**УДК 004.056.2**

# Подходы к устранению вредоносных возмущений посредством визуальных искажений и шумоподавления

**Керимбай А.** (Университет ИТМО), **Пузикова Я.В.** (Университет ИТМО), **Есипов Д.А.** (Университет ИТМО)
**Научный руководитель – аспирант, ассистент ФБИТ Роговой В.** (Университет ИТМО)

**Введение.** Благодаря цифровизации системы обработки изображений получили широкое распространение во всех сферах человеческой деятельности ввиду их высокой эффективности. Их активно внедряют в системы городского видеонаблюдения, контроля и управления доступом, беспилотные транспортные средства, а также в медицину для постановки более точного диагноза. Все перечисленные области на данном этапе курируются человеком, однако, учитывая человеческий фактор, людям в помощь или на смену приходят системы компьютерного зрения. В то же время системы искусственного интеллекта связаны с рядом угроз, описанным в том числе в БДУ УБИ ФСТЭК [1]. Реализация некоторых угроз злоумышленниками может привести к нарушению работы систем искусственного интеллекта и машинного обучения, что может повлечь за собой значительные убытки для компаний, а также угрозу безопасности людей, взаимодействующих с данной системой, что особенно актуально для беспилотного транспорта.

Для дальнейшего развития и распространения систем искусственного интеллекта, а также повышения доверия к ним, необходимо разработать соответствующие механизмы обеспечения безопасности от актуальных угроз. Одним из значимых методов реализации угроз являются состязательные атаки [2-4]. Основные методы защиты предполагают некоторые модификации входного изображения: изменение размерности, поворот, внесение случайных изменений [5].

Следует отметить, что существующие методы имеют ряд ограничений в применении и нуждаются в доработке. При обеспечении высокого уровня безопасности, методы нередко влияют на целевую модель и ее паттерны распознавания, тогда как менее инвазивные методы демонстрируют слабые защитные свойства и имеют различные ограничения.

**Основная часть.** Учитывая вопросы экономической эффективности и затрачиваемых ресурсов, наиболее эффективным решением может быть предобработка входных данных. Состязательные шумовые атаки предполагают внесение в изображение некоторых возмущений, тогда внесение дополнительного искажения потенциально позволит нарушить целостность вредоносной маски, что в свою очередь позволит избежать атаки.

Для предобработки данных возможно использование шумоподавления и внесения дополнительных визуальных искажений, причем для последних следует отметить, что возможно применение как актуальных, так и неактуальных для целевой системы искажений. При обработке изображений реальных (физических) объектов в таких системах, как беспилотный транспорт и видеонаблюдение, актуальными искажениями будут засвечивание, размытие и затемнение, к неактуальным может быть отнесено, например, обесцвечивание. Выбор искажений зависит от целевой системы. Поскольку внесение искажений также будет оказывать влияние на модель, возможны два подхода к дальнейшей обработке изображения:

1) устранение внесенного искажения;
2) обучение целевой модели обработке таких искажений.

В случае восстановления изображения необходима также разработка соответствующего алгоритма. В противном случае для актуальных искажений может потребоваться доработка модели, тогда как для неактуальных искажений – обучение дополнительной модели. Важно отметить, что обучение дополнительной модели вносит значительную избыточность в целевую систему, что может быть критично для некоторых областей применения.

Предложенные манипуляции могут привести к нарушению целостности внесенного вредоносного искажения.

**Выводы.** В текущей работе были предложены подходы к противодействию состязательным шумовым атакам посредством нарушения целостности вредоносного возмущения при помощи визуальных искажений и шумоподавления. В дальнейшей работе предполагается проверка сформулированных гипотез, а также исследование влияния предлагаемых подходов на производительность целевой системы и эффективность устранения различных состязательных атак.

**Список использованных источников**:


1. Банк данных угроз безопасности информации ФСТЭК России [Электронный ресурс]. – URL: https://bdu.fstec.ru/ (дата обращения: 11.01.2023).
2. Huang X. et al. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability //Computer Science Review. – 2020. – Т. 37. – С. 100270.
3. Yao Z. et al. Trust region based adversarial attack on neural networks //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. – 2019. – С. 11350-11359.
4. Moosavi-Dezfooli S. M., Fawzi A., Frossard P. Deepfool: a simple and accurate method to fool deep neural networks //Proceedings of the IEEE conference on computer vision and pattern recognition. – 2016. – С. 2574-2582.
5. Сирота А. А., Дрюченко М. А., Пузатых М. С. Моделирование аппликативных помех на изображениях с использованием глубоких нейронных сетей //Вестник ВГУ. Серия: Системный анализ и информационные технологии. – 2022. – №. 4. – С. 87-98.



Керимбай А. (автор) Подпись

Пузикова Я.В. (автор) Подпись

Есипов Д.А. (автор) Подпись

Роговой В. (научный руководитель) Подпись