Defense Against Adversarial Attacks on No-Reference Image Quality Models with Gradient Norm Regularization

Yujia Liu*, Chenxi Yang*, Dingquan Li, Jianhao Ding, Tingting Jiang

NERCVT PKU, National Key Laboratory for Multimedia Information Processing PKU, SMS PKU, Peng Cheng Laboratory

{yujai_liu, dingquanli, ttjiang}@pku.edu.cn; {yangchenxi, djh01998}@stu.pku.edu.cn

PEKING UNIVERSITY

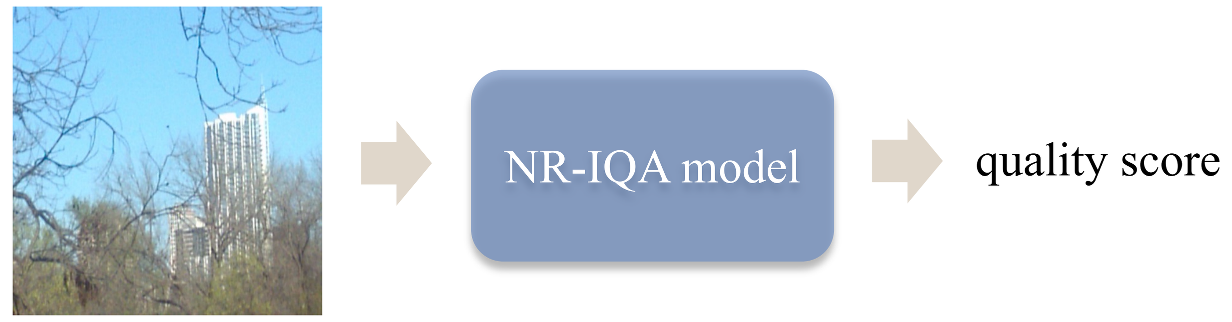
paper

code

北京大学数学科学学院 School of Mathematical Sciences,Peking University

No-Reference Image Quality Assessment

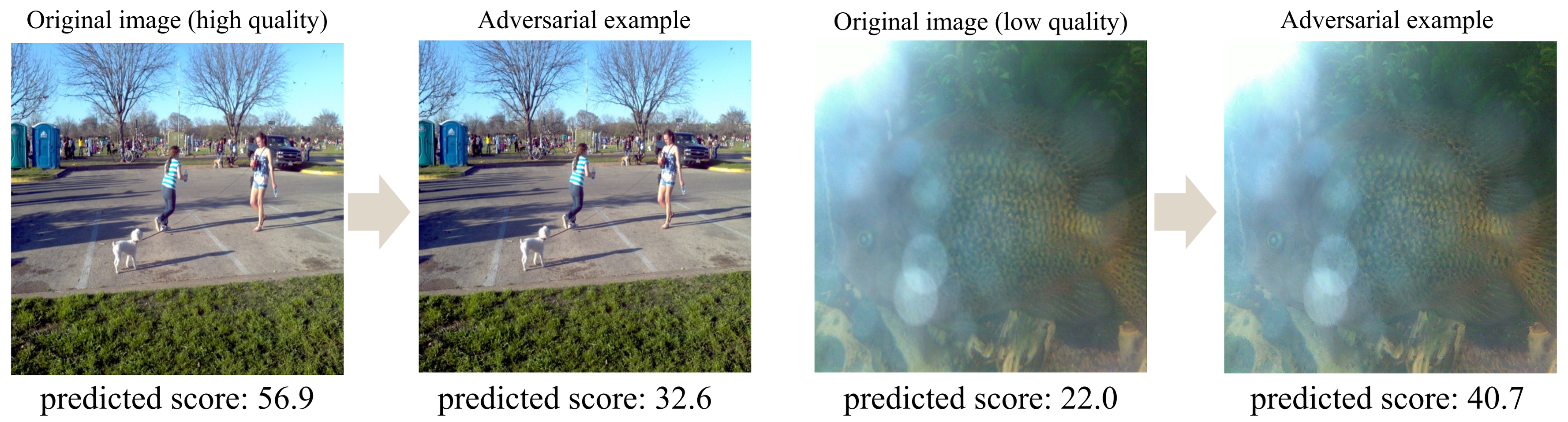
- NR-IQA models: predict the quality score of an image without reference.



- Applications: media industry, performance evaluation, image compression and so on.

Motivation

- NR-IQA models are vulnerable to adversarial attacks, and no IQA-specific defense methods have been explored.



small changes to humans, **large** changes in scores

- The robustness of NR-IQA models is related to the gradient norm.

Problem Definition

- Adversarial attacks on NR-IQA can be described as:

$$\max |f(x+\delta)-f(x)|, \quad \text{s.t. } D(x+\delta, x) \le \varepsilon,$$

f: an NR-IQA model x: an input image δ: perturbation

$D(\cdot,\cdot)$: perceptual distance between two images

ε: the tolerance of human eyes for image differences

Methodology

Why to regularize gradient norm?

- The magnitude of changes in predicted scores can be approximated by $\|\nabla_x f\|_1$ when δ is ℓ_∞-bounded.

Theorem 1. *Suppose f represents an NR-IQA model, ε is the strength of an attack, then*

$$\sup_{\delta:\|\delta\|_\infty \leqslant \varepsilon} |f(x+\delta)-f(x)| \approx \varepsilon \| \nabla_x f(x) \|_1$$

Proof. Taylor expansion

$$f(x+\delta) \approx f(x) + \delta^T \nabla_x f(x) \Rightarrow |f(x+\delta)-f(x)| \approx |\delta^T \nabla_x f(x)|$$

$|\delta^T \nabla_x f(x)|$ is maximized when $\delta = \epsilon \cdot \operatorname{sign}(\nabla_x f)$ ∎

How to regularize gradient norm?

- Directly add the gradient norm into the loss function? **No**

$$L(f,x) = L_{IQA}(f,x) + \lambda \cdot \| \nabla_x f(x) \|_1^2$$ ***Double backpropagation !***

- Finite difference

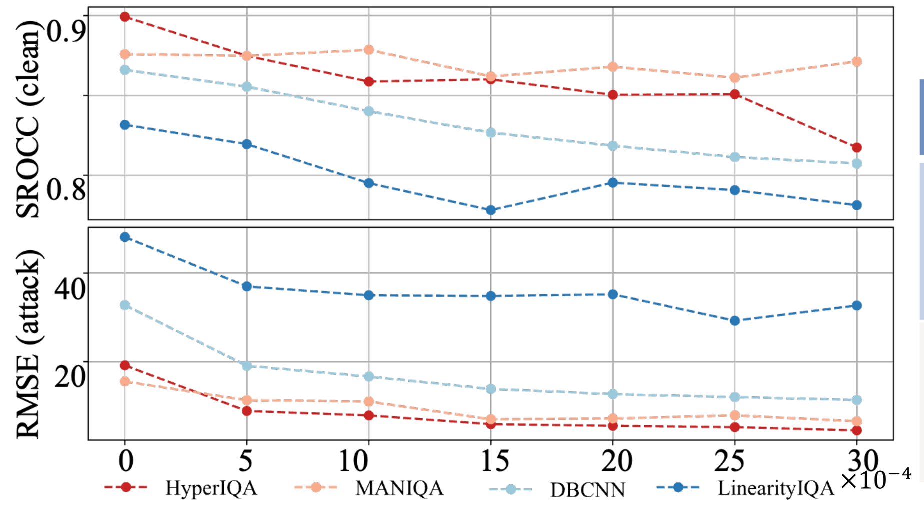
$$\| \nabla_x f(x) \|_1 \approx \left| \frac{f(x+h\cdot d)-f(x)}{h} \right|$$

$h \in \mathbb{R}^+$: small step size
$d = \operatorname{sign}(\nabla_x f)$

- Norm regularization Training strategy (NT) for robust NR-IQA models:

$$L(f,x) = L_{IQA}(f,x) + \lambda \cdot \left| \frac{f(x+h\cdot d)-f(x)}{h} \right|^2.$$

Ablation Studies



Ablation study of λ

Ablation study of h on DBCNN

	h	0.001	0.01	0.1	1
Clean	SROCC↑	0.788	0.856	0.846	0.844
	RMSE↓	16.099	14.138	12.417	14.809
Attack	SROCC↑	0.577	0.200	-0.383	-0.441
	RMSE↓	7.356	19.065	28.785	18.767

Experiments on the LIVEC dataset

- Performance on clean images

Performance calculated between **predicted scores & MOS** (baseline / +NT)

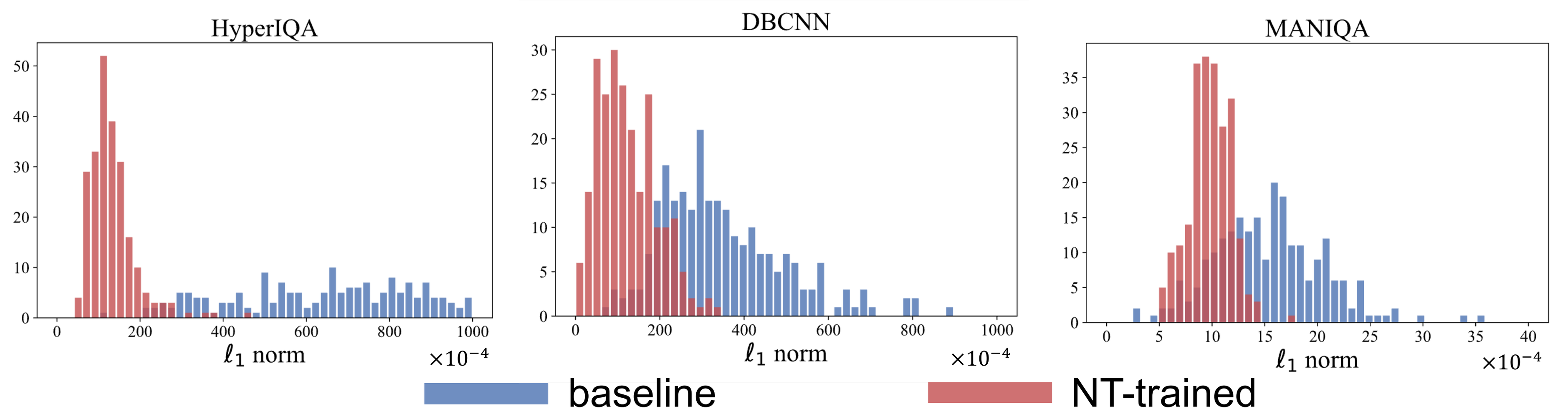
	HyperIQA	DBCNN	LinearityIQA	MANIQA
RMSE↓	**9.913** / 12.575	**10.897** / 13.140	**12.730** / 13.173	26.082 / **23.830**
SROCC↑	**0.899** / 0.859	**0.866** / 0.856	**0.832** / 0.820	**0.876** / 0.871

- Robustness improvement

RMSE calculated between **scores before & after attack** (baseline / +NT)

Attacks		HyperIQA	DBCNN	LinearityIQA	MANIQA
White-box	FGSM	19.174 / **7.885**	32.778 / **19.065**	48.128 / **36.988**	15.549 / **6.562**
	Perceptual	6.360 / **0.130**	63.991 / **14.524**	115.732 / **80.857**	**0.079** / 0.189
Black-box	UAP	10.583 / **8.131**	14.833 / **10.922**	20.813 / **19.434**	5.795 / **5.592**
	Kor	13.698 / **10.107**	6.514 / **5.298**	14.807 / **12.407**	7.759 / **6.680**

- Norm reduction



Conclusion

- In theory, prove that the score changes of NR-IQA models are related to the ℓ_1 norm of the gradient.
- In practice, apply the theory to improve the robustness of NR-IQA models.

Future Works

- More explorations on Full-Reference IQA models.
- More effective defense on SROCC, PLCC and KROCC.
- Less performance drop on clean images.