Разработка универсальных атакующих надбавок черного ящика для методов оценки качества изображений и видео

Бычков Георгий Константинович
Студент
Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия
E-mail: georgy.bychkov@graphics.cs.msu.ru

Научный руководитель *— Ватолин Дмитрий Сергеевич*

Большинство методов оценки качества изображений и видео, основанных на нейронных сетях, демонстрируют более высокую корреляцию с субъективными оценками, чем классические методы и активно используются в различных бенчмарках и сравнениях [1]. Однако недавние исследования выявили их значительную уязвимость к состязательным атакам [3]. Эти атаки манипулируют входными данными для увеличения значений методов оценки качества без соответствующего улучшения визуального качества изображения. В то время как текущие исследования в основном сосредоточены на атаках белого ящика, атаки черного ящика остаются недостаточно изученными, несмотря на их практическую значимость, например, использование при атаках на недифференцируемые модели или в ситуациях с доступом только к значениям модели.

В данной работе исследуется уязвимость современных методов оценки качества изображений и видео к 10 адаптированным состязательным атакам черного ящика, изначально разработанных для классификаторов изображений. Эти атаки были протестированы на 9 безэталонных (NR) метриках и 9 эталонных (FR) метриках, продемонстрировав их подверженность состязательным атакам. Все безреференсные метрики были успешно атакованы с помощью методов черного ящика, однако полнореференсные метрики продемонстрировали большую устойчивость к атакам и не все были взломаны.

Однако практическая польза от этих атак ограничена их вычислительной сложностью. Чтобы решить эту проблему, предложен метод обучения универсальных атакующих возмущений (UAP) [2,3] для 5 наиболее успешных атаках черного ящика. Для этого универсальное возмущение обучается на отдельном обучающем наборе

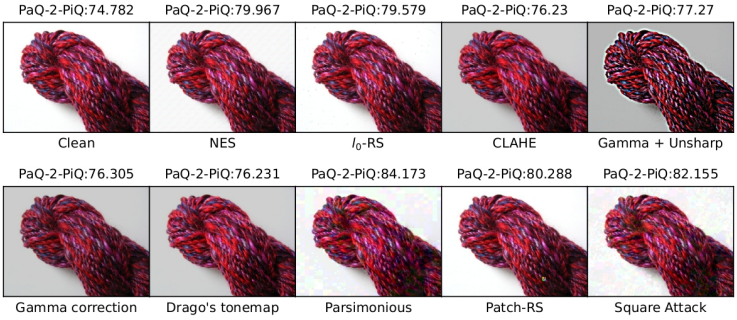
данных со следующей функцией потерь:

$$P = \begin{cases} \underset{P:||P||_\infty \leq \varepsilon}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} J^{\mathrm{NR}}(\mathrm{clip}(x_i + P, 0, 1)) & \text{для NR метрик,} \\ \underset{P:||P||_\infty \leq \varepsilon}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} J^{\mathrm{FR}}(\mathrm{clip}(x_i + P, 0, 1), x_i^{ref}) & \text{для FR метрик,} \end{cases}$$

где n- размер набора данных, $J^{\mathrm{FR}}, J^{\mathrm{NR}}$ - функции потерь, используемые для обучения состязательных атак на FR и NR метрики.

Экспериментальная оценка предлагаемых методов UAP показывает, что хотя эти методы и демонстрируют меньшую силу атаки по сравнению с оригинальными методами, все же эффективно повышают метрики и пригодны для приложений в реальном времени.

Иллюстрации



Влияние состязательных атак на метрику PaQ-2-PiQ

Литература

1. Huang Z. et al. Vbench: Comprehensive benchmark suite for video generative models //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. – 2024. – С. 21807-21818.
2. Shafahi A. et al. Universal adversarial training //Proceedings of the AAAI Conference on Artificial Intelligence. – 2020. – Т. 34. – №. 04. – С. 5636-5643.
3. Shumitskaya E., Antsiferova A., Vatolin D. Universal perturbation attack on differentiable no-reference image-and video-quality metrics //arXiv preprint arXiv:2211.00366. – 2022.