Устойчивость импульсных нейронных сетей к вредоносным атакам

М.Ю. Леонтьев[1], Д.И. Антонов[1], С.В. Сухов[1]

[1] *Ульяновский филиал Института радиотехники и электроники им. В.А. Котельникова РАН, Спасская 14, Ульяновск, Россия, 432011*

Аннотация

Исследована устойчивость к вредоносным атакам аналоговых и импульсных искусственных нейронных сетей. Импульсные нейронные сети были получены из аналоговых путем их конвертации. Протестированы несколько методов увеличения устойчивости нейронных сетей к вредоносным атакам. Эксперименты показали, что импульсные нейронные сети с частотным кодированием информации практически так же подвержены вредоносным атакам, что и аналоговые нейронные сети.

Ключевые слова

Импульсные нейронные сети, спайковые нейронные сети, вредоносные атаки, генеративные сети

1. Введение

Генеративные сети (вариационные автокодировщики, генеративно-состязательные сети и т.д.) могут использоваться для восстановления набора данных (например, изображений), на которых они были ранее обучены. Однако обучение генеративных сетей может представлять собой нетривиальную задачу. Недавно было показано [1], что для генерации изображений может быть использована обычная нейронная сеть-классификатор. Особенность данного подхода в том, что сеть-классификатор должна быть устойчивой к вредоносным атакам (adversarial attacks). Под вредоносной атакой понимается искажение (обычно незначительное) входных изображений таким образом, чтобы они были распознаны ошибочно. Изображения из устойчивого к атакам классификатора могут быть получены путем максимизации отклика выходного нейрона нужного класса градиентным спуском (метод максимизации активации) [1].

Уязвимость к вредоносным атакам является одной из причин осторожного использования нейронных сетей в критических важных областях (например, в медицине и на транспорте). В недавних публикациях [2,3] утверждается, что импульсные (спайковые) нейронные сети (ИмНС), могут быть более устойчивыми к вредоносным атакам. Нейроны в данных сетях обмениваются между собой короткими импульсами одинаковой амплитуды (спайками), что отличает их от обычных (аналоговых) нейронных сетей (АНС).

В данной работе мы провели сравнительный анализ устойчивости аналоговых и импульсных нейронных сетей с частотной кодировкой информации к вредоносным атакам при различных методах обучения нейронных сетей.

2. Эксперимент

На первом этапе экспериментов мы обучали аналоговые свёрточные нейронные сети в среде Keras на наборе изображений (использовались наборы MNIST, EMNIST, Fashion-MNIST). Далее аналоговые сети были преобразованы в импульсные с помощью пакета SNN Toolbox [4]. Полученные данным способом ИмНС используют частотное кодирование информации. Стоит отметить, что спайковые сети, полученные путём конвертации, могут не обладать всеми свойствами, которыми обладают спайковые сети с другим типом кодирования информации (временным, фазовым и т.д.) и/или обученные с нуля.

Устойчивость аналоговых и импульсных нейронных сетей к вредоносным атакам тестировалась в нескольких экспериментах. Изображения для вредоносных атак создавались инструментом FoolBox [5]. Пример вредоносного изображения показан на Рисунке 1.

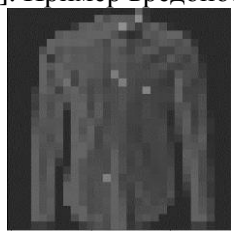
а)  б) 

Рисунок 1: (а) Измененное в FoolBox изображение из набора Fashion-MNIST, ошибочно классифицированное нейронной сетью; (б) пиксельная маска внесённых изменений

Для улучшения способности нейронных сетей противостоять вредоносным атакам, мы добавили в обучающий набор дополнительный класс отрицательных примеров (изображений, не содержащих ни один из желаемых классов). Эксперименты показали, что шумы, поданные на вход сетей, обученных с помощью такого набора данных, с большой уверенностью (свыше 98%) распознаются как «неизвестный объект» в отличие от архитектуры, где такого выхода не было предусмотрено. Такой простой архитектурный трюк открывает таким образом путь к проектированию более устойчивых к атакам нейронных сетей.

3. Заключение

Спайковые нейронные сети, полученные путём конвертации из аналоговых, наследуют уязвимости к различным вредоносным атакам. Разница между точностью классификации в АНС и ИмНС не превышала 1-2%. Без принятия специальных мер обучения и АНС, и ИмНС можно с легкостью обмануть специально подобранными вредоносными примерами. Повышению устойчивости нейросетей к вредоносным атакам может способствовать дополнительное обучение на отрицательных примерах, на атакуемых и/или шумовых изображениях.

4. Благодарности

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта №20-07-00974.

5. Литература

[1] Santurkar, S. Image synthesis with a single (robust) classifier / S. Santurkar, A. Ilyas, D. Tsipras, L. Engstrom, B. Tran, A. Madry // Advances in Neural Information Processing Systems. – 2019. – P. 1262-1273.

[2] Tan, Y.X.M. Exploring the Back Alleys: Analysing The Robustness of Alternative Neural Network Architectures against Adversarial Attacks / Y.X.M. Tan, Y. Elovici, A. Binder // ArXiv preprint arXiv:1912.03609. – 2019.

[3] Sharmin, S. A comprehensive analysis on adversarial robustness of spiking neural networks / S. Sharmin, P. Panda, S.S. Sarwar, C. Lee, W. Ponghiran, K. Roy // International Joint Conference on Neural Networks (IJCNN). – 2019. – P. 1-8.

[4] Rueckauer, B. Conversion of continuous-valued deep networks to efficient event-driven networks for image classification / B. Rueckauer, Y. Hu, I.A. Lungu, M. Pfeiffer, S.-C. Liu // Front. Neurosci. – 2017. DOI: 10.3389/fnins.2017.00682.

[5] Rauber, J. Foolbox: A python toolbox to benchmark the robustness of machine learning models / J. Rauber, W. Brendel, M. Bethge // ArXiv preprint arXiv:1707.04131. – 2017.