МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«МИРЭА – Российский технологический университет»
РТУ МИРЭА

ИКБ направление «Кибербразведка и противодействие угрозам с применением технологий искусственного интеллекта» 10.04.01

Кафедра КБ-4 «Интеллектуальные системы информационной безопасности»

Лабораторная работа №1

по дисциплине

«Анализ защищенности систем искусственного интеллекта»

Группа:
ББМО-02-22
Выполнил:
Запруднов М. С.

Проверил:
Спирин
Андрей
Андреевич

Москва 2023

Цель лабораторной работы

В данной лабораторной работе необходимо выявить закономерность или обнаружить отсутствие влияния параметра fgsm_eps на стойкость моделей к атаке. Закономерности или их отсутствие необходимо выявить для сети FC LeNet на датасете MNIST и для сети NiN LeNEt на датасете CIFAR.

Результаты эксперимента

Для достижения поставленной цели использовался язык Python. Код реализации эксперимента представлен в файле AZSII_lab_1.ipynb.

Отразим выявленные закономерности для сети FC LeNet:

- при маленьких значениях fgsm_eps, например, fgsm_eps=0.001 и fgsm_eps=0.02, ошибка классификации (FGSM Test Error) остаётся низкой, и сеть остаётся относительно устойчивой к атакам;

- ошибка классификации начинает расти при fgsm_eps=0.5 и fgsm_eps=0.9 и достигает высоких значений, что свидетельствует о нарушении стойкости сети к атакам;

- при очень большом значении (fgsm_eps=10) ошибка классификации также высока, и сеть становится непригодной для задач классификации из-за большого искажения входных данных.

Сеть NiN LeNet также начинает демонстрировать увеличение ошибки классификации с ростом параметра fgsm_eps (при fgsm_eps=0.001 и fgsm_eps=0.02 ошибка остаётся низкой, но при fgsm_eps=0.5, fgsm_eps=0.9 и fgsm_eps=10 ошибка резко увеличивается).

Визуализируем описанные закономерности для сетей, участвовавших в эксперименте (рис. 1).

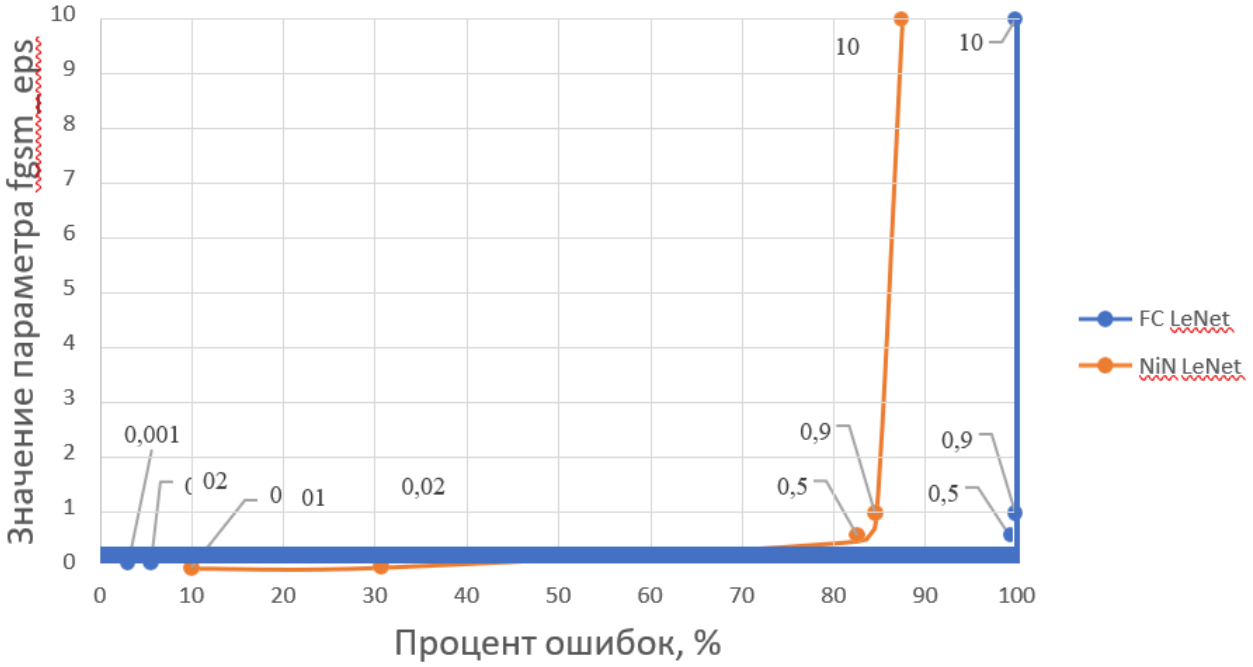

Рисунок 1 – График зависимости параметра fgsm_eps на процент ошибок длясети FC LeNet и NiN LeNet

Заключение

В результате выполнения лабораторной работы было выявлено, чтомаленькие значения fgsm_eps сохраняют стойкость сетей к атакам, и ошибки классификации остаются низкими. При увеличении fgsm_eps сети становятсяболее уязвимыми к атакам и допускают больше ошибок классификации. Для сети FC LeNet на датасете MNIST и для сети NiN LeNEt на датасете CIFAR ненаблюдается отсутствие влияния параметра fgsm_eps. Наоборот, параметр fgsm_eps оказывает существенное влияние на стойкость сетей к атакам.