Секция «Искусственный интеллект в контрольно-надзорной деятельности»

Анализ состязательных атак как реальной угрозы безопасности нейронных сетей

Научный руководитель – Лапина Мария Анатольевна

Дюдюн Глеб Дмитриевич
Студент (специалист)
Северо-Кавказский федеральный университет, Институт информационных технологий и телекоммуникаций, Кафедра информационной безопасности автоматизированных систем, Ставрополь, Россия
E-mail: gleb.dudun@gmail.com

На современном этапе информатизации общества нейронные сети как программный инструмент достаточно активно используется во многих сферах человеческой жизни. Но с ростом популярности возрастает и вероятность совершения противозаконных действий по отношению к информационным системам, использующим нейросети. Наиболее потенциально опасной является угроза состязательных атак.

Состязательные атаки (англ. adversarial attacks) — это методы организации вредоносного вмешательства, обманывающие алгоритмы и системы искусственного интеллекта. Этот тип атак основан на использовании специально созданных входных данных, называемых состязательными примерами, с целью изменения работы системы и снижения её эффективности. Состязательные атаки изучаются с 2013 года, когда исследователи впервые обнаружили потенциальные уязвимости в интеллектуальных алгоритмах классификации изображений. Из-за своей простоты и универсальности состязательные атаки ставят под угрозу огромное количество областей социальной и информационной сфер деятельности, активно применяющих системы машинного обучения и искусственного интеллекта.

На сегодняшний день исследователи в области компьютерной безопасности выделяют следующие типы состязательных атак по характеру воздействия на работу нейронной сети [4]:

Атаки со вводом небольших искажений (adversarial examples) – тип состязательных атак, использующий наложение искажений на входные данные с целью изменить результат их распознавания нейросетью.

Атаки на основе вмешательства в обучение – это атаки, при которых злоумышленник изменяет процесс обучения модели, чтобы сделать ее более уязвимой или заставить выдавать необходимые ему ответы.

Атаки с устранением модели (model inversion attacks) – это атаки, направленные на получение конфиденциальной информации, которую модель использует для принятия решений [7].

Атаки на обнаружение и обход модели (evasion attacks) – это атаки, при которых злоумышленник создает входные данные, специально предназначенные для воздействия на определенную модель и направленные на введение её в заблуждение [n1].

Так же следует упомянуть о методах генерирования состязательных примеров:

- Fast gradient sign method (FGSM) – один из самых известных и простых методов генерации состязательных примеров, основанный на расчете градиентов математических функций нейронов модели и определения с их помощью варианта искажения с наибольшим влиянием на результат распознавания.

- PGD метод – это так же метод генерации, использующий расчет градиентов. Но в отличие от FGSM, PGD выполняет несколько итераций перерасчета искажений входных

данных с помощью градиентов, при этом ограничивая изменения пикселей в пределах допустимого диапазона.

- MIM метод – расширенный вариант метода PGD, также применяющий несколько итераций обновления входных данных с помощью градиентов. Однако при этом каждая итерация учитывает предыдущие изменения, что позволяет повысить сложность состязательных примеров, исключая необходимость постоянной перенастройки генератора под различные нейронные сети.

Существует несколько подходов к защите от состязательных атак:

В исследовании [9] авторы представляют обзор состязательных атак и защиты в различных типах данных, включая изображения, графики и тексты. Авторы рассматривают различные способы создания атак и подходы к защите, а также описывают современные тенденции и проблемы в этой области.

В статье [5] представляют методологию моделей глубоких нейронных сетей, устойчивых к атакам противника. Предлагаемый ими подход связан с усложнением процесса обучения моделей и сосредоточении на учёте потенциальных атак, что способствует повышению устойчивости модели.

В статье [4] предложили новую методику защиты, основанную на интерпретации работы нейронной сети для обнаружения данных, подвергшихся состязательным атакам.

Как понятно из сказанного выше, противодействие состязательным атакам состязательных атак является крайне непростой задачей. Поэтому необходимо продолжить и увеличить дальнейшие исследования на высокоэффективных методах обнаружения и защиты от состязательных атак и, что не менее важно, на оперативном устранении последствий их влияния на работу нейронных сетей.

Источники и литература

1) D. Cires¸an, U. Meier, J. Masci, et al. Multi-column deep neural network for traffic sign classification. Neural Networks, 32:333–338, 2012.
2) Goodfellow I. J., Shlens J., Szegedy C. Explaining and harnessing adversarial examples //arXiv preprint arXiv:1412.6572. – 2014.
3) Huang S. et al. Adversarial attacks on neural network policies //arXiv preprint arXiv:1702.02284. – 2017.
4) Ko G., Lim G. Unsupervised detection of adversarial examples with model explanations //arXiv preprint arXiv:2107.10480. – 2021
5) Madry A. et al. Towards deep learning models resistant to adversarial attacks //arXiv preprint arXiv:1706.06083. – 2017.
6) Papernot N. et al. The limitations of deep learning in adversarial settings //2016 IEEE European symposium on security and privacy (EuroS&P). – IEEE, 2016. – С. 372-387.
7) R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with task learning. In Proceedings of the 25th international conference on Machine learning, pages 160–167. ACM, 2008
8) Szegedy C. et al. Intriguing properties of neural networks //arXiv preprint arXiv:1312.6199. – 2013.
9) Xin, Li, Fuxin, Li: Adversarial Examples Detection in Deep Networks with Convolutional Filter Statistics. In ICCV. IEEE Computer Society, 5775-5783, (2017)