

# Combining Attack Success Rate and Detection Rate for effective Universal Adversarial Attacks

Alina Elena Baia, Alfredo Milani, Valentina Poggioni \*

University of Perugia - Italy

**Abstract.** In the framework of Adversarial Machine Learning, several detection and protection techniques are used to characterize specific attack-defense scenarios. In this paper, we present universal, unrestricted black-box adversarial attacks based on a multi-objective nested evolutionary algorithm able to incorporate the detection rate and a measure of image quality into the attack building phase.

## 1 Introduction and Related works

Deep Neural Networks, despite their superior performance, are remarkably vulnerable to adversarial attacks, creating severe security issues [1].

The majority of the proposed attacks are performed and optimized to add small random perturbations to the pixel values. However, these artificial modifications are often not semantically meaningful and, even if limited to a few pixels, can create unnatural-looking images that are easily detectable [1]. For this reason, researchers started exploring new types of threats models that can significantly change an input while maintaining the semantics.

Such methods require either access to the targeted network architecture [2] or additional resources like pretrained networks to perform image segmentation [3], colorization and style transfer [4]. In some cases, it is necessary to train neural networks from scratch for each image in order to find effective adversarial perturbations[5].

In this scenario, several approaches and systems able to detect adversarial attacks have been proposed and developed. Some of them propose ad-hoc trainable techniques like distillation [6], perturbation rectifying [7] or feature squeezing [8] that can be used also in the universal attack scenario.

In this work, we decided to focus on generating untargeted unrestricted universal adversarial attacks in a black-box scenario, since the limited knowledge and ability of the attacker is more similar to a real-world scenario, making the attack itself more challenging but its applicability more practical.

Moreover we decided to implement the attacks by means of image filters application, such as those that can be found in many apps, social networks or modern cameras. We think that building attacks through image filters makes the attacks more sneaky and inherently more difficult to be detected since images are usually filtered with enhancing objectives and not with malicious intentions. Some results about the use of image filters to generate adversarial examples has

---

\*This is an optional funding source acknowledgement.

been reported very recently in [9], but it is limited to consider the effect of just one filter instead of a combination of them.

In this work we propose a gradient-free method based on nested evolutionary algorithms and multi-objective optimization that, given a set of commonly-used image filters, finds an optimal image-agnostic sequence of them, that, when applied to an image, is hardly detectable and causes the classifier to misclassify the image. The standard universal  $L_p$ -bounded attacks are transformed into universal unrestricted attacks and a multi-objective evolutionary approach is used to build a process able to optimize, at the same time, the attack success rates, the detection rate and to limit the modifications applied to the images.

Given the conflicting nature of the above-mentioned objectives and motivated by the success of multi-objective evolutionary algorithms (MOEA) in other applications [10, 11, 12], we propose to model our method as a multi-objective optimization problem.

## 2 Problem Formulation and Backgrounds

In case of *universal* approaches the objective is to find  $a$ , *single*,  $\delta \in \mathbb{R}^d$  able to fool a classifier function  $F : X \rightarrow L$ , for *almost all* the data points available in  $X$ , that is the returned labels  $F(x + \delta) \neq F(x)$ , for almost all  $x$  in the input image set  $X \subseteq \mathbb{R}^d$ .

Differently from the case of *restricted* attacks, where the objective is to find the smallest perturbation  $\delta \in \mathbb{R}^d$  (s.t.  $\|\delta\|_p \leq \epsilon$ ) able to cause the misclassification, we propose to solve a multi-objective optimization problem where we want to maximize the *Attack Success Rate* while the *Detection Rate*, and the *Perturbations Applied* to the image are minimized.

When solving multi-objective optimization problems (MOPs), the aim is to obtain the Pareto optimal set. Thus, given a MOP, the goal of a multi-objective evolutionary algorithm is to produce a good approximation of its Pareto front.

### 2.1 Image filters

We used several Instagram-inspired image filters to perform the attacks and each filter has different characteristics and effects given by distinct levels of saturation, contrast, brightness, etc. Each filter is regularized by two parameters ( $\alpha$  and  $s$ ) that the algorithm has to optimize. The parameter  $\alpha$  controls the intensity of each effect, i.e how much the contrast has to be increased or how much light to add to the image, while  $s$  represents the parameter of the convex interpolation between the original image and the modified one, and it is used to adjust the strength of the filter application.

## 3 Approach and Algorithm

We propose a nested-evolutionary algorithm for generating universal unrestricted adversarial examples in a black-box scenario inspired by [13]. Given a sequence of image filters as input, the algorithm returns the best image-agnostic filter

configuration which, applied to the images from the dataset, maximizes the classification error of the target model while the detection rate and the perturbation applied are minimized.

Our idea is to provide the attack with the ability to bypass a detection mechanisms that could be used to protect the system while the applied perturbation are automatically controlled. We believe this to be a powerful feature of our method given that the field of adversarial machine learning lacks such approaches.

**Method:** The method consists of two evolutionary nested algorithms: the outer algorithm, in charge of finding the sequence of filters to use, and the inner algorithm used to choose the parameter values. Given a set  $S = \{f_1, f_2, \dots, f_m\}$  of  $m$  image filters, the outer algorithm genotype (with length  $l$ ) is encoded as a list of filters while the inner algorithm genotype is represented by a list containing the parameters used for each selected filter. The associated phenotype, applied to a set of images, generates the adversarial examples by applying the selected sequence of filters, with their corresponding optimized parameters, to legitimate images.

**Outer Algorithm:** For the outer optimization step we employ a genetic algorithm: a population of  $N$  candidate solutions is iteratively evolved towards better solutions. In order to breed a new generation, population members are randomly selected and the crossover (one-point crossover) and mutation operations (one filter change) are performed. The quality of the candidates is evaluated based on their fitness values and, at the end of each iteration, the  $N$  best individuals are chosen for the next generation.

**Inner Algorithm:** We propose to optimize the parameters by using  $(1, \lambda)$  evolution strategy with  $\lambda = 5$ . ES iteratively updates a search distribution by following the natural gradient towards higher expected fitness. In our case, for each list of parameters we compute a batch of  $N$  samples by perturbing the original individual. A gradient towards a better solution is estimated using the fitness values of the  $N$  samples. This gradient is then used to update the original individual. The entire process is repeated until a stopping criterion is met.

**Evaluation:** The fitness function is modeled as a multi-objective problem which accounts for the attack success rate as well as the detection rate and an image quality measure. The attack success rate (ASR) is the portion of successful attacks obtained over the dataset used for the evolution process; similarly the detection rate (DR) is the portion of perturbed images that have been recognized as malicious by any detection algorithm or a combination of more detection mechanism; finally the image quality measure is any measure able to assess the likelihood that a human user would detect that the image has been substantially artificially modified.

## 4 Experiments and Discussion

In order to evaluate the presented method we used as target model a convolutional neural network proposed by Papernot et al. in [6] and used also in [14]

to prove the efficiency of their attack, the model was trained on the CIFAR-10 training set. In order to optimize our multi-objective attack, we used 200 images from the CIFAR-10 test set, while the remaining 9800 images were used to generate actual attacks and assess their effectiveness. We intentionally chose a rather small subset for the training-optimization phase to prove the validity and strength of the universal attack.

**Detection Mechanism:** For the experiments, we chose *Feature Squeezing* (FS) as detection strategy [8] in the fitness function of the optimization algorithm given that it is a well-known, attack-independent and computationally low-cost technique that achieves high detection rates against many state-of-the-art attacks with an overall detection rate of 84.5% on CIFAR-10.

**Image Quality Assessment:** The automatic image quality assessment is probably the most critical part. Several preliminary experiments have been made by using standard No-Reference IQA methods like NIQE, BRISQUE and NIMA but the results proved that these measures cannot be applied in our case because there is not a clear correlation between the measures' values and the perturbations generated by the filters.

There we decided to approach the problem as an anomaly detection problem where we consider as anomalous such images that, after the perturbations, deviate from the distribution generated by clean (non-perturbed) images. The objective is to recognize images that have been excessively perturbed. We chose to employ a U-net autoencoder to automatically assess the quality of the perturbed images by means of its reconstruction error. By training the U-net model to minimize the reconstruction error defined by the Mean Squared Error (MSE) on clean images, we can use the MSE to evaluate how much an image was altered: images heavily modified by filters will have a higher reconstruction error than the non-modified images. The goal of the optimization process is to find an optimal filter configuration that results in a small reconstruction error.

**Experiments setup:** Preliminary tests were conducted in order to select the optimal number of epochs for the optimization phase. Experimental results have shown that, in most cases, 6 epochs are enough for the algorithm to find an optimal adversarial configuration.

**Results:** First of all we analyzed the Attack Success Rate and the Detection Rate obtained with sequences of filters of different lengths. We computed the ASR and DR on both training and test set in order to evaluate also the generalization ability of the algorithm. The results are reported in Table 1. We observed that increasing the number of filters corresponds to an improvement of the ASR values, even if this produced, in some cases, a deterioration of the quality of the image. It is interesting to note that, in all the cases, the detection rate is very low. This indicates that our attack is highly effective since less than 5% of the successful adversarial examples have been identified as illegitimate. Furthermore, we obtain a very good generalization ability given that we only lose between 2% and 6% on ASR while still keeping the DR extremely low.

With respect the quality of the images some considerations have to be made. In all the cases images do not present artifacts like the ones produced by other

algorithms [1, 14] that, in general, can be easily detected by the majority of the detection systems. This is clearly an advantage produced by the use of image filters instead of changing single pixels or adding textures [1, 4]. On the other hand, the composition of more filters sometimes can produce dull and sandblasted effects that are not sufficiently recognized by the U-net, as shown in Table 2.

No. of Filters	Training set		Test set	
	ASR %	DR %	ASR %	DR %
5	74.00	4.00	71.87	2.44
6	80.50	1.20	78.59	1.92
7	82.50	2.40	79.51	2.33
8	83.50	1.19	79.18	1.21

Table 1: Attack success rate (ASR) and Detection Rate (DR).

























No. of Filters	Successful Adversarial Examples					
5						
6						
7						
8						
Label names	0:airplane, 1:automobile, 2:bird, 3:cat, 4:deer, 5:dog, 6:frog, 7:horse, 8:ship, 9:truck					

Table 2: Some successful adversarial attacks showing dull and sandblasted effects.

## 5 Conclusions and Future works

In this paper an evolutionary algorithm able to produce effective unrestricted universal attacks generated by the composition of image filters has been proposed. A multi-objective evolutionary approach is used to build a process able to optimize, at the same time, the attack success rate, the detection rate and to limit the modifications applied to the images. The main contributions given by this work can be summarized in: (i) the integration of different objectives

like the optimization of the attack success rate combined with the detection rate and the image quality assessment, (ii) the use of combinations of image filters to produce sneaky and hardly detectable attacks, (iii) a general multi-objective evolutionary approach, which can be applied and extended to different filters, different attacks and different detection methods.

Experimental results proved that the approach is effective in producing universal attacks that are very easy to move by anyone.

Nevertheless there are some open issues that require further investigations: (i) image filters library has to be extended and improved: the feasibility of the approach relies mainly on the ability to produce nice images, (ii) the image quality assessment has to be improved in order to avoid dull and sandblasted effects on the filtered images, (iii) additional attacks detection systems have to be included in the fitness function order to make even more harder the attacks detectability, (iv) the algorithm should be tested also against other computer vision tasks than image classification like object detection and semantic image segmentation.

## References

- [1] X. Yuan, P. He, Q. Zhu, and X. Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE Trans on Neural Networks and Learning Systems*, 30(9):2805–2824, 2019.
- [2] Zhengyu Zhao, Zhuoran Liu, and M. Larson. Adversarial color enhancement: Generating unrestricted adversarial images by optimizing a color filter. *In Proc. of BMVC 2020*.
- [3] Ali Shahin Shamsabadi, Ricardo Sanchez-Matilla, and Andrea Cavallaro. Colorfool: Semantic adversarial colorization. *In Proc of CVPR 2020*.
- [4] Anand Bhattad, Min Jin Chong, Kaizhao Liang, B. Li, and D. Forsyth. Unrestricted adversarial examples via semantic manipulation. *In In Proc. of ICLR 2020*.
- [5] Ali Shahin Shamsabadi, Changjae Oh, and Andrea Cavallaro. Edgefool: an adversarial image enhancement filter. *In Proc of ICASSP 2020*.
- [6] N. Papernot, P. McDaniel, Xi Wu, S. Jha, and A. Swami. Distillation as a defense to adversarial perturbations against deep neural networks. *In Proc of IEEE SP 2016*.
- [7] N. Akhtar, J. Liu, and A. Mian. Defense against universal adversarial perturbations. *In Proc of IEEE CVPR 2018*, pages 3389–3398.
- [8] Weilin Xu, David Evans, and Y. Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *ArXiv*, abs/1704.01155, 2018.
- [9] Zhe Wu, Zuxuan Wu, Bharat Singh, and Larry Davis. Recognizing instagram filtered images with feature de-stylization. *Proc. of AAAI 2020*, 34:12418–12425.
- [10] A. Zhou, B.Y. Qu, H. Li, S.S. Zhao, P.N. Suganthan, and Q. Zhang. Multiobjective evolutionary algorithms: A survey. *Swarm and Evol. Computation*, 1(1):32 – 49, 2011.
- [11] M. Baiocchi, C.A. Coello Coello, G. Di Bari, and V. Poggioni. Multi-objective evolutionary gan. *In Proc. of GECCO 2020*, pages 1824 – 1831, 2020.
- [12] M. Baiocchi and et al. Can differential evolution be an efficient engine to optimize neural networks? *In Proc of MOD 2018*, pages 401–413. Springer Intern. Publ., 2018.
- [13] Alina Elena Baia, Gabriele Di Bari, and Valentina Poggioni. Effective universal unrestricted adversarial attacks using a moe approach. *In In Proc of EvoAPPS 2021*.
- [14] Nicholas Carlini and D. Wagner. Towards evaluating the robustness of neural networks. *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57, 2017.