УДК 004.056

ПОДХОДЫ К ВОССТАНОВЛЕНИЮ ИСКАЖЕННЫХ ДАННЫХ ПОСЛЕ ОДНОПИКСЕЛЬНОЙ АТАКИ НА СВЕРТОЧНЫЕ НЕЙРОННЫЕ СЕТИ

Сайдумаров С. К. (Университет ИТМО), **Сулименко Н.С.** (Университет ИТМО), **Есипов Д.А.** (Университет ИТМО), **Роговой В.** (Университет ИТМО)

Научный руководитель – кандидат технических наук Попов И.Ю. (Университет ИТМО**)**

Введение. Сверточные нейронные сети (Convolutional Neural Network, CNN) широко используются в различных областях, включая распознавание изображений, обнаружение объектов и обработку естественного языка. Однако CNN уязвимы [1] к состязательным атакам, цель которых - привести нейронную сеть к некорректному отклику, добавляя небольшие возмущения к входным данным[2]. Одним из типов состязательной атаки является однопиксельная атака, при которой модифицируется только один пиксель. Обнаружение таких атак необходимо для обеспечения надежности и безопасности систем искусственного интеллекта.

Наличие искаженных данных может поставить под угрозу точность CNN и сделать их ненадежными в критически важных приложениях, таких как медицинская диагностика, распознавание лиц и автономные транспортные средства. Следовательно, обнаружение искаженных данных имеет решающее значение для обеспечения целостности и надежности сверточных нейронных сетей.

Существующие алгоритмы обнаружения однопиксельной атаки имеют как свои достоинства так и недостатки. [3]

Основная часть. Для обнаружения атак с одним пикселем разработаны подходы на основе анализа статистических свойств входных данных.

Возможность применения методов статистического анализа основана на предположении, что состязательные возмущения вызывают отклонение распределения значений пикселей от того, что можно было бы ожидать для немодифицированных изображений. Обнаружение вредоносного возмущения, характерного для однопиксельной атаки, на изображении возможно посредством методов статистического анализа в том числе в различных комбинациях.

Для обнаружения вредоносных возмущений применяются следующих методов статистического анализа [4]:

1. анализ Z-оценки;
2. анализ гистограммы;
3. расстояние Махаланобиса;
4. обнаружение выбросов.

Ввиду наличия различных недостатков в каждом статистическом методе для решения задачи обнаружения однопиксельной атаки также могут быть использованы их комбинации. Так анализ Z-оценки может не работать на значениях пикселей с ненормальным распределением, анализ гистограммы требует большого набора данных для точного сравнения, также как анализ расстояния Махаланобиса, а анализ обнаружения выбросов может привести к большому количеству ложноположительных или ложноотрицательных результатов.

Для оценки эффективности изложенного подхода к обнаружению однопиксельных атак могут быть использованы следующие метрики: доля ошибок 1 и 2 родов, временная сложность алгоритмов.

Выводы. В текущей работе были определены наиболее подходящие практики применения методов статистического анализа для обнаружения возмущения на изображении,

характерного для однопиксельной атаки. Использование комбинации различных методов статистического анализа даёт возможность более эффективного обнаружения атаки по сравнению с уже существующими методами, за счет уменьшения влияния их недостатков.

Список использованных источников:

1. Su J., Vargas D. V., Sakurai K. One pixel attack for fooling deep neural networks //IEEE Transactions on Evolutionary Computation. – 2019. – Т. 23. – №. 5. – С. 828-841.
2. Kaziakhmedov E. et al. Real-world attack on MTCNN face detection system // 2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON). – IEEE, 2019. – С. 0422-0427.
3. Husnoo M. A., Anwar A. Do not get fooled: defense against the one-pixel attack to protect IoT-enabled deep learning systems //Ad Hoc Networks. – 2021. – Т. 122. – С. 102627.
4. “One pixel attack for fooling deep neural networks” Jiawei Su, Danilo Vasconcellos Vargas, Sakurai Kouichi // IEEE Transactions on Evolutionary Computation. pp. 828–841 2019.

Сайдумаров С.К. (автор) Подпись

Есипов Д.А. (автор) Подпись

Роговой В. (автор) Подпись

Сулименко Н.С. (автор) Подпись

Попов И.Ю. (научный руководитель) Подпись