ИССЛЕДОВАНИЕ ВЗАИМОСВЯЗЕЙ МЕЖДУ ПАРАМЕТРАМИ СОСТЯЗАТЕЛЬНЫХ АТАК И ОШИБОК КЛАССИФИКАЦИИ ИЗОБРАЖЕНИЙ НЕЙРОННЫМИ СЕТЯМИ

Д. М. Войнов

Белорусский государственный университет, Минск;
voynovdd@gmail.com;
науч. рук. – В. А. Ковалев, канд. техн. наук, доц.

Изучено явление состязательных атак, рассмотрены существующие алгоритмы генерации состязательных примеров. Постановлено и проведено экспериментальное исследование, выявляющее ключевые зависимости устойчивости глубоких нейронных сетей от главных параметров алгоритмов генерации. Обнаружены зависимости от исходной вероятности изображения и модальности медицинских изображений.

Ключевые слова: гистологические изображения; рентгеновские изображения; глубокие нейронные сети; состязательные атаки; состязательные примеры.

ВВЕДЕНИЕ

В настоящее время технология глубокого обучения показывает невероятные результаты в широком спектре задач касательно обработки, анализа и классификации изображений. Множество таких задач возникает в анализе медицинских изображений [1]: классификация изображений, как основа для систем автоматического компьютерного диагностирования; сегментация изображений, как вспомогательная задача для анализа изображений; генерация искусственных медицинских изображений, для замены натуральных в связи с недостатком количества таковых и другие.

Однако относительно недавно была обнаружена серьезнейшая уязвимость глубоких нейронных сетей – состязательные атаки. Научное сообщество сильно заинтересовано в этом явлении поскольку, во-первых, оно показало, что модели глубокого обучения абсолютно неустойчивы к специфическим изменениям входных данных [2] и, как следствие, небезопасны в применении в системах с повышенной ответственностью; во-вторых, открыло новый способ познания природы данных, так как некоторые свойства состязательных атак не зависят от архитектуры самой сети, а именно от набора данных, на котором сеть была обучена.

СОСТЯЗАТЕЛЬНЫЕ АТАКИ

Состязательной атакой называется некоторое действие, заставляющее классификатор совершать ошибки во время предсказания. Атаки произ-

водятся посредством, так называемых состязательных примеров – искусственно созданных изображений, незначительно отличающихся от оригинальных изображений, которые классификатор расценивает как совершенно другое. Зачастую алгоритмы генерации состязательных примеров настолько хороши, что различия между примером и исходным изображением незаметны для человеческого глаза (рис. 1).

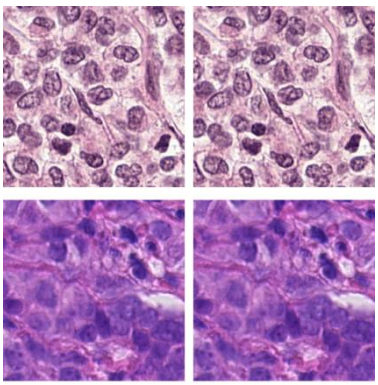

Рис. 1. Оригинальные изображения (слева)
вместе с состязательными примерами (справа)

АЛГОРИТМЫ ГЕНЕРАЦИИ СОСТЯЗАТЕЛЬНЫХ ПРИМЕРОВ

Несмотря на то, что состязательные атаки были впервые обнаружены не так давно, на сегодняшний день существует уже достаточно большое количество разнообразных методов их генерации [3]. Среди них выделяются как основные концепции, так и множество так называемых эвристик. Было найдено, что основная последовательность работы существующих алгоритмов генерации состязательных примеров представлена следующими шагами:

1. Возьмем некоторое изображение из предметной области атакуемой нейронной сети.
2. Сгенерируем специальное состязательное возмущение.
3. Прибавим сгенерированное возмущение к исходному изображению. Результат и будет состязательным примером.

МЕТОД СПРОЕЦИРОВАННОГО ГРАДИЕНТНОГО СПУСКА

Данный метод есть не что иное, как применение самой классической техники градиентного спуска с учетом а) итеративности и б) ограниченности применяемого возмущения.

Генерация состязательного примера этого метода зависит от исходного класса m, от параметра $\alpha > 0$, количества итераций n, максимальной допустимой амплитуды ϵ и определяется следующим образом:

$$x_{k+1} = Clip_{x,\varepsilon}(x_k - \alpha * \nabla y_m(x_k)),$$

где y_m – выход нейронной сети как функция от входного изображения, x_0 – исходное изображение, k меняется от 0 до n-1, $x^* = x_n$ – состязательный пример.

ИСПОЛЬЗУЕМЫЕ НАБОРЫ ДАННЫХ

В проведенном исследовании использовались 5 различных наборов медицинских изображений. Суммарно задействовалось 292000 гистологических изображений и 610080 рентгеновских изображений. Среди них были: набор гистологических изображений лимфоузлов, пораженных метастазами (H-MT); набор гистологических изображений яичников и щитовидной железы с опухолями (H-OV, H-TH, H-OV-TH); набор рентгеновских изображений грудной клетки мужчин и женщин в возрасте от 17 до 80 лет (X-NR2, X-NR3); набор рентгеновских изображений грудной клетки с выделенной аортой (X-AO); набор рентгеновских изображений грудной клетки с сегментированными легкими, который был аугментирован (X-TB).

На основе этих данных было составлено 8 задач классификации. Описание приведено в таблице.

Таблица

Сокращение	Кол-во изображений, тыс.	Кол-во изображений по классам, тыс.	Точность нейронной сети
H-MT	100	50/50	0.97
H-OV	96	48/48	0.92
H-TH	96	48/48	0.94
H-OV-TH	192	48/48/48/48	0.91
X-NR2	200	100/100	0.98
X-NR3	550	183.3/183.3/183.3	0.83
X-AO	27	16/11	0.78
X-TB	28	14/14	0.82

ПРОВЕДЕНИЕ ЭКСПЕРИМЕНТАЛЬНОГО ИССЛЕДОВАНИЯ

Исследование будем проводить следующим образом:

Пусть E = [0.02, … , 0.2] с шагом в 0.02 (итого 10 значений).

1. Для каждого изображения из тестовой выборки, каждого $\varepsilon \in E$ сгенерируем состязательный пример, проведя 20 итераций алгоритма спроецированного градиентного спуска

2. Сохраним вероятности, предсказанные атакуемой нейронной сетью в результате подачи в нее состязательных примеров

Далее проведем анализ полученных вероятностей. В качестве оценки успешности состязательных атак использовался так называемый рейтинг атаки. Вычисляется он как отношение количества состязательных примеров, которые сеть предсказывает с ошибкой к размеру исходной выборки.

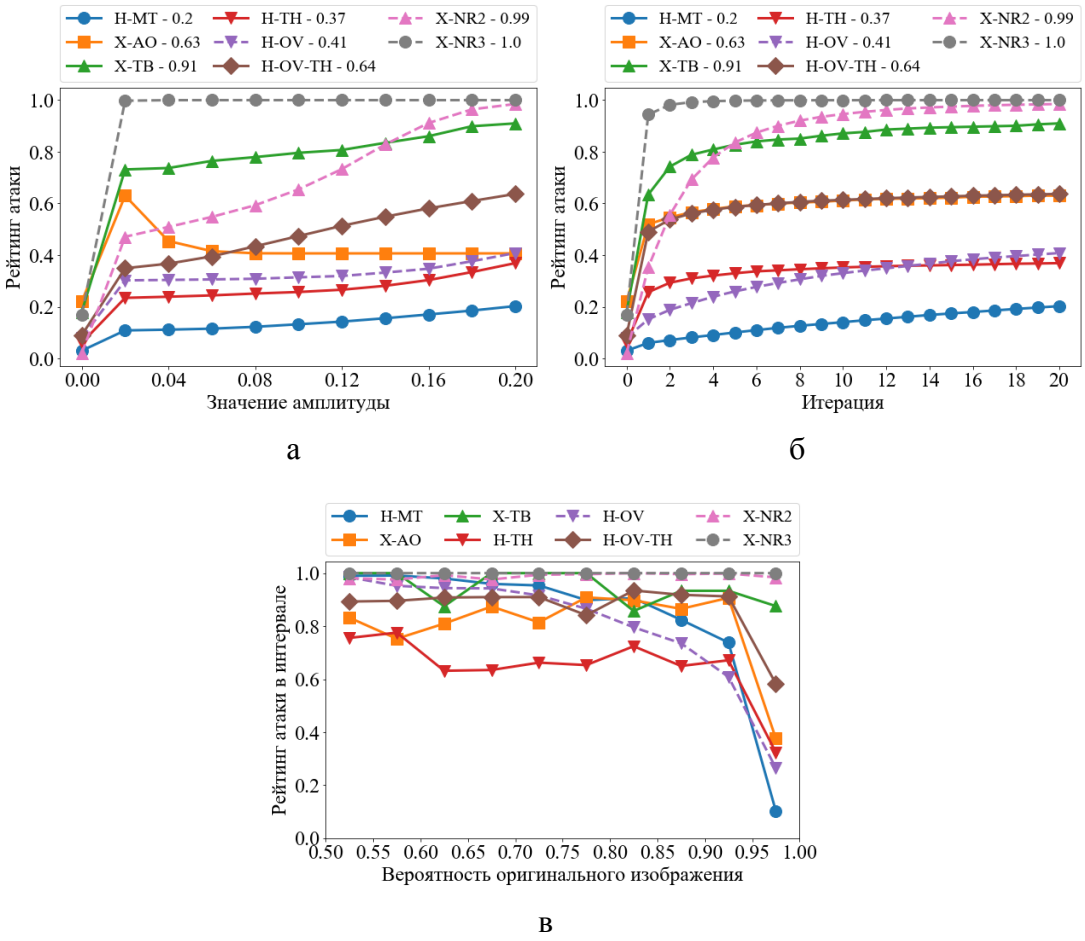

Рис. 2. Графики зависимостей (а) рейтинга атаки от допустимой амплитуды ε; (б) рейтинга атаки от номера итерации; (в) рейтинга атаки от исходной вероятности принадлежности изображения к своему классу

ЗАКЛЮЧЕНИЕ

В ходе проведения экспериментального исследования были обнаружены следующие закономерности:

- С увеличением амплитуды допустимого возмущения вероятность ошибки предсказания состязательного примера растет
- Итерации метода Спроецированного Градиентного Спуска постепенно увеличивают количество ошибок, допускаемых сетью, с асимптотической сходимостью к максимуму

- Изображения, которые классифицируются сетью с уверенностью более 95 %, менее склонны к образованию вредоносных состязательных примеров
- Нейронные сети, обученные для классификации гистологических изображений, оказались более устойчивы к состязательным атакам, нежели сети, обученные для классификации рентгеновских изображений различного рода

Библиографические ссылки

1. A survey on deep learning in medical image analysis / G. Litjens [et al.] // Medical Image Analysis. 2017. Vol. 42. P. 60–88.
2. *Akhtar N., Mian A. S.* Threat of Adversarial Attacks on Deep Learning in Computer Vision // IEEE Access. 2018. Vol. 6. P. 14410–14430.
3. *Xu W., Evans D., Qi Y.* Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks [Electronic resource]. URL: https://arxiv.org/pdf/1704.01155.pdf (дата обращения: 30.04.2019).