Инженерный вестник Дона, №3 (2025)
ivdon.ru/ru/magazine/archive/n3y2025/9942

Сравнительный анализ устойчивости нейронных сетей ResNet18 и ResNet50 к состязательным атакам на обучающие множества

Р.Ю. Демина, А.А. Хмелёва, А.М. Меркулова

Астраханский государственный университет имени В.Н. Татищева

Аннотация: Данная статья посвящена сравнительному анализу устойчивости нейронных сетей архитектуры ResNet18 и ResNet50 к состязательным атакам на обучающие множества. Рассмотрен вопрос важности обеспечения безопасности обучающих множеств с учетом роста сфер применения искусственного интеллекта. Описан процесс проведения состязательной атаки на примере задачи распознавания животных. Проанализированы результаты двух экспериментов. Целью первого эксперимента стало выявление зависимости числа эпох, необходимых для успешного совершения состязательной атаки на обучающее множество, от версии нейронной сети архитектуры ResNet на примере ResNet18 и ResNet50. Целью второго эксперимента стало получение ответа на вопрос: на сколько успешны атаки на одну нейронную сеть с помощью модифицированных изображений второй нейронной сетью. Анализ результатов экспериментов показал, что ResNet50 более стоек к состязательным атакам, но дальнейшее совершенствование все же необходимо.
Ключевые слова: искусственный интеллект, компьютерное зрение, ResNet, ResNet18, ResNet50, состязательные атаки, обучающее множество, безопасность обучающего множества, нейронные сети, сравнительный анализ.

Введение

Использование искусственного интеллекта (ИИ) растет экспоненциально, проникая во все аспекты современной жизни и оказывая значительное влияние на различные отрасли, от финансов и здравоохранения до транспорта и развлечений. Нейронные сети, являясь ключевым компонентом современных систем ИИ, демонстрируют впечатляющие возможности в решении сложных задач, которые ранее считались исключительно прерогативой человеческого интеллекта. Эти задачи варьируются от распознавания лиц на смартфонах и персонализированной диагностики заболеваний в медицине до автономного управления транспортными средствами и создания произведений искусства. В частности, задача распознавания изображений стала одной из наиболее часто решаемых с применением методов ИИ, что обусловлено ее широким спектром

практических применений и высокой востребованностью в различных областях [1-3].

Одной из ключевых областей применения распознавания изображений является обеспечение безопасности и контроля доступа. Системы контроля и управления доступом (СКУД), основанные на технологиях распознавания лиц, обеспечивают эффективную защиту контролируемых зон от несанкционированного доступа, автоматизируя процессы идентификации и снижая риск человеческих ошибок. Такие системы находят применение в офисных зданиях, аэропортах, на границе и в других местах, где требуется высокий уровень безопасности. Однако, несмотря на впечатляющую эффективность и широкое распространение, нейронные сети, используемые в этих системах, уязвимы к так называемым «состязательным атакам», что создает серьезные риски для их надежности и безопасности.

Состязательные атаки представляют собой специально разработанные методы обмана нейронных сетей, основанные на внесении небольших, часто незаметных для человеческого глаза изменений во входные данные, что приводит к неправильной классификации изображений. Эти атаки могут привести к тому, что нейронная сеть будет идентифицировать лицо злоумышленника как лицо авторизованного пользователя, позволяя ему получить несанкционированный доступ к защищенной зоне. Подобные атаки на обучающие множества представляют реальную угрозу для безопасности как отдельных людей, так и целых систем, и требуют разработки эффективных методов защиты.

В данной статье будет более детально рассмотрен вопрос устойчивости к состязательным атакам на обучающие множества двух архитектур сверточных нейронных сетей: ResNet глубиной 18 и 50 слоев (ResNet18 и ResNet50) [4]. ResNet, или остаточная нейронная сеть, представляет собой инновационный подход к построению глубоких нейронных сетей, который

позволяет эффективно обучать сети с большим количеством слоев, решая проблему затухания градиента. Сравнительный анализ устойчивости ResNet18 и ResNet50 к состязательным атакам позволит лучше понять, как глубина сети влияет на ее уязвимость к таким атакам, и определить наиболее перспективные направления для разработки более надежных и безопасных систем распознавания изображений.

Состязательные атаки

Состязательные атаки представляют собой целенаправленное искажение входных данных (например, обучающего множества), спроектированное для того, чтобы нейронная сеть допускала ошибки классификации/детектирования [5-7]. Особенностью данных атак является то, что изменения, вносимые в данные, часто остаются незамеченными для человеческого глаза, приводя при этом к сбоям в работе искусственных нейросетевых моделей.

Механизм проведения состязательных атак

Суть состязательных атак заключается в эксплуатации "слабых мест" нейронной сети, используя тщательно разработанные алгоритмы для создания едва заметных модификаций входных данных, которые приводят к ошибочной классификации. Злоумышленник может внести минимальные изменения в изображение, такие как добавление шума, изменение яркости, контрастности или даже незначительные изменения отдельных пикселей. Эти изменения, по отдельности или в комбинации, не влияют на восприятие изображения человеком, но оказываются достаточными для того, чтобы нейронная сеть отнесла объект к другому классу или не смогла его детектировать на изображении. Модифицированное изображение создается с использованием специализированных алгоритмов, которые анализируют и используют "слабые места" нейронной сети, выявляя наиболее чувствительные к изменениям области.

Процесс осуществления состязательной атаки, как правило, происходит в несколько этапов:

1. Загрузка и нормализация исходного изображения. Злоумышленник начинает с загрузки исходного изображения, которое необходимо подвергнуть атаке. Затем, выполняется нормализация значений пикселей, переводя их в диапазон от 0 до 1. Этот шаг важен для обеспечения стабильности и эффективности процесса атаки.

2. Создание "вектора шума". Следующим шагом является создание так называемого "вектора шума". Этот вектор представляет собой небольшие, но целенаправленные изменения в значениях пикселей, которые будут добавлены к исходному изображению. "Вектор шума" разрабатывается с учетом особенностей нейронной сети, чтобы максимально эффективно воздействовать на ее работу.

3. Формирование состязательного образца. К исходному изображению добавляется "вектор шума", формируя модифицированное изображение, которое также называют состязательным образцом (рис. 1). Этот процесс можно описать следующим уравнением:

$$y_i = x_i + r_i$$

где:

– y_i - модифицированное (состязательное) изображение
– x_i - исходное изображение
– r_i - "вектор шума"

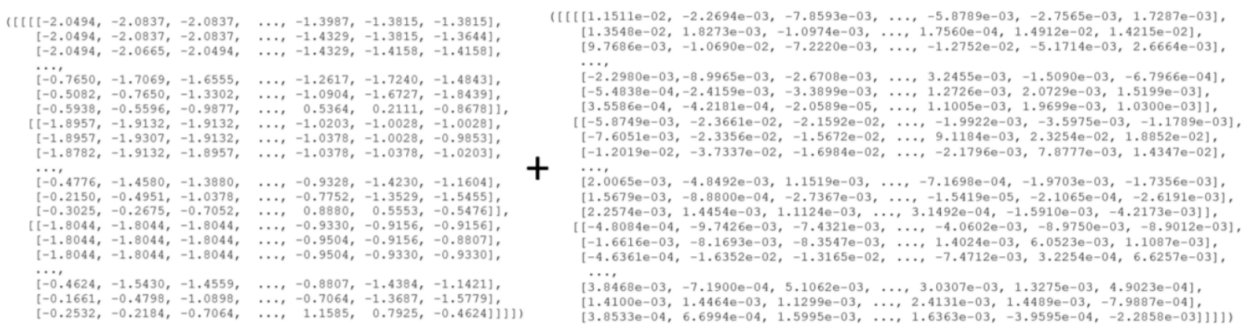

Рис. 1. – Формирование состязательного образца

4. Проверка модифицированного изображения нейронной сетью. Модифицированное изображение подается на вход нейронной сети для оценки результата.

5. Оценка результата и итеративное улучшение "вектора шума". Если нейронная сеть не распознала модифицированное изображение правильно (например, ошибочно классифицировала медведя как петуха), процесс атаки считается успешным и останавливается. Однако, если нейронная сеть продолжает распознавать изображение правильно, алгоритм вычисляет новый "вектор шума" и повторяет процесс, добавляя его к модифицированному изображению. Этот итеративный процесс повторяется до тех пор, пока не будет достигнута цель атаки– ошибочная классификация.

6. Итог атаки. В результате успешно проведенной состязательной атаки получается модифицированное изображение, которое практически не отличается от оригинала для человеческого глаза, но при этом заставляет нейронную сеть ошибочно классифицировать его.

Пример состязательной атаки на ResNet18

Для иллюстрации рассмотрим модель классификации животных, обученную с использованием ResNet18 [8, 9]. Исходное изображение, содержащее фотографию медведя, было подвергнуто целенаправленному изменению в ходе состязательной атаки. В результате, модифицированное изображение для человека все еще выглядит как медведь, но нейронная сеть ошибочно относит его к классу "петух" ("cock") (рис. 2).

Визуально изменения могут быть практически незаметны для восприятия человеческим зрением, но для нейронной сети они оказываются достаточно существенными, чтобы привести к ошибкам классификации или детектирования. Данная особенность приводит к тому, что без проведения специального анализа изображений крайне проблематично выявить факт несанкционированного вмешательства в обучающее множество.

Рис. 2. – Пример состязательной атаки

Состязательные атаки демонстрируют серьезную уязвимость нейронных сетей к злонамеренным и целенаправленным воздействиям. Разработка эффективных методов защиты от таких атак является критически важной задачей, требующей дальнейшего изучения, разработки новых алгоритмов и стратегий, а также повышения устойчивости нейронных сетей к различным видам состязательных воздействий. Понимание механизмов действия состязательных атак является первым шагом к созданию более надежных и безопасных систем машинного обучения.

Сравнение стойкостей ResNet18 и ResNet50 к состязательным атакам

ResNet18 и ResNet50 – это две широко используемые архитектуры глубоких нейронных сетей, входящие в семейство ResNet, разработанные для эффективного решения задач классификации изображений. Ключевым отличием между этими архитектурами является глубина сети, определяемая количеством слоев: ResNet18 содержит 18 слоев, в то время как ResNet50 имеет 50 слоев. Увеличение глубины сети позволяет модели изучать более сложные и абстрактные признаки изображений, что потенциально может повысить точность классификации. Однако, глубина сети также может

влиять на ее устойчивость к различным видам атак, в том числе и к состязательным

Для оценки сравнительной устойчивости [10] ResNet18 и ResNet50 к состязательным атакам была проведена серия экспериментов, целью которых стало:

1. Установить зависимость числа эпох, необходимых для формирования состязательного образца, от версии нейронной сети архитектуры ResNet.

2. Установить можно ли атаковать нейронную сеть ResNet18 с помощью состязательных образцов, сгенерированных с помощью ResNet50, и наоборот.

Эксперимент №1.

Эксперимент по установлению зависимости числа эпох, необходимых для формирования состязательного образца, от версии нейронной сети архитектуры ResNet проходил в рамках нескольких этапов, описанных ниже.

1. Формирование набора данных. В качестве набора данных, для обучения нейронных сетей, были выбраны 10 изображений различных животных, представляющих разнообразие классов и визуальных характеристик:

1 – агама
2 – альбатрос
3 – аксолотль
4 – кот
5 – петух
6 – крокодил
7 – собака
8 – золотая рыбка
9 – сорока
10 – заяц

Выбор небольшого количества изображений на начальном этапе обусловлен необходимостью детального анализа поведения моделей и трудоемкостью процесса генерации состязательных примеров. В дальнейшем планируется расширение набора данных для получения более статистически значимых результатов.

2. Обучение исходных моделей. На выбранном наборе данных были обучены две модели классификации изображений, использующие архитектуры ResNet18 и ResNet50. В процессе обучения применялись стандартные методы оптимизации и регуляризации для достижения высокой точности классификации на исходных изображениях.

3. Генерация состязательных примеров. После обучения исходных моделей был применен алгоритм состязательной атаки для модификации каждого из 10 исходных изображений таким образом, чтобы каждая из обученных моделей относила его к неправильному классу. Важным условием было то, что целевые классы (классы, в которые "превращались" исходные изображения) не должны повторяться, чтобы обеспечить разнообразие состязательных примеров и избежать систематических ошибок.

4. Параметры атаки. Алгоритм состязательной атаки включал в себя итеративное добавление небольших изменений к исходным изображениям до тех пор, пока модель не начинала классифицировать их неправильно. Параметром, который регистрировался в процессе атаки, было количество эпох (итераций), необходимое для успешной генерации состязательного примера. Этот параметр может служить индикатором устойчивости модели к атаке: чем больше эпох требуется для обмана модели, тем более устойчивой она считается.

5. Анализ результатов генерации состязательных образцов. В таблице 1 представлены результаты генерации состязательных примеров для ResNet18 и ResNet50. В столбцах таблицы указаны пары номеров изображений *i*-*j*, где *i* – номер исходного изображения, которое было подвергнуто модификации, а *j* – номер изображения, в которое "превращалось" *i*-ое изображение. В ячейках таблицы указано количество эпох, потребовавшееся для успешной атаки на модели, основанные на ResNet18 и ResNet50.

Таблица № 1

Количество эпох, необходимых для модификации изображений

	1-3	2-4	3-5	4-6	5-7	6-8	7-9	8-10	9-1	10-2
ResNet18	20	15	15	20	45	25	20	15	25	15
ResNet50	11	18	11	15	45	38	12	8	10	15

Анализ результатов, представленных в таблице 1, показывает, что количество эпох, необходимое для успешной атаки, варьируется в широком диапазоне – от 8 до 45. При этом, в большинстве случаев (в шести случаях из десяти), для успешной атаки на модель ResNet18 требовалось большее число эпох, чем для ResNet50. В двух случаях из десяти число эпох для генерации состязательных образцов с помощью разных нейронных сетей было одинаковым.

Эксперимент №2

Для дальнейшей оценки устойчивости моделей к состязательным атакам был проведен эксперимент по "переносу атак". Суть этого эксперимента заключается в том, что состязательные примеры, сгенерированные для одной модели, используются для атаки на другую модель. Если атака успешна, это свидетельствует о том, что модели имеют схожие уязвимости и что состязательные примеры могут быть перенесены с одной модели на другую.

Для проведения эксперимента были обучены еще две модели (ResNet18 и ResNet50) на модифицированных изображениях, полученных на предыдущем этапе.

Итого, для эксперимента использовались

1. Тридцать изображений животных:
    a. 10 исходных.
    b. 10 модифицированных с использованием ResNet18.
    c. 10 модифицированных с использованием ResNet50.
2. Четыре модели:

a. Обученные на исходных оригинальных изображениях:
   i. Модель ResNet18, обученная на 1.a.
   ii. Модель ResNet50, обученная на 1.a.

b. Обученные на соответственно целенаправленных модифицированных изображениях:
   i. Модель ResNet18, обученная на 1.b.
   ii. Модель ResNet50, обученная на 1.c.

Целью данного этапа эксперимента стала оценка количества ошибок, совершаемых каждой моделью при классификации состязательных примеров, сгенерированных для другой модели. В таблицах 2 и 3 представлены результаты классификации состязательных примеров моделями ResNet18 и ResNet50, обученными на соответствующих состязательных примерах. Символ "–" означает, что модель распознала изображение неверно (атака прошла успешно), а символ "+" означает, что модель распознала изображение верно (атака не прошла).

Таблица № 2

Проверка классификации модифицированных изображений на модели ResNet18

	1-18	2-18	3-18	4-18	5-18	6-18	7-18	8-18	9-18	10-18
ResNet18	-	-	-	-	-	-	-	-	-	-
ResNet50	+	+	+	+	+	-	+	+	+	+

Анализ результатов, представленных в таблице 2, показывает, что модель архитектуры ResNet18, обученная модифицированных образцах (с помощью ResNet18), некорректно распознала все тестовые примеры, атака прошла успешно. Модель архитектуры ResNet50, обученная модифицированных образцах (с помощью ResNet18), корректно распознала

все тестовые примеры, кроме одного, атака прошла в большей степени неуспешно.

Аналогичным образом была произведена атака с использованием изображений, модифицированных с помощью ResNet50 (таблица 3).

Таблица № 3

Проверка классификации модифицированных изображений на модели ResNet50

	1-50	2-50	3-50	4-50	5-50	6-50	7-50	8-50	9-50	10-50
ResNet18	-	-	-	-	-	-	-	-	-	-
ResNet50	-	-	-	-	-	-	-	-	-	-

Анализ результатов, представленных в таблице 3, показывает, обе модели не устояли против состязательной атаки с использованием модифицированных образцов, полученных с помощью ResNet50.

Анализ результатов экспериментов

Анализ результатов проведенной серии экспериментов позволяет сформулировать следующие промежуточные выводы:

– Однозначного соответствия между глубиной нейронной сети архитектуры ResNet и числом эпох, необходимых для генерации состязательного образца выявлено не было. Но с большинстве случаев, для атаки на нейронную сеть с меньшим числом слоев потребовалось больше эпох.

– Состязательные примеры, сгенерированные для ResNet18, успешно обманывают модель ResNet18, но в большинстве случаев не обманывают модель ResNet50. Это может свидетельствовать о том, что ResNet50 обладает большей устойчивостью к состязательным атакам, чем ResNet18.

– Состязательные примеры, сгенерированные для ResNet50, успешно обманывают как модель ResNet18, так и модель ResNet50. Это может говорить о том, что атаки, разработанные для более глубокой сети, являются более универсальными и могут быть перенесены на менее глубокие сети.

Результаты проведенного экспериментального исследования показали, что ResNet50 демонстрирует большую устойчивость к состязательным атакам, основанным на представленном в данной статье алгоритме, по сравнению с ResNet18. Однако, следует отметить, что ResNet50 также допускает ошибки при классификации состязательных примеров, что подчеркивает актуальность задачи дальнейшего совершенствования архитектур семейства ResNet и разработки эффективных методов защиты от состязательных атак, эксперименты показали, что ResNet50 более стойка к состязательной атаке, основанной на представленном в данной статье алгоритме. Но следует отметить что она все равно допускает ошибки таким образом актуальна задача дальнейшего совершенствуй архитектур семейства ResNet.

Заключение

Состязательные атаки представляют собой серьезную и растущую угрозу для безопасности и надежности нейронных сетей, особенно в критически важных приложениях, где принятие неверных решений может привести к значительным последствиям. Поэтому, разработка эффективных методов защиты от этих атак является не просто важной, а необходимой задачей для обеспечения доверия к технологиям искусственного интеллекта и их широкого внедрения в различные сферы нашей жизни.

Хотя результаты сравнительного анализа, представленные в данной статье, указывают на то, что архитектура ResNet50 демонстрирует более

высокую устойчивость к определенному типу состязательных атак по сравнению с ResNet18, важно подчеркнуть, что ни одна из этих архитектур не является полностью неуязвимой. Даже небольшие, едва заметные для человеческого глаза изменения во входных данных могут привести к тому, что нейронная сеть примет неверное решение, что ставит под сомнение ее надежность в реальных условиях эксплуатации.

В связи с этим, необходимо продолжать активные исследования и разработки в области повышения защищенности нейронных сетей от состязательных атак, рассматривая эту проблему как многогранную и требующую комплексного подхода. Разработчики должны стремиться к созданию не только более сложных и глубоких архитектур, способных улавливать тонкие закономерности в данных, но и к разработке эффективных методов обучения, позволяющих сети адаптироваться к различным видам атак и сохранять свою точность и надежность в условиях неопределенности.

Особое внимание следует уделить обучению нейронных сетей на больших и разнообразных наборах данных, включающих не только "чистые" примеры, но и состязательные примеры, сгенерированные с использованием различных техник атак. Такой подход позволит сети научиться распознавать и игнорировать состязательные возмущения, повышая ее устойчивость к злонамеренным воздействиям. Кроме того, необходимо разрабатывать и внедрять методы активной защиты, такие, как детекторы состязательных атак и механизмы восстановления входных данных, позволяющие обнаруживать и нейтрализовать атаки в режиме реального времени.

Литература

1. Сасов Д.А., Зубков А.В., Орлова Ю.А., Турицына А.В. Классификация рака молочной железы с помощью сверточных нейронных сетей // Инженерный вестник Дона, 2023, №6. URL: ivdon.ru/ru/magazine/archive/n6y2023/8507.

2. Макаров Р.А. Алгоритм распознавания маркировки грузового контейнера с использованием глубоких нейронных сетей // Инженерный вестник Дона, 2023, №4. URL: ivdon.ru/ru/magazine/archive/n4y2023/8340.

3. Марьенков А.Н., Кузнецова В.Ю., Гелагаев Т.М. Применение технологий распознавания лиц в системах контроля и управления доступом // Прикаспийский журнал: управление и высокие технологии. 2021. №1 (53). С. 83-90.

4. Глебов В.В., Марьенков А.Н. Сравнительный анализ алгоритмов обнаружения человека на изображении // Прикаспийский журнал: управление и высокие технологии. 2023. №2(62). С. 97-106.

5. Чехонина Е.А., Костюмов В.В. Обзор состязательных атак и методов защиты для детекторов объектов. International Journal of Open Information Technologies. 2023;11(7), P. 11–20.

6. Ghaffari Laleh, N., Truhn, D., Veldhuizen, G.P. et al. Adversarial attacks and adversarial robustness in computational pathology // Nature Communications. 2023. №13. doi: 10.1038/s41467-022-33266-0

7. Zhongguo Yang, Irshad Ahmed Abbasi, Fahad Algarni, Sikandar Ali, Mingzhu Zhang An IoT Time Series Data Security Model for Adversarial Attack Based on Thermometer Encoding // Security and communication network. 2021. URL: onlinelibrary.wiley.com/doi/epdf/10.1155/2021/5537041 doi: 10.1155/2021/5537041

8. Sai Abhishek A.V., Gurrala V.R., Sahoo L. Resnet18 Model With Sequential Layer For Computing Accuracy On Image Classification Dataset. // International Journal of Creative Research Thoughts. 2022;10(5). pp. 176–181.

9. Хмелёва А.А., Демина Р.Ю., Ажмухамедов И.М. Проблема компрометации системы распознавания изображений путем целенаправленной фальсификации обучающего множества. Моделирование,

оптимизация и информационные технологии. 2024; №12(2). URL: moitvivt.ru/ru/journal/pdf?id=1535 DOI: 10.26102/2310-6018/2024.45.2.005

10. Durbha K. S., Amuru S., "AutoML Models for Wireless Signals Classification and their effectiveness against Adversarial Attacks", 2022 14th International Conference on COMmunication Systems & NETworkS (COMSNETS), Bangalore, India, 2022, pp. 265-269, doi: 10.1109/COMSNETS53615.2022.9668448.

References

1. Sasov D.A., Zubkov A.V., Orlova J.A., Turicyna A.V. Inzhenernyj vestnik Dona, 2023. №6. URL: ivdon.ru/ru/magazine/archive/n6y2023/8507.

2. Makarov R.A. Inzhenernyj vestnik Dona, 2023. №4. URL: ivdon.ru/ru/magazine/archive/n4y2023/8340.

3. Marenkov A.N., Kuznetsova V.Yu., Gelagaev T.M. Prikaspijskij zhurnal: upravlenie i vysokie tekhnologii, 2021, №1 (53). pp. 83-90.

4. Glebov V.V., Marenkov A.N. Prikaspijskij zhurnal: upravlenie i vysokie tekhnologii, 2023, №2 (62), pp. 97-106.

5. Chekhonina E., Kostyumov V. International Journal of Open Information Technologies. 2023;11(7), pp. 11–20.

6. Ghaffari Laleh, N., Truhn, D., Veldhuizen, G.P. et al. Nature Communications. 2023. №13. doi: 10.1038/s41467-022-33266-0

7. Zhongguo Yang, Irshad Ahmed Abbasi, Fahad Algarni, Sikandar Ali, Mingzhu Zhang Security and communication network. 2021. URL: onlinelibrary.wiley.com/doi/epdf/10.1155/2021/5537041. doi: 10.1155/2021/5537041

8. Sai Abhishek A.V., Gurrala V.R., Sahoo L. International Journal of Creative Research Thoughts. 2022;10(5). pp. 176–181.

9. Hmeljova A.A., Demina R.Ju., Azhmuhamedov I.M. Modelirovanie, optimizaciya i informacionnye tekhnologii 2024; №12(2). URL: moitvivt.ru/ru/journal/pdf?id=1535 doi: 10.26102/2310-6018/2024.45.2.005

10. Durbha K. S., Amuru S., 14th International Conference on COMmunication Systems & NETworkS (COMSNETS), Bangalore, India, 2022, pp. 265-269, doi: 10.1109/COMSNETS53615.2022.9668448.

Дата поступления: 19.01.2025

Дата публикации: 4.03.2025