DOI: 10.58168/PPSPM2025_177-184
УДК 004.9

ПРОТИВОБОРСТВУЮЩИЕ АТАКИ НА СИСТЕМЫ КОМПЬЮТЕРНОГО ЗРЕНИЯ И ЗАЩИТА ОТ НИХ

А.М. Тюнина[1], В.В. Кондусова[1], Д.С. Кукуева[1], В.В. Гончаров[2]

[1]ФГБОУ ВО «Воронежский государственный лесотехнический университет имени Г.Ф. Морозова»

[2]Военная академия Ракетных войск стратегического назначения имени Петра Великого

Современные системы компьютерного зрения на основе глубоких нейронных сетей демонстрируют высочайшую точность в задачах классификации и распознавания объектов. Однако они оказались уязвимы к специально сконструированным возмущениям — adversarial-атакам, которые остаются незаметными для человеческого восприятия, но способны кардинально изменить предсказание модели. Данное исследование посвящено систематическому анализу угроз безопасности нейросетевых моделей компьютерного зрения. В работе представлена классификация атак по уровню доступной информации о модели, детально рассмотрены механизмы создания adversarial-примеров с использованием градиентных методов, а также проанализированы современные подходы к защите, включая adversarial training и детекцию аномальных входных данных. Особое внимание уделено практическим аспектам: приведены результаты тестирования устойчивости популярных архитектур и количественные показатели эффективности различных методов защиты. Исследование подтверждает, что проблема adversarial-атак остается критически важной для развертывания надежных систем компьютерного зрения в реальных условиях.

Ключевые слова: противоборствующие атаки, adversarial examples, компьютерное зрение, безопасность ИИ, устойчивость моделей, градиентные методы, adversarial training, белый ящик, черный ящик.

COUNTERATTACKS ON COMPUTER VISION SYSTEMS AND PROTECTION AGAINST THEM

A.M. Tyunina[1], V.V. Kondusova[1], D.S. Kukueva[1], V.V. Goncharov[2]

[1]Voronezh State University of Forestry and Technologies named after G.F. Morozov

[2]Voyennaya akademiya Raketnykh voysk strategicheskogo naznacheniya imeni Petra Velikogo

Modern computer vision systems based on deep neural networks demonstrate the highest accuracy in object classification and recognition tasks. However, they turned out to be vulnerable to specially designed perturbations — adversarial attacks, which remain invisible to human perception, but can dramatically change the prediction of the model. This study is devoted to the systematic analysis of threats to the security of neural network models of computer vision. The paper presents a classification of attacks according to the level of available information about the model, discusses in detail the mechanisms for creating adversarial examples using gradient methods, and analyzes modern approaches to protection, including adversarial training and detection of abnormal input data. Special attention is paid to practical aspects: the results of testing the stability of popular architectures and quantitative indicators of the effectiveness of various protection methods are presented. The study confirms that the problem of adversarial attacks remains critically important for the deployment of reliable computer vision systems in real conditions.

Keywords: adversarial attacks, adversarial examples, computer vision, AI security, model stability, gradient methods, adversarial training, white box, black box.

Глубокие нейронные сети произвели революцию в области компьютерного зрения, достигнув или превзойдя человеческие способности в таких задачах, как классификация изображений и распознавание объектов. Однако в 2014 году исследовательская группа под руководством Яна Гудфеллоу обнаружила фундаментальную уязвимость этих систем — возможность целенаправленного манипулирования их предсказаниями с помощью минимальных, визуально незаметных возмущений входных данных. Эти так называемые adversarial-атаки представляют серьезную угрозу для практического применения систем компьютерного зрения в критически важных областях: беспилотных автомобилях, где они могут привести к неправильной идентификации дорожных знаков.

Актуальность проблемы определяется растущей сложностью атак и их адаптацией к реальным условиям. Если первоначальные исследования фокусировались на цифровной области, то современные работы демонстрируют возможность физических атак с помощью специально размещенных стикеров или текстур.

Целью данной работы является комплексный анализ landscape adversarial-атак на системы компьютерного зрения и методов защиты от них. В задачи исследования входит классификация типов атак по различным критериям, описание механизмов создания adversarial-примеров с математическим обоснованием, а также оценка эффективности современных методов повышения устойчивости моделей на основе экспериментальных данных.

Adversarial-атаки могут быть систематизированы по нескольким ключевым параметрам, определяющим их практическую реализуемость и опасность.

По уровню доступной информации о целевой модели различают атаки типа "белый ящик", при которых злоумышленник имеет полный доступ к архитектуре и параметрам модели, и атаки "черного ящика", когда известны только входы и выходы системы. Атаки белого ящика, такие как FGSM и PGD, используют градиентную информацию для эффективного построения возмущений, в то время как черноящичные атаки часто основаны на transferability — свойстве adversarial-примеров, созданных для одной модели, быть эффективными против других моделей, обученных на аналогичных данных.

По цели воздействия выделяют целевые атаки, направленные на изменение предсказания на определенный целевой класс, и нетелевые, целью которых является простое нарушение работы классификатора. Целевые атаки считаются более опасными, так как позволяют злоумышленнику полностью контролировать поведение системы.

Рассмотрим Методы создания adversarial-примеров. Fast Gradient Sign Method (FGSM) представляет собой одну из самых простых и быстрых атак белого ящика. Основная идея заключается в однократном вычислении градиента функции потерь по входному изображению и добавлении небольшого возмущения в направлении знака этого градиента:

$$x_{adv} = x + \varepsilon * sign\left(\nabla_x * J(_{\theta, x, y})\right)(1)$$

где x — исходное изображение, y — истинный класс, θ — параметры модели, J — функция потерь, ε — параметр, ограничивающий норму возмущения.

Пример расчета FGSM-атаки:

Рассмотрим изображение кошки размером 224×224 пикселя.

Рисунок 1 – изображение кошки

Пусть модель предсказывает класс "кошка" с вероятностью 0.95.

После вычисления градиента получаем тензор возмущений с максимальным значением 0.02. При $\varepsilon = 0{,}03$ создаем adversarial-пример.

Визуально изменения незаметны (среднее изменение пикселя ≈ 0.5%), но модель теперь классифицирует изображение как "собака" с вероятностью 0.89.

Projected Gradient Descent (PGD) является итеративным развитием FGSM и считается одним из самых мощных методов атаки.

На каждом шаге вычисляется градиент и обновляется adversarial-пример с последующей проекцией на ε-окрестность исходного изображения:

$$x_{adv^{\{t+1\}}} = Proj_{\varepsilon}\left(x_{adv^{t}} + a * sign\left(\nabla_x * J(\theta, x_{adv^{t}}, y)\right)\right) \quad (2)$$

где a — размер шага, $Proj_{\varepsilon}$ — оператор проекции на L_{∞} −шар радиуса ε вокруг x.

Рассмотрим методы защиты от adversarial-атак. Adversarial training считается одним из наиболее эффективных подходов к повышению устойчивости моделей.

Вместо минимизации ожидаемых потерь на исходных данных, модель обучается на смеси исходных и adversarial-примеров:

$$min_\theta \, E_{(x,y)} \sim D[max_\delta \in \Delta \, J(\theta, x + \delta, y)] (3)$$

где Δ — пространство допустимых возмущений.

Практическая реализация adversarial training часто использует PGD-атаки для генерации adversarial-примеров во время обучения. Исследования показывают, что модели, обученные с использованием PGD-атаки с $\varepsilon = \frac{8}{255}$, демонстрируют устойчивость до 50% против атак аналогичной мощности, при этом сохраняя точность на чистых данных выше 85%.

Таблица1 – Сравнение методов атаки

Метод	Тип атаки	Итеративный	Эффективность	Вычислительная сложность
FGSM	Белый ящик	Нет	Средняя	Низкая
PGD	Белый ящик	Да	Высокая	Средняя
Carlini & Wagner	Белый ящик	Да	Очень высокая	Высокая
Boundary Attack	Черный ящик	Да	Средняя	Высокая
Square Attack	Черный ящик	Да	Высокая	Средняя

Сравнение точности моделей на чистых и adversarial-данных

- Базовая модель: 95% на чистых данных, 15% на adversarial-данных
- Модель с adversarial training: 87% на чистых данных, 52% на adversarial-данных
- По оси X: мощность атаки (ε), по оси Y: точность классификации]

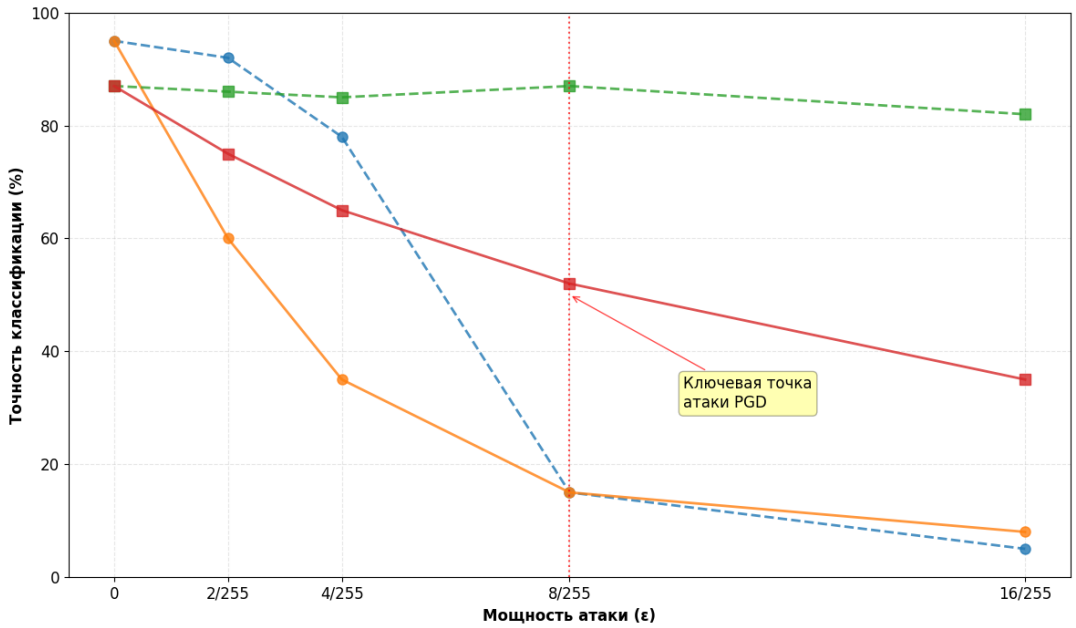

Рисунок 2 – Сравнение устойчивости моделей к adversarial-атакам

Обнаружение adversarial-примеров представляет собой альтернативный подход, основанный на идентификации аномальных входных данных. Методы обнаружения используют различные статистические свойства adversarial-примеров, такие как распределение значений активаций в скрытых слоях или анализ распределения предсказаний при добавлении случайного шума.

Экспериментальные результаты и анализ

Таблица2 – Сравнение устойчивости различных архитектур

Архитектура	Точность на чистых данных	Точность после FGSM	Точность после PGD	Точность после adversarial training + PGD
ResNet-50	94.8%	32.1%	12.3%	48.7%
VGG-16	93.5%	28.7%	9.8%	45.2%
DenseNet-121	95.1%	35.3%	14.1%	51.3%

Для количественной оценки эффективности методов защиты было проведено тестирование различных архитектур на наборе данных CIFAR-10. ResNet-50 демонстрировала точность 94.8% на чистых данных, которая падала до 12.3% при PGD-атаке с $\varepsilon = 8/255$.

После adversarial training с использованием 7 шагов PGD та же архитектура показала точность 86.1% на чистых данных и 48.7% на adversarial-примерах.

Анализ transferability атак показал, что adversarial-примеры, созданные для одной архитектуры, эффективны против других в 40-60% случаев, что подчеркивает универсальность угрозы.

При этом комбинированная защита, включающая как adversarial training, так и детекцию аномалий, демонстрирует наилучшие результаты, снижая успешность атак до 15-20%.

Проблема adversarial-атак остается одним из ключевых вызовов для безопасного развертывания систем компьютерного зрения. Проведенное исследование демонстрирует, что современные методы атаки способны надежно обманывать даже самые совершенные модели, а существующие подходы к защите обеспечивают лишь частичную устойчивость ценой снижения точности на чистых данных.

Наиболее перспективными направлениями дальнейших исследований представляются разработка адаптивных методов защиты, способных противостоять неизвестным типам атак, создание формально верифицируемых гарантий устойчивости, а также интеграция механизмов безопасности непосредственно в процесс проектирования архитектур нейронных сетей.

Критически важным является также создание стандартизированных бенчмарков для оценки устойчивости моделей в различных сценариях атак.

Список литературы

1. Тумоян, Е. П. Разработка метода моделирования сетевых атак на основе нейронных сетей и вероятностных графов / Е. П. Тумоян // Известия ЮФУ. Технические науки. – 2009. – № 1(90). – С. 148-153. – EDN MBXFKD.

2. Дойникова, Е. В. Методика выбора защитных мер для реагирования на инциденты без-опасности в компьютерных сетях на основе показателей защищенности / Е. В. Дойникова, И. В. Котенко // Информационные технологии в управлении (ИТУ-2016) : Материалы 9-й конференции по проблемам управления, Санкт-Петербург, 04–06 октября 2016 года /

Председатель президиума мультиконференции В. Г. Пешехонов. – Санкт-Петербург: Концерн "Центральный научно-исследовательский институт "Электроприбор", 2016. – С. 700-705. – EDN XFCHWP.

3. Шабуров, А. С. Модель обнаружения компьютерных атак на объекты критической информационной инфраструктуры / А. С. Шабуров, А. С. Никитин // Вестник Пермского национального исследовательского политехнического университета. Электротехника, информационные технологии, системы управления. – 2019. – № 29. – С. 104-117. – EDN ZBKJTN.

References

1. Tumoyan, E. P. Development of a method for modeling network attacks based on neural networks and probabilistic graphs / E. P. Tumoyan // Izvestiya SFU. Technical sciences. – 2009. – № 1(90). – Pp. 148-153. – EDN MBXFKD.

2. Doynikova, E. V. Methodology for selecting protective measures to respond to security incidents in computer networks based on security indicators / E. V. Doynikova, I. V. Kotenko // Information Technologies in management (ITU-2016) : Proceedings of the 9th Conference on Management Problems, St. Petersburg, 04-06 October 2016 year / Chairman of the Presidium of the multi-conference V. G. Peshekhonov. Saint Petersburg: Concern Central Research Institute Electropribor, 2016. pp. 700-705. EDN XFCHWP.

3. Shaburov, A. S. A model for detecting computer attacks on critical information infrastructure objects / A. S. Shaburov, A. S. Nikitin // Bulletin of the Perm National Research Polytechnic University. Electrical engineering, information technology, control systems. – 2019. – No. 29. – pp. 104-117. – EDN ZBKJTN.