Станіна О.Д. – старший викладач кафедри економічної та інформаційної безпеки Дніпропетровського державного університету внутрішніх справ, к. т. н.

НЕЙРОННІ МЕРЕЖІ: МОЖЛИВОСТІ ТА ПРОБЛЕМИ ВИКОРИСТАННЯ

Нейронні мережі (НМ) з кожним днем стають все більш розповсюдженими в нашому сьогоденні. Більшість з нас не задумується над тим, що використовує їх кожен день: під час взаємодії з голосовим асистентом Siri, розпізнавання обличчя для розблокування мобільного телефону, при побудові маршруту до найближчої кав'ярні, пошуку фільму тощо. Потенціал НМ дуже високий, тому можна зробити припущення, що їх розповсюдження надалі буде лише збільшуватися. І цьому є об'єктивна причина: НМ – це універсальні апроксиматори функцій. Тобто за використання НМ з великою ємністю можна апроксимувати будь-яку нелінійну функцію.

Проте, вони не бездоганні та мають свої недоліки. І це добре було показано Яном Гудфеллоу в його статті [1], у якій він продемонстрував, що при додаванні шуму до початкового знімку панди НМ розпізнавала остаточну картинку як гібона, хоча з початковим зображенням таких проблем не було. Це здається кумедною особливістю НМ, але в реальності може мати досить суттєві негативні наслідки для поліції, медицини, банківської сфери тощо.

Всі атаки на НМ, в залежності від усвідомленості зловмисника, можна умовно поділити на атаки білої та чорної скриньок. Атаки білої скриньки виникають тоді, коли зловмисник має доступ до базової мережі, а отже - до архітектури. А коли архітектура мережі йому відома, він може керувати окремими нейронами, а отже - привести НМ до помилкового висновку. Атака чорної скриньки виникає в тому випадку, коли зловмисник нічого не знає про будову архітектури мережі. Хоча така атака є більш складною, велика кількість злочинців вдається до неї також.

Найбільш розповсюдженим типом атак, в залежності від способу впливу на мережу, є так звані атаки в обхід (evasion attack), сутність яких полягає в тому, що зловмисник, використовуючи початкові дані та шум, створює змагальні приклади - вхідні умови, які модифіковані таким чином, щоб ввести НМ в оману (наприклад, спуффінг-атаки на біометричні системи, які набувають широке розповсюдження [2]).

Найбільш небезпечною вважається так звана отруйна атака (poisoning attack), сутність якої полягає в порушенні процесу навчання НМ завдяки спеціально згенерованим зразкам. Найчастіше такі атаки виникають у випадку наявності онлайн-навчання мережі і лише у виключних випадках - через отримання інсайдерської інформації. Слід зазначити, що зловмисник, який використовує даний тип атак повинен мати високий рівень компетенції в Data Science.

Всі зазначені вище атаки відносяться до програмного типу, але науковці також виділяють фізичний різновид нападу на НМ. Прикладом таких атак можуть виступати так звані «змагальні стікери», певне розташування яких на фізичному

об’єкті призводить до хибного висновку [3].

Зараз вже існує досить велика різноманітність методів навчання НМ, які дозволяють запобігти атакам різних типів. Крім того, розроблено ряд методів для захисту нейронних мереж; найбільш розповсюдженими серед них є:

- Змагальна підготовка – передбачає самостійне створення змагальних прикладів для навчання НМ. Такий метод захисту можна вважати найкращим способом протидії зловмиснику, адже він робить мережу більш стійкою до кібератак.
- Регулізація – допомагає згладжувати границі прийняття рішень між класами та спрощує класифікацію мереж.
- Змішування – дозволяє збільшувати навчальний набір, а отже - знижує залежність класифікації від невеликої кількості нейронів.
- Тестування на проникнення – комбінує усі вищезгадані методи та передбачає залучення спеціалістів з кібербезпеки для виявлення уразливості мережі та розміру можливих збитків.

Отже, виявлення та запобігання вторгнень є однією з найважливіших задач в області інформаційної безпеки. З ростом зацікавленості світу до глибокого навчання, популяризації біометричних систем та автономних машин, повсюдного впровадження штучного інтелекту та НМ у житті сучасної людини все гостріше стає питання інформаційної безпеки та важливості протистоянню кібератакам. Слід пам’ятати, що світ не стоїть на місці, та разом з виникненням нових можливостей постають проблеми щодо їх використання.

Використані джерела:

1. Goodfellow, Ian & Shlens, Jonathon & Szegedy, Christian. Explaining and Harnessing Adversarial Examples. 2014, URL: https://arxiv.org/abs/1412.6572

2. Мирошниченко В.О. Біометрічна ідентифікація клієнтів в банківській сфері. Міжнародна та національна безпека: теоретичні і прикладні аспекти. Матер III Міжнар. наук-практ. конф. (м. Дніпро, 15 бер.2019 р.) Дніпро: Дніпроп. держ. ун-т внутр. справ, 2019, с. 263 - 265.

3. Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, Dawn Song, Robust Physical-World Attacks on Deep Learning Models, 2017 URL: https://arxiv.org/abs/1707.08945