aci avances en ciencias e ingenierías

Artículo/Article
Vol. 17, nro. 2
e3699

PoI+NBU: A feasibility study in generating high-resolution adversarial images with a black box evolutional algorithm based attack

Enea Mancellari[1]*, Ali Osman Topal[1], Franck Leprévost[1]

[1] University of Luxembourg, Faculty of Science, Technology and Medicine, Computer Science Department, Esch-sur-Alzette, Luxembourg

*Autor para correspondencia/Corresponding author: enea.mancellari@uni.lu

PoI+NBU: Un estudio de viabilidad en la generación de imágenes adversariales de alta resolución con un ataque basado en algoritmos evolutivos de caja negra

Abstract

Adversarial attacks in the digital image domain pose significant challenges to the robustness of machine learning models. Trained convolutional neural networks (CNNs) are among the leading tools used for the automatic classification of images. They are nevertheless exposed to attacks: given an input clean image classified by a CNN in a category, carefully designed adversarial images may lead CNNs to erroneous classifications, although humans would still classify "correctly" the constructed adversarial images in the same category as the input image. In this feasibility study, we propose a novel approach to enhance adversarial attacks by incorporating a pixel of interest detection mechanism. Our method involves utilizing the BagNet model to identify the most relevant pixels, allowing the attack to focus exclusively on these pixels and thereby speeding up the process of adversarial attack generation. These attacks are executed in the low-resolution domain, and then the Noise Blowing-Up (NBU) strategy transforms the low-resolution adversarial images into high-resolution adversarial images. The PoI+NBU strategy is tested on an evolutionary-based black-box targeted attack against MobileNet trained on ImageNet using 100 clean images. We observed that this approach increased the speed of the attack by approximately 65%.

Keywords: Black-box attack, Convolutional Neural Network, High resolution adversarial image, Noise Blowing-Up method, Pixels of Interest.

Resumen

Los ataques adversariales en el dominio de las imágenes digitales plantean desafíos significativos para la robustez de los modelos de aprendizaje automático. Las redes neuronales convolucionales (CNNs) entrenadas están entre las herramientas principales utilizadas para la clasificación automática de imágenes. Sin embargo, están expuestas a ataques: dada una imagen limpia de entrada clasificada por una CNN en una categoría, las imágenes adversariales diseñadas cuidadosamente pueden llevar a las CNNs a clasificaciones erróneas, aunque los humanos seguirían clasificando "correctamente" las imágenes adversariales construidas en la misma categoría que la imagen de entrada. En

Editado por / Edited by:
Dennis Cazar

Recibido / Received:
19/11/2024

Aceptado / Accepted:
24/12/2024

Publicado en línea / Published online:
21/08/2025

 DOI: https://doi.org/10.18272/aci.vi.3699

este estudio de viabilidad, proponemos un enfoque novedoso para mejorar los ataques adversariales mediante la incorporación de un mecanismo de detección de píxeles de interés. Nuestro método implica el uso del modelo BagNet para identificar los píxeles más relevantes, lo que permite que el ataque se enfoque exclusivamente en estos píxeles y, de esta manera, acelere el proceso de generación de ataques adversariales. Estos ataques se ejecutan en el dominio de baja resolución y, luego, la estrategia de Ampliación de Ruido (Noise Blowing-Up, NBU) transforma las imágenes adversariales de baja resolución en imágenes adversariales de alta resolución. La estrategia PoI+NBU se prueba en un ataque dirigido de caja negra basado en evolución contra MobileNet entrenado en ImageNet, utilizando 100 imágenes limpias. Observamos que este enfoque aumentó la velocidad del ataque en aproximadamente un 65%.

Palabras clave: Ataque de caja negra, Red Neuronal Convolucional, Imagen adversarial de alta resolución, Método de Ampliación de Ruido, Píxeles de Interés

INTRODUCTION

Convolutional neural networks (CNNs) have become indispensable in the field of computer vision, showcasing exceptional performance across various tasks, particularly in image classification [1, 2, 3]. By leveraging the power of convolutional layers for feature extraction, CNNs excel in identifying intricate patterns and subtleties within visual data. CNN's classifications are represented by output vectors of length equal to the number of categories the CNN is designed to sort images into (e.g., 1000 for those trained on ImageNet [4]). For each category *c*, the CNN computes a *c*-label value $\in$ [0, 1] that measures the likelihood that the image belongs to *c*.

Recently, the vulnerability of CNNs to adversarial attacks has become a topic of significant interest. Attacks involve finding perturbations in input data, often with imperceptible changes to human observers, that lead to misclassification by the model. These vulnerabilities pose significant safety concerns in real-world applications such as self-driving cars, surveillance of sensitive areas, medical diagnoses, etc. However, they can also be exploited to obscure security and privacy-sensitive information from CNN-based threat models aimed at extracting such data from images [5, 6].

In particular, images used on social media are usually high-resolution large size images (they belong to the so-called HR domain). Leprévost et al. [7, 8], detailed the generic *Noise Blowing-Up strategy* (NBU) for generating high-resolution (HR) adversarial images against CNNs. Additionally, the authors presented in [9] the generic *zone-of-interest strategy* (ZoI) that originally *a priori* works in the low-resolution (LR) domain.

Our contribution (Subsection 1.1)

The present article, on the one hand, addresses issues remained open in [9], in particular an experimental validation, and, on the other hand, provides the design of a new generic attack that combines the Pixels of Interest (PoI) strategy with the Noise Blowing Up (NBU) method. The resulting PoI+NBU method aims at enhancing the effectiveness

 DOI: https://doi.org/10.18272/aci.vi.3699

of any type of attack (white-box or black- box) and of any specific attack on CNNs at the creation of HR adversarial images of exceptional visual quality.

This combination works as follows in practice. A clean high-resolution image is reduced to the LR domain to fit the input size of a CNN to attack. The PoI strategy is applied in the LR domain to identify the most relevant areas of an image for its classification by the considered CNN. Then an attack is performed, focusing on these zones, thereby reducing its search space and enhancing its efficiency. The adversarial noise, created that way in limited zones in the LR domain, is blown-up to the HR domain. This noise is then added to the HR clean image, leading to a high-resolution adversarial image, indistinguishable from the original HR clean image for a human eye.

We validate the combined PoI+NBU approach experimentally. Specifically, we employ a variant of the evolutionary algorithm-based (EA) attack described in [10] on 100 high resolution (HR) clean images, targeting the MobileNet CNN [11] trained on ImageNet.

Organisation of the paper (Subsection 1.2)

Section 2 outlines the key theoretical steps of the PoI+NBU strategy. Section 3 lists the targeted CNN, the HR clean images, and the essential features of the EA-based targeted attack used in the experiments. Section 4 presents the outcome of the experiments, including a visual assessment of the quality of the adversarial images obtained through some illustrative images. The Conclusions section summarizes the findings of this paper.

The algorithms and experiments were implemented in Python 3.9 utilizing the NumPy 1.23.5, TensorFlow 2.14.0, Keras 3, and Scikit 0.22 libraries. Computational tasks were executed on nodes equipped with Nvidia Tesla V100 GPUs within the IRIS HPC Cluster at the University of Luxembourg [12]. Additional material (clean images used, their size, example of adversarial images, and source code) can be retrieved at https://github.com/emancellari/PoI_NBU.git

COMBINING THE PIXELS OF INTEREST AND THE NOISE BLOWING-UP STRATEGIES (SECTION 2)

This section provides a rapid overview of the typology of attacks and of attack scenarios (Subsection 2.1). Then it describes the PoI generic strategy (Subsection 2.2) and the NBU strategy (Subsection 2.3). Finally, it gives the overall scheme of the combined PoI+NBU generic strategy (Subsection 2.4).

Typology of attacks and visual expectations (Subsection 2.1)

Attacks are classified according to the level of knowledge an attacker has about the CNN to deceive. In white-box attacks [13, 14, 15], the attacker has complete knowledge of the target CNN's architecture, parameters, and training data, allowing for precise creation of adversarial images, often with high success rates. In contrast, black-box attacks [10, 16, 17, 18] rely only on observing the input-output behavior of the target model, typically requiring more time and resources.

DOI: https://doi.org/10.18272/aci.vi.3699

Attack scenarios are manifold. Given a clean image classified by the CNN in a category c_a, in the target scenario, one selects a category $c_t \neq c_a$, and one adds adversarial noise to the clean image to create an adversarial image classified by the CNN in c_t. As such, one has defined a *good enough* adversarial image. A τ -strong adversarial image (for $0 < \tau \leq 1$) is an adversarial image classified in c_t with a c_t-label value $\geq \tau$. In the untargeted scenario, the process is similar as in the target scenario, except that one requires the adversarial image to be classified in any category $c \neq c_a$.

Finally, adversarial images can be indistinguishable for a human as compared to the associated clean images, or not. The former requirement is clearly much more challenging than the latter one.

Pixels of Interest (PoI) strategy (Subsection 2.2)

Figure 1 describes the PoI process in the LR domain. One is given a CNN C to deceive, and a clean image A, of size equal to the input size of C (say 224 x 224 if C is trained on ImageNet), classified by C as belonging to the category c_a with c_a-label value equal to τ_a.

One uses BagNet [19] to identify the pixels relevant for a CNN's classification of the image in c_a (b) and in c_t (d), thanks to a heatmap. Note that one does not specify which CNN we are dealing with, so that making use of BagNet is compliant with the requirements set by black-box attacks. Then we sieve these pixels and keep only the x% most significant for c_a on the one hand and for c_t on the other hand, where x is fixed at will ((c) and (e)). One merges this information (without redundancy) in (f). The attack is performed on these pixels of interest, leading to an adversarial image classified by C in the target category c_t with a c_t-label value equal to $_t$.

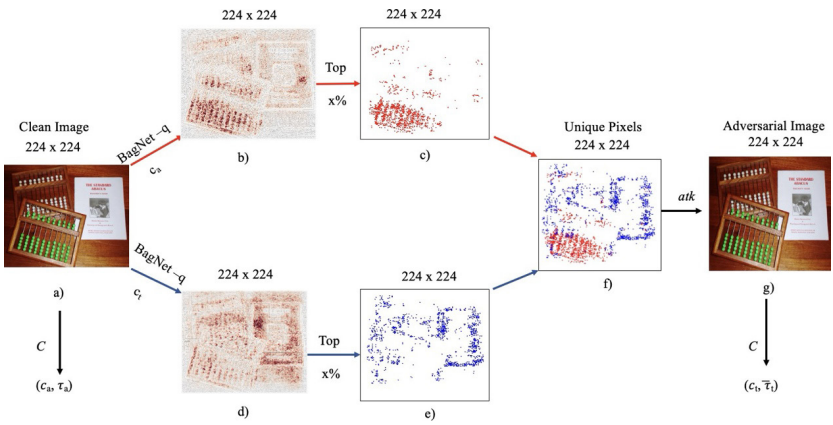

FIGURE 1. PoI process in the LR domain for any attack, any scenario, and any CNN.

Remarks. Firstly, one could use clustering techniques like DBSCAN [20] to encapsulate these top x% most relevant pixels into larger zones of interest before performing the attack. Doing so presents the advantage of a lesser concentration of the attack on individual

 DOI: https://doi.org/10.18272/aci.vi.3699

pixels, what may lead to a better visual quality. However, it does not prove true in practice (essentially because an observer notices rectangles on the adversarial images obtained). Moreover, it often implies that very large proportions of the image are subject of the attack, even if one uses only the top 1% most relevant pixels: our experiments showed that one jumps from 4.70% of the image without DBSCAN (see Table 2 and Figure 3 in additional material file), to 65% with DBSCAN, thereby adding a very large proportion of less-relevant pixels to the attack, leading to a slowing down of the process and lesser success rates. In other words, clustering techniques are unlikely to provide any substantial advantage.

Secondly, BagNet acts as a proxy of the CNN to attack but does not substitute it. Therefore, the usage of BagNet is compatible with a black-box attack scheme.

Thirdly, one can see our PoI strategy as a generalisation of the attacks [21, 22], where one or a few pixels are modified to create adversarial images. However, our aim goes beyond, since, as opposed to the aforementioned attacks where a human immediately sees that an attack occurred, we intend to create adversarial images indistinguishable from the original clean image.

Noise Blowing-Up (NBU) strategy (Subsection 2.3)

In a nutshell, in the Noise Blowing-Up (NBU) generic strategy [8] illustrated in Figure 2, a clean HR image is reduced with a resizing interpolation function to fit the CNN *C*'s input size. c_a denotes the category in which *C* classifies this resized clean image. Then an attack *atk* is performed in the LR domain on this image to create an adversarial image classified in $c \neq c_a$ (which may be a predefined category c_t in the target scenario). The adversarial noise is extracted in the LR domain and then blown-up to the HR domain to fit the original clean image size. This blown-up noise is then added to the HR clean image, leading to a HR tentative adversarial image. This image is again processed to fit *C*'s input size. If *C* classifies it in *c*, one has obtained that way a HR adversarial image.

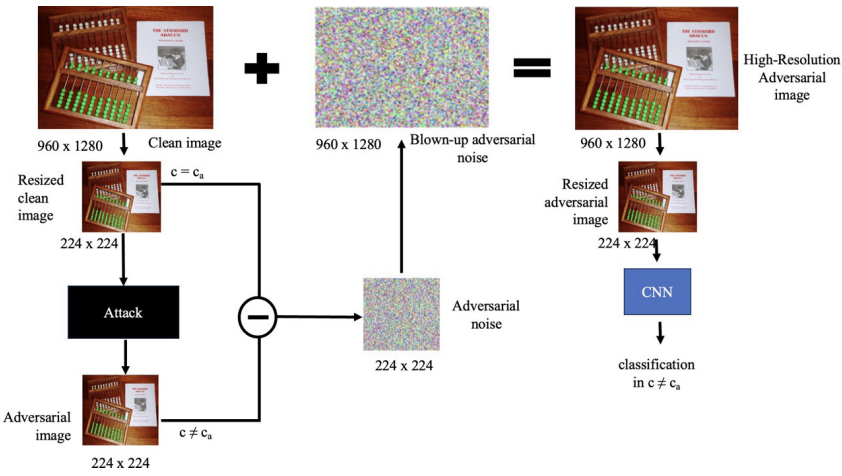

FIGURE 2. The Noise Blowing-Up strategy

DOI: https://doi.org/10.18272/aci.vi.3699

PoI+NBU strategy (Subsection 2.4)

The PoI+NBU method illustrated in Figure 3 integrates the PoI strategy with the NBU method to create high-resolution adversarial images effectively. The PoI strategy initially identifies the relevant regions of the resized clean image in the LR domain on which the attack will occur. Once the attack is applied within these zones, the NBU strategy is used to blow up the obtained adversarial noise to the HR domain, and the process continues as in Subsection 2.3.

A key advantage of this approach combining two generic strategies is that the result is again a generic strategy: It applies *a priori* to any attack, any scenario, and any CNN, and it is still a black-box attack.

For attacks that incorporate randomness, such as evolutionary-based attacks, rather than relying on a single substantial attack round which would create a very strong adversarial noise at once, one could also consider performing multiple rounds of moderate attacks, each leading to the creation of moderate noise [9], where for instance each round

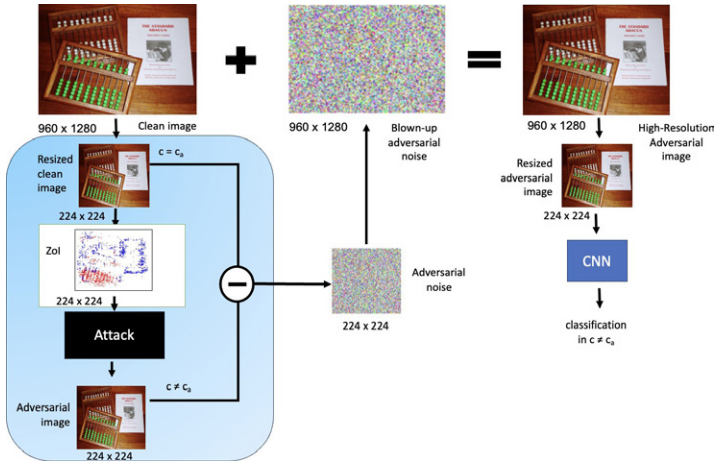

FIGURE 3. The combined PoI+NBU strategy

generates focused adversarial noise within some particular zone of interest. Although none of them would be enough to create a HR adversarial noise, their collaborative efforts may. The successive layers of moderate noise, blown-up and carefully combined, may progressively generate an adversarial image in the HR domain.

FRAMEWORK OF THE EXPERIMENTAL VALIDATION (SECTION 3)

We exposed the PoI strategy on the one hand (working in the LR domain), and the PoI+NBU generic strategy on the other hand (working in the HR domain) to a series of experiments. We specify here the attack scenario and the specific HR images used in the tests (Subsection 3.1), the concrete attack considered (Subsection 3.2), and the CNN to deceive in this feasibility study (Subsection 3.3).

DOI: https://doi.org/10.18272/aci.vi.3699

There are essentially three BagNet models that one can use in the PoI part of the combined strategy, namely BagNet-q with q = 9, 17, 33. We selected q = 33 due to its accuracy and runtime performance, reported in [19].

Regarding Subsection 3.2, let us stress that we were unable to test the strategy against other attacks such as FGSM [23], PGDInf [24], BIM [25], SimBA [26], and AdvGAN [27]. This limitation is due to the lack of full access to the code of these attacks. Note as well that most processes involved can be parallelized, but we did not explore it in the present study.

The attack scenario and the HR clean images (Subsection 3.1)

The experimentation is performed for the target scenario for the 10 pairs (c_a, c_t) of clean-target categories specified in Table 1 (same to those utilized in [10, 28])

TABLE 1. For $1 \leq p \leq 10$, the 2nd row gives the ancestor category c_a and its index number a_p among the categories of ImageNet (Mutatis mutandis for the target categories, 3rd row).

p	1	2	3	4	5	6	7	8	9	10
c_{ap}	abacus	acorn	baseball	broom	brown bear	canoe	hippo	llama	maraca	mountain bike
a_p	398	988	429	462	294	472	344	355	641	671
c_{tp}	bannister	rhinoceros beetle	ladle	dingo	pirate	Saluki	trifle	agama	conch	strainer
t_p	421	306	618	273	724	176	927	42	112	828

For each ancestor category c_a, we picked at random 10 clean ancestor images from the ImageNet validation scheme in c_a, provided that their sizes $h \times w$ satisfy $h \geq 224$ and $w \geq 224$. This ensures that these 100 clean images belong to the HR domain. Additional material (see end of Introduction) contains these images and their original sizes.

Experiments (Subsection 3.2)

Once the pixels of interest are identified, one performs the black-box Evolutionary-based Algorithm (EA) attack [10] (see Algorithm 1 for its pseudo-code) within these regions, while keeping the rest of the pixels untouched.

The attack is executed for the target scenario to create 0.55-strong adversarial images (this ensures a convincing margin ≥ 0.10 with respect to the second best category). The maximum number of generations is set to N = 10, 000, and the population size is set to 40.

ϵ controls the maximum allowable change in pixel values for the entire image, while α determines the magnitude of change for each pixel in each generation of the EA. These parameters play a crucial role in shaping the nature and magnitude of adversarial perturbations generated by the algorithm. Throughout the experiments, the value of α per generation is fixed at 1/255.

DOI: https://doi.org/10.18272/aci.vi.3699

The EA is initially executed without PoI, with $\epsilon = 8$, 12, and 16. Subsequently, it runs with PoI applied to increasing percentages of relevant pixels: the top x% (x: 10, 20, 25, 30, and 35) of both c_a-label and c_t-label values (taken together without any duplication) as measured by BagNet-33. This leads to 1800 attempts to generate adversarial images within the LR domain. Experience shows that the EA is unable to generate a significant number of adversarial images if $x < 10$, since in this case the proportion of the image affected is too narrow. Therefore, the study considers $x \geq 10$.

Algorithm 1 EA attack pseudocode [10, 18]

1: **Input**: CNN *C*, initial image *A*, perturbation magnitude α, max perturbation ϵ, ancestor class c_a, target class index *t*, current generation *g*, max generations N
2: Initialize population: 40 copies of *A*; I_0 as the first individual
3: Compute fitness for all individuals
4: **while** $(OI_0[t] < \tau)$ & $g < N$ do
5: Rank individuals by fitness: top 10 as elite, next 20 as middle class, last 10 as lower class
6: Mutate a random number of pixels in middle and lower class individuals withα; clip mutations to $[-\epsilon, \epsilon]$
7: Replace lower class with mutated elite and middle class individuals
8: Cross-over individuals to form new population
9: Compute fitness for all individuals

CNN: MobileNet (Subsection 3.3)

The feasibility study is performed using MobileNet [11] trained on ImageNet [4]. We selected this CNN because it is optimized (and favored over other CNNs) for applications running on devices with limited processing power, memory, and storage capacity [29]. Examples of recent applications of MobileNet include the classification of freshwater fish on smartphones for farmers [30], the identification of tomato leaf disease in agriculture [31], the detection of skin cancer [32], etc.

Table 2 presents a comparison between MobileNet, the original GoogleNet [33] and VGG16 [34] in terms of the number of parameters, accuracy, and computational resources. MobileNet achieves nearly the same accuracy as VGG16 but with significantly smaller-sized parameters, being 32 times smaller, and requiring 27 times less computational resources (Mult-Adds). MobileNet outperforms GoogleNet in terms of accuracy while being smaller and requiring more than 2.5 times less computational resources.

TABLE 2. MobileNet vs original GoogleNet and VGG16: Details include parameter counts, ImageNet accuracy, and Mult-Adds (M-millions)

Name of the CNN	Parameters	Image-Net Accuracy	Mult-Adds
MobileNet [15]	4.2 M	70.6%	569 M
GoogleNet [29]	6.8 M	69.8%	1550 M
VGG16 [26]	138 M	71.5%	15300 M

DOI: https://doi.org/10.18272/aci.vi.3699

OUTCOME OF THE EXPERIMENTS (SECTION 4)

PoI speed-up of the attack in the LR domain (Subsection 4.1)

TABLE 3. Average number of generations required to generate adversarial images (from *acorn1* and *maraca2*) in the LR domain by EA without and with PoI guidance. Results are for $\epsilon = 8, 12, 16$ and the top x% most relevant pixels for $x = 10, 20, 25, 30, 35$. The speed change is also given in percentages; negative values indicate a slower performance, and positive values a faster performance.

ϵ	EA	EA guided with PoI				
		Top 10%	Top 20%	Top 25%	Top 30%	Top 35%
8	2093	6172 –195.0%	3131 –49.6%	2659 –27.0%	2342 –11.9 %	1720 17.8%
12	1728	1101 36.3%	805 53.4%	827 52.2%	799 53.7 %	767 55.6%
16	1787	752 57.9%	670 62.5%	607 66.0%	667 62.7 %	587 67.2%

Table 3 presents the performance of the EA in generating adversarial images in the LR domain, measured by the number of generations, both with and without PoI guidance. The results are based on *acorn1* and *maraca2* (see Additional material), as the EA successfully generated 0.55-strong adversarial images from these two clean images for all the mentioned settings (top x% and ϵ), both with and without additional PoI guidance. The values are averaged for these two attempts.

When ϵ is increased, the performance of EA increases for all the top x% values. The best performance, in terms of the number of generations, of the EA with PoI is obtained when the top 35% of relevant pixels are used with $\epsilon = 16$. It results in a 67.2% speed increase compared to the EA without PoI guidance. For $\epsilon = 16$, Figure 4 shows how EA converges to the target category without PoI on the one hand, and with PoI using top 35% of the most relevant pixels for the (acorn1-rhinoceros beetle) ancestor-target pair on the other hand. EA's learning period is drastically shortened when one uses PoI. Indeed, using PoI, the EA finds the path to the target category almost 60% faster than without PoI. This acceleration behavior is consistent across most of the ancestor-target pairs.

DOI: https://doi.org/10.18272/aci.vi.3699

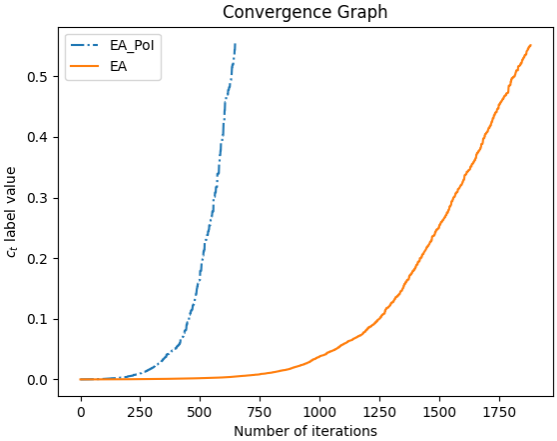

FIGURE 4. With $\epsilon = 16$, EA's convergence pattern from the clean category c_a towards the target category c_t without PoI (EA) and with PoI (EA_PoI, using the top 35% most relevant pixels) is analyzed for c_a = acorn1, c_t = rhinoceros beetle.

Visual quality in the LR domain (Subsection 4.2)

Figure 5 illustrates, with the clean image *acorn1*, the visual quality of low-resolution adversarial images generated by the EA alone (without PoI), and when the EA is guided with PoI using the top 35% of relevant pixels for $\epsilon = 8, 12, 16$. Results for other top x% are provided in Figure 2 of the Additional material. For a human, all obtained adversarial images are challenging to distinguish from the clean image.

PoI+NBU strategy in the HR domain (Subsection 4.3)

In view of what precedes (in terms of speed and visual quality of adversarial images in the LR domain), we used $\epsilon = 16$ and the top 35% most significant pixels identified by BagNet-33 for the remaining experiments combining PoI and NBU.

Using these parameters, the EA generated 56 0.55-strong adversarial images in the LR domain from 100 clean images. Out of the 56, NBU successfully converted 44 of them into HR adversarial images that MobileNet classifies in the target category for the (c_a, c_t) pair and target scenario specified in Table 1. Table 4 summarizes the results for these 44 HR adversarial images; numerical values are averaged. Its first column lists the clean image categories. Note that the *brown_bear* is not included because no 0.55-strong adversarial images were generated from this category. The second column shows the proportion of the image space that is identified by considering the top 35% most relevant pixels. It shows that the EA attack will focus on 70.4% of the clean LR image on average. The third column shows the average number of generations required by the EA to create a 0.55-strong adversarial image in the LR domain (on average, each generation takes between 0.90 and 0.99 seconds). The fourth column gives the average value of $_t$ (which is necessarily ≥ 0.55). The fifth column provides the average *ct*-label value τ_t for degraded adversarial images, and the sixth column gives the resulting average loss $L_C() = _t - \tau_t$, where is the clean HR image classified in the original category c_a.

 DOI: https://doi.org/10.18272/aci.vi.3699

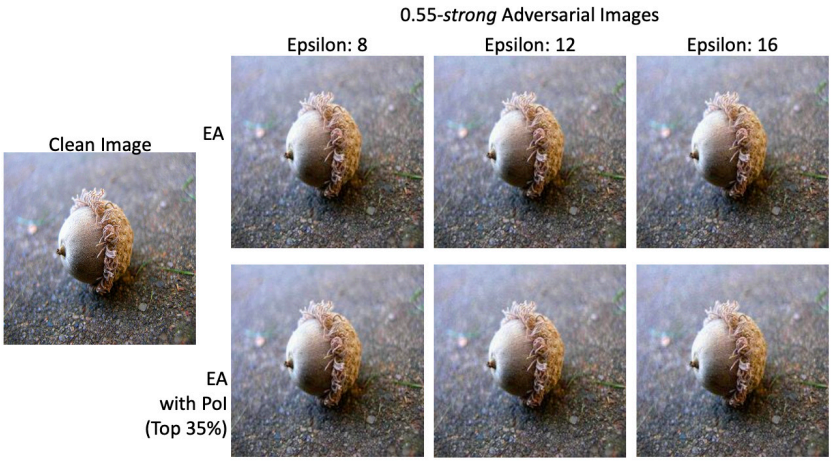

FIGURE 5. Visual quality of the low-resolution adversarial images generated with EA alone and with PoI guidance (using the top 35% of the most relevant pixels) in the LR domain for epsilon values 8, 12, and 16.

On average, the NBU process caused a 0.251 label value loss. Despite this loss, the created HR adversarial images remain adversarial, achieving an average c_t -label value of 0.301.

Table 5 provides the execution time (in seconds) of the main steps of the combined PoI+NBU strategy performed on two representative examples: the largest clean image *canoe4* (2448 x 3264) and the smallest one *llama4* (253 x 380). Using BagNet-33 to find the top 35% most relevant pixels of c_a-label and c_t-label values (combined without any duplication) takes 4.38 seconds. The following step is the attack performed in the LR domain. Its timing varies from one method to another. The EA attack required 23 minutes for one image and 84 minutes for the other. The NBU process blowing up the adversarial noise from the LR domain to the HR domain and adding it to the clean HR image is the last step. It takes less than a second.

Altogether, the PoI+NBU strategy *per se* takes only around 5 seconds and remains completely marginal as compared to the time required by the attack (the EA attack in the present feasibility study). This outcome demonstrates the efficiency of the PoI+NBU approach in generating high-resolution adversarial images with minimal time overhead, apart from the chosen attack method (less than 1% overhead in the case of the EA attack).

DOI: https://doi.org/10.18272/aci.vi.3699

TABLE 4. Average metrics (for top 35% and ϵ=16) for generating 0.55-strong adversarial images in the LR domain, including pixels of interest size (avgPoI), number of generations (avgGens), target label value before (avg_ t) and after NBU (avg_ τ_t), and loss (avg_L).

c_a	avgPoI	avgGens	avg_ $_t$	avg_ τ_t	avg_L
abacus	75%	3774	0.551	0.249	0.302
acorn	69%	1521	0.554	0.316	0.238
baseball	66%	1146	0.554	0.327	0.227
broom	76%	2153	0.550	0.437	0.113
canoe	73%	2383	0.552	0.268	0.284
hippopotamus	69%	4987	0.551	0.304	0.247
llama	67%	3438	0.553	0.236	0.317
maraca	70%	2069	0.552	0.344	0.208
mountain bike	69%	4252	0.553	0.228	0.325
Average	**70.4%**	**2858**	0.552	0.301	0.251

TABLE 5. Time performance of PoI+NBU using the largest and smallest clean images. One uses $\epsilon = 16$, the top 35% most relevant pixels identified by BagNet- 33, and the EA-based attack. Values are in seconds.

HR-Clean Images		PoI	Adversarial Images in the LR domain	NBU	PoI+NBU	Adversarial Images in the HR domain
canoe4	(2448x3264)	4.38	1426	0.74	5.12	1431
llama4	(253x380)	4.37	5071	0.10	4.47	5075

The visual quality of the high-resolution adversarial images (Subsection 4.4)

The visual quality of high-resolution adversarial images generated by the PoI+ NBU strategy for the EA-based attack is assessed on three examples in Figure 6. Its 1st row displays the HR clean images, and its 2nd row their corresponding HR adversarial images. Their names and sizes are at the top of each figure. Despite the added adversarial perturbations, the visual differences between the clean and adversarial HR images are imperceptible to the human eye. To further substantiate this observation, we computed the Fréchet Inception Distance (FID) [35] between clean and adversarial HR images and obtained an average FID score of 54.5.

DOI: https://doi.org/10.18272/aci.vi.3699

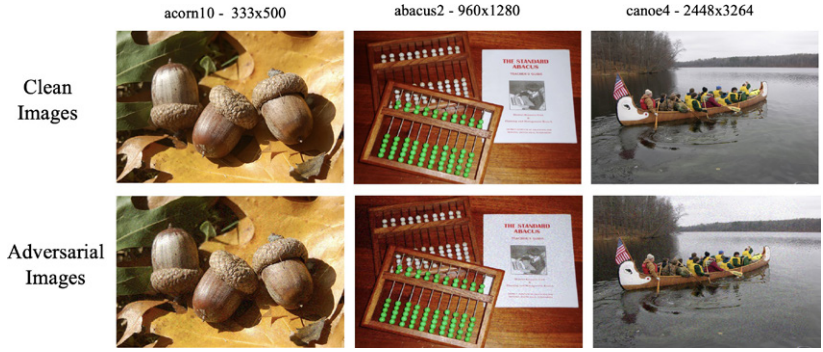

FIGURE 6. Visual comparison between HR clean (1st row) and adversarial (2nd row) images.

CONCLUSION

This paper introduces PoI+NBU, a generic approach that combines the Pixels of Interest (PoI) and Noise Blowing Up (NBU) strategies. The PoI+NBU strategy is designed to enhance the effectiveness of any adversarial attacks, black-box or white-box, against any convolutional neural network for any scenario (targeted or untargeted). The approach is assessed by a feasibility study performed with a black-box evolutionary-based attack on MobileNet for the targeted scenario.

Experiments were performed for different ϵ (measuring the magnitude of values that a pixel value is allowed to be modified), and top x% values (assessing the most significant pixels for the CNN's classification, as assessed by BagNet- 33). Our study showed that $\epsilon = 16$ and $x = 35$ provides a convenient trade-off. With these choices of parameters, the PoI+NBU method created 44 HR adversarial images with the EA-based attack. The visual quality of the adversarial images is outstanding. A human is unable to distinguish the clean HR image from the adversarial one. The overhead of the PoI+NBU strategy is marginal both in absolute and in comparative terms. In absolute terms, its time cost is 5 seconds. It represents less than 1% overhead as compared to the EA-based attack. Future work will focus on testing PoI+NBU with super high-resolution images and exploring its applicability to other adversarial attacks.

AUTHOR CONTRIBUTIONS

Enea Mancellari developed the methodology, performed the coding, experiments, testing, and wrote the original draft. Ali Osman Topal contributed to the conceptualization, supported the methodology, and participated in writing and reviewing. Franck Leprévost supervised the work, contributed significantly to the conceptualization and methodology, and was involved in reviewing and editing the manuscript.

CONFLICT OF INTEREST

All authors declare that they have no conflicts of interest.

DOI: https://doi.org/10.18272/aci.vi.3699

REFERENCES

[1] Koçi, J, Topal, A. O., & Ali, M. (2020). Threat object detection in X-ray images using SSD, R-FCN and Faster R-CNN. *2020 International Conference on Computing, Networking, Telecommunications & Engineering Sciences Applications (CoNTESA),* 10-15. https://doi.org/10.1109/CoNTESA50436.2020.9302863

[2] Ghosh, A., Jana, N. D., Das, S., & Mallipeddi, R. (2023). Two-phase evolutionary convolutional neural network architecture search for medical image classification. *Journal Articles.* https://10.1109/ACCESS.2023.3323705

[3] Khan, M. J., Singh, P. P., Pradhan, B., Alamri, A., & Lee, C.-W. (2023). Extraction of roads using the archimedes tuning process with the quantum dilated convolutional neural network. *Sensors, 23*(21), 8783. https://doi.org/10.3390/s23218783

[4] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Li, F.-F. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition,* 248-255. https://doi.org/10.1109/CVPR.2009.5206848

[5] Meng, W., Xing, X., Sheth, A., Weinsberg, U., & Lee, W. (2014). Your online interests: Pwned! A pollution attack against targeted advertising. *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, 129-140. https://doi.org/10.1145/2660267.2687258

[6] Hardt, M., & Nath, S. (2012) Privacy-aware personalization for mobile advertising. *Proceedings of the 2012 ACM conference on Computer and communications security*, 662-673. https://doi.org/10.1145/2382196.2382266

[7] Leprévost, F., Topal, A. O., & Mancellari, E. (2023). Creating high-resolution adversarial images against convolutional neural networks with the noise blowing-up method. In N. T. Nguyen et al. *Intelligent Information and Database Systems. ACIIDS 2023* (Lecture Notes in Computer Science, Vol. 13995). https://doi.org/10.1007/978-981-99-5834-4_10

[8] Topal, A. O., Mancellari, E., Leprévost, F., Avdusinovic, E., & Gillet, T. (2024). The noise blowing-up strategy creates high-quality, high-resolution adversarial images against convolutional neural networks. *Applied Sciences, 14*(8). https://doi.org/10.3390/app14083493

[9] Leprévost, F., Topal, A. O., Mancellari, E., & Lavangnananda, K. (2023). Zone-of interest strategy for the creation of high-resolution adversarial images against convolutional neural networks. *2023 15th International Conference on Information Technology and Electrical Engineering (ICITEE),* 127-132. https://doi.org/10.1109/ICITEE59582.2023.10317668

[10] Topal, A. O., Chitic, R., & Leprévost, F. (2023). One evolutionary algorithm deceives humans and ten convolutional neural networks trained on ImageNet at image recognition. *Applied Soft Computing, 143.* https://doi.org/10.1016/j.asoc.2023.110397

[11] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861.* https://doi.org/10.48550/arXiv.1704.04861

[12] Varrette, S., Bouvry, P., Cartiaux, H., & Georgatos, F. (2014). Management of an academic HPC cluster: The UL experience. *2014 International Conference on High Performance Computing & Simulation,* 959-967. https://doi.org/10.1109/HPCSim.2014.6903792

[13] Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., & Roli, F. (2013). Evasion attacks against machine learning at test time. *Machine Learning and Knowledge Discovery in Databases,* 387-402. https://doi.org/10.1007/978-3-642-40994-3_25

[14] Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. *2017 IEEE Symposium on Security and Privacy,* 39-57. https://doi.org/10.1109/SP.2017.49

[15] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R. (2013). Intriguing properties of neural networks. *arXiv:1312.6199v4.* https://doi.org/10.48550/arXiv.1312.6199

[16] Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016). The Limitations of Deep Learning in Adversarial Settings. *2016 IEEE European Symposium on Security and Privacy,* 372-387. https://doi.org/10.1109/EuroSP.2016.36

 DOI: https://doi.org/10.18272/aci.vi.3699

[17] Chitic, R., Bernard, N., Leprévost, F. (2020). A proof of concept to deceive humans and machines at image classification with evolutionary algorithms. *Intelligent Information and Database Systems,* 467-480. https://doi.org/10.1007/978-3-030-42058-1_39

[18] Chitic, R., Leprévost, F., Bernard, N. (2020). Evolutionary algorithms deceive humans and machines at image classification: An extended proof of concept on two scenarios. *Journal of Information and Telecommunication,* 5(1), 1-23. https://doi.org/10.1080/24751839.2020.1829388

[19] Brendel, W., & Bethge, M. (2019). Approximating CNNs with bag-of-local-features models works surprisingly well on ImageNet. *International Conference on Learning Representations.* https://doi.org/10.48550/arXiv.1904.00760

[20] Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 226-231. https://dl.acm.org/doi/10.5555/3001460.3001507

[21] Su, J., Vargas, D. V., & Sakurai, K. (2019). One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation, 23*(5), 828-841. https://doi.org/10.1109/TEVC.2019.2890858

[22] Li, Y., Pan, Q., Feng, Z., & Cambria, E. (2023). Few pixels attacks with generative model. *Pattern Recognition, 144,* 109849. https://doi.org/10.1016/j.patcog.2023.109849

[23] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. *arXiv:1412.6572.* https://doi.org/10.48550/arXiv.1412.6572

[24] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2019). Towards deep learning models resistant to adversarial attacks. *arXiv:1706.06083.* https://doi.org/10.48550/arXiv.1706.06083

[25] Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial examples in the physical world. *arXiv:1607:02533.* https://doi.org/10.48550/arXiv.1607.02533

[26] Guo, C., Gardner, J. R., You, Y., Wilson, A. G., & Weinberger, K. Q. (2019). Simple black-box adversarial attacks. *Proceedings of the 36th International Conference on Machine Learning*, 4410-4423. https://doi.org/10.48550/arXiv.1905.07121

[27] Targonski, C. (2019). TensorFlow implementation of generating adversarial examples with adversarial networks. GitHub. https://github.com/ctargon/AdvGAN-tf

[28] Chitic, R., Topal, A. O., & Leprévost, F. (2023). ShuffleDetect: Detecting adversarial images against convolutional neural networks. *Applied Sciences, 13*(6). https://doi.org/10.3390/app13064068

[29] Rybczak, M., & Kozakiewicz, K. (2024). Deep machine learning of MobileNet, efficient, and inception models. *Algorithms, 17*(3), 96. https://doi.org/10.3390/a17030096

[30] Suharto, E., Suhartono, Widodo, A. P., & Sarwoko, E. A. (2020). The use of MobileNet v1 for identifying various types of freshwater fish. *Journal of Physics: Conference Series, 1524.* https://doi.org/10.1088/1742-6596/1524/1/012105

[31] Elhassouny, A., & Smarandache, F. (2019). Smart mobile application to recognize tomato leaf diseases using Convolutional Neural Networks. *2019 International Conference of Computer Science and Renewable Energies*, 1-4. https://www.researchgate.net/publication/343863345_Smart_mobile_application_to_recognize_tomato_leaf_diseases_using_Convolutional_Neural_Networks

[32] Wibowo, A., Adhi Hartanto, C., & Wisnu Wirawan, P. (2020). Android skin cancer detection and classification based on MobileNet v2 model. *International Journal of Advances in Intelligent Informatics, 6*(2), 135-148. https://doi.org/10.26555/ijain.v6i2.492

[33] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *2015 IEEE Conference on Computer Vision and Pattern Recognition*, 1-9. https://doi.org/10.1109/CVPR.2015.7298594

[34] Simonyan, K., & Zisserman, A. (2014) Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556.* https://doi.org/10.48550/arXiv.1409.1556

[35] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems, 30,* 6626-6637. https://doi.org/10.48550/arXiv.1706.08500

 DOI: https://doi.org/10.18272/aci.vi.3699