, and the loop ends. The second condition is that in the backward step if the image distance between the adversarial example and the base image is less than 2.9, it concludes that it can no longer modulate and then terminates the iteration. an adversarial example that generated from the second condition, the evasion attack success rate is low because the feature distance of adversarial image and the target image is not close enough.

We generate an adversarial example using the forward-backward-splitting iterative procedure, that feature of target instance appointed by an attacker

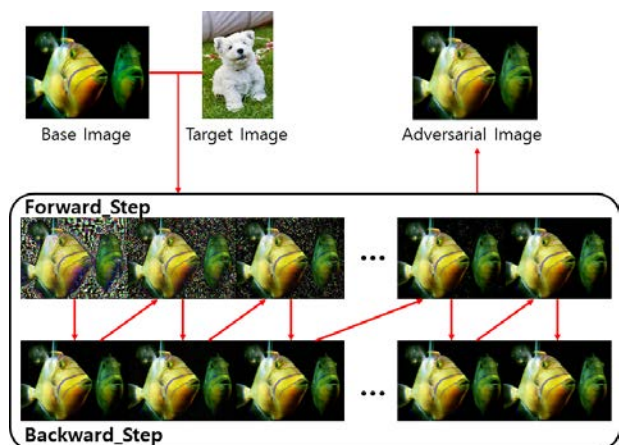


Figure 4. Adversarial example generate

Figure 2 shows the modulation of the base image in the forward and backward steps in the middle of generating an adversarial example. We generated an adversarial example, using a fish image as the base image, and a dog image as the target image. In the forward step, the image has a large modulation. Because we only consider the feature distance between the target image and the adversarial example without considering the image distance of the base image and the adversarial example. Inversely, in the backward step, minimize image modulation with close the image distance of the base image and the adversarial example. Repeat the forward and backward similar to generate an adversarial example that features value is close to the target image feature but whose image distance is close to the base image.

IV. EXPERIMENT

In this chapter, We performed a white-box attack and a targeted attack on the show and tell models. The target model is an Image Caption Model created using Inception V3. First, generate an adversarial example using the Forward-backward-splitting iterative procedure. Then use the adversarial example as input to the model and check the results.

A. Generate adversarial example

In figure 6, (a) shows the image distance between the base image and the adversarial imaged, and (b) shows the feature distance between the target image and the adversarial image. As the step progresses, the image distance from the base image gets closer. And the feature value also gets closer to the target distance from the target image.

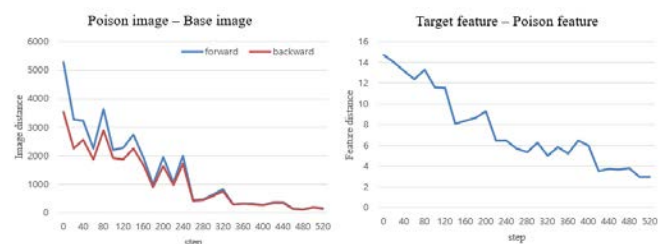


Figure 5. According to the progress of the gradient of the forward and backward steps of the image distance and feature distance values

In Figure 6, the adversarial example is shown in comparison with the target image.

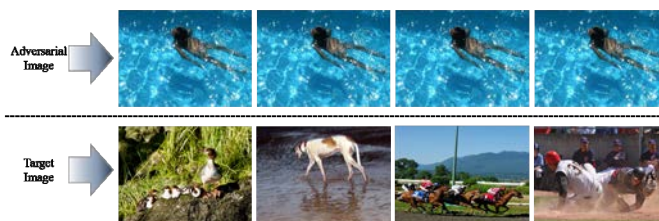


Figure 6. Sample target and adversarial example

the adversarial example looks no different from the base image. However, the feature value of the adversarial example is similar to the target image.

B. Evasion attack result

Five random photographs were selected from the test data, one was selected as the base image, and the other four were selected as the target image.

Using the base image, we made an adversarial example using the feature value of the target image and input it into the image caption model. The result of the attack was that the image caption model described the target image, even though it was the base image for humans. Figure 7 shows the results of the evasion attack.

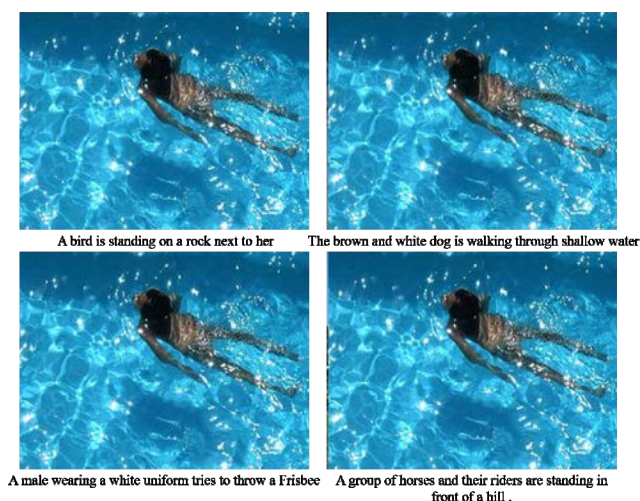


Figure 7. evasion attack result

V. CONCLUSION

In this paper, we generate an adversarial example using forward-backward-splitting iterative procedure and perform an evasion attack on the image caption model using inception V3. When entering the adversarial example, the target model describes the target image rather than the base image, which is the visible image.

Recently, deep learning technology has been applied to various fields with high performance and various services. However, the current deep learning model is very vulnerable to evasion attack. so that, we needed the research on the safety of deep learning models. Currently evasion attack shows 100% success rate in white box environment, but it is very low in black box environment. In our future work will research defense techniques that use random elements to make the white box unpredictable, similar to a black box environment.

ACKNOWLEDGMENT

This research was supported by Korea Electric Power Corporation. (Grant number:R18XA05)

REFERENCES

- [1] M. Wang, W. Deng, "Deep Face Recognition: A Survey," arXiv preprint arXiv:1804.06655, 2018.
- [2] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep Hypersphere Embedding for Face Recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp. 212-220, 2017.
- [3] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-Margin Softmax Loss for Convolutional Neural Networks," Proceedings of the 33rd International Conference on Machine Learning, pp. 507-516, 2016.
- [4] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z.B. Celik, and A. Swami, "The Limitations of Deep Learning in Adversarial Settings," 2016 IEEE European Symposium on Security and Privacy, IEEE, pp. 372-387, 2016.
- [5] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan." Show and Tell: A Neural Image Caption Generator", arXiv:1411.4555,2014
- [6] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, "Going Deeper with Convolutions" 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015
- [7] Tom Goldstein, Christoph Studer, and Richard Baraniuk. "A field guide to forward-backward splitting with a fast implementation". arXiv preprint arXiv:1411.3406, 2014.
- [8] Ali Shafahi, W. Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, Tom Goldstein, "Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks." arXiv:1804.00792, 2018.



DongSeop Lee Graduated from Chungnam National University in Daejeon in 2019. Since March 2019, he has been working on a master's program in computer communication and security at Chungnam National University in Daejeon.

His research interests in cryptography and machine learning.



Hyunjin Kim 2015: Graduated from Chungnam National University, 2017: Master of Computer Engineering, Chungnam National University, 2017 ~ present: Ph.D., Computer Engineering, Chungnam National University.

His research interests in machine learning, cryptography and Computer network.



Jaechol Ryou In February 1985, he graduated with a degree in Industrial Engineering from Hanyang University, and in May 1988 he received a master's degree in computer science from Iowa State University. In December 1990, he received a Ph.D. in Computer Science from Northwestern University.

His research interests in machine learning, cryptography, Mobile security and Block chain.