УДК 629.7.08

ИССЛЕДОВАНИЕ МЕТОДОВ АТАК НА СВЕРТОЧНЫЕ НЕЙРОННЫЕ СЕТИ

Бодунков Н.Е.
Арефин В.В.
Кобринец С.К.
Московский авиационный институт (национальный исследовательский университет), Москва

Ключевые слова: распознавание объектов, сверточные нейронные сети, состязательные атаки

Ведение

В настоящее время беспилотные транспортных средств (БТС), в том числе беспилотные летательные аппараты (БЛА) активно используются для решения широкого спектра целевых задач (ЦЗ), например, поиска объектов, мониторинга, доставки грузов и пр.

Многие задачи, связанные с обнаружением, распознаванием и оценкой положения объектов интереса, эффективно решаются с использованием сверточных нейонных сетей. Однако с ростом популярности нейронных сетей растет интерес к задачам противодействия им. Недавние исследования показали, что точность работы сверточных нейронных сетей можно существенно уменьшить добавлением относительно небольших возмущений к входному изображению. Такие воздействия называются атаками.

1 Обзор методов атак

Атаки можно разделить на две группы – атаки типа «белый ящик», и «черный ящик». В атаке «Белый ящик» атакующая сторона знает все об атакуемой нейронной сети – структуру, веса обученной сети и даже выборку изображений, на которой производилось обучение сети [Wang Y. 2021].

При атаке «черный ящик» атакующая сторона ничего не знает о структуре сети и ее параметрах. Однако доступ к сети все равно необходим. Такая атака часто формируется путем подачи на вход НС различных искаженных изображений и анализом достоверностей и меток классов на ее выходе [Pin-Yu Chen 2017].

Обычно сверточные нейронные сети представляют собой различные комбинации специализированных сверточных слоев (которые

предназначены для выделения признаков) и «классических» полносвязных слоев, выступающих в роли классификаторов. Суть всех атак заключается в том, чтобы на атакуемое изображение добавить признаки существенно более весомые (для классификатора) чем признаки атакуемого объекта. Таким образом, классификатор будет реагировать на искажения, а не на сам объект. Важно отметить, что данные виды атак не только маскируют признаки объекта (существующие сети устойчивы к таким видам искажений), но и формируют новые, ложные признаки.

Существует несколько распространенных методов атак на нейронные сети: «пиксельные атаки», «вредоносное искажение» и «вредоносные патчи» (или «заплатки»). Пиксельная атака заключается в изменении одного или нескольких пикселей исходного изображения. Во время атаки ищутся «чувствительные» для НС области изображения. При изменении цвета этих областей повышается вероятность ошибок классификатора. Суть атаки состоит в определении количества, положения и цвета «чувствительных» областей (пикселей). При этом, даже с изменением одного пикселя возможно реализовать атаку. В статье [Su J. 2019] показана возможность такой атаки. Также, показана принципиальная возможность и реализация данного вида атак с 1 – 5 пикселями. Следует отметить, что данный вид атак относится к классу «черный» ящик, т.е. не обязательно обладать информацией о структуре сети для ее реализации.

2 Методика исследования

Данная работа посвящена исследованию устойчивости пиксельной атаки к различным дестабилизирующим факторам и возможности ее реализации в реальных условиях для противодействия системе мониторинга дорожной обстановки. В ходе эксперимента была проведена пиксельная атака на изображение, на котором представлены объекты класса «Автомобиль» полученные с борта летательного аппарата (вид «сверху»).

В рамках исследования атака проводилась на нейронную сеть класса YOLO v4. В связи с отсутствием обученной на исследуемый класс объектов сетей была подготовлена выборка и произведено обучение.

Чтобы оценить возможность реализации данной атаки в реальных условиях необходимо провести ряд исследований ее устойчивости к различным дестабилизирующим факторам: изменению яркости, контраста, масштаба, повороту, различным шумам. В рамках данной работы рассматривались изменения масштаба и зашумление искаженного изображения. Рассмотрена методика проведения пиксельной атаки для сетей VGG16 и YOLOv4, предварительно обученной на изображения транспортных средств с борта летательного аппарата (как имитации системы дорожного мониторинга).За основу была взята существующая реализация

пиксельной атаки. Однако, для ее использования под нейронную сеть YOLOv4 потребовались некоторые изменения. Запуск доработанной атаки на тестовом изображении показал ее работоспособность. Т.е. все объекты интереса на искаженном изображении перестали распознаваться. Следует отметить, что исходные достоверности распознавания объектов интереса были близки к 1.

3 Анализ результатов

Сравнение изменения достоверности распознавания объектов на эталонном и атакованном изображениях при различных значениях шума, показало, что атака устойчива к рассматриваемым воздействиям на всем диапазоне изменения дисперсии шума.

Показано, что изменение масштаба изображения также мало влияет на результат атаки. При значениях масштабирующего коэффициента больше 0.25 нейронная сеть распознает объекты с достоверностью близкой к 1. Однако, после атаки достоверность класса «автомобиль» для тех же объектов становится близкой к 0.

Выводы

Полученные результаты показали устойчивость атаки в широком диапазоне изменения масштаба и дисперсии шума. Таким образом, несмотря на относительную простоту (по сравнению с другими видами атак), пиксельную атаку в реальных условиях можно считать физически реализуемой.

Список литературы

[Su J. 2019] Su J., Vasconcellos D. V., Sakurai K. One pixel attack for fooling deep neural networks // IEEE Transactions on Evolutionary Computation, 2019, vol. 23, pp. 828 – 841. DOI: 10.1109/TEVC.2019.2890858

[Bhambri S. 2020] Bhambri S., Muku S., Tulasi A., Balaji A. B. A Survey of Black-Box Adversarial Attacks on Computer Vision Models. arXiv:1912.01667, 2020, 33 p.

[Goodfellow I. 2015] Goodfellow I., Shlens J., and Szegedy C. Explaining and harnessing adversarial examples. arXiv:1412.6572, 2015.

[Moosavi-Dezfooli S. 2017] Moosavi-Dezfooli S., Fawzi A., Fawzi O., Frossard P. Universal adversarial perturbations. arXiv:1610.08401, 2017.

[Liu X. 2019] Liu X., Yang H., Liu Z., Song L., Li H., Chen Y. DPATCH: An Adversarial Patch Attack on Object Detectors. arXiv:1806.02299, 2019.

[Tom B. 2018] Tom B. Mané D., Roy A. Adversarial patch. arXiv:1712.09665, 2018.

[S. Das 2011] Das S., Suganthan P. Differential evolution: A survey of the state-of-the-art // IEEE transactions on evolutionary computation, 15(1): pp.4–31, 2011.

[Simonyan K. 2015] Simonyan K., Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556v6, 2015.

[Bochkovskiy A. 2020] Bochkovskiy A., Chien-Yao Wang, Hong-Yuan Mark Liao. YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv:2004.10934, 2020.

[Bodunkov N. 2018] Bodunkov N., Kim N., Mikhaylov N. The ground objects monitoring by uav using a search entropy evaluation // Applied Informatics and Cybernetics in Intelligent Systems pp.443-452. DOI:10.1007/978-3-030-51974-2_42

[Wang Y. 2021] Wang Y., Liu J., Chang X, Mišić J., Mišić Vojislav B. IWA: Integrated Gradient based White-box Attacks for Fooling Deep Neural Networks. arXiv:2102.02128, 2021.

[Pin-Yu Chen 2017] Pin-Yu Chen, Zhang H., Sharma Y., Yi J., Hsieh Cho-Jui. Zoo: Zeroth Order Optimization based Black-box Attacks to Deep Neural Networks without Training Substitute Models. arXiv:1708.03999, 2017.

[Lin Tsung-Yi 2015] Lin Tsung-Yi, Maire M., Belongie S., Bourdev L., Girshick R, Hays J and etc. Microsoft COCO: Common Objects in Context. arXiv:1405.0312, 2015.

[Vargas D. 2019] Vargas D. V., Su J. Understanding the One-Pixel Attack: Propagation Maps and Locality Analysis. arXiv:1902.02947, 2019.