ВСТУП

Упродовж останніх років нейронні мережі досягли значних успіхів у задачах класифікації, особливо в галузі комп'ютерного зору. Різні архітектури демонструють різні рівні точності залежно від складності та специфіки завдання.

Серед згорткових нейронних мереж (CNN) AlexNet, представлена у 2012 році, стала проривом, досягнувши топ-5 помилки у 15,3% на змаганні ImageNet Large-Scale Visual Recognition Challenge (ILSVRC). У 2014 році VGGNet з глибшою архітектурою та меншими фільтрами знизила цю помилку до 7,3%. ResNet, представлена у 2015 році, впровадила архітектуру з пропускними з'єднаннями, досягнувши топ-5 помилки у 3,6% на тому ж наборі даних.

Рекурентні нейронні мережі (RNN), зокрема LSTM та GRU, широко застосовуються в обробці природної мови та часових рядів. Наприклад, у задачах передбачення наступного слова в реченні вони демонструють високу точність, значно перевершуючи традиційні моделі.

Генеративно-змагальні мережі (GAN) показали здатність створювати фотореалістичні зображення. Зокрема, модель StyleGAN генерує зображення облич, які важко відрізнити від реальних фотографій.

Трансформери, такі як BERT та GPT, досягли передових результатів у задачах розуміння та генерації тексту, включаючи системи запитань-відповідей та машинний переклад.

Порівняльний аналіз понад 400 нейронних мереж для задачі класифікації зображень показав, що сучасні архітектури, такі як EfficientNet та Vision Transformers (ViT), досягають точності понад 90% на наборі даних ImageNet.

З моменту виявлення у 2013 році феномену змагальних прикладів (adversarial examples), коли незначні, практично непомітні для людини зміни вхідних даних можуть призвести до хибних висновків моделі, дослідження в галузі атак на штучний інтелект значно просунулися. У 2023 року були розроблені більш витончені методи атак, такі як генеративно-змагальні мережі (GAN), здатні створювати високореалістичні підроблені дані, що обходять системи виявлення.

Окрім того, зловмисники почали активно використовувати штучний інтелект для проведення кібератак. 2023 році щонайменше п'ять кібервугруповань застосовували продукти OpenAI для збору інформації з

відкритих джерел (OSINT), що дозволило їм ефективніше планувати та здійснювати атаки.

У відповідь на ці загрози дослідники розробляють нові методи захисту нейронних мереж. Наприклад, запропоновані підходи, засновані на ідентифікації та нейтралізації тригерів закладок (бекдорів), які дозволяють виявляти та усувати приховані вразливості в моделях. Однак, незважаючи на прогрес у цій галузі, універсального рішення, що забезпечує повний захист систем штучного інтелекту від усіх типів атак, поки не знайдено.

Таким чином, забезпечення безпеки та стійкості систем штучного інтелекту залишається актуальним і складним завданням. розвитком технологій штучного інтелекту з'являються нові загрози, які потребують постійної уваги та розробки ефективних методів захисту.

Цей навчальний посібник присвячений аналізу загроз і атак на системи ШІ, зокрема дослідженню змагальних прикладів (adversarial examples), атак на конфіденційність та підробку даних. Детально розглядаються методи впливу на моделі машинного навчання, способи їхнього обходу та експлуатації вразливостей. Окремо висвітлено актуальні напрямки захисту від атак, включаючи алгоритмічні методи підвищення стійкості моделей, а також правові та етичні аспекти безпеки ШІ.

Навчальний матеріал містить як теоретичні аспекти, так і практичні кейси, що дозволяють глибше зрозуміти механізми атак та засоби їхньої нейтралізації. Посібник буде корисним студентам, дослідникам, розробникам та спеціалістам у сфері безпеки, які прагнуть опанувати фундаментальні принципи атак і захисту систем штучного інтелекту.