

Curls & Whey: Boosting Black-Box Adversarial Attacks

Supplemental Material

Yucheng Shi, Siyu Wang, Yahong Han
College of Intelligence and Computing
Tianjin University, Tianjin, China

{yucheng, syuwang, yahong}@tju.edu.cn

In this supplemental material, we show additional experimental results and more adversarial examples generated by Curls & Whey attack, including tables of untargeted black-box attack on Imagenet [4] and targeted black-box attack on Tiny-Imagenet. Adversarial examples generated on two datasets are listed behind.

A. Untargeted Attack

In Table 1, we report median and average ℓ_2 distance of adversarial perturbations crafted on Imagenet dataset. Four DNN models with different structures are compared: resnet-101 [1], densenet-161 [3], vgg19-bn [5] and senet-154 [2]. In this 4×4 matrix, each element represents the result of substitute model of this row against the target model of this column over the entire 1000×10 images collected from validate set of Imagenet, 10 images for each category. We compare our Curls & Whey attack with four other attack methods, FGSM, I-FGSM, MI-FGSM and vr-IGSM. As can be seen, Curls & Whey achieves smaller median noise magnitude in ℓ_2 norm than other methods on black-box attacks (off-diagonal elements). Because of the gaussian noises introduced, our method as well as vr-IGSM perform not so well on white-box results (diagonal elements), where transferability is guaranteed to be 100%. Fig. 1 illustrates adversarial examples crafted by different attacks. The leftmost images in Fig. 1 are original images. Noise magnitude in ℓ_2 norm is placed below each adversarial example.

B. Targeted Attack

In Table 2, we provide median and average adversarial perturbation on 200×10 images collected from Tiny-Imagenet dataset, 10 images for each category. Four DNN models are compared: resnet-18, inception V3 [7], inception-resnet V2 [6] and nasnet [8]. As we can see, the performance of Curls & Whey is far beyond other methods. Results of iterative attacks like I-FGSM, MI-FGSM and vr-IGSM are all around 80, which means these methods seldom successfully achieve targeted misclassification with

small ℓ_2 distance. Three decision based attacks, boundary attack, pointwise attack and vanilla interpolation are also compared. These methods do not rely on substitute model, but collect a legitimate image that can be classified into the target category by the target model first and then search between original image and this image. Our method significantly reduces the noise magnitude of targeted attack in black-box scenario. Several groups of targeted adversarial examples are shown in Fig. 2, where original image, image of target category, noise and targeted adversarial example are listed from left to right in each group.

References

- [1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1
- [2] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *arXiv preprint arXiv:1709.01507*, 2017. 1
- [3] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *CVPR*, pages 2261–2269, 2017. 1
- [4] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Sathesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 1
- [5] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 1
- [6] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, volume 4, page 12, 2017. 1
- [7] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016. 1
- [8] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. Learning transferable architectures for scalable image recognition. *CoRR*, abs/1707.07012, 2017. 1

Table 1. Median and average ℓ_2 distance of adversarial perturbation for untargeted attack on Imagenet.

		resnet-101		densenet-161		vgg19-bn		senet-154	
	attack methods	median	average	median	average	median	average	median	average
resnet-101	FGSM	0.3167	3.9776	6.7431	16.5051	6.0534	15.1296	7.9076	15.6174
	I-FGSM	0.2045	0.4065	2.0577	3.8353	1.9820	4.0847	3.5190	6.5371
	MI-FGSM	0.2305	0.2864	2.2312	4.3944	2.2569	5.1220	4.1914	8.1522
	vr-IGSM	0.2390	0.2940	1.9114	3.7779	1.8940	4.3436	3.3970	7.0107
	Curls&Whey	0.2295	0.5198	1.8655	3.5871	1.7285	3.5187	2.9872	5.3745
densenet-161	FGSM	6.5304	15.3936	0.3070	4.2659	5.3533	13.0053	6.8372	14.5869
	I-FGSM	1.8307	3.8534	0.2173	0.5338	1.7311	3.6356	2.8061	5.2298
	MI-FGSM	1.9436	4.2498	0.2258	0.2051	1.9164	4.4277	3.2574	6.3578
	vr-IGSM	1.8994	3.7641	0.2576	0.1834	1.6656	3.8322	2.7925	5.6133
	Curls&Whey	1.7041	3.3246	0.2494	0.7397	1.5771	3.1351	2.4977	4.4188
vgg19-bn	FGSM	9.9305	19.8893	8.7631	16.3457	0.1819	2.2736	11.2227	23.8974
	I-FGSM	4.2179	8.7935	3.9970	7.5216	0.1406	0.8352	4.5875	8.7350
	MI-FGSM	4.5438	9.9437	4.1861	8.3386	0.1468	0.2462	5.3552	10.2055
	vr-IGSM	3.6475	8.3765	3.4204	6.9270	0.1537	0.2357	4.2161	8.7974
	Curls&Whey	3.3500	6.9225	3.2049	6.3321	0.1511	0.8173	3.6962	7.1415
senet-154	FGSM	8.3359	15.4190	8.3936	15.1964	7.9624	14.7991	0.6791	5.9169
	I-FGSM	4.2529	7.9353	4.1996	7.4578	2.4439	4.8991	0.3478	0.9178
	MI-FGSM	4.5414	9.9268	4.5520	9.6595	2.9568	6.6679	0.4465	0.4386
	vr-IGSM	3.4674	8.4754	3.5301	8.3745	2.5631	5.9603	0.3226	0.7623
	Curls&Whey	3.0064	5.8348	3.0913	5.5426	1.9326	3.6826	0.2665	0.4206

Table 2. Median and average ℓ_2 distance of adversarial perturbation for targeted attack on Tiny-Imagenet.

		resnet-18		inception V3		inc-resnet V2		nasnet	
resnet-18	FGSM	81.2735	72.7756	82.5241	80.9782	82.5322	81.0193	82.5626	80.9734
	I-FGSM	1.5398	2.9277	81.2345	70.4633	80.9559	70.9525	81.9787	76.3114
	MI-FGSM	5.4267	34.7935	80.9999	68.6765	80.6864	68.8399	81.6331	73.6505
	vr-IGSM	0.3751	0.4328	80.9087	68.7406	80.6796	67.6213	81.6413	73.5731
	Interpolation	27.1537	28.0997	24.8444	25.2685	24.0634	24.9918	24.2808	24.9455
	Pointwise	40.0754	40.8887	39.8188	40.4638	39.9544	40.6741	40.0107	40.6636
	Boundary	31.7736	32.5285	31.2757	31.8612	31.5086	32.0049	31.4495	32.0101
	Curls&Whey	2.9242	3.5819	9.3365	9.8224	9.1087	9.6767	9.2421	9.8868
inception V3	FGSM	82.5945	81.4726	82.5049	80.1433	82.4776	80.8414	82.6294	81.7538
	I-FGSM	81.7994	76.8168	0.3668	0.4005	80.6272	66.7824	81.8072	75.4738
	MI-FGSM	81.7709	75.286	0.6941	0.787	80.6065	65.7128	81.2065	71.1222
	vr-IGSM	81.5944	73.4011	0.7074	0.7944	80.5404	64.9861	81.6332	71.7732
	Interpolation	27.1537	28.0997	24.8444	25.2685	24.0634	24.9918	24.2808	24.9455
	Pointwise	40.0754	40.8887	39.8188	40.4638	39.9544	40.6741	40.0107	40.6636
	Boundary	31.7736	32.5285	31.2757	31.8612	31.5086	32.0049	31.4495	32.0101
	Curls&Whey	9.3832	10.2564	1.2996	1.9423	7.0783	7.8715	7.9913	8.7716
inc-resnet V2	FGSM	82.5945	81.1598	82.4787	80.9113	82.4778	80.6907	82.5397	81.4183
	I-FGSM	81.8539	76.5748	80.5121	66.4706	0.5139	1.8753	81.7709	74.8515
	MI-FGSM	81.7926	75.2021	80.6945	66.7647	1.5027	2.0268	81.4114	71.1565
	vr-IGSM	81.5944	74.2181	80.6342	66.0386	1.5101	2.0193	82.3142	75.3981
	Interpolation	27.1537	28.0997	24.8444	25.2685	24.0634	24.9918	24.2808	24.9455
	Pointwise	40.0754	40.8887	39.8188	40.4638	39.9544	40.6741	40.0107	40.6636
	Boundary	31.7736	32.5285	31.2757	31.8612	31.5086	32.0049	31.4495	32.0101
	Curls&Whey	9.4622	10.0514	7.7449	8.3226	2.6911	3.0239	7.5282	8.2552
nasnet	FGSM	82.5626	81.3046	82.5626	81.4469	82.4840	81.0555	82.4787	80.6151
	I-FGSM	81.6910	75.8341	79.8133	64.4298	79.9032	65.0252	0.3363	0.3741
	MI-FGSM	81.6249	74.6014	80.5495	67.4740	79.8082	63.0784	0.6785	0.7839
	vr-IGSM	81.6166	74.2587	80.5220	66.5961	79.7085	62.8439	0.7582	0.8575
	Interpolation	27.1537	28.0997	24.8444	25.2685	24.0634	24.9918	24.2808	24.9455
	Pointwise	40.0754	40.8887	39.8188	40.4638	39.9544	40.6741	40.0107	40.6636
	Boundary	31.7736	32.5285	31.2757	31.8612	31.5086	32.0049	31.4495	32.0101
	Curls&Whey	11.1441	11.867	8.25630	8.70140	6.9883	7.6393	1.3578	1.9226







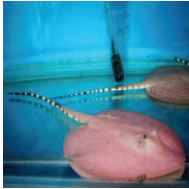
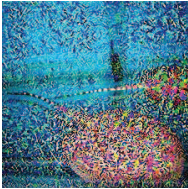
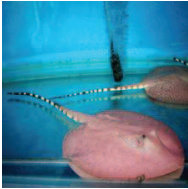
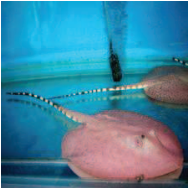
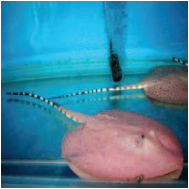
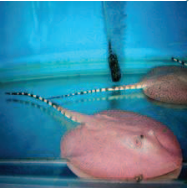






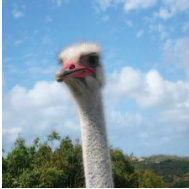












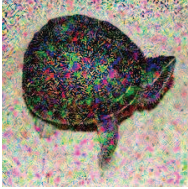




original	FGSM	I-FGSM	MI-FGSM	vr-IGSM	Curls&Whey
 white shark	 goldfish $l_2 = 5.8269$	 goldfish $l_2 = 0.9121$	 goldfish $l_2 = 0.9375$	 goldfish $l_2 = 0.8967$	 goldfish $l_2 = 0.8324$
 crampfish	 eel $l_2 = 45.0341$	 stingray $l_2 = 3.4814$	 stingray $l_2 = 3.8821$	 stingray $l_2 = 3.6347$	 stingray $l_2 = 3.3096$
 cock	 hen $l_2 = 27.9184$	 hen $l_2 = 4.8635$	 hen $l_2 = 6.1862$	 hen $l_2 = 5.2485$	 hen $l_2 = 4.4770$
 ostrich	 maypole $l_2 = 33.9738$	 vulture $l_2 = 7.0560$	 vulture $l_2 = 12.5591$	 vulture $l_2 = 13.4132$	 vulture $l_2 = 6.9842$
 chickadee	 humming bird $l_2 = 14.8459$	 magpie $l_2 = 4.6537$	 magpie $l_2 = 5.1470$	 magpie $l_2 = 5.1279$	 indigo bunting $l_2 = 3.9699$
 mud turtle	 armadillo $l_2 = 41.9588$	 terrapin $l_2 = 4.1295$	 loggerhead $l_2 = 3.8689$	 loggerhead $l_2 = 4.6041$	 terrapin $l_2 = 3.1625$

Figure 1. Six groups of untargeted attack examples on Imagenet. Columns from left to right show original image and the adversarial examples generated by five different methods. The misclassification category and l_2 norm of the noise are below each image.

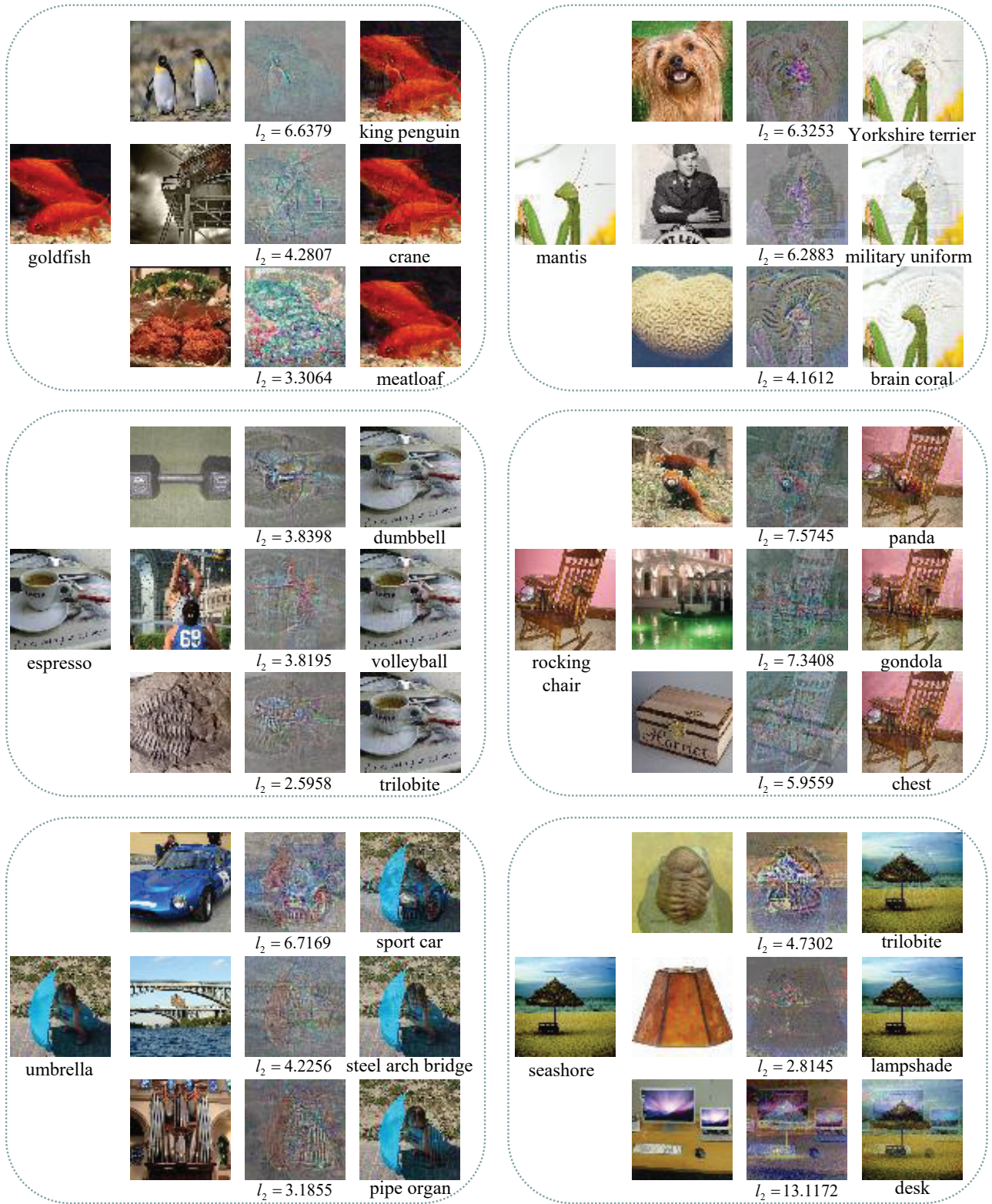


Figure 2. Six groups of targeted attack examples on Tiny-Imagenet. Original image, images of target category, noises and adversarial examples are listed from left to right in each group. By adding three different noises, each original image is misclassified into three other categories.