Секція 2

ЗАБЕЗПЕЧЕННЯ ФУНКЦІОНАЛЬНОЇ БЕЗПЕКИ НЕЙРОННИХ МЕРЕЖ В УМОВАХ НЕВИЗНАЧЕНОСТІ

Заіка В. В.
Національний аерокосмічний університет ім. М. Є. Жуковського «ХАІ»
Науковий керівник: Шостак А. В.

Актуальність. Штучний інтелект та нейронні мережі стали невід'ємною частиною критично важливих систем, таких як автоматизовані транспортні засоби, системи медичної діагностики, розумні енергетичні мережі та робототехніка [1]. Однак їх широке впровадження породжує нові ризики для функціональної безпеки, зокрема в умовах невизначеності. До таких ризиків належать помилки через недостатню кількість або низьку якість навчальних даних, уразливості до атак типу "adversarial examples", а також труднощі з прогнозуванням поведінки нейромереж у нестандартних сценаріях [2].

За даними дослідження Microsoft Research, майже 30% автономних систем, що використовують штучний інтелект, демонструють непередбачувану поведінку в умовах неповних або суперечливих даних [3]. Така непередбачуваність створює загрозу для функціональної безпеки систем і вимагає нових підходів до їх проектування, тестування та забезпечення надійності.

Метою даної роботи є вивчення ключових викликів забезпечення функціональної безпеки нейромереж в умовах невизначеності, аналіз сучасних методів підвищення їх стійкості, а також розробка рекомендацій для впровадження цих методів у практику.

Особливу увагу приділено адаптації алгоритмів машинного навчання до роботи в умовах обмеженої інформації, неконтрольованих змін середовища або навмисних атак. Також досліджуються способи оцінки ризиків і тестування моделей у критичних системах, таких як автономні транспортні засоби та медичні пристрої. Це дозволить не лише підвищити надійність таких систем, але й закласти основу для стандартизації процесів забезпечення функціональної безпеки в умовах стрімкого розвитку технологій штучного інтелекту.

Основні положення. Для підвищення надійності нейромереж запропоновано навчання з урахуванням найгірших сценаріїв. Використання підходів, що дозволяють нейромережам бути стійкими до несподіваних змін у середовищі. Інтерпретація результатів роботи

нейромереж. Використання методів Explainable AI для ідентифікації можливих помилок у критичних сценаріях. Захист від атак. Застосування алгоритмів захисту від атак типу "adversarial examples", зокрема обробка вхідних даних за допомогою розсіювання шуму або детекції аномалій.

Висновки. В умовах стрімкого впровадження нейромереж у критично важливі системи забезпечення їх функціональної безпеки стає пріоритетним завданням. Реалізація запропонованих підходів сприятиме зниженню ризиків, пов'язаних з непередбачуваною поведінкою нейромереж, і підвищить надійність таких систем у нестабільних умовах.

Список літератури

1. He K., Zhang X., Ren S., Sun J. Deep Residual Learning for Image Recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016. DOI: https://doi.org/10.1109/CVPR.2016.90.
2. Goodfellow I. J., Shlens J., Szegedy C. Explaining and Harnessing Adversarial Examples. International Conference on Learning Representations (ICLR). 2015. DOI: https://doi.org/10.48550/arXiv.1412.6572.
3. Amodei D., Olah C., Steinhardt J., Christiano P., Schulman J., Mane, D. (2016). Concrete Problems in AI Safety. DOI: https://doi.org/10.48550/arXiv.1606.06565.

Відомості про авторів

Заіка Владислав Віталійович, магістрант кафедри комп'ютерних систем, мереж і кібербезпеки, НАУ «ХАІ», v.v.zaika@student.csn.khai.edu
Шостак Анатолій Васильович, доцент кафедри комп'ютерних систем, мереж і кібербезпеки, НАУ «ХАІ», к.т.н., доцент, a.shostak@csn.khai.edu