УДК 004.8:614.2

УСТОЙЧИВОСТЬ НЕЙРОСЕТЕЙ К ЦЕЛЕНАПРАВЛЕННЫМ АТАКАМ В МЕДИЦИНСКИХ СИСТЕМАХ

Жуковец Л.И.
бакалавр, Московский физико-технический институт
(Долгопрудный, Россия)

NEURAL NETWORK ROBUSTNESS TO ADVERSARIAL ATTACKS IN MEDICAL SYSTEMS

Zhukovets L.I.
bachelor's degree, Moscow institute of physics and technology
(Dolgoprudny, Russia)

Аннотация

В статье рассматривается проблема устойчивости нейросетевых моделей к целенаправленным (adversarial) атакам в медицинских системах. Проанализированы основные архитектуры, применяемые в задачах классификации медицинских изображений и анализа биосигналов, с точки зрения их уязвимости к различным типам атак. Представлены методы повышения робастности, включая adversarial training, distillation и предобработку входных данных, а также обсуждаются способы детектирования атак с использованием вспомогательных моделей. Подчёркивается значимость комплексного подхода к защите медицинских ИИ-систем с учётом вычислительных и клинических ограничений.

Ключевые слова: нейросети, целенаправленные атаки, робастность, защита моделей, медицинские изображения, искусственный интеллект, биосигналы, adversarial training

Abstract

The article addresses the issue of neural network robustness against adversarial attacks in medical systems. It analyzes common architectures used in medical image classification and biosignal analysis in terms of their vulnerability to various types of attacks. Several robustness enhancement methods are presented, including adversarial training, distillation, and input preprocessing. The use of auxiliary models for attack detection is also discussed. The study highlights the importance of a comprehensive protection strategy for medical AI systems, considering both computational and clinical constraints.

Keywords: neural networks, adversarial attacks, robustness, model protection, medical imaging, artificial intelligence, biosignals, adversarial training.

Введение

Современные медицинские информационные системы всё чаще используют алгоритмы глубокого обучения для диагностики заболеваний, интерпретации изображений и поддержки клинических решений. Особенно широкое распространение получили нейросетевые модели, демонстрирующие высокую точность в задачах классификации и сегментации медицинских изображений. Однако вместе с ростом их применимости возрастает и уязвимость к специфическим типам атак, среди которых особую угрозу представляют целенаправленные (adversarial) вмешательства. Такие атаки, формируемые путём минимальных, зачастую незаметных для человека искажений входных данных, могут существенно изменить вывод модели, что критично в условиях клинической практики.

В отличие от традиционных атак, целенаправленные вмешательства нацелены на эксплуатацию слабых мест архитектуры нейросети с целью изменения её поведения. Это создаёт риски как для достоверности диагностических решений, так и для безопасности пациентов, особенно в автоматизированных или дистанционных сценариях оказания медицинской помощи. Проблема усугубляется тем, что большинство существующих систем не включает механизмов активной защиты от таких угроз, а сами модели часто обучаются без учёта потенциальных атакующих стратегий [1].

Целью настоящей работы является систематический анализ устойчивости нейросетевых моделей, используемых в медицинских системах, к целенаправленным атакам. В исследовании рассматриваются существующие подходы к построению защищённых архитектур, методы повышения робастности моделей, а также стратегии обнаружения и нейтрализации adversarial-примеров. Отдельное внимание уделяется применимости данных решений в клиническом контексте, где цена ошибки может быть чрезвычайно высокой.

Основная часть

Угрозы целенаправленных атак в медицинских нейросетевых системах

С увеличением использования нейросетевых алгоритмов в медицинских информационных системах встаёт вопрос их устойчивости к внешним воздействиям, в том числе - к целенаправленным атакам. В медицинском контексте данные атаки особенно опасны, поскольку могут привести к диагностическим ошибкам, подрыву доверия к автоматизированным системам и прямому риску для жизни пациентов. Атаки могут быть направлены как на визуальные данные (рентгеновские снимки, МРТ, КТ), так и на табличные или сигнальные медицинские данные, что делает их универсальным инструментом подрыва целостности решений на базе искусственного интеллекта [2].

Целенаправленные атаки в медицинских задачах отличаются от аналогичных вмешательств в других сферах своей сложностью и потенциальными последствиями. Например, минимальные изменения в структуре снимка легких, внесённые на уровне пикселей, могут быть достаточно для того, чтобы модель изменила классификацию «здоров» на «пневмония» или наоборот. Такие манипуляции часто незаметны даже для опытного специалиста, поскольку сохраняют визуальную правдоподобность. Это делает традиционные подходы к верификации на основе визуального анализа недостаточными.

Различают несколько видов целенаправленных атак: **атаки в белом ящике (white-box)**, при которых атакующий имеет доступ к архитектуре и весам модели; **атаки в чёрном ящике (black-box)**, где доступ ограничен только входными и выходными данными; а также **переносимые (transferable) атаки**, способные работать на разных моделях без дополнительной адаптации. Все эти сценарии актуальны в условиях распределённых медицинских систем, где модели развёртываются в облачных средах или на периферийных устройствах. Таким образом, анализ устойчивости моделей к таким воздействиям становится необходимым этапом при внедрении нейросетей в здравоохранение.

Сравнительный анализ уязвимости нейросетевых моделей к целенаправленным атакам

Различные архитектуры нейросетей, используемые в медицинских задачах, демонстрируют неодинаковую степень устойчивости к целенаправленным атакам [3]. Наиболее уязвимыми, как правило, являются глубокие сверточные нейронные сети (CNN), применяемые в анализе изображений, в то время как более компактные и специализированные архитектуры, обученные с применением методов регуляризации и защиты, проявляют относительно большую устойчивость. Проведённый обзор позволяет выделить ключевые характеристики, влияющие на восприимчивость модели к adversarial-примерам: глубина сети, наличие механизмов нормализации, тип функции активации, а также используемая стратегия обучения.

В таблице 1 представлено сравнение наиболее распространённых архитектур, применяемых в медицинских системах, с точки зрения их устойчивости к различным видам

атак. В сравнении учитываются параметры архитектуры, тип используемых входных данных, характер атак и тип реакции модели.

Таблица 1

Устойчивость популярных нейросетевых архитектур к целенаправленным атакам в медицинских задачах

Архитектура	**Тип данных и область применения**	**Устойчивость к атакам и уязвимости**
ResNet-50	Медицинские изображения; используется для диагностики по КТ и МРТ	Низкая устойчивость к атакам в белом ящике, средняя устойчивость в чёрном, высокая уязвимость к переносимым атакам
DenseNet-121	Медицинские изображения; применяется при анализе рентгенограмм	Средняя устойчивость к белому и чёрному ящику, умеренная уязвимость к переносимым атакам
EfficientNet-B0	Медицинские изображения; эффективна в мобильных решениях скрининга	Средняя устойчивость к атакам в белом ящике, высокая - в чёрном, средняя уязвимость к переносимым атакам
LSTM	Биомедицинские временные ряды; применяется для анализа ЭКГ и ЭЭГ	Высокая устойчивость ко всем видам атак, особенно надёжна против переносимых атак
TabNet	Табличные клинические данные; используется для оценки медицинских рисков	Высокая устойчивость в белом ящике, средняя - в чёрном, низкая уязвимость к переносимым атакам

Сравнительный анализ показывает, что устойчивость модели может быть существенно повышена за счёт выбора соответствующей архитектуры, а также при использовании адаптированных к задаче методов защиты. Однако универсальных решений, полностью устраняющих угрозу целенаправленных атак, на данный момент не существует, что подчеркивает необходимость комплексного подхода при проектировании безопасных медицинских систем.

Методы повышения робастности нейросетевых моделей в медицинских приложениях

С учётом высокой чувствительности медицинских систем к целенаправленным атакам становится необходимым внедрение механизмов, обеспечивающих устойчивость нейросетей (робастность) без ущерба для их диагностической точности. Одним из наиболее распространённых подходов является adversarial training - обучение модели на примерах, содержащих заранее сгенерированные adversarial-искажения. Этот метод позволяет повысить чувствительность нейросети к аномалиям и сформировать более устойчивое поведение при встрече с атакующими входами. Однако данная стратегия требует значительных вычислительных ресурсов и может снижать обобщающую способность модели.

Другим направлением защиты является внедрение архитектурных модификаций, таких как защитные слои (defensive distillation), нормализация признаков и регуляризация, уменьшающая чувствительность модели к локальным возмущениям. Также применяются методы стохастической активации и обрезки (pruning), уменьшающие сложность модели и тем самым - её восприимчивость к манипуляциям. Подходы на основе энергетических функций и бэйесовских моделей позволяют дополнительно учитывать неопределённость в прогнозах, что особенно актуально в клинических условиях, где требуется высокая интерпретируемость решений.

Кроме того, растёт интерес к методам предварительной обработки входных данных, которые позволяют нивелировать или удалять потенциальные искажения до подачи информации в модель. Такие методы включают фильтрацию шумов, нормализацию

изображений, а также преобразования, устойчивые к adversarial-эффектам (например, JPEG-компрессия или случайное масштабирование) [4]. Совокупность перечисленных стратегий формирует основу комплексного подхода к обеспечению робастности нейросетевых решений, необходимого при разработке надёжных медицинских систем.

Обзор методов защиты от целенаправленных атак в медицинских нейросетях

Защита нейросетей от целенаправленных атак в медицинских приложениях может быть реализована на различных уровнях - от архитектурной модификации моделей до методов обработки данных и механизмов обнаружения аномалий [5]. Каждый из подходов имеет свои преимущества и ограничения в зависимости от специфики задачи, вычислительных ресурсов и уровня допустимого риска. В условиях здравоохранения особенно актуальны методы, способные обеспечивать интерпретируемость решений и не нарушать доверие со стороны врачей.

В таблице 2 представлены основные категории защитных методов с указанием их применимости в медицинских сценариях, вычислительной нагрузки и устойчивости к различным типам атак.

Таблица 2

Сравнение методов защиты нейросетей от целенаправленных атак в медицинских задачах

Метод защиты	**Принцип действия**	**Применимость в медицине**
Adversarial training	Использование искажённых примеров в обучении для устойчивости модели	Высокая, но требует значительных вычислительных ресурсов
Defensive distillation	Снижение чувствительности модели через сглаживание выходов	Средняя, снижает чувствительность модели
Feature squeezing	Удаление шумов и уменьшение вариативности признаков	Средняя, эффективна в условиях ограниченных ресурсов
Input preprocessing	Предварительная фильтрация входных данных	Высокая, подходит для фильтрации артефактов
Randomized smoothing	Использование усреднённых предсказаний модели	Средняя, баланс между надёжностью и затратами
Bayesian networks	Моделирование неопределённости выходных значений	Средняя, повышает интерпретируемость решений

Таблица демонстрирует, что ни один из представленных методов не обеспечивает полной устойчивости ко всем типам атак без дополнительных затрат. Наиболее эффективным в условиях атак с полной осведомлённостью об архитектуре модели остаётся метод обучения на искажённых данных, однако он требует значительных вычислительных ресурсов. Методы предварительной обработки входных данных и снижение чувствительности к шуму обеспечивают базовый уровень защиты при низкой нагрузке на систему. Подходы, основанные на вероятностных моделях, представляют интерес с точки зрения повышения интерпретируемости результатов и оценки доверия к прогнозам, что особенно важно в медицинских задачах. Таким образом, выбор метода защиты должен опираться на характер клинической задачи, доступные ресурсы и допустимый уровень риска.

Детектирование целенаправленных атак с использованием вспомогательных нейросетей

Одним из перспективных направлений обеспечения устойчивости медицинских систем на базе нейросетей является создание вспомогательных детекторов, способных отличать нормальные входные данные от adversarial-примеров. Такие детекторы обучаются параллельно с основной моделью и могут функционировать как фильтр на этапе предобработки либо как компонент системы принятия решений. Преимущество этого подхода

заключается в возможности адаптации к различным типам атак без необходимости модификации базовой модели [6].

В качестве иллюстрации приводится фрагмент кода на Python, реализующий детектор, обучаемый на признаках скрытого слоя основной нейросетевой модели. Для построения примера используется фреймворк PyTorch и упрощённый датасет.

```
import torch
import torch.nn as nn
import torch.optim as optim
from torchvision import datasets, transforms, models
from torch.utils.data import DataLoader, random_split

# Загрузка и трансформация данных
transform = transforms.Compose([
    transforms.Resize((224, 224)),
    transforms.ToTensor()
])
dataset = datasets.FakeData(transform=transform)  # для примера; в медицине: изображения МРТ, КТ и др.
train_set, val_set = random_split(dataset, [800, 200])
train_loader = DataLoader(train_set, batch_size=32)
val_loader = DataLoader(val_set, batch_size=32)

# Предобученная модель (ResNet18) в качестве экстрактора признаков
resnet = models.resnet18(pretrained=True)
resnet.fc = nn.Identity()  # убираем классификатор, получаем выход скрытого слоя
resnet.eval()

# Детектор аномалий (на выходах скрытого слоя)
class Detector(nn.Module):
    def __init__(self, input_dim=512):
        super(Detector, self).__init__()
        self.classifier = nn.Sequential(
            nn.Linear(input_dim, 128),
            nn.ReLU(),
            nn.Linear(128, 2)  # 0 - чистый пример, 1 - adversarial
        )
    def forward(self, x):
        return self.classifier(x)

detector = Detector()
criterion = nn.CrossEntropyLoss()
optimizer = optim.Adam(detector.parameters(), lr=0.001)

# Обучение детектора
for epoch in range(5):
    detector.train()
    for inputs, _ in train_loader:
        with torch.no_grad():
            features = resnet(inputs)
        labels = torch.randint(0, 2, (inputs.size(0),))  # имитация: случайная разметка (заменить на реальные метки)
        outputs = detector(features)
        loss = criterion(outputs, labels)

        optimizer.zero_grad()
        loss.backward()
```

```
        optimizer.step()

print("Обучение завершено.")
```

Этот код демонстрирует базовую архитектуру вспомогательной модели-детектора, обучающейся на выходах скрытого слоя основного классификатора. В реальных медицинских задачах вместо FakeData следует использовать реальные датасеты (например, CheXpert, BraTS), а метки должны отражать факт наличия adversarial-искажений. Такой подход позволяет повысить устойчивость системы без переработки всей архитектуры [7].

Ограничения и перспективы использования робастных нейросетей в медицинских системах

Несмотря на активное развитие методов защиты нейросетей от целенаправленных атак, их интеграция в медицинскую практику сопровождается рядом ограничений [8]. Одним из главных препятствий остаётся компромисс между устойчивостью и точностью модели. Повышение робастности нередко приводит к снижению диагностической чувствительности, что критично в задачах раннего выявления заболеваний [9].

Другим важным аспектом является ограниченность доступных наборов данных, содержащих adversarial-примеры, специфичных для медицинской области. Это затрудняет обучение и валидацию детекторов атак, а также снижает обобщающую способность защитных механизмов. Отсутствие общепринятых бенчмарков для оценки устойчивости моделей в клинических задачах также препятствует стандартизации подходов.

Тем не менее, перспективы развития в этой области остаются значительными. Разрабатываются гибридные архитектуры, сочетающие традиционные методы машинного обучения и нейросетевые решения с элементами логического вывода. Интеграция с технологиями распределённого обучения (federated learning) позволяет учитывать разнообразие клинических данных без нарушения конфиденциальности. Кроме того, усилия исследователей направлены на повышение интерпретируемости робастных моделей, что особенно важно для обеспечения доверия медицинского персонала к системам искусственного интеллекта [10].

Заключение

Устойчивость нейросетей к целенаправленным атакам приобретает особую значимость в контексте медицинских систем, где точность и надёжность выводов непосредственно влияют на здоровье и безопасность пациентов. Проведённый анализ показывает, что большинство современных архитектур подвержены воздействию adversarial-примеров, особенно в задачах классификации изображений и анализа биосигналов. Это требует обязательного включения механизмов защиты при проектировании и внедрении интеллектуальных медицинских решений. Разнообразие подходов к обеспечению робастности - от adversarial-training до детекторов на скрытых слоях - демонстрирует, что комплексные стратегии являются наиболее перспективными. Однако полная защита от атак пока недостижима, и каждый метод сопряжён с определёнными компромиссами, включая снижение производительности и увеличение вычислительной нагрузки.

Таким образом, обеспечение устойчивости нейросетей в медицинских системах должно рассматриваться как приоритетное направление в области цифрового здравоохранения. Дальнейшие исследования должны быть направлены на создание стандартизированных протоколов оценки робастности, разработку энергоэффективных защитных архитектур и формирование доверительной среды взаимодействия между врачом и системой искусственного интеллекта.

References

1. Javed H., El-Sappagh S., Abuhmed T. Robustness in deep learning models for medical diagnostics: security and adversarial challenges towards robust AI applications // Artificial Intelligence Review. 2024. Vol. 58. No. 1. P. 12.

2. Garifullin R. The use of modern web technologies for ensuring compatibility and interoperability in the context of medical information platforms // International Journal of Humanities and Natural Sciences. 2025. Vol. 1-3(100). P. 99-103.
3. Puttagunta M.K., Ravi S., Nelson Kennedy Babu C. Adversarial examples: attacks and defences on medical deep learning systems // Multimedia Tools and Applications. 2023. Vol. 82. No. 22. P. 33773-33809.
4. Hirano H., Minagi A., Takemoto K. Universal adversarial attacks on deep neural networks for medical image classification // BMC medical imaging. 2021. Vol. 21. P. 1-13.
5. Ghaffari Laleh N., Truhn D., Veldhuizen G.P., Han T., van Treeck M., Buelow R.D., Kather J.N. Adversarial attacks and adversarial robustness in computational pathology // Nature communications. 2022. Vol. 13. No. 1. P. 5711.
6. Garifullin R. Analysis and development of interfaces for emergency systems with multitasking processing: approaches to minimizing user errors and enhancing reliability // International scientific journal "Innovative Science". 2025. № 1-2-1. P. 20-24.
7. Kaviani S., Han K.J., Sohn I. Adversarial attacks and defenses on AI in medical imaging informatics: A survey // Expert Systems with Applications. 2022. Vol. 198. P. 116815.
8. Moradi M., Samwald M. Improving the robustness and accuracy of biomedical language models through adversarial training // Journal of Biomedical Informatics. 2022. Vol. 132. P. 104114.
9. Jakkaraju A. Self-Healing Neural Networks Against Adversarial Attacks // International Journal of Intelligent Systems and Applications in Engineering. 2024. Vol. 12. P. 2537-2549.
10. Kwon H., Lee J. Diversity adversarial training against adversarial attack on deep neural networks // Symmetry. 2021. Vol. 13. No. 3. P. 428.