FAST ADVERSARIAL CNN-BASED PERTURBATION ATTACK ON NO-REFERENCE IMAGE QUALITY METRICS

Ekaterina Shumitskaya[1], **Anastasia Antsiferova**[2,3] **& Dmitriy Vatolin**[1,2,3]
[1] Lomonosov Moscow State University, Moscow, Russian Federation
[2] ISP RAS Research Center for Trusted Artificial Intelligence, Moscow, Russian Federation
[3] MSU Institute for Artificial Intelligence, Moscow, Russian Federation
{ekaterina.shumitskaya, aantsiferova, dmitriy}@graphics.cs.msu.ru

ABSTRACT

Modern neural-network-based no-reference image- and video-quality metrics exhibit performance as high as full-reference metrics. These metrics are widely used to improve visual quality in computer vision methods and compare video processing methods. However, these metrics are not stable to traditional adversarial attacks, which can cause incorrect results. Our goal is to investigate the boundaries of no-reference metrics applicability, and in this paper, we propose a fast adversarial perturbation attack on no-reference quality metrics. The proposed attack (FACPA) can be exploited as a preprocessing step in real-time video processing and compression algorithms. This research can yield insights to further aid in designing of stable neural-network-based no-reference quality metrics.

1 INTRODUCTION

State-of-the-art no-reference (NR) metrics commonly use neural-network-based approaches and deliver more accuracy than traditional ones. Quality-metric attacks lead such metrics to increase the quality score without improving visual quality. The developers of image- or video-processing and compression algorithms can exploit that vulnerability. For example, the developers of libaom (Deng et al., 2020) exploited the vulnerability of VMAF (VMAF: Perceptual video quality assessment based on multi-method fusion) by implementing the VMAF-oriented tuning option in the encoder. Our research aims to investigate the possibility of injecting quality-metric attacks into real-time algorithms, which can cause incorrect results in benchmarks. In this paper, we propose a Fast Adversarial CNN-based Perturbation Attack (FACPA) on NR quality metrics which is much faster than previous iterative attack methods and more effective than universal perturbation methods. Using this method, researchers can estimate the potential cheating gain of any NR metric used in benchmarks and video processing algorithms.

2 RELATED WORK

Existing attack methods mainly focus on attacking image classification models (Szegedy et al., 2014), (Goodfellow et al., 2015), (Kurakin et al., 2017), (Carlini & Wagner, 2017), (Madry et al., 2018). Several studies on adversarial attacks related to NR quality metrics have been published. Sang et al. (2022) proposed an iterative algorithm based on adaptive distortions to attack quality metrics. This algorithm is based on the MI-FGSM (Dong et al., 2018) and controls distortions using the NR metric NIQE (Mittal et al., 2012). Korhonen & You (2022) proposed a method for the iterative attack of NR quality metrics that employs the Sobel filter to generate distortions on object edges since it allows to improve the visual quality of attacked images. Zhang et al. (2022) proposed an iterative attack to craft adversarial images using different full-reference metrics to control visual distortions. However, all these attacks are iterative and time-consuming, so their injection into real-time algorithms is unlikely. Shumitskaya et al. (2022) proposed to craft a universal adversarial perturbation to attack an NR image/video quality metric. This approach is much faster than iterative methods. However, this attack does not use information about the input and hence does not use the full potential of the attack.

3 PROPOSED METHOD AND RESULTS

Let x_i be images from the training set of N samples, M — NR image/video quality metric to be attacked (e.g. MDTVSFA, Linearity, etc.). The proposed adversarial attack $f(x)$ can be formulated as follows:

$$\underset{f}{\operatorname{argmax}}\{\frac{1}{N}\sum_{i=1}^{N}(M(x_i + f(x_i)) - M(x_i)))\}, \forall x_i : \|f(x_i)\|_\infty \leq \epsilon, \tag{1}$$

To find a function $f(x)$, we approximated it using U-Net encoder-decoder architecture with the tanh activation and scaling. We tried different approaches and chose U-Net architecture as it showed the best performance on preliminary tests. Constraint ϵ provided scaling. We used ϵ equal to $\frac{10}{255}$. For each target metric, we trained specific CNN weights. For CNNs training, we used 10,000 256×256 images from the COCO dataset (Lin et al., 2014), defined the loss function as the target metric with the opposite sign and employed Adam optimizer (Kingma & Ba, 2014).

We compared the proposed attack with previous attacks on quality metrics using images from NIPS 2017: Adversarial Learning Development Set (Alex K, 2017). We attacked three quality metrics (Linearity (Li et al., 2020), VSFA (Li et al., 2019) and MDTVSFA (Li et al., 2021)) and compared the achieved increase of target metrics scores. For iterative attacks, we used a learning rate of 0.001 and attacked metrics using a different amount of iterations. Table 1 contains comparison results. FACPA is much faster than iterative methods (15 ms versus 110 – 5,000 ms) and more effective than the universal perturbation method. Figure 1 shows attack examples for the Linearity metric for compared methods.

Table 1: Percentage of metrics increase by proposed and other attacks and GPU calculation time.

Attack method	It.	Linearity		VSFA		MDTVSFA	
		Gain ↑	Time ↓	Gain ↑	Time ↓	Gain ↑	Time ↓
Zhang et al. (2022)	1	5.9%	173 ms	11.7%	626 ms	21.3%	631 ms
	10	26.1%	1590 ms	30.9%	2660 ms	**47.9**%	5980 ms
Korhonen & You (2022)	1	5.6%	111 ms	11.1%	483 ms	20.4%	505 ms
	10	21.0%	992 ms	**41.7**%	4750 ms	39.3%	4040 ms
Sang et al. (2022)	1	6.5%	153 ms	4.4%	535 ms	3.7%	540 ms
	3	18.6%	358 ms	32.6%	1580 ms	39.2%	1580 ms
Shumitskaya et al. (2022)	-	22.9%	**11 ms**	22.0%	**11 ms**	29.4%	**11 ms**
FACPA (ours)	-	**32.0**%	15 ms	30.6%	15 ms	40.3%	15 ms

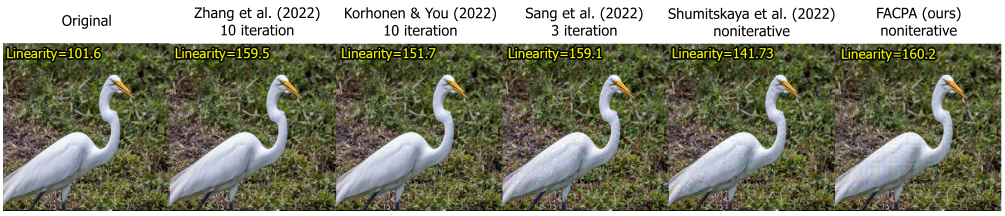

Figure 1: Attack examples for the Linearity quality metric.

4 CONCLUSION

In this paper, we proposed a fast CNN-based attack that increases the scores of NR quality metrics. The comparison with previous methods showed that our attack is much faster than iterative methods and more efficient than the universal adversarial perturbation. The speed of the proposed attack showed that it can be injected into video compression and other real-time algorithms. Therefore proposed attack can serve as an additional verification of metric reliability. Our code is publicly available at `https://github.com/katiashh/FACPA`.

URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2023 Tiny Papers Track.

ACKNOWLEDGEMENTS

The work received support through a grant for research centers in the field of artificial intelligence (agreement identifier 000000D730321P5Q0002, dated November 2, 2021, no. 70-2021-00142 with the Ivannikov Institute for System Programming of the Russian Academy of Sciences).

REFERENCES

Ian Goodfellow Alex K, Ben Hamner. Nips 2017: Defense against adversarial attack, 2017. URL https://kaggle.com/competitions/nips-2017-defense-against-adversarial-attack.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. Ieee, 2017.

Sai Deng, Jingning Han, and Yaowu Xu. Vmaf based rate-distortion optimization for video coding. In *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6. IEEE, 2020.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9185–9193, 2018.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6572.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Jari Korhonen and Junyong You. Adversarial attacks against blind image quality assessment models. In *Proceedings of the 2nd Workshop on Quality of Experience in Visual Multimedia Applications*, pp. 3–11, 2022.

Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=HJGU3Rodl.

Dingquan Li, Tingting Jiang, and Ming Jiang. Quality assessment of in-the-wild videos. In *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 2351–2359, 2019.

Dingquan Li, Tingting Jiang, and Ming Jiang. Norm-in-norm loss with faster convergence and better performance for image quality assessment. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 789–797, 2020.

Dingquan Li, Tingting Jiang, and Ming Jiang. Unified quality assessment of in-the-wild videos with mixed datasets training. *International Journal of Computer Vision*, 129(4):1238–1257, 2021.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

Qingbing Sang, Hongguo Zhang, Lixiong Liu, Xiaojun Wu, and Alan Bovik. On the generation of adversarial samples for image quality assessment. *Available at SSRN 4112969*, 2022.

Ekaterina Shumitskaya, Anastasia Antsiferova, and Dmitriy S Vatolin. Universal perturbation attack on differentiable no-reference image- and video-quality metrics. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022. URL https://bmvc2022.mpi-inf.mpg.de/0790.pdf.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6199.

VMAF: Perceptual video quality assessment based on multi-method fusion. Vmaf: Perceptual video quality assessment based on multi-method fusion, 2017. URL https://github.com/Netflix/vmaf.

Weixia Zhang, Dingquan Li, Xiongkuo Min, Guangtao Zhai, Guodong Guo, Xiaokang Yang, and Kede Ma. Perceptual attacks of no-reference image quality models with human-in-the-loop. *arXiv preprint arXiv:2210.00933*, 2022.