ПРИМЕНЕНИЕ ОБЪЯСНИТЕЛЬНОГО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ ЗАЩИТЫ ОТ АТАК НА НЕЙРОННЫЙ СЕТИ

Шевченко А.В. [1], **Аверкин А.Н.** [1,2]

[1]Университет «Дубна», Институт системного анализа и управления,
Россия, 141982, Московская обл., г. Дубна, ул. Университетская, д. 19,
Тел.: +7 (914) 625-00-47,
E-mail: leviathan0909@gmail.com
[2]ФИЦ «Информатики и управления» РАН
Россия, 12733, г. Москва, ул. Вавилова, д. 19,
Тел.: +7 (910) 422-71-82,
E-mail: averkin2003@inbox.ru

В базовом функционале многих широко применяемых нейросетевых технологий понятия объяснимости, интерпретируемости и прозрачности первоначально не заложены, отчего алгоритмы принятия тех или иных решений для конечного пользователя остаются “черным ящиком”. Вместе с тем при исследовании математических свойств подобных плохо интерпретированных нейросетевых технологии, особенно глубоких нейросетей, была отмечена их неустойчивость в зависимости от входных данных и возможность модификации входных данных, в результате обработки которых нейронные сети дают ложно-положительные или ложно-отрицательные выводы. Проведя манипуляции с входными данными нейронную сеть можно вывести из строя, заставив выдать заведомо неуместное или опасное решение. В тоже время широкая интеграция нейронных сетей в различные критические области применения дает возможности для атак на системы искусственного интеллекта. Результатом таких атак может стать не только материальный и финансовый ущерб, но и возникнуть угроза жизни и здоровью человека.

Над решением задачи объяснимости процесса принятия решения нейронными сетями в настоящее время работает множество проектов и регулярно публикуются их результаты, например, по проекту DARPA [1], и многочисленные статьи по объяснительному искусственному интеллекту [2, 3, 4], проводятся конференции.

Наиболее рациональным путем решения задачи является необходимость дать пользователям нейронных сетей не только возможность оценивать релевантность алгоритма и результата его работы, но и оценивать достоверность ответа путем косвенных методов их анализа. Сомнение реализуется, не только как вероятностная оценка соответствия, но и как набор параметров, позволяющих указать на значительное количество близких к этому ответу классификаций. Конечная визуализация результатов оценки должна производится в доступном понятном для оценивающего эксперта формате. Для этого должны быть разработаны интерфейсы с внутримодельными и с пост-фактумными объяснениями, позволяющие интерпретировать результаты, полученные искусственной нейронной сетью.

Литература.

1. Gunning D, Vorm E, Wang JY, Turek M. DARPA's explainable AI (XAI) program: A retrospective // Defense Advanced Research Projects Agency. Applied AI Letters, Volume 2, Issue 4, 2021
2. Guidotti R, Monreale A, Ruggieri S, Turini F, Pedreschi D, Giannotti F. A Survey Of Methods For Explaining Black Box Models // Cornell University, Computer Science, Computers and Society, 2018
3. Keppel J, Liebers J, Auda J, Gruenefeld U, Schneegass S. ExplAInable Pixels: Investigating One-Pixel Attacks on Deep Learning Models with Explainable Visualizations // MUM '22: Proceedings of the 21st International Conference on Mobile and Ubiquitous Multimedia, 2022
4. Аверкин А.Н. Объяснимый искусственный интеллект как часть искусственного интеллекта третьего поколения // Всемирный Конгресс, «Теория систем, алгебраическая биология, искусственный интеллект: математические основы и приложения», 2023