УДК 004.89

ГЕНЕРАЦИЯ СОСТЯЗАТЕЛЬНЫХ ПРИМЕРОВ В ОГРАНИЧЕННОЙ ОБЛАСТИ ИЗОБРАЖЕНИЯ НА ОСНОВЕ ПРЕДВАРИТЕЛЬНО ОБУЧЕННЫХ МОДЕЛЕЙ

ADVERSARIAL EXAMPLE GENERATION IN A LIMITED IMAGE REGION BASED ON PRE-TRAINED MODELS

Ч. Чжан, В.И. Терехов, А.И. Канев
Москва, МГТУ им. Н.Э. Баумана
C. Zhang, V.I. Terekhov, A.I. Kanev
Moscow, BMSTU

Аннотация. *В данной работе исследуется генерация состязательных примеров (adversarial example) в ограниченной области изображения на основе предварительно обученных моделей. На основе проведенных экспериментов показано, что можно генерировать ненаправленные и направленные состязательные примеры с высокой степенью успешности для нейронных сетей, основанных на определенной предварительно обученной модели, при условии, что установлен подходящий порог для функции потерь. Показано, что метод быстрого градиента имеет более высокий процент успеха, чем методы, основанные на оптимизации. В заключении делается вывод о том, что предложенный метод генерации состязательных примеров может существенно помешать работе нейронных сетей, основанных на предварительно обученных моделях.*

Ключевые слова: *состязательный пример, глубокое обучение, распознавание лиц, сверточная нейронная сеть.*

Abstract. *In this work we investigate an adversarial example generation in a limited image region based on pre-trained models. Experiments proved that it is possible to generate non-targeted and targeted adversarial examples with high success rates for neural networks based on a certain pre-trained model, when a proper threshold for the loss function is set. It is shown that the fast gradient method has a higher success rate than method based on optimization. It is concluded that our adversarial example generation method can significantly interfere with neural networks based on pre-trained models.*

Keywords: *adversarial example, deep learning, face recognition, convolutional neural network.*

Введение

Сверточные нейронные сети (CNN) быстро заменили предыдущие сложные методы распознавания, основанные на инженерии признаков, в качестве основного метода распознавания лиц. Некоторые тщательно построенные данные могут оказать значительное влияние на точность модели, следовательно, и на вопросы безопасности. Повышение устойчивости глубоких нейронных сетей к состязательным примерам [1] стало одним из основных направлений исследований. Для того чтобы изучить производительность моделей глубокого обучения при различных типах сложных входных данных, улучшить модели и повысить их устойчивость, сначала необходимо сгенерировать высококачественные состязательные примеры для имитации атак на модель.

В настоящее время широко используются предварительно обученные модели. Поскольку многие системы распознавания лиц используют одну и ту же предварительно обученную модель, состязательные примеры, созданные против предварительно обученной модели, скорее всего, окажут влияние на эти системы

распознавания лиц. Поэтому авторы предлагают метод генерации состязательных примеров на основе предварительно обученных нейронных сетей и существующих методах генерации «adversarial example». В работе показано, что состязательные примеры, сгенерированные против широко используемой предварительно обученной модели, могут заставить нейронную сеть распознавания лиц, основанную на этой модели, выдать неправильные результаты классификации.

Смежные исследования

Состязательный пример (Adversarial example) – это данные, к которым были добавлены тщательно созданные неслучайные возмущения, которые с большей вероятностью приведут к неправильным выводам модели глубокого обучения. Быстрый градиентный алгоритм (Fast Gradient) и быстрый градиентный символьный алгоритм (Fast Gradient Sign) – это два алгоритма, предложенные Ian J. [2]. В свою очередь Goodfellow предложил алгоритмы генерации враждебных примеров, которые могут быть использованы для направленных и ненаправленных атак. Улучшенными методами, основанными на методе быстрого градиента, являются I-FGSM и MI-IFGSM [3]. Другие методы включают DeepFool [4] и JSMA [5]. Специально созданные изображения, снятые камерой мобильного телефона, были поданы в нейронную сеть классификации изображений Inception v3, значительная часть которых была классифицирована неверно [6]. Наложение специально разработанного изображения на лоб также может привести к тому, что нейронная сеть будет выдавать неправильные результаты классификации [7]. Эти два примера показывают, что состязательные примеры могут существовать в реальной среде. Кроме того, переносимость устойчивых примеров является важной областью исследований [8]. Состязательный пример с переносимостью может оказывать влияние на несколько моделей. В целом, создание изображения состязательного примера с переносимостью и маскировкой части лица является возможной атакой на модели распознавания лиц.

Методы генерации состязательных примеров

Чтобы сделать состязательные примеры более обобщаемыми, авторы предлагают метод генерации состязательных примеров для предварительно обученных моделей. Суть предложения состоит в том, чтобы генерировать состязательные примеры, направленные против правильного распознавания лиц нейронной сетью, основанной на предварительно обученной модели (рис. 1).

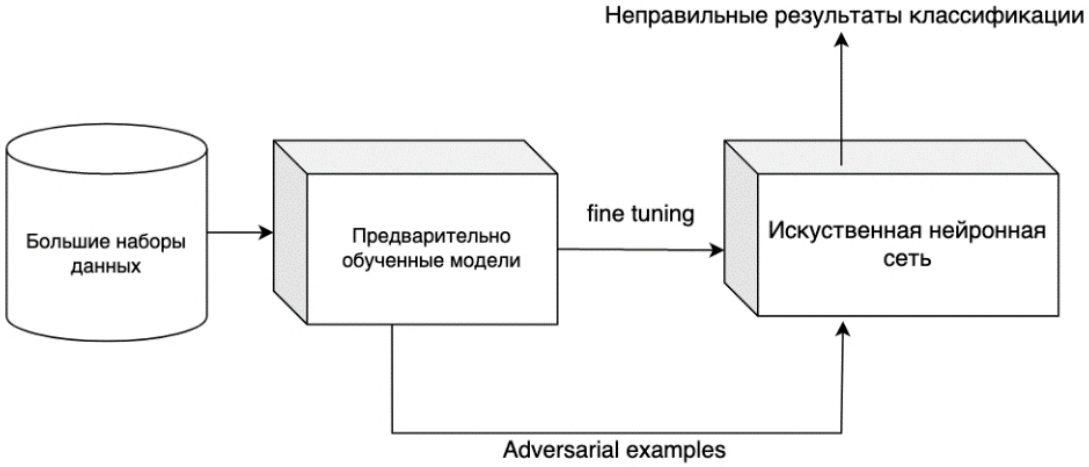

Рис. 1. Методы генерации состязательных примеров
на основе предварительно обученных моделей

Предложенный подход способен влиять на результаты классификации искусственных нейронных сетей на изображениях с ограниченным числом модификаций пикселей, поэтому в работе исследовано влияние различных принципов генерации на способность состязательных примеров нарушать правильное распознавание лиц нейронной сетью.

Состязательные примеры также можно разделить на 2 категории в зависимости от эффекта атаки.

- Направленная атака. Этот тип атаки устанавливает цель заранее, то есть результаты, генерируемые вводящей в заблуждение моделью, предопределены до того, как состязательный пример вводится в модель.
- Ненаправленная атака. Этот тип атаки не устанавливает цель перед атакой, а лишь требует, чтобы вводимый в модель состязательный пример получил результат, отличный от исходной информации.

В нашей работе используются два метода генерации состязательных примеров.

1. Метод на основе оптимизации

$$\underset{x^*}{argmin}\, l\left(f\left(x\right), f\left(x^*\right)\right)$$

где x^* – состязательный пример; x – целевое изображение; f – предварительно обученная модель, которая не содержит полностью связанного слоя; l – функция потерь.

В качестве функции потерь l выбрана среднеквадратичная ошибка (MSE), а оптимизатором является Adam.

2. Метод быстрого градиента (Fast Gradient Method) позволяет состязательный пример двигаться по градиенту. В итеративном методе быстрого градиента значение состязательного примера итеративно изменяется в соответствии с направлением градиента до тех пор, пока не будет выполнено условие остановки.

$$clip(x^* - \eta \frac{\nabla_{x^*} l\left(f(x), f(x^*)\right)}{||\nabla_{x^*} l\left(f(x), f(x^*)\right)||}) \rightarrow$$

где x^* – состязательный пример; x – целевое изображение; f – предварительно обученная модель, которая не содержит полностью связанного слоя; l – функция потерь; η – длина шага; $clip(x)$ – функция, используемая для приведения каждого измерения x к диапазону значений пикселей, т.е. [0, 255] в данной работе.

Эффективность метода, основанного на предварительно обученной модели

Мы использовали MobileFaceNet [9], и нашли предварительно обученную модель с точностью 99,5% на LFW в Интернете [10]. Модели предварительного обучения были построены на основе набора данных CASIA WebFace [11].

Алгоритмы были протестированы на наборе данных LFW - наборе данных лиц LFW (Labled Faces in the Wild) [12], который сегодня является распространенным тестовым набором для распознавания лиц и в котором представлены изображения лиц из естественных сцен в жизни. Были отобраны 300 пар изображений, принадлежащих различным лицам, которые могут быть правильно классифицированы выбранной нами нейронной сетью. Для ненаправленных атак первое изображение каждой пары изображений выбирается в качестве исходных данных. Для направленных атак первое изображение каждой пары изображений выбирается в качестве исходного изображения, а второе изображение - в качестве цели.

После определения ограниченной области оптимизации на изображении, два различных изображения лица одновременно подаются в предварительно обученную

модель без полностью подключенного слоя. Одно изображение 1 (рис. 2а) используется в качестве цели для классификации, а другое изображение 2 (рис. 2б) используется для создания состязательного примера 3 (рис. 2г). Разница между признаками изображения 1 и изображения 3 вычисляется с помощью функции потерь. Информация о градиенте функции потерь на состязательный пример получается методом обратного распространения.

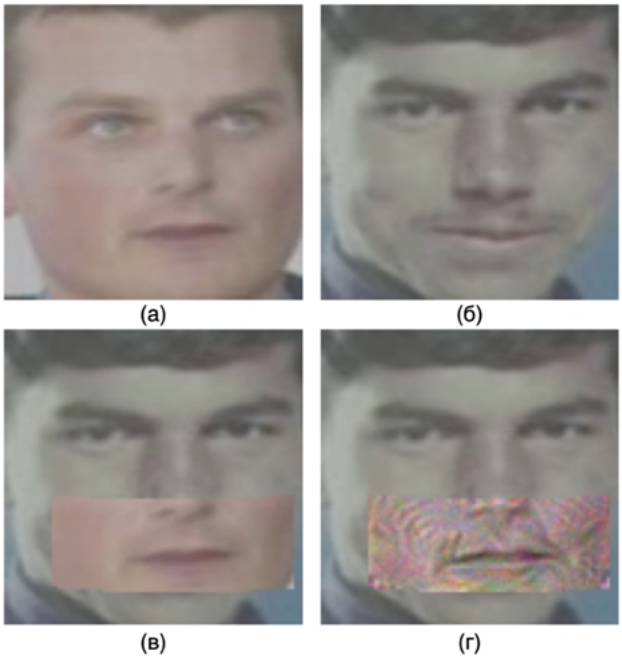

Рис. 2. Изображения, которые были использованы для создания состязательного примера:
а) целевое изображение, б) оригинальное изображение,
в) изображение после частичной замены, г) обработанное изображение частичной замены
(состязательный пример)

Оптимизатор Adam модифицирует состязательный пример, а затем восстанавливает модификации за пределами ограниченной области состязательного примера. Значения, выходящие за пределы диапазона представления пикселей, усекаются. Вышеописанные шаги повторяются до тех пор, пока значение функции потерь не станет меньше установленного нами порога. Результаты классификации полученного состязательного примера (изображение 3) и исходного изображения 2 в нейронной сети (содержащей созданный нами полносвязный слой), построенной на основе предварительно обученной модели, считаются успешными для ненаправленных атак, если они не совпадают. Полученный состязательный пример (изображение 3) считается успешным для направленных атак, если он согласуется с результатом классификации изображения 1 в нейронной сети, построенной на основе модели предварительного обучения. Метод быстрого градиента не требует оптимизатора, а просто добавляет возмущение в направлении градиента к состязательным примерам.

На основании экспериментальных результатов, приведенных в таблице, мы пришли к выводу, что можно генерировать состязательные примеры с высокой степенью успешности для нейронных сетей, основанных на определенной предварительно обученной модели, при условии, что установлен подходящий порог для функции потерь. При этом метод быстрого градиента имеет более высокий процент успеха, чем методы, основанные на оптимизации.

Уровень успешности различных методов на нейронных сетях с MobileFaceNet

Методы		Уровень успешности		
Метод замещения пикселей	Направленная	1,7%		
	Ненаправленная	17,3%		
Порог функции потерь		0,01	0,005	0,001
Метод на основе оптимизации	Направленная	48,0%	50,2%	49,4%
	Ненаправленная	97,2%	98,0%	97,8%
Метод быстрого градиента	Направленная	92,4%	90,8%	92,0%
	Ненаправленная	100%	99,8%	100%

Заключение

В данной работе было проведено исследование ненаправленных и направленных состязательных примеров, созданных с помощью различных методов, на предварительно обученной модели. Полученные результаты подтверждают, что на основе предварительно обученной модели можно построить эффективные состязательные примеры, даже если получены не все параметры нейронной сети. С другой стороны, было обнаружено, что состязательные примеры, созданные нашим методом, более эффективно вмешиваются в работу нейронной сети, чем замещение пикселей. В представленной работе были разработан метод на основе предварительно обученной модели и показано, что они способны генерировать ненаправленные и направленные состязательные примеры для нейронных сетей на основе этой предварительно обученной модели с высоким уровнем успеха. Состязательные примеры, сгенерированные в ограниченных областях, с большей вероятностью могут быть использованы в атаках на системы распознавания лиц в реальных условиях, и соответственно могут быть проведены исследования по построению нейронных сетей распознавания лиц с использованием информации о их глубине или другой информации.

Исследование выполнено за счет гранта Российского научного фонда (проект № XX-XX-XXXXX)

Список литературы

1. Akhtar N., Mian A. Threat of adversarial attacks on deep learning in computer vision: A survey // Ieee Access. 2018. № 6. DOI 1412.6572.
2. Goodfellow I., Shlens J., Szegedy C. Explaining and Harnessing Adversarial Examples // International Conference on Learning Representations. 2015. № 43405.
3. Dong Y., Liao F., Pang T. Boosting adversarial attacks with momentum // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018. P. 9185–9193. DOI 10.1109/CVPR.2018.00957.
4. Moosavi-Dezfooli S.M., Fawzi A., Frossard P. Deepfool: A Simple and Accurate Method to Fool Deep Neural Networks // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016. P. 2574–2582. DOI 10.1109/CVPR.2016.282.
5. Papernot N., McDaniel P., Jha S. The Limitations of Deep Learning in Adversarial Settings // 2016 IEEE European Symposium on Security and Privacy (EuroS&P). 2016. P. 372–387. DOI 10.1109/EuroSP.2016.36/
6. Narodytska N., Kasiviswanathan S.P. Simple Black-Box Adversarial Perturbations for Deep Networks // ArXiv preprint. 2016. DOI 10.48550/ARXIV.1612.06299.
7. Komkov S., Petiushko A. Advhat: Real-world adversarial attack on ArcFace Face ID system // 25th International Conference on Pattern Recognition (ICPR). 2021. P. 819–826. DOI 10.1109/ICPR48806.2021.9412236.
8. Liu Y., Chen X., Liu C., Song, D. Delving into Transferable Adversarial Examples and Black-box Attacks // ArXiv preprint. 2016. DOI 10.48550/ARXIV.1611.02770.

9. Chen S., Liu Y., Gao X., Han Z. MobileFaceNets: Efficient CNNs for Accurate Real-Time Face Verification on Mobile Devices // Chinese Conference on Biometric Recognition. 2018. P. 428–438. DOI 10.1007/978-3-319-97909-0_46.
10. MobileFaceNet // GITHUB.COM. 21 июня 2018. URL: https://github.com/zhanglaplace/MobileFaceNet (дата обращения: 20.03.2022).
11. Yi D., Lei Z., Liao S., Li S.Z. Learning Face Representation from Scratch // ArXiv preprint. 2014. DOI 10.48550/ARXIV.1411.7923.
12. Labeled Faces in the Wild // CS.UMASS.EDU. 9 января 2018. URL: http://vis-www.cs.umass.edu/lfw/ (дата обращения: 20.03.2022).