УДК 004.056

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

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

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

Введение. Сверточные нейронные сети (Convolutional Neural Network, CNN) широко используются для обработки изображений и естественного языка, однако они могут быть подвержены состязательным атакам [1], таким как однопиксельная атака, которая изменяет только один пиксель на входном изображении, но приводит к некорректному отклику модели. Исправление последствий таких атак имеет значение в различных сферах, в которых применяется обработка изображений. Примерами критически значимых сфер являются автономные транспортные средства и системы распознавания лиц в магазинах с автоматическим процессом приобретения товара.

Наличие искаженных данных ставит под угрозу достоверность классификации CNN [2], что повышает риск возникновения неблагоприятного инцидента, например в случае некорректно распознанном автомобилем автомобильного знака, что потенциально приводит к аварии.

Существующие алгоритмы исправления искажений изображения вызванных однопиксельной атакой имеют как достоинства, так и недостатки, связанные с их применимостью и скоростью исправления [3].

Основная часть. Исправление искажений, вызванных однопиксельной атакой, является важным шагом для обеспечения надежности и безопасности CNN.

Один из подходов к исправлению последствий состязательных атак – использование методов статистического анализа, которые могут помочь обнаружить атаку и исправить искаженные данные. Для этого применяются методы анализа статистических свойств входных данных, которые позволяют определить наличие искаженного пикселя и восстановить его изначальное значение [2].

Подход к восстановлению искаженных данных предусматривает использование методов анализа Z-оценки из-за простоты реализации, анализа гистограммы из-за способности обнаруживать ненормально распределенные значения пикселей, анализа расстояния Махаланобиса из-за того, что данный метод учитывает ковариацию значений пикселей, анализ обнаружения выбросов из-за простоты реализации, а также из комбинаций для снижения влияния их недостатков [4].

Применение методов статистического анализа, а также их различных комбинаций позволяют исправлять искажения на изображениях, вызванные однопиксельной атакой.

Выводы. В текущей работе были определены подходы по исправлению искаженных данных, вызванных однопиксельной атакой. Безопасность CNN в критически важных системах может быть обеспечена с использованием методов статистического анализа и их комбинации по исправлению искажений в изображениях. Если гипотезы верны, то при использовании методов математической статистики возможно устранение вредоносного возмущения вызванного однопиксельной атакой на сверточные нейронные сети обработки изображения.

Список использованных источников:

1. 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.

2. 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.

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.

Сулименко Н.С. (автор) Подпись

Есипов Д.А. (автор) Подпись

Роговой В. (автор) Подпись

Сайдумаров С.К. (автор) Подпись

Попов И.Ю. (научный руководитель) Подпись