DOI: 10.15514/ISPRAS-2022-34(6)-7

Влияние трансформаций на успешность состязательных атак для классификаторов изображений Clipped BagNet и ResNet

[1] *Е.О. Курденкова, ORCID: 0000-0001-5871-8179 <kurdenkova@ispras.ru>*
[2] *М.С. Черепнина, ORCID: 0000-0003-1186-6718 <i.knaz@yandex.ru>*
[1,3] *А.С. Чистякова, ORCID: 0000-0003-4896-4418 <a.chistyakova@ispras.ru>*
[1] *К.В. Архипенко, ORCID: 0000-0002-8699-889X <arkhipenko@ispras.ru>*

[1] *Институт системного программирования им. В.П. Иванникова РАН, 109004, Россия, г. Москва, ул. А. Солженицына, д. 25*
[2] *Мюнхенский технический университет, Германия, 80333 Мюнхен, Арсиситрассе 21*
[3] *Московский государственный университет имени М.В. Ломоносова, 119991, Россия, Москва, Ленинские горы, д. 1*

Аннотация. В нашей статье сравнивается точность классической модели ResNet-18 с точностью моделей Clipped BagNet-33 и BagNet-33 с состязательным обучением в разных условиях. Мы провели эксперименты для изображений, атакованных состязательной наклейкой, в условиях трансформаций изображений. Состязательная наклейка представляет из себя небольшую область атакуемого изображения, внутри которой значения пикселей можно неограниченно менять, что может порождать ошибки в предсказании модели. Трансформации атакованных изображений в данной статье моделируют искажения, появляющиеся в физическом мире, когда смена ракурса, масштаба или освещения изменяет распознаваемое изображение. Наши эксперименты показывают, что модели из семейства BagNet плохо справляются с изображениями в низком качестве. Также мы проанализировали влияние разных видов трансформаций на устойчивость моделей к состязательным атакам и переносимость этих атак.

Ключевые слова: состязательная атака; состязательная наклейка; архитектура BagNet; состязательное обучение; проектируемый градиентный спуск

Для цитирования: Курденкова Е.О., Черепнина М.С., Чистякова А.С., Архипенко К.В. Влияние трансформаций на успешность состязательных атак для классификаторов изображений Clipped BagNet и ResNet. Труды ИСП РАН, том 34, вып. 6, 2022 г., стр. 101-116. DOI: 10.15514/ISPRAS-2022-34(6)-7

Effect of transformations on the success of adversarial attacks for Clipped BagNet and ResNet image classifiers

[1] *E.O. Kurdenkova, ORCID: 0000-0001-5871-8179 <kurdenkova@ispras.ru>*
[2] *M.S. Cherepnina, ORCID: 0000-0003-1186-6718 <m.cherepnina@ispras.ru>*
[1,3] *A.S. Chistyakova, ORCID: 0000-0003-4896-4418 <a.chistyakova@ispras.ru>*
[1] *K.V. Arkhipenko, ORCID: 0000-0002-8699-889X <arkhipenko@ispras.ru>*

[1] *Ivannikov Institute for System Programming of the Russian Academy of Sciences, 25, Alexander Solzhenitsyn st., Moscow, 109004, Russia*
[2] *Technical University of Munich Arcisstraße 21, 80333 München, Germany*
[3] *Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow, 119991, Russia*

Abstract. Our paper compares the accuracy of the vanilla ResNet-18 model with the accuracy of the Clipped BagNet-33 and BagNet-33 models with adversarial learning under different conditions. We performed experiments on images attacked by the adversarial sticker under conditions of image transformations. The adversarial sticker is a small region of the attacked image, inside which the pixel values can be changed indefinitely, and this can generate errors in the model prediction. The transformations of the attacked images in this paper simulate the distortions that appear in the physical world when a change in perspective, scale or lighting changes the image. Our experiments show that models from the BagNet family perform poorly on images in low quality. We also analyzed the effects of different types of transformations on the models' robustness to adversarial attacks and the tolerance of these attacks.

Keywords: adversarial attack; adversarial patch; BagNet architecture; adversarial training; projected gradient descent

For citation: Kurdenkova E.O., Cherepnina M.S., Chistyakova A.S., Arkhipenko K.V. Effect of transformations on the success of adversarial attacks for Clipped BagNet and ResNet image classifiers. Trudy ISP RAN/Proc. ISP RAS, vol. 34, issue 6, 2022. pp. 101-116 (in Russian). DOI: 10.15514/ISPRAS-2022-34(6)-7

1. Введение

Предыдущие работы [1, 2] показали, что несмотря на хорошую способность современных нейронных моделей решать задачи компьютерного зрения, эти модели остаются уязвимыми для атак злоумышленников. В данной статье рассматриваются атаки состязательной наклейкой, которая представляет из себя ограниченную область изображения, к которой может быть применено неограниченное возмущение пикселей. Пример атакованных изображений приведен на рис. 1.

Атака состязательной наклейкой находит свое применение в физическом мире. Злоумышленник может распечатать такую наклейку и поместить её в область видимости классификатора (например, во время детекции) и тем самым атаковать систему. Существуют атаки, которые могут генерировать универсальные состязательные наклейки, которые не зависят от сцены на изображении и даже от распознаваемого предмета [3]. Самым популярным примером применения состязательной наклейки является нанесения таковой на дорожный знак «стоп», после чего такой знак распознается классификатором как знак «ограничение скорости». Такая атака может привести к серьезным последствиям в эпоху тотальной автоматизации процессов. Таким образом, защита от подобного вида атак состязательной наклейкой является крайне важной задачей.

Существуют архитектуры моделей, которые лучше других справляются с атаками наклейкой. Например, в статье [4] авторами была предложена модель BagNet, которая более устойчива к атакам наклейкой в сравнении с другими моделями. Затем в статье [5] была предложена модификация этой модели – Clipped BagNet (CBN), которая улучшила устойчивость модели.

Авторы статьи [6] пошли еще дальше и предложили универсальные маски для моделей с небольшими размерами рецептивных полей для повышения устойчивости моделей к атакам наклейкой. Также в статье [6] предложен метод состязательного обучения модели Clipped BagNet с использованием одной из масок.

Рис. 1. Изображения, атакованные состязательной наклейкой PGD на примере модели CBN. В первом ряду изображения до применения трансформаций, второй ряд – гауссовский шум, третий ряд – затемнение, четвертый – поворот с масштабированием, пятый – комбинация всех трансформаций. Над изображениями указаны истинные метки и предсказанные на соответствующем этапе. Число итераций: 40, размер стороны наклейки 40 пикселей, что эквивалентно 3,2% от площади изображения

Fig. 1. Images attacked with a PGD adversarial patch using the CBN model as an example. In the first row of the image before applying the transformations, the second row is Gaussian noise, the third row is darkening, the fourth is rotation with scaling, the fifth is a combination of all transformations. Above the images, the true labels and those predicted at the corresponding stage are indicated. Number of iterations: 40, patch side size 40 pixels, equivalent to 3.2% of the image area

В перечисленных статьях атаки состязательной наклейкой рассматривались статичными, то есть без моделирования эффектов из физического мира. В физическом мире атаки наклейкой могут рассматриваться как непосредственное нанесение наклейки на предмет (например, на дорожный знак), уже после этого получается изображение, которое необходимо классифицировать. В таком случае изображение и наклейка вместе будут подвержены одинаковым изменениям, связанным с ракурсом, выбранным для фотографии, изменениям

освещения, а также шумам, возникающим в связи с техническими особенностями фотоаппарата. Чтобы получить более объективную оценку эффективности моделей при атаке, необходимо учитывать эти изменения изображения, поэтому в данном исследовании рассматриваются трансформации над состязательными изображениями.

Модели на основе архитектуры BagNet обладают маленьким размером рецептивного поля, что в совокупности с эффектами физического мира может привести к низкой точности предсказаний. Поэтому, одна из основных задач данной статьи – это исследовать влияние таких эффектов на устойчивость модели к атакам.

В рамках нашей статьи мы:

- провели и систематизировали более 200 вычислительно трудоемких экспериментов по состязательным атакам наклейкой;
- показали, что модели из семейства BagNet плохо справляются с изображениями в низком качестве;
- проанализировали, насколько сильно четыре разных типа трансформаций физического мира влияют на устойчивость моделей к состязательным атакам и переносимость этих атак.

2. Типы состязательных атак

2.1 Атака состязательной наклейкой

Атака состязательной наклейкой в данном исследовании основана на атаке проектируемого градиентного спуска (PGD – Projected Gradient Descent), которая является атакой белого ящика, то есть предполагается, что злоумышленник имеет доступ к модели, ее весам и градиентам [7]. Успеха в такой атаке часто можно добиться с помощью невидимых человеку изменений, что является несомненным плюсом для злоумышленника.

Идея PGD атаки заключается в решении задачи оптимизации с ограничениями. Атака пытается найти такие возмущения изображения, которые бы максимизировали потери модели на выходе (в случае целевой атаки – на конкретном выходе), при этом искомое возмущение обычно ограничено некоторой величиной ϵ. Ограничение задается с помощью нормы ℓp, которая определяется формулой (1). В данном исследовании рассматривается $\ell 0$-норма, которая измеряет, какое количество пикселей были изменены:

$$\left|\left|x\right|\right|_p = \left(\sum_i |x_i|^p\right)^{1/p}. \quad (1)$$

Если ставится задача изменить предсказание модели на любое неверное, то атака называется нецелевой. Формальная запись преобразований для нецелевой PGD атаки представлена формулой (2):

$$x'_{i+1} = \prod_{x+\epsilon}\left[x'_i + \alpha \cdot \operatorname{sign}\left(\nabla_x \mathcal{L}(x, y, \theta)\right)\right]. \quad (2)$$

Если же цель – обмануть модель, получив от модели метку заранее определенного класса, то атака называется целевой. Преобразования для целевой PGD атаки описываются формулой 3:

$$x'_{i+1} = \prod_{x+\epsilon}\left[x'_i - \alpha \cdot \operatorname{sign}\left(\nabla_x \mathcal{L}(x, \hat{y}, \theta)\right)\right]. \quad (3)$$

В этих формулах θ – параметры модели, x – изображение, y – истинная метка, $\hat{y}$ – целевая метка, $\mathcal{L}$ – функция потерь, i – номер итерации, $\Pi x + \epsilon$ – оператор проекции, который обрезает входные данные в позициях вокруг предопределенного диапазона возмущений, α – шаг, ϵ – максимально возможное возмущение, ℓp – норма.

Атака состязательной наклейкой отличается от обычной PGD атаки тем, что область возмущения на изображении ограничена, при этом значение возмущения ϵ не ограничено [3]. В данном эксперименте рассматриваются квадратные наклейки для атаки, реализованной в библиотеке Adversarial Robustness Toolbox, и для классической PGD-атаки, реализованной авторами в рамках данного исследования и описанной в статье [2].

В классической атаке PGD подбор наклейки, которая бы изменила метку изображения при предсказании, происходит следующим образом, аналогично методу из статьи [4]. Изображение разделяется на участки с помощью квадратной сетки, где длина ячейки сетки равна длине стороны квадратной наклейки. Далее наклейка подбирается отдельно на каждом участке сетки по порядку до тех пор, пока не найдется такое место и такая наклейка, при которой атака успешна. Если после перебора всех мест по сетке не удалось найти наклейку, которая бы меняла метку изображения, то атака считается неуспешной.

В рамках данного исследования реализация указанных методов атаки включала в себя расчёт градиента по всему изображению, а не только по области с состязательной наклейкой. Наклейку, полученную описанным образом, будем называть состязательной наклейкой PGD.

2.2 Состязательные атаки в физическом мире

Многие атаки генерируют состязательные изображения, основываясь на готовом классифицируемом изображении, добавляя к нему некоторые возмущения. Такие атаки обычно используются в случае классификации после получения фотографии, т.е. атака не изменяет реальные объекты. Но если распечатать состязательную наклейку и поместить ее на реальный предмет, то успешность атаки обычно снижается. Это происходит, поскольку добавленные возмущения для создания состязательного изображения изменятся из-за различных эффектов физического мира, несовершенства или особенностей сенсоров камеры, ракурса съёмки и т. д.

В статье [8] предлагается метод генерации состязательных примеров, которые были бы устойчивыми к трансформациям физического мира. Авторы статьи назвали предложенный метод Expectation over transformation (EOT), т.е. ожидание над трансформацией. EOT основывается на моделировании трансформаций физического мира в рамках процедуры генерации состязательной атаки. Задача, которая решается в данном методе, представлена формулой (4):

$$\arg\max_{x'} \mathbb{E}_{t\sim T}\left[\log P\left(y_t|t(x')\right)\right]$$
$$\text{при условии } \mathbb{E}_{t\sim T}\left[d\left(t(x'),t(x)\right)\right] < \varepsilon, \mathrm{x} \in [0,1]^D. \quad (4)$$

В этой формуле обозначено за T – распределение возможных трансформаций, $t(\cdot)$ – функция преобразования из T, $d(\cdot,\cdot)$ – функция расстояния, x – исходное изображение, x' – состязательное изображение, D – размерность пространства изображений, yt – целевой класс, P – вероятность, а ϵ – максимальное возмущение.

Данный метод вместо оптимизации вероятности предсказания целевого класса на одном исходном изображении использует выбранное распределение трансформаций и оптимизирует математическое ожидание по этому распределению. Также вместо обычного расстояния между исходным и состязательным изображениями в данном методе рассматривается ожидаемое расстояние по всем трансформациям. Авторы статьи [8] рассматривают двумерные и трёхмерные трансформации, среди которых повороты, масштабирование, затемнение и гауссовский шум.

2.3 PGD-атака, реализованная в Adversarial Robustness Toolbox

В дополнение к классической PGD атаке в нашей статье исследуется поведение нейронных сетей ResNet и BagNet при PGD атаке наклейкой, реализованной в библиотеке Adversarial

Robustness Toolbox (ART) [9] и описанной в статье [3]. Данная атака является модификацией рассмотренной ранее PGD атаки, отличаясь от неё инициализацией состязательной наклейки и ее местоположением на изображении. В реализованной в ART атаке наклейка инициализируется однотонной картинкой – средним между максимальным и минимальным возможными значениями пикселя. Затем на каждой итерации наклейка помещается в случайное место на атакуемом изображении.

Указанный выше способ генерации состязательной наклейкой позволяет делать её более универсальной, не зависящей от освещения, угла камеры, места и т. д.

Наклейку, полученную в процессе этой атаки, будем называть состязательной наклейкой ART.

3. Защита от состязательных атак на основе изменения архитектуры сети

В этом разделе мы рассмотрим методы защиты от состязательных наклеек с помощью изменения архитектуры нейронной сети.

3.1 Метод защиты на основе модели BagNet

Архитектура BagNet [4] основана на идее модели Bag-Of-Local-Features (мешок слов), которая применяется для задач, связанных с обработкой текстов. Её идея состоит в представлении текста в виде словаря из слов и количеством их вхождений в текст. Так, текст можно представить в виде числового вектора.

Для изображений в качестве визуальных слов используются отдельные кусочки изображения. По каждому из кусочков в отдельности делается предсказание. Итоговые выходные значения модели определяются голосованием по всем кускам изображения.

Таким образом, первым шагом в модели BagNet является получение 2048-мерного представления для каждого фрагмента исходного изображения размером $q\times q$ пикселей. Для этого используется несколько блоков ResNet и линейный классификатор. Так получается значение выхода на каждом классе для каждого фрагмента изображения. Затем значение выходов на всех фрагментах усредняются для получения итогового значения для каждого класса.

BagNet отличается от классического ResNet только заменой большинства свёрток 3×3 на 1×1, что позволяет ограничить размер рецептивного поля самой верхней свёртки до размера $q\times q$.

В экспериментах к данной статье в качестве q для модели BagNet было выбрано q=33.

3.2 Clipped BagNet (CBN)

Слабой стороной BagNet является то, что большое изменение одного выходного значения для хотя бы одного фрагмента изображения приведёт к увеличению среднего значения, в результате чего модель неустойчива к атакам.

Для решения этой проблемы можно использовать модификацию модели – Clipped BagNet (CBN). Данная модификация реализуется с помощью ограничения выходов модели на каждом фрагменте изображения. В статье [5] показано, что в качестве ограничения можно выбрать функцию $f(x)$=tanh(ax+b) с параметрами a=0.05 и b=−1.

На рис. 2 изображена тепловая карта выходов модели BagNet (до ограничения выходных значений) и CBN до трансформаций изображений для случая, когда CBN предсказывает метку верно, а BagNet ошибается. По этому рисунку можно понять, как изображение видит модель BagNet, а как CBN.

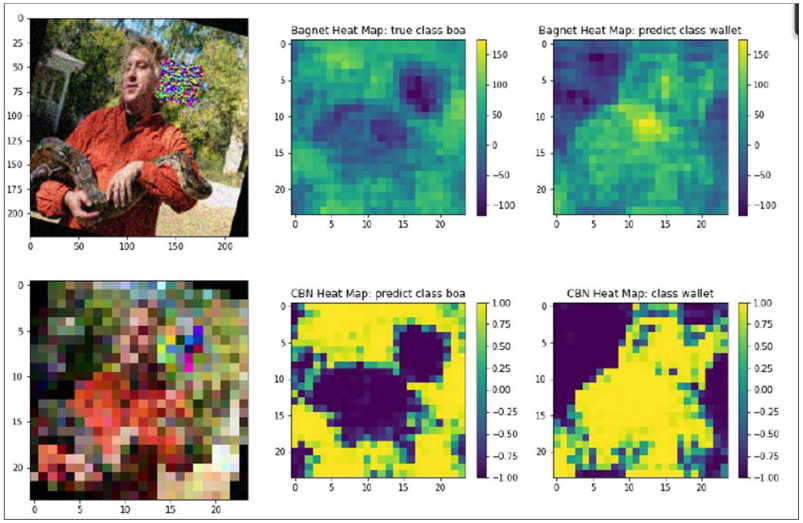

Рис. 2. Тепловая карта выходов моделей для истинного и предсказанного классов на каждом участке рецептивного поля для состязательного изображения с трансформациями. Случай, когда CBN предсказывает верно, а BagNet ошибается. Для модели CBN в нижнем ряду справа изображена тепловая карта для класса, который предсказала модель BagNet

Fig. 2. Heat map of model outputs for true and predicted classes at each region of the receptive field for an adversarial image with transformations. The case when CBN predicts correctly, but BagNet is wrong. For the CBN model, the bottom row on the right shows the heat map for the class predicted by the BagNet model

3.3 Clipped BagNet с состязательным обучением (ADV CBN)

В статье [6] авторами было предложено улучшить устойчивость модели к атакам с помощью PatchGuard масок. Идея одной из масок (маска «m») состоит в том, что необходимо до агрегирования выходов модели пропускать свидетельства для каждого класса через окно определенного размера и маскировать (убирать) максимальное значение, после чего уже агрегировать выходы и делать предсказание. Размер окна зависит от структуры модели и от максимально возможного размера атакующей наклейки. Вторая маска (маска «cbn») заключается в добавлении ограничивающей свидетельства класса функцией перед финальной агрегацией. В качестве ограничивающей функции была выбрана аналогичная [5] функция $f(x)=\tanh(ax+b)$ с параметрами $a=0.05$ и $b=-1$. Авторы статьи [6] предложили использовать такие маски для состязательного обучения модели CBN. Обученная таким способом модель в данной статье обозначается как ADV CBN (Adversarial training for Clipped BagNet).

4. Влияние трансформаций на успешность состязательных атак

В экспериментах, описываемых в данной статье, рассматривались изображения размера 224×224 пикселя, в качестве максимального возмущения для атак было выбрано $\epsilon=1$. Значения компонент пикселей оригинальных изображений заданы числами от 0 до 1.

В данном исследовании проводилась классическая PGD атака наклейкой на модели ResNet, CBN и ADV CBN, а также проводились атаки, реализованные в библиотеке Adversarial Robustness Toolbox [9]. В результате этих атак были получены состязательные изображения, над которыми проводились некоторые трансформации. Эти трансформированные состязательные изображения снова подавались на вход модели. Таким образом проверялась

устойчивость моделей ResNet и CBN на атаках с трансформациями состязательных изображений.

Параметры состязательной атаки ART настраивались подобно параметрам атаки PGD. В эксперименте рассматривались квадратные наклейки без дополнительных поворотов и каких-либо трансформаций при генерации атаки (таким образом, стороны наклейки всегда параллельны сторонам изображения).

В качестве дополнительного результата были получены точности модели на состязательных изображениях для других моделей. Трансформации изображений к ним также применялись.

Помимо обычных моделей, в экспериментах рассматривались модели с масками из статьи [6]. Для модели ResNet рассматривались обе маски, а для моделей CBN и ADV CBN только маска «m», поскольку применение маски «cbn» заключается в ограничении свидетельств модели гиперболическим тангенсом, которое и так содержится в этих моделях.

4.1 Модели и данные

В экспериментах использовались модели ResNet-18 (ResNet), Clipped BagNet-33 (CBN) и Adversarial Clipped BagNet-33 (ADV CBN). В качестве данных была выбрана часть датасета ImageNet [10], состоящая из 10 классов по 1000 изображений в каждом. Тестовая выборка составляла 33% от используемых данных, остальная часть данных использовалась для обучения моделей.

CBN и ADV CBN обучались с помощью SGD оптимизатора с параметром learning rate =0.001, а ResNet с помощью SGD оптимизатора и с параметром learning rate =0.01.

Код к данному исследованию, а также ссылки на используемые данные и модели представлены в репозитории https://github.com/kekaterina/transformation_adversarial_attack.

4.2 Трансформация изображений

Трансформации изображений помогают проверить надежность атаки наклейкой в физическом мире, когда материальная наклейка нанесена на реальный предмет. В экспериментах использовались 3 типа трансформаций: затемнение, гауссовский шум и поворот с масштабированием, а также комбинация из всех этих трансформаций одновременно.

Табл. 1. Значения параметров трансформаций
Table 1. Values of transformation parameters

Трансформация	*Минимум*	*Максимум*
Масштаб (№1)	0.9	1.4
Масштаб (№2)	0.8	1.0
Поворот (№1)	$-22.5°$	$22.5°$
Поворот (№2)	$-10.0°$	$10.0°$
Осветвление и затемнение	-0.05	0.05
Трансформация	*Среднее значение*	*Отклонение*
Гауссовский шум	0.0	0.1

Параметры трансформаций (для масштаба и поворота варианты №1) взяты из статьи [8] и находятся в табл. 1. В указанных границах строились равномерные распределения параметров, из которых для каждого изображения случайным образом подбирались конкретные параметры для трансформаций.

Для экспериментов с состязательной наклейкой ART использовались масштаб №1 и поворот №1, а для экспериментов с состязательной наклейкой PGD использовались масштаб №2 и поворот №2. Это обусловлено тем, что в состязательных наклейках ART местоположение

наклейки обычно находится в середине изображения, поэтому трансформировать такое изображение без большого изменения площади наклейки можно сильнее. Для состязательной наклейки PGD параметры масштабирования и поворота уменьшены, чтобы не обрезать большую часть изображения с состязательной наклейкой. В последующих таблицах трансформации типа осветления/затемнения будут обозначаться как затемнение.

5. Результаты экспериментов

В этом разделе мы сравним устойчивость моделей CBN, ADV CBN и ResNet к состязательным атакам в условиях трансформаций и исследуем переносимость атак.

5.1 Трансформации

Эксперименты над трансформированными состязательными изображениями проводились с помощью двух типов состязательных атак квадратной наклейкой:

- наклейки PGD разных размеров: 60×60, 50×50, 40×40, 32×32, 20×20 пикселей;
- наклейки ART при разных значениях параметра $patchScale$≈0.446,0.268,0.223,0.179,0.142,0.079; параметр $patchScale$ рассчитывался таким образом, чтобы наклейка ART соответствовала доли равной доли наклейки PGD от площади исходного изображения.

На рис. 1 представлены изображения атакованные состязательными PGD наклейками, до и после трансформаций.

Точность моделей рассчитывалась на изображениях из тестового набора изображений – 3273 изображения из 10 классов.

Табл. 2. Атака состязательной наклейкой PGD. Точность моделей на чистых и атакованных изображениях с трансформациями и без них. Размер наклейки указан в пикселях. Красными и зелеными цифрами в скобках обозначена разница в точности CBN и ADV CBN по сравнению с аналогичным экспериментом на ResNet. ***Жирным*** *шрифтом выделен лучший результат в каждом столбце. Результаты для модели ADV CBN получены на 1000 тестовых изображениях, результаты остальных моделей представлены для всего тестового набора*

Table 2. Attack with the adversarial PGD patch. Accuracy of models on clean and attacked images with and without transformations. The patch size is in pixels. The red and green numbers in parentheses indicate the difference in CBN and ADV CBN accuracy compared to the same experiment on ResNet. ***Bold*** *indicates the best result in each column. The results for the ADV CBN model were obtained on 1000 test images, the results of other models are presented for the entire test set*

Модель	*Трансформация*	*Исходное изображение + трансформации*	*Наклейка 20×20*	*Наклейка 32×32*	*Наклейка 40×40*	*Наклейка 50×50*	*Наклейка 60×60*
ResNet-18	Без трансформации	**0.790**	0.128	0.027	0.013	0.009	0.007
	Поворот	0.696	**0.620**	**0.615**	**0.560**	**0.500**	**0.428**
	Шум Гаусса	0.715	0.317	0.324	0.060	0.022	0.010
	Затемнение	0.787	0.218	0.236	0.025	0.011	0.073
	Комбинация	0.626	0.563	0.569	0.512	0.475	0.406
CBN-33	Без трансформации	**0.675** (−0.115)	0.601 (+0.473)	0.555 (+0.528)	0.520 (+0.507)	0.456 (+0.447)	0.398 (+0.391)
	Поворот	0.635 (−0.061)	0.628 (+0.008)	0.609 (−0.006)	**0.609** (+0.049)	**0.588** (+0.088)	**0.567** (+0.139)
	Шум Гаусса	0.262 (−0.453)	0.255 (−0.062)	0.246 (−0.078)	0.241 (+0.181)	0.234 (+0.014)	0.219 (+0.209)
	Затемнение	0.668 (−0.119)	**0.643** (+0.425)	**0.616** (+0.380)	0.565 (+0.540)	0.508 (+0.497)	0.452 (+0.379)
	Комбинация	0.263 (−0.363)	0.262 (−0.301)	0.259 (−0.310)	0.257 (−0.255)	0.250 (−0.225)	0.247 (−0.159)
ADV CBN-33	Без трансформации	**0.696** (−0.094)	0.639 (+0.511)	0.612 (+0.585)			
	Поворот	0.642 (−0.054)	0.616 (−0.004)	**0.613** (−0.002)			
	Шум Гаусса	0.373 (−0.342)	0.377 (−0.062)	0.346 (+0.060)			
	Затемнение	0.690 (−0.097)	**0.642** (+0.424)	0.610 (+0.374)			
	Комбинация	0.390 (−0.236)	0.385 (−0.187)	0.378 (−0.191)			

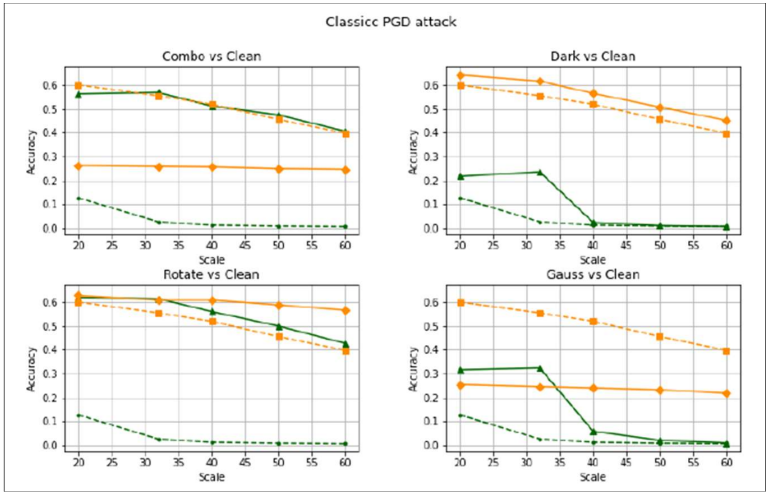

Рис. 3. Изменение точности моделей на атакованных классической PGD-атакой изображениях с трансформациями и без. По горизонтальной оси указан размер состязательной наклейки в пикселях

Fig. 3. Changing the accuracy of models on images attacked by the classical PGD attack with and without transformations. The horizontal axis shows the size of the contest sticker in pixels

Результаты экспериментов с трансформациями и состязательными наклейками PGD представлены в табл. 2 и на рис. 3. Данные результаты показывают, что:

- модели CBN и ADV CBN ожидаемо лучше защищены от атак без трансформаций, чем модель ResNet. Это совпадает с выводами в [4, 5, 11];
- из всех трансформаций без состязательной атаки гауссовский шум больше всех снижает точность BagNet: CBN – на 45%, ADV CBN – на 34%, тогда как точность ResNet упала лишь на 7,5%;
- при всех размерах состязательной наклейки самые низкие результаты CBN и ADV CBN показывают на изображениях с гауссовским шумом;

- при комбинированной трансформации CBN и ADV CBN оказались хуже ResNet во всех проведенных экспериментах: как с состязательными наклейками, так и без;
- поворот и затемнение/осветление состязательных изображений значительно повышает точность моделей ResNet и CBN;
- модель ADV CBN имеет точность выше, чем у обычного CBN, на изображениях без трансформаций. Но в половине экспериментов с трансформированными состязательными изображениями ADV CBN проигрывает CBN;
- гауссовский шум чуть меньше влияет на ADV CBN, чем на CBN – точность на зашумленных состязательных изображениях у ADV CBN выше.

Таким образом, модели CBN и ADV CBN показывают лучшую точность, чем ResNet на атаках наклейками. Однако при атаках с комбинацией трансформаций ResNet оказывается лучше CBN. Добавление состязательного обучения к модели CBN (для получения ADV CBN) дает значимое преимущество только при работе с состязательными изображениями, не подвергшимся трансформациям.

Заметим, что для модели ADV CBN-33 представлены результаты не для всех рассматриваемых размеров наклейки, поскольку исследование этой модели не было основной целью данной работы и проводилось по остаточному принципу, насколько хватило ресурсов. Также, результаты проведенных над ADV CBN-33 экспериментов показывают, что эта модель во многом похожа на модель CBN-33.

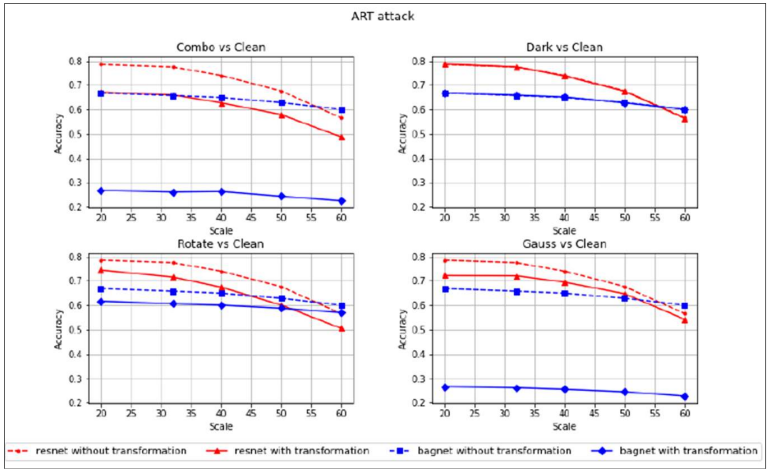

Рис. 4. Изменение точности моделей на атакованных состязательной наклейкой ART изображениях с трансформациями и без. По горизонтальной оси указан размер состязательной наклейки в пикселях
Fig. 4. Changing the accuracy of models on images attacked by an ART adversarial sticker with and without transformations. The horizontal axis shows the size of the contest sticker in pixels

Результаты экспериментов для состязательных наклеек ART показаны в табл. 3 и на рис. 4. При выбранных нами гиперпараметрах атака ART оказалась слабее PGD, т.к. при одинаковых размерах наклейки PGD снижала точность моделей сильнее. Результаты в табл. 3 показывают, что:

- только при размерах наклейки ART 60×60 пикселей CBN показывает результат выше, чем ResNet, в 3 из 5 экспериментах; при более маленьких наклейках CBN всегда остается менее точным;
- лучшие результаты обе модели показывают на изображениях без трансформаций и на изображениях с затемнением;

- среди всех трансформаций гауссовский шум больше всего снижает точность CBN; аналогичный результат получен и в экспериментах с PGD наклейкой.

*Табл. 3. Атака состязательной наклейкой ART. Точность моделей на чистых и атакованных изображениях с трансформациями и без них. Размер наклейки указан в пикселях. Красными и зелеными цифрами в скобках обозначена разница в точности CBN по сравнению с аналогичным экспериментом на ResNet. **Жирным** шрифтом выделен лучший результат в каждом столбце*
Table 3. Attack with the adversarial patch ART. Accuracy of models on clean and attacked images with and without transformations. The patch size is in pixels. The red and green numbers in brackets indicate the difference in CBN accuracy compared to the same experiment on ResNet. Bold indicates the best result in each column

Модель	*Трансформация*	*Исходное изображение + трансформации*	*Наклейка 20×20*	*Наклейка 32×32*	*Наклейка 40×40*	*Наклейка 50×50*	*Наклейка 60×60*
ResNet-18	Без трансформации	**0.790**	0.786	0.775	**0.740**	**0.677**	**0.566**
	Поворот	0.735	0.745	0.716	0.675	0.602	0.506
	Шум Гаусса	0.714	0.722	0.722	0.695	0.647	0.543
	Затемнение	0.787	**0.787**	**0.778**	0.738	0.675	0.564
	Комбинация	0.665	0.672	0.680	0.628	0.580	0.488
CBN-33	Без трансформации	**0.675** (−0.115)	0.669 (−0.117)	0.658 (−0.117)	0.650 (−0.090)	**0.629** (−0.048)	**0.601** (+0.035)
	Поворот	0.623 (−0.112)	0.617 (−0.128)	0.607 (−0.109)	0.601 (−0.074)	0.588 (−0.014)	0.570 (+0.064)
	Шум Гаусса	0.261 (−0.453)	0.266 (−0.456)	0.262 (−0.460)	0.255 (−0.440)	0.245 (−0.402)	0.228 (−0.315)
	Затемнение	0.670 (−0.117)	**0.670** (−0.117)	**0.660** (−0.118)	**0.652** (−0.086)	0.628 (−0.047)	**0.601** (+0.037)
	Комбинация	0.266 (−0.399)	0.269 (−0.403)	0.262 (−0.418)	0.264 (−0.364)	0.244 (−0.336)	0.225 (−0.263)

Таким образом, при слабых атаках, несильно снижающих точность ResNet, модель с архитектурой CBN остается менее точной.

Точность моделей на оригинальных и состязательных изображениях при использовании масок, описанных в статье [6], приводится в табл. 4. По представленным данным видно, что:

- для модели ResNet маски в среднем увеличивают устойчивость модели к атаке, но маска «cbn» в данном случае показывает себя лучше маски «m». ResNet в совокупности с маской «m» имеет более низкую точность на оригинальных изображениях в сравнении с маской «cbn»; зато для моделей CBN и ADV CBN маска «m» улучшает устойчивость к атаке и к трансформациям;
- применение поворота к состязательным изображениям на модели ResNet с применением маски и на обеих версиях CBN с применением маски действует на точность по-разному; для ResNet с маской это повышает точность модели, а для CBN и ADV CBN с маской наоборот понижает.

Заметим, что многих случаях PGD наклейка располагается в верхнем левом углу изображения (рис. 1). Это связано с тем, что алгоритм и гиперпараметры PGD-атаки зачастую позволяют успешно атаковать модель с первой итерации и, следственно, не перебирать другие локации для наклейки. Мы предполагаем, что любое другое положение наклейки

(например, в центре изображений) не изменило бы описанных нами закономерностей. Однако доказательство этой гипотезы мы оставляем для дальнейших исследований.

Табл. 4. Точность моделей на чистых и атакованных изображениях с трансформациями и без них для моделей с масками PatchGuard. Размер наклейки указан в пикселях. Синим цветом отмечена точность на «родных» состязательных изображениях для модели. Результаты для 1000 тестовых изображений. Для состязательных изображений размер наклейки равен 32×32

Table 4. Accuracy of models on clean and attacked images with and without transformations for models with PatchGuard masks. The patch size is in pixels. The blue color marks the accuracy on the "native" adversarial images for the model. Results for 1000 test images. For competitive images, the sticker size is 32×32

Модель	*Трансформация*	*Исходное изображение + трансформации*	*Состязательные изображения для ResNet*	*Состязательные изображения для CBN*	*Состязательные изображения для ADV CBN*
ResNet-18 + маска «cbn»	Без трансформации	0.792	0.567	0.785	0.784
	Поворот	0.730	**0.667**	0.724	0.709
	Шум Гаусса	**0.793**	0.579	**0.788**	**0.788**
	Затемнение	0.731	0.482	0.730	0.732
	Комбинация	0.653	0.610	0.654	0.655
ResNet-18 + маска «m»	Без трансформации	0.382	**0.637**	**0.638**	**0.662**
	Поворот	0.516	**0.427**	0.490	0.478
	Шум Гаусса	0.661	0.377	0.627	0.632
	Затемнение	0.585	0.374	0.584	0.566
	Комбинация	0.289	0.269	0.278	0.278
CBN-33 + маска «m»	Без трансформации	**0.688**	**0.681**	**0.655**	**0.679**
	Поворот	0.652	0.652	0.628	0.642
	Шум Гаусса	0.667	0.667	0.638	0.669
	Затемнение	0.269	0.273	0.271	0.274
	Комбинация	0.278	0.289	0.269	0.278
ADV CBN-33 + маска «m»	Без трансформации	**0.694**	**0.696**	**0.677**	**0.633**
	Поворот	0.632	0.634	0.627	0.604
	Шум Гаусса	0.683	0.690	0.673	0.617
	Затемнение	0.382	0.388	0.375	0.365
	Комбинация	0.394	0.383	0.386	0.385

5.2 Переносимость атак

В табл. 5 представлены точности моделей на оригинальных изображениях без состязательной наклейки и на «чужих» состязательных изображениях с наклейкой размера 32×32 пикселя. Зеленым цветом обозначена точность моделей на оригинальных изображениях, а черным – на состязательных. В этом эксперименте рассматривалась классическая PGD атака. По результатам видно, что:

- атака переносится слабо; на собственных состязательных изображениях точность моделей (табл. 2) гораздо ниже, чем на чужих;
- на чужих состязательных изображениях без трансформаций точность снижается незначительно, зато при повороте этих изображений разница с точностью на оригинальных изображениях без трансформаций становится существеннее.

Табл. 5. Точность моделей на состязательных изображениях других моделей. Размер наклейки 32×32 в пикселях. По горизонтали отмечены названия моделей, для которых генерировались состязательные изображения. На диагональных блоках (где название модели совпадает с названием модели, для которой строились состязательные изображения) зелёным цветом выделена точность на оригинальных изображениях без наклейки. Результаты для 1000 тестовых изображений

Tab. 5. Accuracy of models on competitive images of other models. Patch size is 32×32 in pixels. The names of the models for which competitive images were generated are marked horizontally. On the diagonal blocks (where the name of the model matches the name of the model for which the competitive images were built), the accuracy on the original images without a patch is highlighted in green. Results for 1000 test images

Модель	*Трансформация*	*Состязательные изображения для ResNet*	*Состязательные изображения для CBN*	*Состязательные изображения для ADV CBN*
ResNet-18	Без трансформации	**0.794**	**0.785**	**0.789**
	Поворот	0.709	0.708	0.698
	Шум Гаусса	0.791	0.784	0.786
	Затемнение	0.727	0.722	0.725
	Комбинация	0.628	0.632	0.630
CBN-33	Без трансформации	**0.694**	**0.694**	**0.684**
	Поворот	0.643	0.642	0.633
	Шум Гаусса	0.689	0.678	0.669
	Затемнение	0.276	0.269	0.272
	Комбинация	0.273	0.263	0.272
ADV CBN-33	Без трансформации	**0.693**	0.679	**0.696**
	Поворот	0.647	0.635	0.650
	Шум Гаусса	0.687	**0.685**	0.692
	Затемнение	0.378	0.377	0.374
	Комбинация	0.369	0.373	0.380

Аналогичные результаты переносимости атак получены для моделей с масками, описанными в статье [6], представлены в табл. 4. Результаты в данном случае аналогичны эксперименту для моделей без масок.

По результатам можно сделать вывод, что переносимость состязательных изображений одной модели на другую для данных видов атак и моделей есть, но она очень слабая.

6. Заключение

По результатам экспериментов можно сделать следующие выводы:

- добавление гауссовского шума и комбинации из трансформаций особенно сильно понижают предсказательную точность моделей CBN и ADV CBN даже без состязательных наклеек; таким образом, модель CBN плохо справляется с изображениями низкого качества;
- поворот и затемнение состязательных изображений значительно снижает успешность атак; несмотря на то, что рассматриваемое затемнение/осветление не видно для человеческого глаза, при некоторых размерах состязательных наклеек (например, 20×20, 32×32 при обоих методах генерации наклеек) лучше всего смогли нивелировать атаку;
- переносимость состязательных изображений одной модели на другую для данных видов атак и моделей есть, но она очень слабая.

Таким образом, архитектура CBN и ADV CBN может защитить от состязательных атак только при двух условиях:

- качество изображения хорошее, т.е. не подвержено таким трансформациям как зашумление, поворот, затемнение;
- атака значительно снижает точность незащищенной модели.

В случае же работы с изображениями плохого качества или при слабых атаках лучше предпочесть классическую модель классификации изображений или другие методы защиты от состязательных атак.

Список литературы / References

[1] Goodfellow I.J., Shlens J., Szegedy C. Explaining and Harnessing Adversarial Examples. ArXiv 1412.6572, 2014, 11 p.
[2] Madry A., Makelov A. et al. Towards Deep Learning Models Resistant to Adversarial Attacks. ArXiv 1706.06083, 2017, 28 p.
[3] Brown T.B., Mané D. et al. Adversarial Patch. ArXiv 1712.09665, 2017, 6 p.
[4] Brendel W., Bethge M. Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet. ArXiv 1904.00760, 2019, 13 p.
[5] Zhang Z., Yuan B. et al. Clipped BagNet: Defending Against Sticker Attacks with Clipped Bag-of-features. In Proc. of the 2020 IEEE Security and Privacy Workshops (SPW), 2020, pp. 55-61.
[6] Xiang C., Bhagoji A.N. et al. PatchGuard: Provable Defense against Adversarial Patches Using Masks on Small Receptive Fields. ArXiv 2005.10884, 2020, 23 p.
[7] Dong Y., Liao F. et al. Boosting Adversarial Attacks with Momentum. In Proc. of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 9185-9193.
[8] Athalye A., Engstrom L. et al. Synthesizing Robust Adversarial Examples. ArXiv 1707.07397, 2017, 19 p.
[9] Nicolae M., Sinn M. et al. Adversarial Robustness Toolbox v1.0.0. ArXiv, abs/1807.01069, 2018, 34 p.
[10] Russakovsky O., Deng J. et al. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision, vol. 115, issue 3, 2015, pp. 211-252.
[11] Uesato J., O'Donoghue B. et al. Adversarial Risk and the Dangers of Evaluating Against Weak Attacks. ArXiv, abs/1802.05666, 2018, 13 p.

Информация об авторах / Information about authors

Екатерина Олеговна КУРДЕНКОВА – стажер-исследователь Центра доверенного искусственного интеллекта ИСП РАН (Научные интересы: исследование и разработка нейросетевых архитектур, устойчивых к состязательным атакам.

Ekaterina Olegovna KURDENKOVA – Graduate Research Trainee at ISP RAS Research Center for Trusted Artificial Intelligence. Research interests: research and development of neural network architectures aimed at robustness to adversarial attacks.

Мария Сергеевна ЧЕРЕПНИНА – студентка магистратуры Мюнхенского технического университета. Научные интересы: объяснимый искусственный интеллект, бизнес-информатика.

Maria Sergeevna CHEREPNINA – Master's Student at Technical University of Munich. Research interests: explainable artificial intelligence, business informatics.

Анна Сергеевна ЧИСТЯКОВА – студентка бакалавратуры ф-та ВМК МГУ, лаборант Центра доверенного искусственного интеллекта ИСП РАН. Научные интересы: методы регуляризации моделей глубокого обучения, обеспечивающие устойчивость к состязательным атакам.

Anna Sergeevna CHISTYAKOVA – Bachelor's Student of the faculty of the CMC of Moscow State University, Assistant at ISP RAS Research Center for Trusted Artificial Intelligence. Research interests: regularization methods for deep learning models providing robustness to adversarial attacks.

Константин Владимирович АРХИПЕНКО – младший научный сотрудник Центра доверенного искусственного интеллекта ИСП РАН. Научные интересы: защита моделей машинного обучения от атак, объяснимый искусственный интеллект.

Konstantin Vladimirovich ARKHIPENKO – Junior Research Fellow at ISP RAS Research Center for Trusted Artificial Intelligence. Research interests: defending machine learning models against attacks, explainable artificial intelligence.