НЕКОТОРАЯ ОЦЕНКА РИСКА ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ НЕЙРОСЕТЕЙ В ЗАВИСИМОСТИ ОТ МНОЖЕСТВА РАЗЛИЧНЫХ АТАК

Козлов А.Д., Нога Н.Л.
Институт проблем управления им. В.А. Трапезникова РАН, Москва, Россия
alkozlov@ipu.ru, noga@ipu.ru

Аннотация. Предложена методика оценки значения уровня риска информационной безопасности нейронных сетей в зависимости от уровня угроз от множества различных атак методами нечеткой логики и регрессионного анализа, что дает возможность определять степень влияния атак на уровень риска и исключать из рассмотрения атаки, незначительно влияющие на уровень риска.

Ключевые слова: информационная безопасность, нейронная сеть, атака, множественная регрессия, угроза, риск, нечеткая логика.

Введение

В современном мире понятия нейронной сети, искусственного интеллекта, машинного обучения прочно вошли в нашу жизнь и стали уже обыденной реальностью. Практически ни одна сфера повседневной жизни не обходится без использования таких информационных технологий. Эти технологии постоянно развиваются и в том числе нейронные сети все шире используются в различных областях.

Их основное преимущество – обучаемость, возможность быстрой адаптации к быстро меняющимся условиям. Широкое применение нейронные сети нашли в поисковых интернет-системах, переводчиках с различных естественных языков, управлении беспилотным транспортом, космическими аппаратами, диагностических медицинских системах, роботизированных промышленных системах, системах безопасности, работающих с использованием биометрических данных, компьютерных играх и многих других системах. По данным некоторых аналитиков объем мирового рынка технологий, использующих нейросети, превысит в 2024 году 500 млрд. долларов [1].

Нейросети используются также в системах кибербезопасности для выявления попыток несанкционированного доступа. А злоумышленники используют их в целях поиска уязвимостей в системах защиты информационных ресурсов.

Отдельной сферой использования нейросети является воссоздание образов и речи других людей, как реально живущих, так и давно ушедших. Этими возможностями пользуются и представители шоу-бизнеса для замены живых актеров, а также злоумышленники для доступа к конфиденциальным данным. Эта сфера относится к этическим и правовым отношениям. Требуется соответствующая нормативная база, регулирующая эти отношения.

Чем больше различных отраслей экономики, включая критически важные, используют технологию нейросети, тем острее встает вопрос их уязвимости. Это в первую очередь относится к системам, касающимся как физической безопасности граждан, так и их персональных данных. К первым можно отнести автономный транспорт, ко вторым – банковские, правоохранительные, интернет-торговые и другие системы.

Вот только некоторые уязвимости, свойственные нейронным сетям с машинным обучением: «извлечение модели», «инверсия модели», «отравление», «состязание», «уклонение» и другие [1].

Чем выше ценность ресурса, использующего технологию нейросети, тем выше искушение для злоумышленников атаковать эти сети, используя имеющиеся уязвимости. В этой связи становится актуальной задача оценки риска информационной безопасности конкретной нейронной сети в условиях возможности множества атак из различных источников угроз.

1. Краткий обзор некоторых атак на системы машинного обучения и нейронные сети

В настоящее время существует множество различных публикаций с описанием атак на интеллектуальные системы. Как правило с ростом популярности какого-либо научно-технического направления, в частности развития нейронных сетей, активность злоумышленников увеличивается в соответствии с этим ростом. Состязательные атаки на системы машинного обучения акцентированы на нарушении функций безопасности и надежности при работе различных технологических систем. Часто эти атаки используют не только изменение набора данных машинного обучения, но и нередко изменение самой модели машинного обучения. В большинстве случаев атаки можно различать по объему знаний злоумышленником об атакуемой системе [2]. Т.е., если злоумышленник имеет доступ

к параметрам модели и к данным модели, то это классифицируется как «белый ящик». Если нет такого доступа, то это классифицируется как «черный ящик». Возможны и половинчатые решения.

Например, достаточно распространена среди злоумышленников для обмана нейронных сетей, используемых для распознавания изображений, атака уклонением на основе метода быстрого градиентного знака (FGSM) [3]. Атака FGSM заключается в том, чтобы изменить изображение незначительно таким образом, чтобы обученная модель ошибочно идентифицировала его другим классом.

В работе [4] рассматривается модификация метода FGSM – итеративный градиентный метод IGSM, отличающийся тем что атака проводится за несколько итераций метода FGSM.

В случае применения метода атаки BIM (базовый итерационный метод) [5] вносятся изменения во входные данные, которые используются для машинного обучения моделей. В результате такой атаки обучаемая система может делать ошибочные выводы либо некорректные действия.

Также представляют опасность атаки с нулевым запросом (Zero-Query transfer attacks –ZQA) [6]. Здесь предполагается обмен знаниями между моделями, т.е. используются в качестве объекта атаки выводы моделей, а не входные данные. Вначале идет присваивание неверных меток объектам. Затем разрабатывается и используется новая структура данных, которая позволяет фильтровать планы атак и выбирать наиболее успешный.

Атаки с использованием генеративно-состязательной сети (GAN) [2] заключаются в тренировке нейронной сети – генератора и нейронной сети – дискриминатора, которые передают друг другу данные и таким образом проходят обучение. Вначале генератор передает поддельные примеры вместе с настоящими примерами дискриминатору. При этом дискриминатор обучается отличать настоящие данные от неверных и делится опытом с генератором. Далее генератор, используя полученные знания о структуре данных организует атаки на модель машинного обучения.

Вызывают интерес и атаки типа формирования карты значимости на основе Якобиана (Jacobian Saliency Map Attack - JSMA), которая использует матрицу Якоби для выявления входных параметров, которые могут более всего повлиять на выходной результат в данной нейронной сети, т.е. имеют наибольший вес. [7]. При такой атаке требуется большое время вычислений для перебора всех входных параметров.

Часто приходится иметь дело с атакой типа DeepFool [8], где используется минимальное возмущение для изменения решения классификатора. При этой атаке генерируются небольшие изменения, с минимальными расстояниями между исходными входными данными и ошибочно интерпретируемыми выходными данными. При этом требуется существенно больше вычислительных ресурсов относительно атак типа FGSM и JSMA.

Во множестве случаев для атаки используется уязвимость, свойственная для компьютерных атак с внедрением программных закладок (бэкдор). Такие закладки какое-то время остаются скрытыми, затем при вводе модифицированных данных, содержащих триггеры, активируются и наносят ущерб всем компонентам нейронных сетей, что может привести к полному прекращению функционирования систем искусственного интеллекта [9].

В работах [1, 2, 10] на данный момент представлен наиболее обширный перечень, как атак на интеллектуальные системы, так и методов защиты от этих атак.

В работах [11-12] авторами была предложена методология, которая, используя теорию нечеткой логики и эконометрики, позволяет осуществлять оценку риска информационной безопасности. На основе этой методологии можно проводить и оценку уровня угрозы для нейронных сетей с учетом многообразия атак.

2. Задача оценки риска угрозы информационной безопасности нейронных сетей

В данной работе рассмотрим зависимость уровня риска информационной безопасности нейронных сетей от уровня угроз при реализации различных атак на рассматриваемую систему. В качестве примера рассмотрим некоторую нейронную сеть с машинным обучением, основная задача которой заключается в распознавании образов. К таким системам могут относиться транспортные интеллектуальные системы, беспилотные системы, системы распознавания людей в потоке, различные диагностические системы и другие. Во всех этих системах ошибка при распознавании может привести к значительному ущербу.

Как уже указывалось выше нейронная сеть может быть подвержена одновременно множеству видов атак. Предлагаемая методика позволяет при оценке риска учитывать любое количество различных видов атак. Для упрощения демонстрации самой методики ограничимся в рассматриваемом примере пятью видами атак, приведенными в Таблице 1.

Таблица 1. Виды атак

	Название атаки (англ.)	**Название атаки (рус.)**	**Сокращение**
1	Fast Gradient Sign Method	Метод быстрого градиента	FGSM
2	Iterative Gradient Sign Method	Итеративный градиентный метод	IGSM
3	Basic Iterative Method	Базовый итеративный метод	BIM
4	Zero-Query Attacks	Атака с нулевым запросом	ZQA
5	Generative Adversarial Network	Генеративно-состязательная сеть	GAN

Первые три атаки относятся к типу «белый ящик», две другие – «черный ящик». В общем виде в нашем примере риск информационной безопасности рассматриваемой сети можно представить в виде функции:

$$R=R\ (U_{FGSM},\ U_{IGSM},\ U_{BIM},\ U_{ZQA},\ U_{GAN}), \quad (1)$$

где R – уровень риска, U_{FGSM} – уровень угрозы от атаки FGSM, U_{IGSM} – уровень угрозы от атаки IGSM, U_{BIM} – уровень угрозы от атаки BIM, U_{ZQA} – уровень угрозы от атаки ZQA, U_{GAN} - – уровень угрозы от атаки GAN.

Уровень угрозы от конкретной атаки напрямую связан с возможным ущербом, который может возникнуть в случае успешной реализации той или иной атаки. Чем больше потенциальный ущерб, тем выше уровень угрозы.

Определив на первом шаге перечень переменных, зависимость риска от которых будем определять, на втором шаге определяются границы термов для каждой переменной. Определение границ термов подробно изложены в работах [11, 12].

При этом для упрощения примера приводятся три градации качественного значения уровня угрозы для всех видов атак, а для «уровня риска» четыре градации.

Значения входных и выходных переменных для рассматриваемого примера приведены в Таблицах 2 и 3.

Таблица 2. Значение входных переменных «уровень угрозы» U

Вид атаки	**Качественное значение уровня угрозы**	**Описание угрозы**	**Границы терма**	**Усредненные значения**
FGSM	Низкий	Результаты перестают быть достоверными, но идентификация восстанавливается	0 – 0,25	0,12
	Средний	Значение функции потерь незначительно.	0,2 – 0,45	0,33
	Высокий	Если, независимо от величины шага градиентного спуска значение функции потерь значимо, идентификация нарушена.	0,4 – 1,0	0,7
IGSM	Низкий	Значение функции потерь незначительно.	0 – 0,25	0,12
	Средний	Значение функции после нескольких итераций все еще малозначительно	0,2 – 0,5	0,35
	Высокий	Значение функции потерь после еще нескольких итераций уже значимо, идентификация нарушена.	0,4 – 1,0	0,7
BIM	Низкий	Применяется контроль целостности входных данных и обучение на большом количестве данных. Используются методы дополнительного контроля: применение одноразовых ПИН-кодов или двухфакторной аутентификации.	0 – 0,2	0,1
	Средний	Применяется контроль целостности входных данных и обучение на большом количестве данных.	0,15 – 0,45	0,3
	Высокий	Не применяется контроль целостности входных данных и обучение осуществляется на сравнительно небольшом объеме данных. Не используются методы дополнительного контроля: применение одноразовых ПИН-кодов или двухфакторной аутентификации.	0,4 – 1,0	0,7
ZQA	Низкий	При использовании методов обнаружения аномалий и обучения с «учителем». Обучение	0 – 0,2	0,1

Вид атаки	Качественное значение уровня угрозы	Описание угрозы	Границы терма	Усредненные значения
		моделей и замедление процесса передачи опыта между моделями.		
	Средний	В случае передачи опыта между моделями, используя только выходные данные.	0,15 – 0,45	0,3
	Высокий	Отсутствие обучения и неограниченная передача опыта между моделями.	0,4 – 1,0	0,7
GAN	Низкий	Дискриминатор обучен отличать настоящие данные от поддельных.	0 – 0,2	0,1
	Средний	Генератор научился создавать ложные примеры для дискриминатора, которые трудно отличить от реальных данных.	0,1 – 0,5	0,3
	Высокий	Генератор использует полученные знания о структуре данных и готов сгенерировать атаку на модель машинного обучения.	0,4 – 1,0	0,7

Таблица 3. Значения выходной переменной «уровень риска» R

Переменная	Качественные значения	Границы терма	Усредненные значения
Риск, *R*	Незначительный	0 – 0,21	0,1
	Допустимый	0,16 – 0,41	0,3
	Высокий	0,35 – 0,65	0,5
	Критический	0,60 – 1,0	0,8

Теперь на основе сформированных продукционных правил, приведенных в Таблице 4, получаем совокупность данных (Таблица 5) для исследования влияния атак на нейронную сеть методами регрессионного анализа и решаем нашу задачу по следующему алгоритму, аналогично приведенному в работе [11].

Итак, строим уравнение множественной регрессии с указанными в Таблице 2 переменными. Далее применяем метод наименьших квадратов (МНК) и вычисляем коэффициенты при переменных уравнения регрессии. Затем строим уравнение регрессии в стандартизованном виде и вычисляем стандартизованные коэффициенты нашего уравнения. Наконец проводим анализ на корреляционную зависимость переменных. При необходимости исключаем переменные, слабо влияющие на значение уровня риска. Кроме того, можно вычислить точечную оценку значения уровня риска, а также при необходимости вычислить границы доверительного интервала для этого значения уровня риска.

Таблица 4. Продукционные правила

№ п/п	u_1	u_2	u_3	u_4	u_5	Уровень риска *(R)*
1.	низкий	низкий	низкий	низкий	низкий	Незначительный
2.	низкий	низкий	низкий	низкий	средний	Допустимый
3.	низкий	низкий	низкий	низкий	высокий	Допустимый
4.	низкий	низкий	низкий	средний	низкий	Допустимый
5.	низкий	низкий	низкий	средний	средний	Допустимый
…	…	…	…	…	…	…
154.	средний	высокий	высокий	низкий	низкий	Высокий
155.	средний	высокий	высокий	низкий	средний	Критический
…	…	…	…	…	…	…
241.	высокий	высокий	высокий	высокий	низкий	Критический
242.	высокий	высокий	высокий	высокий	средний	Критический
243.	высокий	высокий	высокий	высокий	высокий	Критический

Для упрощения уравнение (1) представим в следующем виде (2):

$$R = a_0 + m_1u_1 + m_2u_2 + m_3u_3 + m_4u_4 + m_5u_5 + \varepsilon, \quad (2)$$

где $u_1 = U_{FGSM}, u_2 = U_{IGSM}, u_3 = U_{BIM}, u_4 = U_{ZQA}, u_5 = U_{GAN}, \varepsilon$ - погрешность.

Таблица 5. Совокупность данных

	u1	**u2**	**u3**	**u4**	**u5**	**Уровень риска (R)**
1	0,12	0,12	0,1	0,1	0,1	0,1
2	0,12	0,12	0,1	0,1	0,3	0,3
3	0,12	0,12	0,1	0,1	0,7	0,3
…						
155	0,33	0,7	0,7	0,1	0,3	0,8
…						
243	0,7	0,7	0,7	0,7	0,7	0,8

Для вычисления коэффициентов уравнения множественной регрессии применяем МНК. Полученные коэффициенты при переменных несравнимы, поэтому необходимо построить уравнение в стандартизованном виде, где коэффициенты при переменных уже можно сравнивать между собой [13] и, таким образом, выстроить переменные по степени влияния на риск. Что позволяет исключать избыточные переменные.

С помощью опции MS Excel по данным из Таблицы 5 находим стандартизованные коэффициенты уравнение (3).

$$k_R = 0{,}2069k_{u_1} + 0{,}2776k_{u_2} + 0{,}3403tk_{u_3} + 0{,}2491k_{u_4} + 0{,}3880k_{u_5}. \quad (3)$$

где, $k_{u_i} = \frac{u_i - \bar{u}_i}{\sigma_{u_i}}, \bar{k}_R = \bar{k}_{u_i} = 0, \sigma_{k_R} = \sigma_{k_{u_i}} = 1.$

Теперь мы можем определить, какая из объясняющих переменных наименьшим образом влияет на значение уровня риска. В соответствии с этим такую переменную, а именно, угрозу такой атаки можно не рассматривать в дальнейшем.

В нашем примере наименьшее влияние на общее значение риска оказывает уровень угрозы от атаки FGSM, и его можно исключить из рассмотрения, а атака GAN является наиболее опасной для рассматриваемой нейронной сети (системы).

Однако, в нашем примере желательно далее рассматривать все указанные переменные, т.к. все пять угроз существенно влияют на значение уровня риска.

Для оценки совместного влияния указанных параметров на значение уровня риска необходимо вычислить коэффициент множественной корреляции, а также коэффициент множественной детерминации.

$$M_{Ru_1u_2u_3u_4u_5} = 0{,}89565465,\ M^2_{Ru_1u_2u_3u_4u_5} = 0{,}802197252.$$

Вычисленный коэффициент множественной детерминации показывает, что 80,2% изменения риска объясняется влиянием переменных полученного уравнения регрессии, и определяет оценку качества нашей модели регрессии.

Чтобы убедиться в статистической значимости как уравнения регрессии, так и коэффициентов уравнения регрессии можно использовать *F*-критерий Фишера и *t*-критерий Стьюдента.

В примере был рассмотрен вариант оценки риска информационной безопасности нейронной сети в случае осуществления пяти разных атак на систему. В реальной жизни различных атак может быть гораздо больше, и, оценивая риск, необходимо учитывать множество атак на рассматриваемую систему. Предлагаемая методика позволяет за несколько итераций оценить уровень риска при осуществлении различных видов атак, а также установить наиболее опасные из них. Это позволит спланировать комплекс мероприятий по защите, в первую очередь, от наиболее опасных видов атак.

При каждом изменении условий эксплуатации нейронной сети (изменение модели, расширение массива данных, дополнение учителя, установка дополнительных средств защиты и др.) необходимо осуществлять повторную оценку уровня риска. Предлагаемая методика позволяет это делать при приемлемой трудоемкости процесса оценки.

3. Заключение

Совместное использование эконометрических методов и методов нечеткой логики позволяет при решении задачи оценки уровня риска рассмотреть возможность совершения не одной определенной атаки, а множества атак нескольких типов (FGSM, IGSM, BIM, ZQA, GAN) на нейронную сеть. При этом, используя предложенную методику, можно определить степень устойчивости нейронной сети к

различным видам атак, как известным в настоящее время, так и тем, которые могут возникнуть в будущем. Также это позволяет:

- устанавливать степень влияния различных видов атак на безопасность нейронных сетей, включая и их совместное воздействие;
- оценивать риски безопасности нейронной сети при каждом случае выявления новых попыток атаковать данную сеть;
- осуществлять повторную оценку уровня риска при каждом изменении условий эксплуатации нейронной сети;
- определять конкретный комплекс мероприятий по защите нейронной сети от проводимых против нее атак с целью минимизации ущерба.

Литература

1. *Аветисян А.И.* Кибербезопасность в контексте искусственного интеллекта // Вестник Российской академии наук, 2022, том 92, №12, С. 1119–1123
2. *Котенко И. В., Саенко И. Б., Лаута О. С., Васильев Н. А., Садовников В. Е.* Атаки и методы защиты в системах машинного обучения: анализ современных исследований. Вопросы кибербезопасности, 2024. - №1 (59). – С. 24-37.
3. *Дьяченко Р.А., Частикова В.А., Лях А.Р.* Реализация атак уклонением на нейронные сети и методы их предотвращения // Электронный сетевой политематический журнал «Научные труды КубГТУ». 2022. - №5. С. 68–77.
4. *Miyato T., Maeda S.I., Koyama M., Ishii S.* Virtual adversarial training: a regularization method for supervised and semi-supervised learning // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2019. V. 41. N 8. P. 1979–1993.
5. *Mądry A., Makelov A., Schmidt L., Tsipras D., Vladu A.* Towards deep learning models resistant to adversarial attacks // Stat. 2017. V. 1050. P. 9.
6. *Cai Z., Rane S., Brito A.E. and oth.* Zero-Query transfer attacks on context-aware object detectors https://synthical.com/article/dbe0f69d-a12c-4e77-bf01-293e7bf47219/ru (дата обращения 08.04.2024)
7. *Su J., Vargas D.V., Sakurai K.* One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation, 2019, vol. 23, no. 5, pp. 828 – 841.
8. *Carlini N., Wagner D.* Towards evaluating the robustness of neural networks. Proc. of the IEEE Symposium on Security and Privacy (SP), 2017, pp. 39–57. https://doi.org/10.1109/sp.2017.49
9. *Менисов А.Б., Ломако А.Г., Дудкин А.С.* Метод защиты нейронных сетей от компьютерных бэкдор атак на основе идентификации триггеров закладок // Научно-технический вестник информационных технологий, механики и оптики. 2022. Т. 22, № 4. С. 742–750. doi: 10.17586/2226-1494-2022-22-4-742-750
10. *Есипов Д.А., Бучаев А.Я., Керимбай А., Пузикова Я.В., Сайдумаров С.К., Сулименко Н.С., Попов И.Ю., Кармановский Н.С.* Атаки на основе вредоносных возмущений на системы обработки изображений и методы защиты от них // Научно-технический вестник информационных технологий, механики и оптики. 2023. Т. 23, № 4. С.720–733.
11. *Kozlov A. D., Noga N. L.* Applying the Methods of Regression Analysis and Fuzzy Logic for Assessing the Information Security Risk of Complex Systems // Proceedings of the 14th International Conference "Management of Large-Scale System Development" (MLSD). Moscow, IEEE, 2021. URL: https://ieeexplore.ieee.org/document/9600245.
12. *Козлов А.Д., Нога Н.Л.* Методика определения наиболее критичных узлов сетевых информационных инфраструктур с целью обеспечения информационной безопасности // Информационные технологии. - М.: Новые технологии. №6, Т. 29, 2023, С. 296-306.
13. *Елисеева И.И. и др.* Эконометрика // М.: Финансы и статистика, 2003. – С.344.