APPLYING EXPLAINABLE ARTIFICIAL INTELLIGENCE TO PROTECT AGAINST ATTACKS ON NEURAL NETWORKS

Shevchenko A.V.[1], Averkin A.N.[1,2]

[1]University "Dubna", Institute of System Analysis and Management,
Russia, 141982, Moscow region, Dubna, st. Universitetskaya, 19,
Tel.: +7 (914) 625-00-47,
E-mail: leviathan0909@gmail.com
[2]FRC "Informatics and Control" RAS
Russia, 12733, Moscow, st. Vavilova, 19,
Tel.: +7 (910) 422-71-82,
E-mail: averkin2003@inbox.ru

The basic functionality of many widely used neural network technologies does not initially include the concepts of explainability, interpretability and transparency, which is why algorithms for making certain decisions remain a “black box” for the end user. At the same time, when studying the mathematical properties of such poorly interpreted neural network technologies, especially deep neural networks, their instability depending on the input data and the possibility of modifying the input data were noted, as a result of the processing of which the neural networks give false positive or false negative conclusions. By manipulating the input data, the neural network can be disabled, forcing it to produce a solution that is obviously inappropriate or dangerous. At the same time, the widespread integration of neural networks into various critical application areas provides opportunities for attacks on artificial intelligence systems. The result of such attacks can not only be material and financial damage, but also pose a threat to human life and health.

Many projects are currently working on solving the problem of explainability of the decision-making process by neural networks and their results are regularly published, for example, under the DARPA project [1], numerous articles on explainable artificial intelligence are published [2, 3, 4], and conferences are held.

The most rational way to solve the problem is the need to give users of neural networks not only the opportunity to evaluate the relevance of the algorithm and the result of its work, but also to evaluate the reliability of the answer through indirect methods of their analysis. Doubt is realized not only as a probabilistic assessment of conformity, but also as a set of parameters that make it possible to indicate a significant number of classifications close to this answer. The final visualization of the assessment results must be made in an accessible format that is understandable to the assessing expert. To achieve this, interfaces with in-model and post-hoc explanations must be developed to allow interpretation of the results obtained by the artificial neural network.

References.

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