PoI+NBU: A Feasibility study in Generating High-Resolution Adversarial Images with a Black Box Evolutional Algorithm based Attack (Additional Material)

Enea Mancellari$^{1[0000-0002-8562-1433]}$, Ali Osman Topal$^{1[0000-0003-0141-4742]}$, and Franck Leprévost$^{1[0000-0001-8808-2730]}$

[1]University of Luxembourg, House of Numbers, 6, avenue de la Fonte, L-4364
Esch-sur-Alzette, G-D of Luxembourg
enea.mancellari@uni.lu & aliosman.topal@uni.lu & franck.leprevost@uni.lu

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

1 Clean images

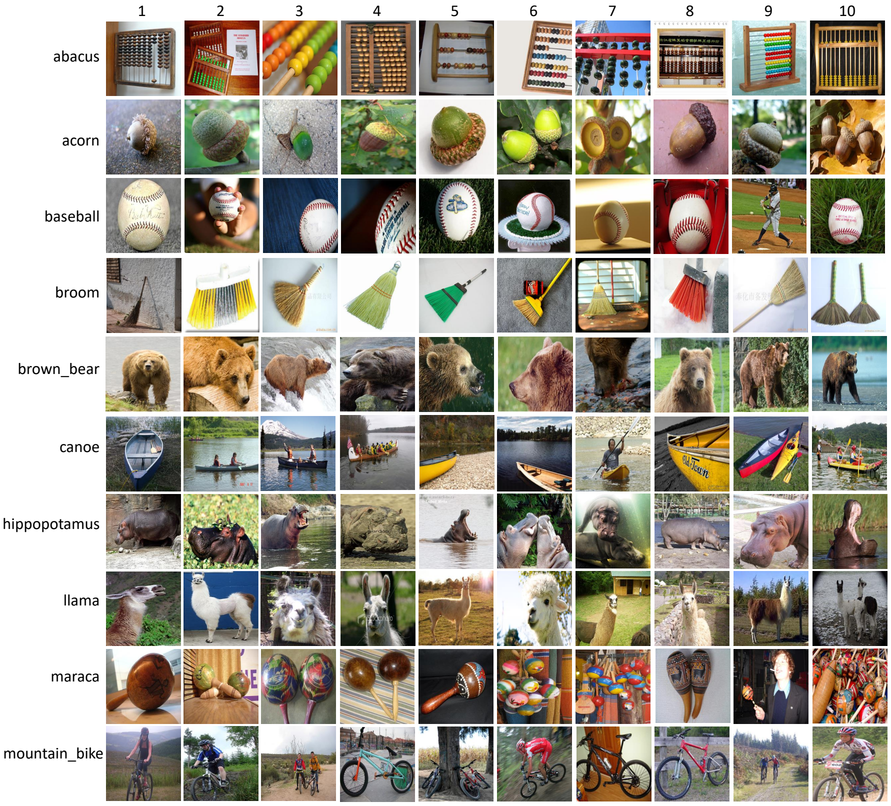

Fig. 1: Representation of the 100 ancestor clean images $\mathcal{A}_q^p$ used in the experiments. $\mathcal{A}_q^p$ pictured in the q^{th} row and p^{th} column $(1 \leq p, q \leq 10)$ is randomly chosen from the ImageNet validation set of the ancestor category c_{a_q} specified on the left of the q^{th} row.

Table 1: Size $h \times w$ (with $h, w \geq 224$) of the 100 clean ancestor images $\mathcal{A}_q^p$.

Ancestor images $\mathcal{A}_q^p$ and their original size $(h \times w)$											
c_{a_q}	q \ p	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
abacus	1	2448x3264	960x1280	262x275	598x300	377x500	501x344	375x500	448x500	500x500	2448x3264
acorn	2	374x500	500x469	375x500	500x375	500x500	500x500	375x500	374x500	461x500	333x500
baseball	3	398x543	240x239	2336x3504	333x500	262x350	310x310	404x500	344x500	375x500	285x380
broom	4	597x400	286x490	360x480	298x298	413x550	366x500	400x400	348x500	346x500	640x480
brown bear	5	700x467	406x500	333x500	500x333	497x750	336x500	480x599	375x500	334x500	419x640
canoe	6	500x332	450x600	360x525	2448x3264	375x500	600x400	1067x1600	333x500	1536x2048	375x500
hippopotamus	7	375x500	1200x1600	333x500	450x291	525x525	375x500	500x457	424x475	500x449	339x500
llama	8	500x333	618x468	500x447	253x380	490x500	333x500	375x1024	375x500	290x345	1920x2560
maraca	9	700x642	375x500	470x627	900x928	960x1280	500x375	500x375	375x500	375x500	375x500
mountain bike	10	768x1024	500x375	375x500	333x500	500x375	300x402	768x1024	446x500	375x500	2065x1335

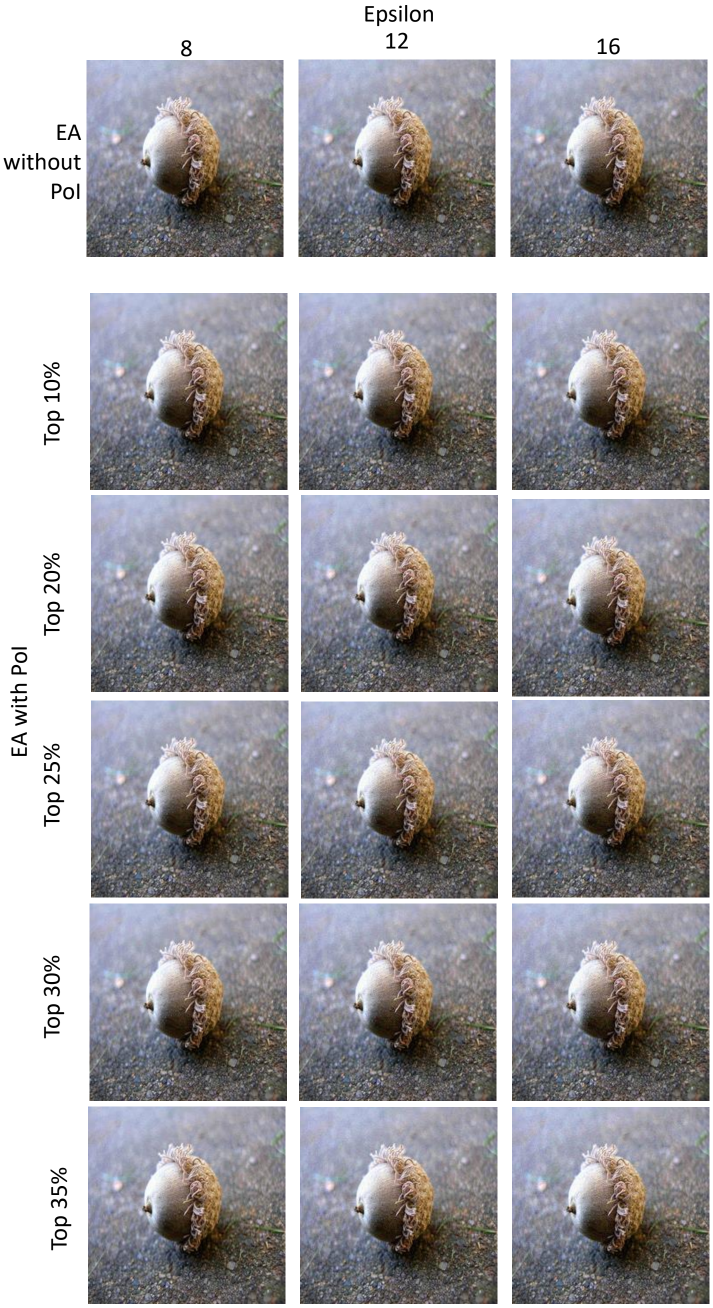

Fig. 2: Visual quality of the EA-generated adversarial images in the LR domain under various settings: without and with PoI; for epsilon values 8, 12, and 16; and with PoI using different Top $x\%$ of the most relevant pixels, where $x = 10$, 20, 25, 30, and 35.

2 Proportion of pixels with various x values for the Top x%.

For each c_a category, Table 2 displays, for various values of x, the average (computed over the 10 clean images selected from c_a, resized to the LR domain 224×224) proportion of pixels, that contribute to the Top $x\%$ of the c_a-label value and c_t-label value (taken together without any duplication) as measured by BagNet-33.

As $x\%$ increases, there is a rapid rise in the occupied image space up to 5%. Beyond 5%, the increase becomes more gradual (see Figure 3). This trend occurs because the most relevant pixels are densely clustered around the object's key structural features, with additional pixels extending into less critical areas, resulting in a smoother increase in occupied space.

Table 2: **Proportion of image space** occupied by the Top $x\%$ most relevant pixels in the LR domain for the c_a and the c_t categories as measured by BagNet-33.

Top $x\%$ c_a	1%	2%	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%	55%	60%
abacus	4.70	8.87	19.49	33.33	44.12	52.82	59.98	65.97	71.04	75.26	78.88	81.98	84.60	86.87
acorn	4.66	8.55	18.35	31.25	41.82	50.70	58.15	64.44	69.85	74.46	78.32	81.56	84.31	86.67
baseball	4.73	8.70	18.55	31.45	41.81	50.32	57.46	63.69	69.07	73.72	77.71	81.14	84.05	86.48
broom	4.71	8.82	19.30	32.73	43.10	51.41	58.23	63.97	68.87	73.02	76.71	79.93	82.83	85.35
brownbear	4.79	9.01	19.91	34.36	45.55	54.49	61.74	67.65	72.54	76.66	80.11	83.03	85.52	87.60
canoe	4.60	8.49	18.40	31.58	42.05	50.70	57.92	64.03	69.21	73.69	77.56	80.95	83.85	86.27
hippopotamus	4.81	8.92	18.95	31.60	41.49	49.67	56.70	62.78	68.07	72.75	76.81	80.35	83.36	85.92
llama	4.55	8.44	18.95	30.48	40.25	48.53	55.57	61.71	67.14	71.87	76.04	79.67	82.81	85.49
maraca	4.75	8.87	18.09	32.51	42.97	51.51	58.68	64.77	70.03	74.51	78.38	81.66	84.47	86.85
mountainbike	4.70	8.68	19.20	32.16	42.67	51.29	58.68	64.48	69.62	73.96	77.71	80.95	83.73	86.10
Average	**4.70**	**8.74**	**18.90**	**32.15**	**42.58**	**51.14**	**58.28**	**64.35**	**69.54**	**73.99**	**77.82**	**81.12**	**83.95**	**86.36**

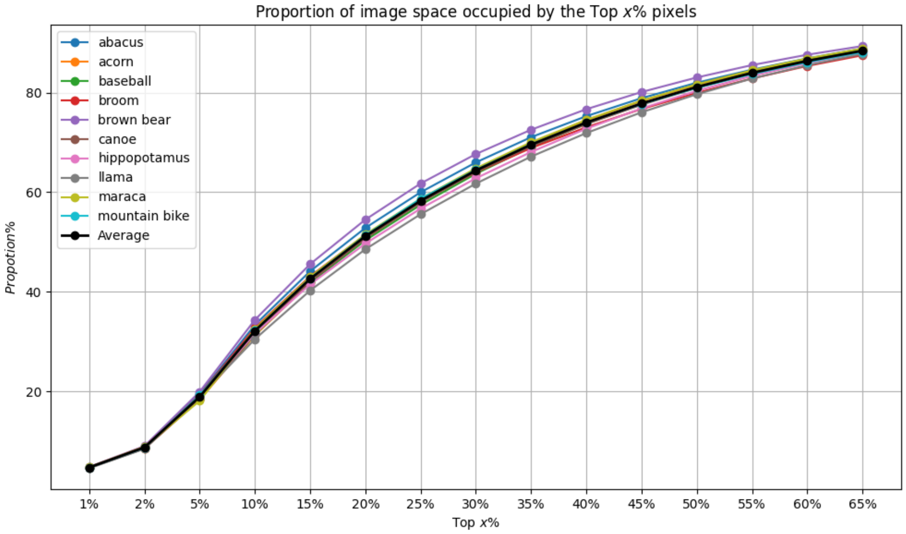

Fig. 3: **Proportion of image space occupied by the Top $x\%$ most relevant pixels in the LR Domain for the c_a category and for the c_t-category as measured by BagNet-33.**