Министерство образования и науки Российской Федерации Федеральное государственное бюджетное образовательное учреждение высшего образования

"Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых"

(ВлГУ)

ОТЧЕТ ПО ПРАКТИКЕ

Дисциплина: Системы, сети и устройства телекоммуникаций

Тема: Детерминированное сетевое исчисление

Исполнитель темы ст. гр. ИБ-120 Грушка А.И.

(подпись, дата)

Научный руководитель ст. пр. Мазурок Д.В.

(подпись, дата)

Владимир 2023

Содержание

[**Введение:** 3](#_Toc145288328)

[**Типы уязвимостей:** 4](#_Toc145288329)

[**Контрмеры:** 6](#_Toc145288330)

[**Обзор моделей угроз систем, использующих нейронные сети:** 7](#_Toc145288331)

[**Методы защиты от угроз нейронных сетей:** 8](#_Toc145288332)

[**Заключение:** 10](#_Toc145288333)

[**Литература:** 11](#_Toc145288334)

# **Введение:**

**Актуальность темы.**

Сетевое исчисление — это набор последних разработок, которые обеспечивают глубокое понимание проблем, связанных с потоками, встречающимися в сетях. В основе сетевого исчисления лежит математическая теория диодов, в частности, диода Min-Plus (также называемая алгеброй Min-Plus). С помощью сетевого исчисления мы можем понять некоторые фундаментальные свойства сетей интегрированных услуг, управление потоком окон, составление расписания и измерение буфера или определение размеров задержки.

**Объект исследования.** Нейросети.

**Предмет исследования.** Методы обнаружения и противодействия состязательным атакам.

**Цель работы:**

* Повышение точности работы нейросетей в условиях проведения состязательных атак.

# **Типы уязвимостей:**

## 1**«Извлечение модели»**  предполагает дублирование модели машинного обучения через API, что не требует знаний обучающих данных и алгоритмов. Основной метод заключается в обучении замещающей модели. Замещающая модель имеет аппроксимированные атрибуты и результаты прогнозов целевой модели. В данном случае целевая модель рассматривается в качестве «чёрного ящика» с неизвестными параметрами.

## **«Инверсия модели»** предполагает вывод из модели обучающих данных, восстановление принадлежности данных или свойств данных. Так, инверсия модели, используемой в системе распознавания лиц, позволяет с удовлетворительным качеством восстановить изображения, использованные при обучении нейросети.

## **«Отравление»** – т.е. искажение обучающих данных. Любое манипулирование данными, которое осталось вне поля зрения разработчика ИИ-системы, может иметь катастрофические последствия или поставить под угрозу сам процесс обучения. В известном примере с распознаванием нейросетью дорожного знака обязательной остановки ИИ «принимает во внимание» области красного цвета, форму букв и ряд других характеристик изображения. Зная, как именно обучена нейросеть, злоумышленник может, например, изменить дорожный знак так, чтобы это было незаметно для человека, но ввело бы в заблуждение нейросеть. Существуют иные, весьма разнообразные, возможности «отравить» нейросеть, обеспечив её неработоспособность в определённых условиях.

## Belfer Сenter **«Состязание»** схоже с отравлением, в этом случае также имеет место неверная классификация моделью вредоносной информации, подлежащей обработке нейросетью. Отличие заключается в том, что состязание предполагает использование слабостей самой модели, а отравление ‑ атаку на обучающие данные.

## **«Уклонение»** предполагает внесение «шумов» в обученные нейросети в целях нарушения их корректной работы. Добавляя определённый случайный «шум» в исходные данные (такой «шум», т.е. помехи, вносятся, например, в изображения на уровне пикселей, так, что человеческий глаз их не замечает) за счёт изменения, например, весовых коэффициентов анализируемых признаков, можно сделать так, что нейросеть при определённых условиях утратит работоспособность.

## 

## 

Рисунок 1. Условная схема «отравления» нейросети на примере распознавания дорожного знака.

Британская компания Roke Manor Research приводит в своем исследовании несколько уровней злоумышленного проникновения в нейросети для эксплуатации уязвимостей перечисленных типов:

* **«белый ящик»** – позволяет генерировать ложные входные данные, реализуется благодаря инсайдерской информации о нейросети и методах её обучения;
* **«чёрный ящик»** – позволяет настраивать желательную реакцию нейросети на вредоносные входные данные;
* **замещение нейросети** – использование замещающей нейросети в качестве копии подлинной нейросети, это позволяет разработать метод атаки и проверить его на замещающей нейросети;
* **копирование нейросети** – то же, что и замещение, но замещающая нейросеть представляет собой не копию подлинной нейросети, а её имитатор («аппроксимацию»).

# **Контрмеры:**

* создание центра мониторинга инцидентов и управления изменениями для ликвидации последствий ошибок, неправомерных действий и целенаправленных атак;
* разработку модели угроз и модели нарушителя для систем, использующих ИИ;
* определение правового статуса интегрированных систем ИИ и анализ полномочий по их практическому использованию;
* создание площадки для обмена лучшими практиками в области обеспечения доверия и безопасности систем, использующих ИИ;
* дальнейшее развитие процессов стандартизации ИИ и научные исследования.

# **Обзор моделей угроз систем, использующих нейронные сети:**

**Внутренние угрозы**: атаки со стороны пользователей или злоумышленников, имеющих доступ к системе. Они могут изменять данные или модели нейронных сетей, что приводит к некорректным результатам и/или ошибкам.

**Внешние угрозы:** атаки со стороны злоумышленников, пытающихся получить несанкционированный доступ к системе, не имея прямой доступ. Они могут включать в себя попытки взлома системы, фишинг и другие методы.

**Угрозы конфиденциальности:** связаны с возможностью получения несанкционированного доступа к конфиденциальным данным, обрабатываемыми нейронными сетями. Злоумышленник может попытаться получить доступ к личной коммерческой информации, которые хранятся в системе.

**Угрозы целостности:** связаны с возможностью изменения данных или моделей нейронных сетей, что может привести к искажению результатов и ошибкам. Злоумышленник может изменить данные, поступающие на вход нейронной сети, или модифицировать ее параметры таким образом, чтобы она давала неверные результаты или принимала неправильные решения.

**Угрозы доступности:** связаны с возможностью отказа в обслуживании или нарушения работы системы, использующей нейронные сети. Злоумышленник может попытаться перегрузить систему, что приведет к недоступности системы для пользователей.

# **Методы защиты от угроз нейронных сетей:**

**Аутентификация и авторизация:** использование механизмов аутентификации и авторизации для предотвращения несанкционированного доступа к системе и моделям нейронных сетей.

**Мониторинг и обнаружение аномалий:** использование систем мониторинга и обнаружения аномалий для выявления необычного поведения нейронных сетей и предотвращения атак.

**Защита данных:** использование методов шифрования и других механизмов для защиты конфиденциальности данных, обрабатываемых нейронными сетями.

**Проверка моделей:** использование методов верификации и проверки моделей нейронных сетей для обнаружения уязвимостей и ошибок.

GitHub как отчёт по практике:

<https://github.com/TherQut/Summer-2023/blob/main/SummerPractice>

# **Заключение:**

В ходе выполнения НИП: были выявлены модели угроз систем, использующих нейронные сети и выявлены оптимальные контрмеры.

Экспоненциальный рост влияния ИИ на важные отрасли, государственную, экономическую и общественную сферы, требует решения проблем безопасности и доверия к ИИ-системам, комплексных мер и всестороннего анализа исследований в области ИИ.

# **Литература:**

1. [Реализация модели анализа защиты информации на основе нейронной сети](https://cyberleninka.ru/article/n/realizatsiya-modeli-analiza-zaschity-informatsii-na-osnove-neyronnoy-seti/viewer)

(Дата обращения 8.07.2023)

1. [МОДЕЛЬ ВЫЯВЛЕНИЯ ВТОРЖЕНИЙ В КОМПЬЮТЕРНЫЕ СЕТИ](https://www.elibrary.ru/download/elibrary_44034215_43414590.pdf)

(Дата обращения 9-10.07.2023)

1. [ОБНАРУЖЕНИЕ ИНЦИДЕНТОВ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ НА ОСНОВЕ ТЕХНОЛОГИИ НЕЙРОННЫХ СЕТЕЙ](https://www.elibrary.ru/download/elibrary_50310108_30980813.pdf)

(Дата обращения 10-12.07.2023)

1. [ИГРЫ ИСКУССТВЕННОГО РАЗУМА: АТАКИ НА МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ И ИХ ПОСЛЕДСТВИЯ](https://kryptonite.ru/articles/ataki-na-ml/)

(Дата обращения 13-14.07.2023).