Attacking Perceptual Similarity Metrics

Abhijay Ghildyal and Feng Liu

Motivation

How robust are perceptual similarity metrics against imperceptible adversarial perturbations?

Perceptual similarity metrics measure the similarity between two images and are widely used in many real-world applications. Thus, having a robust metric is critical.

Currently, there are two popular approaches for examining the robustness of perceptual similarity metrics:

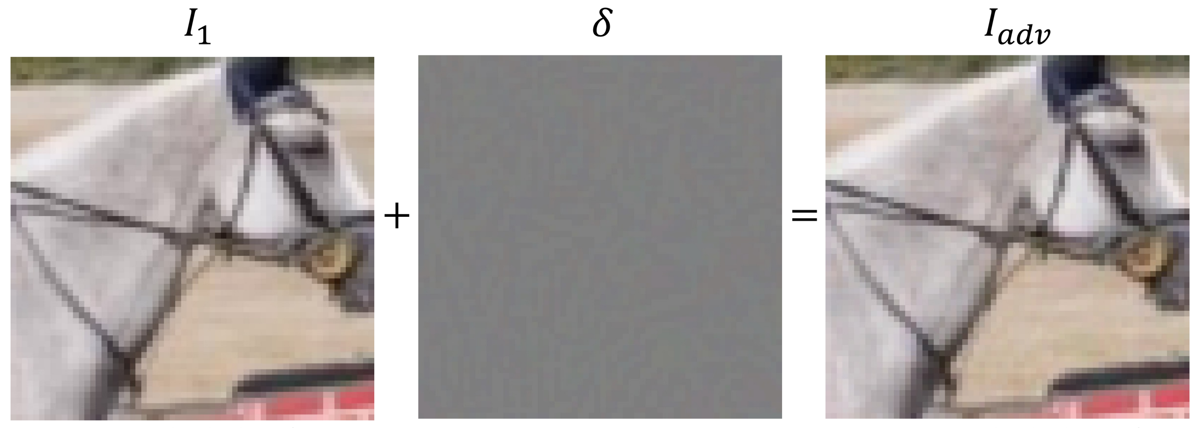
1. Addition of small amounts of hand-crafted distortions such as translation, rotation, dilation, speckle noise, color jittering, JPEG compression, and Gaussian blur.
2. Analysis of more advanced adversarial perturbations.

We focused our efforts on (2), i.e. adversarial attacks, as it has not received considerable attention.

Attack Methods

In real-world scenarios, attackers might lack access to crucial details like a metric's architecture, parameters, or data. To overcome this, they can transfer adversarial examples from one metric, such as LPIPS, to another.

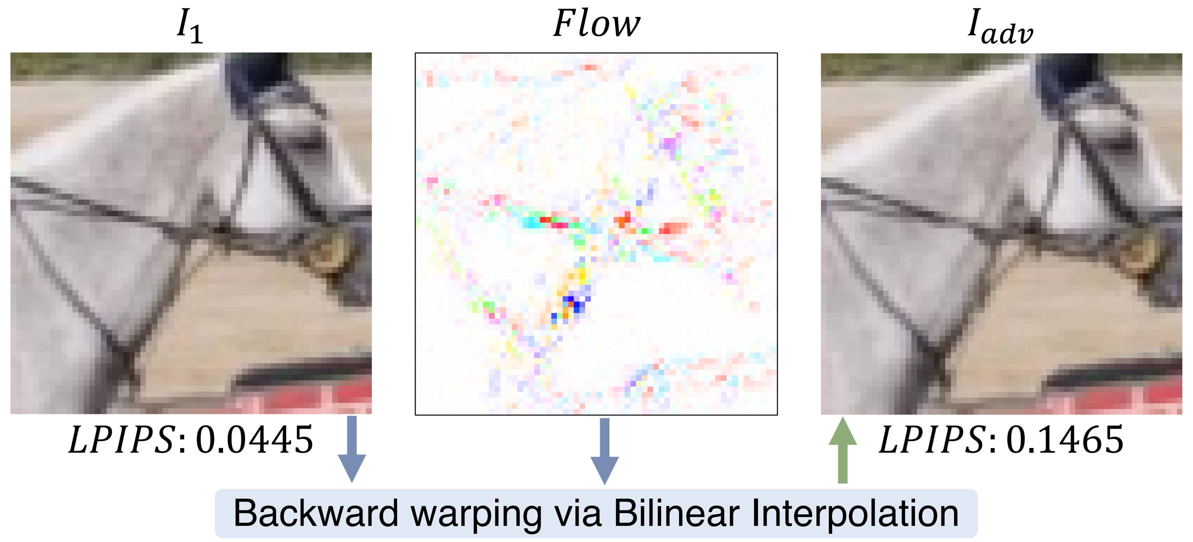
Perturbation attack (PGD) on LPIPS



$$J = ((s_{other}/(s_{other}+s_{adv})) - 1)^2$$

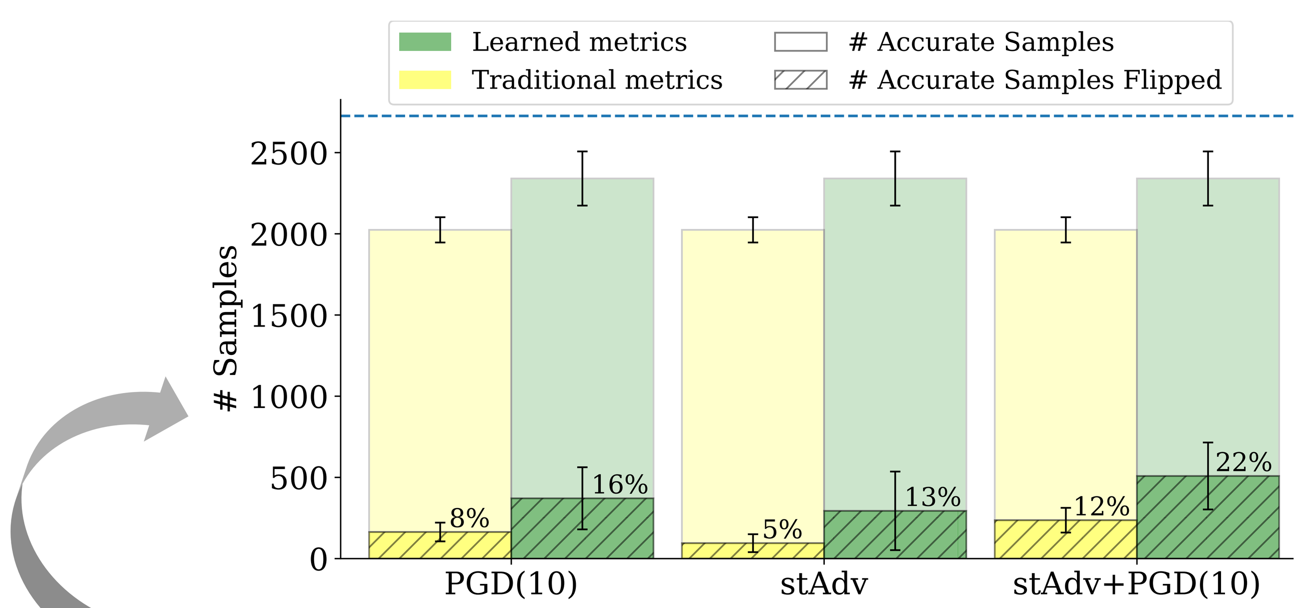
$$I^{t+1}_{adv} = P_c\,(I^t_{adv} + \alpha \cdot sign(\nabla_{I^t_{adv}} J(\theta, I^t_{adv}, I_{other}, I_{ref})))$$

Spatial Attack (stAdv) on LPIPS



Results

1. We successfully demonstrate that a wide variety of perceptual similarity metrics are susceptible to such imperceptible adversarial perturbations.
2. We attack the widely adopted LPIPS using the spatial attack stAdv to create adversarial examples and use them to benchmark the adversarial robustness of other similarity metrics.
3. Combining stAdv (spatial attack) with PGD (ℓ_∞-bounded attack) increases the transferability of the adv. samples.
4. Our investigations also show that although more accurate, learned perceptual similarity metrics may not be more robust than traditional ones.
5. Furthermore, we demonstrate the reverse of our attack (make the less similar distorted image more similar to the reference) and its applicability to higher resolution images.



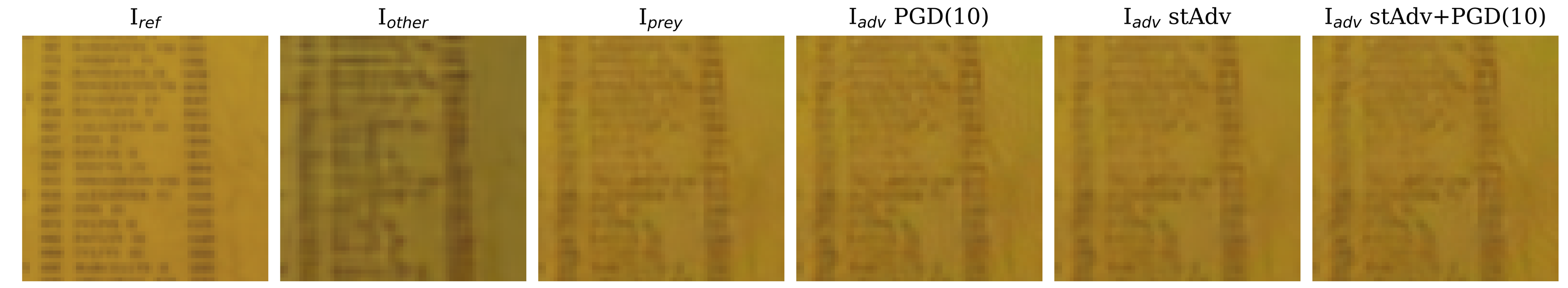
Network	#Accurate Samples	# Accurate Samples Flipped: PGD(10)	PGD(20)	stAdv	stAdv + PGD(5)	stAdv + PGD(10)	stAdv + PGD(15)	stAdv + PGD(20)
L2	2099/77%	101/5%	174/8%	77/4%	134/6%	189/9%	200/10%	257/12%
SSIM (Wang et al., 2004)	2093/77%	237/11%	442/21%	78/4%	180/9%	339/16%	370/18%	540/26%
MS-SSIM (Wang et al., 2003)	2022/74%	158/8%	256/13%	76/4%	162/8%	224/11%	234/12%	333/16%
CWSSIM (Wang & Simoncelli, 2005)	1883/69%	101/5%	172/9%	42/2%	60/3%	128/7%	139/7%	193/10%
FSIMc (Zhang et al., 2011)	2025/74%	222/11%	325/16%	202/10%	233/12%	302/15%	310/15%	393/19%
WaDIQaM-FR (Bosse et al., 2018)	2083/76%	95/5%	186/9%	59/3%	85/4%	146/7%	156/7%	238/11%
GTI-CNN (Ma et al., 2018)	1946/71%	448/23%	480/25%	494/25%	488/25%	504/26%	510/26%	543/28%
LPIPS(Squz.) (Zhang et al., 2018b)	2503/92%	298/12%	656/26%	114/5%	221/9%	519/21%	555/22%	886/35%
LPIPS(VGG) (Zhang et al., 2018b)	2317/85%	435/19%	814/35%	131/6%	288/12%	643/28%	685/30%	992/43%
E-LPIPS (Kettunen et al., 2019b)	2442/90%	503/21%	643/26%	517/21%	552/23%	641/26%	655/27%	817/33%
DISTS (Ding et al., 2020)	2413/89%	311/13%	576/24%	146/6%	257/11%	510/21%	546/23%	801/33%
Watson-DFT (Czolbe et al., 2020)	2179/80%	387/18%	614/28%	216/10%	324/15%	532/24%	562/26%	750/34%
PIM-1 (Bhardwaj et al., 2020)	2468/91%	696/28%	814/33%	756/31%	772/31%	826/33%	852/35%	958/39%
PIM-5 (Bhardwaj et al., 2020)	2457/90%	751/31%	844/34%	765/31%	791/32%	864/35%	893/36%	963/39%
A-DISTS (Ding et al., 2021)	2346/86%	339/14%	661/28%	164/7%	276/12%	561/24%	590/25%	850/36%
ST-LPIPS(Alex) (Ghildyal & Liu, 2022)	2470/91%	104/4%	198/8%	96/4%	123/5%	205/8%	212/9%	310/13%
ST-LPIPS(VGG) (Ghildyal & Liu, 2022)	2493/91%	210/8%	453/18%	103/4%	153/6%	321/13%	360/14%	576/23%
SwinIQA (Liu et al., 2022)	2310/85%	249/11%	357/15%	262/11%	279/12%	342/15%	375/16%	482/21%

Conclusion

The main contribution of this paper is the systematic investigation on whether existing perceptual similarity metrics are susceptible to invisible adversarial distortions. We suggest further research on this topic to investigate methods for mitigating these vulnerabilities.

Transferable Attack on Perceptual Similarity Metrics

	I_{ref}	I_{other}	I_{prey}	I_{adv} PGD(10)	I_{adv} stAdv	I_{adv} stAdv+PGD(10)
L2 ↓	0.0091	0.0127	0.0128	0.0128	0.0128
SSIM ↑	0.8754	0.8823	0.8721	0.8770	0.8635
FSIMc ↑	0.99069	0.99058	0.99061	0.99061	0.99064
WaDIQaM-FR ↓	1.2747	1.3567	1.3730	1.3622	1.3572
GTI-CNN ↓	135.61	255.97	220.48	217.10	217.65
DISTS ↓	0.0996	0.0729	0.0952	0.0873	0.1152
LPIPS(Squeeze) ↓	0.0736	0.0393	0.0421	0.0490	0.0517
LPIPS(VGG) ↓	0.0916	0.0669	0.0802	0.0783	0.1011
E-LPIPS ↓	0.0057	0.0041	0.0069	0.0068	0.0075
Watson-DFT ↓	908.63	922.66	1112.21	1071.77	1136.02
PIM-1 ↓	0.6141	0.4485	1.1852	1.2937	1.2917
PIM-5 ↓	6.2894	5.0282	11.3717	12.0675	12.2006



	I_{ref}	I_{other}	I_{prey}	I_{adv} PGD(10)	I_{adv} stAdv	I_{adv} stAdv+PGD(10)
L2 ↓	0.0361	0.0050	0.0057	0.0056	0.0063
SSIM ↑	0.3163	0.5807	0.5528	0.5646	0.5357
FSIMc ↑	0.98102	0.98274	0.98079	0.98016	0.97770
WaDIQaM-FR ↓	1.3614	1.2760	1.2575	1.2983	1.2943
GTI-CNN ↓	133.18	59.11	77.51	78.95	85.07
DISTS ↓	0.2772	0.2324	0.2739	0.2678	0.3021
LPIPS(Squeeze) ↓	0.0986	0.0761	0.1231	0.1058	0.1762
LPIPS(VGG) ↓	0.2167	0.1601	0.2451	0.2028	0.3269
E-LPIPS ↓	0.0115	0.0103	0.0169	0.0170	0.0178
Watson-DFT ↓	2433.66	1344.98	1415.91	1392.29	1410.53
PIM-1 ↓	2.9635	2.5469	3.2072	3.2161	3.5531
PIM-5 ↓	33.8370	27.0413	35.6628	37.6837	39.1791

	I_{ref}	I_{other}	I_{prey}	I_{adv} PGD(10)	I_{adv} stAdv	I_{adv} stAdv+PGD(10)
L2 ↓	0.0010	0.0010	0.0012	0.0012	0.0015
SSIM ↑	0.9739	0.9779	0.9730	0.9743	0.9681
FSIMc ↑	0.99992	0.99985	0.99983	0.99983	0.99980
WaDIQaM-FR ↓	1.1214	1.1190	1.1177	1.1165	1.1184
GTI-CNN ↓	47.72	11.53	79.21	85.79	84.42
DISTS ↓	0.1180	0.0065	0.0200	0.0129	0.0283
LPIPS(Squeeze) ↓	0.0023	0.0013	0.0025	0.0017	0.0033
LPIPS(VGG) ↓	0.0791	0.0027	0.0069	0.0038	0.0103
E-LPIPS ↓	0.0139	0.0002	0.0045	0.0047	0.0052
Watson-DFT ↓	924.09	541.48	783.71	693.21	861.64
PIM-1 ↓	0.7539	0.0110	1.0787	1.1750	1.1291
PIM-5 ↓	7.0737	0.1121	11.2964	12.0483	11.7169

	I_{ref}	I_{other}	I_{prey}	I_{adv} PGD(10)	I_{adv} stAdv	I_{adv} stAdv+PGD(10)
L2 ↓	0.0121	0.0133	0.0133	0.0133	0.0133
SSIM ↑	0.9068	0.9112	0.9006	0.9103	0.8958
FSIMc ↑	0.99392	0.99181	0.99185	0.99187	0.99183
WaDIQaM-FR ↓	1.1942	1.2634	1.2653	1.2699	1.2813
GTI-CNN ↓	53.66	28.88	62.75	61.31	69.90
DISTS ↓	0.1341	0.1034	0.1121	0.1056	0.1132
LPIPS(Squeeze) ↓	0.0264	0.0371	0.0395	0.0375	0.0411
LPIPS(VGG) ↓	0.0545	0.0462	0.0520	0.0472	0.0571
E-LPIPS ↓	0.0039	0.0033	0.0055	0.0054	0.0065
Watson-DFT ↓	1097.13	901.26	1147.84	1078.05	1157.19
PIM-1 ↓	0.2170	0.2429	1.0924	1.2546	1.2119
PIM-5 ↓	3.4366	2.9138	12.0777	13.0601	13.2696