Efficient Warm Restart Adversarial Attack for Object Detection

Ye Liu
College of Computer Science and Engineering, Chongqing University of Technology
Chongqing, China
liuye_ly94@163.com

Xiaofei Zhu
College of Computer Science and Engineering, Chongqing University of Technology
Chongqing, China
zxf@cqut.edu.cn

Xianying Huang[*]
College of Computer Science and Engineering, Chongqing University of Technology
Chongqing, China
hxy@cqut.edu.cn

ABSTRACT

This article introduces the solution of the champion team ***green hand*** for CIKM2020 Analyticup: Alibaba-Tsinghua Adversarial Challenge on Object Detection. In this work, we propose a new adversarial attack method called Efficient Warm Restart Adversarial Attack for Object Detection. It consists of three modules: 1) Efficient Warm Restart Adversarial Attack, which is designed to select proper top-k pixels ; 2) Connecting Top-k pixels with Lines, which specifies the strategy on how to connect two top-k pixels to reduce the patch number and minimize the number of changed pixels; 3) Adaptive Black Box Optimization, which is used to achieve a better performance of the black box adversarial attack by adjusting only the white box models. The final results show that our model, which only uses two white box models (i.e., YOLOv4 and Faster-RCNN), achieves an evaluation score of 3761 in this competition, which ranks first among all 1,701 teams. Our code will be available at https://github.com/liuye6666/EWR-PGD.

CCS CONCEPTS

• **Computing methodologies → Object detection**;

KEYWORDS

adversarial attacks, neural networks, object detection

1 INTRODUCTION

Deep neural networks have achieved great success in object detection [6–8]. However, recent studies have shown that deep neural networks are vulnerable to attacks from adversarial examples [1, 5, 10]. In order to identify the fragility of the object detection models and better evaluate the model's adversarial robustness, Alibaba and Tsinghua organize the CIKM2020 AnalytiCup Challenge, i.e., Alibaba-Tsinghua Adversarial Challenge on Object Detection. The competition uses the MSCOCO dataset[1], and expects that participants can make the models unable to detect objects while adding fewer adversarial patches.

To make the challenge more competitive, the challenge organizer add two Constraints:

- **Constraint 1**: Maximum Changed Pixel Rate Constraint, which limits the changed pixel rate less than 2% of all image pixels.
- **Constraint 2**: Patch Number Constraint, which requires the number of patches no more than 10.

Existing adversarial attack methods, such as FGSM [3], PGD [4], MultiTargeted-PGD [2], ODI-PGD [9], add adversarial perturbations to the whole image. The shortcomings of these approaches are: (1) Due to the Constraint 1, adding adversarial perturbations to the whole image is not allowed. (2) All these adversarial attack methods are mainly designed in the image classification scenario. As there is a considerable difference between object detection and image classification, directly applying above methods in the object detection scenario would lead to sub-optimal results. (3) These methods do not control the number of adversarial patches, thus it couldn't satisfy the Constraint 2.

To address the above-mentioned problems of existing approaches, in this work, we propose a novel approach, named Efficient Warm Restart Adversarial Attack for Object Detection. It consists of three modules: (1) Efficient Warm Restart Adversarial Attack (EWR), which performs multiple warm restarts during the process of generating adversarial examples and selects the most important top-k pixels based on the gradient value for each warm restart. (2) Connect Top-k pixels with Lines (CTL), which connects these important pixels together with lines to ensure less pixels are modified and patch number satisfies the Constraint 2. (3) Adaptive Black Box Optimization method (ABBO), which attempts to adjust the white box models to implicitly affect the performance of the black box adversarial attack.

The main contribution of this work is summarized as follows:

1) We propose a novel approach which can effectively handle the limitation of existing adversarial attack methods, and satisfy the two constraints given by the challenge.
2) Our method achieves the best performance among all 1,701 teams with utilizing only two white box models, i.e., YOLOv4 and Faster-RCNN.

2 OUR APPROACH

In order to solve the problem given in this competition, we propose a novel method, which contains three modules: (1) Efficient Warm Restart Adversarial Attack; (2) Connecting Top-k Pixels with Lines, and (3) Adaptive Black Box Optimization.

[*] Corresponding author
[1] https://cocodataset.org/

In: Dimitar Dimitrov, Xiaofei Zhu (eds.): Proceedings of the CIKM AnalytiCup 2020, 22 October, 2020, Gawlay (Virtual Event), Ireland, 2020, published at http://ceur-ws.org.

2.1 Efficient Warm Restart Adversarial Attack (EWR)

In an image, there are multiple objects which can be detected. Based on our preliminary analysis, we find that the loss usually don't change in parallel. In particular, in the beginning, some objects will change their corresponding loss considerably, while the loss change of the remaining objects is small. After that, these objects with less loss change in the beginning will change their corresponding loss greatly. If we select top-k pixel only in the beginning stage, then the selected top-k pixels will be biased towards these objects with a high early loss change. It will inevitably result in selecting improper important pixels.

Inspired by the work of PGD[4] and I-FGSM[3], we design a novel module named Efficient Warm Restart Adversarial Attack. In the first few restarts, modifying the selected top-k pixels will increase the loss of some objects. As the number of restarts increases, more important pixels are selected, so that the loss of the remaining objects will increase significantly. This method can effectively solve the problem of selecting improper important pixels as we mentioned before.

Therefore, for a given original image, we use multiple warm restarts. For each warm restart, we start from the result of last warm restart, and then feed previous restarted adversarial examples into the YOLOv4 and Faster-RCNN model. Then we compute the loss, and obtain the gradient value of the input image through back propagation. At last, we select the pixel points according to the new gradient values, and modify these pixels in the direction of loss raising. When the number of restarts reach a specified threshold (e.g., 10) , or the evaluation score of the subsequent restarts doesn't increase. Finally, we obtain the best adversarial example with the highest score.

2.2 Connecting Top-k pixels with Lines (CTL)

In order to satisfy the condition 2, we need to connect the important top-k pixels together to reduce the patch number. In this work, we propose a simple while effective method, called Connecting Top-k pixels with Lines, to make the number changed pixels as small as possible.

Specifically, we iteratively connect two top-k pixels to reduce the patch number and minimize the number of changed pixels. First, we randomly select a pixel from all top-k pixels and connect it to its nearest pixel in the remaining top-k pixels by using a Line. It is worth noting that a line will involve minimum changed pixels, this step can minimize the changed number of pixels. Then we ignore the selected pixel, and run the above process again in the remaining set of pixels. We will conduct this two steps iteratively until all important pixels are in the same connected sets.

2.3 Adaptive Black Box Optimization (ABBO)

For adversarial attack, the black box models are much harder as compared with the white box models. Since in our work we only make use of two white box models for adversarial attack, we will improve our model to achieve a better performance over the black box adversarial attack. In particular, we will adaptively adjust the strategy of connecting top-k pixels as well as the parameter k of top-k. For an image with a small number of changes pixels for the white box models, it will be difficult for the black box attack. Thus, we first select a small k for top-k pixels. Then we restrict the number of changed pixels between two top-k pixel . Inversely, when an image has a large number of changed pixels for the white box models, we will select a bigger k for top-k pixels. In particular, we conduct as follows:

1) When white box score > 3.3, which means a small number of changed pixels, we set k=10 for top-k, and don't connect two top-k pixels if the number of changed pixels between them are more than 100.
2) When white box score is between 3 and 3.3, which means a medium number of changed pixels, we set k=20 for top-k, and don't connect two top-k pixels if the number of changed pixels between them are more than 150.
3) When white box score is < 3, which means a larger number of changed pixels, we set k=35 for top-k, and don't connect two top-k pixels if the number of changed pixels between them are more than 500.

2.4 Loss Function

In our the EWR module, the loss function directly affects the position of selecting important top-k pixels. Since the goal of this competition is to make the model unable to identify the bounding boxes, we only need to consider the loss related to the confidence of bounding boxes. In order to make the confidence of all bounding boxes small than a given threshold, we set different weights for different confidence intervals. Specifically, for bounding boxes with higher confidence, we set a larger weight in order to make it drop faster.

In YOLOv4 model, we set 4 confidence intervals, and set different weights for different confidence intervals as follows:

$$Loss_{YOLO} = \begin{cases} -0.01 \times conf & \text{if } conf \leq 0.2 \\ -0.1 \times conf & \text{if } 0.2 < conf \leq 0.3 \\ -1 \times conf & \text{if } 0.3 < conf \leq 0.4 \\ -10 \times conf & \text{if } 0.4 < conf \leq 0.5 \end{cases}$$

where $conf$ represents the confidence of the detection bounding boxes.

In Faster-RCNN model, since the confidence threshold of the boxes is 0.3, which is smaller than that in YOLOv4 (In YOLOv4, the confidence threshold of the detection bounding boxes is 0.5), we simply modify the loss function of Faster-RCNN as follow:

$$Loss_{RCNN} = \begin{cases} -0.01 \times conf & \text{if } conf \leq 0.1 \\ -0.1 \times conf & \text{if } 0.1 < conf \leq 0.15 \\ -1 \times conf & \text{if } 0.15 < conf \leq 0.2 \\ -10 \times conf & \text{if } 0.2 < conf \leq 0.3 \end{cases}$$

Finally, for the overall loss function, we combine the loss function of YOLOv4 and Faster-RCNN by simply adding both of them:

$$Loss_{all} = Loss_{YOLO} + Loss_{Faster-RCNN} \tag{1}$$

3 EXPERIMENTS

Dataset: This competition selected about 1,000 images from test split of MSCOCO 2017 dataset. Each image has been resized to 500× 500.

Model: we use only the two white box models, i.e., YOLOv4 and Faster-RCNN.

Evaluation Metrics: The goal of the adversarial attack is to make all bounding boxes invisible by adding the adversarial patches to images. Thus we will adopt the following metric for evaluation:

$$\begin{aligned} S(x, x^*, m_i) = &\left(1 - \frac{min(F(x; m_i), F(x^*; m_i))}{F(x; m_i)}\right) \\ &\times \left(2 - \frac{\sum_k R_k}{5000}\right) \end{aligned} \tag{2}$$

where R_k is the k-th patch's area, x is the original image, x^* is the submitted adversarial image, and m_i is the i-th model ($i \in [1, 2, 3, 4]$). $F(x; m_i)$ returns the number of bounding boxes of image x, given by model m_i (a small number of bounding boxes given by the adversarial example indicates a higher score). At last, the final score is the sum of the scores of all images over the 4 models:

$$FinalScore = \sum_{i=1}^{4} \sum_{x} S(x, x^*, m_i) \tag{3}$$

3.1 Results

Table 1 shows the performance of our proposed approach via different combinations of modules.The combination of EWR and CTL achieves evaluation score of 2500+ and 2600+ when attacking YOLOv4 and Faster-RCNN, respectively. When attacking both YOLOv4 and Faster-RCNN, the combination of EWR and CTL will achieve an evaluation score of 3560+. We further combine all three modules (i.e., EWR, CTL and ABBO), we will obtain the highest evaluation score (i.e., 3761+), which ranks first among all 1,701 teams in the challenge of CIKM2020 Analyticup: Alibaba-Tsinghua Adversarial Challenge on Object Detection.

Table 1: Results of Ablation Experiments

Model		Method			Score
YOLO	RCNN	EWR	CTL	ABBO	
√		√	√		2500+
	√	√	√		2600+
√	√	√	√		3560+
√	√	√	√	√	3761+

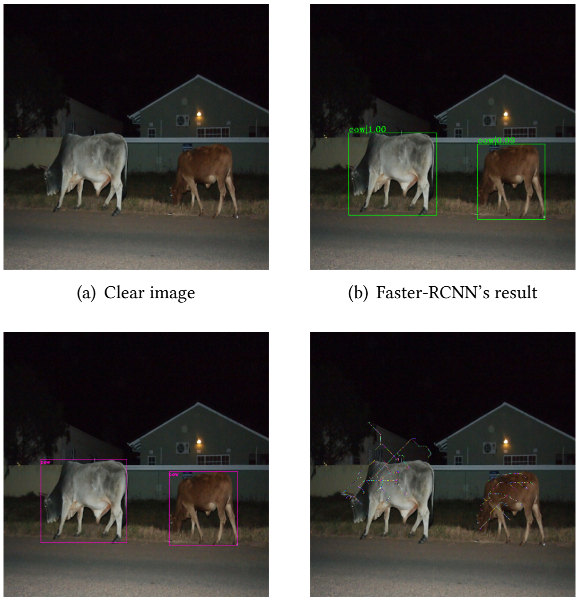
3.2 Case Study

Figure 1 demonstrates the adversarial attack results of an image, where (a) is the original image, (b) is the results of Faster-RCNN model's detection, (c) is the results of YOLOv4 model's detection, and (d) is adversarial example. We can observe that our methods have the following advantages:

- It has a small number of changed pixels.
- Most of the top-k pixels are the key positions of attacked objects.

4 CONCLUSION

In this paper, we proposed an Efficient Warm Restart Adversarial Attack Method for Object Detection, which can modify fewer pixels while maintaining a very high success rate of adversarial attack. Our solution achieves the best performance in all 1701 teams in the challenge of CIKM2020 Analyticup: Alibaba-Tsinghua Adversarial Challenge on Object Detection.



(a) Clear image

(b) Faster-RCNN's result

(c) YOLOv4's result

(d) Adversarial example

Figure 1: result of adversarial attack on the 47.png

REFERENCES

[1] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. 2018. Boosting Adversarial Attacks with Momentum. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9185–9193.

[2] Sven Gowal, Jonathan Uesato, Chongli Qin, Po-Sen Huang, Timothy A. Mann, and Pushmeet Kohli. 2019. An Alternative Surrogate Loss for PGD-based Adversarial Testing. *arXiv preprint arXiv:1910.09338* (2019).

[3] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. 2017. Adversarial examples in the physical world. In *ICLR (Workshop)*.

[4] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *ICLR 2018 : International Conference on Learning Representations 2018*.

[5] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. 2020. Deep Double Descent: Where Bigger Models and More Data Hurt. In *ICLR 2020 : Eighth International Conference on Learning Representations*.

[6] Shaoqing Ren, Kaiming He, Ross Girshick, Xiangyu Zhang, and Jian Sun. 2017. Object Detection Networks on Convolutional Feature Maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 7 (2017), 1476–1481.

[7] Evan Shelhamer, Jonathan Long, and Trevor Darrell. 2017. Fully Convolutional Networks for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 4 (2017), 640–651.

[8] Hirotaka Suzuki and Masato Ito. 2019. Information processing device, information processing method, and program.

[9] Yusuke Tashiro, Yang Song, and Stefano Ermon. 2020. Diversity can be Transferred: Output Diversification for White- and Black-box Attacks. *arXiv: Learning* (2020).

[10] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L. Yuille. 2019. Improving Transferability of Adversarial Examples With Input Diversity. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2730–2739.