А. В. Петров, Т. И. Паюсова
Тюменский государственный университет, г. Тюмень
УДК 004.056

ИССЛЕДОВАНИЕ АТАК НА НЕЙРОСЕТЕВЫЕ МОДЕЛИ ПРИ ПОМОЩИ «ОТРАВЛЕНИЯ» ДАННЫХ И ВЫБОР МЕТОДОВ ЗАЩИТЫ ОТ НИХ

Аннотация. В статье кратко рассмотрены некоторые типы атак путем «отравления» данных на модели машинного обучения, а также проанализированы методы защиты от этих атак с целью выбрать наиболее эффективный и наименее ресурсоемкий.

Ключевые слова: машинное обучение, «отравление» данных, кибератаки, информационная безопасность, искусственный интеллект, нейронные сети.

Введение. На сегодняшний день все большую популярность набирает искусственный интеллект. Даже новичок сегодня способен попробовать натренировать свою собственную модель и попытаться встроить ее в какой-нибудь сервис, однако информация о том, как защищаться от различных атак далеко не так распространена. Даже некоторые коммерческие продукты от крупных компаний не полностью защищены от отравления [1]. Если компании не научатся защищаться от различных атак на их нейросети, то в некоторых сферах применения возможны катастрофические последствия. Например, в области автопилотов и умных автомобилей, медицины, промышленных процессов.

Атака при помощи «отравления» данных — это изменение данных, на которых будет тренироваться новая модель, с целью повлиять на работоспособность этой модели в будущем. Злоумышленник может «отравить» данные так, что модель просто перестанет работать как задумано, либо начнет ошибаться только в случае подачи определенных злоумышленником входных данных. Чтобы проследить как со временем возрастала сложность атак и защитных мер, были рассмотрены статьи, изданные в разное время. К примеру, в наиболее старой [2] статье от Сяо Хана и др. рассмотрена смена меток — простейший метод атаки, не требующий никаких знаний.

В более поздних исследованиях к отравлению подходят гораздо более креативно: к примеру используют признаки изображений [3], как это делал Али Шафахи со своими коллегами, или даже тренируют отдельную нейросеть, чтобы эффективнее создавать яд [4]. Методы защиты также пришли к использованию нейросетей, причем Чэнь Цзянь и др. применили сразу две в своем фреймворке De-Pois [9].

Главная опасность отравления состоит в том, что эффект от него будет постоянным. Обученная нейросеть будет совершать ошибки до тех пор, пока ее не перетренируют.

Испорченные данные получить очень просто. Источником могут быть шутники, умышленно вносящие противоречивые данные при опросах, непроверенные датасеты из интернета, скомпрометированные узлы в распределенных вычислениях и т. д.

Уязвимость к таким атакам присутствует тогда, когда модель проходит обучение, будь то начальная тренировка, или дообучение в процессе эксплуатации.

Проблема исследования. Более наглядно окно уязвимости показывает рис. 1. Существует множество способов защиты, которые эффективны против одних атак, но бесполезны против других. Задачей исследования является обзор существующих методов противодействия отравлению данных, а также выбор наиболее эффективных и универсальных из них.

Виды атак. В данном разделе мы рассмотрим несколько типов атак на примере компьютерного зрения. Очень важно отметить, что по результатам многих тестов видно, эффективность атаки зависит не только от объема отравленных данных и атакуемой модели, но и от датасета, который мы модифицируем.

Первой и самой легкой в исполнении можно считать атаку со сменой метки, или LF атаку [2]. Суть такой атаки в том, что злоумышленник никак не модифицирует сами изображения, но изменяет метки классов лишь на некоторых из них. Если менять метки у изображений одного класса, то вероятность правильного опознания этого класса уменьшится. Если же наоборот — присваивать метку какого-либо класса изображениям, не относящимся к нему, то вероятность ложного отнесения к этому класса повышается.

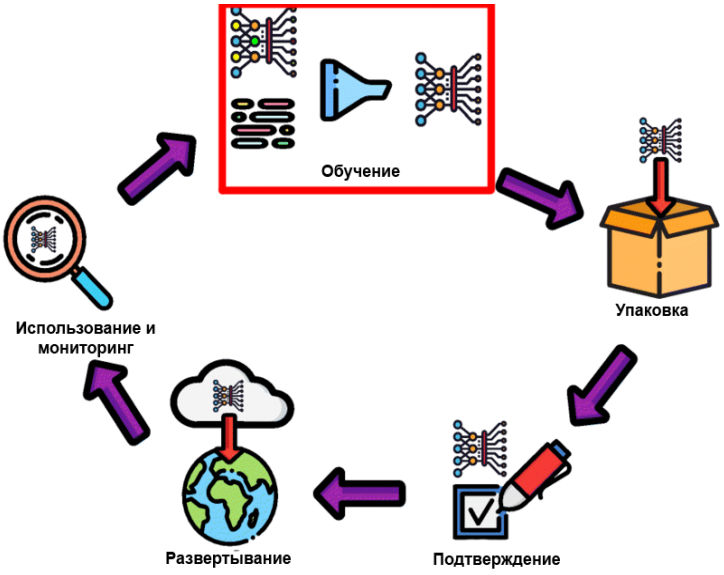

Рис. 1. Уязвимое место в жизненном цикле модели

Различные алгоритмы смены меток могут дать разный результат отравления [2]. Таблица 1 показывает, какой процент данных нужно отравить, чтобы SVM модель начала ошибаться с вероятностью в 50% и выше. Обучение проводилось на датасетах размером 100, 200 и 300 изображений. Впоследствии классификатору выдавали 800 случайных изображений из различных датасетов с реальными данными. Например, для датасета dna нужно случайно изменить 42.5% из 200 изображений. Также видно, что для более точных моделей, использующих радиально-базисные функции для ядра, необходим куда меньший объем отравленных данных.

Следующий тип атак, который нужно рассмотреть — атаки с правильной меткой, или TCL атаки. Они являются гораздо более скрытными, ведь при проверке, с точки зрения человека, все изображения в датасете будут помечены верно. Принцип действия такой атаки заключается в том, что к изображению базового класса добавляется небольшой шум как на рис. 2.

Таблица 1

Зависимость необходимого количества отравленных данных от выбранного датасета и способа смены меток для сильной деградации точности классификации

Data sets	100				200				300			
	Rand.	Near.	Furt.	ALFA	Rand.	Near.	Furt.	ALFA	Rand.	Near.	Furt.	ALFA
SVM with linear kernel												
a9a	41.9	70.4	**29.5**	31.5	43.7	72.2	**27.1**	29.8	44.5	72.9	**26.7**	29.9
acoustic	38.5	77.6	19.2	**17.1**	41.5	77.4	18.8	**17.3**	42.5	76.6	18.8	**17.4**
connect-4	38.2	67.7	**27.7**	29.1	40.1	73.7	**24.4**	27.5	42.2	77.3	**21.4**	25.2
covtype	32.1	73.7	25.0	**23.8**	37.0	74.4	24.6	**22.6**	36.9	75.1	23.9	**21.7**
dna	**43.4**	47.6	50.7	47.8	**42.5**	51.6	45.8	44.2	43.5	54.6	**42.6**	43.2
gisette	47.7	56.6	43.7	**43.6**	47.0	61.8	**37.9**	37.9	47.6	63.8	**35.6**	35.6
ijcnn1	33.9	62.6	26.5	**25.4**	37.9	72.7	21.5	**20.8**	38.2	76.4	19.7	**17.6**
letter	36.7	80.6	**18.2**	19.0	40.2	82.6	**17.1**	18.6	41.5	82.1	**17.4**	19.1
seismic	38.7	73.8	26.3	**25.5**	40.7	71.3	**28.3**	28.7	41.3	70.7	28.8	**28.1**
satimage	44.5	70.5	**30.0**	32.2	45.4	70.3	29.8	**25.5**	46.4	69.2	30.6	**22.3**
SVM with RBF kernel												
a9a	21.6	65.3	12.8	**7.7**	31.5	74.9	18.8	**12.0**	36.1	76.1	20.4	**14.1**
acoustic	6.3	14.7	4.1	**2.9**	16.3	36.8	10.2	**7.1**	22.6	52.7	13.7	**7.8**
connect-4	7.2	33.8	3.7	**2.8**	18.5	68.8	8.7	**5.3**	25.2	76.2	12.3	**6.8**
covtype	2.5	13.2	1.8	**1.4**	6.6	55.8	4.3	**2.2**	11.6	71.2	7.3	**3.9**
dna	27.6	53.6	20.8	**11.6**	40.9	63.7	31.6	**17.0**	46.7	66.5	32.6	**19.2**
gisette	29.4	68.9	23.4	**14.1**	38.7	70.8	28.4	**17.8**	43.4	69.2	29.0	**19.3**
ijcnn1	8.1	27.2	4.2	**3.5**	19.4	41.0	13.6	**8.4**	25.0	40.3	20.4	**10.4**
letter	22.6	78.0	11.7	**8.0**	31.0	84.4	14.1	**10.9**	35.3	84.5	14.2	**11.9**
seismic	11.0	33.4	6.4	**4.3**	24.0	64.4	13.5	**7.4**	29.3	69.0	16.4	**9.6**
satimage	39.1	69.2	25.5	**23.7**	41.8	68.8	28.7	**22.3**	43.4	67.8	30.3	**23.3**

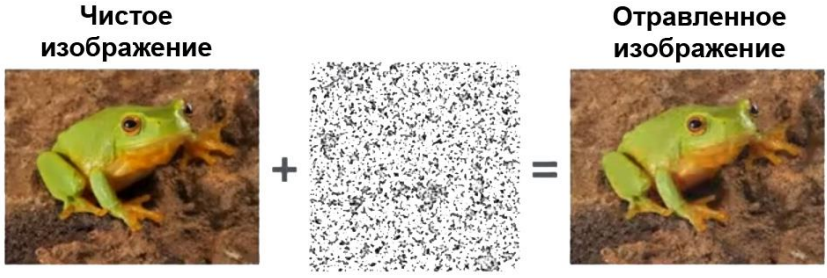

Рис 2. Пример модификации изображения

Такое изображение относится к базовому классу (лягушка), но, благодаря шуму, в поле признаков очень сильно приближено к изображениям целевого класса, в данном случае изображениям самолетов [3]. Следующая функция определяет принадлежность к тому или

иному классу. Первое слагаемое следующей функции показывает сходство в поле признаков, а второе — сходство в пикселях.

$$p = argmin \left\| f(x) - f(t) \right\|_2^2 + \beta \left\| x - b \right\|_2^2$$

Если модели выдать достаточно большое количество таких отравленных изображений, то она начнет ошибочно определять объекты целевого класса, как объекты базового. Проведение такой атаки требует некоторых знаний об алгоритме извлечения признаков, который используется в атакуемой модели, поэтому такие атаки проводить сложнее, а также они требуют от атакующего больше вычислительных мощностей.

Атаки с «чистой» меткой могут быть чрезвычайно эффективны, так как процентный шанс ошибки может в десятки раз превысить процентный объем отравленных данных [1], как это показано на рис. 3. В исследовании были протестированы несколько популярных моделей с разными парами базового и целевого классов.

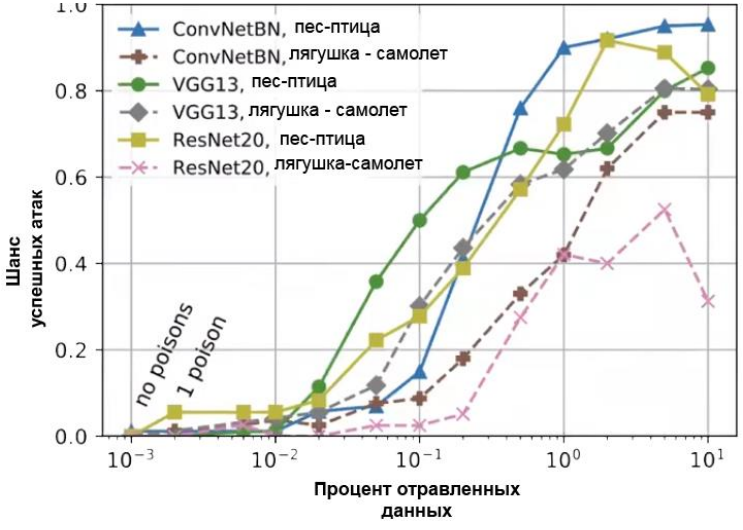

Рис. 3. График зависимости вероятности ошибки от объема отравленных данных

Последний рассматриваемый тип атак — атаки с использованием отравляющих генеративно-состязательных нейронный сетей, или pGAN атаки. Такие сети отличаются от обычных тем, что помимо стандартных генератора и дискриминатора в них присутствует классификатор [4]. Классификатор — алгоритм, который, в идеале, должен быть копией классификатора атакуемой модели, определяет принадлежность полученного объекта к классу. Дискриминатор все также пытается отличить сгенерированное изображение от настоящего. Генератор — алгоритм который должен генерировать отравленные изображения базового класса, которые содержат некоторые искажения, их хорошо видно на рис. 4. Отравленные изображения помечены как пятерки:

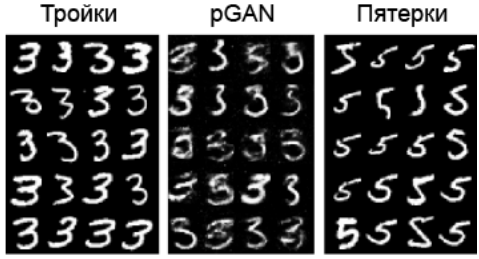

Рис. 4. Пример отравленных изображений, созданных сетью

Задача генератора минимизировать шанс отбрасывания объекта дискриминатором, а также максимизировать шанс ошибки классификатора. Также злоумышленник может управлять соотношением агрессивности и скрытности отравленных данных, получая возможность обходить механизмы защиты, которые рассчитаны на устранение явных выбросов, далеких от границы принятия решений.

Способы защиты. Теперь стоит пройтись по методам защиты от данных атак.

Первый из них — k ближайших соседей [5, 6]. По сути это два метода, которые очень похожи и заключаются в том, чтобы сравнивать метки подозрительных объектов, которые находятся далеко от линии принятия решений, с их ближайшими соседями, как показано на рис 5.

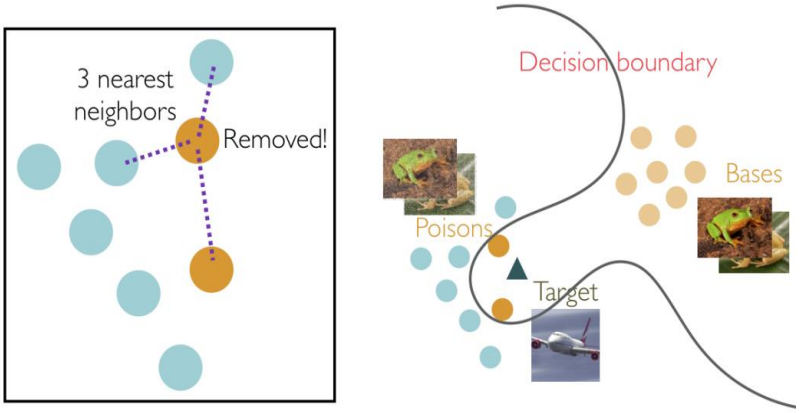

Рис. 5. Наглядная демонстрация сравнения с ближайшими соседями

Это может быть сделано при атаке со сменой метки [5], так и при атаке с правильной меткой [6]. Эффективность такой защиты показана на рис. 6.

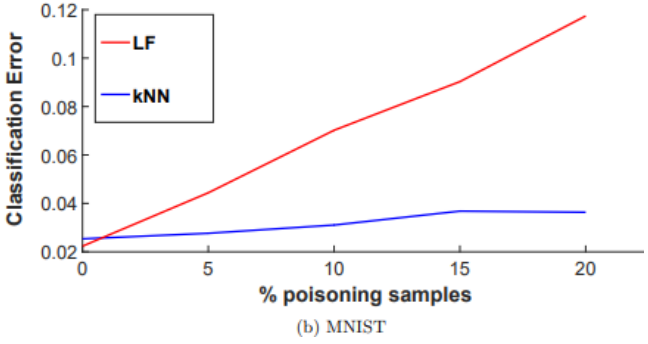

Рис. 6. Сравнение шансов ошибки классификации на датасете MNIST

Проблема заключается в том, что такая защита основана на предположении, что рядом с подозрительным объектом всегда будут «чистые» объекты для сравнения, либо что отравленные объекты будут далеко от линии принятия решений. Этими уязвимостями и воспользовались [7], чтобы обойти оба этих метода защиты.

Следующий метод — комплекс мер под названием certified defenses [8]. Коротко говоря, он нацелен на исключение из датасета всех выбросов и последующее уменьшение ущерба для оставшихся данных. После фильтрации выбросов генерируется верхняя граница эффективности отравления, относительно которой данные фильтруются в дальнейшем. Есть две версии фреймворка, фиксированный и зависящий от данных. Разбирать их подробно не имеет смысла, так как оба строятся на двух допущениях. Первое — один отравленный объект не может оказать сильное влияние на модель. Второе — функция потерь будет выпуклой. Если одно из этих условий будет нарушено, то защита будет гораздо менее эффективной.

Последний рассматриваемый метод — De-Pois [9]. Этот фреймворк призван защищать от любого отравления данных вообще. На доверенном чистом датасете с реальными данными обучается генеративно-состязательная нейросеть, которая создает искусственный датасет с таким же распределением классов. Далее обучается генеративно-состязательная сеть Вассерштайна с целью получить дискриминатор, способный отличить отравленный объект от чистого. В конечном итоге наши потенциально отравленные датасеты должны будут проверяться этим дискриминатором. Эффективность данного метода в сравнении с другими показана в таблице 2. Как видно, De-Pois не превосходит по эффективности узконаправленные методы защиты от конкретных атак, но очень близок к ним.

Таблица 2

Сравнение эффективности различных методов защиты от разных атак при разном объеме отравленных данных

Metric	Attack	Defense	5%$\lvert S_o\rvert$	10%$\lvert S_o\rvert$	15%$\lvert S_o\rvert$	20%$\lvert S_o\rvert$	25%$\lvert S_o\rvert$	30%$\lvert S_o\rvert$
F1-score	TCL-attack	Ours	0.851±0.05	0.904±0.05	0.903±0.03	0.923±0.03	0.892±0.03	**0.956±0.02**
		Deep-*k*NN	**0.948±0.02**	**0.933±0.02**	**0.943±0.03**	**0.939±0.03**	**0.952±0.02**	0.945±0.02
	pGAN-attack	**Ours**	0.678±0.03	0.685±0.03	0.700±0.03	0.703±0.03	**0.713±0.03**	0.722±0.03
		CD	**0.725±0.02**	**0.713±0.02**	**0.721±0.02**	**0.718±0.02**	0.711±0.02	**0.724±0.02**
	LF-attack	**Ours**	**0.930±0.05**	**0.943±0.05**	**0.950±0.05**	**0.954±0.05**	**0.961±0.05**	**0.958±0.05**
		CD	0.826±0.02	0.828±0.02	0.832±0.02	0.837±0.02	0.842±0.01	0.844±0.01

Более поздние исследования показали, что даже этот фреймворк можно обойти. Используя дистилляцию знаний можно создать копию дискриминатора и научиться ее обходить.

Результаты. Авторами были проанализированы некоторые методы атаки при помощи отравления данных, а также некоторые методы защиты. Полный список различных атак огромен и постоянно расширяется, однако на основании проанализированного материала можно сделать вывод, что однозначно эффективной и универсальной защиты не существует. Из всех просмотренных вариантов защиты De-Pois кажется наиболее эффективным. Стоит также отметить, что защита будет эффективной до тех пор, пока потенциальному злоумышленнику неизвестен механизм защиты.

Заключение. В процессе исследования были обнаружены методы защиты, которые являются достаточно эффективными и универсальными, однако полностью защитить модель они не способны. На данный момент наиболее простой и эффективный способ защиты нейросетей от отравления на начальном этапе — избежать получения отравленного датасета, используя доверенные источники и защищая узлы, если имеют место быть распределенные вычисления.

СПИСОК ЛИТЕРАТУРЫ

1. W. Ronny. MetaPoison: Practical General-purpose Clean-label Data Poisoning / Ronny W, Geiping Jonas, Fowl Liam [и др.] // Advances in Neural Information Processing Systems 33 (NeurIPS 2020). — Online : Advances in Neural Information Processing Systems, 2020. — С. 12080-12091. (дата обращения: 25.05.2023). —Text : electronic.
2. Han X. Adversarial Label Flips Attack on Support Vector Machines / Xiao Han, Xiao Huang, Eckert Claudia // Vol. 242: ECAI 2012. — Montpellier : ECAI, 2012. — С. 870-875 (дата обращения: 25.05.2023). —Text : electronic.
3. Ali Shafahi. Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks / Shafahi Ali, Ronny W, Najibi Mahyar [и др.] // Advances in Neural Information Processing Systems 31 (NeurIPS 2018). — Montreal : Advances in Neural Information Processing Systems, 2018 (дата обращения: 25.05.2023). —Text : electronic.

4. Poisoning attacks with generative adversarial nets. Arxiv : [сайт]. — URL: https://arxiv.org/abs/1906.07773 (дата обращения: 25.05.2023). —Text : electronic.
5. Andrea, Paudice Label Sanitization Against Label Flipping Poisoning Attacks / Paudice Andrea, Muñoz-González Luis, C. L. Emil // ECML PKDD 2018 Workshops. — Dublin : ECML PKDD, 2018. — С. 5-15 (дата обращения: 25.05.2023). —Text : electronic.
6. Andrea Paudice. Deep k-NN Defense Against Clean-Label Data Poisoning Attacks / Paudice Andrea, Muñoz-González Luis, C. L. Emil // Computer Vision — ECCV 2020 Workshops. — online : ECCV, 2020. — С. 55-70. (дата обращения: 25.05.2023). —Text : electronic.
7. Pang W. K. Stronger data poisoning attacks break data sanitization defenses / W. K. Pang, Steinhardt Jacob, Liang Percy. // Machine Learning. — Online : Machine Learning, 2022. — С. 1-47. (дата обращения: 25.05.2023). — Text : electronic.
8. Jacob Steinhardt. Certified Defenses for Data Poisoning Attacks / Steinhardt Jacob, Wei,W Pang, S. L. Percy // Advances in Neural Information Processing Systems 30 (NIPS 2017). — Long Beach : Advances in Neural Information Processing Systems, 2017 (дата обращения: 25.05.2023). — Text : electronic..
9. Jian Chen. De-Pois: An Attack-Agnostic Defense against Data Poisoning Attacks / Chen Jian, Zhang Xuxin, Zhang Rui [и др.] // IEEE Transactions on Information Forensics and Security. — Online : IEEE, 2021. — С. 3412-3425 (дата обращения: 25.05.2023). —Text : electronic.
10. Breaking the De-Pois Poisoning Defense. Arxiv : [сайт]. — URL: https://arxiv.org/abs/2204.01090 (дата обращения: 25.05.2023). —Text : electronic.