Анализ алгоритмов защиты методов оценки качества изображений от состязательных атак

Гущин Александр Евгеньевич

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: alexander.gushchin@graphics.cs.msu.ru

Научный руководитель *— Ватолин Дмитрий Сергеевич*

Современные методы оценки качества изображений основанные на нейронных сетях показывают лучшее качество в сравнении с традиционными методами. Однако, нейронные сети как класс моделей уязвимы для атак злоумышленников, которые могут исказить показания моделей. Состязательные атаки в этом классе задач стремятся повысить значения методов оценки без улучшения визуального качества самих изображений. Таким образом, методы завышают показатель визуального качества и становятся непригодными для использования. Для противодействия состязательным атакам к моделям применяются алгоритмы защит.

На текущий момент широко исследуется устойчивость различных методов компьютерного зрения к состязательным атакам (таких, как классификаторы изображений, детекторы объектов). Однако, в области методов оценки качества исследования до сих пор широко не проводились. В данной работе исследуются эти алгоритмы защит методов оценки качества от состязательных атак. По результатам работы выявлены лучшие алгоритмы для различных типов состязательных атак, таких как FGSM [1], MADC[2], UAP, AdvCF[3] и других.

Одним из наиболее распространенных методов защиты от состязательных атак является очищение атакованного изображения. Метод получается на вход изображение, которое содержит атакующую добавку. Выходными данными является очищенное изображение, на котором влияние атакующей добавки нивелировано.

Методы защиты оцениваются по двум критериям: качество удаления атакующей добавки и сходство выходного изображения с оригинальным (до атаки). Качество удаления считается как мера близости ответа модели на очищенном изображении к ответу модели на оригинальном изображении. Сходство выходного изображения с оригинальным считается с помощью методов PSNR и SSIM. Также для оценки качества очищенного изображения было проведено по-

парное субъективное сравнение на платформе Subjectify.com с участием более 200 респондентов.

Иллюстрации

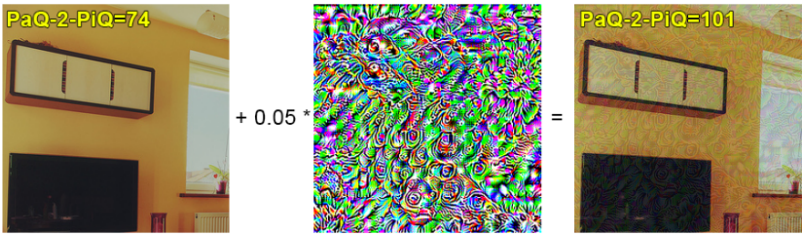

Рис 1. Пример атакованного изображения.

В данной работе было впервые в литературе протестировано 15+ методов защиты моделей оценки качества изображений. Были протестированы как простые методы предобработки (размытие, изменение разрешения, сжатие), так и более сложные методы с использованием нейронных сетей (диффузионные методы, методы очищения шума).

Экспериментальная оценка выявила лучшие методы защиты, которые помогут разработчикам моделей повысить их устойчивость к состязательным атакам.

Литература

1. Hong X. Deep Fusion Network for Image Completion // In Proceedings of the 27th ACM International Conference on Multimedia, 2019, P. 2033–2042.
2. Ronneberger O. U-Net: Convolutional Networks for Biomedical Image Segmentation // In International Conference on Medical image computing and computer-assisted intervention, 2015, P. 234–241.
3. Wang Z. Image quality assessment: from error visibility to structural similarity // In IEEE transactions on image processing, 2004, P. 600–612.