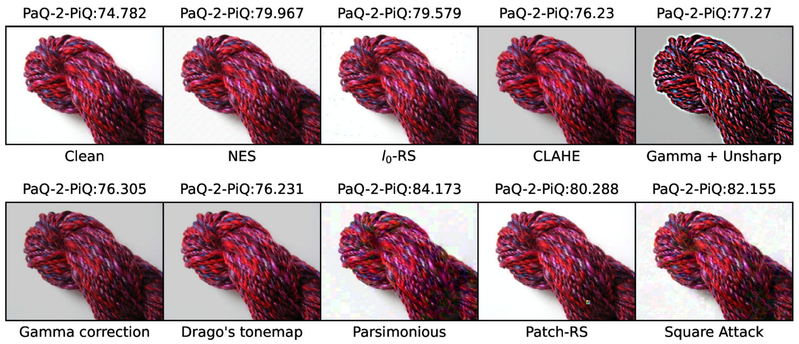
Black-box Universal Adversarial Perturbations for Image and Video Quality Assessment Methods

Georgii Bychkov, Sergey Lavrushkin, Dmitriy Vatolin

MSU Institute for Artificial Intelligence
Lomonosov Moscow State University

Motivation

Image- and video-quality assessment metrics (IQA and VQA) are widely used in benchmarks, comparisons and as a component of different algorithms. However, recent research [1] exposes their vulnerability to adversarial attacks. Although current research predominantly focuses on white-box adversarial attacks, black-box attacks remain underdeveloped despite their practical relevance, such as targeting metrics that are nondifferentiable or whose weights are unavailable. We make comprehensive robustness evaluation of popular image and video-quality metrics to black-box attacks and make the first known to authors attempt to attack IQA and VQA with black-box universal adversarial perturbations (UAPs). This approach greatly speeds up the process of attack application making necessary testing of their robustness to adversarial attacks.



Adversarial examples generated using adapted black-box adversarial attacks on PaQ-2-PiQ, applied to an image from KonIQ-10k.

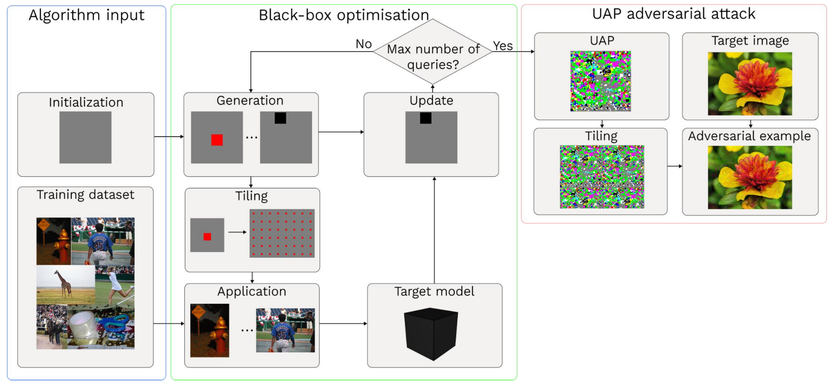
Contributions

1. Adaptation of 10 black-box attacks to IQA and VQA metrics.
2. New methods, based on universal adversarial perturbations (UAPs), that incorporate those black-box attacks, thereby addressing the primary limitation of such approaches: high computational complexity during inference.
3. Robustness evaluation of 18 common metrics against our proposed adversarial attacks.

Code availability

The code is available on GitHub:
https://github.com/georgebychkov/black-box-iqa

Proposed black-box UAP scheme



Scheme of UAP attack based on Square attack [2]

Experiments

To assess metric robustness, we use the normalized gain score:

$$Norm.gain = \frac{1}{n}\sum_{i=1}^{n}\frac{f(x_i^{adv}) - f(x_i^{clean})}{\text{range}(f)},$$

where range(f) is metric's output-value range,
x_i^{adv}, x_i^{clean} — attacked and clean images.

Datasets: KonIQ-10k, NIPS2017, Vimeo-90k and AGIQA-3k.

Subsampling: clustering using spatial information (SI), colorfulness (CF), and mean opinion score (MOS).

Method	NES	Parsimonious	Square Attack	Ran et al. [3]	Black-box UAP (proposed)
Computational complexity (FPS)	0.031	0.024	0.027	0.027	**~240**

Comparison of application speed of various black-box attacks

	No-reference metrics									Full-reference metrics								
	Koncept512	Linearity	PaQ-2-PiQ	VSFA	MANIQA	MDTVSFA	DB-CNN	UNIQUE	TOPIQ-NR	AHIQ	PieAPP	LPIPS AlexNet	LPIPS VGG	MS-SSIM	ASNA-MACS	DISTS	VMAF	VMAF NEG
Frame-RS	0.034	0.119	0.199	0.154	0.083	0.08	0.23	0.163	0.105	-0.003	0.014	-0.022	-0.045	-0.001	0.192	-0.019	0.0	-0.001
NES	**0.49**	0.372	0.304	0.203	0.132	0.105	0.46	**0.326**	0.353	-0.001	0.068	-0.003	-0.026	-0.004	0.136	-0.022	-0.012	-0.01
l_0-RS	0.337	0.255	0.323	0.141	0.199	0.082	0.34	0.178	0.172	-0.003	0.088	-0.007	-0.009	-0.001	0.155	0.011	0.025	0.004
CLAHE	0.07	0.02	0.13	0.024	0.036	0.008	0.066	0.024	0.016	0.011	-0.006	0.001	-0.0	-0.003	0.046	-0.0	0.304	-0.079
Gamma correction	0.028	0.035	0.011	0.008	0.024	0.007	0.054	0.052	0.049	-0.001	-0.0	-0.0	-0.0	**0.0**	0.002	0.0	0.0	0.0
Gamma + Unsharp	0.087	0.058	0.145	0.027	0.061	0.016	0.103	0.05	0.077	0.014	-0.001	0.002	**0.007**	-0.001	0.072	0.006	0.369	-0.025
Drago's tonemap	0.2	0.135	0.185	0.067	0.089	0.037	0.178	0.085	0.13	**0.028**	**0.393**	-0.002	-0.0	-0.01	0.041	-0.001	**0.392**	-0.007
Parsimonious	0.311	0.288	0.425	**0.329**	**0.233**	**0.131**	0.428	0.264	0.254	-0.057	0.284	-0.015	-0.023	-0.015	**0.485**	0.009	0.013	-0.036
Patch-RS	0.376	0.184	0.295	0.08	0.077	0.052	0.304	0.142	0.128	-0.003	0.032	-0.0	-0.0	-0.0	0.021	0.003	0.056	0.0
Square Attack	0.365	0.27	0.441	0.319	0.229	0.127	0.436	0.263	0.231	0.001	0.207	**0.003**	0.005	**0.0**	0.466	**0.028**	0.032	**0.006**
Ran et al.	0.381	**0.419**	**0.509**	-0.037	-0.05	-0.007	**0.614**	0.321	**0.386**	-0.106	-0.304	-0.126	-0.219	-0.137	-0.139	-0.192	-0.146	-0.073

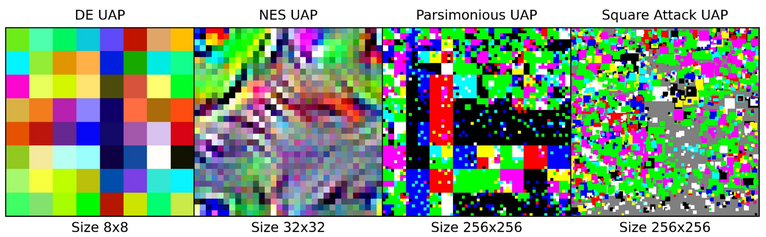
Evaluation of adapted black-box attacks for quality metrics. KonIQ-10k dataset

	No-reference metrics									Full-reference metrics								
	Koncept512	Linearity	PaQ-2-PiQ	VSFA	MANIQA	MDTVSFA	DB-CNN	UNIQUE	TOPIQ-NR	AHIQ	PieAPP	LPIPS AlexNet	LPIPS VGG	MS-SSIM	ASNA-MACS	DISTS	VMAF	VMAF NEG
DE UAP	-0.05	**0.132**	0.263	0.002	-0.082	**0.024**	0.186	**0.112**	-0.014	**0.008**	-0.057	-0.044	-0.174	-0.019	**0.008**	-0.156	-0.025	-0.014
NES UAP	**0.096**	-0.001	**0.55**	0.012	-0.012	0.007	0.057	0.074	-0.002	-0.001	-0.005	-0.0	-0.002	-0.0	-0.001	**-0.0**	-0.001	-0.0
Parsimonious UAP	0.042	-0.017	0.269	0.01	-0.012	0.013	0.119	-0.015	-0.006	-0.076	**0.015**	-0.02	-0.026	-0.013	-0.037	-0.035	-0.049	-0.047
Square Attack UAP	0.053	0.112	0.364	**0.045**	**0.016**	0.02	**0.264**	0.022	**0.017**	-0.003	-0.006	**-0.0**	**-0.0**	**0.0**	-0.064	-0.021	**-0.0**	**0.0**
Ran et al. UAP	-0.095	-0.013	0.206	-0.066	-0.056	-0.027	0.11	-0.076	-0.061	-0.172	-0.377	-0.38	-0.49	-0.241	-0.579	-0.533	-0.315	-0.259
NES	0.596	0.395	0.362	0.212	0.157	0.11	0.483	0.334	0.38	-0.002	0.068	-0.004	-0.03	-0.004	0.149	-0.026	-0.014	-0.011
Parsimonious	0.387	0.385	0.564	0.358	0.278	0.138	0.496	0.275	0.3	-0.099	0.409	-0.04	-0.062	-0.042	0.604	-0.037	-0.048	-0.105
Square Attack	0.429	0.358	0.609	0.35	0.326	0.127	0.466	0.266	0.251	-0.003	0.23	0.001	0.002	0.0	0.487	0.017	0.04	0.003
Ran et al.	0.467	0.508	0.681	-0.041	-0.097	-0.011	0.629	0.34	0.494	-0.169	-0.529	-0.207	-0.306	-0.3	-0.222	-0.291	-0.247	-0.216

Comparison of UAP attacks with base attacks with the same bounds. KonIQ-10k dataset

Results

All tested no-reference metrics are vulnerable to black-box adversarial attacks. The most robust models were MDTVSFA (several training datasets) and MANIQA (ViT). Black-box UAP was able to attack all no-reference metrics, increasing the output values by **10%** for most models, indicating the necessity to check their robustness to such attacks. The most robust models were VSFA, MDTVSFA, MANIQA and TOPIQ-NR due to their architectures.
Most full-reference metrics were resistant to black-box attacks with the exception of VMAF, PieAPP, ASNA-MACS.



Examples of generated UAP for PaQ-2-PiQ metric

References

[1] Shumitskaya, E., Antsiferova, A., Vatolin, D, Universal perturbation attack on differentiable no-reference image-and video-quality metrics, *BMVC*, 2022
[2] Andriushchenko, M., Croce, F., Flammarion, N., Hein, M., Square attack: a query-efficient black-box adversarial attack via random search, *ECCV*, 2020
[3] Ran, Y., Zhang, A. X., Li, M., Tang, W., Wang, Y. G., Black-box adversarial attacks against image quality assessment models, *Expert Systems with Applications*, 2025