Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/53/20241368

Adversarial attack against deep learning algorithms for gun category detection

Fan Lu

Department of Computer Science, Nanjing Technology University, Nanjing, 215000, China

206111102@mail.sit.edu.cn

Abstract. Contemporarily, many deep learning methods have been generated for weapon detection. The weapon detection technology could be used in investigating violent cases. However, the existing gun detection models lack adversarial attack verification for special types of firearms and special picture samples. This study investigates the efficiency of Fast Gradient Sign Method(FGSM) adversarial attack in the field of weapon detection and the influence of weapon category on the attack's result. The dataset is scraped from IMDBF.com and the model being attacked is MobileNetV2, created by HeeebsInc in 2020. As a result, using FGSM methods, adversarial samples generated in film and television graphics containing pistols and rifles can effectively decrease the accuracy of the weapon detection model above. Besides, it is observed the difference of eps needed in attacking different types of gun graphics like film pictures and collection photos. These results verify that some weapon detection models have weak anti-interference, which may provide some ideas for future attacks like BIM or PGD attack.

Keywords: FGSM algorithm, weapon detection, adversarial attacks.

1. Introduction

With the development of various network platforms and the rise of live broadcast and short video fields, manuscript review system is facing an increasing pressure. Combined with deep learning algorithms, real-time detection of network videos, images or live broadcasts is a more efficient review technology, but it also puts forward higher requirements for the accuracy and real-time performance of machine learning algorithms and detection models [1, 2]. In the process of online manuscript review, the detection of gun categories can better help the network platform evaluate the level of manuscripts and provide data support for advanced classification algorithms.

Firearm detection has been a deep learning detection topic. A lot of models and devices have been used for detecting firearms, such as pistols and rifles, in CCTV or network videos. On the whole, the deep learning based methods used in weapon detection can be divided into two-stage and one-stage methods. For two-stage methods, many teams (e.g., González et al. [3], kaya et al. [4], Galab et al. [5]) had used models like Faster R-CNN, VGG-16. For one-stage methods, other teams (e.g., Singh et al. [6], Bhatti et al. [7], Lamas et al. [8]) used models like YOLOv4, YOLOv3 and SSD.

For two-stage methods, González et al. chose Faster R-CNN as their model and applied FPN with ResNet-50 on a database made by themselves, which contained images from CCTV in a campus. The acc score of the architecture reached 88.12% [1]. However, their dataset was unable to be used on

DOI: 10.54254/2755-2721/53/20241368

training or testing programs because of their unsatisfactory evaluation index. Kaya et al. referred to existing models like VGG-16, ResNet-50, ResNet-101 and developed an original model which has seven layers for seven kinds of weapons [1]. The model presented by them could detect many kinds of weapons like knifes, pistols or grenades with 98.40% accurate. But the method was too slow to meet the real-time requirement. Galab et al. managed to develop the model's detection results by increasing the brightness of targets with a specific method [1]. However, their method put a strict limitation on the images, which needed a complicated preprocess.

While for one-stage methods, Singh et al. chose YOLOv4 as their model and created an approach which was computer vision-based to develop their detection result. The model employed showed a satisfactory mean Average Precision but is the model able to meet real-time detection requirements has not been identified. Bhatti et al. put Sliding window technology and region proposal/object detection method into use when developing their model [1]. Among all the models they used, YOLOv4 showed the highest F1_score and mean average accuracy while there are also a lot of false negatives. Lamas et al. developed a weapon detection technology which attached importance to the posture of the figures who hold firearms in a specific scenario [1]. The deep learning architecture trained on their dataset could reach a precision score of 94%, while their model could only be used to detect human-handled firearms.

To conclusion, all the methods and technologies above have their own highlights and application scenarios. The deep learning based methods for weapon detection have reached a satisfactory achievement, both in their diversity and detection results. However, the existing gun detection models lack adversarial attacks verification for special weapon models and special picture samples. The aim of the paper is studying the efficiency of deep learning attack algorithm in the field of firearm category detection and the influence of weapon category on attack. Some special image samples, such as photographs containing guns which are taken from a special angle, may be good subjects of attacks and give ideas to the follow attacks.

2. Data and method

The pictures used for this study is scraped from IMFDB.com, which is a large movie weapons database. There are about 5000 pictures containing 1520 negatives, 1520 pistols and 1370 pistols. Most of these pictures are internet film and television stills, which are taken with special angles and holding positions. Besides of these stills, some collection photos are also included in the dataset which are taken from vertical angle with white background. The model can study the features of guns from collection photos while the film stills can improve the universality of the model.

The model used by this study is MobileNetV2. MobileNetV1 is a pipeline-based lightweight neural network built using depth-level separable convolution, with the introduction of two hyperparameters that allow developers to choose the right model based on their application and resource constraints. MobileNetV2 is an improvement over V1. The network module of MobileNetV2 is shown in Fig. 1. The concrete model structure in this study is shown in Fig. 2.

DOI: 10.54254/2755-2721/53/20241368

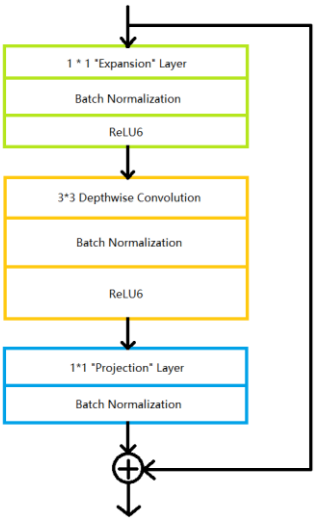

Figure 1. The network module of MobileNetV2 (Photo/Picture credit: Original).

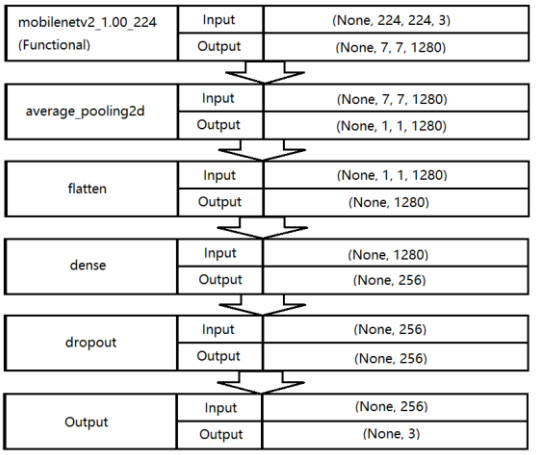

Figure 2. The concrete model structure (Photo/Picture credit: Original).

On the basis of the model structure, the input tensor (None, 224, 224, 3) shows that the input images should be pre-processed to size of 224*224. The output tensor (None, 3) shows the one-hot code of classification results. In the training session, the dataset which is pre-processed is divided into 90% training set and 10% testing set. The random seed is 10, with 50 training epochs and the training was stopped at the 35th epoch because of early stopping. After comparing multiple groups of data, this paper uses the model which is trained without edge and augment pictures to get a better classification result. There are three categories: 0. No-weapon, 1. pistols, 2. rifles. The correspondence of one-hot codes and classification labels are shown in Fig. 3. The loss/acc of the model after 35 epochs of training is shown in Fig. 4.

DOI: 10.54254/2755-2721/53/20241368

[[1. 0. 0.]] ——— no Weapon

[[0. 1. 0.]] ——— Pistols

[[0. 0. 1.]] ——— Rifles

Figure 3. The correspondence of one-hot codes and classification labels

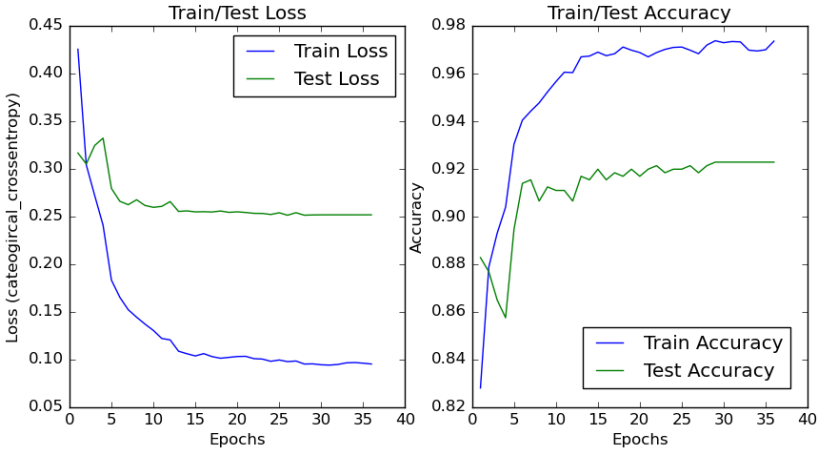

Figure 4. The loss/acc of the model (Photo/Picture credit: Original).

The adversarial attack method used in this paper is Fast Gradient Sign Method (FGSM). FGSM algorithm [9] is explained here. FGSM is one of the white-box attacks, which is presented by Goodfellow, to get adversarial examples quickly [10]. The formula of FGSM principle is expressed in Eq. 1, where x means the original sample, θ is the weight parameter of the model, and y is the true class of x. The adversarial examples are made by inputting the original sample, the weight parameters and the real class, and finding the loss value of the neural network by the J loss function. ∇x means taking the partial derivative of x. Sign() is a symbolic function.

$$ads = \varepsilon sign\left(\nabla_x J(\theta, x, y)\right) \tag{1}$$

The FGSM attack method is fast and simple, with only one step iteration. Implementing FGSM attacks can lay the foundation for future white box attacks like BIM or PGD. Keras has many tools and functions for normal deep learning methods. The most important functions used in this study to implement the FGSM attack are sign() to calculate symbolic function, gradient() to calculate the partial derivative of x and categorical_crossentropy() to calculate loss. The implementation flow is shown in Fig. 5.

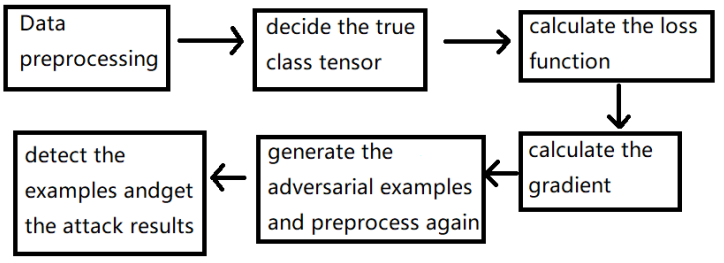

Figure 5. The implementation flow of FGSM in keras (Photo/Picture credit: Original).

DOI: 10.54254/2755-2721/53/20241368

3. Results and discussion

The example shows in Fig. 6 expresses the relation between original examples and adversarial examples. The eps in the sample are settled to 0.003. If the eps is settled to 0.04, the model will detect the adversarial example as 0.noWeapon with 72% prediction. However, the disturbance on the adversarial example will also be obvious because of the white background.

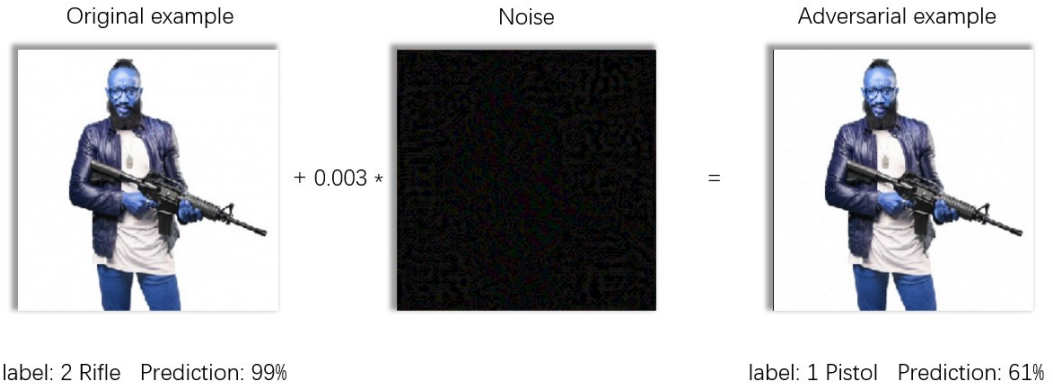

Figure 6. The relation between original examples and adversarial examples (Photo/Picture credit: Original).

Figure 7. Samples from film stills and collection photos (Photo/Picture credit: Original).

According to the detection results, the model showed the highest accuracy of 90.23% when detecting images, which is cropped and pre-processed to special size and resolution ratio, containing pistols. So, the images with pistols are good attacking targets which can show efficiency of the attacking method. The original images are divided into two groups: A group with 20 collection photos of pistols while B group with 20 film stills with pistols. The samples taken from film stills and collection photos are shown in Fig. 7. The two groups, A and B, only contain pistols. The attacking result on A group is shown in Fig. 8. The eps are increased from 0 to 0.35 with 0.05 each step while the acc varies from 1 to 0.05. The attacking result on B group is shown in figure 9. The eps are increased from 0 to 0.0045 with 0.0005 each step while the acc varies from 0.85 to 0.05.

DOI: 10.54254/2755-2721/53/20241368

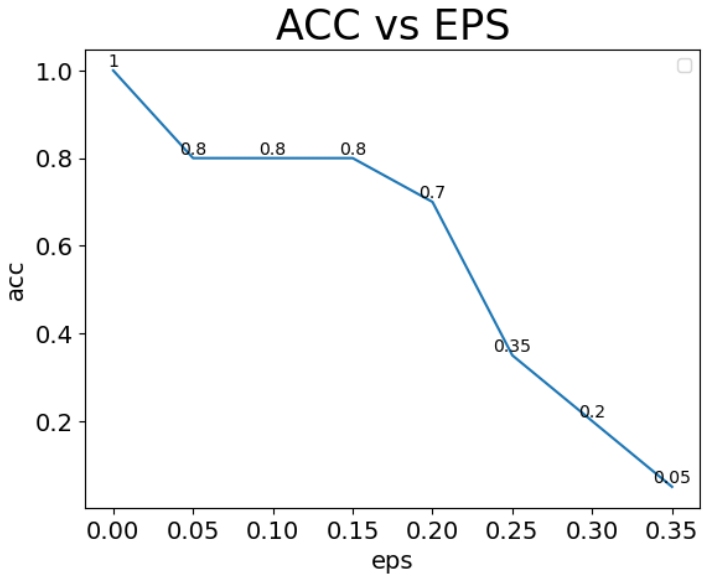

Figure 8. The attack result on group A (Photo/Picture credit: Original).

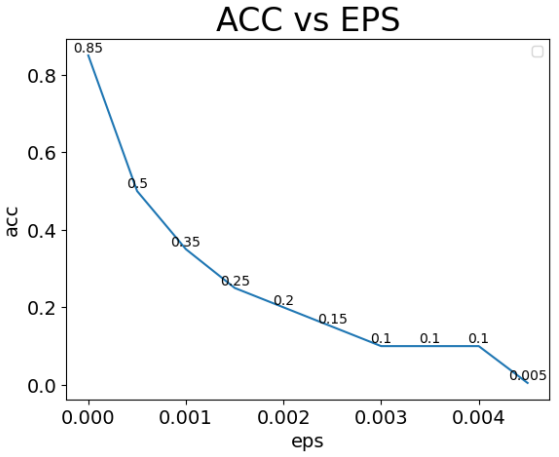

Figure 9. The attack result on group B (Photo/Picture credit: Original).

Comparing the attacking results of group A and B, the eps needed in attacking collection photos is much higher than that in attacking film stills. The collection photos, which are taken with vertical angles and white background, are easier for the model to detect. As a result, it will be much difficult for FGSM to attack these samples, which needs higher eps. For film stills, the attacking method can get the same final attacking result as group A with eps of about 0.004, which means the disturbance on adversarial examples is unable to be noticed by human. Given that most original examples in real situation are pictures of guns with specific backgrounds and holding positions, FGSM can be a viable method in attacking weapon detection models. Besides of the results above, the author found that the shooting angle of guns photos is an important factor which can significantly decrease the accuracy of weapon detection models. The reason of this phenomenon is likely to be the specific shape of guns. Overall, most guns are long and compressed in their shapes, which means guns in photos taken from different angles show great difference. It is much easier for FGSM to attack the samples with a large shooting angle, like a photo with a pistol which is taken from the front of the gun's barrel.

DOI: 10.54254/2755-2721/53/20241368

4. Conclusion

According to the experiment and analysis above, FGSM is a useful adversarial attacking method in attacking weapon detection models, which shows the necessity of increasing weapon detection models' anti-interference capability. In addition, the collection photos of guns are more difficult to attack than normal images, which shows a difficult target for the future attacks. However, the limitations of this study are also significant. First, the number of experimental samples is too small because of the poor performance device. More samples can be tested on a batch basis in future to get a better result. Second, more attacking methods, e.g., BIM or PGD, should be tested to evaluate the efficiency of different attacking methods. Besides of attacking methods, training models with special adversarial examples may be a good way to develop model's accuracy and robustness against deep learning attacks. The method builders can add adversarial samples, which contain weapons and taken from special angles, to their datasets with right labels. Hence, the builders can decrease their models' sensibility to shooting angles and discuss the images with noises made by attacking methods to develop their robustness.

References

[1] Pavinder Y, Nidhi G and Pawan K S 2023 Expert Systems with Applications vol 212 p 118698.
[2] HeeebsInc, WeaponDetection, October 21, 2020. Retrieved on September 26, 2023. Retrieved from: https://github.com/HeeebsInc/WeaponDetection/tree/master
[3] González J L S, Zaccaro C, Álvarez-García J A, Morillo L M S and Caparrini F S 2020 Neural Networks vol 132 pp 297–308.
[4] Kaya V, Tuncer S and Baran A 2021 Applied Sciences vol 11(16) p 7535.
[5] Galab M K, Taha A and Zayed H H 2021 Arabian Journal for Science and Engineering vol 46(4) pp 4049–4058.
[6] Singh A, Anand T, Sharma S and Singh P. 2021 6th International Conference on Communication and Electronics Systems pp. 488–493.
[7] M T Bhatti, M G Khan, M Aslam and M J Fiaz 2021 IEEE Access vol 9 pp 34366–34382.
[8] Lamas A, Tabik S, Montes A C, Pérez-Hernández F, García J and Olmos R 2022 Neurocomputing vol 489 pp 488–503.
[9] Goodfellow J, Shlens J and Szegedy C 2014 arXiv preprint arXiv:1412.6572.
[10] Xu J, Cai Z and Shen W 2019 IEEE 2nd International Conference on Electronics and Communication Engineering (ICECE), Xi'an, China pp 20-25.