УДК 004.932.1

ИССЛЕДОВАНИЕ ЭФФЕКТИВНОСТИ СОСТЯЗАТЕЛЬНЫХ АТАК НА МОДЕЛЬ СЕГМЕНТАЦИИ РАКА ПЕЧЕНИ

Нирян Павел Леонидович, студент, направление подготовки 01.03.02 Прикладная математика и информатика, Оренбургский государственный университет, Оренбург
e-mail: niran908@gmail.com

Гришина Любовь Сергеевна, аспирант, направление подготовки 02.06.01 Компьютерные и информационные науки, Оренбургский государственный университет, Оренбург
e-mail: zabrodina97@inbox.ru

Научный руководитель: **Болодурина Ирина Павловна,** доктор технических наук, профессор, заведующий кафедрой прикладной математики, Оренбургский государственный университет, Оренбург
e-mail: prmat@mail.osu.ru

Аннотация. *В последнее время машинное обучение начало использоваться в разных сферах, в том числе и в медицине. Эти системы в большинстве случаев работают в качестве рекомендательных. Но, как и любая другая система, они подвергаются взлому и атаке. В данной статье рассмотрена задача мультиклассовой сегментации изображений рака печени; построенная интеллектуальная система имеет точность больше 98%. Также были проведены различные состязательные атаки, которые показали, что данная система подвержена взлому. Метод Fast Gradient Sign Method (FGSM) дал хороший результат, с помощью небольшого возмущения удалось обойти систему более чем на 80%.*

Ключевые слова: *машинное обучение, состязательные атаки, мультиклассовая сегментация, ResNet50, FGSM, Input Perturbation.*

Для цитирования: Нирян П. Л., Гришина Л. С. Исследование эффективности состязательности атак на модель сегментации рака печени // Шаг в науку. – 2023. – № 4. – С. 79–83.

STUDY OF THE EFFECTIVENESS OF ADVERSARIAL ATTACKS ON A LIVER CANCER SEGMENTATION MODEL

Niryian Pavel Leonidovich, student, training program 01.03.02 Applied Mathematics and Computer Science, Orenburg State University, Orenburg
e-mail: niran908@gmail.com

Grishina Lyubov Sergeevna, graduate student, training program 02.06.01 Computer and information sciences, Orenburg State University, Orenburg
e-mail: zabrodina97@inbox.ru

Research advisor: **Bolodurina Irina Pavlovna,** Doctor of Technical Sciences, Professor, Head of the Department of Applied Mathematics, Orenburg State University, Orenburg
e-mail: prmat@mail.osu.ru

Abstract. *Machine learning has recently begun to be used in various fields, including medicine. These systems, in most cases, work as a recommendation system. But like any `other system, it is susceptible to hacking and attack. This paper examines the task of multiclass segmentation of liver cancer images; the intelligent system built has an accuracy greater than 98%. Various adversarial attacks have also been conducted, which have shown that this system is susceptible to hacking. Fast Gradient Sign Method (FGSM) showed a good result, with a small perturbation it was possible to fool the system by more than 80%.*

Key words: *machine learning, adversarial attacks, multi-class segmentation, ResNet50, FGSM, Input Perturbation.*

Cite as: Niryian, P. L., Grishina, L. S. (2023) [Study of the effectiveness of adversarial attacks on a liver cancer segmentation model]. *Shag v nauku* [Step into science]. Vol. 4, pp. 79–83.

Введение

Сегментация рака печени – одна из самых сложных задач анализа медицинских изображений, требующая высокой точности и надежности. Точная сегментация злокачественной опухоли играет ключевую роль в ранней диагностике и лечении, что делает эту задачу особенно важной [1]. Однако модели сегментации рака печени могут столкнуться с проблемой атак, когда злоумышленники могут изменить данные изображения или внести шум, чтобы изменить результаты сегментации. Это представляет угрозу для точности и надежности моделей сегментации рака печени. Таким образом, атаки противника на модели сегментации рака печени являются актуальной темой исследований. Они могут помочь в оценке устойчивости модели к изменениям входных данных и повысить ее точность и надежность. Следовательно, исследование состязательных атак на модели сегментации рака печени важно для медицинских исследований и может помочь повысить точность диагностики и лечения злокачественной опухоли.

Обзор исследований

Исследованиями в области построения интеллектуальных медицинских систем, а также исследованиями их устойчивости на состязательных атаках занимаются во всем мире.

В работе [4] ученые продемонстрировали существование неблагоприятных примеров практически для всех типов моделей машинного обучения, которые когда-либо изучались. Например, шум, наложенный на фотографию доброкачественной родинки и незаметный для человеческого глаза, позволяет обойти модель, заставляя ее классифицировать эту родинку как злокачественную с вероятностью 100%. В исследовании [5] авторы пришли к выводу, что в трех наборах данных (фундоскопии, рентгенографии грудной клетки и дерматоскопии) для успешной атаки требуется небольшое возмущение на входное изображение, чтобы обойти классификатор. Но эти атаки могут быть легко обнаружены с помощью простого детектора, который может быть обучен только на глубоких функциях, и при этом достигать точности около 98% AUC. В статье [6] представлена атака со стороны противника для прогнозирования злокачественности легочных узелков, а также стратегия защиты на основе ансамбля, чтобы уменьшить эффект от атаки противника. Также был выбран набор данных National Lung Screening Trial (NLST). Результаты экспериментов показали, что на исходных изображениях (без атаки противника) точность классификации трех CNN-моделей составила 75,1%, 75,5% и 76%. После атаки метода быстрого градиентного знака (FGSM) точность составляла 46,4%, 39,24% и 39,71%. При использовании ансамбля, основанного на мультиинициализации, и включении в обучающий набор неблагоприятных изображений точность классификации после FGSM и однопиксельной атаки составила 82,27% и 81,43% соответственно.

Таким образом, обзор современных исследований показал, что разработка интеллектуальных медицинских систем, а также их исследование на состязательные атаки на текущий момент является актуальной темой.

В данной работе рассмотрена задача мультиклассовой сегментации рака печени, а также построение и проведение состязательных атак на данную модель.

Постановка задачи и набор данных

Пусть X – множество изображений МРТ печени; Y

Пусть множество изображений $X = \{X_1, X_2,..., X_N\}$; множество меток $Y = \{Y_1, Y_2,..., Y_N\}$, где N – количество изображений в наборе данных. Каждое изображение X_1 представляет собой матрицу пикселей размера $W \times H$ с C каналами цвета, а каждая метка Y_1 – матрицу размера $W \times H$ с метками классов, соответствующими каждому пикселю изображения.

Дано: $\{x_1,..., x_N\}$ – обучающая выборка; $y_i = y(x_i)$, $i = \overline{1,N}$ – известные ответы, где – изображение 512 × 512 пикселей, заданное матрицей значений цветов.

Требуется построить алгоритм $a: X \to Y$, способный сегментировать произвольный объект $x \in X$.

Для практической реализации используемых алгоритмов и проведения экспериментов использовался набор данных[1] с сайта Kaggle.

Набор данных представляет собой 5702 изображения формата .jpg. размером 512 × 512. Все эти изображения имеют соответствующие файлы формата .png с размеченными элементами: задний фон, печень и раковые образования.

[1] Liver Tumor Segmentation. Available at: https://www.kaggle.com/datasets/andrewmvd/liver-tumor-segmentation/code (accessed 06.03.2023) (In Eng.).

Рисунок 1. Один из примеров из набора данных[1], где изображено оригинальное изображение, его размеченная маска и наложение изображения и сегментированной области
Источник: разработано автором П. Л. Ниряном

Архитектура модели сегментации и методы атак

Архитектура сегментации ResNet50. Рассмотрим архитектуру *ResNet50*, которая способна решать задачи сегментации. *ResNet50* – это сверточная нейронная сеть, которая используется для задач компьютерного зрения, таких как классификация изображений и семантическая сегментация [7]. Она состоит из 50 слоев, включая сверточные, пулинговые, активационные и полносвязные слои. Основная идея *ResNet50* заключается в использовании «residual connections» или «skip connections», которые позволяют обойти проблему затухания градиентов при обучении глубоких сетей. Эти соединения позволяют проходить градиенты от последующих слоев непосредственно к предыдущим слоям, минуя несколько слоев между ними. Это способствует более эффективному обучению более глубоких сетей и повышает точность классификации. На базе этой архитектуры в рамках данного исследования обучена модель сегментации рака печени.

Fast Gradient Sign Method (FGSM) является одним из простейших атакующих алгоритмов «белого ящика» для моделей машинного обучения, основанных на градиентном спуске [3]. Он работает путем вычисления градиента функции потерь по входным данным (изображениям) и добавляет к ним шум, который пропорционален знаку градиента.

FGSM можно описать как следующее математическое выражение:

$$x' = x + \varepsilon sign(\nabla_x J(\theta, x, y)) \quad (3)$$

где

x' – это возмущенный x, который генерируется путем добавления небольшой константы ε со знаком, равным направлению градиента потерь J (функция потерь) относительно x.

Input Perturbation. Атака Input Perturbation заключается в изменении входных данных для модели таким образом, чтобы модель давала неверный результат, но при этом данные оставались достаточно близкими к исходным [2]. В данном случае генерируется шум, который добавляется к исходному изображению.

Результаты и эксперименты

Построенная модель была обучена на 5 эпохах. Точность модели составляет 0.96 по метрике IoU. Результат работы модели можно увидеть на рисунке 2.

Теперь проведем атаки на полученную модель и посмотрим на полученные результаты.

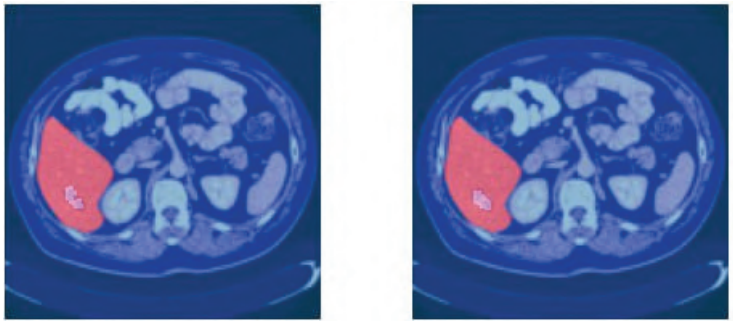
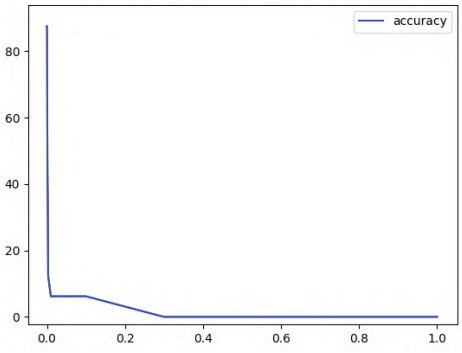


Рисунок 2. Пример работы модели. Рисунок слева – истинная маска, справа – результат работы модели
Источник: разработано автором П. Л. Ниряном

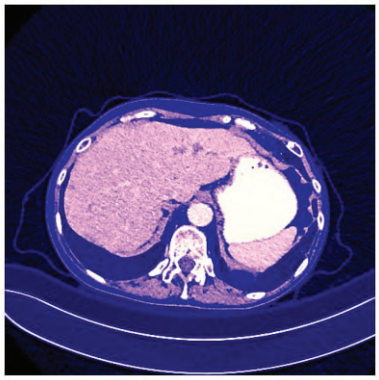
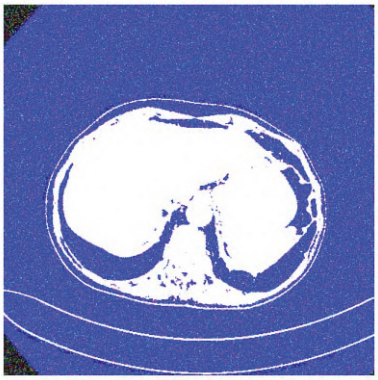
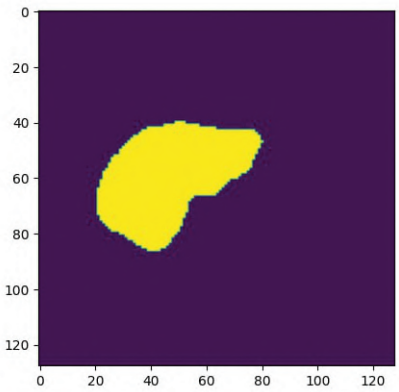
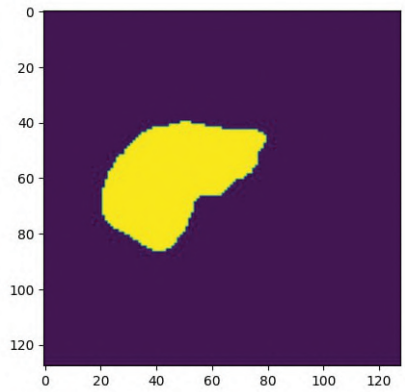
FGSM




```
robust accuracy for perturbations with
  Linf norm ≤ 0.0   : 87.5 %
  Linf norm ≤ 0.0002: 87.5 %
  Linf norm ≤ 0.0005: 75.0 %
  Linf norm ≤ 0.0008: 62.5 %
  Linf norm ≤ 0.001 : 56.2 %
  Linf norm ≤ 0.0015: 43.8 %
  Linf norm ≤ 0.002 : 37.5 %
  Linf norm ≤ 0.003 : 12.5 %
  Linf norm ≤ 0.01  :  6.2 %
  Linf norm ≤ 0.1   :  6.2 %
  Linf norm ≤ 0.3   :  0.0 %
  Linf norm ≤ 0.5   :  0.0 %
  Linf norm ≤ 1.0   :  0.0 %
```

Рисунок 3. Результат атаки методом FGSM
Источник: разработано автором П. Л. Ниряном

Таблица 1. Результат атаки *Input Perturbation*

Коэф. Ампл-ды	0	0.5
nosie_image		
pred_mask		
IoU	1	0.8639

Источник: разработано автором П. Л. Ниряном

Можно заметить, что при увеличении константы ε, точность модели снижается. При достижении всего $\varepsilon = 0.01$, модель теряет в точности 81,3%. При достижении $\varepsilon > 0.3$ составляет 0. Это связано с тем, что чем больше ε, тем заметнее шум. Очевидно, что создаваемые шумы были заметны человеческому глазу.

Input Perturbation. Рассмотрим таблицу 1, в которой изображены результаты атаки *Input Perturbation*.

Рассмотрим зашумленные изображения, которые были отправлены в модель. Изображения с коэффициентом амплитуды от 0 до 0.1 модель сегментирует почти также, как и оригинальное изображение. При увеличении коэффициента амплитуды, модель начинает давать неправильные предсказания, но и входные изображения имеют зашумленный вид, что может быть заметно пользователю.

Заключение

В результате проведенного исследования была разработана интеллектуальная система для задачи мультиклассовой сегментации. Кроме того, было доказано, что данная интеллектуальная система подвержена состязательным атакам. Также были проведены два типа атак на данную модель: FGSM и Input Perturbation. Наилучшие результаты показал метод FGSM, при небольшом возмущении ε модель теряет в точности больше 80%. В дальнейшем будет использован этот метод состязательной атаки для детекции.

Литература

1. Зельтер П. М., Колсанов А. В., Пышкина Ю. С. Сегментация очаговых образований печени и виртуальная резекция на основе данных компьютерной томографии // Бюллетень сибирской медицины. – 2021. – Т. 20, № 1. – С. 39–44.

2. Haohui (2019) Adversarial Attacks in Machine Learning and How to Defend Against Them. Medium, Dec 19 Available at: https://towardsdatascience.com/adversarial-attacks-in-machine-learning-and-how-to-defend-against-them-a2beed95f49c (accessed 03.04.2023) (In Eng.).

3. Tae J. (2021) Fast Gradient Sign Method. Jake Tae, Jan. 5 Available at: https://jaketae.github.io/study/fgsm/ (accessed 03.04.2023) (In Eng.).

4. Finlayson S.G. et al. (2019) Adversarial attacks on medical machine learning. Science. Vol. 363. No. 6433, pp. 1287–1289. – https://doi.org/10.1126/science.aaw4399 (In Eng.).

5. Ma X. et al. (2019) Understanding adversarial attacks on deep learning based medical image analysis systems. Pattern Recognition. Vol. 110, pp. 107332. – https://doi.org/10.1016/j.patcog.2020.107332 (In Eng.).

6. Paul R., Schabath M., Gillies R., Hall L., Goldgof D. (2020) Mitigating adversarial attacks on medical image understanding systems. 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). Iowa City, IA, USA, pp. 1517–1521. – https://doi.org/10.1109/ISBI45749.2020.9098740

7. ResNet and ResNetV2. Available at: https://keras.io/api/applications/resnet/ (accessed 01.04.2023) (In Eng.).

Статья поступила в редакцию: 16.05.2023; принята в печать: 20.11.2023.

Авторы прочитали и одобрили окончательный вариант рукописи.