CS 725 Report

An In-depth Discussion of Adversarial Image Generator (AIG) in Real-life Scenarios

Yezhou Liu

UPI: Yliu442

Abstract

Güera et al. proposed a counter-forensic method which is the adversarial image generator (AIG). By using AIG, people can alter images to make camera model identification convolutional neural networks (CNNs) to misclassify them. AIG exposes vulnerability of existing camera model identification CNNs.

In this report, we introduce the strengths and limitations of AIG, and discuss details of its design and experiments. By analysing three articles, we show that AIG would not be widely used by attackers in the real-world.

Introduction

Güera et al. proposed a counter-forensic method (Adversarial Image Generator, AIG) for convolutional neural network (CNN) based camera model identification. AIG can generate adversarial images which can be misclassified by camera model identification CNNs with high confidence (Güera et al., Jul 2017). In this report, we would discuss the details about AIG, and explain why AIG does not apply to real-life scenarios.

Our discussion is mainly based on three articles: (Güera et al., Jul 2017) , (Böhme & Kirchner, 2013) and (Goljan, Fridrich, & Filler, Feb 5, 2009). Other materials are supporting articles, which we would not go through them in details.

Research question of the report

Güera et al. claimed that CNN-based camera model identification architectures are vulnerable to AIG. Since CNN is widely used in modern world, especially in image processing field, its security is of

great concern. We are interested in if AIG has potential to be used in real-life scenarios by attackers. From Lampson's point of view (Butler W Lampson, 2004) , the real-life security is about value, locks and punishment. We do not assume there is perfect defense in a real-world system, we would only consider and discuss the difficulties for applying AIG in reality.

Our main research question of this report is: can AIG be widely applied by attackers in real-life?

Counter-forensic classification of AIG

AIG can both work with *fast gradient sign method* (FGSM) and *Jacobian-based saliency map attack* (JSMA) to craft adversarial examples to perform targeted and untargeted attacks to camera model identification CNNs (Güera et al., Jul 2017). Güera et al. claimed that the only access required for AIG is the access to the predictions of the target CNNs. A block diagram of AIG is shown in figure 1. AIG can be added to the pipeline of a forensic CNN. AIG first takes the K patches (the first step of Güera et al.'s camera model identification CNN pipeline is to randomly extract K non-overlapping patches of size 32*32 pixels from the original image) as input, then perturb these patches by applying FGSM model or JSMA model to achieve misclassification.

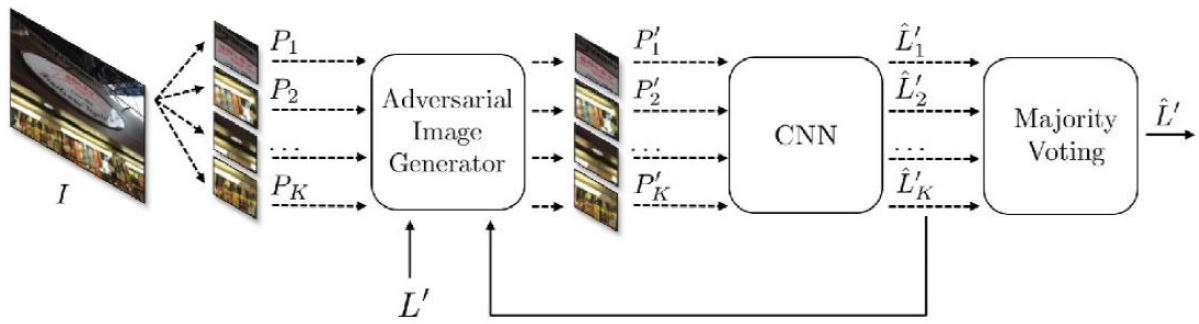

Figure 1. Block diagram of AIG. Reprinted from fig.2 (Güera et al., Jul 2017)

Böhme and Kirchner introduced a set of definitions and the classification of counter-forensic techniques (Böhme & Kirchner, 2013) . The design space of counter-forensic techniques is shown in figure 2. Following Böhme and Kirchner's definitions, we would classify AIG to "post-processing attack", "targeted" and about "security".

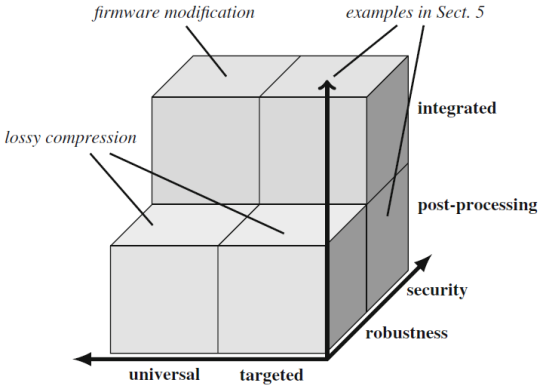

Figure 2. Design space for counter-forensic techniques. (In this figure, "examples in Sect. 5" includes suppressing traces of image processing and adding traces of authentic images.) Reprinted from fig.3 (Böhme & Kirchner, 2013)

A great illustration of post-processing attack and integrated attack is shown in figure 3. In the figure, N is the set of natural views with an infinite size; A is the collection of image acquisition functions; P is the concatenation of image processing functions and I represents the classes of images. Both two types of attacks can make the forensic classifier misclassify the samples to class *l'* (which is written as $\hat{l}$ in the figure). Post-processing attack modifies the generated image I_k to $I_{l'}$ to achieve misclassification; while integrated attack directly generates $I_{l'}$ – which would be classified as *l'*. In other words, post-processing methods perform attack after the original image generation process; and integrated approaches use a tuple of new methods *acquire'* and *process'* to replace the original image generation functions *acquire* and *process*. Since AIG does not acquire and process images from the start, which only adds an additional processing step to modify images, we can find that AIG performs a post-processing attack.

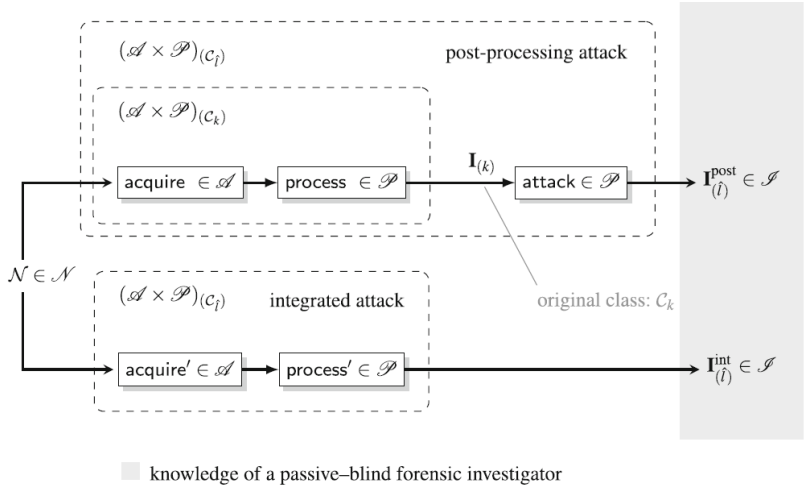

Figure 3. Post-processing and integrated counter-forensic attacks. Reprinted from fig.4 (Böhme & Kirchner, 2013)

A targeted attack takes advantage of the weaknesses of a particular forensic algorithm (the target). This type of attack would directly relate to the forensic algorithm's image model, which can be detected by other forensic algorithms or improved models. Differently, a universal attack maintains statistical properties of an image, which can hide the manipulations of the attack, even faces unknown forensic tools (Böhme & Kirchner, 2013). AIG performs targeted attack rather than universal attack, because it attacks a particular forensic algorithm – camera model identification CNNs. Since AIG visually changes an image when modifying it, an adversarial image generated by AIG can be detected by humans or other forensic tools.

Following Böhme and Kirchner's definition of robustness and security, Güera et al.'s experiments of AIG could prove that most of camera model identification CNNs are insecure; and their work was irrelevant to CNNs' robustness. When people talk on robustness, they consider if a forensic algorithm can handle legitimate cases or post-processing (for instance lossy compression); when people discuss security, they pay attention to if a forensic tool can detect and prevent illegitimate examples created by attackers (Böhme & Kirchner, 2013). In Güera et al.'s research, they intentionally use AIG to produce adversarial examples to attack camera model identification CNNs, thus their study was focused on security of the CNN-based camera model identification models, rather than the robustness.

Interestingly, Güera et al. mentioned in their article that "we will explore viable adversarial example detection methods and defense techniques to increase the robustness of CNN-based camera model detectors" (Güera et al., Jul 2017) . If this sentence was talking about the future AIG work, in Böhme and Kirchner's point of view, it was not about robustness, it was on security, since AIG intends to create illegitimate adversarial examples to spoof the camera model identification CNNs.

Advantages and limitations of AIG

Figure 4 shows AIG's experiment results. We can find that AIG achieved a very high error rate, which means it can make the CNNs to misclassify images with a very high probability (over 91.4% in FGSM, over 99% in JSMA). The "confidence score" in figure 4 refers to the average probability value of the camera mode labels for images in the test set, where the probability is the highest probability value from the softmax layer of the CNN (Güera et al., Jul 2017).

AIG could greatly increase the CNN's false rejection rate (FRR) and false acceptance rate (or false alarm rate, FAR). False rejection is to reject an image actually taken by the target camera model. False acceptance is to accept an image which was not originally from the camera model (Goljan et al., Feb 5, 2009). In a security system, compare to false rejection, false acceptance often causes more damage. For instance, in a fingerprint identification system, a false rejection means an authorized person was rejected by the guard of the system, he or she just needs to try again; while a false acceptance means there is an unauthorized person entered the system! Lampson (Butler W Lampson, 2004) proposed a similar point of view: when we talk about security, we usually pay attention to the negative aspect – to keep the bad guys out.

By using FGSM which performs untargeted image manipulation, AIG can make the CNN classifier reject more than 91.4% images which should be accepted (increase FRR). By using JSMA, AIG can produce a specific misclassification class, which can make more than 99.2% of generated adversarial images to be accepted by the CNN (increase FAR). Similar to FGSM, using JSMA also can increase FRR. Thus, for attackers, AIG with JSMA could be a more powerful weapon.

ϵ value	Error rate (%)	Confidence Score (%)
0.001	91.4	97.7
0.002	91.7	97.2
0.003	92.2	96.7
0.004	92.7	95.8
0.005	93.1	95.3
0.006	94.1	95.1
0.007	94.5	94.2
0.008	95.3	93.6
0.009	95.9	93.0
0.01	96.2	92.3

Target Camera Model	Error rate (%)	Confidence Score (%)
AS-One	99.5	87.7
ES-D5100	99.3	88.6
MK-Powershot	99.3	88.4
MK-s860	99.7	88.5
PAR-1233	99.7	87.9
PAR-1476	99.4	88.1
PAR-1477	99.5	88.2
PAR-A015	99.6	88.4
PAR-A075	99.3	87.8
PAR-A106	99.2	87.9

Figure 4. AIG's experiment results. Left: FGSM, right: JSMA. Reprinted from table 3 and table 4 (Güera et al., Jul 2017)

AIG has many advantages. AIG does not require the access to the network training, which is good for a counter-forensic method – in real-life scenarios attackers obviously cannot touch the training process of a forensic CNN classifier. Since AIG performs post-processing attack, people may use it to modify or forge existing images; which is much useful than other counter-forensic tools which only can generate new adversarial images. As mentioned, AIG can work with both FGSM model and JSMA model, the way they work possibly could be generalized. FGSM uses derivative of the loss function of the target CNN, JSMA exploits the forward derivative of the CNN. These two algorithms not only could be used in attacking camera model identification CNNs, but also could be used to attack other types of forensic CNNs.

Besides the advantages, AIG also has its limitations. Firstly, when people use AIG to alter an image, in order to achieve misclassification, AIG would bring visual changes to the image. Since AIG-altered images are visually different from the original images, these forged images could probably be detected by human detectors or other detection tools; especially when the detectors know the original look of the images. Secondly, although AIG does not need to access to the training process of its target CNNs, it still needs the access to the CNNs' predictions - there is no guarantee that an attacker can get such access.

Problems for applying AIG to real-life scenarios

AIG shows the vulnerabilities of CNN-based camera model identification architectures (Güera et al., Jul 2017), but the authors did not mention if camera model identification CNNs are widely used or

not. Tuama et al. claimed that "from the state of the art mentioned above, CNN approach has not been used for camera identification" in 2016 (Tuama, Comby, & Chaumont, 2016). Although "camera identification" is not exactly same to "camera model identification", it probably implied that using CNN in camera identification or camera model identification was a work-in-progress for researchers.

Another issue for using AIG is that Güera et al. did their experiments only based on 1611 images acquired from 10 camera models. In real-life scenarios, there are more than 10 camera models need to be identified. Goljan et al. (Goljan et al., Feb 5, 2009) did a large-scale test by using one millions images from 6896 cameras in 150 models. They claimed that most of camera sensor identification methods (CSIs) were only evaluated for a limited number of cameras, which was less than 20. Although Güera et al. worked on camera model identification rather than CSI, their experiment result was also from less than 20 cameras, which is not very convincing.

Güera et al.'s 10 camera models were from different types of cameras including digital single-lens reflex camera and phone cameras. They claimed that they used a flat-field image set to test AIG, which was harder to alter without bringing visual changes. Unfortunately, they did not mention if these 10 models are the most popular models in the market – if these 10 models are all popular ones from real-world, their experiment results would be more meaningful.

Compares to Güera et al.'s work, Goljan et al.'s experiments were actually about robustness. Their work was based on images collected from Flickr, which were not modified/forged by attackers. The main reason caused the misclassification in Goljan et al.'s experiments was the quality of test images are low. They wanted to figure out which camera models made camera model identification works less reliably, but they failed, because their dataset could not provide enough evidence for this question. Goljan et al. claimed their large-scale test provided upper bounds of error rates of camera model identification method, which FRR <= 0.0238 and FAR <= $2.4 * 10^{-5}$.

From Güera et al.'s experiments, we can find AIG can destroy both a camera model identification CNN's FAR and FRR, which is a good result for a counter-forensic method; but the authors did all of these experiments on an image set with only 10 camera models. There is no clear evidence can prove that AIG can attack CNNs working on other camera models - for instance, the camera models tested in Goljan et al.'s research.

As mentioned, AIG changes the visual appearance of images. Güera et al. sampled image blocks with 32*32 pixels from the original images to build a test, but real-world images are much larger. For instance, a high-resolution image can be 9600*7200. To make sure this image is totally altered by AIG, by using Güera et al.'s design, we need to cut the image into (9600*7200) / (32*32) = 67500 patches, then use AIG to modify these patches. Thus, for this image, visual changes may appear in 67500 patches/places, which may make the generated image looks unnatural – since AIG almost modified everywhere of the original image.

Another technique questions about using AIG in real-world is: the CNN built by Güera et al. cuts images into 32*32-pixel patches, which the authors mentioned the design was original motivated by

other researchers' solid work on DenseNet on CIFAR-10 image dataset. Do all the real-world camera model identification CNNs cut images into the same size? Unfortunately, the authors did not mention these details.

Besides the discussion on "if AIG could be widely used", we would also love to briefly discuss these questions: if an attack wants to use AIG to attack a forensic CNN, what can be his/her goal? Is AIG cost-effective?

Since AIG is classified as "targeted", attackers who use it often focus on misleading a specific camera mode identification CNN. Assuming there is an attacker plans to use AIG to modify a number of evidence photos to spoof a forensic CNN to deceive the court. The first question for the attacker is to get the target forensic CNN's prediction access, which may need the attacker to pay money for it (for instance, to bribe someone who has the prediction access). In this case, running time of AIG is also a question - if the attacker wants to finish modifying the evidence photos before the trial.

Güera et al. did not mention too much about running time of AIG, but in real-world scenarios, running time is sometimes an issue. For an attacker, maybe there is no big difference between spending 1 second or 20 second to alter a photo by using AIG; but if he/she needs to spend 1 day to modify one image, that could be a problem. The total cost of using AIG includes money-cost (to get the access to the forensic CNN predictions) and time-cost (to alter images). From Güera et al.'s article, we cannot conclude whether AIG is cost-effective or not.

Another interesting question is: if we have a camera model identification CNN, how can we defend it from AIG attacks? There are many possible answers. We should not leak the predictions of the CNN; we could use another forensic tool to detect the traces of the image that have been modified; we may combine the CNN with other camera model identification tools; or simply using multiple different CNNs working together. AIG was designed to attack one specific CNN by modifying an image, there is no evidence shows the images modified by AIG can spoof many CNNs at the same time.

Conclusion

In conclusion, AIG cannot be widely applied in real-life scenarios, reasons behind the conclusion are vary. Firstly, there is no evidence shows that camera model identification CNNs are widely used. Secondly, the test of AIG only included 10 camera models, which could not cover the camera models in real-world. Thirdly, AIG changes images' visual appearances when altering them, which may cause the modified images being detected by humans or algorithms. Finally, real-life camera model identification CNNs may have different structures to Güera et al.'s test CNN (which cuts input images into 32*32-pixel patches).

Although AIG has these limitations, it is a good start point for people to pay attention to the security of the deep neural network forensic models. Even AIG cannot be widely applied in real-world, it does expose CNN-based camera model identification architectures are vulnerable to some extents.

References

Böhme, R., & Kirchner, M. (2013). Counter-forensics: Attacking image forensics. *Digital image forensics* (2013th ed., pp. 327-366). New York, NY: Springer New York. doi:10.1007/978-1-4614-0757-7_12

Butler W Lampson. (2004). *Computer security in the real world*. New York: The Institute of Electrical and Electronics Engineers, Inc. (IEEE).

Goljan, M., Fridrich, J., & Filler, T. (Feb 5, 2009). Large scale test of sensor fingerprint camera identification. Paper presented at the , *7254*(1) 12. doi:10.1117/12.805701 Retrieved from http://dx.doi.org/10.1117/12.805701

Güera, D., Yu Wang, Bondi, L., Bestagini, P., Tubaro, S., & Delp, E. J. (Jul 2017). (Jul 2017). A counter-forensic method for CNN-based camera model identification. Paper presented at the 1840-1847. doi:10.1109/CVPRW.2017.230 Retrieved from https://ieeexplore.ieee.org/document/8014964

Tuama, A., Comby, F., & Chaumont, M. (2016). (2016). Camera model identification with the use of deep convolutional neural networks. Paper presented at the *2016 IEEE International Workshop on Information Forensics and Security (WIFS),* 1-6. doi:10.1109/WIFS.2016.7823908 Retrieved from https://ieeexplore.ieee.org/document/7823908