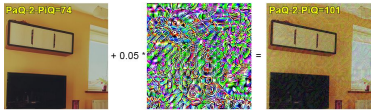
Universal Perturbation Attack on Differentiable No-Reference Image- and Video-Quality Metrics

Ekaterina Shumitskaya, Anastasia Antsiferova, Dmitriy Vatolin

Lomonosov Moscow State University, Moscow, RU
MSU Institute for Artificial Intelligence Moscow, RU

Motivation

No-reference (NR) image- and video-quality metrics are widely used for video processing tasks including different algorithms comparison. Comparison participants can dishonestly increase the metric scores by attacks to get better results. The goal of attacks on quality metric is to increase the quality score of an output image, when visual quality does not improve after the attack. We make the first attempt in attacking differentiable no-reference (NR) image- and video-quality metrics through universal adversarial perturbations (UAPs). We use UAP attack as quickest attack since we are investigating the possibility of injecting attacks on quality metrics in video compression and processing algorithms



Example of attack on NR quality metric PaQ-2-PiQ

Contributions

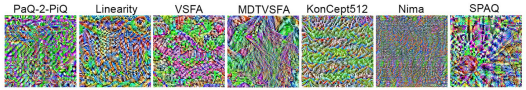
1. We employed a universal perturbation attack against seven differentiable NR metrics (PaQ-2-PiQ, Linearity, VSFA, MDTVSFA, KonCept512, Nima and SPAQ)
2. We applied trained UAPs to Full-HD video frames before compression and proposed a method for comparing metrics stability based on RD curves to identify metrics that are the most resistant to UAP attack

Code availability

The code is available on GitHub:
https://github.com/katiashh/UAP_Attack_on_Quality_Metrics

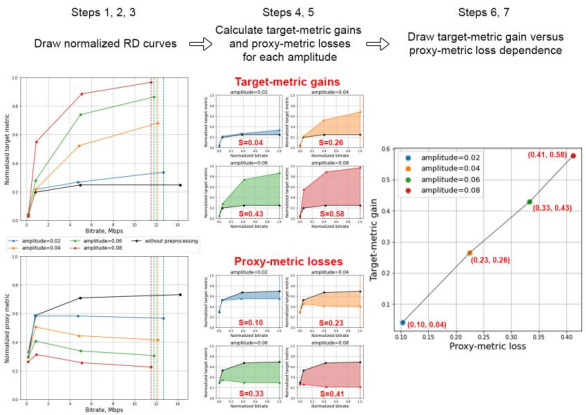
Proposed Method

1. UAP attack: we trained UAP on the dataset of 256x256 images from COCO dataset



Trained universal perturbations for all tested metrics

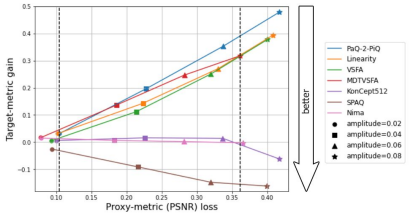
2. Stability score: we applied UAP as preprocessing to the frames of 20 FullHD videos taken from Xiph.org datatet, compress original and attacked videos with four bitrates (200k, 1M, 5M and 12M) and on the basis of RD curver calculate stability score as area under target-metric gain versus proxy-metric loss dependencies multiplied by -100



Calculation of target-metric gain and proxy-metric loss using normalized RD curves for a video at four amplitude levels

Results

SPAQ, Nima and KonCept512 proved to be resistant to UAP attack, while PaQ-2- PiQ, Linearity, VSFA and MDTVSFA proved vulnerable. We recommend the proposed method as an additional verification of NR metric reliability to complement traditional subjective tests and benchmarks.



Target-metric gain versus proxy-metric loss dependencies for all tested NR metrics

PaQ-2-PiQ	MDTVSFA	Linearity	VSFA	KonCept512	Nima	SPAQ
-5.3	-4.8	-4.2	-3.8	**-0.3**	**-0.1**	**2.6**

Stability scores (↑) for all target NR metrics

References

[1] Moosavi-Dezfooli et al. (2017). “Universal adversarial perturbations.” In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
[2] Wang et al. (2008). “Maximum differentiation (mad) competition: A methodology for comparing computational models of perceptual quantities.” In: Journal of Vision, 8(12):8–8
[3] Liu et al. (2016). “Software to stress test image quality estimators.” In: Eighth International Conference on Quality of Multimedia Experience (QoMEX), pages 1–6