Computer Science and Application 计算机科学与应用, 2019, 9(12), 2239-2248
Published Online December 2019 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2019.912249

Hans汉斯

Artificial Intelligence Security

—Analysis on Adversarial Attacks

Kaifan Yi[1,2], Qian Shao[1], Min Chen[2]

[1]Shanghai Jiao Tong University Affiliated High School, Shanghai
[2]Shanghai Jiao Tong University Cyberspace Security Practice Office, Shanghai
Email: yiping@sjtu.edu.cn

Received: Nov. 13th, 2019; accepted: Nov. 25th, 2019; published: Dec. 2nd, 2019

Abstract

With the rapid development of artificial intelligence and its wide application, artificial intelligence security has also begun to attract people's attention. Attackers have added subtle disturbances in normal samples, resulting in errors in the classification and judgment of artificial intelligence deep learning models. It is called adversarial sample attacks. This paper reviews the research status of adversarial sample attacks, and studies the classic algorithms on adversarial sample attacks: FGSM, DeepFool, JSMA, CW. And the paper analyzes the efficiency of these classic attack algorithms and their misleading effect on deep learning model, in order to provide theoretical guidance for the design of adversarial sample detection and defense algorithms.

Keywords

Artificial Intelligence Security, Deep Learning, Adversarial Attacks

人工智能安全

——对抗攻击分析

易楷凡[1,2]，邵　倩[1]，陈　敏[2]

[1]上海交通大学附属中学，上海
[2]上海交通大学网络空间安全实践工作站，上海
Email: yiping@sjtu.edu.cn

收稿日期：2019年11月13日；录用日期：2019年11月25日；发布日期：2019年12月2日

摘　要

随着人工智能的迅速发展及其广泛应用，人工智能安全也开始引起人们的关注，攻击者在正常样本中增

文章引用: 易楷凡, 邵倩, 陈敏. 人工智能安全[J]. 计算机科学与应用, 2019, 9(12): 2239-2248.
DOI: 10.12677/csa.2019.912249

加了细微的扰动，导致人工智能深度学习模型分类判断出现错误，这种行为称为对抗样本攻击。该文综述对抗样本攻击的研究现状，研究了对抗样本攻击的经典算法：FGSM、DeepFool、JSMA、CW，分析了这几种经典对抗算法的生成对抗样本的效率及其对深度学习模型的误导效果，为对抗样本检测和防御算法设计提供理论指导。

关键词

人工智能安全，深度学习，对抗攻击

1. 引言

随着人工智能深度学习成为研究热点，其在医疗[1]、生物[2] [3]，金融[4]、自动驾驶[5]各个领域皆有所应用，并且取得丰硕的成果。深度学习不同于传统的基于特征提取的机器学习，不需要使用者掌握太多的专业知识，只需调节好参数即可取得很好的效果。人工智能深度学习应用极其广泛，此时对人工智能安全的需求也迫在眉睫。早在 2015 年 IanGoodfellow 提出了对抗攻击的概念，对于熊猫图片只需增加一点扰动，在人类视觉中干扰图片与原图差别很小，几乎无法看出，但是人工智能深度学习模型却会以 99.3%的概率将其错判为长臂猿[6]，如图 1。对抗样本对于人脸识别也能导致错误分类造成影响。

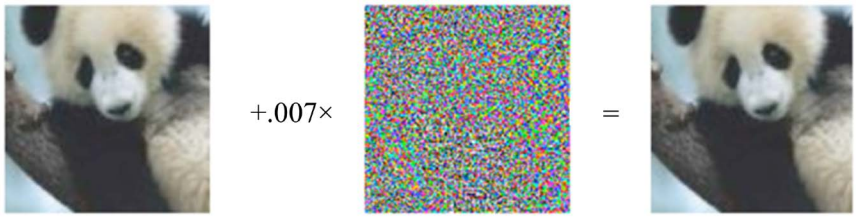

Figure 1. Adversarial Sample
图 1. 对抗样本

近年来，对抗样本攻击成为了研究热点，在深度学习的广泛研究与应用中，对抗攻击对于深度学习的安全威胁也愈发严重。2019 人工智能顶级会议 CVPR 上发表论文[7]，表明对抗攻击可以让一个人在摄像头面前几乎隐形；科恩实验室通过研究发现[8]，在路面部署对抗样本干扰后，可导致车辆经过时对车道线做出错误判断，致使车辆驶入反向车道，对抗样本攻击造成了很大的安全威胁。因此对抗样本的攻击算法与特征的研究刻不容缓。

目前，对抗样本攻击集中于图像识别领域，有许多种对抗样本算法被提出与应用。在多数攻击算法生成的图像中，有多处明显特征：对于图像的处理程度微小，人类几乎无法分辨是否经过处理；攻击算法具有很强的迁移性，针对部分网络结构有效的攻击，在别的网络结构中也具有一定的欺骗效果[9]。

现有的对抗样本攻击大都基于图像模型的分类的问题中，本文主要介绍对抗样本的内容概念与其算法分类，通过研究同种算法不同攻击步长与多种算法对于同一图片的对抗攻击的方式分析主流对抗样本攻击方法的异同，为对抗样本检测与防御研究提供理论指导。

2. 对抗样本简介

首先明确定义，当采用深度学习的方式完成一项分类任务时，需要使用分类器；分类器的定义：利用给定的类别已知的训练数据来学习分类规则和分类器，然后对未知数据进行分类。对抗样本是一种通过指定算法进行处理的图片，在原始样本加入部分扰动，使得分类器改变对于原有的样本的分类[10]。对抗样本产生于训练时的样本的覆盖率的缺陷，识别过程中，训练集不可能完全覆盖所有可能，往往测试集与训练集有部分重复，结果产生了很小几率的分类错误。在分类过程中，基于对抗样本所具有的迁移性，其对于很多网络结构都具有一定欺骗性。

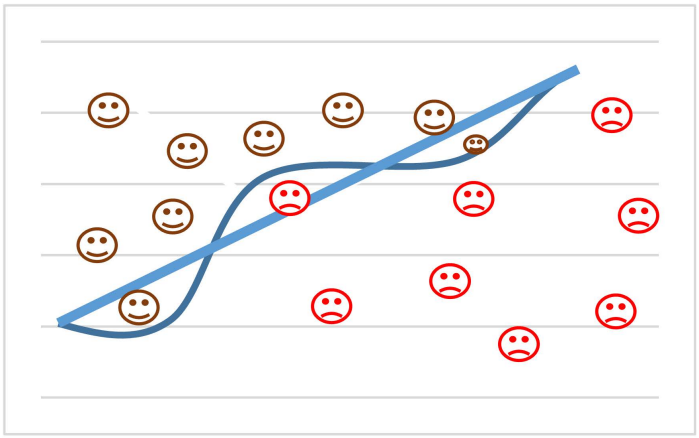

Figure 2. The reason of adversarial sample
图 2. 对抗样本产生的原因

如图 2 所示，笑脸与哭脸是两种样本，因为对抗样本的训练(直线)与真实(曲线)有微小的差别，其中有部分“笑脸”进入了“哭脸”的分类，反之亦然。这中间的“灰色地带”被称为对抗样本的空间区域。而对抗样本算法则是要迅速寻找其中的对抗样本区域样本，利用其细小的扰动，影响最终识别结果，导致识别错误，例如：哭脸进入直线左上区域。

3. 对抗样本的生成算法

本节介绍在深度学习领域中经典的对抗样本攻击生成算法，在新的神经网络模型提出后，同样新的攻击方法不断涌现。本节通过对四种方法的介绍，并进行相互的比较，分析得出具体结论。

3.1. 快速梯度攻击

IanGoodfellow 提出[6]快速梯度攻击(Fast Gradient Sign Method, FGSM)，寻找算法模型梯度变化最大的方向如图 3，在这个方向上增加图像扰动，从而导致网络进行错误的分类。

如图 4 所示，这是一个基于二维向量的模型，在这个上升的函数中，红色箭头表示其中梯度最大的一个方向。这是一个简单的快速梯度模型，其中斜率最大的方向在网络模型中很容易寻找。而快速梯度算法则是在这个方向上做出变化，当 x 值在正方向中改变极小的数值时(对抗样本的扰动)，y 值的改变量极大(分类依据)。

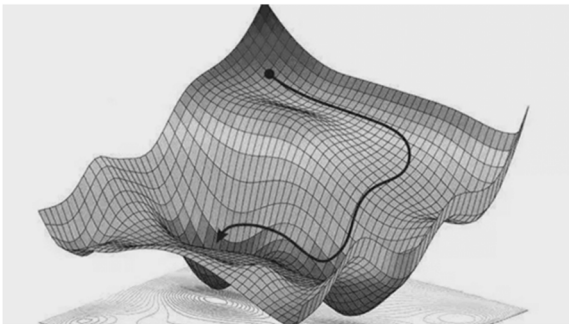

Figure 3. Fast gradient descent
图 3. 快速梯度下降

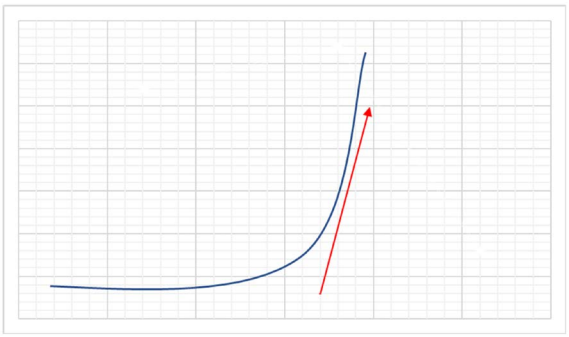

Figure 4. Two-dimensional gradient diagram
图 4. 二维梯度示意图

FGSM 攻击属于白盒攻击，需要得到模型的结构信息，可以实现源/目标误分类的攻击方式。FGSM 在设计之后，产生出了许多以 FGSM 为基础的对抗样本攻击方式。文献[11]提出一种迭代化的 FGSM/the Basic Iterative Method (BIM)。FGSM 在迭代过程中，在梯度上升程度最大时产生一次扰动；BIM 在迭代过程中，沿梯度最大处添加许多步小扰动，并且在每处扰动之后，重新计算梯度。BIM 较 FGSM 可以更加精确地对抗攻击，略去了少数无关扰动，所需计算量与时间也大幅增加[12]。其后 RAND-FGSM 攻击算法也被提出[11]，首先在梯度计算过程前，在样本中添加扰动。通过计算前添加部分扰动可以避免初始值产生时梯度过大，以前的对抗样本攻击算法只适用于某一种模型。在 RAND-FGSM 处理后对抗样本更具有普遍性，适用于更多网络模型。

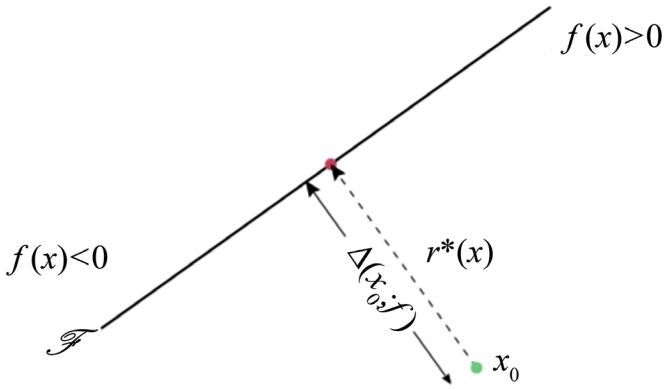

Figure 5. DeepFool attack algorithm schematic [13]
图 5. DeepFool 攻击算法原理图[13]

DeepFool [13]同样是一种基于快速梯度的白盒攻击算法。如图 5 所示，改变分类的方法是通过计算出该样本距离分类边界的最小距离。以点到直线距离公式为例每次迭代时，该点都以很小的移动距离不断逼近分割平面。针对非线性的函数，DeepFool 通过在每一次迭代，在迭代点处对函数做线性逼近，将问题简化为线性函数攻击的问题。在实验结果所示，DeepFool 算法能够得出 FGSM 较更小的扰动，而欺骗率大致相同。

3.2. JSMA 攻击

JSMA 攻击是对原图添加有限个数像素点的扰动。文献[14]中作者介绍了一种非循环前馈神经网络的攻击方法—JSMA (Jacobian Saliency Map)。该方法利用雅克比矩阵，计算从输入到输出的特征，因此只修改小部分的输入特征就能达到改变输出分类的目的[14]。通过计算前向导数，构建特征图，从而确定输入与对抗样本攻击之间的关联性。通过这种方法扰动输入特征，只需要选择数值大的像素点进行扰动，可以很快达到攻击者误分类的目的。下图 6 中显示了 JSMA 算法扰动数从左至右逐渐增加，其对于图像整体扰动较小，在部分点像素附件增加了较大的扰动。

Figure 6. The result of JSMA attack
图 6. JSMA 攻击效果图

3.3. CW 攻击

CW 攻击是一种基于优化的攻击[15]，它在本文介绍的 4 种对抗攻击中产生的对抗样本是与原样本公认差距最小的一种，且其攻击力被认为最强。其实质是在迭代过程中，将对抗样本和原样本的分类的可区分性所结合作为新的优化目标，将迭代与优化的过程结合。如 CW 攻击算法公式[15]所示，c 为引入的一个参数，CW 算法中采用二分法查找 c 最小的数值。c 值用于定义损失函数之间的关系。同理，k 值是置信度，当 k 值增加，样本分类出错概率增加；此时带来的影响是计算量增加，迭代时间增加。

$$r_n = \frac{1}{2}\left(\tanh\left(\omega_n\right)+1\right) - X_n$$

$$\min_{\omega_n} \left\|r_n\right\| + c \cdot f\left(\frac{1}{2}\left(\tanh\left(\omega_n\right)+1\right)\right)$$

$$\text{Where } f\left(x'\right) = \max\left(\max\left\{Z\left(x'\right)_i : i \neq t\right\} - Z\left(x'\right)_t, -k\right)$$

对上述分析的 4 种经典对抗样本生成算法进行原理分析，表 1 是 4 种攻击算法的对比分析，其中攻击强度表示攻击算法生成攻击样本与原样本的差异大小，攻击强度越大的算法生成的对抗样本与原样本差异越小。

4. 算法实验及对比分析

本文实验环境采用英伟达 TITAN V 深度学习 GPU，操作系统采用 Linux，编程语言采用 Python3，深度学习框架采用 PyTorch，人工智能对抗攻击工具包采用 AdvBox [16]，实验测试数据集采用 ImageNet [17]，其中实验测试样本熊猫图片如图 7。

Table 1. Comparison of adversarial sample attack algorithms
表 1. 对抗样本攻击算法对比

攻击方法	学习方式	攻击强度	攻击算法特点
快速梯度攻击	单步、迭代	***	对图像添加整体扰动，生成效率高
JSMA 攻击	迭代	**	对像素点的扰动
DeepFool 攻击	迭代	**	添加的对抗样本的扰动幅度较小
CW 攻击	迭代	****	基于优化，可以调节置信度，产生扰动最小，时间最长

Figure 7. The picture of adversarial sample
图 7. 对抗样本实验图片

我们进行了两组实验测试，第一组分析了 FGSM 算法攻击参数与程序运行时间、攻击成功率的关系，第二组实验对比分析了 FGSM、DeepFool、JSMA、CW 四种攻击算法的运算时间。

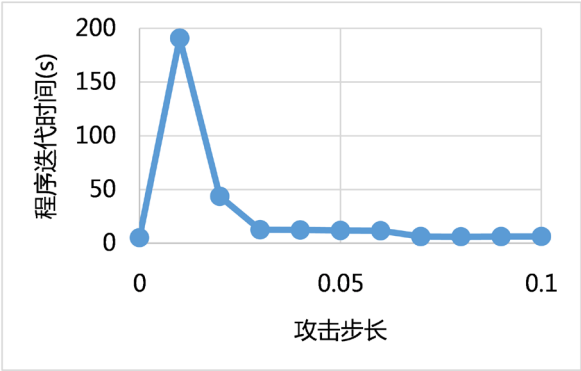

Figure 8. The picture of adversarial sample
图 8. 对抗样本实验图片

如图 8，比较了 FGSM 算法的攻击步长与迭代时间之间的关系。其中 x 轴以攻击步长，y 轴。实验中每次程序迭代生成 50 个对抗样本，统计程序总迭代时间。在攻击步长从 0 增长到 0.01 的过程中，程序迭代所需时间较长；在攻击步长从 0.01 增长到 0.1 的过程中，程序迭代时间逐渐减小；在攻击步长从 0.03 增长到 0.1 的过程中，程序迭代时间保持在 10 秒以下的较快水平。

攻击成功表明生成的对抗样本导致深度模型识别出错，如图 9，比较了攻击成功数量(每 50 个)与攻击步长的关系，如图 10，比较了成功率与攻击步长的关系。随着攻击步长从 0 增长到 0.02 的过程中，成功率与欺骗数量都有较大幅度的增长，成功率与欺骗数量的增长比例大致相同；随着攻击步长从 0.03 增长到 0.1 的过程中，成功率与欺骗数量均保持最大值左右。伴随着攻击步长(0~0.1 范围)的增长过程中，所需时间先呈增长趋势，随后呈下降趋势直至最低，而欺骗数量与成功率始终呈现上升趋势直至最大。可知攻击步长调整在 0.03 以上，对于 FGSM 算法攻击效率较高，成功率较大。

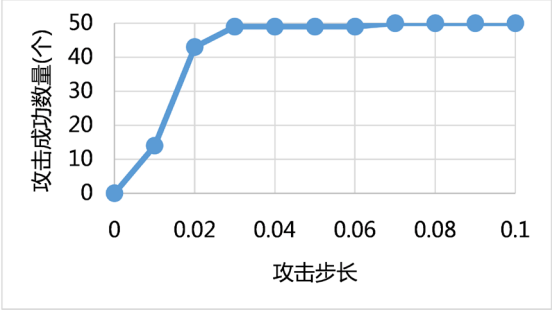

Figure 9. The picture of adversarial sample
图 9. 对抗样本实验图片

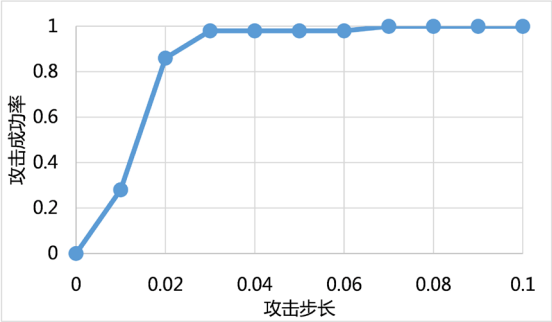

Figure 10. The relationship diagram of attack step size and attack success rate
图 10. 攻击成功率与攻击步长关系图

FGSM 增加对图像分类器损失对图像添加扰动。其以效率较高为优势，生成的图像以全图扰动为主。下图 11 所示 FGSM 算法对于 MNIST 数据集所增加扰动后，生成对抗样本的效果图。

Figure 11. FGSM attack effect in MNIST data set
图 11. 在 MNIST 数据集中 FGSM 攻击效果图

第二种测试在相同实验条件下，测试了 FGSM、DeepFool、JSMA、CW 4 种经典的对抗攻击方法针对图[9]产生对抗样本的时间和 L_2 范数。由图 12 可得，FGSM 攻击的速度最快，据算法可初步得到：FGSM 效率最高，因其通过研究者手动调节攻击步长的方式，不需计算攻击步长；DeepFool 运算时间较长，因其算法是计算得出最小攻击步长，而消耗了部分时间；JSMA 运算时间较长，因其构建特征图需要消耗部分时间；CW 运算时间最长，较其他算法所消耗时间的 7 到 10 倍左右，因其算法是迭代与优化的结合，计算量较大，效率较低。

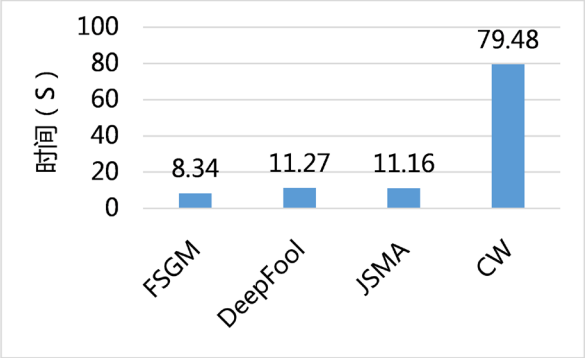

Figure 12. Four attack algorithm operation time comparison
图 12. 四种攻击算法运算时间对比图

计算 L_2 范数，比如欧式距离就是目前常用的一种 L_2 范数，表示为向量元素的平方和再开方，其公式为 $\|x\|_2 = \sqrt{\sum_i x_i^2}$。在对抗攻击中，$L_2$ 范数表示对抗样本相对于原始图片所修改的像素变化量的平方和再开方，通俗地说就是对抗样本与原图的差距，数值越小，对抗样本越难以被检测到并且防御。由图 13 可知，JSMA 的 L_2 范数较大，与原图增加的局部像素点扰动较为明显；FGSM、JSMA、CW 的 L_2 范数都较小，产生的扰动量较小。

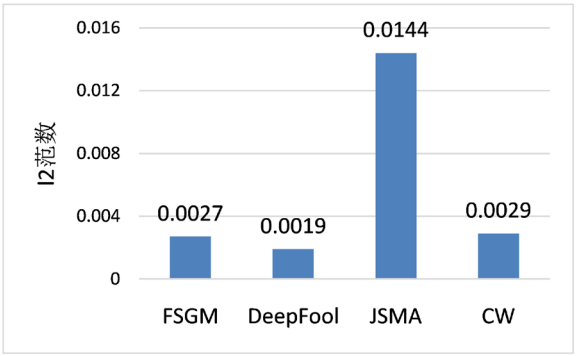

Figure 13. Comparison of L_2 norm of four attack algorithms
图 13. 四种攻击算法的 L_2 范数对比图

5. 结论

本文介绍了对抗样本概念与分类，对比分析了四种经典对抗样本攻击算法。通过实验得出结论，FGSM 效率最高，对抗样本与原图的差距较小，后续研究者在检测与防御中应着重处理有关 FGSM 算法的攻击；CW 算法效率较低，对抗样本与原图的差距较小，难以被人眼识别，且具有较强的迁移性[9]，后续研究者在检测与防御中应注重 CW 攻击的威胁。本论文应对当前对抗攻击方法较多且较缺乏统一的

对比分析与归纳，使得检测与防御对抗攻击的研究者难以权衡研究对象，所以设计并制作了具有针对性的比较实验，为后续研究者提出检测与防御算法所应该针对哪些攻击性较强的对抗攻击提供方向，对人工智能安全的攻击与防范进行具体的理论指导。本文实验分析主要针对 MNIST 数据集，今后可以拓展到 Cifar 和 ImageNet [18]等多个数据集，至今研究者已经提出了多种的攻击、检测和防御方法[19] [20]，在人工智能时代之际，对抗攻击所产生的安全威胁逐渐显露，研究人工智能安全与对抗攻击迫在眉睫；与此同时伴随对抗攻击的检测与防御也是未来所应当发展的领域。

致　谢

感谢上海交通大学网络空间安全实践工作站提供科学研究环境。特别感谢实践工作站陈敏站长的帮助，在课题研究时提供研究方向，并指导研究方法。

感谢上海交通大学网络空间安全学院研究生黄程与张维同学的帮助。黄程引导本人入门并了解深度学习网络模型与基础算法，搭建实验环境。张维帮助本人解决实验中所遇到的困难，并提供测试指导。

参考文献

[1] Goodfellow, I., Yoshua, B. and Aaron, C. (2016) Deep Learning. MIT Press, Boston.

[2] Webb, S. (2018) Deep Learning for Biology. *Nature*, **554**, 555-557. https://doi.org/10.1038/d41586-018-02174-z

[3] Branson, K. (2018) A Deep (Learning) Dive into a Cell. *Nature Methods*, **15**, 253-254. https://doi.org/10.1038/nmeth.4658

[4] Deng, Y., Bao, F., Kong, Y.Y., *et al*. (2017) Deep Direct Reinforcement Learning for Financial Signal Representation and Trading. *IEEE Transactions on Neural Networks and Learning Systems*, **28**, 653-664. https://doi.org/10.1109/TNNLS.2016.2522401

[5] He, Y., Zhao, N. and Yin, H.X. (2018) Integrated Networking, Caching, and Computing for Connected Vehicles: A Deep Reinforcement Learning Approach. *IEEE Transactions on Vehicular Technology*, **67**, 44-55. https://doi.org/10.1109/TVT.2017.2760281

[6] Goodfellow, I., Shlens, J. and Christian, S. (2015) Explaining and Harnessing Adversarial Examples. https://arxiv.org/abs/1412.6572

[7] Thys, S., Van Ranst, W. and Goedemé, T. (2019) Fooling Automated Surveillance Cameras: Adversarial Patches to Attack Person Detection. https://arxiv.org/pdf/1904.08653.pdf

[8] Tencent Keen Security Lab. (2019) Experimental Security Research of Tesla Autopilot.

[9] https://keenlab.tencent.com/en/whitepapers/Experimental_Security_Research_of_Tesla_Autopilot.pdf.

[10] Papernot, N., Mcdaniel, P., Goodfellow, I., *et al*. (2016) Practical Black-Box Attacks against Machine Learning. https://arxiv.org/abs/1602.02697

[11] Kurakin, A., Goodfellow, I. and Bengio, S. (2018) Adversarial Examples in the Physical World. https://arxiv.org/abs/1805.10997

[12] Huang, S., Papernot, N., Goodfellow, I., Duany, Y. and Abbeel, P. (2017) Adversarial Attacks on Neural Network Policies. https://arxiv.org/abs/1702.02284v1

[13] Tramer, F., Goodfellow, I., Boneh, D., *et al*. (2017) Ensemble Adversarial Training: Attacks and Defenses. https://arxiv.org/abs/1705.07204

[14] Moosavidezfooli, S., Fawzi, A. and Frossard, P. (2015) DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. https://arxiv.org/abs/1511.04599

[15] Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., BerkayCelik, Z. and Swami, A. (2016) The Limitations of Deep Learning in Adversarial Settings. *IEEE European Symposium on Security and Privacy*, Saarbrücken, 21-24 March 2016, 372-387. https://doi.org/10.1109/EuroSP.2016.36

[16] Nicholas, D.W. (2017) Towards Evaluating the Robustness of Neural Networks. https://arxiv.org/pdf/1608.04644.pdf

[17] Baidu xlab. AdvBox. https://github.com/baidu/AdvBox

[18] Stanford Vision Lab. ImageNet. http://www.image-net.org

[19] Fawzi, A., Fawzi, O. and Frossard, P. (2015) Fundamental Limits on Adversarial Robustness.

http://www.alhusseinfawzi.info/papers/workshop_dl.pdf

[20] Guo, C., Rana, M., Cisse, M. and Maaten, L. (2018) Countering Adversarial Images Using Input Transformations. https://arxiv.org/abs/1711.00117