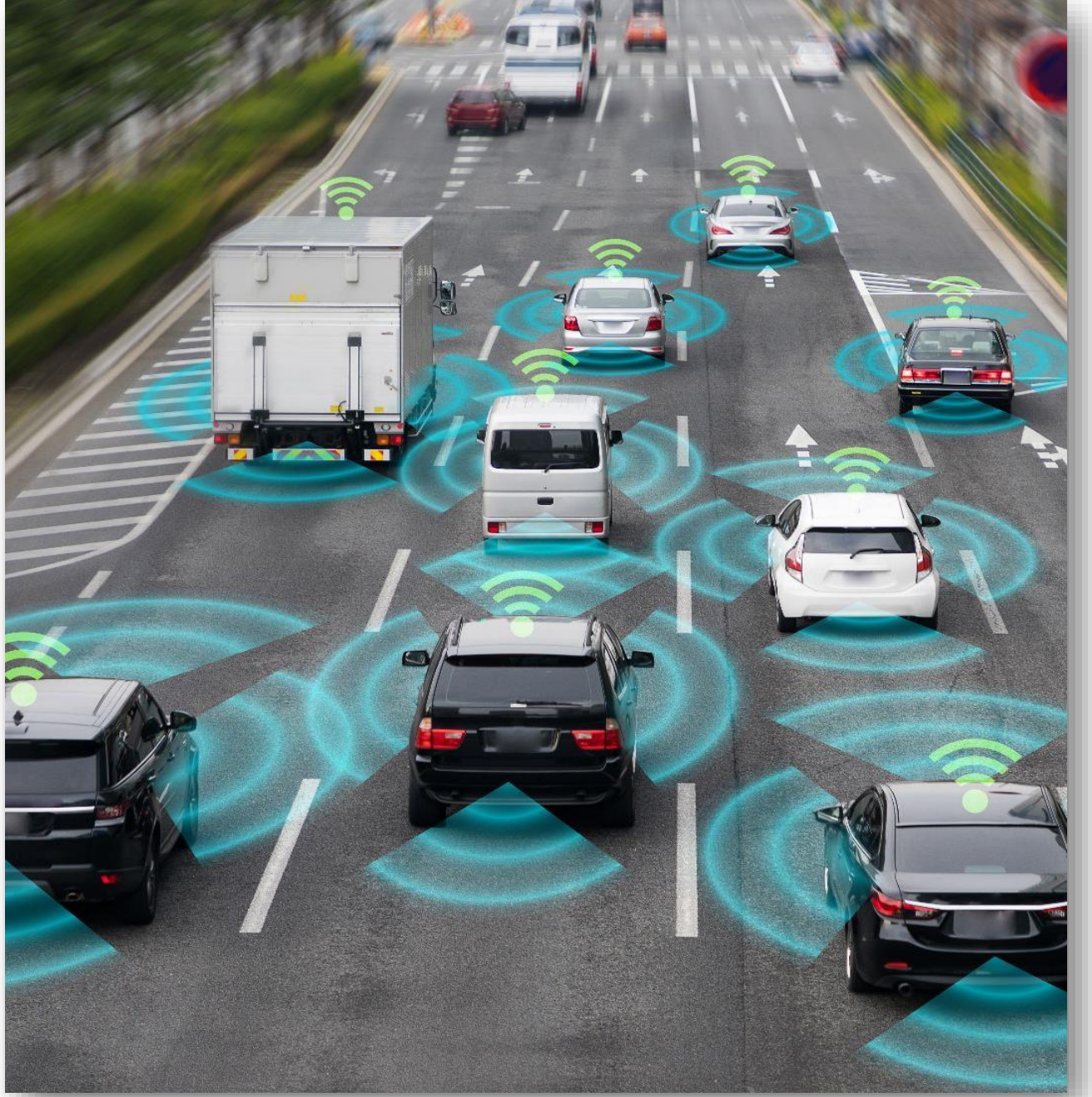
БЕЗОПАСНОСТЬ
информационных технологий
ФОРУМ - САНКТ-ПЕТЕРБУРГ

Технологии машинного обучения и искусственного интеллекта: угрозы безопасности и подходы к защите

Керимбай Акылжан, магистрант ФБИТ
Есипов Дмитрий, аспирант ФБИТ



Сферы применения ИИ

- медицинская диагностика
- управление беспилотным транспортом
- биометрическая аутентификация
- видеонаблюдение
- обработка естественного языка
- распознавание речи
- и др.

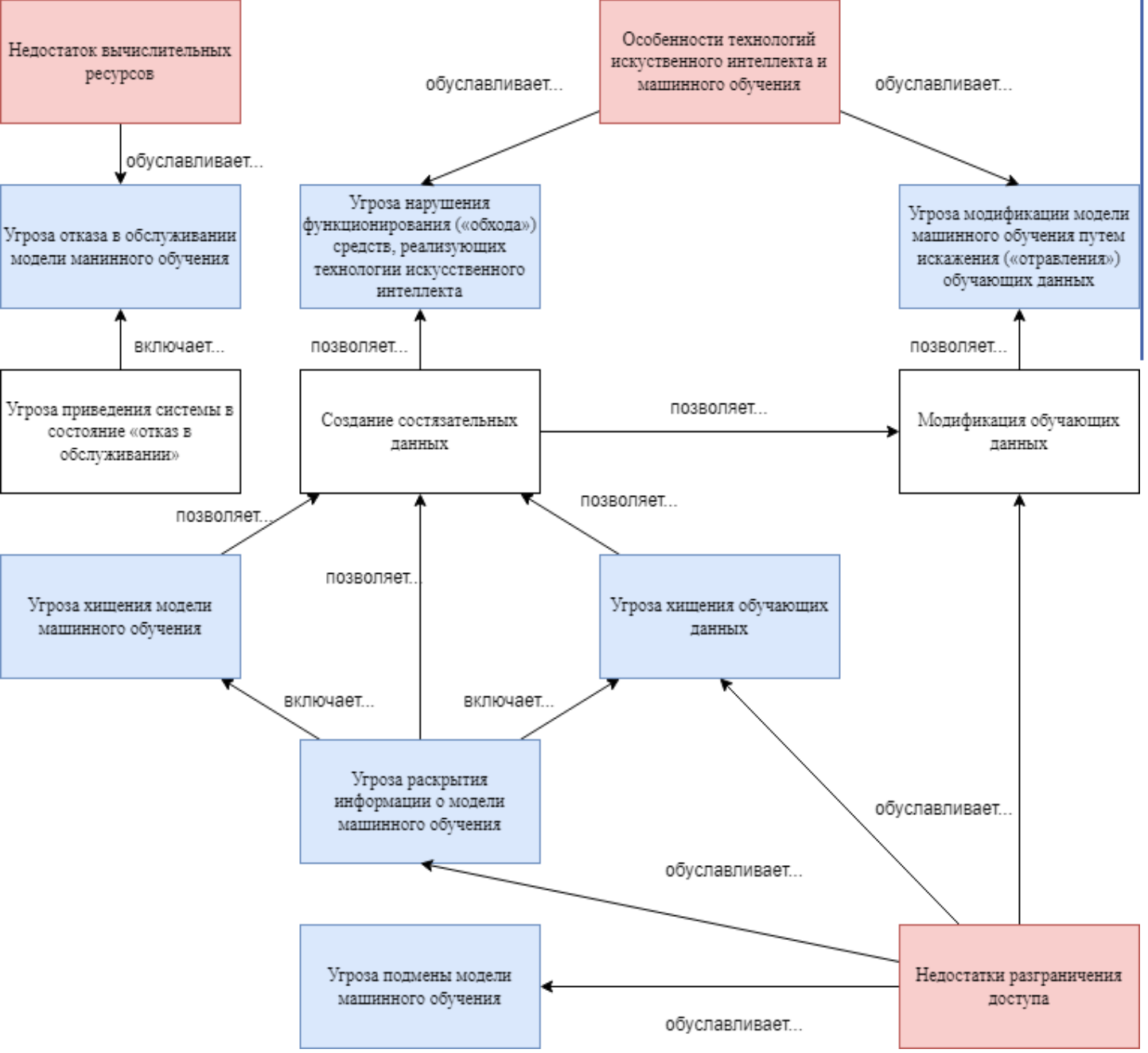
Угрозы, связанные с ИИ

БДУ ФСТЭК

- УБИ.220 – Угроза нарушения функционирования («обхода») средств, реализующих технологии искусственного интеллекта
- УБИ.221 – Угроза модификации модели машинного обучения путем искажения («отравления») обучающих данных

MITRE ATLAS

- Обход модели МО
- Отказ в обслуживании модели МО
- Нарушение целостности модели МО
- Создание бэкдора в модели МО
- Создание состязательных данных
- Отравление обучающих данных



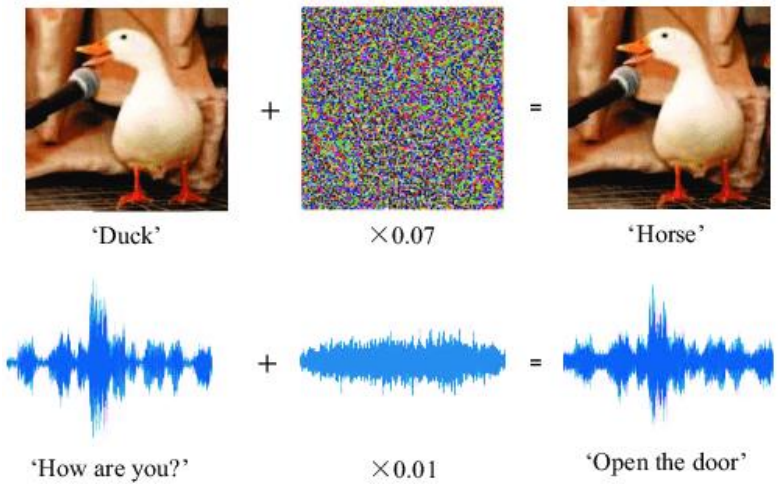
Анализ первопричины угроз

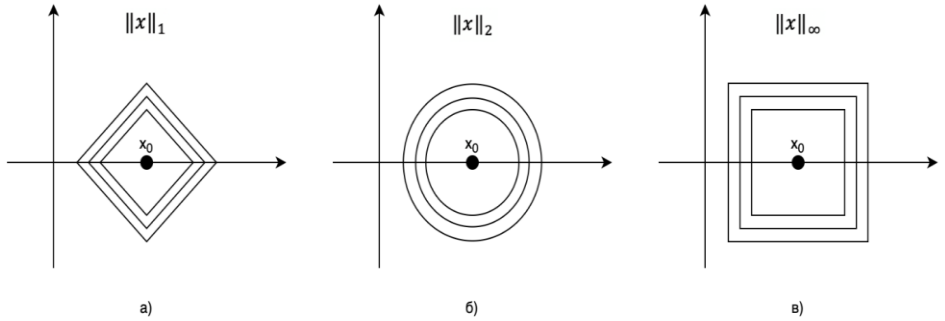
- красный – первопричина
- синий – угроза, связанная с технологиями ИИ и МО
- белый – прочие угрозы

Только для угроз обхода и модификации модели первопричиной являются характерные особенности технологий ИИ и МО.

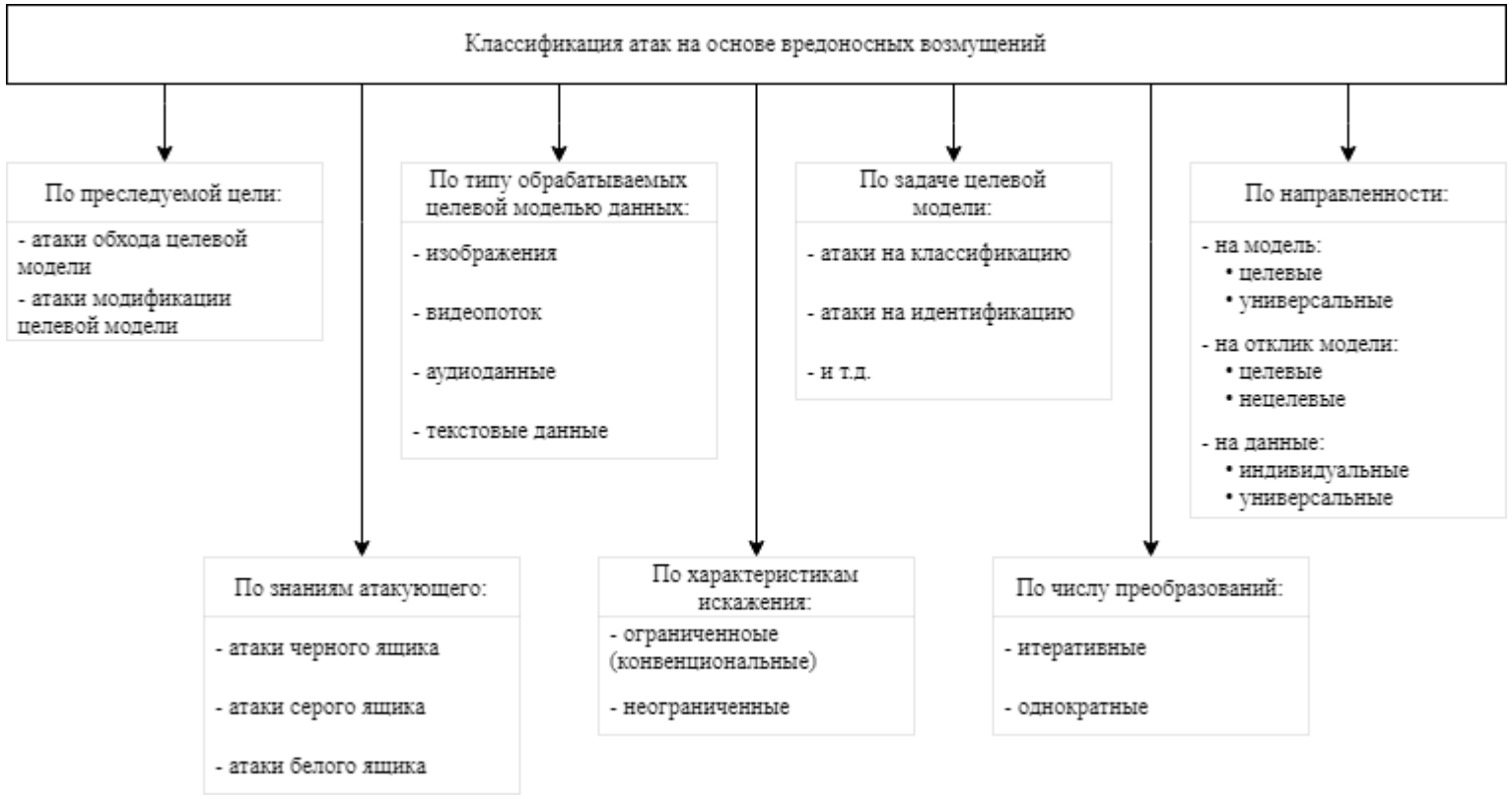
Атаки на основе вредоносных возмущений

Атаки на основе вредоносных возмущений предполагают внесение искажений во входные данные, приводящие к нарушению функционирования целевой системы





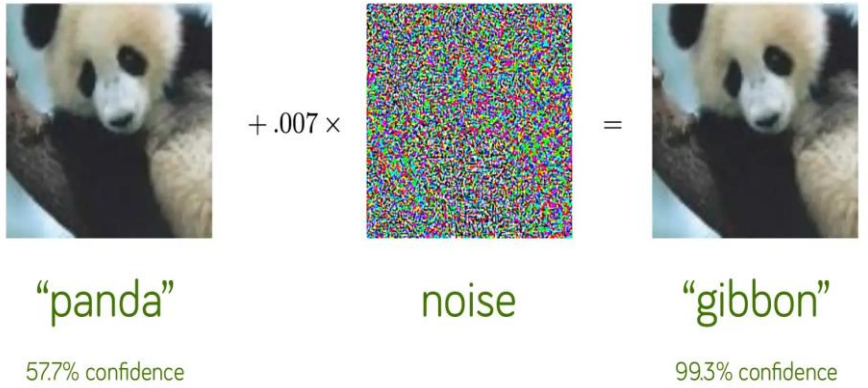
Классификация атак



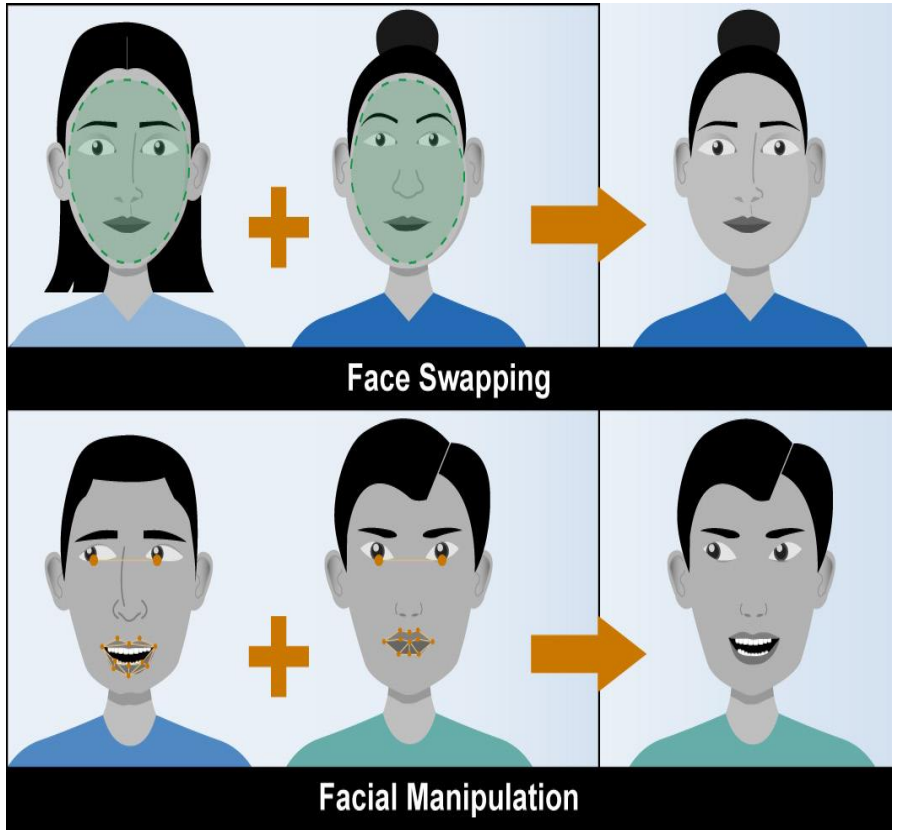
Конвенциональные атаки

Название	Ограничение нормы	Доля успешных атак, %
L-BFGS	-	87,65
FGSM	$\|x\|_2 \leq 6.25$	54,33
PGD	$\|x\|_\infty \leq 0.20$	91,85
DeepFool	$\|x\|_\infty \leq 0.20$	92,6
C&W	$\|x\|_\infty \leq 0.20$	94,8
TR	$\|x\|_\infty \leq 0.10$	94,77

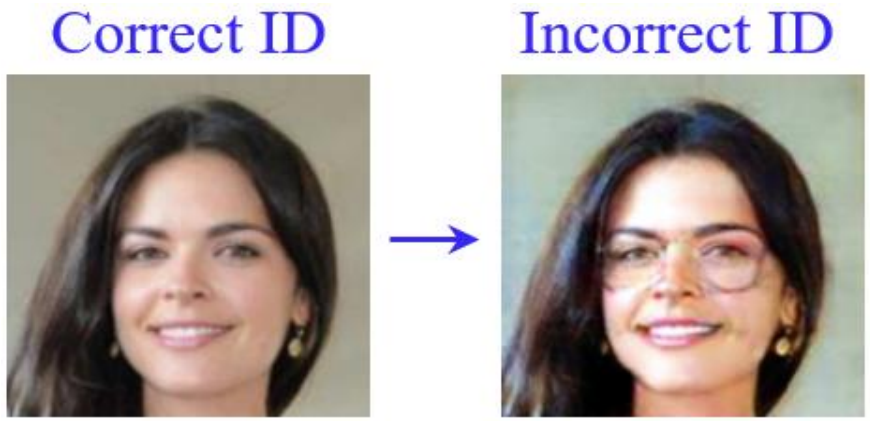
- Внесение возмущения меньшей нормы позволяет атаке оставаться незаметной
- Искажения малой нормы менее устойчивы к возможным помехам и преобразованиям
- Было показано существование универсальных возмущений
- Могут быть реализованы по модели «черного» и «белого» ящика



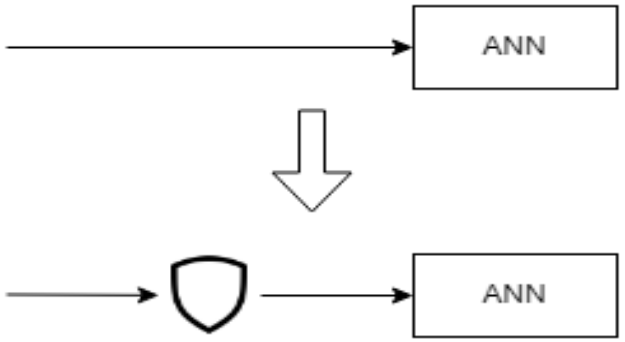
Неограниченные атаки



- Наложение неограниченного возмущения
- Манипуляция цветом
- Манипуляция атрибутами
- Создание дипфейка



Требования к методам защиты



Минимальное воздействие на архитектуру	Минимальное влияние на показатели качества	Минимальное влияние на быстродействие
Воздействие на архитектуру целевой системы для обеспечения защиты от состязательных атак должно быть минимально.	Использование методов защиты от состязательных атак не должно снижать показатели качества системы на нормальных данных.	Увеличение времени работы системы при использовании методов защиты должно быть минимально.

Модификация целевой модели

Достоинства

- позволяет игнорировать вредоносное возмущение
- не предполагает дополнительных вычислений

Недостатки

- снижение показателей качества целевой модели
- необходимость модификации и переобучения целевой нейронной сети
- сложность формирования обучающей выборки*

* для Adversarial Learning

Модификация целевой системы

Достоинства

- не оказывает значительного влияния на качество целевой системы

Недостатки

- использование дополнительной нейронной сети
- сложность адаптации к различным атакам
- снижение показателей качества целевой модели*

* для Certified Defense

Предобработка входных данных

Достоинства

- низкая вычислительная сложность
- потенциально позволяет устранить вредоносное возмущение

Недостатки

- потенциальное снижение качества изображения
- не позволяет гарантировано нивелировать атаку

Внесение случайности

Достоинства

- позволяет устранить вредоносное возмущение
- потенциальное повышение показателей качества целевой модели*

* для Random Self-Ensemble

Недостатки

- модификация архитектуры целевой системы или модели
- не позволяет гарантировано нивелировать атаку
- потенциальное снижение показателей качества целевой модели

Заключение

- Технологии машинного обучения и искусственного интеллекта получили широкое распространение
- Применение указанных технологий в областях КИИ
- Прогнозирование роста популярности и распространения МО и ИИ

- Характерные угрозы для МО и ИИ позволяют нарушить функционирование систем, использующих эти технологии
- Существующие методы обеспечения безопасности не могут обеспечить эффективной защиты от угроз

Возможные подходы к защите:

- Обнаружение вредоносных возмущений посредством анализа статистических характеристик:
  - Атаки, оптимизированные по L_0
  - Цифровые триггеры (патчи)
  - Прочие неограниченные возмущения
  - Конвенциональные атаки
- Устранение возмущений посредством предобработки входных данных:
  - Внесение малых искажений
  - Манипулирование шумом

Спасибо за внимание

Контактные данные

Керимбай Акылжан

@akerimbai@itmo.ru

+7 (952) 367-67-82

Контактные данные

Есипов Дмитрий

some1else.d.ma@gmail.com

+7 (911) 676 - 86 - 16

БЕЗОПАСНОСТЬ
информационных технологий
ФОРУМ - САНКТ-ПЕТЕРБУРГ

Технологии машинного обучения и искусственного интеллекта: угрозы безопасности и подходы к защите

Керимбай Акылжан, магистрант ФБИТ
Есипов Дмитрий, аспирант ФБИТ