Актуальность состязательных атак как реальной угрозы безопасности нейронных сетей

Дюдюн Глеб Дмитриевич

Международная научная конференция студентов, аспирантов и молодых учёных «Ломоносов-2024»,
Секция «Секция цифр трансформация госуправления» г. Москва, 2024г.

ВЫСШАЯ ШКОЛА
ГОСУДАРСТВЕННОГО
АДМИНИСТРИРОВАНИЯ
МГУ имени М.В.Ломоносова

Команда проекта

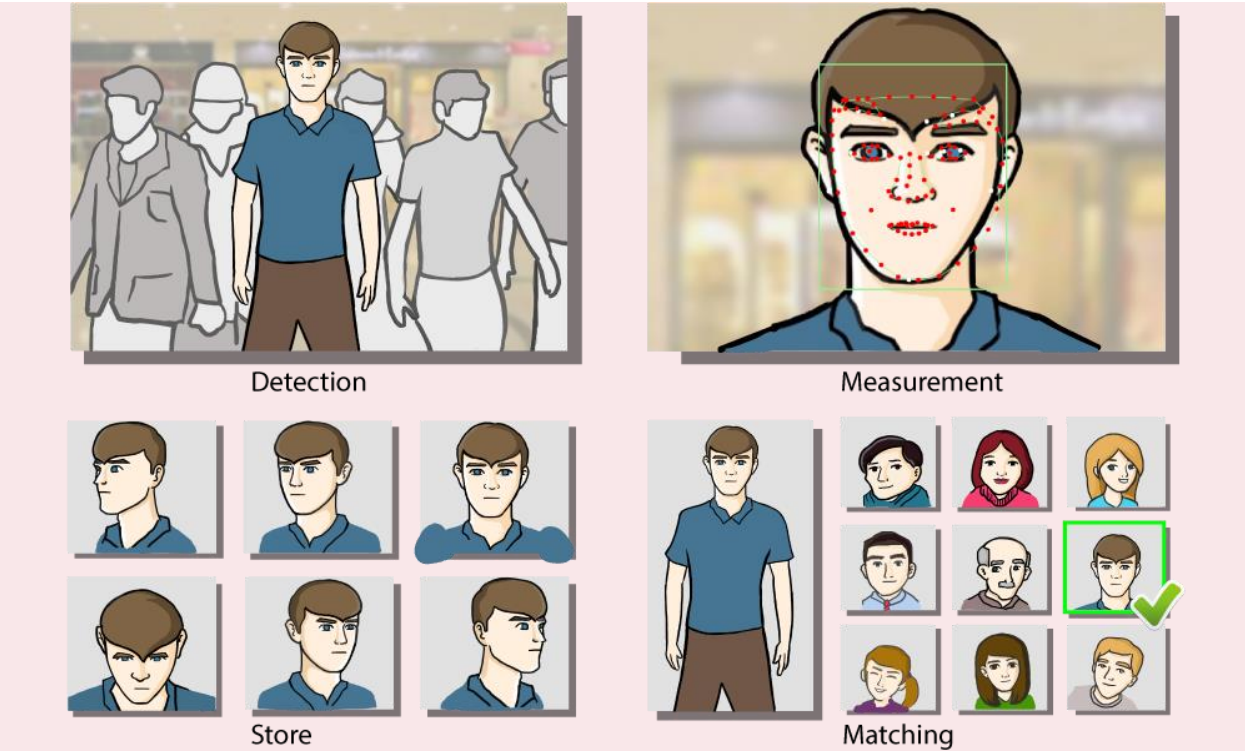
Дюдюн Глеб Дмитриевич – студент 2 курс

Лапина Мария Анатольевна – кандидат физико-математических наук, доцент

Бабенко Михаил Григорьевич – доктор физико-математических наук, доцент

ВЫСШАЯ ШКОЛА ГОСУДАРСТВЕННОГО АДМИНИСТРИРОВАНИЯ
МГУ имени М.В.Ломоносова

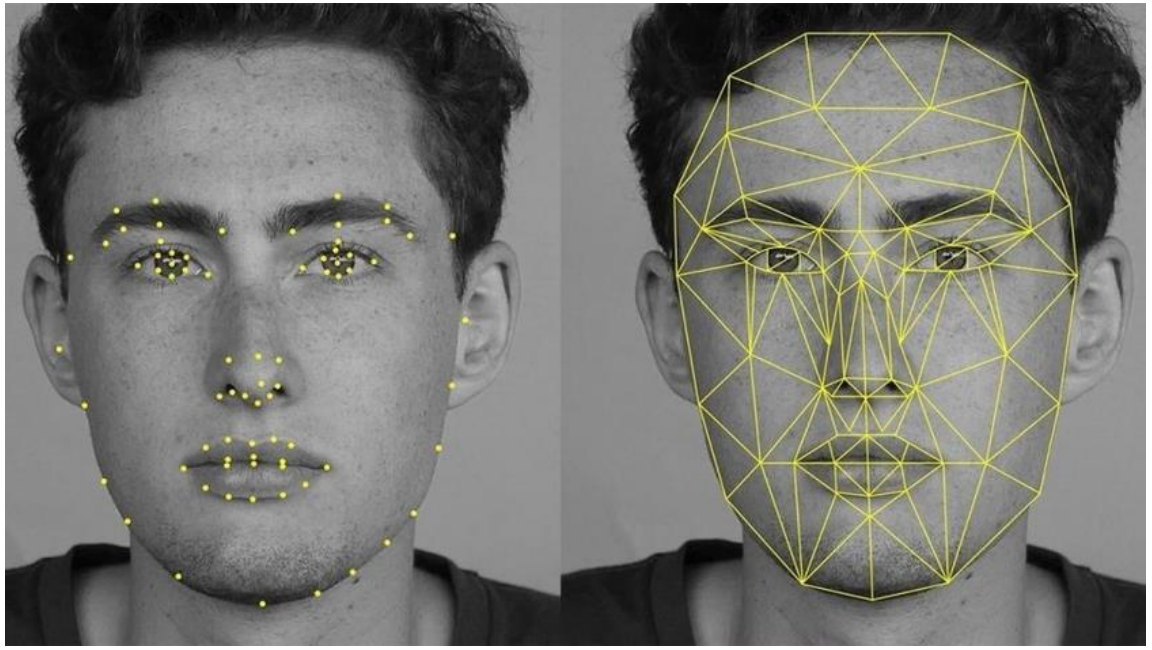
Вступление



Быстрое развитие технологий нейронных сетей привлекло внимание потенциальных мошенников, а совершенствование и распространение интеллектуальных систем аутентификации и видеонаблюдения потребовало от них использования новых методов обхода безопасности.

ВЫСШАЯ ШКОЛА
ГОСУДАРСТВЕННОГО
АДМИНИСТРИРОВАНИЯ
МГУ имени М.В.Ломоносова

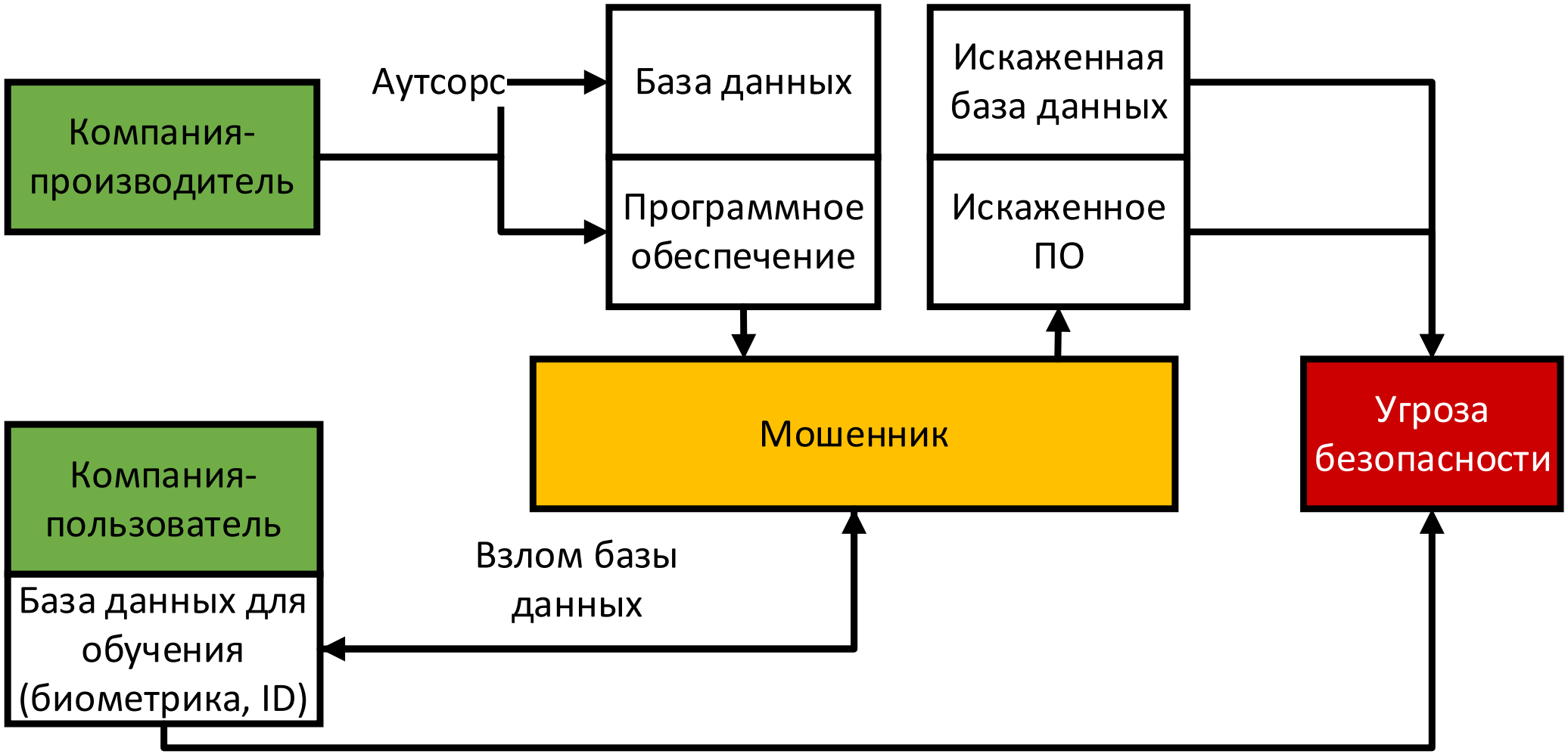
Проблематика



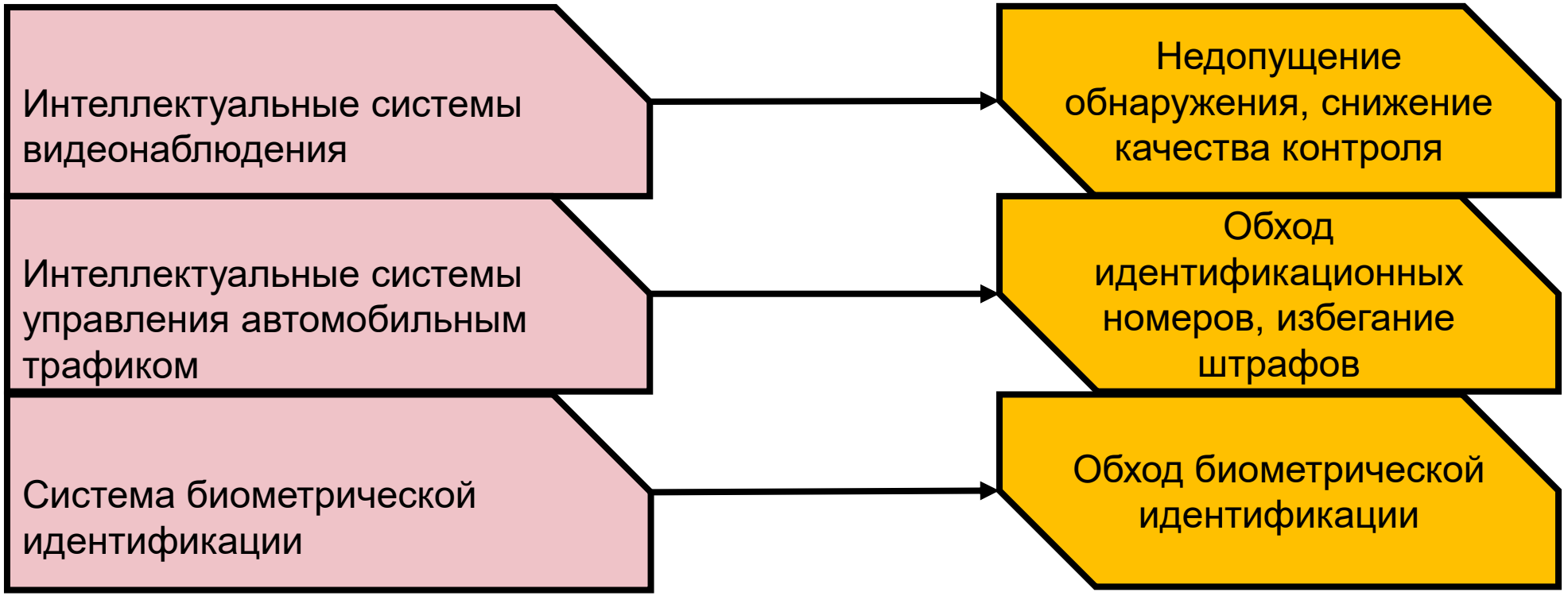
Наиболее выгодным по соотношению простота/эффективность является метод состязательных атак. При правильном применении он позволяет мошенникам стать не опознаваемыми для любой интеллектуальной системы либо сильно снизить её эффективность

ВЫСШАЯ ШКОЛА
ГОСУДАРСТВЕННОГО
АДМИНИСТРИРОВАНИЯ
МГУ имени М.В.Ломоносова

Схема угрозы

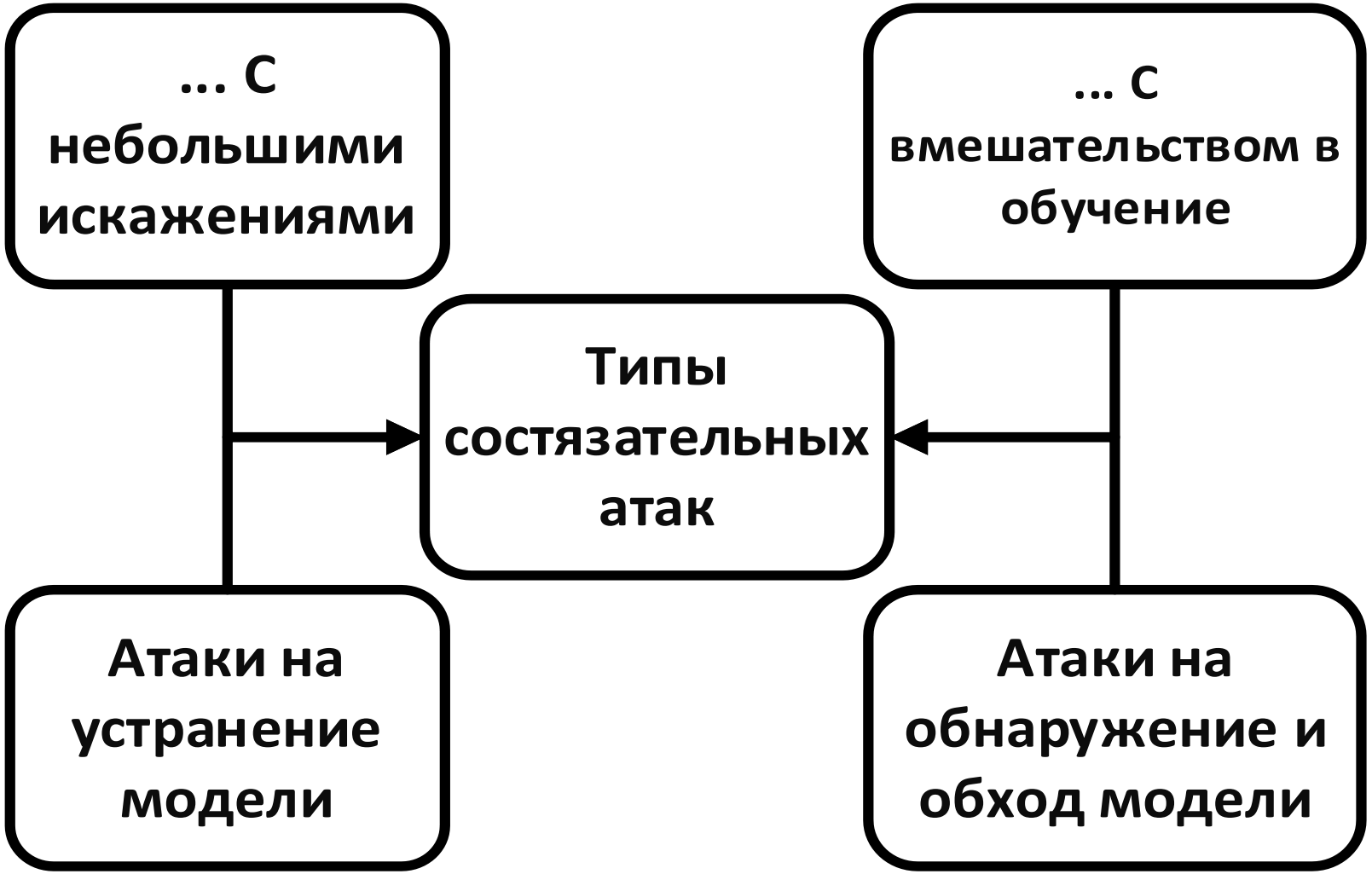


Перечень угроз



ВЫСШАЯ ШКОЛА
ГОСУДАРСТВЕННОГО
АДМИНИСТРИРОВАНИЯ
МГУ имени М.В.Ломоносова

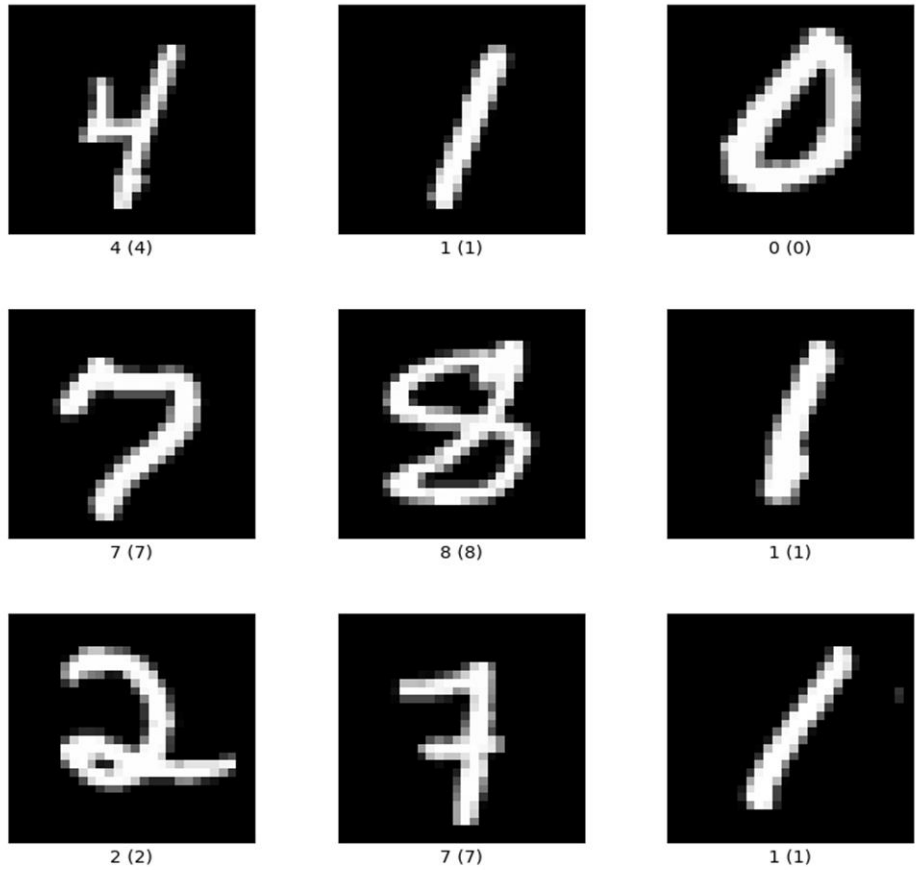
Типы состязательных атак



ИСПОЛЬЗУЕМАЯ МОДЕЛЬ

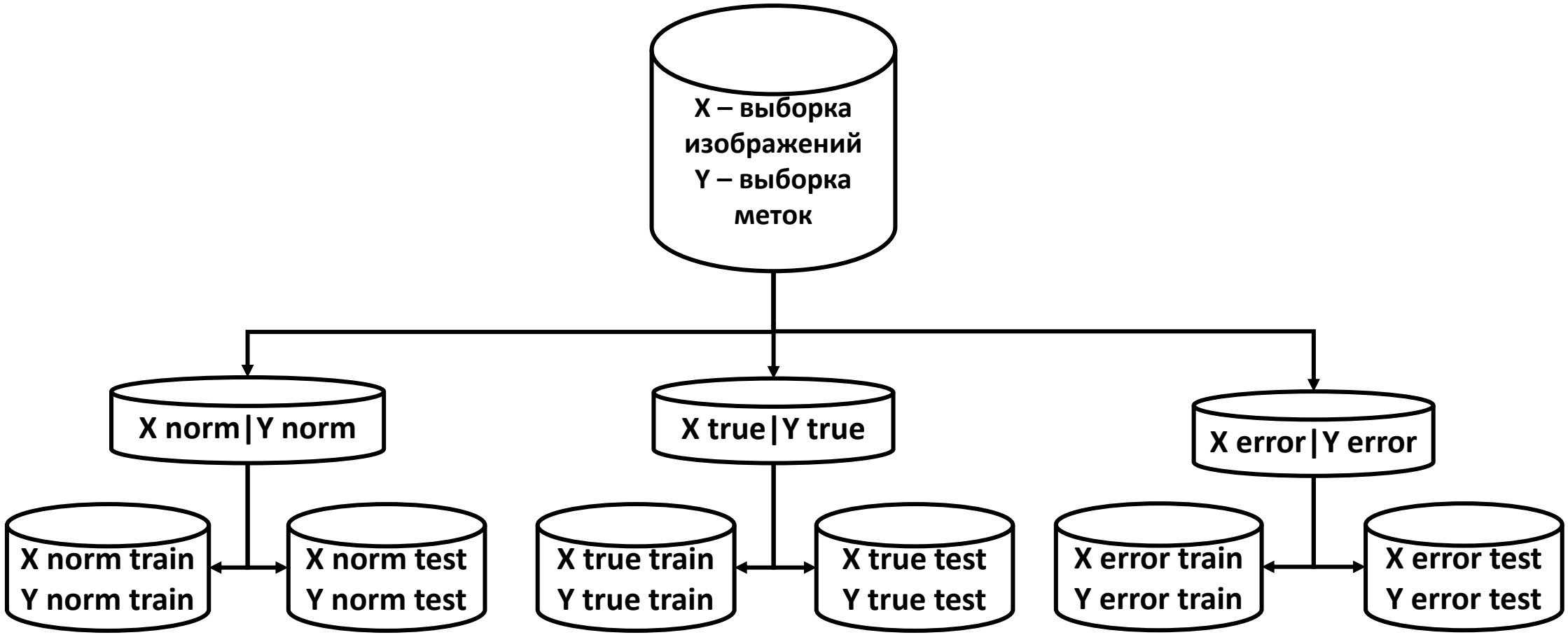
Для исследования была использована простейшая модель нейронной сети, предназначенная для распознавания рукописных символов (в нашем случае, чисел)

Входные данные - DB «Mnist» (Рукописные цифры + метки-значения)



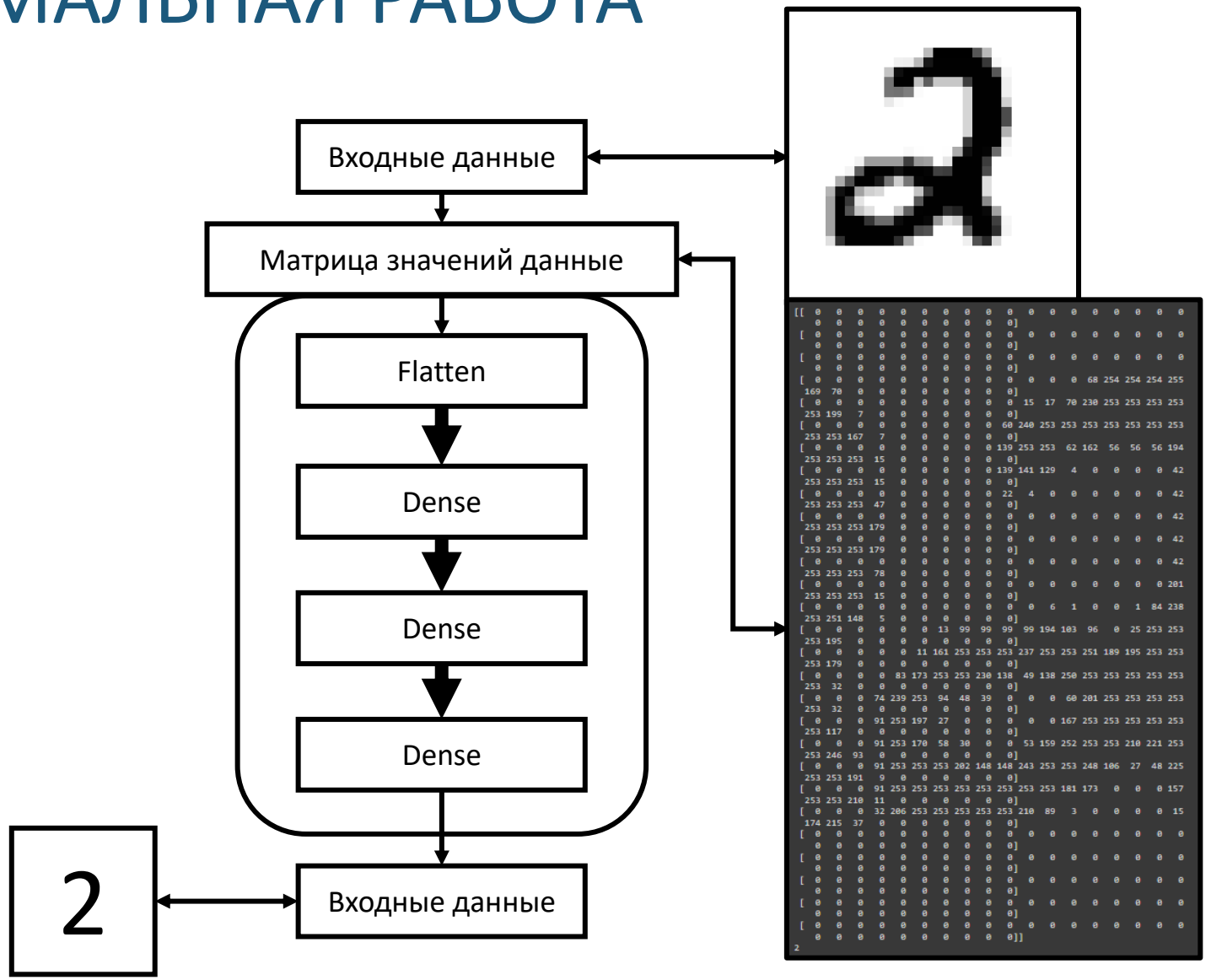
ВЫСШАЯ ШКОЛА
ГОСУДАРСТВЕННОГО
АДМИНИСТРИРОВАНИЯ
МГУ имени М.В.Ломоносова

ИСПОЛЬЗУЕМАЯ ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА ДАННЫХ



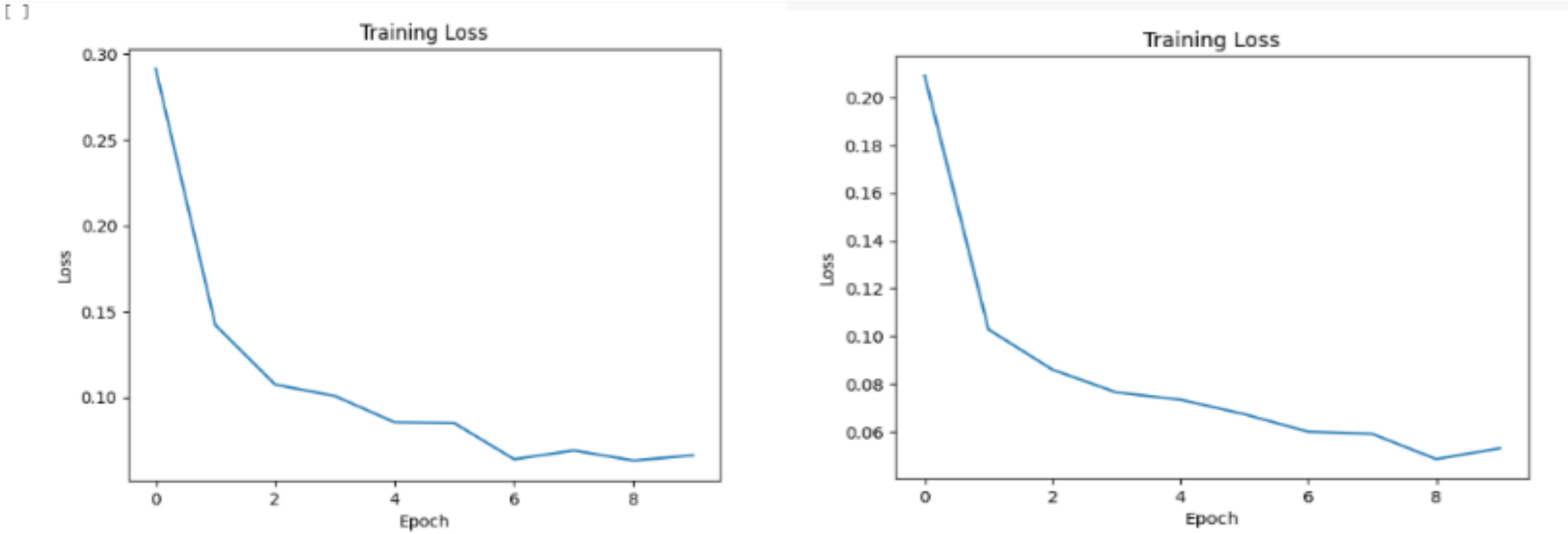
ВЫСШАЯ ШКОЛА ГОСУДАРСТВЕННОГО АДМИНИСТРИРОВАНИЯ
МГУ имени М.В.Ломоносова

СТРУКТУРА НЕЙРОННОЙ СЕТИ НОРМАЛЬНАЯ РАБОТА



ВЫСШАЯ ШКОЛА ГОСУДАРСТВЕННОГО АДМИНИСТРИРОВАНИЯ МГУ имени М.В.Ломоносова

Первые результаты: Общие последствия враждебного нападения



В ходе эксперимента доказана зависимость общей точности работы нейросети при воздействии состязательной атаки (рост значения функции ошибки)

ВЫСШАЯ ШКОЛА ГОСУДАРСТВЕННОГО АДМИНИСТРИРОВАНИЯ
МГУ имени М.В.Ломоносова

РЕЗУЛЬТАТЫ ИСПЫТАНИЙ: зависимость то выбора элемента

	0	1	2	3	4	5	6	7	8	9
0	0.980799973	0.8718000054	0.8744000196	0.8723999858	0.8741999865	0.8754000068	0.8776000142	0.8762000203	0.8713999987	0.8762000203
1	0.8687999845	0.980799973	0.8632000089	0.8705999851	0.8676000237	0.8658000231	0.8623999953	0.8658000231	0.8677999973	0.8655999899
2	0.8795999885	0.8777999878	0.980799973	0.8841999769	0.8820000291	0.8795999885	0.878000021	0.8812000155	0.8790000081	0.8786000013
3	0.8802000284	0.8787999749	0.8809999824	0.980799973	0.8790000081	0.8812000155	0.8799999952	0.8790000081	0.8790000081	0.8787999749
4	0.8862000108	0.8877999783	0.8855999708	0.8835999966	0.980799973	0.8877999783	0.8826000094	0.8863999844	0.8853999972	0.885800004
5	0.8949999809	0.8921999931	0.8916000128	0.893599987	0.8957999945	0.980799973	0.893599987	0.893599987	0.8944000006	0.8880000114
6	0.8808000088	0.8808000088	0.882799983	0.880400002	0.8805999756	0.8776000142	0.980799973	0.8776000142	0.8769999743	0.8781999946
7	0.8791999817	0.8754000068	0.8784000278	0.8805999756	0.8808000088	0.8790000081	0.8759999871	0.980799973	0.8794000149	0.8763999939
8	0.8841999769	0.8845999837	0.882799983	0.8877999783	0.8841999769	0.8863999844	0.8831999898	0.8841999769	0.980799973	0.8862000108
9	0.8798000216	0.8822000027	0.8809999824	0.8848000169	0.8866000175	0.8830000162	0.8798000216	0.8826000094	0.8816000223	0.980799973

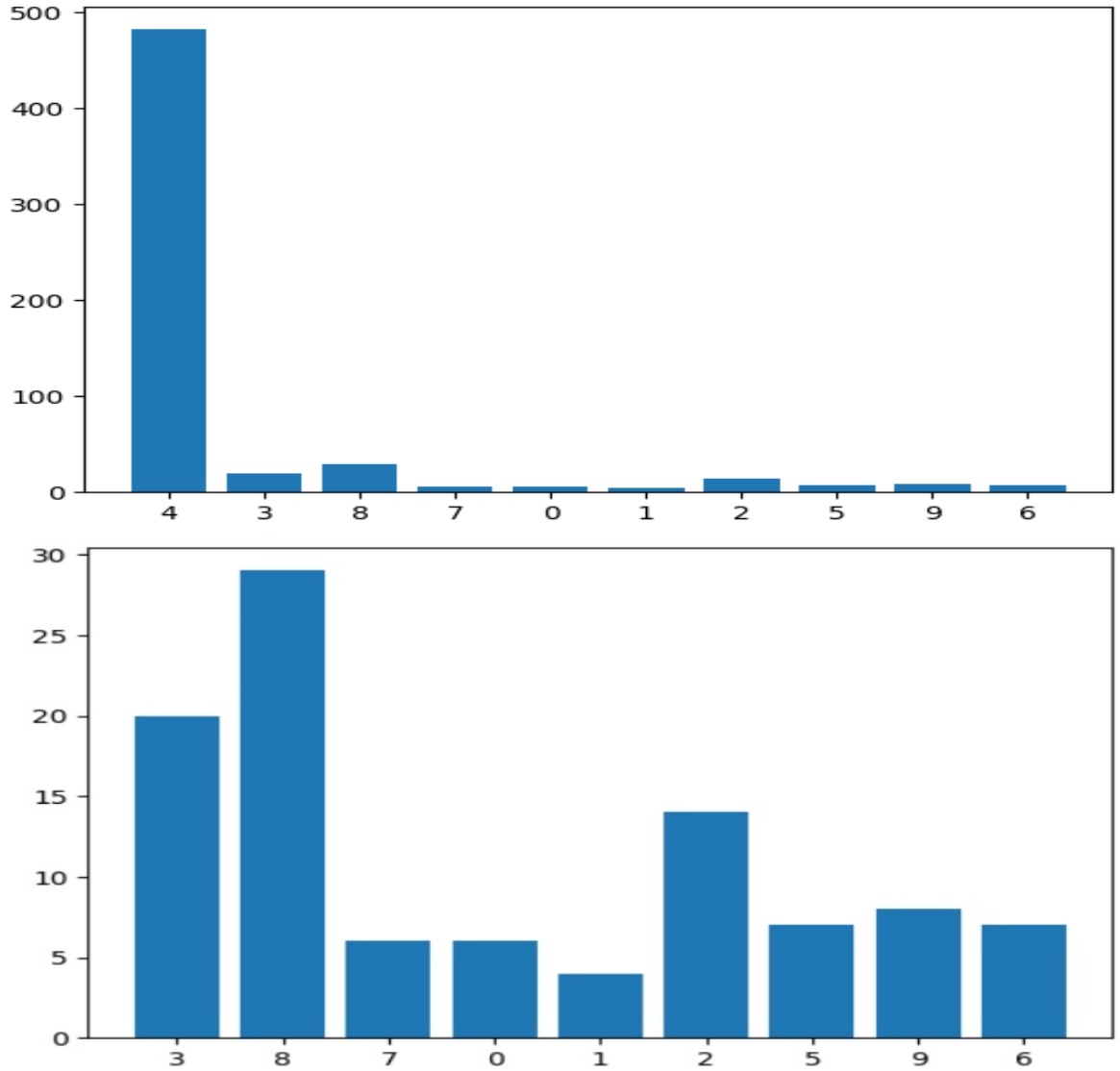
В ходе эксперимента было установлено, что снижение точности работы нейронной сети не зависит от выбора типа элементов для атаки (нет никакой схемы в распределении значений таблицы, кроме главной диагонали)

ВЫСШАЯ ШКОЛА
ГОСУДАРСТВЕННОГО
АДМИНИСТРИРОВАНИЯ
МГУ имени М.В.Ломоносова

РЕЗУЛЬТАТЫ ИСПЫТАНИЙ: Анализ ошибок нейронной сети

Эксперимент заключается в атаке на элементы типа «4» и последующем подсчете ошибочных распознаваний нейросети

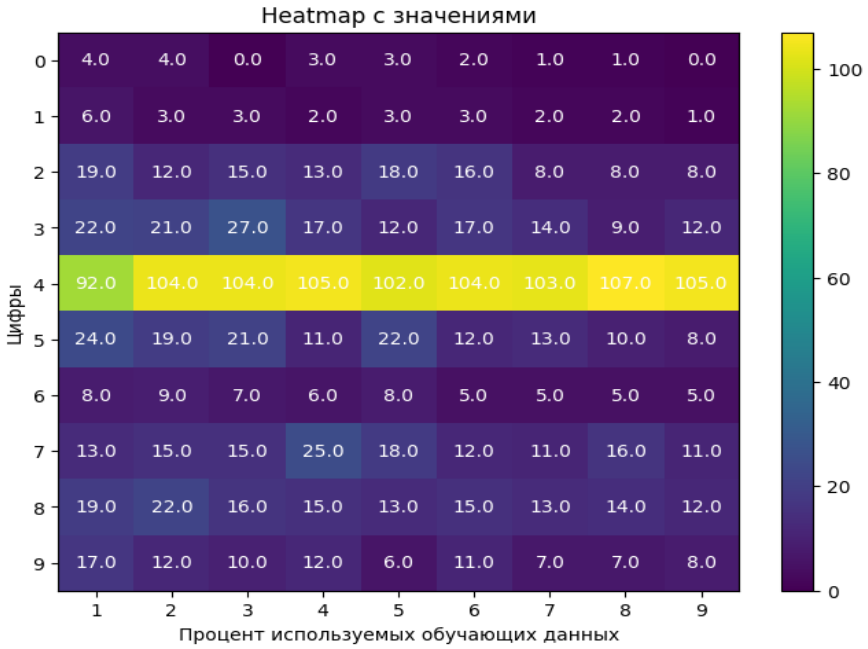
Во время теста было установлено, что можно установить атакуемый элемент по числу ошибок, допущенных при распознавании этого элемента



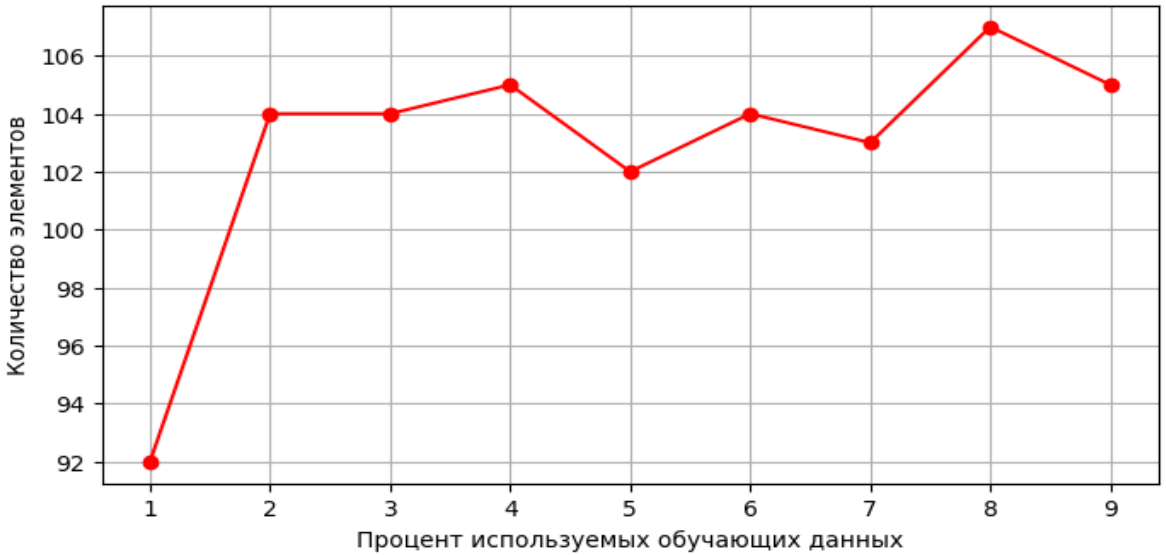
ВЫСШАЯ ШКОЛА ГОСУДАРСТВЕННОГО АДМИНИСТРИРОВАНИЯ МГУ имени М.В.Ломоносова

РЕЗУЛЬТАТЫ ИСПЫТАНИЙ: Применение метода образцов

Методика образцов заключается в копировании исследуемой модели и обучении её на небольшом количестве неискаженных данных и использовании её как образца для сравнения ответов и подсчета ошибок в работе оригинала.



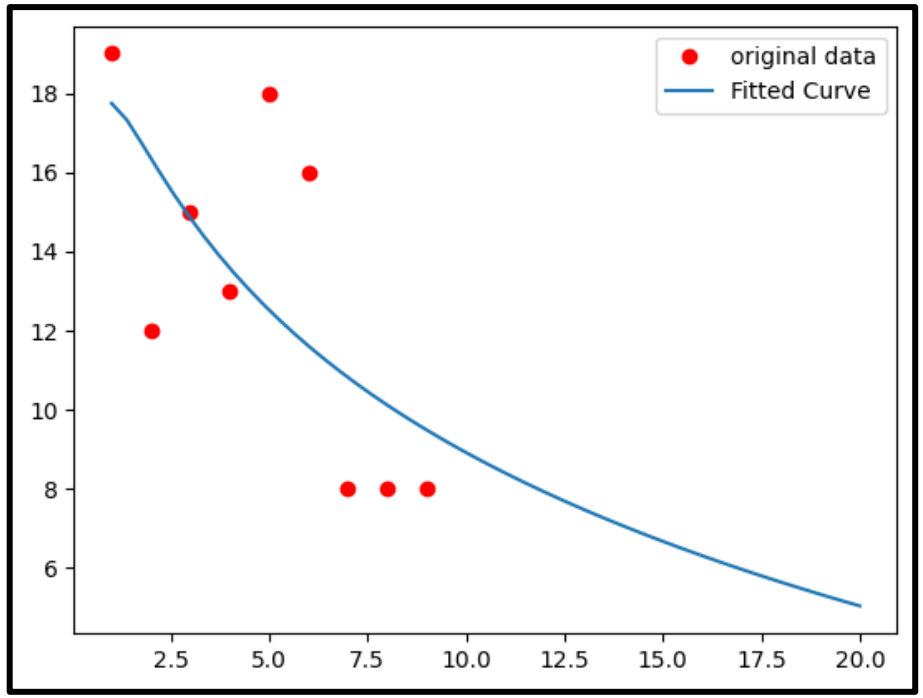
Наибольшая «контрастность» ответов наблюдается на 10% от исходного количества данных



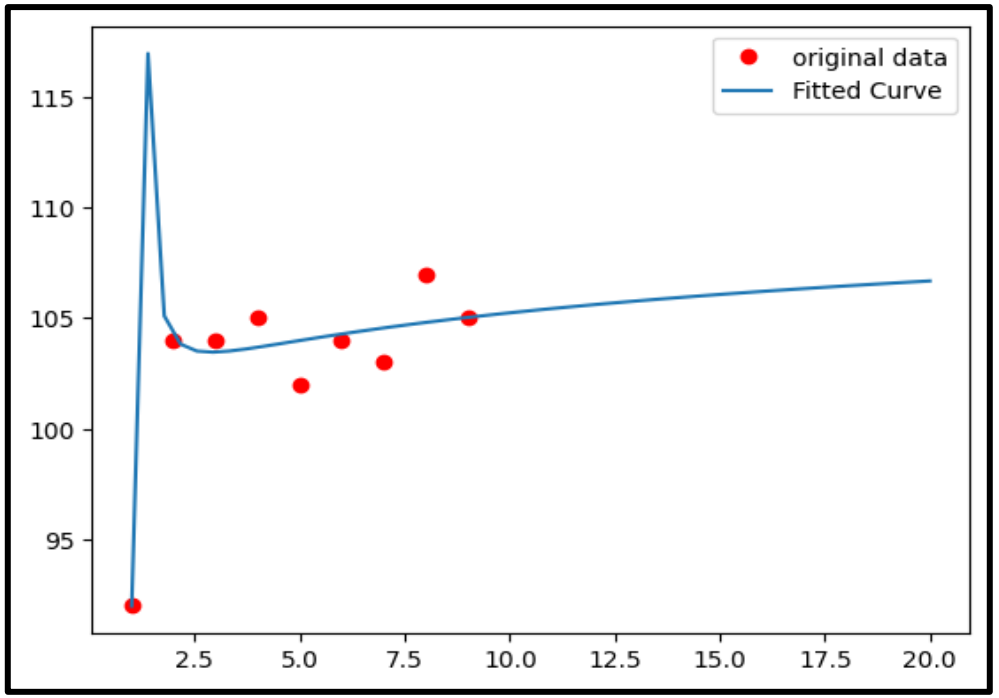
Графики отражают качество выделения атакованного элемента «4» от количества использованных данных

ВЫСШАЯ ШКОЛА ГОСУДАРСТВЕННОГО АДМИНИСТРИРОВАНИЯ МГУ имени М.В.Ломоносова

Построение линий тренда



Вариант линии тренда для распознавания «2»



Вариант линии тренда для распознавания «4»

ВЫСШАЯ ШКОЛА
ГОСУДАРСТВЕННОГО
АДМИНИСТРИРОВАНИЯ
МГУ имени М.В.Ломоносова

Благодарю за внимание

Студент второго курса направления ИБАС, Дюдюн Глеб Дмитриевич

ВЫСШАЯ ШКОЛА
ГОСУДАРСТВЕННОГО
АДМИНИСТРИРОВАНИЯ
МГУ имени М.В.Ломоносова