

11th International Conference "Distributed Computing and Grid
Technologies in Science and Education" (GRID'2025)



Contribution ID: 568

Type: Sectional talk

ЗАЩИТА НЕЙРОСЕТЕВЫХ МОДЕЛЕЙ ОТ СОСТЯЗАТЕЛЬНЫХ АТАК ЧЕРЕЗ ОБЪЯСНИТЕЛЬНУЮ ВИЗУАЛИЗАЦИЮ

Thursday 10 July 2025 15:15 (15 minutes)

В работе представлен анализ современных подходов к обеспечению безопасности и интерпретируемости глубоких нейронных сетей через объяснительные визуализации в контексте противодействия состязательным атакам. Рассмотрены методы объяснимого ИИ, типы состязательных атак и современные стратегии защиты. Практически показано применение визуализаций Grad-CAM для выявления критических областей данных при атаках. В результате работы подтверждена эффективность интеграции объяснительных методов для создания более прозрачных, устойчивых и безопасных моделей искусственного интеллекта, пригодных к внедрению в ответственных сферах, в том числе для повышения надёжности и защищённости распределённых систем.

Authors: SHEVCHENKO, Alexei (Dubna state university); CHEKOLAEV, Dmitry (The Dubna University)

Co-authors: TROFIMOV, Yuri (Государственный университет «Дубна»); АВЕРКИН, Алексей (фиц иу ран)

Presenter: SHEVCHENKO, Alexei (Dubna state university)

Session Classification: Methods and Technologies for Experimental Data Processing