Supplementary Material: You See What I Want You to See: Exploring Targeted Black-Box Transferability Attack for Hash-based Image Retrieval Systems

Yanru Xiao and Cong Wang
Old Dominion University, Norfolk, VA
{yxiao002,c1wang}@odu.edu

In this supplementary material, we present more results of the proposed mechanism by: 1) evaluating the performance on softmax classification tasks; 2) visualizing more examples of "advertising for free"; 3) comparing with the query-based black-box attacks.

1. Performance on Softmax Classification

We conduct additional experiments to evaluate the proposed mechanism on softmax classification tasks using the same setting of ImageNet and compare with the aforementioned benchmarks [8, 10]. Since softmax returns the $\arg\max$ label as the classification result, we revise the objective of NAG slightly,

$$\epsilon^*(\sigma) = \underset{||x'-x||_\infty<\eta}{\arg\min} \left(\mathbb{E}_{r\sim\mathcal{N}(0,\sigma^2 I)} \left(\mathcal{L}_{CE}(x', x_t) + \lambda ||d(x'+r, x')||_1 \right) \right) \tag{1}$$

Given the target label x_t, the first term is the cross-entropy loss of targeting x' at x_t. The second term is the l_1 distance between the output of $x'+r$ and x', that the goal is to keep noise-injected $x'+r$ close to x' so that it remains adversarial [6]. λ is a scaling parameter to balance the two losses.

We randomly select one from the 100 categories as the target class (other than the original source class) and show the transferability rates in Table 1. We add Inception_v3 into the model combination because of improved accuracy on the softmax classification task. We set $\lambda = 50$ to weigh more on the second loss and the noise level is set to 32. From Table 1, the average black-box transferability is 0.93%, 1.11%, 4.19% and 4.69% for PGD, DI, DI-Mom and NAG, i.e., both DI-Mom and NAG offer $4\times$ targeted transferability compared to PGD and DI. Note that DI/DI-Mom were originally proposed and tested on softmax classification, which defeat the winners of NIPS 2017 adversarial competition by a large margin. NAG is slightly better or on par with DI-Mom in softmax.

Discussion. We notice some interesting phenomenons during our experimentation on the hashing and classification networks. The first one is their response to the injected random noise: softmax classification is more robust to random noise, such that: a) the convergence of (1) is much faster than hashing; b) for the same level of random noise, NAG is more effective in deep hashing than softmax (our mechanism is comparable to DI-Mom in softmax, but with more than 15% improvements in deep hashing). In contrast to softmax, hashing learns similarity relations from pairwise inputs. The difference could be investigated along the direction of attentive regions/feature structures learned by deep hashing and softmax. We also notice that the variation of loss curvature in deep hashing is higher than softmax during training, which indicates that softmax may have a smaller Lipchitiz constant overall (sensitivity of the network to perturbations). This may partially illustrate why random noise is less effective on softmax classification. We also notice that the distributions of the perturbation generated by NAG visually retain a Gaussian distribution. Since the solutions are mostly found among the vertices of the l_∞ ball, the rest of the additive Gaussian noise is preserved. The fundamental questions of how much random noise would help learn a randomized smoothing classifier [9], improve black-box transferability (compared to the competitive input diversity methods [10]), and resolve the relations between adversarial perturbations and random noise in different learning tasks (softmax/metric/hashing) are worth future research efforts.

2. More Examples from "Advertising for Free"

We present more examples to advertise for free using the proposed attack. Figs.1 and 2 visualize the two strategies. Recall that in Strategy I, we randomly pick a fixed number of images from the most vulnerable category and generate the corresponding adversarial images (one of them is depicted with the perturbation). It is seen that the vulnerable categories do not appear to be purely random - some of them have obvious semantic relations, e.g., the advertisement of "beer/soda" (third row) is closest to "lotion", because most lotion images contain bottle(s). This allows NAG to realize black-box transfer attacks more easily. Similarly, the fourth advertisement of beverage features a dog in the image and

		ResNet34	ResNet50	ResNet101	ResNet152	ResNext101	SeResNet50	Inception_v3	DenseNet161
Res34	PGD	**100.0**	1.3	1.1	1.0	1.2	1.0	0.9	0.3
	DI	99.9	1.3	1.2	1.1	1.6	0.5	0.5	0.4
	DI-M	**100.0**	4.5	3.7	4.5	4.3	2.9	2.5	2.7
	NAG	98.5	**7.8**	**6.2**	**6.5**	**5.6**	**3.3**	**5.3**	**4.2**
Res50	PGD	1.2	**100.0**	1.7	1.3	1.4	0.9	0.3	0.3
	DI	1.6	**100.0**	1.8	1.7	1.6	1.1	1.0	0.6
	DI-M	**6.5**	99.8	6.0	7.1	5.0	**4.1**	2.6	4.1
	NAG	5.0	99.7	**7.2**	**8.8**	**6.7**	**4.1**	**3.2**	**4.3**
Res101	PGD	1.1	2.1	**100.0**	1.8	1.6	0.9	0.7	0.6
	DI	1.2	2.5	**100.0**	2.6	2.1	0.7	1.2	0.8
	DI-M	**7.4**	**8.8**	**100.0**	**11.9**	6.3	3.7	2.6	**4.0**
	NAG	4.9	8.7	99.6	9.1	**7.5**	**3.8**	**3.6**	3.5
Res152	PGD	1.5	2.3	3.6	**100.0**	1.9	0.9	0.9	0.5
	DI	1.2	3.2	3.8	**100.0**	2.2	0.9	1.0	1.0
	DI-M	**8.0**	8.9	8.0	99.9	5.7	4.9	3.1	4.9
	NAG	7.3	**11.9**	**10.4**	99.8	**9.9**	**5.3**	**4.9**	**6.1**
Next101	PGD	0.8	1.1	1.5	1.1	**100.0**	0.7	0.7	0.3
	DI	1.5	1.5	2.0	1.3	99.9	0.3	0.6	1.0
	DI-M	6.2	6.3	4.4	6.5	**100.0**	**3.3**	2.6	4.2
	NAG	**6.4**	**9.5**	**7.9**	**9.0**	**100.0**	2.9	**3.4**	**7.6**
SeRes50	PGD	0.4	0.4	0.5	0.8	0.7	**100.0**	0.7	0.3
	DI	0.4	0.9	0.7	0.5	0.5	**100.0**	0.9	0.6
	DI-M	3.1	3.0	1.9	**3.0**	**2.8**	**100.0**	1.9	2.5
	NAG	**3.6**	**4.2**	**2.9**	2.8	2.6	99.6	**3.6**	**2.7**
Inc_v3	PGD	0.5	0.3	0.5	0.4	0.4	0.5	**100.0**	0.3
	DI	0.8	0.3	0.4	0.3	0.6	0.2	99.9	0.4
	DI-M	**2.9**	**2.6**	**2.1**	**2.5**	**1.9**	**2.0**	**100.0**	2.0
	NAG	2.1	2.2	1.9	2.0	1.8	**2.0**	91.4	**2.2**
Dense161	PGD	0.8	0.8	0.6	0.8	0.9	0.6	0.4	**100.0**
	DI	0.8	0.8	1.2	0.8	0.8	0.8	0.9	99.9
	DI-M	**4.3**	**2.8**	**2.1**	**3.2**	**2.3**	**1.9**	**2.1**	**100.0**
	NAG	0.6	0.6	1.0	0.6	0.6	0.5	0.8	90.7

Table 1: Targeted attack success rates of softmax classification (%). The diagonal blocks indicate the white-box success rates. On average, both DI-Mom and NAG offer $4\times$ targeted transferability compared to PGD and DI. NAG is slightly better than DI-Mom on softmax classification tasks.

"Welsh Springer" came as the vulnerable category (not directly retrievable within T_h). It is interesting to see that the last advertisement of handbag includes the posture of lying on the side and "studio couch" emerges as the vulnerable class, which also has some connections. Strategy II exploits the top-n most vulnerable categories and selects one image from each category to generate adversarial examples. Except a few vulnerable categories with semantic relations, the rest seem quite random, e.g., one may ask why leopard/clog is in any way similar to the last advertisement of an Android phone. In our experiment, we found that these categories actually have lower chance to succeed.

3. Compare with Query-based Black-Box Attacks

An alternative to make the adversarial example transferrable is through repetitive queries, which does not need any other information except normal access to the black-box model [7, 4, 1]. In this section, we evaluate the success rate of query-based black-box attacks. Most of the query-based attacks leverage the probability score [7, 4] or the decision [5, 2, 3] in softmax classification models. State-of-the-art techniques can achieve nearly 100% success rate in both targeted/untargeted attacks [1]. To adapt these attacks to image retrieval, we provide an extension to transfer the list of retrieved images back to a probability vector via calculating the proportion of each category. The attacker can adopt a pre-trained, auxiliary model to facilitate the classification of retrieved images. We implement both untargeted and targeted attacks based on the method in [1].

We randomly select 1,000 images from the ImageNet in our evaluation. For untargeted attack to succeed, it has to subvert the original query results, i.e., retrieving more images from irrelevant categories or suppressing the number of correctly matched results. Here, if the retrieved images from the original class have been reduced below 10,

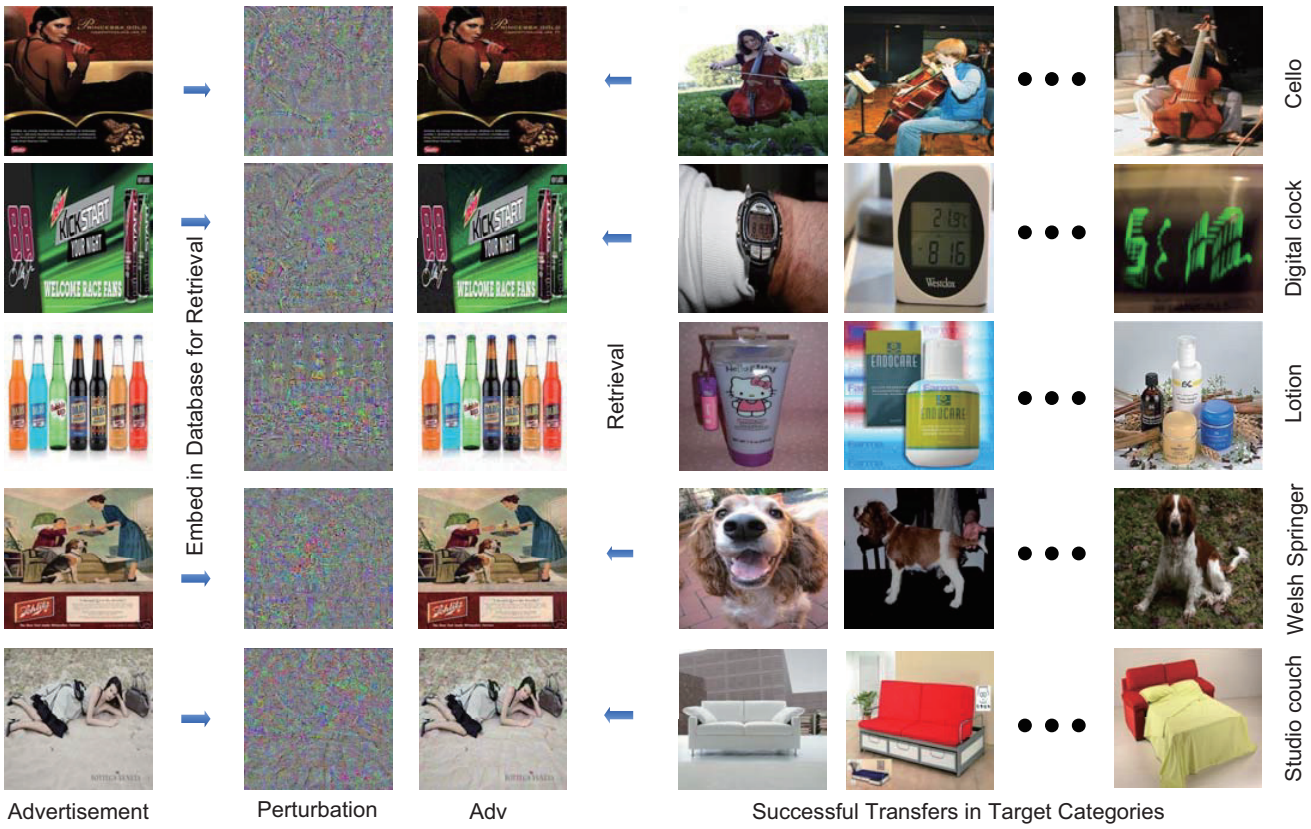


Figure 1: Visualization of Strategy I: exploit the most vulnerable categories. For each advertisement image, randomly pick a fixed number of images from the most vulnerable category and generate corresponding adversarial examples.

Networks	Untargeted			Targeted		
	$L_\infty = 8$	$L_\infty = 16$	$L_\infty = 32$	$L_\infty = 8$	$L_\infty = 16$	$L_\infty = 32$
ResNet101	0.9876	0.9904	0.9986	0.001	0.001	0
ResNet152	0.9854	0.9893	0.9960	0.006	0.006	0.002
ResNext101	0.9838	0.9888	0.9963	0.007	0.006	0.005

Table 2: Success rates of query-based untargeted and targeted attacks [1]

we consider that the untargeted attack is successful (even if the hash code does not provide a match to any of the images). For targeted attack, we randomly select the target class and consider the attack to be successful when the adversarial image retrieves more than 10 images from the targeted class. We test the method on ResNet101, ResNet152 and ResNext101 models and show the success rates in Table 2 and the number of queries in Fig. 3.

We can see that the untargeted and targeted attacks have contrastive results. While the untargeted attacks can reach nearly 100% success, the targeted attack is barely successful in image retrieval, which is even lower than the transferred based attacks. This is due to different levels of difficulty to achieve attack success. In deep hashing, the majority of hash space consists of empty space with no hash code corresponding to a relevant image category, so untargeted attacks are successful as long as the retrieval results are diverted from the original query. Unlike in softmax classification that query-based attacks can achieve high success rate [1], targeted attacks in deep hashing is more difficult. The reason is due to the fact that the images are only mapped to a small portion of hash codes. The loss function of the black-box query would update towards the direction of the adversarial objective. However, once it has stepped into the non-searchable regions of hash codes, the loss function stops progressing because there is no retrievable results from the model, nor is the method able to step out of the non-searchable region given little useful feedback

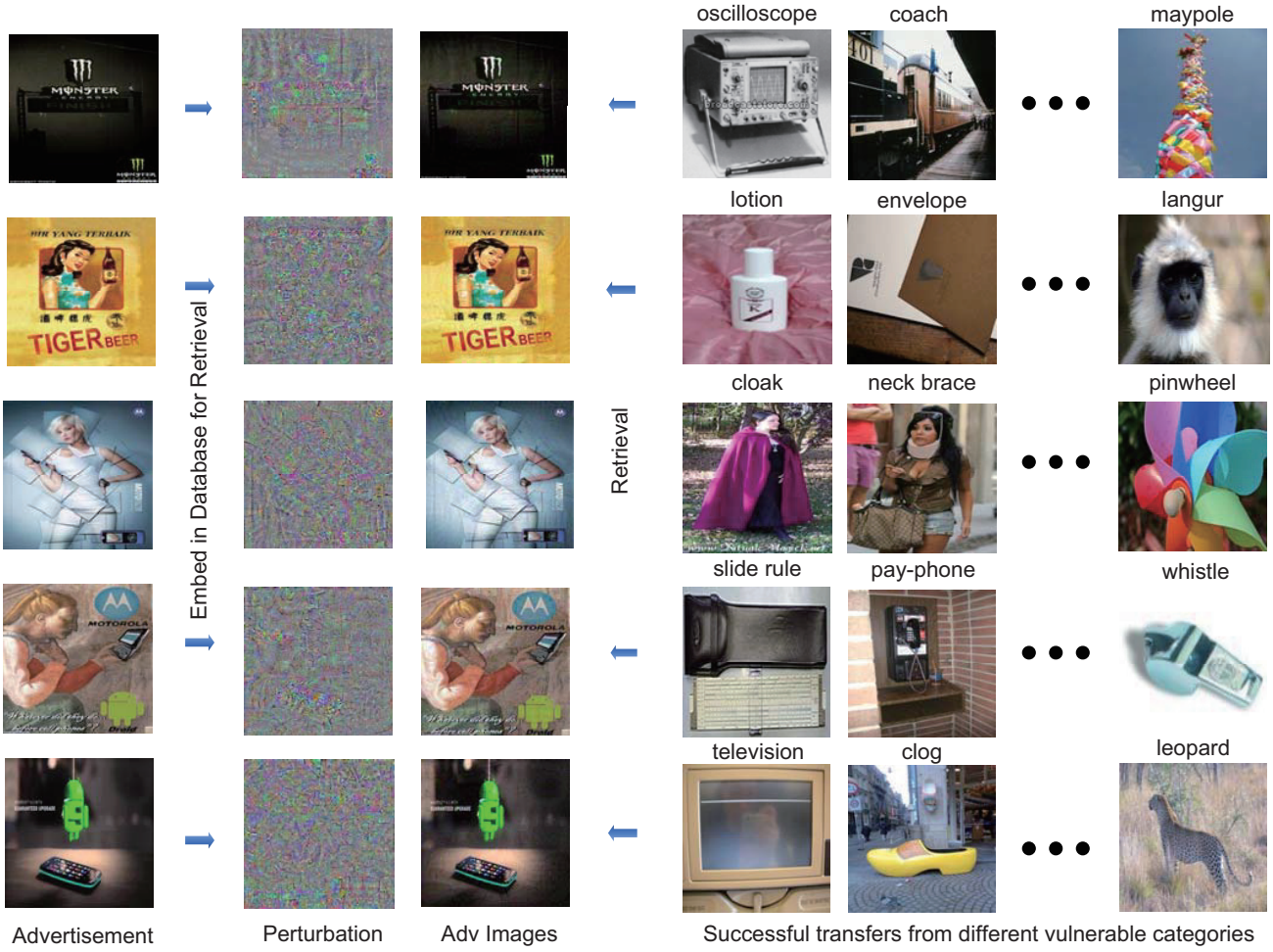
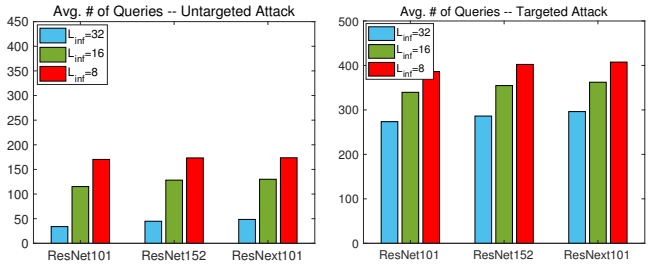


Figure 2: Visualization of Strategy II: exploit top-n vulnerable categories. For each advertisement images, pick the most n vulnerable categories according to the hamming distance and generate an adversarial example for each category.



(a) (b)

Figure 3: Number of queries needed to accomplish (a) untargeted attack (b) targeted attack.

from the model. Thus, the success rate of targeted attacks remains extremely low and the number of queries to accomplish those very few successful cases are much higher.

References

[1] M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *European Conference on Computer Vision*, pages 484–501. Springer, 2020. 2, 3

[2] W. Brendel, J. Rauber, and M. Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *ICLR*, 2018. 2

[3] T. Brunner, F. Diehl, M. T. Le, and A. Knoll. Guessing smart: Biased sampling for efficient black-box adversarial attacks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4958–4966, 2019. 2

[4] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 15–26, 2017. 2

[5] Y. Dong, H. Su, B. Wu, Z. Li, W. Liu, T. Zhang, and J. Zhu. Efficient decision-based black-box adversarial at-

tacks on face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7714–7722, 2019. 2

[6] S. Hu, T. Yu, C. Guo, W.-L. Chao, and K. Q. Weinberger. A new defense against adversarial images: Turning a weakness into a strength. In *Advances in Neural Information Processing Systems*, pages 1635–1646, 2019. 1

[7] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin. Black-box adversarial attacks with limited queries and information. *ICML*, 2018. 2

[8] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 1

[9] H. Salman, J. Li, I. Razenshteyn, P. Zhang, H. Zhang, S. Bubeck, and G. Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems*, pages 11292–11303, 2019. 1

[10] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, and A. L. Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2730–2739, 2019. 1