Метод обнаружения атак на нейросети детектирования лиц

В. Ф. Коновалов
Самарский национальный исследовательский университет им. академика С.П. Королева
Самара, Россия
vitfvk@gmail.com

Е. В. Мясников
Самарский национальный исследовательский университет им. академика С.П. Королева
Самара, Россия
mevg@geosamara.ru

***Аннотация*— В статье рассматривается задача обнаружения атак на нейросети детектирования лиц. Рассматриваются существующие атаки и методы решения указанной задачи, выбирается и реализуется базовый метод защиты для сети MTCNN. Предлагается метод, позволяющий улучшить качество детектирования по сравнению с базовым. Проводится экспериментальное сравнение методов.**

Ключевые слова*— *MTCNN, безопасность нейросетей, FGSM, градиентная атака, детектирование лиц

1. Введение

Современные нейросети детектирования объектов показывают хорошую точность на чистых данных, но являются крайне уязвимыми к атакам. Одной из задач, в которых работоспособность нейросети является критически важной, является детектирование лица. Так, без обнаруженного лица система не сможет провести его распознавание. В связи с этим возникает необходимость выявления возможных атак, создания более устойчивых к таким атакам систем.

На данный момент известны атаки на нейросети, которые позволяют незначительным изменением входных данных получать нужный злоумышленнику выход. Такими атаками являются градиентные атаки, обнаружение которых и будет рассмотрено в настоящей работе.

Работа имеет следующую структуру. В разделе 2 описываются некоторые из существующих цифровых атак на нейросети детектирования лиц. В разделе 3 рассматриваются существующие методы защиты от атак, выбирается базовый метод детектирования. В разделе 4 предлагается предложенный метод защиты, улучшающий базовый. Раздел 5 посвящен сравнительному исследованию базового и предложенного методов.

2. Рассматриваемые атаки на нейросети обнаружения лиц

Цифровые атаки на нейросети заключаются в добавлении к исходному изображению некоторого небольшого возмущения с целью добиться некорректного детектирования/распознавания объекта.

Наиболее часто описываемой атакой является метод быстрого градиента FGSM (Fast Gradient Sign Method) [1]. Суть данной атаки состоит в оптимизации некоторой функции потерь относительно целевого или истинного класса. Это можно делать неизбирательно, «удаляя» изображение от истинного класса в направлении любого другого класса, или уменьшая функцию потерь в отношении определенного целевого класса.

Модификациями метода FGSM [1] являются I-FGSM [2] и MI-FGSM (Moment-Iterative Fast Gradient Sign Method) [3]. Последний метод, согласно описанию, позволяет обходить локальные экстремумы градиента.

3. Существующие способы защиты от атак

Обзор методов атак и защит от них приведен в работе [4]. Ниже приведены некоторые из используемых методов защиты.

Одним из первых способов защиты стало безопасное обучение – “подмешивание” атакованных изображений к обычному тренировочному набору входных данных [5], либо обучение с функцией потерь, учитывающей атаку [1]. К сожалению, при использовании подобных схем атакованные изображения не классифицируются от обычных, и попытки атак остаются незамеченными.

Случайные аугментации изображения (фильтрация, обрезка, изменение размера) могут помочь улучшить качество детектирования лиц [6]. Достоинство данного метода в простоте и отсутствии необходимости обучения, недостатки — аугментация так же ухудшает качество детектирования, искажает изображение. При этом атаки так же остаются необнаруженными.

Еще один способ защиты – добавление классификатора на выходе или входе нейросети, позволяющего отделять атакованные изображения от чистых. Простой и быстрый метод, основанный на таком подходе [7], показал хорошую эффективность обнаружения градиентных атак на наборе данных MNIST. Указанный подход используется в качестве базового подхода в настоящей работе.

4. Предлагаемый метод

В ходе работы было введено понятие приближенного шума — абсолютной разницы между сглаженным оконным фильтром и оригинальным, потенциально атакованным изображением. Было показано, что распределение значений приближенного шума для атакованных и чистых изображений отличается.

По результатам исследования был предложен метод детектирования, основанный на базовом алгоритме и состоящий в следующем. Входное изображение приводится к изображению x в цветовом пространстве YcbCr. Далее с использованием свертки вычисляется приближенное значение шума y:

$$y = |x - \hat{x}|, \text{ где } \hat{x} = \frac{1}{8}\begin{bmatrix}1&1&1\\1&0&1\\1&1&1\end{bmatrix} * x \quad (1)$$

Полученный шум стандартизуется межквантильным алгоритмом, что позволяет уравнять большие и малые внесенные искажения, оставив значимой лишь форму

шума. По цветовым каналам CbCr полученного стандартизованного шума строится гистограмма частот, которая затем преобразуется в плотность распределения делением на сумму частот. Преобразование в плотность распределения позволяет уравнять изображения малого и большого разрешения, а так же приводит все признаки к диапазону [0,1], что улучшает качество последующей классификации. Внешний вид признаков изображен на Рис. 1. Полученные значения плотности являются признаками классификаторов, в качестве которых в настоящей работе исследовались: машина опорных векторов (SVM), случайный лес (Random Forest) и многослойный перцептрон (MLP).

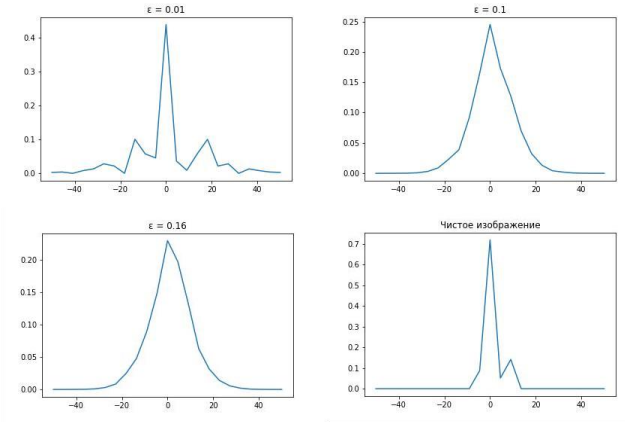

Рис. 1. Распределение приближенного шума для чистого и атакованных изображений с разным уровнем возмущения ε.

5. Экспериментальная часть

Для каждого из методов детектирования (базового и предложенного) было проведено исследование эффективности. Для исследования использовалось несколько наборов данных:

- images-small: 1060 чистых изображений из набора данных WIDER FACE val;
- images-train: 1238 чистых изображений из набора WIDER FACE train;
- attacked_FGSM: 21200 изображений, атакованных FGSM с разными параметрами;
- attacked_train_FGSM: 24760 изображений, атакованных FGSM с разными параметрами.

Наборы attacked_FGSM, attacked_train_FGSM были получены из наборов images_small, images_small_train соответственно. Для получения использовались различные уровни возмущения ε от 0,01 до 0,20 с шагом 0,01.

Эксперимент проводился для изображений как сохраненных без сжатия, так и сжатых алгоритмом jpeg после атаки.

Для оценки методов детектирования использовались оценки качества macro precision, macro recall и macro f1-score. Результаты исследования для базового метода приведены в таблице I, для предложенного метода – в таблицах II и III для нескольких различных алгоритмов машинного обучения.

Таблица I. Результаты для базового алгоритма

	F1-Score	Precision	Recall
Несжатые файлы	0,7544	0,7458	0,8250
Сжатые файлы	0,6902	0,7237	0,7839

Из таблицы I можно заключить, что базовый метод детектирования с меньшим успехом распознает изображения, сохраненные со сжатием.

Таблица II. Результаты для несжатых изображений

Алгоритм	Среднее на кроссвалидации			Результат на тестовой выборке		
	F1-Score	Precision	Recall	F1-Score	Precision	Recall
SVM	0,9984	0,9980	0,9988	0,9996	0,9992	0,9999
Random Forest	0,9990	0,9989	0,9991	0,9996	0,9994	0,9999
MLP	0,9981	0,9978	0,9985	0,9970	0,9948	0,9992

Таблица III. Результаты для сжатых изображений

Алгоритм	Среднее на кроссвалидации			Результат на тестовой выборке		
	F1-Score	Precision	Recall	F1-Score	Precision	Recall
SVM	0,9761	0,9704	0,9830	0,9570	0,9318	0,9872
Random Forest	0,9924	0,9905	0,9925	0,9797	0,9685	0,9918
MLP	0,9761	0,9707	0,9824	0,9789	0,9676	0,9911

Из приведенных результатов можно сделать вывод о том, что предложенный метод значительно улучшает эффективность обнаружения по сравнению с базовым. При этом предложенный метод также менее подвержен снижению качества детектирования на сжатых изображениях.

Литература

[1] Ian J. Goodfellow. Explaining and Harnessing Adversarial Examples / Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy – 2014. – DOI:10.48550/arXiv.1412.6572

[2] Alexey Kurakin. Adversarial Examples in the Physical World / Alexey Kurakin, Ian Goodfellow, Samy Bengio. – 2016. – DOI:10.48550/arXiv.1607.02533

[3] Yinpeng Dong. Boosting Adversarial Attacks with Momentum / Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, Jianguo Li. – 1998. – DOI:10.48550/arXiv.1710.06081

[4] Chakraborty, A. Adversarial Attacks and Defences: A Survey / A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, D. Mukhopadhyay // 2018 IEEE/CVF. – 2018. – P. 9185-9185.

[5] Alexey Kurakin. Adversarial Machine Learning at Scale / Alexey Kurakin, Ian Goodfellow, Samy Bengio. – 2018. – DOI:10.48550/arXiv.1810.00069

[6] Kurakin, A. Adversarial Attacks and Defences Competition / A. Kurakin, I. Goodfellow, S. Bengio, Y. Dong, F. Liao, et al. // The NIPS '17 Competition: Building Intelligent Systems. – 2018. – P. 195-231 DOI:10.48550/arXiv.1804.0009

[7] Schöttle, P. Detecting Adversarial Examples - A Lesson from Multimedia Forensics / P. Schöttle, A. Schlögl, C. Pasquini, R. Böhme. – 2018. – P. 947-951. doi: 10.23919/ EUSIPCO.2018.8553164