



Надежность нейронных сетей

Укрепляем устойчивость ИИ к обману

Кэти Уорр



Санкт-Петербург · Москва · Екатеринбург · Воронеж
Нижний Новгород · Ростов-на-Дону
Самара · Минск

2021

Краткое содержание

Предисловие.....	10
------------------	----

ЧАСТЬ I. ОБЩИЕ СВЕДЕНИЯ ОБ ОБМАНЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Глава 1. Введение.....	19
Глава 2. Мотивация к атакам.....	38
Глава 3. Основные понятия ГНС.....	48
Глава 4. ГНС-обработка изображений, аудио- и видеоданных.....	76

ЧАСТЬ II. ГЕНЕРАЦИЯ ВРЕДНОСНЫХ ВХОДНЫХ ДАННЫХ

Глава 5. Базовые принципы вредоносных входных данных.....	104
Глава 6. Методы генерации вредоносных искажений.....	132

ЧАСТЬ III. ПОНИМАНИЕ РЕАЛЬНЫХ УГРОЗ

Глава 7. Схемы атак против реальных систем.....	168
Глава 8. Атаки в физическом мире.....	185

ЧАСТЬ IV. ЗАЩИТА

Глава 9. Оценка устойчивости модели к вредоносным входным данным.....	202
Глава 10. Защита от вредоносных входных данных.....	221
Глава 11. Дальнейшие перспективы: повышение надежности ИИ.....	261
Приложение. Справочник математических обозначений.....	267
Об авторе.....	269
Об обложке.....	270

Оглавление

Предисловие	10
Для кого предназначена книга	11
Структура издания	12
Условные обозначения	14
Использование примеров программного кода	14
Математические обозначения.....	15
Благодарности.....	15
От издательства	16

ЧАСТЬ I. ОБЩИЕ СВЕДЕНИЯ ОБ ОБМАНЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Глава 1. Введение	19
Неглубокий обзор глубокого обучения	19
Очень краткая история глубокого обучения	21
Неожиданное открытие: оптические иллюзии искусственного интеллекта.....	23
Что такое вредоносные входные данные	26
Вредоносное искажение	28
Неестественные вредоносные входные данные	29
Вредоносная заплатка	31
Вредоносные образы в физическом мире	33
Вредоносное машинное обучение в более широком смысле	35
Последствия воздействия вредоносных входных данных	36
Глава 2. Мотивация к атакам.....	38
Обход веб-фильтров.....	39
Репутация в Интернете и управление брендом	41
Камуфляж против видеонаблюдения	42
Личная конфиденциальность в Интернете.....	43

Дезориентация автономных транспортных средств.....	44
Устройства с голосовым управлением.....	46
Глава 3. Основные понятия ГНС.....	48
Машинное обучение	48
Концептуальные основы глубокого обучения.....	50
Модели ГНС как математические функции.....	55
Входные и выходные данные ГНС.....	58
Внутреннее содержимое ГНС и обработка с прямым распространением	59
Как обучается ГНС	63
Создание простого классификатора изображений	69
Глава 4. ГНС-обработка изображений, аудио- и видеоданных	76
Изображения.....	77
Цифровое представление изображений	78
ГНС для обработки изображений.....	80
Общие сведения о сверточных нейронных сетях	81
Аудиоданные.....	87
Цифровое представление аудиоданных.....	88
ГНС для обработки аудиоданных.....	89
Общие сведения о рекуррентных нейронных сетях	91
Обработка речи.....	94
Видеоданные.....	96
Цифровое представление видеоданных.....	96
ГНС для обработки видеоданных.....	96
Соображения о вредоносности	97
Классификация изображений с помощью сети ResNet50.....	99

ЧАСТЬ II. ГЕНЕРАЦИЯ ВРЕДНОСНЫХ ВХОДНЫХ ДАННЫХ

Глава 5. Базовые принципы вредоносных входных данных	104
Входное пространство	105
Обобщение обучающих данных.....	110
Эксперименты с данными вне распределения	113
Что «думают» ГНС.....	114
Искажающая атака: максимальный эффект при минимальном изменении.....	120

Вредоносная заплатка: максимальное отвлечение внимания	122
Оценка выявляемости атак	123
Математические методы оценки искажения	124
Особенности человеческого восприятия	127
Резюме	129
Глава 6. Методы генерации вредоносных искажений	132
Методы белого ящика	135
Поиск во входном пространстве	136
Использование линейности модели	139
Вредоносная значимость	148
Повышение надежности вредоносного искажения	154
Разновидности методов белого ящика	156
Методы ограниченного черного ящика	157
Методы черного ящика с оценкой	163
Резюме	166
 ЧАСТЬ III. ПОНИМАНИЕ РЕАЛЬНЫХ УГРОЗ 	
Глава 7. Схемы атак против реальных систем	168
Схемы атак	168
Прямая атака	170
Атака с копированием	171
Атака с переносом	173
Универсальная атака с переносом	177
Многократно используемые заплатки и искажения	179
Сводим все вместе: комбинированные методы и компромиссы	183
Глава 8. Атаки в физическом мире	185
Вредоносные объекты	187
Изготовление объекта и возможности камеры	187
Углы обзора и окружение	189
Вредоносный звук	195
Возможности микрофона и системы воспроизведения	196
Положение аудиосигнала и окружение	197
Осуществимость атак с использованием физических вредоносных образов	200

ЧАСТЬ IV. ЗАЩИТА

Глава 9. Оценка устойчивости модели к вредоносным входным данным	202
Цели, возможности, ограничения и знания злоумышленника	204
Цели	204
Возможности, осведомленность и доступ	209
Оценка модели	211
Эмпирические метрики устойчивости	212
Теоретические метрики устойчивости	218
Резюме	219
Глава 10. Защита от вредоносных входных данных.....	221
Улучшение модели	222
Маскирование градиентов	223
Вредоносное обучение	226
OoD-обучение.....	236
Оценка неопределенности случайного отсева	241
Предварительная обработка данных	248
Предварительная обработка в общей последовательности обработки.....	249
Интеллектуальное удаление вредоносного контента	253
Соккрытие информации о целевой системе.....	254
Создание эффективных механизмов защиты от вредоносных входных данных	257
Открытые проекты	257
Получение общей картины	258
Глава 11. Дальнейшие перспективы: повышение надежности ИИ	261
Повышение устойчивости за счет распознавания контуров	262
Мультисенсорные входные данные.....	263
Вложенность и иерархия объектов	265
В заключение	266
Приложение. Справочник математических обозначений	267
Об авторе	269
Об обложке	270